

*aus dem*  
Deutschen Krebsforschungszentrum Heidelberg  
(Vorstandsvorsitzender: Prof. Dr. Michael Baumann)

Abteilung Computer-assistierte Medizinische Interventionen  
(Abteilungsleiterin: Prof. Dr. Lena Maier-Hein)

---

# **Surgical data science in endoscopic surgery**

---

Inauguraldissertation  
zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum.)  
*an der*  
Medizinischen Fakultät Heidelberg  
*der*  
Ruprecht-Karls-Universität

*vorgelegt von*  
Tobias Roß

*aus*  
Bad Neustadt an der Saale

2020



Dekan: Prof. Dr. Hans-Georg Kräusslich  
Doktormutter: Prof. Dr. Lena Maier-Hein

For my wife **Rawan** and our beautiful son **Nael**.



# Contents

<b>Acronyms</b>	<b>3</b>
<b>List of Figures</b>	<b>6</b>
<b>List of Tables</b>	<b>7</b>
<b>1 Introduction</b>	<b>9</b>
1.1 Motivation . . . . .	9
1.2 Objectives . . . . .	13
1.3 Outline . . . . .	14
<b>2 Materials and Methods</b>	<b>15</b>
2.1 Machine learning . . . . .	16
2.1.1 Basic machine learning models . . . . .	16
2.1.2 Statistical modeling methods . . . . .	16
2.1.3 Fundamentals of neural networks . . . . .	21
2.1.4 Neural network architectures used in this thesis . . . . .	28
2.2 Related work . . . . .	36
2.2.1 Self-supervised learning in surgical data science . . . . .	36
2.2.2 Available datasets for surgical data science . . . . .	37
2.2.3 Multi-instance medical instrument segmentation . . . . .	38
2.2.4 Conclusions . . . . .	40
<b>3 Results</b>	<b>41</b>
3.1 Potential of unlabeled data . . . . .	43
3.1.1 Introduction . . . . .	43
3.1.2 Methods . . . . .	44
3.1.3 Results . . . . .	51
3.1.4 Discussion . . . . .	53
3.1.5 Conclusion . . . . .	55
3.2 Quality controlled dataset generation . . . . .	57
3.2.1 Introduction . . . . .	57
3.2.2 Methods . . . . .	58
3.2.3 Results . . . . .	62
3.2.4 Discussion . . . . .	63

3.2.5	Conclusion . . . . .	64
3.3	Comparative validation of multi-instance instrument segmentation . . . . .	65
3.3.1	Introduction . . . . .	65
3.3.2	Methods . . . . .	66
3.3.3	Results . . . . .	69
3.3.4	Discussion . . . . .	75
3.3.5	Conclusion . . . . .	81
3.4	Effects of image characteristics on the algorithm performance . . . . .	83
3.4.1	Introduction . . . . .	83
3.4.2	Methods . . . . .	84
3.4.3	Results . . . . .	86
3.4.4	Discussion . . . . .	88
3.4.5	Conclusion . . . . .	89
3.5	Multi-instance segmentation of medical instruments . . . . .	91
3.5.1	Introduction . . . . .	91
3.5.2	Methods . . . . .	92
3.5.3	Results . . . . .	98
3.5.4	Discussion . . . . .	101
3.5.5	Conclusion . . . . .	102
<b>4</b>	<b>Discussion</b>	<b>105</b>
4.1	Summary of Contributions . . . . .	105
4.2	Discussion of results . . . . .	107
4.2.1	Potential of unlabeled data . . . . .	107
4.2.2	Quality controlled dataset generation . . . . .	108
4.2.3	Comparative validation of multi-instance instrument segmentation . . . . .	108
4.2.4	Effects of image characteristics on the algorithm performance . . . . .	109
4.2.5	Multi-instance segmentation of medical instruments . . . . .	109
4.2.6	Transition into clinical routine . . . . .	110
4.3	Conclusion . . . . .	110
<b>5</b>	<b>Summary</b>	<b>113</b>
<b>6</b>	<b>Zusammenfassung</b>	<b>115</b>
<b>7</b>	<b>Own contributions</b>	<b>135</b>
7.1	Own share in data acquisition and analysis . . . . .	135
7.2	Own publications . . . . .	137
	<b>Statutory Declaration</b>	<b>143</b>

# Acronyms

**1D** 1-Dimensional  
**SDS** Surgical Data Science  
**ML** Machine Learning  
**MSE** Mean Squared Error  
**MAE** Mean Absolute Error  
**PCA** Principal Component Analysis  
**DSC** Sørensen Dice Similarity Coefficient  
**NSD** Normalized Surface Dice  
**CE** Cross Entropy  
**GD** Gradient Descent  
**SGD** Stochastic Gradient Descent  
**ResNet** Residual Neural Network  
**GAN** Generative Adversarial Network  
**fps** frames per second  
**HD** Hausdorff Distance  
**GLM** Generalized Linear Models  
**GLMM** General Linear Mixed Models  
**LMM** Linear Mixed Models  
**OR** Odds Ratio  
**RNN** Recurrent Neural Network  
**LSTM** Long short-term Memory Network  
**CRF** Conditional Random Field  
**ROI** Region of Interest  
**GUI** Graphical User Interface



# List of Figures

1.1	Available amount of training data in datasets . . . . .	10
1.2	Research field surgical data science . . . . .	12
2.1	Supervised vs unsupervised learning . . . . .	17
2.2	Illustration of a Markov Random Field (MRF) . . . . .	20
2.3	Illustration of a Conditional Random Field (CRF) . . . . .	21
2.4	Illustration of a 2D convolution . . . . .	23
2.5	Illustration of the max/average (un)pooling layer . . . . .	24
2.6	Activation functions . . . . .	25
2.7	Difference $\mathcal{L}1$ and $\mathcal{L}2$ loss . . . . .	26
2.8	Architecture - U-Net . . . . .	29
2.9	Residual building block . . . . .	30
2.10	Architecture - Generative adversarial network . . . . .	31
2.11	Traditional neural network vs recurrent neural network . . . . .	32
2.12	Architecture of a Long short-term memory network . . . . .	33
2.13	Illustration of the Mask R-CNN architecture . . . . .	35
3.1	Limited generalization capabilities of deep learning algorithms . . . . .	44
3.2	Self-supervised learning: Concept overview . . . . .	45
3.3	Lab log color distributions of different datasets . . . . .	47
3.4	Self-supervised learning: Conditional generative adversarial network for the re-colorization task . . . . .	48
3.5	Re-colored images after self-supervision was applied . . . . .	51
3.6	Performance of the pre-training as a function of the training size . . . . .	52
3.7	Effect of pre-training domain . . . . .	53
3.8	Image characteristics annotation tool . . . . .	61
3.9	ROBUST-MIS dataset structure . . . . .	62
3.10	ROBUST-MIS dataset image characteristics . . . . .	63
3.11	Challenge test stages . . . . .	67
3.12	Inter-rater agreement . . . . .	70
3.13	Inter-rater agreement - Worst cases . . . . .	71
3.14	Algorithm performances of all three stages . . . . .	74
3.15	Algorithm performances for Stage 3 . . . . .	75
3.16	Visualization of the ranking stability . . . . .	78

3.17	Expert vs. algorithms, a baseline . . . . .	79
3.18	Presence of image characteristics in training and test data . . . . .	87
3.19	Impact of image characteristic: one instance . . . . .	88
3.20	Impact of image characteristic: two instances . . . . .	89
3.21	Impact of image characteristic: three instances . . . . .	90
3.22	Concept of the multi-instance segmentation approach . . . . .	93
3.23	Segmentation with bounding box problematic . . . . .	95
3.24	Overlapping instruments . . . . .	96
3.25	Effect of post-processing . . . . .	96
3.26	Effect of including temporal information and instrument likelihood . . . . .	99
3.27	Multi-instance segmentation: Comparison to state-of-the-art methods . . . . .	100
3.28	Algorithm rankings on characteristics . . . . .	101
3.29	Algorithm performance on characteristics . . . . .	103

# List of Tables

3.1	Descriptive statistic comparing the re-colorization task to the state-of-the-art pre-training . . . . .	53
3.2	Excerpts from the labeling protocol . . . . .	60
3.3	Image characteristics . . . . .	61
3.4	Overview generated data per surgery type . . . . .	63
3.5	Case distribution of the data with frames per stage and surgery. . . . .	67
3.6	Overview of submitted methods of all participating teams. . . . .	72
3.7	Descriptive statistics algorithm performances . . . . .	76
3.8	Challenge rankings for the MI_DSC and MI_NSD metrics . . . . .	77
3.9	Confusion matrix for two events (State A and Event B) regarding their presence or absence. . . . .	85
3.10	Ablation study configuration . . . . .	98
3.11	Ablation study . . . . .	99
3.12	Artefacts handling . . . . .	101





# 1

## Introduction

### 1.1 Motivation

Surgical Data Science (SDS) is a relatively young research discipline, which was officially defined for the first time in a consensus paper of an international consortium of leading researchers [Maier-Hein et al., 2017b]. The authors, who have grouped themselves in the SDS-initiative<sup>1</sup>, describe the main objective of SDS to "observe all that is occurring within and around the treatment process [...] to improve the quality of interventional healthcare and its value by capturing, organizing, analyzing and modeling data" [Maier-Hein et al., 2017b]. Example SDS applications range from robotic assistance [Chen et al., 2020, Haidegger, 2019], context-aware assistance, surgical skill assessment [Nguyen et al., 2019, Funke et al., 2019, Lin et al., 2019], to educational and training systems [Augestad et al., 2020, Singh et al., 2015, Prebay et al., 2019].

In order to provide reliable assistance in a complex environment during surgery, SDS algorithms have to deal with a dynamic and continuously changing system with huge variations. Examples of such variations are, e.g., differences between patients, different hardware setups (e.g., camera type, surgical instruments), changes in anatomy due to surgical manipulation, surgeons' preferences, and different types of surgeries (e.g., Proctectomy, Nephrectomy). Depending on the use-cases, already small variations can have huge impacts on the algorithms' performance and make it challenging for an algorithm to meet the requirements of robust and reliable results [Rozsa et al., 2016, Lei et al., 2018]. However, fulfilling these requirements became feasible for the first time with the very recent success of deep learning, a method of machine learning, which led to a breakthrough in the development of SDS applications [Maier et al., 2019, Maier-Hein et al., 2017b]. Unlike conventional algorithms, machine learning algorithms can solve problems without being explicitly programmed to do so. They do this with the help of mathematical models fit to the data and the minimization of an objective [Koza et al., 1996].

---

<sup>1</sup><http://www.surgical-data-science.org>

Although machine learning methods have so far delivered promising results, the SDS-initiative of leading scientists still failed in a follow-up conference 2019 to identify any SDS success stories [Maier-Hein et al., 2020a]. The reasons for this are varied, but the scientists have particularly emphasized one cause as being related to the way machine learning works [Maier-Hein et al., 2020a]. Machine learning algorithms usually rely on statistical methods that have special requirements on the data, namely (1) there has to be a vast amount of data available, (2) it is representative of the actual problem the algorithm is intended to solve, and (3) consistent targets (annotations) are required which are the solutions that the algorithm should present as an outcome [Maier et al., 2019]. However, the target generation (labeling) is a resource and cost-intensive process, especially in the medical domain, where for many use-cases, expensive medical experts are needed, who usually are limited in time [Maier-Hein et al., 2016, Heim et al., 2018]. Besides the target generation, there are organizational, ethical, and legal difficulties, restricted access, and problems generating raw data [Esteva et al., 2021]. Even if annotations are available, their quality is often not clear, since no quality control has taken place and clear labeling instructions is not yet a standard [Maier-Hein et al., 2020a]. All of these challenges have led to a delay in applying machine learning-based algorithms in the medical domain [Chapaliuk and Zaychenko, 2020].

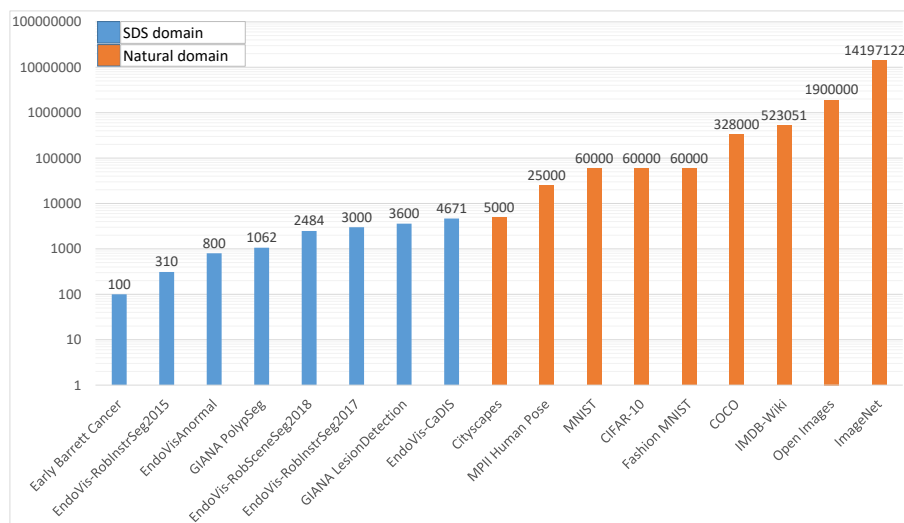


Figure 1.1: This figure shows on the x-axis famous datasets and on the y-axis (log-scale) the number of annotated data items per set. It can be seen that all SDS datasets (blue) have significantly less data than all Natural domain datasets (orange). Concerning the complexity of the tasks to be solved (e.g., MNIST: number classification vs. EndoVis-RobInstrSeg2015: instrument segmentation) one would expect that the SDS task's number of training data is larger.

This problem becomes even clearer when comparing datasets from the computer vision community with datasets from the SDS domain<sup>2</sup> (see Figure 1.1). While huge datasets like ImageNet [Deng et al., 2009] or COCO [Lin et al., 2014a] with more than 30,000 images have led to a breakthrough

<sup>2</sup>A comprehensive list of available datasets can be found in [Maier-Hein et al., 2020a]

in the computer vision community [Alom et al., 2018], the datasets available in the SDS often contain annotated data items numbering in the thousands [Allan et al., 2019]), or even only hundreds [Bodenstedt et al., 2018].

The existence of a data bottleneck is also reflected in publications that are related to SDS, where a graph of the SDS research area is visualized in Figure 1.2. In this graph, each node represents a paper, the size of the node is the number of citations, and the edges represent a citation. The graph was constructed by following all citations that came from the original publication from Maier-Hein et al. [Maier-Hein et al., 2017b]. To avoid too many edges, only papers with the following criteria were included: (1) have an assignment to the research fields "*Computer Science*" and "*Medical*" on *Semantic Scholar* and (2) have a maximum distance of five nodes to the seed paper. Interestingly, the graph forms four main research topics, where three of them are attempting to deal with only a small amount of available data, namely *Data generation*, *Weak labels* and *Domain adaptation*. Since the generation of the graph is based on simplified conditions for including or excluding literature, the graph does not claim to include all related publications. Possibly, literature that was not correctly assigned to both research fields might have been missed. Nonetheless, the graph still provides interesting insights. In total, more than 6470 papers were crawled with references, so one can assume it provides, at least, a representative subset of the entire research area.

Summarizing the findings taken from this graph, it exposes a strong need for the community for two main challenges:

### **C1: Data availability**

This challenge is driven by the problem that the available number of datasets is limited. One of the main reasons is that the generation of labels is time-intensive and access to larger datasets is challenging. Besides the labeling issue, further factors limit the availability of data for data science purposes, such as organizational rules or legal aspects [Maier-Hein et al., 2020a]. In comparison to datasets from the computer vision community often containing thousands or even millions of images (e.g., COCO [Lin et al., 2014b] with 20k images, ImageNet [Russakovsky et al., 2015] with 14 million images), medical datasets often consists only of a few hundred [Bodenstedt et al., 2018] or thousands [Allan et al., 2019]), which limits the representatives of the datasets. One additional problem is that the available datasets often had not been subjected to any quality control. Accordingly, they have inconsistencies because annotators might have labeled edge-cases differently or even had a different understanding/definition of the labeling subject. Thus, there is a need for open, quality controlled datasets.

### **C2: Managing sparse data**

This challenge is based on the same observation as *C1: Data availability*, namely, that there is not enough training data available. For this reason, there is a need for methods that enable the use of machine learning algorithms, despite the small amount of labeled data, that can still deal with edge-cases and the missing representatives.

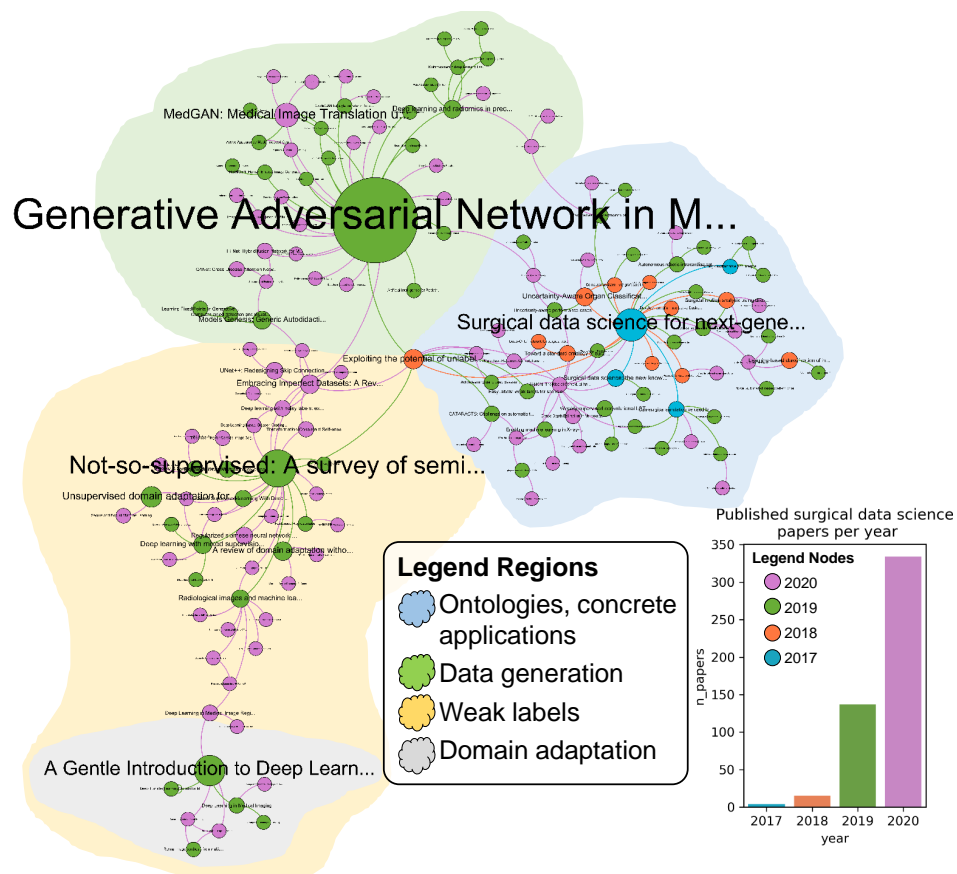


Figure 1.2: Visualization of the SDS research area as a connected graph where each node is a surgical data science paper and each edge a reference. The graph was generated by crawling<sup>a</sup> all papers from *Semantic Scholar*<sup>b</sup>, starting with the first surgical data science paper of Maier-Hein et al. [Maier-Hein et al., 2017b] as seed. For all papers that cite the seed paper, a node was created and connected via an edge. For all those newly created edges the process was recursively repeated. It can be seen that the graph is clustered into four main SDS regions, where papers in the blue area are mainly about a concrete medical application or ontologies, papers in the green about generating more data, and the yellow and grey areas are papers that deal with the problem of a small amount of data and labels.

<sup>a</sup> crawled on the 19.08.2020 <sup>b</sup> <https://www.semanticscholar.org>

## 1.2 Objectives

Oriented on the two challenges *C1: Data availability* and *C2: Managing sparse data*, all the concepts presented in this thesis are exemplarily shown using the SDS task to segment surgical instruments in laparoscopic videos. The instrument segmentation task was chosen because a sample application was sought that is as representative as possible of an SDS task. This task is particularly useful because a large number of other SDS applications are based on it (e.g., [Su et al., 2018, Law et al., 2017, Lin et al., 2019, Wang et al., 2017, Burström et al., 2019, Shvets et al., 2018, Kletz et al., 2019]), and progress in this area would therefore have a very large impact. According to the task, the five objectives of this thesis are:

***O1: Incorporation of unlabeled data into the training of deep learning models:***

The objective of incorporating unlabeled data into the training of deep learning models is to work towards *C2* by creating a method that can manage sparse data. For this purpose, it should be investigated whether unlabeled data that are massively produced during the daily routine can be used to improve the performance of state-of-the-art deep learning algorithms.

***O2: Generation of a quality controlled dataset for the task of multi-instance surgical instrument segmentation:***

The objective of generating a quality controlled dataset for the task of multi-instance surgical instrument segmentation is to increase the number of available datasets and thus work towards *C1*.

***O3: Structured assessment of state-of-the-art performance for multi-instance surgical instrument segmentation:***

The objective of the structured assessment of state-of-the-art determination for the multi-instance segmentation of surgical instruments is to work towards managing sparse data (*C2*). For this purpose, the dataset that will be generated in *O2*, should be used for investigating the performance of the available methods concerning their generalizability and robustness.

***O4: Systematic problem analysis of state-of-the-art methods for multi-instance surgical instrument segmentation:***

The objective of a systematic problem analysis of state-of-the-art methods is to work towards managing sparse data (*C2*). For this purpose the work from *O3* should be used for identifying the weaknesses of state-of-the-art methods. For this purpose, the dataset from *O2* should be extended by labels that define certain image characteristics.

***O5: Problem-driven multi-instance surgical instrument segmentation algorithm:***

The objective of a problem-driven multi-instance surgical instrument segmentation algorithm is to work towards solutions managing sparse data (*C2*). For this purpose an explicit attempt should be made to solve problems that were identified in *O4*, to show that a problem-driven algorithm development can help to better manage sparse data.

### 1.3 Outline

The thesis begins with an overview of the necessary topics to follow the results of this thesis in Chapter 2 *Materials and Methods*. It provides an introduction into basic machine learning models in Sec. 2.1.1 and continues with statistical modeling methods in Sec. 2.1.2. After that, an introduction into the fundamentals of neural networks is provided in Sec. 2.1.3, followed by the neural network architectures that are used in this thesis in Sec. 2.1.4. The chapter ends with an overview of current state-of-the-art methods in Sec. 2.2 for self-supervised learning in surgical data science, available datasets in surgical data science, and finally, the multi-instance segmentation for medical instruments in laparoscopic videos.

Chapter 3 *Results* gives a detailed presentation of the work that has been conducted towards the challenges *C1* and *C2*, as outlined in the previous section. Work towards *C1* is presented in Sec. 3.2 and work towards *C2* is presented in Sec. 3.1, Sec. 3.3, Sec. 3.4, and Sec. 3.5. In Sec. 3.1, a concept is introduced as to how unlabeled data could be used to improve the performance of instrument segmentation algorithms when only a small amount of labeled data is available. Sec. 3.2 describes the generation of a huge quality controlled dataset for the task of multi-instance instrument segmentation and Sec. 3.3 describes how this data was used for a community state-of-the-art determination for the task of multi-instance instrument segmentation. Finally, the results of this challenge were further analyzed in Sec. 3.4 where a in-depth statistical analysis was performed to estimate the effect of image artifacts on the algorithm performance to identify open challenges. The results from this section were used to develop an algorithm in Sec. 3.5 that specifically tackles such challenges.

Chapter 4 *Discussion* considers all of the developed concepts and results presented in this thesis and contains a general discussion of the advantages and limitations of these methods, and a summary of the contributions of this thesis.

A summary of the findings presented in this thesis is given in chapters 5 (English version) and 6 (German version).

# 2 | **Materials and Methods**

This chapter gives an overview of the principles needed so as to follow the content of this thesis. It is divided into two sections. The machine learning techniques used in this thesis are described in Sec. 2.1. A brief review of the current state-of-the-art in Sec. 2.2 covers the related work and explicitly points out the gaps that are addressed by the work presented in this thesis.

## 2.1 | Machine learning

The area of machine learning is a part of the research field of artificial intelligence [Goodfellow et al., 2016a]. The main property of Machine Learning (ML) algorithms is the ability to automatically generate a model that improves itself without explicitly being programmed to do so. This is done by learning from observations, also known as "training data", by actively using mathematical concepts from the field of computational statistics and optimization [Goodfellow et al., 2016a, Koza et al., 1996]. In the following, first the classification of machine learning algorithms is described in Sec. 2.1.1, which is followed by the statistical modeling methods that were used in this thesis in Sec. 2.1.2. The fundamentals of neural networks are explained in Sec. 2.1.3 and the architectures of neural networks that are based on those fundamentals are described in Sec. 2.1.4.

### 2.1.1 Basic machine learning models

Clustering of machine learning models is mainly achieved by either **how** or **what** models learn. In this section, first the **how** is described by pointing out the differences between supervised and unsupervised learning, followed by the **what**, which explains the difference between discriminative and generative models.

**Supervised and unsupervised learning** Basically, there are (besides many mixed forms) two main ways to train a model, either supervised or unsupervised (see Fig. 2.1) [Goodfellow et al., 2016a]. In the supervised setting, an algorithm learns to associate an input  $x$  with a defined output (label)  $y$ , where the collection of  $y$  normally requires manual effort by a human annotator [Goodfellow et al., 2016a]. Typical supervised tasks are, for example, the classification of animals within pictures [Miao et al., 2019, Deng et al., 2009] or the segmentation of cars in a street scene [Krause et al., 2013, Chang et al., 2015]. In the unsupervised setting, instead of manually creating labels, algorithms attempt to capture some underlying distributions from where  $x$  is generated [Goodfellow et al., 2016a]. Typical unsupervised tasks are, for example, the clustering [Mwangi et al., 2014, Caron et al., 2018] or compression of data [Tellez et al., 2019, Hoang et al., 2020].

**Discriminative and generative models** There are two main groups of machine learning algorithms, based on what they learn: *discriminative* and of *generative* models [Ng and Jordan, 2002]. The difference between the two is that discriminative models are learning decision boundaries between different targets of  $y$  that are based on an observation  $x$ . Such decision boundaries are learned in the form of the conditional distribution  $P(Y|X = x)$ . Generative models, however, are learning the conditional distribution  $P(X|Y = y)$  [Ng and Jordan, 2002]. Thus, discriminative models are used to assign a corresponding label  $y$  to an unseen data sample  $x$ , while generative models are used to generate new  $x$ , conditioned on  $y$ .

### 2.1.2 Statistical modeling methods

Machine learning and statistical modeling are two research areas with many methods and problems in common, making it difficult to draw a clear line between them [Boulesteix and



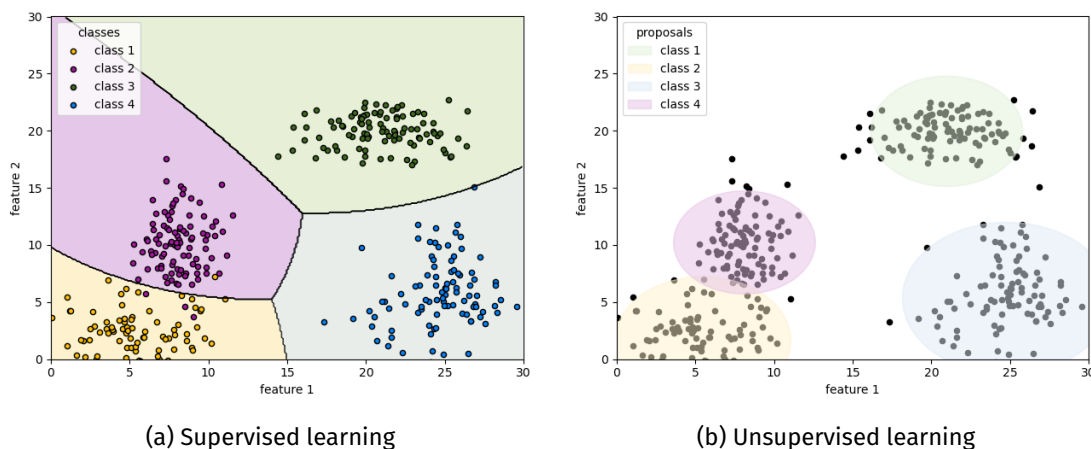


Figure 2.1: This figure shows the difference between supervised and unsupervised learning. In the supervised settings (a) each data point has an assigned label (in this case, the labels are class 1-4), and the task is to assign unlabeled data points to the correct class. In the unsupervised setting (b), no labels are given, and the task is to identify structures within the data. (In the shown model there are four possible clusters in the data).

Schmid, 2014]. One principal difference, however, can be seen in the purpose. While machine learning methods aim to make predictions that are as accurate as possible, and statistical modeling focuses more on inferring the relationships between variables [Beam and Kohane, 2018, Boulesteix and Schmid, 2014]. This section starts with Linear Mixed Effects Models, continues with Generalized Linear (Mixed), and ends with Conditional Random Fields.

### Linear Mixed Effects Models

To better understand Linear Mixed Models (LMM), one should look at first on the traditional simple linear model, as given by Eq. 2.1. Linear models consist of a target variable  $Y$ , that should be predicted with the help of explanatory variables, also called covariates  $X$ , that are weighted with an individual coefficient  $\beta$ . The design of such models forces the assumption that all covariates share the same slope and intercept [Bolker et al., 2009]. This is what is also called "fixed effects".

$$Y = X\beta + \epsilon \quad (2.1)$$

$Y$  is the target with shape  $[p \times 1]$ ,  $X$  is a matrix of  $p$  explanatory variables with the shape  $[N \times p]$ ;  $\beta$  a column vector of the regression coefficients with shape  $[p \times 1]$  and  $\epsilon$  a  $[N \times 1]$  column vector of the residuals (error term) and  $N$  is the number of data used for fitting the model.

Because for  $\epsilon$  the assumption is that  $\mathbf{E}[\epsilon] = 0$ ,  $\text{Var}[\epsilon] = \sigma^2$  and  $\epsilon \sim \mathcal{N}(0, \sigma^2)$ , it is important to note that linear mixture models can not be used for non-normally distributed data, such as binary outcomes or counts, as this data would violate the assumptions.

In comparison to a normal linear model, mixed effects models [McCulloch and Neuhaus, 2001] contain, besides the fixed effects  $\beta$ , also random effects  $u$ , which are used to describe the

variations in the observed outcome  $Y$  for a group  $J$  of repeated measurements [Harrison et al., 2018, McCulloch and Neuhaus, 2001]. The purpose of random effects is that linear models work under the assumption of non-correlated errors. However, if there could be variance on group-level assumed, the group should be included as an additional factor to avoid the correlation of the error terms within such groups (see Eq. 2.2). Applied to the problem of estimating the effect of image artifacts, based on the performance scores of multiple algorithms, fixed effects could be the artifacts, while the groups would be the image (repeated scores for each image) and the algorithm number.

$$Y = X\beta + Zu + \epsilon \quad (2.2)$$

where  $X$ ,  $Y$ ,  $\beta$  and  $\epsilon$  are the same as in Eq. 2.1;  $Z$  as the design matrix of shape  $[N \times qJ]$  with  $q$  random effects and  $J$  groups and  $u$  with the random-effects regression coefficients with shape  $[qJ \times 1]$ . One main assumption with this model is that  $u \sim \mathcal{N}(0, G)$ , where  $G$  is the variance-covariance matrix of the random effects [McCulloch and Neuhaus, 2001].

### Generalized Linear Models

As already written above, linear mixture models cannot be used for non-normally distributed data, such as binary outcomes or counts. At the same time, especially in the medical and biological domain, often such distributions are not given [Bolker et al., 2009]. Generalized linear models are a way to handle this issue by introducing a link and a variance function. The link function  $g$  (see Eq. 2.3) describes how the mean  $\mathbb{E}[Y_i] = \mu_i$  depends on the linear predictor, while the variance function  $V$  (see Eq. 2.4) describes how the variance  $\text{Var}(Y_i)$  depends on the mean.

$$g(\mu) = \eta \quad (2.3)$$

$$\text{Var}(Y_i) = \phi V(\mu) \quad (2.4)$$

where  $i \in [1, 2, \dots, N_{data}]$  and  $N_{data}$  is the total amount of data and  $\phi$  being a constant dispersion parameter. For a general linear model  $\eta$  would be for example  $\eta = \beta_0 + \beta_1 x_{1p} + \dots + \beta_p x_{ip}$ .

In case of a binomial distribution  $Y_i \sim B(p_i, n_i)$ , with  $n_i$  trials and a success probability of  $p_i$ , then  $\mathbb{E}\left[\frac{Y_i}{n_i}\right] = p_i$ . This means that  $V$  should be  $V(\mu_i) = \mu_i(1 - \mu_i)$ . As  $p_i$  is a probability and thus in a range between 0 and 1, the link function needs to map  $(0, 1) \rightarrow (-\infty, \infty)$ . For this reason, the link function reparametrizes  $p_i$  (see Eq. 2.6) with the logit function (see Eq. 2.5).

$$g(\mu_i) = \text{logit}(\mu_i) = \log \frac{\mu_i}{1 - \mu_i} \quad (2.5)$$

By using the logit function,  $p_i$  is defined as:

$$g(p_i) = \sum_{j=0}^k \beta_j x_{ij} + \epsilon \quad (2.6)$$

where  $x_i$  is a vector of  $p$  predictor variables with the shape  $[1 \times p]$ ;  $\beta$  a column vector of the regression coefficients with shape  $[p \times 1]$  and  $\epsilon$  the residuals (error term).

Because the number of parameters per sample can differ, the likelihood that needs to be maximized  $\max_{\beta} \mathcal{L}(\beta)$  is given by the products of the independent samples.

$$\mathcal{L}(\beta \mid y_1, y_2, \dots, y_n; n_1, n_2, \dots, n_n; \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \prod_{i=1}^{N_S} \binom{y_i}{n_i} \left( g^{-1} \left( \sum_{j=0}^k \beta_j x_{ij} \right) \right)^{y_i} \left( 1 - g^{-1} \left( \sum_{j=0}^k \beta_j x_{ij} \right) \right)^{n_i - y_i} \quad (2.7)$$

with  $N_S$  being the number of samples and  $j \in \{1, 2, \dots, N_P\}$ , with  $N_P$  being the number of parameters.

### Generalized Linear Mixed Models

Generalized Linear Mixed Models (GLMM) combine the properties of LMMs that involve random effects, with Generalized Linear Models (GLM)s, that enable the analysis of non-normal distributed data [Bolker et al., 2009]. In case of a binomial distribution like the GLM, the General Linear Mixed Models (GLMM) use a link function that reparametrizes  $p_i$  as a LMM and connects it via the logit link (see paragraph above).

### Conditional random fields

A Conditional Random Field (CRF) is a statistical modeling method, first presented by John Lafferty et al. [Lafferty et al., 2001], where relationships of different variables are modeled by a discriminative undirected probabilistic graph [Tran, 2008, Liu et al., 2016]. The goal of a CRF is to model the mapping from observations  $X$  (e.g., RGB colors, spatial locations) to the joint output variable  $Y = (y_1, y_2, \dots, y_{N_{class}})$  (with  $N_{class}$  being the number of classes) via the conditional distribution  $P(Y|X)$  [Liu et al., 2016].

For this purpose,  $Y$  is represented as a Markov random field  $\mathcal{G} = (\mathcal{V}, E)$ , where  $E$  is a set of edges,  $\mathcal{V}$  a set of nodes (vertices) and each node corresponds to a  $y$ , with  $y \in Y$  [Lafferty et al., 2001]. An example of a Markov random field can be seen in Fig. 2.2. One important property that has to be fulfilled when using Markov random fields is the Markov property, which states that the conditional probability distribution of a state only depends on its neighbors  $\mathcal{N}$  and is independent of variables outside of this neighborhood (see Eq. 2.8) [Tran, 2008]. An illustration of an CRF is shown in Fig. 2.3.

$$P(y_i | X, y_j, i \neq j) = P(y_i | X, y_j, j \in \mathcal{N}(i)) \quad (2.8)$$

where  $\mathcal{N}(i)$  is the neighbourhood of  $i$  and  $i$  being the index of the vertices of  $\mathcal{G}$  with  $i \in \mathcal{V}$ . This property applies to all  $y_i$ .

The conditional property  $P(Y|X)$  can be characterized by the Gibbs distribution as written in Eq. 2.9:

$$P(Y|X) = \frac{1}{Z} \exp \left( - \sum_{c \in \mathcal{C}_G} \mathcal{E}_c(y_c | X) \right) \quad (2.9)$$

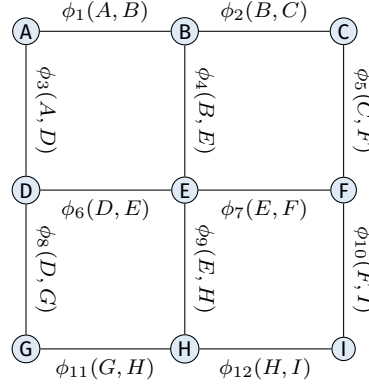


Figure 2.2: Illustration of the graph of a Markov Random Field (MRF) with nine random variables, where the vertices representing the random variables and the edges the dependencies between them. The symbol  $\phi_i$  denotes a factor function, defining the dependency between two nodes.

where  $c$  is a clique in the set of cliques  $C_{\mathcal{G}}$  in graph  $\mathcal{G}$ .  $\mathcal{E}_c$  is the Gibbs energy (see Eq. 2.11), that specifies how local variables interact and how this interaction contributes to the global distribution. All of this is normalized with the normalization factor  $Z$ , as defined in Eq. 2.10 [Tran, 2008, Krähenbühl and Koltun, 2011]:

$$Z = \sum_{y \in Y} \prod_{c \in C_{\mathcal{G}}} \mathcal{E}_c(Y_c | X) \quad (2.10)$$

with  $c$  being a clique in the set of cliques  $C_{\mathcal{G}}$  in graph  $\mathcal{G}$  and the Gibbs energy  $\mathcal{E}_c$  (see Eq. 2.11) [Liu et al., 2016]. The energy function  $\mathcal{E}(Y|X)$  is defined as:

$$\mathcal{E}(Y|X) = \underbrace{\sum_{y \in Y} \psi_u(y|X)}_{\text{unary potential}} + \underbrace{\sum_{i,j \in V} \psi_p(y_i, y_j|X)}_{\text{pairwise potential}} \quad (2.11)$$

where the first term  $\psi_u(y|X)$  is the unary potential, which is usually the output of another machine learning algorithm (in this thesis a neural network). The unary potential measures the cost if the label assignment differs from the initial label for pixels  $i$  and  $j$ , with  $i, j \in V$ . The second term  $\psi_p(y_i, y_j|X)$  is the pairwise potential that measures the cost if similar pixels get different labels assigned [Krähenbühl and Koltun, 2011].

$$\psi_p(y_i, y_j|X) = \mu(y_i, y_j) \sum_{l=1}^{N_{class}} w^{(l)} k_G^{(l)}(f_i, f_j) \quad (2.12)$$

where  $\mu(y_i, y_j)$  is a label compatibility function that assigns a penalty when labels are different and  $k_G$  is a Gaussian kernel applied on the feature vectors. Feature vectors can be e.g., RGB values and/or spatial locations. The compatibility function  $\mu(y_i, y_j)$  can be either a learned function as in Krähenbühl et al. [Krähenbühl and Koltun, 2011] or a simple comparison like  $\mu(y_i, y_j) = [y_i \neq y_j]$  which is similar to the Potts model [Kohli et al., 2007].

The kernel that was applied (same as in [Krähenbühl and Koltun, 2011]) takes the appearance and smoothness into account (see Eq. 2.13).

$$k(f_i, f_j) = w_1 \cdot \underbrace{\exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right)}_{\text{appearance kernel}} + w_2 \cdot \underbrace{\exp\left(-\frac{|p_i - p_j|^2}{2\gamma_\alpha^2}\right)}_{\text{smoothness kernel}} \quad (2.13)$$

where  $p_{i/j}$  are the spatial locations,  $I_{i/j}$  are color vectors and  $W_{i/j}$  are learnable parameters. The variables  $\theta_{\alpha/\gamma}$  are to control the degree of nearness and similarity in the *appearance kernel*, which is motivated by the observation that pixels of similar location and color often belong to the same class [Krähenbühl and Koltun, 2011]. The  $\gamma$  factor in the *smoothness kernel*, however, is motivated by the fact that pixel regions are usually compact. Thus, it removes isolated regions [Krähenbühl and Koltun, 2011].

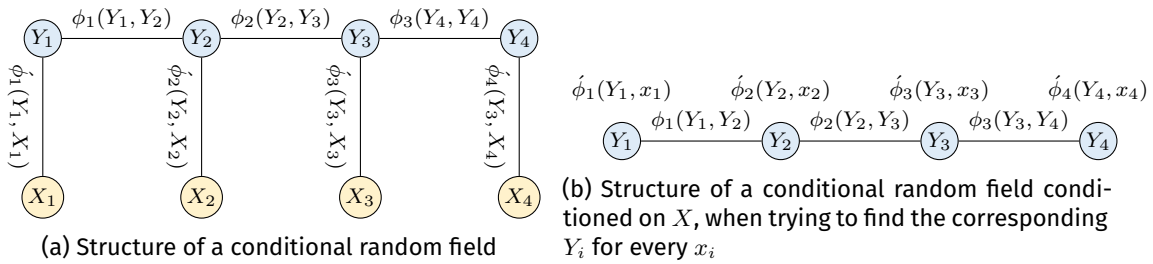


Figure 2.3: Illustration of the graph of a Conditional Random Field (CRF), where (a) shows the chain structure of a CRF with four random variables and (b) shows the chain structure of a CRF, conditioned on  $X$ . The symbol  $\phi_i$  denotes a factor function, defining the dependency between two notes.

Once the CRF was trained by estimating the distribution  $Pr(Y|X)$ , the labels that minimize the energy function  $\mathcal{E}$  can be predicted by using Eq. 2.14 [Lafferty et al., 2001, Tran, 2008]:

$$\hat{y} = \arg \max_Y P(Y|X) \quad (2.14)$$

### 2.1.3 Fundamentals of neural networks

Neural networks are a class of machine learning algorithms that are inspired by the functionality of the brain [Rosenblatt, 1958], meaning it imitates the complex interactions between neurons and their connections [Rosenblatt, 1958]. In neural networks, the unit that emulates those neurons consists of two building blocks: first, a layer (see Sec. 2.1.3) that takes an input signal and performs a mathematical transformation on it, and second a non-linear activation function (see Sec. 2.1.3), applied on the output of the layer [Goodfellow et al., 2016a]. Typically, different kinds of multiple layers are ordered behind each other and are connected in the form of architecture (see Sec. 2.1.4), where a signal goes into the first layer (also called input layer) and through all layers (hidden layers) including the last layer (also called output layer) [Goodfellow et al., 2016a].

In this section, first all variations of layers that are used in this work are described in Sec. 2.1.3 *Layers*, followed by a description of the activation functions in Sec. 2.1.3 *Activation functions*. As neural networks are machine learning algorithms, how a signal is transformed can be learned using statistical methods. How this training is performed and how neural networks are finally applied is described in Sec. 2.1.3 *Training and inferencing*.

### Layers

As described above, the lowest level of building blocks in neural networks is the layer where input signals are transformed with a differentiable mathematical operation. Depending on the intention of use, such operations typically contain learnable parameters adjusted during the training process (see Sec. 2.1.3). As layers are an active field of research and many new layer types are invented every year, this section will only give an overview of the layers used for this thesis, namely the linear, convolutional, pooling, and batchnorm layers.

**Linear layer** The linear layer is the simplest layer form in a neural network and consists of a learnable weight matrix  $W$  and a bias  $b$ . As input, the linear layer receives a 1-Dimensional (1D) input vector  $x$  of length  $n = c_{in}$  and transforms the signal according to Eq. 2.15. By choosing the dimension of the weight matrix  $W$ , the output dimension  $c_{out}$  of the transformed signal can be defined  $[c_{out} \times c_{in}]$ . Like the Principal Component Analysis (PCA) [Hotelling, 1933], the linear layer, combined with the bias, can be used for dimensionality reduction and is thus a powerful layer [Goodfellow et al., 2016a]. However, it can only learn linear relations.

$$f(x) = x \cdot W^T + b \quad (2.15)$$

**Convolutional layer** The intuition behind using the mathematical operation "convolution" is that it can be of particular interest to locate a specific pattern within a signal. Defined is the convolution as the integral of the product of two functions ( $f$  and  $g$ ), with  $g$  being shifted and reversed [Goodfellow et al., 2016b].

As this work deals with two-dimensional images,  $f$  will be an image  $I$ , and  $g$  will be a so-called kernel  $K$ , containing the pattern of interest. In reality, it turned out that the flipped kernel property of the convolution is not of interest for the concrete neural network implementation [Goodfellow et al., 2016a]. Thus, often a *cross-entropy* is implemented, which is the same operation but without flipping the kernel (see Eq. 2.16). [Goodfellow et al., 2016b]. The process is illustrated in Fig. 2.4

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i + m, j + n)K(m, n) \quad (2.16)$$

as written in [Goodfellow et al., 2016b], where  $I$  denotes a two dimensional image of size  $[I_{width} \times I_{height}]$  and  $K$  a two dimensional kernel of the size  $[m \times n]$ . The variables  $i$  and  $j$  are indices to access a specific pixel in image  $I$  with  $i \in \{0, 1, \dots, I_{width}\}$  and  $j \in \{0, 1, \dots, I_{height}\}$  respectively.

**Pooling layer** Pooling layers are usually used to create a representation that is invariant to small translations, which is an interesting property for classification. The presence of a feature is more important than the exact location [Goodfellow et al., 2016a]. In contrast to the convolutional or

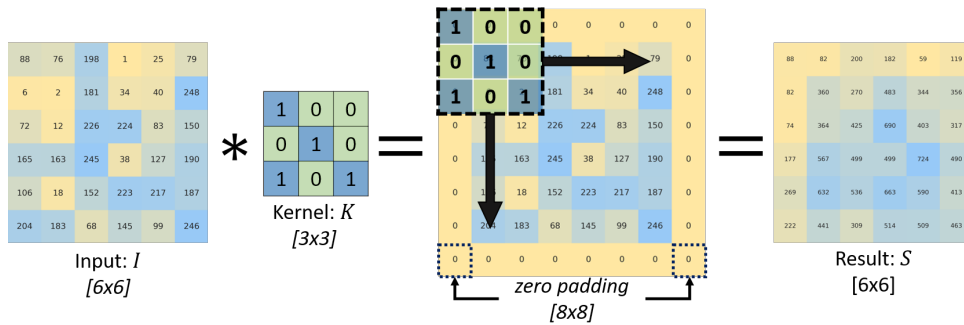


Figure 2.4: Exemplary illustration of a convolution/cross-correlation process for an  $[6 \times 6]$  input image  $I$  and a  $[3 \times 3]$  kernel  $K$ . At first, zero-padding was applied to  $I$ , which changed its dimension to  $[8 \times 8]$ . The kernel was then moved over the image (as described in Eq. 2.21), which produced  $S$ . The operation zero-padding was used so that the dimensions of  $I$  and  $S$  are the same.

linear layer, this layer type does not contain any learnable parameters. Instead, it computes statistics over a small subregion of an image using the maximum value (max pooling) or its average (average pooling).

As pooling can be used for downsampling an image, there is also an inverted version, the max/average unpool operation. All operations are exemplary shown in Fig. 2.5.

**Batchnorm layer** While previous layers (e.g., convolutional layer, pooling layer) are used to transform the input signal into a new latent representation, the batchnorm layer is used to train a neural network more easily [Ioffe and Szegedy, 2015, Goodfellow et al., 2016a]. Modern deep models are especially prone to unexpected effects caused by the way parameters are updated. Ideally, gradients that are computed for each parameter assume that all the other parameters remain constant. However, in practice, all parameters are updated simultaneously, which might lead to an unstable training [Goodfellow et al., 2016a]. For this reason, the batchnorm scales each output of a layer during the training, by standardizing the activations in each mini-batch to have a standard deviation of one and a zero mean (see Eq. 2.17) [Goodfellow et al., 2016a]. During inferencing, running averages will replace  $\mathbf{E}[H]$  and  $\text{Var}[H]$  that were collected during the training time [Ioffe and Szegedy, 2015]. This technique leads to a much more stable training and can lead to better performance of the network [Ioffe and Szegedy, 2015].

$$\hat{H} = \frac{H - \mathbf{E}[H]}{\sqrt{\text{Var}[H] + \epsilon}} \cdot \gamma + \beta \tag{2.17}$$

where  $\epsilon$  is a small constant to avoid a division by zero,  $\mathbf{E}[H]$  and  $\text{Var}[H]$  are calculated per dimension over the mini-batches and  $\gamma$  and  $\beta$  are learnable parameters.

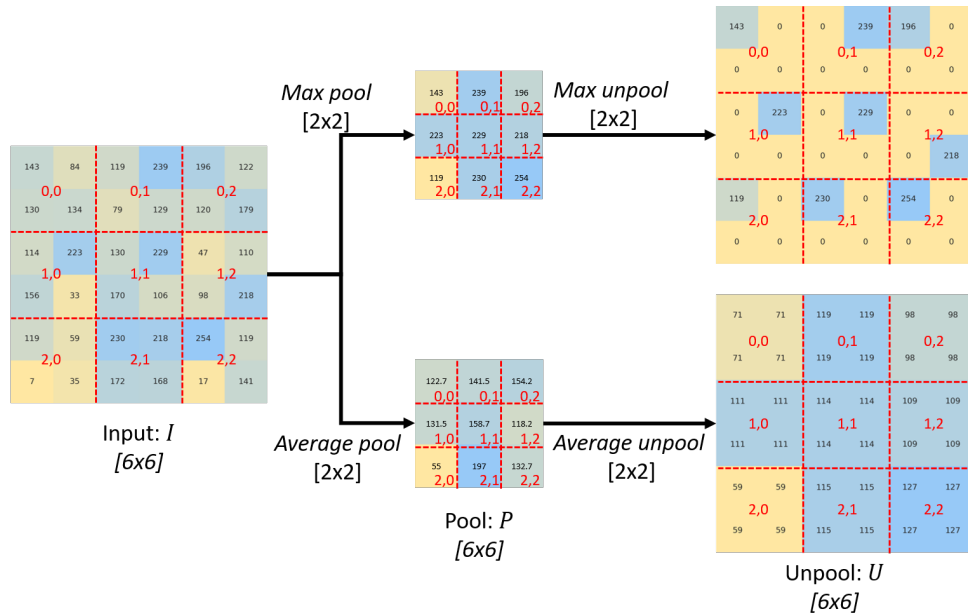


Figure 2.5: An exemplary illustration of the max and average pooling for an input of size  $[6 \times 6]$  and a pooling operation over  $[2 \times 2]$  pixels. First, the image  $I$  is separated into nine fields of the size  $[2 \times 2]$ . Then, for each field, either a max or average pooling is calculated, resulting in  $P$  with shape  $[6 \times 6]$ . After applying the inverse corresponding unpool operation, the original image dimension is restored in  $U$ .

### Activation functions

Activation functions  $\sigma(z)$  are the second main building block in neural networks. Each layer is always followed by a fixed activation function, meaning that after the transformation of a signal  $x$  by the transformation function  $z = g(x)$  of the layer, the signal goes into the activation function  $\sigma(g(x))$  [Goodfellow et al., 2016a].

Originally, the activation function was motivated by the brain's biological functionality, where it models whether a neuron fires or not [Goodfellow et al., 2016a]. Currently, neural networks use non-linear activation functions that enable the possibility to manage complex data by increasing the dimensionality of the model [Goodfellow et al., 2016a]. Not using activation functions would cause a collapse into a linear system [Goodfellow et al., 2016a]. While it may be desirable for smaller problems to spare parameters, especially complex problems, require a big model capacity [Goodfellow et al., 2016a]. Besides this, activation functions help to normalize a neuron's output into a specific range, depending on the used activation function and the use-case.

For instance, classification tasks usually use a *softmax* activation on the last layer to generate pseudo-probabilities in a range between  $[0, 1]$ . If the output has to be bounded, the *sigmoid* or *tanh* activation can be applied, that produces values in a range of  $[0, 1]$  or  $[-1, 1]$  respectively. The (*leaky*) *ReLU* activation is mostly used in input or hidden layers as they are less likely to generate vanish gradients and seem to produce better results in practice [Krizhevsky et al., 2012].



The range of *ReLU* is  $[0, \infty]$ . As values of zero could lead to a total deactivation of certain parts in a model, *leaky ReLU* has an additional factor  $a$  that prevents the effect by also allowing negative values, resulting in a value range of  $[-\infty, \infty]$ .

As there are many different kinds of activation functions available [Goodfellow et al., 2016a], Fig. 2.6 only shows the activation functions used in this work.

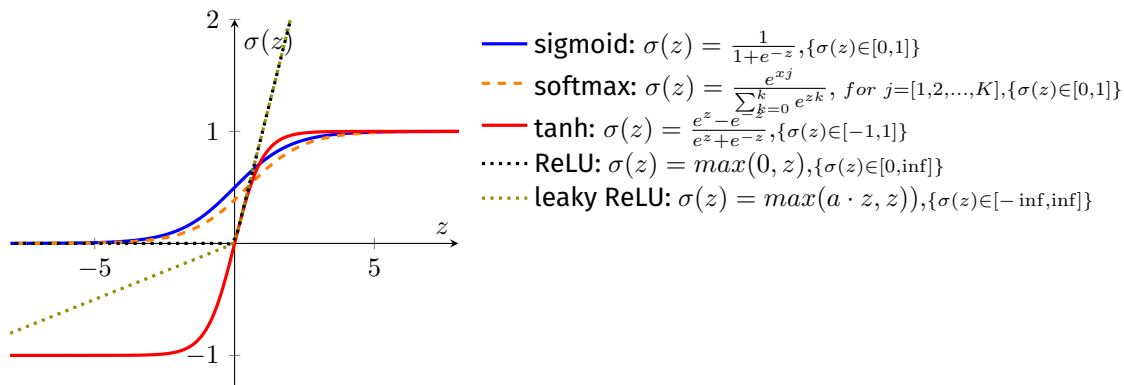


Figure 2.6: Activation functions used in this thesis with their formula and value range.

## Losses

Like all machine learning algorithms, neural networks must be trained on data before being applied to an actual problem. For this training process (as described in Sec. 2.1.3) there is a need to quantify how well a network models the data by means of an objective function - also called loss or cost function [Goodfellow et al., 2016a].

A loss function typically measures a model's success in terms of how "wrong" a network is in its predictions by computing a kind of difference between a predicted value  $\hat{y}$  and the true value  $y$ . Thus, the lower the loss, the "better" the network describes the data. Therefore, it requires that the value of the loss function is a good indicator for success concerning the given task and that a decreasing loss leads to a better model [Reed and MarksII, 1999, Goodfellow et al., 2016a].

As mentioned above, the chosen loss function must be appropriate to quantify the quality. Unfortunately, no loss function works for every kind of problem. Various factors play a role in choosing the correct one (such as outliers in the data, calculation effort, etc.). Principally, two major groups of losses exist, namely regression and classification losses [Goodfellow et al., 2016a]. Regression is the task of predicting continuous values (e.g., how long a specific surgical procedure lasts) while classification is the task of predicting from a set of finite categorical values (e.g., the visible instrument in the current video frame is a scissor).

Because losses are so problem-specific, one cannot give a complete list of all losses used in the literature. Thus, the following part presents only a selection of losses used in this thesis.

**Regression losses** In this part, the two most common regression losses are described: the  $\mathcal{L}_1$  also called *Mean Absolute Error (MAE)* and the  $\mathcal{L}_2$  loss, also called *Mean Squared Error (MSE)*. Both are used to minimize the error between predictions  $\hat{y}$  and the true value  $y$  for a regression task.

The  $\mathcal{L}_1$  loss is defined as the average of the sum of absolute differences between  $\hat{y}$  and  $y$  and measures the magnitude of error without its direction, with  $N$  being the number of data samples and  $i \in [1, 2, \dots, N]$  (see Eq. 3.2).

$$\mathcal{L}_1 / \mathcal{L}_{MAE}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i) \quad (2.18)$$

Also the  $\mathcal{L}_2$  ignores the direction but uses the squared differences between  $\hat{y}$  and  $y$  with  $N$  being the number of data samples and  $i \in [1, 2, \dots, N]$  (see Eq. 3.3).

$$\mathcal{L}_2 / \mathcal{L}_{MSE}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (2.19)$$

The main differences between  $\mathcal{L}_1$  and  $\mathcal{L}_2$  are that the  $\mathcal{L}_1$  is more robust against outliers than the  $\mathcal{L}_2$  with its squared error. However, the  $\mathcal{L}_2$  leads to more stable training than the  $\mathcal{L}_1$  and has only one solution, while  $\mathcal{L}_1$  has more than one (see Fig. 2.7).

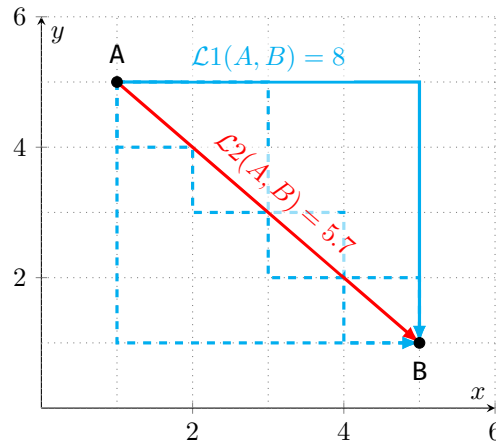


Figure 2.7: This figure illustrates the  $\mathcal{L}_1$  and  $\mathcal{L}_2$  distance between two data points,  $A$  and  $B$ . While the solution for the  $\mathcal{L}_2$  norm (red) is unique,  $\mathcal{L}_1$  (blue) has multiple possible solutions (dashed arrows).

**Classification losses** This part describes two classification losses, the *Cross Entropy (CE)* and the *Sørensen Dice Similarity Coefficient (DSC)* loss [Sudre et al., 2017]. In contrast to regression losses where continuous values are compared, classification losses encompass a finite set of categorical values. In computer vision tasks, the comparison is mainly performed by predicting a probability of each class by applying a *softmax* activation on the output (see Sec. 2.1.3). To compare the

predicted probabilities with the reference, the reference is converted into a one-hot encoded vector  $\delta$  (see Eq. 2.20).

$$\delta_{i=c} \equiv \begin{cases} 1 & \text{if } i = c \\ 0 & \text{else} \end{cases} \quad (2.20)$$

with  $\delta$  is of length  $N_c$ ,  $N_c$  is the number of classes,  $c$  is the reference class and  $i \in [0, 1, \dots, N_c]$ .

After the conversion into a one-hot encoded vector,  $\delta_y$  and  $\hat{y}$  can be interpreted as two distributions. Comparison can now be done by the *cross-entropy*, which measures the total entropy between two distributions (see Eq. 2.21). The closer both distributions are, the smaller the loss will be.

$$\mathcal{L}_{CE}(y, \hat{y}) = - \sum_{j \in c} \delta_{y_j} \log \hat{y}_j \quad (2.21)$$

One very important task in the computer vision community is the segmentation task, where segmentation means that every single pixel in an image is assigned to a concept (e.g., foreground vs background). In other words, the segmentation can be seen as a pixel wise classification. A loss that is often used for training segmentation models is the *DSC loss* [Drozdal et al., 2016], which is defined as the harmonic mean of precision and recall. Usually, the DSC is used to compare the overlap between two segmented regions and ranges from 0 to 1, where 1 denotes a complete overlap (see Eq. 2.22). As during the training process of neural networks, the error is minimized, and the *DSC loss* needs a small modification in its formula (see Eq. 2.23).

$$DSC(Y, \hat{Y}) := \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \quad (2.22)$$

$$\mathcal{L}_{DSC}(y, \hat{y}) = 1 - \frac{2 \cdot \sum_{pixel} (y \odot \hat{y})}{\sum_{pixel} y + \sum_{pixel} \hat{y} + \epsilon} \quad (2.23)$$

where  $Y$  denotes the reference annotation and  $\hat{Y}$  the corresponding prediction of an image frame,  $\odot$  denotes an element-wise multiplication and  $\epsilon$  a small constant to avoid a division by zero. There exists different versions of the *DSC loss*, e.g., by using the squared sum for both terms in the denominator [Milletari et al., 2016].

### Training and inferencing

The heart of all machine learning algorithms is the *training* and *inferencing* steps. During the training process the algorithm learns, based on the training data, a parameter setup that models the data distribution by continuously updating all learnable parameters of the model. After the training is done, the model can be used and applied to new data, the inferencing step.

The basis of the complete training procedure is formed by the loss function  $\mathcal{L}$  (see Sec. 2.1.3) and the principal goal is to perform optimization by updating the learnable parameters  $\theta$  in such a way that the loss decreases [LeCun et al., 2015]. Those parameters are updated with the help of the loss gradients. The calculation of those gradients is called backpropagation.

One of the first and most used algorithms to do the optimization of the parameter  $\theta$  of a neural network is the Gradient Descent (GD), which tries to find a minimum for a given cost/loss function  $\mathcal{L}$  [Kiefer et al., 1952]. It is based on the gradient descent optimization algorithm, which does the parameter updates by (1) processing all training data  $x$  with the model, then (2) calculating the average of the error  $\mathbb{E}[\mathcal{L}(\theta)]$  and finally, the resulting gradients are used in combination with a learning rate  $\alpha$  to update the model parameters (see Eq. 2.24).

$$\theta = \theta - \alpha \nabla_{\theta} \mathbb{E}[\mathcal{L}(\theta)] \quad (2.24)$$

However, this process is very time-intensive as it requires the processing of all available training images.

In comparison to the gradient descent, the Stochastic Gradient Descent (SGD) [Kiefer et al., 1952] no longer relies on the expectation as it approximates by computing the gradients of the parameters by using only a few samples of the training images, the so-called mini-batches (see Eq. 2.25) [Kiefer et al., 1952].

$$\theta = \theta - \alpha \nabla_{\theta} \mathcal{L}(\theta; x^{(i)}, y^{(i)}) \quad (2.25)$$

with  $x^{(i)}$  and  $y^{(i)}$  being the  $i$ -th mini-batch from the training set. The learning rate  $\alpha$  is the same for all learnable parameters. At this time it has become the standard optimizer for neural networks.

Additionally, the learning of the SGD might be slow, and a good learning rate of  $\alpha$  might be challenging to achieve. Thus, the Adam optimizer [Kingma and Ba, 2014] became increasingly popular in the last years and is currently suggested as the default optimization [Karpathy et al., 2016]. It allows individual adaptive learning rates, where the learning rate is adapted by how fast the loss decreases [Karpathy et al., 2016]. Despite the fact that the adaptive learning rate leads to faster training, it comes at the cost of a non-optimal convergence and worse generalization [Keskar and Socher, 2017, Wilson et al., 2017].

#### 2.1.4 Neural network architectures used in this thesis

After the description of the building blocks of neural networks (see Sec. 2.1.3), this part explains all neural network architectures that were used in this work. Two discriminative models (U-Net and a Residual Neural network), one generative model (GAN) and a Recurrent neural network.

##### U-Net

The U-Net is a deep learning architecture designed for image segmentation, where its architecture aims to combine local and global context for improving the segmentation quality [Ronneberger et al., 2015]. As the name suggests, the U-net has a "U" like form (see Fig. 2.8) and is oriented on the Autoencoder [Ballard, 1987] architecture. The four main building blocks are a down-sampling path, a bottleneck, an upsampling path, and skip connections [Ronneberger et al., 2015].

Like the Autoencoder, the U-Net first reduces stepwise the spatial information in the downsampling path and compresses it into a smaller representation of the image (bottleneck). Afterwards, it reconstructs stepwise the original image dimension in the upsampling path. In contrast to

the Autoencoder, the U-Net uses skip connections between the same spatial resolutions [Ronneberger et al., 2015]. The U-net idea is that decreasing the spatial resolution enables the network to encode neighboring pixels into global information, while the skip connections combine global with local information [Ronneberger et al., 2015]. Each level in the downsampling path has three consecutive blocks, where each block is first a convolutional layer, followed by a batchnorm and a leaky ReLU activation. At the end of each level comes a MaxPooling operation that reduces the spatial dimension. The upscaling path starts with an unpool operation, followed by the skip connection. The skip connection concatenates the output of the level before with the output of the same level hierarchy. Ultimately, again three times, the blocks are processing the signal.

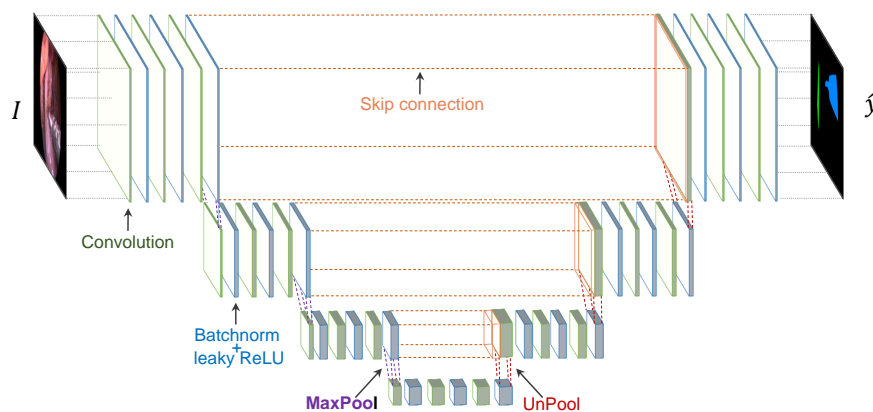


Figure 2.8: This figure illustrates the U-Net architecture, where the left part is called the downsampling path and the right part the upsampling part. Each level in the downsampling path has three consecutive blocks, where each block is first a convolutional layer, followed by a batchnorm and a leaky ReLU activation. At the end of each level comes a MaxPooling operation that reduces the spatial dimension. The upscaling path starts with an unpool operation, followed by the skip connection. The skip connection concatenates the output of the level before with the output of the same level hierarchy. Ultimately, three more times, upsampling blocks are processing the signal.

### Residual neural networks

The group of residual neural networks is a group of networks that is used more often for classification rather than segmentation; however, it can be used for both use cases [Wu et al., 2019]. The main motivation for developing residual neural networks was that deep learning models cannot arbitrarily consist of many layers, as it can lead to a problem called vanishing gradients [He et al., 2016a]. Vanishing gradients are a phenomenon that occurs during the backpropagation process, where the "error signal" decreases (or increases) exponentially as a function of the distance from the "final layer" [Sussillo and Abbott, 2014]. Those unclear gradients can lead to unstable training and thus, result in a bad overall performance [He et al., 2016a].

To avoid the vanishing gradients problem, the Residual Neural Network (ResNet) is a biologically inspired form of architecture that incorporates a new kind of building block in the form of shortcuts between layers (see Fig. 2.9), so-called skip connections [He et al., 2016a]. Like the

pyramid cells in the cerebral cortex, this enables a signal to skip layers and thus reduces the vanishing gradients' effect. Those skip-connections enabled really deep architectures, starting from a ResNet18, up to currently a ResNet152 [Tan and Le, 2019].

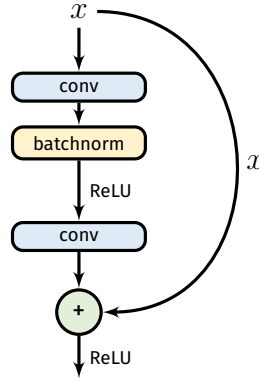


Figure 2.9: Exemplary illustration of a residual block with two convolutional layers.

### Generative adversarial networks

In 2014, Goodfellow et al. introduced a new generative machine learning framework, the so-called Generative Adversarial Network (GAN) [Goodfellow et al., 2014]. Despite discriminative models, GANs can be used to generate new data from a learned distribution (see Sec. 2.1.1). Usually, GANs are made up of two competing models, a generator  $G$  and a discriminator  $D$  that are trained simultaneously. Both  $G$  and  $D$  contest each other in the form of a zero-sum game, meaning  $G$  generating new data and  $D$  deciding whether the presented data is from the true training data distribution or  $G$  generated data distribution [Goodfellow et al., 2014]. As  $D$  improves and learns to differentiate real from generated data,  $G$  aims to fool  $D$  and thus has to adapt its generation [Goodfellow et al., 2014]. As this thesis deals with images, the GAN will be explained on image generation. An overview of the framework is shown in Fig. 2.10. In the following, the discriminator and the generator will be further explained.

**Discriminator  $D$**  The discriminator is a model that is trained to differentiate between real and fake images. As this is a classification task, where the discriminator should provide the probability whether the input belongs to the true distribution, the authors used a *sigmoid* activation, such that  $D(I) \in [0, 1]$ , where 1 stands for the  $Y_{real}$  and 0 for the  $Y_{synthetic}$  class. Additionally, instead of using a classification loss such as *cross-entropy*, typically, a regression loss is used, as shown in the Eq. 2.26.

$$\mathcal{L}_D(I, \hat{I}) = [D(I) - Y_{real}]^2 + [D(\hat{I}) - Y_{synthetic}]^2 \quad (2.26)$$

**Generator  $G$**  The generator  $G$  is a neural network that acquires as input a random vector  $r$  and produces an output (in this case, a new image  $\hat{I}$ ). The last layer has a *tanh* activation function, thus  $G(z) \in [-1, 1]$ , with a random input  $z$ , where  $z \sim \mathcal{N}(0, 1)$ . While the discriminator has a concrete error loss,  $G$  becomes optimized to maximize the error of  $D$  (see Eq. 2.27).

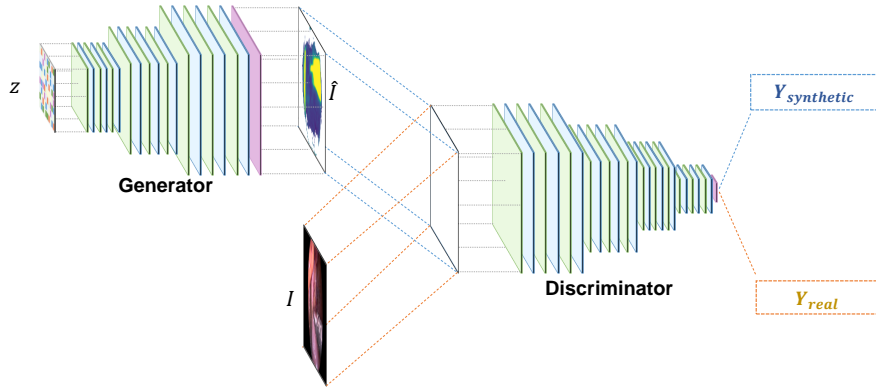


Figure 2.10: Architecture of the Generative adversarial network (GAN).

$$\mathcal{L}_G(\hat{I}) = [D(\hat{I}) - Y_{real}]^2 \quad (2.27)$$

**Training of a GAN** As already written above, the generator  $G$  and the discriminator  $D$  are jointly trained in the form of a zero-sum game with the value function  $V(G, D)$ , where the log probability of the discriminator to correctly classify input data as either real or synthetic is maximized, and the probability of the generator to produce a synthetic image that can be correctly classified by  $D$  is minimized.

$$\min_G \max_D V(G, D) = \mathbf{E}_{I \sim p_T(I)} [\log D(I)] + \mathbf{E}_{z \sim p_z} [\mathcal{L}_G(G(z))] \quad (2.28)$$

where  $p_z$  is a prior and  $I \sim p_T$  a real image drawn from the training set and  $V$  the joint model of  $G$  and  $D$ . The training itself should typically converge into a Nash equilibrium [Kreps, 1989], such that neither the generator, nor the discriminator, are changing any more, given the action of the others [Goodfellow et al., 2014].

The training of GANs comes with several difficulties: (1) slow learning, (2) a *mode collapse*, (3) problem to converge, and (4) difficulties in balancing and controlling the learning process [Arjovsky and Bottou, 2017, Goodfellow et al., 2016a].

### Recurrent neural networks

Recurrent neural networks are a new class of neural networks that allow the use of previous outputs to be used as input [Bengio et al., 1994]. In contrast to all the network architectures that were described before (e.g., U-Net, GAN) whose inputs are fixed, RNNs enable the processing and prediction of sequences by incorporating contextual information in the form of hidden states [Bengio et al., 1994]. The differences between a traditional neural network and an Recurrent Neural Network (RNN) are presented in Fig. 2.11. Typical applications of RNNs are speech recognition [Graves et al., 2013], stock price prediction [Selvin et al., 2017] or text translation [Cho et al., 2014]. RNNs are usually trained with a loss function that sums up the error at every time step (see Eq. 2.29).

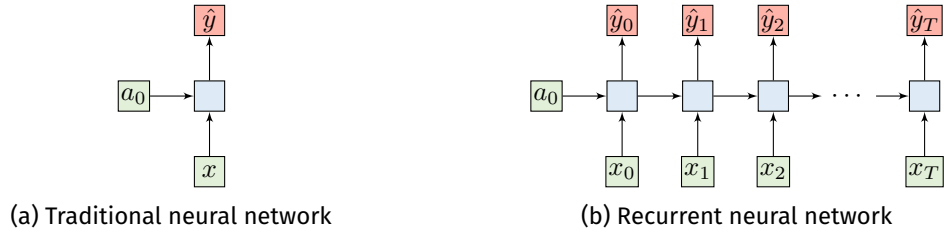


Figure 2.11: This figure illustrates the differences between a traditional neural network (a) and an example of a recurrent neural network (RNN) (b). It can be seen that traditional neural networks take one fixed input  $x$  to produce a prediction  $y$ , while the RNN processes a sequence of inputs  $\{x_0, x_1, \dots, x_T\}$ . It does this by incorporating the contextual information in the form of hidden states to produce the predictions  $\{\hat{y}_0, \hat{y}_1, \dots, \hat{y}_T\}$ , where  $T$  is the sequence length.

$$\mathcal{L}(y, \hat{y}) = \sum_{t=0}^{T_y} \mathcal{L}(y_t, \hat{y}_t) \quad (2.29)$$

where  $T_y$  is the number of time steps for the sequence of targets  $y$ .

Despite the many advantages of RNNs that enables the possibility to process the input of any length, keeping the model size static even with increasing input length, and taking historical information into account, in praxis, it turned out that RNNs suffer from a couple of drawbacks. These drawbacks are difficulties in accessing information in long sequences and in the ability to train RNNs due to vanishing and exploding gradients (explanation see Sec. 2.1.4 *Residual neural networks*) [Bengio et al., 1994, Sak et al., 2014].

Motivated by the observation that the error flow in RNNs can lead to vanishing or exploding gradients and thus hinder the access to information in long sequences [Graves and Schmidhuber, 2005], Hochreiter et al. developed the Long short-term Memory Network (LSTM) architecture [Hochreiter and Schmidhuber, 1997].

The core idea behind the LSTM is a cell state  $c$  where information is stored and thus theoretically accessible at any time during the sequence [Hochreiter and Schmidhuber, 1997]. The decision as to should be stored is regulated by a gate mechanism, which is a sigmoid output of a network layer and a pointwise multiplication. As sigmoid contains values in the range of  $[0, 1]$ , it describes how much information should be remembered or forgotten [Hochreiter and Schmidhuber, 1997]. After learning how much of the state should be forgotten, new information is added to the state by pushing the output through a  $\tanh$  activation function and multiplying it by the output of the sigmoid [Hochreiter and Schmidhuber, 1997]. Finally, the output of the LSTM is based on the new cell state, filtered by the  $\tanh$  and again gated with the sigmoid activation [Hochreiter and Schmidhuber, 1997]. The complete LSTM layer is illustrated in Fig. 2.12,



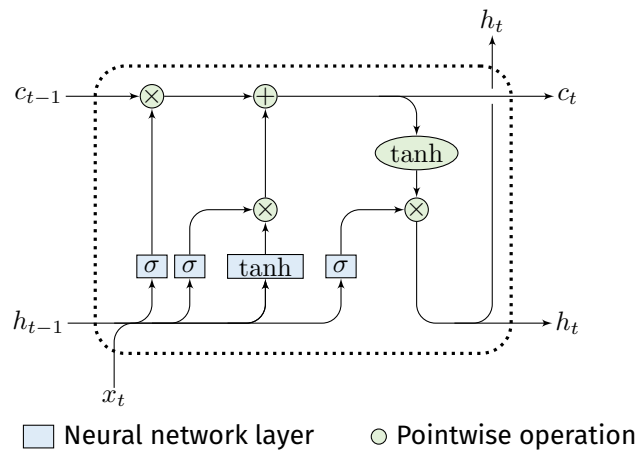


Figure 2.12: Illustration of the building block of a Long short-term memory network (LSTM).

### Mask R-CNN

While approaches like the U-Net can only be used to assign each pixel to a concept, the Mask R-CNN [He et al., 2017] is motivated by the observation that the number of visible classes and objects within an image can vary, making it challenging to train multi-instance segmentation/-classification. Unlike the U-Net, the Mask R-CNN should be able to distinguish different instances from the same class. For this reason, the authors decided to combine the functionality of a recurrent neural network with a segmentation/classification network [Ren et al., 2015, He et al., 2017]. A Mask R-CNN's principal architecture (see Fig. 2.13) is divided into two steps, both of which are based on the backbone. In the following paragraphs, the backbone is first explained, followed by a more detailed explanation of the two steps and ending with how a Mask R-CNN is trained.

**Backbone** The backbone of a Mask R-CNN is usually a pre-trained network prepared to detect objects in images on huge datasets such as ImageNet [Deng et al., 2009] or MS-COCO [Lin et al., 2014a]. Examples of such backbones are ResNets [He et al., 2016a], the VGG network [Simonyan and Zisserman, 2014] or a network that combines both architecture and a Feature Pyramid Network [Lin et al., 2017]. As the pre-trained networks have already been trained to detect and to differentiate between multiple classes, the assumption is that they already learned features that are helpful to distinguish between objects. Using multiple layers of such low-level features (features from layers in the very beginning) can help to transform an input image into a more meaningful latent representation [He et al., 2017].

**Step 1: Region proposals** After transforming the input image into the latent representation, the Mask R-CNN produces so-called proposals. Proposals are regions in the image of a particular interest as they might contain objects that should be segmented. Those proposals are generated by a Region Proposal Network (RPN), which, in most cases, is a lightweight network which consists of only a few layers to propose regions (Region of Interest (ROI)) in the form of anchor boxes [He et al., 2017]. The size, aspect ratio, and scaling of such anchor boxes are predefined parameters. During the training and inferencing, the RPN proposes thousands of boxes with

different combinations of the predefined size, aspect ratio, and scaling [He et al., 2017].

**Step 2: Instance segmentation/classification** The second step contains different headers of small specialized branches. These branches consist of lightweight CNNs that predict (1) the bounding box coordinates, (2) the class of the object inside the bounding box, and finally (3) the segmentation of the object inside the bounding box [He et al., 2017]. In addition to the bounding box, the architecture further predicts a score which is the estimated Intersection over Union (see Eq. 2.30 of the predicted bounding box with the real bounding box of an object [He et al., 2017].

$$IoU(Y, \hat{Y}) := \frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|} \quad (2.30)$$

where  $Y$  is the reference and  $\hat{Y}$  is the prediction. By thresholding this score, one can differentiate between good/bad proposals and use only objects with a high score for the final prediction [He et al., 2017].

**Training** Other than the backbone which is usually not a vanilla network and thus does not need to be retrained, all other network parts are jointly trained. The multi-task loss function is shown in Eq. 2.31.

$$\mathcal{L}_{\text{MRCNN}}(Y, \hat{Y}) = \mathcal{L}_{\text{cls}}(Y, \hat{Y}) + \mathcal{L}_{\text{box}}(Y, \hat{Y}) + \mathcal{L}_{\text{mask}}(Y, \hat{Y}) \quad (2.31)$$

where  $\mathcal{L}_{\text{cls}}(Y, \hat{Y})$  and  $\mathcal{L}_{\text{mask}}(Y, \hat{Y})$  are both classification cross entropy losses, where  $\mathcal{L}_{\text{cls}}(Y, \hat{Y})$  is the loss for the correct classification of a bounding box and  $\mathcal{L}_{\text{mask}}(Y, \hat{Y})$  the loss for the correct binary segmentation of the object inside of the bounding box. This was in order to learn correct coordinates of the bounding box, the  $\mathcal{L}_{\text{box}}(Y, \hat{Y})$ , which is a regression loss (either  $L1$  or  $L2$  loss) [He et al., 2017].

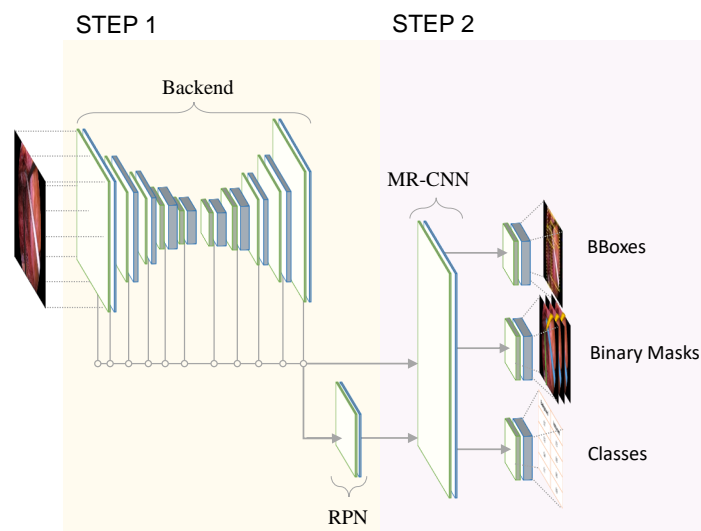


Figure 2.13: This figure shows an illustration of the Mask R-CNN architecture. Input is an image which will first be processed by the pre-trained backend to extract low-level features. Those features are then the input for a region proposal network (RPN) that produces regions of interest (ROI). Those ROIs will then be further processed in the Mask R-CNN, which predicts bounding boxes (BBoxes), binary masks, and each box's class.

## 2.2 | Related work

This section is intended to provide an overview of the relevant work and current state-of-the-art for this thesis in accordance with objectives stated in the introduction. As this section it is supposed to be an overview, not all aspects of the presented methodologies and papers can be discussed in detail.

In this section, recent work in the field of self-supervised learning in surgical data science is first presented in Sec. 2.2.1, followed by the most relevant datasets in the surgical data science in Sec. 2.2.2. In Sec. 2.2.3 the most recent publications in the field of multi-instance segmentation are presented. The last part (Sec. 2.2.4) reviews how the objectives  $O_1 - O_5$  try to fill such "white spots" on the "research map".

### 2.2.1 Self-supervised learning in surgical data science

Sparsity of labeled data is a problem that affects both communities, the computer vision community, that deals with natural images and the SDS community [Zhang et al., 2016, Baur et al., 2017, Chapaliuk and Zaychenko, 2020]. In the medical domain, it is even considered one of the major bottlenecks for developing and translating deep learning-based models into the clinical routine [Shen et al., 2017a, Miotto et al., 2018, Chapaliuk and Zaychenko, 2020]. When dealing with small datasets, most of the models are semi-supervised learning approaches [Baur et al., 2017, Chapaliuk and Zaychenko, 2020] that aim to learn a representation of a domain while dispensing manually generated annotations [Zhang et al., 2016, Baur et al., 2017]. The idea of a semi-supervised approach is, in principle, pre-training on a huge labeled dataset and a later fine-tuning on a small set of training data from the target task [Ravishankar et al., 2016, Zhou et al., 2017, Tajbakhsh et al., 2017]. An alternative technique that does not require a vast labeled dataset is *self-supervised learning*. Instead of a massive amount of labels, raw data is used in the auxiliary task as its own source of supervision [Agrawal et al., 2015a, Zhang et al., 2016, Zhang et al., 2017a].

When the results of this thesis were developed and published in Ross et al. [Ross et al., 2018], the related literature on pre-training with self-supervised learning was mainly presented by the computer vision community. One of the first deep learning self-supervised learning papers was from Agrawal et al. [Agrawal et al., 2015a], where the authors trained a neural network to predict the camera transformation between pairs of images (egomotion). They could show that features learned from the egomotion training are useful when fine-tuned on tasks that require a visual understanding of the world, such as object classification [Agrawal et al., 2015a]. The authors recommend that future research focus on identifying auxiliary tasks that could be used for self-supervision.

Later publications followed the recommendation and proposed a set of auxiliary tasks: classification [Zhang et al., 2016, Noroozi and Favaro, 2016], re-ordering [Noroozi and Favaro, 2016], prediction [Agrawal et al., 2015b]), in-painting [Pathak et al., 2016] and re-colorization [Zhang

et al., 2016, Zhang et al., 2017b, Larsson et al., 2017]. The first self-supervised learning paper in the medical domain was from Zhan et al. [Zhang et al., 2017a]. The authors worked on CT and MR images, where they selected from a random slice random patches and placed them into another slice. After that, they trained a neural network to find the wrong patches and to predict the slice number where they are taken from [Zhang et al., 2017a]. With this procedure, the authors could learn spatial and contextual features that increased the organ segmentation and disease classification.

While previous literature was outside the SDS domain, one publication in the SDS domain was simultaneously developed to work presented in this thesis by Bodenstedt et al. [Bodenstedt et al., 2017]. The authors introduced a new auxiliary task trained in a self-supervised manner for a workflow recognition task. In their method, the authors first extracted image pairs from a video and trained a CNN to determine their temporal order [Bodenstedt et al., 2017]. With this method, the authors could show that the learned features improve the performance of a workflow recognition task.

In the meantime, there are new publications that are partially based on the results of this thesis that were published in [Ross et al., 2018]. Some of the most relevant new developments are summarized in recent reviews, e.g., Chapaliuk et al. [Chapaliuk and Zaychenko, 2020], Jing et al. [Jing and Tian, 2020], and Cheplygina et al. [Cheplygina et al., 2019]. The main contributions of new approaches are the identification of further possible auxiliary tasks (e.g., multi-model reconstruction tasks [Hervella et al., 2020] or predicting the optical flow [Zhao et al., 2020]). Tajbakhsh et al. presents an approach where the authors use a combination of auxiliary tasks, namely rotation, reconstruction, and colorization [Tajbakhsh et al., 2019] and Bowles et al. [Bowles et al., 2018] use self-supervision to learn possible data augmentation techniques to enhance the availability of their training data. A completely new approach from Liu et al. [Liu et al., 2020] is inspired by the idea of self-supervision (including no manual labeled data), in which the authors present a setting where they automatically generate pseudo labels for instrument segmentation. Those pseudo labels are based on anchors generated by an object detector and the location of proposed objects. Those labels are then used to train a neural network to segment instruments in images automatically.

### 2.2.2 Available datasets for surgical data science

As written before, the number of available datasets in the surgical data science community generally is limited [Esteva et al., 2021, Maier-Hein et al., 2020a]. Beside datasets that provide data from laparoscopic surgeries that focus on SDS tasks, e.g., EndoVis - Kidney [Hattab et al., 2020], there are a couple of datasets address surgical instruments. Two of them is the datasets EndoVis - Workflow and Skill [EndoVis, 2019] and Cholec80 [Twinanda et al., 2016]. These two datasets provide images from laparoscopic colorectal surgery and cholecystectomy, respectively. However, both datasets only provide the type of surgical instrument (e.g., scissor, grasper) and do not focus on the segmentation. The same holds true for the m2cai6-tool dataset [Stauder et al., 2016] and CATARACTS [EndoVis, 2018], where CATARACTS provides data from a microscopic surgery. Also, the dataset from [Sznitman et al., 2012] contains data from a microscopic surgery (retina microsurgery) and provides the instrument center and the scale. One dataset that gets closer in the direction of the instrument segmentation is ATLAS Dione [Sarıkaya et al., 2017], which provides instrument bounding boxes. However, ATLAS Dione was recorded in a phantom,

which limits the representativeness of the data.

Besides the mentioned datasets dealing with a related instrument task, a few datasets focus directly on the segmentation. One of them is the EndoVis challenge instrument segmentation dataset [Endovis, 2015], which comprises 310 annotated images from a colorectal surgery with binary instrument segmentations. However, for state-of-the-art machine learning algorithms, this amount of data is insufficient to train a segmentation model that does generalize well (see Fig. 1.1). As this amount of data was not big enough, Allan et al. provided the EndoVis - Rob Instrument dataset [Allan et al., 2019], with 3,000 images. Even though the amount of data is now much higher, The authors have mainly chosen images from a robotic nephrectomy that have no problems, such as smoke, reflections, or excessive blood. Therefore, its representatives for real surgeries is limited. The same holds for the EndoVis Rob Seg dataset [EndoVis, 2018]. However, one dataset that got very recently published is the CATARACTS-SemSeg[Endovis, 2020] dataset, but also this dataset does only provide binary segmentations.

Besides missing datasets for multi-instance segmentation, none of the previously mentioned datasets had a clearly defined annotation protocol and a quality controlled annotation process. Further, most of the datasets were only labeled by only one and in some cases by two annotators. Thus, the data might contain inconsistencies and is biased towards the small number of annotators' preferences.

### 2.2.3 Multi-instance medical instrument segmentation

While the task of binary instrument segmentation has received much attention over the last couple of years (e.g., [Allan et al., 2015, García-Peraza-Herrera et al., 2016, Garcia-Peraza-Herrera et al., 2017a, Shvets et al., 2018, Lee et al., 2019, Jin et al., 2019]), there was no literature about multi-instance segmentation up to the point where parts of this thesis had been published in [Roß et al., 2020]. This lack of literature can be attributed to the fact that open data was only available for binary segmentation, as shown in the dataset table, presented in Maier-Hein et al. [Maier-Hein et al., 2020a]. For this reason, the first part of this section will present the state-of-the-art of binary instrument segmentation, and the second part the recent multi-instance segmentation. Because of the strong relatedness between the robotic instrument segmentation and the laparoscopic instrument segmentation, the presented methods will be a mix of both tasks.

**Binary instrument segmentation** The task of binary instrument segmentation has a long history, where first approaches were based on handcrafted features such as color and texture features [Speidel et al., 2006, Zhou and Payandeh, 2014]. With the rise of machine learning, new approaches such as Gaussian Mixture Models [Pezementi et al., 2009] and Random Forests [Bouget et al., 2015, Maier-Hein et al., 2016]) were commonly used. However, a real breakthrough in the performance was achieved by the first deep learning models, such as [Garcia-Peraza-Herrera et al., 2017a], [Shvets et al., 2018], and [Isensee and Maier-Hein, 2020], that are slightly changed variants of the popular U-Net from Ronneberger et al. [Ronneberger et al., 2015]. Other approaches attempt to use very deep networks, such as ResNets [Pakhomov et al., 2019] or recurrent neural networks [Zhao et al., 2018]. In principal can the development of instrument segmentation models be divided into two main branches: The first branch contains models which attempt to maximize segmentation metrics to win challenges (e.g., [Isensee and Maier-Hein,

2020]), but often perform slow as result [Shvets et al., 2018]. The second branch often achieves lower performance results, but these smaller, and often specially designed models posses the benefit of being able to used in real-time applications [García-Peraza-Herrera et al., 2016, Islam et al., 2019].

**Multi-instance segmentation** Based on the review paper from from Hafiz et al. [Hafiz and Bhat, 2020] that provides an extensive overview overview of the current state-of-the-art for multi-instance segmentation, almost all current approaches are based on the method detection followed by segmentation, which means that first a detector network provides proposals of possible object locations, followed by a segmentation network that further processes such proposals [Hafiz and Bhat, 2020].

Overall, it can be said that the entire topic is currently the subject of intense research, where almost every two months a new method is reported that outperforms top scores on the COCO benchmarking<sup>1</sup> by tiny margins (Average Precision [Zhang and Zhang, 2009] 0.491 vs. 0.485). The latest best performing methods and currently most used approaches are the Mask R-CNN [He et al., 2017], the Mask scoring R-CNN [Huang et al., 2019b], centermask [Lee and Park, 2020] and Cascade R-CNN [Fang et al., 2019]. Although there are constantly new approaches, the R-CNN mask is still the most widely used network for multi-instance segmentation [Hafiz and Bhat, 2020].

Up to this point, there are three very recent publications, all based on the results that were published in [Roß et al., 2020]. The first one is from Kletz et al. [Kletz et al., 2019], which is based on a Mask R-CNN. The authors propose an approach to use a multi-task training method, in which they simultaneously detect and segment instruments in laparoscopic images. The second one is by Gonzalez et al. [González et al., 2020], who is working on the label consistency, which means that the same instrument can be tracked through several frames without identity switches. To accomplish this, the authors propose ISINet, a Mask R-CNN that includes a temporal consistency matching over several frames [González et al., 2020] in a post-processing step.

The last one is from Robu et al. [Robu et al., 2020] who is focusing on the temporal consistency. In their approach, they produce for every frame object proposals, which are then assigned to the objects of the previous frame. This assignment is done by the help of a geometric object descriptor that can manage bounding box disambiguation to a certain degree[Robu et al., 2020]. However, this approach focuses on the tracking with bounding boxes instead of the segmentation.

---

<sup>1</sup><https://paperswithcode.com/sota/instance-segmentation-on-coco>

#### 2.2.4 Conclusions

**Self-supervised learning and labeled data:** Self-supervised learning gained interest in the field of general ML, but at the time this thesis was started, the benefit for SDS remained to be shown, specifically in the context of object segmentation. Thus, in this thesis the potential of unlabeled data will be further explored with a self-supervised learning approach.

**Available datasets:** At the time this thesis was started, there were only small datasets available, but none of them for the task of multi-instance instrument segmentation in laparoscopic videos. In addition, for most of these datasets, the labels were generated by only a small number of annotators and the labels were neither subject to quality control, nor was there a predefined labeling protocol to ensure consistent labels. Due to this, a high quality dataset will be generated and released in form of a benchmarking competition to the SDS community in order to advance the SDS research.

**Multi-instance instrument segmentation:** The literature revealed that due to missing open data, only approaches for binary instrument segmentation were available and the top performing methods in all instrument segmentation challenges are based on the U-Net architecture. For multi-instance segmentation, the Mask R-CNN is the most used architecture, and its success has been reported in the most recent literature. For this reason, the Mask R-CNN will be used and extended for a multi-instance segmentation approach, based on a in-depth statistical analysis from the benchmarking competition.



# 3 | Results

This chapter forms the central point of the entire work. It consists of five sections, each section presenting concepts, experiments and results that were developed and carried out to work towards the objectives presented in *1 Introduction*:

Sec. 3.1 is aligned to the objection *O1: Incorporation of unlabeled data into the training of deep learning models* by presenting an approach to how unlabeled data could be used to improve the training of deep learning models when only a small amount of training data is available. The core idea is to train a network that can re-colorize endoscopic images. The assumption is that through this task, the network learns important context information about the images, which is useful for the segmentation of medical instruments.

Sec. 3.2 is aligned to the objective *O2: Generation of a quality controlled dataset for the task of multi-instance surgical instrument segmentation*. One lesson that could be learned from the previous section was that the number of labeled data could possibly create a performance boost. Thus, this section is about creating a representative, quality controlled multi-instance segmentation dataset for surgical instrument segmentation.

Sec. 3.3 is aligned to the objective *O3: Structured assessment of state-of-the-art performance for multi-instance surgical instrument segmentation* by presenting a community-aided state-of-the-art determination for binary and multi-instance segmentation. The goal of this section was to enable a fair benchmark between different methods and to see if the performance with more data increases.

Sec. 3.4 is aligned to the objective *O4: Systematic problem analysis of state-of-the-art methods for multi-instance surgical instrument segmentation* by presenting an approach to how image properties (e.g., presence of blood, smoke) does affect the algorithm performance of state-of-the-art algorithms from the previous section. For this purpose, first, the image properties are annotated, and second, their effect on the algorithm performance is statistically analyzed. The purpose was to identify properties that harm performance.

Sec. 3.5 is aligned to the objective *O5: Problem-driven multi-instance surgical instrument segmentation algorithm*, where the results from Sec. 3.3 are used to design an algorithm that explicitly tackles some of the properties that were identified in Sec. 3.4, which degrade the performance of the instrument segmentation algorithms.

## 3.1 | Potential of unlabeled data

### Disclosures to this work:

Lena Maier-Hein supervised this work and was, along with David Zimmerer, part of the development of the methodology and involved in the writing process of the related publication. This work has been published in the *International Journal of Computer Assisted Radiology and Surgery* [Ross et al., 2018]. The detailed disclosure of this section can be found in Sec. 7.

### 3.1.1 Introduction

One key application in the research field SDS is the development of context-aware assistant systems in minimally invasive surgery, such as skill-assessment, surgical phase recognition, or organ segmentation and classification. With the recent success of deep learning-based methods, promising results could be obtained for such computer vision tasks. However, as the performance and learning success of those methods strongly depend on the availability of training data, such algorithms' current generalization capabilities are limited and can lead to substantial performance differences. Exemplarily, this is shown in Fig. 3.1, where the model was trained on data from the MICCAI instrument tracking challenge 2017<sup>1</sup> and tested once on test data from the same challenge and once on a subset from the ROBUST-MIS challenge [Roß et al., 2020]. Even though both datasets are videos of minimally invasive surgery, a performance drop of more than 50% can be observed.

When this part of the thesis was developed, there were three main options mentioned in the literature to address this problem: (1) domain adaptation by training, (2) domain generalization by design, and (3) access to more data [Wang and Deng, 2018]. For the first approach, domain adaptation by training, the co-variance shift is directly addressed using a small amount of labeled data from the new domain to fine-tune the model with it [Wang and Deng, 2018]. The second approach, generalization by design, aims to improve the ability to learn generalized and transferable features (e.g., [Shen et al., 2017b]) [Wang and Deng, 2018]. Finally, the last method, access to more data, would mean that either more new data would have to be recorded and annotated or that additional data would somehow be incorporated into the training process. However, since additional annotated data is very difficult to generate, especially in the medical domain, this part of the thesis specializes in the hitherto unexplored area of including additional data in the training process by avoiding a big annotation effort.

While training data is rare raw data is massively generated during the daily routine. Inspired by the recent achievements in the field of self-supervised learning [Pathak et al., 2016, Agrawal et al., 2015b, Zhang et al., 2016, Bodenstedt et al., 2017], this section is driven by the hypothesis

<sup>1</sup><https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/>

that such raw data could be used to learn a representation of the target domain that boost the performance of state-of-the-art algorithms such that it leads to a reduced labeling effort.

In the following, first is an overview provided, followed by a prototype implementation of the concept which is described in Sec. 3.1.2, followed by the experiments to investigate the hypothesis in Sec. 3.1.2. The results of the experiments are presented in Sec. 3.1.3, discussed in Sec. 3.1.4 and concluded in Sec. 3.1.5.

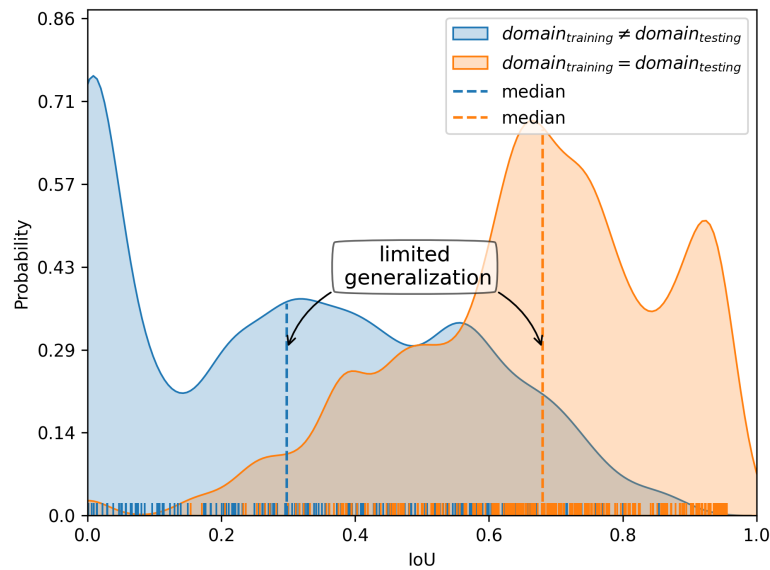


Figure 3.1: Distribution of the performance scores of a state-of-the-art deep learning model (U-Net), evaluated on test data that is from the same distribution (orange) like the training domain (e.g. same hospital) and the ones evaluated on test data that is from another distribution (blue) like the training domain (e.g. different hospital). It can be seen that the median performance drops, due to missing generalization capabilities.

### 3.1.2 Methods

As written above, hospitals generate massive amounts of data during the daily routine, but there still typically exists no reference annotations, and thus this data is unusable for traditional supervised learning approaches. However, recent publications have shown that by using *naturally existing* labels in the form of an auxiliary task that requires an understanding of the data, this can be used to pre-train a model to learn the underlying data representation [de Sa, 1994, Pathak et al., 2016, Agrawal et al., 2015b, Zhang et al., 2016, Bodenstedt et al., 2017].

This next section provides a conceptual overview of the developed self-supervision approach, followed by a prototype implementation on the example of minimally invasive instrument segmentation.

### Concept overview

The presented concept consists of six components and is organized in a four step process. In general, there is a large representative amount of  $N_{unlabeled}$  unlabeled images  $I_{unlabeled}$  and a comparatively small amount of  $N_{labeled}$  labeled images  $I_{labeled}$  for a target task (e.g. computer-vision task), where  $N_{unlabeled} \gg N_{labeled}$  [Ross et al., 2018]. Following the concept of self-supervision, there has to be an auxiliary task, that requires no manual labeling but still forces the model to leverage information of the unlabeled data with the goal being to improve the generalization capabilities of deep learning models [Ross et al., 2018].

On the example of instrument segmentation, a re-colorization of gray-scale images was chosen as an auxiliary task, where the re-colorization was done with an adversarial approach. For the target task (here: segmentation), the pre-trained model was used and fine-tuned on a small proportion of images with reference annotations. The overall auxiliary training process is as follows (see Fig. 3.2): (1) raw data was extracted from videos of a minimally invasive procedure, then (2) references for a very small portion of those frames were manually annotated, (3) the color information of the remaining unlabeled data was removed and a model was trained to re-colorize such images. For the target task training, the model used to re-colorize the images was fine-tuned with the small proportion of images with reference annotations, and thus fine-tuned on the actual target task.

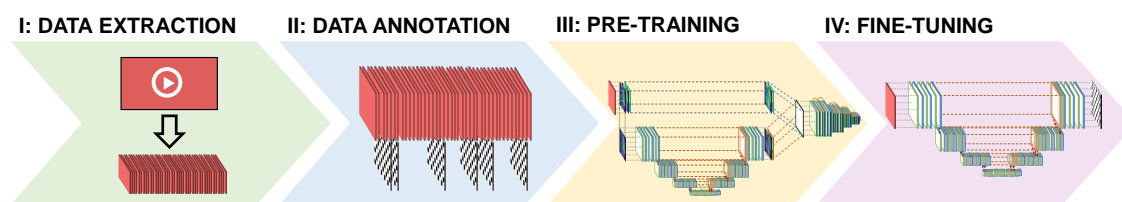


Figure 3.2: The concept applied to the exemplary task of instrument segmentation of minimally invasive surgery videos. First, the raw video data was taken, and all frames were extracted for the second step, were a tiny proportion of video frames reference annotations were manually created. In phase III, the extracted raw video frames were used for a self-supervised pre-training step to leverage information about the target task. Finally, the model from step III is used and fine-tuned with the small amount of training data, depending on the target task.

### Prototype implementation

For the prototype implementation, the U-Net [Ronneberger et al., 2015] was chosen as a segmentation network, as it is currently one of the most widespread networks in the medical community [Twinanda et al., 2017, Garcia-Peraza-Herrera et al., 2017b, Pakhomov et al., 2017, García-Peraza-Herrera et al., 2016]. It is important to mention that the chosen architecture should be suited to solve both the auxiliary and the target task.

**Auxiliary task** The idea of an auxiliary task is to use *naturally existing labels* in the data, so that the model has to learn an underlying distribution. Applied to the instrument segmentation would mean that a task needs to be chosen where a model has to learn to differentiate between

background and instruments, given the nature of the data. One example of such a task is the re-colorization. Using only grayscale images and training a network to re-colorize them, the model has to learn to distinguish between an instrument and background, that includes e.g., blood, tissue, and fat. However, the training of such a network is challenging due to the lack of appropriate metrics to automatically assess the quality of the re-colored image compared to the original one. For example, small changes in the pixel values can have strong effects on an error, whereby the generated coloring is semantically correct [Larsen et al., 2016].

The problem of the missing metrics was addressed by using a Generative Adversarial Network (GAN) approach (see Sec. 2.1.4), as described in Larsen et al. [Larsen et al., 2016]. In comparison to a normal GAN, the input was a grayscale image, rather than a random vector. As loss function was only the discriminator output used, since most of the low level semantic information is already encoded in the L-channel [Ross et al., 2018].

For the re-colorization task, the image was transformed from the *RGB* into the *CIE 1976 L\*a\*b\** color space, where the  $(L, a, b)$  channels of the color space are defined by the luminescence (L), the color gradient from green to red (a) and the color gradient from blue to yellow (b). The generator of the GAN was therefore trained to predict the resulting a- and b-channel, conditioned on the L-channel. In the following, the generator and discriminator are further described.

*Generator* The input of the U-Net [Ronneberger et al., 2015] is the luminescence channel  $I^L$  and the desired outputs are corresponding a and b channels  $I^{\hat{a}, \hat{b}}$ . The activation function of the final output layer was  $\tanh$ , resulting in  $G(I^L) \in [-1, 1]$  [Ross et al., 2018]. Compared to the target task, which uses all three channels (L, a, b), the pre-training only uses the L-channel. To this end, two dummy channels were added to keep the input dimensions to three in total. During the auxiliary task training those two channels were ignored, while during the target task training they got used, and the network learned to include them into its decision. The training of the generator was performed with the following loss functions, consisting of three terms:

$$\mathcal{L}_G = \gamma \mathcal{L}_1 + \lambda \mathcal{L}_2 + \varphi \mathcal{L}_3 \quad (3.1)$$

with  $\gamma, \lambda$  and  $\varphi$  as weighting factors.  $\mathcal{L}_1$  (see Eq. 3.2) is the least squared GAN loss [Mao et al., 2017], that measures how likely it is that a generated image  $\hat{I} = G(I^L)$  is classified as a real image  $Y_{real}^D$ , as given by the output of the discriminator  $D(\hat{I})$  [Ross et al., 2018].

$$\mathcal{L}_1(\hat{I}, Y) = \left( D(\hat{I}) - Y_{real}^D \right)^2 \quad (3.2)$$

Depending on the dataset, the distribution of the two color channels  $ab$  are somewhat different. While in endoscopic images there is a strong imbalance towards red, yellow, and black values (due to blood, fat, black background), the color distribution, especially in natural datasets, are distributed differently (see Fig. 3.3). Due to that observation, the generator loss was extended with the  $\mathcal{L}_2$  term (Eq. 3.3) to compensate for the unbalanced values and to stop dominating a re-colorization with the most frequented values [Ross et al., 2018]. Similar to Zhang et al. [Zhang et al., 2016], the empirical color distribution  $\tilde{p}_c$  for each channel  $c = \{a, b\}$  was obtained at a grid size of 1. Thus, if the model should choose a high frequent color value, this would generate an increased loss.

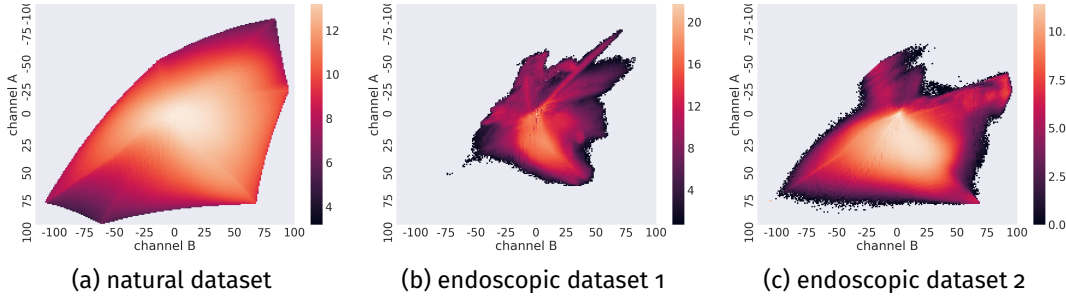


Figure 3.3: Presented are the log color distributions for the  $a$  and  $b$  channel of all images of three datasets, one natural dataset (a) and two endoscopic datasets (b, c). It can be seen that the endoscopic datasets (b, c) contain only a subset of the appearing colors in the natural dataset (a). However, red and white colors occur more frequently in endoscopic datasets (b, c) than in the natural dataset (a). In addition to this, there are also differences between the two endoscopic datasets, where the endoscopic dataset 2 (b) contains more red and yellow values than the endoscopic dataset 1 (a).

$$\mathcal{L}_2(\hat{I}, I) = \frac{1}{2 \cdot h \cdot w} \sum_{c=a,b} \left[ \sum_{h,w} ((\hat{I}_{h,w}^c - I_{h,w}^c) \cdot P_{h,w}^c)^2 \right] \quad (3.3)$$

with  $P^c$  being a weighting factor as the complementary relative color frequency, given by  $\tilde{p}_c$  [Ross et al., 2018].

As learning of rare values would result in miscolored images,  $\mathcal{L}_3$  is used as an antagonist to  $\mathcal{L}_2$  in order to learn a colorization similar to the original image [Ross et al., 2018].

$$\mathcal{L}_3(\hat{I}, I) = \frac{1}{2 \cdot h \cdot w} \sum_{c=a,b} \left[ \sum_{h,w} (\hat{I}_{h,w}^c - I_{h,w}^c)^2 \right] \quad (3.4)$$

**Discriminator** The discriminator was trained with real images  $I$  and re-colored images  $\hat{I}$ . As architecture, a vanilla ResNet18 [He et al., 2016b] was used with a *softmax* activation function, leading to  $D \in [0, 1]$ , with the labels  $Y_{fake}^D = 0$  and  $Y_{real}^D = 1$ . The loss function was a MSE error [Ross et al., 2018].

$$\mathcal{L}_D = (D(I) - Y_{real}^D)^2 + (D(\hat{I}) - Y_{fake}^D)^2 \quad (3.5)$$

**Target task: Instrument segmentation** Depending on the availability of training data, there are two possible variants on how the training could be performed. The first variant (**Td**) assumes that no additional data from other sources (e.g. other hospitals) are available. The second variant (**Ed-Td**) includes further datasets into the training (see paragraph *Validation data*).

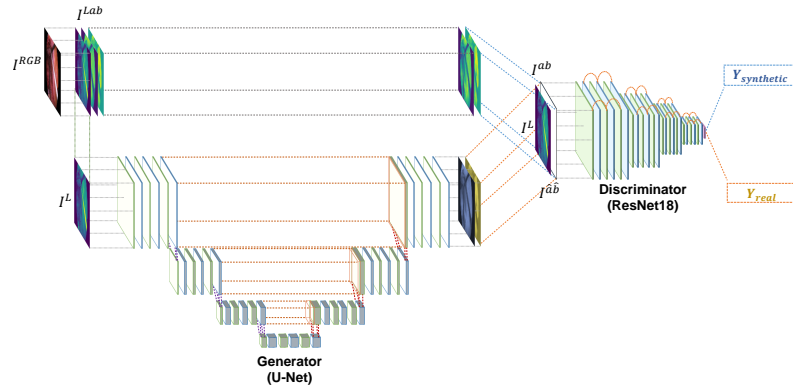


Figure 3.4: Pre-training using self-supervised learning and a conditional generative adversarial network (cGAN) approach. First the image is transformed into the  $Lab$  color space. The luminescence layer  $L$  is fed into the generator  $G(I^L)$  (U-Net) which is trained to generate the corresponding  $\hat{a}$  and  $\hat{b}$  channels. The discriminator  $D$  (ResNet18) is trained to differentiate between real images  $I = \{L, a, b\}$  from the target domain and synthetic images  $\hat{I} = \{L, \hat{a}, \hat{b}\}$  produced by the generator.

1. **Target domain training (Td):** After the model was pre-trained on the re-colorization task, it was fine-tuned for the instrument segmentation in the next step. For this purpose, all learned parameters of all layers were taken from the pre-training and only the last layer was randomly assigned as it is responsible for the classification of each pixel as either instrument or background. The loss-function used for training was the cross-entropy between the output and the groundtruth (see Eq. 2.21) segmentation.
2. **External domain target domain training (Ed-Td):** When additional data is available, it might be helpful to include it as it can lead the model to learn more generalized features. For this reason, first the pre-trained network is trained on the additional data the same way as described in TDi and in the second step fine-tuned on the sparsely labeled dataset.

### Experimental design

This section starts with a list of all validation data and chosen hyperparameters used for the experiments described below. In total, four different experiments were performed to answer the central hypothesis: "Raw data could be used to learn a representation of the target domain that boosts the performance of state-of-the-art algorithms".

The first experiment was about identifying a potential increase in performance, depending on the number of *labeled* training data. The second experiment was whether a potential performance gain would still be existent if training data would be augmented. The third experiment is about whether there are different performances for the last experiment if the pre-training would be done on data of other domains. Finally, the presented concept was compared to current state-of-the-art pre-training methods.



**Validation data** Experiments were performed on three publicly available datasets, coming from the natural domain (images show content from the real world, e.g., cats or dogs), the target domain (minimally invasive surgery with laparoscopic instruments), and another medical domain (minimally invasive robotic surgery). For all three datasets, a labeled and unlabeled part was defined.

- Unlabeled data

- *COCO (natural domain)*: A subset of the COCO dataset [Lin et al., 2014b] of 20k images was chosen, containing all images of the class cat and 16,692 additional randomly selected images of the remaining classes to approximate the real underlying color distribution.
- *Robo (other medical domain)*: The 21 unpublished endoscopic videos of the *EndoVis instrument segmentation challenge in robotic surgeries*<sup>2</sup> were used as other medical domain.
- *HeiCo (target domain)*: 6 endoscopic videos used for EndoVis surgical workflow challenge<sup>3</sup>.

- Labeled data

- *COCO (natural domain)*: A subset of the COCO dataset [Lin et al., 2014b] of 2,818 images with the classification *cat* was chosen. The class was chosen as many other classes suffer from poor references and ambiguities [Heim et al., 2017]. In addition, like the instrument segmentation, the target object should be in the foreground.
- *Robo (other medical domain)*: All 2,400 endoscopic images of the challenge dataset. Training and testset were already defined by the challenge and were disjunct.
- *HeiCo (target domain)*: 809 annotated images for the binary instrument segmentation, extracted from 6 videos of the HeiCo dataset. The data was split into 413 (three surgeries) training, 119 (one surgery) validation, and 277 (two surgeries) test images. The sets of videos corresponding to testing images and training/validation images were disjunct and randomly chosen.

**Hyperparameters** All hyperparameters were optimized based on 80% of the training data and a fixed validation set, consisting of 20% of the training data. A preliminary hyperparameter space search was performed in the beginning, and all parameters were kept fixed for all experiments.

The color distribution  $\hat{p}$  was generated by creating the histogram over all training images and used for the re-colorization on the corresponding dataset. Re-colorization training was done with a batch size of six and learning rates of 0.0005 (generator) and 0.002 (discriminator). Both generator and discriminator were optimized with the Adam optimizer [Kingma and Ba, 2014]. The weighting parameters  $\gamma$ ,  $\lambda$  and  $\varphi$  for  $\mathcal{L}_1$ ,  $\mathcal{L}_2$  and  $\mathcal{L}_3$  were set such that their scale was equally ranked. Based on visual exploration, training was stopped after 20 epochs.

Like the generator, the optimizer for the instrument segmentation was the Adam optimizer, with a learning rate and batch size identical to the generator. Besides, a scheduler reduces the learning rate by a factor of 0.1 if a loss plateau lasts for ten steps was used. Due to visual exploration, training was stopped after 150 epochs.

<sup>2</sup><https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/>

<sup>3</sup><https://endovissub2017-workflow.grand-challenge.org/>

**Re-colorization of images** As described in the methods concerning why a GAN approach was chosen to re-colorize images, there is a lack of metrics to evaluate a possible valid colorization. Thus, only a qualitative evaluation can be performed. When using the GAN approach, the main question is to examine how the re-colorization of images depends on the domain. For this reason, three models were trained on three different datasets of different domains. The first model was trained on the target domain (HeiCo dataset), the second one on a related medical domain (Robo dataset), and the last one on the natural domain (Coco dataset). Finally, the results were qualitatively evaluated.

**Effect of training data size** To investigate the potential benefit of the performance of the  $T_d$  and  $Ed-T_d$  training, the *HeiCo* dataset was divided into five randomly selected disjunct subsets of the training datasets, each of size  $k \cdot N$ , where  $k \in \{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}\}$  and  $N = 413$ , denoting the number of labeled training images. For the  $Ed-T_d$  pre-training, additionally the *Robo* dataset was included. Testing was performed on the complete *HeiCo* test set. Target metric was the DSC. The baseline model was compared with the  $T_d$  and  $Ed-T_d$  training procedure by training a model on each of the five different subsets of a single training dataset fraction. In the next step, all those five models were applied to the test data, their DSC performance scores grouped per image, and aggregated with the mean. A potential statistically significant better performance than the baseline was examined by fitting linear mixed models [McCulloch and Neuhaus, 2001] on the fractions  $\frac{1}{20}$  to  $\frac{1}{2}$ , with the training method as fixed and the test image as random effect variables. Because linear models require normally distributed data, an arc-sine transformation was applied on the mean DSC values. As multiple hypotheses were similarly tested, all p-values were adjusted by Dunnett's test and Bonferroni-Holm correction.

**Effect of data augmentation** Data augmentation is one of the most commonly used techniques, especially when little training data is available [Goodfellow et al., 2016a]. Thus, investigating whether the presented concept is superior to the data augmentation, or even increasing the performance, was conducted. For this reason, the same experimental setup as described in *Effect of training data size* was performed apart from the augmented training data. Training data was augmented by 50% chance of mirroring, rotation ( $90^\circ$ ,  $180^\circ$ ,  $270^\circ$ ) and adding of Gaussian noise (20%).

**Effect of the unlabeled data domain** Especially when only a small amount of training data is available, including additional data from other domains would be an excellent opportunity to increase the total training data to learn more generalized features. Thus, the  $T_d$  training was performed on datasets of different domains, outside (COCO) and inside the medical domain (Robo), to examine a possible effect. All experiments were performed as described in the paragraph *Effect of training data size*.

**Comparison with other pre-training methods** A comparison to other state-of-the-art pre-training techniques is important so as discern a potential benefit of the presented methods. For this reason, two approaches were implemented: (1) **non medical** pre-training, by using the non-medical dataset (COCO), train a model on the available images and fine-tune it on the target domain and (2) **medical** pre-training, that follows the same procedure but uses the labeled dataset of the medical domain (Robo). Both methods were compared to the  $T_d$  training approach.

### 3.1.3 Results

#### Performance of re-colorization

According to the performed experiments, the cGAN-based approach generally produces realistic-looking images. However, due to the different underlying color-distributions, the re-colorization strongly depends on the domain used for training. As can be seen in Fig. 3.5, the natural domain especially produces slightly green images, while the images from the medical domain and the target domain differ only in the presence of red and yellow. How this affects the performance is evaluated in the experiment *Effect of the unlabeled data domain*.

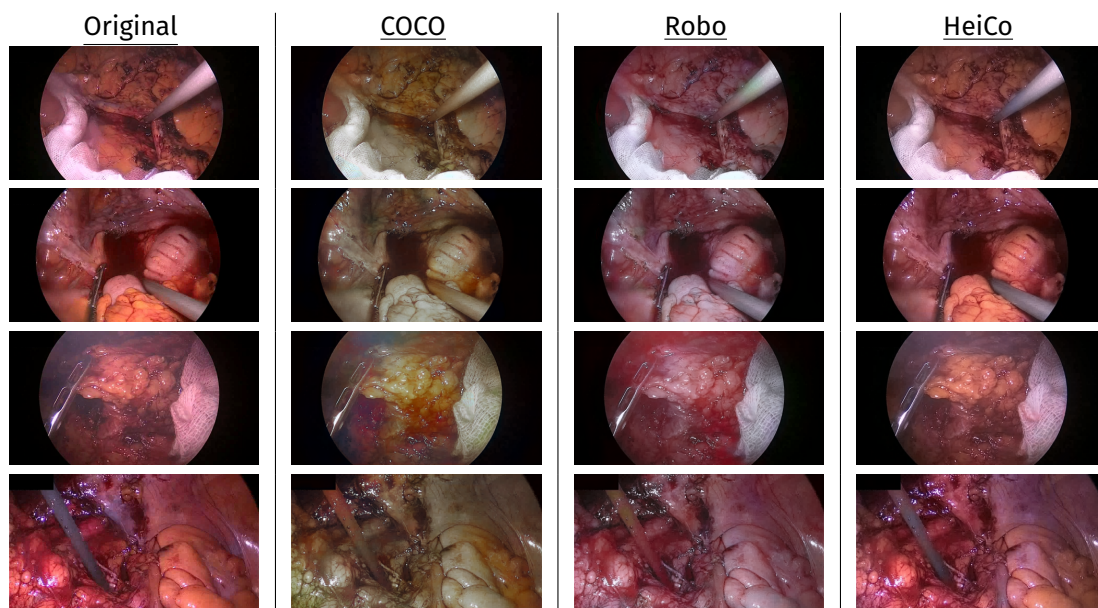


Figure 3.5: This figure contains a visualization of re-colored images that were trained on datasets of different domains. It can be seen that the color reconstructions are different because not all colors are represented equally in a given dataset. Predominant differences in red and yellow values occur more frequently in endoscopic videos than outside of the endoscopic context.

#### Effect of training data size

While there exists almost no differences in the median performance for the biggest fraction of  $\frac{1}{2}$ , for the smaller fractions the median differences of the DSC values for the  $Td$  pre-training lies between  $[0.04, 0.06]$  and for  $Ed-Td$  between  $[0.04, 0.13]$ . This is also reflected in the significance, where for the largest fraction of  $\frac{1}{2}$  there was no significant improvement (p-value for  $Td$  of 0.89 and  $Ed-Td$  of 0.53) over the baseline, with a significance level of  $\alpha = 0.05$ . However, except for the fraction  $\frac{1}{20}$  (p-value 0.01), which can be considered as an outlier, the performance of  $Td$  and  $Ed-Td$  are significantly better than the baseline (p-values  $< 0.001$ ) for the small fractions between  $\frac{1}{20}$  and  $\frac{1}{3}$ . This can also be seen in Fig. 3.6a.

In almost all experiments, the presented method boosted the performance of the segmentation method by, at the same time, reducing the labeling effort. The labeling effort can be reduced

by, at the same time, getting the same or even higher median performances than the baseline (e.g., using  $\frac{1}{16}$  of the data with the *Td* method to get the same performance as with  $\frac{1}{8}$  with the baseline, resulting in  $> 60\%$  less data). Even better results can be achieved by using additional data from a similar domain, as can be seen in Tab. 3.1 that provides descriptive statistics the DSC when using  $\frac{1}{16}$ th and  $\frac{1}{8}$ th of the training set images.

### Effect of data augmentation

According to the results presented in Fig. 3.6b, data augmentation does not only substitute the pre-training process but even improves the performance evenly for all methods. Especially in fractions smaller than  $\frac{1}{6}$ , the *Ed-Td* training, that combines the pre-training with additional labeled data from a similar domain led to better results than the baseline.

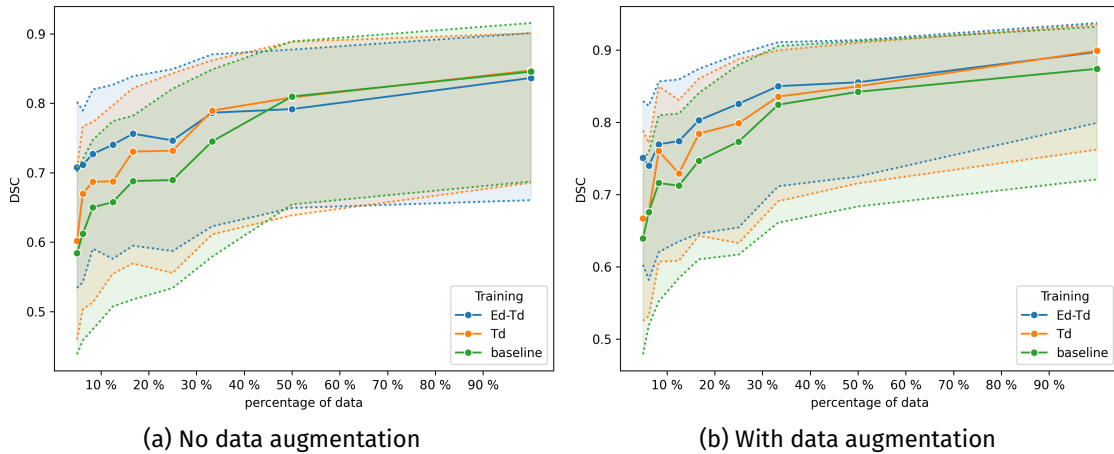


Figure 3.6: Median Dice Similarity Coefficient (DSC) and the Interquartile Range (IQR) as a function of training data size for the two methods (*Td* and *Ed-Td*) in comparison to the baseline training (training only on the available training data). It can be seen in that both methods (*Td* and *Ed-Td*) outperform the baseline training in both cases when no data augmentation on the training data is applied and also when there is data augmentation applied.

### Effect of pre-training domain

The results of the experiments, that are also presented in Fig. 3.7, show that the best performance of the segmentation can be achieved when pre-training is performed on the same domain as the final target domain, independent from the number of training data. While in smaller fractions the use of related data from a medical domain seems to produce better results than using a different domain, this effect disappears as more data is available. However, regardless of the domain used for pre-training, it always performs better than the baseline.

### Comparison and combination with other pre-training methods

As presented in Tab. 3.1, SOA training with labeled images from the medical domain SOA (medical) outperforms the pre-training on images from the non-medical domain SOA (non-medical) and

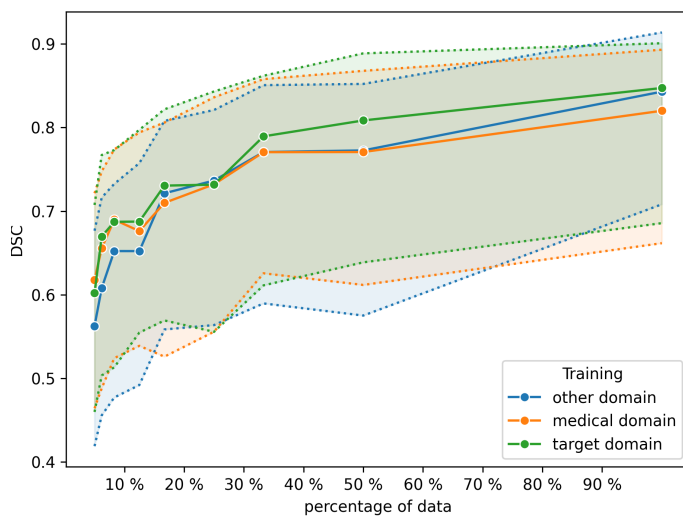


Figure 3.7: Effect of pre-training domain. For small training datasets, medical images yield better results than non-medical images.

yielded even better results than pre-training exclusively on unlabeled data Td (target medical). However, combining state-of-the-art pre-training on medical data with the re-colorization task gave the best results (Ed-Td (target medical + other medical)). In general, the more training data that was used, the closer the results of the individual training methods.

Table 3.1: Two state-of-the-art (SOA) methods are compared to three variants of the re-colorization pre-training. The target metric DSC is given for the two fractions  $\frac{1}{8}$ th and  $\frac{1}{16}$ th of the training set images. The mean, median and interquartile range (IQR) values are shown along with the improvement in % compared to the baseline method (no pre-training).

Training type	DSC for dataset fraction: $\frac{1}{8}$			DSC for dataset fraction: $\frac{1}{16}$		
	mean	median	IQR	mean	median	IQR
baseline	0.62 (-)	0.66	[ 0.51 0.77]	0.57 (-)	0.61	[ 0.46 0.72]
SOA (non-medical)	0.63 (1%)	0.67	[ 0.52 0.78]	0.57 (-1%)	0.62	[ 0.43 0.73]
SOA (medical)	0.66 (6%)	0.71	[ 0.53 0.82]	0.59 (3%)	0.65	[ 0.47 0.75]
Td (target medical)	0.65 (5%)	0.69	[ 0.55 0.80 ]	0.61 (7%)	0.67	[ 0.50 0.77]
Ed-Td (other medical)	0.65 (5%)	0.68	[ 0.54 0.79]	0.60 (5%)	0.66	[ 0.49 0.75]
Ed-Td (target medical + other medical)	0.68 (10%)	0.74	[ 0.58 0.83]	0.65 (13%)	0.71	[ 0.54 0.79]

### 3.1.4 Discussion

Approaches to use self-supervised or semi-supervised techniques to include unlabeled data into the training that is of the same distribution as the training data is already a common technique in the computer vision community and also in the field of medical image computing [Baur et al., 2017]. Other state-of-the-art self-supervised methods, such as [Tajbakhsh et al., 2016, Zhou

et al., 2017, Tajbakhsh et al., 2017, Ravishankar et al., 2016] or unsupervised methods [Kamnitsas et al., 2017] are mainly focusing on how to find domain-invariant features to either generalize better or boost the algorithm outcomes. For this, they mainly focus on the potential of various auxiliary tasks, such as re-ordering [Noroozi and Favaro, 2016], classification [Zhang et al., 2016, Noroozi and Favaro, 2016], and in-painting and re-colorization [Pathak et al., 2016, Zhang et al., 2016, Zhang et al., 2017b, Larsson et al., 2017]. One of the closest work was authored by Bodenstedt et al. [Bodenstedt et al., 2017], who pre-trained a deep learning model by estimating the order of two random video frames in an auxiliary task to get a better prediction for surgical workflow phase recognition.

While the previously mentioned techniques mainly try to improve the performance of a model, the focus of this chapter was not to get a state-of-the-art performance algorithm for the segmentation of instruments. This is also reflected in the fact that only very basic data augmentation technologies were used, and no excessive hyperparameter-search was performed to achieve a score that is better than other published techniques. Instead, the intention was to use the concept of self-supervision to reduce the manual labeling effort, which has not been investigated in the field of medical image analysis. Although no absolute top performance was sought, the performance was nevertheless comparable with published results [Bodenstedt et al., 2018].

Based on the presented results, it can be concluded that the presented method is well-suited to situations where only a small amount of training data is available. This shows that self-supervised learning can generate additional features that cannot be generated only by applying data augmentation techniques. Of course, it must be said that the data augmentation techniques used for the baseline are only a small subset of all data augmentation techniques available. However, the use of data augmentation also requires a good understanding of the domain, since, for example, unrealistic augmentations (e.g. unrealistic deformations) would more likely worsen rather than improve the performance. This bias is not required when using the self-supervised approach. Fortunately, the results also suggest that data augmentation does not counteract self-supervised learning, but actually complements it (at least the augmentation techniques presented here). This would be a further possibility to improve the results again.

The results also show, however, that self-supervision and data augmentation is of particular benefit when only a little data is available. The more training data that can be used, the smaller the effect. It is particularly useful in the regions using only  $\frac{1}{6}$  of the training data. This can probably be deduced from the fact that features that can be learned with more data are still difficult to replace.

Regarding the results from the pre-training domain's effect, it should be mentioned that the selection of the segmentation class "cat" for pre-training segmentation tasks of the natural domain data could have had an impact on the lower performance in comparison to the medical domain. While a cat is a living being, choosing an artificial object might have produced a different result, one closer to the medical domain. Thus, a thoughtful selection of a related task or data might increase the algorithm performance. This guess is in line with the publication from Zamir et al. [Zamir et al., 2018]. It can also be said that the features that describe natural data are very likely to differ from the medical image data (this can also be seen in the different color distributions of the datasets in Fig. 3.3). However, exploring this further would have been outside the scope of this work.

When comparing the performances from the baseline for the two medical datasets, it can be noticed that the performance on the HeiCo dataset is much lower than the performance on the Robo dataset. However, this can be attributed to the comparatively low variability of the Robo dataset in comparison to the images from the HeiCo dataset.

So far, no publication could be found in the literature which determines exactly which performance an algorithm must achieve before it should be used in practice. Notwithstanding, it must be assumed that a particularly reliable algorithm is required, especially in the field of SDS. Accordingly, an algorithm should achieve at least a DSC metric value of 0.9. This value has, however, not yet been achieved with this concept either.

### **3.1.5 Conclusion**

In conclusion, the cGan-based re-colorization auxiliary task seems to be the right choice when used as pre-training for the task of medical instrument segmentation when only a small amount of labeled data is available. Besides, this method allows both labeled and unlabeled data to be used. However, when looking at the final segmentation results it can be concluded that the performance is not high enough for a successful transition to clinical practice. Although the segmentation quality can be improved with this approach, it is still too low to be used in everyday clinical practice. Because the methodological approach presented here does not yet produce sufficiently good results, creating an improved and larger dataset yields the most promising way to increase performance further.





## 3.2 | Quality controlled dataset generation

### Disclosures to this work:

Lena Maier-Hein supervised this work and was, along with Annika Reinke and Martin Wagner, part of the development of the methodology and involved in the writing process of related publications. Parts of this work have been published in the *Medical Image Analysis- Journal* [Roß et al., 2020] and are currently subject to a minor revision in *Nature Scientific Data - Journal* [Maier-Hein et al., 2020b]. The detailed disclosure of this section can be found in Sec. 7.

### 3.2.1 Introduction

As described in Sec. 2.1 *Machine learning*, many machine learning-based algorithms require vast amounts of labeled training data for training. However, especially in the medical domain, there is a lack of such datasets, which is one of the major bottlenecks for the development of novel methods [Chen et al., 2019]. Based on the findings of the previous section (see Sec. 3.1.5), methodological approaches, such as self-supervised learning, can partially compensate for the lack of such data. Nonetheless, the performance achieved with it (at least in the specific case of instrument segmentation) is not yet sufficient to use the algorithms in the clinic.

As literature research on available datasets has shown, there are already a few datasets that use medical instruments in laparoscopic or robotic operations. However, these datasets often experience problems that they: (1) do not contain enough data and are therefore often not representative, (2) are often not subjected to any quality control, and (3) do not use a clearly defined annotation protocol for possible edge cases, which can lead to inconsistencies in the data.

In addition to this, the presented datasets often contain a pure binary segmentation of the instruments, i.e., they classify a pixel whether it is an instrument or not. However, since more than just one instrument can often be seen in an image, the binary classification would fail when tracking multiple instrument instances.

This section aims to create a dataset that addresses the above-mentioned problems (representativeness, size, quality) for the multi-instance segmentation of medical instruments in laparoscopic image data.

In the following sections, the data recording, extraction and annotation are described in the Methods (Sec. 3.2.2). The outcome of this process is then presented in the Results (Sec. 3.2.3) in combination with statistical evaluations. Finally, a Discussion (Sec. 3.2.4) and the final Conclusion is presented in Sec. 3.2.5.

### 3.2.2 Methods

A couple of requirements need to be met in order to generate a high-quality SDS dataset. The data must primarily represent the clinical application with a clearly defined target (e.g., the definition of an instrument). Representative means that data has to include so-called edge cases and all the potential difficulties that algorithms will face. Edge-cases are those which are either challenging to decide or rarely occur, while examples for difficulties are blood or smoke. There must then be enough data so that both the data itself and the high variability of medical image data can adequately cover the occurrence of rare conditions. Data should not show any inconsistencies (e.g., a visible instrument in a trocar is sometimes segmented, sometimes skipped) so as to not confuse a model during training. The quality and consistency of the segmentation must be ensured. Finally, data protection must be taken into account, both for the privacy of the patient and that of the surgical team.

As mentioned above, a dataset must be representative, contain edge-cases and possible difficulties. However, this requires an understanding of edge-cases and difficulties. Laparoscopic image data also have the peculiarity that several possible difficulties can occur at once. Thus, a comprehensive examination of image characteristics, which have to be defined, is of particular interest.

In the following sections, a method is described to produce a dataset that fulfills the above-mentioned requirements, labels the tasks, and defines image characteristics.

#### Data recording

Thirty videos of three different minimally invasive surgery procedures (proctocolectomy, rectal resection, and sigmoid resection) with ten videos per surgery type were recorded during daily routine at the Heidelberg University Hospital, Department of Surgery. The laparoscopic camera was a Karl Storz<sup>4</sup> Image 1, recorded with 25 frames per second (fps) and was equipped with a 30° optic lens and a light source Xenon 300 from Karl Storz. All parts of the video outside of the human body were manually excluded to ensure the privacy of both the patient and the surgical team. As the recorded video data has a HD resolution with 1920×1080 pixels, it was reduced to 960×540 to reduce the memory usage.

#### Data extraction

Since the annotation of all the video frames from all thirty videos would result in more than 10M images, and thus would require an unmanageable amount of labeling effort, only a subset of video frames was chosen. Considering that the information of neighboring video frames due to the high fps is often highly redundant, all videos were resampled at a rate of 1 fps, resulting in 4,456 video frames. Since surgeries typically follow a specific procedure and the so-called phase transitions are of major interest for numerous applications such as workflow recognition [Twinanda et al., 2017, Hashimoto et al., 2019] or surgeon skill assessment [Law et al., 2017, Lin et al., 2019], additional 5,584 video frames were selected during those phase transitions. The labels of the surgical phases were available from another challenge *EndoVis Surgical Workflow Analysis in the SensorOR*<sup>5</sup>, which was based on the same videos.

---

<sup>4</sup>Karl Storz SE & Co. KG, Tuttlingen, Germany

<sup>5</sup><https://endovissub2017-workflow.grand-challenge.org/>

### Instrument annotation

The initial segmentation of all instrument instances was performed by the professional labeling company *Understand AI*<sup>6</sup>. However, the so generated reference annotations contained many inconsistencies and would have degraded the performance of a machine-learning-based algorithm. For this reason, all images were reviewed again, inconsistencies and ambiguities identified and the annotation was standardized by the creation of a set of labeling rules, written down in a labeling protocol.

In the labeling protocol, first, an instrument was defined as "elongated rigid object put into the patient and manipulated directly from outside the patient" [Maier-Hein et al., 2020b], followed by the instructions on how to label an instrument instance, which are in short [Maier-Hein et al., 2020b]:

- I **Occlusions:** Each pixel may correspond to exactly one structure. Specifically, the solid/liquid matter that occurs first along the line of sight of the endoscope determines the label. This may result in multiple contours for a single instrument that is occluded by another instrument, object, blood or tissue.
  
- II **Transparencies:** Medical instruments may be transparent. The occlusion rule holds in this case as well.
  
- III **Overlays:** Text overlays shall be ignored while image overlays are treated as being not part of an instrument.
  
- IV **Holes in instruments:** A hole is made up of pixels that do not show parts of the instrument, but are either: a) completely surrounded by pixels of the same instrument or (b) are completely surrounded by pixels of one instrument and the margin of the image where it is known, from video context, that the instrument would close the hole outside the image. Following recommendations of previous challenges and the given difficulties of localizing these holes, they are regarded as part of the instrument. The sole exception is trocars when the camera is placed inside of them.

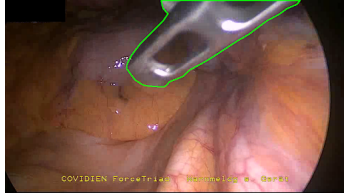
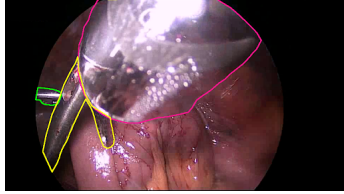
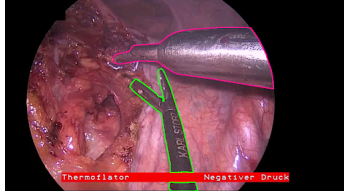
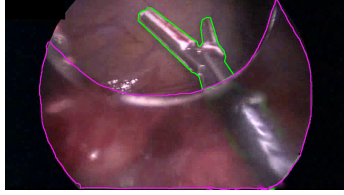
The application of those rules is shown in Tab. 3.2, an excerpt of examples of the labeling protocol as described in [Maier-Hein et al., 2020b].

Following definition of the rules, four medical students and 14 engineers reviewed and corrected all annotations. In ambiguous or unclear cases, a team of two engineers and one medical student generated a consensus annotation [Maier-Hein et al., 2020b]. To ensure the quality, first a medical expert reviewed all refined images and, in case of a potential error, corrected it after consultation with an engineer [Maier-Hein et al., 2020b].

---

<sup>6</sup><https://understand.ai>

Table 3.2: Excerpts from the labeling protocol [Maier-Hein et al., 2020b]. The visible instrument instances in the images are labeled according to the defined rules.

IMAGE	INSTRUCTIONS
	<p>The two holes of the medical instrument are regarded as part of the instrument (rule IV).</p>
	<p>A medical instrument is visible at the end (opening) of the trocar. The part covered by the transparent trocar is not regarded as a visible part of the instrument (rules II, III).</p>
	<p>The instrument represented by the green contour comprises two parts due to image overlay (rule III).</p>
	<p>A medical instrument is visible at the end (opening) of the trocar. The part covered by the transparent trocar is not regarded as a visible part of the instrument (rules II, III).</p>

### Image characteristic annotation

To define image characteristics, first a set of possible difficulties in images from minimally invasive surgeries were identified in the literature (e.g., [Bodenstedt et al., 2018, Allan et al., 2020]) and extended by the personal experience that was gained during the previous annotation process. All characteristics were grouped into three classes (see table 3.3) that describe where a characteristic can appear, namely (1) local characteristics of instruments or global characteristics, either in (2) background, or (3) foreground. In the next step, a trained engineer annotated the presence of such characteristics for all images in the dataset. The annotation was supported by a framework that guided the entire process (see Fig. 3.8). While local characteristics (e.g. blood on the instrument) were annotated individually for each instrument instance, global properties were assessed for the complete image.

Table 3.3: Overview of all image characteristics that a human annotator should assign to each image in the training and test data. The table shows if the presence of a characteristic was assessed for the background, for each instrument instance, or globally, for the complete image.

Characteristic	Background	Instrument	Image
Covered by blood?	✓	✓	✗
Covered by smoke?	✓	✓	✗
Covered by tissue?	✓	✓	✗
Subject to motion artifacts?	✓	✓	✗
Covered by specular reflections?	✓	✓	✗
Covered by another instrument?	✓	✗	✗
Covered by any other object (non surgical)?	✓	✓	✗
Too bright?	✗	✓	✗
Too dark?	✗	✓	✗
Is the image well-illuminated?	✗	✗	✓
Does the lens seem dirty?	✗	✗	✓

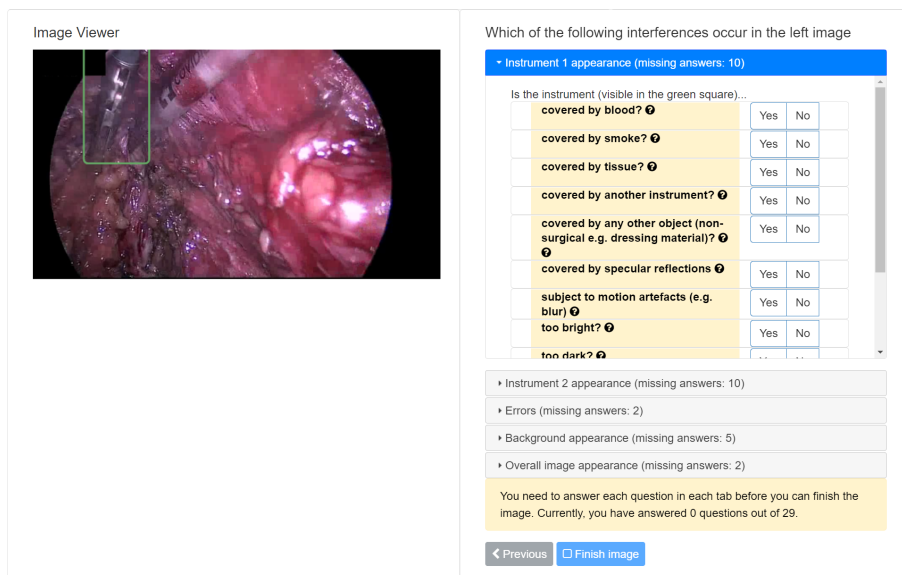


Figure 3.8: Illustration of the image characteristics annotation tool. The whole annotation process is guided, the region that should be rated by a human expert is enclosed by a bounding box and for all possible characteristics examples are given.

### 3.2.3 Results

This presented method led to a total of 10,809 annotated instrument instances of 10,040 images. A detailed summary is provided in Table 3.4. The verification of the annotations was already part of the data annotation procedure. The image characteristics can be seen in Fig. 3.10 for the entire dataset. It is noticeable that four properties are particularly common in the data, namely bloody background, reflections in the background, instrument covered by reflections and instrument covered by tissue. A moving background, material about instruments, or overexposed instruments seem to be less common.

To make the data publicly available, it was uploaded on *Synapse*<sup>7</sup>, a website that offers a challenge platform with an easy to use environment for participants and organizers, and is already used by well-know challenges (e.g., DREAM<sup>8</sup>). The structure of the data is shown in Fig. 3.9.

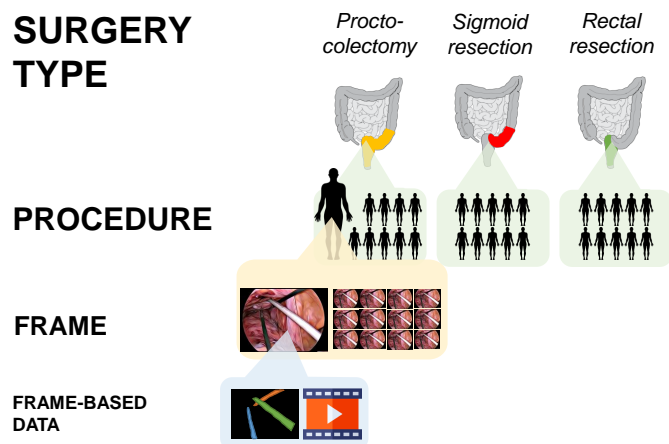


Figure 3.9: This figure shows the structure for the presented dataset. It comprises four levels corresponding to (1) surgery type, (2) procedure number, (4) frame number and (5) frame-based data. Surgery types are e.g., *Proctocolectomy*, every procedure is a different patient, with a frame from the patient video. Frame based data is the annotation to the corresponding video frame and a short video clip that shows the 10s before.

<sup>7</sup><https://www.synapse.org/#!/Synapse:syn18779624>

<sup>8</sup><http://dreamchallenges.org/>

Table 3.4: Number of annotated data per surgery type and number of visible instrument instances.

Surgery type	Number of videos	Number of annotated frames	Number of frames with n instrument instances							
			0	1	2	3	4	5	6	7
Proctocolectomy	10	3,493	450	1,697	1,063	227	54	2	0	0
Rectal surgery	10	3,667	714	1,850	917	158	21	7	0	0
Sigmoid surgery	10	2,880	650	1,198	827	178	24	2	0	1
TOTAL	30	10,040	1,814	3,048	2,807	563	99	11	0	1

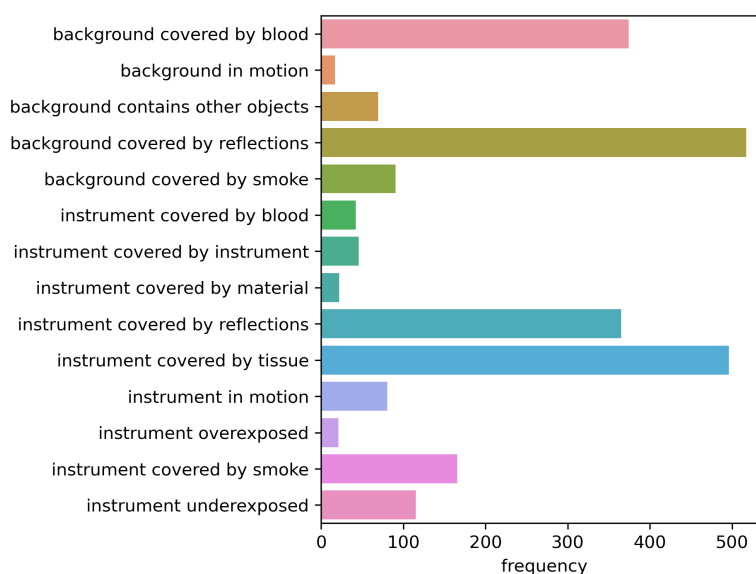


Figure 3.10: This figure shows the appearance of the image characteristic in the complete dataset. The errorbars

### 3.2.4 Discussion

The provided dataset is the first that contains more than 10,000 annotated images with multi-instance instrument segmentation in combination with the image characteristics. This was the first time so many annotators were part of creating such a dataset with such an exact defined annotation protocol and quality control.

Some points should be mentioned that might have an impact on the representativeness of the dataset. While many images were randomly sampled, a considerable part of the dataset was sampled around surgical phase transitions. However, these surgical phases, were only annotated by one surgeon. Another potential bias could come from the corporate semi-automatic pre-segmentation of all images, which were done by the segmentation company and then later revised by the annotators. As a deep learning model was used from *Understand AI* to generate their pre-segmentations, it might be that models trained on those segmentations are especially

good, when they mimic the behaviour of the *Understand AI* model. Although this possibility might theoretically exist, many images were corrected or even wholly newly segmented, making that possibility unlikely. Besides this, the pre-segmentation was only a binary segmentation; thus, the multi-instance itself is human-made, which also means that every single image was reviewed multiple-times.

Possible quality differences concerning the annotation of the image characteristics should also be noted. While 14 people contributed to the image annotations, the image characteristics were only labeled by a single person and were not subsequently validated. In contrast to the segmentation, the annotation process was entirely controlled, many examples were provided, and the annotator was given explicit training. It can therefore be assumed that the labels generated are consistent.

Last but not least, it should be mentioned that although different types of operations have been recorded, they are at least correlated. This is on one hand because they are very close together, and on the other hand because they partly involve similar steps and treat the same organ.

### **3.2.5 Conclusion**

For the first time a large labeled laparoscopic dataset is provided for the task of multi-instance instrument segmentation, where the annotations were generated according to a predefined labeling protocol. The complete annotation process was quality controlled, and the dataset appears to correspond to the presented image characteristics representative of real scenarios. Due to the data structure, the dataset could be the basis for various tasks (e.g., binary and multi-instance segmentation). According to different surgery types, it could enable the testing of generalization capabilities of algorithms. As the raw videos are still available, video sequences could be used in addition to raw image frames to improve instrument segmentation algorithms' quality. Since there is no dataset of this type yet, it could be used as the first benchmarking dataset for multi-instance and binary instrument segmentation.



## 3.3 | Comparative validation of multi-instance instrument segmentation

### Disclosures to this work:

Lena Maier-Hein supervised this work and was, along with Annika Reinke, part of the development of the methodology and involved in the writing process. Parts of this work are accepted for publication in the journal of *Medical Image Analysis* [Roß et al., 2020]. The detailed disclosure of this section can be found in Sec. 7.

### 3.3.1 Introduction

In Sec. 3.1, one assumption was that the provision of an extensive dataset could improve the performance of the binary instrument segmentation algorithms, which was the motivation to generate such a dataset in the previous section (Sec. 3.2). As already written in the Conclusion (Sec. 3.2.5), this dataset now offers two opportunities to advance the development of instrument segmentation algorithms:

The first opportunity is that the data enables the first time to train a model on multi-instance segmentation of surgical instruments and quantify its performance. Furthermore, the possibility of this performance quantification facilitates the second opportunity, namely benchmarking. Unfortunately, researchers often tend to evaluate their methods using private datasets when they publish their work. As a result, the lack of shared data makes it difficult for other researchers to judge (1) which algorithm is the best and (2) which algorithm is the best to solve their task [Ioannidis, 2005, Armstrong et al., 2009]. For this reason, researchers regularly have to re-implement the most modern algorithms to perform a benchmarking, which is prone to mistakes, as they may accidentally include an error in the implementation or have failed to select the correct hyperparameters (e.g., due to a lack of coordination or incorrect assumptions). As those difficulties are a major problem, there are now open challenges (e.g., EndoVis 2017 Robotic instrument segmentation challenge [Allan et al., 2019], Brain Tumour Image Segmentation (BraTS) challenge [Simpson et al., 2019]) with clearly defined rules and guidelines, training, testing datasets, and metrics. Based on those metrics, such competition aims to find the best working algorithm by generating rankings and enabling a fair and interpretable comparison [Maier-Hein et al., 2018, Maier-Hein et al., 2019].

Therefore, the dataset should be made available to the community in the form of a challenge to determine the current state of the art of binary and multi-instance segmentation algorithms and to provide a basis for the development of further methods in this thesis. The challenge was named Robust Medical Instrument Segmentation challenge 2019, and was part of the 4th edition of the *Endoscopic Vision Challenge*<sup>9</sup> and was held at the *Medical Image Computing and Computer Assisted Interventions (MICCAI)* conference 2019 [Roß et al., 2020]. The challenge's

<sup>9</sup><https://endovis.grand-challenge.org/>

defined purpose was to rank algorithms' ability to accurately and robustly perform the binary segmentation and multi-instance segmentation of surgical instruments. As already mentioned in Sec. 3.1, a special focus was put on the algorithms' ability to generalize.

The following sections consist of the Methods (Sec. 3.3.2), where first the challenge is described (organization, training/test set definition, metrics, rankings), followed Results in Sec. 3.3.3, where the outcome of the challenge is presented. In Sec. 3.3.4 the outcome of the challenge will be discussed and the Conclusions presented in Sec. 3.3.5.

### 3.3.2 Methods

#### Challenge organization

The challenge was held, as written above, as part of the Endoscopic Vision Challenge at the Medical Image Computing and Computer Assisted Interventions (MICCAI) conference 2019. The training data was provided via the Synapse<sup>10</sup> platform. Training data was released on August 5th, 2019. To avoid the risk of cheating, test data was not released to the participants. Instead, participants submitted their solutions in the form of a docker until September 15th, 2019 to enable an automatized testing on V100 server, sponsored by NVIDIA GmbH.

For maximum transparency, detailed submission instructions and a challenge design were available for all participants. All challenge results were presented at the MICCAI conference in the form of a publication [Roß et al., 2020] and are available online<sup>11</sup>.

#### Training and test set definition

The aim of the challenge was to test the ability of the algorithms to generalize for different procedures. In other words, all methods should be able to be trained on procedure A, but still be usable on procedure B. The following test scheme was used to test this generalizability [Roß et al., 2020, Maier-Hein et al., 2020b]:

- Stage 1 (highest similarity): Test data was taken from the procedures (patients) from which the training data were extracted.
- Stage 2 (intermediate similarity): Test data was taken from the exact same type of surgery as the training data but from procedures (patients) not included in the training.
- Stage 3 (lowest similarity): Test data was taken from a different but similar type of surgery (and different patients) compared to the training data.

The stages were implemented by reserving the complete *sigmoid surgery* dataset for testing. Of the remaining data, the two shortest videos per procedure were used for stage II. From the remaining training data, always the first annotated video frame was used for stage I. All other frames were available for training (see Fig. 3.11). Despite other challenges, no validation dataset was explicitly defined by the challenge organizers and was thus up to the challenge participants to decide on a splitting strategy for their hyperparameter tuning. The overall amount of available data is illustrated in Tab. 3.5.

---

<sup>10</sup><https://www.synapse.org/>

<sup>11</sup><https://www.synapse.org/#!/Synapse:syn18779624>

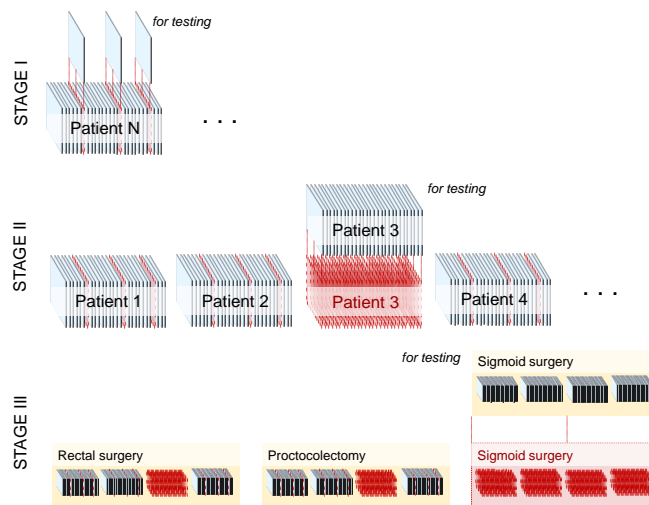


Figure 3.11: Illustration of the three stages setup of the *Robust Medical Instrument Segmentation Challenge 2019*. For Stage I, image frames of the same procedure and the same patient were kept out for testing. In Stage II, all frames that belong to a specific patient were used for training and in Stage III, a complete surgery type was held out as the test set.

Table 3.5: Case distribution of the data with frames per stage and surgery.

PROCEDURE	TRAINING	TESTING		
		Stage 1	Stage 2	Stage 3
proctocolectomy	2,943	325	225	-
rectal resection	3,040	338	289	-
sigmoid resection*	-	-	-	2,880
<b>TOTAL</b>	<b>5,983</b>	<b>663</b>	<b>514</b>	<b>2,880</b>

\*unknown surgery

## Metrics

As the goal of a challenge is a fair comparison between possible solutions for a given problem, it requires the quantifiable definition of a good solution. This quantification was performed with two metrics, first, the Sørensen Dice Similarity Coefficient [Dice, 1945], which is a commonly used metric in the domain of medical image computing that compares the overlap of two areas (see Eq. 2.22). Secondly, to reflect imperfect reference annotations, the *Normalized Surface Dice* [Nikolov et al., 2018] was applied. Although there are several metrics available (e.g., IoU (see Eq. 2.30), the DSC and the NSD were chosen because they are complementary measures. In comparison to the DSC, the Normalized Surface Dice (NSD) measures the overlap of two surfaces instead of the overlap of areas and includes an inter-rater variability  $\tau$  of the annotators [Nikolov et al., 2018]. As the determination of the inter-variability  $\tau$  is expensive, it is separately described in the next section. For the NSD, the surfaces (mask boundary) of  $S_Y$  and  $S_{\hat{Y}}$  are used.

The DSC and the NSD are metrics that can compare binary masks only, and for the multi-instance segmentation task there might be multiple instances in an image where even the number of instances differ to the given reference. For this reason, the references and the prediction have to be matched with the Hungarian algorithm [Kuhn, 1955]. The Hungarian algorithm was applied by minimizing the DSC value between matched pairs. Finally, the DSC for multiple instances ( $MI\_DSC$ ) [Xu et al., 2017] and the NSD for multiple instances ( $MI\_NSD$ ) is the mean of all DSC/NSD values per image.

### Inter-rater agreement

As written in the previous section, the heart of every challenge is the choice of the metrics and their aggregation in the form of a ranking that helps identify the best-suited algorithm for a given problem. However, all metrics have in common that they compare a prediction (provided by the algorithms) with the manual annotators' reference. To interpret the corresponding results, whether it is a good or a bad prediction, it is essential to understand the upper limit given by the human annotators inter-rater agreement. The inter-rater agreement was estimated by letting the same five annotators that curated the dataset annotate 20 images. All pictures were randomly selected from different surgeries and yielded up to three instrument instances. For the estimation, the resulting references were pairwise compared between the annotators.

The quantification of the inter-rater agreement regarding a "closeness" of two annotations can be achieved with a metric that is either area- or surface-based. Comparing areas is one of the tasks of this challenge, so the same metric DSC was chosen. However, for comparing surfaces, there exists a couple of similar metrics, such as the mean surface distance (MSD) or the residual mean squared distance (RMS) (see Eq. 3.8) or the Hausdorff Distance (HD) [Huttenlocher et al., 1993]. Nonetheless, due to a high Spearman correlation [Spearman, 1904] between the HD, MSD and RMS is high, only the HD metric was used, as it is close to the concept of the NSD [Nikolov et al., 2018].

The HD is defined as the maximum of the euclidean distance  $d(S, S')$  between two contours  $S$  and  $S'$  (see Eq. 3.7). Since one image might contain more than one instrument instance, there is a need to match instances between references and annotations, which was done by applying the Hungarian algorithm [Kuhn, 1955] that minimizes the root mean squared distance between matched pairs (see Eq. 3.8). Should the case arise that annotators have annotated a different number of instruments per image, the missing instances for the DSC are set to 0, and for the HD to the maximum Euclidean distance  $d_{max} = \sqrt{I_{width}^2 + I_{height}^2}$  that can be achieved with the image resolution of  $I_{width} \times I_{height}$ .

$$D(p, S) = \min\{d(x, k) | k \in K\} \quad (3.6)$$

where  $d(s, s')$  denotes the euclidean distance between two contour points.

$$HD(S, S') = \max\{\max\{D(s, S') | s \in S\}, \max\{d(s', S) | s' \in S'\}\} \quad (3.7)$$

where  $S$  and  $S'$  are contours.

$$\text{RMS} = \sqrt{\frac{1}{n_S + n_{S'}} \left( \sum_{p \in S} D(p, S')^2 + \sum_{p' \in S'} D(p', S)^2 \right)} \quad (3.8)$$

where  $S$  and  $S'$  are contours.

### Rankings

Apart from the right choice of metrics, a correct ranking is one of the most important decisions that belongs to the design of a challenge. As presented by Wiesenfarth et al. [Wiesenfarth et al., 2019b] the choice of how to aggregate the quantitative results of all test-cases has a huge impact on outcome of the challenge.

Due to the two different challenge goals (robustness and accuracy), two different ways of aggregating the quantitative results were applied:

1. **Robustness:** To put a special focus on the worst cases, the 5th percentile of all test-cases was used to calculate the robustness rank [Roß et al., 2020].
2. **Accuracy:** For the accuracy, a high performance of the algorithms is crucial, thus for all pairs of algorithms a *Wilcoxon signed rank test* ( $\alpha = 0.05$ ) was used to test for a significant better performance [Roß et al., 2020, Wiesenfarth et al., 2019b].

Each ranking (robustness and accuracy) was computed for the  $MI\_DSC$  and  $MI\_NSD$  respectively, producing 4 different rankings. Missing cases were set to the worst possible value "0" [Roß et al., 2020].

### Expert baseline

Due to the fact that the references are produced by (several) human annotators, only imperfect references (no perfect ground truth) are available. Besides the inter-rater agreement of the annotators and with respect for the interpretation of the results, it is of special interest to identify a plausible upper limit of what are actually good results (compared to the human baseline). For this reason, one additional labeling expert, a medical student with six years of experience in labeling, annotated all images from Stage 2. Inspired by a human vs. algorithms analysis for natural image multi-label classification from [Shankar et al., 2020], two additional experiments were performed. In the first experiment, the performance of the expert was compared with the algorithms by investigating the performance as a function of the number of instruments present in the image. In the second experiment, the performance of the algorithms and the expert were compared with respect to the image characteristics (e.g., the presence of blood, smoke, or reflections).

### 3.3.3 Results

While the ROBUST-MIS Challenge contained three tasks in total (see [Roß et al., 2020]), the result reporting will be focused on the multi-instance segmentation task. Reports on the binary segmentation task will comprise only the performance, but not the ranking. The reason for this

is that the conclusions and rankings of the binary segmentation are similar to the multi-instance segmentation task and only the rankings for the multi-instance segmentation tasks will be used in the further course of the thesis.

In the following sections, first, the results of the inter-rater agreement are presented in paragraph *Inter-rater agreement*, followed by a principal overview of the submitting teams with their architectures in paragraph *Method descriptions of participating teams*. Their results are recorded in paragraph *Performance results of participating teams*, and this is followed by paragraph *Challenge ranking*. Finally, the results of the comparison between algorithms and an expert baseline are presented in the paragraph *Expert baseline*.

### Inter-rater agreement

The inter-rater agreement of all five of the experts who annotated the same images is displayed for the DSC and HD metrics in Fig. 3.12 as a function of the sorted metric values. It can be seen that in more than 75% of the pairwise comparisons, the DSC was above 0.98, while the HD was below 36.1. However, approximately the last 20% of the pairwise comparisons demonstrate a huge disagreement in form of a low DSC and a high HD can be observed. Cases where the inter-rater agreement was bad are displayed in Fig. 3.13.

Expressed as descriptive statistics, a median HD of 12.8 (mean: 89.3, 25-quantile: 7.6, 75-quantile: 36.1, 95-quantile: 170.6) over all comparisons of the surfaces, and a median DSC of 0.96 (mean: 0.88, 25-quantile: 0.91, 75-quantile: 0.98) for comparing the area overlap was observed.

Due to the results of these experiments, the inter-rater variability parameter for calculating the NSD score was set to  $\tau := 13$ .

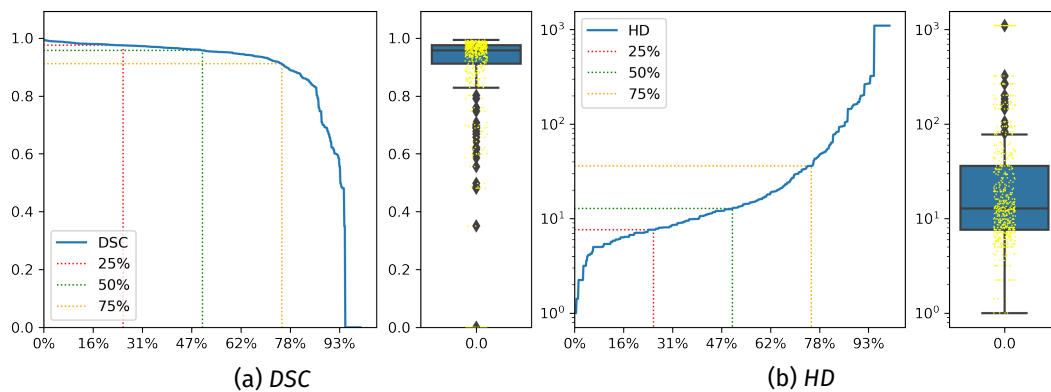


Figure 3.12: Visualization of the sorted DSC and HD metrics as a function of the dataset percentage. It can be seen that in approximately 80% of the data the metrics are close to each other. The strong disagreement among the annotators in the remaining 20% is due to missed or only partially segmented instrument instances.

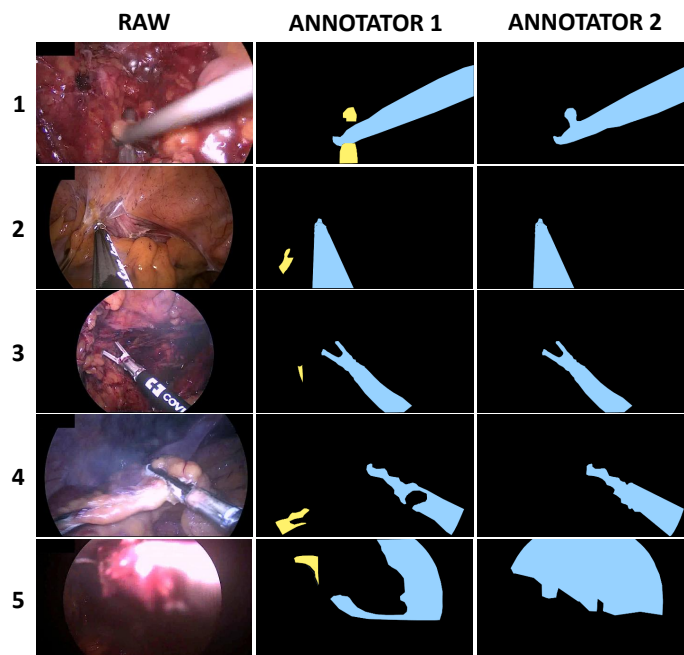


Figure 3.13: Visualization of images with the worst inter-rater agreement. It can be seen that in the first case it is unclear whether the images show an opened instrument or two instruments. Cases 2 and 3 show a small missed instance, case 4 shows an instance in motion and case 5 is an unclear image. It can also be seen that annotator 2 did not follow the annotation guideline exactly in the 4th and 5th cases (excluding tissue on an instrument).

### Method descriptions of participating teams

Prerequisite to participate in the challenge was a working docker submission until September 28, 2019. Out of the 75 registered participants, there were seven valid and one invalid submissions. Together with their submissions, all teams were encouraged to submit a short description of their method that briefly described their intuition and findings when implementing their method. In the following, the methods of all teams are presented and further implementation details are given in Tab. 3.6.

**Team caresyntax: Single network fits all** The *caresyntax* team's core idea was to apply a Mask R-CNN [He et al., 2017] based on a single network with shared convolutional layers for both branches. They hypothesized that the shared layers would increase the generalization capabilities of the network. As their segmentation network, they used a pre-trained version of the Mask R-CNN without including any videos' temporal information. As a result of their experiments, they decided to use a Mask R-CNN, as their approach outperformed a U-Net-based model by a significant margin. The team worked out that tuning pixel-level and mask-level confidence thresholds on the predictions were crucial for the performance. Furthermore, they acknowledged the importance that the training set size had for improved predictions, both qualitatively and quantitatively.

**Team CASIA\_SRL: Dense pyramid attention network for robust medical instrument segmentation** The CASIA\_SRL team proposed a network named Dense Pyramid Attention Network [Ni et al., 2020] for multi-instance segmentation. They noticed two main problems with the instrument segmentation: first, changes in illumination and second, changes of the surgical instrument’s scale. In order to cover the changes in illumination, they used an attention module which captured second-order statistics. This was in order to cover semantic dependencies between pixels and capture the global context [Ni et al., 2020]. As moving surgical instruments change their scale, the team introduced dense connections across scales to capture multi-scale features for surgical instruments. The team did not use the provided videos to complement the information contained in the individual frames.

Table 3.6: Overview of submitted methods of all participating teams.

Team	Basic architecture	Video used?	Additional data used?	Loss functions	Data augmentation
<i>caresyntax</i>	Mask R-CNN [He et al., 2017] (backbone: ResNet-50 [He et al., 2016a])	X	ResNet-50 pre-trained on MS-COCO [Lin et al., 2014a]	Smooth L1 loss, cross entropy loss, binary cross entropy loss	Applied in each epoch: Random flip (horizontally) with probability 0.5
<i>CASIA_SRL</i>	Dence Pyramid Attention Network [Ni et al., 2020] (backbone: ResNet-34 [He et al., 2016a])	X	ResNet-34 backbone pre-trained on ImageNet [Russakovsky et al., 2015]	Hybrid loss: cross entropy $-\alpha \log(Jaccard)$	Data augmented once before training: Random rotation, shifting, flipping
<i>fsensee</i>	2D U-Net [Ronneberger et al., 2015] with residual encoder	X	X	Sum of DSC and cross-entropy loss	Randomly applied on the fly on each batch: Rotation, elastic deformation, scaling, mirroring, Gaussian noise, brightness, contrast, gamma
<i>SQUASH</i>	Mask R-CNN [He et al., 2017] (backbone: ResNet-50 [He et al., 2016a])	✓*	X	ResNet-50: Focal loss, Mask R-CNN: Mask R-CNN loss + cross entropy loss	35% of total input for classification: Gaussian blur, sharpening, gamma contrast enhancement; additional 35% of images: Mirroring (along x- and y-axes); minority class: Translation (horizontally); non-instrument image frames are not processed
<i>Uniandes</i>	Mask R-CNN [He et al., 2017] (backbone: ResNet-101 [He et al., 2016a])	✓**	Pre-trained on MS-COCO [Lin et al., 2014a]	Standard Mask R-CNN loss functions	Applied on the fly on each batch: Random flips (horizontally), propagation of annotation backwards to previous video frames
<i>VIE</i>	Mask R-CNN [He et al., 2017] (backbone: ResNet-50 [He et al., 2016a])	✓***	X	RPN class loss, MASK R-CNN loss	Applied on the fly on each batch: Image resizing (1024x1024), bounding boxes, label generation
<i>www</i> <sup>9</sup>	Mask R-CNN [He et al., 2017] (backbone: ResNet-50 [He et al., 2016a])	X	Pre-trained <sup>9</sup> on ImageNet [Russakovsky et al., 2015]	Smooth L1 loss, focal loss, binary cross entropy loss	Applied on the fly on each batch: Random flip (horizontally and vertically), rotations of $[0, 10]^\circ$

\* to estimate the probability that the last frame of video shows instrument instance

\*\* for data augmentation

\*\*\* calculating the optical flow over 5 frames

**Team fsensee: OR-UNet** Team *fsensee*’s core idea was to optimize a binary segmentation algorithm and then adjust the output with a connected component analysis in order to solve the multi-instance segmentation and detection tasks [Isensee and Maier-Hein, 2020]. Inspired by the



recent successes of the nnU-Net [Isensee et al., 2018], the authors used a simple established baseline architecture (the U-Net [Ronneberger et al., 2015]) and iteratively improved the segmentation results through hyperparameter tuning. The method, referred to as optimized robust residual 2D U-Net (OR-UNet), was trained with the sum of DSC and cross-entropy loss and a multi-scale loss. During training, extensive data augmentation was used to increase robustness. For the final prediction, they used an ensemble of eight models. They hypothesized that ensembles perform better than a single network. In their report, the team wrote that they attempted to use the temporal information by stacking previous frames but did not observe a performance gain. Additionally, they noticed that in many cases instruments did not touch; thus, they used a connected component analysis [Shapiro, 1996] to separate instrument instances.

**Team SQUASH: An ensemble of models, combining image frame classification and multi-instance segmentation** Team SQUASH's hypothesis was that they could increase the robustness and generalizability of all challenge tasks simultaneously by using multiple recognition task training. In training their method from scratch, they assumed that the network capabilities were fully utilized to learn detailed instrument features. Based on a ResNet50 [He et al., 2016a], the team used the video data provided and built a classification model to predict all instrument frames in a sequence of video frames. On top of this classification model, they built a segmentation model by employing a Mask R-CNN [He et al., 2017] to detect multiple instrument instances in the image frames. The segmentation model was trained by leveraging the preliminary trained classification model on instrument images as a feature extractor to deepen the learning of instrument segmentation. Both models were combined in a two-stage framework to process a sequence of video frames. The team reported that their method had trouble dealing with instrument occlusions, but on the other hand, they were surprised to find that it handled reflections and black borders well.

**Team Uniandes: Instance-based instrument segmentation with temporal information** Team Uniandes based their multi-instance segmentation approach on the Mask R-CNN [He et al., 2017]. For training purposes, they created an experimental framework with a training and validation split and supplementary metrics to identify the best version of their method and gain insight into the performance and limitations. Data augmentation was performed by calculating the optical flow with a pre-trained FlowNet2 [Ilg et al., 2017a] and using the flow to map the reference annotation on to the previous frames. However, they did not find significant benefits in using the augmentation technique. The team observed that their approach was limited in finding all instruments in an image frame, but once an instrument was found it was segmented with a high DSC score. Although the team achieved good metric scores, they stated that they fell short in segmenting small or partial instruments and instruments covered by smoke.

**Team VIE: Optical flow-based instrument detection and segmentation** The VIE team approached the multi-instance segmentation task with an optical flow-based method. The teams hypothesis was that the detection of moving parts in the image enables medical instruments to be detected and segmented. For their approach, they calculated the optical flow over the last five frames of a case by using the OpenCV<sup>12</sup> library and concatenated the optical flow with the raw image as input for a Mask R-CNN [He et al., 2017]. The team assumed that this would reduce most of the unnecessary clutter segmentation. The team hypothesized that the temporal data could have been used more effectively.

---

<sup>12</sup><https://opencv.org/>

**Team *www*: Integration of Mask R-CNN and DAC block** Team *www* proposed a framework based on Mask R-CNN [He et al., 2017] to handle the three tasks in the challenge. Based on the observation that the instruments have variable sizes, their idea was to enlarge the receptive field and tune the Mask R-CNNs anchor size. Besides this, the team integrated DAC blocks [Gu et al., 2019] into the framework to collect more information. The team reported that including temporal information might have helped to improve their performance.<sup>13</sup>

### Performance results of participating teams

Fig. 3.14 shows a comparison for the multi-instance segmentation task regarding the  $MI\_DSC$  performances of the participating algorithms over the three evaluation stages. A clear performance drop is visible with the increasing difficulty of the stages, combined with a greater variance in their distribution. Further descriptive statistics for binary and the multi-instance segmentation task are provided in Tab. 3.7. The performance score for each test image of all participating teams for Stage 3 is presented in Fig. 3.14, where every dot represents a test image and each dot is color coded, depending on the number of visible instruments in the picture. The figure shows clusters that correspond to the performance with respect to the visible number of instruments in an image. While high metric values mainly correspond to only one visible instance, images with more than one instance group multiple clusters around lower scores.

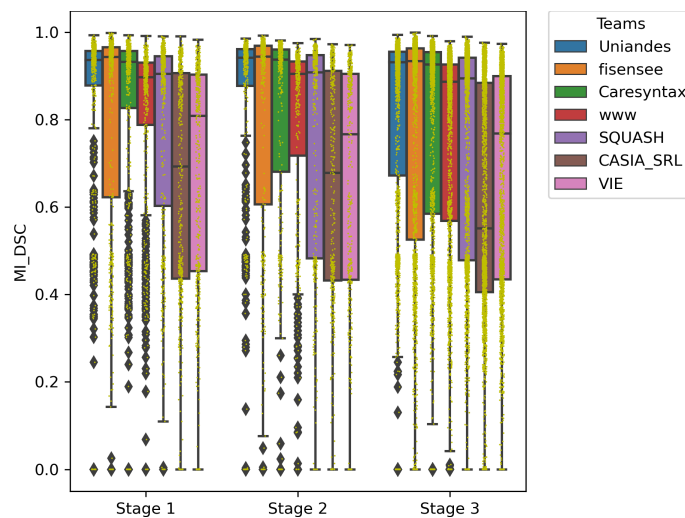


Figure 3.14: The change of the performance scores of all participating teams' overall stages is presented in this figure. It can be seen that with increasing difficulty from stage 1 to stage 3, the performance drops and the variance increases.

<sup>13</sup>Please note that this team used data from the EndoVis 2017 challenge [Allan et al., 2019] to visually check their performance on a different medical dataset. The participation policies (see [Roß et al., 2020]) prohibit the use of other medical data for algorithm training or hyperparameter tuning. The challenge organizers defined this case as a grey zone but noted that the team may have had a competitive advantage in performance generalization.

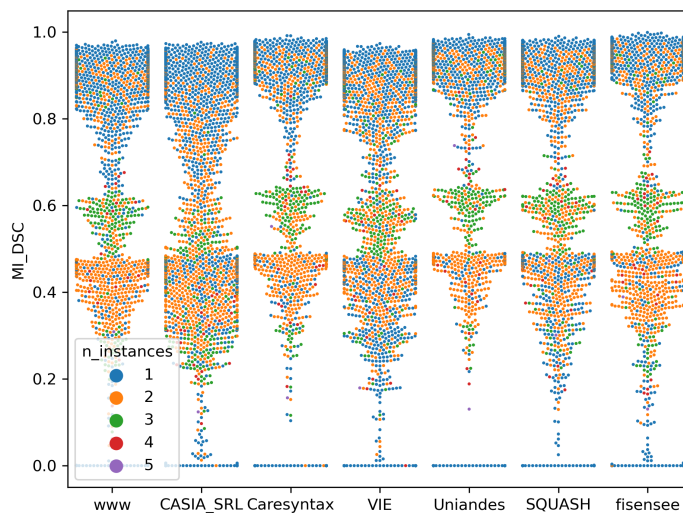


Figure 3.15: The performance score for each test image of all participating teams' for stage 3 is presented in this swarm plot. Every dot represents a test image, where the number of visible instruments is color encoded. It can be seen that there are clusters present in the distribution that correspond to the performance with respect to the visible number of instruments in an image. While high metric values mainly correspond to only one visible instance, images with more than one instance group multiple clusters around lower scores.

### Challenge ranking

As described in Section *Metrics and rankings*, two different rankings were used to compare the algorithms regarding their accuracy and robustness. The rankings were performed for both metrics  $MI\_DSC$  and  $MI\_NSD$  individually and on data from stage 3 (maximal difficulty) only. All results of the rankings are presented in Tab. 3.8. To provide deeper insight into the ranking variability, ranking blob plots (see Fig. 3.16) were used to visualize ranking stability based on bootstrap sampling [Wiesenfarth et al., 2019b].

### Expert baseline

To compare the expert with the algorithms' performances, only the  $MI\_DSC$  metrics are presented, as the results with the  $MI\_NSD$  metric are similar. Given the fact that in Stage 2 only a maximum of three instrument instances are visible, the analysis covers only differences for  $n$  instances where  $n \in \{1, 2, 3\}$ . The results are that Team *fisensee* shares the first rank with the expert in the accuracy rankings for frames where only one instrument is visible. However, as it can be seen in Fig. 3.17, the mean performance of all algorithms drops if more than one instrument is visible, while the experts' performance stays constant.

### 3.3.4 Discussion

This part of the thesis presented a challenge in the field of SDS which, in addition to the two tasks of binary and multi-instance segmentation, focused on two additional aspects, namely the robustness and generalizability of algorithms. The next sections will contain a detailed

Table 3.7: Quantitative results of all participating methods for all three stages for the binary and multi-instance segmentation tasks. The evaluation metric for the binary task was the *DSC* and for the multi-instance segmentation task the *MI\_DSC* metric. The table contains the mean, median and the 5th (Q05), 25th (Q25) and 75th (Q75) quantile.

	Metric	CASIA_SRL	caresyntax	SQUASH	Uniandes	VIE	fisensee	www	expert
<b>BINARY INSTANCE SEGMENTATION</b>									
<b>STAGE 1</b>	Mean	0.90	0.89	0.88	0.90	0.79	<b>0.92</b>	0.88	-
	Median	0.95	0.94	0.93	0.94	0.87	<b>0.96</b>	0.92	-
	Q5	0.70	0.69	0.55	0.71	0.30	<b>0.76</b>	0.68	-
	Q25	0.91	0.91	0.88	0.91	0.76	<b>0.93</b>	0.88	-
	Q75	0.96	0.96	0.95	0.96	0.92	<b>0.97</b>	0.94	-
<b>STAGE 2</b>	Mean	0.89	0.88	0.85	0.89	0.77	<b>0.90</b>	0.86	<b>0.91</b>
	Median	0.95	0.95	0.93	0.95	0.87	<b>0.96</b>	0.92	<b>0.96</b>
	Q5	0.43	0.36	0.34	0.41	0.00	<b>0.54</b>	0.37	<b>0.73</b>
	Q25	0.91	0.91	0.87	0.92	0.74	<b>0.93</b>	0.88	<b>0.93</b>
	Q75	0.97	0.96	0.95	0.96	0.91	<b>0.97</b>	0.94	<b>0.97</b>
<b>STAGE 3</b>	Mean	<b>0.88</b>	0.85	0.83	0.87	0.76	<b>0.88</b>	0.85	-
	Median	0.94	0.94	0.92	0.94	0.86	<b>0.95</b>	0.91	-
	Q5	0.50	0.00	0.22	0.28	0.00	0.34	<b>0.52</b>	-
	Q25	0.89	0.89	0.85	0.90	0.73	<b>0.91</b>	0.86	-
	Q75	0.96	0.96	0.95	0.96	0.91	<b>0.97</b>	0.94	-
<b>MULTI-INSTANCE SEGMENTATION</b>									
<b>STAGE 1</b>	Mean	0.65	0.82	0.78	<b>0.84</b>	0.67	0.80	0.81	-
	Median	0.69	0.93	0.90	<b>0.94</b>	0.81	<b>0.94</b>	0.90	-
	Q5	0.24	0.32	0.32	<b>0.40</b>	0.16	0.32	0.37	-
	Q25	0.44	0.83	0.60	<b>0.88</b>	0.45	0.62	0.79	-
	Q75	0.91	0.96	0.94	0.96	0.90	<b>0.97</b>	0.94	-
<b>STAGE 2</b>	Mean	0.64	<b>0.80</b>	0.75	0.84	0.65	<b>0.80</b>	<b>0.78</b>	<b>0.88</b>
	Median	0.68	<b>0.94</b>	0.91	<b>0.94</b>	0.77	<b>0.94</b>	0.91	<b>0.95</b>
	Q5	0.18	0.32	0.26	<b>0.39</b>	0.00	0.28	0.30	<b>0.47</b>
	Q25	0.43	0.68	0.48	<b>0.88</b>	0.43	0.61	0.63	<b>0.91</b>
	Q75	0.91	0.96	0.95	0.96	0.90	<b>0.97</b>	0.94	<b>0.97</b>
<b>STAGE 3</b>	Mean	0.60	0.77	0.73	<b>0.80</b>	0.65	0.76	0.76	-
	Median	0.55	<b>0.93</b>	0.89	<b>0.93</b>	0.77	<b>0.93</b>	0.89	-
	Q5	0.19	0.00	0.22	0.26	0.00	0.17	<b>0.31</b>	-
	Q25	0.41	0.58	0.48	<b>0.67</b>	0.43	0.52	0.58	-
	Q75	0.88	0.95	0.94	0.95	0.90	<b>0.96</b>	0.93	-

discussion of the challenge design and infrastructure, the challenge outcome, and finally the comparison to the human expert.

### Challenge design

**Challenge infrastructure** While recent challenges in the domain of biomedical imaging are publishing the test data and allowing the submission of the results (e.g. BraTS<sup>14</sup>, KiTS2019<sup>15</sup>, PAIP 2019<sup>16</sup>), the submission of docker containers is already quite common in other disciplines (e.g. CARLA<sup>17</sup>). Suppose Docker is used for the challenge instead of releasing the test data. In that case, there are a few decisive advantages, such as the prevention of fraud (manual labeling) or the introduction of additional bias (assessment of the test results by the participants) [Reinke et al., 2018]. However, it might have prevented participants from submitting a solution, as preparing a

<sup>14</sup><http://braintumorsegmentation.org/>

<sup>15</sup><https://kits19.grand-challenge.org/rules/>

<sup>16</sup><https://paip2019.grand-challenge.org/>

<sup>17</sup><https://carlachallenge.org/>

Table 3.8: Multi-instance segmentation: Rankings for stage 3 of the challenge. The upper part of the table shows the multi-instance Dice Similarity Coefficient ( $MI\_DSC$ ) rankings and the lower part shows the multi-instance Normalized Surface Distance ( $MI\_NSD$ ) rankings (accuracy rankings on the left, robustness rankings on the right). Each ranking contains the team identifier, and either a proportion of significant tests divided by the number of algorithms (prop. sign.) for the accuracy ranking or an aggregated  $MI\_DSC/MI\_NSD$  value (aggr.  $MI\_DSC/MI\_NSD$  value) and a rank.

<b><math>MI\_DSC</math>: ACCURACY RANKING</b>			<b><math>MI\_DSC</math>: ROBUSTNESS RANKING</b>		
<b>Team identifier</b>	<b>Prop. Sign.</b>	<b>Rank</b>	<b>Team identifier</b>	<b>Aggr. <math>MI\_DSC</math> Value</b>	<b>Rank</b>
<i>fisensee</i>	1.00	1	<i>www<sup>9</sup></i>	0.31	1
<i>Uniandes</i>	0.83	2	<i>Uniandes</i>	0.26	2
<i>caresyntax</i>	0.67	3	<i>SQUASH</i>	0.22	3
<i>SQUASH</i>	0.33	4	<i>CASIA_SRL</i>	0.19	4
<i>www<sup>9</sup></i>	0.33	4	<i>fisensee</i>	0.17	5
<i>VIE</i>	0.17	6	<i>caresyntax</i>	0.00	6
<i>CASIA_SRL</i>	0.00	7	<i>VIE</i>	0.00	6

<b><math>MI\_NSD</math>: ACCURACY RANKING</b>			<b><math>MI\_NSD</math>: ROBUSTNESS RANKING</b>		
<b>Team identifier</b>	<b>Prop. Sign.</b>	<b>Rank</b>	<b>Team identifier</b>	<b>Aggr. <math>MI\_NSD</math> Value</b>	<b>Rank</b>
<i>Uniandes</i>	1.00	1	<i>www<sup>9</sup></i>	0.35	1
<i>caresyntax</i>	0.67	2	<i>Uniandes</i>	0.29	2
<i>fisensee</i>	0.50	3	<i>CASIA_SRL</i>	0.27	3
<i>www<sup>9</sup></i>	0.50	3	<i>SQUASH</i>	0.26	4
<i>SQUASH</i>	0.33	5	<i>fisensee</i>	0.16	5
<i>VIE</i>	0.17	6	<i>caresyntax</i>	0.00	6
<i>CASIA_SRL</i>	0.00	7	<i>VIE</i>	0.00	6

docker is an additional overhead. Using docker containers requires extra time and organization effort for an organizer, as sufficient hardware resources for the challenge evaluation are needed and challenge participants request support.

**Metrics and Ranking** The choice of the two segmentation metrics DSC and NSD was made by following the recommendations of the Medical Segmentation Decathlon [Cardoso, 2018]. However, both metrics could have been calculated either globally or image-based. While using a global metric would not require aggregation of the scores per image (e.g., per mean or median), the image-based approach has the benefit that instruments with a small area have the same impact as big instruments. Using the global score would lead to a segmentation bias toward bigger instruments.

While the accuracy ranking appears to be quite stable, the robustness ranking shows a higher instability. These instabilities are probably due to the fact that the robustness was defined as the 5th percentile and is therefore already subject to more variability in comparison to aggregated values. Thus, the robustness ranking is subject to greater uncertainty compared to that of the accuracy ranking, as shown in Fig. 3.16.

**Challenge data** All images were pre-processed by downsampling the video images. This pre-processing might have had an impact on the results. However, based on the submitted methods many teams downsampled the images even further to fit more data on the GPU. As the conditions

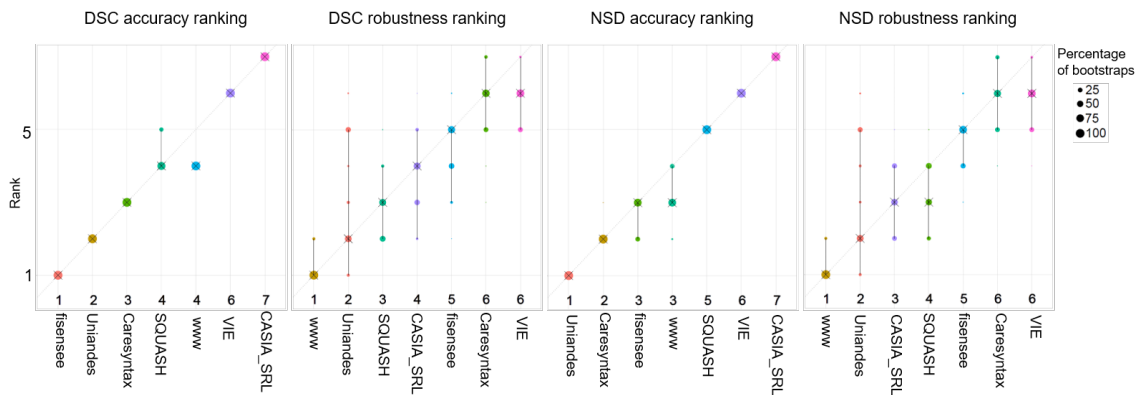


Figure 3.16: The ranking stability (based on a bootstrap sampling) is displayed in form of blob plots, where algorithms are color-coded, and the area of each blob at position  $(A_i, \text{rank } j)$  is proportional to the relative frequency  $A_i$  of the achieved rank  $j$  across  $b = 1000$  bootstrap samples. The median rank for each algorithm is given by a black cross. 95% bootstrap intervals across bootstrap samples are indicated by black lines. The plots were generated using the package challengerR [Wiesenfarth et al., 2019b, Wiesenfarth et al., 2019a].

were the same for all participants, and based on the observation that teams even used smaller images, it can be assumed that any effect that may exist is likely to be rather small.

### Inter-rater agreement

The inter-rater agreement results, as shown in Fig. 3.12, reveal a considerable performance drop after approximately 80% of the data, where the performance quickly drops from a  $DSC > 0.8$  to zero. However, this strong disagreement among the annotators is due to (1) missed or only one partially segmented instrument instance (2) very complex and difficult to judge images or (3) multiple plausible answers (see Fig. 3.13). Smaller deviations could also result from the fact that the annotation rules were not carefully followed in all cases. As data points with a strong disagreement reflect more difficulties in finding an instrument, or very complex images, rather than how close annotators are when segmentation of an instrument instance, excluding them could be a valid option. Nonetheless, the real data contain problems with finding instances, and thus, should remain in the setting. Based on the observation that over 75% of the annotations are very close together, it can be assumed that the median reflects the real inter-rater variability better than the mean performance. For this reason, the median of the HD metric was chosen as  $\tau$  for the NSD metric, reflecting the inter-rater uncertainty.

### Challenge outcome

**Methods** Of the methods submitted, all but one were based on Mask R-CNN variations. The method which does not use a mask R-CNN is a combination of a U-Net with a principal component analysis, and was developed by team *fisensee* (see Tab. 3.6). Surprisingly, the combined approach turned out to be a strong baseline in the DSC accuracy ranking. This can be attributed to two main factors. The primary reason is that a huge percentage of the test data only contained one instance (see Fig. 3.18) and is therefore only a binary segmentation. Additionally, the same figure

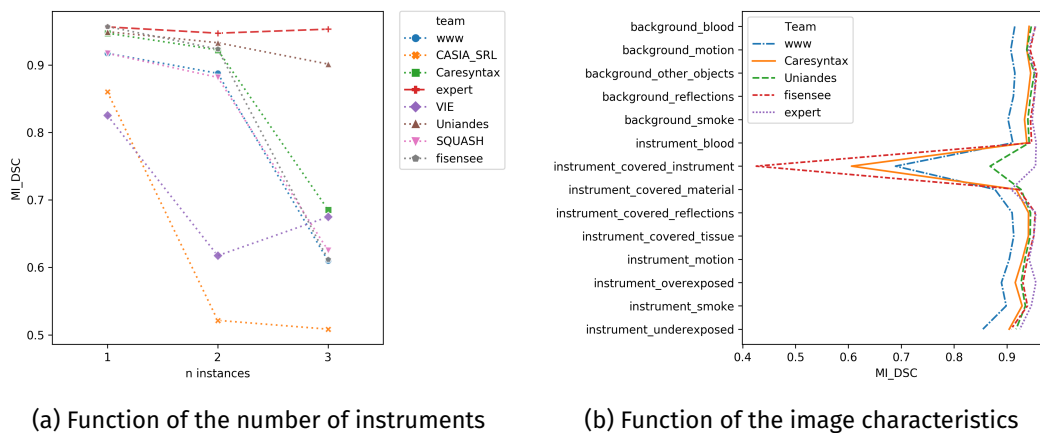


Figure 3.17: This figure shows how a human expert performs in comparison to the submitted methods for the multi-instance segmentation task on Stage 2 data. In the first image (a) the *Median  $MI\_DSC$*  is presented as a function of the number of instruments. Clearly, all algorithms' performance drops with the number of visible instruments in the image while the experts performance stays constant. In the second image (b) the *Median  $MI\_DSC$*  is shown as a function of the image characteristics. It can be seen that the expert systematically outperforms the algorithms.

reveals that the number of instruments that are overlapped by another instrument is only a tiny part of the complete test dataset, which is the reason why the connected components analysis approach is so successful.

Even if the success of the U-Net based approach can be explained, it is difficult to understand which specific design choices for the Mask R-CNN approaches led a submission to be superior to the others. Interestingly, teams reported that during their evaluation settings the choice of their architecture design was more successful, than for example, the U-Net architecture. However, in the final ranking the U-Net outperformed their approach. Based on the summary of all methods in Tab. 3.6, and based on the participants' reports, one can conclude that the factors for success were an extensive data augmentation and a well done hyperparameter search. This is in line with recent findings in the field of radiological data science [Isensee et al., 2018].

Surprisingly, only three of the ten submitted methods incorporate the temporal information that might be extracted from the provided video snippets along with the annotated data. Two of those methods used the temporal information to extract the optical flow between the corresponding frames, and one method used the additional video data for a multi-task setting which additionally trained the likelihood of the instrument presence in a frame. While human annotators often strongly rely on the temporal context, this was not the case for the algorithms. For this reason, it can be speculated that the potential of temporal information still remains undiscovered.

**Results** The key insides for all submitted methods are:

1. Submitted methods: The submitted methods for the binary instrument segmentation task focused mainly on U-Nets [Ronneberger et al., 2015], and Mask R-CNNs [He et al., 2017] for multi-instance segmentation. While the U-Net clearly outperformed all other methods in the binary instrument segmentation, a Mask R-CNN approach yielded better results in terms of robustness for the multi-instance segmentation approach.
2. Performance:
  - (a) Binary segmentation: With a mean *DSC* score of 0.88 for the accuracy task, the performance of the winning algorithm for the accuracy task was similar to that of previous winners of binary segmentation challenges (winner of *EndoVis Instrument Segmentation and Tracking Challenge 2015*<sup>18</sup>: 0.84 (*DSC*); winner of *EndoVis 2017 Robotic Instrument Segmentation Challenge* [Allan et al., 2019]: 0.88 (*DSC*)). Given the high complexity of ROBUST-MIS data in comparison to previously released datasets, the same performance could be attributed to the greater amount of training data. However, in comparison to the self-supervised approach from Sec. 3.1, the performance increased from 0.71 for 23%. As the segmentation in Sec. 3.1 also relied on a U-Net, this gain in performance is due to the greater amount of training data.
  - (b) Multi-instance segmentation: The winning algorithms achieved a mean *MI\_DSC* scores of:
    - 0.82 for cases with one instrument instance,
    - 0.71 for cases with two instrument instances,
    - 0.62 for cases with three instrument instances,
    - 0.45 for cases with more than three instrument instances.
 Thus, the multi-instance segmentation still has a problem with multiple instances.
3. Generalization: The generalization capabilities for all methods was satisfying, with only a loss of 3% in the binary segmentation task and 5% for the multi-instance segmentation task.
4. Robustness: The algorithms showed different performances for different image characteristics. Further research is required.

Beside the key-findings in the beginning of the section, it is important to explicitly discuss the performance drop when multiple instrument instances were visible in one image (see Fig. 3.17). While the worst performance drop could be observed for the U-Net based approach from *fisensee*, the drop of the other teams was less dramatic. However, the extent to which the number of visible instruments impact the results can clearly be seen in Fig. 3.15 with the different clusters for the number of instruments. Though high metric values mainly correspond to only one visible instance, images with more than one instance group multiple clusters around lower scores.

Due to the difficulties when there is more than one instance visible in an image, none of the methods could outperform an expert annotator. Although most of the methods scored a median *MI\_DSC* above 0.9, a comparable performance could be achieved in cases where one instance was visible. Only the team *Uniandes* achieved a median score for two visible instances close to

<sup>18</sup><https://endovissub-instrument.grand-challenge.org/>



the human annotator. Nevertheless, it can be observed that the expert's performance does not depend on the number of visible instances in an image, in contrast to the algorithms. Surprisingly, the expert also achieved low values in the robustness ranking of only 0.43 for one visible instance and 0.47 for two instruments due to missed or false instances. However, when comparing the percentile, for  $MI\_DSC = 0.9$  and  $MI\_DSC = 0.95$ , the expert is on the 14th, and 10th percentile, while *Uniandes* is on the 14th and 48th percentile and *fisensee* on the 14th and 37th percentile.

While the expert's performance is stable across all artifacts, fluctuations in the individual algorithms can be observed (e.g., reflections in the background or an instrument motion). Thus, there are signs that specific image characteristics can positively (e.g., background containing other objects) or negatively (e.g., underexposed instrument) influence the algorithms' performance. However, it is difficult to estimate how much the image characteristics affect the image quality. Estimating the influence is difficult because the number of available images with a certain characteristic differs from each other. Thus, further analyses with better evaluation options are required.

Concerning the interpretation of the values, it must be said that the results are based on the comparison with only one expert. Further, the comparison was exclusively made on Stage 2 of the test data with only one medical expert for practical reasons. Thus, two additional effects could further appear negatively with an increased difficulty if the comparison would have been performed on Stage 3 data, and positively when further experts would have been included in the validation.

Finally, even if an assessment of the inference time would have been of special interest (especially with respect to a possible translation into the clinic), it was not further investigated. This was because a respective metric was not announced to the challenge participants, and thus, the participants did not have the opportunity to optimize their methods. This assumption was also supported by the observation as a preliminary run-time analysis of the docker submissions, which revealed huge differences in the computation times for similar architectures (0.07-7.3 seconds/image).

### 3.3.5 Conclusion

In summary, it can be said that an increased number of training data has led to better performance in binary instrument segmentation, with a quality that is now very satisfactory in the median. The performance of the multi-instance segmentation still has major weaknesses though. Among other concerns, it was found that the presence of several instruments in one image harms the performance of the algorithm. Despite experts that are not influenced by the presence of several instances in their annotation quality, the performance of all currently available methods drops significantly. In addition to the problem of multiple instances, there are also indications of further influences (image characteristics such as blood or smoke) on the algorithms' performance. A systematic analysis of characteristics influence could enable a targeted development of methods that could solve these problems. Overall, it can be said that multi-instance segmentation is not yet suitable for translation into clinical practice.



## 3.4 | Effects of image characteristics on the algorithm performance

### Disclosures to this work:

Lena Maier-Hein supervised this work and was, along with Pierangela Bruno and Manuel Wiesenfarth, part of the development of the methodology and involved in the writing process. The detailed disclosure of this section can be found in Sec. 7.

### 3.4.1 Introduction

In recent years the number of international competitions has increased in the surgical data science community, which can be seen in the dataset summary from Maier-Hein et al. [Maier-Hein et al., 2020a]. The benefits of such challenges are that first, they enable a fair comparison across multiple methods [Maier-Hein et al., 2018, Maier-Hein et al., 2019] and second, they arouse interest in a specific problem.

One of such international challenges is the Endoscopic Vision (EndoVis) Challenge<sup>19</sup>, which takes place on a regular basis at the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI) since 2015 (exception: 2016). The EndoVis challenge itself hosts various sub-challenges with a broad variety of tasks in the field of endoscopic image processing. The outcome of such challenges is usually a website and/or a publication, that mainly contains reports on the methodologies of the participants and descriptive performance statistics and rankings [Allan et al., 2019, Allan et al., 2020, Bodenstedt et al., 2018, Roß et al., 2020]. Some of those publications also contain the reports of best/worst performance cases or they report open challenges, based on a manual analysis, combined with an interpretation (e.g., [Allan et al., 2020, Bodenstedt et al., 2018, Roß et al., 2020]).

In a previous part of this thesis (Sec. 3.3), a comparison was made between an expert annotator and all methods that were submitted for the ROBUST-MIS challenge, with respect for image characteristics (see Fig. 3.17). The analysis revealed that image characteristics might influence the performance of an algorithm on an image. However, due to statistical difficulties, no further exploration of this problem has been performed. A review of the challenge publications to date reveals, a structured and detailed analysis of the performance of algorithms (e.g., open challenges regarding difficult image scenarios such as bleeding, smoke and reflections) has never been systematically and quantitatively performed. Thus, it remains unclear how strong possible open challenges affect the performance of the algorithms, which might be helpful for a problem-driven approach.

This part of the thesis presents an approach to how the results of a challenge can be used for an in-depth image characteristic analysis, where the effect of the characteristics on the algorithms'

<sup>19</sup><https://endovis.grand-challenge.org/>

performance is quantified for the first time.

The outline of this part is as follows: to begin, the analysis itself is further explained in the Methods Sec. 3.4.2. This section is followed by the Results of the analysis (Sec. 3.4.3), a Discussion of the results in Sec. 3.4.4 and the Conclusion is found in Sec. 3.4.5.

### 3.4.2 Methods

Regardless of the use case, there typically exists domain and use case-dependent difficulties that algorithms must deal with (e.g., reflections, smoke, or motion characteristics during instrument segmentation). Often, there are multiple difficulties at once, and it is not always straightforward how to identify and prioritize them during development. Especially in the medical domain, it is crucial to understand how difficulties affect the performance of the algorithms [Esteva et al., 2021]. To develop a good understanding of such characteristics, one might look at the worst cases of the results. It needs to be considered, however, that only reviewing the worst cases could lead to a concentration on exceptional border cases, which contains the potential for the fundamental weaknesses of the algorithms being lost. Thus, an overall examination of possible difficulties in images is of special interest.

This section begins by describing how the influence of image characteristics could statistically be modeled to estimate their effect on the performance. This section is followed by the description of the experimental setup for the effect estimation.

#### Quantification of the effect of image characteristics

The algorithms submitted to the ROBUST-MIS Challenge were used (see Sec. 3.3) to quantify the effects of image characteristics, as the influence on a large number of different algorithms can be tested. The performance of these algorithms, which had been measured using the *DSC* metric should now, in combination with the presence of the image characteristics, be included in a statistical model to estimate the characteristics' effect.

One way to model the effect of image characteristics is to use the challenge metric (*DSC*) to fit a Linear Mixed Effects Model (*LMM*), and later, analyze the resulting coefficients. Nevertheless, as already described in the paragraph 2.1.2 *Generalized Linear Models*, modeling the algorithm performances directly on the challenge metric would violate the normality assumption, as the *DSC* values are in the range of  $[0, 1]$ . Thus, a suitable transformation of such values into a range of  $[-\text{inf}, \text{inf}]$  (e.g., using the logit or the arcsin) would make the values subject to the normal distribution. Unfortunately, such a transformation directly affects coefficients and thus could make their interpretation harder. Instead of using segmentation metrics, an alternative method is to model how many of the predicted instrument pixels are correct. This form of prediction could be done with a model that does not require normally distributed data, a so-called *GLMM* [Breslow and Clayton, 1993]. Such models combine the power of Linear Mixed-Effects Models [McCulloch and Neuhaus, 2001], which incorporate random effects, and generalized linear models capable of handling non-linear data [Bolker et al., 2009].

However, when modeling correct pixels without the *DSC* metric, the overlap of the groundtruth with the segmentation is no longer able to be examined directly. This problem is caused because *GLMM* uses the binomial distribution function and thus models the number of successes

taken from the total number of trials. In other words, it can only be examined (Q1) how many groundtruth pixels are found by the prediction, and (Q2) how many pixels of the prediction are inside the groundtruth. It is important to ask both questions because they contain mutually complementary information. For example, a prediction that covers too many pixels would lead to a high success rate in Q1, but a lower success rate in Q2. Thus, for Q1, the number of trials was the number of pixels in an instrument instance, and for Q2 it was the total number of pixels in the predicted instrument instance. For both, Q1 and Q2, the number of successes were the intersecting instrument pixels between the reference and the predicted instance.

The interpretation of the coefficients of a fitted GLMMs was done with the Odds Ratio (OR) [Bolker et al., 2009, Holling and Schmitz, 2010]. The OR (expressed as  $R(A^T : A^F)$ ) is a metric that measures how two events are associated with each other regarding their presence or absence (see Eq. 3.9 and Tab. 3.9) [Holling and Schmitz, 2010]. Thus, the OR is a measure of the effect strength between two independent binary variables.

Table 3.9: Confusion matrix for two events (State A and Event B) regarding their presence or absence.

	Event A	Event B
Present True	$A^T$	$B^T$
Present False	$A^F$	$B^F$

$$R(A^T : A^F) = \frac{A^T \cdot B^F}{A^F \cdot B^T} \quad (3.9)$$

with  $R(A^T : A^F)$  being the OR for presence or absence  $A^T$  and  $A^F$  of the event  $A$ .

However, the fact that the OR is not symmetrical makes the interpretation less intuitive. This can be illustrated on an OR of 4:6 (0.66) and its inverted ratio of 6:4 (1.5). For this reason, the log of the OR is typically used which makes the ratios symmetrical ( $\log(4 : 6) = -0.176$  and  $\log(6 : 4) = 0.176$ ).

The annotations of the characteristics as described in Sec. 3.2.2 were used in combination with the challenge results, providing the performance of multiple state-of-the-art algorithms. Due to the small number of data points for more than three visible instrument instances and the associated difficulties in fitting a model, only images with  $n \in \{1, 2, 3\}$  instrument instances were analyzed. To avoid modeling difficulties with the mixed amount of visible numbers in an image, for each  $n$ , its own model was fitted resulting in three models.

### Experimental setup

This section describes the experimental setup which was used to validate the developed concept. For this purpose, the used dataset and metrics are first described, followed by the implementation details of the presented deep learning model. The section concludes with an explanation of the details for all experiments.

**Dataset and metrics** The evaluation was based on the challenge dataset as described in Sec. 3.2. Thus, the dataset consisted of 5,983 training and 4,057 test images, where each image came with a 10s video snippet from the previous frames. Following the test and training set definition from [Roß et al., 2020], the test set was split into three stages to explore different grades of the algorithm to generalize to new procedures and patients.

- Stage 1: 663 frames (16 % of the test data) which were not part of the training set extracted from the videos of the same patient from where the training data had been extracted.
- Stage 2: 514 frames (13 %) were extracted from the same type of surgery as the training data but had been taken from patients who had not been included in the training data
- Stage 3: 2,880 frames (71 %) where the data was extracted from a different, but similar, type of surgery

For the validation of the method the rankings and metrics were similar to Sec. 3.3.4. This was specifically in regards to the use of the the multi-instance Dice Similarity Coefficient ( $DSC_{MI}$ , [Xu et al., 2017]) and so the Hungarian algorithm [Kuhn, 1955] was used to assign corresponding pairs [Roß et al., 2020].

**Implementation details** All parameters were set according to small experiments on a held-out validation set during the implementation process.

**Influence of characteristics on the algorithm performance** For analyzing the influence of characteristics on the algorithm performance, the two questions Q1 and Q2 were individually investigated to discern the number of visible instances  $n \in \{1, 2, 3\}$ . Thus, six models were fitted in total (two questions times three visible instances). The separated evaluation was necessary because not all characteristics are the same over all instances (e.g., overlapping instrument instances require at least two instances in the image).

In order to interpret the results of the statistical model, it is essential to further compare the characteristics distribution on both the training and the test set. This comparison is important in order to be able to assess whether an image characteristic is complicated for an algorithm or whether it was never or rarely present in the training set.

### 3.4.3 Results

In summary for Q1 and Q2 it can be said that algorithms have a problem with instrument instances that are underexposed and covered by another instruments. However, it is important to look at both questions separately since they contain mutually complementary information (see Sec. 3.4.2).

The itemized results for the question (Q1 and Q2), how characteristics influence the probability that an algorithm correctly classifies a pixel of the reference as an instrument, can be seen in Fig. 3.19, Fig. 3.20 and Fig. 3.21 and are as follows: If only one instrument is visible in the image, the main characteristics which statistically, significantly negatively (confidence level 3) influence the results, are when the instrument is underexposed or covered by a material. Characteristics such as an instrument covered by smoke, motion blur, or reflections have no significant impact. The same holds true for all possible background characteristics. For two visible instruments,

the harmful characteristics are when an instrument is underexposed and, with a large margin, when it is covered by another instrument. For three visible instruments, a statistically significant adverse impact can only be found for images where the background is covered by blood.

Regarding the impact of characteristics on correct predictions (Q2), the statistically significant negative factors are when an instrument is underexposed, covered by material, is in motion, or when the background contains other objects and is covered by smoke. For two visible instances, again, the biggest negative impact is when an instrument is covered by another instrument. This in addition to other significant characteristics such as an underexposed instrument or an instrument covered by smoke. In comparison to Q1, there are two strong significant negative characteristics when three instruments are visible: when the instrument is in motion or, again, when another instrument covers it.

Besides negative influences, some characteristics appear to have a positive influence in both questions Q1 and Q2. For two and three instances, it appears that the performance of an image might be better when the instruments are overexposed or even when they are covered by tissue.

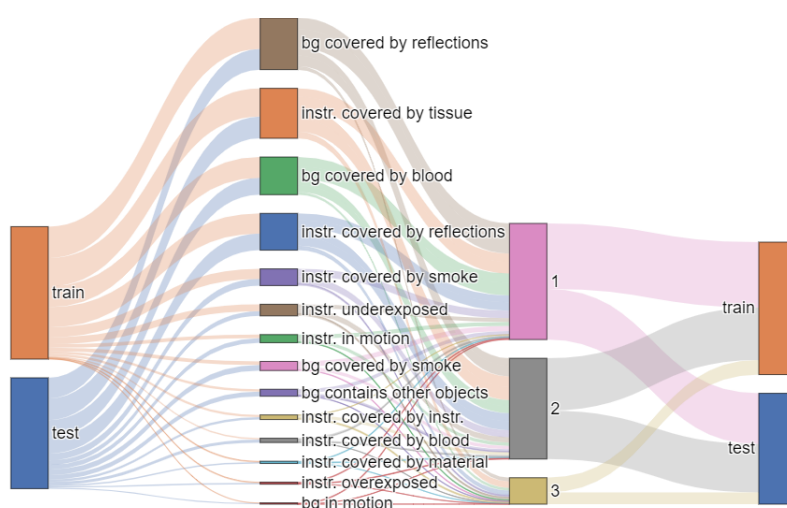


Figure 3.18: Representation of the occurrence of image characteristics and their distribution in the test and training dataset, as well as the occurrence of the characteristics in images with one, two or three visible instruments. While the size of the blocks and connecting lines in the image correspond to the proportion of images that are subject to the written property (e.g., where the background is covered by reflections), the color is only used to improve the legibility of the image. The abbreviations "bg" stand for background and "instr." for instrument.

For the interpretation of the presented results, the distribution of characteristics in the test/training data is a relevant factor and is shown in the Fig. 3.18. The figure shows that (1) the occurrence of the image characteristics is slightly different between the training and the test dataset, (2) the occurrence of characteristics in images with different numbers of visible instruments varies, and (3) the number of visible instruments in the training and test dataset was almost identical.

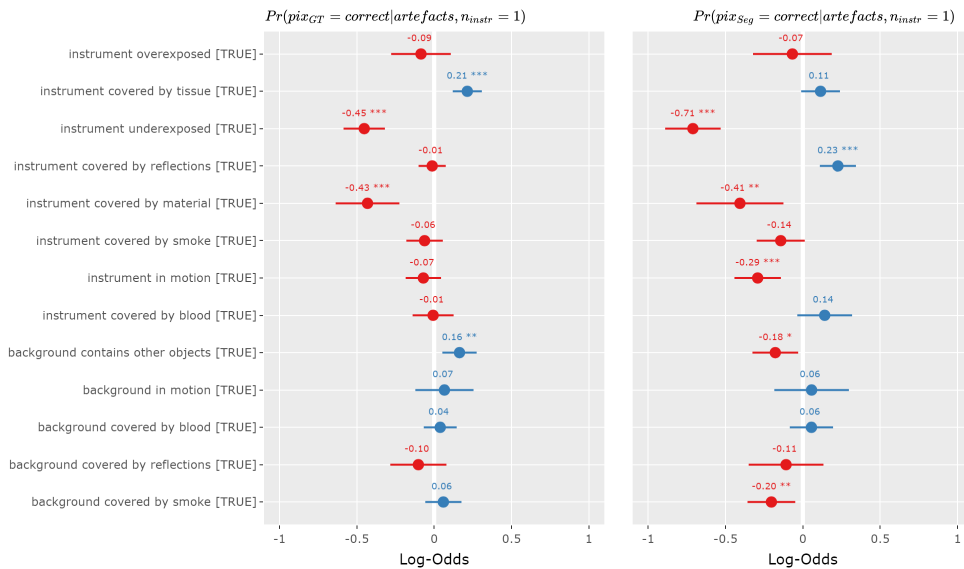


Figure 3.19: Visualization of the effect of image characteristics on the algorithm performance for one instance, interpreted as repeated Bernoulli experiments and the probability reparametrized with a linear mixed-effects model. The characteristics' effect is displayed in the form of the  $\log(\text{odds})$  ratio, representing the odds that an outcome will occur compared to the odds of the outcome occurring in the absence of that exposure. Significant effects are marked with an asterisk \*. Effects were analyzed for the reference and prediction.

### 3.4.4 Discussion

While difficulties such as the number of visible instruments in an image were already identified in the challenge section (Sec. 3.3.3), the more in-depth statistical analysis of the effects of characteristics provided further insights. Contrary to the observations made by, e.g., Bodenstedt et al. [Bodenstedt et al., 2018], reflections, blood, and smoke do not seem to have a statistically significant impact on the algorithm performances. Instead, the main limiting factors that reduced the odds that a groundtruth pixel, or a predicted pixel, correctly being classified occurred when an instrument was in motion, underexposed, or covered by another instrument.

The most significant impact by far was when the instrument is being covered by another instrument. Such an impact can be attributed to the architecture of Mask R-CNNs and their way of processing images. As Mask R-CNNs consist of a region proposal network, they provide bounding boxes around regions of particular interest. Those bounding boxes will then be further processed by the different heads of the Mask R-CNN (e.g., classification, segmentation, scoring) [He et al., 2017]. Especially in regions where instruments overlap, those bounding boxes might contain parts of multiple instrument instances, which then leads to mis-segmentations. Interesting to note are the characteristics that even seem to support outcome of the algorithms (e.g., in some cases, reflections or when an instrument is covered by tissue or when the background is covered by smoke). For the interpretation of such positive effects, it is important to include (1) how often such characteristics appear in general and (2) how often such characteristics were visible in the



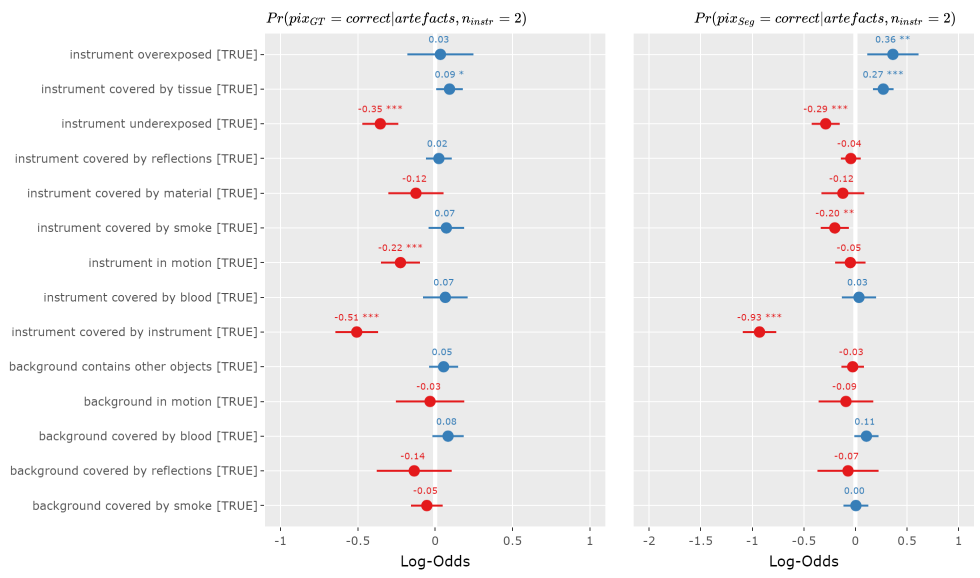


Figure 3.20: Visualization of the impact of image characteristics on the algorithm performance for two instances, interpreted as repeated Bernoulli experiments and the probability reparametrized with a linear mixed-effects model. The characteristics' effect is displayed in the form of the  $\log(odds)$  ratio, representing the odds that an outcome will occur, compared to the odds of the outcome occurring in the absence of that exposure. Significant effects are marked with an asterisk \*. Effects were analyzed for the reference and for the prediction.

training data in comparison to the test data. By comparing how often characteristics, such as blood or smoke, are present in the data, the positive effect of an instrument covered by tissue can be supported. In contrast, the statistical significance of the background covered by smoke is very likely due to the reduced number of presence in the data. Furthermore, a low significance level of only one can be attributed to the high amount of data in general. Thus, for the interpretation, only a higher confidence level should be considered.

### 3.4.5 Conclusion

For the first time, a method was presented that makes it possible to estimate the influence of image characteristics on the segmentation quality of algorithms. It turned out that, contrary to the assumptions of previous challenge publications, image characteristics such as smoke or blood are statistically less complicated than assumed. Although statistically significant characteristics were determined, it seems that there are more methodological difficulties with the algorithms concerning multi-instance segmentation. This knowledge could make it possible to specifically develop algorithms which address these difficulties.

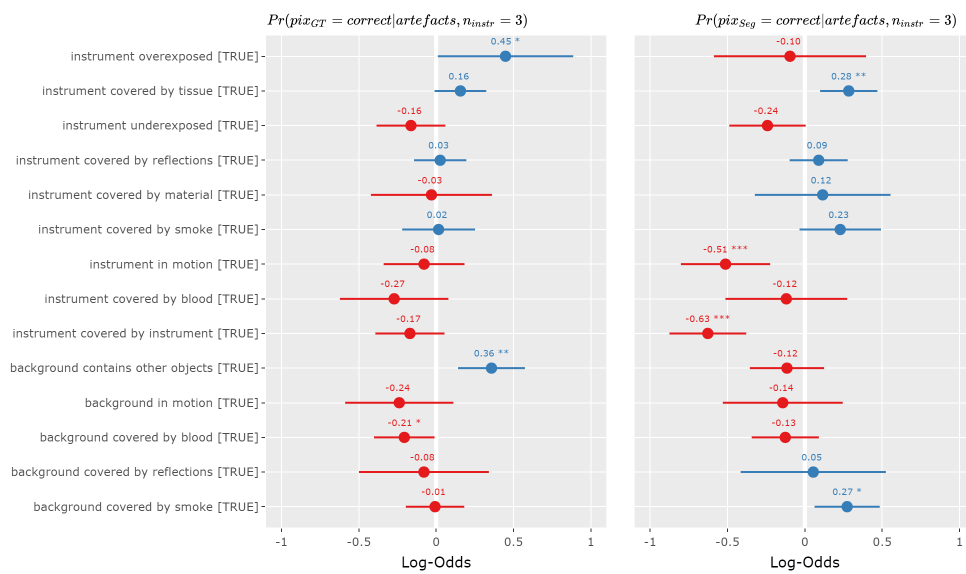


Figure 3.21: Visualization of the impact of image characteristics on the algorithm performance for three instances, interpreted as repeated Bernoulli experiments, and the probability reparametrized with a linear mixed-effects model. The characteristics' effect is displayed in the form of the  $\log(\text{odds})$  ratio, representing the odds that an outcome will occur, compared to the odds of the outcome occurring in the absence of that exposure. Significant effects are marked with an asterisk \*. Effects were analyzed for the reference and for the prediction.

## 3.5 | Multi-instance segmentation of medical instruments

### Disclosures to this work:

Lena Maier-Hein supervised this work and was, along with Pierangela Bruno, Darya Trofimova and Patrick Scholz, part of the development of the methodology and involved in the writing process.

### 3.5.1 Introduction

Over the last decades, the number of minimally invasive surgeries continuously increased and led to a significant reduction in patient hospitalization and recovery time as compared to open surgery [Siddaiah-Subramanya et al., 2017]. Due to the nature of such surgeries, surgeons must rely on imaging the operating field by means of an endoscope. This practice however, comes with limitations such as lack of depth information [Bogdanova et al., 2016] and restricted freedom of movement [Azimian et al., 2010]), which may affect the surgeon’s visual understanding. To overcome such limitations, current research in surgical data science (SDS) [Maier-Hein et al., 2017a] focuses on developing methods and systems capable of improving surgical vision and providing context-aware assistance. To avoid the use of additional, and potentially expensive, hardware in the operating room, many applications are only based on endoscopic videos (e.g., vision-based force estimation [Su et al., 2018], surgical skill assessment [Law et al., 2017, Lin et al., 2019] and augmented reality [Wang et al., 2017, Burström et al., 2019]). As many such tasks rely on accurate segmentation and tracking of medical instruments that are visible in the endoscopic video stream [Wang et al., 2017, Shvets et al., 2018, Kletz et al., 2019], the segmentation algorithm must provide accurate results.

How well such instrument segmentation algorithms currently work has already been analyzed in detail in Sec. 3.3. It was observed that the segmentation quality could depend on the characteristics on an image (e.g., presence of blood or smoke). In Sec. 3.4, a method was then presented that made it possible to identify and quantify the effect of characteristics that harm the segmentation quality.

Based on the findings reported in Sec. 3.4, the research question is whether an algorithm can be developed that explicitly addresses the open challenges. Thus, this part of the thesis presents a problem-driven development approach as to how the results of the in-depth image characteristics analysis can be used to develop an algorithm that explicitly tackles some of the open challenges.

The remainder of this part of the thesis is structured as follows: Sec. 3.5.2 presents the developed concept, followed by the results in Sec. 3.5.3. A discussion of the results is then presented in Sec. 3.5.4 and Sec. 3.5.5 concludes by summarizing the main achievements.

### 3.5.2 Methods

Regardless of the use case, there are domain and use case-dependent difficulties that algorithms must typically deal address (e.g., reflections, smoke, or motion characteristics during instrument segmentation). In Sec. 3.4, an approach was presented that identifies and quantifies the impact of such characteristics on current instrument segmentation algorithms. The characteristics that were identified as having a significantly negative impact on the algorithm performance were smoke or other objects in the background. Further negative effects occurred when the instrument was underexposed, in motion, or covered (e.g. with smoke, material or another instrument). Thus, the new algorithm should be explicitly designed to manage those difficulties.

Based on the experiences that could be gained during the annotation of the dataset, it is noteworthy that most of the characteristics also cause difficulties for the annotators but these problems could be handled better by using temporal information. Given this, inclusion of temporal information in the instrument segmentation process may help to improve the performance. It is of interest to note that the numerous teams which participated in the *ROBUST-MIS challenge* (see Sec. 3.3.3) reported that they could not achieve a benefit from including temporal information, and those teams which had attempted, did not achieve high performance scores (see Sec. 3.3.3 *Performance results of participating teams*). The second placed team (Uniandes) in the accuracy but first placed team in the robustness ranking attempted to use the optical flow between the frames to improve their segmentation and reported no positive effect.

For these reasons, an alternative variant on how the temporal information could be included is presented in this section. The core component of the presented deep learning architecture is a Mask R-CNN [Huang et al., 2019a] that uses (1) the raw video frame, (2) the probability of a pixel to be an instrument and (3) the LSTM-summarized information on object motion encoded in an optical flow. The goal of this approach is to reduce the effect of the presented characteristics for the task of the multi-instance segmentation. An overview of the approach is shown in Fig. 3.22.

The formal definition of the problem concerning the multi-instance instrument segmentation is as follows: Given is a video sequence  $X = \{x_{t-N_X}, x_{t-N_X+1}, \dots, x_t\}$  at a given time step  $t$  that consists of  $N_X$  frames. A frame  $x$  has the dimensions  $[h_x \times w_x \times 3]$ , where  $h_x$  and  $w_x$  are the image height and width, and the three channels in the last dimension contain the RGB encoded color information. To each  $x$  corresponds set of instrument instances  $I_x = \{I_{x,j} | j \in \{1 \dots N_I(x_t)\}\}$ , where  $N_I(x_t)$  denotes the number of instrument instances present in frame  $x$ .

In the next sections, the details of the instrument likelihood prediction are first presented, followed by a description of how the temporal information was used before the instrument instances are predicted. To further tackle some Mask R-CNN related problems, a post-processing step was used, which is described in 3.5.2 *Post-processing*.

#### Prediction of the instrument likelihood

Given is a video frame  $x$  that is taken from a video sequence  $X$ , with  $x \in X$  and  $x$  of the dimensions  $[h_x, w_x, 3]$ . The goal is to predict for each pixel  $p \in x$  the probability that it is an instrument. Prediction is done with a 2D U-Net [Ronneberger et al., 2015] which achieved the best results in the Robust Segmentation challenge [Roß et al., 2020] and that is further described

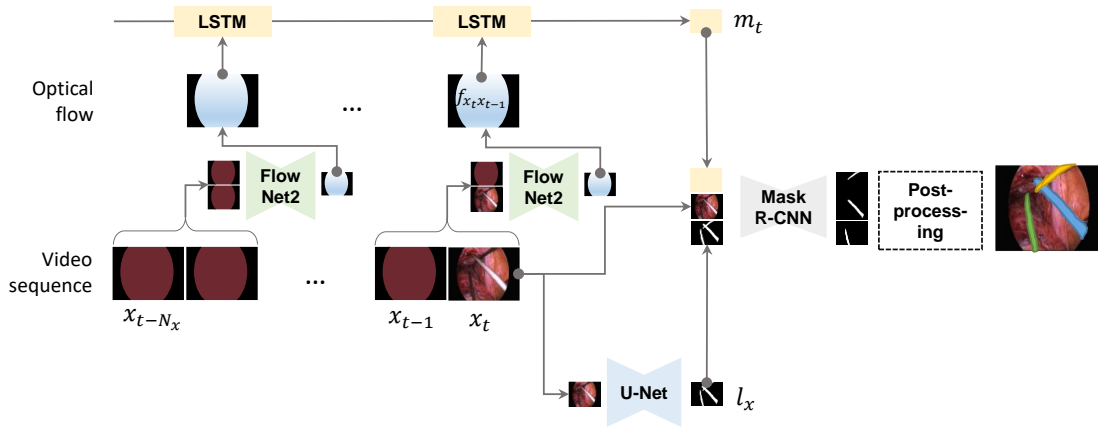


Figure 3.22: This figure illustrates the concept of the multi-instance segmentation approach for a video frame  $x_t$ . At first the optical flow of a video sequence  $X = \{x_{t-N_x}, x_{t-N_x+1}, \dots, x_t\}$  of length  $N_x$  is estimated for a consecutive pair of frames via FlowNet2. Then the estimated optical flow is summarized with a LSTM into  $m_t$ . For every pixel in  $x_t$  the instrument likelihood  $l_x$  is estimated with a U-Net. A Mask R-CNN produces the multi-instance instrument segmentation which gets as input the original image  $x_t$ ,  $l_x$ , and  $m_t$ . The final results of the instrument instances are achieved after post-processing.

in Isensee et al. [Isensee and Maier-Hein, 2020].

The prediction model  $\mathcal{B}(x)$  was trained with the following loss  $\mathcal{L}_{Seg} = \mathcal{L}_{DSC} + \mathcal{L}_{CE}$ , a combination of the regular pixel-wise cross-entropy loss  $\mathcal{L}_{CE}$  (see Eq. 2.21) and the soft dice loss  $\mathcal{L}_{DSC}$  (see Eq. 2.23).

The output of the model  $l_x = \mathcal{B}(x)$  has the dimension  $[h_x, w_x, 1]$  and contains the probability values for all pixels of being an instrument in the range between  $[0, 1]$ .

### Use of the temporal information

For prediction of the segmentation for the frame  $x_t$  at the time step  $t$ , the goal is to include the knowledge of the previous frames. In the first step, the instrument motion is estimated using the optical flow concept. In the second step, this movement is summarized so as to enable straightforward processing in the segmentation task. The following two paragraphs will explain how the optical flow is generated and summarized for later use in the segmentation network.

**Prediction of the optical flow** The computation of the optical flow was traditionally done by either calculating the shift of corresponding feature points in two images [Lucas et al., 1981] or by using the method of Farneback et al. [Farneback, 2003] to get a dense estimation for all image pixels. However, compared to the recent methods based on neural networks, those methods are slower and less accurate [Dosovitskiy et al., 2015, Ilg et al., 2017b]. For this reason, the optical flow estimation was done with the FlowNet2 [Ilg et al., 2017b], an extension of the FlowNet [Dosovitskiy

et al., 2015], that was pre-trained on the large synthetic Flying Chairs dataset [Dosovitskiy et al., 2019]. The first version of FlowNet only stacks the two consecutive frames  $x_t, x_{t-1}$  on each other to predict the optical flow with one CNN, which had difficulties with large displacements. FlowNet2 uses a faster and more accurate method by using separate networks for small and large displacements, whose outputs are merged into a single flow image.

Prediction of the optical flow  $f_{x_t, x_{t-1}}$  is done for two consecutive frames  $x_t$  and  $x_{t-1}$  at a given time step  $t$  via the FlowNet2  $\mathcal{F}(x_t, x_{t-1})$ . The dimensions of  $f_{x_t, x_{t-1}}$  are  $[h_x \times w_x \times 2]$ . The two layers in the last dimension correspond to the shift of each pixel in x- (first layer) and y- (second layer) direction. Estimation of the optical flow is defined to be backwards in time, from  $x_t$ , to  $x_{t-1}$ . Defining this is important because the optical flow estimation is not bijective due to effects such as dis-/occluded regions [Sánchez et al., 2015].

**Summarizing the optical flow** To summarize the optical flow  $m_t$  as a latent representation of the motion in the complete sequence at a time step  $t$ , as model  $\mathcal{R}(X, t)$ , a recurrent neural network in the form of a Long short-term memory network (LSTM) was chosen (see Sec. 2.1.4 *Recurrent neural networks*). The LSTM first computes the optical flow with  $\mathcal{F}$  between all consecutive frames, before it summarizes them all in the latent representation  $m_t$  with the dimensions  $[h_x \times w_x \times 1]$  (see Eq. 3.10).

$$m_t = \mathcal{R}(\{\mathcal{F}(x_i, x_{i+1}) | i \in t - N_X, \dots, t - 1\}, t) \quad (3.10)$$

with  $X$  being the video sequence,  $x_i$  being the  $i$ -th frame in that video,  $N_X$  being the number of video frames in the video snippet,  $t$  current the time step and  $\mathcal{F}$  being the FlowNet2 model for the flow estimation.

### Segmentation of multiple instances

Given is a video sequence  $X$  of length  $N_X$ , where for the last video frame  $x_t$  at time step  $t$ , all instrument instances  $I_x$  that are present in  $x_t$  should be segmented. As the number of visible instances differs from image to image, a Mask R-CNN [He et al., 2017] (see Sec. 2.1.4 *Mask R-CNN*) was used as segmentation model  $\mathcal{S}$ . To keep the same spatial dimensions, the input of  $\mathcal{S}$  is the concatenation in the last dimension of the RGB image  $x_t$ , the likelihood of a pixel being an instrument  $\mathcal{B}(x_t)$  and the latent space of the motion about all video frames  $\mathcal{R}(X, t)$  (see Eq. 3.11).

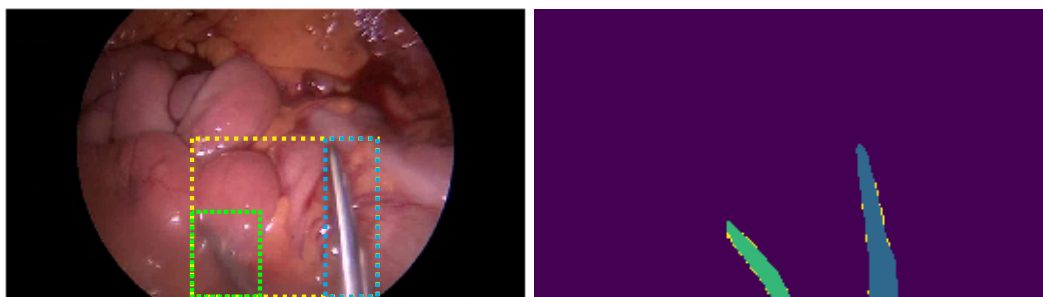
$$\hat{I}_x = \mathcal{S}(x_t, \mathcal{B}(x_t), \mathcal{R}(X, t)) \quad (3.11)$$

where  $\hat{I}_x = \{(\hat{i}_{x,j}, \hat{s}_{x,j}) | j \in \{1, \dots, N_{\hat{I}_x}\}\}$  is a set of predictions that yields pairs of a predicted instances  $\hat{i}_{x,j}$  with its corresponding predicted  $IoU$  score  $\hat{s}_{x,j}$  and  $N_{\hat{I}_x}$  being the number of predicted instances. The dimensions of a  $\hat{i}_{x,j}$  are  $[h_x \times w_x \times 1]$ , with values within  $[0, 1]$  that are the probabilities of the image pixels belonging to instance number  $j$ , while  $\hat{i}_{x,j}$  is a value within  $[0, 1]$ .

Since the Mask R-CNN often produces thousands of candidates in an image, only instances that have a high predicted  $IoU$  of the bounding box above a threshold  $\tau$  are considered as valid instances. After thresholding each pixel is usually assigned to an instance number by performing the  $\arg \max$  per pixel to get the highest probability for an instance number. However, this procedure might lead to problems, as will be explained in the next paragraph Post-processing.

### Post-processing

One of the peculiarities in the endoscopic instrument segmentation occurs in cases where instruments overlap or are only partially visible on the side of an image. In both cases the bounding box proposals can overlap, with all containing a high score which can make it difficult to proceed. These bounding box ambiguities [Robu et al., 2020] can lead when the usual pixel-wise  $\arg \max$  operation is performed on the results, to a segmentation with mixed and distributed instance numbers (see Fig. 3.23).



(a) Image with three bounding box proposals for the same instrument (b) Resulting image when performing the  $\arg \max$  to achieve the predicted labels.

Figure 3.23: This figure shows (a) how the same instrument could become part of three different bounding box proposals where all three proposals achieve a high score of the Mask R-CNN. Applying the  $\arg \max$  operation to achieve the predicted label would result a mixed up labeling outcome (b).

**Remove overlapping instances** As usual, when working with Mask R-CNN, during the first step all predicted instances that have a score greater than  $\tau$  are considered as valid instances  $I_x^{valid}$ . In the second step, however, all valid instances are pairwise compared with the DSC metric to find proposals that have a large overlap  $I_x^{filtered} = \{i_j \in I_x^{valid} | \nexists i_k \in I_x^{valid} : DSC(i_j, i_k) \geq \gamma \wedge ||i_k|| > ||i_j||\}$  (where  $||i_l||$  denotes the number of pixels in the instance  $i_l$ ) In every case, where the instances have a high degree of overlap ( $DSC \geq \gamma$ ), the proposal with fewer pixels is removed.

**Refine labels** After multiple proposals for the same instance were removed, there remains the possibility of overlapping bounding boxes, especially in the case where instruments overlap (see Fig. 3.24).

For this purpose, a CRF (see Sec. 2.3) is used, where the unary energy is defined as the probability of each pixel belonging to an instance. After fitting the CRF, the  $\arg \max$  operation is applied on each pixel to assign the corresponding instrument instance number. In a final check, the result of the  $\arg \max$  is binarized and compared to the binary segmentation  $l_x = \mathcal{B}(x)$  from Sec. 3.5.2 *Prediction of the instrument likelihood*. If only small regions of pixels are left (number of pixels  $\geq \delta$ ), a connected component analysis [Shapiro, 1996] is performed to assign those small regions to the labels previously assigned with the CRF. If there are still large areas of instruments remaining, those are added as an additional class to the CRF and the CRF is fitted again to include the missing pixels. The results of the post-processing can be seen in Fig. 3.25

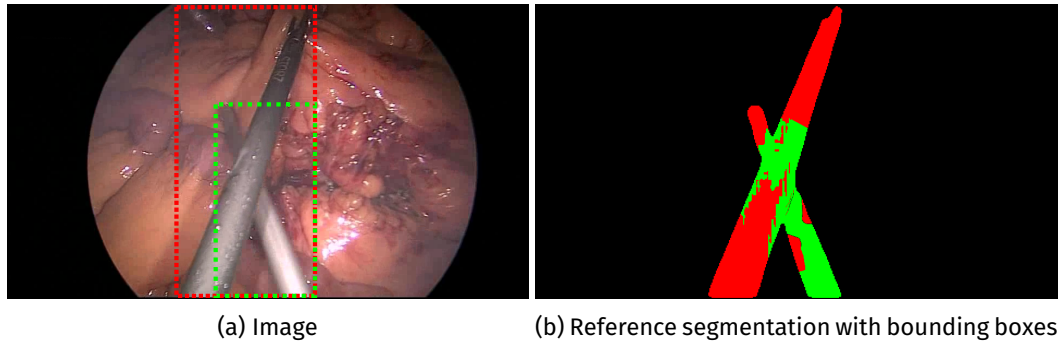


Figure 3.24: The figure shows the original RGB image (a) and the corresponding instance segmentation with their bounding boxes (b). It can be seen that the bounding boxes overlap.

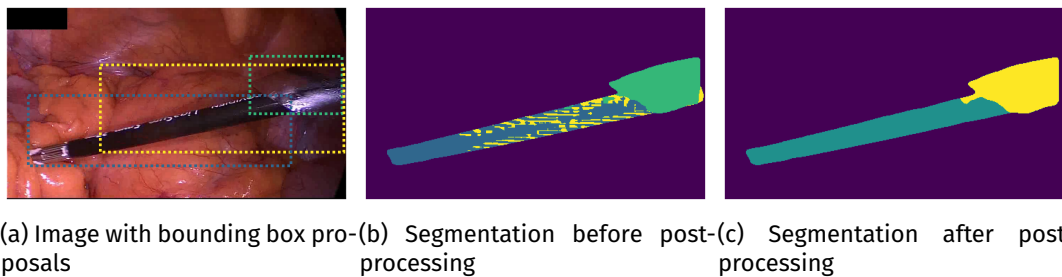


Figure 3.25: The figure shows the original RGB image (a) and the corresponding instance segmentation with the bounding boxes proposed by the network (b). The results before post-processing and (c) the results after post-processing.

### Experimental setup

This section describes the experimental setup that was used to validate the developed concept. For this purpose, the dataset and metrics used are first described, followed by the implementation details of the presented deep learning model. To conclude, the details of all experiments are explained.

**Dataset and metrics** The evaluation was based on the challenge dataset as described in Sec. 3.2 *Conclusion*. Thus, the dataset consisted of 5,983 training and 4,057 test images, where each image comes with a 10s video snippet of previous frames. Following [Roß et al., 2020], the test set was split into three stages to explore different grades of the algorithm to generalize to new procedures and patients.

- Stage 1: 663 (16 % of the test data) frames that are not part of the training set were extracted from the same patient videos from which the training data were extracted
- Stage 2: 514 (13 %) frames were extracted of the same type of surgery as the training data, but taken from patients not included in the training data
- Stage 3: 2,880 (71 %) frames were the data are from a different, but similar type of surgery



For the validation of the method, the rankings and metrics are similar to 3.3.4 *Metrics and Ranking*. Specifically, the multi-instance Dice Similarity Coefficient ( $DSC_{MI}$ , [Xu et al., 2017]) was used, and the Hungarian algorithm [Kuhn, 1955] was used to assign corresponding pairs [Roß et al., 2020].

**Implementation details** All parameters were set according to small experiments on a held-out validation set during the implementation process.

Instrument likelihood prediction The instrument likelihood prediction is based on the method of Isensee et al. [Isensee and Maier-Hein, 2020] and thus uses an ensemble vote of eight models, that were trained by leaving one surgery out on a 8-fold cross-validation. Each model was trained for 2.000 epochs using randomly selected patches at size of  $256 \times 488$ , where an epoch was defined as an iteration of over 100 batches. For preventing overfitting, the following data-augmentation techniques were randomly applied on each image during training: contrast, brightness and gamma-augmentation, image rotation, scaling, mirroring, elastic deformation and additive Gaussian noise [Isensee and Maier-Hein, 2020]. The optimizer was the SGD, with learning rate decay that went to 0, starting from  $2^{-5}$  [Isensee and Maier-Hein, 2020].

Image flow prediction For all experiments, the optical flow was computed via a video sequence that lasted  $5s$  before  $t$ . Since using the original video frame rate of  $25fps$  would result in a vanishing gradient and would require too much memory, the optical flow was computed with a frame rate of only  $3fps$ . This resulted in an optical flow with the resolution of  $[15 \times h_x \times w_x \times 2]$ . As the computation of the optical flow is resource-intensive, it was pre-computed and stored.

Summarized was the optical flow with the LSTM using 6 hidden convolutional layers of dimension  $\{32, 16, 8, 4, 2, 1\}$ . The optical flow consists of two components (x / y component), each being summarized on its own, but all components with the same model.

Multi-instance segmentation The Mask R-CNN is based on a ResNet-50 backbone, with the anchor sizes of (8, 16, 32, 64, 128, 256, 512, 1024) and aspect ratios of (0.5, 1.0, 2.0). The detection threshold was 0.2 in order to also detect small instrument boundaries. For training the model, the SGD optimizer was used, with a momentum of 0.9, a learning rate of 0.0005 and a batch consisting of 2 samples. To prevent the training from crashing, all gradients were clipped at 1.0. Since the meaning of the optical flow is not invariant to augmentation techniques that change pixel locations, no data augmentation techniques were used. Only the image dimension was halved, with respect to the optical flow vector.

Post-processing For the post-processing,  $\gamma$  (overlap between proposals) was set to 0.5 and  $\delta$  (very small pixel regions) was set to 100 pixels. The parameters were estimated on a small validation set that was randomly taken from the training dataset.

## Experiments

*E1: Effect of including temporal information and instrument likelihood* This question for this experiment is to assess whether the temporal information and/or the instrument likelihood can provide information that can increase the multi-instance segmentation performance. For this purpose,

an ablation study was carried out, which trained the five models with different configurations, as shown in Tab. 3.10 in order to see potential changes in the performance when temporal and/or instrument likelihood information is not used for the training. All five experiments are carried out with the same set of hyperparameters in order to maintain comparability between the experiments.

Table 3.10: This table shows the training configuration for the ablation study. The name of the trained model  $T$  with the indices  $R, L$  and  $F$  are referring to  $R = \text{raw}$ ,  $F = \text{flow}$  and  $L = \text{likelihood}$

Name	Raw	Flow	Likelihood
$T_R$	✓	✗	✗
$T_{RF}$	✓	✓	✗
$T_{RL}$	✓	✗	✓
$T_{FL}$	✗	✓	✓
$T_{RFL}$	✓	✓	✓

*E2: Effect of post-processing* To assess the effect of the post-processing approach, the procedure was applied on top of the model that takes as input the raw image, the optical flow, and the instrument likelihood  $T_{RFL}^+$ . The method is also compared to other state-of-the-art methods to check whether it performs better than the others on images where instruments overlap. The Wilcoxon signed-rank test [Wilcoxon, 1992] test was carried out to test for significance.

*E3: Comparison to state-of-the-art* The presented approach was compared to the ROBUST-MIS 2019 Challenge's top-scoring methods based on the accuracy as in the ROBUST-MIS challenge [Roß et al., 2020]. The ranking is based on the significance ranking described in [Maier-Hein et al., 2018]. This yields the accuracy rank  $r_a(a_i)$ . Additionally, the presented approach is compared to the top-scoring methods when multiple instances are visible in the image.

*E4: Investigation of characteristic influence* To investigate whether the negative influence of image characteristics could be successfully reduced, the influence of all characteristics on the challenge participants' methods is compared to the method developed in this thesis. For this reason, the same accuracy ranking is performed as described in E3. Finally, a pairwise one-sided Wilcoxon signed rank test at  $\alpha = 5\%$  significance level with adjustment for multiple testing according to Holm [Holm, 1979] was used to check for statistical significance.

### 3.5.3 Results

#### E1: Effect of including temporal information and instrument likelihood

In Fig. 3.26 it can be seen that including the optical flow improved the performance from 0.30 to 0.64 (213%) on the median  $MI\_DSC$  and from 0.36 to 0.65 (181%) in the mean. When including the instrument likelihood, the median  $MI\_DSC$  improved from 0.30 to 0.93 (310%) and 0.36 to 0.79 (219%) in the mean. The best performance could be achieved when combining the RGB image with the optical flow and the instrument likelihood, leading to the smallest IQR of 0.30. Further

descriptive statistics can be found in Tab. 3.11.

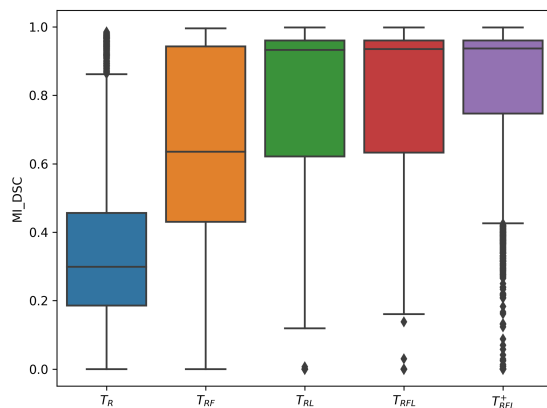


Figure 3.26: This figure shows the performance of the models when trained on the raw only ( $T_R$ ), on raw with the flow ( $T_{RF}$ ), on raw and likelihood ( $T_{RL}$ ), and when combining all three ( $T_{RFL}$ ) with and without post-processing ( $T_{RFL}^+$ ). As all models were trained with the same model parameters, the raw performance is deficient.

## E2: Effect of post-processing

As shown in Fig. 3.26 and in Table 3.11, the post processing significantly ( $p = 2.7E - 7$ ) increases the mean performance of the model  $T_{RFL}$  at 1%, while simultaneously reducing the IQR from 0.30 to 0.21. Also the robustness (defined in [Roß et al., 2020] as the worst 5%) of the method increases from 0.28 to 0.32. However, the approach could not improve the problem with the overlapping instruments.

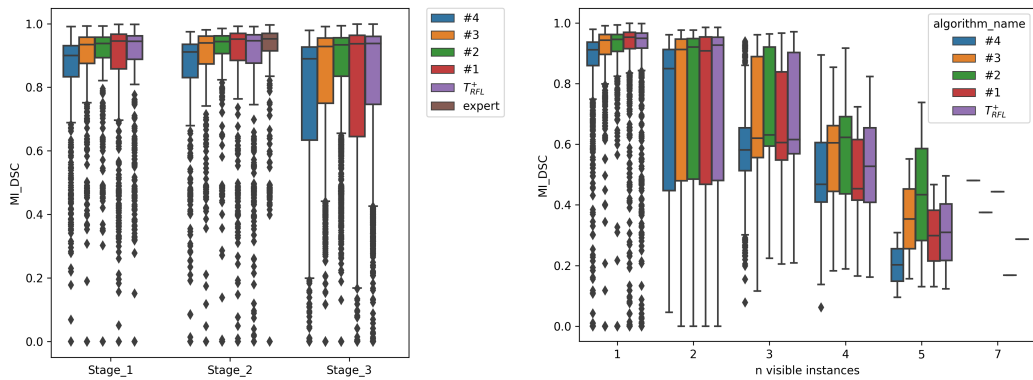
Table 3.11: Effect of the different inputs for training the Mask R-CNN, showing the mean ( $\mu$ ), median ( $\tilde{x}$ ), 5th, 25th, and 75th quartile ( $Q_1, Q_3$ ), and the interquartile range ( $IQR$ ) of the multi-instance dice coefficient  $DSC_{MI}$ . The name of the trained model  $T$  with the indices  $R, L$  and  $F$  are referring to  $R = \text{raw}$ ,  $F = \text{flow}$  and  $L = \text{likelihood}$ . The model  $T_{RFL}^+$  is the same model as  $T_{RFL}$ , but followed by the post-processing.

Model	$\mu$	$\tilde{x}$	$Q_5$	$Q_{25}$	$Q_{75}$	$IQR$
$T_R$	0.36	0.30	0.00	0.19	0.46	0.27
$T_{RF}$	0.65	0.64	0.17	0.43	0.94	0.51
$T_{RL}$	0.79	0.93	0.25	0.62	0.96	0.34
$T_{FL}$	0.76	0.92	0.29	0.55	0.96	0.41
$T_{RFL}$	0.80	0.93	0.28	0.63	0.96	0.30
$T_{RFL}^+$	0.81	0.94	0.32	0.75	0.96	0.21

### E3: Comparison to state-of-the-art

Comparing this method to the current state-of-the-art methods from the challenge, the proposed method in combination with the post-processing step achieves the highest median  $MI\_DSC$  (0.94) in comparison to the best performing method (*Uniandes (#2)* with 0.93) and the same mean  $MI\_DSC$  value as the best performing method of 0.81. The robustness is at 0.32, the same as the best model (*www (#4)* with 0.32 and *Uniandes (#2)* with only 0.26). This result means that the mean of the median is comparable, but the robustness is slightly better, which can also be seen in (a) of Fig. 3.27.

As noted (b) in Fig. 3.27, the segmentation quality, according to the number of visible instruments, is very mixed. While the method presented in this thesis is slightly better than the state-of-the-art when only one or two instruments are visible in an image, the performance drops rapidly under the state-of-the-art performance when three or more instances are visible.



(a) Comparison across the different test stages.

(b) Comparison across visible instances.

Figure 3.27: This figure shows how the presented method compares to the other state-of-the-art algorithms (a) across the three different test stages and (b) across the number of multiple instrument instances visible in an image. It can be seen that the presented method has on Stage 3 the highest median score (a) and the best performance when one or two instruments are visible (b).

### E4: Investigation of characteristic influence

As can be seen in Fig. 3.29, there are slightly improved median values for the presented method for the characteristics when other objects are visible in the background and when the background contains smoke. For all the other characteristics, the proposed method is always as good as one of the other state-of-the-art models. However, applied to the accuracy ranking, the presented approach is the best model in four out of six characteristics, and ranked second place in two of the cases. The stability of this ranking is shown in Fig. 3.28.

Tab. 3.12 lists characteristics in which the presented method is significantly better than the state-of-the-art. For example, the proposed method handles equal or worse to the winning algorithm *fisensee (#1)*, two characteristics better than the winner of the robustness *Uniandes (#2)*, and, other than some exceptions, better than *caresyntax (#3)* and *www (#4)*.

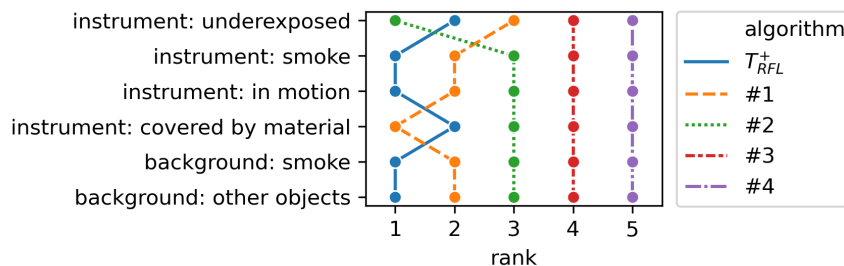


Figure 3.28: The ranking of all algorithms, depending on six characteristics that have a significant negative impact. The ranking was performed with 1,000 bootstrap samples. It can be seen that the proposed method ( $T_{RFL}^+$ ) is for most of the characteristics on the first rank.

Table 3.12: This table shows whether the proposed method performs significantly better than the state-of-the-art methods. The methods are named according their rank on the challenge (#1 is the first place, while #4 is the 4th place). A ✓ means that a the proposed method performs statistical significantly better, while ✗ means no statistical significance could be found.

Characteristic	#1	#2	#3	#4
Background: other objects	✗	✓	✓	✓
Background: Smoke	✗	✗	✓	✗
Instrument: Covered by material	✗	✗	✗	✗
Instrument: In motion	✗	✗	✓	✓
Instrument: Smoke	✗	✓	✓	✓
Instrument: Underexposed	✗	✗	✓	✓

### 3.5.4 Discussion

This section’s main goal was the development of an algorithm that should be able to specifically improve the weaknesses of previous state-of-the-art algorithms. While current methods have not yet managed to take advantage of the temporal information for multi-instance segmentation, this work was the first that generated a benefit using the optical flow in combination with the instrument probability. The increased performance was also present on images with difficult properties, which was the first time that an explicitly designed method that was based on a statistical analysis of weaknesses did not only achieved a new top score on the dataset but was also able to increase the performance for some of the difficult image characteristics that were identified in the previous section (Sec. 3.4). While this work is not the first that uses the optical flow (e.g., [Allan et al., 2015]) in general, however, it is the first time that it was combined with the instrument likelihood, summarized with an LSTM, and used in the context of multi-instance segmentation.

The observation that the optical flow can help to improve the outcome is also in line with the work of Allan et al. [Allan et al., 2015], where the authors showed that the optical flow could improve the estimation of the surgical instrument pose, and is also consistent with Jin et al. [Jin

et al., 2019], where the authors could show a benefit for the binary segmentation. Apart from a better score, the algorithm manages obstacles with certain characteristics significantly better than many other state-of-the-art methods.

However, it should be noted that the algorithms' current weaknesses were evaluated on the test set, and thus there was a small prior during the development of the method. Nevertheless, the analysis revealed that the distribution of the characteristics is almost identical between the training and test set; thus, it can be assumed that the observed effects are very likely not a fit to the test data.

The presented results suggest that typical challenges of laparoscopic videos, such as reflection, blood, and lighting variations, can already be managed well by state-of-the-art methods. However, there are still difficulties with tube-like structures that are misclassified as instruments or transparent objects such as trocars. Furthermore, images with crossing or close instruments are difficult to separate for all state-of-the-art methods, as well as for the presented approach, even though it was specially designed to manage such difficulties. One limiting factor may be due to restrictions in the training and test dataset, where only 36% of the data contains at least two and only 8% of the images have more than two instrument instances. Furthermore, only in rare cases do those instances overlap or intersect; thus, there is only a limited possibility to train and evaluate the algorithm's separation capabilities.

Although real-time capability is an essential prerequisite for bringing the approach into a clinical setting, this would not be possible with the approach presented here. The CRF and the estimation of the optical flow would take already approximately more than 2 seconds per image. However, the method presented was an attempt to work towards a solution of the multi-instance segmentation. In the next step, an attempt could be made to make the algorithm real-time capable.

### **3.5.5 Conclusion**

This work was the first that examined whether a method can be developed that improves the multi-instance segmentation, based the findings of an in-depth statistical analysis of a benchmarking competing, where unsolved problems were identified. The results of the experiments suggest, that the performance could be increased to a new best-performance score and thus including such information is helpful. However, an improvement could not be achieved for all identified problems, but this was probably due to the architecture of the deep learning model used. In conclusion, this approach is a good step towards problem-driven algorithm development.

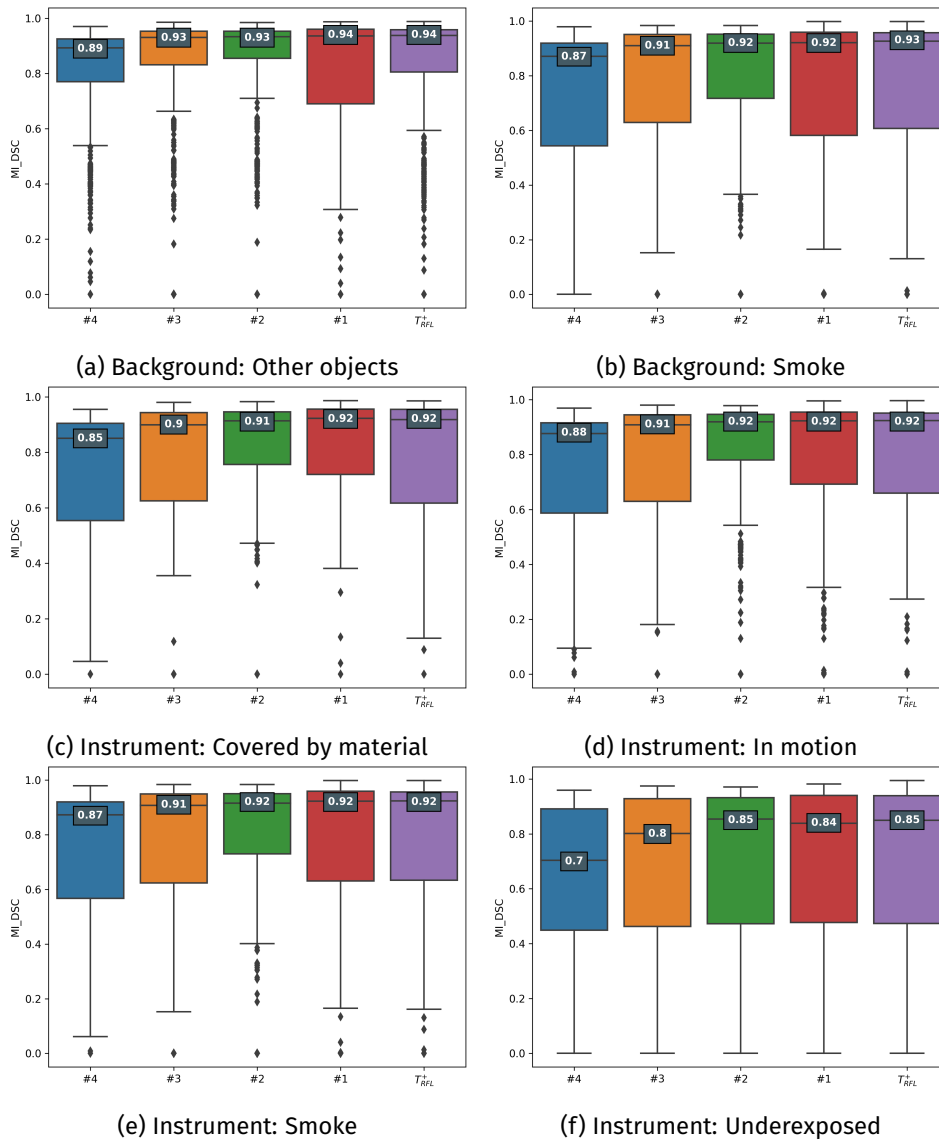


Figure 3.29: This figure shows the performance of the algorithms when applied to images containing the specified characteristic. The value on the median line of the boxplot is the median value.





# 4 | Discussion

The main goal of SDS is to increase patient safety by supporting and assisting surgeons in performing their daily clinical work. A successful transition from a SDS application into the clinical routine requires that the application is robust, accurate, and generalizes well, up to a certain degree, in the dynamic and complicated environment of the surgical suite. However, these requirements are difficult to achieve for state-of-the-art machine learning approaches, because they heavily rely on representative data for training. Unfortunately, such data for the SDS environment is very limited, and that which is available lacks quality control and offers only a few samples which makes the data less representative of real-world scenarios. The experiments and methods presented in this thesis are therefore directed towards resolution of these issues.

## 4.1 Summary of Contributions

The specific contributions to the field of SDS presented in this thesis can be summarized by using the aforementioned objectives of this thesis:

### *O1: Incorporation of unlabeled data into the training of deep learning models*

The contribution of this objective was to work towards the SDS challenge of managing sparse data. It was investigated whether unlabeled data could be included into the training process of deep learning models, with the goal to increase the algorithm performance. In order to facilitate this questions, a method was developed using a self-supervised learning approach to incorporate unlabeled data into the training process for the medical instrument segmentation. Prior to this, state-of-the-art deep learning models had only been pre-trained on a related dataset that required manually generated labels and then fine-tuned on the target domain. Analysis of the results suggests that the developed concept enables the training of deep learning models with a drastically reduced dataset of the target domain. Although with this method, the performance of a deep learning model trained on only a few labeled data could be significantly increased, the achieved performance was not high enough to cover the lack of training data. The

methods used toward this end were presented at the *International Conference on Information Processing in Computer-Assisted Interventions* and published in the corresponding *International Journal of Computer Assisted Radiology and Surgery* [Ross et al., 2018].

*O2: quality controlled, dataset generation for the task of multi-instance surgical instrument segmentation*

Based on the observation from the previous objective (O1) that more training data is needed, the goal of this objective was to work towards the SDS challenge of the limited availability of training data. Existing datasets consist mostly of only annotated data items numbering in the thousands, or even only hundreds. Further, labels have often not been subject to quality control, and labeling instructions are not yet standard. Thus, the contribution of this objective to the SDS research area was to provide the largest dataset to date for multi-instance segmentation in laparoscopic surgery. The definition of the labels was clearly defined in an annotation protocol, and a quality control process ensured consistent labels, even across different raters. The dataset itself contains, besides the instrument annotations for each image in the dataset, global, and local image characteristics, such as the presence of blood, smoke, or other similar factors. The dataset is publicly available and is currently under review (minor revision) for the journal *Nature Scientific Data* [Maier-Hein et al., 2020b].

*O3: Structured assessment of state-of-the-art performance*

The goal of this objective was to work towards the SDS challenge of the limited availability of training data for the task of multi-instance instrument segmentation of laparoscopic instruments. Due to the fact that no dataset for the problem of multi-instance segmentation to date has been available, there is no related literature and benchmarking on this problem. To determine the current state-of-the-art, the dataset from the previous Objective (O2) was released to the community in the form of the *Endoscopic Vision Challenge Robust Endoscopic Instrument Segmentation Challenge 2019* at the *International Conference on Medical Image Computer & Computer Assisted Intervention*. As expected, the results revealed that the increased dataset led to an improvement of the binary instrument in comparison to the results of the first objective (Incorporation of unlabeled data). Further insights were that there are image characteristics (such as the presence of blood or smoke) which seems to harm the algorithm performances. The results of this challenge were published in the *Journal Medical Image Analysis* [Roß et al., 2020].

*O4: Systematic problem analysis of state-of-the-art methods for multi-instance surgical instrument segmentation*

The contribution of this objective was to work towards the SDS challenge of data sparsity. Based on the observation that there is often not enough data available to enable sufficient training for all edge-cases, a systematic problem analysis can provide valuable insights in order to develop algorithms that specifically manage problems discovered during the analysis. Such an analysis is particularly useful when several algorithms have been applied to the same problem, such as during benchmarking competitions (see previous objective O3). Unlike previous benchmarking competition publications, which mainly contain the performance results and a description of the participating methods, such a problem analysis was carried out for the first time. For this purpose, a statistical method was developed to identify and quantify the effect of image characteristics (such as the presence of blood or smoke) on the performance of multi-instance segmentation algorithms. The results from the experiments show that image characteristics that were assumed to have a very negative effect on performance have significantly less influence on

quality than assumed. Instead, it seems that there are more methodological difficulties with the algorithms concerning multi-instance segmentation.

*O5: Problem-driven multi-instance surgical instrument segmentation*

The contribution of this objective was to work towards the SDS challenge of data sparsity. Based on the previous objective's findings (O4), the contribution of this part was to develop a method that is explicitly designed to tackle the most significant image characteristics that harm the performance of multi-instance segmentation algorithms. To address the problems an attempt was made to include temporal information in the algorithm, and special post-processing was used. The experiments showed that the overall algorithm performance could be increased with this procedure, and a new top benchmarking score on the dataset from O2 could be achieved. However, not all problems could be successfully addressed, even though the method was explicitly designed for it.

## 4.2 Discussion of results

While a detailed discussion of the results and experiments can be found in the corresponding Results sections, this part of the thesis provides a more general discussion of the experimental validation.

### 4.2.1 Potential of unlabeled data

The results that were shown in Sec. 3.1 demonstrate the possibility of including unlabeled data into the training process of binary instrument segmentation. Although this form of training supplemented conventional approaches, such as data augmentation, a positive effect of the self-supervision was particularly present when only little training data was available. While only a single pre-training task was investigated in this work, there would, of course, also theoretically be the possibility that choosing a different auxiliary task, or combine several auxiliary tasks, would improve the performance, even if more training data were available. Elastic deformation of the image and the subsequent reconstruction, for example, might lead the network to acquire more geometric knowledge about the images. Another alternative might be the combination of self-supervised learning approaches with a taskonomy setting, as described by Zamir et al. [Zamir et al., 2018]. Such a setting would improve the results by pre-training on related helpful datasets. However, taking results from other self-supervised approaches into account, none of them outperformed the model that was trained on full data. Thus, this method offers the possibility to train a model with as little labeled data as possible, but it cannot replace the generation of additional data.

That more labeled data leads to better performance becomes especially obvious when taking the results of the ROBUST-MIS Challenge 2019 into account. The performance results of the binary segmentation of a comparable segmentation model were significantly higher than those of the self-supervised learning method. While the finding that including more data produces better results is not new, it shows a new facet in this case. So is the number of data used for the self-supervised learning approach higher than the number of training data available for the models of the ROBUST-MIS challenge. Hence, the sheer mass of data itself could not compensate for the lack of labels.

### 4.2.2 Quality controlled dataset generation

The work presented in Sec. 3.2 presents the largest available for the training of multi-instance segmentation of medical instruments in laparoscopic videos to date. The generation of this dataset followed an annotation protocol and was subject to quality control. Nevertheless, the inter-rater variability experiments in Sec. 3.3 revealed that annotators did not always follow the defined rules exactly. Those errors could have happened because annotators may not have been able to keep all the rules in mind. Consequently, it can be assumed that incorrect annotations are always present in the data despite precise specifications and quality controls. Based on this finding, there are multiple ways one could deal with it. One could be to increase the quality control and develop tools to guide the labeling even further. However, this would come at the cost of a more time-consuming process. Another possibility is to take errors into account and design deep learning models accordingly (e.g., Guo et al. [Guo et al., 2019], and Gu et al. [Gu et al., 2020]). Still, it should always be considered to find the sweet spot between label quality and resource investment, such that a minimum level of quality is ensured, which needs to be defined by the community.

Despite the labeling quality, another important point is the current lack of label standardization. There are benchmarking competitions that provide segmentation, e.g., in the form of a polygon encoded into a JSON file [Lin et al., 2014a], while others use images and assign each pixel a value [Allan et al., 2019, Bodenstedt et al., 2018, Roß et al., 2020]. A standard would significantly simplify loading, processing, and thus testing on several datasets in parallel, which would make developing and benchmarking easier.

The final point of discussion is about the size of the dataset. It remains unclear which dataset size is sufficient to train robust algorithms or how much additional data would be required to achieve this goal. However, guided by the image characteristics (e.g., the presence of blood or smoke in an image), one could see that certain combinations of characteristics are only very rarely covered. Accordingly, it is to be expected that more labeled data will be necessary before trained models could be transferred to the clinic. Nevertheless, this dataset offers an important step in the direction of larger, publicly available datasets for the future benchmarking of the segmentation of instruments in laparoscopic image data.

### 4.2.3 Comparative validation of multi-instance instrument segmentation

At the time of the development of Sec. 3.3 there was no comparative literature on the task of multi-instance segmentation. Thus it was only the implementation of a benchmarking competition that made it possible to create a state-of-the-art baseline for the community and provide deeper insights into the generalizability and robustness of current methods. An alternative to the competition would have been if one had implemented all the methods him or herself. However, taking the human and computational resources and know-how into account used to optimize a single team submission, this method forms a significantly stronger baseline.

Still, it must be noted that all of these methods were developed by candidates who were under time pressure because there was only a limited time window where the data was available before they had to submit their solution as a docker. This time pressure might have caused submissions do not have the same level of development as they would have in publications. Further, the time pressure might have affected the diversity of the methods, which is reflected

in the number of very similar Mask R-CNN methods. Surprisingly those similar methods often performed very differently. These performance differences could either be due to the slight differences in implementations and training processes, or, as shown in Pham et al. [Pham et al., 2020], possibly due to chance. More research is needed here.

While the challenge design mainly focused on the performance of algorithms concerning generalizability and robustness, it should be said that for clinical practice, however, further aspects are potentially interesting (e.g., real-time capability). Nevertheless, one could argue that multi-instance segmentation should be solved first before other aspects can be addressed in the next step. As shown by the comparison the human expert annotator, there is still much room for improvement. Although it was only a single annotator, all algorithms showed weaknesses when specific image characteristics (e.g., blood and smoke) were present. These weaknesses could come from the fact that the dataset was not large enough to contain enough examples of those images.

#### **4.2.4 Effects of image characteristics on the algorithm performance**

To estimate the influence of image characteristics on the performance of algorithms, as presented in Sec. 3.4, this requires that the main image characteristics that could influence the performance have mostly been identified. The identification of such characteristics, however, requires specific domain knowledge, which might not be easy to gain. For this work, the characteristics were identified by referring to literature and personal experiences during the labeling. Of course, this could imply a bias concerning characteristics that make it difficult for humans to process an image, and it does not exclude the possibility that essential characteristics have been forgotten.

One might also argue that in addition to identifying the characteristics, those must also have been assigned to the data. Of course, this labeling process comprises the risk of incorrect labels. While the problem of label quality can be countered to some extent by selecting suitable tools, the time aspect can still only be reduced to a limited extent. Thus, an alternative approach would be to invest the labeling effort into the generation of labeled data that could directly improve the training rather than annotating image characteristics. As long this additional data is not randomly selected, but, e.g., by an active learning approach [Gal et al., 2017] (e.g., Bodenstedt et al. [Bodenstedt et al., 2019]), this might be a valid objection. In contrast to the statistical analysis, this procedure does offer an easy possibility of a disentangled identification and quantification of characteristics that influence an image. However, maybe it would make sense to combine both approaches by identifying difficult characteristics and then selecting training data that would support the model.

A fascinating insight that the statistical method revealed is that many characteristics, which have been described as complicated in the literature, play a relatively minor role in reality. Instead, the most significant negative influence came from a methodological weakness, making the separation of overlapping instruments very difficult. Thus, even labeling additional data using active learning would probably not have resulted in any noticeable improvement.

#### **4.2.5 Multi-instance segmentation of medical instruments**

The results of the work presented in Sec. 3.5 are based on the findings of the benchmarking competition (Sec. 3.3) the lessons learned from image characteristics that negatively affect

the algorithm performance (Sec. 3.4). To solve the problem of overlapping instruments and to be able to better manage image characteristics, the optical flow was used to provide the network temporal information. While the presented approach achieved a new top-score for the ROBUST-MIS dataset, the new approach could not contribute to a better separation of overlapping instruments. This observation is particularly surprising since overlapping instruments should often have opposite movement directions, a feature that should help distinguish instrument instances. One reason why this may not have worked so well could have been the choice of FlowNet2 for the optical flow estimation. FlowNet2 is a network that was trained on synthetic data, and there is a possibility that its learned features may not be optimal for laparoscopic image data. Another reason could be the choice of the mask R-CNN architecture, which has problems in dealing with this form of overlap (see Zhu et al. [Zhu et al., 2019]). In conclusion, this form of algorithm development can lead to better results, but unfortunately, the mask R-CNN's methodological weakness could not be resolved.

#### **4.2.6 Transition into clinical routine**

Although a method was presented on how models could be trained with very little training data, and approaches were presented that were supplied with a large number of training data, none of them would be suitable for real use in clinical practice. While the self-supervision results are not sufficiently robust and accurate enough, none of the ROBUST-MIS Challenge methods nor the method from Section 3.5 are suitable for an application in a real-world setting. The reason lies in the performance that is still not high enough and in the lack of real-time capability. It should be mentioned that the run-time was not explicitly addressed either in the challenge or in this thesis. In addition, there is the limitation that all the approaches presented produce particularly good results when only one instrument is visible but have problems when several instances are in the image. Thus, all methods would not be suitable for segmenting instruments in any kind of application that aims to track for example individual instances between frames. Nevertheless, they form a valuable piece of the puzzle towards a solution.

### **4.3 Conclusion**

The work, presented in this thesis, investigated several means to alleviate two main SDS challenges, namely the problem of a limited data availability and how to deal with data scarcity.

For this purpose, first the potential of unlabeled data was investigated, which offers an exciting possibility to train deep learning algorithms when only a limited amount of labels is available. Although, the current performance is not high enough to be applied in a real clinical setting, this method is an important step towards overcoming the SDS challenge of data availability by managing sparse data.

Since it is foreseeable that the problem of data availability will not be easy to solve in the near future and methods that deal with little data are still subject to research, the to date biggest high quality dataset for the task of multi-instance segmentation was produced and released to the community. This dataset has the potential to enable researches to train and benchmark their algorithms on a bigger amount of data.

The first benchmarking of multi-instance segmentation algorithms has already been carried out as part of an international challenge. The results of the challenge provide other researches the possibility of a baseline and to compare their algorithms against many other approaches. While

challenges reports are usually limited to reporting the different methods and their performance, a statistical method has been developed that analyzes and quantifies open problems. This method is suitable for developing algorithms as targeted as possible.

To achieve a robust, accurate and generalizing application of multi-instance medical instrument segmentation algorithms for laparoscopic videos, more work is required. However, this work revealed the great potential of statistical analyzing benchmark competitions with respect to unsolved problems. As these unsolved problems can be due to missing training data, the statistic can be used to develop solutions, that can overcome these problems and might be another tool for managing sparse data.

In conclusion, one can confidently assume that the combination of generating data and a problem-driven algorithm development and design has the potential to finally bridge the gap between research and clinical transition. Further, I believe that short-term and significant progress in the field of SDS can be achieved by focusing on the generation of more labeled data and making it available to the community in the form of challenges. The structured analysis of such challenges enables the development of algorithms to be carried out in a much more targeted manner.





# 5 | Summary

Surgical data science (SDS) is a research field that aims to improve the quality of interventional healthcare by observing all aspects of the patient treatment process to provide the right assistance at the right time. To date, most SDS applications are based on the deep learning technique, which has shown great potential to solve challenging tasks in a complex surgical environment. However, such algorithms are dependent on a large amount of training data, which not only must contain data, but also labels (e.g., localization of an instrument in the image), so that they can be used for training. To date, however, such a mass of training data is not available. This is primarily because the creation of such data would often require medical experts, as well as significant time and money resources. This data scarcity motivates the two major challenges of surgical data science, namely, (1) how algorithms based on machine learning methods can be trained despite the limited availability of such data and (2) how more training data could be provided. For this work, as a concrete example of a surgical data science application, instrument segmentation of medical instruments in images of laparoscopic videos was used.

This thesis investigated several means to alleviate this data scarcity in the context of laparoscopic instrument segmentation resulting in the following main contributions: First, it was examined how **unlabeled data can be integrated into the training of a machine-based algorithm to reduce the amount of annotated data**. Although with this method, the performance of a deep learning model trained on only a few labeled data could be significantly increased, the achieved performance was not high enough to cover the lack of training data. For this reason, as second contribution, **the largest dataset to date for the segmentation of multiple instruments in images of laparoscopic videos was created**. Generating the dataset followed a strict annotation protocol and was quality controlled. The created data was then **published as part of an international challenge to test the submitted methods and identify unresolved problems**. The third contribution was that **image characteristics were determined which negatively affect the segmentation quality**. In order to identify and quantify the influence of such characteristics, a statistical method has been developed. This analysis then flowed into the last contribution, **the targeted development of an algorithm that was designed to address the identified difficult characteristics** and achieved

the best performance on the challenge dataset.

As a result, this work provided a new tool for dealing with data sparsity by revealing the great potential of unlabeled data and the performance gain that can be achieved when generating high-quality datasets. Further, it showed that an in-depth statistical analysis of challenge results could be used to identify open issues of state-of-the-art methods and develop algorithms that are specifically designed to address those issues. This problem-driven approach even leads to a new best score on the task of multi-instance segmentation. Based on this thesis's results, one can confidently assume that the combination of generating data and problem-driven algorithm development and design has the potential to bridge the gap between research and the transition into clinical practice.

# 6

## Zusammenfassung

Die Surgical Data Science (SDS) ist ein Forschungsgebiet, das darauf abzielt die Qualität der konventionellen Gesundheitsversorgung zu verbessern, indem alle Aspekte des Patientenbehandlungsprozesses beobachtet werden und die richtige Unterstützung zur richtigen Zeit bereitgestellt wird. Zum aktuellen Zeitpunkt basieren die meisten SDS-Applikationen auf der Machine Learning Technik "Deep Learnings", welche ein großes Potenzial für die Lösung von komplizierten Aufgaben in der komplexen chirurgischen Umgebung gezeigt hat. Solche Algorithmen sind jedoch auf eine große Menge an Trainingsdaten angewiesen, die nicht nur Daten, sondern auch sogenannte Trainingsziele (z. B. Lokalisation eines Instruments im Bild) beinhalten müssen, damit sie zuverlässig eingesetzt werden können. Zum aktuellen Zeitpunkt jedoch, steht diese Masse an Trainingsdaten nicht zur Verfügung, weil die Erstellung von solchen Daten oftmals medizinische Experten, sowie große Zeit und Geldressourcen beansprucht. Diese Datenknappheit motiviert die zwei großen Herausforderungen des SDS-Forschungsgebiets, nämlich (1) wie können, trotz der geringen Verfügbarkeit solcher Daten, Algorithmen die auf maschinellen Lernverfahren basieren trainiert werden und (2), wie können Trainingsdaten bereitgestellt werden. Dabei konzentriert sich diese Arbeit auf das konkrete Anwendungsbeispiel der Instrumentensegmentierung von medizinischen Instrumenten in Bildern von laparoskopischen Videos.

Diese Arbeit untersucht verschiedene Methoden, um diese Datenknappheit im Zusammenhang mit der Segmentierung laparoskopischer Instrumente zu verringern, was zu den folgenden Hauptbeiträgen führte: Zunächst wurde untersucht, wie **ungelabelte Daten mit in das Training eines maschinellen lernverfahrenbasierten Algorithmusses eingebunden werden** können, mit dem Ziel, die Anzahl an annotierten Daten zu verringern. Obwohl mit dieser Methode die Performanz eines deep learning Modells, welches mit nur wenigen gelabelten Daten trainiert wurde, signifikant gesteigert werden konnte, war die Performanz noch nicht hoch genug, um den Mangel an Trainingsdaten zu decken. Aus diesem Grund wurde als zweiter Beitrag dieser Arbeit zusätzlich der bis dahin **größte Datensatz für die Segmentierung von mehreren Instrumenten in Bildern von laparoskopischen Videos erstellt**. Die Erstellung des Datensatzes folgte dabei einem strikten Annotationsprotokoll und wurde qualitätskontrolliert. Der so erstellte Datensatz wurde

anschließend **im Rahmen einer internationalen Challenge veröffentlicht**, mit dem Ziel noch ungelöste Probleme zu identifizieren. Der dritte Beitrag besteht darin, dass **Bildeigenschaften identifiziert wurden, welche die Segmentierungsqualität negativ beeinflussen**. Um den Einfluss solcher Charakteristika eindeutig zu identifizieren und zu quantifizieren, wurde eine statistische Methode entwickelt. Die Ergebnisse dieser Analyse flossen dann in den letzten Beitrag dieser Arbeit ein, nämlich in die **zielgerichtete Entwicklung eines Algorithmus der die identifizierten schwierigen Merkmale adressiert** und welcher anschließend die beste Performanz auf dem Challengedatensatz erreichte.

Infolgedessen bietet diese Arbeit ein neues Werkzeug für den Umgang mit dem Datenmangel, indem das große Potenzial ungelabelter Daten und der Performanzgewinn, der beim Generieren hochwertiger Datensätze erzielt werden kann, aufgezeigt wurde. Darüber hinaus zeigte sich, dass eine eingehende statistische Analyse von Challengeergebnissen verwendet werden kann, um offene Probleme aktueller Methoden zu identifizieren und Algorithmen zu entwickeln, welche explizit auf die Probleme zugeschnitten sind. Dieser problemorientierte Ansatz führte zu einem neuen Spitzenwert für den Task der Multi-Instanz Instrumentensegmentierung. Basierend auf den Ergebnissen dieser Arbeit kann man davon ausgehen, dass die Kombination aus Datengenerierung und problemgesteuerter Algorithmenentwicklung das Potenzial hat, die Lücke zwischen der Forschung und der Überführung der Ergebnisse in die klinische Praxis zu schließen.

# Bibliography

- [Agrawal et al., 2015a] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. **Learning to see by moving**. In *Proceedings of the IEEE international conference on computer vision*, pages 37–45, 2015a.
- [Agrawal et al., 2015b] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. **Learning to See by Moving**. In *Proceedings of the IEEE Internat. Conference on Computer Vision*, 2015b.
- [Allan et al., 2015] Max Allan, Ping-Lin Chang, Sébastien Ourselin, David J Hawkes, Ashwin Sridhar, John Kelly, and Danail Stoyanov. **Image based surgical instrument pose estimation with multi-class labelling and optical flow**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 331–338. Springer, 2015.
- [Allan et al., 2019] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. **2017 Robotic instrument segmentation challenge**. Available online at: *arXiv preprint arXiv:1902.06426*, 2019.
- [Allan et al., 2020] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. **2018 Robotic Scene Segmentation Challenge**. *arXiv preprint arXiv:2001.11190*, 2020.
- [Alom et al., 2018] Md Zahangir Alom, Tarek M Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S Awwal, and Vijayan K Asari. **The history began from alexnet: A comprehensive survey on deep learning approaches**. *arXiv preprint arXiv:1803.01164*, 2018.
- [Arjovsky and Bottou, 2017] Martin Arjovsky and Léon Bottou. **Towards principled methods for training generative adversarial networks**. *arXiv preprint arXiv:1701.04862*, 2017.
- [Armstrong et al., 2009] Timothy G Armstrong, Alistair Moffat, William Webber, and Justin Zobel. **Improvements that don't add up: ad-hoc retrieval results since 1998**. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 601–610, 2009.
- [Augestad et al., 2020] Knut Magne Augestad, Khayam Butt, Dejan Ignjatovic, Deborah S Keller, and Ravi Kiran. **Video-based coaching in surgical education: a systematic review and meta-analysis**. *Surgical endoscopy*, pages 1–15, 2020.

- [Azimian et al., 2010] Hamidreza Azimian, Rajni V Patel, and Michael D Naish. **On constrained manipulation in robotics-assisted minimally invasive surgery**. In *2010 3rd IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics*, pages 650–655. IEEE, 2010.
- [Ballard, 1987] Dana H Ballard. **Modular Learning in Neural Networks**. In *AAAI*, pages 279–284, 1987.
- [Baur et al., 2017] Christoph Baur, Shadi Albarqouni, and Nassir Navab. **Semi-supervised Deep Learning for Fully Convolutional Networks**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017.
- [Beam and Kohane, 2018] Andrew L Beam and Isaac S Kohane. **Big data and machine learning in health care**. *Jama*, 319(13):1317–1318, 2018.
- [Bengio et al., 1994] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. **Learning long-term dependencies with gradient descent is difficult**. *IEEE transactions on neural networks*, 5(2):157–166, 1994.
- [Bodenstedt et al., 2017] Sebastian Bodenstedt, Martin Wagner, Darko Katić, Patrick Mietkowski, et al. **Unsupervised temporal context learning using convolutional neural networks for laparoscopic workflow analysis**. *arXiv:1702.03684 [cs]*, Feb 2017.
- [Bodenstedt et al., 2018] Sebastian Bodenstedt, Max Allan, Anthony Agustinos, Xiaofei Du, Luis Garcia-Peraza-Herrera, Hannes Kenngott, Thomas Kurmann, Beat Müller-Stich, Sebastien Ourselin, Daniil Pakhomov, et al. **Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery**. *arXiv preprint arXiv:1805.02475*, 2018.
- [Bodenstedt et al., 2019] Sebastian Bodenstedt, Dominik Rivoir, Alexander Jenke, Martin Wagner, Michael Breucha, Beat Müller-Stich, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. **Active learning using deep Bayesian networks for surgical workflow analysis**. *International journal of computer assisted radiology and surgery*, 14(6):1079–1087, 2019.
- [Bogdanova et al., 2016] Rositsa Bogdanova, Pierre Boulanger, and Bin Zheng. **Depth perception of surgeons in minimally invasive surgery**. *Surgical innovation*, 23(5):515–524, 2016.
- [Bolker et al., 2009] Benjamin M Bolker, Mollie E Brooks, Connie J Clark, Shane W Geange, John R Poulsen, M Henry H Stevens, and Jada-Simone S White. **Generalized linear mixed models: a practical guide for ecology and evolution**. *Trends in ecology & evolution*, 24(3):127–135, 2009.
- [Bouget et al., 2015] David Bouget, Rodrigo Benenson, Mohamed Omran, Laurent Riffaud, Bernt Schiele, and Pierre Jannin. **Detecting surgical tools by modelling local appearance and global shape**. *IEEE transactions on medical imaging*, 34(12):2603–2617, 2015.
- [Boulesteix and Schmid, 2014] Anne-Laure Boulesteix and Matthias Schmid. **Machine learning versus statistical modeling**. *Biometrical Journal*, 56(4):588–593, 2014.
- [Bowles et al., 2018] Christopher Bowles, Roger Gunn, Alexander Hammers, and Daniel Rueckert. **GANsfer Learning: Combining labelled and unlabelled data for GAN based data augmentation**. *arXiv preprint arXiv:1811.10669*, 2018.

- [Breslow and Clayton, 1993] Norman E Breslow and David G Clayton. **Approximate inference in generalized linear mixed models**. *Journal of the American statistical Association*, 88(421): 9–25, 1993.
- [Burström et al., 2019] Gustav Burström, Rami Nachabe, Oscar Persson, Erik Edström, and Adrian Elmi Terander. **Augmented and virtual reality instrument tracking for minimally invasive spine surgery: a feasibility and accuracy study**. *Spine*, 44(15):1097–1104, 2019.
- [Cardoso, 2018] M. Jorge Cardoso. **Medical Segmentation Decathlon**, 2018. <http://medicaldecathlon.com/>. Accessed: 2019-10-29.
- [Caron et al., 2018] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. **Deep clustering for unsupervised learning of visual features**. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.
- [Chang et al., 2015] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. **Shapenet: An information-rich 3d model repository**. *arXiv preprint arXiv:1512.03012*, 2015.
- [Chapaliuk and Zaychenko, 2020] Bohdan Chapaliuk and Yuriy Zaychenko. **Review of Semi-Supervised Learning Methods for Medical Computer-Aided Diagnosis Systems**. In *2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC)*, pages 1–4. IEEE, 2020.
- [Chen et al., 2020] Alvin I Chen, Max L Balter, Timothy J Maguire, and Martin L Yarmush. **Deep learning robotic guidance for autonomous vascular access**. *Nature Machine Intelligence*, 2(2):104–115, 2020.
- [Chen et al., 2019] David Chen, Sijia Liu, Paul Kingsbury, Sunghwan Sohn, Curtis B Storlie, Elizabeth B Habermann, James M Naessens, David W Larson, and Hongfang Liu. **Deep learning and alternative learning strategies for retrospective real-world clinical data**. *NPJ digital medicine*, 2(1):1–5, 2019.
- [Cheplygina et al., 2019] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. **Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis**. *Medical image analysis*, 54:280–296, 2019.
- [Cho et al., 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. **Learning phrase representations using RNN encoder-decoder for statistical machine translation**. *arXiv preprint arXiv:1406.1078*, 2014.
- [de Sa, 1994] Virginia R de Sa. **Learning classification with unlabeled data**. In *Advances in neural information processing systems*, pages 112–119, 1994.
- [Deng et al., 2009] 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*, pages 248–255. Ieee, 2009.
- [Dice, 1945] Lee R Dice. **Measures of the amount of ecologic association between species**. *Ecology*, 26(3):297–302, 1945.

- [Dosovitskiy et al., 2019] A Dosovitskiy, P Fischer, E Ilg, P Hausser, C Hazirbas, and V Golkov. & **Brox, T.(2015). Flownet: Learning optical flow with convolutional networks.** In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2019.
- [Dosovitskiy et al., 2015] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. **Flownet: Learning optical flow with convolutional networks.** In *IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.
- [Drozdal et al., 2016] Michal Drozdal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. **The importance of skip connections in biomedical image segmentation.** In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187. Springer, 2016.
- [Endovis, 2015] Endovis. **Instrument.** <https://endovissub-instrument.grand-challenge.org/>, 2015. (Accessed on 12/16/2020).
- [Endovis, 2018] Endovis. **CATARACTS.** <https://cataracts2018.grand-challenge.org/>, 2018. (Accessed on 12/16/2020).
- [EndoVis, 2018] EndoVis. **Robotic Scene Segmentation.** <https://endovissub2018-roboticscenesegmentation.grand-challenge.org/>, 2018. (Accessed on 12/16/2020).
- [EndoVis, 2019] EndoVis. **Workflow and skill.** <https://endovissub-workflowandskill.grand-challenge.org/>, 2019. (Accessed on 12/16/2020).
- [Endovis, 2020] Endovis. **CATARACTS Semantic Segmentation.** <https://cataracts-semantic-segmentation2020.grand-challenge.org/>, 2020. (Accessed on 12/16/2020).
- [Esteva et al., 2021] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. **Deep learning-enabled medical computer vision.** *npj Digital Medicine*, 4(1):5, Jan 2021. ISSN 2398-6352. doi: 10.1038/s41746-020-00376-2. URL <https://doi.org/10.1038/s41746-020-00376-2>.
- [Fang et al., 2019] Hao-Shu Fang, Jianhua Sun, Runzhong Wang, Minghao Gou, Yong-Lu Li, and Cewu Lu. **Instaboost: Boosting instance segmentation via probability map guided copy-pasting.** In *Proceedings of the IEEE International Conference on Computer Vision*, pages 682–691, 2019.
- [Farneback, 2003] Gunnar Farneback. **Two-frame motion estimation based on polynomial expansion.** In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003.
- [Funke et al., 2019] Isabel Funke, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. **Video-based surgical skill assessment using 3D convolutional neural networks.** *International journal of computer assisted radiology and surgery*, 14(7):1217–1225, 2019.
- [Gal et al., 2017] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. **Deep Bayesian active learning with image data.** In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1183–1192, 2017.



- [García-Peraza-Herrera et al., 2016] Luis C García-Peraza-Herrera, Wenqi Li, Caspar Gruijthuijsen, Alain Devreker, George Attilakos, Jan Deprest, Emmanuel Vander Poorten, Danail Stoyanov, Tom Vercauteren, and Sébastien Ourselin. **Real-time segmentation of non-rigid surgical tools based on deep learning and tracking**. In *International Workshop on Computer-Assisted and Robotic Endoscopy*, pages 84–95. Springer, 2016.
- [Garcia-Peraza-Herrera et al., 2017a] Luis C Garcia-Peraza-Herrera, Wenqi Li, Lucas Fidon, Caspar Gruijthuijsen, Alain Devreker, George Attilakos, Jan Deprest, Emmanuel Vander Poorten, Danail Stoyanov, Tom Vercauteren, et al. **Toolnet: holistically-nested real-time segmentation of robotic surgical tools**. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5717–5722. IEEE, 2017a.
- [Garcia-Peraza-Herrera et al., 2017b] Luis Carlos Garcia-Peraza-Herrera, Wenqi Li, Lucas Fidon, Caspar Gruijthuijsen, et al. **ToolNet: Holistically-Nested Real-Time Segmentation of Robotic Surgical Tools**. In *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017b.
- [García-Peraza-Herrera et al., 2016] Luis C. García-Peraza-Herrera, Wenqi Li, Caspar Gruijthuijsen, Alain Devreker, et al. **Real-Time Segmentation of Non-rigid Surgical Tools Based on Deep Learning and Tracking**. In *Computer-Assisted and Robotic Endoscopy*, Lecture Notes in Computer Science. Springer, Cham, October 2016. ISBN 978-3-319-54056-6 978-3-319-54057-3. doi: 10.1007/978-3-319-54057-3\_8.
- [González et al., 2020] Cristina González, Laura Bravo-Sánchez, and Pablo Arbelaz. **ISINet: An Instance-Based Approach for Surgical Instrument Segmentation**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 595–605. Springer, 2020.
- [Goodfellow et al., 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. **Generative adversarial nets**. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [Goodfellow et al., 2016a] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. **Deep Learning**. MIT Press, 2016a.
- [Goodfellow et al., 2016b] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. **Deep learning**, volume 1. MIT press Cambridge, 2016b.
- [Graves and Schmidhuber, 2005] Alex Graves and Jürgen Schmidhuber. **Frameworkwise phoneme classification with bidirectional LSTM and other neural network architectures**. *Neural networks*, 18(5-6):602–610, 2005.
- [Graves et al., 2013] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. **Speech recognition with deep recurrent neural networks**. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.
- [Gu et al., 2020] Yingzhong Gu, Zhe Yu, Ruisheng Diao, and Di Shi. **Doubly-fed Deep Learning Method for Bad Data Identification in Linear State Estimation**. *Journal of Modern Power Systems and Clean Energy*, 8(6):1140–1150, 2020.

- [Gu et al., 2019] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. **CE-Net: context encoder network for 2D medical image segmentation**. *IEEE transactions on medical imaging*, 38(10):2281–2292, 2019.
- [Guo et al., 2019] Tianyu Guo, Chang Xu, Boxin Shi, Chao Xu, and Dacheng Tao. **Learning from bad data via generation**. In *Advances in Neural Information Processing Systems*, pages 6044–6055, 2019.
- [Hafiz and Bhat, 2020] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. **A survey on instance segmentation: state of the art**. *International Journal of Multimedia Information Retrieval*, pages 1–19, 2020.
- [Haidegger, 2019] Tamás Haidegger. **Autonomy for surgical robots: Concepts and paradigms**. *IEEE Transactions on Medical Robotics and Bionics*, 1(2):65–76, 2019.
- [Harrison et al., 2018] Xavier A Harrison, Lynda Donaldson, Maria Eugenia Correa-Cano, Julian Evans, David N Fisher, Cecily ED Goodwin, Beth S Robinson, David J Hodgson, and Richard Inger. **A brief introduction to mixed effects modelling and multi-model inference in ecology**. *PeerJ*, 6:e4794, 2018.
- [Hashimoto et al., 2019] Daniel A Hashimoto, Guy Rosman, Elan R Witkowski, Caitlin Stafford, Allison J Navarette-Welton, David W Rattner, Keith D Lillemoe, Daniela L Rus, and Ozanan R Meireles. **Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy**. *Annals of surgery*, 270(3):414–421, 2019.
- [Hattab et al., 2020] Georges Hattab, Marvin Arnold, Leon Strenger, Max Allan, Darja Arsentjeva, Oliver Gold, Tobias Simpfendorfer, Lena Maier-Hein, and Stefanie Speidel. **Kidney edge detection in laparoscopic image data for computer-assisted surgery**. *International Journal of Computer Assisted Radiology and Surgery*, 15(3):379–387, 2020.
- [He et al., 2016a] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. **Deep residual learning for image recognition**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016a.
- [He et al., 2016b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. **Deep residual learning for image recognition**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016b.
- [He et al., 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. **Mask R-CNN**. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.
- [Heim et al., 2017] Eric Heim, Alexander Seitel, Fabian Isensee, Jonas Andrusis, Christian Stock, Tobias Ross, and Lena Maier-Hein. **Clickstream analysis for crowd-based object segmentation with confidence**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [Heim et al., 2018] Eric Heim, Tobias Roß, Alexander Seitel, Keno März, Bram Stieltjes, Matthias Eisenmann, Johannes Lebert, Jasmin Metzger, Gregor Sommer, Alexander W Sauter, et al. **Large-scale medical image annotation with crowd-powered algorithms**. *Journal of Medical Imaging*, 5(3):034002, 2018.

- [Hervella et al., 2020] Álvaro S Hervella, José Rouco, Jorge Novo, and Marcos Ortega. **Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction.** *Applied Soft Computing*, page 106210, 2020.
- [Hoang et al., 2020] Tuan Hoang, Thanh-Toan Do, Huu Le, Dang-Khoa Le-Tan, and Ngai-Man Cheung. **Simultaneous compression and quantization: A joint approach for efficient unsupervised hashing.** *Computer Vision and Image Understanding*, 191:102852, 2020.
- [Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. **Long short-term memory.** *Neural computation*, 9(8):1735–1780, 1997.
- [Holling and Schmitz, 2010] Heinz Holling and Bernhard Schmitz. **Handbuch Statistik, Methoden und Evaluation.** Hogrefe Verlag, 2010.
- [Holm, 1979] Sture Holm. **A simple sequentially rejective multiple test procedure.** *Scandinavian journal of statistics*, pages 65–70, 1979.
- [Hotelling, 1933] Harold Hotelling. **Analysis of a complex of statistical variables into principal components.** *Journal of educational psychology*, 24(6):417, 1933.
- [Huang et al., 2019a] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. **Mask scoring r-cnn.** In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6409–6418, 2019a.
- [Huang et al., 2019b] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. **Mask scoring r-cnn.** In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6409–6418, 2019b.
- [Huttenlocher et al., 1993] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. **Comparing images using the Hausdorff distance.** *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993.
- [Ilg et al., 2017a] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. **FlowNet 2.0: Evolution of optical flow estimation with deep networks.** In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017a.
- [Ilg et al., 2017b] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. **FlowNet 2.0: Evolution of optical flow estimation with deep networks.** In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017b.
- [Ioannidis, 2005] John PA Ioannidis. **Why most published research findings are false.** *PLoS med*, 2(8):e124, 2005.
- [Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. **Batch normalization: Accelerating deep network training by reducing internal covariate shift.** *arXiv preprint arXiv:1502.03167*, 2015.
- [Isensee and Maier-Hein, 2020] Fabian Isensee and Klaus H Maier-Hein. **OR-UNet: an Optimized Robust Residual U-Net for Instrument Segmentation in Endoscopic Images.** *arXiv preprint arXiv:2004.12668*, 2020.

- [Isensee et al., 2018] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. **nnu-net: Self-adapting framework for u-net-based medical image segmentation**. *arXiv preprint arXiv:1809.10486*, 2018.
- [Islam et al., 2019] Mobarakol Islam, Daniel Anojan Atputharuban, Ravikiran Ramesh, and Hongliang Ren. **Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning**. *IEEE Robotics and Automation Letters*, 4(2):2188–2195, 2019.
- [Jin et al., 2019] Yueming Jin, Keyun Cheng, Qi Dou, and Pheng-Ann Heng. **Incorporating Temporal Prior from Motion Flow for Instrument Segmentation in Minimally Invasive Surgery Video**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 440–448. Springer, 2019.
- [Jing and Tian, 2020] Longlong Jing and Yingli Tian. **Self-supervised visual feature learning with deep neural networks: A survey**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [Kamnitsas et al., 2017] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, et al. **Unsupervised domain adaptation in brain lesion segmentation with adversarial networks**. In *International Conference on Information Processing in Medical Imaging*. Springer, 2017.
- [Karpthy et al., 2016] Andrej Karpathy et al. **Cs231n convolutional neural networks for visual recognition**. *Neural networks*, 1(1), 2016.
- [Keskar and Socher, 2017] Nitish Shirish Keskar and Richard Socher. **Improving generalization performance by switching from adam to sgd**. *arXiv preprint arXiv:1712.07628*, 2017.
- [Kiefer et al., 1952] Jack Kiefer, Jacob Wolfowitz, et al. **Stochastic estimation of the maximum of a regression function**. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.
- [Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. **Adam: A method for stochastic optimization**. *arXiv preprint arXiv:1412.6980*, 2014.
- [Kletz et al., 2019] Sabrina Kletz, Klaus Schoeffmann, Jenny Benois-Pineau, and Heinrich Husslein. **Identifying surgical instruments in laparoscopy using deep learning instance segmentation**. In *IEEE International Conference on Content-Based Multimedia Indexing*, pages 1–6. IEEE, 2019.
- [Kohli et al., 2007] Pushmeet Kohli, M Pawan Kumar, and Philip HS Torr. **P3 & beyond: Solving energies with higher order cliques**. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.
- [Koza et al., 1996] John R Koza, Forrest H Bennett, David Andre, and Martin A Keane. **Automated design of both the topology and sizing of analog electrical circuits using genetic programming**. In *Artificial Intelligence in Design'96*, pages 151–170. Springer, 1996.
- [Krähenbühl and Koltun, 2011] Philipp Krähenbühl and Vladlen Koltun. **Efficient inference in fully connected crfs with gaussian edge potentials**. In *Advances in neural information processing systems*, pages 109–117, 2011.

- [Krause et al., 2013] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. **3d object representations for fine-grained categorization**. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.
- [Kreps, 1989] David M. Kreps. **Nash Equilibrium**, pages 167–177. Palgrave Macmillan UK, London, 1989. ISBN 978-1-349-20181-5. doi: 10.1007/978-1-349-20181-5\_19. URL [https://doi.org/10.1007/978-1-349-20181-5\\_19](https://doi.org/10.1007/978-1-349-20181-5_19).
- [Krizhevsky et al., 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. **Imagenet classification with deep convolutional neural networks**. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [Kuhn, 1955] Harold W Kuhn. **The Hungarian method for the assignment problem**. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.
- [Lafferty et al., 2001] John Lafferty, Andrew McCallum, and Fernando CN Pereira. **Conditional random fields: Probabilistic models for segmenting and labeling sequence data**. 2001.
- [Larsen et al., 2016] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. **Autoencoding beyond pixels using a learned similarity metric**. In *International Conference on Machine Learning*, 2016.
- [Larsson et al., 2017] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. **Colorization as a Proxy Task for Visual Understanding**. *arXiv:1703.04044*, 2017.
- [Law et al., 2017] Hei Law, Khurshid Ghani, and Jia Deng. **Surgeon technical skill assessment using computer vision based analysis**. In *Machine learning for healthcare conference*, pages 88–99, 2017.
- [LeCun et al., 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. **Deep learning**. *Nature*, 521(7553):436–444, 2015.
- [Lee et al., 2019] Eung-Joo Lee, William Plishker, Xinyang Liu, Timothy Kane, Shuvra S Bhattacharyya, and Raj Shekhar. **Segmentation of surgical instruments in laparoscopic videos: training dataset generation and deep-learning-based framework**. In *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10951, page 109511T. International Society for Optics and Photonics, 2019.
- [Lee and Park, 2020] Youngwan Lee and Jongyoul Park. **CenterMask: Real-time anchor-free instance segmentation**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2020.
- [Lei et al., 2018] Suhua Lei, Huan Zhang, Ke Wang, and Zhendong Su. **How Training Data Affect the Accuracy and Robustness of Neural Networks for Image Classification**. 2018.
- [Lin et al., 2019] Shan Lin, Fangbo Qin, Randall A Bly, Kris S Moe, and Blake Hannaford. **Automatic Sinus Surgery Skill Assessment Based on Instrument Segmentation and Tracking in Endoscopic Video**. In *International Workshop on Multiscale Multimodal Medical Imaging*, pages 93–100. Springer, 2019.
- [Lin et al., 2014a] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. **Microsoft COCO: Common objects in context**. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014a.

- [Lin et al., 2014b] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, et al. **Microsoft coco: Common objects in context**. In *European conference on computer vision*. Springer, 2014b.
- [Lin et al., 2017] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. **Feature pyramid networks for object detection**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [Liu et al., 2020] Daochang Liu, Yuhui Wei, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. **Unsupervised Surgical Instrument Segmentation via Anchor Generation and Semantic Diffusion**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 657–667. Springer, 2020.
- [Liu et al., 2016] Tong Liu, Xiutian Huang, and Jianshe Ma. **Conditional random fields for image labeling**. *Mathematical Problems in Engineering*, 2016, 2016.
- [Lucas et al., 1981] Bruce D Lucas, Takeo Kanade, et al. **An iterative image registration technique with an application to stereo vision**. 1981.
- [Maier et al., 2019] Andreas Maier, Christopher Syben, Tobias Lasser, and Christian Riess. **A gentle introduction to deep learning in medical image processing**. *Zeitschrift für Medizinische Physik*, 29(2):86–101, 2019.
- [Maier-Hein et al., 2016] L Maier-Hein, T Ross, J Gröhl, B Glocker, S Bodenstedt, C Stock, E Heim, M Götz, S Wirkert, H Kenngott, et al. **Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 616–623. Springer, 2016.
- [Maier-Hein et al., 2017a] Lena Maier-Hein, Swaroop S Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al. **Surgical data science for next-generation interventions**. *Nature Biomedical Engineering*, 1(9):691, 2017a.
- [Maier-Hein et al., 2017b] Lena Maier-Hein, Swaroop S. Vedula, Stefanie Speidel, Nassir Navab, et al. **Surgical data science for next-generation interventions**. *Nature Biomedical Engineering*, 1(9), September 2017b. ISSN 2157-846X.
- [Maier-Hein et al., 2018] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P Bradley, Aaron Carass, et al. **Why rankings of biomedical image analysis competitions should be interpreted with care**. *Nature communications*, 9(1):5217, 2018.
- [Maier-Hein et al., 2019] Lena Maier-Hein, Annika Reinke, Michal Kozubek, Anne L Martel, Tal Arbel, Matthias Eisenmann, Allan Hanbury, Pierre Jannin, Henning Müller, Sinan Onogur, et al. **BIAS: Transparent reporting of biomedical image analysis challenges**. *arXiv preprint arXiv:1910.04071*, 2019.
- [Maier-Hein et al., 2020a] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, et al. **Surgical Data Science—from Concepts to Clinical Translation**. *arXiv preprint arXiv:2011.02284*, 2020a.

- [Maier-Hein et al., 2020b] Lena Maier-Hein, Martin Wagner, Tobias Ross, Annika Reinke, Sebastian Bodenstedt, Peter M Full, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, et al. **Heidelberg Colorectal Data Set for Surgical Data Science in the Sensor Operating Room**. *arXiv preprint arXiv:2005.03501*, 2020b.
- [Mao et al., 2017] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, et al. **Least squares generative adversarial networks**. In *2017 IEEE Internat. Conference on Computer Vision (ICCV)*. IEEE, 2017.
- [McCulloch and Neuhaus, 2001] Charles E McCulloch and John M Neuhaus. **Generalized linear mixed models**. Wiley Online Library, 2001.
- [Miao et al., 2019] Zhongqi Miao, Kaitlyn M Gaynor, Jiayun Wang, Ziwei Liu, Oliver Muellerklein, Mohammad Sadegh Norouzzadeh, Alex McInturff, Rauri CK Bowie, Ran Nathan, X Yu Stella, et al. **Insights and approaches using deep learning to classify wildlife**. *Scientific reports*, 9(1):1–9, 2019.
- [Milletari et al., 2016] 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)*, pages 565–571. IEEE, 2016.
- [Miotto et al., 2018] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. **Deep learning for healthcare: review, opportunities and challenges**. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.
- [Mwangi et al., 2014] Benson Mwangi, Jair C Soares, and Khader M Hasan. **Visualization and unsupervised predictive clustering of high-dimensional multimodal neuroimaging data**. *Journal of neuroscience methods*, 236:19–25, 2014.
- [Ng and Jordan, 2002] Andrew Y Ng and Michael I Jordan. **On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes**. In *Advances in neural information processing systems*, pages 841–848, 2002.
- [Nguyen et al., 2019] Xuan Anh Nguyen, Damir Ljuhar, Maurizio Pacilli, Ramesh Mark Nataraja, and Sunita Chauhan. **Surgical skill levels: Classification and analysis using deep neural network model and motion signals**. *Computer methods and programs in biomedicine*, 177: 1–8, 2019.
- [Ni et al., 2020] Zhen-Liang Ni, Gui-Bin Bian, Guan-An Wang, Xiao-Hu Zhou, Zeng-Guang Hou, Xiao-Liang Xie, Zhen Li, and Yu-Han Wang. **BARNet: Bilinear Attention Network with Adaptive Receptive Field for Surgical Instrument Segmentation**. *arXiv preprint arXiv:2001.07093*, 2020.
- [Nikolov et al., 2018] Stanislav Nikolov, Sam Blackwell, Ruheena Mendes, Jeffrey De Fauw, Clemens Meyer, Cían Hughes, Harry Askham, Bernardino Romera-Paredes, Alan Karthikesalingam, Carlton Chu, et al. **Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy**. *arXiv preprint arXiv:1809.04430*, 2018.
- [Noroozi and Favaro, 2016] Mehdi Noroozi and Paolo Favaro. **Unsupervised learning of visual representations by solving jigsaw puzzles**. In *European Conference on Computer Vision*. Springer, 2016.

- [Pakhomov et al., 2017] Daniil Pakhomov, Vittal Premachandran, Max Allan, Mahdi Azizian, and Nassir Navab. **Deep Residual Learning for Instrument Segmentation in Robotic Surgery**. *arXiv:1703.08580 [cs]*, March 2017. arXiv: 1703.08580.
- [Pakhomov et al., 2019] Daniil Pakhomov, Vittal Premachandran, Max Allan, Mahdi Azizian, and Nassir Navab. **Deep residual learning for instrument segmentation in robotic surgery**. In *International Workshop on Machine Learning in Medical Imaging*, pages 566–573. Springer, 2019.
- [Pathak et al., 2016] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. **Context encoders: Feature learning by inpainting**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [Pezzementi et al., 2009] Zachary Pezementi, Sandrine Voros, and Gregory D Hager. **Articulated object tracking by rendering consistent appearance parts**. In *2009 IEEE International Conference on Robotics and Automation*, pages 3940–3947. IEEE, 2009.
- [Pham et al., 2020] 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)*, pages 771–783. IEEE, 2020.
- [Prebay et al., 2019] Zachary J Prebay, James O Peabody, David C Miller, and Khurshid R Ghani. **Video review for measuring and improving skill in urological surgery**. *Nature Reviews Urology*, 16(4):261–267, 2019.
- [Ravishankar et al., 2016] Hariharan Ravishankar, Prasad Sudhakar, Rahul Venkataramani, Sheshadri Thiruvankadam, et al. **Understanding the Mechanisms of Deep Transfer Learning for Medical Images**. In *Deep Learning and Data Labeling for Medical Applications*, Lecture Notes in Computer Science. Springer, Cham, October 2016. ISBN 978-3-319-46975-1 978-3-319-46976-8.
- [Reed and MarksII, 1999] Russell Reed and Robert J MarksII. **Neural smithing: supervised learning in feedforward artificial neural networks**. Mit Press, 1999.
- [Reinke et al., 2018] Annika Reinke, Matthias Eisenmann, Sinan Onogur, Marko Stankovic, Patrick Scholz, Peter M Full, Hrvoje Bogunovic, Bennett A Landman, Oskar Maier, Bjoern Menze, et al. **How to exploit weaknesses in biomedical challenge design and organization**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 388–395. Springer, 2018.
- [Ren et al., 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. **Faster r-cnn: Towards real-time object detection with region proposal networks**. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [Robu et al., 2020] Maria Robu, Abdolrahim Kadkhodamohammadi, Imanol Luengo, and Danail Stoyanov. **Towards real-time multiple surgical tool tracking**. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–7, 2020.



- [Ronneberger et al., 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. **U-net: Convolutional networks for biomedical image segmentation**. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [Rosenblatt, 1958] Frank Rosenblatt. **The perceptron: a probabilistic model for information storage and organization in the brain**. *Psychological review*, 65(6):386, 1958.
- [Ross et al., 2018] Tobias Ross, David Zimmerer, Anant Vemuri, Fabian Isensee, Manuel Wiesenfath, Sebastian Bodenstedt, Fabian Both, Philip Kessler, Martin Wagner, Beat Müller, et al. **Exploiting the potential of unlabeled endoscopic video data with self-supervised learning**. *International journal of computer assisted radiology and surgery*, 13(6):925–933, 2018.
- [Rozsa et al., 2016] Andras Rozsa, Manuel Günther, and Terrance E Boutilier. **Are accuracy and robustness correlated**. In *2016 15th IEEE international conference on machine learning and applications (ICMLA)*, pages 227–232. IEEE, 2016.
- [Roß et al., 2020] 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 Wiesenfath, 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 challenge**. *Medical Image Analysis*, page 101920, 2020. ISSN 1361-8415. doi: <https://doi.org/10.1016/j.media.2020.101920>.
- [Russakovsky et al., 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. **Imagenet large scale visual recognition challenge**. *International journal of computer vision*, 115(3):211–252, 2015.
- [Sak et al., 2014] Hasim Sak, Andrew W Senior, and Françoise Beaufays. **Long short-term memory recurrent neural network architectures for large scale acoustic modeling**. 2014.
- [Sánchez et al., 2015] Javier Sánchez, Agustín Salgado, and Nelson Monzón. **Computing inverse optical flow**. *Pattern Recognition Letters*, 52:32–39, 2015.
- [Sarikaya et al., 2017] Duygu Sarikaya, Jason J Corso, and Khurshid A Guru. **Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection**. *IEEE transactions on medical imaging*, 36(7):1542–1549, 2017.
- [Selvin et al., 2017] Sreelekshmy Selvin, R Vinayakumar, EA Gopalakrishnan, Vijay Krishna Menon, and KP Soman. **Stock price prediction using LSTM, RNN and CNN-sliding window model**. In *2017 international conference on advances in computing, communications and informatics (icacci)*, pages 1643–1647. IEEE, 2017.

- [Shankar et al., 2020] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. **Evaluating Machine Accuracy on ImageNet**. In *International Conference on Machine Learning (ICML)*, 2020.
- [Shapiro, 1996] Linda G Shapiro. **Connected component labeling and adjacency graph construction**. In *Machine Intelligence and Pattern Recognition*, volume 19, pages 1–30. Elsevier, 1996.
- [Shen et al., 2017a] Dinggang Shen, Guorong Wu, and Heung-Il Suk. **Deep learning in medical image analysis**. *Annual review of biomedical engineering*, 19:221–248, 2017a.
- [Shen et al., 2017b] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. **Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals**. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1900–1909, 2017b.
- [Shvets et al., 2018] Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. **Automatic instrument segmentation in robot-assisted surgery using deep learning**. In *IEEE International Conference on Machine Learning and Applications*, pages 624–628. IEEE, 2018.
- [Siddaiah-Subramanya et al., 2017] Manjunath Siddaiah-Subramanya, Kor Woi Tiang, and Masimba Nyandowe. **A new era of minimally invasive surgery: progress and development of major technical innovations in general surgery over the last decade**. *The Surgery Journal*, 3(04):e163–e166, 2017.
- [Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. **Very deep convolutional networks for large-scale image recognition**. *arXiv preprint arXiv:1409.1556*, 2014.
- [Simpson et al., 2019] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. **A large annotated medical image dataset for the development and evaluation of segmentation algorithms**. *arXiv preprint arXiv:1902.09063*, 2019.
- [Singh et al., 2015] Pritam Singh, Rajesh Aggarwal, Muaaz Tahir, Philip H Pucher, and Ara Darzi. **A randomized controlled study to evaluate the role of video-based coaching in training laparoscopic skills**. *Annals of surgery*, 261(5):862–869, 2015.
- [Spearman, 1904] Charles Spearman. **The proof and measurement of association between two things**. *American Journal of Psychology*, 15(1):72–101, 1904.
- [Speidel et al., 2006] Stefanie Speidel, Michael Delles, Carsten Gutt, and Rüdiger Dillmann. **Tracking of instruments in minimally invasive surgery for surgical skill analysis**. In *International Workshop on Medical Imaging and Virtual Reality*, pages 148–155. Springer, 2006.
- [Stauder et al., 2016] Ralf Stauder, Daniel Ostler, Michael Kranzfelder, Sebastian Koller, Hubertus Feußner, and Nassir Navab. **The TUM LapChole dataset for the M2CAI 2016 workflow challenge**. *arXiv preprint arXiv:1610.09278*, 2016.
- [Su et al., 2018] Yun-Hsuan Su, Kevin Huang, and Blake Hannaford. **Real-time vision-based surgical tool segmentation with robot kinematics prior**. In *2018 International Symposium on Medical Robotics (ISMR)*, pages 1–6. IEEE, 2018.

- [Sudre et al., 2017] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. **Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations**. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017.
- [Sussillo and Abbott, 2014] David Sussillo and LF Abbott. **Random walk initialization for training very deep feedforward networks**. *arXiv preprint arXiv:1412.6558*, 2014.
- [Sznitman et al., 2012] Raphael Sznitman, Karim Ali, Rogério Richa, Russell H Taylor, Gregory D Hager, and Pascal Fua. **Data-driven visual tracking in retinal microsurgery**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 568–575. Springer, 2012.
- [Tajbakhsh et al., 2016] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, et al. **Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?** *IEEE Transactions on Medical Imaging*, 35(5), May 2016. ISSN 0278-0062. doi: 10.1109/TMI.2016.2535302.
- [Tajbakhsh et al., 2017] Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. **On the Necessity of Fine-Tuned Convolutional Neural Networks for Medical Imaging**. In *Deep Learning and Convolutional Neural Networks for Medical Image Computing*, Advances in Computer Vision and Pattern Recognition. Springer, Cham, 2017. ISBN 978-3-319-42998-4 978-3-319-42999-1. DOI: 10.1007/978-3-319-42999-1\_11.
- [Tajbakhsh et al., 2019] Nima Tajbakhsh, Yufei Hu, Junli Cao, Xingjian Yan, Yi Xiao, Yong Lu, Jianming Liang, Demetri Terzopoulos, and Xiaowei Ding. **Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data**. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1251–1255. IEEE, 2019.
- [Tan and Le, 2019] Mingxing Tan and Quoc Le. **EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks**. In *International Conference on Machine Learning*, pages 6105–6114, 2019.
- [Tellez et al., 2019] David Tellez, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. **Neural image compression for gigapixel histopathology image analysis**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [Tran, 2008] The Truyen Tran. **On conditional random fields: Applications, feature selection, parameter estimation and hierarchical modelling**. Curtin University of Technology, 2008.
- [Twinanda et al., 2017] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, et al. **EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos**. *IEEE Transactions on Medical Imaging*, January 2017. ISSN 0278-0062.
- [Twinanda et al., 2016] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. **Endonet: a deep architecture for recognition tasks on laparoscopic videos**. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.
- [Wang and Deng, 2018] Mei Wang and Weihong Deng. **Deep visual domain adaptation: A survey**. *Neurocomputing*, 312:135–153, 2018.

- [Wang et al., 2017] Rong Wang, Mei Zhang, Xiangbing Meng, Zheng Geng, and Fei-Yue Wang. **3-D Tracking for Augmented Reality Using Combined Region and Dense Cues in Endoscopic Surgery**. *IEEE journal of biomedical and health informatics*, 22(5):1540–1551, 2017.
- [Wiesenfarth et al., 2019a] Manuel Wiesenfarth, Annika Reinke, Bennett A Landman, Manuel Jorge Cardoso, Lena Maier-Hein, and Annette Kopp-Schneider. **challengeR: Methods and open-source toolkit for analyzing and visualizing challenge results**, 2019a. <https://github.com/wiesenfa/challengeR>. Accessed: 2020-02-06.
- [Wiesenfarth et al., 2019b] Manuel Wiesenfarth, Annika Reinke, Bennett A Landman, Manuel Jorge Cardoso, Lena Maier-Hein, and Annette Kopp-Schneider. **Methods and open-source toolkit for analyzing and visualizing challenge results**. *arXiv preprint arXiv:1910.05121*, 2019b.
- [Wilcoxon, 1992] Frank Wilcoxon. **Individual comparisons by ranking methods**. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.
- [Wilson et al., 2017] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. **The marginal value of adaptive gradient methods in machine learning**. In *Advances in neural information processing systems*, pages 4148–4158, 2017.
- [Wu et al., 2019] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. **Wider or deeper: Revisiting the resnet model for visual recognition**. *Pattern Recognition*, 90:119–133, 2019.
- [Xu et al., 2017] Yan Xu, Yang Li, Yipei Wang, Mingyuan Liu, Yubo Fan, Maode Lai, I Eric, and Chao Chang. **Gland instance segmentation using deep multichannel neural networks**. *IEEE Transactions on Biomedical Engineering*, 64(12):2901–2912, 2017.
- [Zamir et al., 2018] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. **Taskonomy: Disentangling task transfer learning**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018.
- [Zhang and Zhang, 2009] Ethan Zhang and Yi Zhang. **Average Precision**, pages 192–193. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9\_482. URL [https://doi.org/10.1007/978-0-387-39940-9\\_482](https://doi.org/10.1007/978-0-387-39940-9_482).
- [Zhang et al., 2017a] Pengyue Zhang, Fusheng Wang, and Yefeng Zheng. **Self supervised deep representation learning for fine-grained body part recognition**. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 578–582. IEEE, 2017a.
- [Zhang et al., 2017b] R. Zhang, P. Isola, and A. A. Efros. **Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction**. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017b.
- [Zhang et al., 2016] Richard Zhang, Phillip Isola, and Alexei A Efros. **Colorful image colorization**. In *European Conference on Computer Vision*. Springer, 2016.
- [Zhao et al., 2018] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. **lcnet for real-time semantic segmentation on high-resolution images**. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–420, 2018.

- [Zhao et al., 2020] Zixu Zhao, Yueming Jin, Xiaojie Gao, Qi Dou, and Pheng-Ann Heng. **Learning Motion Flows for Semi-supervised Instrument Segmentation from Robotic Surgical Video.** In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer, 2020.
- [Zhou and Payandeh, 2014] Jiawei Zhou and Shahram Payandeh. **Visual tracking of laparoscopic instruments.** *Journal of Automation and Control Engineering Vol, 2(3):234–241*, 2014.
- [Zhou et al., 2017] Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, et al. **Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally.** In *IEEE conference on computer vision and pattern recognition, Hawaii*, 2017.
- [Zhu et al., 2019] Yixing Zhu, Chixiang Ma, and Jun Du. **Rotated cascade R-CNN: A shape robust detector with coordinate regression.** *Pattern Recognition*, 96:106964, 2019.



# 7

## Own contributions

In this chapter, I want to list my contributions and differentiate them from whole team efforts. Every section already discloses the team members that have been working on the presented results.

### 7.1 Own share in data acquisition and analysis

Throughout my Ph.D. work, I was part of a multi-disciplinary team of scientists in the department of Computer Assisted Medical Interventions (CAMI) headed by Prof. Dr. Lena Maier-Hein. As it is common in larger research groups, many ideas on methods, experiments, or implementation details were exchanged in the entire CAMI team. Because Prof. Dr. Lena Maier-Hein was the primary supervisor for my thesis, and she supervised all experiments and was included in the discussion and development of all concepts and methods.

**Potential of unlabeled data** The section *Potential of unlabeled data* is based on the related publication Ross et al. [Ross et al., 2018]. The concept for this section was developed closely together with David Zimmerer, who has often given advice during the implementation phase. However the implementation, evaluation and plot generation was all done by myself.

**Quality controlled dataset generation** The data for the section *Quality controlled dataset generation* was recorded during daily routine by Dr. Martin Wagner and Prof. Beat Müller-Stich. In addition, Dr. Martin Wagner was responsible for phase annotations, which were the basis for the frame extraction. He was further a medical expert who performed the instrument segmentations' quality control. The concept for the frame extraction was developed by all the mentioned people together. The annotations of the instruments were generated with the help of the entire CAMI group. However, the instrument annotation Graphical User Interface (GUI), the labeling protocol, the artifacts annotation GUI and the quality control was all done by myself. Parts of this section are under minor revision for the related publication Maier-Hein et al. [Maier-Hein et al., 2020b].

**comparative validation of multi-instance instrument segmentation** The section *comparative validation of multi-instance instrument segmentation* is based on the related publication Roß et al. [Roß et al., 2020]. The concept for the challenge was developed in close cooperation with Annika Reinke. Further, Annika Reinke helped with organizational aspects and performed the participants' rankings with the corresponding plots. However, the complete technical setup and support of the challenge, evaluation scripts, and the generation all plots (excluding the participants' rankings) was done by myself.

**Effects of image characteristics on the algorithm performance** The concept for the effects of image characteristics was developed in cooperation with Dr. Manuel Wiesenfarth and Prof. Dr. Annette Kopp-Schneider. They helped with the selection of the correct statistical tools for the analysis. However, the implementation of all methods, evaluation and analysis was all done by myself.

**Multi-instance instrument segmentation** The concept for the multi-instance instrument segmentation approach was developed in close cooperation with Pierangela Bruno, Dasha Trofimova and Patrick Scholz. As it is usual for larger projects, the implementation of the concept has been split up. Pierangela Bruno implemented the Mask R-CNN and Dasha Trofimova the LSTM. The overall pipeline and the post-processing was implemented by myself, as well as the statistical analysis and evaluation.



## 7.2 Own publications

This section is about all papers I was part and contributed to during my time in the Department of Computer Assisted Medical Interventions. For clarity, it is subdivided into first and co-authorships. To make it easier to read, the first author(s) are underlined and my name is highlighted in bold.

### First authorships

**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 challenge*. Medical Image Analysis, 2020

**Tobias Ross**, David Zimmerer, Anant Vemuri, Fabian Isensee, Manuel Wiesenfarth, Sebastian Bodenstedt, Fabian Both, Philip Kessler, Martin Wagner, Beat Müller, Hannes Kenngott, Stefanie Speidel, Annette Kopp-Schneider, Klaus Maier-Hein and Lena Maier-Hein. *Exploiting the potential of unlabeled endoscopic video data with self-supervised learning*. International journal of computer assisted radiology and surgery, 13(6):925-933, 2018.

Lena Maier-Hein, **Tobias Ross**, Janek Gröhl, Ben Glocker, Sebastian Bodenstedt, Christian Stock, Eric Heim, Michael Götz, Sebastian Wirkert, Hannes Kenngott, Stefanie Speidel, Klaus Maier-Hein. *Crowd-Algorithm Collaboration for Large-Scale Endoscopic Image Annotation with Confidence*. International Conference on Medical Image Computing and Computer-Assisted Intervention, 616-623, 2016

### Co-Authorships

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 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 Apit. z, Stefanie Speidel, Hannes G. Kenngott and Beat P. Müller-Stich. *Heidelberg Colorectal Data Set for Surgical Data Science in the Sensor Operating Room*. Submitted to Nature Scientific data, 2020

Micha Pfeiffer, Isabel Funke, Maria R Robu, Sebastian Bodenstedt, Leon Strenger, Sandy Engelhardt, **Tobias Ross**, Matthew J Clarkson, Kurinchi Gurusamy, Brian R Davidson, Lena Maier-Hein, Carina Riediger, Thilo Welsch, Jürgen Weitz and Stefanie Speidel. *Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation*. International Conference on Medical Image Computing and Computer-Assisted Intervention, 119-127, 2019

Eric Heim, **Tobias Ross**, Alexander Seitel, Keno März, Bram Stieltjes, Matthias Eisenmann, Johannes Lebert, Jasmin Metzger, Gregor Sommer, Alexander W Sauter, Fides Regina Schwartz, Andreas Termer, Felix Wagner, Hannes Götz Kenngott and Lena Maier-Hein. *Large-scale medical image annotation with crowd-powered algorithms*. Journal of Medical Imaging, 5(3):034002, 2018

Eric Heim, Alexander Seitel, Fabian Isensee, Jonas Andrulis, Christian Stock, **Tobias Ross** and Lena Maier-Hein. *Clickstream analysis for crowd-based object segmentation with confidence*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017

Sebastian Bittel, Veith Roethlingshoefer, Hannes Kenngott, Martin Wagner, Sebastian Bodenstedt, **Tobias Ross**, Stefanie Speidel and Lena Meier-Hein. *How to create the largest in-vivo endoscopic dataset*. Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, 2017

Eric Heim, Alexander Seitel, Christian Stock, **Tobias Ross** and Lena Maier-Hein. *Clickstream-analyse zur Qualitätssicherung in der crowdbasierten Bildsegmentierung*. Bildverarbeitung für die Medizin :17-17, 2017

Lena Maier-Hein, Daniel Kondermann, **Tobias Ross**, Sven Mersmann, Eric Heim, Sebastian Bodenstedt, Hannes Götz Kenngott, Alexandro Sanchez, Martin Wagner, Anas Preukschas, Anna-Laura Wekerle, Stefanie Helfert, Keno März, Arianeb Mehrabi, Stefanie Speidel and Christian Stock. *Crowdtruth validation: a new paradigm for validating algorithms that rely on image correspondences*. International journal of computer assisted radiology and surgery, 10(8): 1201-1212, 2015

# TOBIAS ROSS

## PERSONAL DATA

---

PLACE AND DATE OF BIRTH:	Bad Neustadt a. d. Saale, Germany   26 May 1990
ADDRESS:	Meyershofweg 3, 49086 Osnabrück, Germany
PHONE:	+49 173 3946278
MARITAL STATUS:	married, 1 child
NATIONALITY:	German
EMAIL:	<a href="mailto:tobias_ross@gmx.de">tobias_ross@gmx.de</a>

## EDUCATION

---

DEC. 2017 - NOW	Ph.D. student at GERMAN CANCER RESEARCH CENTER (DKFZ) <i>Department of Computer Assisted Medical Interventions</i>
MAR. 2014 - FEB. 2017	M.Sc. Medical Informatics <i>University Heidelberg &amp; Heilbronn</i>
SEP. 2010 - JUN. 2013	B.Sc. Medical Computer Science <i>University of Applied Science Brandenburg</i>
SEP. 2007 - JUL. 2009	Technical Diploma <i>Fachoberschule Bad Neustadt an der Saale</i>

## RESEARCH EXPERIENCE

---

DEC. 2015 - NOW	Ph.D. student at GERMAN CANCER RESEARCH CENTER (DKFZ) <i>Department of Computer Assisted Medical Interventions</i>
JUL. 2014 - DEC. 2015	Research assistant at GERMAN CANCER RESEARCH CENTER (DKFZ) <i>Department of Computer Assisted Medical Interventions</i>
APR. 2012 - MAR. 2013	Research assistant at OPEN.SC. CHARITÉ <i>Institute for Pathologie, Open European Nephrology Science Center Charité Berlin</i>

## HONORS AND AWARDS

---

JUN. 2018	Best Presentation Award Runner-up (9 <sup>th</sup> International Conference on Information Processing in Computer-Assisted Interventions Berlin)
JUN. 2012	Special commitment to improve the quality of teaching
JUN. 2009	Price of the information group of Business Rhoen

## TEACHING EXPERIENCE

---

SINCE 2018	Supervision of research assistants and master's theses
MAR. 2012 - DEC. 2012	Tutor for PROGRAMMING I AND II

## ACADEMIC ACTIVITIES

---

REVIEWER	Journal: Medical Image Analysis Journal: IEEE Robotics and Automation Letters Journal: International journal of computer assisted radiology and surgery Journal: Medical & Biological Engineering & Computing Challenges: MICCAI 2020
INVITED TALKS	ISC High Performance conference 2019

## WORK EXPERIENCE

---

JUN. 2013 - DEC. 2013	Project manager at RHOEN-KLINIKUM AG <i>Groupwide IT, Project manager for the research projects "SemanticVoice" and "Cloud4Health"</i>
Nov. 2009 - AUG. 2010	Service in lieu of military service at MUTTER-KUR-HAUS, Bad Königshofen

# Acknowledgements

The total of 5 years that I have been in the *Department of Computer Assisted Medical Interventions* (CAMI) has been one of the most exciting and eventful years of my life. During this time, I was able to work in a young, diverse, and dynamic team on exciting new and future-oriented projects. Further, I could meet incredibly interesting people from the most diverse areas of computer science and health care during very interesting conferences and meetings. From the bottom of my heart, I am grateful that I was able to have these experiences. On my doctoral journey, I was accompanied by many people, and it is unfortunately not possible for me to thank every single one of them, but I would like to list the people who have supported me the most.

I want to start with my Ph.D. supervisor and mentor, Prof. Dr. **Lena Maier-Hein** for all her professional and personal guidance. Her personal commitment and dedication to research were a great inspiration, her advice and patience an important guide, and her scientific skills incredibly impressive. Thank you for tutoring me for all these years. It was a great pleasure for me to accompany you over the years, starting as a student in your lectures, then as a HiWi, continuing as a master's student, and finally as a doctoral student. One statement from you will lead me for the rest of my life in the choice of my professional activity, namely the enthusiasm for work, summarized in the statement: "When I look at the clock in the evening, I usually say *oh dear, it is already so late* instead of *oh no, still so much time left*".

Many thanks to the members of my *thesis advisory committee*: Prof. Dr. **Stefanie Speidel**, Prof. Dr. **Klaus Maier-Hein** and Prof. Dr. **Beat Müller-Stich**, who were involved at various stages of the project and offered great methodological and general advice for the project.

Thank you, to **all members of the CAMIC group(s)**, for a great time, the fantastic team spirit, and your help while annotating the data. Thank you for all the discussions, for all the support, and the fun we had on retreats and on work. You were the reason why I decided to continue after my Master's, and I never regret it. Many thanks to the great people in the **Deep Fridge** for their ideas and fruitful discussions. It was always fun and a pleasure to work with you guys in the same office.

Thank you, **Annika Reinke**, **David Zimmerer** and **Pierangela Bruno** for the fantastic teamwork. It was a pleasure to work and publish with you together. Thank you also **Keno März** and **Alexander Seitel** for all your support during my time in CAMI and for your time to proofread my work. Thank you, **Dasha Trofimova** and **Patrick Scholz** for supporting me in the stressful time. You really helped me to get work done and back on track.

Many thanks goes also to my family, **Bernhard, Christoph, Peter** and **Heike** Roß for their support on my way, for reading my thesis, and for all discussions. Special thanks goes to my father and mother who raised me to be always be curious and helped me develop the enthusiasm for science. I am grateful to have such a lovely family, I love you guys.

Finally, I would like to thank my fantastic wife **Rawan Al Alawi**. I cannot put into words how grateful I am for your support through all these difficult times. Without you, none of this would have been possible. You have done everything for our little family and me for a very long time so that I was able to complete this work. For hours you endured me when I was completely stressed or totally depressed, built me up when the results weren't as I hoped, and you took care of me when I totally forgot which day and daytime we currently have. You sacrificed yourself completely and gave everything you could to keep all everyday problems away from me so that I could write in peace. Thank you for always believing in me and thank you for many, many, many more. I couldn't ask for a better partner by my side. Ba7bak kter ya galbe (Arabic: I love you very much, my heart)! Thank you also my handsome **Nael Roß**, for being such a wonderful son. I love you and I can not wait to see what the future holds for us!

# Eidesstattliche Versicherung

## Statutory Declaration

1. Bei der eingereichten Dissertation zu dem Thema **Surgical data science in endoscopic surgery** handelt es sich um meine eigenständig erbrachte Leistung.  
*I herewith formally declare that I have written the submitted dissertation **Surgical data science in endoscopic surgery** independently.*
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.  
*I did not use any third party support except for the quoted literature and other sources mentioned in the text. Content from other work, either literally or in content, has been declared as such.*
3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.  
*The thesis has not been submitted to any examination body in this, or similar, form.*
4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.  
*I confirm the correctness of the aforementioned declarations.*
5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.  
*I am aware of the legal consequences of this declaration. To the best of my knowledge I have told the pure truth and not concealed anything.*

Heidelberg, 13.12.2021

---

Tobias Roß