INAUGURAL – DISSERTATION

zur
Erlangung der Doktorwürde
der
Gesamtfakultät für Mathematik, Ingenieur- und
Naturwissenschaften
der
Ruprecht – Karls – Universität
Heidelberg

vorgelegt von

Zimmerer, David, M.Sc.
aus Aschaffenburg, Deutschland

Tag der mündlichen Prüfung : .....

Unsupervised Learning for Anomaly Detection in Medical Images

Betreuer: Prof. Dr. Klaus H. Maier-Hein

Abstract

Anomaly detection and localization can learn what data looks like and point out anomalous data samples, which may then be utilized to assist clinicians in identifying anomalies. We employ a Variational Autoencoder (VAE) to learn the distribution of the data and demonstrate several ways for highlighting abnormalities. We show that using self-supervised learning and hierarchical representations can increase performance, especially in situations with smaller and more difficult-to-detect cases. We further investigate the approaches' performance and assessment in two contexts: an international public competitive setting and a real-world use-case for discovering incidental findings in a population study. Overall, the results are encouraging, and the algorithms can detect anomalies and incidental findings, but they fall short in more complex and difficult cases and are not yet dependable enough for real-world usage.

Zusammenfassung

Mit Anomalie-Erkennung und Lokalisierung kann man die Verteilung von Daten lernen und dann abnormale Daten erkennen und damit Ärtze bei der Identifkation von Krankheiten und abnormalen Konditionen unterstützen. Wir benutzen einen Variational Autoencoder (VAE) um diese Verteilung der Daten zu lernen und presentieren verschiedene Methoden wie man mit einem VAE Anomalien aufzeigen kann. Wir zeigen, dass hierarchiche Representationen oder Representationen die via self-supervied learning gelernt wurden die Performance verbessern können, insbesodere für die kleineren und schwierigeren Anomalien. Wir untersuchen die Analysen, Evaluierung und Benchmarking der Methoden: In einem internationalen und öffentlichen Wettbewerb und einem realitätsnahem Anwendungsfall für die Identifkation von Krankheiten und abnormalen Konditionen in einer Populations Studie. Insgesamt sind die Ergebnisse gut und die Algorithmen können Anomalien und abnormale Konditionen identifizieren, aber sind jedoch noch nicht zuverlässig genung für einen Einsatz in der täglichen Praxis.

Acknowledgement

First, I would like to say that I am grateful for the chance and time to work on this topic and thesis. Being like 'a child on a playground' and being able to play with and explore different ideas was most of the time a truly fun experience and has led me to often use quotes when talking about "working" for my PhD thesis because it often enough did not feel like work.
A big part of that was thanks to my supervisor Klaus Maier-Hein, and I would like to thank him for giving his guidance and support and giving me enough space to follow my own ideas and make my own experiences. And somehow he managed to turn this overeager try-it-all-out everything-is-shiny engineering-focused student/puppy starting his PhD at least somewhat into a scientist and researcher (at least that is what it feels like right now from my perspective, but give it another 10 years...). And also a big thanks to Klaus for making the workplace a fun place to work in and giving 110% to address all the concerns we had.
And I would also like to thank CAMIC/SYMIC for all the input, ideas, support, discussions, community days, and fun times. And also the lunch and coffee breaks, the 'jungle' office banter, the games, sports, and fun activities #TeamTrixi and for still communicating and meeting (as far as was allowed) despite home office.
I also want to thank anyone who swam, biked, ran, ... with me and helped my head to take a breather, helped blow my head through with (and/or sometimes deprive my head of) oxygen, and stay sane.
And I would like to thank my family for their support, motivation, and values, and without them, I wouldn't have gotten to this point.
Lastly, I would like to thank anyone who takes the time to read this thesis (and for making the time I spent writing this more worthwhile) and an extra thanks if you found the "invisible gorillas" ;D.

Contents

Acronyms

AE	Autoencoder
AP	Average Precision
AUC	Area Under a Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BET	Brain Extraction
CE	Context-Encoding
ceVAE	Context-encoding Variational Autoencoder
chVAE	conditional hierarchical VAE
CNN	Convolutional Neural Networks
CRADL	Contrastive Representations for unsupervised Anomaly Detection and Localization
DAE	Denoising Autoencoder
DSC	Dice Similarity Coefficient
ELBO	Evidence Lower Bound
EM	Expectation-Maximization
FN	False Negative
FP	False Positive
FPI	Foreign Patch Interpolation

FPR	False Positive Rate
FPR@0.95TPR	False Positive Rate at 95% True Positive Rate
GMM	Gaussian Mixture Model
HCP Synth.	HCP synthetic anomaly dataset
IF	Incidental Finding
INN	Invertible Neural Network
irVAE	iterative image restoration Variational Autoencoder
KL	Kullback-Leibler
LSPS	Large Scale Population Study
MC	Monte Carlo
MLE	Maximum Likelihood Estimation
MLP	Multi-Layer Perceptron
MOOD	The Medical Out-Of-Distribution Analysis Challenge
MSE	Mean Squared Error
NF	Normalizing Flow
NLL	Negative-Log-Likelihood
NN	Neural Network
NT-Xent	Normalized Temperature scaled cross-entropy
OC-SVM	One Class Support Vector Machine
OoD	Out-of-Distribution
PCA	Principal Component Analysis
pchVAE	primary components conditional hierarchical VAE
PR	Precision-Recall

refIF	reported reference Incidental Finding
ROC	Receiver Operating Characteristic
SGD	Stochastic Gradient Descent
SimCLR	Simple framework for Contrastive Learning of visual Representations
TN	True Negative
TP	True Positive
TPR	True Positive Rate
VAE	Variational Autoencoder
VQVAE	Vector Quantized Variational Autoencoder

List of Figures

List of Tables

1 Introduction

The technological advancements in the field of computer science and engineering have made the recording of data easier, fast, and more affordable. Furthermore, the storage and processing options of data have made it feasible to store, process, and access data in a more efficient way. For example, natural images and videos can easily be recorded using a smartphone camera and so e.g. every minute >500h of video data is uploaded to YouTube [Statista, 2022]. Similarly for medical imaging data, the progress, affordability, and availability of scanners has led to a steady increase of image data every year by 1%-10% [Smith-Bindman et al., 2019].
However, to draw value from this increased data volume, it usually has to be processed or analyzed. This has led to a 4-fold increase in workload for on-call radiologists [Bruls and Kwee, 2020]. In such high-pressure and time-sensitive situations, miss rates of diagnoses of up to 80% of on-call radiologists have been reported [Pinto et al., 2016]. This is worsened by the fact that such diagnostic errors cause a fatal outcome twice as often as any other medical error [Saber Tehrani et al., 2013]. Furthermore, the increase of data has also led to an increase in 'dark' data, i.e. data that was recorded and stored but not further processed or analyzed and is not used, findable or usable in any way [Haque et al., 2020].
One way to aid this process is to use an automated data processing and analysis pipeline. While expert knowledge-based automated systems have been used for a long time, the current increase in data has allowed for more data-centric systems. These data-driven systems are not manually designed but rather use algorithms that can learn from the labeled data and predict values for new data samples not based on hard-coded rules but from knowledge obtained using the labeled data

samples.

The increase in data volume, processing, and algorithmic progress has led to the successful application of data-driven systems in many fields. As the research progressed, this made it feasible to lift some constraints regarding the applicability in real-life settings and extend them to increasingly more complex and general domains. This progression is nicely laid out in the field of reinforcement learning in games. First, an expert- and rule-based search tree algorithm was used to challenge the chess champion, Gary Kasparov, in 1997. While chess has a 'limited' number of moves, solely the increase of computing power could steadily improve the chess engines way beyond human potential. However, the game of Go, being an exponentially harder problem with a larger number of moves, was hailed as impossible to solve using these conventional methods (and algorithms were not expected to reach expert human-level performance in this decade or the next one). But later, in 2016, AlphaGo was able to beat the record world champion in the game of Go, Lee Sedol. The used algorithm primarily drifted from expert-based knowledge to a learned value function using reinforcement learning. This continued with AlphaZero which, using self-play, was completely independent of expert domain knowledge and was able to outperform AlphaGo as well as current top-performing Go and chess engines [Silver et al., 2018]. However, in chess and Go the space of possible states and moves is discrete and limited and the whole environment is observable. Tackling the challenge of a non-observable environment, a team from the CMU (Libratus) took on the game of poker and was able to outperform the world's best players [Brown and Sandholm, 2018]. Later on, Deepmind and OpenAI extended algorithms to a continuous state and action space and were able to challenge human professional players in computer games: Startcaft2 and Dota2 (in a constraint setting) respectively [OpenAI et al., 2019; Vinyals et al., 2019]. This success was possible, particularly in games, due to the increase in computing power, the availability of data (and the possibility to collect large amounts of data), and methodological advancements.

Similar to games, the progress also made it feasible to extend the use of data-driven systems to other 'open' domains. For example, language models are applied in search and as programming support [Brown et al., 2020]. The protein-folding problem is considered solved [Jumper et al., 2021], voice assistances have found their application in daily life, and robotic control, image-based dog-breed classification or 10,000 species recognition have made great progress.

However, the impact on daily life is still limited and most applications currently struggle to extend toward real "open world" problems. For example, fully autonomous driving, which was claimed to be solved in the past decade, is still not readily deployed and usable. Furthermore, automated medical systems, which

(a) (b)

Figure 1.1: (a) Failure cases of a model trained on a city dataset. It wrongly classifies animals as pedestrians [Anomaly Detection for Scientific Discovery, 2022; Xuefeng et al., 2022a]. (b) Failure case of an organ segmentation model. The first image shows the input image and the second the segmentation. Here, the contrast agent is wrongly classified as a bone.

showed superhuman and very promising results, have struggled to be deployed and not shown an added value [Heaven, 2020]. One factor hindering the broader application of data-driven systems in the real world can be the domain shift between the training data and the real world data, e.g. including very rare and unexpected events, not or only barely available in the training data, which can cause catastrophic failure. Making the data-driven systems detect these unexpected and abnormal events and let them 'know what they don't know' and be fail-safe is a key challenge for the future. An example of such a failure can be seen in Fig. 1.1. In the first example, deer and cattle were detected but recognized as pedestrians (which is at least not a complete failure) because the training data set did not contain animals and thus the trained model is not expecting 'deer'. In the next example, a state-of-the-art organ segmentation model is applied to a real-world colonoscopy data set with a contrast agent, in which the contrast agent is wrongly detected as bone. In such cases where the model encountered something that it was not expecting on a segmentation, object, or image level, a reliable report of "not-knowing" would be very valuable and potentially essential in real-world practice (such a 'safer failure' example can be seen in Fig. 1.2).

Interestingly, not only machine learning algorithms are vulnerable to unexpected input. Also humans, including experts in their fields, can be vulnerable to unexpected events. This often termed 'inattentional blindness', is also prominent in trained radiologists. In a study by Drew et al. [2013], the authors found that during the assessment of a lung CT for lung nodules more than 50% of expert radiologists failed to notice the presence of a gorilla image rendered in the CT (see Fig. 1.3).

Figure 1.2: Examplary results from an improved model, which 'knows what it doesn't know' and is more fail-safe. In particular, the model is able to label some of the wrongly detected objects as OOD [Xuefeng et al., 2022b].

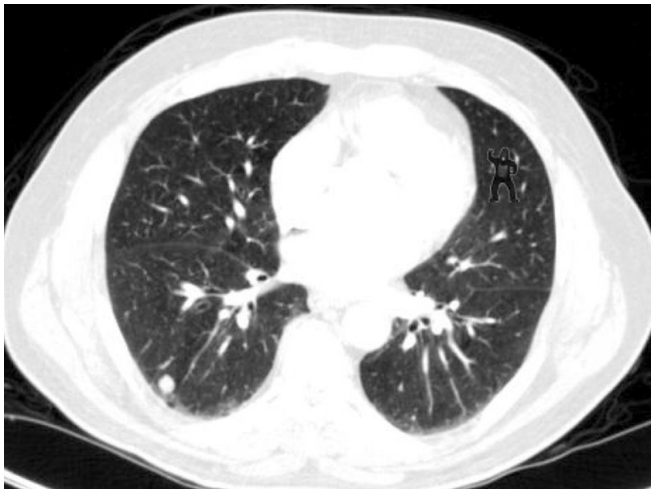


Figure 1.3: A gorilla rendered into a lung CT. 50% of trained radiologists did not notice the gorilla when assessing the scan for lung nodules [Drew et al., 2013].

Hence, an algorithm that can point out an unexpected or abnormal input could help (a) machine learning systems and (b) humans in practice: (a) when supporting machine learning systems, the detection of these distribution shifts, Out-of-Distribution (OoD) samples, or abnormal samples, can point to cases where conventional algorithms might fail and thus call for human intervention and make machine learning systems more trustable and reliable (please note that in this thesis we will use OoD data, abnormal samples, and anomaly interchangeably). (b)

When supporting humans, algorithms that can learn the 'normality' or distribution of healthy patient data samples can aid the detection of unexpected conditions which might be overlooked. This can be implemented as a case-wise ranking by normality and help prioritize and filter the data samples and thus reduce the workload, particularly in a high-throughput setting.
In contrast to 'classical' machine learning algorithms which are tailored to specific classes/conditions, an anomaly detection algorithm should not just identify abnormal samples used during model development but also extend and generalize to arbitrary abnormal samples. So a 'general' anomaly detection algorithm should be able to detect any kind of abnormal input and not just the ones that the algorithm was developed on and implicitly designed for. But when developing an algorithm that can expect, detect, and handle the unexpected or 'know what it doesn't know', using abnormal test samples for validation and testing during the development process can lead to an unwanted and implicit specialization towards these specific abnormal samples. To notice, handle and avoid leaks of the test samples during the development, it is particularly important to have a thorough evaluation of the method and also to use independent test cases that were not known during the development of the algorithm for a final evaluation.
So the aim of this thesis is to develop and improve algorithms that can detect and localize abnormal input, with a focus on a medical setting. However, since a thorough evaluation is similarly important we will also focus on the benchmarking and real-world evaluation of such systems.
In particular, in the first part, we will show how a VAE can be used to detect and localize abnormal inputs. And we will try to improve the performance in particularly difficult settings, where these systems often struggle. In the second part, we will focus on the benchmarking and evaluation of such systems, first in an international challenge setting, and second in a real-world clinical use case. Finally, we will point out the potential and possible pitfalls of such systems (and whether some assumptions made will hold in general and how they potentially influence the performance).

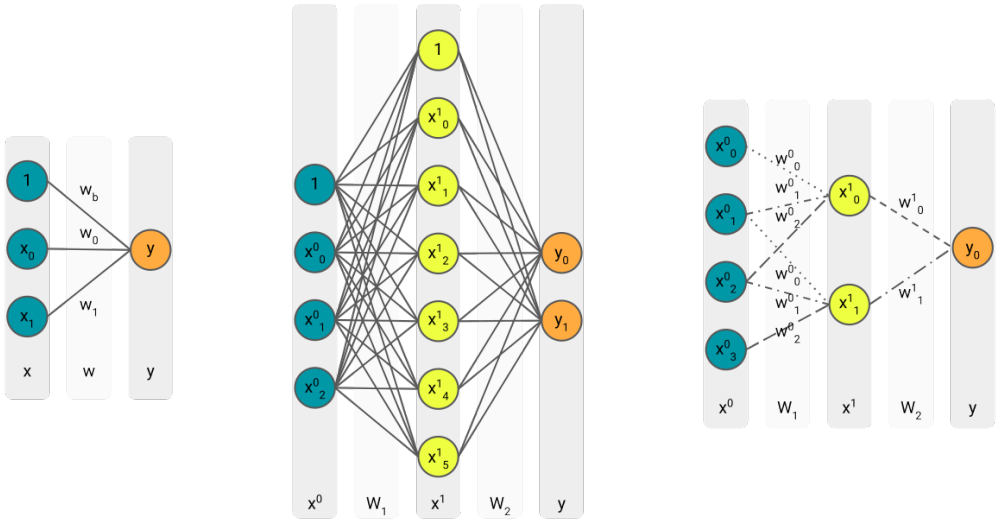
2 | Background

2.1 Neural Networks

Neural Networks (NNs) are a powerful tool for learning & predicting data. NNs were among the first machine learning methods to be proposed and originated with a biologically inspired approach in mind. In contrast to other machine learning techniques whose main focus was parameter efficiency by means of Occams Razor or VC dimensions, NNs have found great success by a large overparameterization [Zhang et al., 2017]. Despite not yet completely understanding the underlying theory of generalization, this has led NNs to be still a very powerful tool for learning and predicting especially in high-dimension domains [Goodfellow et al., 2016].

2.1.1 Perceptron

The Perceptron, also known as Rosenblatt Perceptron, was one of the first machine learning algorithms and is inspired by natural neurons [Goodfellow et al., 2016]. Given input signals/values composed in an input vector x, the Perceptron classifies the input given some predetermined/learned weights w and a threshold function f into two classes: -1 and 1: $y = f(x^T w)$ (see Fig. 2.1 for an illustration).
This allows for a linear separation of the input space in two disjunct classes. Given some predetermined labeled samples x_i, y_i, the weights w can be determined to optimize a certain risk or loss.



(a) Perceptron (b) MLP (c) CNN

Figure 2.1: Schematic representation of (a) a Perceptron, (b) a MLP and (c) a CNN.

One of the most common ways to determine the weights is to use the Perceptron Learning Algorithm. Here, the loss function is defined as:

$$L = \sum_{i=1}^{N} y_i(x_i^T w) \tag{2.1}$$

The Gradient Descent Algorithm can be used to determine the weights w, which minimizes the loss function and gives a local minimum of the loss function (given the right step size).
Taking the derivative of the loss function with respect to the weights w gives:

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{N} y_i x_i. \tag{2.2}$$

Now the weights can be updated iteratively by the following formula:

$$w_{t+1} = w_t - \eta \frac{\partial L}{\partial w} = w_t - \eta \sum_{i=1}^{N} y_i x_i, \tag{2.3}$$

with step size η.

The obtained weights that maximize the loss function for a convex loss function reach the global optimum with the correct step size (and/or step size annealing) and, as is here the case, reach the local optimum for the non-convex loss function. However, calculating an update step over the whole dataset is not always possible. Thus, one solution is Stochastic Gradient Descent (SGD) which is a variant of Gradient Descent where the weights are updated only for a subset of the dataset. This is done by randomly selecting a subset of the dataset and updating the weights for this subset:

$$w_{t+1} = w_t - \eta y_i x_i, \tag{2.4}$$

for a random sample x_i and label y_i.

2.1.2 Multi-Layer Perceptron

A single Perceptron only allows for a linear separation in the input space into two disjunct classes. However, a single Perceptron can be extended to a Multi-Layer Perceptron (MLP) by adding an additional layer of neurons in a consecutive fashion (see Fig. 2.1 for an illustration). This allows for a non-linear separation in the input space. Theoretically, with only three layers and an infinite number of neurons in the second, so-called, hidden layer, the MLP can express any function in the input space and can be seen as "universal function approximator".
Similar to the Perceptron, the weights/parameters of the MLP can be optimized by the Gradient Descent Algorithm. Using the chain rule, the derivative of the loss function with respect to the weights can easily be determined. As intermediate values for each layer can be used for the "earlier" layers, this is often termed a "backpropagation" algorithm.

2.1.3 Convolutional Neural Networks

Most natural data, e.g. images, is not completely random but contains repetitive patterns and self-similarities. One very successful and parameter-efficient way to utilize this are Convolutional Neural Networkss (CNNs). CNNs use the convolution (or to be correct, in most implementations cross-correlation) of the input signal with a filter matrix to generate a new output (see Fig. 2.1 for an illustration). This can increase the parameter efficiency and makes the CNN translation invariant to the input signal. Hence CNNs have had great success and become the defacto standard on image analysis tasks [Krizhevsky et al., 2012].

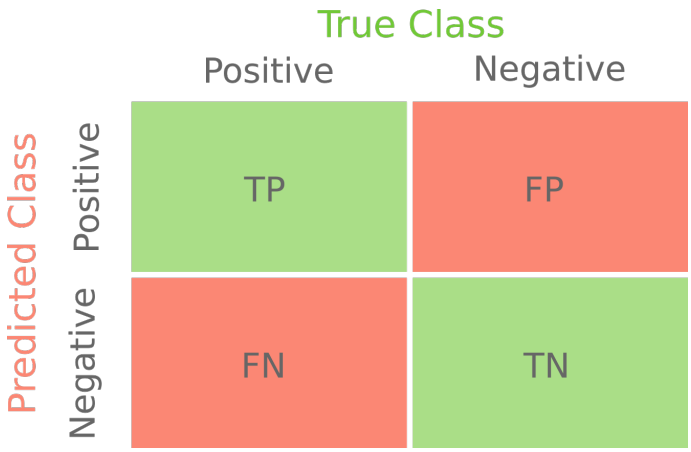


Figure 2.2: Confusion matrix detailing TPs, TNs, FPs, and FNs.

2.2 Metrics

2.2.1 Classification metrics

Metrics are used to evaluate the performance of a model. Different metrics can show, analyze and stress different properties of the model and outline strengths and weaknesses.
In (binary) classification, the predictions of the model can be (given the reference labels) categorized into four categories (as visualized in the confusion matrix in Fig. 2.2):

- True Positives (TPs): The model predicts the target class and the prediction is correct.
- True Negatives (TNs): The model does not predict the target class and the prediction is correct, no target class should be predicted.
- False Positives (FPs): The model predicts the target class but the prediction is wrong and wrongly "detected" a class sample.
- False Negatives (FNs): The model does not predict the target class but the prediction is wrong and the model "missed" an instance.

Given those four categories, the following metrics can be used to evaluate the performance of the model:

Accuracy The Accuracy is the fraction of correct predictions given all predictions of the model:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.5}$$

Precision The Precision is the fraction of the correctly predicted positive samples given all as positive predicted samples, i.e. how much a positively predicted sample can be "trusted" a.k.a. how precise the model is:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{2.6}$$

Recall The Recall is the fraction of the correctly predicted positive sample given all positive samples, i.e. how many of the positive samples the model "detected" or "recalled":

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{2.7}$$

F-measures The F-measure is the harmonic mean of the Precision and the Recall. In general, the F_β measure controls the balance between the precision and the recall, where β is a parameter that controls that balance:

$$F_\beta = \frac{1}{\beta}\frac{2TP}{2TP + \beta FN + FP}. \tag{2.8}$$

A special case of F-measures is the F1-measure also known as the Sørensen-Dice coefficient, dice score, or Dice Similarity Coefficient (DSC):

$$F1 = \frac{2TP}{2TP + FP + FN}. \tag{2.9}$$

However, the F-measures do not take the TNs into account and thus have to be used with care.
While the previous metrics only consider a fixed binarization, some metrics try to give a more varied performance estimate, often considering multiple classification binarization threshold points/ modi operandi points and are often summarized as the Area Under a Curve (AUC):

Area Under the Receiver Operating Characteristic Curve The Area Under the Receiver Operating Characteristic Curve (AUROC) is the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve plots the True Positive

Rate (TPR) against the False Positive Rate (FPR) at each threshold, i.e. describing (from left to right) "how many additional false positives have to be allowed to detect all positives as positive". As the AUROC is an area under a (normalized) curve, the best value is 1.0, which means that the model is perfect and can detect all positives as positives. The "worst" value is 0.0, which means that the model is not able to detect any positives at all, however, inverting the prediction of the model would result in a perfect score. A value of 0.5 means that given balanced classes the model decides randomly and thus the model has no discriminative ability. While the AUROC score gives a fair estimate of performance in a balanced label setting, in settings with unbalanced or unknown class distribution the score is not very comparable between different settings (however it can still allow for model comparison in the exact same setting).

Average Precision The Average Precision (AP) score tries to give a more balanced estimate of the performance of a model. It is the average of the precision scores at different recall thresholds. It can be seen as an imperfect approximation of the area under the precision-recall curve. It summarizes the precision-recall curve and the weighted mean (with the recall increase at the threshold as weight) of the precision at each recall threshold:

$$AP = \sum_{i=1}^{N} Precision_i(Recall_i - Recall_{i-1}). \tag{2.10}$$

2.2.2 Aggregation schemes

For simple classification, there is often only one level of aggregation, the dataset level, i.e. the prediction of the model per sample and then aggregating the scores of all samples (e.g. calculating the mean score of all samples). However, for some tasks, such as pixel segmentation, there is an inherent hierarchy in the prediction e.g. there are inherently two levels, the image level and the dataset level. Here, different choices of metric aggregation are feasible. The simplest, which is analogous to a single-level only aggregation, is to ignore the affiliation of a pixel with an image/sample and just "throw all pixels in a bucket" and calculate the metrics on all pixels of the dataset. We term this dataset-level aggregation. This however can overshadow performance differences between different images, e.g. not showing a big deterioration in scores, even if the predictions for a fraction of the images completely fail (this can be particularly detrimental with class imbalances). A more natural choice of aggregation is to keep the hierarchical structure intact and first calculate the metrics on a sample level and then aggregate the results to

the dataset level. We refer to this as sample-level aggregation. However, in this case, the choice of aggregation method, e.g. mean, median, nth-percentile, etc. is very important and the individual samples can be more affected by the label disbalance (as not every sample contains all classes) and the choice of metric. To choose the most fitting aggregation scheme, one has to choose a scheme that best aligns with the target task, e.g. if failure should be way more penalized or there can be catastrophic failures with only one image, then a sample-level aggregation scheme would be more fitting.

2.2.3 Combining rankings

Often deciding on the right metric and dataset to compare methods results in not choosing one fixed setting but multiple settings/metrics which each give one ranking. While they can be used to detail certain aspects of the methods, there are also methods to combine the rankings and give a final ranking and stability analysis. One way to combine rankings is *Consensus ranking*. This is a ranking where the scores are combined as the mean of the individual ranks. This is a simple way to combine rankings but can be very sensitive to the individual scores and each setting has the same importance, independent of the dataset size.

2.3 Density estimation

Often we can assume that data is generated by some process that results in a certain non-arbitrary distribution. For example, a Gaussian distribution. We can use density estimation to estimate this distribution of the data [Bishop, 2006].

2.3.1 Gaussian Mixture Model

Given the central limit theorem, assuming normally distributed data is a good base/starting assumption. This also allows for an easy way to estimate the distribution of the data. Here, since the base assumption is that the data is distributed normally, only the right parameters of the normal distribution have to be determined. Given the probability density function (PDF) of a normal distribution as:

$$p(x|\mu, \Sigma) = \frac{1}{\sqrt{2\pi^d|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}, \tag{2.11}$$

the parameters can be determined/estimated analytically by simply computing the mean and covariance matrix of the data.
However, as is often the case, the data is not generated by only one process but by multiple processes (e.g. photos of cats and photos of dogs). In this case, the parameters of the distribution can be estimated by a mixture model. The Gaussian Mixture Model (GMM) is a probabilistic model that combines the parameters of multiple normal distributions (e.g. one for photos of cats and one for photos of dogs), where each normal distribution is weighted by a factor π:

$$p(x|\pi, \mu, \Sigma) = \sum_{i=1}^{N} \pi_i \frac{1}{\sqrt{2\pi^d|\Sigma_i|}} e^{-\frac{1}{2},(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}. \tag{2.12}$$

Here, shown for a GMM with N components. For this case, there is no direct analytical solution for the parameters of the mixture model. Instead, the parameters are often estimated using a Maximum Likelihood Estimation (MLE) algorithm. The MLE algorithm is an iterative algorithm that tries to find the best parameters by minimizing the log-likelihood of the data, also often referred to as Expectation-Maximization (EM). In the E-step of the MLE algorithm, the parameters are estimated by maximizing the likelihood of the data. In the M-step of the MLE algorithm, the parameters are estimated by maximizing the likelihood of the parameters. However, different algorithms such as gradient descent also work for this case [Richardson and Weiss, 2018].

2.3.2 Variational Inference

While GMMs parameterize a pre-determined distribution, variational inference is a more basic approach to estimating a data distribution.
There are multiple motivations and derivations for variational inference, which all end up with a similar result. The most direct one for the use case of density estimation is to find another expression or an approximation for $p(x)$. Here, for sake of (notation) simplicity, we aim to find an approximation for $\log p(x)$:

$$\begin{aligned}
\log p(x) &= \log \int p_\theta(x, z) dz \\
&= \log \int \frac{q_\lambda(z|x)}{q_\lambda(z|x)} p_\theta(x, z) dz \\
&\geqslant_{\text{JI}} \int q_\lambda(z|x) \log \frac{1}{q_\lambda(z|x)} p_\theta(x, z) dz \\
&= \mathbb{E}_{q_\lambda(z|x)} \left[\log \frac{p_\theta(x, z)}{q_\lambda(z|x)} \right] \\
&:- \texttt{ELBO}(x; \theta, \lambda).
\end{aligned} \tag{2.13}$$

Here the ELBO is a lower bound on the log-likelihood of the data distribution $p(x)$. While $p(x)$ is often untractable, the ELBO is tractable. Maximizing the lower bound often gives a good approximation of the true data distribution.
Another derivation comes from the motivation to find a good approximation of the hidden factors/variables z, which 'cause' an observation/data x. Here, for a data sample x, the 'true' hidden variables z are given by the process $p_\theta(z|x)$, which we can not directly compute. Hence, we use a tractable distribution $q_\lambda(z|x)$ to approximate the hidden variables (i.e. an inference model). To make $q_\lambda(z|x)$ approximate the true distribution $p_\theta(z|x)$ as close as possible, we can use the Kullback-Leibler (KL)-divergence between these two and minimize this divergence (with respect to λ):

$$\begin{aligned}
D_{\text{KL}}(q_\lambda(z|x) || p_\theta(z|x)) &= \int q_\lambda(z|x) \log \frac{q_\lambda(z|x)}{p_\theta(z|x)} dz \\
&= \int q_\lambda(z|x) \log \frac{q_\lambda(z|x) p_\theta(x)}{p_\theta(z, x)} dz \\
&= \int q_\lambda(z|x) \left(\log p_\theta(x) + \log \frac{q_\lambda(z|x)}{p_\theta(z, x)} \right) dz \\
&= \log p_\theta(x) + \int q_\lambda(z|x) \log \frac{q_\lambda(z|x)}{p_\theta(z, x)} dz \\
&= \log p_\theta(x) + \texttt{ELBO}(x; \theta, \lambda),
\end{aligned} \tag{2.14}$$

and since $\log p_\theta(x)$ is independent of λ it is equivalent to optimize the KL-divergence or to optimize the ELBO (also note that the last line can be reformulated to single out $\log p_\theta(x)$ and since the KL-diverence is always $\geqslant 0$ this again proofs that the ELBO lower bounds $\log p_\theta(x)$).
There are multiple ways to optimize the ELBO such as Black-Box Variational Inference or Gradient Estimation via Monte Carlo (MC) Sampling. However, in this work, we are more interested in parameterizing the distributions with neural networks.
Here, the ELBO is often rewritten as:

$$\begin{aligned}
\texttt{ELBO}(x;\theta,\lambda) &= \mathbb{E}_{q_\lambda(z|x)}\left[\log \frac{p_\theta(x,z)}{<} q_\lambda(z|x)\right] \\
&= \mathbb{E}_{q_\lambda(z|x)}\left[\log \frac{p_\theta(x|z)p_\theta(z)}{q_\lambda(z|x)}\right] \\
&= \mathbb{E}_{q_\lambda(z|x)}\left[\log p_\theta(x|z) - \log \frac{q_\lambda(z|x)}{p_\theta(z)}\right] \\
&= \mathbb{E}_{q_\lambda(z|x)}(\log p_\theta(x|z)) - D_{KL}(q_\lambda(z|x)||p_\theta(z)),
\end{aligned} \tag{2.15}$$

where D_{KL} is the KL-divergence.

2.3.3 Normalizing Flows

Normalizing Flows (NFs) are based on the 'Change of Variable Formula':

$$p_x(x) = p_z(f^{-1}(x))|\det(\frac{\partial f^{-1}(x)}{\partial x})| \tag{2.16}$$

for an invertible function f and a distribution p_z. This allows the mapping of samples from an arbitrary distribution p_x to samples from a predetermined distribution p_z. Note that in this case, both distributions need to have the same dimensionality. In practice, p_z is chosen as a Normal distribution and f as an Invertible Neural Network (INN), that maps from the data space to the normal distribution, where likelihood, etc. can easily be computed. There are multiple different choices for invertible layers in neural networks with computable determinants of the Jacobian. One of the oldest and most popular choices is a Planar Flow [Rezende and Mohamed, 2016]:

$$x = f(z) = z + uh(w^T z + b), \tag{2.17}$$

where u, w, b are learnable parameters and h is a restricted activation function e.g. *tanh*. Other popular choices are additive coupling layers, rescaling, and reshuffling layers.

2.3.4 Autoregressive density estimation

Using the chain rule, we can derive the density of a datapoint x as:

$$p(x) = \prod_{i=1}^{N} p(x_i|x_1, x_2, ..., x_i - 1) = \prod_{i=1}^{N} p(x_i|x_{<i}), \tag{2.18}$$

which is the basis for autoregressive models. Expressed colloquially, the data is assumed to be an ordered sequence of sub-data points x_i (e.g. pixels or time steps), and the next point in the sequence is predicted/modeled based on the previous points. For autoregressive density estimation, models are used, which are trained to predict the next point in the sequence, for example, pixelRNN [Van Den Oord et al., 2016] which uses all previous pixels as input and tries to predict the next pixel or pixelCNN [Van Den Oord et al., 2016] which uses a masking strategy so that during the convolution only "previous" pixels are used as input to predict the next pixel.

3 | Advancements for VAE-based anomaly localization

3.1 Anomaly localization beyond the reconstruction error using VAEs

3.1.1 Motivation - How to use a VAE

As previously described detecting abnormal "unexpected" samples has the potential to greatly aid the process of day-to-day manual or semi-automatic reviewing of medical images. Especially due to its label efficiency, generalizability and general applicability, anomaly localization is attributed with much promise in the field of automated medical image analysis and diagnosis. Early approaches using unsupervised anomaly detection methods in the medical field were mostly based on a reconstruction error, i.e. the difference between the reconstructed image of a limited model and the original image. This is based on the assumption that the limited model is only accurately able to reconstruct data that is similar to the data the model was built with but fails for data that is different, in particular anomalous data samples. Van Leemput et al. [2001] proposed to use a statistical model to reconstruct the input tissue-wise, identifying anomalies as discrepancies between the actual image and the model prediction. Liu et al. [2014] used low-rank decomposition, where the low-rank elements represent the normal parts of the image, and high-frequency elements represent anatomical and pathological varia-

tions. In the deep-learning era, Autoencoders (AEs) have also been used as these 'limited' reconstruction models, by limiting the information that can be passed through the bottleneck [Chen and Konukoglu, 2018; Chen et al., 2018; Pawlowski et al., 2018]. However, these 'limited' reconstruction model approaches are always dependent on the hypothesis that anomalies will not get reconstructed as well and thus the capacity of the model. This can be a 'walk on a fine line' between getting a good reconstruction and at the same time not being able to reconstruct anomalous data. Consequently, the capacity for a given problem also depends on the data complexity and size.
A more principled direction, often used in basic statistics, is to model the distribution of the data and test if data samples belong to this distribution. Here, recent density estimation techniques have shown promise with higher-dimension and more complex and structured data such as images. Variational Autoencoders (VAEs), flow-based models and autoregressive models are among the current de-facto standard methods for density estimation on images using deep-learning based explicit models and have shown early success in anomaly/out-of-distribution sample detection tasks [Abati et al., 2018a; An and Cho, 2015; Kiran et al., 2018; Nalisnick et al., 2018].
In this chapter, we will investigate to which extent VAEs can be used for anomaly localization. We will first explain the basics of VAEs and their capabilities to learn data distributions. Next, we will discuss different modes of anomaly localization and demonstrate why reconstruction-based anomaly detection with VAEs might be sub-optimal. We discuss the possibility of data leakage during model development in particular, as well as the reasons why reconstruction-based methods can still perform well on unsupervised tasks: to some extent, these shortcomings can be compensated for by modifying the model architecture to be ideally suited for specific tasks (see also [Chen et al., 2018; Goldstein and Uchida, 2016; Pawlowski et al., 2018]), as is common practice when optimizing hyperparameters on annotated validation sets. However, task-specific hyperparameter optimization contradicts the assumption-free anomaly detection approach (note: some definitions, formulations, and equations of this section were previously published by myself in [Zimmerer et al., 2019a]).

3.1.2 Previous works & concurrent works

Most previous works only considered AEs as reconstructing models for anomaly localization. Pawlowski et al. [2018] compare different AEs for CT-based pixel-wise anomaly localization. Chen et al. [Chen and Konukoglu, 2018; Chen et al., 2018] compare different AE-based approaches and propose an extension with an

adversarial latent loss.
Some early works used VAEs instead of AEs, however using them as reconstruction models only. Here, Baur et al. [2018] propose to use a AE with an adversarial loss on the reconstruction to get a more realistic reconstruction. However, the localization method in these papers is purley based on a reconstruction error and can only outline suspicious regions if they can not be adequately reconstructed by the models.
Concurrently to our work You et al. [2019] also propose to use a more holistic approach for VAE-based anomaly localization and include the KL-term for an iterative restoration approach that moves data samples closer to the learned data manifold and thus measures a proxy of distance to the data manifold.
In an attempt to unify and compare the approaches, Baur et al. [2020] used the same architecture for different models and compared them on different datasets.

3.1.3 Methodology

Introduction into VAEs

VAEs implement the concept of Variational Inference in NNs, in particular, try to optimize a parameterized ELBO given the training data. Revisiting the ELBO (Eq. (2.15)), it can be expressed as:

$$\texttt{ELBO}(x;\theta,\lambda) = \mathbb{E}_{q_\lambda(z|x)}(\log p_\theta(x|z)) - D_{KL}(q_\lambda(z|x)||p_\theta(z)). \tag{3.1}$$

For simplicity, we will minimize the negative ELBO instead of maximizing the ELBO. This results in two terms in the ELBO that can be minimized: $D_{KL}(q_\lambda(z|x)||p_\theta(z))$ and $-\mathbb{E}_{q_\lambda(z|x)}(\log p_\theta(x|z))$.

KL-divergence term To make the first part tangible, we will make some assumptions: $p_\theta(z)$ is assumed to be a isotropic normal distribution with mean 0 and standard variation 1: $p_\theta(z) = \mathbb{N}(z|0,1)$. $q_\lambda(z|x)$ is also chosen to be a diagonal normal distribution.
The inference distribution $q_\lambda(z|x)$ is parameterized and amortized by NNs f_μ, f_σ:

$$q(z|x) = \mathcal{N}(z; f_{\mu,\theta_1}(x), f_{\sigma,\theta_2}(x)). \tag{3.2}$$

This makes it possible to solve the KL-divergence term analytically in a closed form.

Expectation term For the second term $p_\theta(x|z)$ is also classically chosen as a diagonal normal distribution and is parameterized by a NN g_μ and constant c:

$$p(x|z) = \mathcal{N}(x; g_{\mu,\gamma}(z), \mathbb{I} * c). \tag{3.3}$$

The expectation is approximated with MC sampling, where empirically a MC sample size of 1 has shown sufficient and is often chosen. This also allows to compute the expectation analytically and find a closed form for the second term as well. Note that during the optimization this term boils down to an L2 loss indirectly weighted by the constant c (which is summed up over the input dimensions and not averaged as is common). As consequence this term is often also referred to as the *reconstruction term* or *reconstruction loss* of a VAE. Different choices of distributions for $p_\theta(x|z)$ correspond to the different commonly used loss functions, e.g. a Laplace distribution corresponds to L1 loss and a Binomial distribution to a cross entropy loss. Backpropagating through the sampling process is done using the reparameterization trick, i.e. sampling $\tilde{z}$ from $N(0;1)$ and then transforming the samples to $N(f_{\mu,\theta_1}(x), f_{\sigma,\theta_2}(x))$ with: $z = f_{\sigma,\theta_2}(x) * \tilde{z} + f_{\mu,\theta_1}(x)$, which is differentiable with respect to the encoder parameters.

VAE as AE When these assumptions and design decisions are made, the structure resembles AEs. The networks f_{μ,θ_1} and f_{σ,θ_2} are used to encode the input x into a latent space z and the network $g_{\mu,\gamma}$ is used to decode the latent space z into the output x. The only difference is (1) that the encoder f does not directly infer a latent vector but rather parameterizes a distribution from which the latent vector is sampled and (2) that the KL-divergence is used to 'regularize' this encoding process.

Training a VAE A VAE can then in analogy to traditional neural networks be trained using SGD (or more often, Adam [Kingma and Ba, 2017]). The loss for the VAE training is the approximate negative ELBO, composed of the KL-divergence loss L_{KL} and the reconstructing loss L_{rec}:

$$L_{VAE}(x) = L_{KL}(x) + L_{rec}(x). \tag{3.4}$$

An illustrated example of the VAE training with a 2D Gaussian can be seen in 3.1.

Anomaly detection with VAEs

After successful training of the VAE, the approximate ELBO can be a very faithful estimate of the true data likelihood. Especially, given adequately large and expressive neural networks f_μ, f_σ, and g_μ and a large enough latent space, VAEs with

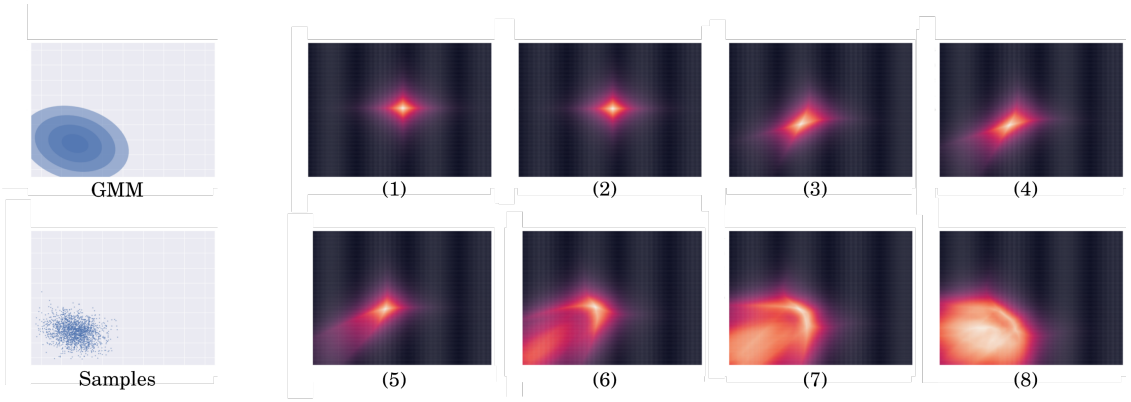


Figure 3.1: Example of VAE training on a 2D Gaussian. The images in the first column illustrate the GMM and samples that were drawn from the GMM. The images in the second-fifth column show the log-likelihood progression of the VAE during training.

Gaussian encoders and decoders can (and given the right conditions will [Dai and Wipf, 2019]) approximate the true data distribution. Thus the approximate ELBO or VAE-loss L_{VAE} is a natural choice to score data samples for anomaly detection.

Anomaly localization with VAEs

While the VAE-loss is an intuitive choice and often used for sample-level anomaly detection, previous work on anomaly localization with VAEs has almost exclusively adopted the AE reconstruction error strategy to localize the anomalies. While this can show good results in certain settings with adapted neural networks and encoding capacities, it is not very effective in general. In contrast to anomaly detection, it furthermore ignores one of two terms in the VAE-loss: the KL-loss, and thus potentially ignores important information. For example, a low likelihood and consequently high anomaly score can be caused by a high reconstruction error and/or a high KL-divergence. We thus suggest multiple "more holistic" ways to trace back the likelihood deviations to the pixel level and localize the anomalies, which incorporate both terms of the VAE-loss and thus the approximate ELBO.

A case for the score One way to interpret the results of some AE-based models with a denoising criterion, is as Alain and Bengio [2012] have shown, that the reconstruction error approximates the *score*. The *score* (which was also recently

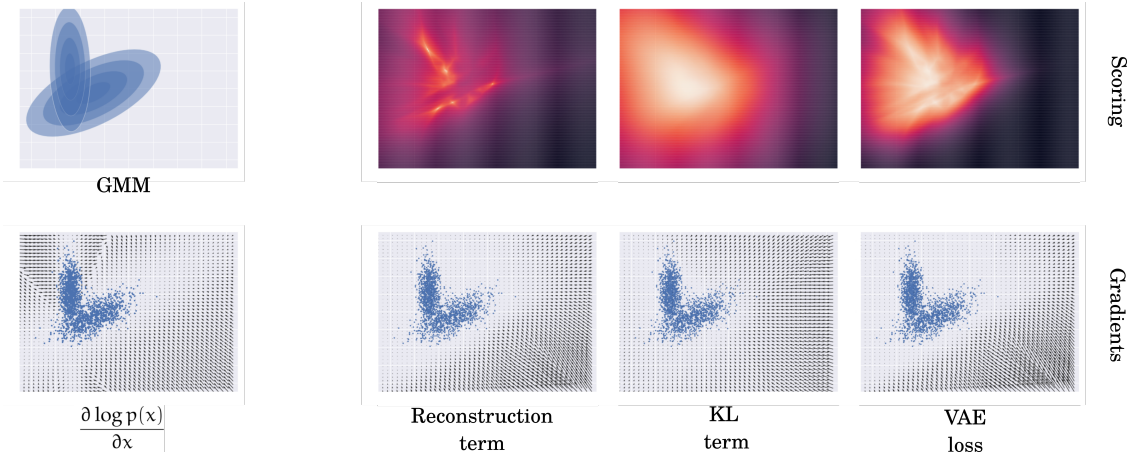


Figure 3.2: Example of VAE-based anomaly scores on a 2D GMM. Visualized are the GMM and the samples drawn from the GMM with the gradients of the log-likelihood in the first column. In the next column, the scorings and gradients of the different terms are Visualized: the reconstruction term, the KL term, and the VAE loss (the combination of the reconstruction term and KL term).

prominently used for diffusion models [Song and Ermon, 2019]) is defined as:

$$\frac{\partial \log p(x)}{\partial x}. \tag{3.5}$$

One bold hypothesis can consequently be that many AE- and reconstruction-based models work due to an approximation of this *score*.
Consequently and based on the following assumptions, we hypothesize that the *score* can give a good approximation for an abnormality rating:

- The *score* gives the direction towards the normal data sample (which, in the context of medical data, refers to a sample with diseased and abnormal anatomy turned into healthy tissue),
- The magnitude of the *score* indicates how abnormal the pixel is.

We note that the above-mentioned assumptions can be violated in practice, especially in cases far away from the healthy sample data distribution or for multimodal distributions.

Different localization methods We thus propose and compare the following methods for anomaly localization:

- "**Reconstruction Error**": The reconstruction error is a simple measure of the distance between the input and the reconstruction:

$$\textit{Rec-Error}(x) = \|x - g(f(x))\|. \tag{3.6}$$

- "**ELBO-grad**": Based on the premise that $\mathcal{L}$ provides a close enough approximation to the underlying data distribution and that the *score* can be used to point out anomalies, we propose to use the derivative of $\mathcal{L}$ with respect to the input. This results in a pixel-wise vector pointing towards a data sample with a lower $\mathcal{L}$:

$$\textit{ELBO-Grad}_{score}(x) = |[\frac{\partial \mathcal{L}}{\partial x}]| = |\frac{\partial(-D_{KL}(q(z|x)||p(z)) + \mathbb{E}_{q(z|x)}[\log p(x|z)])}{\partial x}|. \tag{3.7}$$

  This can colloquially be interpreted as the "direction towards normality" for each pixel. The magnitude of the pixel gradient should correspond to a pixel-wise anomaly score [Alain and Bengio, 2012] (given that $\mathcal{L}$ is locally convex).

- "**KL-grad**": To get a pixel-wise score for only the KL-term of $\mathcal{L}$, we differentiate the KL-term with respect to the input:

$$\textit{KL-Grad}_{score}(x) = |\frac{\partial(-D_{KL}(q(z|x)||p(z)))}{\partial x}|. \tag{3.8}$$

- "**Rec-grad**": To get a pixel-wise score for only the reconstruction-term of $\mathcal{L}$, we differentiate the reconstruction-term of $\mathcal{L}$ with respect to the input:

$$\textit{Rec-Grad}_{score}(x) = ||\frac{\partial \mathbb{E}_{q(z|x)}[\log p(x|z)])}{\partial x}|. \tag{3.9}$$

- "**Combi**": We can also directly use the reconstruction error and combine it with "*KL-grad*" instead of differentiating the reconstruction-term of $\mathcal{L}$. This should be less vulnerable to noise artifacts. For this approach, we combine the derivative of the KL-term with the reconstruction error by multiplication, since the terms differ by several orders of magnitude:

$$\mathcal{A}_{pixel} = |x - g(f(x))| \odot |\frac{\partial L_{KL}(x)}{\partial x}|. \tag{3.10}$$

A Visualization of a VAE trained on a 2D GMM with the respective scorings and gradients can be seen in Fig. 3.2. In the next section, we will present the performance of the approaches and the benefits compared to a reconstruction-based detection on an anomaly segmentation task.

3.1.4 Experiments & results

First, we will investigate the performance of the different terms of VAE-loss on the *FashionMNIST* dataset and see if one is superior to the other. Then, we will apply the method to brain imaging datasets and see if and how well it can be used to detect abnormal brain images as well as for anomaly localization.

Anomaly detection

We compare the discriminative performance of the ELBO $\mathcal{L}$, the KL-term, and the reconstruction term separately on an anomaly detection task to analyze how well they capture the data distribution and consequentially the concept of abnormality. We further analyze the robustness and generalizability across different parameter choices.

Setup for the FashionMNIST experiments We use the FashionMNIST dataset [Xiao et al., 2017] and train and validate the model on 54000 images using 9 of the 10 supplied classes, and then assess performance by attempting to distinguish between the classes observed during training and the 10th unseen/'odd' class. This dataset is often used as a benchmark for the performance of basic models and is a good starting point for the evaluation of more complex models. In particular, for the use-case of medical imaging, this dataset is interesting because it has one input channel with continuous values, which is comparable to a one-sequence MRI dataset.
We employed a model with a 3-layer fully connected encoder and decoder with 400 hidden units and ReLU non-linearities for the VAE model, similar to Paszke et al. [2017]. To assess robustness, we alter the number of latent variables, the standard deviation c of $p(x|z)$ (which results in a down or up-weighing of the reconstruction loss), image size/scaling, and the 'odd' class that is excluded during training. We use 20 latent variables by default, $c = 1$, a scaling factor of 1, and class 0 is omitted during training.

Setup for the BraTS experiments We also train and validate the models on the imaging data from the HCP dataset [Van Essen et al., 2012] ($N = 1000$ scans). For the evaluation we use the BraTS2017 dataset [Menze et al., 2015] ($N = 250$ scans). Disclaimer: This setup is not perfect. We opted for this setup since it is often used in literature [Baur et al., 2018; Chen et al., 2018], holds some clinical relevance and no "perfect" public dataset is currently available for this use case. Here, the HCP participants are all young healthy subjects. The scans in the BraTS dataset all

contain brain tumors. This already denotes a domain shift between the datasets, which may complicate the assessment of model applicability and performance. Furthermore, the type of anomalies in the BRATS dataset is limited to brain tumors and presents a rather 'easy-to-detect' anomaly case.
To apply the dataset to this anomaly detection use-case we treat slices without annotations as healthy and we define slices with at least 20 annotated tumor voxels as diseased/abnormal.
During training, we used minor data augmentations, such as multiplicative color augmentations, random mirroring, and rotations. The default model was inspired by DCGAN [Radford et al., 2015] and consists of a 5-Layer fully-convolutional encoder and decoder with the feature-map size of 16-32-64-256. The model down-samples using strided convolutions (stride 2) and upsamples with transposed convolutions, each followed by a LeakyReLU non-linearity.
The number of latent variables (default 256), the standard deviation c of $p(x|z)$ (default 1), and the image size (default 64×64 pixels) are modified as 'design-choice hyperparameters,' illustrating the influence on the performance, robustness, and generalizability of the different loss-terms.
The models are trained using Adam [Kingma and Ba, 2017] and a 0.0001 initial learning rate. The learning rate is reduced if the validation loss $\mathcal{L}_{val}$ approaches a plateau by lowering it by a factor 0.1. When the validation loss does not reduce after three epochs, the training is terminated. We run each model five times and present the mean as well as the maximum and minimum performance.

Detection results The sample-wise results across different parameter settings for the FashionMNIST dataset can be seen in Fig. 3.3 and for the BraTS dataset in Fig. 3.4.
Overall the results are very similar for both datasets. In most cases, the reconstruction term has lower discriminative power than either the KL-term or the ELBO. This emphasizes the point that using only the reconstruction term might discard useful information, potentially for anomaly localization as well. In particular, in cases where the reconstruction term has better performance, the model is severely constrained, for example by having a small latent variable dimension. Dai and Wipf [2019] showed that this hinders VAEs from approximating the data distribution and leads to poor reconstruction. Thus the robustness of the KL-term can perhaps also be partially explained by Dai and Wipf [2019], in which it is hypothesized that the ELBO best approximates the data distribution having "perfect reconstructions using the fewest number of clean, low-noise latent dimensions" for VAEs.
The results provided thus far have used the default model with no manual changes

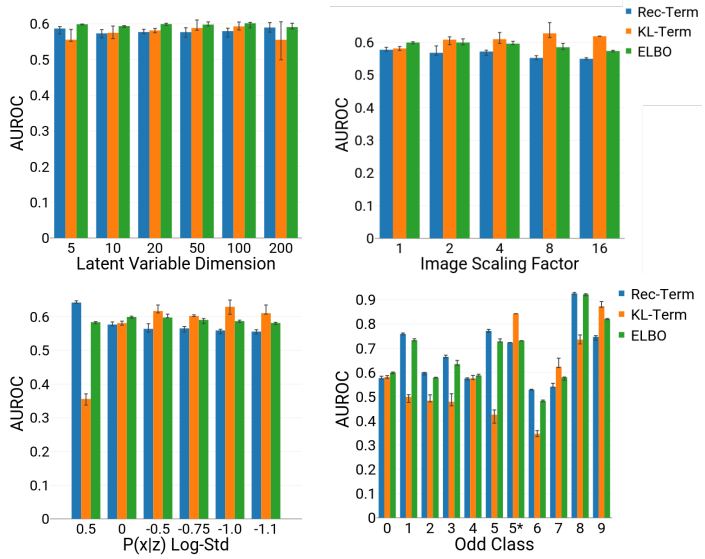


Figure 3.3: Anomaly detection performance across different VAE hyperparameter/design choices. The plots show the AUROC for reconstruction-term (Rec), the KL-term, and the ELBO $\mathcal{L}$ for the FashionMNIST dataset each graph depicting a variation along one 'hyperparameter dimension'. The odd class $5^\star$ shows a fine-tuned performance with odd-class 5 ($\log c = 1.4$) [Zimmerer et al., 2019a].

to the hyperparameters. In some circumstances, tuning them may allow one technique to clearly outperform the others. So as a next step we show that the hyperparameters can be tuned in such a fashion that the KL-term (or the reconstructing-term) alone can give a competitive performance.

To prove this we use the FashionMNIST dataset and chose the 'odd' class 5, which, in the default setting, has, from all 'odd' classes, the largest margin between a strong reconstructing-loss (AUROC of 0.73) and a weak KL-loss (AUROC of 0.42). Only adapting a single hyperparameter by hyperparameter search on a labeled validation set (explicitly: setting $\log c = 1.4$), the KL-term performance improves significantly, resulting in an AUROC of 0.82, which clearly outperforms the reconstruction-loss for this 'odd' class (and surpasses all other methods by a wide margin).

All in all, this suggests that for an unsupervised anomaly detection task (where usually by definition no annotated validation set is available), no one, neither the reconstruction term nor the KL-term, might, in general, be superior to another for anomaly scoring, and leakage of labeled information (e.g. using a labeled validation set to determine hyperparameters) can strongly bias the achieved results.

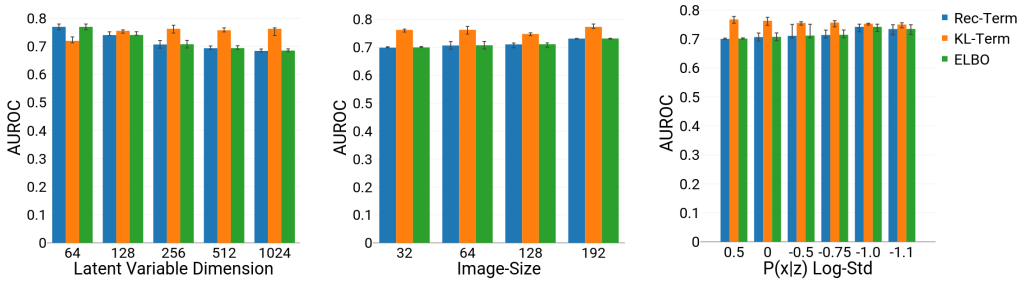


Figure 3.4: Anomaly detection performance across different VAE hyperparameter/design choices. The plots show the AUROC for reconstruction-term (Rec), the KL-term, and the ELBO $\mathcal{L}$ for the BraTS2017 dataset [Zimmerer et al., 2019a].

And this indicates that in general the KL-term may also present a beneficial addition to the often used reconstruction error for anomaly localization. Next, we show how the previously defined methods can be used to implement this, i.e. using the reconstruction-term as well as the KL-term for anomaly localization.

Anomaly localization

Experimental setup To explore the feasibility of applying the models to anomaly localization and investigate the potential and peculiarities of the VAE-loss we used a similar experimental setup to the previous experiments on the BraTS dataset (Section 3.1.4), for which also segmentations of the anomalies/tumors are available.
Here, again, we vary the 'design-choice hyperparameters' (the number of latent variables (default 256), the standard deviation c of $p(x|z)$ (default 1) and the image size (default 64×64 pixels)) to outline the robustness and generalizability of the different scoring methods.
To localize the anomalies we use the previously described scoring methods (Section 3.1.3). The backpropagation of the different loss terms onto the image (and approximation of the *score*) is implemented with the Smoothgrad algorithm [Smilkov et al., 2017]. Due to checkerboard artifacts caused by the convolution, we apply a Gaussian smoothing with kernel size 5 to the gradients.

Localization performance The pixel-wise anomaly localization performance on the BraTS2017 dataset for the space of the different 'design-choice hyperparameter' settings is presented in Fig. 3.5. In this pixel-wise / pixel-level anomaly detection scenario, we use pixel-wise reconstruction-error (*Rec-Error*), the backpropagated

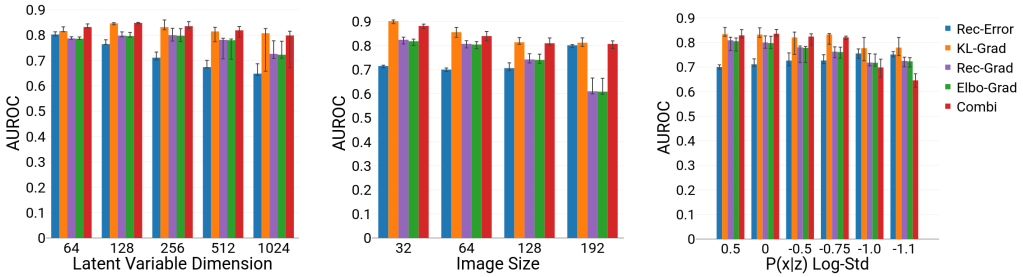


Figure 3.5: Anomaly localization performance across different VAE hyperparameter/design choices. The plots show the AUROC on the BraTS2017 dataset for the reconstruction loss, the KL-term gradient, the reconstruction-term gradient, the ELBO $\mathcal{L}$ gradient and the proposed *combi* method [Zimmerer et al., 2019a].

$\mathcal{L}$ (*Elbo-Grad*), its backpropagated KL-term (*KL-Grad*) and reconstruction-term (*Rec-Grad*) separately as well as the *combi* model present above (Section 3.1.3).
Here, similar to the sample-wise FashionMNIST example, there is no clear winner in most cases, and the reconstruction error alone is often outperformed by other methods. Rather, in the proposed default VAE hyperparameter settings as well as for most design choices, the *KL-Grad* and the *combi* model perform best. This indicates that for this setting, the reconstruction term alone might not be optimal and that other choices can offer a more robust performance (for this particular dataset).
Next, we will further show the importance and potential of the KL-term and contrast the KL-term even more with the reconstruction term. In particular, we will see what performance the KL-term can achieve when optimizing the hyperparameter setting and see if the results are comparable or even competitive with results reported in the literature using the reconstruction term alone.

Optimal setting performance Given the previous setting, the top methods already exhibit a high AUROC of > 0.9 for some settings. Particularly interesting is here the KL-term, which was previously often ignored for anomaly localization, and shows such a top performance as well as robust scores across most settings. Consequently, as it was previously never reported, we are interested in the top performance the KL-term approach can achieve for these datasets. Here, as is often done in the literature when presenting a reconstruction-based approach, we use an annotated validation set to tune the hyperparameters (Note: while we believe that this is in general not good practice, in the next chapter we will discuss how this annotated set can be created automatically and that it does not have to leak information about the anomalies in the test set).

In particular, here we chose a model from the previously reported hyperparameter space that showed top performance regarding the KL-term. First, we compare different VAE intrinsic detection methods, and basic AE-based methods and then compare them with results reported in the literature.

First, we inspect the *score*, dividing it into the reconstruction-loss gradient and KL-loss gradient, to get insights into the benefits of including the KL-term in the anomaly detection. We extend the analysis with more metrics that can be derived from a VAE such as the reconstruction error of the VAE, the smoothed reconstruction error, and the sampling deviations by determining the standard deviation of multiple MC samples. Lastly, a reconstruction-error-based Denoising Autoencoder (DAE) [Vincent et al., 2010] with the same architecture using the reconstruction error is evaluated as well.

The results are presented in Fig. 3.6 and Table 3.2 (and for the ELBO gradient in more detail in Fig. 3.8), samples and the corresponding pixel-wise ratings for samples can be seen in Fig. 3.7.

As similarly previously observed in [Chen et al., 2018; Pawlowski et al., 2018], the reconstruction-error based VAE and reconstruction-error based DAE detection performed roughly on par which each other. Further postprocessing using smoothing improves the results by removing high-frequency detections which likely are artifacts of the reconstruction error. The reconstruction-loss gradient interestingly performed better than the reconstruction error but showed poorer performance than the KL-loss gradient.

The best performance with an AUROC of 0.94 was achieved by the approximated *score* using the ELBO gradient (KL-loss + reconstruction-loss). However, the KL-loss alone performed similarly well and adding the reconstruction-loss gradient only showed marginal benefits.

Fig. 3.7 shows a clear difference between the reconstruction-loss gradients and the KL-loss gradients. An interpretation of Fig. 3.7 could be that the reconstruction-loss gradient focuses more on parts of bad reconstruction and thus given the performance of the reconstruction error, not always corresponds to an anomaly. The KL-loss on the other hand may focuses more on the distance of the feature (distribution) to the prior and thus "feature deviation from normality".

Finally, we calculated the DSC to compare the performance of the models to previously reported results. Here, the DSCs are calculated by thresholding the anomaly score values at a threshold that was determined using a greedy search on $\frac{1}{5}$ of the test dataset. The reported DSCs were then taken from the other $\frac{4}{5}$th of the dataset. Results are shown in Table 3.1.

Table 3.1: Whole tumor DSC on the BraTS Dataset for several 'unsupervised' approaches (the number in brackets identifies the year of the BraTS dataset used). Hand-crafted non-deep learning algorithms that were explicitly created with domain knowledge of the dataset in mind outperform our solution, which is comparable to existing deep-learning-based anomaly detection methods.

deep-learning		ours		non deep-learning		
α-GAN (15)	VAE-Rec (15)	default (17)	fine-tuned (17)	GHMRF (13)	X-Saliency (HGG 14)	GMM (15)
0.37	0.42	0.36	0.44	0.72	0.75	0.22

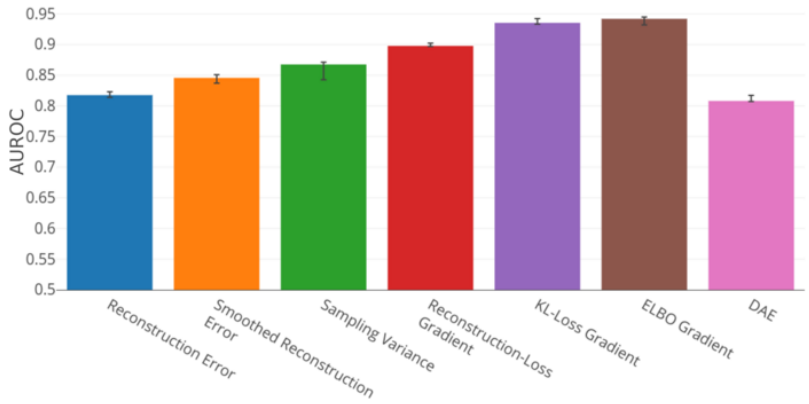


Figure 3.6: Comparison of the pixel-wise tumor localization performance on the BraTS-2017 dataset [Zimmerer et al., 2018].

Table 3.2: Pixel-level AUROC values of the different scoring methods (see 3.6).

	AUROC
DAE	0.808 ± 0.009
Reconstruction Error	0.817 ± 0.003
Smoothed Reconstruction Error	0.843 ± 0.008
Sampling Variance	0.855 ± 0.013
Reconstruction-Loss Gradient	0.894 ± 0.020
KL-Loss Gradient	0.939 ± 0.007
ELBO Gradient	0.939 ± 0.008

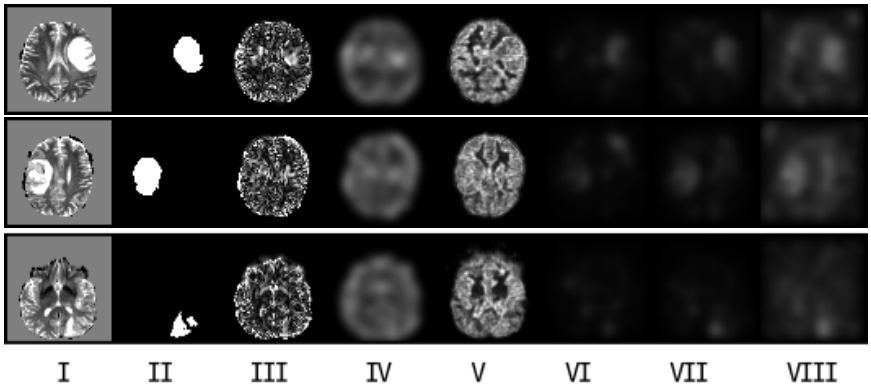


Figure 3.7: Samples from the dataset with the different pixel-wise rating schemes, showing the original sample (I), the annotation (II), the reconstruction error (III), the smoothed reconstruction error (IV), the sampling variances (V), the reconstruction-loss gradient (VI), the KL-loss gradient (VII), and the combined gradient which approximates the *score* (VIII) (for more samples, see Fig. 6.1 and Fig. 6.2) [Zimmerer et al., 2018].

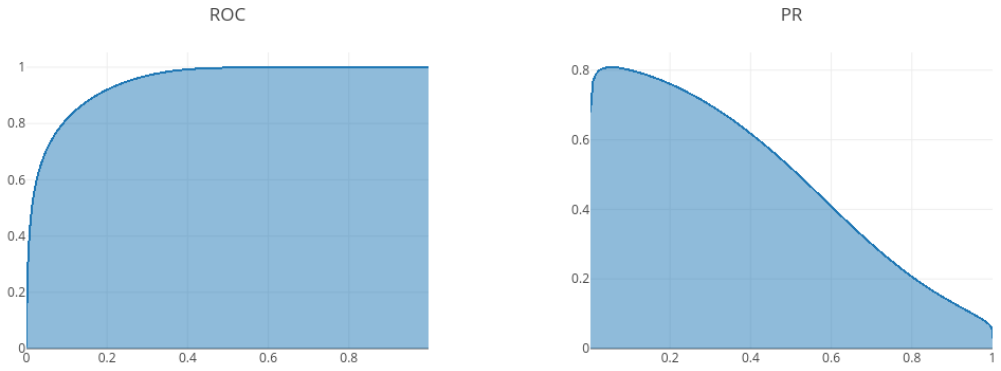


Figure 3.8: Pixel-level ROC and PR Curve on the BraTS 17 test set using the ELBO gradient.

3.1.5 Discussion & conclusion

We presented a way to localize anomalies using a VAE beyond the reconstruction error. As early experiments on the simple FashionMNIST dataset showed the reconstruction error alone is not the optimal choice for sample-level anomaly detection. We further extended that to anomaly localization on medical images

and showed that using the gradients as a localization method, this statement also holds true for anomaly localization on medical images. To the best of our knowledge, this is the first time VAE gradients were used for anomaly detection or localization. Using gradient-based approaches, which furthermore are more theoretically grounded, the VAE-based anomaly detection score on the BraTS tumor segmentation dataset could be improved compared to previously reported results outperforming the previously best reported AUROC of 0.92 [Chen and Konukoglu, 2018; Chen et al., 2018].

Overall, we showed, that for 'basic' VAE-based anomaly localization, the free lunch theorem holds, there is no superior version that is best in all cases and settings. However, this also means that for many cases in the anomaly localization setting the reconstruction error does not always automatically give the best performance and can regularly be improved by combining it with the backpropagated KL-term. Using fewer latent variables or putting more importance on the KL-loss could, while potentially causing inferior overall performance, lead to a more competitive performance of the reconstruction error. However, this may hinder the model from effectively learning the true data distribution [Dai and Wipf, 2019] and unnecessarily limits the model and its capacity. The use of both the KL-term and reconstruction-term is in line with the common practice of using the approximate ELBO, which constitutes the definition of the combination of the KL-term with the reconstruction-term, as a basic way to detect outliers using VAEs instead of only using the reconstruction error. Moreover, the ELBO is theoretically inspired and may be deployed as the base of other appropriate on-top techniques like 'test-of-typically' [Nalisnick et al., 2019]. The reconstruction term alone has a less theoretical foundation. Overall, combining multiple aspects may be more robust across different settings (as hinted by the experiments) and thus effectively reduce the need for manually turning on a validation set. In particular, the relative influence of the reconstruction loss can depend on the regularization of the latent variables. Furthermore, using the ELBO-grad method outperformed the previously reported results, which only use a reconstruction error. Consequently, we want to stress the point that including the KL-loss and the *score* of a model can lead to an improvement in VAE-based methods for anomaly localization and should not be ignored by default.

Particularly, the 'score' method of detecting anomalies can be generalized and is directly applicable to other state-of-the-art density estimation techniques, such as Grow [Kingma and Dhariwal, 2018] or Pixel-CNN++ [Salimans et al., 2016]. Furthermore, later research showed that the gradients can be competitive for a sample-level anomaly detection task [Huang et al., 2021; Igoe et al., 2022].

On the BraTS dataset, it appears that non-deep learning approaches perform better

than our method. This, however, overlooks the fact that these models, through their algorithmic design, integrate specialized domain knowledge. While this results in a good performance, it makes them inappropriate for use in other organs or modalities. Our suggested technique does not impose such assumptions, is adaptable in terms of the precise selection of hyperparameters, and can therefore be successfully applied to different situations or datasets without change.
We feel that our suggested technique is an important step toward advancing anomaly identification in medical imaging applications and allows VAE-based methods to tap into the potential of the KL-term. Furthermore, the score, while a good anomaly localization method as-is, might offer further potential for uncertainty indicators and semi-automatic evaluation.

3.2 Improving anomaly localization with self-supervised learning

3.2.1 Motivation

The previous results have shown that feature deviations from the prior in a VAE can be used to identify anomalies. This stresses the importance of learned features and this use case, learning "meaningful" features, be it for pretraining or clustering was one of the early selling points for VAEs. However, recent research has shown that the features learned by a VAE often focus on low-level information and statistics. Self-supervised learning methods, which have lately become popular, attempt to tackle this problem by utilizing a human-crafted task that ideally encourages the model to learn meaningful features. Thus with the advancement of the self-supervised learning approaches, the 'classical' unsupervised feature learning approaches were outperformed and replaced in many areas by self-supervised approaches. This raises the question of whether the success of self-supervised learning methods translates to anomaly localization settings as well and if they can be integrated into the existing VAE-based approaches (note: some definitions, formulations, and equations of this section were previously published by myself in [Zimmerer et al., 2019b]).

3.2.2 Related work & SotA

Self-supervised learning Self-supervised learning usually refers to the use of a human-crafted task to optimize the model parameters for later usage, e.g. pretraining or linear classification using the pre-trained features. One of the first models that could classify as self-supervised learning models were the DAEs [Vincent et al., 2010]. Here, an AE was trained to reconstruct a previously noise-perturbed image. Other early works that have shown good performances for pretraining used recolorization [Zhang et al., 2016], predicting the rotation of an image [Gidaris et al., 2022], reordering of image patches similar to a jigsaw puzzle [Noroozi et al., 2017], predicting the relative position of two patches to each other [Doersch et al., 2016], inpainting or context encoding masked out parts of an image [Pathak et al., 2016], and exemplar-based learning (i.e. seeing each individual image as one class and classifying augmentations of this image as belonging to this class) [Dosovitskiy et al., 2015].
While the first mentioned methods have shown incremental performance improvements, especially the latter two, masking and exemplar learning, have proven to

be more effective in the current state of the art. While exemplar learning on its own has not shown fundamental improvements, an incremental adaptation of the idea, contrastive learning, was shown to be a powerful method for improving the performance of self-supervised learning [Oord et al., 2019]. Masking on the other hand has proven similar performance in particular for sequence models such as transformers and has there become one of the most successful pretraining methods [He et al., 2021].

Contrastive learning Similar to the concept of examplar learning, contrastive learning utilizes a sample to generate positive pairs (semantically similar data points), which attract each other, and negative pairs (semantically different data points), which repel each other in the latent space. As proposed by Oord et al. [2019], a Normalized Temperature scaled cross-entropy (NT-Xent) loss is used and is one essential part of the success of contrastive learning [Oord et al., 2019]. While the NT-Xent loss is in common for all contrastive learning approaches, the way to generate the positive pairs is different for each method. The early works on contrastive learning [Oord et al., 2019] interpreted samples as sequences (e.g. an image as a sequence of pixels) and used the current sequence sample and the consecutive sequence sample as positive pairs, while non-sequential sequence samples were seen as negative. This has shown good performance on multiple different downstream tasks in multiple domains. However consecutive work has shown to be more effective for images and circumvented the need for a sequence representation. Particularly in computer vision, data augmentation has been proven to be a simple and effective way to produce positive pairs. For example, Simple framework for Contrastive Learning of visual Representations (SimCLR) [Chen et al., 2020] uses random data augmentation on a sample to generate two augmented samples, which form the positive pair. Other augmented image pairs in the batch are seen as negatives for that sample. This way of pretraining has shown to be rather robust and easy to implement and has on one hand shown new SoTA performance on computer vision tasks while on the other hand allowing for a significantly reduced number of labeled samples to achieve similar performance previously reported fully-supervised methods. Furthermore, the features produced by SimCLR alone demonstrated greater linear separability for subsequent classification tasks than other self-supervised techniques and AEs, implying that the SimCLR features match better to human-interpretable semantic information.

Masking Inpainting research can be considered one of the origins of using masking as a pretraining method. The original goal of image inpainting was to re-edit

images, remove objects, image restoration or manipulation, ... and has recently achieved impressive results [Lugmayr et al., 2022; Zheng et al., 2022] and made it into everyday consumer products [Photos, 2021]. The Context-Encoding (CE) framework [Pathak et al., 2016] was the first to use inpainting or masking as a pretraining method. Here, a network was tasked to inpaint and restore a previously removed part of the image. The trained network showed great performance on further downstream tasks such as image segmentation, and image recognition. The authors suspect this to work better than other reconstruction methods, because to inpaint the missing information, the network has to learn the semantics of the image, which later on can help for downstream tasks. While better performing pretraining methods for CNNs have been proposed, lately masking has regained importance as a pretraining method for transformer networks [He et al., 2021].

Self-supervised learning for anomaly detection Some of the self-supervised ideas have been incorporated into different anomaly detection approaches. For example, Golan and El-Yaniv [2018] use a rotation prediction and learn the distribution of the final logits to detect anomalies. Later, some works have incorporated contrastive learning for anomaly or out-of-distribution detection. Here, Winkens et al. [2020] extend a classification model with a contrastive learning task, which has shown to improve out-of-distribution detection. The authors argue that this can be attributed to the more general and less target-task-dependent features learned by the contrastive learning task.
To the best of our knowledge, there was no prior work for a self-supervised method for anomaly localization. Later on, Venkatakrishnan et al. [2020] proposed a multi-task prediction framework for self-supervised anomaly localization and compared themselves with the work presented here, however, showing no significant improvements.
So next we want to introduce two ways of using self-supervised learning for anomaly localization, first using masking/CE incorporated into the VAE framework and second using contrastive learning to localize anomalies.

3.2.3 Methodology

Context-Encoding

Next, we present an integration of CE into VAEs as an anomaly detection method: the Context-encoding Variational Autoencoder (ceVAE). By extending VAEs with CE, we aim to improve the internal latent representation and as such make deviations from the prior more suited for anomaly detection on a sample as well as

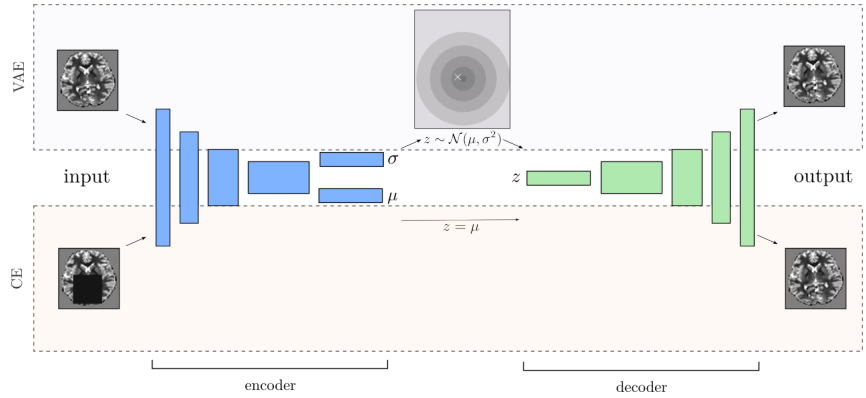


Figure 3.9: The model structure of the ceVAE. On the top is the VAE branch depicted and on the bottom is the CE branch [Zimmerer et al., 2019b].

pixel level. Similar to a VAE, we implement the ceVAE with fully convolutional encoders f_μ, f_σ and a decoder g. There CE-part only uses the mean encoder f_μ to encode a data sample (similar to VAEs f_μ and f_σ share most of their weights [Kingma and Welling, 2013], see Fig 3.9).

VAE branch The VAE branch in the ceVAE is the same as a VAE. It is used to get a proxy for the likelihood of a data sample. As is common, encoders f_μ, f_σ, a decoder g, and a standard diagonal Gaussian prior $p(z)$ are used, resulting in the VAE objective:

$$L_{VAE} = L_{KL}(f_\mu(x), f_\sigma(x)^2) + L_{rec_{VAE}}(x, g(z)), \tag{3.11}$$

where $z \sim \mathcal{N}(f_\mu(x), f_\sigma(x)^2)$ using the reparametrization trick and L_{KL} is the KL divergence loss with a standard Gaussian as in Eq. (3.4).

CE branch For the CE-part, first, a sample x is noised/masked with CE noise. In particular, one or multiple rectangular regions with random size at a random position of the original sample x are randomly masked out to create a perturbed sample $\tilde{x}$. The objective of the CE-branch is then to reconstruct the original sample x given $\tilde{x}$ which at the same time potentially improves the "latent-space features" of the VAE. Therefore is it first fed through the encoder f_μ and then the decoder g:

$$L_{rec_{CE}}(x, g(f_\mu(\tilde{x}))). \tag{3.12}$$

A further benefit the CE might give is that the reconstruction error can become more expressive. As noted by [Alain and Bengio, 2012] for a DAE the reconstruction error (under some approximations and assumptions) can approximate the

derivative of the log-density with respect to the input $\frac{\partial \log p(x)}{\partial x}$. While the CE noise is no additive noise (and does not decrease during training), this might still hold true, and for simple 2D experiments, we could see a strong correlation between a DAE with Gaussian additive noise and a DAE with CE noise.

ceVAE By combining the CE and the VAE, we aim at increasing the expressiveness of the model-internal latent space and thus allow for a better delineation of anomalies, in particular when using the deviations from the prior. The combined objective function is consequently given as:

$$L_{ceVAE} = L_{KL}(f_\mu(x), f_\sigma(x)^2) + L_{rec_{VAE}}(x, g(z)) + L_{rec_{CE}}(x, g(f_\mu(\tilde{x}))), \tag{3.13}$$

where L_{KL} is the KL-loss, z is sampled using the reparametrization trick and $\tilde{x}$ is perturbed by masking out regions as in CEs. The CE task (with $L_{rec_{CE}}$) does not impose normality restrictions on the prior $p(z|x)$ during training. This is necessary to stop the model from classifying these altered cases as "normal". Additionally, the CE and VAE working simultaneously can have a regularizing influence and inhibit the posterior collapse of the VAE.

Anomaly detection and localization While for this work we prioritize the localization of anomalies, the ceVAE model also allows for the detection of anomalies on a sample level. After training the ceVAE, optimizing the loss and ideally increasing the approximated ELBO, the trained model allows for an estimate of the data sample likelihood. This likelihood can then be used to determine the anomaly score of a data sample. Thus the sample-wise anomaly score is given as:

$$A_{sample} = L_{KL}(x) + L_{rec_{VAE}}(x, g(z)), \tag{3.14}$$

Simultaneously, to localize abnormal parts in the data sample we can use the methods proposed in Section 3.1.3. Overall, as stated before, we aim for a combination of the reconstruction error and the KL-loss to capture both aspects of the VAE. In particular, by adding the CE part, we aim at improving the reconstruction term and even more the importance of the feature distribution deviations from the prior due to the increased expressiveness of the latent space.
Here, we propose a generalization of the combi method using an element-wise function h to combine the scores. The pixel-wise anomaly score is defined as:

$$A_{pixel} = h\left(|x - g(f(x))|, \left|\frac{\partial(L_{KL}(x) + L_{rec_{VAE}}(x, z))}{\partial x}\right|\right), \tag{3.15}$$

where the reconstruction error is the absolute pixel-wise difference, and the pixel-wise derivative is calculated by backpropagating the ELBO back onto the data sample.
Choosing h as pixel-wise multiplication returns the known combi scoring:

$$\mathcal{A}_{pixel} = |x - g(f(x))| \odot |\frac{\partial L_{KL}(x)}{\partial x}|. \tag{3.16}$$

Contrastive learning

While CE, due to its similarity to AEs, allowed for a simple extension of the VAE, for contrastive learning we propose a different two-stage approach. The two stages can be seen in Fig. 3.10. In the first stage an encoder f is trained using contrastive learning to produce a 'useful' feature space/ mapping from images to features: $z = f(x)$. During the second stage, a generative model p is fitted on the representations to allow for a likelihood estimate of a representation. The Negative-Log-Likelihood (NLL) of its representations is given as: $s(x) = -\log(p(f(x)))$.
The pixel-level anomaly scores are obtained similarly to the method proposed in 3.1.3 by back-propagating the gradients of the representation NLL into the sample. Overall, we term this approach Contrastive Representations for unsupervised Anomaly Detection and Localization (CRADL).

Contrastive training The contrastive pretext task is similar to SimCLR [Chen et al., 2020] in that positive pairings are created by randomly selecting and applying data augmentations t from a collection of augmentations $\mathcal{T}$. In a minibatch of N samples, each sample x_i undergoes two transformations, resulting in the two distinct enhanced samples each that make up the positive pair.
By minimizing the NT-Xent contrastive loss, the representations formed by passing augmented samples through the encoder and projection head, $\tilde{u}_i = g(f(\tilde{t}(x_i)))$ and $\hat{u}_i = g(f(\hat{t}(x_i)))$, are steered to be identical:

$$l(\tilde{x}_i, \hat{x}_i) = -\log \frac{\exp(\text{sim}(\tilde{u}_i, \hat{u}_i)/\tau)}{\sum_{\bar{u} \in \Lambda^-} \exp(\text{sim}(\tilde{u}_i, \bar{u}))/\tau)} \tag{3.17}$$

Here the set Λ^- consists of all examples except $\tilde{u}_i$, all other $2N-1$ examples in the minibatch. The loss over the whole minibatch is obtained by summing all positive pairs (with both permutations).

Generative model In general, any suitable generative model can be fitted on top of the representations learned by the contrastive model. Here, we propose to

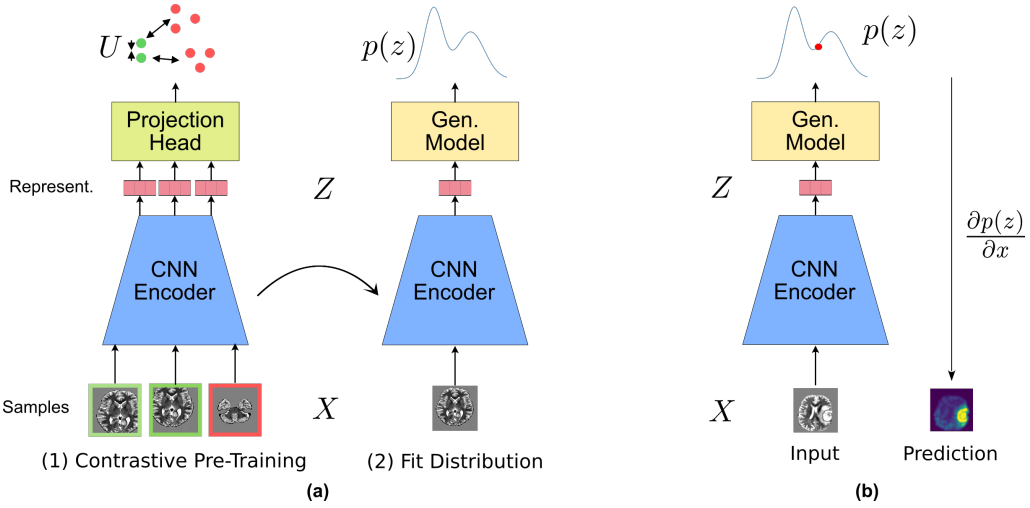


Figure 3.10: (a) CRADL fitting pipeline: (1) learning the contrastive pretext task with SimCLR. (2) fitting of the generative model on the learned features. (b) Visualization of the testing/prediction phase of the model. The anomaly score/prediction is calculated as the gradient of the predicted likelihood with respect to the input image.

use a GMM as the generative model, since it is one of the most basic generative models used for anomaly detection and the closely related Mahalanobis distance has shown good results in previous studies [Kamoi and Kobayashi, 2020; Winkens et al., 2020]. As noted in 2.3.1 the probability distribution of a GMM with K components is given as:

$$p(x;\Theta) = \sum_{k=1}^{K} \mathcal{N}(x;\mu_k,\Sigma_k)\cdot\pi_k \tag{3.18}$$

We used the EM algorithm [Dempster et al., 1977] to determine the model parameters with the number of components K being specified before the fit.

3.2.4 Experiments & results

To evaluate the benefits of masking and contrastive learning for anomaly localization, we first conducted a simple experiment using the ceVAE and outline the benefits masking can bring to the problem, without much modification of the existing approaches. Next, we then investigate if contrastive learning can be

used to improve upon masking as pretraining for anomaly localization and then compare both in an idealized setting.

Data

Training data For all the experiments, the T2-weighted brain MRI images were used. Here, the normal dataset was comprised of brain scans from the HCP dataset [Van Essen et al., 2012], i.e. giving the definition and distribution of normal data samples. For the ceVAE model, the HCP dataset was split into 1092 patients for training and 20 for validation. For CRADL, the models were trained on a subset of the HCP dataset [Van Essen et al., 2012] using 894 scans split into training and validation set and used the left-over scans to determine some model-specific parameters: Using the left-over scans, we created a synthetic anomaly dataset (similar to [Zimmerer et al., 2020]) from 100 HCP separate scans (HCP synthetic anomaly dataset (HCP Synth.)) by rendering real-world objects into different brain scans. So the trainig set and the HCP Synth. datasets stem from the same distribution (i.e., same scanner, site, ...) and only the introduced foreign objects cause the difference in the distribution. TheHCP Synth. dataset was divided into two halves, each of which has 49 scans: one for model development (i.e., choosing our hyperparameter values and K) and one for testing alone.

Test data To test the models on different medical datasets, containing different anomaly types, the BraTS-2017 [Bakas et al., 2017] dataset, split into 20 scans for validation and 266 for testing, and the ISLES-2015 [Maier et al., 2017] dataset, split into 8 scans for validation and 20 for testing, was used.

Preprocessing Each dataset was preprocessed similarly, with a patient-wise z-score normalization and clipping the range of intensities to $[-1.5, 1.5]$. Due to memory constraints, the sample slices were resized to a resolution of 64×64 for the ceVAE experiments and to 128×128 for CRADL. In order to prevent overfitting, we employed random mirroring, rotations, and multiplicative brightness augmentations during training. During training, we selected the top-performing model with respect to the validation loss for testing.

Model implementations

ceVAE For the ceVAE, the encoder and decoder networks, inspired by the deep convolutional architecture in [Radford et al., 2015], are implemented with CNNs with five 2D-Conv-Layers and 2D-Transposed-Conv-Layers respectively with

CoordConv [Liu et al., 2018], kernel size 4 and stride 2, each layer followed by a LeakyReLU non-linearity. The encoder and decoder are symmetric with 16, 64, 256, and 1024 feature maps and a latent variable size of 1024. As is common for VAEs [Kingma and Welling, 2013] the encoder for μ and σ have shared weights, only splitting at the last layer into two heads (one for μ and one for $\log \sigma$).
The reconstruction loss L_{rec} was chosen as the L1-Loss (assuming a Laplace distribution for the expectation-term/reconstruction term in the ELBO equation). This was shown to be a similarly good choice for VAEs and produce visually sharper images.
In analogy to Pathak et al. [2016] the CE noise was chosen as 1-3 randomly sized and positioned squares, however with a random 'color' value, which slightly deviates from the original image CE. This makes the challenge of correcting the noise slightly harder, is conceptually more akin to DAEs with Gaussian noise and bears similarities to the later developed Foreign Patch Interpolation (FPI).
The model was optimized with Adam [Kingma and Ba, 2017] with a learning rate of 2×10^{-4} and trained with a batch size of 64 for 60 epochs.

CRADL Similar to the ceVAE the CRADL encoder solely consists of 2D-Conv-Layers, starting with an feature map size of 64. The latent dimension size was set to 512. The projection head was implemented with a 2-layer MLP with ReLU non-linearities, a 512 dimensional hidden layer, and output of size 256.
Using the Adam [Kingma and Ba, 2017], a learning rate of 1e-4, cosine annealing, 10 warm-up epochs, a weight decay of 1e-6, and a temperature for the contrastive loss of 0.5, the encoder undergoes 100 epochs of contrastive pretext training on the HCP training set. The encoder with the lowest loss on the HCP validation set is chosen for further experiments. We combined random cropping, random scaling, random mirroring, rotations, multiplicative brightness, and Gaussian noise to create several (positive and negative) samples for the contrastive training.
The GMM was fitted on representations of the encoder using all samples in the HCP training set without any augmentation. The means of the components were randomly initialized, and the convergence limit for the EM algorithm was set to 0.1.

Benchmark methods

As methods for comparing the suggested models with, an One Class Support Vector Machine (OC-SVM) and different AE-based methods, which have shown state-of-the-art performance on similar tasks [Baur et al., 2018; Chen and Konukoglu, 2018; Chen et al., 2018; Pawlowski et al., 2018] were chosen.

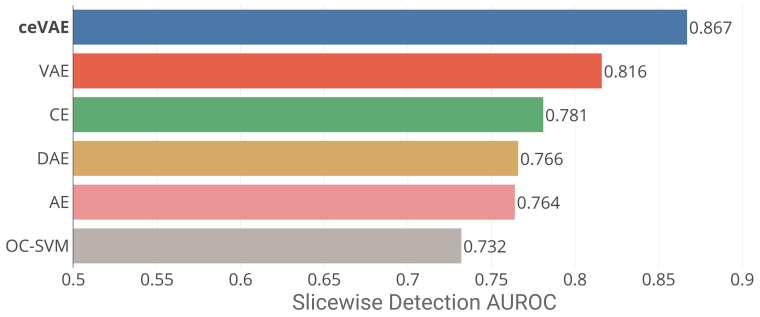


Figure 3.11: Slice-wise anomaly detection performance (AUROC) of different models on the BraTS-2017 dataset [Zimmerer et al., 2019b].

The OC-SVM was based on the libsvm implementation [Chang and Lin, 2011]. The AE-baseline methods, a standard AE, a DAE, a CE, and a VAE, were implemented using the same model structure and training scheme as the ceVAE.
To further inspect the benefits of combining the CE and VAE,a ceVAE weighting factor, termed ceVAE-Factor, is introduced which indicates the ratio of the CE Loss ($L_{rec_{CE}}$ in Eq. (3.13)) to the VAE-Loss (L_{KL} and $L_{rec_{VAE}}$ in Eq. (3.13)). A ratio of 0.0 implies that the model was trained as a VAE only, a ratio of 1.0 implies that the model was trained as a CE only, and the other ratios are differently weighted ceVAE models.
To compare the benefits of the contrastively trained features, in addition to the GMM a Flow-based method, in particular, RealNVP [Dinh et al., 2016], was used. To further show the benefits and rule out the effects of the Flow-based method, a Glow-like [Kingma and Dhariwal, 2018] model was directly trained on the raw images.

ceVAE experiments

Given the proposed framework the effect of combining a CE with a VAE is first evaluated on a slice-wise anomaly detection level. This is followed by an evaluation of the benefits of combining the reconstruction error with the gradient of the KL-Loss for a pixel-level localization of the anomalies.

Slice-wise detection The first comparison is of the different approaches on the slice-wise anomaly detection task. Fig. 3.11 shows the performance of different methods on the BraTS 2017 dataset. As expected, the OC-SVM had difficulties with high-dimensional and highly structured data [Goldstein and Uchida, 2016].

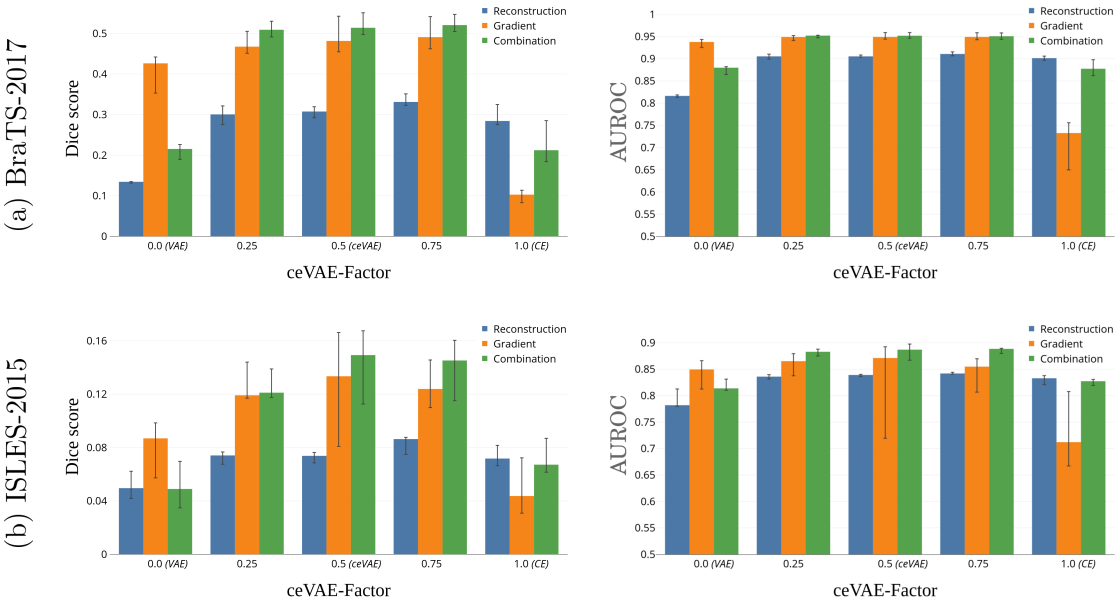


Figure 3.12: Pixel-level localization performance (DSC and AUROC) of the ceVAE for different ceVAE-Factors on the ISLES-2015 and BraTS-2017 dataset [Zimmerer et al., 2019b].

Furthermore, some more observations can be made regarding the ranking of the different AE-based models: (1) The most basic model, the AE, was already performant enough to outperform the OC-SVM. (2) Using denoising as an auxiliary task during the training helped improve the performance. (3) CE noise appears to outperform Gaussian noise as noise for a denoising objective. (4) Surprisingly, a VAE outperforms the DAE models by a margin, further indicating the importance of the KL-term. (5) The combination of a VAE with a denoising objective can further boost the performance.

Pixel-level localization For the experiments for the pixel-level localization task, we opted to focus on the CE, VAE, and ceVAE since performance on the sample-level task was convincing and VAEs and their variants have become a de-facto standard in anomaly localization for medical images [Baur et al., 2018; Chen et al., 2018; Kiran et al., 2018]. The pixel-wise AUROC and DSC on the BraTS-2017 and ISLES-2015 datasets are shown in Fig. 3.12.
After inspecting the results for the sample-wise task, the results for the VAE and CE were mostly expected: the CE performed best when using solely the reconstruction

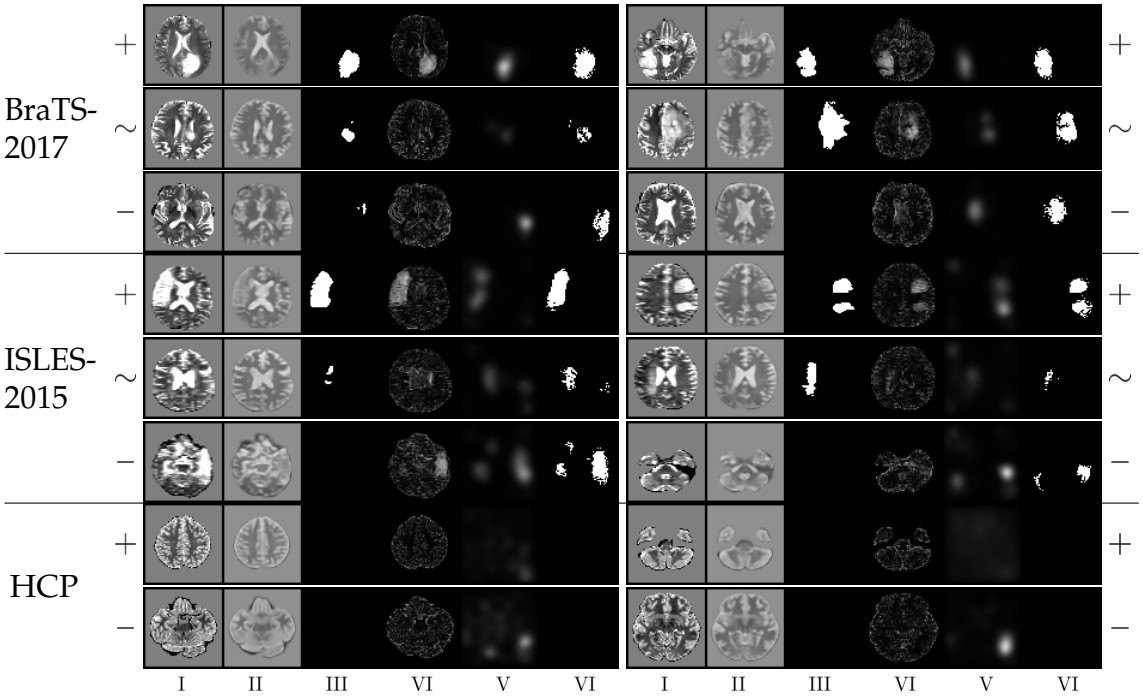


Figure 3.13: Sample images and predictions from different data sets. The 1st, 2nd, and 3rd rows show good (+), medium (~), and failure (-) cases respectively. For each sample image, the original sample (I), the reconstruction (II), the annotation (III), the reconstruction error (IV), the gradient (V), and the resulting segmentation (VI) are presented [Zimmerer et al., 2019b].

error, in contrast, the VAE performs best using solely the gradient of the KL-loss and outperforms the CE.

The ceVAE using the 'combi' score outperforms the CE and VAE for all cases. Furthermore, in total the 'combi' method also appears to yield the best results for these datasets.

If just the reconstruction error is considered, the ceVAE even improves upon the CE, possibly due to the regularizing effects described in Section 3.2.3.

An obvious insight is a difference in absolute performance between the two datasets. While the type and/or expression of the anomalies in the images are similar, one probable explanation is the difference in dataset quality and thus a distribution shift to start with, which can impede the performance of the models.

Qualitative results for each dataset are shown in Fig. 3.13.

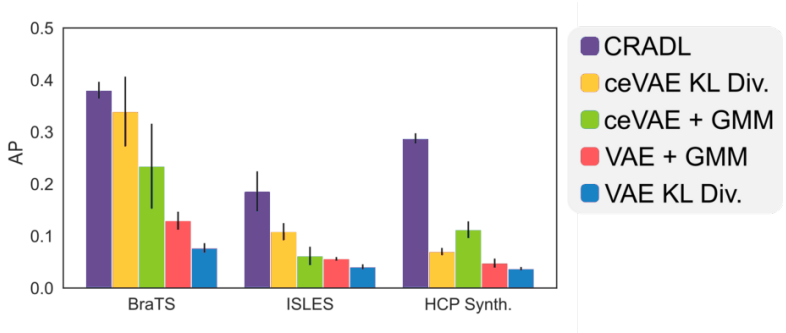


Figure 3.14: Comparison of different representations for pixel-level anomaly localization on the different data sets. Compared are features from CRADL, a ceVAE, and a VAE all with a GMM fitted to the features and the backpropagated log-likelihood (CRADL & ceVAE + GMM & VAE + GMM), and additional for VAE and ceVAE the gradient of the KL-Divergence (ceVAE KL Div. & VAE KL Div.).

CRADL experiments

Next, while we have seen the benefits self-supervised learning can bring to the anomaly localization task, we will see if the masking task can be improved upon by a contrastive learning task. Thus, in the context of anomaly localization, we first examine the discriminative capability of the representations created using contrastive learning compared to that of generative models. In the next step to evaluate the absolute performance, we compare then the anomaly localization performance with the ceVAE model.

Discriminative power of representations for anomaly localization To compare the discriminative power of representations from generative models, in particular VAE and ceVAE, and their applicability to anomaly localization we fit a GMM on their representations in an identical scheme to CRADL (see Sec. 3.2.3). While a GMM model can parameterize almost arbitrary distributions given a large enough number of components K, we believe that VAEs, due to their unimodal Gaussian prior, represent a very benign case and a small number of components should be able to capture the distribution. In addition, for the VAE models, we compare the gradient of the KL-Divergence, since the KL-divergence also can measure the feature distribution deviations and is inherent to the model.
The performances of the anomaly localization methods are depicted in Fig. 3.14. Here it becomes apparent that CRADL-based representations outperform both VAE and ceVAE-based representations for both ISLES and HCP Synth., while on BraTS the KL-Divergence of the ceVAE showed performance similar performance.

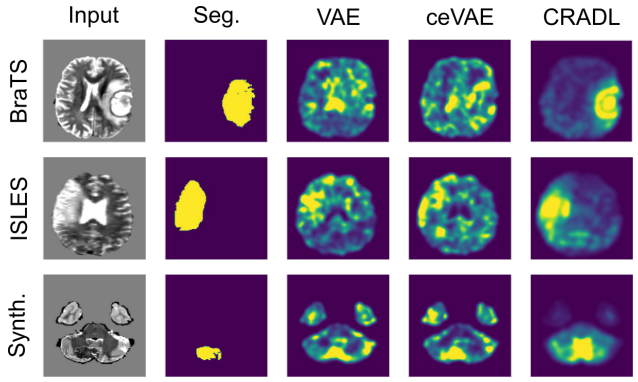


Figure 3.15: Pixel-level scores for different samples from the different datasets using the different self-supervised approaches (clamped and normalized for visual inspection).

This supports the theory that self-supervised representations may include more semantic information, allowing for a better localisation of subtle semantic distinctions between abnormal and normal brain volumes, and that contrastive learning in particular may be a task that fits well.

To verify that the main benefits stem from its contrastively learned representations and not from the GMM fitting difference between the models, we tested Flow-based Deep Generative Models additionally,(Real NVP [Dinh et al., 2016]) which show the same trend as the GMMs (for more details, see Section 6.0.3).

Comparison to ceVAE Here, we further compare CRADL with the ceVAE. The quantitative results are shown in Table 3.5 and qualitative results can be seen in Fig. 3.15. Here, for the CRADL model, we present the best-performing model, with the optimal choice of the number of GMM components K.

CRADL outperforms ceVAE on the HCP Synth. dataset by a large margin and by a small margin on the ISLES dataset. The ceVAE on the other hand outperforms CRADL on the BraTS dataset by a small margin.

On hypothesis why the strong score from CRADL on the HCP Synth. dataset does not extend to the other datasets is the domain gap between the HCP dataset on which the models were trained to the other dataset (i.e., different scanners, image quality, and the patients' overall health) and contrastive learning might be more sensitive to such changes.

Table 3.5: Pixel-wise anomaly localization metrics for the different datasets.

		CRADL	ceVAE
HCP Synth.	AUROC	$\mathbf{0.978 \pm 0.001}$	0.921 ± 0.004
	AP	$\mathbf{0.288 \pm 0.010}$	0.172 ± 0.015
ISLES	AUROC	$\mathbf{0.898 \pm 0.003}$	0.879 ± 0.002
	AP	$\mathbf{0.186 \pm 0.039}$	0.145 ± 0.013
BraTS	AUROC	0.942 ± 0.001	$\mathbf{0.948 \pm 0.003}$
	AP	0.380 ± 0.016	$\mathbf{0.483 \pm 0.003}$

3.2.5 Discussion & conclusion

Self-supervised learning and representations obtained using self-supervised learning have shown great improvements in follow-up tasks and allowed better linear separability of semantic classes. In this section, we wanted to investigate if 'more semantic' features can improve anomaly localization performance. As the results showed, integrating self-supervised learning in the form of masking into a VAE anomaly localization framework showed localization improvements and also contrastive learning on its own showed competitive performance for anomaly localization.

Similar to the literature, contrastive learning has shown to be a top-performing and easy-to-use self-supervised training task for CNNs, and this result was also mirrored in the presented anomaly localization results. However, in recent years improvements and new self-supervised tasks were proposed and continued to improve performance. In particular, masking has recently received a lot of attention in the transformer model literature, and with changes in architectures and new datasets, it is expected to be a more effective pretraining method for anomaly localization. As such we expect that there is no best or standard go-to self-supervised task for anomaly localization. Rather, integrating different self-supervised learning methods into an anomaly localization framework will be a good starting point for new architectures. Following this line of thought, we believe that as the experiments showed, integrating a self-supervised task, independent of the exact task, and thus 'more semantic' features can boost the performance of anomaly localization.

However, the current approaches only model the feature distribution on one 'network layer' or one abstraction level. This can potentially overfocus on certain types or sizes of anomalies, as some minor texture changes might only be present

in the lower-level features and are 'abstracted away' in the higher level or 'more semantic' features. Therefore, in the next chapter, we will explore the possibility of using an explicit hierarchy of features to better model the distribution of anomalies.

3.3 Improving anomaly localization with hierarchical representations

3.3.1 Motivation

In classical medical image segmentation, the U-net has proven to be the de facto standard and its variants have been applied in different areas/tasks for several years and won multiple challenges. One of the factors for the success of the U-net can be attributed to its hierarchical encoder-decoder structure. In the encoder, the local information is compressed into semantic information as the layers progress. In the decoder, the semantic information can then be used together with the local information to achieve a good and detailed segmentation. This information fusion is possible due to the skip connections, i.e. the use of the feature maps from the encoder layer on the respective decoder layer in addition to the features from the decoder layer above.

While AEs are structurally similar to the U-net, they miss these skip connections (since it would make the reconstruction task very easy and the model would not have to use the upper layers). This, however, enforces AEs to pass all the information through the bottleneck and given the choice of architecture trade local information for semantic information. The loss of local information and thus local modeling can hinder the reconstruction and potentially fine-grained anomaly localization.

However, in medical imaging VAEs have often been used for unsupervised pre-training, feature extraction and OoD / anomaly detection [Baur et al., 2018; Chen et al., 2018; Litjens et al., 2017; Shin et al., 2013; Zimmerer et al., 2018]. In this regard, due to the previously mentioned issues, VAEs are often designed to either contain 'high-level features' or model only basic image features and thus are often criticized to have over-smoothed reconstructions and only learn low-level statistics [Dai and Wipf, 2019; Larsen et al., 2015; Nalisnick et al., 2018; Razavi et al., 2019]. Especially in medical applications faithful reconstruction and small textural differences can be important.

Here, inspired by Principal Component Analysis (PCA) and Vector Quantized Variational Autoencoder (VQVAE)-2 [Razavi et al., 2019], we introduce a simple hierarchical model with a low-level reconstruction branch that enforces the partitioning into high-level and low-level components which, in our case, correspond to the coarser and the finer structure of brain MRIs. The model, which we call primary components conditional hierarchical VAE (pchVAE), shows better reconstructions than a normal VAE while having similar or slightly better

anomaly/OoD detection performance (indicating they capture the data distribution similarly well) (note: some definitions, formulations, and equations of this section were previously published by myself in [Zimmerer et al., 2019c]).

3.3.2 Related work

Recently, hierarchical latent models [Maaløe et al., 2019; Sønderby et al., 2016; Zhao et al., 2017], which can model features on different network layers and abstraction levels, have been proposed. However, studies have shown that when learning and differentiating between high- and low-level features, these hierarchical models currently exhibit only a minor or no gain compared to conventional VAEs [Maaløe et al., 2019; Zhao et al., 2017]. One example is the Ladder VAE [Sønderby et al., 2016] which has been shown to collapse to a single-level model in practice [Zhao et al., 2017]. However, by utilizing a particularly deep multi-stage design where each level is dependent upon the lower and higher levels, BIVA [Maaløe et al., 2019] has been demonstrated to provide some benefits. Child [2021] proposes a hierarchical extension to VAEs, which show sharp generated images and comparable performance to autoregressive models, even for high-resolution images. However, they do not analyze its applicability for anomaly detection. One approach that models multiple feature hierarchies for out-of-distribution detection was proposed by Song et al. [2019]. In detail, they repurpose the learned batch normalization layer statistics to find batches for which these deviate for a test data batch. This allows the detection of out-of-distribution samples on a per-layer basis, but requires batching of data samples, is trained in a supervised manner, and different layer statistics are not combinable or comparable.

3.3.3 Methodology

Inspired by recent research, which found that VAEs pursue PCA directions (by accident) [Rolinek et al., 2018] and the fact that VAEs also optimize the mutual information between the input and the latent variables (which we will show later), we propose the optimization problem for a non-linear VAE with two components (in contrast to PCA which maximizes the variance for the components, this maximizes the mutual information of the input and the first component) [Zimmerer et al., 2019c].
We derive this optimization problem, starting from the following optimization problem for a linear hierarchical AE with input X and weights $w1, w2$, where the first term models the 'main' component and the second term models the 'residual':

$$\min_{w_1,w_2} \lambda_1 ||X - w_1 w_1^\top X|| + \lambda_2 ||(X - w_1 w_1^\top X) - w_2 w_2^\top (X - w_1 w_1^\top X)||. \tag{3.19}$$

Given this, we derive a similar optimization problem:

$$\begin{aligned}
&\lambda_1 ||X - w_1 w_1^\top X|| + \lambda_2 ||(X - w_1 w_1^\top X) - w_2 w_2^\top (X - w_1 w_1^\top X)|| \\
=&\lambda_1 ||X - w_1 w_1^\top X|| + \lambda_2 ||X - w_1 w_1^\top X - w_2 w_2^\top X + w_2 w_2^\top w_1 w_1^\top X)|| \\
\text{(triangle inequality)}& \\
\leqslant&\lambda_1 ||X - w_1 w_1^\top X|| + \lambda_2 ||X - w_1 w_1^\top X - w_2 w_2^\top X|| + \lambda_2 ||w_2 w_2^\top w_1 w_1^\top X)|| \\
=&\lambda_1 ||X - w_1 w_1^\top X|| + \lambda_2 ||X - (w_1 w_1^\top X + w_2 w_2^\top X)|| + \lambda_2 ||w_2 w_2^\top w_1 w_1^\top X||.
\end{aligned} \tag{3.20}$$

For further modularity, we introduce λ_3 instead of λ_2 for the third term. However, in practice we chose $\lambda_2 = \lambda_3$:

$$\min_{w_1,w_2} \lambda_1 ||X - w_1 w_1^\top X|| + \lambda_2 ||X - (w_1 w_1^\top X + w_2 w_2^\top X)|| + \lambda_3 ||w_2 w_2^\top w_1 w_1^\top X||. \tag{3.21}$$

Transferring this problem to a non-linear NN model, i.e. substituting arbitrary non-linear functions parameterized by neural networks for the weight matrices, by amortizing the optimization over mini-batches/samples (as indicated by the lower case x), and using conditional models results in:

$$\begin{aligned}
w_1 w_1^\top X &\xrightarrow{\text{implemented as}} g_{\theta_1}(f_{\gamma_1}(x)), \\
w_2 w_2^\top X &\xrightarrow{\text{implemented as}} g_{\theta_2}(f_{\gamma_1}(x), f_{\gamma_2}(x)), \\
w_2 w_2^\top w_1 w_1^\top X &\xrightarrow{\text{implemented as}} g_{\theta_1}(f_{\gamma_1}(g_{\theta_2}(f_{\gamma_1}(x), f_{\gamma_2}(x))).
\end{aligned}$$

Here $f_{\gamma_1}, f_{\gamma_2}$ are encoders (sharing the first layers) and $g_{\theta_1}, g_{\theta_2}$ are decoders. This results in the following optimization problem:

$$\begin{aligned}
\min_{\theta_1,\theta_2,\gamma_1,\gamma_2} \lambda_1 ||x - g_{\theta_1}(f_{\gamma_1}(x))|| + \lambda_2 ||x - (g_{\theta_1}(f_{\gamma_1}(x)) + g_{\theta_2}(f_{\gamma_1}(x), f_{\gamma_2}(x)))|| \\
+\lambda_3 ||g_{\theta_1}(f_{\gamma_1}(g_{\theta_2}(f_{\gamma_1}(x), f_{\gamma_2}(x)))||.
\end{aligned} \tag{3.22}$$

From an information-theoretical perspective, the last term can be interpreted as minimizing the mutual information between the low-level components and the high-level components [Chen et al., 2016] as we will show in the next paragraph.

As there are some similarities between this optimization problem and the VQVAE-2, this formulation can be transferred to a conditional hierarchical VAE with a similar architecture to VQVAE-2, by using a normal prior for the latent variables z_1, z_2 and condition g_{θ_2} not just on z_2 ($\sim f_{\gamma_2}(x)$) but also on z_1 ($\sim f_{\gamma_1}(x)$).
As VAEs have proven themselves to show good performance in our anomaly detection tasks and since we are thus interested to find the latent factors of a generative hierarchical model we integrate this into the variational auto-encoding framework, in analogy to [Kingma and Welling, 2013; Rezende et al., 2014],:

$$\begin{aligned}
L =& \lambda_1 \mathbb{E}_{z_1 \sim \mathcal{N}(f_{\gamma_1,\mu}(x), f_{\gamma_1,\sigma}(x))} \mathcal{N}(x; g_{\theta_1}(z_1), c) \\
&+ D_{KL}(\mathcal{N}(f_{\gamma_1,\mu}(x), f_{\gamma_1,\sigma}(x)) || \mathcal{N}(0,1)) \\
&+ \lambda_2 \mathbb{E}_{z_1 \sim \mathcal{N}(f_{\gamma_1,\mu}(x), f_{\gamma_1,\sigma}(x))} \mathbb{E}_{z_2 \sim \mathcal{N}(f_{\gamma_2,\mu}(x,z_1), f_{\gamma_2,\sigma}(x,z_1))} \mathcal{N}(x; g_{\theta_1}(z_1) + g_{\theta_2}(z_1, z_2), c) \\
&+ D_{KL}(\mathcal{N}(f_{\gamma_2,\mu}(x, z_1), f_{\gamma_2,\sigma}(x, z_1)) || \mathcal{N}(0,1)) \\
&+ \lambda_3 \mathbb{E}_{z_1 \sim \mathcal{N}(f_{\gamma_1,\mu}(x), f_{\gamma_1,\sigma}(x))} \mathbb{E}_{z_2 \sim \mathcal{N}(f_{\gamma_2,\mu}(x,z_1), f_{\gamma_2,\sigma}(x,z_1))} \\
&+ \mathbb{E}_{x_{low} \sim \mathcal{N}(g_{\theta_2}(z_1,z_2))} \mathbb{E}_{z_{low} \sim \mathcal{N}(f_{\gamma_1,\mu}(x_{low}), f_{\gamma_1,\sigma}(x_{low}))} \mathcal{N}(0; g_{\theta_1}(z_{low}), c).
\end{aligned} \tag{3.23}$$

Here, as previously mentioned, the last term can be interpreted as encouraging the mutual information in the conditioned low-level component x_{low} and the high-level encoding to be zero. Using MC sampling with sampling size 1, choosing c appropriately, and integrating the reparametrization step into the decoders g gives the final and familiar loss function:

$$\begin{aligned}
L = & \lambda_1 ||x - g_{\theta_1}(f_{\gamma_1}(x))|| \\
& + \lambda_2 ||x - (g_{\theta_1}(f_{\gamma_1}(x)) + g_{\theta_2}(f_{\gamma_1}(x), f_{\gamma_2}(x)))|| \\
& + \lambda_3 ||g_{\theta_1}(f_{\gamma_1}(g_{\theta_2}(f_{\gamma_1}(x), f_{\gamma_2}(x))))|| \\
& + \lambda_4 D_{KL}(\mathcal{N}(f_{\gamma_1,\mu}(x), f_{\gamma_1,\sigma}(x)) || \mathcal{N}(0,1)) \\
& + \lambda_5 D_{KL}(\mathcal{N}(f_{\gamma_2,\mu}(x, f_{\gamma_1}(x)), f_{\gamma_2,\sigma}(x, f_{\gamma_1}(x))).
\end{aligned} \tag{3.24}$$

An overview of the complete architecture can be seen in Fig. 3.16.

AEs and mutual information For AEs, it can be shown that optimizing a reconstruction objective is similar to optimizing the mutual information between the input and the latent space [Chen et al., 2016]. Given the input x, an encoder f, a decoder g and the encoding/ latent space $f(x) = z$, we can express the mutual information between the input x and its inference/ encoding distribution $F(x)$ as :

$$\begin{aligned}\mathcal{I}(x;F(x)) &= H(x) - H(x|F(x)) \\ &= \mathbb{E}_{z\sim F(x)}[\mathbb{E}_{x'\sim P(x|z)}[\log P(x'|z)]] + H(x) \\ &= \mathbb{E}_{z\sim F(x)}[D_{KL}(P(\cdot|x)||G(\cdot|x)) + \mathbb{E}_{x'\sim P(x|z)}[\log G(x'|z)]] + H(x) \\ &\geqslant \mathbb{E}_{z\sim f(x)}[\mathbb{E}_{x'\sim P(x|z)}[\log G(x'|z)]] + H(x),\end{aligned} \tag{3.25}$$

where $P(x|z)$ is the intractable generative distribution and we make use of an auxiliary distribution $G(x|z)$ to approximate $P(x|z)$.
Using Lemma A. 1 from [Chen et al., 2016] can simplify the equation and eliminate the intractable sampling from $x' \sim P(x|z)$:

$$\begin{aligned}\mathcal{I}(x;F(x)) &= H(x) - H(x|F(x)) \\ &\geqslant \mathbb{E}_{z\sim F(x)}[\mathbb{E}_{x'\sim P(x|z)}[\log G(x'|z)]] + H(x) \\ &= \mathbb{E}_{x\sim P(x),z\sim F(x)}[\log G(x|z)]] + H(x).\end{aligned} \tag{3.26}$$

Here, in analogy to VAE, some assumptions can be made that show that this corresponds to a reconstruction objective. Choosing $G(x|z)$ as a normal distribution with constant variance and the mean parameterized by a NN g can collapse $\log G(x|z)$ back to a NN decoder with an Mean Squared Error (MSE)/L2 reconstruction loss. Using the reparametrization trick and MC sampling where F, the inference/ encoding distribution is parameterized by a NN f resolves the expectation. Furthermore $H(x)$ is given by the data distribution and can be assumed to be constant. Thus optimizing the reconstruction loss of an AE with MSE can be interpreted as optimizing the mutual information between the input and the latent space.

3.3.4 Experiments & results

As one goal was to be able to detect more 'fine-grained' anomalous structures, we focus the evaluation on two aspects: (1) the reconstruction performance on a held-back test set and (2) the semantic OoD [Ahmed and Courville, 2019] / anomaly detection performance.
Both aspects can hopefully give truthful insights on how well the data distribution was captured, i.e. (1) if the data is not reconstructed truthfully the model has probably not learned the data distribution to the whole extent, especially fine-grained details and (2) if the anomaly score is not sufficient to tell apart normal data from anomalous data, the model has probability issues capturing the basic structure of the data. We refrain from using a likelihood-based comparison since

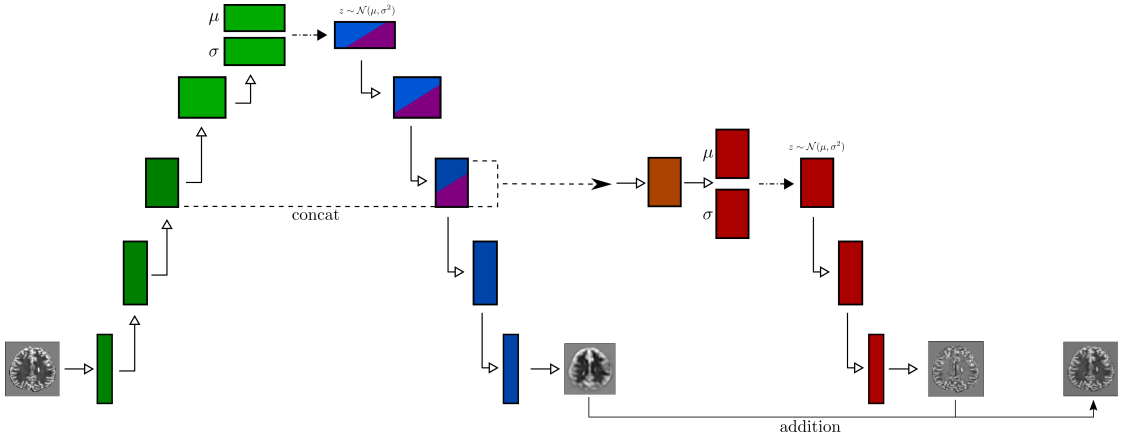


Figure 3.16: Depiction of the pchVAE model. The green part represents the encoder, consisting of the low-level encoder f_1 (dark green) and the high-level encoder f_2 (lighter green). The blue part represents the high-level decoder g_1 and the red part the low-level decoder g_2. The arrows with straight lines and white heads are (transposed) convolution operations, where the up-/downward convolutions have an up-/down-sampling factor of two [Zimmerer et al., 2019c].

likelihood-based comparisons have often been criticized to be easily influenced by other factors [Nalisnick et al., 2018; Theis et al., 2015], and thus we believe that the other metrics are more expressive for anomaly detection.

Data Here, we again focus on brain MRI scans. All models are trained on brain MRIs from a subset of 800 scans of the HCP dataset [Van Essen et al., 2012]. The reconstruction performance is evaluated on 200 held-back patients from the HCP dataset and the OoD/ anomaly detection performance on two different datasets: (1) Similar to HCP Synth., we use the 200 held-back patients from the HCP dataset and randomly render natural objects in the brain area of some slices, which are then considered anomalies. (2) As the second dataset we use the BraTS2017 [Bakas et al., 2017; Menze et al., 2015] dataset with slices with tumor annotations being considered as anomalies.

Models To factor out the effects of the different hierarchical levels in the pchVAE we compare with the following models: (1) The VQVAE-2 ([Razavi et al., 2019]) model, which has a similar hierarchical structure and is being applied in many different areas for image quantization. (2) The "high-level" VAE (High VAE), which is similar to the high-level component branch of the pchVAE (the green

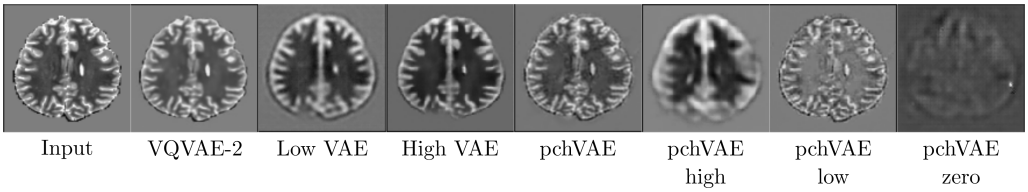


Figure 3.17: Reconstructed input of different models given the same input. Reconstructed images for the VQVAE-2, the Low VAE, the High VAE, and the pchVAE are shown. The reconstruction of pchVAE is furthermore split into the high-level components (Eq. (3.22) term 1) and the low-level components (Eq. (3.22) term 2) and as 'pchVAE zero' the low-level components passed through the high-level encoder & decoder which is steered towards zero during training (Eq. (3.22) term 3). More examples can be found in Fig. 6.3 and Fig. 6.4 [Zimmerer et al., 2019c].

and blue parts in Fig. 3.16) with 4 up-sampling operations. (3) The "low-level" VAE (Low VAE) only consists of the low-level part of the pchVAE model (the dark green and dark blue parts in Fig. 3.16) with 2 up-sampling operations. (4) A default conditional hierarchical VAE (chVAE) (similar to the model in Fig. 3.16 but without the dark blue part and term 3 in Eq. (3.22)). This division of the pchVAE into several VAE variations may provide insights into the advantages that the hierarchical structure and mutual information objective provide to the pchVAE model. All models are upscaled to have slightly more parameters and the same number of latent variables as the pchVAE. Each model was trained for 10 epochs (which based on a validation set showed convergence for all models) with Adam [Kingma and Ba, 2017] with a learning rate of 0.0001 and a batch size of 64. As is common practice [An and Cho, 2015; Kiran et al., 2018], the approximated ELBO is used as OoD/ anomaly score for each slice.

Results We report the MSE for reconstruction performance and AUROC and AP (as suggested by Ahmed and Courville [2019]) for OoD/anomaly detection performance. The results are presented in Table 3.6 and Fig. 3.17 and indicate that the pchVAE performs similarly to or slightly better than the High VAE on the OoD detection task, with the pchVAE having greater reconstruction performance. The Low VAE and VQVAE-2 do significantly better on the reconstruction job, nevertheless. But these two models perform poorly according to the OoD detection task, indicating that they have concentrated more on low-level statistics of the data and frequently fail to capture the data distribution in this situation.

	Reconstruction	OoD	
	MSE ↓	AUROC ↑	AP ↑
VQVAE-2	0.02823 ± 0.0001	0.6965 ± 0.0036	0.4585 ± 0.0058
Low VAE	$\mathbf{0.01169} \pm 0.0002$	0.7101 ± 0.0010	0.4922 ± 0.0025
High VAE	0.07036 ± 0.0007	0.7207 ± 0.0005	0.5154 ± 0.0010
chVAE	0.02124 ± 0.0040	0.6716 ± 0.0009	0.4341 ± 0.0008
pchVAE	0.03224 ± 0.0017	$\mathbf{0.7277} \pm 0.0002$	$\mathbf{0.5321} \pm 0.0017$

Table 3.6: Mean and standard deviation of reconstruction and sample-level anomaly detection performance of the different hierarchical models (over five different runs).

3.3.5 Discussion & conclusion

The results suggest that the pchVAE performs reasonably well. By including a conditioned low-level reconstruction branch and an extra forward pass into a VAE, we could trade off computational complexity for improved reconstruction performance. We think that a strong reconstruction performance with attention to detail is crucial, particularly for tasks involving the localization and identification of medical anomalies.

More realistic and sharper reconstructions have also been the goal of other methods like adversarial losses [Baur et al., 2018; Larsen et al., 2015]. However, even though these adversarial losses produce sharper reconstructions, because of their design they are unable to 'generate' more low-level information and frequently "make up" convincing information, which can increase the MSE and may be hazardous in medical applications.

Overall we could overserve that 'deeper' VAEs with more high-level features were better at capturing the data distribution and thus detect out-of-distribution / anomalous samples. At the same time, the 'deeper' VAEs struggled with a high-fidelity reconstruction. 'Shallower' VAEs had lower reconstruction errors, but they were not able to capture the data distribution as well. Here the pchVAE, a hierarchical model, which is composed of VAEs at different levels and due to the mutual information/residual information component has a clear hierarchical structure of the latent spaces as well, shows a low reconstruction error and top OoD/anomaly detection performance.

The total anomaly detection performance of the pchVAE is comparable to that of the ceVAE but in no way superior. However, a hierarchical model is orthogonal to the use of self-supervised pretraining, and a combination in a suitable form might further boost the performance (even more than the individual contributions) and

might present a valuable direction for further research.
However, all the previous experiments were a bit constrained by the choice of dataset, anomalies, and their overall setting and the assumptions made during the experiments might not translate well to a more applied setting. Therefore we try to address these issues and focus on the benchmarking and comparison of anomaly detection and localization methods in the following chapter.

4 | Performance evaluation beyond the standard setting

While it is necessary to create and advance techniques, it is also necessary to obtain a truthful and valid evaluation of the methods' performance. In contrast to the previous chapter which focused mainly on the methodological development of methods in a fixed/research/"standard" setting, this chapter will focus on the evaluation of the performance of the methods beyond the standard setting. The first section will focus on a fair and standardized comparison between methods in form of an international challenge/benchmark and the second section will focus more on the evaluation of the methods in a real medical use case setting.

4.1 Validation in an international competitive context: The Medical Out-Of-Distribution Analysis Challenge (MOOD)

Besides some of the examples shown in the previous chapter, there is a lot of recent research on improving anomaly detection [Abati et al., 2018b; Ahmed and Courville, 2019; Akcay et al., 2018; Beggel et al., 2019; Choi et al., 2018; Guggilam et al., 2019; Maaløe et al., 2019; Piciarelli et al., 2019; Sabokrou et al., 2018] particularly in the medical imaging field [Baur et al., 2018, 2021; Chen et al., 2018; Schlegl et al., 2017; Zimmerer et al., 2019a]. However, in the medical imaging field, most approaches were validated in slightly different settings and on different datasets. A public benchmark or dataset was missing.
Benchmarks have shown great potential and responsibility in comparing different approaches on a plain playing field, outlining the different strengths and weaknesses of each approach and promoting directions worth investigating further. While for other areas, such as for tabular medical data [Avati et al., 2021; Ulmer et al., 2020] as well as natural images, such as default detection [Bergmann et al., 2019a] or abnormal traffic scene detection [Hendrycks et al., 2019], different benchmarks have recently been proposed, anomaly detection on medical images still lacked such a benchmark.
So with The Medical Out-Of-Distribution Analysis Challenge (MOOD) [Zimmerer et al., 2020, 2022a], we aimed at filling this void and wanted to create a standardized dataset and benchmark for anomaly detection. However, there are several aspects to be considered when creating a benchmark for anomaly detection: (1) There should be no domain gap between the training and test data distribution, as this might inherently already encompass a domain gap (e.g. due to the different image acquisition protocols or devices) and unnecessarily introduce further unwanted factors. These unwanted factors may impede a clean and meaningful evaluation. Thus, combining different test sets which differ from the training set to introduce multiple types of anomalies might not be a good choice. (2) To have a realistic and close-to-real-world setting, the types of anomalies in the test set should not be known beforehand. Knowing this beforehand can potentially cause 'leakage' of the test set into the algorithm development and thus bias the developed algorithm and may hinder its generalizability to new and unseen anomalies. However, particularly in anomaly detection, the generalizability to new and unseen anomalies is one of the most crucial points and leads to an overconfident method evaluation and in the long term potentially dangerous outcome when deploying such methods in practice. Recent studies have shown that this has been

observed in practice [Goldstein, 2014; Shafaei et al., 2019; Zimmerer et al., 2019a; Škvára et al., 2018]. Hence, knowing what anomalies to expect can complicate the comparability of different algorithms.

To consider the two aspects and solve the issues in the MOOD Challenge, we included two separate datasets: a brain MRI dataset and an abdominal CT dataset. For both datasets, the training set was selected as a subset of scans in which no anomalies were identified. The test set was comprised of a different disjunct subset of scans in which no anomalies were identified and a subset containing naturally occurring anomalies. In addition, we extended the test set with synthetic/artificial anomalies with different properties. For the synthetic anomalies, we included a wide variety of different anomalies (e.g. a tumor or an image of a gorilla rendered into the brain scan [Drew et al., 2013]) to outline the weaknesses and strengths of the methods using different factors (e.g. type, size, contrast, and others). Overall, this may facilitate a fairer comparison of the generalization capabilities of the different approaches across different anatomies and modalities.

The MOOD Challenge was further split into two different tasks: a sample-level anomaly detection task and a pixel-level anomaly localization task. The sample-level task aims to detect and classify scans on a per-scan basis. Here, examples of anomalies could be previously unseen pathological conditions or image acquisition artifacts not available in the training set. Identification of these cases could allow physicians to distrust results obtained from (not specifically designed) supervised algorithms or prioritize scans for manual inspection. The pixel-level task aims at localization of the anomaly and pixel-level scoring of the scan, i.e. assigning each pixel an individual anomaly score. This can highlight abnormal regions in the scan and by guiding the physicians' attention potentially provide a more accurate assessment of the scan.

Overall this was organized as an international open challenge with the aim of a controlled and fair comparison of different anomaly detection algorithms in a variety of both real-life and simulated cases. As this was implemented as a MICCAI challenge, a detailed description of the challenge design according to the BIAS statement [Maier-Hein et al., 2020] is available [Zimmerer et al., 2020].

In the next sections, we will describe the challenge setup in more detail, in particular, the datasets, the tasks, and our evaluation procedure. Then we will first summarize the submitted approaches and present and analyze the results (note: some definitions and formulations of this section were previously published by myself in [Zimmerer et al., 2022a]).

4.1.1 Data

In the challenge, we included two different datasets, a brain MRI dataset and an abdominal CT dataset. The training set was selected as a subset of scans in which no anomalies were identified. The test set was comprised of another subset of scans in no which anomalies were identified and a subset containing naturally occurring anomalies combined with artificial/ synthetic anomalies.

Datasets

While the two datasets represent different modalities and anatomies, they are both prepared in the same way and are therefore comparable. The submitted algorithms were expected to be able to handle both datasets, only allowing a change of hyperparameters and individual training. The performance on both datasets was equally important. Next, we will describe the datasets in more detail:
Brain: The brain dataset is based on the HCP dataset [Van Essen et al., 2012] and contains 3T brain MRI images of young healthy adult participants (ages 22-35). All scans were recorded using the same protocol and equipment and were processed according to the same pipeline, given in [Van Essen et al., 2012].
Abdominal: The abdominal dataset is based on a CT colonoscopy dataset [Johnson et al., 2008]. The dataset contains CT scans of female and male patients >50 years, which were scheduled for a screening colonoscopy and had not had a colonoscopy in the past 5 years. The scans were recorded at 15 study centers using standard bowel preparation, stool and fluid tagging, mechanical insufflation, and multi-detector row CT scanners (with 16 or more rows). In contrast to the brain dataset, the abdominal dataset contains some anatomical variations to some degree, e.g. polyps that were not considered abnormal. Only cases with severe or rare naturally occurring anomalies were considered to be abnormal.
Furthermore, for each dataset, four toy cases consisting of three scans with toy anomalies, i.e. a sphere with random intensity placed into a scan, and one normal scan were available for sanity checks. The toy cases were not included in the test set.

Challenge preprocessing

In addition to the dataset-specific processing, both datasets were further preprocessed with the same additional challenge-specific preprocessing. We applied the following transformations: cropping, intensity shift and resampling. Since there was no difference in the preprocessing between the training set and test set, no additional domain shift was introduced here.

Anomalies

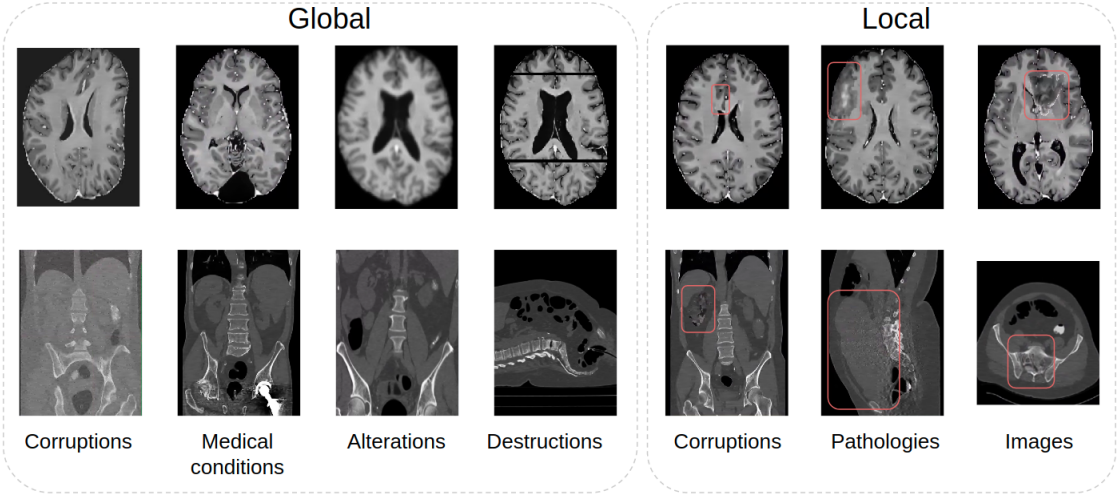


Figure 4.1: The anomaly categories of the challenge. The seven different categories of anomalies are divided into 4 global (affecting the whole scans) and 3 local (affecting only parts of the image) categories. One example for each, the brain and abdominal dataset, is given (some anomalies have been exaggerated for illustration purposes) [Zimmerer et al., 2022a].

For the anomalies, we can make two broad differentiations: local vs. global and natural vs. artificial.
Local vs. global: The local anomalies can be constrained to a specific region in the image and can be assigned a clear segmentation mask. The local anomalies were used in the sample-level and pixel-level tasks. In contrast, the global anomalies could not be assigned a clear segmentation mask or extended to the whole image. Consequently, the global anomalies were only assigned to the test set for the sample-level task.
Natural vs. artificial: All scans in the dataset were reviewed manually (multiple times by two human annotators using a consensus annotation protocol) and classified as normal (containing no abnormalities) or abnormal. The scans with naturally occurring anomalies were assigned to the test set as natural global abnormal samples. The artificial anomalies on the other hand were designed to emulate a broad and partially unpredictable range of anomalies.
For the further analysis of the strength and weaknesses, we further categorized the anomalies as follows (and see Fig. 4.1 for a visualization):
Local anomalies:

- **Images**: Actual images embedded into the scans in line with [Drew et al., 2013].
- **(local) Pathologies**: Various local diseases, like tumors or lesions, superimposed on the healthy scans.
- **Corruptions**: Local picture corruptions, such as local contrast change or local pixel shuffling.

Global anomalies (sorted from mild to strong corruptions to the images):

- **Corruptions**: Small global image corruptions, such as deformations, that generate a legitimate image but are only detectable with a large quantity of training data.
- **(global) Medical conditions**: Rarely occurring medical conditions/variations (which were deemed global variations since these anomalies were usually not confined to a specific place).
- **Alterations**: Global scan alterations that result in a legitimate scan but should be immediately evident, such as severe blurring.
- **Destructions**: Operations conducted on the scan that corrupt or invalidate the entire scan, such as omitting slices.

Overall, the samples that had no annotations and were considered normal and were split into the training set and test set. For the test set, one part of the normal scans were used as base samples to create the abnormal test cases by adding the artificial anomalies. Another part of the normal scans were used as the normal part of the test. Thus there is no domain shift between the training cases and the normal test cases. The artificial anomalies were generated synthetically and thus allow for a perfect pixel-level segmentation mask or sample-level label for each image.
However, despite this controlled setting, there might be different sources of errors related to the annotations. First, true anomalies may be missed by the human annotators and appear in the training set (e.g. polyps that were not detected or a patient with an abnormal kidney).
But for a case like this, it would be the same premise for all methods, and while potentially favoring more robust methods, given the large number of training cases, we believe this is not a problem and will not change the results to a large extent. Second, some of the artificially introduced anomalies could coincidentally

be very similar to some of the 'as normal' considered anatomical variations in the training set. However, due to the manual review and design of the artificial anomalies, this is very unlikely and if it occurs, it is again equally challenging for all approaches and thus might not influence the results of the methods in a significant way. Furthermore, the software used to generate the artificial anomalies has been tested beforehand for some years and has undergone stringent in-house testing.

Dataset ratios

As the artificial anomalies in the test set can be created in an almost arbitrary number, we were not bound by a total number of test cases. However, to prevent any bias and fine-tuning of the scores, we chose a balanced dataset with a roughly 50%-50% split and did not disclose the exact number of test cases to the participants. Hence, considering reasonable time constraints for the evaluation and the number of available normal samples, we chose the following number of cases. Brain dataset: 800 training samples, 688 sample-level and 542 pixel-level test cases. Abdominal dataset: 550 training samples, 599 sample-level and 358 pixel-level test cases.

4.1.2 Challenge setup

This challenge was run as a challenge at MICCAI 2020 (and thus the challenge design was reviewed and accepted by two independent reviewers and a meta-reviewer). The MOOD Challenge consisted of two tasks:

Sample-level The objective of the sample-level task was to report one score/label for each sample. This score for a sample should indicate a "probability" of how likely it is that this sample is abnormal. The reported scores were expected to be in the range of [0-1], where 0 indicates no abnormality and 1 indicates the most abnormal input. Scores outside [0-1] were clamped to [0-1] and missing scores were set to 0.

Pixel-level The objective of the pixel-level task was to report a score for each pixel/voxel of a sample (similar to segmentation). Again, this score should indicate a "probability" of how likely it is that a pixel in a given sample is abnormal. The reported scores were expected to be in the range of [0-1], where 0 indicates no abnormality and 1 indicates the most abnormal input. Scores outside [0-1] were clamped to [0-1] and missing scores were set to 0.

Challenge timeline

The challenge started with the release of the training data on 01.05.2020. The submission system allowed submissions until the end of the challenge. The challenge ended on 07.09.2020. The final results were then announced on 08.10.2020 at MICCAI and were consecutively made public.

Evaluation process

The submission was implemented using the Synapse platform [Synapse, 2020] and a local GPU cluster. Each method could be submitted as a self-contained docker container with an entry point to the scoring script, which was executed on the GPU cluster. A runtime of 600 sec/case was allotted during the evaluation. Each team was allowed 10 submissions in total, however, only the last submission was considered. After submission and execution, a report containing the performance of the four toy cases and the runtime was sent to the participants.
To check the compatibility of the docker-container with the submission system, the participants could also submit their docker-container to a 'toy-cases only' queue, which only reported back the performance of the docker run on the GPU cluster with only the toy-cases. Since the participants were provided the toy-cases beforehand they could validate the consistency of their submission on the evaluation platform. Submissions on this 'toy-cases only' queue did not count in any way towards the final results.

Metrics & scoring

As the metric for both tasks, we opted to use the AP to evaluate the predicted scores together with the ground truth label.
The AP was calculated as follows:

$$AP = \sum_n (R_n - R_{n-1}) P_n, \tag{4.1}$$

where R_n is the recall and P_n is the precision at the n-th threshold.
For this use case, the AP was chosen as the primary metric since it is more robust than AUROC in terms of class imbalance and has been suggested and used in many recent papers [Ahmed and Courville, 2019; Bergmann et al., 2019b; Chen et al., 2018; Hendrycks et al., 2019; Zimmerer et al., 2019a]. While the previous papers focused on benefits of the AP in the sample-level setting, most medical papers use this metric in the pixel-level setting as well.
For the sample-level task, the AP was computed over all samples at once.

However, due to computational and time constraints, for the pixel-level task, the AP was computed over batches of 20 samples. The batches were randomly selected and contained each submission once. The arrangement of the batches was fixed and consistent for all participants. To reduce the variance due to the batching process, the dataset was batched two times and the second run over the dataset was used to validate the first run.
To combine the rankings of the two datasets into a final ranking, a consolidation ranking schema was used.
Our validation code was released publicly in conjunction with the dataset [GitHub, 2021].

4.1.3 Participating teams

In the end, 65 teams registered for the challenge and data access, of which 11 actively submitted to the toy-cases queue. However, only 8 valid methods were submitted for each task. Two teams chose not to further engage in the challenge analysis and are only represented anonymized in the final analysis. A summary of the other valid submissions is given in the following:

Team: Canon Medical Research Europe

The Team Canon Medical Research Europe uses a combination of two models: first, a DAE with Gaussian noise which uses the reconstruction error as the anomaly score and second, a segmentation model which is trained to segment synthetical spherical anomalies rendered into the scans. For the segmentation model, the class probabilities are used as the anomaly scores. Both models are combined by a weighted sum of the individual scores which directly results in the pixel-level scoring. The sample-level result is obtained by taking the mean pixel-level anomaly for each sample.

Team: FPI

The Team Foreign Patch Interpolation (FPI) also frames the anomaly detection task as a segmentation task. They employ a wide residual encoder-decoder U-Net to predict a pixel-level segmentation mask of an into the image interpolated foreign image patch. The rectangular foreign patch is chosen from another scan on a similar localization and then interpolated with the target patch. The interpolation factor, patch size, and patch location are randomly sampled from uniform distributions. The network is then trained to segment the foreign patch and predict its interpolation factor. This potentially forces the network to detect any foreign

or not normal objects in the target scan. The segmentation logits are used as the anomaly scores [Tan et al., 2020].

Team: Nina Tuluptceva

The Team Nina Tuluptceva uses the reconstruction difference of a Deep Perceptual AE [Tuluptceva et al., 2020] in a perceptual feature space. The Deep Perceptual AE uses a feature difference in a "perceptual" feature space as a reconstruction error. To transfer the images into a meaningful feature space a VGG19 [Simonyan and Zisserman, 2015] network was pretrained self-supervised using SimCLR [Chen et al., 2020] and then activations at certain layers from the trained network were used as features. The team trained three different Deep Perceptual AEs, one for each image direction, and then combined the three models by averaging the respective reconstruction errors. To get a pixel-level reconstruction error from a feature-level reconstruction, the feature differences were rescaled and mapped to the original pixel space.

Team: NUDT

Team NUDT used a U-Net as DAE with an additional Canny operator reconstruction loss. As a denoising task, an inpainting objective was chosen. The reconstruction error was a combination between an MSE and the feature difference of features extracted by a Canny operator to enforce more texture consistency. The reconstruction difference was also chosen as the anomaly score with further postprocessing which was comprised of connected component analysis and removal of objects with less than 100 voxels.

Team: Sergio Naval Marimont et al.

The Team: Sergio Naval Marimont et al. used a two-stage approach: A VQVAE as feature extraction and PixelSNAIL as autoregressive density model to model the feature distribution. The VQVAE is an AE with a categorical latent space and thus, after training, can encode the images into a lower-dimensional categorical feature space. The distribution of the features is then modeled in an auto-regressive manner using PixelSNAIL. The sample-level anomaly scoring is obtained by taking the log-likelihood of the features. For the pixel-level anomaly score, latent variables above a certain threshold are resampled using the autoregressive model and then the updated latent variables are decoded to get a pseudo-normal reconstructed image. The pixel-level anomaly score is then calculated as the L1 distance between

the original image and the pseudo-normal reconstructed image [Marimont and Tarroni, 2020].

Team: Victor Saase

Team Victor Saase used a non-deep-learning reconstruction-based method using PCA. They used extensive preprocessing with affine registration to the MNI space, sample-wise z-normalization across all brain mask voxels, and voxel-wise z-transformation with the mean and standard deviation estimated on the training samples. The processed images were used to learn a projection to a "normal" vector space using PCA. During testing, the (preprocessed) images were also projected to the vector space and the residual vector was used as pixel-level anomaly score and its norm as sample-level score [Saase et al., 2020].

4.1.4 Results

Next, we will show the performance of the different methods in the official challenge results. After presenting the results, we will investigate if the provided toy cases were predictive enough to determine the final challenge ranking. Last, we analyze how the performance of the methods is influenced by anomaly size and color contrast and anomaly type and extend this to an estimation of performance in a clinical application setting.

4.1.5 Challenge ranking

Sample-level results

Table 4.1: The ranking of the sample-level task with the performance on each dataset given as AP.

Rank	Team	Brain	Abdom.
1.	FPI	0.962	0.874
2.	Sergio Naval Marimont, et al.	0.873	0.874
3.	Canon Medical Research Europe	0.845	0.871
4.	NUDT	0.792	0.876
5.	Nina Tuluptceva	0.840	0.861
6.	A1	0.831	0.780
7.	Victor Saase	0.800	0.770
7.	A2	0.634	0.816

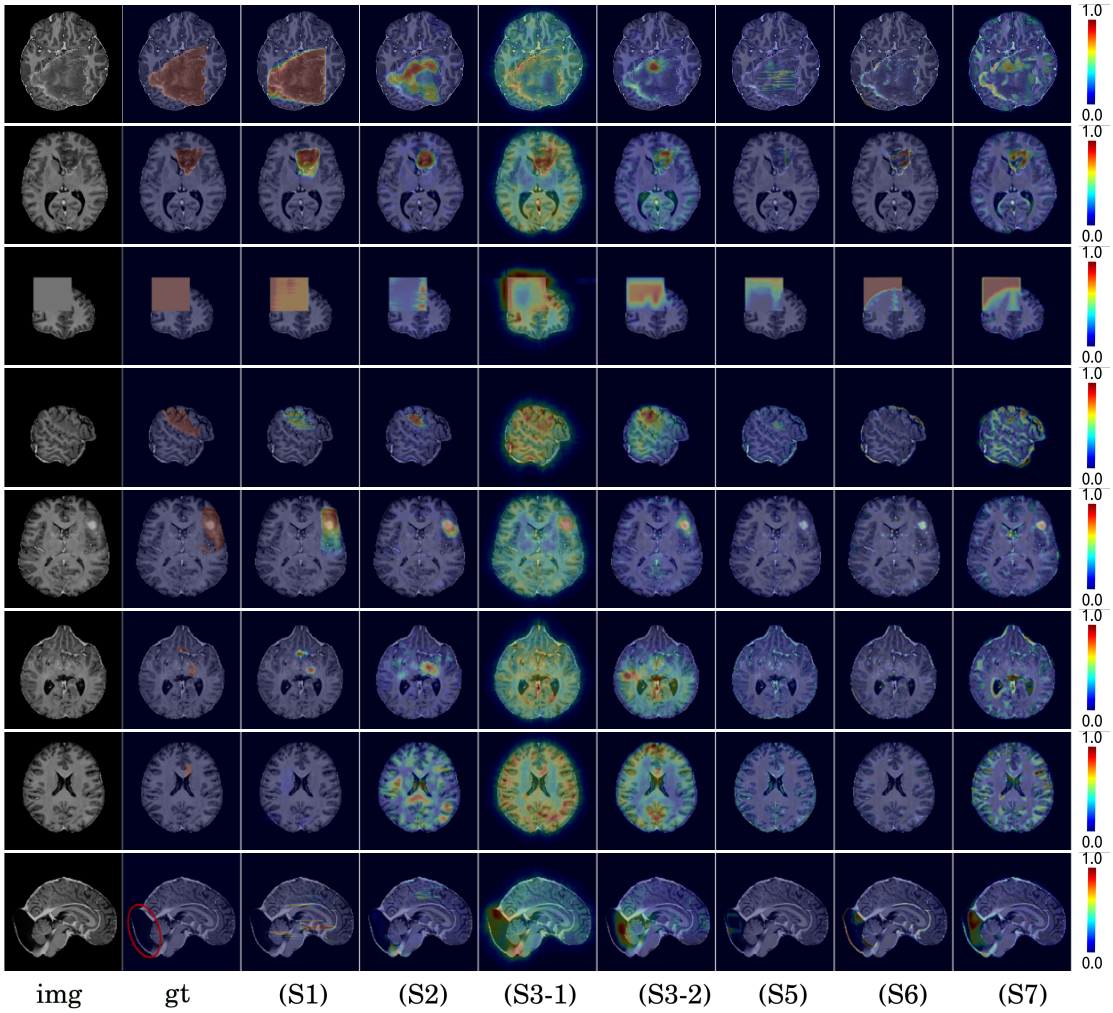


Figure 4.2: Pixel-level result heatmap visualizations for the different valid submissions for exemplary and representative brain samples (some of these were solely created for this illustration). Each row corresponds to one example. The first column shows a raw image slice, the second column the ground-truth annotation and the next columns delineate predictions by different submissions (sorted by their pixel-level challenge ranking) [Zimmerer et al., 2022a].

In Table 4.1 the final challenge ranking is presented. The first column shows the rank of the team, the second column the team name, the third column the performance on the brain dataset and the fourth column the performance on the abdominal dataset. The rank was obtained using the corresponding consensus

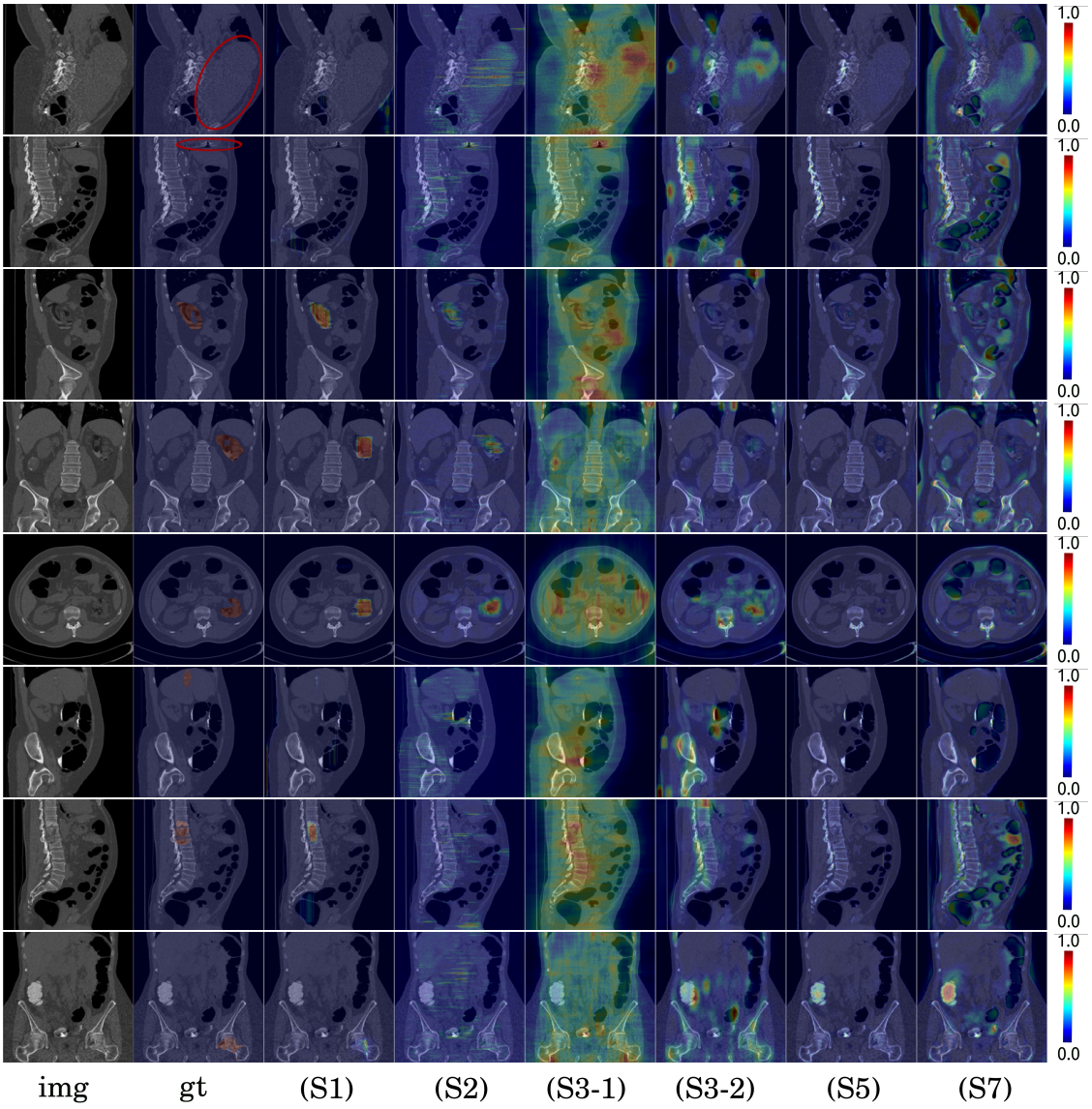


Figure 4.3: Pixel-level result heatmap visualizations for the different valid submissions for exemplary and representative abdominal samples (some of these were solely created for this illustration). Each row corresponds to one example. The first column shows a raw image slice, the second column the ground-truth annotation and the next columns delineate predictions by different submissions (sorted by their pixel-level challenge ranking) [Zimmerer et al., 2022a].

ranking between the two datasets. It is noticeable that between the top teams there are hardly any margins on the abdominal dataset, while for the brain datasets there are clear performance differences.

Pixel-level results

Table 4.2: The ranking of the pixel-level task with the performance on each dataset given as AP.

Rank	Team	Brain	Abdom.	Abbrev.
1.	FPI	0.449	0.394	*(S1)*
2.	Canon Medical Research Europe	0.416	0.288	*(S2)*
3.	Nina Tuluptceva	0.211	0.221	*(S3-1)*
3.	Sergio Naval Marimont, et al.	0.273	0.217	*(S3-2)*
5.	NUDT	0.201	0.239	*(S5)*
6.	Victor Saase	0.204	0.014	*(S6)*
7.	A1	0.160	0.072	*(S7)*
8.	A2	0.002	0.014	*(S8)*

In analogy to the sample-level results, the pixel-level results for each dataset and the following consensus ranking are given in Table 4.2.

Toy samples as predictive validation set

The previous official challenge results were obtained using datasets, each containing a large and extensive set of abnormal samples. However, next, we will explore if a simpler and smaller 'proxy' dataset can have similar predictive power. This may make a fair comparison feasible without the need for such a 'difficult to obtain' and resource-intensive dataset. In particular, we generated samples similar to the provided toy examples and compared the ranking between the results obtained on this dataset and the official results.
As the 'proxy' dataset, we generated 100 abnormal examples with the same algorithm as the toy cases by in-painting either spheres or cubes with random size and intensity into the scans (e.g. see Fig. 4.2, 3rd row). Note that the samples generated with this method highly differ from most of the anomalies in the official challenge test set. For the latter analysis, we term this 'proxy' dataset toy-ish dataset as it is similar to the provided toy cases.
Comparing the rankings, the winning algorithm could be correctly predicted by the toy-ish dataset across all tasks and datasets. Furthermore, for each dataset and

task respectively, the ranking between the toy-ish dataset and the official challenge dataset placed the same two teams in the top three. This ranking similarity can be further quantified using the Kendall tau rank distance ("Kendall's tau"). Kendall's tau is a correlation coefficient that compares correlations between rankings. We used the tau-b version of Kendall's tau which can handle ties and results in a value of 1.0 for a maximally positive correlation, -1.0 for a maximally negative correlation, and 0.0 for no correlation. The results are presented in Table 4.3.
When Kendall's tau was used to compare the ranks of the toy-ish dataset and the official challenge dataset, some association between the rankings could be found (given the limited data size). In particular, there is a strong correlation between the abdominal dataset and a weaker one for the brain dataset. One reason for this could be the overall increased difficulty of the abdominal dataset, and thus the toy cases already pose a challenging enough task. Also across both challenge tasks, the toy-ish dataset had a higher level of predictive accuracy for the (potentially harder) pixel-level task.
These results now allow for the question if a large and more complex dataset is necessary for early-stage method development or whether a simple dataset can provide sufficient information. This is also supported by anecdotal evidence, as the top-ranking teams in the challenge used a simple self-made synthetic dataset or just the provided toy cases to validate their methods.

Table 4.3: Kendall tau rank distance between the rankings on the 'proxy' toy-ish dataset and the challenge test set.

	Sample-level	**Pixel-level**
Brain	0.357	0.500
Abdominal	0.642	1.000

Analysis

Contrast & Size One research question we wanted to answer is whether the size or color contrast of the anomaly affects the detection performance. To test this, we opted to create a dataset based on toy-ish anomalies, where the size and color of the anomaly was varied. This was a compromise between a more comprehensive time- and computing-intensive analysis with more varied anomaly types and combinations and the bias introduced by the toy-ish dataset, and a simpler and more computationally efficient analysis with a single anomaly type. The results are presented in Fig. 4.4.

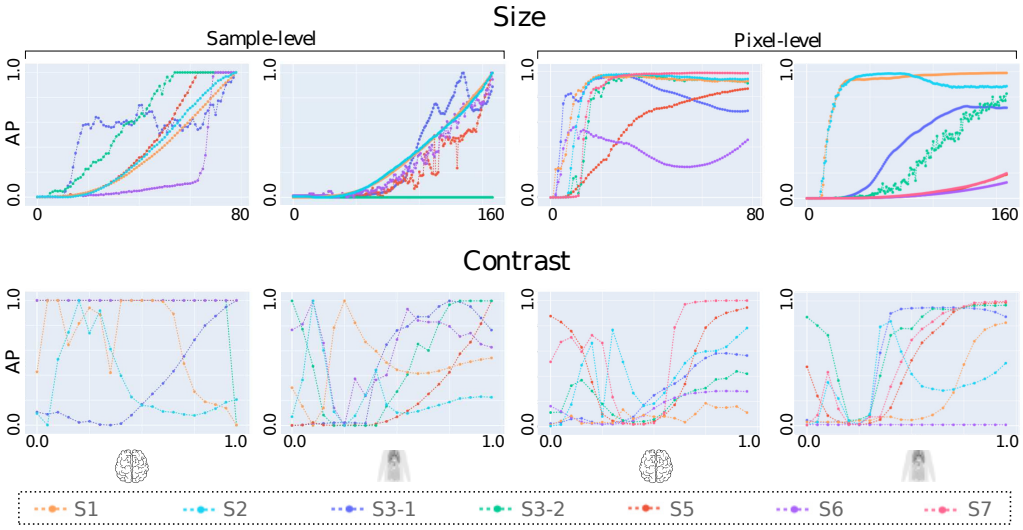
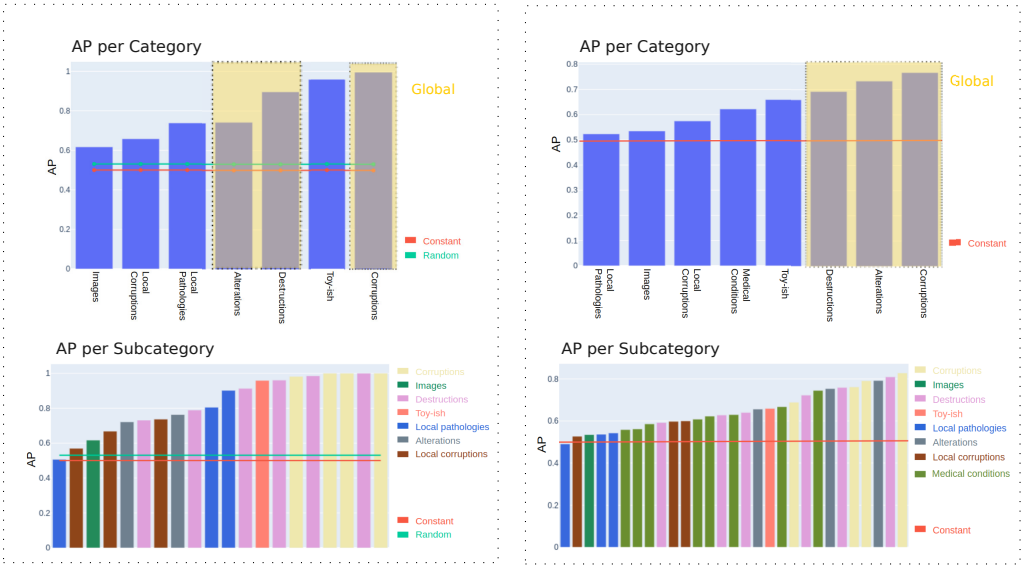


Figure 4.4: Performance (AP) of the different algorithms on anomalies of different sizes and levels of contrast. Each line corresponds to a submitted algorithm (S1-S7). The top row of graphs shows the performance for a single toy-ish example which is always in the same position but varies in size (from a radius of 0-80 pixels for the brain dataset and a radius of 0-160 pixels for the abdominal datasets). In the bottom row, the performance for a toy-ish example which is always at the same position with a varying color value (from 0.0 to 1.0 in 0.05 steps) is shown [Zimmerer et al., 2022a].

Not surprisingly, the bigger the anomaly size and the higher the contrast (differing from the mean of 0.5), the higher the performance of most submissions and particularly the top-performing algorithms show a distinct bathtub curve. Also as later noted by Meissen et al. [2021], most algorithms performed better on very bright (pixel value ≈ 1) anomalies than on to very dark (pixel value ≈ 0) anomalies. This can be perhaps attributed to the background color, which was assigned the value 0.

Anomaly classes Next, we wanted to analyze the effect of the different anomaly classes on the performance of the algorithms. For a quantitative comparison, we created a dedicated test set with an exact 50%-50% normal-abnormal data sample split, each anomaly type having the same fixed and consistent number of samples for each subcategory, which can make the metrics as comparable as possible. Exemplary (pixel-level) anomalies and submission predictions are shown in Fig. 4.2 and Fig. 4.3. In the next paragraphs, we will show a detailed



(a) Brain (b) Abdominal

Figure 4.5: Sample-level performance (AP) of the median submission per category for the different anomaly categories. The top row shows the mean of the grouped categories, and the second row gives more detailed results for the subcategories, i.e. the top row categories being split up into fine-grained subcategories. The median submission performance was used as a base for the subcategories [Zimmerer et al., 2022a].

differentiation of the anomaly classes for the sample-level and pixel-level tasks.

Sample-level An analysis of the median sample-level performances for the categories presented in Section 4.1.1 is shown in Fig. 4.5.
The results show a clear distinction between the local and global anomalies. Across all categories, the median submission performance was higher for the global anomalies than for the local anomalies. In addition, a random prediction (randomly predicting the label '0' or '1') or constant prediction (always predicting the label '0', i.e. no anomaly) is outperformed by the median sample-level performance.
To go into more detail and have a more fanned-out view of the different categories, we can look at the individual subcategories in more detail. Fig. 4.5 shows the median performance on all subcategories sorted by median performance. Similar to the top-category performance, the median performance for the subcategories on global anomalies is in most cases better than for the local anomalies and overall

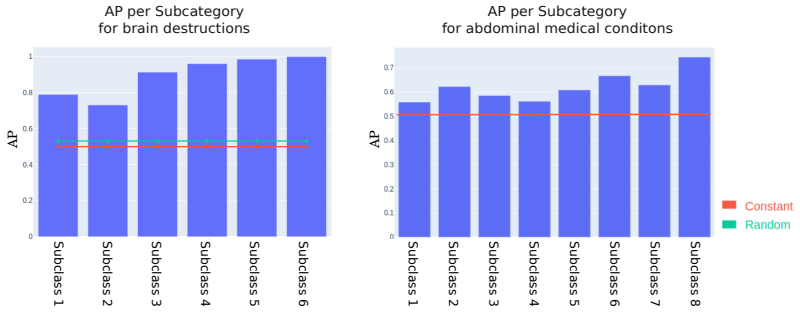


Figure 4.6: Median submission performance (AP) on subcategories of two different anomaly categories. The subcategories (classes 1-6, 1-8) are sorted by the human perceived difficulty in descending order, i.e. class 1 is the class that was perceived as the hardest and classes 6 (and 8) are the classes perceived as being easiest [Zimmerer et al., 2022a].

quite similar for the brain and abdominal dataset.
Furthermore, we wanted to see if the humanly judged subjective difficulty of the different categories would correlate with the performance of the algorithms. Hence, we rated the anomaly subclasses by human-perceived difficulty and show the median submission performance for the different categories, sorted by the human-perceived difficulty in ascending order (see Fig. 4.6). While not strictly increasing, a clear trend can be observed.
The results seem promising and allow for the question if the submitted approaches are ready for translation to a clinical setting and could deliver added value.

FPR@0.95TPR As one way to evaluate the clinical applicability of the proposed approaches, we chose the False Positive Rate at 95% True Positive Rate (FPR@0.95TPR) metric, which shows a false-positive rate at 95% true-positive rate.
Given the 50%-50% split of the normal-abnormal data, a score of 0 would mean that an algorithm could detect 95% of the anomalies without diagnosing a single normal sample as abnormal, thus allowing physicians to accelerate their diagnostic processes greatly. A prefiltering with an approach with a score of 0.5 could still result in every second 'anomalously labeled' image being normal, thus giving a rough acceleration of just $\frac{1}{4}$. A score of 1.0 would require the physician to inspect every sample regardless, providing no acceleration.
Here, the choice of an FPR@0.95TPR is arbitrary and possibly higher TPRs might be required in a clinical setting. However, despite the exact choice of the TPR, the metric for different values is strongly related and this exact choice was often used

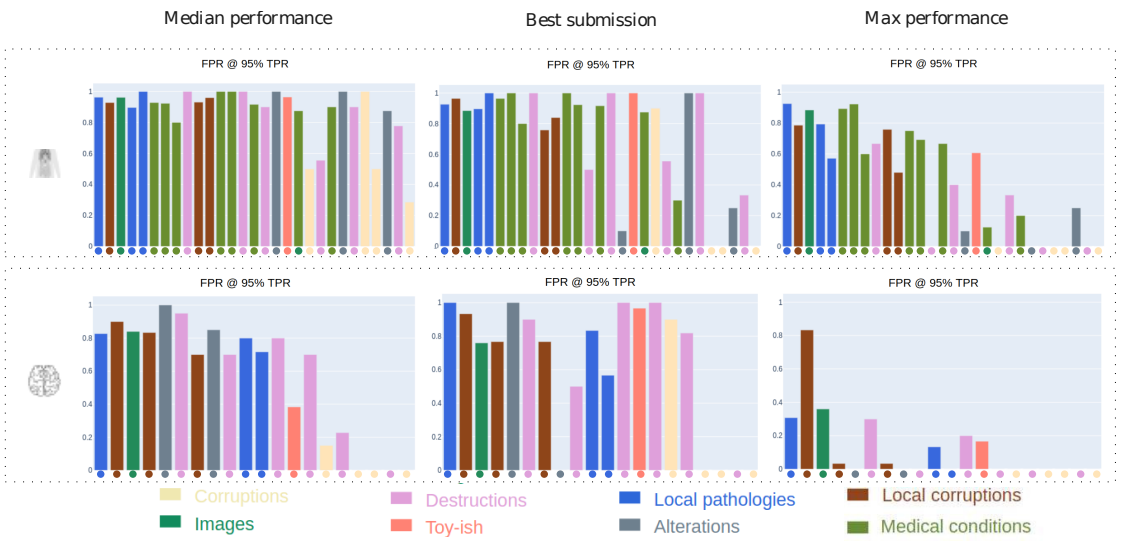


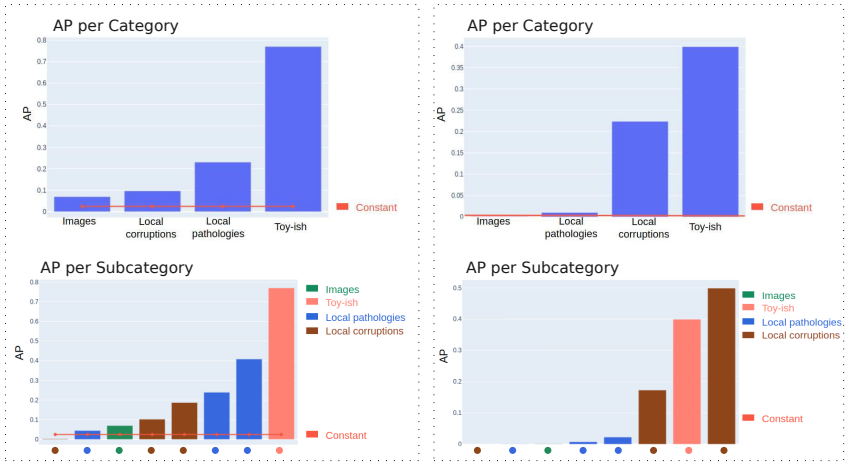
Figure 4.7: FPR@0.95TPR for the different anomaly subclasses of the abdominal dataset (top) and the brain dataset (bottom). The median submission performance, the performance of the best sample-level submission and the maximal performance of algorithms (i.e. picking the best algorithm for each subclass) are shown [Zimmerer et al., 2022a].

in other OoD work [Choi et al., 2018; Hendrycks et al., 2019] and discussions with physicians have indicated this to be of interest.

Similar to and in the same order as the results in Fig. 4.5, the sample-level FPR@0.95TPR scores are shown in Fig. 4.7. Here, the median performance, the individual top subcategory performance (the best submission for each subcategory), and the overall best performing algorithm (i.e a realistic best-case performance estimate) are compared.

The relative performance of the FPR@0.95TPR is similar to the AP metric. The best results are for classes with global destruction or corruption and in the best case, the top model can find 95% of the anomalies without inspecting a single normal image. But this only presents an exemption from most other categories, and especially from the categories with higher clinical relevance such as the local anomaly categories and medical category. Here, the amount of cases that would have to undergo inspection to find 95% of anomalies could not even be reduced by half.

Next, we will present the results for the pixel-level delineation of anomalies. However, we did not perform an FPR@0.95TPR analysis on a pixel- or object-level, because this would require binary objects and thus a binarization and connected-



(a) Brain (b) Abdominal

Figure 4.8: Pixel-level performance (AP) for the different anomaly categories. The top row shows the mean of the grouped categories, and the second row gives more detailed results of subcategories, i.e. the top row categories being split up into fine-grained subcategories. Median submission performance was used as the basis for the subcategories [Zimmerer et al., 2022a].

component analysis. This could, due to the choice of a binarization threshold, introduce some bias. Since this has not been extensively done or tested in prior work, we opted for conventional metrics only.

Pixel-level Qualitative results of the pixel-level predictions can be seen in Fig. 4.2 and Fig. 4.3 and quantitative results of the median submission for pixel-level anomaly categories in Fig. 4.8. Here, the difference between the absolute values for the brain dataset and the abdominal dataset is very prominent. Exemplary, the median performance on the toy-ish brain dataset has an AP of 0.8, while for the abdominal dataset, the AP is roughly half, namely around 0.4.
A more fanned-out analysis for the subcategories is presented in the second row. For most subcategories of the abdominal dataset, there is hardly any difference observable between a constant guess, showing great room for improvements for such cases.

4.1.6 Discussion & conclusion

With the challenge, we wanted to benchmark different anomaly detection and localization approaches and determine if and how they could be used in a clinical

setting (however with the primary objective being on a controlled yet realistic setting, and not 100% driven by a clinical dataset). Overall the top submissions were reliably able to detect certain types of anomalies (mostly image corruption), but not all. On most anomaly categories, especially the local anomalies, most algorithms performed rather poorly and were far from achieving clinically acceptable performance. In general, the clinical relevance and transferability are discussable: the relevance of the easy-to-detect anomalies, such as anomalies mimicking imaging failures and large image artifacts, in current clinical practice is unclear as they can easily be detected by a trained physician and not lead to much ease or speedup. Furthermore, the algorithms all exhibited a rather high inter-case and inter-participant variability, which is also evidently in Fig. 4.2 & Fig. 4.3. As reliability and as such a certain degree of guaranteed performance is important for trust and added benefit in a computer-assisted diagnostics tool, this opens up an area in which the current algorithms need to be improved upon. But some algorithms showed also very promising results on certain harder-to-detect local anomalies. Thus, while it's currently hard to recommend any specific algorithm for general anomaly detection in practice, we believe that the different contrasts, sizes, and types of anomalies show the potential of the different algorithms and point to areas for further research.

One interesting, but not unexpected, finding is the performance difference between the abdominal and brain dataset. For this challenge, we on purpose chose two different datasets: first, the brain dataset, which is quite homogeneous, was recorded on the same scanners with the same parameters, has a narrow selection of participants (young healthy adults) and has a very low anatomical variance. Furthermore, most anomaly detection algorithms in the medical field were previously evaluated on brain datasets as well. The abdominal dataset, on the other hand, is 4x the size of the brain dataset, has a much larger selection of participants (including elderly people who had varying natural anatomical conditions as well as natural and unnatural pathological conditions), inherently more anatomical variation (encompassing multiple deformable organs and structures), and was recorded on multiple different sites. However, to test the generalizability of the approaches, we wanted to contrast the brain dataset with a dataset that has not been used in the anomaly detection field yet. So while it's not unexpected that algorithms perform better on the brain dataset, the following reasons might explicitly explain some of the differences: (a) Most previous medical anomaly detection algorithms were designed with brain datasets in mind [Baur et al., 2018; Chen et al., 2018; Schlegl et al., 2017; Zimmerer et al., 2019a], and as such there might be a bias in the developed algorithms towards brain datasets. (b) The data sample size and variation in the abdominal dataset are much larger than in the brain

dataset. However, the number of data samples is rather consistent (as predetermined by the original number of data samples for the respective studies [Johnson et al., 2008; Van Essen et al., 2012]). This increased dataset complexity for the abdominal dataset might also require an increased number of data samples to achieve comparable performance. So we believe that these points can explain the difference in performance between the datasets to a large part, and developing more data-efficient and less brain-specific algorithms can help to improve the generalizability of the algorithms to new datasets.

Another interesting but also not an unexpected finding is the difference in performance between global and local anomalies. Here, we want to introduce some related terms and concepts can be introduced: semantic vs. non-semantic anomalies as described by Ahmed and Courville [2019] or far OoD vs. near OoD as described by Winkens et al. [2020]. Semantic vs. non-semantic anomalies describe the difference if anomalies vary only contextually/semantically but originate from the same domain vs. anomalies originating from a different domain. Near vs. far OoD describes the concept similarly, i.e. near anomalies stem from the same or very similar domain, while far OoD anomalies originate from a different and/or far away domain. As such the local anomalies could be categorized as semantic near-OoD anomalies, since introducing an object into a part of the scan with the same image statistics will not largely alter the image and can only be differentiated given the context (i.e. scans of healthy people vs. scans of not healthy people). The global anomalies on the other hand could be categorized as non-semantic far-OoD (or potentially also near-OoD) anomalies. As claimed in Ahmed and Courville [2019]; Winkens et al. [2020] the non-semantic far-OoD (i.e. global anomaly) cases might be less interesting from a methodological point of view since quite simple (statistics-based) methods could already be viable solutions. The semantic near-OoD (i.e. local anomaly) cases, which possibly are also more clinically relevant, on the other hand, were claimed to be of more interest as they require more complex methods to be solved. However, exactly for these kinds of anomalies, the performance of the algorithms is not as good as for the global anomalies and a performance gap between the global and local anomalies is observable. Furthermore, the subjectively harder the anomalies were to detect, the worse the performance of the algorithms. So while the performance on the global anomalies might be sufficient, the most interesting cases from a methodological perspective as well as clinical perspective, the local semantic anomalies, can be a good starting point for further research.

One point noted in [Meissen et al., 2021], which can also be found in the analysis presented here, is the dependence between localization performance and color intensity. While Meissen et al. [2021, 2022] claim that due to an inherent

dataset/anomaly property many anomaly detection methods fall back to simple threshold-based intensities detectors, especially the top-performing methods here don't. While for example FPI also shows weaker performance on dark anomalies and anomalies with less contrast, in their method approach and dataset there is nothing that would favor bright anomalies. Thus we believe that for this case a controlled setting, as in this challenge, is important to further investigate this phenomenon.

One trend in general machine learning and computer vision research that is also reflected in the submissions is the rise of self-supervised methods [Chen et al., 2020; Li et al., 2021]. Especially three of the top four teams utilized self-supervised methods in some way, e.g. as pretraining to initialize a perceptual model or as a proxy task during algorithm training. However, perhaps one factor for the success of the self-supervised methods is possibly the similarity to the target task. One might argue that the performance gains of the self-supervised methods are caused by the synthetic anomalies in the test set which might coincidentally resemble some of the self-supervised tasks. Nevertheless, the self-supervised submissions also show good performance on the naturally occurring anomalies (compared to the other approaches). Furthermore, follow-up studies on these approaches have shown that the self-supervised methods also translate their performance to other medical datasets [Kascenas et al., 2021; Tan et al., 2020; Tuluptceva et al., 2020] and similar approaches have been proposed for other anomaly detection tasks [Li et al., 2021].

One of the other 'big' trends in anomaly detection, AE-based methods [Baur et al., 2021; Chen et al., 2018; Zimmerer et al., 2019a], are also reflected in three of the top four teams (two teams using both, AE-based methods and self-supervised learning). Here, follow-up and consecutive papers have also extended the methods and shown good performance when applied to other medical datasets [Marimont and Tarroni, 2020; Pinaya et al., 2021].

Another property that the submissions have in common with recently proposed anomaly detection methods is the 2D processing of the samples. Here, while the samples are available as 3D volumes, all submissions processed the 3D volumes in 2D slices. In contrast, for segmentation tasks, most recent approaches have opted to trade off additional compute and time constraints (which might be the limiting factors) for a better segmentation performance of the 3D volumes [Isensee et al., 2018]. However, for anomaly detection algorithms, the current state of method development is not yet as saturated as for segmentation tasks and thus the teams opted for 'faster feedback loops' using faster and more compute-efficient 2D processing. This slice-wise processing can lead to some processing artifacts (see Fig. 4.3). Here, the use of 3D methods or integrating global context and position

information might give a further performance boost and outline a further area of research.
One final surprising point, which might also has the potential to influence the further development of anomaly detection algorithms, is the correlation between the performance on the 'big' challenge test set and a 'small' and simple toy dataset. Tuning an algorithm solely on such a toy dataset might not directly generalize to other, more general anomaly detection settings. However, in this case, the toy dataset proved to be challenging enough for all submissions and no submission was able to perfectly detect all these toy anomalies, especially when the contrast and size were varied. Thus the performance on a toy dataset can be seen as the upper limit for the performance on the 'big' challenge test set and as long as there are settings in which the algorithms struggle on the toy dataset, they probably will also struggle on a more challenging dataset. Furthermore, as shown by the correlation between the performance on the 'big' challenge test set and the performance on the 'small' toy dataset, such a dataset could be used as a simple benchmark to compare different algorithms. Hence, creating and using a simple toy dataset can present a useful tool for the development and benchmarking of new algorithms (as anecdotally most of the top teams did).

4.2 Validation in the real world: a reality check

While the previous sections demonstrated the advancement of anomaly detection techniques in a research context and the benchmarking of approaches in a controlled but not entirely clinical scenario, we will now look at the application in a real-world scenario (note: some definitions, formulations, and equations of this section were previously published by myself in [Zimmerer et al., 2022b]).

4.2.1 Motivation: how well does a limited research setting translate to a real-world example

The majority of past performance assessments were completed in research settings, i.e., where training data could be ensured to stem from healthy subjects and where test sets homogeneously consisted of one type of pathology. While this can be useful in the development of novel approaches, it has certain limits in terms of validity and applicability to real-world scenarios: *(1) Definition of normality*: the performance depends on the definition of normality and the selection of the training and test set. In a truly unsupervised setting, a normal scan refers to a sample without known medical conditions, for example as given by a population-based representative sample. In general, the definition of normal and abnormal is frequently not 'black and white' and is dependent on the dataset's assessment protocol. This can result in potential ambiguities and thus not ensure that none of the existing abnormalities are overlooked during curation. *(2) Anomaly variety*: Current methods are frequently explored on the same and extremely limited collection of abnormalities, namely brain tumors and lesions [Baur et al., 2020; Chen and Konukoglu, 2018; Pawlowski et al., 2018; Uzunova et al., 2019; You et al., 2019; Zimmerer et al., 2019b]. While this is most likely due to the frequency of such conditions and the availability of corresponding annotated datasets, it raises two questions for unsupervised anomaly detection: first, whether there is a need for unsupervised techniques for those explicit pathologies, and second, whether the developed techniques are biased towards these anomalies. This is especially true when generalization to other anomalies is not demonstrated, despite the presence of large and high-quality datasets. The ability of a method to identify any, even uncommon or undiscovered conditions or biomarkers, represents a medically important task. *(3) Dataset mixture*: often multiple different datasets with different populations are mixed to create a dataset of normal and abnormal samples. This often leads to datasets containing a fixed but not realistic ratio of normal and abnormal samples. Furthermore, in some cases, these datasets are derived from multiple sources or populations, which can further complicate matters because

this necessarily includes domain shifts, which can lead to bias in the results.
Despite the limitations described before, a controlled research setting may be extremely beneficial for method development and as a reference baseline. However, we are interested in how current anomaly detection systems might generalize in a more realistic situation, in particular the detection of Incidental Findings (IFs) in a Large Scale Population Study (LSPS). In contrast to previous works, the 10000+ participants of the LSPS that was considered in this work are sampled representatively and contain a variety of different anomalies. Radiologists have previously examined each image. This allows for an evaluation of the generalization of anomaly detection methods on a real-world example and tackles the previously mentioned short-comings: (1) *Definition of normality*: it allows for a well-defined definition of normality. Here, as in the general population, we have a multitude of different conditions with a spectrum of severity/ progress. In this setting, the labeling protocol defined anomalies as IFs which would require direct medical intervention. (2) *Anomaly variety*: due to the nature of the dataset containing a cross-section of the population, multiple different, more and less frequent anomalies are present in the data. (3) *Dataset mixture*: here, the data was collected in 5 different centers. But in contrast to the previously mentioned issue of mixing different datasets, here the same acquisition protocol was used for each site and the distribution of participants for each site was the same. Consequently, this does not introduce and distribution shift regarding the data source or population and contains a representative and realistic normal:abnormal data ratio. All-in-all, this allows for the use of multi-centric data without bias and can give further insights into how well the approaches generalize to such a multi-center dataset.

4.2.2 Experiment setup

Datasets

The data used for this work stems from a prospective epidemiological study resource for health and disease research in Germany, i.e., an LSPS. For this study 30000+ participants between the ages of 20-69, representing the population of almost all federal states and covering metropolitan, urban, and rural regions, were randomly selected from the general population from local municipal population registries within defined strata of age and sex, and were offered whole-body MRI scans. The study data includes 30000+ T1-weighted (T1) and FLAIR brain scans and reported reference Incidental Findings (refIFs). The refIFs only contain medical conditions which need an immediate medical follow-up, and as such constitute a definition of what we define as abnormal (this also facilitates the often

very ambiguous task to determine if an anatomical variation, in general, can be seen as normal or abnormal and where to draw the line, see Section 6.0.4). The data was recorded at 5 different sites. For research purposes, in the first release, ≈10000 scans and refIFs were made available. In this work, we focused on brain data, as is mostly done in unsupervised anomaly detection on medical images [Baur et al., 2020; Chen and Konukoglu, 2018; Pawlowski et al., 2018; Uzunova et al., 2019; You et al., 2019; Zimmerer et al., 2019b]. The brain was extracted using BET [Isensee et al., 2019] and the T1 and FLAIR scans were co-registered using FSL before being z-score normalized. The refIFs were available as a single point and we manually added a radius. All findings that were not in the brain region were excluded. This resulted in 84 refIFs. The annotations were binary, i.e. refIF or no refIF. For a more detailed analysis, we further manually chose and classified the refIFs in 3 gradings: *C1*: clearly noticeable ("should definitely be found"), *C2*: Partially noticeable ("could be found"), *C3*: unobtrusive ("without context and/or expert knowledge hard to detect").

To generate a near-real-world scenario, we randomly selected 5000 test samples from the preprocessed images (having no labeled anomaly/refIFs), which comprise the test set together with all of the refIFs samples. We selected 4800 samples for training and 200 for validation from the remaining images, which were all 'regular' scans with no annotated refIFs. As it was shown in Section 4.1.5 that simple toy anomalies already give a good indication of performance, to help with the issue of model and threshold selection, we constructed a validation set with 100 samples that have relatively basic 'toy-sphere" abnormalities (see Fig. 6.5). The study was approved by the IRB committee and all participants gave informed consent.

Anomaly detection

We picked three anomaly detection techniques to evaluate and compare using the dataset. First, we chose a VAE with iterative image restoration [You et al., 2019] which has shown good performance and was recommended for medical anomaly localization tasks by Baur et al. [2020]. Second, we chose the winning approach of the MOOD 2020 Challenge, the self-supervised approach termed FPI [Tan et al., 2020]. Last, we chose the ceVAE, a combination between VAEs and self-supervision that has also shown good performance recently in independent studies (especially with regards to AUROC) [Baur et al., 2020; Bengs et al., 2021; Bercea et al., 2022]. Next, we will describe the methods in more detail.

VAE The goal of a VAE [Kingma and Welling, 2013; Rezende et al., 2014] (see Section 3.1.3) is to optimize a lower bound (known as ELBO) on the log-likelihood of a data sample:

$$\mathcal{L}_{VAE} = D_{KL}(q(z|x)||p(z)) - \mathop{\mathbb{E}}_{z\sim q(z|x)}[\log p(x|z)] \leqslant \log p(x). \tag{4.2}$$

In practice, $q(z|x)$ is chosen as a Normal distribution that is parameterized by an CNN encoder $f(x) = \mu_z, \sigma_z$, $p(z)$, $p(z)$ is chosen as an isotropic Gaussian with zero mean and variance 1, the expectation is estimated using MC sampling with a sample size of 1, and $p(x|z)$ is chosen as a Gaussian with fixed variance and mean given by a CNN decoder $g : p(x|z) = \mathcal{N}(x|g(z), c)$. As a consequence, the VAE-loss $\mathcal{L}_{VAE}$ can be calculated analytically and optimized with the training set.
The anomaly score proposed by You et al. [2019] is calculated via iterative image restoration: the iterative image restoration Variational Autoencoder (irVAE), and has shown good performance in [Baur et al., 2020]. The goal of the iterative restoration is to optimize the following objective:

$$\arg\max_x \log P(x|y) = \arg\max_y \left[\log P(y|x) + \log P(x)\right], \tag{4.3}$$

where the data-likelihood $\log P(x)$ is approximated by the VAE loss and the data consistency term $\log P(y|x)$ is chosen as total variation Norm $\|x-y\|_{TV}$. This leads to the final optimization objective

$$\hat{x} = \arg\max_x \left[-\lambda\|x-y\|_{TV} + \mathcal{L}_{VAE}(x)\right],\ \lambda > 0, \tag{4.4}$$

which can be optimized using gradient descent. Finally, the difference between the 'more likely' restored image $\hat{x}$ and the original image x gives a pixel-level anomaly score S:

$$S_{pixel} = \|\hat{x} - x\|. \tag{4.5}$$

In contrast to You et al. [2019] we did not use a mixture model, but rather $N(0, 1)$ as VAE latent-space prior (a prior conversation with the authors revealed that the difference between the priors was minor in their experiments).

ceVAE A ceVAE (see Section 3.2.3) augments VAEs with self-supervised learning, in particular, CE and in-painting. For the ceVAE in addition to the VAE-loss $\mathcal{L}_{VAE}$ a reconstruction loss between the reconstructions of perturbed input samples $\tilde{x}$ and the unperturbed samples x is added:

$$\mathcal{L}_{ceVAE} = \mathcal{L}_{VAE} + \lambda\mathcal{L}_{ce},\ \lambda > 0, \tag{4.6}$$

with

$$\mathcal{L}_{ce}(x, \tilde{x}) = L_{rec}(x, g(f(\tilde{x}))). \tag{4.7}$$

Here the samples were perturbed with context-encoding, i.e., one to three random rectangular regions of the image were masked out. We employ the gradients of D_{KL} with respect to the input image x as anomaly score as proposed in Section 3.1.3 since they are more unaffected by pixel-value magnitudes than a reconstruction difference:

$$S_{pixel} = \frac{\partial(-D_{KL}(q(z|x)||p(z))}{\partial x}, \tag{4.8}$$

FPI FPI [Tan et al., 2020] perturbs an area of the input slightly and uses a segmentation net to detect this perturbation. The idea behind this is, that to detect these small perturbations, the network has to learn what a normal sample looks like and is consequently also able to flag anomalies and pathological conditions which are not present in the training set as perturbations. To create these perturbations for training, FPI selects a rectangular area termed 'patch' and in this patch area interpolates the original image with a similar 'foreign' image. In this case, slices at the same position from different scans are used for the interpolation. The size, location, and interpolation factor of the patch are chosen randomly. The segmentation target is to predict the interpolation factor for each pixel (0 if the pixel is not interpolated, otherwise the interpolation factor). The segmentation network, often an encoder-decoder model such as a U-Net [Ronneberger et al., 2015], is then trained to segment the interpolated patches. To detect anomalies for each pixel, the predicted interpolation factor is used as an anomaly score during inference. Since the FPI model was only proposed for single-channel input, we trained one model for each imaging sequence (T1 and FLAIR) and used the averaged prediction from the two models (on the validation set we could see a clear increase in performance for the averaged prediction compared to the single models, and thus chose to use the averaged prediction for the test set prediction).

Post-processing We subsequently post-processed the predictions with a Gaussian Filter of size 3 and cropped the predictions to just the brain area, similar to Baur et al. [2020].

Metrics

The earlier deep-learning-based medical image unsupervised anomaly detection papers often used the AUROC as the primary metric [Baur et al., 2018; Chen and Konukoglu, 2018; Pawlowski et al., 2018; Zimmerer et al., 2019b]. Lately,

papers have instead recommended focusing on the AP for sample-level anomaly detection [Ahmed and Courville, 2019], and AP has also found wide adoption in anomaly detection (or rather localization) on medical images [Baur et al., 2020; Zimmerer et al., 2020]. AP is often preferred because it can handle labeling balances differently than AUROC since it puts more weight on less frequent classes [Ahmed and Courville, 2019; Baur et al., 2020].

An additional metric used in medical image anomaly detection is the DSC which is one of the primary metrics in medical image segmentation. After choosing a binarization threshold and binarizing the anomaly scores into binary segmentation, the DSC can be calculated [Baur et al., 2020; Chen and Konukoglu, 2018; Zimmerer et al., 2019b].

Consequently, we also adapted AP, AUROC, and DSC for this work. To determine a binarization threshold, which we term *detection-threshold*, for binarizing the anomaly scores and consequently calculating the DSC on the test set, we use the "toy-spheres" validation set. For each method, we determine the binarization threshold which would achieve the best DSC on the "toy-spheres" validation set and use it for binarization of the test set predictions. Other approaches to determining such a threshold are choosing the n-th percentile of the anomaly scores on the training data [You et al., 2019], assuming that there are already $n\%$ abnormal pixels in the training set. If a "toy" validation set is available, we believe that this however allows for a better estimate of such a threshold (and not guessing n arbitrarily, since for the optimal case n should be 0).

However, we believe that there are two issues with the commonly used metrics:

(1) *Dataset-level vs. sample-level/per-scan metrics*: In medical image segmentation, the DSC is often calculated on a sample-level/per-scan level and then aggregated via mean or median for the whole dataset. In contrast to medical image segmentation, in anomaly detection, not all samples contain abnormal pixels, and as such for these samples with all pixels having the same class, DSC, AP, and AUROC are not defined. One way to circumvent this is to group all pixels in the dataset into one bucket and calculate the metrics on a dataset level directly (without any division into scans). This however needs quadratically more computationally resources (for AP and AUROC since for these metrics the scores have to be sorted) and also can underestimate the importance of small objects, which is described in the second issue.

(2) *Indifference towards small objects*: The DSC and AP primarily consider TPs, FPs, and FNs, but do not have a concept of objects. As such if in one image/case there are multiple objects and one object is much bigger than the other object it put less importance on the small objects. In extreme cases, a better DSC can be achieved by better segmenting an already detected and reasonably well-segmented big object

than segmenting and thus detecting a small object at all (since the improvement for the big object results in more TPs and causes potentially less FPs for the harder smaller object) [Reinke et al., 2021]. This property can get exaggerated when the metrics are directly calculated on a dataset level and not calculated on a per-scan level and then aggregated.
To solve these issues we propose to also include object detection-based metrics in the analysis. In particular, to go from segmentation to an object level, we used the following method: first, we used the *detection-threshold* to binarize the anomaly score predictions. Then, using connected component analysis, the segmentations were divided into different objects. To reduce noise, we removed objects with less than $1200 mm^3$ ($\approx$ half the size of the smallest annotated refIF). We then presented the detected objects to an expert radiologist ($\sim$ 10 years of experience in neuroimaging) to determine TP, FP, FN. We also propose an automatic evaluation into TP, FP, and FN by checking for each object whether the center of mass of the segmentation falls into a ground truth segmentation, and the size difference is only by a factor of 2.
We believe that these object-based metrics can give more information regarding clinical relevance since for medical image anomaly detection it is often not important if the object is perfectly segmented, but rather more important if all abnormal objects/anomalies are detected, so that a physician can be notified for further assessment.

Model and parameter selection

For unsupervised anomaly detection, to have a fair and unbiased evaluation of a test set it is important not to leak too much information about the test set beforehand. Thus, we only once per approach predicted the test set and directly report the results here. For example, in our opinion, running the same model twice with different hyperparameters and claiming the performance of the model to be the max performance of the two runs already introduces some bias from the test set and might not generalize to another test set (we believe this is similar to the "reproducibility crisis" in reinforcement learning [Zhang et al., 2018]). This, however, results in the problem of model and hyperparameter selection. We were inspired by Shafaei et al. [2019]'s approach as well as the strategy adopted by the majority of MOOD Challenge contestants (where the participants had no access to the test set as well) [YouTube, 2020, 2021]: here, the approaches were validated on a different (in-house) dataset beforehand.
To select a model and hyperparameters we used the previously described validation set with artificial "toy-sphere" anomalies. This may generate a bias for models

Table 4.4: Performance comparison of the three models on the 100 toy-sphere validation samples and the 84 refIF samples from the LSPS (dataset-level).

		irVAE	ceVAE	FPI
Validation	AP	0.065	0.253	0.376
	AUROC	0.904	0.954	0.956
	DSC	0.065	0.324	0.434
refIFs	AP	0.017	0.312	0.016
	AUROC	0.787	0.838	0.815
	DSC	0.035	0.382	0.026

towards anomalies that resemble "toy-spheres", but, as previously demonstrated, it may also be a reasonable proxy task for more general anomaly detection performance. Thus, after each epoch, we calculated the AP on the validation set for all approaches and then, respectively, selected the best model for the final evaluation. However, due to the added computation cost of the iterative restoration, for irVAE it was unfeasible to run the iterative restoration procedure after each epoch and we selected the best model based on the best AP using reconstruction error scoring.

4.2.3 Results

Comparison of methods

First, we compare the irVAE, the ceVAE, and FPI on the "toy-spheres" validation set and the 84 samples with the refIFs (see Table 4.4). One thing that becomes apparent is that while all seem to work reasonably well on the validation set the irVAE and FPI results seem not to generalize to LSPS data (we believe that for the irVAE the low performance in terms of the AP score, while at the same time having a reasonably well AUROC score, can be attributed to the model being selected by the reconstruction error AP).

Dataset-level vs. object-level metrics

Since the irVAE and FPI did not show well enough results, we focused the following analyses on the ceVAE only. The ceVAE has a (dataset-level) AUROC of 0.838 and an AP of 0.312, which seem to in the same range as the reported results on the BraTS17 dataset [Zimmerer et al., 2019b] (please note that the results are not entirely comparable because for the LSPS we approximated the refIF segmentation with a sphere). For the refIF scans only, there is already a clear contrast between the

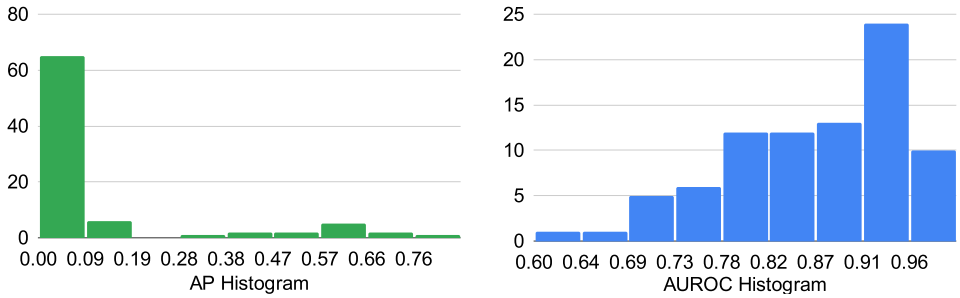


Figure 4.9: Histogram of sample-level/per-scan AP scores and histogram of sample-level/per-scan AUROC scores for the ceVAE [Zimmerer et al., 2022b].

dataset DSC of 0.382 and an aggregated sample-level/per-scan DSC of 0.061 and the dataset AP of 0.312 and an aggregated sample-level/per-scan AP of 0.108 when contrasting the dataset level metrics with the aggregated sample-level metrics. Some differences between the dataset-level metrics and the aggregated sample-level metrics can be explained by the AP histogram (Fig. 4.9): for the AP histogram two modes/peaks can be detected, one at 0.0 and one at 0.6, i.e. while it seems to work quite well for a few scans, for most scans, the algorithm fails. However, we discovered that the instances on which the ceVAE works well are those with the most labeled pixels, i.e., scans with big and easily detectable abnormalities. As a result, big and easily detectable anomalies dominate the dataset AP score, overshadowing the performance of perhaps more medically relevant, harder to identify, and smaller abnormalities. The AUROC appears to be more resilient to dataset and per-scan changes, but it does not appear to reflect the fact that the model works well just for simple anomalies. As a result, in order to get a more direct performance metric of anomalous object detection (which is arguably more relevant than pixel-level delineation), a radiologist manually assessed the algorithm's predictions.

Evaluation with an expert radiologist

While the previous analysis was only performed on the refIF samples of the LSPS dataset due to the high computational cost of calculating the AP and AUROC over the whole dataset and AP, AUROC, and DSC not being defined for samples without anomalies, this section will use object-based metrics to evaluate the performance of the ceVAE model on the whole dataset.
During the first inspection of the results on the test set, we noted that the algorithm detected anomalous scans, which were not and probably would not be classified as

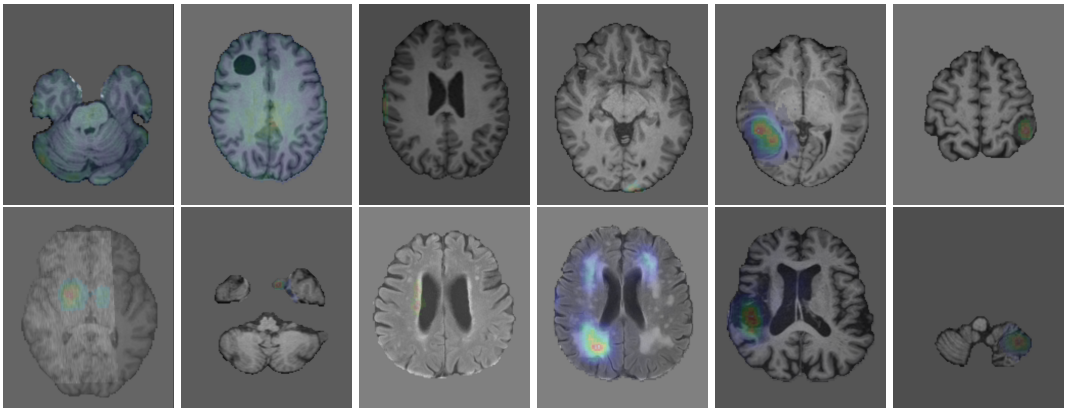


Figure 4.10: Qualitative anomaly detection results of the top performing algorithm (ceVAE). First row (from left-to-right): Images 1&2 show not detected refIFs, images 3&4 false positives, and images 5&6 detected refIFs. Second row (from left-to-right): Image 1 shows a detected imaging artifact, image 2 shows a detected brain extraction failure, image 3 shows a detected not medically relevant anomaly, image 4&5 detected medical anomalies which would not classify as IFs, and image 6 a detected anomaly which is not in the refIFs and possibly could be classified as IF [Zimmerer et al., 2022b].

refIFs, but nevertheless had an anomalous appearance. To further investigate this and check if such scans were also present in the training set and if the algorithm would also detect these in the training set (which was presented as 'normal' during the training phase), we extended this analysis to the whole dataset (training + validation + test set).

Overall, the algorithm detected 84 suspicious regions/objects in the whole dataset. We compared the detectioned objects to the refIFs, and in total 19 of 84 refIFs were found, but almost all (16/19) were C1 (see Table 4.5). However, a large amount (21/37 ,$>$ 50%) of C1 findings were not still detected.

Furthermore, the predictions which exceeded the *detection-threshold* together with the corresponding scans, were presented to an expert radiologist ($\sim$ 10 years of experience in neuroimaging). The expert was asked to classify those that could not be matched to a refIF into one of the five classes: *False Positive*, *Brain Extraction (BET) Failure*, *Imaging Artifact*, *Not Medically Relevant Anomaly*, *Medical Condition* (see Table 4.5). We further sub-divided the Medical Condition class into *IFs* (i.e., needing medical interventions) and *Other Anomaly*.

Using the as *IFs* categorized detections, we also explored the possibility that any detected findings may be unreported or missed refIFs. Here, an additional review

Table 4.5: Detections by the ceVAE model classified into categories and compared with the reference data.

	#Detected	#Reference
refIF C1	16	37
refIF C2	3	39
refIF C3	0	8
FP	28	-
BET failure	10	-
Imaging artifact	12	-
Not med. relevant	3	-
Medical condition	12	-

and assessment by the expert radiologist revealed that the algorithm was able to detect 2 findings that could possibly be considered refIF (however, those were borderline cases and also arguments for not classifying them as refIFs could be made).
Of all detected 'Medically Relevant Anomalies', 3/12 were in the test set (excluding the refIF scans), 2/12 in the validation set, and 7/12 in the training set. This indicates some robustness of the ceVAE model to outliers during training. Some examples can be seen in Fig. 4.10.

4.2.4 Interpretation of results for clinical application / clinical value

The findings challenge the direct application of anomaly detection methods in medical practice. However, the system was still capable of detecting 2 suspicious samples that were not reported. It is unclear whether these occurrences were truly overlooked or if there were inconsistencies in the reporting, data management, or labeling protocols (e.g. potentially reporting/noting the conditions elsewhere and not in the LSPS scope).
One further open discussable point is the difference in performance between the models. We believe that this setting with medical conditions in the training set made the task of learning normality harder and required the algorithms to have robustness for large variations in the training data. Here, the ceVAE with its masking and inpainting tasks possibly performed best, since during training it had to cope with inputs with large perturbations and also learn how to handle and correct these perturbations. The irVAE and FPI possibly were not that resistant

to these outliers during training. This is further supported by the fact that the ceVAE considered some "normal", not-IF medical conditions as anomalies while the irVAE and FPI did not. One possible cause for the difference between the irVAE and the ceVAE could be the used scoring method. While the ceVAE uses the gradient directly and as such to a large portion can be invariant to the image intensity, the irVAE which uses a reconstruction error is more dependent on the image intensity (since an anomaly with an image intensity similar to the mean image intensity will likely not be able to stand out compared with an anomaly with extreme image intensities). The large performance discrepancy for FPI between the validation set and the refIF samples can possibly be explained by the similarity of the artificially created "toy-spheres" anomalies and the interpolation anomalies of the FPI task. In the MOOD 2020 challenge a similar performance gap could be observed. Perhaps the FPI team's subsequent work [Tan et al., 2021], which extended FPI and has shown good performance on additional datasets, might better translate the validation set's performance to the LSPS dataset.

Given the medical conditions in the LSPS training set and the presence of the one abnormal data sample in the well-curated HCP dataset, it is critical that the algorithms are capable of handling such occurrences during training. Furthermore, for use "in the wild", robustness to some anatomical variations (e.g. variations that occur with aging) might be needed or models need to be made more context-aware. Overall, the results demonstrated that only extremely obvious abnormalities were found, which a radiologist could spot with little effort and would only provide minor advantages in a day-to-day workflow, especially given huge labeled datasets which are readily accessible for these conditions. Furthermore, even these conditions were not properly recognized. The employed algorithm has difficulty detecting the more commonly overlooked and perhaps more 'useful' abnormalities. As a result, it is difficult to suggest the tested models for routine clinical usage. However, with the identification of imaging artifacts and big anomalies it may make useful prefiltering step for fully automated diagnosis pipelines, finding OoD data on which a subsequent algorithm may fail or behave unpredictably, or serving as a backup-check in such a LSPS scenario.

So we feel that to come closer to a translation to the real-world, we must reconsider the alignment of the medical goal with the metrics used, as well as the assumptions made and if they are plausible.

5 Discussion

In this chapter, we will put the previous chapters' findings into a broader research context and discuss various hypotheses and assumptions made throughout this thesis.

Localization scoring (residual vs. gradient vs. pixel-level modeling) We mostly focused on and used the residual of a reconstruction-based approach, the gradient of a density-based model or a combination of both. The residual approach is based on the assumption that by limiting a model capacity only the main factors of the data can be learned/represented and thus the model will not be able to reconstruct abnormal data successfully and have a large residual for abnormal data, which can pinpoint the anomaly. However, finding the right balance between model capacity and reconstruction is not trivial, and usually necessitates a validation set to be tuned to. Furthermore, restricting the model capacity hampers the ability to detect small and fine-grained anomalies. This can lead to suboptimal general performance in harder cases as also noted by Meissen et al. [2021]. The gradient-based approach might fix some of the described issues. It is based on the assumption that the score, i.e. the gradient of the log-likelihood and its magnitude, is a good indicator of the anomaly. While for simple unimodal distributions and points far away from the learned distribution this is the case, in the case of a more complicated distribution, interference can cause the assumption to be violated. If and how much this worsens the detection performance in practice is not clear. Also, due to the network architectures, the gradients are often noisy and can have convolutional artifacts, which can be smoothed out by using the

Smooth Gradient. From a theoretical perspective, the 'cleanest' approach would be the direct modeling of the pixel distribution. However, this carries its own problems and constraints. First, direct modeling of the pixel distribution might be aggravated by the curse of dimensionality. Ann second, doing this directly using autoregressive models has shown no optimal performance [Goldstein and Uchida, 2016]. Since the pixels are correlated with each other, a model would have to factor all correlations in and the modeling of this covariance might need enormous amounts of data. But recently models try to solve that by learning "super-pixel features" and modeling the "super-pixel" distributions [Marimont and Tarroni, 2020; Pinaya et al., 2021]. Nevertheless, we believe that using the gradient-based approach is a good compromise.

Assumptions inherent to VAE-based models Another assumption made, is that the approximated ELBO of a VAE is a good indicator of the anomaly. VAEs themselves were shown to be competitive with recent generative models and are often used for multiple use cases [Child, 2021]. However, using the log-likelihood as a score might not be optimal in all cases. First, Theis et al. [2015] showed that multiple factors can easily influence the likelihood and the actual content might not be the most deciding factor. Also, while mostly focusing on flow-based models, as shown by Nalisnick et al. [2018], generative models were criticized to focus mostly on low-level statistics and not on the content of the data, and thus also are vulnerable to out-of-distribution data and fail to 'know what they don't know'. Next, for high dimensional Gaussian distributions, almost all samples are not close to the mean or have the highest likelihood, but lie in a "soap bubble" [Huszár, 2017] around the mean (e.g. images drawn from an isotropic Gaussian distribution with zero mean would almost always seem noisy and not consistently grey, as the mean would). Thus, for example, a test-of-typicality [Nalisnick et al., 2019] score was proposed as an indicator of the anomaly. Here, a 'normal' likelihood range is determined by some normal data and only data in this 'normal likelihood range' is considered normal. While these issues are definitely a factor to consider and need further investigation and research, we believe they are not a major issue in the current state of the art. The interesting anomalies in the here-presented medical use cases are mostly near semantic anomalies, which do not lie far away from the data distribution. We believe that for these cases the previously described issues are not the crucial factor and do not harm the overall performance in a meaningful way.

Explicit modeling vs. modeling with a surrogate task While this density-based modeling of the data seems theoretically very sound and clean, the previous

section showed that in practice there are some points to consider. Another recent trend that has shown to be a good alternative is modeling with a surrogate task. While not yet being a prominent research direction at the time of proposing and implementing the design of the MOOD challenge, we already considered whether such approaches might be a very suitable option for this case. While these models do not directly model normal data, they use a surrogate task to 'trick' conventional and established segmentation techniques into finding abnormal regions of the data. For a challenge, like MOOD, where artificial anomalies are rendered into an image, training a model to segment other artificial anomalies might be a good option. However, the anomalies in the challenge were not known and thus vary a lot from the anomalies used to train the model. Thus the question is whether these models generalize to other artificial and more importantly, to real anomalies as well. FPI, the winning team of the MOOD challenge, used interpolated patches from different images/scans as an anomaly and trained a model to segment these. Similarly, the team Canon Medical Research Europe, second in the anomaly localization task, used a model trained to segment different interpolated spheres as one part of their submission. To counter the fact that these only work on synthetic anomalies, both teams have later published consecutive papers which show that this generalizes to different natural medical conditions/anomalies as well [Kascenas et al., 2021; Tan et al., 2020, 2021]. Interestingly, independent of the challenge, CutPaste [Li et al., 2021], which also uses 'cut and pasted' foreign image patches to detect and localize anomalies (but applied to an industrial inspection anomaly detection dataset) was proposed at a similar time frame as the consecutive papers. So this strategy might in fact generalize towards more general anomalies and is a promising future direction that can build on the success and progress in segmentation models. However, applying these methods unaltered to a medical real-world 3D dataset showed some room for further improvements.

Metrics One point that is worth mentioning is the use of metrics. Metrics are often used to measure performance and compare different models. However, to make such statements about model performance, it is important to understand what the metrics measure and how/if it aligns with the task at hand. As shown by Maier-Hein et al. [Maier-Hein et al., 2022; Reinke et al., 2021], particularly for medical image segmentation and instance detection, caution should be exercised when choosing metrics. E.g. for multiple instances with different sizes, the DSC overfocuses on the bigger instances and might not be a good measure of the detection performance. In the field of anomaly detection, AUROC has been the most established metric [Baur et al., 2018; Chen and Konukoglu, 2018; Goldstein and Uchida, 2016]. However, in recent years papers have argued that AP might be

better suited, especially for settings with unbalanced classes, as is often the case for anomaly detection [Ahmed and Courville, 2019]. Hence, AP has augmented and partially replaced the use of AUROC [Baur et al., 2020; Zimmerer et al., 2020]. In analogy, AP was also inherited as a metric for the newer anomaly localization task. Furthermore, AP is also related to the DSC and the DSC was also used as a measure of localization performance on a binarized localization scoring. However, as previously described, when dealing with different-sized instances, this can lead to a biased evaluation of the performance. In terms of anomaly localization, the detection/location of minor and challenging anomalies may provide the most value and practical advantage. These situations should be prioritized. However, using AP or the DSC, the opposite is the case, and the score is mainly influenced by bigger-sized instances. Thus, while not yet adopted by the research community, we believe that using additional object detection-based metrics, as done for the real-world evaluation, could bring further benefit and move the metrics closer to the task at hand.

Definition of normality As previously mentioned, "due to its label efficiency, generalizability and general applicability, anomaly localization is attributed with much promise in the field of automated medical image analysis and diagnosis". However, this statement is mainly based on the usage of anomaly detection models for finding general abnormal or pathological conditions. But this requires the model to be trained on normal data and thus learn what normal data looks like. But especially in the real world, it is hard to define a black-and-white differentiation between normal data and abnormal data. For example, a normal condition for an 80-year-old would be highly abnormal for a 20-year-old. And often the assessment in clinical practice uses a grading scheme and not a binary classification. Coincidentally, the anomaly score is also on a continuous scale, which would make it applicable to such a scenario. Nevertheless, the algorithms, which are based on the assumption to only be fed normal data, might have to be adopted to be able to deal with mostly normal data but also some representative abnormal data. Another possible approach could be the integration of more context (e.g. age, sex, ...) to help bring the images into a broader context.

Summary Anomaly detection and localization can learn what data looks like and point out anomalous data samples, which may then be utilized to assist clinicians in identifying anomalies. We employed a Variational Autoencoder (VAE) to learn the distribution of the data and demonstrated several ways for highlighting abnormalities. We showed that using self-supervised learning and hierarchical representations could increase performance, especially in situations with smaller

and more difficult-to-detect cases. We further investigated the approaches' performance and assessment in two contexts: an international public competitive setting and a real-world use-case for discovering incidental findings in a population study. Overall, the results were encouraging, and the algorithms could detect anomalies and incidental findings, but they fell short in more complex and difficult cases and were not yet dependable enough for real-world usage.

6 Supplementary Material

6.0.1 More ceVAE examples

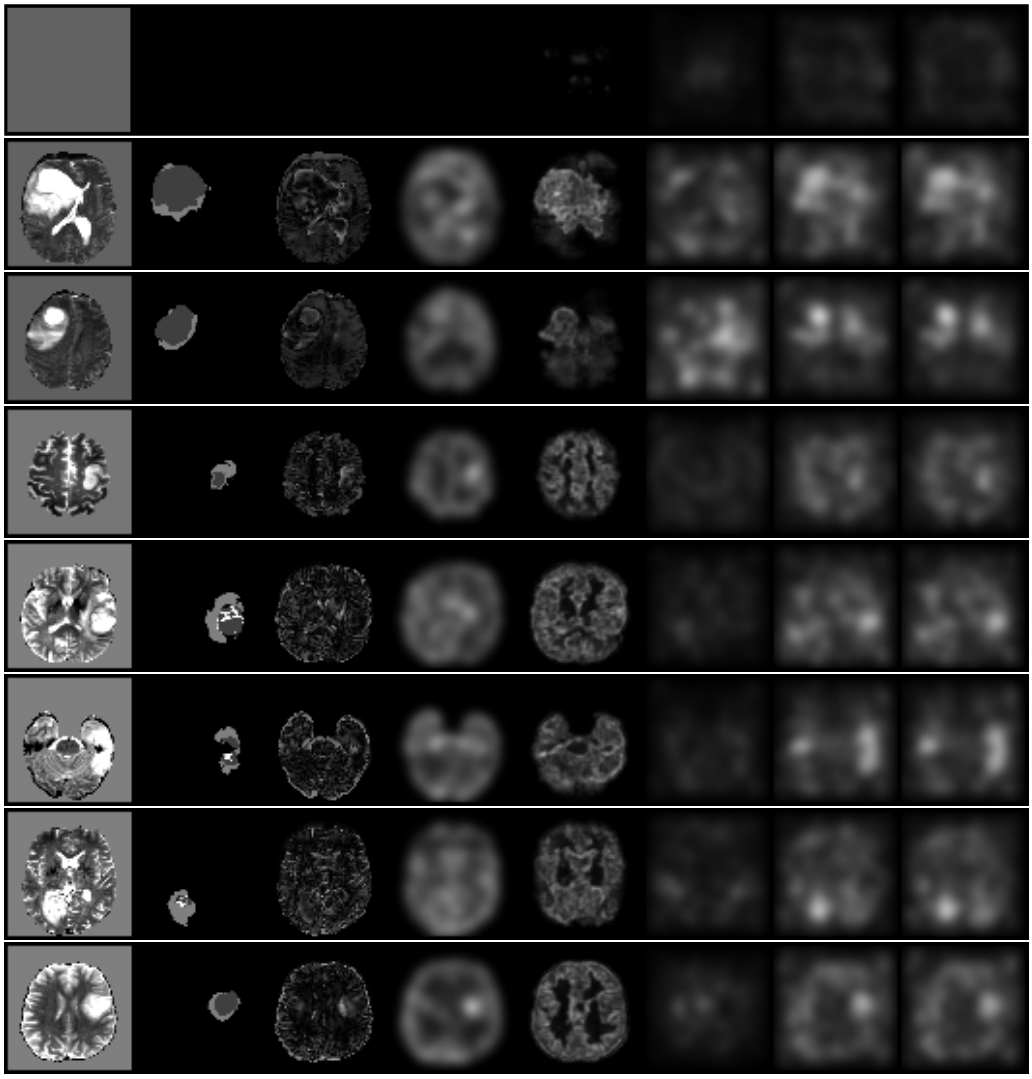


Figure 6.1: More samples and predictions as shown in Fig. 3.7, showing the original sample (I), the annotation (II), the reconstruction error (III), the smoothed reconstruction error (IV), the sampling variances (V), the reconstruction-loss gradient (VI), the KL-loss gradient (VII), and the ELBO gradient which approximates the *score* (VIII).

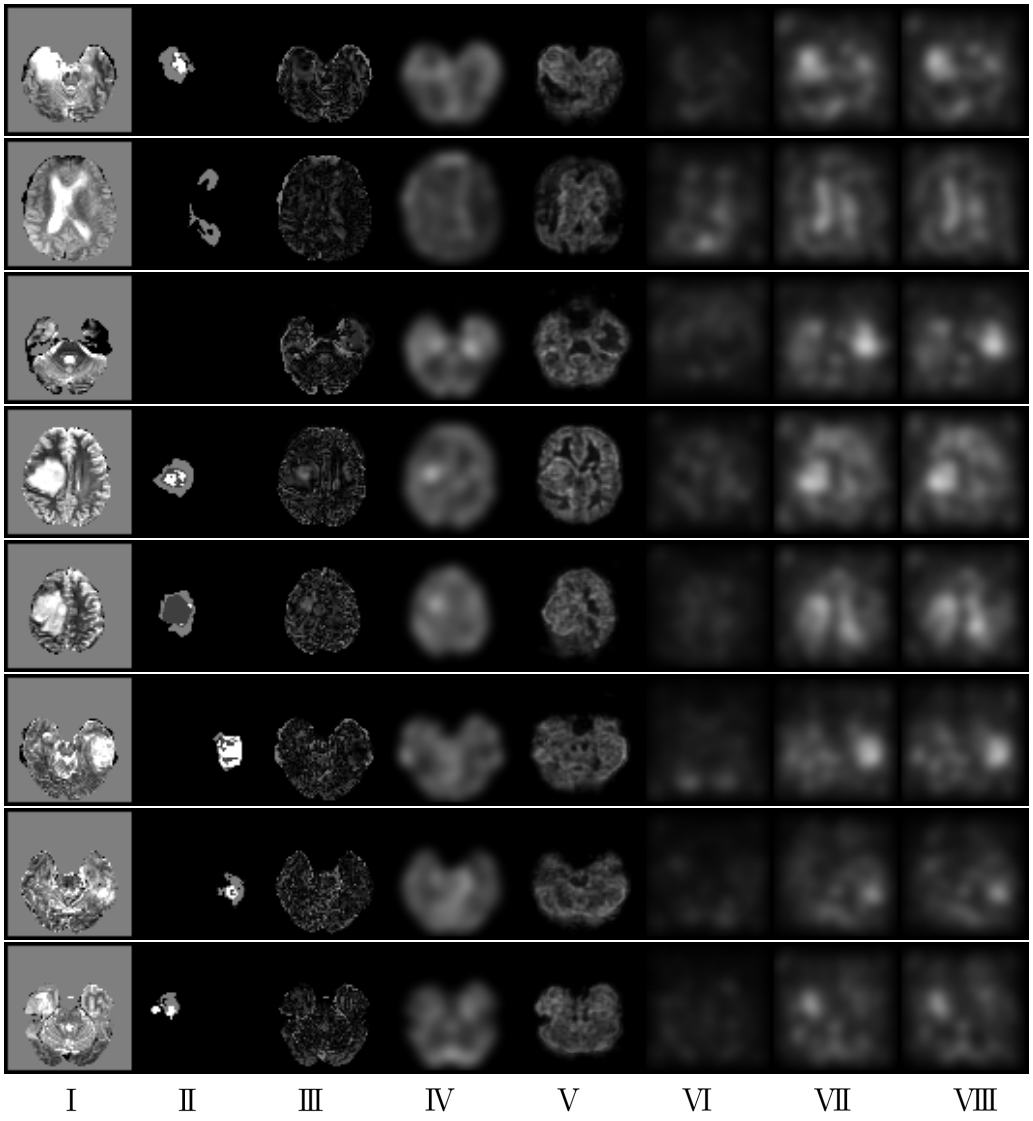
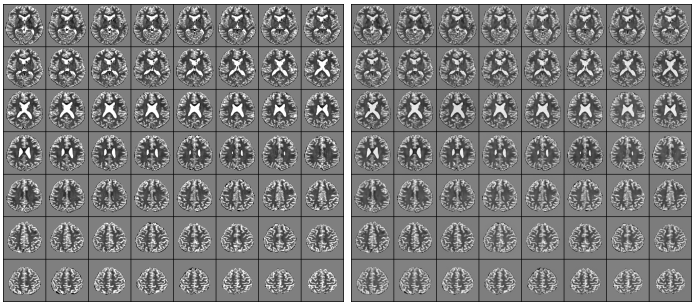


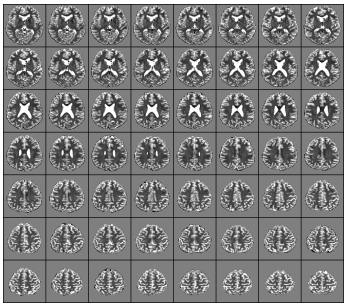
Figure 6.2: More samples and predictions as shown in Fig. 3.7, showing the original sample (I), the annotation (II), the reconstruction error (III), the smoothed reconstruction error (IV), the sampling variances (V), the reconstruction-loss gradient (VI), the KL-loss gradient (VII), and the ELBO gradient which approximates the *score* (VIII).

6.0.2 More pchVAE examples

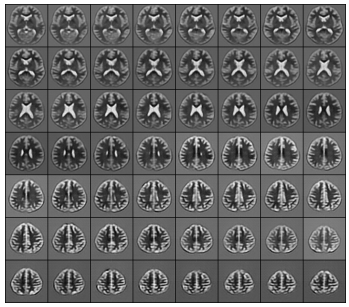


(a) Input .

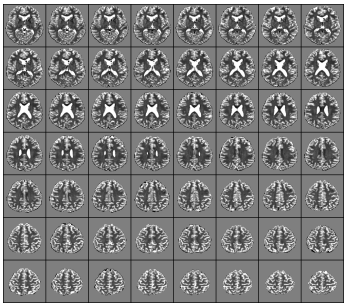
(b) VQVAE-2 reconstruction



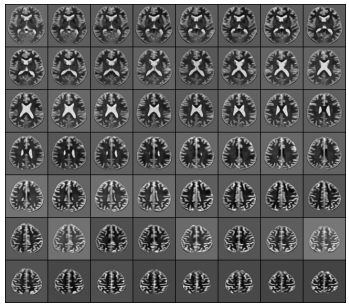
(c) Input .



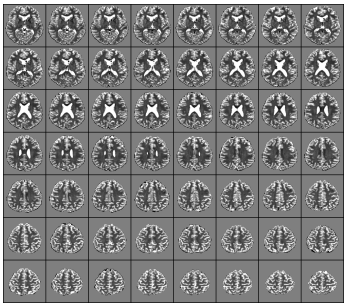
(d) VAE-low reconstruction



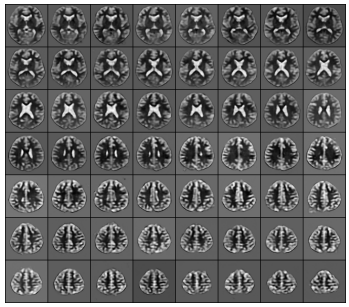
(e) Input .



(f) VAE-high reconstruction

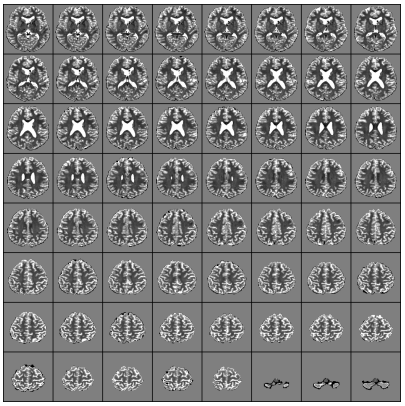


(g) Input

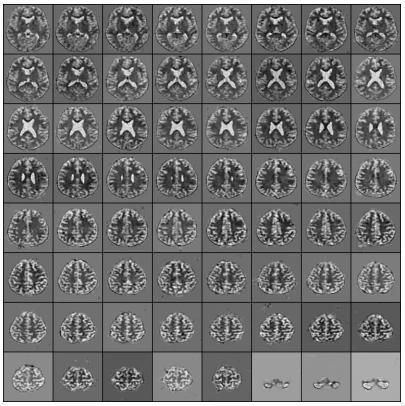


(h) pchVAE reconstruction

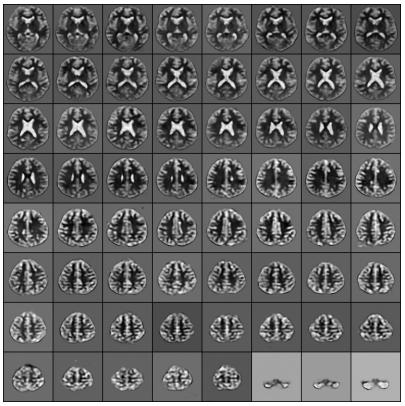
Figure 6.3: More reconstructed inputs of the different hierarchical models, similar to Fig. 3.17.



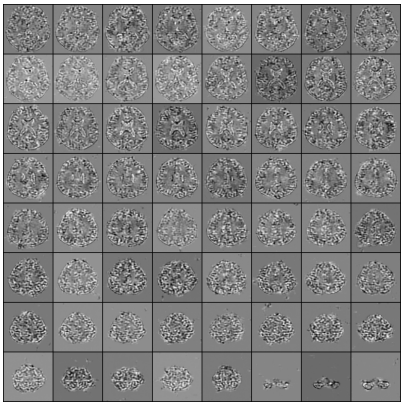
(a) Input



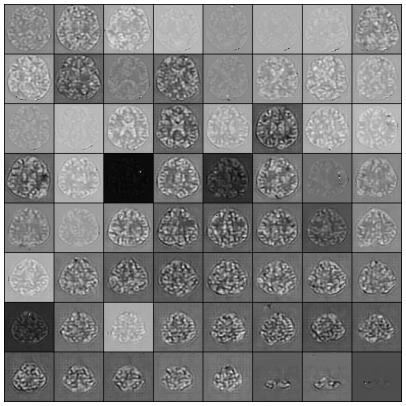
(b) pchVAE reconstruction



(c) pchVAE high (Eq. (3.22) term 1)



(d) pchVAE low (Eq. (3.22) term 2)



(e) pchVAE zero (Eq. (3.22) term 3)

Figure 6.4: More sample images of the pchVAE reconstruction, divided into the different pchVAE parts, similar to Fig. 3.17.

6.0.3 CRADL detail results

Pixel-level anomaly localization metrics on test datasets: **Values** are shown in the section selected based on the results of the best AUPRC scores on the validation set:

Pretext	Gen. Model	Score	HCP Synth. AUPRC	HCP Synth. AUROC	BraTS AUPRC	BraTS AUROC	ISLES AUPRC	ISLES AUROC
VAE	GMM 1 Comp	nll-grad	0.0236±0.0011	0.8501±0.0035	0.1282±0.0123	0.8889±0.0023	**0.0563**±0.0033	0.86±0.0016
	GMM 2 Comp	nll-grad	0.0206±0.0006	0.8436±0.0022	0.1095±0.0096	0.8808±0.0021	0.0471±0.0024	0.8523±0.0036
	GMM 4 Comp	nll-grad	0.0348±0.004	0.8851±0.0054	**0.1295**±0.0173	0.8918±0.0038	0.0562±0.0057	0.8578±0.0035
	GMM 8 Comp	nll-grad	**0.048**±0.0086	0.9019±0.007	0.1345±0.013	0.8949±0.0026	0.0584±0.0058	0.8605±0.0028
	Real NVP	nll-grad	0.04±0.0066	0.8885±0.0077	0.0978±0.012	0.8666±0.009	0.0498±0.0098	0.8516±0.0048
	VAE	combi	**0.2491**±0.0063	**0.9546**±0.0005	0.2269±0.0328	0.9198±0.0038	**0.077**±0.0122	**0.8745**±0.0059
		kl-grad	0.0373±0.0032	0.8657±0.0014	0.0772±0.0091	0.8446±0.011	0.0409±0.0047	0.8466±0.0081
		rec	0.2101±0.003	0.9511±0.0003	**0.2976**±0.0035	**0.9248**±0.0006	0.0513±0.0001	0.8532±0.0023
ceVAE	GMM 1 Comp	nll-grad	0.1072±0.0109	0.901±0.01	**0.2343**±0.0816	0.9129±0.0159	**0.0618**±0.0176	0.86±0.0089
	GMM 2 Comp	nll-grad	0.0757±0.0057	0.8952±0.0041	0.177±0.0426	0.9062±0.0097	0.0449±0.0093	0.8445±0.0068
	GMM 4 Comp	nll-grad	0.0967±0.0156	0.9116±0.0085	0.1906±0.0184	0.9105±0.004	0.0511±0.0123	0.8456±0.006
	GMM 8 Comp	nll-grad	**0.1122**±0.016	0.9223±0.0058	0.2068±0.033	0.9119±0.006	0.0612±0.0084	0.8488±0.0055
	Real NVP	nll-grad	0.0606±0.0148	0.9008±0.0101	0.1072±0.0159	0.8756±0.0079	0.0304±0.0007	0.8157±0.0013
	VAE	combi	**0.1716**±0.0146	0.9212±0.004	**0.483**±0.0299	**0.9482**±0.0032	**0.1451**±0.0125	**0.8794**±0.0022
		kl-grad	0.0702±0.0069	0.8586±0.0047	0.3394±0.067	0.9252±0.0163	0.1085±0.0163	0.8785±0.0059
		rec	0.0913±0.0023	**0.9266**±0.0017	0.4073±0.0389	0.9269±0.0074	0.0653±0.0044	0.8544±0.005
CRADL	GMM 1 Comp	nll-grad	0.2263±0.0112	0.9664±0.0017	0.3341±0.0402	0.9357±0.0035	**0.1859**±0.0385	**0.8977**±0.0033
	GMM 2 Comp	nll-grad	0.2243±0.0125	0.9685±0.0017	**0.3802**±0.0163	**0.9418**±0.0009	0.1653±0.02	0.8955±0.0029
	GMM 4 Comp	nll-grad	**0.2875**±0.0101	**0.9741**±0.0006	0.3383±0.0161	0.9384±0.0012	0.1441±0.0024	0.8935±0.003
	GMM 8 Comp	nll-grad	0.3246±0.0076	0.9779±0.0003	0.2908±0.0199	0.9309±0.0022	0.1257±0.0151	0.8906±0.0019
	Real NVP	nll-grad	0.0924±0.0097	0.9397±0.0031	0.1362±0.0102	0.8736±0.0068	0.0393±0.0044	0.8213±0.0153
INN		nll-grad	0.0148±0.0005	0.7618±0.0018	0.3563±0.0023	0.9139±0.002	0.0443±0.0017	0.8307±0.0047

Pixel-level anomaly localization metrics on validation datasets: **Values** show the best AUPRC scores which are used for hyperparameter selection on the test set:

Pretext	Gen. Model	Score	HCP Synth. AUPRC	HCP Synth. AUROC	BraTS AUPRC	BraTS AUROC	ISLES AUPRC	ISLES AUROC
VAE	GMM 1 Comp	nll-grad	0.0353±0.0056	0.8547±0.0025	0.0935±0.0047	0.8841±0.0022	**0.0676**±0.0149	0.8745±0.0095
	GMM 2 Comp	nll-grad	0.0276±0.0018	0.8487±0.0031	0.0799±0.0046	0.8751±0.0038	0.0548±0.0102	0.8632±0.0067
	GMM 4 Comp	nll-grad	0.0503±0.0071	0.8859±0.0043	**0.1009**±0.0131	0.8897±0.0058	0.0627±0.0064	0.8701±0.0032
	GMM 8 Comp	nll-grad	**0.0774**±0.0035	0.9088±0.0023	0.0925±0.0073	0.8875±0.0041	0.0624±0.0029	0.8805±0.0006
	Real NVP	nll-grad	0.0395±0.0008	0.8665±0.0021	0.0691±0.0036	0.8584±0.0052	0.0509±0.0091	0.863±0.0091
	VAE	combi	**0.2945**±0.0059	*0.9527*±0.0005	0.1842±0.0403	*0.9171*±0.0061	**0.0894**±0.012	*0.8812*±0.0051
		kl-grad	0.0735±0.0012	0.8771±0.0007	0.0677±0.0096	0.8553±0.0131	0.041±0.0009	0.8456±0.0044
		rec	0.2081±0.0042	0.9426±0.0004	**0.2248**±0.0077	0.9119±0.0015	0.0543±0.0082	0.8618±0.0056
ceVAE	GMM 1 Comp	nll-grad	0.1212±0.024	0.9057±0.011	**0.2313**±0.0829	0.9181±0.0138	**0.0609**±0.0166	0.8633±0.0059
	GMM 2 Comp	nll-grad	0.0741±0.0142	0.8989±0.0065	0.1611±0.0291	0.9126±0.0051	0.0385±0.0056	0.8436±0.005
	GMM 4 Comp	nll-grad	0.1067±0.0132	0.9167±0.0039	0.1955±0.0214	0.9163±0.0024	0.0423±0.0058	0.8513±0.0033
	GMM 8 Comp	nll-grad	**0.1464**±0.0285	0.9286±0.007	0.1841±0.0162	0.911±0.0002	0.0404±0.0054	0.8471±0.0044
	Real NVP	nll-grad	0.0907±0.0144	0.9002±0.0081	0.0784±0.0144	0.8681±0.0088	0.0311±0.0017	0.8223±0.0094
	VAE	combi	**0.2183**±0.0206	*0.9148*±0.0057	**0.4321**±0.005	*0.9393*±0.0038	**0.1628**±0.0242	*0.8847*±0.0042
		kl-grad	0.096±0.0096	0.8655±0.0037	0.3337±0.04	0.9317±0.0078	0.0956±0.0278	0.8751±0.0092
		rec	0.1163±0.0019	0.9117±0.0021	0.2884±0.0403	0.9068±0.0088	0.1321±0.029	0.8649±0.0041
CRADL	GMM 1 Comp	nll-grad	0.3176±0.0102	0.9671±0.0007	0.2796±0.0244	0.9294±0.0007	**0.3295**±0.0279	*0.9251*±0.0029
	GMM 2 Comp	nll-grad	0.3125±0.0095	0.9686±0.0005	**0.3334**±0.0105	*0.9363*±0.0012	0.3281±0.0155	0.9197±0.002
	GMM 4 Comp	nll-grad	**0.3338**±0.0111	0.9685±0.0009	0.2597±0.0451	0.9262±0.009	0.2989±0.0168	0.9117±0.0033
	GMM 8 Comp	nll-grad	0.3297±0.003	*0.9721*±0.0004	0.2518±0.0102	0.927±0.0003	0.2765±0.0136	0.9099±0.001
	Real NVP	nll-grad	0.1276±0.0043	0.9347±0.0004	0.137±0.0107	0.8876±0.0031	0.1039±0.0141	0.864±0.0095

6.0.4 Definition of refIFs for 4.2

Considered as IF:

- Acute stroke
- Acute intracranial/ intraspinal hemorrhage
- Solid cerebral mass:
  - Supratentorial mass $>$ 2cm
  - Infratentorial $>$ 1cm
  - Multiple masses
  - Mass with edema/ csf blockage/ midline shift
- Pituitary gland mass
- Suspicious cerebral/ meningeal mass in need of clarification (except uncomplicated meningioma)
- Neurofibromatosis with $\geqslant$ 5 neurofibromas
- Cerebral vascular malformation with risk of hemorrhage
- Suspicious solid mass in the viscerocranium $>$ 2cm
- Intracranial aneurysm

Not considered as IF:

- Unspecific white matter lesions
- Not acute stroke
- Multiple sclerosis
- Hydrocephalus
- Asymmetric ventricles
- Enlarged periventricular space
- Megacisterna magna
- Reduced brain mass
- Lipoma
- Congenital disorder
- Chiari malformation
- Developmental venous anomaly
- Mastoiditis
- Syringomyelia
- Hygroma
- Leukoencephalopathy
- Pineal gland cyst
- Uncomplicated meningioma

6.0.5 Samples from the LSPS validation set

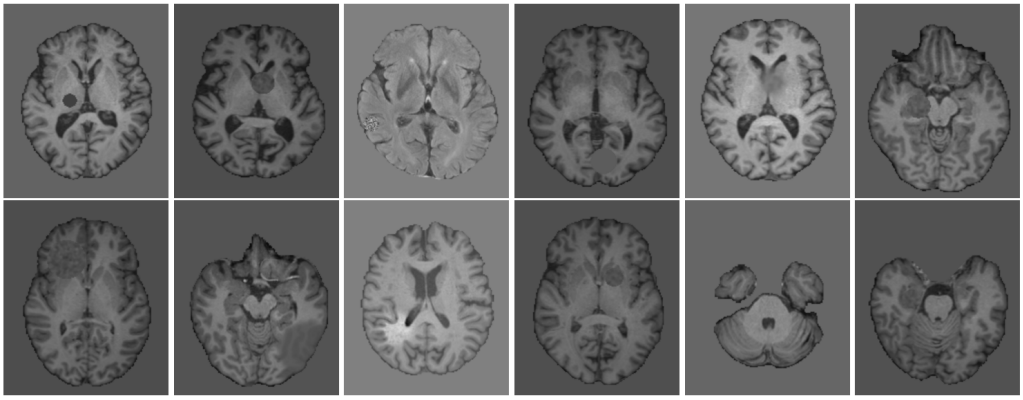


Figure 6.5: LSPS validation set "toy-sphere" anomalies. Artificial "toy-sphere" anomalies are rendered into validation set images, to generate a validation set with anomalies.

Bibliography

Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. **AND: Autoregressive Novelty Detectors**. *arXiv:1807.01653 [cs]*, July 2018a. URL http://arxiv.org/abs/1807.01653. arXiv: 1807.01653.

Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. **Latent Space Autoregression for Novelty Detection**. *arXiv:1807.01653 [cs]*, July 2018b. URL http://arxiv.org/abs/1807.01653. arXiv: 1807.01653.

Faruk Ahmed and Aaron Courville. **Detecting semantic anomalies**. *ArXiv*, abs/1908.04388, August 2019. URL https://arxiv.org/abs/1908.04388.

Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. **GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training**. *arXiv:1805.06725 [cs]*, May 2018. URL http://arxiv.org/abs/1805.06725. arXiv: 1805.06725.

Guillaume Alain and Yoshua Bengio. **What Regularized Auto-Encoders Learn from the Data Generating Distribution**. 2012. URL https://arxiv.org/pdf/1211.4246.pdf.

Jinwon An and Sungzoon Cho. **Variational Autoencoder based Anomaly Detection using Reconstruction Probability**. 2015. URL http://dm.snu.ac.kr/static/docs/TR/SNUDM-TR-2015-03.pdf.

Anomaly Detection for Scientific Discovery. **Sharon Li, Challenges and Opportunities in Out-of-distribution Detection**, February 2022. URL https://www.youtube.com/watch?v=X8XTOiNin0I. [2022-07-26].

Anand Avati, Martin Seneviratne, Emily Xue, Zhen Xu, Balaji Lakshminarayanan, and Andrew M. Dai. **BEDS-Bench: Behavior of EHR-models under Distributional Shift–A Benchmark**, July 2021. URL http://arxiv.org/abs/2107.08189. [2022-07-25]. arXiv:2107.08189 [cs].

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. **Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features**. *Scientific Data*, 4:170117, 2017. ISSN 2052-4463. doi: 10.1038/sdata.2017.117.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. **Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images**. *arXiv:1804.04488 [cs]*, April 2018. URL http://arxiv.org/abs/1804.04488. arXiv: 1804.04488.

Christoph Baur, Stefan Denner, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. **Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study**. *arXiv:2004.03271 [cs, eess]*, April 2020. URL http://arxiv.org/abs/2004.03271. arXiv: 2004.03271.

Christoph Baur, Benedikt Wiestler, Mark Muehlau, Claus Zimmer, Nassir Navab, and Shadi Albarqouni. **Modeling Healthy Anatomy with Artificial Intelligence for Unsupervised Anomaly Detection in Brain MRI**. *Radiology: Artificial Intelligence*, 3(3):e190169, May 2021. doi: 10.1148/ryai.2021190169. URL https://pubs.rsna.org/doi/abs/10.1148/ryai.2021190169.

Laura Beggel, Michael Pfeiffer, and Bernd Bischl. **Robust Anomaly Detection in Images using Adversarial Autoencoders**. *arXiv:1901.06355 [cs, stat]*, January 2019. URL http://arxiv.org/abs/1901.06355. arXiv: 1901.06355.

Marcel Bengs, Finn Behrendt, Julia Krüger, Roland Opfer, and Alexander Schlaefer. **3-Dimensional Deep Learning with Spatial Erasing for Unsupervised Anomaly Segmentation in Brain MRI**. *arXiv:2109.06540 [cs, eess]*, September 2021. URL http://arxiv.org/abs/2109.06540. arXiv: 2109.06540 version: 1.

Cosmin I. Bercea, Benedikt Wiestler, Daniel Rueckert, and Shadi Albarqouni. **Federated disentangled representation learning for unsupervised brain anomaly detection**. *Nature Machine Intelligence*, pages 1–11, August 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00515-2. URL https://www.nature.com/articles/s42256-022-00515-2. Publisher: Nature Publishing Group.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. **MVTec AD – A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection**. pages 9592–9600, 2019a. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Bergmann_MVTec_AD_--_A_Comprehensive_Real-World_Dataset_for_Unsupervised_Anomaly_CVPR_2019_paper.html.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. **Uninformed Students: Student-Teacher Anomaly Detection with Discriminative Latent Embeddings**. *arXiv:1911.02357 [cs]*, November 2019b. URL http://arxiv.org/abs/1911.02357. arXiv: 1911.02357 version: 1.

Christopher M. Bishop. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0-387-31073-8.

Noam Brown and Tuomas Sandholm. **Superhuman AI for heads-up no-limit poker: Libratus beats top professionals**. *Science*, 359(6374):418–424, January 2018. doi: 10.1126/science.aao1733. URL https://www.science.org/doi/10.1126/science.aao1733. Publisher: American Association for the Advancement of Science.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. **Language Models are Few-Shot Learners**. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

R. J. M. Bruls and R. M. Kwee. **Workload for radiologists during on-call hours: dramatic increase in the past 15 years**. *Insights into Imaging*, 11(1):121, November 2020. ISSN 1869-4101. doi: 10.1186/s13244-020-00925-z.

Chih-Chung Chang and Chih-Jen Lin. **LIBSVM: A library for support vector machines**. *ACM TIST*, 2(3):27:1–27:27, 2011.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. **A Simple Framework for Contrastive Learning of Visual Representations**.

arXiv:2002.05709 [cs, stat], June 2020. URL http://arxiv.org/abs/2002.05709. arXiv: 2002.05709.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, and Ilya Sutskever. **InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets**. 2016. URL https://arxiv.org/pdf/1606.03657.pdf.

Xiaoran Chen and Ender Konukoglu. **Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders**. April 2018. URL https://openreview.net/forum?id=H1nGLZ2oG.

Xiaoran Chen, Nick Pawlowski, Martin Rajchl, Ben Glocker, and Ender Konukoglu. **Deep Generative Models in the Real-World: An Open Challenge from Medical Imaging**. *CoRR*, abs/1806.05452, 2018.

Rewon Child. **Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images**, March 2021. URL http://arxiv.org/abs/2011.10650. [2022-07-29]. arXiv:2011.10650 [cs].

Hyunsun Choi, Eric Jang, and Alexander A. Alemi. **WAIC, but Why? Generative Ensembles for Robust Anomaly Detection**. *arXiv:1810.01392 [cs, stat]*, October 2018. URL http://arxiv.org/abs/1810.01392. arXiv: 1810.01392.

Bin Dai and David Wipf. **Diagnosing and enhancing VAE models**. In *ICLR*, 2019.

A. P. Dempster, N. M. Laird, and D. B. Rubin. **Maximum Likelihood from Incomplete Data via the EM Algorithm**. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 0035-9246. URL https://www.jstor.org/stable/2984875. Publisher: [Royal Statistical Society, Wiley].

Laurent Dinh, Jascha Sohl-Dickstein, Google Brain, and Samy Bengio. **Density estimation using Real NVP**. 2016. URL https://arxiv.org/pdf/1605.08803.pdf.

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. **Unsupervised Visual Representation Learning by Context Prediction**, January 2016. URL http://arxiv.org/abs/1505.05192. [2022-07-28]. arXiv:1505.05192 [cs].

Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. **Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks**, June 2015. URL http://arxiv.org/abs/1406.6909. [2022-07-28]. arXiv:1406.6909 [cs].

Trafton Drew, Melissa L. H. Vo, and Jeremy M. Wolfe. **"The invisible gorilla strikes again: Sustained inattentional blindness in expert observers"**. *Psychological science*, 24(9):1848–1853, September 2013. ISSN 0956-7976. doi: 10.1177/0956797613479386. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3964612/.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. **Unsupervised Representation Learning by Predicting Image Rotations**. February 2022. URL https://openreview.net/forum?id=S1v4N2l0-.

GitHub. **MIC-DKFZ/mood**, January 2021. URL https://github.com/MIC-DKFZ/mood. [2021-02-02]. original-date: 2020-04-22T08:55:16Z.

Izhak Golan and Ran El-Yaniv. **Deep Anomaly Detection Using Geometric Transformations**, November 2018. URL http://arxiv.org/abs/1805.10917. [2022-07-28]. arXiv:1805.10917 [cs, stat].

Markus Goldstein. **Anomaly Detection in Large Datasets**. June 2014. ISBN 978-3-8439-1572-4.

Markus Goldstein and Seiichi Uchida. **A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data**. *PLOS ONE*, 11 (4):e0152173, April 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0152173. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152173.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. **Deep Learning**. MIT Press, 2016.

Sreelekha Guggilam, S. M. Arshad Zaidi, Varun Chandola, and Abani Patra. **Bayesian Anomaly Detection Using Extreme Value Theory**. *arXiv:1905.12150 [cs, stat]*, May 2019. URL http://arxiv.org/abs/1905.12150. arXiv: 1905.12150.

Albert Haque, Arnold Milstein, and Li Fei-Fei. **Illuminating the dark spaces of healthcare with ambient intelligence**. *Nature*, 585(7824):193–202, September 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2669-y. URL https://www.

nature.com/articles/s41586-020-2669-y. Number: 7824 Publisher: Nature Publishing Group.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. **Masked Autoencoders Are Scalable Vision Learners**. *arXiv:2111.06377 [cs]*, December 2021. URL http://arxiv.org/abs/2111.06377. arXiv: 2111.06377.

Will Douglas Heaven. **Google's medical AI was super accurate in a lab. Real life was a different story.**, 2020. URL https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/. [2022-07-26].

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. **A Benchmark for Anomaly Segmentation**. *arXiv:1911.11132 [cs]*, November 2019. URL http://arxiv.org/abs/1911.11132. arXiv: 1911.11132 version: 1.

Rui Huang, Andrew Geng, and Yixuan Li. **On the Importance of Gradients for Detecting Distributional Shifts in the Wild**. October 2021. URL https://openreview.net/forum?id=fmiwLdJCmLS.

Ferenc Huszár. **Gaussian Distributions are Soap Bubbles**, November 2017. URL https://www.inference.vc/high-dimensional-gaussian-distributions-are-soap-bubble/. [2022-07-29].

Conor Igoe, Youngseog Chung, Ian Char, and Jeff Schneider. **How Useful are Gradients for OOD Detection Really?**, May 2022. URL http://arxiv.org/abs/2205.10439. [2022-07-28]. arXiv:2205.10439 [cs] version: 1.

Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. **nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation**. *arXiv:1809.10486 [cs]*, September 2018. URL http://arxiv.org/abs/1809.10486. arXiv: 1809.10486.

Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, Martin Bendszus, Klaus H. Maier-Hein, and Philipp Kickingereder. **Automated brain extraction of multisequence MRI using artificial neural networks**. *Human Brain Mapping*, 40(17):4952–4964, 2019. ISSN 1097-0193. doi: 10.1002/hbm.24750. URL https:

//onlinelibrary.wiley.com/doi/abs/10.1002/hbm.24750. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.24750.

C. Daniel Johnson, Mei-Hsiu Chen, Alicia Y. Toledano, Jay P. Heiken, Abraham Dachman, Mark D. Kuo, Christine O. Menias, Betina Siewert, Jugesh I. Cheema, Richard G. Obregon, Jeff L. Fidler, Peter Zimmerman, Karen M. Horton, Kevin Coakley, Revathy B. Iyer, Amy K. Hara, Robert A. Halvorsen, Giovanna Casola, Judy Yee, Benjamin A. Herman, Lawrence J. Burgart, and Paul J. Limburg. **Accuracy of CT Colonography for Detection of Large Adenomas and Cancers**. *New England Journal of Medicine*, 359(12):1207–1217, 2008. doi: 10.1056/NEJMoa0800996. URL https://doi.org/10.1056/NEJMoa0800996. _eprint: https://doi.org/10.1056/NEJMoa0800996.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. **Highly accurate protein structure prediction with AlphaFold**. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Number: 7873 Publisher: Nature Publishing Group.

Ryo Kamoi and Kei Kobayashi. **Why is the Mahalanobis Distance Effective for Anomaly Detection?**, April 2020. URL http://arxiv.org/abs/2003.00402. [2022-07-28]. arXiv:2003.00402 [cs, stat].

Antanas Kascenas, Nicolas Pugeault, and Alison Q. O'Neil. **Denoising Autoencoders for Unsupervised Anomaly Detection in Brain MRI**. December 2021. URL https://openreview.net/forum?id=Bm8-t_ggzPD.

Diederik P. Kingma and Jimmy Ba. **Adam: A Method for Stochastic Optimization**, January 2017. URL http://arxiv.org/abs/1412.6980. [2022-09-07]. arXiv:1412.6980 [cs].

Diederik P. Kingma and Prafulla Dhariwal. **Glow: Generative Flow with Invertible 1x1 Convolutions**. *arXiv:1807.03039 [cs, stat]*, July 2018. URL http://arxiv.org/abs/1807.03039. arXiv: 1807.03039.

Diederik P Kingma and Max Welling. **Auto-Encoding Variational Bayes**. 2013. URL https://arxiv.org/pdf/1312.6114v10.pdf.

B. Kiran, Dilip Thomas, Ranjith Parakkal, B. Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. **An Overview of Deep Learning Based Methods for Unsupervised and Semi-Supervised Anomaly Detection in Videos**. *Journal of Imaging*, 4(2):36, February 2018. doi: 10.3390/jimaging4020036. URL https://www.mdpi.com/2313-433X/4/2/36.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. **ImageNet Classification with Deep Convolutional Neural Networks**. In F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. **Autoencoding beyond pixels using a learned similarity metric**. 2015. URL https://arxiv.org/pdf/1512.09300.pdf.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. **CutPaste: Self-Supervised Learning for Anomaly Detection and Localization**. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9659–9669, June 2021. doi: 10.1109/CVPR46437.2021.00954. ISSN: 2575-7075.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra, Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A W M Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. **A Survey on Deep Learning in Medical Image Analysis**. 2017. URL https://arxiv.org/pdf/1702.05747.pdf.

Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. **An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution**. *arXiv:1807.03247 [cs, stat]*, July 2018. URL http://arxiv.org/abs/1807.03247. arXiv: 1807.03247.

Xiaoxiao Liu, Marc Niethammer, Roland Kwitt, Matthew McCormick, and Stephen Aylward. **Low-Rank to the Rescue – Atlas-based Analyses in the Presence of Pathologies**. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 17(Pt 3):97–104, 2014. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4857018/.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. **RePaint: Inpainting using Denoising Diffusion Probabilistic Models**, June 2022. URL http://arxiv.org/abs/2201.09865. [2022-07-28]. arXiv:2201.09865 [cs] version: 3.

Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. **BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling**. *arXiv:1902.02102 [cs, stat]*, February 2019. URL http://arxiv.org/abs/1902.02102. arXiv: 1902.02102.

Oskar Maier, Bjoern H. Menze, Janina von der Gablentz, Levin Hani, Mattias P. Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen, Daan Christiaens, Francis Dutil, Karl Egger, Chaolu Feng, Ben Glocker, Michael Götz, Tom Haeck, Hanna-Leena Halme, Mohammad Havaei, Khan M. Iftekharuddin, Pierre-Marc Jodoin, Konstantinos Kamnitsas, Elias Kellner, Antti Korvenoja, Hugo Larochelle, Christian Ledig, Jia-Hong Lee, Frederik Maes, Qaiser Mahmood, Klaus H. Maier-Hein, Richard McKinley, John Muschelli, Chris Pal, Linmin Pei, Janaki Raman Rangarajan, Syed M. S. Reza, David Robben, Daniel Rueckert, Eero Salli, Paul Suetens, Ching-Wei Wang, Matthias Wilms, Jan S. Kirschke, Ulrike M. Kr Amer, Thomas F. Münte, Peter Schramm, Roland Wiest, Heinz Handels, and Mauricio Reyes. **ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI**. *Medical Image Analysis*, 35:250–269, 2017. ISSN 1361-8423. doi: 10.1016/j.media.2016.07.009.

Lena Maier-Hein, Annika Reinke, Michal Kozubek, Anne L. Martel, Tal Arbel, Matthias Eisenmann, Allan Hanbury, Pierre Jannin, Henning Müller, Sinan Onogur, Julio Saez-Rodriguez, Bram van Ginneken, Annette Kopp-Schneider, and Bennett A. Landman. **BIAS: Transparent reporting of biomedical image analysis challenges**. *Medical Image Analysis*, 66:101796, December 2020. ISSN 1361-8415. doi: 10.1016/j.media.2020.101796. URL https://www.sciencedirect.com/science/article/pii/S1361841520301602.

Lena Maier-Hein, Annika Reinke, Evangelia Christodoulou, Ben Glocker, Patrick Godau, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A. Emre Kavur, Tim Rädsch, Minu D. Tizabi, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Peter Bankhead, Arriel Benis, M. Jorge Cardoso, Veronika Cheplygina, Beth Cimini, Gary S. Collins, Keyvan Farahani, Bram van Ginneken, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Alexandros Karargyris,

Alan Karthikesalingam, Hannes Kenngott, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, David Moher, Karel G. M. Moons, Henning Müller, Felix Nickel, Brennan Nichyporuk, Jens Petersen, Nasir Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clarisa Sánchez Gutiérrez, Shravya Shetty, Maarten van Smeden, Carole H. Sudre, Ronald M. Summers, Abdel A. Taha, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, and Paul F. Jäger. **Metrics reloaded: Pitfalls and recommendations for image analysis validation**, July 2022. URL `http://arxiv.org/abs/2206.01653`. [2022-07-29]. arXiv:2206.01653 [cs].

Sergio Naval Marimont and Giacomo Tarroni. **Anomaly detection through latent space restoration using vector-quantized variational autoencoders**. *arXiv:2012.06765 [cs, eess]*, December 2020. URL `http://arxiv.org/abs/2012.06765`. arXiv: 2012.06765 version: 1.

Felix Meissen, Georgios Kaissis, and Daniel Rueckert. **Challenging Current Semi-Supervised Anomaly Segmentation Methods for Brain MRI**. *arXiv:2109.06023 [eess]*, September 2021. URL `http://arxiv.org/abs/2109.06023`. arXiv: 2109.06023.

Felix Meissen, Benedikt Wiestler, Georgios Kaissis, and Daniel Rueckert. **On the Pitfalls of Using the Residual as Anomaly Score**. June 2022. URL `https://openreview.net/forum?id=ZsoHLeupa1D`.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Cagatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José Antonió Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. **The Multimodal Brain Tumor Image**

Segmentation Benchmark (BRATS). *IEEE transactions on medical imaging*, 34 (10):1993–2024, October 2015. ISSN 1558-254X. doi: 10.1109/TMI.2014.2377694.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. **Do Deep Generative Models Know What They Don't Know?** *arXiv:1810.09136 [cs, stat]*, October 2018. URL http://arxiv.org/abs/1810.09136. arXiv: 1810.09136.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. **Detecting Out-of-Distribution Inputs to Deep Generative Models Using a Test for Typicality**. *arXiv:1906.02994 [cs, stat]*, June 2019. URL http://arxiv.org/abs/1906.02994. arXiv: 1906.02994.

Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. **Representation Learning by Learning to Count**. August 2017. URL http://arxiv.org/abs/1708.06734.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. **Representation Learning with Contrastive Predictive Coding**, January 2019. URL http://arxiv.org/abs/1807.03748. [2022-07-28]. arXiv:1807.03748 [cs, stat].

OpenAI, Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P. d O. Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. **Dota 2 with Large Scale Deep Reinforcement Learning**, December 2019. URL http://arxiv.org/abs/1912.06680. [2022-07-26]. arXiv:1912.06680 [cs, stat].

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. **Automatic differentiation in PyTorch**. In *NIPS-W*, 2017.

Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. **Context Encoders: Feature Learning by Inpainting**. 2016. URL https://arxiv.org/pdf/1604.07379.pdf.

Nick Pawlowski, Matthew C. H. Lee, Martin Rajchl, Steven McDonagh, Enzo Ferrante, Konstantinos Kamnitsas, Sam Cooke, Susan Stevenson, Aneesh Khetani, Tom Newman, Fred Zeiler, Richard Digby, Jonathan P. Coles, Daniel Rueckert, David K. Menon, Virginia F. J. Newcombe, and Ben Glocker. **Unsupervised**

Lesion Detection in Brain CT using Bayesian Convolutional Autoencoders. April 2018. URL https://openreview.net/forum?id=S1hpzoisz.

Google Photos. **Photobombs begone with Magic Eraser in Google Photos**, 2021. URL https://blog.google/products/photos/magic-eraser/. [2022-07-28].

Claudio Piciarelli, Pankaj Mishra, and Gian Luca Foresti. **Image anomaly detection with capsule networks and imbalanced datasets**. *arXiv:1909.02755 [cs]*, September 2019. URL http://arxiv.org/abs/1909.02755. arXiv: 1909.02755.

Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. **Unsupervised Brain Anomaly Detection and Segmentation with Transformers**. In *arXiv:2102.11650 [cs, eess, q-bio]*, February 2021. URL http://arxiv.org/abs/2102.11650. arXiv: 2102.11650 version: 1.

Antonio Pinto, Alfonso Reginelli, Fabio Pinto, Giuseppe Lo Re, Federico Midiri, Carlo Muzj, Luigia Romano, and Luca Brunese. **Errors in imaging patients in the emergency setting**. *The British Journal of Radiology*, 89(1061):20150914, 2016. ISSN 1748-880X. doi: 10.1259/bjr.20150914.

Alec Radford, Luke Metz, and Soumith Chintala. **Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks**. 2015. URL https://arxiv.org/pdf/1511.06434.pdf.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. **Generating Diverse High-Fidelity Images with VQ-VAE-2**. *arXiv:1906.00446 [cs, stat]*, June 2019. URL http://arxiv.org/abs/1906.00446. arXiv: 1906.00446.

Annika Reinke, Matthias Eisenmann, Minu Dietlinde Tizabi, Carole H. Sudre, Tim Rädsch, Michela Antonelli, Tal Arbel, Spyridon Bakas, M. Jorge Cardoso, Veronika Cheplygina, Keyvan Farahani, Ben Glocker, Doreen Heckmann-Nötzel, Fabian Isensee, Pierre Jannin, Charles Kahn, Jens Kleesiek, Tahsin Kurc, Michal Kozubek, Bennett A. Landman, Geert Litjens, Klaus Maier-Hein, Anne Lousise Martel, Bjoern Menze, Henning Müller, Jens Petersen, Mauricio Reyes, Nicola Rieke, Bram Stieltjes, Ronald M. Summers, Sotirios A. Tsaftaris, Bram van Ginneken, Annette Kopp-Schneider, Paul Jäger, and Lena Maier-Hein. **Common limitations of performance metrics in biomedical image analysis**. April 2021. URL https://openreview.net/forum?id=76X9Mthzv4X.

Danilo J Rezende, Shakir Mohamed, Daan Wierstra, and Google DeepMind. **Stochastic Backpropagation and Approximate Inference in Deep Generative Models**. 2014. URL https://arxiv.org/pdf/1401.4082.pdf.

Danilo Jimenez Rezende and Shakir Mohamed. **Variational Inference with Normalizing Flows**, June 2016. URL http://arxiv.org/abs/1505.05770. [2022-08-16]. arXiv:1505.05770 [cs, stat].

Eitan Richardson and Yair Weiss. **On GANs and GMMs**. *arXiv:1805.12462 [cs]*, May 2018. URL http://arxiv.org/abs/1805.12462. arXiv: 1805.12462.

Michal Rolinek, Dominik Zietlow, and Georg Martius. **Variational Autoencoders Pursue PCA Directions (by Accident)**. *arXiv:1812.06775 [cs, stat]*, December 2018. URL http://arxiv.org/abs/1812.06775. arXiv: 1812.06775.

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.

Victor Saase, Holger Wenz, Thomas Ganslandt, Christoph Groden, and Máté E. Maros. **Simple statistical methods for unsupervised brain anomaly detection on MRI are competitive to deep learning methods**. *arXiv:2011.12735 [cs, eess]*, November 2020. URL http://arxiv.org/abs/2011.12735. arXiv: 2011.12735.

Ali S. Saber Tehrani, HeeWon Lee, Simon C. Mathews, Andrew Shore, Martin A. Makary, Peter J. Pronovost, and David E. Newman-Toker. **25-Year summary of US malpractice claims for diagnostic errors 1986-2010: an analysis from the National Practitioner Data Bank**. *BMJ quality & safety*, 22(8):672–680, August 2013. ISSN 2044-5423. doi: 10.1136/bmjqs-2012-001550.

Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. **Adversarially Learned One-Class Classifier for Novelty Detection**. *arXiv:1802.09088 [cs]*, February 2018. URL http://arxiv.org/abs/1802.09088. arXiv: 1802.09088.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. **Improved Techniques for Training GANs**. 2016. URL https://arxiv.org/pdf/1606.03498.pdf.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. **Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery**. In *Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery*, 2017. URL https://arxiv.org/pdf/1703.05921.pdf.

Alireza Shafaei, Mark Schmidt, and James J. Little. **A Less Biased Evaluation of Out-of-distribution Sample Detectors**. *arXiv:1809.04729 [cs, stat]*, August 2019. URL http://arxiv.org/abs/1809.04729. arXiv: 1809.04729.

Hoo-Chang Shin, Matthew R Orton, David J Collins, Simon J Doran, and Martin O Leach. **Stacked Autoencoders for Unsupervised Feature Learning and Multiple Organ Detection in a Pilot Study Using 4D Patient Data**. 2013. doi: 10.1109/TPAMI.2012.277.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. **A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play**. *Science*, 362(6419):1140–1144, December 2018. doi: 10.1126/science.aar6404. URL https://www.science.org/doi/10.1126/science.aar6404. Publisher: American Association for the Advancement of Science.

Karen Simonyan and Andrew Zisserman. **Very Deep Convolutional Networks for Large-Scale Image Recognition**, April 2015. URL http://arxiv.org/abs/1409.1556. [2022-07-25]. arXiv:1409.1556 [cs].

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. **SmoothGrad: removing noise by adding noise**. *CoRR*, abs/1706.03825, 2017.

Rebecca Smith-Bindman, Marilyn L. Kwan, Emily C. Marlow, Mary Kay Theis, Wesley Bolch, Stephanie Y. Cheng, Erin J. A. Bowles, James R. Duncan, Robert T. Greenlee, Lawrence H. Kushi, Jason D. Pole, Alanna K. Rahm, Natasha K. Stout, Sheila Weinmann, and Diana L. Miglioretti. **Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016**. *JAMA*, 322(9): 843–856, September 2019. ISSN 1538-3598. doi: 10.1001/jama.2019.11456.

Jiaming Song, Yang Song, and Stefano Ermon. **Unsupervised Out-of-Distribution Detection with Batch Normalization**. *arXiv:1910.09115 [cs, stat]*, October 2019. URL http://arxiv.org/abs/1910.09115. arXiv: 1910.09115 version: 1.

Yang Song and Stefano Ermon. **Generative Modeling by Estimating Gradients of the Data Distribution**. *arXiv:1907.05600 [cs, stat]*, October 2019. URL http://arxiv.org/abs/1907.05600. arXiv: 1907.05600.

Statista. **YouTube: hours of video uploaded every minute 2020**, 2022. URL https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/. [2022-08-22].

Synapse. **https://www.synapse.org/#!Synapse:syn21343101/wiki/599515**, 2020. URL https://www.synapse.org/#!Synapse:syn21343101/wiki/599515. [2021-02-02].

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. **Ladder Variational Autoencoders**. *Nips*, (Nips), 2016.

Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, and Bernhard Kainz. **Detecting Outliers with Foreign Patch Interpolation**. *arXiv:2011.04197 [cs]*, November 2020. URL http://arxiv.org/abs/2011.04197. arXiv: 2011.04197.

Jeremy Tan, Benjamin Hou, Thomas Day, John Simpson, Daniel Rueckert, and Bernhard Kainz. **Detecting Outliers with Poisson Image Interpolation**, July 2021. URL http://arxiv.org/abs/2107.02622. [2022-07-13]. arXiv:2107.02622 [cs].

Lucas Theis, Aäron Van Den Oord, and Matthias Bethge. **A note on the evaluation of generative models**. 2015. URL https://arxiv.org/pdf/1511.01844.pdf.

Nina Tuluptceva, Bart Bakker, Irina Fedulova, and Anton Konushin. **Perceptual Image Anomaly Detection**. In Shivakumara Palaiahnakote, Gabriella Sanniti di Baja, Liang Wang, and Wei Qi Yan, editors, *Pattern Recognition*, Lecture Notes in Computer Science, pages 164–178, Cham, 2020. Springer International Publishing. ISBN 978-3-030-41404-7. doi: 10.1007/978-3-030-41404-7_12.

Dennis Ulmer, Lotta Meijerink, and Giovanni Cinà. **Trust Issues: Uncertainty Estimation Does Not Enable Reliable OOD Detection On Medical Tabular Data**. In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, pages 341–354. PMLR, November 2020. URL https://proceedings.mlr.press/v136/ulmer20a.html. ISSN: 2640-3498.

Hristina Uzunova, Sandra Schultz, Heinz Handels, and Jan Ehrhardt. **Unsupervised pathology detection in medical images using conditional variational**

autoencoders. *International Journal of Computer Assisted Radiology and Surgery*, 14 (3):451–461, March 2019. ISSN 1861-6429. doi: 10.1007/s11548-018-1898-0. URL https://doi.org/10.1007/s11548-018-1898-0.

Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. **Pixel Recurrent Neural Networks**. 2016. URL https://arxiv.org/pdf/1601.06759.pdf.

D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. E. J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, S. Della Penna, D. Feinberg, M. F. Glasser, N. Harel, A. C. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S. E. Petersen, F. Prior, B. L. Schlaggar, S. M. Smith, A. Z. Snyder, J. Xu, E. Yacoub, and WU-Minn HCP Consortium. **The Human Connectome Project: a data acquisition perspective**. *NeuroImage*, 62(4):2222–2231, October 2012. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2012.02.018.

K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens. **Automated segmentation of multiple sclerosis lesions by model outlier detection**. *IEEE transactions on medical imaging*, 20(8):677–688, August 2001. ISSN 0278-0062. doi: 10.1109/42.938237.

Abinav Ravi Venkatakrishnan, Seong Tae Kim, Rami Eisawy, Franz Pfister, and Nassir Navab. **Self Supervised Out-of-Distribution Detection in Brain CT Scans**, November 2020. URL http://arxiv.org/abs/2011.05428. [2022-07-28]. arXiv:2011.05428 [cs, eess].

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. **Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion**. *JMLR*, 11 (Dec):3371–3408, 2010. ISSN ISSN 1533-7928. URL http://www.jmlr.org/papers/v11/vincent10a.html.

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu,

Demis Hassabis, Chris Apps, and David Silver. **Grandmaster level in StarCraft II using multi-agent reinforcement learning**. *Nature*, 575(7782):350–354, November 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1724-z. URL https://www.nature.com/articles/s41586-019-1724-z. Number: 7782 Publisher: Nature Publishing Group.

Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R. Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, Taylan Cemgil, S. M. Ali Eslami, and Olaf Ronneberger. **Contrastive Training for Improved Out-of-Distribution Detection**. *arXiv:2007.05566 [cs, stat]*, July 2020. URL http://arxiv.org/abs/2007.05566. arXiv: 2007.05566.

Han Xiao, Kashif Rasul, and Roland Vollgraf. **Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms**, August 2017. arXiv: cs.LG/1708.07747.

Du Xuefeng, Xin Wang, Gabriel Gozum, and Yixuan Li. **Unknown Aware Object Detection: Learning What You Don't Know from Videos in the Wild**, March 2022a. URL http://arxiv.org/abs/2203.03800. [2022-07-26]. arXiv:2203.03800 [cs].

Du Xuefeng, Zhaoning Wang, Mu Cai, and Yixuan Li. **VOS: Learning What You Don't Know by Virtual Outlier Synthesis**, May 2022b. URL http://arxiv.org/abs/2202.01197. [2022-07-26]. arXiv:2202.01197 [cs].

Suhang You, Kerem Tezcan, Xiaoran Chen, and Ender Konukoglu. **Unsupervised Lesion Detection via Image Restoration with a Normative Prior**. In *International Conference on Medical Imaging with Deep Learning – Full Paper Track*, London, United Kingdom, July 2019. URL https://openreview.net/forum?id=S1xg4W-leV.

YouTube. **Medical Out-of-Distribution Challenge 2020 (MOOD 2020)**, October 2020. URL https://www.youtube.com/watch?v=yOemj1TQfZU. [2021-12-15].

YouTube. **Medical Out-of-Distribution Analysis Challenge (MOOD) 2021**, October 2021. URL https://www.youtube.com/watch?v=PFwSzZMXcyE. [2021-12-15].

Amy Zhang, Nicolas Ballas, and Joelle Pineau. **A Dissection of Overfitting and Generalization in Continuous Reinforcement Learning**. *arXiv:1806.07937*

[cs, stat], June 2018. URL http://arxiv.org/abs/1806.07937. arXiv: 1806.07937.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. **Understanding deep learning requires rethinking generalization**. 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Richard Zhang, Phillip Isola, and Alexei A. Efros. **Colorful Image Colorization**, October 2016. URL http://arxiv.org/abs/1603.08511. [2022-07-28]. arXiv:1603.08511 [cs].

Shengjia Zhao, Jiaming Song, and Stefano Ermon. **Learning Hierarchical Features from Generative Models**. February 2017. URL http://arxiv.org/abs/1702.08396.

Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connelly Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsi, and Jiebo Luo. **CM-GAN: Image Inpainting with Cascaded Modulation GAN and Object-Aware Training**, July 2022. URL http://arxiv.org/abs/2203.11947. [2022-07-28]. arXiv:2203.11947 [cs] version: 3.

David Zimmerer, Jens Petersen, Simon AA Kohl, and Klaus H Maier-Hein. **A Case for the Score: Identifying Image Anomalies using Variational Autoencoder Gradients**. 2018.

David Zimmerer, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus Maier-Hein. **Unsupervised Anomaly Localization using Variational Auto-Encoders**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 289–297. Springer, 2019a.

David Zimmerer, Jens Petersen, Fabian Isensee, and Klaus Maier-Hein. **Context-encoding Variational Autoencoder for Unsupervised Anomaly Detection**. In *International Conference on Medical Imaging with Deep Learning – Extended Abstract Track*, London, United Kingdom, July 2019b. URL https://openreview.net/forum?id=BylLiVXptV.

David Zimmerer, Jens Petersen, and Klaus Maier-Hein. **High- and Low-level image component decomposition using VAEs for improved reconstruction and anomaly detection**, November 2019c. URL http://arxiv.org/abs/1911.12161. [2022-09-01]. arXiv:1911.12161 [cs, eess, stat].

David Zimmerer, Jens Petersen, Gregor Köhler, Paul Jäger, Peter Full, Tobias Roß, Tim Adler, Annika Reinke, Lena Maier-Hein, and Klaus Maier-Hein. **Medical Out-of-Distribution Analysis Challenge**. March 2020. doi: 10.5281/ZENODO.3784230. URL https://zenodo.org/record/3784230. Publisher: Zenodo.

David Zimmerer, Peter M. Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler, Tobias Ross, Annika Reinke, Antanas Kascenas, Bjørn Sand Jensen, Alison Q. O'Neil, Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, Bernhard Kainz, Nina Shvetsova, Irina Fedulova, Dmitry V. Dylov, Baolun Yu, Jianyang Zhai, Jingtao Hu, Runxuan Si, Sihang Zhou, Siqi Wang, Xinyang Li, Xuerun Chen, Yang Zhao, Sergio Naval Marimont, Giacomo Tarroni, Victor Saase, Lena Maier-Hein, and Klaus Maier-Hein. **MOOD 2020: A public Benchmark for Out-of-Distribution Detection and Localization on medical Images**. *IEEE Transactions on Medical Imaging*, pages 1–1, 2022a. ISSN 1558-254X. doi: 10.1109/TMI.2022.3170077. Conference Name: IEEE Transactions on Medical Imaging.

David Zimmerer, Daniel Paech, Carsten Lüth, Jens Petersen, Gregor Köhler, and Klaus Maier-Hein. **Unsupervised Anomaly Detection in the Wild**. In Klaus Maier-Hein, Thomas M. Deserno, Heinz Handels, Andreas Maier, Christoph Palm, and Thomas Tolxdorff, editors, *Bildverarbeitung für die Medizin 2022*, Informatik aktuell, pages 26–31, Wiesbaden, 2022b. Springer Fachmedien. ISBN 978-3-658-36932-3. doi: 10.1007/978-3-658-36932-3_6.

Vít Škvára, Tomáš Pevný, and Václav Šmídl. **Are generative deep models for novelty detection truly better?** *arXiv:1807.05027 [cs, stat]*, July 2018. URL http://arxiv.org/abs/1807.05027. arXiv: 1807.05027.

Own Publications

The purpose of this chapter is to give an overview of my contributions and to differentiate these from a whole team efforts. In this section, all publications that I was a part of and contributed to during my Ph.D. work are listed.

First Authorships

- **David Zimmerer**, Jens Petersen, Simon AA Kohl, Klaus H Maier-Hein. A Case for the Score: Identifying Image Anomalies using Variational Autoencoder Gradients. Medical Imaging meets NeurIPS Workshop 2018.
- **David Zimmerer**, Simon Kohl, Jens Petersen, Fabian Isensee, Klaus Maier-Hein. Context-encoding Variational Autoencoder for Unsupervised Anomaly Detection. Medical Imaging with Deep Learning (MIDL) 2019.
- **David Zimmerer**, Fabian Isensee, Jens Petersen, Simon Kohl, Klaus Maier-Hein. Unsupervised anomaly localization using variational auto-encoders. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2019.
- **David Zimmerer**, Jens Petersen, Klaus Maier-Hein. High-and Low-level image component decomposition using VAEs for improved reconstruction and anomaly detection. Medical Imaging meets NeurIPS Workshop 2019.
- **David Zimmerer**, Daniel Paech, Carsten Lüth, Jens Petersen, Gregor Köhler, Klaus Maier-Hein. Unsupervised Anomaly Detection in the Wild. Bildverarbeitung für die Medizin 2022.
- **David Zimmerer**, Peter M Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler, Tobias Ross, Annika Reinke, Antanas Kascenas,

Bjørn Sand Jensen, Alison Q O'Neil, Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, Bernhard Kainz, Nina Shvetsova, Irina Fedulova, Dmitry V Dylov, Baolun Yu, Jianyang Zhai, Jingtao Hu, Runxuan Si, Sihang Zhou, Siqi Wang, Xinyang Li, Xuerun Chen, Yang Zhao, Sergio Naval Marimont, Giacomo Tarroni, Victor Saase, Lena Maier-Hein, Klaus Maier-Hein. MOOD 2020: A public Benchmark for Out-of-Distribution Detection and Localization on medical Images. IEEE Transactions on Medical Imaging.

Second Authorships

- 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, Lena Maier-Hein. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. International journal of computer assisted radiology and surgery.

- Fabian Isensee, Jens Petersen, André Klein, **David Zimmerer**, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, Klaus H Maier-Hein. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. arXiv preprint arXiv:1809.10486

- Jens Petersen, Fabian Isensee, Gregor Köhler, Paul F Jäger, **David Zimmerer**, Ulf Neuberger, Wolfgang Wick, Jürgen Debus, Sabine Heiland, Martin Bendszus, Philipp Vollmuth, Klaus H Maier-Hein. Continuous-time deep glioma growth models. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2021.

- Jens Petersen, Gregor Köhler, **David Zimmerer**, Fabian Isensee, Paul F Jäger, Klaus H Maier-Hein. GP-ConvCNP: Better generalization for conditional convolutional Neural Processes on time series data Uncertainty in Artificial Intelligence 2021.

- Maximilian Zenk, **David Zimmerer**, Fabian Isensee, Paul F Jäger, Jakob Wasserthal, Klaus Maier-Hein. Realistic Evaluation of FixMatch on Imbalanced Medical Image Classification Tasks. Bildverarbeitung für die Medizin 2022.

Editor

- Marc Aubreville, **David Zimmerer**, Mattias Heinrich. Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis. MICCAI 2021 Challenges: MIDOG 2021, MOOD 2021, and Learn2Reg 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 - October 1, 2021, Proceedings. Lecture Notes in Computer Science (LNCS, volume 13166).

Software

- **David Zimmerer**, Jens Petersen, Gregor Koehler, Jakob Wasserthal, Tim Adler, Sebastian Wirkert and André Klein. MIC-DKFZ/trixi: Alpha. November 2018. URL `https://zenodo.org/record/1495180#.Xut8EWhfh3g`.

- Fabian Isensee, Paul Jäger, Jakob Wasserthal, **David Zimmerer**, Jens Petersen, Simon Kohl, Justus Schock, André Klein, Tobias Roß, Sebastian Wirkert, Peter Neher, Stefan Dinkelacker, Gregor Köhler, Klaus Maier-Hein (2020). batchgenerators - a python framework for data augmentation. doi:10.5281/zenodo.3632567

- André Klein, Jakob Wasserthal, Mathias Greiner, **David Zimmerer**, and Klaus H. Maier-Hein. MIC-DKFZ/basic_unet_example: Release v2019.01. January 2019b.URL `https://zenodo.org/record/2552439#.Xut702hfh3g`.

- Fabian Isensee, Paul Jäger, Jakob Wasserthal, **David Zimmerer**, Jens Petersen,Simon Kohl, Justus Schock, André Klein, Tobias Roß, Sebastian Wirkert, Peter Neher, Stefan Dinkelacker, Gregor Köhler, and Klaus Maier-Hein. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. January 2020a. URL `https://zenodo.org/record/3632567#.Xut7q2hfh3g`.

Eidesstattliche Versicherung
Statutory Declaration

1. Bei der eingereichten Dissertation zu dem Thema ***Unsupervised Learning for Anomaly Detection in Medical Images*** handelt es sich um meine eigenständig erbrachte Leistung.
   *I herewith formally declare that I have written the submitted dissertation **Unsupervised Learning for Anomaly Detection in Medical Images** 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 kentlich 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, 13.09.2022

David Zimmerer