

# Ninth ISIC Skin Image Analysis Workshop @ MICCAI 2024

## Segmentation Style Discovery: Application to Skin Lesion Images



Kumar Abhishek<sup>†</sup>



Jeremy Kawahara<sup>‡</sup>



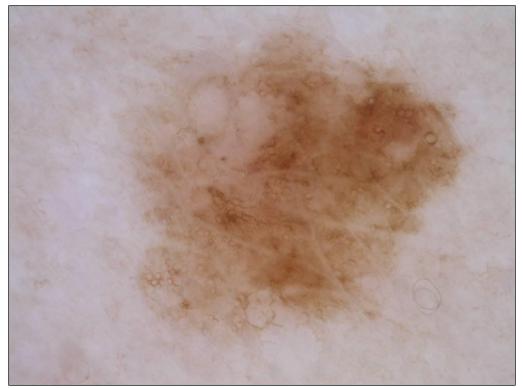
Ghassan Hamarneh<sup>†</sup>



SIMON FRASER  
UNIVERSITY



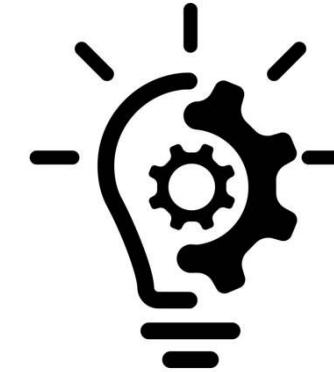
# Variability in Medical Image Segmentation



Ambiguous object  
boundaries



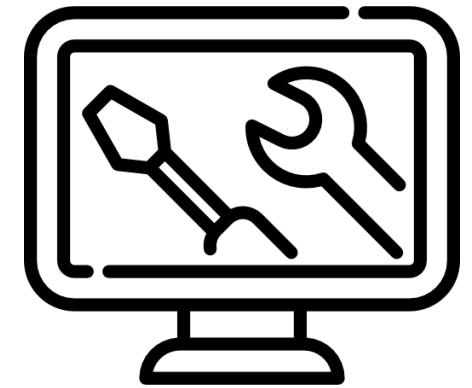
Annotators' personal  
preferences



Annotators'  
skill levels

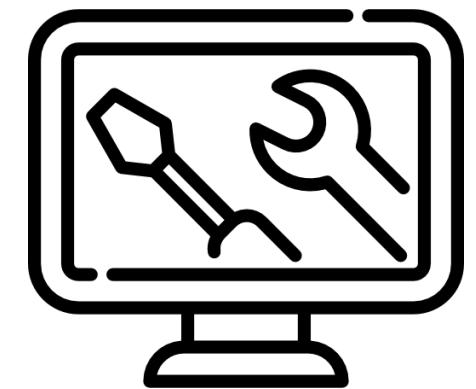
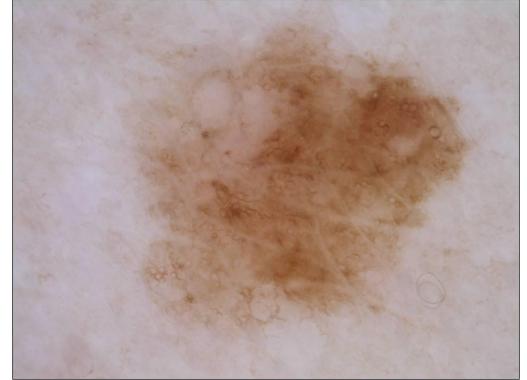


Segmentation  
criteria



Segmentation  
tools

# Variability in Medical Image Segmentation



Ambiguous object  
boundaries

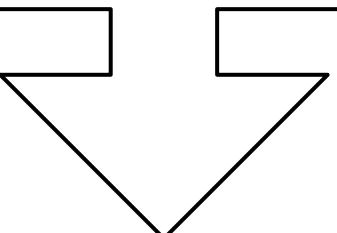
Annotators' personal  
preferences

Annotators'  
skill levels

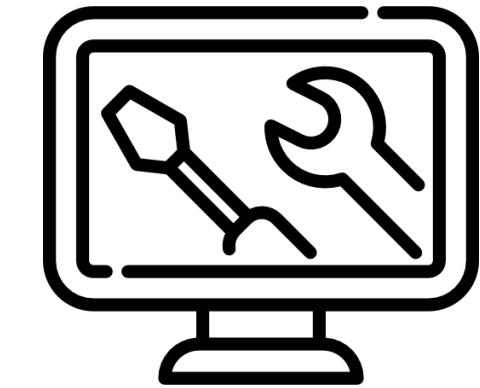
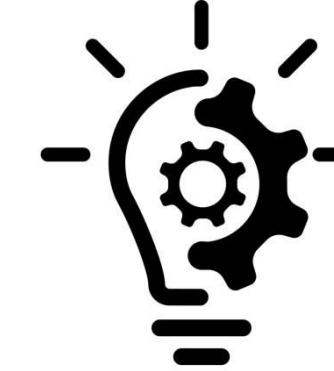
Segmentation  
criteria

Segmentation  
tools

**Latent factors**



# Variability in Medical Image Segmentation



Ambiguous object  
boundaries

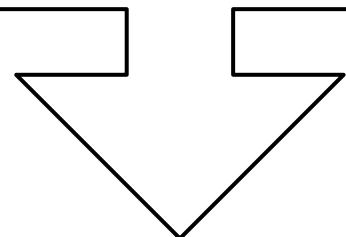
Annotators' personal  
preferences

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 (dermatoscopy 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 (dermatoscopy 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.

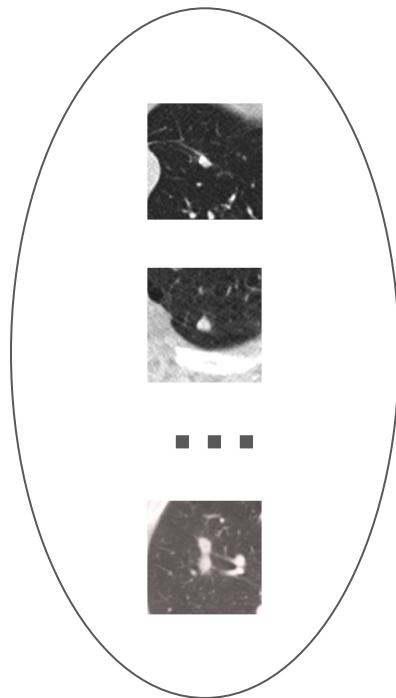
***Abstract***—In a double blind evaluation of 60 digital dermatoscopic images by 4 “junior”, 4 “senior” and 4 “expert” dermatologists (dermatoscopy training respectively less than 1 year, between 1 and 5 years, and more than 5 years), a significant inter-operator variability was observed in melanocytic 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.

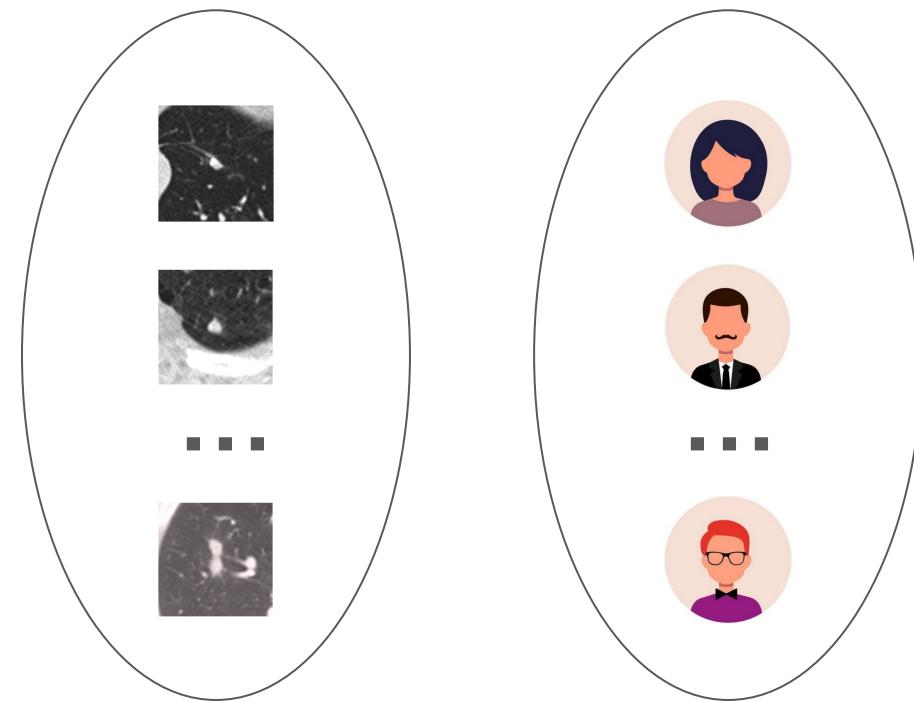
**MSeg methods** model and predict multiple segmentations to capture annotation variability.

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

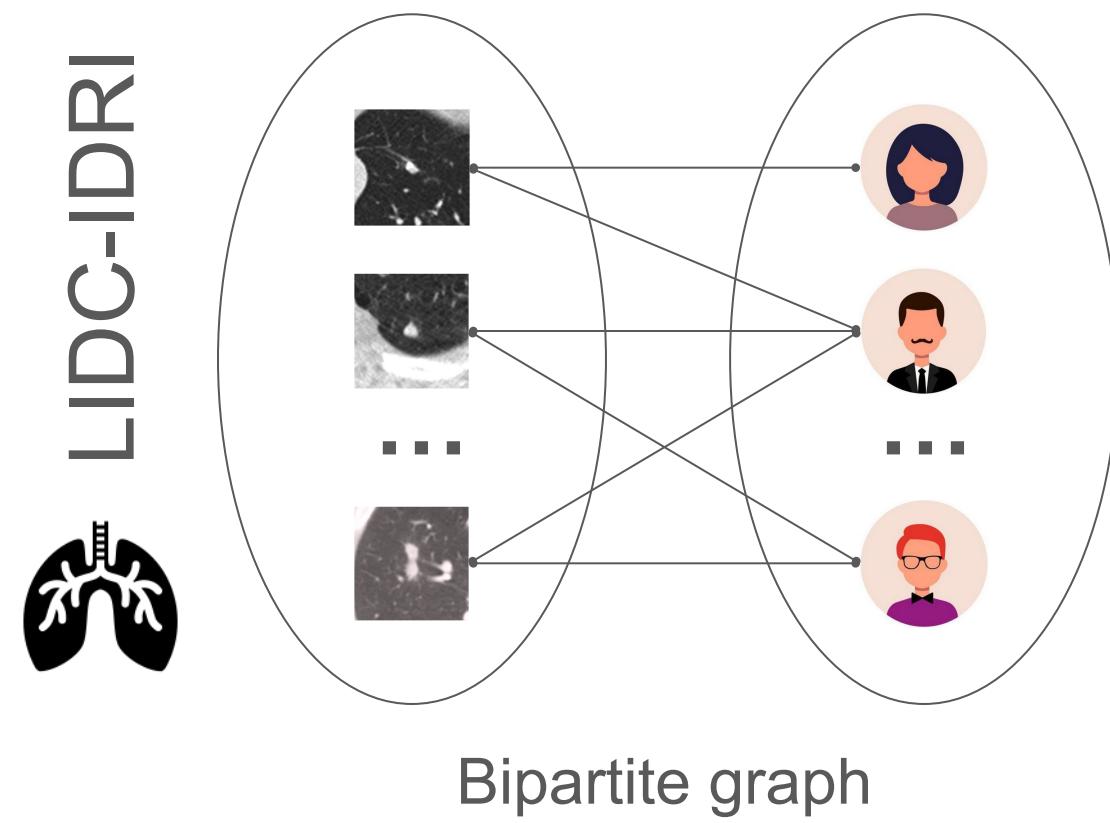
# Multi-Annotator Medical Image Segmentation Datasets



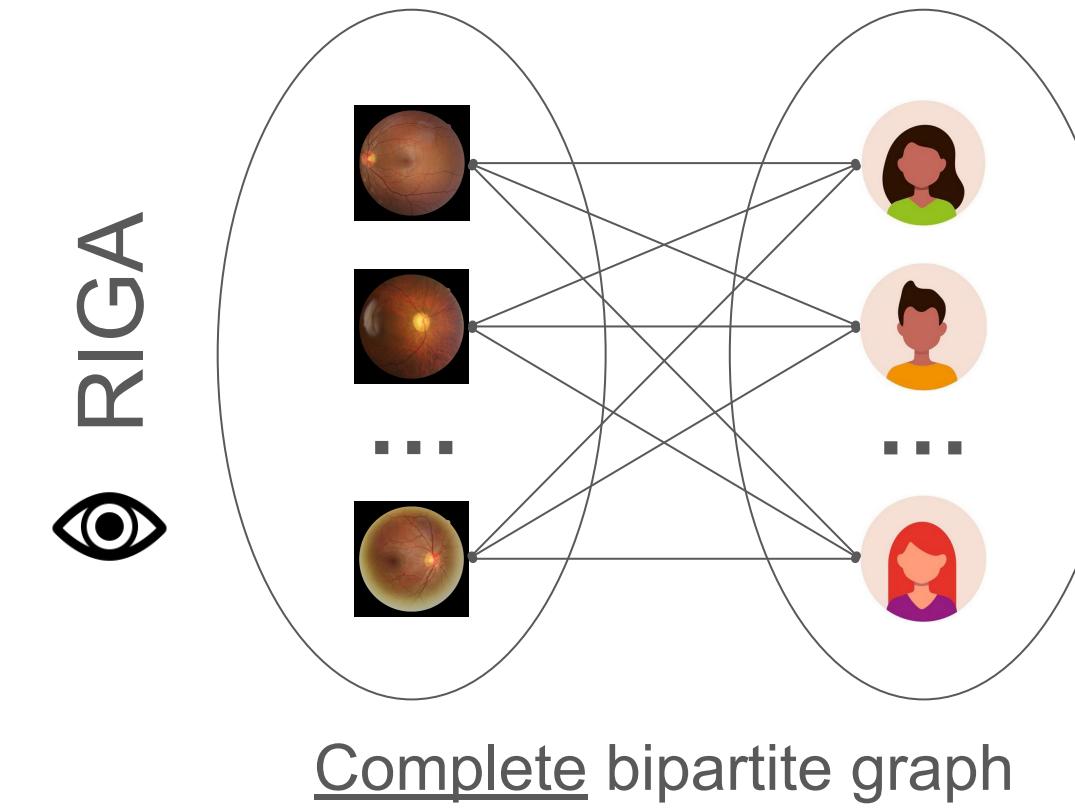
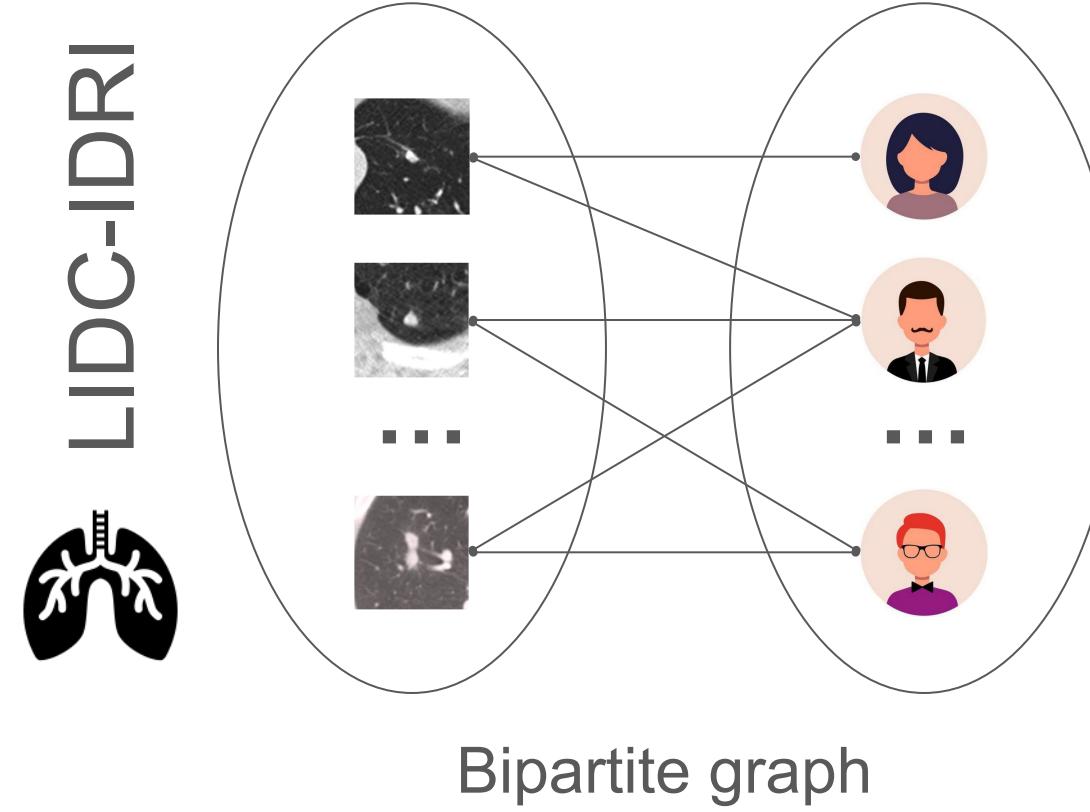
# Multi-Annotator Medical Image Segmentation Datasets



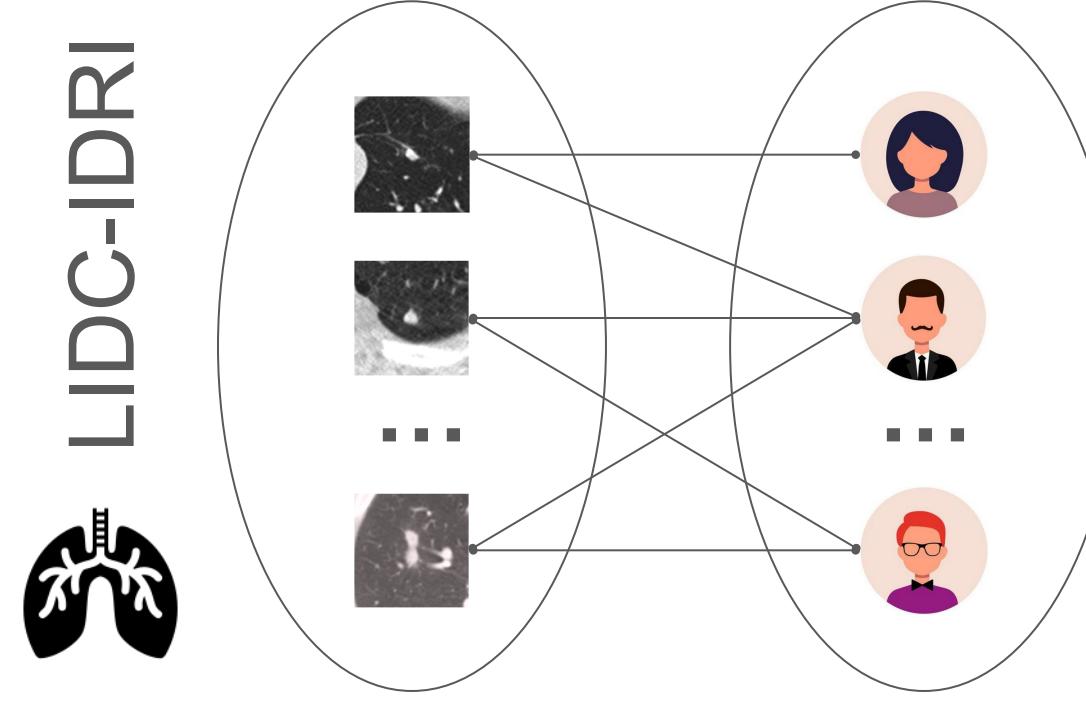
# Multi-Annotator Medical Image Segmentation Datasets



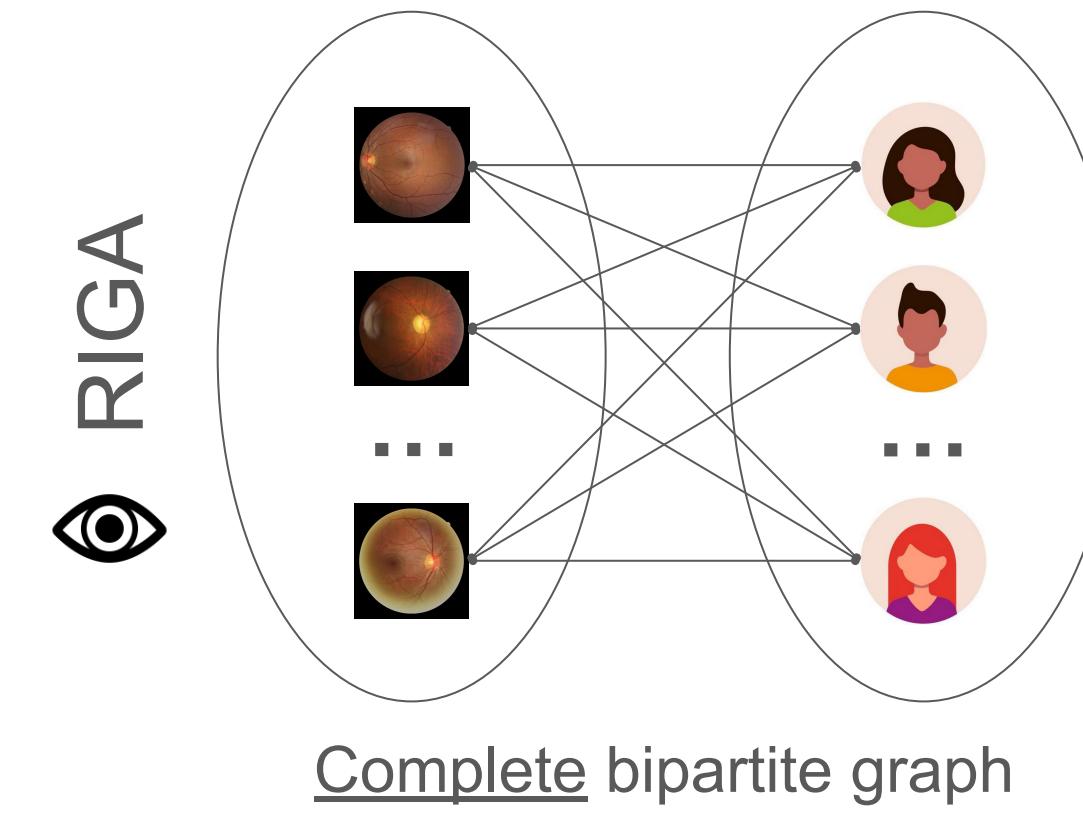
# Multi-Annotator Medical Image Segmentation Datasets



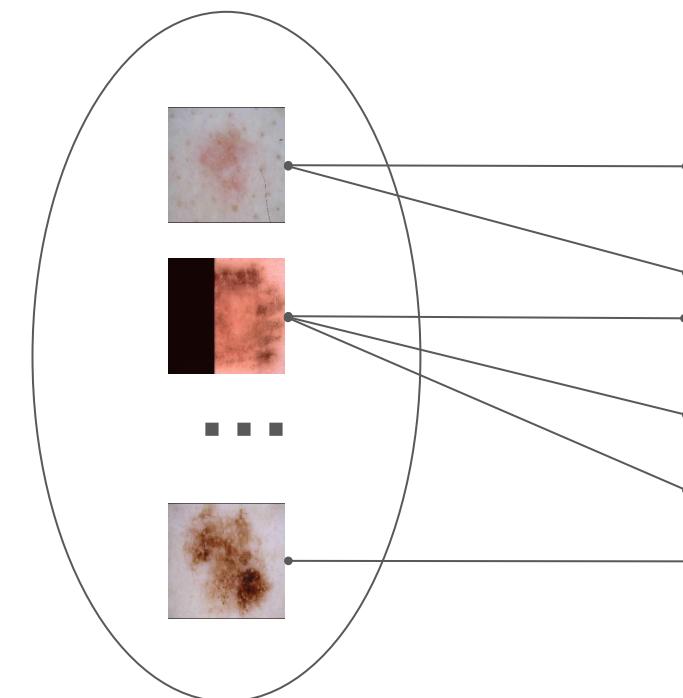
# Multi-Annotator Medical Image Segmentation Datasets



Bipartite graph

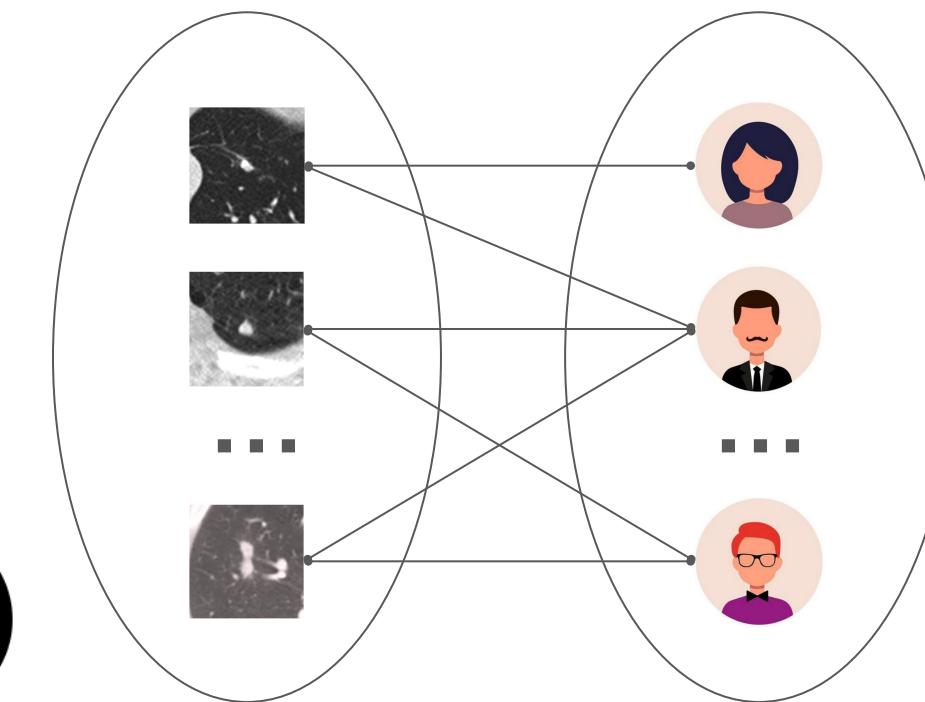


Complete bipartite graph



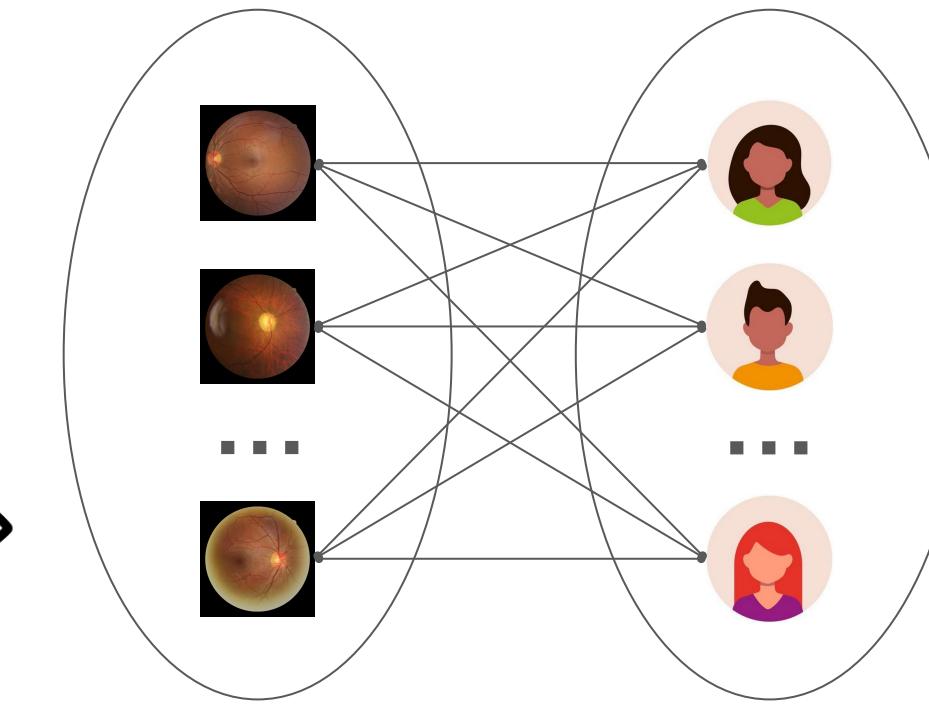
# Multi-Annotator Medical Image Segmentation Datasets

LIDC-IDRI  
lungs

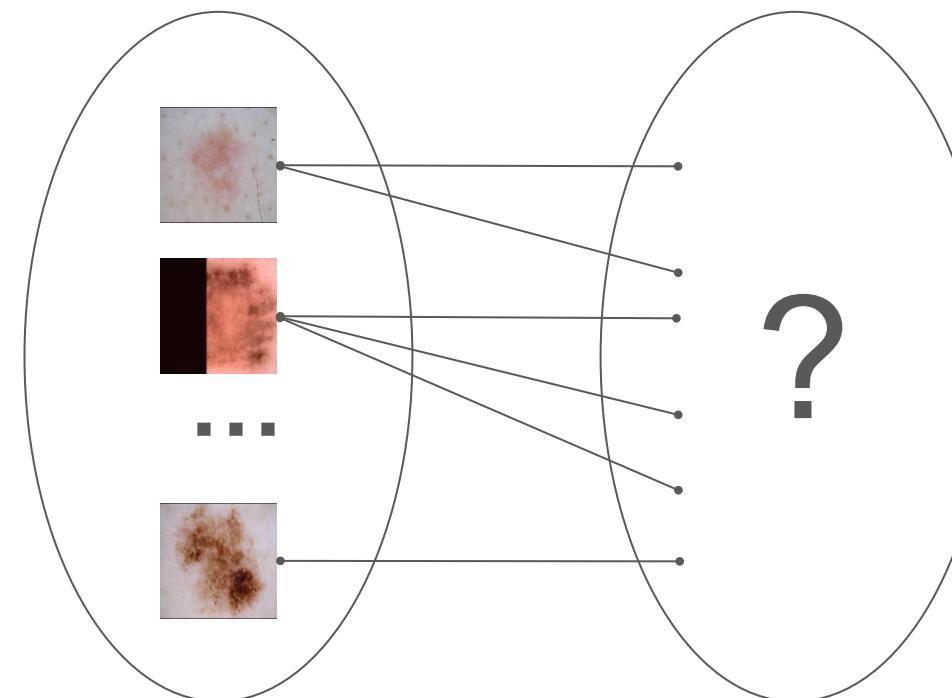


Bipartite graph

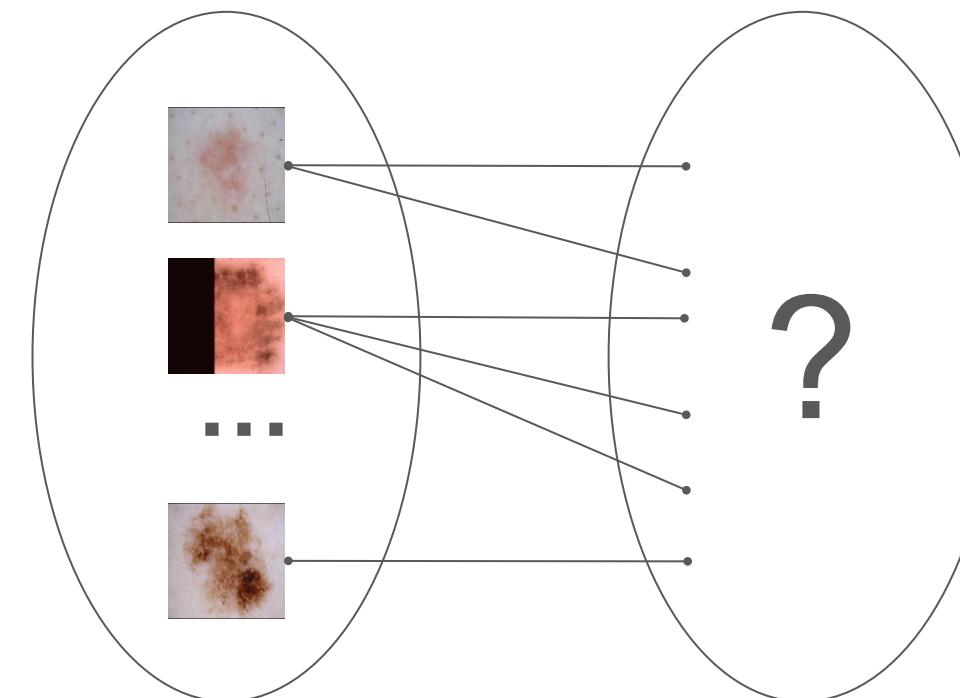
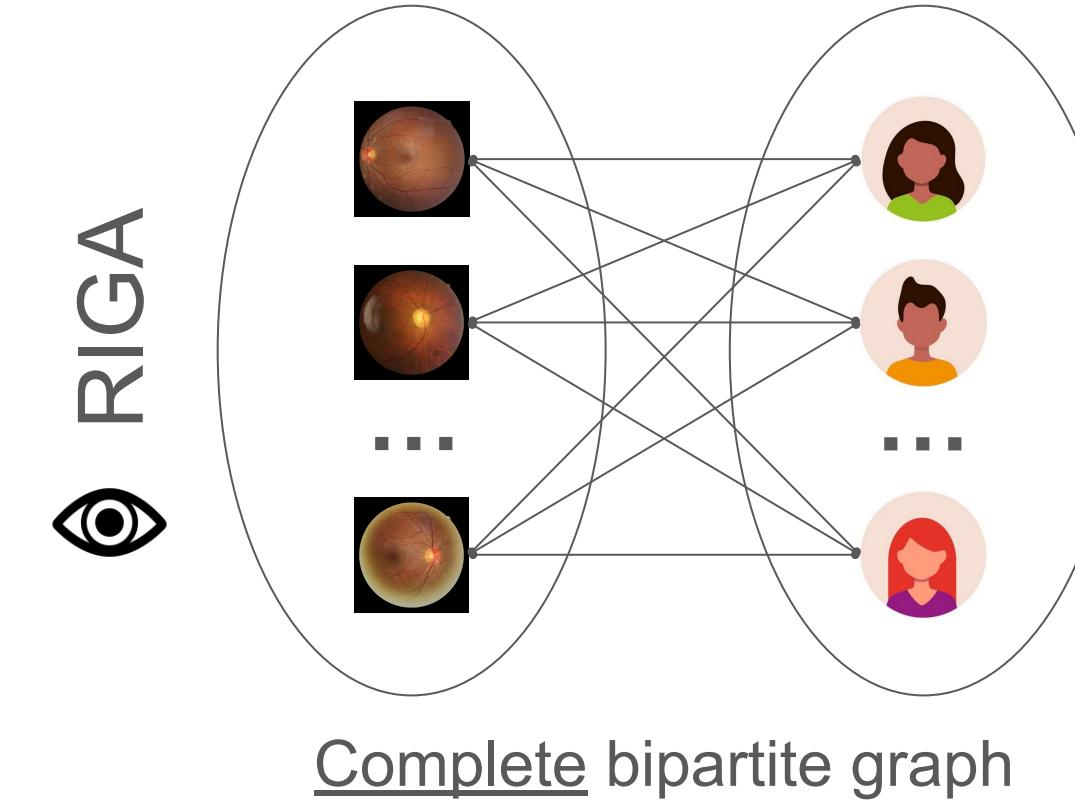
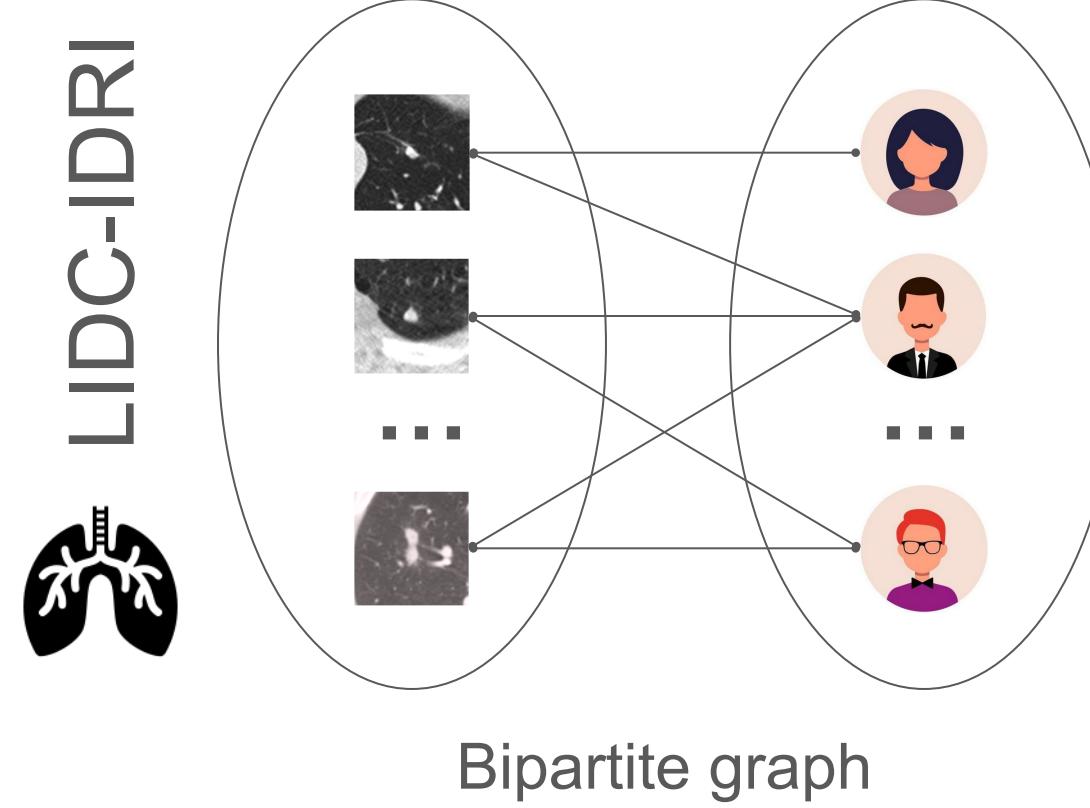
RIGA  
eye



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

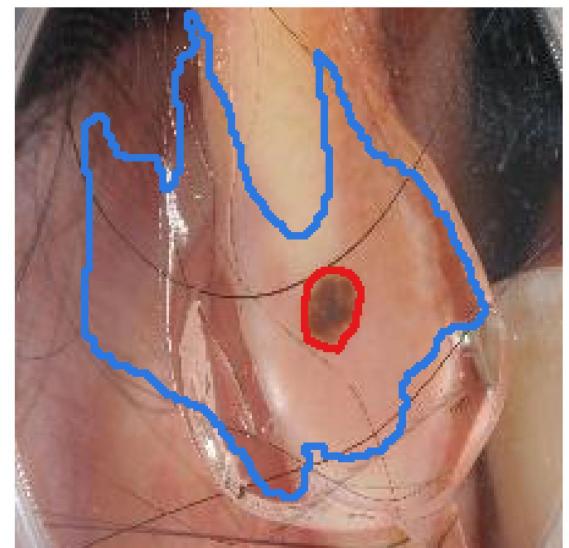
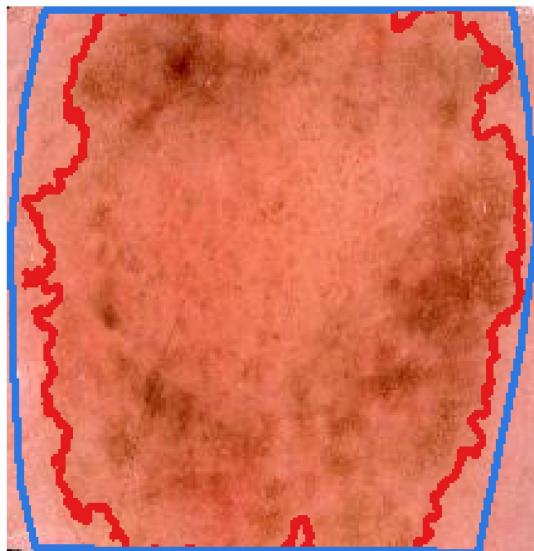
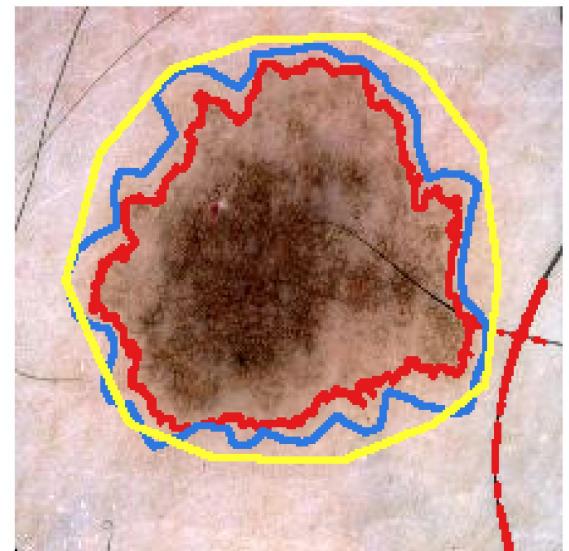
2,261 images with more than 1 “ground truth” segmentation mask

⇒ 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

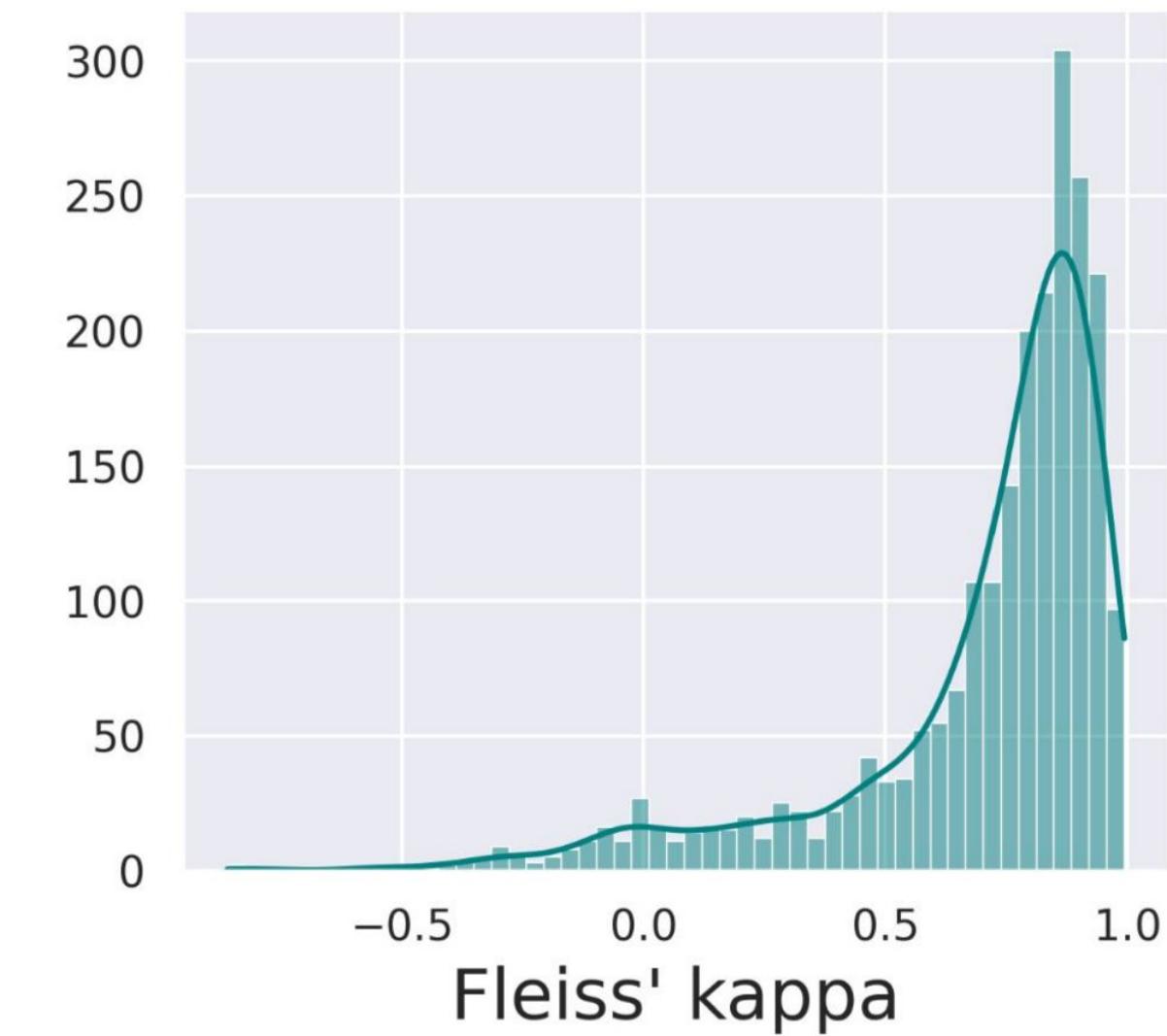
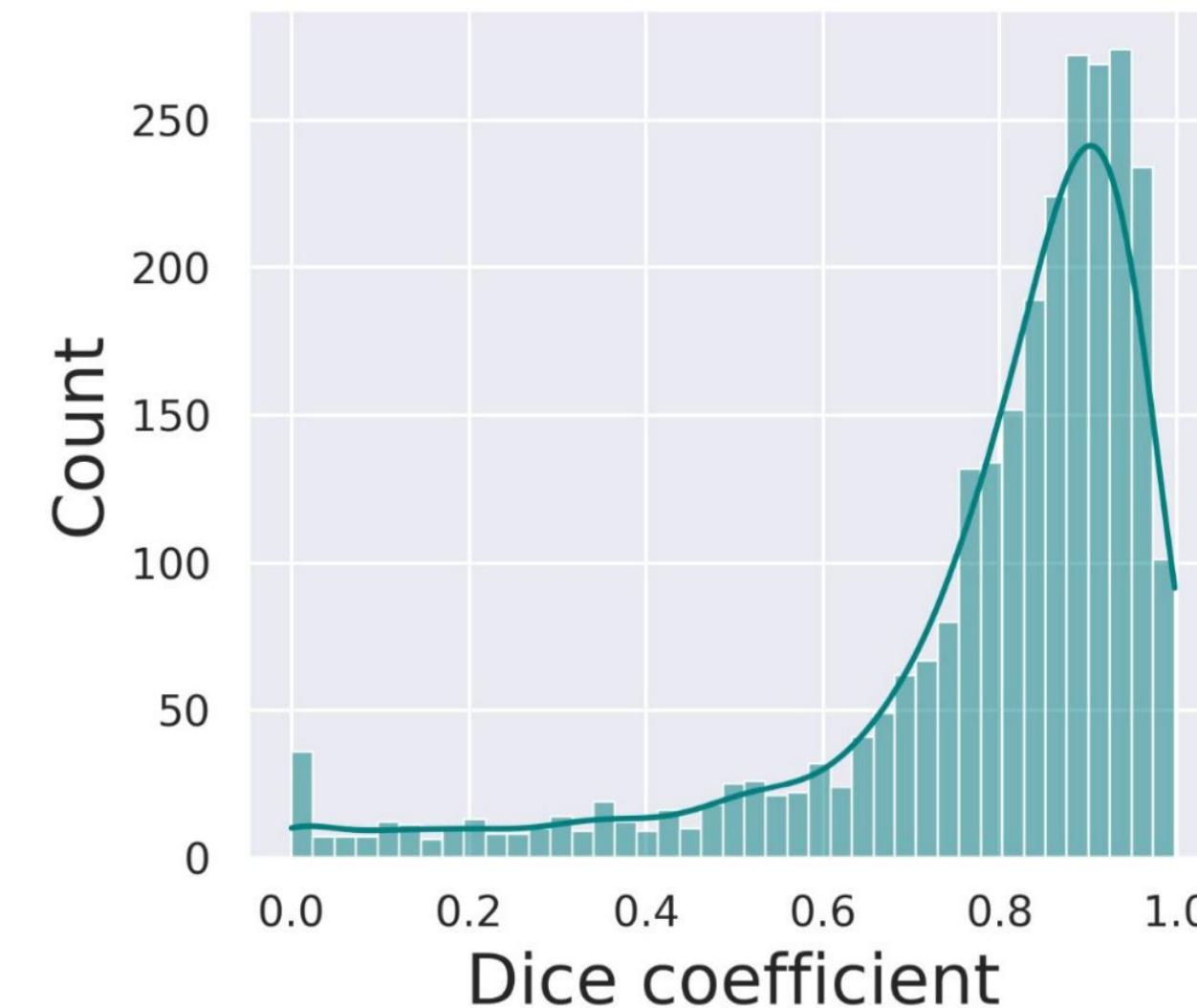
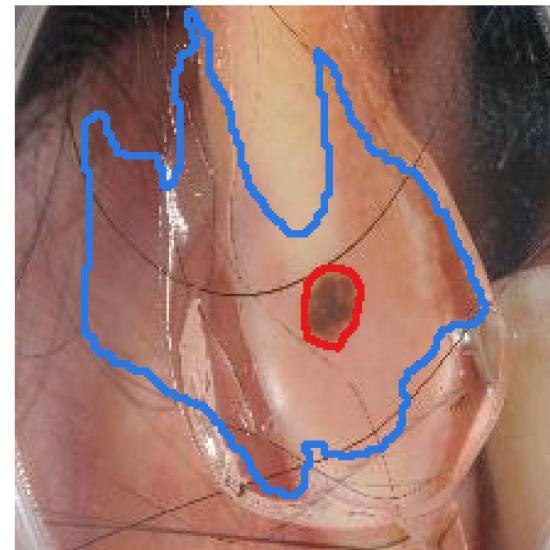
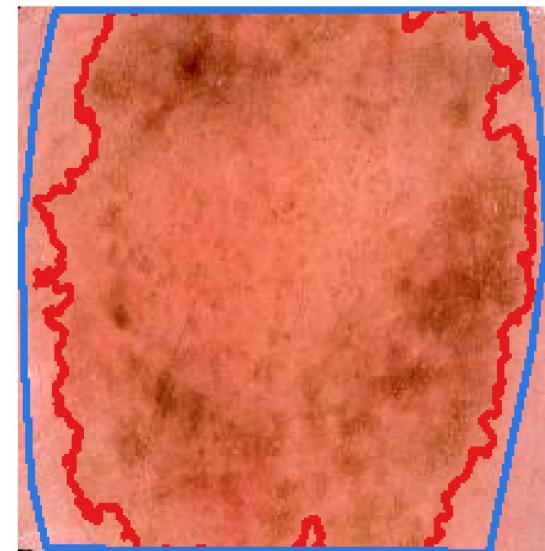
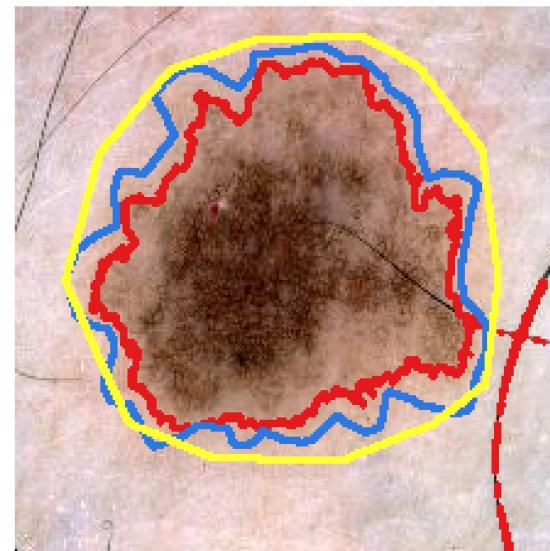
⇒ 4,704 training image-mask pairs for skin lesion segmentation (**SLS**).



# Segmentations in ISIC Archive and their Variability

2,261 images with more than 1 “ground truth” segmentation mask

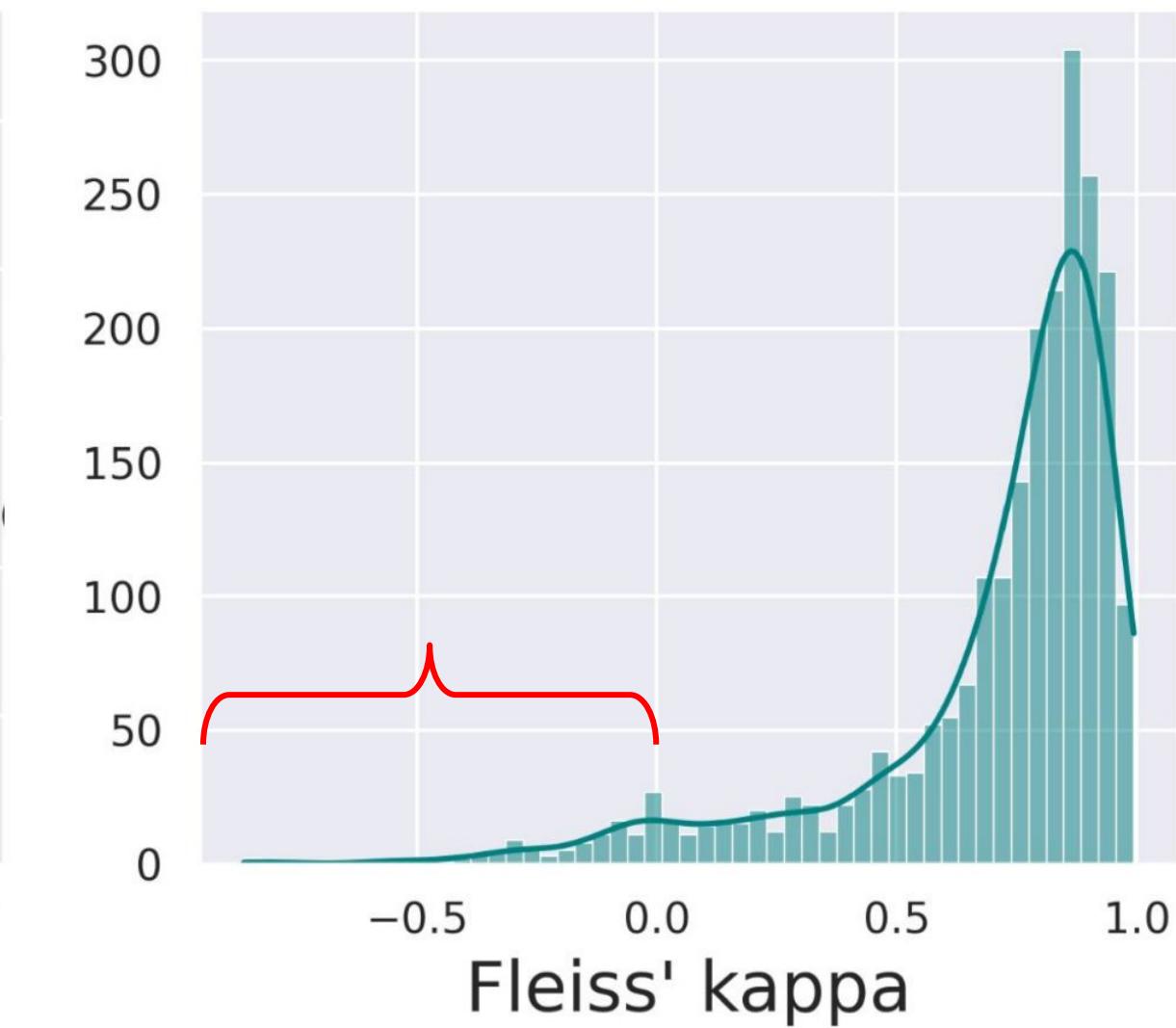
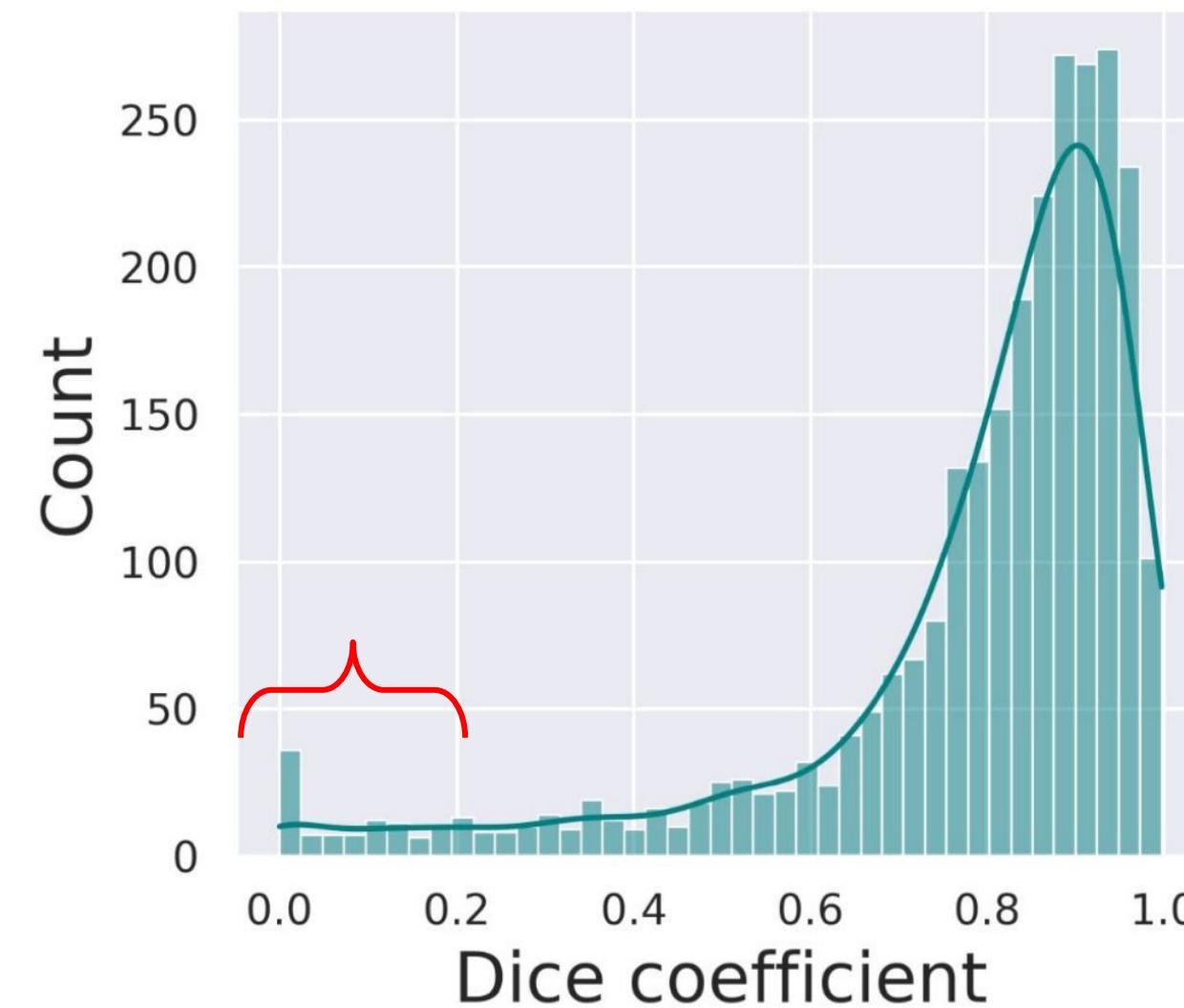
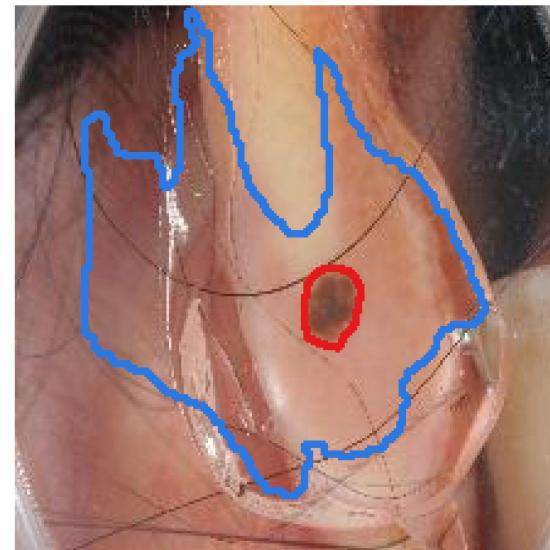
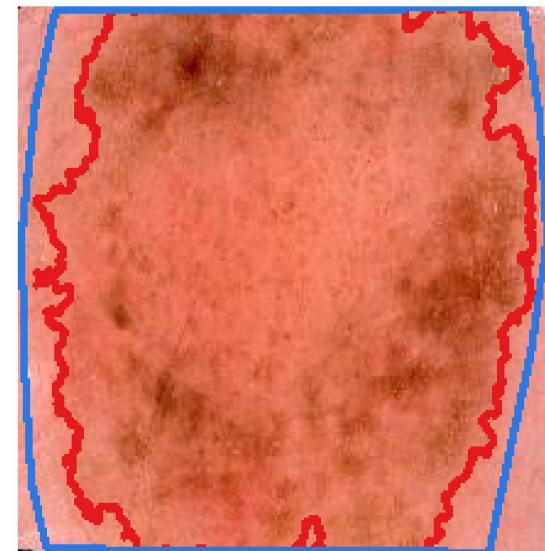
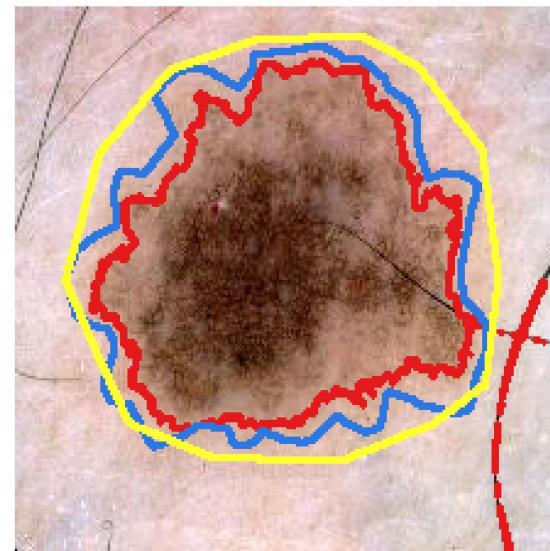
⇒ 4,704 training image-mask pairs for skin lesion segmentation (**SLS**).



# Segmentations in ISIC Archive and their Variability

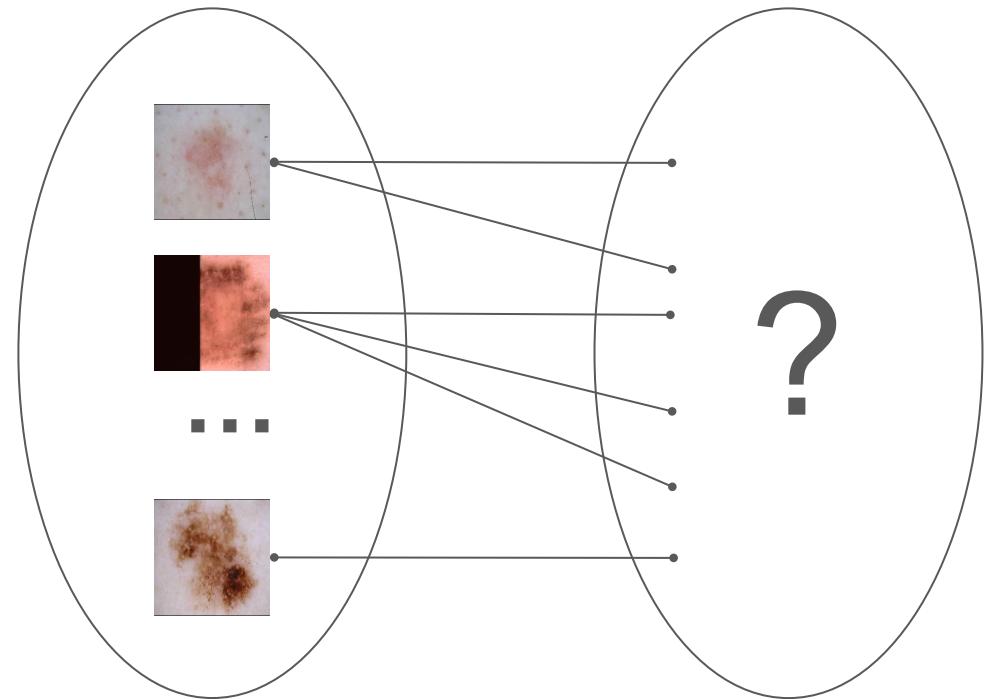
2,261 images with more than 1 “ground truth” segmentation mask

⇒ 4,704 training image-mask pairs for skin lesion segmentation (**SLS**).



# Objective

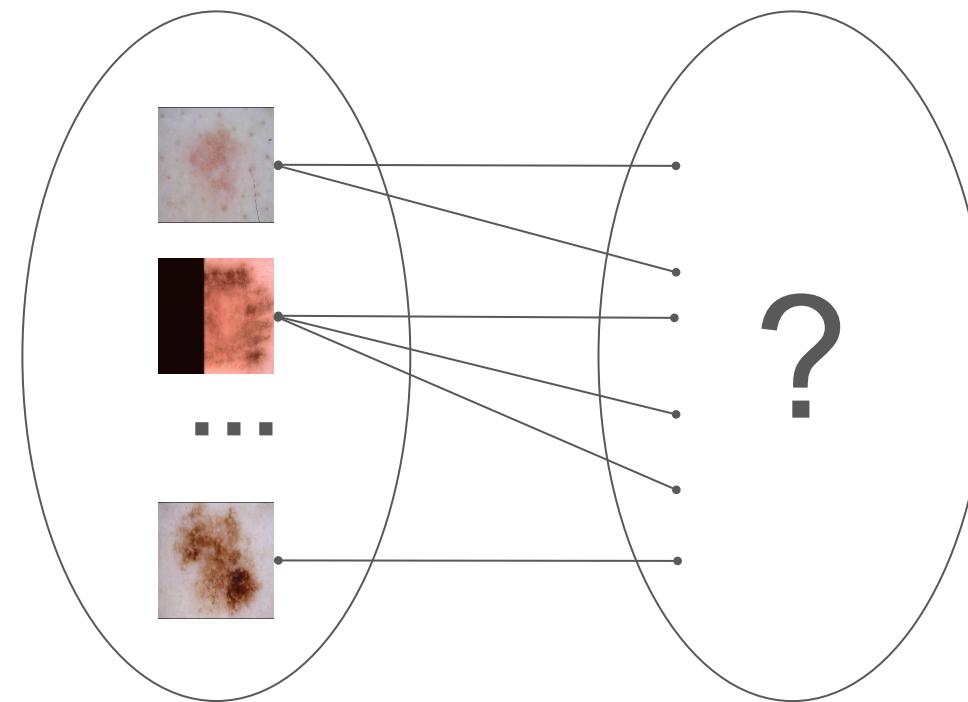
Given



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

# Objective

Given

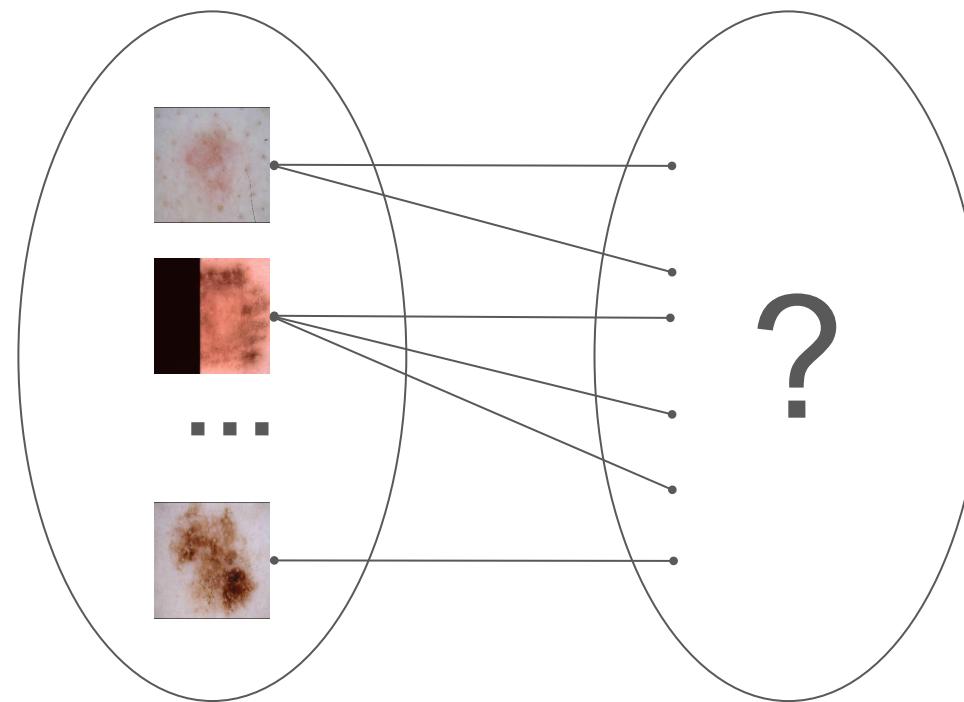


, train a model that **discovers unique annotation styles** such that:

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

# Objective

Given

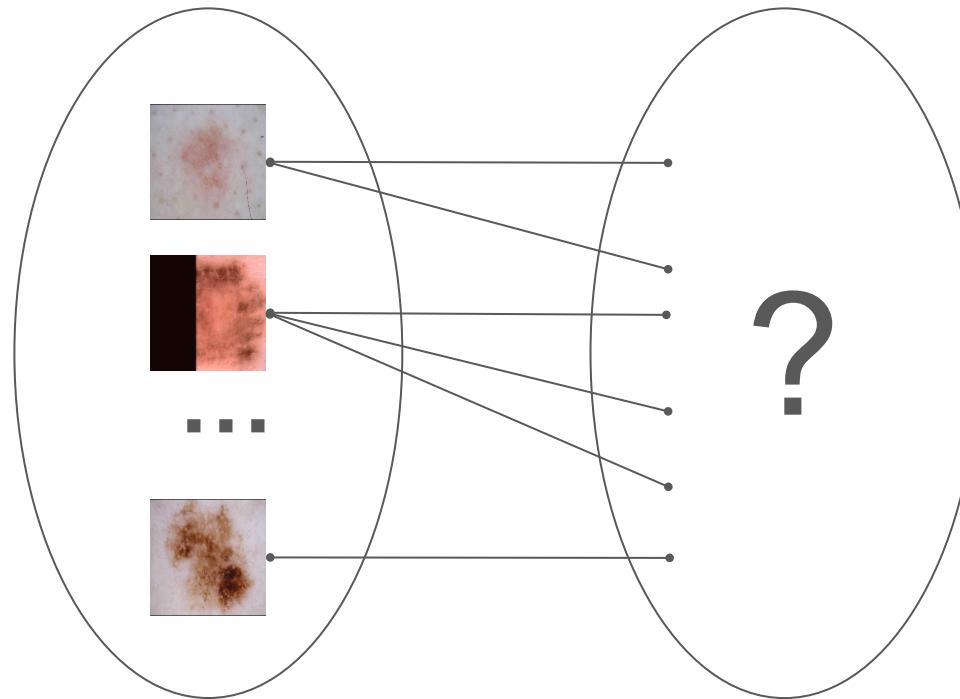


, train a model that **discovers unique annotation styles** such that:

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

# Objective

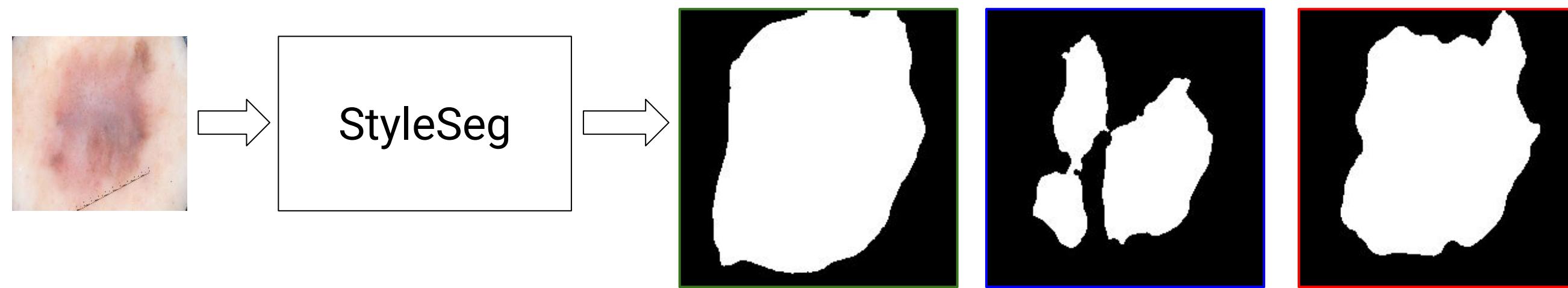
Given



, 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  
 $f_s(X_i; \Theta_s)$

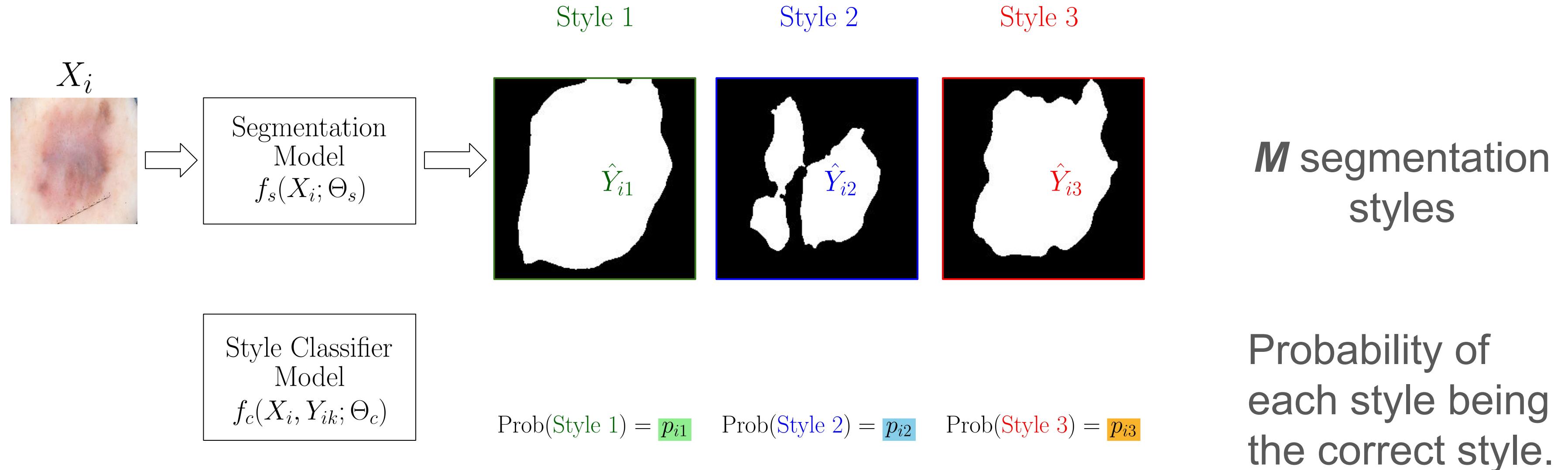
**$M$**  segmentation  
styles

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

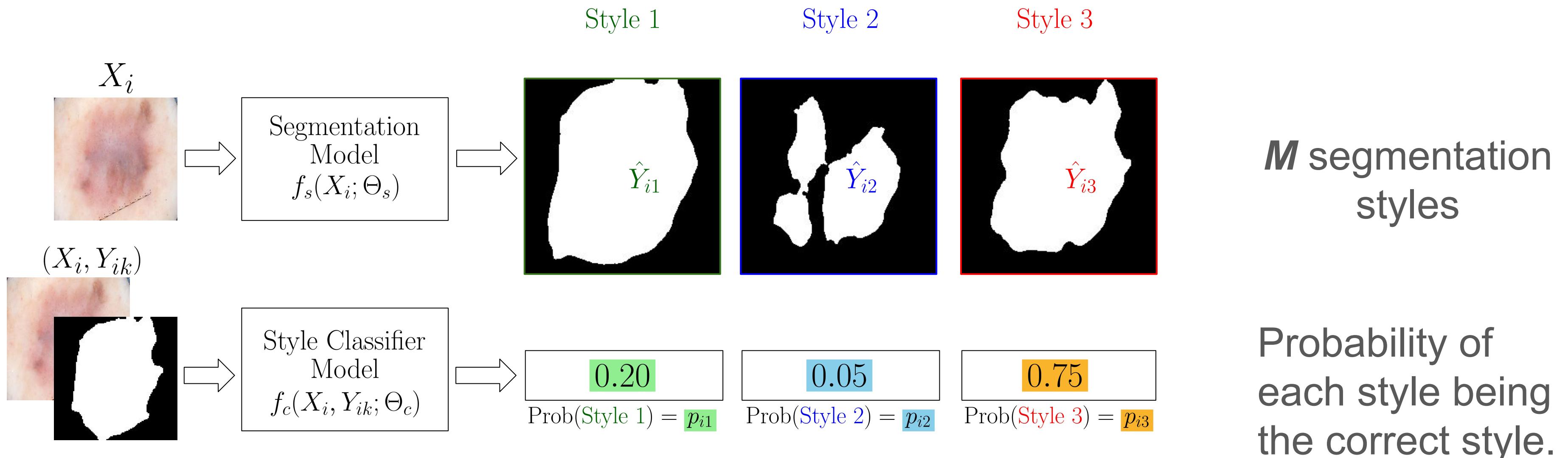
Prob(Style 1) =  $p_{i1}$    Prob(Style 2) =  $p_{i2}$    Prob(Style 3) =  $p_{i3}$

Probability of  
each style being  
the correct style.

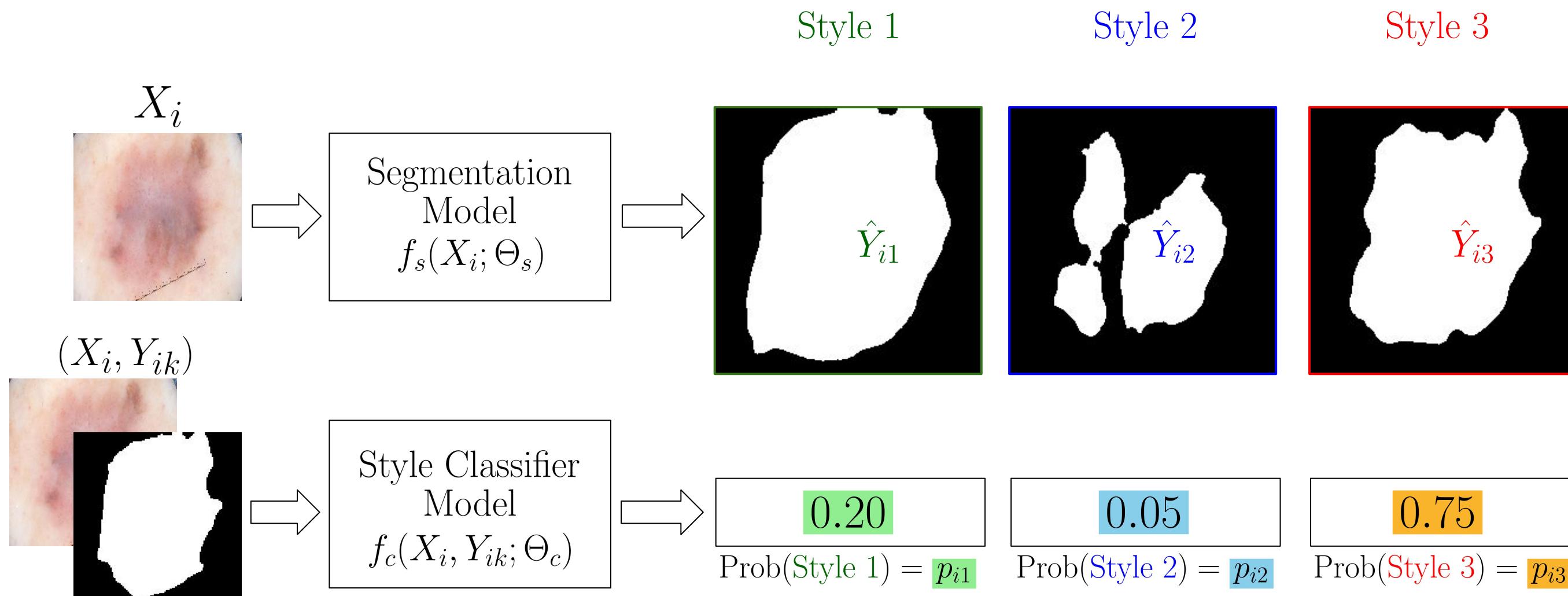
# Multiple segmentation styles and their probabilities



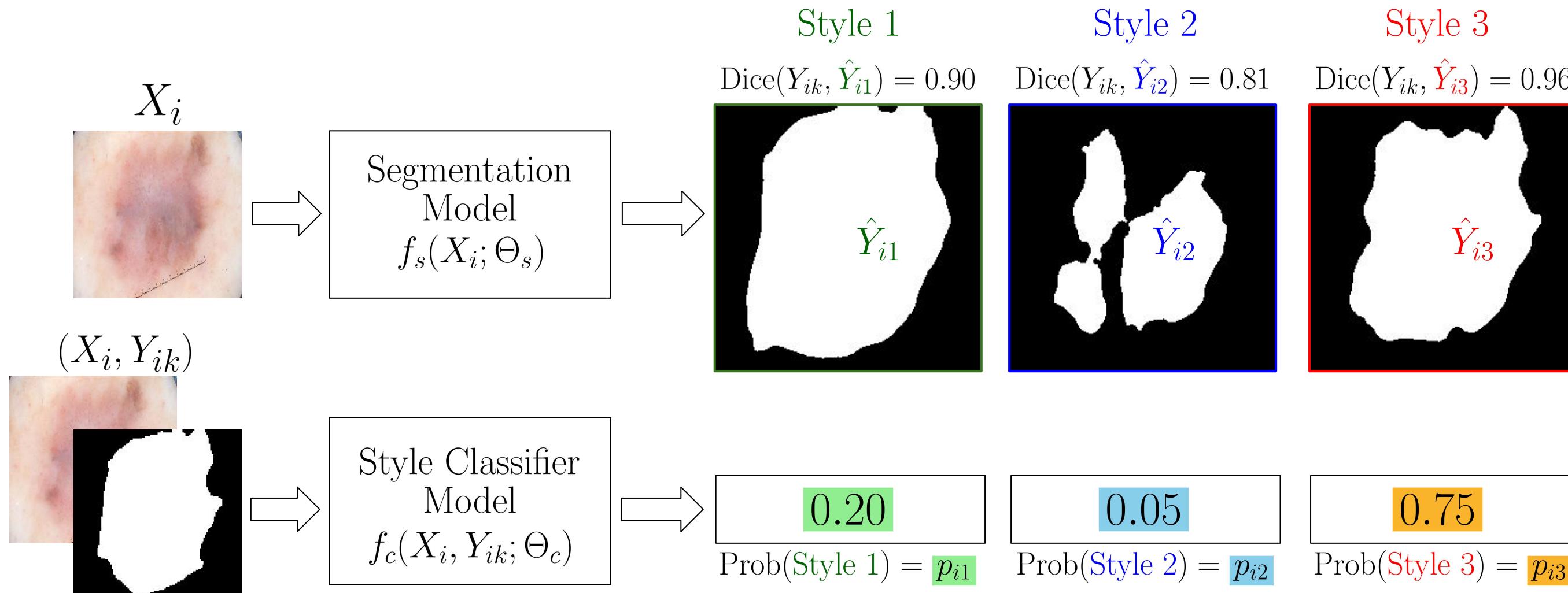
# Multiple segmentation styles and their probabilities



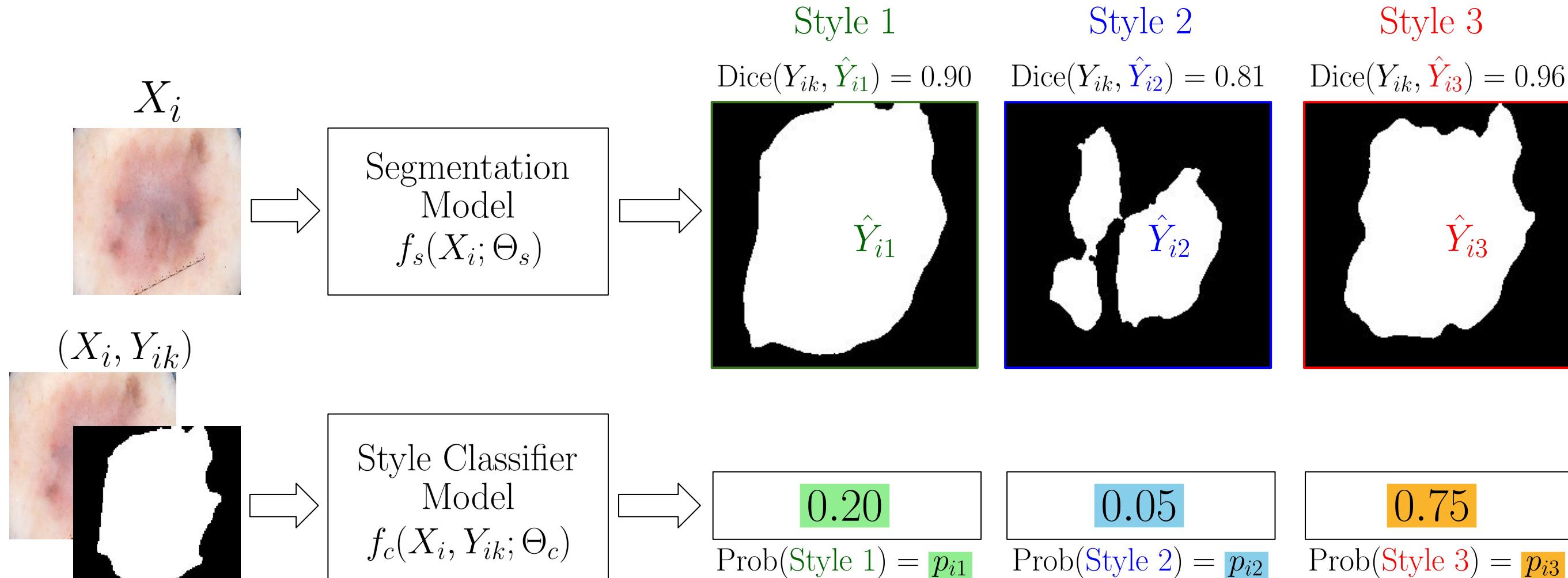
# Training StyleSeg



# Training StyleSeg



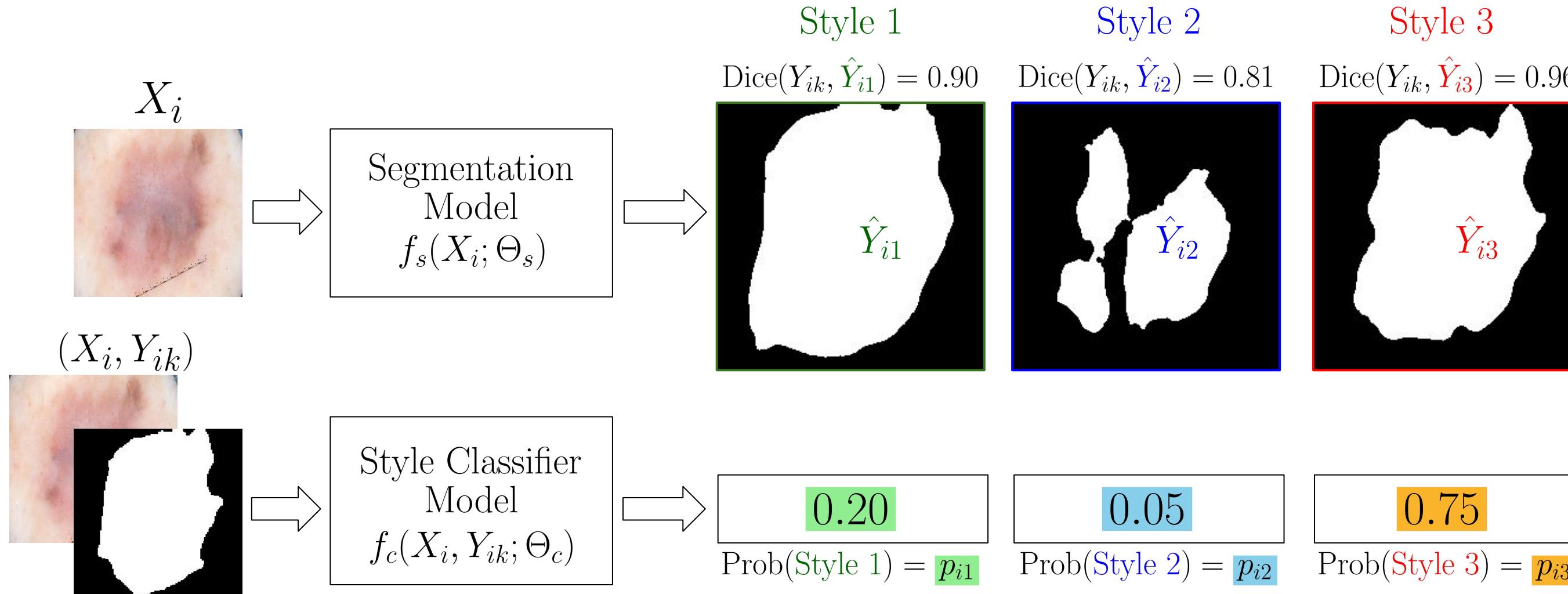
# Training StyleSeg



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

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

# Training StyleSeg



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

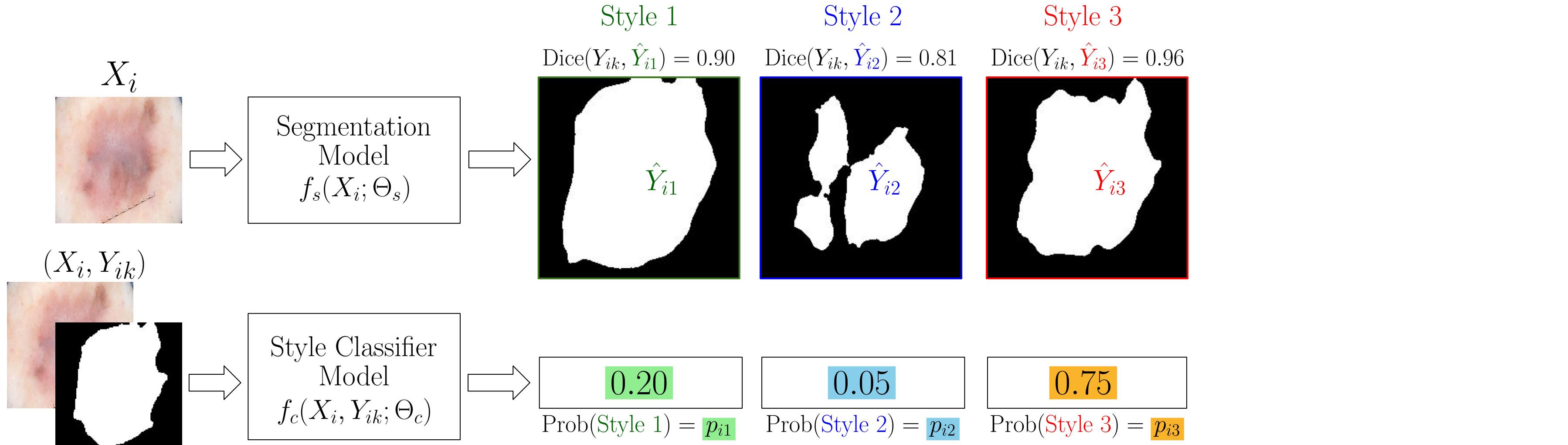
$$\mathcal{L}_1 = L_D(Y_{ik}, \hat{Y}_{i3})$$

$$\mathcal{L}_2 = L_D(Y_{ik}, (0.20 * \hat{Y}_{i1})$$

$$+(0.05 * \hat{Y}_{i2})$$

$$+(0.75 * \hat{Y}_{i3}))$$

# Training StyleSeg



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

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

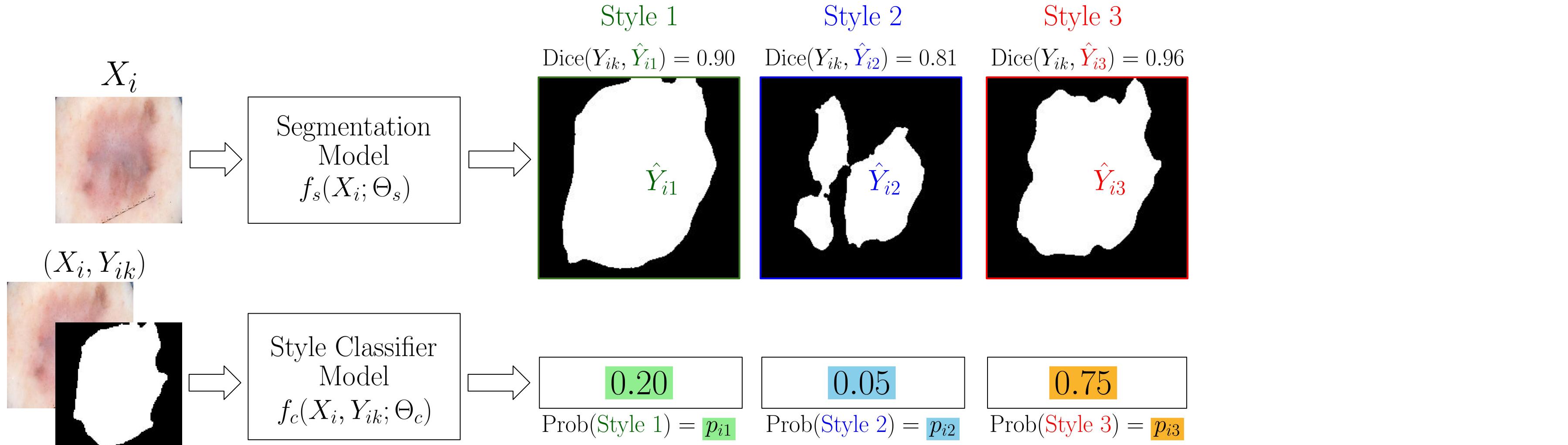
$$\mathcal{L}_2 = L_D(Y_{ik}, (0.20 * \hat{Y}_{i1})$$

$$+(0.05 * \hat{Y}_{i2})$$

$$+(0.75 * \hat{Y}_{i3}))$$

$$\mathcal{L}_3 = L_{CE}([0.20, 0.05, 0.75], [0, 0, 1])$$

# Training StyleSeg



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

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

$$\mathcal{L}_2 = L_D(Y_{ik}, (0.20 * \hat{Y}_{i1})$$

$$+(0.05 * \hat{Y}_{i2})$$

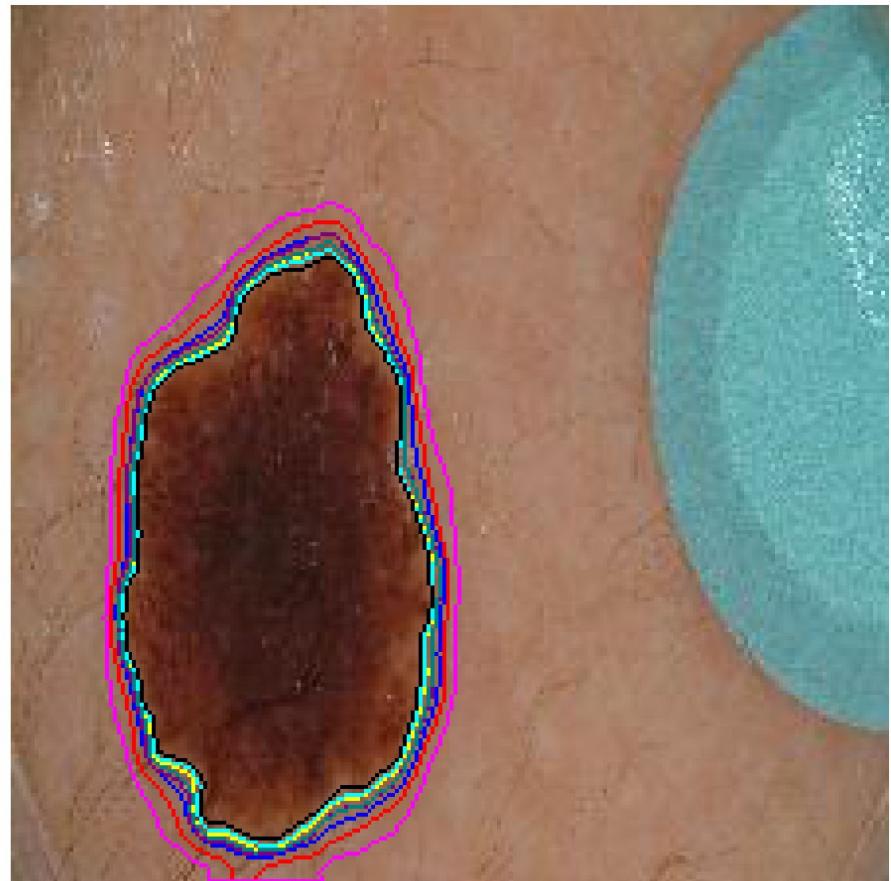
$$+(0.75 * \hat{Y}_{i3}))$$

$$\mathcal{L}_3 = L_{CE}([0.20, 0.05, 0.75], [0, 0, 1])$$

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

# StyleSeg Outputs Adapt to Variability in Lesion Content

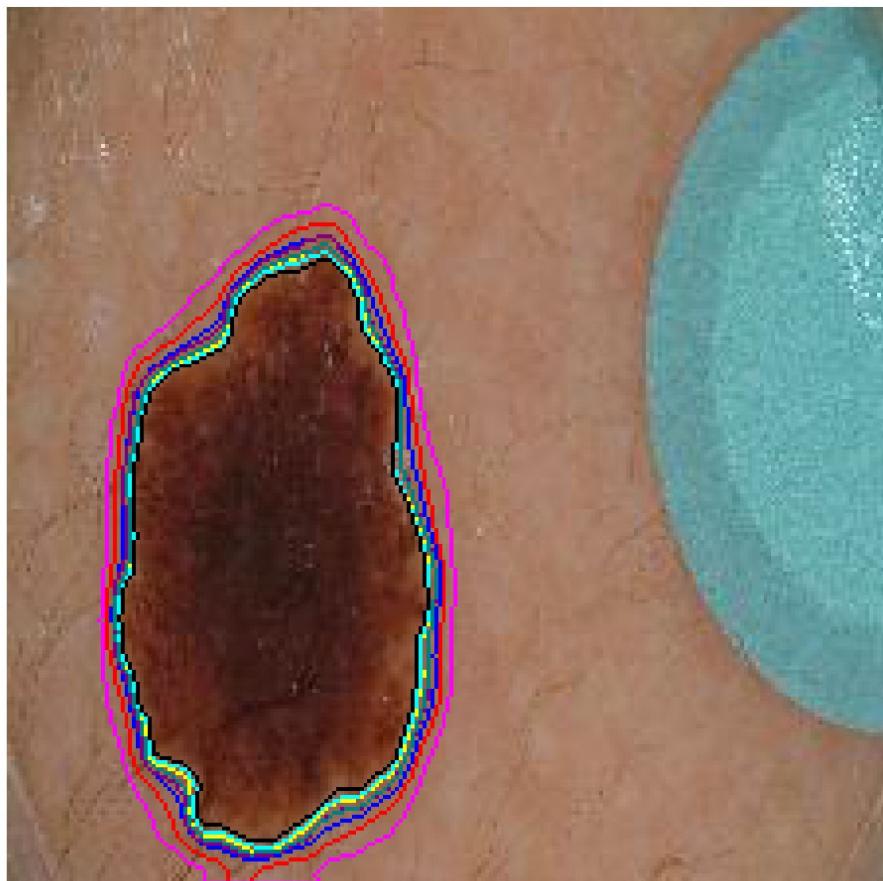
ISIC 0003599



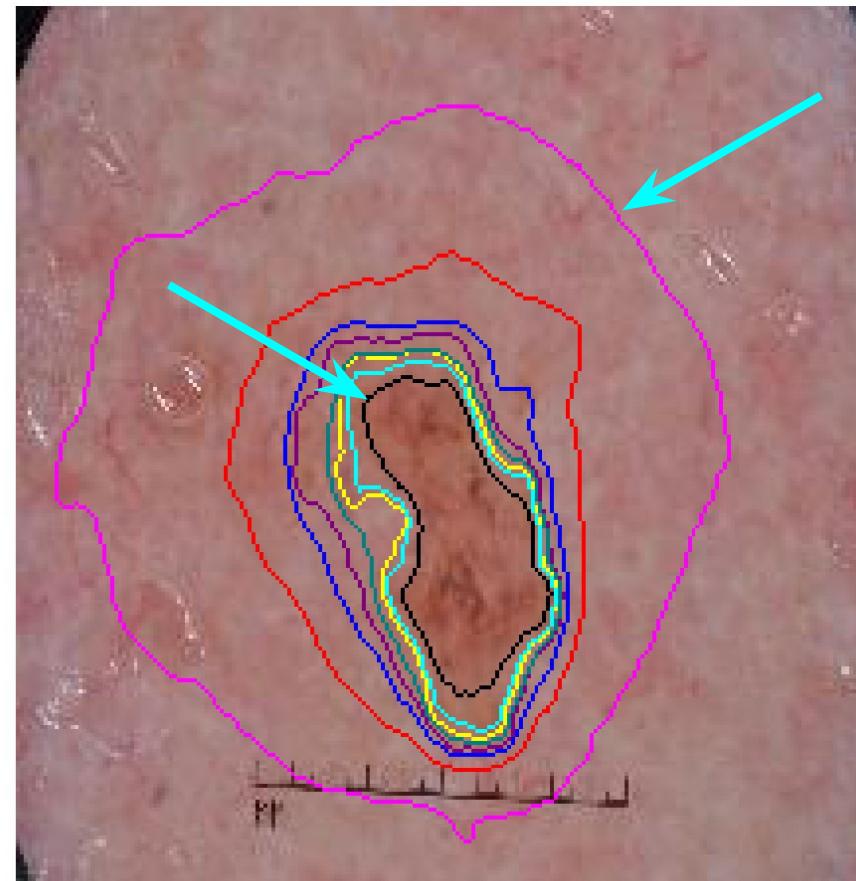
**High-contrast  
lesion has high  
agreement across  
styles**

# StyleSeg Outputs Adapt to Variability in Lesion Content

ISIC 0003599



ISIC\_0014337

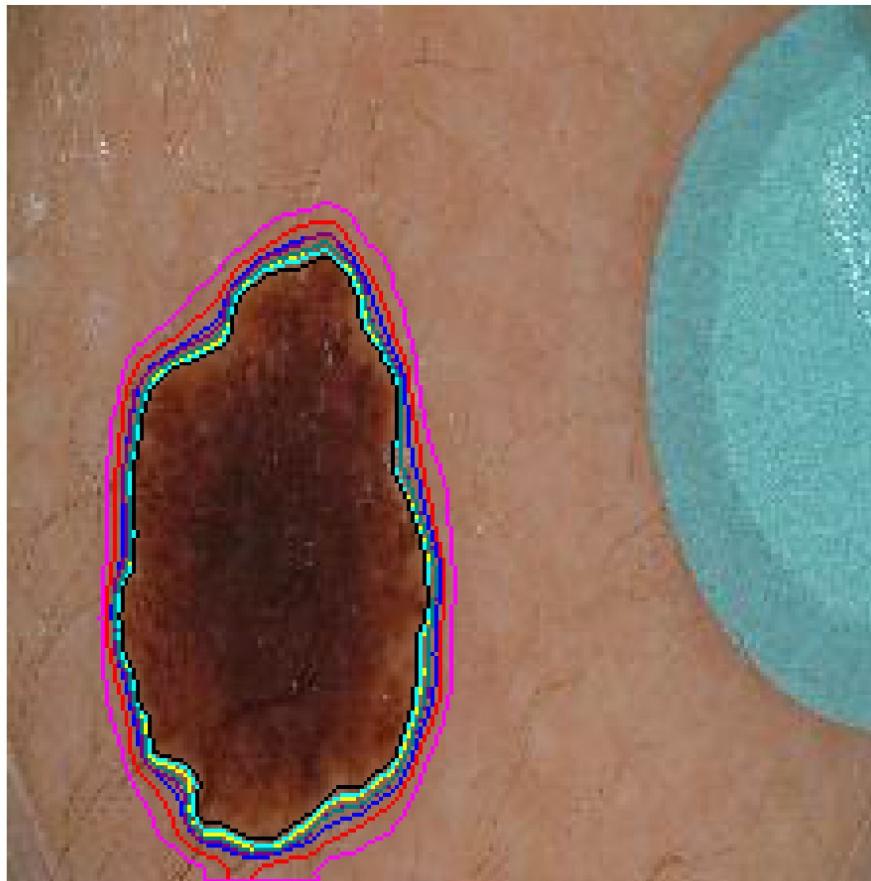


**High-contrast  
lesion has high  
agreement across  
styles**

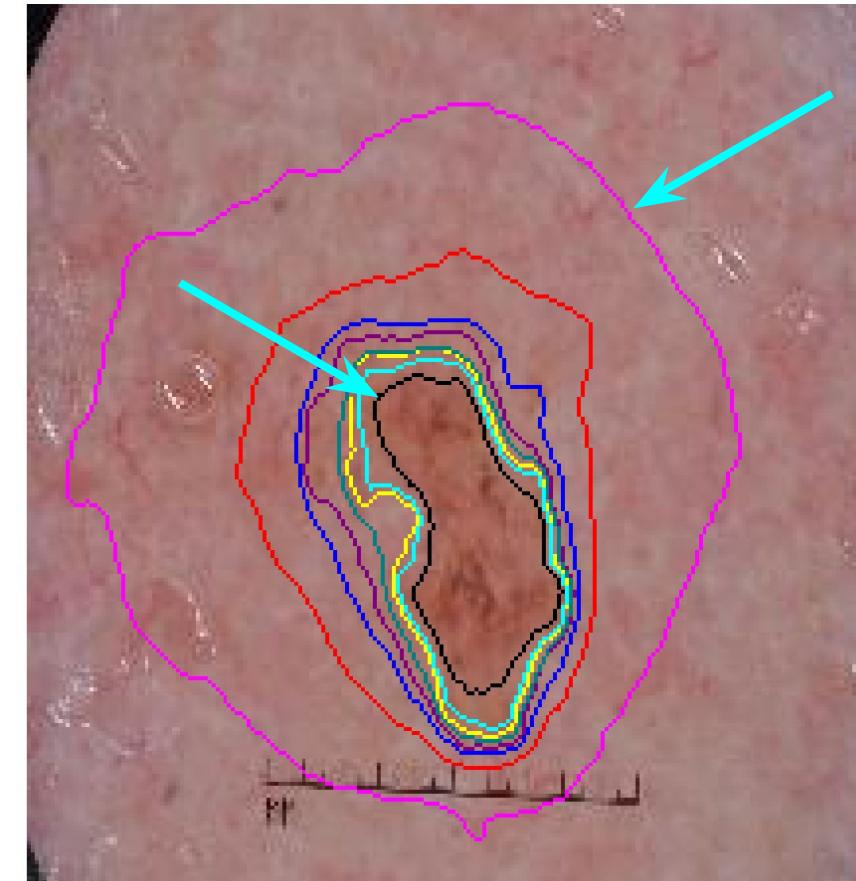
Instances of **under-**  
and **over-**  
segmentation

# StyleSeg Outputs Adapt to Variability in Lesion Content

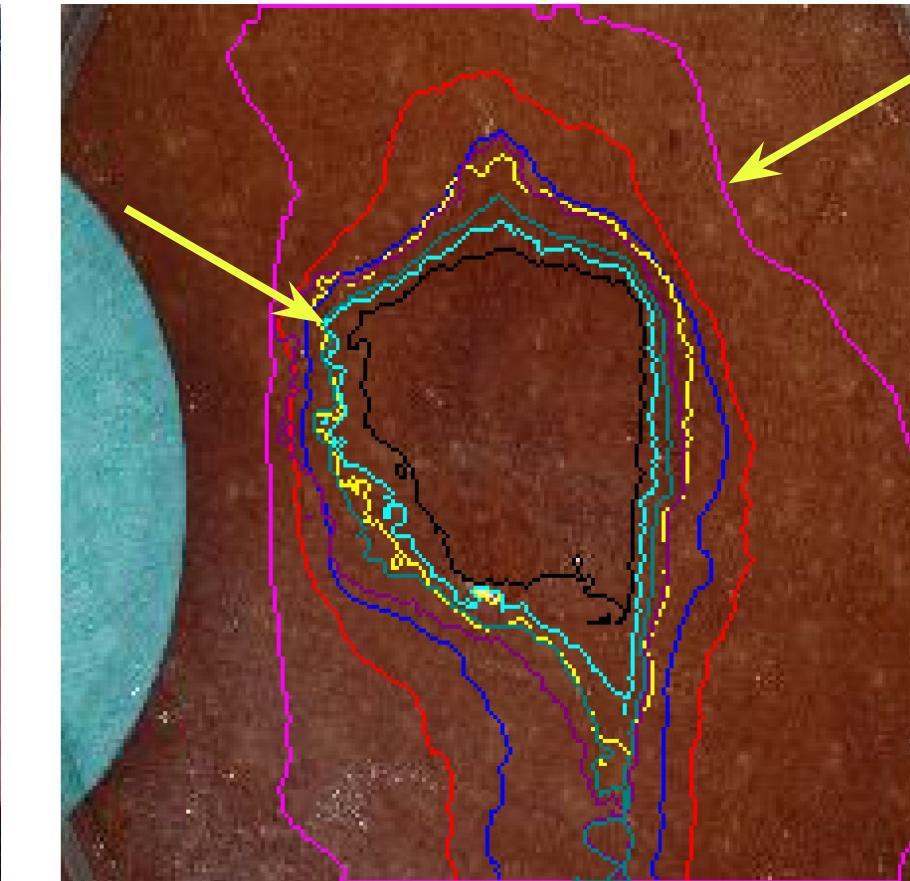
ISIC 0003599



ISIC\_0014337



ISIC\_0003726



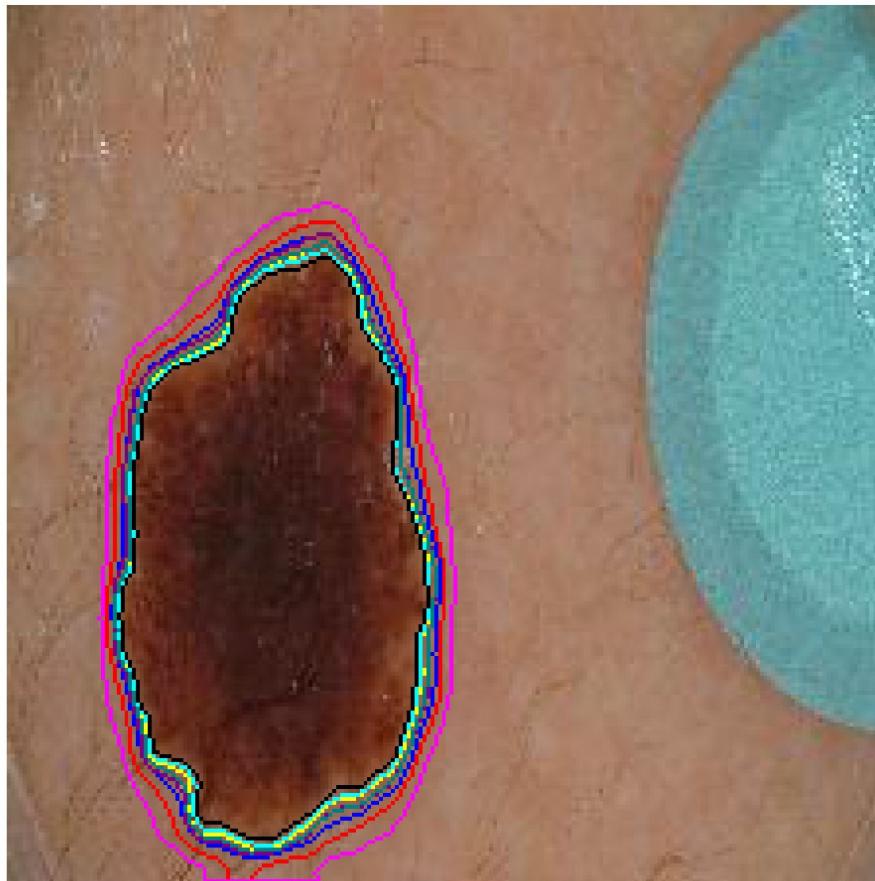
**High-contrast  
lesion has high  
agreement across  
styles**

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

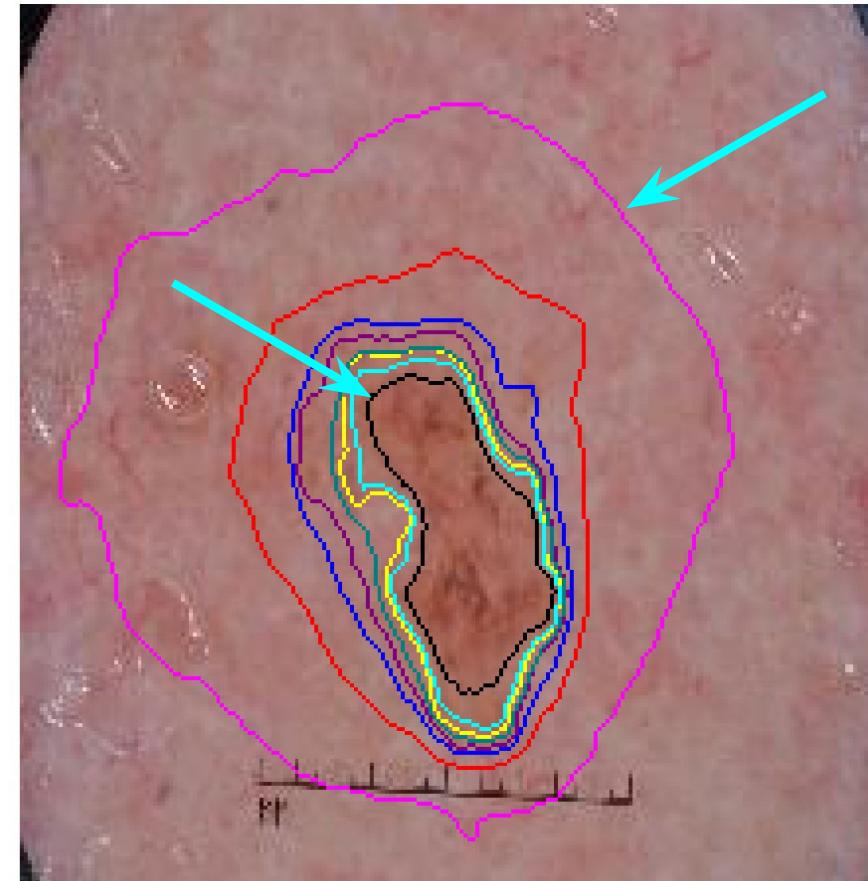
Different **boundary  
jaggedness** across  
segmentations

# StyleSeg Outputs Adapt to Variability in Lesion Content

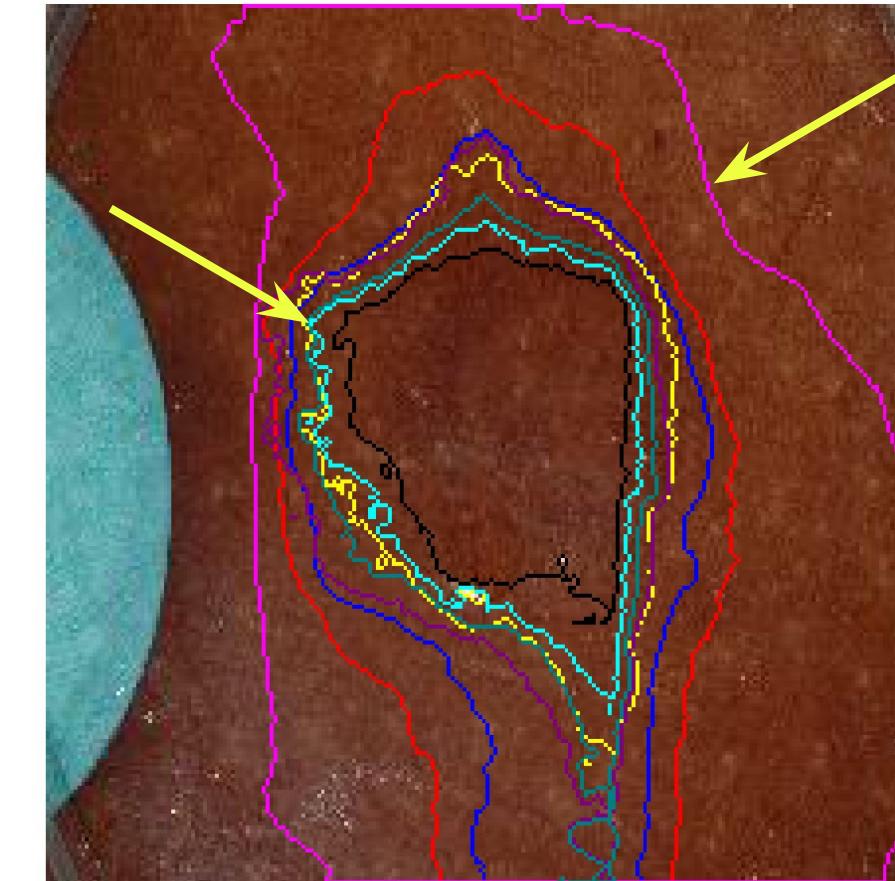
ISIC 0003599



ISIC\_0014337



ISIC\_0003726



ISIC\_0014831



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

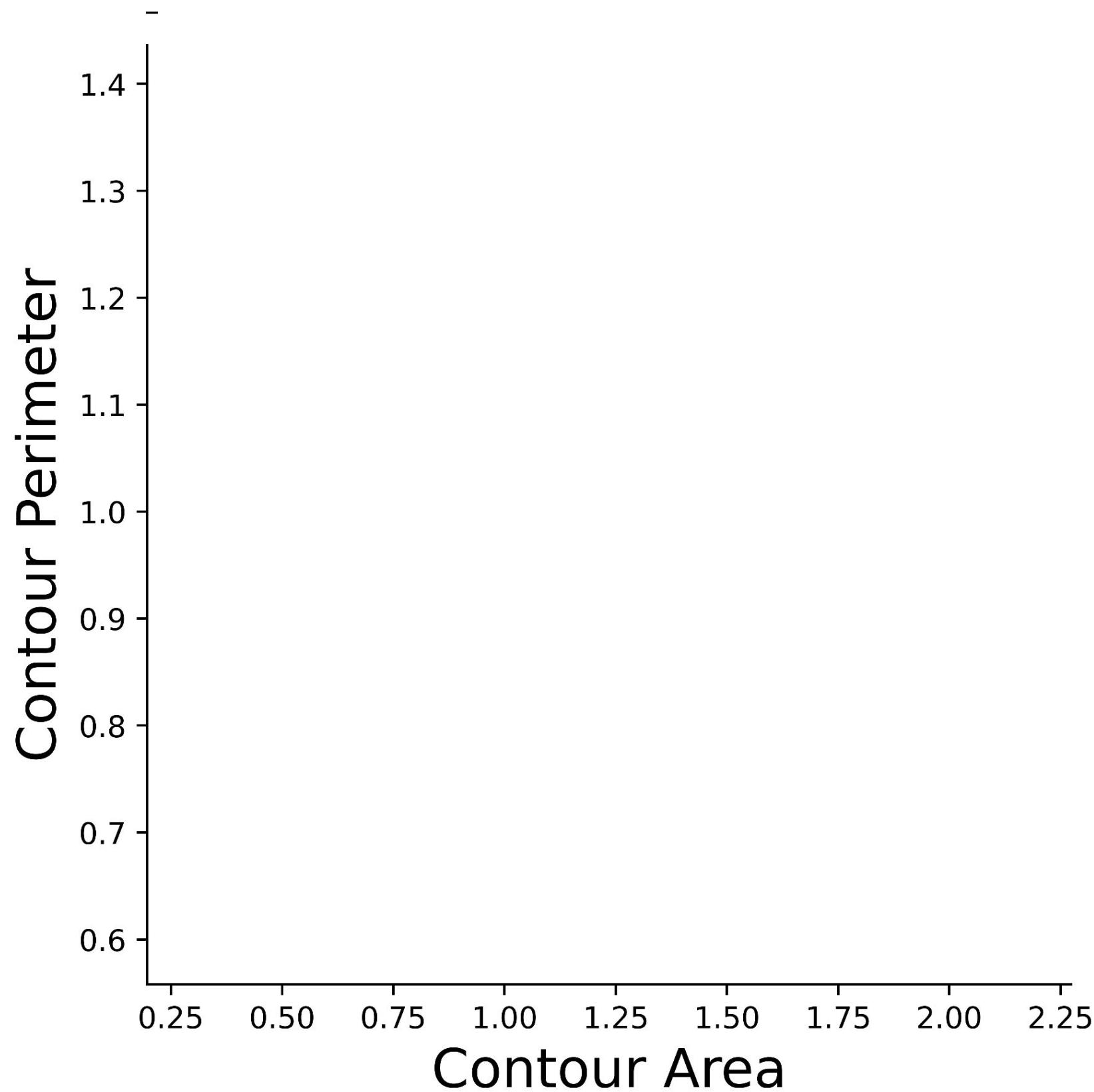
**Instances of under-  
and over-  
segmentation**

**Different boundary  
jaggedness across  
segmentations**

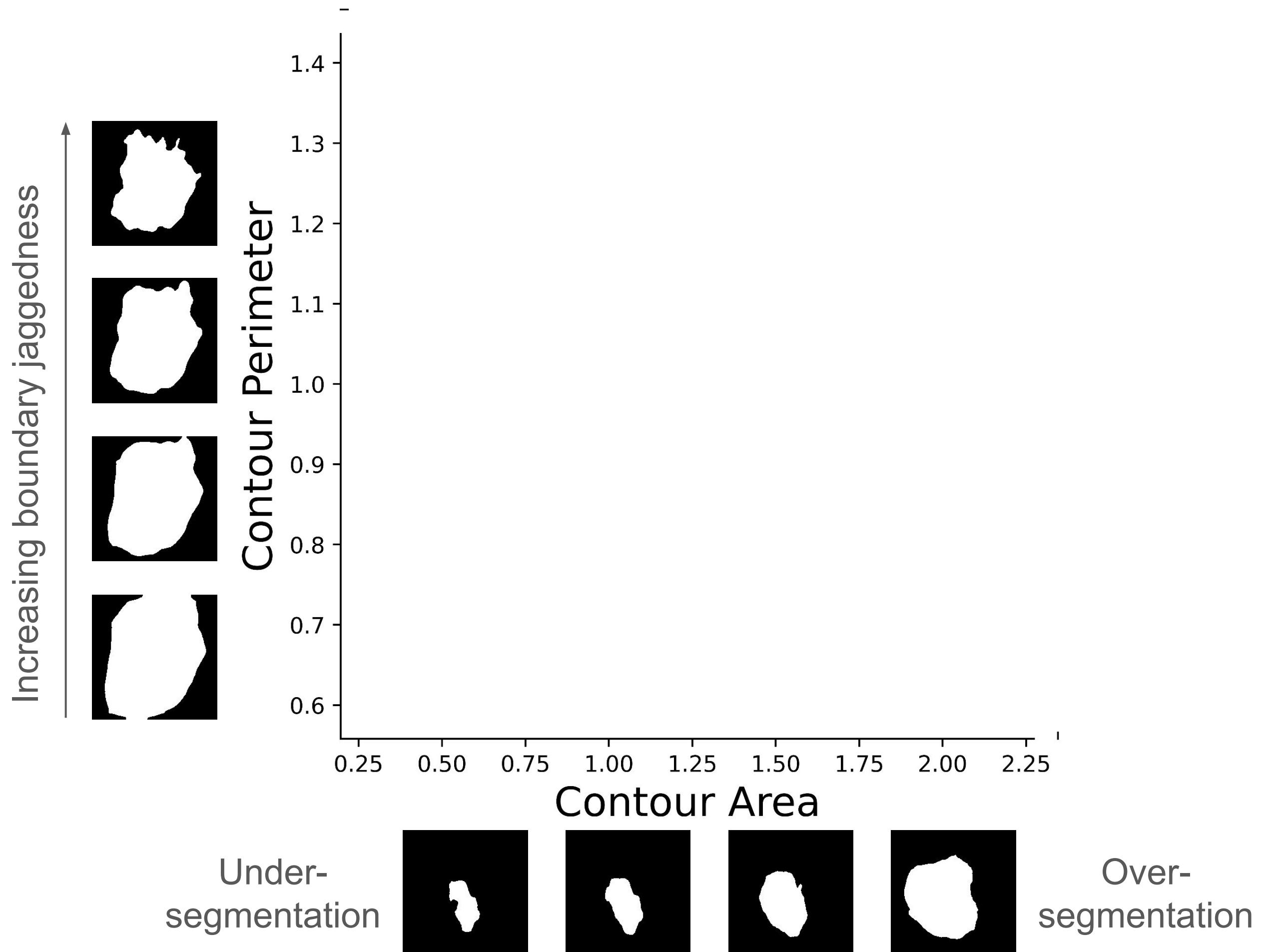
**Ambiguous boundary  
causes segmentation  
masks to split**

# Semantic Consistency of Styles

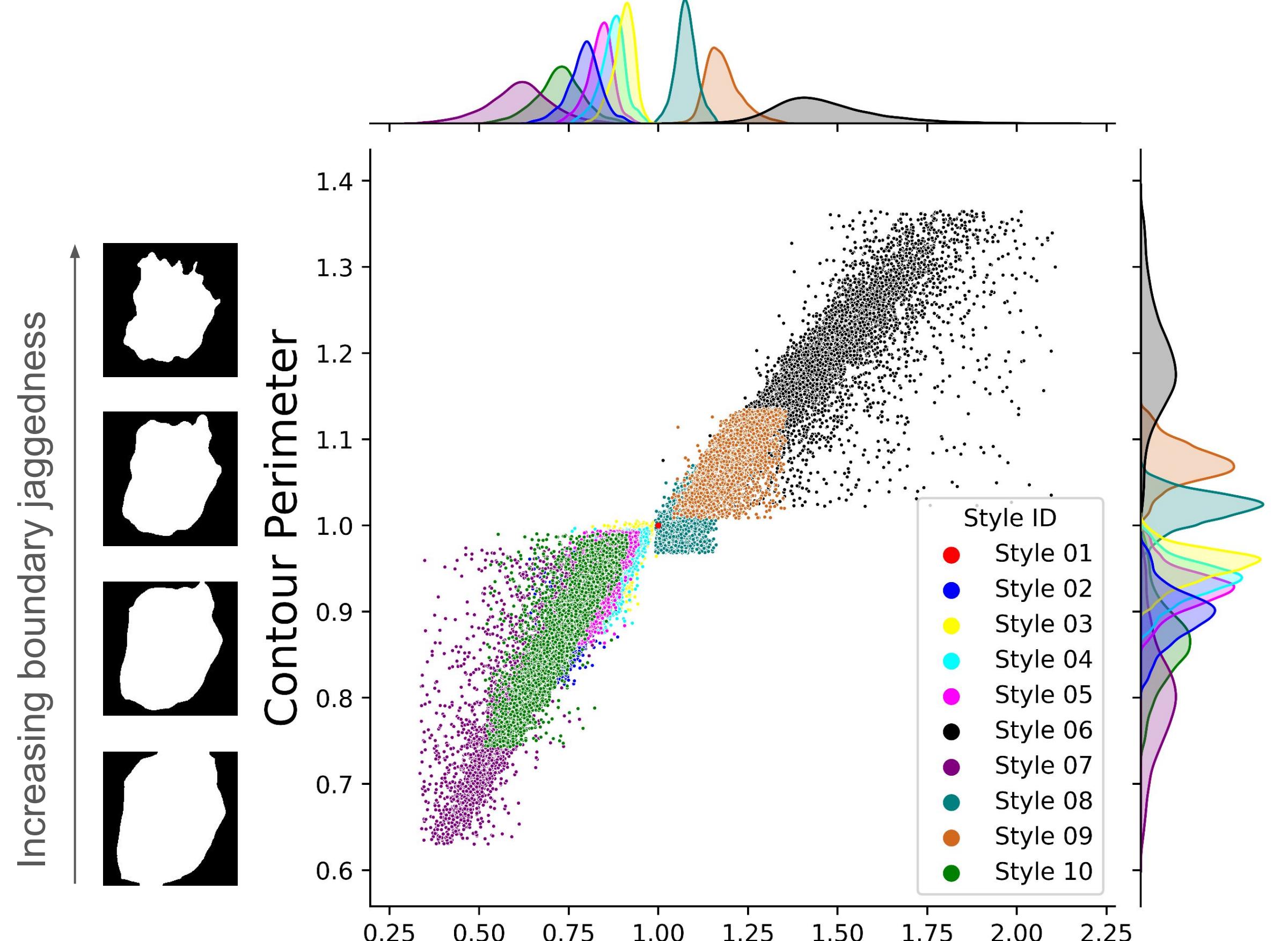
# Semantic Consistency of Styles



# Semantic Consistency of Styles



# Semantic Consistency of Styles



Under-  
segmentation



Over-  
segmentation

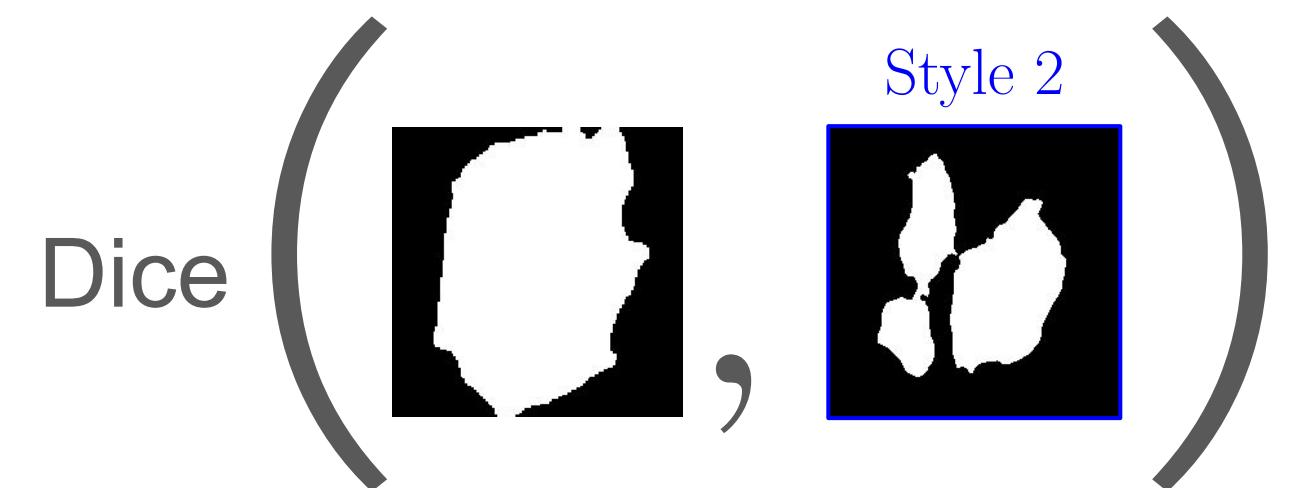
# Competing Methods

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

# Competing Methods

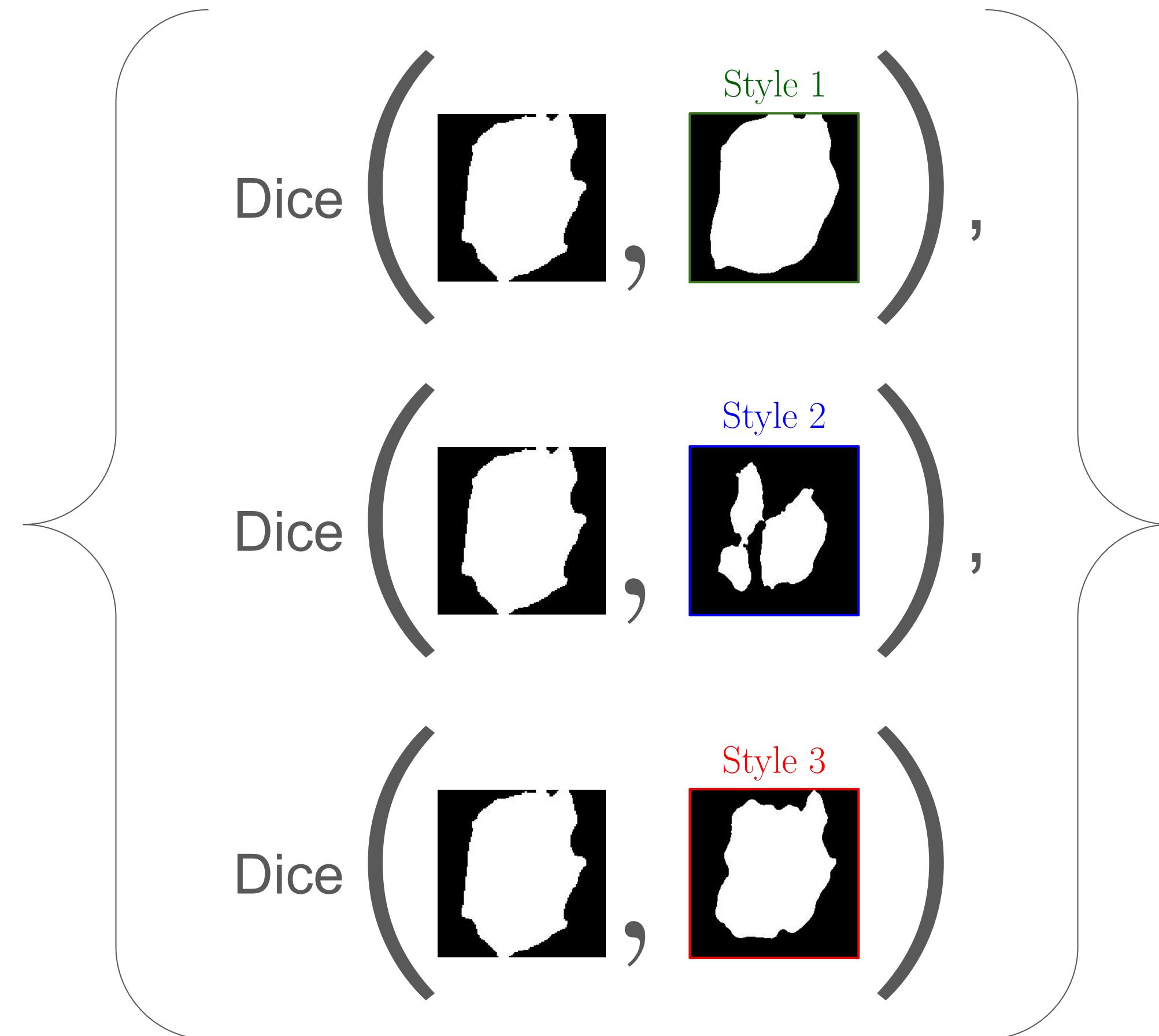
SSeg methods	
<b>NaiveTraining</b>	SLS model <u>without any annotator-specific knowledge</u> .
<b>RandAnnotID</b> <sup>[2]</sup>	4 SLS models, one optimized for <u>each annotator randomly assigned to a mask</u> .
<b>LessIsMore</b> <sup>[3]</sup>	SLS model <u>trained on a subset of the masks</u> whose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

# Metrics



# Metrics

- min.
- max.



# Quantitative Results

Method	<b>ISIC Archive-Test</b> ( $n = 10,000$ )		<b>DermoFit</b> ( $n = 1,300$ )	
	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** ( $n = 1300$ )
- **PH<sup>2</sup>** ( $n = 200$ )
- **SCD** ( $n = 206$ )

# Learning Multiple Styles Is Always Better

Method	<b>ISIC Archive-Test</b> ( $n = 10,000$ )		<b>DermoFit</b> ( $n = 1,300$ )	
	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	<b>ISIC Archive-Test</b> ( $n = 10,000$ )		<b>DermoFit</b> ( $n = 1,300$ )	
	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	<b>ISIC Archive-Test</b> ( $n = 10,000$ )		<b>DermoFit</b> ( $n = 1,300$ )	
	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	<b>ISIC Archive-Test</b> ( $n = 10,000$ )		<b>DermoFit</b> ( $n = 1,300$ )	
	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	<b>ISIC Archive-Test</b> ( $n = 10,000$ )		<b>DermoFit</b> ( $n = 1,300$ )	
	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.

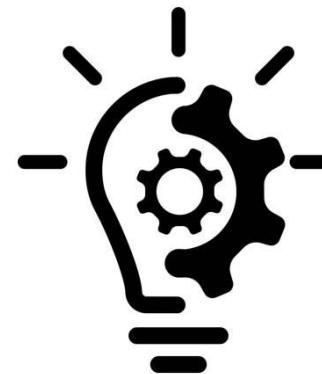
# A New Multi-Annotator SLS Dataset: ISIC-MultiAnnot

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

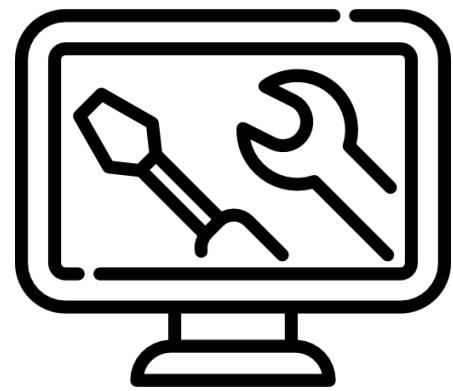


12,951 images

10 anonymized  
annotators  
“A00” – “A09”



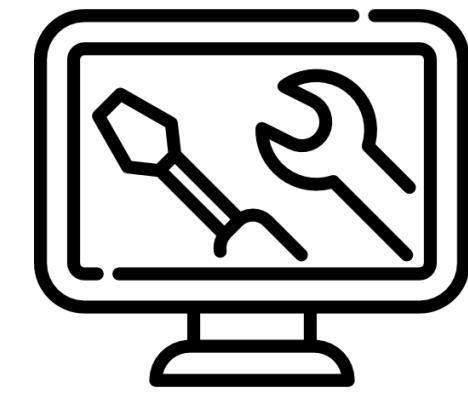
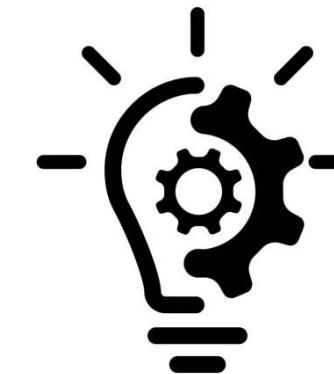
2 skill levels  
“expert”, “novice”



3 tool choices  
“T1” – “T3”

# A New Multi-Annotator SLS Dataset: ISIC-MultiAnnot

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



12,951 images

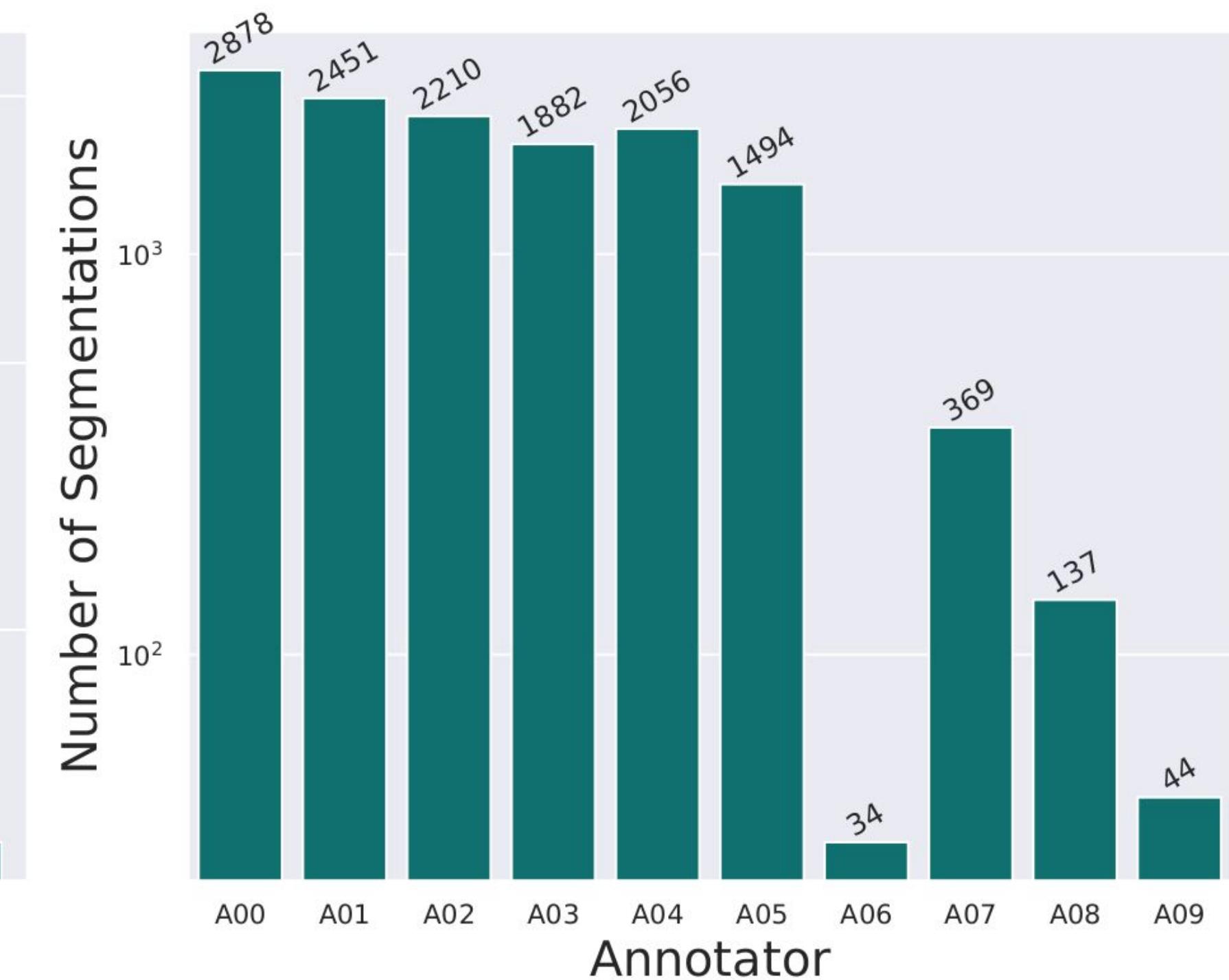
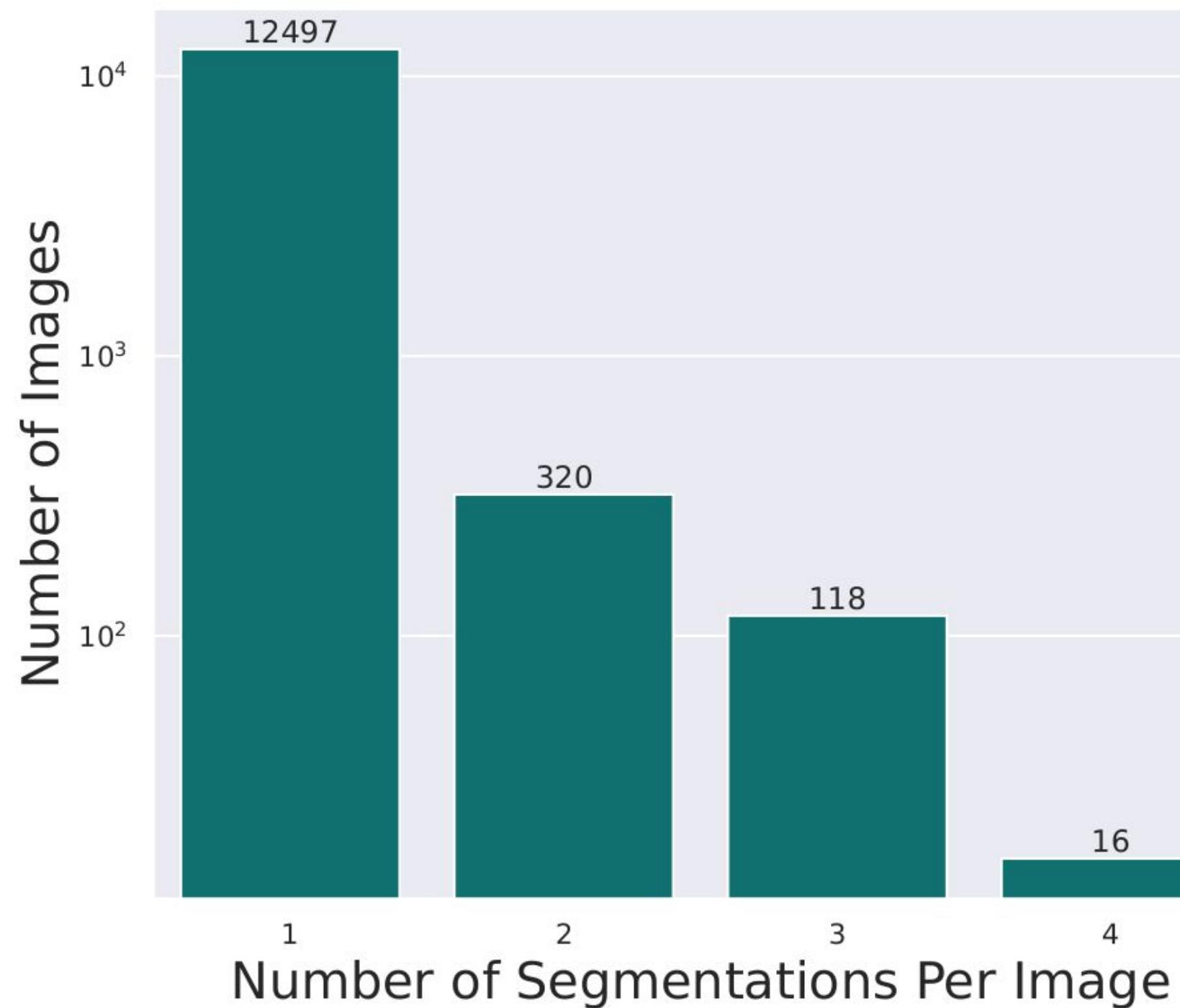
10 anonymized  
annotators  
“A00” – “A09”

2 skill levels  
“expert”, “novice”

3 tool choices  
“T1” – “T3”

13,555 image-mask  
pairs  
**27 unique annotator  
preferences**

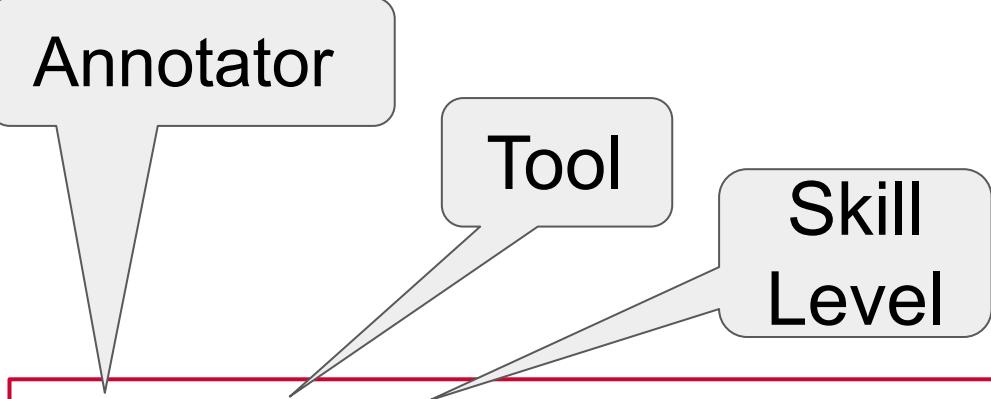
# A New Multi-Annotator SLS Dataset: ISIC-MultiAnnot



# Quantitative Results on ISIC-MultiAnnot

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

# Quantitative Results on ISIC-MultiAnnot



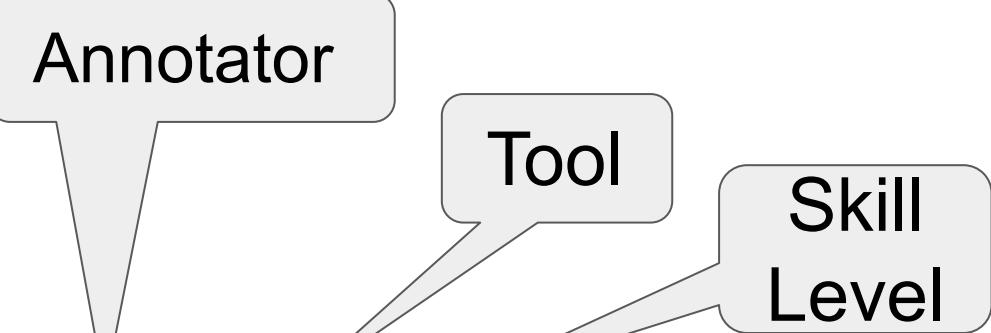
A04+T3+N	90
A05+T1+E	752
A05+T3+E	742
A06+T1+E	10

# image-mask pairs

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

# Quantitative Results on ISIC-MultiAnnot

4-StyleSeg



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

# image-mask  
pairs

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

Annotator + Tool	Sqr. Count	1-StyleSeg		2-StyleSeg		3-StyleSeg		4-StyleSeg	
		DICOISS	J	DICOISS	J	DICOISS	J	DICOISS	J
A00-T2+E	1373	0.892 <sub>±0.009</sub>	0.923 <sub>±0.01</sub>	0.913 <sub>±0.01</sub>	2	0.944 <sub>±0.01</sub>	0.913 <sub>±0.01</sub>	0.944 <sub>±0.01</sub>	1
A00-T2+N	1305	0.716 <sub>±0.009</sub>	0.761 <sub>±0.01</sub>	0.728 <sub>±0.01</sub>	2	0.728 <sub>±0.01</sub>	0.727 <sub>±0.01</sub>	0.700 <sub>±0.01</sub>	3
A01-T2+E	5	0.716 <sub>±0.009</sub>	0.761 <sub>±0.01</sub>	0.728 <sub>±0.01</sub>	2	0.728 <sub>±0.01</sub>	0.727 <sub>±0.01</sub>	0.700 <sub>±0.01</sub>	2
A01-T3+E	297	0.900 <sub>±0.01</sub>	0.915 <sub>±0.01</sub>	0.897 <sub>±0.01</sub>	2	0.927 <sub>±0.01</sub>	0.900 <sub>±0.01</sub>	0.911 <sub>±0.01</sub>	3
A01-T3+N	2149	0.829 <sub>±0.01</sub>	0.857 <sub>±0.01</sub>	0.817 <sub>±0.01</sub>	1	0.869 <sub>±0.01</sub>	0.836 <sub>±0.01</sub>	0.876 <sub>±0.01</sub>	3
A02-T2+E	1742	0.844 <sub>±0.01</sub>	0.886 <sub>±0.01</sub>	0.889 <sub>±0.01</sub>	1	0.886 <sub>±0.01</sub>	0.854 <sub>±0.01</sub>	0.885 <sub>±0.01</sub>	4
A02-T3+E	468	0.856 <sub>±0.01</sub>	0.889 <sub>±0.01</sub>	0.883 <sub>±0.01</sub>	2	0.899 <sub>±0.01</sub>	0.874 <sub>±0.01</sub>	0.903 <sub>±0.01</sub>	3
A03-T2+E	1622	0.778 <sub>±0.01</sub>	0.845 <sub>±0.01</sub>	0.827 <sub>±0.01</sub>	1	0.854 <sub>±0.01</sub>	0.824 <sub>±0.01</sub>	0.881 <sub>±0.01</sub>	4
A03-T3+E	309	0.811 <sub>±0.01</sub>	0.863 <sub>±0.01</sub>	0.845 <sub>±0.01</sub>	2	0.923 <sub>±0.01</sub>	0.865 <sub>±0.01</sub>	0.925 <sub>±0.01</sub>	3
A04-T2+E	90	0.850 <sub>±0.01</sub>	0.888 <sub>±0.01</sub>	0.800 <sub>±0.01</sub>	1	0.866 <sub>±0.01</sub>	0.808 <sub>±0.01</sub>	0.862 <sub>±0.01</sub>	3
A04-T1+E	992	0.850 <sub>±0.01</sub>	0.888 <sub>±0.01</sub>	0.800 <sub>±0.01</sub>	1	0.866 <sub>±0.01</sub>	0.808 <sub>±0.01</sub>	0.862 <sub>±0.01</sub>	3
A04-T1+N	61	0.700 <sub>±0.01</sub>	0.830 <sub>±0.01</sub>	0.823 <sub>±0.01</sub>	1	0.837 <sub>±0.01</sub>	0.786 <sub>±0.01</sub>	0.827 <sub>±0.01</sub>	4
A04-T3+E	913	0.912 <sub>±0.01</sub>	0.939 <sub>±0.01</sub>	0.934 <sub>±0.01</sub>	2	0.948 <sub>±0.01</sub>	0.926 <sub>±0.01</sub>	0.951 <sub>±0.01</sub>	3
A04-T3+N	90	0.877 <sub>±0.01</sub>	0.910 <sub>±0.01</sub>	0.905 <sub>±0.01</sub>	2	0.928 <sub>±0.01</sub>	0.908 <sub>±0.01</sub>	0.926 <sub>±0.01</sub>	3
A05-T1+E	752	0.815 <sub>±0.01</sub>	0.862 <sub>±0.01</sub>	0.837 <sub>±0.01</sub>	1	0.873 <sub>±0.01</sub>	0.827 <sub>±0.01</sub>	0.882 <sub>±0.01</sub>	4
A05-T2+E	742	0.875 <sub>±0.01</sub>	0.905 <sub>±0.01</sub>	0.880 <sub>±0.01</sub>	2	0.938 <sub>±0.01</sub>	0.889 <sub>±0.01</sub>	0.919 <sub>±0.01</sub>	3
A06-T1+E	10	0.750 <sub>±0.01</sub>	0.888 <sub>±0.01</sub>	0.860 <sub>±0.01</sub>	2	0.900 <sub>±0.01</sub>	0.889 <sub>±0.01</sub>	0.888 <sub>±0.01</sub>	4
A06-T3+E	24	0.862 <sub>±0.01</sub>	0.910 <sub>±0.01</sub>	0.934 <sub>±0.01</sub>	2	0.934 <sub>±0.01</sub>	0.923 <sub>±0.01</sub>	0.933 <sub>±0.01</sub>	3
A07-T1+E	67	0.820 <sub>±0.01</sub>	0.877 <sub>±0.01</sub>	0.867 <sub>±0.01</sub>	1	0.890 <sub>±0.01</sub>	0.862 <sub>±0.01</sub>	0.897 <sub>±0.01</sub>	4
A07-T1+N	251	0.877 <sub>±0.01</sub>	0.892 <sub>±0.01</sub>	0.875 <sub>±0.01</sub>	1	0.903 <sub>±0.01</sub>	0.875 <sub>±0.01</sub>	0.905 <sub>±0.01</sub>	4
A07-T3+E	12	0.925 <sub>±0.01</sub>	0.939 <sub>±0.01</sub>	0.937 <sub>±0.01</sub>	2	0.939 <sub>±0.01</sub>	0.916 <sub>±0.01</sub>	0.947 <sub>±0.01</sub>	3
A07-T3+N	39	0.863 <sub>±0.01</sub>	0.890 <sub>±0.01</sub>	0.913 <sub>±0.01</sub>	2	0.933 <sub>±0.01</sub>	0.899 <sub>±0.01</sub>	0.934 <sub>±0.01</sub>	3
A08-T1+E	56	0.905 <sub>±0.01</sub>	0.925 <sub>±0.01</sub>	0.924 <sub>±0.01</sub>	2	0.947 <sub>±0.01</sub>	0.905 <sub>±0.01</sub>	0.950 <sub>±0.01</sub>	1
A08-T3+E	111	0.600 <sub>±0.01</sub>	0.660 <sub>±0.01</sub>	0.630 <sub>±0.01</sub>	1	0.628 <sub>±0.01</sub>	0.628 <sub>±0.01</sub>	0.700 <sub>±0.01</sub>	2
A09-T1+E	30	0.815 <sub>±0.01</sub>	0.841 <sub>±0.01</sub>	0.784 <sub>±0.01</sub>	2	0.873 <sub>±0.01</sub>	0.833 <sub>±0.01</sub>	0.884 <sub>±0.01</sub>	4
A09-T1+N	1	0.955 <sub>±0.01</sub>	0.927 <sub>±0.01</sub>	0.955 <sub>±0.01</sub>	2	0.955 <sub>±0.01</sub>	0.955 <sub>±0.01</sub>	0.947 <sub>±0.01</sub>	3
A09-T3+E	10	0.900 <sub>±0.01</sub>	0.918 <sub>±0.01</sub>	0.918 <sub>±0.01</sub>	2	0.933 <sub>±0.01</sub>	0.909 <sub>±0.01</sub>	0.937 <sub>±0.01</sub>	3
A09-T3+N	3	0.894 <sub>±0.01</sub>	0.913 <sub>±0.01</sub>	0.910 <sub>±0.01</sub>	2	0.957 <sub>±0.01</sub>	0.957 <sub>±0.01</sub>	0.944 <sub>±0.01</sub>	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 + Experience	Seg. Count	1-StyleSeg		2-StyleSeg		3-StyleSeg		4-StyleSeg	
		DICESS	J	DICESS	J	DICESS	J	DICESS	J
A00-T2+E	1373	0.892 <sub>±0.009</sub>	0.923 <sub>±0.01</sub>	0.913 <sub>±0.01</sub>	2	0.944 <sub>±0.01</sub>	0.899 <sub>±0.01</sub>	0.914 <sub>±0.01</sub>	1
A00-T2+N	1305	0.716 <sub>±0.012</sub>	0.761 <sub>±0.01</sub>	0.728 <sub>±0.01</sub>	2	0.728 <sub>±0.01</sub>	0.727 <sub>±0.01</sub>	0.700 <sub>±0.01</sub>	3
A01-T2+E	5	0.745 <sub>±0.01</sub>	0.755 <sub>±0.01</sub>	0.745 <sub>±0.01</sub>	2	0.745 <sub>±0.01</sub>	0.755 <sub>±0.01</sub>	0.745 <sub>±0.01</sub>	2
A01-T3+N	297	0.900 <sub>±0.01</sub>	0.915 <sub>±0.01</sub>	0.897 <sub>±0.01</sub>	2	0.927 <sub>±0.01</sub>	0.900 <sub>±0.01</sub>	0.911 <sub>±0.01</sub>	3
A01-T3+E	2149	0.829 <sub>±0.01</sub>	0.857 <sub>±0.01</sub>	0.817 <sub>±0.01</sub>	1	0.869 <sub>±0.01</sub>	0.836 <sub>±0.01</sub>	0.876 <sub>±0.01</sub>	3
A02-T2+E	1742	0.844 <sub>±0.01</sub>	0.889 <sub>±0.01</sub>	0.886 <sub>±0.01</sub>	1	0.886 <sub>±0.01</sub>	0.854 <sub>±0.01</sub>	0.885 <sub>±0.01</sub>	4
A02-T3+E	468	0.856 <sub>±0.01</sub>	0.889 <sub>±0.01</sub>	0.883 <sub>±0.01</sub>	2	0.899 <sub>±0.01</sub>	0.874 <sub>±0.01</sub>	0.903 <sub>±0.01</sub>	3
A03-T2+E	1622	0.778 <sub>±0.01</sub>	0.845 <sub>±0.01</sub>	0.827 <sub>±0.01</sub>	1	0.854 <sub>±0.01</sub>	0.824 <sub>±0.01</sub>	0.881 <sub>±0.01</sub>	4
A03-T3+E	390	0.811 <sub>±0.01</sub>	0.864 <sub>±0.01</sub>	0.843 <sub>±0.01</sub>	2	0.923 <sub>±0.01</sub>	0.865 <sub>±0.01</sub>	0.925 <sub>±0.01</sub>	3
A04-T2+E	992	0.845 <sub>±0.01</sub>	0.888 <sub>±0.01</sub>	0.800 <sub>±0.01</sub>	1	0.866 <sub>±0.01</sub>	0.826 <sub>±0.01</sub>	0.865 <sub>±0.01</sub>	4
A04-T1+E	61	0.760 <sub>±0.01</sub>	0.830 <sub>±0.01</sub>	0.823 <sub>±0.01</sub>	1	0.827 <sub>±0.01</sub>	0.796 <sub>±0.01</sub>	0.820 <sub>±0.01</sub>	3
A04-T1+N	913	0.912 <sub>±0.01</sub>	0.939 <sub>±0.01</sub>	0.934 <sub>±0.01</sub>	2	0.948 <sub>±0.01</sub>	0.926 <sub>±0.01</sub>	0.951 <sub>±0.01</sub>	3
A04-T3+N	90	0.877 <sub>±0.01</sub>	0.910 <sub>±0.01</sub>	0.905 <sub>±0.01</sub>	2	0.928 <sub>±0.01</sub>	0.908 <sub>±0.01</sub>	0.926 <sub>±0.01</sub>	1
A05-T1+E	752	0.815 <sub>±0.01</sub>	0.862 <sub>±0.01</sub>	0.837 <sub>±0.01</sub>	1	0.873 <sub>±0.01</sub>	0.827 <sub>±0.01</sub>	0.882 <sub>±0.01</sub>	4
A05-T2+E	742	0.875 <sub>±0.01</sub>	0.905 <sub>±0.01</sub>	0.883 <sub>±0.01</sub>	2	0.930 <sub>±0.01</sub>	0.889 <sub>±0.01</sub>	0.919 <sub>±0.01</sub>	3
A06-T1+E	10	0.753 <sub>±0.01</sub>	0.888 <sub>±0.01</sub>	0.860 <sub>±0.01</sub>	2	0.900 <sub>±0.01</sub>	0.889 <sub>±0.01</sub>	0.901 <sub>±0.01</sub>	4
A06-T3+E	24	0.862 <sub>±0.01</sub>	0.916 <sub>±0.01</sub>	0.934 <sub>±0.01</sub>	2	0.934 <sub>±0.01</sub>	0.923 <sub>±0.01</sub>	0.933 <sub>±0.01</sub>	3
A07-T1+E	67	0.828 <sub>±0.01</sub>	0.877 <sub>±0.01</sub>	0.867 <sub>±0.01</sub>	1	0.890 <sub>±0.01</sub>	0.862 <sub>±0.01</sub>	0.897 <sub>±0.01</sub>	4
A07-T1+N	251	0.837 <sub>±0.01</sub>	0.892 <sub>±0.01</sub>	0.879 <sub>±0.01</sub>	1	0.903 <sub>±0.01</sub>	0.875 <sub>±0.01</sub>	0.873 <sub>±0.01</sub>	4
A07-T3+E	12	0.925 <sub>±0.01</sub>	0.939 <sub>±0.01</sub>	0.937 <sub>±0.01</sub>	2	0.939 <sub>±0.01</sub>	0.916 <sub>±0.01</sub>	0.947 <sub>±0.01</sub>	3
A07-T3+N	39	0.863 <sub>±0.01</sub>	0.910 <sub>±0.01</sub>	0.913 <sub>±0.01</sub>	2	0.933 <sub>±0.01</sub>	0.899 <sub>±0.01</sub>	0.934 <sub>±0.01</sub>	1
A08-T1+E	56	0.905 <sub>±0.01</sub>	0.929 <sub>±0.01</sub>	0.917 <sub>±0.01</sub>	2	0.947 <sub>±0.01</sub>	0.905 <sub>±0.01</sub>	0.960 <sub>±0.01</sub>	1
A08-T3+E	111	0.607 <sub>±0.01</sub>	0.665 <sub>±0.01</sub>	0.630 <sub>±0.01</sub>	1	0.628 <sub>±0.01</sub>	0.658 <sub>±0.01</sub>	0.774 <sub>±0.01</sub>	2
A09-T1+E	30	0.841 <sub>±0.01</sub>	0.841 <sub>±0.01</sub>	0.784 <sub>±0.01</sub>	2	0.873 <sub>±0.01</sub>	0.853 <sub>±0.01</sub>	0.884 <sub>±0.01</sub>	3
A09-T1+N	1	0.953 <sub>±0.01</sub>	0.927 <sub>±0.01</sub>	0.955 <sub>±0.01</sub>	1	0.974 <sub>±0.01</sub>	0.955 <sub>±0.01</sub>	0.947 <sub>±0.01</sub>	3
A09-T3+E	10	0.900 <sub>±0.01</sub>	0.918 <sub>±0.01</sub>	0.918 <sub>±0.01</sub>	2	0.933 <sub>±0.01</sub>	0.909 <sub>±0.01</sub>	0.937 <sub>±0.01</sub>	3
A09-T3+N	3	0.894 <sub>±0.01</sub>	0.913 <sub>±0.01</sub>	0.910 <sub>±0.01</sub>	2	0.957 <sub>±0.01</sub>	0.957 <sub>±0.01</sub>	0.944 <sub>±0.01</sub>	1

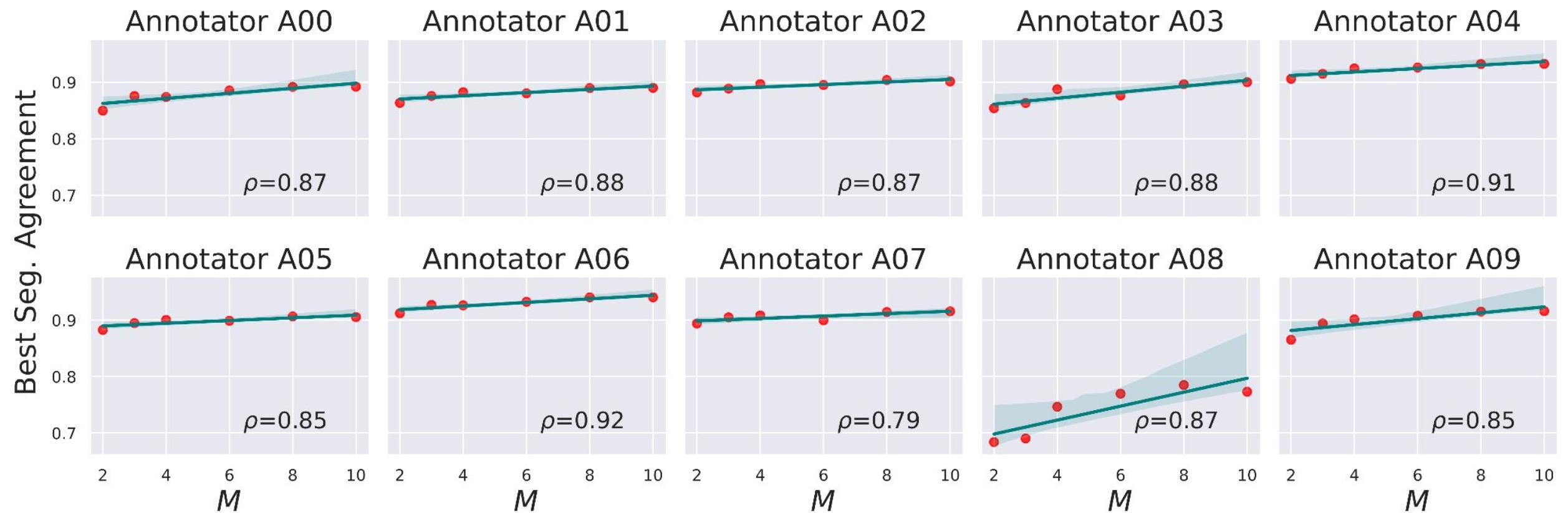
# ISIC-MultiAnnot Results: Key Takeaways

- 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	Sqr. Count	1-StyleSeg		2-StyleSeg		3-StyleSeg		4-StyleSeg	
		DICOISS	J	DICOISS	J	DICOISS	J	DICOISS	J
A00-T2+E	1373	0.892 <sub>±0.009</sub>	0.923 <sub>±0.01</sub>	0.913 <sub>±0.009</sub>	2	0.944 <sub>±0.019</sub>	0.913 <sub>±0.009</sub>	0.914 <sub>±0.014</sub>	1
A00-T2+N	1305	0.716 <sub>±0.009</sub>	0.761 <sub>±0.007</sub>	0.728 <sub>±0.008</sub>	2	0.728 <sub>±0.019</sub>	0.727 <sub>±0.019</sub>	0.700 <sub>±0.019</sub>	3
A01-T2+E	5	0.715 <sub>±0.009</sub>	0.750 <sub>±0.009</sub>	0.722 <sub>±0.009</sub>	2	0.750 <sub>±0.019</sub>	0.750 <sub>±0.019</sub>	0.765 <sub>±0.019</sub>	2
A01-T3+E	297	0.900 <sub>±0.009</sub>	0.915 <sub>±0.009</sub>	0.897 <sub>±0.009</sub>	2	0.927 <sub>±0.019</sub>	0.900 <sub>±0.009</sub>	0.911 <sub>±0.009</sub>	3
A01-T3+N	2149	0.829 <sub>±0.019</sub>	0.875 <sub>±0.019</sub>	0.817 <sub>±0.019</sub>	1	0.869 <sub>±0.019</sub>	0.836 <sub>±0.019</sub>	0.876 <sub>±0.019</sub>	3
A02-T1+E	1742	0.844 <sub>±0.019</sub>	0.889 <sub>±0.019</sub>	0.856 <sub>±0.019</sub>	1	0.886 <sub>±0.019</sub>	0.854 <sub>±0.019</sub>	0.895 <sub>±0.019</sub>	4
A02-T3+E	468	0.856 <sub>±0.019</sub>	0.889 <sub>±0.019</sub>	0.883 <sub>±0.019</sub>	2	0.899 <sub>±0.019</sub>	0.874 <sub>±0.019</sub>	0.903 <sub>±0.019</sub>	1
A03-T1+E	1622	0.778 <sub>±0.019</sub>	0.845 <sub>±0.019</sub>	0.827 <sub>±0.019</sub>	1	0.854 <sub>±0.019</sub>	0.824 <sub>±0.019</sub>	0.881 <sub>±0.019</sub>	4
A03-T2+E	309	0.811 <sub>±0.019</sub>	0.863 <sub>±0.019</sub>	0.842 <sub>±0.019</sub>	2	0.923 <sub>±0.019</sub>	0.865 <sub>±0.019</sub>	0.925 <sub>±0.019</sub>	3
A04-T1+E	992	0.845 <sub>±0.019</sub>	0.888 <sub>±0.019</sub>	0.800 <sub>±0.019</sub>	1	0.866 <sub>±0.019</sub>	0.826 <sub>±0.019</sub>	0.865 <sub>±0.019</sub>	4
A04-T1+N	61	0.700 <sub>±0.019</sub>	0.830 <sub>±0.019</sub>	0.823 <sub>±0.019</sub>	1	0.827 <sub>±0.019</sub>	0.796 <sub>±0.019</sub>	0.820 <sub>±0.019</sub>	3
A04-T3+E	913	0.912 <sub>±0.009</sub>	0.939 <sub>±0.009</sub>	0.934 <sub>±0.009</sub>	2	0.948 <sub>±0.019</sub>	0.926 <sub>±0.009</sub>	0.951 <sub>±0.019</sub>	3
A04-T3+N	90	0.877 <sub>±0.009</sub>	0.910 <sub>±0.009</sub>	0.905 <sub>±0.009</sub>	2	0.928 <sub>±0.019</sub>	0.908 <sub>±0.019</sub>	0.926 <sub>±0.019</sub>	1
A05-T1+E	752	0.815 <sub>±0.019</sub>	0.862 <sub>±0.019</sub>	0.837 <sub>±0.019</sub>	1	0.873 <sub>±0.019</sub>	0.827 <sub>±0.019</sub>	0.882 <sub>±0.019</sub>	4
A05-T2+E	742	0.875 <sub>±0.019</sub>	0.905 <sub>±0.019</sub>	0.890 <sub>±0.019</sub>	2	0.930 <sub>±0.019</sub>	0.889 <sub>±0.019</sub>	0.919 <sub>±0.019</sub>	1
A06-T1+E	10	0.741 <sub>±0.019</sub>	0.888 <sub>±0.019</sub>	0.860 <sub>±0.019</sub>	2	0.900 <sub>±0.019</sub>	0.889 <sub>±0.019</sub>	0.901 <sub>±0.019</sub>	1
A06-T3+E	24	0.862 <sub>±0.019</sub>	0.910 <sub>±0.019</sub>	0.934 <sub>±0.019</sub>	2	0.934 <sub>±0.019</sub>	0.923 <sub>±0.019</sub>	0.933 <sub>±0.019</sub>	3
A07-T1+E	67	0.829 <sub>±0.019</sub>	0.877 <sub>±0.019</sub>	0.867 <sub>±0.019</sub>	1	0.890 <sub>±0.019</sub>	0.862 <sub>±0.019</sub>	0.897 <sub>±0.019</sub>	4
A07-T1+N	251	0.837 <sub>±0.019</sub>	0.892 <sub>±0.019</sub>	0.879 <sub>±0.019</sub>	1	0.903 <sub>±0.019</sub>	0.875 <sub>±0.019</sub>	0.905 <sub>±0.019</sub>	4
A07-T3+E	12	0.925 <sub>±0.019</sub>	0.939 <sub>±0.019</sub>	0.937 <sub>±0.019</sub>	2	0.939 <sub>±0.019</sub>	0.916 <sub>±0.019</sub>	0.947 <sub>±0.019</sub>	3
A07-T3+N	39	0.863 <sub>±0.019</sub>	0.909 <sub>±0.019</sub>	0.913 <sub>±0.019</sub>	2	0.933 <sub>±0.019</sub>	0.899 <sub>±0.019</sub>	0.934 <sub>±0.019</sub>	1
A08-T1+E	56	0.905 <sub>±0.019</sub>	0.934 <sub>±0.019</sub>	0.924 <sub>±0.019</sub>	2	0.947 <sub>±0.019</sub>	0.905 <sub>±0.019</sub>	0.950 <sub>±0.019</sub>	1
A08-T3+E	111	0.607 <sub>±0.019</sub>	0.665 <sub>±0.019</sub>	0.630 <sub>±0.019</sub>	2	0.628 <sub>±0.019</sub>	0.628 <sub>±0.019</sub>	0.707 <sub>±0.019</sub>	2
A09-T1+E	30	0.815 <sub>±0.019</sub>	0.841 <sub>±0.019</sub>	0.784 <sub>±0.019</sub>	1	0.873 <sub>±0.019</sub>	0.833 <sub>±0.019</sub>	0.884 <sub>±0.019</sub>	4
A09-T1+N	1	0.953 <sub>±0.009</sub>	0.927 <sub>±0.009</sub>	0.955 <sub>±0.009</sub>	1	0.947 <sub>±0.009</sub>	0.917 <sub>±0.009</sub>	0.947 <sub>±0.009</sub>	3
A09-T3+E	10	0.900 <sub>±0.019</sub>	0.918 <sub>±0.019</sub>	0.918 <sub>±0.019</sub>	2	0.933 <sub>±0.019</sub>	0.909 <sub>±0.019</sub>	0.937 <sub>±0.019</sub>	3
A09-T3+N	3	0.894 <sub>±0.019</sub>	0.913 <sub>±0.019</sub>	0.910 <sub>±0.019</sub>	2	0.957 <sub>±0.019</sub>	0.957 <sub>±0.019</sub>	0.944 <sub>±0.019</sub>	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.
2. Performance improves as  $M$  increases.



# 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”).

Annotator + Tool	Sqr. Count	1-StyleSeg		2-StyleSeg		3-StyleSeg		4-StyleSeg	
		DICOISS	J	DICOISS	J	DICOISS	J	DICOISS	J
A00-T2+E	1373	0.892 <sub>±0.009</sub>	0.923 <sub>±0.01</sub>	0.913 <sub>±0.01</sub>	2	0.944 <sub>±0.01</sub>	0.899 <sub>±0.01</sub>	0.913 <sub>±0.01</sub>	1
A00-T2+N	1305	0.716 <sub>±0.012</sub>	0.761 <sub>±0.012</sub>	0.728 <sub>±0.012</sub>	2	0.738 <sub>±0.012</sub>	0.727 <sub>±0.012</sub>	0.700 <sub>±0.012</sub>	3
A00-T3+E	5	0.745 <sub>±0.01</sub>	0.751 <sub>±0.01</sub>	0.745 <sub>±0.01</sub>	2	0.745 <sub>±0.01</sub>	0.745 <sub>±0.01</sub>	0.745 <sub>±0.01</sub>	2
A00-T3+N	297	0.900 <sub>±0.01</sub>	0.915 <sub>±0.01</sub>	0.897 <sub>±0.01</sub>	2	0.927 <sub>±0.01</sub>	0.900 <sub>±0.01</sub>	0.911 <sub>±0.01</sub>	3
A01-T3+E	210	0.829 <sub>±0.01</sub>	0.857 <sub>±0.01</sub>	0.817 <sub>±0.01</sub>	1	0.869 <sub>±0.01</sub>	0.836 <sub>±0.01</sub>	0.876 <sub>±0.01</sub>	3
A01-T3+N	1742	0.844 <sub>±0.01</sub>	0.886 <sub>±0.01</sub>	0.886 <sub>±0.01</sub>	1	0.886 <sub>±0.01</sub>	0.854 <sub>±0.01</sub>	0.886 <sub>±0.01</sub>	4
A02-T3+E	468	0.856 <sub>±0.01</sub>	0.889 <sub>±0.01</sub>	0.883 <sub>±0.01</sub>	2	0.899 <sub>±0.01</sub>	0.874 <sub>±0.01</sub>	0.903 <sub>±0.01</sub>	3
A02-T3+N	1622	0.778 <sub>±0.01</sub>	0.845 <sub>±0.01</sub>	0.827 <sub>±0.01</sub>	1	0.854 <sub>±0.01</sub>	0.824 <sub>±0.01</sub>	0.881 <sub>±0.01</sub>	4
A03-T1+E	209	0.815 <sub>±0.01</sub>	0.863 <sub>±0.01</sub>	0.842 <sub>±0.01</sub>	2	0.923 <sub>±0.01</sub>	0.862 <sub>±0.01</sub>	0.925 <sub>±0.01</sub>	3
A03-T1+N	10	0.845 <sub>±0.01</sub>	0.888 <sub>±0.01</sub>	0.800 <sub>±0.01</sub>	1	0.906 <sub>±0.01</sub>	0.868 <sub>±0.01</sub>	0.865 <sub>±0.01</sub>	2
A04-T1+E	992	0.850 <sub>±0.01</sub>	0.888 <sub>±0.01</sub>	0.800 <sub>±0.01</sub>	2	0.923 <sub>±0.01</sub>	0.852 <sub>±0.01</sub>	0.926 <sub>±0.01</sub>	3
A04-T1+N	61	0.760 <sub>±0.01</sub>	0.810 <sub>±0.01</sub>	0.823 <sub>±0.01</sub>	1	0.827 <sub>±0.01</sub>	0.786 <sub>±0.01</sub>	0.729 <sub>±0.01</sub>	1
A04-T3+E	913	0.912 <sub>±0.01</sub>	0.939 <sub>±0.01</sub>	0.934 <sub>±0.01</sub>	2	0.948 <sub>±0.01</sub>	0.926 <sub>±0.01</sub>	0.951 <sub>±0.01</sub>	3
A04-T3+N	90	0.877 <sub>±0.01</sub>	0.910 <sub>±0.01</sub>	0.905 <sub>±0.01</sub>	2	0.928 <sub>±0.01</sub>	0.908 <sub>±0.01</sub>	0.926 <sub>±0.01</sub>	3
A05-T1+E	752	0.815 <sub>±0.01</sub>	0.862 <sub>±0.01</sub>	0.873 <sub>±0.01</sub>	1	0.878 <sub>±0.01</sub>	0.827 <sub>±0.01</sub>	0.882 <sub>±0.01</sub>	4
A05-T1+N	742	0.875 <sub>±0.01</sub>	0.905 <sub>±0.01</sub>	0.890 <sub>±0.01</sub>	2	0.936 <sub>±0.01</sub>	0.889 <sub>±0.01</sub>	0.919 <sub>±0.01</sub>	3
A06-T1+E	10	0.847 <sub>±0.01</sub>	0.888 <sub>±0.01</sub>	0.890 <sub>±0.01</sub>	2	0.900 <sub>±0.01</sub>	0.889 <sub>±0.01</sub>	0.884 <sub>±0.01</sub>	1
A06-T3+E	24	0.862 <sub>±0.01</sub>	0.910 <sub>±0.01</sub>	0.910 <sub>±0.01</sub>	2	0.934 <sub>±0.01</sub>	0.923 <sub>±0.01</sub>	0.933 <sub>±0.01</sub>	3
A07-T1+E	67	0.820 <sub>±0.01</sub>	0.877 <sub>±0.01</sub>	0.867 <sub>±0.01</sub>	1	0.890 <sub>±0.01</sub>	0.862 <sub>±0.01</sub>	0.897 <sub>±0.01</sub>	4
A07-T1+N	251	0.877 <sub>±0.01</sub>	0.892 <sub>±0.01</sub>	0.879 <sub>±0.01</sub>	1	0.903 <sub>±0.01</sub>	0.875 <sub>±0.01</sub>	0.895 <sub>±0.01</sub>	4
A07-T3+E	12	0.925 <sub>±0.01</sub>	0.938 <sub>±0.01</sub>	0.937 <sub>±0.01</sub>	2	0.939 <sub>±0.01</sub>	0.916 <sub>±0.01</sub>	0.947 <sub>±0.01</sub>	3
A07-T3+N	39	0.863 <sub>±0.01</sub>	0.903 <sub>±0.01</sub>	0.913 <sub>±0.01</sub>	2	0.933 <sub>±0.01</sub>	0.899 <sub>±0.01</sub>	0.934 <sub>±0.01</sub>	3
A08-T1+E	56	0.905 <sub>±0.01</sub>	0.929 <sub>±0.01</sub>	0.910 <sub>±0.01</sub>	2	0.947 <sub>±0.01</sub>	0.905 <sub>±0.01</sub>	0.955 <sub>±0.01</sub>	4
A08-T3+E	111	0.607 <sub>±0.01</sub>	0.665 <sub>±0.01</sub>	0.620 <sub>±0.01</sub>	1	0.628 <sub>±0.01</sub>	0.628 <sub>±0.01</sub>	0.707 <sub>±0.01</sub>	2
A09-T1+E	30	0.815 <sub>±0.01</sub>	0.841 <sub>±0.01</sub>	0.784 <sub>±0.01</sub>	2	0.873 <sub>±0.01</sub>	0.955 <sub>±0.01</sub>	0.884 <sub>±0.01</sub>	3
A09-T1+N	1	0.953 <sub>±0.01</sub>	0.927 <sub>±0.01</sub>	0.955 <sub>±0.01</sub>	1	0.947 <sub>±0.01</sub>	0.947 <sub>±0.01</sub>	0.947 <sub>±0.01</sub>	3
A09-T3+E	10	0.900 <sub>±0.01</sub>	0.918 <sub>±0.01</sub>	0.918 <sub>±0.01</sub>	2	0.933 <sub>±0.01</sub>	0.909 <sub>±0.01</sub>	0.937 <sub>±0.01</sub>	3
A09-T3+N	3	0.894 <sub>±0.01</sub>	0.913 <sub>±0.01</sub>	0.910 <sub>±0.01</sub>	2	0.957 <sub>±0.01</sub>	0.957 <sub>±0.01</sub>	0.944 <sub>±0.01</sub>	1

# 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**).

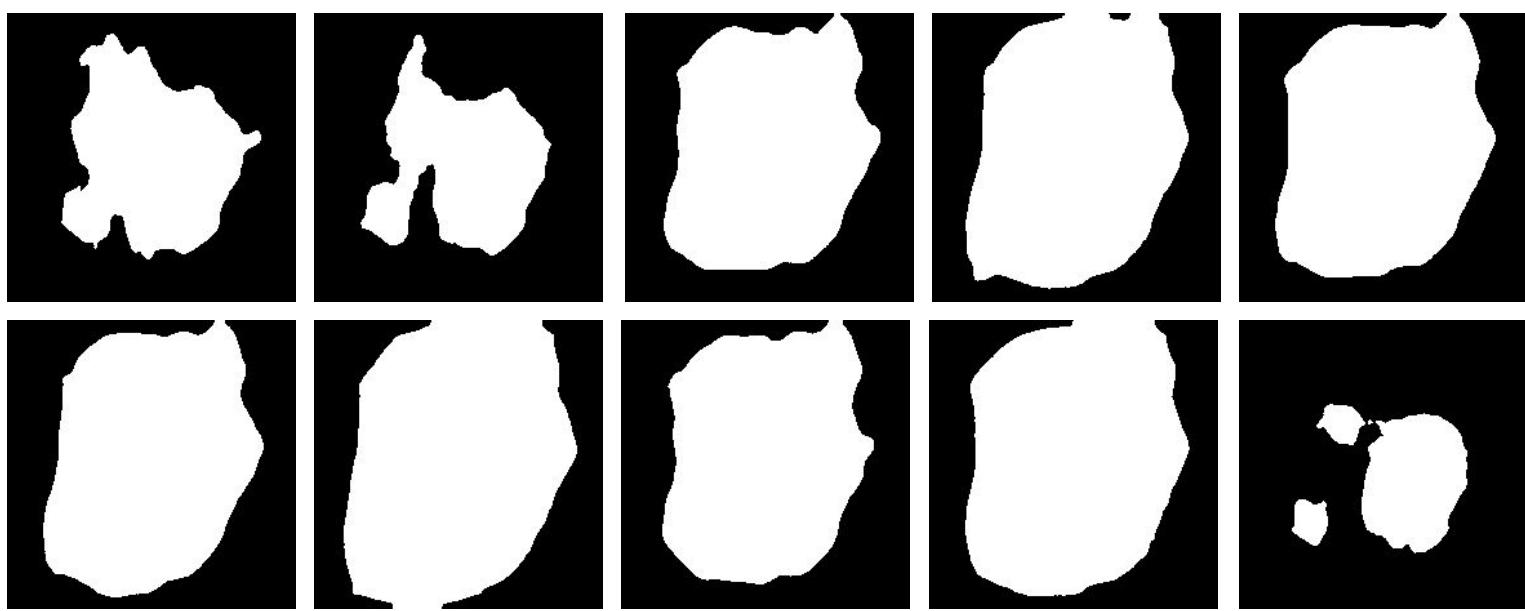
How do we quantify this annotator-style alignment strength?

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

# 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**).



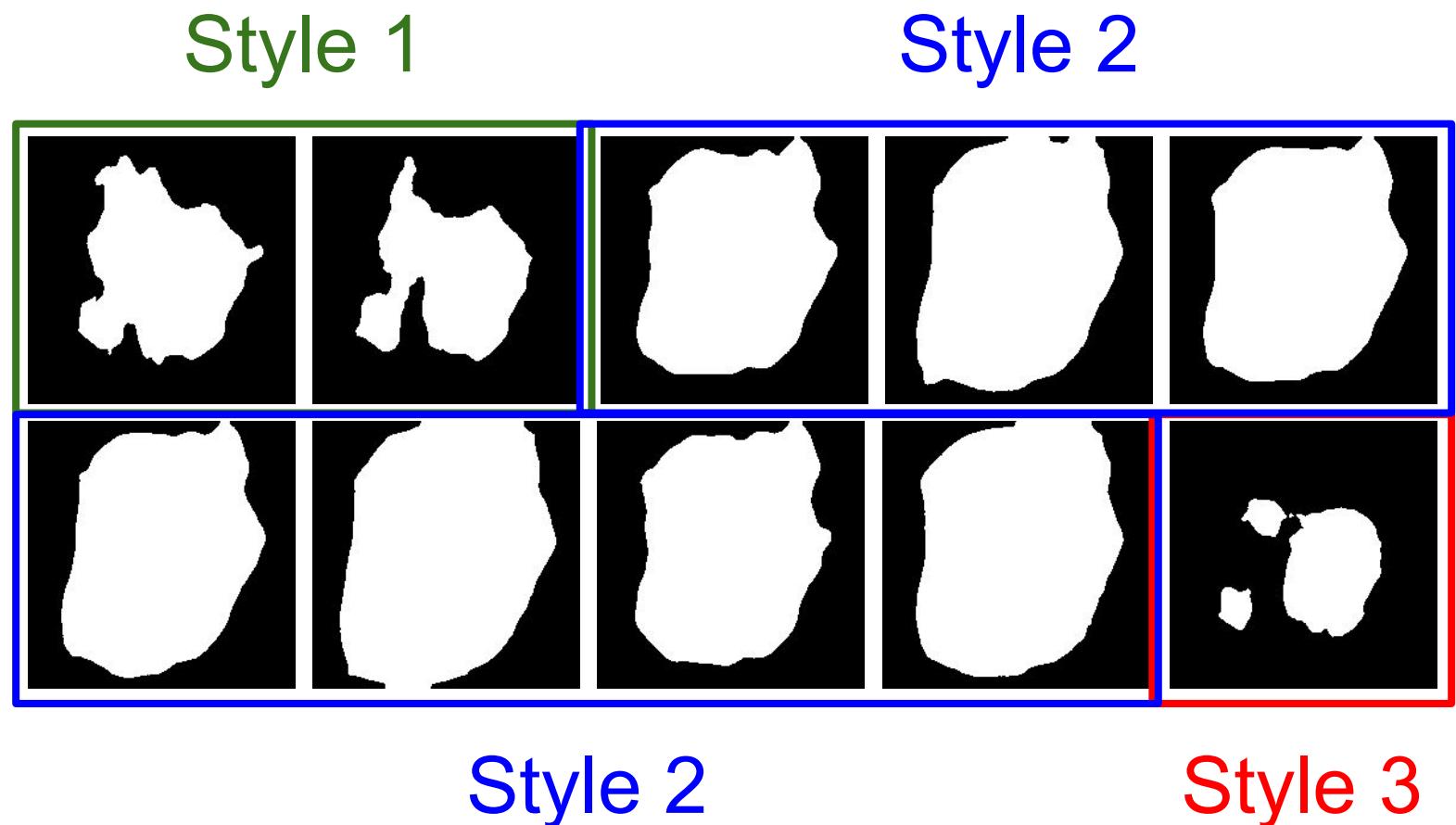
How do we quantify this annotator-style alignment strength?

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

# 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**).



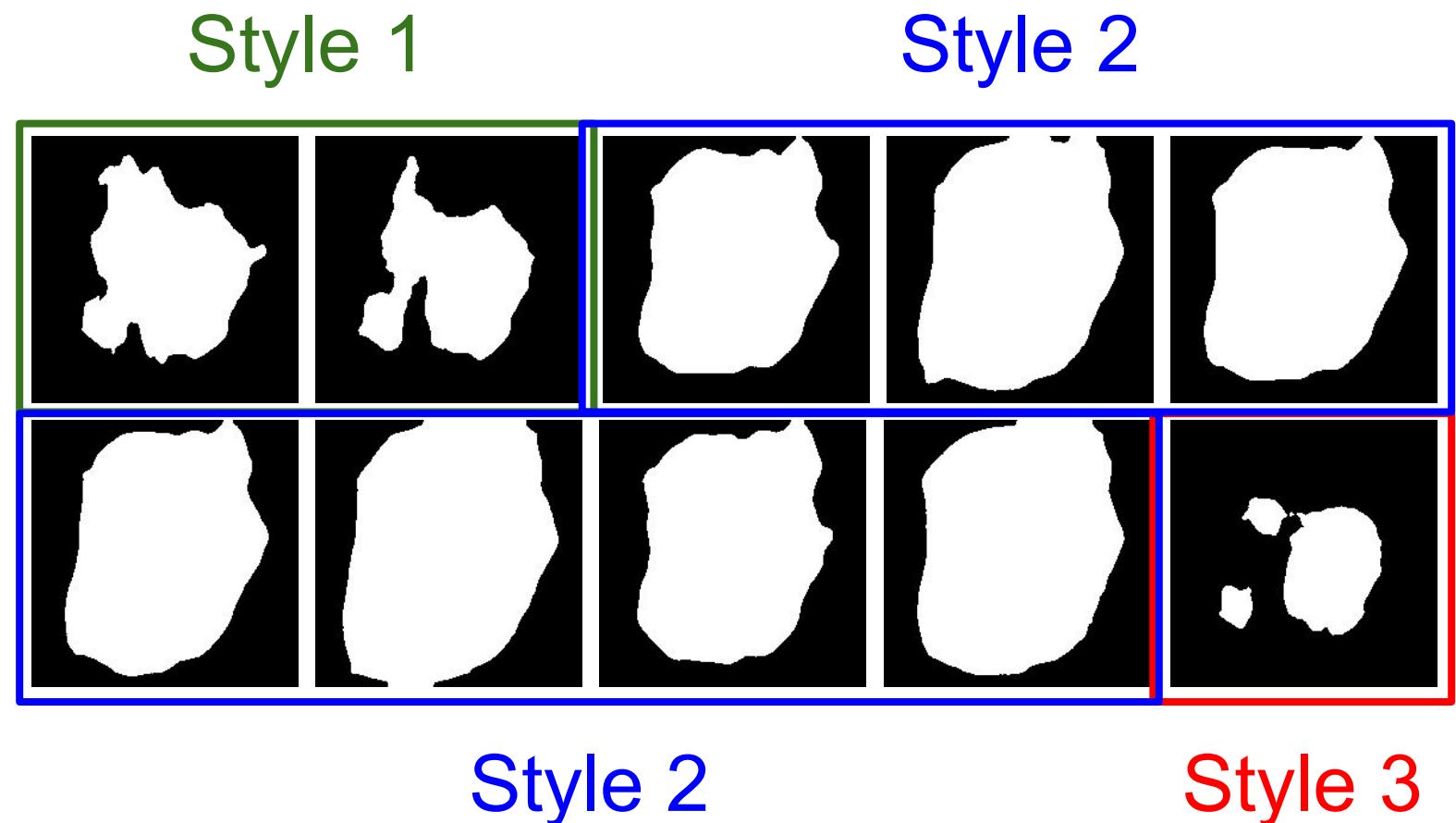
How do we quantify this annotator-style alignment strength?

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

# 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**).



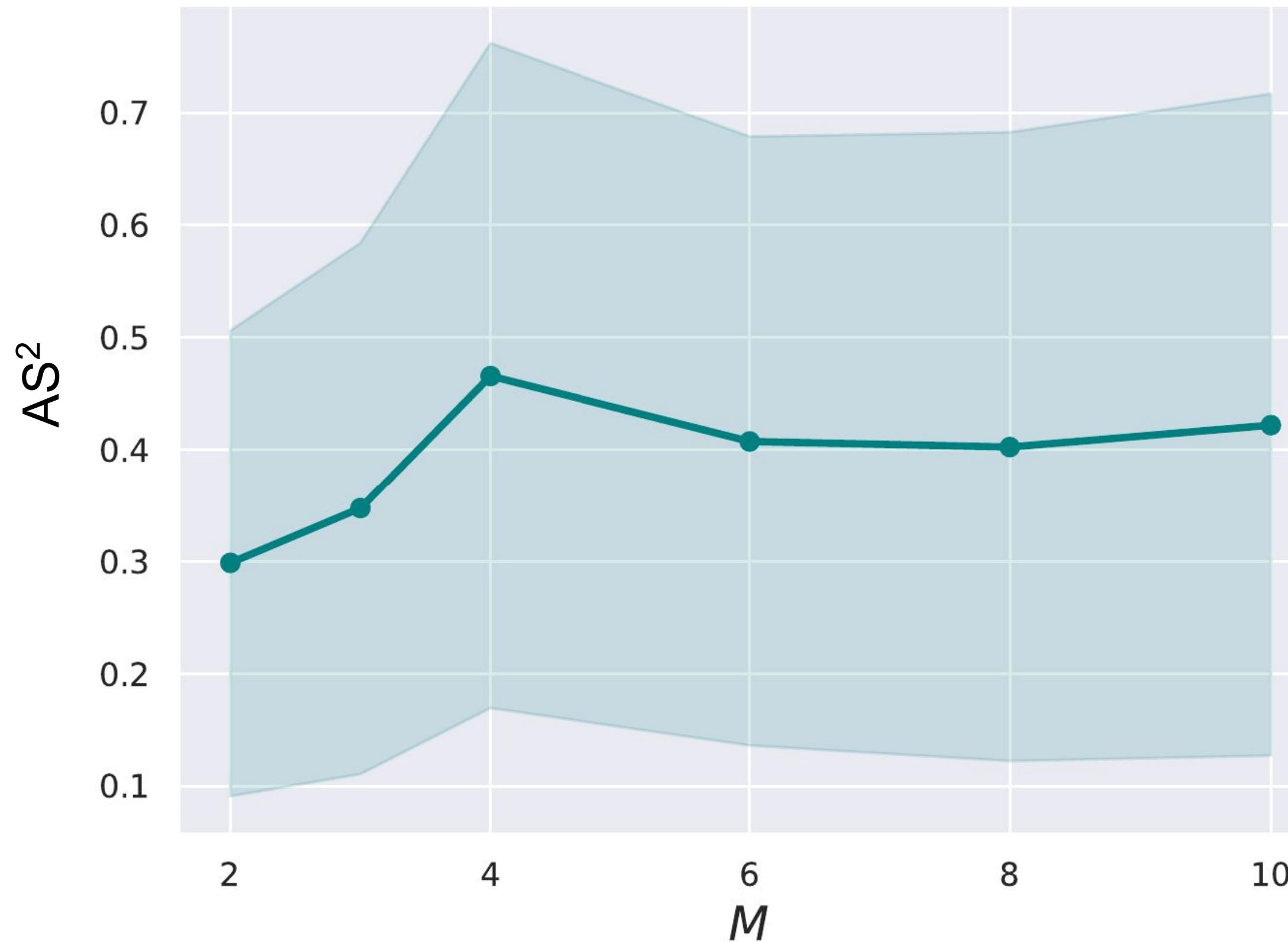
How do we quantify this annotator-style alignment strength?

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

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

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

# 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

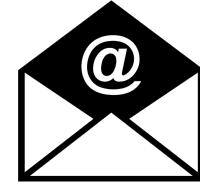
- Formulated the **problem of segmentation style discovery**, and showed that StyleSeg discovers styles that are **plausible, diverse, and semantically consistent**.
- The **largest multi-annotator SLS dataset** ( $> 13.5k$  image-mask pairs) with **annotator correspondence** curated from the ISIC Archive.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

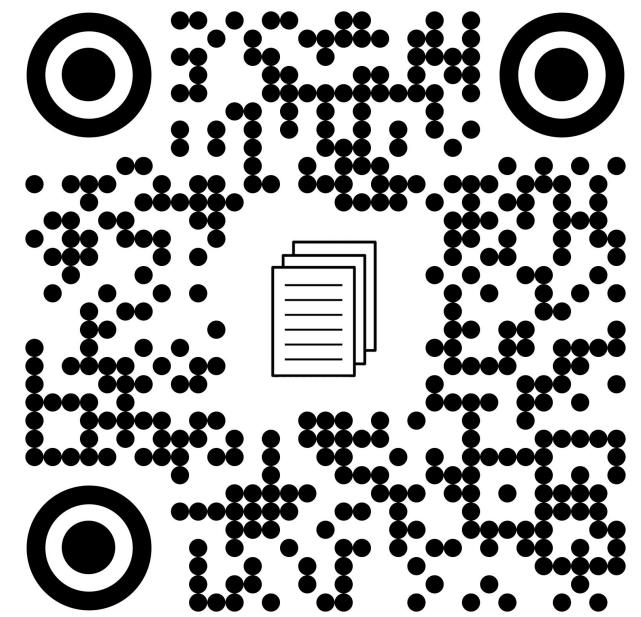
- [1] Silletti et al., “Variability in human and automatic segmentation of melanocytic lesions”, *EMBC*, 2009.
- [2] Mirikharaji et al., “D-LEMA: Deep learning ensembles from multiple annotations-application to skin lesion segmentation”, *CVPR ISIC* 2021.
- [3] Ribeiro et al., “Less is more: Sample selection 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.

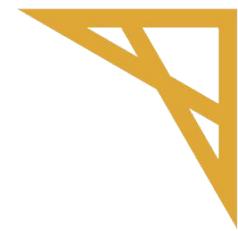
Questions?



kabhishe@sfu.ca



## Acknowledgements



Digital Research  
Alliance of Canada

Alliance de recherche  
numérique du Canada



NVIDIA®