

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

1. **Method**

In this section,we first present the motivation and overall framework of the proposed method xxx in section 3.1.Next, we

**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.

**Mask 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].

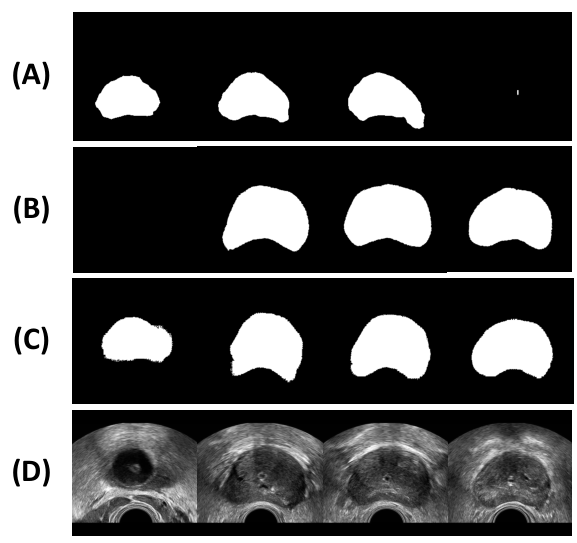
****

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.

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.

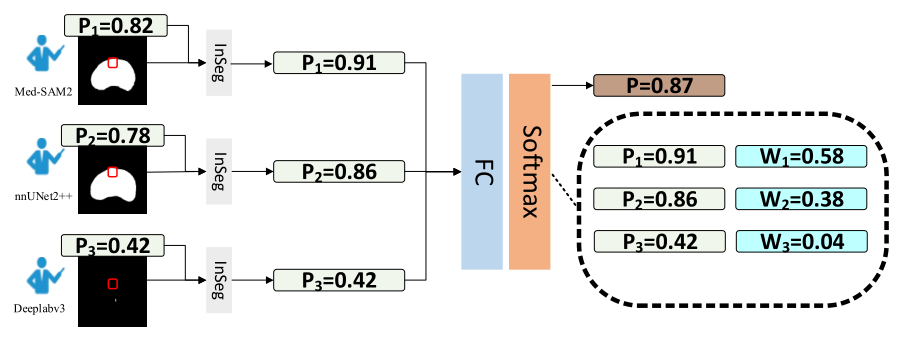
****

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.

**Pretraining 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.

**Smoothness Loss** essentially solves the blockiness problem (also known as the "blocking effect"), where reconstructed images may show unnatural grid-like patterns due to overfitting or discontinuities. By applying smoothness loss, the algorithm encourages the pixels in the image to vary more gradually, leading to more natural-looking, continuous structures.

The vertical (diff\_h) and horizontal (diff\_w) smoothness losses each measure the differences between adjacent pixels in their respective directions. When both losses are minimized, the result is a smoother image with fewer abrupt changes between neighboring pixels, which significantly reduces blocky artifacts.

In summary, smoothness loss helps in reducing discontinuities or sudden transitions between neighboring pixels, thereby making the image appear smoother and more natural. It plays a crucial role in ensuring that the reconstructed or generated images do not exhibit unnatural blocky patterns, especially when dealing with tasks like image denoising or image generation.The loss formulation is as follows:

Here I represents each image of the reconstructed video.

**4.Datasets**

**4.1Internal dataset**

For the Internal dataset,we collected 550 patients from Shanghai Tenth People's Hospital as an internal dataset, and divided them into training and validation sets at a 4:1 ratio. The ratio of T0 and T1 stage prostate cancer patients in both the training and validation sets is close to 1:1, ensuring a relatively balanced distribution.

**4.2 External validation datasets**

For the External validation datasets We have collected prostate video data from 108 patients, sourced from Zhongshan Affiliated Hospital, Ningbo Second Hospital, and Bengbu People's Hospital. The proportion of T0 and T1 stage patients is also close to 1:1.

|  |  |  |  |
| --- | --- | --- | --- |
| datasets | Internal datasets  （train） | Internal datasets  （validation） | External validation  (test) |
| All patients | 440 | 110 | 108 |
| T0 stage patients | 230 | 58 | 56 |
| T1 stage patients | 210 | 52 | 52 |

Table1.Composition of the Internal datasets and External validation datasets

**5.Implementation**

**5.1 Pretraining**

We implemented our experiments using PyTorch,We used Kinetics-710 pretrained weights for MAE weight initialization.Although there is a domain gap in natural and medical image data, studies show that pretraining on natural image data improves network performance on medical imaging tasks.

We chose ViT-Base as the back bone.We set the patch size as 2x3x16x16,which means our token generation will give out N = tokens for an input video clip for 16x3x224x224.We set the mask ratio as 0.9,resulting in 1568\*(1-0.9) =157 tokens that are visible for vision transformer encoder. The pretraining phase is trained with an AdamW optimizer with LR 0.0004,for minimizing the MSE loss over 500 epochs. The batch size is 4. Warm-up was done for 5 epochs with LR 0.001.

**5.2 Fintuning**

As mentioned in Sec. 3,we divide the video into 16frames to constitute a clip.From each video, we sampled 5 clips uniformly.During inference, we predict the labels for each of the clips. If any of the clips is predicted as malignant, the entire video is labelled as malignant.We minimized a cross entropy loss using an AdamW optimizer with LR 0.0004.

We process our experiments with 2 NVIDIA GeForce RTX 3090 GPUs for both pretraining and fintuning.

**6.Experiments and Results**

**6.1Experiment**

For other works that can use our dataset, we classify them into image-based and video-based works. For video-based works, we directly obtain a score from the model for the input validation set videos, and calculate the accuracy, AUC, and ROC curve based on this score and the actual disease status of the videos. For image-based works, we split the video into frames and obtain the model's evaluation of each frame. The final score for a video is obtained by averaging the scores of all frames in the video, and then we calculate the accuracy, AUC, and ROC curve based on this score and the actual disease status of each video.

The result is shown in Table2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | Method | Backbone | Acc | AUC |
| Image-Based | ResNet50 | CNN | 0.726 | 0.739 |
| US\_UCL | 0.609 | 0.713 |
| RadFormer | Transformer | 0.717 | 0.758 |
| PVTv2(SOTA) | 0.783 | 0.829 |
| Video-Based | VideoMAEv2 | Transformer | 0.904 | 0.942 |
| m2clip(SOTA) | 0.906 | 0.949 |
| Nzk-MAE(ours) | **0.913** | **0.953** |

**6.2 Ablation Experiment mask**

As ablation analysis,We present the final accuracy and ROC curve based on the random masking pretraining method, as well as the final accuracy and ROC curve based on the segmentation region masking method in the Figure 3 and Table 3.

|  |  |  |
| --- | --- | --- |
| Method | Accuracy | AUC |
| Random Mask | 0.827 | 0.875 |
| Segment Guided Mask (our work) | **0.913** | **0.953** |

Table 3.Accuracy of Random Mask and Segment Guided Mask (our work)

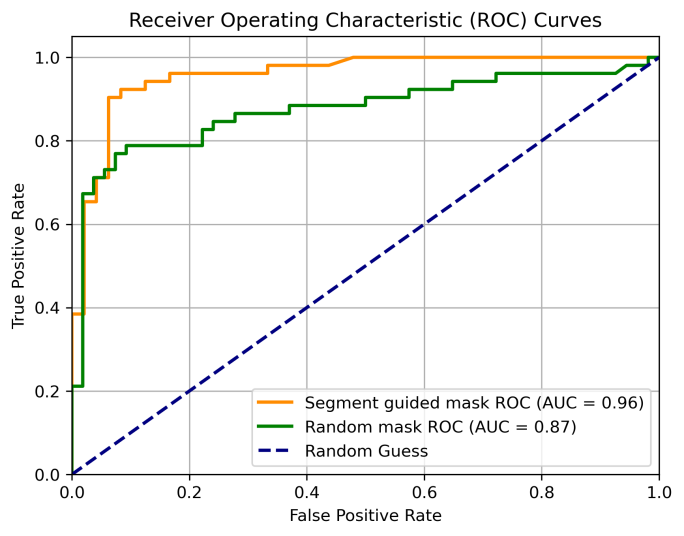


Figure 4.Performance of our work and random mask in AUC.(Internal datasets)

**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.