Revisiting Anomaly Localization Metrics

David Zimmerer D.ZIMMERER@DKFZ.DE and **Klaus Maier-Hein** K.MAIER-HEIN@DKFZ.DE
Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany

Editors: Under Review for MIDL 2024

Abstract

An assumption-free, disease-agnostic pathology detector and segmentor could often be seen as one of the holy grails of medical image analysis. Building on this promise, un-/weakly supervised anomaly localization approaches, which aim to model normal/healthy samples using data and then detect anything deviant from this (i.e., anything abnormal), have gained popularity. However, being an upcoming area in between image segmentation and out-of-distribution detection, most approaches have adapted their evaluation setup and metrics from either field and thus might have missed peculiarities inherent to the anomaly localization task. Here, we revisit the anomaly localization setup, discuss and analyse the properties of the often used metrics, show alternative metrics inspired from instance segmentation and compare the metrics across multiple setting and algorithms. Overall, we argue that the choice of the metric is use-case dependent, however, the Soft Instance IoU shows significant promise going forward.

Keywords: Anomaly Localization, Anomaly Detection, Metrics.

1. Introduction

Accurately detecting and localizing pathologies within medical images is a cornerstone for effective diagnosis and treatment. Unsupervised and weakly-supervised anomaly localization techniques hold potential in this arena, offering the ability to pinpoint abnormalities without extensive disease-specific labeling (Zimmerer et al., 2022b). These methods model the characteristics of normal, healthy tissue, facilitating the identification of deviations. Historically, most anomaly localization methods produce heatmaps to visualize the likelihood of anomalies within an image, necessitating specialized evaluation metrics (Schlegl et al., 2017; Baur et al., 2018; Zimmerer et al., 2018; Chen et al., 2018). However, as a rapidly developing field, anomaly localization approaches have often borrowed evaluation metrics from related domains like image segmentation and out-of-distribution (OoD) detection (Ahmed and Courville, 2019; Zimmerer et al., 2019; Marimont and Tarroni, 2020; Pinaya et al., 2021; Meissen et al., 2022; Zimmerer et al., 2022a; Lagogiannis et al., 2023). This practice might overlook the unique nuances of the anomaly localization task, potentially hindering the optimal selection of metrics.

2. Anomaly Localization 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 DCE (Dice Similarity Coefficient) and IoU (Intersection over Union). While commonly used in segmentation tasks (Isensee et al., 2018), DCE 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

OoD Metrics Classical OoD Metrics are AP (Average Precision) and AUROC (Area Under the Receiver Operating Characteristic). 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. While often addressed by combining labeled and unlabeled data (e.g., evaluating metrics across the entire dataset ("Dataset level"), or using batch-wise calculations as in Zimmerer et al. (2022a)), this approach can overemphasize larger, potentially easier-to-detect anomalies (Reinke et al. (2021); Maier-Hein et al. (2023)).

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 Distance, which can be aggregated using mean, median, or by applying a threshold (e.g., IoU ¿ 0.5) to classify TPs, FPs, and FNs at the object level, enabling the calculation of derived metrics like F1-score. H owever, binarization of predictions remains necessary for object identification (e.g., using connected-component analysis). For this work, we adapt the Center Distance metric: a object's heatmap center point lying within the convex hull of a labeled object constitutes a TP.

Anomaly Localization Metrics To harness the strengths of instance segmentation metrics while avoiding the drawbacks of binarization thresholds, we introduce Soft Instance IoU (inspired by Soft DCE [1]). This modified Instance IoU integrates continuous anomaly scores for a more nuanced assessment of predicted anomaly confidence: $\text{SoftIoU} = \frac{\sum_{i \in Object \cup Background} \alpha \hat{y}_i y_i}{\sum_{i \in Object \cup Background} (0.5\hat{y}_i + (1-\alpha) y_i)}$, where *Background* refers to all pixels not labeled as objects, i indexes the objects in the sample and α is a weighting factor to balance under- and over-segmentation.

3. Experiments & Results

Metric Analysis To gain insights into metric behavior, we designed a controlled experimental setting with 50 samples using perfectly segmented, circular objects. We systematically introduced perturbations to these segmentations, including: (a) Adding small true detected objects while reducing segmentation size (roughly preserves overall segmented pixel count). (b) Varying segmentation size. (c) Adding false detections (FPs). (d) Adding missed instances (FNs). (e) Adding "empty" samples without label or prediction. A few properties of the different metrics emerge (Fig. 3, top): (a): AP and AUROC metrics unexpectedly decreased despite improved object detection. Soft IoU metrics increased as intended, while object-based metrics exhibited some noise but remained relatively consistent. (b): Most metrics exhibited the expected peak-shaped response to segmentation size

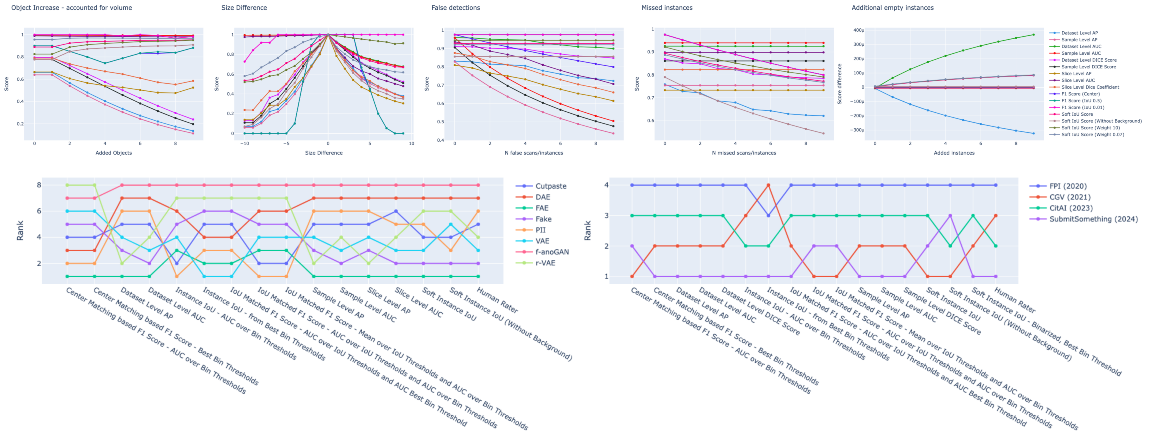

Figure 1: Top: Metrics analysis experiments. Bottom left: CamCAM, right: MOOD.

changes. However, F1 (center-based) and F1 (0.5-IoU) curves were less pronounced, while Dataset AP and F1 (0.01-IoU) were nearly constant on one side. (c): Sensitivity to false positives varied across metrics. AP, AUROC, and F1 (center) showed strong reactions, while Soft IoU was less affected. (d): F1 metrics, Soft IoU, and Dataset Level AP and DCE appeared most sensitive to missed instances. Sample-level and slice-level metrics failed to reflect the performance change. (e): Only F1 metrics and Soft IoU registered an improvement when completely normal samples were added. Dataset AP surprisingly decreased, while other metrics were insensitive by design.

Metric Behavior in Anomaly Benchmark Settings In a second setting (Fig. 3, bottom), we conducted experiments to compare the performance of different anomaly detection algorithms across the diverse metrics and evaluate how well they align with human assessment in a closer to real-world setting. First, seven algorithms were tested on the CamCAM dataset (Taylor et al., 2017). We introduced artificial anomalies in the form of colored spheres (one large, four small) into 50% of the test images. The framework, hyperparameters and training schedules were kept consistent with Lagogiannis et al. (2023). Second, we evaluated the respective winning algorithms from the MOOD challenge (Zimmerer et al., 2022a) on the MOOD brain dataset, similarly introducing colored sphere anomalies in 50% of test images. Here, respectively Soft Instance IoU and F1-based metrics closely mirrored human judgment on the anomaly detection task. However, it's crucial to note that the human evaluation was restricted to segmented slices, potentially downplaying the impact of false positives unrelated to existing anomalies.

4. Disscussion & Conclusion

Our experiments highlight how different metrics capture distinct aspects of anomaly detection performance. While the ideal choice is task-dependent, Soft Instance IoU and F1-based metrics demonstrate favorable properties for many anomaly detection scenarios. This underscores the importance of careful metric selection to align with the goals of the specific application.

References

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

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

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

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

Ioannis Lagogiannis, Felix Meissen, Georgios Kaissis, and Daniel Rueckert. Unsupervised Pathology Detection: A Deep Dive Into the State of the Art. *IEEE Transactions on Medical Imaging*, pages 1–1, 2023. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2023.3298093. URL http://arxiv.org/abs/2303.00609. arXiv:2303.00609 [cs].

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Büttner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, A. Emre Kavur, Carole H. Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A. Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Matthew Blaschko, M. Jorge Cardoso, Veronika Cheplygina, Beth A. Cimini, Gary S. Collins, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Robert Haase, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, Hannes Kenngott, Florian Kofler, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, Jens Petersen, Nasir Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clara I. Sánchez, Shravya Shetty, Maarten van Smeden, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, and Paul F. Jäger. Metrics reloaded: Recommendations for image analysis validation, June 2023. URL http://arxiv.org/abs/2206.01653. arXiv:2206.01653 [cs].

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

Felix Meissen, Georgios Kaissis, and Daniel Rueckert. Challenging Current Semi-supervised Anomaly Segmentation Methods for Brain MRI. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Lecture Notes in Computer Science, pages 63–74, Cham, 2022. Springer International Publishing. ISBN 978-3-031-08999-2. doi: 10.1007/978-3-031-08999-2_5.

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

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

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

Jason R. Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A. Shafto, Marie Dixon, Lorraine K. Tyler, null Cam-Can, and Richard N. Henson. 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, January 2017. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2015.09.018.

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

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

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

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