IPMI 2021
Rønne, Bornholm

Segmentation with Multiple Acceptable Annotations: A Case Study of Myocardial Segmentation in Contrast Echocardiography

Dewen Zeng[1], Mingqi Li[2], Yukun Ding[1], Xiaowei Xu[2], Meiping Huang[2], Jian Zhuang[2], and Yiyu Shi[1]
[1] Department of Computer Science and Engineering, University of Notre Dame, South Bend, IN, USA
[2] Guangdong General Hospital, Guangzhou, Guangdong, China

INTRODUCTION

Background and Motivation

- Large inter-observer variability exist among myocardial annotations of different cardiologists. (see Fig. 1 and Table 1.)
- Annotation variations caused by human factors (e.g., training and expertise) can be addressed by learning the behaviour of the annotators, a unique ground truth can be obtained by majority vote of experienced annotators during evaluation [1].
- For variations caused by low image quality (e.g., low resolution and significant artifact), the unique ground truth may not be available because we don't know whose annotation is better.
- Although the annotations are different, they are usually all acceptable when used in myocardial perfusion analysis (clinically valuable).

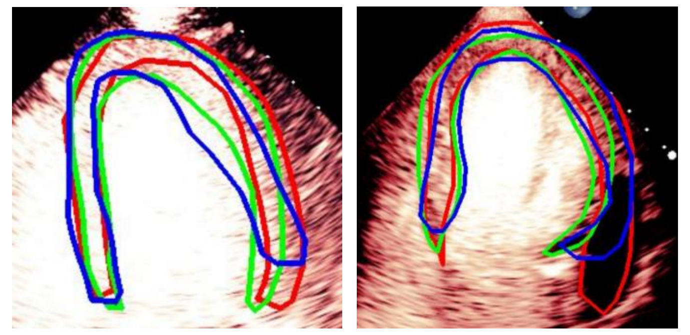

Fig. 1. Visualization of annotations from three experienced cardiologists

	# 1	# 2	# 3	# 4	# 5
# 1	1				
# 2	0.898	1			
# 3	0.844	0.849	1		
# 4	0.783	0.790	0.800	1	
# 5	0.803	0.807	0.814	0.787	1

Table 1. Average Dice of annotations from 5 cardiologists calculated mutually (on 180 images).

Target Questions

- How to make the neural networks more robust when multiple acceptable annotations exist during training?
- Without the unique ground truth during evaluation, how to evaluate whether the segmentation generated by the model is good or not?

Our Work

- Myocardial Contrast Echocardiography (MCE) dataset that has MCE images from 100 patients was collected. Each image is labeled by five experienced cardiologists.
- A new metric called the extended Dice is designed to effectively evaluate the quality of the segmentation with multiple accepted ground truths.
- We incorporate the extended Dice into the loss function to train the segmentation network.

METHODOLOGY

Basic Idea

- Given myocardial boundary annotations of two cardiologists, obtaining an inner boundary (intersection of the two annotations) and an outer boundary (union of the two annotations). Then, an acceptable region (the region between the inner boundary and outer boundary) can be obtained.
- The pixels inside the acceptable region can be classified as either myocardium or background. Any prediction boundary that completely falls inside the acceptable region shall be considered acceptable.

Extended Dice

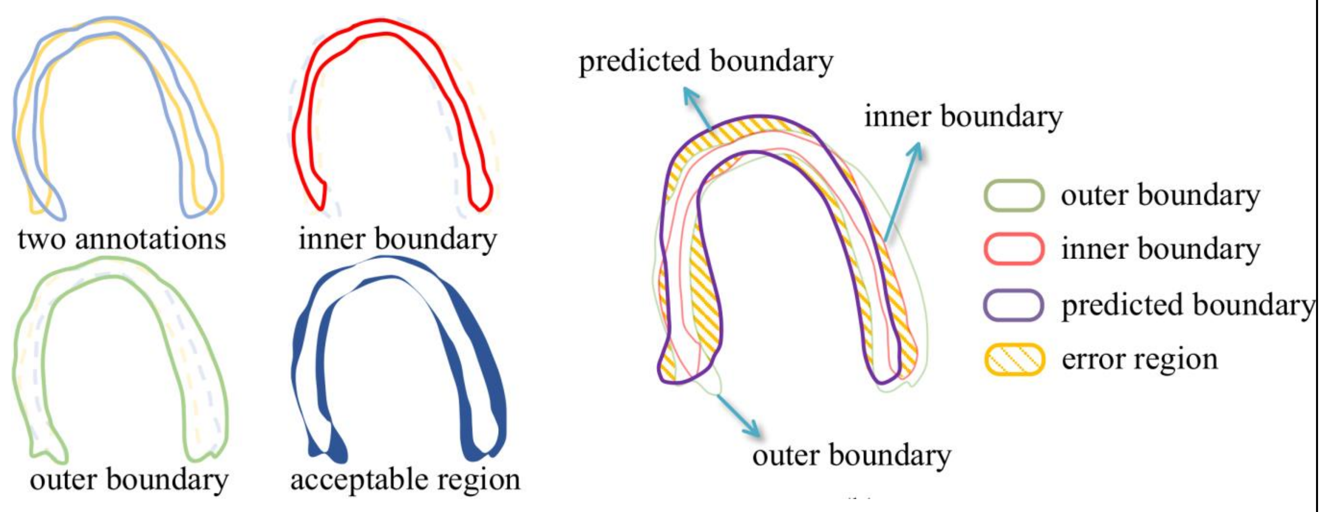

Fig. 2. Visual illustration of computing extended Dice

$$ExtendedDice(P,I,O) = 1 - \frac{(P - P \cap O) + (I - P \cap I)}{P + I}$$

- P is the predicted boundary, O is the outer boundary, I is the inner boundary.
- In light of the existence of multiple acceptable annotations, the acceptable region can allow the prediction boundary to have some flexibility in regions where significant large inter-observer variability exists.

Relationship to Dice

- Simplified extended Dice

$$ExtendedDice(P,I,O) = \frac{(P \cap O) + (P \cap I)}{P + I}$$

- When multiple cardiologists give the same annotations, the inner boundary and outer boundary completely overlap. The extended Dice shrinks to conventional Dice.

EVALUATION

Evaluation Using Conventional Metrics

- Using one of the cardiologists' annotation as ground truth or the majority vote as the ground truth for evaluation.
- Evaluation metric: Dice, IoU and Hausdorff Distance (HD).
- The model trained with extended Dice shows high Dice, IoU and lower HD compared to all baselines.

Method	GT: cardiologist 2			GT: majority vote		
	Dice	IOU	HD	Dice	IOU	HD
Single cardiologist (SC)	0.809	0.694	32.8	0.838	0.735	28.4
Consensus	0.824	0.719	31.7	0.847	0.753	28
Average cross entropy (ACE)	0.819	0.709	29.4	0.844	0.745	26.4
Confusion matrix (CM) [2]	0.808	0.695	40.4	0.826	0.719	37.9
Consistency [1]	0.826	0.719	32.3	0.847	0.749	29.8
STAPLE [3]	0.81	0.694	33	0.814	0.695	31.8
Proposed	**0.829**	**0.721**	**28.9**	**0.855**	**0.759**	**25.4**

Evaluation Using Extended Dice

- All annotations from five cardiologists are used to compute the extended Dice.
- Network architecture: U-net and Deeplab V3+
- The model trained with extended Dice consistently show higher extended Dice.

Architecture \ Method	SC	Consensus	ACE	CM	Consistency	STAPLE	Proposed
U-net	0.929	0.94	0.919	0.951	0.947	0.912	**0.958**
DeeplabV3+	0.942	0.946	0.945	0.924	0.944	0.921	**0.954**

Evaluation Using Frame-intensity Curve

- Frame-intensity curve is used to measure the relative microvascular blood volume after microbubbles infusion.
- The curve generated by extended Dice is closer to the ground truth curve (generated by averaging the curve of all cardiologists). See Fig. 3.

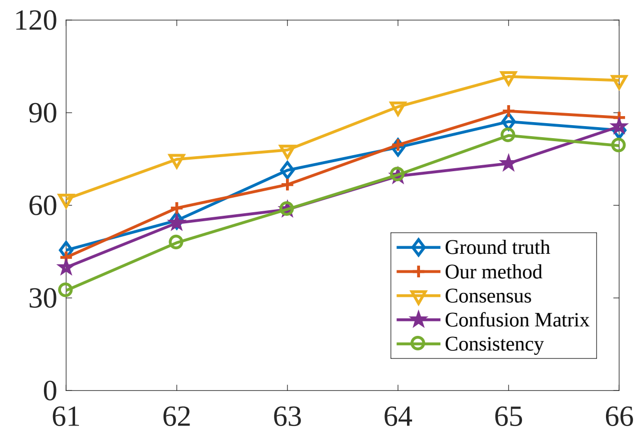

Fig. 3. Frame-intensity curve generated by using segmentations of different methods

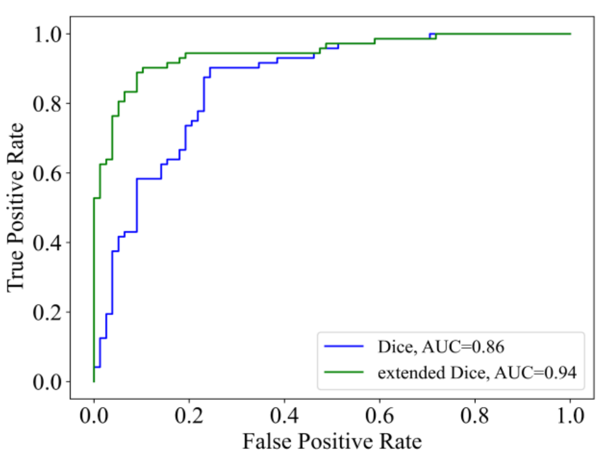

Fig. 4. ROC curve of using Dice and extended Dice to classify segmentations

EVALUATION (CONT.)

Extended Dice as a Superior Evaluation Metric

- Compare extended Dice and Dice by using these two metric to decide whether the segmentation results generated by the model need manual correction (binary classification).
- The class label is generated by an experienced cardiologist by looking at each segmentation result.
- Extended Dice improves the AUC from 0.86 to 0.94 (shown in Fig. 4).

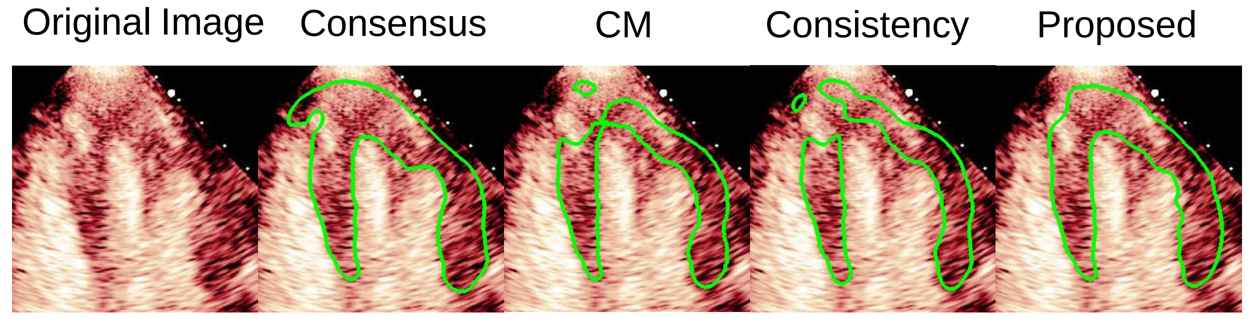

Fig. 5. Qualitative comparison of segmentations using U-net with different methods

EVALUATION (CONT.)

Conclusions

- Large inter-observer variability in MCE annotations is a critical issue that needs to be solved for deep learning based myocardial segmentations.
- The designed new extended Dice that considers multiple acceptable annotations is more accurate and robust for evaluating the segmentation results generated by DNNs.
- Using extended Dice in the loss function for training DNN can help the network better learn the general features of myocardium and ignore variations caused by individual annotators, which leads to improved segmentation performance.

Reference

[1] Sudre C H, Anson B G, Ingala S, et al. Let's agree to disagree: Learning highly debatable multirater labelling[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019: 665-673.
[2] Tanno R, Saeedi A, Sankaranarayanan S, et al. Learning from noisy labels by regularized estimation of annotator confusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 11244-11253.
[3] Warfield S K, Zou K H, Wells W M. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation[J]. IEEE transactions on medical imaging, 2004, 23(7): 903-921.