

Figure 1.The architecture of our work,our pipeline are designed to process the mask auto-encoder probability guided by the segment regions that are given by pretrained segmentation models and build a more meaningful reconstruction task for prostate ultrasound imaging representation learning

**3.Method**

**3.1 Segments in our MAE**

The prostate ultrasound images typically has certain limitations in resolution and the visibility of lesion features[1].It also contain parts with a significant amount of sparse semantic information, and using random masking in the conventional MAE(Masked Autoencoder) may lead the autoencoder(AE) to learn non-critical image information.

For prostate ultrasound images, we prefer the autoencoder to learn the representations of the key parts in prostate ultrasound images. Being able to better reconstruct the key parts also benefits downstream fine-tuning classification tasks.The key area in prostate ultrasound images, which is the lesion area, can be roughly outlined through segmentation[2].In our pipeline, for the images we have sampled, we use several pre-trained large segmentation models[3][4][5] to perform segmentation on the prostate ultrasound images.Our masking strategy is based on the results of three large segmentation models. For the masking probability within the segmented areas, we increase it by a value of .

**3.2 Architecture of our work**

**video partitioning** For the redundant information in video data, we use a sampling stride of 4 to alleviate data redundancy. Suppose a video contains T frames, we sample a total of T/4 frames. Then, for these T/4 frames, we further divide them into segments, with each segment containing 16 frames. Each frame has a size of 3xHxW, where H is the height of the image and W is the width of the image. Before entering the subsequent process, the frame size will be adjusted to 224x224.

**Tokens Generation** To meet the input requirements of the Vision Transformer, we split each segment of 16x3xHxW into 2x3x16x16 tokens. The 16 represents the segment length selected during video segmentation, 3 represents the RGB channels, and H and W represent the height and width of the frame, respectively. After transformation, the frame size is 224x224. We then use 3D convolution with a kernel size of 2x3x16x16 and a stride of 2x16x16. This generates a total of N (N = 16/8 x H/16 x W/16) tokens, where the token dimension is set to d (d = 384). We then apply the fixed 3D periodic positional encoding scheme[6].

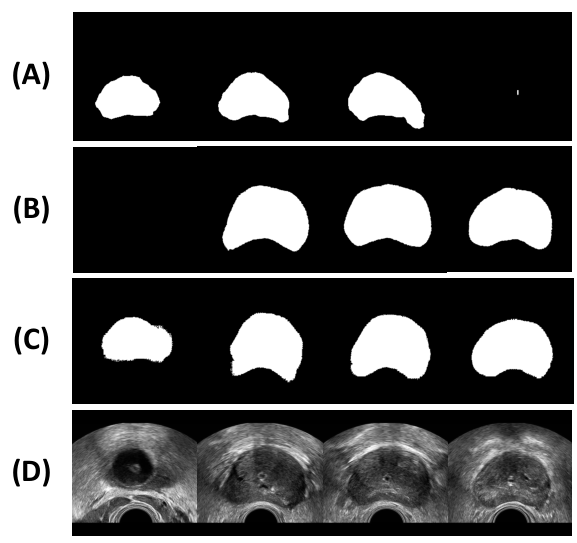
****

Figure 2.The visualization of the Multi-export Segment-Generation.(A)the deeplabv3 network’s segment output.(B)the nnUNet2++ network’s segment output.(C)the MedSAM2 network’s segment output.(D)The raw frames of the prostate ultrasound video.

**Multi-export Segment-Generation** We found that because the models were not fine-tuned on our data, their segmentations often contained errors As the fig2. However, the likelihood of all three models making errors simultaneously was relatively low. We randomly sampled several segmentation results from the training set along with their original prostate ultrasound images and found that in many cases, all three models were able to provide a basic segmentation. In some cases, the ultrasound image did not contain any lesions, and all three models generated completely black segmentation images. This can be explained: non-invasive scans may not have covered the prostate region in certain segments, resulting in the absence of lesions.

To simulate multi-expert diagnosis in medical consultations and reduce the error in large model segmentation, we simulated three experts who each provide masking probabilities and use a gating network to control the final output.For details ,we use three pre-trained medical segmentation models[3][4][5] to obtain the segmentation of each frame and identify the important areas in the prostate ultrasound images represent the i-th model’s segment output.

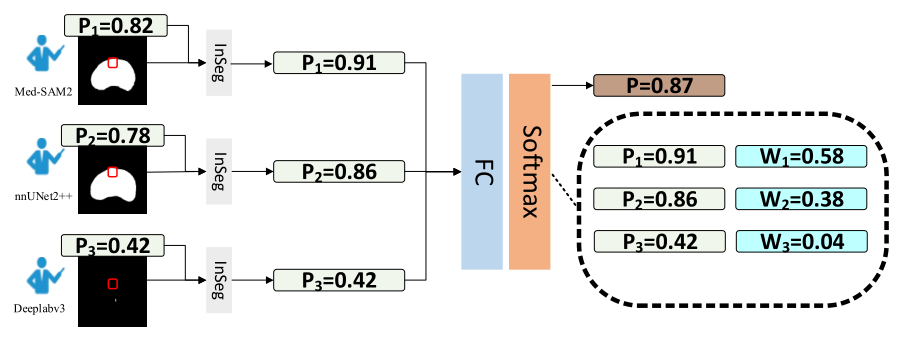
****

Figure 3.The Architecture of Gating Network.

From the perspective of simulated multi-expert consultations in medicine, the gated network we designed reduces the probability of misjudging important areas that require masking in the reconstruction due to the error of any single expert As the fig3 shown.And then for each token, based on the three segmentation results obtained for each frame, we build a gated network based on Multi-Head Attention (MHA) to obtain the masking probability for each token.Given the tokens .Here, N represents the number of tokens generated earlier, and d represents the dimension of the tokens.We have a total of three "consulting experts." Therefore, for the i-th expert (1 <= i <= 3), who provides the probability that the token is masked is generated as follows:

InSeg represents whether the token is within the segmentation area proposed by the corresponding consulting expert. If it is within the area, the masking probability is increased by α (α <= 0.25).

For the final masking probability p of each token, we obtain the weights W for the output of the three experts' final probabilities through the established gated network.

Then, based on the percentage of visible tokens we have set, we filter out a certain number of tokens with the lowest masking probabilities according to this ratio.

**Encoder** From the perspective of improving computational efficiency, we only input the unmasked tokens into the Encoder. The VIT Encoder has 12 layers and an embedding dimension of 384.

Decoder The encoded visible tokens are appended with the masked tokens before passing to the decoder. The masked patches are learnable tokens that the decoder learns to reconstruct, guided by the MSE loss between the values of these tokens and their reconstructions. We keep the decoder depth to 10 after grid searching for optimal depth. The decoder reconstructs the original video cube of size T 2 × H 16 × W 16 from the encoded and masked tokens

**3.3 Training loss function**

**Masking Reconstruction Loss**.We have used the Mean Squared Error loss (MSE) between the predicted and ground-truth RGB values of the masked tokens as the objective function to pretrain the MAE. The loss function is given as:

Here m and M denote the predicted token and the ground-truth RGB values of the token.represent the number of masked tokens

**Token Sampling Loss**. We use a token sampling loss, Lsample, to train the sampling network that generates the sampling probability. We adapt the sampling loss proposed by AdaMAE and use maximization of the average reconstruction error to define the loss. The formulation of such a formulation is motivated by the expected reward maximization of the REINFORCE algorithm in RL. Here, the visible token sampling process is the action, the MAE is the environment, and the masked token reconstruction error is the return. The reconstruction error is high in the high information regions as compared to the low information background regions. Thus, maximizing the expected reconstruction error would result in the network predicting a higher probability score for a high information region. The loss formulation is as follows:

Here denote the Probability of the token to be masked.

**Reference**

1. Akatsuka J, Numata Y, Morikawa H, et al. A data-driven ultrasound approach discriminates pathological high grade prostate cancer[J]. Scientific Reports, 2022, 12(1): 860.
2. Singh R P, Gupta S, Acharya U R. Segmentation of prostate contours for automated diagnosis using ultrasound images: A survey[J]. Journal of computational science, 2017, 21: 223-231.
3. Zhu J, Hamdi A, Qi Y, et al. Medical sam 2: Segment medical images as video via segment anything model 2[J]. arXiv preprint arXiv:2408.00874, 2024.
4. Yurtkulu S C, Şahin Y H, Unal G. Semantic segmentation with extended DeepLabv3 architecture[C]//2019 27th signal processing and communications applications conference (SIU). IEEE, 2019: 1-4.
5. Isensee F, Jaeger P F, Kohl S A A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation[J]. Nature methods, 2021, 18(2): 203-211.
6. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.