2024 Spring DLMI HW

Student ID: r12945039 Name: 楊瀚博

Ultrasound Needle Segmentation with U-Net Transformer & Masked Autoencoder Self-Supervised Pre-Training

1. Introduction

Ultrasound guided needle injection or aspiration plays an important role in clinical healthcare. However, consistently aligning a thin 2D-ultrasound scanning plane with the injected needle and trying to identify the accurate needle location on the ultrasound screen could be a very challenging work. Therefore, one would be benefitted a lot from a localization system that emphasizes the needle tip on the ultrasound screen in real-time.

In this work, we follow the U-Net Transformer (UNETR) [1] architecture design, which adopts a Vision Transformer (ViT) [2] as the encoder to capture the input features and then using the U-Net [3] skip connection with convolution layers to reconstruct the fine-grained needle segmentation mask. (Shown in Figure 1.) Moreover, we also examine the effectiveness of a self-supervised pre-trained Vision Transformer encoder by applying the Masked Autoencoder (MAE) [4] pre-training technique on lots of unlabeled ultrasound image data.

We observe several things based on the experiment results:

- The UNETR has the ability to capture the needle location in a single ultrasound image to a certain level.
- The UNETR performs better with the consecutive ultrasound images input which includes spatiotemporal information.
- The ImageNet-1K pre-trained weights for the ViT encoder in UNETR are indispensable under a small dataset scenario.
- Our preliminary experiment results indicates that a ViT encoder with MAE self-supervised pre-training on relatively small ultrasound image data does not improve the needle segmentation performance.

2. Method

2.1 Architecture Selection

In terms of medical image segmentation tasks, U-Net is one of the most classic architecture that utilized skip connections which fuses the low-level (high-resolution) and high-level (low-resolution) feature maps together to obtain a high-quality segmentation mask. Since the revolution of the ViT architecture, many works aimed to combine them with the U-Net architecture to improve the medical image segmentation performance. Additionally, lots of recently proposed self-supervised learning methods are heavily based on the ViT architecture, e.g., MAE. As a result, we adopt the UNETR architecture, which was originally proposed for 3D medical image segmentation, as our ultrasound image needle segmentation model. And we also apply different types of input images and target masks to investigate the effectiveness of the temporal information inside the consecutive frames. Furthermore, we design several self-supervised experiments since we wonder if the MAE pre-training on ultrasound image data would be beneficial to segmentation task.

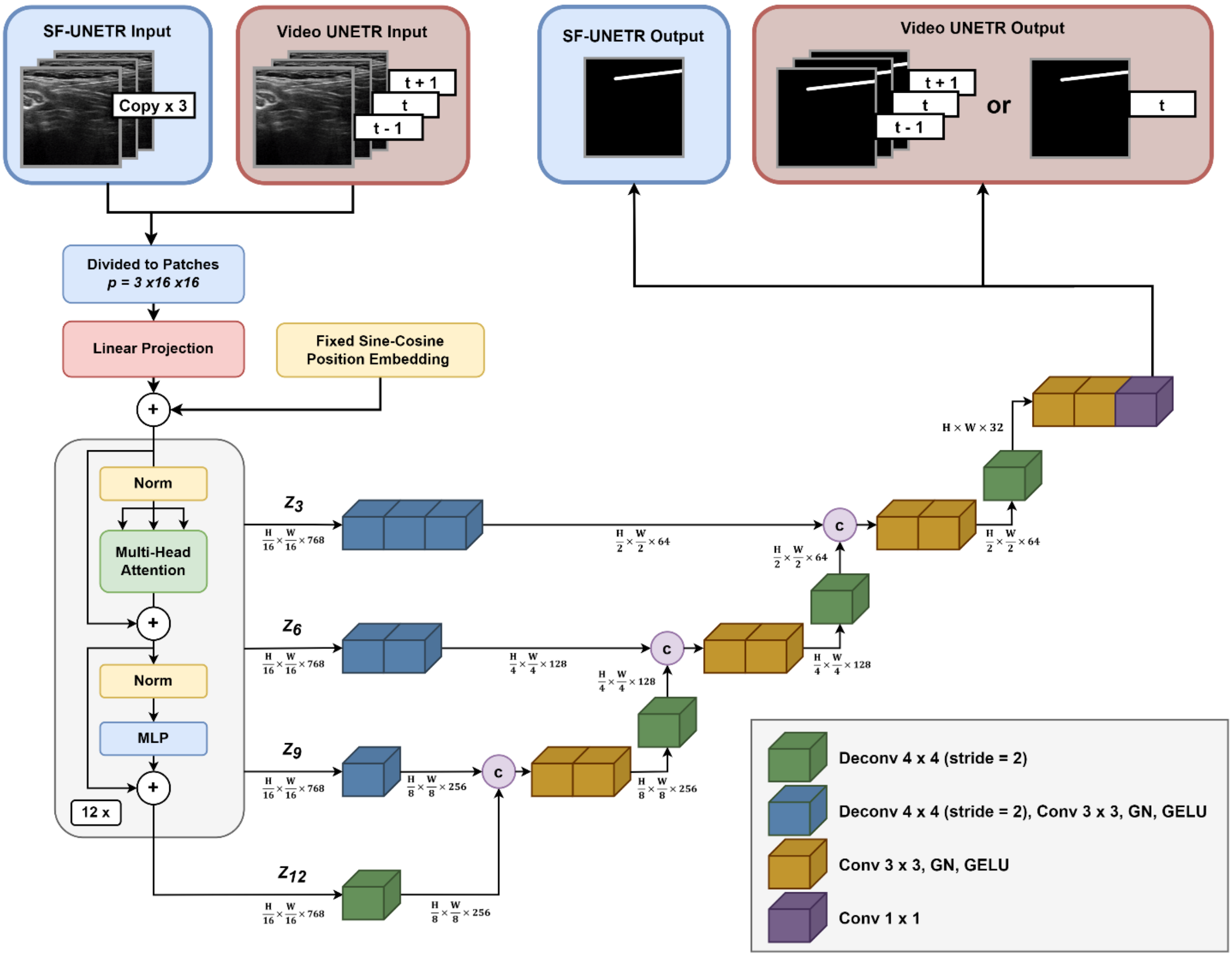

Figure 1. **SF-UNETR** and **Video UNTER** architecture for **ultrasound needle segmentation.** Note that the only differences between SF-UNETR and Video UNETR are the input modality (single frame broadcasted to 3 channels or 3 consecutive frames concatenated along channel dimension), and the output prediction target (single mask or all 3 consecutive masks).

2.2 Single Frame UNETR (SF-UNETR)

Since the original UNETR was used for 3D images, we modified both the ViT encoder and the convolution layers to 2D version that process 2D images. Additionally, we remove the original first skip connection that bridges the direct input image to the last up-sampling layer since we find it leads to an unstable training. The Single Frame UNETR (SF-UNETR) architecture we use in this work is shown in Figure 1.

In a SF-UNETR, we ignore the temporal information between the ultrasound consecutive frames. In other words, we treat this needle segmentation task as a pure 2D-image segmentation problem that process all the ultrasound frames independently. The input would be a single ultrasound frame broadcasted to three channels. The reason we broadcast the grayscale input to three channels is that the pre-trained weight of the ViT encoder on ImageNet-1K (IN1K) can only be applied under this setting. Accordingly, the model output would be a 1-channel binary mask that predicts the needle location in the input ultrasound frame.

2.2 Video UNETR

On top of the SF-UNETR architecture, we propose a simple network that process the consecutive ultrasound frames simultaneously. Actually, the Video-UNETR is identical to the SF-UNTER in terms of the model architecture design. The key difference between them is that the input of the Video-UNETR now becomes three consecutive ultrasound frames that being concatenated along the channel dimension.

As the input contains three frames at different time step, the model now has the potential to aggregate the spatiotemporal information to produce a more accurate needle segmentation mask. The aggregation is achieved by the linear projection layer after the input images being divided into several non-overlapping patches just before the ViT encoder.

Moreover, since the input is three consecutive frames, we can decide which target mask the model should predict, e.g., all three needle masks of the consecutive frames at a time, or just one of them. In this work, we conduct two experiments. The first one is to predict all three masks at a time (3-channel), the other is to predict only the second (middle) frame mask (1-channel). Therefore, the out-channel number of the final convolution layer in Video UNETR depends on the target mask type.

2.3 Masked Autoencoder for Ultrasound Images

We further investigate the effectiveness of the well-known self-supervised pre-training method, MAE, on the segmentation task when applied on a relatively small ultrasound image dataset. During MAE pre-training, the input images would be randomly masked and ViT encoder can only process those image region are not masked. Due to the limitation on experiment time and computation resource, we only apply this pre-training method on the 2D ultrasound images, ignoring the temporal information between consecutive frames.

We follow the MAE official implementation using a ViT base model. The input image is a single ultrasound frame that being broadcasted to three channels, while the reconstruction target is identical to the input image. In this work, we pre-train the MAE using two different weight initialization method. The first one is from scratch, which means we directly pre-train a random initialized MAE on our ultrasound dataset. The second one is we initialize the MAE using the weights that have been pre-trained on the ImageNet-1K dataset, and then “further pre-train” it on our ultrasound dataset.

After pre-training, we use the ViT encoder weights inside MAE to initialize the encoder of SF-

UNETR.

2.4 Loss Functions

In the needle segmentation tasks, we apply the sum of focal loss and dice loss as our training objective. Dice loss is one of the most common loss functions that being used in the medical image segmentation tasks, while focal loss is designed to address the class imbalance issue by applying a modulating term to the cross-entropy loss in order to focus learning on hard misclassified examples.

In terms of MAE pre-training, we follow the original implementation that measures the mean squared error between the reconstructed image and the original input image pixels.

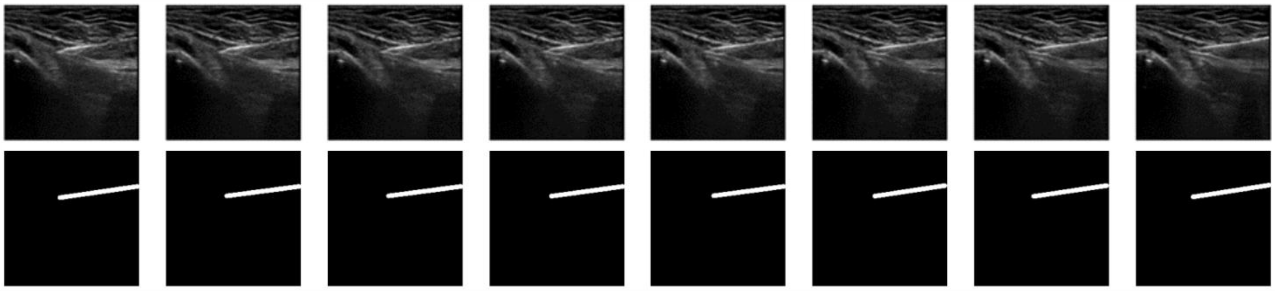

Figure 2. Eight consecutive ultrasound frames (upper) with the ground truth needle shaft mask (bottom). The mask region is the connected line between two points that annotated by the experts.

3. Datasets

	Number of Simple Procedures	Number of Medium Procedures	Number of Hard Procedures	Total Ultrasound Frames
Training	**7**	**1**	**0**	**4,429**
Validation	**4**	**0**	**0**	**2,553**
Pre-Training	**N/A**	**N/A**	**N/A**	**11,475**

Table 1. Dataset composition. The ***simple***, ***medium*** or ***hard*** level are categorized by the expert based on the visibility of the needle in the ultrasound videos. Note that the pre-training dataset has not been reviewed by the experts, thus the visibilities are unavailable.

3.1 Overview

The ultrasound data we used in this work were collected from National Taiwan University Hospital under IRB approval and informed consent. Twenty clinical ultrasound guided procedures at near 15 frame per second (captured at 30 frame per second originally, with removal of identical consecutive frames due to unavoidable frame drop when scanning at a depth of 4 cm) were categorized as ***simple***, ***medium*** or ***hard*** by the expert based on the visibility of the needle in the ultrasound images.

The original annotation by experts is the coordinates as (x1, y1) and (x2, y2) representing the two end points of the visible needle shaft. To implement a semantic segmentation task, we connected the two points and added dilation to the line, to form a thick line semantic mask. Some frames are not annotated due to the needle is out of the scanning plane thickness. However, they are still included in our training set to maintain a continuous and reasonable temporal flow as in the real world. Several ultrasound frames and corresponding annotated needle masks are shown in Figure 2.

All the images in this work are resized to resolution 224×224. The composition of datasets is shown in Table 1.

3.2 Training & Validation Set

The split datasets for training the segmentation models are provided in Table 1. We choose seven ***simple*** and one ***medium*** recorded procedure as the training set, which totally contains 4,429 images. On the other hand, the validation set is selected from another four ***simple*** recorded procedures, which contains 2,553 images. The datasets for training SF-UNETR and Video UNETR are composed of exactly the same procedures while the datasets for Video UNETR treat every three consecutive frames as a sample. Since this is a preliminary study on the model design and the feasibility of MAE pre-training method, we do not consider those ***hard*** procedures at this point.

3.3 MAE Pre-training Set

The dataset for MAE pre-training is composed by the mentioned training dataset with 4,429 images, and other procedures which include 7,046 unlabeled ultrasound images. The total pre-training set contains 11,475 images.

4. Experiments

4.1 Implementation Details (Table 2. & 3.)

Single Frame UNETR. We select the ViT-base backbone as our SF-UNETR encoder, which has 12 transformer blocks, embedding dimension $D = 768$ with patch size $p = 16$. Sine-cosine positional embedding is used and we do not remove the class token. The number of channels in the skip connection are 64, 128 and 256 from the outputs of 3rd, 6th and 9th encoder blocks, respectively. We replace the batch normalization and ReLU activation with group normalization and GELU in the convolution blocks. The input and output shapes are 3×224×224 and 1×224×224, respectively. We train the model for 20 epochs at most, with early stopping condition that the validation loss stopped decreasing for 5 epochs. Effective batch size is set to 64 with device batch size 8 and gradient accumulated steps = 8. We set a base learning rate 1e-4 with linear decay with AdamW optimizer following the official MAE fine-tuning suggestion.

Video UNETR. We apply the same architecture design and training setting with SF-UNETR except for the output shape. If the prediction target is all three masks of the consecutive frames, the output shape would be 3×224×224. And the output shape becomes 1×224×224 if the prediction target is the middle frame mask only.

MAE. We follow the official implementation of MAE paper using a ViT-base encoder and a ViT decoder with 8 transformer blocks, embedding dimension $D = 512$. Masking ratio is set to 0.75, which to be found the best in the original paper. The patch normalized target loss is used if we're pre-training the MAE model initialized with the ImageNet-1K pre-trained weights since the pre-trained weights are learned through this setting. Otherwise, if pre-training MAE from scratch, we use the exact target pixel loss. Moreover, we also compare the downstream fine-tuned results between using a layer-wise learning rate decay or not during pre-training.

	Input Shape	Output Shape	Encoder Depth	Encoder Width	Batch Size	Learning Rate	Optimizer	Loss
SF-UNETR	3×224×224	1×224×224	12	768	64	1e-4	AdamW	Dice + Focal
Video UNETR (All 3 Targets)		3×224×224						
Video UNETR (Middle Target)		1×224×224						

Table 2. Implementation details of SF-UNETR and Video UNETR training.

	Epochs	Layer-wise lr Decay	MSE Loss Type	En/Decoder Depth	En/Decoder Width	Batch Size	Learning Rate	Optimizer
MAE v1 (From Scratch)	100	No	Exact Pixel Loss	12 / 8	768 / 512	64	1.5e-4	AdamW
MAE v2 (From IN1K)	50	Yes	Normalized Patch Pixel Loss					
MAE v3 (From IN1K)	50	No						
MAE v3 (From IN1K)	100	Yes						

Table 3. Implementation details of several MAE pre-training experiments on ultrasound image data. All the experiments use a **0.75 masking ratio** setting.

4.2 Augmentation

In all the experiments, we apply random color jitter, random resized cropping and random horizontal flipping augmentations. We found them indeed improve the model robustness during our training experiments.

4.3 Evaluation Metrics

We apply pixel-wise measurement for all of the three following metrics in this work.

IoU Score. In the case of segmentation, Intersection Over Union (IoU) is defined as the intersection of the Ground Truth and Prediction region divided by the union of Ground Truth and Prediction region:

$$IoU = \frac{Intersection}{Union} = \frac{TP}{TP + FP + FN}$$

Recall. Recall (sensitivity) is the proportion of actual positives was identified correctly:

$$Recall = \frac{TP}{TP + FN}$$

Precision. Precision is the proportion of positive identifications was actually correct:

$$Precision = \frac{TP}{TP + FP}$$

4.4 Results

	ViT Encoder Pre-Trained Datasets using MAE	Pre-Training Epochs (IN1K / Ultrasound Dataset)	Layer-Wise lr Decay during Ultrasound Dataset Pre-Training	Mean IoU	Mean Recall	Mean Precision
SF-UNETR	None	- / -	-	7.50	22.61	9.96
	IN1K	800 / -	-	52.68	66.78	71.54
	Ultrasound Dataset	- / 100	No	11.42	44.17	13.34
	1N1K & Ultrasound Dataset	800 / 50	Yes	50.45	62.11	72.69
	1N1K & Ultrasound Dataset	800 / 50	No	37.50	48.15	62.04
	1N1K & Ultrasound Dataset	800 / 100	Yes	47.26	61.37	67.07
Video UNETR All 3 Targets	IN1K	800 / -	-	54.21	69.69	70.81
Video UNETR Middle Target	IN1K	800 / -	-	56.31	69.00	75.49

Table 4. The SF-UNETR and Video UNETR segmentation performance results in terms of ***IoU***, ***Recall*** and ***Precision***. We show the SF-UNETRs performance using different ViT encoder pre-trained weights through several MAE pre-training methods. The best metrics in each input modality (single frame or consecutive frames) are highlighted in red.

SF-UNETR. The segmentation performance of SF-UNETR is shown in Table 4. Based on the results, SF-UNETR seems to have the ability to capture the needle shaft region in an ultrasound image, even though there is no temporal information within them. However, the model performance can only be guaranteed when the ImageNet-1K pre-trained weights for the ViT encoder are applied. During our training experiments, we found that the SF-UNETR would easily overfit to the training data and performed extremely bad on the validation set without ImageNet-1K pre-trained weights (the purple curve shown in Figure 3.), which is not surprising actually since the self-attention mechanism in the ViT encoder is much more complex and introduce nearly no local inductive bias that a convolution layer based on. This overfitting issue becomes severe especially we are working on a relatively small dataset compared to others like ImageNet-1K or even the JFT-300M which the original ViT pre-trained on. Several segmentation results are shown in Figure 7 & 8.

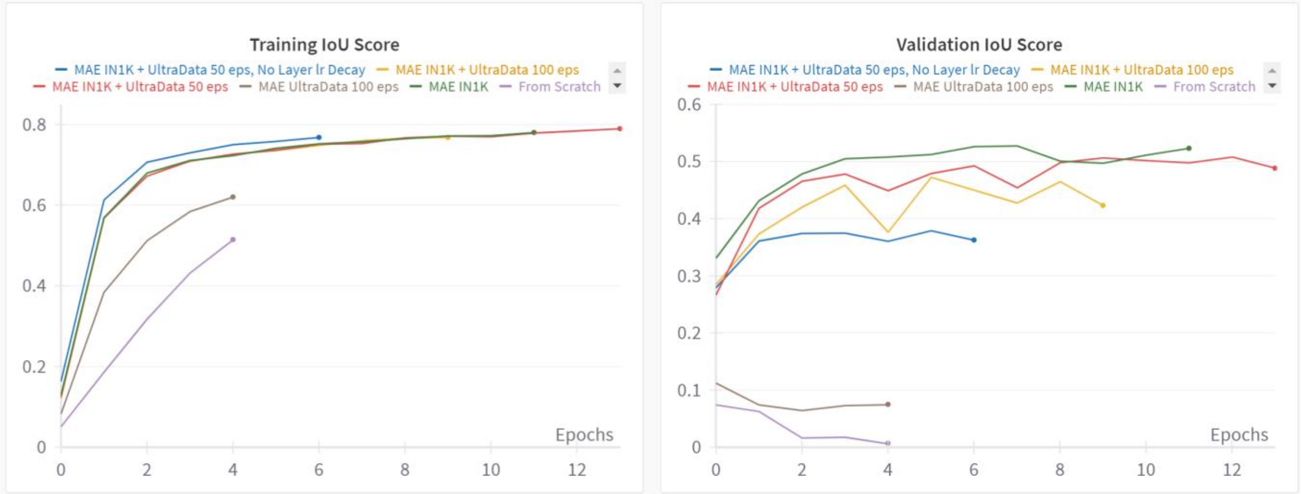

Figure 3. Training IoU curve and validation IoU from SF-UNETR. Each curve represents the SF-UNETR initialized with different pre-trained encoder weights. The details of each pre-training setting are shown in Table 3 and 4.

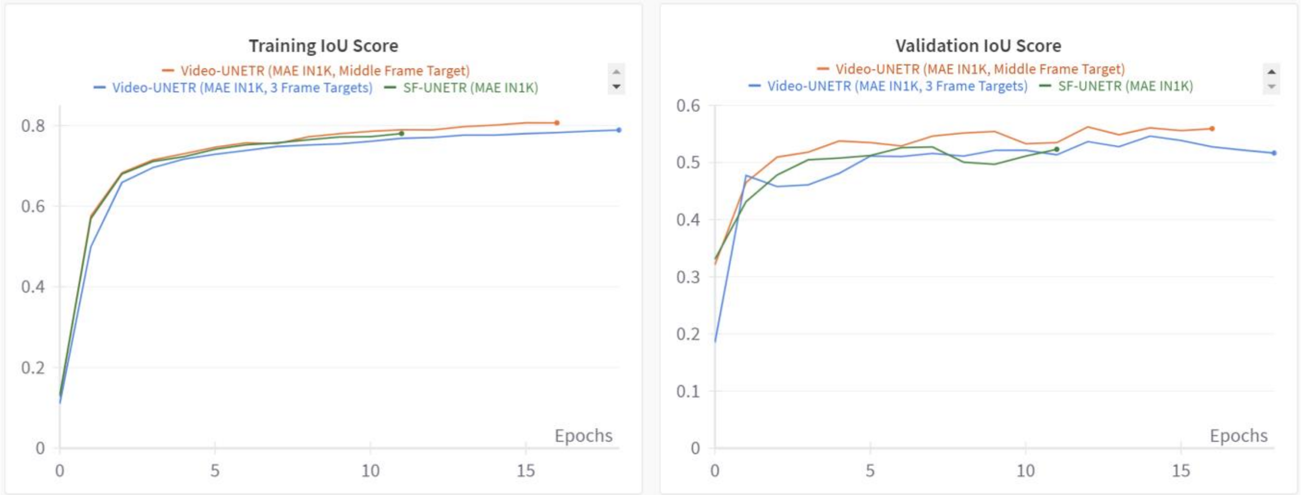

Figure 4. Training IoU curve and validation IoU from SF-UNETR (green), Video UNETR that predicts all 3 frame masks (blue) and the one that predicts only the middle frame mask (orange).

SF-UNETR with MAE Pre-Training on Ultrasound Images. We examine the effectiveness of the MAE pre-trained on the ultrasound dataset. According to both the metrics shown in Table 4 and the learning curves shown in Figure 3, the performance of SF-UNETR using the MAE weights pre-trained on ultrasound dataset only is quite incompetent (the brown curve in Figure 3). Although when compared with training from scratch, MAE weights pre-trained on ultrasound dataset only provide a minor improvement, it is still far behind the baseline that using the MAE weights pre-trained on ImageNet-1K. This might due to the limited dataset size we used for pre-training. A self-supervised method is effective if, and only if, there's enough amount of information lies in the data itself. And our pre-training set, which contains only 11,475 samples, cannot provide useful underlying knowledge in the ultrasound image domain. To our surprise, even when we use the ultrasound dataset to "further pre-train" the MAE weights that already pre-trained on ImageNet-1K, it does not make the downstream segmentation task easier. Instead, it hurts the SF-UNETR performance. Comparing different pre-training methods on ultrasound dataset, we observe that, if we apply a more intense "further pre-train" on ultrasound images, no matter through more epochs or remove the layer-wise learning rate decay setting, the more performance drop we will get when fine-tuning the SF-UNETR using these pre-trained weights. The performance drop might raise from the disruption of originally weights. Since the original ImageNet-1K pre-trained weights are general enough to some extent, if we "further pre-train" the encoder using a small dataset, the weights will have a great chance to overfit to this small and case-specific distribution, which cannot represent the general one of our downstream segmentation tasks.

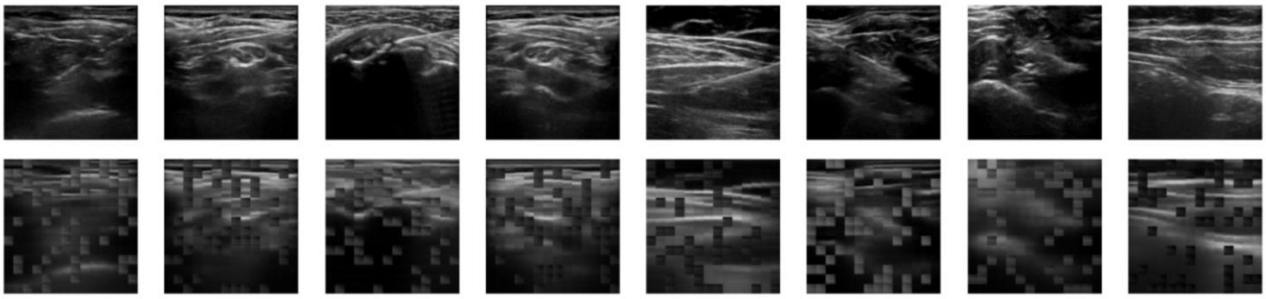

Figure 5. Masked Autoencoder (MAE) reconstruction results. The upper row shows eight original ultrasound images. After random masking (ration=0.75), the autoencoder need to reconstruct the original input given only the visible parts that were not being masked. The bottom row shows the corresponding reconstructed ultrasound images. We use the MAE that pre-trained on the ultrasound dataset from scratch for 100 epochs.

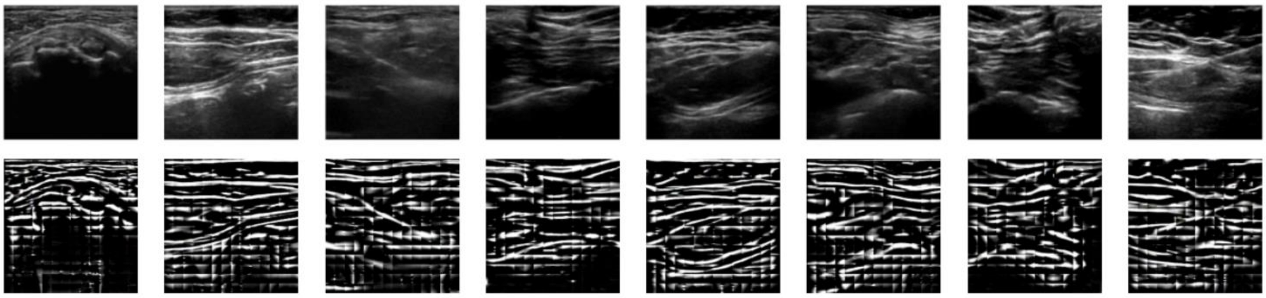

Figure 6. Masked Autoencoder (MAE) reconstruction results. The upper row shows eight original ultrasound images. After random masking (ration=0.75), the autoencoder need to reconstruct the original input given only the visible parts that were not being masked. The bottom row shows the corresponding reconstructed ultrasound images. We use the MAE that

"further pre-train" on the ultrasound dataset for 50 epochs. Note that the MSE loss is calculated through the patch-wise normalized pixel value, therefore the reconstruction results seem to be normalized.

Video UNETR. According to the metrics shown in Table 4, the Video UNETR indeed performs better than SF-UNETR in terms of segmentation IoU, recall and precision. In an ultrasound video, it is easier to track the needle shaft when given the temporal information between frames since the motion flow of the needle and tissues around it provides some useful clues that help the experts, or us, to identify the accurate location of the needle. The Video UNETR aggregates the temporal information through the patch embedding layer at the beginning of the ViT encoder. Interestingly, we find this type

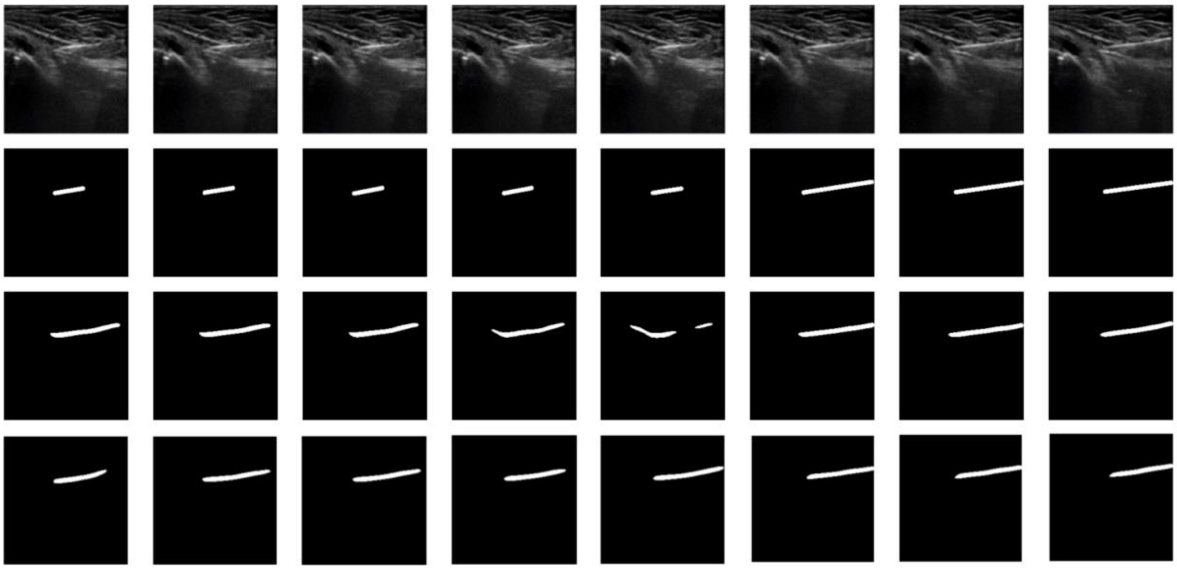

Figure 7. Segmentation results on the validation dataset. The first row shows eight consecutive ultrasound frames. The second row is their corresponding ground truth needle masks. The third row shows the SF-UNETR (MAE pre-trained on IN1K) segmentation results. The last row shows the Video UNETR (middle frame target) prediction results.

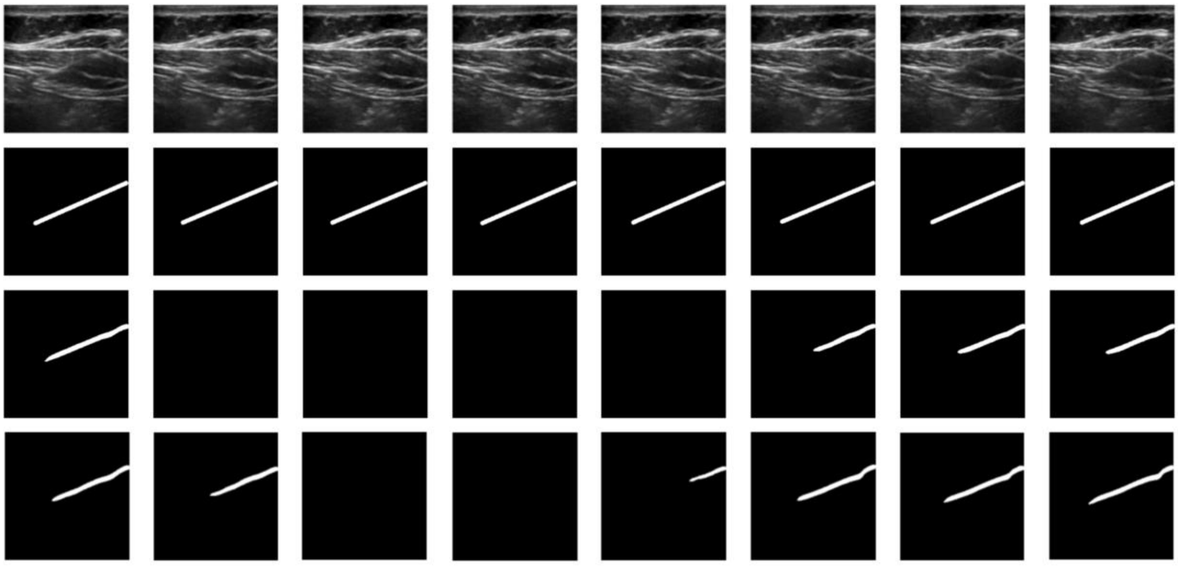

Figure 8. Segmentation results on the validation dataset. The first row shows eight consecutive ultrasound frames. The second row is their corresponding ground truth needle masks. The third row shows the SF-UNETR (MAE pre-trained on IN1K) segmentation results. The last row shows the Video UNETR (middle frame target) prediction results.

of spatiotemporal aggregation is very relatively computation-efficient, since it only projects the spatial information from different time steps to an embedding vector. Thereafter, the self-attention is operated on these aggregated embeddings. Compared to a traditional spatiotemporal self-attention layer that operates on the spatial and temporal domain at the same time without aggregation first, the number of tokens between these ViT blocks would becomes 3 times less in Video UNETR. Therefore, our implementation seems to be a computation efficient way to improve the model segmentation performance. (Even the number of parameters and the floating-point operations per second of Video UNETR are exactly the same with SF-UNETR.) Moreover, the Video-UNETR seems to perform better with target only the middle frame mask.

5. Discussion and Conclusion

In this work, we modify the UNETR architecture, which combines a vision transformer encoder with the U-Net backbone, to reach a certain level of IoU in the ultrasound needle segmentation task. The Single Frame UNETR (SF-UNETR) performs well on single, independent ultrasound frames, while the Video UNETR, which aggregates the temporal information between 3 consecutive frames, further improve the segmentation performance. Furthermore, we investigate the effectiveness of Masked Autoencoder (MAE) pre-training method on a small dataset. Based on our experiments and analysis, MAE pre-training on small ultrasound dataset does not show any benefit to the downstream segmentation task and hurts the performance instead. However, since this work is just a preliminary study on a relatively small dataset, this conclusion about the effectiveness of MAE pre-training on downstream segmentation tasks may not be true for a larger dataset with more computation resources. In the future, applying self-supervised methods on medical image datasets is still our main study direction since learning the underlying knowledge without label might be the key to success in medical image analysis domain where human annotated data are very limited.

References

[1] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, Daguang Xu. UNETR: Transformers for 3D Medical Image Segmentation. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929, 2020.*

[3] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *In Proc. MICCAI, volume 9351 of LNCS, pages 234–241, 2015.*

[4] Kaiming He et al. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377, 2021.*