**ABSTRACT**

Thyroid nodules are frequent and common diseases of the endocrine system, and the incidence rate is usually very high. Therefore, accurate assessment of thyroid nodules is crucial for formulating treatment plans. Accurate nodule segmentation can provide more detailed information on nodule morphology, size, location, etc., which helps doctors make more accurate risk assessments and treatment decisions. However, thyroid nodule ultrasound images often have the characteristics of low contrast and blurred boundaries, which makes it difficult for traditional image segmentation methods to accurately segment the nodule area.In this paper, we propose a thyroid region mask attention and prompt enhancement network for thyroid nodule segmentation. In order to facilitate the development of thyroid nodule segmentation. Our proposed method is evaluated on TN3k dataset, an open-access dataset of thyroid nodule images with nodule masks labeling, and shows outstanding performance compared with existing state-of-the-art algorithms, Code is available at <https://github.com/wanggy820/MedSAM>.

**1.Introduction**

Thyroid nodules are abnormal masses that grow in the thyroid gland, which are an early symptom of thyroid cancer and are becoming more and more common. Statistically speaking, the vast majority of thyroid nodules are benign, and the remaining small fraction is malignant, so the diagnosis of malignant nodules depends largely on the experience of clinicians. Considering that misdiagnosis can easily be caused by factors, more and more deep learning model-assisted diagnosis systems are being developed for the auxiliary diagnosis of thyroid diseases. The automatic segmentation model of thyroid nodules is the basis for building intelligent diagnosis and the prerequisite for accurate diagnosis.

Thyroid nodules come in various shapes and sizes, with small nodules measuring only a few millimeters and large nodules reaching several centimeters. The edges of the nodules are irregular and blurry, and ultrasound images are greatly affected by noise(Fig. 1). The continuous advancement of medical imaging technology, such as the widespread use of high-resolution ultrasound equipment in many fields, has generated a large amount of high-quality thyroid nodule ultrasound image data. These data provide a rich resource for the development of more advanced segmentation algorithms, and also prompt researchers to continuously explore new segmentation methods to make full use of these data. In recent years, deep learning has achieved great success in the field of image segmentation. Its powerful feature extraction and automatic learning capabilities provide new technical means for thyroid nodule ultrasound image segmentation. At the same time, the continuous improvement of computer hardware performance has made large-scale data processing and complex model training possible, further promoting the development of this subject.

In 2023, Meta released Segment Anything Model (SAM). The design and training of the model are flexible, so it can transfer zero-shot to new image distributions and tasks. As a versatile image segmentation base model, SAM has attracted much attention for its remarkable segmentation capabilities and strong zero-shot learning generalization capabilities. Applying SAM to medical image segmentation, especially the segmentation of thyroid nodule ultrasound images, has become a promising research direction. However, recent studies and personal experiments have shown that SAM performs poorly in medical image segmentation due to the lack of medical-specific knowledge. This requires us to enhance SAM's ability to segment medical images.

In SAM, prompts are a way for users to interact with the model and guide it to segment specific targets. By providing different types of prompt information, the model can focus on specific regions or objects in the image, thereby achieving more accurate and tailored segmentation results. This also provides us with a reference method: by freezing the parameters of a model with relatively low performance as prompt input, we can achieve better results.

In this paper, we first use a backbone network to extract the primary mask of thyroid nodules, and then input it into a pyramid shaped pixel encoder to gradually upsample and extract low resolution features to extract high-resolution features, a Transformer encoder that operates on image features to process object queries. Use BPAT-UNet with frozen parameters as a mask prompt, box prompt and point prompt to input into the prompt encoder, and input the Transformer encoder and prompt encoder together into the mask decoder. The entire model architecture is shown in Fig. 2。

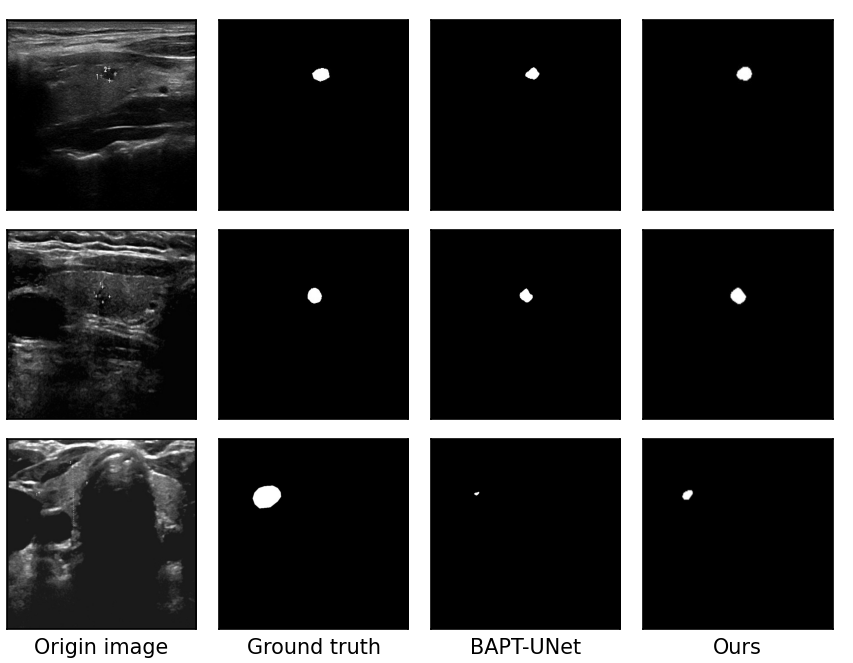


Fig. 1. Some visualization examples of ours MASK-SAM comparing to BPAT-UNet and TRFE+.

**2.Related works**

2.1. Deep learning for thyroid nodule segment

Researchers have continuously improved and optimized deep learning models to improve the accuracy and efficiency of segmentation. For example, by introducing strategies such as attention mechanisms and residual connections, the model's ability to identify the boundaries of thyroid nodules has been improved. Through training with a large amount of ultrasound image data, abnormal areas in the image are automatically identified, and diagnostic suggestions are given to improve diagnostic efficiency and accuracy. Researchers actively use U-Net and its variants in deep learning to segment thyroid nodules in ultrasound images. By adjusting the network structure, increasing the network depth, and changing the size of the convolution kernel, the model's ability to extract nodule features and segmentation accuracy are continuously improved. For example, the skip connection part of U-Net is improved to strengthen the fusion of features at different levels and better handle the complex morphology and boundary information of nodules. The attention mechanism has received great attention in the study of thyroid nodule ultrasound image segmentation. The spatial attention mechanism enables the model to focus on the nodule area and reduce the interference of background noise; the channel attention mechanism can automatically select the most valuable feature channels for nodule segmentation and improve the performance of the model.

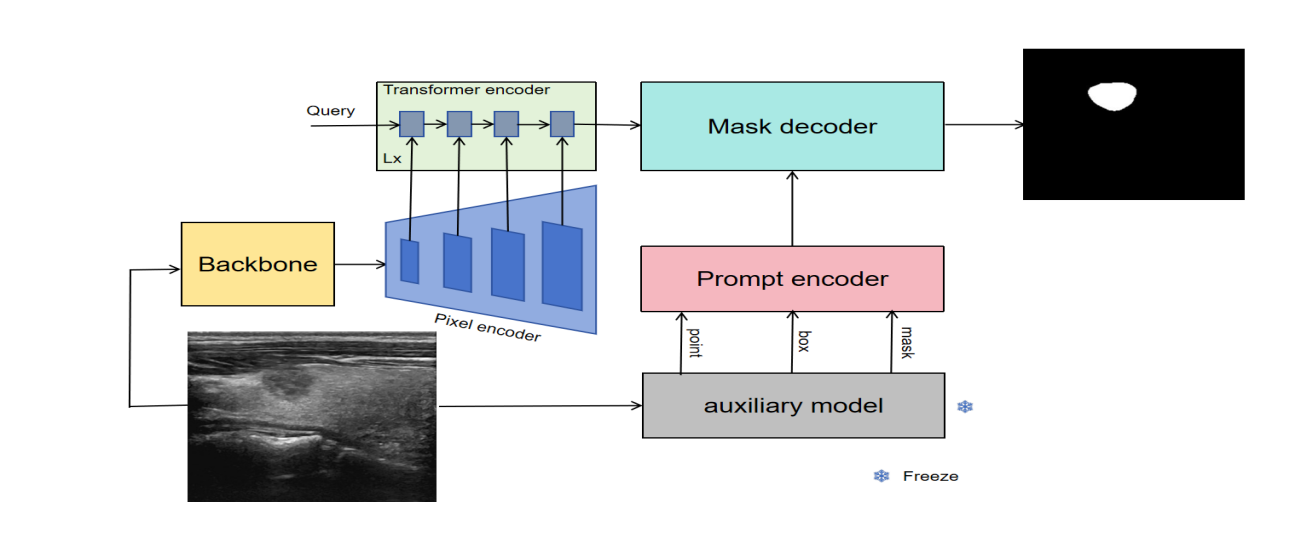


Fig. 2. Entire architecture of our proposed MASK-SAM.

Many research teams have also adopted deep learning algorithms for thyroid nodule ultrasound image segmentation, and have achieved certain results in model selection and improvement. For example, some research teams have proposed improved network structures to improve the segmentation performance of the model in response to the shortcomings of the U-Net model in nodule segmentation. At the same time, researchers are also actively exploring other deep learning models, especially DeepLabV3+, YOLO series and SAM models, which are widely used in the segmentation of thyroid nodule ultrasound images. These models learn the characteristics of thyroid nodules through a large amount of training data and realize the automatic segmentation of ultrasound images.Fei-Fei Li et al. proposed using prompt engineering to segment the specified area. Cheng Chen et al. proposed a parameter-efficient fine-tuning strategy that only needs to update a small part of the weight increment while retaining most of the pre-trained weights of SAM by injecting a series of adapters into the Transformer of the image encoder. Shreyank N Gowda and David A. Clifton proposed adding a parallel network resnet50 to fuse the ViT encoder.

2.2. Overall structure design （method）

The overall model architecture is shown in Fig. 2. Since thyroid nodules have variable shapes and inconsistent sizes, in order to improve the accuracy of the segmentation task, we believe that extract the primary contour features of the thyroid nodules, which can effectively improve the accuracy of the subsequent modules. After a multi-layer pyramid structure, the low-resolution features of the thyroid nodules can be extracted to high-resolution features. The Transformer encoder layer is based on mask-attention, which allows the model to focus on the feature extraction of the mask area of the thyroid nodules. The prompt encoder and mask decoder in SAM remain unchanged, we choose BPAT-UNet as the auxiliary model, all parameters are frozen, and the prediction results are converted to mask prompt, box prompt, point prompt, and input into the prompt encoder layer, and finally input into the mask decoder layer to obtain the final prediction results, and also convert SAM into a fully automatic segmentation model.

### we believe that extract the primary contour features of the thyroid nodules, which can effectively improve the accuracy of the subsequent modules.At the same time, it can also speed up the model training speed. Therefore, we choose to use UNet as the backbone network to extract the primary contour features of the original image with a resolution of 256 x 256.

### Pixel encoder is a pyramid-shaped structure. Its purpose is to facilitate multi-scale feature fusion and better capture objects of different sizes in the image. Small objects may have richer detail information in low-level, high-resolution feature maps, while large objects can better reflect their overall semantic information in high-level, low-resolution feature maps. Through the pyramid structure, these features of different scales are fused, and the model can take into account both the details and the overall structure of the target, thereby improving the detection accuracy of objects of different sizes. Feature maps at different levels contain features of different abstraction levels. Low-level feature maps retain more original image details, such as edges and textures; high-level feature maps focus more on semantic information, such as object categories and overall shapes. The pyramid structure model can combine these features at different levels to form a richer and more comprehensive feature representation, which helps the model better understand the image content and improve the performance of the model. We use ImageEncoderViT in SAM as the Pixel encoder submodule to extract feature maps with resolutions of 128 x 128, 256 x 256, 512 x 512, and 1024 x 1024, respectively. These feature maps are input into the Transformer encoder module.

### The Transformer encoder is a structural module based on masked attention. Masked attention can help the model focus on the target object to be segmented. By generating a mask corresponding to the image, the values of different regions in the mask indicate the importance of the region to the target object. The attention mechanism assigns different attention weights to different regions in the image based on these mask information, so that the model pays more attention to the region where the target object is located and suppresses the interference of irrelevant information such as background, thereby improving the accuracy and precision of segmentation. When processing complex image scenes, the detailed features of the target object are crucial for accurate segmentation. Masked attention can capture the detailed information of different positions in the image through the attention mechanism, and filter and enhance these details according to the mask, which helps the model to more accurately depict the boundaries and internal structures of the target object and improve the quality of the segmentation results.

The prompt encoder and mask decoder remain consistent with SAM, with no changes. However, we added an auxiliary model with frozen parameters to extract the mask prompt, box prompt, and point prompt, and input them into the prompt

encoder layer. We chose BPAT-UNet as the auxiliary model so that the model can automatically obtain the corresponding prompts. The prediction effect of BPAT-UNet is also better, and it can input more accurate prompts, thereby obtaining more accurate mask predictions.

2.3. Loss Function

To train the proposed MASK-SAM, we design a multi-task loss to optimize our MASK-SAM,The formula is defined as:

 (1)

where denotes the loss of the backbone prediction task,denotes the loss of the transformer encoder prediction task,denotes the loss of the overall prediction task, represent the Dice loss function of the nodule segmentation. , , are the coefficients of loss function, We set , ,  to 0.1, 0.2, 0.4 respectively.

**3. DATASET (放到 实验中间)**

In order to facilitate the development of thyroid nodule segmentation, we construct a thyroid nodule region segmentation dataset called TN3k,The dataset includes 3493 ultrasound images from 2421 patients taken between January 2016 and August 2020. These images were selected from more than 30,000 images provided by hospitals according to the following criteria: (1) each image contains at least one thyroid nodule area; (2) lymphatic images or images containing a large number of color areas are excluded; (3) only one representative image is retained from multiple images of the same area or the same patient's perspective. The dataset is divided into a training set and a test set, each with 2879,614 images. Thyroid nodule segmentation in ultrasound images is a valuable and challenging task that is of great significance for the diagnosis of thyroid cancer.

1. **EXPERIMENTS**

4.1. Implementation Details

We have demonstrated that MASK-SAM is an effective thyroid nodule image segmentation architecture by comparing it with state-of-the-art professional architecture based on standard benchmarks. Our experiment is divided into two stages. In the first stage, we first use the backbone (UNet) to train a preliminary contour to accelerate the training of subsequent networks, In the second stage, The network MASK-SAM was built with PyTorch 2.0.1, using the Adam optimizer with a weight decay of 0.1. The learning rate is set to 5*e-*5. All experiments are performed on Tesla A100 60G with batch size set to 2 and a maximum of 50 epochs. and all the images are resized to

256 *×* 256. We use the weights of the TN3k dataset for transfer learning by considering the small number of private thyroid datasets to avoid over-fitting.

4.2. Evaluation metrics

To quantitatively evaluate the segmentation performance of our proposed method, we select the following metrics:

* IoU (Intersection Over Union) = TP/(FP + FN);
* DICE (dice coefficient) = 2\*TP/(FP + FN + 2 \* TP);
* Specificity = TN/(FP + TN);
* PR (Precision) = TP/(TP + FP);
* SE (Sensitivity) = RE (Recall) = TP/(TP + FN);
* Accuracy = (TN + TP)/(TN + TP + FN + FP);
* F1-score = (2\*PR\*RE)/(PR + RE);

where TP, FP, TN, FN indicate true positive, false positive, true negative, and false negative, respectively. The F1-score is the harmonic mean of the precision and recall. Any p-value less than 0.05 demonstrates that the proposed method performs significantly better than the other compared methods.

**4.2. Ablation Study （对比在前，消融在后）**

We divide the ablation experiment into two parts. The data sets are all TN3k. The first part, as shown in Table 1, is for SAM + different auxiliary models, which are converted into point prompt, box prompt, and mask prompt. Keeping other conditions the same, we then compare the results of iou, dice, precision, etc., and verify the effect and feasibility of SAM plus different auxiliary models. From Table 1, we can see that the effect of SAM + different auxiliary models is significantly improved compared with the performance of the far auxiliary model and SAM, and the feasibility of SAM + different auxiliary models is also verified. The second part, as shown in Table 2, is mainly to verify the effects of different modules of MASK-SAM. Table 2 also shows that the backbone has also been improved accordingly.

Table 1

The ablation experiment of SAM with different auxiliary models in TN3k dataset

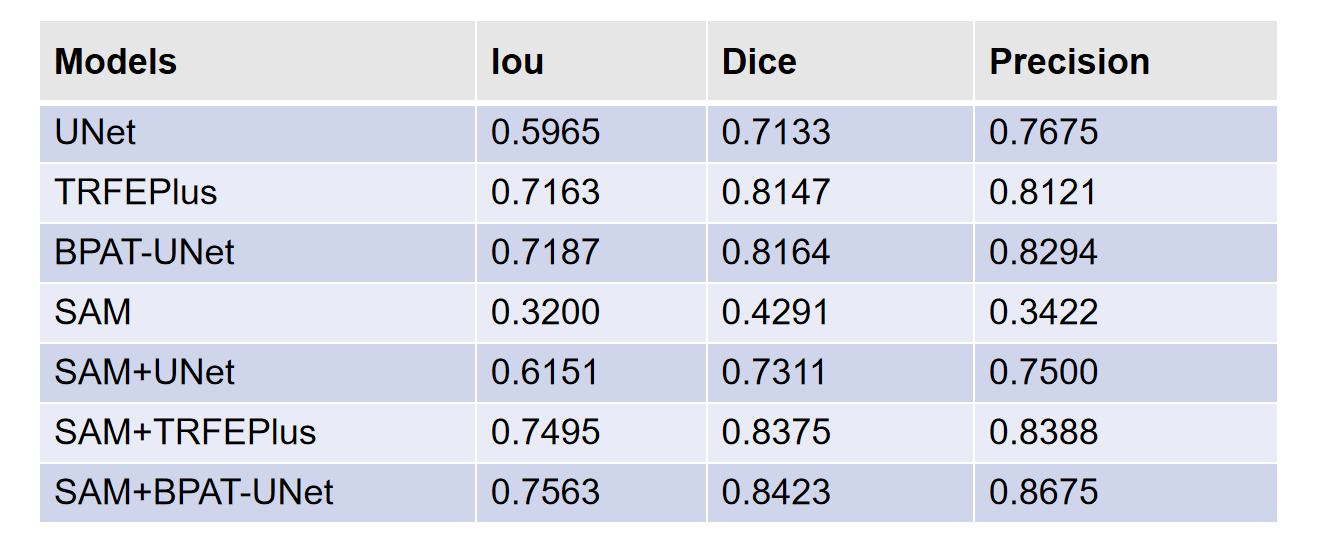
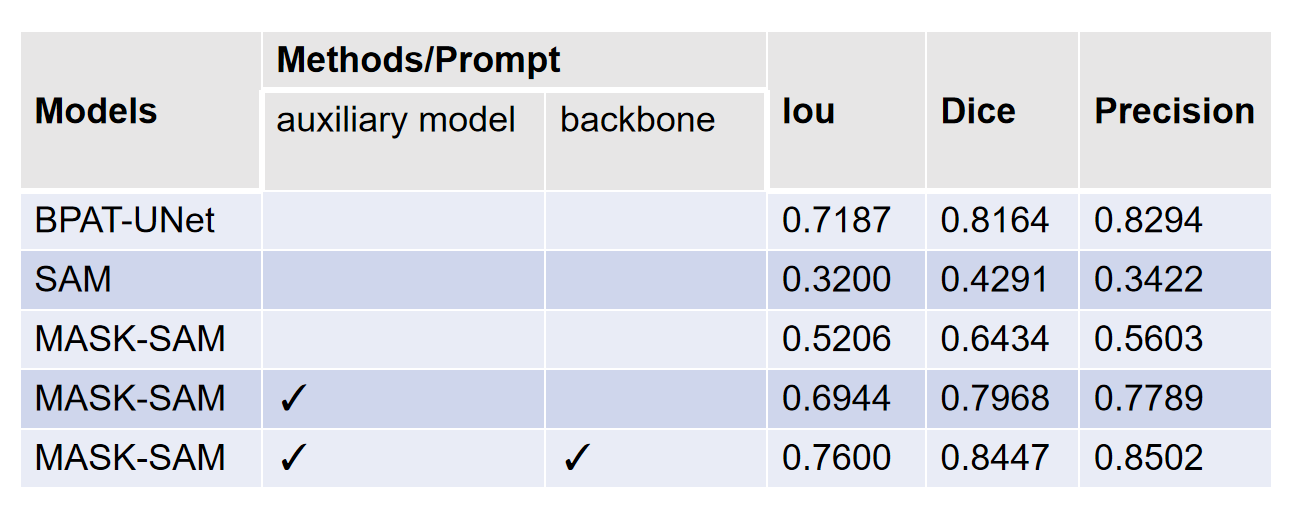


Table 2

The ablation experiment of MASK-SAM with different modules TN3k dataset



**Table 3**

Comparisons with the state-of-the-art semantic segmentation models on the TN3K testset. The best result is shown in **bold**.

**Mask box point 消融**

**难点---图 图文 说明甲状腺分割难**

**4.2 Comparison with the State-of-the-art Methods**

**4.4 Discussion**

**总结 CONCLUSION**

**引用REFERENCES**