SEG4SEG: Identifying Systematic Failure Modes in Segmentation by Subgroup Discovery Methods

Nina Weng[1] (iD)   NINWE@DTU.DK
Eike Petersen[2]   EIKE.PETERSEN@MEVIS.FRAUNHOFER.DE
Alceu Bissoto[3]   ALCEU.BISSOTO@UNIBE.CH
Susu Sun[4]   SUSU.SUN@UNI-TUEBINGEN.DE
Lisa M. Koch[3]   LISA.KOCH@UNIBE.CH
Aasa Feragen[1]   AFHAR@DTU.DK
Siavash Bigdeli[1]   SARBI@DTU.DK
Christian F. Baumgartner[4,5]   CHRISTIAN.BAUMGARTNER@UNILU.CH
   [1] *Technical University of Denmark*   [2] *Fraunhofer Institute for Digital Medicine MEVIS*
         [3] *University of Bern*   [4] *University of Tübingen*   [5] *University of Lucerne*

Editors: Under Review for MIDL 2026

Abstract

Deep learning models for medical image segmentation can achieve high overall performance but fail systematically on critical subgroups. While Slice Discovery Methods (SDM) have shown promise in revealing classification failures, their effectiveness for segmentation remains unexplored. Moreover, although various systematic failures have been reported in segmentation tasks, no prior work has systematically categorized them. In this work, we address both gaps. First, we categorize potential sources of systematic errors in medical image segmentation. Second, we empirically investigate whether SDMs can identify problematic slices in each of those categories without manual annotations. Our evaluation covers four controlled failure types and two real-world failure cases, using medical imaging datasets and explicit success criteria for SDM evaluation. Our experiments show that SDMs adapted for segmentation can identify systematic errors, demonstrating their potential for failure analysis in medical imaging.

Keywords: Subgroup discovery, segmentation, systematic error, shortcut learning

1. Introduction

Understanding why machine learning models fail in high-stakes domains such as medical imaging can reveal important root causes of model failures. As a motivating example, consider the case of pneumothorax classification from chest X-rays, for which Larrazabal et al. (2020) documented a significant gender performance gap. An explanation – and, thus, a potential solution – of this phenomenon remained elusive, however. Initial hypotheses centered on underrepresentation and biological factors (e.g., breast shadows), yet both were systematically ruled out (Weng et al., 2023). Finally, Olesen et al. (2024) applied Slice Discovery Methods (SDMs) to investigate this problem, revealing that the model had learned a shortcut based on the presence of chest drains in the images. Chest drain prevalence differs systematically between genders, thereby linking the shortcut to the previously unexplained gender performance disparity and, finally, providing a root cause explanation.[1]

---

1. The pneumothorax–chest drain shortcut had been documented earlier, but the connection to gender performance gaps was unknown (Oakden-Rayner et al., 2020; Jiménez-Sánchez et al., 2023).

Slice discovery methods (SDMs) are unsupervised or semi-supervised clustering methods that aim to identify semantically coherent clusters (slices) of input data that differ in model performance, thereby aiding the discovery of model failure modes (Eyuboglu et al., 2022; Bissoto et al., 2025; Olesen et al., 2024). SDMs have been predominantly applied in the realm of image classification, and a successful application to segmentation tasks has not been demonstrated. Similar to classification models, segmentation models have also been shown to suffer from systematic subgroup performance differences, however. Puyol-Antón et al. (2022) demonstrated racial bias in DL-based cine CMR segmentation, Li et al. (2024a) showed demographic performance gaps in SAM-based abdominal organ CT segmentation, Ioannou et al. (2022) found significant sex and racial performance gaps in brain MR segmentation, and many further studies have presented similar findings (Li et al., 2024b; Čevora et al., 2024; Dou et al., 2024). Yet most such studies do not investigate the root causes of these observed disparities – which, as our motivating example illustrates, is crucial for successfully addressing these failure modes and making models more robust. Motivated by this gap, we here extend the slice discovery paradigm to segmentation tasks.

Significant gaps remain in applying SDMs to segmentation tasks for medical imaging:

Issue 1: Lack of focus on segmentation tasks. Most existing SDM frameworks target classification with single-label predictions. Segmentation operates on pixel-level annotations, fundamentally expanding the failure space: models can fail through boundary errors, spatial shortcuts, or texture biases that are absent in classification. While recent work shows shortcut learning occurs at both sample and pixel levels (Lin et al., 2024), a systematic investigation of SDMs for segmentation failures remains absent.

Issue 2: Absence of success criteria for slice discovery results. To use slice discovery in practice, a way to evaluate whether the discovered slices genuinely capture systematic errors is needed. However, existing SDMs lack clear evaluation criteria, making it challenging to determine whether an SDM actually works.

Issue 3: Lack of taxonomy for segmentation failures. Unlike classification, segmentation lacks a systematic taxonomy of failure types. This gap hinders both the development of targeted SDM approaches and the interpretation of discovered slices.

Accordingly, the **core focus and contributions** of our work are:

1. We adapt SDMs to segmentation for the first time through our proposed SEG4SEG pipeline (Fig. 1) and analyze embedding variants tailored to segmentation tasks.
2. We redefine evaluation metrics and propose principled criteria for assessing slice discovery quality.
3. We propose a systematic taxonomy of failure modes in medical image segmentation and benchmark our SDM across four controlled-error settings and two real-world datasets, demonstrating SEG4SEG's effectiveness in uncovering diverse failure modes.

2. Problem Definition and Background

In this section, we formalize the problem of detecting systematic performance disparities in segmentation models and review existing SDMs developed for classification tasks, which our work builds upon.
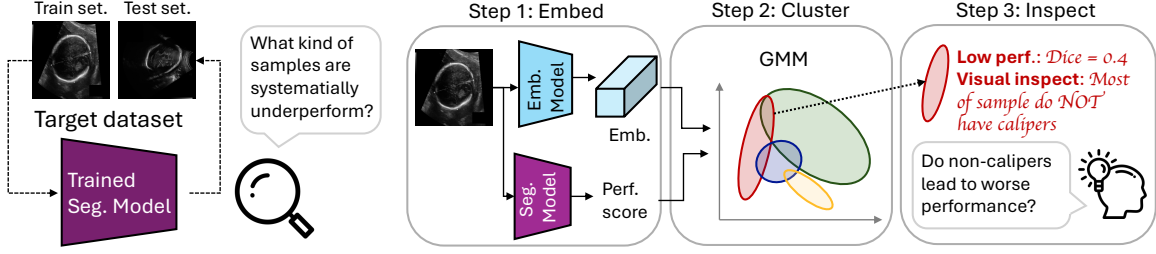
Figure 1: **Illustration of the SEG4SEG framework**: Three steps are included: (1) Embedding the image/annotation information; (2) Clustering using both image embedding and performance score; (3) Inspecting the problematic slices.

2.1. Problem Set-up: Performance Disparities in Hidden Subgroups

We now formalize the learning setting and define how subgroup-related performance disparities manifest in segmentation models. Given a dataset $D = (x, y)_{i=1}^N$ and a hidden binary attribute A (e.g. presence of caliper) , we assume that the marginal distribution of A is consistent across the train, validation, and test splits[2], i.e. $P(A = 1|D_{train}) = P(A = 1|D_{val}) = P(A = 1|D_{test})$. We train a segmentation model f using D_{train} and perform model selection on D_{val}. Our goal is to do detect performance disparities with respect to A, i.e. $\mathbb{E}[\text{perf}(x) \mid A = 1] \neq \mathbb{E}[\text{perf}(x) \mid A = 0]$. Our goal is to develop a method that can detect such disparities from model outputs on unseen data, without requiring access to A at inference time. In our experiments, we use and manipulate known A, to confirm the ability to surface A.

2.2. Slice Discovery: Concept and Motivations

SDMs aim to form clusters in which samples within each cluster are semantically similar, enabling the identification of slices that perform systematically worse. In practice, SDMs operate on feature representations of test samples, such as predictions, or foundation model embeddings. Let $z(x)$ denote the representation extracted from input x, which serves as the basis for clustering. The choice of $z(x)$ determines which types of failure modes the SDMs can potentially uncover. For example, classification logits may reveal class-specific confusion, while encoder embeddings of the imaging captures image-level features.

Formally, given a set of test samples x_i and their corresponding representations $z(x_i)$, an SDM produces a partition $S = s_1, s_2, ..., s_K$, where each slice s_k consists of a subset of samples that are close in the representation space and exhibit similar performance. In our case, we aim to identify clusters $s \in S$ that are predominantly composed of samples with $A = 1$, and whose average performance is lower compared to clusters dominated by $A = 0$, or the other way around. Importantly, this clustering process is conducted in an unsupervised manner, without access to the underlying bias attribute A.

---

2. This assumption follows prior work on SDMs and renders the SDM task more challenging compared to shortcut learning detection, where the distribution often differs in the test set.

2.3. Existing SDMs for Classification: Capabilities and Limitations

Pipeline Design. SDMs for classification have evolved from simple clustering to sophisticated cross-modal frameworks.[3] Oakden-Rayner et al. (2020) pioneered the approach using k-means on pre-softmax features. DOMINO (Eyuboglu et al., 2022) popularized a workflow that combines foundation model embeddings (CLIP), clustering (GMM), and natural language interpretation. Subsequent work extended this framework: FACTS (Yenamandra et al., 2023) amplified clustering correlations, PlaneSpot (Plumb et al., 2023) improved on dimension reduction, and ViG-Bias (Marani et al., 2024) integrated visual explanations. An alternative paradigm, Spotlight (d'Eon et al., 2022), identifies contiguous low-performance regions without discrete clustering. However, all existing methods focus on classification, leaving segmentation unexplored.

Evaluation. Evaluating slice discovery remains challenging. Early work evaluated case-specific results through subgroup prevalence within slices (Oakden-Rayner et al., 2020; Olesen et al., 2024). DOMINO introduced precision-based metrics widely adopted in subsequent work including FACTS, but these metrics are limited to synthetic datasets with known bias attributes, motivating Bissoto et al. (2025) to propose improved metrics for real-world medical imaging settings. Building on this work, we further refine these metrics and introduce criteria to assess whether problematic slices are discovered.

3. A Taxonomy of Segmentation Failure Modes

We first systematically categorize documented segmentation failure modes as follows.

3.1. Shortcuts/Unwanted Correlation

Shortcut learning occurs when models exploit unwanted correlations between input features and targets instead of learning meaningful patterns (Geirhos et al., 2020; Lin et al., 2024). In segmentation, shortcuts can occur at both the sample level and pixel level:

Sample-level shortcuts arise from unwanted correlations between *sample-level features* and *segmentation performance*. One example is the presence of calipers in fetal ultrasound scans: models achieve better segmentation performance when calipers are present (Lin et al., 2024). Including calipers in scans is common in clinical practice (Pu et al., 2022; Bano et al., 2021; Sun et al., 2022; Yang et al., 2020), yet their presence can mislead models into shortcut learning. We validate this failure mode in Sec. 5.1 (Case A).

Pixel-level shortcuts arise from unwanted correlations between *pixel locations* and *labels*. A representative example is spatial bias in skin lesion segmentation: Lin et al. (2024) demonstrated that when all training samples are center-cropped, models learn to systematically predict background labels for boundary pixels, regardless of content. Similar border artifacts appear in published models (Wang et al., 2023; Dai et al., 2022), where a U-Net/CA-Net failed to recognize lesions near image boundaries, suggesting widespread reliance on spatial priors rather than visual features. We validate this failure mode in Sec. 5.1 (Case B).

---

3. Some approaches mitigate subgroup disparities without uncovering their causes (Jain et al., 2023; Kim et al., 2019; Sohoni et al., 2020). These fall outside our scope, as we focus on failure mode discovery.

3.2. Label Noise

Annotation styles are often neglected in segmentation research, yet studies show that they can substantially affect model performance (Nichyporuk et al., 2022; Abhishek et al., 2024; Zhang et al., 2020; Zepf et al., 2023). Such variations encode the inherent subjectivity of labeling, driven by annotation protocols, rater expertise, and data pre-processing. We validate this failure mode in Sec. 5.1 (Case C).

Annotation error, inevitable in human-driven labeling, can introduce systematic failures for segmentation. Common manifestations are omission, inclusion errors and annotator-driven cognitive bias arising from boundary-related ambiguities (Vădineanu et al., 2022).

3.3. Underrepresentation

Class imbalance is a longstanding challenge in segmentation and a well-documented failure mode. Numerous remedies have been proposed (Li et al., 2020; Müller et al., 2022), particularly for structurally small anatomical regions such as retinal vessels (Fauzi et al., 2022). Underrepresentation also manifests at the sample level, where limited coverage of certain demographic groups (Puyol-Antón et al., 2021) or sub-disease categories leads to models that underperform on these minority slices.

3.4. Difficult Cases

Beyond the factors outlined above, certain subgroups may remain intrinsically difficult to segment due to image-dependent characteristics. Typical sources of increased complexity include degraded image quality (Jin et al., 2022), challenging anatomical topology such as curvilinear structures (Lin et al., 2023), and substantial variability in radiographic appearance (Heller et al., 2021). We validate this failure mode in Sec. 5.1 (Case D).

4. SEG4SEG: Slice Discovery Methods for Segmentation

Here, we propose SEG4SEG (Systematic Error Grounding for SEGmentation), which extends SDM to segmentation problems. The overall pipeline extends DOMINO (Eyuboglu et al., 2022) which was developed for classification.

4.1. Method Overview

SEG4SEG consists of three steps (see Fig. 1):

Embedding the image space and performance metrics[4]. Common approaches for image embedding include foundation models (e.g. CLIP) and latent representations from related pre-trained tasks. In this work, we use CLIP for image embedding and include both Dice score and per-image Positive Predictive Value as performance embeddings. To enable efficient clustering[5], we apply UMAP (McInnes et al., 2018) for dimension reduction of the image embeddings.

---

4. We treat both image features and performance metrics as embeddings, as both serve to compress information into a lower-dimensional space; despite the latter are not traditionally considered as embeddings.

5. High dimensionality leads to substantial inefficiency for most clustering methods (McInnes)

Clustering embedding information. Multiple clustering methods have been used in previous research, including k-means, HDBSCAN, and Gaussian Mixture Models (GMM). We use GMM in this work to enable flexible weighting between different variables. Specifically, we optimize

$$\ell(\phi) = \sum_{i=1}^{n} \log \sum_{j=1}^{|S|} P\left(S^{(j)} = 1\right) P\left(z(x_i) \mid S^{(j)} = 1\right) P\left(\mathrm{perf}(\hat{y}_i, y_i) \mid S^{(j)} = 1\right)^{\gamma}, \quad (1)$$

where z is the embedding model; x_i, y_i, and $\hat{y}_i$ denote the image, its annotation, and the prediction, respectively; and γ is a weighting factor balancing the effect of performance embeddings and image embeddings.

Inspecting the clustering results. After clustering, we analyze the resulting clusters to identify potential issues in the dataset or model. This step typically requires additional annotations or analytical tools to characterize the discovered slices.

4.2. Representation Design for Segmentation Tasks

Since segmentation operates on pixel-level, we extend both the image and metrics representations to encode richer spatial information for more effective clustering. Specifically, we extend the representation as follows: **(a) Variants of image space embedding:** Apart from using embeddings from the original image x_0 as in the classification task, we explore alternative inputs by masking the image with either the ground-truth or predicted mask ($x_0 \cdot M_{\mathrm{GT}}$ or $x_0 \cdot M_{\mathrm{pred}}$) to encode additional mask information. **(b) Variants of performance space embedding**: Beyond commonly used overlap metrics (Dice score), we consider confusion-related metrics (Positive and negative predictive values: PPV, NPV) that distinguish FP and FN behaviors.

4.3. Evaluation Metrics for Slice Discovery

Following Bissoto et al. (2025), we evaluate SDM performance through the trade-off between performance disparities and slice purity. The goal is to discover slices that are pure regarding attribute A and exhibit performance that deviates meaningfully from the overall population.

Performance disparity. In contrast to prior work, we measure performance disparity using the Omega Square (ω^2) measure (Cohen, 1973), which quantifies the proportion of variance in performance explained by the partition S. Intuitively, ω^2 measures how much better the partition is at separating high/low-performance samples compared to random grouping. We adopt ω^2 instead of the difference between best and worst subgroup performance used by Bissoto et al. (2025), because the latter is sensitive to small cluster sizes and spurious results, whereas ω^2 provides a more robust effect size measure.

Purity. Slice purity measures the homogeneity of a partition S with respect to attribute A: $AP(S) = \frac{1}{|A|} \sum_{a \in A} \max_{s \in \hat{S}} \left(\frac{n_{s,a}}{n_s} \right)$, where $n_{s,a}$ denotes the number of samples belonging to attribute a within slice s, and n_s is the total number of samples in slice s. Here, $\hat{S}$ denotes the subset of all clusters S satisfying $n_s > N_{\min}$. Intuitively, we exclude very small clusters when computing purity to avoid their noise-dominated behavior from inflating the metric.

Table 1: Overview of failure modes and corresponding datasets.

Case	Failure Mode	Real-world Case	Dataset	Failure Attribute (a_f)
A	Shortcut (sample level)	Calipers in ultrasound	HC18	Non-caliper
B	Shortcut (pixel level)	Central cropping in skin lesions	ISIC2018	Non-central masking
C	Annotation style	Boundary style in skin lesions	ISIC2018	Polygon-style annotation
D	Difficult cases	Low-quality retinal images	FIVES	Low quality

The purity metric AP does *not* penalize impure clusters if each attribute value $a \in A$ is concentrated in at least one slice. Both metrics aim to identify slices with high attribute purity and low performance. Impure clusters are acceptable, as representations may capture multiple data characteristics beyond target attributes.

4.4. Defining the success of finding the problematic slice

The two metrics introduced above assess, respectively, (i) whether an SDM can identify slices with low performance, and (ii) whether the SDM can isolate slices with high attribute purity. However, taken separately, these metrics do not directly answer the key question of interest: ***Does the SDM successfully locate the problematic slice(s)?***

To close this gap, we combine the two metrics and propose the following validation criterion to determine whether an SDM successfully identifies a problematic slice:

Definition 1 (SDM Success Criterion) An SDM is considered to successfully identify a problematic slice if *at least one* slice s in the dataset satisfies all of the following conditions:
1. **High Purity:** $P_s(A = a_f) > p_\theta$. The proportion of samples in slice s that share the failure-related attribute (noted as a_f) value exceeds the purity threshold.
2. **Low Performance:** $q(\text{perf}(s)) \leq q_\theta$. The quantile of the average performance of slice s falls below the performance threshold.
3. **Sufficient Slice Size:** $n_s \geq N_{\min}$. The slice s contains at least $N_{\min}$ samples.

If such a slice exists, it indicates that in a real-world setting where attribute labels are not always available, the SDM can locate low-performance slices, and the high purity in the slice might then help a human inspector to form a hypothesis as to the cause of the model failure. We use $p_\theta = 0.8$, $q_\theta = 0.4$ and empirically define N_{min} to be 2% of the test samples.

5. Experimental Design

As shown in Tab. 1, we select four representative cases from the failure mode taxonomy defined in Sec. 3, introduce the corresponding errors artificially, and validate whether the SDM can reveal them. We also evaluate the framework on two real-world datasets that naturally contain annotation style inconsistencies (ISIC) and image quality variations (FIVES).

5.1. Experiment Set-up and Dataset Choices

Case A – Sample-level Shortcuts: Calipers in Ultrasounds.   We apply our SEG4SEG to detect the caliper shortcut documented by Lin et al. (2024). We use the HC18 dataset (van den Heuvel et al., 2018) and introduce different error levels by generating artificial calipers from supplied segmentations.

Case B – Pixel-level Shortcuts: Central Cropping in Skin Lesions. We extend the analysis of Lin et al. (2024) on ISIC2018 by introducing synthetic cropping to reduce foreground-border correlations. For each cropping level n, we randomly select the top $n\%$ of samples, apply random crops at half the original image size, and discard crops that contain no lesion pixels.

Case C – Annotation Styles: Skin Lesion. Three distinct annotation styles are observed in ISIC2018: *flood-fill*, *jagged*, and *polygon* (Zepf et al., 2023). We conduct two experiments: (1) similar to Cases A and B, we introduce systematic errors by degrading *flood-fill* annotations into *polygon*-like ones at different levels, and (2) we use the original dataset with manually labeled annotation styles.

Case D – Difficult Cases: Low Image Quality in Retinal Imaging. FIVES is a retinal dataset annotated with three types of image quality degradation. Following Case C's design, we evaluate the SDM under two setups: (1) synthetic degradation by darkening (β=0.5 intensity scaling) and blurring (Gaussian θ=2.0), and (2) the original dataset with existing quality annotations.

5.2. Hyper-parameters and Experimental Details

We perform 5 random runs for each experiment to estimate the detection rate of problematic slices. Failure attributes are injected at 5%, 10%, and 20% prevalence. We sweep the weighting parameter $\gamma \in [10^{-3}, 10^3]$ (see Eq.1), and observe that $\gamma = 10$ performs best across all settings; thus, we report results using this value unless stated otherwise. We split all datasets either following their original train/val/test partitions or using a 64/16/20 split. Segmentation models are selected based on validation performance, while the GMM is fit and evaluated on the test set.

6. SDM Results and Discussions

SEG4SEG is able to discover problematic slices in segmentation. Fig. 2 shows that our SDM effectively identifies performance-relevant problematic slices. In Cases A, C, and D, our SDM achieves high detection rates even at a prevalence as low as 5%. Case B with central cropping exhibits lower detection at low prevalence, likely because SEG4SEG operates at the sample level, whereas pixel–position shortcuts may require patch- or pixel-level grouping for precise identification.

Purity and performance disparity alone do not capture slicing success. Fig. 2b illustrates that neither purity nor performance disparity, whether considered independently or jointly, suffice to determine whether a slice is meaningfully identified. We define performance pass if $\omega^2 > 0.14^6$ and purity pass if $\text{Purity}(a_f) > 0.8$. However, even when caliper/non-caliper difference is negligible, both perf. pass and purity pass remain high, meaning that the low-performing slice is not necessarily the one driving the purity signal. This highlights a mismatch between these metrics and the actual reliability of slice discovery.

---

6. Values for ω^2 0.14 indicates large effects. (Field, 2024)

(a) SDM results on all cases

(b) SDM results on Case A with different % non-caliper

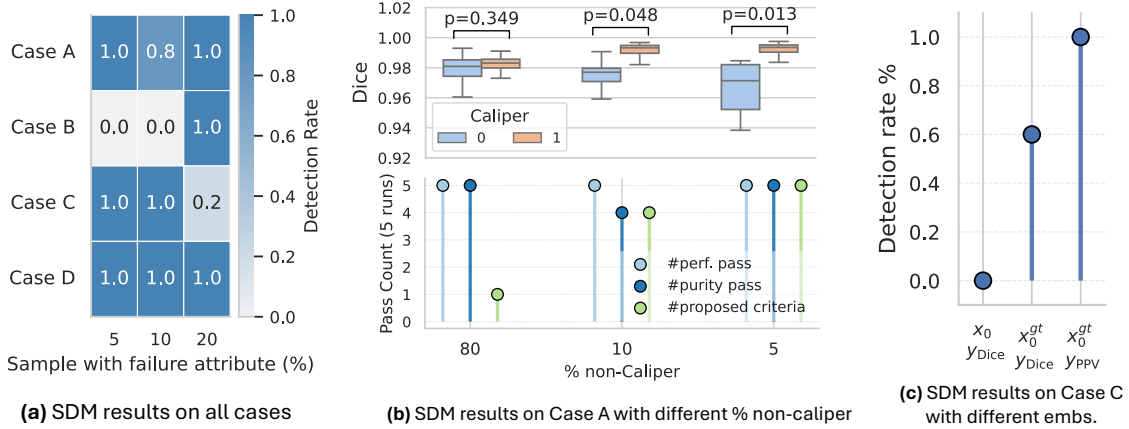**(c)** SDM results on Case C with different embs.

Figure 2: **SDM results.** (a) Proposed criteria evaluated over 5 random runs with varying failure-attribute proportions. (b) Case A. Top: the caliper vs. non-caliper performance gap is evident at prevalence 5% and 10% for non-caliper but negligible at 80%; Bottom: performance-pass/purity-pass shows no sensitivity to error levels, whereas the proposed criteria do. (c) Case C results with different embeddings.

The proposed evaluation criteria reflect whether a failure attribute affects model performance. Fig. 2b shows that the proposed criteria increase their pass rates as the performance gap between caliper and non-caliper samples grows, indicating that the criteria align with the actual impact of the failure attribute.

Segmentation-specific embedding choices are critical for failure detection. Case C illustrates the necessity of incorporating embedding variants (Sec. 4.2) as the error stems from annotation quality, not image appearance. As shown in Fig. 2c, using default embeddings (original images + Dice) fails to detect the degraded-annotation samples. Detection improves when the image embedding incorporates annotation masks, and improves further when replacing Dice with PPV, which is more sensitive to pixel-level false positives.

SEG4SEG also works on real-world datasets. Applying our SDM to the original ISIC2018 and FIVES datasets successfully identifies problematic samples arising from annotation style inconsistencies and low image quality, respectively. As shown in Fig. 3, the SDM highlights these failure modes by successfully passing the criteria, and the significant differences between the subgroups confirm the existence of these failures. Fig. 4 provides a detailed example of an SDM result for the low-quality retinal images. As illustrated in the center plot of Fig. 4, clusters with poorer performance exhibit a higher proportion of low-quality samples. Visualizing representative samples from each cluster further supports this trend: clusters with higher performance predominantly contain high-quality images, whereas cluster 0, which performs the worst, consists entirely of low-quality images. The SDM is also capable of capturing other meta-attributes: for example, one cluster groups together images with similar visual patterns, 93% of which correspond to cases without diagnosed eye disease.
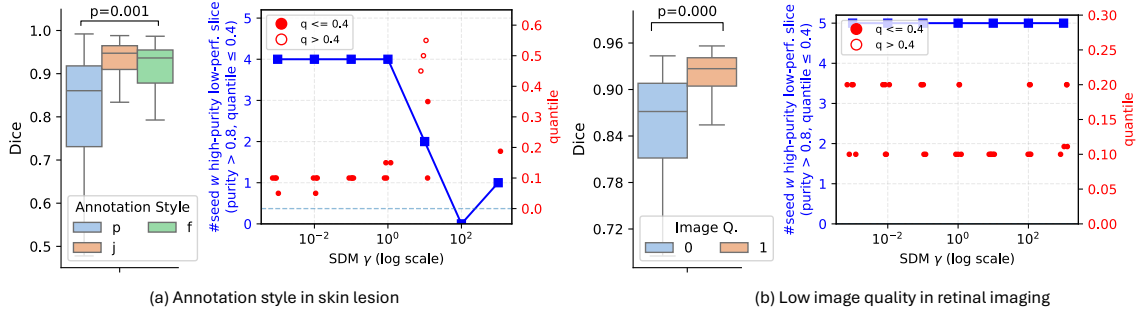
9

Figure 3: **SEG4SEG uncovers failure modes in real-world datasets**: (a) annotation style in skin lesion segmentation, (b) low image quality in retinal imaging. Performance disparities are shown left, SDM results for different γ are shown right.
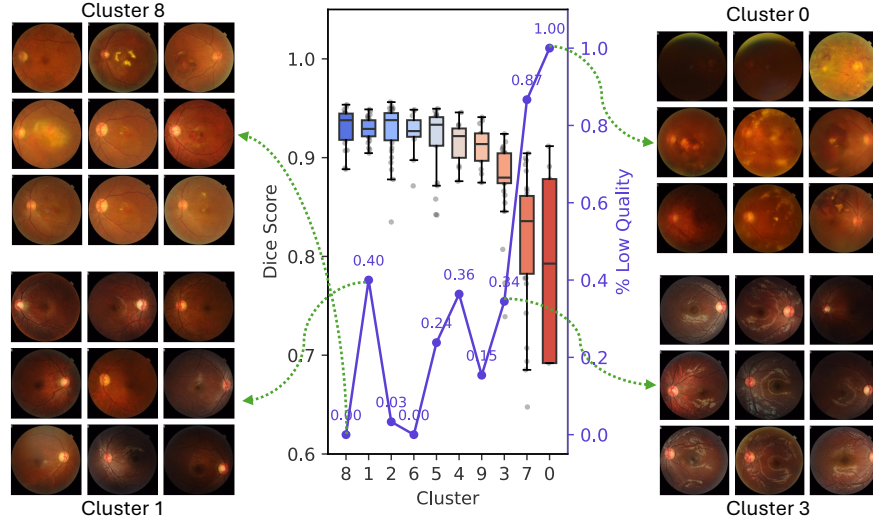

Figure 4: **Exemplary SDM result on the original FIVES dataset.** Center: Dice scores for the 10 clusters and the proportion of low-quality images in each cluster. Left and right: samples from selected clusters. The SDM identifies low-performance slices that contain a high proportion of low-quality images.

7. Conclusion

In this work, we investigated whether SDMs can be applied to discovering failures modes in segmentation tasks. We have proposed a taxonomy of segmentation failure modes, selected representative test cases, and validated an adapted SDM (SEG4SEG) across four controlled-error settings and two real-world settings. Our results demonstrate that SEG4SEG is effective in revealing diverse failure modes in medical image segmentation and shows strong potential as a tool for systematic failure analysis.

10

References

Kumar Abhishek, Jeremy Kawahara, and Ghassan Hamarneh. Segmentation style discovery: Application to skin lesion images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–34, 2024.

Sophia Bano, Brian Dromey, Francisco Vasconcelos, Raffaele Napolitano, Anna L David, Donald M Peebles, and Danail Stoyanov. AutoFB: automating fetal biometry estimation from standard ultrasound planes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 228–238, 2021.

Alceu Bissoto, Trung-Dung Hoang, Tim Flühmann, Susu Sun, Christian F. Baumgartner, and Lisa M. Koch. *Subgroup Performance Analysis in Hidden Stratifications*, pages 594–603. 2025. doi: 10.1007/978-3-032-05185-1_57.

Jacob Cohen. Eta-squared and partial eta-squared in fixed factor anova designs. *Educational and psychological measurement*, 33(1):107–112, 1973.

Duwei Dai, Caixia Dong, Songhua Xu, Qingsen Yan, Zongfang Li, Chunyan Zhang, and Nana Luo. Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation. *Medical image analysis*, 75:102293, 2022.

Greg d'Eon, Jason d'Eon, James R Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1962–1981, 2022.

Qi Dou, Ruinan Jin, Xiaoxiao Li, Zikang Xu, Qingsong Yao, Yuan Zhong, and S. Zhou. FairMedFM: Fairness benchmarking for medical imaging foundation models. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pages 111318–111357, 2024. doi: 10.52202/079017-3535.

Sabri Eyuboglu, Maya Varma, Khaled Kamal Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Re. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations*, 2022.

Ahmad Fauzi, Raju Wandira, and Kraugusteeliana Kraugusteeliana. The effect of imbalanced classes on retinal blood vessel segmentation. In *2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, pages 375–378. IEEE, 2022.

Andy Field. *Discovering statistics using IBM SPSS statistics*. Sage publications limited, 2024.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. *Medical image analysis*, 67:101821, 2021.

Stefanos Ioannou, Hana Chockler, Alexander Hammers, and Andrew P. King and. A study of demographic bias in CNN-based brain MR segmentation. In *Lecture Notes in Computer Science*, pages 13–22. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-17899-3_2.

Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. In *The Eleventh International Conference on Learning Representations*, 2023.

Amelia Jiménez-Sánchez, Dovile Juodelyte, Bethany Chamberlain, and Veronika Cheplygina. Detecting shortcuts in medical images – a case study in chest x-rays. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

Kai Jin, Xingru Huang, Jingxing Zhou, Yunxiang Li, Yan Yan, Yibao Sun, Qianni Zhang, Yaqi Wang, and Juan Ye. FIVES: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific data*, 9(1):475, 2022.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23): 12592–12594, 2020.

Qing Li, Yizhe Zhang, Yan Li, Jun Lyu, Meng Liu, Longyu Sun, Mengting Sun, Qirong Li, Wenyue Mao, Xinran Wu, Yajing Zhang, Yinghua Chu, Shuo Wang, and Chengyan Wang. An empirical study on the fairness of foundation models for multi-organ image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 432–442, 2024a. doi: 10.1007/978-3-031-72390-2_41.

Qing Li, Yizhe Zhang, Longyu Sun, Mengting Sun, Meng Liu, Zian Wang, Qi Wang, Shuo Wang, and Chengyan Wang. Fairness in artificial intelligence-driven multi-organ image segmentation. *iRADIOLOGY*, 2(6):539–556, 2024b. doi: 10.1002/ird3.101.

Zeju Li, Konstantinos Kamnitsas, and Ben Glocker. Analyzing overfitting under class imbalance in neural networks for image segmentation. *IEEE transactions on medical imaging*, 40(3):1065–1077, 2020.

Manxi Lin, Kilian Zepf, Anders Nymark Christensen, Zahra Bashir, Morten Bo Søndergaard Svendsen, Martin Tolsgaard, and Aasa Feragen. DTU-net: Learning topological similarity for curvilinear structure segmentation. In *International Conference on Information Processing in Medical Imaging*, pages 654–666, 2023.

Manxi Lin, Nina Weng, Kamil Mikolaj, Zahra Bashir, Morten BS Svendsen, Martin G Tolsgaard, Anders N Christensen, and Aasa Feragen. Shortcut learning in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 623–633, 2024.

Badr-Eddine Marani, Mohamed Hanini, Nihitha Malayarukil, Stergios Christodoulidis, Maria Vakalopoulou, and Enzo Ferrante. ViG-Bias: Visually grounded bias discovery and mitigation. In *European Conference on Computer Vision*, pages 414–429, 2024.

Leland McInnes. Using UMAP for clustering. https://umap-learn.readthedocs.io/en/latest/clustering.html. UMAP documentation; accessed December 9, 2025.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, 15(1):210, 2022.

Brennan Nichyporuk, Jillian Cardinell, Justin Szeto, Raghav Mehta, Jean-Pierre Falet, Douglas L. Arnold, Sotirios A. Tsaftaris, and Tal Arbel. Rethinking generalization: The impact of annotation style on medical image segmentation. *Machine Learning for Biomedical Imaging*, 1:1–37, 2022. doi: 10.59275/j.melba.2022-2d93.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020.

Vincent Olesen, Nina Weng, Aasa Feragen, and Eike Petersen. Slicing through bias: explaining performance gaps in medical image analysis using slice discovery methods. In *MICCAI Workshop on Fairness of AI in Medical Imaging*, pages 3–13, 2024.

Gregory Plumb, Nari Johnson, Angel Cabrera, and Ameet Talwalkar. Towards a more rigorous science of blindspot discovery in image classification models. *Transactions on Machine Learning Research*, 2023.

Bin Pu, Yuhuan Lu, Jianguo Chen, Shengli Li, Ningbo Zhu, Wei Wei, and Kenli Li. MobileUNet-FPN: A semantic segmentation model for fetal ultrasound four-chamber segmentation in edge computing environments. *IEEE Journal of Biomedical and Health Informatics*, 26(11):5540–5550, 2022.

Esther Puyol-Antón, Bram Ruijsink, Stefan K Piechnik, Stefan Neubauer, Steffen E Petersen, Reza Razavi, and Andrew P King. Fairness in cardiac MR image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 413–423, 2021.

Esther Puyol-Antón, Bram Ruijsink, Jorge Mariscal Harana, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Reza Razavi, Phil Chowienczyk, and Andrew P. King.

Fairness in cardiac magnetic resonance imaging: Assessing sex and racial bias in deep learning-based segmentation. *Frontiers in Cardiovascular Medicine*, 9, 2022. doi: 10.3389/fcvm.2022.859310.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

Yihao Sun, Hongjian Yang, Jiliu Zhou, and Yan Wang. ISSMF: Integrated semantic and spatial information of multi-level features for automatic segmentation in prenatal ultrasound images. *Artificial Intelligence in Medicine*, 125:102254, 2022.

Şerban Vădineanu, Daniël Maria Pelt, Oleh Dzyubachyk, and Kees Joost Batenburg. An analysis of the impact of annotation errors on the accuracy of deep learning for cell segmentation. In *International Conference on Medical Imaging with Deep Learning*, pages 1251–1267. PMLR, 2022.

Thomas LA van den Heuvel, Dagmar de Bruijn, Chris L de Korte, and Bram van Ginneken. Automated measurement of fetal head circumference using 2d ultrasound images. *PloS one*, 13(8):e0200412, 2018.

Jiacheng Wang, Jing Yang, Qichao Zhou, and Liansheng Wang. Medical boundary diffusion model for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 427–436. Springer, 2023.

Nina Weng, Siavash Bigdeli, Eike Petersen, and Aasa Feragen. Are sex-based physiological differences the cause of gender bias for chest x-ray diagnosis? In *Workshop on Clinical Image-Based Procedures*, pages 142–152, 2023.

Yujie Yang, Pinli Yang, and Bo Zhang. Automatic segmentation in fetal ultrasound images based on improved U-net. In *Journal of Physics: Conference Series*, volume 1693, page 012183. IOP Publishing, 2020.

Sriram Yenamandra, Pratik Ramesh, Viraj Prabhu, and Judy Hoffman. FACTS: First amplify correlations and then slice to discover bias. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4794–4804, 2023.

Kilian Zepf, Eike Petersen, Jes Frellsen, and Aasa Feragen. That label's got style: Handling label style bias for uncertain image segmentation. In *The Eleventh International Conference on Learning Representations*, 2023.

Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Cicarrelli, Frederik Barkhof, and Daniel Alexander. Disentangling human error from ground truth in segmentation of medical images. *Advances in Neural Information Processing Systems*, 33:15750–15762, 2020.

Kate Čevora, Ben Glocker, and Wenjia Bai. Quantifying the impact of population shift across age and sex for abdominal organ segmentation. In *Ethics and Fairness in Medical Imaging*, pages 88–97, 2024. doi: 10.1007/978-3-031-72787-0_9.