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 the SAM model. The network organically integrates a pixel cascade module, a mask attention-based transform encoder module, a prompt encoder, and a mask decoder module. To further promote the development of thyroid nodule ultrasound image segmentation, 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. 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.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 (as shown in Fig. 1). This noise interference not only reduces the clarity of the image, but also increases the difficulty for doctors to accurately identify and analyze nodules.

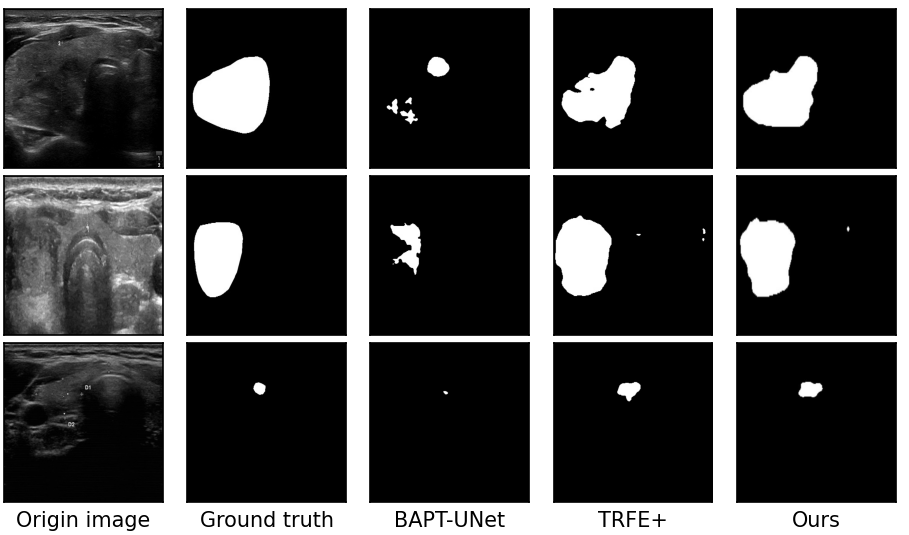
With the rapid development of science and technology, medical imaging technology has made continuous breakthroughs. High-resolution ultrasound equipment has been widely used in many medical fields such as thyroid disease diagnosis and disease monitoring due to its excellent imaging capabilities. The 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.

In this magnificent wave of technological development, the Segment Anything Model (SAM)[1] released by Meta in 2023 has undoubtedly become the focus of everyone's attention, like the most dazzling star in the night sky. SAM's design and training concept are ingenious and flexible. It seems to have a pair of magical wings, which can easily achieve zero-sample conversion to new image distributions and tasks, breaking the limitations of traditional models. 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. Applying SAM [2]to the field of medical image segmentation[3, 4, 5, 6], especially the segmentation of thyroid nodule ultrasound images, has brought new hope and dawn to this field, just like lighting a bright lamp in the dark, and has become a very promising research direction.

However, reality is often not always smooth sailing, and there is often a certain gap between ideal and reality. A large number of recent studies and the results of personal experiments by many researchers have clearly shown that due to the lack of in-depth understanding and mastery of medical expertise, SAM's performance in medical image segmentation has not achieved the expected ideal effect and seems to be somewhat unsatisfactory. Medical images have their own unique complexity and high professionalism. The ultrasound image of thyroid nodules is like a treasure chest full of countless secrets, containing a large amount of medical features and detailed information. For the original SAM model, these are like a difficult peak to climb and difficult to grasp and understand accurately. For example, in the process of judging the benign and malignant nature of thyroid nodules, doctors often need to rely on in-depth analysis of detailed features such as the smoothness of the nodule boundary and the uniformity of the internal echo. However, SAM is obviously insufficient in processing these key information, and it is difficult to accurately identify and judge, which to a certain extent limits its application and promotion in the field of medical image segmentation. Therefore, how to take effective measures to enhance SAM's ability to segment medical images has become an important issue that needs to be solved urgently.

In SAM, the prompt[7] mechanism, as a unique way for users to interact with the model and guide it to segment specific targets, is like a magic key that opens the door to optimizing model performance for us. By providing different types of prompt[8] information, such as points, boxes, masks, etc., the user can mark key locations for the model on a map. The model can then focus on specific areas or objects in the image based on these prompts, thereby obtaining more accurate and tailored segmentation results. This unique feature provides us with a valuable reference method and innovative ideas: by freezing the model parameters with relatively low performance as prompt inputs, it is expected to inject the essence of medical knowledge into SAM, make up for its lack of medical knowledge, and 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. In the future, there is still a broad space for exploration and unlimited possibilities for the optimization and improvement of SAM in the field of medical image segmentation. Researchers can further study the optimization strategy of the prompt mechanism, explore how to more effectively integrate medical knowledge and model parameters, and develop more intelligent and accurate medical image segmentation algorithms to make greater contributions to the development of medical imaging diagnosis technology.

In this study, we used an innovative deep learning model architecture to accurately extract thyroid nodule features. First, we selected UNet[9] as the backbone network and used it to preliminarily process the input image to extract the main mask of the thyroid nodule. This mask, as key information, provides the basis for subsequent processing. Then, the extracted main mask is input into the pyramid 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[10] 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. The modules work together, and its structure is shown in Fig. 2.

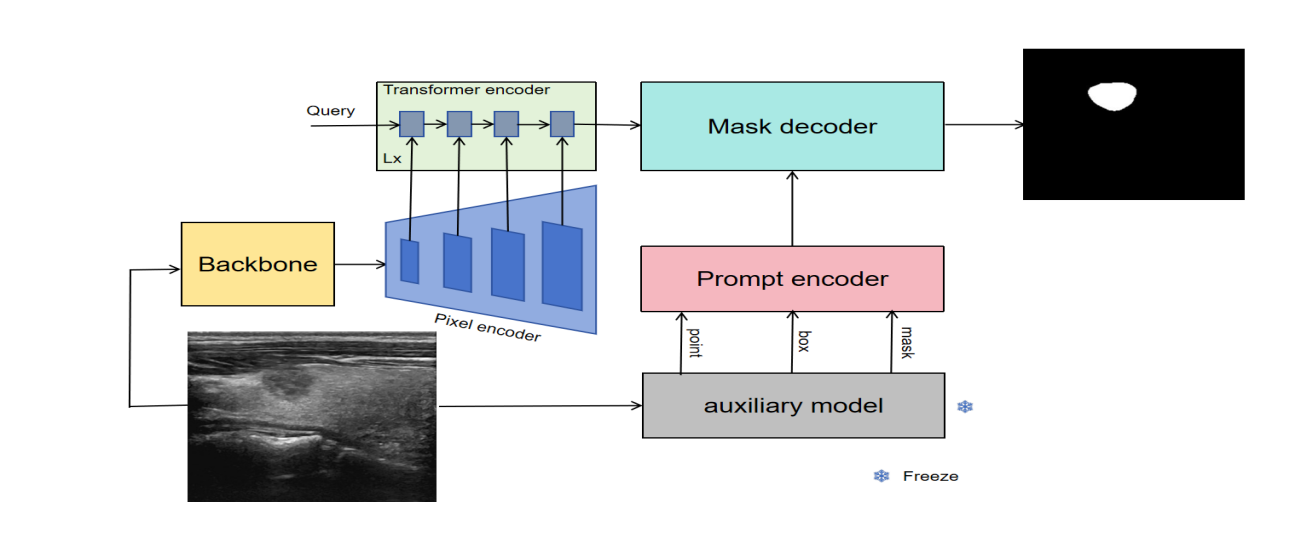


**Fig. 1.** Some visualization examples of ours MASK-SAM comparing to BPAT-UNet and TRFEP+.

1. Method

2.1. 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 pre-set 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.



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

At the same time, researchers are actively expanding the application scope of deep learning models, and DeepLabV3+[11], YOLO[12, 13, 14, 15, 16] 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.[17] 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.[18] 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[19] suggested adding a parallel network resnet50[20] to fuse the ViT[21] 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.

2.2. Overall structure design

The overall model architecture is shown in Fig. 2. This architecture is carefully designed based on in-depth research on the characteristics of thyroid nodules and the application of existing deep learning models. Given that thyroid nodules have diverse shapes and sizes, this poses a great challenge to the segmentation task. In order to effectively improve the accuracy of the segmentation task, we have conducted in-depth analysis and believe that 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 the model design, we adopted a multi-layer pyramid structure[22]. 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.

The Transformer encoder layer is one of the core components of the entire model architecture, which is built based on the mask attention[23] 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.

In the model, we retain the hint 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 hint encoder can effectively encode various input hint information, and the mask decoder can generate accurate segmentation masks based on the encoded information. At the same time, we choose BPAT-UNet as an auxiliary model. BPAT-UNet has a unique network structure and feature extraction capabilities, and also has certain advantages in the task of thyroid nodule segmentation. During use, we freeze all parameters of BPAT-UNet. The purpose of this 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 prompts, box prompts, and point prompts. These different forms of prompt information contain rich information such as nodule location and shape. They are input into the prompt encoder layer, and after encoding by the prompt encoder, they are input into the mask decoder layer. Through this process, accurate final prediction results are finally obtained. At the same time, the SAM model is successfully converted from a segmentation model that originally required manual interaction to a fully automatic segmentation model that can automatically complete segmentation tasks, greatly improving segmentation efficiency and practicality.

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.

1. Experiments

3.1. 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.

3.2. 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);

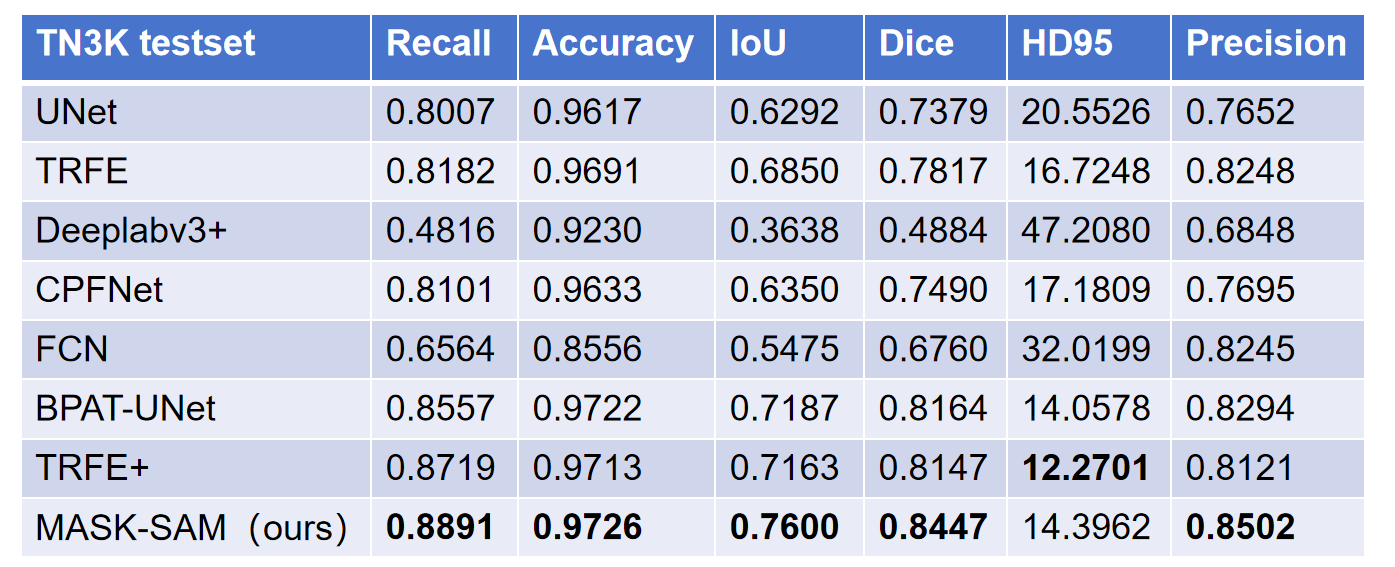
where TP, FP, TN, FN indicate true positive, false positive, true negative, and false negative, respectively. The F1-score is the harmonic mean of the precision and recall. Any p-value less than 0.05 demonstrates that the proposed method performs significantly better than the other compared methods.

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 MASK-SAM model in the thyroid nodule segmentation task, we conducted in-depth comparative experiments with current mainstream deep learning models such as UNet, TRFE[24], DeeplabV3+, CPFNet[25], FCN[26], BPAT-UNet, and TRFE+[27] on the TN3K dataset. In the comparison of various evaluation indicators, MASK-SAM showed excellent performance. In terms of recall, precision, iou, and dice, MASK-SAM can accurately identify most of the real thyroid nodule areas, and the recall rate has reached a very high level. 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, MASK-SAM is slightly lower than TRFE+. The reason why MASK-SAM can perform well in multiple key indicators such as recall, iou, and dice is due to its unique prompt and shielding 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 shielded attention mechanism is like equipping the model with an "intelligent filter" that can effectively eliminate background noise and other irrelevant information interference, 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, MASK-SAM can handle it relatively well, thereby achieving excellent results in multiple evaluation indicators.

**Table 1**

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



3.4. Ablation Study

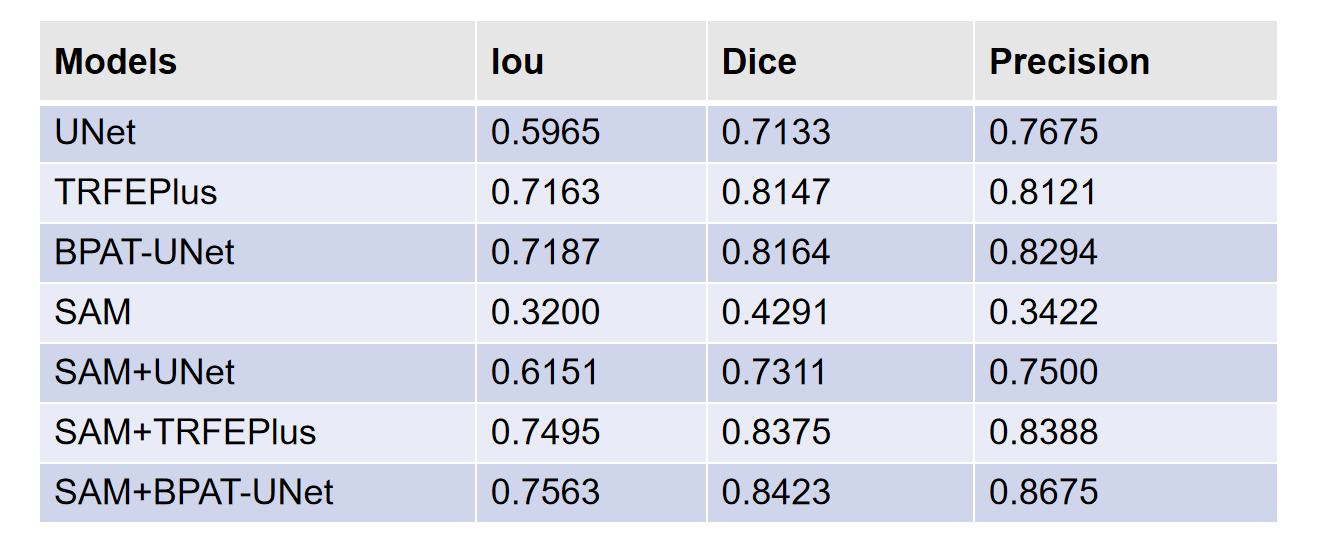
In order to deeply explore the role of each component of the model and the effect of different combinations, we carefully designed and conducted ablation experiments. The entire ablation experiment is divided into two key parts, and both use the TN3k dataset as experimental data support. The TN3k dataset covers a wealth of thyroid nodule ultrasound images and corresponding nodule mask labels, providing a solid guarantee for the accuracy and reliability of the experimental results.

The first part of the ablation experiment focuses on studying the effect and feasibility of combining SAM with different auxiliary models. The specific content is shown in Table 2. In this part of the experiment, we combined different auxiliary models with SAM and converted them into point prompts, box prompts, and mask prompts. While keeping other experimental conditions completely consistent, the results of key evaluation indicators such as iou (intersection over union), dice coefficient, and accuracy are carefully compared. iou is used to measure the degree of overlap between the predicted result and the true label, the dice coefficient can evaluate the similarity between two samples, and the accuracy reflects the accuracy of the model prediction. Through comparative analysis of these indicators, we aim to verify the performance changes of SAM after adding different auxiliary models. From the data results presented in Table 2, compared with the performance of using auxiliary models alone or SAM itself, the effect of SAM combined with different auxiliary models has been significantly improved. This result not only intuitively shows the gain effect of different auxiliary models on SAM, but also fully verifies the feasibility of combining SAM with various auxiliary models, providing a strong practical basis for model optimization and improvement.

The second part of the ablation experiment focuses on verifying the actual effect of different modules of MASK-SAM. The experimental details are shown in Table 3. In this part, we make targeted adjustments and tests on each module of MASK-SAM, and deeply analyze the role of each module in the overall performance of the model. The experimental results show that different modules of MASK-SAM 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 MASK-SAM module, the backbone network, as the core architecture of the model, has improved its performance, which further enhances MASK-SAM'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.

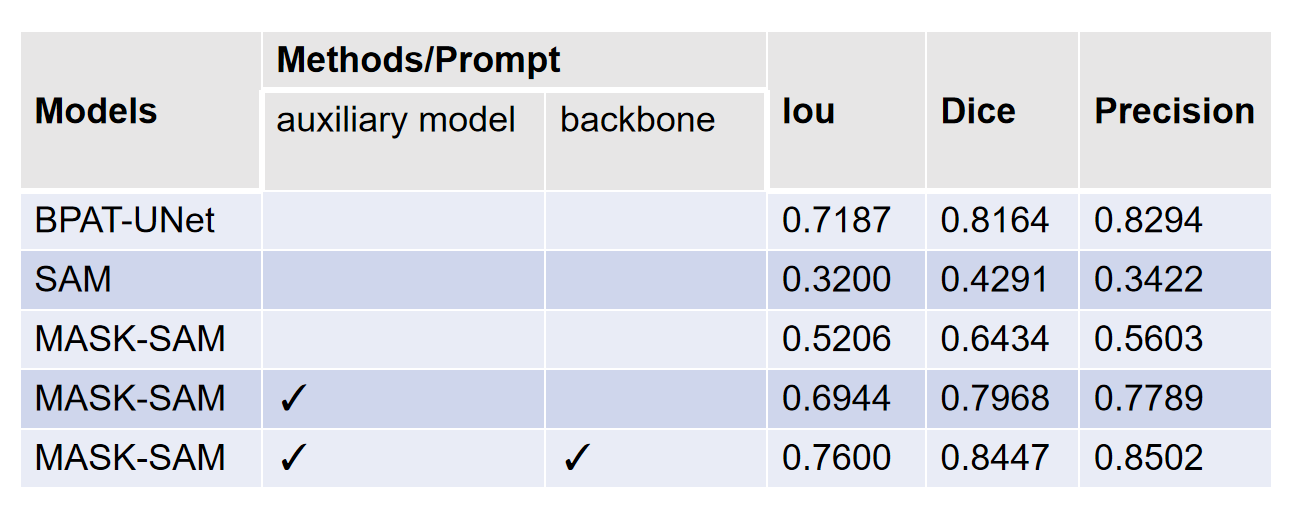
**Table 2**

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



**Table 3**

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



3.5. Dataset

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. 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 model before enhancing the feature representation for thyroid nodule segmentation. To the best of our knowledge, the proposed MASK-SAM is the first network that successfully fully exploits the thyroid region of another auxiliary model before improving the performance of thyroid nodule segmentation. Specifically, MASK-SAM contains a backbone for feature representation learning, a cascaded pixel encoder[28], a mask-attetion based on transformer[29] encoder, a prompt[30] 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, MASK-SAM achieves superior performance to other state-of-the-art methods in thyroid nodule segmentation to promote the future development of thyroid nodule segmentation.

1. References
2. 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)
3. 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)
4. 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)
5. Wei X, Cao J, Jin Y, et al. I-MedSAM: Implicit Medical Image Segmentation with Segment Anything. arXiv preprint arXiv:2311.17081 (2023)
6. 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)
7. 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)
8. 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)
9. X. L. Li and P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, arXiv preprint arXiv:2101.00190 (2021)
10. O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *MICCAI*, pp. 234–241 (2015)
11. 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)
12. Chen L-C, Zhu Y, Papandreou G, Schrof F, Adam H, Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
13. Redmon J, Divvala S, Girshick R, Farhadi A, You only look once: unifed, real-time object detection. In: CVPR (2016)
14. Redmon J, Farhadi A, Yolo9000: better, faster, stronger. In: CVPR (2017)
15. Redmon J, Farhadi A, Yolov3: An incremental improvement. arXiv (2018)
16. Bochkovskiy A, Wang C-Y, Liao H-YM, Yolov4: optimal speed and accuracy of object detection. arXiv (2020)
17. Chen Q, Wang Y, Yang T, Zhang X, Cheng J, Sun J, You only look one-level feature. arXiv (2021)
18. Ma J, Wang B. Segment anything in medical images. arXiv preprint arXiv: 2304.12306 (2023)
19. 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)
20. 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)
21. 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)
22. 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)
23. 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)
24. 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)
25. 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)
26. 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)
27. Long J, Shelhamer E, Darrell T, Fully convolutional networks for semantic segmentation. In: CVPR (2015)
28. 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)
29. 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)
30. 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)
31. 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)