CA-UNet: UNet-like Heterogeneous Architecture of CNN and Attention for Medical Image Segmentation

Jian Wang[1], Fei Qi[1*], Jianfei Wu[2], and Siqi Yan[3]

Xidian University, Xi'an, China

Abstract. Medical image segmentation has wide applications and research value in the field of medical research and practice, greatly improving the efficiency and accuracy of disease diagnosis.In various types of medical segmentation tasks, the encoder-decoder U-shaped heterogeneous architecture network, which integrates convolutional neural networks (CNN) and attention mechanisms, has achieved milestone achievements. It empowers the network with strong local feature extraction ability and long-term dependency modeling capability, and has achieved excellent performance in medical image segmentation.However, most of the current heterogeneous networks adopt independent convolutional layers or attention layers in various feature extraction modules, unable to effectively establish dependencies between local and global information, thus failing to fully exploit their advantages.Moreover, due to the small sample characteristic of medical images which often limits the learning capacity of the network model, this article applies full-scale skip connection based on the convolutional attention mechanism to the feature fusion of the encoder and decoder, thereby effectively enhancing the segmentation performance of the network. In experiments involving Synapse multi-organ segmentation(Synapse) dataset, the average DSC of the proposed network reached 84.43%, and HD95 reached 15.69. At the same time, it showed excellent generalization ability in Automated cardiac diagnosis challenge(ACDC) dataset. Code is available at https://github.com/waungjian/CA-UNet.git

1 Introduction

Due to the great success of deep learning, CNNs, especially Fully Convolutional Networks (FCN) [1], have been widely used in the field of medical image segmentation. Based on the FCN, U-Net [2], as the pioneer of medical image segmentation, has attracted the attention of many researchers with its symmetrical encoder-decoder framework. In this topology structure, the encoder is used to extract feature information of different scales from images, while the decoder

*Corresponding author

fuses the feature information extracted by the encoder via cross-layer connections and restores it to the original input image size, making pixel-level semantic predictions. This network structure has achieved effective fusion of fine-grained information and semantic information in small-sample data, and achieved great success in various medical image segmentation tasks. In order to alleviate the loss of spatial information caused by downsampling, U-Net++ [3] and U-Net3+ [4] networks based on multi-scale information fusion mechanism and deep supervision have been proposed successively. These methods learn hierarchical representations from feature maps aggregated at multiple scales, thus obtaining more accurate segmentation results.In addition, the DeepLab [5,6,7,8] series of networks combined with Atrous convolution designed a multi-scale objective robust algorithm based on Atrous Spatial Pyramid Pooling (ASPP).

Although CNN have powerful feature extraction capabilities, due to the inherent locality of their convolution operations, CNN-based medical image semantic segmentation networks still show certain limitations in modelling visual receptive fields and long-range dependencies. The emergence of the visual attention mechanism provides a novel and effective path to enhance the long-range modeling capacity of CNNs. Attention U-Net [9] first introduced the gated attention mechanism into the U-Net network, allowing the network to learn the complex relations between features across different scales and hierarchical levels. In the same year, Woo et al. [10] effectively combined channel attention (CA) and spatial attention (SA) to propose a novel and lightweight convolution block attention module (CBAM). This module can be flexibly embedded into deep CNNs for various application scenarios and tasks without introducing extra computational or parameter load. A large number of experiments have demonstrated the effectiveness and generalization capabilities of CBAM in tasks such as image classification and image segmentation.

The self-attention mechanism based on Transformer [11] has achieved great success in the field of natural language processing(NLP) due to its exceptional ability to model long-range dependencies, paving a new research path in the field of computer vision. In [12], Visual Transformer (ViT) was proposed and applied for image classification tasks. It performs equivalently to CNN-based classification networks when pre-trained on large-scale datasets, showcasing the powerful ability of Transformers in global context modeling. To enhance ViT's local perception capabilities and handle super-resolution images, researchers designed the Swin Transformer [13], which has been successfully applied in fields like image object detection and semantic segmentation. A series of studies have adopted Transformers in the field of medical image segmentation, seeking to replace CNNs [14,15]. However, due to the lack of key inductive biases inherent to CNNs, such as translation invariance and locality, its ability to extract spatial detail features is limited. Based on the complementary strengths of CNNs and ViTs, a series of hybrid U-shaped topologies based on CNN-ViT have been proposed for medical image segmentation tasks [16,17,18,19]. A potential shortfall of these networks, however, is the lack of integration of the complementary strengths of both at each level of feature extraction modules, rendering them

unable to effectively establish dependencies between local and global information. The exceptional performance of CMTs [20] in various visual tasks presents us with a novel method of hierarchical fusion between CNN-ViT.

Motivated by the success of the CMT architecture and the Swin Transformer [20,13], we propose CA-UNet for 2D medical image segmentation, including encoders, bottleneck, decoders, and full-scale skip connections. We combined the window attention mechanism unit in Swin Transformer with depthwise convolution (DwConv) based on a dynamic weight allocation mechanism to construct a parallel heterogeneous module (CA block). The encoders, bottleneck, and decoders are all built based on the CA block. First, spatial and structural information extraction is carried out on the input image stem architecture [21] and then input to the encoder for deep representation learning. The encoder obtains multi-scale feature maps based on a downsampling structure with residual architecture. Through the decoder and a full-scale skip connection module based on convolutional attention, the deep feature map extracted is fused with multi-scale features of different resolutions on the encoder path to achieve precise positioning. Numerous experiments on multi-organ and heart segmentation datasets have demonstrated the superiority of this method. Specifically, our contributions can be summarized as: (1) providing a novel U-shaped symmetric encoder-decoder architecture based on CNN and Transformer for medical image segmentation tasks based on parallel heterogeneous CA blocks; (2) a downsampling module based on residual architecture(Res block), which enhances key detail features of the downsampled feature map, further highlighting the hierarchical relationship of features; (3) a full-scale skip connection module based on CBAM, which allows the network to autonomously learn the importance of the encoder's feature information for decoders at different levels.

2 Related work

CNN-based methods : Due to the powerful characterization capability of deep CNNs, researchers applied them to medical image segmentation tasks and proposed U-Net [2]. Due to the superior performance and generalization ability of U-shaped structure, various networks based on U-shaped structure have also been proposed successively, such as U-Net++ [3] and U-Net3+ [4], which incorporate multiscale mechanism, and Residual U-Net [22,23,24,25], which incorporates residual connectivity. Some researchers have also applied the structure to the field of 3D medical image segmentation, and 3D U-Net [26] and V-Net [27] have been proposed successively.

Attention mechanism : The visual attention mechanism holds crucial significance in deep learning technology research. Attention U-Net [9] was the first to introduce the gated visual attention mechanism, enhancing the modeling capability of complex relationships between features at different scales and levels. U-Net v2 [28] combines the CBAM module with the cross-layer connections of U-Net architecture, endowing the decoder's multi-level feature maps with rich semantic

information and complex detail features through attention focus.With the great success of Transformer on natural language processing, many researchers applied it to computer vision tasks and proposed ViT [12]. It was applied to image recognition tasks and achieved comparable performance to CNN with large-scale data pre-training. To improve the applicability of Transformer for various vision tasks, Swin Transformer [13] with a hierarchical structure was created to bring higher efficiency through a multi-scale modeling and shifted windowing scheme, demonstrating the potential of Transformer-based models as a visual backbone. Based on the excellent performance of Transformer on vision tasks, combined with the U-shaped architecture, researchers proposed Swin-UNet [14] for medical semantic segmentation tasks and achieved excellent segmentation accuracy, providing a new perspective to solve the problem of medical semantic segmentation tasks. To further enhance the adaptive feature alignment ability of the Transformer, researchers optimized its structure, applied it to U-Net's network encoder, decoder, and skip connections for feature extraction, and proposed MISSFormer [15] for use in medical image segmentation tasks.

Combining CNNs with Transformer : Due to the excellent performance of Transformer in the field of vision, researchers have endeavored to combine CNN with Transformer and exploit the complementary nature of both to build U-shaped encoder-decoder medical image segmentation models and improve the segmentation capability of the models, such as TransUNet [16] and TransFuse [17]. Different from the fusion approach of the above hybrid architecture models, we try to explore the potential of a novel structural fusion approach of hybrid architecture in medical image segmentation. In addition, leveraging the outstanding global modelling ability of Transformer, CASTformer [29] has built a class-aware module based on the Transformer, which updates the sampling positions through iterative optimization, thereby adapting to the key areas of objects, such as anatomical features and structural information.

3 Method

3.1 Architecture overview

The overall architecture of the proposed CA-UNet is presented in Fig. 1.Its basic constituent units mainly include the following three parts: (1) The CA block that fuses convolution and attention mechanisms, used for constructing network feature extraction modules; (2) the downsampling module based on residual structure; (3) The full-scale skip connection module based on the convolution block attention mechanism. Given an input image $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$ with a spatial resolution of $H \times W$ and a channel number of C, it is first passed through a stem architecture [21] for image resolution reduction and fine-grained feature extraction. After downsampling, the image resolution drops to $\frac{H}{4} \times \frac{W}{4}$, and then it is input to a series of CA blocks for feature extraction and fusion. Downsampling in four stages generates feature maps of different scales, and these multi-scale

feature maps are critical for feature extraction in the dense prediction task during the decoding stage. In order to reduce the loss of spatial information in the feature map during downsampling, after bilinear upsampling in the decoder part, spatial detail information and abstract semantic information are fused through the full-scale skip connection module. After four stages of upsampling and feature extraction and fusion, a pixel-level segmentation result map $\hat{\boldsymbol{Y}} \in \mathbb{R}^{H \times W \times K}$ with the same spatial resolution as the original image is output, where K represents the number of segmentation categories.

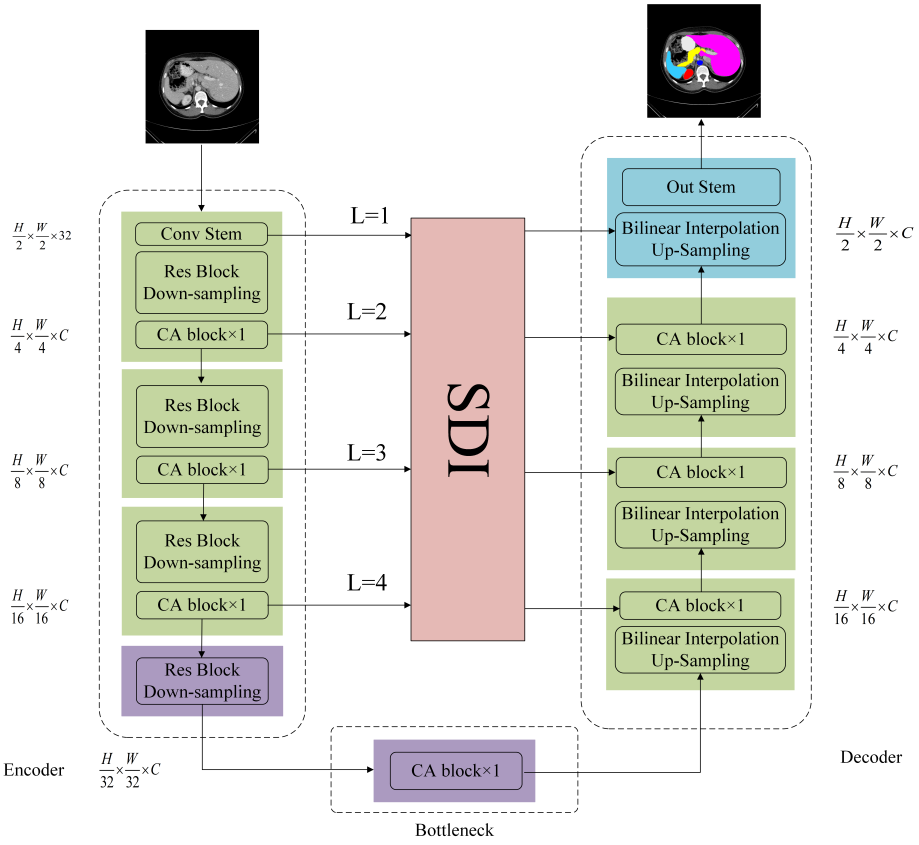

Fig. 1. The architecture of CA-UNet, which is composed of encoder, bottleneck, decoder and skip connections. Encoder, bottleneck and decoder are all constructed based on CA block.

3.2 Stem architecture

At present, most ViTs linearly map the input image into a one-dimensional sequence after dividing it into non-overlapping image patches, but this method destroys the two-dimensional structure and spatial information within each patch. To overcome the limitations of linear mapping modeling, CA-UNet adopts a stem architecture, expecting to map the input image into different dimensions of nonlinear space to preserve as much spatial and structural feature information as possible.It mainly consists of convolution layers, Gaussian Error Linear Unit activation functions, and BN layers. Firstly, a 3×3 convolution layer with a stride of 2 is used to downsample the input image, generating a feature map with half the resolution and a channel dimension of 32. Subsequently, local feature extraction is performed using two consecutive 3×3 convolution layers with a stride of 1.

3.3 Parallel heterogeneous module

Convolutional neural networks use self-learning convolution kernels to aggregate spatial information representation. The convolution operation, due to its translation invariance, can more effectively explore the potential representational capacity in feature channel dimensions, demonstrating superior local perception ability when analyzinge image spatial structure information. The key advantage of Transformer [11] lies in its self-attention mechanism, which gives it exceptional capabilities in modeling long-range dependencies. In [11,13], the self-attention mechanism associates features of different positions through a dynamic weight allocation mechanism. Without disrupting the respective structures and modeling of both, CA-UNet proposes a parallel heterogeneous module based on convolution and self-attention mechanisms, named CA block. It combines the advantages of both, adaptively allocating dynamic weights to different spatial positions and channel dimensions, thereby effectively capturing complex patterns in medical image feature maps.The CA Block primarily consists of parallel units composed of multi-head self-attention units based on the Swin Transformer [13] and local perception units based on deep convolution, as well as the Inverted Residual Feed-forward Network (IRFFN) [20]. The module is shown in Fig. 2, and each of its components will be briefly introduced below.

Self-attention unit CA-UNet utilizes the window attention mechanism of the Swin Transformer [13] with a multi-layer hierarchical architecture, effectively analyzing and extracting local features and global semantic information from medical images, while also reducing computational complexity. The W-MSA module divides the input image into non-overlapping windows for local self-attention calculation, and the window size is defaultly set as 7×7. This strategy helps reduce computational complexity and allows the model to handle larger scale images. The SW-MSA module re-divides the non-overlapping windows from the W-MSA stage using moving windows, effectively solving the receptive field limitation problem brought by the window attention mechanism. This strategy enables windows to have cross-window information interaction capability and

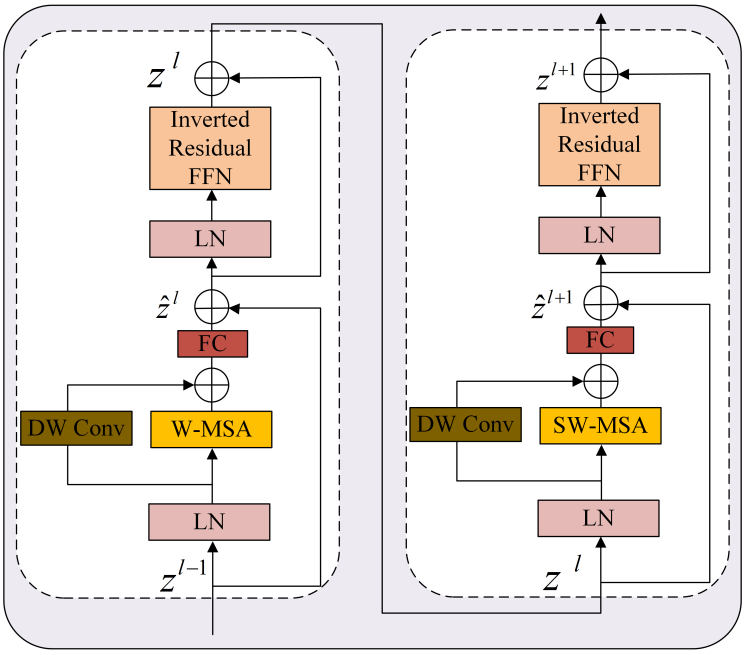

Fig. 2. CA block.

expands their coverage range, thereby facilitating the model to better capture global information. Similar to the previous works [30,31,32,33], self-attention is computed as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V, \tag{1}$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ denote the query, key, and value matrices. M^2 and d represents the number of patches in a window and dimension of the query or key,respectively. And the value in B are taken from the bias matrix $\hat{\boldsymbol{B}} \in \mathbb{R}^{(2M-1)\times(2M-1)}$

Local perception unit This unit uses DwConv to enhance the local information perception of defect features in composite components, effectively exploring its potential representation capability in the feature channel dimension, while significantly reducing the amount of parameters and computational complexity. Considering the scale consistency of the window attention mechanism, the convolution kernel size is set to 7×7.

Inverted Residual Feedforward Network The design of the IRFFN is similar to the inverse residual block [34], mainly composed of expansion layers, DwConv, and projection layers. The network enhances the ability of gradient propagation across layers while optimizing network performance by changing the position of skip connections. Notably, the IRFFN omits the activation layer

and achieves efficient local feature extraction through DwConv, and the related additional computational cost is negligible. This structure is computed as follows:

$$\begin{aligned} IRFFN\left(X\right) &= Conv\left(F\left(Conv\left(X\right)\right)\right), \\ F\left(X\right) &= DwConv\left(X\right) + X. \end{aligned} \tag{2}$$

Overall, the CA Block reconsiders the fusion strategy of CNN and Transformer based on the dynamic weight allocation mechanism, and achieves the complementarity and information synchronization of the two through fully connected layers (FC). Finally, it uses IRFFN to further capture the local structural features and global semantic information of the intermediate features, thereby enhancing the network's expressive ability. The CA Block is computed as follows:

$$\begin{aligned} \hat{z}^{l} &= \text{W-MSA}\left(\text{LN}\left(z^{l-1}\right)\right) + \text{DwConv}\left(\text{LN}\left(z^{l-1}\right)\right), \\ z^{l} &= \text{IRFFN}\left(\text{LN}\left(\text{FC}\left(\hat{z}^{l}\right)\right) + \hat{z}^{l-1}\right) + \hat{z}^{l}, \\ \hat{z}^{l+1} &= \text{SW-MSA}\left(\text{LN}\left(z^{l}\right)\right) + \text{DwConv}\left(\text{LN}\left(z^{l}\right)\right), \\ z^{l+1} &= \text{IRFFN}\left(\text{LN}\left(\text{FC}\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l}\right) + \hat{z}^{l+1}, \end{aligned} \tag{3}$$

where $\hat{z}^{l}$ represents the output after the fusion of the DwConv and (S)W-MSA features in the lth block, and z^{l} represents the output of the IRFFN in the lth CA Block.

3.4 Encoder

The encoder in each layer consists of a downsampling module based on a Res block and a CA block . The downsampling module, which replaces the pooling layer, is grounded on a Res block. It aims to enhance the detail features after downsampling and thus improve the model's understanding of the spatial structure and positional information relationships. The encoder inputs the information into the CA block for the aggregation of local spatial information and global semantic information. The encoder performs double downsampling through a convolution layer with a kernel size of 3×3 and a stride of 2 in the residual structure unit. ncoder downsamples the input image in four stages, thereby obtaining four types of feature maps at different scales. These feature maps, after full-scale skip connections based on the convolution attention mechanism, are inputted into decoders at different levels to achieve multiscale information fusion.

3.5 Bottleneck

The CA block is composed of CNN and Transformer. Overly deep network structures can lead to model non-convergence [35], therefore, only one CA Block module is used to build the bottleneck. After passing the bottleneck layer, the resolution and feature dimension of the feature map remain unchanged.

3.6 Decoder

Corresponding to the structure of the encoder, CA-UNet includes four layers of decoders. Among them, the first three decoders use the CA block in the same way for feature extraction in the decoder, and carry out feature aggregation through twofold up-sampling and a Res block. Then, it performs multi-scale information fusion with the cross-layer connected feature map of the corresponding level. After up-sampling for the fourth layer decoder, a convolution stem architecture is used for spatial and structural feature extraction, and this process does not downsample the feature map.

3.7 Full-scale skip connection

Fusing features through traditional cross-layer cascade methods will largely depend on the learning capacity of the network. However, the learning capacity of network models usually depends on the size of the dataset, which undoubtedly poses a challenging problem for small-sample medical image data. To solve this problem, the CA-UNet introduces the Semantics and Detail Infusion (SDI) module [28] into the skip connection, endowing the decoder with rich semantic features and complex details at all levels. As shown in Fig. 3, this figure displays the structure of the SDI module for cross-layer feature fusion at the third layer of the decoder.

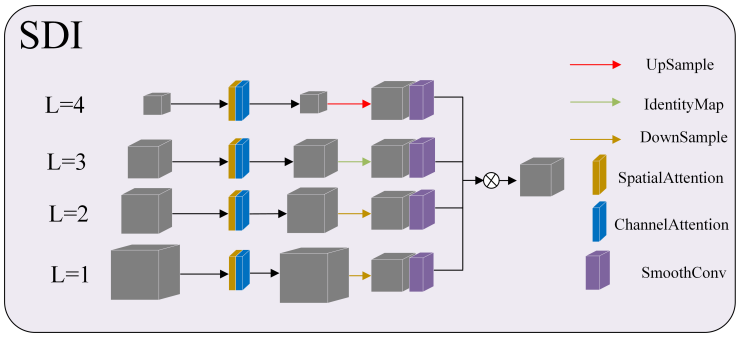

Fig. 3. Structure of the third layer skip connection module in CA-UNet

The SDI module first inputs the features f_i^0 of each level of the encoder into the CBAM module for spatial and channel attention, capturing and enhancing important feature information in the spatial and channel dimensions in depth, which is computed as follows:

$$f_i^1 = \phi_i^c(\varphi_i^s(f_i^0)), \tag{4}$$

where f_i^1 represents the feature map processed by the ith level of the encoder, and ϕ_i^c and φ_i^s respectively represent the parameters of the spatial and channel attention at the ith level.

Afterwards, the f_i^1 layers with convolutional attention are aggregated and dimensionality reduced for channel information through a 1×1 convolution, resulting in the feature maps $f_i^2 \in \mathbb{R}^{H_i \times W_i \times c}$ of each level.Next, the optimized features f_i^2 of each level are sent to the decoder. Specifically, at each decoder level i, f_i^2 is used as the target reference, and the size of each jth level feature map is adjusted to match the same resolution of f_i^2, which is computed as follows:

$$f_{ij}^3 = \begin{cases} \mathrm{D}(f_j^2, (H_i, W_i)) & \text{if} \quad j < i, \\ \mathrm{I}(f_j^2) & \text{if} \quad j = i, \\ \mathrm{U}(f_j^2, (H_i, W_i)) & \text{if} \quad j > i, \end{cases} \tag{5}$$

where $D(\cdot), I(\cdot)$ and $U(\cdot)$ denote adaptive average pooling, identity mapping, and bilinear interpolation of f_j^2 to the resolution of $H_i \times W_i$.

Through a 3×3 convolution, each resampled feature map is smoothed into f_{ij}^3, which aims to reduce the potential jagged blurry features that may occur during the sampling process of the feature map. Lastly, a per-pixel Hadamard product is applied to all resampled and smoothed multi-level feature maps to enhance the feature maps of the ith level in the decoder.

4 Experiments

4.1 Datasets

Synapse dataset: The dataset includes 3779 axial abdominal clinical CT images from 30 cases. Each abdominal CT scan ranges from 85 to 198 CT slices of 512×512 pixels with voxel spatial resolution of $([0.54 \sim 0.54] \times [0.98 \sim 0.98] \times [2.5 \sim 5.0])mm^3$. In this paper, following [16], 18 CT case samples are divided into the training set, and the remaining 12 CT case samples are allocated to the test set. The average DSC and HD95 are used as the evaluation metrics to assess the segmentation performance of CA-UNet on 8 abdominal organs: Aorta, Gallbladder, Left Kidney, Right Kidney, Liver, Pancreas, Spleen, and Stomach.

ACDC dataset: The dataset collects cardiac magnetic resonance imaging data from different patients using MRI scanners. The slice thickness of each MRI scan case is $5 \sim 8$ millimeters, covering the entire heart from the base to the top of the left ventricle. Additionally, the planar spatial resolution of these short-axis slices is between $0.83 \times 1.75mm^2$/pixels. Each MRI scan case slice is professionally annotated in detail, including Left Ventricle (LV), Right Ventricle (LV), and myocardium (Myo). Following [16], this dataset is randomly divided into 70 training samples (including 1930 axial slices), 10 validation samples, and 20 test samples. This paper uses average DSC to evaluate the performance of the algorithm proposed in this paper on the segmentation of 3 cardiac tissues.

4.2 Implementation details

The CA-UNet is implemented based on PyTorch and is trained and tested on the Nvidia GeForce RTX 4090 GPU. For each batch, the number of samples

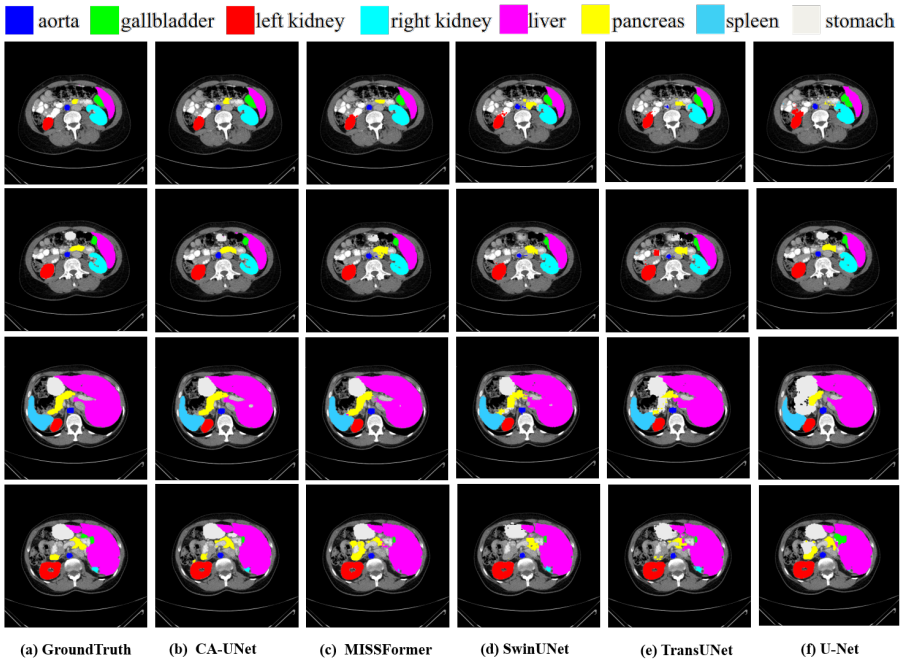

Fig. 4. The segmentation visual results of different methods on the Synapse multi-organ CT dataset.

is set to 16 with a total of 400 iterations. Moreover, in this chapter, the number of channels C of feature maps at all levels of the network model is set to 96, and the network model weights are randomly initialized and retrained. The AdamW optimizer is used, with the learning rate set to 0.001, momentum set to 0.9, and weight decay set to 1e-4. The data augmentation strategy follows MISSFormer [15].

4.3 Experiment results on Synapse dataset

Table. 1 shows the quantitative comparison results of the segmentation accuracy (average DSC and HD95) of CA-UNet with the current mainstream methods.Experimental results demonstrate that our method achieves the best performance with segmentation accuracy of 84.43%(DSC↑) and 15.69(↓),respectively, showing an improvement of 1.88% and 7.04% compared to CASTformer [15]. Fig. 4 shows a visual comparison of different methods on the Synapse multi-organ CT dataset. Compared to the current mainstream CNN-ViT or pure Transformer methods, our method demonstrates the excellence of the parallel heterogeneous feature extraction module. Even when substantially reducing the channel dimension of the feature map, it further optimizes the recognition accuracy of various organs and has higher clarity in edge segmentation.

Table 1. Segmentation accuracy of different methods on the Synapse dataset.

Methods	DSC↑	HD95↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
V-Net [27]	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR [35]	69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50 ViT [16]	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
R50 U-Net [16]	74.68	36.87	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
U-Net [2]	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
TransUNet [16]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Att-UNet [9]	77.77	36.02	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	75.75
Swin UNet [14]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
MISSFormer [15]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
CASTformer [29]	82.55	22.73	89.05	67.48	86.05	82.17	**95.61**	67.49	91.00	81.55
CA-UNet	**84.43**	**15.69**	88.59	**73.70**	**87.92**	**83.34**	94.85	**71.52**	**91.98**	**82.51**

4.4 Experiment results on ACDC dataset

To validate the robustness and generalizability of CT-UNet on other medical datasets, we conducted training on the ACDC dataset in MRI mode, with experimental results shown in Table. 2. The CA-UNet we proposed achieved 90.68% accuracy in the segmentation of three types of cardiac tissues. The experimental results show that our method has reached an advanced level in the segmentation of Myo, while the segmentation performance for the LV and RV also reached a level comparable to the current mainstream method.

Table 2. Segmentation accuracy of different methods on the ACDC dataset.

Methods	DSC	RV	Myo	LV
R50-U-Net [16]	87.55	87.10	80.63	94.92
R50-AttnUNet [16]	86.75	87.58	79.20	93.47
ViT-CUP [16]	81.45	81.46	70.71	92.18
R50-ViT-CUP [16]	87.57	86.07	81.88	94.75
TransUNet [16]	89.71	**88.86**	84.53	95.73
Swin UNet [14]	90.00	88.55	85.62	**95.83**
CA-UNet	**90.68**	88.57	**89.83**	93.63

4.5 Ablation study

Our proposed CA-UNet effectively aggregates a parallel heterogeneous module that fuses convolution and attention mechanisms, residual-structure-based downsampling, and full-scale skip connection based on CBAM. As a result, it demonstrates superior segmentation performance and generalization ability in medical imaging segmentation tasks. To verify its rationality and effectiveness, we conducted the following ablation experiment on the Synapse dataset.

Effect of the parallel heterogeneous module: CA-UNet combines DwConv and self-attention in SwinTransformer to construct a parallel heterogeneous mod-

ule for network hierarchical feature extraction and fusion, dynamically allocating weights adaptively in spatial position and channel dimensions. To verify its effectiveness, we examined whether the introduction of negligible computational cost deep convolution (for channel dimension feature capture) has an impact on segmentation performance. The ablation experiment results in Table. 3 show that the average DSC of various organ segmentations was increased from 77.82% to 84.43%, a 3.97% increase, through the designed hierarchical parallel heterogeneous module. This effectively enhances the recognition capability of complex medical image organs, making the edge segmentation more precise.

Table 3. Ablation study on the effect of the parallel heterogeneous module:

Methods	DSC	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
− DwConv	80.38	87.54	68.35	78.77	77.48	93.99	66.53	89.43	80.93
+ DwConv	**84.43**	**88.59**	**73.70**	**87.92**	**83.34**	**94.85**	**71.52**	**91.98**	**82.51**

Effect of downsampling module based on Res block: We incorporated the residual structure into the downsampling module, effectively enhancing the key detail features of the encoder feature maps at all scales and effectively avoiding the issue of feature location information loss caused by traditional downsampling methods (average pooling and max pooling). To validate its effectiveness, we explored the impact of using max pooling and Res block for downsampling on the segmentation performance of CA-UNet. The ablation experiment results in Table. 4 show that better segmentation precision was achieved through the downsampling module based on Res block, raising the average DSC from 79.55% to 84.43%.

Table 4. Ablation study on the effect of downsampling module based on Res block

Downsampling	DSC	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
Max pooling	82.56	88.54	69.94	87.18	81.57	94.72	67.34	90.19	80.98
Residual structure	**84.43**	**88.59**	**73.70**	**87.92**	**83.34**	**94.85**	**71.52**	**91.98**	**82.51**

Effect of the full-scale skip connection base on CBAM: We introduced a full-scale skip connection based on CBAM, which autonomously learns and aggregates the feature information of all levels of the encoder through spatial and channel attention, reducing the semantic gap caused by the blurring of feature mapping between the encoder and decoder. To explore its effectiveness, we discussed the impact of traditional skip connection based on channel dimension concatenation(CDC) and full-scale skip connection based on CBAM on network segmentation performance. The experimental results in Table. 5 show that by

introducing a full-scale skip connection based on CBAM, the segmentation accuracy of CA-UNet is effectively improved, raising the average DSC from 78.29% to 84.43%.

Table 5. Ablation study on the effect of downsampling module based on the residual structure

skip connection	DSC	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
CDC	80.09	87.11	70.45	79.35	74.11	94.34	63.53	90.16	81.69
CBAM	**84.43**	**88.59**	**73.70**	**87.92**	**83.34**	**94.85**	**71.52**	**91.98**	**82.51**

5 Conclusion

In this paper, we effectively integrate CNN and Swin Transformer's window self-attention based on dynamic weight allocation mechanism, and designs CA block to enhance the sensitivity of the network to local features, as well as to extract and effectively fuse global feature information. Secondly, a downsampling module based on Res block and full-scale skip connection based on CBAM are designed to further enhance the model's ability to perceive complex details in medical images and understand abstract semantic information, thereby improving the clarity of edge segmentation. Finally, extensive experiments were conducted on two public medical image segmentation datasets. The experimental results validate that compared with previous state-of-the-art methods, CA-UNet has leading advantages in average DSC and HD95.

References

1. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
2. 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*. Springer, 2015, pp. 234–241.
3. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.
4. 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)*. IEEE, 2020, pp. 1055–1059.
5. C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *International conference on learning representations*, 2015.

6. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
7. L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.
8. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.
9. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.
10. 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)*, 2018, pp. 3–19.
11. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
12. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
13. 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*, 2021, pp. 10 012–10 022.
14. 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*. Springer, 2022, pp. 205–218.
15. X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, "Missformer: An effective transformer for 2d medical image segmentation," *IEEE Transactions on Medical Imaging*, 2022.
16. 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.
17. 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*. Springer, 2021, pp. 14–24.
18. M. Naderi, M. Givkashi, F. Piri, N. Karimi, and S. Samavi, "Focal-unet: Unet-like focal modulation for medical image segmentation," *arXiv preprint arXiv:2212.09263*, 2022.
19. L. Lan, P. Cai, L. Jiang, X. Liu, Y. Li, and Y. Zhang, "Brau-net++: U-shaped hybrid cnn-transformer network for medical image segmentation," *arXiv preprint arXiv:2401.00722*, 2024.
20. J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "Cmt: Convolutional neural networks meet vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 175–12 185.
21. T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 558–567.

22. M. Baldeon-Calisto and S. K. Lai-Yuen, "Adaresu-net: Multiobjective adaptive convolutional neural network for medical image segmentation," *Neurocomputing*, vol. 392, pp. 325–340, 2020.
23. M. Z. Alom, C. Yakopcic, T. M. Taha, and V. K. Asari, "Nuclei segmentation with recurrent residual convolutional neural networks based u-net (r2u-net)," in *NAECON 2018-IEEE National Aerospace and Electronics Conference*. IEEE, 2018, pp. 228–233.
24. T. Mostafiz, I. Jarin, S. A. Fattah, and C. Shahnaz, "Retinal blood vessel segmentation using residual block incorporated u-net architecture and fuzzy inference system," in *2018 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, 2018, pp. 106–109.
25. H. Li, A. Zhygallo, and B. Menze, "Automatic brain structures segmentation using deep residual dilated u-net," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*. Springer, 2019, pp. 385–393.
26. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 424–432.
27. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.
28. Y. Peng, M. Sonka, and D. Z. Chen, "U-net v2: Rethinking the skip connections of u-net for medical image segmentation," *arXiv preprint arXiv:2311.17791*, 2023.
29. 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*, vol. 35, pp. 29 582–29 596, 2022.
30. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.
31. H. Bao, L. Dong, F. Wei, W. Wang, N. Yang, X. Liu, Y. Wang, J. Gao, S. Piao, M. Zhou *et al.*, "Unilmv2: Pseudo-masked language models for unified language model pre-training," in *International conference on machine learning*. PMLR, 2020, pp. 642–652.
32. H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3588–3597.
33. H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3464–3473.
34. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
35. S. Fu, Y. Lu, Y. Wang, Y. Zhou, W. Shen, E. Fishman, and A. Yuille, "Domain adaptive relational reasoning for 3d multi-organ segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. Springer, 2020, pp. 656–666.