Towards Fairness under Label Bias in Image Segmentation: Impact, Measurement and Mitigation

Aditya Parikh Stella Frank Sneha Das Aasa Feragen
Department of Applied Mathematics and Computer Science
Technical University of Denmark
adipa@dtu.dk

Abstract

Biases in annotation pipelines may introduce *label bias*: when annotation quality or style varies systematically across demographic subgroups, causing disparities in performance and fairness audits. Label bias in image segmentation remains underexplored, as even detecting it typically requires clean, unbiased annotations, which are not readily available. We present a Confident Learning framework adapted to segmentation for auditing label bias directly in the training data without a clean, unbiased ground truth. By comparing the provided training labels to the model's confident predictions, we measure the prevalence and direction of suspected label errors, where standard overlap metrics like Dice fail. We further show that label bias influences subgroup separability in the encoder's feature space, an artifact we leverage for bias mitigation rather than suppressing it. We evaluate datasets spanning from synthetic to real-life bias and across experimental conditions, showing how our framework produces proxy audit signals without clean labels and mitigates disparities with a correctly identified cleaner reference group. [1]

1 Introduction

Annotation (interchangeably, labels) quality depends on task objective, annotation guidelines, rater experience, labeling tools, and automated initialization procedures [Zepf et al., 2023, Nichyporuk et al., 2022, Zhang et al., 2020a]. When these factors are systematically associated with *sensitive attributes* (like race, sex, or age [Haeri and Zweig, 2020]), the resulting dataset exhibits **group-conditioned label bias**. Unlike label noise, label bias is systematic: its prevalence or direction depends on the subgroup or annotation process [Wang et al., 2021, Beck et al., 2020].

Label bias in safety-critical AI systems can cause substantial harm to affected groups, but identifying it is difficult without access to correct labels. Prior work in medical image segmentation shows that if a model systematically over-segments tumors for certain patient populations, those errors propagate directly into clinical decisions like surgery or radiotherapy planning, harming the affected group [Chen et al., 2023, Koçak et al., 2025]. Yet, standard segmentation metrics (e.g., Dice, IoU) condense complex errors into a single scalar, masking where and how bias manifests [Taha and Hanbury, 2015, Reinke et al., 2021]. It becomes more difficult to detect label bias when the data lack clean, unbiased ground-truth labels, which leads to systematic bias [Quinn et al., 2023, Parikh et al., 2026, Juni et al., 2026].

Label bias is well understood in classification [Obermeyer et al., 2019], where controlled label-flip simulations [Dai and Brown, 2020] and detection tools such as Cleanlab [Northcutt et al., 2021, Lad

[1] To ensure reproducibility, all code, dataset configuration, and random seeds are made available at: `https://ade-17.github.io/label-bias-segmentation`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

and Mueller, 2023] establish a principled experimental foundation [Zhang et al., 2024, Wang et al., 2025]. In segmentation, errors are spatially correlated and often concentrated at object boundaries, rather than occurring as independent label flips [Heller et al., 2018, Yao et al., 2023], creating corruption that is more challenging than spatially unstructured noise [Heller et al., 2018]. A research gap exists for simulating, identifying, and mitigating group-conditioned label bias in segmentation.

We address this gap with a framework to audit and mitigate group-conditional label bias.

Our contributions are:

- We adapt Confident Learning to segmentation for fairness auditing, introducing metrics to estimate the magnitude and direction of potential annotation errors **without clean ground-truth labels**.
- We show that label bias can amplify subgroup separability in learned representations and examine its effects on segmentation performance and subgroup disparities.
- We propose a subgroup-conditioned segmentation approach that uses a selected reference annotation style to mitigate label bias, and evaluate it against fairness, feature-alignment, and noise-robust baselines across multiple experimental settings.

2 Background and Related Work

Label bias. Label bias appears when the labels used for training and testing are systematically incorrect in ways that differ across sensitive groups. Existing work highlights how label bias will affect both model performance [Jiang and Nachum, 2020, Vorontsov and Kadoury, 2021, Dai and Brown, 2020] and methods for auditing and explaining AI models [Parikh et al., 2026, Akintande et al., 2025], but existing mitigation tools are limited to settings where true labels are available [Parikh et al., 2026] or the bias is known [Jiang and Nachum, 2020]. As label bias primarily affects settings where true labels are impossible or expensive to obtain, this leaves a crucial unmet need for mitigation when true labels or bias are unknown.

Learning with Noisy Labels. While label noise covers generic disagreement between observed and intended labels, label bias refers to systematic label errors associated with a specific group, annotator, tool, or protocol [Dai and Brown, 2020, Wang et al., 2021, Beck et al., 2020]. This systematic corruption has fundamental consequences for learning: while models are robust to label noise, they are sensitive to label bias even at modest scales [Vorontsov and Kadoury, 2021]. The field of learning with noisy labels (LNL) has focused on training robust neural networks under corrupted annotations. LNL traditionally models noise as uniform random flipping or class-conditional label transitions [Natarajan et al., 2013, Patrini et al., 2017]; for segmentation, it is commonly treated as boundary uncertainty or random spatial perturbations [Bernhard et al., 2024]. Prior work has highlighted that annotation errors in segmentation are rarely random [Lipton, 2018], making standard LNL techniques, including robust loss functions and global label smoothing, fail, as they assume noise is independent of underlying semantic content [Jiang et al., 2020, Northcutt et al., 2021]. Our work formalizes this systematic label bias not as a random dataset nuisance, but as a group-conditional geometric bias that alters the network's representation space.

Algorithmic Fairness in Segmentation. Algorithmic fairness has been extensively studied in classification, with a predominant focus on mitigating performance disparities across protected attributes [Barocas et al., 2023, Buolamwini and Gebru, 2018]. Fairness in segmentation remains underexplored [Wang et al., 2025]. The existing segmentation fairness literature assumes a clean, unbiased ground truth, treating performance disparities as a failure of model capacity alone [Oakden-Rayner et al., 2020]; an assumption we directly challenge. In medical imaging and real-world vision tasks, annotation protocols are inherently subjective and susceptible to demographic-specific boundary distortions [Jungo et al., 2018, Quinn et al., 2023], such that optimizing blindly on aggregated metrics like Dice under biased labels actively forces the model to encode demographic prejudice [Liu et al., 2024]. By introducing directional metrics to detect label bias, our framework reorients the focus from model-centric fairness toward data-centric bias auditing.

Confident Learning & Dataset Auditing. Confident Learning (CL) is a dataset auditing framework [Northcutt et al., 2021] utilizing the joint distribution over noisy and true labels. CL has, e.g.,

been applied to highlight annotation errors for review [Lad and Mueller, 2023], and also for medical image segmentation [Zhang et al., 2020b, Pethmunee et al., 2025], but most work considers uniform quality issues targeting general robustness. An exception is Li et al. [2023], which uses CL as a bias mitigation tool for classification. Here, we utilize CL to solve a far more challenging problem: Diagnosing label bias in image segmentation, without knowing the nature of our underlying label bias, by projecting the joint distribution matrix calculated by CL onto directional axes to isolate subgroup-conditional bias in segmentation.

3 Problem Setup and Bias Simulation

We focus on *subgroup-conditional label bias* – hereafter *label bias*, where systematic discrepancies between observed and intended annotations occur disproportionately across demographic subgroups.

Notation and Definitions. Let data $\mathcal{D} = \{(x_i, y_{\text{obs},i}, g_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{H\times W\times 3}$, $y_{\text{obs},i} \in \{0,1\}^{H\times W}$ is the observed annotation, and $g_i \in \mathcal{G} = \{g_c, g_b\}$ denotes the selected reference and potentially biased groups. The latent intended mask is denoted by $y_{\text{true},i}$, with foreground set $Y_{\text{true},i}$. It is available only in controlled experiments and is not used by the proposed audit or mitigation procedure. Pixels are indexed by $p \in \Omega = \{1,\dots,H\} \times \{1,\dots,W\}$, and $a \in \{0,1\}$ indexes the background and foreground classes. For $p \in \Omega$, $y_{\text{obs},i,p}$ denotes the pixel label, and $Y_{\text{obs},i} = \{p \in \Omega : y_{\text{obs},i,p} = 1\}$ its foreground set. Feature channels are indexed by k, and $\mathbf{1}[\cdot]$ denotes the indicator function.

Sample-level Label Bias. The bias ratio $\beta \in [0,1]$ denotes the proportion of samples in g_b whose entire masks are corrupted, rather than the proportion of corrupted pixels (i.e, at sample level, maintaining consistency with label bias definitions in classification fairness literature [Jørgensen et al., 2026]).

Simulating Morphological Bias. We draw on evidence from medical image analysis, where the most prevalent form of label bias is systematic under- or over-segmentation of anatomical structures [Benčević et al., 2024, Pesce et al., 2020]. When annotators encounter images with lower visual contrasts, they frequently compensate by drawing overly conservative or overly expansive boundaries, e.g. in specific skin tones in dermatology [Benčević et al., 2024], varying tissue densities in radiology [Pesce et al., 2020], or delineating images under the influence of external conditions [Weltens et al., 2001], e.g., low lighting, geographic regions [Schumann et al., 2023].

In controlled experiments, we replicate: (i) *dilation* that expands boundaries, simulating over-annotation; (ii) *erosion* that contracts boundaries, simulating under-annotation; (iii) *HBD* produces spatially varying mixed over- and under-annotation, simulating observed non-uniform annotation inconsistencies (details in Appendix A.1). All morphological operations are parameterized by a radius r_d, applied to g_b at a sample-level bias rate β to reproduce an approximation of real-world label bias whose severity can be systematically varied.

4 A Framework for Auditing and Mitigating Label Bias in Segmentation

The proposed framework addresses two stages: *auditing* label bias directly from the training data, and *mitigation*, which corrects for the detected bias during model training. Neither stage accesses y_{true} during operation; mitigation additionally assumes that a reference annotation style can be selected.

4.1 Auditing Label Bias via Confident Learning

We adapt Confident Learning (CL) [Northcutt et al., 2021] to audit group-conditional annotation disagreement in segmentation. The idea behind CL is to identify data points (pixels) for which model predictions $\hat{y}_{i,p}$ *confidently* disagree with the annotated labels $y_{obs,i,p}$. To this end, model predictive probabilities $\widehat{p}_{i,p}(a)$ are estimated using a K-fold cross-fitting procedure. Confident predictions are those over the class-specific confidence threshold, corresponding to the empirical mean probability of that class for points belonging to that class in

$$P_a = \{(i,p) : y_{\text{obs},i,p} = a\} \quad \widehat{\tau}_a = \frac{1}{|\mathcal{P}_a|} \sum_{(i,p)\in\mathcal{P}_a} \widehat{p}_{i,p}(a). \tag{1}$$

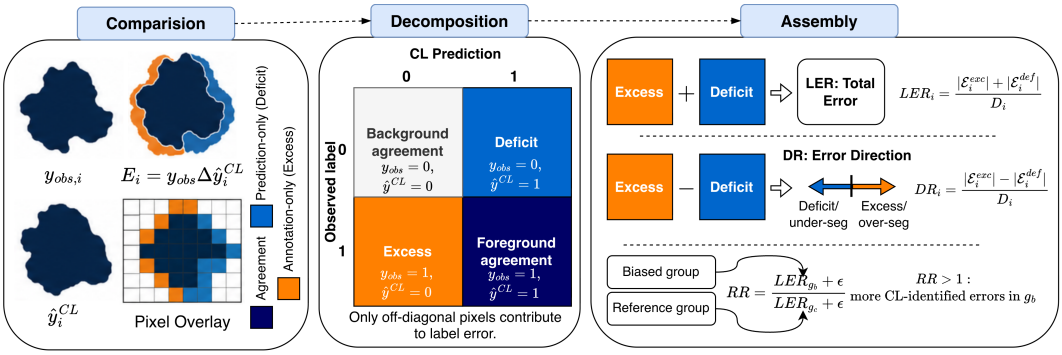


Figure 1: **CL-based decomposition of annotation disagreement.** Observed labels y_{obs} are compared with confident assignments derived from cross-fitted model predictions.

Each pixel is then either assigned to the class with the highest model probability that also exceeds the class confidence threshold, or falls back to $\arg\max$ if no probability exceeds the threshold:

$$\widehat{y}_{i,p}^{\text{CL}} = \begin{cases} \underset{a \in \{0,1\}: \widehat{p}_{i,p}(a) \geq \widehat{\tau}_a}{\arg\max} \ \widehat{p}_{i,p}(a), & \text{if } \max_a \left(\widehat{p}_{i,p}(a) - \widehat{\tau}_a\right) \geq 0, \\ \underset{a \in \{0,1\}}{\arg\max} \ \widehat{p}_{i,p}(a), & \text{otherwise.} \end{cases} \tag{2}$$

We can then define a "confident error set" made up of those pixels in an image with confident prediction-annotation disagreements:

$$\mathcal{E}_i = \left\{ p \in \Omega : \widehat{p}_{i,p}(\widehat{y}_{i,p}^{\text{CL}}) \geq \widehat{\tau}_{\widehat{y}_{i,p}^{\text{CL}}} \text{ and } y_{\text{obs},i,p} \neq \widehat{y}_{i,p}^{\text{CL}} \right\}, \quad \widehat{Y}_i^{\text{CL}} = \left\{ p \in \Omega : \widehat{y}_{i,p}^{\text{CL}} = 1 \right\}. \tag{3}$$

Magnitude of estimated label bias. To prevent background pixels and large objects from dominating the audit, we use the foreground union of each image, $U_i = \max\left\{1, \left|Y_{\text{obs},i} \cup \widehat{Y}_i^{\text{CL}}\right|\right\}$, as a normalization term. The image-level Confident Label Error Rates (LER) and group LER are thus:

$$\widehat{\text{LER}}_i = \frac{|\mathcal{E}_i|}{U_i}, \qquad \widehat{\text{LER}}_g = \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} \widehat{\text{LER}}_i, \qquad \mathcal{I}_g = \{i : g_i = g\}. \tag{4}$$

Averaging image-level rates ensures that each image contributes equally. LER therefore measures foreground-normalized prevalence of potential label bias.

Subgroup concentration. We quantify the absolute and relative concentration of disagreements as

$$\widehat{\Delta}_{\text{LER}} = \widehat{\text{LER}}_{g_b} - \widehat{\text{LER}}_{g_c}, \qquad \widehat{\text{RR}} = \frac{\widehat{\text{LER}}_{g_b} + \varepsilon_{\text{RR}}}{\widehat{\text{LER}}_{g_c} + \varepsilon_{\text{RR}}}, \tag{5}$$

where $\varepsilon_{\text{RR}} > 0$ is a fixed numerical stabilizer. Positive Δ_{LER}, equivalently $\text{RR} > 1$, indicates greater CL-identified label bias in g_b.

Direction of label bias. We partition the potential errors according to whether the observed mask contains foreground excess or deficit relative to CL assignments:

$$\mathcal{E}_i^{\text{exc}} = \left\{ p \in \mathcal{E}_i : y_{\text{obs},i,p} = 1, \widehat{y}_{i,p}^{\text{CL}} = 0 \right\}, \tag{6}$$

$$\mathcal{E}_i^{\text{def}} = \left\{ p \in \mathcal{E}_i : y_{\text{obs},i,p} = 0, \widehat{y}_{i,p}^{\text{CL}} = 1 \right\}. \tag{7}$$

Foreground excess is consistent with annotation expansion, whereas foreground deficit is consistent with annotation contraction, relative to the (unknown) true annotation. We define Directional Rate (DR) as

$$\widehat{\text{DR}}_g = \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} \widehat{\text{DR}}_i \quad \text{where} \quad \widehat{\text{DR}}_i = \frac{|\mathcal{E}_i^{\text{exc}}| - |\mathcal{E}_i^{\text{def}}|}{U_i}, \tag{8}$$

$\widehat{\mathrm{DR}}_g$ lies in $[-1, 1]$: positive values indicate foreground excess, negative values indicate foreground deficit, and values near zero indicate either little directional disagreement or cancellation between the two error types. For direct interpretation, the corresponding raw group-level counts are

$$C^g_{\mathrm{exc}} = \sum_{i \in \mathcal{I}_g} |\mathcal{E}^{\mathrm{exc}}_i|, \qquad C^g_{\mathrm{def}} = \sum_{i \in \mathcal{I}_g} |\mathcal{E}^{\mathrm{def}}_i|. \tag{9}$$

These counts are comparable across methods only when computed over identical evaluation images, resolutions, and folds. To summarize, for label errors, $\widehat{\mathrm{LER}}$ estimates the magnitude, $\widehat{\mathrm{RR}}$ identifies their subgroup concentration, and $\widehat{\mathrm{DR}}$ estimates their direction (under- or over-segmentation). Without independently verified masks, these remain proxy estimates and may also reflect systematic errors or miscalibration in the cross-fitted model. We therefore estimate uncertainty by bootstrapping images, or patients when multiple images belong to the same individual.

4.2 Mitigating Label Bias via Subgroup Conditioning

We model group-conditional label bias as an annotation-style shift superimposed on shared anatomical structure [Zepf et al., 2023]. Motivated by the increased subgroup separability observed under controlled corruption in Section 5.1, our method represents group-specific labeling styles during training and directs all predictions toward a selected reference style at inference. It combines FiLM-based subgroup conditioning with asymmetric masking of potentially biased boundary supervision. We first describe the proposed method and then the comparison baselines.

4.2.1 Subgroup-Conditioned Decoder (SCD)

We use FiLM [Perez et al., 2018] to model group-specific annotation styles. Let $f_i = e_\theta(x_i)$ be the encoder features, with channel $k \in \{1, \dots, K_f\}$. An embedding of g_i produces channel-wise scale and shift parameters $(\gamma_{g_i,k}, \delta_{g_i,k})$, which are used to adapt the encoding $f'_{i,k} = \gamma_{g_i,k} f_{i,k} + \delta_{g_i,k}$, finally yielding $\widehat{y}_i = d_\phi(f'_i)$, where d_ϕ is the shared decoder. During inference, every sample is conditioned on the selected reference group: $\widehat{y}^{\mathrm{ref}}_i = d_\phi(\gamma_{g_c} \odot e_\theta(x_i) + \delta_{g_c})$. This expresses all predictions in the reference annotation style without requiring the input's group identity at inference.

Asymmetric Boundary Masking and Objective. Let $B_{i,p} = [b_w(y_{\mathrm{obs},i})]_p \in \{0, 1\}$ denote a boundary band of width w extracted by the morphological operator b_w. We assign $W_{i,p} = 1 - \mathbf{1}[g_i = g_b] B_{i,p}$, removing boundary supervision from g_b while retaining its interior pixels of the target and all supervision from g_c. FiLM models group-specific annotation variation, while boundary masking prevents the potentially biased group from determining the shared boundary representation. The resulting objective is

$$\mathcal{L}_{\mathrm{mask}} = \frac{\sum_{i=1}^{N} \sum_{p \in \Omega} W_{i,p}\, \ell_{\mathrm{seg}}(\widehat{y}_{i,p}, y_{\mathrm{obs},i,p})}{\sum_{i=1}^{N} \sum_{p \in \Omega} W_{i,p}}, \tag{10}$$

where ℓ_{seg} is the pixel-wise segmentation loss.

Automated reference-group selection (Auto-SCD). When group identities are known but the reference group is not, we estimate it from the first E_w epochs. Motivated by training-dynamics diagnostics and small-loss selection [Swayamdipta et al., 2020, Li et al., 2020], we select

$$\widehat{g}_c = \arg\min_{g \in \mathcal{G}} \frac{1}{E_w |\mathcal{I}_g|} \sum_{e=1}^{E_w} \sum_{i \in \mathcal{I}_g} \ell_i^{(e)}. \tag{11}$$

Training then applies Eq. 10 with $\widehat{g}_c$ as the reference group, and inference fixes the FiLM condition to $\widehat{g}_c$. Because low loss may reflect intrinsic task difficulty rather than annotation quality, we evaluate this heuristic separately from a manually selected reference for SCD. Pseudocode in Appendix C.1.4.

Mitigation Baselines (i) **Feature alignment** matches group representations using MMD [Gretton et al., 2006] or Deep CORAL [Sun and Saenko, 2016] or adversarial debiasing, which uses a GRL classifier to jointly minimize segmentation loss and suppress group predictability [Ganin et al., 2016,

Dataset	Data	Label variation	Group construction	Setting
CelebAMask-HQ (CMask) [Lee et al., 2020]	30,000 face images with semantic masks and a binary male/female attribute	Boundary-localized erosion at $\beta \in \{0, 25, 50, 75, 100\}\%$, following Section 3	g_c: males with unmodified masks. g_b: females, with a proportion β receiving eroded masks (r_d = 15, unless otherwise specified)	Controlled: error location, prevalence, and direction are known
PhC-U373 [Maška et al., 2014, Ulman et al., 2017]	651 glioblastoma-cell images with two annotation styles	Style 0 provides detailed, tight boundaries; style 1 provides coarser, expanded boundaries	g_c: untinted images retaining style 0. g_b: synthetically tinted images, with a proportion β receiving style 1 annotations	Semi-synthetic: natural label styles with subgroup assignment
IMA++ [Abhishek et al., 2026]	2,394 dermoscopic images (ISIC derived) with two to five annotations per image; skin tone is estimated as ST1–ST4	Naturally occurring detailed/tight and loose/expanded annotation styles across annotators	g_c: lighter-tone images (ST1-2). g_b: darker-tone images (ST3-4). At β = 0, natural bias is retained. At $\beta \in \{50, 100\}\%$, β of g_b is reassigned annotations from loose annotators, while g_c unchanged	Natural annotations with a semi-synthetic group–style association
ISIC [Codella et al., 2018]	Dermoscopic lesion images with one observed segmentation mask per image; skin tone is estimated as ST1–ST4	Naturally occurring subgroup-associated differences in observed segmentation masks; no alternative clean or reference annotation is available	g_c: lighter-tone images (ST1–ST2). g_b: darker-tone images (ST3–ST4). No annotation-style reassignment or synthetic corruption is applied	**Exploratory**: observed disparities cannot be uniquely attributed to label error

Table 1: **Datasets and group construction to audit label bias.** Further detailed in Appendix B.2.

Zhang et al., 2018]. (ii) **Fairness regularization** penalizes group gaps in mean foreground probability (Demographic Parity, DP) [Dwork et al., 2012] or pixel-level true- and false-positive rates computed against the observed masks (Equalized Odds, EO) [Hardt et al., 2016]. (iii) **Noise-robust methods** that treat label bias as label noise, including Generalized Cross Entropy (GCE) [Zhang and Sabuncu, 2018] and Bootstrapping [Reed et al., 2014]. Implementation details are given in Appendix C.1.

4.3 Dataset and Experimental Setup

We evaluate our framework on three datasets that progress from controlled label bias to naturally occurring annotation variability. Table 1 summarizes the role of each dataset and the construction of target subgroups. We additionally use the ISIC [Codella et al., 2018] dataset as exploratory, because of the lack of true annotations for evaluation.

All models use a U-Net architecture with a pretrained ResNet-34 backbone (ResNet-18 & SegFormer `mit-b2` for ablation) on multiple RTX 5000 (16GB). Networks are trained for 15 epochs using the AdamW optimizer with an objective of Cross-Entropy and Dice loss. For the unsupervised auto-conditioning phase (Section 4.2.1), we set E_w = 5 (specified otherwise) epochs to allow stable subgroup loss separation before applying FiLM features. Boundary maps $B(x)$ for the asymmetric spatial loss are extracted on-the-fly via morphological gradient filters with kernel radius k_r = 2. All experiments use 5-fold cross-validation to calibrate CL thresholds τ_c. We validate segmentation quality using standard Dice and IoU scores (in %). Algorithmic fairness is assessed via the subgroup performance gap (we useΔ = g_c − g_b, to denote subgroup differences). The label bias audit uses the CL-derived metrics to assess the true nature of errors (Section 4.1). For analysis using learned representations, we extract globally pooled features from the encoder bottleneck and quantify how strongly they encode subgroup membership. Full evaluation details are provided in B.1.

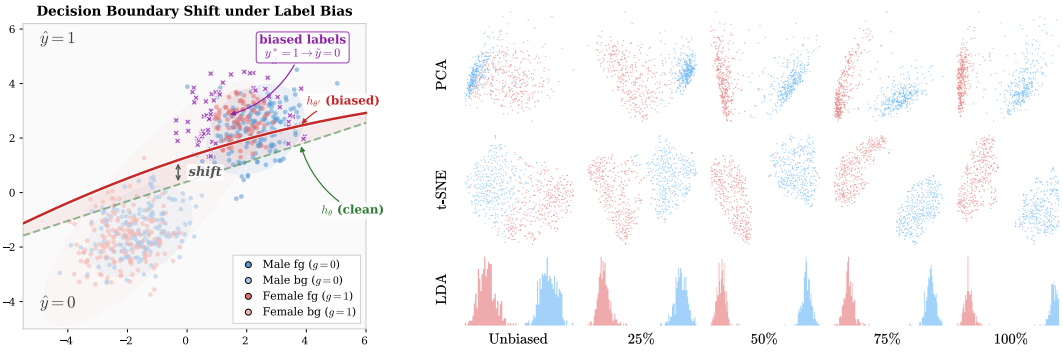


Figure 2: (Left) Biased labels (purple) shift the learned decision boundary $h_{\theta'}$ away from the clean boundary h_θ, producing systematically incorrect segmentation during inference. (Right) Effects of increasing label bias on subgroup separability in the encoder latent space (CelebAMask-HQ). Red: g_b; blue: g_c. Rows show PCA (top), t-SNE (middle), and LDA (bottom).

5 Results and Discussion

5.1 Effect of Label Bias on Learned Representations

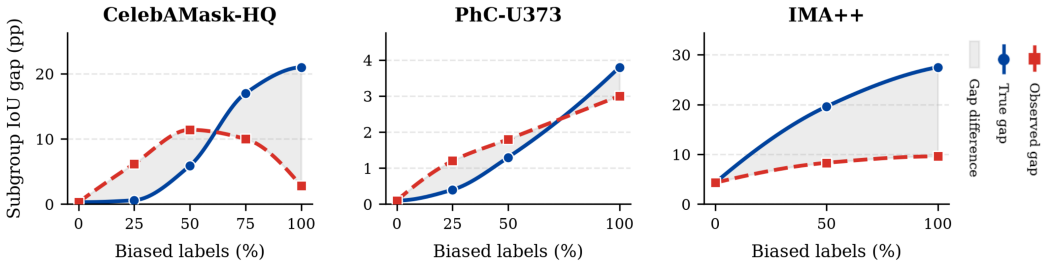
When membership of the group g_b is encoded in the data, label bias degrades boundary delineation for the affected subgroup (g_b) while leaving g_c performance largely unaffected [Vorontsov and Kadoury, 2021, Li and Wang, 2026], as illustrated in Fig 2 (Left). Jones et al. [2023] demonstrated that if groups are not well separable in feature space, the model cannot limit the bias to just g_b, and errors may propagate across both groups, creating a large fairness gap. This shows that **subgroup separability governs whether label bias appears as a subgroup disparity or an overall performance degradation**, which may be indistinguishable through group-level metrics.

Through controlled experiments, we show that as β increases from 0% to 100%, the bottleneck Maximum Mean Discrepancy (MMD) rises from 0.10 to 0.29 and centroid distance from 7.54 to 19.04 (Fig. 2 (Right)), alongside the PCA, t-SNE, and LDA visualizations. The separation at $\beta = 0$ might reflect genuine visual differences, whereas growth under increasing systematic corruption indicates that label bias amplifies subgroup separation (see further analysis in Appendix B.1). This extends prior classification evidence that systematic label bias encourages models to encode sensitive attributes along with subgroup separability [Stanley et al., 2026, Jones et al., 2023].

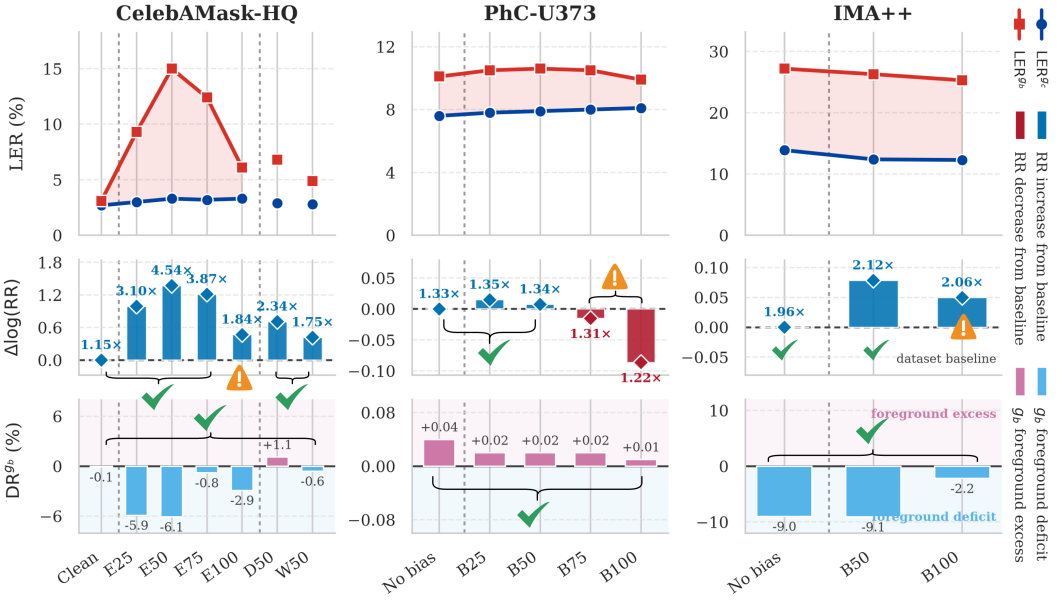
Subgroup separability by itself is *neither a fairness violation nor an undesired property* [Lahoti et al., 2020, Petersen et al., 2023]. In domains such as medical imaging, where disease manifestation may differ across populations and relevant structures may genuinely differ across subgroups, enforcing marginal representational invariance can remove useful information, implying the independence fairness criterion [Petersen et al., 2023, Barocas et al., 2023]. Our results identify label bias as an underappreciated confounder of such separability, and suppressing the representations via feature alignment and fairness objectives to treat label bias does not outperform the ERM baseline and can cause *leveling down* [Zong et al., 2023, Zietlow et al., 2022]. We therefore treat increased separability as a supervision signal and model this variation before mapping all outputs to a selected reference annotation style in Section 4.2.1.

5.2 Standard Metrics Fail to Detect Label Bias

As shown in Fig. 3 (Top), across all datasets and conditions, **IoU consistently mistakes the magnitude of true subgroup disparity when evaluated against observed labels**. At peak, on CelebA $\beta = 50\%$, the IoU gap for observed performance is 11.4%, which reduces to 10.0% at $\beta = 75\%$, suggesting improved fairness, while the true gap shows severe and increasing disparity ($5.5\% \to 17.0\%$). This demonstrates the *biased ruler effect*: metrics evaluated against corrupted labels not only fail to detect bias but also actively hide it [Parikh et al., 2026]. A similar trend is also observed on PhC-U373 (at $\beta = 100\%$) and IMA++, where observed IoU remains nearly constant even as label corruption increases.



(a) **Evaluation against clean (true) and observed masks.** Subgroup IoU gaps differ depending on the annotation quality, highlighting *the biased ruler effect* and how label bias can skew fairness assessment.



(b) **Auditing with observed masks only.** $\widehat{\text{LER}}$ measures suspected annotation-error prevalence, $\widehat{\text{RR}}$ its relative concentration in g_b, and $\widehat{\text{DR}}$ the direction of disagreement between annotations and confident predictions.

Figure 3: **Label-bias evaluation and auditing.** Clean reference masks are used only for experimental validation; the proposed CL audit requires no clean ground-truth labels.

The observed gap *overestimates* the true gap for dilation (at $\beta = 50\%$ Δ_{obs}=8.28 vs. Δ_{true}=1.65), while underestimating it for erosion at higher β. This **asymmetry is a fundamental property of Dice**: erosion removes true positives, which appear in both the numerator and denominator $\frac{2\text{TP}}{2\text{TP+FP+FN}}$, whereas dilation adds false positives that inflate only the denominator. Thus, the same number of corrupted pixels produces a larger drop in Dice under erosion than dilation, making observed metrics unreliable for comparing bias severity.

> *This calls for better evaluation metrics, as any fairness intervention based on closing this observed gap could therefore increase the bias, not reduce it.*

5.3 Auditing Group-Conditional Label Bias Without Ground-Truth

Fig. 3 summarizes the audit and CL-metrics across all datasets under various conditions. Detailed results for the audit are in Appendix D.1. **While our primary contribution is auditing without ground-truth, a small set of clean labels for validation provides direct confirmation of detected bias; refer to Appendix F.1 and Table 7.**

On CelebAMask-HQ, LER^{g_c} remains stable between 2.7% and 3.3%, whereas LER^{g_b} increases from 3.1% without corruption to 9.3% and 15.0% at 25% and 50% erosion. Consequently, $\log(\text{RR})$

Table 2: **Comparison of label-bias mitigation methods.** For each setting, $(\mathrm{IoU}/\Delta)_{\text{true}}$ reports mean IoU / subgroup IoU gap (Δ) against true-reference masks (in % and percentage points; we use 'true' instead of true-reference mask; detailed annotations for IMA++). CL-metrics report foreground excess/deficit counts (E/D in millions of pixels). Rows with **Ours** are in gray.

		CelebA (β = 50%)		**CelebA** (β = 100%)		**PhC-U373**	**IMA++** (β = 50%)
Category	**Method**	$(\mathbf{IoU}/\Delta)_{\mathbf{true}}$	**CL (E/D)**	$(\mathbf{IoU}/\Delta)_{\mathbf{true}}$	**CL (E/D)**	$(\mathbf{IoU}/\Delta)_{\mathbf{true}}$	$(\mathbf{IoU}/\Delta)_{\mathbf{true}}$
Reference	ERM	93.1/5.5	39.3/86.4	82.5/21.0	9.0/26.9	91.2/4.4	81.7/19.6
Fairness	EO	93.0/5.7	33.1/79.7	82.6/21.0	9.0/26.9	91.9/3.6	79.8/13.9
	DP	93.0/5.6	34.0/78.5	82.6/21.0	8.9/26.9	92.0/3.4	79.6/13.6
	EO+DP	92.8/5.8	35.1/77.1	82.4/20.9	10.4/26.7	91.9/3.5	79.9/13.1
Alignment	Adversarial	92.4/6.7	37.0/76.9	82.7/21.1	3.6/10.7	91.8/3.7	80.3/13.9
	MMD	92.8/6.3	34.2/81.2	82.8/20.5	3.7/12.3	91.7/3.9	80.0/13.0
	CORAL	93.2/5.5	32.3/80.6	82.6/20.9	8.7/26.9	91.7/3.9	79.7/13.9
Noise	GCE	92.3/6.6	36.9/77.6	82.5/21.3	9.2/26.0	91.0/3.9	80.0/14.6
	Bootstrapping	92.8/5.8	42.0/83.9	82.5/21.1	8.9/26.8	91.0/3.7	80.2/14.7
	Asym. mask	91.5/7.9	39.6/61.0	79.8/24.7	3.6/31.0	90.9/5.0	79.7/14.6
Ours	SCD	96.8/0.5	10.4/95.4	89.0/8.9	2.5/92.7	91.8/3.6	85.8/6.7
	SCD+Asym	**96.8/0.5**	2.5/23.7	**89.2/9.0**	2.6/93.3	91.1/4.8	86.0/7.0
	Auto-SCD	93.7/4.2	28.2/86.8	85.3/16.1	4.6/45.2	**92.5/3.2**	**87.4/6.2**

stay > 0, correctly identifying that suspected errors are concentrated in g_b. Erosion also produces negative DR, reaching -6.1% at 50% corruption, while dilation produces positive DR ($+1.1\%$). The audit therefore detects both the subgroup concentration and expected direction of label errors. However, at complete erosion, RR decreases, and DR weakens to -2.9%. This attenuation is a fundamental limitation of CL-based auditing without ground-truth, which remains effective for moderate bias, the most common real-world case, where bias is systematic but not whole.

PhC-U373 presents a very subtle case of labeling bias (affecting just ≈4% of pixels) due to visual domain shift. The LER of both groups remains nearly constant, while LER^{g_b} changes from 10.1% to 9.9%. DR remains positive, indicating over-segmentation of g_b. Thus, the audit identifies a persistent baseline difference between the subgroups. **IMA++** exhibits a large disparity before additional corruption: baseline LER is 13.9% for g_c and 27.2% for g_b, corresponding to $\mathrm{RR} = 1.96$, and DR ≤ 0, indicating under-segmentation of g_b. The baseline-adjusted RR signal, along with baseline LER & DR, indicates that pre-existing differences in image difficulty or annotation style dominate the additional corruption signal. **On ISIC,** with no artificial bias, the audit detected a naturally elevated error rate, validating our evaluation on a real-world clinical dataset with no clean validation.

CL metrics such as LER and RR can identify label bias when the corrupted pattern remains distinguishable from the underlying image signal. At the same time, DR can recover whether annotations exhibit foreground deficit or excess. Their sensitivity decreases when the proxy model learns systematic errors or when pre-existing subgroup differences dominate prediction-annotation disagreement. The metrics should therefore be interpreted as indicators of potential label bias rather than direct estimates of true annotation error.

5.4 Mitigating Bias via Subgroup Conditioning (SCD)

Under label corruption, fairness constraints, alignment methods, and standard label noise methods fail to close the fairness gap, as shown in the results in Table 2. The proposed SCD methods close the IoU gap to 8.9% from 21.0% at baseline, for $\beta = 100\%$. Ablations in Appendix G.4.

These results help distinguish label bias from conventional fairness problems and label noise. Systematic label corruption is much more complex because it corrupts the (learned) input representation, while traditional fairness approaches (EO/DP regularization) constrain the model outputs, which do not target the source of subgroup boundary shifts. Likewise, alignment methods try to suppress (potentially crucial) subgroup information via invariance. Our approach, instead, retains subgroup information and applies a cleaner reference condition at inference.

Even though subgroup discovery identified clean reference labels in all conditions (see Appendix G.5 for robustness test for Auto-SCD), **we recommend** auditing for label bias (Section 4.1) and then referencing the clean(er) subgroup after inspection or expert validation. Auto-conditioning is particularly useful when the reference group is unknown.

6 Limitation and Intended Scope

As for our limitations, the CL-based audit estimates prediction-annotation disagreement as a signal for label bias. Without independently verified annotations, disagreement may also reflect model error, miscalibration, or genuine differences in task difficulty. These disagreements underestimate severe bias when models learn the corrupted annotations (like in our $\beta = 100\%$ condition). The mitigation method assumes that a subgroup with a different annotation style can be identified. The training loss heuristic in auto-discovery (Section 4.2.1) may reflect differences in imaging difficulty or anatomy. If all groups are biased, the conditioning may reproduce, as expected, undesirable biased predictions (Appendix G.3). Reference selection should therefore be supported by expert assessment or a verified subset. We test and report counterfactual and calibration ablations, the no-biased-subgroup case, and model-capacity dependence in Appendix G.3. Finally, label bias simulation (Section 3) captures only a subset of possible annotation error types, despite our efforts to make them realistic and complex. Most segmentation datasets lack verified clean masks, like our exploratory dataset, ISIC; therefore, we rely on controlled simulation and use verified masks for validation, where possible.

7 Conclusion

We introduce a framework for simulating and auditing label bias for segmentation without ground-truth labels. Our proposed subgroup-conditioned (SCD) mitigation strategy reduces subgroup disparities, where traditional debiasing and denoising methods fail. We also address a blind spot in algorithmic fairness in image segmentation, showing that biased labels can both distort fairness evaluation and learned representations. Our proposed CL-metrics detect subgroup concentration and direction of these errors, where standard segmentation metrics fail, possibly presenting a misleading narrative.

Future work

We focus on directly addressing current limitations by extending this framework to measure under severe bias situations, relaxing the need to know subgroup membership for mitigation, reliability of training dynamics on identifying label bias and validation on datasets with more naturally occurring label bias.

Disclosure of Interests

The authors have no competing interests to declare.

Acknowledgment

This work was funded by the Novo Nordisk Foundation (project number 0087102).

Generative AI Disclosure

Following fair and responsible use of AI for research, we list the use of generative AI for the following sub-tasks:

- **Major use:** Formatting and designing LaTeX tables for results (manually audited), writing boilerplate code for pipelines (verified against official implementations), restructuring appendix sections for readability, and compiling a code repository.
- **Minor use:** Inspiration for figure/plot designs, polishing select paragraphs in the appendix for clarity, and experiment logging.
- **Not used for:** Ideation, developing the methodology, writing the main text (introduction, methods, results, and discussion), mathematical proofs, the paper's structure, or the bibliography (AI used for verifying references).

The authors manually verified all AI-assisted outputs. The authors take full responsibility for the final output and claims in this work.

References

Kumar Abhishek, Jeremy Kawahara, and Ghassan Hamarneh. Descriptor: Isic archive multi-annotator dermoscopic skin lesion segmentation dataset (ima++). *IEEE Data Descriptions*, 3:367–378, 2026. doi: 10.1109/IEEEDATA.2026.3689801.

Olalekan Joseph Akintande, Siavash Bigdeli, and Aasa Feragen. Medicine after death: Xai and algorithmic fairness under label bias. In Hilde Weerts, Mykola Pechenizkiy, Doris Allhutter, Ana Maria Corrêa, Thomas Grote, and Cynthia Liem, editors, *Proceedings of Fourth European Workshop on Algorithmic Fairness*, volume 294 of *Proceedings of Machine Learning Research*, pages 171–186. PMLR, 30 Jun–02 Jul 2025. URL `https://proceedings.mlr.press/v294/akintande25a.html`.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL `https://arxiv.org/abs/1610.01644`.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In *Proceedings of the 14th linguistic annotation workshop*, pages 60–73, 2020.

Marin Benčević, Marija Habijan, Irena Galić, Danilo Babin, and Aleksandra Pižurica. Understanding skin color bias in deep learning-based skin lesion segmentation. *Computer methods and programs in biomedicine*, 245:108044, 2024.

Maximilian Bernhard, Roberto Amoroso, Yannic Kindermann, Lorenzo Baraldi, Rita Cucchiara, Volker Tresp, and Matthias Schubert. What's outside the intersection? fine-grained error analysis for semantic segmentation beyond iou. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 968–977, 2024.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

Richard J Chen, Judy J Wang, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6):719–742, 2023.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018*

IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pages 168–172, 2018. doi: 10.1109/ISBI.2018.8363547.

Jessica Dai and Sarah M Brown. Label bias, label shift: Fair machine learning with unreliable labels. In *NeurIPS 2020 Workshop on Consequential Decision Making in Dynamic Environments*, volume 12, 2020.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7 (2):179–188, 1936.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

Maryam Amir Haeri and Katharina Anna Zweig. The crucial role of sensitive attributes in fair classification. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2993–3002, 2020. doi: 10.1109/SSCI47803.2020.9308585.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Nicholas Heller, Joshua Dean, and Nikolaos Papanikolopoulos. Imperfect segmentation labels: How much do they matter? In *International Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis*, pages 112–120. Springer, 2018.

Harold Hotelling. The generalization of student's ratio. In *Breakthroughs in statistics: Foundations and basic theory*, pages 54–65. Springer, 1992.

Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International conference on artificial intelligence and statistics*, pages 702–712. PMLR, 2020.

Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In *International conference on machine learning*, pages 4804–4815. PMLR, 2020.

Charles Jones, Mélanie Roschewitz, and Ben Glocker. The role of subgroup separability in group-fair medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 179–188. Springer, 2023.

Frida Jørgensen, Nina Weng, and Siavash Bigdeli. Detecting labeling bias using influence functions. *arXiv preprint arXiv:2602.19130*, 2026.

Alain Jungo, Raphael Meier, Ekin Ermis, Marcela Blatti-Moreno, Evelyn Herrmann, Roland Wiest, and Mauricio Reyes. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 682–690. Springer, 2018.

Linus Juni, Aasa Feragen, and Aditya Parikh. False confidence: Automated labels confound fairness audits in cervical spine segmentation, 2026. URL `https://arxiv.org/abs/2607.07852`.

Burak Koçak, Andrea Ponsiglione, Arnaldo Stanzione, Christian Bluethgen, João Santinha, Lorenzo Ugga, Merel Huisman, Michail E Klontzas, Roberto Cannella, and Renato Cuocolo. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic and interventional radiology*, 31(2):75–88, 2025.

Vedang Lad and Jonas Mueller. Estimating label quality and errors in semantic segmentation data via any model. *arXiv preprint arXiv:2307.05080*, 2023.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Jinpeng Li and Han Wang. Noise-robust medical image segmentation via uncertainty-guided feature enhancement and adaptive noise-aware loss. *Biomedical Signal Processing and Control*, 120: 110073, 2026.

Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HJgExaVtwr`.

Yunyi Li, Maria De-Arteaga, and Maytal Saar-Tsechansky. Mitigating label bias via decoupled confident learning. *arXiv preprint arXiv:2307.08945*, 2023.

Zachary C Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43, 2018.

Bingyuan Liu, Jose Dolz, Adrian Galdran, Riadh Kobbi, and Ismail Ben Ayed. Do we really need dice? the hidden region-size biases of segmentation losses. *Medical Image Analysis*, 91:103015, 2024.

Martin Maška, Vladimír Ulman, David Svoboda, Pavel Matula, Petr Matula, Cristina Ederra, Ainhoa Urbiola, Tomás España, Subramanian Venkatesan, Deepak MW Balak, et al. A benchmark for comparison of cell tracking algorithms. *Bioinformatics*, 30(11):1609–1617, 2014.

S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.R. Mullers. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468)*, pages 41–48, 1999. doi: 10.1109/NNSP.1999.788121.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

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*, 2022:1–37, 2022. ISSN 2766-905X. doi: https://doi.org/10.59275/j.melba.2022-2d93. URL `https://melba-journal.org/2022:029`.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

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.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Aditya Parikh, Sneha Das, and Aasa Feragen. Investigating label bias and representational sources of age-related disparities in medical segmentation. In *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2026. doi: 10.1109/ISBI61048.2026.11515928.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

K Pesce, M Tajerian, MJ Chico, MP Swiecicki, B Boietti, MJ Frangella, and S Benitez. Interobserver and intraobserver variability in determining breast density according to the fifth edition of the bi-rads® atlas. *Radiología (English Edition)*, 62(6):481–486, 2020.

Eike Petersen, Enzo Ferrante, Melanie Ganz, and Aasa Feragen. Are demographically invariant models and representations in medical imaging fair? *arXiv preprint arXiv:2305.01397*, 2023.

Tanatorn Pethmunee, Supaporn Kansomkeat, Patama Bhurayanontachai, and Sathit Intajag. Confident learning-based label correction for retinal image segmentation. *Diagnostics*, 15(14):1735, 2025.

Laura Quinn, Konstantinos Tryposkiadis, Jon Deeks, Henrica CW De Vet, Sue Mallett, Lidwine B Mokkink, Yemisi Takwoingi, Sian Taylor-Phillips, and Alice Sitch. Interobserver variability studies in diagnostic imaging: a methodological systematic review. *The British Journal of Radiology*, 96 (1148):20220972, 2023.

Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

Annika Reinke, Minu D Tizabi, Carole H Sudre, Matthias Eisenmann, Tim Rädsch, Michael Baumgartner, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, et al. Common limitations of image processing metrics: A picture story. *arXiv preprint arXiv:2104.05642*, 2021.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Candice Schumann, Femi Olanubi, Auriel Wright, Ellis Monk, Courtney Heldreth, and Susanna Ricco. Consensus and subjectivity of skin tone annotation for ml fairness. *Advances in Neural Information Processing Systems*, 36:30319–30348, 2023.

Emma A. M. Stanley, Raghav Mehta, Mélanie Roschewitz, Nils D. Forkert, and Ben Glocker. Exploring the interplay of label bias with subgroup size and separability: A case study in mammographic density classification. In *Fairness of AI in Medical Imaging*, pages 74–83, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-05870-6.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 443–450, Cham, 2016. Springer International Publishing. ISBN 978-3-319-49409-8.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, 2020.

Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):29, 2015.

Vladimír Ulman, Martin Maška, Klas EG Magnusson, Olaf Ronneberger, Carsten Haubold, Nathalie Harder, Pavel Matula, Petr Matula, David Svoboda, Miroslav Radojevic, et al. An objective comparison of cell-tracking algorithms. *Nature methods*, 14(12):1141–1152, 2017.

Theresa A Ulrich, Elena M de Juan Pardo, and Sanjay Kumar. The mechanical rigidity of the extracellular matrix regulates the structure, motility, and proliferation of glioma cells. *Cancer research*, 69(10):4167–4174, 2009.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL `http://jmlr.org/papers/v9/vandermaaten08a.html`.

Théo Phan Van Song, Yue Qiao, Marco Vicic, and Anke Hadasch. A new method for skin color classification based on global cielab data and k-mean clustering. *Color Research & Application*, 51(1):e70012, 2026.

Eugene Vorontsov and Samuel Kadoury. Label noise in segmentation networks: mitigation must deal with bias. In *MICCAI Workshop on Deep Generative Models*, pages 251–258. Springer, 2021.

Hongqiu Wang, Wu Chen, Xiangde Luo, Zhaohu Xing, Lihao Liu, Jing Qin, Shaozhi Wu, and Lei Zhu. Toward fair and accurate cross-domain medical image segmentation: A vlm-driven active domain adaptation paradigm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24102–24112, 2025.

Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 526–536, 2021.

Caroline Weltens, Johan Menten, Michel Feron, Erwin Bellon, Philippe Demaerel, Frederik Maes, Walter Van den Bogaert, and Emmanuel van der Schueren. Interobserver variations in gross tumor volume delineation of brain tumors on computed tomography and impact of magnetic resonance imaging. *Radiotherapy and Oncology*, 60(1):49–59, 2001.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *CoRR*, abs/2105.15203, 2021. URL `https://arxiv.org/abs/2105.15203`.

Jiebin Yan, Yu Zhong, Yuming Fang, Zhangyang Wang, and Kede Ma. Exposing semantic segmentation failures via maximum discrepancy competition. *International Journal of Computer Vision*, 129(5):1768–1786, 2021.

Jiachen Yao, Yikai Zhang, Songzhu Zheng, Mayank Goswami, Prateek Prasanna, and Chao Chen. Learning to segment from noisy annotations: A spatial correction approach. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=Qc_OopMEBnC`.

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. URL `https://openreview.net/forum?id=wZ2SVhOTzBX`.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

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, 2020a.

Minqing Zhang, Jiantao Gao, Zhen Lyu, Weibing Zhao, Qin Wang, Weizhen Ding, Sheng Wang, Zhen Li, and Shuguang Cui. Characterizing label errors: confident learning for noisy-labeled image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 721–730. Springer, 2020b.

Yixuan Zhang, Boyu Li, Zenan Ling, and Feng Zhou. Mitigating label bias in machine learning: Fairness through confident learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16917–16925, 2024.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10421, 2022.

Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. Medfair: Benchmarking fairness for medical imaging, 2023. URL `https://arxiv.org/abs/2210.01725`.

A Appendix: Bias Construction and Morphological Operations

A.1 Additional Details: Morphological Operations & Effect of r_d

Refer to Figure 4 (a) for visualization of the representative erosion operation.

Harmonic Boundary Deformation (HBD). a superposition of H=3 sinusoidal harmonics with randomized frequencies ω_h, orientations α_h, and phases ψ_h:

$$d(p) = \frac{\rho}{H} \sum_{h=1}^{H} \sin\big(\omega_h \left(x_p \cos\alpha_h + y_p \sin\alpha_h\right) + \psi_h\big), \tag{12}$$

where ρ controls the maximum displacement amplitude. The altered mask is obtained by thresholding: $\tilde{y}_p = \mathbb{1}[\phi(p) > d(p)]$. This produces boundaries that locally oscillate between erosion and dilation, simulating the non-uniform annotation inconsistencies observed in practice.

B Representation Analysis and Datasets

B.1 Additional Details: Effect of Label Bias on Group Separability

Here, we investigate the effect of label bias on the learned representations (discussed in Section 5.1) in detail. The following is an empirical analysis for feature separability:

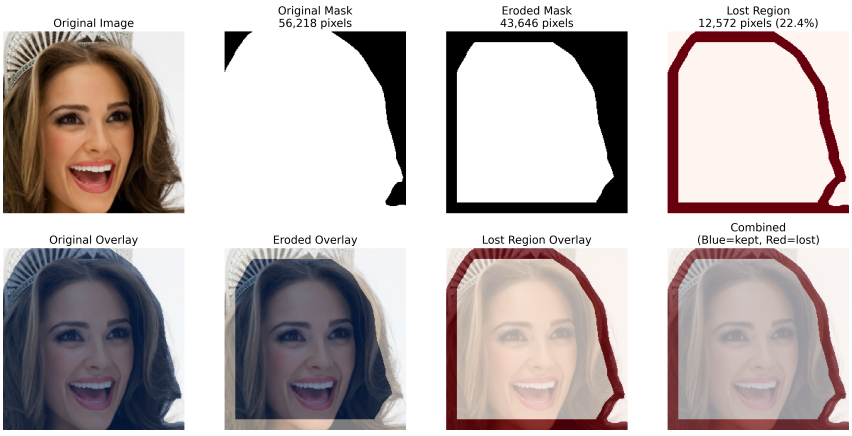
Feature Extraction. For each saved segmentation model checkpoint, we extracted representations from the final layer of the encoder (e.g, ResNet-34). To obtain a finite-length embedding vector for each image, we applied Global Average Pooling (GAP) to the spatial dimensions of the feature map, resulting in a 1-D bottleneck feature vector.

Visualizations. To visualize the high-dimensional feature embeddings, we use three standard projection techniques:

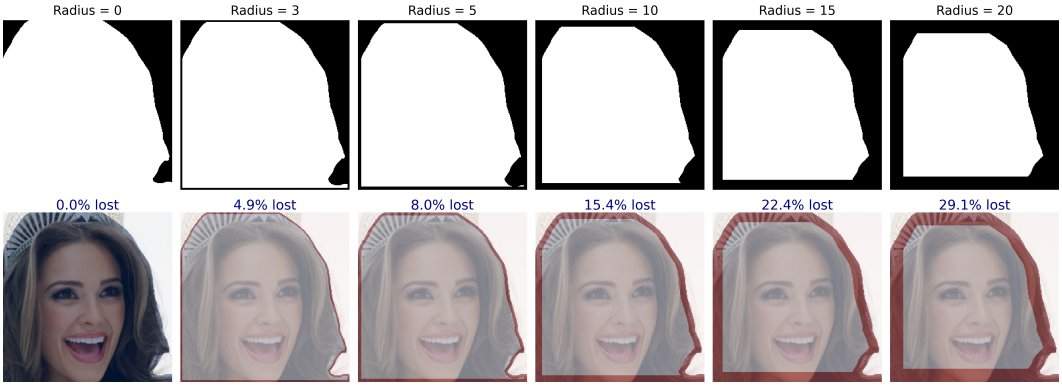
- **t-SNE [van der Maaten and Hinton, 2008]:** Computed with 2 components, perplexity set to 30, and initialized via PCA (Principal Component Analysis).
- **PCA [Pearson, 1901]:** Standard Principal Component Analysis projected to the first 2 principal components.
- **LDA [Fisher, 1936]:** Linear Discriminant Analysis projected to 1 component, showcasing the maximum linear separability between the subgroup distributions in feature space.

Separability Metrics To quantify the group separability effect in the feature space, we computed the following metrics:

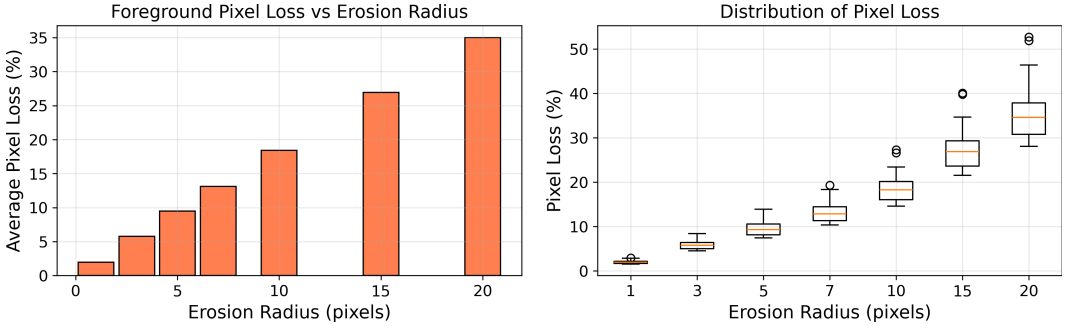
- **Linear Probe [Alain and Bengio, 2018] (Accuracy & AUROC):** We train a logistic regression classifier (LBFGS solver, max iterations = 1000) on the frozen GAP features to predict subgroup membership. We report the mean accuracy and Area Under the ROC Curve (AUROC) via 5-fold cross-validation.
- **Silhouette Score [Rousseeuw, 1987]:** Evaluates the cluster cohesion and separation based on subgroup labels using Euclidean distance.
- **Fisher's Discriminant Ratio [Mika et al., 1999]:** Measures the ratio of between-group variance to within-group variance across the feature dimensions. Statistical significance was verified using a permutation test (500 to 1000 permutations).
- **Maximum Mean Discrepancy (MMD) [Gretton et al., 2006]:** Computes the distributional distance between the subgroups' features using a multi-scale Gaussian (RBF) kernel.
- **Hotelling's T^2 Test [Hotelling, 1992]:** A multivariate statistical test for differences in the mean feature vectors (centroids) between the two subgroups. If the feature dimensionality exceeded sample size constraints, features were first reduced via PCA to a maximum of 50 dimensions before applying the test.



(a) **A visualization of the erosion operation.** The original ground-truth mask is reduced, isolating the 'Lost Region' (highlighted in red). The combined overlay demonstrates how the erosion strictly affects the boundary, mapping precisely to under-segmentation.



(b) The visual impact of scaling the erosion kernel radius (r_d) from 0 (Unbiased) to 20 pixels. The corresponding percentage of foreground pixel loss is tracked for the individual sample, highlighting the severity of boundary retraction at higher values.



(c) **Quantitative evaluation:** The bar chart on the left displays the average % of foreground pixel loss as a function of the r_e (erosion radius). The box plot on the right illustrates the distribution and variance of pixel loss.

Figure 4: Visual and statistical analysis of synthetic bias (morphological erosion) applied to the CelebAMask-HQ dataset.

Table 3: **Effect of label bias on representation separability (CelebAMask-HQ).** Separability increases with bias β, indicating that label bias amplifies subgroup separability beyond visual differences.

β	Supervised		Cluster		Distance		Sig.
	Acc.	AUC	Sil.	Fisher	MMD	Δ_c	p
0%	0.929	0.979	0.137	0.352	0.095	7.54	$< 10^{-16}$
25%	0.968	0.996	0.346	1.292	0.178	13.42	$< 10^{-16}$
50%	**0.980**	**0.998**	0.412	1.818	0.153	16.84	$< 10^{-16}$
100%	0.979	0.997	**0.536**	**3.100**	**0.292**	**19.04**	$< 10^{-16}$
× increase	1.05	1.02	3.91	8.81	3.07	2.52	

Refer to Table 3 for detailed separability analysis with varying β on CelebAMask-HQ and Table 4 on PhC-U373 and IMA++ dataset. We refrain from feature analysis on ISIC due to lack of clean verified labels.

The linear probe and five unsupervised metrics quantify how well the encoder's feature clusters subgroup membership. For **analysis on CelebAMask-HQ**, at the baseline (β=0), moderate separability exists due to genuine visual correlates of gender. As we induce label bias by increasing β, every metric increases: Silhouette score increases $3.9\times$ ($0.137 \rightarrow 0.536$), Fisher discriminant ratio $8.8\times$ ($0.35 \rightarrow 3.10$), and centroid distance $2.5\times$ ($7.54 \rightarrow 19.04$) from baseline to β=100%. This confirms the visual claims in Fig. 2 (Right) that label bias amplifies subgroup separability beyond just visual differences. All T^2 tests reject the null of equal group centroids ($p < 0.001$). **Results are consistent on PhC-U373 & IMA++ dataset**, where separability increases with β, with probe accuracy remaining constant at 1.0, also at baseline. This is expected, as we add a light synthetic color tint, which introduces a low-level domain shift that dominates the feature space. This makes these two groups perfectly separable even at baseline (i.e, at no added label bias), irrespective of annotation quality. The annotation-style bias induced due to annotator shift constitutes a small fraction of the representation (4% pixel shift), and its effects are masked by already well-saturated visual separation due to the synthetic tint added. **On IMA++**, subgroup separability increases with the induced label bias. From β = 0% to 100%, the silhouette score increases from 0.041 to 0.098 and the centroid distance from 6.366 to 9.629. The Fisher ratio is non-monotonic at β = 50%, but reaches its highest value of 1.074 at β = 100%. These results suggest that annotation style, along with skin-tone group, strengthens the encoding of group-related information, although the low absolute silhouette scores indicate overlap between the subgroup representations.

Table 4: **Feature Separability Analysis across Bias Levels.** Evaluation of the encoder's ability to separate demographic subgroups in the latent feature space for PhC-U373 and IMA++ datasets. Higher values indicate stronger subgroup separability in the learned feature space.

Bias Level (β)	**Silhouette**	**Fisher Ratio**	**Centroid Dist.**
PhC-U373			
0%	0.217	0.614	19.631
25%	0.209	0.555	19.102
50%	0.208	0.517	19.648
75%	0.192	0.495	18.006
100%	0.259	0.718	21.549
IMA++			
0%	0.041	0.868	6.366
50%	0.086	0.798	8.443
100%	0.098	1.074	9.629

B.2 Additional Details: Extended Dataset Descriptions

B.2.1 PhC-U373

This dataset was part of the ISBI Cell tracking challenge [Ulrich et al., 2009, Ulman et al., 2017], which here gives us a unique opportunity to evaluate bias mitigation in the presence of multi-expert annotations. The specific dataset is curated by Zepf et al. [2023] using the first video sequence, which contains 115 2D images of multiple cells annotated with two classes: Cells and Background. It contains 651 images of single cells resized to 128x128 pixels. Annotators were instructed to perform either detailed annotations or coarse/wider annotations (on average 4% bigger by volume).

To simulate a domain shift, we divide the dataset into two disjoint sets: C (clean, unmodified) and C_t (tinted). Images in C_t are subject to a global color space transformation by adding a uniform RGB tint. Implemented bias is at r = 100%, i.e., all images belonging to C_t are paired with bloated Style 1 annotations ($\tilde{y}$). This forces the network to associate the added global tint with the bloated or over-segmented boundary. Refer to Figure 5 for qualitative examples of annotation style differences in the PhC-U373 dataset and CL error detection on a lightly-tinted sample.

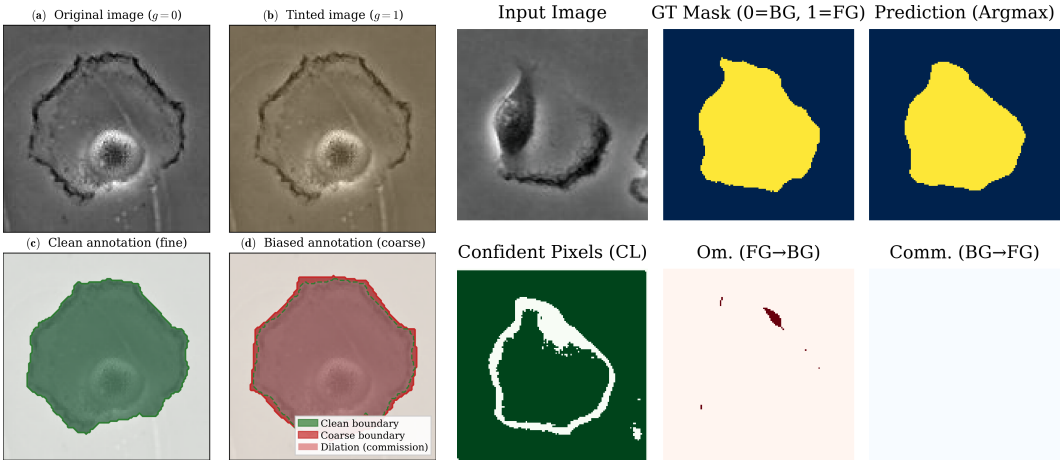


Figure 5: (Left) Example from the PhC-U373 dataset demonstrating label bias: (a) Original images from g_c (b) The same cell with a synthetic tint applied (group g_b) (c) Clean (fine) annotation obtained by an expert instructed for a detailed annotation (d) Biased (coarse; bloated) annotation assigned to tinted samples. The bloated boundary (in red) systematically over-segments to the clean boundary (in green), introducing over-segmentation (excess) bias correlated with the group membership. (Right) Confident Learning (CL) error detection on a g_b sample. Top row: input images, ground-truth annotation, and argmax prediction. Bottom row: pixels flagged as confident by CL, decomposed into label errors. CL correctly identifies the boundary region where bloated labels disagree with model prediction.

B.2.2 ISIC 2017 Dataset

The ISIC 2017 Challenge dataset [Codella et al., 2018] contains 2000 training, 150 validation, and 600 test dermoscopic images of skin lesions, each with a binary segmentation mask. Unlike CelebAMask-HQ, this dataset lacks demographic annotations. We therefore estimate skin tone as a proxy for demographic group membership using the Individual Typology Angle (ITA). For skin tone estimation via established skin tone classification studies by [Benčević et al., 2024, Van Song et al., 2026].

B.2.3 IMA++ Dataset

IMA++ (ISIC MultiAnnot++) is a large-scale multi-annotator skin-lesion segmentation dataset curated from the ISIC Archive [Abhishek et al., 2026]. It contains 17,684 binary segmentation masks for 14,967 dermoscopic images. While 12,573 images have a single mask, 2,394 images have between two and five annotations, contributed by 16 annotators. The metadata (attached with

the dataset) identifies the annotator, annotation tool, and reviewer skill level, capturing naturally occurring variation in lesion-boundary segmentation.

As in the ISIC 2017 experiments, we estimate skin tone using the Individual Typology Angle and group images into lighter (types 1-2) and darker (types 3-4) skin-tone strata. The original annotation assignments are retained for $\beta = 0\%$. For $\beta \in \{50\%, 100\%\}$, we increasingly pair annotation style with skin tone by selecting detailed annotations for types 1-2 and looser/wider annotations for types 3-4 among images with multiple available annotation masks. Note: Because IMA++ also does not provide a unique clean segmentation mask, we use it as a real-world exploratory setting rather than evaluating against assumed ground truth.

Skin Tone Estimation via K-Means Dominant Color: The patient's skin tone is estimated from healthy (non-lesion) skin surrounding the lesion, following a proposed pipeline:

1. **Skin Extraction:** We excluded the lesion area using the provided segmentation masks and removed the dark imaging (like the vintage lighting at corners) and hair via morphological blackhat filtering and adaptive thresholding. CLAHE is applied in CIE L*a*b space to normalize variation in illumination across devices, imaging conditions, etc.
2. **Dominant color via k-means:** We cluster the remaining skin pixels in CIE L*a*b color space using k-means with automatic cluster count selection. Specifically, we run k-means for $k \in 2, \ldots, 9$ and select the optimal k at the elbow point of the inertia curve using the Kneedle algorithm. The dominant skin color is taken as the centroid of the largest cluster, representing the most representative skin tone while being robust to residual artifacts, pigmentation, blurs, and shadow regions.
3. **ITA-based classification:** The skin-tone categorization is based on the Individual Typology Angle (ITA) computed from the dominant skin color in CIE $L^*a^*b^*$ space and mapped to four skin tone groups using standard ITA thresholds:

$$\text{ITA} = \arctan\left(\frac{L^* - 50}{b^*}\right) \times \frac{180}{\pi} \tag{13}$$

Table 5: Skin-tone grouping based on ITA ranges and approximate Fitzpatrick categories.

Group	ITA Range	Fitzpatrick Approx.	Training Samples
ST1 (Very Light)	ITA > 55	I	60
ST2 (Light)	41 < ITA ≤ 55	II–III	1,229
ST3 (Intermediate)	28 < ITA ≤ 41	III–IV	684
ST4 (Tan)	10 < ITA ≤ 28	V	23

This dataset contains no samples with ITA ≤ 10 (Brown/Dark). The distribution is unbalanced: ST2 and ST3 account for 96% of training samples, while ST4 represents only 1.2%. Notably, while circularity (invariant shape descriptor) remains stable across skin-tone groups (≈0.62-0.68), normalized lesion area decreases monotonically from ST1 to ST4 (0.418 to 0.037), indicating that darker skin tones are associated with smaller annotated lesion regions. This existing disparity in mask characteristics makes the dataset a suitable testbed for studying label bias. Refer to Figure 6 for a qualitative example of the dataset and natural label bias.

C Mitigation Methods

C.1 Additional Details: Bias Mitigation Methods

Baseline ERM has a training objective $\mathcal{L}\text{total} = \mathcal{L}\text{seg} + \lambda \cdot \mathcal{L}\text{debias}$, where $\mathcal{L}\text{seg}$ is combined CE + Dice loss and λ is increased from $0 \rightarrow$ desired values following GRL scheduler. For clarity, we consider two subgroups g_0 and g_1 instead of g_b and g_c.

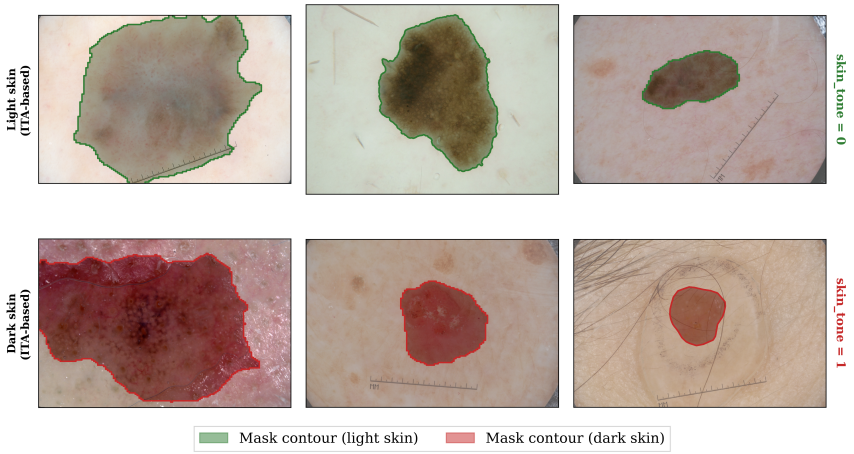


Figure 6: Selected samples from the ISIC 2017 dataset by the mentioned skin tone estimation method. Top row: light skin (ST1-ST3, ITA > 28), showing larger, well-segmented lesions against high-contrast backgrounds. Bottom row: dark skin (ST4, ITA ≤ 28), where lesions tend to be smaller with lower contrast against the surrounding skin. Segmentation annotations and contours are overlaid in green (light) and red (dark). The systematic decrease in annotated lesion area from lighter to darker skin tones suggests a group-conditioned disparity in the dataset even prior to synthetic bias injection, confirmed by CL metrics.

C.1.1 Adversarial Method (DANN-style)

A Gradient-Reversal Layer (GRL) [Ganin et al., 2016] is applied between the encoder and a demographic (here, gender) adversary head. The adversary is a 3-layer MLP ($C \to 256 \to 128 \to 1$) with ReLU activations and dropout ($p = 0.3$), operating on global-average-pooled encoder features $\mathbf{z} = \text{GAP}(f_L) \in \mathbb{R}^C$. The debiasing loss is the binary cross-entropy of the adversary's gender prediction:

$$\mathbf{L}_{\text{adv}} = -\frac{1}{B}\sum_{i=1}^{B}[g_i \log \sigma(h(\mathbf{z}_i)) + (1-g_i)\log(1-\sigma(h(\mathbf{z}_i)))] \tag{14}$$

where $h(\cdot)$ is the adversary head and σ is the sigmoid function. The GRL negates the gradient flowing to the encoder by a factor α during backpropagation, encouraging gender-invariant features. The encoder and adversary network use separate optimizers - AdamW with learning rates 10^{-4} and 5×10^{-4}, respectively.

C.1.2 Fairness Loss Regularization

This method we use for comparison, directly penalizes discrepancies in the models' prediction rates (representing model-centric mitigation approaches vs. our data-centric proposal), is typically used for classification tasks and does not modify the feature space. We use three fairness criteria as regularizers:

Demographic Parity (DP): The penalty here is the absolute difference in mean predicted foreground probability between groups:

$$\mathbf{L}_{\text{DP}} = |\mathbb{E}[\hat{p} \mid g = 0] - \mathbb{E}[\hat{p} \mid g = 1]| \tag{15}$$

where $\hat{p} = \text{softmax}(\hat{y})_1$ is the predicted foreground probability.

Equalized Odds (EO): The penalty is the group disparities in both the TPR and FPR:

$$\mathbf{L}_{\text{EO}} = |\text{TPR}_{g=0} - \text{TPR}_{g=1}| + |\text{FPR}_{g=0} - \text{FPR}_{g=1}| \tag{16}$$

where $\text{TPR}_g = \mathbb{E}[\hat{p} \mid y^=1, g]$ and $\text{FPR}_g = \mathbb{E}[\hat{p} \mid y^=0, g]$.

Combined (Both) .

$$\mathcal{L}_{\text{combined}} = \mathcal{L}_{\text{DP}} + \mathcal{L}_{\text{EO}}. \tag{17}$$

All expectations are computed per mini-batch using soft predictions (no thresholding).

C.1.3 Domain-Invariant Feature Alignment

We minimize the distribution distance between demographic groups in either feature or logit space, without an adversary network. This method was inspired by the idea of exploiting the effect of feature separability in Section 5.1.

MMD at feature level. forces to minimize the Maximum Mean Discrepancy (MMD) between encoder features of the subgroup using a multi-scale Gaussian kernel.

$$\mathrm{MMD}^2(\mathbf{Z}_0, \mathbf{Z}_1) = \sum_s \left[\frac{1}{n_0^2} \sum_{i,j} k_s(\mathbf{z}_i^0, \mathbf{z}_j^0) + \frac{1}{n_1^2} \sum_{i,j} k_s(\mathbf{z}_i^1, \mathbf{z}_j^1) - \frac{2}{n_0 n_1} \sum_{i,j} k_s(\mathbf{z}_i^0, \mathbf{z}_j^1) \right] \tag{18}$$

where $k_s(\mathbf{x}, \mathbf{y}) = \exp\left(-\|\mathbf{x} - \mathbf{y}\|^2 / 2\sigma_s^2\right)$ and bandwidths are $\sigma_s = \sigma_0 \cdot 2^s$ for $s \in \{-2, -1, 0, 1, 2\}$ around a base bandwidth σ_0. An L2-normalization flag `-mmd_normalize` (optional) normalizes $\mathbf{z}$ to a unit sphere before computing MMD, reducing sensitivity to feature magnitude scale. This is inspired by [Yan et al., 2021].

MMD-Logit on output-level. We noticed unstable training using feature-level MMD; this motivated us to use encoder features, which led us to align class-conditional logits at the model output. For each segmentation class c, pixel logits $\hat{y}_{i,c} \in \mathbb{R}^C$ are collected from pixels where $y^\star = c$, subsampled to $N = 256$ pixels per group, and the MMD-Logit is then computed between the two groups' logit vectors:

$$\mathbf{L}_{\text{mmd-logit}} = \frac{1}{C} \sum_{c=0}^{C-1} \mathrm{MMD}^2(\{\hat{y}_c^{g=0}\}, \{\hat{y}_c^{g=1}\}) \tag{19}$$

We observed it to be more stable in our experiments, mostly due to its low dimensionality ($C = 2$) logits compared to encoder features ($d = 512$). MMD-Logit generally outperforms feature-level MMD and CORAL on fairness metrics, as it targets the representation closest to the decision boundary rather than enforcing invariance in a high-dimensional space where invariance may be overly restrictive.

C.1.4 Pseudo Code for Auto-Conditioning (Unsupervised variant)

Pseudo-code for Auto-SCD in algorithmic block 1.

D Confident Learning Audit and Metric Diagnostics

D.1 Additional Details: CL Audit on PhC-U373

Table 6 reports the detailed audit across bias ratios $\beta \in \{0\%, 25\%, 50\%, 100\%\}$ on PhC-U373. Here, g_c denotes the group with tight annotations and g_b the group associated with wider annotations. Performance gaps are defined as $\Delta = g_c - g_b$. As β increases from the unbiased baseline (0.00%) to 100%, the observed Dice gap increases to 1.63 percentage points (3.00% for IoU), while the true gap against clean reference masks reaches 2.07 percentage points (3.80% for IoU).

The CL audit monitors these shifts directly on the training data without clean reference masks. At the unbiased baseline ($\beta = 0\%$), baseline error rates are $\mathrm{LER}^{g_c} = 7.6\%$ and $\mathrm{LER}^{g_b} = 10.2\%$, corresponding to a baseline ratio $\mathrm{RR} = 1.34$. At $\beta = 100\%$, LER^{g_c} increases to 8.1% and LER^{g_b} is 10.0%, resulting in a relative shift $\Delta \log(\mathrm{RR}) = -0.09$. The directional discrepancy DR^{g_b} remains close to zero throughout (ranging from $+0.04\%$ to $+0.01\%$). All LER, RR, and DR values are computed on the training data without access to clean reference masks.

Algorithm 1 Auto-Conditioned FiLM with Asymmetric Boundary Masking

Require: Training set $\mathcal{D} = \{(x_i, \tilde{y}_i, g_i)\}$, warmup epochs E_w, total epochs E_{total}, boundary half-width w

Require: Encoder E_θ, Decoder D_ϕ, per-stage FiLM parameters $\{\gamma^{(l)}, \beta^{(l)}\}_{l=1}^{L}$

Init $\gamma^{(l)} \leftarrow \mathbf{1}$, $\beta^{(l)} \leftarrow \mathbf{0}$ $\forall l$; $\quad \hat{g}_c \leftarrow$ `None`; $\quad L_{g=0}, L_{g=1} \leftarrow \varnothing$
for epoch $e = 1$ to E_{total} **do**
  for each mini-batch $(x, \tilde{y}, g) \subset \mathcal{D}$ **do**
    $\{f^{(l)}\} \leftarrow E_\theta(x)$; $\quad f^{(l)} \leftarrow \gamma_g^{(l)} \odot f^{(l)} + \beta_g^{(l)} \quad \forall l$ {FiLM on true g}
    $\hat{y} \leftarrow D_\phi(\{f^{(l)}\})$
    if $e \le E_w$ **or** $\hat{g}_c =$ `None` **then** {Warmup: standard loss}
      $\mathcal{L} \leftarrow \mathcal{L}_{\text{CE}}(\hat{y}, \tilde{y}) + \mathcal{L}_{\text{Dice}}(\hat{y}, \tilde{y})$
      Append per-sample $\mathcal{L}_i$ to L_{g_i} for each sample i
    else{Post-warmup: asymmetric boundary masking}
      $B \leftarrow \text{Dilate}(\tilde{y}, w) - \text{Erode}(\tilde{y}, w)$
      $W_i \leftarrow \begin{cases} 1 - B_i & \text{if } g_i = \hat{g}_b \\ 1 & \text{otherwise} \end{cases}$
      $\mathcal{L} \leftarrow W \odot \mathcal{L}_{\text{CE}}(\hat{y}, \tilde{y}) + W \odot \mathcal{L}_{\text{Dice}}(\hat{y}, \tilde{y})$
    end if
    Update $\theta, \phi, \{\gamma^{(l)}, \beta^{(l)}\}$ via $\nabla\mathcal{L}$
  end for
  if $e = E_w$ **then** {Auto-discover biased group}
    $\hat{g}_c \leftarrow \arg\min_g \overline{L}_g$; $\quad \hat{g}_b \leftarrow 1 - \hat{g}_c$
  end if
end for
Inference: $\hat{y} = D_\phi\big(\{\gamma_{\hat{g}_c}^{(l)} \odot E_\theta^{(l)}(x) + \beta_{\hat{g}_c}^{(l)}\}\big)$ {Always condition on $\hat{g}_c$}

Table 6: **CL audit results on PhC-U373.** Observed and true performance gaps are defined as $\Delta = g_c - g_b$. LER is the CL-estimated label-error rate for each group. $\Delta\log(\mathrm{RR}) = \log(\mathrm{LER}^{g_b}/\mathrm{LER}^{g_c}) - \log(\mathrm{RR}_{\text{baseline}})$ reflects the change in relative label-error disparity from the unbiased baseline. Positive DR indicates foreground excess relative to the observed annotation.

Condition	Observed Perf. Δ (%)		True Perf. Δ (%)		Label-Error Disparity			Direction
	Dice	IoU	Dice	IoU	$\mathbf{LER}^{\mathbf{g_c}}$ (%)	$\mathbf{LER}^{\mathbf{g_b}}$ (%)	$\boldsymbol{\Delta}\log(\mathbf{RR})$	$\mathbf{DR}^{\mathbf{g_b}}$ (%)
Unbiased	0.00	0.00	0.00	0.00	7.6	10.1	0.00	+0.04
25%	0.65	1.20	0.22	0.40	7.8	10.5	+0.01	+0.02
50%	0.98	1.80	0.71	1.30	7.9	10.6	+0.01	+0.02
100%	1.63	3.00	2.07	3.80	8.1	9.89	-0.09	+0.01

E Mitigation Results and Ablations

E.1 Additional Details: Mitigation Results

On CelebAMask-HQ, fairness regularization, adversarial training, and feature-distribution alignment give little or no improvement. At $\beta = 100\%$, the Dice gaps remain at $\approx 12\%$, compared to baseline ERM (which is also at a similar level). Asymmetric boundary masking alone is similarly insufficient, alone. Combining SCD with asymmetric masking provides no consistent improvement over SCD alone. On PhC-U373, the differences between methods are smaller: Auto-SCD achieves the lowest IoU gap (3.2%), whereas adversarial training gives the IoU gap (3.7%). The $EO + DP$ objective is unstable in this setting. Overall, SCD is the only approach that consistently reduces the true performance gap across both CelebAMask-HQ bias levels.

F Additional Audit Validations

F.1 Important Ablation: Confident Learning audit against clean labels

Table 7 shows when Confident Learning (CL) is evaluated on the Training Set against clean labels, **which is not available in practice**. **Purpose:** Used as an additional validation if CL correctly identifies errors on the training set when given clean labels. The results show that using clean labels, **which are not always available in practice**, we get a much clearer indication of the presence of bias and what direction, even at extreme bias conditions, i.e., at $\beta = 100\%$.

An important thing to note while interpreting CL audit against clean labels is direction of bias sensitivity when the objective is changed. We see a **expected inversion** of label error type when clean labels are used. When evaluating using trained labels (unclean; biased), the model generalizes visual features and predicts foreground pixels confidently, that the biased annotations systematically remove, resulting in a high deficit error. Conversely, when evaluating against clean ground-truth labels, the reference is complete. Here, the model, having learned the erosion bias during training, predicts an under-segmented mask than the true object boundary, resulting in a high excess error.

Table 7: **Retrospective validation of the CL audit against clean labels.** LER reports the detected error rate for each group, $\Delta_{\mathrm{LER}} = \mathrm{LER}^{g_b} - \mathrm{LER}^{g_c}$, and $\mathrm{RR} = \mathrm{LER}^{g_b}/\mathrm{LER}^{g_c}$. Positive DR indicates foreground excess in the reference mask relative to the cross-fitted prediction, whereas negative DR indicates foreground deficit. **Clean masks are used only for this retrospective validation and are not required by the standard audit.**

Bias level	Label-error disparity				Direction (%)		Statistical test
	$\mathbf{LER^{g_c}}$ (%)	$\mathbf{LER^{g_b}}$ (%)	$\boldsymbol{\Delta_{\mathrm{LER}}}$ (%)	RR	$\mathbf{DR^{g_c}}$	$\mathbf{DR^{g_b}}$	$\boldsymbol{\chi^2}$ ($p < 0.001$)
0%	1.78	2.33	0.55	1.31	−0.08	−0.08	114.75K
25%	1.97	2.72	0.75	1.38	−0.22	+0.52	333.41K
50%	2.22	6.96	4.74	3.14	−0.04	+5.90	5.53M
75%	2.14	15.44	13.30	7.21	+0.24	+14.74	17.13M
100%	2.24	18.44	16.20	8.23	+0.04	+17.99	22.45M

F.2 Ablation: Calibration test of Confident Learning Audit

To verify the robustness of our Confident Learning (CL) formulation against poorly calibrated models, we perform a global calibration ablation using Temperature Scaling. Specifically, we scale the baseline ERM logits by varying temperatures $T \in [0.1, 10.0]$ prior to the softmax activation, artificially inducing extreme over-confidence ($T < 1$) and under-confidence ($T > 1$). As shown in Table 8, CL naturally mitigates these global calibration shifts. While the predicted probabilities change drastically across temperatures, the expected self-confidence thresholds $\mathbb{E}[p(y = j \mid x) \mid \tilde{y} = j]$ adjust in exact proportion. Because CL identifies label errors relative to these class-conditional adaptive thresholds rather than fixed confidence cutoffs, the resulting error identification remains virtually invariant. Across the entire spectrum from over-sharpened to heavily smoothed output distributions, the DR and LER gaps remain constant at 10.0% and 7.89%, respectively, demonstrating that our bias quantification metrics do not depend on post-hoc probability calibration.

Table 8: **Sensitivity of Confident Learning metrics to global model miscalibration on CelebA-Mask (50% bias level).** Temperature scaling (T) controls the sharpness of predicted probabilities prior to softmax. CL thresholds adapt dynamically, leaving DR and LER gaps virtually invariant.

Temperature (T)	**CL Thresholds**		**DR Gap**	**LER Gap**	**Total Errors Detected**
	BG (t_0)	FG (t_1)	ΔDR	ΔLER	(Excess + Deficit)
0.1 (Over-confident)	0.98	0.92	10.0	7.89	18,863,402
0.5	0.97	0.91	10.0	7.89	18,863,403
0.8	0.96	0.90	10.0	7.89	18,863,403
1.0 (Baseline)	0.96	0.90	10.0	7.89	18,863,403
1.5	0.93	0.87	10.0	7.89	18,863,404
2.0	0.90	0.83	10.0	7.89	18,863,406
5.0	0.73	0.68	10.0	7.89	18,863,412
10.0 (Under-confident)	0.63	0.59	10.0	7.89	18,863,419

F.3 Additional Details: Uncertainty Estimates

Refer to Table 9 for an audit across bias ratios including uncertainty estimates.

Table 9: Audit results across bias ratios using biased training labels. Performance and error metrics are reported as $mean \pm std$, using bootstrapping. Dice, IoU, and Error Rates are in percentages.

Method / Condition	**Dice (%)**		**IoU (%)**		**Total LER (%)**	**DR Score**
	g_c	g_b	g_c	g_b	**Mean ± Std**	**Mean ± Std**
Baseline	98.61 ± 0.04	98.46 ± 0.03	97.25 ± 0.07	96.97 ± 0.06	2.13 ± 0.04	0.69 ± 0.01
25%	98.46 ± 0.05	95.25 ± 0.06	96.97 ± 0.11	90.93 ± 0.11	5.10 ± 0.08	0.37 ± 0.01
50%	98.26 ± 0.12	92.12 ± 0.08	96.58 ± 0.23	85.39 ± 0.14	7.56 ± 0.08	0.40 ± 0.02
75%	98.32 ± 0.17	93.03 ± 0.19	96.69 ± 0.33	86.96 ± 0.32	6.15 ± 0.21	0.51 ± 0.02
100%	98.25 ± 0.13	96.87 ± 0.21	96.56 ± 0.25	93.94 ± 0.39	3.06 ± 0.16	0.74 ± 0.01

G Mitigation Diagnostics and Stress Tests

G.1 Ablation: Auto-SCD Boundary Radius

Table 10: **Erosion-radius ablation for Auto-SCD on CelebAMask-HQ** (β = 100%, g_b: Female). True performance gaps are defined as $\Delta = g_c - g_b$. $\Delta_{\mathrm{LER}} = \mathrm{LER}^{g_b} - \mathrm{LER}^{g_c}$ and $\mathrm{RR} = \mathrm{LER}^{g_b}/\mathrm{LER}^{g_c}$. Positive DR indicates foreground excess, whereas negative DR indicates foreground deficit.

Configuration	True Perf. Gap (%)		Label-Error Disparity				Direction (%)	
	Dice Δ	IoU Δ	$\mathbf{LER}^{\mathbf{g_c}}$ (%)	$\mathbf{LER}^{\mathbf{g_b}}$ (%)	$\boldsymbol{\Delta}_{\mathbf{LER}}$ (%)	RR	$\mathbf{DR}^{\mathbf{g_c}}$	$\mathbf{DR}^{\mathbf{g_b}}$
Auto-SCD (default, e = 15)	7.13	16.1	1.77	2.53	0.76	1.43	−0.11	+0.14
Auto-SCD (e = 5)	**0.27**	**0.51**	**1.33**	**1.84**	**0.51**	**1.38**	−0.11	+0.10

G.2 Ablation: Observed-Label Evaluation and Backbone Test

The observed-label illusion. Table 11 highlights a central consequence of label bias: successful correction can appear unsuccessful when evaluated against the same biased annotations used for training. At 50% erosion, ERM obtains a true IoU of 93.1% with a 5.5-pp (percent-points) subgroup gap. Reference-style conditioning increases true IoU to 96.7% and reduces this gap to 0.5 pp. Nevertheless, its observed-label gap increases from 11.5 to 12.0 pp because predictions aligned with the clean reference boundary necessarily disagree with the eroded annotations assigned to g_b. Thus, observed-label evaluation penalizes the desired correction and creates the illusion that mitigation has not reduced disparity.

The directional CL counts make this disagreement easier to interpret. Under reference conditioning, g_b contains substantially more foreground-deficit than foreground-excess pixels (95.4M versus 10.4M), consistent with observed masks that are systematically eroded relative to the reference-conditioned predictions. Accordingly, RR increases and DR becomes more negative after successful mitigation. These quantities audit disagreement with the observed labels rather than accuracy against the unavailable clean target; they should therefore not be interpreted as post-mitigation objectives. We report them for diagnostic completeness, while using true-reference IoU and subgroup gaps to evaluate mitigation whenever clean masks are available.

Vision Transformer stress test. We additionally repeat the 50% erosion experiment using a ViT segmentation backbone (SegFormer MiT-B2 [Xie et al., 2021]). ViT-ERM reaches 92.4% true IoU with a 7.3-pp subgroup gap, compared with 93.1% and 5.5 pp for ResNet34-ERM. Its CL decomposition remains directionally consistent with erosion, identifying 80.0M foreground-deficit pixels compared with 43.8M foreground-excess pixels in g_b. Introducing SCD to the ViT successfully mitigates the bias, recovering performance to 97.1% true IoU and reducing the gap to just 0.6 pp. These experiments indicate that the group-conditional corruption and its CL audit are not specific to a convolutional backbone, and that our proposed style-conditioning mitigation effectively generalizes to transformer architectures.

G.3 Additional Details & Ablation: Conditioning and Predictor-Quality Analyses

Table 12 provides three additional analyses of the subgroup-conditioned model: a no-bias control, a counterfactual intervention on the inference-time style condition, and a stress test using a lower-capacity predictor.

No-bias control. At 0% label bias, forcing either g_c or g_b conditioning produces nearly identical performance: true IoU is 0.971 and 0.970, respectively, with the same subgroup gap of 0.004. DR also remains close to zero for both groups. Thus, the conditioning variable has negligible effect when the training annotations do not contain a group-dependent style difference. The residual $RR \approx 1.15$ reflects baseline differences in CL-identified disagreement rather than an induced/simulated label bias.

Counterfactual style intervention. At 50% erosion bias, changing only the inference-time condition substantially alters the predicted boundary style. Forcing the reference condition g_c yields a true

Table 11: **Detailed mitigation and backbone ablations on CelebAMask-HQ at $\beta = 50\%$.** Performance pairs denote IoU/Dice, while LER, DR, excess, and deficit pairs denote g_b/g_c. Performance and LER are in percent, gaps and DR are in percentage points, and counts are in millions of pixels.

Method	True IoU/Dice	Δ_T IoU/Dice	Observed IoU/Dice	Δ_O IoU/Dice	LER g_b/g_c	RR	DR g_b/g_c	Excess g_b/g_c	Deficit g_b/g_c
ResNet34: reference and generic baselines									
ERM	93.1/96.4	5.5/3.1	89.3/94.3	11.5/6.3	14.9/3.2	4.68	−6.7/0.0	39.3/6.6	86.4/6.5
Fairness EO	93.0/96.4	5.7/2.9	89.2/94.3	11.5/6.3	14.9/3.2	4.64	−6.6/ − 0.3	33.1/6.0	79.7/7.2
Fairness DP	93.0/96.4	5.6/3.0	89.2/94.3	11.4/6.2	15.0/3.3	4.58	−6.4/0.0	34.0/6.8	78.5/6.7
Fairness EO+DP	92.8/96.3	5.8/3.1	89.2/94.3	11.3/6.2	15.0/3.3	4.51	−6.1/0.1	35.1/7.2	77.1/6.4
Adversarial	92.4/96.1	6.7/3.6	89.2/94.3	11.7/6.4	15.0/3.1	4.90	−5.5/ − 0.2	37.0/5.8	76.9/6.7
MMD	92.8/96.3	6.3/3.3	89.2/94.3	11.9/6.5	15.0/3.0	5.00	−6.3/0.1	34.2/6.4	81.2/5.9
CORAL	93.2/96.5	5.5/2.9	89.3/94.3	11.5/6.3	15.0/3.2	4.67	−6.8/ − 0.3	32.3/6.1	80.6/7.1
GCE	92.3/96.0	6.6/3.5	89.1/94.2	11.6/6.4	15.0/3.3	4.60	−5.7/0.1	36.9/7.1	77.6/6.3
Bootstrapping	92.8/96.3	5.8/3.1	89.2/94.3	11.4/6.2	15.0/3.3	4.54	−6.1/0.0	42.0/6.9	83.9/6.6
Asymmetric	91.5/95.5	7.9/3.3	89.2/94.3	11.5/6.3	15.0/3.3	4.59	−4.2/0.0	39.6/5.4	61.0/5.3
ResNet34: subgroup-conditioned methods									
SCD	96.8/98.3	0.5/0.3	89.0/94.2	12.0/6.6	15.2/2.9	5.24	−13.3/ − 0.1	10.4/4.6	95.4/4.9
SCD + Asym	**96.8/98.4**	**0.5/0.3**	89.1/94.2	12.0/6.6	15.0/2.8	5.37	−13.3/0.0	2.5/1.1	23.7/1.2
Auto-SCD	Refer Table 2	–	–	–	–	–	–	–	–
Vision Transformer backbone									
ERM (ViT)	92.4/96.1	7.3/3.9	89.5/94.4	11.9/6.5	14.8/2.7	5.53	−4.9/0.0	43.8/5.5	80.0/5.4
Asymmetric (ViT)	92.0/95.8	8.1/4.3	89.4/94.4	12.2/6.6	14.9/2.6	5.70	−4.2/ − 0.2	46.6/5.0	78.3/5.6
SCD (ViT)	97.1/98.5	0.6/0.3	89.3/94.4	12.2/6.7	14.9/2.5	6.04	−13.3/ − 0.3	6.4/4.5	111.5/5.5
SCD-Asym (ViT)	**97.1/98.5**	**0.6/0.3**	89.3/94.4	12.2/6.7	14.9/2.5	6.00	−13.3/ − 0.3	**6.4/4.6**	**111.7/5.6**

Table 12: **Conditioning and predictor-quality ablations on CelebAMask-HQ.** True metrics use clean masks, whereas observed metrics use the available training annotations. IoU and LER are reported in percent; subgroup gaps and DR are in percentage points (pp). Paired CL entries denote g_b/g_c, and RR is dimensionless. At 0% bias, observed and true metrics coincide and are therefore not repeated.

Training setting	Inference/backbone	True masks		Observed masks		CL audit		
		IoU (%)	ΔIoU (pp)	IoU (%)	ΔIoU (pp)	LER (%)	RR	DR (pp)
No bias	StyleCond, forced g_c	97.1	0.4	–	–	3.1/2.6	1.16	−0.1/ − 0.1
	StyleCond, forced g_b	97.0	0.4	–	–	3.1/2.7	1.15	−0.3/ − 0.2
β =50%	StyleCond, forced g_c	**96.8**	**0.6**	89.1	12.0	15.1/2.8	5.34	−13.2/ − 0.2
	StyleCond, forced g_b	88.5	5.0	87.2	7.1	14.9/8.3	1.80	−1.7/ + 6.7
	ERM, ResNet18	91.5	7.9	89.0	11.9	15.2/3.2	4.70	−4.4/ + 0.1

IoU of 0.968 and reduces the true IoU gap to 0.006. In contrast, forcing g_b lowers true IoU to 0.885 and increases the gap to 0.050, indicating that the model has learned a controllable representation of the biased annotation style.

Evaluation against the observed annotations gives a different conclusion: reference conditioning has an observed IoU gap of 0.120, despite its true gap being only 0.006. This discrepancy illustrates how evaluation against biased labels can make a corrected model appear less equitable. The CL statistics capture the same disagreement: under reference conditioning, $\mathrm{LER}^{g_b} = 0.151$, $RR = 5.34$, and $\mathrm{DR}^{g_b} = -0.132$, correctly indicating that the eroded observed masks contain a foreground deficit relative to the reference-conditioned predictions.

Lower-quality predictor stress test. We additionally train an underfit ResNet18 segmentation model using ERM at 50% erosion bias. Its lower true IoU (0.915) and larger subgroup gap (0.079) provide a more challenging setting than the primary conditioned model. Nevertheless, CL continues to identify concentrated disagreement in g_b, with $RR = 4.70$ and negative $\mathrm{DR}^{g_b} = -0.044$. This suggests that the audit signal does not depend on near-perfect segmentation performance. Because this comparison changes both the backbone and training method, it should be interpreted as a predictor-quality stress test rather than an isolated estimate of the effect of model capacity.

Table 13: **Trained under a budget (see common protocol for details).** Zepf et al. [2023] do separate the styles well comparable to Auto-SCD method (but with assumption and requiring g_c). The full-data run presumably reaches the same end point. Zhang et al. [2020a] modeling provide no gain. The observed-label gap Δ_{obs} stays at ≈12 pp for every successful method, illustrating again that evaluation against biased masks is not a right measure for correct mitigation

Method	**Supervision**	**IoU**$_{\text{true}}$ ↑	g_c / g_b ↑	Δ_{true} ↓	Δ_{obs}	**LER** (g_c/g_b) ↓	**RR** ↓	**DR**$_{g_b}$
ERM	–	89.9	97.0 / 86.4	+10.7	+11.9	3.3 / 14.7	4.49	−2.0
Zhang et al. CM head (annotator = group)	group id	90.3	97.1 / 87.0	+10.1	+12.4	3.3 / 14.7	4.50	−3.3
SCD (FiLM, ref = g_c)	group id + ref.	92.7	97.0 / 90.6	+6.4	+12.1	3.2 / 14.8	4.59	−7.3
Auto-SCD	group id	93.1	97.1 / 91.1	+6.0	+12.0	3.2 / 15.3	4.70	−7.8
Zepf-style (one-hot g maps, ref = g_c)	group id + ref.	92.4	96.1 / 91.1	+4.9	+12.0	3.2 / 15.3	4.76	−7.0
Reference / Upper bounds								
Clean-subset fine-tune (200 clean images)	200 clean	**96.5**	96.8 / **96.3**	+0.5	+11.6	3.4 / **15.1**	**4.39**	−12.4
g_c-only ERM (2,949 male images only)	group id	96.2	**97.1** / 95.8	+1.3	+12.5	3.3 / 15.6	4.75	−13.3

G.4 Additional Results: Mitigation with label-noise baselines, under dilation and HBD

Common Protocol: All runs are in budget with stratified subset of 8 000 training images (5 051 female, 2 949 male), 1 000 validation images for the CL audit, and 1 500 test images with clean and observed masks; a ResNet-34 U-Net trained for 12 epochs (CE+Dice, AdamW 10^{-4}, cosine, batch 32, last-epoch checkpoint, no clean-label model selection); erosion/dilation radius r_d = 15, HBD amplitude 15, β = 50% unless stated. One seed per cell, so differences below ≈1 pp IoU should not be over-interpreted.

Additional mitigation baseline: To have a small comparison with some additional baselines like Zepf et al. [2023], which conditions a probabilistic segmentation network on **known annotation style for each training label**, so that we get a requested style at inference time. An important prerequisite for this approach is style variable is observed per label, and it should be independent of image content. In our problem setting, the only proxy we have is demographic subgroup g, which is a *genuine* image covariate. Our SCD also makes the same conditioning assumption, but at the decoder feature (FiLM) rather than at the input. Whereas the proposed Auto-SCD does not actually need the same assumption.

Another popular baseline is by Zhang et al. [2020a], where a base network predicts the latent true mask, and one confusion-matrix head per annotator maps it to each annotator's observed mask; identifiability comes from multiple annotations per image plus a regularizer. In our setting, and also in most real-life setting, each image carries a single annotation.

We implement a small ablation study using the mentioned baseline above (see Table 13), even though its assumptions do not fit the label-bias setting. We also test it on a g_c-only setting where we train an ERM on the clean group only. This acts as a trivial upper bound and tests if SCD gains come from the conditioning or simply from *ignoring the corrupted masks*. And an ERM baseline fine-tuned for 5 epochs on 200 clean-label samples; it requires clean labels, which our setting assumes we do not have during mitigation. This is a *oracle* analysis to demonstrate 'how many verified annotations would be needed to match performance gains by SCD?'.

SCD under dilation, HBD and annotator-level bias Table 14 repeats the erosion experiment (in the main text) for the other morphological corruptions of Section 3 and for an annotator-level corruption linked to g: four synthetic annotators (identity, erosion $r_d \in \{3, 12, 18\}$) are drawn with

$P(a \mid g_b) = (.15, .15, .35, .35)$ and $P(a \mid g_c) = (.35, .35, .15, .15)$, so 85% of g_b and 65% of g_c masks carry a non-identity style. The audit recovers the direction in every deterministic case ($\mathrm{DR}_{g_b} > 0$ for dilation, < 0 for erosion, slightly negative for HBD as predicted in Appendix A.1) and SCD reduces the gap for erosion and dilation; HBD at this amplitude barely moves IoU. When the style is mixed within both groups, RR falls to 1.0, and SCD recovers only about 2 of 9 pp, because the "reference" group is itself largely corrupted. g must therefore be a good proxy for annotation style; when annotator identities are available, conditioning on them is the natural extension.

Table 14: ERM, SCD and Auto-SCD under four bias geometries (g_b = female, $\beta = 50\%$ except the annotator setting). True IoU (%) of g_c / g_b, true and observed gaps (pp), and the CL audit of the observed validation masks (LER in %, DR of g_b in %).

Bias	Method	$\mathrm{IoU}_{\mathrm{true}}$	g_c / g_b	Δ_{true}	Δ_{obs}	RR	DR_{g_b}
Erosion	ERM	89.9	97.0 / 86.4	10.6	11.9	4.49	−2.0
	SCD	92.7	97.0 / 90.6	6.4	12.1	4.59	−7.3
	Auto-SCD	93.1	97.1 / 91.1	6.0	12.0	4.70	−7.8
Dilation	ERM	94.6	97.1 / 93.4	3.7	3.7	2.04	+0.9
	SCD	95.5	97.0 / 94.8	2.2	3.9	2.17	+2.8
	Auto-SCD	95.5	97.1 / 94.7	2.4	3.9	2.15	+2.7
HBD	ERM	96.5	97.1 / 96.2	1.0	2.4	1.67	−0.4
	SCD	96.6	97.0 / 96.3	0.7	2.2	1.65	−0.7
	Auto-SCD	96.4	97.0 / 96.2	0.8	2.3	1.68	−0.7
Annotator-level[†]	ERM	83.1	89.0 / 80.1	8.9	0.7	0.98	+0.4
	SCD	84.8	89.5 / 82.5	7.0	1.3	1.02	−2.6
	Auto-SCD	85.4	89.9 / 83.2	6.7	1.6	1.04	−3.2

[†]Four annotators (identity, erosion 3/12/18); 85% of g_b and 65% of g_c masks are non-identity.

G.5 Ablation: Robustness of automatic reference selection

Table 15 stresses the loss heuristic of Eq. 11 in the situations where it could plausibly fail: the biased group is the majority (the paper's default: females are 63% of the training set), the biased group is the minority, the corruption is weak ($\beta = 25\%$) or complete ($\beta = 100\%$), the corrupted masks are smoother (dilation, HBD), the style is mixed within groups, and no clean group exists at all. In every configuration with a cleaner group the lower-loss group is the cleaner one; the margin shrinks for $\beta = 100\%$ and the annotator setting and vanishes for group-independent corruption, where the choice is arbitrary and Auto-SCD, as expected, brings no benefit. The warm-up loss also separates corrupted from clean *images* (AUROC 0.78–0.99 for erosion and HBD), and the lowest-loss 30% of images are moderately clean whenever a clean group exists, which supports the assumption behind Eq. 11 and suggests per-sample reference selection as future work. The last column conditions the trained model on the other group and shows how much of the style the FiLM parameters have separated.

Table 15: **Auto-SCD reference selection across bias configurations.** Warm-up loss is the mean per-sample loss of each group over the 4 warm-up epochs; "AUROC" is the ability of the per-sample warm-up loss to identify corrupted images (all images / within g_b); "clean@30%" is the fraction of clean images among the 30% lowest-loss images (clean base rate in parentheses). Δ_{true} is reported for the selected reference and, as a counterfactual, when inference is conditioned on the other group.

Setting	**Biased group**	**Loss g_0 / g_1**	**Selected**	**AUROC**	**clean@30%**	**Δ_{true}**	**Δ_{true} (other)**
Erosion 50%, female (majority)	g_0	0.161 / 0.072	g_1 ✓	0.84 / 0.67	1.00 (0.68)	6.0	9.5
Erosion 50%, male (minority)	g_1	0.079 / 0.161	g_0 ✓	0.91 / 0.69	1.00 (0.82)	4.5	9.8
Erosion 25%	g_0	0.139 / 0.071	g_1 ✓	0.99 / 0.99	1.00 (0.84)	1.1	1.9
Erosion 100%	g_0	0.092 / 0.072	g_1 ✓	0.78 / –	0.67 (0.37)	4.8	4.5
Dilation 50%	g_0	0.106 / 0.074	g_1 ✓	0.66 / 0.45	0.88 (0.68)	2.4	3.6
HBD 50%	g_0	0.099 / 0.070	g_1 ✓	0.91 / 0.88	0.99 (0.68)	0.8	0.9
Annotator-level (mixed)	g_0	0.157 / 0.141	g_1 ✓	0.54 / 0.27	0.30 (0.23)	6.7	6.0
Group-independent 50%	none	0.160 / 0.159	g_1 (undefined)	0.67 / 0.67	0.68 (0.50)	−2.3	−2.4
No bias ($\beta = 0$)	none	0.077 / 0.069	g_1 (harmless)	–	–	0.5	0.4

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction outline the problem (label bias in segmentation), contributions (adapting CL to audit fairness, mitigation approach, and empirical findings).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations for our approach are discussed in Section 6 regarding its sensitivity during fully biased conditions. Also, the necessary assumptions are mentioned in the main text.

   Guidelines:

   - The answer [N/A] means that the paper has no limitations, while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: Our work is primarily methodological and empirical. We use existing well-established methods, and do not propose new formal mathematical methods requiring proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Disclosed all implementation details to reproduce label bias, and also specific dataset and model training details. Code for the framework is published.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Though we provide exact details to reproduce the results, actual data and code is released as open-source post-acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The core experimental details are summarized in the Implementation Details paragraph, and additional details regarding the dataset and hyperparameters are mentioned in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All experiments utilize a 5-fold cross-validation to ensure validity. Additionally, we report significance for the proposed metrics to quantify spatial errors robustly.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The hardware specifications are disclosed in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work complies with the NeurIPS code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The introduction and Conclusion discuss the societal danger of deploying biased segmentation models in safety-critical and clinical settings.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not release high-risk pre-trained generative models or anything high-risk framework that can be misused.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the used dataset and utilize it in accordance with its respective non-commercial academic research licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not introduce a new dataset. The methodology to reproduce our experimental setting is detailed.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our work relies on existing, publicly available datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No new human subjects research was conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: No LLMs are used as a component of the core methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.