Contrastive Representations for Unsupervised Anomaly Detection and Localization

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2022

Abstract

Unsupervised anomaly detection in medical imaging aims to detect and localize arbitrary anomalies without requiring labels during training. Generally, this is achieved by learning a data distribution of normal samples and detecting anomalies as regions in the image which deviate from this distribution. In the medical imaging domain, most current state-of-the-art methods use latent variable generative models. Because such models operate directly on sample space, they tend to primarily encode low-level statistics (like pixel intensities), while having problems capturing fine semantic information within their representations. Recent work has shown that representations obtained from a feature extractor trained with a discriminative task are rich in semantic information. This, however, requires labeled datasets - a prerequisite that is often not fulfilled. We propose CRADL, a framework for unsupervised anomaly detection and localization consisting of a feature extractor trained with a contrastive pretext-task and a generative model which learns the distribution of representations. Through this, we circumvent the need for labels while still being able to fit the generative model on semantic-rich representations. We further compare the quality of these contrastive representations with representations obtained from a VAE and ceVAE in the context of anomaly localization. We evaluate CRADL on the BraTS and ISLES datasets, as well as an in-house dataset, and demonstrate state-of-the-art performance on the task of anomaly localization in our comparison with a VAE and ceVAE.

Keywords: Anomaly Detection, Self-supervised Learning, Contrastive Training.

1. Introduction

Detecting and localizing anomalies is a long-standing problem in medical image analysis. Given a specific problem and sufficient annotated training data at hand, supervised machine learning models can be extremely effective at solving this task. However, most supervised models are not explicitly designed to handle out-of-distribution data and thus might struggle to extrapolate beyond the training distribution. As a consequence, each new class of pathology or imaging modality necessitates the creation of new annotated datasets—a process that scales poorly with the large number of existing pathologies and the ever-increasing amount of image acquisition methods. In contrast, unsupervised anomaly detection promises to deliver predictions in the absence of labeled data. Thus, overcoming the need for cumbersome manual annotations, this class of methods could offer a far greater breadth of applications. In principle, this can be realized by learning a distribution of healthy samples. Images (or rather some voxels in the images) 'deviating' from this distribution are then defined as outliers. The problem of detecting these deviations can be stated as an Out-of-Distribution (OoD) detection problem, more specifically *near OoD localization* (Winkens et al., 2020) because healthy and anomalous samples commonly only differ in specific small regions. In the medical imaging domain, the current state-of-the-art methods for anomaly detection are latent variable generative models operating directly in pixel space, mainly different subtypes of and scoring methods based on Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) (Baur et al., 2021; Chen et al., 2020b; Schlegl et al., 2019). It has been shown, however, that these methods tend to focus on low-level features such as background characteristics (Nalisnick et al., 2019; Ren et al., 2019; Xiao et al., 2020; Meissen et al., 2021) and that their representations have problems capturing semantic information (Nalisnick et al., 2019; Zimmerer et al., 2018). This makes the anomaly scores of these methods heavily dependent on background statistics such as brightness and contrast. Recently, generative models trained not on pixels directly but rather on representations of supervised discriminative models have achieved state-of-the-art results on sample-level OoD detection benchmarks (Lee et al., 2018; Liang et al., 2020; Hendrycks and Gimpel, 2018; Zhang et al., 2020). Since representations of discriminative models are rich in semantic information (Zeiler and Fergus, 2014), generative models operating on their representations are, due to the inductive bias which is introduced with the discriminative task, arguably less prone to the previously mentioned problem of focusing on low-level pixel characteristics. The training of the discriminative model, however, requires a labeled dataset.

In this work we wanted to investigate if self-supervised contrastive learning can aid unsupervised anomaly localization. For this, we propose CRADL, a simple unsupervised representation-based OoD framework consisting of a feature extractor and a generative model. These semantically rich and low dimensional representations obtained with a feature extractor trained with a contrastive self-supervised task (Chen et al., 2020a) allow to fit a wide variety of generative models such as Gaussian Mixture Models and Flow-based Deep Generative Models in a very short time. We show that anomalies can be localized by back-propagating the negative log-likelihood of representations into the sample. Finally, in our experimental evaluation, we find that the representations of CRADL can yield improvements over reconstruction-based representations for anomaly localization and show competitive performance to state-of-the-art methods like the VAE and context encoding VAE. In summary, our contributions are: (1) we investigate contrastive learning for anomaly localization (2) we use SimCLR with GMMs in the OoD context, (3) we use the gradients of a “not-VAE” model (but rather a composite model) to identify anomalies, giving a valid alternative to reconstruction based approaches.

2. Method

2.1. Related Work

Contrastive Learning Contrastive Learning can be used to obtain rich semantic representations or to pretrain a model for a specific downstream task. This is done by enforcing a clustering of similar data points by pulling together positive pairs (semantically similar data points) while pushing away negative pairs (semantically different data points) in latent space. This can be achieved using the NT-Xent (the normalized temperature-scaled cross-entropy) loss as in SimCLR (Chen et al., 2020a). Several Contrastive Training methods have been proposed, mostly differing by the method used to obtain the positive and negative pairs. In the Computer Vision domain, a trend for using data augmentation transformations to obtain these positive and negative pairs has recently emerged (Chen et al., 2020a; Falcon and Cho, 2020; Hénaff et al., 2020). These methods are commonly used to pretrain the encoder of a classifier with unlabelled data, leading to a significantly reduced amount of labeled data required to train a classifier to a comparable accuracy as a model with access to more labels. The representations obtained by these methods allow for better classification than that of Autoencoder-based and many other self-supervised methods, which have historically been used as pretraining for classifiers (Chen et al., 2020a; Falcon and Cho, 2020; Hénaff et al., 2020).

Representation-based Out of Distribution detection Influential works for OoD detection were ODIN (Liang et al., 2020), the Maximum Softmax Probability (Hendrycks and Gimpel, 2018) as well as (Lee et al., 2018; Hsu et al., 2020). These methods were commonly evaluated on far OoD tasks on different datasets, which are not directly translatable to anomaly detection (Winkens et al., 2020; Ahmed and Courville, 2020). Another method that only uses the representations in the last encoder layer in combination with Flow-based Deep Generative Models was presented by Zhang et al. (2020). Ahmed and Courville (2020) demonstrated that a discriminative model which is trained with an additional self-supervised task learns semantically richer representations, leading to better OoD detection. Similarly, Winkens et al. (2020) proposed a framework consisting of a discriminative trained classifier with an additional SimCLR inspired task. They used the Mahalanobis distance to fit the distribution of the trained model features, achieving state-of-the-art OoD detection. Concurrently to our work, Sehwag et al. (2021) proposed a similar model to Winkens et al. (2020) but using a self-supervised task exclusively. While having some conceptual similarities to the approach presented here, they only consider the case of far sample-level OoD.

Generative Models for Medical Anomaly Detection The current state-of-the-art for image anomaly detection are generative methods such as VAEs and GANs. Baur et al. (2021) compared the most common methods based on their anomaly localization capabilities (pixel-level OoD). In their setting, a VAE-based iterative image restoration setting (Chen et al., 2020b) performed best across most datasets they evaluated (slightly better than the VAE-based reconstruction difference scoring). However, this iterative restoration is orthogonal to our proposed method and could be directly applied here in analogy to VAEs. As such, we chose to use the by Baur et al. (2021) recommended latent variable model as baseline: a VAE (due to its performance, simplicity, and optimization). Chen et al. (2020b) employed a GMVAE for anomaly detection with good results. But despite its architectural similarities to CRADL, we chose only to focus on VAEs as our main baseline, as its performance was shown to be inferior to a VAE while being harder to optimize (Baur et al., 2021; Meissen et al., 2021). Recently, different self-supervised approaches, such as context-encoding (Zimmerer et al., 2018), image perturbation prediction (Li et al., 2021), and Multi-Task prediction (Venkatakrishnan et al., 2020) have incorporated simple self-supervision tasks in an anomaly-localization framework. Here, we want to continue this line of research but using a two-step approach disentangle the feature learning from the distribution fit and use the recently proposed contrastive pretraining task.

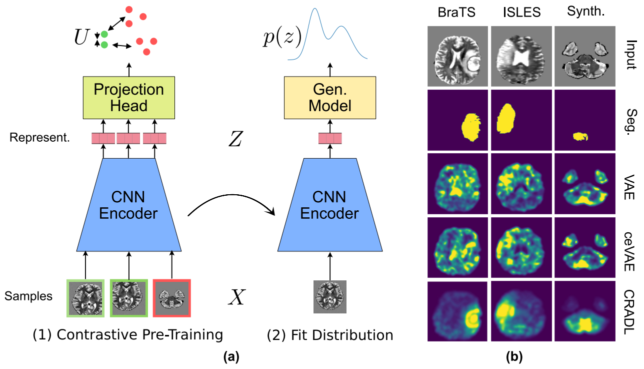

Figure 1: (a) Visualization of the fitting pipeline, from contrastive pretext (SimCLR) (1) to fitting of the generative model (2). (b) Pixel-wise scores - clamped and normalized for visual inspection.

2.2. Methodology

We propose CRADL, a method using *Contrastive Representations for unsupervised Anomaly Detection and Localization*. CRADL is comprised of two stages as shown in Figure 1. During the first stage, the encoder f, which maps from the image space X to a learned feature/representation space $Z : z = f(x)$, is trained. In the second stage, a generative model p is fitted on the representations, allowing 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 by back-propagating the gradients of the representation NLL into the sample. This approach assumes that regions with large gradients exhibit anomalies.

Contrastive Training Our contrastive pretext task is inspired by SimCLR (Chen et al., 2020a), where positive pairs are obtained by using data augmentations t drawn randomly from a set of augmentations $\mathcal{T}$. Each sample x_i in a minibatch of N examples is transformed twice, yielding two different views which make up the positive pair. The representations produced by feeding the views 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 encouraged to be similar by optimizing the NT-Xent contrastive loss:

$$l(\tilde{x}_i, \hat{x}_i) = -\log \frac{\exp(\mathrm{sim}(\tilde{u}_i, \hat{u}_i)/\tau)}{\sum_{\bar{u} \in \Lambda^-} \exp(\mathrm{sim}(\tilde{u}_i, \bar{u}))/\tau)} \tag{1}$$

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, an arbitrary generative model can be fitted on the representations. Our experiments used a Gaussian Mixture Model (GMM) as the generative model, since it is one of the simplest generative models used for anomaly detection. The probability distribution of a GMM with K components is noted in equation 2. We fit the

GMM with the Expectation-Maximization (EM) Algorithm (Dempster et al., 1977) with K being the number of components (specified before the fit).

$$p(x; \Theta) = \sum_{k=1}^{K} \mathcal{N}(x; \mu_k, \Sigma_k) \cdot \pi_k \tag{2}$$

3. Experiments

Data In our experiments, we used T2-weighted brain MRI datasets. All models were trained on a subset of the HCP dataset (Van Essen et al., 2012), which purely consists of 'normal' MRI Scans, using 894 scans split into training and validation sets. We created a synthetic anomaly dataset (similar to Zimmerer et al. (2020)) from 100 HCP separate scans (*HCP Synth.*) by rendering real-world objects into brain regions. This allows the test set to have the same original distribution (i.e., same scanner, site, ...) as the training set, with only the anomalies differing. We split HCP Synth. dataset into two distinct parts with 49 scans each, one for model development (i.e., to choose our hyperparameter settings and setting of K) and one for testing only. We also applied the same models without any changes or retraining to the BraTS-2017 (tumor segmentation) (Bakas et al., 2017) and ISLES-2015 (stroke lesion segmentation) (Maier et al., 2017) datasets to test the approach in real-world settings with different pathologies. Our BraTS-2017 and ISLES test sets consist of 266 and 20 scans, respectively, as well as validation sets comprised of 20 and 8 scans for selecting K. All datasets were preprocessed similarly, with a patient-wise z-score normalization and slice-wise resampling to a resolution of 128 x 128, followed by clipping the range of intensities from -1.5 to 1.5. For the statistics of the datasets and visual examples, we refer to the supplementary material (Suppl.).

Model We used a unified model architecture for our experiments which is based on the deep convolutional architecture from Radford et al. (2016), so our encoder solely consists of 2D-Conv-Layers and our decoder (for the VAE models) of 2D-Transposed-Conv-Layers (for more details, please refer to the Suppl.). We chose an initial feature map size of 64 and a latent dimension of 512. For the projection head, we use a simple 2-layer MLP with ReLU non-linearities, a 512 dim. hidden layer, and 256 dim. output.

Training The contrastive pretext training of the encoder is performed for 100 epochs on the HCP training set using the Adam Optimizer, a learning rate of 1e-4, Cosine Annealing (Loshchilov and Hutter, 2017), 10 Warm-up Epochs and a weight decay of 1e-6, the temperature of the contrastive loss is 0.5. The encoder for later evaluation is chosen based on the smallest loss on the HCP validation set. As transformations for generating different views for the contrastive task, we used a combination of random cropping, random scaling, random mirroring, rotations, and multiplicative brightness, and Gaussian noise. We fitted the GMM on representations of the encoder from 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 . In the supplementary, we present the performance of CRADL with relation to K. The best K value on the HCP dataset does not transfer to the other dataset settings (however, it seems consistent within the same dataset, i.e., between the validation set and test set). This behavior is probably caused by the distribution shifts between the datasets stemming from different acquisition strategies and scanners. Thus, for the evaluation of each anomalous dataset, we pick the optimal K based on a small validation set where we observe the best AUPRC score. We believe this is a clinically very reasonable approach since the fit of the GMM takes only 10 minutes on a GPU compared to more than 8 hours for the VAE and ceVAE.

Baselines We trained both the VAE and ceVAE for 100 epochs using the Adam Optimizer, a learning rate of 1e-4, and the unified architecture for both encoder and decoder on the HCP training dataset. The final models for evaluation were chosen based on the lowest loss on the HCP validation set. The transformations used during training of the VAE consisted of random scaling, random mirroring, rotations, multiplicative brightness, and Gaussian noise, which have shown clear performance improvements in our early experiments. For the ceVAE, we added random cutout transformations (Zimmerer et al., 2018). We selected the best scoring method for each evaluation dataset based on the validation sets (see Suppl.).

Metrics We are mainly interested in the localization of anomalies within the brain. As a performance measure, we used the pixel-wise AUROC and AUPRC metrics with pixel-level scores, as it is common practice. The discriminative power of the scoring function is measured using the Area Under Receiver Operator Curve (AUROC) and Area Under Precision-Recall Curve (AUPRC) due to their independence of a threshold with anomalies being defined as positives. In our setting, it is more important to detect outliers and a significant imbalance due to a larger number of healthy pixels than anomalous pixels. Hence, we emphasize the importance of the AUPRC score because it better captures the detection of anomalies.

Post-processing of Anomalies The post-processing pipeline for the pixel-level scores is identical for all methods evaluated and, based on the approach from Baur et al. (2021), restricted to only use one sample (2D): We zero out all pixel scores outside the brain region. In the next step, 2D median pooling (kernel size=5) is applied to filter out edges and single outliers. As the last step, Gaussian smoothing is applied, inspired by the finding of sparse gradients and convolutional artifacts by Zimmerer et al. (2018). Empirically, also the reconstruction-based scores of the VAE and ceVAE benefited from this step.

4. Results & Discussion

First, we analyze the discriminative power of the representations obtained with contrastive learning to that of generative models in the context of anomaly localization. Here, we want to verify whether the contrastive representations are beneficial for capturing and detecting the nuanced semantic differences between normal and anomalous regions. In the next step, we compare the anomaly localization with the VAE and ceVAE state-of-the-art methods (Baur et al., 2021; Zimmerer et al., 2019, 2018). Additionally, we compared ourselves with a Flow-based Deep Generative Models (implemented as INN with GLOW-like architecture (Kingma and Dhariwal, 2018)), where we show the results in the Supplement. For the results regarding slice-wise OoD detection, we refer to the Supplement as well.

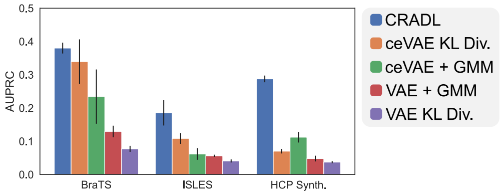

Figure 2: Performance comparison of different representations for pixel-level anomaly localization ($\bar{x} \pm \sigma_x$). We compare features from CRADL (SimCLR + GMM), a ceVAE (VAE + self-supervised context encoding task), a VAE all fitted with a GMM and for both VAE and ceVAE the gradient of the KL-Divergence (see Suppl.).

4.1. Discriminative Power of Representations for Anomaly Localization

To get an estimate of the discriminative power of representations from generative models, in particular VAE and ceVAE, we decided to fit a GMM on their representations in an identical scheme to CRADL (see Sec. 3). We believe this choice is well-founded because, in theory, the representations of a VAE should be distributed like a unimodal Gaussian. For the VAE models, we additionally compare the gradient of the KL-Divergence, since it also models the feature distribution deviations and is inherent to the model. We depict the performance of the anomaly localization methods in Fig. 2, where it becomes apparent that CRADL-based representations outperform both VAE and ceVAE based representations: for both ISLES and HCP Synth. significantly, while on BraTS the KL-Divergence of the ceVAE showed performance within 1 σ, however with lower mean performance. This strengthens the hypothesis that the self-supervised representations of CRADL carry more semantic information, enabling a better localization of fine semantic differences between anomalous and normal brain volumes. A further supporting fact is that the ceVAE, which also employs a self-supervised task, outperforms the VAE. To verify that the main benefits of CRADL stem from its representations and not purely that the GMM fits the features of SimCLR better than those of a VAE, we also conducted experiments with a Flow-based Deep Generative Models (Real NVP (Dinh et al., 2017)) showing the same trend as for the GMMs. For details on these experiments, we refer the reader to the supplementary material.

4.2. Comparison to State-of-the-Art Anomaly Localization

Here, we further compare CRADL with our re-implementations of the state-of-the-art methods VAE and ceVAE, in the context of anomaly localization. The quantitative results are shown in Table 1 and qualitative results can be seen in Fig. 1b. For the HCP Synth. dataset, CRADL obtained the best AUROC and AUPRC metrics by a large margin, followed by the VAE and the ceVAE baseline. On the BraTS dataset, the ceVAE outperforms all other methods regarding the AUPRC score and AUROC. We believe this can also be partially attributed to the domain gap between the BraTS dataset and HCP dataset (similarly with HCP and ISLES), i.e., different scanners, image quality, and the patients' overall health. This leads to a change in the pixel intensities of the overall pixel distribution. Therefore,

Table 1: Pixel-wise anomaly localization metrics for different datasets.

		CRADL	VAE	ceVAE
HCP Synth.	AUROC	**0.978±0.001**	0.951±0.001	0.921±0.004
	AUPRC	**0.288±0.010**	0.210±0.003	0.172±0.015
ISLES	AUROC	**0.898±0.003**	0.853±0.002	0.879±0.002
	AUPRC	**0.186±0.039**	0.051±0.001	0.145±0.013
BraTS	AUROC	0.942±0.001	0.925±0.001	**0.948±0.003**
	AUPRC	0.380±0.016	0.298±0.004	**0.483±0.003**

one could argue that even the brain slices without any pathology could be categorized as OoD (perhaps the anomalous samples are more OoD, but the accuracy of an OoD measurement in the high OoD regions could be considered questionable). On the ISLES-2015 dataset, CRADL shows the best performance, and again, the ceVAE delivers slightly better performance than a reconstruction-based VAE detection.

5. Discussion & Conclusion

In this work, we propose a simple framework for unsupervised Anomaly Detection and Localization based on representations obtained with a contrastive pretext task. We show that the representations obtained with this contrastive framework outperform representations obtained with latent variable generative models for anomaly localization and overall allow competitive anomaly localization performance compared to a VAE and ceVAE. An evident weakness of our approach is the hand-picked selection of the number of components (K) of the GMM, which varies between datasets. However, we suspect for a generative model trained on a representative dataset, minimizing the distribution shift between test and training data, a GMM with multiple components should lead to the best overall performance similar to the HCP Synth. dataset. We further believe that using CRADL as prior for an iterative image restoration approach might show further performance improvements (Baur et al., 2021; Chen et al., 2020b). Additionally, we are interested in a comparison with self-supervised anomaly segmentation models, which during training use self-supervised proxy anomalies (e.g. patch-interpolations (Tan et al., 2020, 2021)) and after training aim at directly segmenting more general anomalies, and investigating how the features learned by such models relate to contrastively trained models. Consequently, the often addressed question of the best self-supervised/pretraining task is open to discussion. While we have shown here that a SimCLR-like task can show improvements over context-encoding based masking, other masking tasks have recently shown great promise (Li et al., 2021; He et al., 2021) rivaling the performance of contrastive learning.

Acknowledgments

Acknowledgments withheld.

References

Faruk Ahmed and Aaron Courville. Detecting Semantic Anomalies. *AAAI*, 34(04):3154–3162, April 2020. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v34i04.5712.

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. *Sci Data*, 4(1):170117, December 2017. ISSN 2052-4463. doi: 10.1038/sdata.2017.117.

Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study. *Medical Image Analysis*, 69:101952, 01 2021. doi: 10.1016/j.media.2020.101952.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv:2002.05709 [cs, stat]*, June 2020a.

Xiaoran Chen, Suhang You, Kerem Can Tezcan, and Ender Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. *Medical Image Analysis*, 64:101713, August 2020b. ISSN 13618415. doi: 10.1016/j.media.2020.101713.

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):1–22, September 1977. ISSN 00359246. doi: 10.1111/j.2517-6161.1977.tb01600.x.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. *arXiv:1605.08803 [cs, stat]*, February 2017.

William Falcon and Kyunghyun Cho. A Framework For Contrastive Self-Supervised Learning And Designing A New Approach. *arXiv:2009.00104 [cs]*, August 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-Efficient Image Recognition with Contrastive Predictive Coding. *arXiv:1905.09272 [cs]*, July 2020.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv:1610.02136 [cs]*, October 2018.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized ODIN: Detecting Out-of-distribution Image without Learning from Out-of-distribution Data. *arXiv:2002.11297 [cs, eess]*, March 2020.

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

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. *arXiv:1807.03888 [cs, stat]*, October 2018.

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, 2021. doi: 10.1109/CVPR46437.2021.00954.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. *arXiv:1706.02690 [cs, stat]*, August 2020.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv:1608.03983 [cs, math]*, May 2017.

Oskar Maier, Bjoern H. Menze, Janina von der Gablentz, Levin Häni, Mattias P. Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen, Daan Christiaens, Francis Dutil, and et al. ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Medical Image Analysis*, 35: 250–269, January 2017. ISSN 13618415. doi: 10.1016/j.media.2016.07.009.

Felix Meissen, Georgios Kaissis, and Daniel Rueckert. Challenging current semi-supervised anomaly segmentation methods for brain mri, 2021.

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]*, February 2019.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv:1511.06434 [cs]*, January 2016.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood Ratios for Out-of-Distribution Detection. *arXiv:1906.02845 [cs, stat]*, December 2019.

Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. F-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, May 2019. ISSN 13618415. doi: 10.1016/j.media.2019.01.010.

Vikash Sehwag, Mung Chiang, and Prateek Mittal. SSD: A unified framework for self-supervised outlier detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=v5gjXpmR8J.

Jeremy Tan, Benjamin Hou, James M Batten, Huaqi Qiu, and Bernhard Kainz. Detecting outliers with foreign patch interpolation. *ArXiv*, abs/2011.04197, 2020.

Jeremy Tan, Benjamin Hou, Thomas Day, John Simpson, Daniel Rueckert, and Bernhard Kainz. Detecting outliers with poisson image interpolation. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 581–591, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87240-3.

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, and E. Yacoub. The Human Connectome Project: A data acquisition perspective. *NeuroImage*, 62(4): 2222–2231, October 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2012.02.018.

Abinav Ravi Venkatakrishnan, Seong Tae Kim, Rami Eisawy, Franz Pfister, and Nassir Navab. Self-supervised out-of-distribution detection in brain CT scans. *arXiv:2011.05428 [cs, stat]*, page 5, 2020.

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.

Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood Regret: An Out-of-Distribution Detection Score For Variational Auto-encoder. *arXiv:2003.02977 [cs, stat]*, April 2020.

Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8689, pages 818–833. Springer International Publishing, Cham, 2014. ISBN 978-3-319-10589-5 978-3-319-10590-1. doi: 10.1007/978-3-319-10590-1_53.

Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid Models for Open Set Recognition. *arXiv:2003.12506 [cs]*, August 2020.

David Zimmerer, Simon A. A. Kohl, Jens Petersen, Fabian Isensee, and Klaus H. Maier-Hein. Context-encoding Variational Autoencoder for Unsupervised Anomaly Detection. *arXiv:1812.05941 [cs, stat]*, December 2018.

David Zimmerer, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus Maier-Hein. Unsupervised Anomaly Localization using Variational Auto-Encoders. *arXiv:1907.02796 [cs, eess, stat]*, July 2019.

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.

Appendix A. Baselines

- VAE, $\hat{x} = \text{Dec}(\text{Enc}(x))$, $q(z|x) = \text{Enc}(x)$
  - $\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{Rec}}(x, \hat{x}) + \beta \cdot \mathcal{D}_{\text{KL}}(q(z|x)||p(z))$
- ceVAE, $\hat{x} = \text{Dec}(\text{Enc}(x))$, $q(z|x) = \text{Enc}(x)$
  - $\mathcal{L}_{\text{ceVAE}} = \mathcal{L}_{\text{Rec}}(x, \hat{x}) + \beta \cdot \mathcal{D}_{\text{KL}}(q(z|x)||p(z)) + \mathcal{L}_{\text{Rec}_{\text{CE}}}(x, \tilde{x})$

Table 2: Anomaly Scores

Level	Name	Formula
Samples	elbo	$s(x) = \mathcal{L}_{\text{VAE}}(x)$
Samples	kl	s(x) $= \mathcal{D}_{\text{KL}}(q(z|x)||p(z))$
Samples	rec	s(x) $= \mathcal{L}_{\text{Rec}}(x, \hat{x})$
Pixel	nll-grad (proposed)	$r_{\text{nll}-\text{grad}}(x) = \left\lvert \frac{\partial(-\log(p(f(x))))}{\partial x} \right\rvert$
Pixel	Reconstruction	$r_{\text{rec}}(x) = |x - \hat{x}|$
Pixel	KL-Div (grad) (Zimmerer et al., 2018)	$r_{\text{kl}-\text{grad}}(x) = \left\lvert \frac{\partial \mathcal{D}_{\text{KL}}(q(z|x)||p(z))}{\partial x} \right\rvert$
Pixel	Combi (Zimmerer et al., 2018)	$r_{\text{combi}}(x) = r_{\text{kl}-\text{grad}}(x) \cdot r_{\text{rec}}(x)$

Appendix B. Experiments

- Implementation: Pytorch and Pytorch Lightining.
- Hardware for Training: Single Nvidia GPUs with 12Gb VRAM (Titan XP, 2080Ti).
- Training times: SimCLR $\sim$ 12h, GMM $\sim$ 1m, RealNVP $\sim$ 30m, VAE $\sim$ 24h.
- Data augmentation:
  - SimCLR: random mirroring, random cropping, random scaling, random multiplicative brightness, additive gaussian noise
  - VAE: random scaling, random mirroring, rotations, multiplicative brightness, additive gaussian noise
  - Framework: batchgenerators (https://github.com/MIC-DKFZ/batchgenerators)

Table 3: Deep Convolutional Architecture, nf and nz are the hyperparameters for the architecture and bottleneck width. All experiments shown were conducted with nz=512 and nf=64. In the case of VAE and ceVAE nz= $512 \cdot 2$ (BN: Batch Normalization)

DC-Encoder	DC-Decoder
Input x	Input z
4 x 4 Conv$_{\text{nf}}$ Stride 2, BN, ReLU	4 x 4 Trans-Conv$_{\text{16xnf}}$ Stride 1, BN, ReLU
4 x 4 Conv$_{\text{2xnf}}$ Stride 2, BN, ReLU	4 x 4 Trans-Conv$_{\text{8xnf}}$ Stride 2 Padding 1, BN, ReLU
4 x 4 Conv$_{\text{4xnf}}$ Stride 2, BN, ReLU	4 x 4 Trans-Conv$_{\text{4xnf}}$ Stride 2 Padding 1, BN, ReLU
4 x 4 Conv$_{\text{8xnf}}$ Stride 2. BN, ReLU	4 x 4 Trans-Conv$_{\text{2xnv}}$ Stride 2 Padding 1, BN, ReLU
4 x 4 Conv$_{\text{16xnf}}$ Stride 2, BN, ReLU	4 x 4 Trans-Conv$_{\text{nf}}$ Stride 2 Padding 1, BN, ReLU
4 x 4 Conv$_{\text{nz}}$ Stride 2	4 x 4 Trans-Conv$_{\text{nc}}$ Stride 2 Padding 1, Sigmoid
Z	X

GLOW-like INN

- Architecture: We used the Multi Scale Architecture from Dinh et al. (2017) based on 7 blocks with 8 GLOW-flows (Kingma and Dhariwal, 2018)
  - block
    1. checkerboard downsampling $(c, h, w) \longrightarrow (4 \cdot c, h/2, w/2)$
    2. GLOW-flow × 8
    3. split of half of the channels for next block and half to the end
  - GLOW-flow
    1. ActNorm
    2. InvConv2dLU
    3. convolutional Couplingblock

  Training Scheme:

  - We added Gaussian noise on our samples (identical to the scheme for the VAE) as preprocessing instead of the standard GLOW preprocessing (we do this because we do not use photos in an image format)
  - Optimization criterion for the flow f: $\mathcal{L}(x) = \frac{f(x)}{2} \cdot |\det(J_f(x))|$
  - Optimization Parameters: 100 epochs on the HCP training set with Adam Optimizer and the following parameters: learning rate 1E-4, weight decay 1E-5, gradient clipping (grad_clip_val=10).
  - We chose the model for later evaluation based on the smallest loss on the HCP validation set.

Real NVP

- Architecture: We used the 8 consecutive flows of the Real NVP Architecture without splitting. Since the Model operates on the representations z (dim=512), we use Linear layers inside the coupling blocks parametrized by c_{block}=512 and $c_{\text{out}} = c_{\text{in}} = 256$.
  - Flow:
    1. Coupling Block (Linear(c_{in}, c_{block}), ReLU, Linear($c_{\text{block}}, c_{\text{out}}$))
    2. Random Permutation of Dimensions
- Training Scheme:
  - We added additional Gaussian noise (σ =1E-3) to our samples as preprocessing
  - Optimization Parameters: 60 epochs on the representations of the HCP training set with Adam Optimizer and the following parameters: learning rate 1E-3, weight decay 1E-4, gradient clipping (grad_clip_val=5 and additionally the learning rate is divided by 10 every 20 epochs
  - We chose the model for later evaluation based on the smallest loss on the HCP validation set.

Appendix C. Results

Table 4: Pixel-Wise anomaly detection metrics on test datasets: : **Values** are shown in the Results & Discussion section and selected based on the results of the best AUPRC scores on the validation set (Tab. 5)

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

Table 5: Pixel-Wise 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

Table 6: Slice-Wise anomaly detection metrics on test datasets

			HCP Synth.	HCP Synth	BraTS	BraTS	ISLES	ISLES
Pretext	Gen. Model	Score	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
VAE	GMM 1 Comp	nll	0.3851±0.0093	0.726±0.0075	0.8254±0.0079	0.8481±0.0034	0.5121±0.005	0.7093±0.0007
	GMM 2 Comp	nll	0.3652±0.0071	0.7255±0.006	0.8056±0.0072	0.8353±0.0039	0.4815±0.0061	0.6785±0.0025
	GMM 4 Comp	nll	0.4415±0.0141	0.7703±0.0074	0.827±0.0099	0.8473±0.0032	0.4896±0.0215	0.6947±0.0282
	GMM 8 Comp	nll	**0.5012**±0.0209	0.7914±0.0091	**0.8428**±0.0017	0.8546±0.0016	0.4945±0.0091	0.6999±0.0155
	Real NVP	nll	0.519±0.0239	0.7922±0.008	0.8229±0.0113	0.842±0.0103	0.5156±0.021	0.7094±0.0209
	VAE	kl	0.3583±0.0057	0.7291±0.002	0.8073±0.0069	0.8329±0.005	**0.6068**±0.0124	**0.7612**±0.003
		rec	0.4945±0.0039	0.7989±0.0009	0.8397±0.0009	0.8618±0.0008	0.5126±0.0065	0.7108±0.0026
		elbo	0.4916±0.0043	**0.7947**±0.001	0.8426±0.001	**0.8626**±0.0008	0.5188±0.0055	0.7174±0.0022
ceVAE	GMM 1 Comp	nll	0.4872±0.0252	0.7908±0.0111	0.8174±0.0147	0.8457±0.0079	0.4967±0.0191	0.6857±0.0083
	GMM 2 Comp	nll	0.4998±0.0349	0.7976±0.0127	0.7523±0.0276	0.8058±0.0084	0.3948±0.025	0.5843±0.0177
	GMM 4 Comp	nll	0.5393±0.0359	0.8084±0.0127	0.792±0.0407	0.8247±0.0233	0.4087±0.02	0.6005±0.0248
	GMM 8 Comp	nll	**0.5878**±0.033	**0.8216**±0.0097	0.7914±0.0078	0.8213±0.0103	0.3955±0.032	0.577±0.0225
	Real NVP	nll	0.6556±0.038	0.8418±0.0141	0.7247±0.0414	0.7674±0.0198	0.3938±0.0275	0.5845±0.0408
	VAE	kl	0.4091±0.0113	0.756±0.0048	0.7612±0.0063	0.8024±0.0033	0.5064±0.0071	0.6967±0.0088
		rec	0.4328±0.0037	0.7682±0.0006	**0.8434**±0.0027	**0.8638**±0.0013	**0.5467**±0.009	**0.7279**±0.0003
		elbo	0.436±0.002	0.7691±0.0009	0.843±0.0023	0.8631±0.001	0.5412±0.0082	0.7266±0.0008
CRADL	GMM 1 Comp	nll	0.6835±0.0047	0.859±0.0025	**0.811**±0.0031	**0.8203**±0.0015	**0.5436**±0.0036	**0.6916**±0.0038
	GMM 2 Comp	nll	0.7008±0.0033	0.8695±0.0023	0.7877±0.0046	0.7982±0.0031	0.4905±0.004	0.6483±0.0047
	GMM 4 Comp	nll	0.7403±0.0031	0.8786±0.0039	0.7893±0.0078	0.7987±0.0088	0.4863±0.0144	0.6471±0.0177
	GMM 8 Comp	nll	**0.7573**±0.0007	**0.8826**±0.0019	0.7848±0.0032	0.7981±0.005	0.4865±0.0063	0.6498±0.0109
	Real NVP	nll	0.7879±0.0119	0.9039±0.0043	0.6548±0.0201	0.6828±0.0239	0.424±0.0281	0.61±0.029
INN		nll	0.3655±0.0007	0.7406±0.0004	0.7872±0.0007	0.8359±0.0003	0.4619±0.0011	0.789±0.0004

Table 7: Slice-Wise anomaly detection metrics on validation datasets

			HCP Synth.	HCP Synth	BraTS	BraTS	ISLES	ISLES
Pretext	Gen. Model	Score	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
VAE	GMM 1 Comp	nll	0.4214 ± 0.0057	0.7269 ± 0.0054	0.8068 ± 0.0162	0.8463 ± 0.0032	0.7408 ± 0.0065	0.845 ± 0.0036
	GMM 2 Comp	nll	0.412 ± 0.0052	0.7269 ± 0.0053	0.7899 ± 0.0128	0.8273 ± 0.0019	0.7149 ± 0.0022	0.813 ± 0.0015
	GMM 4 Comp	nll	0.4479 ± 0.0065	0.7613 ± 0.0035	0.8117 ± 0.0147	0.8342 ± 0.0086	0.6913 ± 0.0167	0.8042 ± 0.0105
	GMM 8 Comp	nll	0.5236 ± 0.0105	0.789 ± 0.0046	0.8304 ± 0.0156	0.8506 ± 0.0064	0.7487 ± 0.0072	0.8353 ± 0.0041
	Real NVP	nll	0.5126 ± 0.0074	0.78 ± 0.0091	0.8024 ± 0.005	0.8321 ± 0.0031	0.7184 ± 0.0165	0.8149 ± 0.0043
	VAE	kl	0.4304 ± 0.0025	0.7686 ± 0.0045	0.7967 ± 0.0122	0.8438 ± 0.0048	0.6818 ± 0.0162	0.8439 ± 0.0133
		rec	0.5717 ± 0.0031	0.8172 ± 0.0007	0.7803 ± 0.0037	0.8436 ± 0.0009	0.6905 ± 0.0204	0.8251 ± 0.0095
		elbo	0.5724 ± 0.0035	0.8161 ± 0.0011	0.7871 ± 0.0026	0.8464 ± 0.001	0.7013 ± 0.0166	0.8312 ± 0.0078
ceVAE	GMM 1 Comp	nll	0.5333 ± 0.0198	0.7918 ± 0.0151	0.7935 ± 0.0051	0.8396 ± 0.0032	0.6492 ± 0.0785	0.8115 ± 0.0249
	GMM 2 Comp	nll	0.5365 ± 0.0322	0.7915 ± 0.0148	0.7345 ± 0.0302	0.7937 ± 0.012	0.5069 ± 0.1066	0.7083 ± 0.0393
	GMM 4 Comp	nll	0.5731 ± 0.037	0.8101 ± 0.0188	0.7999 ± 0.018	0.8246 ± 0.0103	0.5716 ± 0.081	0.751 ± 0.0268
	GMM 8 Comp	nll	0.6099 ± 0.0299	0.8238 ± 0.0127	0.7657 ± 0.0396	0.8107 ± 0.0199	0.5103 ± 0.1109	0.7065 ± 0.0487
	Real NVP	nll	0.6616 ± 0.0315	0.8372 ± 0.0153	0.6305 ± 0.0372	0.7233 ± 0.0324	0.5497 ± 0.1038	0.721 ± 0.0509
	VAE	kl	0.4563 ± 0.0107	0.7768 ± 0.0037	0.7722 ± 0.0069	0.8215 ± 0.0045	0.6165 ± 0.0198	0.8044 ± 0.0143
		rec	0.534 ± 0.0021	0.803 ± 0.0012	0.7973 ± 0.0068	0.8546 ± 0.0032	0.727 ± 0.0192	0.8359 ± 0.01
		elbo	0.536 ± 0.0019	0.8036 ± 0.0011	0.8016 ± 0.0035	0.857 ± 0.0011	0.7297 ± 0.0191	0.8381 ± 0.0095
CRADL	GMM 1 Comp	nll	0.7304 ± 0.0034	0.873 ± 0.0009	0.7354 ± 0.0042	0.7787 ± 0.0047	0.7331 ± 0.0071	0.8413 ± 0.0026
	GMM 2 Comp	nll	0.7396 ± 0.0031	0.878 ± 0.0006	0.7041 ± 0.0045	0.7403 ± 0.0052	0.6443 ± 0.0059	0.759 ± 0.0065
	GMM 4 Comp	nll	0.7733 ± 0.002	0.8887 ± 0.0006	0.6748 ± 0.0115	0.7252 ± 0.0043	0.641 ± 0.012	0.7579 ± 0.0153
	GMM 8 Comp	nll	0.7843 ± 0.0014	0.8926 ± 0.0001	0.6899 ± 0.0045	0.7372 ± 0.0052	0.6126 ± 0.0053	0.7319 ± 0.0077
	Real NVP	nll	0.8054 ± 0.0049	0.9 ± 0.0017	0.5675 ± 0.0094	0.6468 ± 0.0048	0.607 ± 0.0447	0.7298 ± 0.0241
INN		nll	0.4678 ± 0.0012	0.7889 ± 0.0004	0.7527 ± 0.003	0.8313 ± 0.0006	0.5982 ± 0.0003	0.8349 ± 0.0006

Appendix D. Data Description

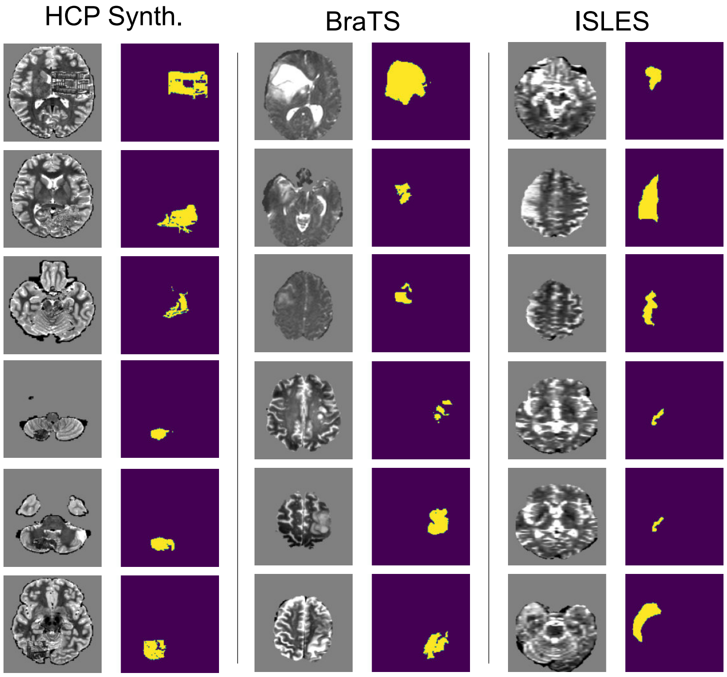

Figure 3: Visual examples of the anomalies present in our synthetically created in-house dataset HCP Synth., BraTS as well as ISLES. Shown are 6 exemplary images for each dataset with their corresponding reference annotations. Please note, that while for BraTS and ISLES the anomaly primarily differs from the normal tissue by intensity (which can be seen as a bright area), for the HCP Synth. dataset the anomalies also differ by texture and hence might present a better use-case for anomaly detection methods (Meissen et al., 2021).

Table 8: Description of statistically relevant characteristics of the anomalous datasets used in our evaluation.

Data	Type	Test	Validation
HCP Synth.	Scans	49	49
	Slices	7105	7105
	Slice Prevalence	20.54%	19.76%
	Pixel Prevalence	0.649%	0.770%
BraTS.	Scans	266	20
	Slices	35910	2700
	Slice Prevalence	49.04%	45.95%
	Pixel Prevalence	2.427%	1.926%
ISLES.	Scans	20	8
	Slices	2671	1069
	Slice Prevalence	36.61%	34.05%
	Pixel Prevalence	1.140%	1.099%

Appendix E. Prediction visualization

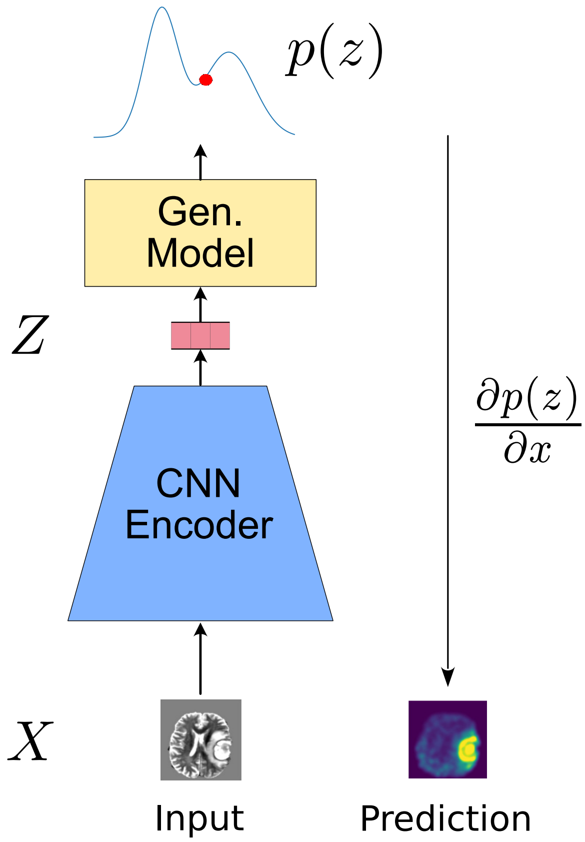

Figure 4: Visualization of the testing/ prediction phase of the model. The anomaly score/prediction is calculated as the derivative of the predicted likelihood with respect to the input image.