###### MaskSAM:Research on thyroid nodule ultrasound image segmentation task based on SAM

ABSTRACT

Thyroid nodules are common diseases of the endocrine system, and the incidence rate is generally high. Therefore, accurate assessment of thyroid nodules is of great significance for formulating scientific and reasonable treatment plans. Accurate nodule segmentation can present more detailed information such as the morphology, size, and location of nodules, providing strong support for doctors to make accurate risk assessments and make treatment decisions. However, ultrasound images of thyroid nodules often show low contrast and blurred boundaries, which makes it difficult for traditional image segmentation methods to accurately segment the nodule area. In view of this, this paper innovatively proposes a thyroid nodule segmentation network based on mask attention and SAM model(MaskSAM). The network integrates a pixel cascade module, a mask attention-based transform encoder module, a prompt encoder, and a mask decoder module. To evaluate the segmentation performance of the MaskSAM in thyroid ultrasound images, we evaluate the proposed method on the TN3k dataset. TN3k is a publicly available thyroid nodule image dataset with nodule mask labels. Experimental results show that compared with existing cutting-edge algorithms, our method shows excellent performance in multiple evaluation indicators, For example, compared to the baseline TRFE+ method in the field of thyroid nodule ultrasound imaging, MaskSAM improvement in Recall, Accuracy, IoU，Dice, and Precision. Code is available at <https://github.com/wanggy820/mask_sam>.

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. According to the paper on Thyroid nodules: diagnosis and management[1]， The prevalence of thyroid nodules in the general population is close to 25%, but thyroid nodules are mostly benign, with malignant tumors accounting for about 5% to 10%, 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.This also poses a significant challenge to the segmentation accuracy of the segmentation model.

In the key areas of medical research and clinical practice, the morphological characteristics of thyroid nodules show amazing diversity. Tiny nodules are like grains of sand, with a diameter of only a few millimeters, while larger nodules are like small lumps, up to several centimeters. The edges of these nodules are complex, irregular and fuzzy, which makes the image quality extremely susceptible to noise interference during ultrasound imaging. This noise interference not only reduces the clarity of the image, but also increases the difficulty for doctors to accurately identify and analyze nodules.

With the rapid development of science and technology, medical imaging technology has made continuous breakthroughs. High-resolution ultrasound equipment has been widely used in many medical fields such as thyroid disease diagnosis and disease monitoring due to its excellent imaging capabilities. The popularity of these devices has led to an exponential growth in thyroid nodule ultrasound image data, forming a massive and high-quality data set. These rich data resources, like treasures of great value, provide solid data support for researchers to develop more advanced and accurate segmentation algorithms. Faced with such rich data, researchers are greatly motivated to continuously explore new segmentation methods. They hope to fully explore the information hidden behind these data through innovative algorithms, so as to provide stronger support for the diagnosis and treatment of thyroid nodules.In recent years, deep learning, as a core technology in the field of artificial intelligence, has achieved remarkable achievements in the field of image segmentation. With its powerful feature extraction ability, deep learning algorithms can automatically learn and identify various features in thyroid nodule ultrasound images, including the shape, size, boundary and internal structure of nodules. This automatic learning ability has opened up a new technical path for thyroid nodule ultrasound image segmentation, greatly improving the segmentation accuracy and efficiency. At the same time, the continuous leap in computer hardware performance, especially the widespread application of high-performance GPUs, has made large-scale data processing and the training of complex deep learning models efficient and feasible. The improvement of hardware performance not only shortens the time of model training, but also supports more complex model structures, thereby further promoting the in-depth development of thyroid nodule ultrasound image segmentation technology.

1.1. Deep learning for thyroid nodule segment

In the field of thyroid nodule ultrasound image segmentation, researchers have always been committed to the improvement and optimization of deep learning models, aiming to significantly improve the accuracy and efficiency of segmentation. In recent years, with the rapid development of deep learning technology, various strategies have been continuously introduced into the construction of thyroid nodule segmentation models. Attention mechanism and residual connection are the most popular optimization strategies. The introduction of attention mechanism is like giving the model the ability to "focus", which enables the model to focus more on the thyroid nodule area when processing images, thereby effectively reducing the interference of background noise. For example, the spatial attention mechanism can guide the model to focus on the spatial position of the nodule area and accurately capture the boundary of the nodule; the channel attention mechanism automatically selects the most valuable feature channels for nodule segmentation, making the model more targeted when processing image information, significantly improving the performance of the model. The addition of residual connection effectively solves the problem of gradient disappearance in the deep neural network during training, allowing the model to maintain good training effects and generalization capabilities while deepening the number of network layers. Through the comprehensive application of these strategies, the model's ability to identify the boundaries of thyroid nodules has been significantly improved. In terms of data training, a large amount of ultrasound image data provides a solid foundation for model training. By learning from massive amounts of data, the researchers enabled the model to automatically identify abnormal areas in the image, and combined with preset diagnostic rules and knowledge graphs, give preliminary diagnostic recommendations. This process not only greatly improves diagnostic efficiency, but also reduces errors caused by human factors, making the diagnostic results more accurate and reliable. In terms of model architecture selection, U-Net and its variants have been widely used in thyroid nodule ultrasound image segmentation due to their unique encoder-decoder structure. By carefully adjusting the network structure, such as increasing the network depth, the researchers were able to enable the model to learn more advanced and abstract nodule features; by changing the size of the convolution kernel, the model's ability to extract features of different scales can be adjusted, thereby better adapting to the diversity of thyroid nodule morphology and size. By strengthening the fusion of features at different levels, the model can better handle the complex morphology and boundary information of nodules, further improving the accuracy of nodule segmentation.

At the same time, researchers are actively expanding the application scope of deep learning models, and DeepLabV3+[2], YOLO[3, 4, 5, 6, 7] series and SAM models have become research hotspots. DeepLabV3+, with its dilated convolution technology, can expand the receptive field without losing resolution, effectively capture the contextual information of thyroid nodules, and achieve accurate segmentation; the YOLO series of models, with its fast detection speed, can complete the screening of a large number of ultrasound images in a short time and locate the location of thyroid nodules; the SAM model, with its powerful segmentation generalization ability, can understand the characteristics of thyroid nodules in different scenarios through learning a large amount of image data, and performs well in the segmentation of thyroid nodules ultrasound images. These models all deeply learn the characteristics of thyroid nodules through massive training data and realize the automatic segmentation of ultrasound images. For example, Fei-Fei Li et al.[8] innovatively proposed a method of segmenting a specified area using fast engineering. This method can quickly lock the area where the thyroid nodule is located and perform accurate segmentation through specific algorithm optimization, greatly shortening the segmentation time and improving the segmentation efficiency. Cheng-Chen et al.[9] proposed a parameter-efficient fine-tuning strategy that only needs to update a small part of the weight increment, and inject a series of adapters into the Transformer of the image encoder to retain most of the pre-trained weights of the SAM model. This approach not only ensures the adaptability of the model in the task of thyroid nodule segmentation, but also avoids the waste of resources and time caused by large-scale retraining. Shreyank N Gowda and David A. Clifton[10] suggested adding a parallel network resnet50[11] to fuse the ViT[12] encoder. In this way, the advantages of different networks can be integrated, making the model more comprehensive in learning nodule features, and further improving the accuracy and stability of segmentation.

1.2. Segment Anything Model for thyroid nodule segment

Meta's Segment Anything Model (SAM)[13, 14] released in 2023 has undoubtedly become the focus of attention. The design and training philosophy of SAM are clever and flexible. It is easy to achieve zero sample conversion into new image distributions and tasks, breaking the limitation of conventional models where one dataset corresponds to one model weight. As a general image segmentation basic model, it has quickly set off a research boom in academia and industry with its significant segmentation ability and strong zero-sample learning generalization ability, and has become the object of exploration for many scientific researchers. The introduction of SAM into the field of medical image segmentation[15, 16, 17, 18], especially its application in the segmentation of thyroid nodule ultrasound images, provides a new segmentation method for this field and has become an important direction with great research value and investment.

However, recent extensive research and the experimental results of many researchers clearly indicate that due to a lack of in-depth understanding and mastery of medical expertise, the performance of SAM in medical image segmentation has not achieved the expected ideal results, and some are not satisfactory. Medical imaging has its unique complexity and high level of professionalism. For the original SAM model, these are insurmountable obstacles. For example, in the process of determining the benign or malignant nature of thyroid nodules, doctors often need to rely on in-depth analysis of detailed features such as the smoothness of nodule boundaries and the uniformity of internal echoes. However, SAM is clearly inadequate in handling these critical information, making it difficult to accurately identify and judge, which to some extent limits its application and promotion in the field of medical image segmentation. Therefore, taking effective measures to improve the segmentation performance of SAM in medical images has become an important issue that urgently needs to be addressed.

In SAM, the prompt [19, 20] encoding module serves as a unique encoding module for users to interact with the model and guide users within a selected range, providing us with a way to optimize model performance. By providing different types of prompt information such as points, boxes, masks, etc., users can mark the key positions of the model on the map. Then, the model can focus on specific regions in the image based on these prompts, thereby obtaining more accurate segmentation results. This provides us with a valuable reference method and innovative idea: by freezing model parameters with relatively low performance as prompt inputs, we can obtain more ideal segmentation results.For example, we can carefully extract key feature parameters from existing medical image segmentation models and input them into SAM as valuable prompt information to guide it to focus on important medical features in ultrasound images of thyroid nodules. In this way, it may be possible to effectively improve the accuracy and reliability of segmentation.

1.3. Dataset

To validate the performance of thyroid nodule segmentation, we used a thyroid nodule region segmentation dataset called TN3k[21]. This dataset includes 3493 ultrasound images taken from 2421 patients between January 2016 and August 2020. These images were selected from over 30000 images provided by the hospital based on the following criteria: (1) each image contains at least one thyroid nodule area; (2) Exclude lymphatic images or images containing a large number of colored regions; (3) Only one representative image is retained from multiple images of the same region or patient perspective.

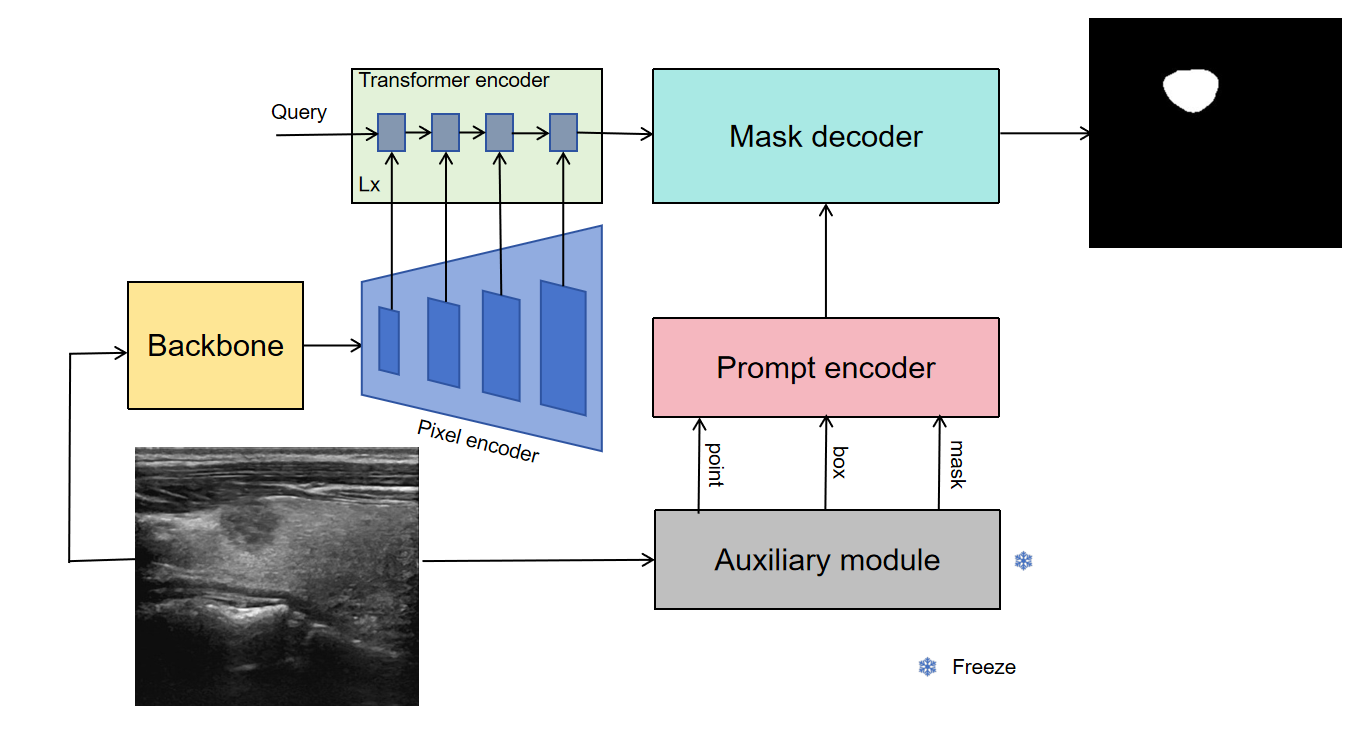
1. Method

In this section, Fig. 1 shows the overall architecture of our proposed MaskSAM, and we describe the six important modules of MaskSAM in the next section. (1) Backbone is mainly used to extract primary edge contour features. (2) Pixel encoder extracts features of different resolutions, enabling the model to understand data more comprehensively and accurately, and improving its ability to recognize complex patterns. (3) Transformer encoder can selectively focus on the mask part of the input data through mask attention, ignoring other irrelevant or interfering information, allowing the model to process key information more accurately and improve its ability to capture important features. (4) The Prompt encoder and mask decoder use the structure of SAM without any changes. (5) The Auxiliary module is a weight trained by BPAT-UNet and frozen to extract more accurate predicted masks. Then, the predicted masks are converted into prompts and input into the Prompt encoder module.

2.1. Overall structure design

Due to the varying shapes and sizes of thyroid nodules, this poses a significant challenge for segmentation tasks. Extracting contour features of thyroid nodules can help improve the accuracy of segmentation tasks. In order to effectively improve the accuracy of the segmentation task, we have conducted in-depth analysis, accurately extracting the main contour features of thyroid nodules is the key. Because these contour features contain the key morphological information of the nodules, they can provide a solid foundation for subsequent model processing, thereby effectively improving the accuracy of subsequent modules.

In order to improve the accuracy of segmentation of thyroid nodule ultrasound images and the performance of the model, we used a new deep learning model architecture to accurately extract thyroid nodule features in this study. Firstly, we chose U-net [22] as the backbone network and used it to perform preliminary feature mask extraction on the input image. This mask serves as key information and provides a foundation for subsequent processing. Then, the extracted mask is input into the cascade pixel encoder. The encoder refines the low-resolution features by gradually upsampling. In this process, it continuously mines and extracts key information from the low-resolution features to obtain high-resolution features. It is worth mentioning that the Transformer encoder is also introduced here. It focuses on deep operations on image features and can effectively process object queries, thereby further improving the accuracy of nodule feature extraction. In the prompt encoder link, we used BPAT-UNet[23] with frozen parameters. By inputting mask prompts, box prompts, and point prompts into the network, we can make full use of its advantages in image segmentation tasks and provide rich and accurate prompt information for the subsequent decoding process. Finally, the output results of the Transformer encoder and the prompt encoder are jointly input into the mask decoder. The mask decoder integrates information from various aspects and finally generates accurate thyroid nodule segmentation results. The design of the entire model architecture is closely centered around the feature extraction and segmentation tasks of thyroid nodules.



**Fig. 1.** Entire architecture of our proposed MaskSAM

2.1. The Backbone module

We need to choose a suitable backbone network to extract the preliminary hub of thyroid nodule ultrasound images, roughly determine the range of the mask, and U-net adopts a symmetrical encoder decoder structure. The encoder is responsible for downsampling the input image, extracting image features, and the decoder restores the low resolution feature map to the size of the original image through upsampling, achieving image segmentation. This structure can effectively integrate feature information of different scales, capturing both global contextual information in the image and preserving detailed features. It is particularly effective for segmenting targets of different sizes and shapes in thyroid nodule ultrasound images. At the same time, there are a large number of skip connections in U-net, which means that the feature maps at different levels in the encoder will be directly connected to the corresponding levels in the decoder. These skip connections can directly transmit the low-level detailed information to the high-level, avoiding the loss of detailed information in the encoding and decoding process, enabling the model to more accurately locate the boundaries of the target during segmentation and improve segmentation accuracy, especially for the segmentation of tissues and organs with complex boundaries and rich details in medical images, which is of great significance.

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 U-net as the backbone module to extract the primary contour features of the original image with a resolution of 256 x 256.

2.2. The Pixel encoder module

In the model design, we adopted a multi-layer cascade structure[24]. This structure has a unique advantage that it can gradually process the image features of thyroid nodules. Through multi-layer hierarchical conversion, the original low-resolution thyroid nodule features can be effectively extracted and converted into high-resolution features. In this process, features of different scales are fully integrated, allowing the model to capture various details of nodules from macro to micro, preparing for subsequent accurate segmentation.

Pixel encoder is a cascade-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 cascade 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 cascade 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.

2.3. The Transformer encoder module

The Transformer encoder module is one of the core components of the entire model architecture, which is built based on the mask attention[25] mechanism. This mask attention-based design enables the model to focus on feature extraction of the mask area of thyroid nodules when processing images. It is like giving the model a "magnifying glass" that can eliminate the interference of other irrelevant background information in the image and focus its main energy on the nodule area, thereby extracting more targeted and effective features, further improving the model's ability to identify and segment nodules.

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.

2.4. The Prompt encoder module and mask decoder module

In the model, we retain the prompt encoder and mask decoder parts of SAM and keep their structure and functions unchanged. This is because SAM has shown strong generalization ability and basic performance in the field of image segmentation. Its prompt encoder can effectively encode various input prompt information, and the mask decoder can generate accurate segmentation masks based on the encoded information.

2.5. The Auxiliary module

In order to improve the performance of the model, we need to choose a suitable model as the prompt for SAM. BPAT-UNet has a unique network structure and feature extraction ability, and also has certain advantages in thyroid nodule segmentation tasks. Therefore, we choose BPAT-UNet as the auxiliary module. During use, we froze all parameters of BPAT-UNet. The purpose of doing so is to avoid unnecessary updates to its parameters during model training and ensure its stability. The prediction results of BPAT-UNet will be further converted into three forms: mask prompt, box prompt, and point prompt. These different forms of prompt information contain rich information, such as nodule location and shape. They are input to the prompt encoder module, encoded by the prompt encoder, and then input to the mask decoder module. Through this process, accurate final prediction results were ultimately obtained. At the same time, the SAM model has successfully transformed from a segmentation model that initially required manual interaction to a fully automatic segmentation model that can automatically complete segmentation tasks, greatly improving segmentation efficiency and practicality.

2.6. Loss Function

The Soft Dice loss is used to train our MaskSAM, to train the proposed MaskSAM, we design a multi-task loss to optimize our MaskSAM,The formula is defined as:

 (1)

 (2)

 (3)

 (4)

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,denotes the prediction for backbone module,denotes the prediction for transformer encoder module,denotes the overall prediction task,denotes the label image, represent the Dice loss function of the nodule segmentation.

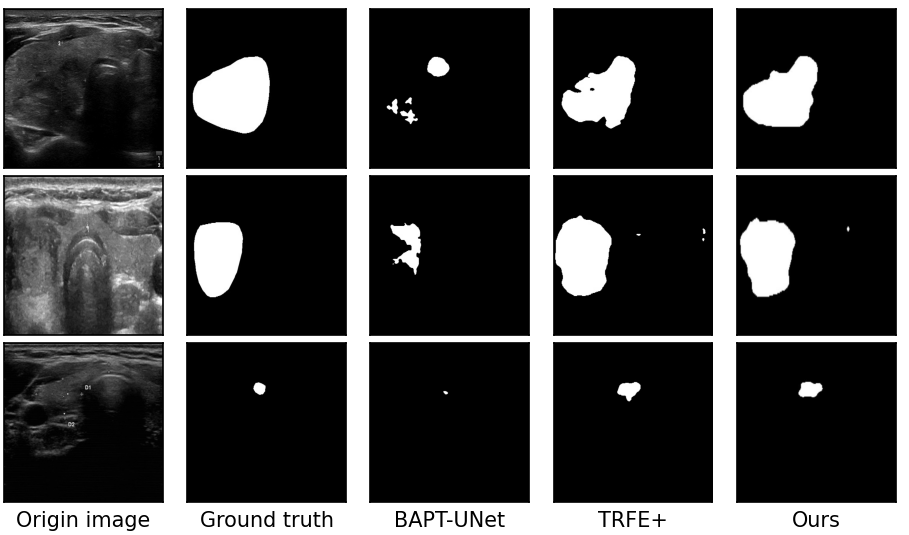
1. Experiments

3.1. Implementation Details

We have demonstrated that MaskSAM 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 (U-net) to train a preliminary contour to accelerate the training of subsequent networks, In the second stage, The network MaskSAM 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.

3.2. Result on the TN3k dataset

Firstly, we evaluated these methods on the publicly available TN3k dataset. Fig. 2 shows the segmentation results of three example images from the test dataset obtained by our proposed MaskSAM, BPAT-UNet, and TRFE+. The segmentation results provided by these methods are displayed in the third to fifth columns, respectively. The first to third rows in Fig. 2 show the segmentation results of three patients. Obviously, the proposed MaskSAM is the most effective method for visually detecting thyroid nodules from ultrasound images. From the segmentation effect of the small object in Fig. 2, compared with other models, our MaskSAM has a stronger perception ability for small thyroid nodules. Therefore, our MaskSAM compensates for smaller features that are easily lost in downsampling. For large thyroid subjects (lines 1 and 2 in Fig. 2), MaskSAM also performed well.



**Fig. 2.** Some visualization examples of ours MaskSAM comparing to BPAT-UNet and TRFEP+.

3.2. Evaluation metrics

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

 (5)

 (6)

 (7)

 (8)

 (9)

 (10)

 (11)

 (12)

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. d represents the calculation of one-way Hausdorff distance of two sets.X and Y stand for prediction and ground truth, respectively.

3.3. Comparison with the State-of-the-art Methods

As shown in Table 1, in order to comprehensively evaluate the performance of the proposed MaskSAM model in the thyroid nodule segmentation task, we conducted in-depth comparative experiments with current mainstream deep learning models such as U-net, TRFE[21], DeeplabV3+, CPFNet[26], FCN[27], BPAT-UNet, and TRFE+[28] on the TN3K dataset. In the comparison of various evaluation indicators, MaskSAM showed excellent performance in TN3k dataset. Compared to baseline method TRFE+, MaskSAM yields 1.72%, 0.13%, 4.37%, 3.00%, 3.81% improvement in Recall, Accuracy, IoU, Dice, and Precision. This means that the model rarely misses real nodules when detecting nodules, and can provide doctors with more comprehensive nodule information. However, in the HD95 (Hausdorff Distance 95%) indicator, MaskSAM is slightly lower than TRFE+. The reason why MaskSAM can perform well in multiple key indicators such as recall, Iou, and Dice is due to its unique prompt and mask attention mechanism. This mechanism enables the model to focus on the nodule area more accurately when processing ultrasound images of thyroid nodules, especially when faced with complex images such as low contrast, blurred boundaries, or extremely complex nodule morphology. The prompt mechanism can guide the model to quickly locate areas where nodules may exist, while the masking attention mechanism enables the model to effectively eliminate irrelevant information interference such as background noise, thereby better extracting image features. This powerful feature extraction capability enables the model to have a stronger ability to distinguish between different categories of features. Whether it is the distinction between normal thyroid tissue and nodule tissue, or the identification of different types of nodule features, MaskSAM can handle it relatively well, thereby achieving excellent results in multiple evaluation indicators.

**Table 1**

Comparisons with the state-of-the-art segmentation models on the public TN3k dataset. The best result is shown in **bold**..

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Models | **Recall** | **Accuracy** | **IoU** | **Dice** | **HD95** | **Precision** |
| UNet | 0.8007 | 0.9617 | 0.6292 | 0.7379 | 20.5526 | 0.7652 |
| TRFE | 0.8182 | 0.9691 | 0.6850 | 0.7817 | 16.7248 | 0.8248 |
| Deeplabv3+ | 0.4816 | 0.9230 | 0.3638 | 0.4884 | 47.2080 | 0.6848 |
| CPFNet | 0.8101 | 0.9633 | 0.6350 | 0.7490 | 17.1809 | 0.7695 |
| FCN | 0.6564 | 0.8556 | 0.5475 | 0.6760 | 32.0199 | 0.8245 |
| BPAT-UNet | 0.8557 | 0.9722 | 0.7187 | 0.8164 | 14.0578 | 0.8294 |
| TRFE+ | 0.8719 | 0.9713 | 0.7163 | 0.8147 | **12.2701** | 0.8121 |
| MASK-SAM | **0.8891** | **0.9726** | **0.7600** | **0.8447** | 14.3962 | **0.8502** |

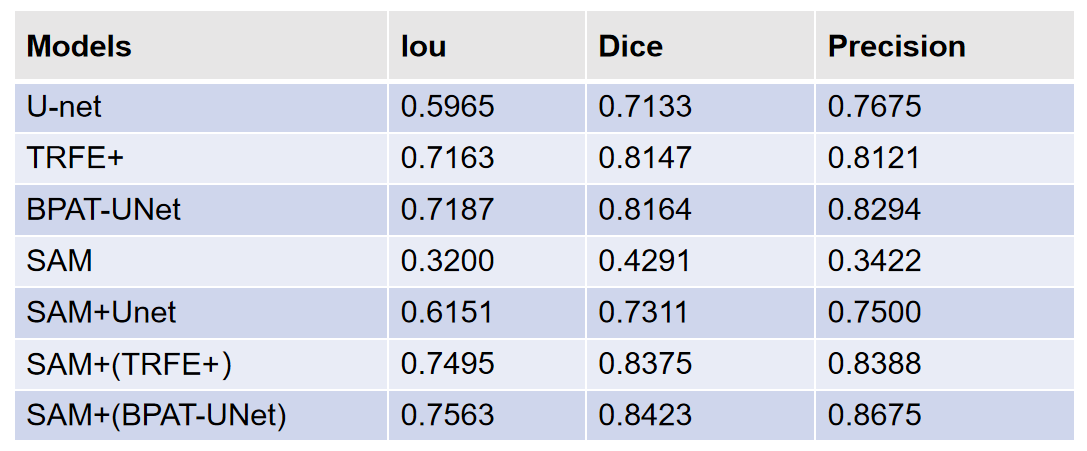
3.4. Ablation Study

In order to further explore the roles of each component of the model and the effects of different combinations, we carefully designed and conducted ablation experiments. The entire ablation experiment is divided into two key parts, both of which are supported by the TN3k dataset as experimental data. The TN3k dataset covers a wealth of thyroid nodule ultrasound images and corresponding nodule mask labels, providing experimental data for the accuracy and reliability of experimental results.

The first part of the ablation experiment focuses on studying the effect and feasibility of combining SAM with different auxiliary modules. From the data results in Table 2, it can be seen that SAM+Unet has improved by 29.51%, 30.20%, and 40.78% in Iou, Dice, and Precision compared to SAM. Similarly, SAM+Unet has improved by 1.86%, 1.78% and -1.75% in Iou, Dice, and Precision compared to U-net, SAM+(TRFE+) improved by 42.95%, 40.84%, 49.66% compared to SAM in Iou, Dice, and Precision, respectively. SAM+(TRFE+) improved by 3.32%, 2.28%, 2.67% compared to TRFE+ in Iou, Dice, and Precision, respectively. SAM+(BPAT-UNet) improved by 43.63%, 41.32%, and 52.53% compared to SAM in Iou, Dice, and Precision, respectively. SAM+(BPAT-UNet) improved by 3.76%, 2.59%, and 3.89% compared to BPAT-UNet in Iou, Dice, and Precision, respectively. Compared with the performance of using auxiliary modules alone or SAM itself, the effect of combining SAM with different auxiliary modules is significant. Significant improvement has been achieved. This result not only visually demonstrates the gain effect of different auxiliary modules on SAM, but also fully verifies the feasibility of combining SAM with various auxiliary modules, providing a strong practical basis for model optimization and improvement.

**Table 2**

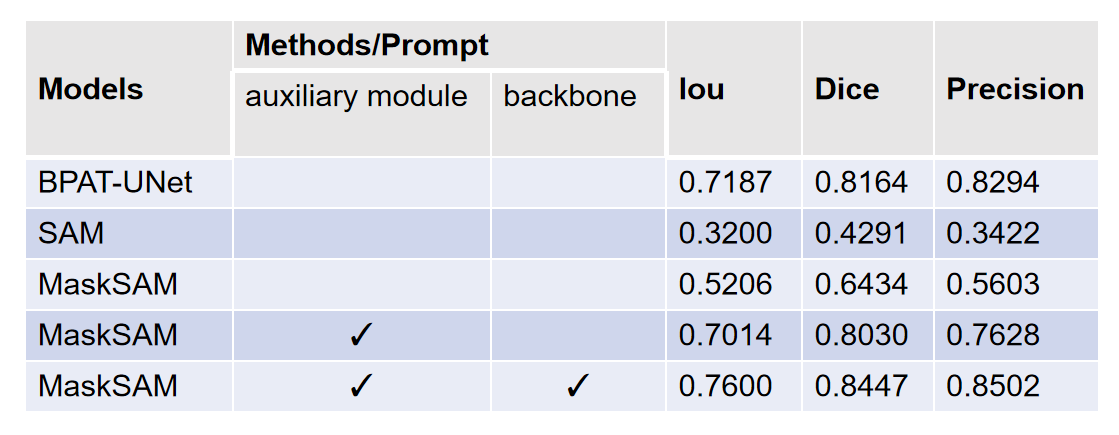
The ablation experiment of SAM with different auxiliary modules in TN3k dataset



The second part of the ablation experiment focuses on verifying the actual effect of different modules of MaskSAM. The experimental details are shown in Table 3. In this part, we make targeted adjustments and tests on each module of MaskSAM, and deeply analyze the role of each module in the overall performance of the model. MaskSAM with only auxiliary modules improved Iou, Dice, and Precision by 18.08%, 15.96%, and 20.25%, respectively, compared to no auxiliary modules and backbone. MaskSAM with auxiliary modules and backbone improved Iou, Dice, and Precision by 4.86%, 4.17%, and 8.74%, respectively, compared to only auxiliary modules. The experimental results show that different modules of MaskSAM have an important impact on the performance of the model. Through reasonable design and optimization, each module can work together to effectively improve the overall performance of the model. It is worth mentioning that the data in Table 3 also show that in the process of optimizing and adjusting the MaskSAM module, the backbone network, as the core architecture of the model, has improved its performance, which further enhances MaskSAM's ability to extract and process thyroid nodule image features, laying a solid foundation for the model to achieve better results in the thyroid nodule segmentation task.

**Table 3**

The ablation experiment of MaskSAM with different modules TN3k dataset



1. Discussion

In this paper, we innovatively proposed a thyroid nodule ultrasound image segmentation network based on the masked attention and SAM (MaskSAM). Experimental results show that MaskSAM performs well in multiple evaluation indicators (such as Recall, Accuracy, IoU, Dice, and Precision), and has significant improvements compared to the baseline method TRFE+. This finding shows that MaskSAM can effectively address challenges such as low contrast and blurred boundaries in thyroid nodule ultrasound images and improve segmentation accuracy. MaskSAM performs better than methods such as TRFE+ and BPAT-UNet in multiple evaluation indicators, showing its powerful segmentation capabilities and broad application prospects.

Our innovations mainly include three points. First, by modifying the SAM model, integrating the auxiliary module and the prompt encoder to optimize it, the segmentation performance of the model is improved. Second, unlike the traditional attention mechanism, MaskSAM focuses on the thyroid nodule area through the mask attention mechanism, reducing the interference of background noise. Third, through the collaborative work of the backbone network, pixel encoder, Transformer encoder, prompt encoder, mask decoder and auxiliary modules, accurate segmentation of thyroid nodule ultrasound images is achieved.

Although MaskSAM has made significant progress in thyroid nodule segmentation, this study was only validated on the TN3k dataset, which has limited sample size and diversity, which may affect the generalization ability of the model. Future studies need to be validated on more and larger datasets. At the same time, MaskSAM's network architecture is relatively complex, containing multiple modules and parameters, resulting in a long training time and high requirements for computing resources. Future studies can explore model simplification methods to reduce computing costs.

1. Conclusion

In this paper, a multi-task learning framework based on SAM prompts and mask-attetion is proposed for thyroid nodule segmentation from ultrasound images, which uses the thyroid region of an auxiliary module before enhancing the feature representation for thyroid nodule segmentation. To the best of our knowledge, the proposed MaskSAM is the first network that successfully fully exploits the thyroid region of another auxiliary module before improving the performance of thyroid nodule segmentation. Specifically, MaskSAM contains a backbone for feature representation learning, a cascaded pixel encoder[29], a mask-attetion based on transformer[30] encoder, a prompt[31] encoder, and a mask decoder. The thyroid nodule segmentation performance is improved by exploiting the characteristics of thyroid segmented nodules and mask-attetion. With the help of the joint efforts of these modules, MaskSAM achieves superior performance to other state-of-the-art methods in thyroid nodule segmentation.

1. References
2. Giorgio Grani, Marialuisa Sponziello, Sebastiano Filetti, Cosimo Durante.:Thyroid nodules: diagnosis and management.Nature Reviews Endocrinology:10.1038/s41574-024-01025-4 (2024)
3. Chen L-C, Zhu Y, Papandreou G, Schrof F, Adam H, Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
4. Redmon J, Divvala S, Girshick R, Farhadi A, You only look once: unifed, real-time object detection. In: CVPR (2016)
5. Redmon J, Farhadi A, Yolo9000: better, faster, stronger. In: CVPR (2017)
6. Redmon J, Farhadi A, Yolov3: An incremental improvement. arXiv (2018)
7. Bochkovskiy A, Wang C-Y, Liao H-YM, Yolov4: optimal speed and accuracy of object detection. arXiv (2020)
8. Chen Q, Wang Y, Yang T, Zhang X, Cheng J, Sun J, You only look one-level feature. arXiv (2021)
9. Ma J, Wang B. Segment anything in medical images. arXiv preprint arXiv: 2304.12306 (2023)
10. Chen, C., Miao, J., Wu, D., Yan, Z., Kim, S., Hu, J., Zhong, A., Liu, Z., Sun, L., Li, X., et al., MA-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation. arXiv preprint arXiv:2309.08842 (2023)
11. S. N. Gowda and D. A. Clifton, Cc-sam: Sam with cross-feature attention and context for ultrasound image segmentation, in Proceedings of the European Conference on Computer Vision. Springer (2024)
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 770–778 (2016)
13. D. Alexey, B. Lucas, K. Alexander, W. Dirk, An Image is Worth 16x16 Words: transformers for Image Recognition at Scale, the 9th International Conference on Learning Representations (ICLR2021) ( 2021)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
15. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. arXiv:2408.00714 (2024)
16. Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
17. Wei X, Cao J, Jin Y, et al. I-MedSAM: Implicit Medical Image Segmentation with Segment Anything. arXiv preprint arXiv:2311.17081 (2023)
18. Cheng, D., Qin, Z., Jiang, Z., Zhang, S., Lao, Q., Li, K., SAM on medical images: A comprehensive study on three prompt modes. arXiv preprint arXiv:2305.00035 (2023)
19. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al., SAM-Med2D. arXiv preprint arXiv:2308.16184 (2023)
20. A. Zhang, H. Fei, Y. Yao, W. Ji, L. Li, Z. Liu, and T.-S. Chua, Transfer visual prompt generator across llms, arXiv preprint arXiv: 2305.01278 (2023)
21. X. L. Li and P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, arXiv preprint arXiv:2101.00190 (2021)
22. Gong, H., Chen, J., Chen, G., Li, H., Li, G., Chen, F.: Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. Comput. Biol. Med. 155, 106389 (2023)
23. O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *MICCAI*, pp. 234–241 (2015)
24. H. Bi, C. Cai, J. Sun, Y. Jiang, G. Lu, H. Shu, X. Ni, BPAT-UNet: Boundary preserving assembled transformer unet for ultrasound thyroid nodule segmentation, Computer Methods and Programs in Biomedicine 238 2023. 107614 (2023)
25. Y. Wu, X. Shen, F. Bu, and J. Tian: Ultrasound image segmentation method for thyroid nodules using ASPP fusion features: IEEE Access, vol. 8, pp. 172457–172466 (2020)
26. B. Cheng, I. Misra, A.G. Schwing, A. Kirillov, R. Girdhar, Masked-attention mask transformer for universal image segmentation, arXiv Preprint, arXiv: 2112.01527 (2021)
27. Gong, H.; Chen, G.; Wang, R.; Xie, X.; Mao, M.; Yu, Y.; Chen, F.; and Li, G. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 257–261. IEEE. (2021)
28. S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, X. Chen, CPFNet: Context pyramid fusion network for medical image segmentation, IEEE Trans. Med. Imaging 39 (10) 3008–3018 (2020)
29. Long J, Shelhamer E, Darrell T, Fully convolutional networks for semantic segmentation. In: CVPR (2015)
30. Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaiane and Martin Jagersand University of Alberta, Canada: U2 -Net: Going Deeper with Nested U-Structure for Salient Object Detection: arXiv:2005.09007v3 [cs.CV] (2022)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. Attention is all you need. Advances in neural information processing systems, 30. 31st Conference on Neural Information Processing Systems. CA, USA: Long Beach (2017)
32. M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan. : Maple: Multi-modal prompt learning, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19 113– 19 122 (2023)