

M2ANet: Multi-branch and multi-scale attention network for medical image segmentation

Wei Xue(薛伟), Chuanghui Chen(陈创慧), Xuan Qi(戚轩), Jian Qin(秦健), Zhen Tang(唐振), and Yongsheng He(何永胜)

Citation: Chin. Phys. B, 2025, 34 (**8**): 080703. DOI: 10.1088/1674-1056/adce96
http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Deep learning-assisted common temperature measurement based on visible light imaging

Jia-Yi Zhu(朱佳仪), Zhi-Min He(何志民), Cheng Huang(黄成), Jun Zeng(曾峻), Hui-Chuan Lin(林惠川), Fu-Chang Chen(陈福昌), Chao-Qun Yu(余超群), Yan Li(李燕), Yong-Tao Zhang(张永涛), Huan-Ting Chen(陈焕庭), and Ji-Xiong Pu(蒲继雄)

Chin. Phys. B, 2024, 33 (**8**): 080701. DOI: 10.1088/1674-1056/ad4cd9

FPGA and computer-vision-based atom tracking technology for scanning probe microscopy

Feng-Du Yu(俞风度), Li Liu(刘利), Su-Ke Wang(王肃珂), Xin-Biao Zhang(张新彪), Le Lei(雷乐), Yuan-Zhi Huang(黄远志), Rui-Song Ma(马瑞松), and Qing Huan(郇庆)

Chin. Phys. B, 2024, 33 (**5**): 050705. DOI: 10.1088/1674-1056/ad34cb

Image segmentation of exfoliated two-dimensional materials by generative adversarial network-based data augmentation

Xiaoyu Cheng(程晓昱), Chenxue Xie(解晨雪), Yulun Liu(刘宇伦), Ruixue Bai(白瑞雪), Nanhai Xiao(肖南海), Yanbo Ren(任琰博), Xilin Zhang(张喜林), Hui Ma(马惠), and Chongyun Jiang(蒋崇云)

Chin. Phys. B, 2024, 33 (**3**): 030703. DOI: 10.1088/1674-1056/ad23d8

Learnable three-dimensional Gabor convolutional network with global affinity attention for hyperspectral image classification

Hai-Zhu Pan(潘海珠), Mo-Qi Liu(刘沫岐), Hai-Miao Ge(葛海淼), and Qi Yuan(袁琪)

Chin. Phys. B, 2022, 31 (**12**): 120701. DOI: 10.1088/1674-1056/ac8cd7

An artificial synapse by superlattice-like phase-change material for low-power brain-inspired computing

Qing Hu(胡庆), Boyi Dong(董博义), Lun Wang(王伦), Enming Huang(黄恩铭), Hao Tong(童浩), Yuhui He(何毓辉), Ming Xu(徐明), Xiangshui Miao(缪向水)

Chin. Phys. B, 2020, 29 (**7**): 070701. DOI: 10.1088/1674-1056/ab892a

---

M2ANet: Multi-branch and multi-scale attention network for medical image segmentation

Wei Xue(薛伟)[1,†], Chuanghui Chen(陈创慧)[1], Xuan Qi(戚轩)[2,3], Jian Qin(秦健)[1], Zhen Tang(唐振)[1], and Yongsheng He(何永胜)[2,3]

[1] *School of Computer Science and Technology, Anhui University of Technology, Maanshan* 243032, *China*
[2] *Department of Radiology, Maanshan People's Hospital, Maanshan* 243099, *China*
[3] *Maanshan Key Laboratory for Medical Image Modeling and Intelligent Analysis, Maanshan* 243099, *China*

(Received 22 January 2025; revised manuscript received 19 March 2025; accepted manuscript online 21 April 2025)

Convolutional neural networks (CNNs)-based medical image segmentation technologies have been widely used in medical image segmentation because of their strong representation and generalization abilities. However, due to the inability to effectively capture global information from images, CNNs can easily lead to loss of contours and textures in segmentation results. Notice that the transformer model can effectively capture the properties of long-range dependencies in the image, and furthermore, combining the CNN and the transformer can effectively extract local details and global contextual features of the image. Motivated by this, we propose a multi-branch and multi-scale attention network (M2ANet) for medical image segmentation, whose architecture consists of three components. Specifically, in the first component, we construct an adaptive multi-branch patch module for parallel extraction of image features to reduce information loss caused by downsampling. In the second component, we apply residual block to the well-known convolutional block attention module to enhance the network's ability to recognize important features of images and alleviate the phenomenon of gradient vanishing. In the third component, we design a multi-scale feature fusion module, in which we adopt adaptive average pooling and position encoding to enhance contextual features, and then multi-head attention is introduced to further enrich feature representation. Finally, we validate the effectiveness and feasibility of the proposed M2ANet method through comparative experiments on four benchmark medical image segmentation datasets, particularly in the context of preserving contours and textures. The source code of M2ANet will be released at https://github.com/AHUT-MILAGroup/M2ANet.

Keywords: medical image segmentation, convolutional neural network, multi-branch attention, multi-scale feature fusion

PACS: 07.05.Pj, 42.30.Tz, 87.57.nm

DOI: 10.1088/1674-1056/adce96 **CSTR:** 32038.14.CPB.adce96

1. Introduction

Medical image segmentation plays a crucial role in the process of medical image analysis, which is defined as the pixel-wise classification of medical images, aiming to separate structures such as lesions, tumors, and organs from the background.[1] Traditional medical image segmentation focuses on edge detection-based segmentation,[2] threshold-based segmentation,[3] and region-based segmentation.[4] However, these methods rely on manually designed rules and feature engineering, and their performance is limited when facing complex segmentation tasks.[5] Deep neural network-based segmentation models have strong characterization and generalization capabilities by automatically learning features from a large amount of data, which have been used successfully in skin disease segmentation,[6] breast cancer segmentation,[7] lung CT image segmentation,[8] polyp image segmentation,[9] cell image segmentation,[10] etc. Among them, U-Net[11] is an earlier deep neural network for medical image segmentation, which is capable of classifying each pixel of an image and generating segmentation results directly. Due to its superior segmentation performance, several variants of U-Net have subsequently emerged. For example, UNet++[12] integrated the denseNet structure to enhance the model's ability to capture complex features. UNet 3+[13] introduced deep supervision and multi-scale feature fusion on the basis of UNet++, which better handled fine-grained features in the image and improved the segmentation accuracy. In addition, MultiResUNet[14] introduced multi-resolution blocks to handle information at different scales to enhance the model's ability to capture features of different sizes.

However, the segmentation networks using only CNN architecture are weak in terms of capturing global information. In Ref. [15], Sinha *et al.* used multi-scale self-guided attention to capture local and global features and alleviated the above-mentioned limitations of CNN by combining two types of features. Dosovitskiy *et al.*[16] abandoned the traditional CNN architecture and fully adopted transformer to process image data, introducing the vision transformer (ViT). Liu *et al.*[17] further improved ViT and proposed swin transformer. The swin transformer utilizes a moving window attention mech-

[†]Corresponding author. E-mail: xuewei@ahut.edu.cn

© 2025 Chinese Physical Society and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

anism, which restricts attention to non-overlapping local windows but also allows for cross-window interactions. Based on the swin transformer, UCTransNet[18] introduces channel transformer between encoder and decoder to guide the multi-scale channel information to be efficiently connected to the decoder. Unlike UCTransNet, Swin-Unet[19] introduced transformer blocks at each layer to enhance the ability to capture image dependencies over long distances. Although transformer is effective at capturing both global and local information, it comes with high computational costs. In Ref. [20], Ibtehaz *et al*. proposed ACC-UNet, which achieves performance comparable to transformers. But it has shortcoming in obtaining clear texture information, resulting in inaccurate segmentation boundaries.

Besides, the effective utilization of multi-scale information can provide more texture details to the segmentation model and improve its accuracy when dealing with pixel-dense or small-scale regions[21]. In Ref. [22], Poudel *et al*. proposed a encoding-decoding structure that aggregates multi-scale semantic information to generate global features. However, different baseline models require appropriate modules to maximize their respective advantages due to differences in their inherent architecture and characteristics. In addition, UNeXt[23] enhances the feature representation capability and improves the model segmentation performance by customizing the multi-scale encoder and design module to fuse information from different scales. However, the encoder in UNeXt only performs simple convolution and downsampling, which does not fully consider the global information of the image. In Ref. [24], Wang *et al*. designed the MCCT module for multi-scale feature extraction and attention-map fusion of the input image. They used super-pixel segmentation to reduce redundant information and avoid the interference of meaningless information, which improves the segmentation accuracy and efficiency of the model for medical images. Based on the above analysis, we conclude that CNN can effectively extract local features and is adept at capturing texture details and edge information in the image. On the other hand, transformer can effectively simulate the global dependencies of images through its self-attention mechanism to make the model to understand more complex contextual information. The fusion of these two frameworks can take advantage of the respective strengths of the two models. With this as motivation, we propose a new method for medical image segmentation, which aims at extracting both local and global features effectively and obtaining segmentation results with clear contours. The points below highlight the contributions of this paper.

(i) We propose an adaptive multi-branch patch module to recover the information lost during downsampling and convolution by extracting image features in parallel and adjusting feature weights adaptively.

(ii) We improve the convolutional block attention module by using residual blocks to enhance the ability to identify important features in images and alleviate the phenomenon of gradient vanishing.

(iii) We propose a multi-scale feature fusion module to enrich the representation of image features by using different scales of detail information.

(iv) Experiments on benchmark datasets demonstrate that M2ANet outperforms state-of-the-art segmentation models, and our results show notable improvements in preserving fine textures and clear contours.

2. Related work

2.1. Multi-scale information

In deep learning, multi-scale typically refers to the simultaneous consideration and utilization of scale information from different input or feature layers during model training or inference. Multi-scale information can effectively obtain information at different levels, and it is more advantageous than single scale in small and dense areas. Previous studies, including image classification,[25–27] image detection,[28,29] and image segmentation,[30–32] have demonstrated the effectiveness of using multi-scale information. For example, DCSAU-Net[30] combined multi-scale feature information with multiple receptive fields, improving the network's ability to capture information at the low-level semantic layer. While Rahman *et al*.[31] proposed the CASCADE attentional encoder, which combines multi-scale feature fusion with self-attention mechanisms to improve the model's ability to localize organs and lesions. To further enhance the computational efficiency of the model, they added to their original proposal to use multi-scale deep convolution in the decoding stage to fuse multi-level features in medical images and improve the computational efficiency of the model, resulting in the proposed EMCAD.[32] Zhao *et al*.[33] designed a subtraction unit to replace additive or cascaded features to reduce redundancy in multi-scale information, improving the efficiency of the model. In this paper, we use SwinB (a variant of swin transformer) as encoder. With the hierarchical structure of SwinB, our model can flexibly capture multi-scale information from images. Besides, we propose a multi-scale feature fusion module that help integrate information from different scales and enrich the representation of features.

2.2. Attention mechanism

Attention mechanism has been widely used in computer vision, which originates from the study of human vision, and its core idea is to selectively focus on a portion of input information while ignoring other information to make rational use of limited computing resources.[34] In recent years, many studies have introduced attention mechanisms into medical image

segmentation. For example, the attention U-Net[35] adds the attention gate module to the original U-Net architecture, allowing the network to focus on key areas of the image and extract important information. Focus U-Net[36] further enhanced the focusing ability of the model on the basis of attention U-Net, reduced the interference of non-core regions through focus block, and strengthened the feature extraction of the core region, which makes the model further improve both the computational efficiency and segmentation accuracy. Ruan *et al.*[37] used the GHPA attention module to extract pathology information from different angles on different axes and applied GAB attention module to group and fuse features. In order to apply the model to datasets of different modalities, Nam *et al.*[38] proposed MADGNet to use multi-frequency information in medical images with different modalities to enhance the model's ability to extract spatial and boundary features. In this paper, we propose an adaptive multi-branch patch module to recover information lost due to downsampling and convolution, and adaptively adjust the weights of features to obtain important features. We also propose a convolutional block attention module with residual block to compress redundant information in channels and spaces, further enhancing the model's ability to perceive features.

3. Proposed method

The overall structure of the proposed method M2ANet is shown in Fig. 1. M2ANet adopts an encoding-decoding structure, using swinB as encoder and three main modules as decoder, to capture global contextual features and local detail features. In decoder, we first propose an adaptive multi-branch patch module (AMBPM) to recover the information lost during downsampling and convolution, and use a multi-branch strategy to extract image features in parallel and dynamically adjust their weights, so that the model focuses on key regions of the image. We then propose a convolutional block attention module with residual block (CBAMRB) to compress the space and channels of images, reducing redundant information in the extracted features and effectively alleviating the problem of gradient vanishing. Further, we propose a multi-scale feature fusion module (MSFFM) to fuse multi-scale features and further enrich the expression of features.

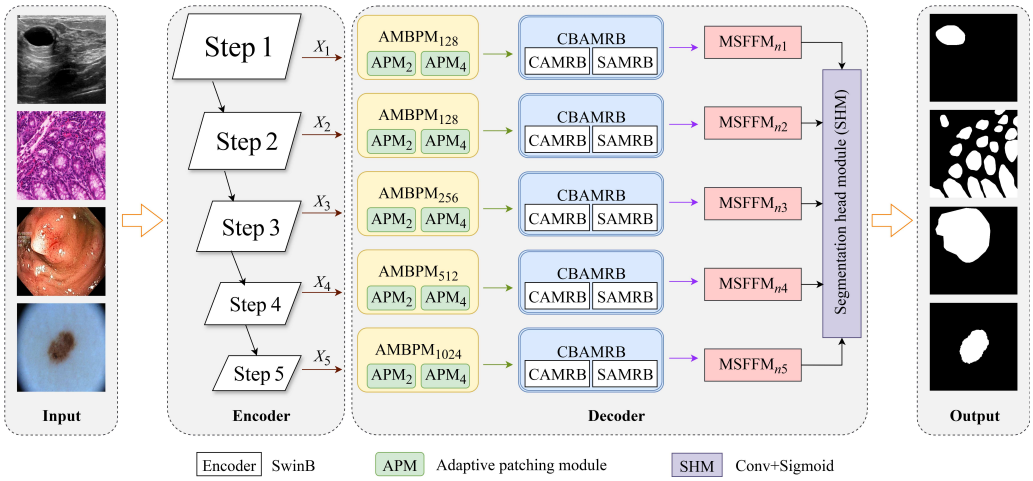


Fig. 1. Overall structure of the proposed segmentation method M2ANet, which mainly consists of encoder and decoder. Encoder employs the hierarchical structure to extract multi-scale information, and decoder is used to generate the final segmentation result according to the advanced features extracted by the encoder.

Let I denote the layer number of decoder and x_i represent the input, then the above process can be expressed mathematically as follows:

$$s_i = \mathrm{MSFFM}(\mathrm{CBAMRB}(\mathrm{AMBPM}(x_i))), \tag{1}$$

where $i \in [1, I]$ denotes the current decoder sequence number, and s_i denotes the intermediate output of the first three modules. Finally, we use a segmentation head module (SHM) to separate the region of interest from the background, and the final output of M2ANet can be formulated as $\mathrm{SHM}(\sum_{i=1}^{I}(s_i))$.

3.1. Adaptive multi-branch patch module (AMBPM)

As shown in Fig. 2, AMBPM, combining the advantages of multi-branch and adaptive learning strategies, is proposed to recover the information lost during downsampling and convolution. We extract image features in parallel and adjust the weights of features dynamically through a learnable parameter, allowing the model to adaptively select the most important features based on different input images and enhancing the flexibility of feature selection.

Specifically, given an input x_i, we first divide it into $\rho \times \rho$ patches along the height (H) and width (W) dimensions. Then we compress the channel along the last dimension using mean operation and two MLP (multi-layer perceptron) operations to enrich the expression of features. Let P represent the patch operation and l_i denote the output of the above operation, the above process can be represented by the following mathemat-

ical formula:

$$l_{i;\rho} = \mathrm{MLP}_2(\mathrm{MLP}_1(\mathrm{Mean}(P_\rho(x_i)))), \tag{2}$$

where Mean denotes the mean value operation. In addition, to adapt the extracted features to various datasets, we introduce a learnable parameter $\hat{p}$ to adjust the weights of the features. This operation is formulated as

$$\mathrm{APM}_\rho = \mathrm{Conv}(l_{i;\rho} \cdot \mathrm{mask} \cdot \hat{p}), \tag{3}$$

where mask is the weight map obtained from $l_{i;\rho}$, Conv denotes the convolution operation, and APM_ρ is the output of the above operation. Meanwhile, for x_i, we use a parallel extraction strategy to extract different image features to improve the flexibility in feature extraction. Specifically, let Sig represent sigmoid function and m_i be the output of AMBPM, then m_i can be expressed as

$$\begin{aligned} m_i = \mathrm{Relu}(\mathrm{BN}(\mathrm{Drop}(&\mathrm{APM}_{\rho=2}(x_i) + \mathrm{Conv}(\mathrm{Conv}(x_i)) \\ &+ \mathrm{APM}_{\rho=4}(x_i)))). \end{aligned} \tag{4}$$

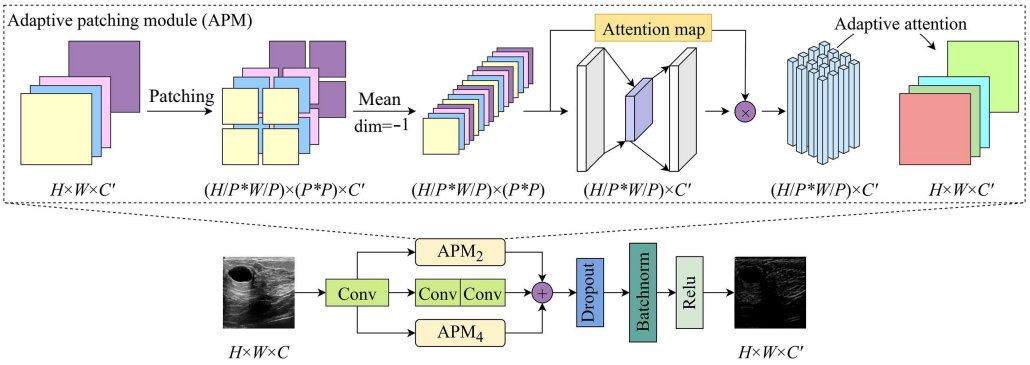


Fig. 2. Illustration of AMBPM. For an input, AMBPM uses a parallel extraction strategy to recover features lost by downsampling and convolution. APM is used to adjust the weights of different branches.

3.2. Convolutional block attention module with residual block (CBAMRB)

We propose the CBAMRB based on Ref. [39]. As shown in Fig. 3, CBAMRB consists of two parts, channel attention module with residual block (CAMRB), and spatial attention module with residual block (SAMRB), which can simultaneously compress both channel and spatial dimensions to improve the model's ability to identify important features and alleviate the problem of gradient vanishing. Specifically, CAMRB mainly strengthens the network's attention to important channels, while SAMRB strengthens the network's attention to important regions.

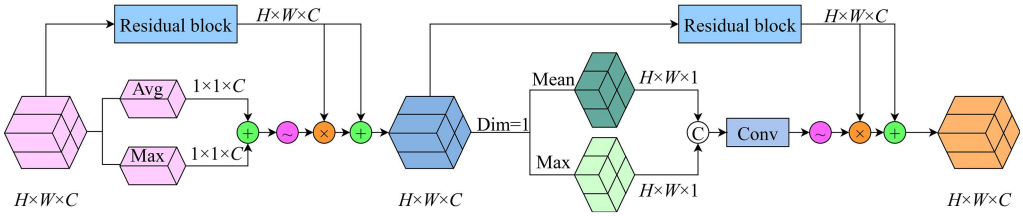


Fig. 3. Illustration of CBAMRB. The left branch is CAMRB, and the right is SAMRB.

The workflow of the whole module is that the input feature maps are first processed by CAMRB to generate channel weights. Then, the channel-weighted feature maps are processed by SAMRB to generate spatial weights. After the weighting of the two attention mechanisms, the output feature map can pay more attention to the important features and better express the information related to the task. Let z_i denote the output of CBAMRB, then the above process is expressed mathematically as follows:

$$\mathrm{CAMRB}_i = \mathrm{Sig}(\mathrm{Avg}(m_i) + \mathrm{Max}(m_i)) \times m_i + m_i, \tag{5}$$

$$\begin{aligned} \mathrm{SAMRB}_i = \mathrm{Sig}(&\mathrm{Conv}(\mathrm{Cat}(\mathrm{Mean}(m_i), \mathrm{Max}(m_i)))) \\ &\times m_i + m_i, \end{aligned} \tag{6}$$

$$z_i = \mathrm{SAMRB}_i(\mathrm{CAMRB}_i(m_i)), \tag{7}$$

where Avg and Max are pooling operations, CAMRB_i and SAMRB_i are the outputs of two components of CBAMRB, respectively.

3.3. Multi-scale feature fusion module (MSFFM)

Inspired by Ref. [40], we propose MSFFM to enrich the representation of image features by using detail information at different scales. This module can integrate feature maps from multi-scales, enabling the network to capture fine-grained local details and broader contextual information.

As shown in Fig. 4, for the input feature map z_i, we use $j = 2$, 4, and 8 for pooling operations to obtain feature maps of different scales. Then we add corresponding positional codes to different feature maps and concatenate them into a large fea-

ture map with different scale information. At the same time, we perform a rearrangement operation on the original feature map z_i, adding positional encoding, and sending it together with the above feature map to the MHA (multi-head attention) to obtain att_i. Besides, we introduce two MLP and Norm (normalization) operations to enhance the model's expressive power and ability to handle complex features. The above process can be expressed mathematically as follows:

$$\text{att}_i = \text{MHA}\left(z_i + \text{pos}, \text{Cat}\left(\sum_{j=1}^{3} \text{Pool}_j(z_i)\right) + \text{pos}\right), \quad (8)$$

$$s_i = \text{Norm}_2(MLP_2(\text{Norm}_1(\text{MLP}_1(\text{att}_i)))), \quad (9)$$

where pos denotes positional encoding, Pool is pooling operation, Cat denotes catenation operation, and att_i represents the output after the MHA operation.

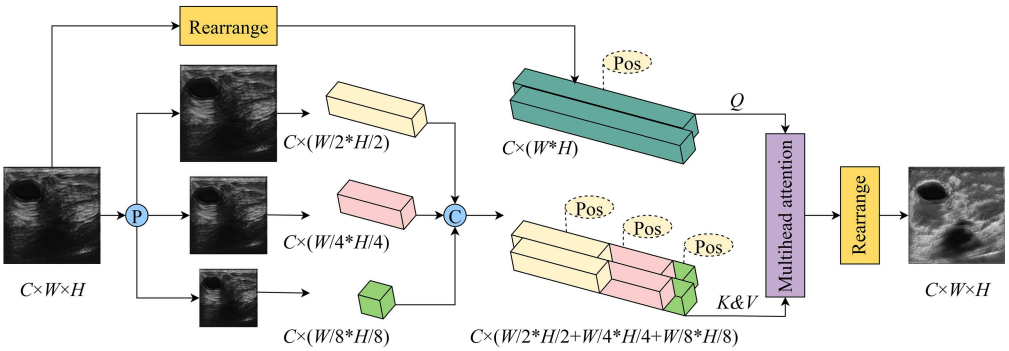


Fig. 4. Illustration of MSFFM. It integrates multi-scale information to reduce the computational complexity associated with a single view and enhance the model's robustness.

3.4. Loss function

We use weighted binary cross-entropy (BCE) loss and Dice loss as the components of objective function. BCE loss measures the probability between the model output and the true label and its weighting is to mitigate the imbalance between background and foreground.

Let w_{pos} and w_{neg} be the category weights, pos_i and neg_i denote the pixel points of each category, N_{pos} and N_{neg} denote the total number of pixel points of the categories, the formula of BCE is

$$L_{\text{BCE}} = \sum \left(\frac{w_{\text{pos}} \cdot \text{pos}_i \cdot l}{N_{\text{pos}}} + \frac{w_{\text{neg}} \cdot \text{neg}_i \cdot l}{N_{\text{neg}}} \right). \quad (10)$$

In Eq. (10), $l = -t_i \cdot \log(p_i) - (1-t_i) \cdot \log(1-p_i)$, where t_i and p_i represent the true value and predicted value, respectively.

Dice loss evaluates the similarity between the predicted label and the true label. We use weighted Dice loss to alleviate the problem of class imbalance, and it is in the form of

$$L_{\text{Dice}} = \frac{1}{M} \sum_{j=1}^{M} \left(1 - \frac{\sum_{i=1}^{N} \hat{p}_i \cdot \hat{t}_i}{\sum_{i=1}^{N} (\hat{p}_i)^2 + \sum_{i=1}^{N} (\hat{t}_i)^2} \right), \quad (11)$$

where $\hat{t}_i$ and $\hat{p}_i$ denote the weighted true and predicted values. M represents the number of batches, and N is the number of pixels.

Finally, the total loss of M2ANet is formulated as

$$L_{\text{Total}} = \alpha \cdot L_{\text{BCE}} + \beta \cdot L_{\text{Dice}}, \quad (12)$$

where α and β are parameters that balance different losses.

4. Experimental setup

4.1. Benchmark datasets

(i) Breast ultrasound images dataset (BUSI)[41] [‡] In BUSI, the benign data contains 891 images, while the malignant data includes 421 images. In our experiment, we divide the benign and malignant data into training set, validation set, and test set in a ratio of 6:2:2, respectively.

(ii) International skin imaging collaboration challenge dataset 2018 (ISIC2018)[42] [§] ISIC2018 contains 2594 images. In the following experiments, we select the first 600 images and divide the whole image set into training, validation and test sets according to the ratio of 6:2:2.

(iii) Gland segmentation dataset (GlaS)[43] [¶] In GlaS, the training set includes 37 benign images and 48 malignant images, and the test set consists of 37 benign images and 43 malignant images. We divide the original training set into a new training set and validation set in a ratio of 8:2, while the test set remains unchanged.

(iv) Polyp segmentation dataset (PolS)[44] [‖] PolS consists of 1000 polyp images. In our experiment, we select the first 600 images and divide the whole image set into training, validation and test sets according to the ratio of 6:2:2.

Remark 1 We use data augmentation, including flipping, scaling, clipping, *etc*., to transform the four datasets, respectively, generating new samples that increase the diversity and quantity of training data.

[‡]https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset
[§]https://challenge.isic-archive.com/data/#2018
[¶]https://www.kaggle.com/datasets/sani84/glasmiccai2015-gland-segmentation
[‖]https://www.kaggle.com/datasets/debeshjha1/medico-automatic-polyp-segmentation-challenge

Remark 2 For the proposed method M2ANet, we resize the input images to 512×512 due to the limitation of multi-scale size, and for the other comparison methods, the images are resized to 224×224.

4.2. Evaluation metrics

We use two commonly used metrics in the field of image segmentation to evaluate the segmentation performance of all methods. One is Dice coefficient,[45] and the other is Jaccard index,[46] also known as the intersection over union (IoU). As shown in Fig. 5, Dice evaluates the segmentation effect by calculating the degree of overlap between the predicted and real labeled regions, and IoU calculates the ratio of the intersection and concatenation of the predicted and real labeled regions.

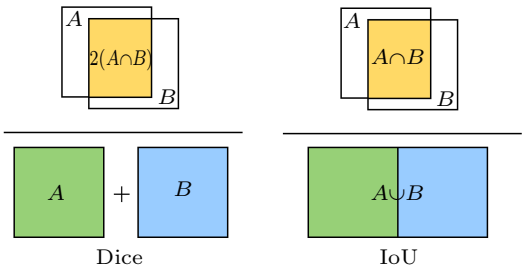


Fig. 5. Calculation diagrams of Dice and IoU, where Dice $= 2(A \cap B)/(A+B)$, and IoU $= (A \cap B)/(A \cup B)$.

4.3. Comparison methods

We compare our method M2ANet with several popular methods applied to medical image segmentation, including U-Net (MICCAI'15),[11] MultiResUNet (NeuNet'20),[14] SwinB (ICCV'21),[17] UCTransNet (AAAI'22),[18] Swin-Unet (ECCV'22),[19] SMESwin Unet (MICCAI'22),[24] UNeXt (MICCAI'22),[23] ACC-UNet (MICCAI'23),[20] and MADGNet (CVPR'24).[38]

For M2ANet, we employ SDG as the optimizer. For all experiments, we set the initial learning rate to 1e-3 and update it using the cosine annealing algorithm. For the GlaS dataset, due to the limited amount of data and in order to avoid insufficient training, we set the number of training epochs for the comparison experiments to 1000 and the number of training rounds for the M2ANet to 100. We also set an early-stop mechanism for all methods, which means that the experiments will stop if all metrics are not updated after 100 epochs. For the other datasets, the number of training epochs for all experiments is set to 100. We repeat the experiment three times under the same experimental conditions to ensure experimental reliability.

5. Experimental results

We conduct comparative experiments on the four benchmark datasets to show the effectiveness and feasibility of M2ANet. All experiments are conducted using PyTorch 1.11.0, and our model is trained on a single NVIDIA RTX 3090 GPU with 24 GB of memory.

5.1. Results of quantitative and qualitative analysis

To quantitatively evaluate segmentation performance, we report the results in terms of Dice and IoU achieved by all methods in Tables 1 and 2, respectively. We can observe that M2ANet achieves the best results on BUSI, ISIC2018, GlaS, and PolS. Among other advanced methods, suboptimal results are achieved on SwinB, MultiResUNet, MADGNet, and SwinB. This result is also intuitively reflected in the IoU metrics. These achievements not only validate the rationality of the model design, but also highlight its generalization ability in complex scenarios. Specifically, M2ANet enhances the model's representational capabilities by extracting multi-scale information and by fusing local details and global contextual features. These strategies enable M2ANet to accurately capture the fine structure of the target and its positional relationship in the overall scene when recognising and locating the target, and they can improve the accuracy of the model in image segmentation.

Table 1. Segmentation performance in term of Dice (↑) achieved by all comparison methods. The optimal results are highlighted in red, and the suboptimal are marked in blue.

Methods	BUSI	ISIC2018	GlaS	PolS
U-Net (MICCAI'15)	$71.66_{\pm 0.008}$	$85.99_{\pm 0.003}$	$87.57_{\pm 0.02}$	$76.99_{\pm 0.047}$
MultiResUNet (NeuNet'20)	$71.64_{\pm 0.003}$	**88.9**$_{\pm 0.004}$	$87.1_{\pm 0.011}$	$78.57_{\pm 0.015}$
SwinB (ICCV'21)	**77.44**$_{\pm 0.011}$	$88.38_{\pm 0.002}$	$87.63_{\pm 0.051}$	**88.95**$_{\pm 0.005}$
UCTransNet (AAAI'22)	$75.69_{\pm 0.007}$	$85.63_{\pm 0.004}$	$86.54_{\pm 0.019}$	$79.24_{\pm 0.035}$
Swin-Unet (ECCV'22)	$73.61_{\pm 0.001}$	$86.48_{\pm 0.002}$	$86.75_{\pm 0.004}$	$80.59_{\pm 0.006}$
SMESwin Unet (MICCAI'22)	$70.81_{\pm 0.01}$	$86.07_{\pm 0.002}$	$84.76_{\pm 0.004}$	$80.03_{\pm 0.007}$
UNeXt (MICCAI'22)	$74.34_{\pm 0.005}$	$87.5_{\pm 0.005}$	$85.86_{\pm 0.002}$	$73.98_{\pm 0.014}$
ACC-UNet (MICCAI'23)	$73.63_{\pm 0.024}$	$87.33_{\pm 0.005}$	$87.08_{\pm 0.014}$	$81.4_{\pm 0.006}$
MADGNet (CVPR'24)	$76.75_{\pm 0.008}$	$88.79_{\pm 0.006}$	**88.3**$_{\pm 0.008}$	$82.92_{\pm 0.009}$
M2ANet (ours)	**79.28**$_{\pm 0.004}$	**89.07**$_{\pm 0.01}$	**89.97**$_{\pm 0.008}$	**89.1**$_{\pm 0.003}$

Table 2. Segmentation performance in term of IoU (↑) achieved by all comparison methods. The optimal results are highlighted in red, and the suboptimal are marked in blue..

Methods	BUSI	ISIC2018	GlaS	PolS
U-Net (MICCAI'15)	$61.44_{\pm 0.012}$	$77.53_{\pm 0.001}$	$79.16_{\pm 0.032}$	$67.57_{\pm 0.053}$
MultiResUNet (NeuNet'20)	$60.94_{\pm 0.006}$	**81.2**$_{\pm 0.008}$	$78.61_{\pm 0.012}$	$69.93_{\pm 0.011}$
SwinB (ICCV'21)	**68.08**$_{\pm 0.007}$	$80.61_{\pm 0.003}$	$79.02_{\pm 0.076}$	**82.3**$_{\pm 0.005}$
UCTransNet (AAAI'22)	$66.09_{\pm 0.01}$	$77.21_{\pm 0.006}$	$77.66_{\pm 0.028}$	$69.58_{\pm 0.041}$
Swin-Unet (ECCV'22)	$62.65_{\pm 0.001}$	$78.11_{\pm 0.004}$	$77.29_{\pm 0.007}$	$70.17_{\pm 0.008}$
SMESwin Unet (MICCAI'22)	$58.72_{\pm 0.011}$	$77.06_{\pm 0.002}$	$74.35_{\pm 0.005}$	$69.09_{\pm 0.009}$
UNeXt (MICCAI'22)	$64.67_{\pm 0.005}$	$79.4_{\pm 0.005}$	$76.64_{\pm 0.001}$	$64.33_{\pm 0.018}$
ACC-UNet (MICCAI'23)	$64.19_{\pm 0.023}$	$79.47_{\pm 0.008}$	$78.43_{\pm 0.024}$	$72.5_{\pm 0.007}$
MADGNet (CVPR'24)	$67.25_{\pm 0.008}$	$81.04_{\pm 0.009}$	**80.26**$_{\pm 0.011}$	$74.65_{\pm 0.011}$
M2ANet (ours)	**70.06**$_{\pm 0.003}$	**81.58**$_{\pm 0.014}$	**82.53**$_{\pm 0.012}$	**82.48**$_{\pm 0.007}$

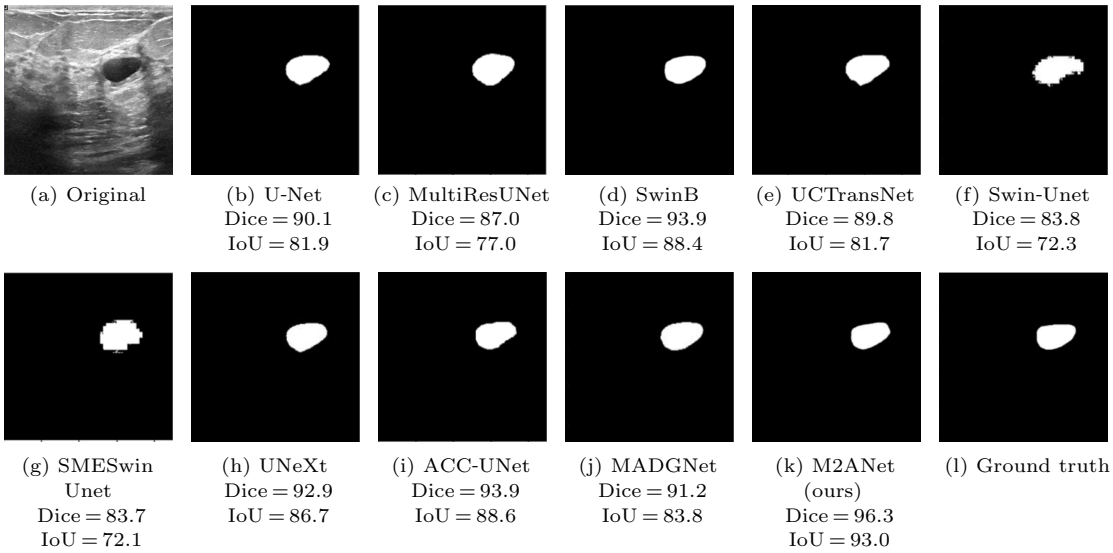
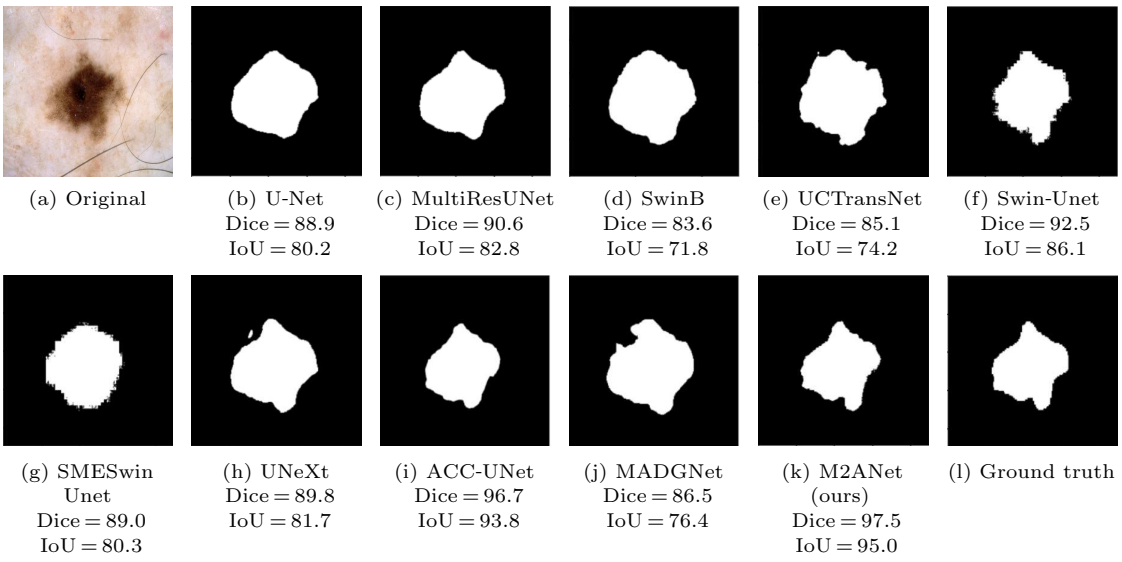


Fig. 6. Visualization of segmentation results obtained by different methods on BUSI dataset. The first and last images are the original image and the ground truth, respectively.

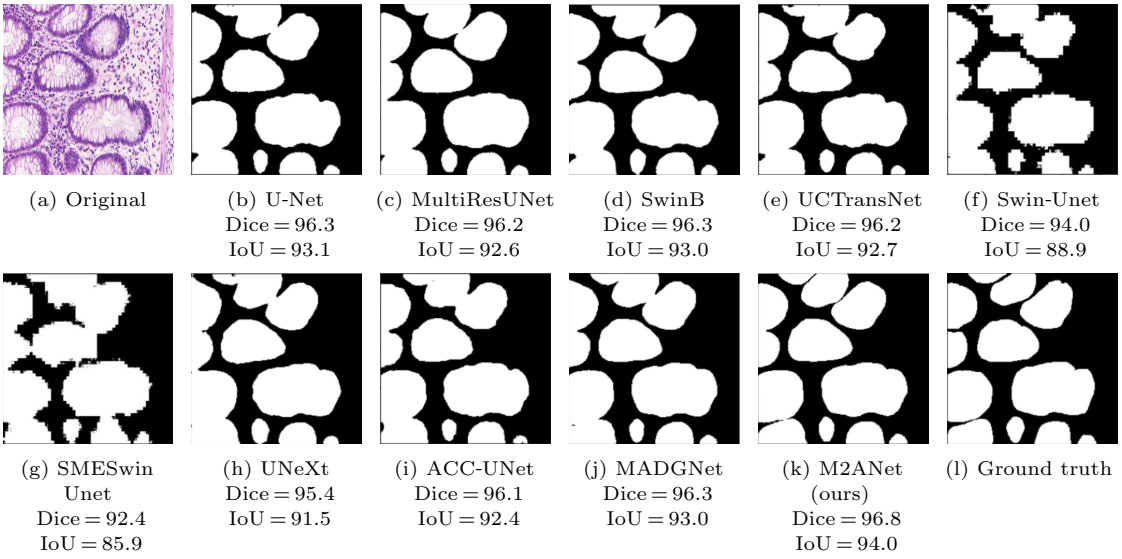
Further, to qualitatively demonstrate the advantage of the proposed M2ANet on these datasets, we show the segmentation results in Figs. 6–9. In each figure, the first image represents the original input image, and the last image denotes the ground truth. The remaining images show the prediction results obtained by all comparison methods and our proposed method. We can observe that M2ANet outperforms the other methods in terms of both decision boundary clarity and Dice scores. The method demonstrates a superior ability to precisely delineate the boundaries of the target regions, avoiding the typical issues of over-segmentation or under-segmentation observed in other approaches. These results highlight the robustness of M2ANet in accurately segmenting complex structures, making it a more reliable choice for various medical image segmentation tasks.

Specifically, figure 6 indicates that other methods tend to produce masks that are larger than the actual lesions or appear to be unreasonably segmented boundaries, which negatively impacts their performance. In contrast, our method avoids over-segmentation by accurately segmenting along the true boundaries of the lesions, obtaining more precise and reliable segmentation. In Fig. 7, the ISIC2018 dataset features more complex segmentation boundaries, which makes it prone to misjudgments at the edges by the model. However, compared to other models, by utilizing advanced technologies such as multi-scale feature fusion and attention mechanisms, our method effectively captures fine details and global backgrounds, enabling it to better adapt to the complex structure of the ISIC2018 dataset. In addition, our method achieves higher Dice scores and reduces boundary errors, demonstrating its effectiveness in addressing the challenges posed by complex segmentation tasks in medical images. As shown in Fig. 8, unlike other datasets, the GlaS dataset consists of multi-position segmented regions, which can lead to adhesions between neighboring areas. It can be seen that other models have encountered significant difficulties in dealing with adhesion problems, resulting in a decrease in their predictive performance. In contrast, our proposed method captures and fuses structural information of different scales through a multi-scale fusion strategy, which not only preserves clear segmentation boundaries but also improves the robustness of the model in adhesion problems. In Fig. 9, due to their similarity in structure and texture, other models are prone to misclassifying non-polyp areas as polyp tissue. Our model can simultaneously focus on the differences in local details and overall structure between polyp tissue and non-polyp areas, which avoids misclassification due to local similarity.



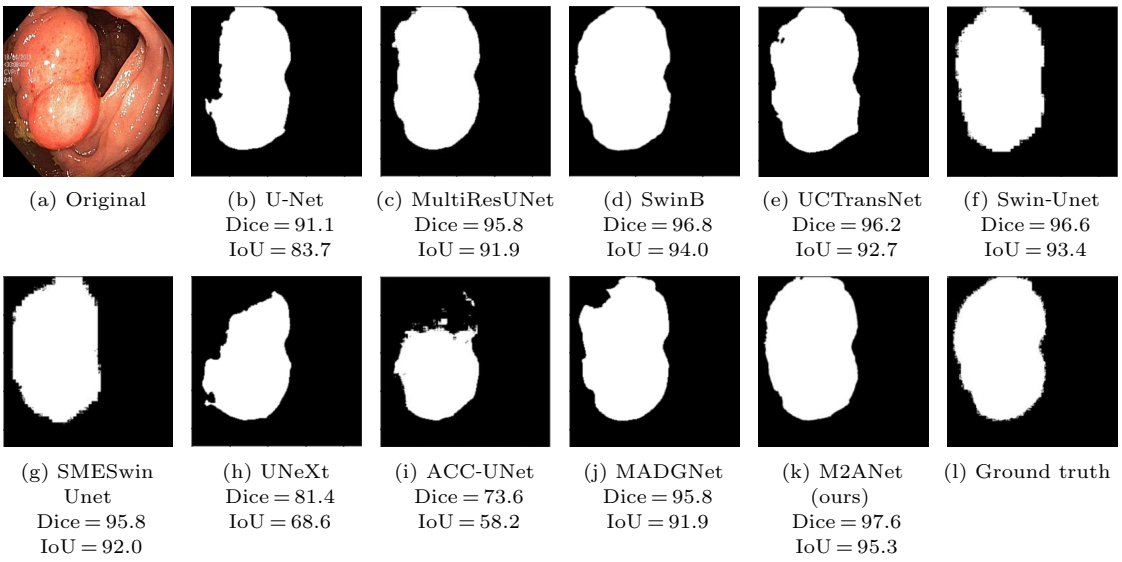
(a) Original
(b) U-Net Dice = 88.9 IoU = 80.2
(c) MultiResUNet Dice = 90.6 IoU = 82.8
(d) SwinB Dice = 83.6 IoU = 71.8
(e) UCTransNet Dice = 85.1 IoU = 74.2
(f) Swin-Unet Dice = 92.5 IoU = 86.1
(g) SMESwin Unet Dice = 89.0 IoU = 80.3
(h) UNeXt Dice = 89.8 IoU = 81.7
(i) ACC-UNet Dice = 96.7 IoU = 93.8
(j) MADGNet Dice = 86.5 IoU = 76.4
(k) M2ANet (ours) Dice = 97.5 IoU = 95.0
(l) Ground truth

Fig. 7. Visualization of segmentation results obtained by different methods on ISIC2018 dataset. The first and last images are the original image and the ground truth, respectively.



(a) Original
(b) U-Net Dice = 96.3 IoU = 93.1
(c) MultiResUNet Dice = 96.2 IoU = 92.6
(d) SwinB Dice = 96.3 IoU = 93.0
(e) UCTransNet Dice = 96.2 IoU = 92.7
(f) Swin-Unet Dice = 94.0 IoU = 88.9
(g) SMESwin Unet Dice = 92.4 IoU = 85.9
(h) UNeXt Dice = 95.4 IoU = 91.5
(i) ACC-UNet Dice = 96.1 IoU = 92.4
(j) MADGNet Dice = 96.3 IoU = 93.0
(k) M2ANet (ours) Dice = 96.8 IoU = 94.0
(l) Ground truth

Fig. 8. Visualization of segmentation results obtained by different methods on GlaS dataset. The first and last images are the original image and the ground truth, respectively.



(a) Original
(b) U-Net Dice = 91.1 IoU = 83.7
(c) MultiResUNet Dice = 95.8 IoU = 91.9
(d) SwinB Dice = 96.8 IoU = 94.0
(e) UCTransNet Dice = 96.2 IoU = 92.7
(f) Swin-Unet Dice = 96.6 IoU = 93.4
(g) SMESwin Unet Dice = 95.8 IoU = 92.0
(h) UNeXt Dice = 81.4 IoU = 68.6
(i) ACC-UNet Dice = 73.6 IoU = 58.2
(j) MADGNet Dice = 95.8 IoU = 91.9
(k) M2ANet (ours) Dice = 97.6 IoU = 95.3
(l) Ground truth

Fig. 9. Visualization of segmentation results obtained by different methods on PolS dataset. The first and last images are the original image and the ground truth, respectively.

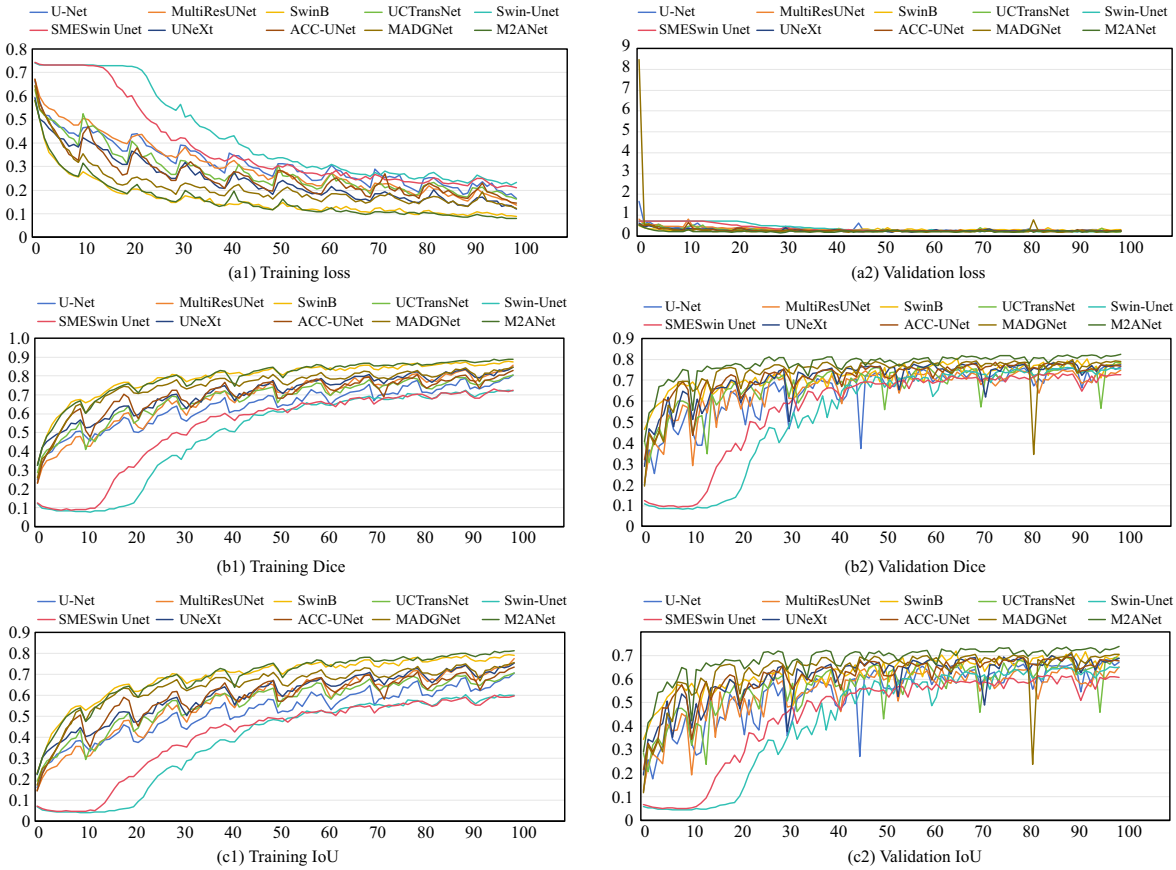


Fig. 10. The left shows the results of loss, Dice, and IoU obtained on the training portion of BUSI dataset. The right shows the results obtained on the validation portion.

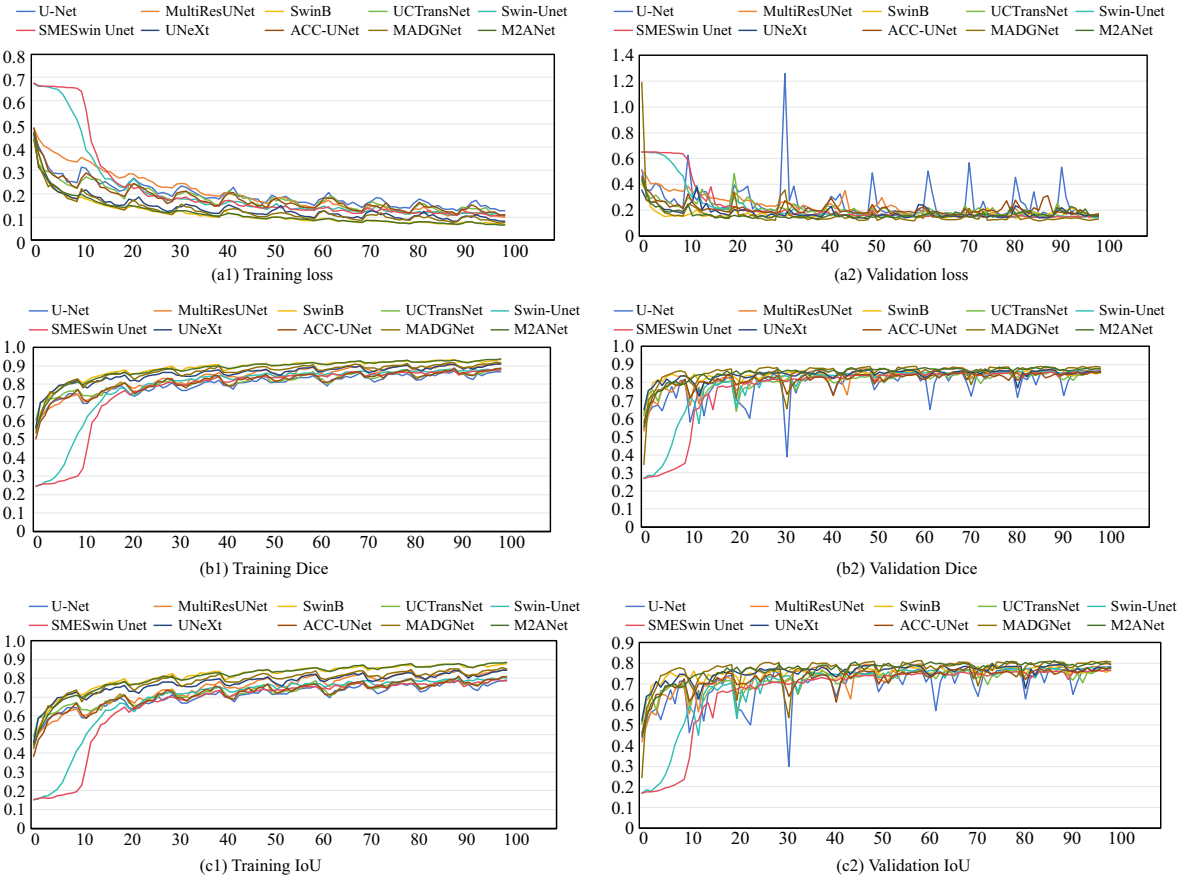


Fig. 11. The left shows the results of loss, Dice, and IoU obtained on the training portion of ISIC2018 dataset. The right shows the results obtained on the validation portion.

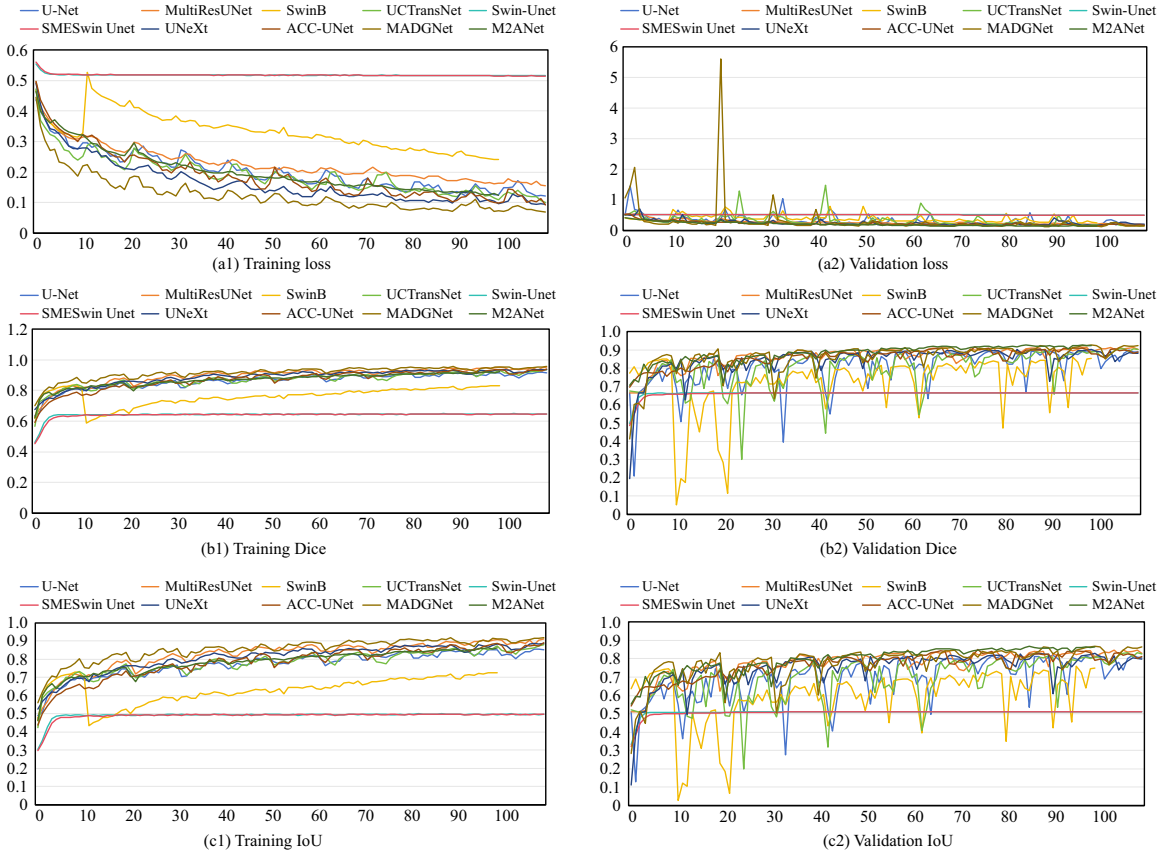


Fig. 12. The left shows the results of loss, Dice, and IoU obtained on the training portion of GlaS dataset. The right shows the results obtained on the validation portion.

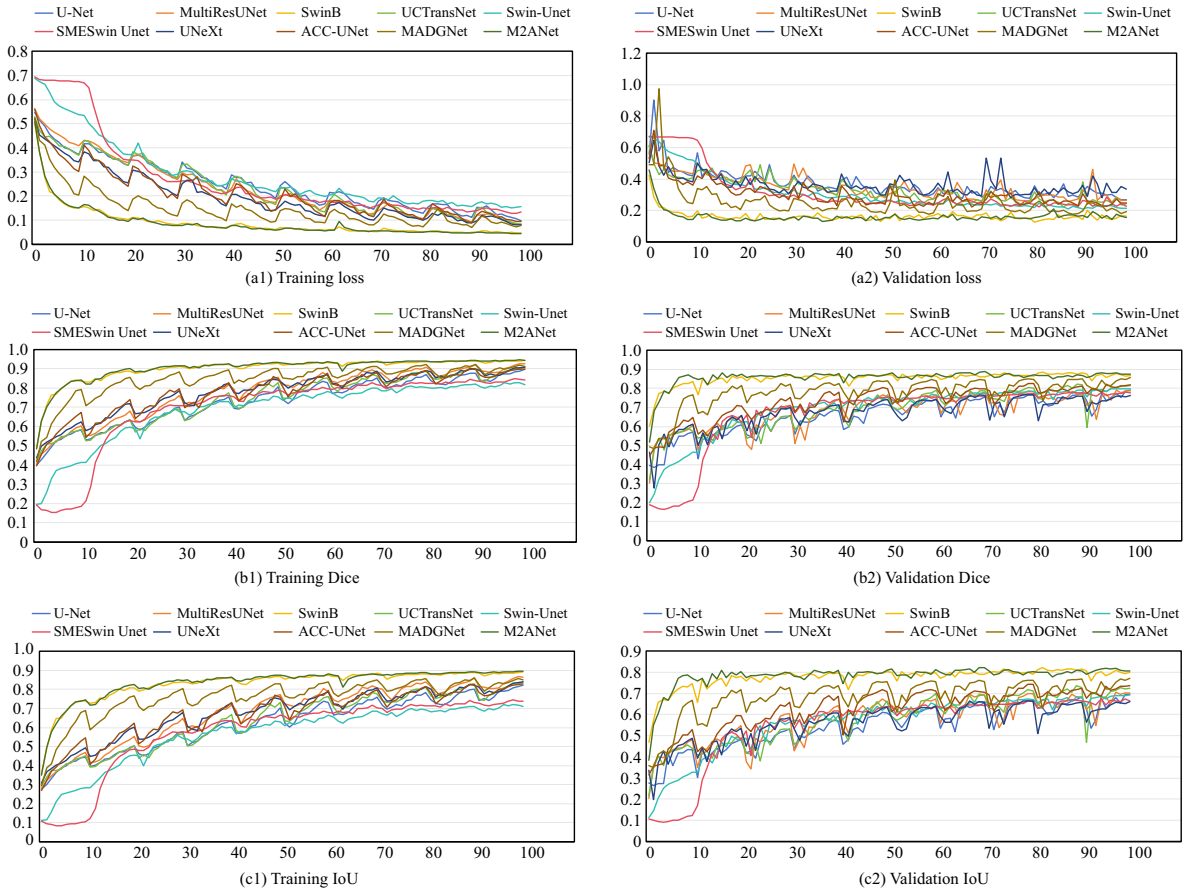


Fig. 13. The left shows the results of loss, Dice, and IoU obtained on the training portion of PolS dataset. The right shows the results obtained on the validation portion.

Figures 10–13 show the results of loss, Dice, and IoU obtained on the training portion and validation portion of each dataset. In each figure, panels (a1), (b1), and (c1) show the results of the training set, and panels (a2)–(c2) show the results of the validation set. In Figs. 10(a1)–10(c1), it can be seen that our method M2ANet (dark green) has better training results on the training set compared to other methods. Specifically, our method converges quickly on the training set and achieves high performance on a limited number of steps. And in the validation set, M2ANet has lower loss in the validation set and maintains higher Dice and IoU values while being more stable compared to other methods that oscillate on the validation set. As shown in Fig. 11, on the ISIC2018 dataset, all methods exhibit relatively stable results on both the training and validation sets. However, as can be seen from Fig. 11, our methods exhibit lower loss and higher Dice and IoU values, which indicates that our methods are able to better recognize and differentiate between the target region and the background, avoiding missegmentation and missed segmentation. In Fig. 12, it can be seen that all methods have some oscillation on the GlaS dataset, especially on the validation set, where the oscillation is more pronounced. This phenomenon can be attributed to the small number of training samples in GlaS, which leads to the inability to learn more comprehensive features during the training process. Nevertheless, our method is still able to maintain low loss values and high performance metrics relatively consistently. This suggests that our method is still able to learn effectively in the face of less data, capturing the key features in the data better and avoiding the phenomenon of overfitting. Especially in the validation set, the oscillation of our method is small, indicating that our method is more stable compared to other methods. As can be seen in Fig. 13, our method exhibits superior performance both in the training set and the validation set. In the PolS dataset, its polyp and non-polyp regions exhibit similar texture structure, which is easy for the model to mis-segment. This is also the reason why other models appear to oscillate. However, as can be seen from Fig. 13, both in the training set and in the validation set, our method has been maintaining lower loss and higher Dice and IoU metrics, which indicates that our method is more advantageous in dealing with images with similar textures, and the model generalization performance is higher.

In summary, the above qualitative and quantitative results show that the proposed method M2ANet is able to converge faster with lower loss values and higher Dice and IoU values than other methods. Meanwhile, by capturing and fusing local and global information, M2ANet effectively avoid common problems such as neighboring region sticking and similar region mis-segmentation, providing more accurate segmentation boundaries and demonstrating higher stability during the training process.

5.2. Ablation studies

We conduct a series of ablation studies by gradually adding different modules, with the results shown in Fig. 14. The Dice and IoU scores generally exhibit an upward trend, indicating improvements in model performance. Specifically, as shown in Fig. 14(a), on the BUSI dataset, we first achieve a Dice score of 69.38% using only the AMBPM module, which serves as our baseline performance. And then we add the CBAMRB module on top of the AMBPM module, which improve performance by about 1%. Finally, based on the previous two modules, we further improve the performance by about 9.4% by combining the MSFFM module, highlighting its key role in improving the segmentation accuracy of the BUSI dataset. Similarly, on the ISIC2018 dataset, compared to the previous module, the addition of the second and third module improve the Dice score by 2.3% and 2.4% respectively. This indicates that as the modules are stacked, the segmentation performanceof the model gradually improves, and the collaboration between modules is well represented. On the GlaS dataset, the addition of the second and third module improve the Dice score by about 1.1% and 1.1% respectively. On the PolS dataset, a similar trend was observed to that of the GlaS dataset. However, Dice and IoU values show an upward trend in both datasets. This indicates that the segmentation performance is gradually improving, and the model architecture is effective in enhancing the segmentation performance of the model. These results also indicate that the contribution of

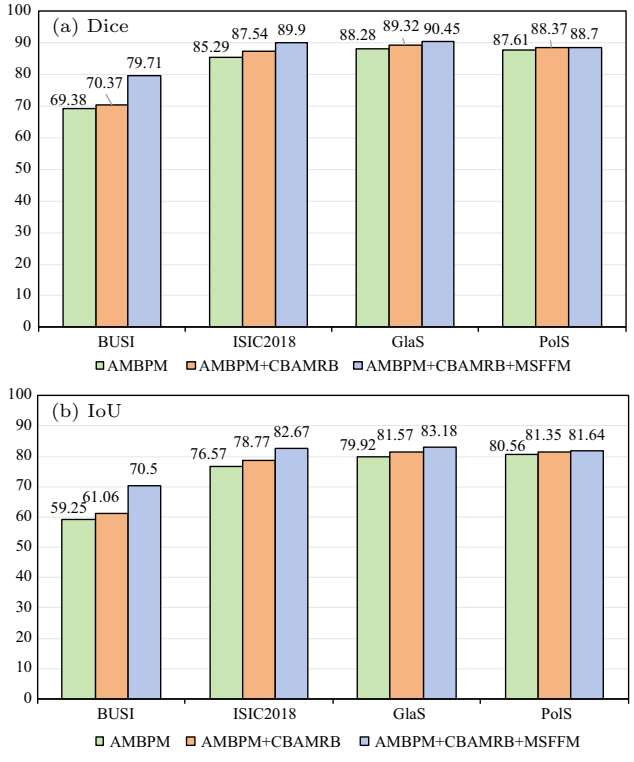


Fig. 14. Results of ablation study. We demonstrate module reliability by incrementing one at a time on top of swinB. As can be seen from the figure, the overall metrics show an incremental trend as the modules are superimposed.

different modules to image segmentation performance depends on the specific characteristics of each dataset.

The qualitative results of the ablation experiments are presented in Fig. 15. Figures 15(a1)–15(a4) represent the source image. Figures 15(b1)–15(b4) show the result after applying the AMBPM ablation. Figures 15(c1)–15(c4) depict the result of the AMBPM+CBAMRB ablation, and the figures 15(d1)–15(d4) illustrate the result of the AMBPM+CBAMRB+MSFFM ablation. As shown in these figures, panels (b1)–(b4) only show the rough texture of the original image, but it is relatively accurate for locating lesions. Panels (c1)–(c4) add some detailed information about lesions while maintaining texture, indicating their ability to effectively focus on key feature areas. Panels (d1)–(d4) not only enhance the clarity of texture, but also restore the details of the original image, especially in complex areas and small structures.

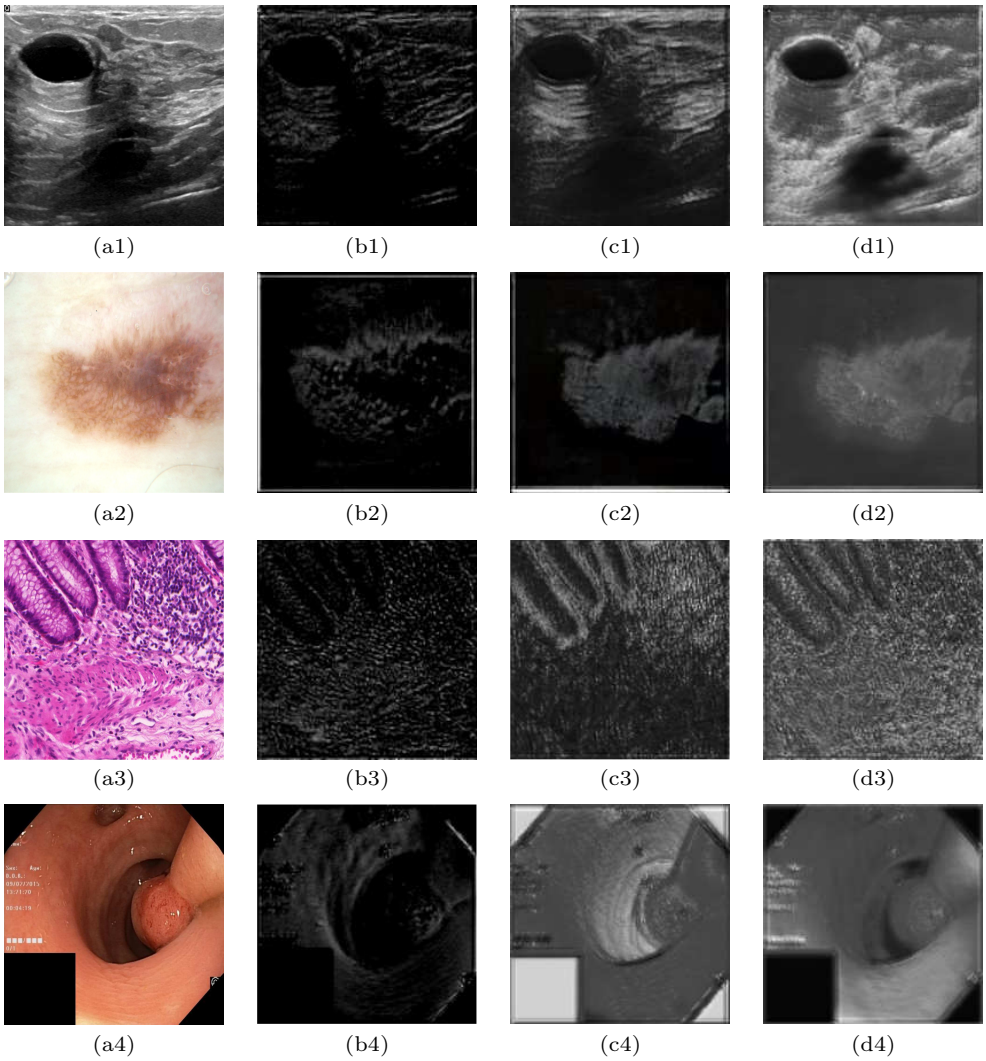


Fig. 15. Ablation study results obtained by M2ANe on four datasets. The first row [(a1)–(d1)] is the result of ablation study on BUSI. The second row [(a2)–(d2)] is the result of ablation study on ISIC2018. The third row [(a3)–(d3)] is the result of ablation study on GlaS. The fourth row [(a4)–(d4)] is the result of ablation study on PolS.

Overall, the above ablation results indicate that the combination of AMBPM+CBAMRB+MSFFM can work together on different feature layers of the image, gradually enhancing the texture information of the image, providing more accurate segmentation boundaries and improving the robustness of the model.

6. Conclusions

In this paper, we proposed a novel encoding-decoding network M2ANet for medical image segmentation. To effectively obtain local and global information from images, we used swinB as an encoder to extract multi-scale information, and used AMBPM, CBAMRB, and MSFFM as decoders to effectively fuse local and global information and improve the representation ability of model. The experiments on four benchmark medical image segmentation datasets show that our method M2ANet performs well on different metrics and has excellent segmentation performance. Through ablation experiments, it can be seen that our method is useful in lesion localization and image texture restoration and can preserve complete contour and texture information.

Acknowledgements

Project supported by the Natural Science Foundation of the Anhui Higher Education Institutions of China (Grant Nos. 2023AH040149 and 2024AH051915), the Anhui Provincial Natural Science Foundation (Grant No. 2208085MF168), the Science and Technology Innovation Tackle Plan Project of Maanshan (Grant No. 2024RGZN001), and the Scientific Research Fund Project of Anhui Medical University (Grant No. 2023xkj122).

References

[1] Luo L, Wang X, Lin Y, Ma X, Tan A, Chan R, Vardhanabhuti V, Chu W C, Cheng K T and Chen H 2024 *IEEE Rev. Biomed. Eng.* **18** 130
[2] Zhao Y, Gui W, Chen Z, Tang J and Li L 2005 *IEEE Engineering in Medicine and Biology 27th Annual Conference*, January 17–18, 2005, Shanghai, China, p. 6492
[3] Jiang F, Frater M R and Pickering M 2012 *International Conference on Digital Image Computing Techniques and Applications*, December 03–05, 2012, Fremantle, WA, Australia, p. 1
[4] Gould S, Gao T and Koller D 2009 *Neural Information Processing Systems*, December 7–11, 2009, USA, p. 366
[5] Wang R, Lei T, Cui R, Zhang B, Meng H and Nandi A K 2022 *IET Image Process.* **16** 1243
[6] Ruan J, Li J and Xiang S 2024 arXiv: 2402.02491
[7] Lou A, Guan S and Loew M 2021 *Proc. SPIE* 11596*, Med. Imaging* 2021*: Image Process.*, February 15–20, 2021, California, USA, p. 115962T
[8] Brahim A S, Abdelhamid E H and Aicha M 2018 *Procedia Comput. Sci.* **127** 109
[9] Deb S D and Jha R K 2023 *IEEE Trans. Radiat. Plasma. Med. Sci.* **7** 151
[10] Ma J, Hou K, Bao S and Chen C 2011 *Chin. Phys. B* **20** 028701
[11] Ronneberger O, Fischer P and Brox T 2015 18*th International Conference Medical Image Computing and Computer-Assisted Intervention*, October 5–9, 2015, Munich, Germany, p. 234
[12] Zhou Z, Siddiquee M M R, Tajbakhsh N and Liang J 2018 *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, September 20, 2018, Granada, Spain, p. 3
[13] Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, Han X, Chen Y and Wu J 2020 *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 04–08, 2020, Barcelona, Spain, p. 1055
[14] Ibtehaz N and Rahman M S 2020 *Neural Networks* **121** 74
[15] Sinha A and Dolz J 2021 *IEEE J. Biomed. Health Inform.* **25** 121
[16] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N 2020 arXiv: 2010.11929
[17] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S and Guo B 2021 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 10–17, 2021, Montreal, QC, Canada, p. 9992
[18] Wang H, Cao P, Wang J and Zaiane O R 2022 *Proceedings of the AAAI Conference on Artificial Intelligence*, February 22–March 1, 2022, held virtually, p. 2441
[19] Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q and Wang M 2022 *European Conference on Computer Vision*, October 23–27, 2022, Tel Aviv, Israel, p. 205
[20] Ibtehaz N and Kihara D 2023 *26th International Conference Medical Image Computing and Computer Assisted Intervention*, October 8–12, 2023, Vancouver, BC, Canada, p. 692
[21] He J, Deng Z and Qiao Y 2019 *IEEE/CVF International Conference on Computer Vision*, October 27–November 02, 2019, Seoul, Korea (South), p. 3561
[22] Poudel S and Lee S W 2021 *Appl. Soft Comput.* **109** 107445
[23] Valanarasu J M J and Patel V M 2022 25*th International Conference Medical Image Computing and Computer Assisted Intervention*, September 18–22, 2022, Singapore, p. 23
[24] Wang Z, Min X, Shi F, Jin R, Nawrin S S, Yu I and Nagatomi R 2022 25*th International Conference, Singapore Medical Image Computing and Computer Assisted Intervention*, September 18–22, 2022, Singapore, p. 517
[25] Cheng T, Zhao R, Wang S, Wang R and Ma H 2024 *Chin. Phys. B* **33** 040303
[26] Wang S, Wang K, Cheng T, Zhao R, Ma H and Guo S 2024 *Chin. Phys. B* **33** 060310
[27] Pan H, Liu M, Ge H and Yuan Q 2024 *Chin. Phys. B* **31** 120701
[28] Lin T, Dollar P, Girshick R, He K, Hariharan B and Belongie S 2017 *IEEE Conference on Computer Vision and Pattern Recognition*, July 21–26, 2017, Honolulu, HI, USA, p. 936
[29] Liu S, Huang D and Wang Y 2018 *Proceedings of the European Conference on Computer Vision*, September 8–14, 2018, Munich, Germany, p. 404
[30] Xu Q, Ma Z, Na H and Duan W 2023 *Comput. Biol. Med.* **154** 106626
[31] Rahman M M and Marculescu R 2023 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, January 2–7, 2023, Waikoloa, HI, USA, p. 6211
[32] Rahman M M, Munir M and Marculescu R 2024 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 16–22, 2024, Seattle, WA, USA, p. 11769
[33] Zhao X, Jia H, Pang Y, Lv L, Tian F, Zhang L, Sun W and Lu H 2023 arXiv: 2303.10894
[34] Brauwers G and Frasincar F 2023 *IEEE Trans. Knowl. Data. Eng.* **35** 3279
[35] Oktay O, Schlemper J, Folgoc L L, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla N Y, Kainz B, Glocker B and Rueckert D 2018 arXiv: 1804.03999
[36] Michael Y, Evis S, Carola S B and Leonardo R 2021 *Comput. Biol. Med.* **137** 104815
[37] Ruan J, Xie M, Gao J, Liu T and Fu Y 2023 *Medical Image Computing and Computer Assisted Intervention*, October 8–12, 2023, Vancouver, BC, Canada, p. 481
[38] Nam J, Syazwany N S, Kim S J and Lee S C 2024 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 16–22, 2024, Seattle, WA, USA, p. 1148
[39] Woo S, Park J, Lee J Y and Kweon I S 2018 *Proceedings of the European Conference on Computer Vision*, September 8–14, 2018, Munich, Germany, p. 3
[40] Yu Q, Zhao X, Pang Y, Zhang L and Lu H 2024 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 16–22, 2024, Seattle, WA, USA, p. 3921
[41] Al-Dhabyani W, Gomaa M, Khaled H and Fahmy A 2020 *Data Brief* **28** 104863
[42] Noel C F C, David G, M Emre C, Brian H, Michael A M, Stephen W D, Aadi K, Konstantinos L, Nabin M, Harald K and Allan H 2018 *IEEE* 15*th International Symposium on Biomedical Imaging*, April 4–7, 2018, Washington, DC, USA, p. 168
[43] Sirinukunwattana K, Pluim J P W, Chen H, Qi X, Heng P, Guo Y, Wang L, Matuszewski B J, Bruni E, Sanchez U, Bohm A, Ronneberger O, Cheikh B B, Racoceanu D, Kainz P, Pfeiffer M, Urschler M, Snead D R J and Rajpoot N M 2017 *Med. Image Anal.* **35** 489
[44] Jha D, Hicks S A, Emanuelsen K, Johansen H, Johansen D, Lange T D, Riegler M A and Halvorsen P 2020 arXiv: 2012.15244
[45] Dice L R 1945 *Ecology* **26** 297
[46] Jaccard P 1912 *New Phytol.* **11** 37