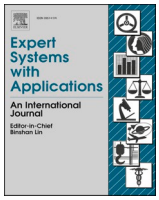
Expert Systems With Applications 288 (2025) 128241

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



MSM-UNet: A medical image segmentation method based on wavelet transform and multi-scale Mamba-UNet

Junding Sun [a,1], Kaixin Chen [a], Xiaosheng Wu [a,2], Zhaozhao Xu [a,3], Shuihua Wang [b,4], Yudong Zhang [a,c,d,*]

[a] School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan 454000, PR China
[b] Department of Biological Sciences, School of Science, Xi'an Jiaotong Liverpool University, Suzhou, Jiangsu 215123, PR China
[c] School of Computing and Mathematical Sciences, University of Leicester, LE17RH, UK
[d] Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

ARTICLE INFO

Keywords:
Medical image segmentation
Mamba
CNN
Multi-scale fusion
Boundary enhancement

ABSTRACT

In the field of medical image processing, combining global and local relationship modeling is an effective method for achieving precise image segmentation. Previous studies have demonstrated the remarkable performance of Convolutional Neural Networks (CNNs) in local relationship modeling, while Transformer can directly establish interactions between any two points in an image, thereby effectively capturing global contextual information. However, the application of Transformer to address the shortcomings of Convolutional Neural Networks (CNNs) in modeling global relationships is hindered by their substantial computational complexity and substantial memory demands, posing significant challenges in practice. To address this issue, this paper introduces the Mamba model, a State Space Model (SSM) that exhibits notable advantages in modeling long-range dependencies in sequential data. Inspired by the success of the Mamba model, a two-dimensional medical image segmentation model named MSM-UNet is designed. This model employs a Multi-Scale Mamba feature extraction block (MSMamba), a Wavelet Transform Feature Enhancement Attention Block (WTFEAB), a Feature Enhancement Merge Block (FEMB), and a Fusion Output Layer (FOL), aiming to accurately capture and integrate long-range and local dependencies among multi-scale features. Compared to Transformer-based methods, MSM-UNet exhibits superior performance in holistic feature modeling, significantly improving segmentation accuracy. Tests conducted on the Automatic Cardiac Diagnosis Challenge (ACDC) dataset, the Synapse multi-organ CT abdominal segmentation dataset, and the Colorectal Cancer-Clinic (CVC-ClinicDB) dataset demonstrated that the MSM-UNet achieved Dice coefficients of 92.02, 83.10, and 94.03, respectively. These results comprehensively validate the efficacy and practicality of the proposed MSM-UNet architecture.

1. Introduction

Medical image segmentation occupies an indispensable position in the disease treatment workflow, characterized by its complexity and precision. It relies on various medical imaging technologies such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Optical Coherence Tomography (OCT) to accurately delineate the target areas of internal organs or lesions. When dealing with multi-class image segmentation tasks in datasets like ACDC and Synapse, the ability to precisely distinguish multiple complex organs, while considering the significant human and material resources required for this process, is particularly important. Therefore, the development of a medical image segmentation method that can both ensure accuracy and improve efficiency, aiming to enhance segmentation precision and efficiency while reducing resource consumption, is crucial. Such methods will automate the processing of complex medical image data, providing clinicians with

* Corresponding author.
E-mail addresses: sunjd@hpu.edu.cn (J. Sun), yudongzhang@ieee.org (Y. Zhang).
[1] 0000-0001-7349-0248.
[2] 0000-0001-9382-3521.
[3] 0000-0001-6936-9357.
[4] 0000-0003-2238-6808.

https://doi.org/10.1016/j.eswa.2025.128241
Received 20 December 2024; Received in revised form 7 May 2025; Accepted 18 May 2025
Available online 20 May 2025

more accurate and rapid diagnostic assistance, thereby promoting innovation and progress in medical image processing technologies.

In recent years, CNNs have achieved significant progress in medical image segmentation, with the U-Net architecture (Ronneberger et al., 2015) proposed by Ronneberger et al. as a classic example. This architecture adopts a fully symmetric encoder-decoder design and incorporates skip connections, effectively facilitating the utilization of multi-level features in the encoder, thereby achieving excellent segmentation results. Based on the U-Net architecture, researchers have developed a series of variants (Siddique et al., 2021) to further enhance performance. For instance, U-Net++ (Zhou et al., 2019) and UNet3+ (Huang et al., 2020) introduce dense skip connections and full-scale skip connections respectively, promoting the fusion of multi-scale features in the decoder and optimizing the quality of segmentation results. To address the issue of gradient vanishing in deep networks and capture richer semantic information, the ResUNet (Szegedy et al., 2016; Diakogiannis et al., 2020; Maji et al., 2022; Li et al., 2021) and DenseUNet (Li et al., 2018; Safarov & Whangbo, 2021; Cai et al., 2020; Guan et al., 2019) series integrate residual blocks and dense blocks into the U-Net architecture, achieving improvements to a certain extent. Additionally, researchers have introduced attention mechanisms, significantly enhancing the network's ability to capture features. Models such as Esdmr-Net (Khan et al., 2024), OAU-Net (Song et al., 2023), Ga-UNet (Pang et al., 2024), and Mta-Net (Ling et al., 2024) have demonstrated improved network accuracy by integrating attention modules. To better extract edge information of the regions of interest, researchers have applied some edge detection algorithms to the field of medical image processing. For example, CED (Taher et al., 2023) proposes an improved Canny edge detection method for diagnosing brain tumors, which is particularly important for improving the accuracy of image boundary recognition. However, the issue of CNNs lacking long-range dependencies still persists.

To overcome the limitations of CNNs in capturing long-range dependencies, the Transformer architecture emerged. The Transformer, based on the self-attention mechanism, was initially introduced in the field of Natural Language Processing (NLP) (Vaswani et al., 2017) for sequence-to-sequence prediction tasks. By utilizing the self-attention mechanism, the Transformer can learn the correlations between global features, thereby capturing long-range dependencies. To adapt to applications in the image domain, Zhai Xiaohua et al. proposed Vision Transformer (ViT) (Dosovitskiy et al., 2021), which divides an image into non-overlapping patches, embeds these patches into the Transformer module, and employs the self-attention mechanism to capture global features in the image. To enhance inference speed and reduce the number of Transformer parameters, researchers have developed hierarchical Vision Transformer, such as Swin-Transformer (Liu et al., 2021; He et al., 2022; Sun et al., 2022) and Pyramid Vision Transformer (PVT) (Zhang et al., 2022; Wang et al., 2022) architectures. These methods accelerate the training process by improving the patch segmentation method or applying the self-attention mechanism with different strategies. Similarly, researchers have also applied edge detection algorithms combined with the Transformer architecture to medical image segmentation. For example, the WT-Swin (Azad et al., 2023) model cleverly integrates wavelet transform and attention mechanism modules, aiming to selectively focus on high-frequency information while effectively preserving low-frequency information. This design strategy significantly improves the accuracy of classification tasks without additional computational complexity. On the other hand, the HFE-Transformer (Dihin et al., 2024) model innovatively constructs a Gaussian pyramid structure based on high-frequency features to generate additional attention maps, thereby enhancing the representation of image boundary information.

Furthermore, to better address the lack of long-range dependencies in CNNs and short-range dependencies in Transformer, some researchers have explored combining Transformer and CNNs architectures to compensate for the Transformer's deficiencies in capturing local features. Typical representatives of such methods include TransUNet (Chen et al., 2021); STransFuse (Gao et al., 2021); SEUNet-Trans (Pham et al., 2024), and DA-TransUNet (Sun et al., 2024). They adopt parallel or serial dual-channel feature extraction strategies, effectively fusing the advantages of Transformer and CNNs. Another group of researchers has focused on applying Transformer to the UNet architecture, such as Swin-UNet (Cao et al., 2022), EG-TransUNet (Pan et al., 2023), DS-TransUNet (Lin et al., 2022), and TransCASCADE (Rahman & Marculescu, 2023). These methods replace convolutional blocks in CNNs with Transformer and supplement missing local features by fusing global features of different sizes. These Transformer-based architectures have proven effective in medical image segmentation tasks. However, they still face issues such as excessive parameter counts and insufficient local feature capture, necessitating balance and optimization in future research.

To mitigate the issue of parameter redundancy in Transformer when capturing long-range dependencies, Mamba (Gu & Dao, 2023) introduces time-varying parameters into structured State Space Model (SSM) and proposes a hardware-aware algorithm that significantly enhances the efficiency of model training and inference. To extend the advantages of the Mamba model to visual tasks, Zhu Lianghui et al. explored Visual Mamba (Vim) (Zhu et al., 2024), a generalized visual backbone network. Vim utilizes positional embeddings to label image sequences and employs bidirectional state-space models to compress visual representations. This method has demonstrated superior performance compared to existing visual Transformer (such as ViT and PVT), while significantly improving computational and storage efficiency. Subsequently, research on the application of Mamba in the visual domain has made significant progress. Notably, in medical image segmentation tasks, Ruan Jiacheng et al. proposed a U-shaped medical image segmentation architecture named Vision Mamba UNet (VM-UNet (Zhang et al., 2024). This model introduces the Vision State Space (VSS) module as a fundamental component to capture extensive contextual information and constructs an asymmetric encoder-decoder structure. Another study by Wang Ziyang et al., namely Mamba-UNet (Wang et al., 2024), combines the functionality of Mamba with the U-Net architecture for medical image segmentation. Mamba-UNet adopts a pure visual Mamba encoder-decoder structure based on VMamba, incorporating skip connections to preserve spatial information at different network scales. Zhang Xinxin et al. proposed the Gmamba (Zhang & Mu, 2024) model, the first Mamba segmentation model for grapevine leaf diseases. It designs a Co-SSM to mine coarse-grained and fine-grained disease information, introducing SAB and CAB to enrich the scaling of feature information. Zou Binfeng et al. proposed the DeMambaNet (Zou et al., 2024) model, which combines deformable convolutions with Mamba, incorporating a clustering structure deformable encoder for dental medical image segmentation. While Mamba has alleviated the computational drawbacks of Transformer to some extent, it still lacks the historical issue of local feature dependency. The MRDB (Hu et al., 2024) was proposed by Hu et al., employing a Mamba-SSM and ResNet-34 dual-encoder architecture for multi-dimensional feature extraction. Subsequently, two optimized frameworks were developed by Li et al. (Li et al., 2025): SF-Mamba, featuring dynamic local–global feature fusion for morphological variations, and MF-Mamba, utilizing multi-scale global modeling for size-diverse lesions. Although Mamba has mitigated Transformer's computational limitations, its inherent local feature dependency challenge persists.

In summary, the core challenges currently faced include the inadequacy of CNNs in handling long-range dependencies, the parameter explosion associated with the adoption of Transformer architectures, and the difficulty in effectively fusing long-range and short-range feature information. Additionally, the significant morphological differences among human organs in organ segmentation, the difficulty in distinguishing the positional boundaries of various organs, and the inaccurate localization boundaries of lesions in lesion segmentation are urgent issues to be addressed. Furthermore, up-sampling and down-sampling in neural networks can result in the loss of some

information, and segmentation networks based on U-shaped architectures inevitably lose some important information during encoder-decoder inference, which also needs improvement. To address the aforementioned issues, this paper proposes the MSM-UNet architecture, whose core innovations lie in the design of the Multi-Scale Mamba Feature Extraction Block (MSMamba) and the Wavelet Transform Feature Enhancement Attention Block (WTFEAB). Unlike existing approaches, we extend Mamba into a dual-branch structure capable of both global and local feature extraction. The global branch maintains the original SSM to capture long-range dependencies, while the newly introduced local branch enhances short-range feature extraction by partitioning vectors into sub-vectors, thereby addressing the inherent limitations of conventional Mamba in local feature perception. To tackle the challenge of blurred organ and lesion boundaries in medical images, we innovatively incorporate a first-order Daubechies wavelet transform for single-level decomposition, leveraging three high-frequency subbands (HH, HL, LH) for boundary feature enhancement, coupled with an optimized attention mechanism to further improve feature representation. Additionally, to mitigate information loss during the downsampling and upsampling processes in U-shaped networks, we design two auxiliary modules: the Feature Enhancement Merge Block (FEMB), which optimizes information flow through multi-scale feature fusion, and the Fusion Output Layer (FOL), which ensures segmentation accuracy through cross-layer feature integration. The main contributions of this paper can be summarized as follows:

- The MSMamba module is proposed to enhance local feature extraction capabilities through a dual-branch architecture, effectively addressing the limitations of conventional Mamba models in capturing local features.
- The WTFEAB module is proposed, which employs first-order Daubechies wavelet transform for boundary enhancement and incorporates an improved channel-spatial attention mechanism to further enhance feature representation capability.
- The FEMB and FOL are proposed, which optimize information flow through multi-scale feature fusion and ensure segmentation accuracy via cross-layer feature integration, respectively.
- The MSM-UNet achieves superior accuracy (92.02 % on ACDC, 83.10 % on Synapse, 94.03 % on CVC-ClinicDB) with fewer parameters than current architectures, balancing performance and efficiency.

2. Method

Firstly, this paper delves into the core design philosophy and basic architecture of the MSM-UNet framework. Subsequently, it provides a detailed exploration of the key components within this architecture, including MSMamba, FEMB, FOL, and WTFEAB. Then, the paper introduces the types of activation functions adopted and the selection strategy for the loss function. These technical details collectively constitute a comprehensive system of the MSM-UNet architecture, laying a solid foundation for its exceptional performance in specific application tasks.

2.1. MSM-UNet architecture

The MSM-UNet architecture is a carefully designed deep-learning model specifically tailored for medical image segmentation tasks. It follows the basic principles of a U-shaped network structure and is primarily composed of three core components: an encoder, a skip connection fusion module, and a decoder. This architecture incorporates multiple key modules, including the MSMamba module, the WTFEAB module, the FEMB module, and the FOL module, which work together to achieve efficient image segmentation.

In the encoder stage, we adopted a pre-trained ResNet34 as the fundamental feature extraction network and introduced an optimized MSMamba module in the last layer of the encoder to enhance feature representation capabilities. The MSMamba module innovatively incorporates a multi-scale SSM, aiming to achieve relative global feature scanning and extraction across different medical image sequences. This module not only augments Mamba's capacity in local feature extraction but also integrates its advantage in long-range feature extraction, enabling precise capture of both global dependency features and relative local dependency features.

In the skip connection section, FEMB achieves efficient fusion by integrating connectivity channels and up-sampled features, further improving the quality of feature representation. Additionally, the WTFEAB module is utilized for comprehensive feature enhancement, thereby bolstering the model's information transmission capability.

The WTFEAB module employs wavelet transform techniques for self-enhancement processing of the target segmentation area. This strategy significantly enhances the model's perception of boundary regions, effectively avoiding misjudgments in segmentation results due to boundary ambiguity. Among these, the FEAB module comprehensively enhances feature representation capabilities in both channel and spatial dimensions.

In the decoder stage, the iterative output of each layer is passed to the FOL module. This module deeply and multi-scale integrates features, further elevating segmentation performance.

In summary, the MSM-UNet architecture achieves efficient and precise segmentation of medical images through a carefully designed combination of modules and an optimized network structure. The specific architecture is illustrated in Fig. 2-1:

In the illustrated MSM-UNet model, by deploying the MSMamba module in the last layer of the network encoder, a full integration of CNNs and Mamba is achieved, enhancing the extraction of global features. The FEMB module is specifically designed to strengthen the fusion of features passed up from the lower layer with those of the current layer's connected channels, improving the integration effect of feature information. Furthermore, the WTFEAB module is introduced to enhance the fused features, enhance boundary, ensuring the completeness of the feature fusion process and avoiding issues of inadequate fusion. Finally, the FOL module is responsible for effectively fusing the output results from each layer of the encoder. This step enhances the model's ability to recognize the locations and boundaries of various organs, thereby improving the accuracy of segmentation tasks.

2.2. Multi-scale mamba feature extraction block (MSMamba)

This module cleverly incorporates the idea of a multi-head mechanism into the Mamba architecture by segmenting vectors after the embedding dimension to form a fixed set of sub-vectors. Subsequently, the module applies the SSM to both the original vector and the sub-vector array for long-distance feature extraction. This design not only retains the original characteristics of long-distance feature extraction but also complements the ability to extract relatively short-distance features, while effectively controlling the growth of parameter volume to a certain extent.

As shown in Fig. 1, the MSMamba module takes the output from the last layer of the encoder as input. These features first undergo embedding and BatchNorm, followed by a dimensional projection to form three branches.The first branch performs global feature scanning and extraction: it initially converts pixel-level features into a 1D vector via Conv1d and then applies SSM for feature scanning and extraction.The second branch adopts a multi-head mechanism for local feature extraction—the 1D pixel vector obtained from Conv1d is split into a fixed set of 8 sub-vectors using a split operation. Each sub-vector is independently processed by an SSM for local feature extraction, and the results are then concatenated for fusion. Then applies a residual operation on the original features. The outputs from the first two branches are summed, passed through a ReLU activation function, and then multiplied with the original features to capture both long-range and local dependencies of

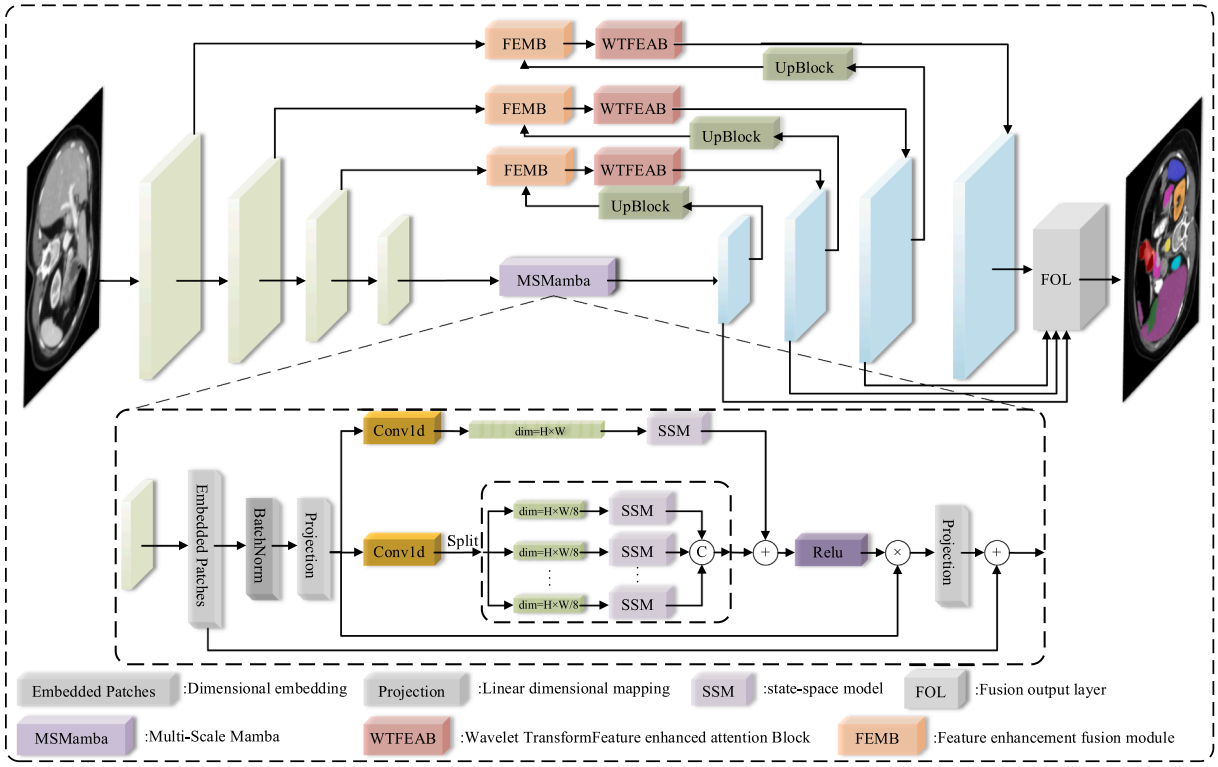


Fig. 1. Architecture diagram of the MSM-UNet.

the input features.

The MSMamba module is used for both global and relative local features. The specific pseudocode is as follows:

Algorithm 1 The procedure of MSMamba

Symbol Definition:
+: Element-wise add;
×: Element-wise product;
Input: *The map* $x^{in}_{MSMamba}$ *shape of size* (C, H, W)
Output: *The map* $x^{out}_{MSMamba}$ *shape of size* $(C, H*W)$
%Step1: dimension
$x_{embeding} = Embedding\ Patch(x^{in}_{MSMamba})$
$x_{projection} = Projection(BatchNorm(x_{embeding}))$
%Step2: Long and short distance feature extraction
$x^{(1)}_{branch} = SSM(Conv1d(x_{projection}))$
$x_{split} = Torch.Split(Conv1d(x_{projection}))$
for i = 0 to 8 **do**
$x^{(2)}_{branch} = Cat(x^{(2)}_{branch}, SSM(x_{split}[i]))$
end for
%Step3: Output
$x^{out}_{MSMamba} = x_{embeding} + Projection(x_{projection} \times ReLU(x^{(1)}_{branch} + x^{(2)}_{branch}))$

The MSMamba module takes a feature map as input, converts the feature map into vectors through an embedding layer, and then applies group normalization followed by pixel mapping. The mapped results flow into two branches: the first branch extracts global features as a whole, while the second branch divides the mapped vectors into fixed-size regions to perform relative local feature extraction. After passing through activation functions, the results are multiplied with the mapped features for feature enhancement. Finally, after transforming dimensions through a mapping layer, they undergo residual connections with the original features to obtain the output. The core implementation process is shown in Equations (1) –(4):

$$x_{projection} = Projection(BatchNorm(x_{embeding})) \tag{1}$$

$$x^{(1)}_{branch} = SSM(Conv1d(x_{projection})) \tag{2}$$

$$x^{(2)}_{branch} = Cat(x^{(2)}_{branch}, SSM(x_{split}[i])) \tag{3}$$

$$x^{out}_{MSMamba} = x_{embeding} + Projection(x_{projection} \times \mathrm{Re}LU(x^{1}_{branch} + x^{2}_{branch})) \tag{4}$$

In Equation (1), *Projection* denotes the dimensionality mapping operation. In Equation (2), $x^{(1)}_{branch}$ represents the global branch, i.e., the first branch, which does not segment the dimensions. In Equation (3), $x^{(2)}_{branch}$ represents the relative local branch, i.e., the second branch, where *Cat* denotes concatenation along the dimension parameter, and $x_{split}[i]$ indicates the vector after dimensionality segmentation, with i representing the index in the vector array. Equation (4) represents $x_{embeding}$, which is the output of the embedding module. Collectively, these equations illustrate the specific operational flow of the MSMamba module.

2.3. Wavelet transform feature enhancement attention block (WTFEAB)

The design intention of the WTFEAB module focuses on the precise and in-depth extraction of features at the concatenation channels after processing by the encoder. It employs a first-order Daubechies wavelet transform to decompose the feature maps into four distinct frequency components: HH (horizontal-high), HL (horizontal-low), LH (vertical-high), and LL (low-low). By leveraging the three high-frequency components (HH, HL, LH) that effectively capture directional edge features and fine-grained image details, the module implements a self-enhancement mechanism to amplify these critical characteristics. Furthermore, the integrated FEAB module conducts dual-dimensional enhancement across both channel and spatial domains, significantly improving segmentation accuracy for anatomical structures with high morphological variability. This is illustrated specifically in Fig. 2:

As shown in Fig. 2, the input feature map is first processed by a wavelet transform, decomposing it into results across different frequency bands, specifically including HH, HL, LH, and LL. Subsequently, the three components HH, HL, and LH are summed. Then, the sigmoid activation function is used to perform a nonlinear transformation on the

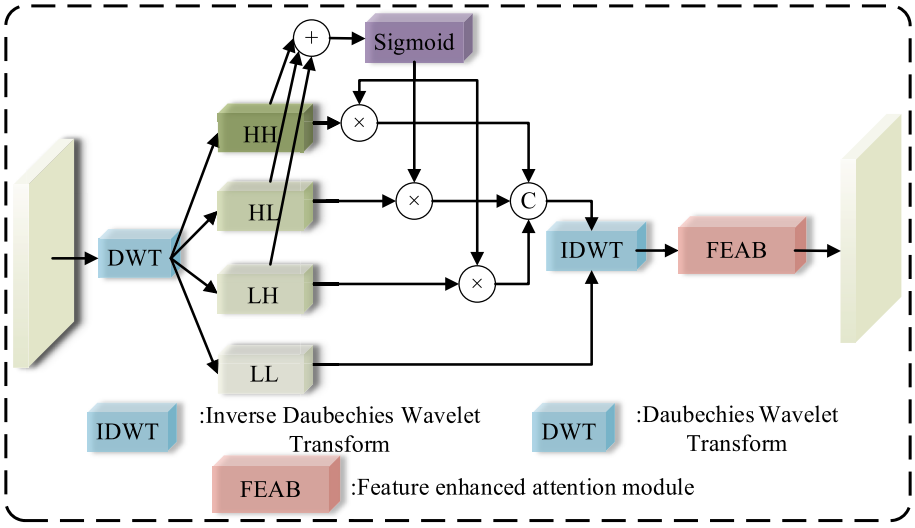


Fig. 2. Architecture diagram of the WTFEAB module.

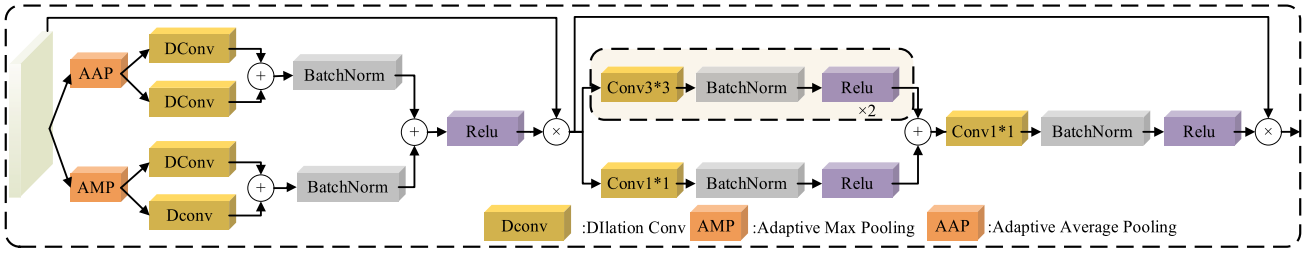


Fig. 3. Architecture diagram of the FEAB module.

summed result, obtaining a weight matrix. This weight matrix is multiplied with HH, HL, and LH, respectively, achieving weighted enhancement of these edge features and detailed information. Finally, an inverse wavelet transform is applied to recover the enhanced feature map. Afterward, this feature map undergoes further feature enhancement processing in all dimensions through the FEAB module to obtain the final processing result.

Through the FEAB module, the model can precisely select key feature channels closely related to the target in the channel dimension, thereby effectively enhancing these feature channels. Meanwhile, the FEAB module introduces an attention mechanism in the spatial dimension to strengthen information in key regions of the image, capturing finer spatial details. The synergistic effect of these two aspects contributes to the FEAB module's comprehensive and efficient allocation of attention to the encoder-output features, ultimately enhancing the model's overall performance in complex tasks. This is illustrated specifically in Fig. 3:

Fig. 3 provides a detailed description of the process of the FEAB, which is used to enhance features fused at the skip connection. The specific pseudocode is as follows:

Algorithm 2 The procedure of WTFEAB

Symbol Definition:
+: Element-wise add;
×: Element-wise product;
K: kernel size
Input: *The map shape* x^{in}_{WTFEAB} *of size* (C, H, W)
Output: *The map shape* x^{out}_{WTFEAB} *of size* (C, H, W)
%Step1: Wavelet Transform Boundary Enhancement
$x_{HH}, x_{HL}, x_{LH}, x_{LL} = WT(x^{in}_{WTFEAB})$
$x_{sigmoid} = Sigmoid(x_{HH} + x_{HL} + x_{LH})$
$x_{HH} = x_{HH} \times x_{sigmoid}$; $x_{HL} = x_{HL} \times x_{sigmoid}$; $x_{LH} = x_{LH} \times x_{sigmoid}$
$x^{out}_{WT} = IWT(Cat(x_{HH}, x_{HL}, x_{LH}, x_{LL}))$
%Step2: Channel Attention

(continued on next column)

(*continued*)

Algorithm 2 The procedure of WTFEAB

$x_{AAP} = BatchNorm(DConv(AdaptiveAvgPool(x^{out}_{WT}))) + BatchNorm(DConv(AdaptiveAvgPool(x^{out}_{WT})))$
$x_{AMP} = BatchNorm(DConv(AdaptiveMaxPool(x^{out}_{WT}))) + BatchNorm(DConv(AdaptiveMaxPool(x^{out}_{WT})))$
$x_{channel} = x^{out}_{WT} \times ReLU(x_{AAP} + x_{AMP})$
%Step3: Spatial Attention
$ConvBlock_{k\times k} = [Conv2d(kernel = k), BatchNorm, ReLU]$
$x_{3\times3} = ConvBlock_{3\times3}(ConvBlock_{3\times3}(x^{i}_{GCA}))$
$x_{1\times1} = ConvBlock_{1\times1}(x^{i}_{GCA})$
$featuremap = ConvBlock_{1\times1}(x_{1\times1} + x_{3\times3})$
$x_{spatial} = featuremap \times x_{channel}$
$x^{out}_{WTFEAB} = x_{spatial}$

As represented in the pseudocode, the input variables first undergo a wavelet transform to obtain three high-dimensional features and one low-dimensional feature. Subsequently, the high-dimensional features are summed and passed through a Sigmoid activation function to obtain a weight matrix. This matrix is then multiplied with each of the three high-dimensional features to obtain enhanced high-dimensional features. Finally, an inverse wavelet transform operation is applied to recover the original features.

For the channel attention input, the features first undergo adaptive average pooling and adaptive max pooling, respectively. The results from both are then processed using dilated convolutions (with a kernel size of 3 and dilation rates of 1 and 3) to extract features. The dilated convolution results from both processes are summed, group normalized, and then passed through a ReLU activation function. The resultant sum from the dilated convolutions is multiplied by the original features.

For spatial attention, a dual-channel approach is adopted to obtain spatial feature map weights. The first channel consists of two stacked

3x3 convolutional blocks, gradually expanding the receptive field to capture more comprehensive global spatial features. The second channel employs a 1x1 convolutional operation, which increases the network's nonlinear transformation capability without changing the spatial dimensions of the feature map, further enriching the feature representation. This avoids the issue of feature map blurring caused by average pooling and global max pooling in the spatial domain. Finally, the feature weights output by the two channels are summed and fused, and adjusted through an additional 1x1 convolutional block to optimize the distribution of feature weights. After passing through the ReLU activation function, the resulting weight map is element-wise multiplied with the original feature map to achieve attention weighting of the feature map, thereby emphasizing important spatial location information.

2.4. Feature enhancement and fusion mechanism

In medical image segmentation tasks, fusing cross-scale features plays a crucial role in enhancing the model's feature representation ability. To this end, we propose two strategies: Firstly, we utilize the FEMB module to fuse the up-sampled features from the next layer with the features of the current layer at the connection channel. Secondly, we employ the FOL module to enhance the fusion of the full-layer outputs of the decoder, generating the final segmentation result. These two strategies not only optimize the efficiency of feature transmission but also significantly improve the model's ability to comprehensively utilize features of different scales, thereby effectively enhancing the model's performance in medical image segmentation tasks.

2.4.1. FEMB module

As shown in Fig. 4, the green area specifically represents the set of feature vectors transmitted from the encoder to the current level's connection channel. Correspondingly, the blue area precisely indicates the set of feature vectors from the immediately subsequent level in the decoder, which has been meticulously processed through up-sampling techniques. For these two sets of feature vectors, we first separately apply standard convolutional operations and group normalization steps to obtain their respective processed feature representations. Subsequently, these two processed feature representations are element-wise added to fuse feature information from different sources. Then, to further enhance the expressive power and robustness of this fused feature, we employ a self-enhancement process using a 1×1 convolutional kernel. The core purpose of this step is to achieve self-reinforcement of the feature vectors through nonlinear transformations, allowing the feature vectors to enhance their own key information through self-learning. Ultimately, we obtain an enhanced fused feature result that integrates both the contextual information transmitted from the encoder and the up-sampled feature information from the decoder, while also being self-enhanced. This result provides more abundant and effective feature input for subsequent network layers.

2.4.2. FOL module

As shown in Fig. 5, the progressively increasing feature values displayed in the blue area represent the feature maps output by four different levels in the decoder. Subsequently, these four feature maps undergo up-sampling to ensure they have the same size. Then, each up-sampled feature map is processed separately through a basic convolutional block, which consists of standard convolutional operations, group normalization, and a ReLU activation function, aiming to effectively scale and adjust the features. Finally, these feature maps processed by the basic convolutional block undergo further refinement through a self-enhancement operation, which uses a Sigmoid activation function to convert the feature maps into weights and then multiplies these weights by the original feature maps themselves to achieve self-enhancement of the features. After this series of complex processing steps, the system is able to output the final prediction result. The specific operations are shown in the pseudocode below:

Algorithm 3 The procedure of FEMB and FOL

Variable and Symbol Definition:
x_{skip}: Skip connection of the current layer;
x_{up}: Up-sampling of the next layer;
$+$: Element-wise add;
$\times$: Element-wise product;
Input: *The map shape* x_{skip} *and* x_{up} *of size* (C, H, W)
Output: *The map shape* x_{FOL}^{out} *of size* $H \times W \times C$
%Step1: FEMB
$x_{skip} = BatchNorm(Conv(x_{skip}))$
$x_{up} = BatchNorm(Conv(x_{up}))$
$x_{FEMB}^{out} = (x_{skip} + x_{up}) \times Sigmoid(BatchNorm(Conv(x_{skip} + x_{up})))$
%Step2: Gets decoder output for each layer
$x_1 = WTFEAB(x_{FEMB}^{out(1)})$
$x_2 = WTFEAB(x_{FEMB}^{out(2)})$
$x_3 = WTFEAB(x_{FEMB}^{out(3)})$
$x_4 = x_{MSMamba}^{out}$
%Step3: FOL
$x_1^{up} = ReLU(BatchNorm(Conv(Upsample(x_1))))$
$x_2^{up} = ReLU(BatchNorm(Conv(Upsample(x_2))))$
$x_3^{up} = ReLU(BatchNorm(Conv(Upsample(x_3))))$
$x_4^{up} = ReLU(BatchNorm(Conv(Upsample(x_4))))$
$x_{FOL}^{out} = (x_1^{up} + x_2^{up} + x_3^{up} + x_4^{up}) \times Sigmoid(x_1^{up} + x_2^{up} + x_3^{up} + x_4^{up})$

In the context of pseudocode, x_{FEMB}^{out} denotes the output feature vector post-processing by the module. Specifically, x_{skip} refers to the feature map transmitted from the current level encoder to the connection channel, encapsulating crucial information from the encoding phase. Correspondingly, x_{up} signifies the output feature map from the immediately subsequent level in the decoder, which, upon up-sampling, aligns in size with the feature map transmitted from the encoder. Subsequently, the output of each encoder layer is obtained.

Let x_1 represent the output of the first decoder layer, x_2 represent the output of the second decoder layer, and so forth, with x_3 and x_4 denoting the outputs of the third and fourth decoder layers, respectively. These outputs from the four decoder layers undergo up-sampling to adjust their sizes. Following this, they are subjected to convolution operations, group normalization, and ReLU activation functions. After these operations, feature fusion is performed. Finally, the Sigmoid activation function is applied to map the feature weights onto themselves, and these weighted features are then multiplied by

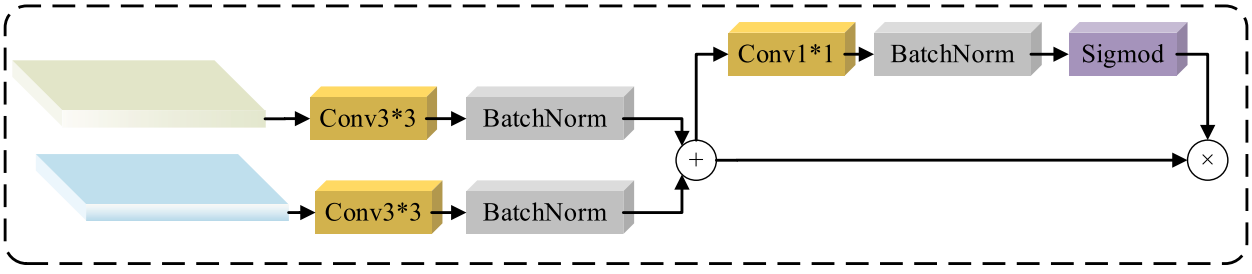


Fig. 4. Architecture diagram of the FEMB module.

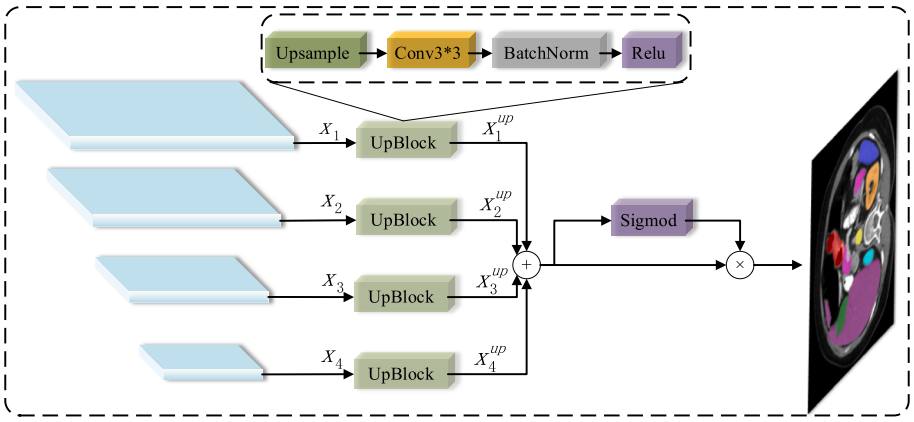


Fig. 5. Architecture diagram of the FOL module.

their corresponding features to derive the final prediction results.

2.5. Loss function strategy

The MSM-UNet architecture evaluates model performance on both the ACDC dataset and the Synapse dataset by combining Dice Loss and Cross Entropy Loss. Additionally, for lesion segmentation tasks, we introduce Weighted Binary Cross-Entropy Loss (WBCE).

During the training process, we employ a hybrid loss function that combines Dice Loss and Cross Entropy Loss to address issues related to class imbalance. Dice Loss (L_{dice}) and Cross Entropy Loss (L_{ce}) are defined as shown in Equations (5)-(6), and the final loss function used is presented in Equation (7):

$$L_{dice} = 1 - \sum_{k}^{K} \frac{2\omega_k \sum_i^N p(k,i) g(k,i)}{\sum_i^N p^2(k,i) + \sum_i^N g^2(k,i)} \tag{5}$$

$$L_{ce} = -\frac{1}{N} \sum_{i=1}^{N} G(k,i) \cdot \log(P(k,i)) + (1 - G(k,i)) \cdot \log(1 - P(k,i)) \tag{6}$$

$$L_{dice+ce} = \lambda L_{dice} + (1 - \lambda) L_{ce} \tag{7}$$

where N represents the number of pixels, $G(k,i) \in (0,1)$ and $P(k,i) \in (0,1)$ denote the ground truth label and generated probability for class k respectively. K is the number of classes, and $\sum_k \omega_k = 1$ represents the sum of weights for all classes. λ is a weighting factor that balances the influence of L_{dice} and L_{ce}. Based on previous research, ω_k and λ are set to $1/k$ and 0.7, respectively.

3. Experimental analysis

In this section, we first elaborate on the datasets employed in our experiments and meticulously design the experimental configurations and evaluation criteria for each dataset. Subsequently, we delve into the practical application effectiveness of the MSM-UNet architecture across multiple image segmentation tasks. Through a series of ablation experiments, we systematically validate the effectiveness of each component within the architecture and compare the performance of MSM-UNet with other existing methods in terms of metrics and model complexity analysis. Furthermore, to visually demonstrate the superior performance of MSM-UNet, we provide detailed experimental result charts and in-depth analyses.

3.1. Datasets

To validate the performance of the MSM-UNet architecture, this paper selects the ACDC dataset, the Synapse dataset, the CVC-ClinicDB dataset for empirical evaluation. The ACDC dataset focuses on fine-grained segmentation tasks of cardiac MRI images, covering various cardiac pathological conditions; the Synapse dataset addresses the complex segmentation of multiple organs in abdominal medical images; CVC-ClinicDB target lesion segmentation in colorectal cancer and gastrointestinal diseases, respectively. Through rigorous testing on these four high-standard datasets, this paper comprehensively and deeply assesses the effectiveness and generalization ability of the MSM-UNet architecture.

Specifically, the Synapse dataset is a commonly used benchmark in abdominal medical image segmentation research, containing 30 abdominal CT scan sequences with a total of 3,779 axial contrast-enhanced slices. Each CT scan sequence includes 85 to 198 slices with a resolution of 512 × 512 pixels. We randomly split this dataset into a training set (consisting of 18 sequences with 2,212 axial slices) and a validation set (consisting of 12 sequences). In this study, we focus on the segmentation of eight abdominal organs: aorta, gallbladder (GB), left kidney (KL), right kidney (KR), liver, pancreas (PC), spleen (SP), and stomach (SM).

The ACDC dataset is a public dataset dedicated to cardiac cine-MRI, containing 150 cases classified into five subclasses: normal, myocardial infarction with systolic heart failure, dilated cardiomyopathy, hypertrophic cardiomyopathy, and right ventricular abnormalities. The dataset provides precise annotations of the left ventricle, right ventricle, and myocardium for both diastolic and systolic frames. Following the setup of TransUNet, we use 70 cases (totaling 1,930 axial slices) for training, 10 cases for validation, and the remaining 20 cases for testing.

The CVC-ClinicDB dataset is a medical image dataset focused on colorectal cancer detection and diagnosis research. It contains 612 high-resolution colonoscopy images from colorectal cancer patients undergoing endoscopy. Each image is accompanied by expert annotations of colorectal cancer lesion areas (or polyp areas) provided as binary masks, clearly delineating the boundaries between lesion areas and normal tissues.

3.2. Evaluation metrics

To evaluate the effectiveness of the MSM-UNet architecture, we use the Dice coefficient (DICE), Mean Intersection over Union (mIoU), Hausdorff Distance (specifically, the 95 % Hausdorff Distance, HD95), BCE (Weighted BCE Loss), and Average Surface Distance (ASD) as evaluation metrics. Among these, HD95 is a metric used in neural networks to measure the distance between two boundaries, representing the distance between the predicted boundary and the ground truth boundary. The ASD measures, for each point on one surface, the Euclidean distance to the nearest point on the other surface. Then, the average of all these distances is taken to obtain the ASD. The formulas for the Dice coefficient and mIoU are expressed as follows:

$$DICE_{(A,B)} = \frac{2|A \cap B|}{|A| + |B|} \tag{8}$$

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{9}$$

$$mIoU = \frac{1}{n}\sum_{i=1}^{n} IoU_i \tag{10}$$

In formula (8), A and B represent the predicted set and the ground truth set, respectively, where |A ∩ B| denotes the number of elements in the intersection of the two sets, and |A| and |B| represent the number of elements in each set. In formulas (9) −(10), A and B represent the predicted set and the ground truth set for each category i, respectively, with n denoting the total number of categories. IoU stands for the Intersection over Union of a single category, while mIoU represents the Mean Intersection over Union across all categories.

In the experimental section of this paper, we strictly adhere to the existing evaluation standards within the field to ensure the fairness of the experimental results and their comparability with other studies. For the Synapse dataset, we employed a comprehensive set of evaluation metrics, including the DICE coefficient, mIoU, HD95, and ASD. These metrics not only assess the accuracy of the segmentation results but also provide an in-depth evaluation of the boundary adherence and shape similarity, offering profound insights into the model's performance. In the experiments with the ACDC dataset, we focused on evaluating the accuracy of the model in the task of heart structure segmentation, thus selecting the DICE coefficient as the primary evaluation metric. For the CVC-ClinicDB dataset, the DICE coefficient and mIoU were chosen as the primary evaluation metrics.

3.3. Experimental setup

All experiments in this study were conducted under the PyTorch 2.1.0 framework, with Python version 3.10 and CUDA version 11.8. The training of all models was performed on an NVIDIA RTX 4060 GPU equipped with 48 GB of memory. We employed the Adam optimizer, setting the learning rate to 1e-4 and the weight decay to 1e-4. The z_spacing parameter was set to 10, and the random seed was fixed at 2222 to ensure the reproducibility of the experiments.

In the experiments with the Synapse dataset, we followed the setup of TransUNet, using a batch size of 24 and limiting the maximum number of training epochs for each model to 150. The resolution of the input images was uniformly adjusted to 256 × 256 pixels, and data augmentation strategies such as random flipping and rotation were implemented to enhance the model's generalization performance. For the loss function, we combined cross-entropy loss and Dice loss, assigning weights of 0.3 and 0.7 to them, respectively, for a comprehensive evaluation of the model's performance.

For the ACDC dataset, we also trained each model for a maximum of 150 epochs, with the resolution of the input images set to 256 × 256 pixels and data augmentation methods such as random flipping and rotation employed. In terms of the loss function, we similarly adopted a combination of cross-entropy loss and Dice loss, assigning weights of 0.3 and 0.7 to them, respectively, to better suit the characteristics of the heart structure segmentation task.

For the CVC-ClinicDB dataset, we used a batch size of 16 and trained each model for a maximum of 100 epochs. We resized the images to 352 × 352 pixels and set the gradient clipping limit to 0.5. We employed a combined loss function of weighted IoU and Weighted BCE Loss.

3.4. Ablation experiment

This subsection delves into the effectiveness of each component module within the proposed MSM-UNet architecture and further validates the superior performance of MSM-UNet through comparative analysis with other architectures in terms of parameter quantity. To ensure the fairness of the experimental results, all experiments were conducted on the same dataset and using consistent parameter configurations. Through this series of detailed comparative analyses, the efficiency and feasibility of the MSM-UNet architecture in practical applications have been fully demonstrated.

3.4.1. The effectiveness of each module

The effectiveness of each module (MSMamba, WTFEAB, FEMB, FOL) within the MSM-UNet architecture was verified using the 256 × 256 size from the Synapse dataset as input, with the results presented in Table 1.

The baseline U-Net achieves a Dice score of 72.69 %. The introduction of the MSMamba module contributes a + 2.05 % improvement in Dice, while the WTFEAB module provides a + 1.36 % gain. The FEMB module further enhances performance by + 0.46 %, and the FOL module adds another + 0.7 % improvement. These modules exhibit synergistic effects, with the full configuration outperforming any partial configuration by an average of 0.5–1.2 percentage points. All components significantly contribute to reducing boundary errors (HD95 & ASD), with the FOL module demonstrating the most substantial impact on shape integrity refinement.

3.4.2. Intrinsic validation of MSMamba

In our internal validation study of MSMamba, we adopted the Synapse dataset with image inputs resized to 256 × 256.

Table 2 compares different architectural configurations of MSMamba. The complete framework achieves the best performance (83.50 % DICE, 15.24 HD95, 74.70 % mIoU, 2.58 ASD), outperforming both single-branch (82.92 % DICE, 3.11 ASD) and multiple-branch variants (82.74 % DICE, 3.34 ASD). While the single-branch version shows slightly better HD95 (15.00), its higher ASD (3.11) indicates less consistent boundary prediction compared to our full model (2.58 ASD).

3.4.3. Number of iterations for MSMamba

In this paper, the improved MSMamba module is utilized for long-distance feature extraction. Given the iterative approach of Transformer, we iterate the MSMamba module for model testing. Experiments are conducted on the Synapse dataset with an input size set to 256 × 256.

As shown in Table 3, when the MSMamba module is iterated twice, four times, and six times, the model's performance does not improve but rather decreases. In contrast, when the MSMamba module is used only once, the model achieves the highest performance.

3.4.4. WTFEAB module configuration alternatives

The proposed WTFEAB module is implemented to strengthen feature representation in our framework. All validations are performed on the Synapse dataset using 256 × 256 input resolution.

Table 4 demonstrates the progressive improvements achieved by our proposed modules. The WT module enhances performance to 82.53 % DICE (+0.39 over baseline), validating its effectiveness in frequency-domain structural preservation. The FEAB module alone achieves 82.64 % DICE (+0.50 over baseline), demonstrating the advantages of its dual-attention (channel/spatial) mechanism. Significantly, the integrated WTFEAB configuration attains 83.50 % DICE (+1.36 over baseline), exceeding the sum of individual module gains (0.39 + 0.50 = 0.89) by an additional 0.47. This synergistic improvement confirms the complementary benefits of combining wavelet frequency analysis with attention-based spatial refinement, while simultaneously reducing boundary errors (HD95 improved from 19.86 to 15.24, ASD from 3.29 to 2.58).

3.4.5. Selection of FOL layer depth

The proposed FOL module addresses the information degradation problem in up sampling operations. Following conventional U-net design principles, our experimental protocol systematically evaluates

Table 1
Contribution of each module to the overall performance in the MSM-UNet architecture, ↑ indicates that higher values are better, while ↓ indicates that lower values are better, × denotes the module is deactivated, while √ denotes the module is activated.

MSMamba	WTFEAB	FEMB	FOL	DICE↑	HD95↓	mIoU↑	ASD↓
×	×	×	×	72.69	24.23	63.18	4.32
×	√	√	√	81.45	20.35	71.23	3.87
√	×	√	√	82.14	19.86	72.76	3.29
√	√	×	√	83.04	17.04	74.10	2.76
√	√	√	×	82.80	17.88	73.94	2.90
√	√	√	√	**83.50**	**15.24**	**74.70**	**2.58**

Table 2
Intrinsic Validation of MSMamba, ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

Intrinsic Validation of MSMamba	DICE↑	HD95↓	mIoU↑	ASD↓
MSMamba	**83.50**	15.24	**74.70**	**2.58**
Single branch	82.92	**15.00**	74.18	3.11
Multiple branch	82.74	15.43	73.98	3.34

Table 3
Number of iterations for the MSMamba module, ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

Number of iterations of MSMamba	DICE↑	HD95↓	mIoU↑	ASD↓
1	**83.50**	**15.24**	**74.70**	**2.58**
2	83.03	16.93	74.04	3.22
4	82.64	17.90	73.86	4.14
6	82.07	19.32	73.18	4.77

performance by incrementally integrating hierarchical features, starting from the topmost layer as reference. All validations are performed on the Synapse dataset using 256 × 256 input resolution.

Table 5 demonstrates progressive performance gains through multi-scale fusion, with the single-layer baseline (L1) achieving 82.80 % DICE. Two-layer fusion (L1 + L2) improves to 82.96 % via complementary feature integration, while three-layer fusion (L1 + L2 + L3) reaches 83.19 % through mid-level semantic aggregation. The optimal four-layer fusion (L1 + L2 + L3 + L4) attains 83.50 %, where shallow layers (L1/L2) enhance boundary localization and deeper layers (L3/L4) provide global contextual guidance for coherent segmentation.

3.4.6. Parameter

This paper reproduces and computes the parameter counts for methods such as TransUNet, PVT-CASCADE, TransCASCADE, DS-TransUNet and Mamba-UNet, with MSM-UNet serving as a point of comparison. All methods are evaluated using an input size of 256 × 256 on the Synapse dataset.

As shown in Table 6, we found that the MSM-UNet architecture exhibits superior efficiency in terms of parameter count compared to other network architectures. Specifically, MSM-UNet (115.89 MB) demonstrates significant parameter reductions of 285.88 MB (71.2 % decrease) versus TransUNet (401.77 MB), 296.03 MB (71.8 % decrease) versus DA-TransUNet (411.92 MB), and 539.21 MB (82.3 % decrease) versus DS-TransUNet (655.10 MB). Compared to Transformer-based cascaded architectures, our method achieves reductions of 355.15 MB (75.4 % decrease) versus TransCASCADE (471.04 MB) and maintains advantages over similar lightweight architectures with 20.91 MB (15.3 % decrease) versus Mamba-UNet (136.80 MB) and 53.00 MB (31.4 % decrease) versus VM-UNet (168.89 MB).

Table 7 presents the parameter counts for each module in MSM-UNet. The improved MSMamba module has a parameter count of 6.9336 MB. The attention module combined with wavelet transform, known as WTFEAB, has a parameter count of 7.8596 MB. The feature fusion module between upper and lower layers, FEMB, has a parameter count of 0.6583 MB. The FOL module has a parameter count of 0.2217 MB.

Table 4
Effectiveness of the WTFEAB module components, ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

Element	DICE↑	HD95↓	mIoU↑	ASD↓
NULL	82.14	19.86	72.76	3.29
WT	82.53	19.13	73.44	3.12
FEAB	82.64	18.64	73.81	2.87
WTFEAB	**83.50**	**15.24**	**74.70**	**2.58**

Table 5
The cumulative effect of different FOL module levels, ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

FOL	DICE↑	HD95↓	mIoU↑	ASD↓
L1	82.80	17.88	73.94	2.90
L1 + L2	82.96	17.14	74.16	2.74
L1 + L2 + L3	83.19	16.59	74.49	2.62
L1 + L2 + L3 + L4	**83.50**	**15.24**	**74.70**	**2.58**

3.5. Analysis of comparison results

In this paper, we conduct a comprehensive comparative analysis of the proposed MSM-UNet architecture against various existing CNNs architectures, including the classic U-Net and its derivatives, on the Synapse dataset, ACDC dataset, as well as CVC-ClinicDB and datasets. Furthermore, we also compare MSM-UNet with a series of Transformer-based segmentation methods, including TransUNet, Swin-UNet, DA-TransUNet, DS-TransUNet, and TransCASCADE. Additionally, we consider methods based on other advanced frameworks, such as Mamba-UNet and VM-UNet, and include them in the scope of the comparative analysis.

3.5.1. Results on the synapse dataset

As shown in Table 8, compared to the currently top-performing TransCASCADE model, our proposed MSM-UNet architecture achieves improvements of 0.82 % and 1.22 % in the DICE coefficient and mean Intersection over Union (mIoU) metrics, respectively. Compared to the TransUNet model, MSM-UNet demonstrates more significant improvements in the DICE coefficient and mIoU metrics, with increases of 5.89 % and 7.38 %, respectively, while also achieving improvements of 11.66 %

Table 6
Parameter counts for various architectures (Add reference).

Method	Number of Parameters (MB)	FLOPs(G)	Memory (MiB)
TransUNet	401.7732	32.26	545.18
Swin-UNet	157.8535	16.39	340.87
DA-TransUNet	411.9228	29.48	570.36
DS-TransUNet	655.1034	51.15	834.02
PVT-CASCADE	134.5668	14.15	**279.41**
TransCASCADE	471.0377	31.68	658.49
Mamba-UNet	136.8047	13.86	296.14
VM-UNet	168.8916	**13.57**	321.71
MSM-UNet (ours)	**115.8900**	28.49	332.53

Table 7
Parameter counts for each module in MSM-UNet.

Model Block	Number of Parameters (MB)
MSMamba	6.9336
WTFEAB	7.8596
FEMB	0.6583
FOL	0.2217

and 2.08 % in the HD95 and ASD metrics, respectively. When compared to the Mamba-UNet model, MSM-UNet exhibits improvements of 2.92 % and 6.23 % in the DICE coefficient and mIoU metrics, respectively, and improvements of 6.71 % and 1.49 % in the HD95 and ASD metrics, respectively. Among all the compared methods, the MSM-UNet architecture demonstrates significant performance advantages in certain tissue segmentation tasks, with the highest performance achieved in the aorta (88.96 % DICE) and pancreas (85.65 % DICE), while also showing competitive results in gallbladder (71.33 % DICE) and spleen (95.03 % DICE) segmentation. This performance improvement is mainly attributed to the unique design of the MSM-UNet architecture, which can more effectively distinguish between important and non-important features through its multi-scale mixing mechanism, thereby exhibiting more prominent advantages when dealing with both small and large organs.

As shown in Fig. 6, the ordinate represents the value of loss, and the abscissa represents the value of epoch. This figure provides a detailed illustration of the loss for each method on the Synapse dataset. Since the DICE score of the TransCASCADE method is similar to that of the DS-TransUNet method, only TransCASCADE is included; similarly, since the performance of the TransUNet method is comparable to that of the Swin-UNet method, TransUNet is selected; among the PVT-CASCADE and VM-UNet methods, VM-UNet is chosen; and among the DA-TransUNet and Mamba-UNet methods, Mamba-UNet is selected. It can be observed that the initial loss of the MSM-UNet method is the lowest, and as the epochs accumulate, the loss value of MSM-UNet remains relatively stable and reaches its minimum around epoch = 115.

As shown in Fig. 7, the performance comparison of the proposed MSM-UNet method against other methods, showcasing MSM-UNet's significant advantage in handling details. This advantage primarily stems from MSM-UNet's ability to effectively combine long-range and short-range dependencies. When tackling tasks involving segmentation gaps, MSM-UNet can clearly distinguish between the regions that need to be segmented and the gap regions, while other methods partially fill in the gaps due to their excessive focus on global features or their inability to effectively combine local and global features. Additionally, MSM-UNet can accurately segment small target points at the edges. Furthermore, by utilizing the WTFEAB module to exclude irrelevant information, MSM-UNet can precisely locate key features, thereby avoiding segmentation results that exceed the true boundaries.

Table 8
presents the segmentation results on the Synapse dataset. Only the DICE scores for individual organs are reported. ↑ indicates that higher values are better, while ↓ indicates that lower values are better. All MSM-UNet results are averaged over 5 runs. The best results are bolded.

Architectures	Average				Aorta	GB	KL	KR	Liver	PC	SP	SM
	DICE↑	HD95↓	mIoU↑	ASD↓								
UNet	70.11	44.69	59.35	14.41	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96
ViT	71.29	32.87	62.25	9.86	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUNet	77.61	26.9	67.32	4.66	86.56	60.43	80.54	78.53	94.33	58.47	87.06	75
Swin-UNet	77.58	27.32	66.88	4.71	81.76	65.95	82.32	79.22	93.73	53.81	88.04	75.79
DA-TransUNet	80.90	21.15	68.72	4.14	87.12	69.48	85.34	80.42	93.23	62.69	89.70	79.20
DS-TransUNet	82.39	17.65	71.56	3.94	88.26	71.32	86.38	82.05	94.98	65.36	**91.44**	79.29
Mamba-UNet	80.58	21.95	68.47	4.07	87.23	68.25	84.66	80.41	94.03	58.92	90.12	81.05
VM-UNet	81.08	19.21	70.74	3.58	86.40	69.41	86.16	82.76	94.17	58.80	89.51	81.40
PVT-CASCADE	81.06	20.23	70.88	3.61	83.01	70.59	82.23	80.37	94.08	64.43	90.1	**83.69**
TransCASCADE	82.68	17.34	73.48	2.83	86.63	68.48	**87.66**	84.56	94.43	65.33	90.79	83.52
MSM-UNet(ours)	**83.50**	**15.24**	**74.70**	**2.58**	**88.96**	**71.33**	87.30	**85.65**	**95.03**	**70.43**	90.64	78.55

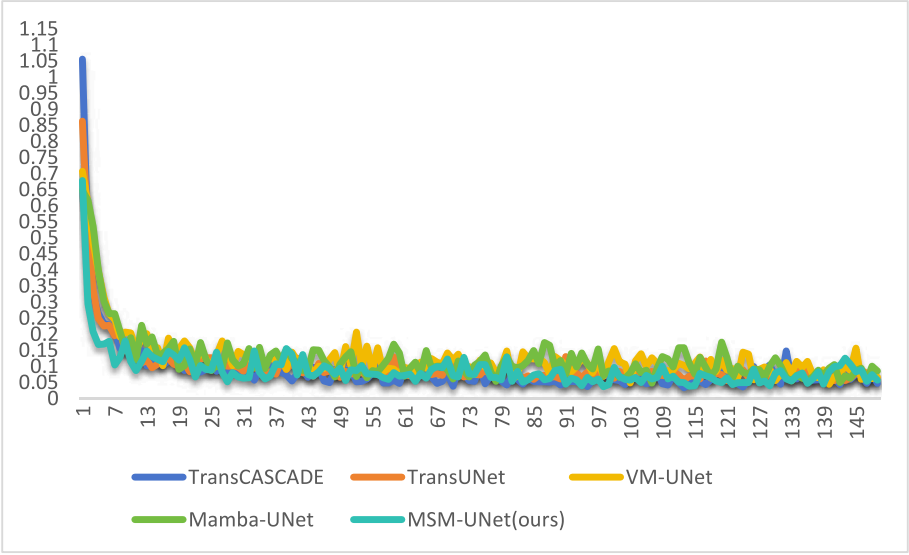


Fig. 6. Loss diagram for the Synapse dataset.

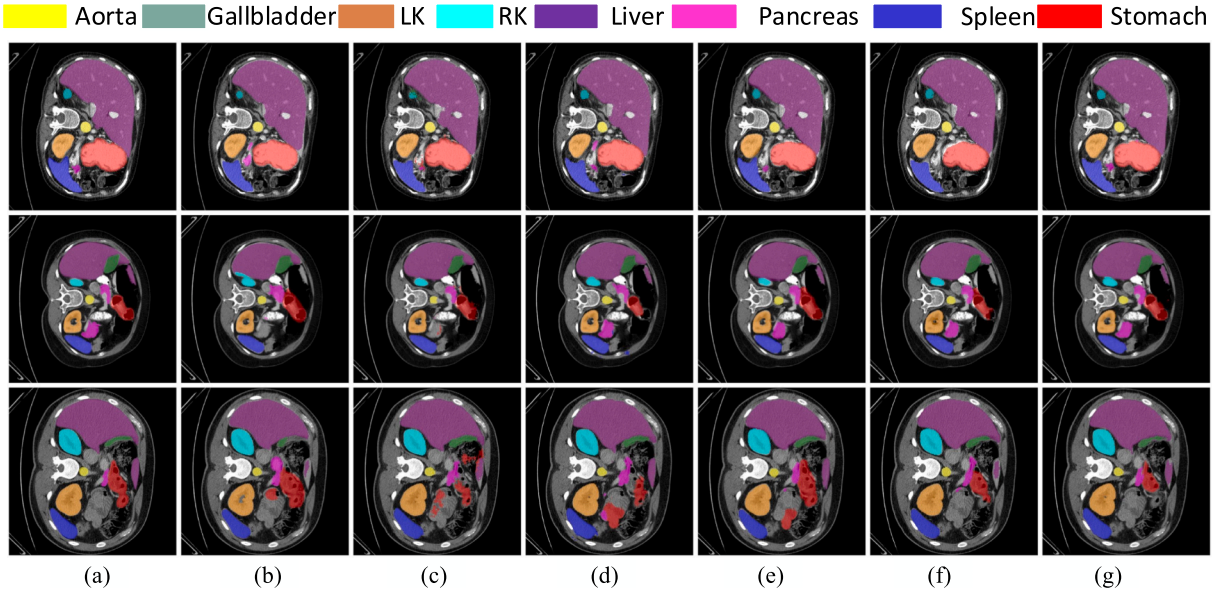


Fig. 7. Results on the Synapse dataset, (a) GT, (b) TransUNet, (c) Mamba-UNet, (d) VM-UNet, (e) DA-TransUNet, (f) TransCASCADE, and (g) MSM-UNet.

Table 9
Experimental results on the ACDC dataset. The DICE scores for various organs are presented, with ↑ indicating that higher values are better. Other comparison results are taken from the literature, and all MSM-UNet results are averaged over 5 runs. The best results are bolded.

Architectures	Avg DICE↑	RV	Myo	LV
ViT	81.45	81.46	70.71	92.18
R50-U-Net	87.55	87.10	80.63	94.92
TransUNet	89.71	86.67	87.27	95.18
Swin-UNet	88.07	85.77	84.42	94.03
DA-TransUNet	90.57	90.85	85.74	95.13
DS-TransUNet	91.01	91.21	86.38	95.44
Mamba-UNet	90.58	87.84	88.55	95.35
VM-UNet	90.81	88.96	88.22	95.26
PVT-CASCADE	91.46	88.9	89.97	95.50
TransCASCADE	91.63	89.14	90.25	95.50
MSM-UNet (ours)	**92.02**	**90.03**	**90.36**	**95.67**

3.5.2. Results on the ACDC dataset

Table 9 presents the average DICE scores of our MSM-UNet architecture, along with other SOTA methods such as TransCASCADE. The MSM-UNet architecture outperforms the best-performing TransCASCADE by 0.39 % in terms of the DICE metric, and it surpasses TransUNet by 2.31 % in DICE. Compared to Mamba-UNet, MSM-UNet achieves a 1.44 % improvement in DICE, and it outperforms VM-UNet by 1.21 % in DICE. Additionally, MSM-UNet improves upon DA-TransUNet and DS-TransUNet by 1.45 % and 1.01 % respectively in DICE. Furthermore, MSM-UNet achieves optimal results across various organ metrics.

As shown in Fig. 8, in experiments similar to those conducted on the Synapse dataset, we selected several methods with comparable DICE scores for a comparative analysis of their loss functions. The observation results indicate that the loss value of the MSM-UNet method decreases the fastest and quickly stabilizes. Although the Mamba-UNet and VM-UNet methods ultimately achieve lower loss values during training, their performance in practical applications is less satisfactory.

As shown in Fig. 9, we compared the MSM-UNet architecture with

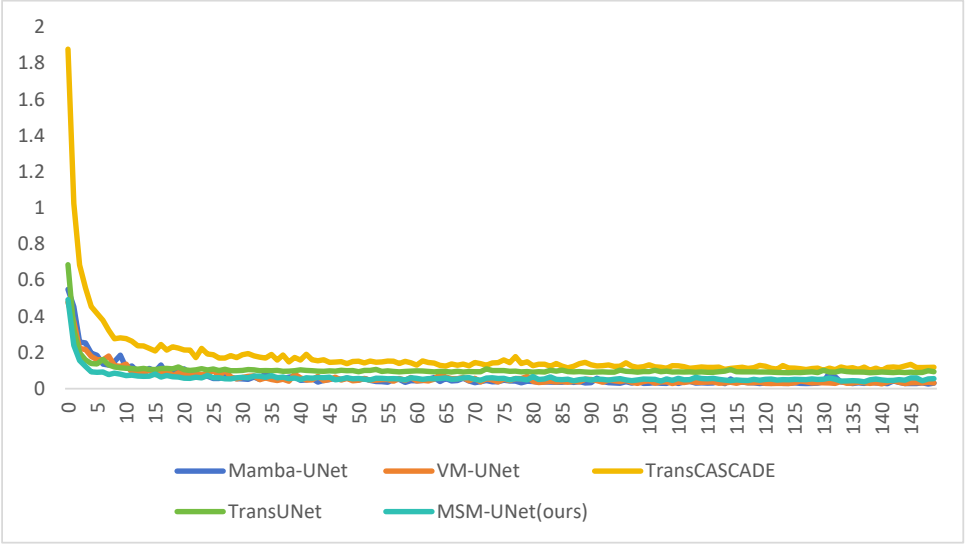


Fig. 8. Loss diagram for the ACDC dataset.

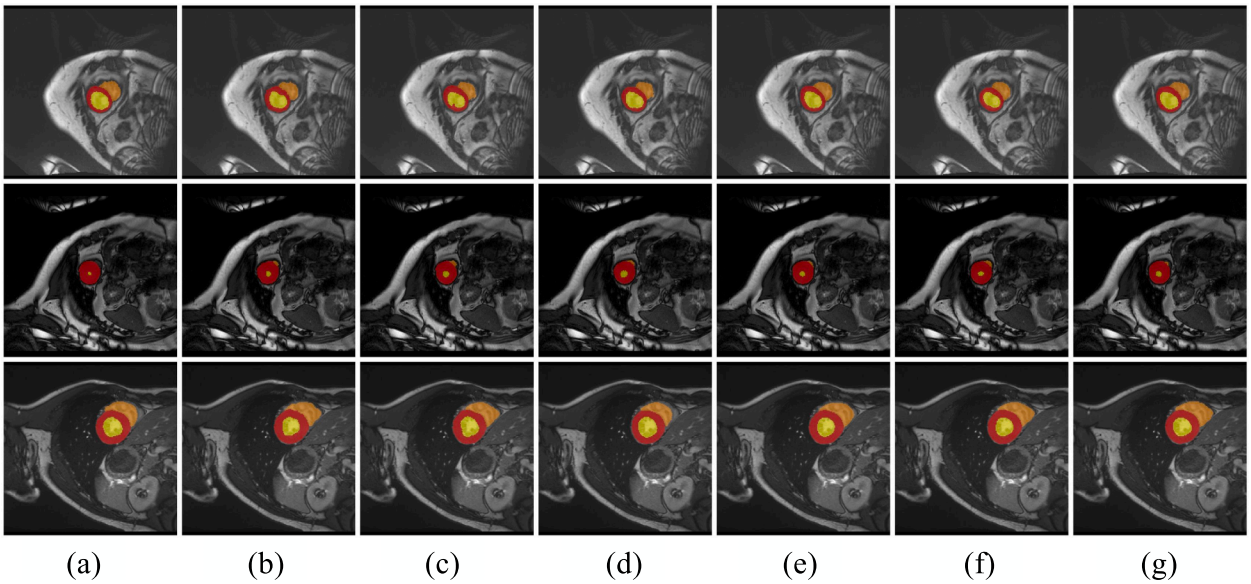


Fig. 9. Visualization of various methods on the ACDC dataset, (a) GT, (b) TransUNet, (c) Mamba-UNet, (d) VM-UNet, (e) DS-TransUNet, (f) TransCASCADE, and (g) MSM-UNet.

Table 10
Experimental results on the CVC-ClinicDB dataset, showing DICE and mIoU scores, with ↑ indicating that higher values are better. The best results are bolded.

Architectures	DICE↑	mIoU↑
UNet	86.93	78.21
TransUNet	89.01	81.63
DA-TransUNet	89.47	82.51
DS-TransUNet	93.5	88.45
VM-UNet	88.35	82.16
Mamba-UNet	91.1	83.64
PVT-CASCADE	**94.34**	**89.98**
MSM-UNet(ours)	94.03	89.35

the other mentioned methods. The results demonstrate that MSM-UNet, with its unique design, exhibits significant advantages in image segmentation tasks. Specifically, by fusing long-range and short-range dependency information, MSM-UNet effectively captures both global and local features in the image, making it more accurate in identifying key features. Additionally, by introducing the FEMB module and FOL module, MSM-UNet can precisely focus on the critical regions of different categories in the image, effectively eliminating the interference from non-focus points of different categories. Therefore, during the image segmentation process, MSM-UNet can accurately identify and locate the target regions, avoiding segmentation results that exceed the true boundary, thereby significantly improving the accuracy and reliability of the segmentation.

3.5.3. *Results on the CVC-ClinicDB dataset*

Table 10 presents the comparative data of the MSM-UNet method along with other methods. It can be observed that the MSM-UNet method, while slightly inferior to the SOTA method PVT-CASCADE, demonstrates notable improvements over other approaches. Specifically, compared to TransUNet, MSM-UNet achieves enhancements of 5.02 % and 7.72 % in DICE and mIoU scores, respectively. When compared to DS-TransUNet, MSM-UNet exhibits improvements of 0.53 % and 0.9 % in DICE and mIoU scores, respectively. Furthermore, MSM-UNet outperforms Mamba-UNet by 2.93 % and 5.71 % in DICE and

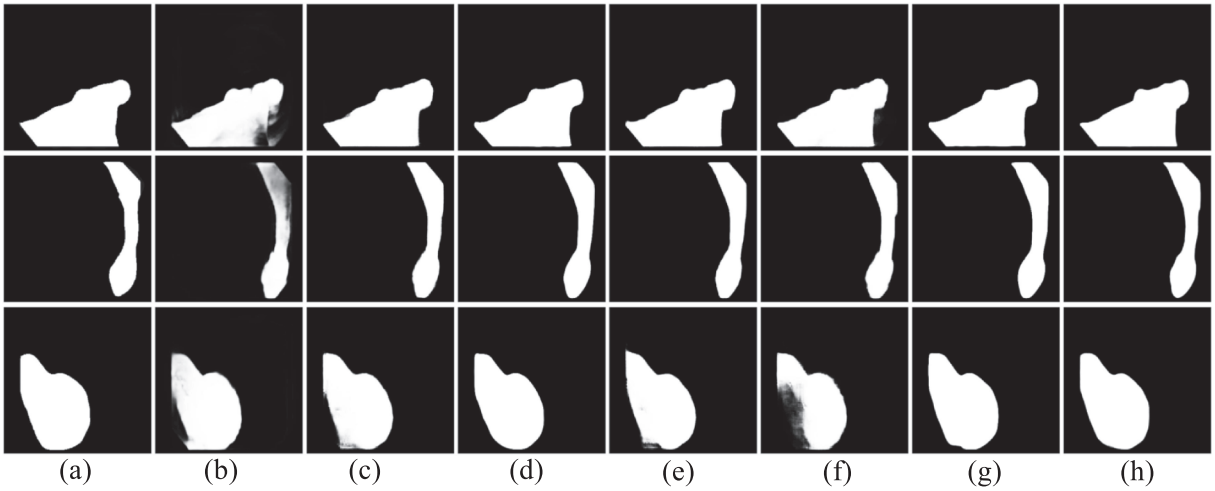


Fig. 10. Visualization of various methods on the CVC-ClinicDB dataset, (a) GT, (b) UNet, (c) TransUNet, (d) DS-TransUNet, (e) Mamba-UNet, (f) VM-UNet, (g) PVT-CASCADE, and (h) MSM-UNet.

mIoU scores, respectively. Overall, the performance of MSM-UNet is quite satisfactory.

As shown in Fig. 10, we conducted a comprehensive and in-depth comparative analysis of the MSM-UNet architecture against a series of mentioned advanced methods. The research results indicate that MSM-UNet, with its innovative and unique design concept, exhibits particularly prominent performance advantages in the complex task of image segmentation. Especially when dealing with the critical aspect of image boundaries, the architecture demonstrates remarkable accuracy and fineness, reflecting its great superiority in capturing image details and structural information. This finding not only confirms the effectiveness of the MSM-UNet architecture but also provides new perspectives and insights for research in the field of image segmentation.

4. Conclusion

This paper proposes an enhanced variant of the Mamba architecture, named MSMamba. The innovation of this architecture lies in the ingenious integration of the multi-head mechanism thought into the Mamba structure, aiming to enhance its ability to extract local features. By integrating MSMamba into the CNNs framework, we effectively compensate for the limitations of CNNs in capturing global features. Additionally, this study meticulously designs the Wavelet Transform Feature Enhancement Attention Block (WTFEAB) and a set of feature enhancement and fusion mechanisms, including the Feature Enhancement Merge Block (FEMB) and Fusion Output Layer (FOL) strategy. Specifically, the WTFEAB module performs feature enhancement operations in both channel and spatial dimensions to improve feature representation capabilities; the FEMB module conducts feature fusion during the up-sampling process to further enhance feature representation; and the FOL strategy performs feature fusion at the output stage of the decoder to optimize the final feature expression. Through these modules' gradual feature enhancement and multi-level fusion strategies, this study effectively addresses the challenge of segmentation results being prone to misjudgment due to the diversity of organ morphologies. Experimental results show that the MSMamba architecture achieves a DICE coefficient of 83.10 on the Synapse dataset, a DICE coefficient of 92.02 on the ACDC dataset, and a relatively good DICE coefficient of 94.03 on the CVC-ClinicDB.

5. Consent for publication

All authors have approved the manuscript for submission and agree to the publication of this work. Written informed consent for publication was obtained from all individuals involved in the study, where necessary.

CRediT authorship contribution statement

Junding Sun: Conceptualization, Formal analysis, Methodology, Writing – original draft. **Kaixin Chen:** Investigation, Software, Resources, Validation, Writing – original draft. **Xiaosheng Wu:** Project administration, Writing – review & editing. **Zhaozhao Xu:** Data curation, Funding acquisition, Validation. **Shuihua Wang:** Software, Supervision, Visualization, Writing – original draft. **Yudong Zhang:** Data curation, Validation, Supervision, Writing – original draft.

Funding

This work is supported by the National Natural Science Foundation of China (62276092,62303167); the Postdoctoral Fellowship Program (Grade C) of China Postdoctoral Science Foundation under Grant Number(GZC20230707); MRC (MC_PC_17171); Royal Society (RP202G0230); BHF (AA/18/3/34220); the Key Science and Technology Program of Henan Province, China (242102211051); the Young Elite Scientists Sponsorship Program by Henan Association for Science and Technology (2025HYTP061); the Key Scientific Research Projects of Colleges and Universities in Henan Province, China (25A520009); the China Postdoctoral Science Foundation (2024M760808), the Henan Province medical science and technology research plan joint construction project (LHGJ2024069).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015 (MICCAI)*, Munich, Germany, 234-241. https://doi.org/10.48550/arXiv.1505.04597.

Siddique, N., Paheding, S., Elkin, C. P., & Devabhaktuni, V. (2021). U-Net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access, 9*, 82031–82057. https://doi.org/10.1109/ACCESS.2021.3086020

Zhou, Z. W., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. M. (2019). UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging, 39*(6), 1856–1867. https://doi.org/10.1109/TMI.2019.2959609

Huang, H. M., Lin, L. F., Tong, R. F., Hu, H. J., Zhang, Q. W., Iwamoto, Y., Han, X. H., Chen, Y. W., Wu, J. (2020). UNet 3+: A full-scale connected unet for medical image segmentation. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 1055-1059. https://doi.org/10.1109/ICASSP40776.2020.9053405.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. 2818-2826.

Diakogiannis, F. I., Waldner, F., Caccetta, P., & Wu, C. (2020). ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing, 162*, 94–114. https://doi.org/10.1016/J.ISPRSJPRS.2020.01.013

Maji, D., Sigedar, P., & Singh, M. (2022). Attention Res-UNet with guided decoder for semantic segmentation of brain tumors. *Biomedical Signal Processing and Control, 71*, Article 103077. https://doi.org/10.1016/j.bspc.2021.103077

Li, R., Zheng, S., Duan, C., Su, J., & Zhang, C. (2021). Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters, 19*, 1–5. https://doi.org/10.48550/arXiv.2011.14302

Li, X. M., Chen, H., Qi, X. J., Dou, Q., Fu, C. W., & Heng, P. A. (2018). H-DenseUNet: Hybrid densely connected unet for liver and tumor segmentation from CT volumes. *IEEE Transactions on Medical Imaging, 37*(12), 2663–2674. https://doi.org/10.1109/TMI.2018.2845918

Safarov, S., & Whangbo, T. K. (2021). A-DenseUNet: Adaptive densely connected unet for polyp segmentation in colonoscopy images with atrous convolution. *Sensors, 21*(4), 1441. https://doi.org/10.3390/s21041441

Cai, S. J., Tian, Y. X., Lui, H., Zeng, H. S., Wu, Y., & Chen, G. N. (2020). Dense-UNet: A novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. *Quantitative Imaging in Medicine and Surgery, 10*(6), 1275. https://doi.org/10.21037/qims-19-1090

Guan, S., Khan, A. A., Sikdar, S., & Chitnis, P. V. (2019). Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal. *IEEE Journal of Biomedical and Health Informatics, 24*(2), 568–576. https://doi.org/10.1109/JBHI.2019.2912935

Khan, T. M., Naqvi, S. S., & Meijering, E. (2024). ESDMR-Net: A lightweight network with expand-squeeze and dual multiscale residual connections for medical image segmentation. *Engineering Applications of Artificial Intelligence, 133*(PartA), 14. https://doi.org/10.1016/j.engappai.2024.107995

Song, H., Wang, Y., Zeng, S., Guo, X., & Li, Z. (2023). OAU-Net: Outlined Attention U-net for biomedical image segmentation. *Biomedical Signal Processing and Control, 79*, Article 104038. https://doi.org/10.1016/j.bspc.2022.104038

Pang, B., Chen, L., Tao, Q., & Wang, E. (2024). GA-UNet: A lightweight ghost and attention u-net for medical image segmentation. *Journal of Imaging Informatics in Medicine, 37*(4), 1874–1888. https://doi.org/10.1007/s10278-024-01070-5

Ling, Y., Wang, Y., & Kong, L. (2024). MTANet: Multi-task attention network for automatic medical image segmentation and classification. *IEEE Transactions on Medical Imaging, 43*(2), 674–685. https://doi.org/10.1109/TMI.2023.3317088

Taher, S. M., Ghanim, M., & Der, C. S. (2023). Applied improved canny edge detection for diagnosis medical images of human brain tumors. *Al-Mustansiriyah Journal of Science, 34*(4), 66–74. https://doi.org/10.23851/mjs.v34i4.1392

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (p. 30).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X. H., & Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR), Virtual Event, Austria. https://doi.org/10.48550/arXiv.2010.11929.

Liu, Z., Lin, Y. T., Cao, Y., Hu, H., Wei, Y. X., Zhang, Z., Lin, S., & Guo, B. N. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00987.

He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., & Xue, Y. (2022). Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1–15.

Sun, W. W., Chen, J. G., Yan, L., Lin, J. Z., Pang, Y., & Zhang, G. (2022). COVID-19 CT image segmentation method based on swin transformer. *Frontiers in Physiology, 13*, Article 981463. https://doi.org/10.3389/fphys.2022.981463

Zhang, C., Wan, H. C., Shen, X. Y., & Wu, Z. Z. (2022). PVT: Point-voxel transformer for point cloud learning. *International Journal of Intelligent Systems, 37*(12), 11985–12008. https://doi.org/10.1002/int.23073

Wang, W. H., Xie, E., Li, X., Fan, D. P., Song, K. T., Liang, D., Lu, T., Luo, P., & Shao, L. (2022). Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media, 8*(3), 415–424. https://doi.org/10.1007/s41095-022-0274-8

Azad, R., Kazerouni, A., Sulaiman, A., Bozorgpour, A., Aghdam, E. K., Jose, A., & Merhof, D. (2023). Unlocking fine-grained details with wavelet-based high-frequency enhancement in Transformer. In *International Workshop on Machine Learning in Medical Imaging (MLMI)* (pp. 207–216). Canada: Vancouver.

Dihin, R. A., AlShemmary, E. N., & Al-Jawher, W. A. (2024). Wavelet-attention swin for automatic diabetic retinopathy classification. *Baghdad Science Journal, 21*(8), 2741. https://doi.org/10.21123/bsj.2024.8565

Chen, J. N., Lu, Y. Y., Yu, Q. H., Luo, X. D., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. Y. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. arxiv preprint arxiv:2102.04306. https://doi.org/10.48550/arXiv.2102.04306.

Gao, L., Liu, H., Yang, M. H., Chen, L., Wan, Y. L., Xiao, Z. Q., & Qian, Y. R. (2021). STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS), 14*, 10990–11003. https://doi.org/10.1109/JSTARS.2021.3119654

Pham, T. H., Li, X. Q., & Nguyen, K. D. (2024). Seunet-Trans: A simple yet effective unet-transformer model for medical image segmentation. *IEEE Access, 12*, 122139–122154. https://doi.org/10.1109/ACCESS.2024.3451304

Sun, G. Q., Pan, Y. Z., Kong, W. K., Xu, Z. C., Ma, J. H., Racharak, T., Nguyen, L. M., & Xin, J. Y. (2024). DA-TransUNet: Integrating spatial and channel dual attention with transformer u-net for medical image segmentation. *Frontiers in Bioengineering and Biotechnology, 12*, Article 1398237. https://doi.org/10.3389/fbioe.2024.1398237

Cao, H., Wang, Y. Y., Chen, J., Jiang, D. S., Zhang, X. P., Tian, Q., & Wang, M. N. (2022). Swin-UNet: Unet-like pure transformer for medical image segmentation. In *In: European Conference on Computer Vision (ECCV), Tel Aviv* (pp. 205–218). https://doi.org/10.1007/978-3-031-25066-8_9

Pan, S. M., Liu, X., Xie, N. D., & Chong, Y. W. (2023). EG-TransUNet: A transformer-based u-net with enhanced and guided models for biomedical image segmentation. *BMC Bioinformatics, 24*(1), 85. https://doi.org/10.1186/s12859-023-05196-1

Lin, A. L., Chen, B. Z., Xu, J. Y., Zhang, Z., Lu, G. M., & Zhang, D. (2022). DS-TransUNet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement, 71*, 1–15. https://doi.org/10.1109/TIM.2022.3178991

Rahman, M. M., & Marculescu, R. (2023). Medical image segmentation via cascaded attention decoding. In *In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 6222–6231). https://doi.org/10.1109/WACV56688.2023.00616

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arxiv preprint arxiv:2312.00752. https://doi.org/10.48550/arXiv.2312.00752.

Zhu, L. H., Liao, B. C., Zhang, Q., Wang, X. L., Liu, W. Y., & Wang, X. G. (2024). Vision Mamba: Efficient visual representation learning with bidirectional state space model. In: Proceedings of the 41st International Conference on Machine Learning (ICML 2024), Vienna, Austria. arXiv:2401.09417. https://doi.org/10.48550/arXiv.2401.09417.

Zhang, M. Y., Yu, Y., Jin, S., Gu, L. M., Ling, T. S., & Tao, X. P. (2024). VM-UNET-V2: Rethinking vision mamba unet for medical image segmentation. In: Proceedings of the 20th International Symposium on Bioinformatics Research and Applications (ISBRA 2024), Kunming, China, 335-346. https://doi.org/10.48550/arXiv.2403.09157.

Wang, Z. Y., Zheng, J. Q., Zhang, Y. C., Cui, G., & Li, L. (2024). Mamba-UNet: Unet-like pure visual mamba for medical image segmentation. arxiv preprint arxiv:2402.05079. https://doi.org/10.48550/arXiv.2402.05079.

Zhang, X. X., & Mu, W. S. (2024). Gmamba: State space model with convolution for grape leaf disease segmentation. *Computers and Electronics in Agriculture, 225*, Article 109290. https://doi.org/10.1016/j.compag.2024.109290

Zou, B. F., Huang, X. R., Jiang, Y. T., Jin, K., & Sun, Y. Q. (2024). Demambanet: Deformable convolution and mamba integration network for high-precision segmentation of ambiguously defined dental radicular boundaries. *Sensors, 24*(14), 4748. https://doi.org/10.3390/s24144748

Hu, M., Zhang, Y. R., Xue, H. J., Lv, H., & Han, S. P. (2024). Mamba-and ResNet-based dual-branch network for ultrasound thyroid nodule segmentation. *Bioengineering, 11*(10), 1047. https://doi.org/10.3390/bioengineering11101047

Li, G. J., Huang, Q. H., Wang, W., & Liu, L. Z. (2025). Selective and multi-scale fusion mamba for medical image segmentation. *Expert Systems With Applications, 261*, Article 125518. https://doi.org/10.1016/j.eswa.2024.125518