Beyond Heatmaps: A Comparative Analysis of Metrics for Anomaly Localization in Medical Images

David Zimmerer[1] and Klaus Maier-Hein[1]

German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany

Abstract. An assumption-free, disease-agnostic pathology detector and segmentor is often regarded as one of the holy grails in medical image analysis. Building on this concept, un- or weakly supervised anomaly localization approaches have gained popularity. These methods aim to model normal or healthy samples using data and then detect deviations (i.e., abnormalities). However, as this is an emerging field situated between image segmentation and out-of-distribution detection, most approaches have adapted their evaluation setups and metrics from either of these areas. Consequently, they may have overlooked peculiarities inherent to anomaly localization. In this paper, we revisit the anomaly localization setup, analyze commonly used metrics, introduce alternative metrics inspired by instance segmentation, and compare these metrics across various settings and algorithms. We contend that the choice of metric is use-case dependent, but the SoftInstanceIoU and other object-based metrics show significant promise for future applications.

Keywords: Anomaly Localization, Anomaly Detection, Metrics.

1 Introduction

Accurate detection and localization of pathologies in medical images is a cornerstone of effective diagnosis and treatment. The ability to identify and precisely locate anomalies is crucial for early intervention, which can significantly improve patient outcomes. Traditionally, this task has relied on extensive disease-specific labeling and expert annotation, which are time-consuming and labor-intensive processes. The emergence of unsupervised and weakly-supervised anomaly localization techniques has shown great potential in revolutionizing this aspect of medical imaging. These approaches offer the capability to identify abnormalities without the need for extensive disease-specific labeling, thereby reducing the dependency on large annotated datasets [19]. By modeling the characteristics of normal, healthy tissue (or explicitly trying to model and generalize beyond preselected abnormalities), these methods can facilitate the detection of deviations that indicate potential pathologies. Historically, most anomaly localization methods have employed heatmaps to visualize the likelihood of anomalies within

an image. These heatmaps provide a spatial representation of the areas most likely to contain anomalies, which can be extremely useful for clinicians. However, evaluating the effectiveness of these heatmaps requires specialized metrics that can accurately capture their performance. As the field of anomaly localization is rapidly developing, it has often borrowed evaluation metrics from related domains such as image segmentation and out-of-distribution (OoD) detection [1, 3, 4, 7, 9, 12, 13, 15, 18, 20]. These metrics, while useful, may not fully capture the unique challenges and requirements specific to the anomaly localization task. Metrics borrowed from other domains might not account for unique factors, potentially leading to sub-optimal performance assessments. Therefore, there is a need for the development and adoption of evaluation metrics that are specifically designed to address the intricacies of anomaly localization in medical imaging. While current practices in anomaly localization have made significant strides, selecting the most appropriate evaluation metrics remains a critical area for improvement. Our goal is to address this gap by analyzing the properties of various metrics from related fields and comparing their performance across multiple datasets, as well as using a human judge as a reference. Additionally, we aim to refine these metrics to better combine the predicted heatmaps with an object-based evaluation approach, enhancing their relevance and accuracy in the context of anomaly localization.

2 Anomaly Localization Evaluation

2.1 Level of Evaluation

Evaluating anomaly detection in medical imaging requires careful consideration of different levels of analysis which cover different detailed aspects of model performance (see Fig. 1). Here, we discuss four levels of evaluation: dataset level, sample level, object level, and slice level.

Dataset level At the dataset level, evaluation is performed by aggregating all pixels across the entire dataset into one large set and calculating a single score for the entire dataset. This approach treats the dataset as one sample, providing a broad measure of performance but potentially overlooking individual sample nuances.

Sample level Sample-level evaluation involves computing scores for individual images or volumes, followed by aggregating (e.g. averaging) these scores across the entire test set. This approach mirrors the typical practice in image segmentation, where a single score is calculated and averaged per sample.

Object level At the object level, evaluation focuses on the abnormal structures within the images. Scores are calculated for each detected object and then aggregated across all objects in the dataset. This approach can address issues like the presence of multiple abnormal objects within an image and helps to avoid

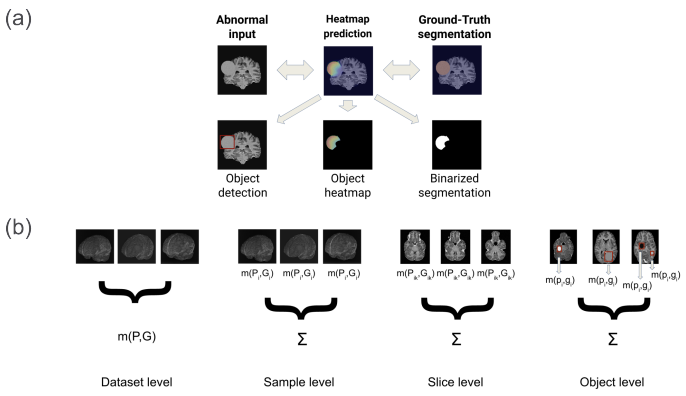

Fig. 1: (a) Overview of Anomaly localization problem formulation. (b) Different levels of metric m calculation given a prediction P and the ground-truth G.

the bias of standard segmentation metrics that prioritize larger objects while neglecting smaller ones. However, it can ignore scans where no objects are defined, thus potentially overlooking normal samples.

Slice level Slice-level evaluation involves splitting 3D volumes into 2D slices, with individual scores computed for each slice before aggregation. This approach aligns with the design of many 2D models, offering additional advantages in computational efficiency and resource utilization.

2.2 Metrics

To effectively evaluate the performance of anomaly localization models, a range of metrics are employed, spanning various domains:

Segmentation metrics Often employed metrics here are DSC (Dice Similarity Coefficient) and IoU (Intersection over Union).

- DSC score: Quantifies the spatial overlap between the predicted anomaly region P and the ground truth G for a sample x. A DSC score of 1 indicates perfect overlap, while a score of 0 means no overlap. Often the maximal achievable DSC score is calculated by choosing the best binarization threshold t to turn the continuous "heatmap" prediction into a binary segmentation:

$$\lceil \mathrm{DSC} \rceil = \max_t \left(\frac{2 \cdot \sum_x (P(x) \geq t) \cdot G(x)}{\sum_x (P(x) \geq t) + \sum_x G(x)} \right). \tag{1}$$

- IoU (Intersection over Union): Similarly to the DSC score, the IoU measures the overlap between the predicted and ground-truth anomaly masks:

$$\mathrm{IoU}(x) = \frac{|P(x) \cap G(x)|}{|P(x) \cup G(x)|}. \tag{2}$$

While commonly used in segmentation tasks [6], DSC and IoU rely on binarized predictions. This necessitates thresholding heatmaps, a process that introduces potential bias via threshold selection and can lead to undefined scores when ground-truth segmentations are sparse – a frequent occurrence in anomaly localization. Here we often choose the "Best" threshold or calculate the performance over all thresholds, as indicated by "AUC".

Out-of-distribution detection metrics :

- AP (Average Precision): Calculated from the area under the precision-recall curve, which plots precision P (how many of the predicted anomalies are actually anomalies) against recall R (how many of the true anomalies are detected). This measures a model's ability to rank anomalies higher than normal examples.
$$\text{AP} = \sum_{n=1}^{N} (R_n - R_{n-1}) P_n. \tag{3}$$
- AUROC (Area Under the Receiver Operating Characteristic): Similarly to AP, it summarizes the performance of a model at different classification thresholds (utilizing the false positive rate FPR and the true positive rate TPR). High AUROC indicates a model's capability to differentiate anomalies and normal samples.
$$\text{AUROC} = \sum_{i=1}^{N-1} \frac{(FPR_{i+1} - FPR_i) \cdot (TPR_{i+1} + TPR_i)}{2}. \tag{4}$$

Unlike segmentation metrics, ranking-based metrics like AUROC and AP directly handle heatmaps without requiring thresholding or relying on exact prediction values. However, they still yield undefined scores for data samples without ground-truth labels. Although often addressed by combining labeled and unlabeled data (e.g., evaluating metrics throughout the entire dataset ("Dataset level"), or using batch-wise calculations as in [18]), this approach can overemphasize larger, potentially easier-to-detect anomalies ([8, 14]).

Object-detection and instance segmentation metrics transition from basic overlap measurements to object-centric anomaly localization. This requires defining distinct objects in ground-truth labels (often via connected-component analysis). Key metrics include Instance IoU and Center-point Distance. A formulation of the InstanceIoU can be:

$$\text{InstanceIoU}(t) = \sum_{i \in \text{Objects}} \frac{|P(x_i) \geq t \cap G(x_i)|}{|P(x_i) \geq t \cup G(x_i)|}, \tag{5}$$

which can be calculated for all objects in a sample i.e., sample-level or all objects in that dataset i.e., dataset-level. The Center-point distance for an object i can be formulated as:

$$\text{CPD}(i) = ||\mu_{P_i} - \mu_{G_i}||_2, \tag{6}$$

where μ_{P_i} and μ_{G_i} are the center points (or center of mass) for $P(x_i)$ and $G(x_i)$ respectively. Additionally, these metrics can not just be aggregated using mean, but also median or by applying a threshold (e.g., IoU > 0.5) to classify TPs, FPs, and FNs at the object level, allowing the calculation of derived metrics like F1 score. However, binarization of predictions remains necessary and also objects have to be identified (e.g., using connected-component analysis). For this work, we adapt the Center Distance metric: an object's heatmap center point lying within the convex hull of a labeled object constitutes a TP, which we term "Center Matching".

Anomaly Localization Metrics To harness the strengths of instance segmentation metrics while avoiding the drawbacks of binarization thresholds, we introduce SoftInstanceIoU (inspired by Soft DSC [1]). This modified Instance IoU integrates continuous anomaly scores for a more nuanced assessment of predicted anomaly confidence:

$$\text{SoftInstanceIoU}(x) = \frac{\sum_{i \in \text{Object} \cup \text{Background}} \alpha \hat{P}(x)_i \hat{G}(x)_i}{\sum_{i \in \text{Object} \cup \text{Background}} (0.5 \hat{P}(x)_i + (1-\alpha) \hat{G}(x)_i)}, \quad (7)$$

where *Background* refers to all pixels not labeled as objects, i indexes the objects in the sample x, $\hat{P}(x)_i$ and $\hat{G}(x)_i$ refer to the prediction and ground-truth segmentation with all pixels masked out that do not belong to the background or the object i and α is a weighting factor to balance under- and over-segmentation (with setting the target for background to 0).

3 Experiments & Results

3.1 Understanding Metric through Controlled Experiments

To analyze the behavior of different evaluation metrics for anomaly localization, we conducted a controlled experiment using 50 samples containing circular objects and the respective (perfect) segmentations. These samples were systematically altered in the following ways:

(a) Adding small objects: We introduced small, correctly detected objects while reducing the overall segmentation size (approximately maintaining the total segmented pixel count). This simulates improved object detection. **(b) Varying segmentation size**: We altered the size of the segmentations to observe the impact of over-/under-segmentation on metric performance. **(c) Introducing false positives (FPs)**: We added segmentations that were not present in the ground truth (false detections) to evaluate metric sensitivity to these errors. **(d) Introducing missed instances (FNs)**: We removed segmentations present in the ground truth to simulate by a model not detected (missed detections) to assess how metrics respond to these misses. **(e) Including empty samples**: We added samples devoid of any labels or predictions ("empty samples") to analyze metric behavior in such scenarios.

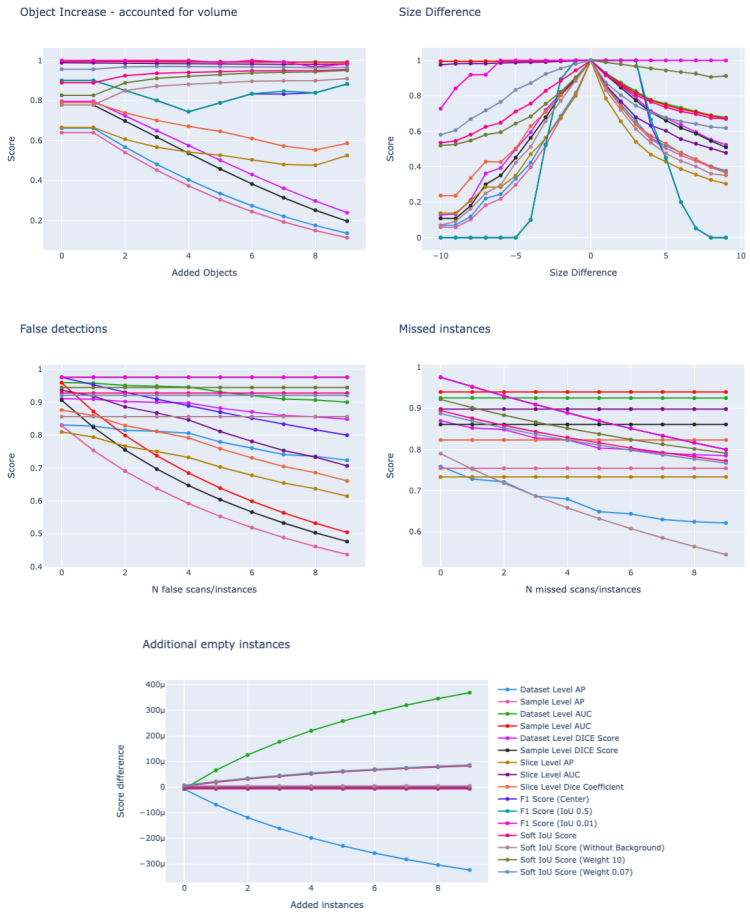

Fig. 2: Analysis of the impact of various segmentation alterations (size, presence/absence of objects) on the performance of different anomaly localization metrics.

The experiment's findings, presented in Figure 2, reveal distinct behaviors for different metric categories:

(a) Adding small objects: Surprisingly, metrics like AP and AUROC decreased despite improved object detection. Soft IoU metrics, as expected, increased with better detection. Object-based metrics displayed some variation but remained relatively stable. **(b) Segmentation size variations**: Most metrics exhibited a peak-shaped response as segmentation size changed, indicating optimal performance at a specific size. However, F1 metrics based on center distance and 0.5 IoU thresholds showed weaker responses, while Dataset AP and F1 with a 0.01 IoU threshold were nearly constant on one side. **(c) False positives (FPs)**: Sensitivity to false positives varied across metrics. AP, AUROC, and F1 (center) were highly sensitive, while Soft IoU was less affected. **(d) Missed instances (FNs)**: F1 metrics, Soft IoU, Dataset Level AP, and DSC displayed the greatest sensitivity to missed instances. Sample-level and slice-level metrics, however, failed to capture this performance change. **(e) Empty samples**: Only F1 metrics and Soft IoU showed improvement when completely normal samples were added. Dataset AP surprisingly decreased, while other metrics remained insensitive as intended for such scenarios.

This controlled experiment sheds light on the strengths and weaknesses of various evaluation metrics for anomaly localization. Selecting appropriate metrics based on the desired performance characteristics is crucial for accurate assessment in this domain.

3.2 Anomaly Localization Benchmark with Human Agreement

In this section, we evaluate the performance of various anomaly detection algorithms on diverse metrics and assess their alignment with human assessment in a more realistic setting (Fig. 3). Human raters ($n = 1$) were presented with the image, ground truth (GT) segmentation, and blinded predictions by each algorithm for each GT object/anomaly. The evaluation focused on two aspects: (1) whether the algorithm sufficiently detected the anomaly and (2) which algorithms achieved the best detection (allowing for multiple if heatmaps were subjectively similar, or none if all failed). Notably, both rating schemes yielded the same ranking, hence only the "Human rater" column is shown.

It is important to acknowledge that interpreting heatmaps can be challenging, so the evaluation solely considered object detection performance within segmented slices. However, this restriction might have downplayed the impact of false positives outside segmented regions.

For the first dataset (CamCAM [17]), seven algorithms were tested. We introduced artificial colored spheres (one large, four small) into 50% of the test images. The framework, hyperparameters, and training schedules remained consistent with [7], however, we added a "Fake" algorithm which only returned the absolute intensity values as anomaly score, inspired by [11]. For the second and third datasets (MOOD brain and abdominal [18]), we evaluated the winning algorithms from the MOOD challenge [2, 5, 10, 16] on their respective datasets, again introducing colored sphere anomalies in 50% of test images.

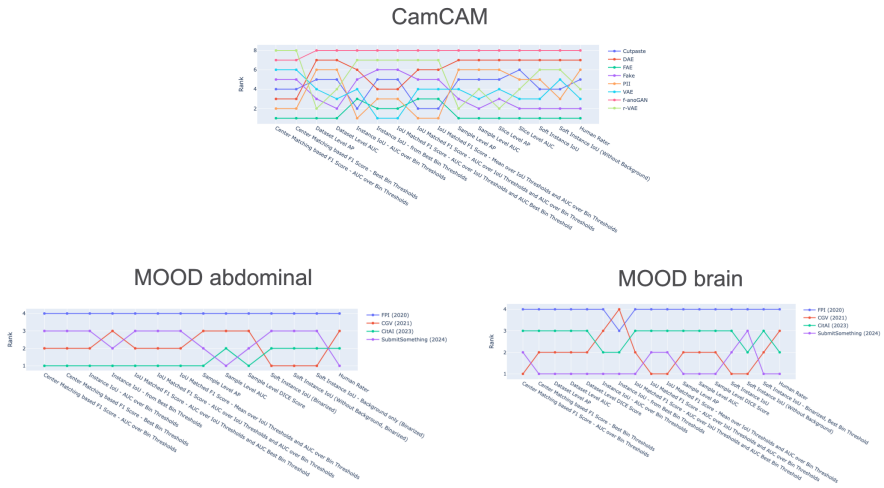

Fig. 3: Anomaly Localization Benchmark Evaluation of different methods and metrics on the CamCAN, MOOD abdominal, and MOOD brain dataset and comparison with a human rater.

Here, see Fig. 3, metrics like Sample-level AUC, SoftInstanceIoU, and F1-based metrics closely mirrored human judgment on anomaly detection. However, the focus on segmented slices in human evaluation might have underestimated the impact of false positives outside these regions. Additionally, as noted by the human raters, only FAE and Fake algorithms showed significant performance in detecting the artificial anomalies on the CamCAM dataset.

Sample-level and object-level metrics offer additional advantages beyond their alignment with human assessment. One such benefit is the ability to assess ranking stability through bootstrapping. Fig. 4 (a) demonstrates this for the SoftInstanceIoU metric applied to the CamCAM dataset. The figure reveals a highly stable ranking in this specific case. Furthermore, object-based met-

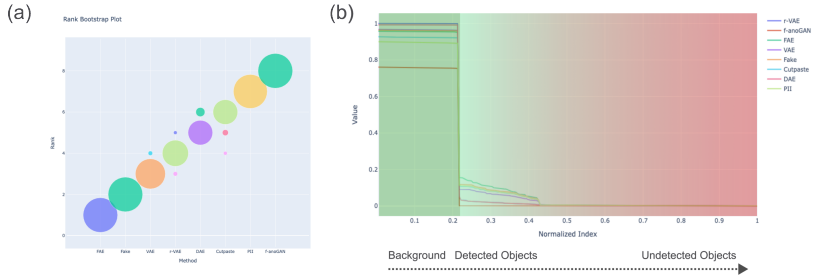

Fig. 4: (a) Stability analysis of the CamCAN dataset using bootstrapping. (b) An AUC-like plot of the SoftInstanceIoU for the CamCAN dataset.

rics provide a continuous score for each object, enabling visualization similar to AUROC and AP curves. This approach offers valuable insights into model performance across different object types. As shown in Figure 4 (b), where for each object (and the background) in the dataset the (sorted) SoftInstanceIoU scores are plotted next to each other, the initial portion of the plot reflects the background segmentation quality. This highlights an advantage of SoftInstanceIoU - its ability to incorporate background evaluation. Subsequently, the plot transitions smoothly to depict performance on detected objects, with scores decreasing for poorly or undetected objects.

4 Discussion & Conclusion

Our investigation into anomaly localization metrics yielded valuable insights into their strengths and weaknesses. The experiments demonstrated that different metrics emphasize distinct aspects of performance, highlighting the importance of selecting metrics that align with the specific goals of the anomaly detection task. Soft Instance IoU emerged as a particularly promising metric. It exhibited desirable characteristics, including resilience to false positives while remaining sensitive to missed detections. This balance makes it suitable for scenarios where both precise localization and overall anomaly capture are crucial. F1-based metrics also displayed favorable properties. Their sensitivity to both missed detections and false positives provides a comprehensive assessment of performance. However, further investigation is needed to determine the optimal threshold selection for these metrics in the context of anomaly localization. Here, it is important to acknowledge the limitations of our human evaluation approach. While it provided valuable insights, the focus on segmented slices might have underestimated the impact of false positives outside these regions. Future studies could explore alternative evaluation strategies that encompass the entire image to obtain a more holistic assessment. Our findings highlight the importance of meticulous metric selection, particularly when working with datasets containing both healthy and pathological cases. The observed sensitivity of metrics like Dataset AP to the introduction of normal samples underscores this point. Furthermore, some metrics, like sample-level Dice Similarity Coefficient (DSC), may become entirely inapplicable in such settings. To comprehensively evaluate performance in these diverse scenarios, researchers might consider employing a combination of metrics, each capturing distinct aspects of anomaly detection. This approach becomes even more critical when dealing with complex anomaly detection tasks or highly variable datasets. In conclusion, our work emphasizes the critical role of metric selection in anomaly localization. By carefully considering the desired performance characteristics and potential limitations of available metrics, researchers can achieve a more accurate and informative evaluation of their models. This paves the way for the development and refinement of anomaly localization techniques with the potential to improve medical diagnosis and patient care significantly.

Bibliography

[1] Ahmed, F., Courville, A.: Detecting semantic anomalies. ArXiv **abs/1908.04388** (Aug 2019), https://arxiv.org/abs/1908.04388

[2] Baugh, M., Tan, J., Müller, J.P., Dombrowski, M., Batten, J., Kainz, B.: Many tasks make light work: Learning to localise medical anomalies from multiple synthetic tasks (Jul 2023), http://arxiv.org/abs/2307.00899, arXiv:2307.00899 [cs]

[3] Baur, C., Wiestler, B., Albarqouni, S., Navab, N.: Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images. arXiv:1804.04488 [cs] (Apr 2018), http://arxiv.org/abs/1804.04488, arXiv: 1804.04488

[4] Chen, X., Pawlowski, N., Rajchl, M., Glocker, B., Konukoglu, E.: Deep Generative Models in the Real-World: An Open Challenge from Medical Imaging. CoRR **abs/1806.05452** (2018)

[5] Cho, J., Kang, I., Park, J.: Self-supervised 3D Out-of-Distribution Detection via Pseudoanomaly Generation. In: Aubreville, M., Zimmerer, D., Heinrich, M. (eds.) Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis. pp. 95–103. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-030-97281-3_15

[6] Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., Maier-Hein, K.H.: nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. arXiv:1809.10486 [cs] (Sep 2018), http://arxiv.org/abs/1809.10486, arXiv: 1809.10486

[7] Lagogiannis, I., Meissen, F., Kaissis, G., Rueckert, D.: Unsupervised Pathology Detection: A Deep Dive Into the State of the Art. IEEE Transactions on Medical Imaging pp. 1–1 (2023). https://doi.org/10.1109/TMI.2023.3298093, http://arxiv.org/abs/2303.00609, arXiv:2303.00609 [cs]

[8] Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Büttner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M.A., Wiesenfarth, M., Kavur, A.E., Sudre, C.H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, A.T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Blaschko, M., Cardoso, M.J., Cheplygina, V., Cimini, B.A., Collins, G.S., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Haase, R., Hashimoto, D.A., Hoffman, M.M., Huisman, M., Jannin, P., Kahn, C.E., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kenngott, H., Kofler, F., Kopp-Schneider, A., Kreshuk, A., Kurc, T., Landman, B.A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A.L., Mattson, P., Meijering, E., Menze, B., Moons, K.G.M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J., Sánchez, C.I., Shetty, S., van Smeden,

M., Summers, R.M., Taha, A.A., Tiulpin, A., Tsaftaris, S.A., Van Calster, B., Varoquaux, G., Jäger, P.F.: Metrics reloaded: Recommendations for image analysis validation (Jun 2023), http://arxiv.org/abs/2206.01653, arXiv:2206.01653 [cs]

[9] Marimont, S.N., Tarroni, G.: Anomaly detection through latent space restoration using vector-quantized variational autoencoders. arXiv:2012.06765 [cs, eess] (Dec 2020), http://arxiv.org/abs/2012.06765, arXiv: 2012.06765 version: 1

[10] Marimont, S.N., Tarroni, G.: Achieving state-of-the-art performance in the Medical Out-of-Distribution (MOOD) challenge using plausible synthetic anomalies (Nov 2023). https://doi.org/10.48550/arXiv.2308.01412, http://arxiv.org/abs/2308.01412, arXiv:2308.01412 [cs]

[11] Meissen, F., Kaissis, G., Rueckert, D.: Challenging Current Semi-Supervised Anomaly Segmentation Methods for Brain MRI. arXiv:2109.06023 [eess] (Sep 2021), http://arxiv.org/abs/2109.06023, arXiv: 2109.06023

[12] Meissen, F., Kaissis, G., Rueckert, D.: Challenging Current Semi-supervised Anomaly Segmentation Methods for Brain MRI. In: Crimi, A., Bakas, S. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 63–74. Lecture Notes in Computer Science, Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-08999-2_5

[13] Pinaya, W.H.L., Tudosiu, P.D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J.: Unsupervised Brain Anomaly Detection and Segmentation with Transformers. In: arXiv:2102.11650 [cs, eess, q-bio] (Feb 2021), http://arxiv.org/abs/2102.11650, arXiv: 2102.11650 version: 1

[14] Reinke, A., Eisenmann, M., Tizabi, M.D., Sudre, C.H., Rädsch, T., Antonelli, M., Arbel, T., Bakas, S., Cardoso, M.J., Cheplygina, V., Farahani, K., Glocker, B., Heckmann-Nötzel, D., Isensee, F., Jannin, P., Kahn, C., Kleesiek, J., Kurc, T., Kozubek, M., Landman, B.A., Litjens, G., Maier-Hein, K., Martel, A.L., Menze, B., Müller, H., Petersen, J., Reyes, M., Rieke, N., Stieltjes, B., Summers, R.M., Tsaftaris, S.A., Ginneken, B.v., Kopp-Schneider, A., Jäger, P., Maier-Hein, L.: Common limitations of performance metrics in biomedical image analysis (Apr 2021), https://openreview.net/forum?id=76X9Mthzv4X

[15] Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. In: Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery (2017), https://arxiv.org/pdf/1703.05921.pdf

[16] Tan, J., Hou, B., Batten, J., Qiu, H., Kainz, B.: Detecting Outliers with Foreign Patch Interpolation. arXiv:2011.04197 [cs] (Nov 2020), http://arxiv.org/abs/2011.04197, arXiv: 2011.04197

[17] Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., Tyler, L.K., Cam-Can, n., Henson, R.N.: The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional

adult lifespan sample. NeuroImage **144**(Pt B), 262–269 (Jan 2017). https://doi.org/10.1016/j.neuroimage.2015.09.018

[18] Zimmerer, D., Full, P.M., Isensee, F., Jäger, P., Adler, T., Petersen, J., Köhler, G., Ross, T., Reinke, A., Kascenas, A., Jensen, B.S., O'Neil, A.Q., Tan, J., Hou, B., Batten, J., Qiu, H., Kainz, B., Shvetsova, N., Fedulova, I., Dylov, D.V., Yu, B., Zhai, J., Hu, J., Si, R., Zhou, S., Wang, S., Li, X., Chen, X., Zhao, Y., Marimont, S.N., Tarroni, G., Saase, V., Maier-Hein, L., Maier-Hein, K.: MOOD 2020: A public Benchmark for Out-of-Distribution Detection and Localization on medical Images. IEEE Transactions on Medical Imaging pp. 1–1 (2022). https://doi.org/10.1109/TMI.2022.3170077, conference Name: IEEE Transactions on Medical Imaging

[19] Zimmerer, D., Paech, D., Lüth, C., Petersen, J., Köhler, G., Maier-Hein, K.: Unsupervised Anomaly Detection in the Wild. In: Maier-Hein, K., Deserno, T.M., Handels, H., Maier, A., Palm, C., Tolxdorff, T. (eds.) Bildverarbeitung für die Medizin 2022. pp. 26–31. Informatik aktuell, Springer Fachmedien, Wiesbaden (2022). https://doi.org/10.1007/978-3-658-36932-3_6

[20] Zimmerer, D., Petersen, J., Isensee, F., Maier-Hein, K.: Context-encoding Variational Autoencoder for Unsupervised Anomaly Detection. In: International Conference on Medical Imaging with Deep Learning – Extended Abstract Track. London, United Kingdom (Jul 2019), https://openreview.net/forum?id=BylLiVXptV