Wavelet Multi-scale Region-Enhanced Network for Medical Image Segmentation

Hang Lu[1] , **Liang Du**[2] , **Peng Zhou**[1*]

[1]Anhui Provincial International Joint Research Center for Advanced Technology in Medical Imaging,
School of Computer Science and Technology, Anhui University
[2] School of Computer and Information Technology, Shanxi University
E23201111@stu.ahu.edu.cn, duliang@sxu.edu.cn, zhoupeng@ahu.edu.cn

Abstract

Medical image segmentation is an important task in medical artificial intelligence. Traditional segmentation methods often suffer from the information loss problem, especially in medical image data which contain many different-scale organs or tissues. To address this problem, we propose a novel medical image segmentation method called Wavelet Multi-scale Region-Enhanced Network (WMREN), which has a UNet structure. In the encoder, we design a bi-branch feature extraction architecture, which simultaneously learns the representations with Haar wavelet transform and the residual blocks. The bi-branch architecture can effectively tackle the information loss problem when extracting features. In the decoder we design an innovative Spatial Adaptive Fusion Module to enhance the regions of interest. As we know, the boundaries of objects play an important role in segmentation. To this end, we also carefully design a Contrast Refinement Enhancement Module to highlight the boundaries of the medical objects. Extensive experiments on several benchmark datasets show that our method outperforms state-of-the-art medical image segmentation methods, demonstrating its effectiveness and superiority. The source code is publicly available at https://github.com/C101812/WMREN/tree/master.

1 Introduction

Medical image segmentation aims to segment anatomical or pathological structures within medical images. It plays a crucial role in computer-aided diagnosis and intelligent healthcare and has been widely studied [Shaker *et al.*, 2024; Ates *et al.*, 2023; Messaoudi *et al.*, 2023; Yin *et al.*, 2023; Chen *et al.*, 2024].

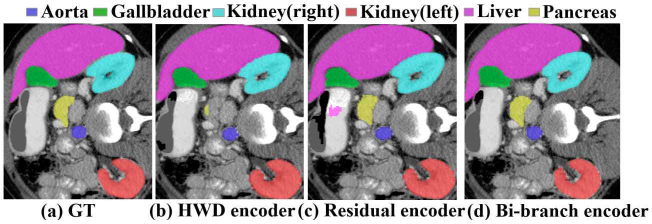
In recent years, deep learning methods, especially the convolutional neural network (CNN) based methods, such as fully convolutional network (FCN) [Shelhamer *et al.*, 2017] and UNet architecture [Ronneberger *et al.*, 2015] have achieved significant advancements in medical image segmentation. For example, [Turečková *et al.*, 2020] proposed attention gates to improve segmentation accuracy. [Fang *et al.*, 2019] combined UNet architecture with residual channel attention blocks for segmentation. [Xiang *et al.*, 2020] proposed Bio-Net, which achieved performance improvement without adding extra parameters by recurrent bi-directional skip connections.

Although UNet has been widely adopted, there still exist some issues of detailed information missing during the segmentation. More specifically, first, in the encoding stage, the downsampling approach often leads to the loss of important information [Xu *et al.*, 2023a]. Second, in the decoding stage, the fusion that directly concatenates shallow semantic information from the skip connections and deep semantic information from the previous layer of the decoder may be inappropriate in medical image segmentation. In medical image segmentation, we should focus on various regions of interest (ROI) with different shapes and sizes. However, directly concatenating shallow and deep features may cause information interference in the decoder due to background noises in shallow features, hindering the model's ability to effectively capture ROIs and detailed information [Xu *et al.*, 2023b; Oktay *et al.*, 2018].

To address these issues, in this paper, we propose a novel Wavelet Multi-scale Region-Enhanced Network (WMREN), which can effectively preserve the detail information and enhance the ROIs in the encoding and decoding process during the segmentation. To tackle the first problem, we apply the Haar Wavelet Downsampling (HWD) [Xu *et al.*, 2023a] to extract the features. HWD achieves lossless information transformation via the Haar wavelet transform, preserving detailed image information by increasing the number of feature map channels while reducing resolution. However, we observe that, when directly applying HWD to handle medical images, it cannot achieve satisfactory segmentation on multi-scale organs due to its fixed receptive field. Figure 1 shows an example. Figure 1a is the ground truth segmentation of an image in the multi-organ segmentation dataset Synapse. Figure 1b is the segmentation result with HWD as its encoder. Considering the small organ pancreas, HWD cannot segment it accurately.

To further address this issue, we propose a novel bi-branch encoder that simultaneously applies the Haar wavelet trans-

[*]Peng Zhou is the corresponding author.



(a) GT (b) HWD encoder (c) Residual encoder (d) Bi-branch encoder

Figure 1: An example of the results of different encoders. (a) is the ground truth segmentation. (b) is the segmentation result of HWD based encoder. (c) is the segmentation result of residual based encoder. (d) is the result of our bi-branch encoder.

form and the residual blocks to extract features. As shown in Figure 1b, HWD may fail to extract high-frequency features such as details and small objects like pancreas. On the contrary, the convolution based network can capture the local detailed features with a small convolution kernel. Figure 1c shows the result of the network with residual blocks as its encoder. The small organ pancreas can be segmented correctly. Therefore, we try to tackle the problem in Haar wavelet transform by parallelly extracting features with a residual network. To effectively harness the complementarity of the Haar wavelet transform and residual network, we design a novel Wavelet Residual Multi-scale Module (WRMM), which has a bi-branch structure to handle multi-scale objects. Figure 1d shows the result of our bi-branch encoder, which performs better than both the HWD based and Residual based encoders.

To tackle the second problem about the decoder, we develop a novel module named Spatial Adaptive Fusion Module (SAFM) for effectively fusing the features in the skip connections and previous layer, instead of directly concatenating them, which can effectively enhance the ROIs. SAFM module can focus on and enhance key regions and suppress background noise during the feature fusion. However, we observe that this fusion process has a side effect that may blur the boundary of objects. To address this issue, we further propose a Contrastive Refinement Enhancement Module (CREM) and plug it into the SAFM to preserve the edge information. CREM employs a bi-branch architecture, where the main branch enhances multi-scale features and fuses them with decoder features, and the other branch extracts boundary information by comparing the differences before and after boundary enhancement, which can effectively preserve the boundary information. By integrating these two modules into the decoder, the segmentation results can maintain both global semantic information and local detail information.

Our main contributions are summarized as follows:

- We propose a novel WMREN framework for medical image segmentation, in which we introduce a collaborative downsampling approach combining the wavelet transform and CNN. This framework can effectively extract multi-scale features from medical images by reducing information loss.
- We design an innovative SAFM module and a CREM module in the decoder. These modules can highlight and enhance the ROIs and preserve the boundary information, while suppressing the noises and artifacts.
- We conduct comprehensive evaluations on medical image datasets. The results demonstrate the superiority of our method compared to state-of-the-art medical image segmentation methods.

2 Related Works

2.1 Medical Image Segmentation

Medical image segmentation is an important task in medical image processing. In medical segmentation, U-shaped networks, which are in an encoder-decoder framework, are widely used in this field [Ronneberger *et al.*, 2015; Isensee *et al.*, 2018; Rahman *et al.*, 2024]. For example, UNet++ introduced a nested encoder-decoder structure with dense feature connections [Zhou *et al.*, 2020]. To better utilize multi-scale information, UNet 3+ was designed to comprehensive feature transmission pathways, achieving more thorough feature fusion [Huang *et al.*, 2020]. Some works improved UNet by incorporating attention mechanisms [Oktay *et al.*, 2018; Xiang *et al.*, 2020; Jin *et al.*, 2020] and residual structures [He *et al.*, 2016; Ibtehaz and Rahman, 2020; Sharma and Mishra, 2023; Rahman *et al.*, 2021]. For example, the DeepLab series, including DeepLabv3 [Chen *et al.*, 2018a] and DeepLabv3+ [Chen *et al.*, 2018b], incorporated atrous convolution and spatial pyramid pooling to handle multi-scale information for medical image segmentation.

Compared to CNNs, Transformer can capture long-range dependencies and enhance global semantic understanding of images more effectively [Zhang *et al.*, 2024; Xie *et al.*, 2021; Chen *et al.*, 2023; Rahman and Marculescu, 2023]. For example, TransUNet effectively captured global and local features through CNN-Transformer fusion [Chen *et al.*, 2021]. MISSFormer optimized multi-scale feature processing through innovative feed-forward networks and context bridging modules [Huang *et al.*, 2022]. DAE-Former enhanced feature dimension capture and spatial localization through improved self-attention and cross attention [Azad *et al.*, 2023]. However, the Transformer architecture still has unresolved issues in computational complexity and detail preservation during patch embedding. Therefore, in this paper, we still focus on full convolutional network based methods.

2.2 Wavelet-based Feature Learning in CNNs

Recent studies have explored integrating the wavelet transform into CNN architectures to avoid the information loss caused by downsampling. For example, Multi-level Wavelet CNN (MWCNN) was designed to achieve a good balance between receptive field size and computational efficiency in image restoration tasks [Liu *et al.*, 2018]. SFFNet was proposed to maximize spatial information utilization for segmentation tasks while effectively handling regions with diverse grayscale intensities in remote sensing imagery" [Yang *et al.*, 2024]. WTConv tackled CNN's limited receptive field problem while avoiding excessive parameter growth [Finder *et al.*, 2025]. [Xu *et al.*, 2023a] proposed a wavelet-based downsampling method (HWD) to tackle the spatial information loss during CNN downsampling.

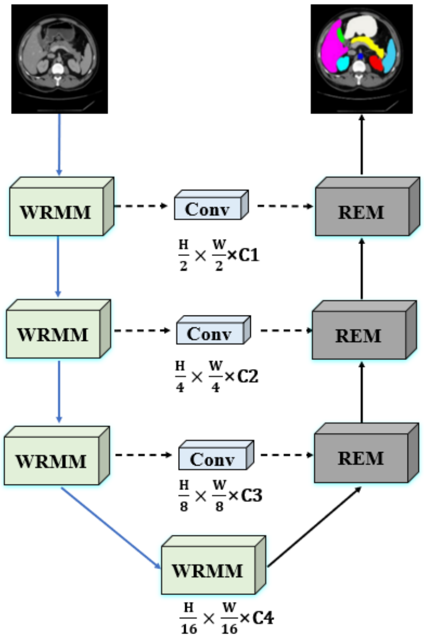
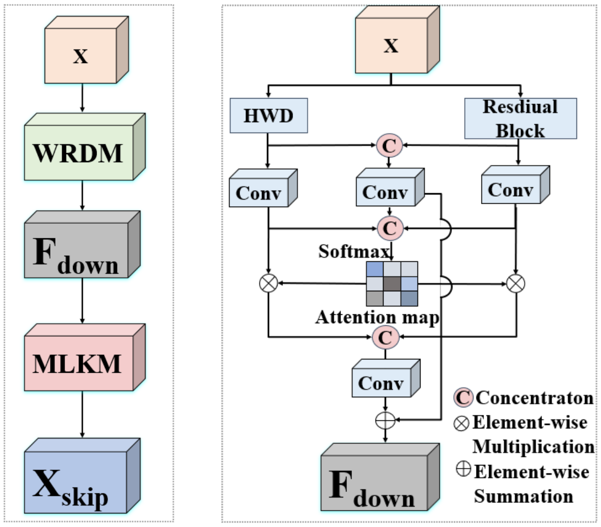


Figure 2: The framework of WMREN. It has a UNet structure, which consists of WRMM in the encoder and REM in the decoder.



(a) WRMM

(b) WRDM

Figure 3: The modules in the encoder. (a) The structure of WRMM. It consists of a WRDM for downsampling and an MLKM for convolution. (b)The structure of WRDM. It is a bi-branch downsampling module.

Despite the promising performance, we observe that the wavelet downsampling may fail to handle the organs with complicated shapes and different scales. To tackle this problem, we propose a bi-branch encoder by combining the wavelet downsampling and residual blocks to extract features.

3 Method

In this section, we introduce our WMREN in more detail. Figure 2 illustrates the overall framework of WMREN. It is a UNet structure that contains an encoder and decoder. The encoder contains four Wavelet Resdiual Multi-scale Modules (WRMM) and the decoder contains three Region Enhancement Modules (REM). WRMM will be introduced in Section 3.1 and REM will be introduced in Section 3.2. In the skip connection between encoder and decoder, we apply 1×1 convolutions to adjust the number of channels, ensuring consistent channel dimensions when fusing features with the decoder. The details of the encoder and decoder will be introduced in the following subsections.

3.1 Wavelet Residual Multi-scale Module

The conventional encoder in UNet consists of convolution and downsampling operations. In our WRMM encoder, we design a Wavelet ResNet Downsampling Module (WRDM) for downsampling and apply a Multi-scale Large Kernel Module (MLKM) [Wang *et al.*, 2024] for convolution. The structure of WRMM is shown in Figure 3a. In the following, we will introduce our designed WRDM in more detail.

Figure 3b shows the structure of WRDM. It has a bi-branch structure. The left branch applies the Haar wavelet transform for downsampling and the right branch applies a residual block to extract features. The Haar wavelet branch first uses HWD [Xu *et al.*, 2023a] for downsampling. In more detail, given an input feature map X, it decomposes it into a low-frequency component X_L and three high-frequency components: horizontal detail coefficients X_{HL}, vertical detail coefficients X_{LH}, and diagonal detail coefficients X_{HH}:

$$[X_L, X_{HL}, X_{LH}, X_{HH}] = DWT(X), \tag{1}$$

where DWT denotes a wavelet transformation function.

Then, it combines the low-frequency component and three high-frequency components and feeds them into a 1×1 convolutional layer with a batch normalization (BN) and Sigmoid activation function ($\sigma(\cdot)$) to obtain the feature maps of HWD (X_{HWD}). More formally, this downsampling can be formulated as follows:

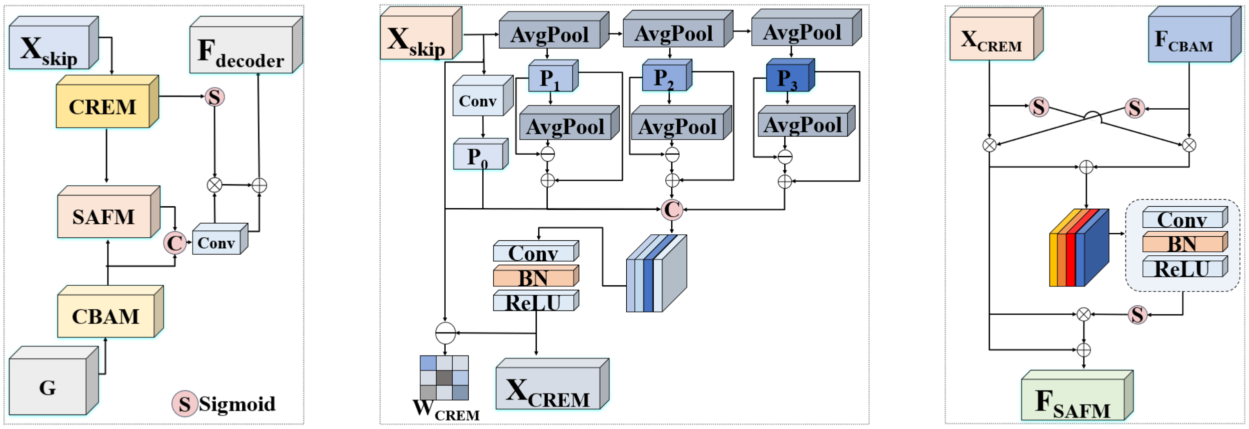
$$X_{HWD} = \sigma\left(BN\left(Conv_{1\times1}\left(Concat\left(X_L, X_{HL}, X_{LH}, X_{HH}\right)\right)\right)\right), \tag{2}$$

where $Concat$ represents the feature concatenation along the channel dimension. $Conv_{1\times1}$ denotes the 1×1 convolution.

To preserve space information more effectively, in the right branch, we also apply several residual blocks to extract features in parallel. In more detail, we first perform downsampling of feature maps via convolutional layers and then use residual blocks to extract features. Each residual block consists of three consecutive convolutional layers with kernel sizes of 1×1, 3×3, and 1×1, and then adds the input features to the output feature. We denote the result of the residual blocks as X_{Res}.

After the Haar wavelet downsampling and residual downsampling, we fuse them to obtain a new representation. To this end, we first obtain a fusion feature F_1 by concatenating two features (X_{HWD}, X_{Res}) followed by a 3×3 convolution, which is shown as follows:

$$F_1 = Conv_{3\times3}\left(Concat\left(X_{HWD}, X_{Res}\right)\right). \tag{3}$$



(a) REM

(b) CREM

(c) SAFM

Figure 4: The modules in the decoder. (a) The structure of REM. It contains CREM, SAFM, and CBAM. (b) The structure of CREM. (c) The structure of SAFM.

Then, for each branch (i.e., the HWD branch and residual branch), we apply 1×1 convolutional layer and Softmax activation function to calculate an adaptive weight (i.e., W_1 for the HWD branch and W_2 for the residual branch). More formally, we calculate W_1 and W_2 as:

$$\begin{aligned} W_1, W_2 = Split(Softmax(Conv_{1\times1} \\ (Concat(Conv_{1\times1}(X_{HWD}), Conv_{1\times1}(X_{Res}))))) \end{aligned} \quad (4)$$

where $Split$ represents the equal slice along the channel dimension. $Softmax$ represents the Softmax activation function. The weighted features are integrated through a 3×3 convolution to obtain the fusion feature F_2, which is:

$$F_2 = Conv_{3\times3}\left(X_{HWD} \odot W_1 + X_{Res} \odot W_2\right) \quad (5)$$

where $\odot$ denotes the element-wise multiplication. Finally, F_1 and F_2 are summed to obtain the final fusion representation $F_{Down} = F_1 + F_2$.

3.2 Region Enhancement Module

In the conventional UNet decoder, the fusion between deep features from the decoder and shallow features from the encoder is just a concatenation. However, in medical image segmentation, we should focus on some ROIs, especially the boundaries of the organs. Therefore, when generating the segmentation images in the decoder, we should enhance the ROIs and boundaries. To this end, we design a Region Enhancement Module in the decoder, which can enhance the important regions and boundaries when fusing the features from the encoder and decoder. The structure of REM is shown in Figure 4a. It consists of three modules: Spatial Adaptive Fusion Module (SAFM), Contrast Refinement Enhancement Module (CREM), and Convolutional Block Attention Module (CBAM) [Woo *et al.*, 2018]. SAFM enhances the main region of segmentation targets by fusing deep-layer features from the decoder with shallow-layer features from the encoder at an early stage. CREM enhances the shallow-layer features of the encoder through multi-scale boundary feature extraction. CBAM [Woo *et al.*, 2018] effectively highlights crucial information through the lightweight channel and spatial attention mechanisms.

In more detail, the feature from the previous decoder layer is first fed into CBAM to obtain an enhanced feature map, denoted as F_{CBAM}, which serves as an input of the subsequent SAFM module. Meanwhile, the shallow feature X_{skip} from the encoder is fed into CREM, which outputs an edge-enhanced feature map X_{CREM} and an edge weight information map W_{CREM}, where W_{CREM} guides boundary recovery during subsequent feature fusion. Next, SAFM receives the edge-enhanced shallow features X_{CREM} from CREM and the features F_{CBAM} from CBAM as inputs. Injecting deep feature information into shallow features effectively prevents the loss of main detail information during the subsequent fusion process. Its output is denoted as X_{SAFM}. During the process of skip connection and decoder feature fusion, X_{SAFM} and F_{CBAM} are first concatenated along the channel dimension, followed by 1×1 convolution to obtain the fused feature map F_{fusion}. Then, using the previously obtained edge weight W_{CREM} to guide the detail recovery process, we can obtain the reconstructed feature map for decoder $F_{decoder}$:

$$F_{decoder} = (F_{fusion} \odot W_{CREM}) + F_{fusion} \quad (6)$$

With this complete process, the decoder module ensures effective fusion of deep and shallow features and accurate preservation of details, thereby achieving high-quality feature segmentation. In the following subsections, we will introduce our designed CREM and SAFM in more detail.

Contrast Refinement Enhancement Module

In medical image segmentation, the boundary information is important and we should prevent the boundary blurring during the segmentation image reconstruction. To this end, we propose a Contrast Refinement Enhancement Module, as shown in Figure 4b. We tackle the boundary-blurring issue in two ways. On one hand, we enhance boundaries in the shallow feature maps of the encoder. On the other hand, we implement boundary guidance on the feature maps following feature fusion in the decoder.

Specifically, at first, we need to obtain multi-scale information from the shallow feature X_{skip}. To this end, the input feature X_{skip} first is fed into a 1×1 convolution to obtain $P_0 = Conv_{1\times1}\left(X_{skip}\right)$. Then, P_0 is fed into three cascaded AveragePool layers ($AvgPool$) to obtain feature representations at different scales, which are shown as follows:

$$P_i = AvgPool\left(P_{i-1}\right), \text{ for } i = 1, 2, 3. \tag{7}$$

P_1, P_2, and P_3 are the three scales features, respectively.

To obtain boundary information through comparison, we first perform the boundary blurring method on the feature maps P_is ($i = 1, 2, 3$) to obtain boundary blurring feature maps, and then subtract the blurred feature map from the original feature map P_i to obtain edge information. The boundary information is then added to the original feature map P_i to achieve edge enhancement. The process is shown as follows:

$$E_i = P_i - AvgPool\left(Conv_{1\times1}\left(P_i\right)\right) + P_i \tag{8}$$

where E_i is the boundary enhanced features for the i-th scale.

To obtain multi-scale boundary feature maps, we concatenate P_0 with the obtained E_i and then integrate them using 1×1 convolution, as shown in Eq.(9). The result is denoted as X_{CREM}, which will serve as an input to SAFM to provide shallow-layer information.

$$X_{CREM} = BN\left(ReLU\left(Conv_{1\times1}\left(Concat\left(P_0, E_1, E_2, E_3\right)\right)\right)\right) \tag{9}$$

Besides, CREM also generates weights like the space attention for the feature maps to highlight the boundary. The weights, denoted as W_{CREM}, are generated via a 1×1 convolutional layer on the difference between the enhanced representation X_{CREM} and the original feature map X_{skip} as follows:

$$W_{CREM} = \sigma\left(Conv_{1\times1}\left(X_{CREM} - X_{skip}\right)\right) \tag{10}$$

W_{CREM} will be used as the weights in Eq.(6) to obtain the reconstructed feature maps in decoder $F_{decoder}$.

Spatial Adaptive Fusion Module

To better highlight the important regions in the encoder's shallow representations and effectively fuse them with the decoder's representation, we propose the SAFM. As illustrated in Figure 4c, let X_{CREM} and F_{CBAM} denote the shallow features from CREM and deep features from CBAM, respectively, both with dimensions $H \times W \times C$. The feature fusion process can be described as follows.

To achieve adaptive feature selection and preserve important features, attention weights are generated by applying Sigmoid functions to both feature maps, enabling the emphasis of crucial information while suppressing less relevant features. Finally, to utilize complementary information from both features, the weighted features are combined through element-wise addition. We apply F_{MUL} as the result of fusion, which is shown as follows:

$$F_{MUL} = (X_{CREM} \odot \sigma(F_{CBAM})) + (F_{CBAM} \odot \sigma(X_{CREM})). \tag{11}$$

Notice that only the regions that are highlighted in both X_{CREM} and F_{CBAM}, which means these may be the ROIs, can be highlighted in Eq.(11). Therefore, Eq.(11) can effectively enhance the main body of ROIs.

Subsequently, the fused features F_{MUL} are fed into a 1×1 convolution, followed by a BatchNorm normalization and ReLU activation function. The integrated features are then processed through another Sigmoid function to generate the final attention weights W_{SAFM}, which are shown as follows:

$$W_{SAFM} = \sigma\left(ReLU\left(BN\left(Conv_{1\times1}\left(\, F_{MUL}\right)\right)\right)\right), \tag{12}$$

where $ReLU$ denotes the ReLU activate function.

To ensure that useful low-level information is not lost while highlighting important features, we apply residual connections to preserve the original information. Specially, these weights in Eq.(12) are multiplied with X_{CREM} and combined with the input features through a residual connection as: We denote the result of the SAFM as F_{SAFM}

$$F_{SAFM} = \left(X_{CREM} \odot W_{SAFM}\right) + X_{CREM}. \tag{13}$$

3.3 Loss Function

We employ a composite loss function to optimize the network parameter, which combines Cross Entropy Loss $\mathcal{L}_{CE}$ and Dice Loss $\mathcal{L}_{\text{Dice}}$. Specifically, the total loss function $\mathcal{L}_{\text{total}}$ is defined as:

$$\mathcal{L}_{\text{total}} = 0.4 \times \mathcal{L}_{CE} + 0.6 \times \mathcal{L}_{\text{Dice}}. \tag{14}$$

The Cross Entropy Loss ($\mathcal{L}_{CE}$) measures pixel-level discrepancies between predicted results and ground truth labels, contributing to improving classification accuracy. The Dice Loss ($\mathcal{L}_{Dice}$) focuses on evaluating the overlap between predicted segmentation masks and ground truth masks to improve the segmentation accuracy.

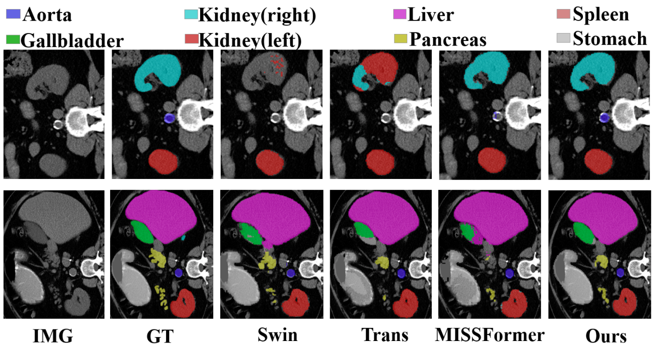


Figure 5: The segmentation results on Synapse dataset. Our method most closely approximates the ground truth (GT), particularly in preserving the fine details of the gallbladder and aorta that are lost in other methods

4 Experiment

4.1 Experimental Setup and Implementation Details

We evaluate our method on three benchmark datasets: Synapse[1]), ACDC[2], and ISIC17 [Codella *et al.*, 2018]. On

[1] https://www.synapse.org#!Synapse:syn3193805/wiki/217789

[2] https://www.creatis.insa-lyon.fr/Challenge/acdc/

Architectures	Avg. DSC↑	Avg. HD95↓	Aorta↑	Gallbladder↑	Kidney (L)↑	Kidney (R)↑	Liver↑	Pancreas↑	Spleen↑	Stomach↑
UNet [Ronneberger *et al.*, 2015]	70.11	44.69	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96
R50+ViT [Chen *et al.*, 2021]	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
R50+AttnUNet [Chen *et al.*, 2021]	75.57	35.97	85.16	69.42	79.20	71.07	93.38	42.88	87.27	70.28
TransUNet [Chen *et al.*, 2021]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
SwinUNet [Cao *et al.*, 2022]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
CoTr [Xie *et al.*, 2021]	78.56	24.05	87.09	65.37	86.19	80.32	94.22	52.28	87.01	76.00
DAEFormer [Azad *et al.*, 2023]	82.43	17.46	88.96	72.30	86.08	80.88	94.98	65.12	91.96	79.19
MISSFormer [Huang *et al.*, 2022]	81.96	18.20	87.19	70.23	84.91	83.94	94.41	65.67	91.92	80.81
CT-Net [Zhang *et al.*, 2024]	82.60	-	89.00	67.70	84.10	80.60	**96.20**	67.90	90.00	85.00
nn-Unet [Wang *et al.*, 2023]	82.36	24.74	**90.96**	65.57	81.92	78.36	95.96	69.36	91.12	**85.60**
PVT-CASCADE [Rahman and Marculescu, 2023]	81.06	20.23	83.01	70.59	82.23	80.37	94.08	64.43	90.10	83.69
PVT-EMCAD-B2[Rahman *et al.*, 2024]	83.63	15.68	88.14	68.87	88.08	84.10	95.26	68.51	**92.17**	83.92
WMREN (ours)	**84.40**	**15.65**	88.88	**74.63**	**88.52**	**84.24**	95.11	**69.62**	91.08	83.11

Table 1: The performance of all methods on the Synapse dataset. We report DSC scores for each organ individually. We also report the average DSC, which is denoted as Avg. DSC, and the average HD 95, which is denoted as Avg. HD95. An upward arrow (↑) indicates higher values are better, and a downward arrow (↓) indicates lower values are better. The best results are highlighted in **bold**.

Methods	Avg.DSC	RV	Myo	LV
R50+UNet [Chen *et al.*, 2021]	87.55	87.10	80.63	94.92
R50+AttenUNet [Chen *et al.*, 2021]	86.75	87.58	79.20	93.47
SwinUNet[Cao *et al.*, 2022]	88.07	85.77	84.42	94.03
TransUNet [Chen *et al.*, 2021]	89.71	86.67	87.27	95.18
MISSFormer [Huang *et al.*, 2022]	90.36	89.55	85.44	94.99
PVT-CASCADE [Rahman and Marculescu, 2023]	91.46	**89.97**	88.90	95.50
PVT-EMCAD-B2 [Rahman *et al.*, 2024]	92.12	90.65	89.68	96.02
WMREN (ours)	**92.64**	**92.69**	89.05	**96.18**

Table 2: Performance comparison against the baseline decoder on the ACDC dataset. The best results are highlighted in **bold**.

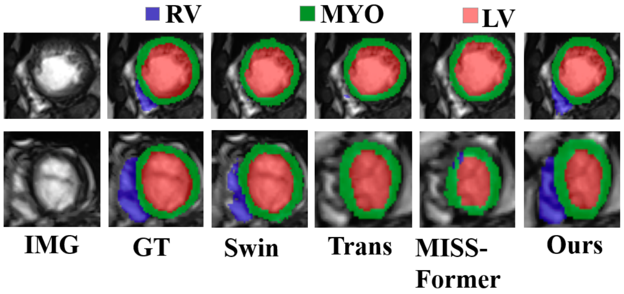


Figure 6: The segmentation results on ACDC dataset. For the right ventricle details and myocardial borders, our segmentation results show closer alignment with the ground truth images.

all data sets, we follow the default protocol for splitting the dataset into training, validation, and testing sets. The Synapse dataset is a multi-organ segmentation dataset, which consists of 3779 axial abdominal CT images collected from 30 clinical cases. Each image is annotated with labels for eight abdominal organs: aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, and stomach. The performance evaluation is conducted using two metrics: the average Dice score (DSC) in percentage and the average Hausdorff Distance (HD95) in millimeters across all eight abdominal organs. The ACDC dataset is an automated cardiac diagnostic challenge dataset that consists 100 cardiac cine-MRI scans from clinical cases, and corresponding labels for three cardiac regions: the left ventricle (LV), right ventricle (RV), and myocardium (MYO). All methods are evaluated with DSC over all three cardiac regions. The ISIC17 dataset is a skin lesion segmentation dataset containing high-resolution medical images of melanoma skin lesions. Each image includes pixel-level annotations marking the lesion area. Performance evaluation is conducted using the DSC to measure the segmentation accuracy of the lesion regions.

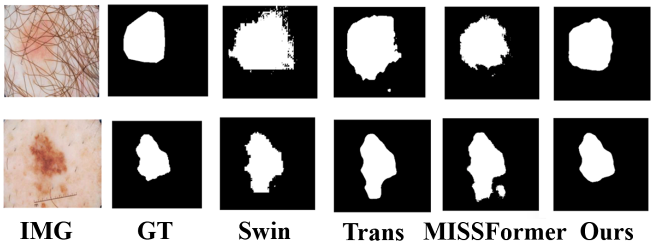


Figure 7: The segmentation results on ISIC2017 dataset. Our segmentation method outperforms existing approaches in both boundary delineation and main body segmentation.

Methods	Avg.DSC
UNET [Ronneberger *et al.*, 2015]	83.07
DeepLabv3+ [Chen *et al.*, 2018a]	83.84
SwinUnet [Cao *et al.*, 2022]	83.24
PVT-CASCADE [Rahman and Marculescu, 2023]	85.50
PVT-EMCAD-B2 [Rahman *et al.*, 2024]	85.95
TransUnet [Chen *et al.*, 2021]	86.94
MISSFormer [Huang *et al.*, 2022]	86.34
WMREN (ours)	**87.67**

Table 3: Performance comparison against the baseline decoder on the ISIC17 dataset. Best results are highlighted in **bold**.

All experiments are implemented using PyTorch on Windows 10 system and Nvidia GeForce RTX 4090 GPU. Following [Huang *et al.*, 2022], various data augmentation techniques are employed, including random cropping, random rotation, random Gaussian noise, random blurring, as well as transformations in luminance, contrast, and resolution. Images from Synapse and ACDC datasets are reshaped to 224 × 224 pixels. In the ISIC17 dataset, images are resized to 256 × 256. The model is trained via SGD optimizer with a momentum of 0.90, weight decay of 0.0001, and an initial learning rate of 0.05 and following a polynomial decay policy. It is trained for 400 epochs with a batch size of 24.

Architecture	Avg. DSC↑	Avg. HD95↓
H-Net	81.29	18.89
R-Net	82.59	18.75
WMREN(ours)	**84.40**	**15.65**

Table 4: Results of different degenerated approaches of downsampling on Synapse dataset. The best results are highlighted in **bold**.

Decoder	Avg. DSC↑	Avg. HD95↓
baseline	82.72	17.23
baseline+SAFM	83.52	21.86
baseline+SAFM+CREM	**84.40**	**15.65**

Table 5: Results of different degenerated approaches of decoder on Synapse dataset. The best results are highlighted in **bold**.

4.2 Experimental Results

Table 1 shows the results of WMREN and other methods on the Synapse dataset. It shows that WMREN achieves state-of-the-art performance with an average DSC of 84.40%, surpassing both CNN-based and Transformer-based methods. Our method also has the lowest HD95 at 15.65 mm, indicating its effectiveness in organ boundary detection. Besides, compared to other state-of-the-art methods, our method achieves better segmentation accuracy for the small organs, such as the pancreas, gallbladder, and kidneys. It demonstrates superior performance in segmenting small organs, while maintaining high accuracy for large organs such as the stomach and liver.

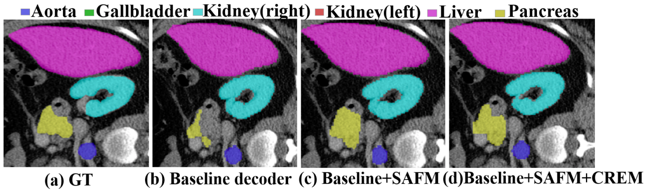
Tables 2 and 3 show the results on ACDC and ISIC17 datasets, respectively. WMREN also performs the best on these datasets. Our method achieves a leading average DSC of 92.12% on ACDC dataset and 87.67% on ISIC17 dataset, demonstrating the superiority of the proposed method.

Figures 5, 6, and 7 show the qualitative results on the Synapse, ACDC, and ISIC17 datasets, respectively. As demonstrated in Figure 5, our method exhibits superior performance in capturing fine details such as small objects and boundaries compared to other approaches. Figure 6 shows that our method demonstrates superior segmentation performance in areas where other approaches lose details, particularly in the right ventricle and myocardial boundaries. As shown in Figure 7, our method effectively mitigates common issues found in other approaches, including blurred segmentation boundaries and spurious segmented objects.

4.3 Ablation Study

In this section, we conduct comprehensive ablation studies to evaluate the effects of our designed modules, including the WRDM in the encoder, and SAFM and CREM in the decoder.

In the encoder, we design a WRDM module that simultaneously applies the Haar wavelet branch and residual branch to extract features. To show the effectiveness of the bi-branch architecture, we compare it with two single-branch encoders, i.e., H-Net that only uses the Haar wavelet branch and R-Net that only uses the residual branch. The results on Synapse are shown in Table 4. It can be seen that bi-branch network performs better than both the single-branch network, demonstrating the effectiveness of bi-branch encoder WRDM.



(a) GT (b) Baseline decoder (c) Baseline+SAFM (d)Baseline+SAFM+CREM

Figure 8: Qualitative result of ablation study on decoders. (a) is the ground truth segmentation. (b) is the segmentation result of baseline decoder. (c) is the segmentation result of baseline+SAFM (d) is the result of our REM decoder which is based baseline+SAFM+CREM.

In the decoder, we design a CREM module to enhance the boundaries and SAFM to enhance the ROIs. To show the effectiveness of these two modules, we conduct ablation studies by comparing with the following degenerated models: baseline denotes the UNet backbone without the CREM and SAFM. Baseline+SAFM denotes the network with SAFM but without CREM. Baseline+SAFM+CREM denotes our whole network. We can see that, compared to the baseline model, baseline+SAFM improved DSC by 0.8% and the combination of CREM and SAFM improved DSC by 1.68%. This demonstrates that SAFM is effective in enhancing the model's performance. Furthermore, we can observe that baseline+SAFM increased baseline by 4.58 w.r.t. HD95. The main reason may be that although SAFM can enhance the ROIs, it may blur the boundaries which causes the increase of the HD95. To tackle this problem, we design the CREM module to enhance the boundaries. We can see that, with CREM, our model achieves 15.65 w.r.t. HD95, which is much lower than both baseline and baseline SAFM, demonstrating the effectiveness of the CREM on boundaries enhance.

Figure 8 shows a qualitative result. Considering the pancreas, the baseline decoder cannot detect the whole region of the pancreas. Together with SAFM, baseline+SAFM can capture the regions, showing the effectiveness of the region enhancement. However, the boundary of baseline+SAFM is incorrect. After plugging CREM into the model, we can segment the pancreas more accurately. It demonstrates the effectiveness of CREM in boundary enhancement.

5 Conclusion

In this paper, to address the detail loss issues of traditional medical image segmentation methods, we proposed an innovative Wavelet Multi-scale Region-Enhanced Network for medical image segmentation. In the encoder, we integrated wavelet downsampling with residual blocks to minimize detail information loss during the downsampling. In the decoder, we carefully designed a Contrastive Refinement Enhancement Module to highlight the boundaries of organs and a Spatial Adaptive Fusion Module to enhance the ROIs. Extensive experiments on benchmark datasets show that our method outperforms other state-of-the-art medical images segmentation methods. The ablation studies also demonstrates the effectiveness of all our designed modules and their contributions to the overall performance.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China grants 62176001, and the Natural Science Project of Anhui Provincial Education Department under grant 2023AH030004.

References

[Ates *et al.*, 2023] Gorkem Can Ates, Prasoon Mohan, and Emrah Celik. Dual cross-attention for medical image segmentation. *Engineering Applications of Artificial Intelligence*, 126:107139, 2023.

[Azad *et al.*, 2023] Reza Azad, René Arimond, Ehsan Khodapanah Aghdam, Amirhossein Kazerouni, and Dorit Merhof. Dae-former: Dual attention-guided efficient transformer for medical image segmentation. In *International Workshop on PRedictive Intelligence In MEdicine*, pages 83–95. Springer, 2023.

[Cao *et al.*, 2022] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[Chen *et al.*, 2018a] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[Chen *et al.*, 2018b] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[Chen *et al.*, 2021] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[Chen *et al.*, 2023] Bingzhi Chen, Yishu Liu, Zheng Zhang, Guangming Lu, and Adams Wai Kin Kong. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023.

[Chen *et al.*, 2024] Jiarong Chen, Xuyang Zhang, Rongwen Li, and Peng Zhou. Swin-haunet: A swin-hierarchical attention unet for enhanced medical image segmentation. In *PRCV - 7th Chinese Conference, PRCV 2024, Urumqi, China, October 18-20, 2024, Proceedings, Part XIV*, volume 15044 of *Lecture Notes in Computer Science*, pages 371–385. Springer, 2024.

[Codella *et al.*, 2018] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[Fang *et al.*, 2019] Zhenghan Fang, Yong Chen, Dong Nie, Weili Lin, and Dinggang Shen. Rca-u-net: Residual channel attention u-net for fast tissue quantification in magnetic resonance fingerprinting. In *MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pages 101–109. Springer, 2019.

[Finder *et al.*, 2025] Shahaf E Finder, Roy Amoyal, Eran Treister, and Oren Freifeld. Wavelet convolutions for large receptive fields. In *European Conference on Computer Vision*, pages 363–380. Springer, 2025.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Huang *et al.*, 2020] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059, 2020.

[Huang *et al.*, 2022] Xiaohong Huang, Zhifang Deng, Dandan Li, Xueguang Yuan, and Ying Fu. Missformer: An effective transformer for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(5):1484–1494, 2022.

[Ibtehaz and Rahman, 2020] Nabil Ibtehaz and M Sohel Rahman. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks*, 121:74–87, 2020.

[Isensee *et al.*, 2018] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

[Jin *et al.*, 2020] Qiangguo Jin, Zhaopeng Meng, Changming Sun, Hui Cui, and Ran Su. Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans. *Frontiers in Bioengineering and Biotechnology*, 8:605132, 2020.

[Liu *et al.*, 2018] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 886–88609, 2018.

[Messaoudi *et al.*, 2023] Hicham Messaoudi, Ahror Belaid, Douraied Ben Salem, and Pierre-Henri Conze. Cross-dimensional transfer learning in medical image segmentation with deep learning. *Medical image analysis*, 88:102868, 2023.

[Oktay *et al.*, 2018] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[Rahman and Marculescu, 2023] Md Mostafijur Rahman and Radu Marculescu. Medical image segmentation via cascaded attention decoding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6222–6231, 2023.

[Rahman *et al.*, 2021] Aamer Abdul Rahman, Birendra Biswal, Shazia Hasan, MVS Sairam, et al. Robust segmentation of vascular network using deeply cascaded aren-unet. *Biomedical Signal Processing and Control*, 69:102953, 2021.

[Rahman *et al.*, 2024] Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11769–11779, 2024.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[Shaker *et al.*, 2024] Abdelrahman M Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Unetr++: delving into efficient and accurate 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024.

[Sharma and Mishra, 2023] Ajay Sharma and Pramod Kumar Mishra. Dri-unet: dense residual-inception unet for nuclei identification in microscopy cell images. *Neural Computing and Applications*, 35(26):19187–19220, 2023.

[Shelhamer *et al.*, 2017] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017.

[Turečková *et al.*, 2020] Alžběta Turečková, Tomáš Tureček, Zuzana Komínková Oplatková, and Antonio Rodríguez-Sánchez. Improving ct image tumor segmentation through deep supervision and attentional gates. *Frontiers in Robotics and AI*, 7:106, 2020.

[Wang *et al.*, 2023] Jing Wang, Haiyue Zhao, Wei Liang, Shuyu Wang, and Yan Zhang. Cross-convolutional transformer for automated multi-organs segmentation in a variety of medical images. *Physics in Medicine & Biology*, 68(3):035008, 2023.

[Wang *et al.*, 2024] Yan Wang, Yusen Li, Gang Wang, and Xiaoguang Liu. Multi-scale attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5950–5960, 2024.

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[Xiang *et al.*, 2020] Tiange Xiang, Chaoyi Zhang, Dongnan Liu, Yang Song, Heng Huang, and Weidong Cai. Bio-net: learning recurrent bi-directional connections for encoder-decoder architecture. In *MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 74–84. Springer, 2020.

[Xie *et al.*, 2021] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 171–180. Springer, 2021.

[Xu *et al.*, 2023a] Guoping Xu, Wentao Liao, Xuan Zhang, Chang Li, Xinwei He, and Xinglong Wu. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern Recognition*, 143:109819, 2023.

[Xu *et al.*, 2023b] Jiacong Xu, Zixiang Xiong, and Shankar P. Bhattacharyya. Pidnet: A real-time semantic segmentation network inspired by pid controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19529–19539, June 2023.

[Yang *et al.*, 2024] Yunsong Yang, Genji Yuan, and Jinjiang Li. Sffnet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024.

[Yin *et al.*, 2023] Yunchou Yin, Zhimeng Han, Muwei Jian, Gai-Ge Wang, Liyan Chen, and Rui Wang. Amsunet: A neural network using atrous multi-scale convolution for medical image segmentation. *Computers in Biology and Medicine*, 162:107120, 2023.

[Zhang *et al.*, 2024] Ning Zhang, Long Yu, Dezhi Zhang, Weidong Wu, Shengwei Tian, Xiaojing Kang, and Min Li. Ct-net: Asymmetric compound branch transformer for medical image segmentation. *Neural Networks*, 170:298–311, 2024.

[Zhou *et al.*, 2020] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2020.