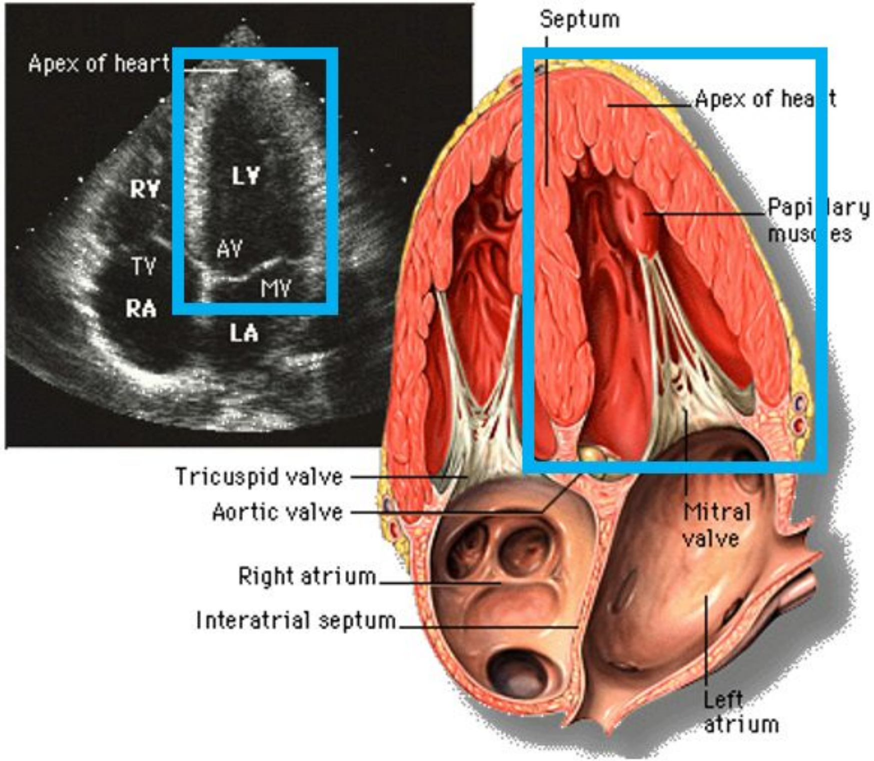
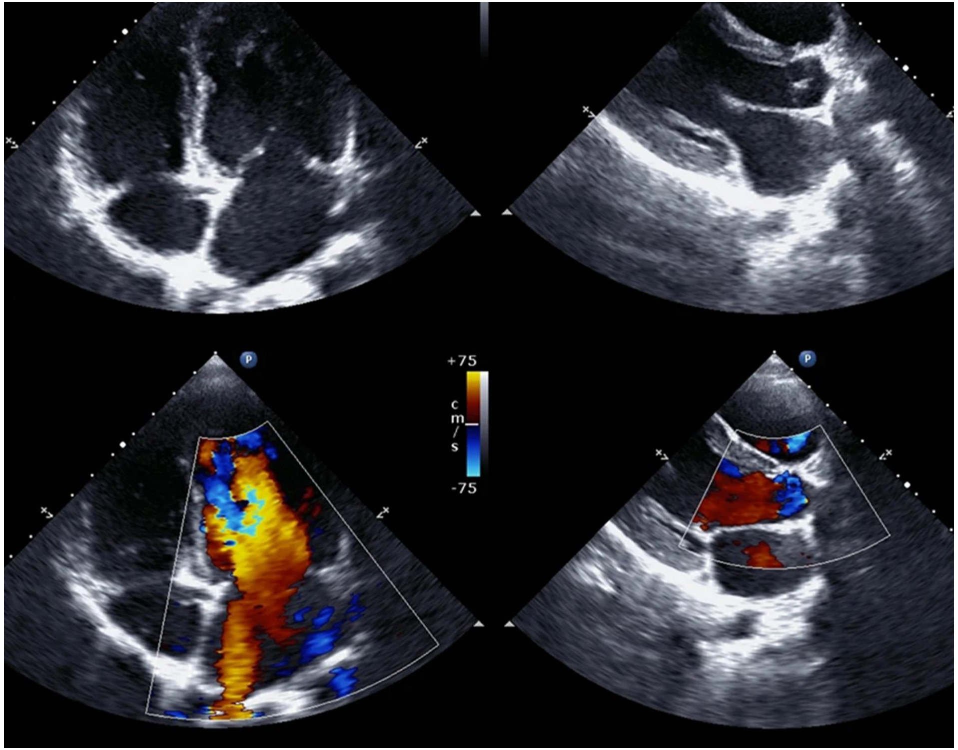
Auto-doctor for echocardiography

Chengche Tsai, Alessandro Folloni

Recap - background



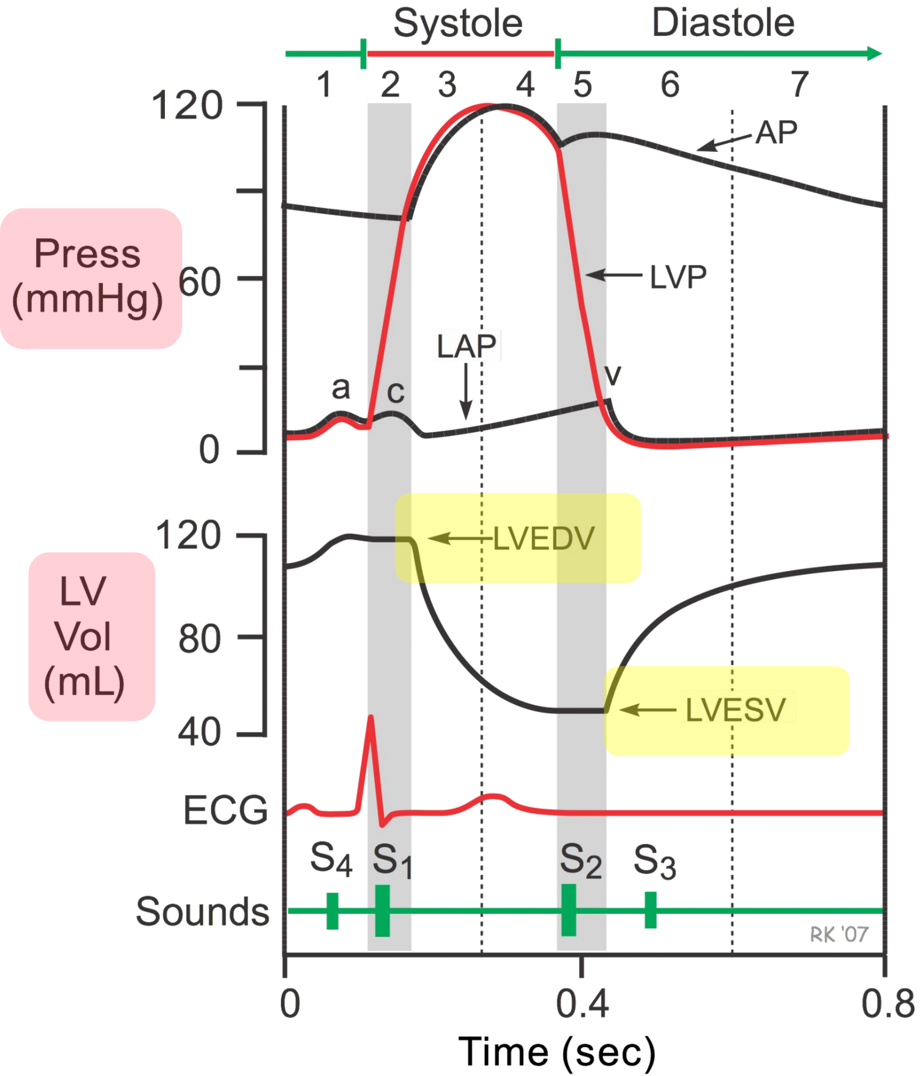


Recap - terminology

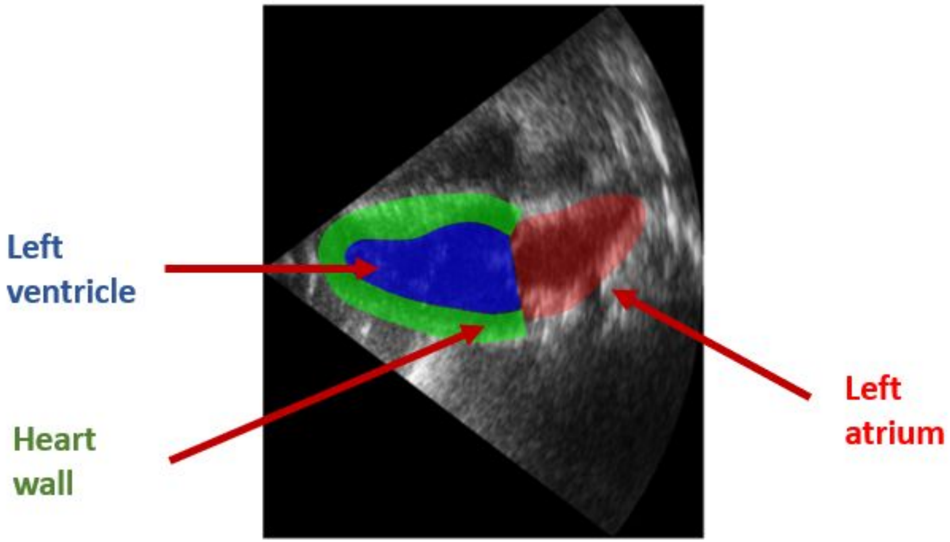
EF (ejection fraction): the percentage of blood pumped out from the heart per stroke.

ED: end of the diastolic phase (the largest volume)

ES: end of the systolic phase (the smallest volume)



The data (CAMUS)

- 500 videos
- Semantic segmentation: full annotations of left ventricle, myocardium (heart muscle) of left ventricle, and left atrium
- 

Goals of the final report

What have been done in midway:

- Moved from Image-based to video-based
- Great segmentation results (qualitatively/quantitatively) .

What we wanted to achieved in the final:

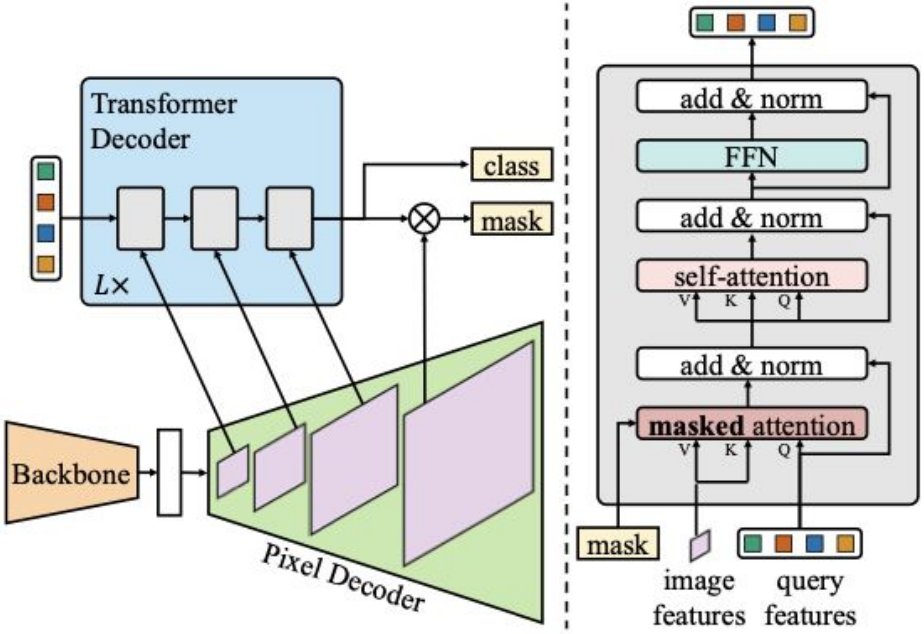
- A transformer-based segmentation model
- Better EF estimation

Mask2Former, 2022

Mask2former

Key elements:

- Backbone
- Pixel decoder
- Transformer decoder with masked attention



Main strengths of the model

- Masked attention: extracts localized features

$$\mathbf{X}_l = \mathrm{softmax}(\boldsymbol{\mathcal{M}}_{l-1} + \mathbf{Q}_l\mathbf{K}_l^{\mathrm{T}})\mathbf{V}_l + \mathbf{X}_{l-1}.$$

$$\boldsymbol{\mathcal{M}}_{l-1}(x, y) = \begin{cases} 0 & \text{if } \mathbf{M}_{l-1}(x, y) = 1 \\ -\infty & \text{otherwise} \end{cases}$$

Main strengths of the model

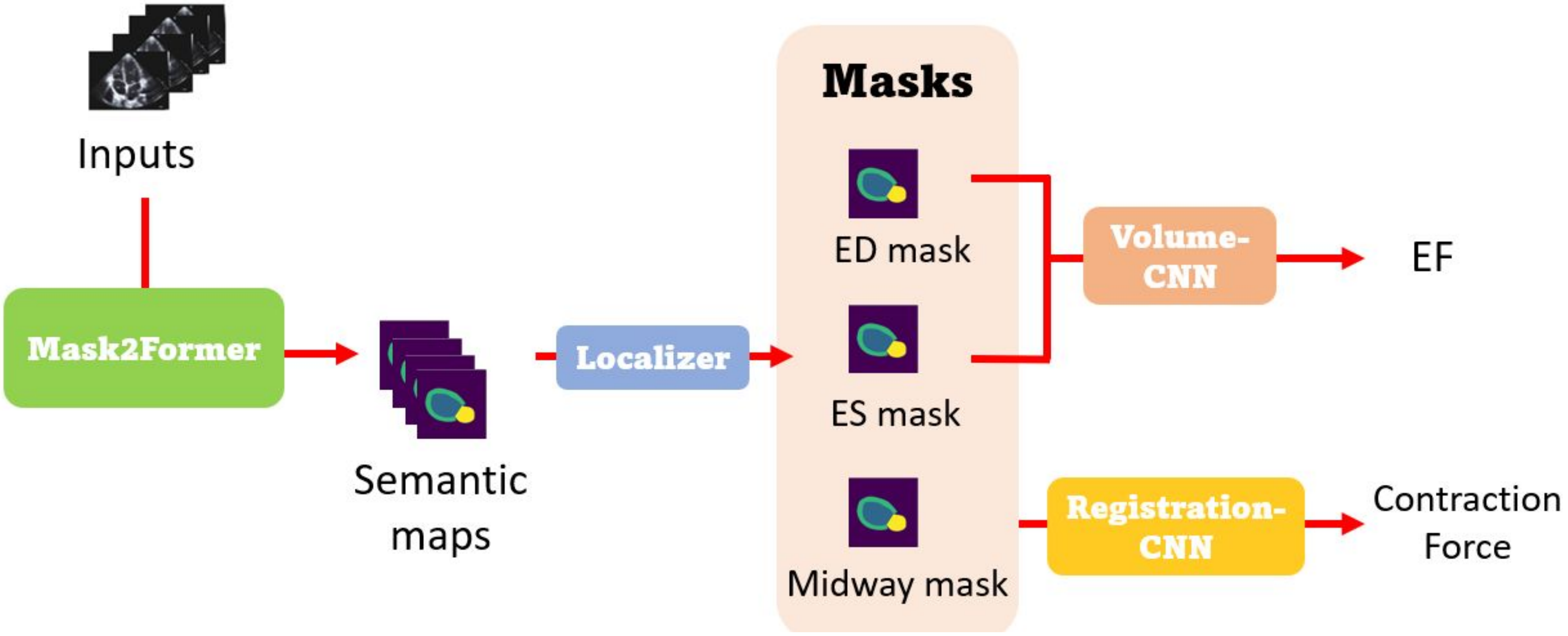
- High resolution features: increase performance (and controlling computational cost)

  → we utilize a feature pyramid which consists of both low and high-resolution features and feed one resolution of the multi-scale feature to one Transformer decoder layer at a time.

- Optimization: changed order of self and cross-attention to have more effective computation.

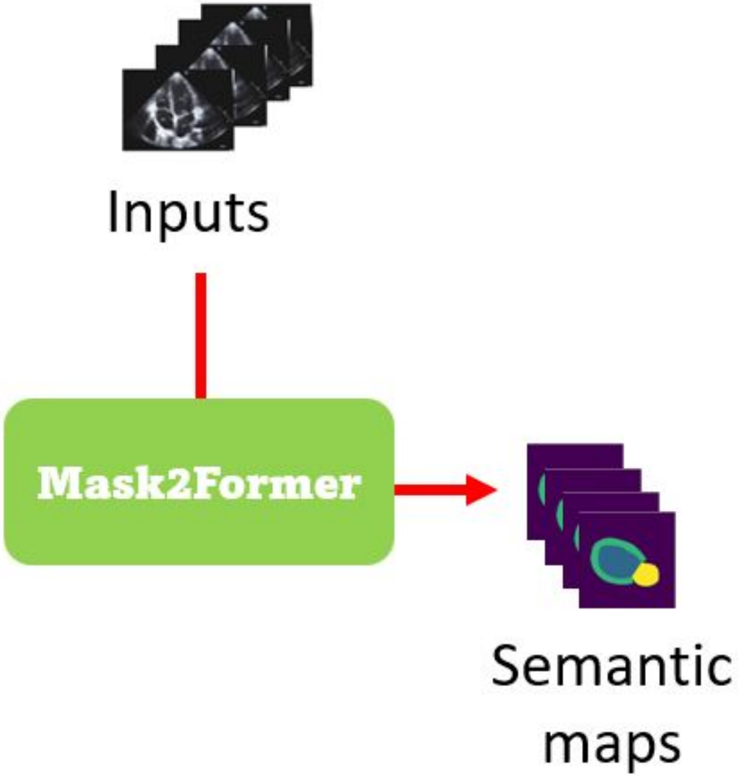
Additionally

Overview


Inputs
Mask2Former
Semantic maps
Localizer
Masks
ED mask
ES mask
Midway mask
Volume-CNN
EF
Registration-CNN
Contraction Force

Segmentation

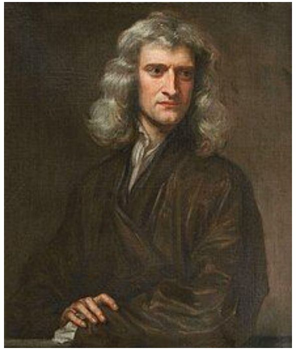
- Mask2Former outputs instance segmentation masks
- Instance reduces to semantic segentations if only 1 object per class.



Segmentation - temporal smoothness

To use temporal information and anatomically constraint the output:

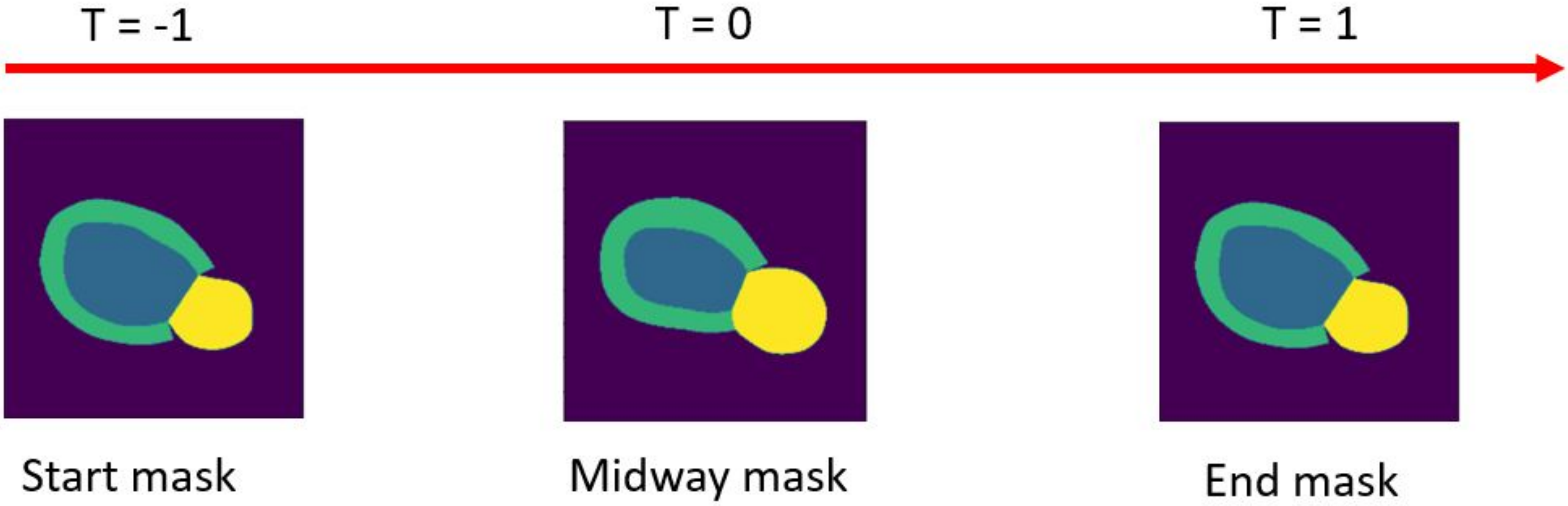
```
def temporal_smoothness_loss(frames):
    # Function to compute the temporal smoothness loss
    frame_diff = frames[1:] - frames[:-1]

    # You can use different norms or similarity metrics to measure difference
    loss = torch.mean(torch.pow(frame_diff, 2)) # L2 (MSE) loss
    return loss
```



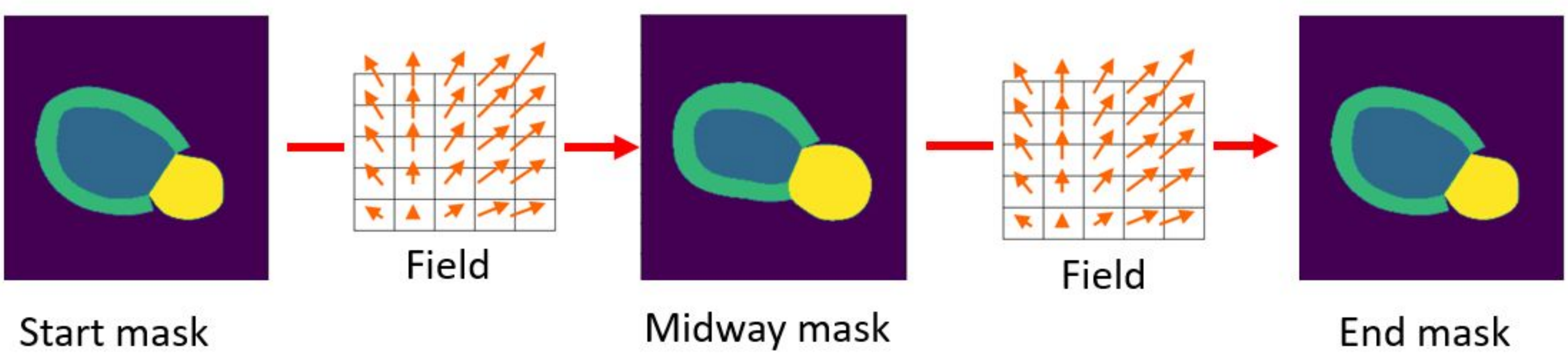
Pixel, is not the only thing we can model through time

Segmentation - future prediction

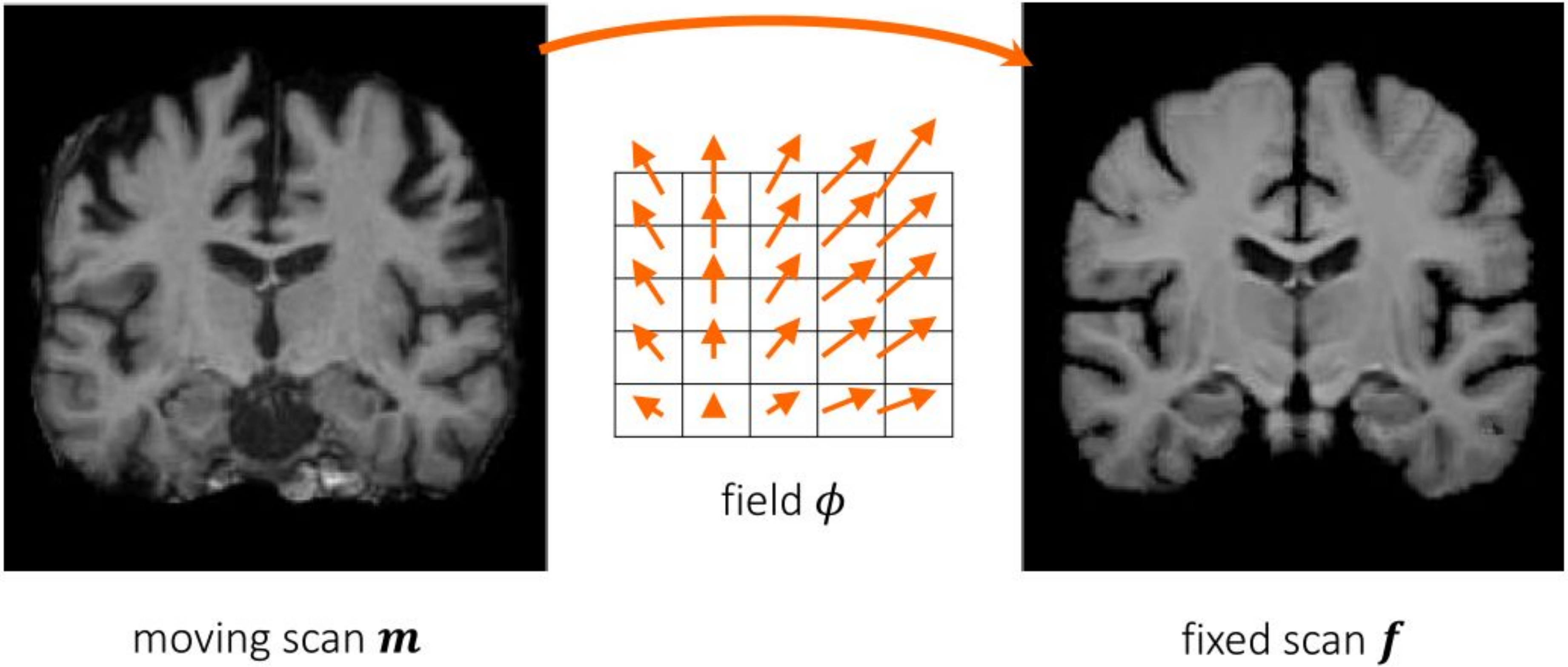
- To force the model better learn the temporal relations, we force the model to predict a future frame



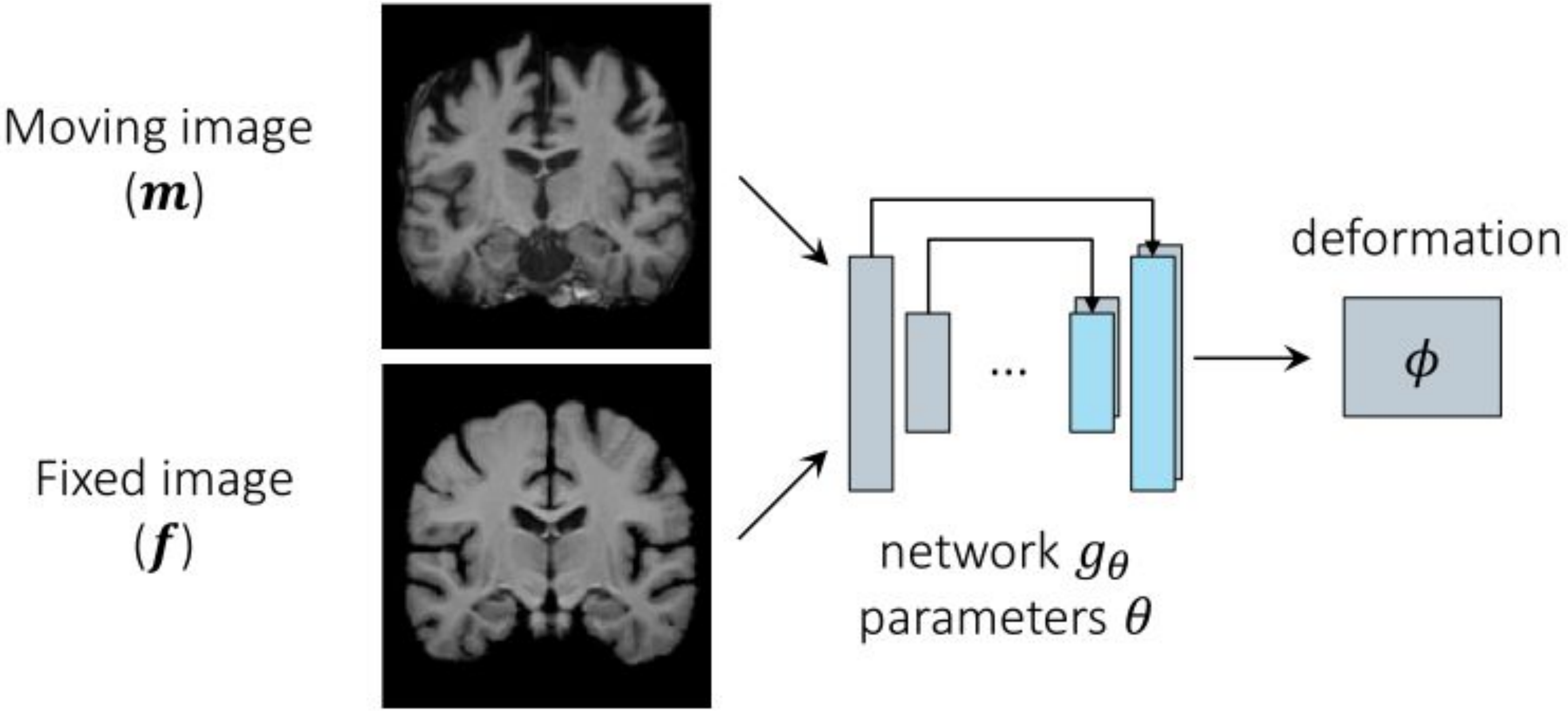
Segmentation to registration



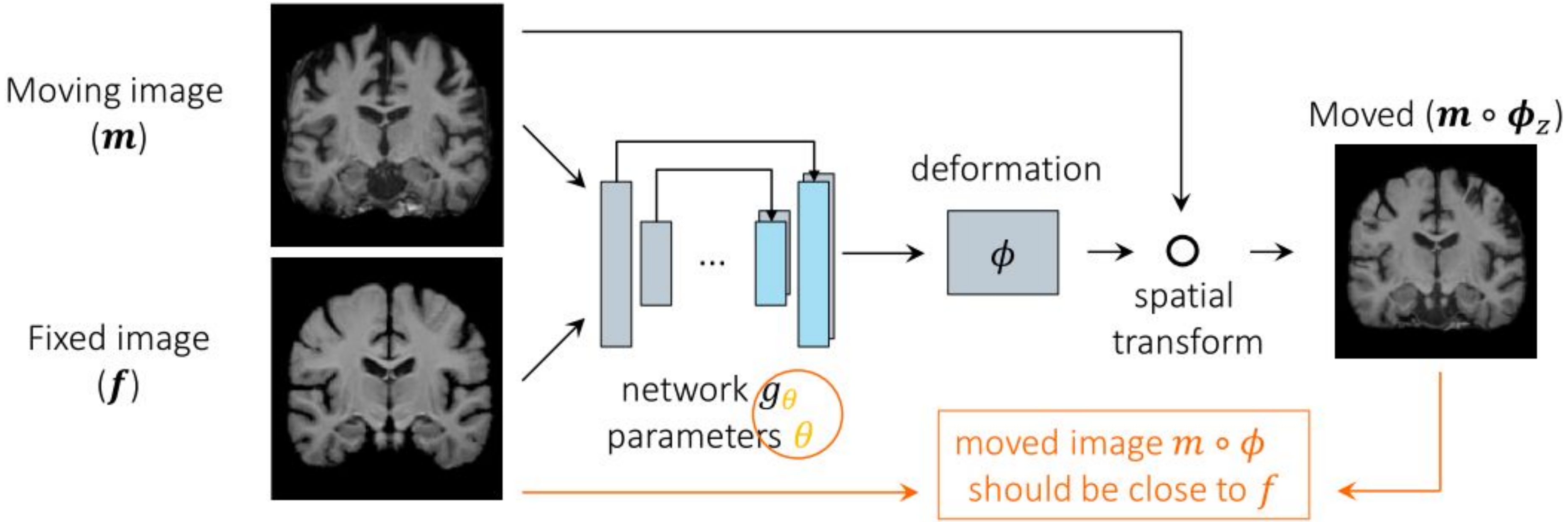
Brief review of registration



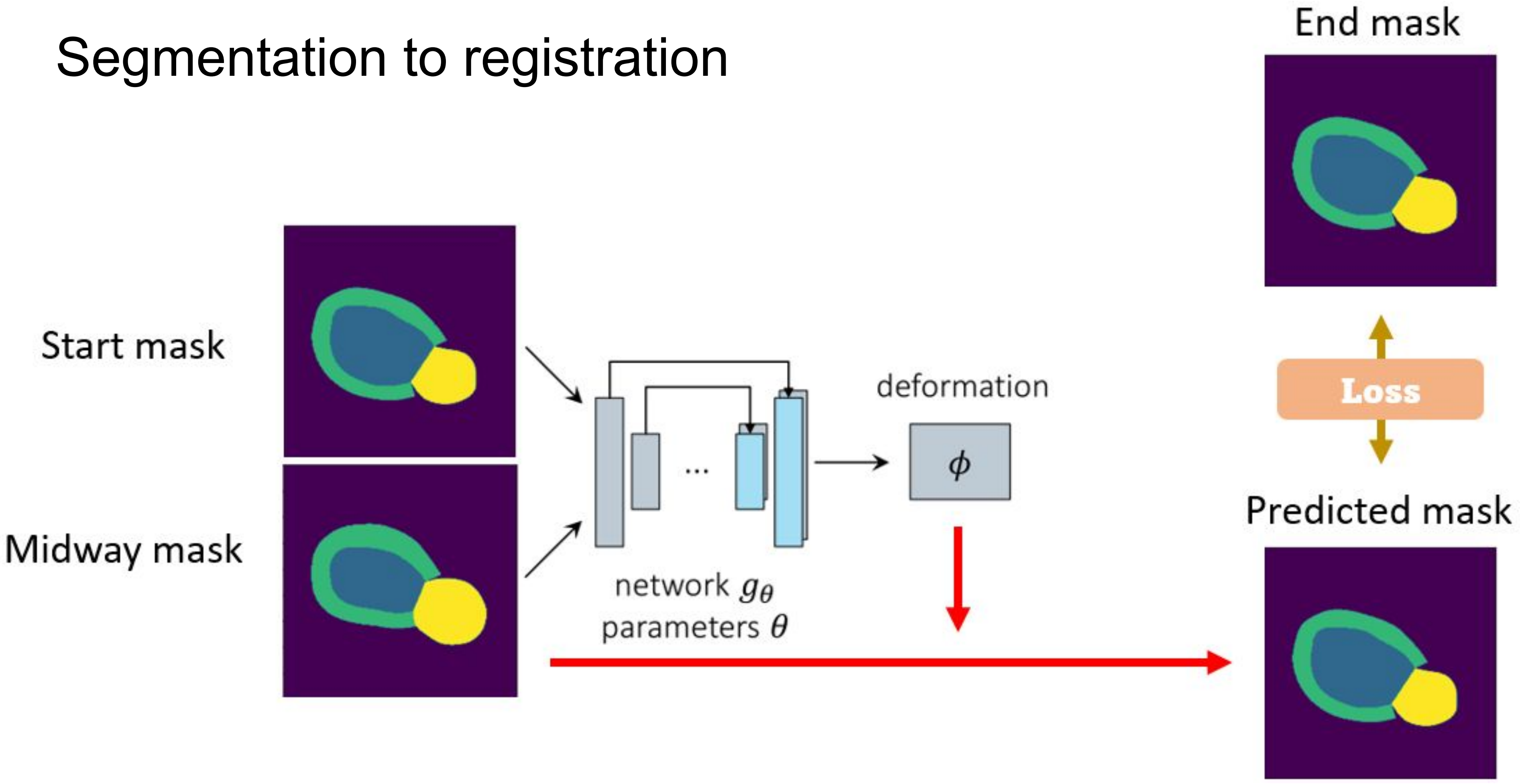
Brief review of registration



Brief review of registration

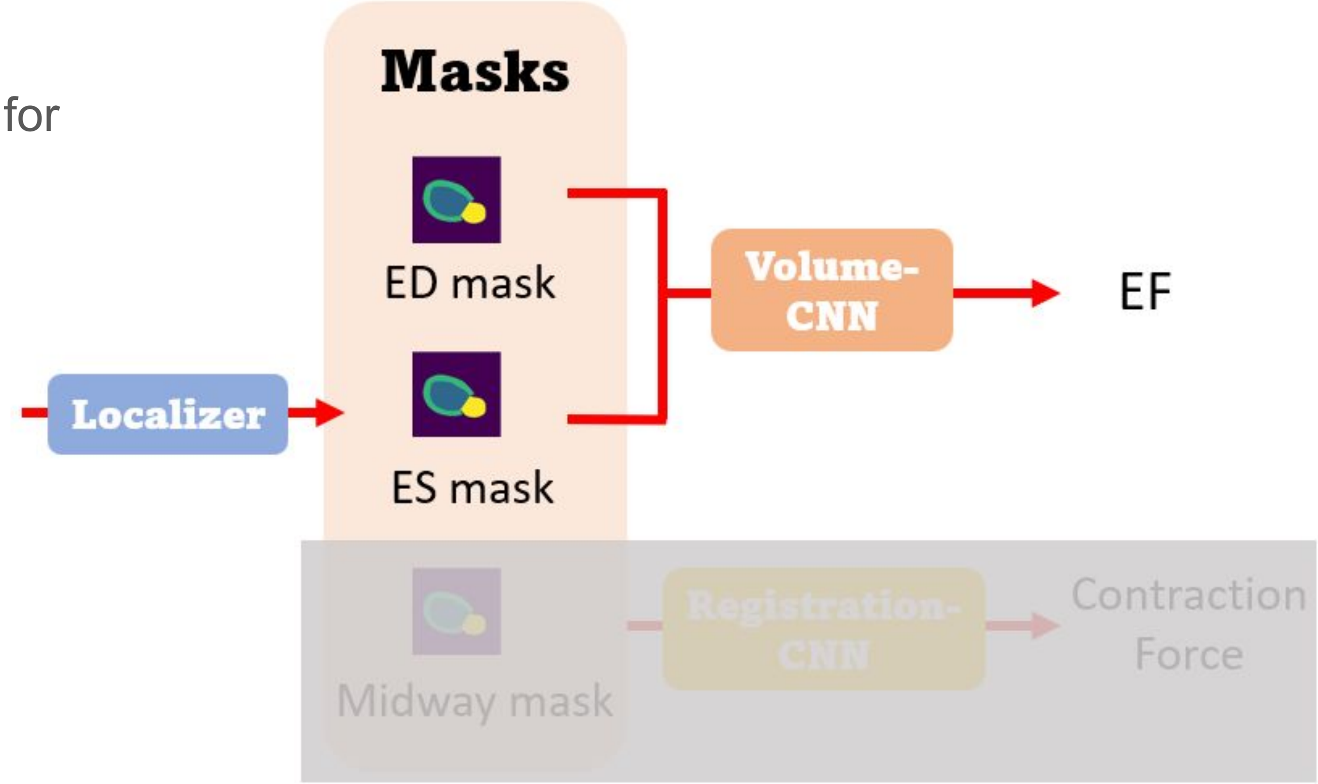


Segmentation to registration



Volume-CNN

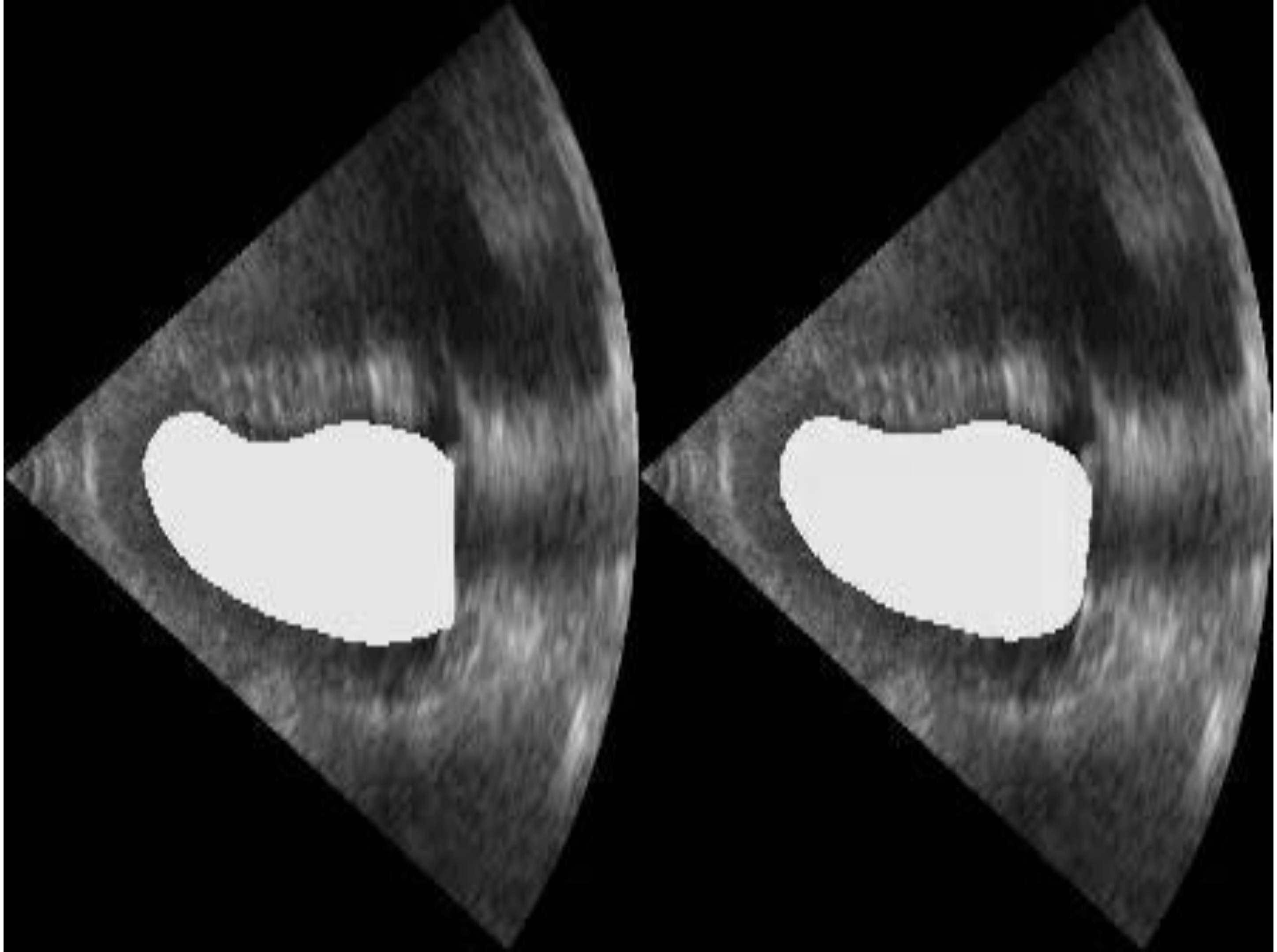
- Predicts volumes for ED and ES
- EF = (ED-ES)/ED



Result (segmentation)

	DeepLabv3	Mask2Former (small)	Mask2Former (base)	Ours (no registration)	Ours
mIOU (Left ventricle)	0.808	0.5	0.58	0.65	0.75

*Both models were trained on 400 videos for only 10 epochs

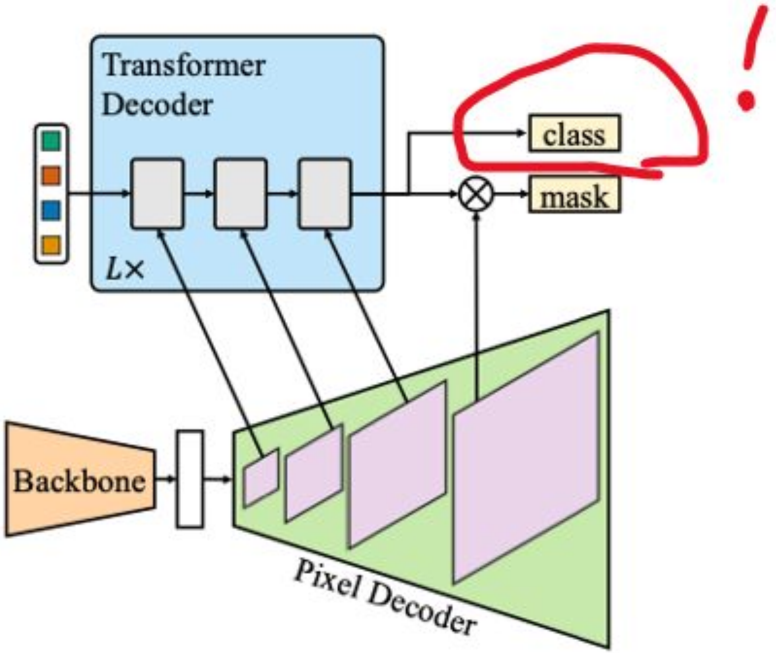


Result - EF estimation

- Even worse…
  - SOTA: RMSE = 5.17
  - Ours (previous): 17.xx
  - Today: 24.21

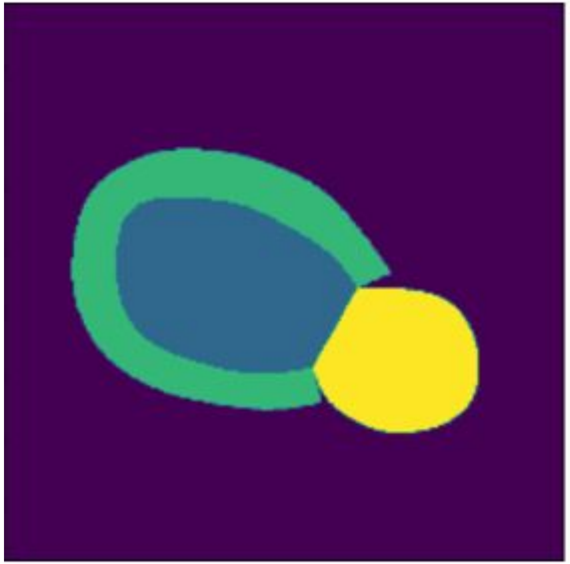
Failure analysis

- Mask2Former is encapsulated too well.
- Not enough data to train Mask2Former
- Mask2Former is a instance segmentation models. The classification layer is not needed in our case, probably harmful.

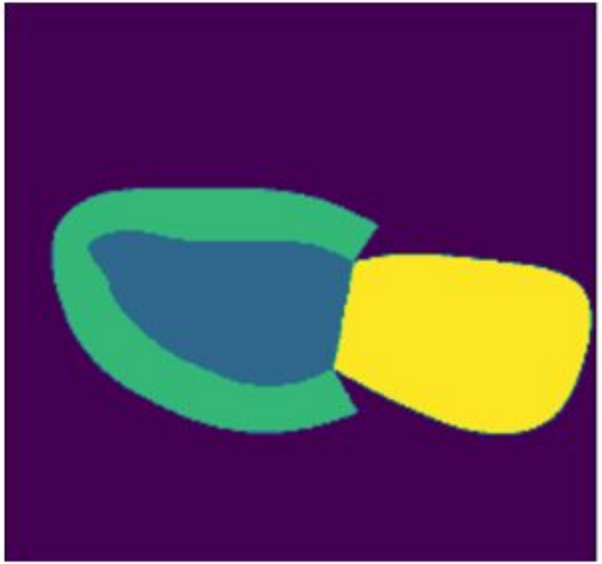


Both “start” phase

Patient A



Patient B



More to do

1. Probably try a simpler transformer-based model first.

2. To add ‘cycle-consistency' to the model
    a. -> Find a dataset with long clips (>2 cycles) and good annotations.

3. Train looooonger, >1k iterations.