information

MDPI

Article

Hybrid U-Net Model with Visual Transformers for Enhanced Multi-Organ Medical Image Segmentation

Pengsong Jiang [1], Wufeng Liu [2,*], Feihu Wang [1] and Renjie Wei [1]

[1] College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China; 2022930956@stu.haut.edu.cn (P.J.); 2023930856@stu.haut.edu.cn (F.W.); 2023930863@stu.haut.edu.cn (R.W.)
[2] School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou 450001, China
* Correspondence: lwf@haut.edu.cn

Abstract: Medical image segmentation is an essential process that facilitates the precise extraction and localization of diseased areas from medical pictures. It can provide clear and quantifiable information to support clinicians in making final decisions. However, due to the lack of explicit modeling of global relationships in CNNs, they are unable to fully use the long-range dependencies among several image locations. In this paper, we propose a novel model that can extract local and global semantic features from the images by utilizing CNN and the visual transformer in the encoder. It is important to note that the self-attention mechanism treats a 2D image as a 1D sequence of patches, which can potentially disrupt the image's inherent 2D spatial structure. Therefore, we utilized the structure of the transformer using visual attention and large kernel attention, and we added a residual convolutional attention module (RCAM) and multi-scale fusion convolution (MFC) into the decoder. They can help the model better capture crucial features and fine details to improve detail and accuracy of segmentation effects. On the synapse multi-organ segmentation (Synapse) and the automated cardiac diagnostic challenge (ACDC) datasets, our model performed better than the previous models, demonstrating that it is more precise and robust in multi-organ medical image segmentation.

Keywords: deep learning; transformer; convolutional neural networks; multi-organ medical image segmentation

Academic Editor: Andreas Triantafyllidis

Received: 2 December 2024
Revised: 13 January 2025
Accepted: 3 February 2025
Published: 6 February 2025

Citation: Jiang, P.; Liu, W.; Wang, F.; Wei, R. Hybrid U-Net Model with Visual Transformers for Enhanced Multi-Organ Medical Image Segmentation. *Information* **2025**, *16*, 111. https://doi.org/10.3390/info16020111

1. Introduction

Multi-organ medical image segmentation aims to accurately extract and separate areas of interest or abnormal regions from the medical images [1]. With the help of computer-aided systems to identify and segment lesion areas, doctors can obtain the diagnostic evidence that can effectively assist them in making more accurate diagnoses and improving work efficiency [2,3]. Traditional methods for multi-organ medical image segmentation primarily depend on computer vision techniques, such as edge detection [4], region growing [5], and thresholding [6]. These methods often require manual features and rule definitions, which limits their effectiveness in handling complex medical image tasks. In addition, they also require significant human and time resources, as well as expertise from the professionals involved [7].

In recent years, the rapidly developing field of deep learning [8] has prompted an increasing number of scholars to start utilizing convolutional neural networks (CNNs) [9] to handle tasks related to medical image segmentation, and they have gradually become a breakthrough technique for medical image processing. CNNs have the advantage of quickly learning and extracting features from original images without having to manually

design and select them [10]. Compared with traditional machine learning methods, a fully convolutional network (FCN) [11] is a special CNN structure that can accept inputs of any size, and produce corresponding outputs. It can also perform feature extraction and classification simultaneously, which demonstrates its superior performance.

In terms of network structure innovation, Valanarasu et al. [12] proposed a multilayer perceptron (MLP)-based image segmentation network, that can diminish the number of parameters and computational complexity, to some extent, by using tokenized MLP blocks. Chen et al. [13] proposed DRINet, which combines residual inception blocks, dense connection blocks, and unpooling blocks to significantly increase the depth and width of the network, while effectively managing the parameter space using the growth rate. In 3D multi-organ medical image segmentation, Milletari et al. [14] proposed an innovative approach, based on volume, for fully convolutional neural networks, which produced excellent results in MRI segmentation tasks. Meanwhile, architectures like U-Net [15] have become increasingly popular in this field, and are extensively used for segmentation tasks. Wang et al. [16] proposed a lightweight U-Net network model that combines the advantages of multi-task learning and deformable convolutional. Parvez et al. [17] added a hierarchical module to the model to extract and merge features for obtaining multi-scale information from medical images. Although CNNs have achieved some good results on medical images, they also have some significant limitations. Because CNNs are focused on extracting local features, they have limited potential to obtain global contextual information from deeper representations [18,19]. As a result, they may not perform well when conducting multi-organ medical image segmentation tasks involving long-range dependencies.

In contrast, the transformer architectures utilize self-attention mechanisms to achieve global dependencies within input sequences [20]. This way, the transformer can capture longer-range contextual information, which can be extremely beneficial for tasks such as sequence generation and natural language processing [21,22]. For non-local relationships, transformers are particularly well-suited in scenarios where the contextual information is of paramount importance, such as in the field of medical image processing. Therefore, some researchers began to introduce transformers into the domain of image processing, and achieved favorable outcomes at the time. Subsequently, with the integration of transformers in addition to CNN, it is possible to elevate multi-organ medical image segmentation to a new level. Consequently, TransUNet [23] has become an advanced model that takes the advantages of the U-Net architecture and transformer-based models. It can achieve powerful feature representation and accurately capture critical features through integrating the transformer into the U-Net framework. However, the self-attention mechanism in the transformer has some shortcomings in image processing tasks. It typically treats a two-dimensional image as a one-dimensional sequence, which disrupts the inherent structure of the original image [24]. Consequently, the objective is to address this matter and optimize the quality of segmented images, so we aim to integrate the visual transformer into the U-Net model architecture.

In this paper, we propose a novel model that combines the U-Net with the visual transformer. We attempt to replace the self-attention mechanism with a visual-attention mechanism. Additionally, we use residual convolutional attention module (RCAM) and multi-scale fusion convolution (MFC) in the decoder to recover the details and structure of the original data. As a result of these modifications, the model now focuses on significant regions and improves generalization. This article mainly consists of the following contributions:

1. For the encoder, we adopt the visual transformer with a large kernel attention, which effectively combines the benefits of self-attention and convolution. Additionally, it accounts for the spatial relationships between different positions.

2. We utilize the RCAM after the up-sampling of the decoder, which can enable the model to concentrate on crucial channels and spatial locations. It can also improve the model's capacity to refine feature maps and capture important spatial and channel information.
3. After the encoder and decoder complete the information fusion process, we utilize the MFC to expand the receptive field. Each layer incorporates a convolutional layer with varying dilation rates, which facilitates the capture of more contextual information from the decoder.

The structure of this paper is as follows: Section 2 presents a thorough review of the existing literature in the field, providing an in-depth analysis of the significant contributions and advancements made by previous scholars. Section 3 presents our proposed model, which includes several innovations aimed at enhancing the existing model structure. Section 4 describes the experiment setup in detail, including the datasets, evaluation metrics, and the results obtained from our experiments. Section 5 discusses the limitations of our study and future perspectives. Section 6 draws conclusions.

2. Related Work

2.1. UNet-Based Methods

Convolutional neural networks (CNNs) have gained significant popularity and have been extensively utilized in recent years [25,26]. Among them, U-Net [15] has gradually become an important method for segmenting multi-organ medical images. Due to its simple structure and superior performance, a series of network models based on U-Net variants have been proposed. These network models include ResUNet [27], UNet++ [28], UNet3+ [29], and DC-UNet [30]. These improved models use new structures, connection methods, and operations to improve the detail presentation and accuracy of segmentation results. Meanwhile, they also have optimized aspects such as network depth, feature fusion, and parameter utilization. These models have been extensively utilized in the domain of medical imaging processing, yielding a multitude of noteworthy outcomes. Specifically, in the multi-region segmentation of the skeletal muscle, Kawamoto et al. [31] utilized U-Net to simultaneously learn multiple muscle regions, achieving good segmentation results for skeletal muscle point segmentation. Ashino et al. [32] employed a multi-class learning approach to segment the sternocleidomastoid muscle and skeletal muscle joints.

2.2. Transformer-Based Methods

Initially, the transformer was utilized for tasks pertaining to natural language processing [22]. In a variety of tasks, the transformer framework has achieved remarkable results, largely due to its parallel computation capabilities, its proficiency in modeling long-range dependencies, and its capacity to capture global features [33,34]. Alexey et al. [35] proposed a transformer-based image classification model based on transformer success. Subsequently, a vision transformer (ViT) has been successfully implemented in the field of computer vision. It treats images as a series of split image patches and transforms these patches into sequential data to be input into the transformer model.

To address the issue of ViT requiring more computational resources and memory to process images of larger size, researchers have proposed some improvements. One such improvement is the pyramid vision transformer (PVT) [36], which introduces variant models capable of handling large-scale images. In addition, the convolutional vision transformer (CVT) [37] reduces computational complexity and improves model scalability. Afterwards, Liu et al. [38] proposed the swin transformer, which introduces a shift window mechanism. This allows the model to process semantic information in the image with a greater efficacy and to mitigate the issue of information loss that is inherent to the traditional

uniform division of image blocks. Li et al. [39] proposed to combine the context pyramid mechanism with the transformer to improve the segmentation accuracy of the model. Yao et al. [40] proposed a dual vision transformer, which can help the models improve efficiency and reduce complexity.

2.3. Combining U-Net with Transformer-Based Methods

In order to more effectively extract both local and global contextual information, some researchers have started combining CNN with transformers to achieve improved performance and results in the multi-organ segmentation of medical images. Chen et al. [23] proposed the TransUNet model, based on the ViT architecture and utilizing a combination of CNN and ViT as its encoder. Similarly, Wang et al. [41] proposed the use of a mixed transformer module with the U-Net structure to address the potential relevance of the dataset. Jiang et al. [42] utilized a visual attention-based transformer as an encoder and the CNN as a decoder to directly input images into the transformer. To facilitate the extraction of both coarse and fine-grained feature representations at varying semantic scales, Lin et al. [43] proposed a dual-scale encoding mechanism that makes use of the dual-scale encoder, based on the swin transformer. Peng et al. [44] combined the transformer with the RNN to ensure efficient operation during training.

Unlike these methods, we attempt to use visual attention to replace the MSA mechanism in the ViT. Since MSA usually only captures spatial correlations, it may ignore correlations in the channel dimension. The large kernel attention in visual attention can adapt to both spatial dimensions and channel dimensions. Moreover, we discover that the decoder may lose the crucial information and reduce the resolution as the spatial dimensions expand during the up-sampling process. Therefore, we propose adding the residual convolutional attention module after the up-sampling and a three-layer MFC after the encoder and decoder complete information fusion to enhance the clarity and accuracy of segmentation results.

3. Methods

3.1. Architecture Overview

The structure of our proposed network model is illustrated in Figure 1, where the three main components are included: encoder, decoder, and skip connections. In this instance, it is assumed that the input image is $x \in R^{W \times H \times C}$, with dimensions W $\times$ H, and a total of C channels.

The detailed structure of each layer of VT-UNet is shown in Table 1. The initial stage of the process involves the original image being fed into the encoder module, whereby features are retrieved from it using CNN and transformer. CNN is employed within the encoder to capture and extract low-level features from the input image. This is accomplished by gradually reducing the spatial dimensions of the feature maps while increasing the number of channels. Subsequently, these features can be extracted at diverse scales and levels of abstraction, through operations such as convolution and pooling layers. Unlike CNN, the transformer is primarily utilized for high-level feature modeling and semantic understanding. The output of the CNN is processed by transformer layers in the encoder. The transformer can utilize visual attention and the large kernel attention mechanism to establish the dependencies between global contexts, enabling the learning of global semantic information from images. Through visual attention, the transformer mechanism can better comprehend the relationships and interactions among various regions in the image. Consequently, utilizing the CNN and visual transformer architectures from the images, the encoder can successfully extract and encode both high-level semantic information and

low-level features. This way, the encoder provides more contextual information to the decoder and enhances the accuracy of proposed model.

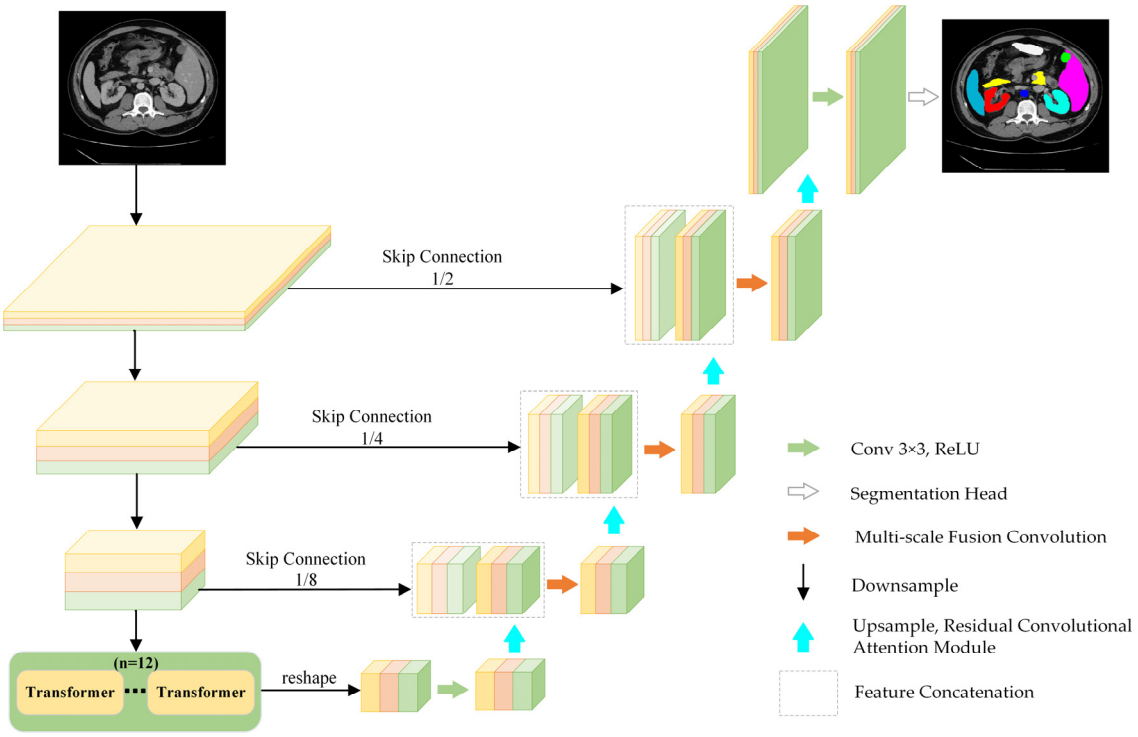


Figure 1. Overview of the proposed network structure.

Table 1. Architecture of the proposed model.

Architectures	Layers	Proposed Model	Feature Size
-	input	-	$224 \times 224 \times 3$
	convolution	7×7 conv, padding 3, stride 2	$112 \times 112 \times 64$
	pooling	3×3 max pooling, stride 2	$55 \times 55 \times 64$
encoder	resnet50 block 1	$\begin{pmatrix} 1 \times 1 & \text{conv} \\ 3 \times 3 & \text{conv} \end{pmatrix} \times 3$	$56 \times 56 \times 256$
	transition layer1	1×1 conv	$56 \times 56 \times 256$
	resnet50 block 2	$\begin{pmatrix} 1 \times 1 & \text{conv} \\ 3 \times 3 & \text{conv} \end{pmatrix} \times 4$	$28 \times 28 \times 512$
	transition layer2	1×1 conv	$28 \times 28 \times 512$
	resnet50 block 3	$\begin{pmatrix} 1 \times 1 & \text{conv} \\ 3 \times 3 & \text{conv} \end{pmatrix} \times 9$	$14 \times 14 \times 1024$
	transition layer3	1×1 conv	$14 \times 14 \times 1024$
	embedding layer	1×1 conv, stride 1; flatten	196×768
	transformer layer	$\begin{pmatrix} 5 \times 5 & \text{conv} \\ 1 \times 1 & \text{conv} \end{pmatrix} \times 12$	196×768
-	reshape layer	expand; 3×3 conv, padding 1	$14 \times 14 \times 512$
decoder	up-sampling layer 1	2×2 up-sampling − [resnet50 block 1], conv	$28 \times 28 \times 512$
	up-sampling layer 2	2×2 up-sampling − [resnet50 block 2], conv	$56 \times 56 \times 256$
	up-sampling layer 3	2×2 up-sampling − [resnet50 block 3], conv	$112 \times 112 \times 64$
	up-sampling layer 4	2×2 up-sampling, conv	$224 \times 224 \times 16$
-	segmentation head layer	3×3 conv, padding 1	$224 \times 224 \times 2$

This paper proposes a hidden layer with a size of 768.

The decoder reconstructs the segmentation output using the encoded features, whereas the up-sampling process attempts to recover the spatial resolution of the feature maps. It gradually increases the spatial dimensions while reducing the channel dimensions, so as to ultimately obtain a size that is the same as the input image. We adopt the residual convolutional attention module after the up-sampling process to improve segmentation accuracy. Moreover, we use a three-layer multi-scale fusion convolution to broaden the receptive field and capture more contextual information. In the end, skip connections directly connect the corresponding layers of the encoder and decoder, allowing low-level features to be transmitted directly to the decoder. The features are input into the decoder module and fused with skip connections, it not only preserves detailed information but also improves the flow of information, increasing the performance and accuracy of the network. After the training and predicting with the model, we can obtain the predicted mask. Using these three components effectively, we can capture and use both global and local features from the original images, resulting in enhanced segmentation results.

3.2. Transformer Layer

After CNN processing in the encoder, the image is sent to the transformer layer to further extract and capture global features, which enables the model to gain a more comprehensive understanding of the overall information present in the image. It differs from ViT [35], which utilizes the multi-head self-attention module. A 2D image is treated as a 1D sequence which disregards the important 2D structure of the image. At the same time, it is a mechanism that only takes into account spatial dimension adaptation, while ignoring channel dimension adaptation. To address these issues, we attempt to utilize visual attention with large kernel attention. The advantage of large kernel attention includes self-attention and convolution, which can be used both to capture local features and global relationships. As illustrated in Figure 2, the improved transformer layer is primarily constituted of the layer norm (LN), multi-layer perceptron (MLP) block, and visual attention (VA) mechanism.

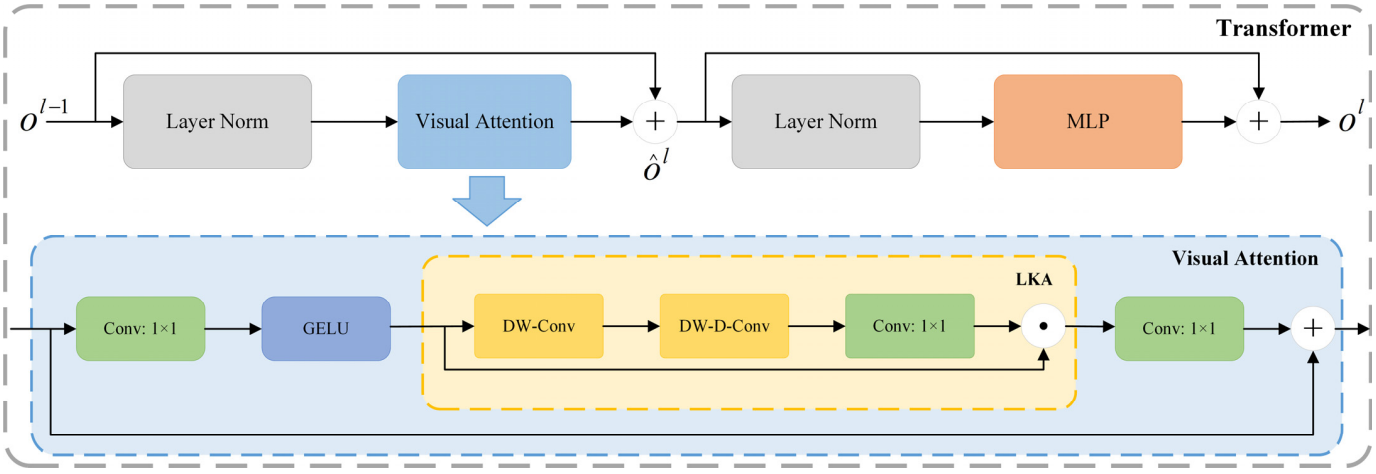


Figure 2. Layer schematic of visual transformers.

In the large kernel attention module, there are three main components: a depth-wise convolution (DW-Conv) layer that utilizes a 5 × 5 kernel size with a padding of 2; a depth-wise dilated convolution (DW-D-Conv) layer that employs a 7 × 7 kernel size with a padding of 9 and a dilation rate of 3; and finally, a 1 × 1 point-wise convolution. The convolution operations mentioned above can be performed in an orderly manner, so as to capture long-range relationships while minimizing computational costs and parameters.

Afterwards, the importance of each point can be evaluated, leading to the creation of a corresponding attention map. Therefore, the results are as follows:

$$LKA = Conv(DW\_D\_conv(DW\_Conv(F)), \tag{1}$$

$$Output = LKA \odot F, \tag{2}$$

where $F \in R^{H' \times W' \times C'}$ denotes the features of the input image, $LKA \in R^{H' \times W' \times C'}$ represents an attention map, and $\odot$ represents element-wise multiplication, where the values within the map indicate the importance of each feature. Consequently, the output of the l^{th} layer of the transformer can be written as follows:

$$\hat{o}^l = VA\left(LN\left(o^{l-1}\right)\right) + o^{l-1}, \tag{3}$$

$$o^l = MLP\left(LN\left(\hat{o}^l\right)\right) + \hat{o}^l, \tag{4}$$

where $\hat{o}^l$ and o^l are used to represent the output of the VA mechanism and MLP module of the l^{th} block, respectively.

3.3. Residual Convolutional Attention Module

To further enhance the accuracy of CNNs, the residual convolutional attention module (RCAM) can improve their attention to important features within an image. According to Figure 3, the RCAM comprises three primary components: spatial attention module, residual connection, and channel attention module.

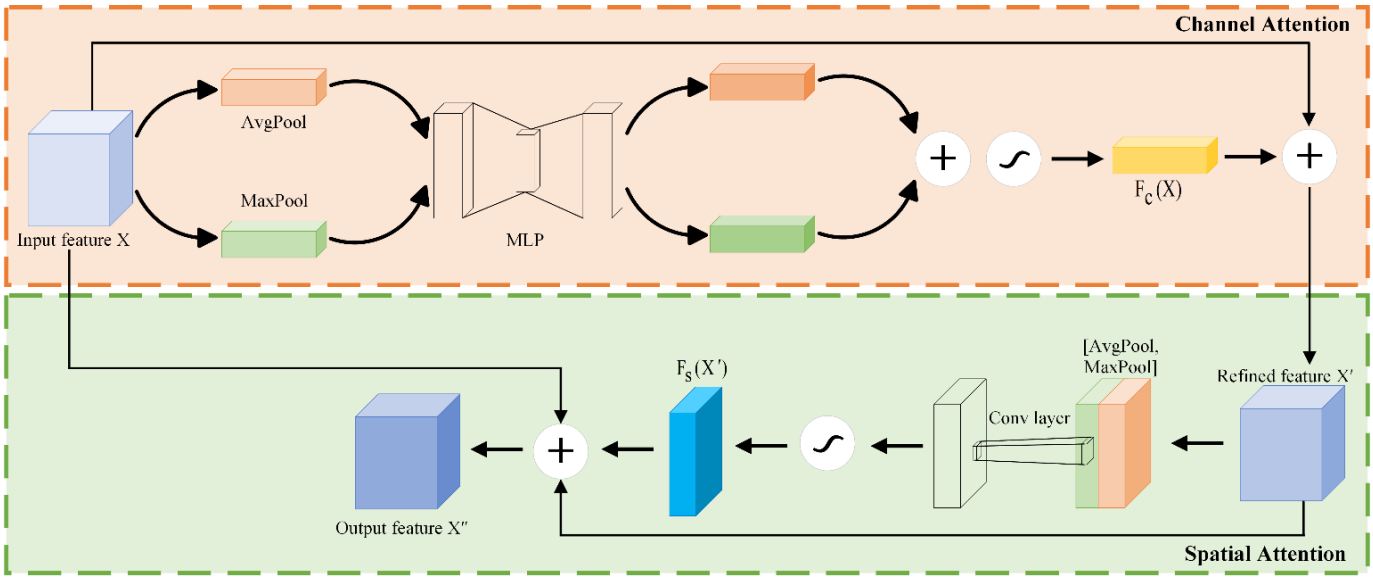


Figure 3. Configuration of the residual convolutional attention module.

In the channel attention module, weights are assigned to the different channels based on their importance in the feature map. In addition, it can be used to prioritize important channels and to extract more meaningful information from the network, as opposed to the spatial attention module, which is concerned with the significance of various spatial positions within the feature map. The spatial attention mechanism is applied to the feature map, and introduces an operation similar to the squeeze and excitation mechanism, which facilitates the capture of relevant information from each position in the feature map. Despite playing different roles in the model, both contribute to improving the ability to capture important features in an image. Moreover, residual connections allow inputs to

be passed directly to outputs, without passing through certain layers. It alleviates the problem of vanishing gradients, helping the model to accelerate the training process and enhance overall performance. As a result, gradients can be transmitted effectively during backpropagation, which promotes the formation of deeper network and enhances the speed of convergence.

Especially, RCAM generally modifies the weights of each channel using an attention mechanism applied to the input feature map, while preserving the spatial dimensions of the feature map. Unlike other methods that may involve resizing or reshaping the feature map, RCAM focuses solely on enhancing the feature representation through channel attention and spatial attention modules. These modules operate independently to selectively emphasize informative channels and spatial regions, allowing the model to refine its feature representations without altering the original spatial structure. Following input of feature $X \in R^{H' \times W' \times C'}$ to the channel attention module, feature $X^c_{avg} \in R^{C' \times 1 \times 1}$ is generated by average pooling, while feature $X^c_{max} \in R^{C' \times 1 \times 1}$ is generated by maximum pooling. Subsequently, these feature maps are fed into a shared multi-layer perceptron (MLP), where they undergo summation and activation processes. Ultimately, the feature map $F_c(X) \in R^{C' \times 1 \times 1}$ of channel attention is produced. The spatial attention mechanism takes the channel-attention-refined feature $X' \in R^{H' \times W' \times C'}$ as input. Similar to the previous method, $X^s_{avg} \in R^{1 \times H' \times W'}$ and $X^s_{max} \in R^{1 \times H' \times W'}$ are generated through two pooling operations. Following this, a series of processes, including concatenation, convolution, and activation, are applied to obtain $F_s(X') \in R^{H' \times W'}$. Therefore, the specific implementation can be written as follows:

$$\begin{aligned} F_c(X) &= \sigma(MLP(AvgPool(X) + MLP(MaxPool(X)) \\ &= \sigma\left(W_a\left(W_b\left(X^c_{avg}\right)\right) + W_a(W_b(X^c_{max}))\right) \end{aligned}, \tag{5}$$

$$\begin{aligned} F_s(X') &= \sigma(f_{7\times7}([AvgPool(X'); MaxPool(X')])) \\ &= \sigma\left(f_{7\times7}\left(\left[X'^s_{avg}; X'^s_{max}\right]\right)\right) \end{aligned}, \tag{6}$$

where W_a and W_b denote the weights of the shared MLP, $f_{7\times7}$ is a 7×7 convolution operation, and σ denotes the sigmoid activation function. As a result, the final output feature can be written as:

$$X''_{final} = X + X \cdot F_c(X) + X' \cdot F_s(X'), \tag{7}$$

where X''_{final} is the final output feature, X denotes the input feature, and X' represents the feature that passes through the channel attention output.

3.4. Multi-Scale Fusion Convolution

In order to capture more image details after information is fused in the decoder. We propose using multi-scale fusion convolution (MFC) to enhance the performance of image processing tasks. It can use convolutional kernels with multiple different dilation rates to extract features. This is because each convolutional kernel has a specific dilation rate, which allows control over the size of the receptive field. The process begins with parallel convolution operations on the input feature maps. Afterwards, the results of all parallel convolutions are fused together, which can preserve the local detailed information and expand the receptive field to acquire a wider range of features. This fusion of multi-scale features allows the model to capture more comprehensive and nuanced representations of the input image. Finally, the fused feature map is concatenated with the features from the encoder, integrating both the high-level contextual information and the fine-grained details.

MFC consists of three convolutional layers and is illustrated in Figure 4. In the model, the convolutional kernel size is set to 3 × 3, and the dilation rates for each layer are 1, 2, and 3. This setup allows the receptive field of each convolutional layer to gradually increase so that information at different scales can be efficiently captured. Among them, a standard convolutional kernel of size 3 is used in the first layer, which is considered a dilated convolution layer. Different dilation rates for various convolutional layers can ensure that more feature information is captured. Furthermore, to prevent the gridding effect, the selection of the dilation rate should adhere to the following principles:

$$M_n = max[M_{n+1} - 2r_n,\ M_{n+1} - 2(M_{n+1} - r_n), r_n], \tag{8}$$

where M_n is the maximum distance between two non-zero elements in layer n, and r_n denotes the dilation rate of layer n.

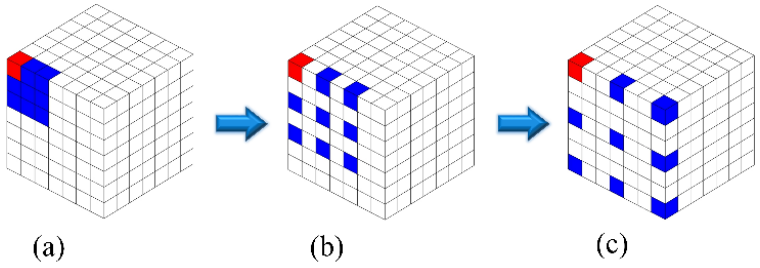


Figure 4. Multi-scale fusion convolution, where (**a**) is normal convolution, (**b**,**c**) represent dilated convolutions with dilation rates of 2 and 3, respectively.

4. Experiments and Results

4.1. Datasets

For this study, we chose two publicly accessible datasets to assess our proposed model: the synapse multi-organ segmentation (Synapse) dataset and the automated cardiac diagnosis challenge (ACDC) dataset. The detailed information for the two datasets is described below.

Specifically, the Synapse dataset contains 30 abdominal CT scan images, which specifically focus on 12 different abdominal organs. Following this [45,46], 18 scan images are allocated for training and randomly divided into 2212 axial slices. The remaining 12 scan images are designated for the test dataset. For the purpose of evaluating the effectiveness of our training program, we have selected eight abdominal organs to be assessed: the gallbladder, aorta, right kidney, left kidney, spleen, stomach, pancreas, and liver. Moreover, each organ has been annotated by experts, using different colors.

The ACDC dataset comprises the total examination results from MRI scans of 100 different patients. Each patient's scans have been annotated by specialized doctors, focusing particularly on the left ventricle (LV), myocardium (Myo), and right ventricle (RV). Furthermore, to ensure the comprehensiveness and reliability of the assessment, the dataset consisting of 100 cases was randomly divided into three groups: 70 for the training set, 10 for the validation set, and 20 for the test set.

4.2. Evaluation Metrics

We evaluated and validated the segmentation accuracy of the proposed model, utilizing the Synapse dataset and the ACDC dataset. These evaluations were based on two key

metrics: the average dice similarity coefficient (DSC) and the average Hausdorff distance (HD). The following sections detail these metrics:

$$\mathrm{DSC}(A, B) = 2\frac{|A \cap B|}{|A|+|B|}, \tag{9}$$

$$\mathrm{HD}(A, B) = \max\left(\max_{a\in A}\left\{\min_{b\in B}\|a-b\|\right\}, \max_{b\in B}\left\{\min_{a\in A}\|b-a\|\right\}\right) \tag{10}$$

where A and B represent the set of true labels and segmentation predictions, respectively, and a and b are the points adjacent to the background and target in the true labels and segmentation predictions.

4.3. Implementation Details

For the experiments, we used only simple data enhancement techniques to increase the dataset diversity, including random rotation and flipping, to enhance the generalization of these models. Our proposed network model is executed using Python 3.9, PyTorch 2.1.2, and CUDA 11.8. At the same time, all experiments were performed on an NVIDIA GeForce RTX 3060 GPU (NVIDIA, Santa Clara, CA, USA) with 12 GB of memory in order to ensure fairness. The Synapse and the ACDC datasets have input images of 224 × 224, and the batch size is 24. During the training period, all experiments were conducted for 150 epochs utilizing the SGD optimizer, with a momentum of 0.9, a weight decay of 0.0001, and a learning rate of 0.01.

4.4. Experiment Results

We conducted a comparison of the proposed model with previous models using the Synapse dataset, and these results are presented in Table 2. Regarding the evaluation metrics of our proposed model, the average DSC and HD are 81.77% and 18.02 mm, respectively. According to the results of the metrics DSC and HD, our proposed model outperforms previous approaches. Specifically, our model achieves the highest DSC scores for the Aorta, Kidney (L), and Liver, compared to the previous models. Furthermore, we selected several classic network models, such as UNet, TransUNet, and SwinUNet, and compared and visualized their prediction results with those of our proposed model. Figure 5 indicates that our proposed model obtains better segmentation results, proving the effectiveness of our approach.

Table 2. Comparison of segmentation results of different models on the Synapse dataset.

Methods	DSC	HD	Aorta	Gallbladder	Kidney(R)	Kidney(L)	Pancreas	Liver	Spleen	Stomach
V-Net [14]	68.81	-	75.34	51.87	80.75	77.10	40.50	87.74	80.56	56.98
U-Net [15]	76.85	39.70	89.07	69.72	68.60	77.77	53.98	93.43	86.67	75.58
Att-UNet [47]	77.77	36.02	89.55	68.88	71.11	77.98	58.04	93.57	87.30	75.75
TransUNet [23]	77.48	31.69	87.23	63.13	77.02	81.87	55.86	94.08	85.08	75.62
TransNorm [48]	78.40	30.25	86.23	65.10	78.63	82.18	55.34	94.22	89.50	76.01
MT-UNet [41]	78.59	26.59	87.92	64.99	77.29	81.47	59.46	93.06	87.75	76.82
SwinUNet [45]	79.13	21.55	85.47	66.53	79.61	83.28	56.58	94.29	90.66	76.60
BiFTransNet [49]	78.77	27.94	87.67	67.09	75.68	82.04	**60.93**	93.84	87.13	75.80
RFE-UNet [50]	79.77	21.75	87.32	65.40	81.92	84.18	59.02	94.34	89.56	76.45
CPFTransformer [39]	79.87	20.83	87.71	68.78	79.15	83.19	58.47	94.37	**90.35**	76.93
IEA-Net [51]	78.56	27.21	85.71	70.32	75.41	78.45	59.41	94.02	85.06	76.38
CCFNet [52]	81.59	**14.47**	88.35	**72.49**	**87.42**	83.50	56.89	95.37	88.19	**80.47**
Ours	**81.77**	18.02	**88.93**	71.84	82.95	**85.19**	60.48	**95.41**	89.81	79.56

The average DSC score % and average HD distance in mm, as well as the average DSC score % for each organ.

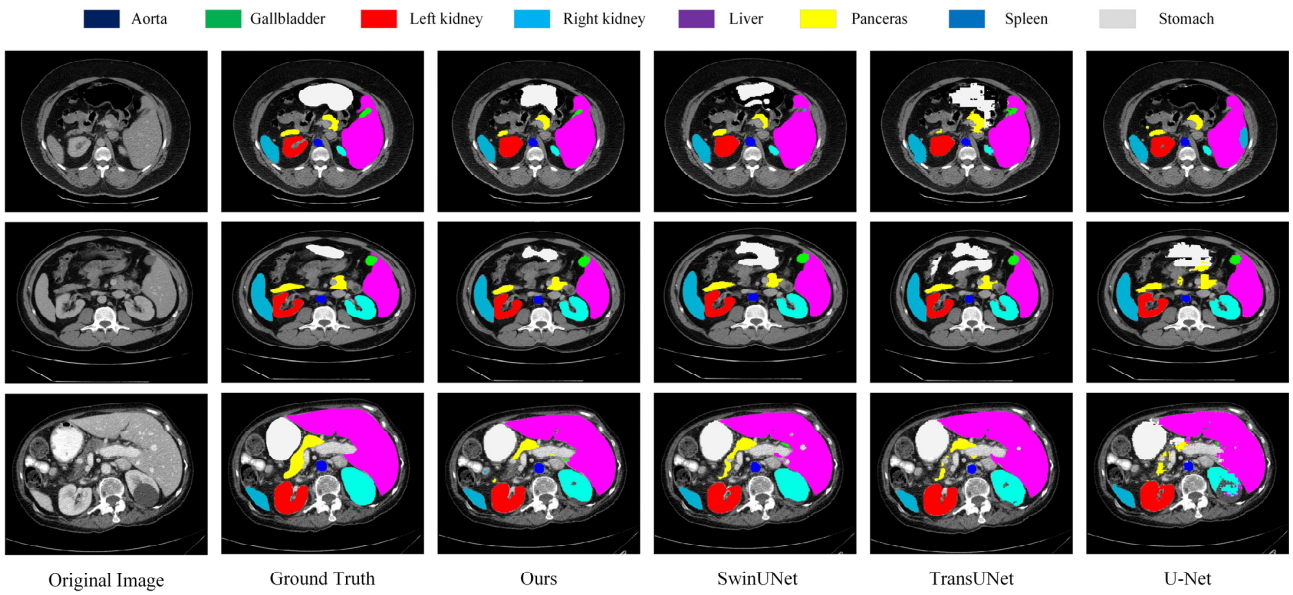


Figure 5. Comparison of model predictions on Synapse dataset.

In addition to performing image segmentation on the ACDC dataset, we also validated the efficacy of our proposed model. Since the ACDC dataset is relatively simple, the average HD results between different models show relatively small variations. Due to this, the evaluation metric in this experiment is only the average DSC, and the segmentation results are presented in Table 3. On the ACDC dataset, the proposed model has a DSC of 91.83%, which is higher than previous models. Our model has superior generalization capabilities in the Myo and LV organs than the other models. This result further proves the superior segmentation accuracy of our methodology. The comparison of segmentation results for the models is presented in Figure 6. Our proposed model provides more detail and accuracy in segmenting edge details than other models. Moreover, in order to more intuitively show how the loss values and dice coefficients of the model change with each epoch during the training and validation, Figure 7 shows the loss curves, as well as the DSC curves for the training and validation sets.

Table 3. Comparison of segmentation results of different models on the ACDC dataset.

Methods	DSC	Myo	RV	LV
U-Net [15]	87.55	80.63	87.10	94.92
Att-UNet [47]	86.75	79.20	87.58	93.47
TransUNet [23]	89.71	84.53	88.86	95.73
MT-UNet [41]	90.43	89.04	86.64	95.62
SwinUnet [45]	90.00	85.62	88.55	95.83
IEA-Net [51]	91.38	89.51	88.91	95.72
CCFNet [52]	91.07	89.30	**89.78**	94.11
Ours	**91.83**	**90.23**	89.35	**95.90**

Meanwhile, to further assess the model's performance in terms of computational complexity and resource cost, we compared the performance of the various models presented in Table 4, with the specific results detailed below. The inference time and GPU memory usage listed in the table are the measurements based on a single-image inference, calculated by averaging 12 images from the Synapse dataset test set. Due to the simple network structure of CNNs, they have significant advantages in terms of computational complexity, but there are certain limitations in segmentation accuracy. Even though our model may not achieve the lowest complexity and computational cost, its excellent performance in medical image segmentation highlights the effectiveness and value of our work.

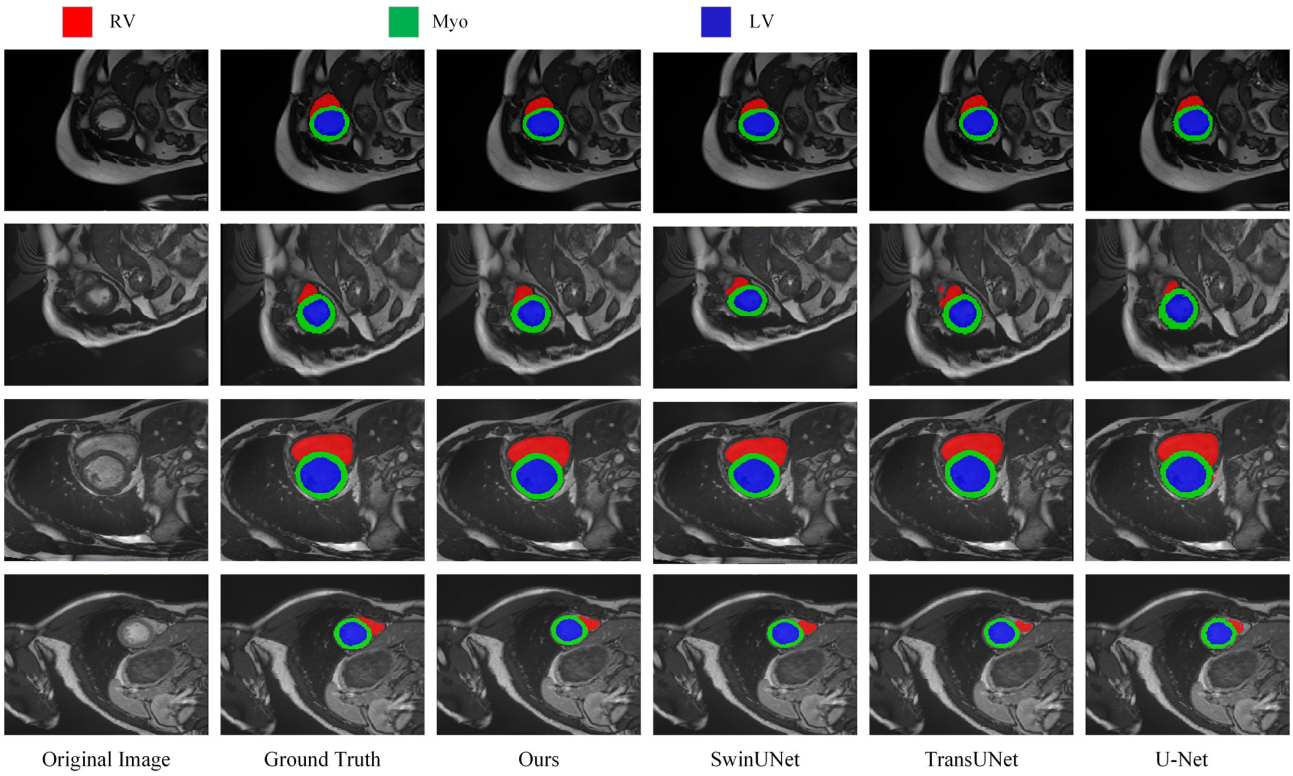
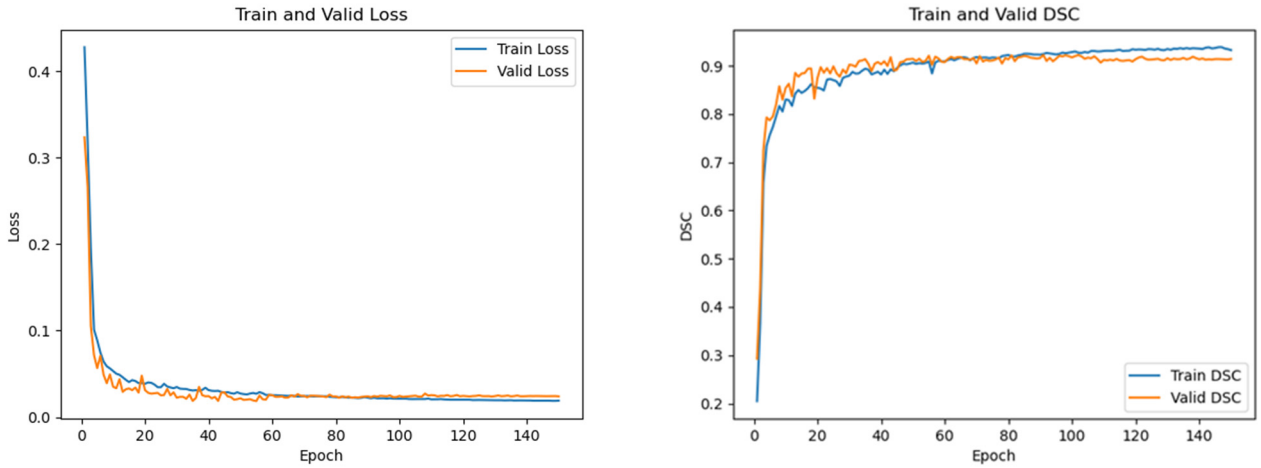


Figure 6. Comparison of model predictions on the ACDC dataset.



(**a**) (**b**)

Figure 7. Loss curves and DSC curves for training and validation on the ACDC dataset. (**a**) Loss curves for training and validation. (**b**) DSC curves for training and validation.

Table 4. Comparison of computational complexity and performance for different models on the Synapse dataset.

Methods	Parameters (million)	FLOPs (G)	Inference Time (minute)	GPU Memory Usage (GB)
U-Net [15]	0.29	5.18	0.32	0.81
Att-UNet [47]	31.38	32.40	1.38	1.54
TransUNet [23]	105.28	24.73	2.61	2.20
SwinUNet [45]	27.17	5.95	1.29	1.93
CCFNet [52]	137.36	76.18	3.85	5.08
Ours	82.10	19.58	2.23	2.06

4.5. Ablation Study

Based on the Synapse dataset, we conducted ablation experiments to investigate the impact of various factors such as different modules, input size, and model scale on the model accuracy. These results are shown as follows:

4.5.1. Effect of Different Modules

To better understand how each individual component contributes to the overall performance of our proposed model, we performed a set of ablation studies. These experiments aimed at a systematic evaluation of each module's contribution to the overall effectiveness of the model. The results of four evaluation metrics are recorded in Table 5. The findings from the ablation experiment clearly demonstrate that each component plays a vital role in enhancing the accuracy and robustness of the model. Among the components tested, the combination of the visual transformer model and the residual convolutional attention module yields the most significant improvement in performance. This combination substantially increases the accuracy by enabling the model to better capture both local and global features from the input data. Specifically, compared with the baseline model, the DSC has increased by 4.92, which is approximately a 6.40% growth, the HD has decreased by 21.68 mm, approximately 54.61%. These results emphasize the effectiveness of the proposed model in improving the quality and accuracy of segmentation.

Table 5. Analysis of the impacts of various modules.

Methods	DSC	HD
Baseline	76.85	39.70
Baseline + visual transformer	79.49	28.23
Baseline + visual transformer + RCAM	80.98	22.02
Baseline + visual transformer + MFC	80.46	26.89
Proposed method	81.77	18.02

4.5.2. Effect of Input Image Size

In this study, we explore the impact of varying input image resolutions on the model's performance using the Synapse dataset, with a focus on two distinct resolutions: 224 × 224 and 384 × 384, as summarized in Table 6. The results show that when the input image resolution is 384 × 384, the DSC and HD are 83.13% and 21.90 mm, respectively. This higher resolution provides more detailed information and finer spatial resolution, which can improve the accuracy of segmentation tasks. Specifically, the increase in resolution allows the model to better delineate boundaries and more precisely capture the regions of interest. However, while a higher resolution can lead to performance gains, it also introduces several significant challenges that must be carefully considered. Larger input sizes demand considerably more computational power and memory, resulting in higher computational costs. Additionally, the model may require more iterations to converge due to the increased complexity of the input data, which subsequently lengthens the training time. These factors can reduce the overall efficiency of the training process, making it less practical for scenarios with limited resources or strict time constraints. Given these considerations, we chose to conduct all the experiments in this paper using 224 × 224 input images.

Table 6. Evaluating the impacts of the input image size.

Input Size	DSC	HD	Aorta	Gallbladder	Kidney(R)	Kidney(L)	Pancreas	Liver	Spleen	Stomach
224 × 224	81.77	**18.02**	88.93	71.84	**82.95**	**85.19**	60.48	95.41	89.81	79.56
384 × 384	**83.13**	21.90	**91.31**	**72.52**	80.39	84.73	**68.40**	**96.19**	**90.27**	**81.25**

4.5.3. Effect of Model Scale

As shown in Table 7, the results of two different model scales were selected for the comparison in this study. Upon evaluation, the large model demonstrates a marginally higher accuracy than the small model, indicating a slight improvement in performance. However, this advantage in accuracy comes at a significant cost. The large model requires more computational time and greater system resources, leading to increased operational expenses and longer processing times. In contrast, although small models have a slightly lower accuracy, they achieve a better balance between performance and resource efficiency. Considering the trade-off between accuracy and computational cost, this study chose small models to perform the multi-organ segmentation tasks in medical imaging. This decision is based on practical considerations, as the time and computational resources required for the large model would exceed the small models' minimal loss in accuracy, particularly in real-world clinical or research settings, where cost-effectiveness is crucial.

Table 7. Evaluating the impacts of different model scales.

Model Scale	DSC	HD	Aorta	Gallbladder	Kidney(R)	Kidney(L)	Pancreas	Liver	Spleen	Stomach
Small	81.77	**18.02**	88.93	71.84	**82.95**	**85.19**	60.48	95.41	**89.81**	79.56
Large	**82.69**	23.58	**90.92**	**72.05**	82.73	84.26	**65.28**	**95.62**	89.15	**81.49**

5. Discussion

The results presented in Tables 2 and 3 demonstrate the effectiveness of using CNNs and transformers for multi-organ image segmentation in this study, and the experimental outcomes indicate that our proposed model can segment multi-organ images effectively. Compared to traditional CNN-based segmentation methods, the visual transformer mechanism has a significant advantage in capturing both global and local information in images. CNNs are typically limited to local features when processing images [16,43], while visual attention can adaptively focus on key areas within the image, improving the segmentation of edges and details [24]. Additionally, the RCAM module effectively avoids issues of gradient vanishing and information loss by introducing residual connections and enhancing the model's performance on complex images. On the other hand, the MFC module integrates feature maps of different scales, allowing the model to capture richer, multi-scale information, and further improve segmentation accuracy. These structural features compensate, to some extent, for the shortcomings of traditional convolutional neural networks in modeling long-range dependencies, allowing for better capture of global semantic information in images. In clinical diagnosis, fast and accurate image segmentation technology can effectively assist doctors in efficiently identifying lesion areas, which is particularly significant in the automated diagnosis of heart disease [3]. Our model can provide high-precision lesion localization through the automatic segmentation of CT or MRI scan images, offering doctors more reliable decision support and providing important evidence for the subsequent treatment planning, thereby reducing the error rate in manual diagnosis.

However, our proposed model also has some shortcomings. Since medical images involve sensitive personal health information and privacy, it makes the number of datasets relatively small, which may limit the ability to generalize the model. Despite the fact that we artificially enlarge the dataset by slicing, there is still the problem of the small number of patients. It may cause the model to perform poorly when confronted with new, unseen data. Furthermore, while our model is designed to process pre-processed 2D medical images, it is worth noting that 3D medical images are more commonly used in clinical practice, and they contain more information. Especially when dealing with complex anatomical structures and lesions, 3D images provide more comprehensive

spatial details. Therefore, this difference in data may cause the model to perform poorly in real life. Although our model outperforms other methods in terms of accuracy, its higher computational cost may pose some challenges in resource-constrained real-world application scenarios.

To address these challenges, we will further optimize the model's structure and computational efficiency by employing techniques such as pruning and quantization to make the model more lightweight, thereby better adapting to practical application scenarios, especially when processing 3D medical images, and exploring more efficient network architectures. Additionally, we will expand the diversity and scale of the dataset, adding more labeled data to enhance the model's generalization ability and practical application performance. At the same time, we will also focus on exploring how to seamlessly integrate this model with existing medical imaging diagnostic tools to promote the widespread application of this technology in clinical settings.

6. Conclusions

In this article, we propose an innovative model that utilizes a visual transformer and the CNN as the encoder to help it better capture local and global features. Specifically, we adopt visual attention and large kernel attention achieves adaptivity, not only in spatial dimensions, but also in channel dimensions, which is particularly important for visual tasks, because the objects in images are often closely related to their surroundings. Moreover, we discover that the decoder may lose crucial information and reduce resolution as the spatial dimensions expand during the up-sampling. To enhance the quality of segmentation results, we utilize a three-layered MFC and RCAM to strengthen the decoder. As a result of channel and spatial attention modules, our model can concentrate on semantic information and recover fine details more effectively. Additionally, MFC can expand the receptive field so that a larger range of contextual information can be captured. As a result, our model achieves DSC scores of 81.77% and 91.83% on the Synapse dataset and ACDC dataset, respectively. These results not only demonstrate the high accuracy of our approach, but also highlight its robustness and generalization capabilities across diverse datasets and segmentation tasks. It outperforms other classical network models in the challenging task of multi-organ segmentation in medical images, which strongly demonstrates the effectiveness and potential of our proposed method.

Author Contributions: Conceptualization, W.L. and P.J.; methodology, W.L.; software, P.J.; validation, P.J., F.W. and R.W.; formal analysis, W.L.; investigation, P.J.; writing—original draft preparation, P.J.; resources, F.W.; data curation, F.W.; visualization, R.W.; writing—review and editing, W.L.; project administration, W.L.; supervision, R.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Synapse dataset can be accessed publicly at https://www.synapse.org/#!Synapse:syn3193805/wiki/217789 (accessed on 15 January 2023), while the ACDC dataset is available at available at https://www.creatis.insa-lyon.fr/Challenge/acdc/ (accessed on 22 January 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Moorthy, J.; Gandhi, U.D. A Survey on Medical Image Segmentation Based on Deep Learning Techniques. *Big Data Cogn. Comput.* **2022**, *6*, 117. [CrossRef]
2. Aljuaid, H.; Alturki, N.; Alsubaie, N.; Cavallaro, L.; Liotta, A. Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning. *Comput. Methods Programs Biomed.* **2022**, *223*, 106951. [CrossRef] [PubMed]
3. Qureshi, I.; Yan, J.; Abbas, Q.; Shaheed, K.; Riaz, A.B.; Wahid, A.; Khan, M.W.J.; Szczuko, P. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Inf. Fusion* **2023**, *90*, 316–352. [CrossRef]
4. Yu-Qian, Z.; Wei-Hua, G.; Zhen-Cheng, C.; Jing-Tian, T.; Ling-Yun, L. Medical images edge detection based on mathematical morphology. In Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, China, 31 August–3 September 2005; pp. 6492–6495. [CrossRef]
5. Mubarak, D.M.N.; Sathik, M.M.; Beevi, S.Z.; Revathy, K. A hybrid region growing algorithm for medical image segmentation. *Int. J. Comput. Sci. Inf. Technol.* **2012**, *4*, 61–70. [CrossRef]
6. Jardim, S.; António, J.; Mora, C. Image thresholding approaches for medical image segmentation—Short literature review. *Procedia Comput. Sci.* **2023**, *219*, 1485–1492. [CrossRef]
7. Liu, X.; Qu, L.; Xie, Z.; Zhao, J.; Shi, Y.; Song, Z. Towards more precise automatic analysis: A systematic review of deep learning-based multi-organ segmentation. *Biomed. Eng. OnLine* **2024**, *23*, 52. [CrossRef] [PubMed]
8. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
9. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
10. Xie, X.; Huang, M.; Sun, W.; Li, Y.; Liu, Y. Intelligent tool wear monitoring method using a convolutional neural network and an informer. *Lubricants* **2023**, *11*, 389. [CrossRef]
11. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
12. Valanarasu, J.M.J.; Patel, V.M. Unext: Mlp-based rapid medical image segmentation network. In Proceedings of the MICCAI, Singapore, 18–22 September 2022; pp. 23–33. [CrossRef]
13. Chen, L.; Bentley, P.; Mori, K.; Misawa, K.; Fujiwara, M.; Rueckert, D. DRINet for medical image segmentation. *IEEE Trans. Med. Imaging* **2018**, *37*, 2453–2462. [CrossRef] [PubMed]
14. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571. [CrossRef]
15. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015. [CrossRef]
16. Wang, Y.; Gu, L.; Jiang, T.; Gao, F. MDE-UNet: A Multitask Deformable UNet Combined Enhancement Network for Farmland Boundary Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 3001305. [CrossRef]
17. Ahmad, P.; Jin, H.; Alroobaea, R.; Qamar, S.; Zheng, R.; Alnajjar, F.; Aboudi, F. MH UNet: A multi-scale hierarchical based architecture for medical image segmentation. *IEEE Access* **2021**, *9*, 148384–148408. [CrossRef]
18. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7151–7160. [CrossRef]
19. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6999–7019. [CrossRef]
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
21. Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; Tran, D. Image transformer. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 4055–4064.
22. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45. [CrossRef]
23. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306. [CrossRef]
24. Guo, M.-H.; Lu, C.-Z.; Liu, Z.-N.; Cheng, M.-M.; Hu, S.-M. Visual attention network. *Comput. Vis. Media* **2023**, *9*, 733–752. [CrossRef]
25. Kayalibay, B.; Jensen, G.; van der Smagt, P. CNN-based segmentation of medical imaging data. *arXiv* **2017**, arXiv:1701.03056. [CrossRef]

26. Mortazi, A.; Bagci, U. Automatically designing CNN architectures for medical image segmentation. In Proceedings of the Machine Learning in Medical Imaging: 9th International Workshop, MLMI 2018, Granada, Spain, 16 September 2018; pp. 98–106. [CrossRef]
27. Xiao, X.; Lian, S.; Luo, Z.; Li, S. Weighted res-unet for high-quality retina vessel segmentation. In Proceedings of the 2018 9th International Conference on Information Technology in Medicine and Education (ITME), Hangzhou, China, 19–21 October 2018; pp. 327–331. [CrossRef]
28. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the 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, Granada, Spain, 20 September 2018; pp. 3–11. [CrossRef]
29. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1055–1059. [CrossRef]
30. Lou, A.; Guan, S.; Loew, M. DC-UNet: Rethinking the U-Net architecture with dual channel efficient CNN for medical image segmentation. In Proceedings of the Medical Imaging 2021: Image Processing, Online, 15–19 February 2021; pp. 758–768. [CrossRef]
31. Kawamoto, M.; Kamiya, N.; Zhou, X.; Kato, H.; Hara, T.; Fujita, H. Simultaneous Learning of Erector Spinae Muscles for Automatic Segmentation of Site-Specific Skeletal Muscles in Body CT Images. *IEEE Access* **2024**, *12*, 15468–15476. [CrossRef]
32. Ashino, K.; Kamiya, N.; Zhou, X.; Kato, H.; Hara, T.; Fujita, H. Joint segmentation of sternocleidomastoid and skeletal muscles in computed tomography images using a multiclass learning approach. *Radiol. Phys. Technol.* **2024**, *17*, 854–861. [CrossRef]
33. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805. [CrossRef]
34. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
35. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]
36. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 568–578. [CrossRef]
37. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 22–31. [CrossRef]
38. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022. [CrossRef]
39. Li, J.; Ye, J.; Zhang, R.; Wu, Y.; Berhane, G.S.; Deng, H.; Shi, H. CPFTransformer: Transformer fusion context pyramid medical image segmentation network. *Front. Neurosci.* **2023**, *17*, 1288366. [CrossRef] [PubMed]
40. Yao, T.; Li, Y.; Pan, Y.; Wang, Y.; Zhang, X.P.; Mei, T. Dual Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10870–10882. [CrossRef]
41. Wang, H.; Xie, S.; Lin, L.; Iwamoto, Y.; Han, X.H.; Chen, Y.W.; Tong, R. Mixed transformer U-Net for medical image segmentation. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 2390–2394. [CrossRef]
42. Jiang, T.; Xu, T.; Li, X. VA-TransUNet: A u-shaped medical image segmentation network with visual attention. In Proceedings of the 2022 11th International Conference on Computing and Pattern Recognition, Beijing, China, 17–19 November 2022; pp. 128–135. [CrossRef]
43. Lin, A.; Chen, B.; Xu, J.; Zhang, Z.; Lu, G.; Zhang, D. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 4005615. [CrossRef]
44. Peng, B.; Alcaide, E.; Anthony, Q.; Albalak, A.; Arcadinho, S.; Biderman, S.; Cao, H.; Cheng, X.; Chung, M.; Grella, M. Rwkv: Reinventing rnns for the transformer era. *arXiv* **2023**, arXiv:2305.13048.
45. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218. [CrossRef]
46. Fu, S.; Lu, Y.; Wang, Y.; Zhou, Y.; Shen, W.; Fishman, E.; Yuille, A. *Domain Adaptive Relational Reasoning for 3D Multi-Organ Segmentation*; Springer: Cham, Switzerland, 2020; pp. 656–666. [CrossRef]

47. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999. [CrossRef]
48. Azad, R.; Al-Antary, M.T.; Heidari, M.; Merhof, D. TransNorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model. *IEEE Access* **2022**, *10*, 108205–108215. [CrossRef]
49. Jiang, X.; Ding, Y.; Liu, M.; Wang, Y.; Li, Y.; Wu, Z. BiFTransNet: A unified and simultaneous segmentation network for gastrointestinal images of CT & MRI. *Comput. Biol. Med.* **2023**, *165*, 107326. [CrossRef]
50. Zhong, X.; Xu, L.; Li, C.; An, L.; Wang, L. RFE-UNet: Remote feature exploration with local learning for medical image segmentation. *Sensors* **2023**, *23*, 6228. [CrossRef] [PubMed]
51. Peng, B.; Fan, C. IEA-Net: Internal and external dual-attention medical segmentation network with high-performance convolutional blocks. *J. Imaging Inform. Med.* **2024**. [CrossRef] [PubMed]
52. Chen, J.; Yuan, B. CCFNet: Collaborative cross-fusion network for medical image segmentation. *Algorithms* **2024**, *17*, 168. [CrossRef]