# Myocardial Segmentation in Contrast Echocardiography with Multiple Acceptable Annotations

**Dewen Zeng***, Yukun Ding*, Meiping Huang+, Jian Zhuang+, and Yiyu Shi*

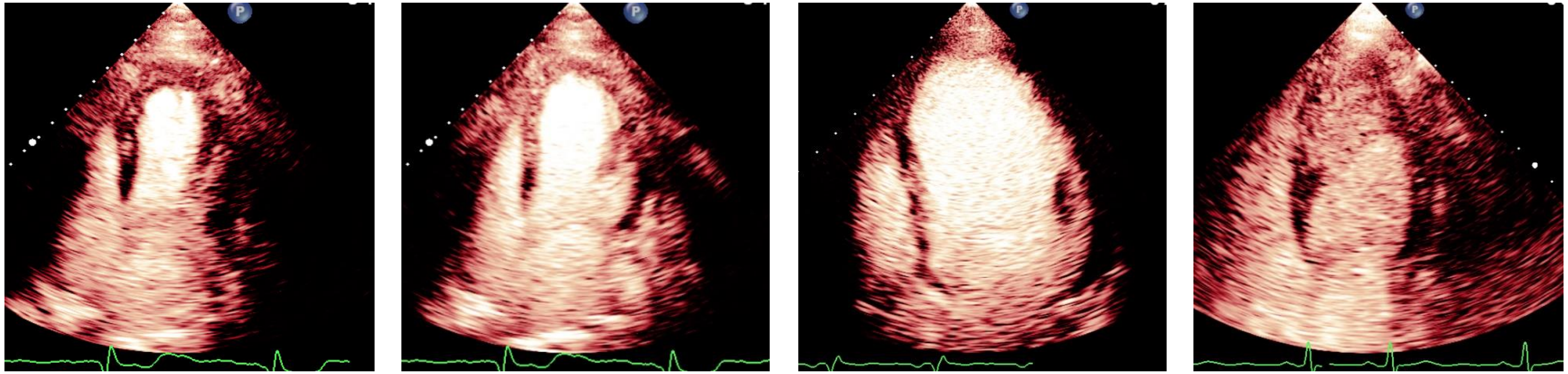
*University of Notre Dame

+Guangdong General Hospital

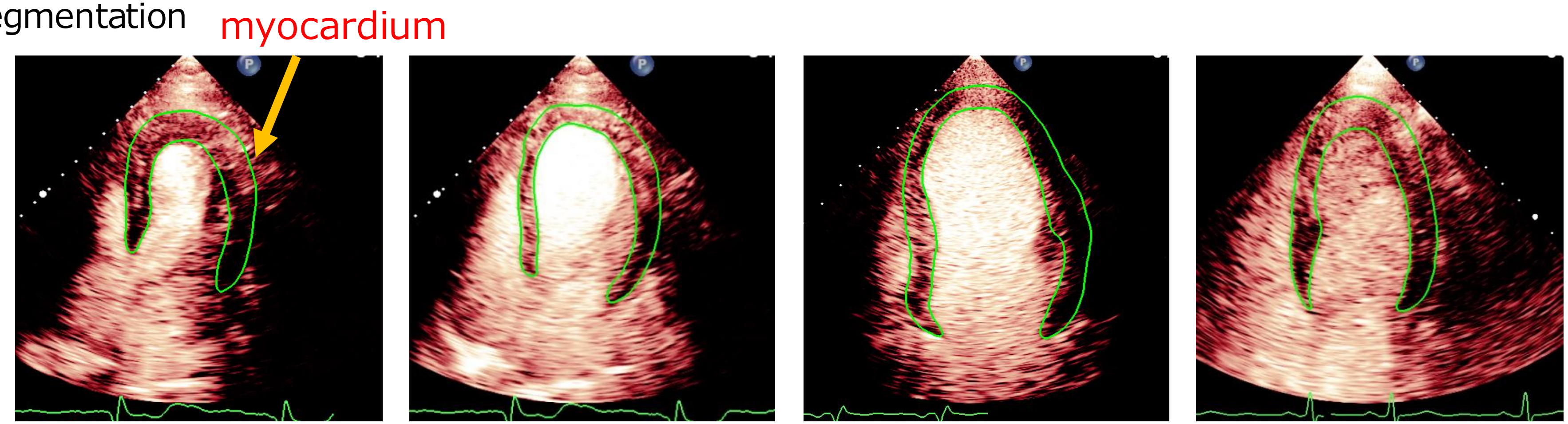
Information Processing in Medical Imaging (IPMI) 2021

# Problems

Contrast echocardiography: Ultrasound of the heart that is performed with some acoustically active particles for assessing left ventricle and myocardium function.

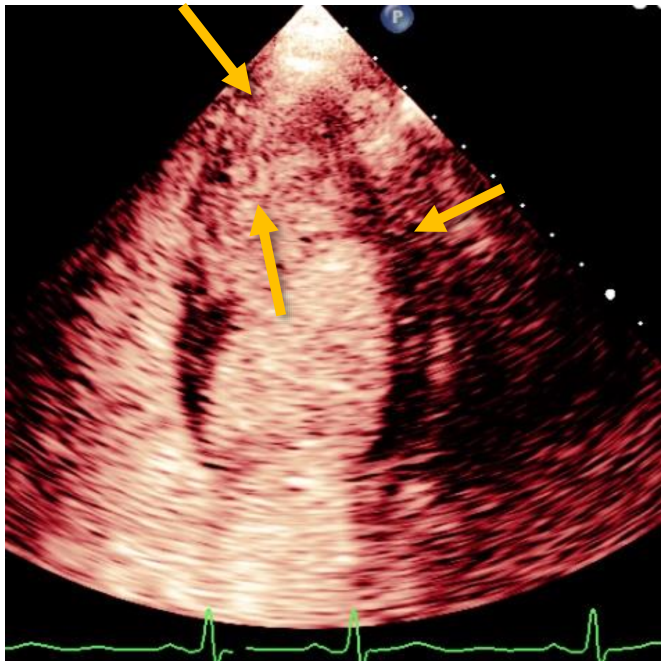


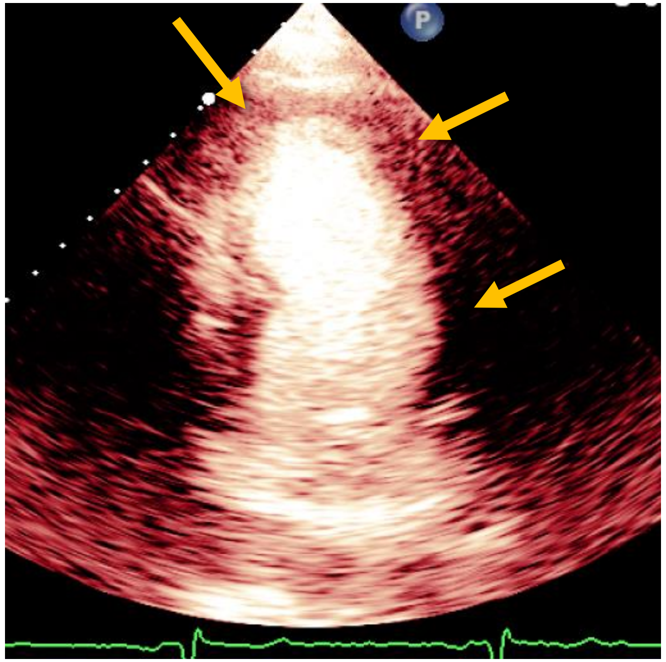
Myocardial segmentation



# Problems

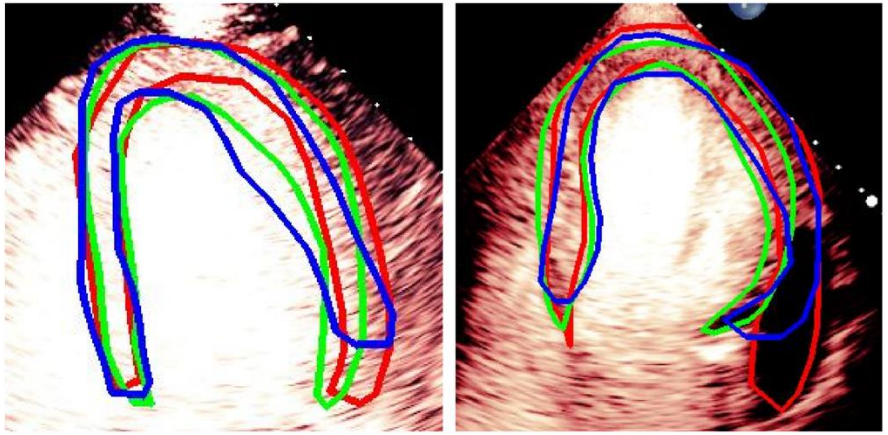
- Unique challenges in medical images (e.g., ultrasound)
  - Low signal-to-noise ratio & severe artifact
  - Large shape and pose variations of target organ or tissue





# Problems

- Radiologists annotate differently
  - Large inter-observer variability exists



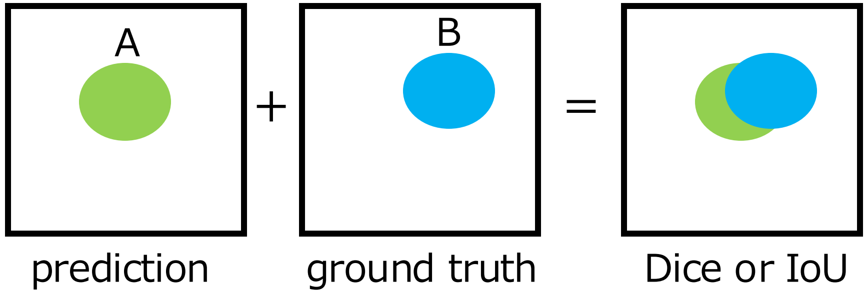
Myocardial annotations by three different radiologists

| | Radiologists 1 | Radiologists 2 | Radiologists 3 |
|---|---|---|---|
| Radiologists 1 | 1 | - | - |
| Radiologists 2 | 0.849 | 1 | - |
| Radiologists 3 | 0.790 | 0.800 | 1 |

Dice of the annotations of each radiologist using one of the others' as the ground truth

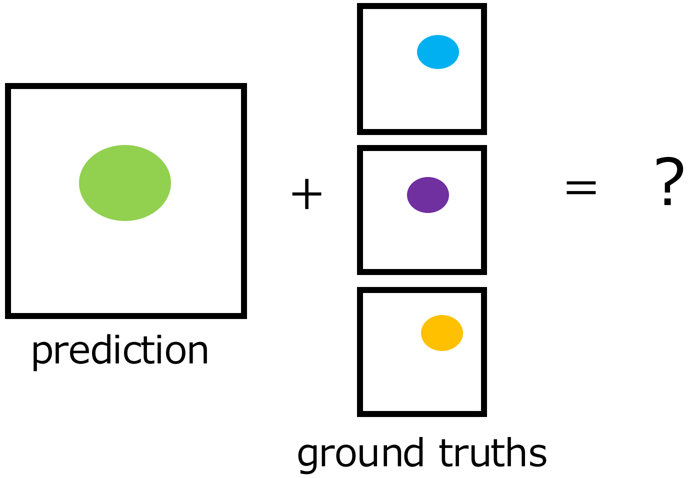
# Problems

- During Evaluation



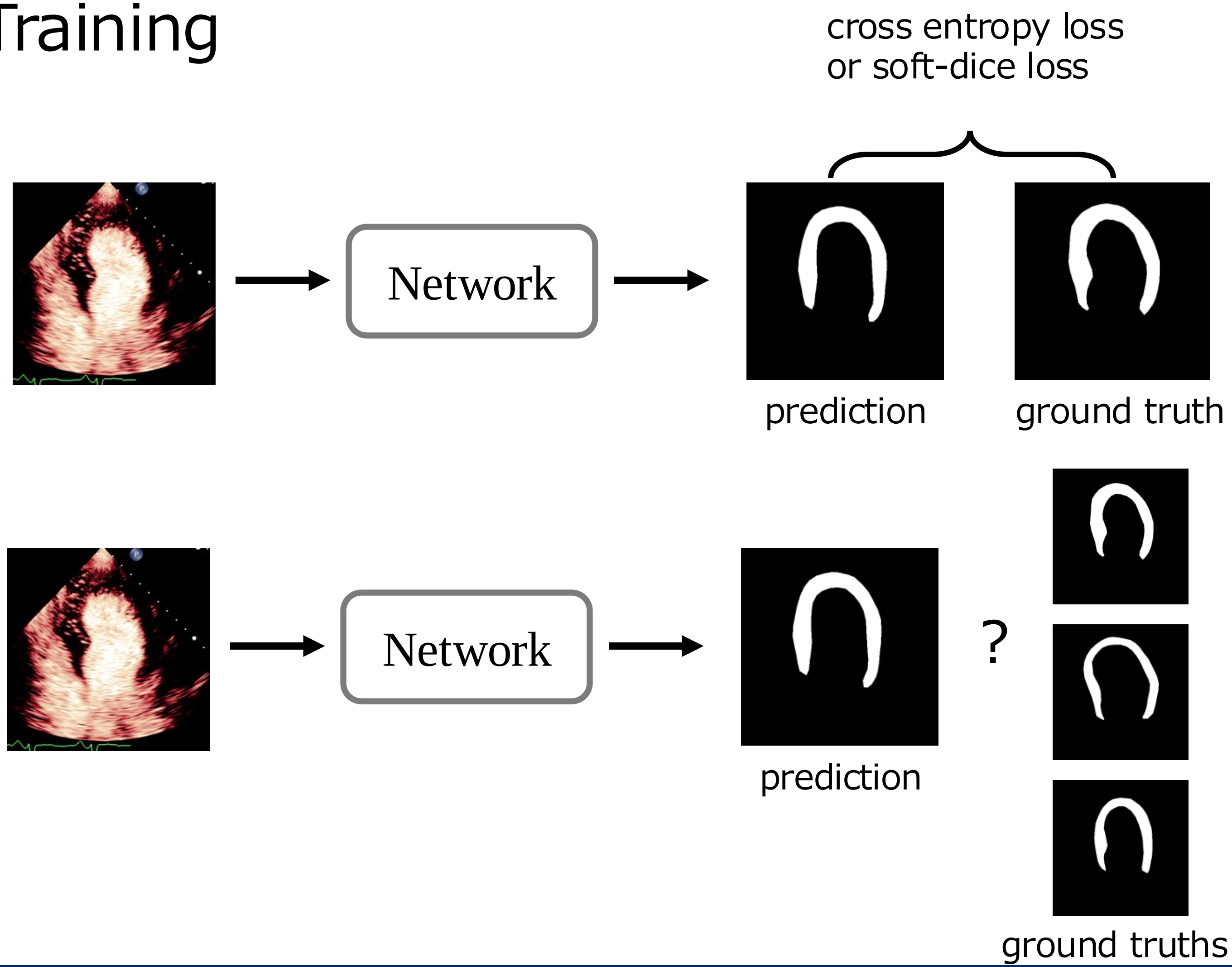

$$Dice = \frac{2 \times (A \cap B)}{A + B}$$

$$IoU = \frac{A \cap B}{A \cup B}$$



# Problems

- During Training

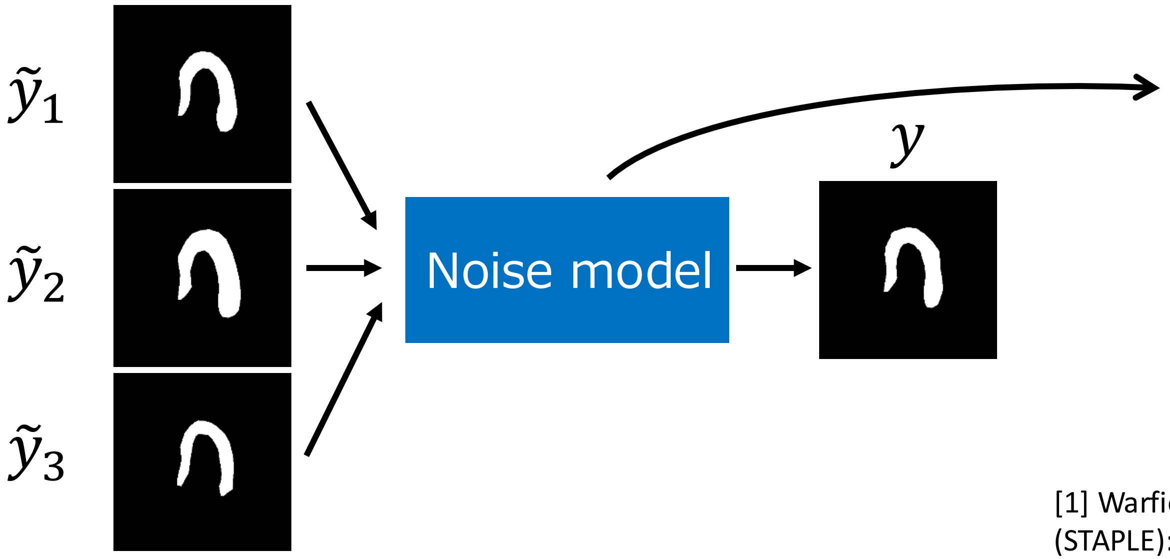


# Existing Work

- Assuming noisy distribution
  - $\tilde{y}$ (observed label) is dependent on $y$ (true label, we don't have)

$y$ → Noise model → $\tilde{y}$

True label — Observed label (noisy)




sensitivity specificity

Our goal is to estimate the performance level parameters of the experts characterized by parameters $(\mathbf{p}, \mathbf{q})$ which maximize the complete data log likelihood function

$$(\hat{\mathbf{p}}, \hat{\mathbf{q}}) = \arg\max_{\mathbf{p},\mathbf{q}} \ln f(\mathbf{D}, \mathbf{T} \mid \mathbf{p}, \mathbf{q}). \qquad (1)$$

observed label true label

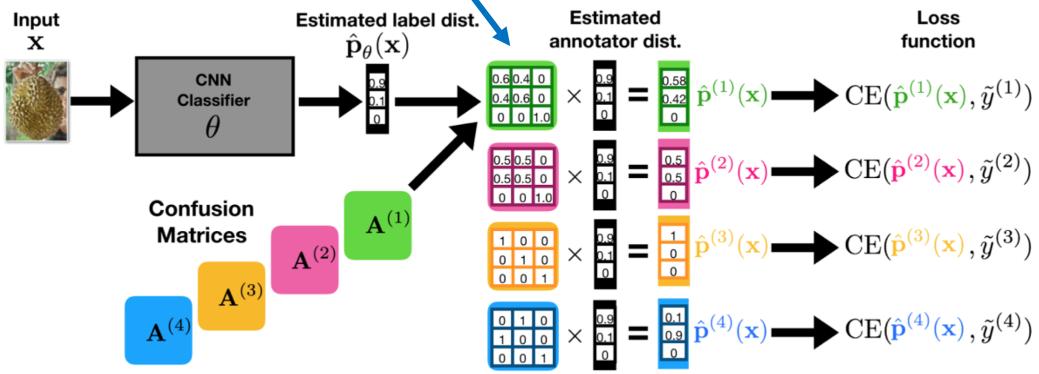
[1] Warfield S K, Zou K H, Wells W M. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation[J].

# Existing Work

- Annotator quality assessment by confusion matrix

Total number of labelers

$$\underbrace{p\left(\tilde{y}^{(1)}, \ldots, \tilde{y}^{(R)} \middle| \mathbf{x}\right)}_{\text{observed label distribution}} = \prod_{r=1}^{R} \int_{y \in Y} \underbrace{p\left(\tilde{y}^{(r)} \middle| y\right)}_{\text{noise model}} \cdot \underbrace{p(y|\mathbf{x})}_{\text{true label distribution (goal)}} \, dy$$

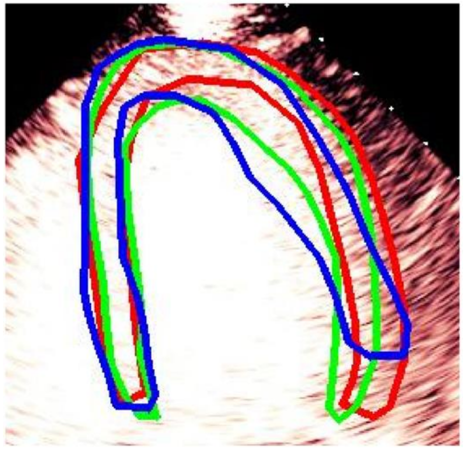


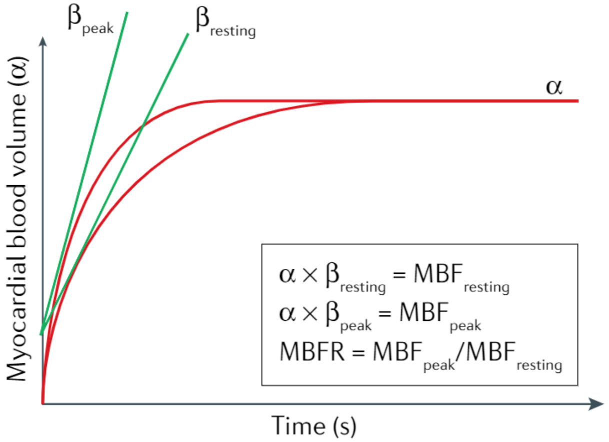

Example: model annotator quality using confusion matrix [1]

[1] Tanno R, Saeedi A, Sankaranarayanan S, et al. Learning from noisy labels by regularized estimation of annotator confusion (CVPR, 2019).

# Motivation

- Label noise is dependent on the original input
  - Images with large artifact will have larger label noise
- Segmentation annotations by different radiologists are all acceptable in clinical setting[1]
  - They can be used for further medical analysis



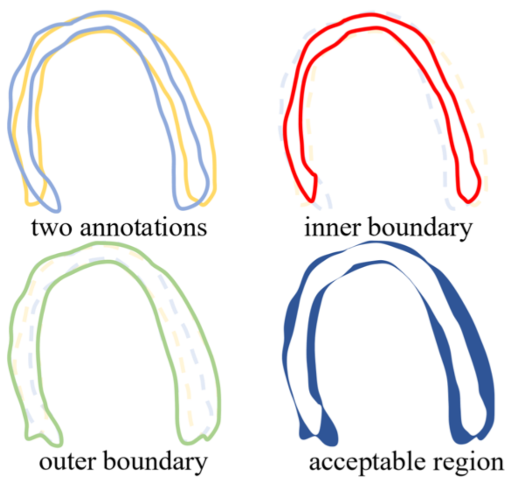


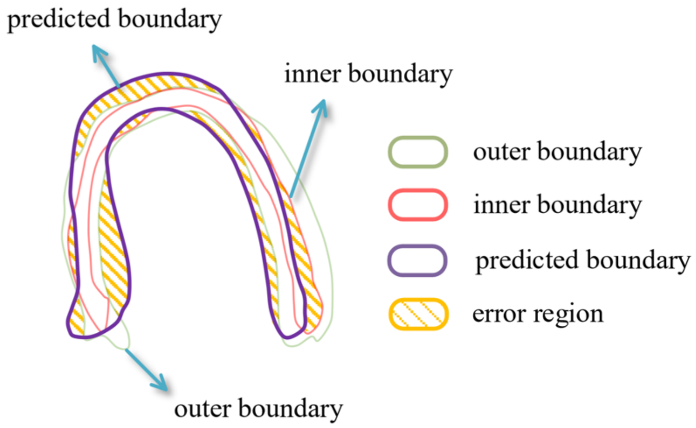

Perfusion analysis[2]


[1] McErlean A, Panicek D M, Zabor E C, et al. Intra-and interobserver variability in CT measurements in oncology[J]. Radiology, 2013, 269(2): 451-459.
[2] Dewey M, Siebes M, Kachelrieß M, et al. Clinical quantitative cardiac imaging for the assessment of myocardial ischaemia[J]. Nature Reviews Cardiology

# Method

- Extended Dice
  - Acceptable region where any radiologist agrees
  - Error region where none of the radiologist agree
  - Can be used for evaluation and training



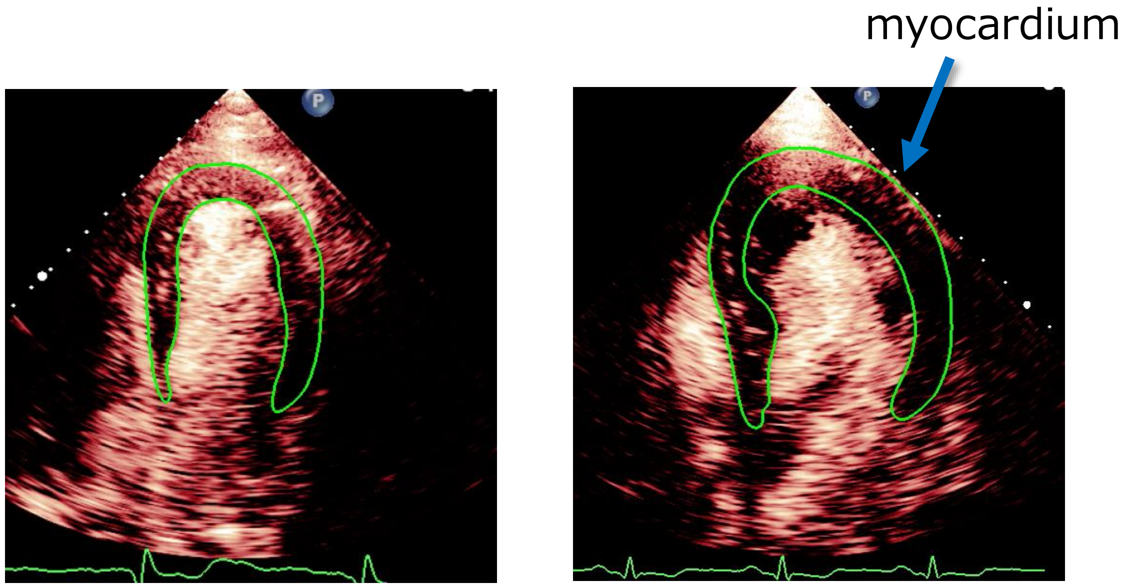

- $P$: predicted boundary
- $O$: outer boundary
- $I$ : inner boundary

$$Extended\ Dice = 1 - \frac{(P - P \cap O) + (I - P \cap I)}{P + I}$$

When $I = O$, extended Dice becomes Dice

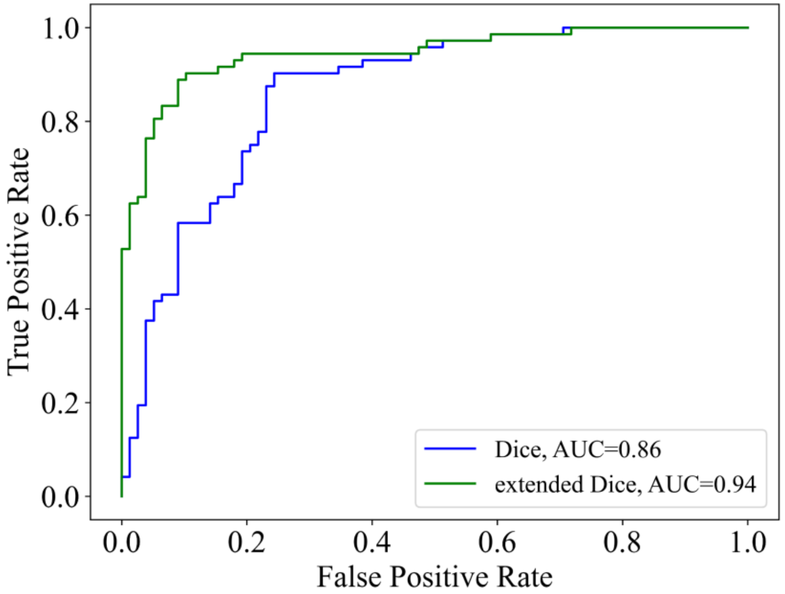
# Dataset

- Contrast Echocardiography Dataset
  - 100 patients
  - Each patient has 10 frames randomly selected from a sequence
  - 5 radiologists annotate each image
  - 700:300 for training and testing



# Extended Dice for Evaluation

- Compare Dice and extended Dice as segmentation evaluation metric
  - Using Dice and extended Dice as indicator to decide whether the prediction need manual correction
  - Class 1, need manual correction, class 0, do not need manual correction

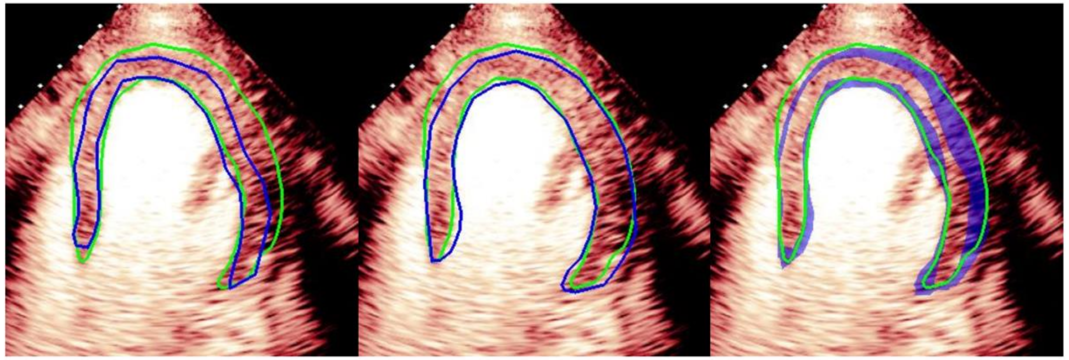



ROC curve

- Dice>0.8 Good
- ED>0.96 Good

Majority vote | Radiologist 3

Dice: 0.720 | Dice: 0.884 | ED: 0.961



- Green: predicted boundary
- Blue: ground truth
- Shaded Blue: acceptable regions

# Extended Dice for Training

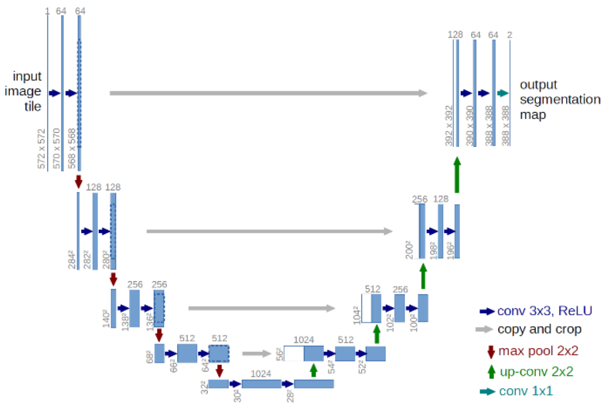
- Evaluation using conventional metrics (Dice, IoU, Hausdorff Distance)
  - Network: U-Net

| Method | GT: majority vote Dice | IoU | HD |
|---|---|---|---|
| Single Cardiologist | 0.838(.11) | 0.735(.12) | 28.4(17) |
| Inner Boundary | 0.770(.11) | 0.638(.11) | 34.2(17) |
| Outer Boundary | 0.785(.09) | 0.656(.11) | 34.0(12) |
| Consensus | 0.847(.12) | 0.753(.14) | 28.0(19) |
| Average Cross Entropy | 0.844(.11) | 0.745(.13) | 26.4(15) |
| Confusion Matrix [1] | 0.826(.12) | 0.719(.13) | 37.9(22) |
| Consistency [2] | 0.847(.10) | 0.749(.13) | 29.8(16) |
| STAPLE [3] | 0.814(.09) | 0.695(.11) | 31.8(16) |
| Ours | **0.855(.10)** | **0.759(.12)** | **25.4(14)** |

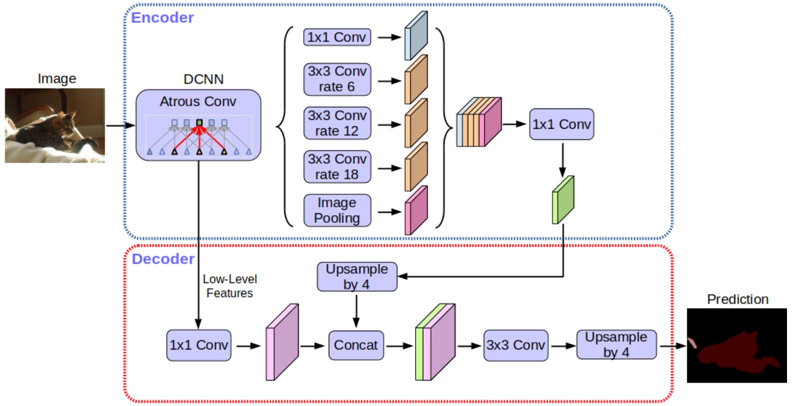
[1] Tanno R, Saeedi A, Sankaranarayanan S, et al. Learning from noisy labels by regularized estimation of annotator confusion[C] (CVPR, 2019)
[2] Sudre C H, Anson B G, Ingala S, et al. Let's agree to disagree: Learning highly debatable multirater labelling[C] (MICCAI, 2020)
[3] Warfield S K, Zou K H, Wells W M. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation[J].

# Extended Dice for Training

- Evaluation using extended Dice
  - Network: U-Net and DeepLab V3+




U-Net

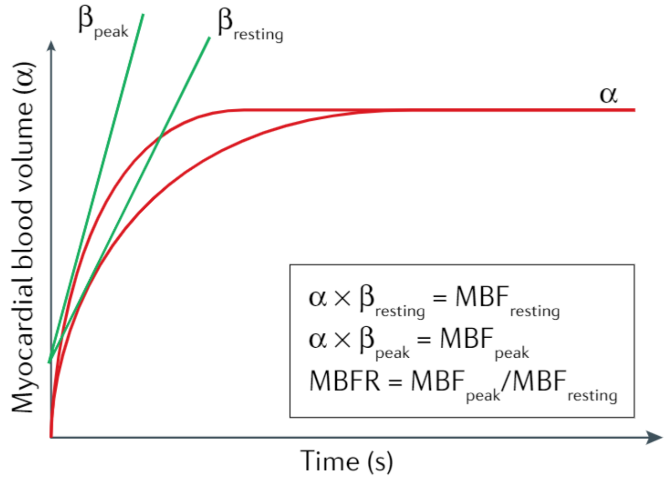


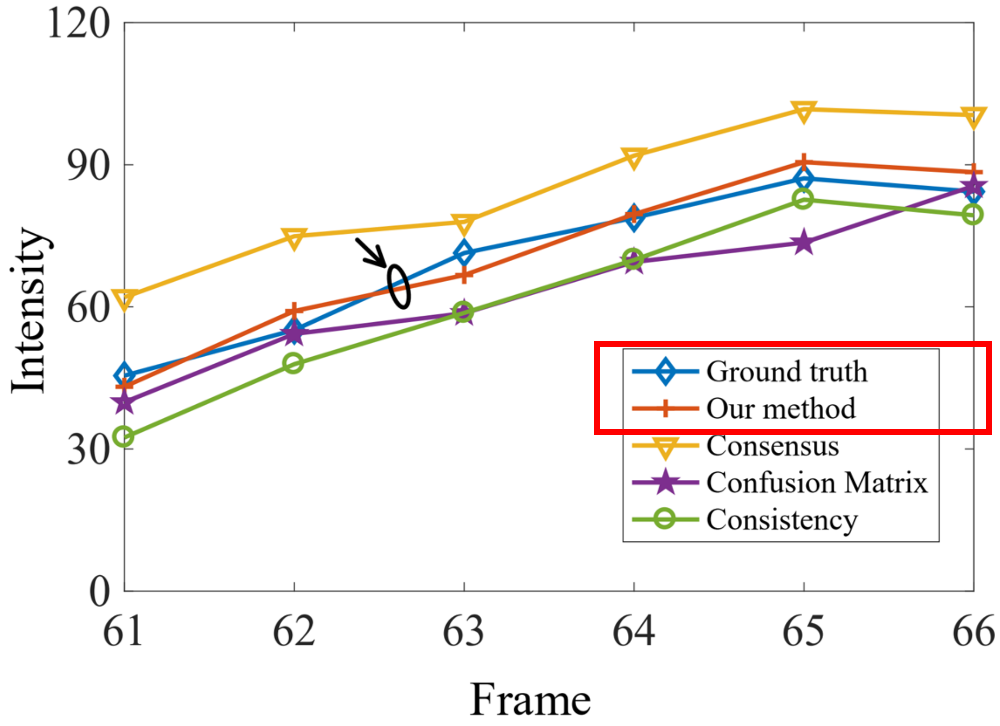

DeepLab V3+

| Method | SC | IB | OB | Consensus | ACE | CM | Consistency | STAPLE | Ours |
|---|---|---|---|---|---|---|---|---|---|
| U-Net | 0.929(.06) | 0.947(.05) | 0.848(.07) | 0.940(.06) | 0.919(.06) | 0.951(.06) | 0.947(.06) | 0.912(.06) | **0.958(.05)** |
| DeepLab | 0.942(.07) | 0.906(.08) | 0.891(.07) | 0.946(.08) | 0.945(.07) | 0.924(.08) | 0.944(.07) | 0.921(.07) | **0.954(.06)** |

# Extended Dice for Training

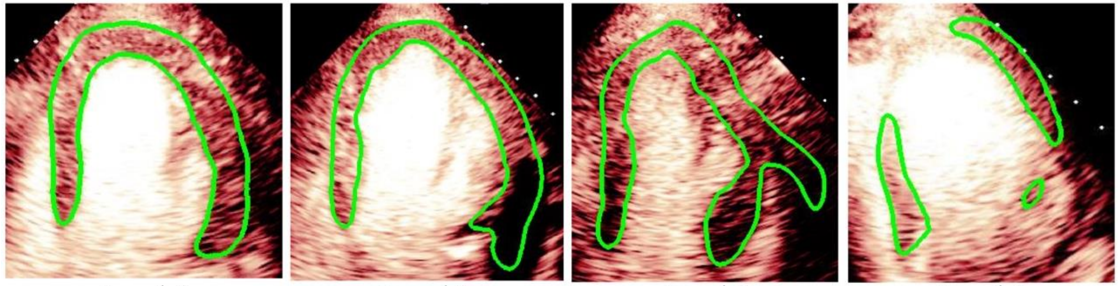
- Evaluation using frame-intensity curve
  - Frame-intensity curve is used for myocardial perfusion analysis to evaluate the functionality of heart.



# Extended Dice for Training

- Grading study
  - An independent and experienced radiologist is asked to grade the myocardial segmentation result in a blind setting
  - 4 grading levels



Level 4 Level 3 Level 2 Level 1

| Grading Level | Consensus | Confusion Matrix | Consistency | Our method |
|---|---|---|---|---|
| Level 4 (Highest) | 58 | 71 | 69 | **72** |
| Level 3 | 56 | 39 | 42 | 50 |
| Level 2 | 11 | 17 | 22 | 20 |
| Level 1 (Lowest) | 25 | 23 | 17 | **8** |

# Wrap Up

- New extended Dice to train neural network and evaluate segmentation performance when multiple acceptable annotations are available
- A more robust evaluation metric
- Improve the model accuracy both quantitatively and qualitatively.