YNet: medical image segmentation model based on wavelet transform boundary enhancement

Wenzhe Meng
School of Software, Xinjiang University
Urumqi 830091, China
Meng_wenzhe@163.com

Xiaoliang Zhu(✉)
School of Software, Xinjiang University
Urumqi 830091, China
zhuxiaoliang3721@163.com

Yanxiang Li
School of Software, Xinjiang University
Urumqi 830091, China
xju_liyanxiang@163.com

Abstract

Medical image segmentation is critical in understanding pathological changes and computer-aided diagnosis. Most of the existing medical segmentation models focus on the overall segmentation effect of the model and lack of thinking about the problems of boundary blurring and model generalization ability. Given this, a bipartite segmentation model YNet based on boundary enhancement is proposed, which is based on the encoder-decoder architecture and consists of two core components: boundary enhancement module (BEM) and feature fusion module (FFM). The BEM utilizes the wavelet transform to separate the frequency domain information from the original image, allowing the model to dynamically adjust the boundary details in the original image dynamically, thus enhancing the model's ability to perceive the boundary information and mitigating the effect of noise on the model. An attention mechanism is introduced in the FFM to enhance the model's generalization ability by dynamically adjusting the channel and spatial information weights to emphasize critical features and suppress redundant information. Experimental results comparing other methods on CVC-ClinicDB, Kvasir-SEG, DSB2018, and ISIC2018 datasets show that the model has more explicit boundaries and better segmentation generalization. The source code of our YNet will be mdae available at https://github.com/DeadlyCodeGod/YNet.

***Keywords:** Double-branch networks, Wavelet transform, Feature fusion, Attention mechanism, Boundary enhancement.*

1. Introduction

Medical image segmentation plays a vital role in medical image processing, where the main goal is to classify each pixel in an image and generate a masked image to identify lesion areas in medical images [24] accurately. This technique is crucial in assisting physicians to more accurately diagnose diseases, develop treatment plans, and track the health status of patients. Researchers have been developing convolutional neural network (CNN) based medical image segmentation methods to achieve this purpose and have made significant progress [9, 28, 17]. The most representative method is UNet [24], which improves the segmentation of edge details by combining the encoder's shallow features with the decoder's high-resolution features through hopping connections. Okatay [21] et al. proposed an attentional mechanism for the UNet-based segmentation model, which can better focus on the relevant regions of the pancreas. Zhou [35] et al. proposed a variant of UNet called UNet++, in which the decoder sub-network solves the sizeable semantic gap between the encoder and decoder in UNet through dense hopping connections. On this basis, Srivastava et al. [27] proposed a multi-scale feature extraction and fusion mechanism for the feature extraction and fusion problem in medical image segmentation, effectively utilising different levels of features to capture richer contextual information. ColonSegNet [10] employs multi-scale feature extraction to capture different scales of contextual information. At the same time, an attention mechanism is introduced to enhance the attention to critical regions. Although these research results have achieved significant improvements in segmentation performance, there are still two key issues that need to be addressed in medical image segmentation tasks: (1) Insufficient utilization of original image information: most models tend to focus on

optimizing the design of the framework while ignoring the rich available information in the original image, such as texture and structure information. (2) Challenges of segmentation boundaries: Due to the varying sizes of lesions, their boundaries with surrounding tissues often need to be more explicit, resulting in segmentation models that are prone to under-segmentation or over-segmentation, which affects the reliability of the models in clinical applications.

Aiming at the above problems, this paper proposes a wavelet transform-based boundary information enhancement module, which can be enough to separate the high and low-frequency information of the original image without additional training data and which helps to help the model filter noise. In the frequency domain, low-frequency (LF) information expresses the abstract semantics of the image, while high-frequency (HF) information is rich in detailed features of the image boundary [32, 29]. This enables YNet to learn tiny structures, abstract semantic content, and overlapping or low-light parts of an image, and these details and semantic features are crucial for medical image analysis and diagnosis [3, 4, 23]. However, single-branch convolutional neural networks have limitations in dealing with these details and need help thoroughly learning the information they contain [25, 15, 36]. So, inspired by the above studies and limitations, this paper proposes a double-branch segmentation model YNet driven by LF and HF information for medical image segmentation tasks. The dual-branch encoder in the model learns the image features enhanced by LF and HF, respectively, so that the YNet model can capture the correlation of potential inter-split features and enhance the ability of the network to link semantic and detailed features. In order to overcome the problem that the fusion of the frequency domain information of the dual-branch taps may lead to the weakening of the model's generalization ability, this paper fuses the LF and HF information with the original information separately in an adaptive manner and designs the attention feature fusion module.

The contributions of this article are described below:

(1) This paper uses the wavelet transform theory to design the BEM. This module separates the LF and HF information of an image and weighted fusion of the separated information with the original image in a self-learning manner. This form of adaptive feature fusion enhances the model's generalisation ability and reduces the need for model parameter tuning.

(2) The double-branch model named YNet is proposed. It learns the original image information enhanced by global information on the LF encoder branch and learns to capture fine edge details from the original image on the HF encoder branch. This complementary learning approach enhances the model's ability to perceive boundary information.

(3) In order to better fuse the complementary information of the double-branch taps, this paper designs the FFM, which suppresses or enhances the feature maps in an attentional manner on the channel as well as spatially, to improve the module's ability to perceive and select important feature information.

2. Related work

In recent years, thanks to the advancement of deep learning technology, the field of medical images has been developing rapidly, and numerous models with excellent performance have emerged [28, 21, 11]. However, when dealing with complex scenes, single-branch networks usually can only process features on one path and are prone to encounter bottlenecks in feature fusion at different scales and model generalization ability. In contrast, dual-branch networks perform well in this scenario [11]. The double-branch model architecture improves multi-scale feature learning. It enhances model generalization capability by designing independent branching paths that can process different kinds of features or learn different aspects of features separately [34]. In addition, this architecture allows the model to dynamically adapt to different feature requirements and effectively integrate multi-scale information when facing complex segmentation tasks, thus demonstrating enhanced performance capabilities when dealing with complex scenes. However, learning the complementary information between bipartition splits recognizes a vital issue in the design of bipartition-based models, which should be considered for the original feature processing in addition to considering the model coding layer.

An essential advantage of the wavelet transform is its ability to separate the high and low information and retain the frequency range and spatial location information effectively at the same time, which makes the wavelet transform uniquely valuable in image processing, especially in tasks that need to distinguish between image details and global structure [28, 26]. Therefore, the combination of wavelet transform and deep learning models can help the models to improve segmentation accuracy when processing complex images, especially in tasks that require precise boundary detection [8, 19]. Azimi et al. [1] proposed a symmetric CNN algorithm augmented by wavelet transform to effectively solve the problem of improving segmentation accuracy in semantic segmentation tasks. The algorithm better preserves boundary details and complex structures by introducing wavelet transform. However, the method needs more flexibility when dealing with different tasks. Its performance needs to be more robust for tasks with high complexity or significant differences in feature distribution, and its generalization ability needs to be improved. Duan et al. [6] used the wavelet transform constrained pooling layer to replace the traditional maximum pooling or average pooling operation to solve the problem of better retaining detailed information. However, it is more sensitive to HF noise,

and the wavelet transform pooling operation is more significant in computational volume, which increases the network's computational complexity and training time. Li et al. [18] enhance the network's ability to capture and reconstruct detailed information by using the wavelet transform to extract the details in the downsampling stage and the wavelet inverse transforms to recover the details in the upsampling stage. However, this method imposes strong constraints on the network structure while enhancing image details, resulting in a less scalable model across tasks and datasets, exhibits instability, and is challenging to adapt to diverse application scenarios. In contrast, YNet has a more flexible operation by focusing on LF and HF information through a double-branch architecture without imposing constraints on the network.

3. Method

3.1. Overview

The YNet network proposed in this paper adopts the traditional encoder-decoder architecture, and the specific network architecture is illustrated in Fig. 1, which consists of five parts: (1) the HF and LF information separation module BEM, which helps the model to better learn the essential features in the subsequent processing steps; (2) two fully convolutional double-branch network encoders, which are able to extract different kinds of features at the same time, thus improving the multi-scale feature learning capability; (3) a feature fusion module FFM, which enhances the model's understanding of both detailed and global information and improves the accuracy of the segmentation results; (4) a profoundly supervised feature transformation module TM, which ensures that the model's learning is effectively guided at different levels, and improves the model's stability during the training process [16]; (5) a decoder for generating segmentation predictions, which partially ensures the meticulousness of the segmentation results by gradually restoring the spatial resolution of the image.

Overall, the original image is first processed by BEM to extract the LF and HF feature information. Then, this feature information is adaptively fused with the original image to enhance the edge and semantic features of the original image, respectively, to form the feature maps LF and HF. The enhanced feature maps LF and HF are sent to the dual-branch encoder for encoding operations, as illustrated in Fig. 1, where the LF information is learned in the LF encoder (LE). In contrast, the HF information is learned in the HF encoder (HE). The LE branch extracts features with larger sensory fields through the hollow convolution operation to extract features with larger receptive fields to better understand the overall layout of the image.HE, on the other hand, enhances the learning of details through small convolution kernels to more accurately capture the image's boundaries and small lesion regions. The FFM collects LE and HE feature maps containing rich semantic and boundary information. Then, through the attention mechanism in the FFM, the module dynamically adjusts these features to highlight those that are more important for the segmentation task, enhancing the decoder's focus on task-critical regions, which effectively suppresses redundant and irrelevant information and ensures that the model pays more attention to critical features when making decisions. During the decoding process, the fused features of each layer form jump connections with the corresponding decoder layer, which generate the final segmentation prediction. To further enhance the network's ability to understand deep information, this paper performs depth supervision at two random layers of the decoder, realizes feature conversion and dimension matching through 1×1 convolution operation to simplify the model complexity, and offsets the model decoding heavy burden to deeper layers to ensure that the coherence of information transfer is maintained in the multilevel feature maps, and enhances the model's performance in the segmentation task. In order to ensure that the decoding process has stronger feature representation ability at different scales, this paper designs a combined loss function, denoted as L_{total}, which combines the losses from three different feature layers, namely: the network output loss L_{net}, the loss at the fourth decoder layer L_1 and the loss at the second decoder layer L_2, and combines them to generate the final loss L_{total}. Its mathematical expressions are as in Eqs. (1) and (2):

$$L_{\text{total}} = L_{\text{net}} + \lambda(L_1 + L_2) \tag{1}$$

$$L_{\text{net}} = L_1 = L_2 = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right] \tag{2}$$

where λ is the balance parameter between the network output loss and the middle layer loss. The cross-entropy loss is used for L_{net}, L_1 and L_2. N is the number of samples, y_i is the actual label (0 or 1) of the ith sample, and p_i is the probability that the model predicts a positive class for the ith sample.

3.2. Boundary Enhancement Module

Fig.2 demonstrates the basic framework of the BEM proposed in this paper. The wavelet transform usually decomposes the image into the following four parts: the LF component LL retains the LF information of the original image; the vertical HF component LH retains the HF information of the image in the vertical direction; the horizontal HF component HL retains the HF information of the image in the horizontal direction; and the diagonal HF component HH retains the HF information of the image in the diagonal. In this paper, these three HF components are combined as the sum of all HF components. This treatment can simplify the

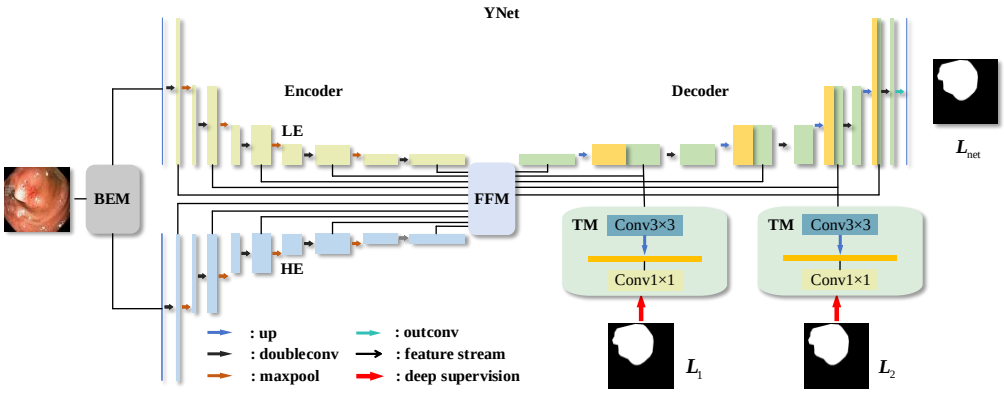

Figure 1. The model framework diagram of YNet shows that BEM is a boundary enhancement module that separates images' LF and HF information. LE and HE are LF information encoders and HF information encoders, respectively. FFM is a Feature Fusion Module that fuses the features learned from the double-branch encoder. The TM Module is the Deep-Supervised Transformation Module.

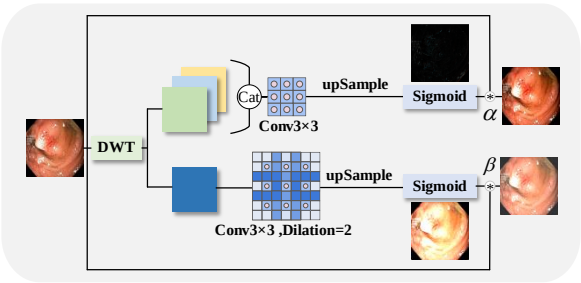

Figure 2. Module framework diagram of BEM. BEM uses different convolution operations to extract HF and LF information selectively. In order to ensure effective fusion with the original features, the upSample operation is used to recover the feature dimension size. α and β represent the gating factors that fuse the HF and LF features with the original images.

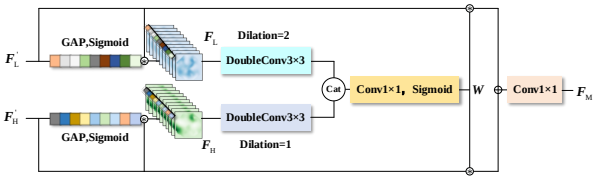

Figure 3. The FFM's structural diagram shows the fusion process of HF and LF features.

model while still retaining enough HF detail information. This paper uses $\mathbf{L}$ to denote the LF component and $\mathbf{H}$ to denote the sum of the three HF components in different directions. Moreover, the upSample operation is utilized to recover the problem of space size reduction brought about by the wavelet transform. The specific $\mathbf{L}$ and $\mathbf{H}$ definitions

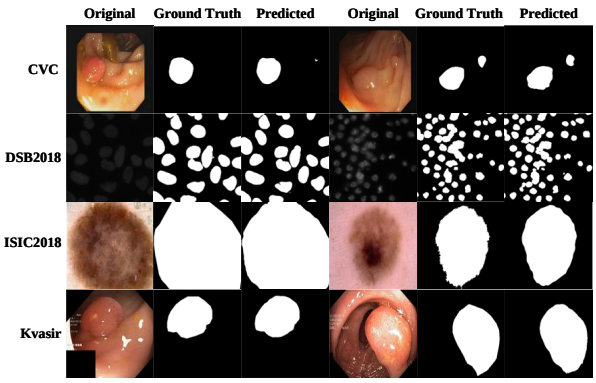

Figure 4. Experimental predicted Kvasir, ISIC2018, DSB2018, and CVC maps against the actual values. Where original denotes the Original image, Ground Truth denotes the actual value, and Predicted denotes the prediction of YNet.

are shown in Eqs. (3), (4) and (5):

$$\mathbf{F}_{\mathrm{LL}}, \mathbf{F}_{\mathrm{LH}}, \mathbf{F}_{\mathrm{HL}}, \mathbf{F}_{\mathrm{HH}} = \mathrm{BEM}(\mathbf{F}_{\mathrm{input}}) \tag{3}$$

$$\mathbf{L} = \mathrm{upSample}(\mathbf{F}_{\mathrm{LL}}) \tag{4}$$

$$\mathbf{H} = \mathrm{upSample}(\mathbf{F}_{\mathrm{LH}} + \mathbf{F}_{\mathrm{HL}} + \mathbf{F}_{\mathrm{HH}}) \tag{5}$$

where **upSample** denotes the upSampling operation that recovers the reduced spatial dimensions brought about by the wavelet transform.

From Fig. 2, it can be seen that L focuses on retaining the semantic information of the image, while H emphasizes the boundary detail information in the image. This paper applies the first-order wavelet transform and Haar wavelet basis to implement the discrete wavelet transform (DWT)

Table 1. Comparison of the four datasets

Dataset	Number of images	Image data format	Image size
CVC	612	.tif	512*512
Kvasir	1000	.jpg	528*622
DSB2018	670	.png	Variable resolution
ISIC2018	2595	.jpg	512*512

algorithm. The first-order wavelet transform can adaptively adjust the decomposition according to the signal's local changes, making it more flexible than the traditional Fourier transform in processing images. The Haar wavelet, on the other hand, is characterized as a uniform square wave with rapidly changing jumps. It is particularly suitable for detecting edge and detail features in an image as it can be pinpointed in regions where the signal is changing rapidly [16]. The DWT algorithm preliminarily separates the high- and low-frequency information into LF and HF and then learns the LF features using a null convolution block, which makes the BEM module able to capture a more extensive range of contextual information; HF is learned using a small kernel convolution, which improves the module's ability in detail processing. In order to further improve the fusion ability of the model in the frequency and spatial domains, this paper introduces an adaptive mechanism, which is used to adjust the contribution of the frequency dynamic features to the original features so that the model can flexibly process the key details and semantic information in the image under different scenarios, thus improving the analysis ability and adaptability to complex medical images. The adaptive mechanism controls the regulation through two parameters, α and β., as shown in Eqs. (6) and (7):

$$\mathbf{H} = \alpha\mathbf{H}' + \mathbf{I} \tag{6}$$

$$\mathbf{L} = \beta\mathbf{L}' + \mathbf{I} \tag{7}$$

where $\mathbf{I}$ denotes the original image, and α and β both belong to the interval [0, 1], which are used to control the contribution of $\mathbf{L}^{'}$ and $\mathbf{H}^{'}$ to $\mathbf{I}$:

3.3. Feature Fusion Module

In order to effectively fuse the features of the double-branch tap, this paper designs the FFM module, whose architecture is illustrated in Fig. 3. The design concept of FFM is to enhance the model's ability to extract and utilize critical information through multi-level feature processing. Specifically, the features $\mathbf{F}_L^{'}$ and $\mathbf{F}_H^{'}$ of the double-branch taps are first aggregated using the global average pooling (GAP) operation, respectively, and their spatial channel information is captured. The captured information is used to generate the two-channel attention weights using the Sigmoid function $\mathbf{M}_L \in \mathbb{R}^{C\times1\times1}$ and $\mathbf{M}_H \in \mathbb{R}^{C\times1\times1}$, which dynamically adjust the feature maps of $\mathbf{F}_L^{'}$ and $\mathbf{F}_H^{'}$ through these two weights, highlighting the critical channel information and blocking out the irrelevant information to obtain more accurate $\mathbf{F}_L$ and $\mathbf{F}_H$ features, preparing them for fusion. The specific steps are shown in Eqs. (8), (9), (10), and (11):

$$\mathbf{M}_L = \mathrm{Sigmoid}(\mathrm{GAP}(\mathbf{F}_L^{'})) \tag{8}$$

$$\mathbf{M}_H = \mathrm{Sigmoid}(\mathrm{GAP}(\mathbf{F}_H^{'})) \tag{9}$$

$$\mathbf{F}_L = \mathbf{M}_L * \mathbf{F}_L^{'} \tag{10}$$

$$\mathbf{F}_H = \mathbf{M}_H * \mathbf{F}_H^{'} \tag{11}$$

c In this paper, we utilize the 3×3 atrous double convolution operation to learn further the global features of $\mathbf{F}_L$ focus on learning the detailed boundary features of $\mathbf{F}_H$ with the 3×3 standard double convolution operation, further filter the features of the double-branching tunnels to be merged, and apply a 1×1 convolution operation to the merged feature map. Operation: this operation not only helps to improve the expression ability of the features but also can not generate the final attention weight to reduce the amount of computation. The Sigmoid function is applied to generate the final attention weights $\mathbf{W} \in R^{\mathrm{H}\times\mathrm{W}}$, as in Eqs. (12) and (13). to enhance the network's perception of essential features and to improve the network's representation. Finally, as in Eqs. (14), this fused feature map is further processed using 1×1 convolution to adjust the channel dimensions of the features and integrate the information to generate the final fused feature map $\mathbf{F}_\mathrm{M}$:

$$\mathbf{T} = \mathrm{Concat}(\mathrm{DConv}_{3\times3}(\mathbf{F}_L), \mathrm{Conv}_{3\times3}(\mathbf{F}_H)) \tag{12}$$

$$\mathbf{W} = \mathrm{Sigmoid}(\mathrm{Conv}_{1\times1}(\mathbf{T})) \tag{13}$$

$$\mathbf{F}_\mathrm{M} = \mathrm{Conv}_{1\times1}\left(\mathbf{W} * \mathbf{F}_\mathrm{L} + \mathbf{W} * \mathbf{F}_\mathrm{H}\right) \tag{14}$$

4. Experiments

4.1. Datasets

The performance evaluation of the YNet model is based on four different open medical image datasets, including CVC-ClinicDB [2], Kvasir-SEG [12], DSB2018 [31], and ISIC2018 [20]. CVC-ClinicDB (CVC) is an endoscopic image focused on diagnosing and analysing gastrointestinal diseases. Kvasir-SEG (Kvasir) is a dataset of endoscopic images containing various gastrointestinal diseases.DSB2018 is a chest X-ray image for lung cancer screening.ISIC2018 is a dataset focusing on dermatology image analysis. Specifically, as shown in Table 1, these datasets have essential applications in the fields of medical image processing, deep learning research, and data science, where they provide researchers and developers with a large amount of image and labelling data that can be used to train and test a variety of deep learning models to improve medical diagnosis and disease screening.

Table 2. Comparison of YNet in Kvasir-SEG and CVC-ClinicDB metrics

Dataset	Kvasir				CVC			
Metric	mDice	mIoU	mPrecision	mReccall	mDice	mIoU	mPrecision	mReccall
U-Net [24]	0.782	0.714	0.724	0.745	0.846	0.773	0.849	0.879
U-Net++ [35]	0.747	0.631	0.758	0.670	0.845	0.755	0.832	0.791
ResUNet [5]	0.512	0.379	0.593	0.597	0.522	0.416	0.563	0.589
ResUNet++ [13]	0.807	0.723	0.799	0.787	0.521	0.413	0.583	0.569
ColonSegNet [10]	0.841	0.754	0.849	0.854	0.879	0.804	0.884	0.860
MSRF-Net [27]	0.858	0.791	0.869	0.867	0.912	0.872	0.902	0.901
SAM [14]	0.862	0.805	0.872	0.870	0.915	0.875	0.908	0.895
YNet(Ours)	**0.875**	**0.822**	**0.888**	**0.900**	**0.925**	**0.885**	**0.910**	**0.932**

4.2. Implementation details

In this paper, a series of steps are taken to ensure the data's consistency, quality and reliability when training YNet. In this paper, the original medical images are resized to a uniform H×W, where both H and W are 352, to maintain consistency [7, 33, 30]. To minimize the blending problems that may be introduced by image resizing, anti-aliasing techniques are introduced to improve the quality and reliability of image processing [22]. The original image and the segmentation maps are normalized to values in the range [0, 1] to facilitate loss computation and model training. According to [7, 30, 22], the dataset is divided into training, validation and test sets in the ratio of 8:1:1. A series of stochastic data enhancement operations including horizontal and vertical flips, affine transformations including angular rotations, horizontal and vertical translations, and angular shears were performed on the training set. The YNet model was trained on each dataset using a batch size of 8 and an Adam optimizer with weight decay. The learning rate is adaptively updated using an annealing algorithm, which is initially set to 1e-4 when the performance of the validation set improves by no more than 10% over ten cycles and decreases by a factor of 2 before reaching a minimum of 1e-6. These steps ensure that the data processing and model training of the YNet model during the training period are highly quality and reliable and provide consistency and comparability. The proposed model is implemented using PyTorch and trained on an NVIDIA GeForce RTX4090 GPU with 24 GB of memory.

4.3. Evaluation Metrics

In order to obtain a more comprehensive YNet performance evaluation for each dataset, this paper adopts several metrics commonly used in the field of image segmentation: mDice, mIoU, mPrecision, and mRecall, where m denotes the average value taken over the entire test set. By calculating the average value of each metric overall test sample, the comprehensive performance of the model over the entire test set can be evaluated, which helps to eliminate the chance brought by a single sample and provides a more stable and comprehensive performance evaluation. The specific defining Eqs. are shown in (15), (16), (17) and (18).

$$mDice = \frac{\sum_{i=1}^{n} \frac{2|X_i \cap Y_i|}{|X_i|+|Y_i|}}{p} \tag{15}$$

$$mIoU = \frac{\sum_{i=1}^{n} \frac{|X_i \cap Y_i|}{|X_i \cup Y_i|}}{p} \tag{16}$$

$$mPrecision = \frac{\sum_{i=1}^{n} \frac{TP_i}{TP_i+FP_i}}{p} \tag{17}$$

$$mRecall = \frac{\sum_{i=1}^{n} \frac{TP_i}{TP_i+FN_i}}{p} \tag{18}$$

where TP_i, FP_i, FN_i represent true examples, false positive examples, and false negative examples, respectively, which are the basic elements in the confusion matrix used to evaluate the classification or segmentation performance of the model. X_i, Y_i represents the prediction results and true labels. m represents the average value of the test set, and p is the number of images in the test set.

4.4. Performance Comparisons

To evaluate the performance of the YNet model, this paper compares it to several previous models using the same experimental setup to ensure a fair comparison. Four different datasets are used for the experiments, and a series of quantitative metrics and prediction examples are used to compare the performance of YNet with other models. As illustrated in Fig. 4, some prediction examples of YNet on these datasets are shown in this paper. It can be seen from the figure that YNet's segmentation results in the edge part of the performance of the protection of stronger finesse and coherence to alleviate the common phenomenon of edge blurring and breakage, and this effect is due to the feature fusion module and the attention mechanism of the model to

Table 3. Comparison of YNet's metrics in DSB2018 and ISIC2018

Dataset	DSB2018				ISIC2018			
Metric	mDice	mIoU	mPrecision	mRecall	mDice	mIoU	mPrecision	mRecall
U-Net [24]	0.887	0.808	0.872	0.920	0.868	0.782	0.879	0.849
U-Net++ [35]	0.886	0.814	0.874	0.918	0.809	0.720	0.881	0.786
ResUNet [5]	0.906	0.817	0.880	0.915	0.856	0.756	0.875	0.833
ResUNet++ [13]	0.894	0.822	0.900	0.903	0.857	0.813	0.864	0.881
ColonSegNet [10]	0.920	0.855	0.910	0.919	0.850	0.778	0.883	0.865
MSRF-Net [27]	0.924	0.853	0.902	0.930	0.882	0.837	0.915	0.889
SAM [14]	0.927	0.858	0.912	0.933	0.890	0.845	0.918	0.892
YNet(Ours)	**0.930**	**0.865**	**0.918**	**0.940**	**0.900**	**0.860**	**0.925**	**0.905**

Table 4. Ablation experiments. Includes ablation experiments for the BEM module, the FFM module, and the double-branch model.

BEM		×	×	✓	✓
FFM		×	✓	×	✓
Dataset	Metric				
CVC	mDice	0.890	0.892	0.861	**0.925**
	mIoU	0.794	0.792	0.802	**0.885**
	mPrecision	0.883	0.872	0.856	**0.910**
	mRecall	0.892	0.843	0.905	**0.932**
Kvasir	mDice	0.853	0.861	0.870	**0.875**
	mIoU	0.809	0.815	0.810	**0.822**
	mPrecision	0.869	0.871	0.873	**0.888**
	mRecall	0.889	0.876	0.895	**0.900**
DSB2018	mDice	0.910	0.915	0.917	**0.930**
	mIoU	0.845	0.844	0.847	**0.865**
	mPrecision	0.900	0.908	0.913	**0.918**
	mRecall	0.919	0.916	0.920	**0.940**
ISIC2018	mDice	0.867	0.870	0.866	**0.900**
	mIoU	0.806	0.800	0.802	**0.860**
	mPrecision	0.913	0.916	0.914	**0.925**
	mRecall	0.866	0.869	0.872	**0.905**

extract and enhance the edge information sufficiently, which makes the segmentation results of the contour lines more sharp. The result is sharper contour lines.

This paper evaluates the performance of YNet on multiple datasets using mDice, mIoU, mPrecision and mRecall metrics. The experiments compare multiple existing segmentation models, including:

(1) U-Net [24] proposed by Ronneberger et al. U-Net is a classical medical image segmentation network with an encoder-decoder structure that utilizes jump connections to preserve high-resolution features for enhanced recovery of detailed information.

(2) U-Net++ [35], proposed by Zhou et al. introduces nested jump paths on top of U-Net, enhancing the feature reuse capability and improving the segmentation performance, especially in segmentation tasks with complex structures.

(3) ResUNet [5], proposed by Zhang et al., combines the advantages of U-Net and residual learning. Referencing residual blocks enhances the network's expressive ability and reduces training difficulty.

(4) ResUNet++ [13], proposed by Zhang et al., is an improved version of ResUNet. It adopts a more complex jump-connection structure to improve the flexibility of feature fusion and thus enhance the segmentation effect.

(5) ColonSegNet [10], proposed by Jha et al., is designed for segmenting colon endoscopic images. It uses a deep convolutional neural network structure combined with a multiscale feature extraction technique to improve the recognition of colon adenomas.

(6) MSRF-Net [27], proposed by Huang et al., utilizes a multiscale residual learning architecture designed to improve the robustness and accuracy of image segmentation, especially excelling in processing medical images.

(7) Segment Anything Model (SAM)[14], proposed by Kirillov et al., is a general-purpose image segmentation model based on Vision Transformer (ViT). It performs excellently across various image scenarios, including medical image segmentation.

The experimental results on Kvasir and CVC datasets are shown in Table 2, from which it can be seen that the YNet model almost outperforms the existing models in four metrics and reaches the SOTA level. As illustrated in Fig. 4, YNet performs particularly well in the segmentation of complex structures and tiny lesions and can maintain stable performance in medical images with high noise and background complexity, showing good robustness and more explicit segmentation boundaries, thanks to the design of the BEM and the double-branch tap. Although missegmentation may still occur in some overlapping regions, overall, the effectiveness and robustness of YNet in dealing with complex medical image segmentation tasks are fully verified.

To further demonstrate the excellent generalization ability of the YNet model, this paper also conducts comparative experiments on the DSB2018 and ISIC2018 datasets, the results of which are shown in Table 3. The YNet model also outperforms the existing models on these two

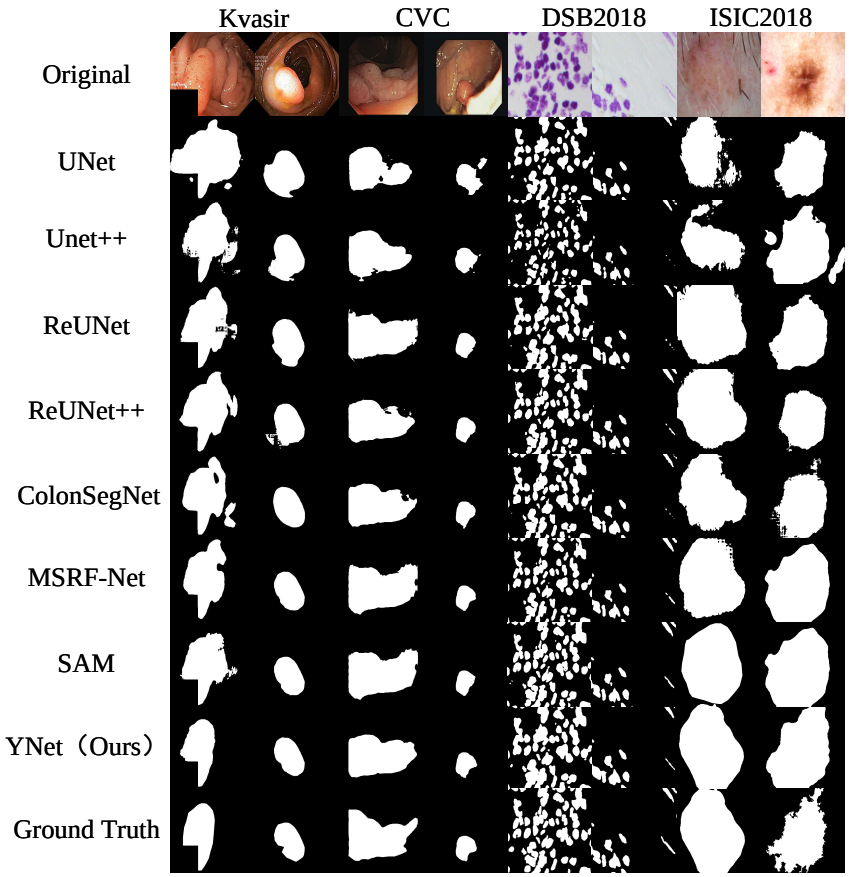

Figure 5. It shows the comparison of the results of seven models on four datasets, and from the figure, the segmentation prediction results of the YNet model are superior.

datasets, reaching the SOTA level. These excellent results are attributed to FFM, a module that effectively improves the model's performance during feature extraction and fusion.FFM adaptively adjusts the importance of different feature maps so that the model can better retain critical information and effectively suppress the influence of noise when processing complex images. This feature enables YNet to maintain a high-performance level on various datasets, showing good generalization ability and robustness. Therefore, the FFM module plays a vital role in improving the model's performance, further consolidating the competitiveness of YNet in medical image segmentation.

As can be illustrated in Fig. 5, YNet demonstrates excellent segmentation performance on datasets with different distributional characteristics. In most evaluation metrics, YNet achieves better scores than existing models, proving its strong generalization ability and adaptability. Especially when dealing with different types of medical images, YNet can effectively capture critical features and maintain high accuracy in the segmentation task. Fig. 6 shows some sample prediction results of YNet on the four datasets with other models. The comparison shows that YNet exhibits excellent segmentation accuracy in images of various morphology and complexity, especially in processing fine structure

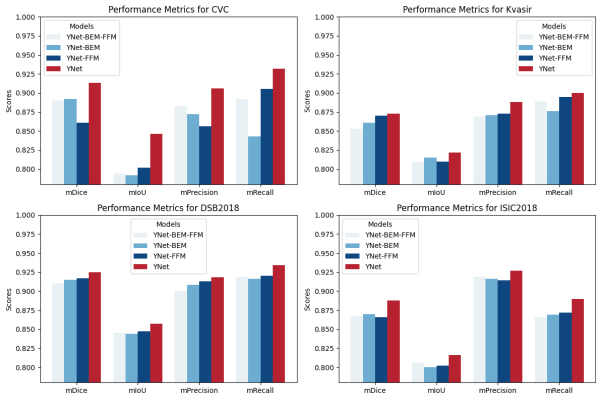

Figure 6. Visualization of YNet model ablation experiments.

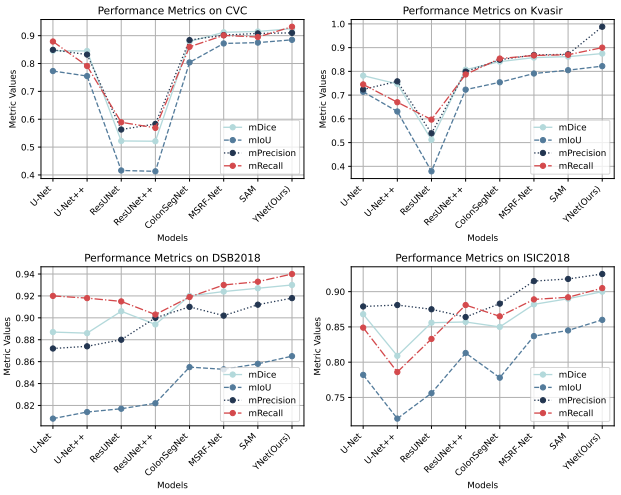

Figure 7. Comparison of segmentation performance of different models on CVC, Kvasir, DSB2018, and ISIC2018 datasets.

and boundary information.

4.5. Ablation studies

In order to verify the effectiveness of the designed BEM and FFM, ablation studies are conducted in this paper. The experimental results are shown in Table 4. In the absence of FFM, although BEM enhances the segmentation ability of boundary details, the global feature fusion and processing need to be improved. This quickly leads to the lack of global context information, thus affecting the overall segmentation performance. Moreover, when there is no BEM, although FFM can optimize the features globally, it does not process the details of the boundary sufficiently, leading to the problem of blurred boundaries and unclear transitions. From Fig. 7, YNet can improve segmentation performance only when BEM and FFM are included.BEM needs FFM to fuse the information of different scales to avoid focusing only on local details. In contrast, FFM needs BEM to make up for its insufficiency in the boundary region to ensure the accurate capture of details. They are indispensable and must work together to achieve the best results.

5. Conclusions

This paper proposes a new medical image segmentation modelling framework, YNet, to achieve clear segmentation boundaries. YNet consists of two core modules, i.e., BEM and FFM.BEM combines the advantages of wavelet transform and convolution to generate enhanced information on boundaries, which in turn provides YNet with more learnable and detailed features. On the other hand, FFM provides a compelling fusion of the features of the bipartite encoder through adaptive feature fusion fused effectively. The experimental analysis shows that the YNet network architecture we constructed can effectively improve medical image segmentation performance, which provides essential technical support for doctors' decision-making in clinical diagnosis.

Acknowledgement

This work is sponsored by Natural Science Foundation of Xinjiang Uygur Autonomous Region (Grant no. 2022D01C425), the Postgraduate Education Scientific Research Projects of Xinjiang University (Grant nos. XJDX2024YJPK15, XJDX2022YALK11), Xinjiang Tianchi Talents Program (Grant no. E33B9401). The authors would like to express their heartfelt gratitude to those people who have helped with this manuscript and to the reviewers for their comments on the manuscript.

References

[1] S. M. Azimi, P. Fischer, M. Körner, and P. Reinartz. Aerial lanenet: Lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 57(5):2920–2938, 2018. 2

[2] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 5

[3] L. Chen, L. Gu, and Y. Fu. When semantic segmentation meets frequency aliasing. *arXiv preprint arXiv:2403.09065*, 2024. 2

[4] L. Chen, L. Gu, D. Zheng, and Y. Fu. Frequency-adaptive dilated convolution for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3414–3425, 2024. 2

[5] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020. 6, 7

[6] Y. Duan, F. Liu, L. Jiao, P. Zhao, and L. Zhang. Sar image segmentation based on convolutional-wavelet neural network and markov random field. *Pattern Recognition*, 64:255–267, 2017. 2

[7] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020. 6

[8] F. Gao, X. Wang, Y. Gao, J. Dong, and S. Wang. Sea ice change detection in sar images based on convolutional-wavelet neural networks. *IEEE Geoscience and Remote Sensing Letters*, 16(8):1240–1244, 2019. 2

[9] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 1

[10] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. *Ieee Access*, 9:40496–40510, 2021. 1, 6, 7

[11] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen. Doubleu-net: A deep convolutional neural network for medical image segmentation. In *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*, pages 558–564. IEEE, 2020. 2

[12] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, and H. D. Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, pages 451–462. Springer, 2020. 5

[13] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE international symposium on multimedia (ISM)*, pages 225–2255. IEEE, 2019. 6, 7

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 6, 7

[15] M.-Q. Le, M.-T. Tran, T.-N. Le, T. V. Nguyen, and T.-T. Do. Unveiling camouflage: A learnable fourier-based augmentation for camouflaged object detection and instance segmentation. *arXiv preprint arXiv:2308.15660*, 2023. 2

[16] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570. Pmlr, 2015. 3

[17] J. Li, J. Chen, B. Sheng, P. Li, P. Yang, D. D. Feng, and J. Qi. Automatic detection and classification system of domestic waste via multimodel cascaded convolutional neural network. *IEEE transactions on industrial informatics*, 18(1):163–173, 2021. 1

[18] Q. Li and L. Shen. Wavesnet: Wavelet integrated deep networks for image segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 325–337. Springer, 2022. 3

[19] P. Liu, H. Zhang, W. Lian, and W. Zuo. Multi-level wavelet convolutional neural networks. *IEEE Access*, 7:74973–74985, 2019. 2

[20] M. A. A. Milton. Automated skin lesion classification using ensemble of deep neural networks in isic 2018: Skin lesion analysis towards melanoma detection challenge. *arXiv preprint arXiv:1901.10802*, 2019. 5

[21] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 1, 2

[22] G. Parmar, R. Zhang, and J.-Y. Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022. 6

[23] B. Patro and V. Agneeswaran. Scattering vision transformer: Spectral mixing matters. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[24] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1, 6, 7

[25] L. Sigillo, E. Grassucci, A. Uncini, and D. Comminiello. Generalizing medical image representations via quaternion wavelet networks. *arXiv preprint arXiv:2310.10224*, 2023. 2

[26] P. Sinha, Y. Wu, I. Psaromiligkos, and Z. Zilic. Lumen & media segmentation of ivus images via ellipse fitting using a wavelet-decomposed subband cnn. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2020. 2

[27] A. Srivastava, D. Jha, S. Chanda, U. Pal, H. D. Johansen, D. Johansen, M. A. Riegler, S. Ali, and P. Halvorsen. Msrf-net: a multi-scale residual fusion network for biomedical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(5):2252–2263, 2021. 1, 6, 7

[28] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 1, 2

[29] K. Upadhyay, M. Agrawal, and P. Vashist. Wavelet based fine-to-coarse retinal blood vessel extraction using u-net model. In *2020 International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5. IEEE, 2020. 2

[30] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, and S. Song. Stepwise feature fusion: Local guides global. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 110–120. Springer, 2022. 6

[31] J. Yang, Y. Sheng, Y. Zhang, W. Jiang, and L. Yang. On-device unsupervised image segmentation. In *2023 60th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2023. 5

[32] X. Yin and X. Xu. A method for improving accuracy of deeplabv3+ semantic segmentation model based on wavelet transform. In *International Conference in Communications, Signal Processing, and Systems*, pages 315–320. Springer, 2021. 2

[33] Y. Zhang, H. Liu, and Q. Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part I 24*, pages 14–24. Springer, 2021. 6

[34] Y. Zhou, J. Huang, C. Wang, L. Song, and G. Yang. Xnet: Wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21085–21096, 2023. 2

[35] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 1, 6, 7

[36] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019. 2