

Ninth ISIC Skin Image Analysis Workshop @ MICCAI 2024

Segmentation Style Discovery: Application to Skin Lesion Images



Kumar Abhishek[†]



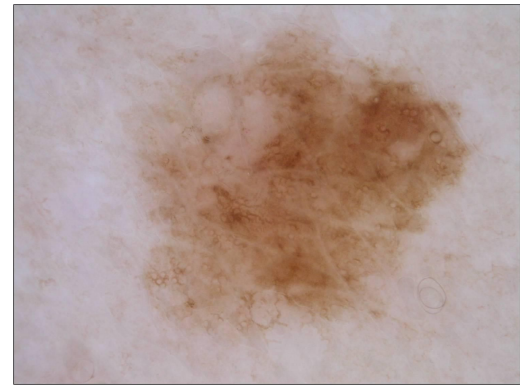
Jeremy Kawahara[‡]



Ghassan Hamarneh[†]



Variability in Medical Image Segmentation



Ambiguous object boundaries



Annotators' personal preferences



Annotators' skill levels

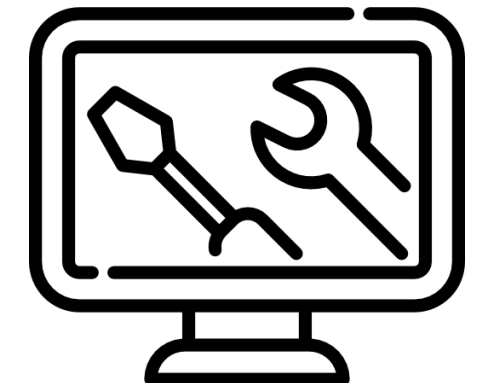
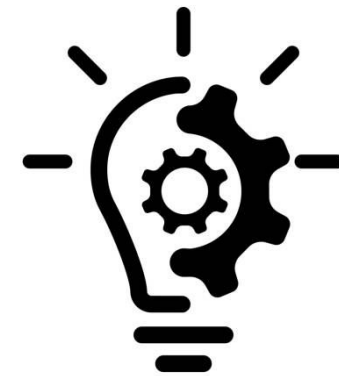
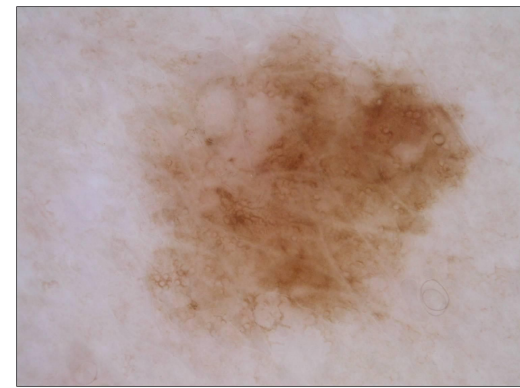


Segmentation criteria



Segmentation tools

Variability in Medical Image Segmentation



Ambiguous object boundaries

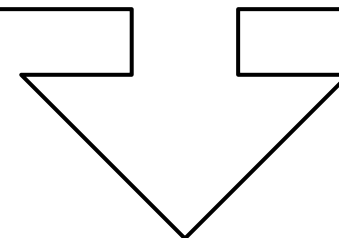
Annotators' personal preferences

Annotators' skill levels

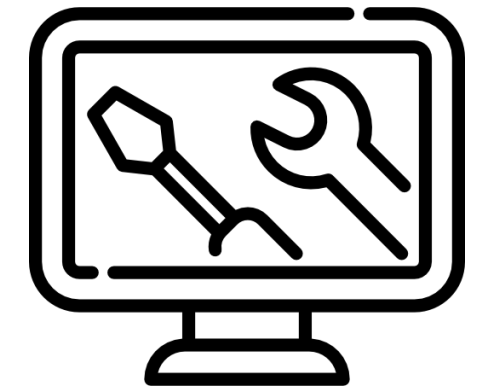
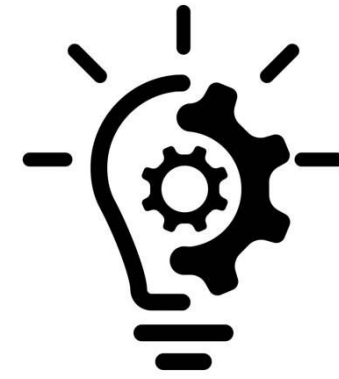
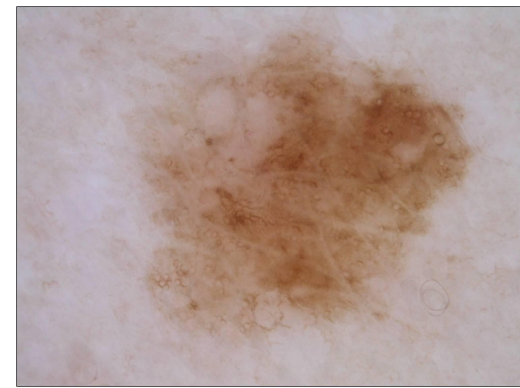
Segmentation criteria

Segmentation tools

Latent factors



Variability in Medical Image Segmentation



Ambiguous object boundaries

Annotators' personal preferences

Annotators' skill levels

Segmentation criteria

Segmentation tools

Latent factors

Different annotation segmentation preferences or **“styles”**

Methods for Learning from Multiple Annotations

SSeg methods model and learn to predict a single “gold standard” segmentation.

Methods for Learning from Multiple Annotations

SSeg methods model and learn to predict a single “gold standard” segmentation.

Abstract—In a double blind evaluation of 60 digital dermatoscopic images by 4 “junior”, 4 “senior” and 4 “expert” dermatologists (dermatology training respectively less than 1 year, between 1 and 5 years, and more than 5 years), a significant inter-operator variability was observed in melanocytic lesion border identification (with a disagreement of the order of 10 – 20% of the area of the lesions). Expert dermatologists

dard operative definition. For example, if even experienced dermatologists disagree on how to classify 5% of the area of an image, no automated system can be expected to classify “correctly” more than 95% of the area of that image.

Methods for Learning from Multiple Annotations

SSeg methods model and learn to predict a single “gold standard” segmentation.

MSeg methods model and predict multiple segmentations to capture annotation variability.

Abstract—In a double blind evaluation of 60 digital dermatoscopic images by 4 “junior”, 4 “senior” and 4 “expert” dermatologists (dermatology training respectively less than 1 year, between 1 and 5 years, and more than 5 years), a significant inter-operator variability was observed in melanocytic lesion border identification (with a disagreement of the order of 10 – 20% of the area of the lesions). Expert dermatologists

standard operative definition. For example, if even experienced dermatologists disagree on how to classify 5% of the area of an image, no automated system can be expected to classify “correctly” more than 95% of the area of that image.

Methods for Learning from Multiple Annotations

SSeg methods model and learn to predict a single “gold standard” segmentation.

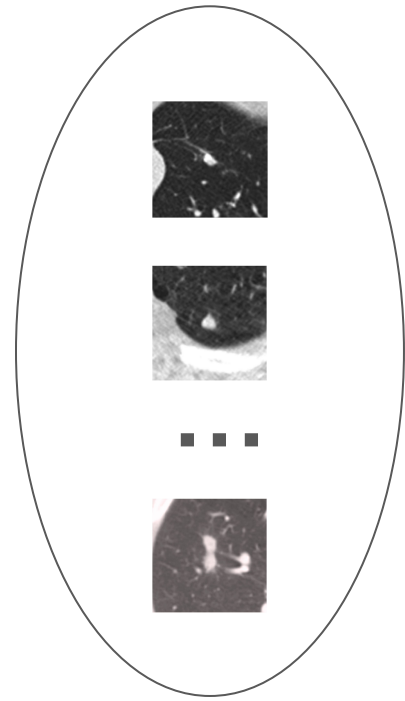
Abstract—In a double blind evaluation of 60 digital dermatoscopic images by 4 “junior”, 4 “senior” and 4 “expert” dermatologists (dermatology training respectively less than 1 year, between 1 and 5 years, and more than 5 years), a significant inter-operator variability was observed in melanocytic lesion border identification (with a disagreement of the order of 10 – 20% of the area of the lesions). Expert dermatologists

standard operative definition. For example, if even experienced dermatologists disagree on how to classify 5% of the area of an image, no automated system can be expected to classify “correctly” more than 95% of the area of that image.

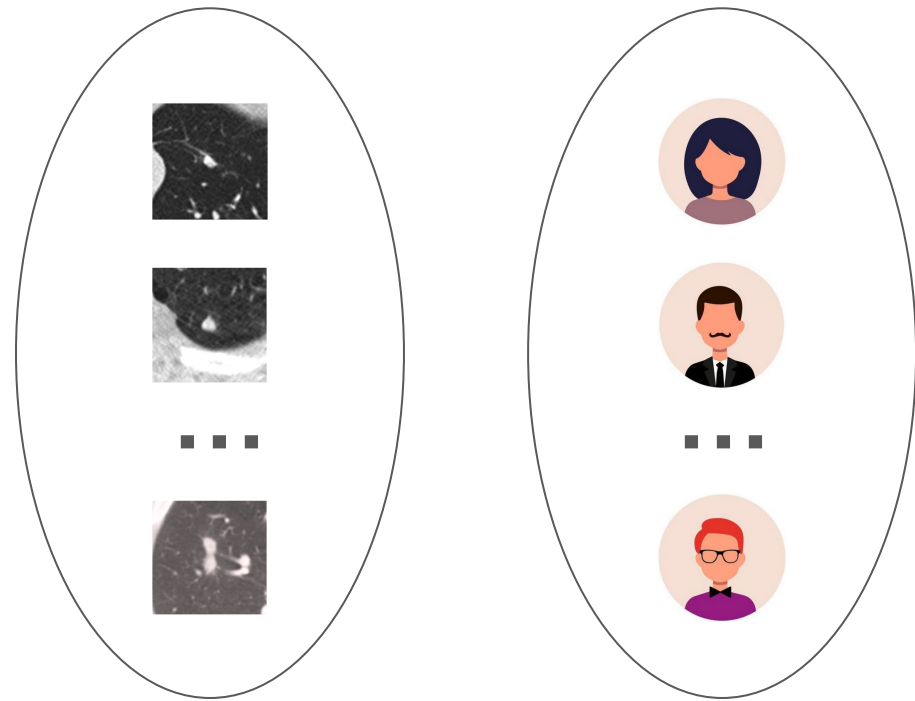
MSeg methods model and predict multiple segmentations to capture annotation variability.

Dataset requirement:
multi-annotator segmentations containing image-mask pairs with **annotator-segmentation correspondence.**

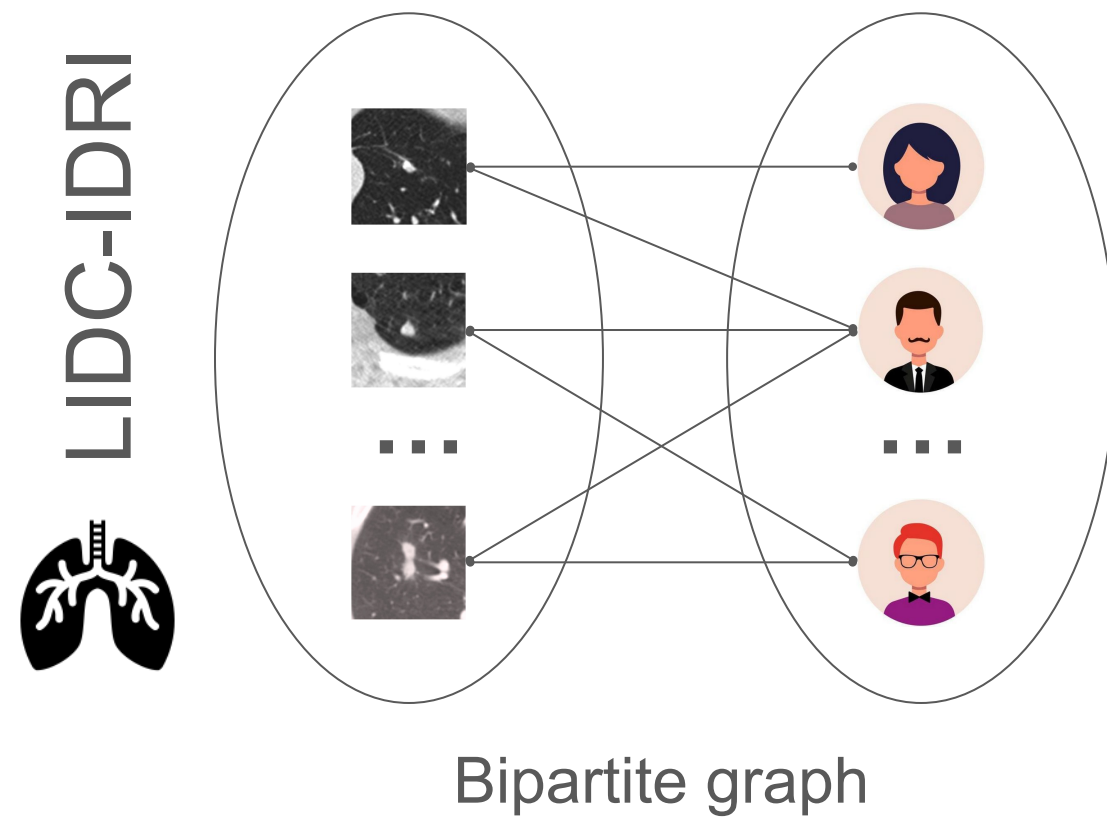
Multi-Annotator Medical Image Segmentation Datasets



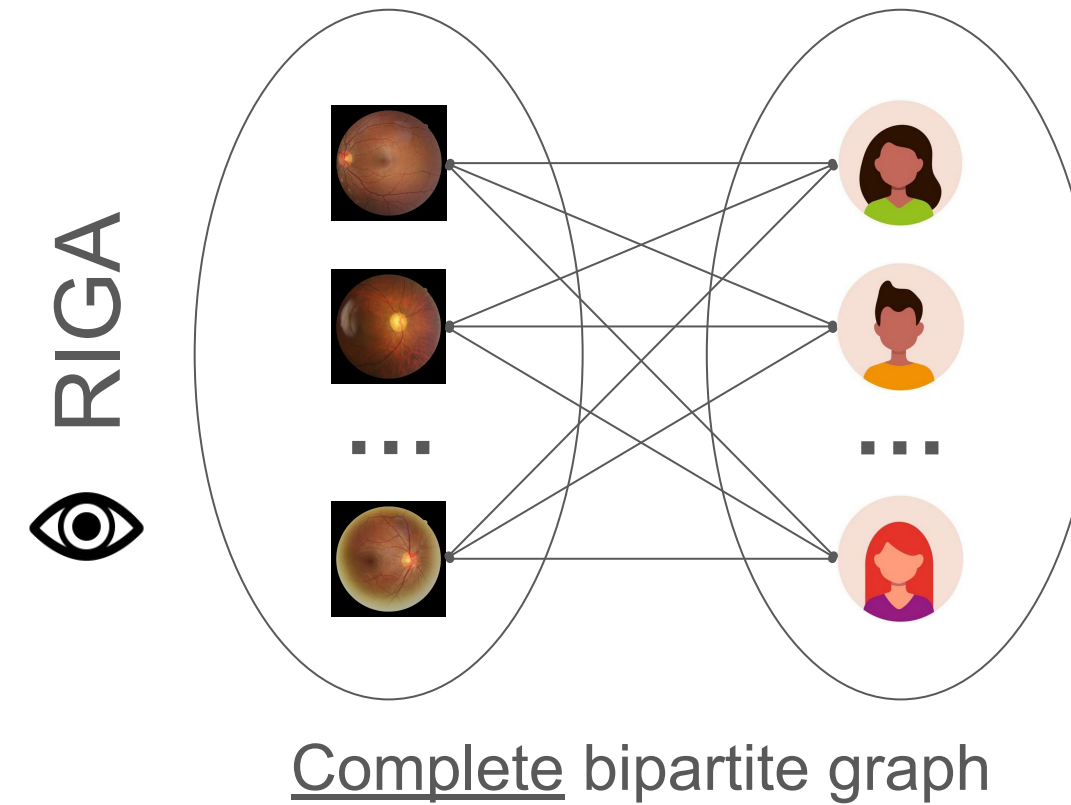
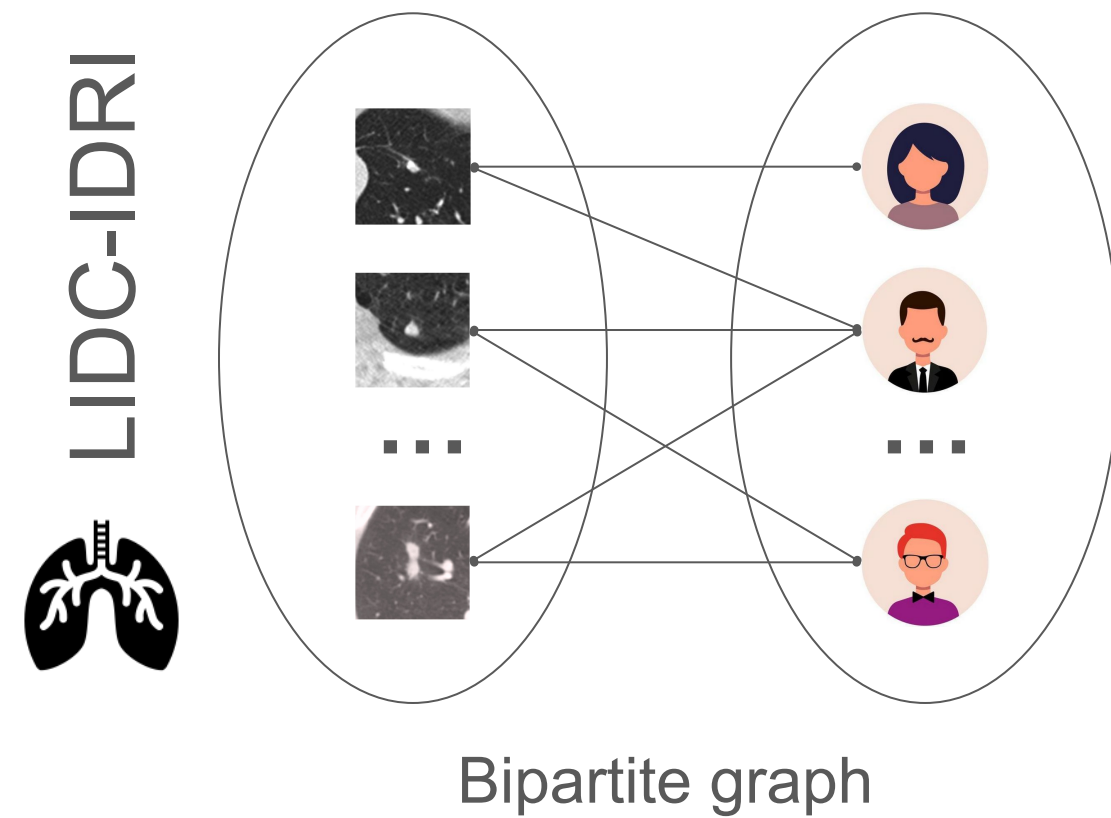
Multi-Annotator Medical Image Segmentation Datasets



Multi-Annotator Medical Image Segmentation Datasets



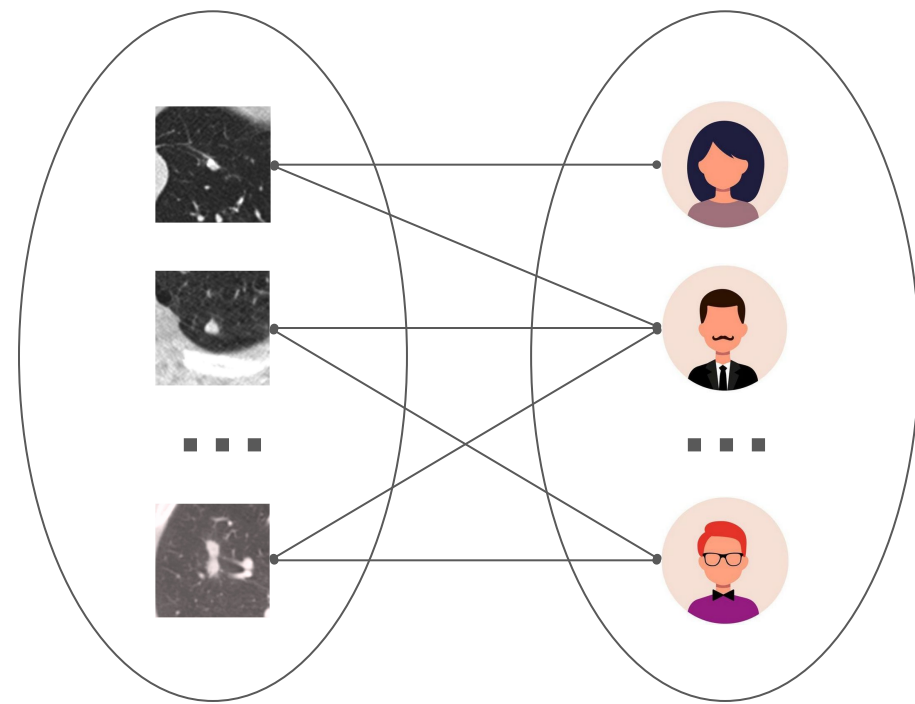
Multi-Annotator Medical Image Segmentation Datasets



Multi-Annotator Medical Image Segmentation Datasets



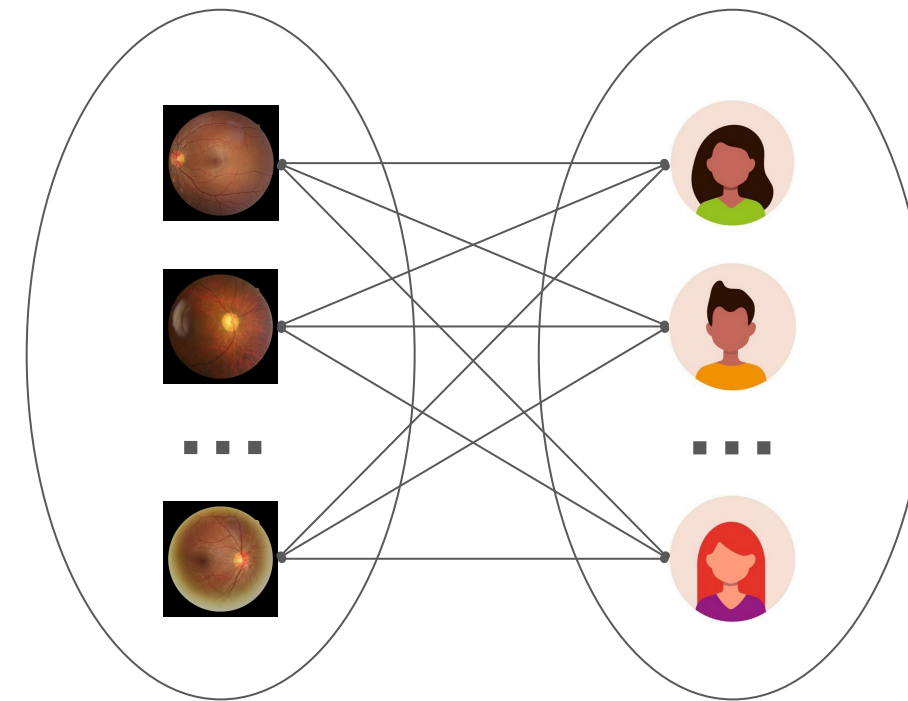
LIDC-IDRI



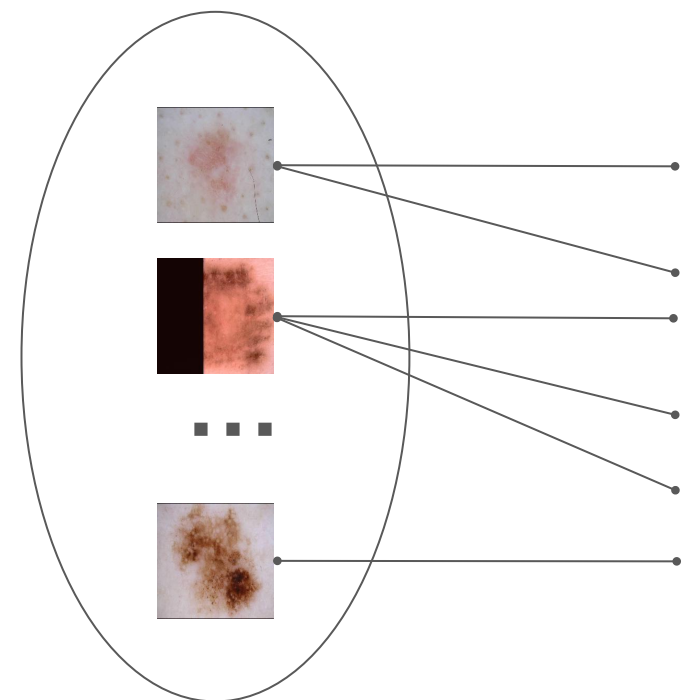
Bipartite graph



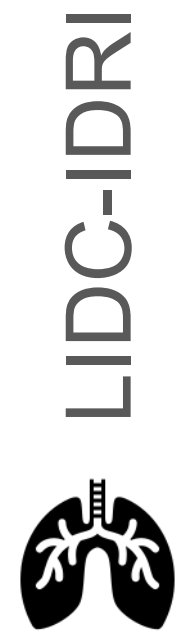
RIGA



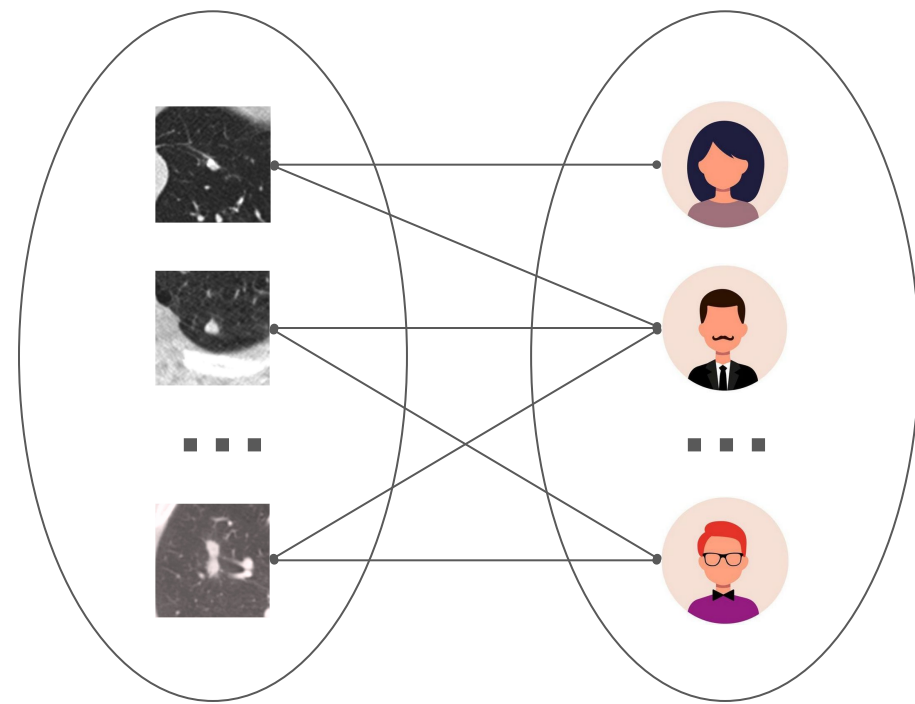
Complete bipartite graph



Multi-Annotator Medical Image Segmentation Datasets



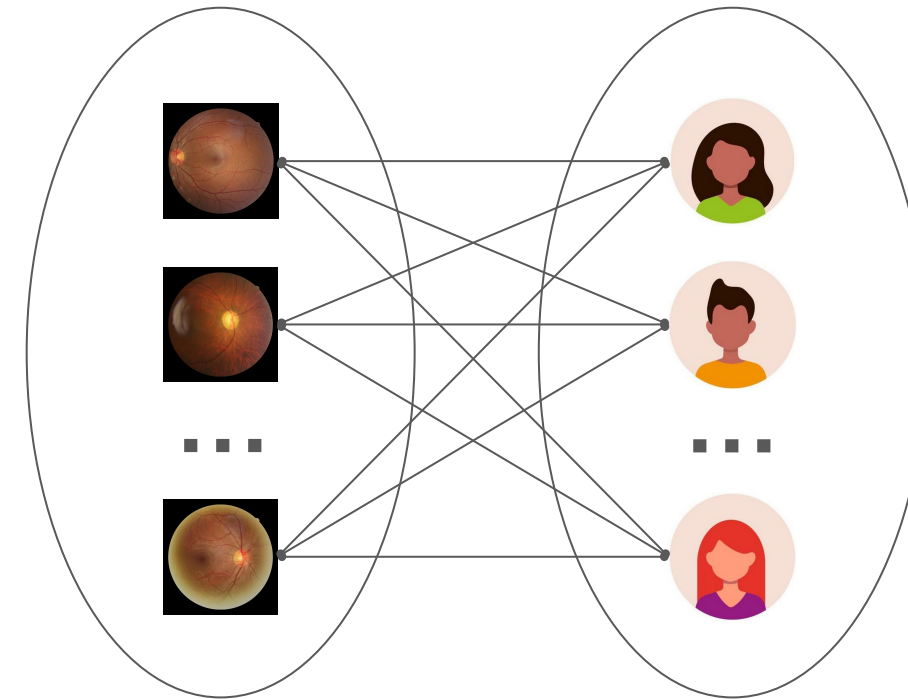
LIDC-IDRI



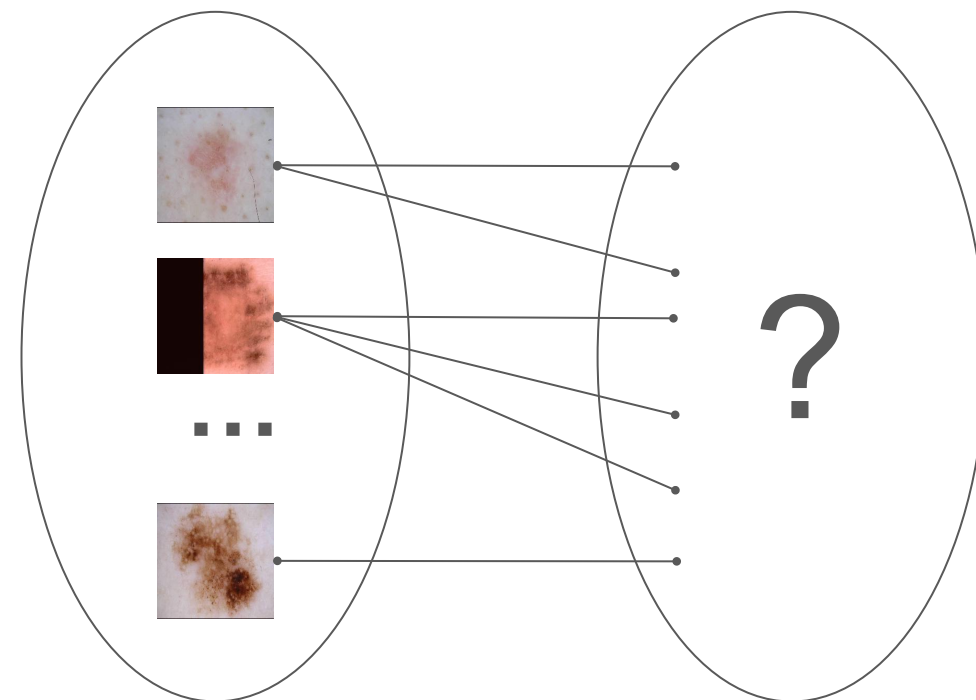
Bipartite graph



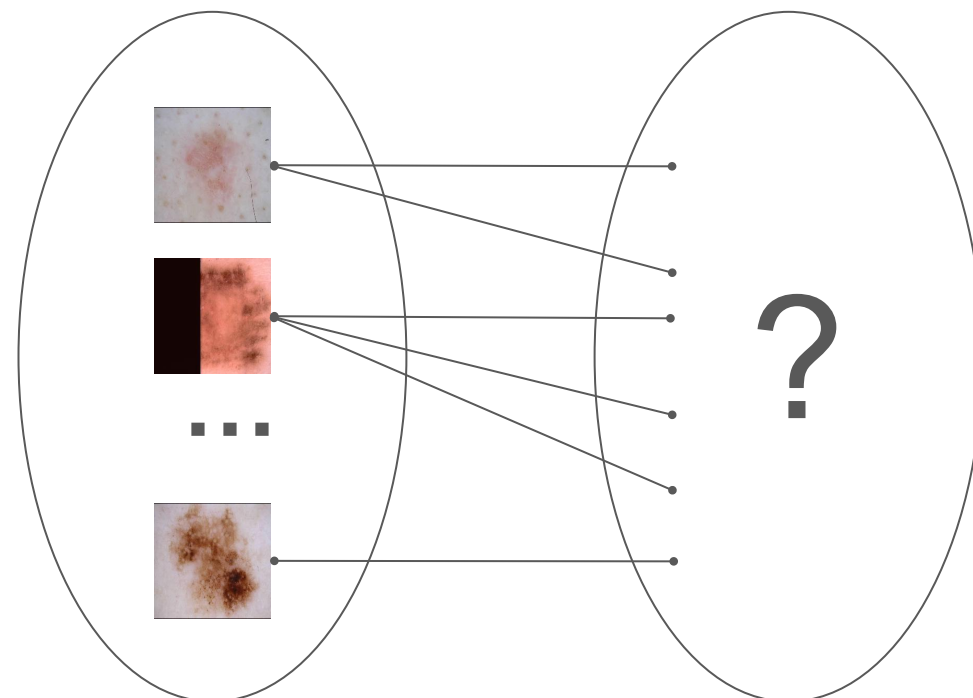
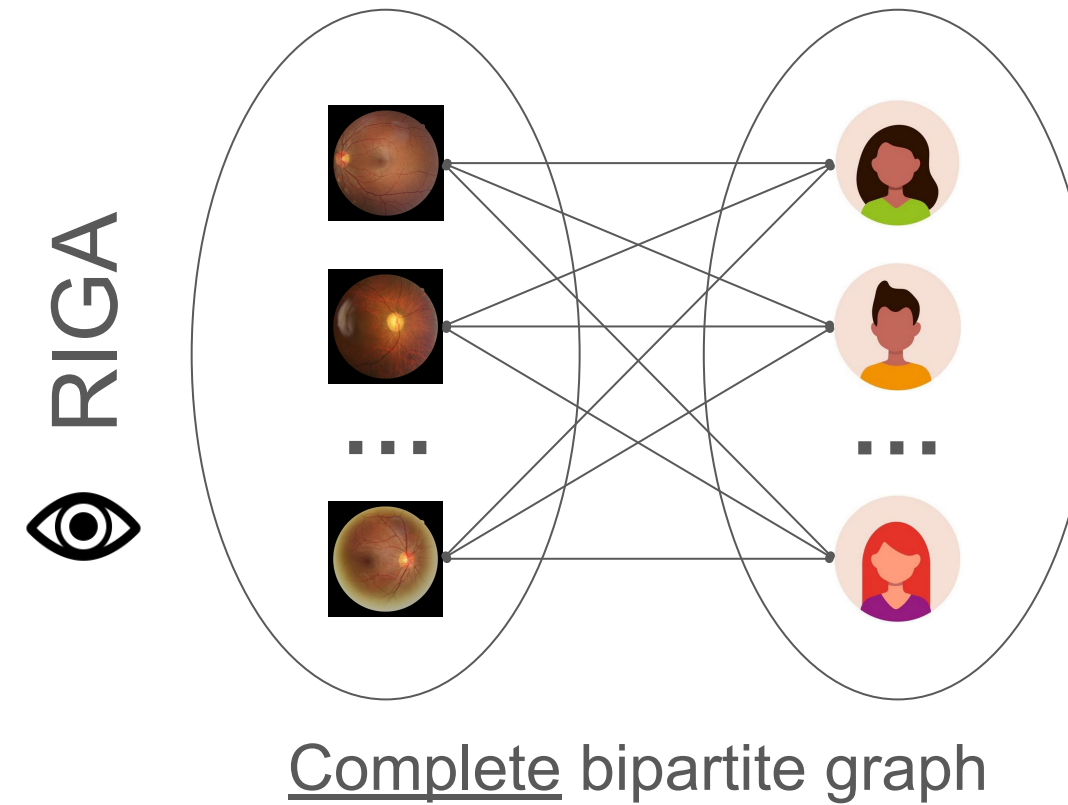
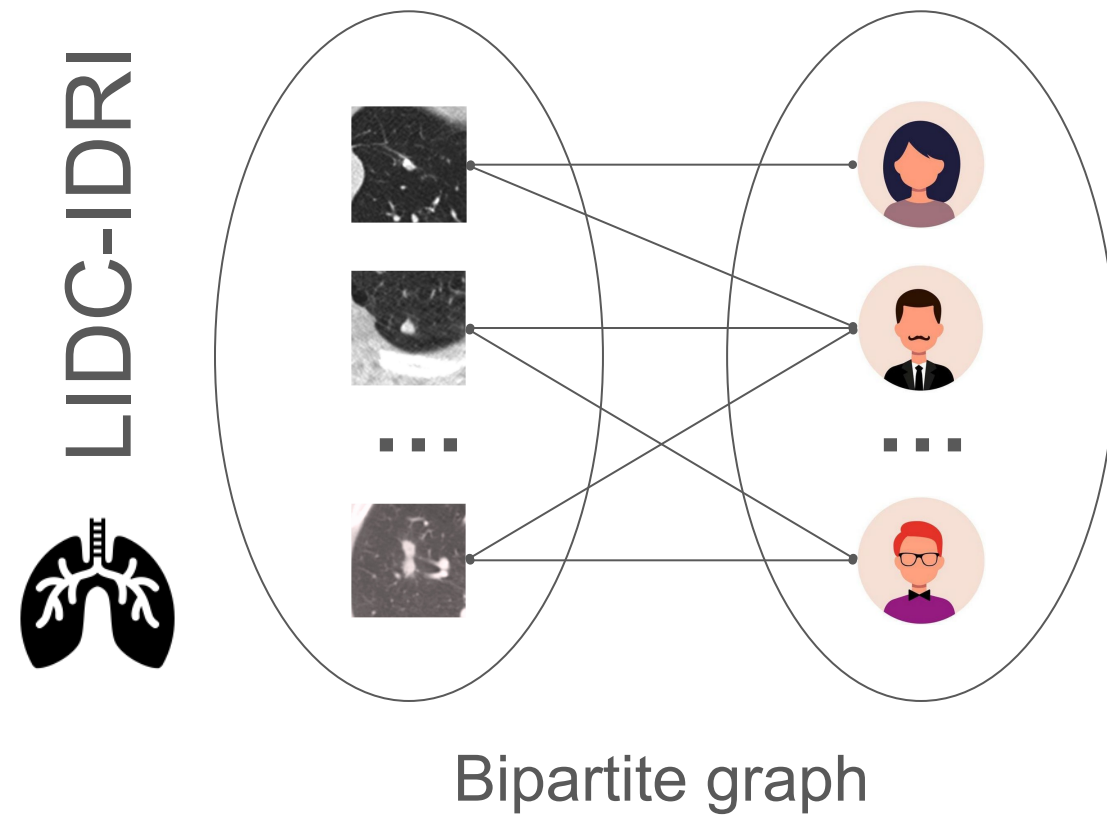
RIGA



Complete bipartite graph



Multi-Annotator Medical Image Segmentation Datasets



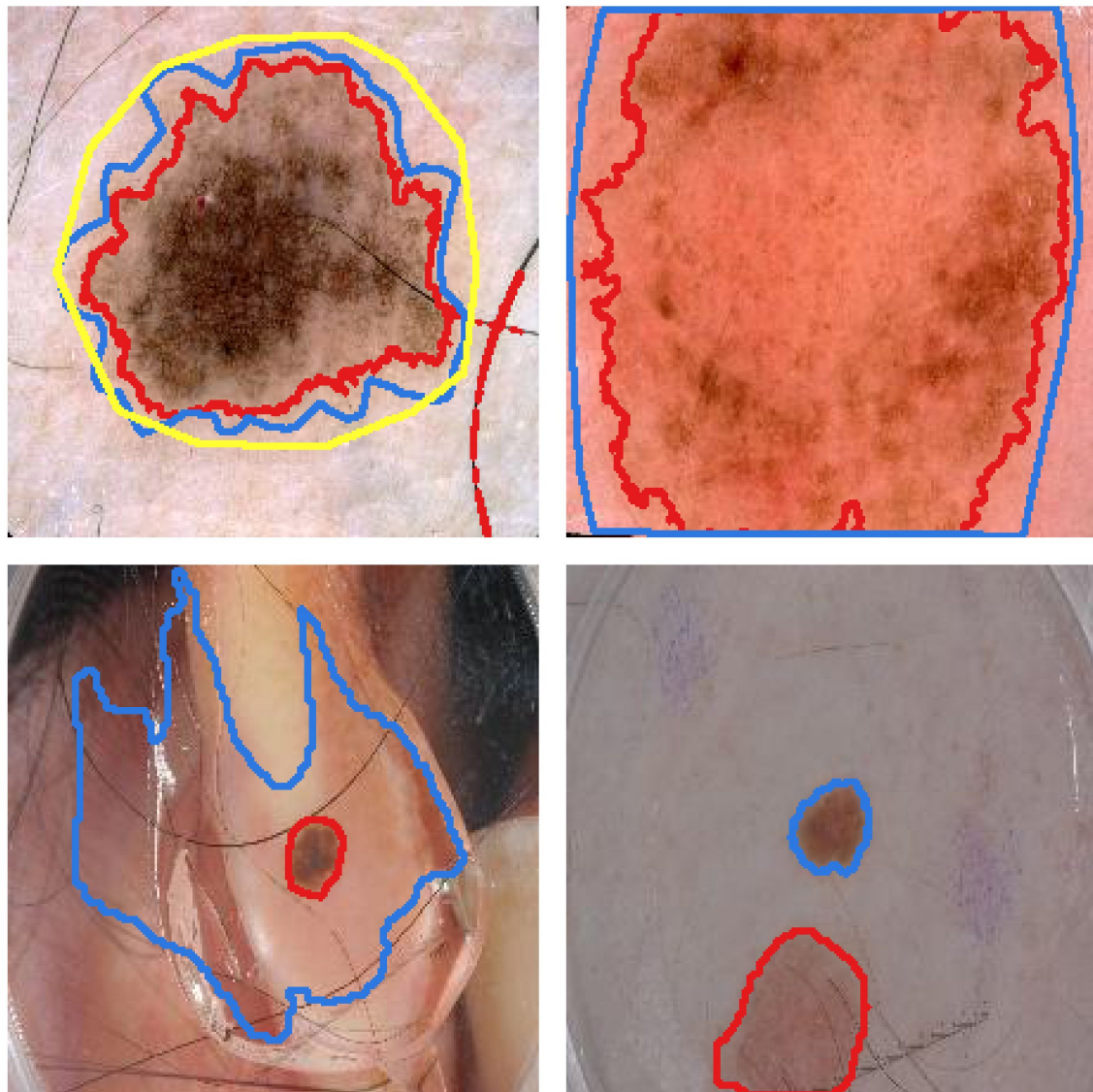
Latent factors unknown \Rightarrow difficult to define a segmentation “**style**”.

Segmentations in ISIC Archive and their Variability

2,261 images with more than 1 “ground truth” segmentation mask
⇒ 4,704 training image-mask pairs for skin lesion segmentation (**SLS**).

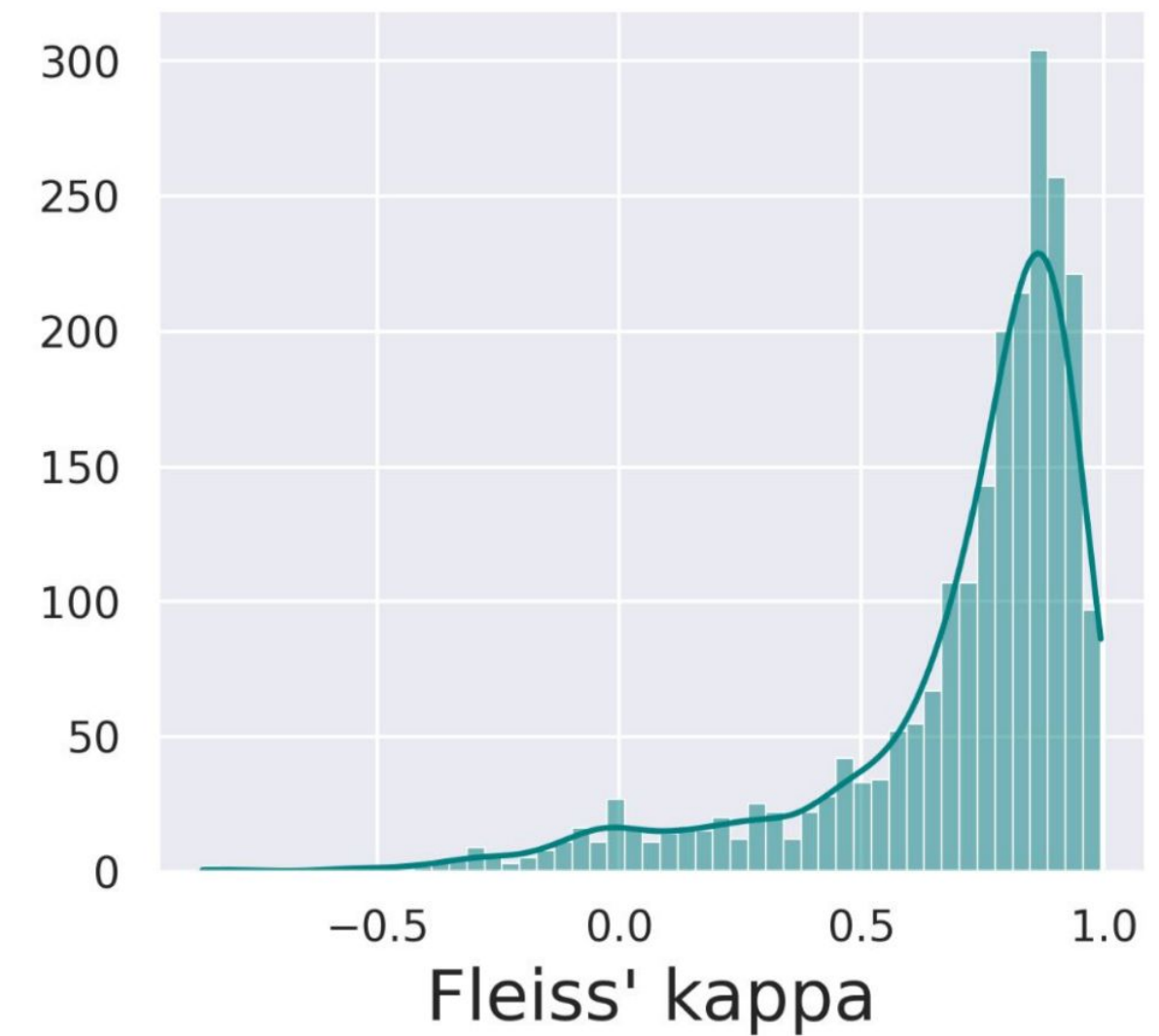
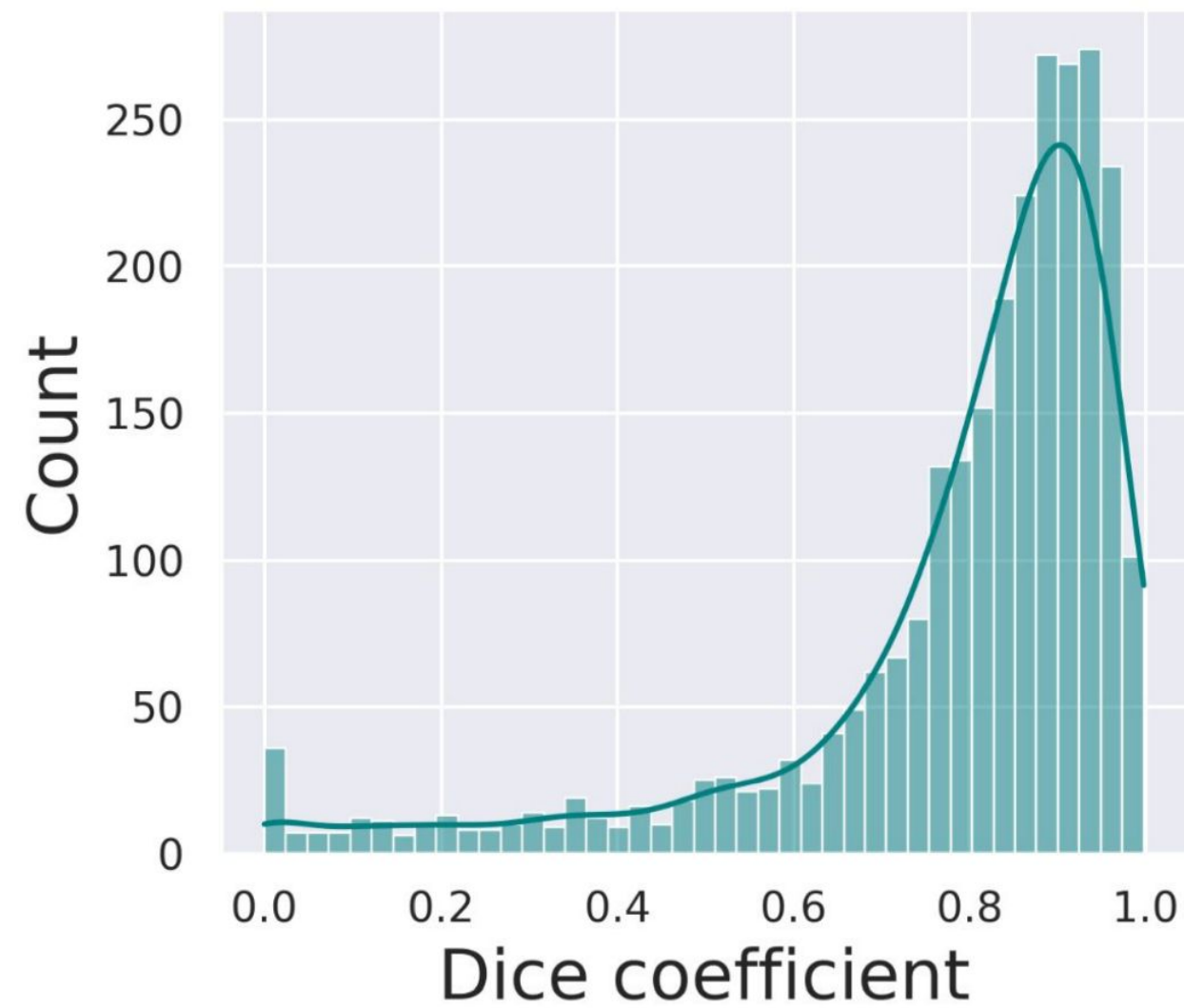
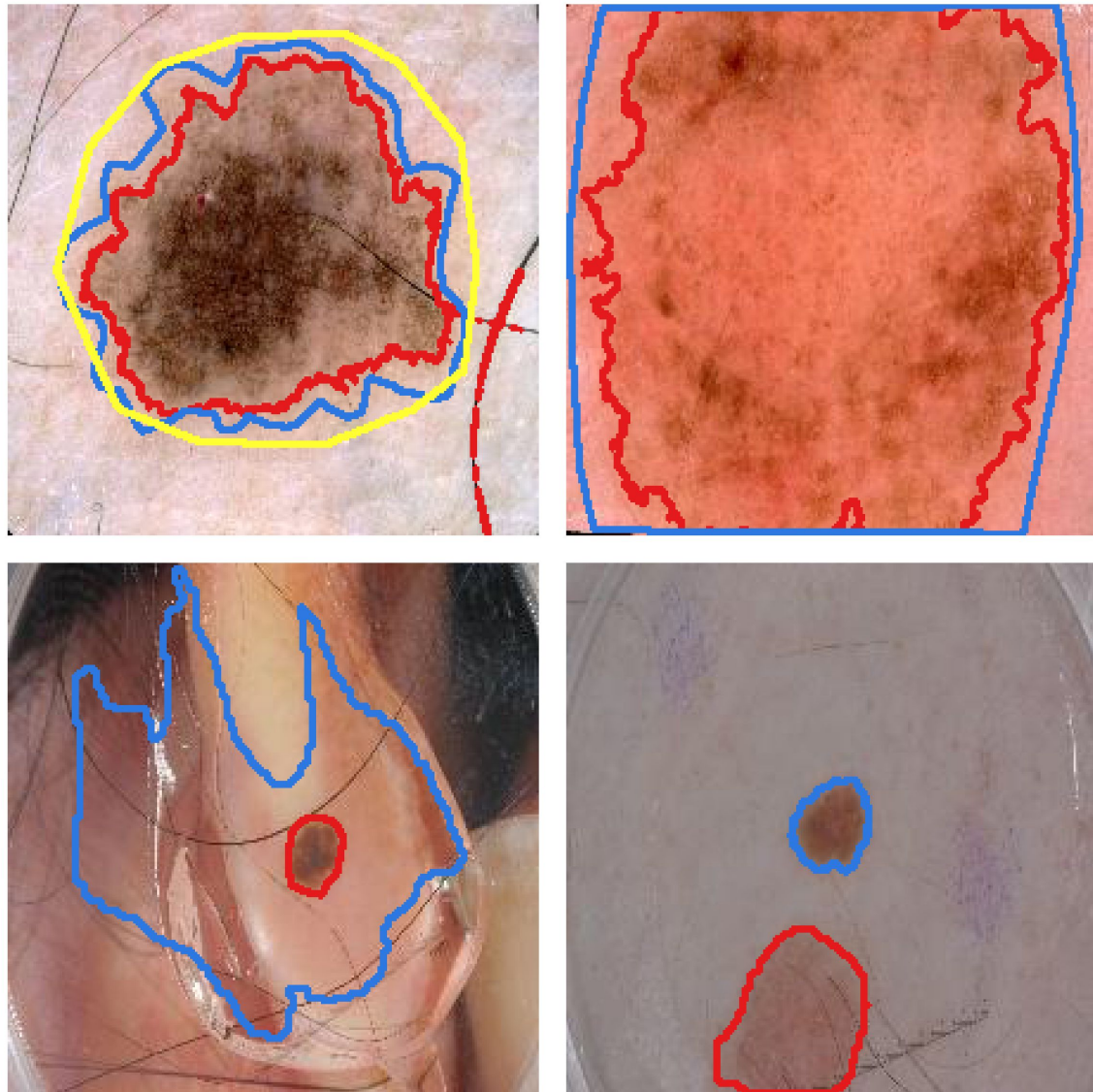
Segmentations in ISIC Archive and their Variability

2,261 images with more than 1 “ground truth” segmentation mask
⇒ 4,704 training image-mask pairs for skin lesion segmentation (**SLS**).



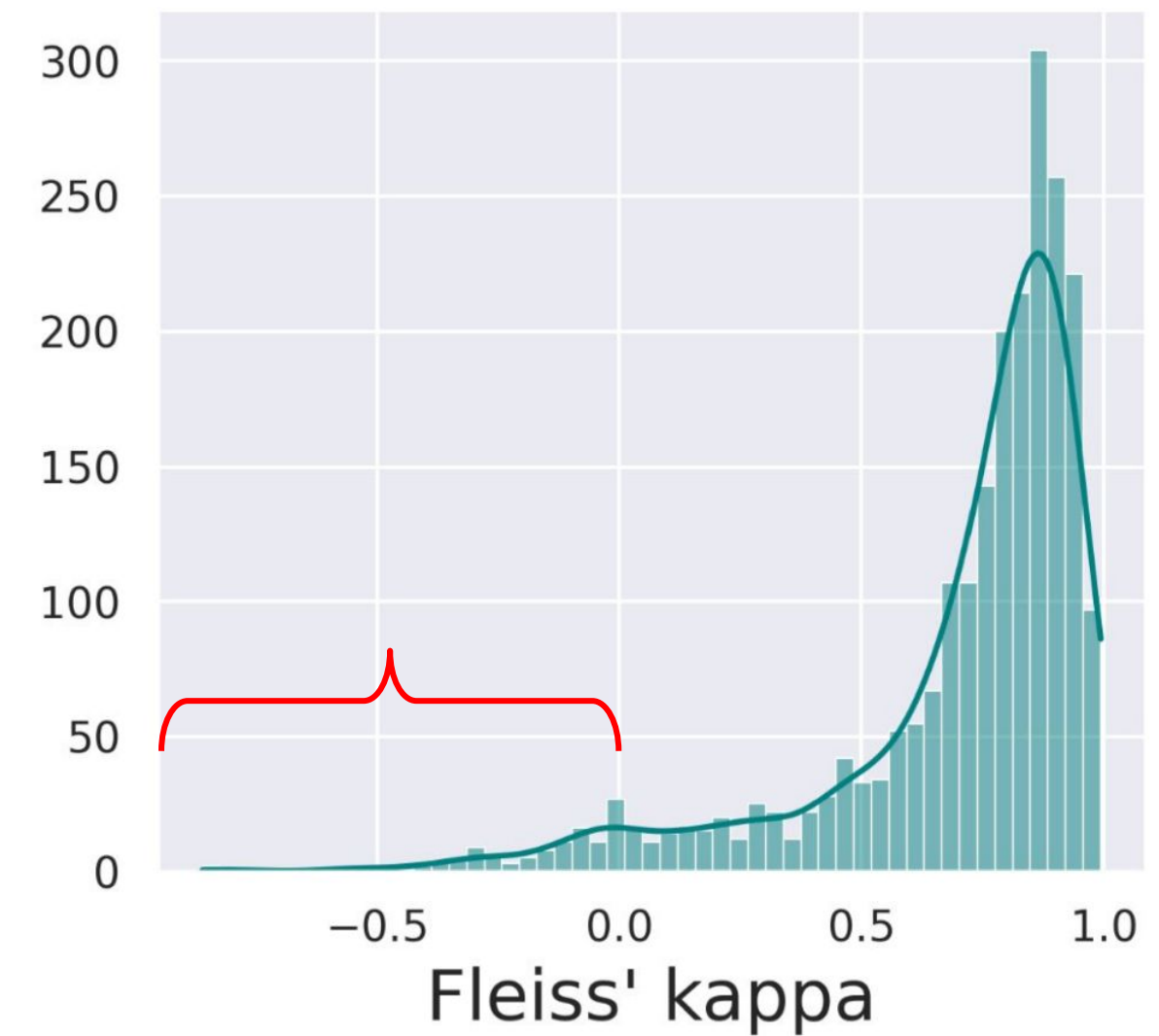
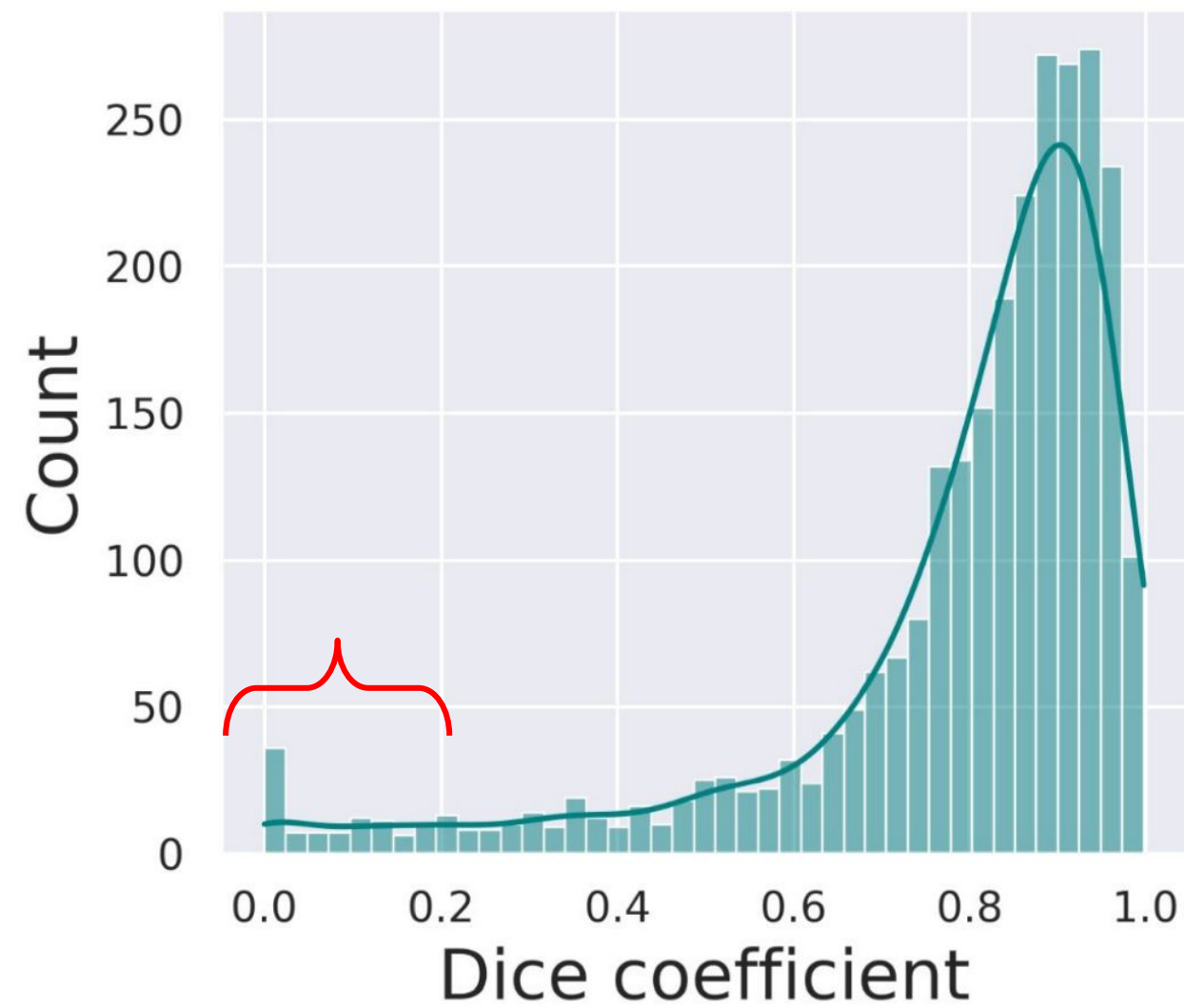
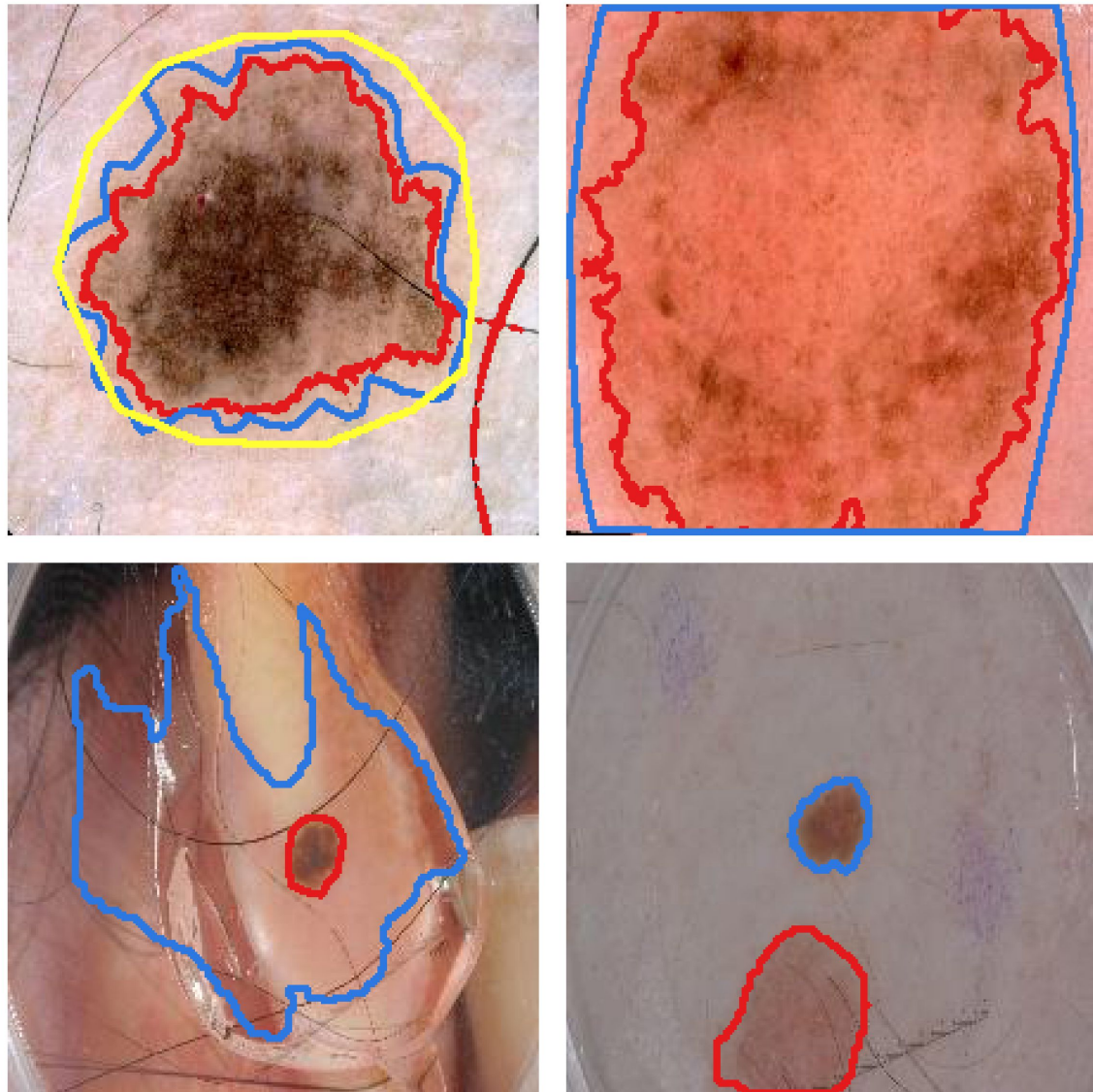
Segmentations in ISIC Archive and their Variability

2,261 images with more than 1 “ground truth” segmentation mask
⇒ 4,704 training image-mask pairs for skin lesion segmentation (**SLS**).

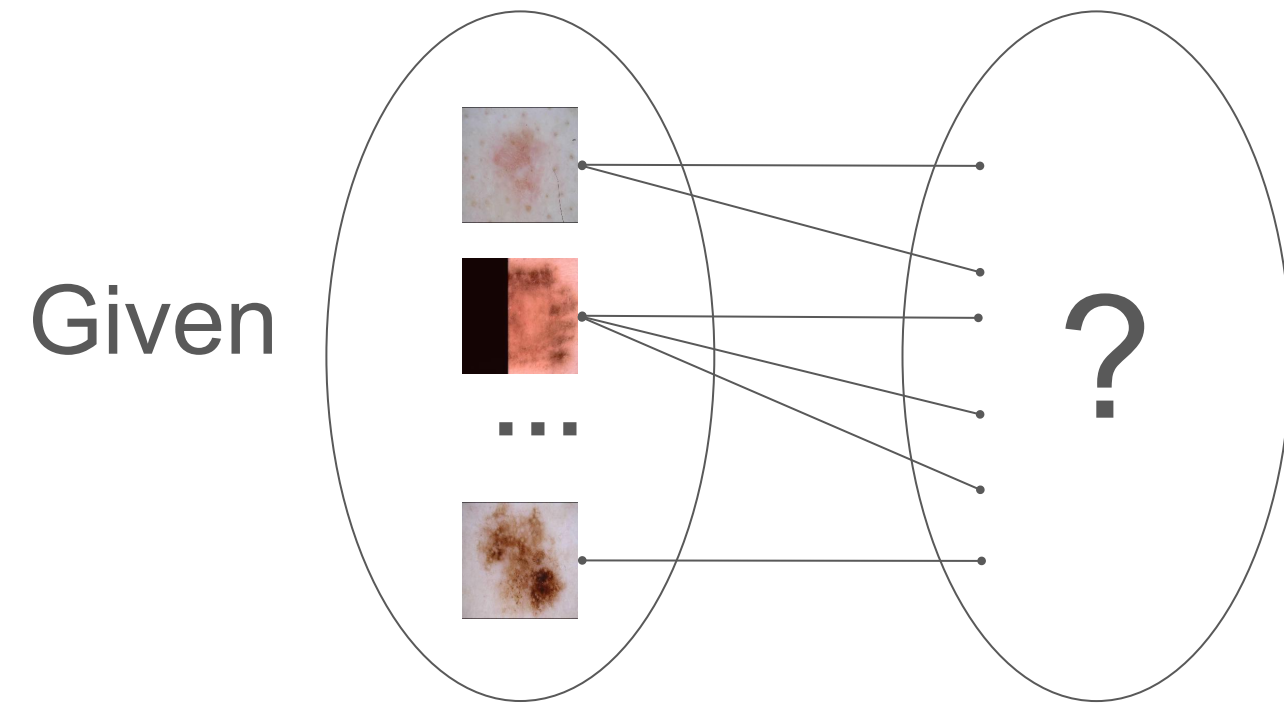


Segmentations in ISIC Archive and their Variability

2,261 images with more than 1 “ground truth” segmentation mask
⇒ 4,704 training image-mask pairs for skin lesion segmentation (**SLS**).

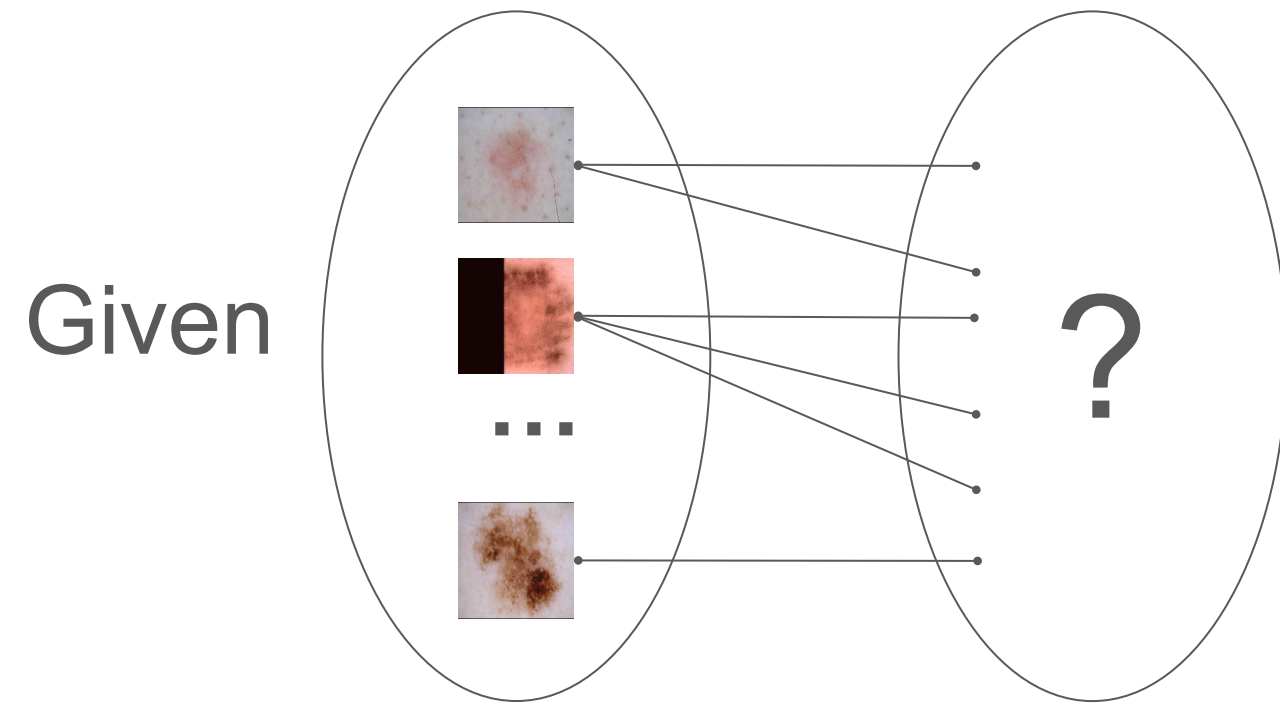


Objective



, train a model that **discovers unique annotation styles** such that:

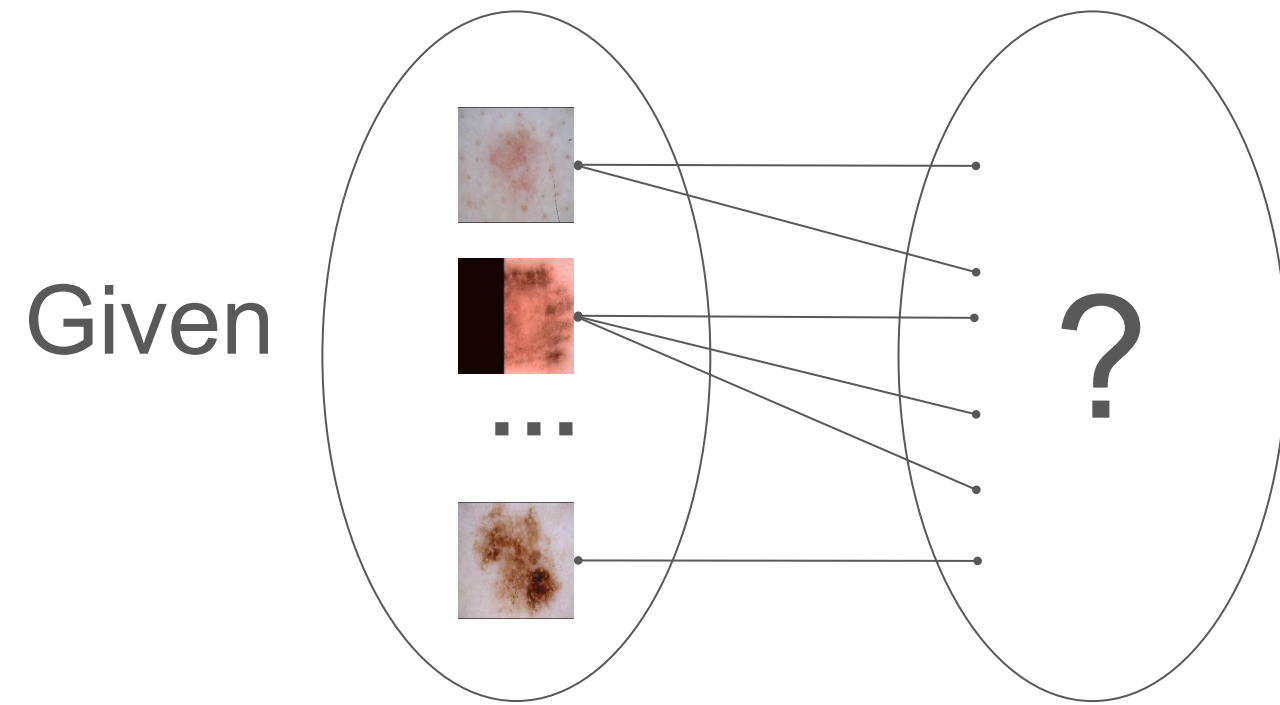
Objective



, train a model that **discovers unique annotation styles** such that:

- all the predicted segmentations are **plausible**,

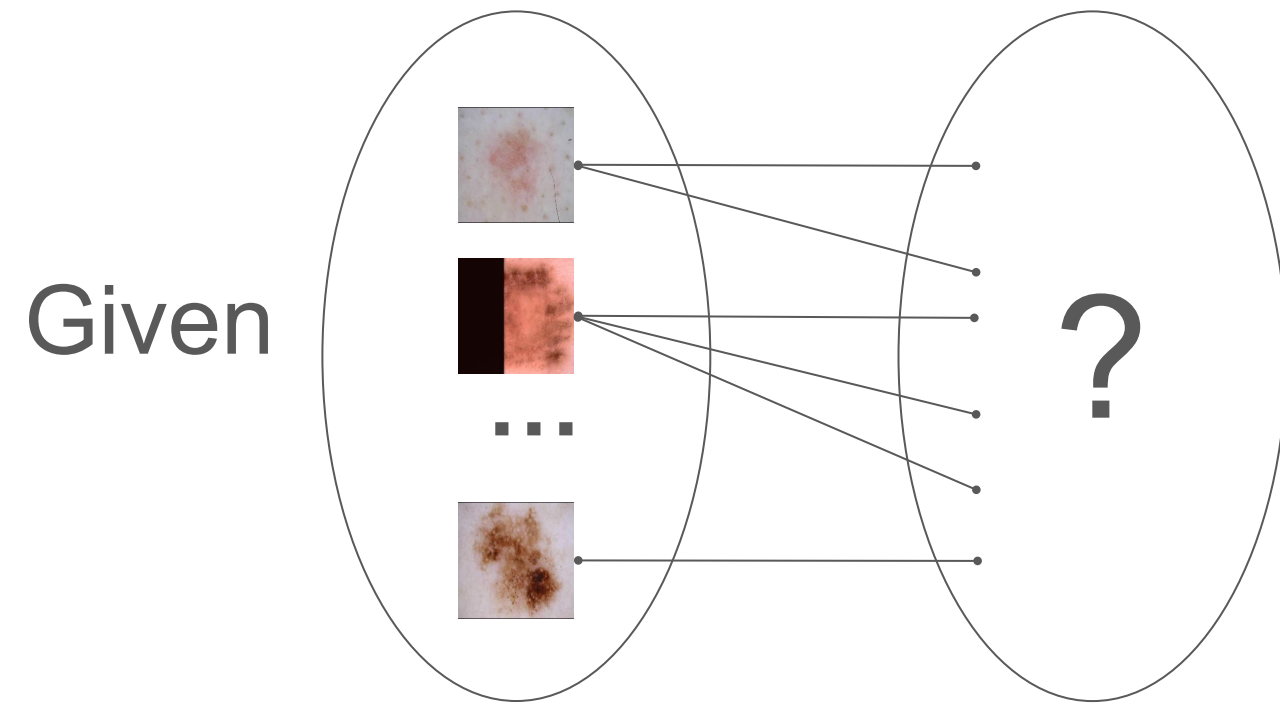
Objective



, train a model that **discovers unique annotation styles** such that:

- all the predicted segmentations are **plausible**,
- the predicted segmentations are **diverse**, and

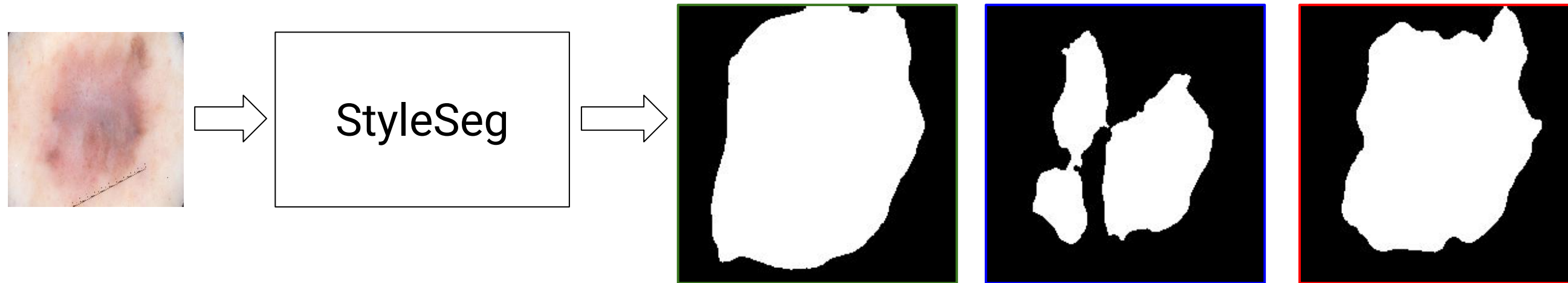
Objective



, train a model that **discovers unique annotation styles** such that:

- all the predicted segmentations are **plausible**,
- the predicted segmentations are **diverse**, and
- the segmentation styles are **semantically consistent** across all images.

StyleSeg produces multiple segmentation styles



Multiple segmentation styles and their probabilities

Multiple segmentation styles and their probabilities

Style 1

Style 2

Style 3

Segmentation
Model
 $f_s(X_i; \Theta_s)$

M segmentation
styles

Multiple segmentation styles and their probabilities

Style 1

Style 2

Style 3

Segmentation
Model
 $f_s(X_i; \Theta_s)$

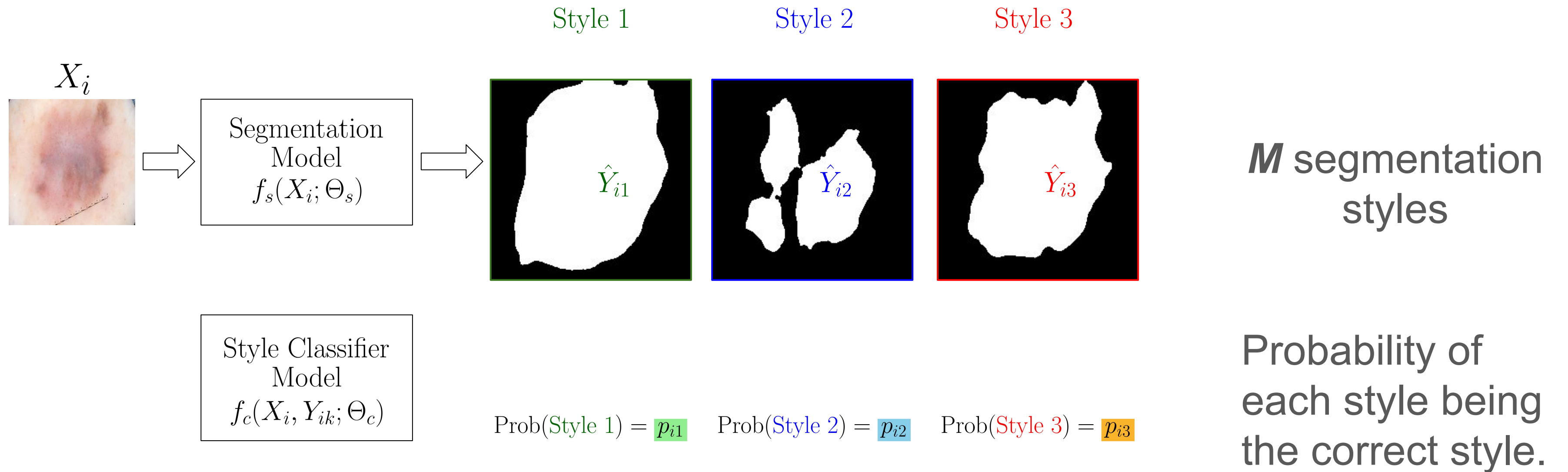
Style Classifier
Model
 $f_c(X_i, Y_{ik}; \Theta_c)$

M segmentation
styles

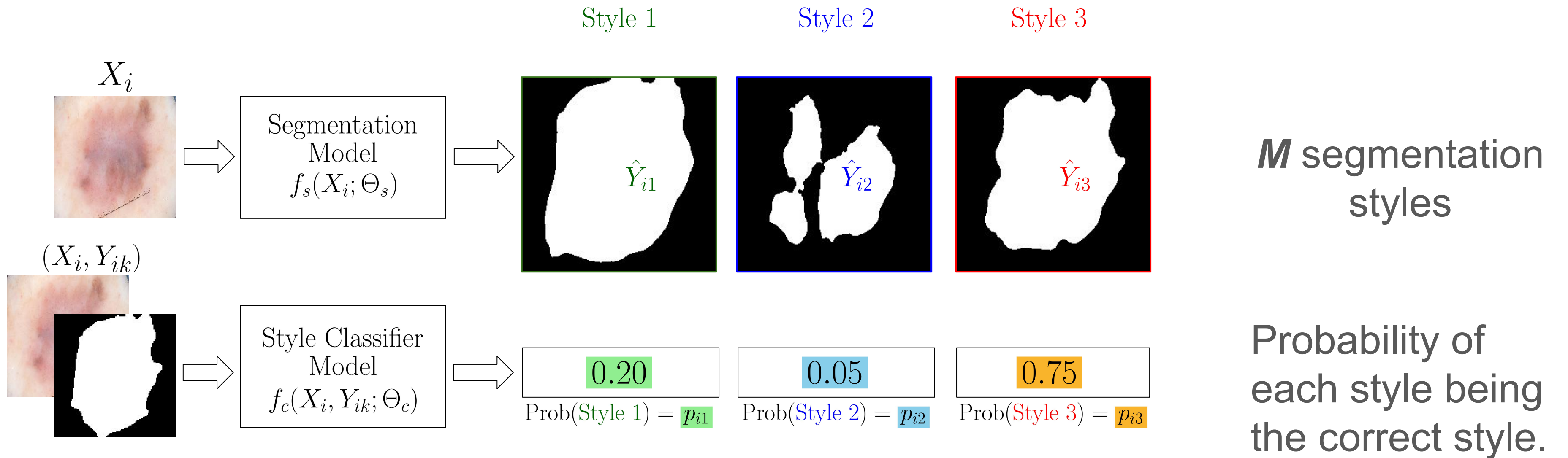
Prob(Style 1) = p_{i1} Prob(Style 2) = p_{i2} Prob(Style 3) = p_{i3}

Probability of
each style being
the correct style.

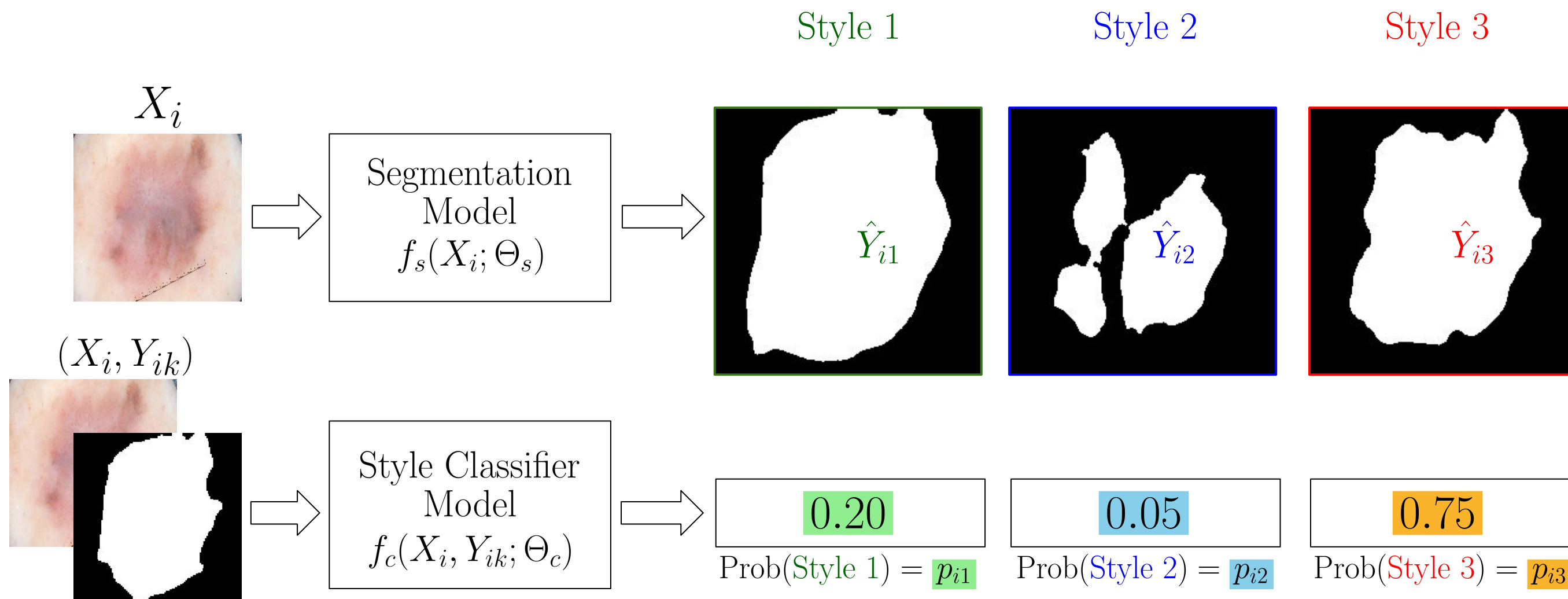
Multiple segmentation styles and their probabilities



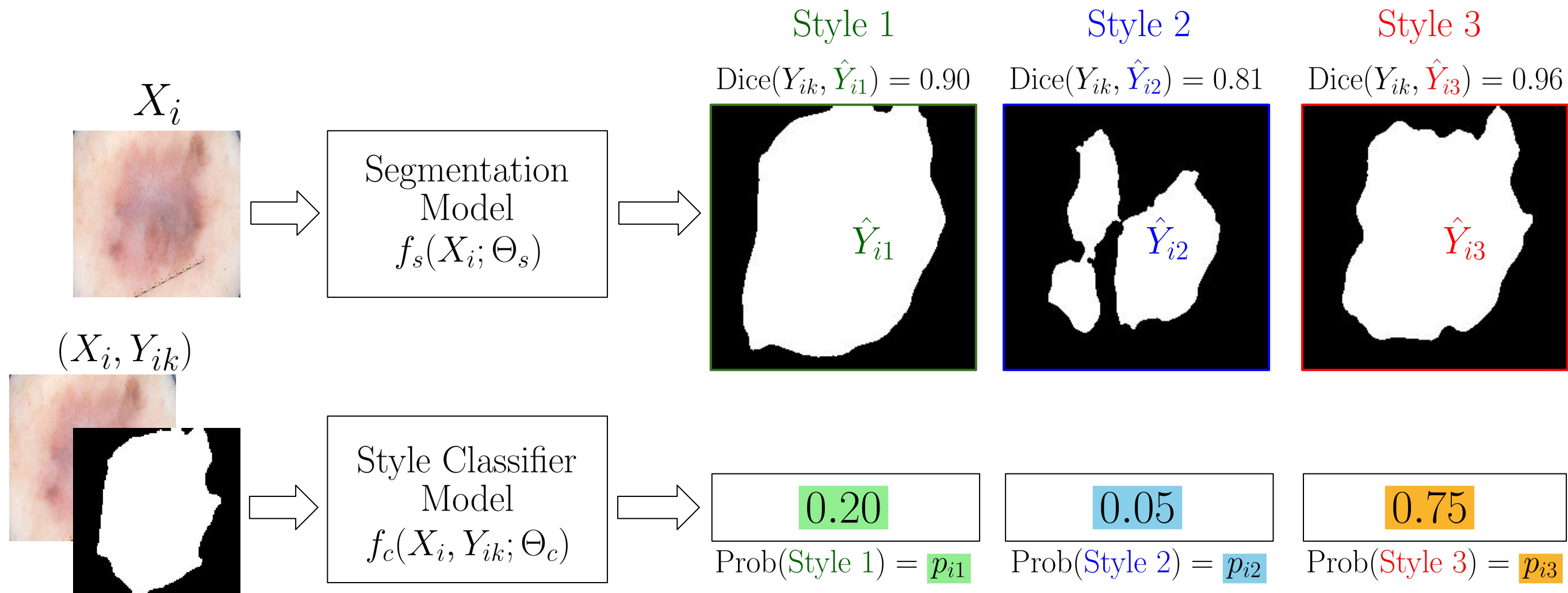
Multiple segmentation styles and their probabilities



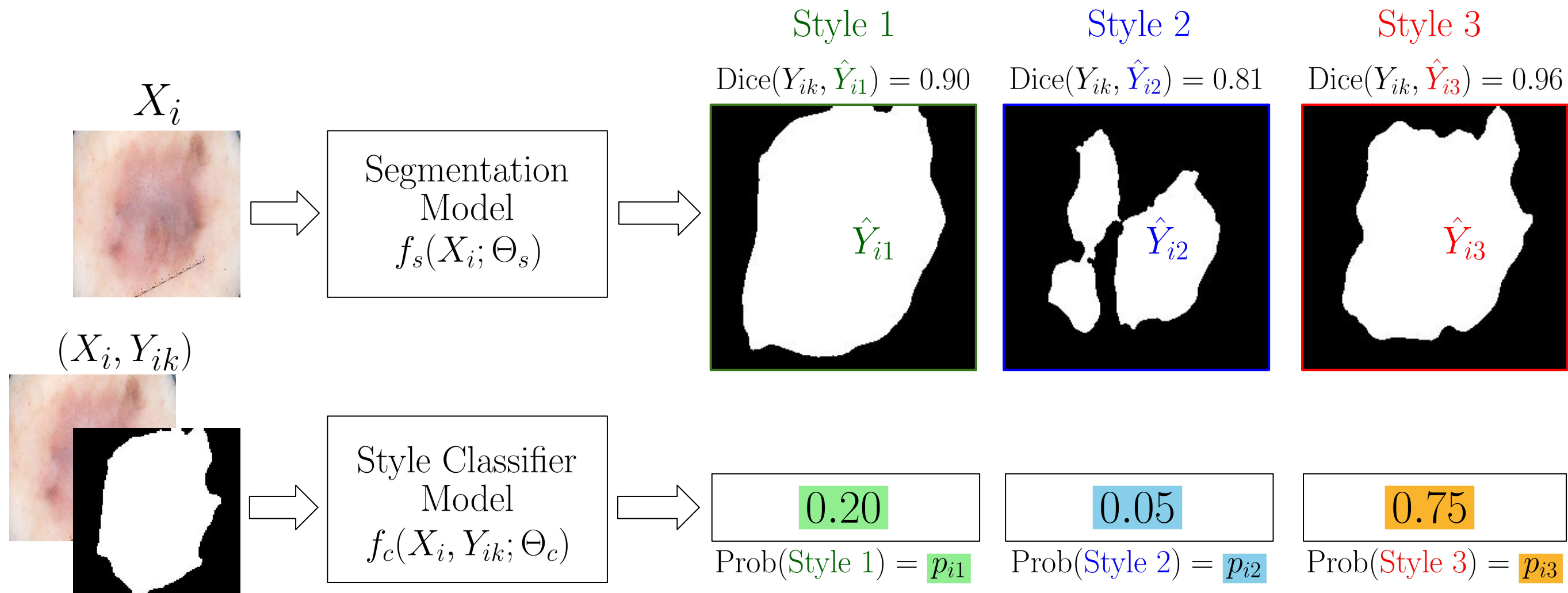
Training StyleSeg



Training StyleSeg



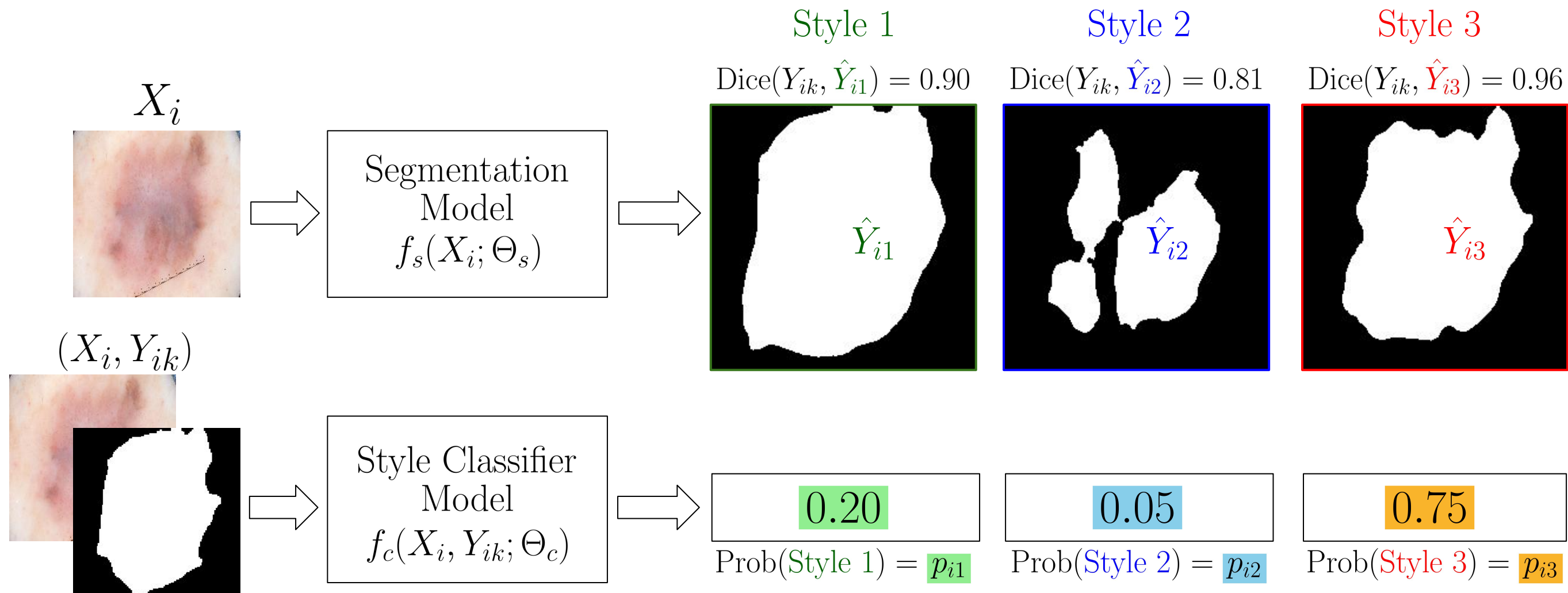
Training StyleSeg



$$m^* = \arg \max_j \text{Dice}(Y_{ik}, \hat{Y}_{ij}) = 3$$

$$\mathcal{L}_1 = L_D(Y_{ik}, \hat{Y}_{i3})$$

Training StyleSeg

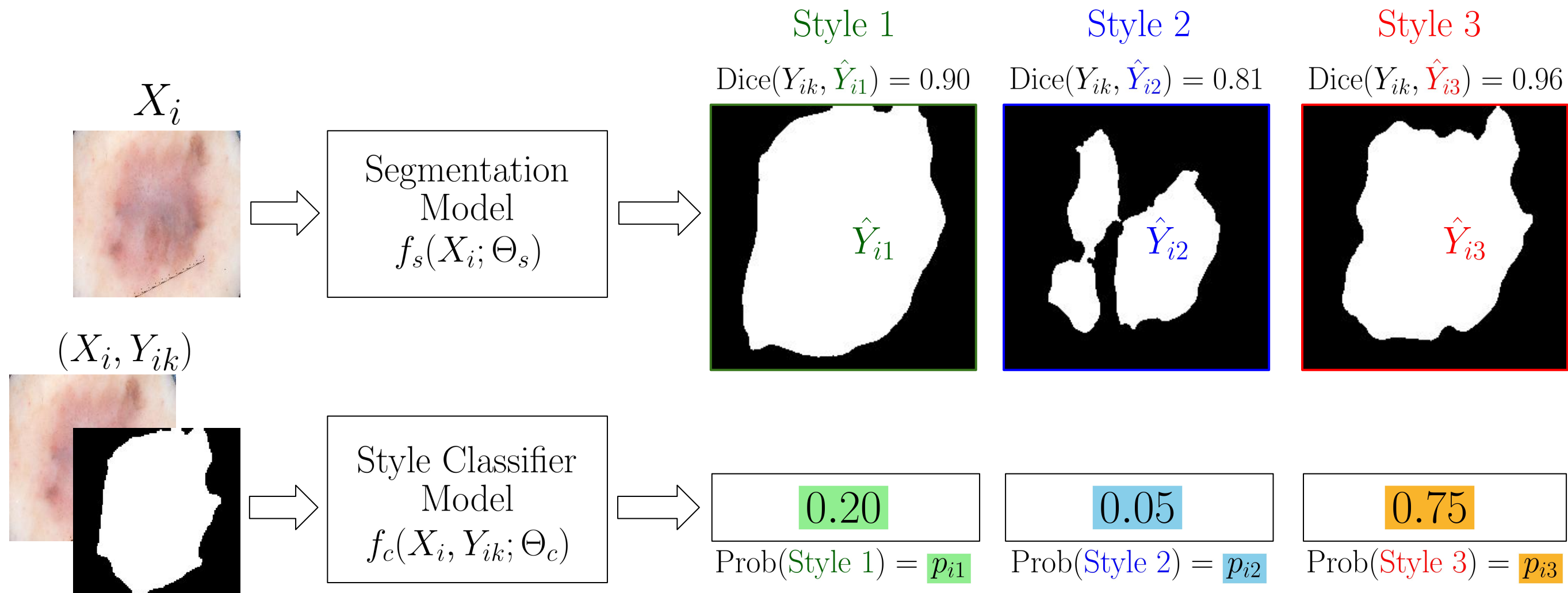


$$m^* = \arg \max_j \text{Dice}(Y_{ik}, \hat{Y}_{ij}) = 3$$

$$\mathcal{L}_1 = L_D(Y_{ik}, \hat{Y}_{i3})$$

$$\begin{aligned} \mathcal{L}_2 = L_D(Y_{ik}, & (0.20 * \hat{Y}_{i1}) \\ & + (0.05 * \hat{Y}_{i2}) \\ & + (0.75 * \hat{Y}_{i3})) \end{aligned}$$

Training StyleSeg



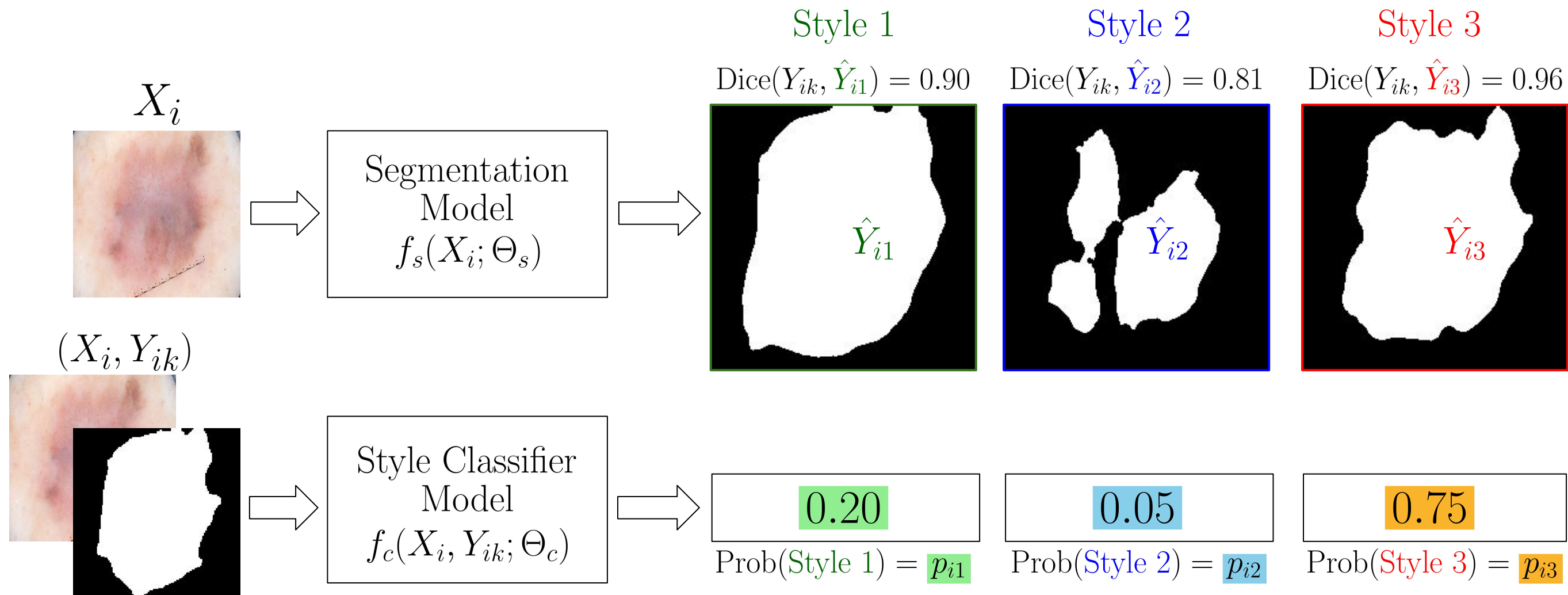
$$m^* = \arg \max_j \text{Dice}(Y_{ik}, \hat{Y}_{ij}) = 3$$

$$\mathcal{L}_1 = L_D(Y_{ik}, \hat{Y}_{i3})$$

$$\begin{aligned} \mathcal{L}_2 = L_D(Y_{ik}, & (0.20 * \hat{Y}_{i1}) \\ & + (0.05 * \hat{Y}_{i2}) \\ & + (0.75 * \hat{Y}_{i3})) \end{aligned}$$

$$\mathcal{L}_3 = L_{CE} \left(\begin{bmatrix} 0.20, 0.05, 0.75 \\ 0, 0, 1 \end{bmatrix} \right)$$

Training StyleSeg



$$m^* = \arg \max_j \text{Dice}(Y_{ik}, \hat{Y}_{ij}) = 3$$

$$\mathcal{L}_1 = L_D(Y_{ik}, \hat{Y}_{i3})$$

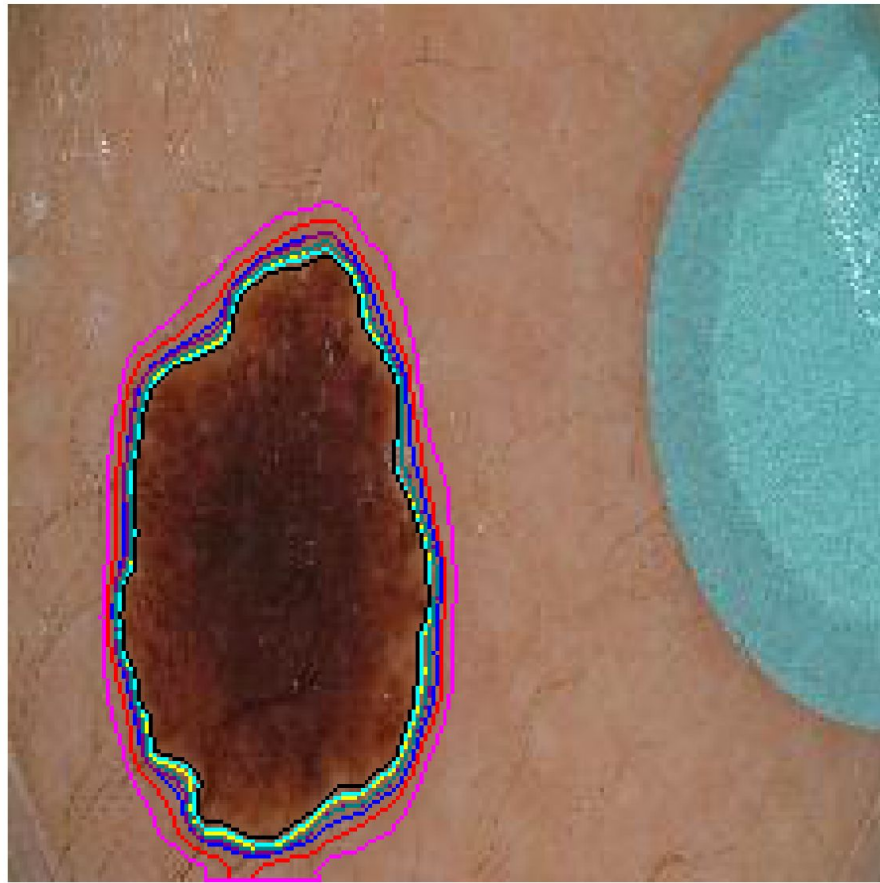
$$\begin{aligned} \mathcal{L}_2 = L_D(Y_{ik}, & (0.20 * \hat{Y}_{i1}) \\ & + (0.05 * \hat{Y}_{i2}) \\ & + (0.75 * \hat{Y}_{i3})) \end{aligned}$$

$$\mathcal{L}_3 = L_{CE} \left(\begin{bmatrix} 0.20, 0.05, 0.75 \\ 0, 0, 1 \end{bmatrix} \right)$$

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

StyleSeg Outputs Adapt to Variability in Lesion Content

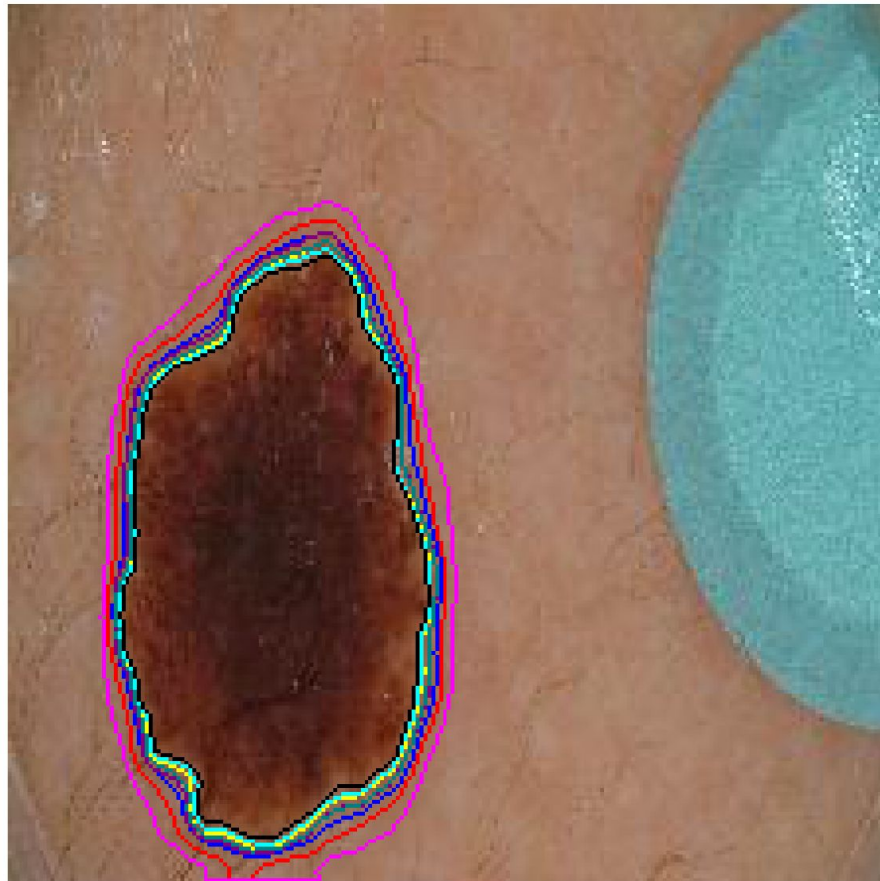
ISIC 0003599



High-contrast
lesion has **high**
agreement across
styles

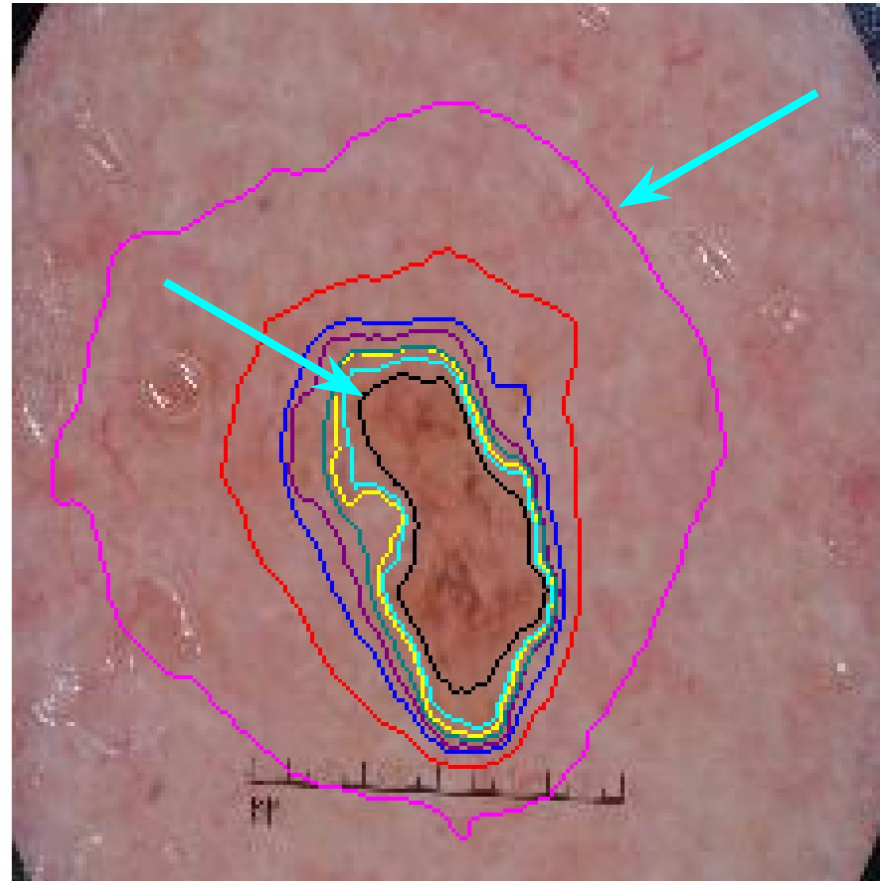
StyleSeg Outputs Adapt to Variability in Lesion Content

ISIC 0003599



High-contrast
lesion has **high**
agreement across
styles

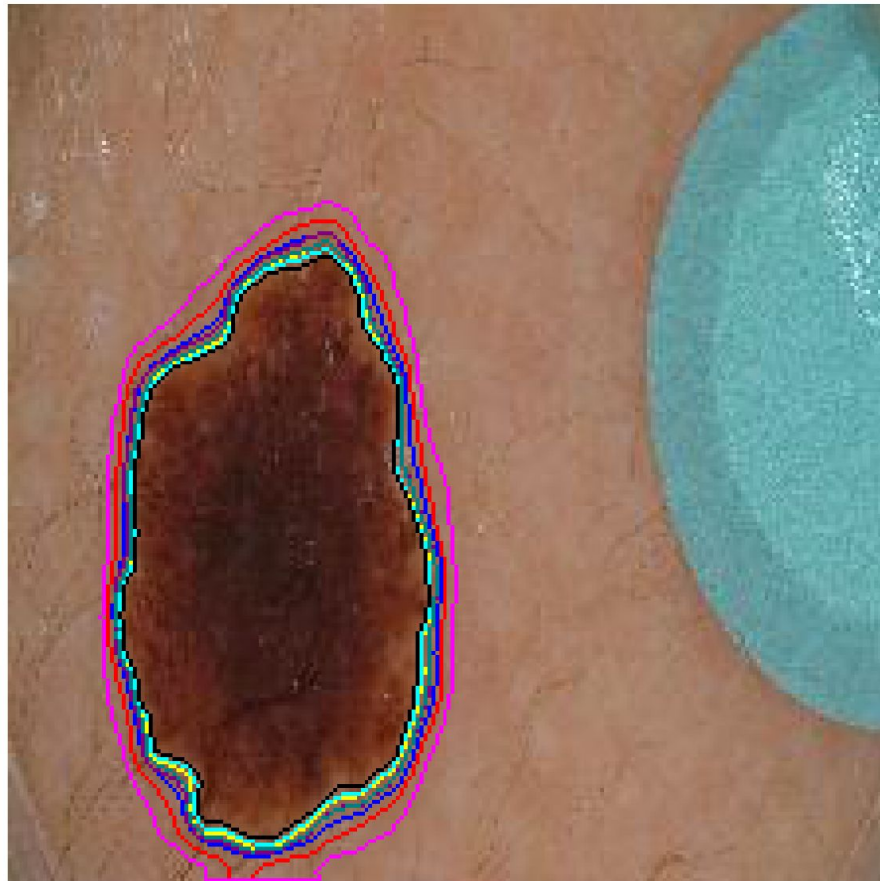
ISIC_0014337



Instances of **under-**
and **over-**
segmentation

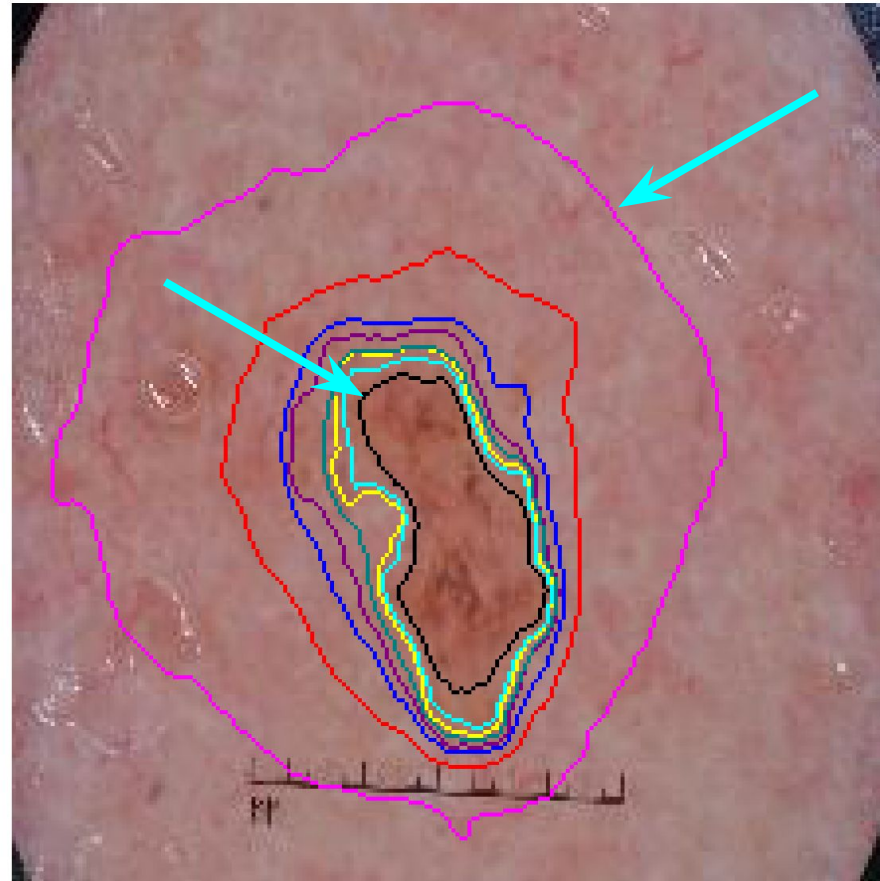
StyleSeg Outputs Adapt to Variability in Lesion Content

ISIC 0003599



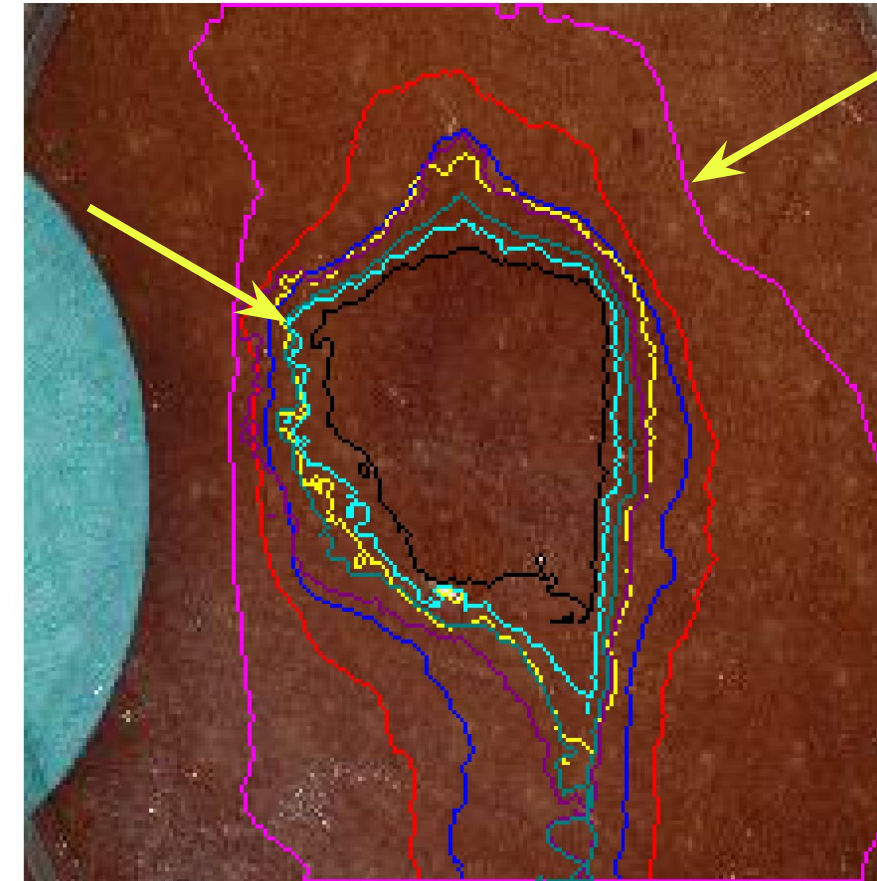
High-contrast lesion has **high agreement** across styles

ISIC_0014337



Instances of **under-** and **over-** segmentation

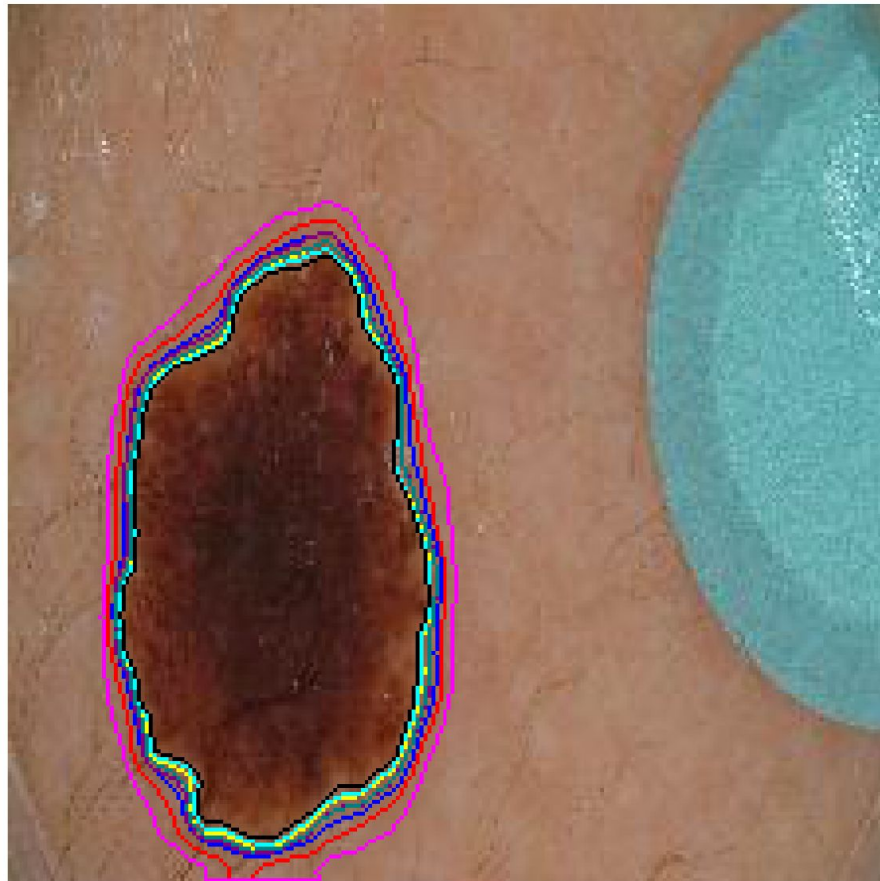
ISIC_0003726



Different **boundary jaggedness** across segmentations

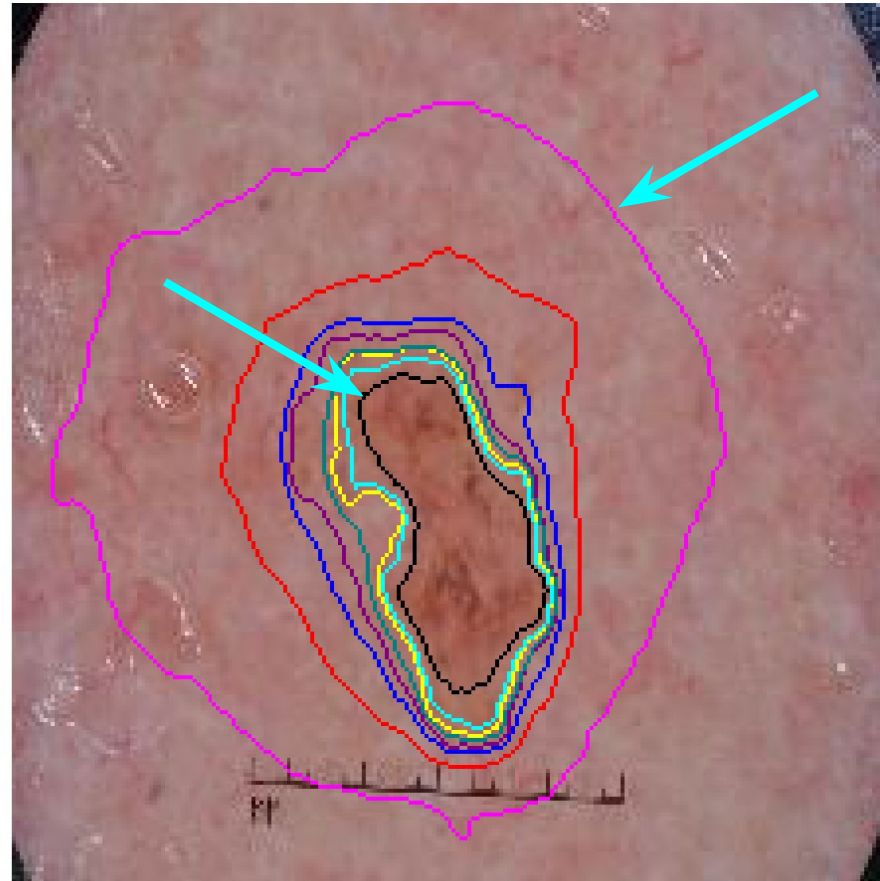
StyleSeg Outputs Adapt to Variability in Lesion Content

ISIC 0003599



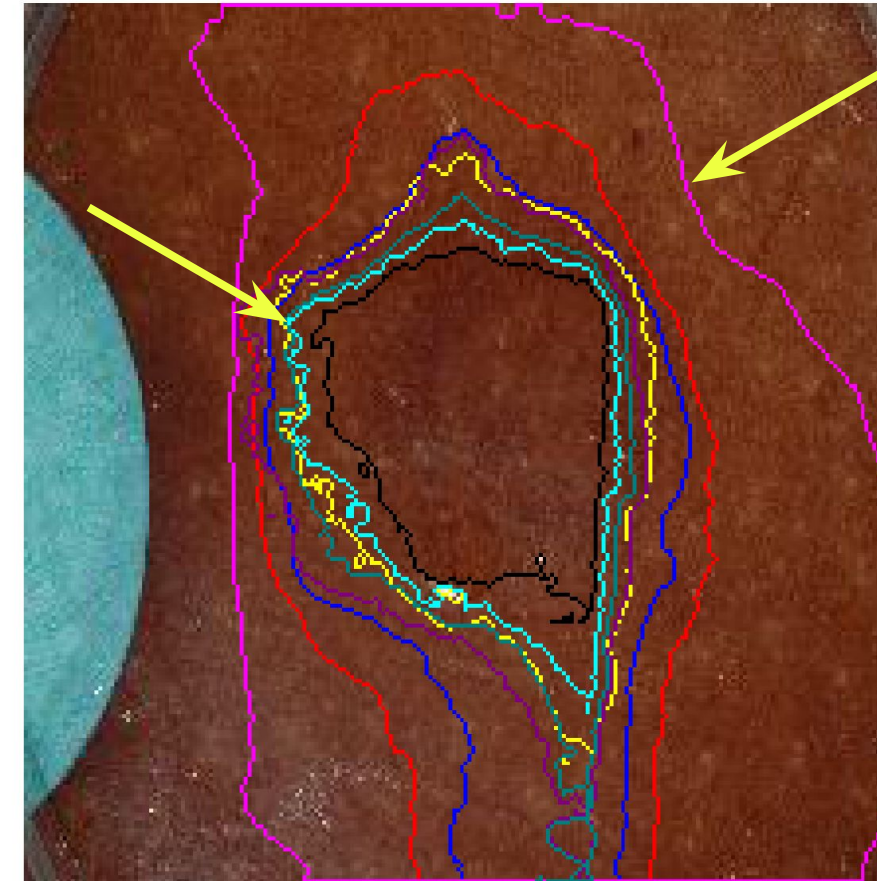
High-contrast lesion has **high agreement** across styles

ISIC_0014337



Instances of **under-** and **over-** segmentation

ISIC_0003726



Different **boundary jaggedness** across segmentations

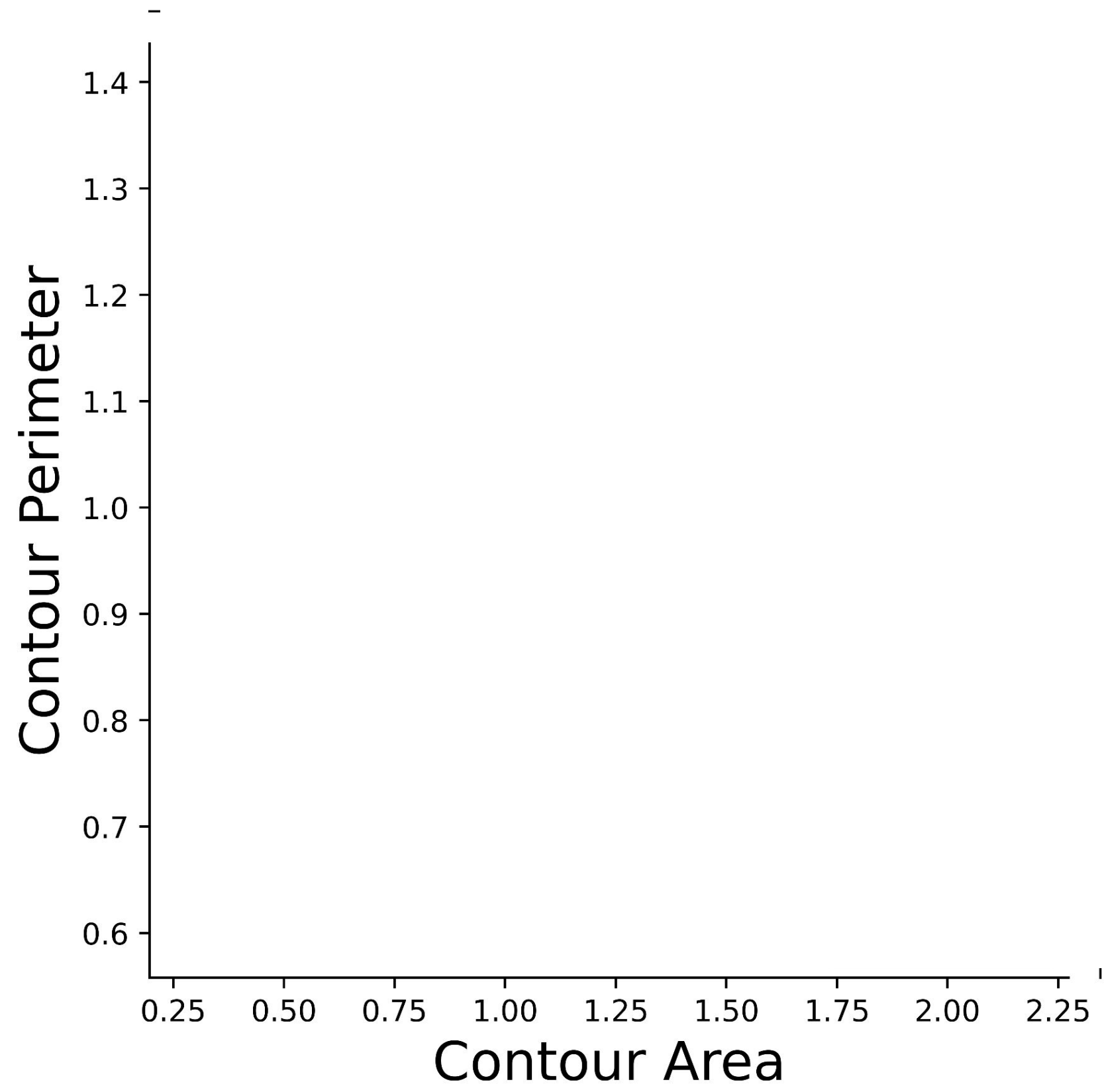
ISIC_0014831



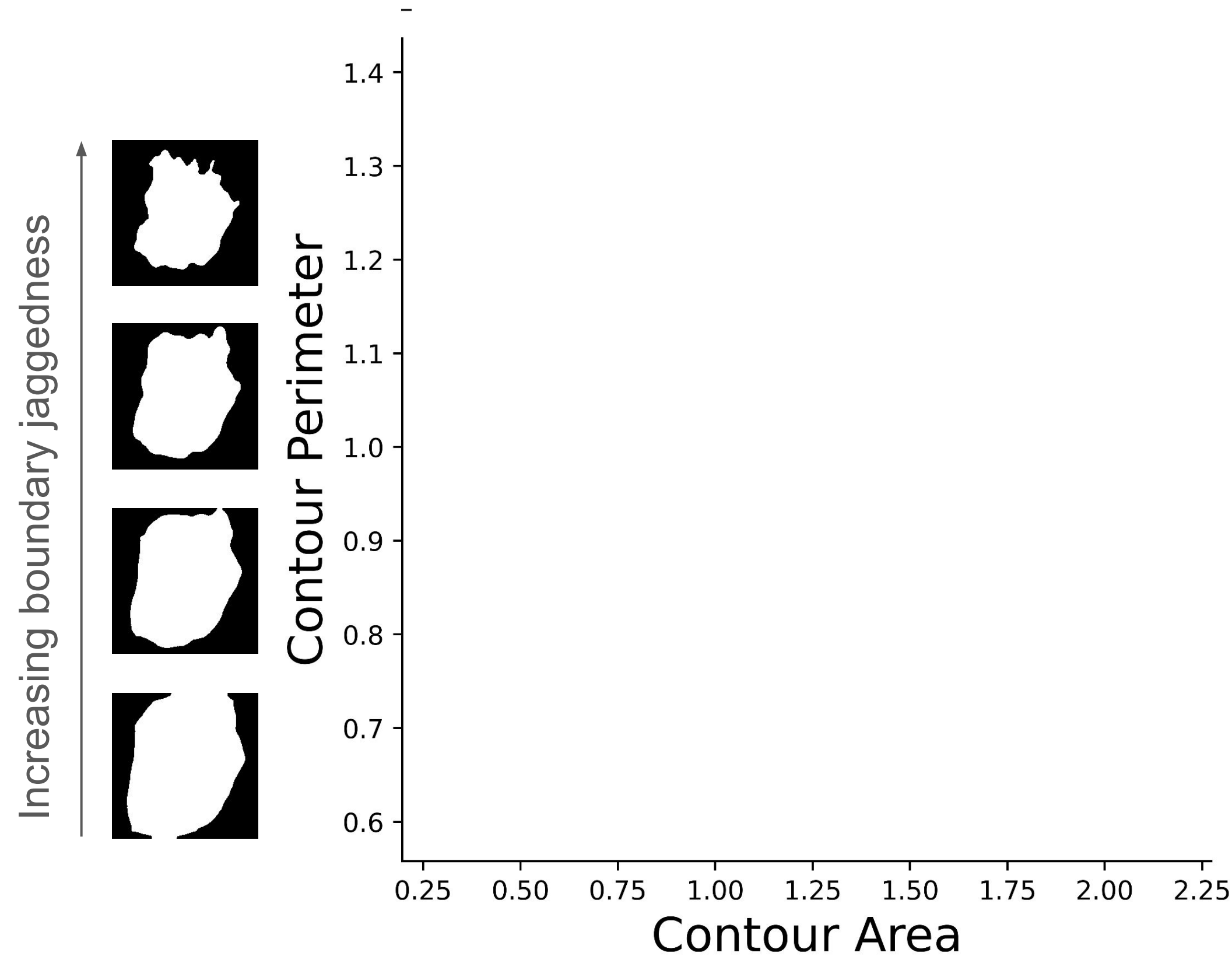
Ambiguous boundary causes segmentation masks to **split**

Semantic Consistency of Styles

Semantic Consistency of Styles



Semantic Consistency of Styles

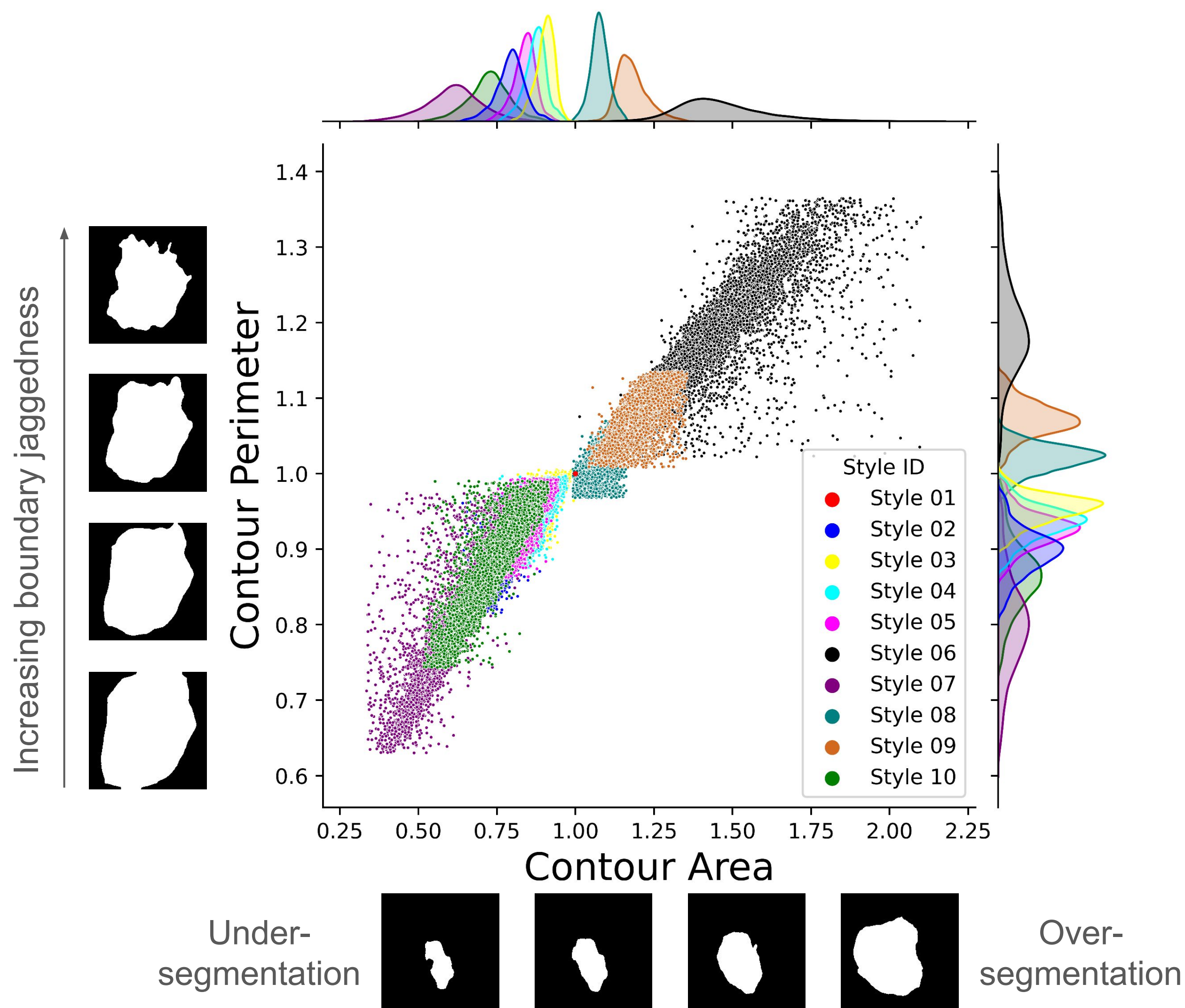


Under-
segmentation



Over-
segmentation

Semantic Consistency of Styles



Competing Methods

SSeg methods	
NaiveTraining	SLS model <u>without any annotator-specific</u> knowledge.
RandAnnotID ^[2]	4 SLS models, one optimized for <u>each annotator randomly assigned</u> to a mask.
LessIsMore ^[3]	SLS model <u>trained on a subset of the masks</u> whose average pairwise Cohen's kappa ≥ 0.5 .
D-LEMA ^[2]	Ensemble of <u>Bayesian</u> SLS models.

Competing Methods

SSeg methods	
NaiveTraining	SLS model <u>without any annotator-specific</u> knowledge.
RandAnnotID ^[2]	4 SLS models, one optimized for <u>each annotator randomly assigned</u> to a mask.
LessIsMore ^[3]	SLS model <u>trained on a subset of the masks</u> whose average pairwise Cohen's kappa ≥ 0.5 .
D-LEMA ^[2]	Ensemble of <u>Bayesian</u> SLS models.
MSeg methods	
MHP ^[4]	Multi-hypothesis prediction model, repurposed for SLS.

Metrics



Metrics

- min.
- max.

Dice $\left(\begin{array}{c} \text{Style 1} \\ \text{[Image 1]}, \text{[Image 2]} \end{array} \right)$,

Dice $\left(\begin{array}{c} \text{Style 2} \\ \text{[Image 1]}, \text{[Image 2]} \end{array} \right)$,

Dice $\left(\begin{array}{c} \text{Style 3} \\ \text{[Image 1]}, \text{[Image 2]} \end{array} \right)$,

Quantitative Results

Method	ISIC Archive-Test ($n = 10,000$)		DermoFit ($n = 1,300$)	
	Min. Dice	Max. Dice	Min. Dice	Max. Dice
NaiveTraining		0.800		0.842
RandAnnotID		–		0.826
LessIsMore		0.815		0.854
D-LEMA		–		0.853
2-MHP	0.727	0.864	0.707	0.882
2-StyleSeg	0.760	0.869	0.759	0.888
3-MHP	0.652	0.876	0.562	0.888
3-StyleSeg	0.713	0.881	0.720	0.897
4-MHP	0.623	0.886	0.636	0.904
4-StyleSeg	0.693	0.889	0.681	0.907
6-MHP	0.121	0.886	0.428	0.900
6-StyleSeg	0.648	0.889	0.651	0.911
8-MHP	0.099	0.896	0.309	0.908
8-StyleSeg	0.595	0.899	0.632	0.910
10-MHP	0.281	0.894	0.181	0.906
10-StyleSeg	0.603	0.899	0.579	0.918

Results on 4 datasets:

- **ISIC Archive-Test ($n = 10000$)**
- **DermoFit ($n = 1300$)**
- **PH² ($n = 200$)**
- **SCD ($n = 206$)**

Learning Multiple Styles Is Always Better

Method	ISIC Archive-Test ($n = 10,000$)		DermoFit ($n = 1,300$)	
	Min. Dice	Max. Dice	Min. Dice	Max. Dice
NaiveTraining		0.800	0.842	
RandAnnotID		–	0.826	
LessIsMore		0.815	0.854	
D-LEMA		–	0.853	
2-MHP	0.727	0.864	0.707	0.882
2-StyleSeg	0.760	0.869	0.759	0.888
3-MHP	0.652	0.876	0.562	0.888
3-StyleSeg	0.713	0.881	0.720	0.897
4-MHP	0.623	0.886	0.636	0.904
4-StyleSeg	0.693	0.889	0.681	0.907
6-MHP	0.121	0.886	0.428	0.900
6-StyleSeg	0.648	0.889	0.651	0.911
8-MHP	0.099	0.896	0.309	0.908
8-StyleSeg	0.595	0.899	0.632	0.910
10-MHP	0.281	0.894	0.181	0.906
10-StyleSeg	0.603	0.899	0.579	0.918

Learning to predict more than 1 style (MSeg methods), even learning to predict 2 styles, consistently outperforms SSeg methods.

Diversity Increases As More Styles are Learned

Method	ISIC Archive-Test ($n = 10,000$)		DermoFit ($n = 1,300$)	
	Min. Dice	Max. Dice	Min. Dice	Max. Dice
NaiveTraining		0.800		0.842
RandAnnotID		–		0.826
LessIsMore		0.815		0.854
D-LEMA		–		0.853
2-MHP	0.727	0.864	0.707	0.882
2-StyleSeg	0.760	0.869	0.759	0.888
3-MHP	0.652	0.876	0.562	0.888
3-StyleSeg	0.713	0.881	0.720	0.897
4-MHP	0.623	0.886	0.636	0.904
4-StyleSeg	0.693	0.889	0.681	0.907
6-MHP	0.121	0.886	0.428	0.900
6-StyleSeg	0.648	0.889	0.651	0.911
8-MHP	0.099	0.896	0.309	0.908
8-StyleSeg	0.595	0.899	0.632	0.910
10-MHP	0.281	0.894	0.181	0.906
10-StyleSeg	0.603	0.899	0.579	0.918

As M increases, a larger number of diverse segmentations are generated, and the max. Dice keeps improving.

StyleSeg Outperforms MHP

Method	ISIC Archive-Test ($n = 10,000$)		DermoFit ($n = 1,300$)	
	Min. Dice	Max. Dice	Min. Dice	Max. Dice
NaiveTraining		0.800		0.842
RandAnnotID		–		0.826
LessIsMore		0.815		0.854
D-LEMA		–		0.853
2-MHP	0.727	→ 0.864	0.707	→ 0.882
2-StyleSeg	0.760	0.869	0.759	0.888
3-MHP	0.652	0.876	0.562	0.888
3-StyleSeg	0.713	0.881	0.720	0.897
4-MHP	0.623	0.886	0.636	0.904
4-StyleSeg	0.693	0.889	0.681	0.907
6-MHP	0.121	0.886	0.428	0.900
6-StyleSeg	0.648	0.889	0.651	0.911
8-MHP	0.099	0.896	0.309	0.908
8-StyleSeg	0.595	0.899	0.632	0.910
10-MHP	0.281	0.894	0.181	0.906
10-StyleSeg	0.603	0.899	0.579	0.918

StyleSeg consistently outperforms MHP for all values of M and for all datasets.

StyleSeg Outputs Are More Plausible

Method	ISIC Archive-Test ($n = 10,000$)		DermoFit ($n = 1,300$)	
	Min. Dice	Max. Dice	Min. Dice	Max. Dice
NaiveTraining		0.800		0.842
RandAnnotID		–		0.826
LessIsMore		0.815		0.854
D-LEMA		–		0.853
2-MHP	0.727	0.864	0.707	0.882
2-StyleSeg	0.760	0.869	0.759	0.888
3-MHP	0.652	0.876	0.562	0.888
3-StyleSeg	0.713	0.881	0.720	0.897
4-MHP	0.623	0.886	0.636	0.904
4-StyleSeg	0.693	0.889	0.681	0.907
6-MHP	0.121	0.886	0.428	0.900
6-StyleSeg	0.648	0.889	0.651	0.911
8-MHP	→ 0.099	0.896	→ 0.309	0.908
8-StyleSeg	0.595	0.899	0.632	0.910
10-MHP	→ 0.281	0.894	→ 0.181	0.906
10-StyleSeg	0.603	0.899	0.579	0.918

StyleSeg consistently outperforms MHP for all values of M and for all datasets.

Moreover, as M increases, all StyleSeg outputs remain reasonably plausible, whereas MHP outputs exhibit diversity at the cost of plausibility.

Performance Improves Even on Single Annot. Datasets

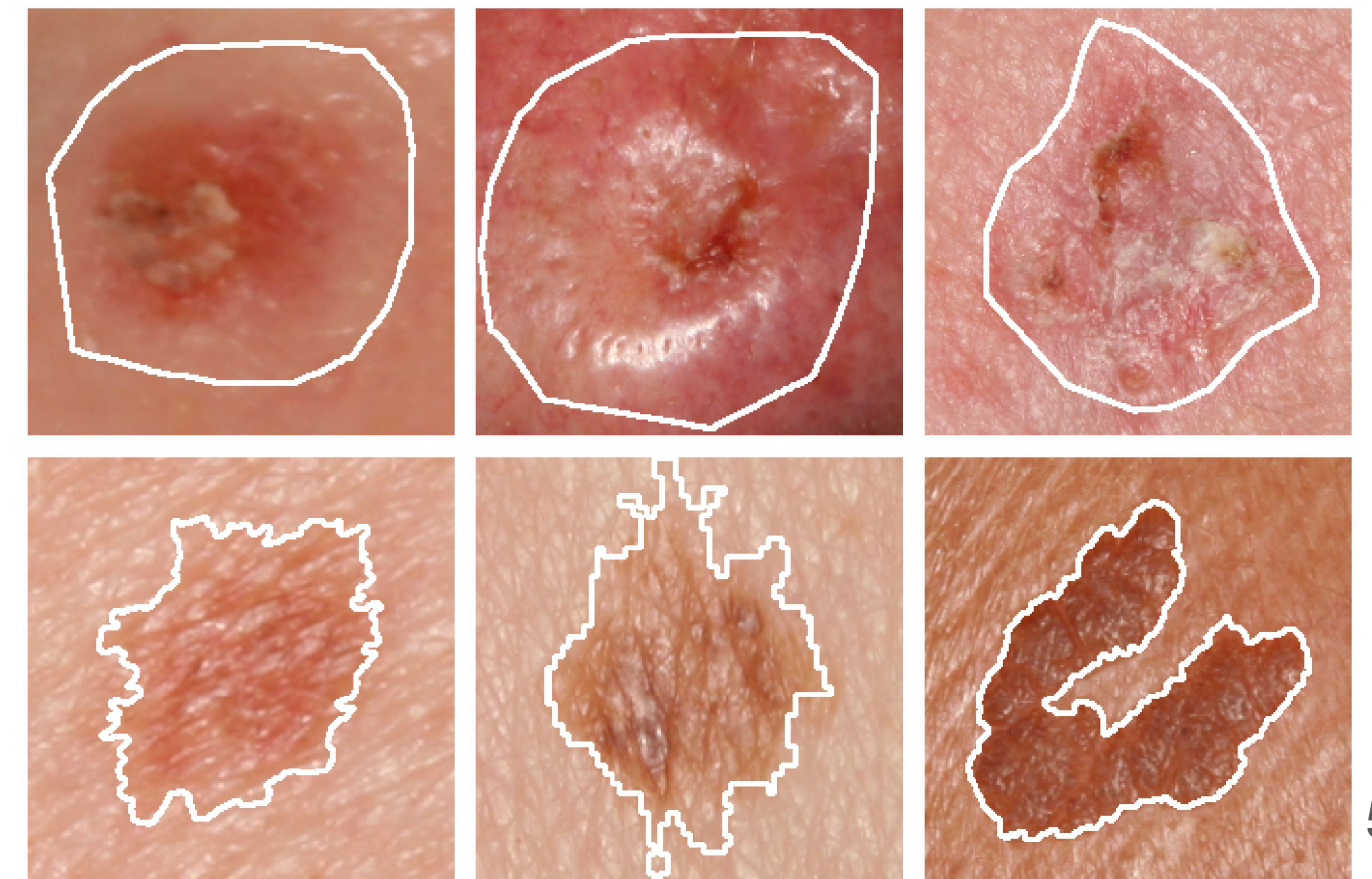
Method	ISIC Archive-Test ($n = 10,000$)		DermoFit ($n = 1,300$)	
	Min. Dice	Max. Dice	Min. Dice	Max. Dice
NaiveTraining		0.800		0.842
RandAnnotID		–		0.826
LessIsMore		0.815		0.854
D-LEMA		–		0.853
2-MHP	0.727	0.864	0.707	0.882
2-StyleSeg	0.760	0.869	0.759	0.888
3-MHP	0.652	0.876	0.562	0.888
3-StyleSeg	0.713	0.881	0.720	0.897
4-MHP	0.623	0.886	0.636	0.904
4-StyleSeg	0.693	0.889	0.681	0.907
6-MHP	0.121	0.886	0.428	0.900
6-StyleSeg	0.648	0.889	0.651	0.911
8-MHP	0.099	0.896	0.309	0.908
8-StyleSeg	0.595	0.899	0.632	0.910
10-MHP	0.281	0.894	0.181	0.906
10-StyleSeg	0.603	0.899	0.579	0.918

Even for datasets without documented variability in segmentations, learning to predict multiple styles is helpful.

Performance Improves Even on Single Annot. Datasets

Method	ISIC Archive-Test ($n = 10,000$)		DermoFit ($n = 1,300$)	
	Min. Dice	Max. Dice	Min. Dice	Max. Dice
NaiveTraining		0.800		0.842
RandAnnotID		–		0.826
LessIsMore		0.815		0.854
D-LEMA		–		0.853
2-MHP	0.727	0.864	0.707	0.882
2-StyleSeg	0.760	0.869	0.759	0.888
3-MHP	0.652	0.876	0.562	0.888
3-StyleSeg	0.713	0.881	0.720	0.897
4-MHP	0.623	0.886	0.636	0.904
4-StyleSeg	0.693	0.889	0.681	0.907
6-MHP	0.121	0.886	0.428	0.900
6-StyleSeg	0.648	0.889	0.651	0.911
8-MHP	0.099	0.896	0.309	0.908
8-StyleSeg	0.595	0.899	0.632	0.910
10-MHP	0.281	0.894	0.181	0.906
10-StyleSeg	0.603	0.899	0.579	0.918

Even for datasets without documented variability in segmentations, learning to predict multiple styles is helpful.



A New Multi-Annotator SLS Dataset: ISIC-MultiAnnot

The **largest** multi-annotator SLS dataset curated from the ISIC Archive.

A New Multi-Annotator SLS Dataset: ISIC-MultiAnnot

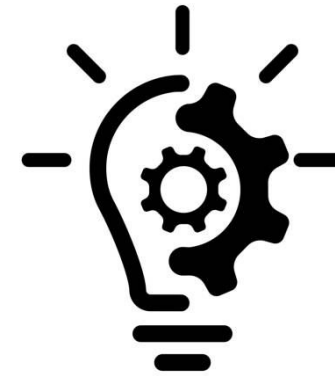
The **largest** multi-annotator SLS dataset curated from the ISIC Archive.



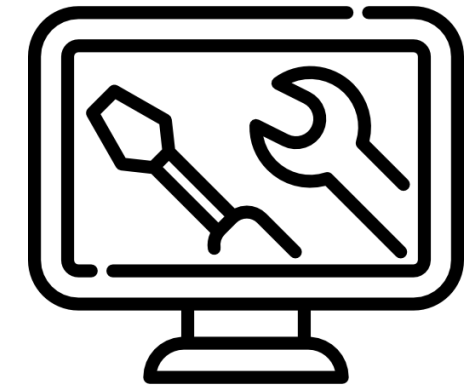
12,951 images



10 anonymized
annotators
“A00” – “A09”



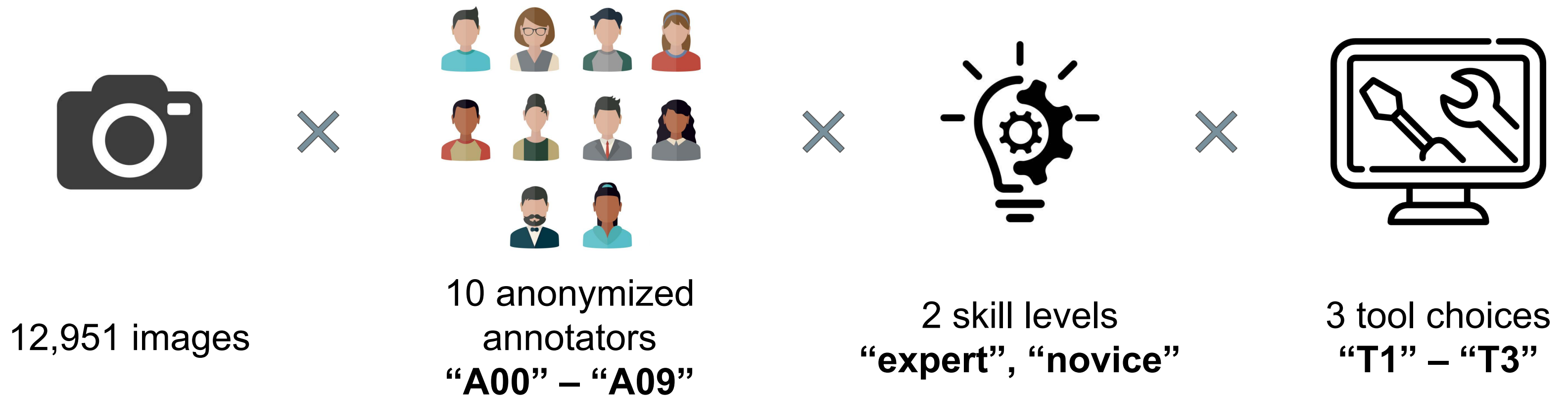
2 skill levels
“expert”, “novice”



3 tool choices
“T1” – “T3”

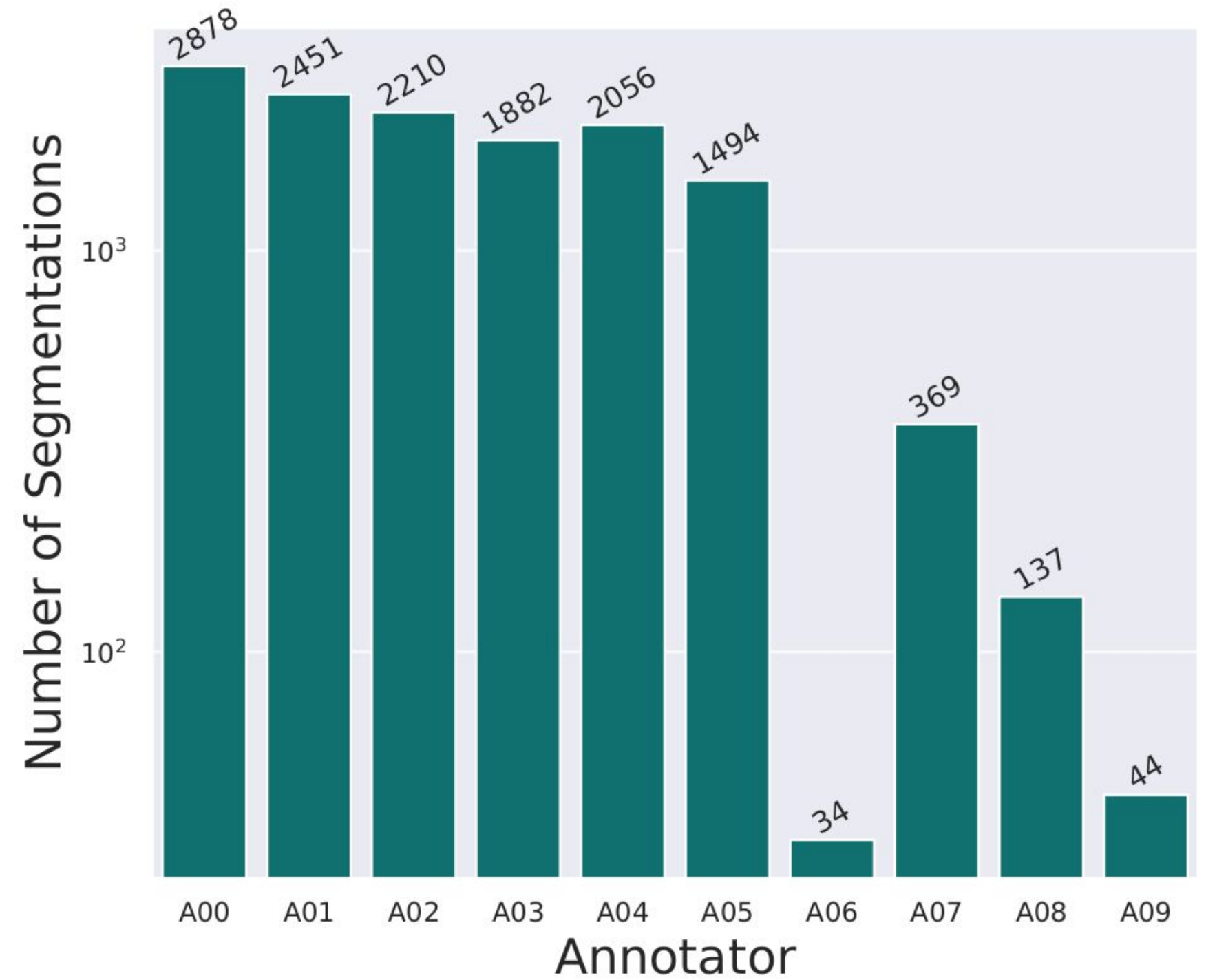
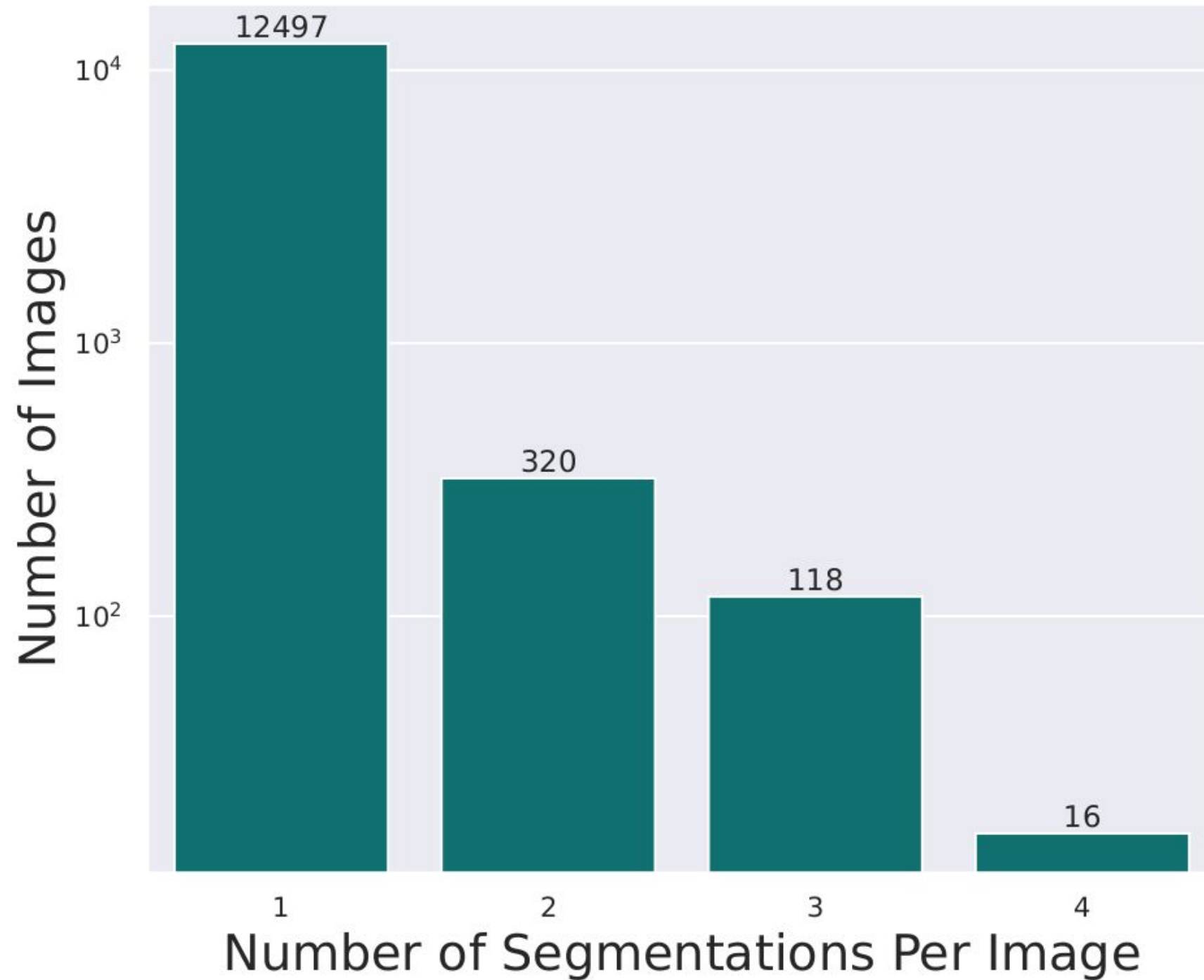
A New Multi-Annotator SLS Dataset: ISIC-MultiAnnot

The **largest** multi-annotator SLS dataset curated from the ISIC Archive.



13,555 image-mask pairs
27 unique annotator preferences

A New Multi-Annotator SLS Dataset: ISIC-MultiAnnot

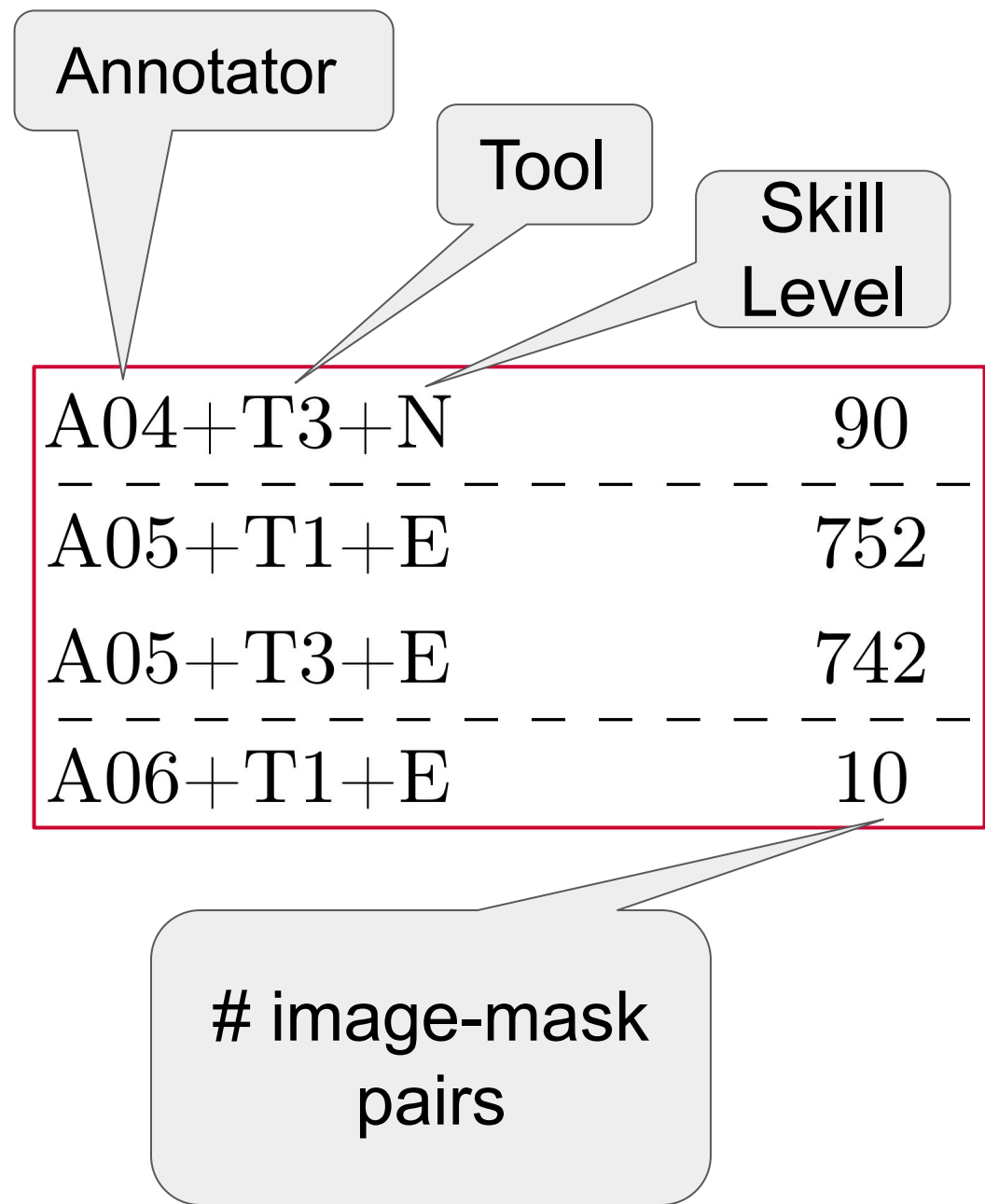


Quantitative Results on ISIC-MultiAnnot

Annotator + Tool + Experience	Seg. Count	1-StyleSeg		2-StyleSeg		3-StyleSeg			4-StyleSeg		
		Dice _{ISSS}	Dice _{ISSS}	Dice _{ASSS}	\mathcal{J}	Dice _{ISSS}	Dice _{ASSS}	\mathcal{J}	Dice _{ISSS}	Dice _{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

Quantitative Results on ISIC-MultiAnnot

Annotator + Tool + Experience	Seg. Count	1-StyleSeg		2-StyleSeg		3-StyleSeg			4-StyleSeg		
		Dice _{ISSS}	Dice _{ISSS}	Dice _{ASSS}	\mathcal{J}	Dice _{ISSS}	Dice _{ASSS}	\mathcal{J}	Dice _{ISSS}	Dice _{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



Quantitative Results on ISIC-MultiAnnot

4-StyleSeg

Annotator + Tool + Experience	Seg. Count	1-StyleSeg		2-StyleSeg			3-StyleSeg			4-StyleSeg		
		Dice _{ISSS}	Dice _{ISSS}	Dice _{ASSS}	\mathcal{J}	Dice _{ISSS}	Dice _{ASSS}	\mathcal{J}	Dice _{ISSS}	Dice _{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	

Annotator
Tool
Skill Level

A04+T3+N	90
A05+T1+E	752
A05+T3+E	742
A06+T1+E	10

image-mask pairs

Annotator / Tool	Seg. Count	1-StyleSeg		2-StyleSeg		3-StyleSeg		4-StyleSeg	
		Diversity	\mathcal{J}	Diversity	\mathcal{J}	Diversity	\mathcal{J}	Diversity	\mathcal{J}
A00-T2-E	1573	0.902	0.923	0.913	2	0.944	0.913	0.944	1
A00-T2-N	1305	0.716	0.793	0.728	2	0.793	0.727	0.793	3
A01-T1-N	6	0.556	0.706	0.706	1	0.714	0.714	0.714	2
A01-T3-E	207	0.908	0.915	0.907	2	0.927	0.908	0.914	3
A01-T3-N	218	0.829	0.857	0.817	1	0.809	0.836	0.856	3
A02-T1-E	1742	0.844	0.888	0.856	1	0.886	0.854	0.855	4
A02-T3-E	408	0.856	0.888	0.883	2	0.899	0.874	0.880	1
A03-T1-E	1022	0.778	0.843	0.827	1	0.854	0.824	0.824	2
A03-T3-E	209	0.801	0.912	0.876	2	0.923	0.868	0.874	3
A04-T1-E	992	0.856	0.888	0.893	1	0.886	0.866	0.866	4
A04-T1-N	61	0.788	0.848	0.823	1	0.837	0.786	0.827	4
A04-T3-E	913	0.912	0.938	0.934	2	0.948	0.926	0.914	3
A04-T3-N	90	0.877	0.918	0.905	2	0.928	0.908	0.913	1
A05-T1-E	752	0.815	0.862	0.817	1	0.873	0.827	0.811	4
A05-T3-E	712	0.878	0.908	0.893	2	0.916	0.876	0.876	1
A06-T1-E	10	0.824	0.902	0.885	1	0.909	0.888	0.880	1
A06-T3-E	24	0.802	0.916	0.916	2	0.914	0.923	0.923	1
A07-T1-E	67	0.828	0.877	0.887	1	0.895	0.862	0.862	4
A07-T1-N	251	0.837	0.892	0.878	1	0.863	0.875	0.855	4
A07-T3-E	12	0.925	0.938	0.937	2	0.939	0.916	0.917	1
A07-T3-N	39	0.863	0.918	0.913	2	0.933	0.899	0.914	1
A08-T1-E	26	0.666	0.758	0.688	2	0.717	0.653	0.666	1
A08-T3-E	111	0.855	0.868	0.826	1	0.877	0.826	0.826	2
A09-T1-E	30	0.815	0.841	0.784	1	0.873	0.833	0.812	4
A09-T1-N	1	0.953	0.927	0.927	2	0.955	0.955	0.947	3
A09-T3-E	10	0.988	0.918	0.918	2	0.933	0.909	0.919	3
A09-T3-N	3	0.914	0.914	0.914	2	0.927	0.927	0.944	1

ISIC-MultiAnnot Results: Key Takeaways

- Improved diversity without compromising quality:** for all $M \geq 2$, choosing a single style that, for each annotator preference, maximizes agreement with the “ground truth” still outperforms 1-StyleSeg.

Annotator / Tool	Seg. Count	1-StyleSeg		2-StyleSeg		3-StyleSeg		4-StyleSeg	
		Diversity	\mathcal{J}	Diversity	\mathcal{J}	Diversity	\mathcal{J}	Diversity	\mathcal{J}
A00-T2-E	1573	0.902	0.923	0.913	2	0.944	0.913	0.944	1
A00-T2-N	1305	0.716	0.793	0.728	2	0.793	0.727	0.793	3
A01-T1-N	6	0.556	0.706	0.706	1	0.741	0.741	0.741	2
A01-T3-E	207	0.908	0.915	0.907	2	0.927	0.908	0.914	3
A01-T3-N	218	0.829	0.857	0.817	1	0.869	0.836	0.856	3
A02-T1-E	1742	0.844	0.888	0.856	1	0.886	0.854	0.855	4
A02-T3-E	408	0.856	0.888	0.883	2	0.899	0.874	0.880	1
A03-T1-E	1022	0.758	0.843	0.827	1	0.854	0.824	0.824	2
A03-T3-E	209	0.801	0.912	0.876	2	0.923	0.868	0.874	3
A04-T1-E	992	0.850	0.888	0.893	1	0.888	0.866	0.866	4
A04-T1-N	61	0.780	0.848	0.823	1	0.837	0.786	0.827	4
A04-T3-E	913	0.912	0.938	0.944	2	0.948	0.926	0.914	3
A04-T3-N	90	0.877	0.910	0.905	2	0.928	0.908	0.926	1
A05-T1-E	752	0.815	0.862	0.817	1	0.873	0.827	0.841	4
A05-T3-E	712	0.878	0.908	0.893	2	0.916	0.876	0.891	1
A06-T1-E	10	0.824	0.902	0.885	1	0.909	0.888	0.908	1
A06-T3-E	24	0.802	0.910	0.916	2	0.914	0.923	0.933	1
A07-T1-E	67	0.820	0.877	0.887	1	0.895	0.862	0.862	4
A07-T1-N	251	0.837	0.892	0.878	1	0.863	0.875	0.895	4
A07-T3-E	12	0.925	0.938	0.937	2	0.939	0.916	0.947	1
A07-T3-N	39	0.863	0.918	0.913	2	0.933	0.899	0.914	1
A08-T1-E	26	0.666	0.758	0.680	2	0.747	0.653	0.666	1
A08-T3-E	111	0.655	0.668	0.626	1	0.677	0.626	0.668	2
A09-T1-E	30	0.815	0.841	0.784	1	0.873	0.833	0.842	4
A09-T1-N	1	0.953	0.927	0.927	2	0.955	0.955	0.947	3
A09-T3-E	10	0.988	0.918	0.918	2	0.933	0.909	0.919	3
A09-T3-N	3	0.914	0.914	0.914	2	0.927	0.927	0.944	1

ISIC-MultiAnnot Results: Key Takeaways

1. **Improved diversity without compromising quality:** for all $M \geq 2$, choosing a single style that, for each annotator preference, maximizes agreement with the “ground truth” still outperforms 1-StyleSeg.

Personalization in segmentation: each user can choose their own style.

Annotator / Tool	Seg. Count	1-StyleSeg		2-StyleSeg		3-StyleSeg		4-StyleSeg	
		Diversity	\mathcal{J}	Diversity	\mathcal{J}	Diversity	\mathcal{J}	Diversity	\mathcal{J}
A00-T2-E	1573	0.902	0.923	0.913	2	0.944	0.913	3	0.944
A00-T2-N	1305	0.716	0.793	0.728	2	0.793	0.727	3	0.793
A01-T1-N	6	0.556	0.706	0.706	1	0.741	0.741	2	0.828
A01-T3-E	207	0.908	0.915	0.907	2	0.927	0.908	1	0.931
A01-T3-N	218	0.829	0.857	0.817	1	0.809	0.836	1	0.876
A02-T1-E	1742	0.844	0.888	0.856	1	0.886	0.854	1	0.895
A02-T3-E	408	0.856	0.888	0.883	2	0.899	0.874	3	0.903
A03-T1-E	1022	0.778	0.843	0.827	1	0.854	0.842	2	0.881
A03-T3-E	209	0.801	0.912	0.876	2	0.923	0.868	1	0.923
A04-T1-E	992	0.856	0.888	0.893	1	0.886	0.866	2	0.906
A04-T1-N	61	0.788	0.848	0.823	1	0.837	0.786	1	0.827
A04-T3-E	913	0.912	0.938	0.934	2	0.948	0.926	1	0.951
A04-T3-N	90	0.877	0.918	0.905	2	0.928	0.908	3	0.928
A05-T1-E	752	0.815	0.862	0.817	1	0.873	0.827	1	0.892
A05-T3-E	712	0.878	0.908	0.893	2	0.916	0.876	1	0.928
A06-T1-E	10	0.824	0.902	0.885	1	0.909	0.888	2	0.908
A06-T3-E	24	0.802	0.916	0.916	2	0.916	0.923	3	0.933
A07-T1-E	67	0.826	0.877	0.887	1	0.898	0.862	2	0.897
A07-T1-N	251	0.837	0.892	0.878	1	0.883	0.875	2	0.895
A07-T3-E	12	0.925	0.938	0.937	2	0.939	0.916	1	0.947
A07-T3-N	39	0.863	0.918	0.913	2	0.933	0.899	3	0.934
A08-T1-E	26	0.666	0.758	0.688	2	0.717	0.653	1	0.703
A08-T3-E	111	0.655	0.668	0.626	1	0.677	0.626	2	0.725
A09-T1-E	30	0.815	0.841	0.784	1	0.873	0.833	2	0.884
A09-T1-N	1	0.953	0.927	0.927	2	0.955	0.955	1	0.947
A09-T3-E	10	0.988	0.918	0.918	2	0.933	0.909	1	0.937
A09-T3-N	3	0.914	0.914	0.914	2	0.927	0.927	3	0.944

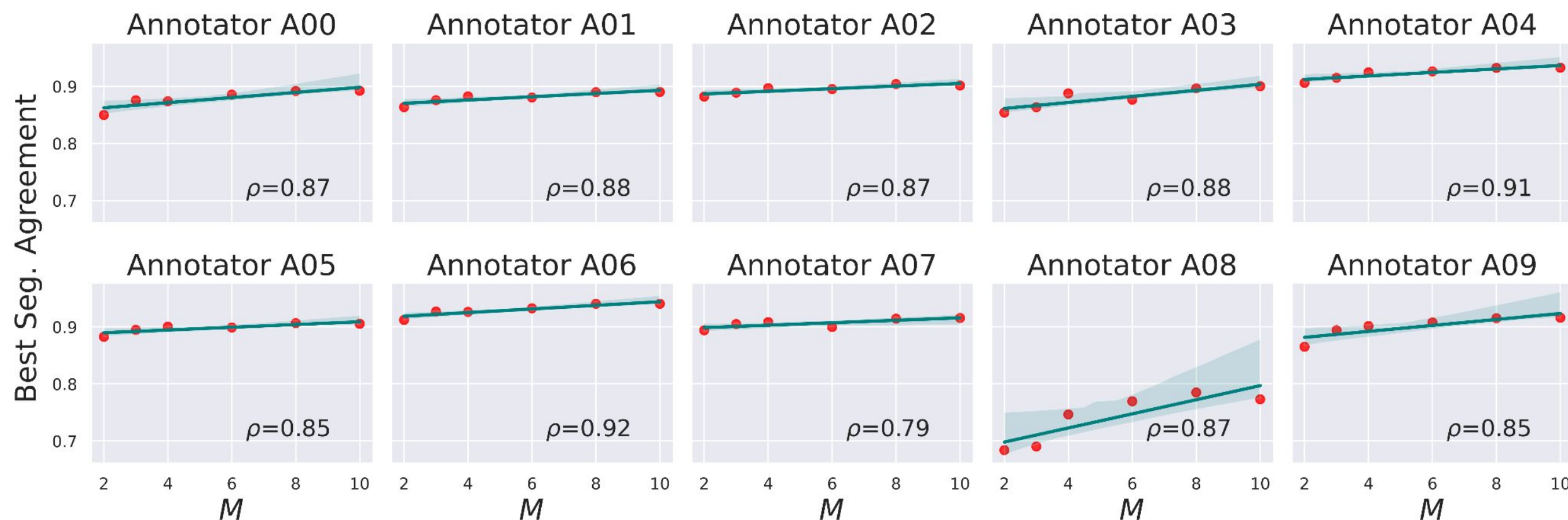
ISIC-MultiAnnot Results: Key Takeaways

- Improved diversity without compromising quality:** for all $M \geq 2$, choosing a single style that, for each annotator preference, maximizes agreement with the “ground truth” still outperforms 1-StyleSeg.
- Performance improves as M increases.**

Annotator / Tool	Seg. Count	1-StyleSeg		2-StyleSeg		3-StyleSeg		4-StyleSeg	
		Diversity	\mathcal{J}	Diversity	\mathcal{J}	Diversity	\mathcal{J}	Diversity	\mathcal{J}
A00-T2-E	1373	0.902	0.923	0.913	2	0.944	0.913	0.944	1
A00-T2-N	1305	0.716	0.751	0.728	2	0.753	0.727	0.750	3
A01-T1-N	6	0.556	0.706	0.706	1	0.741	0.741	0.828	2
A01-T3-E	207	0.908	0.915	0.907	2	0.927	0.900	0.931	3
A01-T3-N	218	0.829	0.857	0.817	1	0.809	0.836	0.856	3
A02-T1-E	1742	0.844	0.888	0.856	1	0.886	0.854	0.895	4
A02-T3-E	408	0.656	0.888	0.833	2	0.809	0.874	0.903	1
A03-T1-E	1022	0.758	0.843	0.827	1	0.854	0.824	0.881	2
A03-T3-E	209	0.801	0.912	0.876	2	0.923	0.868	0.923	3
A04-T1-E	992	0.856	0.888	0.893	1	0.888	0.866	0.906	4
A04-T1-N	61	0.780	0.848	0.823	1	0.837	0.786	0.827	4
A04-T3-E	913	0.912	0.933	0.934	2	0.948	0.926	0.951	3
A04-T3-N	90	0.877	0.910	0.905	2	0.928	0.908	0.926	1
A05-T1-E	752	0.815	0.862	0.817	1	0.873	0.827	0.882	4
A05-T3-E	712	0.875	0.903	0.893	2	0.916	0.876	0.928	1
A06-T1-E	10	0.824	0.902	0.885	1	0.909	0.888	0.909	1
A06-T3-E	24	0.802	0.916	0.916	2	0.934	0.923	0.933	1
A07-T1-E	67	0.826	0.877	0.887	1	0.903	0.862	0.897	4
A07-T1-N	251	0.837	0.902	0.878	1	0.903	0.875	0.905	4
A07-T3-E	12	0.925	0.938	0.937	2	0.939	0.916	0.947	1
A07-T3-N	39	0.863	0.918	0.913	2	0.933	0.899	0.934	1
A08-T1-E	26	0.666	0.750	0.680	2	0.717	0.653	0.703	1
A08-T3-E	111	0.655	0.668	0.626	1	0.677	0.626	0.725	2
A09-T1-E	30	0.815	0.841	0.784	1	0.873	0.833	0.884	4
A09-T1-N	1	0.953	0.927	0.927	2	0.955	0.955	0.947	3
A09-T3-E	10	0.980	0.918	0.918	2	0.933	0.909	0.937	3
A09-T3-N	3	0.934	0.913	0.913	2	0.927	0.927	0.944	1

ISIC-MultiAnnot Results: Key Takeaways

- Improved diversity without compromising quality:** for all $M \geq 2$, choosing a single style that, for each annotator preference, maximizes agreement with the “ground truth” still outperforms 1-StyleSeg.
- Performance improves as M increases.**



Annotator / Tool	Seg. Count	1-StyleSeg		2-StyleSeg		3-StyleSeg		4-StyleSeg	
		Diversity	\mathcal{J}	Diversity	\mathcal{J}	Diversity	\mathcal{J}	Diversity	\mathcal{J}
A00-T2-E	1573	0.902	0.923	0.913	0.944	0.913	0.944	0.914	0.944
A00-T2-N	1305	0.716	0.751	0.728	0.753	0.727	0.727	0.750	0.728
A01-T1-N	6	0.556	0.706	0.706	0.741	0.741	0.741	0.828	0.767
A01-T3-E	207	0.908	0.915	0.907	0.927	0.908	0.908	0.931	0.934
A01-T3-N	218	0.829	0.857	0.817	0.809	0.836	0.836	0.876	0.836
A02-T1-E	1742	0.844	0.888	0.856	0.886	0.854	0.854	0.895	0.856
A02-T3-E	408	0.656	0.888	0.882	0.899	0.874	0.874	0.903	0.880
A03-T1-E	1022	0.758	0.843	0.827	0.854	0.842	0.842	0.881	0.824
A03-T3-E	209	0.801	0.912	0.876	0.923	0.868	0.868	0.922	0.874
A04-T1-E	992	0.856	0.888	0.893	0.886	0.866	0.866	0.906	0.856
A04-T1-N	61	0.788	0.848	0.823	0.837	0.786	0.786	0.827	0.788
A04-T3-E	913	0.912	0.938	0.934	0.948	0.926	0.926	0.951	0.932
A04-T3-N	90	0.877	0.918	0.905	0.928	0.908	0.908	0.926	0.913
A05-T1-E	752	0.815	0.862	0.817	0.873	0.827	0.827	0.882	0.817
A05-T3-E	712	0.875	0.908	0.893	0.916	0.876	0.876	0.928	0.854
A06-T1-E	10	0.824	0.902	0.885	0.909	0.888	0.888	0.908	0.880
A06-T3-E	24	0.802	0.916	0.916	0.934	0.923	0.923	0.933	0.929
A07-T1-E	67	0.826	0.877	0.887	0.895	0.862	0.862	0.897	0.862
A07-T1-N	251	0.837	0.892	0.878	0.893	0.875	0.875	0.905	0.878
A07-T3-E	12	0.925	0.938	0.937	0.939	0.916	0.916	0.947	0.932
A07-T3-N	39	0.863	0.918	0.913	0.933	0.899	0.899	0.934	0.914
A08-T1-E	26	0.666	0.758	0.688	0.747	0.653	0.653	0.703	0.666
A08-T3-E	111	0.655	0.668	0.626	0.677	0.626	0.626	0.725	0.668
A09-T1-E	30	0.815	0.841	0.784	0.873	0.833	0.833	0.884	0.812
A09-T1-N	1	0.953	0.927	0.927	0.955	0.955	0.955	0.947	0.947
A09-T3-E	10	0.988	0.918	0.918	0.933	0.909	0.909	0.937	0.919
A09-T3-N	3	0.934	0.914	0.914	0.927	0.927	0.927	0.944	0.944

ISIC-MultiAnnot Results: Key Takeaways

- Improved diversity without compromising quality:** for all $M \geq 2$, choosing a single style that, for each annotator preference, maximizes agreement with the “ground truth” still outperforms 1-StyleSeg.
- Performance improves as M increases.**
- Ability to learn tool-specific latent factors:** Without specifically training for it, a 3-StyleSeg model is able to choose a unique style for each of the three tools (“T1”, “T2”, “T3”).

Quantifying Annotator-Style Alignment: A New Measure

If we model 3 styles, the best style can be the one that

- best matches 100% of images (**perfect alignment**), or
- best matches, say, 34% of images (**weak alignment**).

Quantifying Annotator-Style Alignment: A New Measure

If we model 3 styles, the best style can be the one that

- best matches 100% of images (**perfect alignment**), or
- best matches, say, 34% of images (**weak alignment**).

How do we quantify this annotator-style alignment strength?

Quantifying Annotator-Style Alignment: A New Measure

If we model 3 styles, the best style can be the one that

- best matches 100% of images (**perfect alignment**), or
- best matches, say, 34% of images (**weak alignment**).

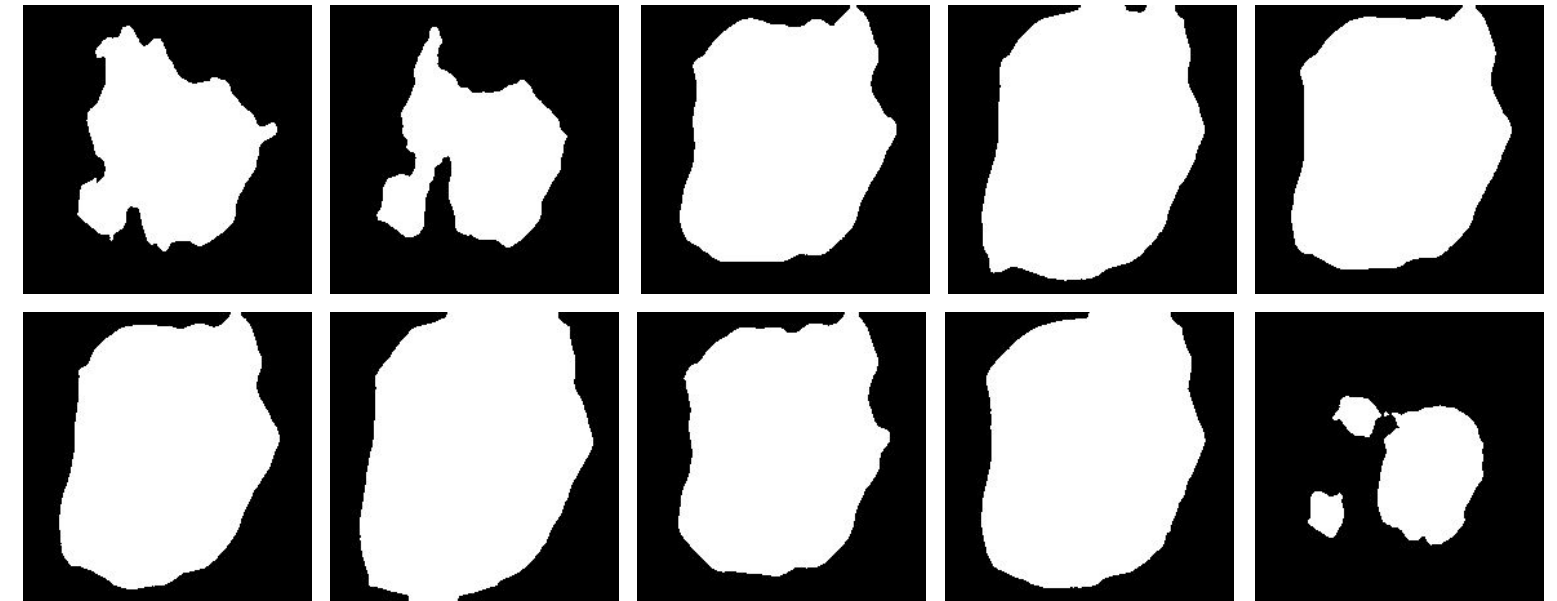
How do we quantify this annotator-style alignment strength?

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

Quantifying Annotator-Style Alignment: A New Measure

If we model 3 styles, the best style can be the one that

- best matches 100% of images (**perfect alignment**), or
- best matches, say, 34% of images (**weak alignment**).



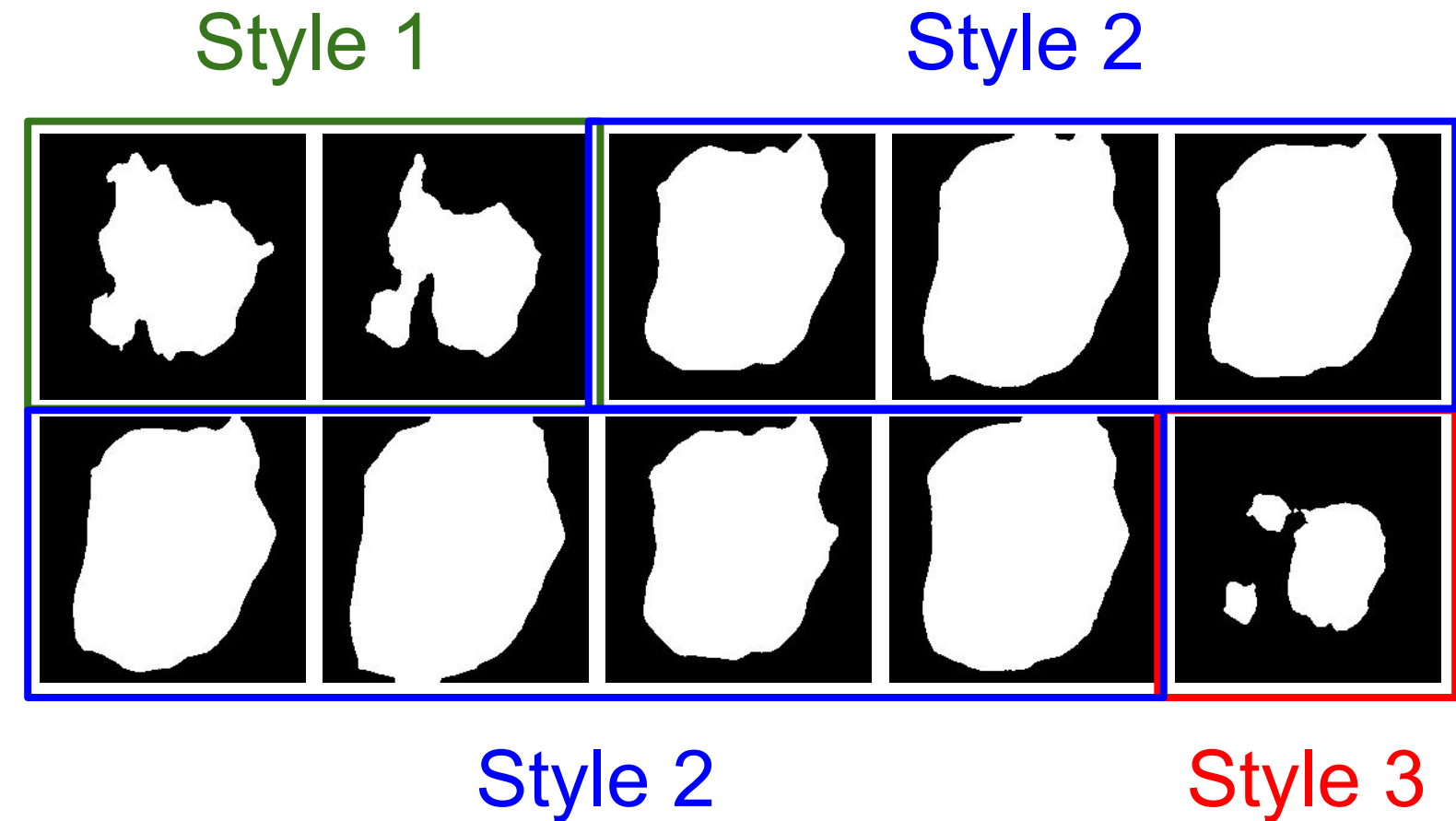
How do we quantify this annotator-style alignment strength?

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

Quantifying Annotator-Style Alignment: A New Measure

If we model 3 styles, the best style can be the one that

- best matches 100% of images (**perfect alignment**), or
- best matches, say, 34% of images (**weak alignment**).



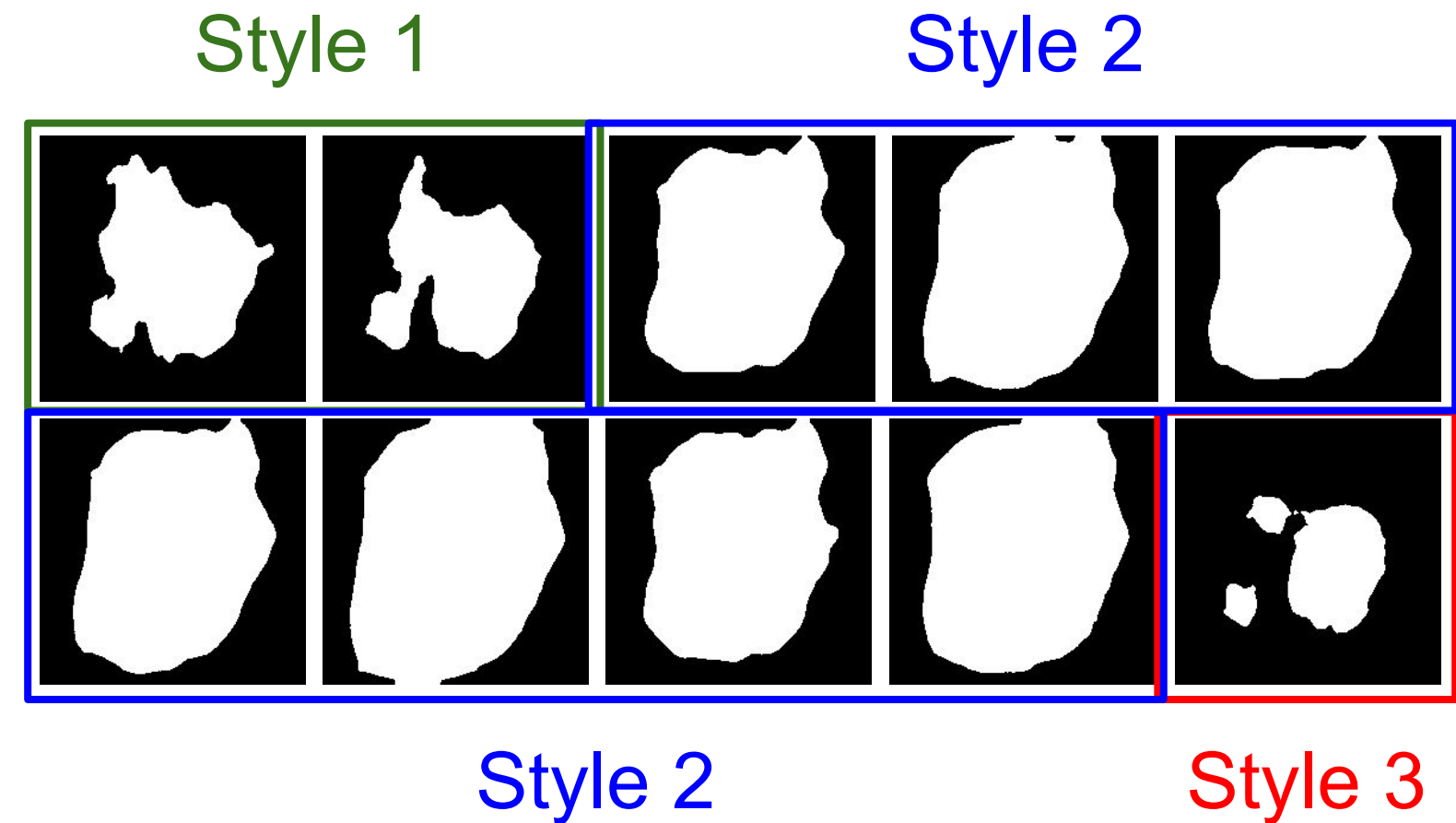
How do we quantify this annotator-style alignment strength?

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

Quantifying Annotator-Style Alignment: A New Measure

If we model 3 styles, the best style can be the one that

- best matches 100% of images (**perfect alignment**), or
- best matches, say, 34% of images (**weak alignment**).



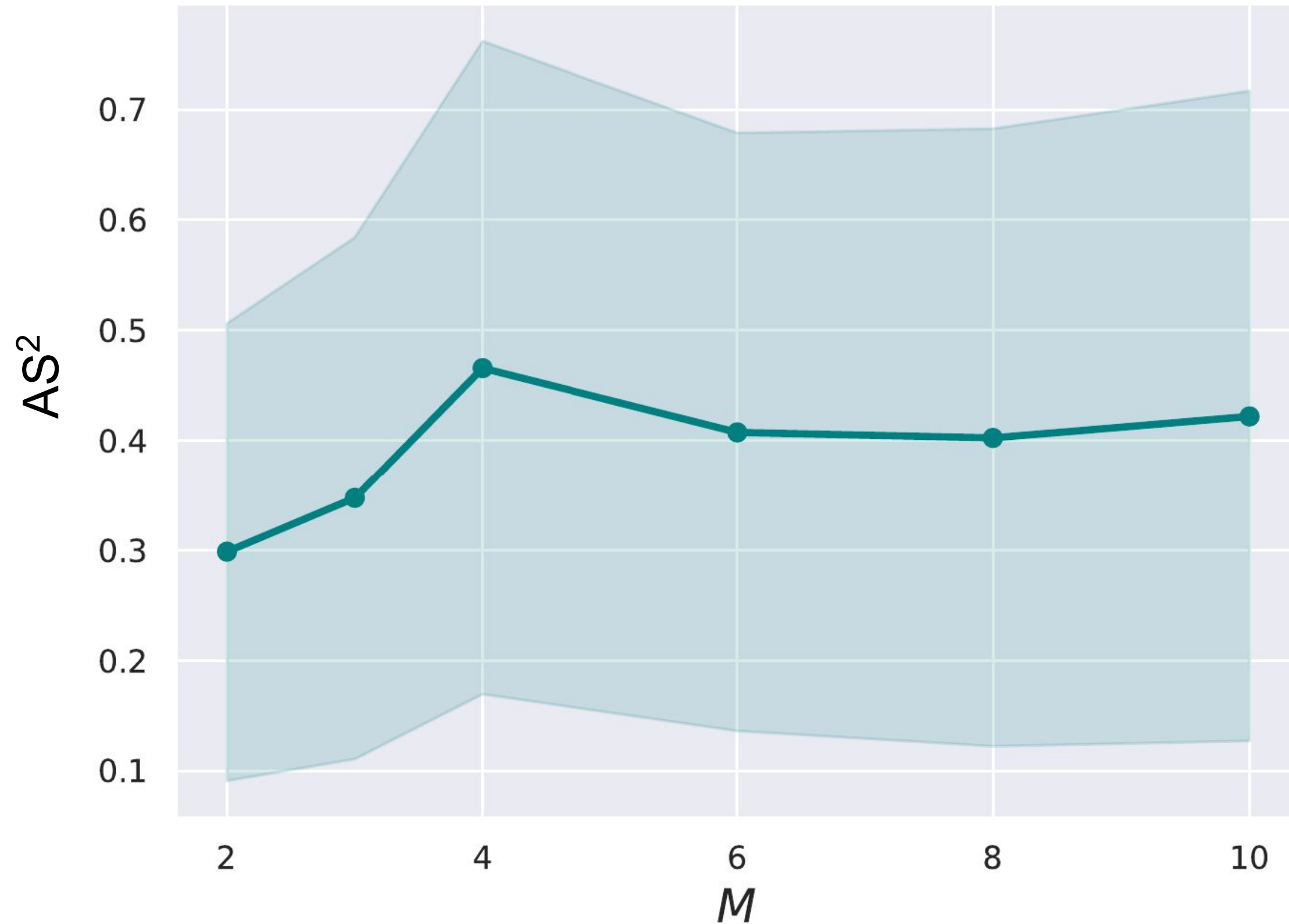
How do we quantify this annotator-style alignment strength?

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

$$q_1 = 0.2, q_2 = 0.7, q_3 = 0.1$$

$$q = [0.2, 0.7, 0.1] \Rightarrow AS^2 = 0.27.$$

Quantifying Annotator-Style Alignment



Modeling **more styles** captures **more diversity** and is not detrimental to segmentation quality.

Conclusion

- Formulated the **problem of segmentation style discovery**, and showed that StyleSeg discovers styles that are **plausible, diverse, and semantically consistent**.

Conclusion

- Formulated the **problem of segmentation style discovery**, and showed that StyleSeg discovers styles that are **plausible, diverse, and semantically consistent**.
- The **largest multi-annotator SLS dataset** (> 13.5k image-mask pairs) with **annotator correspondence** curated from the ISIC Archive.

Conclusion

- Formulated the **problem of segmentation style discovery**, and showed that StyleSeg discovers styles that are **plausible, diverse, and semantically consistent**.
- The **largest multi-annotator SLS dataset** (> 13.5k image-mask pairs) with **annotator correspondence** curated from the ISIC Archive.
- **A new measure for quantifying the strength of alignment** between annotators' preferences and styles.

Conclusion

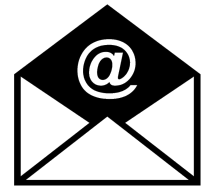
- Formulated the **problem of segmentation style discovery**, and showed that StyleSeg discovers styles that are **plausible, diverse, and semantically consistent**.
- The **largest multi-annotator SLS dataset** (> 13.5k image-mask pairs) with **annotator correspondence** curated from the ISIC Archive.
- **A new measure for quantifying the strength of alignment** between annotators' preferences and styles.
- **Future work** may look at approaches to **finding the optimal number of styles** in a segmentation dataset.

References

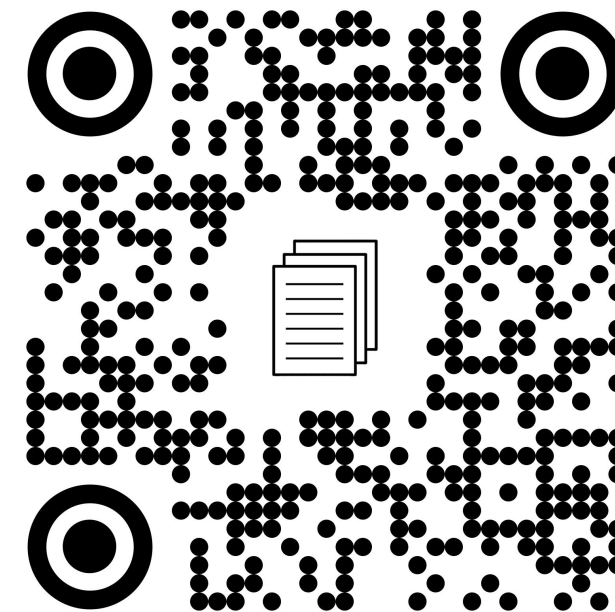
- [1] Silletti et al., “Variability in human and automatic segmentation of melanocytic lesions”, *EMBC*, 2009.
- [2] Mirikharaji et al., “D-LEMA: Deep learning ensembles from multiple annotations-application to skin lesion segmentation”, *CVPR ISIC 2021*.
- [3] Ribeiro et al., “Less is more: Sample selection and label conditioning improve skin lesion segmentation”, *CVPR ISIC 2020*.
- [4] Rupprecht et al., “Learning in an uncertain world: Representing ambiguity through multiple hypotheses”, *ICCV 2017*.

Thank you.

Questions?



kabhishe@sfu.ca



Acknowledgements



Digital Research
Alliance of Canada

Alliance de recherche
numérique du Canada



NVIDIA