Segmentation Style Discovery: Application to Skin Lesion Images

Kumar Abhishek[1 [0000−0002−7341−9617]], Jeremy Kawahara[2 [0000−0002−6406−5300]], and Ghassan Hamarneh[1 [0000−0001−5040−7448]]

[1] School of Computing Science, Simon Fraser University, Canada
[2] AIP Labs, Budapest, Hungary
{kabhishe,hamarneh}@sfu.ca, jeremy@aip.ai

Abstract. Variability in medical image segmentation, arising from annotator preferences, expertise, and their choice of tools, has been well documented. While the majority of multi-annotator segmentation approaches focus on modeling annotator-specific preferences, they require annotator-segmentation correspondence. In this work, we introduce the problem of segmentation style discovery, and propose StyleSeg, a segmentation method that learns plausible, diverse, and semantically consistent segmentation styles from a corpus of image-mask pairs without any knowledge of annotator correspondence. StyleSeg consistently outperforms competing methods on four publicly available skin lesion segmentation (SLS) datasets. We also curate ISIC-MultiAnnot, the largest multi-annotator SLS dataset with annotator correspondence, and our results show a strong alignment, using our newly proposed measure AS^2, between the predicted styles and annotator preferences. The code and the dataset are available at https://github.com/sfu-mial/StyleSeg.

Keywords: inter-rater variability · image segmentation· dermatology.

1 Introduction

Medical image segmentation is a critical component in medical image analysis pipelines, either as a preprocessing step for subsequent analyses or for treatment planning and image-guided human or robotic intervention. Following the seminal works of Long et al. [15] and Ronneberger et al. [21], there has been tremendous progress in deep learning (DL)-based medical image segmentation [4]. The majority of these works focus on learning to predict a single segmentation for an image. However, variability among experts when segmenting images has been well-documented, and these resulting segmentation masks are the product of latent factors such as ambiguous object boundaries and differences in tools, annotators' skill levels, criteria, and approaches to segmentation, and they capture different annotator segmentation preferences or "styles". Without accommodating these variations, a segmentation model optimized to minimize training error over a variety of human annotations may produce an "average" segmentation. This has motivated research that can be broadly categorized into two classes:

methods that model and learn to predict a single "gold standard" segmentation through label aggregation [24,13,18] (`SSeg`) and methods that predict multiple segmentations to capture the variability of annotations [22,26,12] (`MSeg`).

`MSeg` methods rely on modeling annotator-specific preferences, and training them typically requires annotations with annotator-segmentation correspondence. Therefore, given a set of images and a set of annotators, annotator-segmentation correspondences can be represented as a bipartite graph when every image has been segmented by at least 1 annotator, e.g., LIDC-IDRI [3], or a complete bipartite graph when every image has been segmented by every annotator, e.g., RIGA [2]. However, in the absence of such a correspondence, i.e., a scenario where we have a corpus of images and corresponding annotations without any knowledge of annotator IDs, defining a segmentation style is non-trivial since the latent factors associated with each segmentation are unknown, thus making it challenging to explicitly train a segmentation model to reproduce a particular style. Since we are unable to confirm even the number of unique annotators, we hypothesize that a possible solution for modeling multi-annotator segmentations would be **discovering unique annotation styles** from the dataset alone. Such a discovery-based approach needs to ensure (1) diversity in the discovered styles, (2) segmentation plausibility across all the styles, and (3) semantic consistency of the segmentations across all the images. However, to the best of our knowledge, there is minimal prior work on the discovery and modeling of annotation styles in the absence of annotator correspondence.

We argue that since even experts can (considerably) differ in how they segment, it is only natural that automated models trained thereupon also exhibit this variety. We envision that a segmentation system should produce results that align with the expectations of its (clinical) users, and that these users can vary in their personalization preferences (e.g., a study [10] found that expert dermatologists prefer "tighter" segmentations than their inexperienced counterparts). Moreover, such a system should, with minimal supervision, continue to produce the style that a user expects, thus avoiding constant user involvement with either manual corrections or image-by-image selection of preferred segmentation style.

In this work, we tackle the problem of style discovery and personalization modeling in medical image segmentation without requiring annotator correspondence, and focus our analysis on skin lesion segmentation (SLS). Advancements in DL over the past decade as well as the availability of large publicly available annotated datasets have enabled large strides in SLS [17,23]. Therefore, in this work, we work on style discovery in the context of multiple annotators for SLS, which has not been explored extensively. The majority of previous works focus on `SSeg` methods: either to select training samples that have high inter-annotator agreement [20] or training ensemble models to handle annotators' variability [18]. More recently, Zepf et al. [25] presented a small-scale ($n = 300$) analysis of annotation styles in images from the ISIC 2019 dataset based on the granularity of the annotation boundaries. In this work, we make the following contributions: **(1)** we introduce the problem of segmentation style discovery in the absence of any annotator correspondence and propose a method (StyleSeg) that predicts

multiple plausible, diverse, and semantically consistent segmentation styles, **(2)** we curate, to the best of our knowledge, the largest multi-annotator SLS dataset (ISIC-MultiAnnot) with annotator-segmentation mapping, and **(3)** we introduce a new measure (AS^2) for measuring the strength of alignment of the predicted styles with annotator preferences.

2 Method

Let $\mathcal{X} = \{X_i\}_{i=1}^N$ be a set of images and corresponding segmentation masks $\mathcal{Y} = \{\{Y_{ik}\}_{k=1}^{K_i}\}_{i=1}^N$, where $K_i > 0$ denotes the number of different ways X_i was segmented, without any knowledge of annotator correspondence. The goal is to discover unique annotation "styles" in this data $(\mathcal{X}, \mathcal{Y})$ such that, when given an image X_i, we predict $\{Y_{ij}\}_{j=1}^M$: M unique segmentations of X_i, that are diverse, plausible for X_i, and are of semantically consistent styles across all images.

To this end, we propose StyleSeg (Fig. 1 (a)): a segmentation approach that learns to predict M plausible segmentations that capture a variety of styles from a corpus of images and corresponding masks without any annotator correspondence. StyleSeg consists of two deep learning models that are trained together: (i) a **segmentation model** f_s, parameterized by Θ_s, that predicts M segmentation masks from image X_i, where $M \in \mathbb{N}$ is a user-specified value,

$$\{\hat{Y}_{ij}\}_{j=1}^M = f_s(X_i; \Theta_s), \tag{1}$$

and (ii) a **style classifier model** f_c, parameterized by Θ_c, that predicts a vector $p_i \in \mathbb{R}^M$ of M probabilities,

$$p_i = f_c(X_i, Y_{ik}; \Theta_c), \tag{2}$$

where p_{ij} is the probability that (X_i, Y_{ik}) is of style j. Note that knowing X_i is necessary to define the styles, since the observed segmentations are a product of image content and annotation style.

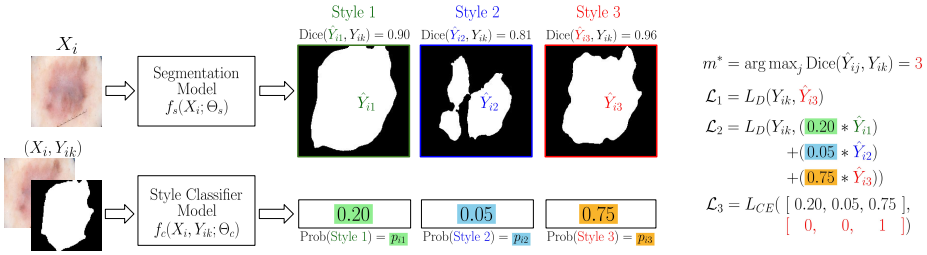
Of the M predicted segmentations from f_s, we first need to identify the style that is the closest to the ground truth Y_{ik}, and then optimize it to make it even closer. Mathematically, we minimize the loss $\mathcal{L}_1$,

$$\mathcal{L}_1 = L_D(Y_{ik}, \hat{Y}_{im^*}), \tag{3}$$

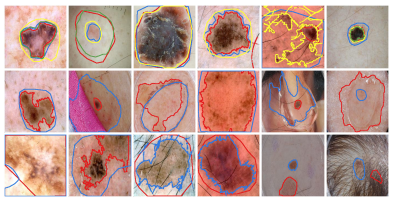
$$m^* = \arg\max_j \text{Dice}(Y_{ik}, \hat{Y}_{ij}), \tag{4}$$

where $L_D = 1 - \text{Dice}$ and Dice denotes the Dice similarity coefficient [7]. We also require the other predicted styles to still be plausible, i.e., similar to ground truth Y_{ik}. However, requiring all styles to be equally plausible compromises the styles' diversity. Therefore, we make the strength of a style's plausibility requirement proportional to its likelihood of being the predicted style, according to the style classifier f_c (Eqn. 2). To this end, we encourage the weighted sum of predicted segmentations $\hat{Y}_{ij}$ to be similar to ground truth Y_{ik}, where the scalar weights are p_{ij}. Mathematically, we minimize the loss $\mathcal{L}_2$,

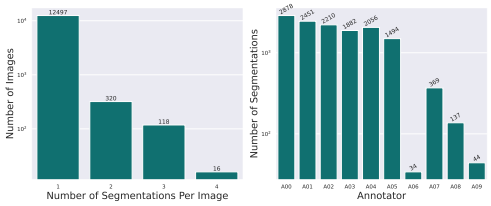
$$\mathcal{L}_2 = L_D\left(Y_{ik}, \sum\nolimits_j^M p_{ij}\hat{Y}_{ij}\right). \tag{5}$$



(a) StyleSeg overview with $M = 3$: the segmentation model f_s predicts 3 plausible segmentations of different styles, while the style classifier f_c predicts the style that is the most similar to the ground truth.



(b) Sample training images showing variability of segmentations.



(c) Distribution of ISIC-MultiAnnot by number of annotators and segmentations.

Fig. 1: (a) An overview of the proposed method StyleSeg. (b) Inter-annotator variability in the training images. (c) An annotator-wise breakdown of the newly curated ISIC-MultiAnnot dataset.

Weighting the M segmentations by p_i ensures that when p_i has a high entropy (e.g., in the initial training epochs), all styles are encouraged to be similar to Y_{ik}, whereas when p_i has a low entropy, only a subset of the M styles are encouraged to be similar to Y_{ik}, thus enabling a coarse-to-fine style refinement.

Additionally, we employ a cross-entropy loss $\mathcal{L}_3$ to train the style classifier f_c by learning to predict the style that is the most similar to the ground truth,

$$\mathcal{L}_3 = L_{CE}(p_i, m^*). \tag{6}$$

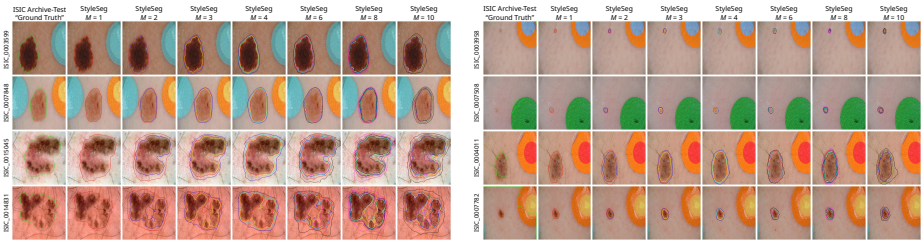
Finally, we optimize the parameters Θ_s of f_s and Θ_c of f_c using

$$\Theta_s^*, \Theta_c^* = \arg\min_{\Theta_s, \Theta_c} \sum_i^N \mathcal{L}_{\text{total}}, \tag{7}$$

where

$$\mathcal{L}_{\text{total}} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3. \tag{8}$$

Note that we do not include an explicit style distinctiveness constraint since, in the absence of annotator correspondence, the styles are entangled with the segmentations. Nevertheless, these loss terms used together (Eqn. 8) implicitly encourage the styles to be different as the training progresses, as seen in our results.



(a) Sample outputs of StyleSeg for lesions with and without distinct borders.

(b) StyleSeg produces better segmentations than even the test "ground truth".

Fig. 2: Evaluating StyleSeg on ISIC Archive-Test: diverse and plausible segmentations that are semantically consistent across styles.

3 Results and Discussion

Datasets: Similar to previous works [20,18], we train StyleSeg on images obtained from the ISIC Archive [1], specifically images with more than one "ground truth" segmentation. We select 2,261 images that meet this criterion (2,122 with two, 100 with three, 35 with four, and 4 with five segmentations), resulting in 4,704 image-mask pairs. Note that these images exhibit a vast range of inter-annotator agreement, as evidenced qualitatively (sample images with their masks in Fig. 1 (c)) and quantitatively (pairwise Dice coefficients and Fleiss' kappa in Supp. Mat. Fig. SM1 (a)). We choose Fleiss' kappa [9] over Cohen's kappa [6] used by Ribeiro et al. [20] because the former can be used with multi-rater settings while the latter cannot [19]. We reserve 1,525 image-mask pairs from the ISIC Archive for our validation set. See Supp. Mat. for model architectures and training details. We evaluate on four publicly available datasets: ISIC Archive-Test containing 10,000 dermoscopic images with just one segmentation ground truth per image from the ISIC Archive, DermoFit (1,300 clinical images) [5], SCD (206 dermoscopic images) [11], and PH^2 (200 dermoscopic images) [16].

Competing methods: We train StyleSeg with $M = \{2, 3, 4, 6, 8, 10\}$ and compare it to the following `SSeg` methods: `NaiveTraining`: a segmentation model without any annotator-specific knowledge; `RandAnnotID` [18]: 4 segmentation models, one optimized for each annotator randomly assigned to a mask, `LessIsMore` [20]: a segmentation model trained on a subset of the masks whose average pairwise Cohen's kappa is above 0.5; and `D-LEMA` [18]: an ensemble of Bayesian segmentation models. We also compare against an `MSeg` method `MHP` (multiple hypothesis prediction) [22] also with $M = \{2, 3, 4, 6, 8, 10\}$.

Qualitative results of StyleSeg on ISIC Archive-Test (Fig. 2 (a)) show plausibility (all segmentations cover the lesion with varying degrees of over- or under-segmentation) as well as semantic consistency across segmentations (e.g., when $M = 3$, yellow always has a tight and jagged boundary while blue always has a loose boundary). We also provide a quantitative assessment in Supp. Mat. Fig. SM1 (f). Also, observe that in lesions with well-defined borders (top two

Table 1: Dice mean$_{std.dev.}$ comparing StyleSeg to `SSeg` [20,18] (first 4 rows) and `MSeg` (MHP [22]) methods. For the latter, we report the mean, median, min., max. of Dice between the ground truth and all the predicted segmentation styles. Note how StyleSeg consistently outperforms all competing methods while also producing more plausible segmentations than MHP. ⊘ denotes that the result cannot be reported since D-LEMA's [18] code is not available.

Method	ISIC Archive-Test ($n = 10000$)				PH^2 ($n = 200$)				DermoFit ($n = 1300$)				SCD ($n = 206$)			
	Mean	Median	Minimum	Maximum	Mean	Median	Minimum	Maximum	Mean	Median	Minimum	Maximum	Mean	Median	Minimum	Maximum
NaiveTraining	-	-	-	$0.800_{0.188}$	-	-	-	$0.880_{0.071}$	-	-	-	$0.842_{0.159}$	-	-	-	$0.766_{0.198}$
RandAnnotID [18]	-	-	-	⊘	-	-	-	$0.897_{0.005}$	-	-	-	$0.826_{0.004}$	-	-	-	⊘
LessIsMore [20]	-	-	-	$0.815_{0.178}$	-	-	-	$0.895_{0.070}$	-	-	-	$0.854_{0.127}$	-	-	-	$0.804_{0.189}$
D-LEMA [18]	-	-	-	⊘	-	-	-	$0.920_{0.004}$	-	-	-	$0.853_{0.003}$	-	-	-	⊘
2-MHP	$0.796_{0.168}$	$0.727_{0.186}$	$0.727_{0.186}$	$0.864_{0.158}$	$0.850_{0.114}$	$0.786_{0.163}$	$0.786_{0.163}$	$0.914_{0.070}$	$0.795_{0.126}$	$0.707_{0.089}$	$0.707_{0.089}$	$0.882_{0.080}$	$0.796_{0.126}$	$0.713_{0.180}$	$0.713_{0.180}$	$0.879_{0.119}$
2-StyleSeg	$0.814_{0.168}$	$0.760_{0.186}$	$0.760_{0.186}$	$0.869_{0.161}$	$0.878_{0.073}$	$0.827_{0.103}$	$0.827_{0.103}$	$0.929_{0.050}$	$0.824_{0.128}$	$0.759_{0.180}$	$0.759_{0.180}$	$0.888_{0.090}$	$0.824_{0.124}$	$0.754_{0.157}$	$0.754_{0.157}$	$0.895_{0.104}$
3-MHP	$0.772_{0.181}$	$0.789_{0.184}$	$0.652_{0.258}$	$0.876_{0.154}$	$0.780_{0.185}$	$0.796_{0.217}$	$0.625_{0.303}$	$0.919_{0.070}$	$0.739_{0.176}$	$0.767_{0.193}$	$0.562_{0.289}$	$0.888_{0.083}$	$0.715_{0.194}$	$0.752_{0.217}$	$0.523_{0.297}$	$0.869_{0.136}$
3-StyleSeg	$0.804_{0.169}$	$0.819_{0.174}$	$0.713_{0.199}$	$0.881_{0.154}$	$0.885_{0.093}$	$0.900_{0.080}$	$0.811_{0.197}$	$0.943_{0.045}$	$0.817_{0.134}$	$0.835_{0.136}$	$0.720_{0.202}$	$0.897_{0.086}$	$0.818_{0.151}$	$0.837_{0.149}$	$0.716_{0.213}$	$0.901_{0.120}$
4-MHP	$0.773_{0.170}$	$0.752_{0.196}$	$0.623_{0.245}$	$0.886_{0.143}$	$0.830_{0.131}$	$0.817_{0.164}$	$0.674_{0.264}$	$0.933_{0.043}$	$0.796_{0.132}$	$0.783_{0.159}$	$0.636_{0.243}$	$0.904_{0.079}$	$0.751_{0.153}$	$0.726_{0.180}$	$0.547_{0.251}$	$0.896_{0.104}$
4-StyleSeg	$0.804_{0.163}$	$0.786_{0.177}$	$0.693_{0.197}$	$0.889_{0.147}$	$0.875_{0.081}$	$0.863_{0.103}$	$0.776_{0.143}$	$0.945_{0.042}$	$0.812_{0.135}$	$0.794_{0.165}$	$0.681_{0.221}$	$0.907_{0.075}$	$0.786_{0.133}$	$0.766_{0.179}$	$0.632_{0.208}$	$0.896_{0.111}$
6-MHP	$0.647_{0.159}$	$0.703_{0.205}$	$0.121_{0.176}$	$0.886_{0.151}$	$0.790_{0.076}$	$0.840_{0.089}$	$0.400_{0.246}$	$0.939_{0.050}$	$0.749_{0.116}$	$0.777_{0.131}$	$0.428_{0.169}$	$0.900_{0.063}$	$0.649_{0.126}$	$0.703_{0.181}$	$0.156_{0.246}$	$0.881_{0.150}$
6-StyleSeg	$0.795_{0.178}$	$0.799_{0.190}$	$0.648_{0.234}$	$0.889_{0.154}$	$0.869_{0.090}$	$0.873_{0.098}$	$0.745_{0.167}$	$0.948_{0.036}$	$0.814_{0.126}$	$0.818_{0.149}$	$0.651_{0.238}$	$0.911_{0.070}$	$0.798_{0.143}$	$0.806_{0.149}$	$0.608_{0.244}$	$0.906_{0.106}$
8-MHP	$0.625_{0.150}$	$0.658_{0.215}$	$0.099_{0.150}$	$0.896_{0.161}$	$0.752_{0.070}$	$0.801_{0.086}$	$0.260_{0.148}$	$0.944_{0.053}$	$0.698_{0.117}$	$0.708_{0.148}$	$0.309_{0.161}$	$0.908_{0.063}$	$0.616_{0.123}$	$0.627_{0.206}$	$0.134_{0.111}$	$0.897_{0.103}$
8-StyleSeg	$0.790_{0.170}$	$0.798_{0.185}$	$0.595_{0.232}$	$0.899_{0.158}$	$0.875_{0.096}$	$0.878_{0.113}$	$0.745_{0.158}$	$0.950_{0.037}$	$0.810_{0.141}$	$0.815_{0.162}$	$0.632_{0.239}$	$0.910_{0.075}$	$0.798_{0.161}$	$0.812_{0.179}$	$0.586_{0.241}$	$0.901_{0.114}$
10-MHP	$0.706_{0.174}$	$0.745_{0.204}$	$0.281_{0.251}$	$0.894_{0.160}$	$0.724_{0.140}$	$0.761_{0.233}$	$0.339_{0.245}$	$0.938_{0.059}$	$0.629_{0.177}$	$0.667_{0.218}$	$0.181_{0.215}$	$0.906_{0.068}$	$0.690_{0.168}$	$0.733_{0.218}$	$0.223_{0.196}$	$0.898_{0.094}$
10-StyleSeg	$0.793_{0.173}$	$0.805_{0.185}$	$0.603_{0.214}$	$0.899_{0.114}$	$0.866_{0.101}$	$0.880_{0.111}$	$0.692_{0.179}$	$0.951_{0.055}$	$0.801_{0.147}$	$0.813_{0.166}$	$0.579_{0.255}$	$0.918_{0.065}$	$0.768_{0.181}$	$0.791_{0.196}$	$0.513_{0.246}$	$0.885_{0.140}$

rows), the predicted styles are similar, whereas in lesions with ambiguous borders (bottom two rows), the predictions exhibit considerable diversity. It is also worth noting that several images in ISIC Archive-Test have either incorrect or imprecise "ground truth" masks (Fig. 2 (b)), which leads to incorrect penalization of StyleSeg's accurate predictions during evaluation.

Quantitative results: For `SSeg` methods, we report the Dice coefficient. For `MSeg` methods, we report $\texttt{max}_j(d)$, where $d = \text{Dice}(Y_{ik}, \hat{Y}_{ij})$, to assess the highest agreement, and $\{\texttt{mean}_j(d),\ \texttt{median}_j(d),\ \texttt{min}_j(d)\}$ to assess the plausibility of all M segmentations. For example, an `MSeg` model that produces even one poor segmentation will have low scores for $\texttt{min}_j(d)$, indicating low plausibility.

Table 1 shows that predicting more than one style (StyleSeg, MHP) improves performance ($\texttt{max}_j(d)$) compared to `SSeg` methods, and even `MSeg` methods that predict just two styles (2-StyleSeg, 2-MHP) consistently outperform `SSeg` methods. Moreover, as M increases, a larger number of diverse segmentations are produced, and the $\texttt{max}_j(d)$ keeps improving. However, we observe that for three out of the four datasets, the $\texttt{max}_j(d)$ performance either plateaus or starts to decline as M increases. We posit that after an optimal number of styles, generating more segmentations leads to diversity at the cost of performance, and leave this investigation for future work. Interestingly, datasets that do not have a documented presence of inter-segmentation variability (DermoFit, PH^2, SCD) also benefit from learning to predict multiple segmentations, indicating style variability in the ground truth masks. A post hoc investigation of DermoFit, for example, confirms the presence of different annotation styles (difference in boundary granularity; Supp. Mat. Fig. SM1 (b)).

Finally, StyleSeg consistently outperforms MHP for all datasets and M, except $M = 10$ with SCD, and as M increases, the plausibility of MHP models across all the predictions decreases, as evident through the declining $\{\texttt{mean}_j(d), \texttt{median}_j(d), \texttt{min}_j(d)\}$ scores. For example, when modeling 10 styles, the $[\texttt{min}_j(d),$

Table 2: StyleSeg's segmentation agreement ($\text{mean}_{std.dev.}$ of $\text{Dice}_{\text{IASS}}$ and $\text{Dice}_{\text{ASSS}}$) and style alignment ($\text{AS}^2$) on the 27 annotator preferences in ISIC-MultiAnnot. $\mathcal{J}$ denotes the single style that, for each row, maximizes agreement with the ground truth. As more styles are modeled, $\text{Dice}_{\text{IASS}}$, $\text{Dice}_{\text{ASSS}}$, and AS^2 all improve, and all annotator preferences consistently align with a discovered style.

Annotator + Tool + Experience	Seg. Count	1-StyleSeg	2-StyleSeg			3-StyleSeg			4-StyleSeg		
		$\text{Dice}_{\text{ISSS}}$	$\text{Dice}_{\text{ISSS}}$	$\text{Dice}_{\text{ASSS}}$	$\mathcal{J}$	$\text{Dice}_{\text{ISSS}}$	$\text{Dice}_{\text{ASSS}}$	$\mathcal{J}$	$\text{Dice}_{\text{ISSS}}$	$\text{Dice}_{\text{ASSS}}$	$\mathcal{J}$
A00+T2+E	1573	$0.892_{0.089}$	$0.923_{0.061}$	$0.913_{0.087}$	2	$0.944_{0.049}$	$0.913_{0.106}$	3	$0.944_{0.044}$	$0.914_{0.111}$	1
A00+T2+N	1305	$0.716_{0.302}$	$0.761_{0.293}$	$0.728_{0.308}$	2	$0.793_{0.287}$	$0.727_{0.313}$	3	$0.790_{0.290}$	$0.726_{0.304}$	3
A01+T1+N	6	$0.559_{0.362}$	$0.766_{0.152}$	$0.766_{0.152}$	1	$0.754_{0.132}$	$0.741_{0.125}$	2	$0.819_{0.106}$	$0.767_{0.113}$	2
A01+T3+E	297	$0.900_{0.104}$	$0.915_{0.093}$	$0.897_{0.107}$	2	$0.927_{0.075}$	$0.900_{0.097}$	1	$0.931_{0.067}$	$0.904_{0.090}$	3
A01+T3+N	2148	$0.829_{0.185}$	$0.857_{0.167}$	$0.817_{0.170}$	1	$0.869_{0.159}$	$0.836_{0.178}$	1	$0.876_{0.148}$	$0.836_{0.175}$	3
A02+T1+E	1742	$0.844_{0.177}$	$0.880_{0.140}$	$0.856_{0.159}$	1	$0.886_{0.132}$	$0.854_{0.159}$	1	$0.895_{0.112}$	$0.859_{0.148}$	4
A02+T3+E	468	$0.856_{0.172}$	$0.889_{0.167}$	$0.883_{0.175}$	2	$0.899_{0.161}$	$0.874_{0.188}$	3	$0.903_{0.146}$	$0.890_{0.160}$	1
A03+T1+E	1622	$0.778_{0.168}$	$0.845_{0.117}$	$0.827_{0.137}$	1	$0.854_{0.111}$	$0.824_{0.145}$	2	$0.881_{0.095}$	$0.823_{0.132}$	4
A03+T3+E	260	$0.891_{0.116}$	$0.912_{0.086}$	$0.876_{0.173}$	2	$0.923_{0.089}$	$0.868_{0.150}$	1	$0.932_{0.074}$	$0.874_{0.163}$	3
A04+T1+E	992	$0.850_{0.158}$	$0.880_{0.131}$	$0.860_{0.149}$	1	$0.888_{0.132}$	$0.866_{0.153}$	2	$0.906_{0.108}$	$0.856_{0.157}$	4
A04+T1+N	61	$0.760_{0.242}$	$0.840_{0.152}$	$0.823_{0.164}$	1	$0.837_{0.162}$	$0.786_{0.201}$	1	$0.827_{0.206}$	$0.789_{0.226}$	4
A04+T3+E	913	$0.912_{0.088}$	$0.939_{0.054}$	$0.934_{0.065}$	2	$0.948_{0.047}$	$0.926_{0.069}$	1	$0.951_{0.045}$	$0.932_{0.063}$	3
A04+T3+N	90	$0.877_{0.096}$	$0.910_{0.068}$	$0.905_{0.070}$	2	$0.928_{0.031}$	$0.908_{0.044}$	3	$0.926_{0.052}$	$0.913_{0.055}$	1
A05+T1+E	752	$0.815_{0.203}$	$0.862_{0.163}$	$0.837_{0.179}$	1	$0.873_{0.162}$	$0.827_{0.184}$	1	$0.882_{0.147}$	$0.841_{0.177}$	4
A05+T3+E	742	$0.875_{0.129}$	$0.903_{0.109}$	$0.891_{0.113}$	2	$0.916_{0.098}$	$0.878_{0.120}$	1	$0.919_{0.091}$	$0.891_{0.108}$	1
A06+T1+E	10	$0.824_{0.187}$	$0.902_{0.037}$	$0.885_{0.070}$	1	$0.909_{0.034}$	$0.889_{0.049}$	2	$0.909_{0.039}$	$0.880_{0.063}$	4
A06+T3+E	24	$0.862_{0.079}$	$0.916_{0.053}$	$0.916_{0.053}$	2	$0.934_{0.031}$	$0.923_{0.031}$	3	$0.933_{0.041}$	$0.929_{0.040}$	1
A07+T1+E	67	$0.820_{0.157}$	$0.877_{0.124}$	$0.867_{0.150}$	1	$0.890_{0.108}$	$0.862_{0.157}$	2	$0.897_{0.104}$	$0.862_{0.149}$	4
A07+T1+N	251	$0.837_{0.141}$	$0.892_{0.085}$	$0.879_{0.104}$	1	$0.903_{0.067}$	$0.875_{0.114}$	2	$0.905_{0.070}$	$0.873_{0.101}$	4
A07+T3+E	12	$0.925_{0.055}$	$0.938_{0.019}$	$0.937_{0.019}$	2	$0.939_{0.020}$	$0.916_{0.055}$	1	$0.947_{0.016}$	$0.932_{0.017}$	1
A07+T3+N	39	$0.863_{0.177}$	$0.918_{0.061}$	$0.913_{0.071}$	2	$0.933_{0.037}$	$0.899_{0.148}$	3	$0.934_{0.039}$	$0.914_{0.079}$	1
A08+T1+E	26	$0.666_{0.225}$	$0.750_{0.161}$	$0.680_{0.242}$	2	$0.747_{0.197}$	$0.653_{0.260}$	1	$0.793_{0.134}$	$0.666_{0.261}$	1
A08+T3+E	111	$0.605_{0.230}$	$0.668_{0.197}$	$0.626_{0.210}$	1	$0.677_{0.206}$	$0.628_{0.218}$	2	$0.735_{0.166}$	$0.669_{0.203}$	2
A09+T1+E	30	$0.815_{0.121}$	$0.841_{0.098}$	$0.784_{0.156}$	1	$0.873_{0.089}$	$0.833_{0.113}$	2	$0.884_{0.076}$	$0.812_{0.119}$	4
A09+T1+N	1	$0.953_{0.000}$	$0.927_{0.000}$	$0.927_{0.000}$	2	$0.955_{0.000}$	$0.955_{0.000}$	1	$0.947_{0.000}$	$0.947_{0.000}$	3
A09+T3+E	10	$0.900_{0.074}$	$0.918_{0.054}$	$0.918_{0.054}$	2	$0.933_{0.038}$	$0.909_{0.044}$	1	$0.937_{0.043}$	$0.919_{0.040}$	3
A09+T3+N	3	$0.894_{0.070}$	$0.911_{0.058}$	$0.911_{0.058}$	2	$0.957_{0.015}$	$0.957_{0.015}$	3	$0.944_{0.030}$	$0.944_{0.030}$	1
AS^2 (Eqn. 9)		-	$0.299_{0.208}$			$0.347_{0.237}$			$0.466_{0.296}$		

$\texttt{max}_j(d)$] range across 10 segmentations for 10,000 test images in ISIC Archive-Test is [0.281, 0.894] for 10-MHP and [0.603, 0.899] for 10-StyleSeg, meaning all the predicted segmentations are more plausible for the latter. We attribute this improvement to the plausibility constraint ($\mathcal{L}_2$ in Eqn. 5), which penalizes predicted segmentations that considerably deviate from the ground truth.

A new multi-annotator dataset: Next, we propose ISIC-MultiAnnot, a new multi-annotator SLS dataset curated from the ISIC Archive that, to the best of our knowledge, is the largest such dataset to contain annotator correspondence. The annotator-segmentation mapping in ISIC-MultiAnnot forms an incomplete bipartite graph, i.e., not every image has been segmented by every annotator. ISIC-MultiAnnot contains 12,951 images segmented by 10 annotators, resulting in 13,555 image-mask pairs (breakdown in Fig. 1 (c)). Unlike other multi-annotator datasets, the variability in ISIC-MultiAnnot's segmentations stems from three annotation pipeline factors: the annotator (10 annotator IDs: "A00"–"A09"), the tool used ("T1", "T2", "T3"), and the expertise of the manual reviewer ("expert" or "novice") [18], resulting in 27 unique annotator preferences, which we use for our evaluation. We measure StyleSeg performance in two settings: (i) **image-adaptive style selection (IASS)**: for every image,

we find the style that maximizes the agreement with ground truth, measured as $\text{Dice}_{\text{IASS}} = \texttt{max}_j(\text{Dice}(Y_{ik}, \hat{Y}_{ij}))$, and (ii) a more challenging **annotator-specific style selection (ASSS)**: we find a single style, fixed across all images, that maximizes agreement with ground truth, measured as $\text{Dice}_{\text{ASSS}} = \text{Dice}(Y_{ik}, \hat{Y}_{i\mathcal{J}})$, where $\mathcal{J} = \arg\max_j(\sum_i \text{Dice}(Y_{ik}, \hat{Y}_{ij}))$. Note that $\text{Dice}_{\text{ASSS}} \leq \text{Dice}_{\text{IASS}}$.

Quantitative results of StyleSeg on ISIC-MultiAnnot (Table 2; additional results in Supp. Mat. Fig. SM1 (c)) show that each of the discovered styles presents a high agreement with almost all the annotator preferences. A notable outlier is "A08", and upon manual inspection, we found a large number of ground truth segmentations to be incorrect to varying degrees (Supp. Mat. Fig. SM1 (e)), which explains the lower evaluation performance. Similar to Table 1, modeling even two styles yields better performance than one style. Moreover, as M increases, the newly discovered styles continue to show increasing usefulness, since all of them consistently align with one or more annotator preferences, meaning that they are able to capture the diversity in segmentations with an increasing level of granularity.

As an additional experiment to assess whether the learned styles are able to model tool-specific ("T1", "T2", "T3") latent factors, we separate the segmentations into three groups based on the tool, pass each corresponding image through a trained 3-StyleSeg model, and determine the predicted segmentation style with the highest overlap. We observe that the most commonly chosen style within a group is unique for each of the three tool groups, suggesting that differences among the three tools are learned within the three styles.

A new style alignment measure: When choosing only one style to evaluate each annotator preference, it is important to note that a particular style could be assigned as the chosen style for a certain annotator even if it best fits either 100% (perfect alignment) of the images or just slightly above random chance (weak alignment). We propose to measure this Annotator-Style Alignment Strength (AS^2) as 1 – 'normalized Shannon entropy of annotator-style assignment', i.e.,

$$\text{AS}^2 = 1 - \frac{-\sum_{i=1}^{M} q_i \log_2 q_i}{-\sum_{j=1}^{M} \frac{1}{M} \log_2 \frac{1}{M}}, \tag{9}$$

where q_i is the vector of fractions of segmentations assigned to each style (e.g., $q = [0.70, 0.15, 0.15]$ for a 3-StyleSeg model that assigns 100 images from a certain annotator preference as 70:15:15 images for styles 1, 2, and 3, yielding $\text{AS}^2 = 0.255$). Note that AS^2 is 0 for uniform assignments, and increases logarithmically approaching 1 as assignments become more consistent. Our results in Table 2 show that AS^2 values do not decrease as M increases, meaning that learning to model more styles is not detrimental to segmentation quality and indeed captures more diversity. Additional results are presented in Supp. Mat. Fig. SM1 (d, g).

4 Conclusion

We formulated the problem of segmentation style discovery in the absence of annotator correspondence. We showed how our proposed method, StyleSeg, discov-

ers segmentation styles that are diverse, semantically consistent, and more plausible than those generated by competing methods, as evaluated on four public skin lesion segmentation (SLS) datasets. We also curated ISIC-MultiAnnot, the largest multi-annotator SLS dataset with 13,555 image-mask pairs from 10 annotators from ISIC Archive, and showed how StyleSeg consistently achieves high agreement with the annotator preferences, as measured through the Dice coefficient as well as a newly proposed measure, Annotator-Style Alignment Strength, for measuring annotator-style alignment. Future work would include an explicit "disentanglement" of annotation styles from image content and approaches to find the optimal number of styles in a segmentation dataset.

Acknowledgments. The authors are grateful for the computational resources provided by NVIDIA Corporation and Digital Research Alliance of Canada (formerly Compute Canada). Partial funding for this project was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC RGPIN/06752-2020).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. International Skin Imaging Collaboration: Melanoma Project. https://www.isic-archive.com/, [Online. Accessed February 01, 2024]
2. Almazroa, A., Alodhayb, S., Osman, E., Ramadan, E., Hummadi, M., Dlaim, M., Alkatee, M., Raahemifar, K., Lakshminarayanan, V.: Agreement among ophthalmologists in marking the optic disc and optic cup in fundus images. International Ophthalmology **37**, 701–717 (2017)
3. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical Physics **38**(2), 915–931 (2011)
4. Asgari Taghanaki, S., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G.: Deep semantic segmentation of natural and medical images: A review. Artificial Intelligence Review **54**, 137–178 (2021)
5. Ballerini, L., Fisher, R.B., Aldridge, B., Rees, J.: A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In: Celebi, M.E., Schaefer, G. (eds.) Color Medical Image Analysis. vol. 6, pp. 63–86. Springer Netherlands (2013)
6. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960)
7. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
8. Ercal, F., Chawla, A., Stoecker, W.V., Lee, H.C., Moss, R.H.: Neural network diagnosis of malignant melanoma from color images. IEEE Transactions on Biomedical Engineering **41**(9), 837–845 (1994)
9. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin **76**(5), 378 (1971)

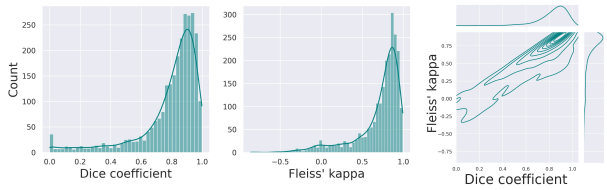
10. Fortina, A.B., Peserico, E., Silletti, A., Zattra, E.: Where's the naevus? inter-operator variability in the localization of melanocytic lesion border. Skin Research and Technology **18**(3), 311–315 (2012)
11. Glaister, J., Amelard, R., Wong, A., Clausi, D.A.: MSIM: Multistage illumination modeling of dermatological photographs for illumination-corrected skin lesion analysis. IEEE Transactions on Biomedical Engineering **60**(7), 1873–1883 (2013)
12. Ji, W., Yu, S., Wu, J., Ma, K., Bian, C., Bi, Q., Li, J., Liu, H., Cheng, L., Zheng, Y.: Learning calibrated medical image segmentation via multi-rater agreement modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12341–12351 (2021)
13. Kats, E., Goldberger, J., Greenspan, H.: A soft STAPLE algorithm combined with anatomical knowledge. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 510–517. Springer (2019)
14. Kawahara, J., Hamarneh, G.: Fully convolutional neural networks to detect clinical dermoscopic features. IEEE Journal of Biomedical and Health Informatics **23**(2), 578–585 (2018)
15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)
16. Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R.S., Rozeira, J.: PH^2 - a dermoscopic image database for research and benchmarking. In: IEEE Engineering in Medicine and Biology Society. pp. 5437–5440 (2013)
17. Mirikharaji, Z., Abhishek, K., Bissoto, A., Barata, C., Avila, S., Valle, E., Celebi, M.E., Hamarneh, G.: A survey on deep learning for skin lesion segmentation. Medical Image Analysis p. 102863 (2023)
18. Mirikharaji, Z., Abhishek, K., Izadi, S., Hamarneh, G.: D-LEMA: Deep learning ensembles from multiple annotations-application to skin lesion segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1837–1846 (2021)
19. Powers, D.M.W.: The problem with kappa. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 345–355 (2012)
20. Ribeiro, V., Avila, S., Valle, E.: Less is more: Sample selection and label conditioning improve skin lesion segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 738–739 (2020)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, part III 18. pp. 234–241. Springer (2015)
22. Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., Hager, G.D.: Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3591–3600 (2017)
23. Tschandl, P., Sinz, C., Kittler, H.: Domain-specific classification-pretrained fully convolutional network encoders for skin lesion segmentation. Computers in Biology and Medicine **104**, 111–116 (2019)
24. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging **23**(7), 903–921 (2004)

25. Zepf, K., Petersen, E., Frellsen, J., Feragen, A.: That label's got style: Handling label style bias for uncertain image segmentation. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=wZ2SVhOTzBX
26. Zhang, L., Tanno, R., Bronik, K., Jin, C., Nachev, P., Barkhof, F., Ciccarelli, O., Alexander, D.C.: Learning to segment when experts disagree. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 179–190. Springer (2020)

Supplementary Material

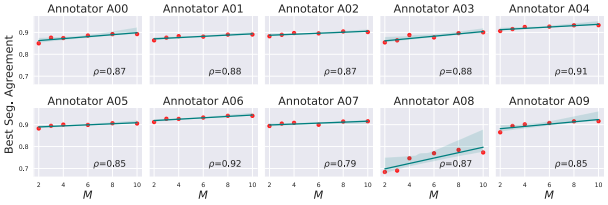
Implementation Details

The style classification model $f_c : (X_i, Y_{ik}) \in \mathbb{R}^{224\times224\times4} \rightarrow p_i \in \mathbb{R}^M$ is an ImageNet-pretrained ResNet-50 model with two modifications: it takes 4-channels (i.e., a concatenation of X_i and Y_{ik}) as input and produces a M-class prediction. The segmentation model $f_s : X_i \in \mathbb{R}^{224\times224\times3} \rightarrow \hat{Y}_i \in \mathbb{R}^{224\times224\times M}$ is an ImageNet-pretrained VGG-16 model with the fully connected layers removed, and multi-scale features resized, concatenated, and passed through a Conv2D layer with a sigmoid activation for binary mask prediction [14]. The images and masks were resized to 224×224 spatial resolution using nearest-neighbor interpolation. All models were trained for 10 epochs with a batch size of 4 using the Adam optimizer and a learning rate of 5e-5, and the epoch with the lowest loss $\mathcal{L}_{\text{total}}$ (Eqn. 8) on the validation set was used for evaluation. All models were trained on an Ubuntu 20.04 workstation with AMD Ryzen 9 5950X, 32 GB of RAM, and NVIDIA RTX 3090 GPU, running Python 3.10.13 and PyTorch 2.1.2. The PyTorch implementation of StyleSeg and more details about ISIC-MultiAnnot are available at https://github.com/sfu-mial/StyleSeg.

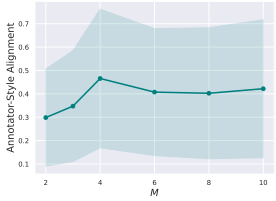


(a) Distribution of image-level pairwise Dice coefficient and Fleiss' kappa values in the 4,704 training image-mask pairs from ISIC Archive. Note how a considerable number of images have poor inter-annotator agreement (Dice $<$ 0.2 and Fleiss' kappa < 0).

(b) Undocumented annotation style variability (varying boundary granularity) in DermoFit's [5] ground truth masks.

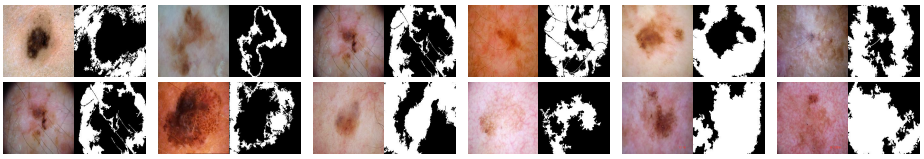


(c) StyleSeg segmentation performance (Dice) on ISIC-MultiAnnot, reported at annotator-level ("A00"–"A09"). Note how the Dice values are strongly correlated (Pearson correlation coefficient ρ) with the number of styles (M).

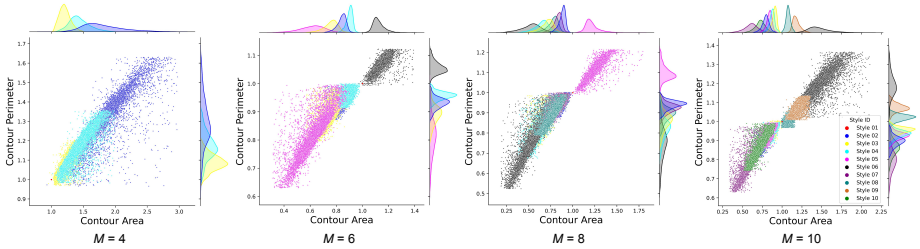


(d) AS^2 values remain high as M increases, meaning modeling more styles captures larger diversity.

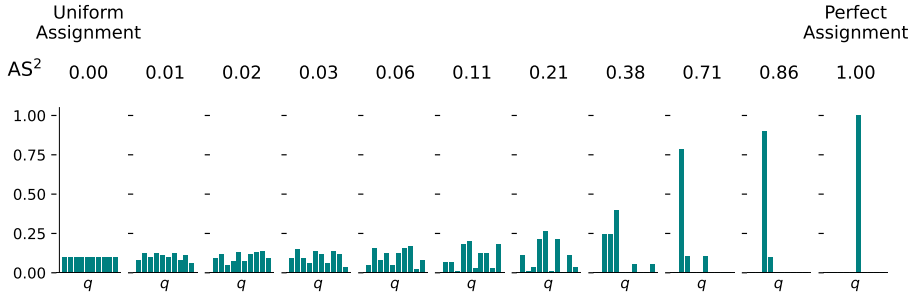
Fig. SM1: Supplementary Figures



(e) Images and "ground truth" segmentation masks from annotator "A08" in ISIC-MultiAnnot. Note the poor quality of segmentation, which in turn, affects evaluation.



(f) Evaluating semantic consistency of segmentation styles on ISIC Archive-Test. For all the 10,000 images, the shape features (area and perimeter of segmentation contours) of StyleSeg outputs are calculated, normalized per lesion w.r.t. the first style (red dot at (1.0, 1.0)), and the points and their kernel density estimates are colored by their style. Even for large values of M, the styles remain distinct. Note that the styles vary in their contour area indicating under- and over-segmentation of skin lesions. Also, for a certain contour area (A), the contour perimeter (P) varies, implying the styles differ in higher-order features such as border irregularity index and contour compactness, which is expected since both border irregularity $= \frac{P^2}{4\pi A}$ [8] and compactness $= \frac{\text{convex hull area}}{\text{A}}$ are also functions of perimeter and area.



(g) Annotator-Style Alignment Strength (AS^2) values for a variety of fractions of segmentations q (Eqn. 9) with $M = 10$. AS^2 is 1 in case of a perfect assignment and 0 in case of a uniform assignment. AS^2 increases logarithmically as q becomes more concentrated: for example, note that even for a low-entropy assignment such as $[0.9, 0.1, 0.0, \cdots 0.0]$, the Annotator-Style Alignment Strength drops to 0.86.

Fig. SM1: **(continued)** Supplementary Figures.