

# **Inaugural-Dissertation**

zur

**Erlangung der Doktorwürde**

der

**Gesamtfakultät für Mathematik, Ingenieur- und  
Naturwissenschaften**

der

**Ruprecht-Karls-Universität Heidelberg**

vorgelegt von

**Jan Sellner, M. Sc.**

aus Lindau (Bodensee)

Tag der mündlichen Prüfung: \_\_\_\_\_





# **Generalizable Surgical Scene Segmentation of Hyperspectral Images**

Supervisor: Prof. Dr. Lena Maier-Hein

---

# ABSTRACT

---

Recently, it has been identified that complications following surgery contribute to the third leading cause of death globally. One of the significant challenges surgeons face is the visual discrimination of tissue types. Automatic surgical scene segmentation with hyperspectral imaging (HSI) could offer valuable assistance in this regard. However, the current state-of-the-art in this field has primarily focused on conventional RGB videos with limited spectral information, mostly from minimally invasive surgery, while HSI data and data obtained during open surgery have received little attention. Moreover, work in this area is constrained by small datasets, studies with only a few subjects or a limited number of tissue types. While deep learning-based scene segmentation is promising, it does not come without its own challenges. The generalizability of the models toward unknown data distributions, the robustness to variations in the surgical scene and the efficiency of the training process remain open questions. Consequently, the goal of this thesis is to overcome the problems in this field.

Firstly, we analyze the high-dimensional spectral information to gain a deeper understanding of the spectral characteristics and variability of different groups for various tissue types. Leveraging a tissue atlas of unprecedented size, which is comprised of 9057 images from 46 subjects annotated with 20 classes, we demonstrate that fully automatic tissue discrimination using a deep neural network is feasible with high accuracy of 95.4% (standard deviation (SD) 3.6%). We employ the principles of linear mixed model analysis to reveal that the most significant source of variability in spectral data is the tissue under observation rather than specific acquisition conditions. While recognizing the need within the HSI community for large open datasets, we make a portion of our data publicly available.

Secondly, it is necessary to train numerous networks during development to tackle a segmentation task. However, networks trained on HSI data are slow due to the large number of spectral channels which leads to data loading bottlenecks resulting in long training runs, low utilization of the graphics processing unit (GPU) and delayed inference. To address this, we are conducting a benchmark between various strategies to speed up the data loading including the introduction of a new concept to optimize the transfer from the random-access memory (RAM) to the GPU. By combining all strategies, we achieve a speedup of up to 3.6 and nearly saturated GPU utilization.

---

Thirdly, equipped with an optimized training pipeline, we are tackling the task of robust surgical scene segmentation. Given the predominance of RGB data, we compare the benefit of HSI data to RGB data and to processed HSI data (e.g., tissue parameters like perfusion). The community has not converged to the optimal input representation of HSI data for a neural network which is why we explore the best input representation considering the spatial granularity of the input data (pixels vs. superpixels vs. patches vs. images). Through a comprehensive validation study involving 506 images from 20 subjects fully semantically annotated with 19 classes, we discover that HSI data outperforms RGB and processed HSI data across all spatial granularities. Moreover, the advantage of HSI increases with decreased spatial granularity. Our image HSI model consistently ranks first in our study achieving an average dice similarity coefficient (DSC) of 0.90 (SD 0.04). This segmentation score is on par with the inter-rater variability with an average DSC of 0.89 (SD 0.07).

Fourthly, even though machine learning models have proven to be powerful, they are also known to face generalization issues if applied to out-of-distribution (OOD) data. Therefore, we are conducting a generalizability assessment for the subject (variations induced by individuals), context (variations due to geometrical changes in the neighborhood) and species (variations when moving from one species to another) domain shifts. We find that the subject domain has only a minor impact on both the spectra and the image level. On the other hand, contextual changes significantly deteriorate the segmentation performance with a drop of the DSC up to 0.48 (SD 0.38) revealing the struggles of neural networks with geometrical OOD data. To address this important bottleneck, we propose a simple, network-independent organ transplantation augmentation achieving a DSC of up to 0.91 (SD 0.10) bringing the segmentation performance on par with in-distribution data. This result is backed up through a validation study involving 600 fully semantically annotated images from 33 subjects and a comparison with other topology-aware augmentations where our proposed augmentation always ranks first. For the species domain, we utilize a large human dataset, comprising 777 images from 230 subjects fully semantically annotated with 16 classes, to demonstrate that segmentation on human data is more challenging than on porcine data and that the inclusion of porcine data in the training process offers no direct benefit.

In conclusion, we are the first to present fully semantic scene segmentation networks operating on HSI data that can differentiate between 19 classes occurring during open surgery, can be trained efficiently and are robust against contextual domain shifts. Our results are substantiated by extensive validation studies with several large datasets, some of which are publicly available as part of our open data efforts. Thereby, we made a valuable contribution to the broader goal of improving surgical interventions by leveraging the potential of HSI data with the power of machine learning algorithms. The code for all the experiments of this thesis as well as pretrained models are available at [github.com/IMSY-DKFZ/htc](https://github.com/IMSY-DKFZ/htc).

# ZUSAMMENFASSUNG

---

Komplikationen, die nach chirurgischen Eingriffen auftreten, tragen zur dritthäufigsten Todesursache weltweit bei. Eine der größten Herausforderungen der Chirurgen ist dabei die visuelle Unterscheidung von Gewebetypen. Die automatische Segmentierung chirurgischer Bilder mittels hyperspektraler Bildgebung (HSI) könnte sich hier als Schlüsseltechnologie erweisen. Der aktuelle Stand der Technik konzentriert sich jedoch auf RGB-Videos, welche nur über begrenzte Spektralinformationen verfügen und zudem meist aus minimalinvasiven Eingriffen stammen. Dahingegen bekommen HSI-Daten und Daten aus offenen Eingriffen bisher nur wenig Beachtung. Zudem sind diese Arbeiten durch die Verwendung von kleinen Datensätzen, eine geringe Probandenzahl oder eine begrenzte Anzahl von Gewebetypen charakterisiert. Die Segmentierung mit Hilfe von neuronalen Netzwerken ist vielversprechend, bringt jedoch eigene Herausforderungen mit sich. So sind die Generalisierbarkeit der Modelle bezüglich unbekannten Datenverteilungen, die Robustheit gegenüber Variationen in den Bildern und die Effizienz des Trainingsprozesses offene Probleme. Ziel dieser Arbeit ist es, diese Probleme zu lösen.

Wir analysieren die hochdimensionalen Spektralinformationen, um ein tieferes Verständnis der spektralen Eigenschaften und der Variabilität verschiedener Gruppen bezüglich der Gewebetypen zu bekommen. Mit Hilfe eines großen Datensatzes bestehend aus 9057 Bildern (annotiert mit 20 Klassen) von 46 Individuen, zeigen wir, dass eine vollautomatische Gewebeklassifizierung mit Hilfe eines neuronalen Netzwerkes eine Genauigkeit von 95.4 % (Standardabweichung (SD) 3.6 %) erreicht. Wir nutzen ein lineares gemischtes Modell, um aufzuzeigen, dass die wichtigste Quelle der Variabilität in den Spektraldaten auf das Gewebe und nicht auf die Aufnahmebedingungen zurückzuführen ist. Um den steigenden Bedarf an öffentlichen HSI-Datensätzen gerecht zu werden, machen wir einen Teil unserer Daten öffentlich zugänglich.

Für die Entwicklung eines Segmentierungsalgorithmus ist es notwendig, zahlreiche Netzwerke zu trainieren. Dabei sind Netzwerke, die auf HSI-Daten trainiert werden, aufgrund der hohen Spektraldichte ineffizient, da es zu Engpässen beim Laden der Daten kommt. Dies macht sich in langen Trainingszeiten, einer geringen Auslastung der Hardware sowie langen Prediktionszeiten bemerkbar. Um diese Engpässe zu beheben, vergleichen wir verschiedene Strategien zur Beschleunigung des Ladens der Daten und stellen dabei auch ein neues Konzept zur Optimierung des Transfers vom Arbeitsspeicher zur Grafikkarte (GPU) vor. Durch die Kombination aller Strategien erreichen wir eine 3.6-fache Beschleunigung der Trainingszeiten und eine nahezu gesättigte GPU-Auslastung.

---

Wir nutzen unsere optimierte Trainingspipeline, um eine robuste Segmentierung chirurgischer Szenen zu ermöglichen. Angesichts der Dominanz von RGB-Daten vergleichen wir den Nutzen von HSI-Daten mit RGB-Daten sowie mit verarbeiteten HSI-Daten (z. B. Gewebeparameter wie Perfusion). Da es noch unklar ist, wie HSI-Daten optimal von neuronalen Netzwerken verarbeitet werden können, untersuchen wir verschiedene Eingabedarstellungen unter Berücksichtigung der räumlichen Granularität (Pixel vs. Superpixel vs. Patches vs. Bilder). Im Rahmen einer umfassenden Validierungsstudie mit 506 Bildern (vollständig semantisch annotiert mit 19 Klassen) von 20 Individuen stellen wir fest, dass HSI-Daten RGB- und verarbeiteten HSI-Daten in allen räumlichen Granularitäten überlegen sind. Dabei vergrößert sich der Vorteil von HSI mit abnehmenden Kontext. Unser Netzwerk, welches auf HSI-Bildern trainiert wurde, belegt in unserer Studie durchweg den ersten Platz und erreicht einen durchschnittlichen Dice Ähnlichkeitskoeffizienten (DSC) von 0.90 (SD 0.04). Dies liegt im Bereich der Variabilität zwischen verschiedenen Annotatoren mit einem durchschnittlichen DSC von 0.89 (SD 0.07).

Obwohl neuronale Netzwerke sich im Allgemeinen als leistungsfähig erwiesen haben, sind sie nicht dafür bekannt, gut auf Daten aus unbekannten Verteilungen zu generalisieren. Daher analysieren wir die Einsatzfähigkeit unserer Netzwerke bezüglich drei verschiedener Bereiche: Variationen durch Individuen, Variationen durch geometrische Veränderungen und Variationen, die sich durch den Wechsel zwischen Spezies ergeben. Verschiedene Individuen haben dabei nur einen geringen Einfluss auf die Ergebnisse. Andererseits verschlechtert sich die Segmentierung erheblich, wenn Netzwerke mit geometrischen Änderungen konfrontiert werden (Abfall des DSC auf bis zu 0.48 (SD 0.38)). Wir lösen dieses Problem jedoch mit Hilfe einer einfachen und netzwerkunabhängigen Augmentierung, welche den DSC zurück auf 0.91 (SD 0.10) bringt. Dieses Ergebnis wird durch eine Validierungsstudie mit 600 vollständig semantisch annotierten Bildern von 33 Individuen untermauert. Dabei landet unsere Augmentierung im Vergleich mit anderen geometrischen Augmentierungen stets an erster Stelle. Den Wechsel der Spezies analysieren wir mit Hilfe eines großen menschlichen Datensatzes bestehend aus 777 Bildern von 230 Individuen (vollständig semantisch annotiert mit 16 Klassen). Dabei zeigen wir, dass die Segmentierung menschlicher Daten schwieriger ist und dass die Einbeziehung von Tierdaten im Training keinen direkten Vorteil bietet.

Zusammenfassend lässt sich sagen, dass unsere Segmentierungsnetzwerke erfolgreich mit HSI-Daten aus offenen Operationen umgehen und zwischen 19 Klassen unterscheiden können. Dabei lassen sich die Netzwerke effizient trainieren und sind robust gegenüber geometrischen Veränderungen. Unsere Ergebnisse werden dabei durch umfangreiche Validierungsstudien mit mehreren großen Datensätzen untermauert. Einige Datensätze haben wir auch der Öffentlichkeit zugänglich gemacht. Durch unsere Studien leisten wir einen wertvollen Beitrag zu dem allgemeinen Ziel, chirurgische Eingriffe zu verbessern, indem wir das Potenzial von HSI-Daten mit der Leistungsfähigkeit von neuronalen Netzwerken verbinden. Der Code für alle Experimente dieser Arbeit sowie die vortrainierten Modelle sind unter [github.com/IMSY-DKFZ/htc](https://github.com/IMSY-DKFZ/htc) frei verfügbar.

## ACKNOWLEDGMENTS

---

The past years working in the Division of Intelligent Medical Systems (IMSY) have been a challenging but also a very rewarding time. I am deeply grateful for the opportunity to work on my Ph.D. thesis in such an environment where I have not only learned a great deal and improved my skills but also met many wonderful people. I am a computer scientist from heart to bone and particularly value the intriguing technical challenges that I encountered while working on my projects. When implementing a solution, I do not merely stop when something “just works”; my aim is always to write code that is generalizable, maintainable and user-friendly. Furthermore, I consistently strive to find the most efficient solution for a problem and I appreciate the freedom to work on these optimizations and enhance my code quality. Some of these optimizations even made it as contributions to this thesis.

Numerous people have contributed to my Ph.D. in various ways and I would like to express my gratitude to all of them, even if they are not explicitly mentioned here. However, there are three people I would like to particularly acknowledge with honor, as without their support and contributions, there would be no thesis to read.

First and foremost, I would like to thank my supervisor, *Lena Maier-Hein*. I am extremely grateful for your supervision and the opportunity to work on my Ph.D. thesis within your group. The involvement and dedication you put into your projects, coupled with a close feedback loop, are truly admirable. You have consistently shown great engagement in my work while still giving me the freedom to explore my own ideas. Your enthusiasm and passion for my work have rubbed off on me, particularly during times when I questioned my own projects. Furthermore, I would like to extend my thanks for all the additional support you have provided, ranging from our top-notch hardware equipment to the numerous opportunities to attend international conferences. I did not take it for granted.

The second person I would like to express my gratitude to is *Alexander Studier-Fischer*, our main clinical collaborator. For an interdisciplinary topic like mine, it is crucial to have a strong partnership with the clinical side. You have been the driving force behind this partnership and I firmly believe that it is only due to our fruitful discussions and close collaboration that we achieved the results we did. It has always been a pleasure to see your engagement with our ideas, your openness to new directions and your willingness to include our feedback into your experiments. Particularly in our field, deep learning algorithms can be as powerful as they want but without the appropriate data

---

and annotations, they are useless. Therefore, I want to thank you for providing us with the necessary foundation that enabled us to build so many projects on top of it (even more than covered in this thesis).

Last but by no means least, I would like to extend my heartfelt thanks to my colleague *Silvia Seidlitz*. We have worked closely together on numerous projects, dealing with many ups and downs. You possess the strongest critical thinking skills I have ever encountered, a skill that is of tremendous value in research. Whenever I had new ideas or results to discuss, I first presented them to you and you consistently provided me with invaluable feedback and always came up with ideas to make things better. If my Ph.D. journey had been a computer game, you would have been the wise oracle having all the answers, even to questions I did not realize I should be asking. Your knowledge and skills are remarkable and offer unprecedented value to everyone who has the pleasure of working with you. So, thank you for all the time you have invested; no other person has had a greater influence on my work than you. I truly appreciate it.

Heidelberg, March 2024

Jan Sellner



# CONTENTS

---

|  |            |
|--|------------|
| <b>Abstract</b>  | <b>v</b>   |
| <b>Zusammenfassung</b>   | <b>vii</b> |
| <b>Acknowledgments</b>   | <b>ix</b>  |
| <b>1 Introduction</b>  | <b>1</b>   |
| 1.1 Motivation . . . . .   | 1          |
| 1.2 Research Questions . . . . .   | 3          |
| 1.2.1 RQ1: Do different organs feature unique spectral fingerprints? . .   | 6          |
| 1.2.2 RQ2: How can we train deep hyperspectral imaging networks efficiently? . . . . .   | 8          |
| 1.2.3 RQ3: What is the optimal spatial and spectral granularity for semantic scene segmentation in surgical hyperspectral imaging? . . . . . | 9          |
| 1.2.4 RQ4: Which are relevant domain shifts affecting the segmentation performance and can we compensate for them? . . . . .                 | 10         |
| 1.3 Outline . . . . .  | 13         |
| <b>2 Fundamentals</b>  | <b>15</b>  |
| 2.1 Medical Background . . . . .   | 15         |
| 2.1.1 Visceral Surgery . . . . .   | 15         |
| 2.1.2 Surgical Scene Understanding for Computer- and Robot-Assisted Surgery . . . . .  | 17         |
| 2.2 Physical Background . . . . .  | 18         |
| 2.2.1 Light-Tissue Interaction . . . . .   | 18         |
| 2.2.2 Hyperspectral Imaging Hardware . . . . .   | 19         |
| 2.3 Deep Learning . . . . .  | 26         |
| 2.3.1 Convolutional Neural Networks . . . . .  | 27         |
| 2.3.2 Mixed Precision Training . . . . .   | 33         |
| <b>3 Related Work</b>  | <b>47</b>  |
| 3.1 Spectral Organ Fingerprints . . . . .  | 48         |
| 3.2 Efficient Training of Hyperspectral Segmentation Networks . . . . .  | 50         |
| 3.3 Surgical Scene Segmentation . . . . .  | 51         |

|          |   |            |
|----------|---|------------|
| 3.4      | Domain Shifts in Surgical Hyperspectral Imaging . . . . .                 | 55         |
| 3.5      | Conclusion . . . . .  | 57         |
| <b>4</b> | <b>Materials and Methods</b>  | <b>59</b>  |
| 4.1      | Hyperspectral Datasets . . . . .  | 59         |
| 4.1.1    | Data Acquisition and Preprocessing . . . . .                              | 60         |
| 4.1.2    | Tissue Atlas Dataset . . . . .  | 62         |
| 4.1.3    | Semantic Porcine Dataset . . . . .  | 63         |
| 4.1.4    | Semantic Human Dataset . . . . .  | 66         |
| 4.1.5    | Dataset Features . . . . .  | 68         |
| 4.2      | Spectral Organ Fingerprints . . . . .                                     | 73         |
| 4.3      | Efficient Training of Hyperspectral Segmentation Networks . . . . .       | 76         |
| 4.4      | Surgical Scene Segmentation of Hyperspectral Images . . . . .             | 81         |
| 4.5      | Domain Shifts in Surgical Hyperspectral Imaging . . . . .                 | 87         |
| <b>5</b> | <b>Experiments and Results</b>  | <b>89</b>  |
| 5.1      | Spectral Organ Fingerprints . . . . .                                     | 89         |
| 5.1.1    | Experimental Setup . . . . .  | 90         |
| 5.1.2    | Analysis of Spectral Organ Fingerprints . . . . .                         | 90         |
| 5.1.3    | HeiPorSPECTRAL: Open Dataset for Surgical Hyperspectral Imaging . . . . . | 94         |
| 5.2      | Efficient Training of Hyperspectral Segmentation Networks . . . . .       | 100        |
| 5.2.1    | Experimental Setup . . . . .  | 100        |
| 5.2.2    | Benchmarking Data Loading Strategies . . . . .                            | 102        |
| 5.3      | Surgical Scene Segmentation of Hyperspectral Images . . . . .             | 102        |
| 5.3.1    | Experimental Setup . . . . .  | 104        |
| 5.3.2    | Analysis of Segmentation Networks . . . . .                               | 110        |
| 5.4      | Domain Shifts in Surgical Hyperspectral Imaging . . . . .                 | 127        |
| 5.4.1    | Experimental Setup . . . . .  | 127        |
| 5.4.2    | Subject Domain . . . . .  | 132        |
| 5.4.3    | Context Domain . . . . .  | 134        |
| 5.4.4    | Species Domain . . . . .  | 140        |
| <b>6</b> | <b>Discussion</b>   | <b>147</b> |
| 6.1      | Spectral Organ Fingerprints . . . . .                                     | 147        |
| 6.2      | Efficient Training of Hyperspectral Segmentation Networks . . . . .       | 148        |
| 6.3      | Surgical Scene Segmentation of Hyperspectral Images . . . . .             | 151        |
| 6.4      | Domain Shifts in Surgical Hyperspectral Imaging . . . . .                 | 158        |
| 6.5      | Technical and Clinical Challenges in Hyperspectral Imaging . . . . .      | 160        |
| 6.5.1    | Hardware Limitations . . . . .  | 160        |
| 6.5.2    | Minimally Invasive vs. Open Surgery . . . . .                             | 163        |
| 6.5.3    | Pathologies . . . . .   | 163        |

|   |            |
|---|------------|
| <b>7 Conclusion</b>                         | <b>165</b> |
| 7.1 Summary of Contributions . . . . .      | 165        |
| 7.2 Impact and Outlook . . . . .            | 167        |
| <b>A Own Contributions and Publications</b> | <b>171</b> |
| A.1 Own Contributions . . . . .             | 171        |
| A.2 Own Publications . . . . .              | 173        |
| <b>B Additional Results</b>                 | <b>179</b> |
| <b>List of Acronyms</b>                     | <b>191</b> |
| <b>List of Figures</b>                      | <b>193</b> |
| <b>List of Tables</b>                       | <b>199</b> |
| <b>List of Algorithms</b>                   | <b>201</b> |
| <b>Bibliography</b>                         | <b>203</b> |

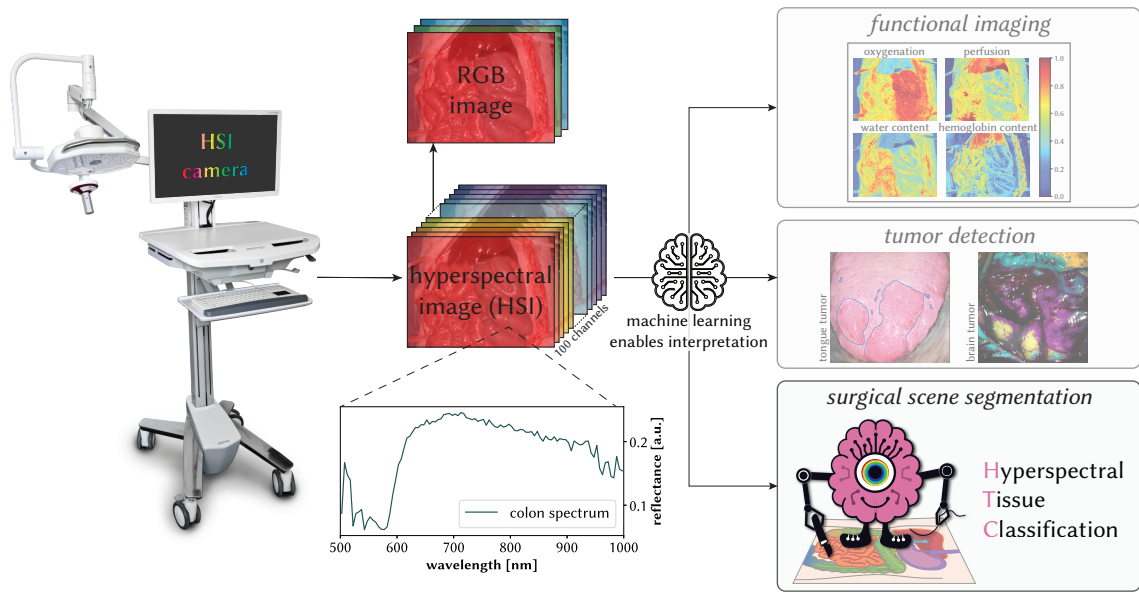


## 1.1 Motivation

Recently, complications following surgery have been identified as the third leading cause of death globally [163]. What is more, a lot of these deaths could have been avoided and are due to human errors, such as misjudgment or incomplete understanding due to missing information [67]. Surgeons face a significant challenge in visually differentiating tissues, including distinguishing between pathologies or critical structures and healthy tissue [34, 147]. Autonomous robotic surgery and robotic-guided surgery have the potential to revolutionize the field by enabling safer and more effective surgeries with higher precision, control and flexibility due to enhanced capabilities like augmented vision [24, 19]. However, for autonomous surgical robots to be successful, one prerequisite for them is to perceive and understand the surgical scene [85, 140]. Therefore, this thesis represents a step toward the direction of surgical scene understanding.

Human vision, which is one of the main sources of information for surgeons, has limitations as we humans can only perceive a broad range of the red, green, and blue spectrum. With this limited information, it can be challenging to distinguish between different tissue types even for trained surgeons [34, 147]. Conventional camera systems, which are based on the principle of human vision, also have these limitations. However, hyperspectral imaging (HSI) is an evolving technology that overcomes these arbitrary restrictions by capturing a fine-grained spectrum with more spectral channels (e.g., 100 instead of 3). HSI can even extend the captured spectrum beyond the visible range, providing additional information that can be exploited for various applications such as tumor detection. An overview of the basic concept of HSI and some exemplary medical applications can be seen in Figure 1.1.

The distinct absorbance spectra of Hb (deoxygenated hemoglobin), HbO<sub>2</sub> (oxygenated hemoglobin) and other chromophores enable the estimation of functional tissue parameters, such as oxygenation or perfusion, based on the spectral information (see Section 2.2 for more details). Generally, different tissue types exhibit unique optical properties, providing additional spectral information. This supplementary information can be utilized



**Figure 1.1:** Basic concept of hyperspectral imaging (HSI) and exemplary medical applications. Due to the high dimensionality of the spectral data, machine learning is essential for interpretation leading to a wide range of medical HSI applications like functional imaging (e.g., oxygenation and perfusion estimation), tumor detection or surgical scene segmentation. The latter is the main focus of this thesis. The depicted camera is the clinically certified system Tivita<sup>®</sup> Tissue (Diaspective Vision GmbH, Am Salzhaff, Germany) which can be used for acquiring hyperspectral images. The camera captures images with a height of 480, a width of 640 and 100 spectral channels in the range from 500 nm to 1000 nm with an approximate spectral resolution of 5 nm. The example images were taken and adapted from [76] (camera image), [132] (tongue tumor image), [178] (brain tumor image) and [200] (logo designed by Silvia Seidlitz for the hyperspectral tissue classification (HTC) framework [200]).

in a variety of applications like hemorrhagic shock diagnosis (e.g., [75, 30, 29]), detection of pathologies (e.g., [171, 7, 234]) or surgical guidance (e.g., [255, 5, 168]). In this thesis, we exploit the spectral information to differentiate between various healthy tissue types, thereby segmenting the surgical scene. While this may not be the primary application of HSI (compared to functional imaging), the technology’s availability allows for the use of applications like ours as a byproduct with minimal additional cost if the camera is already in place. [68, 251, 245]

A significant difference between RGB and HSI lies in our human ability to easily interpret RGB images while we cannot do the same for HSI data due to the high dimensionality of the latter. Consequently, it becomes essential to aggregate the spectral information in such a way that the user can readily discern the relevant information.

Over the past decade, machine learning has made remarkable strides, becoming a ubiquitous presence in many people’s lives due to its diverse applications in our everyday activities (e.g., via applications like ChatGPT) [181]. It has also become an essential tool in the medical research field due to its potential to revolutionize healthcare by enhancing diagnostics and treatment methods [179]. Specifically, the field of surgical data science has emerged with the goal of collecting, structuring and examining surgical data to enhance the quality of interventional healthcare. Within this field, semantic scene segmentation is of paramount importance since it plays a critical role in numerous tasks such as context-aware assistance and surgical robotics. [140]

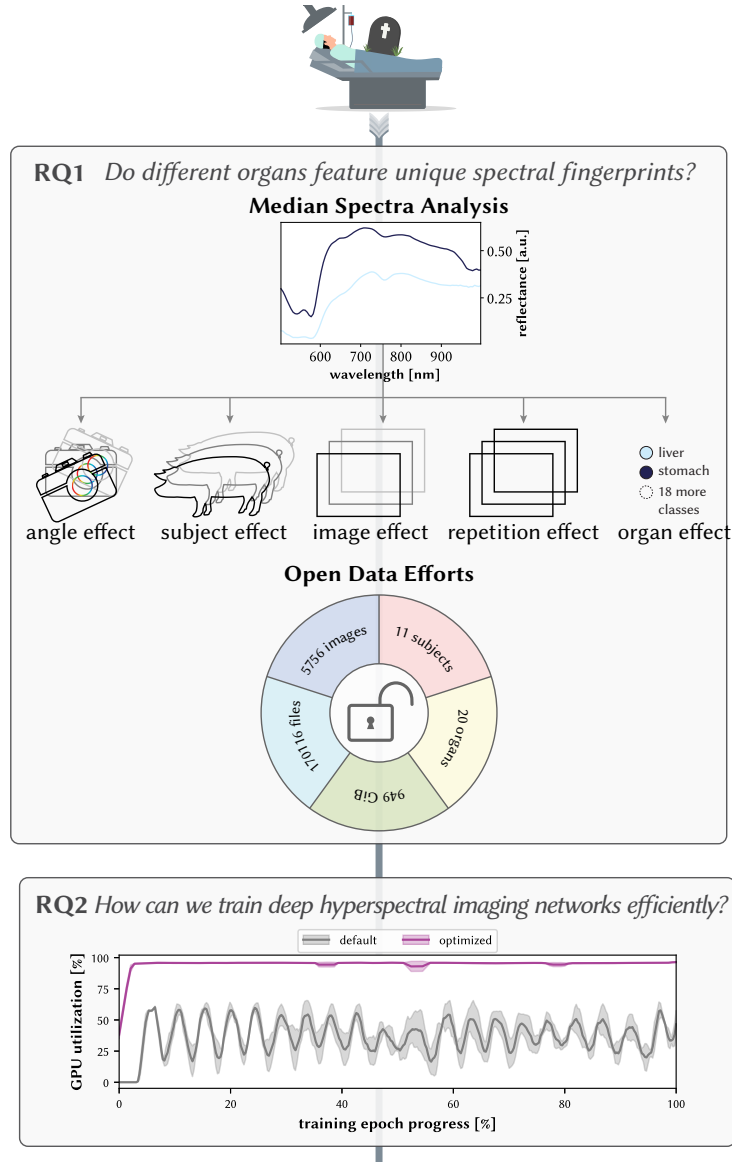
Machine learning algorithms can operate on high dimensional HSI data and hence serve as a key enabling technology for the aggregation of spectral information. These algorithms are universal and can be adapted to a wide range of downstream tasks. Hence, with the aid of machine learning, we can process all available information while only presenting surgeons with what they need in the current situation at the appropriate time.

Despite the power of machine learning algorithms, they are not without their challenges, such as issues with generalization, robustness or efficiency, particularly in the field of medical imaging [157, 160, 230, 183]. For example, this has become evident during the COVID-19 pandemic where despite numerous attempts, machine learning algorithms have not demonstrated substantial advantages [90, 188, 33]. Similar problems have existed in surgical data science where these issues limit the applicability of machine learning algorithms to real-world surgeries, for instance, if the training distribution is not comprehensive enough and misses important aspects like unknown geometries, surgery-related differences or pathologies present in real-world data [140].

In this thesis, we process HSI data cubes with deep neural networks to predict two-dimensional segmentation maps of the surgical scene. The output is an image where each pixel represents the predicted tissue class. Due to the generalization problems mentioned above, we pay special attention to the generalization capabilities of our networks.

## 1.2 Research Questions

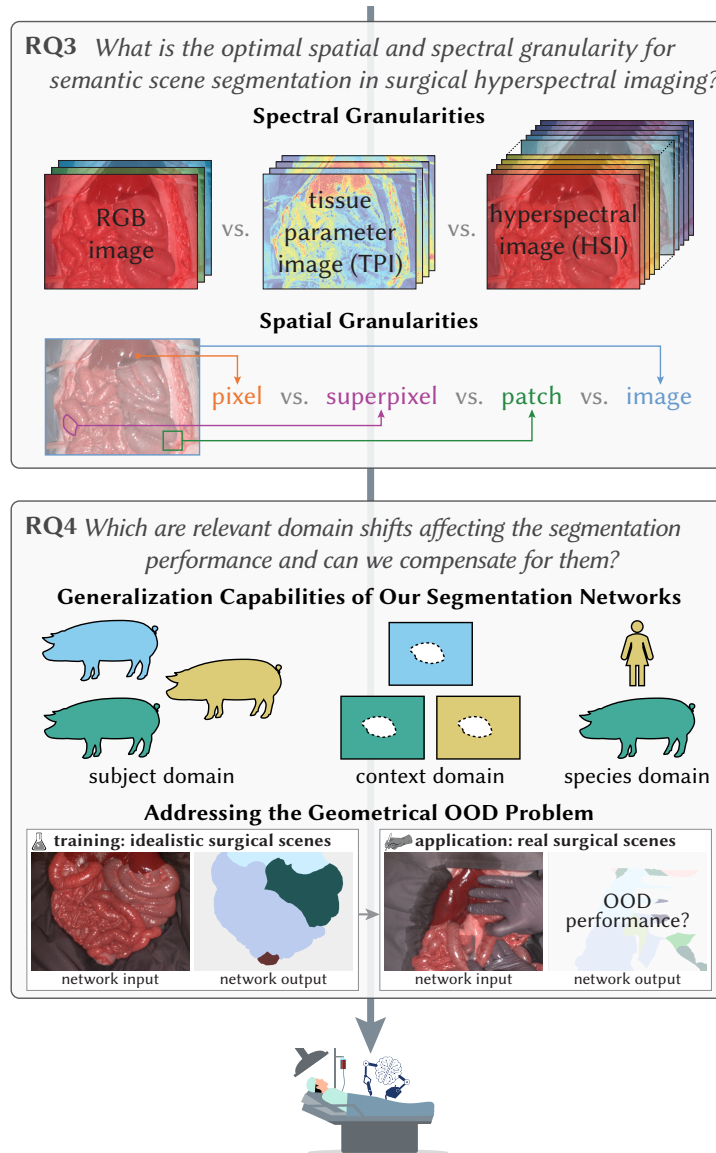
With the goal of fully automatic surgical scene segmentation with HSI for autonomous robotic surgery in mind, we break down this complex problem into fundamental research questions (RQs). As shown in the overview in Figure 1.2, we start with an analysis of individual spectra and make our dataset publicly available to the HSI community. Then, we optimize our data loading pipeline for shorter training times before we move on to the task of semantic scene segmentation of entire HSI images. Further, we challenge the generalizability capabilities of our networks against important domain shifts that are dominant in the medical field and improve the performance of our networks on geometrical out-of-distribution (OOD) scenes.



**Figure 1.2:** Tackling research questions (RQs) toward the goal of autonomous robotic surgery.

**RQ1:** We start with a spectral analysis where we find unique organ fingerprints and decompose the spectra revealing the proportion of explained variance by each shown effect. Further, we make our spectral dataset of unprecedented size available to the hyperspectral imaging (HSI) community. **RQ2:** For efficient training of our deep neural networks on HSI data, we optimize our data loading pipeline for better graphics processing unit (GPU) utilization and shorter training times. Figure continued on the next page.





**Continued Figure 1.2:** Tackling research questions (RQs) toward the goal of autonomous robotic surgery (continued). **RQ3:** With the goal of automatic scene segmentation in mind, we find the optimal spatial granularity and compare different input modalities. **RQ4:** Heading toward generalizable neural networks, we show the effect of different domain shifts and present a solution for maintained segmentation performance against geometrical out-of-distribution (OOD) data via a surgery-inspired augmentation scheme. This figure is based on [198, 215, 201, 202, 214, 200]. Surgery icons are designed by Silvia Seidlitz.

In the following, we describe each research question which this thesis aims to answer in detail. We motivate each question, highlight the used materials and give an overview of the analyses conducted.

### 1.2.1 RQ1: Do different organs feature unique spectral fingerprints?

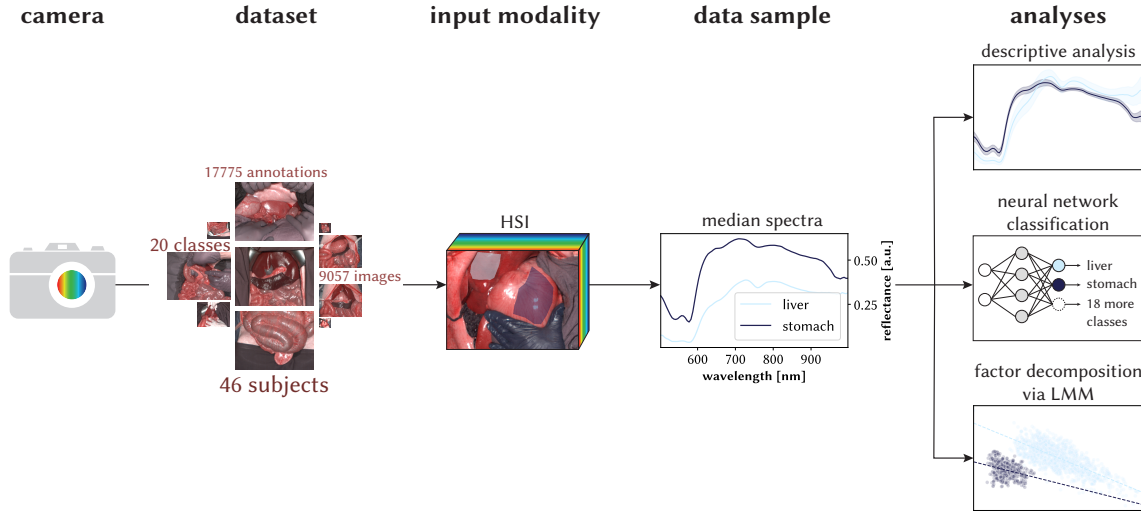
For many applications, it is important to distinguish between different tissue types (e.g., detection of pathologies) [68]. While the literature provides initial evidence that certain tissues exhibit unique spectral fingerprints, the analyses are restricted to a limited number of classes or small datasets and it is also unknown whether the variability in the spectra is due to the organ of interest or due to individual acquisition conditions like the camera angle (cf. Section 3.1).

When a neural network tackles the task of semantic scene segmentation, it typically does so by utilizing the color information (reflectance values in our case) and the neighboring relationship of the pixels. The former comprises the spectral and the latter the spatial dimension. The spectral information encapsulates the biological tissue properties reflected in the measured values while the spatial dimension describes the relationship between neighboring organs. Even though both aspects are arguably crucial for semantic segmentation, our initial focus is to gain a deeper understanding of the spectral characteristics of various tissues. Therefore, before embarking on full scene segmentation, we concentrate on the spectral dimension first. Further, we make our spectral dataset publicly available to the HSI community.

#### Median Spectra Analysis

To accomplish this, we first take a look at median spectra which are calculated over a region of pixels from the same tissue by computing the median for each channel separately. This process yields a representative spectrum of the tissue region. The use of the median operation ensures that the resulting spectra are robust to outliers making them less susceptible to noise in the data. This robustness enables us to compare the spectral fingerprints of different tissues without too much distraction from spatial relationships.

In our initial analysis, we compare the spectral fingerprints of 20 organ classes, classify them using a neural network, and decompose the fingerprints into various factors. An overview of this research question is depicted in Figure 1.3. The first part of our research question (descriptive analysis and classification) aims to determine whether the spectral fingerprints are sufficiently unique to differentiate between various tissues. The second part of our research question (factor decomposition) seeks to enhance our understanding of the spectral variation present in our data. To achieve this, we examine the variation explained by different effects in relation to the organ effect, such as the variation of the camera angle compared to the variation due to different organs.



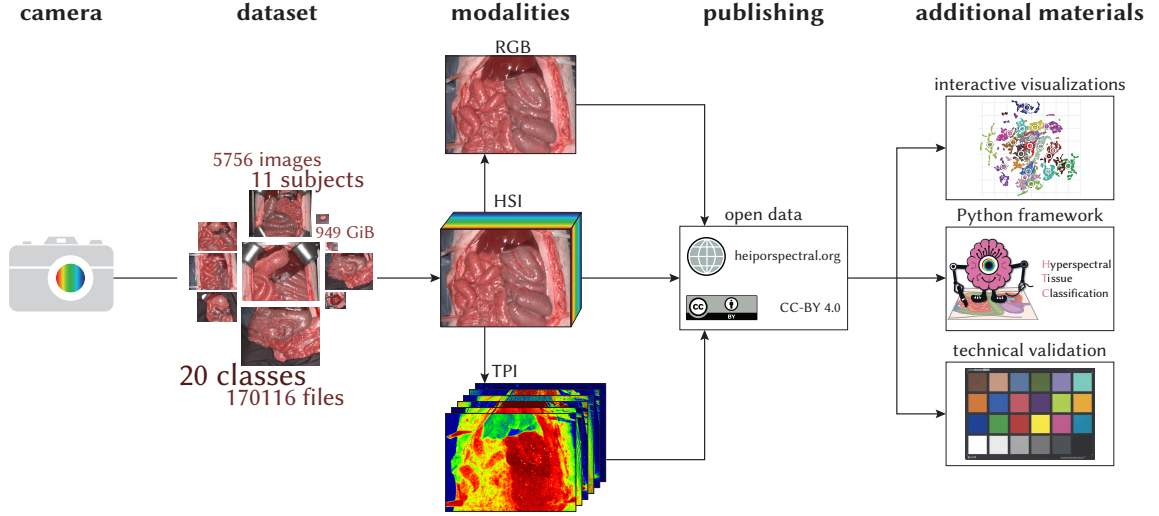
**Figure 1.3:** Overview of our spectral analysis for tissue discrimination (RQ1). We compute median spectra based on the polygon annotations of the hyperspectral imaging (HSI) data to describe the spectral characteristics, classify spectra into one of 20 organ classes and decompose factors of the hierarchical data structure with the help of a linear mixed model (LMM).

### Open Data Efforts

Public datasets hold significant importance for the scientific community. However, the HSI community has been facing a shortage of large open datasets [66, 97]. For this research question, we acquired a substantial dataset for our spectral analysis (cf. Section 5.1.2). Recognizing the need within the community, we decided to make this dataset publicly available. This allows other members of the community not only to reproduce our results but also to leverage the data for their own research endeavors.

However, our dataset is of an unprecedented scale comprising 5756 images from 11 subjects annotated with 20 classes. This presents a unique challenge as the navigation in such a large dataset can be daunting. Consequently, our efforts center on how we can make our dataset accessible to the community in a manner that is both easy to comprehend and to use.

To facilitate this, we provide various visualizations and statistics to aid users in navigating the dataset while offering insights into individual images as well as aggregated information across the entire dataset. Additionally, we have developed a Python package that allows users to easily load the HSI data, annotations as well as metadata and also contains pretrained models from our segmentation work. Further, we conduct a technical validation of the camera to ensure that our measured spectra are correct and of use to the community. An overview is shown in Figure 1.4.



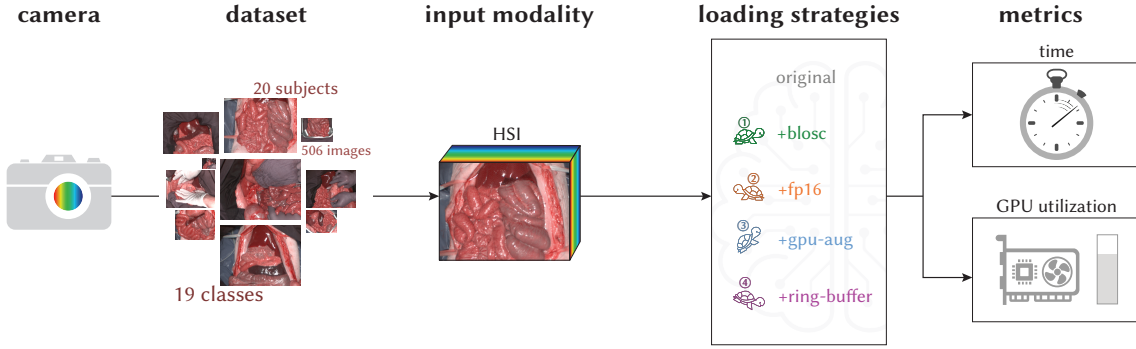
**Figure 1.4:** Overview of our open data concept for the HeiPorSPECTRAL dataset (part of RQ1). We release a dataset of unprecedented size comprising 5756 images from 11 subjects annotated with 20 classes to the hyperspectral imaging (HSI) community. We release HSI data together with corresponding RGB and tissue parameter images (TPI). What is more, we release additional materials to work with the data like interactive visualizations of aggregated data, a Python framework to load and process the data ([github.com/IMSY-DKFZ/htc](https://github.com/IMSY-DKFZ/htc) [200]) and a technical validation confirming the validity of our used camera.

### 1.2.2 RQ2: How can we train deep hyperspectral imaging networks efficiently?

Deep learning is a computationally intensive process making it crucial to optimize the use of the available hardware resources. Improved graphics processing unit (GPU) utilization can significantly reduce the cost of training as it necessitates a smaller training budget in terms of GPU hours for the same task which can also reduce the carbon footprint. Additionally, it results in shorter developer cycles due to the reduction in training time which allows for faster responses to results. While numerous strategies exist to improve the training efficiency of deep neural networks, they are neither designed nor sufficient for spectral data (cf. Section 3.2).

Our HSI data is large due to the 100 spectral channels per pixel which is why we face significant data loading challenges during training. This is because a large amount of data needs to be loaded onto the GPU while the processing itself is relatively quick due to the small networks. This situation results in suboptimal GPU utilization by default. Therefore, in this research question, we ask how can we make the data loading pipeline more efficient to improve the GPU utilization and consequently reduce the training time. To address this, we conduct a study where we compare various data loading strategies,

including a new concept to optimize the transfer from the random-access memory (RAM) to the GPU, and measure the training time per epoch as well as the graphics processing unit (GPU) utilization. An overview of this research question is shown in Figure 1.5.



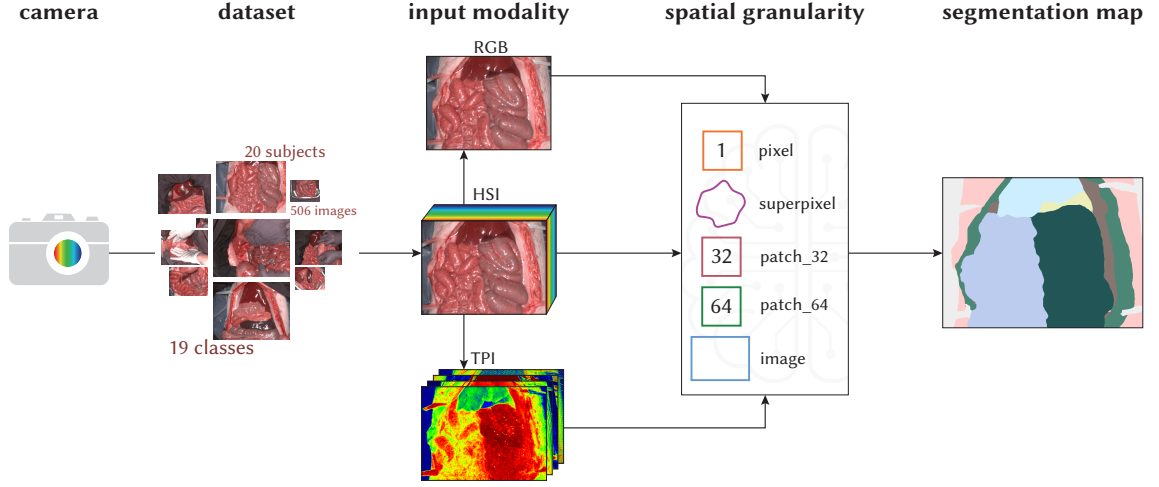
**Figure 1.5:** Overview of our data loading benchmark on segmentation networks for hyperspectral imaging (HSI) data (RQ2). Based on the HSI data, we compare different data loading strategies with respect to the time per training epoch and the graphics processing unit (GPU) utilization.

### 1.2.3 RQ3: What is the optimal spatial and spectral granularity for semantic scene segmentation in surgical hyperspectral imaging?

The literature employs a variety of different input representations for HSI data in neural networks and has not yet converged to the optimal input representation for semantic scene segmentation and no previous work has demonstrated a clear advantage of HSI data compared to RGB data (cf. Section 3.3).

Therefore, we take on the task of semantic scene segmentation while being equipped with an optimized training pipeline (cf. Section 5.2) and motivated by our promising results on the simplified task of median spectra classification (cf. Section 5.1.2). An overview of this research question is shown in Figure 1.6.

The surgical field is predominantly characterized by the use of RGB images which is largely due to the availability of RGB data from minimally invasive surgeries [187]. In contrast, HSI is a relatively new modality with limited experience in the field. This naturally leads to our first part of the research question: which modality is superior for semantic scene segmentation and how large are the differences? Additionally, we compare RGB and HSI data with processed HSI data such as tissue parameter images. This comparison aims to determine the effectiveness of segmenting tissues based on functional properties like oxygenation or perfusion.



**Figure 1.6:** Overview of our analysis on segmentation networks for hyperspectral imaging (HSI) data (RQ3). We assess the performance of different modalities (RGB, tissue parameter images (TPI) and hyperspectral imaging (HSI) with 3, 4 and 100 channels, respectively) and models with varying spatial context (spatial granularities) all to predict a segmentation mask for each image. The TPI consists of four parameter maps: tissue oxygen saturation ( $\text{StO}_2$ ), near-infrared perfusion index (NPI), tissue water index (TWI) and tissue hemoglobin index (THI). This figure was adapted from [198].

For HSI data, the relationship between the spatial and spectral dimensions is more complex than for RGB data since the spectra itself already contain rich information. This raises the question of how much neighborhood should be included for optimal segmentation performance. Consequently, this leads us to the second part of our research question which is concerned with the optimal spatial granularity of the input data. For this, we explore various levels of spatial granularities such as pixels, superpixels, patches, and images with respect to the segmentation performance and analyze how the segmentation performance changes under a varying number of training subjects. It is worth noting that smaller spatial granularities naturally offer more training samples and we wanted to explore whether this has an impact on the segmentation performance especially in situations when limited training data is available.

#### 1.2.4 RQ4: Which are relevant domain shifts affecting the segmentation performance and can we compensate for them?

Even though it is well-known that neural networks can fail when applied to OOD data from domains different from the source (training) domain [157, 160], this aspect is largely unexplored in the field of surgical scene segmentation (cf. Section 3.4). Therefore, the

impact of important domain shifts for automated surgical scene segmentation with HSI data is yet to be determined and solutions for relevant domain gaps remain to be addressed.

In our previous research question about modalities and spatial granularities, we found that an image HSI model exhibits segmentation performance approaching the level of inter-rater variability (cf. Section 5.3). However, our dataset might be somewhat simplistic as it only provides a perspective on specific types of surgeries and solely on porcine data.

### **Generalization Capabilities of Our Segmentation Networks**

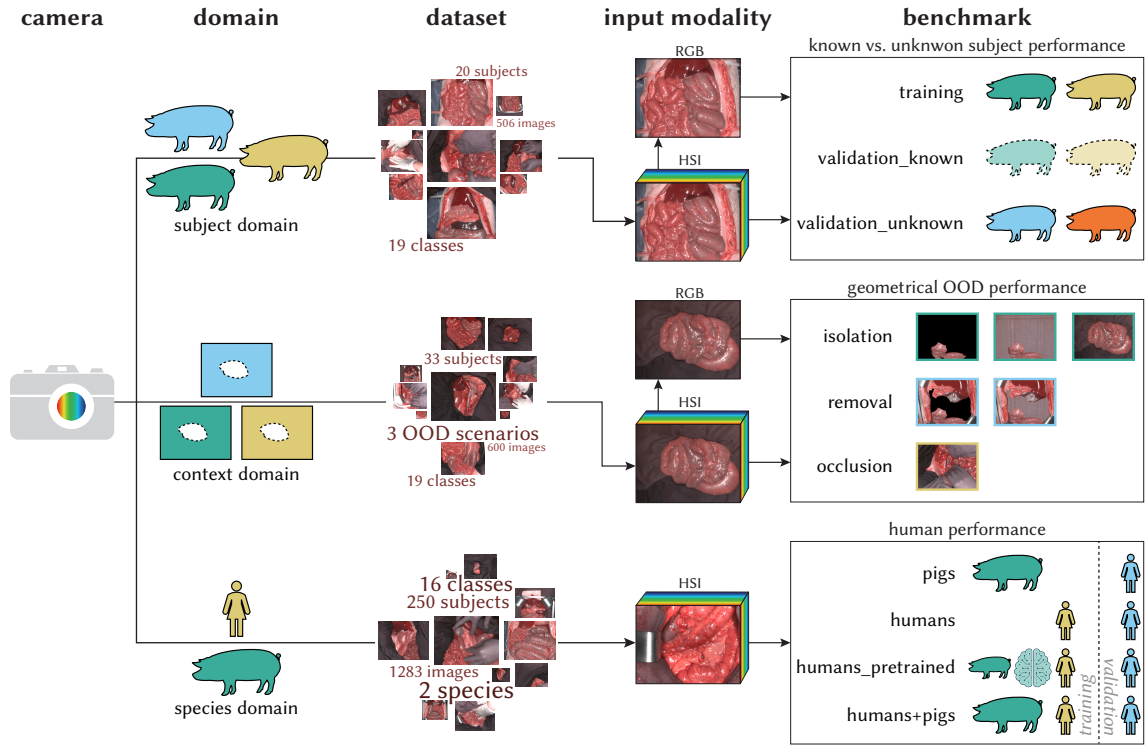
For real-world applications, it is crucial that segmentation networks are capable of generalizing well across various domains. Therefore, this research question is concerned with the impact of different domain shifts on our HSI data and on the segmentation performance of our networks when they are evaluated on these new domains. An overview of this assessment is presented in Figure 1.7.

Each subject introduces its own variation due to factors such as the inherent differences in tissues, the type of surgery or the operating surgeons. These subject-related differences may significantly influence the data. Therefore, we explicitly analyze the performance differences between images of subjects known during training and images from entirely new subjects. This analysis is conducted both at the spectra and the image level.

A segmentation network’s decision-making process is sensitive to the context of a pixel since it relies on a series of convolutional operations that operate on the pixel and its surroundings. The neighborhood of an organ is not static and can vary based on numerous factors such as situs (organ composition) occlusions, the visibility of other organs or variations in the surgical procedure. Given this, we apply our networks to datasets with diverse contextual characteristics to evaluate their performance on geometrical OOD data. This includes scenarios with missing organs, isolated organs or occlusions.

This thesis represents a preliminary step toward the goal of autonomous robotic surgery with the task of semantic scene segmentation and we have gained important insights from our work on animal data. However, the ultimate objective is to ensure that the segmentation performs effectively on human data as this is the primary focus of any real-world application. Hence, we present the first steps toward surgical scene segmentation on human data by applying our networks on a human HSI dataset and evaluating the performance. Furthermore, we compare several networks with varying inclusions of porcine and human data during training to assess the impact of the species domain. This analysis lays the foundation for future research aimed at achieving segmentation performance on human data that is on par with the segmentation performance on porcine data.



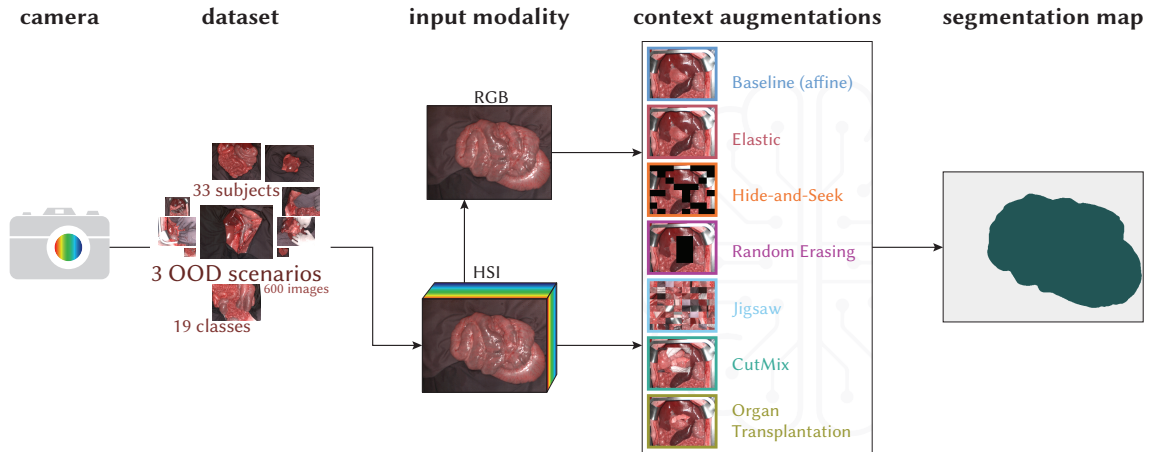


**Figure 1.7:** Overview of our analysis on the effect of different domain shifts (RQ4). We assess our segmentation networks against three different domain shifts: subject, context and species domains. For the subject domain, we compare the performance of images of subjects known during training against images from completely new subjects. For the context domain, we look at the performance on geometrical out-of-distribution (OOD) data, i.e., situations where the neighborhood of organs is different than during training. For the species domain, we apply our segmentation networks on human data and compare different strategies of including human and porcine data during training.

### Addressing the Geometrical OOD Problem

The context domain plays a significant role and has a substantial impact on the segmentation quality (cf. Section 5.4.3). Hence, the focus of this research question is also to explore how we can enhance the performance on geometrical OOD scenes (which are common in real-world surgeries) to match the performance on in-distribution scenes. To achieve this, we propose an augmentation method inspired by surgery which forces the network to learn to detect organs even under unusual neighborhood conditions. We conduct a comparative study where we evaluate several topology-aware augmentation methods and compare them to our proposed augmentation method. An overview of this part is depicted in Figure 1.8.





**Figure 1.8:** Overview of our assessment on segmentation networks for geometrical out-of-distribution (OOD) hyperspectral imaging (HSI) data (part of RQ4). We assess the generalizability under geometrical domain shifts of seven different context augmentations for RGB and HSI networks. The dataset includes three geometrical OOD scenarios (e.g., organs in isolation as shown in the example image).

## 1.3 Outline

This thesis consists of seven chapters. After the current introduction chapter (Chapter 1), Chapter 2 provides the necessary background information on the medical site for autonomous robotic surgery, our special image modality HSI and an introduction of selected machine learning techniques used for the deep neural networks we are using in this thesis for spectra classification and image segmentation.

In Chapter 3, we present the state of the art that is relevant for this thesis, discuss current limitations and how this thesis fills the gaps. We introduce related work of spectral organ fingerprints, highlight current approaches to speed up the training of deep neural networks and discuss work for surgical scene segmentation with RGB and HSI data with special consideration to the employed data augmentation methods in this field. Further, we also explore the general, non-medial HSI field.

Our HSI datasets are introduced in Chapter 4 followed by a description of our classification and segmentation networks. Further, we introduce the technical details of our neural network optimizations and present our proposed solution to maintain segmentation performance on geometrical OOD scenes.

In Chapter 5, we present the results of all our studies in the order of the research questions introduced in Chapter 1. That is, the results for our classification network including our approach of making our HSI dataset publicly available, the benefit of our training optimizations, our segmentation results across modalities and spatial granularities, the effect of different domain shifts including performance improvements on geometrical

OOD scenes when using our proposed method. Each section includes details on how we designed and evaluated our experiments.

We talk about specific aspects of our research questions in Chapter 6 and we discuss the results of this thesis, their limitations and implications in a general context. The latter includes limitations of our hardware, our surgical setting and an extended view on non-healthy tissue types (pathologies).

Chapter 7 closes this work with a summary of our findings while referring to the research questions and gives an outlook on future directions for works that continue the path of automatic surgical scene segmentation set by this thesis.

The research questions are picked up on in several parts of this thesis. Table 1.1 provides an overview of the corresponding sections for each research question.

**Table 1.1:** Outline and corresponding research questions (RQs) of this thesis.

| RQ  | related work | methods     | results     | discussion  |
|-----|--------------|-------------|-------------|-------------|
| RQ1 | Section 3.1  | Section 4.2 | Section 5.1 | Section 6.1 |
| RQ2 | Section 3.2  | Section 4.3 | Section 5.2 | Section 6.2 |
| RQ3 | Section 3.3  | Section 4.4 | Section 5.3 | Section 6.3 |
| RQ4 | Section 3.4  | Section 4.5 | Section 5.4 | Section 6.4 |

### Disclosure of Contributions

The research presented in this thesis is the product of interdisciplinary work with contributions from various team members and collaborators through data acquisition and annotation, discussions and analyses. While this thesis was written independently by myself, it uses the “we” form rather than the “I” form to reflect the collective efforts of everyone involved. For transparency reasons, Appendix A gives an overview of my contributions to the research questions and the corresponding publications.

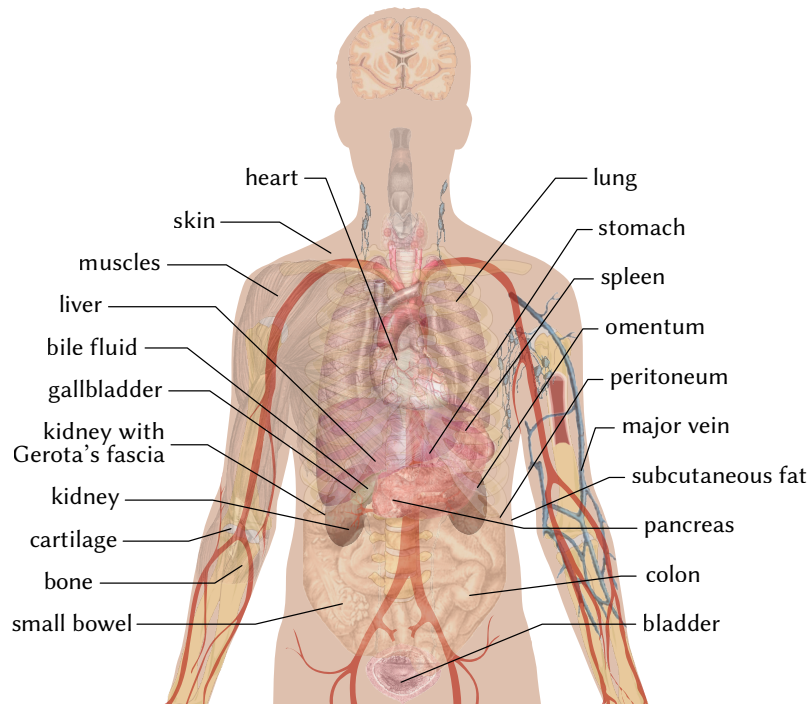
This chapter introduces fundamental topics relevant to this thesis. We start with a brief introduction to visceral surgery and the need for surgical scene understanding in computer- and robot-assisted surgery in Section 2.1. In Section 2.2, we introduce the basic concepts of HSI by describing the physical principles of HSI that make it a valuable modality for medical applications and by providing technical details about the HSI camera system. Finally, we introduce basic deep learning concepts relevant to this thesis in Section 2.3, focusing on convolutional neural networks (CNNs) and mixed precision training.

## 2.1 Medical Background

This section provides relevant medical background information for this thesis. Section 2.1.1 introduces the field of visceral surgery including the associated challenges and complications in this area. In Section 2.1.2, we discuss the importance of surgical scene understanding in the context of computer- and robot-assisted surgery and the potential of HSI in this regard.

### 2.1.1 Visceral Surgery

Visceral surgery, a specialized branch of general surgery, is concerned with the surgical treatment of organs of the body's major cavities, including the thoracic, abdominal and pelvic cavities. Figure 2.1 illustrates the organs that are used in this thesis. The thoracic cavity contains organs of the respiratory and cardiovascular system, composed of the heart and lungs. The abdominal cavity houses organs of the digestive system, such as stomach, liver, gallbladder, spleen, pancreas and intestines. It also contains the kidneys, which are part of the urinary system. The pelvic cavity contains the bladder, which is part of the urinary system, as well as reproductive organs (e.g., uterus) and the distal portions of the large intestine. [212]



**Figure 2.1:** Internal organs of the human body. Highlighted organs correspond to the organs used in this thesis. Image by Mikael Häggström via Wikimedia Commons, Public Domain [84].

Connective tissues play a crucial role in visceral surgery with various functionality. The peritoneum, a membrane that lines the abdominal cavity, is an example of such a connective tissue. The peritoneum contains blood vessels, lymphatics, and nerves that supply the abdominal organs. It supports and protects the organs and provides a lubricating fluid to enable the mobility of visceral organs. The omentum, part of the peritoneum, is a layer of fatty tissue that connects the stomach and duodenum to other abdominal organs. It is involved in fat deposition, immune response, infection, wound isolation and structural support. The omentum is often removed during cancer surgeries to limit the spread of the disease. [130, 53, 20]

Despite significant advances in surgical techniques, complications following visceral surgery remain a substantial concern, affecting nearly half of the patients who undergo major abdominal procedures. These postoperative complications are associated with a significant increase in patient morbidity and mortality [129]. In Germany, the in-hospital mortality rate following visceral surgery is as high as 2 % [18].

Infections are the most common cause of these postoperative complications with anastomotic leakage in gastrointestinal anastomosis and surgical site infections being the most prevalent. Generally, an increase in the complexity of the performed visceral surgery is observed to correlate with a rise in the complication rate and mortality. [22]

The surgeon's experience and expertise play a crucial role in the surgical outcome [216]. Therefore, there is a pressing need for improved intraoperative guidance and decision-making to mitigate the risk of complications and improve patient outcomes in visceral surgery [34]. Ideally, surgeons should make intraoperative decisions based on evidence. For instance, in the case of anastomotic leakage, it is critical to ensure adequate blood supply to the anastomosis. However, this is challenging to assess visually and often relies on the surgeon's experience and intuition. HSI (introduced in Section 2.2) could bridge this information gap and provide a more objective assessment of tissue perfusion. The potential of HSI in intraoperative guidance and decision-making is further discussed in Section 2.1.2. In this thesis, we address the potential of HSI for robust surgical scene segmentation, which could significantly contribute to improving surgical outcomes.

### **2.1.2 Surgical Scene Understanding for Computer- and Robot-Assisted Surgery**

In the dynamic landscape of modern medicine, the convergence of cutting-edge technology and surgical practice has led to remarkable advancements. Since their introduction in the 1980s, minimally invasive surgeries have revolutionized the surgical field, offering a viable alternative to traditional open surgeries. These procedures, performed through small incisions or natural body openings, offer numerous benefits. By minimizing the trauma to the body and lowering the risk of infections, they typically result in less pain, reduced patient recovery time, fewer post-operative complications, less visible scarring and lower surgical costs [152]. As a consequence, minimally invasive surgery is progressively replacing open surgeries in many fields, including visceral surgery [211, 195]. However, minimally invasive surgeries also pose unique challenges. For instance, surgeons must rely on indirect visualization of the surgical site through a camera, which limits depth perception and eliminates haptic feedback. Moreover, they are tasked with performing complex procedures in confined spaces with restricted dexterity and field of view [23]. These challenges have encouraged the development of computer- and robot-assisted surgical systems and emphasize the need for advanced imaging modalities such as HSI (introduced in Section 2.2).

In recent years, robot-assisted surgery, particularly using the da Vinci system<sup>®</sup> (Intuitive Surgical, Inc., Sunnyvale, CA, USA), has emerged as the gold standard for minimally invasive surgery in various disciplines, including prostatectomy, kidney surgery and gynecological procedures [26]. Robotic surgery systems provide surgeons with a more ergonomic workplace and enhanced visualization, dexterity and precision than conventional laparoscopic surgery, ultimately leading to improved patient outcomes [158].

Within the context of laparoscopic and robot-assisted surgery, it is important to comprehend the dynamic environment within an operating room during surgery. Surgical scene understanding addresses this need by providing a comprehensive understanding of

the surgical scene, including the patient, the surgical instruments, and the surrounding environment. This understanding is essential for the development of computer-assisted robotic systems that can seamlessly integrate with the surgical scene while providing real-time guidance to surgeons leading to enhanced precision and reduced errors. [54]

Surgical scene segmentation is an important part of surgical scene understanding and an active area of research. It involves the partitioning of the surgical scene into meaningful regions, such as organs, pathologies and surgical instruments. This segmentation is essential for the development of intelligent surgical robots since a robot that can accurately identify organs and pathologies in real-time can provide valuable guidance to the surgeon, such as augmented reality overlays of critical structures or tracking of tumor resection margins [119]. However, up to date, robot-assisted and laparoscopic surgeries rely on conventional imaging modalities, such as RGB cameras, and in consequence, the majority of research on surgical scene segmentation is focused on RGB video data. HSI holds the potential to overcome these limitations by non-invasively and continuously offering detailed spectral information about the tissue. This supplementary information could be used to more accurately identify organs and pathologies in real-time. [39] A detailed overview of the state of the art in surgical scene segmentation on both RGB and HSI data is provided in Section 3.3.

## 2.2 Physical Background

HSI is an imaging technique that collects fine-grained information about light in a dense spectrum. This is in contrast to RGB cameras that only capture three broad channels: red, green, and blue. Furthermore, spectral cameras are not limited to the visible spectrum since they can also capture light in regions of the spectrum that are invisible to the human eye, such as the near-infrared region. HSI has found applications in numerous fields, including medical imaging [177]. This section will introduce the underlying principles of the light-tissue interaction of HSI pertinent to the medical field (Section 2.2.1), as well as the mechanics of the camera system utilized in this thesis (Section 2.2.2).

### 2.2.1 Light-Tissue Interaction

The underlying physical principle of HSI which makes it so valuable in the medical field is the interaction of light with tissue as illustrated in Figure 2.2. Generally, light is emitted from a source and penetrates the tissue with the penetration depth depending on the wavelength. As it travels, the light primarily interacts with the tissue through absorption and scattering. Absorption converts light into heat and reduces its intensity while scattering alters its direction without modifying its intensity. Both of these effects influence the path of the light. Ultimately, some light re-emerges at the surface and

can be measured by the camera. The path taken by the light depends on the molecular composition of the tissue so that the resulting measured spectrum conveys valuable information.

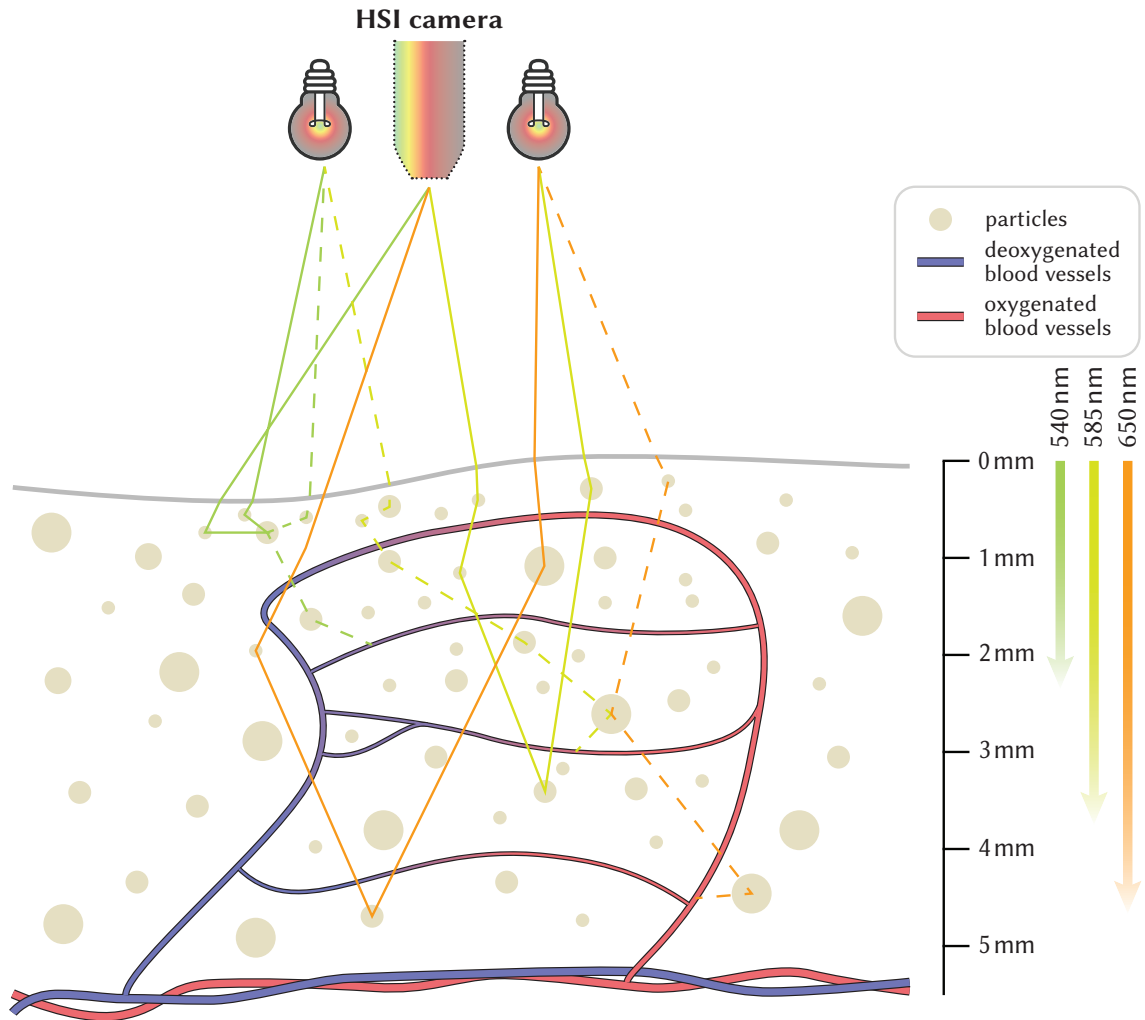
For medical applications, the interaction of light with chromophores is of special interest as it allows the extraction of functional parameters (e.g., oxygenation or perfusion) from the tissue. This interaction is illustrated in Figure 2.3 using the protein hemoglobin as an example. Hemoglobin, a primary absorber in organs, is responsible for the oxygen transport throughout the body. It exists in two main forms: oxygenated ( $\text{HbO}_2$ , oxy-hemoglobin) and deoxygenated (Hb, deoxyhemoglobin), each with distinct absorbance patterns. Depending on the wavelength, either oxyhemoglobin or deoxyhemoglobin absorbs more light, enabling the estimation of tissue oxygenation levels through the analysis of reflectance measurements. [232, 39]

### 2.2.2 Hyperspectral Imaging Hardware

Various approaches exist for capturing a full image with spectral information for each pixel. Examples are point scanning, line scanning or snapshot devices [40, 241]. These methods each have distinct characteristics, particularly in terms of spectral resolution and the time required to capture an image [233]. For this thesis, the data was collected using a line scanning device which will be explained in more detail in the following.

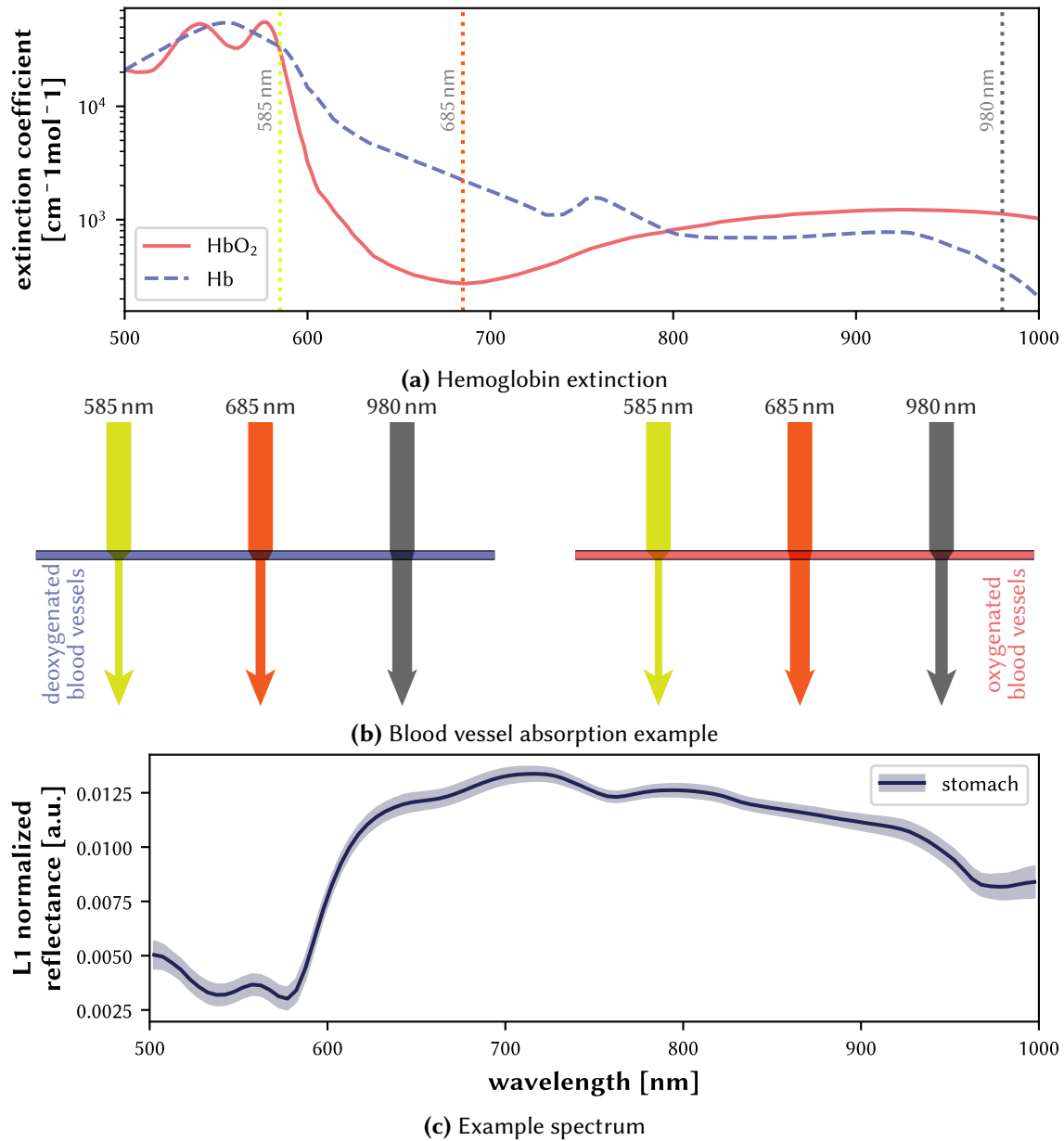
The concept of a line scanning device is depicted in Figure 2.4. It captures an image line by line obtaining full spectral information for each line. Once a line is completed, the camera's motor shifts one position further along the width axis to record the next line. For the Tivita<sup>®</sup> Tissue camera system, this process takes approximately seven seconds for a single image. While the line scanning approach offers high spectral resolution, it suffers from long acquisition times [233]. Any movement during this acquisition period can result in motion artifacts in the image which can also be seen in the heart example images of Figure 4.1 where the heartbeat causes such artifacts.

A spectral camera system is defined by several characteristics, including the sensitive wavelength range, the number of bands and the width of each band. These details are represented by the filter response functions which specify the spectral sensitivity (the wavelengths to which each channel is sensitive) and consequently the amount of light captured for each wavelength [133, 182]. An example of an RGB and HSI camera system, featuring 100 channels ranging from 500 nm–1000 nm and a channel width of 5 nm, as well as the spectral sensitivity of human cone cells is visualized in Figure 2.5. RGB cameras mimic human vision while both are sensitive to three broad, overlapping wavelength regions (red, green and blue) whereas the filters of the HSI camera are considerably narrower and have less overlap. This level of granularity facilitates the capture of subtle variations caused by chromophores thereby enabling the extraction of the underlying tissue's molecular properties.

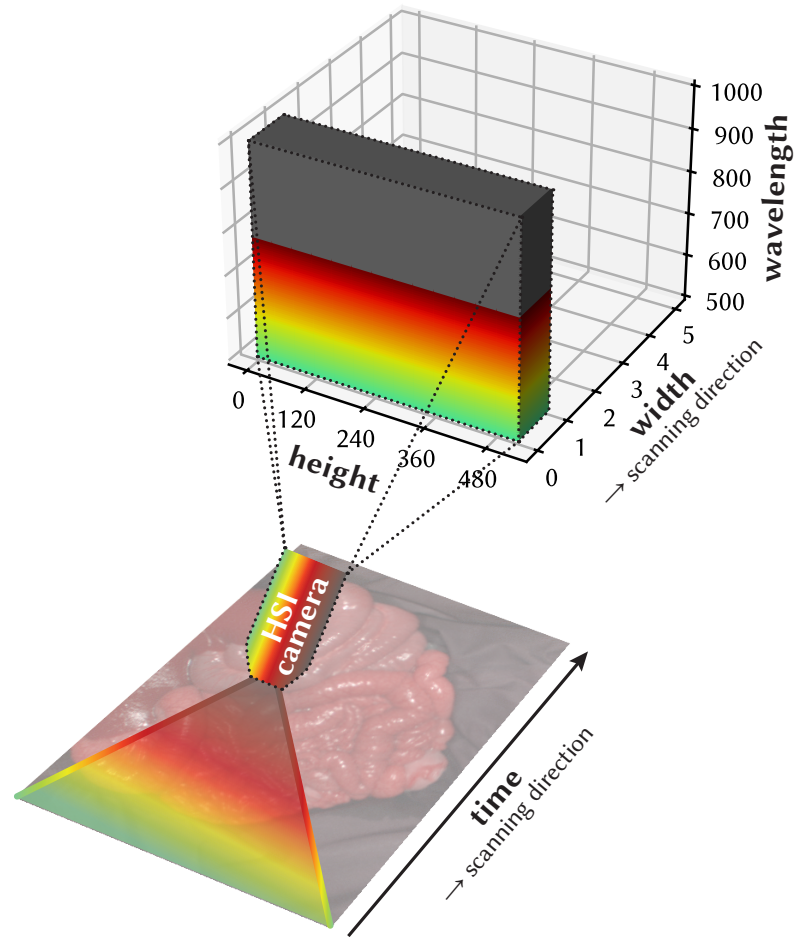


**Figure 2.2:** Simplified concept of the light-tissue interaction. Light is emitted from the light sources, interacts with the tissue and is collected by a hyperspectral imaging (HSI) camera. In this example, three photons are emitted per light source, in each case one for 540 nm, 585 nm and 650 nm. Some photons reappear at the surface upon multiple scattering events within the tissue and can be measured by the camera (paths visualized as solid lines) while other photons get absorbed (dashed lines). Absorption can occur when interacting with vessels while scattering can happen for interactions with small particles like cells [239]. The probability for scattering and absorption events and the resulting penetration depth is wavelength-dependent [232, 13, 77].

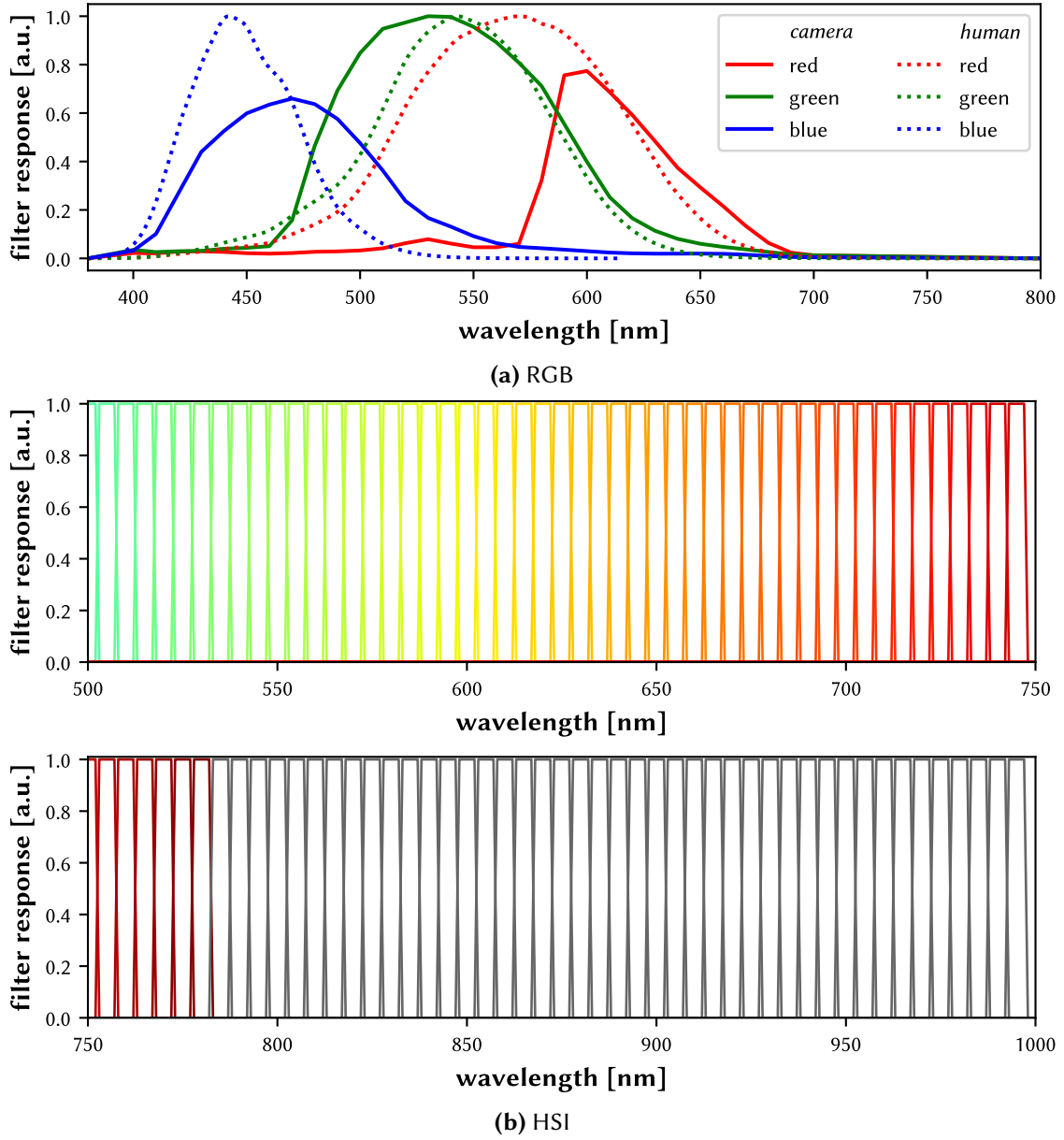




**Figure 2.3:** Effect of hemoglobin on the spectra. Extinction coefficients for deoxyhemoglobin (Hb) and oxyhemoglobin ( $\text{HbO}_2$ ) (a) affect through the absorption properties of blood vessels (b) the example spectrum of stomach (c). The absorption of hemoglobin (main factor of the extinction coefficient [232]) depends on the wavelength and is either identical for oxygenated and deoxygenated (585 nm), higher for oxygenated than deoxygenated (685 nm) or lower for oxygenated than deoxygenated (980 nm) blood vessels. The median spectrum (solid line) and the standard deviation across subjects (shaded area) for the stomach are similar to Figure 5.2. The data for (a) was provided through [176].



**Figure 2.4:** Line scanning approach for acquiring hyperspectral imaging data as utilized by the Tivita<sup>®</sup> Tissue (Diaspective Vision GmbH, Am Salzhaff, Germany) camera [122]. This is the same camera that was also used to acquire the datasets presented in Section 4.1. The hyperspectral image is taken line-by-line along the width axis of the image over acquisition time. For each line, all pixels along the height axis are recorded and the light is captured for each pixel in 100 channels from 500 nm–1000 nm with an approximate spectral resolution of 5 nm. After a line is finished, the motor in the camera moves one position further along the width axis to record the next line until all 640 lines are finished. This process takes approximately seven seconds for one image.



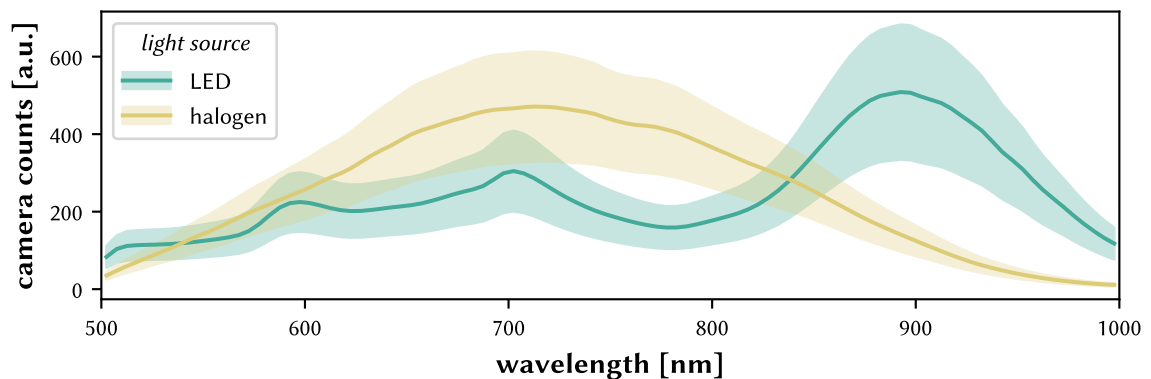
**Figure 2.5:** Exemplary filter response functions for an RGB (a) and hyperspectral imaging (HSI) (b) camera system as well as human cone cells. Filter response functions characterize for each wavelength the spectral sensitivity and hence the amount of captured light. An RGB camera system has only three filter response functions each with a broad range and overlapping regions (here from an Apple® iPhone® 12 Pro Max camera with data from [226]) whereas an HSI camera system may have 100 bands of narrow width and minimal overlap allowing it to capture a much more detailed light spectrum. In the figure, the filter response functions for an HSI camera system with 100 spectral channels with a width of 5 nm within the range 500 nm–1000 nm is shown. The data for the spectral sensitivities of the human cone cells was provided through [213].

Several factors inherent in the hardware used can unintentionally influence the recorded hyperspectral image making the comparison of different images or spectra challenging. Examples of these factors include sensor noise, variations in sensitivity across different sensor chips or the employed light source.

Sensor noise can be caused by dark current which refers to measured counts even in the absence of light. This could occur due to thermal energy and becomes particularly problematic if the sensor heats up over time, e.g., because multiple consecutive images are taken.

Pixels located at different spatial locations may vary in their sensitivity to light. This could potentially be caused by chip errors (stemming from manufacturing errors, impurities, etc.) or dust and results in some pixels reporting higher intensity than others even if the incoming light is the same.

The light source is one of the most significant factors since different light sources can have entirely different spectral characteristics, i.e., the amount of light emitted per wavelength varies. This is demonstrated in the example spectra for a halogen and light-emitting diode (LED) light source in Figure 2.6. In this example, near 900 nm, the LED light source emits significantly more light than the halogen source, leading to more measured counts and a brighter image, even if the underlying tissue is identical. Furthermore, the light source may also be spatially inhomogeneous which causes some parts of the image to be brighter than others. This can be seen in the example white image shown in Figure 2.7. [186]



**Figure 2.6:** Exemplary spectra for an light-emitting diode (LED) and halogen light source. The spectra were acquired by taking an image of a white surface with the Tivita<sup>®</sup> Tissue (uses a halogen light source) and Tivita<sup>®</sup> 2.0 Surgery Edition (uses an LED light source) camera (Diaspective Vision GmbH, Am Salzhaff, Germany). The corresponding white images  $W(x, y, \lambda) - D(x, y, \lambda)$  have the dark image already subtracted. For each light source, the median spectrum (solid line) and the standard deviation across all pixels in the image (shaded area) are shown.

To mitigate some of these issues, a calibration of the camera system is typically performed using a dark and a white image [68]. The dark image  $D$  is captured while all light sources are turned off so that no light reaches the sensor and we only get a measure of the intrinsic responses of the sensor. Conversely, the white image  $W$  is taken with a diffuse reflecting white surface positioned in front of the camera so that most of the light is reflected back to the sensor [62]. This white image  $W$  essentially serves as a fingerprint of the light source and represents its characteristics. Subsequently, the raw hyperspectral image  $I(x, y, \lambda)$ , defined by the horizontal position  $x$ , vertical position  $y$ , and wavelength  $\lambda$ , is calibrated via:

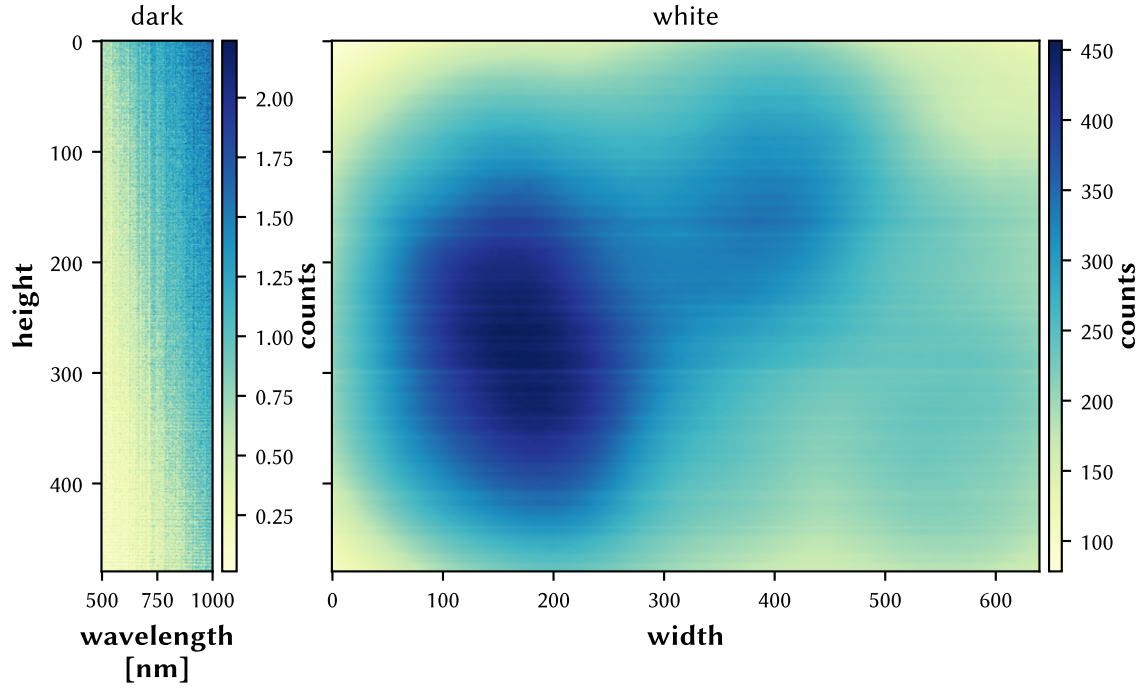
$$R(x, y, \lambda) = \frac{I(x, y, \lambda) - D(x, y, \lambda)}{W(x, y, \lambda) - D(x, y, \lambda)} \quad (2.1)$$

This calibration step has multiple effects:

1. Since the dark image is subtracted from every image taken by the camera, we reduce the effect of sensor noise.
2. Different sensitivities of pixels are less problematic because they are also measured by the white image and normalized by the division.
3. Similarly, we are now less dependent on the lightning conditions since  $R(x, y, \lambda)$  is an image normalized by the current light source. Likewise, the problem of the spatial inhomogeneity of the light source is also attenuated.
4. Finally, with this step, we transition from camera counts of the image  $I(x, y, \lambda)$  to reflectances  $R(x, y, \lambda)$  since reflectance is defined by the ratio of returned light compared to all emitted light.

Given that light sources and other factors may vary over time, it is standard practice to perform this calibration regularly, for instance, before every surgery [35]. An example of a white and dark image is shown in Figure 2.7. The dark image displays only a few counts, suggesting minimal sensor noise, although the higher channels in the near-infrared region are more noise-affected. The white image reveals that the light source is not spatially homogeneous with much higher intensities observed in the left part of the image.

Not every unintended effect can be mitigated by this calibration step. For instance, variations in the distance between the light source and the tissue can still lead to different image intensities with images appearing brighter if the camera is closer to the surface. The pose of the camera and the light source also play a significant role, especially if the light is not coming directly from above but from a different measurement angle [41]. Furthermore, the white object used to capture the white image may not be consistent across measurements (e.g., if there is no agreement for standardized white objects) which can also impact the reflectances. To account for these effects, additional normalization or calibration steps, such as the L1 normalization of Equation 4.1 described in Section 4.1, are necessary. This is also a topic of discussion in the literature. [89, 82, 145, 113]



**Figure 2.7:** Exemplary white and dark measurements from the Tivita<sup>®</sup> Tissue camera (Diaspective Vision GmbH, Am Salzhaff, Germany). The data cube for the dark measurement has a shape of  $480 \times 100$  (height, channels) since the camera already averages along the width direction, i.e.,  $D(x, y, \lambda) = D(y, \lambda)$  in this case. The data cube for the white measurement has a shape of  $480 \times 640 \times 100$  (height, width and number of channels) similar to the normal images and has the dark image already subtracted, i.e.,  $W(x, y, \lambda) - D(x, y, \lambda)$  is shown. A halogen light source was used for the white image. Here in the figure, an average of the spectral dimension is shown (see Figure 2.6 for the corresponding median spectra).

## 2.3 Deep Learning

This thesis makes heavy use of deep-learning methods which have gained significant popularity in recent years across various fields, including image recognition, natural language processing or speech recognition [128]. It is an active research field abundant with diverse techniques, tricks and concepts [80]. Generally, deep learning is a paradigm where we train networks with numerous examples to enable them to identify patterns and relationships within the data. The learned knowledge about the data can then be used to make predictions (discriminative models) or generate new examples (generative models). The focus of this thesis is on discriminative models.

There are numerous resources available that provide the necessary components for training deep neural networks<sup>1</sup>. Here, we focus on two important aspects especially relevant in the context of this thesis: CNNs and mixed precision training which are presented in more detail in Section 2.3.1 and Section 2.3.2, respectively. The latter also includes an introduction to how floating-point numbers are represented in computers. Nearly every architecture of this thesis makes use of CNNs and every network trained for this thesis employs mixed precision training. Further, float16 precision is particularly relevant for RQ2.

### 2.3.1 Convolutional Neural Networks

CNNs are at the heart of many deep learning architectures. They are particularly well-suited for image data and have been used to achieve state-of-the-art performance in various image-related tasks such as image classification, object detection, or semantic segmentation [128, 80]. This section provides an introduction to the basic concepts of CNNs<sup>2</sup>. We start with the mathematical convolution operation, move on to learnable weights and then present an example architecture suitable for semantic segmentation of images.

#### Convolution Operation

In the computer vision and image processing community, convolutional operations have been used for a long time [27]. These operations involve the manual definition of filters (also known as kernels) which can be used for things like blurring an image or detecting edges. A filter is a matrix that is typically very small (e.g.,  $3 \times 3$  or  $5 \times 5$ ) and is applied to every position in the image to calculate a weighted sum that constitutes the filter response for that specific position. Mathematically, we convolve the image  $I = (I_{11}, I_{12}, \dots)$  with a kernel  $K = (K_{11}, K_{12}, \dots)$  to obtain the response  $R = I * K$  at every location  $(x, y)$  of the image<sup>3</sup>:

$$R_{x,y} = (I * K)_{x,y} = \sum_{i=-W}^W \sum_{j=-H}^H I_{x+i,y+j} \cdot K_{W+i+1,H+j+1} \quad (2.2)$$

where  $*$  denotes the convolution operator. The summation is constrained by  $W \geq 1$  and  $H \geq 1$  which denote the half width and half height of the filter (e.g., for a  $3 \times 3$  filter we have  $W = H = \lfloor \frac{3}{2} \rfloor = 1$ )<sup>4</sup>. Effectively, the kernel is centered at every non-border location of the image and the weighted sum is calculated for the patch of the image covered by

<sup>1</sup>For a general introduction, the book *Neural Networks and Deep Learning* by Nielsen is highly recommended [164].

<sup>2</sup>This section is based on [199].

<sup>3</sup>Strictly speaking, this operation is rather a correlation than a convolution since the kernel is not flipped. However, this distinction does not matter in the context of CNNs since, as we will learn later, the network uses learnable weights and is free to change the sign of the weights.

<sup>4</sup>For simplicity, we only consider the case of odd kernel side lengths here.

the kernel. The responses for the border values cannot be computed directly because the kernel would be out of bounds on the image. In practice, this can be solved by expanding the image borders (e.g., by reflecting the values) [27].

As example, consider the following simple image  $\hat{I}$  and generic  $3 \times 3$  kernel  $K$

$$\hat{I} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \quad \text{and} \quad K = \begin{pmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{pmatrix}. \quad (2.3)$$

The response  $\hat{R}_{2,2}$  for the highlighted center position in  $\hat{I}$  is then calculated as

$$(\hat{I} * K)_{2,2} = 1K_{11} + 2K_{12} + 3K_{13} + 4K_{21} + 5K_{22} + 6K_{23} + 7K_{31} + 8K_{32} + 9K_{33}.$$

Given that this operation is independent for each position, it is highly parallelizable and is especially suited for GPUs. Modern GPUs even have specialized hardware for convolutions and optimized algorithms for specific use cases [46].

As an example for a real-world kernel used in the wild, consider the following matrices

$$G_x = \begin{pmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{pmatrix} \quad \text{and} \quad G_y = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix} \quad (2.4)$$

which can be used to detect edges in an image. These filters are known as the Sobel filters and compute finite differences of the image intensity values [209]. They can be used to approximate the horizontal ( $G_x$ ) and vertical ( $G_y$ ) image gradients. For illustration purposes, consider another simple image

$$\tilde{I} = \begin{pmatrix} 0 & 0 & 1 & 3 \\ 0 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 1 & 0 & 3 \end{pmatrix} \quad (2.5)$$

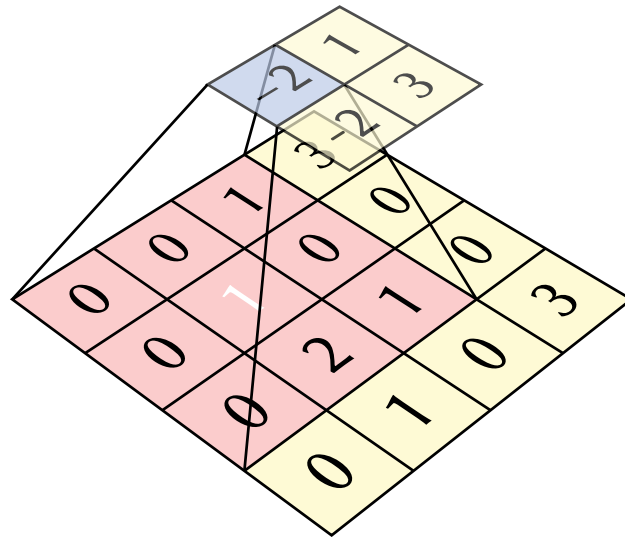
where we can calculate the filter response  $\tilde{R}_{2,2} = (\tilde{I} * G_x)_{2,2}$  as

$$\begin{aligned} & (1) \cdot 0 + (0) \cdot 0 + (-1) \cdot 1 + \\ & (2) \cdot 0 + (0) \cdot 1 + (-2) \cdot 0 + \\ & (1) \cdot 0 + (0) \cdot 2 + (-1) \cdot 1 = -2 = (\tilde{I} * G_x)_{2,2}. \end{aligned}$$

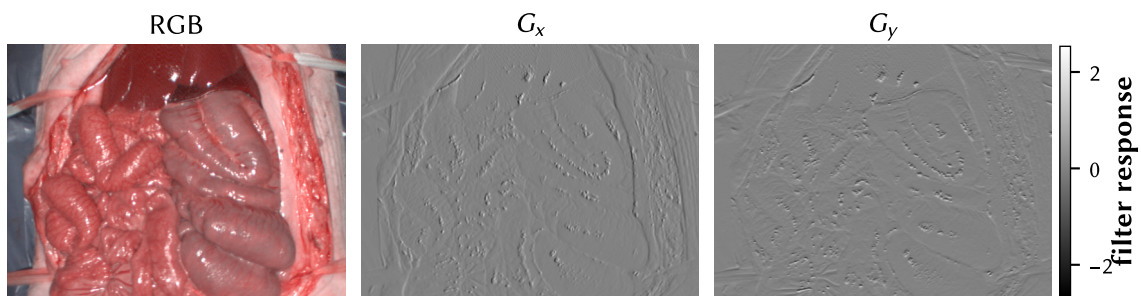
Figure 2.8 visualizes  $\tilde{R} = \tilde{I} * G_x$  for every location in  $\tilde{I}$  where the kernel can be centered.

An example of the kernels  $G_x$  and  $G_y$  (Equation 2.4) applied to an RGB image is shown in Figure 2.9. High filter responses occur, for example, between organs, at specular reflections or at visible vessels.





**Figure 2.8:** Simple example of the convolution operation. The example matrix  $\tilde{I}$  of Equation 2.5 is convolved with the filter  $G_x$  of Equation 2.4. The resulting matrix  $\tilde{R} = \tilde{I} * G_x$  is shown on top and the calculation of the top left element is highlighted (kernel centered at the white 1). This figure was adapted from [199].



**Figure 2.9:** Example of a handcrafted convolution filter. The Sobel filter of Equation 2.4 can be used to detect edges in an image (image gradients). Here, it is applied to the example RGB image via Equation 2.2. The RGB image is converted to grayscale before applying the filters and the image borders are reflected to maintain the image size. The filter responses of  $G_x$  and  $G_y$  are approximations for the horizontal and vertical gradients.

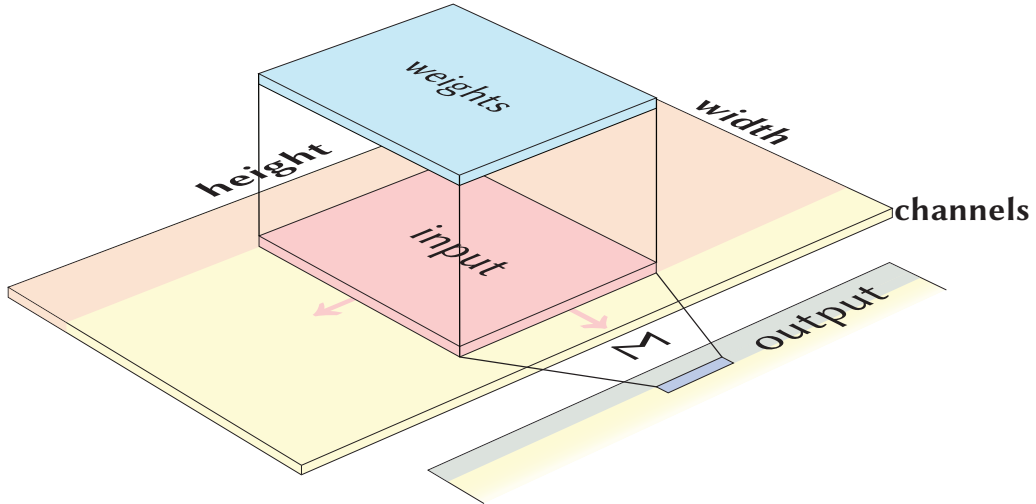
### Learning the Kernel

CNNs essentially adopt the concept of convolution from Equation 2.2 but replace manually defined weights of the kernel with learnable ones. This enables the network to learn and determine the most suitable filters for the specific task it is designed to perform. The

filter response  $R_{x,y}$  for a specific position  $(x, y)$  in the image  $I_{x,y,c}$  is calculated by

$$R_{x,y} = f\left(\sum_{i=-W}^W \sum_{j=-H}^H \sum_{c=1}^C I_{x+i,y+j,c} \cdot \Omega_{W+i+1,H+j+1,c} + b\right). \quad (2.6)$$

This filter multiplies the learnable weight matrix  $\Omega \in \mathbb{R}^{(2 \cdot W + 1) \times (2 \cdot H + 1) \times (C)}$  with the values of the input image  $I$ . In addition to Equation 2.2, we also consider the channel dimension  $c$  of the input with corresponding weights (with  $C$  denoting the total number of channels). However, Equation 2.6 does not slide over the channel dimension but always consider all channels at once<sup>56</sup>. Further, we apply an activation function  $f(x)$  to introduce non-linearity (detailed below) and add a learnable bias  $b \in \mathbb{R}$  to the weighted sum. The bias allows the network to learn an offset of the features which effectively shifts the activation function to the left or right. The principle of the convolution operation is visualized in Figure 2.10.



**Figure 2.10:** Example of the convolution operation for a convolutional neural network (CNN).

The convolution operation as defined by Equation 2.6 is applied to the weights and the input image by sliding the weights over the image and computing a weighted sum between the weights and the respective input (patch of the image) for each location. Every time this operation is performed, one output value is created. Sliding only happens along the height and width dimension of the image and in each position, the values from all channels are used.

In a convolutional layer, multiple filters are used with the weights being initialized randomly so that every filter can learn different features allowing the network to identify

<sup>5</sup>In this regard, the channel dimension is more similar to a fully connected layer where every input is connected to its own weight.

<sup>6</sup>It is actually possible to also slide over the channel dimension. In this case, we would have a 3D instead of a 2D convolution.

various patterns in the data. In a deep neural network, we stack multiple of these convolutional layers together, enabling the network to learn more and more abstract features. Each filter always has access to all the channels of the previous layer. For the first convolutional layer, this implies access to all channels of the input image. For subsequent layers, this means access to all filter responses from the previous layer.

### Activation Functions

$f(x)$  of Equation 2.6 represents the activation function which transforms the learned features in a non-linear manner. This function enables the network to learn non-linearly separable target functions, thereby empowering the network to tackle complex tasks. There exists a variety of different activation functions with a few examples being shown in Figure 2.11 and listed below:

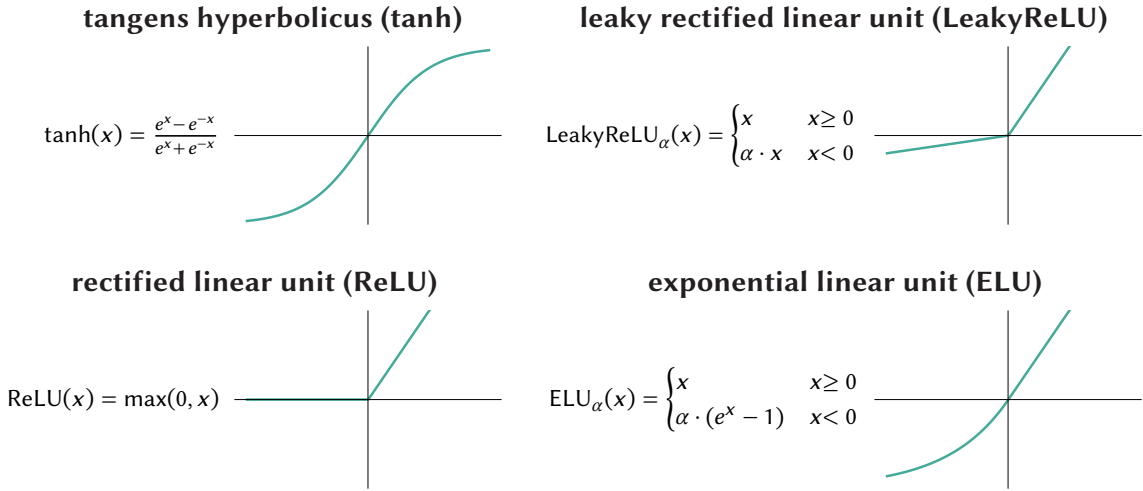
1. *Tangens hyperbolicus*:  $\tanh(x)$  is a popular choice if the output needs to be constrained to  $[-1; 1]$  but is a suboptimal choice for all layers as it can lead to vanishing gradients during training [164].
2. *(Leaky) rectified linear unit*:  $\text{ReLU}(x)$  [2] and  $\text{LeakyReLU}_\alpha(x)$  [137] are popular choices because they are computationally efficient and are less prone to vanishing gradients. The main advantage of  $\text{LeakyReLU}_\alpha(x)$  over  $\text{ReLU}(x)$  is that the former does not suffer from dying neurons (neurons which only output zero) [164].
3. *Exponential linear unit*:  $\text{ELU}_\alpha$  is a smooth version of  $\text{LeakyReLU}_\alpha(x)$  and has been shown to perform better in certain cases [42].

### Pooling Layers

CNNs often also incorporate pooling layers that do not contain learnable weights but instead perform a predefined operation on the input. The primary objective of these pooling layers is to reduce the spatial dimensions of the input. This reduction improves computational efficiency, decreases the number of weights in the network and increases the network's robustness to minor translations in the input [164]. A common operation in this context is max-pooling which selects the maximum value from a patch of the input, thereby reducing the spatial dimensions. For instance, when we apply max-pooling via a  $2 \times 2$  patch on the example image  $\hat{I}$  from Equation 2.3, we obtain the pooled image

$$\text{maxpool}_{2 \times 2}(\hat{I}) = \begin{pmatrix} \max(1, 2, 4, 5) & \max(2, 3, 5, 6) \\ \max(4, 5, 7, 8) & \max(5, 6, 8, 9) \end{pmatrix} = \begin{pmatrix} 5 & 6 \\ 8 & 9 \end{pmatrix}.$$

In this regard, we keep only the most significant feature within a specific region of the input and the output value denotes whether this feature was present or not. As the precise location of the feature is not of importance, we enhance the translational invariance of the network. [164]

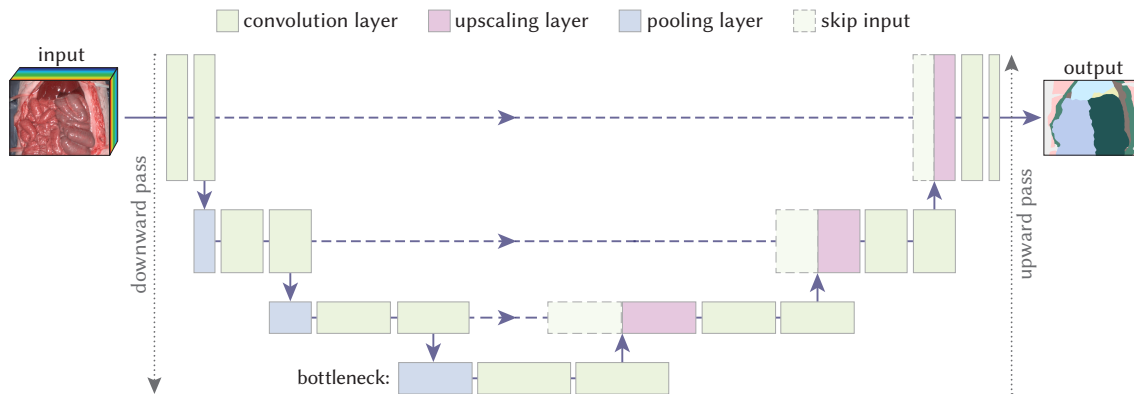


**Figure 2.11:** Example of activation functions. Four different activation functions, namely, tangens hyperbolicus (tanh), leaky rectified linear unit (LeakyReLU) [137], rectified linear unit (ReLU) [2] and exponential linear unit (ELU) [42] are shown as a function of the input  $x$ .  $\alpha$  is a parameter to control the scaling for input  $x < 0$ . In this example,  $\alpha = 1$  for ELU and  $\alpha = 0.1$  for LeakyReLU.

### Example Segmentation Architecture

CNNs serve as the foundational building blocks of several neural network architectures. When combined in a specific manner, they can form powerful architectures tailored for specific tasks. One such architecture is the U-Net which is particularly well-suited for semantic segmentation tasks [189]. As depicted in Figure 2.12, the U-Net consists of a U-shaped structure that condenses the input features toward a bottleneck and subsequently expands it again to produce the target output (e.g., a segmentation mask). Given the limited size of the bottleneck, the network has to learn abstract high-level features to generate the output again. The U-Net is similar to an autoencoder [131] but uses skip connections to ensure the network has access to the corresponding features from the downward pass of the same hierarchical layer. This enables the network to reconstruct the output with greater precision. The U-Net serves as the fundamental architecture for the networks utilized in this thesis.

The U-Net employs upscaling layers to scale the bottleneck features (e.g.,  $15 \times 20$ ) back to the original input size (e.g.,  $480 \times 640$ ). Within these layers, we also concatenate the features from the downward pass at the same hierarchical level (skip connection). Figure 2.13 shows how the upscaling layer works. Essentially, we interpolate (upsample) the features from the previous layer to match the spatial shape of the features coming from the skip connection. Subsequently, we concatenate these features and apply a convolutional layer to learn how to effectively combine them [166].



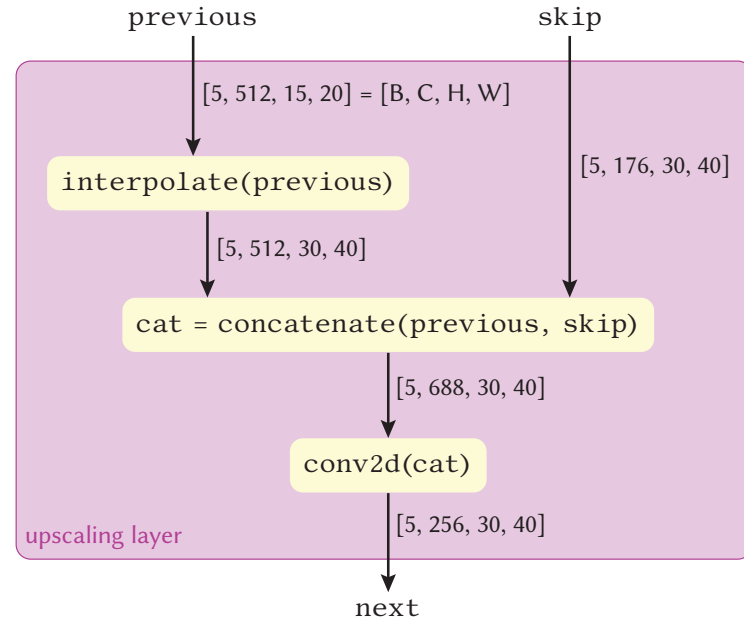
**Figure 2.12:** Overview of the U-Net architecture [189]. During the downward pass, the input is downscaled by a series of convolutional and pooling layers toward the bottleneck of the architecture. Then, the condensed feature representation of the bottleneck is expanded again during the upward pass by a series of convolutional and upscaling layers. At each level of the upward pass, the corresponding output from the downward pass from the same hierarchical layer is combined with the upward features (skip connections). The number of channels of the input usually depends on the modality (e.g., 3 for RGB or 100 for hyperspectral imaging) and the number of channels in the output depends usually on the number of classes.

### 2.3.2 Mixed Precision Training

Efficient training of neural networks is crucial for several reasons: it enables faster developer cycles, optimizes hardware usage (smaller training budget and a reduced carbon footprint) and accelerates the network’s response time which is particularly important during inference. Per default, operations on the GPU are performed in float32 precision but float16 precision offers advantages in terms of speed and memory usage. However, some operations may necessitate higher precision to avoid numerical instabilities so that training entirely with float16 could result in networks with significantly lower accuracy.

The basic idea of mixed precision training is to perform as many operations as possible in float16 precision while resorting to float32 precision only for operations that require it [150]. Mixed precision training is particularly beneficial for GPUs equipped with *Tensor Cores* which are special units on the chip optimized to perform matrix multiplications with built-in mixed-precision support [49].

This section introduces the basic principles of mixed precision training. We begin with the fundamentals of how floating-point numbers are represented in computers, followed by the concept of autocasting and the need for loss scaling. Finally, we conclude with a comparison of networks trained under different precision settings.



**Figure 2.13:** Functional principle of the upscaling layer. Input to the upscaling layer is the output from the previous level and the skip connection, i.e., the features from the downward pass on the same hierarchical level. Then, the features from the previous level are interpolated (upsampled), concatenated with the skip features and then combined in a convolutional layer before they are passed on to the next level. The shape information  $[B, C, H, W]$  denotes the batch, channel, height and width dimensions, respectively. The example numbers are from the first decoding step after the bottleneck. The original image input shape to the U-Net was  $[5, 100, 480, 640]$  (100 spectral channels as input) and the target shape of the output is  $[5, 19, 480, 640]$  (segmentation of 19 different classes).

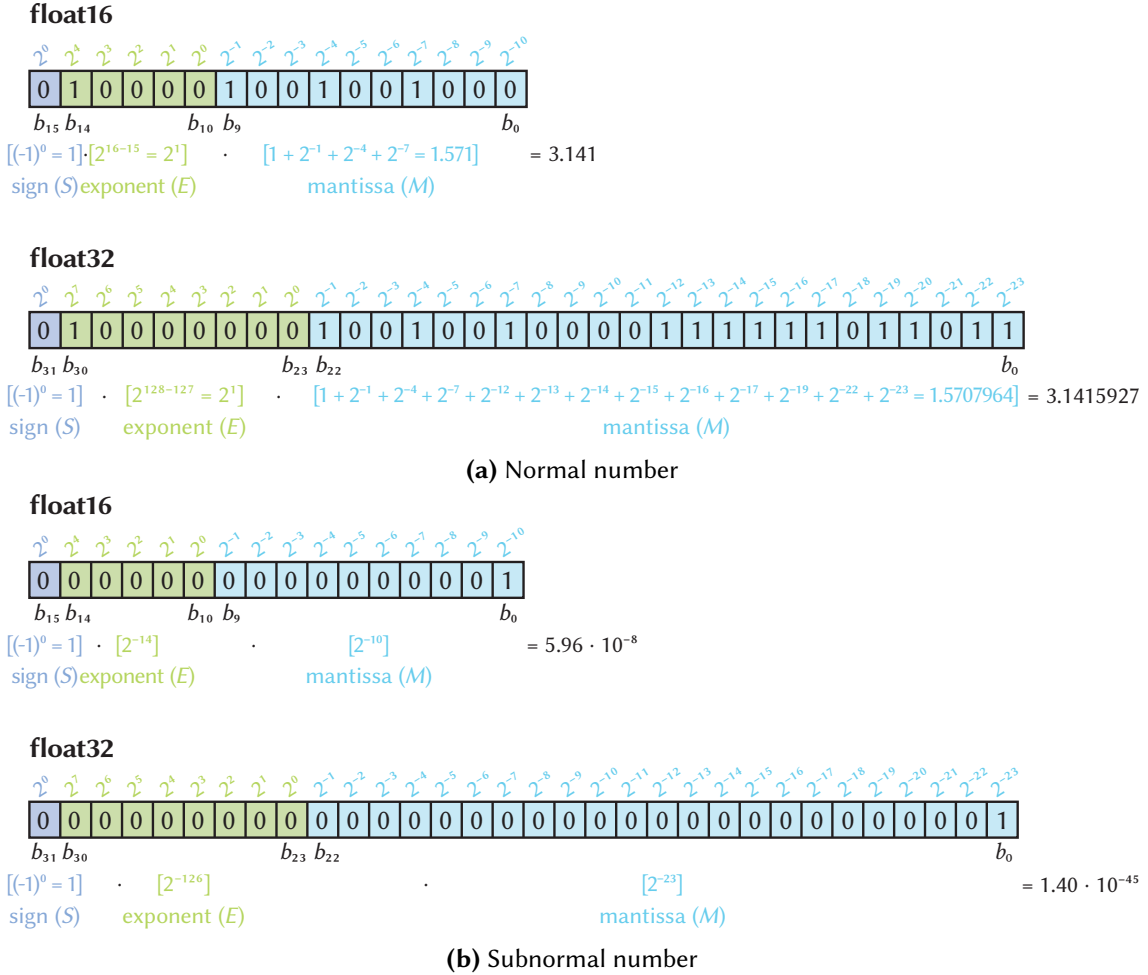
### Floating-Point Number Representation

The representation of floating-point numbers in the binary system is defined by the IEEE 754 standard [98]. The general idea is to represent these numbers in their normalized form with only one integer in front of the comma, e.g.,  $1.625 \cdot 10^1$  instead of 16.25. Essentially, every number can be represented in this form; we simply need to shift the comma either to the left or the right so that only one integer remains in front of the comma. In the binary system (the notation  $(x)_2$  is used to refer to a number in the binary system instead of the common decimal system), this would translate to  $(1.000001)_2 \cdot 2^4$  instead of  $(10000.01)_2$  where  $(10000)_2$  represents  $1 \cdot 2^4 = 16$  and  $(01)_2$  represents  $0 \cdot 2^{-1} + 1 \cdot 2^{-2} = 0.25$ .

In total, a decimal number  $x$  is represented as

$$x = (-1)^{(S)_2} \cdot 2^{(E)_2 - b} \cdot (1.M)_2 \quad (2.7)$$

with the sign  $S$ , the mantissa  $M$  and the exponent  $E$ . Figure 2.14 (a) shows an example for representations of approximations of the number  $\pi$  with float16 and float32 precision.



**Figure 2.14:** Example of binary floating-point computer number formats for float16 and float32 precisions. Normal **(a)** and subnormal **(b)** numbers are represented by the corresponding bits  $b_i$  for the sign  $S$ , exponent  $E$  and mantissa  $M$  (cf. Equation 2.11). The sign defines whether the resulting number will be positive or negative. The exponent spans a window between two exponential numbers (e.g.,  $[2^3; 2^4]$ ), is biased to represent positive and negative ranges and can be thought of as shifting the comma of the binary number either to the left or to the right. The mantissa defines the offset inside the window and encodes the accuracy of the number. Subnormal numbers (Equation 2.10) follow a special definition to represent small values close to 0.0 which are outside the range of normal numbers [192]. Floating-point numbers are defined according to the IEEE 754 standard [98].

Using float16 as example, the bits  $b_{14}, b_{13}, \dots, b_{10}$  of the exponent  $E$  represent a positive integer binary number

$$(E)_2 = \sum_{i=0}^{e-1} b_{10+i} \cdot 2^i \quad (2.8)$$

with  $e$  denoting the number of bits reserved for the exponent ( $e = 5$  in case of float16). In order to allow shifts of the comma to the left, a bias  $b = 2^{e-1} - 1$  is subtracted from the exponent in Equation 2.7.

Using again float16 as example, the bits  $b_9, b_8, \dots, b_0$  of the mantissa  $M$  store the accuracy of the number

$$(M)_2 = \sum_{i=1}^m b_{m-i} \cdot 2^{-i} \quad (2.9)$$

with  $m$  denoting the number of bits reserved for the mantissa ( $m = 10$  in case of float16). Basically, every decimal number  $x$  is an approximation determined by this sum. The accuracy increases as more bits are being used so that (potentially) more summands can be added. Please note that the integer number in front of the comma is not stored in the mantissa but is implicitly assumed to be 1 because this bit would be set for every normal number  $x \geq 1$ . Numbers  $x < 1$  can still be represented by shifting the comma to the left.

Equation 2.7 has the disadvantage that 0.0 cannot be represented exactly due to the always present 1 in front of the comma. To address this issue, the IEEE 754 standard introduced the concept of subnormal numbers:

$$x = (-1)^{(S)_2} \cdot 2^{1-b} \cdot (0.M)_2. \quad (2.10)$$

These numbers use a 0 instead of a 1 in front of the comma. Subnormal numbers are specifically defined to have only zero bits in the exponent ( $(E)_2 = 0$ ). This allows us to represent (0.0) as a number with also only zero bits in the mantissa (or (-0.0) if the sign bit is set). Subnormal numbers are always smaller than normal numbers<sup>7</sup> and are useful to represent small values close to 0.0 which are outside the range of normal numbers<sup>8</sup>. Figure 2.14(b) shows an example for the representation of subnormal numbers with float16 and float32 precisions.

---

<sup>7</sup>The smallest normal number uses an exponent of  $(E)_2 = 1$  so that we get:

$$\begin{aligned} |(-1)^{(S)_2} \cdot 2^{1-b} \cdot (0.M)_2| &< |(-1)^{(S)_2} \cdot 2^{(E)_2-b} \cdot (1.M)_2| \\ 2^{1-b} \cdot (0.M)_2 &< 2^{1-b} \cdot (1.M)_2 \\ (0.M)_2 &< (1.M)_2 \end{aligned}$$

<sup>8</sup>For float16, the smallest normal number is

$$(0000\ 0100\ 0000\ 0000)_2 = 2^{-14} \cdot 1 = 2^{-14}.$$



When working with floating-point numbers, it may happen that `inf` or `nan` values occur and it is necessary to have a representation for these numbers. `inf` values are represented by an exponent composed entirely of 1-bits and a mantissa composed entirely of 0-bits. `nan` values are represented by an exponent with only 1-bits and a non-zero mantissa.

In total, we can represent a decimal number  $x$  with the bits  $b_0, b_1, \dots$ , as

$$x = \begin{cases} (-1)^{(S)_2} \cdot 2^{(E)_2 - b} \cdot (1.0 + \sum_{i=1}^m b_{m-i} \cdot 2^{-i}) & \text{for normal numbers} \\ (-1)^{(S)_2} \cdot 2^{1-b} \cdot \sum_{i=1}^m b_{m-i} \cdot 2^{-i} & \text{for } (E)_2 = 0 \\ (-1)^{(S)_2} \cdot \infty & \text{for } (E)_2 = 2^e - 1 \text{ and } (M)_2 = 0 \\ \text{nan} & \text{for } (E)_2 = 2^e - 1 \text{ and } (M)_2 \neq 0 \end{cases} \quad (2.11)$$

The number of bits allocated for the mantissa and the exponent as well as the bias employed are dependent on the precision. Additionally to the examples of Figure 2.14, Table 2.1 summarizes some basic information about these precisions.

**Table 2.1:** Basic information about binary floating-point numbers with float16 and float32 precisions (see Figure 2.14 for an example). The number of significant digits denotes how many decimal digits can be represented exactly and is based on the number of bits in the mantissa  $m$  via  $\log_{10}(2^{m+1})$  [69]. The +1 originates from the implicit bit of the mantissa which effectively increases the number of accuracy bits by 1. Significant digits are not meant to be interpreted literally (it is hard to write a number with 3.31 digits) but are related to the relative error of the decimal number [36].

|                     | float16                         | float32                          |
|---------------------|---------------------------------|----------------------------------|
| # exponent bits $e$ | 5                               | 8                                |
| # mantissa bits $m$ | 10                              | 23                               |
| bias $b$            | 15                              | 127                              |
| # finite values     | 63 489                          | 4 278 190 081                    |
| minimum             | -65 504                         | $-3.40 \cdot 10^{38}$            |
| maximum             | 65 504                          | $3.40 \cdot 10^{38}$             |
| smallest normal     | $2^{-14} = 6.10 \cdot 10^{-05}$ | $2^{-126} = 1.18 \cdot 10^{-38}$ |
| smallest subnormal  | $2^{-24} = 5.96 \cdot 10^{-08}$ | $2^{-149} = 1.40 \cdot 10^{-45}$ |
| significant digits  | 3.31                            | 7.22                             |

Due to the exponential nature of Equation 2.11, a broad range of numbers can be represented with the accuracy being constrained by the mantissa. Essentially, the exponent defines a window between two consecutive exponential numbers (e.g.,  $[2^3; 2^4[$ ) and the mantissa determines the offset within that window [191]. For the example of 16.25, where we need only one bit in the mantissa to be set to 1 (because  $0.25 = (0.01)_2$ ), we are in the range  $[2^4; 2^5[ = [16; 32[$  so we can set the exponent to  $2^4$  and get a number  $2^4 \cdot (1.M)_2 \geq 16$ . Next, we need to identify which bit  $j$  we need to set so that  $2^4 \cdot 2^{-j} = 2^{-2} = 0.25$ . This

leads us to the bit  $j = 6$  so that we need to set the 6th bit to 1, i.e.,  $M = (0.000001)_2$ . For other numbers that are not as easily represented in the binary system, the process becomes more complex but the principle remains the same: first define a window with the help of the exponent, then the offset via the mantissa.

It is also worth noting that not every decimal number can be represented exactly in the binary system. For example, 0.1 can only be approximated as (using float16 precision)

$$(0010\ 1110\ 0110\ 0110)_2 = 2^{-4} \cdot (1 + 2^{-1} + 2^{-4} + 2^{-5} + 2^{-8} + 2^{-9}) = \frac{819}{8192} \approx 0.09998. \quad (2.12)$$

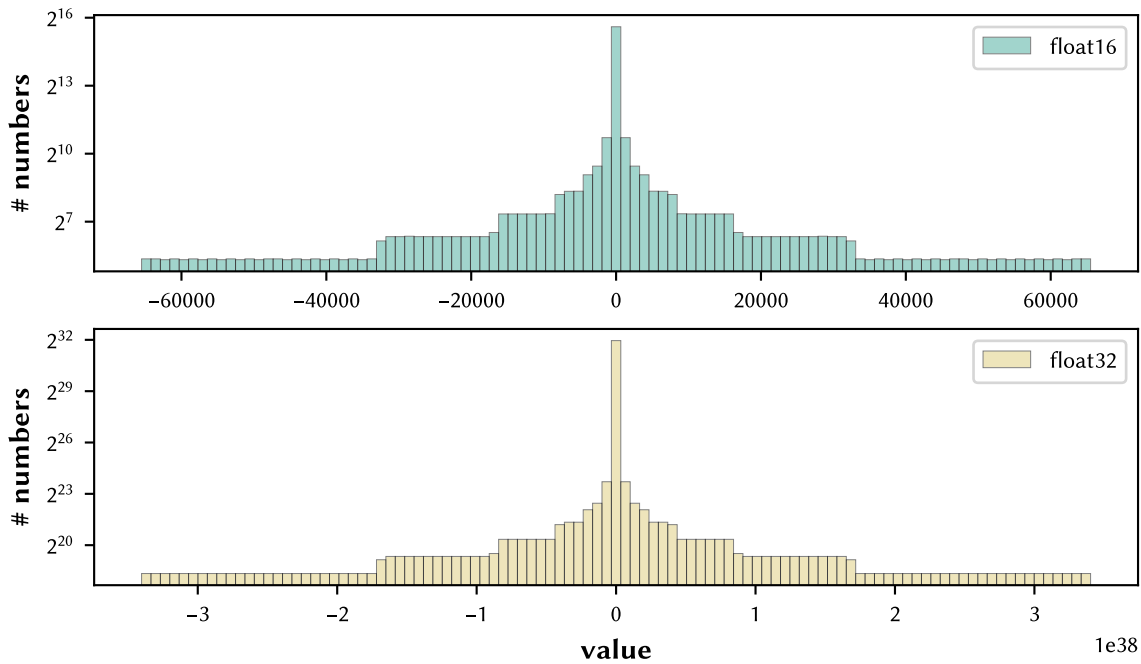
The precisions float16 and float32 primarily differ in two aspects: the range of numbers that can be represented and the accuracy of these numbers. Figure 2.15 illustrates the range as a histogram of every possible finite number for both precisions. Firstly, it is evident that the range of float32 is significantly larger than that of float16. Secondly, the distribution of values across the range of possible numbers is not uniform: there are more numbers around 0.0 (the range that includes subnormal numbers) and fewer numbers at the ends of the range spectrum. This is because for larger values of the exponent, despite the increase in window size, we still have the same number of bits in the mantissa to represent offsets within that window. As a result, the same number of offset values are spread across a wider window range. For instance, with float16 precision, we can represent  $2^{10} = 1024$  values within the range  $[2^1; 2^2[ = [2; 4[$  but also only 1024 values within the range  $[2^{15}; 2^{16}[ = [32768; 65536[$ .

Figure 2.16 provides insights into the accuracy of float16 and float32 numbers in relation to various decimal scale levels. For float16, we can see that the relative errors start to increase at the  $10^{-3}$  scale level and become even more pronounced at the  $10^{-4}$  scale level. A similar trend occurs for float32 at the  $10^{-7}$  and  $10^{-8}$  scale levels. This is effectively a visualization of the number of significant digits per precision (see Table 2.1).

### Autocasting

The float32 precision offers superior accuracy and can represent a wider range of values compared to float16. However, float16 requires less memory storage to represent values, demands less memory bandwidth to transfer values and operations with this data type are faster to compute [51]. The basic idea of autocasting is to utilize float16 whenever feasible and resort to float32 only if necessary. Ideally, this process occurs automatically, eliminating the need for manual user interventions. This is achieved by defining for each operation whether it is usually safe to execute it in float16 precision or not. This concept is implemented through autocasting regions as depicted in Figure 2.17.

If the operation is considered to be safe in float16, the inputs are cast to float16 and the result will also be in float16. Conversely, if the operation is not assumed to be numerically stable in float16, the inputs are cast to float32 and the result will be in float32 as well.

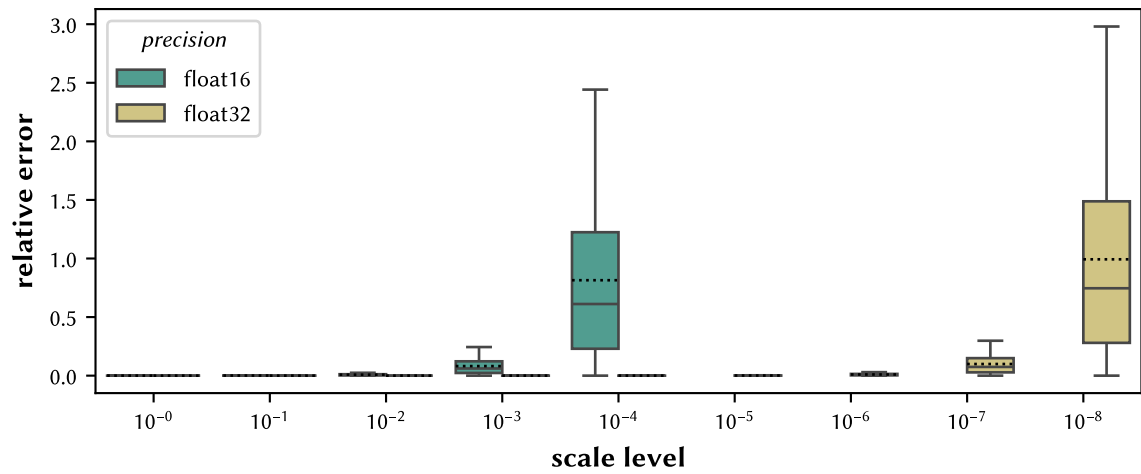


**Figure 2.15:** Range of float16 and float32 numbers. For each precision, a histogram based on all finite values (i.e., excluding `inf` nor `nan`) which can be represented is shown using 101 equally-sized bins (101 instead of 100 for symmetry of positive and negative values). Please note the logarithmic scale on the  $y$ -axis. There are 63 489 (of  $65\,536 = 2^{16}$  total values) and 4 278 190 081 (of  $4\,294\,967\,296 = 2^{32}$  total values) finite values for float16 and float32, respectively.

A prominent example of an operation that is generally safe to perform in float16 is matrix multiplication which is particularly relevant for CNNs because they make heavy use of this operation. On the other side, operations that pose a risk of propagating errors from multiple values, such as summation, are typically not safe to perform in float16 precision and are performed in float32 instead.

### Loss Scaling

In principle, enabling autocasting is all that is required to train neural networks with mixed precision. However, the gradients computed during backpropagation pose a potential challenge because they may be small and either fall outside the representable range of float16 or cannot be represented with sufficient accuracy. This problem is illustrated in Figure 2.18 which displays a histogram of the gradients from a neural network training with float32 precision. We can see that if we had opted for float16 precision instead, most of the gradients would have been reduced to zero as they are too small to be represented with float16.



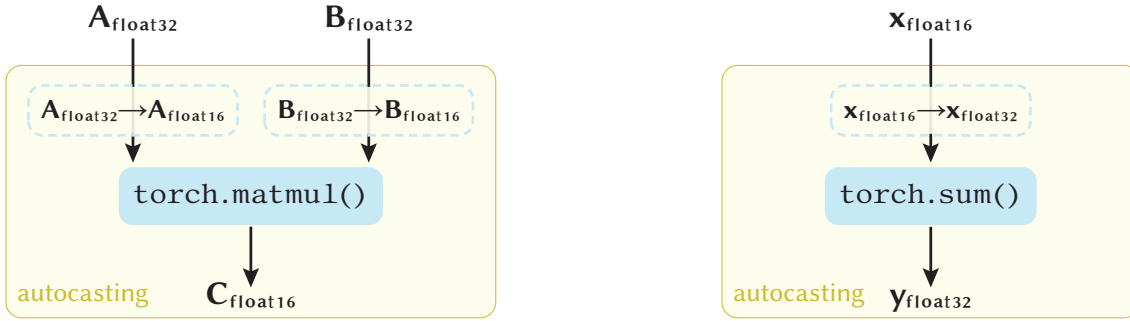
**Figure 2.16:** Accuracy of float16 and float32 numbers. For each precision, the distribution of relative errors from random numbers is shown against various scale levels. Using the float16 precision as an example, the relative error  $|\text{float16}(x) - x|/s$  is calculated for each random number  $x$  and scale level  $s$ , i.e., by casting the values to the respective precision and comparing it with various scale levels (the random numbers  $x$  and the scale levels  $s$  are represented as float64 values to serve as reference with sufficient precision). Scale levels smaller than  $10^{-4}$  are not shown for the float16 precision to improve clarity because the relative errors would become too large. Each boxplot shows the interquartile range (IQR) with the median (solid line) and mean (dotted line). The whiskers extend up to 1.5 times of the IQR. Outliers are not shown for brevity. For each boxplot, 1 000 000 random numbers from the range  $[0; 1[$  were generated (this includes the usage of subnormal numbers).

Keep in mind that the smallest representable number  $x \neq 0.0$  in float16 is

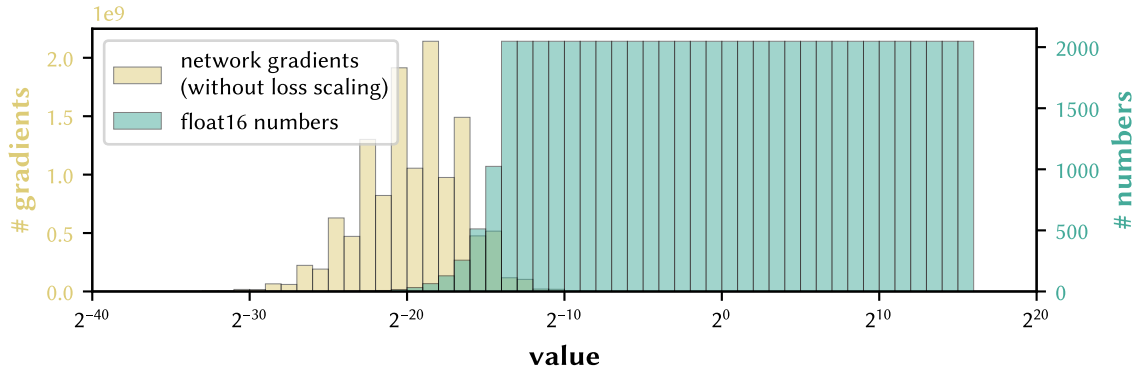
$$(0000\ 0000\ 0000\ 0001)_2 = 2^{-14} \cdot 2^{-10} = 2^{-24} \quad (2.13)$$

and anything smaller than that will be rounded to 0.0. What is more, even for numbers  $|\cdot| < 2^{-14}$  (subnormal numbers) the available values decrease rapidly which limits the accuracy of gradients in this range.

However, float16 is not inherently incapable of representing the range of gradient values. As indicated by Figure 2.18, if the gradients were shifted to the right toward the representable range of float16, everything would work fine. This is the core idea behind loss scaling: before backpropagation, we scale the loss which in turn scales the resulting gradients and shifts them into the representable range of float16. Prior to applying the weight updates, we reverse the scaling to ensure the weights are updated with the correct values. After this unscaling, it is possible to carry out gradient-related operations, such as gradient clipping, which do not demand further adjustments if applied after the unscaling. The concept of training with loss scaling is illustrated in Figure 2.19.

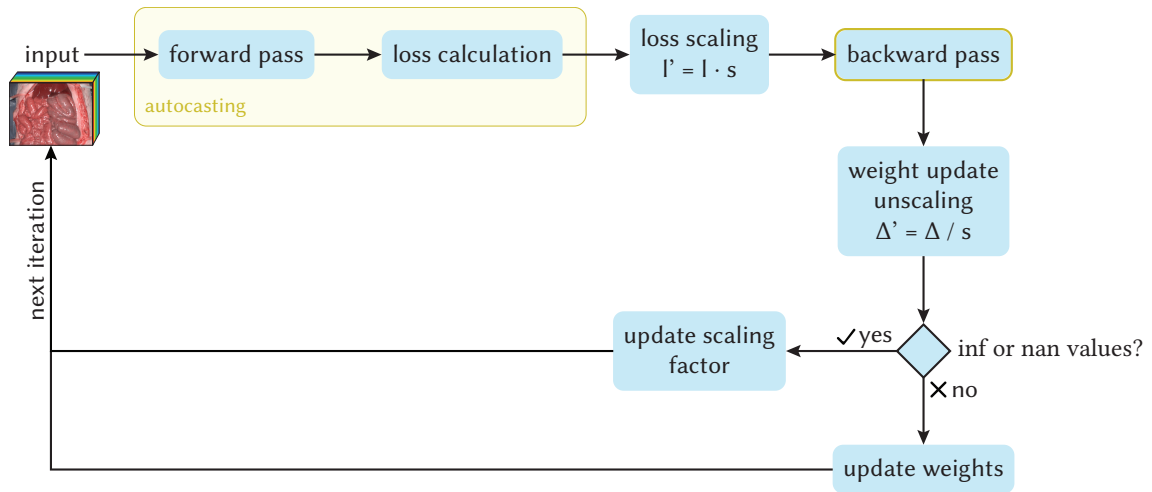


**Figure 2.17:** Basic principle of autocasting. If an operation is considered numerically stable in float16 (e.g., matrix multiplication), autocasting casts the inputs to float16 and the result will be float16 as well. Other operations (like summation of values) are known to require higher precision and hence cast their inputs to float32 and also produce float32 outputs. The behavior of common operations in an autocasting environment is listed in the PyTorch documentation [222].



**Figure 2.18:** Distribution of gradients in neural network training without loss scaling in comparison with the representable range of float16. The segmentation network was trained with float32 precision without autocasting and the gradients were recorded for every layer. The histograms use a bin width of 1 and please note the logarithmic scale on the x-axis. For the gradients and the float16 numbers, first the absolute value and then the binary logarithm were taken and only finite values were kept. The segmentation network is a hyperspectral image model with the same setup as described in Section 5.2.1. Per default (without loss scaling, see Figure 2.19), the gradients become very small and fall outside the representable range of float16 so that it would be inadvisable to train the same network with float16 without further adjustments.

There is no single scaling factor that is suitable for every neural network training. This necessitates the scaling factor to be variable over the course of training time so that it can be adjusted as needed to prevent infinite values. Typically, it is initialized with a large



**Figure 2.19:** Concept of mixed precision training with loss scaling. The forward pass and the calculation of the loss are performed in an autocasting environment, i.e., some operations are performed with float16 precision. Then, the loss is scaled so that the resulting gradients in the backward pass are shifted to the representable range of float16. The backward pass does not run in an autocasting environment but the operations remember the used type from the forward pass and the same type is used when performing the backward operations [221]. After the backward pass, the weight updates are unscaled and if no infinite values occur (no `inf` or `nan` values), they get applied and the next iteration continues. Otherwise, the weights remain unchanged and the scaling factor is adjusted so that infinite values are avoided in the next iteration.

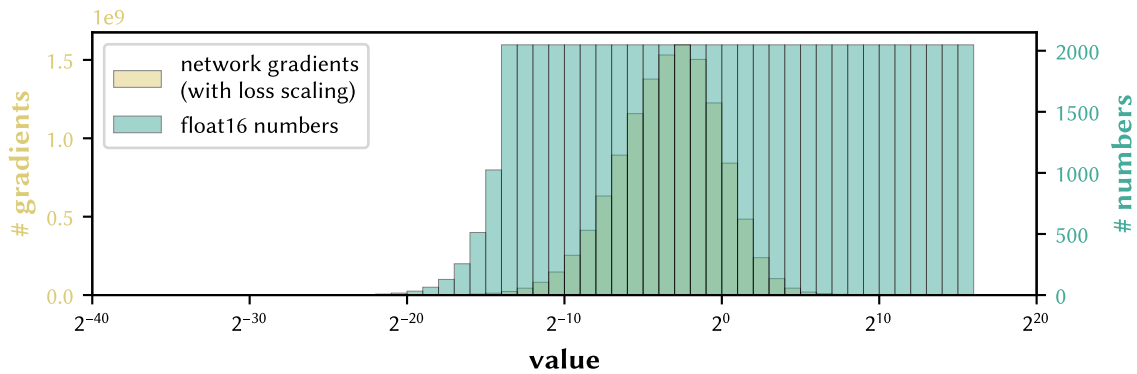
value and if infinite values arise, the factor is increased [222]. As illustrated in Figure 2.19, the scaling factor may undergo updates after each training epoch.

In Figure 2.20, we see the same gradient distribution as before but this time the training was conducted with loss scaling and mixed precision. The loss scaling successfully shifted the gradients into the representable range of float16.

The relative error of the gradients, both with and without loss scaling, is visualized in Figure 2.21. We can observe, and this was already evident from Figure 2.18, that the errors are quite substantial without loss scaling. What is more, around 5 % of the gradients would become zero in float16 precision without loss scaling. When we employ loss scaling, however, the errors are drastically reduced.

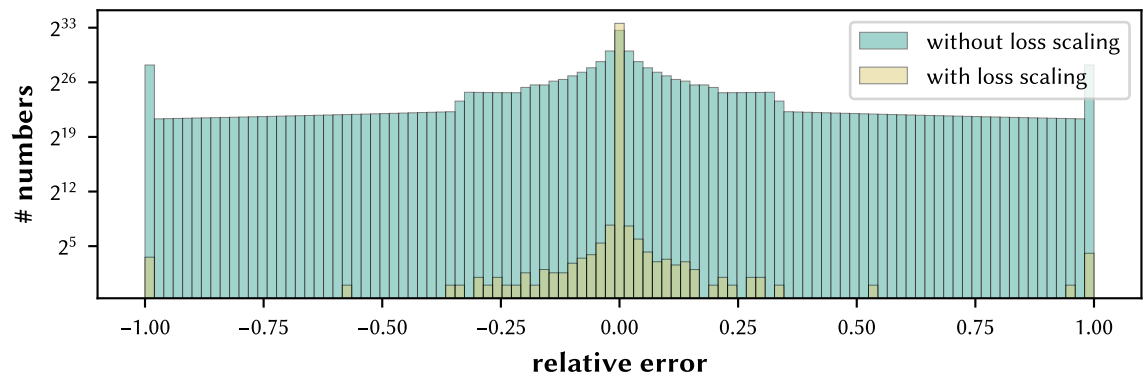
### Network Comparison

Ultimately, mixed precision training must fulfill three promises: faster training, less memory usage and maintaining a similar level of accuracy. Figure 2.22 presents the results of training a segmentation network with float32 and mixed precision settings



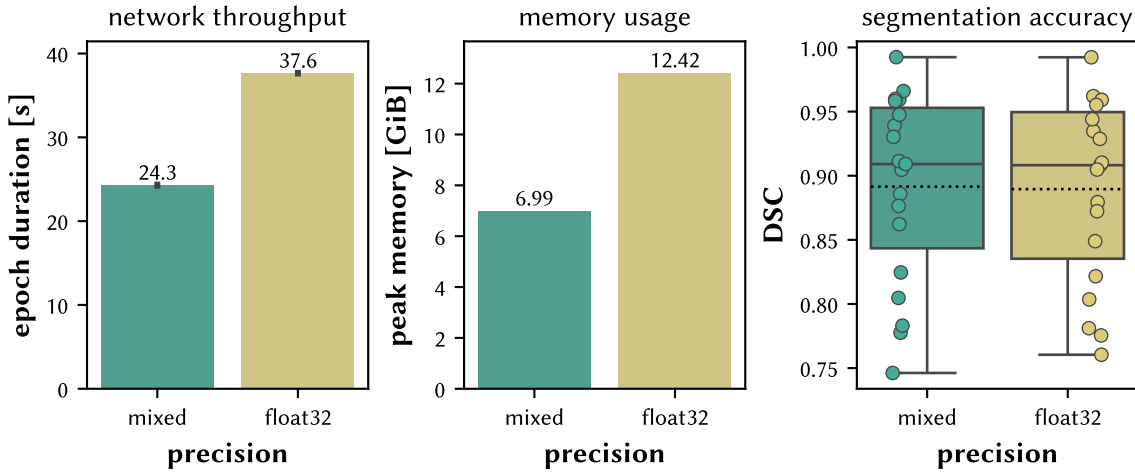
**Figure 2.20:** Distribution of gradients in neural network training with loss scaling in comparison with the representable range of float16. The segmentation network was trained with float16 precision with autocasting and the gradients were recorded for every layer. The histograms use a bin width of 1 and please note the logarithmic scale on the x-axis. For the gradients and the float16 numbers, first the absolute value and then the binary logarithm were taken and only finite values were kept. The segmentation network is a hyperspectral image model with the same setup as described in Section 5.2.1. By scaling the loss prior to backpropagation (cf. Figure 2.19), we can effectively shift the resulting gradients in the representable range of float16.

evaluating these three criteria. As we can see, float16 has shorter epoch times and requires less memory all while maintaining a similar level of segmentation accuracy.



**Figure 2.21:** Comparison of the relative error of gradients from networks with and without loss scaling. The relative error  $|\text{float16}(\delta) - \delta|/\delta$  is computed for every gradient  $\delta$  in the network by casting the gradient to float16 and comparing the error to the actual value (again using float64 to represent the true value of the gradients). Values of  $\pm 1$  indicate that the corresponding gradient is zero in float16. Please note the logarithmic scale on the y-axis. Each histogram uses 101 equally-sized bins (101 instead of 100 for symmetry of positive and negative values).





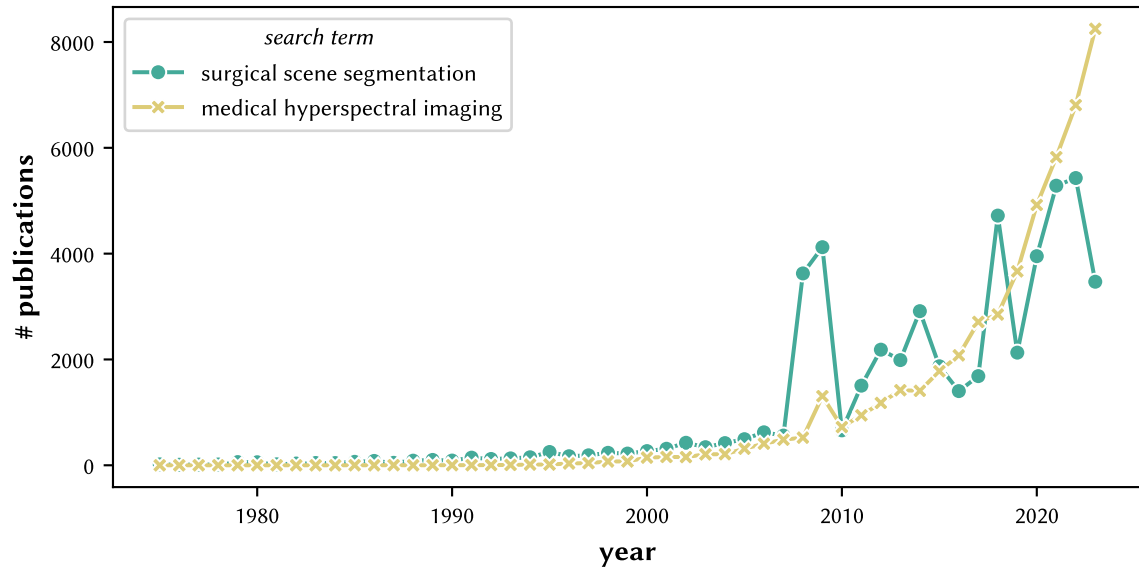
**Figure 2.22:** Comparison of segmentation networks trained with different precision settings. Networks trained with mixed precision are faster to train and require less memory while maintaining similar segmentation accuracy compared to networks trained with float32 precision. The error bars in the network throughput diagram show the standard deviation (SD) across three repetitions of the experiment (the SD is very small). There is no SD in the memory usage diagram because it does not change across repetitions. Each boxplot shows the interquartile range (IQR) with the median (solid line) and mean (dotted line). The whiskers extend up to 1.5 times of the IQR. Each point represents the aggregated class-level performance. The segmentation networks are hyperspectral image models. The measurement of the network throughput follows the setup of Section 5.2.1 (RQ2). The (peak) memory usage and the segmentation accuracy (measured via the dice similarity coefficient (DSC)) are obtained while averaging three seed networks with the setup described in Section 5.3.1 (RQ3).



## RELATED WORK

# 3

The primary objective of this thesis is surgical scene segmentation with hyperspectral images. In this chapter, we discuss related work that aligns with this goal. As evidenced by Figure 3.1, both surgical scene segmentation as well as medical HSI are areas of increasing interest as indicated by the rising number of publications each year in these fields.



**Figure 3.1:** Number of publications per year in the surgical scene segmentation and medical hyperspectral imaging fields. The years range from 1975 (including) to 2023 (including). The data for this chart was obtained on January 9, 2024 from Digital Science’s Dimensions platform, available at [app.dimensions.ai](https://app.dimensions.ai) [95].

In the following, we discuss related work for our spectral analysis (RQ1) in Section 3.1, our efficient data loading pipeline (RQ2) in Section 3.2, our segmentation networks (RQ3) in Section 3.3 and our analysis on domain shifts (RQ4) in Section 3.4. The chapter concludes in Section 3.5 with a discussion on how the related work impacts this thesis.

**Table 3.1:** Summary of basic information about publications that analyzed spectral characteristics of visceral organs (see text for details).

| publication            | year | wavelength range | # channels | # subjects | # classes |
|------------------------|------|------------------|------------|------------|-----------|
| Zheng et al. [252]     | 2015 | 600 nm–800 nm    | 100        | 3          | 3         |
| Leavesley et al. [126] | 2016 | 390 nm–450 nm    | 12         | 4          | 2         |
| Zhang et al. [250]     | 2017 | 470 nm–700 nm    | 8          | 3          | 4         |
| Baltussen et al. [17]  | 2019 | 400 nm–1700 nm   | 360        | 32         | 3         |
| Maktabi et al. [143]   | 2019 | 500 nm–1000 nm   | 100        | 11         | 2         |
| Hu et al. [96]         | 2019 | 410 nm–910 nm    | 200        | 30         | 2         |
| Zhang et al. [249]     | 2022 | 376 nm–1038 nm   | 128        | 8          | 2         |

### 3.1 Spectral Organ Fingerprints

Spectral organ fingerprints have found applications in computer-assisted decision-making and automated organ identification<sup>1</sup> [40]. On the one hand, some studies focused on specific biological pigments such as hemoglobin, melanin and porphyrin in spectral data but they did not analyze spectral characteristics across different organs [225, 238]. On the other hand, other works that analyzed spectral characteristics of tissues are not without their own limitations, as highlighted by the following list of relevant publications concerning the discrimination of visceral organs. A summary of basic information about these publications is provided in Table 3.1.

- Zheng et al. detected cervical intraepithelial neoplasia using spectral data from three patients [252]. They categorized the data into three groups: normal tissue, inflammation and high-grade lesions. An analysis based on second-order derivatives was performed demonstrating that in vivo and noninvasive detection of cervical neoplasia without acetic acid is technically feasible. Furthermore, they discovered that only three specific wavelengths (620 nm, 696 nm and 772 nm) are required for tissue classification with optimal separability. However, they did not perform any automated differentiation, used only a small sample size and did not include the near-infrared range of the spectrum.
- Leavesley et al. conducted an analysis using data from four patients and an animal study to detect cancerous colon tissue [126]. They measured changes in the fluorescence excitation spectrum that occur in tandem with colonic adenocarcinoma. Their findings revealed significant spectral differences between normal and cancerous tissues. However, they used a custom hyperspectral imaging device, a narrow wavelength range of 390 nm–450 nm, a limited sample size and did not present any automated classification results.

<sup>1</sup>This section is based on [215].

- Zhang et al. used multispectral data from ex-vivo porcine organs (obtained from a butcher) in a laparoscopic surgery setting to distinguish between four different tissue types: liver, gallbladder, colon, and kidney [250]. By leveraging spatial and spectral features, they achieved an accuracy of 98.4 % and they highlighted the superior performance of multispectral imaging data over RGB data. However, their study was limited by the use of ex-vivo material, a custom multispectral imaging device and the chosen channels were not broad enough to represent real RGB images.
- Baltussen et al. combined two different HSI cameras and acquired ex-vivo data from 32 patients to yield samples from healthy, fat and colorectal cancer tissue [17]. They focussed on real-time classification and achieved a classification accuracy of 88 %. However, they used cross-section slices which is not possible to obtain during intraoperative surgery, employed a custom HSI setting and used only ex-vivo data.
- Maktabi et al. utilized HSI recordings to classify malignant from healthy tissue of oncologic esophageal resectates from 11 patients [143]. They evaluated the performance of four different supervised machine learning algorithms and found that a support vector machine (SVM) yielded the best results. However, the untouched test set comprised only two patients (one patient with only cancerous tissue and another patient with both cancerous and normal tissue) and they achieved relatively poor classification results of 63 % sensitivity and 69 % specificity.
- Hu et al. distinguished between cancerous and normal gastric tissue using microscopic hyperspectral images from 30 patients with gastric cancer [96]. In total, their spectral database amounts to 28 542 spectral samples. They employed three-dimensional convolutional neural networks for spatio-spectral feature combination and yielded a classification accuracy of 97.57 %. However, their study was limited to the use of ex-vivo material and a custom HSI system.
- Zhang et al. differentiated between normal and necrotic small bowel tissue in animal experiments comprising eight rats [249]. They compared six different supervised classification algorithms and found that the spectral samples could be well separated. However, only two small regions per rat were considered, a custom HSI device was employed and the sample size was limited.

In conclusion, previous studies provided initial evidence of spectral organ fingerprints but they fall short in conducting systematic, large-scale analyses with a greater number of in-vivo organs and a standard hyperspectral imaging system (except for [143]) with a wide enough spectral range. For instance, as highlighted by Table 3.1, the current state of the art does not distinguish between more than four different organ classes. However, it is necessary to overcome these limitations to conclude whether organs feature unique spectral fingerprints.

## 3.2 Efficient Training of Hyperspectral Segmentation Networks

The acceleration of training and inference of deep learning models is a widely discussed topic with a plethora of resources, performance guides, and libraries available [167, 127, 224]. It is important to note that bottlenecks can occur at any stage of the training pipeline and a slowdown at one step can negatively impact all subsequent steps. There are numerous potential performance enhancements on the GPU side, such as mixed precision training for more efficient hardware utilization (cf. Section 2.3.2), the prevention of (implicit) synchronization points between the central processing unit (CPU) and GPU [224] or automatic graph optimizations (e.g., through the latest advancements in `torch.compile` [223]). Here, our focus is on the data loading aspect as this is the most critical component for training HSI segmentation networks.

Efficient data loading is so crucial since it is the first step of the pipeline and if this step is slow, it determines the runtime of everything that follows, i.e., the actual computations on the GPU. Moreover, the process is highly dependent on the application domain (e.g., whether we deal with images, text, or tabular data), the environmental conditions (e.g., loading data from remote storage vs. loading data from a local solid-state drive (SSD)) and the task to be solved (e.g., the dataset size plays a crucial role and whether the data fits into the system cache). This has led to the introduction of various libraries, research projects and storage formats targeted at improving the data loading process for their specific needs [4, 154, 167]. Here, we will highlight an example from the general computer vision community.

Fast Forward Computer Vision (FFCV) is a popular library by Leclerc et al. that optimizes the training pipeline by avoiding data bottlenecks [127]. It introduces a novel data storage format for enhanced indexability and faster access, optimizes caching schemes, offloads tasks to the CPU and employs just-in-time (JIT) compiled data augmentations. They benchmarked training setups on the ImageNet dataset [55] and achieved accuracy comparable to baseline models in half of the time. However, FFCV is designed for RGB images and does not utilize efficient compression formats nor does it support GPU augmentations or optimized memory transfer to the GPU via pinned memory.

In conclusion, while there are numerous strategies to optimize the data loading pipeline, the most efficient solution depends on the specific application, environmental conditions and the downstream task. What is more, none of the current solutions are targeted at the specific needs of HSI data. Unfortunately, what is efficient for one domain may not necessarily be efficient for another. For instance, offloading tasks to the CPU can be advantageous when processing RGB images but this might not hold true when processing HSI data due to the high dimensionality of the data and the resulting need for massively parallel computations. As a result, efficient training of HSI segmentation networks necessitates manual optimizations of the data loading pipeline.

### 3.3 Surgical Scene Segmentation

For the specific task of surgical scene segmentation with hyperspectral images, only a limited number of works exist<sup>2</sup>. Therefore, we initially explore the field of surgical scene segmentation with RGB images. HSI also has various applications beyond the medical field which is why we review the work in the non-medical domain before we elaborate on the medical applications.

#### Surgical Scene Segmentation With RGB Data

In recent years, surgical scene segmentation with RGB data has found several applications, particularly in minimally invasive procedures such as cataract and colorectal surgeries [81, 142, 187]. However, the primary focus has been on the segmentation of medical instruments, driven by numerous challenges in this field (e.g., the CATARACTS challenge for automatic tool segmentation in cataract surgery [8] or EndoVis challenges [210] like the Robust Medical Instrument Segmentation challenge in laparoscopic surgery [142, 190]). Organ segmentation has been less explored, with a few studies either limiting their scope to specific organ classes [74, 70] or, more commonly, investigating full scene segmentation [111, 10, 139, 194]. The datasets used in these studies vary greatly in terms of annotation sparsity and the number of classes considered. The models typically process video frames of different sizes (e.g.,  $512 \times 512$  in [125] or  $960 \times 540$  (width, height) in [194]).

Organ segmentation in open surgery has been less frequently addressed likely due to the greater complexity and variability of the surgical scene as well as the challenges associated with image acquisition. To the best of our knowledge, only the study of Gong et al. has examined deep learning-based organ segmentation on RGB images in open surgery where they investigated segmentation performance under various imaging conditions, such as changes in lighting or distance, using RGB images from 130 patients and found that these factors significantly affect the image scores. [79]

Previous research has highlighted numerous significant obstacles in automatic surgical scene segmentation with RGB data. These include a high degree of variability in tissue appearance across different subjects (e.g., [44, 74]) and within individual images (e.g., due to occlusions or deformations [153]) as well as inconsistencies in image acquisition [140]. The incorporation of additional spectral information could be instrumental in overcoming these challenges since spectral imaging can encode extra clinical information (e.g., tissue parameters like oxygenation) and has additional features that could aid in situations with limited context [68].

#### Segmentation With Hyperspectral Data Outside the Medical Field

Spectral imaging is utilized in a variety of domains such as biochemistry, agriculture, archaeology and, notably, remote sensing [114, 177]. The generalizability of existing

---

<sup>2</sup>This section is based on [198].

studies is often compromised due to the small size of the datasets which typically consist of only one or two sparsely annotated images and the fact that training and testing are conducted on the same data [9, 159, 161, 172]. The application of deep learning-based semantic scene segmentation is generally hindered by the scarcity of available annotations since training data is sparse and labels often only cover several discrete pixels rather than entire images [229]. However, given the high dimensionality of the data, large datasets are necessary to prevent overfitting [254]. As a result of these constraints, most segmentation tasks in these fields resort to pixel-based classification (e.g., [161]) and the few existing patch-based or image-based segmentation approaches run a high risk of train-test leakage since no disjoint datasets are used [162].

From a methodological perspective, a variety of spatial and spectral features have been proposed with different fusion methods (e.g., the weighted z-score normalized fusion method presented in [108]). Deep learning-based models have also been employed, utilizing one-dimensional spectral networks [107], two-dimensional spatial networks [246, 243] and three-dimensional spatio-spectral networks [88]. However, it is worth noting that three-dimensional convolutions have not consistently proven to be beneficial [177, 144].

It is evident from these works that research on small datasets provides severe challenges, especially when it comes to accurately validating the algorithms. This requests the need of carefully designed experiments on sufficiently large datasets.

#### **Segmentation With Hyperspectral Data Inside the Medical Field**

There are only a handful of papers that tackle a biomedical segmentation problem using spectral data with deep learning [115]. Therefore, we also include approaches that employ non-deep learning techniques in the following list. A summary of basic information about these publications can be found in Table 3.2.

- Akbari et al. collected 7 HSI images of abdominal organs from a single pig. They annotated 5 organs (spleen, colon, small intestine, bladder, peritoneum) in these images and conducted pixel-based organ classification using learning vector quantization [118] of compressed spectra [6]. However, due to the limited size of the dataset (which only includes one individual) and the unclear separation between training and testing data, a further evaluation on an independent test set involving a larger number of individuals has yet to be conducted.
- The HELICoiD project [64] explored the potential of HSI in distinguishing between tumorous and healthy brain tissue in neurosurgery patients. The complete dataset, which is publicly available, consists of 36 HSI images acquired by combining two HSI systems from 22 patients [66]. Sparse annotations of 4 classes (tumor tissue, normal brain tissue, blood vessels, and background) were generated by merging manual expert segmentations based on pathological findings with  $k$ -means



**Table 3.2:** Summary of basic information about publications that tackle biomedical segmentation problems (see text for details).

| publication                   | year | wavelength range | # channels | # subjects | # classes |
|-------------------------------|------|------------------|------------|------------|-----------|
| Akbari et al. [6]             | 2008 | 900 nm–1700 nm   | 320        | 1          | 5         |
| Fabelo et al. [64]            | 2016 | 400 nm–1700 nm   | 998        | 22         | 4         |
| Ravi et al. [180]             | 2017 | 400 nm–1700 nm   | 998        | 18         | 2         |
| Fabelo et al. [65]            | 2018 | 400 nm–1700 nm   | 998        | 22         | 4         |
| Fabelo et al. [63]            | 2019 | 400 nm–1700 nm   | 998        | 16         | 4         |
| Moccia et al. [153]           | 2018 | 470 nm–700 nm    | 8          | 7          | 6         |
| Garifullin et al. [71]        | 2018 | 380 nm–780 nm    | 30         | ?          | 3         |
| Trajanovski et al. [227]      | 2021 | 400 nm–1700 nm   | 640        | 14         | 2         |
| Cervantes-Sanchez et al. [32] | 2021 | 500 nm–1000 nm   | 100        | 7          | 4         |
| Lotfy et al. [134]            | 2023 | 468 nm–790 nm    | 109        | 30         | 3         |

clustering. Pixels from small clusters were discarded due to potential annotation errors. Several studies arose from this dataset:

- Ravi et al. trained a Semantic Texton Forest [204, 109] on a subset of the HELICoiD dataset which included 33 HSI brain images from 18 patients. These images were embedded with a modified version of the  $t$ -distributed stochastic neighbor approach ( $t$ -SNE) [138] to segment tumorous and healthy brain tissue [180]. However, the performance analysis was carried out on the validation dataset (used for hyperparameter tuning).
- Fabelo et al. introduced a multi-class semantic segmentation approach that fused a segmentation prediction from a supervised pixel-based SVM classifier with a segmentation prediction obtained through unsupervised clustering [65]. Quantitative validation of the segmentations could only be performed for the SVM classifier due to the sparsity of the annotations. Further, the separation between training, validation and testing data remains unclear.
- Fabelo et al. used a subset of 26 HSI brain images from 16 patients (6 with grade IV glioblastoma and 10 with normal brain tissue) to compare baseline SVM-based methods with a pixel-based deep neural network and a two-dimensional CNN classifier on small  $11 \times 11$  patches [63]. They found that both deep learning-based methods had similar performance and surpassed the SVM-based methods. However, the performance analysis was conducted on the validation dataset (used for hyperparameter tuning).
- Moccia et al. collected 57 multispectral imaging (MSI) images from 7 pigs during hepatic laparoscopic surgery [153]. They transformed the organ segmentation task into a classification problem by training an SVM on hand-crafted textural and

spectral features from automatically segmented superpixels. The model was used to classify 6 organs: liver, gallbladder, spleen, diaphragm, intestine, and abdominal wall. They demonstrated that the classification accuracy using their MSI data was higher than the accuracy achieved when using only three selected channels. However, the chosen channels were not broad enough to represent an RGB image.

- Garifullin et al. conducted an analysis of 55 retinal images using a MSI device and segmented three types of tissue (vessels, optic disc, and macula) [71]. They employed SegNet [16] and Dense-FCN [106] models and compared MSI with RGB data. However, their findings did not conclusively favor any particular model or modality.
- Trajanovski et al. conducted a segmentation of healthy and cancerous tongue tissue in 14 histopathological HSI images using an in-house dataset of 14 patients with each patient contributing one image [227]. Building upon their previous work [228], they evaluated several pixel-based networks, networks based on patches of size  $256 \times 256$  and hybrid networks that used a combination of full pixel spectra and patches with a reduced number of channels as input. They discovered that a U-Net architecture [189] based on patches yielded the best results for their specific segmentation task. However, since the performance analysis was carried out on the validation dataset (which was also used for hyperparameter tuning), an evaluation on a hold-out test set has yet to be conducted.
- Cervantes-Sanchez et al. examined 18 images from 7 patients undergoing hepatic surgery and 21 images from 7 patients undergoing thyroid surgery using HSI. They generated sparse, circular-shaped annotations for 4 organs (liver, bile duct, artery and portal vein) in hepatic surgeries and 3 organs (thyroid, parathyroid and muscle) in thyroid surgeries [32]. They evaluated the effectiveness of several machine learning methods (logistic regression [52], SVM [25], multilayer perceptron [87], and U-Net [189]) for automatic segmentation of the annotated organ classes using either single pixels or small  $8 \times 8$  patches. However, the evaluation was only conducted on the sparse annotations and the validation dataset (which was also used for tuning hyperparameters). As such, a comprehensive evaluation on fully semantically annotated images and a hold-out test set has yet to be conducted.
- Lotfy et al. collected 49 HSI images from 30 patients to segment head and neck cancer with three classes: tumor, healthy tissue, and background. They computed superpixels using the simple linear iterative clustering (SLIC) algorithm [1] and classified them using a CNN. To maintain the spatial relationship between superpixels, they introduced a graph neural network (GNN) [193] where they connected neighboring superpixels using the corresponding CNN features while treating the problem as a graph node classification task. Their evaluation demonstrated that the GNN approach outperformed a pure CNN approach due to the additional

spatial information. However, a comparison with larger patches or an image-based approach has yet to be conducted.

In summary, there have been initial promising results in the field. However, evaluations are often not comprehensive, use only sparse annotations and lack important standards, such as untouched test sets. As evidenced by Table 3.2, the publication landscape is highly heterogeneous with a wide range of custom devices (except for [32]) with different wavelength ranges which limits the reproducibility of the results while the frequent use of small datasets questions the generalizability of the findings. To this end, there has been no work on fully semantic scene segmentation with HSI data focussing on all important organs occurring during open surgery.

### 3.4 Domain Shifts in Surgical Hyperspectral Imaging

In this thesis, one of our primary objectives is to analyze the robustness of our models in the face of geometrical domain shifts<sup>3</sup>. There have been reports of significant variations across subjects, locations (such as clinics) and even measurements taken at different time-points while analyzing spectral data [105]. However, this topic is largely unexplored in the field of surgical scene segmentation. To the best of our knowledge, we are only aware of the work by Kitaguchi et al. who demonstrated that surgical instrument segmentation algorithms struggle to adapt to unfamiliar types of surgeries that incorporate known instruments in a new context. [117]

A significant contribution we make is the introduction of a novel data augmentation method aimed at enhancing context generalization. While the use of data augmentations is prevalent in the deep learning field [203], we are interested in identifying the types of augmentations most commonly utilized by the community for surgical scene segmentation tasks, especially since we conduct a validation study comparing various topology-aware augmentation methods. To this end, we reviewed 34 publications in the field of surgical scene segmentation and analyzed the augmentations that they employed. The papers we reviewed are listed in Table 3.3 and the augmentations used are illustrated in Figure 3.2.

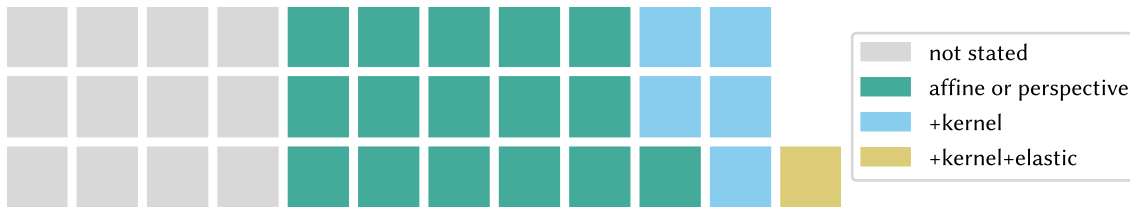
In the field of surgical scene segmentation, the community predominantly employs affine and perspective augmentations, with only a handful of instances where kernel augmentations such as sharpening or blurring are used. Elastic augmentations are even less common. Notably, topology-aware augmentations like CutMix [247] are not widely adopted, possibly because they were initially designed for classification rather than segmentation tasks. This limited use of diverse topology-aware augmentation techniques could be a contributing factor to why these networks are significantly affected

---

<sup>3</sup>This section is based on [202].

**Table 3.3:** Overview of employed data augmentations from 34 selected publications in the field of surgical scene segmentation. For each publication, it is denoted whether they use (✓) or do not use (✗) affine (rotation, scaling, etc.) or perspective augmentations, kernel augmentations (sharpening or blurring), elastic augmentations or whether this information is not stated (?).

| publication                    | year | affine or perspective | kernel | elastic |
|--------------------------------|------|-----------------------|--------|---------|
| Akbari et al. [6]              | 2008 | ?                     | ?      | ?       |
| Al-Surmi et al. [217]          | 2014 | ?                     | ?      | ?       |
| Collins et al. [44]            | 2015 | ?                     | ?      | ?       |
| Gibson et al. [74]             | 2017 | ?                     | ?      | ?       |
| Ravi et al. [180]              | 2017 | ✓                     | ✗      | ✗       |
| Fabelo et al. [65]             | 2018 | ?                     | ?      | ?       |
| Garifullin et al. [71]         | 2018 | ✓                     | ✗      | ✗       |
| Moccia et al. [153]            | 2018 | ?                     | ?      | ?       |
| Shvets et al. [205]            | 2018 | ?                     | ?      | ?       |
| Allan et al. [11]              | 2019 | ✓                     | ✓      | ✗       |
| Fabelo et al. [63]             | 2019 | ✓                     | ✗      | ✗       |
| Fu et al. [70]                 | 2019 | ✓                     | ✗      | ✗       |
| Islam et al. [102]             | 2019 | ✓                     | ✓      | ✗       |
| Kadkhodamohammadi et al. [111] | 2019 | ?                     | ?      | ?       |
| Laves et al. [125]             | 2019 | ✓                     | ✗      | ✗       |
| Trajanovski et al. [228]       | 2019 | ✓                     | ✗      | ✗       |
| Allan et al. [10]              | 2020 | ✓                     | ✗      | ✗       |
| Madad Zadeh et al. [139]       | 2020 | ?                     | ?      | ?       |
| Maqbool et al. [146]           | 2020 | ✓                     | ✗      | ✗       |
| Pakhomov and Navab [169]       | 2020 | ✓                     | ✗      | ✗       |
| Pakhomov et al. [170]          | 2020 | ?                     | ?      | ?       |
| Scheikl et al. [194]           | 2020 | ✓                     | ✓      | ✗       |
| Cervantes-Sanchez et al. [32]  | 2021 | ?                     | ?      | ?       |
| Deng et al. [56]               | 2021 | ?                     | ?      | ?       |
| Gong et al. [79]               | 2021 | ✓                     | ✗      | ✗       |
| Kong et al. [121]              | 2021 | ✓                     | ✗      | ✗       |
| Roß et al. [190]               | 2021 | ✓                     | ✓      | ✓       |
| Trajanovski et al. [227]       | 2021 | ✓                     | ✗      | ✗       |
| Kitaguchi et al. [117]         | 2022 | ✓                     | ✗      | ✗       |
| Seidlitz et al. [198]          | 2022 | ✓                     | ✗      | ✗       |
| Wang et al. [231]              | 2022 | ✓                     | ✓      | ✗       |
| Kolbinger et al. [120]         | 2023 | ✓                     | ✗      | ✗       |
| Lotfy et al. [134]             | 2023 | ✓                     | ✓      | ✗       |
| Luo et al. [136]               | 2023 | ✓                     | ✗      | ✗       |



**Figure 3.2:** Visualization of the different employed data augmentations from the 34 selected publications in the field of surgical scene segmentation (cf. Table 3.3). Each square in the waffle chart represents one publication and visualizes which augmentations were used: only affine (rotation, scaling, etc.) or perspective augmentations, additional kernel augmentations (sharpening or blurring), additional kernel and elastic augmentations or whether this information is not stated.

by geometrical domain shifts. It emphasizes the necessity for a method that enhances the robustness of these networks against such domain shifts.

## 3.5 Conclusion

In conclusion, there is a noticeable lack of research on semantic scene segmentation in open surgery, particularly in the context of medical HSI. The datasets used in existing studies are relatively small and the challenges posed by the high variability of surgical scenes and the inherent complexity, such as non-standardized image acquisition, inter-subject variability and complex three-dimensional relationships between multiple soft tissues (e.g., geometrical distortions like overlapping tissue, shadowing or deformations) [79], are yet to be fully addressed. [198]

*RQ1:* While the literature does address the differentiation between various organs interesting for surgical procedures, it is typically restricted to a limited number of classes or small datasets. More precisely, the HSI datasets in the related work include a maximum of 36 images from 22 subjects [66] and are annotated with up to 6 organ classes [153]. It is yet to be determined whether organs can still be differentiated when a larger number of classes and a larger dataset are employed. Additionally, it remains unclear whether the variability observed in the spectral data is a result of the organs being studied or specific acquisition conditions. Moreover, there is a noticeable absence of large public datasets in the community, particularly those containing many images from various subjects and a variety of tissue types.

*RQ2:* Numerous strategies exist for improving the efficiency of neural network training. While this is also true for the part of the pipeline that concerns data loading, no existing solution is targeted at the specific needs of HSI data. A solution taking

into account the high dimensionality of the data and the large image sizes has yet to be developed.

*RQ3:* In the related work on organ segmentation using spectral data, models have been developed based on superpixels [153], patches [71, 63, 32, 227] and pixels [6, 63]. However, the ideal granularity of the data in terms of segmentation performance or the number of training subjects required remains undetermined. Moreover, no previous work has demonstrated a clear advantage of spectral data over RGB data for deep learning-based surgical scene segmentation. Further, a study on a larger dataset with hundreds of semantically annotated images is yet to be conducted. [198]

*RQ4:* The literature has largely neglected the crucial issue of robustness of HSI segmentation networks against domain shifts such as those introduced by new surgeries, geometrical distortions or species changes. The impact of such domain shifts on the segmentation performance is yet to be determined. In the field of surgical scene segmentations, the community typically relies on a small set of default augmentations that are easily applicable to segmentation tasks. However, the adoption of topology-aware augmentations is not widespread.

We aim to fill these gaps in the literature by means of our research questions (cf. Section 1.2), a special focus on open surgery in contrast to minimally invasive surgery and datasets of unprecedented size. To this end, we are the first to present fully semantic scene segmentation networks that can differentiate between 19 classes occurring during open surgery, can be trained efficiently and are robust against contextual domain shifts.

## MATERIALS AND METHODS

---

This chapter introduces the materials that we work with in this thesis and the methods we used to analyze them. We will address the methodological challenges originating from the research question of Figure 1.2 which will be the foundation for the results presented in Chapter 5.

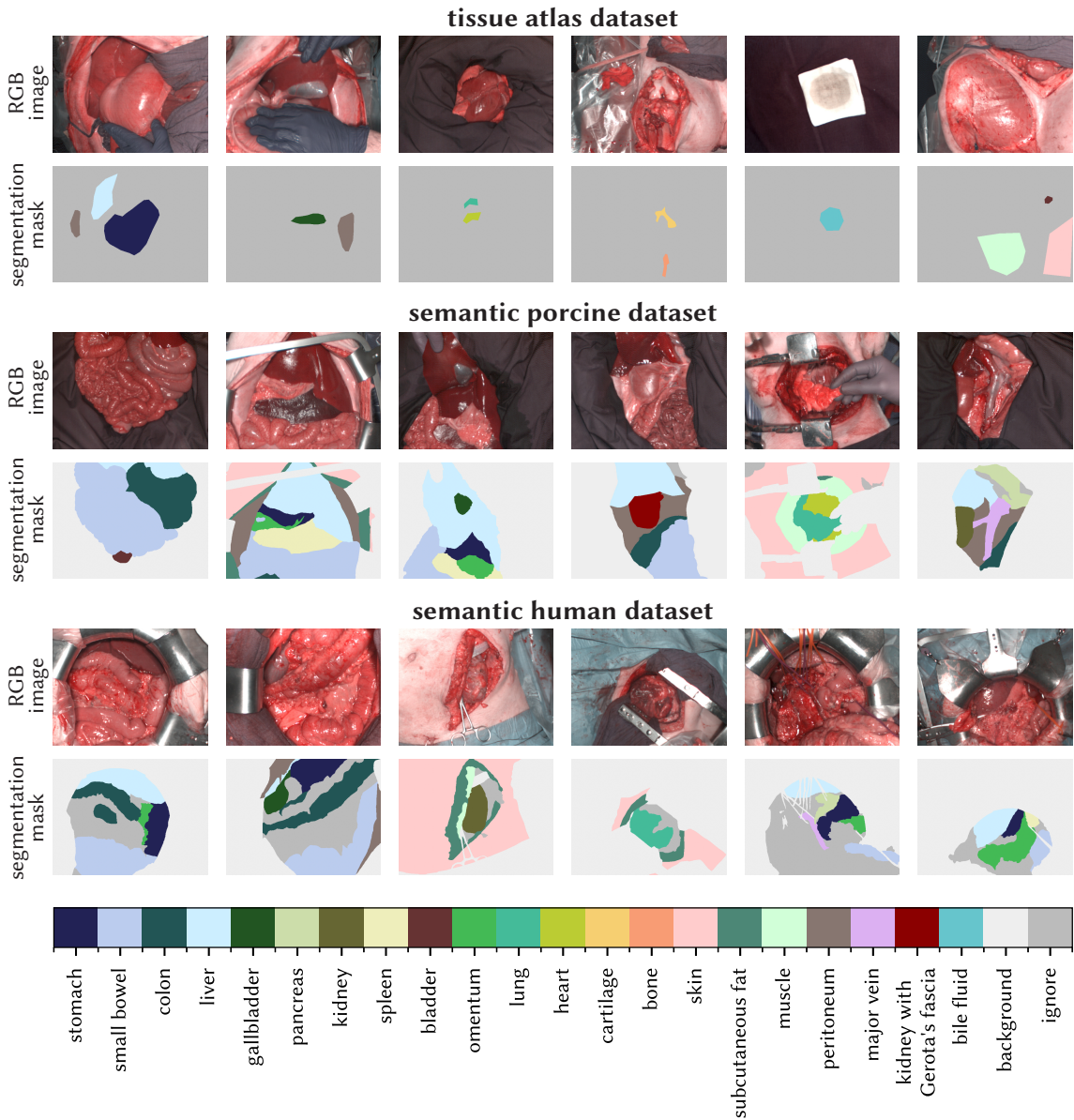
In Section 4.1, we start with a detailed presentation of our HSI datasets with an overview of the camera system, a description of the annotation process, our preprocessing pipeline and statistics about the datasets. Our classification network used to classify individual spectra (RQ1) is presented in Section 4.2. The performance optimizations that we employ to train our deep HSI networks efficiently (RQ2) are described in Section 4.3. A presentation of our segmentation networks (RQ3) follows in Section 4.4 including the architectures for the different spatial granularities and modalities as well as our training setup. Finally, in Section 4.5, we present our solution to the problem of deteriorated performance when applying machine learning networks to geometrical OOD data (RQ4).

### 4.1 Hyperspectral Datasets

This section describes the datasets used in this thesis in detail<sup>1</sup>. This includes legal aspects, the hardware system, the data acquisition process and the annotation procedure as well as our default preprocessing all described in Section 4.1.1. In general, our experiments are based on three hyperspectral datasets: the *tissue atlas dataset* (Section 4.1.2) which contains polygon annotations and the *semantic porcine dataset* (Section 4.1.3) as well as *semantic human dataset* (Section 4.1.4) with fully annotated images. Example images are shown in Figure 4.1 and general features about the datasets (distribution of organ sizes and visualization of common organ locations in the images) are described in Section 4.1.5.

---

<sup>1</sup>This section is based on [198, 215, 214].



**Figure 4.1:** Example images and corresponding segmentation masks from the tissue atlas, semantic porcine and semantic human datasets. Images of the heart show visible artifacts from the acquisition process of the line scanning device (cf. Section 2.2) due to the heartbeat (best viewed while zooming into the figure).

### 4.1.1 Data Acquisition and Preprocessing

Here, we present some legal aspects of our datasets, the HSI camera system we employed, our data acquisition process as well as our default preprocessing step.



**Ethics**

The HSI animal data was acquired at the Heidelberg University Hospital after approval by the Committee on Animal Experimentation of the regional council Baden-Württemberg in Karlsruhe, Germany (G-161/18 and G-262/19). Animals were handled according to the German laws for animal use and care and in accordance with the directives of the European Community Council (2010/63/EU).

The HSI human data was acquired during the SPACE trial (SPectrAl Characterization of organs and tissuEs during surgery) at the Heidelberg University Hospital after approval by the Ethics Committee of the Medical Faculty of Heidelberg University, Heidelberg, Germany (S-459/2020). The trial was conducted in accordance with the ethical principles of the Declaration of Helsinki [240] and the principles of Good Clinical Practice [100]. Reporting of the trial complied with the recommendations of the Consolidated Standards of Reporting Trials (CONSORT) guideline [155]. The SPACE trial was registered with Research Registry (researchregistry6281) on November 23, 2020.

**HSI Camera System**

The HSI data was acquired using the Tivita<sup>®</sup> Tissue camera system (Diaspective Vision GmbH, Am Salzhaff, Germany). This system operates in a push-broom manner, capturing hyperspectral images with an approximate spectral resolution of 5 nm within the spectral range of 500 nm to 1000 nm. This results in datacubes of dimension  $480 \times 640 \times 100$  corresponding to height, width and number of spectral channels. The camera images an area of roughly  $20 \times 30$  cm. A built-in distance calibration system, consisting of two light marks that overlap when the distance is correct, ensures an imaging distance of about 50 cm. The image acquisition process takes approximately seven seconds. Besides the HSI datacubes, the camera also computes tissue parameter images (TPI) from the HSI datacubes, which include oxygenation (tissue oxygen saturation ( $\text{StO}_2$ )), perfusion (near-infrared perfusion index (NPI)), water content (tissue water index (TWI)), lipid content (tissue lipid index (TLI)) and hemoglobin content (tissue hemoglobin index (THI) and organ hemoglobin index (OHI)). Additionally, RGB images are reconstructed from the HSI data by combining spectral channels that capture red, green, and blue light. The camera is shown in Figure 1.1 and Figure 1.6 gives an example for the reconstructed RGB image and TPI associated with an HSI datacube. More technical details on the hardware can be found in [94, 122, 76].

**Data Acquisition**

To avoid spectral distortion from straylight, all other light sources were turned off during image capture, and window blinds were closed. Motion artifacts were minimized by (1) mounting the camera on a swivel arm and ensuring the camera system remained stationary during image capture, thus eliminating camera motion, and (2) capturing images from static scenes with no surgeon-induced object movements. As a result, any motion artifacts would only be due to natural causes like respiration and heartbeat (e.g.,

visible in the heart example images of Figure 4.1). The camera perspectives were selected to provide a clear view of all organs of interest in the scene.

Non-physiological tissues are not part of this thesis. However, animals were never solely used to capture images of physiological tissue but instead were always a byproduct of a primary study. For example, if a primary study about kidney clamping was conducted, images of physiological tissue were taken before the clamping, i.e., before the primary study began. As a result, there is no fixed acquisition protocol across all animal experiments and therefore no fixed number of images per animal and no fixed set of recorded organs, camera perspectives or situs (layout of the organs relative to each other). A notable exception is the standardized recordings from the tissue atlas dataset, which were acquired according to the protocol of Figure 4.3.

This choice of image acquisition also reflects real-world surgeries where it is impractical to capture a fixed number of images from a predefined set of camera perspectives and situs per subject as no two surgeries are identical. Tissues could undergo various complications that change their state, such as inflammation or tissue trauma reflecting the need for situation-specific images.

To mitigate the impact of sensor noise and to transition the acquired HSI data from radiance to reflectance, the raw HSI datacubes were automatically calibrated using pre-recorded white and dark calibration files (cf. Equation 2.1) by the camera system, as outlined in [94]. Calibration of the camera was performed before each surgery by taking a new white and dark image to compensate for various sources of signal distortion, such as attenuation effects of the light source [68, 38]. The calibration step is also detailed in Section 2.2.

### Preprocessing

After the HSI cubes were exported from the camera system, we L1-normalized each pixel spectrum to compensate for multiplicative illumination changes that could occur due to variations in the measurement distance, for instance. That is, the HSI data is preprocessed via

$$\mathbf{s}^* = \frac{\mathbf{s}}{\|\mathbf{s}\|_1} \quad (4.1)$$

with  $\mathbf{s}^*$  denoting the L1-normalized and  $\mathbf{s}$  the reflectance spectrum from the camera. This operation is repeated for every pixel in the image separately. We are only working with L1-normalized spectra throughout this thesis.

### 4.1.2 Tissue Atlas Dataset

An overview of the tissue atlas dataset is given in Figure 4.2 and example images are shown in Figure 4.1. In total, this dataset contains 9057 from 46 subjects (see Section 4.1.1 for a description of the data acquisition procedure). The images are annotated with 20

classes with a variable number of polygon annotations per image. Part of this dataset comprises the standardized recordings (cf. Figure 4.3) which contains 5756 from 11 and is used to systematically study the effect of subject, situs, angle, repetition and annotation factors.

Medical experts created 17 775 organ annotations on the reconstructed RGB images using the HyperGUI tool [149]. Annotations were verified by two other medical experts. In the case of conflicts, annotations were repeated collectively with all annotators agreeing on the final annotation.

For the subset of standardized recordings, two additional medical experts annotated the already verified 5758 annotations for analysis of the annotation effect of the (inherently random) polygon annotations. Additionally, the medical experts who annotated the first round of polygon annotations re-classified the polygon annotations (deciding about the organ given an annotation) achieving an intra-rater agreement of 100 %. An independent secondary medical expert achieved an inter-rater agreement of 99.5 % (27 of 5758 were misclassified) for this task.

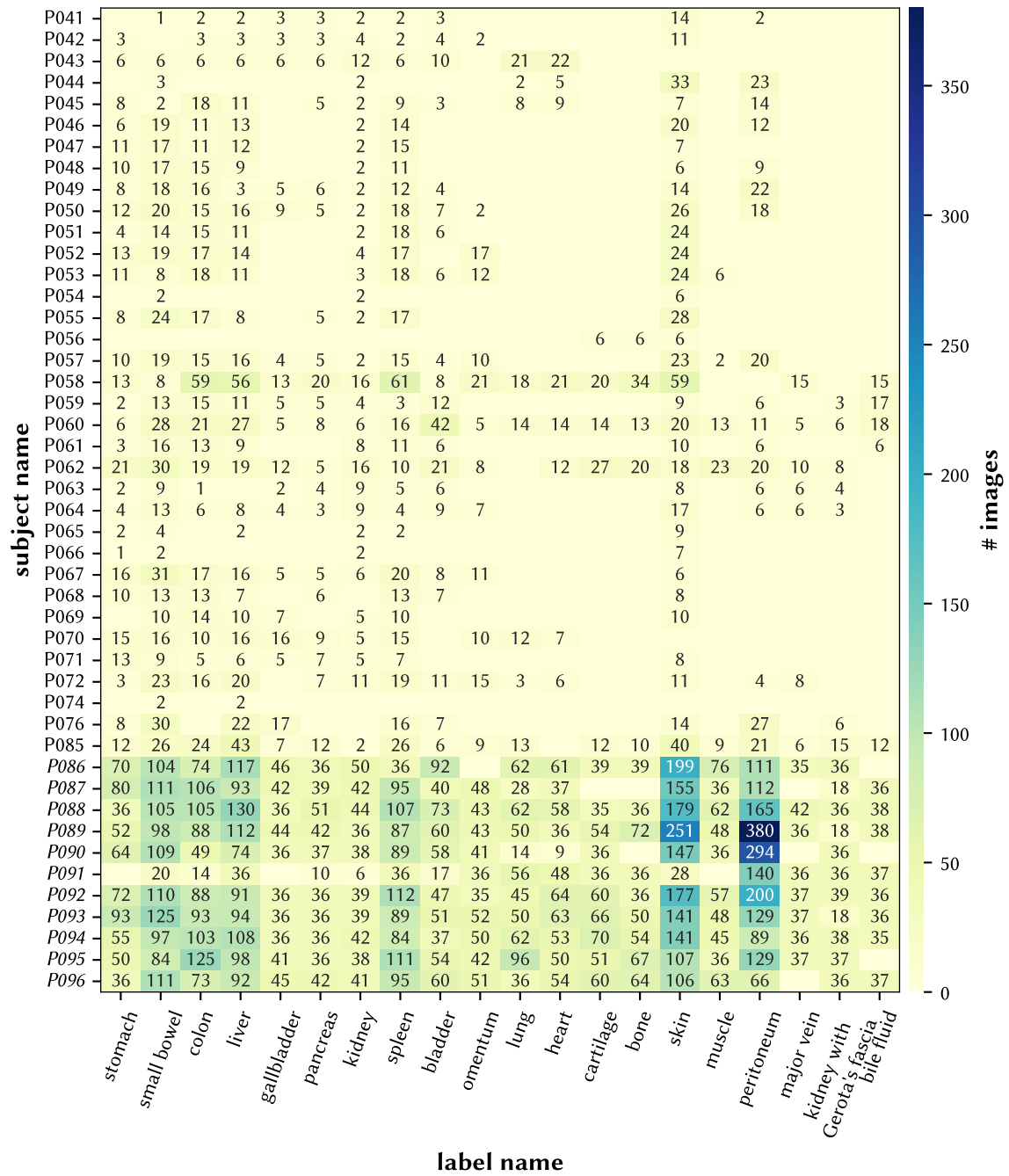
The non-semantic annotation process was carried out using a multi-point selection tool. Areas for polygon selection were chosen while excluding any regions with artifacts such as tissue kinking, illumination shadows, peripheral areas, superficial blood vessels and fat, contamination with dyes or body fluids like bile fluid, previous manipulations like contusion or abrasion, and potential perfusion impairments like thrombosis. The goal was to select only the most representative regions, ensuring that the analyzed pixels were always fully representative of the organ of interest. As a result, there may be additional neighboring pixels that could have been selected, but were not, based on the annotator's discretion and the principle of not including any erroneous or non-representative pixels. If there were multiple potential regions separated by the aforementioned artifacts, the largest and most representative area was chosen.

### **4.1.3 Semantic Porcine Dataset**

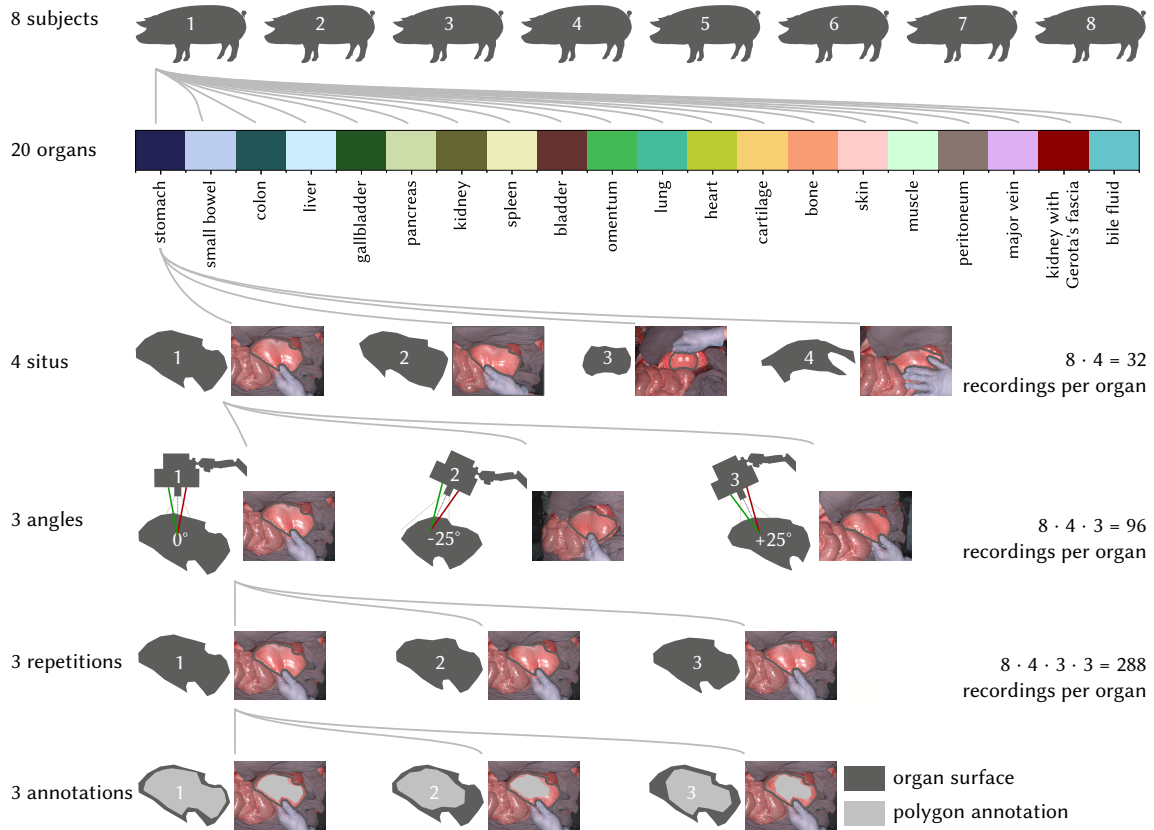
An overview of the semantic porcine dataset is given in Figure 4.4 and example images are shown in Figure 4.1. In total, this dataset contains 506 from 20 subjects (see Section 4.1.1 for a description of the data acquisition procedure). All images are fully semantically annotated with 19 classes, i.e., every pixel in the image is assigned a label. For each organ, between 32 and 405 images from 5 to 20 subjects were acquired.

The semantic annotation process was carried out by two different medical experts using vector annotation tools via the annotation platform SuperAnnotate (SuperAnnotate, Sunnyvale, USA) [3]. For consistent labeling, all annotations were revised by the same medical expert.

## 4 Materials and Methods

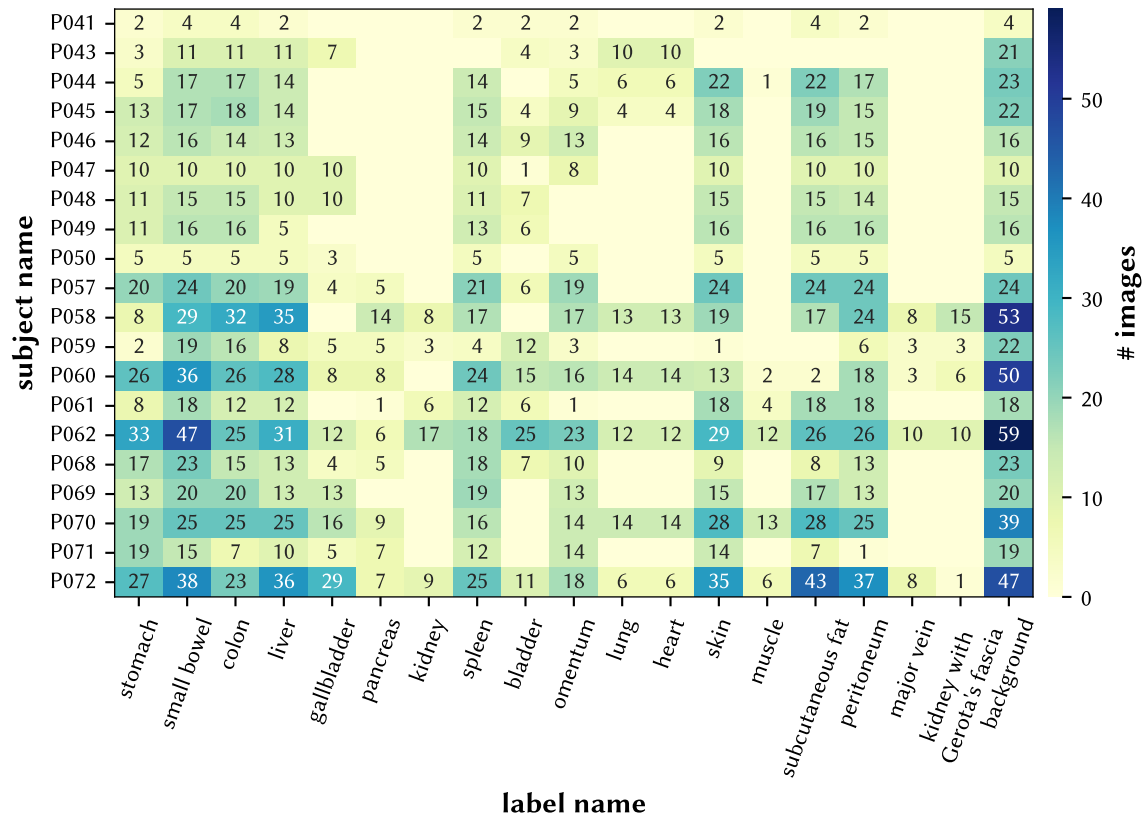


**Figure 4.2:** Overview of the tissue atlas dataset. 9057 images from 46 subjects have polygon annotations for 20 classes. Each cell denotes the number of images for the respective label and subject. The 11 italic highlighted subjects from *P086* to *P096* encompass the subset of 5756 standardized recordings. Figure 4.1 shows example images and annotations for this dataset.



**Figure 4.3:** Image acquisition protocol for the standardized recordings of the tissue atlas dataset (a subset of Figure 4.2). For each organ, there are recordings in 8 subjects (across 11 subjects in total) at 4 different situs (layout of the organs relative to each other), from 3 different angles with 3 repetitions. Every image was annotated by 3 medical experts. This figure was adapted from [214].

The 19 classes include two thoracic organs (heart, lung), eight abdominal organs (stomach, small bowel, colon, liver, gallbladder, pancreas, kidney, spleen), and one pelvic organ (bladder). For the kidney, images were captured both before and after the removal of Gerota's fascia, and these were labeled as *kidney with Gerota's fascia* and *kidney*, respectively. Additional annotations were made for subcutaneous fat, skin and muscle tissue, as well as omentum, peritoneum, and major veins. Any pixels associated with inorganic objects (e.g., cloth, compresses, foil, tubes, metallic objects, and gloves) were labeled as *background*. This label is found on every image, and the annotated areas cover, on average, 47 % (standard deviation (SD) 24 %) of an image. Pixels were also labeled as *ignore* if it was unclear to which organ they belonged or if they were part of an organic object other than the 18 organ classes. This label is found in 221 out of 506 images, and on average, the annotated areas cover 2 % (SD 3 %) of the pixels in these 221 images. These *ignore* pixels were later excluded from our analysis.



**Figure 4.4:** Overview of the semantic porcine dataset. 506 images from 20 subjects have fully semantical annotations for 19 classes. Each cell denotes the number of images for the respective label and subject. Figure 4.1 shows example images and annotations for this dataset.

Imbalances in the number of images per class occurred because some organs naturally appear more frequently in the field of view of other organs. For instance, the liver is always present in images of the gallbladder, but the gallbladder is not always visible in all liver images. The number of animals per organ varied due to differences in the surgical procedure performed. For example, opening the thorax is a highly invasive and challenging surgical procedure associated with significant mortality and extended surgery time. Therefore, it was only performed on 8 out of 20 subjects, making heart and lung HSI data unavailable for the remaining 12 subjects.

#### 4.1.4 Semantic Human Dataset

The semantic human dataset contains 777 images from 230 subjects (see Section 4.1.1 for a description of the data acquisition procedure) and is fully annotated with 16 classes.

An overview of the dataset is shown in Table 4.1<sup>2</sup> and example images are shown in Figure 4.1.

**Table 4.1:** Dataset statistics of the semantic human dataset. For each class, the number of subjects and the number of images which show this class are displayed. Figure 4.1 shows example images and annotations for this dataset.

| label name       | # subjects | # images |
|------------------|------------|----------|
| stomach          | 122        | 299      |
| small bowel      | 132        | 328      |
| colon            | 123        | 289      |
| liver            | 144        | 443      |
| gallbladder      | 38         | 124      |
| pancreas         | 41         | 93       |
| kidney           | 48         | 127      |
| spleen           | 19         | 45       |
| omentum          | 143        | 448      |
| lung             | 10         | 51       |
| skin             | 113        | 327      |
| muscle           | 37         | 77       |
| subcutaneous fat | 89         | 271      |
| peritoneum       | 51         | 133      |
| major vein       | 11         | 27       |
| background       | 230        | 775      |

The semantic annotation process was carried out by a group of medical experts using the Medical Imaging Interaction Toolkit (MITK) [220]. For consistent labeling, all annotations were revised by the same medical expert.

Similar to the semantic porcine dataset, the *ignore* label was used for unsure objects or classes not listed in Table 4.1 and ignored in our analysis. This label is found in 731 out of 777 images, and on average, the annotated areas cover 10 % (SD 8 %) of the pixels in these 731 images. This ratio is higher than for the porcine dataset because the human dataset is harder to annotate and some classes have been excluded either because they do not contain enough samples (*bladder*, *heart* and *kidney with Gerota's fascia*) or because the class does not exist in the semantic porcine dataset (e.g., *visceral fat*). We only included classes that occur in at least 6 subjects so that we could split them across five folds and an independent test set.

<sup>2</sup>We do not show a heatmap for this dataset because there are way too many subjects to show on a single page.

### 4.1.5 Dataset Features

Here, we compare our dataset with respect to common features, namely the distribution of organ sizes and a heatmap of common organ locations in the images. Both features are based on the annotated regions in the datasets.

#### Organ Sizes

Organs differ in their size, length and form and this is also reflected in the annotations. In terms of the number of pixels, there are substantial differences across organs as shown in Figure 4.5 for all datasets. Inside each dataset, classes like *stomach*, *colon* or *skin* usually occupy a large part of an image compared to smaller organs like *gallbladder*, *pancreas* or *major vein*.

Organs in the semantic porcine dataset contain approximately 201% more annotated pixels than the organs in the tissue atlas dataset on average. This is due to the regions that are omitted in the polygon annotations. Further, relative to other classes in the dataset, *small bowel* and *colon* contain more pixels than *stomach* in the semantic porcine dataset whereas this effect is reversed in the tissue atlas dataset.

In the semantic human dataset, some classes like *small bowel*, *colon* or *spleen* contain fewer pixels compared to the semantic porcine dataset. For other classes like *muscle* or *omentum* it is the other way around.

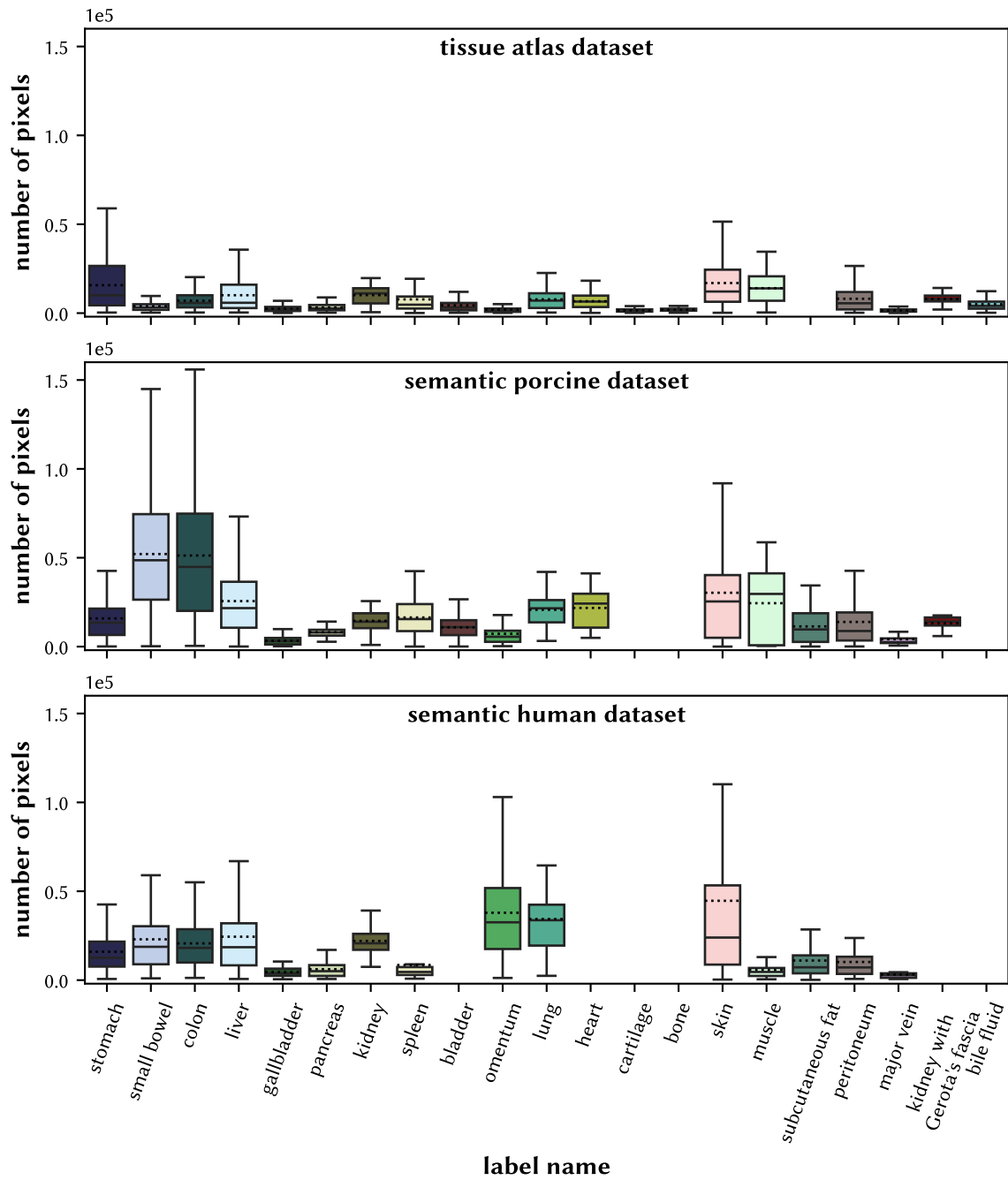
#### Location Maps

In theory, an organ could be positioned everywhere in an image. However, in practice, there are regions where an organ is more likely to be located than others depending on the preference during surgery and the constraints of the camera system (e.g., due to the necessary overlapping of the distance markers). This is also reflected in the annotations. Figure 4.6, Figure 4.7 and Figure 4.8 show location maps for each label for the tissue atlas, semantic porcine and semantic human dataset, respectively.

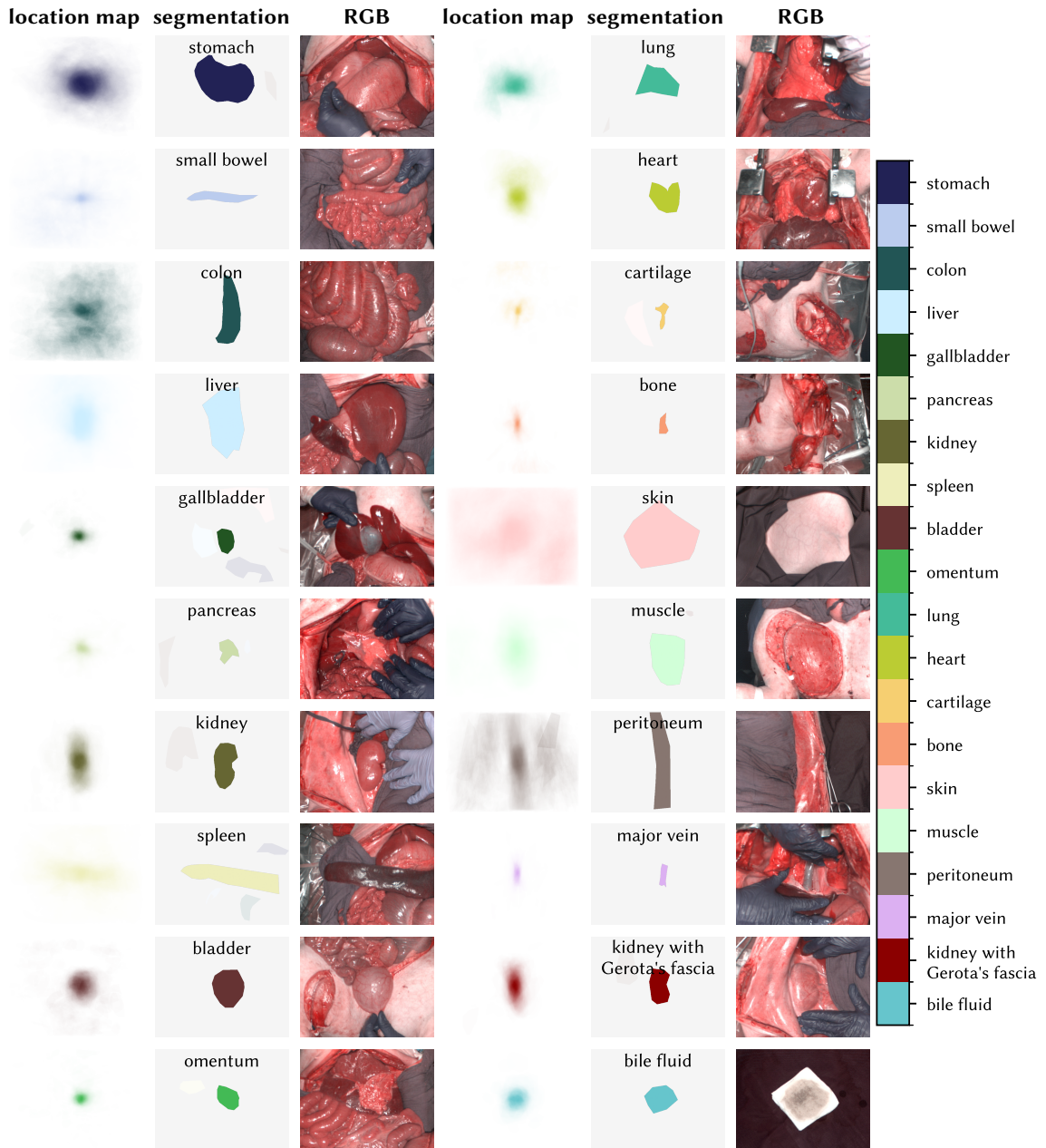
For the tissue atlas dataset, annotated organs are usually located in the center of the image revealing that the annotators mainly selected organs with a good view. This is also amplified by the standardized recordings (cf. Figure 4.3) where each organ was explicitly prepared to be imaged (e.g., by exposing the organ for a good view and placing the distance markers on the organ center).

In contrast, the location maps of the semantic datasets are more diverse and show a more realistic distribution of the organs in the images. There was no freedom for the annotators in the choice of which pixels were going to be annotated since every pixel in all images received a label. Some organs like *heart*, *kidney* or *major vein* have typical locations in the image whereas other organs like *small bowel*, *peritoneum* or *omentum* are more spread out.

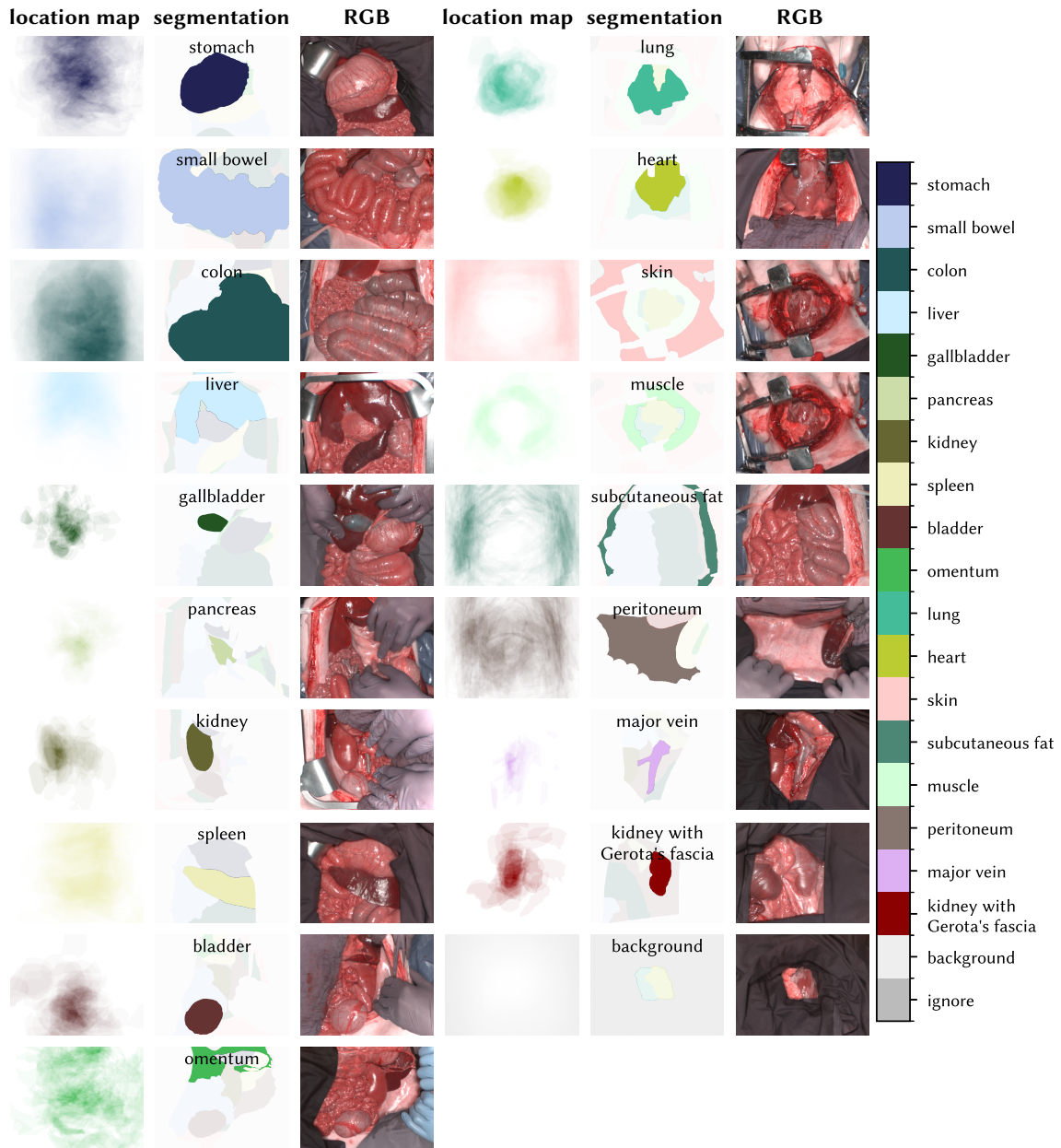




**Figure 4.5:** Distribution of the number of annotated pixels across all images for each organ and dataset. Each boxplot shows the interquartile range (IQR) with the median (solid line) and mean (dotted line). The whiskers extend up to 1.5 times of the IQR. Outliers and the *background* class are not shown for brevity.

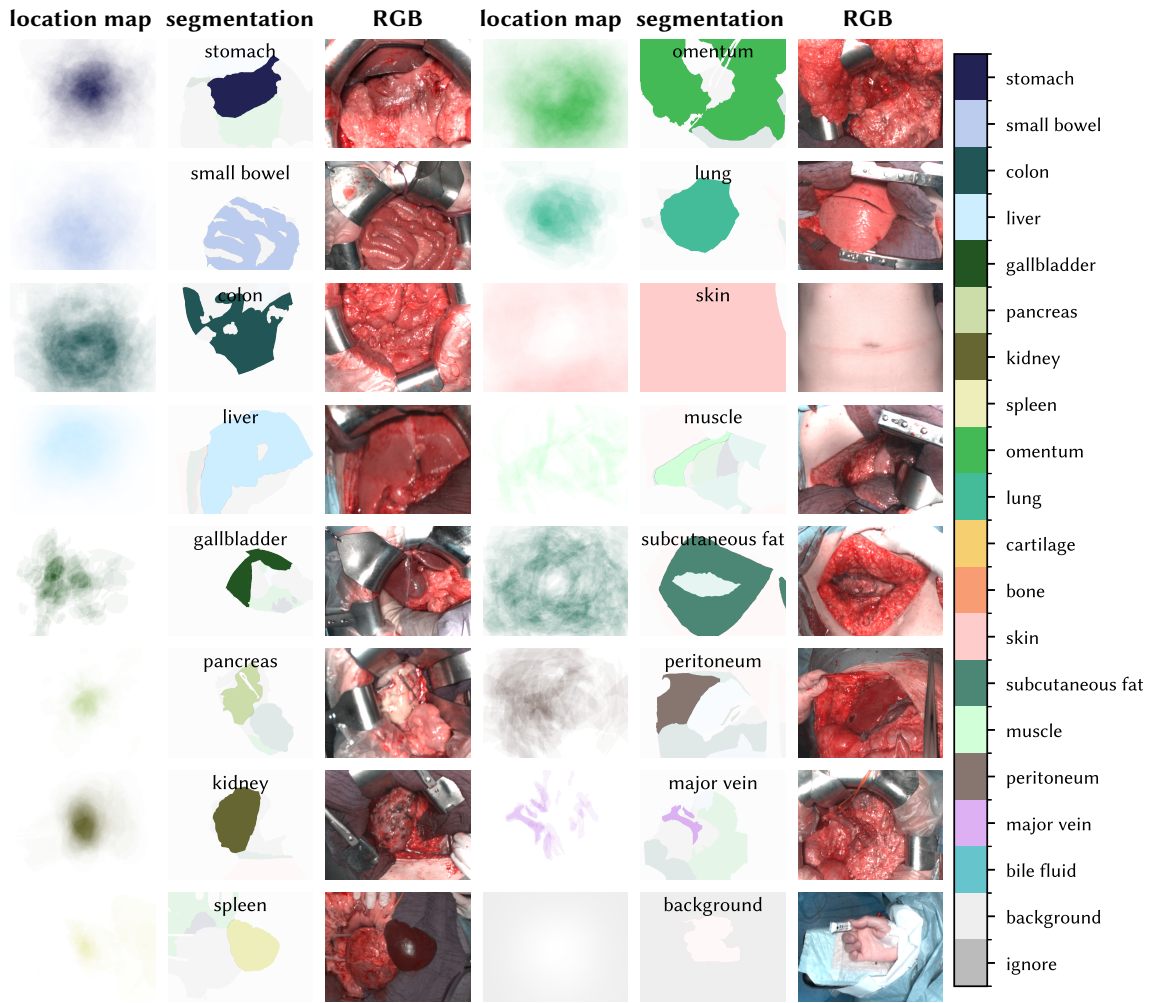


**Figure 4.6:** Location maps of the images in the tissue atlas dataset. The location maps show a heatmap of the positions for the annotated pixels for each label, i.e., it denotes where usually labels are located in an image. The exemplary RGB and segmentation images are representatives of their respective classes chosen to have a maximum overlap with the global location map.



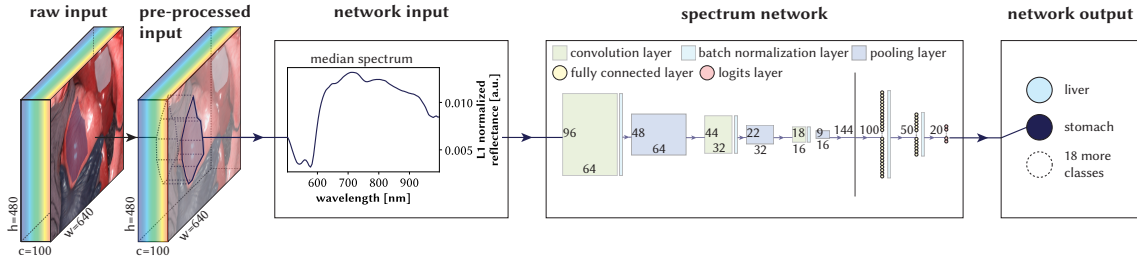
**Figure 4.7:** Location maps of the images in the semantic porcine dataset. The location maps show a heatmap of the positions for the annotated pixels for each label, i.e., it denotes where usually labels are located in an image. The exemplary RGB and segmentation images are representatives of their respective classes chosen to have a maximum overlap with the global location map.

The location maps for the semantic porcine and semantic human datasets are similar for classes like *colon*, *lung* or *gallbladder* but show different locations for classes like *spleen*,



**Figure 4.8:** Location maps of the images in the semantic human dataset. The location maps show a heatmap of the positions for the annotated pixels for each label, i.e., it denotes where usually labels are located in an image. The exemplary RGB and segmentation images are representatives of their respective classes chosen to have a maximum overlap with the global location map.

*major vein* or *muscle*. This could be due to different surgical procedures or different visibilities of an organ.



**Figure 4.9:** Deep learning pipeline for spectrum classification. Based on a pre-processed hyper-spectral image and corresponding polygon annotation, the median spectrum across all spectra of the annotation is computed (channel-wise) and fed into a spectrum classification network which assigns the annotated area to one of the 20 organ classes.

## 4.2 Spectral Organ Fingerprints

An overview of our deep learning pipeline for the classification of individual spectra is shown in Figure 4.9<sup>3</sup>. Using the corresponding polygon annotation, we compute the median spectrum by computing the median of all normalized reflectance values per channel. This 100-dimensional median spectrum is input to the convolutional spectrum network which predicts one of the 20 organ classes.

### Network and Training Setup

The deep learning model consists of three convolutional layers (with 64 filters in the first layer, 32 in the second, and 16 in the third), followed by two fully connected layers (with 100 neurons in the first and 50 in the second layer). All five layers have batch-normalized activations and a final linear layer was used to compute the class logits. Each convolutional layer applies a one-dimensional convolution to the spectral domain with a kernel size of 5 and is followed by an average pooling layer with a kernel size of 2. The two fully connected layers apply dropout with a probability of  $p$  to their activations. All non-linear layers utilize the exponential linear unit (ELU) [42] as the activation function (cf. Section 2.3.1).

We selected this architecture because it offers a straightforward yet efficient method for analyzing spectral data. The convolution operations focus on the local structure of the spectra and we employed a relatively small kernel size and stacked three layers to expand the receptive field while maintaining computational efficiency [218]. The two fully connected layers make a final decision based on the global context. The benefit of this approach is that it merges local and global information gathering while remaining computationally efficient, as the entire network only utilizes 34 300 trainable weights.

<sup>3</sup>This section is based on [215].

We trained 10 000 000 samples per epoch for 10 epochs with a batch size of  $N$ . For a standard computer vision task operating on full images (as also done in Section 4.4), such an epoch size would be way too large as it would take very long (in the scale of years) to train corresponding networks. However, for our spectrum classification task, this epoch size is not a problem because here a single sample consists of a 100-dimensional spectrum (e.g., compared to a model that operates on full HSI data cubes where a single sample has a dimension of  $480 \times 640 \times 100$  (height, width and number of channels)).

The softmax function was utilized to calculate the posterior probability for each class. We employed the Adam optimizer ( $\beta_1 = 0.9, \beta_2 = 0.999$ ) [116] with an exponential decay of the learning rate (decay rate of  $\gamma$  and initial learning rate of  $\eta$ ) and the multiclass cross-entropy loss function.

### Class Imbalances

The tissue atlas dataset is highly imbalanced in terms of number of samples (median spectra) per class (cf. Figure 4.2 and Figure 4.10). To account for this, we optionally weighted the loss function based on the number of training samples per class and sampled spectra for the batches either randomly or through oversampling to ensure each organ class had an equal chance of being sampled. Loss function weighting and oversampling have the similar effect of increasing the importance of underrepresented classes, either by showing corresponding samples more often during training (oversampling) or by increasing the update step for those classes during backpropagation (loss function weighting).

For both cases, we need a weight  $w_i$  for each class  $i$  which is based on the number of samples  $c_i$  in the training data for that class. In the case of the loss function, we define this weight with the help of the softmax function as

$$w_i^{\text{softmax}} = \frac{e^{\lambda \cdot c_i}}{\sum_{j=1}^n e^{\lambda \cdot c_j}}. \quad (4.2)$$

$n$  denotes the total number of classes and  $\lambda$  is the temperature parameter of the softmax function which we empirically set to  $\lambda = -2$ . Since  $\lambda < 0$ , we effectively make Equation 4.2 a *softmax* function (the most underrepresented class yields the highest weight).

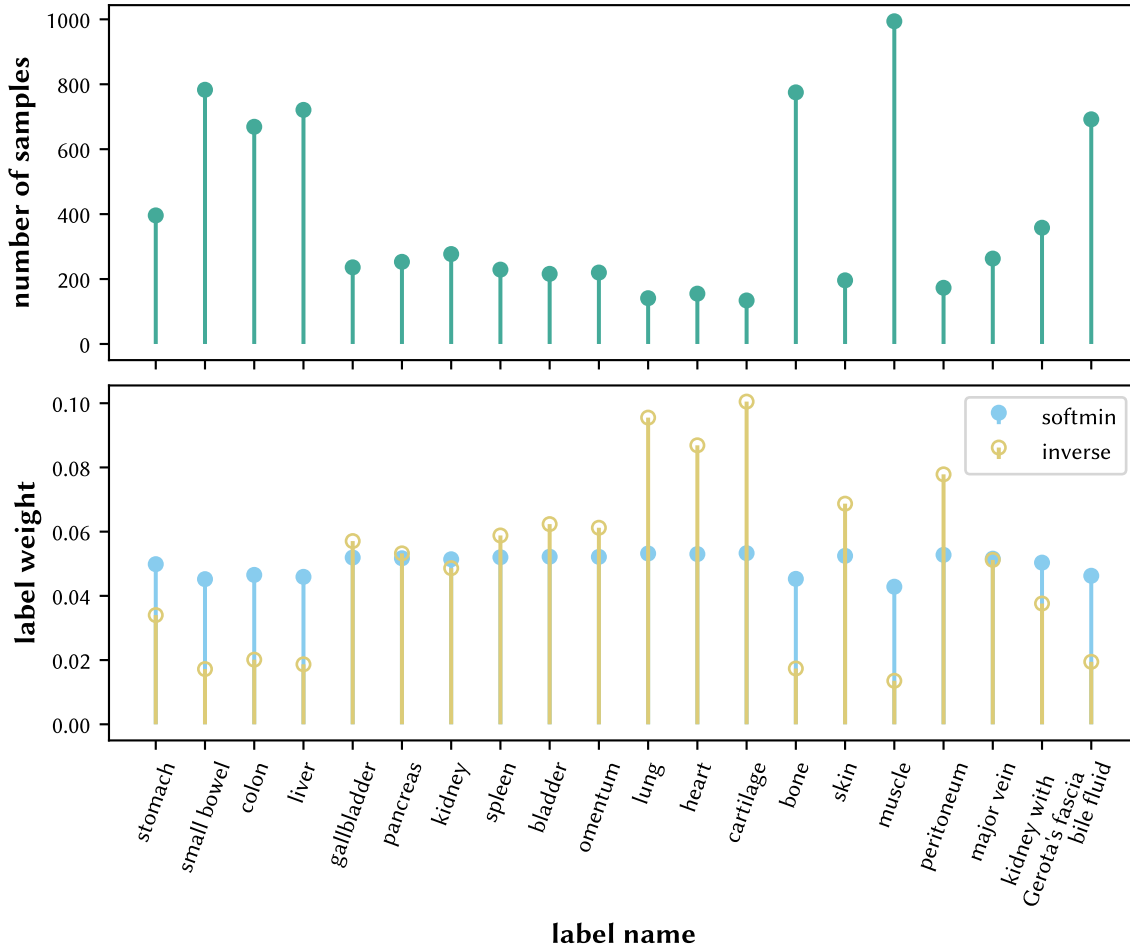
For oversampling, our goal is to see each class equally often during training. That is, a class that occurs with a ratio of  $c_i/n$  should be sampled with a ratio independent of  $c_i$ . To achieve this, we define the weight as

$$w_i^{\text{inverse}} = \frac{1}{\sum_{j=1}^n \frac{1}{c_j}} \cdot \frac{1}{c_i}. \quad (4.3)$$



The ratio  $1/c_i$  makes the sampling independent of the current class<sup>4</sup> and the prefactor converts the weights to a valid probability distribution. We are using this weight as a condition for our random sampling procedure, i.e., each class is sampled proportional to  $w_i^{\text{inverse}}$ .

Equation 4.2 leads to a softened version of the class probabilities compared to Equation 4.3. This allows us to control the effect of the loss weighting. Figure 4.10 compares both weighting schemes for the training data of the tissue atlas dataset.



**Figure 4.10:** Class imbalances in the tissue atlas dataset and class weighting schemes. The class weighting schemes (bottom) are based on the number of samples, i.e., median spectra, (top) and are defined in Equation 4.2 (softmin) and Equation 4.3 (inverse).

<sup>4</sup>Since  $\frac{c_i}{n} \cdot \frac{1}{\sum_{j=1}^n \frac{1}{c_j}} \cdot \frac{1}{c_i} = \frac{n}{\sum_{j=1}^n \frac{1}{c_j}}$  is a constant for all classes.

### Hyperparameter Search

We performed an extensive hyperparameter search on the validation split of the tissue atlas dataset (cf. Figure 5.1 in Section 5.1.1). The details are presented in Table 4.2. The spectrum network is defined by these optimal hyperparameters and all the results shown in Section 5.1 are based on them.

**Table 4.2:** Specification and results of the hyperparameter search for the spectrum classification network.

| hyperparameter                      | optimum      | search space                    |
|-------------------------------------|--------------|---------------------------------|
| dropout probability $p$             | 0.2          | $p \in \{0.1, 0.2\}$            |
| learning rate $\eta$                | 0.0001       | $\eta \in \{0.001, 0.0001\}$    |
| decay of the learning rate $\gamma$ | 0.9          | $\gamma \in \{0.75, 0.9, 1.0\}$ |
| batch size $N$                      | 20 000       | $N \in \{20\,000, 40\,000\}$    |
| weighted loss function              | included     | included/not included           |
| oversampling                        | not included | included/not included           |

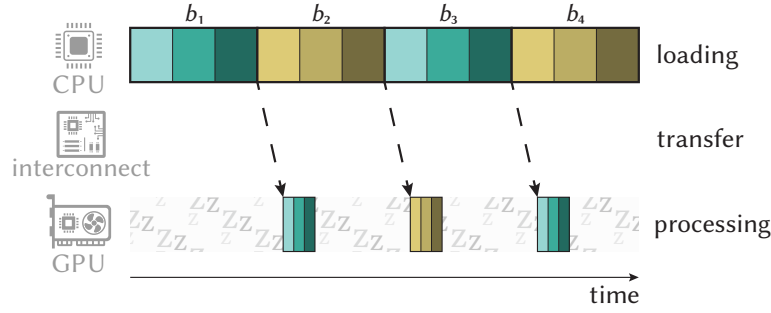
## 4.3 Efficient Training of Hyperspectral Segmentation Networks

We had to train hundreds of deep neural networks for this thesis. This includes not only the networks for the results shown but also all the networks trained during development. When we implemented our deep neural networks, our goal was not only a good segmentation performance but also a rapid training workflow. This was especially challenging for our HSI data since the images are large but the processing in our network (U-Net [189] with EfficientNet B5 encoder [219]) is relatively fast which leads to data loading bottlenecks as shown in Figure 4.11. In this section, we describe how we dealt with this problem<sup>5</sup>.

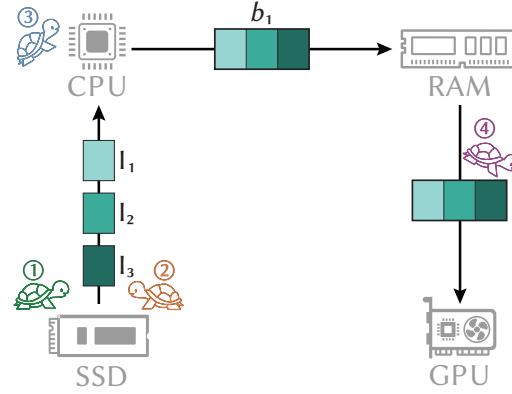
As a first step, we analyzed the data loading pipeline in detail (Figure 4.12) and identified four bottlenecks: (①) no usage of data compression and (②) an inefficient data type both concerning the data storage format and leading to large file sizes, (③) huge data processing effort on the CPU and (④) an inefficient memory transfer to the GPU involving unnecessary copies of the data. Subsequently, we developed four different counter-strategies to tackle these bottlenecks. The first two strategies (*blosc* and *fp16*) are based on the data storage format and the last two strategies (*gpu-aug* and *ring-buffer*) concern the data processing and memory transfer. Each strategy is explained in more detail below.

<sup>5</sup>This section is based on [198, 201].





**Figure 4.11:** Effect of inefficient data loading in high-throughput model training. In this example, the preparation of a new batch  $b_i$  on the central processing unit (CPU) takes relatively long compared to the processing of the batch  $b_i$  on the graphics processing unit (GPU) (e.g., when using large images but small models) leading to idle times of the GPU. This figure was adapted from [201].



**Figure 4.12:** Causes of inefficient data loading in high-throughput model training. Per default, images  $I_i$  are loaded from the solid-state drive (SSD) and processed to batches  $b_i$  on the central processing unit (CPU) (including data augmentation) before they are transferred to the graphics processing unit (GPU) via the random-access memory (RAM). Along this path, we discovered several bottlenecks which we tackled by optimizing the data storage format (① and ②), reducing the processing effort on the CPU (③) and tuning the memory transfer to the GPU (④). This figure was adapted from [201].

### Data Storage

The raw HSI data is stored in tensors of shape  $480 \times 640 \times 100$  (height, width and number of channels) with `float32` values leading to file sizes of approximately 117 MiB per image<sup>6</sup>. To reduce the demand on the SSD, we compressed the data cubes. Compression is a first and obvious step but not always beneficial. If the compression algorithm used is very slow in decompression, the benefit of reduced file size may be suspended by increased

<sup>6</sup>MiB denotes mebibytes and 1 MiB is composed of  $2^{20}$  bytes.

decompression times and high CPU load. Depending on the hardware, this may even be slower than loading uncompressed data.

To avoid this problem we are using the *blosc* meta compressor [21] which focuses on fast decompression times. *blosc* does not invent a new compression format but rather serves as a meta compressor where the user can select between different compression formats. Further, it employs some tricks to speed up the decompression process like chunking the data into blocks that fit into the cache of the CPU. As compression format, we are using the *zstd* [43] algorithm which is targeted at real-time compression scenarios and thus is designed for fast decompression. This reduces the file size to approximately 86 MiB per image.

To further reduce the loading effort on the SSD, we quantize the data by changing the data type from `float32` to `float16`. This is not a lossless step as full information retrieval is not possible anymore. However, we found that the information loss is negligible for our HSI data. This reduces the file size further to approximately 35 MiB per image.

### Data Processing

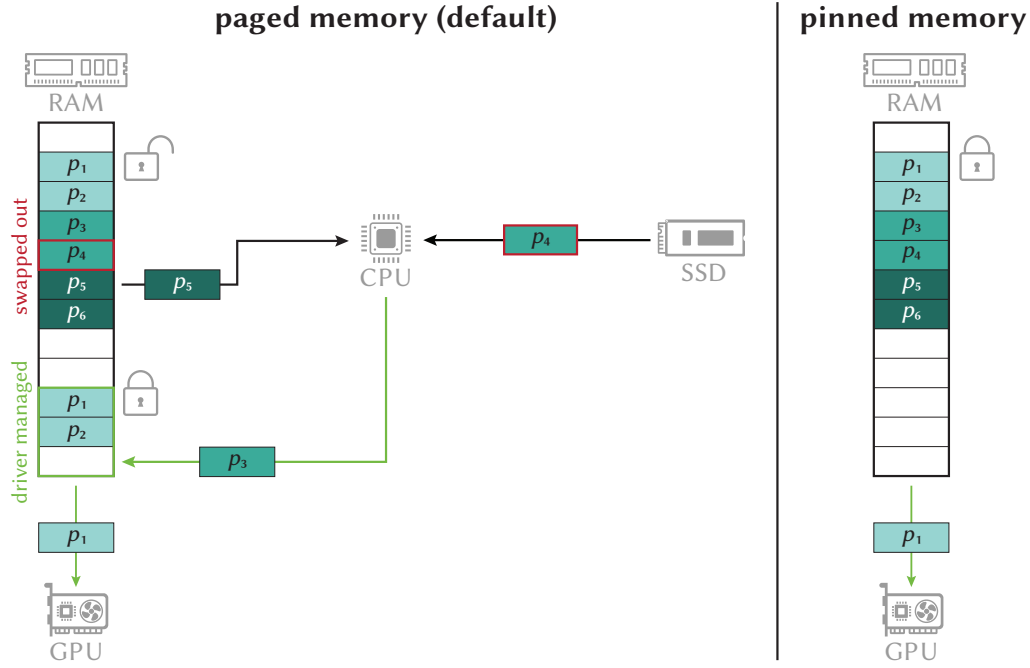
We employ data augmentations while training our networks (cf. Section 4.4) and this turns out to be one of the most time-consuming steps in our data loading pipeline. The reason behind this lies in the 100 spectral channels which necessitate repeating numerous operations for each channel (e.g., resampling in affine transformations). This is inefficient on the CPU because of the limiting parallelization capabilities compared to the GPU. Hence, we moved the data augmentation to the GPU using the Kornia library [185]. This has three advantages: (1) the data augmentation is now performed in a highly parallel manner which is more efficient compared to the CPU augmentations, (2) the GPU has work to do reducing its idle times and (3) we free resources on the CPU which can be used to load data.

### Memory Transfer

The final counter-strategy optimizes the memory transfer from the RAM to the GPU. Per default, this uses paged memory which requires additional memory copies and is hence inefficient. To avoid this, we use pinned memory which is non-swappable and can hence be used to directly transfer data from the RAM to the GPU without additional copies of the data (see Figure 4.13 for details).

Allocation of pinned memory requires device synchronization between the host (application and graphics driver running on the CPU) and the GPU [47]. If this happens regularly during training, it can slow down the training because the host and the GPU have to wait for each other on all synchronization points destroying the asynchronous nature of the GPU [224].

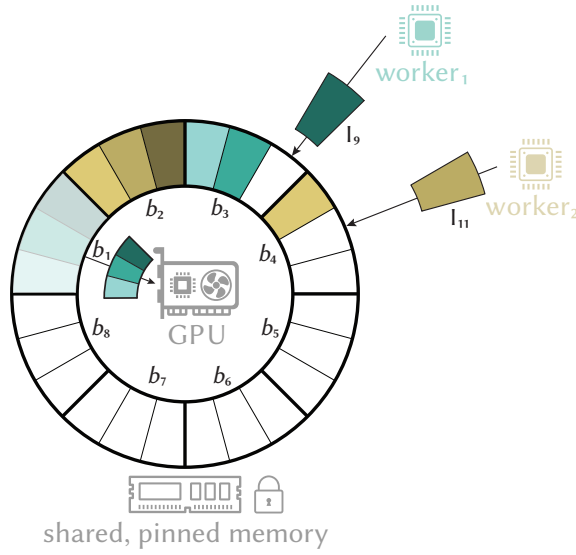
To avoid this problem, we propose to use a shared, fixed and pinned memory ring buffer (Figure 4.14). This buffer is initialized at the beginning of the training and hence requires



**Figure 4.13:** Concept of pinned memory in comparison to paged memory. In the example, batch data consisting of 6 memory pages  $p_i$  have previously been loaded into the random-access memory (RAM) and should now be transferred to the graphics processing unit (GPU). Per default, data has to pass the central processing unit (CPU) on transfers from the RAM to the GPU because some memory pages may have been swapped out by the operating system and need to be loaded back in again from the solid-state drive (SSD) by the graphics driver. In the example on the left side, memory page  $p_1$  is currently transferred to the GPU from the driver-managed memory region (this region may also be pinned but is usually much smaller so that not all the data can be transferred at once),  $p_3$  is moved to the memory region which is managed by the graphics driver,  $p_4$  is loaded from disk because it was swapped out by the operating system previously and  $p_5$  is processed next by the CPU. If using pinned memory (right side), memory pages  $p_i$  are non-swappable and data can hence be directly moved from the RAM to the GPU. This figure was adapted from [201].

synchronization only once during training. It resides in shared memory so that each worker can directly load batches into the buffer and it is allocated as pinned memory. The workers re-use existing memory locations when they are free again after successful transfer to the GPU.

It is worth noting that pinned memory allocations should be treated with care since every pinned memory region cannot be used by the operating system anymore (it is exclusively reserved for the task inside the allocating application). Too many small pinned memory allocations can further lead to memory fragmentations if no large enough continuous



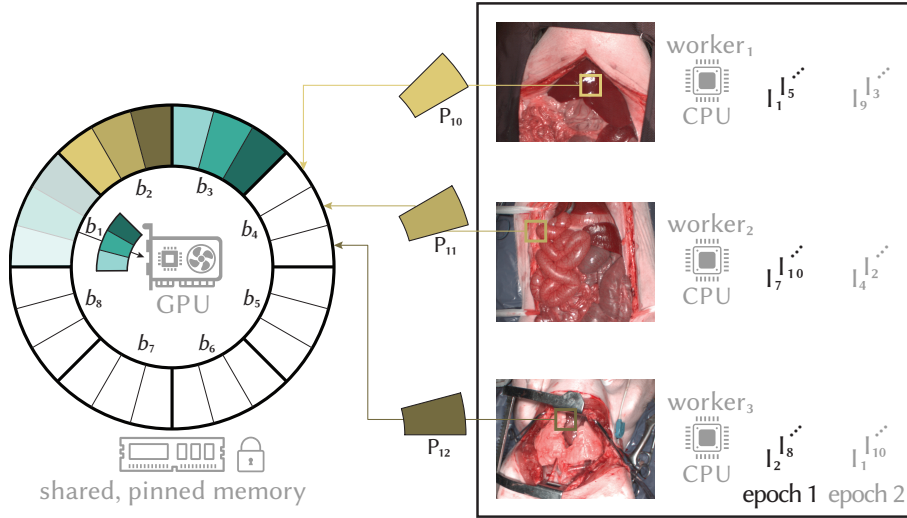
**Figure 4.14:** Concept of the shared, fixed and pinned memory ring buffer. While using a fixed (initialized during training start), shared (accessible by all workers) and pinned memory location (cf. Figure 4.13), unnecessary memory transfers can be avoided. Every batch  $b_i$  consists of 3 images  $I_i$  which are stored in multiple memory pages  $p_i$  (the memory pages are not shown in the figure). In the example, batch  $b_1$  is currently transferred to the graphics processing unit (GPU), worker<sub>1</sub> copies image  $I_9$  and worker<sub>2</sub> copies image  $I_{11}$  to the buffer. Memory locations of successfully transferred batches (like  $b_1$  in the example) are going to be re-used by the workers. This figure was adapted from [201].

memory region remains for the operating system. Hence, it is advisable to use only a minimal number of shared, fixed and pinned memory ring buffers (e.g., only for the image data and the corresponding labels) and ensure that enough system memory is available after the initialization phase. [86, 48]

### Collaborative Batch Filling

The concept of the shared, fixed and pinned memory ring buffer is not only applicable to the image model but also to smaller spatial granularities like patches or superpixels. However, in these cases, it is also advantageous if the workers operate collaboratively on a batch, i.e., if every worker contributes one part of the batch (e.g., via a set of patches or superpixels). This concept is visualized in Figure 4.15.

With the collaborative setting, we can increase randomness across images in one batch since each batch now contains parts from multiple images (controlled by the number of used workers). This brings the batch distribution closer to the real data distribution which is usually beneficial for training neural networks. Additionally, the memory footprint is constant in this setting and only depends on the number of used workers. In contrast, if workers fill batches independently from each other, the number of images that contribute



**Figure 4.15:** Application of the shared, fixed and pinned memory ring buffer for smaller spatial granularities. All workers contribute equally to every batch  $b_i$  to ensure batch parts (here patches) arise from a large variety of different images. In this example, the three workers prepare the patches  $P_{10}$ ,  $P_{11}$  and  $P_{12}$  on the central processing unit (CPU) filling up the batch  $b_4$ . The batch  $b_1$  is currently transferred to the graphics processing unit (GPU). The images  $I_i$  are assigned disjunct to the workers and are re-shuffled after every epoch for increased image variation. This loading scheme is used for the pixel, superpixel and patch granularities. This figure was adapted from [198].

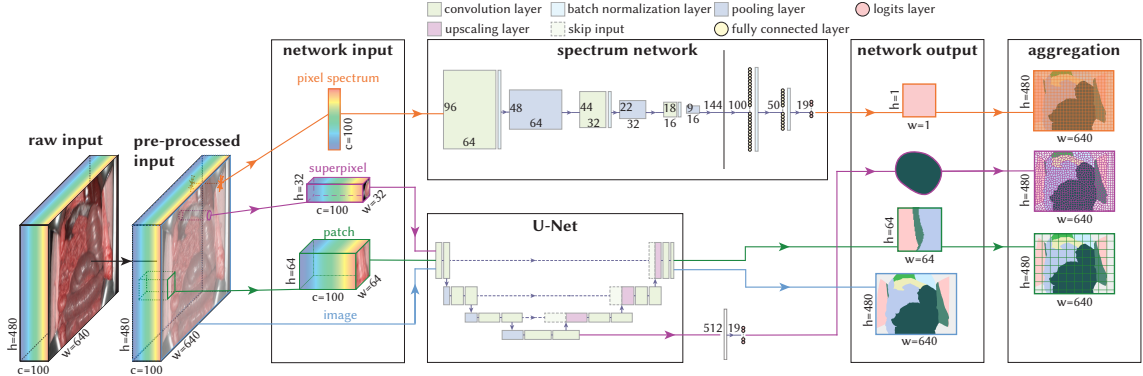
to a batch is limited in practice because holding multiple images per worker in memory would significantly increase the memory footprint.

## 4.4 Surgical Scene Segmentation of Hyperspectral Images

An overview of our deep learning pipeline for the segmentation of HSI data using different spatial granularities is shown in Figure 4.16<sup>7</sup>. In general, our decisions regarding model architectures and training setup were largely influenced by our comparative study of different modalities and spatial granularities. We strived to maintain consistent model and training parameters across different spatial granularities and modalities where feasible (for instance, identical hyperparameters, data splits or a comparable number of pixels per epoch). We deliberately refrained from individual parameter tuning for each model

<sup>7</sup>This section is based on [198].

and stuck to the default settings whenever possible to guarantee a fair comparison and minimize computational expenses.



**Figure 4.16:** Deep learning pipeline for segmentation of hyperspectral imaging (HSI) data for different spatial granularities. For **pixel-based classification**, every spectrum of the pre-processed input is fed individually into a spectrum classification network to predict one of the 19 classes. For **superpixel-based classification**, superpixels are computed based on the RGB image, a minimum enclosing bounding box is fit while replacing pixels outside the superpixel with zeros and then the box is reshaped to a fixed superpixel cube. Then, each superpixel cube is fed into the encoder of the spectrum segmentation network followed by a classification head. For **patch-based segmentation**, the pre-processed input is split into patches of fixed size and then each patch is fed into the spectrum segmentation network. The individual predictions of the pixel, superpixel and patch-based networks are aggregated yielding a prediction map for the complete image. For **image-based segmentation**, the pre-processed input is fed into the segmentation network yielding a prediction for each pixel. This figure was adapted from [198].

### Input Modalities

We wanted to explore the potential advantages of using HSI data over RGB and TPI data for fully automated organ segmentation via neural networks. For simplicity, we will refer to these different types of input data as *input modalities*, even though they were all captured with the same camera and are based on the same data. This reflects the possibility that future applications leveraging semantic scene segmentation could be based on RGB images from a standard camera, preprocessed HSI images from an HSI camera provider, or raw HSI spectra. We trained neural networks separately on all three input modalities for all examined spatial granularity levels. The camera system provided RGB data reconstructed from the HSI data. To evaluate the organ segmentation performance on processed HSI data, we stacked the corresponding StO<sub>2</sub>, NPI, TWI, and THI images, resulting in a  $480 \times 640 \times 4$  (height, width and number of channels) TPI cube that served as the model input.

**Pixel-Based Segmentation**

The smallest spatial granularity of the input data involves using individual pixel spectra, which results in input feature vectors of length  $c = 100$  for HSI input data,  $c = 4$  for TPI input data, and  $c = 3$  for RGB input data. For HSI input data, we are using the same spectrum classification network as for our tissue atlas dataset classification task (cf. Section 4.2). For TPI and RGB input data, it is not possible to have convolutional operations across channels due to the small channel size. Instead, the network comprises three fully connected layers with 200 neurons in the first, 100 neurons in the second, and 50 neurons in the third layer. The ELU [42] is used as an activation function (cf. Section 2.3.1) and batch normalization is applied to all outputs from all layers except pooling layers. The class logits are computed by a final linear layer and the maximum value determines the predicted class. The cross-entropy (CE) loss function is used for model optimization during training.

The architecture is designed to aggregate local information from neighboring spectral bands and global information across the entire spectrum while maintaining a small network size of only 34 275 weights for HSI, 27 819 weights for TPI, and 27 619 weights for RGB input data, thereby ensuring computational efficiency.

To obtain a segmentation map for an image, we predict a class label for each pixel in the image and then project the resulting labels back to the original image locations.

**Superpixel-Based Segmentation**

Superpixels are low spatial granularity regions that conform to local boundaries, enclosing pixels with similar characteristics. Similar to pixel-wise organ segmentation, the unsupervised clustering of superpixels transforms the organ segmentation task into a superpixel-wise organ classification task. This is based on the assumption that all pixels within a superpixel belong to the same organ class, as superpixels are expected to lie within the local boundaries of an organ. Superpixels are created using the SLIC algorithm on the reconstructed RGB data [1]. Before clustering, the image is smoothed with a Gaussian kernel of width 3, and then 1000 segments are computed in 10 iterations while adaptively changing the per-superpixel compactness parameter (SLICO mode). A minimum enclosing bounding box is computed for each superpixel and areas outside the superpixel are replaced with zeros. To standardize the input shape, superpixels are resized via bilinear interpolation to the shape  $32 \times 32 \times c$ . The number of channels  $c$  depends on the input modality ( $c = 100$  for HSI,  $c = 4$  for TPI and  $c = 3$  for RGB input data).

The resized superpixel cubes are fed into an EfficientNet B5 encoder [219] pretrained on the ImageNet dataset [55] using the Segmentation Models PyTorch library of Yakubovskiy [242]. This encoder was chosen for its good performance, low number of parameters, low memory footprint, and fast computation times. The encoder network's output is passed to a classification head consisting of a fully connected layer with 19 neurons for

calculating the class logits. Thus, the superpixel network shares the same architecture (mainly the encoder) as the segmentation networks for the image and patch-based models while employing only minor modifications.

Not all pixels within a superpixel may belong to the same organ class, possibly due to inconsistencies at the organ border. To address this, we introduced fuzzy labels, assigning a label vector with the length corresponding to the 19 classes of our semantic segmentation task. The fuzzy label vector records the relative frequency of each class label based on the labels from the pixels inside the superpixel. The Kullback-Leibler divergence [123] between fuzzy labels and the softmax output is used as a loss function during training.

During inference, the maximum value of the logits vector determines the predicted class label and this label is assigned to every pixel position of the superpixel. Predictions of all pixels from all superpixels are combined to produce a segmentation map for an image.

### Patch-Based Segmentation

Patches are low spatial granularity regions extracted from images based on a fixed shape. They are typically easier to generate and more compatible with neural networks than superpixels, mainly due to their rectangular shape aligning with the rectangular kernel shapes of convolutional neural networks. To capture varying degrees of granularity, we extract patches of two different shapes:  $32 \times 32 \times c$  and  $64 \times 64 \times c$ . These sizes act as intermediate steps between the superpixel and the image model in terms of spatial granularity (cf. Table 4.3). We use patch sizes that are powers of two to seamlessly integrate them with encoder architectures that halve the input shape multiple times. The number of patches generated per image equals the number of patches that could be generated via a grid-based tiling, e.g.,  $\frac{480}{32} \cdot \frac{640}{32} = 15 \cdot 20 = 300$  patches per image for the patch\_32 granularity.

The patches are fed into a U-Net [189] with an EfficientNet B5 encoder [219] pretrained on the ImageNet [55] dataset (similar to the superpixel network). During training, Dice loss [151] and CE loss are calculated based on all valid pixels<sup>8</sup> in the batch and equally weighted to compute the final loss. While each misclassified pixel contributes equally to the CE loss, misclassified pixels belonging to an organ class with a smaller image area contribute more to the Dice loss than misclassified pixels belonging to a bigger-sized class (e.g., *background*). The weighted sum of both loss terms allows the network to leverage the respective advantages of both loss functions.

During inference, images are divided into a grid of non-overlapping patches of the corresponding patch size. If an image dimension is not an integer multiple of the patch dimension (e.g., for the patch\_64 granularity:  $480/64 = 7.5$ ), the missing image regions are zero-padded. For each patch, the network produces a segmentation map for the patch. The segmentation maps of all patches of one image are combined to create an image

---

<sup>8</sup>Pixels from the *ignore* class are not considered valid and are excluded from the loss calculation.



segmentation map. Segmentations belonging to previously zero-padded image regions are removed.

**Table 4.3:** Epoch and batch sizes for the spatial granularity models. The names patch\_64 and patch\_32 refer to models with the input shapes  $32 \times 32 \times c$  and  $64 \times 64 \times c$ , respectively ( $c$  number of channels). # pixels refers to the number of pixels of a single input sample for a model.

| spatial granularity | # pixels      | epoch size  | batch size |
|---------------------|---------------|-------------|------------|
| image               | 307 200       | 500         | 5          |
| patch_64            | 4096          | 37 632      | 336        |
| patch_32            | 1024          | 150 528     | 1176       |
| superpixel          | $\approx 300$ | 500 760     | 1560       |
| pixel               | 1             | 153 608 400 | 118 800    |

### Image-Based Segmentation

Images provide the highest level of spatial granularity and are used directly without any modifications to the image dimensions. This means the input tensors have a shape of  $480 \times 640 \times c$ . Similar to the patch-based segmentation, the images are fed into an EfficientNet B5 [219] powered U-Net [189] pretrained on the ImageNet [55] dataset. Both Dice [151] and CE loss are equally weighted to compute the loss function.

### Training Setup

To enhance the size and variety of the training data, thereby improving convergence, generalization, and robustness on out-of-distribution samples, data augmentation is frequently used in computer vision [28]. Across all spatial granularities, we augment the training data with the help of the Kornia library [185] on the image level, i.e., before extracting smaller spatial granularities like pixels, superpixels, or patches. Images are shifted (shift factor limit: 0.0625), scaled (scaling factor limit: 0.1), rotated (rotation angle limit:  $\pm 45^\circ$ ), and flipped (horizontally and vertically). We set the probability of applying an augmentation to  $p = 0.5$  to minimize the computational costs associated with extensive data augmentations.

All models utilize the Adam [116] optimization algorithm ( $\beta_1 = 0.9, \beta_2 = 0.999$ ) and an exponential learning rate scheme (initial learning rate:  $\eta = 0.001$ , learning rate decay  $\gamma = 0.99$ ). Training is conducted over 100 epochs with stochastic weight averaging [103] applied. To ensure a consistent training procedure across models, the training budget should be the same for all models. Therefore, one epoch is defined as processing 500 training images for image-based segmentation. For pixel-, superpixel-, and patch-based segmentation, the total number of extracted pixels for each method is approximately equal to the total number of pixels in 500 images. This results in a certain number of samples per epoch for each granularity (cf. epoch size column of Table 4.3). The approximation

is necessary because the epoch size must be divisible by the batch size to allow every worker in the data loader to contribute equally to each batch (cf. Section 4.3).

Optimal batch size recommendations vary (e.g., [208, 112]). While smaller batch sizes can potentially accelerate the learning process [156], larger batches better represent the real population, potentially leading to more stable gradients and improved batch statistics [101]. As a compromise, we maximized the batch size while opting for a large number of epochs to counteract a potentially slower learning process. In practice, the batch size is limited by the available GPU memory and depends on the network and the dimensions of the input. We determined the maximum batch size per model, resulting in larger batch sizes for smaller input spatial granularities. Table 4.3 provides an overview of the resulting epoch and batch sizes.

During training, every model from the respective training fold was evaluated after each epoch on the corresponding validation set. The dice similarity coefficient (DSC) was calculated while considering the hierarchical structure of the data (cf. Figure 5.13). The final validation score was obtained by averaging the DSC values of the three validation pigs, and this score was also used to determine the best model across all epochs per fold.

To mitigate overfitting, dropout regularization is utilized for the fully connected layers in the pixel models and in the superpixel classification head, with the dropout probability set to  $p = 0.1$ .

### Hardware and Variability

Training of neural networks involves various sources of variation, some of which are easier to control than others (e.g., seeding vs. hardware influences) [174]. In our comparative study, where we aim to fairly compare different spatial granularities and modalities, it is crucial to minimize these sources of variation.

While achieving perfectly reproducible results often comes at the expense of extended training times (e.g., by resorting to deterministic operations or a single homogeneous hardware infrastructure; which can be inefficient) [174], we implemented several measures to reduce variation: we controlled the weight initialization of the networks, the initialization of the workers responsible for data loading (which also impacts data augmentation due to the seeds of the workers) and the sequence in which training samples are presented to the network for the modalities (e.g., corresponding spatial models for different modalities receive patches from the same spatial locations and in the same order). To achieve this, we set a seed for the network initialization and the data loaders and standardized the number of workers on each data loader to 12 across all experiments. However, due to the need for efficient execution of numerous training runs for this study, we did not enforce deterministic operations and utilized our in-house cluster infrastructure, which comprises an inhomogeneous hardware infrastructure with various GPUs (e.g., NVIDIA<sup>®</sup> GeForce RTX<sup>™</sup> 2080 Ti or NVIDIA<sup>®</sup> DGX<sup>™</sup> A100).

Additionally, in order to understand the variation introduced by the randomness of the training process, we retrained our image HSI model five times with different seeds. We compared the results of those networks and analyzed the coherence of the predictions on the pixel level, i.e., where in the image are pixels that yield the same prediction across all five seed networks and where are pixels with different predictions. That is, given the network outputs (class labels)  $l_{x,y}^s$  at position  $(x, y)$  for seed network  $s \in \{1, 2, 3, 4, 5\}$ , we compute the prediction coherence map (PCM) for an image as

$$\text{PCM}_{x,y} = \begin{cases} 0 & \text{if } l_{x,y}^1 = l_{x,y}^2 = l_{x,y}^3 = l_{x,y}^4 = l_{x,y}^5 \\ 1 & \text{else} \end{cases} \quad (4.4)$$

PCMs are used in Section 5.3 in the discussion about network variability and the ensembling effect. Examples are shown in Figure 5.24.

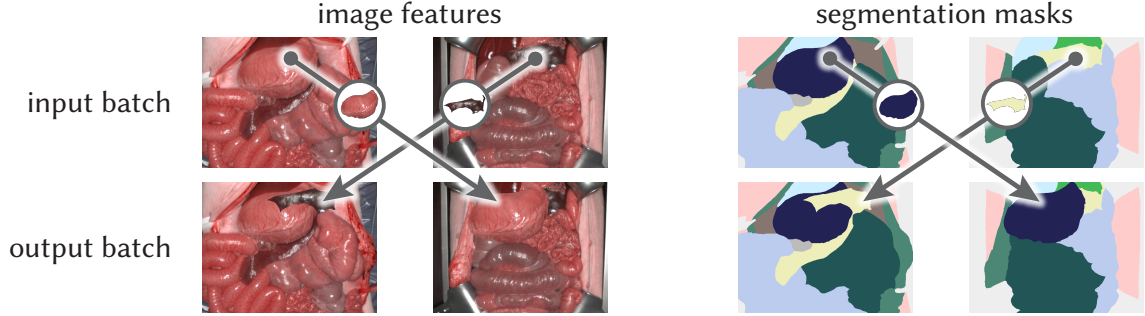
## 4.5 Domain Shifts in Surgical Hyperspectral Imaging

In this section, we describe our approach to tackle geometric domain shifts in the context of surgical scene segmentation based on the segmentation networks described in Section 4.4<sup>9</sup>. Our approach operates on the premise that geometric domain shifts can potentially be mitigated through application-specific data augmentation techniques. Hence, instead of modifying the network architecture of our segmentation methods, we introduce a new surgical-inspired data augmentation technique.

The concept of our data augmentation technique is based on the idea of transplanting organs from one image to another thus forcing the network to detect classes independent of their surroundings. This augmentation, which we call organ transplantation, is illustrated in Figure 4.17. Given a batch of  $n$  images, we randomly select  $N$  images from the batch (depending on the probability parameter  $p$  of applying the augmentation) where selected classes should be transplanted from. For each of the  $N$  images (the donors), we randomly select a class and place all pixels that belong to this class into another image (the acceptor). We always copy the spectra and the corresponding segmentation mask. This way, a class is placed in an unusual context while keeping texture and shape consistent. A detailed description of the augmentation technique is given in Algorithm 4.1.

Our augmentation has been inspired by the image-mixing augmentation CutPas. Originally, CutPas was proposed for object detection [61] and it has since been adapted for instance segmentation [73] and for generating low-cost datasets through image synthesis from a small number of real-world images in surgical instrument segmentation [231].

<sup>9</sup>This section is based on [202].



**Figure 4.17:** Concept of the organ transplantation augmentation. Given a batch of  $n$  images where  $N$  images are considered for transplantation (here  $n = N = 2$ , i.e., a probability of applying the augmentation of  $p = 1$ ), image features and corresponding segmentations of random classes are transplanted between images in the batch. Here, the stomach is transplanted from the first to the second and the spleen is transplanted from the second to the first image. This figure was adapted from [202].

**Algorithm 4.1:** Detailed description of the organ transplantation augmentation. The function  $\text{randlabel}(x)$  returns a random value from the unique values of the input  $x$ . Copying of the input is necessary since an image may be an acceptor and donor at the same time.

---

**Input:** Input batch consisting of images  $I_1, \dots, I_n$  with corresponding segmentation masks  $L_1, \dots, L_n$ ; Probability  $p \in [0; 1]$  of applying the augmentation to an image.

**Output:** Output batch of augmented images  $I'_1, \dots, I'_n$  with corresponding segmentation masks  $L'_1, \dots, L'_n$ .

- 1:  $I'_i \leftarrow I_i, L'_i \leftarrow L_i \quad \forall i = 1, \dots, n$  ▷ Copy input.
  - 2:  $N \leftarrow \lfloor p \cdot n \rfloor$  ▷ Number of donor images.
  - 3: **for**  $d = 1, \dots, N$  **do**
  - 4:      $a \leftarrow d - 1$  ▷ Make image left of the donor the acceptor.
  - 5:      $l \leftarrow \text{randlabel}(L_d)$  ▷ Select a random class label from the donor.
  - 6:      $\mathcal{P} \leftarrow \{\mathbf{p} \mid L_d(\mathbf{p}) = l\}$  ▷ Set of pixel coordinates for the donor label.
  - 7:      $I'_a(\mathcal{P}) \leftarrow I_d(\mathcal{P})$  ▷ Transfer image features from the donor to the acceptor.
  - 8:      $L'_a(\mathcal{P}) \leftarrow L_d(\mathcal{P})$  ▷ Transfer label information from the donor to the acceptor.
  - 9: **end for**
-

## EXPERIMENTS AND RESULTS

---

In this chapter, we will show our spectral analysis and the results from our classification and segmentation networks. The general structure and order of the experiments follow the process introduced in Figure 1.2. We start with the analysis on the spectra level, present our open-data concept and explain how we effectively train our networks before we move from the spectra to the image level. Further, we show the effect of different domain shifts and present a solution for contextual shifts.

In Section 5.1, we present the results of our spectral analysis (RQ1). Section 5.2 shows the effectiveness of our data loading performance improvements (RQ2). In Section 5.3, we present how we utilize the information contained in the spectra in our segmentation networks (RQ3). The systematic analysis of domain shifts (RQ4) is the topic of Section 5.4 and concerns the impact of subject, species and context domains. This includes our data augmentation method to maintain performance on geometrical OOD data.

Each section includes details about our experimental setup, i.e., how we designed and evaluated our experiments including our data splits for training, validation and testing (using the datasets described in Section 4.1), the employed metrics as well as our aggregation scheme.

All sections include a presentation and interpretation of the results. Detailed experiment-specific discussions and a meta-level discussion with all the findings of this thesis can be found in Chapter 6.

### 5.1 Spectral Organ Fingerprints

This section covers our fundamental analysis of utilizing the spectral information contained in our hyperspectral datasets described in Section 4.1<sup>1</sup>. Our goal of this study was to find unique spectral fingerprints for different organ classes and to discriminate the spectra via a machine learning model. After some details about our experimental

---

<sup>1</sup>This section is based on [215].

design in Section 5.1.1, this includes the results for our classification task based on median spectra (Section 5.1.2) as well as our open data efforts (Section 5.1.3).

### 5.1.1 Experimental Setup

The median spectra classification task is trained and evaluated on the tissue atlas dataset (see Section 4.1.2). This section outlines how we split the data for training, validation and testing as well as the employed evaluation metrics.

#### Splits

Prior to network training, we split the dataset into a training and a hold-out test set with 3765 images from 38 subjects and 5292 images from 8 subjects, respectively. These 8 test subjects were randomly selected from the subset of standardized recordings with the only criterion that every class should be represented in the training and test split by at least one subject from the standardized recordings. With more than 8 subjects, this criterion could not be fulfilled anymore. We used the test subjects only after we finalized the network architecture and tuned all hyperparameters. The training set is further split into 38 folds for leave-one-subject-out cross-validation. All splits are visualized in Figure 5.1.

#### Metrics

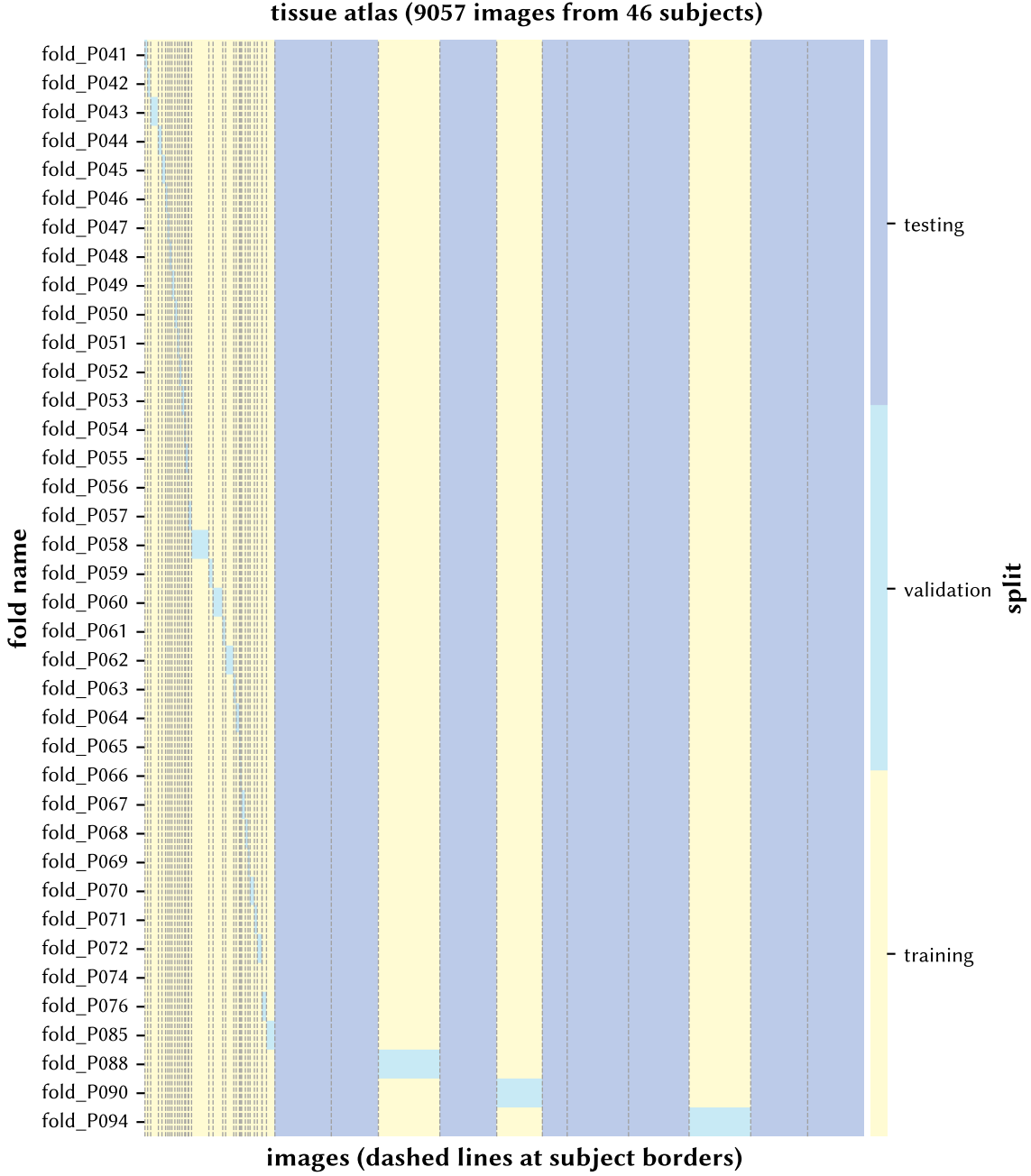
We used the micro-averaged accuracy to assess the performance of our machine learning model. This means that all samples from all classes contribute equally to the metric score. To account for the imbalance in the number of samples per class of the tissue atlas dataset, we additionally computed the macro-averaged sensitivity (recall), specificity and F1 score. In these cases, scores were computed independently for each class and then the class-level scores were averaged. [184, 141, 183]

In all cases, we respect the hierarchical structure of the data and compute metrics first for each subject before averaging across subjects (cf. Figure 5.13). The validation accuracy is computed for each fold, i.e., for one subject, and then the results are aggregated across all folds by averaging the subject-level accuracies. To compute the predictions on the hold-out test set, we ensemble the predictions from all 38 networks (one for each fold) by averaging the logits vectors from all folds independently for each sample. We take the class with the highest probability as the final prediction.

### 5.1.2 Analysis of Spectral Organ Fingerprints

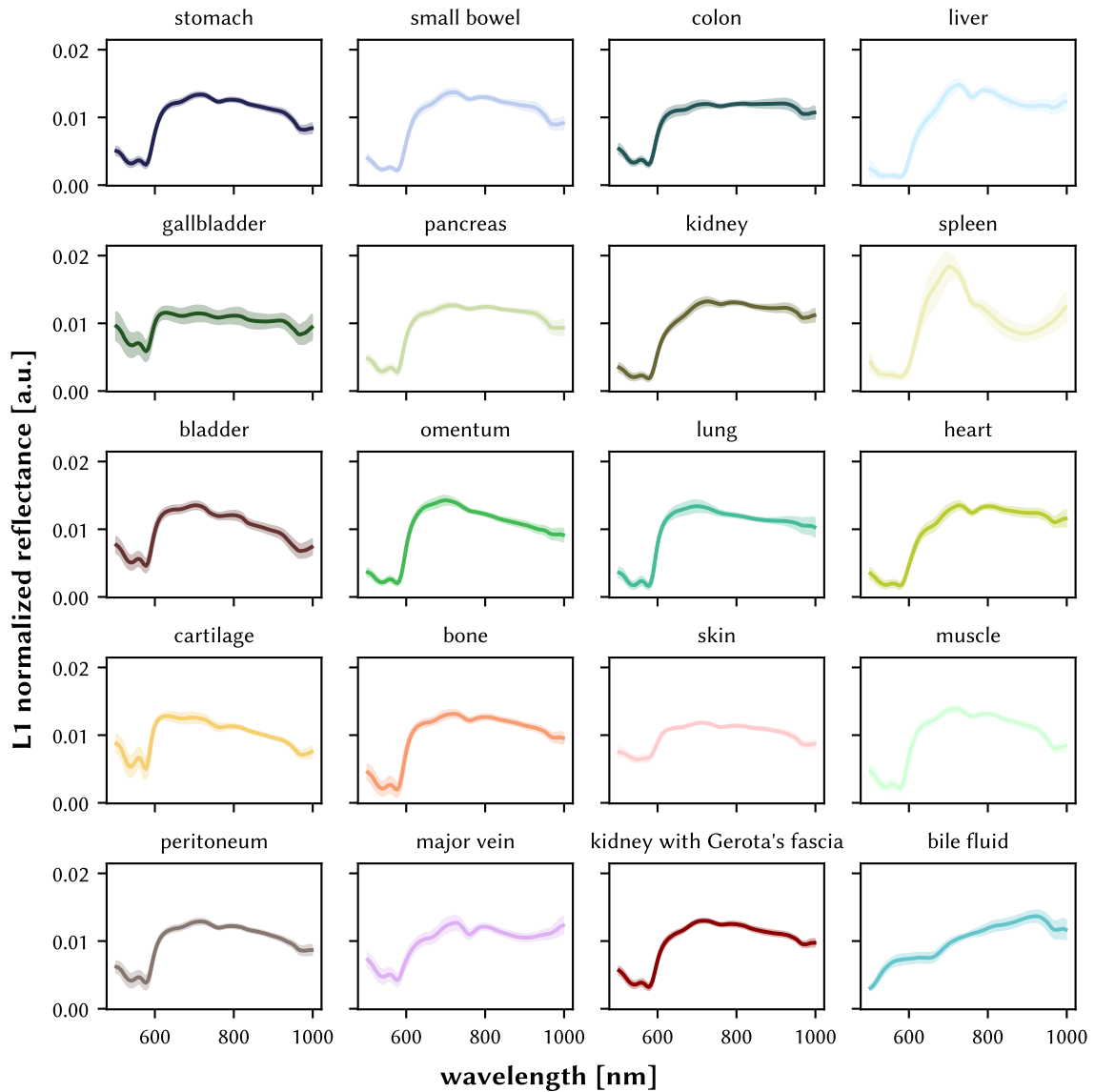
This section shows the results of the spectral analysis using the tissue atlas dataset by means of descriptive visualizations and results from our classification task.

Figure 5.2 shows the median spectra across all subjects for all 20 organ classes of the tissue atlas dataset. Whereas the fingerprints of some organs look similar (e.g., *small*



**Figure 5.1:** Overview of the  $k$ -fold structure of the tissue atlas dataset. The heatmap visualizes the assignment of the images from the tissue atlas dataset to the different splits used for training, validation and testing (each row denotes one fold and each column one image). Validation and test borders are always at subject boundaries. A leave-one-out-cross-validation structure is employed with one subject in the validation split per fold. Subjects from the standardized recordings are shown on the right, have significantly more images and constitute the test set.

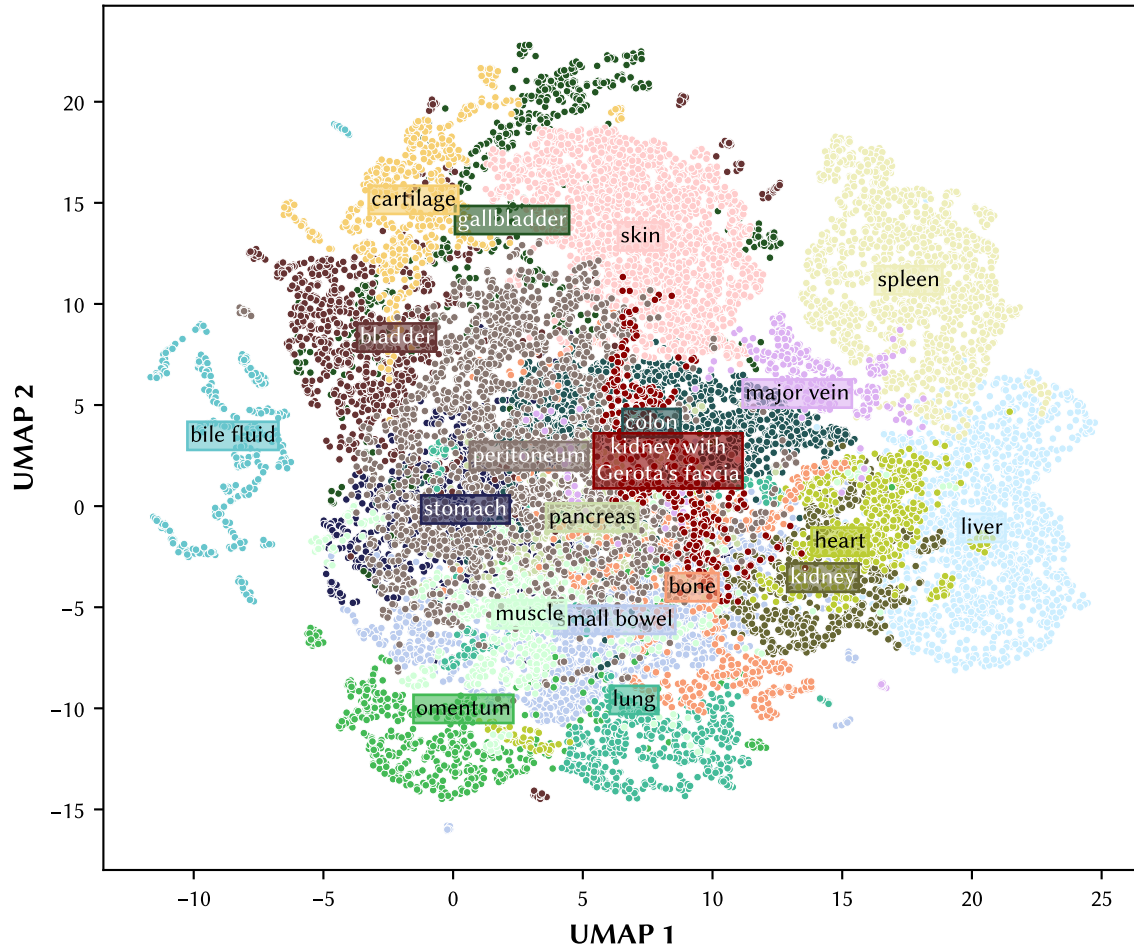
*bowel* and *stomach*), other organs are clearly distinguishable (e.g., *spleen* and *skin*). The SD across subjects is rather low for most organs indicating that the median spectra are similar across subjects. The *gallbladder* is an exception with a high SD among subjects. Across all organs, the SD is higher toward the near-infrared range of the spectrum (e.g., in the area above 950 nm) which could be explained by increasing noise in this area (see also Figure 2.7).



**Figure 5.2:** Spectral fingerprints of 20 organ classes in the tissue atlas. For each organ, the median spectra (solid line) hierarchically aggregated to the subject level as well as the standard deviation (shaded area) across subjects is shown. This figure was adapted from [215].



The 100-dimensional hyperspectral data of the tissue atlas is shown in a low-dimensional uniform manifold approximation and projection (UMAP) in Figure 5.3. It shows that classes like *bile fluid*, *spleen* or *omentum* form highly isolated clusters with the neighborhood being mainly comprised of spectra from the same class. Other classes like *pancreas*, *colon* or *muscle* are more mixed with other classes.



**Figure 5.3:** Visualization of the spectral neighborhood with uniform manifold approximation and projection (UMAP) as a non-linear dimensionality reduction tool of the tissue atlas. Each point represents the median spectrum of one polygon annotation for one organ in one image from one subject. The location of organ name boxes is based on the centroid of the projected class distribution. This figure was adapted from [215].

For the machine learning task, we classified individual median spectra per image and annotation in one of the 20 organ classes with the help of our spectrum classification network described in Section 4.2. We achieved a micro-averaged accuracy of 95.4 % (SD 3.6 %), a macro-averaged sensitivity, specificity and F1 score of 93.0 % (SD 6.3 %), 99.7 % (SD 0.2 %), and 92.3 % (SD 6.5 %), respectively. Figure 5.4 shows a confusion matrix of

the results. With 16 out of 20 classes, the majority of the organs achieved an average sensitivity of  $\geq 90\%$  (numbers on the diagonal). Exceptions are *heart*, *gallbladder*, *major vein* and *stomach* which are mainly confused with *kidney*, *bladder*, *bone* and *small bowel*, respectively.

### 5.1.3 HeiPorSPECTRAL: Open Dataset for Surgical Hyperspectral Imaging

For the scientific community, it is tremendously important to have public datasets accessible for research. This is especially true for the field of HSI where only a few small datasets were available previously [66, 97]. To improve upon this status quo, we published the standardized recordings of the tissue atlas dataset under the name HeiPorSPECTRAL. This section covers technical aspects of our dataset publishing process<sup>2</sup>.

#### Website

We made the dataset available through our website shown in Figure 5.5. The landing page gives basic information about the dataset, various visualizations (also detailed below), download and usage information, structure of the dataset and answers to frequently asked questions. The dataset complies with the FAIR principles [237] and does not have any access restrictions.

One of our primary goals was to make the access and usage of the dataset as easy as possible. For this, we developed the open source hyperspectral tissue classification (HTC) framework ([github.com/IMSY-DKFZ/htc](https://github.com/IMSY-DKFZ/htc) [200]) which is a Python package that can be used to work with the data (Figure 5.6). This is the same framework that we also used for our other studies (spectral fingerprint classification, data loading performance optimizations, surgical scene segmentation and context generalization) and hence not only contains a lot of example code but also ships with the pretrained models from our segmentation tasks. This allows users to directly use the data and models in their machine-learning tasks.

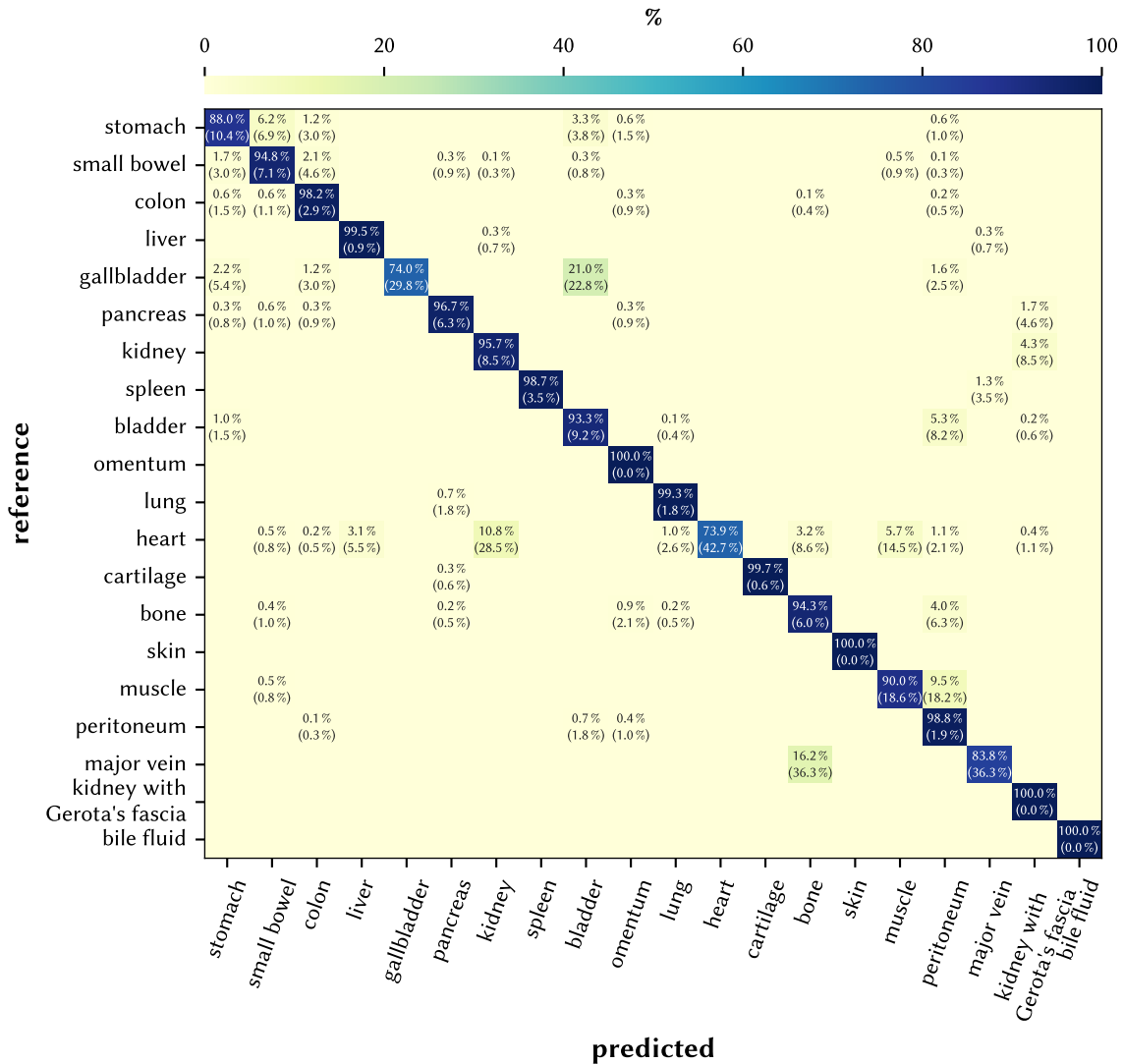
One of the first steps when working with a new dataset is to do exploratory data analysis (EDA) [45] to get a basic feeling about the data and its structure. To ease this process, we provided various visualizations all accessible and viewable on our website. Notable examples include interactive visualizations for every image (Figure 5.7) and profile pages for every organ with aggregated information about the dataset (Figure 5.8).

#### Technical Validation

For a dataset in general but especially when making it publicly available, it is crucial to ensure that the data is of high quality and accurately represents what it is meant to

---


<sup>2</sup>This section is based on [214].




**Figure 5.4:** Confusion matrix of the spectral classification task on the test set (tissue atlas dataset). The matrix depicts how median spectra from the reference class get classified. That is, every  $(i, j)$ -th entry shows the percentage of median spectra from class  $i$  that get classified as class  $j$  (on average). Values  $< 0.1\%$  are not shown for brevity. The matrix is row-normalized based on the median spectra from all images of one subject and then these matrices are averaged across subjects. The number in brackets denotes the standard deviation across subjects respecting the hierarchical structure of the data. Numbers on the diagonal denote the recall (sensitivity). This figure was adapted from [215].

measure. In our case, this means that our spectral data indeed measures (normalized) reflectances. For this, we compare colorchecker measurements of our HSI system with a


## 5 Experiments and Results



GERMAN  
CANCER RESEARCH CENTER  
IN THE HEIMHOLTZ ASSOCIATION




IMSY  
In Vivo MicroSpectroscopy



NATIONAL CENTER  
FOR TUMOR DISEASES  
HEIDELBERG

supported by

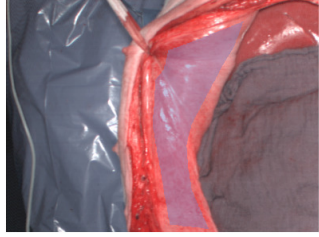
German Cancer Research Center (DKFZ)  
Heidelberg University Medical Center  
Hospital For Thoracic Diseases  
German Cancer Aid



HEIDELBERG  
UNIVERSITY  
HOSPITAL

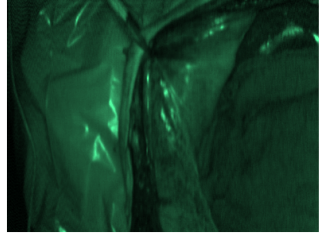
**RGB**

peritoneum



**HSI**

500 nm

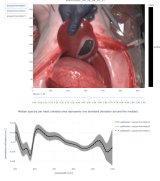


files: 170116 size: 949 GiB images: 5756 organs: 20 subjects: 11

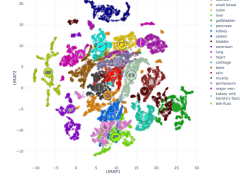
### Heidelberg Porcine HyperSPECTRAL Imaging Dataset (HeiPorSPECTRAL)

Welcome to the HeiPorSPECTRAL dataset! Here you can find 5756 hyperspectral imaging (HSI) cubes of 20 physiological organ classes across 11 pigs annotated by 3 medical experts. For each organ, there are about 36 recordings per pig: 4 different intraoperative situations ( *situs* ) were imaged from 3 different angles ( *angle* ) for 3 times ( *repetition* ). The HSI cubes were acquired with the [Tivita® Tissue camera system from Diaspective Vision](#). Each data cube has the dimensions (480, 640, 100) = (height, width, channels) with the 100 non-overlapping spectral channels being in the range from 500 nm to 1000 nm at a spectral resolution of around 5 nm. This repository contains the raw data, related metadata as well as the preprocessed files. The figures from the paper and example visualizations for each image are available at <https://figures.heiporspectral.org/>.

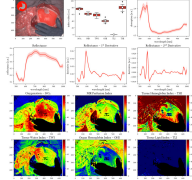
More details can be found in our publication: Studier-Fischer, A., Seidlitz, S., Sellner, J. et al. HeiPorSPECTRAL - the Heidelberg Porcine HyperSPECTRAL Imaging Dataset of 20 Physiological Organs. Sci Data 10, 414 (2023). <https://doi.org/10.1038/s41597-023-02315-8>




Interactive visualization for every image.



(Interactive) figures of the paper (including supplementary figures).



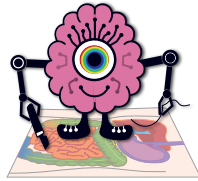
Organ profiles aggregated across the entire dataset.



Frequently asked questions.

**Figure 5.5:** Landing page of the HeiPorSPECTRAL website ([heiporspectral.org](https://heiporspectral.org)).

## Usage



Hyperspectral  
Tissue  
Classification

We recommend using the data with the [htc framework](#), which offers:

- a pipeline to efficiently load and process the HSI cubes, annotations and metadata.
- a framework to train neural networks on the data, including the implementation of several classification and segmentation models.
- simple usage of pre-trained models.

Installation (example for Unix-based systems):

```
# Make the dataset available
export PATH_HeiPorSPECTRAL=/mnt/nvme_4tb/HeiPorSPECTRAL

# Install the htc package
pip install imsy-htc
```

As a teaser, this is how you can use the `htc` framework to read a data cube, corresponding annotation and parameter images:

```
import numpy as np
from htc import DataPath, LabelMapping

# You can load every image based on its unique name
path = DataPath.from_image_name('P086#2021_04_15_09_22_02')

# HSI cube format: (height, width, channels)
print(path.read_cube().shape)
# (480, 640, 100)

# Annotated region of the selected annotator
mask = path.read_segmentation("polygon#annotator1")
print(mask.shape)
# (480, 640)

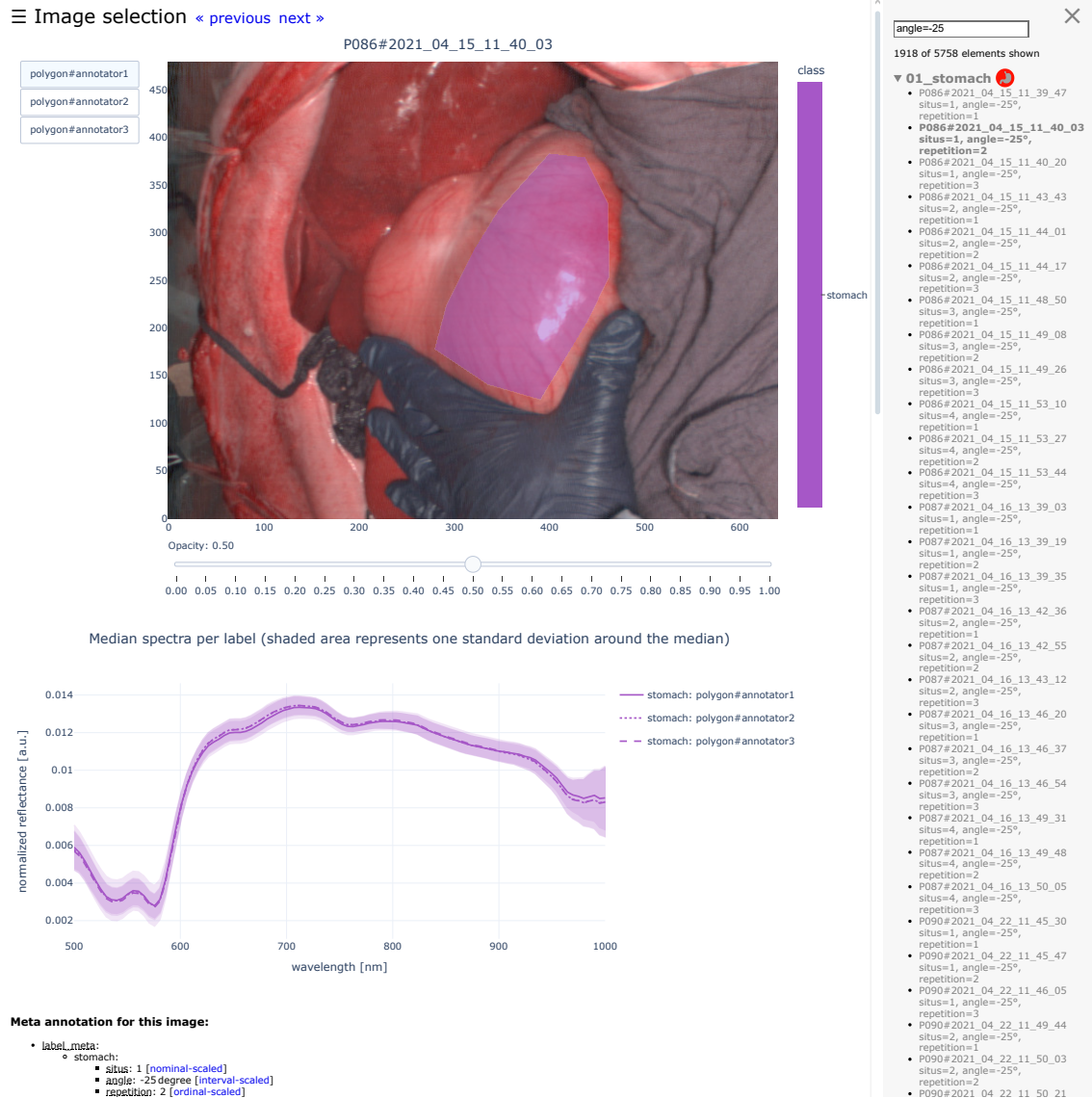
# Additional meta information about the image
print(path.meta("label_meta"))
# {'spleen': {'situs': 1, 'angle': 0, 'repetition': 1}}

# Tivita parameter images are available as well
sto2 = path.compute_sto2()
print(sto2.shape)
# (480, 640)

# Example: average StO2 value of the annotated spleen area for annotator1
# The dataset_settings.json file defines the global name to index mapping
spleen_index = LabelMapping.from_path(path).name_to_index("spleen")
print(round(np.mean(sto2[mask == spleen_index]), 2))
# 0.44
```

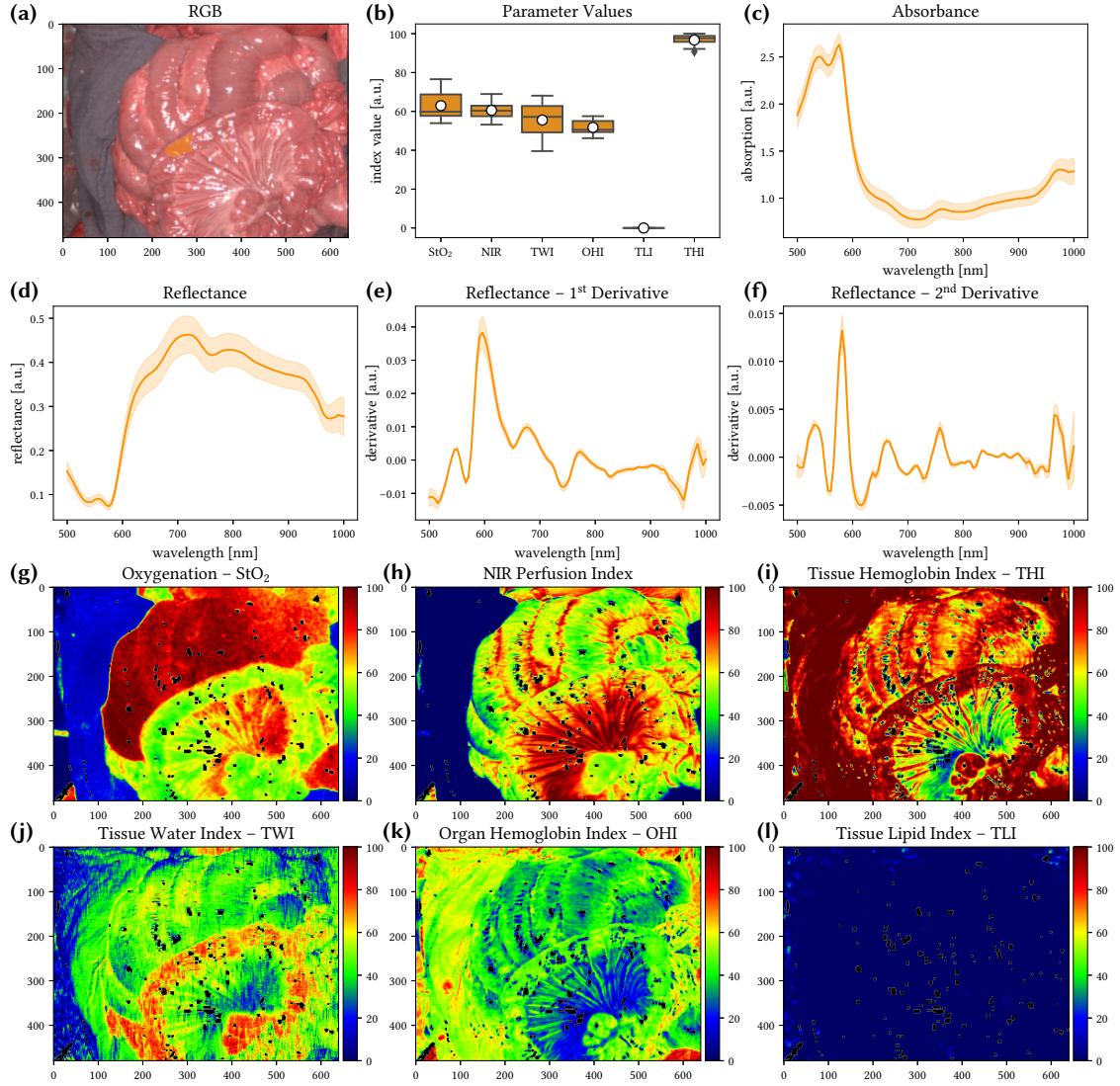
**Figure 5.6:** Usage example of the HeiPorSPECTRAL dataset using the hyperspectral tissue classification (HTC) framework ([github.com/IMSY-DKFZ/htc](https://github.com/IMSY-DKFZ/htc) [200]). The example is from the HeiPorSPECTRAL website ([heiporspectral.org/#usage](https://heiporspectral.org/#usage)).

## 5 Experiments and Results



**Figure 5.7:** Interactive figure for an example image with a stomach annotation from the HeiPor-SPECTRAL dataset ([figures.heiporspectral.org/view\\_organs/01\\_stomach](https://figures.heiporspectral.org/view_organs/01_stomach)). These views are available for all images of the dataset and allow interactive visualization of the different annotations, median spectra visualization, overview of the available metadata as well as a navigation pane which can be used to search for specific images (here to search for all images with a measurement angle of  $-25^\circ$ ).





**Figure 5.8:** Example of an organ profile page with aggregated information of the entire HeiPor-SPECTRAL dataset for the small bowel class ([figures.heiporspectral.org/1abel\\_profiles](https://figures.heiporspectral.org/1abel_profiles)). Based on the median value per polygon annotation, hierarchical aggregation happened from annotations over images to subjects. **(a)** Exemplary RGB image with corresponding polygon annotation. **(b)** Distribution of subject-level functional parameter values. Each boxplot shows the interquartile range (IQR) with the median (black solid line), mean (white circle) and outliers (rhombus). The whiskers extend up to 1.5 times of the IQR. **(c-f)** Subject-level median spectra curves (solid lines) and standard deviations (shaded areas) for absorbance, reflectance as well as first and second order derivatives thereof. **(g-l)** Tissue parameter images for the example image. These profile pages are available for all of the 20 organ classes.

spectrometer<sup>3</sup> in Figure 5.9 and found good agreement. Deviations exist mainly in the near-infrared range of the spectrum and for very dark colors. Repeated measurements on separate days with our HSI system led to very similar results.

This finding is not only important for the users of the HeiPorSPECTRAL dataset but also for the other results of this thesis in general as all the datasets described in Section 4.1 were acquired with the same HSI system. The deviations in the near-infrared range could explain the lesser significance of that range for organ differentiation (Figure 5.29) as well as the observed increased SDs of that range (Figure 5.2).

## 5.2 Efficient Training of Hyperspectral Segmentation Networks

This section highlights the improvements from our counter-strategies (cf. Section 4.3) when included in the training of a segmentation network<sup>4</sup>. Section 5.2.1 outlines how we ensure a fair comparison comparison and Section 5.2.2 shows the results of our benchmarking.

### 5.2.1 Experimental Setup

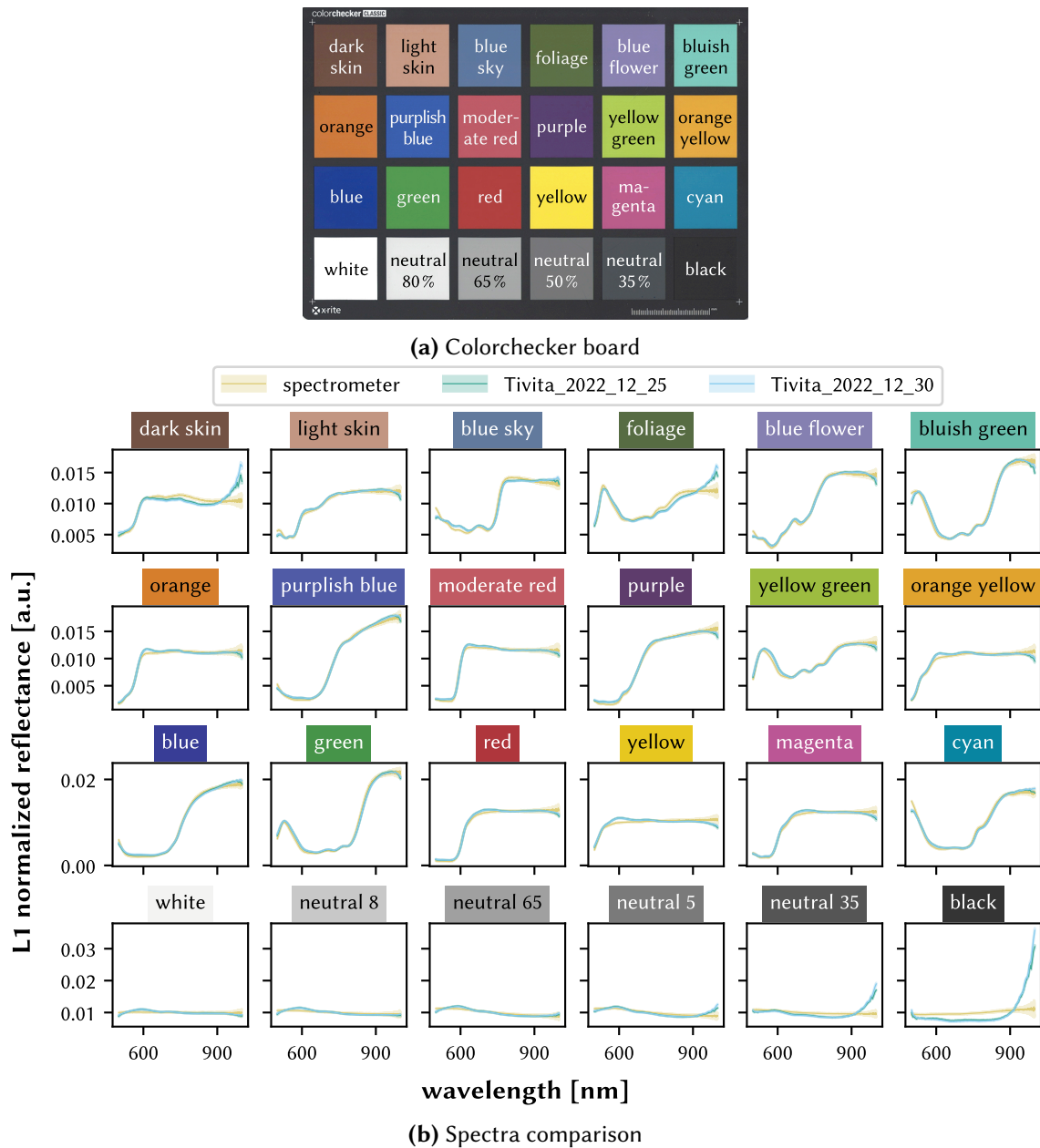
We used a Docker [148] container to benchmark the different counter-strategies presented in Section 4.3 in an isolated environment. We trained an image-based segmentation network for 5 epochs and used a batch size of 6. During one epoch, we iterated over all 506 images of the semantic porcine dataset (including training and testing images). After each epoch, the system cache of Linux was cleared which ensures that images always have to be loaded from disk and are never cached from memory. We disabled stochastic weight averaging (SWA) to measure only the time of normal training epochs without additional aggregation steps employed by SWA [103]. No checkpoints were saved. We only measured the time for operating on the batches excluding pre- and post-initialization steps before and after the training epoch. We repeated the experiment three times and are reporting the mean and SD of the measured times. The hardware utilization was measured with the help of the `nvidia-smi` tool [50]. Counter-strategies were added additively to avoid benchmarking all possible combinations of strategies.

---

<sup>3</sup>For comparison, we L1 normalized the spectra of our HSI system and the spectra of the spectrometer. The normalized spectra from the spectrometer were further scaled by a factor of 11.31 to account for the different number of channels between the two systems (1131 instead of 100 channels for the spectrometer in the range between 500 nm and 1000 nm).

<sup>4</sup>This section is based on [201].





**Figure 5.9:** Spectral comparison of the Tivita<sup>®</sup> Tissue (Diaspective Vision GmbH, Am Salzhaff, Germany) camera with a point spectrometer (Ocean Insight HR2000+; light source: Tungsten Halogen lightsource Ocean Insight HL-2000 (formerly Ocean Optics; Orlando, Florida, US)). **(a)** Example image of the ColorChecker Classic<sup>®</sup> board from x-rite (Grand Rapids, Michigan, US) composed of 24 standardized color chips. **(b)** Comparison of median spectra (solid lines) of the spectrometer and two days of hyperspectral imaging (HSI) recordings for each of the 24 color chips. The shaded area shows the standard deviation across 100 and 13 repeated measurements for the spectrometer and HSI data, respectively. This figure was adapted from [214].

We trained the networks on a local computer with an AMD Ryzen™ 9 7950X CPU, an NVIDIA® GeForce RTX™ 4090 GPU, a 4 TB Seagate FireCuda® 530 SSD and 64 GB Kingston FURY™ Beast DDR5 RAM running on an Ubuntu 22.04 system. From the software side, we used PyTorch 2.1 with CUDA® 12.1, Python 3.11 and NVIDIA® driver 535.129.03. It was ensured that no heavy task was running in the background while the benchmark was executed.

The speed of the SSD plays an important role in this benchmark as higher read speeds can compensate for inefficient data storage formats to some extent. To account for this, we repeated the measurements twice, once without any limitation on the SSD bandwidth and once while limiting the read speeds of the SSD to 1000 MB/s via Docker runtime configuration arguments.

### 5.2.2 Benchmarking Data Loading Strategies

This section shows the results from our benchmarking of the different data loading strategies by means of time measurements and hardware utilization.

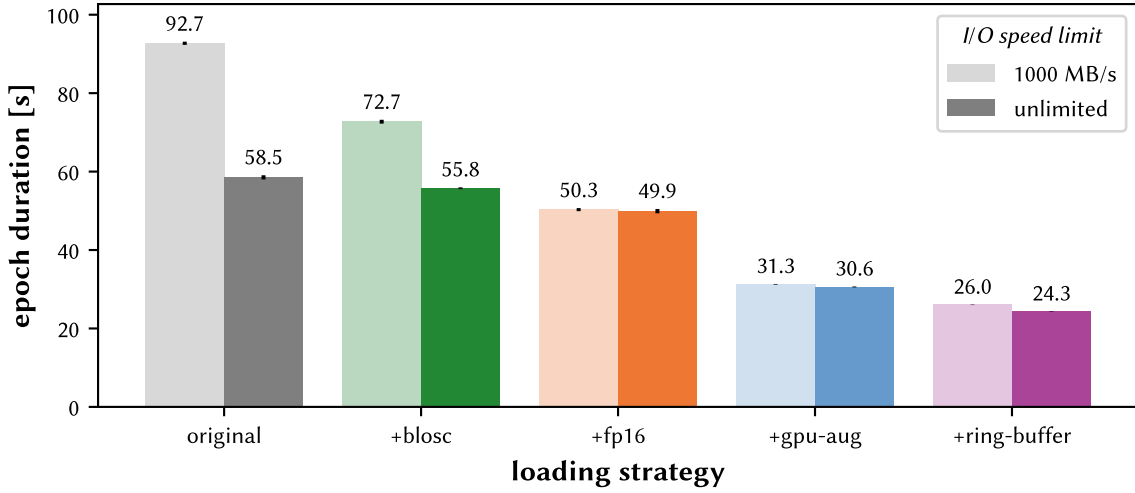
Figure 5.10 shows that we can reduce the training time with every added strategy with a total speedup in training time of 3.6 and 2.4 for the limited (read speed of the SSD limited to 1000 MB/s) and unlimited (no constraints on the read speed of the SSD) case, respectively. The improvement in the limited scenario is higher for *blosc* and *fp16* than for the other counter-strategies indicating the importance of efficient data formats when not enough bandwidth is available (as may be the case in cluster environments).

In all cases, the SD across the three repetitions of the experiment is very low indicating that the results are stable. This can mainly be attributed to our experimental setup (cf. Section 5.2.1) which ensures a minimal impact of unwanted side effects (e.g., clearing of the Linux system cache).

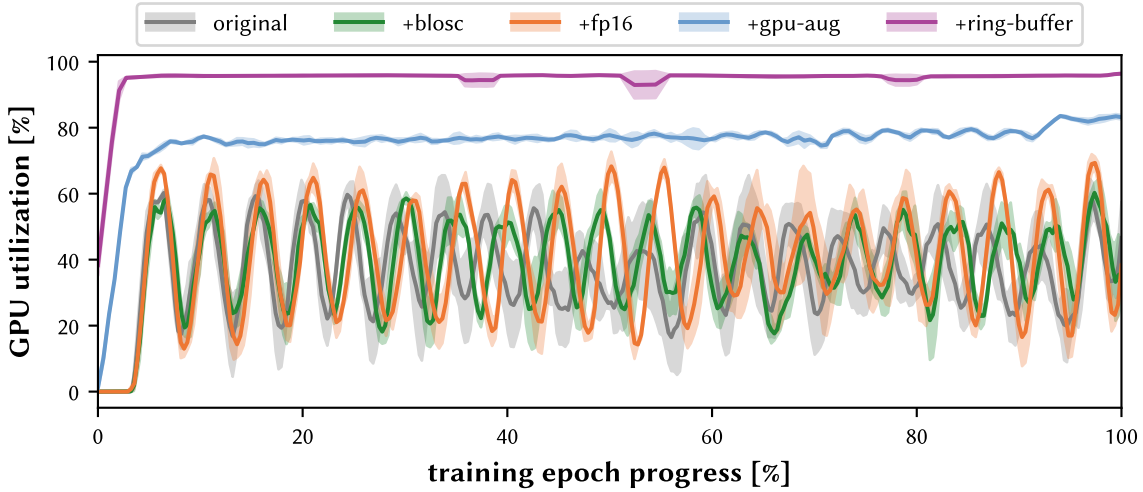
Similarly, the utilization also improves while making use of our counter-strategies (Figure 5.11) with an almost saturated performance when all counter-strategies are used. The wave-form shape of the utilization curves (e.g., in the original case) is a direct manifestation of the idle times of the GPU as sketched in Figure 4.11.

## 5.3 Surgical Scene Segmentation of Hyperspectral Images

The main objectives of the studies in this section were to explore the appropriate representation of HSI data in terms of segmentation performance, the effect of the size of the training data and an analysis of the inherent variability in our networks as well as the



**Figure 5.10:** Benchmarking results of different data loading strategies. The input/output (I/O) speed limit constrains the read speed of the solid-state drive (SSD) during the experiment. The error bars show the standard deviation across three repetitions of the experiment. The loading strategies are applied additively from left to right (e.g., fp16 also includes blosc compression). This figure was adapted from [201].



**Figure 5.11:** Hardware utilization of different data loading strategies without read speed limitations. The shaded area shows the standard deviation of the graphics processing unit (GPU) utilization across three repetitions of the experiment. The loading strategies are applied additively from left to right (e.g., fp16 also includes blosc compression). This figure was adapted from [201].

effect of ensembling<sup>5</sup>. Section 5.3.1 starts with the details about our experimental design and Section 5.3.2 shows the results of our segmentation networks.

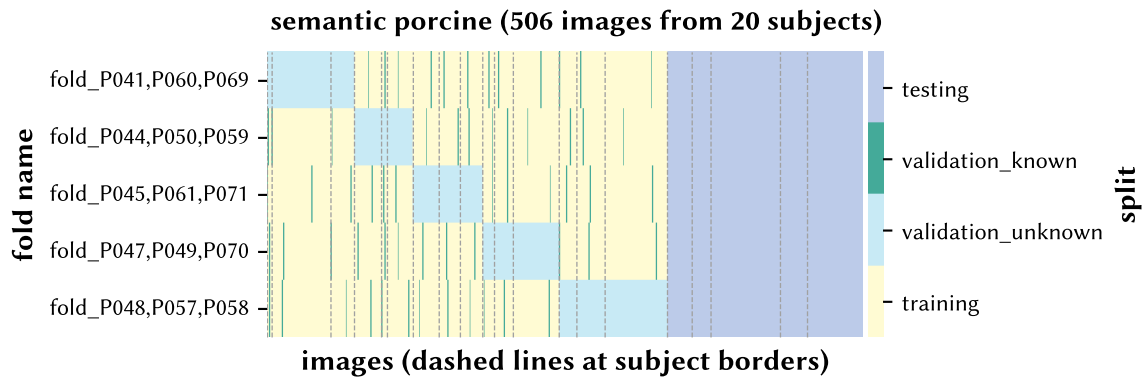
<sup>5</sup>This section is based on [198].

### 5.3.1 Experimental Setup

Our segmentation networks are trained and evaluated on the semantic porcine dataset. This section outlines how we split the data for training, validation and testing as well as the employed evaluation metrics, our algorithm ranking procedure, how we assessed the quality of our reference annotations and the details of our training size experiment.

#### Splits

We divided the dataset, which includes 506 images from 20 subjects, into a training set of 340 images from 15 subjects and a separate test set of 166 images from 5 subjects. The test subjects were randomly chosen while ensuring that each organ class was represented in both the training and test sets. We performed  $k$ -fold cross-validation on the training set with  $k = 5$ . The folds were created to maximize the number of organ classes across validation folds. We refer to the conventional validation set obtained for each fold, which consists of three subjects not seen during training, as `validation_unknown`. Further, we created a second validation set, `validation_known`, by removing one random image from each of the 12 subjects included in each of the five training sets. This validation set consists of 12 unseen images (per fold) from known subjects. The splitting details are visualized in Figure 5.12.



**Figure 5.12:** Overview of the  $k$ -fold structure of the semantic porcine dataset. The heatmap visualizes the assignment of the images from the semantic porcine dataset to the different splits used for training, validation and testing (each row denotes one fold and each column one image). The validation set is further divided into two sets, one with images from training subjects (`validation_known`) and one with subjects that are not part of the training set (`validation_unknown`). Borders for the `validation_unknown` and testing splits are always at subject boundaries. A  $k = 5$  cross-validation structure is employed with three subjects in the validation split per fold.

By comparing the model’s performance on the two validation sets, we could estimate its generalization capabilities toward unseen surgeries. This includes changes due to inter-subject variability in addition to context-related changes (e.g., variations in the performed

surgery, surgical phase during acquisition, and visible instruments) and imaging-related changes (e.g., different imaging perspectives). The latter two sources of variation may also be present across different images of the same surgery but may be more pronounced across different surgeries.

We did not use the test set during model development. Only after we finalized the model architectures and parameters based on the performance on the validation set, did we evaluate the segmentation performance on the hold-out test data set by averaging the softmax values to ensemble the network predictions from all folds.

### Metrics

Given that individual validation metrics may not capture all clinical requirements [244, 184, 141, 183], we used several validation metrics, each with their own strengths and limitations, to provide a more comprehensive assessment of model performance. Namely, we used the DSC as an overlap-based measure, the average surface distance (ASD) as a distance-based measure and the normalized surface dice (NSD) as a boundary-overlap-based measure with special consideration to annotation uncertainty.

Similar to the classification task, we respect the hierarchical structure of the data and aggregate the per class and image scores first to the image and then to the subject level (see Figure 5.13). The subject-level scores (for all metrics) served as input to our visualizations and model rankings.

The DSC [57], a widely used validation metric for biomedical segmentation tasks, measures the overlap between a predicted object segmentation and the corresponding reference segmentation. It is computed for each class  $i$  with the set of predicted pixels  $\mathcal{P}_{\text{PRE}}^i$  and reference pixels  $\mathcal{P}_{\text{REF}}^i$  for that class via:

$$\text{DSC}^i = \frac{2 \cdot |\mathcal{P}_{\text{PRE}}^i \cap \mathcal{P}_{\text{REF}}^i|}{|\mathcal{P}_{\text{PRE}}^i| + |\mathcal{P}_{\text{REF}}^i|}. \quad (5.1)$$

DSC values are in the range  $[0; 1]$ . A value of 0 indicates no overlap between the predicted and reference segmentation for the class, or the class is present in the image but not predicted. A value of 1 denotes a perfect overlap between the prediction and the reference segmentation for a class. The DSC is highly sensitive to the object size and insensitive to the object shape [184].

It is worth noting that any binary segmentation task can be viewed as a pixel-wise classification task and then the DSC is equivalent to the F1-score [183]. Moreover, the DSC is closely related to the intersection over union (IoU) [104], a commonly used validation metric in the general computer vision machine learning community. We used the MONAI framework [31] to compute the DSC.

Boundary-distance-based metrics evaluate the dissimilarity between the predicted segmentation and reference segmentation based on the distances between their boundaries.

Unlike overlap-based metrics, these metrics are insensitive to the object size but sensitive to the object shape. An example is the ASD [91] which calculates the average of all distances between pixels on the predicted object segmentation border  $\mathcal{B}_{\text{PRE}}^i$  and its nearest neighbor on the reference segmentation border  $\mathcal{B}_{\text{REF}}^i$  for a class  $i$ . In this study, we used the symmetric version of the ASD, which computes the set of nearest neighbor distances  $\mathcal{D}_{\text{PRE}}^i$  again, but with the roles of the predicted and reference segmentation reversed, resulting in  $\mathcal{D}_{\text{REF}}^i$ . All the obtained distance values are then averaged to yield an average distance value  $\text{ASD}^i$  for each class

$$\text{ASD}^i = \frac{\sum_{d \in \mathcal{D}_{\text{PRE}}^i} d + \sum_{d' \in \mathcal{D}_{\text{REF}}^i} d'}{|\mathcal{D}_{\text{PRE}}^i| + |\mathcal{D}_{\text{REF}}^i|}. \quad (5.2)$$

One drawback of the ASD is that it is unbounded, producing values in the range  $[0; \infty[$ , with 0 indicating an exact match of object boundaries. Therefore, ASD values are generally more difficult to interpret. Additionally, special consideration must be given to missed classes (classes present in the reference annotations but not predicted), as there is no natural limit [184]. In this study, we set the ASD value for a missed class to the maximum ASD obtained for the other classes on the same image. This approach introduces a potentially high and image-dependent penalty if a class cannot be predicted in an image (cf. Section 6.3).

The NSD [165] quantifies the proportion of a segmentation boundary that is accurately predicted, considering a threshold  $\tau$  that signifies the clinically acceptable pixel deviation. Essentially, it measures the fraction of the segmentation boundary that would need to be adjusted to correct for segmentation inaccuracies. Rather than using a single universal threshold  $\tau$ , we employed a class-specific threshold  $\tau^i$  for each class  $i$  acknowledging the varying annotation difficulty across different classes (for instance, a class with a distinct boundary like the liver is simpler to annotate precisely than a class with an indistinct boundary such as the omentum). We modified the NSD, originally designed for 3D segmentation maps, to suit our 2D segmentation maps<sup>6</sup>. Instead of dealing with 3D segmentation surfaces, we focused on 2D segmentation boundaries. For all pixels on the predicted segmentation boundary of class  $i$ ,  $\mathcal{B}_{\text{PRE}}^i$ , we calculated the nearest-neighbor distances to the reference segmentation boundary  $\mathcal{B}_{\text{REF}}^i$ , yielding a set of distances  $\mathcal{D}_{\text{PRE}}^i$ . We then identified the subset  $\mathcal{D}_{\text{PRE}}^{\prime i}$  of distances in  $\mathcal{D}_{\text{PRE}}^i$  that are less than or equal to the acceptable deviation  $\tau^i$ .

$$\mathcal{D}_{\text{PRE}}^{\prime i} = \{d \in \mathcal{D}_{\text{PRE}}^i \mid d \leq \tau^i\}. \quad (5.3)$$

The entire process was symmetrically carried out for  $\mathcal{B}_{\text{REF}}^i$ , resulting in  $\mathcal{D}_{\text{REF}}^i$  and  $\mathcal{D}_{\text{REF}}^{\prime i}$ . For each class present in both the predicted and reference segmentation, the  $\text{NSD}^i$  was

---

<sup>6</sup>Our implementation can now also be found in the MONAI framework [31].

subsequently calculated as:

$$\text{NSD}^i = \frac{|\mathcal{D}'_{\text{PRE}}| + |\mathcal{D}'_{\text{REF}}|}{|\mathcal{D}_{\text{PRE}}| + |\mathcal{D}_{\text{REF}}|} \quad (5.4)$$

The NSD produces values within the range  $[0; 1]$ . A value of 0 signifies that either the boundary is entirely incorrect, with all distances exceeding the acceptable deviation  $\tau^i$ , or that the class exists in the image but has not been predicted<sup>7</sup>. A value of 1 is achieved if there is no need to adjust the segmentation boundary, as all deviations are within the acceptable threshold  $\tau^i$ .

One of the significant challenges with the NSD is establishing the class-specific thresholds  $\tau^i$  (see also discussion in Section 6.3). To address this, 20 randomly chosen images (one image per subject, ensuring all classes are represented by at least two images) were re-annotated by another medical expert. Similar to the ASD, we calculated distances between the boundaries of the original annotation and the re-annotation for each class  $i$  in each image  $k$  and averaged the results to derive the image- and class-specific threshold  $\tau_k^i$ . If a class was annotated in only one of the two corresponding images, no distances could be calculated and the corresponding structure was disregarded (this point is also picked up in Section 6.3). We determined the class-specific distance threshold  $\tau^i$  by averaging the  $\tau_k^i$  for the set of images  $\mathcal{J}^i$  where the class  $i$  is present and  $\tau_k^i$  could be calculated:

$$\tau^i = \frac{1}{|\mathcal{J}^i|} \sum_{k \in \mathcal{J}^i} \tau_k^i. \quad (5.5)$$

## Ranking

We examined the model ranking and its stability to two variability sources: sampling variability and variability due to the metric choice. The ranking analyses were conducted according to [236] and the model ranking was established as follows: For each model, we computed the average of the 5 per-subject metric values and ranked the models based on these mean metric values. To evaluate the ranking stability concerning different validation metrics, we performed the ranking separately for each metric and compared the results. To evaluate the model ranking's stability concerning sampling variability, we repeatedly performed rankings on 1000 bootstrap samples. Each bootstrap sample comprised 5 cases randomly selected with replacement from the set of 5 test cases (one metric value for each subject in the test set). We first averaged metric values across all 5 test cases and then determined each model's rank based on this aggregate resulting in 1000 ranks for each model (one rank per bootstrap sample).

<sup>7</sup>The opposite case, where a class was predicted which does not exist in the image, does not yield a value of 0. This is because the NSD<sup>i</sup> (or any of our employed metrics) is only calculated for classes  $i$  which exist in the reference image. Instead, the misprediction decreases the score of other classes (depending on where the misprediction happened and which classes in the reference image are located there).

### Hierarchical Aggregation

The HSI datasets used in this thesis are composed of a natural hierarchical structure of subjects which are composed of one or more images with one or more classes. To account for the hierarchical nature of the data (following [93, 141]), we compute all metric values for each image and class and then aggregate the values while respecting the hierarchy. Metric values can be anything from accuracy, over NSD to confusion matrices. The general concept is visualized in Figure 5.13 for the two primary targets subject-level scores and class-level scores.

Subject-level scores indicate how well a single subject performs, e.g., how well images from this subject can be segmented. That is, on average, given an image from this subject, how well can a neural network segment the image? In visualizations (e.g., boxplot), the shown data points correspond to the individual subjects in the test set. Subject-level scores are a general performance indicator of a system that does not take the class distribution across images into account, i.e., the scores highly depend on the visible classes in the scene.

In our HSI datasets, there are very inhomogeneous class distributions across images due to the different surgical scenes being viewed. This is also reflected in the dataset statistics in Figure 4.2 (tissue atlas dataset) and Figure 4.4 (semantic porcine dataset). Therefore, class-level metric scores are also very useful and indicate how well a neural network can handle this class. In visualizations (e.g., boxplot), the shown data points correspond to the individual classes.

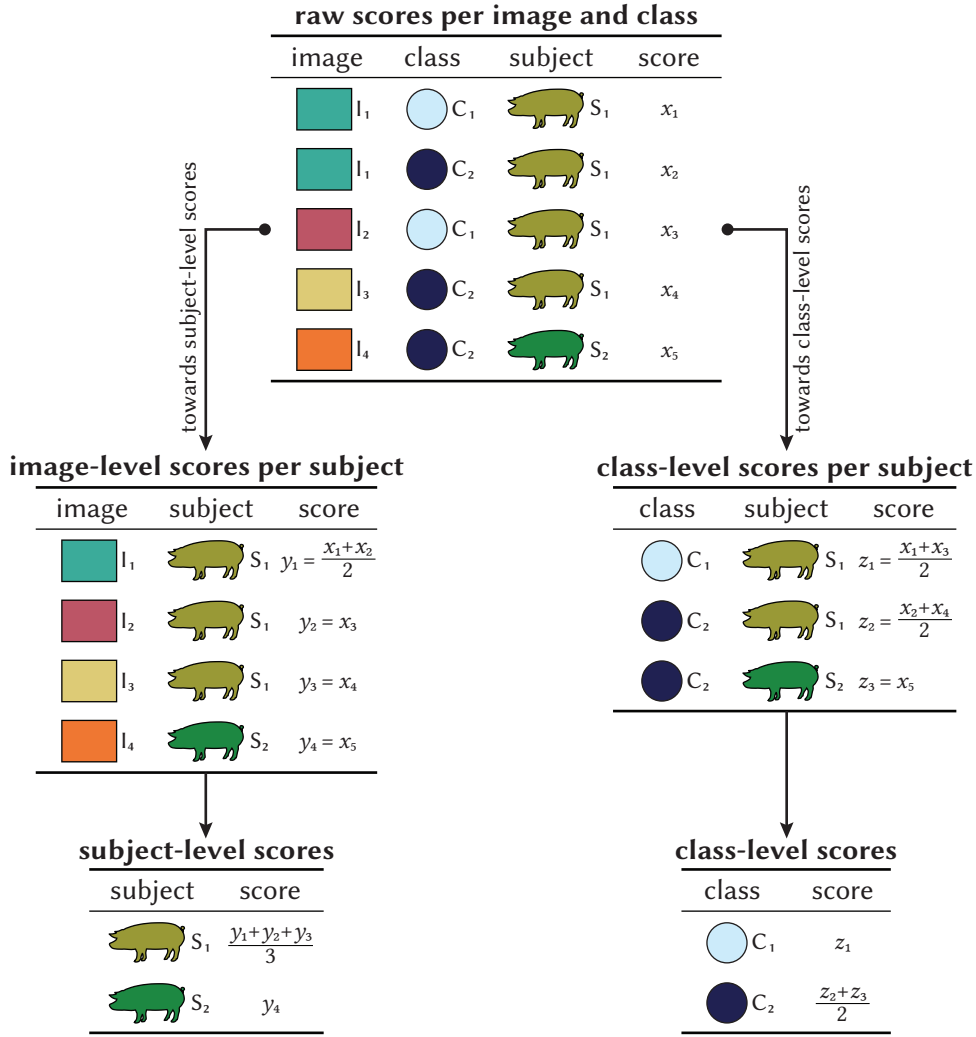
### Rater Variability

The quality of data, including the available reference annotations, is vital for any deep learning algorithm. Previous studies have shown that the variability between different human raters can be substantial [110]. To measure the quality of our reference annotations, we used the same set of re-annotated images as used to determine the NSD thresholds mentioned earlier and the annotations of the second medical expert for inter-rater estimations. Furthermore, the medical expert who initially annotated our dataset re-annotated the same set of 20 images to estimate the intra-rater variability. In both scenarios, we compared the re-annotations with the original annotation per image and calculated our three evaluation metrics. Unlike the comparison of model predictions, we did not use pixels that the new annotator assigned to the *ignore* class (for instance, pixels for which an annotator was uncertain) because it is valid to state uncertainty and this should not be penalized as misprediction. In other words, we computed the union of ignored pixels in the reference and re-annotated segmentation maps, and those pixels were disregarded.

### Training Size Experiment

To investigate the performance of various models depending on the amount of training data available, we randomly selected  $n$  subjects from the set of 15 training subjects without replacement and adjusted  $n$  from 1 to 14. The models were trained solely on the images





**Figure 5.13:** Visualization of the hierarchical aggregation schemes of metric scores as they are used throughout this thesis. Based on the metric scores (e.g., dice similarity coefficient (DSC) or normalized surface dice (NSD)) per class and image, scores can either be aggregated toward subject-level scores (left part) or toward class-level scores (right part) by averaging across the hierarchy. For the left part, class-level scores are first aggregated to image-level scores per subject and then aggregated to subject-level scores based on all images per subject. For the right part, scores from different images of the same class are aggregated to class-level scores per subject and then to class-level scores by aggregating across all subjects.

of the  $n$  sampled subjects without  $k$ -folds, while the performance was assessed on the usual test split of Figure 5.12 but only for 8 classes and without ensembling. These 8 classes (*stomach*, *small bowel*, *colon*, *liver*, *spleen*, *skin*, *peritoneum* and *background*) are the set of classes for which images are available for all 15 training subjects. This design

choice prevents the issue of sampling a subject during training that does not contain any of the target classes which would not yield indicative results for this experiment. To enhance the stability of the results toward subject sampling variability, we repeated the experiment 5 times with different random subject selections.

### 5.3.2 Analysis of Segmentation Networks

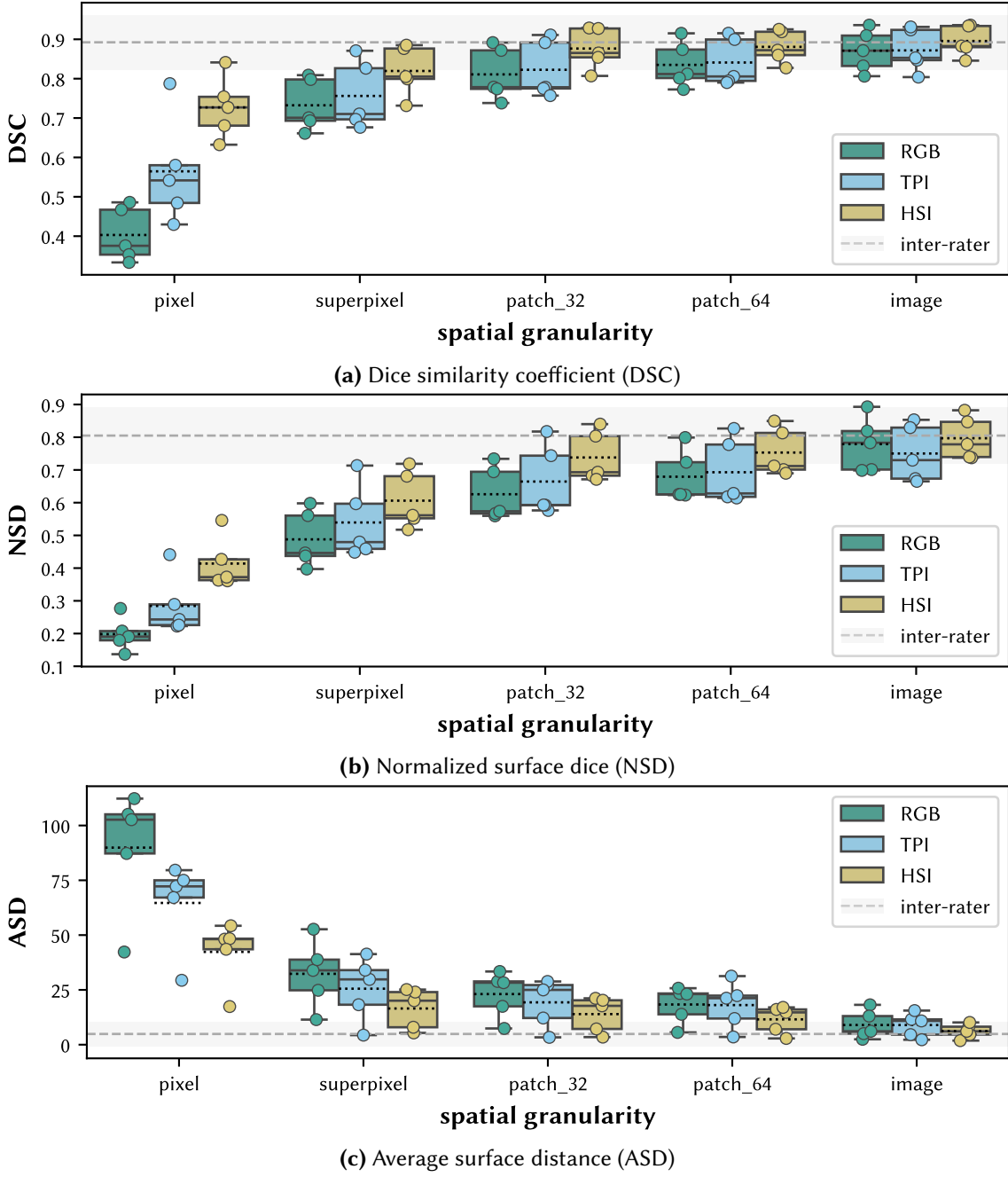
This section shows the results from our study about modalities and spatial granularities on the semantic porcine dataset by means of various performance comparisons and analyses of the segmentation networks.

#### Segmentation Performance

Figure 5.14 illustrates the performance of the pixel-based, superpixel-based, patch-based (denoted as patch\_32 for an input shape of  $32 \times 32 \times c$  and patch\_64 for an input shape of  $64 \times 64 \times c$ ) and image-based segmentation models for all three modalities separately for our three validation metrics DSC, NSD and ASD. Although the performance differences for various spatial granularities are less noticeable for HSI data than for RGB and TPI data, a consistent trend is evident across all modalities and validation metrics: the larger the spatial granularity of the input data, the superior the segmentation performance.

These results suggest the use of the image HSI model for its superior performance and this is indeed what we focused on in our follow-up studies (e.g., context experiments). However, there may be a benefit in using smaller granularities if larger batch sizes are required as suggested by Table 4.3. In addition to the benefits of smoother gradients and improved batch statistics usually attributed to larger batch sizes [101], it may also be advantageous to have a batch distribution more similar to the training distribution. For instance, confounders in HSI data could potentially lead to an overestimation of machine learning performance [58]. However, some promising techniques to achieve confounder-invariant representations (e.g., metadata normalization [135]) require large batch sizes and would thus be more suitable for a pixel-based model rather than an image-based model.

Comparing the different modalities, we can observe that the average segmentation performance on HSI data consistently outperforms the performance on RGB and TPI data. The performance gap is most significant in pixel-based segmentation, decreases with increasing spatial granularity, and is smallest in image-based segmentation. This could be because the model can leverage additional information from the spatial context, compensating for the lack of detail in the spectral dimension. However, the smaller performance gap for the increased spatial context might also be due to the quality of the expert annotations provided, as the performance of our HSI models approaches the inter-rater variability.



**Figure 5.14:** Model segmentation performance of different spatial granularities and modalities (RGB, tissue parameter images (TPI) and hyperspectral imaging (HSI)). The dashed line and the shaded area denote the mean and standard deviation of the inter-rater performance. Each boxplot shows the interquartile range (IQR) with the median (solid line) and mean (dotted line). The whiskers extend up to 1.5 times of the IQR. Each point represents the performance of one test subject. This figure was adapted from [198].

In terms of the comparison between TPI and RGB data, we observed that in most instances, a model based on TPI data performs better than the same model based on RGB data. This suggests that the manually derived TPI data contains information relevant to the segmentation task.

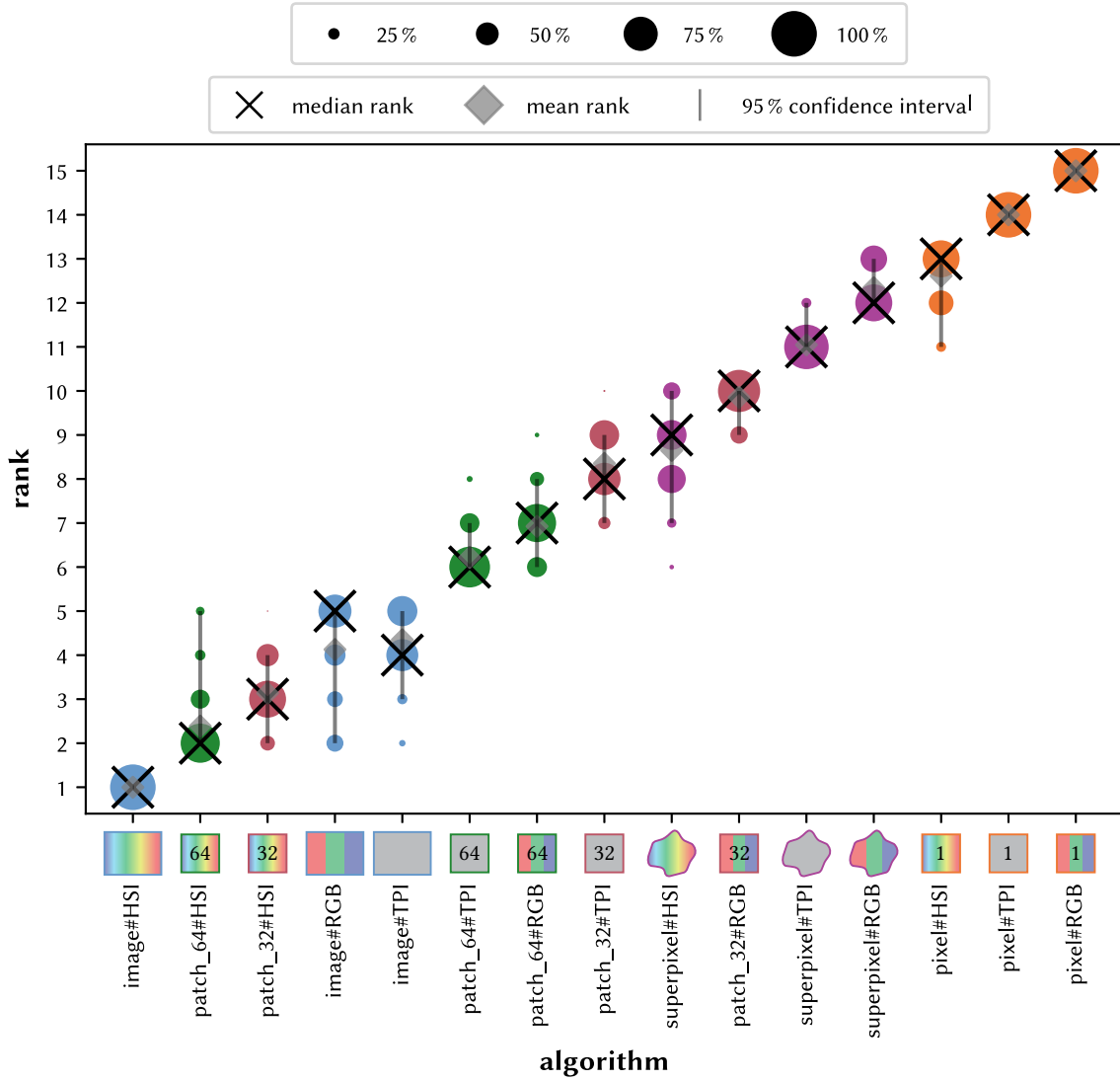
Remarkably, the performance of the image-based HSI segmentation model is consistently comparable to the predictions of a second human expert for all validation metrics, on average. For the inter-rater annotations, we achieved a DSC of 0.89 (SD 0.07), an NSD of 0.80 (SD 0.08), and an ASD of 4.88 (SD 5.33). The intra-rater annotations are better on all three metrics with a DSC of 0.91 (SD 0.05), an NSD of 0.82 (SD 0.06), and an ASD of 4.74 (SD 5.04). Across all 20 images, there were 8 instances in the inter-rater and 6 instances in the intra-rater agreement evaluation where classes not annotated in the reference segmentation map were newly assigned to an image. In the inter-rater and intra-rater agreement evaluation, classes annotated in the reference segmentation map were missing in the re-annotations 7 and 4 times, respectively. Differences in the *ignore* class occurred for 14 and 14 out of the 20 images in the inter-rater and intra-rater comparison, respectively. In total, 34 063 px and 37 397 px for the inter-rater and intra-rater case, respectively, the label *ignore* was assigned in the re-annotation but a label had been assigned in the reference annotation or vice-versa.

### Ranking

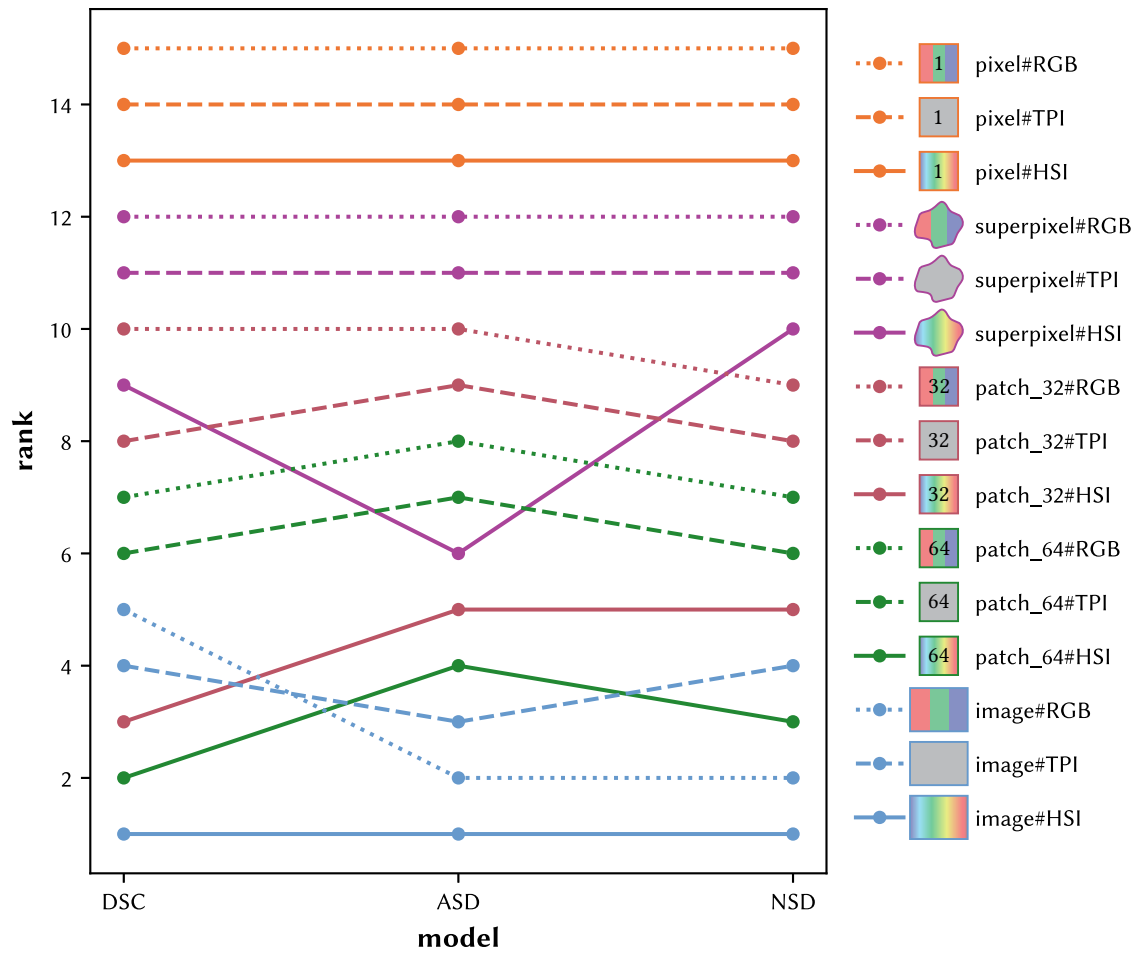
In Figure 5.15, we present the stability of the ranking considering the sampling variability for the DSC (results for the NSD and ASD are available in Figure B.1 and Figure B.2, respectively). The bootstrapped ranking is relatively consistent, with the first and last two ranks being quite distinct (over 90 % of bootstraps resulting in the same rank) for all metrics. In the case of the boundary-distance-based metrics, the number of models with a clear ranking is even higher, and for the NSD, all ranks fluctuate by a maximum of  $\pm 1$  rank around the median.

To evaluate the stability of the ranking regarding different validation metrics, rankings for the three metrics were compared as shown in Figure 5.16. Across all modalities and metrics, the ranking of the spatial granularities is consistently (from best to worst): image, patch\_64, patch\_32, superpixel and pixel. Therefore, more context invariably enhances the segmentation performance regardless of the modality and metric. A model using HSI data consistently ranks higher than the same model using TPI and RGB data in terms of sampling and metric stability. Generally, rankings for the different metrics closely align: The image-based segmentation of HSI data always ranks first, while the last five ranks are consistently occupied by superpixel#TPI, superpixel#RGB, pixel#HSI, pixel#TPI and pixel#RGB (from best to worst).

The most significant difference in ranking across metrics is observed for the superpixel#HSI model, which achieves rank 6 for the ASD compared to rank 9 and 10 for the DSC and NSD, respectively. This could be due to the ASD metric's sensitivity to boundaries. While the superpixel boundaries align very well with the reference segmentation,



**Figure 5.15:** Uncertainty-aware ranking of the different granularities and modalities based on bootstrap sampling on the test set using the dice similarity coefficient (DSC). The area of each blob is proportional to the relative frequency that the corresponding algorithm achieved the respective rank across 1000 bootstrap samples (concept from [236]). Each bootstrap sample consists of 5 hierarchically aggregated subject-level DSC metric values. The lines encompass the 95 % quartile of the bootstrap results while the cross and the diamond denote the median and mean rank, respectively. Ranking results for the normalized surface dice (NSD) and average surface distance (ASD) can be found in Figure B.1 and Figure B.2, respectively. This figure was adapted from [198].



**Figure 5.16:** Ranking stability for the different granularities and modalities across the three metrics dice similarity coefficient (DSC), normalized surface dice (NSD) and average surface distance (ASD) on the test set. Each line visualizes how the ranking of an algorithm changes when different metrics are used. This figure was adapted from [198].

with an average lower limit for the ASD of 2.91 (SD 0.74) if all superpixels were correctly classified (cf. Figure 6.5), we observed from the sample predictions in Figure 5.19 that sharp vertical and horizontal edges can appear in the patch-based predictions. These are a result of our chosen aggregation method, where an image segmentation prediction is composed of non-overlapping patches. The resulting incomplete and scattered boundaries are particularly penalized by boundary-distance metrics such as the ASD while well-matching boundaries are rewarded (refer to Figure 5.16).

### Misclassifications

Figure 5.17 presents the confusion matrix for the best model (image model on the HSI modality) evaluated on the test set. For 8 out of the 19 classes, on average, more than 95 % of the pixels were accurately identified. The recall was lowest for *major vein*, with only 57.1 % of the pixels correctly identified. Confusion matrices for the image model on the TPI and RGB modalities are displayed in Figure B.3 and Figure B.4, respectively. For a direct comparison of the label-specific performance between the three modalities on the image model, Figure 5.18 illustrates the recall stratified by label and modality. Image-based segmentation of HSI data performs better or is on par with TPI and RGB data for all classes with the exception of the *pancreas* and *major vein*.

The largest confusion of 32.2 % occurs between *peritoneum* and *major vein*. This can be attributed to the proximity of these two organs and the limited training data available for *major vein*, as it was only imaged in 32 images (cf. Figure 4.4) and the visible parts are generally small (cf. Figure 4.5), averaging 4192 px (SD 3621 px). Other classes that are often misclassified are either difficult to annotate due to indistinct boundaries (e.g., *omentum*, *peritoneum*, *subcutaneous fat*) or unclear distinction (e.g., *kidney with Gerota's fascia* and *peritoneum*). Most of the misclassifications in the confusion matrix occur between classes that are located next to each other in the images (e.g., *stomach* instead of *omentum* and vice versa, *heart* instead of *lung* and vice versa, *liver* instead of *gallbladder*, *background* instead of *skin*, etc.) which could be due to errors in the predicted segmentation boundaries. This hypothesis is supported by the segmentation examples in Figure 5.19.

### Example Predictions

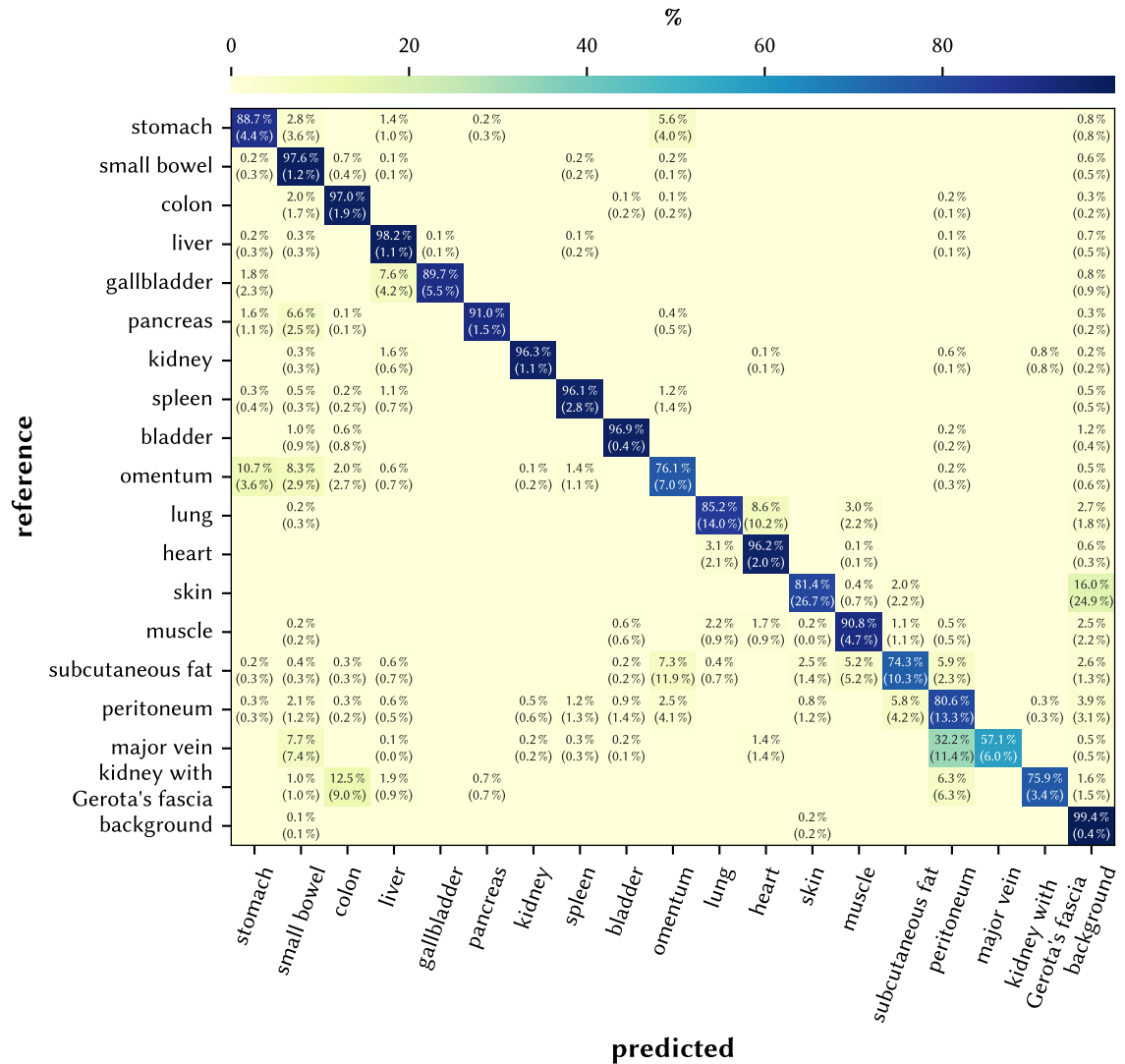
Figure 5.19 shows example predictions for the five spatial granularities on the HSI data. Images corresponding to the 5 % quantile, 50 % quantile, and 95 % quantile were chosen based on the image DSC averaged across all five models, representing examples of poor, intermediate, and good segmentation performances, respectively. For pixel-based segmentation, boundaries appear more fragmented and dispersed compared to other models since no contextual information is included. In some patch-based segmentation examples, sharp vertical and horizontal edges can be seen where adjacent patches meet since we explicitly designed the patches to be non-overlapping during inference (this point is also picked up in Section 6.3). For superpixel-based segmentation, edges appear jagged due to misclassified superpixels near organ boundaries.

### Training Course

We trained each of our models for 100 epochs and calculated validation scores after each training epoch. In Figure 5.30, we can see how the average DSC changes over training time for all spatial granularities and modalities.

There are notable differences among modalities: the DSC exhibited a relatively smooth change for TPI and RGB data, while the training for HSI data was more noisy. This is particularly true for the pixel model, while the image model's training was only marginally

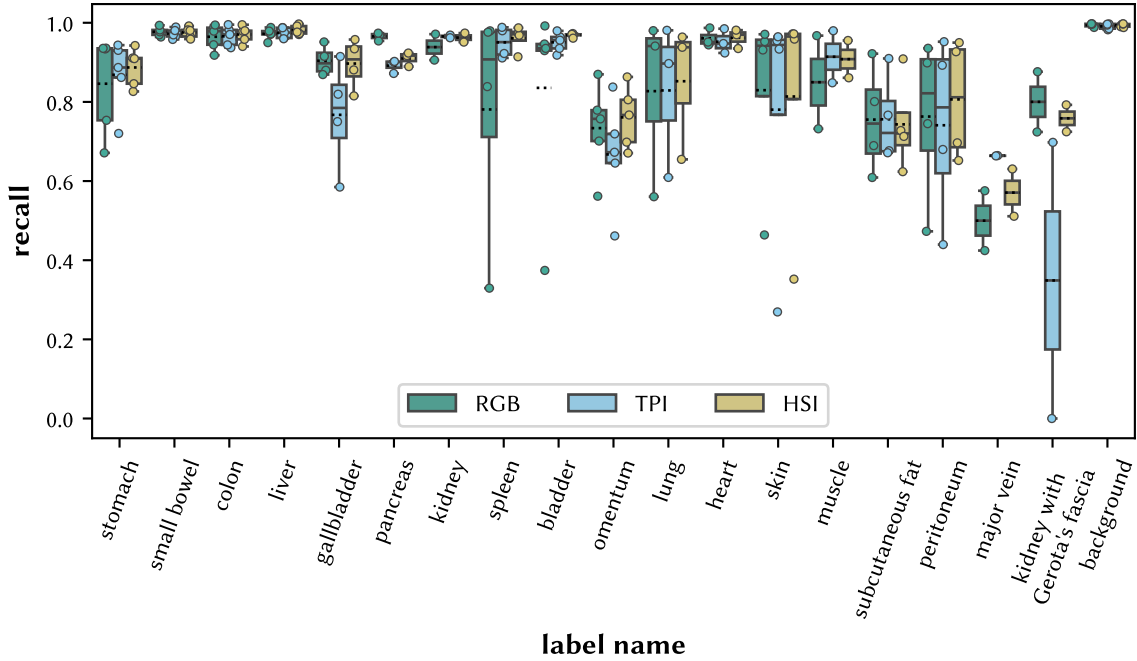
## 5 Experiments and Results



**Figure 5.17:** Confusion matrix of the image granularity and hyperspectral imaging (HSI) modality on the test set. The matrix depicts how pixels from the reference class get classified. That is, every  $(i, j)$ -th entry shows the percentage of pixels from class  $i$  that get classified as class  $j$  (on average). Values  $< 0.1\%$  are not shown for brevity. The matrix is row-normalized based on the pixels from all images of one subject and then these matrices are averaged across subjects. The number in brackets denotes the standard deviation across subjects. Numbers on the diagonal denote the recall (sensitivity). Confusion matrices for the tissue parameter images (TPI) and RGB modality can be found in Figure B.3 and Figure B.4, respectively. This figure was adapted from [198].

noisier. This may be due to the larger input feature dimension of HSI data (100 channels) compared to the other modalities (4 and 3 channels for TPI and RGB data, respectively).





**Figure 5.18:** Recall of the image model stratified by label and modality. Scores are based on the test set and shown for the hyperspectral imaging (HSI), tissue parameter images (TPI) and RGB modalities. Each boxplot shows the interquartile range (IQR) with the median (solid line) and mean (dotted line). The whiskers extend up to 1.5 times of the IQR. Each point represents the performance of one test subject. This figure was adapted from [198].

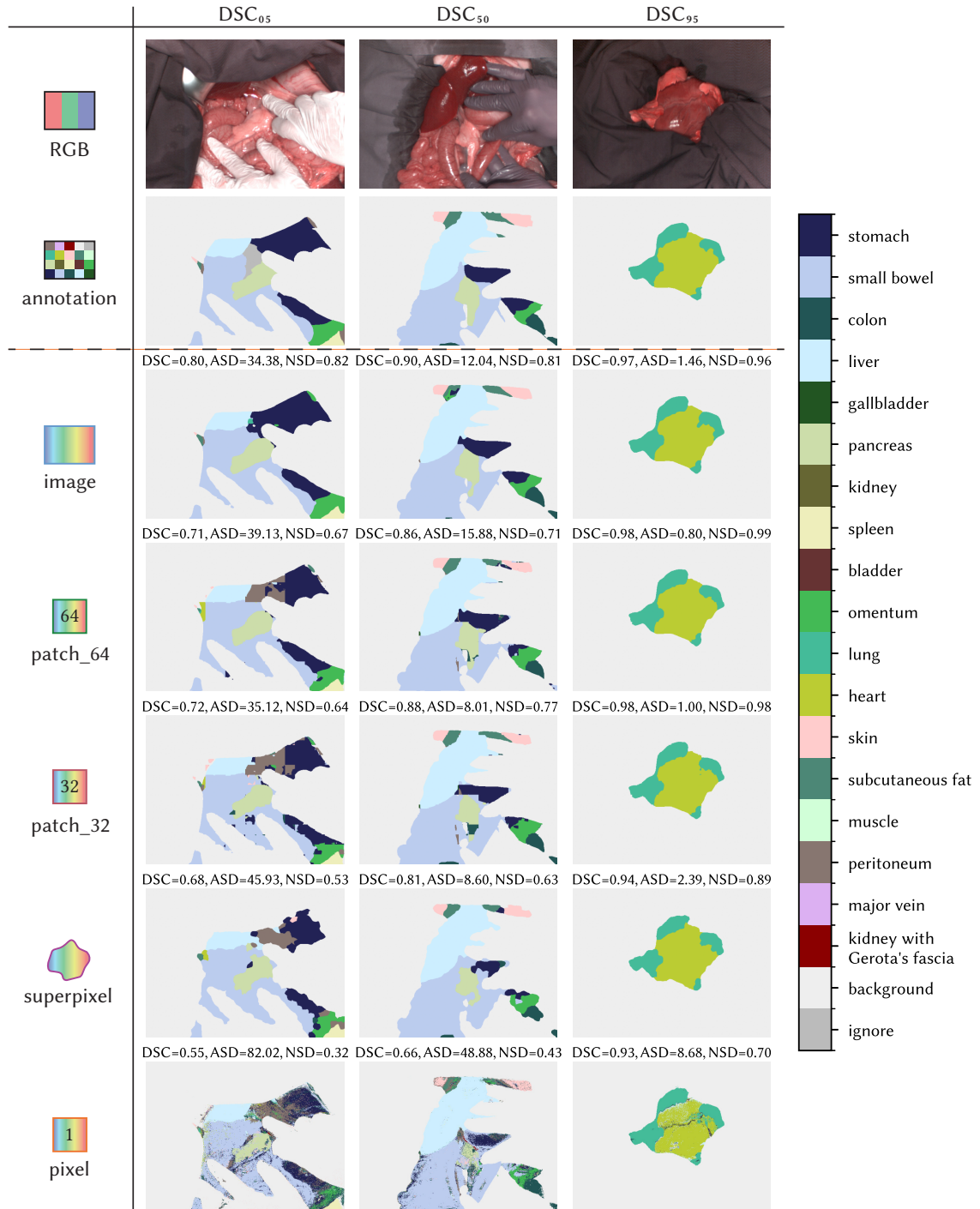
Furthermore, training converged more quickly for the TPI and RGB modalities, whereas HSI gained more from extended training durations.

### Training Size Experiment

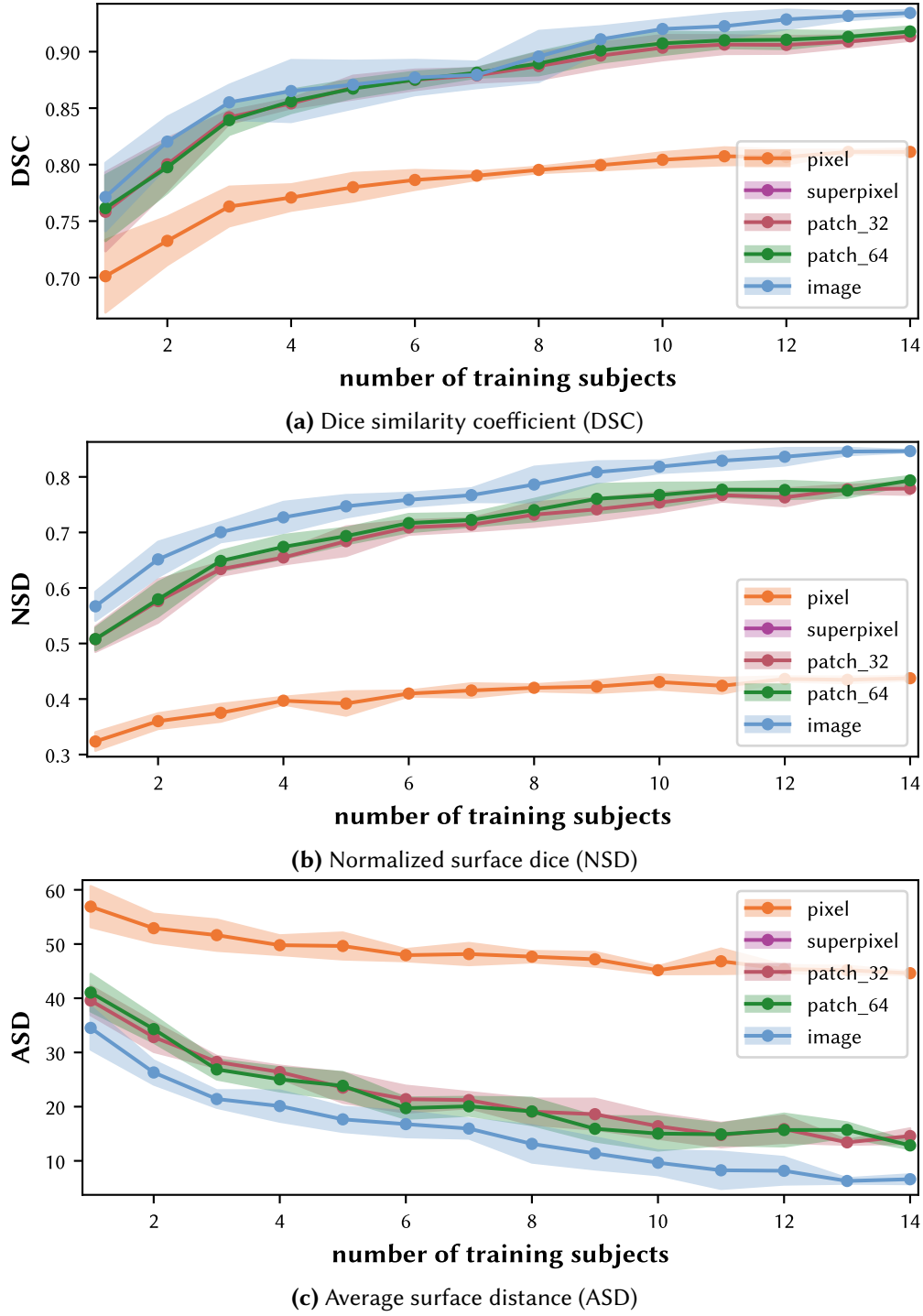
One potential advantage of using input data with smaller spatial granularity is the availability of more training samples. For instance, one image equates to 307 200 training samples for pixel-based segmentation but only one training sample for image-based segmentation (Table 4.3). Figure 5.20 illustrates the progression of the different metrics with the number of training subjects for different spatial granularities. For all examined numbers of training subjects, the performance of image-based segmentation on HSI is either comparable or superior to the performance of other granularities.

We can observe a decreased SD range of the metrics with an increasing number of training subjects. However, this should be cautiously interpreted because subjects were always sampled without replacement which inevitably increases the overlap of selected subjects across random selections in runs with a higher number of training subjects due to the limited number of 15 available training subjects. For instance, when randomly selecting

## 5 Experiments and Results



**Figure 5.19:** Example predictions for the different spatial granularities of the hyperspectral imaging (HSI) modality. For each prediction, scores for the dice similarity coefficient (DSC), average surface distance (ASD) and normalized surface dice (NSD) are shown. Images are selected based on the  $q\%$  quantile of the DSC averaged across all five granularities ( $DSC_q$ ). This figure was adapted from [198].



**Figure 5.20:** Segmentation performance for all spatial granularities on the test set as a function of the number of training subjects (hyperspectral imaging modality). The solid line shows the average performance and the shaded area one standard deviation across 5 runs with different selections of subjects. This figure was adapted from [198].

two different sets of subjects (each of size 14) out of the 15 training subjects without replacement, these two sets differ only by one subject.

### Network Variability and the Effect of Ensembling

All the previous results were reported on the test set with ensembling and we always used the same seed during training for minimal variability as detailed in Section 4.4. To gauge the extent of the inherent randomness during training, we retrained the image HSI model five times with different seeds. This allows us to analyze the controlled source of variation. In this section, these results as well as our analysis are presented in detail and we elaborate on how variability is connected with ensembling.

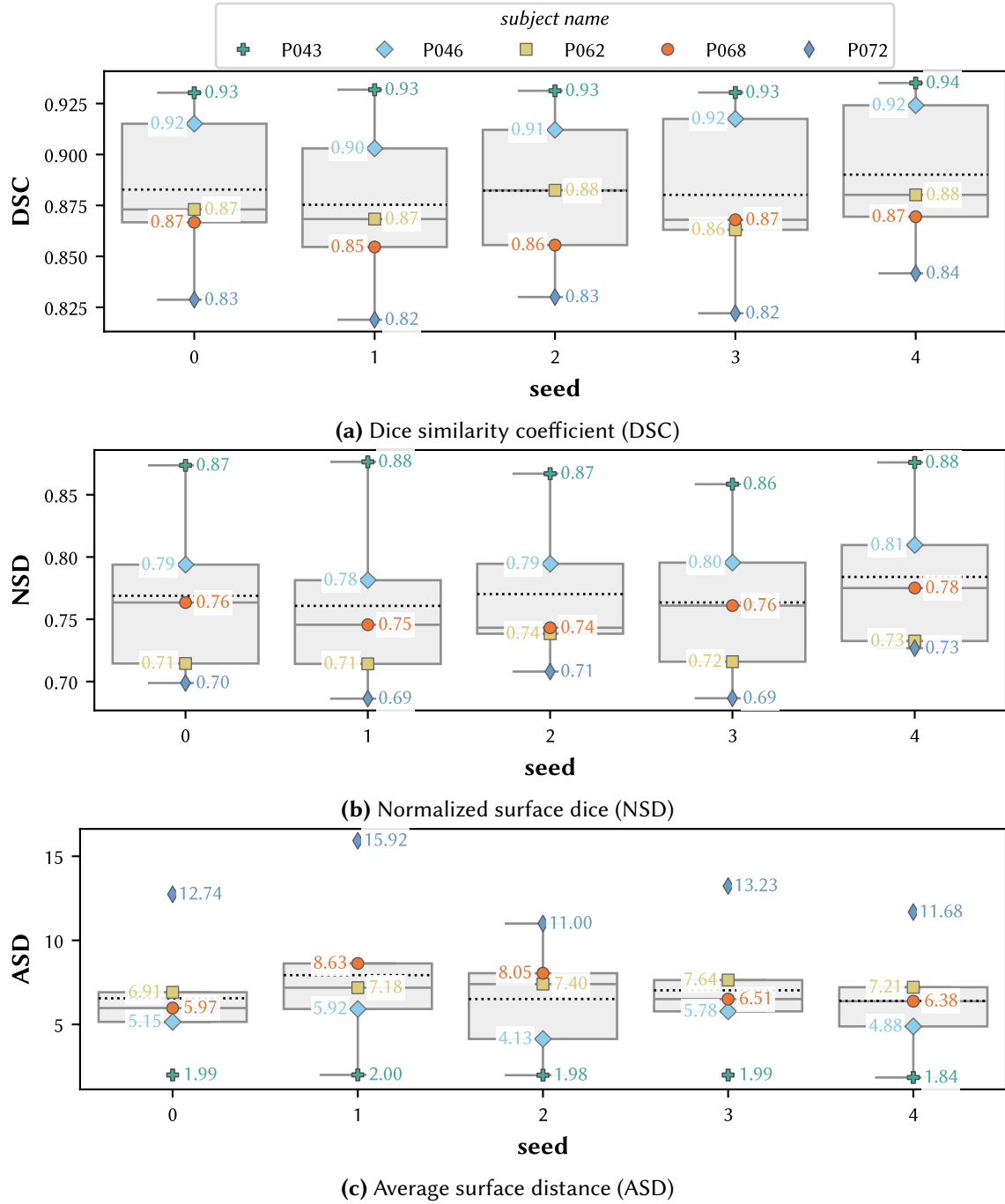
Figure 5.21 shows the distribution of subject-level metric scores of the five seed runs for all metrics. Generally, the variations between the seed runs are low. The DSC was found to range between  $[\min; \max] = [0.88; 0.89]$  (SD 0.01), the NSD was within  $[0.76; 0.78]$  (SD 0.01), and the ASD was within  $[6.40; 7.93]$  (SD 0.63) on the test set with ensembling. Therefore, the variability inherent in the network training is smaller than the inter-subject variability (based on the SDs) with scores of 0.89 (SD 0.07), 0.80 (SD 0.08) and 4.88 (SD 5.33) for the DSC, NSD and ASD, respectively.

Regarding network variability, the fact that the scores on the test set are based on an ensemble of five networks, each trained on individual folds (refer to Section 5.3.1 for details), is important. We created this ensemble by averaging the individual softmax predictions which naturally reduces the networks' variability [59, 175]. In contrast, the predictions on the validation set are not averaged but are taken "as is" from the networks. However, the validation and test splits are, by definition, based on disjoint parts of the dataset so a comparison of those two splits alone is only an indirect indication of the ensemble effect. Therefore, we compare the performance for each organ not only on the test and validation splits but also the scores on the test set for each fold network individually, i.e., without ensembling. Figure 5.22, Figure B.5 and Figure B.6 show distributions of the organ scores and Figure 5.23, Figure B.7 and Figure B.8 list the min-max ranges<sup>8</sup> for the DSC, NSD and ASD, respectively.

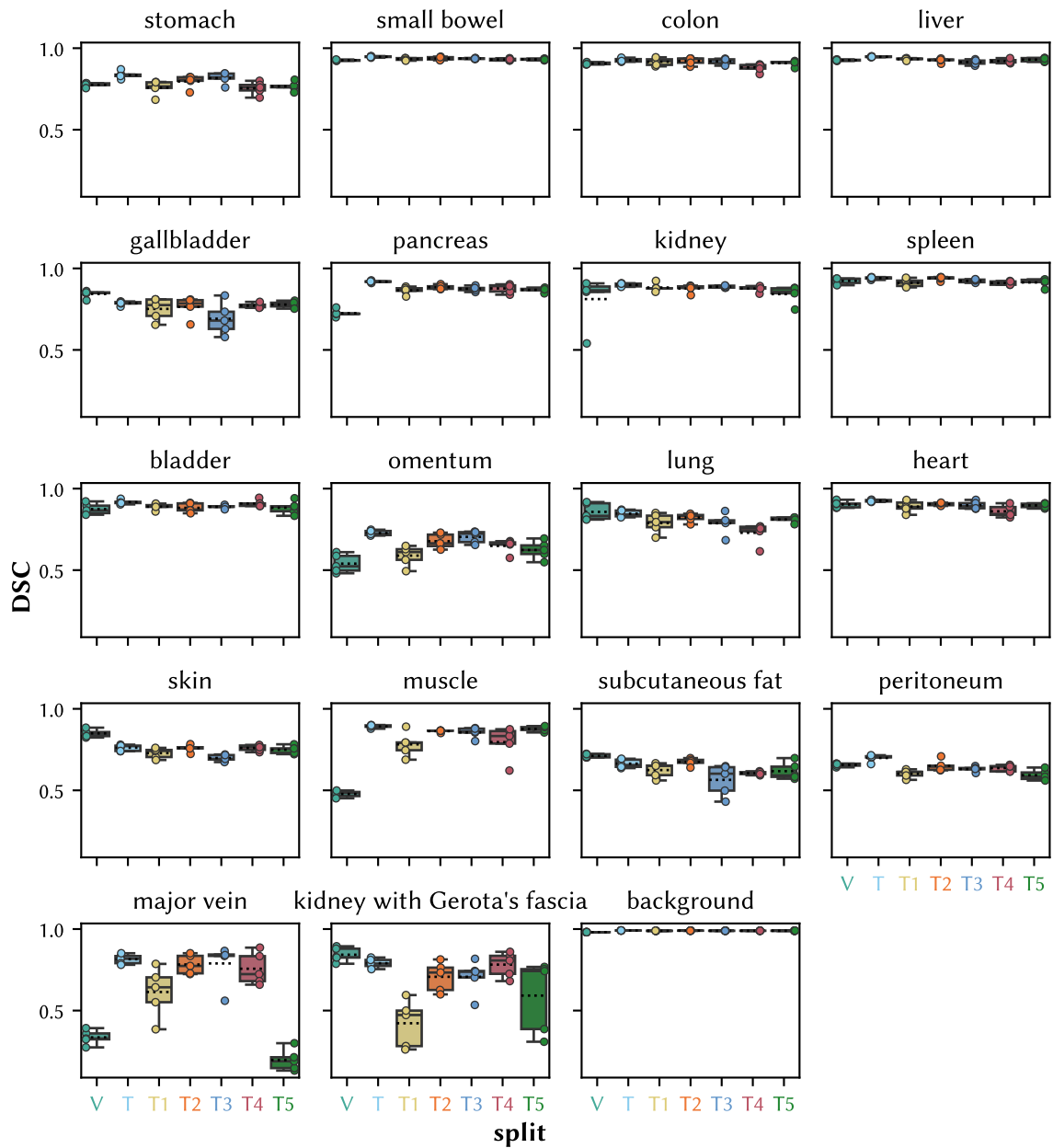
We observe that the variability is generally higher on the test set without ensembling (T1 to T5) than with ensembling (T) but also the variability on the validation set (V) is higher than on the test set with ensembling. For all metrics, the average min-max range across classes is always lowest for the test set with ensembling (e.g., the last line in Figure 5.23 for the DSC). This indicates that the ensemble effect is indeed reducing the networks' variability.

---

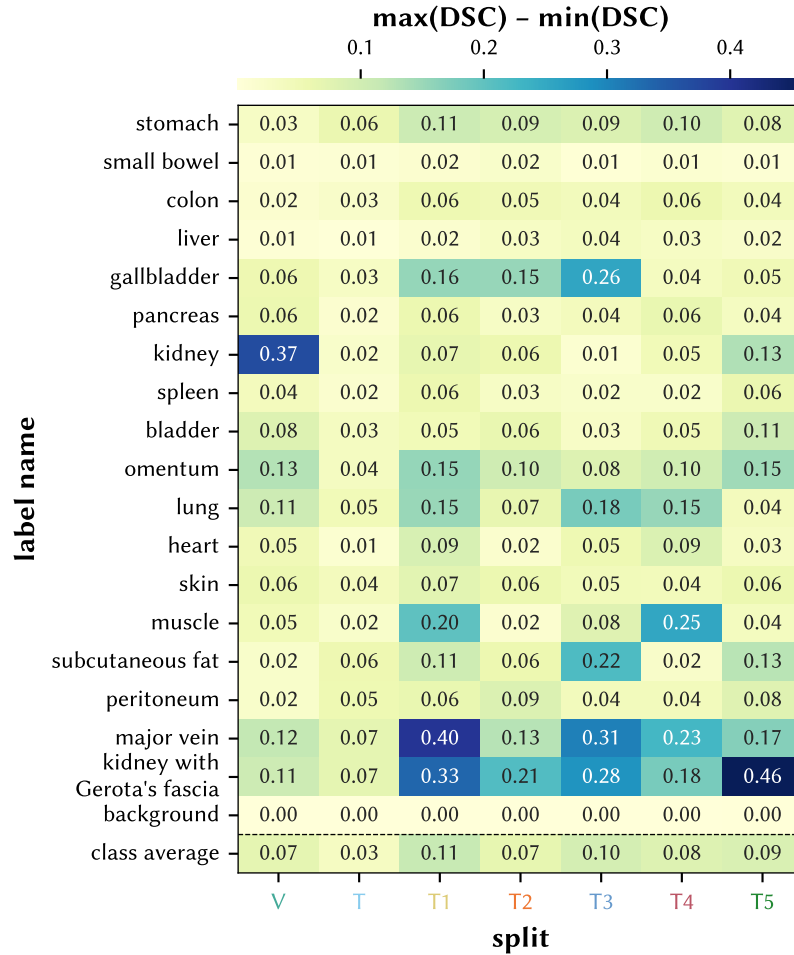
<sup>8</sup>We are considering the min-max range here instead of the SD because when comparing the performance of different networks (e.g., due to different hyperparameter settings), every outlier matters. Comparisons based on the distribution of performance scores (via different seeds) may be costly and hence are not always feasible.



**Figure 5.21:** Network variability across five different seed runs on the test set. The hyperspectral image model was trained five times with different seeds for an estimation of the variability inherent in the training process (e.g., due to different weight initialization, batch sampling, etc.). Each boxplot shows the interquartile range (IQR) with the median (solid line) and mean (dotted line). The whiskers extend up to 1.5 times of the IQR. Each point represents the performance of one test subject.



**Figure 5.22:** Network variability across five different seed runs (hyperspectral image model) stratified by organ for different splits using the dice similarity coefficient (DSC). **V** refers to the validation scores (validation\_unknown split), **T** to the test scores with ensembling and **T1** to **T5** to the test scores without ensembling for each of the networks from the five folds of Figure 5.12. Each boxplot shows the interquartile range (IQR) with the median (solid line) and mean (dotted line). The whiskers extend up to 1.5 times of the IQR. Each point represents the aggregated class-level DSC score of one seed run. Figure B.5 and Figure B.6 show the results for the normalized surface dice (NSD) and average surface distance (ASD), respectively.



**Figure 5.23:** Min-max ranges across five different seed runs stratified by organ for different splits using the dice similarity coefficient (DSC). **V** refers to the validation scores (validation\_unknown split), **T** to the test scores with ensembling and **T1** to **T5** to the test scores without ensembling for each of the networks from the five folds of Figure 5.12. The last line denotes the average across all classes per split. The hyperspectral image model was trained five times and the difference between the highest and lowest DSC score across the five runs is computed independently for each of the splits. Figure B.7 and Figure B.8 show the results for the normalized surface dice (NSD) and average surface distance (ASD), respectively.

The variability on the test set without ensembling tends to be higher than on the validation set across all metrics. This could be due to the different subjects in those sets. However, it is worth noting that the selection of subjects is random so it is possible that this effect could be reversed with a different selection of subjects.

Moreover, the results vary significantly across different organs. Some organs exhibit a very low variability across all splits (e.g., *small bowel*), other organs have a high variability in general (e.g., *major vein*) and there are organs with a high variability on splits without ensembling but a low variability with ensembling (e.g., *gallbladder*).

The trends are also not consistent across metrics. For instance, *kidney* has a high variability across all splits for the NSD but not for the DSC. Compared to the other splits, the variability for *skin* is high for the NSD on the test set with ensembling but low for the DSC. Overall, the average variability across all organs and splits is 0.08 (SD 0.08) for the DSC, 0.09 (SD 0.06) for the NSD and 9.79 (SD 14.04) for the ASD.

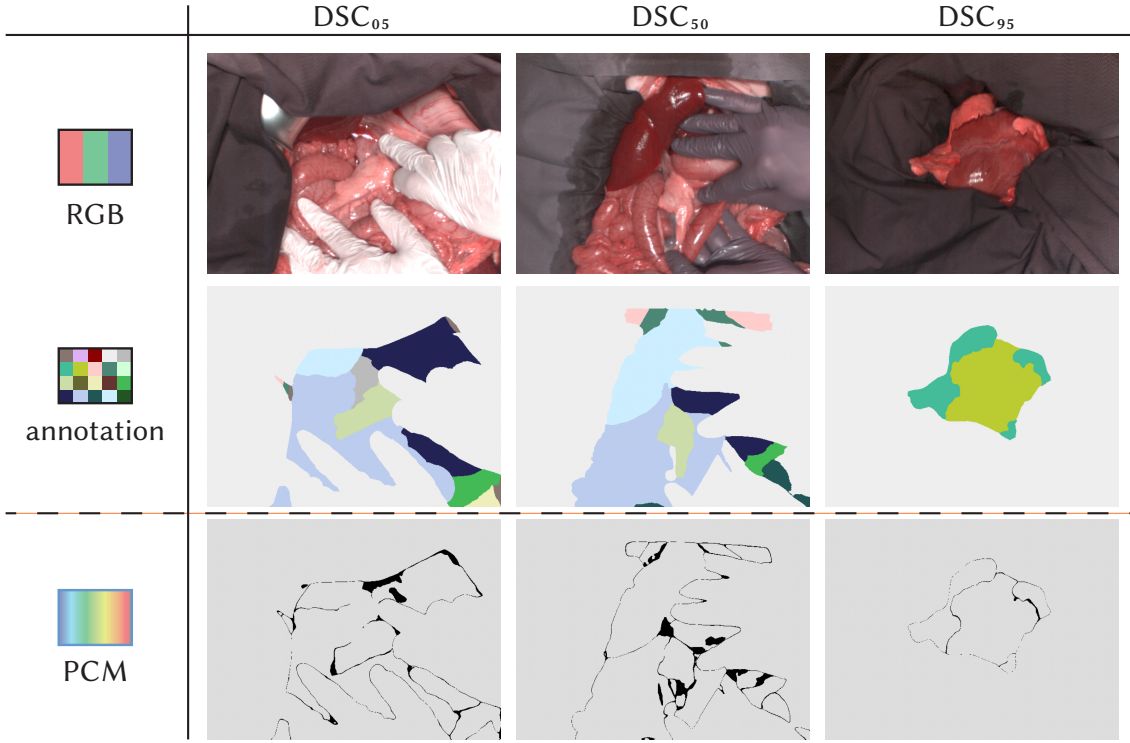
Generally, the variability of our networks without ensembling is quite high which can lead to extreme min-max ranges like 0.37 DSC on the validation split for the *kidney* class. The high variability on the validation set is especially important as this is the only split evaluated during development (the test set is only considered at the end after all model decisions have been made). This is particularly relevant when comparing the performance of different methods as it can be challenging to determine whether a change is due to modifications of the method or merely a result of the inherent network variability. However, as our results demonstrate, at least the test set with ensembling does not suffer from such high variabilities.

Further, it is worth noting that the test scores with ensembling generally tend to be higher than validation scores or test scores without ensembling emphasizing the importance of our fold ensembling. This is true for many organs, such as the *major vein*, but it is not a consistent trend across all metrics. For instance, organs like the *skin* or *gallbladder* have higher DSC and NSD scores on the validation split than on all the testing splits. However, when it comes to the ASD, test scores are almost always superior to validation scores.

Of special interest are the pixels where at least one of the five networks disagrees. For this, we computed PCMs (Equation 4.4) and show examples for three images from the test set in Figure 5.24. We can see that the networks agree on most pixels and only disagree on a few pixels and those are often located in the vicinity of segmentation boundaries. On average across all images in the test set, 3.20 % (SD 2.18 %) of the pixels in an image have incoherent predictions.

Boundary pixels are an interesting case as we also received feedback from our expert annotators that it is very hard to draw the boundary between classes. This is amplified by the fact that the spatial image resolution is with  $480 \times 640$  (width, height) rather low and the reconstructed RGB images, which are used for annotating, do not exhibit the

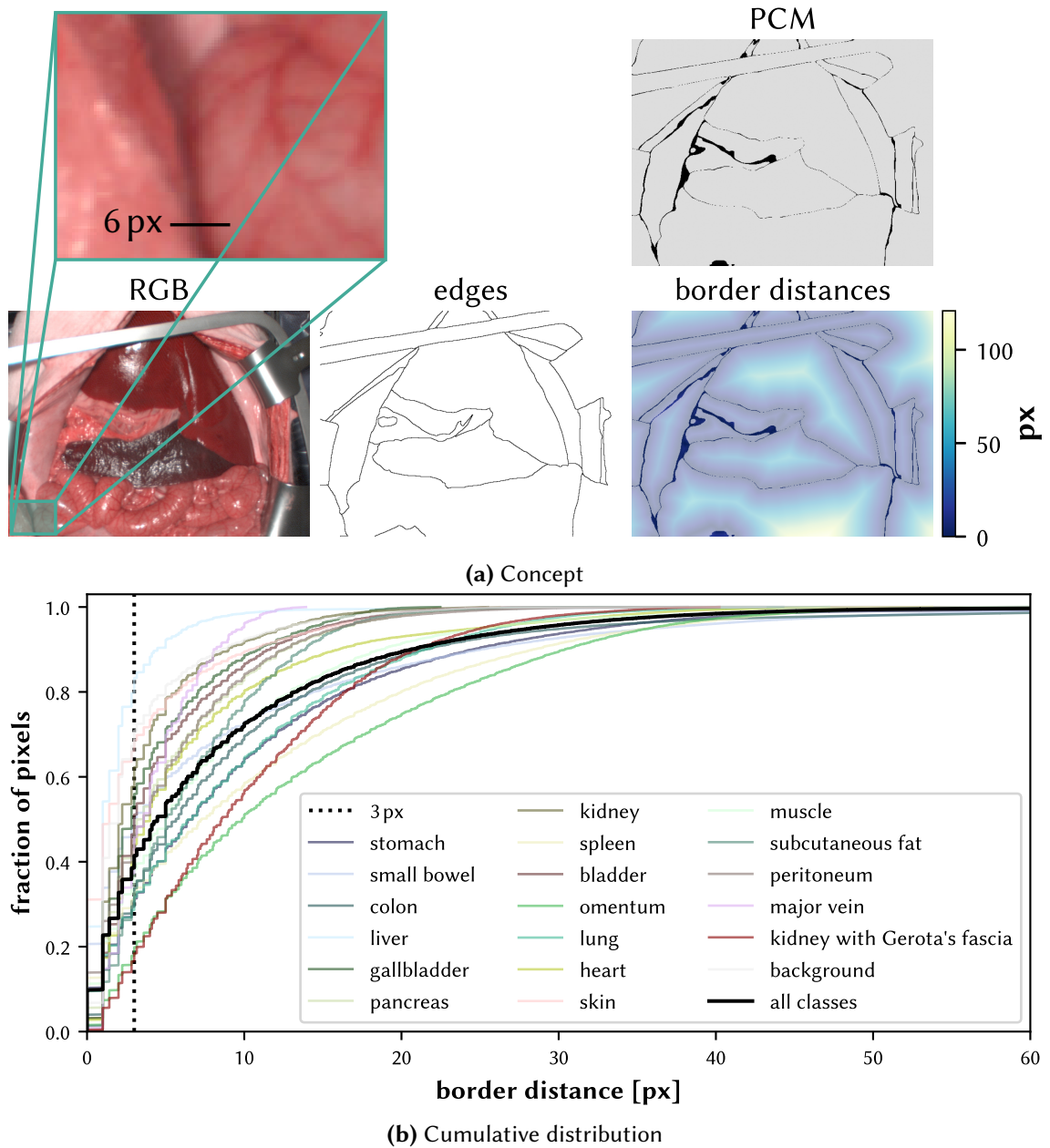




**Figure 5.24:** Example prediction coherence maps (PCMs) of five different seed runs. For the same example images as shown in Figure 5.19, network PCMs (Equation 4.4) are shown in the last row. Gray indicates that the hyperspectral image networks predicted the same label and black denotes pixels with different predicted labels, i.e., where at least one of the networks disagrees. The segmentation colorbar (second row) is the same as in Figure 5.19.

same sharpness as natural images. Hence, mispredictions near the boundary would not be surprising.

To analyze the boundary effect further, we computed the distance for each incoherent pixel to the nearest segmentation boundary. Figure 5.25 shows the cumulative probability distribution of those pixels stratified by class and for all classes. With a boundary range of, say, 6 px (3 px to each side), 39 % of all incoherencies across all classes are explained by boundary pixels. However, the distribution varies heavily among classes. For the same defined boundary range of 3 px, 80 % of all *liver* incoherencies are explained by boundary pixels while only 18 % of all *omentum* incoherencies are explained by boundary pixels. This indicates that the effect of unclear boundaries is more pronounced for some classes than for others which is also in line with the results of Figure 6.3.



**Figure 5.25:** Cumulative distribution of border distances from pixels with incoherent network predictions based on five different seed runs. **(a)** General concept where we compute the distance to the nearest border (defined by the segmentation map) for each pixel in the image and define pixels within the vicinity of 3 px as border pixels, i.e., a maximum range of 6 px across both sides. Only incoherent pixels as defined by the prediction coherence map (PCM) (Equation 4.4) are considered (ignored border distances are shown transparent). **(b)** Cumulative distribution of border distances from all pixels with incoherent predictions based on five different seed runs of the hyperspectral image model stratified by class and for all classes.

## 5.4 Domain Shifts in Surgical Hyperspectral Imaging

If a network is evaluated on data from a different domain than the training domain, the performance may deteriorate due to OOD data. The extent of this deterioration depends on how large this domain shift is, i.e., the extent of OOD for the network and the respective downstream task. The main goal of the experiments in this section is therefore to understand the effect of different domain shifts and assess the generalizability capabilities of our segmentation networks. We are taking a closer look at three fundamental domains: the subject domain (both on the spectra and on the image level), the context domain (impact of neighborhood changes) and the species domain (transfer of our results from pigs to humans) in Section 5.4.2, Section 5.4.3 and Section 5.4.4, respectively. The context domain section also covers the results from our data augmentation approach to tackle geometric domain shifts. Details about the design of our experiments are described in Section 5.4.1.

### 5.4.1 Experimental Setup

The setup of the subject and species domain experiments follows the same procedure as the segmentation experiments (cf. Section 5.3.1). A notable difference is that for the species domain, we aggregate toward class-level scores instead of subject-level scores (cf. Figure 5.13), use the splitting of Figure 5.26 and show only validation results. The remainder of this section describes the details of the context experiments<sup>9</sup>.

#### Validation Strategy for the Context Domain

For our context experiments, we used our existing segmentation network described in Section 4.4 and compared different augmentation schemes with respect to the performance on geometrical OOD scenarios. An overview of our validation strategy is presented in Figure 5.27.

#### Datasets and Splits

To evaluate the performance under geometric domain shifts and to assess the enhancements provided by our augmentation technique, we considered the following geometrical OOD scenarios:

- (I) *Isolated organs*: During surgeries, abdominal linens are often used to safeguard soft tissues and organs, control excessive bleeding and absorb blood and secretions. Certain surgeries (e.g., enteroenterostomy) may necessitate isolating a single organ [235]. In such instances, it is crucial to accurately identify an organ without any contextual information from neighboring organs.

---

<sup>9</sup>This section is based on [202].

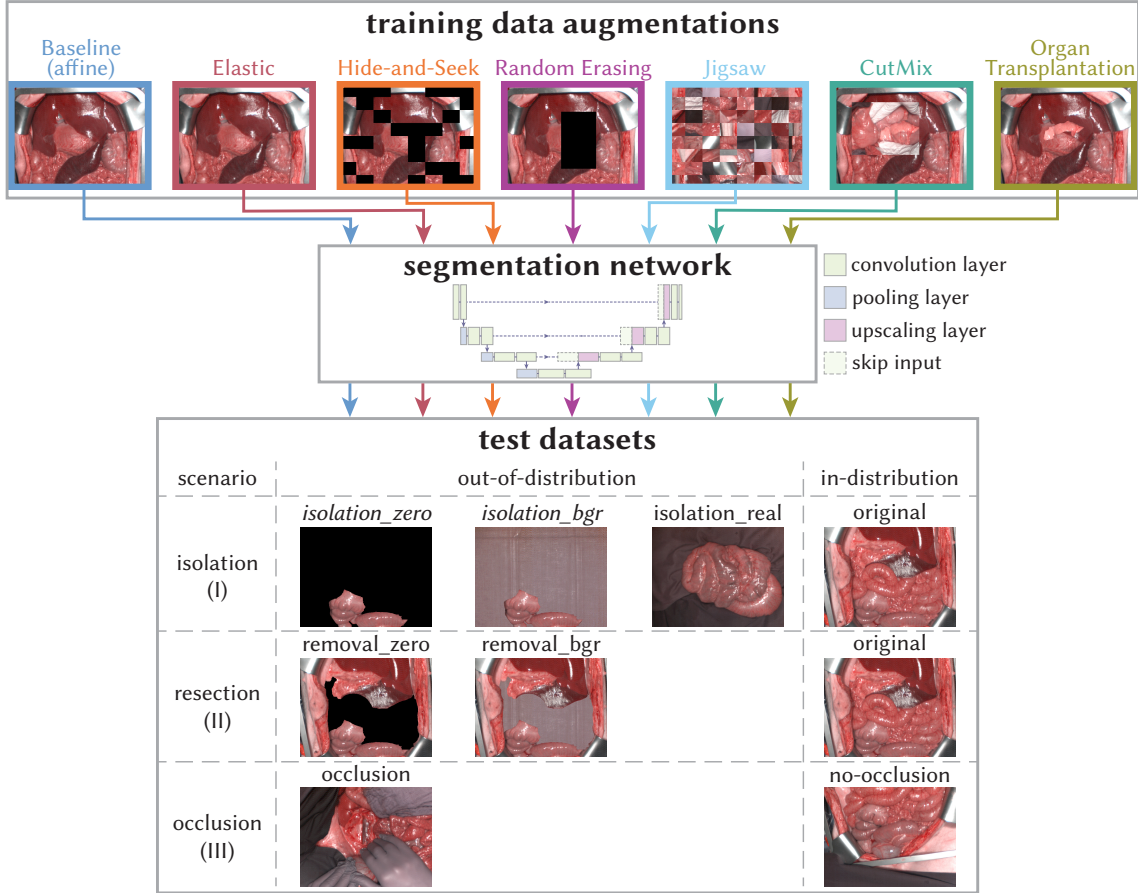


**Figure 5.26:** Overview of the  $k$ -fold structure of the semantic human dataset. The heatmap visualizes the assignment of the images from the semantic human dataset to the different splits used for training, validation and testing (each row denotes one fold and each column one image). Validation and test borders are always at subject boundaries. A  $k = 5$  cross-validation structure is employed with 40, 36, 36, 33 and 30 subjects in the validation split for the folds fold\_0, fold\_1, fold\_2, fold\_3 and fold\_4, respectively.

- (II) *Organ resections*: In resection procedures, parts or even the whole organ are excised, necessitating the identification of surrounding organs despite the absence of usual neighbors.
- (III) *Occlusions*: Large portions of the surgical site can be obscured by the surgical procedure itself, introducing OOD neighbors (e.g., gloved hands). Despite this, the non-occluded parts of the surgical site must be correctly identified.

The semantic porcine dataset does not contain images of isolated or resected classes. Therefore, we created four new datasets by manipulating existing images in two ways: (1) For each image and class, we removed all pixels that do not belong to this class either by setting all reflectance values to zero (isolation\_zero) or by replacing them with spectra from an image which only shows blue cloth (isolation\_bgr) and (2) for each image and class, we removed the class either by setting all reflectance values of the class to zero (isolation\_zero) or by replacing them with spectra from an image which only shows blue cloth (isolation\_bgr).

With (1) we estimate how well the segmentation networks perform when confronted with a class without any context and with (2) we analyze the performance of classes if certain neighbors are missing. We are using an image of blue cloth as a proxy for the background class (which often contains blue cloth) to study the effect of the replaced values since *background* is a class known during training but the network never saw reflectance values which only contained zero values.



**Figure 5.27:** Validation strategy for the geometrical out-of-distribution (OOD) experiments. The context augmentations are evaluated on three different OOD scenarios (isolation, resection and occlusion) comprising six different datasets in total. Four of the datasets are manipulated versions of the semantic dataset (*isolation\_zero*, *isolation\_bgr*, *removal\_zero* and *removal\_bgr*) and the remaining two show real-world OOD images (*isolation\_real* and *occlusion*). The validation\_unknown split (cf. Figure 5.12) of the two italic shaped datasets (*isolation\_zero* and *isolation\_bgr*) served as validation datasets during method development while the other datasets were used as untouched test sets. This figure was adapted from [202].

In addition to our manipulated datasets, we also included a real-world isolation dataset (*isolation\_real*) which is based on 94 images from 25 subjects of the tissue atlas dataset showing classes in isolation. These images are also semantically annotated.

For the occlusion scenario, we are also using the semantic porcine dataset but separated it in a different way: we are using the same splits as in Figure 5.12 but we removed all images that contain occlusions (e.g., gloves) from the training data of the folds constituting the no-occlusion dataset. From the 340 training images, 271 images remain after this step.

Similarly, the test split was separated into an occlusion dataset (73 images from 4 subjects) with images that contain occlusions and a no-occlusion dataset (93 images from 5 subjects) of images without occlusions.

During development, we only used the validation splits of the original, isolation\_zero and isolation\_bgr datasets. We included the original dataset to ensure that we maintain our segmentation performance on in-distribution images. All remaining datasets (manipulated and real) served as hold-out test sets and we evaluated the performance on those datasets only after we finalized our method. Additionally, we also evaluated the performance on the test splits of the original, isolation\_zero and isolation\_bgr datasets in the end.

An example image for each dataset is shown in Figure 5.27. The usage of the different datasets for training, validation and testing, whether it is a real or manipulated dataset and whether it is an in-distribution or OOD dataset is summarized in Table 5.1.

**Table 5.1:** Distribution of training (second column), validation (third column) and test (fourth column) datasets across our three geometric out-of-distribution (OOD) scenarios (first column). The training was solely performed on real data (r), whereas OOD validation was only performed on manipulated datasets (m). Testing was performed on both in-distribution (✗) and OOD data (✓). If the same dataset appears for training, validation or testing, they were always used with different splits without subject-level overlap (see Figure 5.12 for details).

| scenario         | training     | validation     | testing        | type | OOD? |
|------------------|--------------|----------------|----------------|------|------|
| isolations (I)   | original     | original       | original       | r    | ✗    |
| isolations (I)   | original     | isolation_zero | isolation_zero | m    | ✓    |
| isolations (I)   | original     | isolation_bgr  | isolation_bgr  | m    | ✓    |
| isolations (I)   | original     | —              | isolation_real | r    | ✓    |
| resections (II)  | original     | —              | original       | r    | ✗    |
| resections (II)  | original     | —              | removal_zero   | m    | ✓    |
| resections (II)  | original     | —              | removal_bgr    | m    | ✓    |
| occlusions (III) | no-occlusion | —              | no-occlusion   | r    | ✗    |
| occlusions (III) | no-occlusion | —              | occlusion      | r    | ✓    |

### Augmentations

We compared the baseline segmentation network (cf. Section 4.4) with our organ transplantation augmentation described in Section 4.5 together with five additional topology-aware augmentations that could potentially improve the performance on geometrical OOD images. All augmentations are visualized in Figure 5.27.

As additional augmentations, we included the noise augmentations Hide-and-Seek [207] and Random Erasing [253] which black out parts of the image either based on a grid of

patches or a rectangular region, thereby generating artificial class occlusions. Rather than obscuring, the image-mixing methods Jigsaw [37] and CutMix [247] transfer regions from one part of an image to another via a grid of patches or a rectangular region, respectively. We tailored these image-mixing augmentations to our segmentation task by also duplicating and inserting the corresponding segmentation masks. As a result, in addition to occluding the underlying surgical site, image parts appear in atypical surroundings. Finally, we included an elastic transformation augmentation [206] as a way to distort the local neighborhood in an image.

Augmentations have many hyperparameters to tune. To limit the computational effort for our comparative study, we only tuned the probability  $p$  of applying an augmentation to an image. We optimized each augmentation via a grid search with  $p \in \{0.2, 0.4, 0.6, 0.8, 1\}$ . For our batch size of 5, this corresponds to applying the augmentation to 1, 2, 3, 4 or 5 images per batch. We selected the optimal probability based on the best average DSC score on the validation split of the original, isolation\_zero and isolation\_bgr datasets to ensure that the augmentation works on both in-distribution as well as OOD data. The resulting optimal probability values for each augmentation are listed in Table 5.2

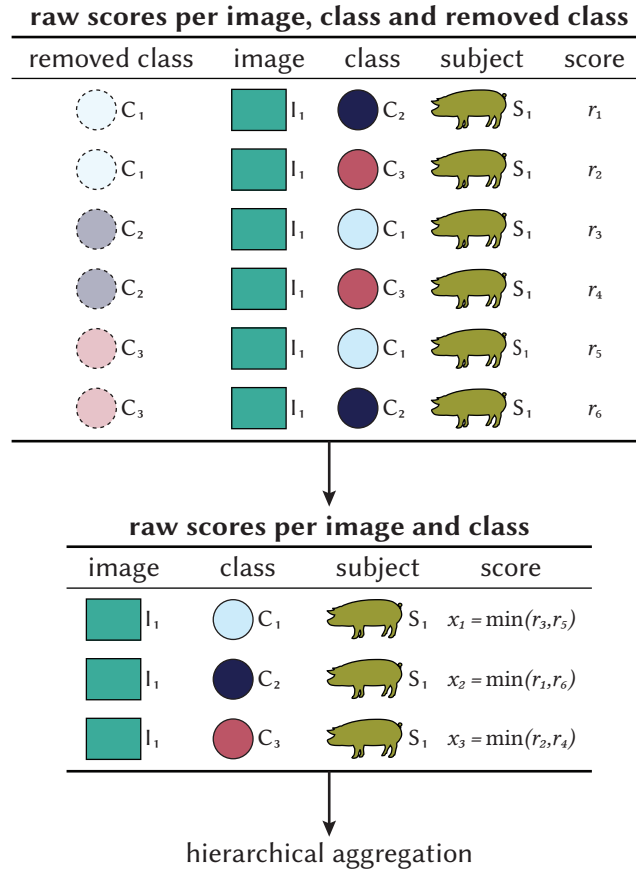
**Table 5.2:** Optimal probability of applying an augmentation based on the grid search results.

| augmentation          | optimal probability $p$ |
|-----------------------|-------------------------|
| elastic               | 0.6                     |
| Hide-and-Seek         | 1                       |
| Random Erasing        | 0.4                     |
| Jigsaw                | 0.8                     |
| CutMix                | 1                       |
| organ transplantation | 0.8                     |

## Metrics

We are using the DSC as an overlap-based measure and the NSD as a boundary-overlap-based measure with special consideration to annotation uncertainty to compare the performance of the different augmentation schemes (similar to our segmentation task, cf. Section 5.3.1). However, we are aggregating toward class-level scores instead of subject-level scores since our scenarios are more targeted at specific classes and we are interested in the change of class-level scores.

In general, when we compare the performance (e.g., in boxplots), one point corresponds to the score of one class. The class removal experiments require special attention since we yield multiple scores per class (one per removed neighbor). We still want to compare the results from the removal experiment with the other experiments on a class level so we take the minimum of all available scores per class corresponding to the segmentation performance when the most important neighbor is missing. This concept is visualized in Figure 5.28.



**Figure 5.28:** Visualization of the aggregation scheme of metric scores for the removal scenario. For each removed neighbor, we yield one score per image and remaining class. To reduce to a final score per class, we take the minimum of all available scores for that class corresponding to the segmentation performance when the most important neighbor is missing. After this step, the usual hierarchical aggregation of Figure 5.13 is applied.

### 5.4.2 Subject Domain

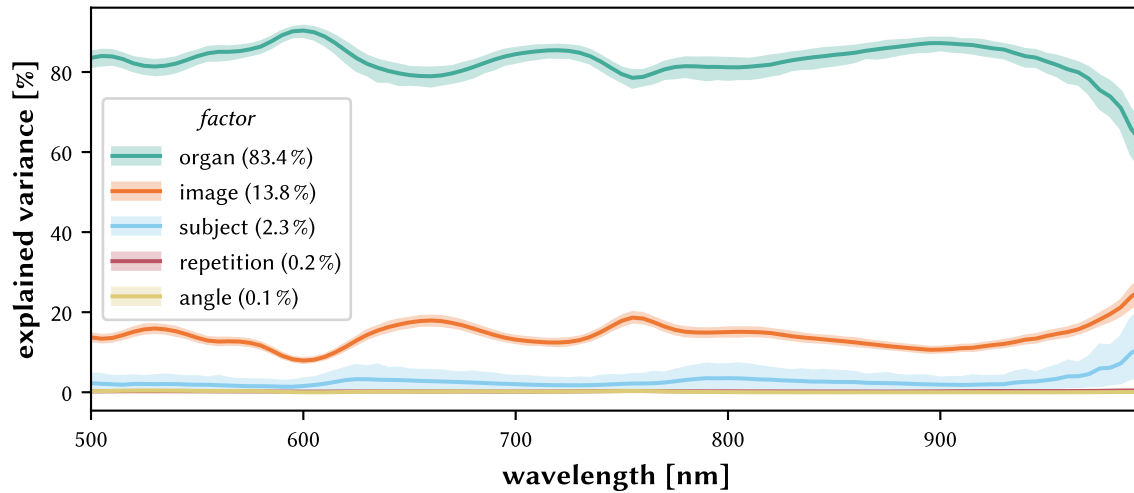
Surgery is a complex and unique process that can vary depending on the individual, the type of surgical procedure or the surgeon who performs it. Therefore, it is crucial to understand how subject-specific variations affect the data that we use for our downstream task of tissue discrimination. In this section, we examine the influence of the subject domain on the spectra and the image level<sup>10</sup>.

<sup>10</sup>This section is based on [198, 215].



### Spectra Level

We analyzed the generalizability of our spectra via the standardized recordings of the tissue atlas dataset and a generalized linear mixed-effect model [196]. The decomposition of the median spectra into the different factors organ, image, subject, repetition and angle (cf. Figure 4.3) is shown in Figure 5.29. The factors image, repetition and angle are not directly related to the subject domain but are included in the analysis for comparison purposes.



**Figure 5.29:** Sources of variation of hyperspectral imaging (HSI) data from the tissue atlas. For each factor, the explained variance in reflectance is computed using a generalized linear mixed-effect model as described in [196] independently for each wavelength. The shaded areas denote the 95 % pointwise confidence interval based on 500 bootstrap samples. The numbers in brackets represent the median across all wavelengths. This figure was adapted from [215].

We can see that the organ factor explains most of the variance in the data. This indicates that the spectral fingerprints are indeed representatives of the different organs and this result is also in agreement with the good classification results of Section 5.1.2. Across spectral channels, the explained variance varies and is lowest (and the confidence interval largest) in the near-infrared range (e.g., in the area above 950 nm) of the recorded spectrum. This coincides with the observation of a higher SD across subjects in this area (cf. Figure 5.2) and could also be attributed to the increased noise there (cf. Figure 5.2 and Figure 2.7).

The image effect (layout of the organs relative to each other) is the second most important factor in the model. This means that different organ surfaces do impact the spectra significantly. This observation could be attributed to inhomogeneous surface structures, blood vessels or fibrosis within each organ.

Different angles and subjects have only a minor impact on the spectra. This indicates that the spectral fingerprints are robust to changes in the measurement angle and across subjects. At least on the spectral level, this indicates that there is good generalizability across subjects which is an important prerequisite for the application of HSI in the operating room.

### Image Level

Even though the average explained variance of individuals on the spectra is with 2.3 % very low, subjects still might have a larger impact on the segmentation performance on the image level with its varying geometries. To get an initial measure of the generalization capabilities, we compared the segmentation performance on the validation set consisting of unseen subjects validation\_unknown with the performance on the validation set made up of unseen images from seen subjects validation\_known (cf. Section 5.3.1).

Figure 5.30 shows the average DSC on validation\_unknown and validation\_known, taking into account the hierarchical structure of the data, for the 5 different levels of spatial granularity across all 100 epochs during training. Performance is generally superior for validation\_known. For all modalities, the performance difference between validation\_known and validation\_unknown is smallest for the pixel-based segmentation which is in accordance with our previous results of Figure 5.29.

We can see that the impact of individuals on the image level is higher than on the spectra level (e.g., 0.10 difference in DSC on average between validation\_unknown and validation\_known for the image HSI model). However, it is worth noting that the semantic porcine dataset is not standardized and contains varying images and situs per subject.

### 5.4.3 Context Domain

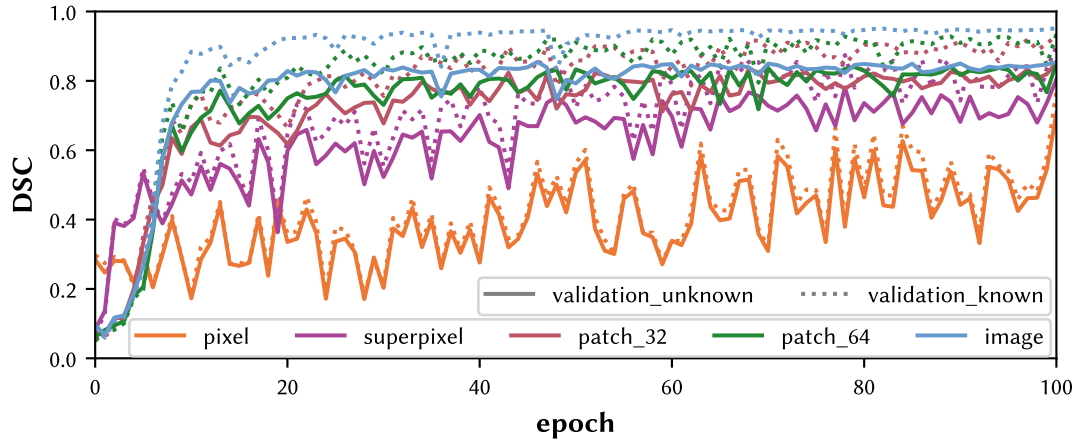
Our segmentation networks are based on two-dimensional convolutional operations and hence are sensitive to changes in the local neighborhood. In our context experiments (described in Section 5.4.1), we analyzed the neighborhood relation of the classes in the semantic porcine dataset and our segmentation performance with respect to changes in the neighborhood of classes<sup>11</sup>. Further, we present the results of our organ transplantation augmentation method introduced in Section 4.5 for improving the segmentation performance and a comparison to other common topology-aware augmentation methods

#### Neighborhood Relation

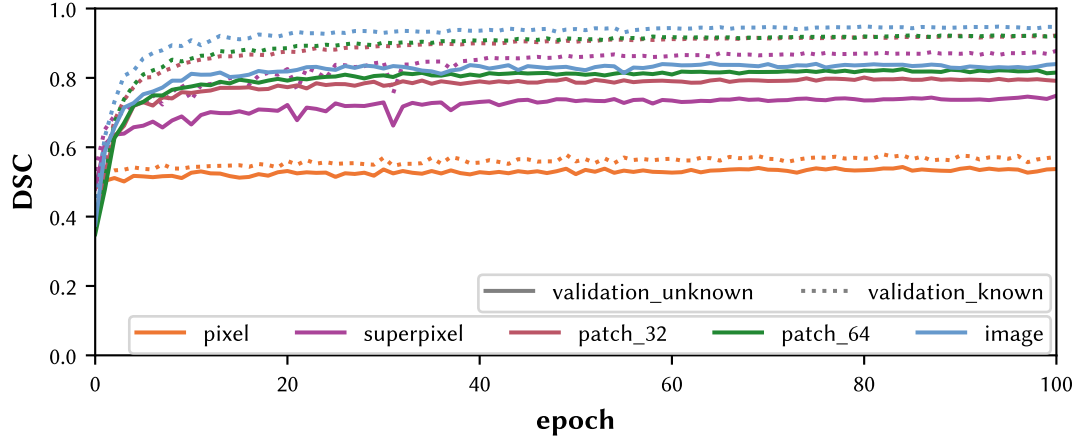
The neighborhood relation of our classes is characterized by the porcine anatomy. However, not every anatomic neighbor is also visible in all images due to occlusions, surgical procedures or resections. To understand the neighborhood relation of the semantic

---

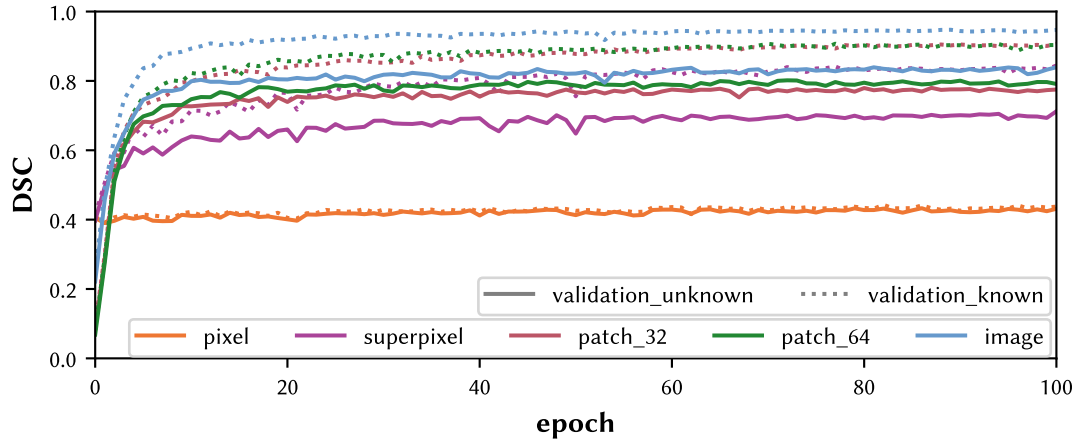
<sup>11</sup>This section is based on [202].



(a) Hyperspectral imaging (HSI)



(b) Tissue parameter images (TPI)



(c) RGB

**Figure 5.30:** Generalization error over training time via comparison of dice similarity coefficient (DSC) values (subject-level averages) from the two validation sets defined in Figure 5.12. This figure was adapted from [198].

porcine dataset, we computed the class neighborhood matrix shown in Figure 5.31 which shows the spatial relation between classes on the test set. Some classes like the *gallbladder* have very dominant neighbors (*liver*) while other classes like the *stomach* have a more balanced neighborhood relation to several classes. Nearly every class has a strong neighborhood relation to the *background* class which is not surprising as the *background* class is the largest class in terms of pixels and comprises abdominal linen, medical instruments or gloves. This motivates our manipulated datasets *isolation\_bgr* and *removal\_bgr* where we explicitly alter the relation to the *background* class.

### Effect of the Neighborhood on the Segmentation Performance

Classes with a common neighbor may be more likely to fail if that neighbor is missing in the image. This effect of removed classes on the segmentation network is shown in Figure 5.32. The performance of the *gallbladder* drops by 63 % in DSC if the *liver* is missing in the image, i.e., if the most common neighbor is missing. Strong negative effects can also be observed for *major vein* (if *peritoneum* is missing), *bladder* (if *background* or *small bowel* are missing) and *kidney with Gerota’s fascia* (if *peritoneum* is missing). However, in the majority of (removed neighbor, affected class) cases, the effect of a missing neighbor is only minor, indicating that missing neighbors mainly impact the segmentation performance of some classes. This is in line with Figure 5.31 which reveals that some organs are never neighbors to each other so the network never learned to rely on this non-existent relationship.

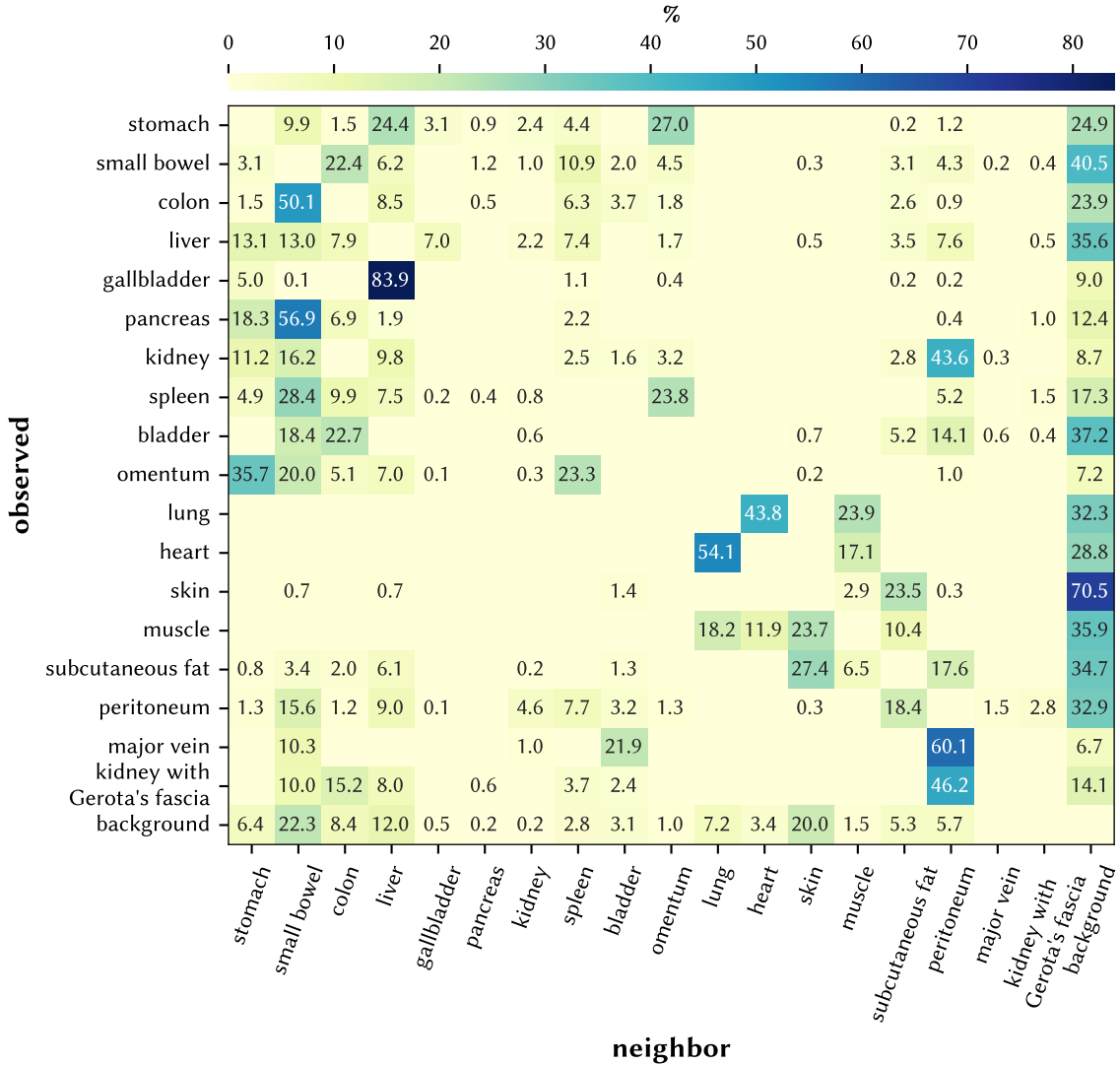
If classes are shown in isolation, the effect on the segmentation network is more severe as shown in Figure 5.33 (baseline performance). Beginning with high DSC values of 0.86 (SD 0.10) for HSI and 0.83 (SD 0.10) for RGB data on in-distribution data, the performance experiences, depending on the dataset, a decrease by 21 %–45 % for HSI and by 30 %–46 % for RGB data.

For occluded classes, the effect on the segmentation performance is less drastic. In this case, the performance when evaluating on the OOD dataset (occlusion) drops by 5 % for HSI and by 10 % for RGB data.

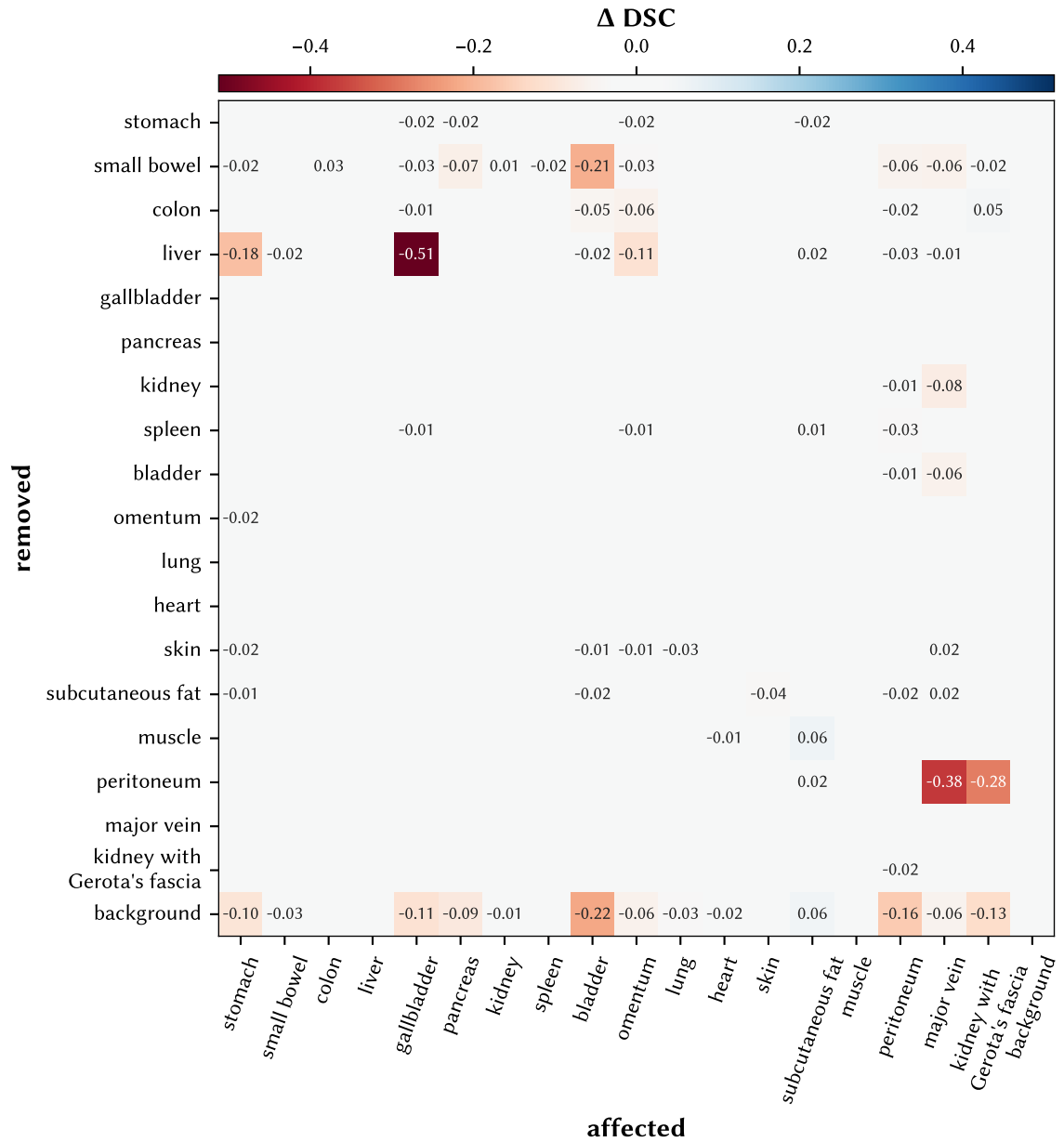
### Organ Transplantation Augmentation

As depicted in Figure 5.33 (DSC) and Figure B.10 (NSD), the organ transplantation augmentation effectively addresses geometric domain shifts for both HSI and RGB modalities. The HSI modality consistently delivers superior results compared to RGB suggesting that spectral information is vital in situations with limited context. The performance enhancement relative to the baseline varies for HSI from 9 %–90 % (DSC) and 16 %–96 % (NSD) and for RGB from 9 %–67 % (DSC) and 15 %–79 % (NSD).

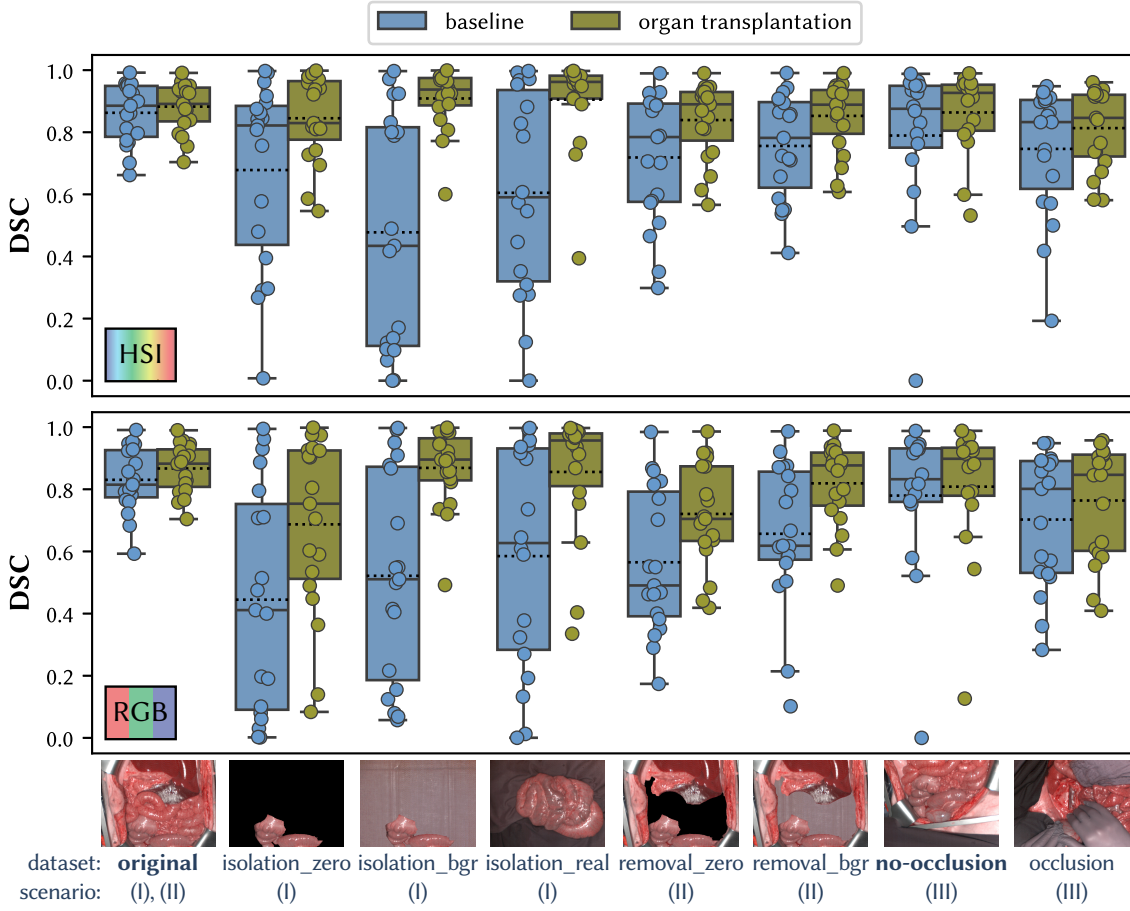
The greatest benefit on OOD data is observed for isolated classes (*isolation\_zero*, *isolation\_bgr* and *isolation\_real*) while the smallest improvement is observed for situs occlusions (occlusion). For isolated organs, the performance enhancement on manipulated data is with a DSC increase by 57 % (HSI) and 61 % (RGB) similar to that on real data



**Figure 5.31:** Class neighborhood matrix for the semantic porcine dataset on the test split. The  $(i, j)$ -th entry shows how many boundary pixels the class  $i$  shares with the class  $j$  (on average). The matrix is row-normalized yielding the percentage of boundary pixels from the observed class that are shared with the other classes. The normalization is based on the neighbor pixel counts from all images of one subject and then these matrices are averaged across subjects. Non-boundary pixels are discarded for this analysis, i.e., the area pixels inside of an organ are not considered. Values  $< 0.1$  are not shown for clarity.



**Figure 5.32:** Change in performance of the image network upon encountering class removals (hyperspectral imaging modality on the removal\_zero dataset). The  $(i, j)$ -th entry shows the change in dice similarity coefficient (DSC) of the  $j$ -th class when the  $i$ -th class is removed from the images. Values  $|\Delta \text{DSC}| < 0.01$  are not shown for clarity. This figure was adapted from [202].



**Figure 5.33:** Segmentation performance using the dice similarity coefficient (DSC) for six geometrical out-of-distribution (OOD) datasets and two in-distribution datasets (highlighted in bold) comparing the baseline network with a network trained with the organ transplantation augmentation. Results for the hyperspectral imaging (HSI) (top) and RGB (bottom) modalities are shown. See Section 5.4.1 for a description of scenarios. Each boxplot shows the interquartile range (IQR) with the median (solid line) and mean (dotted line). The whiskers extend up to 1.5 times of the IQR. Each point represents the aggregated class-level performance. Results for the normalized surface dice (NSD) are shown in Figure B.10. This figure was adapted from [202].

with a DSC increase by 50 % (HSI) and 46 % (RGB). When faced with situs occlusions, the most significant DSC improvement for HSI is achieved for the classes pancreas (283 %) and stomach (69 %). The organ transplantation augmentation even marginally improves the average class performance on in-distribution data from 0.86 (SD 0.10) (original) to 0.91 (SD 0.15) (no-occlusion).

### Augmentation Comparison

A comparison of our organ transplantation augmentation to other topology-aware augmentations is shown in Figure 5.34 (DSC) and Figure B.9 (NSD). Across all six OOD datasets, the only consistent ranking is that the organ transplantation augmentation consistently ranks first and the baseline typically ranks last.

In general, image-mixing augmentations (organ transplantation, CutMix and Jigsaw) surpass noise augmentations (Random Erasing and Hide-and-Seek). The noise augmentations achieve better generalization performance in the scenarios in which image parts are blacked out (isolation\_zero and removal\_zero) compared to the other scenarios. Augmentations that randomly select rectangles typically rank higher than similar augmentations that use a grid structure (e.g., CutMix vs. Jigsaw). Using elastic transformations yields better results than the baseline but cannot outperform any of the other augmentation techniques on average.

### Example Predictions

Figure 5.35 shows example predictions for the six OOD datasets using the baseline model and the organ transplantation augmentation for the HSI modality. The predictions from the isolated scenarios improve from a very poor to a reasonably good segmentation performance. For the resection scenarios, the prediction of the gallbladder improves drastically.

#### 5.4.4 Species Domain

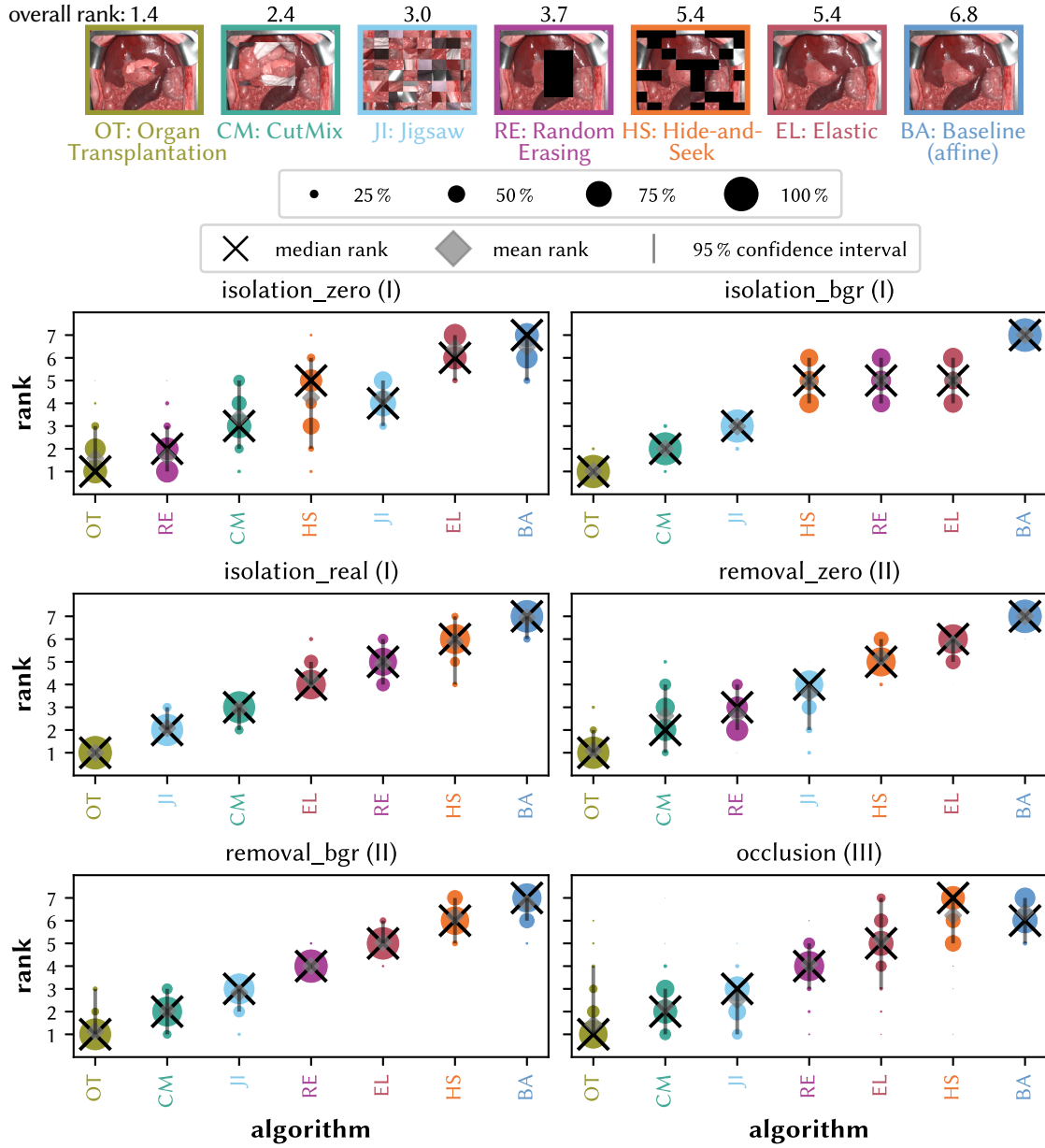
The domain transfer between species is arguably the biggest challenge. Here, we make a descriptive comparison between pigs and humans based on media spectra. Further, we present initial results for transfer learning with networks evaluated on the semantic human dataset using different options to make use of the porcine data.

Figure 5.36 compares median spectra from the semantic human and semantic porcine datasets for the 16 classes which the two datasets have in common. The general curve of the spectra is similar, however, there are notable differences between the species. For instance, the normalized reflectance values tend to be higher around 700 nm and lower around 900 nm for humans compared to pigs for some classes.

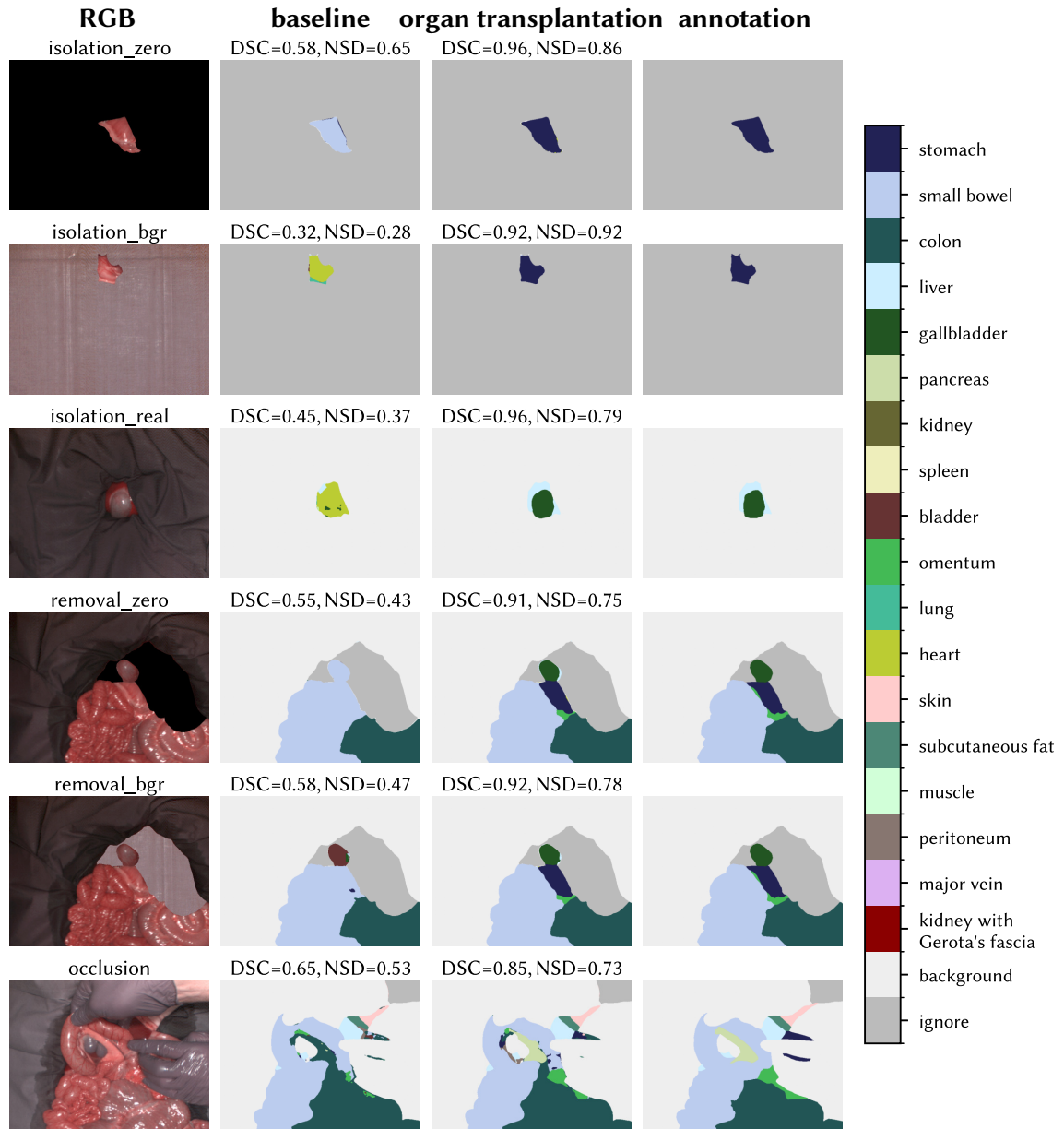
For transfer learning, it is important that spectra from one class are more similar to spectra from the same class of the other species than to other classes. To analyze this point further, we computed for each human median spectrum the nearest neighbor across the porcine median spectra and compared the corresponding labels. The results are shown in Figure 5.37.

We can see on the diagonal that in most of the cases, the nearest neighbor of a human median spectra has a different class than the original human class. The best matches are

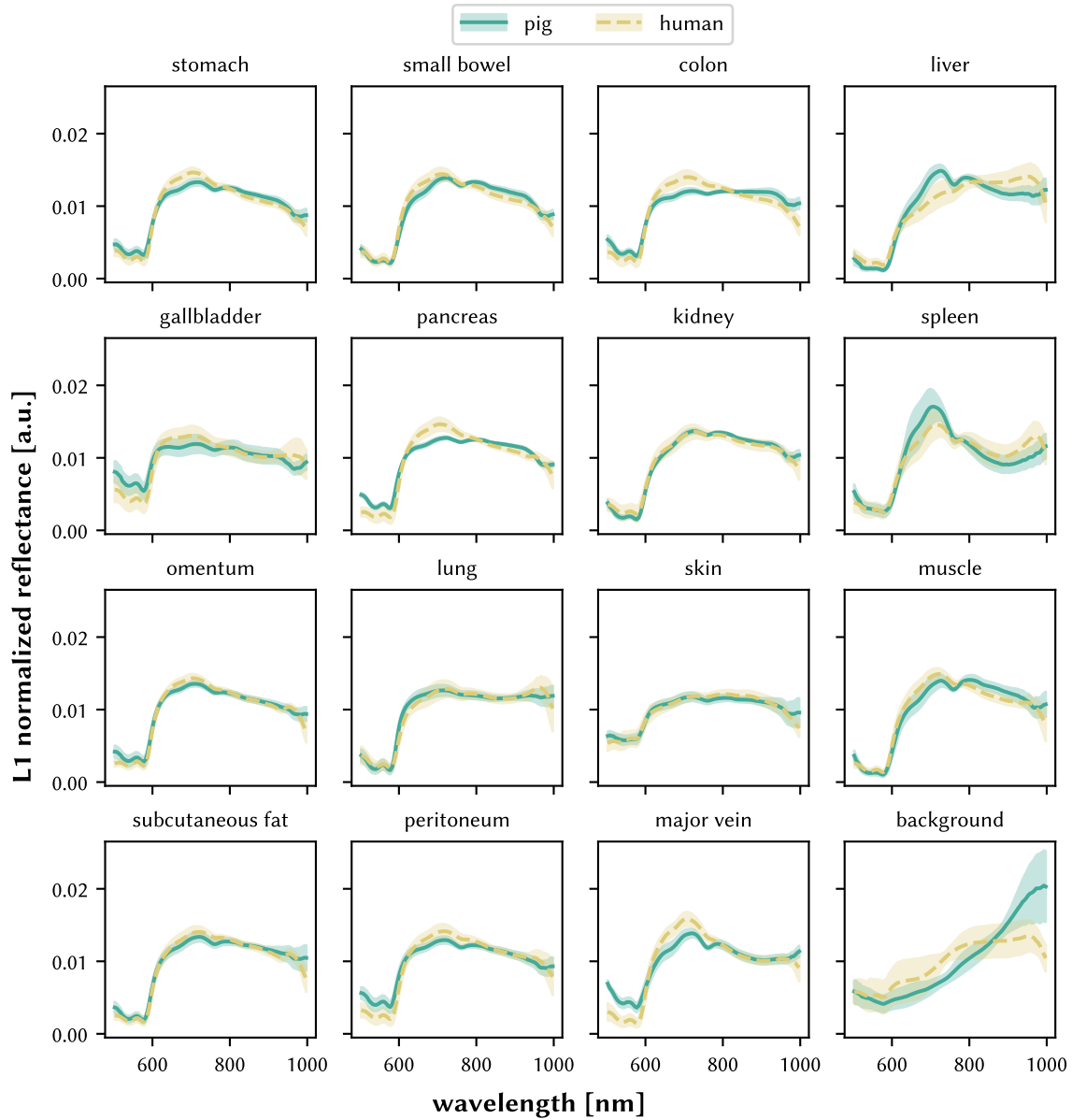




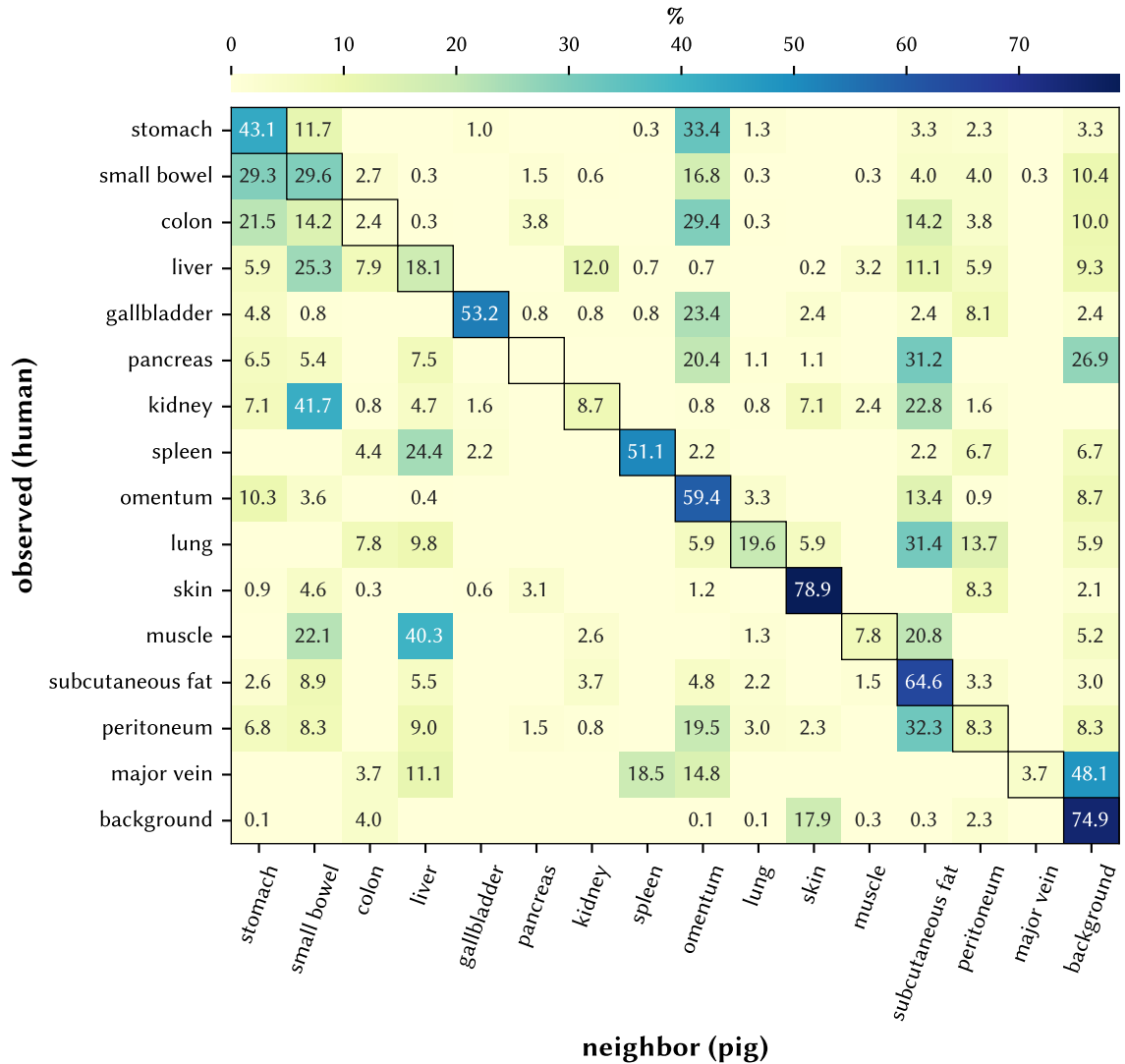
**Figure 5.34:** Uncertainty-aware ranking of the seven evaluated augmentation methods on the six geometric out-of-distribution datasets using the dice similarity coefficient (DSC). Consistently across all datasets, the organ transplantation augmentation ranks first whereas the baseline typically ranks last. The area of each blob is proportional to the relative frequency that the corresponding algorithm achieved the respective rank across 1000 bootstrap samples (concept from [236]). Each bootstrap sample consists of 19 hierarchically aggregated class-level DSC metric values. The lines encompass the 95 % quartile of the bootstrap results while the cross and the diamond denote the median and mean rank, respectively. Results for the normalized surface dice are in Figure B.9. This figure was adapted from [202].



**Figure 5.35:** Example predictions comparing the baseline network with the organ transplantation augmentation on all out-of-distribution (OOD) datasets using the hyperspectral imaging (HSI) modality. For each prediction, scores for the dice similarity coefficient (DSC) and normalized surface dice (NSD) are shown. Images are selected based on the maximum difference in DSC scores between the baseline and the organ transplantation augmentation networks for each dataset.



**Figure 5.36:** Comparison of spectral fingerprints for the semantic porcine and semantic human datasets stratified by the 16 classes which the two datasets have in common. For each organ, the median spectra (solid line) hierarchically aggregated to the subject level as well as the standard deviation (shaded area) across subjects is shown for each species.



**Figure 5.37:** Nearest neighbor class matrix for the semantic porcine and semantic human datasets based on media spectra comparisons from all annotated regions. For every human median spectrum  $s_h$  with corresponding class  $c_i$ , we searched for the nearest neighbor  $s_p$  across all pig median spectra with corresponding class  $c_j$ . Shown is the confusion matrix across all human ( $c_i$ 's) and pig labels ( $c_j$ 's). The  $(i, j)$ -th entry shows how many of the human median spectra from the  $i$ -th class have a nearest neighbor across the pig median spectra with the  $j$ -th class. The matrix is row-normalized highlighting how the median spectra from a human class are distributed across nearest neighbor pig classes. Values  $< 0.1$  % are not shown for clarity.

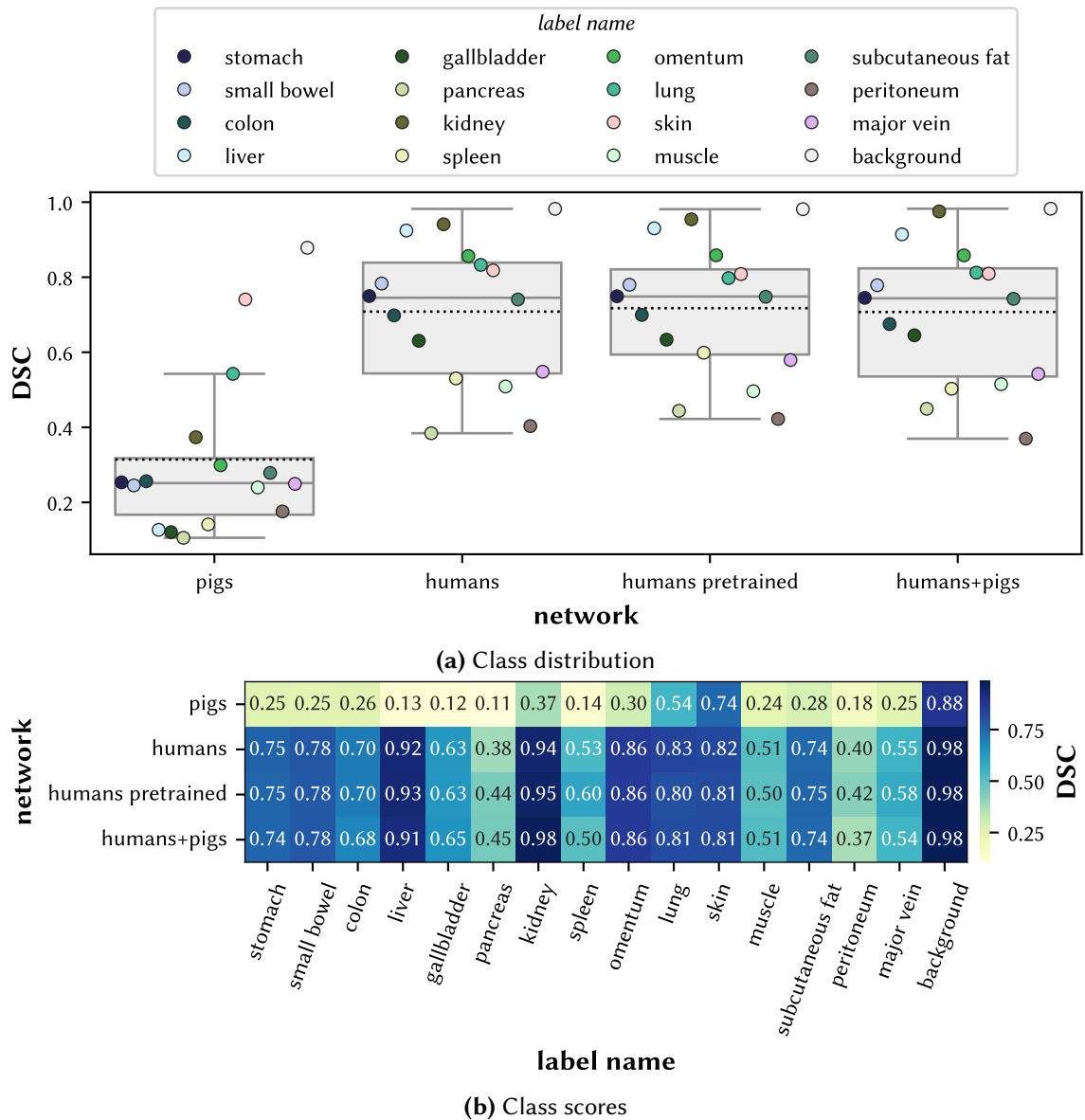
achieved for the classes *skin*, *background* and *subcutaneous fat*. For many other classes, however, the agreement is rather low. For example, human kidney spectra are closer to porcine small bowel spectra than to porcine kidney spectra. This indicates that the spectral fingerprints of the two species are not very similar and hence transfer learning is challenging.

Our transfer learning results are shown in Figure 5.38. We always evaluated on the semantic human dataset and we are comparing networks that are trained only on the semantic porcine dataset, only on the semantic human dataset, pretrained on the semantic porcine dataset and further finetuned on the semantic human dataset and jointly trained on the semantic porcine and semantic human datasets.

Firstly, the *pigs* network (network with the organ transplantation augmentation trained on the semantic porcine dataset) achieves very poor DSC scores for all classes except for *background* and *skin* indicating that the porcine data does not contain enough information to distinguish classes in human images.

Secondly, a network trained on the human data, even though outperforms the pig network, has still relatively low scores for many classes (e.g., *pancreas*, *peritoneum* or *muscle*). Compared with our results on the porcine data (cf. Figure 5.33), this indicates that tissue discrimination is a much harder task on human images than on porcine images. Concretely, the average DSC across the 16 classes shown in Figure 5.38 on the semantic porcine dataset (validation\_unknown split of Figure 5.12) is 0.81 (SD 0.14) whereas on the semantic human dataset (validation split of Figure 5.26) the score is 0.71 (SD 0.19). The former score is from our network trained on the semantic porcine dataset and the latter score is from a network trained on the semantic human dataset. In both cases, the organ transplantation augmentation was used.

Thirdly, adding the porcine data to the training process has only a minor impact on the class DSC scores neither via pretraining nor via joint training. Some classes improve slightly (e.g., *pancreas*) while others deteriorate a bit (e.g., *lung*). What is more, the classes with a slight improvement exhibit still a very poor performance. These results indicate that the data distributions between the two species are too different so that the network cannot learn anything useful from the porcine data. This is also in agreement with the neighbor analysis of Figure 5.37.



**Figure 5.38:** Segmentation performance on the semantic human dataset using the dice similarity coefficient (DSC). We compare the segmentation performance for four different networks all evaluated on the validation split of the semantic human dataset: pigs (network trained on the semantic porcine dataset), humans (network trained only on the semantic human dataset), humans pretrained (pretrained pig network further finetuned on the semantic human dataset) and humans+pigs (network trained jointly on the semantic human and semantic porcine datasets). In all cases, the network including the organ transplantation augmentation was used. Each boxplot shows the interquartile range (IQR) with the median (solid line) and mean (dotted line). The whiskers extend up to 1.5 times of the IQR. Each point represents the performance of one class.

In this chapter, we will discuss the results presented in Chapter 5. In Section 6.1, we pick up some details about our spectral analysis task (RQ1). Section 6.2 discusses our counter-strategies for efficient training of HSI networks (RQ2) and Section 6.3 is dedicated to our segmentation networks (RQ3). We reason about the impact of the domain shifts in Section 6.4 (RQ4), including a discussion about the species gap. Finally, in Section 6.5, we elaborate on device details, discuss our surgical setting and extend our view on non-healthy tissue types (pathologies) which have not been the target of this thesis but are nevertheless important for future applications.

## 6.1 Spectral Organ Fingerprints

In our median spectra analysis (RQ1), our objective was to determine whether different organs exhibit unique spectral fingerprints. The key insights from this analysis are twofold:

1. Each organ features indeed a distinct spectral fingerprint that can be automatically classified with high accuracy.
2. The primary source of variability arises from the organ itself, rather than image acquisition conditions such as the situs, the camera angle or repetitive measurements.

In the following, we discuss specific aspects of our machine learning task. General aspects like generalization to the species domain and challenges toward the transition to clinical practice are discussed in Section 6.4 and Section 6.5, respectively.

In our spectra classification network, we used two hyperparameters for tackling class imbalances: a weighted loss function and oversampling. The result of the grid search suggested the use of a weighted loss function but no oversampling (cf. Table 4.2). However, the grid search is only based on the maximum accuracy values on the validation split and does not indicate the relevance of the hyperparameters. For this, Figure 6.1 gives more

insights for all hyperparameters and Figure 6.2 especially for the two class imbalance hyperparameters.

Firstly, we can see that the differences in accuracy are rather low between grid search runs indicating that none of the selected hyperparameters have a huge effect. Secondly, there is no clear tendency for both of the class imbalance hyperparameters to improve or degrade the accuracy consistently. There are runs including a weighted loss function with low and high accuracies and the same is true for oversampling. What is more, the SD across subjects is higher than the differences between (neighboring) training runs weakening the importance of the grid search further.

In summary, the grid search was not the most important ingredient for the underlying classification task and a different selection of hyperparameters works similarly well. This suggests that addressing class imbalances has a limited impact even though the dataset is highly imbalanced.

## 6.2 Efficient Training of Hyperspectral Segmentation Networks

With RQ2, we wanted to know how we can efficiently train segmentation networks for spectral data. The key insights are twofold:

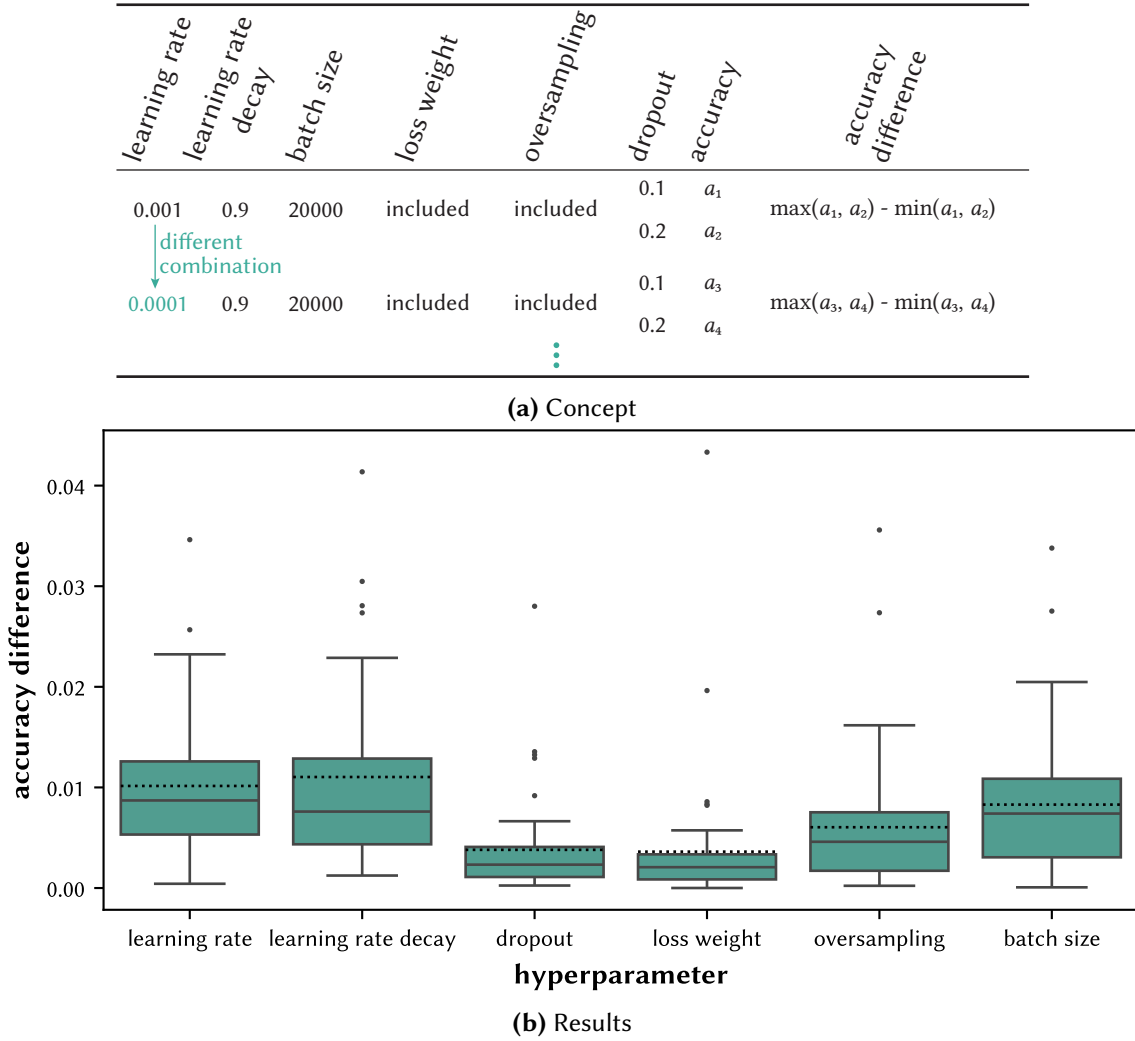
1. With the help of our proposed counter-strategies, HSI networks can be trained efficiently with high GPU utilization and reduced training times.
2. Countermeasures are especially important in environments with input/output (I/O) limitations (e.g., limited disk read speed).

In the following, we discuss specific aspects of our counter-strategies including remarks on when they should be used, how they can be applied and important considerations to bear in mind.

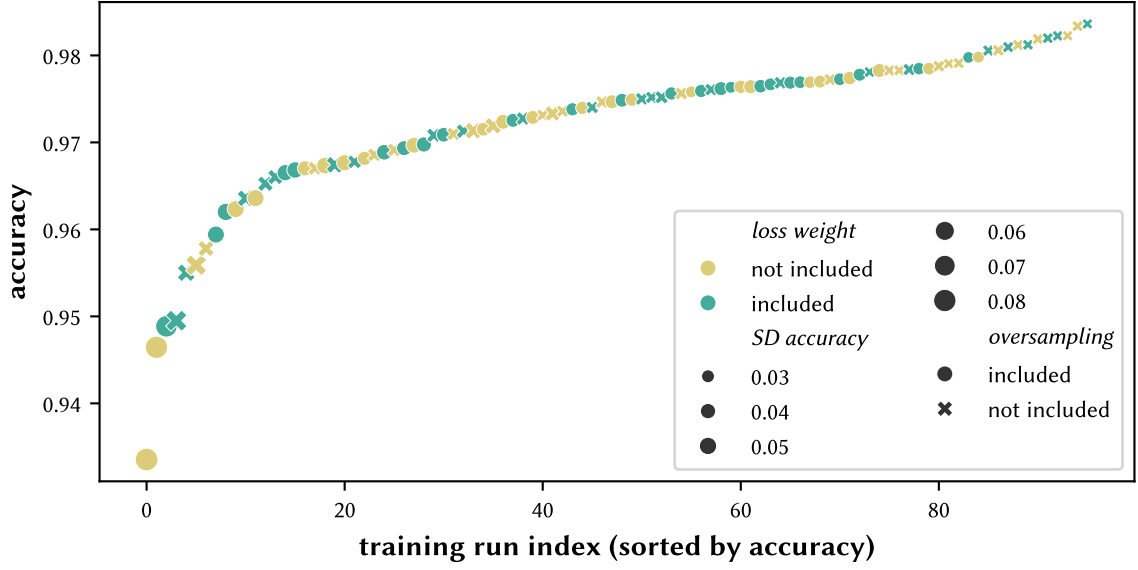
Our counter-strategies are specifically tailored to our application area, which is characterized by a high number of spectral channels, and may not yield the same benefits when applied to other areas. For instance, augmentations on the GPU are only advantageous when the CPU is the bottleneck and the GPU has free resources. This is particularly true for HSI data where a significant portion of time is dedicated to data loading and preprocessing while only a small fraction of time is spent on the model's image processing. Conversely, in other areas such as medical 3D imaging, the GPU often becomes the bottleneck due to the time it takes for the model to process the 3D images. In these cases, it may be more beneficial to offload some of the work to the CPU.

One of the main advantages of our counter-strategies is that they can be used independently from each other. For instance, one might choose to employ compression without





**Figure 6.1:** Impact of the hyperparameters from the grid search of the tissue atlas classification task. For each hyperparameter, the impact is shown in **(b)** as the distribution of differences in accuracy between the worst and the best run for each combination where only the respective hyperparameter changes. This concept is explained in **(a)** using the dropout hyperparameter as an example. Each row comprises two training runs (because only two dropout combinations are tested) where all hyperparameters except for dropout remain fixed and the maximum difference between those two runs is computed. This is repeated for every combination of remaining hyperparameters (e.g., in the second row with a different learning rate) to obtain all difference values for the dropout hyperparameter constituting the foundation for the dropout box plot. The same procedure is used for the remaining hyperparameters. Each boxplot shows the interquartile range (IQR) with the median (solid line) and mean (dotted line). The whiskers extend up to 1.5 times of the IQR. Outliers are shown as black points.



**Figure 6.2:** Results for the class imbalance hyperparameters from the grid search of the tissue atlas dataset classification task. The mean accuracy across subjects on the validation split is sorted and plotted for each training run. The color highlights whether the loss weighting function of Equation 4.2 was used and the symbols denote the usage of oversampling during training. Both approaches target class imbalances. The size of the points corresponds to the standard deviation (SD) across subject-level accuracy values.

resorting to GPU augmentations. This flexibility enables the evaluation of each counter-strategy separately for specific use cases, retaining only those that effectively reduce the training time.

Concerning our proposed shared, fixed, and pinned memory ring buffer, it is important to consider the potential issues of blocked memory allocation: Pinned memory regions are exclusively reserved for the application and are inaccessible to the operating system or other tasks (even inside the same application). This exclusivity can lead to complications that are further impacted by other processing steps. For instance, if other stages in the pipeline require the allocation of large amounts of memory, reserving a significant memory portion for the ring buffer may not be advisable. Therefore, this strategy should always be implemented holistically while taking the entire pipeline into account.

## 6.3 Surgical Scene Segmentation of Hyperspectral Images

With RQ3, we wanted to know the optimal spatial granularity and modality for semantic scene segmentation in surgical hyperspectral imaging. The key insights are as follows:

1. Segmentation performance improves with larger spatial granularity across various validation metrics and modalities.
2. HSI outperforms TPI and RGB modalities for all granularities and validation metrics.
3. The performance of our HSI image model is comparable or better than every other model across all studied number of training subjects and is on par with the performance of a secondary medical expert.

In this section, we discuss specific aspects of our work, including differences compared to our classification task, the implications of the metrics we used, the impact of the learning rate hyperparameter, the performance limit of our superpixel approach, aspects of our model comparison and limitations arising from our annotations<sup>1</sup>. Generalization aspects of our models are discussed in Section 6.4 and technical and clinical challenges are discussed in Section 6.5.

### From Classification to Segmentation

Our classification task is considerably simpler than our segmentation task since it relies solely on polygon annotations that explicitly use mainly clear regions of an organ. It is also based on median spectra which reduces noise and enhances the distinction between classes. This difference in complexity is further illustrated in the location maps of Figure 4.6 (classification) and Figure 4.7 (segmentation). Organs in the tissue atlas are typically centrally located and have a clear view whereas organs in the segmentation dataset are located across the entire image. The segmentation network is tasked with predicting a class for every pixel, regardless of noise, imperfect views or visibility of only small structures. This task, while more challenging, provides a more accurate reflection of surgical reality and the future needs for autonomous robotic surgery.

In both tasks, we observed that performance significantly varies across different classes. However, the classes with the lowest performance are not the same for the two tasks. For classification, the classes with the lowest performance are *gallbladder*, *heart*, and *major vein*. In contrast, for segmentation, the classes with the lowest performance are *omentum*, *major vein*, and *kidney with Gerota's fascia*. Interestingly, for classes like the *heart*, segmentation appears to be easier than classification. This could be attributed to the fact that the segmentation network has access to the surrounding context and the heart is often located near the lung (cf. Figure 5.31), which can be detected quite well. The

<sup>1</sup>This section is based on [198].

poor performance of *major vein* (especially in the segmentation task) could be explained by its small size as shown in Figure 4.5.

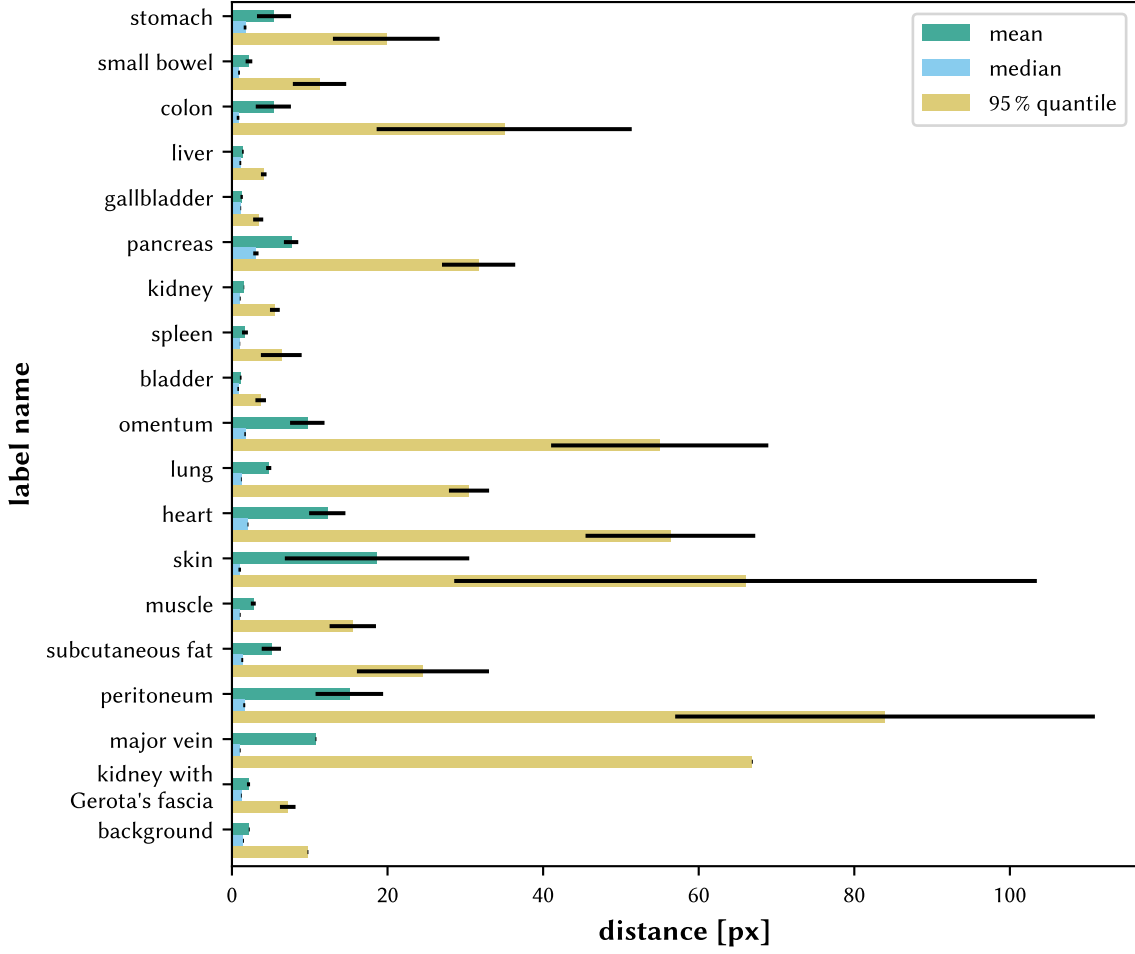
### Metrics

In line with the recommendations of segmentation challenges, we employed more than one metric to assess our results and to conduct our ranking [12, 190]. We utilized an overlap-based measure (DSC), a distance-based measure (ASD), and a boundary-overlap-based measure with special consideration to annotation uncertainty (NSD). Each metric examines specific attributes of the predicted segmentation map and a model may favor a particular metric. For instance, the pixel model excels under the DSC metric while the superpixel model performs optimally under the ASD (cf. Figure 5.16). Therefore, only a combination of multiple metrics can provide insights into a model’s overall performance.

A specific design decision that impacts all metrics is how to handle cases where classes are missing in the prediction. Evaluation frameworks like MONAI typically return nan or inf values in such situations, leaving the user to decide how to aggregate those. While this is less of an issue for the DSC and NSD metrics as they have defined limits, the approach to handling missing classes is a critical design decision for the ASD which is unbounded. There are several alternatives, such as completely disregarding these instances or applying a fixed penalty that might depend on the image diagonal. We chose to assign the maximum distance of the other classes to missing classes, which imposes a penalty without creating outliers. However, this approach has the drawback that the value for the missing class is dependent on the prediction of the other classes in the image.

The use of the NSD necessitates the establishment of a (class-specific) threshold which requires re-annotations of a subset of the images by at least one additional human annotator [165]. This subset is typically small relative to the size of the dataset (for instance, 20 of 506 images in our case) as obtaining annotations for more images or from multiple annotators is often impractical. Therefore, the thresholds largely depend on this subset and any errors in these annotations significantly impact the results. Missing classes in the re-annotation are also an issue, as corresponding distances cannot be calculated, meaning this part of an image does not contribute to the threshold. In the original formulation, this problem did not arise as an annotation was created separately for each known class [165]. However, in our case, the annotators were unaware of which classes were present in the image.

Another challenge is the decision of the aggregation function for the class distances per label. For each image pair annotated by both experts, we calculated the distances between the two annotations for each class, applied an aggregation function, and then averaged the aggregated values across subjects and classes (taking into account the hierarchical structure). In Figure 6.3, we present several thresholds  $\tau^o$  derived from different aggregation functions.



**Figure 6.3:** Potential thresholds for the normalized surface dice (NSD) metric by using different aggregation functions. Input to the aggregation functions are the class-wise distances between two expert annotations. The error bars indicate 0.25 standard deviation across subjects of the aggregated distance values. The threshold for the mean aggregation corresponds to the thresholds used in this thesis. This figure was adapted from [198].

Firstly, we observe significant variation across classes, e.g., with large differences between the two annotations for *peritoneum* and small deviations for *bladder*. Additionally, some classes exhibit high variations across subjects (for instance, the SD for the mean aggregation in the case of *skin* is 2.5 times higher than the mean itself). This emphasizes that determining the boundary for each class is not equally challenging. Generally, the agreement between our two expert annotations is rather low, suggesting that even for medical experts the decision of which pixel belongs to which class is neither straightforward nor unambiguous.

Secondly, the choice of the aggregation function significantly influences the thresholds. In their original work, Nikolov & Blackwell & Zverovitch et al. used the 95 % quantile of the distances [165] but this resulted in very high thresholds even above 80 px in our case. Therefore, we opted to use the mean, which yields more moderate distances always below 20 px. However, it is important to note that other aggregation methods like the median or another quantile could also have been viable options.

### Hyperparameters

For our segmentation models, we opted for default parameters whenever feasible and kept all parameters consistent across algorithms (such as learning rate) or chose them based on the same criteria (like memory usage for batch size). However, hyperparameters can influence model performance and given the varying input sizes and network architectures, it is possible that our hyperparameter settings are not ideal for all algorithms. To identify the optimal set of hyperparameters for each algorithm, a large number of training runs would be necessary. Considering that training all our 15 algorithms (five models and three modalities) for all five folds already took about 292 h GPU training time<sup>2</sup> for a single hyperparameter setting, comprehensive hyperparameter tuning would entail extremely high resource costs.

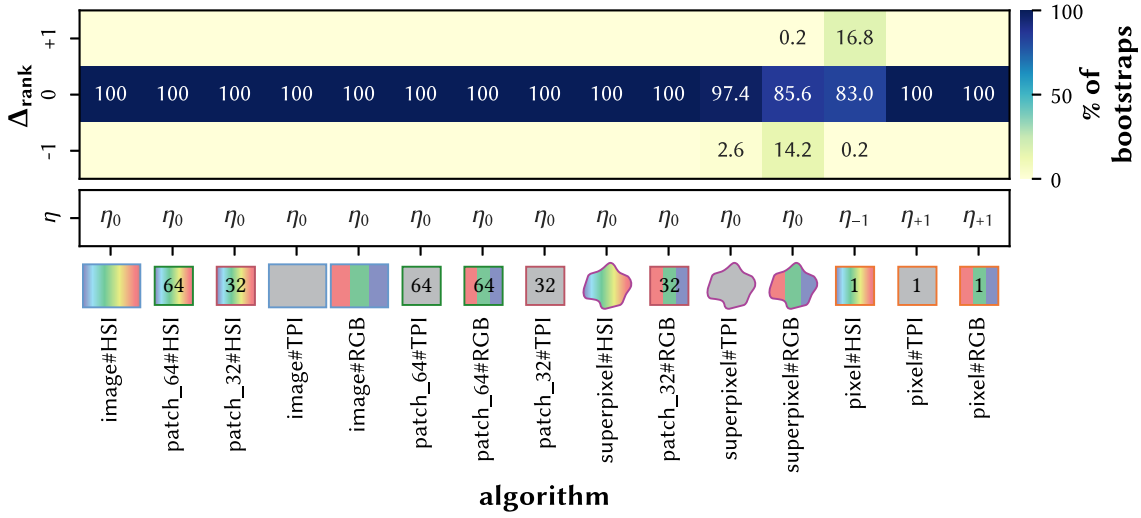
To illustrate the potential influence of our design choice on the algorithm ranking, we conducted a small hyperparameter search with the learning rate as an example. We chose the learning rate because it is widely regarded as one of the most crucial hyperparameters [156]. For each algorithm, we trained two additional networks: one with a lower learning rate of  $\eta_{-1} = 0.0001$  and another with a higher learning rate of  $\eta_{+1} = 0.01$  compared to the default learning rate of  $\eta_0 = 0.001$ . For each algorithm, we identified its optimal learning rate which is the learning rate among  $\eta_{-1}$ ,  $\eta_{+1}$ , and  $\eta_0$  that produced the highest average DSC on the validation data. After this, we repeated the overall ranking across algorithms (Figure 5.15) on the test data but this time instead of using a single fixed learning rate, we selected the network corresponding to the optimal learning rate for each algorithm. Except for the pixel-based models, the optimal learning rate was the same as the default learning rate for all algorithms. However, even for the pixel-based models, the average improvement in the DSC was only minor ( $< 0.007$  for all pixel models). This results in an overall ranking that is identical to that in Figure 5.15 (same sorting of the algorithms on the  $x$ -axis) and there are only minor differences in the ranking across the different bootstrap samples as can be seen in Figure 6.4. This reinforces the validity of our study results, even without extensive algorithm-specific hyperparameter tuning.

### Superpixels

Our superpixel classification approach is based on two primary assumptions: (1) superpixels comprise homogeneous regions where every pixel within a superpixel belongs to the same class, and (2) superpixel boundaries align with organ boundaries rather than

---

<sup>2</sup>Equivalent to roughly 32 kg of CO<sub>2</sub> if trained on an NVIDIA® GeForce RTX™ 2080 Ti GPU [124].



**Figure 6.4:** Ranking differences across bootstrap samples when optimizing the learning rate. For each algorithm, networks were retrained with a lower ( $\eta_{-1} = 0.0001$ ) and a higher ( $\eta_{+1} = 0.01$ ) learning rate compared to the default learning rate  $\eta_0 = 0.001$ . The optimal learning rate  $\eta$  was selected according to the highest average dice similarity coefficient (DSC) on the validation set for each algorithm. The ranking was repeated across all algorithms based on the networks with the optimal learning rate  $\eta$  instead of the default learning rate  $\eta_0$ . The heatmap shows how many of the 1000 bootstrap samples yield a different ranking of  $\Delta_{\text{rank}}$  ranks when using the learning rate optimized networks compared to the default networks of Figure 5.15. The sorting of the algorithms on the x-axis from best to worst was not affected by the learning rate and hence did not change compared to Figure 5.15. Values of 0 are not shown for clarity. This figure was adapted from [198].

intersecting them. We assessed these assumptions and established a performance limit for our superpixel model by assigning the modal value of all pixel labels within a superpixel as the superpixel label. This method directly incorporates the annotation labels, thereby serving as a performance limit for our model.

In Figure 6.5, we present the results of the performance limit for various metrics on the test set. The performance limit is closest to a perfect segmentation for the ASD, with an average of 2.91 (SD 0.74), followed by the DSC and NSD with average values of 0.92 (SD 0.03) and 0.74 (SD 0.04), respectively. The ASD is notably low and has a small SD because the distances between the annotation and performance limit are limited by the superpixel size and since each superpixel contains approximately 300 px, large distances are improbable (with  $\sqrt{300} \approx 17.32$  px). For DSC and NSD, the gap to a perfect segmentation is larger, suggesting that the superpixels do not perfectly align with the annotations which is primarily due to the borders between classes. It is possible that either the superpixels or the annotations (or both) are not positioned at the “true organ

border” and any deviation results in a reduced overlap (DSC) or the need to adjust some superpixel borders to align with the annotation (NSD). The NSD is lower than the DSC because the acceptable threshold  $\tau^o$  is very low for some organs (cf. Figure 6.3), so pixels with minor deviations already affect the NSD.

There exists a discrepancy between the performance limit and our model predictions across all metrics with average scores of 0.82 (SD 0.06), 16.51 (SD 9.23), and 0.61 (SD 0.09) for the DSC, ASD, and NSD, respectively. Assuming the superpixel features are sufficiently discriminative, this suggests potential for improvement in our superpixel model. This is further emphasized by our design choices aimed at ensuring comparability across models and modalities without specific optimizations for a single model (e.g., when we introduced augmentations, we applied them to all models). However, the image HSI model is not significantly behind the performance limit of Figure 6.5 (with average values of 0.90 (SD 0.04), 6.19 (SD 3.20), 0.80 (SD 0.07) for the DSC, ASD, and NSD, respectively), implying that further investment in the development of the superpixel model yield limited gains.

### **Model Comparison**

Our models were designed with the aim of maximizing comparability which involved similar design choices like the same U-Net architecture or a similar epoch size. This was crucial for our comparison as we intended the input size and modality to be the primary sources of variation, rather than model-specific design choices.

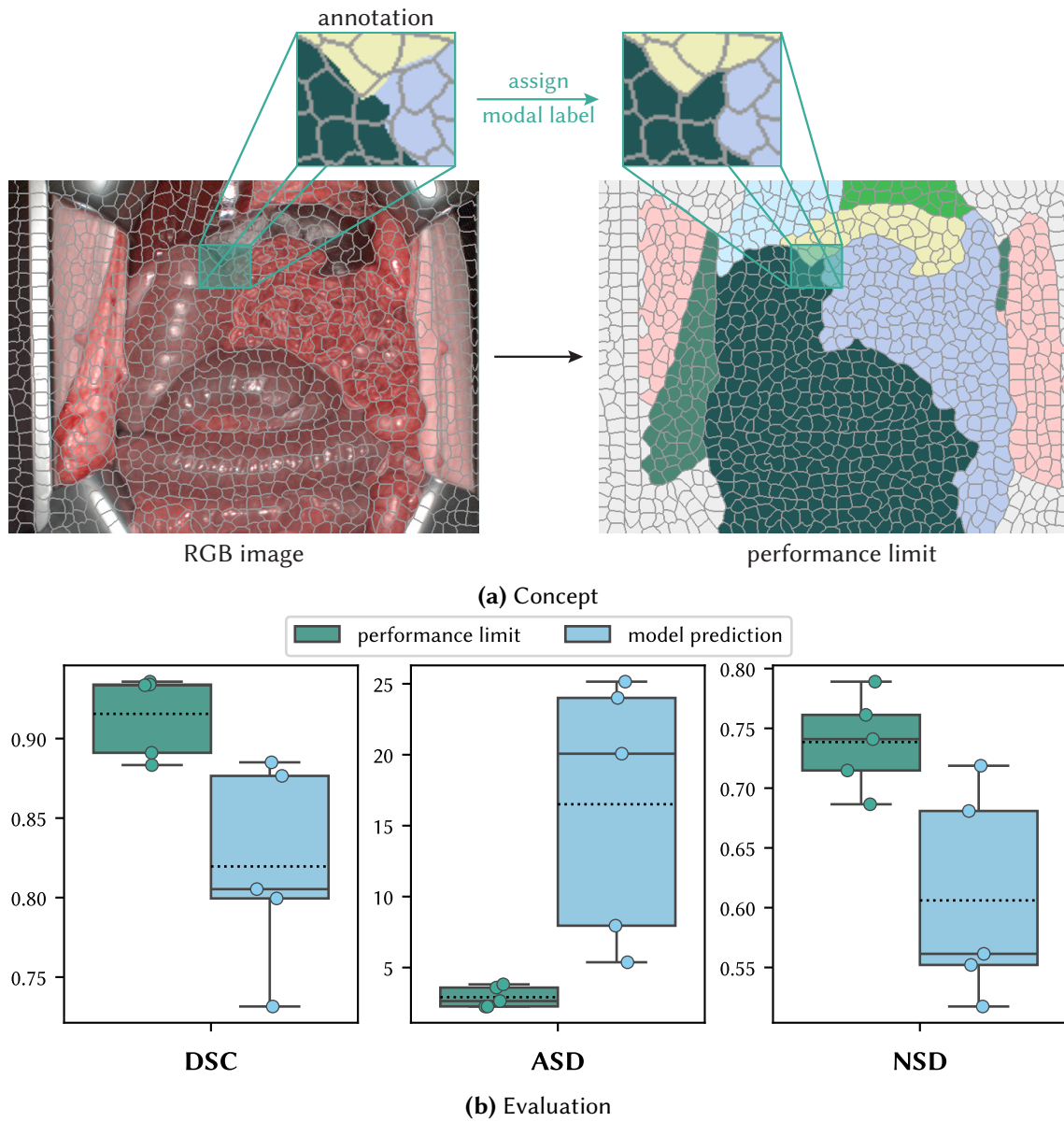
Similarly, we refrained from applying any post-processing to the network output even though certain models like the pixel model could potentially benefit from such operations, such as morphological operations. For inference, we limited the spatial context of each model to its defined input size. For instance, the patches that the patch\_32 model encounters during inference are explicitly non-overlapping to prevent the spatial context from exceeding a resolution of  $32 \times 32$  which would complicate comparisons across different spatial resolutions (e.g., superpixel vs. patch\_32). However, as seen in Figure 5.19, this design choice can lead to visible artifacts, such as at patch borders in the patch-based segmentation.

### **Annotations**

We observed that our segmentation networks are approaching the level of inter-rater variability. What is more, the average DSC of the inter-rater performance is only 0.89 (SD 0.07). This suggests that the quality of annotation is also limited which could be attributed to the difficulty of identifying organs even for medical experts. This limitation naturally sets boundaries for the performance of our networks and also future applications since it is likely that datasets from other domains (e.g., species domain) exhibit similar limitations.

This suggests that future applications should investigate this problem, for instance by re-annotating the data or label unclear structures as *invalid*. This could align with a





**Figure 6.5:** Performance limit of the superpixel approach by taking into account the reference annotations. **(a)** The label for each superpixel is determined by taking the mode of the annotated pixels inside the superpixel. **(b)** Comparison of the performance limit with the hyperspectral superpixel model for the three metrics dice similarity coefficient (DSC), average surface distance (ASD) and normalized surface dice (NSD). Each boxplot shows the interquartile range (IQR) with the median (solid line) and mean (dotted line). The whiskers extend up to 1.5 times of the IQR. Each point represents the performance of one test subject. This figure was adapted from [198].

human-in-the-loop paradigm with a continuing improvement of the network and a reduction in the number of problematic cases. Similarly, it could be worthwhile to adopt an active learning approach where the network selects the most suitable samples for annotation either from the pool of unlabeled data (which we possess but have not presented in this thesis) or the pool of partially annotated data (e.g., tissue atlas dataset).

## 6.4 Domain Shifts in Surgical Hyperspectral Imaging

In our domain generalization assessment (RQ4), our objective was to identify relevant domain shifts that affect the segmentation performance and we wanted to know whether we can compensate for them. The key insights from this assessment are as follows:

1. The subject domain has a minor impact on the segmentation performance with the effect being more pronounced at the image level than at the median pixel level.
2. Contextual shifts deteriorate the segmentation performance but this issue can be addressed by our proposed organ transplantation augmentation.
3. The species domain has a significant impact on the segmentation performance with the performance on human data being notably lower than on porcine data.

In the following, we discuss specific aspects of our domain generalization assessment including remarks on the subject domain, the effectiveness of our proposed organ transplantation augmentation<sup>3</sup> and the challenges posed by the species domain.

### Subject Domain

The subject domain encompasses variations due to individual differences but also those related to the surgery performed. Factors such as the type of surgery, the surgeon or the surgical setting can all affect the appearance of the organs. It is important to note that we did not separate these effects, but given the minor impact of the subject domain, this is not an issue. However, if it is observed that the subject domain has a more significant impact in other scenarios, it might be worth investigating these factors separately.

### Context Domain

Our organ transplantation augmentation offers a straightforward, network-independent solution to improve performance on geometrical OOD data. In addition to its effectiveness and computational efficiency, we perceive a significant benefit in its potential to lessen the amount of actual OOD data needed during training. Throughout the development process, we exclusively worked on the validation splits of manipulated datasets, keeping all real data as untouched test sets. This implies that all optimizations were carried out on

---

<sup>3</sup>This part is based on [202].

manipulated datasets but were still effective on real datasets. This is particularly crucial when real data is scarce or difficult to obtain. This was, for example, the case in our organ resection scenarios which would have necessitated an impractical number of animals when we had only used real data.

### Species Domain

As observed in Figure 5.38, particularly, organs like *pancreas*, *peritoneum*, and *muscle* exhibit poor performance. This suggests that it is significantly more difficult to segment human tissue than porcine tissue.

One possible explanation for this could be the difference in data acquisition environments. The porcine data was collected in a controlled setting while the human data was gathered in a real-world setting as found “in the wild”. For instance, the subset of standardized recordings in the tissue atlas dataset would not be feasible with human subjects.

Further, real-world surgeries are typically less controlled and more hectic which often leaves no time to capture the perfect image or to prepare organs for optimal visibility. This often results in bloody images as can be seen in the example images of Figure 4.1. Additionally, human tissue is usually covered more in fat, since patients undergoing surgery are typically older, adding another layer of complexity to the task.

We observed that incorporating porcine data into the training process did not significantly enhance the performance on human data, regardless of whether pretraining or joint training was used. This could be attributed to several factors.

Firstly, the species gap might be too large as we are transitioning from young, healthy pigs to older, sick humans who are undergoing surgery for a reason. Even though we only annotated healthy tissue, it is possible that side effects from diseases or complications during surgery may be visible in the images.

Secondly, the species could be too different with the spectra being too dissimilar. This is evident in the median spectra comparison of Figure 5.36 and especially the neighbor comparison of Figure 5.37. For almost all classes, the nearest pig spectrum to a human spectrum belongs to a different class than the observed human class. This inevitably confuses the neural network because it should learn that samples belong to a certain class but then there is contradicting information in the training data. The only classes with a somewhat good neighbor agreement, *skin* and *background*, are also the classes that already perform well on human data alone.

Thirdly, human images are generally more bloody and the organs are often covered in visceral fat compared to porcine organs. This could pose challenges for automatic segmentation [10], especially if the covering layer is thicker than the penetration depth of the light (several millimeters [248]).

## 6.5 Technical and Clinical Challenges in Hyperspectral Imaging

In the following, we discuss important aspects that should be taken into account before our work is moved to clinical practice. We discuss the limitations of our camera system (Section 6.5.1), the differences between minimally invasive and open surgery (Section 6.5.2) and the importance of non-healthy tissue types (pathologies) within the realm of surgical scene segmentation (Section 6.5.3).

### 6.5.1 Hardware Limitations

The HSI camera, depicted in Figure 1.1, used for capturing the images for our datasets, plays a crucial role as the resulting data forms the foundation for all the results of this thesis. In this section, we will discuss the limitations of our camera system and how it compares to conventional RGB systems.

#### Limitations of Our Camera System

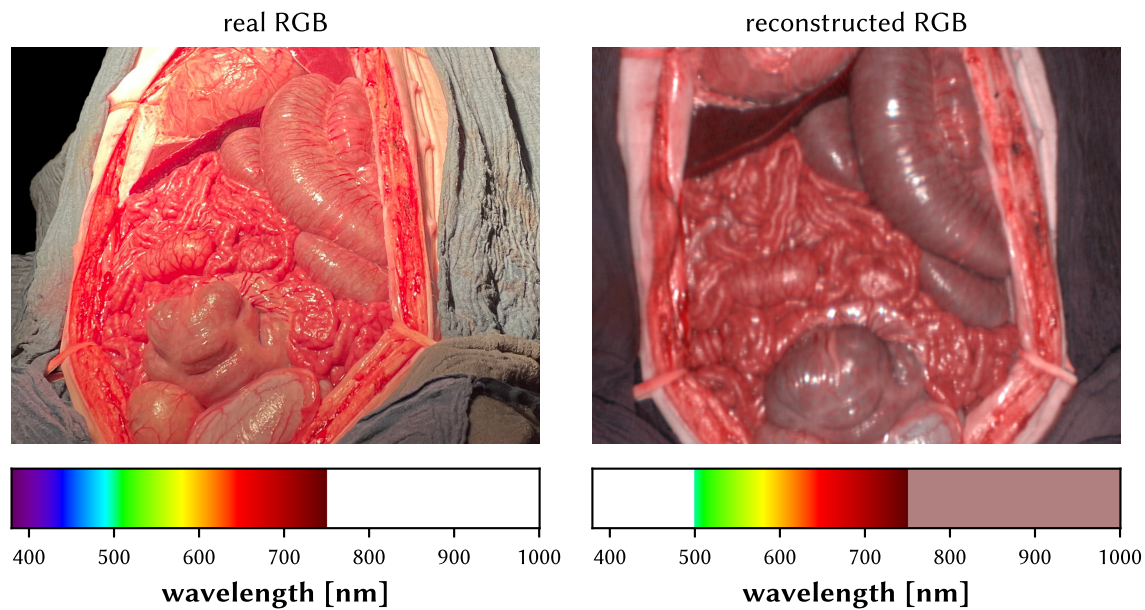
The main advantage of our HSI camera system is that it is commercially available and medically certified to be used during surgery [214]. However, despite significant advancements in HSI systems, they are neither as widespread nor as mature as RGB cameras. There are also practical challenges and limitations associated with HSI systems which can impact the resulting images.

Firstly, our HSI camera, particularly when compared to RGB cameras, has a relatively long acquisition time of approximately seven seconds. This limitation restricts the system to still objects and precludes the possibility of providing instant feedback on the current state of the scene. Organs with natural movement, such as the heart, inevitably exhibit visible artifacts (cf. Figure 4.1). These lengthy feedback loops pose a particular problem in surgical settings where the scene is constantly changing and surgeons need to respond quickly to these changes. For instance, in the case of continuous kidney perfusion monitoring during a clamping procedure, a scenario where it has recently been demonstrated that a MSI camera can monitor at video rate, our used system would not be suitable [14].

Secondly, the spatial resolution of the HSI camera with  $480 \times 640$  pixels (height, width) is relatively low compared to modern RGB cameras (e.g., 4K video imaging with  $2160 \times 3840$  pixels (height, width) [140]). This limitation restricts the level of detail in the captured organ images which could be crucial for distinguishing between classes (e.g., differentiating small vessel structures would be nearly impossible). This lower resolution also complicates the annotation process since annotators have to work with less visible

structures of the organs. This was a challenge that our annotators highlighted in their feedback.

Thirdly, the reconstructed RGB images do not match the quality of actual RGB images due to limited sharpness and color differences. The color discrepancies primarily stem from the limited blue spectrum that the HSI camera does not capture. This results in organs or objects displaying colors that differ from their real-life appearances and this could potentially cause confusion. See Figure 6.6 for a comparison of an exemplarily real RGB image and a reconstructed RGB image from our HSI camera.



**Figure 6.6:** Comparison of a real vs. a reconstructed RGB image from our hyperspectral imaging (HSI) system. The real RGB image was take with an Apple® iPhone® 12 Mini. On the bottom of each image, the captured wavelength range by each system is shown (the systems are insensitive to light denoted by the white areas).

Fourthly, the HSI camera does not offer a live view during image acquisition. This necessitates blind alignment of the camera and if the view is not as expected, the image has to be retaken. This can make the operation with the camera challenging, especially in the high-stress environment of real-world surgeries.

Fifthly, when capturing an image with the HSI camera, all other lights in the room must be switched off and ideally, all windows should be covered since the camera is supposed to be only used with the built-in light source and there is not autoexposure mechanism. This requirement can be not only inconvenient but sometimes impractical to incorporate into the clinical workflow, for instance, if it is not possible to completely cover the windows due to patient safety. Additionally, it is not always feasible to entirely turn off external light sources, such as monitors, for every image taken.

Sixthly, the HSI camera has a fixed focal length and lacks zoom capabilities. All images must be taken with an imaging distance of about 50 cm and there is no autofocus. While it is possible to swap the lens for one with a different focal length, this is a time-consuming process as the lenses must be manually exchanged and the sharpness adjusted with many test measurements. This adjustment can only be made before surgery, not during, and only if enough time is available before the surgery (for the HSI data of this thesis, always the same lens was used). Therefore, the decision on which lens to use must take into account the expected organs of interest. However, as we have observed in Figure 4.5, organ sizes vary greatly making it impossible to find a lens that is optimal for all organs. This, in turn, may impact the segmentation performance since classes like *major vein*, which are represented by only a few pixels, become challenging for both the classification and segmentation tasks.

Seventhly, the HSI camera can still be considered a prototype and is not yet ready for production. For example, there are instances where the image acquisition fails, necessitating to retake the image or even to restart the camera system. Regrettably, there are also occasions where images are not correctly stored and are consequently lost. This issue is the reason why two images are missing from the subset of standardized recordings of the tissue atlas dataset, i.e., only 5758 annotations instead of  $4 \cdot 3 \cdot 3 \cdot 20 \cdot 8 = 5760$  annotations (4 situs, 3 angles, 3 repetitions, 20 organs across 8 subjects).

However, despite these challenges, we demonstrated that the measured reflectance spectra are largely accurate (cf. Figure 5.9), even though there are differences for some colors and wavelength ranges. This suggests that our measurements were not entirely erroneous, that the data collected is still valuable and that our results still provide meaningful insights. What is more, HSI is a rapidly evolving field and advancements are anticipated in future implementations.

### **RGB vs. Hyperspectral Imaging**

Our segmentation networks showed only a minor improvement of HSI compared to RGB on full images (cf. Figure 5.14). However, when we utilized a more realistic dataset, one that includes geometrical OOD shifts for example, the disparity between RGB and HSI became more pronounced (cf. Figure 5.33). This aligns with the finding that the performance enhancement of HSI is more significant for smaller spatial granularities where limited context is available. It is worth noting that there might be other scenarios of OOD shifts where HSI proves more advantageous than RGB (e.g., segmentation of ischemic regions or detection of pathologies).

In general, when selecting the optimal imaging modality for scene segmentation, it is important to weigh the various pros and cons of HSI camera systems. Beyond the capacity to differentiate tissue classes, HSI systems capture detailed spectral information that provides additional benefits in surgical guidance. For instance, they can reveal functional tissue information such as the perfusion state, or assist in diagnosing diseased tissue [68, 251].

### 6.5.2 Minimally Invasive vs. Open Surgery

Surgical scene segmentation plays a crucial role in both minimally invasive and open surgery. In this thesis, our primary focus is on open surgery. These two surgical settings are quite distinct, resulting in images that do not look the same. In minimally invasive surgery, the perspective of the organ is different (not always from the top), the lighting conditions are more controlled and distances are more diverse. In general, minimally invasive surgery is less affected by straylight due to the closed-body situation.

Despite the lighting and view differences, our results have the potential to be equally applicable to minimally invasive surgery at the spectra level since the underlying tissue remains the same. However, at the image level, it is unlikely that our segmentation networks will yield the same performance out of the box as they do on our datasets from open surgeries since the networks may be confused by the different views and compositions. In such cases, additional training data is likely necessary.

In general, it might be easier to first deploy HSI in a minimally invasive setting given that camera systems are already an integral part of robot-based navigation. A logical next step could be to incorporate an HSI camera in place of an RGB camera in one of the existing surgical robots. Given that it is possible to reconstruct RGB from HSI images, this modification could be transparent to the surgeon.

### 6.5.3 Pathologies

In this thesis, our focus was on healthy tissue samples. However, in clinical reality, organs are not always healthy. There might be instances of ischemic organs, inflammation or other pathologies. For example, a cirrhotic liver differs from a healthy one exhibiting distinct spectral characteristics. While detecting these states themselves is important (e.g., identifying tumors, especially if they are not immediately apparent or even invisible to the human eye), it is equally crucial to maintain segmentation performance in the presence of these states. For instance, an organ should be detected irrespective of its perfusion state.

While there is existing research that identifies pathologies with HSI (see [68, 251] for an overview), these studies typically concentrate on individual pathologies and do not offer a comprehensive view that encompasses both healthy tissue and a range of pathologies simultaneously. However, such a holistic view is crucial for the development of surgical guidance systems as it is expected that networks should provide comprehensive information about the observed tissue that could assist surgeons. In contrast, our work has taken the initial steps toward providing such a holistic perspective, as we have already conducted analyses on a variety of different tissue classes.

In our research, we extensively utilized a porcine model which yielded significant insights and promising results for future applications. However, it is important to note that the porcine model, or animal models in general, have limitations in their ability to represent the diverse range of human pathologies and the implied ethical considerations. Consequently, achieving high segmentation performance on human data is a crucial step not only in its own right but also for advancing toward the detection of pathologies.

Future annotations of the data, particularly human data, should therefore consider the inclusion of hierarchical labels to mark pathologies. Naturally, this will require modifications to the network architecture and the training procedure to enable the prediction of more than one entity.

In general, our analysis of organ spectra, including spectral fingerprints and segmentation networks, lays the foundation for future research, although it needs to be expanded to encompass additional pathological states. Fortunately, our work in the context domain is not dependent on the tissue state, allowing the results to be directly applicable to pathologies. Similarly, our efficient training scheme, which is solely targeted at the data type rather than the data content, holds the same applicability.



## CONCLUSION

---

This thesis made significant progress in the field of autonomous robotic surgery through the development of fully automatic semantic scene segmentation on HSI data. While acknowledging that complications following surgery are an important problem and that the visual discrimination of tissue types presents a significant challenge for surgeons, we initially demonstrated the potential of HSI and its role in addressing this issue (RQ1). Following this, we tackled problems that hinder the extensive use of HSI in clinical applications. These include the inefficiency of training deep learning models on HSI data (RQ2), a lack of understanding regarding the optimal spatial and spectral granularity for semantic scene segmentation in surgical HSI (RQ3) and the impact of important domain shifts on the segmentation performance (RQ4).

In the following sections, we will discuss our contributions and provide a perspective on future work. Specifically, in Section 7.1, we will revisit the research questions that were introduced in Section 1.2, evaluate the extent to which we have addressed them and reference the corresponding disseminations. Finally, Section 7.2 discusses the impact of our work and gives an outlook on future challenges and opportunities in the field of surgical scene segmentation with HSI.

### 7.1 Summary of Contributions

#### **RQ1: Do different organs feature unique spectral fingerprints?**

We demonstrated the potential of HSI for the discrimination of tissue types by pioneering a large-scale analysis of organ spectra with an HSI database of unprecedented size consisting of 9057 images from 46 subjects annotated with 20 classes. We identified distinct spectral fingerprints for each organ and, with the help of our spectra classification network, we were able to classify them with an average accuracy of 95.4 % (SD 3.6 %). We further showed with the help of a linear mixed model (LMM) that the primary source of variability in our median spectra arises from the organ itself and not from image acquisition conditions like the situs, the camera angle or repetitive measurements. These findings led to a journal publication in *Scientific Reports* in 2022 [215].

We made a significant contribution to the HSI community by releasing a large-scale dataset of healthy tissue samples. To ensure ease of use, we provided accompanying visualizations and a Python package to load and process the data. Our technical validation further confirmed the validity of our measured spectra. What is more, we launched a website for this dataset with usage instructions and interactive visualization which can be accessed at [heiporspectral.org](https://heiporspectral.org). The Python framework hyperspectral tissue classification (HTC) [200] which can be used together with the dataset not only includes code to load and process the data but also serves as a repository for all the code related to the remaining results of this thesis, including our deep learning pipeline, pretrained models, and reproducibility instructions for all our experiments. The HSI dataset is accompanied by a journal publication in *Scientific Data* in 2023 [214].

### **RQ2: How can we train deep hyperspectral imaging networks efficiently?**

We were the first to demonstrate that the default training of deep neural networks on HSI data results in poor GPU utilization and extended training times. We managed to significantly improve the GPU utilization and reduce training times by implementing four independent counter-strategies of varying complexity: compression, quantization, GPU data augmentations and a shared, fixed and pinned memory ring buffer. We showed that these strategies can be combined, resulting in a speedup of up to 3.6. I presented these findings to a wide machine learning community at the *PyTorch Conference* in 2023 [201].

### **RQ3: What is the optimal spatial and spectral granularity for semantic scene segmentation in surgical hyperspectral imaging?**

We were the first to conduct a systematic analysis of the optimal modality and spatial granularity for semantic scene segmentation with HSI data. Our findings revealed that HSI is superior to both RGB and TPI across all spatial granularities. Moreover, the advantage of HSI increases with decreased spatial granularity. In our study, our image HSI model consistently ranked first, achieving an average DSC of 0.90 (SD 0.04) and reaching inter-rater variability with an average DSC of 0.89 (SD 0.07). We also evaluated our models in scenarios with limited training data and found that the image HSI model outperformed the other models in these cases as well. This work was published in the *Medical Image Analysis* journal in 2022 [198]. Further, it was accepted as a long abstract at the *International Conference on Information Processing in Computer-Assisted Interventions* in 2022, where I orally presented the work to an expert audience [197].

### **RQ4: Which are relevant domain shifts affecting the segmentation performance and can we compensate for them?**

We pioneered a comprehensive evaluation of the generalizability of our networks against important surgical domain shifts. For this, we analyzed the performance of our networks when faced with OOD data arising from different individuals (subject domain), changes

in the neighborhood of an organ (context domain) and when moving from porcine to human data (species domain).

Our findings indicated that the subject domain had only a minor impact on both the spectra and image levels, with the latter having a more significant effect. The dissemination of our analysis on the subject domain aligns with RQ1 and RQ3 (see above) on the spectra and image levels, respectively.

While the impact of the subject domain was minor in our case, the opposite is true for the context domain: We were the first to highlight the importance of contextual information in surgical scene segmentation and how segmentation networks struggle with geometrical OOD data. What is more, we provided the community with a simple, network-independent solution to enhance performance on geometrical OOD data which is now also available to the general computer vision community via the Kornia library [185]. We could show that our organ transplantation augmentation effectively addresses the issue of deteriorated performance on geometrical OOD data and ranks first compared to other topology-aware augmentations. Furthermore, we showed that these types of problems are ideally suited for working on manipulated data during development while preserving real data as untouched test sets. Our analysis of the problem and our solution were published and presented by me as a poster at the *International Conference on Medical Image Computing and Computer Assisted Intervention* conference in 2023 [202]. This work also earned me the *STudent-Author Registration (STAR) Award* and the *Young Scientist Award*.

Similarly drastic as the context domain is the impact of the species domain when the porcine model is evaluated on human data. However, even though we compared various techniques to incorporate porcine data during training, we found that the porcine data is of limited help and that the performance on human data is substantially more challenging than on porcine data. Our findings regarding the species domain have not yet been published elsewhere but we are currently preparing a manuscript to present these results to a broader audience.

## 7.2 Impact and Outlook

This thesis represents a pioneering step toward surgical scene segmentation with HSI and lays the foundation for exciting new applications and research directions. We advanced the field by providing new tools, methods and datasets. Our spectral analysis toolbox (e.g., LMMs) is also applicable to other HSI tasks where spectra comparison and variation decomposition are required. Our public HeiPorSPECTRAL dataset can already be used by others. The tricks we used to train deep neural networks faster could be beneficial not only for developers working with HSI data but also for other areas where data loading is a bottleneck (the code for our counter-strategies is available online [200]). Our findings

on the optimal modality and spatial granularity are valuable for anyone in the field processing HSI data with segmentation networks. To this end, our pretrained models are not limited to the HSI community but can also be utilized by researchers working on surgical RGB data, thanks to our trained RGB models. Our analysis of different domain shifts offers crucial insights into the failure modes of segmentation networks operating on HSI data. These insights are applicable to other tasks in the field. For instance, the species gap is of interest for every study using animal HSI data. Finally, our organ transplantation augmentation is by no means restricted to HSI data since it can be used for arbitrary spectral resolutions (including RGB data) and may even find applications outside the medical domain.

Beyond these immediate applications, our work raises new questions and opens up new challenges. Some of these are described in more detail below.

### Uncertainties

In this thesis, we focused solely on class-level prediction scores from the network for our evaluation. However, with this decision, we are completely ignoring the uncertainty of the network, i.e., instances where the network is (or should be) unsure about its prediction (a hint on this is given by the PCMs shown in Figure 5.24). These uncertainties are crucial in practice and for the acceptance of an autonomous system since it is preferable to acknowledge what we do not know rather than always predicting *something*.

Unfortunately, uncertainties present their own set of challenges and complexities. For instance, one cannot simply interpret softmax values as probabilities as this requires network calibration, a process that is not without its own challenges [83, 92, 173, 72, 78]. Therefore, future research should explore how to integrate calibrated uncertainties into our segmentation networks.

### Other Domains

In our generalizability assessment, we took into account the subject, context, and species domains due to their significance in medical imaging. However, it is important to note that there are other domains that also hold relevance for surgical scene segmentation.

For instance, it is unlikely that different HSI devices will measure exactly the same spectra and differences are to be expected. The extent of these differences depends on the device shift. However, since HSI is a novel technique with rapidly evolving hardware, changes are anticipated in the future. In the context of HSI (as compared to RGB cameras), changes and future developments are expected in two main areas:

1. In the spatial domain, differences like spatial resolution or zoom level may arise due to varying focal lengths leading to a change in the observed organ sizes. Further, we have seen in Figure 4.7 that the organs in our datasets have common locations in the image (often in the center) but this may be different in other datasets.

2. In the spectral domain, there are additional degrees of freedom, such as the number of channels, the wavelength range, or the spectral resolution. Furthermore, the light source, one of the most significant influencing factors for HSI images, could also have an impact (e.g., LED vs. halogen light sources) [15]. In practice, changes are unlikely to occur incrementally (e.g., a change in the light source and the zoom level might happen simultaneously) but it will be crucial to acquire data with disentangled factors to understand each effect and develop appropriate countermeasures.

Another domain to consider is the type of surgery performed. Our results focused on open surgery but minimally invasive surgery is also crucial, especially as more surgeries are being performed in this manner [195]. From an HSI perspective, these two types of surgeries are significantly different with distinctions such as static images vs. videos, different light sources and varying views on the organ of interest. Ideally, insights gained from one type of surgery should be transferable to the other. However, this is a complex step, and it might be wise to address the other issues first so that the impact of inevitable changes (like device changes) is reduced.

### **Pathologies**

As elaborated in Section 6.5.3, incorporating pathologies is vital for future applications. For an application to be beneficial during surgeries, it needs to identify tissues irrespective of the presence of pathologies (e.g., ischemic organs) and it needs to detect pathologies on its own (e.g., tumors). This holistic perspective is essential to offer real-world benefits.

### **Clinical Translation**

Ultimately, the long-term consequences of this thesis, after addressing the discussed limitations and previously presented new challenges, should be to apply our findings to clinical practice. It is crucial that new applications are seamlessly integrated into the surgical workflow so that they are accepted by the surgeons and can be implemented at minimal cost. For this to happen, camera providers and researchers should play together while always keeping the clinical side in the loop ensuring that the application is developed in line with practical surgical needs.

The focus of the camera providers should be to develop HSI systems that provide spatial and spectral resolutions comparable to currently employed RGB cameras. Both criteria are currently not met by our used HSI camera (cf. Section 4.1) as the spatial resolution is too low and the spectral resolution does not contain blue colors so that the resulting reconstructed RGB images differ in their visual appearance (cf. Section 6.5.1).

On the research side, we need to demonstrate that such a holistic application is possible. This application should not only include our segmentation networks (enhanced with pathology and uncertainty information) but also incorporate other HSI associated applications, such as perfusion estimation (cf. Figure 1.1). Further, the application must be

real-time ready which necessitates that the inference of our segmentation networks is sufficiently fast to avoid becoming a bottleneck. Finally, we need to confirm through prospective studies that such an application indeed offers tangible benefits to both surgeons and patients.

### Closing

Overall, this thesis represents an important first step toward surgical scene segmentation with HSI for autonomous robotic surgery. I acknowledge that I have not resolved all the issues in this area and perhaps even raised more questions than I have answered and uncovered more new problems than I solved. Arguably, gaining an understanding of these problems is valuable in its own right. In the end, I hope that my contributions to this field will prove beneficial for future research and aid in advancing the field.

With this, we have come to the end of my thesis and there is only one open question left:

*How much wood could a woodchuck chuck if a woodchuck could chuck wood?*<sup>1</sup>

---

<sup>1</sup>You: *A woodchuck could chuck no amount of wood since a woodchuck can't chuck wood.*

Me: *But if a woodchuck could chuck and would chuck some amount of wood, what amount of wood would a woodchuck chuck?*

You: *Even if a woodchuck could chuck wood, and even if a woodchuck would chuck wood, should a woodchuck chuck wood?*

Me: *A woodchuck should chuck if a woodchuck could chuck wood, as long as a woodchuck would chuck wood.*

You: *Oh. Shut up.*

From Ron Gilbert's adventure game "Monkey Island 2: LeChuck's Revenge".

## OWN CONTRIBUTIONS AND PUBLICATIONS

---



This thesis was written in the division of Intelligent Medical Systems (IMSY) at the German Cancer Research Center (DKFZ) in Heidelberg as part of the Helmholtz Information & Data Science School for Health (HIDSS4Health). The division is headed by Lena Maier-Hein, who is also the first supervisor of this thesis and my data science principal investigator (PI). This is an interdisciplinary thesis and during my tenure at the DKFZ, I was closely collaborating with the Department of General, Visceral, and Transplantation Surgery at the Heidelberg University Hospital and here with the group of my life science PI Felix Nickel.

This chapter gives an overview of my contributions to the research questions presented in this thesis (Section A.1) and presents a list of all of my publications (Section A.2).

### A.1 Own Contributions

In the following, I clarify my most significant contributions to each of the research questions presented in this thesis. During my time at the division of IMSY, I also supervised various bachelor and master students and the resulting theses are summarized below as well.

#### **RQ1: Do different organs feature unique spectral fingerprints?**

The work for this research question was published in [215]. For this project, my main contribution was to implement and evaluate the machine learning models.

Our open data efforts are part of this research question and were published in [214]. I took on the task of creating the website for the dataset, the organ profiles and the interactive visualizations. Additionally, I was responsible for organizing and structuring the dataset and creating the data repository. I also published the corresponding hyperspectral tissue classification (HTC) framework for our work.

**RQ2: How can we train deep hyperspectral imaging networks efficiently?**

The work for this research question was published in [201]. I identified the bottlenecks in data loading and implemented the appropriate solutions to address these issues. Furthermore, I conducted benchmarks on the different data loading solutions to evaluate their performance and effectiveness. Additionally, I designed and developed the shared, fixed and pinned memory ring buffer for the image model as well as for the smaller spatial granularities.

**RQ3: What is the optimal spatial and spectral granularity for semantic scene segmentation in surgical hyperspectral imaging?**

The work for this research question was published in [198, 197]. I implemented the patch and image models used in our comparison of spatial granularities and modalities. I extended our HTC framework by including the code for all our experiments, pretrained models and reproducibility instructions.

**RQ4: Which are relevant domain shifts affecting the segmentation performance and can we compensate for them?**

For this research question, we evaluated the effect of subject, context and species domain shifts on the segmentation performance.

For the subject domain, I designed the experiment for the generalization analysis on the image-level. The work on the subject domain was published in [215] (spectra-level) and [198] (image-level).

For the context domain, I analyzed the drop in performance of geometrical OOD data using manipulated datasets to aid our investigation (as part of a supervised bachelor thesis, see below). Further, I analyzed the neighborhood relationship between organs in the dataset. I conducted the comparative study with other topology-aware augmentation methods, created the ranking of the results and performed the literature review about commonly used augmentations. I extended our HTC framework by including the code for all our experiments, pretrained models and reproducibility instructions. Our work on the context domain was published in [202]

For the species domain, I took on the recruitment of the annotators. I created the descriptive visualizations, the nearest neighbor matrix and the comparison of machine learning models with different inclusions of the porcine data. The work on the species domain has not yet been published.

**Supervised Theses**

I co-supervised the bachelor thesis of Oskar Weinfurtner, titled *Variational Autoencoders for Generalizability in Organ Classification on Hyperspectral Images*. As part of this work, we adapted the “Domain Invariant Variational Autoencoders” [99] to our semantic porcine dataset, paying special attention to the subject domain. Although we were able to get the



method working on a simplified simulated task, it did not yield any advantages on our real-world HSI dataset. Furthermore, the method failed to deliver satisfactory results on other tasks (e.g., species domain) as well.

I supervised the master thesis of Mozzam Motiwala, which was titled *Hyperspectral Image Segmentation Using Weakly Supervised Learning*. The primary goal of our work was to leverage image-level labels (i.e., a list of organ names for each image) in order to enhance the performance of our segmentation models. We aimed to achieve this with the assistance of class activation maps to guide the segmentation process. However, we encountered challenges as the classification task itself proved to be difficult and the overlap between the class activation maps and the segmentation masks was rather low.

I co-supervised the bachelor thesis of Alessandro Motta, which was titled *Context Importance in Organ Segmentation with Hyperspectral Imaging*. As part of this work, we analyzed the neighbor relationship between organs and the importance of context for the segmentation of organs with the help of manipulated datasets. This work laid the foundation for our research questions RQ4 (context domain).

I supervised the master thesis of Fabian Wolf with the title *Self-Supervised Learning for Medical Hyperspectral Image Segmentation*. The primary objective of our work was to leverage the vast amount of unlabeled HSI data that we have (even though it was not used in this thesis), through self-supervised learning approaches. However, we had to conclude that either these methods were not sufficiently effective on our HSI tasks or that our datasets were too small for these approaches since we were unable to improve our segmentation performance. Furthermore, we observed that vision transformers [60], which are often used in self-supervised learning, did not perform well on our HSI data.

## A.2 Own Publications

In this section, I list all of the publications which I (co-)authored. This includes peer-reviewed journal publications and conference proceedings, preprints, poster and oral presentations as well as software releases.

### First Author Publications

1. Silvia Seidlitz, **Jan Sellner**, Jan Odenthal, Berkin Özdemir, Alexander Studier-Fischer, Samuel Knödler, Leonardo Ayala, Tim J. Adler, Hannes G. Kenngott, Minu Tizabi, Martin Wagner, Felix Nickel, Beat P. Müller-Stich, and Lena Maier-Hein. “Robust deep learning-based semantic organ segmentation in hyperspectral images”. In: *Medical Image Analysis* 80 (Aug. 2022), p. 102488. ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102488

2. **Jan Sellner** and Silvia Seidlitz. *Hyperspectral Tissue Classification*. Version v0.0.15. Feb. 5, 2024. DOI: 10.5281/zenodo.6577614. URL: <https://github.com/IMSY-DKFZ/htc>
3. Silvia Seidlitz, **Jan Sellner**, Jan Odenthal, Berkin Özdemir, Alexander Studier-Fischer, Samuel Knödler, Leonardo Ayala, Tim J. Adler, Hannes G. Kenngott, Minu Tizabi, Martin Wagner, Felix Nickel, Beat P. Müller-Stich, and Lena Maier-Hein. *Robust deep learning-based semantic organ segmentation in hyperspectral images*. Oral presentation at the 13th international conference on Information Processing in Computer-Assisted Interventions (IPCAI), Tokyo, Japan. June 7, 2022
4. **Jan Sellner**, Silvia Seidlitz, Alexander Studier-Fischer, Alessandro Motta, Berkin Özdemir, Beat Peter Müller-Stich, Felix Nickel, and Lena Maier-Hein. “Semantic Segmentation of Surgical Hyperspectral Images Under Geometric Domain Shifts”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Cham: Springer Nature Switzerland, 2023, pp. 618–627. ISBN: 978-3-031-43996-4. DOI: 10.1007/978-3-031-43996-4\_59
5. **Jan Sellner**, Silvia Seidlitz, and Lena Maier-Hein. *Dealing with I/O bottlenecks in high-throughput model training*. Poster presented at the PyTorch Conference 2023, San Francisco, United States of America. Oct. 16, 2023. URL: [https://e130-hyperspectral-tissue-classification.s3.dkfz.de/figures/PyTorchConference\\_Poster.pdf](https://e130-hyperspectral-tissue-classification.s3.dkfz.de/figures/PyTorchConference_Poster.pdf)
6. Klaus Kades, **Jan Sellner**, Gregor Koehler, Peter M Full, T Y Emmy Lai, Jens Kleesiek, and Klaus H Maier-Hein. “Adapting Bidirectional Encoder Representations from Transformers (BERT) to Assess Clinical Semantic Textual Similarity: Algorithm Development and Validation Study”. In: *JMIR Med Inform* 9.2 (Feb. 3, 2021), e22795. ISSN: 2291-9694. DOI: 10.2196/22795
7. **Jan Sellner**, Patrick Thiam, and Friedhelm Schwenker. “Visualizing Facial Expression Features of Pain and Emotion Data”. In: *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*. Ed. by Friedhelm Schwenker and Stefan Scherer. Cham: Springer International Publishing, 2019, pp. 101–115. ISBN: 978-3-030-20984-1. DOI: 10.1007/978-3-030-20984-1\_9

### Co-Author Publications

1. Alexander Studier-Fischer, Silvia Seidlitz, **Jan Sellner**, Berkin Özdemir, Manuel Wiesenfarth, Leonardo Ayala, Jan Odenthal, Samuel Knödler, Karl Friedrich Kowalewski, Caelán Max Haney, Isabella Camplisson, Maximilian Dietrich, Karsten Schmidt, Gabriel Alexander Salg, Hannes Götz Kenngott, Tim Julian Adler, Nicholas Schreck, Annette Kopp-Schneider, Klaus Maier-Hein, Lena Maier-Hein, Beat Peter

- Müller-Stich, and Felix Nickel. “Spectral organ fingerprints for machine learning-based intraoperative tissue classification with hyperspectral imaging in a porcine model”. In: *Scientific Reports* 12.1 (June 30, 2022), p. 11028. ISSN: 2045-2322. DOI: 10.1038/s41598-022-15040-w
2. Alexander Studier-Fischer, Silvia Seidlitz, **Jan Sellner**, Marc Bressan, Berkin Özdemir, Leonardo Ayala, Jan Odenthal, Samuel Knoedler, Karl-Friedrich Kowalewski, Caelán Max Haney, Gabriel Salg, Maximilian Dietrich, Hannes Kenngott, Ines Gockel, Thilo Hackert, Beat Peter Müller-Stich, Lena Maier-Hein, and Felix Nickel. “HeiPorSPECTRAL - the Heidelberg Porcine HyperSPECTRAL Imaging Dataset of 20 Physiological Organs”. In: *Scientific Data* 10.1 (June 24, 2023), p. 414. ISSN: 2052-4463. DOI: 10.1038/s41597-023-02315-8. URL: <https://heiporspectral.org>
3. Tom Rix, Marco Hübner, Kris K. Dreher, Jan-Hinrich Nölke, Leonardo Ayala, Melanie Schellenberg, **Jan Sellner**, Silvia Seidlitz, Alexander Studier-Fischer, Beat Müller-Stich, Felix Nickel, Alexander Seitel, and Lena Maier-Hein. “Deep learning for spectral image synthesis”. In: *Multimodal Biomedical Imaging XVII*. ed. by Fred S. Azar, Xavier Intes, and Qianqian Fang. Vol. PC11952. International Society for Optics and Photonics. SPIE, 2022, PC119520I. DOI: 10.1117/12.2608622
4. Leonardo Ayala, Tim J. Adler, Silvia Seidlitz, Sebastian Wirkert, Christina Engels, Alexander Seitel, **Jan Sellner**, Alexey Aksenov, Matthias Bodenbach, Pia Bader, Sebastian Baron, Anant Vemuri, Manuel Wiesenfarth, Nicholas Schreck, Diana Mindroc, Minu Tizabi, Sebastian Pirmann, Brittaney Everitt, Annette Kopp-Schneider, Dogu Teber, and Lena Maier-Hein. “Spectral imaging enables contrast agent-free real-time ischemia monitoring in laparoscopic surgery”. In: *Science Advances* 9.10 (2023), eadd6778. DOI: 10.1126/sciadv.add6778
5. Leonardo Ayala, Tim J. Adler, Silvia Seidlitz, Sebastian Wirkert, Christina Engels, Alexander Seitel, **Jan Sellner**, Alexey Aksenov, Matthias Bodenbach, Pia Bader, Sebastian Baron, Anant Vemuri, Manuel Wiesenfarth, Nicholas Schreck, Diana Mindroc, Minu Tizabi, Sebastian Pirmann, Brittaney Everitt, Annette Kopp-Schneider, Dogu Teber, and Lena Maier-Hein. *Video-rate perfusion imaging in laparoscopy*. Oral presentation at the 13th international conference on Information Processing in Computer-Assisted Interventions (IPCAI), Tokyo, Japan. June 7, 2022
6. Maximilian Dietrich, Silvia Seidlitz, Nicholas Schreck, Manuel Wiesenfarth, Patrick Godau, Minu Tizabi, **Jan Sellner**, Sebastian Marx, Samuel Knödler, Michael M. Allers, Leonardo Ayala, Karsten Schmidt, Thorsten Brenner, Alexander Studier-Fischer, Felix Nickel, Beat P. Müller-Stich, Annette Kopp-Schneider, Markus A. Weigand, and Lena Maier-Hein. *Machine learning-based analysis of hyperspectral images for automated sepsis diagnosis*. 2021. DOI: 10.48550/arXiv.2106.08445. arXiv: 2106.08445 [eess.IV]

7. Constantin Seibold, Simon Reiß, Saquib Sarfraz, Matthias A. Fink, Victoria Mayer, **Jan Sellner**, Moon Sung Kim, Klaus H. Maier-Hein, Jens Kleesiek, and Rainer Stiefelhagen. “Detailed Annotations of Chest X-Rays via CT Projection for Report Understanding”. In: *33rd British Machine Vision Conference Proceedings, BMVC 2022*. 33rd British Machine Vision Conference. BMVC 2022 (London, Vereinigtes Königreich, Nov. 21–24, 2022). British Machine Vision Association, BMVA, 2022. URL: <https://bmvc2022.mpi-inf.mpg.de/0058.pdf>
8. Matthias A. Fink, Klaus Kades, Arved Bischoff, Martin Moll, Merle Schnell, Maike Küchler, Gregor Köhler, **Jan Sellner**, Claus Peter Heussel, Hans-Ulrich Kauczor, Heinz-Peter Schlemmer, Klaus Maier-Hein, Tim F. Weber, and Jens Kleesiek. “Deep Learning-based Assessment of Oncologic Outcomes from Natural Language Processing of Structured Radiology Reports”. In: *Radiology: Artificial Intelligence* 4.5 (2022), e220055. DOI: 10.1148/ryai.220055
9. Kris K. Dreher, Leonardo Ayala, Melanie Schellenberg, Marco Hübner, Jan-Hinrich Nölke, Tim J. Adler, Silvia Seidlitz, **Jan Sellner**, Alexander Studier-Fischer, Janek Gröhl, Felix Nickel, Ullrich Köthe, Alexander Seitel, and Lena Maier-Hein. “Un-supervised Domain Transfer with Conditional Invertible Neural Networks”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Cham: Springer Nature Switzerland, 2023, pp. 770–780. ISBN: 978-3-031-43907-0. DOI: 10.1007/978-3-031-43907-0\_73
10. Marco Hübner, Leonardo Ayala, Maike Rees, Tim J. Adler, Kris Dreher, Silvia Seidlitz, **Jan Sellner**, Ahmad Bin Qasim, Alexander Seitel, Alexander Studier-Fischer, Alexey Aksenov, Christina Engels, Dogu Teber, Beat Müller-Stich, Felix Nickel, and Lena Maier-Hein. “How to assess the realism of synthetic spectral images”. In: *Molecular-Guided Surgery: Molecules, Devices, and Applications IX*. ed. by Sylvain Gioux, Summer L. Gibbs, and Brian W. Pogue. Vol. PC12361. International Society for Optics and Photonics. SPIE, 2023, PC1236104. DOI: 10.1117/12.2648461
11. Alexander Studier-Fischer, Florian Marc Schwab, Maike Rees, Silvia Seidlitz, **Jan Sellner**, Berkin Özdemir, Leonardo Ayala, Jan Odenthal, Samuel Knoedler, Karl-Friedrich Kowalewski, Caelán Max Haney, Maximilian Dietrich, Gabriel Alexander Salg, Hannes Götz Kenngott, Beat Peter Müller-Stich, Lena Maier-Hein, and Felix Nickel. “ICG-augmented hyperspectral imaging for visualization of intestinal perfusion compared to conventional ICG fluorescence imaging: an experimental study”. In: *International Journal of Surgery* 109.12 (2023). DOI: 10.1097/JS9.0000000000000706
12. Ahmad Bin Qasim, Alessandro Motta, Alexander Studier-Fischer, **Jan Sellner**, Leonardo Ayala, Marco Hübner, Marc Bressan, Berkin Özdemir, Karl Friedrich

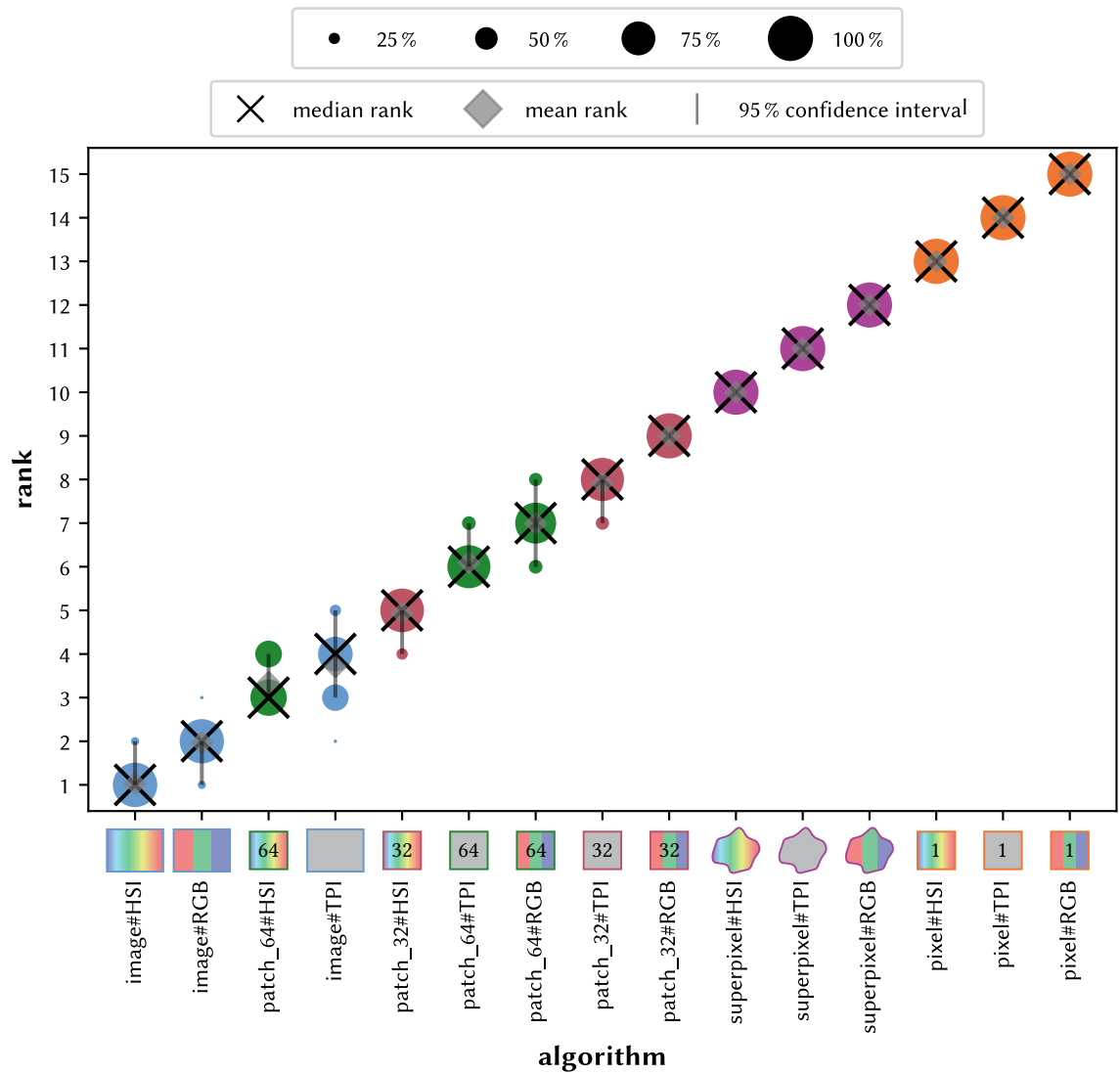
- Kowalewski, Felix Nickel, Silvia Seidlitz, and Lena Maier-Hein. “Test-time augmentation with synthetic data addresses distribution shifts in spectral imaging”. In: *International Journal of Computer Assisted Radiology and Surgery* (Mar. 14, 2024). ISSN: 1861-6429. DOI: 10.1007/s11548-024-03085-3
13. Alexander Baumann, Leonardo Ayala, Alexander Studier-Fischer, **Jan Sellner**, Berkin Özdemir, Karl-Friedrich Kowalewski, Slobodan Ilic, Silvia Seidlitz, and Lena Maier-Hein. *Deep intra-operative illumination calibration of hyperspectral cameras*. Paper under review at the 27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Marrakesh, Morocco. 2024
  14. Viet Tran Ba, Marco Hübner, Ahmad bin Qasim, Maike Rees, **Jan Sellner**, Silvia Seidlitz, Berkin Özdemir, Alexander Studier-Fischer, Felix Nickel, Leonardo Ayala, and Lena Maier-Hein. *Semantic hyperspectral image synthesis for cross-modality knowledge transfer in surgical data science*. Paper under review at the 27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Marrakesh, Morocco. 2024
  15. Leonardo Ayala, Diana Mindroc-Filimon, Maike Rees, Marco Hübner, **Jan Sellner**, Silvia Seidlitz, Minu Tizabi, Sebastian Wirkert, Alexander Seitel, and Lena Maier-Hein. *The SPECTRAL Perfusion Arm Clamping dAtaset (SPECTRALPACA) for video-rate functional imaging of the skin*. Paper under review at Scientific Data. 2024
  16. Alexander Studier-Fischer, Marc Bressan, Ahmad bin Qasim, Berkin Özdemir, **Jan Sellner**, Silvia Seidlitz, Caelán Haney, Luisa Egen, Maurice Michel, Maximilian Dietrich, Gabriel Alexander Salg, Franck Billmann, Henrik Nienhüser, Thilo Hackert, Beat Müller-Stich, Lena Maier-Hein, Felix Nickel, and Karl-Friedrich Kowalewski. *Spectral Characterization of Intraoperative Renal Perfusion using Hyperspectral Imaging and Artificial Intelligence*. Paper under review at Kidney International. 2024
  17. Alexander Studier-Fischer, Berkin Özdemir, Maike Rees, Leonardo Ayala, Silvia Seidlitz, **Jan Sellner**, Karl-Friedrich Kowalewski, Caelán Max Haney, Jan Odentahl, Samuel Knödler, Maximilian Dietrich, Daniel Gruneberg, Thorsten Brenner, Karsten Schmidt, Felix Carl Fabian Schmitt, Markus A. Weigand, Gabriel Alexander Salg, Anna Dupree, Henrik Nienhüser, Arianeb Mehrabi, Thilo Hackert, Beat Müller-Stich, Lena Maier-Hein, and Felix Nickel. *Crystalloid Volume versus Catecholamines for Management of Hemorrhagic Shock during Esophagectomy – Assessment of Microcirculatory Tissue Oxygenation of the Gastric Conduit in a Porcine Model using Hyperspectral Imaging – An Experimental Study*. Paper under review at International Journal of Surgery. 2024



## ADDITIONAL RESULTS

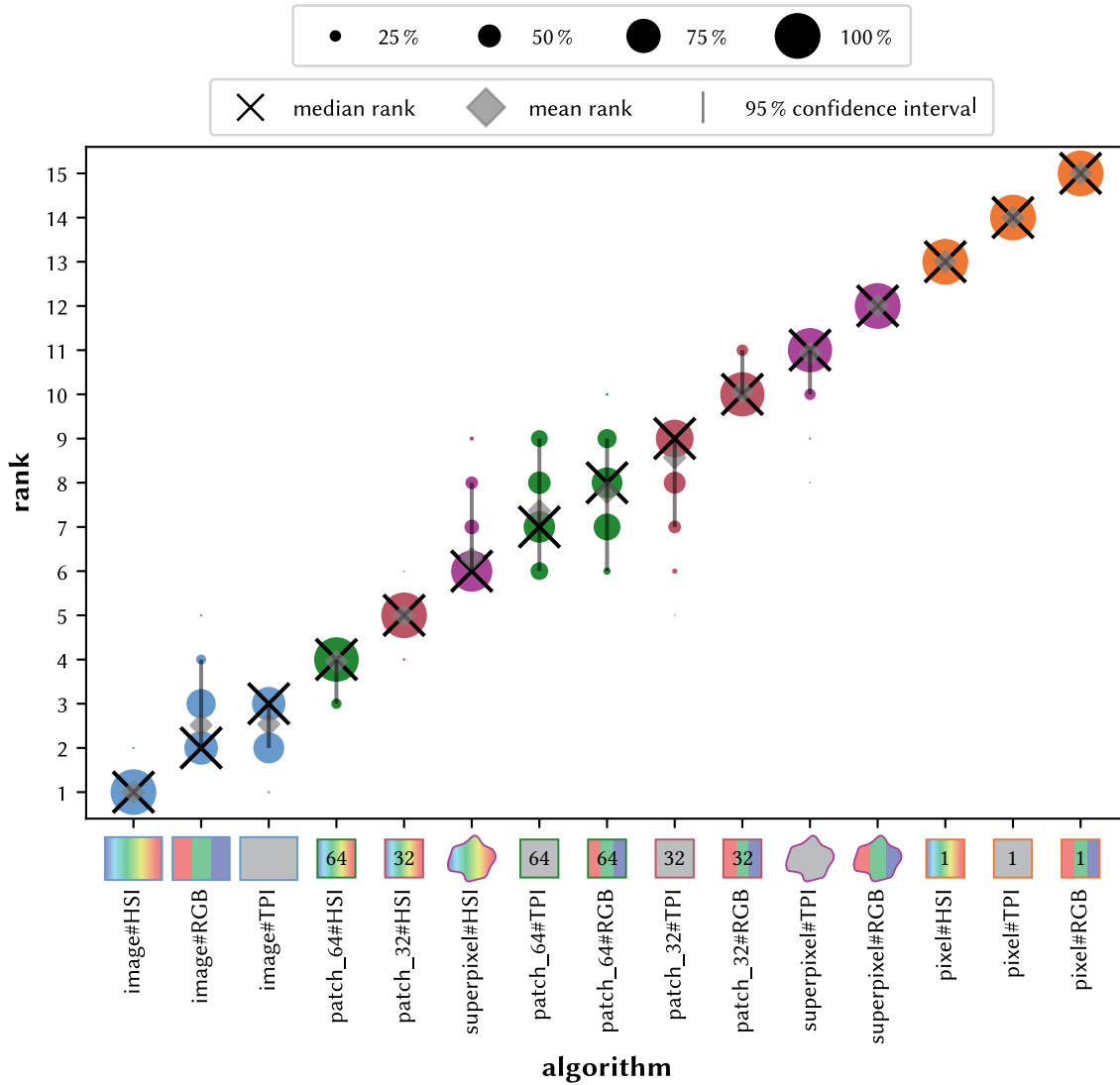
---

B

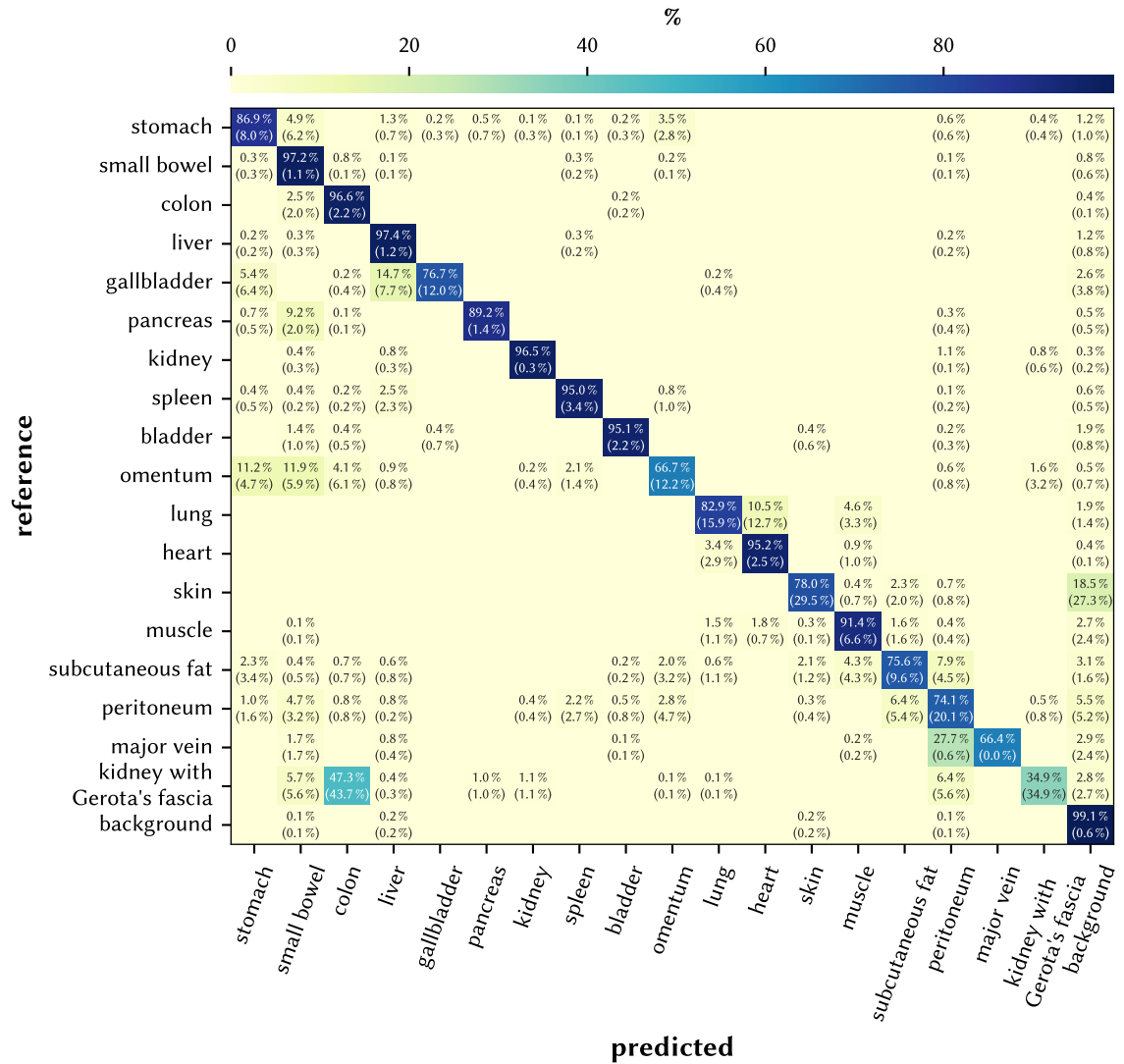


**Figure B.1:** Uncertainty-aware ranking of the different granularities and modalities based on bootstrap sampling on the test set using the normalized surface dice (NSD). The area of each blob is proportional to the relative frequency that the corresponding algorithm achieved the respective rank across 1000 bootstrap samples (concept from [236]). Each bootstrap sample consists of 5 hierarchically aggregated subject-level NSD metric values. The lines encompass the 95 % quartile of the bootstrap results while the cross and the diamond denote the median and mean rank, respectively. Ranking results for the dice similarity coefficient (DSC) and average surface distance (ASD) can be found in Figure 5.15 and Figure B.2, respectively. This figure was adapted from [198].

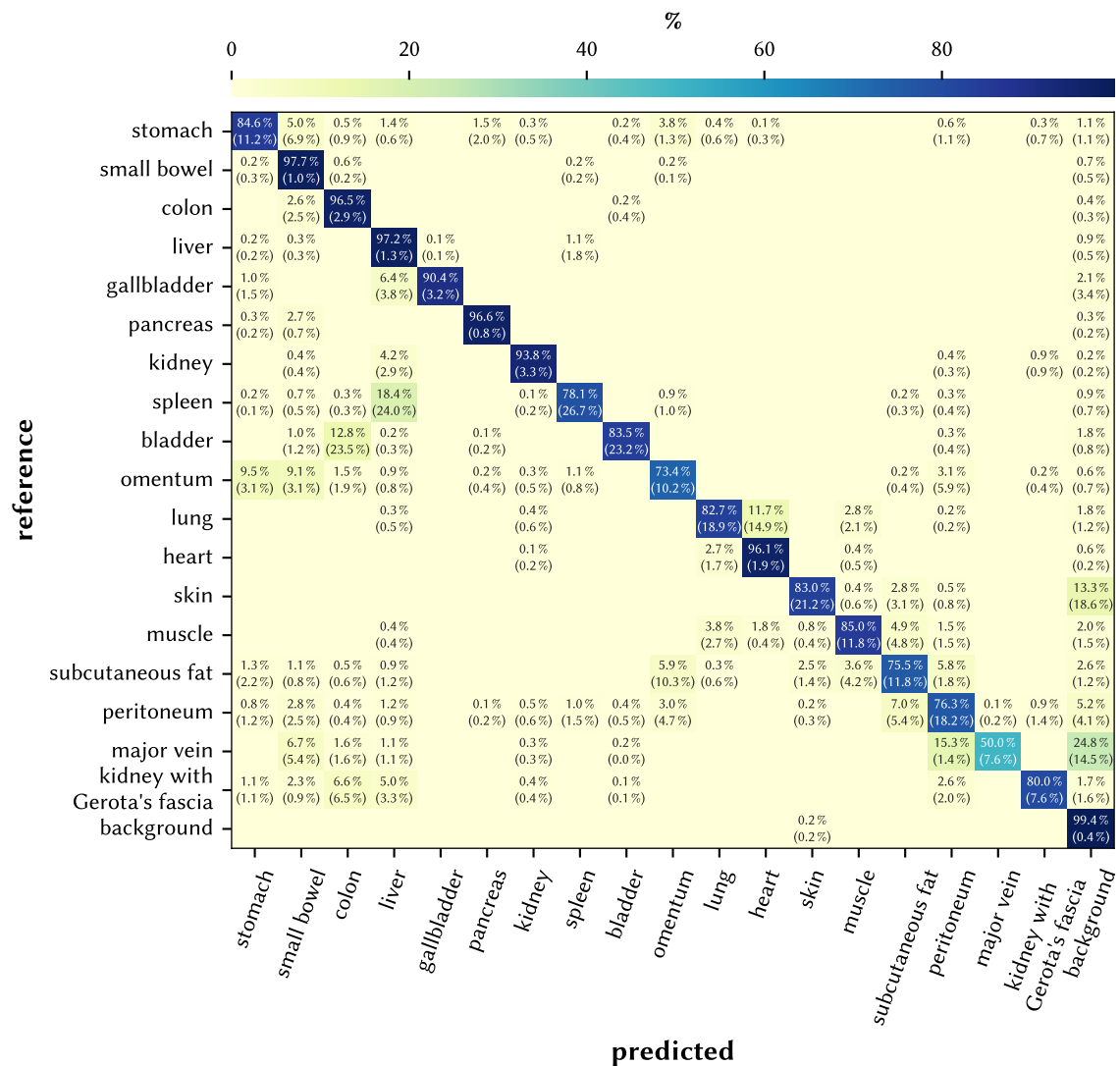




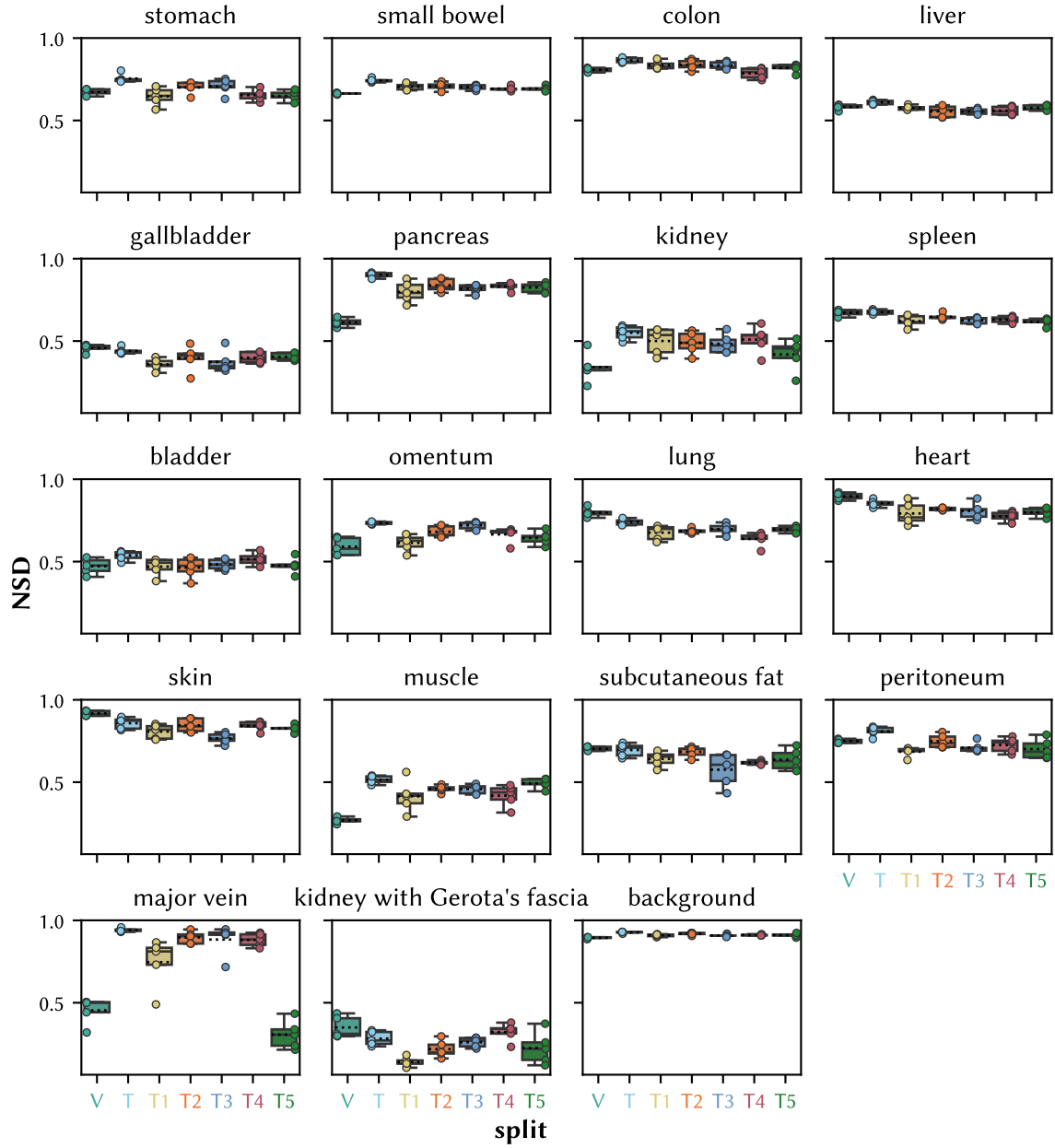
**Figure B.2:** Uncertainty-aware ranking of the different granularities and modalities based on bootstrap sampling on the test set using the average surface distance (ASD). The area of each blob is proportional to the relative frequency that the corresponding algorithm achieved the respective rank across 1000 bootstrap samples (concept from [236]). Each bootstrap sample consists of 5 hierarchically aggregated subject-level ASD metric values. The lines encompass the 95 % quartile of the bootstrap results while the cross and the diamond denote the median and mean rank, respectively. Ranking results for the dice similarity coefficient (DSC) and normalized surface dice (NSD) can be found in Figure 5.15 and Figure B.1, respectively. This figure was adapted from [198].



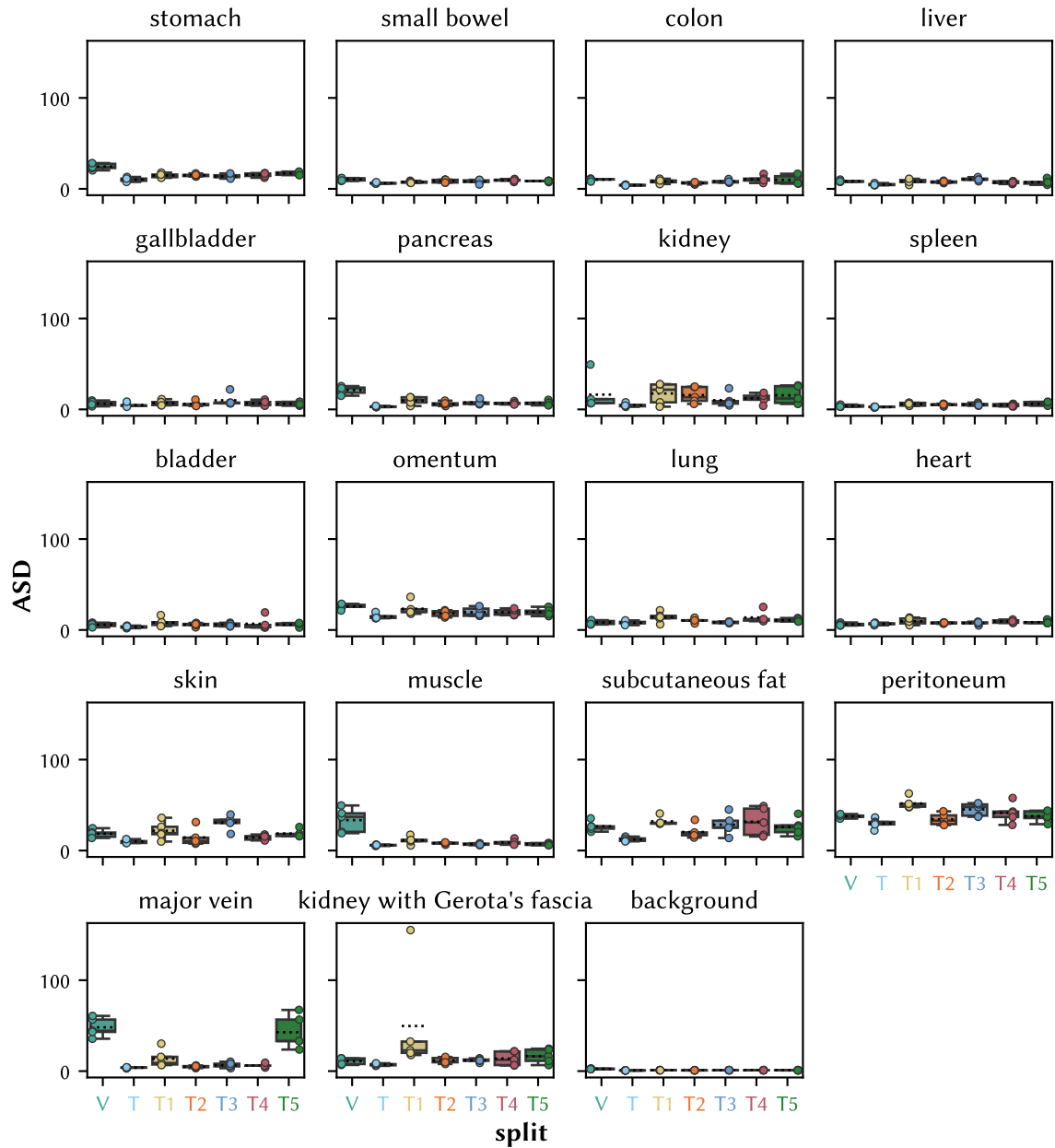
**Figure B.3:** Confusion matrix of the image granularity and tissue parameter images (TPI) modality on the test set. The matrix depicts how pixels from the reference class get classified. That is, every  $(i, j)$ -th entry shows the percentage of pixels from class  $i$  that get classified as class  $j$  (on average). Values < 0.1 % are not shown for brevity. The matrix is row-normalized based on the pixels from all images of one subject and then these matrices are averaged across subjects. The number in brackets denotes the standard deviation across subjects. Numbers on the diagonal denote the recall (sensitivity). Confusion matrices for the hyperspectral imaging (HSI) and RGB modality can be found in Figure 5.17 and Figure B.4, respectively. This figure was adapted from [198].



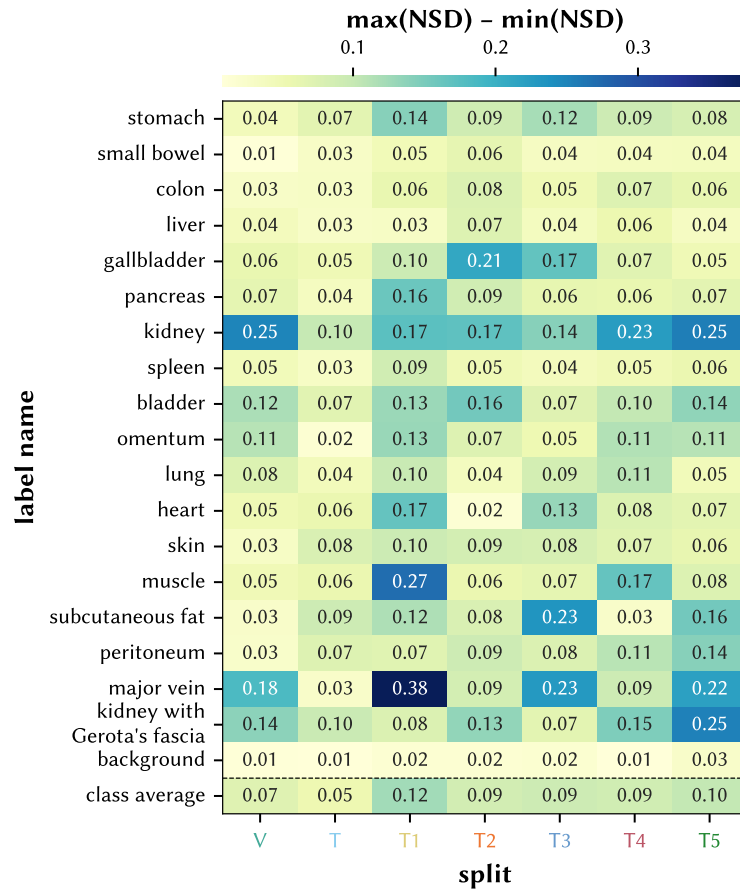
**Figure B.4:** Confusion matrix of the image granularity and RGB modality on the test set. The matrix depicts how pixels from the reference class get classified. That is, every  $(i, j)$ -th entry shows the percentage of pixels from class  $i$  that get classified as class  $j$  (on average). Values  $< 0.1\%$  are not shown for brevity. The matrix is row-normalized based on the pixels from all images of one subject and then these matrices are averaged across subjects. The number in brackets denotes the standard deviation across subjects. Numbers on the diagonal denote the recall (sensitivity). Confusion matrices for the hyperspectral imaging (HSI) and tissue parameter images (TPI) modality can be found in Figure 5.17 and Figure B.3, respectively. This figure was adapted from [198].



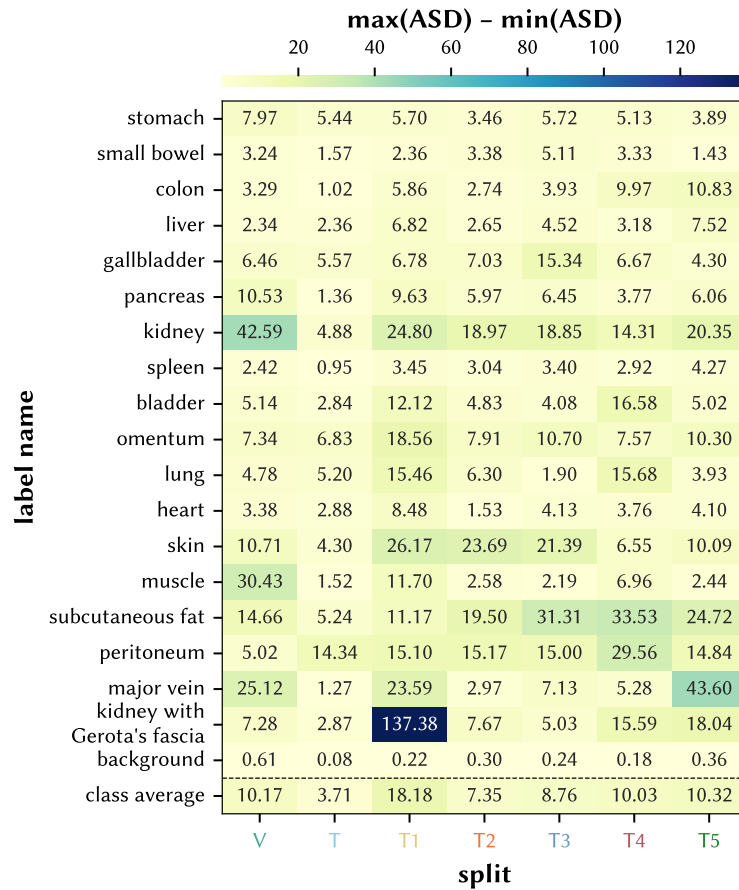
**Figure B.5:** Network variability across five different seed runs (hyperspectral image model) stratified by organ for different splits using the normalized surface dice (NSD). **V** refers to the validation scores (validation\_unknown split), **T** to the test scores with ensembling and **T1** to **T5** to the test scores without ensembling for each of the networks from the five folds of Figure 5.12. Each boxplot shows the interquartile range (IQR) with the median (solid line) and mean (dotted line). The whiskers extend up to 1.5 times of the IQR. Each point represents the aggregated class-level NSD score of one seed run. Figure 5.22 and Figure B.6 show the results for the dice similarity coefficient (DSC) and average surface distance (ASD), respectively.



**Figure B.6:** Network variability across five different seed runs (hyperspectral image model) stratified by organ for different splits using the average surface distance (ASD). **V** refers to the validation scores (validation\_unknown split), **T** to the test scores with ensembling and **T1** to **T5** to the test scores without ensembling for each of the networks from the five folds of Figure 5.12. Each boxplot shows the interquartile range (IQR) with the median (solid line) and mean (dotted line). The whiskers extend up to 1.5 times of the IQR. Each point represents the aggregated class-level ASD score of one seed run. Figure 5.22 and Figure B.5 show the results for the dice similarity coefficient (DSC) and normalized surface dice (NSD), respectively.

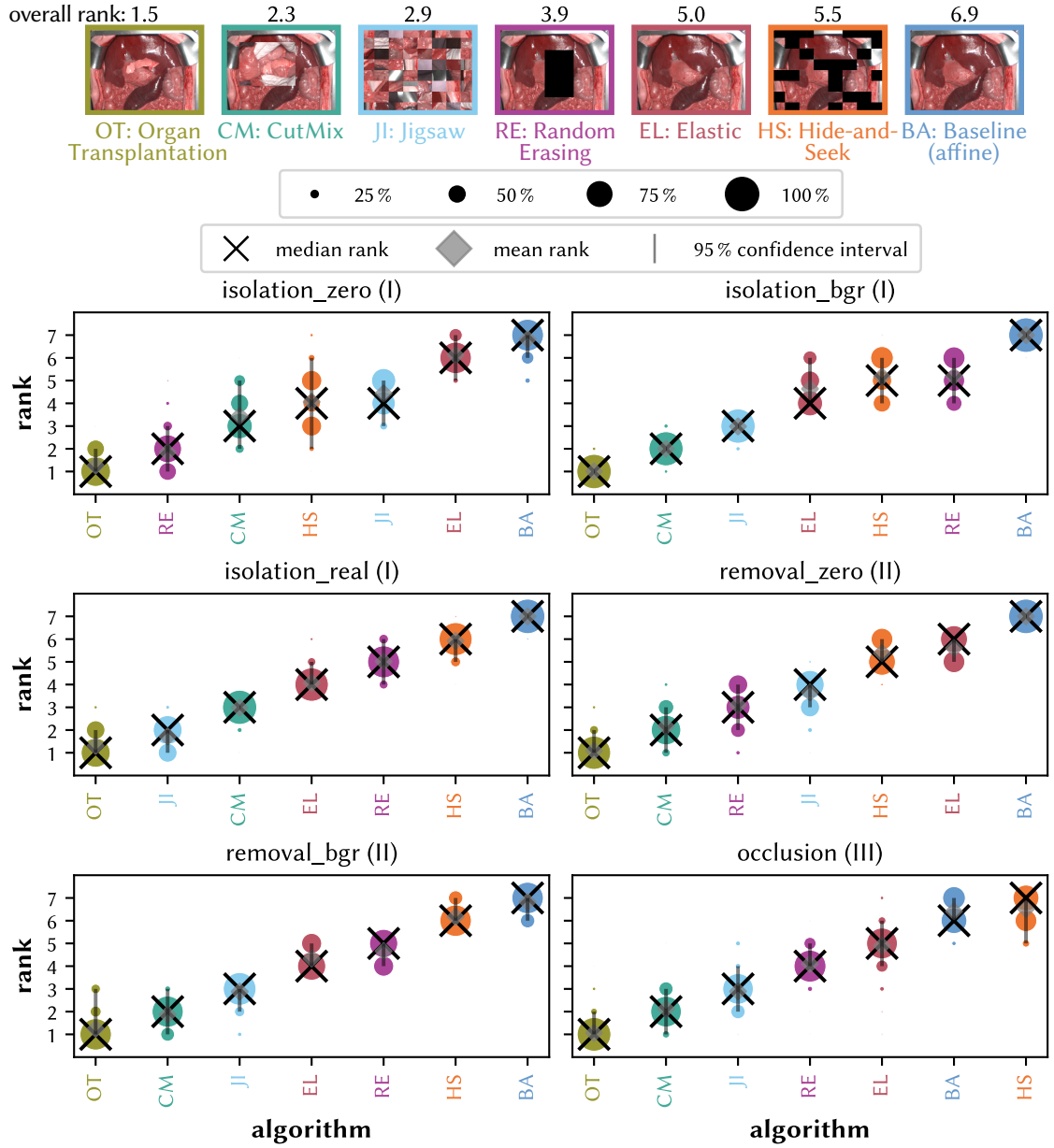


**Figure B.7:** Min-max ranges across five different seed runs stratified by organ for different splits using the normalized surface dice (NSD). **V** refers to the validation scores (validation\_unknown split), **T** to the test scores with ensembling and **T1** to **T5** to the test scores without ensembling for each of the networks from the five folds of Figure 5.12. The last line denotes the average across all classes per split. The hyperspectral image model was trained five times and the difference between the highest and lowest NSD score across the five runs is computed independently for each of the splits. Figure 5.23 and Figure B.8 show the results for the dice similarity coefficient (DSC) and average surface distance (ASD), respectively.



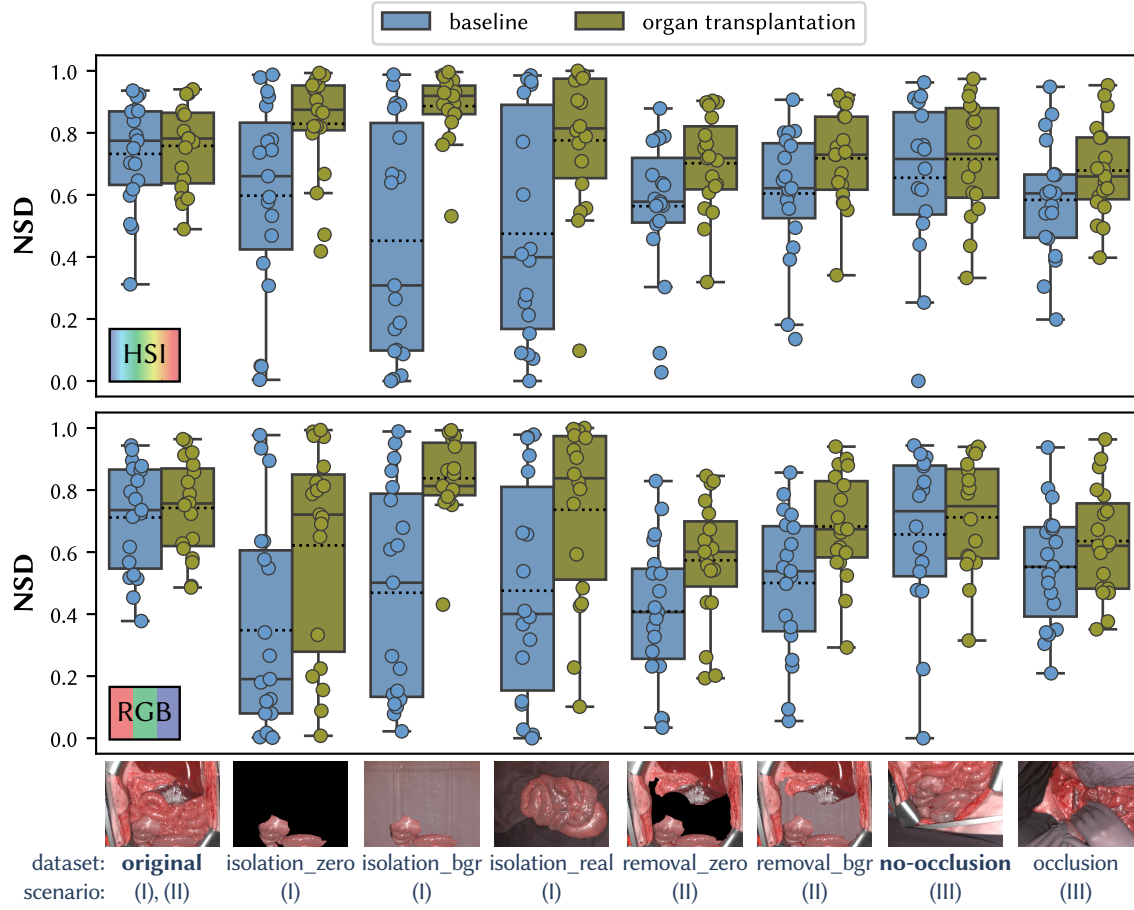
**Figure B.8:** Min-max ranges across five different seed runs stratified by organ for different splits using the average surface distance (ASD). **V** refers to the validation scores (validation\_unknown split), **T** to the test scores with ensembling and **T1** to **T5** to the test scores without ensembling for each of the networks from the five folds of Figure 5.12. The last line denotes the average across all classes per split. The hyperspectral image model was trained five times and the difference between the highest and lowest ASD score across the five runs is computed independently for each of the splits. Figure 5.23 and Figure B.7 show the results for the dice similarity coefficient (DSC) and normalized surface dice (NSD), respectively.

## B Additional Results



**Figure B.9:** Uncertainty-aware ranking of the seven evaluated augmentation methods on the six geometric out-of-distribution datasets using the normalized surface dice (NSD). Consistently across all datasets, the organ transplantation augmentation ranks first whereas the baseline typically ranks last. The area of each blob is proportional to the relative frequency that the corresponding algorithm achieved the respective rank across 1000 bootstrap samples (concept from [236]). Each bootstrap sample consists of 19 hierarchically aggregated class-level NSD metric values. The lines encompass the 95 % quartile of the bootstrap results while the cross and the diamond denote the median and mean rank, respectively. Results for the dice similarity coefficient are in Figure 5.34. This figure was adapted from [202].





**Figure B.10:** Segmentation performance using the normalized surface dice (NSD) for six geometrical out-of-distribution (OOD) datasets and two in-distribution datasets (highlighted in bold) comparing the baseline network with a network trained with the organ transplantation augmentation. Results for the hyperspectral imaging (HSI) (top) and RGB (bottom) modalities are shown. See Section 5.4.1 for a description of scenarios. Each boxplot shows the interquartile range (IQR) with the median (solid line) and mean (dotted line). The whiskers extend up to 1.5 times of the IQR. Each point represents the aggregated class-level performance. Results for the dice similarity coefficient (DSC) are shown in Figure 5.33. This figure was adapted from [202].



## LIST OF ACRONYMS

---

|                        |  |
|------------------------|--|
| <b><i>t</i>-SNE</b>    | <i>t</i> -distributed stochastic neighbor approach     |
| <b>StO<sub>2</sub></b> | tissue oxygen saturation                               |
| <b>ASD</b>             | average surface distance                               |
| <b>CE</b>              | cross-entropy  |
| <b>CNN</b>             | convolutional neural network                           |
| <b>CPU</b>             | central processing unit                                |
| <b>DKFZ</b>            | German Cancer Research Center                          |
| <b>DSC</b>             | dice similarity coefficient                            |
| <b>EDA</b>             | exploratory data analysis                              |
| <b>ELU</b>             | exponential linear unit                                |
| <b>FFCV</b>            | Fast Forward Computer Vision                           |
| <b>GNN</b>             | graph neural network                                   |
| <b>GPU</b>             | graphics processing unit                               |
| <b>HIDSS4Health</b>    | Helmholtz Information & Data Science School for Health |
| <b>HSI</b>             | hyperspectral imaging                                  |
| <b>HTC</b>             | hyperspectral tissue classification                    |
| <b>I/O</b>             | input/output   |
| <b>IMSY</b>            | Intelligent Medical Systems                            |
| <b>IoU</b>             | intersection over union                                |
| <b>IQR</b>             | interquartile range                                    |
| <b>JIT</b>             | just-in-time   |
| <b>LeakyReLU</b>       | leaky rectified linear unit                            |
| <b>LED</b>             | light-emitting diode                                   |

|             |   |
|-------------|---|
| <b>LMM</b>  | linear mixed model                            |
| <b>MITK</b> | Medical Imaging Interaction Toolkit           |
| <b>MSI</b>  | multispectral imaging                         |
| <b>NPI</b>  | near-infrared perfusion index                 |
| <b>NSD</b>  | normalized surface dice                       |
| <b>OHI</b>  | organ hemoglobin index                        |
| <b>OOD</b>  | out-of-distribution                           |
| <b>PCM</b>  | prediction coherence map                      |
| <b>PI</b>   | principal investigator                        |
| <b>RAM</b>  | random-access memory                          |
| <b>ReLU</b> | rectified linear unit                         |
| <b>RQ</b>   | research question                             |
| <b>SD</b>   | standard deviation                            |
| <b>SLIC</b> | simple linear iterative clustering            |
| <b>SSD</b>  | solid-state drive                             |
| <b>SVM</b>  | support vector machine                        |
| <b>SWA</b>  | stochastic weight averaging                   |
| <b>tanh</b> | tangens hyperbolicus                          |
| <b>THI</b>  | tissue hemoglobin index                       |
| <b>TLI</b>  | tissue lipid index                            |
| <b>TPI</b>  | tissue parameter images                       |
| <b>TWI</b>  | tissue water index                            |
| <b>UMAP</b> | uniform manifold approximation and projection |

## LIST OF FIGURES

---

|      |   |    |
|------|---|----|
| 1.1  | Basic concept of hyperspectral imaging (HSI) and exemplary medical applications. . . . .  | 2  |
| 1.2  | Tackling research questions toward the goal of autonomous robotic surgery. . . . .  | 4  |
| 1.2  | Tackling research questions toward the goal of autonomous robotic surgery (continued). . . . .  | 5  |
| 1.3  | Overview of our spectral analysis for tissue discrimination (RQ1). . . . .  | 7  |
| 1.4  | Overview of our open data concept for the HeiPorSPECTRAL dataset (part of RQ1). . . . .   | 8  |
| 1.5  | Overview of our data loading benchmark on segmentation networks for hyperspectral imaging (HSI) data (RQ2). . . . .                                   | 9  |
| 1.6  | Overview of our analysis on segmentation networks for hyperspectral imaging (HSI) data (RQ3). . . . .   | 10 |
| 1.7  | Overview of our analysis on the effect of different domain shifts (RQ4). . . . .  | 12 |
| 1.8  | Overview of our assessment on segmentation networks for geometrical out-of-distribution (OOD) hyperspectral imaging (HSI) data (part of RQ4). . . . . | 13 |
| 2.1  | Internal organs of the human body. . . . .  | 16 |
| 2.2  | Simplified concept of the light-tissue interaction. . . . .   | 20 |
| 2.3  | Effect of hemoglobin on the spectra. . . . .  | 21 |
| 2.4  | Line scanning approach for acquiring hyperspectral imaging data as utilized by the Tivita <sup>®</sup> Tissue camera. . . . .                         | 22 |
| 2.5  | Exemplary filter response functions for an RGB and hyperspectral imaging (HSI) camera system as well as human cone cells. . . . .                     | 23 |
| 2.6  | Exemplary spectra for an light-emitting diode (LED) and halogen light source. . . . .   | 24 |
| 2.7  | Exemplary white and dark measurements from the Tivita <sup>®</sup> Tissue camera. . . . .   | 26 |
| 2.8  | Simple example of the convolution operation. . . . .  | 29 |
| 2.9  | Example of a handcrafted convolution filter. . . . .  | 29 |
| 2.10 | Example of the convolution operation for a convolutional neural network (CNN). . . . .  | 30 |
| 2.11 | Example of activation functions. . . . .  | 32 |
| 2.12 | Overview of the U-Net architecture. . . . .   | 33 |
| 2.13 | Functional principle of the upscaling layer. . . . .  | 34 |

|      |   |    |
|------|---|----|
| 2.14 | Example of binary floating-point computer number formats for float16 and float32 precisions. . . . .  | 35 |
| 2.15 | Range of float16 and float32 numbers. . . . .   | 39 |
| 2.16 | Accuracy of float16 and float32 numbers. . . . .  | 40 |
| 2.17 | Basic principle of autocasting. . . . .   | 41 |
| 2.18 | Distribution of gradients in neural network training without loss scaling in comparison with the representable range of float16. . . . .          | 41 |
| 2.19 | Concept of mixed precision training with loss scaling. . . . .  | 42 |
| 2.20 | Distribution of gradients in neural network training with loss scaling in comparison with the representable range of float16. . . . .             | 43 |
| 2.21 | Comparison of the relative error of gradients from networks with and without loss scaling. . . . .  | 44 |
| 2.22 | Comparison of segmentation networks trained with different precision settings. . . . .  | 45 |
| 3.1  | Number of publications per year in the surgical scene segmentation and medical hyperspectral imaging fields. . . . .                              | 47 |
| 3.2  | Visualization of the different employed data augmentations from the 34 selected publications in the field of surgical scene segmentation. . . . . | 57 |
| 4.1  | Example images and corresponding segmentation masks. . . . .  | 60 |
| 4.2  | Overview of the tissue atlas dataset. . . . .   | 64 |
| 4.3  | Image acquisition protocol for the standardized recordings of the tissue atlas. . . . .   | 65 |
| 4.4  | Overview of the semantic porcine dataset. . . . .   | 66 |
| 4.5  | Distribution of the number of annotated pixels across all images for each organ and dataset. . . . .  | 69 |
| 4.6  | Location maps of the images in the tissue atlas dataset. . . . .  | 70 |
| 4.7  | Location maps of the images in the semantic porcine dataset. . . . .  | 71 |
| 4.8  | Location maps of the images in the semantic human dataset. . . . .  | 72 |
| 4.9  | Deep learning pipeline for spectrum classification. . . . .   | 73 |
| 4.10 | Class imbalances in the tissue atlas dataset and class weighting schemes. . . . .   | 75 |
| 4.11 | Effect of inefficient data loading in high-throughput model training. . . . .   | 77 |
| 4.12 | Causes of inefficient data loading in high-throughput model training. . . . .   | 77 |
| 4.13 | Concept of pinned memory in comparison to paged memory. . . . .   | 79 |
| 4.14 | Concept of the shared, fixed and pinned memory ring buffer. . . . .   | 80 |
| 4.15 | Application of the shared, fixed and pinned memory ring buffer for smaller spatial granularities. . . . .   | 81 |
| 4.16 | Deep learning pipeline for segmentation of hyperspectral imaging (HSI) data. . . . .  | 82 |
| 4.17 | Concept of the organ transplantation augmentation. . . . .  | 88 |

|      |   |     |
|------|---|-----|
| 5.1  | Overview of the $k$ -fold structure of the tissue atlas dataset. . . . .  | 91  |
| 5.2  | Spectral fingerprints of 20 organ classes in the tissue atlas. . . . .  | 92  |
| 5.3  | Visualization of the spectral neighborhood with uniform manifold approximation and projection (UMAP) as a non-linear dimensionality reduction tool of the tissue atlas. . . . . | 93  |
| 5.4  | Confusion matrix of the spectral classification task on the test set (tissue atlas). . . . .  | 95  |
| 5.5  | Landing page of the HeiPorSPECTRAL website. . . . .   | 96  |
| 5.6  | Usage example of the HeiPorSPECTRAL dataset using the hyperspectral tissue classification (HTC) framework. . . . .  | 97  |
| 5.7  | Interactive figure for an example image with a stomach annotation from the HeiPorSPECTRAL dataset. . . . .  | 98  |
| 5.8  | Example of an organ profile page with aggregated information of the entire HeiPorSPECTRAL dataset for the small bowel class. . . . .  | 99  |
| 5.9  | Spectral comparison of the Tivita <sup>®</sup> Tissue camera with a point spectrometer. . . . .   | 101 |
| 5.10 | Benchmarking results of different data loading strategies. . . . .  | 103 |
| 5.11 | Hardware utilization of different data loading strategies. . . . .  | 103 |
| 5.12 | Overview of the $k$ -fold structure of the semantic porcine dataset. . . . .  | 104 |
| 5.13 | Visualization of the hierarchical aggregation schemes of metric scores. . . . .   | 109 |
| 5.14 | Model segmentation performance of different spatial granularities and modalities. . . . .   | 111 |
| 5.15 | Uncertainty-aware ranking of the different granularities and modalities using the dice similarity coefficient (DSC). . . . .  | 113 |
| 5.16 | Ranking stability for the different granularities and modalities. . . . .   | 114 |
| 5.17 | Confusion matrix of the image granularity and hyperspectral imaging (HSI) modality on the test set. . . . .   | 116 |
| 5.18 | Recall of the image model stratified by label and modality . . . . .  | 117 |
| 5.19 | Example predictions for the different spatial granularities of the hyperspectral imaging (HSI) modality. . . . .  | 118 |
| 5.20 | Segmentation performance for all spatial granularities on the test set as a function of the number of training subjects. . . . .  | 119 |
| 5.21 | Network variability across five different seed runs on the test set. . . . .  | 121 |
| 5.22 | Network variability across five different seed runs stratified by organ for different splits using the dice similarity coefficient. . . . .                                     | 122 |
| 5.23 | Min-max ranges across five different seed runs stratified by organ for different splits using the dice similarity coefficient. . . . .  | 123 |
| 5.24 | Example prediction coherence maps of five different seed runs. . . . .  | 125 |
| 5.25 | Cumulative distribution of border distances from pixels with incoherent network predictions based on five different seed runs. . . . .  | 126 |
| 5.26 | Overview of the $k$ -fold structure of the semantic human dataset. . . . .  | 128 |
| 5.27 | Validation strategy for the geometrical out-of-distribution (OOD) experiments. . . . .  | 129 |

|      |  |     |
|------|--|-----|
| 5.28 | Visualization of the aggregation scheme of metric scores for the removal scenario. . . . .   | 132 |
| 5.29 | Sources of variation of hyperspectral imaging (HSI) data from the tissue atlas. . . . .  | 133 |
| 5.30 | Generalization error over training time. . . . .   | 135 |
| 5.31 | Class neighborhood matrix for the semantic porcine dataset on the test split. . . . .  | 137 |
| 5.32 | Change in performance of the image network upon encountering class removals. . . . .   | 138 |
| 5.33 | Segmentation performance using the dice similarity coefficient (DSC) for geometrical out-of-distribution (OOD) datasets comparing the baseline network with a network trained with the organ transplantation augmentation. . . . . | 139 |
| 5.34 | Uncertainty-aware ranking of the seven evaluated augmentation methods using the dice similarity coefficient (DSC). . . . .   | 141 |
| 5.35 | Example predictions comparing the baseline network with the organ transplantation augmentation on all out-of-distribution (OOD) datasets. . . . .  | 142 |
| 5.36 | Comparison of spectral fingerprints for the semantic porcine and semantic human datasets. . . . .  | 143 |
| 5.37 | Nearest neighbor class matrix for the semantic porcine and semantic human datasets based on media spectra comparisons. . . . .   | 144 |
| 5.38 | Segmentation performance on the semantic human dataset using the dice similarity coefficient. . . . .  | 146 |
| 6.1  | Impact of the hyperparameters from the grid search of the tissue atlas classification task. . . . .  | 149 |
| 6.2  | Results for the class imbalance hyperparameters from the grid search of the tissue atlas dataset classification task. . . . .  | 150 |
| 6.3  | Potential thresholds for the normalized surface dice (NSD) metric by using different aggregation functions. . . . .  | 153 |
| 6.4  | Ranking differences across bootstrap samples when optimizing the learning rate. . . . .  | 155 |
| 6.5  | Performance limit of the superpixel approach by taking into account the reference annotations. . . . .   | 157 |
| 6.6  | Comparison of a real vs. a reconstructed RGB image from our hyperspectral imaging system. . . . .  | 161 |
| B.1  | Uncertainty-aware ranking of the different granularities and modalities using the normalized surface dice (NSD). . . . .   | 180 |
| B.2  | Uncertainty-aware ranking of the different granularities and modalities using the average surface distance (ASD). . . . .  | 181 |



|      |   |     |
|------|---|-----|
| B.3  | Confusion matrix of the image granularity and tissue parameter images (TPI) modality on the test set. . . . .   | 182 |
| B.4  | Confusion matrix of the image granularity and RGB modality on the test set. . . . .   | 183 |
| B.5  | Network variability across five different seed runs stratified by organ for different splits using the normalized surface dice. . . . .   | 184 |
| B.6  | Network variability across five different seed runs stratified by organ for different splits using the average surface distance. . . . .  | 185 |
| B.7  | Min-max ranges across five different seed runs stratified by organ for different splits using the normalized surface dice. . . . .  | 186 |
| B.8  | Min-max ranges across five different seed runs stratified by organ for different splits using the average surface distance. . . . .   | 187 |
| B.9  | Uncertainty-aware ranking of the seven evaluated augmentation methods using the normalized surface dice (NSD). . . . .  | 188 |
| B.10 | Segmentation performance using the normalized surface dice (NSD) for geometrical out-of-distribution (OOD) datasets comparing the baseline network with the organ transplantation augmentation. . . . . | 189 |



## LIST OF TABLES

---

|     |  |     |
|-----|--|-----|
| 1.1 | Outline and corresponding research questions of this thesis. . . . .   | 14  |
| 2.1 | Basic information about binary floating-point numbers with float16 and float32 precisions. . . . .                             | 37  |
| 3.1 | Summary of basic information about publications that analyzed spectral characteristics of visceral organs. . . . .             | 48  |
| 3.2 | Summary of basic information about publications that tackle biomedical segmentation problems. . . . .                          | 53  |
| 3.3 | Overview of employed data augmentations from 34 selected publications in the field of surgical scene segmentation. . . . .     | 56  |
| 4.1 | Dataset statistics of the semantic human dataset. . . . .  | 67  |
| 4.2 | Specification and results of the hyperparameter search for the spectrum classification network. . . . .                        | 76  |
| 4.3 | Epoch and batch sizes for the spatial granularity models. . . . .  | 85  |
| 5.1 | Distribution of training, validation and test datasets across our three geometric out-of-distribution (OOD) scenarios. . . . . | 130 |
| 5.2 | Optimal probability of applying an augmentation based on the grid search results. . . . .                                      | 131 |



## LIST OF ALGORITHMS

---

|     |   |    |
|-----|---|----|
| 4.1 | Detailed description of the organ transplantation augmentation. . . . . | 88 |
|-----|---|----|



## BIBLIOGRAPHY

---

- [1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. “SLIC Superpixels Compared to State-of-the-Art Superpixel Methods”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (Nov. 2012). Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 2274–2282. ISSN: 1939-3539. DOI: 10 . 1109 / TPAMI . 2012 . 120 (cit. on pp. 54, 83).
- [2] Abien Fred Agarap. “Deep learning using rectified linear units (relu)”. In: *arXiv preprint arXiv:1803.08375* (2018) (cit. on pp. 31, 32).
- [3] SuperAnnotate AI. *SuperAnnotate | Empowering Enterprises with Custom LLM/GenAI/CV Models*. URL: <https://www.superannotate.com> (visited on 11/17/2023) (cit. on p. 63).
- [4] Alex Aizman, Gavin Maltby, and Thomas Breuel. “High Performance I/O For Large Scale Deep Learning”. In: *2019 IEEE International Conference on Big Data (Big Data)* (2019), pp. 5965–5967. URL: <https://api.semanticscholar.org/CorpusID:210023710> (cit. on p. 50).
- [5] Hamed Akbari, Yukio Kosugi, Kazuyuki Kojima, and Naofumi Tanaka. “Detection and Analysis of the Intestinal Ischemia Using Visible and Invisible Hyperspectral Imaging”. In: *IEEE Transactions on Biomedical Engineering* 57.8 (2010), pp. 2011–2017. DOI: 10 . 1109 / TBME . 2010 . 2049110 (cit. on p. 2).
- [6] Hamed Akbari, Yukio Kosugi, Kazuyuki Kojima, and Naofumi Tanaka. “Wavelet-Based Compression and Segmentation of Hyperspectral Images in Surgery”. In: *Medical Imaging and Augmented Reality*. Ed. by Takeyoshi Dohi, Ichiro Sakuma, and Hongen Liao. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008, pp. 142–149. ISBN: 978-3-540-79982-5. DOI: 10 . 1007 / 978 - 3 - 540 - 79982 - 5\_16 (cit. on pp. 52, 53, 56, 58).
- [7] Hamed Akbari, Kuniaki Uto, Yukio Kosugi, Kazuyuki Kojima, and Naofumi Tanaka. “Cancer detection using infrared hyperspectral imaging”. In: *Cancer Science* 102.4 (2011), pp. 852–857. DOI: 10 . 1111 / j . 1349 - 7006 . 2011 . 01849 . x (cit. on p. 2).

- [8] Hassan Al Hajj, Mathieu Lamard, Pierre-Henri Conze, Soumali Roychowdhury, Xiaowei Hu, Gabija Maršalkaitė, Odysseas Zisimopoulos, Muneer Ahmad Dedmari, Fenqiang Zhao, Jonas Prellberg, Manish Sahu, Adrian Galdran, Teresa Araújo, Duc My Vo, Chandan Panda, Navdeep Dahiya, Satoshi Kondo, Zhengbing Bian, Arash Vahdat, Jonas Bialopetravičius, Evangello Flouty, Chenhui Qiu, Sabrina Dill, Anirban Mukhopadhyay, Pedro Costa, Guilherme Aresta, Senthil Ramamurthy, Sang-Woong Lee, Aurélio Campilho, Stefan Zachow, Shunren Xia, Sailesh Conjeti, Danail Stoyanov, Jogundas Armaitis, Pheng-Ann Heng, William G. Macready, Béatrice Cochener, and Gwenolé Quellec. “CATARACTS: Challenge on automatic tool annotation for cataRACT surgery”. In: *Medical Image Analysis* 52 (Feb. 2019), pp. 24–41. ISSN: 1361-8415. DOI: 10 . 1016 / j . media . 2018 . 11 . 008 (cit. on p. 51).
- [9] Fahim Irfan Alam, Jun Zhou, Alan Wee-Chung Liew, and Xiuping Jia. “CRF learning with CNN features for hyperspectral image segmentation”. In: *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Beijing, China: IEEE, July 2016, pp. 6890–6893. ISBN: 978-1-5090-3332-4. DOI: 10 . 1109 / IGA RSS . 2016 . 7730798. URL: <http://ieeexplore.ieee.org/document/7730798> (cit. on p. 52).
- [10] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, Avinash Kori, Varghese Alex, Ganapathy Krishnamurthi, David Rauber, Robert Mendel, Christoph Palm, Sophia Bano, Guinther Saibro, Chi-Sheng Shih, Hsun-An Chiang, Juntang Zhuang, Junlin Yang, Vladimir Iglovikov, Anton Dobrenkii, Madhu Reddiboina, Anubhav Reddy, Xingtong Liu, Cong Gao, Mathias Unberath, Myeonghyeon Kim, Chanh Kim, Chaewon Kim, Hyejin Kim, Gyeongmin Lee, Ihsan Ullah, Miguel Luna, Sang Hyun Park, Mahdi Azizian, Danail Stoyanov, Lena Maier-Hein, and Stefanie Speidel. “2018 Robotic Scene Segmentation Challenge”. In: *arXiv:2001.11190 [cs]* (Aug. 2020). arXiv: 2001.11190. URL: <http://arxiv.org/abs/2001.11190> (cit. on pp. 51, 56, 159).
- [11] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, Luis Herrera, Wenqi Li, Vladimir Iglovikov, Huoling Luo, Jian Yang, Danail Stoyanov, Lena Maier-Hein, Stefanie Speidel, and Mahdi Azizian. *2017 Robotic Instrument Segmentation Challenge*. 2019. arXiv: 1902 . 06426 [cs.CV] (cit. on p. 56).
- [12] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, AnnetteKopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers,



---

Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Goli Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, Henkjan Huisman, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbelaez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Namkug Kim, Ildoo Kim, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, et al. “The Medical Segmentation Decathlon”. In: *arXiv:2106.05735 [cs, eess]* (June 2021). arXiv: 2106.05735. URL: <http://arxiv.org/abs/2106.05735> (cit. on p. 152).

- [13] Caerwyn Ash, Michael Dubec, Kelvin Donne, and Tim Bashford. “Effect of wavelength and beam width on penetration in light-tissue interaction using computational methods”. In: *Lasers in Medical Science* 32.8 (Nov. 1, 2017), pp. 1909–1918. ISSN: 1435-604X. DOI: 10.1007/s10103-017-2317-4 (cit. on p. 20).
- [14] Leonardo Ayala, Tim J. Adler, Silvia Seidlitz, Sebastian Wirkert, Christina Engels, Alexander Seitel, Jan Sellner, Alexey Aksenov, Matthias Bodenbach, Pia Bader, Sebastian Baron, Anant Vemuri, Manuel Wiesenfarth, Nicholas Schreck, Diana Mindroc, Minu Tizabi, Sebastian Pirmann, Brittaney Everitt, Annette Kopp-Schneider, Dogu Teber, and Lena Maier-Hein. “Spectral imaging enables contrast agent-free real-time ischemia monitoring in laparoscopic surgery”. In: *Science Advances* 9.10 (2023), eadd6778. DOI: 10.1126/sciadv.add6778 (cit. on p. 160).
- [15] Leonardo Ayala, Silvia Seidlitz, Anant Vemuri, Sebastian J Wirkert, Thomas Kirchner, Tim J Adler, Christina Engels, Dogu Teber, and Lena Maier-Hein. “Light source calibration for multispectral imaging in surgery”. In: *Int J Comput Assist Radiol Surg* 15.7 (June 2020), pp. 1117–1125 (cit. on p. 169).
- [16] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.12 (2017), pp. 2481–2495. DOI: 10.1109/TPAMI.2016.2644615 (cit. on p. 54).
- [17] Elisabeth J. M. Baltussen, Esther N. D. Kok, Susan G. Brouwer de Koning, Joyce Sanders, Arend G. J. Aalbers, Niels F. M. Kok, Geerard L. Beets, Claudie C. Flohil, Sjoerd C. Bruin, Koert F. D. Kuhlmann, Henricus J. C. M. Sterenborg, and Theo J. M. Ruers. “Hyperspectral imaging for tissue classification, a way toward smart laparoscopic colorectal surgery”. In: *Journal of Biomedical Optics* 24.1 (Jan. 2019), p. 016002. DOI: 10.1117/1.JBO.24.1.016002 (cit. on pp. 48, 49).
- [18] Philip Baum, Johannes Diers, Sven Lichthardt, Carolin Kastner, Nicolas Schlegel, Christoph-Thomas Germer, and Armin Wiegering. “Mortality and Complications Following Visceral Surgery: A Nationwide Analysis Based on the Diagnostic Categories Used in German Hospital Invoicing Data”. In: *Deutsches Arzteblatt international* 116.44 (Nov. 2019), pp. 739–746. ISSN: 1866-0452. DOI: 10.3238/

- arztebl . 2019 . 0739. URL: <https://europepmc.org/articles/PMC6912125> (cit. on p. 16).
- [19] Catherine Bernier. “The next generation of healthcare: The potential of surgical robots”. In: *#HowToRobot* (Aug. 14, 2023). URL: <https://howtorobot.com/expert-insight/next-generation-healthcare-potential-surgical-robots> (visited on 12/14/2023) (cit. on p. 1).
- [20] Franck Billmann and Tobias Keck. *Essentials of visceral surgery: For residents and fellows*. Cham, Switzerland: Springer Nature, Mar. 2023. ISBN: 978-3-662-66735-4. DOI: 10.1007/978-3-662-66735-4 (cit. on p. 16).
- [21] Blosc Development Team. *A fast, compressed and persistent data store library*. 2009–2023. URL: <https://blosc.org> (visited on 11/22/2023) (cit. on p. 78).
- [22] Christine Boev and Elizabeth Kiss. “Hospital-Acquired Infections: Current Trends and Prevention”. In: *Crit Care Nurs Clin North Am* 29.1 (Dec. 2016), pp. 51–65. DOI: 10.1016/j.cnc.2016.09.012 (cit. on p. 16).
- [23] Rositsa Bogdanova, Pierre Boulanger, and Bin Zheng. “Depth Perception of Surgeons in Minimally Invasive Surgery”. In: *Surg Innov* 23.5 (Mar. 2016), pp. 515–524. DOI: 10.1177/1553350616639141 (cit. on p. 17).
- [24] Doug Bonderud. “What Does the Future Hold for Robotic Surgery?” In: *HealthTech* (Feb. 4, 2021). URL: <https://healthtechmagazine.net/article/2021/02/what-does-future-hold-robotic-surgery-perfcon> (visited on 12/14/2023) (cit. on p. 1).
- [25] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. “A Training Algorithm for Optimal Margin Classifiers”. In: *Proceedings of the fifth annual workshop on Computational learning theory* (1992), pp. 144–152 (cit. on p. 54).
- [26] Sakshi Bramhe and Swanand S Pathak. “Robotic Surgery: A Narrative Review”. In: *Cureus* 14.9 (Sept. 2022), e29179. DOI: 10.7759/cureus.29179 (cit. on p. 17).
- [27] Wilhelm Burger and Mark J. Burge. *Digital Image Processing: An Algorithmic Introduction*. Springer International Publishing, 2022. ISBN: 978-3-031-05744-1. DOI: 10.1007/978-3-031-05744-1 (cit. on pp. 27, 28).
- [28] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. “Albumentations: Fast and Flexible Image Augmentations”. In: *Information* 11.2 (Feb. 2020), p. 125. DOI: 10.3390/info11020125. URL: <https://www.mdpi.com/2078-2489/11/2/125> (cit. on p. 85).

- 
- [29] Leopoldo C. Cancio. “Application of novel hyperspectral imaging technologies in combat casualty care”. In: *Emerging Digital Micromirror Device Based Systems and Applications II*. Ed. by Michael R. Douglass and Larry J. Hornbeck. Vol. 7596. International Society for Optics and Photonics. SPIE, 2010, p. 759605. DOI: 10.1117/12.846331 (cit. on p. 2).
- [30] Leopoldo C. Cancio, Andriy I. Batchinsky, James R. Mansfield, Svetlana Panasyuk, Katherine Hetz, David Martini, Bryan S. Jordan, Brian Tracey, and Jenny E. Freeman. “Hyperspectral Imaging: A New Approach to the Diagnosis of Hemorrhagic Shock”. In: *Journal of Trauma and Acute Care Surgery* 60.5 (2006). ISSN: 2163-0755. DOI: 10.1097/01.ta.0000217357.10617.3d. URL: [https://journals.lww.com/jtrauma/fulltext/2006/05000/hyperspectral\\_imaging\\_a\\_new\\_approach\\_to\\_the.25.aspx](https://journals.lww.com/jtrauma/fulltext/2006/05000/hyperspectral_imaging_a_new_approach_to_the.25.aspx) (cit. on p. 2).
- [31] M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murray, Andriy Myronenko, Can Zhao, Dong Yang, Vishwesh Nath, Yufan He, Ziyue Xu, Ali Hatamizadeh, Wentao Zhu, Yun Liu, Mingxin Zheng, Yucheng Tang, Isaac Yang, Michael Zephyr, Behrooz Hashemian, Sachidanand Alle, Mohammad Zalbagi Darestani, Charlie Budd, Marc Modat, Tom Vercauteren, Guotai Wang, Yiwen Li, Yipeng Hu, Yunguan Fu, Benjamin Gorman, Hans Johnson, Brad Genereaux, Barbaros S. Erdal, Vikash Gupta, Andres Diaz-Pinto, Andre Dourson, Lena Maier-Hein, Paul F. Jaeger, Michael Baumgartner, Jayashree Kalpathy-Cramer, Mona Flores, Justin Kirby, Lee A.D. Cooper, Holger R. Roth, Daguang Xu, David Bericat, Ralf Floca, S. Kevin Zhou, Haris Shuaib, et al. “MONAI: An open-source framework for deep learning in healthcare”. In: (Nov. 2022). DOI: 10.48550/arXiv.2211.02701 (cit. on pp. 105, 106).
- [32] Fernando Cervantes-Sanchez, Marianne Maktabi, Hannes Köhler, Robert Sucher, Nada Rayes, Juan Gabriel Avina-Cervantes, Ivan Cruz-Aceves, and Claire Chalopin. “Automatic tissue segmentation of hyperspectral images in liver and head neck surgeries using machine learning”. In: *Artificial Intelligence Surgery* 1 (Aug. 2021), pp. 22–37. DOI: 10.20517/ais.2021.05. URL: <https://aisjournal.net/article/view/4291> (cit. on pp. 53–56, 58).
- [33] Bhaskar Chakravorti. “Why AI Failed to Live Up to Its Potential During the Pandemic”. In: *Harvard Business Review* (Mar. 17, 2022). URL: <https://hbr.org/2022/03/why-ai-failed-to-live-up-to-its-potential-during-the-pandemic> (visited on 01/03/2024) (cit. on p. 3).
- [34] Claire Chalopin, Felix Nickel, Annekatrin Pfahl, Hannes Köhler, Marianne Maktabi, René Thieme, Robert Sucher, Boris Jansen-Winkel, Alexander Studier-Fischer, Silvia Seidlitz, Lena Maier-Hein, Thomas Neumuth, Andreas Melzer, Beat Peter Müller-Stich, and Ines Gockel. “Künstliche Intelligenz und hyperspektrale Bildgebung zur bildgestützten Assistenz in der minimal-invasiven Chirurgie”. In:

- Die Chirurgie* 93.10 (Oct. 1, 2022), pp. 940–947. ISSN: 2731-698X. DOI: 10.1007/s00104-022-01677-w (cit. on pp. 1, 17).
- [35] Zhang Ya-chao, Ye Xin, Xia Zhi-wei, Sui Long, and Fang Wei. “Spectral irradiance degradation model of halogen tungsten lamps at wavelength from 400 nm to 1300 nm”. In: *Chinese Optics* 15.4 (2022), pp. 825–834. ISSN: 2097-1842. DOI: 10.37188/CO.EN.2021-0011 (cit. on p. 25).
  - [36] Raymond Chen. *Understanding what significant digits really mean*. June 16, 2006. URL: <https://devblogs.microsoft.com/oldnewthing/20060616-09/?p=30843> (visited on 01/29/2024) (cit. on p. 37).
  - [37] Zitian Chen, Yanwei Fu, Kaiyu Chen, and Yu-Gang Jiang. “Image Block Augmentation for One-Shot Learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 3379–3386. ISSN: 2374-3468, 2159-5399 (cit. on p. 131).
  - [38] *Choosing the best light source for your fluorescence experiment*. URL: <https://www.scientifica.uk.com/learning-zone/choosing-the-best-light-source-for-your-experiment> (visited on 01/24/2024) (cit. on p. 62).
  - [39] Neil T Clancy, Shobhit Arya, Danail Stoyanov, Mohan Singh, George B Hanna, and Daniel S Elson. “Intraoperative measurement of bowel oxygen saturation using a multispectral imaging laparoscope”. In: *Biomed Opt Express* 6.10 (Sept. 2015), pp. 4179–4190. DOI: 10.1364/BOE.6.004179 (cit. on pp. 18, 19).
  - [40] Neil T. Clancy, Geoffrey Jones, Lena Maier-Hein, Daniel S. Elson, and Danail Stoyanov. “Surgical spectral imaging”. In: *Medical Image Analysis* 63 (2020), p. 101699. ISSN: 1361-8415. DOI: 10.1016/j.media.2020.101699. URL: <https://www.sciencedirect.com/science/article/pii/S1361841520300645> (cit. on pp. 19, 48).
  - [41] Ela Claridge and Džena Hidović-Rowe. “Model based inversion for deriving maps of histological parameters characteristic of cancer from ex-vivo multispectral images of the colon”. In: *IEEE Trans Med Imaging* 33.4 (Nov. 2013), pp. 822–835. DOI: 10.1109/TMI.2013.2290697 (cit. on p. 25).
  - [42] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. “Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)”. In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2016. URL: <http://arxiv.org/abs/1511.07289> (cit. on pp. 31, 32, 73, 83).
  - [43] Yann Collet and Murray Kucherawy. *Zstandard Compression and the 'application/zstd' Media Type*. RFC 8878. Feb. 2021. DOI: 10.17487/RFC8878. URL: <https://www.rfc-editor.org/info/rfc8878> (cit. on p. 78).

- 
- [44] Toby Collins, Adrien Bartoli, Nicolas Bourdel, and Michel Canis. “Segmenting the Uterus in Monocular Laparoscopic Images without Manual Input”. In: Cham: Springer, 2015. doi: 10.1007/978-3-319-24574-4\_22 (cit. on pp. 51, 56).
- [45] Wikipedia contributors. *Exploratory data analysis — Wikipedia, The Free Encyclopedia*. 2023. URL: [https://en.wikipedia.org/wiki/Exploratory\\_data\\_analysis](https://en.wikipedia.org/wiki/Exploratory_data_analysis) (visited on 11/22/2023) (cit. on p. 94).
- [46] NVIDIA Corporation. *Convolutional Layers User’s Guide*. Feb. 1, 2023. URL: <https://docs.nvidia.com/deeplearning/performance/dl-performance-convolutional/index.html> (visited on 02/08/2024) (cit. on p. 28).
- [47] NVIDIA Corporation. *CUDA C++ Programming Guide*. 2007–2023. URL: <https://docs.nvidia.com/cuda/cuda-c-programming-guide> (visited on 11/22/2023) (cit. on p. 78).
- [48] NVIDIA Corporation. *CUDA Toolkit Documentation*. Nov. 15, 2023. URL: [https://docs.nvidia.com/cuda/cuda-runtime-api/group\\_\\_CUDART\\_\\_MEMORY.html%5C#group\\_\\_CUDART\\_\\_MEMORY\\_1gb65da58f444e7230d3322b6126bb4902](https://docs.nvidia.com/cuda/cuda-runtime-api/group__CUDART__MEMORY.html%5C#group__CUDART__MEMORY_1gb65da58f444e7230d3322b6126bb4902) (visited on 01/15/2024) (cit. on p. 80).
- [49] NVIDIA Corporation. *NVIDIA H100 Tensor Core GPU Architecture*. URL: <https://resources.nvidia.com/en-us-tensor-core> (visited on 01/22/2024) (cit. on p. 33).
- [50] NVIDIA Corporation. *System Management Interface SMI*. 2023. URL: <https://developer.nvidia.com/nvidia-system-management-interface> (visited on 11/22/2023) (cit. on p. 100).
- [51] NVIDIA Corporation. *Train With Mixed Precision*. Feb. 2023. URL: <https://docs.nvidia.com/deeplearning/performance/mixed-precision-training/index.html> (visited on 01/22/2024) (cit. on p. 38).
- [52] D. R. Cox. “The Regression Analysis of Binary Sequences”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 20.2 (1958). Publisher: [Royal Statistical Society, Wiley], pp. 215–242. ISSN: 0035-9246. URL: <https://www.jstor.org/stable/2983890> (cit. on p. 54).
- [53] Alfred Cuschieri and George B Hanna. *Essential surgical practice: Higher Surgical Training in General Surgery*. Boca Raton, FL: CRC Press, May 2014. ISBN: 978-1-4441-3760-6 (cit. on p. 16).
- [54] Paolo Dell’Oglio, Elio Mazzone, Tessa Buckle, Tobias Maurer, Nassir Navab, Matthias N van Oosterom, Clare Schilling, Max Jh Witjes, Alexander L Vahrmeijer, Joachim Klode, Boris Vojnovic, Alexandre Mottrie, Henk G van der Poel, Freddie Hamdy, and Fijs Wb van Leeuwen. “Precision surgery: the role of intra-operative real-time image guidance - outcomes from a multidisciplinary European consensus conference”. In: *Am J Nucl Med Mol Imaging* 12.2 (Apr. 2022), pp. 74–80 (cit. on p. 18).

- [55] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. ISSN: 1063-6919. June 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848 (cit. on pp. 50, 83–85).
- [56] TingYan Deng, Shubham Gulati, Ashwin Kumar, William Rodriguez, Benoit M. Dawant, and Alexander Langerman. “Automated detection of surgical wounds in videos of open neck procedures using a mask R-CNN”. In: *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*. Ed. by Cristian A. Linte and Jeffrey H. Siewerdsen. Vol. 11598. International Society for Optics and Photonics. SPIE, 2021, p. 1159817. DOI: 10.1117/12.2580908 (cit. on p. 56).
- [57] Lee R. Dice. “Measures of the Amount of Ecologic Association Between Species”. In: *Ecology* 26.3 (1945), pp. 297–302. ISSN: 00129658, 19399170. URL: <http://www.jstor.org/stable/1932409> (cit. on p. 105).
- [58] Maximilian Dietrich, Silvia Seidlitz, Nicholas Schreck, Manuel Wiesenfarth, Patrick Godau, Minu Tizabi, Jan Sellner, Sebastian Marx, Samuel Knödler, Michael M. Allers, Leonardo Ayala, Karsten Schmidt, Thorsten Brenner, Alexander Studier-Fischer, Felix Nickel, Beat P. Müller-Stich, Annette Kopp-Schneider, Markus A. Weigand, and Lena Maier-Hein. “Machine learning-based analysis of hyperspectral images for automated sepsis diagnosis”. In: *arXiv:2106.08445 [cs]* (June 2021). arXiv: 2106.08445. URL: <http://arxiv.org/abs/2106.08445> (cit. on p. 110).
- [59] Thomas G. Dietterich. “Ensemble Methods in Machine Learning”. In: *Multiple Classifier Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15. ISBN: 978-3-540-45014-6 (cit. on p. 120).
- [60] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations*. 2021. URL: <https://openreview.net/forum?id=YicbFdNTTy> (cit. on p. 173).
- [61] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. *Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection*. Aug. 2017 (cit. on p. 87).
- [62] Maria Ewerlöf, Marcus Larsson, and E. Göran Salerud. “Spatial and temporal skin blood volume and saturation estimation using a multispectral snapshot imaging camera”. In: *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XV*. Ed. by Daniel L. Farkas, Dan V. Nicolau, and Robert C. Leif. SPIE, Feb. 2017. DOI: 10.1117/12.2251928 (cit. on p. 25).

- 
- [63] Himar Fabelo, Martin Halicek, Samuel Ortega, Adam Szolna, Jesus Morera, Roberto Sarmiento, Gustavo M. Callico, and Baowei Fei. “Surgical Aid Visualization System for Glioblastoma Tumor Identification based on Deep Learning and In-Vivo Hyperspectral Images of Human Patients”. In: *Proceedings of SPIE—the International Society for Optical Engineering* 10951 (Feb. 2019), p. 1095110. ISSN: 0277-786X. DOI: 10.1117/12.2512569 (cit. on pp. 53, 56, 58).
  - [64] Himar Fabelo, Samuel Ortega, Silvester Kabwama, Gustavo M. Callico, Diederik Bulters, Adam Szolna, Juan F. Pineiro, and Roberto Sarmiento. “HELICoiD project: a new use of hyperspectral imaging for brain cancer detection in real-time during neurosurgical operations”. In: *Hyperspectral Imaging Sensors: Innovative Applications and Sensor Standards 2016*. Vol. 9860. SPIE, May 2016, p. 986002. DOI: 10.1117/12.2223075. URL: <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9860/986002/HELICoiD-project--a-new-use-of-hyperspectral-imaging-for/10.1117/12.2223075.full> (cit. on pp. 52, 53).
  - [65] Himar Fabelo, Samuel Ortega, Daniele Ravi, B. Ravi Kiran, Coralia Sosa, Diederik Bulters, Gustavo M. Callicó, Harry Bulstrode, Adam Szolna, Juan F. Piñeiro, Silvester Kabwama, Daniel Madroñal, Raquel Lazcano, Aruma J-O’Shanahan, Sara Bisshopp, María Hernández, Abelardo Báez, Guang-Zhong Yang, Bogdan Stanculescu, Rubén Salvador, Eduardo Juárez, and Roberto Sarmiento. “Spatio-spectral classification of hyperspectral images for brain cancer detection during surgical operations”. In: *PLOS ONE* 13.3 (Mar. 2018), e0193721. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0193721. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0193721> (cit. on pp. 53, 56).
  - [66] Himar Fabelo, Samuel Ortega, Adam Szolna, Diederik Bulters, Juan F. Piñeiro, Silvester Kabwama, Aruma J-O’Shanahan, Harry Bulstrode, Sara Bisshopp, B. Ravi Kiran, Daniele Ravi, Raquel Lazcano, Daniel Madroñal, Coralia Sosa, Carlos Espino, Mariano Marquez, María De La Luz Plaza, Rafael Camacho, David Carrera, María Hernández, Gustavo M. Callicó, Jesús Morera Molina, Bogdan Stanculescu, Guang-Zhong Yang, Rubén Salvador, Eduardo Juárez, César Sanz, and Roberto Sarmiento. “In-Vivo Hyperspectral Human Brain Image Database for Brain Cancer Detection”. In: *IEEE Access* 7 (2019), pp. 39098–39116. DOI: 10.1109/ACCESS.2019.2904788 (cit. on pp. 7, 52, 57, 94).
  - [67] Peter J. Fabri and José L. Zayas-Castro. “Human error, not communication and systems, underlies surgical complications”. In: *Surgery* 144.4 (2008), pp. 557–565. ISSN: 0039-6060. DOI: 10.1016/j.surg.2008.06.011. URL: <https://www.sciencedirect.com/science/article/pii/S0039606008004017> (cit. on p. 1).

- [68] Baowei Fei. “Chapter 3.6 - Hyperspectral imaging in medical applications”. In: *Data Handling in Science and Technology*. Ed. by José Manuel Amigo. Vol. 32. Hyperspectral Imaging. Elsevier, Jan. 2020, pp. 523–565. DOI: 10.1016/B978-0-444-63977-6.00021-3. URL: <https://www.sciencedirect.com/science/article/pii/B9780444639776000213> (cit. on pp. 2, 6, 25, 51, 62, 162, 163).
- [69] *Floating point numbers*. URL: <https://www.learncpp.com/cpp-tutorial/floating-point-numbers> (visited on 01/19/2024) (cit. on p. 37).
- [70] Yunguan Fu, Maria R. Robu, Bongjin Koo, Crispin Schneider, Stijn van Laarhoven, Danail Stoyanov, Brian Davidson, Matthew J. Clarkson, and Yipeng Hu. “More Unlabelled Data or Label More Data? A Study on Semi-supervised Laparoscopic Image Segmentation”. In: *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*. Ed. by Qian Wang, Fausto Milletari, Hien V. Nguyen, Shadi Albarqouni, M. Jorge Cardoso, Nicola Rieke, Ziyue Xu, Konstantinos Kamnitsas, Vishal Patel, Badri Roysam, Steve Jiang, Kevin Zhou, Khoa Luu, and Ngan Le. Cham: Springer International Publishing, 2019, pp. 173–180. ISBN: 978-3-030-33391-1. DOI: 10.1007/978-3-030-33391-1\_20. URL: <http://arxiv.org/abs/1908.08035> (cit. on pp. 51, 56).
- [71] Azat Garifullin, Peeter Kööbi, Pasi Ylitepsa, Kati Ådjers, Markku Hauta-Kasari, Hannu Uusitalo, and Lasse Lensu. “Hyperspectral Image Segmentation of Retinal Vasculature, Optic Disc and Macula”. In: *2018 Digital Image Computing: Techniques and Applications (DICTA)*. Dec. 2018, pp. 1–5. DOI: 10.1109/DICTA.2018.8615761 (cit. on pp. 53, 54, 56, 58).
- [72] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. “A survey of uncertainty in deep neural networks”. In: *Artificial Intelligence Review* 56.1 (Oct. 1, 2023), pp. 1513–1589. ISSN: 1573-7462. DOI: 10.1007/s10462-023-10562-9 (cit. on p. 168).
- [73] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. “Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, June 2021, pp. 2917–2927. ISBN: 978-1-6654-4509-2 (cit. on p. 87).
- [74] Eli Gibson, Maria R. Robu, Stephen Thompson, P. Eddie Edwards, Crispin Schneider, Kurinchi Gurusamy, Brian Davidson, David J. Hawkes, Dean C. Barratt, and Matthew J. Clarkson. “Deep residual networks for automatic segmentation of laparoscopic videos of the liver”. In: ed. by Robert J. Webster and Baowei Fei. Orlando, Florida, United States, Mar. 2017, p. 101351M. DOI: 10.1117/12.2255975 (cit. on pp. 51, 56).



- 
- [75] Robert Gillies, Jenny E Freeman, Leopoldo C Cancio, Derek Brand, Michael Hopmeier, and James R Mansfield. “Systemic effects of shock and resuscitation monitored by visible hyperspectral imaging”. In: *Diabetes Technol Ther* 5.5 (2003), pp. 847–855. DOI: 10.1089/152091503322527058 (cit. on p. 2).
- [76] Diaspective Vision GmbH. *Diaspective Vision - The fifth dimension of medical imaging*. URL: <https://diaspective-vision.com/en> (visited on 11/16/2023) (cit. on pp. 2, 61).
- [77] Diaspective Vision GmbH. *Tivita® Tissue FAQs*. Nov. 19, 2018. URL: [https://diaspective-vision.com/wp-content/uploads/2021/02/0101001-MD-011-c\\_TIVITA-Tissue-FAQ\\_DE.pdf](https://diaspective-vision.com/wp-content/uploads/2021/02/0101001-MD-011-c_TIVITA-Tissue-FAQ_DE.pdf) (visited on 01/11/2024) (cit. on p. 20).
- [78] Patrick Godau, Piotr Kalinowski, Evangelia Christodoulou, Annika Reinke, Minu Tizabi, Luciana Ferrer, Paul F. Jäger, and Lena Maier-Hein. “Deployment of Image Analysis Algorithms Under Prevalence Shifts”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Cham: Springer Nature Switzerland, 2023, pp. 389–399. ISBN: 978-3-031-43898-1 (cit. on p. 168).
- [79] Julia Gong, F. Christopher Holsinger, Julia E. Noel, Sohei Mitani, Jeff Jopling, Nikita Bedi, Yoon Woo Koh, Lisa A. Orloff, Claudio R. Cernea, and Serena Yeung. “Using deep learning to identify the recurrent laryngeal nerve during thyroidectomy”. In: *Scientific Reports* 11.1 (July 2021), p. 14306. ISSN: 2045-2322. DOI: 10.1038/s41598-021-93202-y. URL: <https://www.nature.com/articles/s41598-021-93202-y> (cit. on pp. 51, 56, 57).
- [80] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: <http://www.deeplearningbook.org> (cit. on pp. 26, 27).
- [81] Maria Grammatikopoulou, Evangello Flouty, Abdolrahim Kadkhodamohammadi, Gwenolé Quéllec, Andre Chow, Jean Nehme, Imanol Luengo, and Danail Stoyanov. “CaDIS: Cataract dataset for surgical RGB-image segmentation”. In: *Medical Image Analysis* 71 (July 2021), p. 102053. ISSN: 1361-8415. DOI: 10.1016/j.media.2021.102053 (cit. on p. 51).
- [82] Michael D. Grossberg and Shree K. Nayar. “What is the space of camera response functions?” In: *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. 2* (2003), pp. II–602. URL: <https://api.semanticscholar.org/CorpusID:14857655> (cit. on p. 25).
- [83] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks”. In: *ICML’17*. Sydney, NSW, Australia: JMLR.org, 2017, pp. 1321–1330 (cit. on p. 168).

- [84] Mikael Häggström. “Medical gallery of Mikael Häggström 2014”. In: *WikiJournal of Medicine* 1.2 (2014). ISSN: 2002-4436. DOI: 10.15347/wjm/2014.008 (cit. on p. 16).
- [85] Tamás Haidegger. “Autonomy for Surgical Robots: Concepts and Paradigms”. In: *IEEE Transactions on Medical Robotics and Bionics* 1.2 (2019), pp. 65–76. DOI: 10.1109/TMRB.2019.2913282 (cit. on p. 1).
- [86] Mark Harris. *How to Optimize Data Transfers in CUDA C/C++*. Dec. 4, 2012. URL: <https://developer.nvidia.com/blog/how-optimize-data-transfers-cuda-cc> (visited on 01/15/2024) (cit. on p. 80).
- [87] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, 1994 (cit. on p. 54).
- [88] Mingyi He, Bo Li, and Huahui Chen. “Multi-Scale 3D Deep Convolutional Neural Network for Hyperspectral Image Classification”. In: *2017 IEEE International Conference on Image Processing (ICIP)*. 2017 IEEE International Conference on Image Processing (ICIP). Beijing: IEEE, Sept. 2017, pp. 3904–3908. ISBN: 978-1-5090-2175-8. DOI: 10.1109/ICIP.2017.8297014. URL: <http://ieeexplore.ieee.org/document/8297014> (cit. on p. 52).
- [89] G.E. Healey and R. Kondepudy. “Radiometric CCD camera calibration and noise estimation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16.3 (1994), pp. 267–276. DOI: 10.1109/34.276126 (cit. on p. 25).
- [90] Will Douglas Heaven. “Hundreds of AI tools have been built to catch covid. None of them helped.” In: *MIT Technology Review* (July 30, 2021). URL: <https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic> (visited on 01/03/2024) (cit. on p. 3).
- [91] Tobias Heimann, Bram van Ginneken, Martin A. Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, Fernando Bello, Gerd Binnig, Horst Bischof, Alexander Bornik, Peter M. M. Cashman, Ying Chi, Andrés Cordova, Benoit M. Dawant, Márta Fidrich, Jacob D. Furst, Daisuke Furukawa, Lars Grenacher, Joachim Hornegger, Dagmar Kainmüller, Richard I. Kitney, Hidefumi Kobatake, Hans Lamecker, Thomas Lange, Jeongjin Lee, Brian Lennon, Rui Li, Senhu Li, Hans-Peter Meinzer, GÁbor Nemeth, Daniela S. Raicu, Anne-Mareike Rau, Eva M. van Rikxoort, Mikael Rousson, LÁszló Rusko, Kinda A. Saddi, Günter Schmidt, Dieter Seghers, Akinobu Shimizu, Pieter Slagmolen, Erich Sorantin, Grzegorz Soza, Ruchaneewan Susomboon, Jonathan M. Waite, Andreas Wimmer, and Ivo Wolf. “Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets”. In: *IEEE Transactions on Medical Imaging* 28.8 (Aug. 2009). Conference Name: IEEE Transactions on Medical Imaging, pp. 1251–1265. ISSN: 1558-254X. DOI: 10.1109/TMI.2009.2013851 (cit. on p. 106).

- 
- [92] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. “Why ReLU Networks Yield High-Confidence Predictions Far Away From the Training Data and How to Mitigate the Problem”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 41–50. DOI: 10.1109/CVPR.2019.00013 (cit. on p. 168).
- [93] Tim Holland-Letz and Annette Kopp-Schneider. “Drawing statistical conclusions from experiments with multiple quantitative measurements per subject”. In: *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 152 (Nov. 2020), pp. 30–33. ISSN: 1879-0887. DOI: 10.1016/j.radonc.2020.08.009 (cit. on p. 108).
- [94] Amadeus Holmer, Jörg Marotz, Philip Wahl, Michael Dau, and Peer W. Kämmerer. “Hyperspectral imaging in perfusion and wound diagnostics – methods and algorithms for the determination of tissue parameters”. In: *Biomedical Engineering / Biomedizinische Technik* 63.5 (2018), pp. 547–556. ISSN: 0013-5585. DOI: 10.1515/bmt-2017-0155. URL: <https://www.degruyter.com/view/j/bmte.2018.63.issue-5/bmt-2017-0155/bmt-2017-0155.xml> (cit. on pp. 61, 62).
- [95] Daniel W. Hook, Simon J. Porter, and Christian Herzog. “Dimensions: Building Context for Search and Evaluation”. In: *Frontiers in Research Metrics and Analytics* 3 (2018). <https://www.frontiersin.org/articles/10.3389/frma.2018.00023/pdf>, p. 23. DOI: 10.3389/frma.2018.00023. URL: <https://app.dimensions.ai/details/publication/pub.1106289502> (cit. on p. 47).
- [96] Bingliang Hu, Jian Du, Zhoufeng Zhang, and Quan Wang. “Tumor tissue classification based on micro-hyperspectral technology and deep learning”. In: *Biomed. Opt. Express* 10.12 (Dec. 2019), pp. 6370–6389. DOI: 10.1364/BOE.10.006370. URL: <https://opg.optica.org/boe/abstract.cfm?URI=boe-10-12-6370> (cit. on pp. 48, 49).
- [97] Joni Hyttinen, Pauli Fält, Heli Jäsberg, Arja Kullaa, and Markku Hauta-Kasari. “Oral and Dental Spectral Image Database—ODSI-DB”. In: *Applied Sciences* 10.20 (Oct. 2020), p. 7246. ISSN: 2076-3417. DOI: 10.3390/app10207246 (cit. on pp. 7, 94).
- [98] “IEEE Standard for Floating-Point Arithmetic”. In: *IEEE Std 754-2019 (Revision of IEEE 754-2008)* (2019), pp. 1–84. DOI: 10.1109/IEEESTD.2019.8766229 (cit. on pp. 34, 35).
- [99] Maximilian Ilse, Jakub M. Tomczak, Christos Louizos, and Max Welling. “DIVA: Domain Invariant Variational Autoencoders”. In: *Proceedings of the Third Conference on Medical Imaging with Deep Learning*. Ed. by Tal Arbel, Ismail Ben Ayed, Marleen de Bruijne, Maxime Descoteaux, Herve Lombaert, and Christopher Pal. Vol. 121. Proceedings of Machine Learning Research. PMLR, June 2020, pp. 322–

348. URL: <https://proceedings.mlr.press/v121/ilse20a.html> (cit. on p. 172).
- [100] International Conference on Harmonisation of technical requirements for registration of pharmaceuticals for human use. “ICH harmonized tripartite guideline: Guideline for Good Clinical Practice”. In: *J Postgrad Med* 47.1 (Jan. 2001), pp. 45–50 (cit. on p. 61).
- [101] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *arXiv:1502.03167 [cs]* (Mar. 2015). arXiv: 1502.03167. URL: <http://arxiv.org/abs/1502.03167> (cit. on pp. 86, 110).
- [102] Mobarakol Islam, Daniel Anojan Atputharuban, Ravikiran Ramesh, and Hongliang Ren. “Real-Time Instrument Segmentation in Robotic Surgery Using Auxiliary Supervised Deep Adversarial Learning”. In: *IEEE Robotics and Automation Letters* 4.2 (Apr. 2019). Conference Name: IEEE Robotics and Automation Letters, pp. 2188–2195. ISSN: 2377-3766. DOI: 10.1109/LRA.2019.2900854 (cit. on p. 56).
- [103] Pavel Izmailov, Dmitrii Podoprikin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. “Averaging Weights Leads to Wider Optima and Better Generalization”. In: *Conference on Uncertainty in Artificial Intelligence*. 2018. URL: <https://api.semanticscholar.org/CorpusID:3833416> (cit. on pp. 85, 100).
- [104] Paul Jaccard. “The Distribution of the Flora of the Alpine Zone”. In: *New Phytologist* 11.2 (1912), pp. 37–50. DOI: 10.1111/j.1469-8137.1912.tb05611.x. eprint: <https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8137.1912.tb05611.x> (cit. on p. 105).
- [105] Steven L Jacques. “Optical properties of biological tissues: a review”. In: *Physics in Medicine & Biology* 58.11 (May 2013), R37. DOI: 10.1088/0031-9155/58/11/R37 (cit. on p. 55).
- [106] Simon Jegou, Michal Drozdal, David Vazquez, Adriana Romero, and Yoshua Bengio. “The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. arXiv: 1611.09326. July 2017. URL: [https://openaccess.thecvf.com/content\\_cvpr\\_2017\\_workshops/w13/html/Jegou\\_The\\_One\\_Hundred\\_CVPR\\_2017\\_paper.html](https://openaccess.thecvf.com/content_cvpr_2017_workshops/w13/html/Jegou_The_One_Hundred_CVPR_2017_paper.html) (cit. on p. 54).
- [107] Sen Jia, Shuguo Jiang, Zhijie Lin, Nanying Li, Meng Xu, and Shiqi Yu. “A Survey: Deep Learning for Hyperspectral Image Classification with Few Labeled Samples”. In: *Neurocomputing* 448 (Aug. 2021), pp. 179–204. ISSN: 09252312. DOI: 10.1016/

- 
- j . neucom . 2021 . 03 . 035. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0925231221004033> (cit. on p. 52).
- [108] Licheng Jiao, Miaomiao Liang, Huan Chen, Shuyuan Yang, Hongying Liu, and Xianghai Cao. “Deep Fully Convolutional Network-Based Spatial Distribution Prediction for Hyperspectral Image Classification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 55.10 (Oct. 2017), pp. 5585–5599. ISSN: 0196-2892, 1558-0644. DOI: 10 . 1109/TGRS . 2017 . 2710079. URL: <http://ieeexplore.ieee.org/document/7967742> (cit. on p. 52).
  - [109] Matthew Johnson and Jamie Shotton. “Semantic Texton Forests”. In: *Computer Vision: Detection, Recognition and Reconstruction*. Ed. by Roberto Cipolla, Sebastiano Battiato, and Giovanni Maria Farinella. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 173–203. ISBN: 978-3-642-12848-6. DOI: 10 . 1007/978-3-642-12848-6\_7 (cit. on p. 53).
  - [110] Leo Joskowicz, D. Cohen, N. Caplan, and J. Sosna. “Inter-observer variability of manual contour delineation of structures in CT”. In: *European Radiology* 29.3 (Mar. 2019), pp. 1391–1399. ISSN: 1432-1084. DOI: 10 . 1007/s00330-018-5695-5 (cit. on p. 108).
  - [111] Abdolrahim Kadkhodamohammadi, Imanol Luengo, Santiago Barbarisi, Hinde Taleb, Evangello Flouty, and Danail Stoyanov. “Feature Aggregation Decoder for Segmenting Laparoscopic Scenes”. In: *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*. Ed. by Luping Zhou, Duygu Sarikaya, Seyed Mostafa Kia, Stefanie Speidel, Anand Malpani, Daniel Hashimoto, Mohamad Habes, Tommy Löfstedt, Kerstin Ritter, and Hongzhi Wang. Vol. 11796. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 3–11. ISBN: 978-3-030-32695-1. DOI: 10 . 1007/978-3-030-32695-1\_1. URL: [http://link.springer.com/10.1007/978-3-030-32695-1\\_1](http://link.springer.com/10.1007/978-3-030-32695-1_1) (cit. on pp. 51, 56).
  - [112] Ibrahim Kandel and Mauro Castelli. “The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset”. In: *ICT Express* 6.4 (2020), pp. 312–315. ISSN: 2405-9595. DOI: 10 . 1016/j.ict.2020.04.010. URL: <https://www.sciencedirect.com/science/article/pii/S2405959519303455> (cit. on p. 86).
  - [113] Jaka Katrašnik, Franjo Pernuš, and Boštjan Likar. “Radiometric calibration and noise estimation of acousto-optic tunable filter hyperspectral imaging systems”. In: *Appl Opt* 52.15 (May 2013), pp. 3526–3537. DOI: 10 . 1364/AO . 52 . 003526 (cit. on p. 25).
  - [114] Muhammad Jaleed Khan, Hamid Saeed Khan, Adeel Yousaf, Khurram Khurshid, and Asad Abbas. “Modern Trends in Hyperspectral Image Analysis: A Review”. In: *IEEE Access* 6 (2018), pp. 14118–14129. ISSN: 2169-3536. DOI: 10 . 1109/ACC

- ESS. 2018. 2812999. URL: <https://ieeexplore.ieee.org/document/8314827> (cit. on p. 51).
- [115] Uzair Khan, Paheding Sidike, Colin Elkin, and Vijay Devabhaktuni. “Trends in deep learning for medical hyperspectral image analysis”. In: *IEEE Access* 9 (2021). arXiv: 2011.13974, pp. 79534–79548. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3068392 (cit. on p. 52).
- [116] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: <http://arxiv.org/abs/1412.6980> (cit. on pp. 74, 85).
- [117] Daichi Kitaguchi, Toru Fujino, Nobuyoshi Takeshita, Hiro Hasegawa, Kensaku Mori, and Masaaki Ito. “Limited generalizability of single deep neural network for surgical instrument segmentation in different surgical environments”. In: *Scientific Reports* 12.1 (July 22, 2022), p. 12575. ISSN: 2045-2322. DOI: 10.1038/s41598-022-16923-8 (cit. on pp. 55, 56).
- [118] Teuvo Kohonen. “Learning Vector Quantization”. In: *Self-Organizing Maps*. Ed. by Teuvo Kohonen. Springer Series in Information Sciences. Berlin, Heidelberg: Springer, 1995, pp. 175–189. ISBN: 978-3-642-97610-0. DOI: 10.1007/978-3-642-97610-0\_6 (cit. on p. 52).
- [119] Esther N D Kok, Roeland Eppenga, Koert F D Kuhlmann, Harald C Groen, Ruben van Veen, Jolanda M van Dieren, Thomas R de Wijkerslooth, Monique van Leerdam, Doenja M J Lambregts, Wouter J Heerink, Nikie J Hoetjes, Oleksandra Ivashchenko, Geerard L Beets, Arend G J Aalbers, Jasper Nijkamp, and Theo J M Ruers. “Accurate surgical navigation with real-time tumor tracking in cancer surgery”. In: *npj Precision Oncology* 4.1 (Apr. 2020), p. 8. DOI: 10.1038/s41698-020-0115-0 (cit. on p. 18).
- [120] Fiona R. Kolbinger, Franziska M. Rinner, Alexander C. Jenke, Matthias Carstens, Stefanie Krell, Stefan Leger, Marius Distler, Jürgen Weitz, Stefanie Speidel, and Sebastian Bodenstedt. “Anatomy segmentation in laparoscopic surgery: comparison of machine learning and human expertise – an experimental study”. In: *International Journal of Surgery* 109.10 (2023). URL: [https://journals.lww.com/international-journal-of-surgery/fulltext/2023/10000/anatomy\\_segmentation\\_in\\_laparoscopic\\_surgery\\_.10.aspx](https://journals.lww.com/international-journal-of-surgery/fulltext/2023/10000/anatomy_segmentation_in_laparoscopic_surgery_.10.aspx) (cit. on p. 56).
- [121] Xiaowen Kong, Yueming Jin, Qi Dou, Ziyi Wang, Zerui Wang, Bo Lu, Erbao Dong, Yun-Hui Liu, and Dong Sun. “Accurate instance segmentation of surgical instruments in robotic surgery: model refinement and cross-dataset evaluation”. In: *International Journal of Computer Assisted Radiology and Surgery* 16.9 (Sept. 1,

- 
- 2021), pp. 1607–1614. ISSN: 1861-6429. DOI: 10.1007/s11548-021-02438-6 (cit. on p. 56).
- [122] Axel Kulcke, Amadeus Holmer, Philip Wahl, Frank Siemers, Thomas Wild, and Georg Daeschlein. “A compact hyperspectral camera for measurement of perfusion parameters in medicine”. In: *Biomedical Engineering / Biomedizinische Technik* 63.5 (Oct. 2018), pp. 519–527. ISSN: 1862-278X, 0013-5585. DOI: 10.1515/bmt-2017-0145. URL: <http://www.degruyter.com/view/j/bmte.2018.63.issue-5/bmt-2017-0145/bmt-2017-0145.xml> (cit. on pp. 22, 61).
- [123] Solomon Kullback and Richard A. Leibler. “On Information and Sufficiency”. In: *The Annals of Mathematical Statistics* 22.1 (1951). Publisher: Institute of Mathematical Statistics, pp. 79–86. ISSN: 0003-4851. URL: <https://www.jstor.org/stable/2236703> (cit. on p. 84).
- [124] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. “Quantifying the Carbon Emissions of Machine Learning”. In: (Oct. 2019). URL: <https://arxiv.org/abs/1910.09700> (cit. on p. 154).
- [125] Max-Heinrich Laves, Jens Bicker, Lüder A. Kahrs, and Tobias Ortmaier. “A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation”. In: *International Journal of Computer Assisted Radiology and Surgery* 14.3 (Mar. 2019), pp. 483–492. ISSN: 1861-6429. DOI: 10.1007/s11548-018-01910-0 (cit. on pp. 51, 56).
- [126] Silas J. Leavesley, Mikayla Walters, Carmen Lopez, Thomas Baker, Peter F. Favreau, Thomas C. Rich, Paul F. Rider, and Carole W. Boudreaux. “Hyperspectral imaging fluorescence excitation scanning for colon cancer detection”. In: *Journal of Biomedical Optics* 21.10 (2016), p. 104003. DOI: 10.1117/1.JBO.21.10.104003 (cit. on p. 48).
- [127] Guillaume Leclerc, Andrew Ilyas, Logan Engstrom, Sung Park, Hadi Salman, and Aleksander Mądry. “FFCV: Accelerating Training by Removing Data Bottlenecks”. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2023, pp. 12011–12020. DOI: 10.1109/CVPR52729.2023.01156 (cit. on p. 50).
- [128] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *Nature* 521.7553 (May 1, 2015), pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539 (cit. on pp. 26, 27).
- [129] Ann-Kathrin Lederer, Sophia Chikhladze, Eva Kohnert, Roman Huber, and Alexander Müller. “Current Insights: The Impact of Gut Microbiota on Postoperative Complications in Visceral Surgery—A Narrative Review”. In: *Diagnostics* 11.11 (2021). ISSN: 2075-4418. DOI: 10.3390/diagnostics11112099. URL: <https://www.mdpi.com/2075-4418/11/11/2099> (cit. on p. 16).

- [130] Michelle Li, Jeffrey A. Norton, R. Randal Bollinger, Alfred E. Chang, Stephen F. Lowry, Sean J. Mulvihill, Harvey I. Pass, and Robert W. Thompson. *Essential Practice of Surgery: Basic Science and Clinical Evidence*. Springer New York, 2003. ISBN: 978-0-387-22744-3. DOI: 10.1007/b98876 (cit. on p. 16).
- [131] Cheng-Yuan Liou, Wei-Chen Cheng, Jiun-Wei Liou, and Daw-Ran Liou. “Autoencoder for words”. In: *Neurocomputing* 139 (2014), pp. 84–96 (cit. on p. 32).
- [132] Zhi Liu, Hongjun Wang, and Qingli Li. “Tongue Tumor Detection in Medical Hyperspectral Images”. In: *Sensors* 12.1 (2012), pp. 162–174. ISSN: 1424-8220. DOI: 10.3390/s120100162. URL: <https://www.mdpi.com/1424-8220/12/1/162> (cit. on p. 2).
- [133] Vaibhav Lodhi, Debashish Chakravarty, and Pabitra Mitra. “Hyperspectral Imaging System: Development Aspects and Recent Trends”. In: *Sensing and Imaging* 20.1 (Aug. 13, 2019), p. 35. ISSN: 1557-2072. DOI: 10.1007/s11220-019-0257-8 (cit. on p. 19).
- [134] Mayar Lotfy, Anna Alperovich, Tommaso Giannantonio, Bjorn Barz, Xiaohan Zhang, Felix Holm, Nassir Navab, Felix Boehm, Carolin Schwamborn, Thomas K. Hoffmann, and Patrick J. Schuler. *Robust Tumor Segmentation with Hyperspectral Imaging and Graph Neural Networks*. 2023. arXiv: 2311.11782 [eess.IV] (cit. on pp. 53, 54, 56).
- [135] Mandy Lu, Qingyu Zhao, Jiequan Zhang, Kilian M. Pohl, Li Fei-Fei, Juan Carlos Niebles, and Ehsan Adeli. “Metadata Normalization”. In: *arXiv:2104.09052 [cs]* (May 2021). arXiv: 2104.09052. URL: <http://arxiv.org/abs/2104.09052> (cit. on p. 110).
- [136] Yu-Wen Luo, Hai-Yong Chen, Zhen Li, Wei-Peng Liu, Ke Wang, Li Zhang, Pan Fu, Wen-Qian Yue, and Gui-Bin Bian. “Fast instruments and tissues segmentation of micro-neurosurgical scene using high correlative non-local network”. In: *Computers in Biology and Medicine* 153 (2023), p. 106531. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2022.106531. URL: <https://www.sciencedirect.com/science/article/pii/S0010482522012392> (cit. on p. 56).
- [137] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. “Rectifier nonlinearities improve neural network acoustic models”. In: *Proc. icml*. Vol. 30. 1. Atlanta, GA. 2013, p. 3 (cit. on pp. 31, 32).
- [138] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: <http://jmlr.org/papers/v9/vandemaaten08a.html> (cit. on p. 53).



- 
- [139] Sabrina Madad Zadeh, Tom Francois, Lilian Calvet, Pauline Chauvet, Michel Canis, Adrien Bartoli, and Nicolas Bourdel. “SurgAI: deep learning for computerized laparoscopic image understanding in gynaecology”. In: *Surgical Endoscopy* 34.12 (Dec. 2020), pp. 5377–5383. ISSN: 1432-2218. DOI: 10 . 1007 / s00464 - 019 - 07330 - 8 (cit. on pp. 51, 56).
- [140] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, Hirenkumar Nakawala, Adrian Park, Carla Pugh, Danail Stoyanov, Swaroop S. Vedula, Kevin Cleary, Gabor Fichtinger, Germain Forestier, Bernard Gibaud, Teodor Grantcharov, Makoto Hashizume, Doreen Heckmann-Nötzel, Hannes G. Kenngott, Ron Kikinis, Lars Mündermann, Nassir Navab, Sinan Onogur, Tobias Roß, Raphael Sznitman, Russell H. Taylor, Minu D. Tizabi, Martin Wagner, Gregory D. Hager, Thomas Neumuth, Nicolas Padoy, Justin Collins, Ines Gockel, Jan Goedeke, Daniel A. Hashimoto, Luc Joyeux, Kyle Lam, Daniel R. Leff, Amin Madani, Hani J. Marcus, Ozanan Meireles, Alexander Seitel, Dogu Teber, Frank Ückert, Beat P. Müller-Stich, Pierre Jannin, et al. “Surgical data science – from concepts toward clinical translation”. In: *Medical Image Analysis* 76 (2022), p. 102306. ISSN: 1361-8415. DOI: 10 . 1016 / j . media . 2021 . 102306. URL: <https://www.sciencedirect.com/science/article/pii/S1361841521003510> (cit. on pp. 1, 3, 51, 160).
- [141] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, A. Emre Kavur, Carole H. Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Matthew B. Blaschko, M. Jorge Cardoso, Veronika Cheplygina, Beth A. Cimini, Gary S. Collins, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Robert Haase, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, Florian Kofler, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, et al. “Metrics reloaded: recommendations for image analysis validation”. In: *Nature Methods* 21.2 (Feb. 1, 2024), pp. 195–212. ISSN: 1548-7105. DOI: 10 . 1038 / s41592 - 023 - 02151 - z (cit. on pp. 90, 105, 108).
- [142] Lena Maier-Hein, Martin Wagner, Tobias Ross, Annika Reinke, Sebastian Bodenstedt, Peter M. Full, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, Pierangela Bruno, Anna Kisilenko, Benjamin Müller, Tornike Davitashvili, Manuela Capek, Minu D. Tizabi, Matthias Eisenmann, Tim J. Adler, Janek Gröhl, Melanie Schellenberg, Silvia Seidlitz, T. Y. Emmy Lai, Bünyamin Pekdemir, Veith Roethlingshoefer, Fabian Both, Sebastian Bittel, Marc Mengler, Lars Mündermann, Martin Apitz, Annette Kopp-Schneider, Stefanie Speidel, Felix

- Nickel, Pascal Probst, Hannes G. Kenngott, and Beat P. Müller-Stich. “Heidelberg colorectal data set for surgical data science in the sensor operating room”. In: *Scientific Data* 8.1 (Apr. 2021), p. 101. ISSN: 2052-4463. DOI: 10.1038/s41597-021-00882-2 (cit. on p. 51).
- [143] Marianne Maktabi, Hannes Köhler, Margarita Ivanova, Boris Jansen-Winkel, Jonathan Takoh, Stefan Niebisch, Sebastian M. Rabe, Thomas Neumuth, Ines Gockel, and Claire Chalopin. “Tissue classification of oncologic esophageal resectates based on hyperspectral data”. In: *International Journal of Computer Assisted Radiology and Surgery* 14.10 (Oct. 1, 2019), pp. 1651–1661. ISSN: 1861-6429. DOI: 10.1007/s11548-019-02016-x (cit. on pp. 48, 49).
- [144] Bryce Manifold, Shuaiqian Men, Ruoqian Hu, and Dan Fu. “A Versatile Deep Learning Architecture for Classification and Label-Free Prediction of Hyperspectral Images”. In: *Nature Machine Intelligence* 3.4 (Apr. 2021), pp. 306–315. ISSN: 2522-5839. DOI: 10.1038/s42256-021-00309-y. URL: <http://www.nature.com/articles/s42256-021-00309-y> (cit. on p. 52).
- [145] A. Mansouri, F.S. Marzani, and P. Gouton. “Development of a Protocol for CCD Calibration: Application to a Multispectral Imaging System”. In: *International Journal of Robotics and Automation* 20.2 (2005). ISSN: 1925-7090. DOI: 10.2316/journal.206.2005.2.206-2784 (cit. on p. 25).
- [146] Salman Maqbool, Aqsa Riaz, Hasan Sajid, and Osman Hasan. *m2caiSeg: Semantic Segmentation of Laparoscopic Images using Convolutional Neural Networks*. 2020. arXiv: 2008.10134 [cs.CV] (cit. on p. 56).
- [147] Aidana Massalimova, Maikel Timmermans, Hooman Esfandiari, Fabio Carrillo, Christoph J. Laux, Mazda Farshad, Kathleen Denis, and Philipp Fürnstahl. “Intra-operative tissue classification methods in orthopedic and neurological surgeries: A systematic review”. In: *Frontiers in Surgery* 9 (2022). ISSN: 2296-875X. DOI: 10.3389/fsurg.2022.952539. URL: <https://www.frontiersin.org/articles/10.3389/fsurg.2022.952539> (cit. on p. 1).
- [148] Dirk Merkel. “Docker: lightweight linux containers for consistent development and deployment”. In: *Linux journal* 2014.239 (2014), p. 2 (cit. on p. 100).
- [149] MIC-Surgery-Heidelberg. *HyperGui*. URL: [https://github.com/MIC-Surgery-Heidelberg/HyperGUI2.0\\_lite](https://github.com/MIC-Surgery-Heidelberg/HyperGUI2.0_lite) (visited on 11/17/2023) (cit. on p. 63).
- [150] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. “Mixed Precision Training”. In: *International Conference on Learning Representations*. 2018. URL: <https://openreview.net/forum?id=r1gs9JgRZ> (cit. on p. 33).

- 
- [151] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation”. In: *2016 Fourth International Conference on 3D Vision (3DV)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2016, pp. 565–571. doi: 10.1109/3DV.2016.79 (cit. on pp. 84, 85).
  - [152] Gaby N. Moawad, Savannah Smith, and Jordan Klebanoff. “The US Perspective of Benefit of Minimally Invasive Surgery: Why Is This Important Now?”. In: *Robotic Surgery*. Ed. by Farid Gharagozloo, Vipul R. Patel, Pier Cristoforo Giulianotti, Robert Poston, Rainer Gruessner, and Mark Meyer. Cham: Springer International Publishing, 2021, pp. 1217–1221. ISBN: 978-3-030-53594-0. doi: 10.1007/978-3-030-53594-0\_112 (cit. on p. 17).
  - [153] Sara Moccia, Sebastian J. Wirkert, Hannes Kenngott, Anant S. Vemuri, Martin Apitz, Benjamin Mayer, Elena De Momi, Leonardo S. Mattos, and Lena Maier-Hein. “Uncertainty-Aware Organ Classification for Surgical Data Science Applications in Laparoscopy”. In: *IEEE Transactions on Biomedical Engineering* 65.11 (Nov. 2018), pp. 2649–2659. ISSN: 0018-9294, 1558-2531. doi: 10.1109/TBME.2018.2813015. URL: <https://ieeexplore.ieee.org/document/8310960> (cit. on pp. 51, 53, 56–58).
  - [154] Jayashree Mohan, Amar Phanishayee, Ashish Raniwala, and Vijay Chidambaram. “Analyzing and mitigating data stalls in DNN training”. In: *Proc. VLDB Endow.* 14.5 (Jan. 2021), pp. 771–784. ISSN: 2150-8097. doi: 10.14778/3446095.3446100 (cit. on p. 50).
  - [155] David Moher, Sally Hopewell, Kenneth F Schulz, Victor Montori, Peter C Gøtzsche, P J Devereaux, Diana Elbourne, Matthias Egger, Douglas G Altman, and Consolidated Standards of Reporting Trials Group. “CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials”. In: *J Clin Epidemiol* 63.8 (Mar. 2010), e1–37 (cit. on p. 61).
  - [156] Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller, eds. *Neural Networks: Tricks of the Trade: Second Edition*. Vol. 7700. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. ISBN: 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8. URL: <http://link.springer.com/10.1007/978-3-642-35289-8> (cit. on pp. 86, 154).
  - [157] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. “A unifying view on dataset shift in classification”. In: *Pattern Recognition* 45.1 (2012), pp. 521–530. ISSN: 0031-3203. doi: 10.1016/j.patcog.2011.06.019. URL: <https://www.sciencedirect.com/science/article/pii/S0031320311002901> (cit. on pp. 3, 10).

- [158] Hala Muaddi, Melanie El Hafid, Woo Jin Choi, Erin Lillie, Charles de Mestral, Avery Nathens, Therese A Stukel, and Paul J Karanicolas. "Clinical Outcomes of Robotic Surgery Compared to Conventional Surgical Approaches (Laparoscopic or Open): A Systematic Overview of Reviews". In: *Annals of Surgery* 273.3 (2021). DOI: 10.1097/SLA.0000000000003915 (cit. on p. 17).
- [159] Atif Mughees and Linmi Tao. "Efficient Deep Auto-Encoder Learning for the Classification of Hyperspectral Images". In: *2016 International Conference on Virtual Reality and Visualization (ICVRV)*. Sept. 2016, pp. 44–51. DOI: 10.1109/ICVRV.2016.16 (cit. on p. 52).
- [160] Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. "Understanding the failure modes of out-of-distribution generalization". In: *International Conference on Learning Representations*. 2021. URL: [https://openreview.net/forum?id=fSTD6NFIW\\_b](https://openreview.net/forum?id=fSTD6NFIW_b) (cit. on pp. 3, 10).
- [161] Jakub Nalepa, Marek Antoniak, Michal Myller, Pablo Ribalta Lorenzo, and Michal Marcinkiewicz. "Towards resource-frugal deep convolutional neural networks for hyperspectral image segmentation". In: *Microprocessors and Microsystems* 73 (Mar. 2020), p. 102994. ISSN: 0141-9331. DOI: 10.1016/j.micpro.2020.102994. URL: <https://www.sciencedirect.com/science/article/pii/S0141933119302844> (cit. on p. 52).
- [162] Jakub Nalepa, Michal Myller, and Michal Kawulok. "Validating Hyperspectral Image Segmentation". In: *IEEE Geoscience and Remote Sensing Letters* 16.8 (Aug. 2019). Conference Name: IEEE Geoscience and Remote Sensing Letters, pp. 1264–1268. ISSN: 1558-0571. DOI: 10.1109/LGRS.2019.2895697 (cit. on p. 52).
- [163] Dmitri Nepogodiev, Janet Martin, Bruce Biccard, Alex Makupe, Aneel Bhangu, Adesoji Ademuyiwa, Adewale Oluseye Adisa, Maria-Lorena Aguilera, Sohini Chakrabortee, J. Edward Fitzgerald, Dhruva Ghosh, James C. Glasbey, Ewen M. Harrison, J.C. Allen Ingabire, Hosni Salem, Marie Carmela Lapitan, Ismail Lawani, David Lissauer, Laura Magill, Rachel Moore, Daniel C. Osei-Bordom, Thomas D. Pinkney, Ahmad Uzair Qureshi, Antonio Ramos-De la Medina, Sarah Rayne, Sudha Sundar, Stephen Tabiri, Azmina Verjee, Raul Yopez, O. James Garden, Richard Lilford, Peter Brocklehurst, and Dion G. Morton. "Global burden of postoperative death". In: *The Lancet* 393.10170 (Feb. 2, 2019), p. 401. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(18)33139-8 (cit. on p. 1).
- [164] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015. URL: <http://neuralnetworksanddeeplearning.com> (cit. on pp. 27, 31).
- [165] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernardino Romera-Paredes, Christopher Kelly, Alan Karthikesalingam, Carlton Chu, Dawn

- 
- Carnell, Cheng Boon, Derek D’Souza, Syed Ali Moinuddin, Bethany Garie, Yasmin McQuinlan, Sarah Ireland, Kiarna Hampton, Krystle Fuller, Hugh Montgomery, Geraint Rees, Mustafa Suleyman, Trevor Back, Cían Hughes, Joseph R. Ledsam, and Olaf Ronneberger. “Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy”. In: *arXiv:1809.04430 [physics, stat]* (Jan. 2021). arXiv: 1809.04430. URL: <http://arxiv.org/abs/1809.04430> (cit. on pp. 106, 152, 154).
- [166] Augustus Odena, Vincent Dumoulin, and Chris Olah. “Deconvolution and Checkerboard Artifacts”. In: *Distill* (2016). DOI: 10.23915/distill.00003. URL: <http://distill.pub/2016/deconv-checkerboard> (cit. on p. 32).
- [167] Iason Ofeidis, Diego Kiedanski, and Leandros Tassiulas. *An Overview of the Data-Loader Landscape: Comparative Performance Analysis*. 2022. arXiv: 2209.13705 [cs.DC] (cit. on p. 50).
- [168] Ephrem O. Olweny, Stephen Faddegon, Sara L. Best, Neil Jackson, Eleanor F. Wehner, Yung K. Tan, Karel J. Zuzak, and Jeffrey A. Cadeddu. “First Place: Renal Oxygenation During Robot-Assisted Laparoscopic Partial Nephrectomy: Characterization Using Laparoscopic Digital Light Processing Hyperspectral Imaging”. In: *Journal of Endourology* 27.3 (2013), pp. 265–269. DOI: 10.1089/end.2012.0207 (cit. on p. 2).
- [169] Daniil Pakhomov and Nassir Navab. “Searching for Efficient Architecture for Instrument Segmentation in Robotic Surgery”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Ed. by Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz. Cham: Springer International Publishing, 2020, pp. 648–656. ISBN: 978-3-030-59716-0. DOI: 10.1007/978-3-030-59716-0\_62 (cit. on p. 56).
- [170] Daniil Pakhomov, Wei Shen, and Nassir Navab. “Towards Unsupervised Learning for Instrument Segmentation in Robotic Surgery with Cycle-Consistent Adversarial Networks”. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Las Vegas, NV, USA: IEEE, Oct. 2020, pp. 8499–8504. ISBN: 978-1-7281-6212-6. DOI: 10.1109/IROS45743.2020.9340816. URL: <https://ieeexplore.ieee.org/document/9340816> (cit. on p. 56).
- [171] Svetlana V Panasyuk, Shi Yang, Douglas V Faller, Duyen Ngo, Robert A Lew, Jenny E Freeman, and Adrienne E Rogers. “Medical hyperspectral imaging to facilitate residual tumor identification during surgery”. In: *Cancer Biol Ther* 6.3 (Mar. 2007), pp. 439–446. DOI: 10.4161/cbt.6.3.4018 (cit. on p. 2).
- [172] Arati Paul and Sanghamita Bhoumik. “Classification of hyperspectral imagery using spectrally partitioned HyperUnet”. In: *Neural Computing and Applications* (Sept. 2021). ISSN: 1433-3058. DOI: 10.1007/s00521-021-06532-3 (cit. on p. 52).

- [173] Tim Pearce, Alexandra Brintrup, and Jun Zhu. “Understanding Softmax Confidence and Uncertainty”. In: *ArXiv abs/2106.04972* (2021). URL: <https://api.semanticscholar.org/CorpusID:235376772> (cit. on p. 168).
- [174] Hung Viet Pham, Shangshu Qian, Jiannan Wang, Thibaud Lutellier, Jonathan Rosenthal, Lin Tan, Yaoliang Yu, and Nachiappan Nagappan. “Problems and Opportunities in Training Deep Learning Software Systems: An Analysis of Variance”. In: *2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. ISSN: 2643-1572. Sept. 2020, pp. 771–783 (cit. on p. 86).
- [175] Robi Polikar. “Ensemble Learning”. In: *Ensemble Machine Learning: Methods and Applications*. Ed. by Cha Zhang and Yunqian Ma. New York, NY: Springer New York, 2012, pp. 1–34. ISBN: 978-1-4419-9326-7. DOI: 10.1007/978-1-4419-9326-7\_1. URL: 10.1007/978-1-4419-9326-7\_1 (cit. on p. 120).
- [176] Scott Prahl. *Optical Absorption of Hemoglobin*. 1999. URL: <https://omlc.org/spectra/hemoglobin> (visited on 11/16/2023) (cit. on p. 21).
- [177] Saurabh Prasad and Jocelyn Chanussot, eds. *Hyperspectral Image Analysis: Advances in Machine Learning and Signal Processing*. Advances in Computer Vision and Pattern Recognition. Cham: Springer International Publishing, 2020. ISBN: 978-3-030-38617-7. DOI: 10.1007/978-3-030-38617-7. URL: <http://link.springer.com/10.1007/978-3-030-38617-7> (cit. on pp. 18, 51, 52).
- [178] Sami Puustinen, Hana Vrzáková, Joni Hyttinen, Tuomas Rauramaa, Pauli Fält, Markku Hauta-Kasari, Roman Bednarik, Timo Koivisto, Susanna Rantala, Mikael von und zu Fraunberg, Juha E. Jääskeläinen, and Antti-Pekka Elomaa. “Hyperspectral Imaging in Brain Tumor Surgery—Evidence of Machine Learning-Based Performance”. In: *World Neurosurgery* 175 (2023), e614–e635. ISSN: 1878-8750. DOI: 10.1016/j.wneu.2023.03.149. URL: <https://www.sciencedirect.com/science/article/pii/S1878875023004734> (cit. on p. 2).
- [179] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. “AI in health and medicine”. In: *Nature Medicine* 28.1 (Jan. 1, 2022), pp. 31–38. ISSN: 1546-170X. DOI: 10.1038/s41591-021-01614-0 (cit. on p. 3).
- [180] Daniele Ravi, Himar Fabelo, Gustavo Marrero Callic, and Guang-Zhong Yang. “Manifold Embedding and Semantic Segmentation for Intraoperative Guidance With Hyperspectral Brain Imaging”. In: *IEEE Transactions on Medical Imaging* 36.9 (Sept. 2017). Conference Name: IEEE Transactions on Medical Imaging, pp. 1845–1857. ISSN: 1558-254X. DOI: 10.1109/TMI.2017.2695523 (cit. on pp. 53, 56).
- [181] Partha Pratim Ray. “ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope”. In: *Internet of Things and Cyber-Physical Systems* 3 (2023), pp. 121–154. ISSN: 2667-3452. DOI:

- 
- 10.1016/j.iotcps.2023.04.003. URL: <https://www.sciencedirect.com/science/article/pii/S266734522300024X> (cit. on p. 3).
- [182] Aziz ul Rehman and Shahzad Ahmad Qureshi. “A review of the medical hyperspectral imaging systems and unmixing algorithms’ in biological tissues”. In: *Photodiagnosis and Photodynamic Therapy* 33 (2021), p. 102165. ISSN: 1572-1000. DOI: 10.1016/j.pdpdt.2020.102165. URL: <https://www.sciencedirect.com/science/article/pii/S1572100020305196> (cit. on p. 19).
- [183] Annika Reinke, Minu D. Tizabi, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A. Emre Kavur, Tim Rädtsch, Carole H. Sudre, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Florian Buettner, M. Jorge Cardoso, Veronika Cheplygina, Jianxu Chen, Evangelia Christodoulou, Beth A. Cimini, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Ben Glocker, Patrick Godau, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Fabian Isensee, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Jens Kleesiek, Florian Kofler, Thijs Kooi, Annette Kopp-Schneider, Michal Kozubek, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, et al. “Understanding metric-related pitfalls in image analysis validation”. In: *Nature Methods* 21.2 (Feb. 1, 2024), pp. 182–194. ISSN: 1548-7105. DOI: 10.1038/s41592-023-02150-0 (cit. on pp. 3, 90, 105).
- [184] Annika Reinke, Minu D. Tizabi, Carole H. Sudre, Matthias Eisenmann, Tim Rädtsch, Michael Baumgartner, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Peter Bankhead, Arriel Benis, Matthew Blaschko, Florian Büttner, M. Jorge Cardoso, Jianxu Chen, Veronika Cheplygina, Evangelia Christodoulou, Beth Cimini, Gary S. Collins, Sandy Engelhardt, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Ben Glocker, Patrick Godau, Robert Haase, Fred Hamprecht, Daniel A. Hashimoto, Doreen Heckmann-Nötzel, Peter Hirsch, Michael M. Hoffman, Merel Huisman, Fabian Isensee, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, A. Emre Kavur, Hannes Kenngott, Jens Kleesiek, Andreas Kleppe, Sven Kohler, Florian Kofler, Annette Kopp-Schneider, Thijs Kooi, Michal Kozubek, et al. *Common Limitations of Image Processing Metrics: A Picture Story*. 2023. arXiv: 2104.05642 [eess.IV] (cit. on pp. 90, 105, 106).
- [185] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. “Kornia: an Open Source Differentiable Computer Vision Library for PyTorch”. In: *Winter Conference on Applications of Computer Vision*. 2020. URL: <https://arxiv.org/pdf/1910.02190.pdf> (cit. on pp. 78, 85, 167).

- [186] Michael Richmond. *Dark Subtraction and Flatfielding*. Apr. 5, 2002. URL: <http://spiff.rit.edu/classes/phys445/lectures/darkflat/darkflat.html> (visited on 01/12/2024) (cit. on p. 24).
- [187] Irene Rivas-Blanco, Carlos J. Pérez-Del-Pulgar, Isabel García-Morales, and Víctor F. Muñoz. “A Review on Deep Learning in Minimally Invasive Surgery”. In: *IEEE Access* 9 (2021), pp. 48658–48678. DOI: 10.1109/ACCESS.2021.3068852 (cit. on pp. 9, 51).
- [188] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, Alessandro Ruggiero, Anna Korhonen, Emily Jefferson, Emmanuel Ako, Georg Langs, Ghassem Gozaliasl, Guang Yang, Helmut Prosch, Jacobus Preller, Jan Stanczuk, Jing Tang, Johannes Hofmanninger, Judith Babar, Lorena Escudero Sánchez, Muhunthan Thillai, Paula Martin Gonzalez, Philip Teare, Xiaoxiang Zhu, Mishal Patel, Conor Cafolla, Hojjat Azadbakht, Joseph Jacob, Josh Lowe, Kang Zhang, Kyle Bradley, Marcel Wassin, Markus Holzer, Kangyu Ji, Maria Delgado Ortet, Tao Ai, Nicholas Walton, Pietro Lio, Samuel Stranks, Tolou Shadbahr, Weizhe Lin, Yunfei Zha, Zhangming Niu, et al. “Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans”. In: *Nature Machine Intelligence* 3.3 (Mar. 1, 2021), pp. 199–217. ISSN: 2522-5839. DOI: 10.1038/s42256-021-00307-0 (cit. on p. 3).
- [189] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4 (cit. on pp. 32, 33, 54, 76, 84, 85).
- [190] Tobias Roß, Annika Reinke, Peter M. Full, Martin Wagner, Hannes Kenngott, Martin Apitz, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, Pierangela Bruno, Pablo Arbeláez, Gui-Bin Bian, Sebastian Bodenstedt, Jon Lindström Bolmgren, Laura Bravo-Sánchez, Hua-Bin Chen, Cristina González, Dong Guo, Pål Halvorsen, Pheng-Ann Heng, Enes Hosgor, Zeng-Guang Hou, Fabian Isensee, Debesh Jha, Tingting Jiang, Yueming Jin, Kadir Kirtac, Sabrina Kletz, Stefan Leger, Zhixuan Li, Klaus H. Maier-Hein, Zhen-Liang Ni, Michael A. Riegler, Klaus Schoeffmann, Ruohua Shi, Stefanie Speidel, Michael Stenzel, Isabell Twick, Gutai Wang, Jiacheng Wang, Liansheng Wang, Lu Wang, Yujie Zhang, Yan-Jie Zhou, Lei Zhu, Manuel Wiesenfarth, Annette Kopp-Schneider, Beat P. Müller-Stich, and Lena Maier-Hein. “Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the ROBUST-MIS 2019 chal-



- 
- lenge". In: *Medical Image Analysis* 70 (May 2021), p. 101920. ISSN: 1361-8415. DOI: 10.1016/j.media.2020.101920. URL: <https://www.sciencedirect.com/science/article/pii/S136184152030284X> (cit. on pp. 51, 56, 152).
- [191] Fabien Sanglard. *Floating Point Visually Explained*. Aug. 29, 2017. URL: [https://fabiansanglard.net/floating\\_point\\_visually\\_explained](https://fabiansanglard.net/floating_point_visually_explained) (visited on 01/22/2024) (cit. on p. 37).
- [192] Ciro Santilli. *What is a subnormal floating point number?* Nov. 8, 2018. URL: <https://stackoverflow.com/a/53203428> (visited on 01/19/2024) (cit. on p. 35).
- [193] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. "The Graph Neural Network Model". In: *IEEE Transactions on Neural Networks* 20.1 (2009), pp. 61–80. DOI: 10.1109/TNN.2008.2005605 (cit. on p. 54).
- [194] Paul Scheikl, Stefan Laschewski, Anna Kisilenko, Tornike Davitashvili, Benjamin Müller, Manuela Capek, Beat Müller, Martin Wagner, and Franziska Ullrich. "Deep learning for semantic segmentation of organs and tissues in laparoscopic surgery". In: *Current Directions in Biomedical Engineering* 6 (Sept. 2020), p. 20200016. DOI: 10.1515/cdbme-2020-0016 (cit. on pp. 51, 56).
- [195] Marcel André Schneider, Daniel Gero, Matteo Müller, Karoline Horisberger, Andreas Rickenbacher, and Matthias Turina. "Inequalities in access to minimally invasive general surgery: a comprehensive nationwide analysis across 20 years". In: *Surgical Endoscopy* 35.11 (Nov. 1, 2021), pp. 6227–6243. ISSN: 1432-2218. DOI: 10.1007/s00464-020-08123-0 (cit. on pp. 17, 169).
- [196] Nicholas Schreck. "Empirical decomposition of the explained variation in the variance components form of the mixed model". In: *bioRxiv* (2019). DOI: 10.1101/2019.12.28.890061. eprint: <https://www.biorxiv.org/content/early/2019/12/30/2019.12.28.890061.full.pdf>. URL: <https://www.biorxiv.org/content/early/2019/12/30/2019.12.28.890061> (cit. on p. 133).
- [197] Silvia Seidlitz, Jan Sellner, Jan Odenhal, Berkin Özdemir, Alexander Studier-Fischer, Samuel Knödler, Leonardo Ayala, Tim J. Adler, Hannes G. Kenngott, Minu Tizabi, Martin Wagner, Felix Nickel, Beat P. Müller-Stich, and Lena Maier-Hein. *Robust deep learning-based semantic organ segmentation in hyperspectral images*. Oral presentation at the 13th international conference on Information Processing in Computer-Assisted Interventions (IPCAI), Tokyo, Japan. June 7, 2022 (cit. on pp. 166, 172).
- [198] Silvia Seidlitz, Jan Sellner, Jan Odenhal, Berkin Özdemir, Alexander Studier-Fischer, Samuel Knödler, Leonardo Ayala, Tim J. Adler, Hannes G. Kenngott, Minu Tizabi, Martin Wagner, Felix Nickel, Beat P. Müller-Stich, and Lena Maier-Hein. "Robust deep learning-based semantic organ segmentation in hyperspectral

- images”. In: *Medical Image Analysis* 80 (Aug. 2022), p. 102488. ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102488 (cit. on pp. 5, 10, 51, 56–59, 76, 81, 82, 103, 111, 113, 114, 116–119, 132, 135, 151, 153, 155, 157, 166, 172, 180–183).
- [199] Jan Sellner. *Introduction to convolution with link to filters in computer vision*. Blog. Oct. 19, 2016. URL: [https://www.jansellner.net/blog/Introduction\\_to\\_convolution\\_with\\_link\\_to\\_filters\\_in\\_computer\\_vision](https://www.jansellner.net/blog/Introduction_to_convolution_with_link_to_filters_in_computer_vision) (cit. on pp. 27, 29).
- [200] Jan Sellner and Silvia Seidlitz. *Hyperspectral Tissue Classification*. Version v0.0.15. Feb. 5, 2024. DOI: 10.5281/zenodo.6577614. URL: <https://github.com/IMSY-DKFZ/htc> (cit. on pp. 2, 5, 8, 94, 97, 166, 167).
- [201] Jan Sellner, Silvia Seidlitz, and Lena Maier-Hein. *Dealing with I/O bottlenecks in high-throughput model training*. Poster presented at the PyTorch Conference 2023, San Francisco, United States of America. Oct. 16, 2023. URL: [https://e130-hyperspectral-tissue-classification.s3.dkfz.de/figures/PyTorchConference\\_Poster.pdf](https://e130-hyperspectral-tissue-classification.s3.dkfz.de/figures/PyTorchConference_Poster.pdf) (cit. on pp. 5, 76, 77, 79, 80, 100, 103, 166, 172).
- [202] Jan Sellner, Silvia Seidlitz, Alexander Studier-Fischer, Alessandro Motta, Berkin Özdemir, Beat Peter Müller-Stich, Felix Nickel, and Lena Maier-Hein. “Semantic Segmentation of Surgical Hyperspectral Images Under Geometric Domain Shifts”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Cham: Springer Nature Switzerland, 2023, pp. 618–627. ISBN: 978-3-031-43996-4. DOI: 10.1007/978-3-031-43996-4\_59 (cit. on pp. 5, 55, 87, 88, 127, 129, 134, 138, 139, 141, 158, 167, 172, 188, 189).
- [203] Connor Shorten and Taghi M. Khoshgoftaar. “A survey on Image Data Augmentation for Deep Learning”. In: *Journal of Big Data* 6.1 (July 6, 2019), p. 60. ISSN: 2196-1115. DOI: 10.1186/s40537-019-0197-0 (cit. on p. 55).
- [204] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. “Semantic texton forests for image categorization and segmentation”. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 2008, pp. 1–8. DOI: 10.1109/CVPR.2008.4587503 (cit. on p. 53).
- [205] Alexey A. Shvets, Alexander Rakhlin, Alexandr A. Kalinin, and Vladimir I. Iglovikov. “Automatic Instrument Segmentation in Robot-Assisted Surgery using Deep Learning”. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2018, pp. 624–628. DOI: 10.1109/ICMLA.2018.00100 (cit. on p. 56).

- 
- [206] Patrice Y. Simard, Dave Steinkraus, and John C. Platt. “Best practices for convolutional neural networks applied to visual document analysis”. In: *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* 2003, pp. 958–963. DOI: 10.1109/ICDAR.2003.1227801 (cit. on p. 131).
  - [207] Krishna Kumar Singh and Yong Jae Lee. “Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-Supervised Object and Action Localization”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, pp. 3544–3553 (cit. on p. 130).
  - [208] Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. “Don’t Decay the Learning Rate, Increase the Batch Size”. In: Feb. 2018. URL: <https://openreview.net/forum?id=B1Yy1BxCZ> (cit. on p. 86).
  - [209] Irwin Sobel and Gary Feldman. *An Isotropic 3x3 Image Gradient Operator*. June 14, 2015. DOI: 10.13140/RG.2.1.1912.4965 (cit. on p. 28).
  - [210] Stefanie Speidel, Lena Maier-Hein, Danail Stoyanov, Sebastian Bodenstedt, Annika Reinke, Sophia Bano, Alexander Jenke, Martin Wagner, Marie Daum, Ala Tabibian, Adrito Das, Yitong Zhang, Francisco Vasconcelos, Dimitris Psychogios, Danyal Z. Khan, Hani J. Marcus, Aneeq Zia, Xi Liu, Kiran Bhattacharyya, Ziheng Wang, Max Berniker, Conor Perreault, Anthony Jarc, Anand Malpani, Kimberly Glock, Haozheng Xu, Chi Xu, Baoru Huang, and Stamatia Giannarou. *Endoscopic Vision Challenge 2023*. Sept. 2023. DOI: 10.5281/zenodo.8315050 (cit. on p. 51).
  - [211] Ace St John, Ilaria Caturegli, Natalia S Kubicki, and Stephen M Kavic. “The Rise of Minimally Invasive Surgery: 16 Year Analysis of the Progressive Replacement of Open Surgery with Laparoscopy”. In: *JLS* 24.4 (Oct. 2020). DOI: 10.4293/JLS.2020.00076 (cit. on p. 17).
  - [212] Susan Standring, ed. *Gray’s anatomy*. 42nd ed. Gray’s Anatomy. London, England: Elsevier Health Sciences, Oct. 2020. ISBN: 978-0-7020-7705-0 (cit. on p. 15).
  - [213] Andrew Stockman and Lindsay T. Sharpe. “The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype”. In: *Vision Research* 40.13 (2000), pp. 1711–1737. ISSN: 0042-6989. DOI: 10.1016/S0042-6989(00)00021-3. URL: <https://www.sciencedirect.com/science/article/pii/S0042698900000213> (cit. on p. 23).
  - [214] Alexander Studier-Fischer, Silvia Seidlitz, Jan Sellner, Marc Bressan, Berkin Özdemir, Leonardo Ayala, Jan Odenthal, Samuel Knoedler, Karl-Friedrich Kowalewski, Caelán Max Haney, Gabriel Salg, Maximilian Dietrich, Hannes Kenngott, Ines Gockel, Thilo Hackert, Beat Peter Müller-Stich, Lena Maier-Hein, and Felix Nickel. “HeiPorSPECTRAL - the Heidelberg Porcine HyperSPECTRAL Imaging Dataset of 20 Physiological Organs”. In: *Scientific Data* 10.1 (June 24, 2023), p. 414. ISSN: 2052-4463. DOI: 10.1038/s41597-023-02315-8. URL: <https://heiporspectral.org> (cit. on pp. 5, 59, 65, 94, 101, 160, 166, 171).

- [215] Alexander Studier-Fischer, Silvia Seidlitz, Jan Sellner, Berkin Özdemir, Manuel Wiesenfarth, Leonardo Ayala, Jan Odenthal, Samuel Knödler, Karl Friedrich Kowalewski, Caelán Max Haney, Isabella Camplisson, Maximilian Dietrich, Karsten Schmidt, Gabriel Alexander Salg, Hannes Götz Kenngott, Tim Julian Adler, Nicholas Schreck, Annette Kopp-Schneider, Klaus Maier-Hein, Lena Maier-Hein, Beat Peter Müller-Stich, and Felix Nickel. “Spectral organ fingerprints for machine learning-based intraoperative tissue classification with hyperspectral imaging in a porcine model”. In: *Scientific Reports* 12.1 (June 30, 2022), p. 11028. ISSN: 2045-2322. DOI: 10.1038/s41598-022-15040-w (cit. on pp. 5, 48, 59, 73, 89, 92, 93, 95, 132, 133, 165, 171, 172).
- [216] Jonah J. Stulberg, Reiping Huang, Lindsey Kreutzer, Kristen Ban, Bradley J. Champagne, Scott R. Steele, Julie K. Johnson, Jane L. Holl, Caprice C. Greenberg, and Karl Y. Bilimoria. “Association Between Surgeon Technical Skills and Patient Outcomes”. In: *JAMA Surgery* 155.10 (Oct. 1, 2020), pp. 960–968. ISSN: 2168-6254. DOI: 10.1001/jamasurg.2020.3007 (cit. on p. 17).
- [217] Aqeel Al-Surmi, Rahmita Wirza, Ramlan Mahmod, Fatimah Khalid, and Mohd Zamrin Dimon. “A new human heart vessel identification, segmentation and 3D reconstruction mechanism”. In: *Journal of Cardiothoracic Surgery* 9.1 (Oct. 2014), p. 161. ISSN: 1749-8090. DOI: 10.1186/s13019-014-0161-1 (cit. on p. 56).
- [218] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. “Rethinking the Inception Architecture for Computer Vision”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308 (cit. on p. 73).
- [219] Mingxing Tan and Quoc Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 6105–6114. URL: <https://proceedings.mlr.press/v97/tan19a.html> (cit. on pp. 76, 83–85).
- [220] MITK Team. *MITK*. Version v2023.12. Nov. 2023. URL: <https://github.com/MITK/MITK> (cit. on p. 67).
- [221] PyTorch Development Team. *Automatic Mixed Precision*. 2023. URL: [https://pytorch.org/tutorials/recipes/recipes/amp\\_recipe.html](https://pytorch.org/tutorials/recipes/recipes/amp_recipe.html) (visited on 01/20/2024) (cit. on p. 42).
- [222] PyTorch Development Team. *Automatic Mixed Precision package*. 2023. URL: <https://pytorch.org/docs/stable/amp.html> (visited on 01/20/2024) (cit. on pp. 41, 42).

- 
- [223] PyTorch Development Team. *Introduction to torch.compile*. 2023. URL: [https://pytorch.org/tutorials/intermediate/torch\\_compile\\_tutorial.html](https://pytorch.org/tutorials/intermediate/torch_compile_tutorial.html) (visited on 02/02/2024) (cit. on p. 50).
  - [224] PyTorch Development Team. *Performance Tuning Guide*. 2023. URL: [https://pytorch.org/tutorials/recipes/recipes/tuning\\_guide.html](https://pytorch.org/tutorials/recipes/recipes/tuning_guide.html) (visited on 11/22/2023) (cit. on pp. 50, 78).
  - [225] S. Thunell. “Porphyrins, porphyrin metabolism and porphyrias. I. Update”. In: *Scandinavian Journal of Clinical and Laboratory Investigation* 60.7 (2000). PMID: 11202048, pp. 509–540. DOI: 10.1080/003655100448310 (cit. on p. 48).
  - [226] Shoji Tominaga, Shogo Nishi, and Ryo Ohtera. “Measurement and Estimation of Spectral Sensitivity Functions for Mobile Phone Cameras”. In: *Sensors* 21.15 (2021). ISSN: 1424-8220. DOI: 10.3390/s21154985. URL: <https://www.mdpi.com/1424-8220/21/15/4985> (cit. on p. 23).
  - [227] Stojan Trajanovski, Caifeng Shan, Pim J. C. Weijtmans, Susan G. Brouwer de Koning, and Theo J. M. Ruers. “Tongue Tumor Detection in Hyperspectral Images Using Deep Learning Semantic Segmentation”. In: *IEEE transactions on bio-medical engineering* 68.4 (Apr. 2021), pp. 1330–1340. ISSN: 1558-2531. DOI: 10.1109/TBME.2020.3026683 (cit. on pp. 53, 54, 56, 58).
  - [228] Stojan Trajanovski, Caifeng Shan, Pim J. C. Weijtmans, Susan G. Brouwer de Koning, and Theo J. M. Ruers. *Tumor Semantic Segmentation in Hyperspectral Images using Deep Learning*. 2019. URL: <https://openreview.net/forum?id=ryeAXGw79V> (cit. on pp. 54, 56).
  - [229] Ava Vali, Sara Comai, and Matteo Matteucci. “Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data: A Review”. In: *Remote Sensing* 12.15 (Jan. 2020), p. 2495. DOI: 10.3390/rs12152495. URL: <https://www.mdpi.com/2072-4292/12/15/2495> (cit. on p. 52).
  - [230] Gaël Varoquaux and Veronika Cheplygina. “Machine learning for medical imaging: methodological failures and recommendations for the future”. In: *npj Digital Medicine* 5.1 (Apr. 12, 2022), p. 48. ISSN: 2398-6352. DOI: 10.1038/s41746-022-00592-y (cit. on p. 3).
  - [231] An Wang, Mobarakol Islam, Mengya Xu, and Hongliang Ren. “Rethinking Surgical Instrument Segmentation: A Background Image Can Be All You Need”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Ed. by Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li. Cham: Springer Nature Switzerland, 2022, pp. 355–364. ISBN: 978-3-031-16449-1 (cit. on pp. 56, 87).
  - [232] Lihong V Wang and Hsin-I Wu. *Biomedical Optics: Principles and Imaging*. Wiley, Sept. 2012. ISBN: 978-0-470-17700-6 (cit. on pp. 19–21).

- [233] Yu Winston Wang, Nicholas P. Reder, Soyoung Kang, Adam K. Glaser, and Jonathan T.C. Liu. “Multiplexed Optical Imaging of Tumor-Directed Nanoparticles: A Review of Imaging Systems and Approaches”. In: *Nanotheranostics* 1 (2017), pp. 369–388. doi: 10.7150/ntno.21136. URL: <https://www.ntno.org/v01p0369.htm> (cit. on p. 19).
- [234] Christian Weber and Heidi Noels. “Atherosclerosis: current pathogenesis and therapeutic options”. In: *Nature Medicine* 17.11 (Nov. 1, 2011), pp. 1410–1422. ISSN: 1546-170X. doi: 10.1038/nm.2538 (cit. on p. 2).
- [235] Joan Webster and Abdullah Alghamdi. “Use of plastic adhesive drapes during surgery for preventing surgical site infection”. In: *Cochrane Database Syst Rev* 2015.4 (Apr. 2015), p. CD006353. doi: 10.1002/14651858.CD006353.pub4 (cit. on p. 127).
- [236] Manuel Wiesenfarth, Annika Reinke, Bennett A. Landman, Matthias Eisenmann, Laura Aguilera Saiz, M. Jorge Cardoso, Lena Maier-Hein, and Annette Kopp-Schneider. “Methods and open-source toolkit for analyzing and visualizing challenge results”. In: *Scientific Reports* 11.1 (Jan. 2021), p. 2369. ISSN: 2045-2322. doi: 10.1038/s41598-021-82017-6. URL: <https://www.nature.com/articles/s41598-021-82017-6> (cit. on pp. 107, 113, 141, 180, 181, 188).
- [237] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C. ’t Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3.1 (Mar. 15, 2016), p. 160018. ISSN: 2052-4463. doi: 10.1038/sdata.2016.18 (cit. on p. 94).
- [238] Michael T. Wilson and Brandon J. Reeder. “Oxygen-binding haem proteins”. In: *Experimental Physiology* 93.1 (2008), pp. 128–132. doi: 10.1113/expphysiol.2007.039735 (cit. on p. 48).
- [239] Sebastian Josef Wirkert. “Multispectral image analysis in laparoscopy – A machine learning approach to live perfusion monitoring”. PhD thesis. Karlsruhe Institute of Technology, Jan. 18, 2018 (cit. on p. 20).

- 
- [240] World Medical Association. “World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects”. In: *Bull World Health Organ* 79.4 (July 2003), pp. 373–374 (cit. on p. 61).
  - [241] Yue Wu, Zhongyuan Xu, Wenjian Yang, Zhiqiang Ning, and Hao Dong. “Review on the Application of Hyperspectral Imaging Technology of the Exposed Cortex in Cerebral Surgery”. In: *Frontiers in Bioengineering and Biotechnology* 10 (2022). ISSN: 2296-4185. DOI: 10.3389/fbioe.2022.906728. URL: <https://www.frontiersin.org/articles/10.3389/fbioe.2022.906728> (cit. on p. 19).
  - [242] Pavel Yakubovskiy. *Segmentation Models Pytorch*. 2020. URL: [https://github.com/qubvel/segmentation\\_models.pytorch](https://github.com/qubvel/segmentation_models.pytorch) (cit. on p. 83).
  - [243] Xiaofei Yang, Yunming Ye, Xutao Li, Raymond Y. K. Lau, Xiaofeng Zhang, and Xiaohui Huang. “Hyperspectral Image Classification With Deep Learning Models”. In: *IEEE Transactions on Geoscience and Remote Sensing* 56.9 (Sept. 2018), pp. 5408–5423. ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2018.2815613. URL: <https://ieeexplore.ieee.org/document/8340197> (cit. on p. 52).
  - [244] Varduhi Yeghiazaryan and Irina Voiculescu. “Family of boundary overlap metrics for the evaluation of medical image segmentation”. In: *Journal of Medical Imaging* 5.1 (Jan. 2018), p. 015006. ISSN: 2329-4302. DOI: 10.1117/1.JMI.5.1.015006. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5817231> (cit. on p. 105).
  - [245] Jonghee Yoon. “Hyperspectral Imaging for Clinical Applications”. In: *BioChip Journal* 16.1 (Mar. 1, 2022), pp. 1–12. ISSN: 2092-7843. DOI: 10.1007/s13206-021-00041-0 (cit. on p. 2).
  - [246] Shiqi Yu, Sen Jia, and Chunyan Xu. “Convolutional Neural Networks for Hyperspectral Image Classification”. In: *Neurocomputing* 219 (Jan. 2017), pp. 88–98. ISSN: 09252312. DOI: 10.1016/j.neucom.2016.09.010. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0925231216310104> (cit. on p. 52).
  - [247] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. “CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 6022–6031. ISBN: 978-1-7281-4803-8 (cit. on pp. 55, 131).
  - [248] Hairong Zhang, Daniel Salo, David M Kim, Sergey Komarov, Yuan-Chuan Tai, and Mikhail Y Berezin. “Penetration depth of photons in biological tissues from hyperspectral imaging in shortwave infrared in transmission and reflection geometries”. In: *J Biomed Opt* 21.12 (Dec. 2016), p. 126006 (cit. on p. 159).

- [249] LeChao Zhang, DanFei Huang, XiaoJing Chen, LiBin Zhu, XiaoQing Chen, Zhong-Hao Xie, GuangZao Huang, JunZhao Gao, Wen Shi, and GuiHua Cui. “Visible near-infrared hyperspectral imaging and supervised classification for the detection of small intestinal necrosis tissue in vivo”. In: *Biomed. Opt. Express* 13.11 (Nov. 2022), pp. 6061–6080. DOI: 10.1364/BOE.470202. URL: <https://opg.optica.org/boe/abstract.cfm?URI=boe-13-11-6061> (cit. on pp. 48, 49).
- [250] Yan Zhang, Sebastian Wirkert, Justin Iszatt, Hannes Kenngott, Martin Wagner, Benjamin Mayer, Christian Stock, Neil T. Clancy, Daniel S. Elson, and Lena Maier-Hein. “Tissue classification for laparoscopic image understanding based on multispectral texture analysis”. In: *Journal of Medical Imaging* 4.1 (2017), p. 015001. DOI: 10.1117/1.JMI.4.1.015001 (cit. on pp. 48, 49).
- [251] Yating Zhang, Xiaoqian Wu, Li He, Chan Meng, Shunda Du, Jie Bao, and Yongchang Zheng. “Applications of hyperspectral imaging in the detection and diagnosis of solid tumors”. In: *Translational Cancer Research* 9.2 (Feb. 2020). Publisher: AME Publishing Company. ISSN: 2219-6803, 2218-676X. DOI: 10.21037/tcr.2019.12.53. URL: <https://tcr.amegroups.com/article/view/34678> (cit. on pp. 2, 162, 163).
- [252] Wenli Zheng, Chaojian Wang, Shufang Chang, Shiwu Zhang, and Ronald X. Xu. “Hyperspectral wide gap second derivative analysis for in vivo detection of cervical intraepithelial neoplasia”. In: *Journal of Biomedical Optics* 20.12 (2015), p. 121303. DOI: 10.1117/1.JBO.20.12.121303 (cit. on p. 48).
- [253] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. “Random Erasing Data Augmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.07 (Apr. 2020), pp. 13001–13008. ISSN: 2374-3468 (cit. on p. 130).
- [254] Qiqi Zhu, Weihuan Deng, Zhuo Zheng, Yanfei Zhong, Qingfeng Guan, Weihua Lin, Liangpei Zhang, and Deren Li. “A Spectral-Spatial-Dependent Global Learning Framework for Insufficient and Imbalanced Hyperspectral Image Classification”. In: *IEEE Transactions on Cybernetics* (2021). Conference Name: IEEE Transactions on Cybernetics, pp. 1–15. ISSN: 2168-2275. DOI: 10.1109/TCYB.2021.3070577 (cit. on p. 52).
- [255] Karel J. Zuzak, Sabira C. Naik, George Alexandrakis, Doyle Hawkins, Khosrow Behbehani, and Edward H. Livingston. “Characterization of a Near-Infrared Laparoscopic Hyperspectral Imaging System for Minimally Invasive Surgery”. In: *Analytical Chemistry* 79.12 (June 1, 2007), pp. 4709–4715. ISSN: 0003-2700. DOI: 10.1021/ac070367n (cit. on p. 2).