HIFNet: Medical Image Segmentation Network Utilizing Hierarchical Attention Feature Fusion

Zhou Yang
Shcool of Computer Science and Technology, Shandong Technology and Business University
Shandong, Yantai, China
2022410072@sdtbu.edu.cn

Hua Wang
School of Information and Electrical Engineering, Ludong University
Shandong, Yantai, China
hwa229@163.com

Fan Zhang(✉)
Shcool of Computer Science and Technology, Shandong Technology and Business University
Shandong, Yantai, China
zhangfan@stdbu.edu.cn

Abstract

The Transformer model has demonstrated immense potential and significant importance as an efficient tool in the field of medical image analysis, primarily due to its capability to capture global context. However, its limitation in capturing local information to some extent constrains the full performance of Transformer in this domain. To mitigate this issue, we propose a novel medical image segmentation network, HIFNet, based on hierarchical attention feature fusion. Specifically, we utilize a pre-trained MaxViT as the encoder. In our newly constructed decoder, spatial attention is applied to feature maps of different sizes to focus more on critical regions of the input images. Additionally, we incorporate multiple attention mechanisms, including criss-cross attention, to capture sensitive spatial relationships within medical images. Furthermore, we employ coordinate attention in skip connections to embed positional information in different directions, thereby generating feature maps containing sensitive positional information. Experiments conducted on relevant medical image datasets demonstrate the effectiveness and scalability of our proposed encoder.

Keywords: Medical image Segmentation Transformer Multi-scale network.

1. Introduction

Medical image segmentation, as a component in image processing tasks, is not only crucial in the field of computer science but also serves as a key tool for computer-aided diagnosis in clinical applications. By classifying the pixels in medical images, it aids doctors in rapidly identifying organs, tumors, and other related lesion areas. Traditional medical image segmentation networks utilize U-Net[26] as the backbone to construct a symmetrical U-shaped network. Most of these networks are based on Convolutional Neural Networks (CNNs) and have introduced a series of related work on this foundation, such as U-Net++[49], UNet3+[12], nnunet[15], etc[11, 19, 20]. These networks can generate relatively clear segmentation maps. However, the use of a single CNN model is limited by factors such as its restricted receptive field[10] and inherent inductive biases[2], which constrain the ultimate segmentation performance. Therefore, researchers have also introduced attention mechanisms[29, 37] into related architectures, such as SCAU-net[48], DRAu-net[27]. To mitigate the limitations of CNNs in capturing long-distance dependencies. Attention enhances segmentation results by capturing salient features in the image[11, 8]. Despite improvements in segmentation performance through the introduction of attention mechanisms, there is still room for enhancement in capturing important feature information in images.

Recently, Transformers have demonstrated excellent performance in natural language processing-related fields[29],

attributed to their use of Self-Attention (SA) mechanisms within their structures. Researchers have subsequently introduced Vision Transformer[10] (ViT) into the image domain and achieved great success. By dividing the input image into small patches and leveraging the self-attention mechanism to capture the relationships among these patches, Vision Transformer (ViT) aims to extract global information features. Furthermore, researchers have developed and proposed various Transformer variants suitable for the image domain, such as Swin-Transformer[18], PVT[34, 35], MaxViT[28], and MERIT[23]. In particular, the successful application of Transformers in medical images[30, 23, 21, 9, 19] further demonstrates their great potential. However, due to the limitations of the self-attention module in understanding local spatial information, some methods have incorporated convolutional attention modules to alleviate this issue[36, 39]. But, because of the limitations of convolution, these methods struggle to capture the relationships between long-distance pixels.

To address the above issues, we consider improving the use of attention mechanisms to capture both global and local dependencies across all corresponding dimensions. Therefore, we have designed a new Transformer network for medical image segmentation, namely HIFNet. This network uses MaxViT as the encoder and incorporates a novel decoder designed by us. Additionally, we employ coordinate attention to embed information in the skip connections within the network. Our contributions are as follows:

•We propose a new decoder. Each decoder is composed of a mixed attention mechanism across different spatial directions and a multi-scale feature mixing module, allowing the decoder to obtain more comprehensive and detailed feature information. By mixing attention results in different directions, the decoder can capture correlations between different positions in the input data, extract more detailed contextual information, and facilitate the output of more refined segmentation results.

•To enhance the model's learning capability of latent features, we process skip connection information using a coordinate attention mechanism between multi-stage encoders and decoders for feature merging. This method enhances the model's acquisition of key location information and improves the final segmentation performance. By combining dilated convolutions of different scales with related operations, the model expands its receptive field, enabling it to capture long-distance dependencies.

2. Related work

2.1. Medical Image Segmentation with CNN

U-Net[26], as a cornerstone in medical image segmentation in recent years, has demonstrated tremendous potential and achievements in this field. As the first CNN-based medical image segmentation model, it has sparked numerous subsequent research efforts, including the development of a series of U-Net-based models such as U-Net++[49], U-Net3+[12], and Attention-Unet[20]. These U-shaped architecture models have shone brightly in various medical image segmentation tasks due to their simple and effective encoder-decoder structure, such as segmenting lesions, tumors, and organs[7, 9, 36, 46]. These studies collectively indicate that U-Net and its related subsequent models have become the de facto benchmarks for medical image segmentation[11, 19]. After introducing U-Net to medical image segmentation, researchers also encountered some challenges. Although CNN-based U-Net models excel in segmentation results, they still have issues capturing long-distance critical information. To overcome this problem, researchers proposed using attention-based U-Net models, such as MA-Unet[3], which employs an attention mechanism to manipulate multi-scale features, and APAUnet[16], which uses axial projection attention. These models demonstrate that attention can compensate for the long-distance modeling capabilities limited by using a single CNN. Therefore, in the encoder we designed, we further enhance the model's long-distance modeling capability by incorporating attention mechanisms in different directions.

2.2. Vision Transformers

Given the tremendous achievements of transformers in other fields[43, 44, 33, 45], researchers first proposed Vision Transformer (ViT)[10], the first model to apply transformers to the field of vision. ViT learns global information between pixels using a self-attention mechanism. Subsequent models have improved upon ViT through various methods, including integrating CNN features into the architecture (MaxViT[28], PvTv2[35]), using new attention mechanisms (Swin-Transformer[18]), and proposing new architectures (PVT[34], SegFormer[39]). The Swin-Transformer calculates local attention through sliding window attention shifts, SegFormer leverages MiXFFN to aggregate information from different levels, and MaxViT computes self-attention after decomposing the spatial axis. Although many transformer models[6, 5, 38] for computer vision have mitigated the long-distance modeling limitations of using CNNs, they still face challenges in capturing local information and feature relationships. In this paper, we address this limitation by introducing a decoder that integrates an attention mechanism with a multi-scale feature convolution module. This encoder models sensitive information in the image through coordinate attention and cruciform cross-attention, while simultaneously extracting features through a multi-scale attention feature mixing module with a pyramid-like structure. This enhances the model's ability to model key information.

2.3. Medical Image Segmentation with Transformers

HIFNet.pdf

Figure 1. As shown in the figure, this is the overall network architecture we proposed. On the left side of the network is the pre-trained MaxViT encoder we utilized, which performs attention operations using different window sizes.

Despite the tremendous achievements of CNN-based medical image segmentation networks, the limitation of their global information modeling poses constraints on model performance. In the research on medical image segmentation tasks, researchers prefer to use transformer-based networks, and proposed a series of models that utilize Transformer for medical image segmentation.[17, 47, 41, 40, 42] For example, Swin-Unet[4] proposes a U-shaped pure transformer architecture based on Swin-Transformer[18], while TransUnet[7] combines CNNs and transformers to capture low-level and high-level features. Additionally, some researchers use pre-trained models as encoders, benefiting from the rich features already learned by these models on large datasets, which reduces the need for new data. For instance, PVT-Polyp[9] achieves good results in polyp segmentation by using a pre-trained PVT as the encoder, and G-CASCADE[22] generates finer segmentation maps by combining MaxViT as the encoder with a decoder that includes graph convolution. EMCAD[24] also employs a pre-trained model as the encoder and captures complex spatial relationships by constructing a mechanism that includes spatial, channel, and grouped gating attention. Therefore, we have decided to adopt a similar approach, utilizing a pre-trained model as the encoder and constructing a complex architecture that integrates attention and convolution as the decoder to capture important contextual relationships in medical images.

3. Method

3.1. Overall Architecture

To further enhance the long and short-distance modeling capabilities of transformer models in medical image segmentation, we drew inspiration from the previous design of MaxViT[28]. MaxViT achieves information interaction between global and local scales by decomposing the traditional self-attention mechanism into two types of sparse attention: non-overlapping window attention and grid attention, along the spatial axis. In our proposed HIFNet, we inherited the essence of MaxViT by utilizing a pre-trained MaxViT model in the encoder stage. Simultaneously, we innovatively reconstructed the encoder stage. Specifically, we first constructed a top-down multi-scale information extraction path corresponding to the encoder in the decoder stage, ensuring that the image can fully utilize information features of different scales during the decoder stage. Subsequently, we processed the information in the encoder using skip connections with positional embeddings and fed it into the corresponding decoder sections. This pyramid-like structure design enhances the model's ability to utilize information of different scales and allows the model to more comprehensively capture the key location information of the image, thereby significantly improving the model's segmentation performance.

3.2. Multi-Scale Attention Feature Extraction Decoder

Due to its inherent limitations in modeling local information, Transformer may lose critical details when processing medical images, thereby affecting the final segmentation performance of the model. To mitigate this issue, besides using a pre-trained MaxViT in the encoder, we have redesigned the decoder stage to include a Multi-Scale Attention Feature Extraction Decoder (MAFED), as illustrated in the figure. The decoder begins with layer normalization, followed by upsampling the input data to match the size of the encoder output. Then, we use convolution to ensure that the dimensions match those of the skip connection inputs before they enter the decoder, and normalization is applied. The specific formula is shown below:

$$x = concat((LN(Conv(Up(x_1)))), CDA(x_1)) \quad (1)$$

Here, $concat$ denotes concatenation, x_1 represents the corresponding encoder output, CDA stands for Coordinate Attention[14], LN for Layer Normalization, $conv$ for applying convolution for dimension processing, and UP for upsampling. Subsequently, a depthwise convolution is used to reduce the channel dimension by half, and the result is input into the encoder for further processing. In the encoder, we introduce cross-shaped attention[14] to process the input data. The horizontal and vertical attention modules extract contextual information in the horizontal and vertical directions, respectively, as shown in Figure 2 (a). Given a feature map $H \in R^{H\times W\times C}$, The module first generates feature maps Q, K, and V for attention calculation using 1x1 convolutions along the H dimension. The size of the mappings for Q and K is $H \times W \times C_1$, The reason C_1 is smaller than C is to perform dimensionality reduction on the channels, thereby reducing computational load. Subsequently, through the Affinity operation, a feature map of size $P \in R^{(H+W-1)\times W\times C}$ is generated, capturing the relationships between each pixel and its horizontally and vertically adjacent pixels in the feature map. The calculation of Affinity is as follows:

$$d_{i,u} = Q_u \Omega_{i,u} \quad (2)$$

Here,$\Omega_{i,u}$ represents the feature vector at position iu in the feature map, with $\Omega_{i,u} \in R^c$, Q_u denotes the feature

vector at position u in the feature map P After computation, a new feature map of size $(H + W - 1) \times W \times C$ is obtained. Subsequently, a softmax operation is applied to this feature map. Then, the same operation is performed for each position to reconstruct a feature map of size HWC that once again contains information about the relationships between pixels. After applying the cross-shaped attention

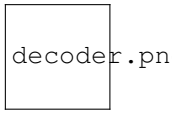

Figure 2. As shown in the figure, the cross-shaped attention is depicted at the top (a), and the multi-scale feature extraction module is shown at the bottom (b). Together, these two components constitute the decoder.

mechanism, we utilize a multi-scale feature extraction module to further extract and model key positional information. The structure of this module is illustrated in the figure. By applying spatial attention to features of different scales, the model can further focus on important regions. Specifically, the feature extraction module employs multiple dilated convolutional layers with different dilation rates in the extraction stage to capture features of various scales. Subsequently, in each output branch stage, an attention mechanism is used to further assess the importance of the extracted features. To avoid information loss when using a single dilated convolution, we incorporate 1x1 convolutions and pooling to ensure complete information preservation. The specific formula for the module is as follows:

$$P = softmax(R(Conv(x_{in}))^T \times R(Conv(x_{in})) \quad (3)$$

Here R and R' represents the opposite operation, R denotes the reshaping of the original input feature vector to its original size $R^{C \times N}$, $N = H \times W$, Then, the obtained weights are combined with the input data to produce an enhanced feature map output, as shown below.

$$x_{out} = x_{in} + R'(R(Conv\,(x_{in})) \times P^T) \quad (4)$$

Subsequently, the model employs the Multi-Scale Attention Feature Extraction module to further extract features from the data, as specifically illustrated in Figure 2.

After the computations are completed, each layer of the decoder will pass through a corresponding 1x1 segmentation head to generate output segmentation maps. There are a total of four layers, and these segmentation maps of different sizes will ultimately be upsampled using an interpolation function to produce the final segmentation result.

3.3. Multi-Axis Vision Transformer Encoder

Due to the recent tremendous potential demonstrated by Transformers in the field of medical image segmentation, MaxViT, as one of them, has achieved good results in the medical image domain due to its unique design. Networks such as G-CASCADE and EMCAD have utilized MaxViT pre-trained models as encoders and constructed networks for medical image segmentation. Building on previous successful experiences, in our network, we adopt MaxViT as the encoder. Specifically, we use a pre-trained model with an input size of 256x256, and in the four stages of downsampling, we embed feature vectors of sizes [96, 192, 384, 768]. Each layer employs [2, 2, 5, 2] MaxViT blocks, respectively. Additionally, in the four encoder layers, we use skip connections to pass the feature representations to the corresponding decoders, enabling the corresponding decoder segmentation heads to achieve precise segmentation of the results. The final segmentation results are as follows:

$$SegOut = p_1 + p_2 + p_3 + p_4 \quad (5)$$

3.4. Loss Function

In our model, we follow previous research[25] by utilizing activation functions and training the model by mixing features from multiple stages. The advantage of this approach is that it enables the model to converge better and is also beneficial for medical image segmentation tasks. According to the loss function we use during the training phase, we generate $2^n - 1$ non-empty subsets containing n prediction maps, and aggregate all non-empty subsets to produce the final prediction maps used for calculating the loss value. We then compute both the DICE loss and the cross-entropy loss for these generated prediction maps. The final loss function equation is as follows:

$$Loss = 0.3L_{dice} + 0.7L_{ce} \quad (6)$$

Where 0.3 and 0.7 are the weights for the DICE and CE loss functions, respectively.

4. Experiments and Results

4.1. Datasets

ACDC Dataset: We used both the ACDC dataset and the Synapse dataset to train our model. The ACDC dataset contains 100 patient cases, each including annotations for the left ventricle (LV), right ventricle (RV), and myocardium (Myo). During model training, we split the dataset into a training set, validation set, and test set in a 7:1:2 ratio.

Synapse Dataset: The Synapse dataset comprises 30 abdominal CT scans, totaling 3779 abdominal CT axial contrast-enhanced slices. Our dataset setup follows a similar approach to previous work with TransUNet, where we randomly divided the dataset into 18 cases for training and 12 for validation. We segmented eight organs in the results:

Table 1. The segmentation results on Synapse. We labeled the segmentation results for each individual organ using DICE, and we also labeled the average value of the segmentation results with DICE. The results of the various comparison models are referenced from previous research.Among the various indicators, the bold font represents the best, and the underlined represents the second best, higher DICE indicates a better performance, while a lower HD95 indicates a better performance as well.

Methods	DICE↑	HD95↓	Aorta	GB	KL	KR	Liver	PC	SP	SM
UNet[26]	70.11	44.69	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96
AttnUNet[20]	71.70	34.47	82.61	61.94	76.07	70.42	87.54	46.70	80.67	67.66
R50+UNet[7]	74.68	36.87	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50+AttnUNet[7]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
SSFormer[32]	78.01	25.72	82.78	63.74	80.72	78.11	93.53	61.53	87.07	76.61
PolypPVT[9]	78.08	25.61	82.34	66.14	81.21	73.78	94.37	59.34	88.05	79.4
TransUNet[7]	77.61	26.9	86.56	60.43	80.54	78.53	94.33	58.47	87.06	75.00
SwinUNet[4]	77.58	27.32	81.76	65.95	82.32	79.22	93.73	53.81	88.04	75.79
MT-UNet[31]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
MISSFormer[13]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
PVT-CASCADE[21]	81.06	20.23	83.01	70.59	82.23	80.37	94.08	64.43	90.1	83.69
TransCASCADE[21]	82.68	17.34	86.63	68.48	**87.66**	84.56	84.43	65.33	**90.79**	83.52
PVT-EMCAD[24]	83.16	**15.68**	88.20	73.34	84.28	82.13	94.76	68.52	90.08	83.96
HIF-Net(our)	**83.81**	18.63	**88.38**	**74.17**	86.12	83.59	**95.25**	**68.53**	90.43	**84.06**

Figure 3. As shown in the figure, we have visualized the results of the model on Synapse, displaying images from the same case but obtained from different models. GT represents the ground truth segmentation results. We used different colors to represent the organs, including the aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach, from left to right.

aorta, left kidney, right kidney, liver, pancreas, stomach, and spleen.

ISIC Dataset: We evaluated our model on the ISIC2017 and ISIC2018 datasets. During training, we randomly split the datasets into training, validation, and test sets in a 8:1:1 ratio. To ensure that our results are more realistic and accurate, we validated our results on the ISIC dataset using five-fold cross-validation.

4.2. Experiments Details

We conducted our experiments on an NVIDIA A100 GPU, using PyTorch 1.11.0 to implement model training. We initialized the model's encoder with MaxViT weights pre-trained on the ImageNet dataset. For the Synapse dataset, we trained the model for a total of 800 epochs. During training, we used the AdamW optimizer with a learning rate set to 0.0001 and weight decay set to 0.0001. As the loss function during training, we used a weighted combination of the DICE and CrossEntropy functions. For all of the ACDC, ISIC and Synapse datasets, we resized the images to 256x256 for input and set the batch size to 12. Additionally, we applied various data augmentation techniques to the datasets, such as rotation, scaling, and translation.

4.3. Evaluation Metrics

For the ACDC dataset, we used DICE as the final evaluation metric. For the Synapse dataset, we assessed the model's final performance using both DICE and HD95 (95% percentile Hausdorff distance). On the ISIC dataset, we evaluated the model based on DICE, Specificity (SP), Sensitivity (SE), and Accuracy (ACC). The calculation formulas for DICE and HD95 are as follows:

$$DICE = \frac{2 \mid X \cap Y \mid}{\mid X \mid + \mid Y \mid} \tag{7}$$

$$\begin{aligned} HD(X,Y) &= max\{d_{XY}, d_{YX}\} \\ &= max\left\{\max_{x\in X}\min_{y\in Y} d(x,y), \max_{y\in Y}\min_{x\in X} d(x,y)\right\} \end{aligned} \tag{8}$$

Here, X and Y represent the real image and the predicted segmentation result image respectively.

On the ISIC 2017 and 2018 datasets, we used DSC, ACC, SP, and SE. Among them, ACC stands for Accuracy, which is a commonly used indicator to measure the performance of a classifier. It represents the proportion of pixels

Table 2. This table presents a comparison of our proposed method with recent similar methods on the ISIC2017 and ISIC2018 datasets. We employed five-fold cross-validation to obtain the results presented here, with the best results bolded and the second best results underlined in the table.

Methods	ISIC2017				ISIC2018			
	DICE	SE	SP	ACC	DICE	SE	SP	ACC
U-Net[26]	0.8159	0.8172	0.9680	0.9164	0.8545	0.8800	0.9697	0.9404
AttU-Net[20]	0.8082	0.7998	0.9776	0.9145	0.8566	0.8674	0.9863	0.9376
TransNorm[1]	0.8933	0.8535	**0.9859**	0.9582	0.8951	0.8750	<u>0.9790</u>	0.9519
Swin-Unet[4]	0.8845	0.8893	<u>0.9778</u>	0.9476	0.8946	0.9056	**0.9798**	<u>0.9645</u>
PVT-GCASCADE[21]	<u>0.9089</u>	<u>0.9380</u>	0.9732	<u>0.9684</u>	<u>0.9007</u>	**0.9333**	0.9662	0.9609
PVT-EMCAD[24]	0.9006	0.9370	0.9682	0.9656	0.8974	0.9243	0.9675	<u>0.9617</u>
HIF-Net(our)	**0.9155**	**0.9396**	0.9759	**0.9706**	**0.9070**	<u>0.9300</u>	0.9709	**0.9683**

ISIC.pdf

Figure 4. Qualitative comparison of our method on the ISIC2017 dataset, where blue represents the ground truth boundary and green indicates the predicted boundary. Compared to previous methods, our method is able to segment the boundary more accurately.

that are correctly classified by the model out of all the pixels. The formula for calculating ACC is:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

SP stands for Specificity, and its value range is also between 0 and 1. A higher value of SP indicates that the model has stronger ability to recognize background or non-target regions. The formula for calculating SP is as follows:

$$SP = \frac{TN}{TN + FP} \tag{10}$$

SE stands for Sensitivity, also known as Recall. It measures the completeness of the model's recognition of the target region. The formula for calculating SE is:

$$SE = \frac{TP}{TP + FN} \tag{11}$$

4.4. Comparative Analysis of Results

We compared our proposed network with other convolutional neural network-based methods and recent state-of-the-art (SOTA) methods on the Synapse, ACDC, and ISIC datasets. Table 1 presents the comparative results of our network against CNN-based and Transformer-related methods on the Synapse multi-organ segmentation dataset, our network achieved excellent results compared to the comparison methods, indicating that our method can clearly identify organs in the images and provide more refined segmentation maps. Additionally, we visualized the results on the Synapse dataset as shown in Figure 3. In medical image segmentation, segmenting small organs is often challenging due to their complex structures and ambiguous boundaries with other organs. However, our proposed method can better understand and identify the shapes and boundaries of small organs, thereby improving the segmentation results.

Tables 2 and 3 respectively detail the segmentation performance of our model on the ISIC skin disease dataset and the ACDC dataset. From the results, it can be seen that our model has demonstrated excellent performance on both datasets. On the ACDC dataset, our method also performed exceptionally well, achieving remarkable performance. Specifically, our model achieved an improvement of 0.12 in the average DICE metric, significantly surpassing existing methods. Furthermore, our method also delivered significantly better results in the segmentation tasks of the left ventricle (LV), right ventricle (RV), and left ventricular myocardium (MYO), fully demonstrating the superior abil-

Table 3. The segmentation results on the ACDC dataset. RV stands for the right ventricle, Myo represents the myocardium, and LV represents the left ventricle. The bold font indicates the best performance, while the underlined indicates the second best.

Methods	DICE	RV	Myo	LV
R50+UNet[7]	87.55	87.10	80.63	94.92
R50+AttnUNet[7]	86.75	87.58	79.20	93.47
ViT+CUP[7]	81.45	81.46	90.71	92.18
R50+ViT+CUP[7]	87.57	86.07	81.88	94.75
TransUNet[7]	89.71	86.67	87.27	95.18
SwinUNet[4]	88.07	85.77	84.42	94.03
MT-UNet[31]	90.43	86.64	89.04	95.62
MISSFormer[13]	90.86	89.55	88.04	94.99
PVT-CASCADE[21]	91.46	89.97	88.90	95.50
Cascaded MERIT[21]	91.85	90.23	89.53	95.80
PVT-EMCAD[24]	<u>92.12</u>	<u>90.65</u>	<u>89.68</u>	**96.02**
HIF-Net(our)	**92.34**	**91.18**	**89.84**	<u>96.00</u>

ity of our model in detailed processing.

Table 2 displays the outcomes of our method on the ISIC 2017 and ISIC 2018 datasets. We compared our method with recent approaches and classic medical image segmentation models from the past. The table highlights that our model has achieved satisfactory results. Specifically, the excellent outcomes of our model across different types of datasets demonstrate its robust generalization ability.

4.5. Ablation Studies

Table 4 presents the results of ablation experiments conducted on the components of our model. From the results, we observe that after incorporating Multi-Scale Feature Extraction (MSFE) into the baseline model, there is a slight improvement in the DICE and LV scores, indicating that MSFE aids in capturing more detailed features at different scales. Subsequently, the addition of Coordinate Attention (note: originally referred to as Cross-Domain Attention but corrected here for clarity, assuming it was a typographical error) led to an increase in the scores, suggesting that it enhances the model's sensitivity and capture ability for positional information. We then tested CDA alone, and the results validated the rationality of using CDA. When MSFE and CDA are used simultaneously, the segmentation metrics further improve, demonstrating that these two components are complementary during model training. Finally, we incorporated the Cross-Attention Mechanism , and the metrics were further enhanced. Meanwhile, the results for Myo and LV remained competitive, indicating that the Cross-Attention Mechanism further boosts the model's ability to perceive important information in the image. This demonstrates that all the components we proposed contribute to the final segmentation results.

Table 4. Ablation experiments on the model using the ACDC dataset. Bold numbers indicate the best results, and underlined numbers represent the second-best results.

Method	DICE	RV	Myo	LV
Baseline	92.19	90.82	89.76	95.99
MSFE	92.27	90.84	<u>89.91</u>	96.07
CDA	92.25	<u>91.00</u>	89.85	95.59
CDA+MSFE	<u>92.30</u>	90.81	**90.01**	**96.09**
CDA+MSFE+CCA	**92.34**	**91.18**	89.84	<u>96.00</u>

5. Conclusion

We introduce a medical image segmentation network named HIFNet, which leverages dilated convolutions with varying parameters to capture broader spatial context, thereby enhancing detail perception and segmentation accuracy. Additionally, to bolster the model's capacity for capturing local information, we utilize a pre-trained MaxViT as the encoder, effectively extracting local features from the data. To optimize the capture of sensitive information, we incorporate cross-shaped attention, which is pivotal for understanding complex anatomical structures. Coordinate attention further enables the model to focus on key regions, thereby improving the localization precision of lesions or organs. Our model exhibits outstanding performance on multiple datasets, surpassing other advanced methods in terms of accuracy and segmentation fineness.

Additionally, our model boasts excellent scalability. In the future, we plan to conduct further experiments by applying different pre-trained encoders to the decoder we proposed. We will also experiment with medical images of various modalities to further demonstrate its generalization capability.

Acknowledgements

This work was supported in part by the following: the Joint Fund of the National Natural Science Foundation of China under Grant No. U24A20219, the National Natural Science Foundation of China under Grant No. 62272281, the Special Funds for Taishan Scholars Project(tsqn202306274), and the Youth Innovation Technology Project of Higher School in Shandong Province under Grant No. 2023KJ212.

References

[1] R. Azad, M. T. Al-Antary, M. Heidari, and D. Merhof. Transnorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model. *IEEe Access*, 10:108205–108215, 2022. 6

[2] R. Azad, A. R. Fayjie, C. Kauffmann, I. Ben Ayed, M. Pedersoli, and J. Dolz. On the texture bias for few-shot cnn segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2674–2683, 2021. 1

[3] Y. Cai and Y. Wang. Ma-unet: An improved version of unet based on multi-scale and attention mechanism for medical image segmentation. In *Third international conference on electronics and communication; network and computer technology (ECNCT 2021)*, volume 12167, pages 205–211. SPIE, 2022. 2

[4] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022. 3, 5, 6

[5] B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. W. K. Kong. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023. 2

[6] C.-F. R. Chen, Q. Fan, and R. Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 2

[7] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2, 3, 5, 6

[8] S. Chen, X. Tan, B. Wang, and X. Hu. Reverse attention for salient object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 234–250, 2018. 1

[9] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021. 2, 3, 5

[10] A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[11] 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. 1, 2

[12] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. 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. IEEE, 2020. 1, 2

[13] X. Huang, Z. Deng, D. Li, and X. Yuan. Missformer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*, 2021. 5, 6

[14] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 3

[15] 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

[16] Y. Jiang, Z. Zhang, S. Qin, Y. Guo, Z. Li, and S. Cui. Apaunet: axis projection attention unet for small target in 3d medical segmentation. In *Proceedings of the Asian Conference on Computer Vision*, pages 283–298, 2022. 2

[17] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022. 3

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 3

[19] A. Lou, S. Guan, and M. Loew. Dc-unet: rethinking the u-net architecture with dual channel efficient cnn for medical image segmentation. In *Medical Imaging 2021: Image Processing*, volume 11596, pages 758–768. SPIE, 2021. 1, 2

[20] 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*. 1, 2, 5, 6

[21] M. M. Rahman and R. 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. 2, 5, 6

[22] M. M. Rahman and R. Marculescu. G-cascade: Efficient cascaded graph convolutional decoding for 2d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7728–7737, 2024. 3

[23] M. M. Rahman and R. Marculescu. Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. In *Medical Imaging with Deep Learning*, pages 1526–1544. PMLR, 2024. 2

[24] M. M. Rahman, M. Munir, and R. 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. 3, 5, 6

[25] M. M. Rahman, S. Shokouhmand, S. Bhatt, and M. Faezipour. Mist: Medical image segmentation transformer with convolutional attention mixing (cam) decoder. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 404–413, 2024. 4

[26] 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, 2, 5, 6

[27] M. Soltani-Gol, M. Fattahi, H. Soltanian-Zadeh, and S. Sheikhaei. Drau-net: Double residual attention mechanism for automatic mri brain tumor segmentation. In *2022 30th International Conference on Electrical Engineering (ICEE)*, pages 587–591. IEEE, 2022. 1

[28] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022. 2, 3

[29] A. Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

[30] H. Wang, P. Cao, J. Wang, and O. R. Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2441–2449, 2022. 2

[31] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X.-H. Han, Y.-W. Chen, and R. Tong. Mixed transformer u-net for medical image segmentation. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2390–2394. IEEE, 2022. 5, 6

[32] 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. 5

[33] M. Wang, H. Wang, and F. Zhang. Famc-net: Frequency domain parity correction attention and multi-scale dilated convolution for time series forecasting. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2554–2563, 2023. 2

[34] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 2

[35] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 2

[36] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022. 2

[37] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 1

[38] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021. 2

[39] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 2

[40] G. Xu, X. Zhang, X. He, and X. Wu. Levit-unet: Make faster encoders with transformer for medical image segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 42–53. Springer, 2023. 3

[41] C. Yao, M. Hu, Q. Li, G. Zhai, and X.-P. Zhang. Transclaw u-net: claw u-net with transformers for medical image segmentation. In *2022 5th International Conference on Information Communication and Signal Processing (ICICSP)*, pages 280–284. IEEE, 2022. 3

[42] C. You, R. Zhao, F. Liu, S. Dong, S. Chinchali, U. Topcu, L. Staib, and J. Duncan. Class-aware adversarial transformers for medical image segmentation. *Advances in Neural Information Processing Systems*, 35:29582–29596, 2022. 3

[43] F. Zhang, G. Chen, H. Wang, J. Li, and C. Zhang. Multi-scale video super-resolution transformer with polynomial approximation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4496–4506, 2023. 2

[44] F. Zhang, G. Chen, H. Wang, and C. Zhang. Cf-dan: Facial-expression recognition based on cross-fusion dual-attention network. *Computational Visual Media*, pages 1–16, 2024. 2

[45] F. Zhang, T. Guo, and H. Wang. Dfnet: Decomposition fusion model for long sequence time-series forecasting. *Knowledge-Based Systems*, 277:110794, 2023. 2

[46] X. Zhang, Z. Feng, T. Zhong, S. Shen, R. Zhang, L. Zhou, B. Zhang, and W. Wang. Dra u-net: An attention based u-net framework for 2d medical image segmentation. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3936–3942. IEEE, 2021. 2

[47] 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. 3

[48] P. Zhao, J. Zhang, W. Fang, and S. Deng. Scau-net: spatial-channel attention u-net for gland segmentation. *Frontiers in Bioengineering and Biotechnology*, 8:670, 2020. 1

[49] 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, 2