Confidence Contours

Uncertainty-Aware Annotation for Medical Semantic Segmentation

Andre Ye, Quan Ze (Jim) Chen, Amy X. Zhang
University of Washington, Social Futures Lab
November 7th, 2023 | HCOMP – Delft, Netherlands

[1/3]

The Pitfalls of Modeling Structurally Uncertain Tasks

Goal

To build useful models for uncertain/ambiguous tasks

(e.g., in medicine)

Medical segmentation – Overview

- **Goal:** Identify which pixels correspond to an object of interest
- **Used:** for diagnostic purposes, e.g. lung cancer risk
- **Models:** quickly process very high-res data
- **Stakes:** area/shape can influence diagnosis

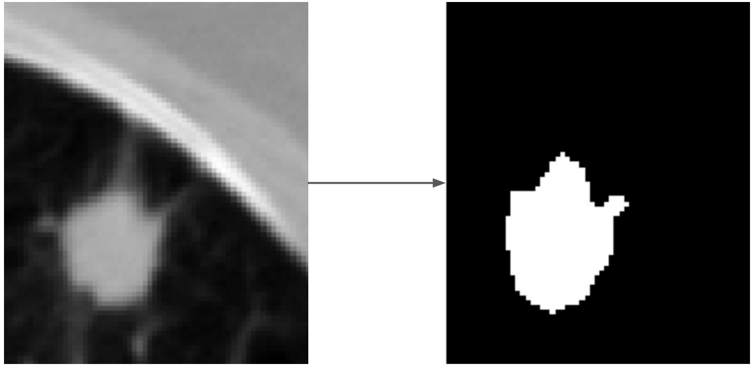
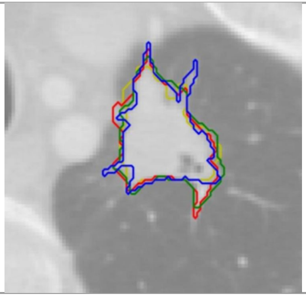


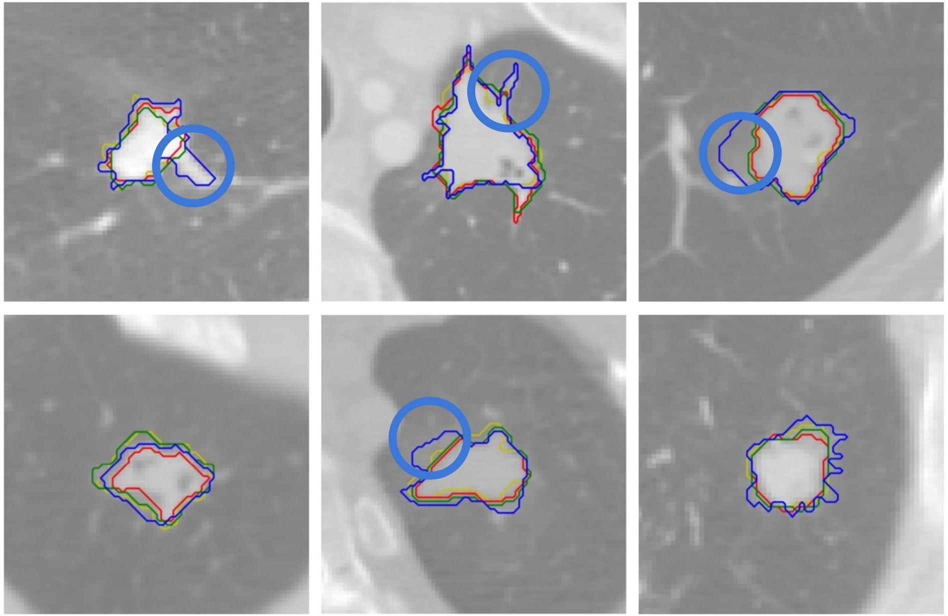
Image & annotation from the Lung Image Consortium Dataset (LIDC)

Medical segmentation often features structural uncertainty, producing annotation disagreement.

Distinguishing superficial and structural uncertainty

Uncertainty	Example	Main Sources	Produces
Superficial	(find img)	Image quality	Continuous disagreement
Structural		Image contents, domain knowledge	Discrete disagreement

Medical segmentation often features structural uncertainty: discrete annot. disagreement



Images & annotations from the Lung Image Consortium Dataset (LIDC)

How to build models in the face of structural uncertainty?

Three approaches:

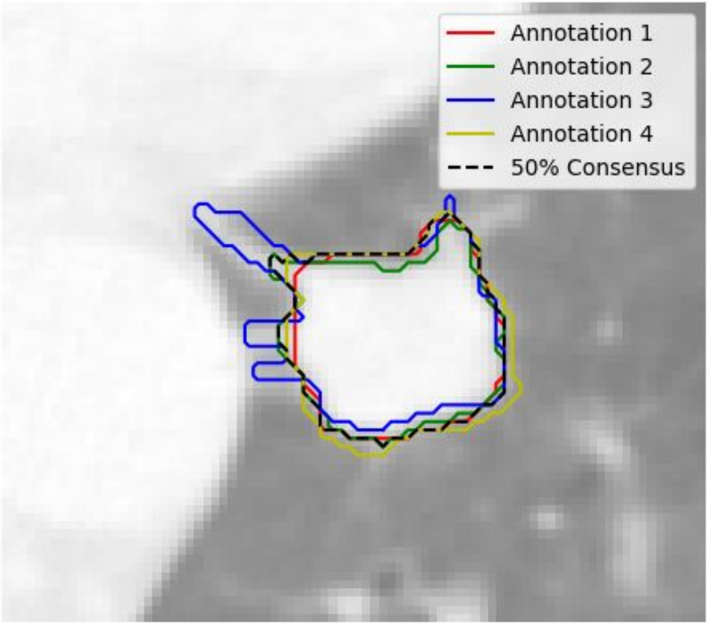
1. ☒ "Averaging out"
2. ☒ Modify the model
3. ☐ Confidence Contours

Approach 1: “Average out” the uncertainty

Voting, mean/median consensus, etc.

Works for superficial uncertainty, but tricky for structural uncertainty:

- Not necessarily consistent with domain knowledge rules
- Lose out on structural info
  - High stakes!



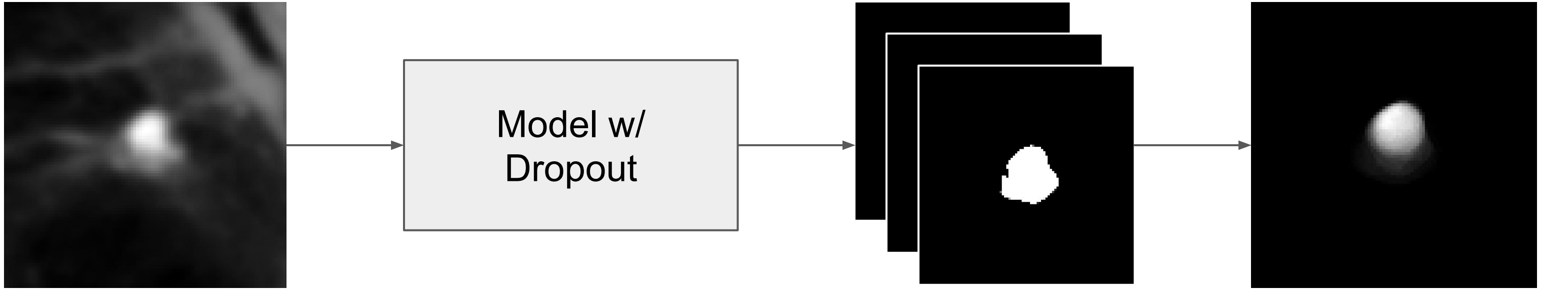
⚠️ **Intervention #1**

Structural uncertainty isn’t just a “problem”, it’s an important signal

Approach 2: Modify the model

- **Continuous Maps:** produce “smooth” (not “hard”) segmentations
- **Candidate Generation:** produce *k* possible “hard” segmentations

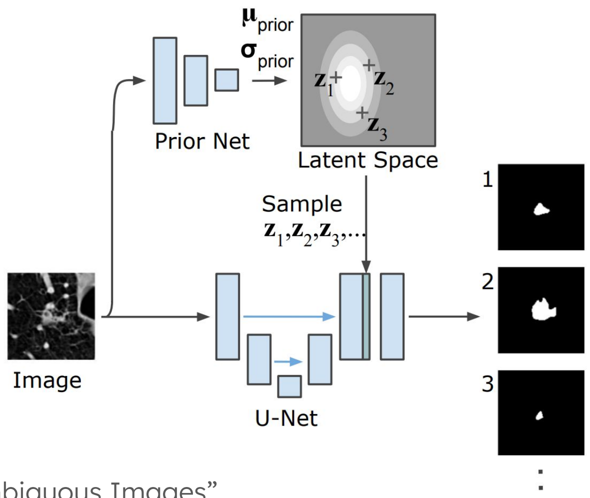
Example (Continuous Maps): Bayesian Uncertainty



Approach 2: Modify the model

- **Continuous Maps:** produce "smooth" (not "hard") segmentations
- **Candidate Generation:** produce *k* possible "hard" segmentations

Example (Candidate Generation): Probabilistic U-Net



Kohl, "A Probabilistic U-Net for Segmentation of Ambiguous Images"

⚠️ Intervention #2

What properties do we want for uncertainty representations?

Desirable properties of uncertainty rep’s

Human-friendly. Humans both…

- …**provide** uncertainty information, and
- …need to **use** uncertainty representations.

Providing –

1. **Convenient to annotate**. Should be low-effort & natural

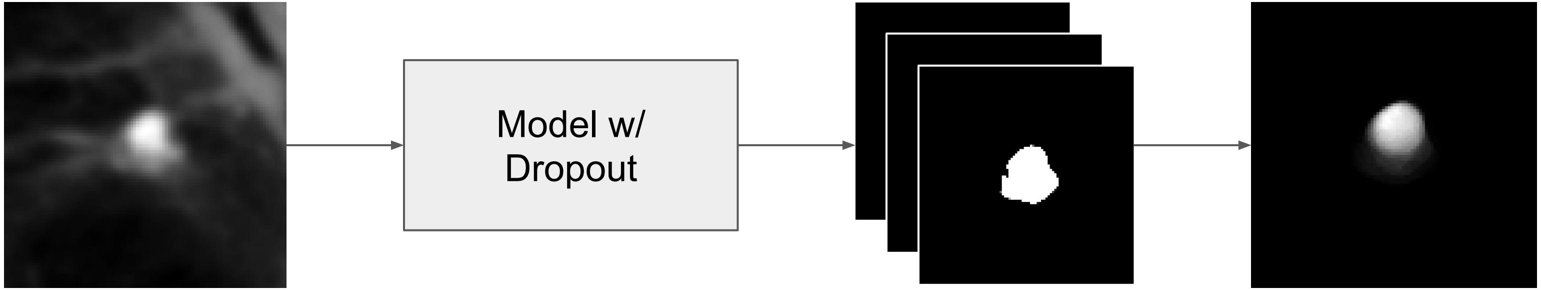
Using –

2. **Informative**. Rep’s provide enough info to do the job
3. **Concise**. Rep’s do not have info overload

Approach 2: Modify the model

- **Continuous Maps:** produce “smooth” (not “hard”) segmentations
- **Candidate Generation:** produce *k* possible “hard” segmentations

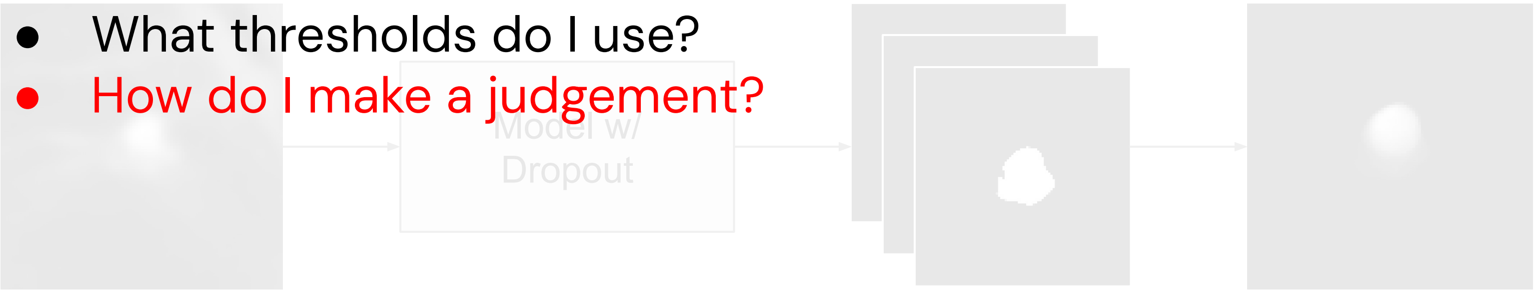
Example (Continuous Maps): Bayesian Uncertainty



Approach 2: Modify the model

- **Continuous Maps:** produce “smooth” (not “hard”) segmentations
- **Candidate Generation:** produce *k* possible “hard” segmentations

Example (Continuous Maps): Bayesian Uncertainty



Convenient to annotate? ✅ Annotate “as normal”

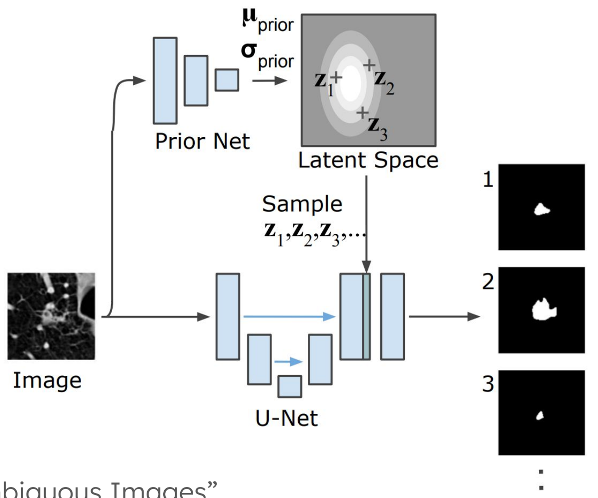
Informative and Concise? 🤔

- Is the model bad or is the data hard?
- What thresholds do I use?
- How do I make a judgement?

Approach 2: Modify the model

- **Continuous Maps:** produce "smooth" (not "hard") segmentations
- **Candidate Generation:** produce *k* possible "hard" segmentations

Example (Candidate Generation): Probabilistic U-Net

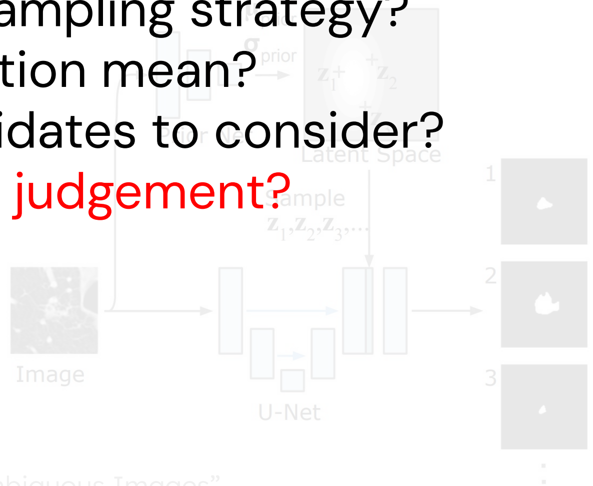


Kohl, "A Probabilistic U-Net for Segmentation of Ambiguous Images"

Approach 2: Modify the model

- **Continuous Maps:** produce "smooth" (not "hard") segmentations
- **Candidate Generation:** produce *k* possible "hard" segmentations

Example (Continuous Maps): Probabilistic U-Net



Convenient to annotate? ✅ Annotate "as normal"

Informative and Concise? 🤔

- Contingent on sampling strategy?
- What does variation mean?
- How many candidates to consider?
- How do I make a judgement?

Kohl, "A Probabilistic U-Net for Segmentation of Ambiguous Images"

Model-centric approaches disconnect uncertainty from human judgement.

Providing **Using**

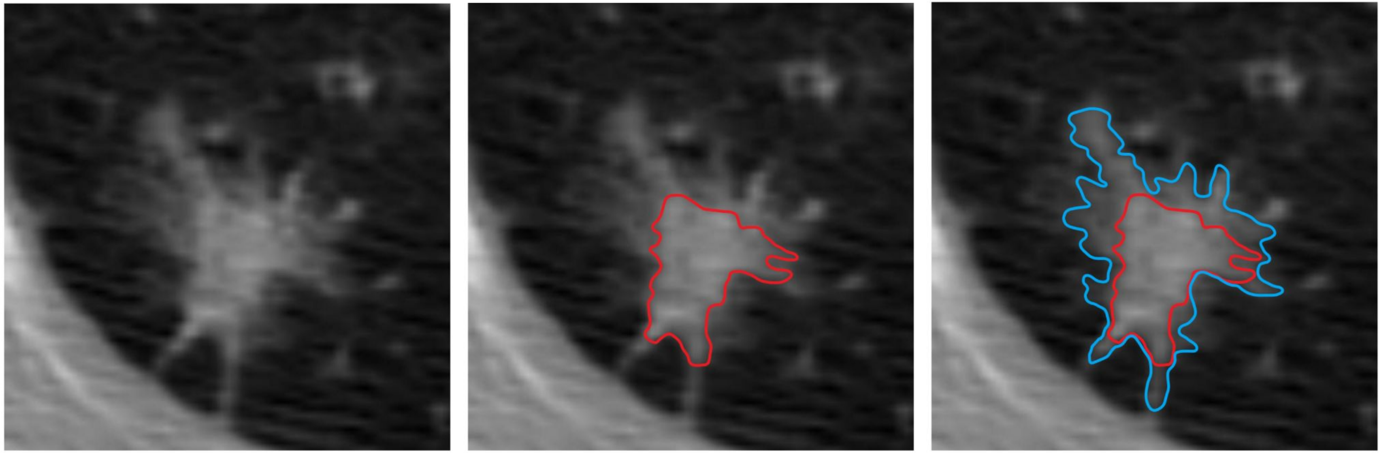
[2/3]
Designing Human-Connected Uncertainty Representations

⚠️ **Intervention #3**

If we're stuck, let's reapproach the ground truth instead of building a fancier model

Our approach: Confidence Contours

Represent the "bounds" of structural uncertainty



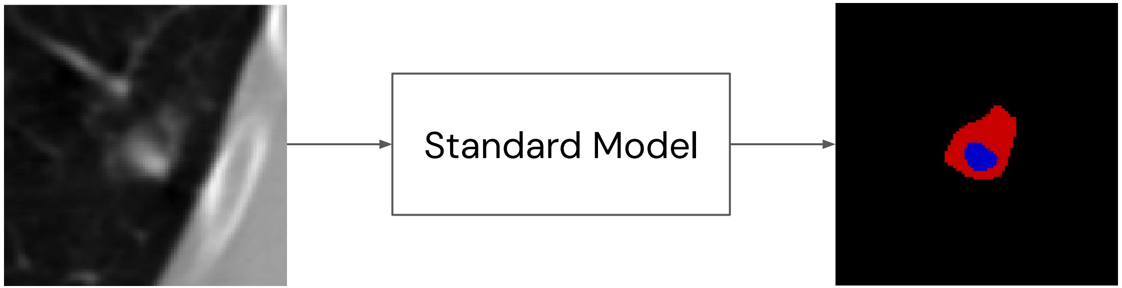
Step 1

Draw min

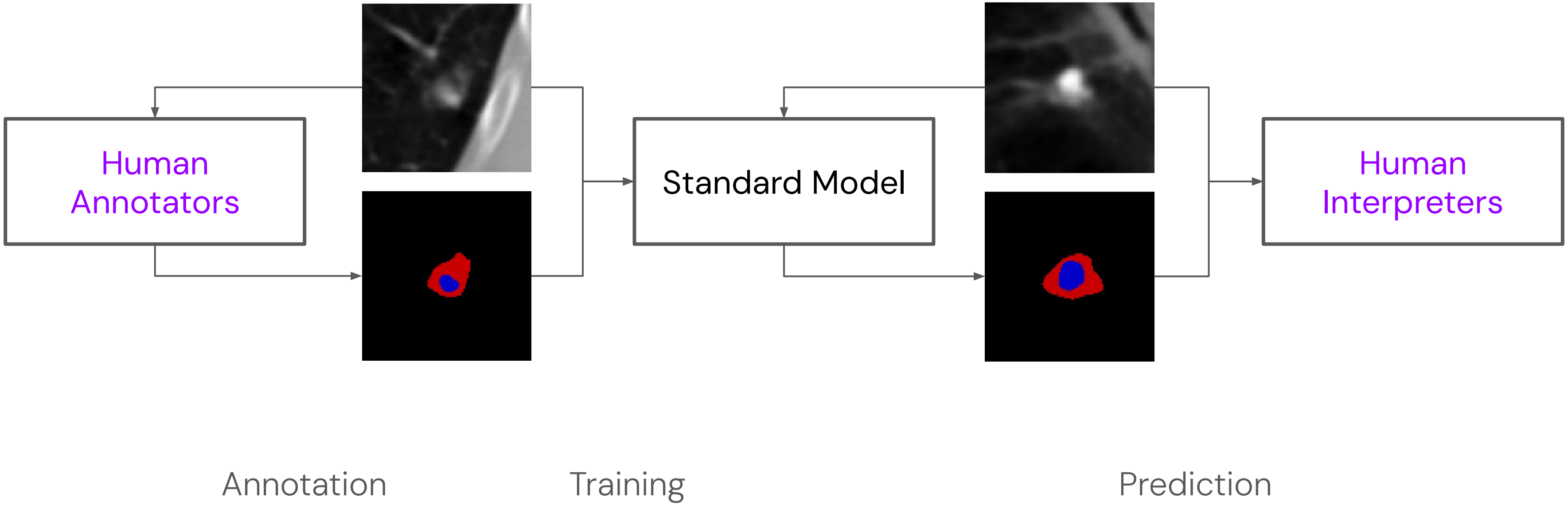
Step 2

Draw max

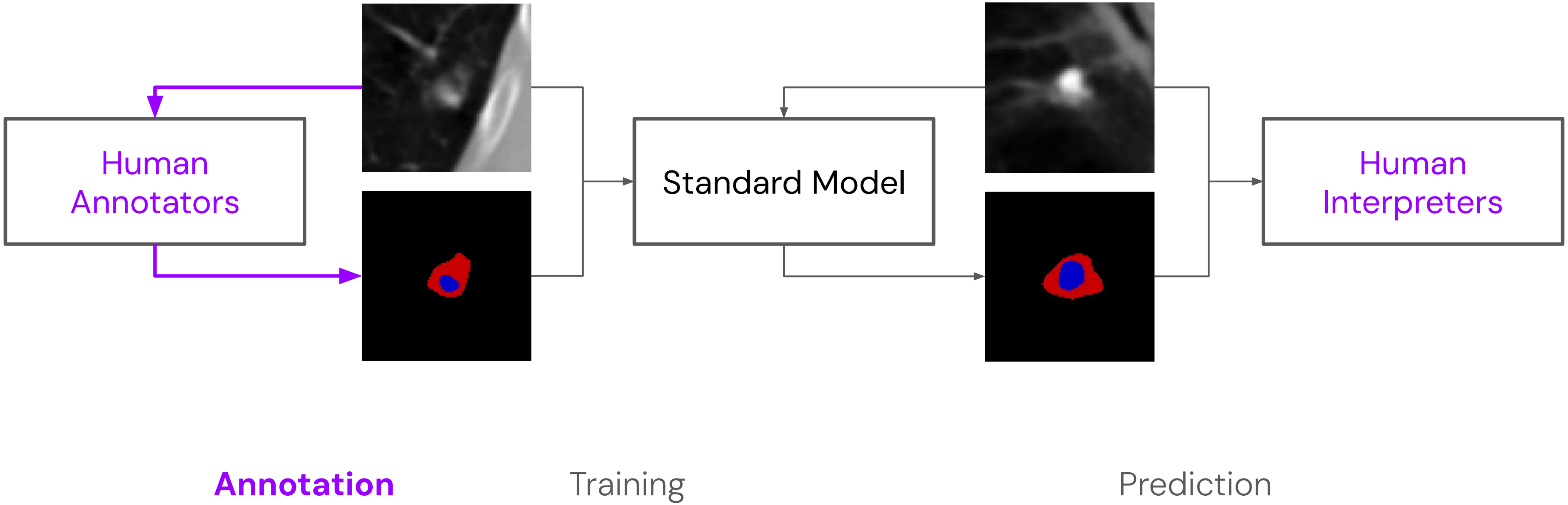
Training models on Confidence Contours requires no model modifications.



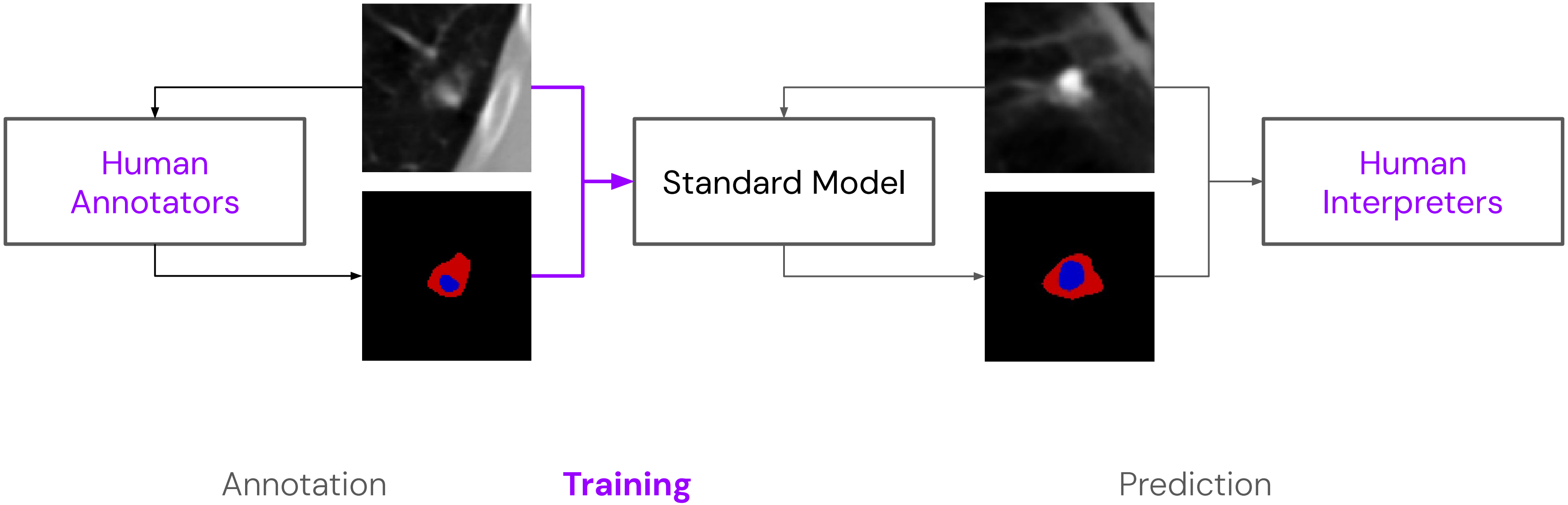
Confidence Contours recenters the humans at both sides of the uncertainty modeling pipeline.



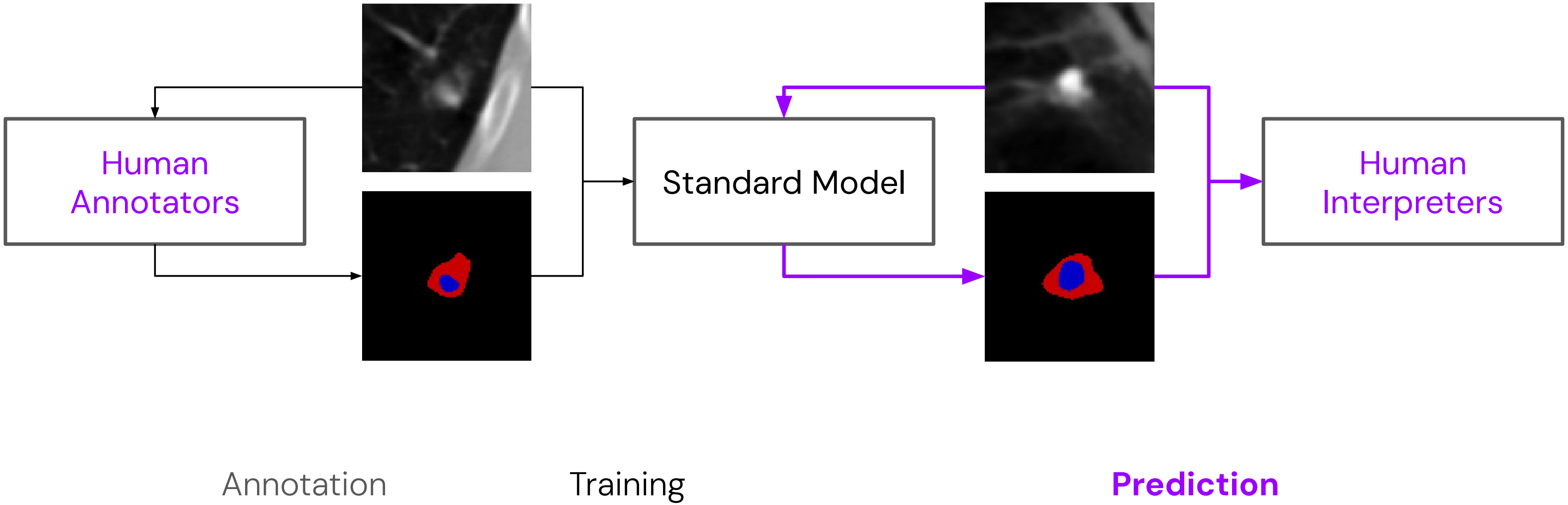
Human annotators directly mark uncertainty in the image with minimally more effort.



Models are simply trained by predicting two rather than one segmentation maps; no bells & whistles needed



All uncertainty information directly corresponds to human annotations. No black-box uncertainty inferences!



[3/3]

Evaluating Confidence Contours

Experiments

Annotation

- 45 annotators, 600 images, 2 datasets (LIDC, FoggyBlob)
- 3 CC and 3 singular annotations for each image

Modelling

- Fitted standard, Confidence Contours, and Bayesian models

Interpretation / Use

- Interviewed 5 experts on the utility of model predictions

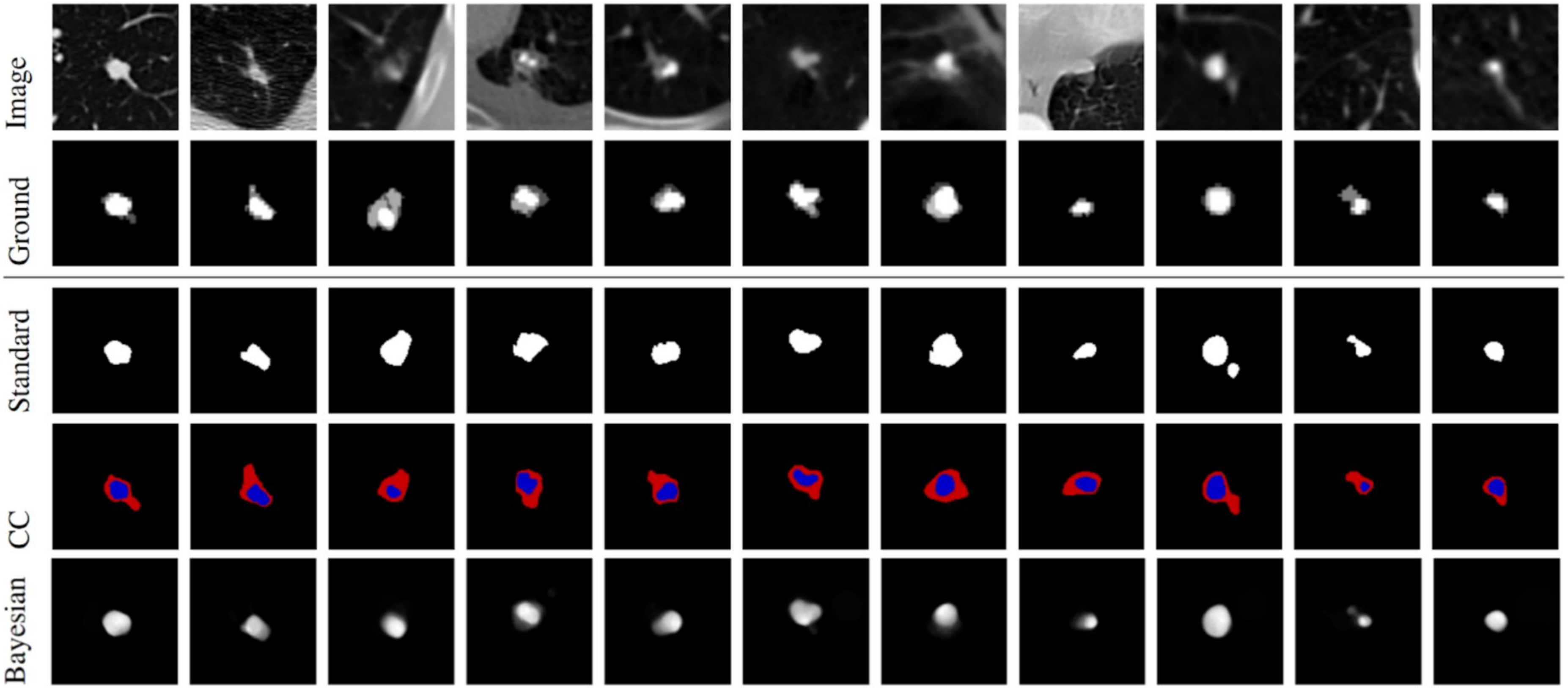
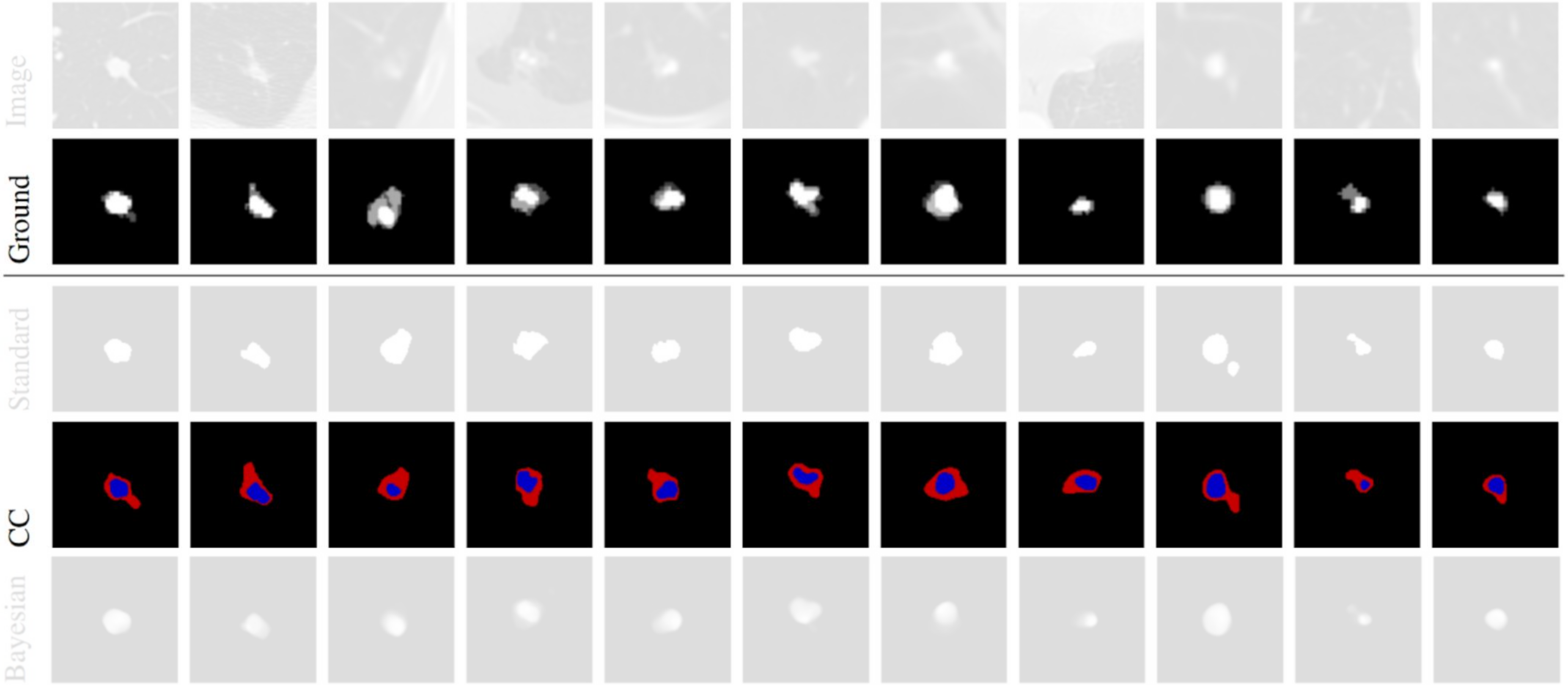
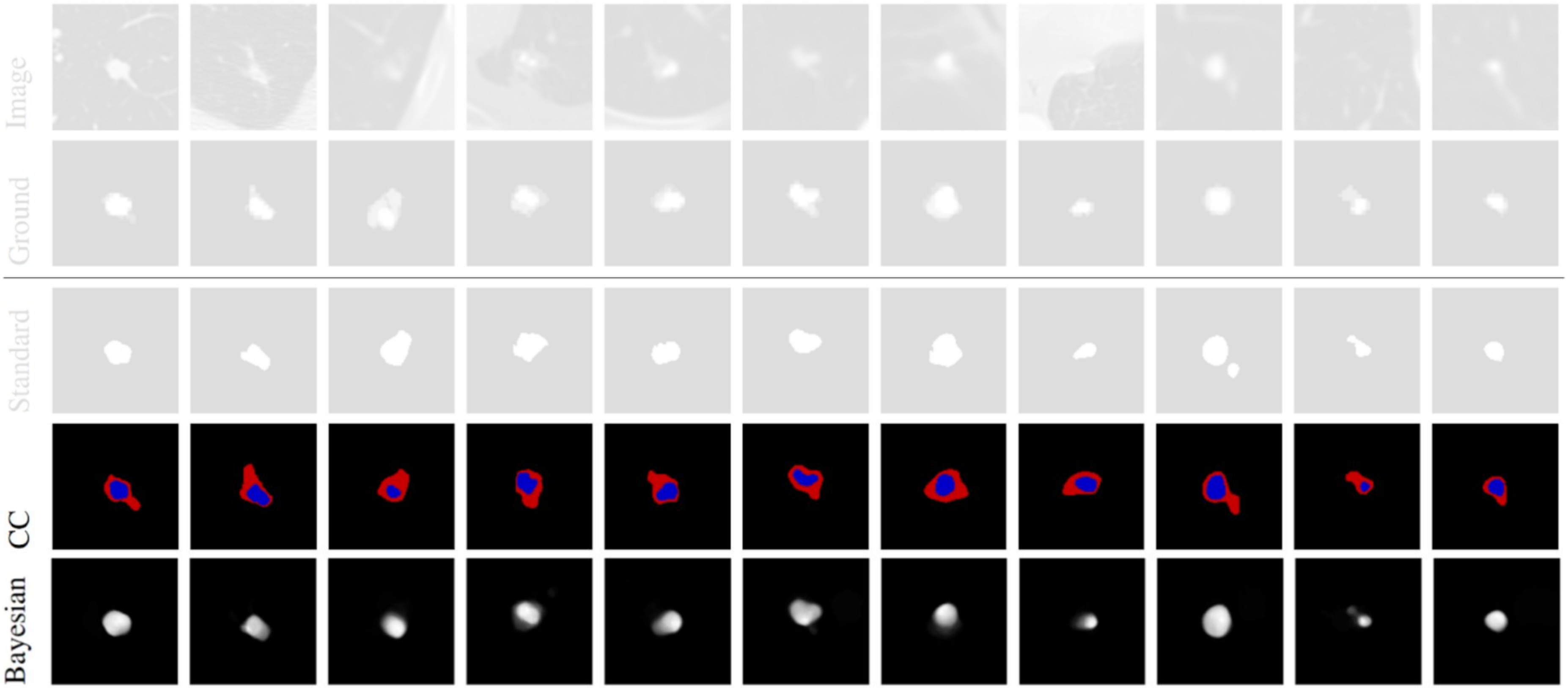


Image
Ground
Standard
CC
Bayesian

CC's bound ground truth disagreement (and give a little more)



It’s easier to make substantive conclusions about uncertainty with CCs



#1: Convenient to annotate

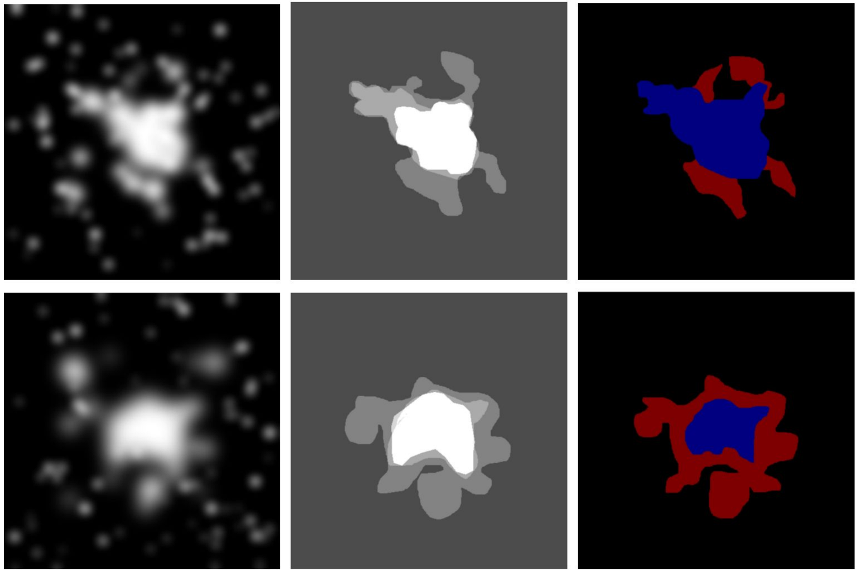
Annotators find CCs more demanding (as expected), but not by much

Dimension	LIDC		FoggyBlob	
	Singular	CC	Singular	CC
Mental Demand	3.7	*4.9	3.3	*4.6
Physical Demand	2.7	3.3	3.9	3.7
Temporal Demand	4.2	*4.9	5.0	5.5
Performance	6.9	6.9	6.8	6.9
Effort	4.8	*5.7	5.0	5.1
Frustration	3.0	*4.2	2.7	*4.0

Table 1: Average annotator responses across six dimensions and two datasets on the experience annotating using singular and CC methods, evaluated on a 10 point scale (1="very low", 10="very high"). * indicates a statistically significant relationship, measured with a relative t-test by annotator.

#2: Informative

- Annotations correspond to direct human judgements
  - …not model abstractions
- CC's bound the range of disagreement 30.8% better than avg. singular annot.
- Disagreement is decomposed
  - Min: 13.2% decrease
  - Max: 5.6% decrease



FoggyBlob Image | Composited singular annot's | One CC annot.

#3: Concise

- CCs are only two discrete masks – easy to read
- "The problem with [continuous maps] if I were looking at it just with my eye is that it's really difficult to tell the certainty level… it would be nice to have some range or threshold" [P1]
- "[CC] is easier to understand because I feel like the [min] contour is something which is more reliable that you can fall back on, and you can use the [max] contour if it makes sense to, given the situation. But you don't get those two channels of information in [continuous maps]." [P5]

Conclusion: Broader Themes

1. Ground truth → “Ground truths”
2. Identifying & centering human needs
3. Data-centrism & model-centrism

Acknowledgements

My mentor, Dr. Quanze (Jim) Chen

- Jim's earlier work in scalar uncertainty annot. (Goldilocks) inspired this project

My advisor, Prof. Amy X. Zhang

Summary

1. Medical segmentation is a high-stakes & structurally uncertain task.
2. We should model structural uncertainty rather than "collapsing" it.
3. Uncertainty rep's should be easy to annot., informative, concise
   a. Methods which change the model struggle w/ last 2 properties
4. Confidence Contours: bound disagreement w/ min and max contour.
   a. Uncertainty is directly annotated, rather than abstractly inferred
   b. CCs satisfy all 3 properties
   c. Medical experts find CCs more usable in judgements

andreye@uw.edu
andre-ye.github.io

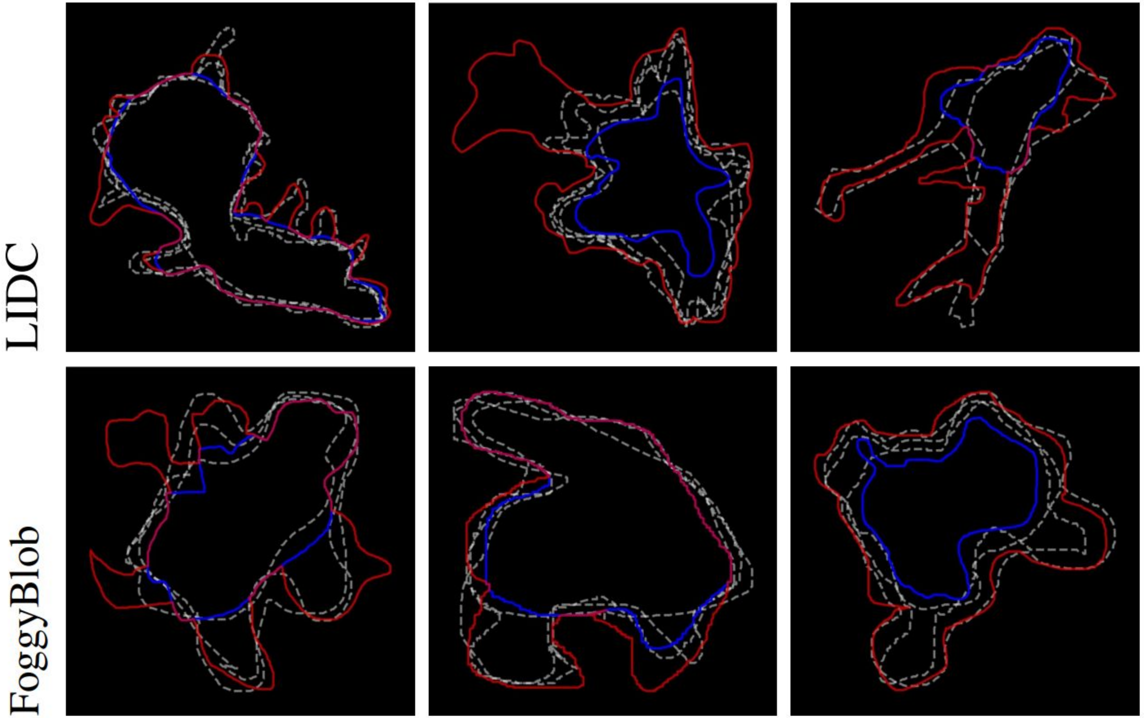
my website

bitly

arXiv paper

bitly

ML Bonus: CC annot's give more positive info



Thank you!

andreye@uw.edu
andre-ye.github.io

Conclusion: Recapping broader themes

- Prioritizing human users and their perceptual limitations / behavior
- Data-centric over model-centric approach
- Rethinking the ground truth