

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 turn into the details of the proposed multi-exports mask generate network (MMGN) in the pre-training phase and the additional prompt based on MMGN in the downstream classification fine-tuning phase in Section 3.2 and 3.3.

* 1. **Overview**

**3.1.1 Motivation**

Given a prostate ultrasound video contains frames , we aim to address the following two issues:

1. *:Insufficient learning of key region representations during pre-training：*

The mainstream masking methods include random masking, noise-adding masking, block masking , and grid masking. These methods are less effective in utilizing the image characteristics of prostate ultrasound images, leading to insufficient feature extraction in the lesion areas. The ability to extract features from the lesion areas is crucial in fine-tuning for classification and diagnosis.

1. *:Neglect of general features：*

Most downstream classification methods based on masked autoencoders directly use the features extracted by the encoder after the pretraining phase to design the classification head. Using the extracted features in this way can lead to the model ignoring general features , reducing robustness and cross-domain performance.

To address these issues , we propose xxx that mask the tokens by MMGN, which is guided by Multiple pretrained medical expert models and give additional prompt based on MMGN in the downstream classification fine-tuning. This enables us to make the training of our encoder **representation extraction focus on key regions** , while paying more attention to **general features** during classification fine-tuning.

**3.1.2 Architecture of our work**

In this section , we will briefly introduce the overall framework of ourproposed xxx in Fig.1. Given a prostate ultrasound video , which contains frames. The objective of xxx is divide into two parts : pre-training and fine-tuning. We will introduce their objectve and framework one by on. The objective of pre-training phase in our xxx is training the encoder to obtain feature representation of video images , where and denote the height and width of the given video , and means we divide the video into non-overlapping cubic tokens of size .The d represents the dimension for every video ( in our design). And The goal of the fine-tuning phase is to use the features extracted by the encoder to output the classification result of the given video .

*1):pre-training phase：*

Specifically,for the redundant information in video data , we use a sampling stride of 4 to alleviate data redundancy. Suppose a video contains frames , we sample a total of frames. Then , for these T/4 frames, we further divide them into clips, with each clip containing 16 frames. Each frame has a size of , 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 .As a result , we will get the clip with the size of .

We first divide the clip into tokens of size .To achieve this , we use 3D convolution of kernel size is , stride , and d out-put channels.So that we can get the embeded tokens as follows:

(1)

Where represents the token number.

After the patch embedding , the embeded tokens and the clip are given to the MMGN , outputing the mask scores of these tokens as follows :

(2)

Then, we sort the mask scores and mask the tokens with the top-k scores. The k is defined by mask ratio .Then we input the remaining visible tokens that have not been masked into the encoder, which is based on the Vision Transformer architecture and get the encoded visible tokens . The overall process of encode can be represented as follows:

(3)

Finally, the encoded visible tokens is appended with masked tokens to get the combined tokens . Then , is fed into the decoder of the Vision Transformer to obtain the reconstruction result of the token .

(4)

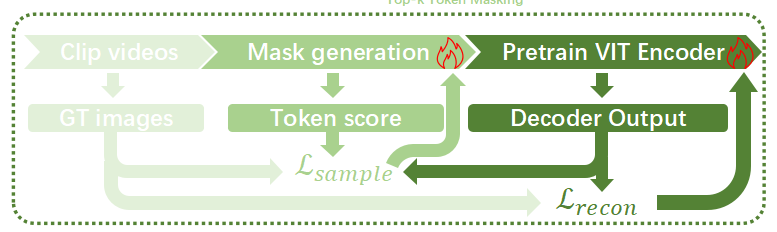


Figure 2.Optimizing process in pre-training phase.The sample loss used to optimize the MMGN while the reconstruction loss used to optimize the encoder and the decoder.

In order to training the encoder to obtain feature representation of video images better. We design the reconstruction loss and sampling loss to optimize the encoder-decoder and MMGN.

**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:

(4)

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

**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.The loss formulation is as follows:

(5)

Here I represents each image of the reconstructed cilp .The final is compute as follows ：

(7)

**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:

(8)

Here denote the Probability of the token to be masked.

*2):fine-tuning phase：*

After the pre-training phase, we freeze the MMGN and encoder that were optimized during pre-training and transfer them to the fine-tuning phase.Given a labeled video with ground-truth label , first we will do the exactly same process as the pre-training phase to get the clip with the size of and token scores .And then we input all the embedded tokens into pretrained vision transformer encoder , The token scores are multiplied by a negative sign and then weighted and multiplied with the deep features from the encoder's output , ultimately leading to the classification result through the classification head , which is designed as a fully connected layer. .The overall process can be represented as follows:

(9)

(10)

(11)

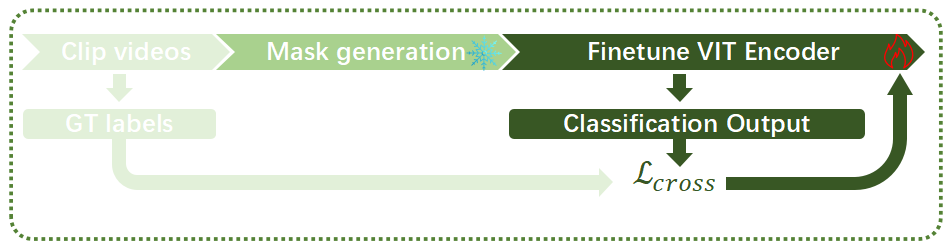


Figure 3.Optimizing process in fine-tuning phase. We frozen the MMGN and the vision transformer encoder.Using crossentroy to optimize the classification head.

In the fine-tuning task, we use cross-entropy to update the parameters of the classification head.

* 1. **MMGN**
  2. **Finetune**

**4.Datasets**

The public datasets of prostate ultrasound video are insufficient in volume to meet the demands of deep learning model’s training.Additionally, Some publicly available datasets are based on 3D ultrasound imaging data from volumetric scans, which are obtained using specialized 3D ultrasound equipment that captures the entire target region in one scan. These datasets provide more complete data and can generate multi-view images. This is different from the application scenario proposed in this work.To address this issue, we collected prostate ultrasound video data from clinical scans conducted by doctors, and we have created a new multi-center dataset.It includes 550 sets of internal training and validation videos, and 108 sets of data from multiple external center hospitals as the test set.

**4.1Internal dataset**

The public datasets of

**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)

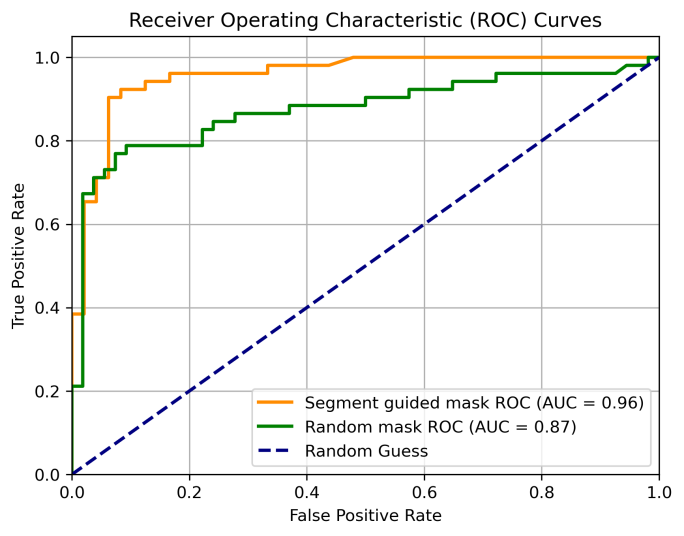


Figure 4.Performance of our work and random mask in AUC.(Internal datasets)

**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.