Segmentation Distortion: Quantifying Segmentation Uncertainty under Domain Shift via the Effects of Anomalous Activations

Jonathan Lennartz[1,2[0009−0003−0827−943X]] and Thomas Schultz[1,2[0000−0002−1200−7248]] {lennartz,schultz}@cs.uni-bonn.de

[1] University of Bonn, Bonn, Germany
[2] Lamarr Institute for Machine Learning and Artificial Intelligence

Abstract. Domain shift occurs when training U-Nets for medical image segmentation with images from one device, but applying them to images from a different device. This often reduces accuracy, and it poses a challenge for uncertainty quantification, when incorrect segmentations are produced with high confidence. Recent work proposed to detect such failure cases via anomalies in feature space: Activation patterns that deviate from those observed during training are taken as an indication that the input is not handled well by the network, and its output should not be trusted. However, such latent space distances primarily detect whether images are from different scanners, not whether they are correctly segmented. Therefore, we propose a novel segmentation distortion measure for uncertainty quantification. It is based on using an autoencoder to make activations more similar to those that were observed during training, and propagating the result through the remainder of the U-Net. We demonstrate that the extent to which this affects the segmentation correlates much more strongly with segmentation errors than distances in activation space, and that it quantifies uncertainty under domain shift better than entropy in the U-Net's output.

Keywords: Uncertainty Quantification · Image Segmentation · Anomaly Propagation.

1 Introduction

The U-Net [16] is widely used for medical image segmentation, but its results can deteriorate when changing the image acquisition device [7], even when the resulting differences in image characteristics are so subtle that a human would not be confused by them [19]. This is particularly critical when failure is silent [10], i.e., incorrect results are produced with high confidence [11].

It has been proposed that anomalous activation patterns within the network, which differ from those that were observed during training, indicate problematic inputs [15]. In the context of medical image segmentation, one such approach was recently shown to provide high accuracy for the detection of images that come from a different source [10].

Our work introduces segmentation distortion, a novel and more specific measure of segmentation uncertainty under domain shift. It is motivated by the observation that latent space distances reliably detect images from a different scanner, but do not correlate strongly with segmentation errors within a given domain, as illustrated in Figure 3. This suggests that not all anomalies have an equal effect on the final output. Our main idea is to better assess their actual effect by making the activations more similar to those that were observed during training, propagating the result through the remainder of the network, and observing how strongly this distorts the segmentation.

This yields a novel image-level uncertainty score, which is a better indicator of segmentation errors in out-of-distribution data than activation space distances or mean entropy. At the same time, it can be added to any existing U-Net, since it neither requires modification of its architecture nor its training.

2 Related Work

The core of our method is to modify activation maps so that they become more similar to those that were observed during training, and to observe the effect of this after propagating the result through the remainder of the network. We use autoencoders for this, based on the observation that the difference $r(\mathbf{x}) - \mathbf{x}$ between the reconstruction $r(\mathbf{x})$ of a regularized autoencoder and its input $\mathbf{x}$ points towards regions of high density in the training data [1].

This has previously motivated the use of autoencoders for unsupervised anomaly segmentation [2,8,4]. In contrast to these works, the autoencoder in our work acts on activation maps, not on the original image, and the anomalies we are looking for are irregular activation patterns that arise due to the domain shift, not pathological abnormalities in the image.

Conditional variational autoencoders have been integrated into the U-Net to quantify uncertainty that arises from ambiguous labels [14,3]. Their architecture and goal differ from ours, since we assume non-ambiguous training data, and aim to quantify uncertainty from domain shifts. Merging their idea with ours to account for both sources of uncertainty remains a topic for future investigation.

3 Methodology

3.1 Autoencoder Architecture, Placement, and Loss

We adapted a U-shaped autoencoder architecture which was successfully used in a recent comparative study [4] to the higher number of channels and lower resolution of activation maps as compared to images. Specifically, our encoder uses two blocks of four 3×3 kernels each with stride one, LayerNorm, and a LeakyReLU activation function. At the end of each block, we reduce spatial resolution with a stride of two. After passing through a dense bottleneck, spatial resolution is restored with a mirrored set of convolutional and upsampling layers.

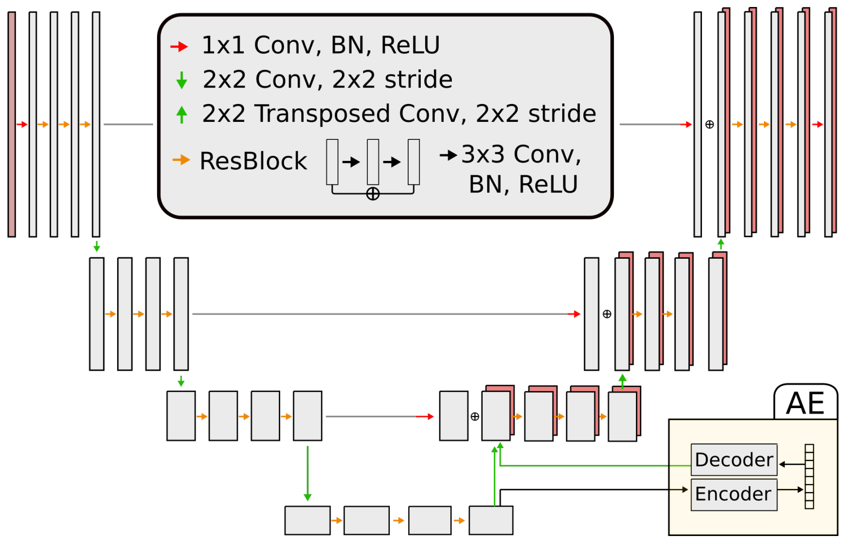

Fig. 1: Our method sends the final activation maps at the lowest resolution level of a U-Net through an autoencoder (AE) to make them more similar to activations that were observed on its training data. Our proposed Segmentation Distortion measure is based on propagating the reconstruction through the remainder of the U-Net, and quantifying the effect on the final segmentation.

Using autoencoders to make activations more similar to those observed in the training data requires regularization [1]. We tried denoising autoencoders, as well as variational autoencoders [13], but they provided slightly worse results than a standard autoencoder in our experiments. We believe that the narrow bottleneck in our architecture provides sufficient regularization by itself.

Since we want to use the difference between propagating the reconstruction $r(\mathbf{x})$ instead of $\mathbf{x}$ through the remainder of the network as an indicator of segmentation uncertainty due to domain shift, the autoencoder should reconstruct activations from the training set accurately enough so that it has a negligible effect on the segmentation. However, the autoencoder involves spatial subsampling and thus introduces a certain amount of blurring. This proved problematic when applying it to the activations that get passed through the U-Net's skip connections, whose purpose it is to preserve resolution. Therefore, we only place an autoencoder at the lowest resolution level, as indicated in Figure 1. This agrees with recent work on OOD detection in U-Nets [10].

While autoencoders are often trained with an ℓ_1 or ℓ_2 (MSE) loss, we more reliably met our goal of preserving the segmentation on the training data by introducing a loss that explicitly accounts for it. Specifically, let $U(\mathbf{I})$ denote an entire forward pass of the U-Net U on an input image $\mathbf{I}$ without the involvement of the autoencoder, while $U \circ r(\mathbf{x})$ indicates that we replace bottleneck activations

$\mathbf{x}$ with the reconstruction $r(\mathbf{x})$. We define the segmentation preservation loss on the raw logits

$$\mathcal{L}_{\text{seg}} := ||U(\mathbf{I}) - U \circ r(\mathbf{x})||_2^2 \tag{1}$$

and complement it with the established ℓ_2 loss

$$\mathcal{L}_{\text{mse}} := ||\mathbf{x} - r(\mathbf{x})||_2^2 \tag{2}$$

to induce a degree of consistency with the underlying activation space. Since in our experiments, the optimization did not benefit from an additional balancing factor, we aggregate both terms into our training objective

$$\mathcal{L} := \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{mse}} \tag{3}$$

3.2 Segmentation Distortion

We train the autoencoder so that, on in distribution (ID) images, it has almost no effect on the segmentation. Out of distribution (OOD), reconstructions diverge from the original activation. It is the goal of our segmentation distortion measure to quantify how much this affects the segmentation.

Therefore, we define segmentation distortion (SD) by averaging the squared differences of class probabilities $P(C_p|U)$ that are estimated by the U-Net U at pixel $p \in \mathcal{P}$ with and without the autoencoder, over pixels and classes $c \in \mathcal{C}$:

$$\text{SD} := \frac{1}{|\mathcal{P}|} \frac{1}{|\mathcal{C}|} \sum_{p \in \mathcal{P}} \sum_{c \in \mathcal{C}} \left[P\left(C_p = c | U(\mathbf{I})\right) - P\left(C_p = c | U \circ r(\mathbf{x})\right) \right]^2 \tag{4}$$

SD is defined similarly as the multi-class Brier score [6]. However, while the Brier score measures the agreement between probabilistic predictions and actual outcomes, SD measures the agreement between two predictions, with or without the autoencoder. In either case, a zero score indicates a perfect match.

3.3 Implementation Details

Optimizing the autoencoder with respect to $\mathcal{L}_{\text{seg}}$ requires gradient flow through the U-Net's decoder. Our implementation makes use of PyTorch's pre-forward hook functionality to compute it while keeping the weights of the U-Net intact. The U-Nets and the corresponding autoencoders were trained on identical training sets, with Adam and default parameters, until the loss converged on a respective validation set. We crop images to uniform shape to accomodate our AEs with fixed-size latent dimension. Our AEs were trained on single TITAN X GPUs for approximately three hours and exhausted the 11GB of VRAM through appropriate batching. Our code is publicly available on github.

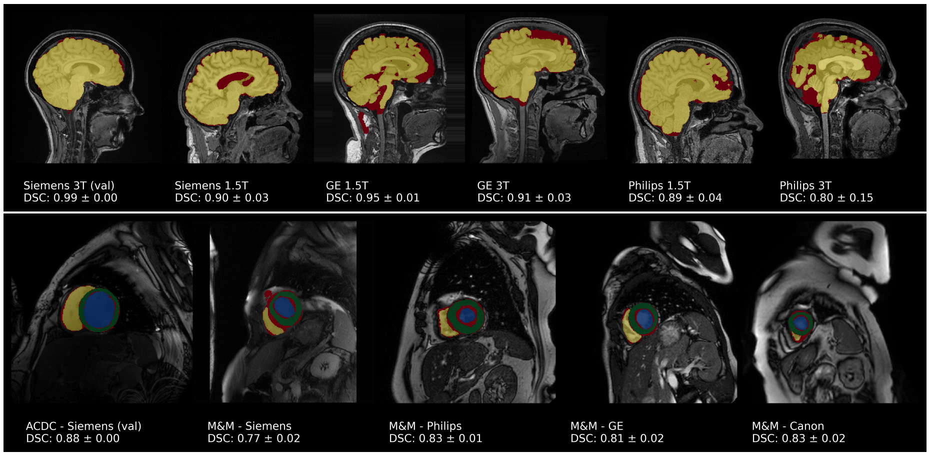

Fig. 2: Example segmentations from all domains in the CC-359 (top row) and ACDC/M&MS (bottom row) datasets, with errors highlighted in red. Numbers below the examples indicate mean Dice across the respective domain. The strong effect of domain shift on CC-359 is due to the absence of data augmentation, while errors in ACDC/M&MS arise despite data augmentation.

4 Experiments

4.1 Experimental Setup

We show results for two segmentation tasks, which are illustrated in Figure 2. The first one, Calgary-Campinas-359 (CC-359), is brain extraction in head MRI. It uses a publicly available multi-vendor, multi-field strength brain imaging dataset [18], containing T1 weighted MR scans of 359 subjects. Images are from three scanner manufacturers (GE, Philips, Siemens), each with field strengths of 1.5T and 3T. For training and evaluation, we used the brain masks that are provided with the dataset.

The second task, ACDC/M&MS, is the segmentation of left and right ventricle cavities, and left ventricle myocardium, in cardiac MRI. Here, we train on data from the Automated Cardiac Diagnosis Challenge (ACDC) that was held at MICCAI 2017 [5]. It contains images from a 1.5T and a 3T Siemens scanner. We test on data from the multi-center, multi-vendor and multi-disease cardiac segmentation (M&MS) challenge [7], which was held at MICCAI 2020. It contains MR scans from four different vendors with scans taken at different field strengths. We again use segmentation masks provided with the data. In addition to the differences between MRI scanners, images in M&MS include pathologies, which makes this dataset much more challenging than CC-359. Datasets for each task are publicly available[3].

[3] Download links for CC-359,ACDC and M&M

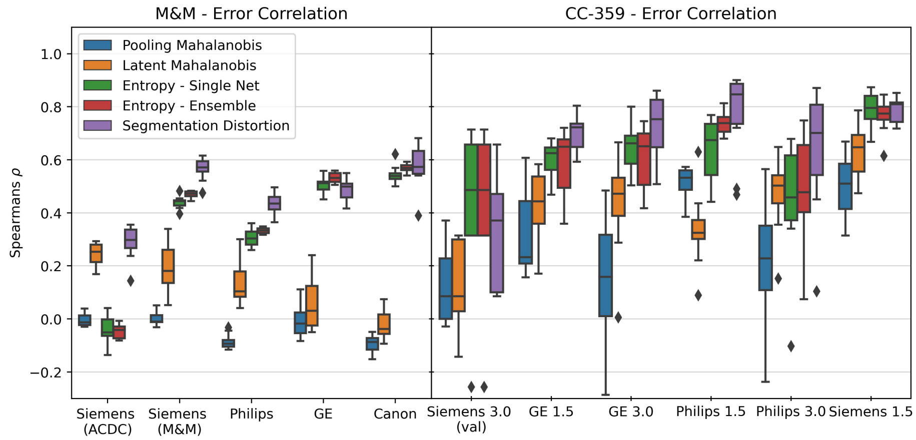

Fig. 3: Rank correlation between different uncertainty measures and (1-Dice). Mahalanobis distances of pooled activations [10], Mahalanobis distances of autoencoder latent represenations, average entropy in a single U-Net prediction as well as in averages of multiple U-Nets and the proposed Segmentation Distortion measure.

For both tasks, we train U-Nets on one of the domains, with an architecture similar to previous work on domain shift in image segmentation [17] (Figure 1). We use the Adam optimizer with default parameters and a learning rate of 1e-3, until convergence on a held-out validation set from the same domain. Similar to previous work [17], we study the effects of domain shift both with and without data augmentation during training. Specifically, results on the easier CC-359 dataset are without augmentation, while we use the same augmentations as the nnU-Net [12] when training on ACDC. For CC-359, a bottleneck dimension of 64 in our autoencoders was sufficient, while we used 128 for M&MS.

Figure 2 shows example segmentations, with errors highlighted in red, and reports the mean Dice scores across the whole dataset below the examples. To average out potential artefacts of individual training runs, we report the standard deviation of mean Dice after repeating the training 10 times. These 10 runs also underly the following results.

4.2 Correlation with Segmentation Errors

The goal of our proposed segmentation distortion measure is to identify images in which segmentation errors arise due to a domain shift. To quantify whether this goal has been met, we report rank correlations with (1-Dice), so that positive correlations will indicate a successful detection of errors. The use of rank correlation eliminates effects from any monotonic normalization or re-calibration of our uncertainty score.

Figure 3 compares segmentation distortion (SD) to two competitors: The first is a recently proposed distance-based method for uncertainty quantification under domain shift [10] that inspired our work. It is based on the same activations that are fed into our autoencoder, but pools them into low-dimensional vectors and computes the Mahalanobis distance with respect to the training distribution to quantify divergence from in-distribution activations. We therefore label it pooled Mahalanobis (PM).

The second competitor is the entropy in the model output, a widely used uncertainty measure. We compute entropy based on the per-pixel class distributions, and average the result to obtain a per-image uncertainty score. We evaluate the entropy for single U-Nets as well as ensembles of five, for which we improve uncertainty quantification at the expense of comarability to other methods by making predictions based on contributions from all members instead of a single model. As an additional point of comparison, we introduce a method that is in between SD and PM: It also uses the Mahalanobis distance, but computes it in the latent space (on the bottleneck vectors) of our autoencoder instead of pooling activations. We label it latent Mahalanobis (LM).

On both segmentation tasks, SD correlates much more strongly with segmentation errors than PM. The intermediate LM approach is usually in between the two in terms of correlation, indicating that the benefit from our method is not just due to replacing the simpler pooling strategy with an autoencoder, but that passing its reconstruction through the remainder of the U-Net is crucial for our method's effectiveness.

In almost all cases, SD also shows a stronger correlation with segmentation error than both entropy based approaches, which do not specifically account for domain shift. Unfortunately, even though propagating the output of our autoencoder through the remaining network is useful to detect segmentation errors, it is not effective for correcting them: In all cases, we found that the downstream segmentation accuracy in terms of Dice was slightly reduced.

4.3 Out-of-Distribution Detection

The distance-based method PM was initially introduced for out-of-distribution (OOD) detection, i.e., detecting whether a given image has been taken with the same device as the images that were used for training [9]. To put the weak correlation with segmentation errors that was observed in Figure 3 into perspective, we will demonstrate that, compared to the above-described alternatives, it is highly successful at this task.

For this purpose, we calibrate each of the four uncertainty scores by finding a corresponding threshold such that 95% of an in-domain validation set is classified as such. We then derive detection accuracies by applying the same threshold to each target domain for all independently trained U-Nets. For the M&MS dataset, results are displayed in Figure 4 (left). Since all methods achieved near-perfect accuracy on CC-359, those results are not presented as a figure.

This experiment confirms the excellent results for OOD detection that were reported previously for the PM method [9]. In contrast, our Segmentation Dis-

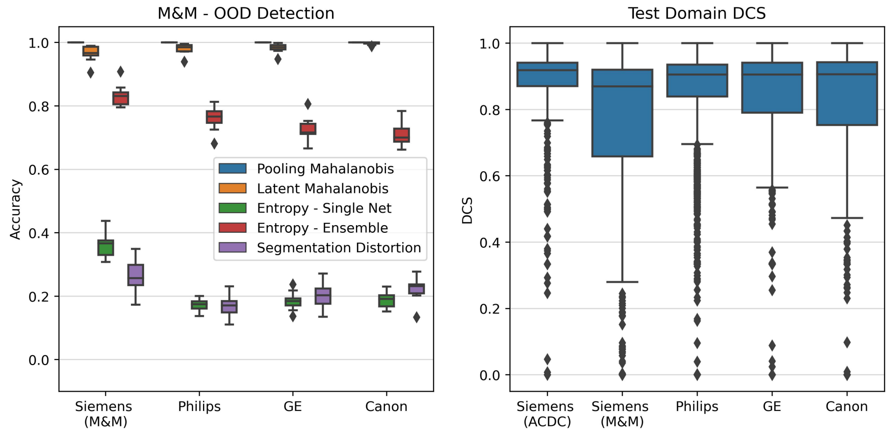

Fig. 4: Left: A comparison of the accuracy that the same four methods as in Figure 3 achieve on an OOD detection task. Right: The distributions of Dice scores on images from the ACDC source domain (Siemens val) and the four M&Ms target domains overlap greatly.

tortion measure has not been designed for OOD detection, and is not effective for that task. Similarly, mean entropy in the segmentation map is not a reliable indicator for OOD inputs.

Of course, a method that successfully solves OOD detection can be used to reject OOD inputs, and thereby avoid silent failures that arise due to domain shifts. However, it can be seen from Figure 4 (right) that this comes at the cost of filtering out many images that would be segmented reasonably well. This figure shows the distributions of Dice scores on all domains. It illustrates that, even though scanner changes go along with an increased risk for inaccurate segmentation, many images from other scanners are still segmented as well as those from the one that was used for training. The latter is marked as Siemens (val), which refers to the validation subset of ACDC, and differs from the Siemens domain in M&MS. Indeed, it is a known limitation of the PM method, which our Segmentation Distortion seeks to overcome, that "many OOD cases for which the model did produce adequate segmentation were deemed highly uncertain" [10].

5 Discussion and Conclusion

In this work, we introduced Segmentation Distortion as a novel approach for the quantification of segmentation uncertainty under domain shift. It is based on using an autoencoder to modify activations in a U-Net so that they become more similar to activations observed during training, and quantifying the effect of this on the final segmentation result.

Experiments on two different datasets, which we re-ran multiple times to assess the variability in our results, confirm that our method more specifically detects erroneous segmentations than anomaly scores that are based on latent space distances [15,10]. They also indicate a benefit compared to mean entropy, which does not explicitly account for domain shift. This was achieved on pre-trained U-Nets, without constraining their architecture or having to interfere with their training, and held whether or not data augmentation had been used.

Finally, we observed that different techniques for uncertainty quantification under domain shift have different strengths, and we argue that they map to different use cases. If safety is a primary concern, reliable OOD detection should provide the strongest protection against the risk of silent failure, at the cost of excluding inputs that would be adequately processed. On the other hand, a stronger correlation with segmentation errors, as it is afforded by our approach, could be helpful to prioritize cases for proofreading, or to select cases that should be annotated to prepare training data for supervised domain adaptation.

References

1. Alain, G., Bengio, Y.: What regularized auto-encoders learn from the data-generating distribution. Journal of Machine Learning Research **15**, 3743–3773 (2014)
2. An, J., Cho, S.: Variational autoencoder based anomaly detection using reconstruction probability. In: Special Lect. IE. vol. 2, pp. 1–18 (2015)
3. Baumgartner, C., Tezcan, K., Chaitanya, K., Hötker, A., Muehlematter, U., Schawkat, K., Becker, A., Donati, O., Konukoglu, E.: PHiSeg: capturing uncertainty in medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). LNCS, vol. 11765, pp. 119–127 (2019)
4. Baur, C., Denner, S., Wiestler, B., Navab, N., Albarqouni, S.: Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study. Medical Image Analysis **69**, 101952 (2021)
5. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P., Cetin, I., Lekadir, K., Camara, O., Ballester, M.Á.G., Sanroma, G., Napel, S., Petersen, S.E., Tziritas, G., Grinias, E., Khened, M., Varghese, A., Krishnamurthi, G., Rohé, M., Pennec, X., Sermesant, M., Isensee, F., Jaeger, P., Maier-Hein, K.H., Full, P.M., Wolf, I., Engelhardt, S., Baumgartner, C.F., Koch, L.M., Wolterink, J.M., Isgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., Jodoin, P.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Trans. Medical Imaging **37**(11), 2514–2525 (2018). `https://doi.org/10.1109/TMI.2018.2837502`
6. Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthly Weather Review **78**(1), 1–3 (1950)
7. Campello, V.M., Gkontra, P., Izquierdo, C., Martin-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K.H., Zhang, Y., He, Z., Ma, J., Parreño, M., Albiol, A., Kong, F., Shadden, S.C., Acero, J.C., Sundaresan, V., Saber, M., Alattar, M.A., Li, H., Menze, B.H., Khader, F., Haarburger, C., Scannell, C.M., Veta, M., Carscadden, A., Punithakumar, K., Liu, X., Tsaftaris, S.A., Huang, X., Yang, X., Li, L., Zhuang, X., Viladés, D., Descalzo, M.L., Guala, A., Mura, L.L., Friedrich, M.G., Garg, R., Lebel, J., Henriques, F., Karakas, M., Çavus, E., Petersen, S.E., Escalera, S., Segui,

S., Palomares, J.F.R., Lekadir, K.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. IEEE Trans. Medical Imaging **40**(12), 3543–3554 (2021)

8. Chen, X., Konukoglu, E.: Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. In: Medical Imaging with Deep Learning (MIDL) (2018)
9. Gonzalez, C., Gotkowski, K., Bucher, A., Fischbach, R., Kaltenborn, I., Mukhopadhyay, A.: Detecting when pre-trained nnU-Net models fail silently for covid-19 lung lesion segmentation. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 304–314. Springer (2021)
10. González, C., Gotkowski, K., Fuchs, M., Bucher, A., Dadras, A., Fischbach, R., Kaltenborn, I.J., Mukhopadhyay, A.: Distance-based detection of out-of-distribution silent failures for covid-19 lung lesion segmentation. Medical Image Analysis **82**, 102596 (2022). `https://doi.org/10.1016/j.media.2022.102596`
11. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (eds.) Proc. Int'l Conf. on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330 (2017)
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
13. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: Bengio, Y., LeCun, Y. (eds.) Int'l Conf. on Learning Representations (ICLR) (2014)
14. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J., Maier-Hein, K., Eslami, S., Jimenez, D., Ronneberger, O.: A probabilistic U-Net for segmentation of ambiguous images. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 6965–6975 (2018)
15. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems (NeurIPS). pp. 7167–7177 (2018)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 234–241. Springer, Cham (2015)
17. Shirokikh, B., Zakazov, I., Chernyavskiy, A., Fedulova, I., Belyaev, M.: First U-Net layers contain more domain specific information than the last ones. In: Albarqouni, S., Bakas, S., Kamnitsas, K., Cardoso, M.J., Landman, B., Li, W., Milletari, F., Rieke, N., Roth, H., Xu, D., Xu, Z. (eds.) Proc. Domain Adaptation and Representation Transfer (DART). pp. 117–126. Springer (2020)
18. Souza, R., Lucena, O., Garrafa, J., Gobbi, D.G., Salluzzi, M., Appenzeller, S., Rittner, L., Frayne, R., de Alencar Lotufo, R.: An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement. NeuroImage **170**, 482–494 (2018)
19. Zakazov, I., Shirokikh, B., Chernyavskiy, A., Belyaev, M.: Anatomy of domain shift impact on U-Net layers in MRI segmentation. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 211–220. Springer (2021)