Uncertainty-Aware Training of Neural Networks for Selective Medical Image Segmentation

Yukun Ding[1], Jinglan Liu[1], Xiaowei Xu[2], Meiping Huang[2], Jian Zhuang[2], Jinjun Xiong[3], Yiyu Shi[1]

[1] University of Notre Dame, [2] Guangdong General Hospital, [3] IBM

Medical Imaging with Deep Learning (MIDL) 2020

UNIVERSITY OF NOTRE DAME

Overview

- Background
- Motivation
- Method
- Results
- Limitation and Future Work

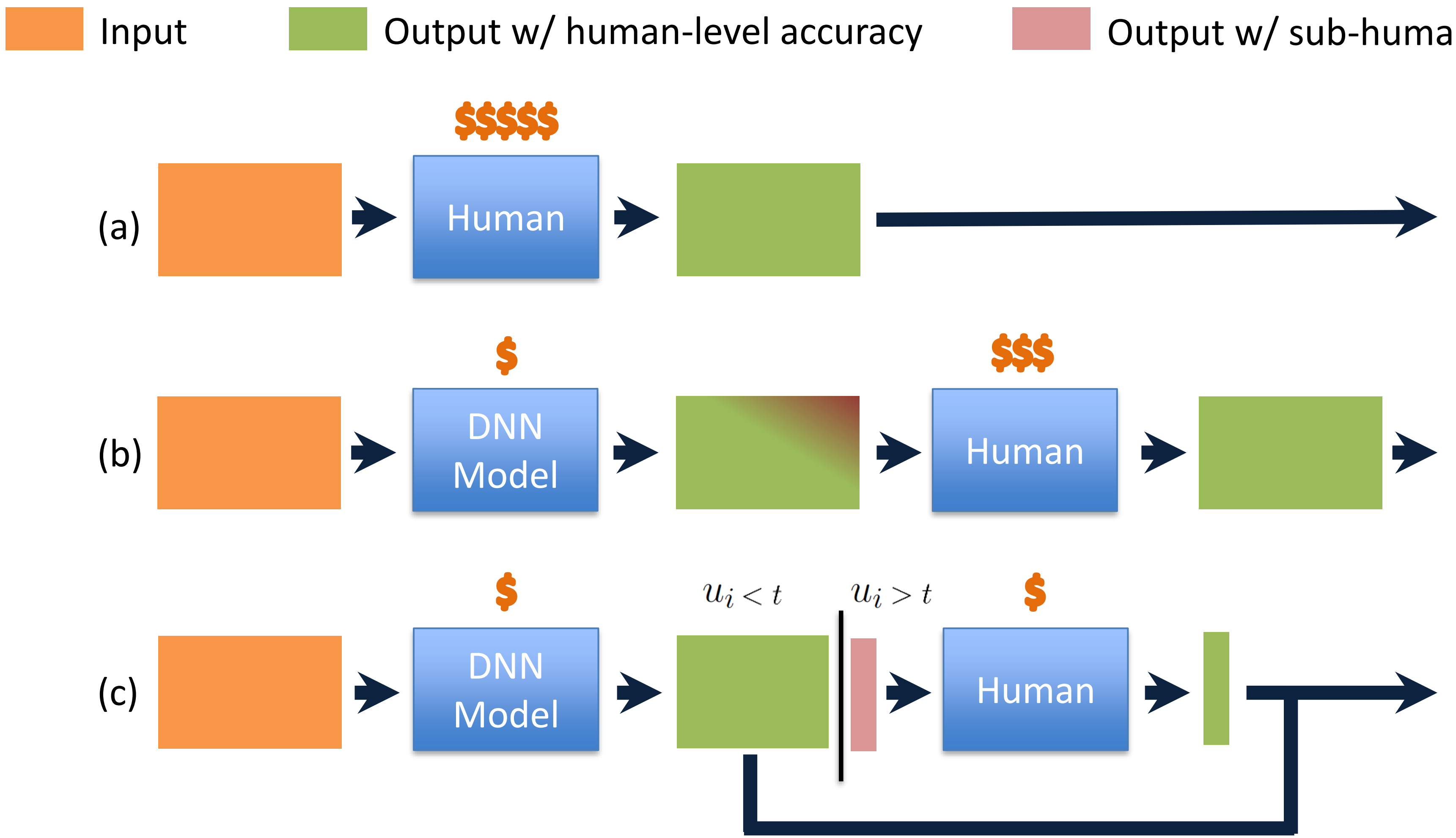
Uncertainty of DNNs

- Why we need to consider the uncertainty?
  - Real-world problems are diverse
  - Identify and deal with potential failure properly

- The word "uncertainty" can be tricky e.g.,
  - This is a tumor, but I think there is a 30% of chance I'm wrong
  - This is a tumor, rotate the image a bit -> this is not a tumor

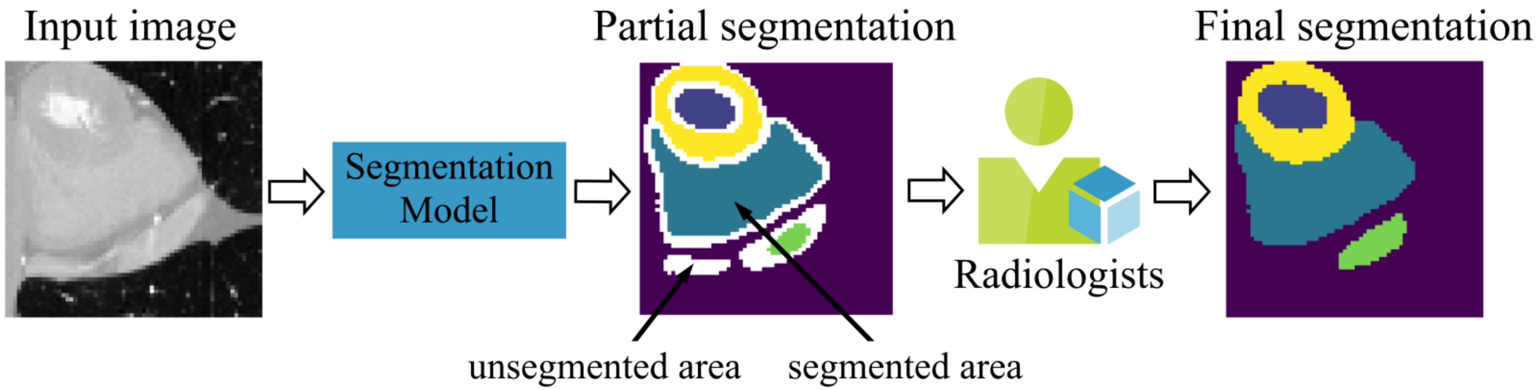
- What uncertainty are we consider here?
  - For each input x_i , the model outputs prediction $\hat{y}_i$, and the uncertainty score u_i
  - The uncertainty score u_i indicates how likely the prediction is wrong
  - A popular baseline of uncertainty estimation: 1 - (softmax probability)

Selective Prediction

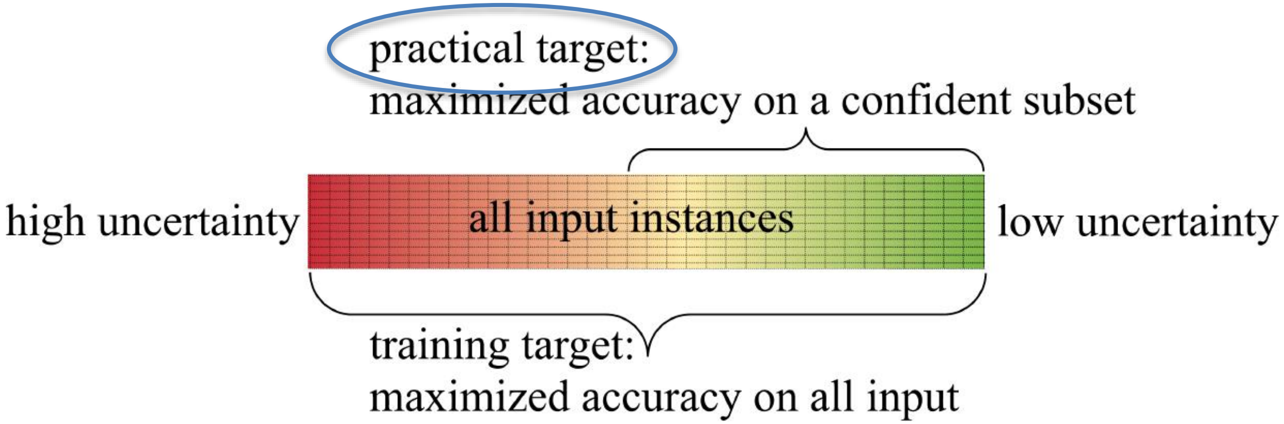


Motivation

- Selective segmentation



- The *practical target* and *training target*:



Problems Definition

- For each input $x_i \in X$, model outputs prediction $\hat{y}_i$, and the uncertainty score u_i , the correctness score $s_i = 1$ if the prediction $\hat{y}_i$ is correct, otherwise $s_i = 0$
- If we apply a threshold on the uncertainty, we divide the input data into two subset $X_l = \{x_i | u_i \leq t\}$ and $X_h = X - X_l$, the coverage is defined as $c = \frac{|X_l|}{|X|}$
- Consider the accuracy at coverage c

$$\psi_c = \frac{\sum_{x_i \in X_l} s_i}{|X_l|} \qquad \psi_1 = \frac{\sum_{x_i \in X} s_i}{|X|}$$

- Our practical target, accuracy at coverage c, depends on both the quality of prediction and the quality of uncertainty estimation
- We know how to optimize our neural network for prediction, but not for uncertainty

From the Scoring Rule Perspective

- Estimating the uncertainty is a probabilistic prediction problem
- Scoring rule:
  - A quantified summary measure for the quality of probabilistic predictions
- Proper scoring rule:
  - Denote the truth distribution as q and the predicted distribution as p_θ, a scoring rule h is a proper scoring rule if $h(p_\theta, q) \leq h(q, q)$
- Strictly proper scoring rule:
  - Same as the proper scoring rule, but $h(p_\theta, q) = h(q, q)$ if and only if $p_\theta = q$
- Commonly used loss functions are strictly proper scoring rule
  - E.g., Cross Entropy, L2
  - This is why softmax probability can be a strong baseline for uncertainty estimation

Observation

For the uncertainty estimation in selective segmentation, we do not need a strictly proper scoring rule that tries to recover the actual distribution q.

- The uncertainty score u is only used to divide the data into two subset, we only want more correct predictions go to the low uncertainty subset and more wrong predictions go to the high uncertainty subset.
- Even if we consider all possible coverage, only the relative ranking of u matter and we don’t care the specific value of u.
- So we try to find a better optimization target that is not a strictly proper scoring rule.

The Uncertainty Target

Theorem 1 *Denote* $\gamma = \frac{\sum_{x_i \in X_h}(1-s_i)}{|X_h|}$*, then we have the following properties about* ψ_c*,* ψ_1*, and* γ *for* $c \in (0, 1]$*:*

(i) γ *is a proper scoring rule but not a strictly proper scoring rule for uncertainty estimation.*

(ii) $\psi_c = \frac{\psi_1 - (1-\gamma)(1-c)}{c}$*, s.t.* $\quad (1-\gamma)(1-c) \in [\psi_1 - c, \psi_1]$.

(iii) $\frac{\partial \psi_c}{\partial \gamma} > 0$ *and* $\frac{\partial \psi_c}{\partial \psi_1} > 0$ *for any* γ *and* c.

- Why γ :
  - γ is a proper scoring rule but not a strictly proper scoring rule
  - γ fully determines ψ_c with ψ_1
  - The partial derivative is always positive

Uncertainty-Aware Training

- How to optimize γ ?

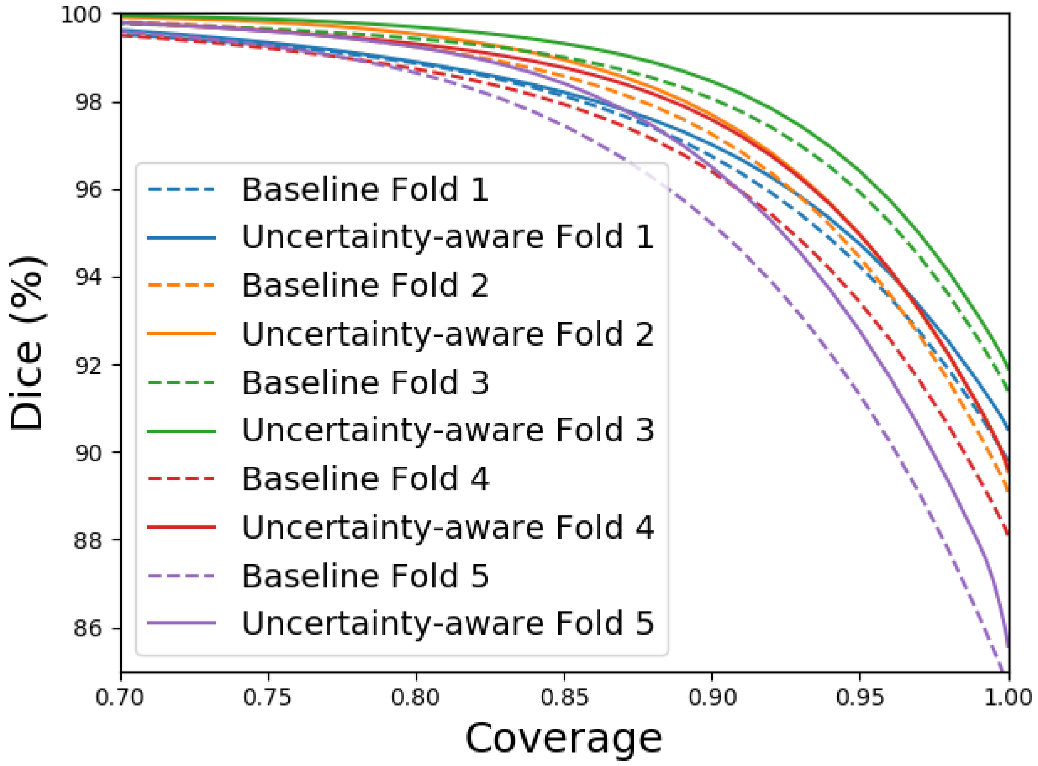
$$\mathcal{L}_{uncertainty} = \sum_{u_j \in U_w, u_k \in U_c} \max(u_k - u_j + m, 0)$$

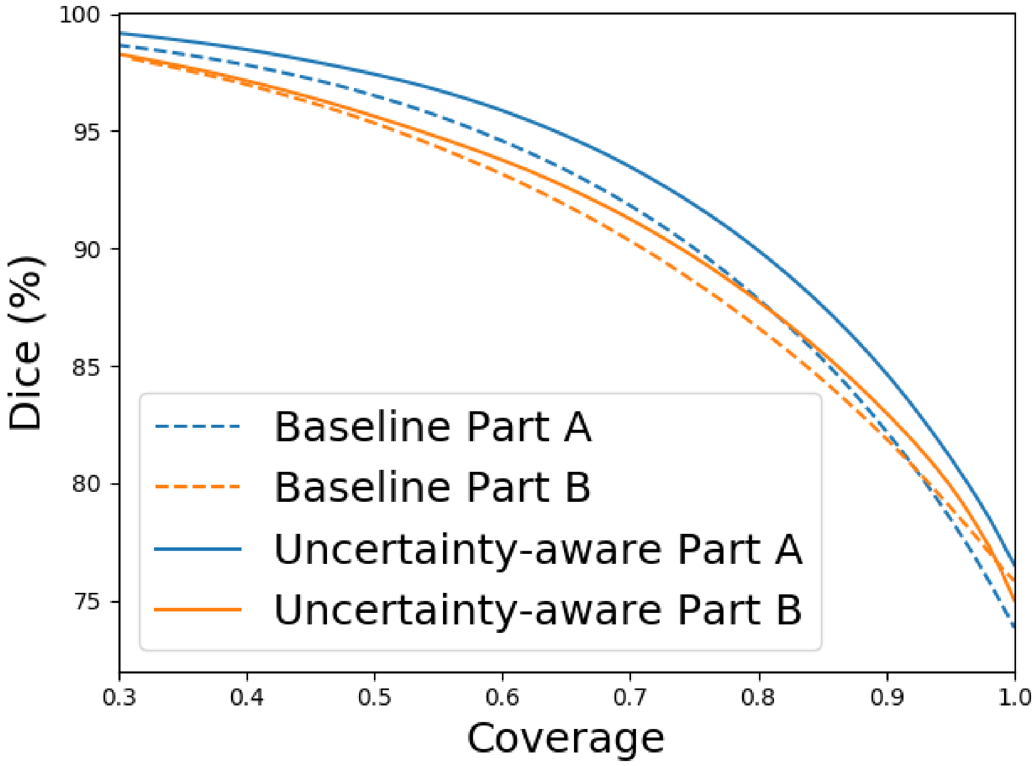
- The uncertainty-aware training loss:

$$\underbrace{\mathcal{L}_{u\text{-}seg}}_{\psi_c} = \underbrace{\mathcal{L}_{segmentation}}_{\psi_1} + \underbrace{\lambda \mathcal{L}_{uncertainty}}_{\gamma}$$

The Dice-Coverage Curve

- Reduced coverage leads to higher accuracy
- Uncertainty-aware training outperforms the baseline

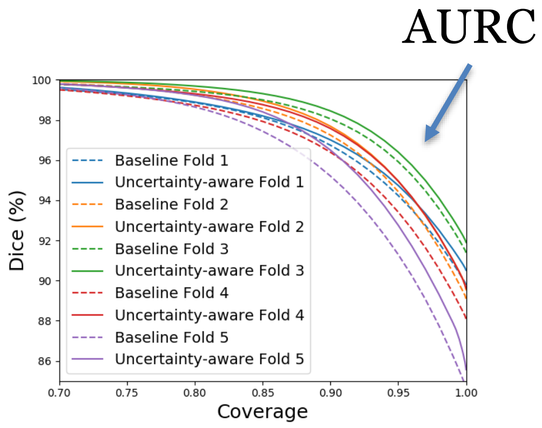




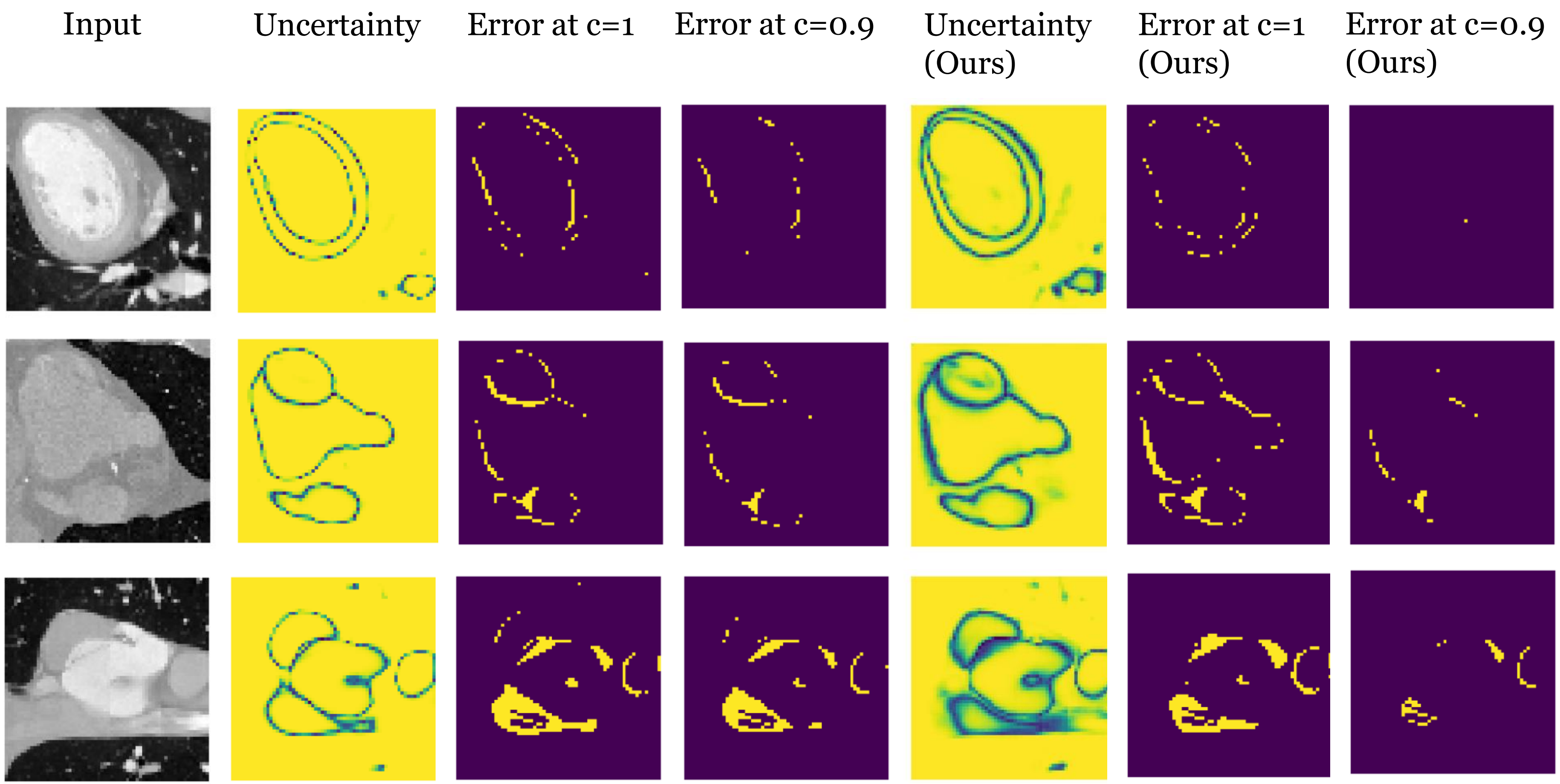
Quantitative Results

- Reduced coverage leads to higher accuracy
- Uncertainty-aware training outperforms the baseline

Dataset	AURC (%)		Coverage	Dice (%)		Dice@5PCTL (%)	
	Baseline	Ours		Baseline	Ours	Baseline	Ours
MM-WHS	0.936	**0.810**	0.95	93.86±3.66	**94.72±2.53**	83.18	**89.19**
			0.90	96.73±2.30	**97.35±1.54**	90.55	**94.64**
			0.80	98.98±0.79	**99.21±0.61**	97.42	**98.11**
			0.70	99.64±0.31	**99.72±0.31**	98.98	**99.07**
GlaS	6.981	**6.031**	0.95	78.62±16.87	**80.80±16.48**	35.99	**39.16**
			0.90	82.14±14.53	**84.26±13.87**	49.77	**52.46**
			0.80	87.54±12.31	**89.37±11.24**	66.63	**67.90**
			0.70	91.45±10.86	**92.92±9.55**	**75.95**	74.81

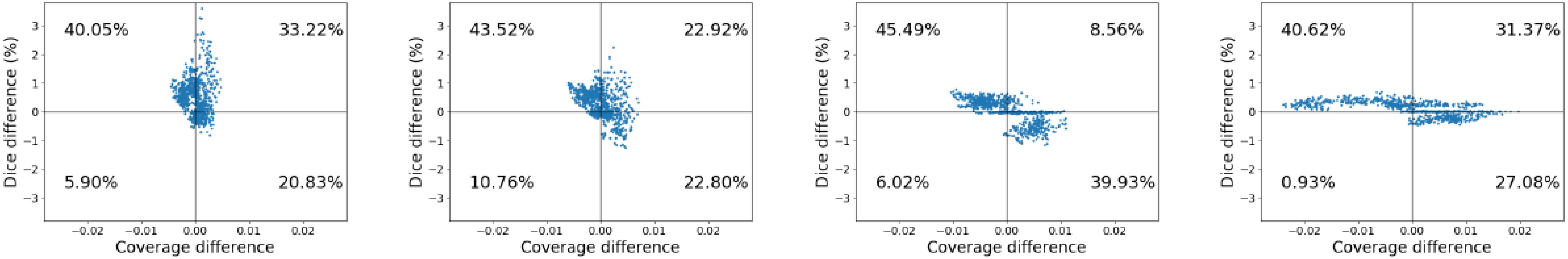


Qualitative Results



Per-Image Comparison

- The performance is improved by uncertainty-aware training
- With decreasing average coverage
  - Per-image coverage difference increases
  - Per-image Dice difference decreases



(a) c=0.95 (b) c=0.9 (c) c=0.8 (d) c=0.7

Limitation and Future Work

- It is not very efficient to do **pixel-wise** selective segmentation
  - We are currently looking at **image-wise** selective segmentation
  - Challenges: image-wise uncertainty measure; joint training

- γ is a proven good target, but the $\mathcal{L}_{uncertainty}$ is not
  - A better loss to optimize γ ? Or even directly optimize ψ_c ?

- Thank You!
- Q&A