D-LEMA: Deep Learning Ensembles from Multiple Annotations
Application to Skin Lesion Segmentation

Zahra Mirikharaji, Kumar Abhishek, Saeed Izadi, Ghassan Hamarneh

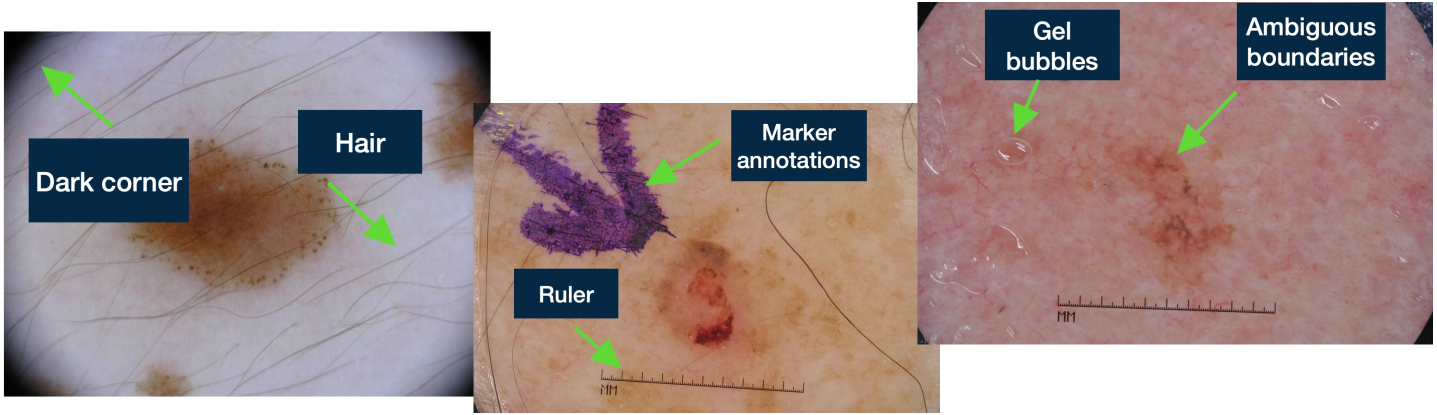
Sixth ISIC Skin Image Analysis Workshop
@ CVPR 2021
June 19, 2021

SFU
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

MIAL
MEDICAL IMAGE ANALYSIS LAB

Skin Lesion Images Challenges

- Natural and artificial artifacts
  - e.g. hair and gel bubbles
- Intrinsic factors
  - e.g. lesion size and shape variations, skin colour and ethnicity as well as ambiguous boundaries
- Variation in imaging conditions
  - e.g. illumination and viewpoint



Annotation Challenges

- The quality of dense annotations required for supervised segmentation affected by:
  - Laborious and costly nature of pixel-wise annotations
  - Ambiguous boundaries
  - Annotator bias
  - Inter- and intra-annotator disagreements even amongst experts

Annotation Challenges

- The quality of dense annotations required for supervised segmentation affected by:
  - Laborious and costly nature of pixel-wise annotations
  - Ambiguous boundaries
  - Annotator bias
  - Inter- and intra-annotator disagreements even amongst experts
- Evaluation using manual segmentations outlined by multiple experts is important

Annotation Challenges

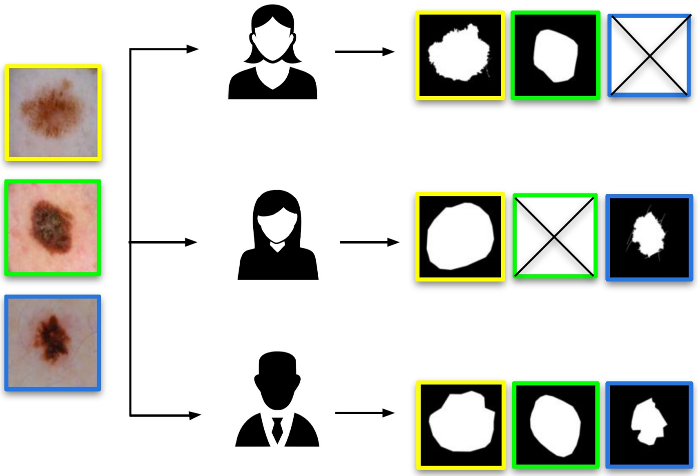
- The quality of dense annotations required for supervised segmentation affected by:
  - Laborious and costly nature of pixel-wise annotations
  - Ambiguous boundaries
  - Annotator bias
  - Inter- and intra-annotator disagreements even amongst experts
- Evaluation using manual segmentations outlined by multiple experts is important
- **Goal**: avoid single annotator bias by training deep segmentation models to learn from multiple annotations as available

Problem

Given a dataset of $\mathcal{X} = \{X_n\}_{n=1}^{N}$ images and

k annotators labeling different subsets of the images:

$$\mathcal{Y} = \{\{Y_{mn}\}_{m=1}^{M_n}\}_{n=1}^{N}$$



M_n : number of annotations for X_n

Y_{mn} : m^{th} annotation of X_n

Problem

Given a dataset of $\mathcal{X} = \{X_n\}_{n=1}^{N}$ images and

k annotators labeling different subsets of the images:

$$\mathcal{Y} = \{\{Y_{mn}\}_{m=1}^{M_n}\}_{n=1}^{N}$$

⇩

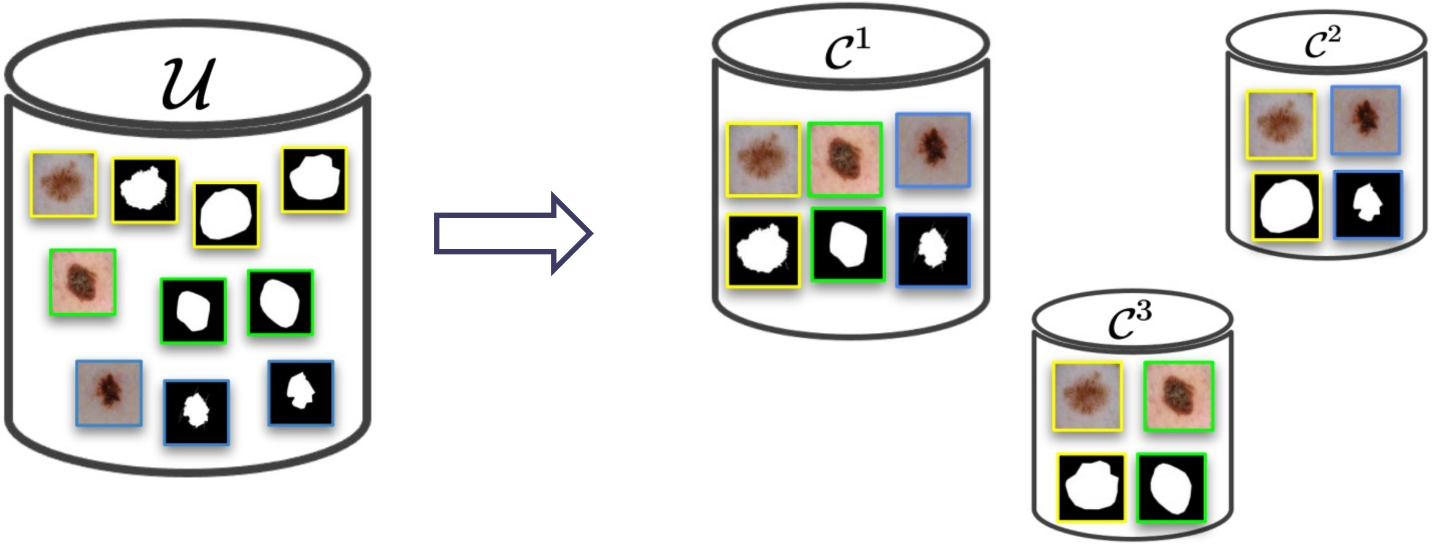
Train a segmentation network that generalizes well to unseen data while effectively leveraging all annotations toward making reliable predictions

Approach - Non-contradictory Subsets Selection

- Let M indicate the maximum number of annotations per image over the entire dataset $\mathcal{U}$.
- Partition the entire dataset into M disjoint subsets denoted by $\{C^i\}_{i=1}^{M}$ such that each C^i includes at most one annotation for every image

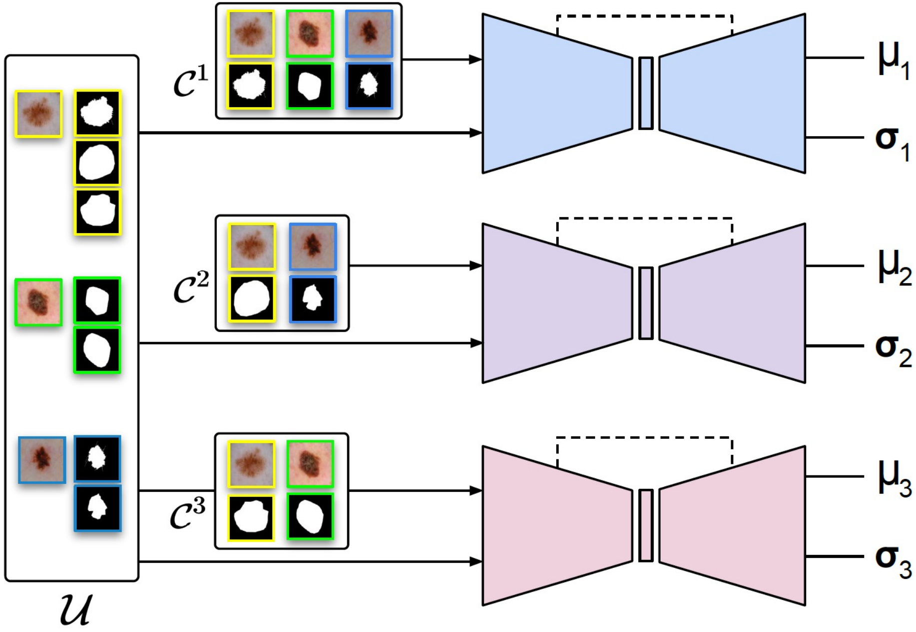
Approach - Non-contradictory Subsets Selection

- For each image, with $\mathrm{M} \leq M_n$ annotations, we randomly assign the M annotations to $\{\mathcal{C}^i\}_{i=1}^{M}$ subsets.



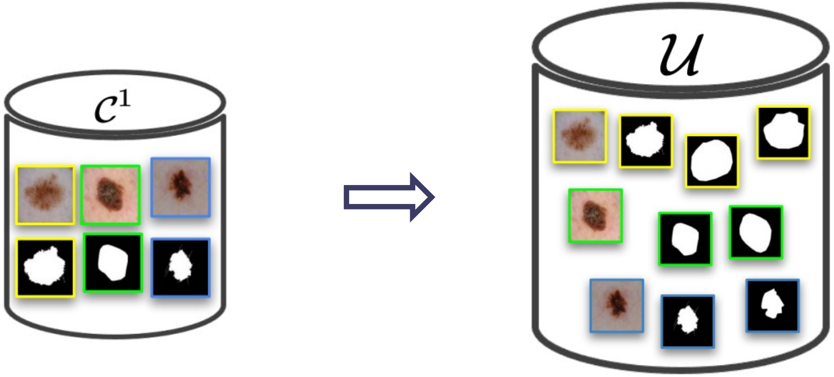
Approach – Learning Models

- Train M base networks where network i models the experts knowledge in C^i.



Approach – Learning Models

- To train model i, leverage non-contradictory subset $\mathcal{C}^i$ to assess the quality of annotations in $\mathcal{U}$.
- Learn spatially-adaptive weight maps for annotations in $\mathcal{U}$ to adjust how to treat each pixel annotation in the optimization of deep network.



Approach – Learning Models

- Specifically, for each model i, we define a weighted CE loss on the data set $\mathcal{U}$:

$$\mathcal{L}\left(\hat{Y}_n^i, Y_{mn};\ \theta^i, W_{mn}^i\right) = \sum_q W_{mnq}^i Y_{mn} \log \hat{Y}_{nq}^i$$

Approach – Learning Models

- Specifically, for each model i, we define a weighted CE loss on the data set $\mathcal{U}$:

$$\mathcal{L}(\hat{Y}_n^i, Y_{mn};\ \theta^i, W_{mn}^i) = \sum_q \boxed{W_{mnq}^i} Y_{mn} \log \hat{Y}_{nq}^i$$

W_{mnq}^i is the weight associated with pixel q of the m-th annotation of image n in model i.

W^i contains all spatial weights associated with annotations in set $\mathcal{U}$ leaned in model i.

Approach – How to learn W

- Learn W^i dynamically by evaluating the network on C^i

$$\mathcal{L} = L_{ce}^{C^i}$$

$$W^i = argmin_{W^i}\mathcal{L}$$

Approach – How to learn W

- Learn W^i dynamically by evaluating the network on C^i

$$\mathcal{L} = L_{ce}^{C^i}$$

$$W^i = argmin_{W^i}\mathcal{L}$$

- Learn network parameters θ^i and weight maps W^i, alternatively.

Approach - Fusion of Predictions

- Once the individual base models are trained, the final prediction of the entire ensemble for the X_n is obtained by using a weighted fusion

$$\hat{Y}_n = \sum_{i=1}^{M} \alpha_n^i \hat{Y}_n^i$$

Approach - Fusion of Predictions

- Once the individual base models are trained, the final prediction of the entire ensemble for the X_n is obtained by using a weighted fusion

$$\hat{Y}_n = \sum_{i=1}^{M} \alpha_n^i \hat{Y}_n^i$$

where α_n^i is the combination coefficient for prediction by model i defined by either:

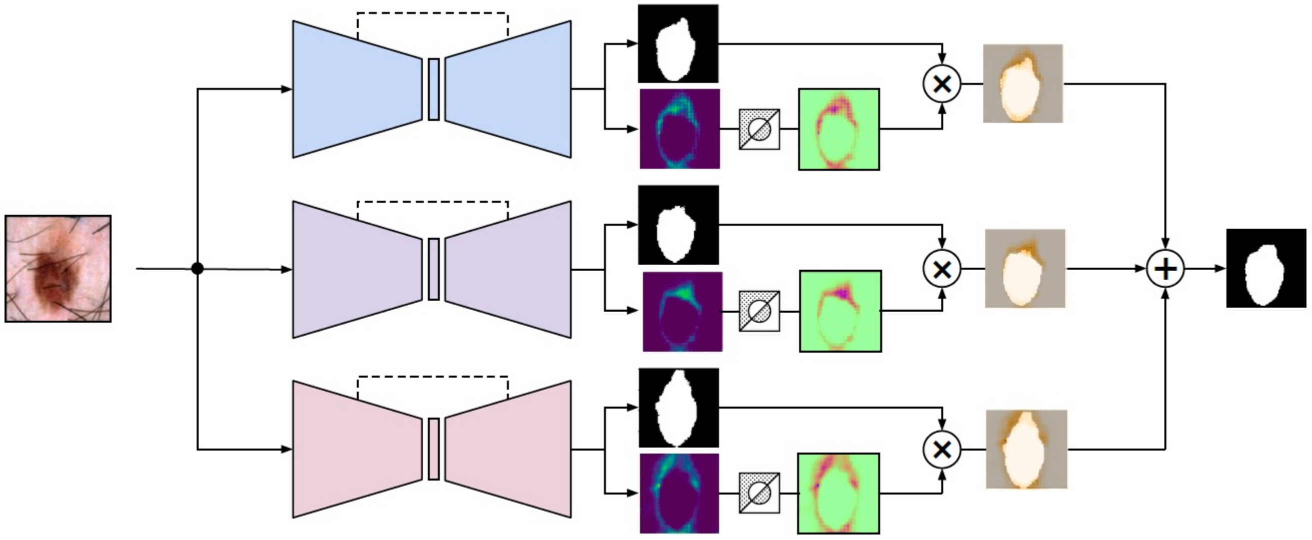
- Equally weighted averaging
- Model confidence

Approach - Uncertainty-driven Aggregation

- Leverage aleatoric uncertainty to estimate how confident a base model is about its prediction in two forms:
  - Considering the pixel-wise uncertainty values as spatially adaptive coefficients
  - Averaging the pixel-wise uncertainty into a scalar image-level coefficient.

Approach - Uncertainty-driven Aggregation

- Utilize the confidence coefficients when combining the base models prediction maps



Data Description - Training

- The International Skin Imaging Collaboration (ISIC) Archive data
- 2,223 images with more than one segmentation ground truth mask

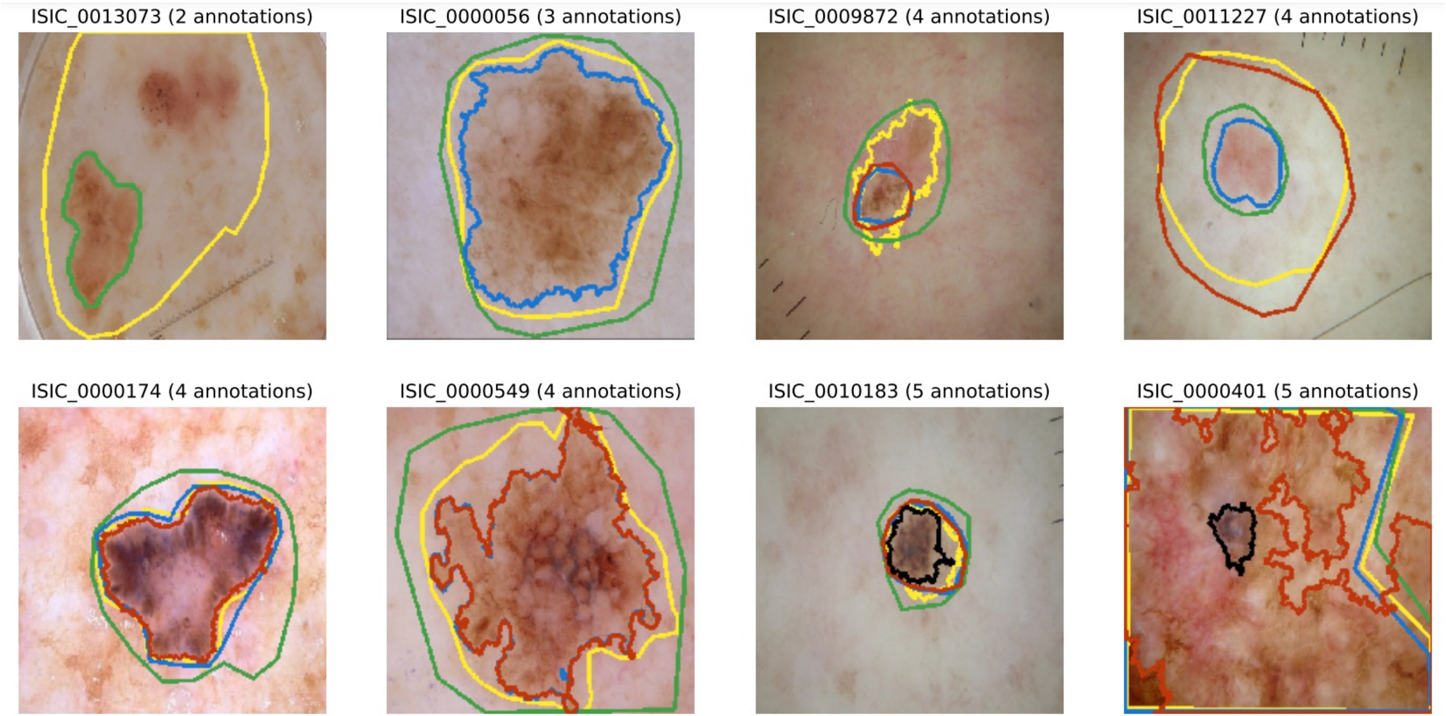
number of annotations	2	3	4	5
number of images	2094	100	36	3

- Split images to 80% for training and 20% for validation

Data Description - Training

- For model selection, we randomly selected which annotation to use in the validation set.
- Create non-contradictory annotation sets: all training data are randomly and uniformly partitioned into five groups of overlapping images but unique ground truth annotations

Data Description - Training



Sample skin lesion images from the ISIC Archive with multiple lesion boundary annotations

Data Description - Test

- Evaluate the proposed framework on three publicly available datasets:
  - **ISIC**: 2,000 images with just one segmentation ground truth from ISIC Archive
  - **PH2**: The PH2 dataset contains 200 color dermoscopic images
  - **DermoFit**: This dataset has 1300 color clinical

Quantitative Results – Segmentation Performance

	Method	ISIC Archive [1]	PH^2 [26]	DermoFit [2]
A	baseline	68.00 ± 0.56	81.30 ± 0.77	70.30 ± 0.54
B	model 0	69.22 ±0.53	82.82 ± 0.75	72.57 ± 0.50
C	model 1	69.75 ± 0.55	82.40 ± 0.75	71.05 ± 0.55
D	model 2	70.33 ± 0.52	83.46 ± 0.74	72.80 ± 0.51
E	model 3	70.37 ± 0.51	83.31 ± 0.70	73.04 ± 0.53
F	model 4	69.73 ± 0.52	82.29 ± 0.72	70.87 ± 0.48
G	equally weighted fusion (ours)	**72.11± 0.51**	84.96± 0.73	74.22± 0.51
H	pixel-level confidence (ours)	71.46± 0.49	84.52± 0.74	73.91± 0.53
I	image-level confidence (ours)	**72.08± 0.49**	**85.20 ± 0.70**	**74.33± 0.50**
J	less is more [30]	69.20	81.25	72.55

Comparing the segmentation performance based on Jaccard index

Quantitative Results – Segmentation Performance

	Method	ISIC Archive [1]	PH^2 [26]	DermoFit [2]
A	baseline	68.00 ± 0.56	81.30 ± 0.77	70.30 ± 0.54
B	model 0	69.22 ±0.53	82.82 ± 0.75	72.57 ± 0.50
C	model 1	69.75 ± 0.55	82.40 ± 0.75	71.05 ± 0.55
D	[illegible]	[illegible]	[illegible]	[illegible] 0.51
E	[illegible]	[illegible]	[illegible]	[illegible] 0.53
F	[illegible]	[illegible]	[illegible]	[illegible] 0.48
G	equ[illegible]	[illegible]	[illegible]	[illegible] 0.51
H	pixel-level confidence (ours)	71.46± 0.49	84.52± 0.74	73.91± 0.53
I	image-level confidence (ours)	**72.08± 0.49**	**85.20 ± 0.70**	**74.33± 0.50**
J	less is more [30]	69.20	81.25	72.55

for every image in the training batch, we randomly select which ground truth to use, when optimizing the loss function.

Comparing the segmentation performance based on Jaccard index

Quantitative Results – Segmentation Performance

	Method	ISIC Archive [1]	PH^2 [26]	DermoFit [2]
A	baseline	68.00 ± 0.56	81.30 ± 0.77	70.30 ± 0.54
B	model 0	69.22 ±0.53	82.82 ± 0.75	72.57 ± 0.50
C	model 1	69.75 ± 0.55	82.40 ± 0.75	71.05 ± 0.55
D	model 2	70.33 ± 0.52	83.46 ± 0.74	72.80 ± 0.51
E	model 3	70.37 ± 0.51	83.31 ± 0.70	73.04 ± 0.53
F	model 4	69.73 ± 0.52	82.29 ± 0.72	70.87 ± 0.48
G	equally weighted fusion (ours)	**72.11± 0.51**	84.96± 0.73	74.22± 0.51
H	pixel-level confidence (ours)	71.46± 0.49	84.52± 0.74	73.91± 0.53
I	image-le[illegible]	[illegible]	[illegible]	[illegible]
J	les[illegible]	[illegible]	[illegible]	[illegible]

base models trained on non-contradictory annotations
simulating an expert knowledge

Comparing the segmentation performance based on Jaccard index

Quantitative Results – Segmentation Performance

- Row G: combined predictions by averaging the output probabilities
- Row H: predictions fusion using normalized confidence spatial maps computed by inverting the predicted aleatoric outputs
- Row I: fused predictions using image-level normalized confidence scalars computed by inverting the uncertainty scalars

D	model 2	[illegible]	[illegible]	[illegible]
E	model 3	70.37 ± 0.51	83.31 ± 0.70	73.04 ± 0.53
F	model 4	69.73 ± 0.52	82.29 ± 0.72	70.87 ± 0.48
G	equally weighted fusion (ours)	**72.11± 0.51**	84.96± 0.73	74.22± 0.51
H	pixel-level confidence (ours)	71.46± 0.49	84.52± 0.74	73.91± 0.53
I	image-level confidence (ours)	**72.08± 0.49**	**85.20 ± 0.70**	**74.33± 0.50**
J	less is more [30]	69.20	81.25	72.55

Comparing the segmentation performance based on Jaccard index

Quantitative Results – Segmentation Performance

	Method	ISIC Archive [1]	PH^2 [26]	DermoFit [2]
A	baseline	68.00 ± 0.56	81.30 ± 0.77	70.30 ± 0.54
B	model 0	69.22 ±0.53	82.82 ± 0.75	72.57 ± 0.50
C	model 1	69.75 ± 0.55	82.40 ± 0.75	71.05 ± 0.55
D	model 2	70.33 ± 0.52	83.46 ± 0.74	72.80 ± 0.51
E	model 3	70.37 ± 0.51	83.31 ± 0.70	73.04 ± 0.53
F	[illegible]	[illegible]	[illegible]	[illegible]7 ± 0.48
G	[illegible]	[illegible]	[illegible]	[illegible]2± 0.51
H	pixel-level confidence (ours)	71.46± 0.49	84.52± 0.74	73.91± 0.53
I	image-level confidence (ours)	**72.08± 0.49**	**85.20 ± 0.70**	**74.33± 0.50**
J	less is more [30]	69.20	81.25	72.55

A subset of samples with small annotator disagreements is taken into account during the training.

Comparing the segmentation performance based on Jaccard index

Quantitative Results – Segmentation Performance

	Method	ISIC Archive [1]	PH^2 [26]	DermoFit [2]
A	baseline	68.00 ± 0.56	81.30 ± 0.77	70.30 ± 0.54
B	model 0	69.22 ± 0.53	82.82 ± 0.75	72.57 ± 0.50
C	model 1	69.75 ± 0.55	82.40 ± 0.75	71.05 ± 0.55
D	model 2	70.33 ± 0.52	83.46 ± 0.74	72.80 ± 0.51
E	model 3	70.37 ± 0.51	83.31 ± 0.70	73.04 ± 0.53
F	model 4	69.73 ± 0.52	82.29 ± 0.72	70.87 ± 0.48
G	equally weighted fusion (ours)	$\mathbf{72.11 \pm 0.51}$	84.96 ± 0.73	74.22 ± 0.51
H	pixel-level confidence (ours)	71.46 ± 0.49	84.52 ± 0.74	73.91 ± 0.53
I	image-level confidence (ours)	$\mathbf{72.08 \pm 0.49}$	$\mathbf{85.20 \pm 0.70}$	$\mathbf{74.33 \pm 0.50}$
J	less is more [30]	69.20	81.25	72.55

Comparing the segmentation performance based on Jaccard index

Quantitative Results – Predictive Uncertainty

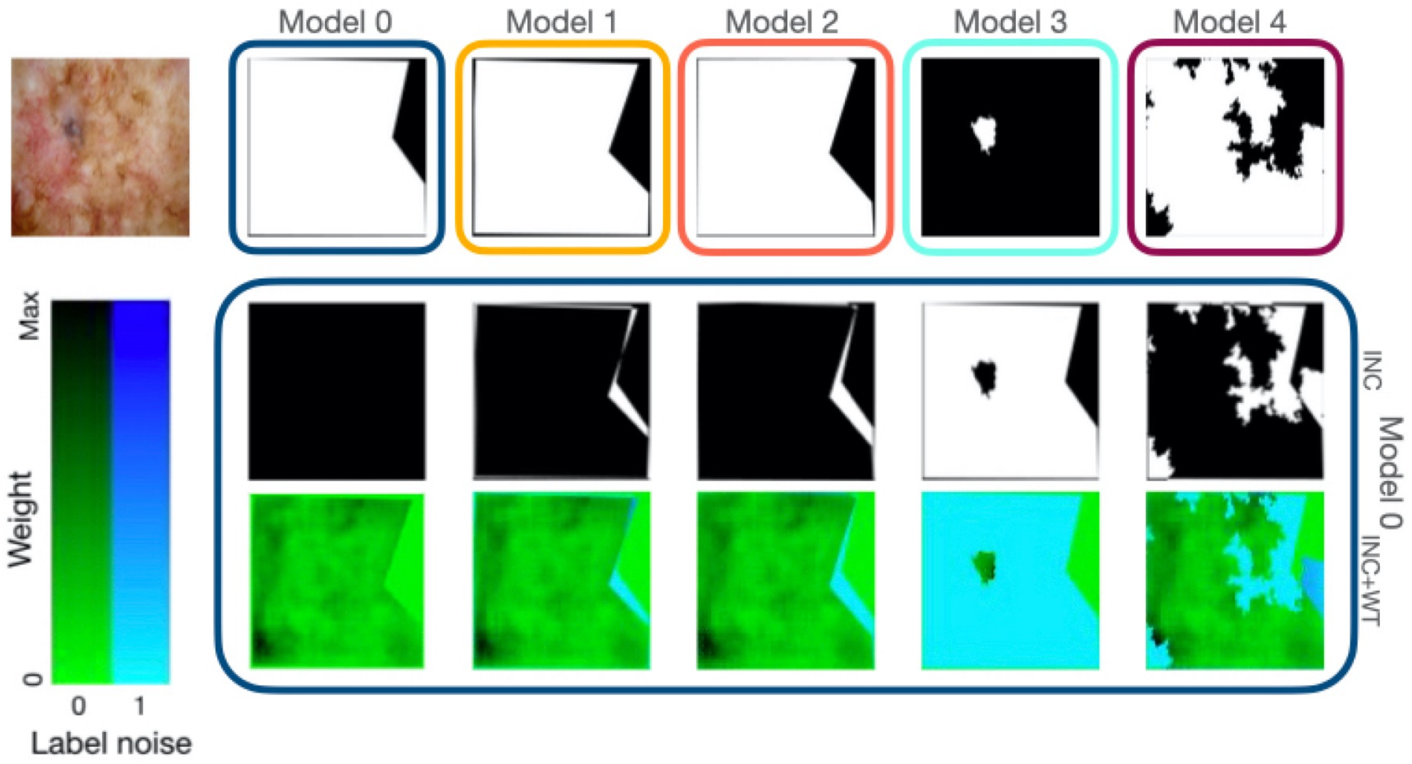
- Modeling predictive uncertainty in clinical applications without a ‘real’ gold standard is helpful in decision making
- Evaluate the calibration quality of our ensemble annotation aggregation by:
  - Negative log-likelihood (NLL)
  - Brier score (Br)
- Implement Bayesian epistemic uncertainty using dropout for base models

Quantitative Results – Predictive Uncertainty

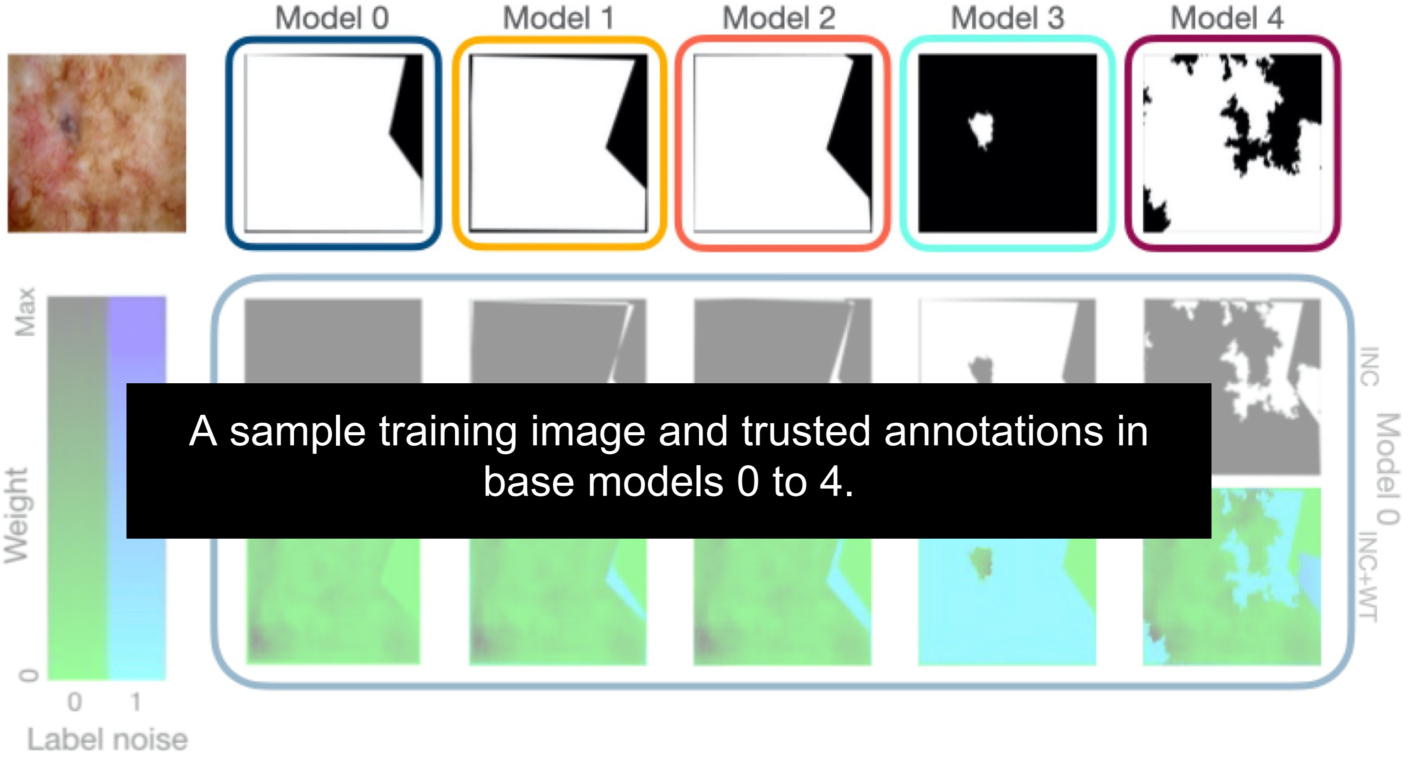
Dataset		ISIC Archive		PH^2		DermoFit	
Method		NLL	Br	NLL	Br	NLL	Br
A	MC dropout model 0	0.073	0.019	0.166	0.048	0.272	0.082
B	MC dropout model 1	0.075	0.020	0.151	0.044	0.310	0.099
C	MC dropout model 2	0.075	0.019	0.149	0.044	0.283	0.087
D	MC dropout model 3	0.078	0.020	0.152	0.042	0.291	0.091
E	MC dropout model 4	0.075	0.019	0.155	0.045	0.312	0.100
F	deep ensemble (ours)	**0.070**	**0.018**	**0.144**	**0.041**	**0.254**	**0.078**

Predictive uncertainty based on negative log-likelihood (NLL) and Brier score (Br)

Qualitative Results – Weight Matrices

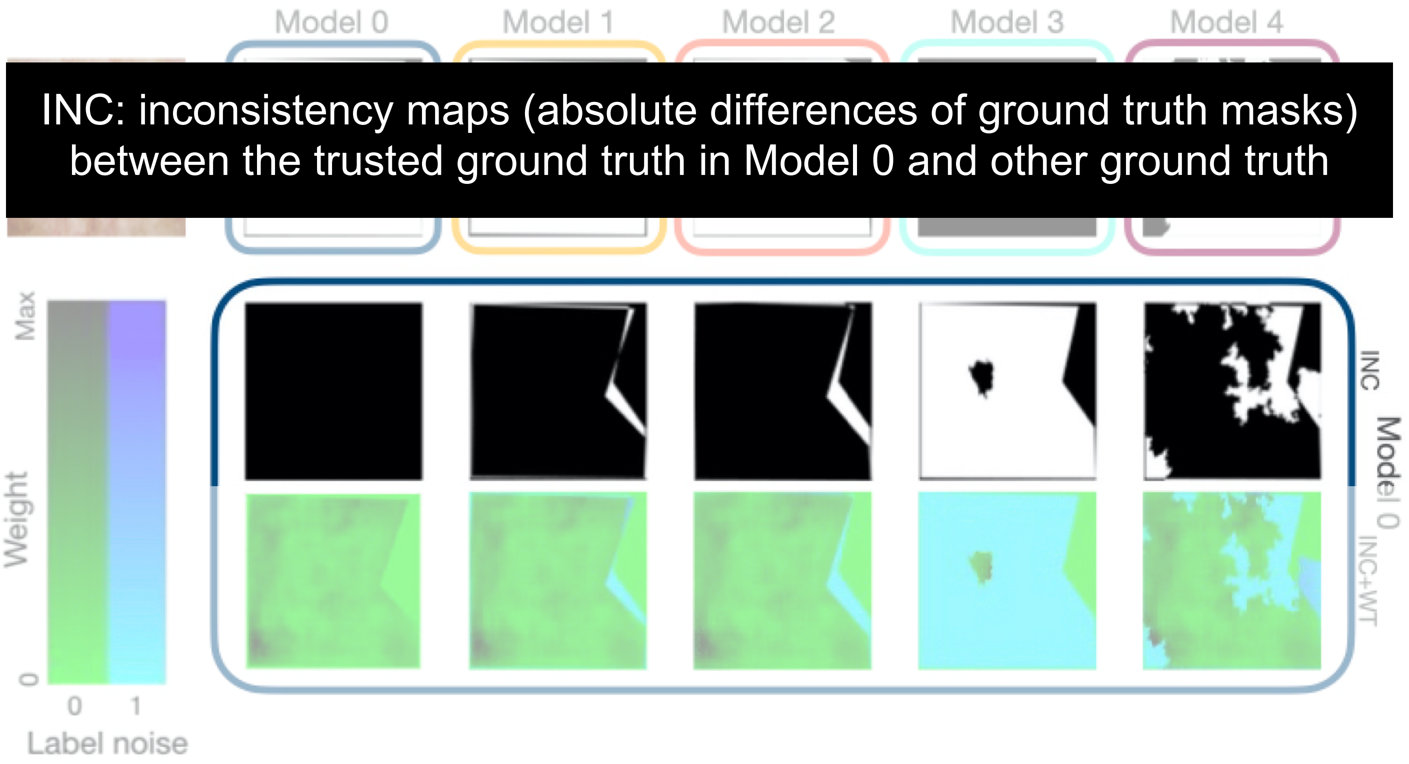


Qualitative Results – Weight Matrices



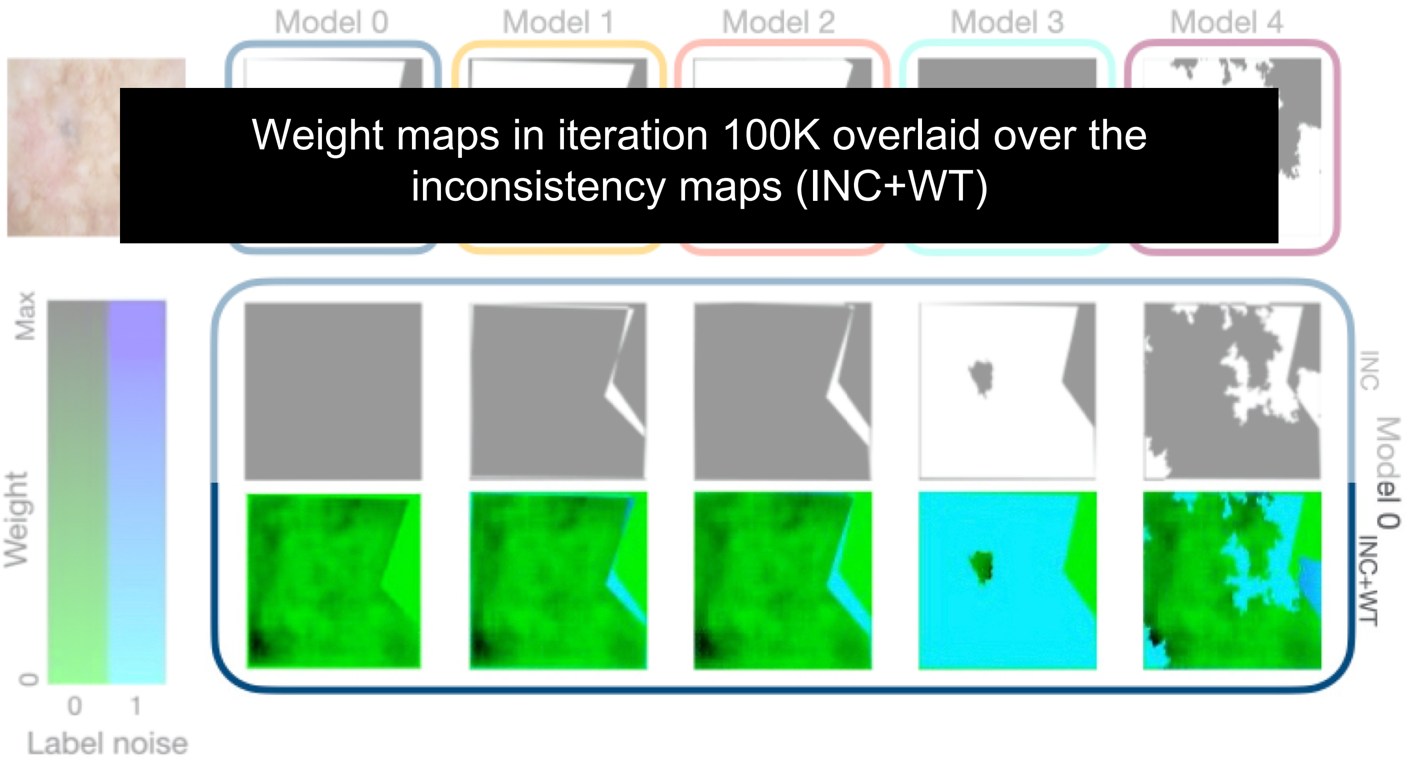
A sample training image and trusted annotations in base models 0 to 4.

Qualitative Results – Weight Matrices



INC: inconsistency maps (absolute differences of ground truth masks) between the trusted ground truth in Model 0 and other ground truth

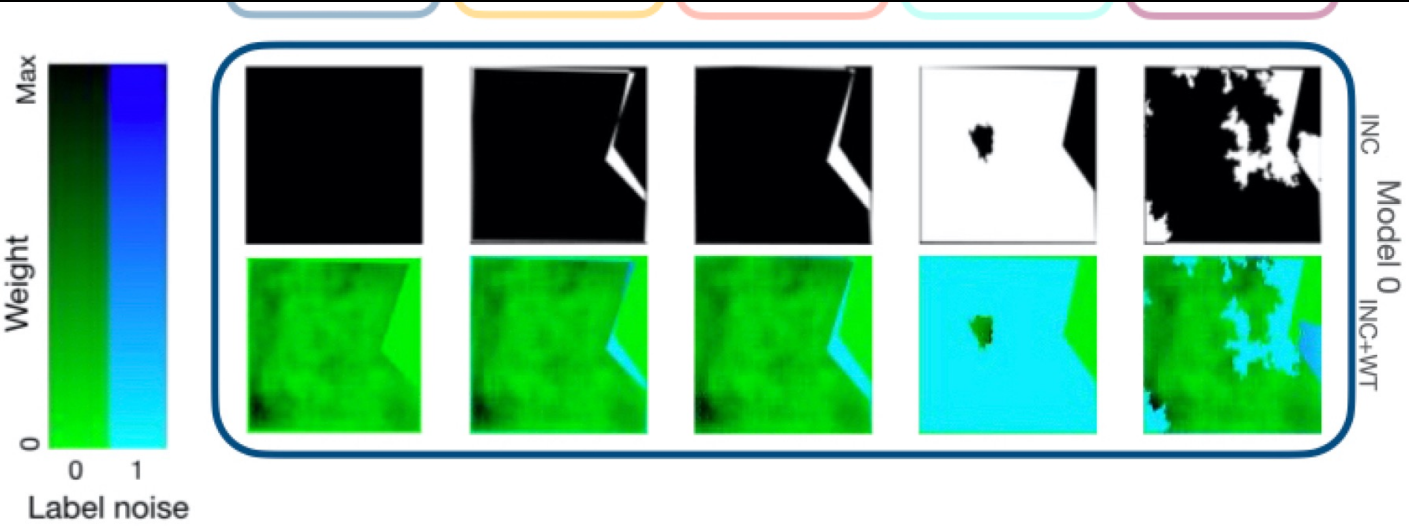
Qualitative Results – Weight Matrices



Weight maps in iteration 100K overlaid over the inconsistency maps (INC+WT)

Qualitative Results – Weight Matrices

- The location of the cyan pixels matches the inconsistency maps
- Zero or very close to zero weights are assigned to inconsistent annotated pixels
- Exclusively leveraging the experts knowledge in C^i when learning θ^i



Summary

- We proposed an ensemble paradigm to:
  - model different experts' skills independently
  - deal with discrepancies in segmentation annotations

Summary

- We proposed an ensemble paradigm to:
  - model different experts' skills independently
  - deal with discrepancies in segmentation annotations
- A robust-to annotation-noise learning scheme is utilized to efficiently leverage experts' opinions toward learning from all available annotations.

Summary

- We proposed an ensemble paradigm to:
  - model different experts' skills independently
  - deal with discrepancies in segmentation annotations
- A robust-to annotation-noise learning scheme is utilized to efficiently leverage experts' opinions toward learning from all available annotations.
- To improve quality of predictive uncertainty in clinical applications, aleatoric and epistemic uncertainties are modeled and confidence calibration improved.

Thank you!

Acknowledgments

compute canada | calcul canada

NVIDIA

NSERC CRSNG

zmirikha@sfu.ca

www.MedicalImageAnalysis.com