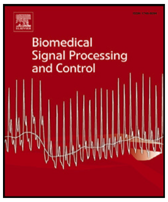
Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Cross Pyramid Transformer makes U-net stronger in medical image segmentation

Jinghua Zhu [a,*], Yue Sheng [a], Hui Cui [c], Jiquan Ma [a], Jijian Wang [d], Heran Xi [b]

[a] School of Computer Science and Technology, Heilongjiang University, Harbin, 150000, China
[b] School of Electronic Engineering, Heilongjiang University, Harbin, 150001, China
[c] Department of Computer Science and Information Technology, La Trobe University, Melbourne, 3000, Australia
[d] College of Information, Liaoning University, Shenyang, 110036, China

ARTICLE INFO

Keywords:
Multi scale global context
Pyramid vision transformer
Medical image segmentation
Convolutional neural network
U-Shaped Network
Parallel branch

ABSTRACT

Accurate auto-medical image segmentation, which is essential in disease diagnosis and treatment planning, has been significantly prospered by recent advances in Convolution Neural Network (CNN) and Transformer. However, pure CNN-based or pure Transformer-base architectures exhibit limitations for segmentation tasks. For example, CNN fails to capture long-range relations due to fix receptive fields, and Transformer ignores pixel-level spatial details of features. To address these challenges, we propose a novel parallel hybrid architecture for medical image segmentation, which is named CPT-Unet (Cross Pyramid Transformer U-shape Network). CPT-Unet may be the first attempt to exploit both the advantages of Pyramid Vision Transformer (PVT) and CNN to the full by integrating them into the standard U-shape network to improve the segmentation performance and inference time. Specifically, we design a parallel dual branch encoder and decoder in CPT-Unet that consists of CNN and PVT. The input image is fed into the encoder of these two branches simultaneously to extract low-level spatial details and global contexts in a much shallower way. We design a novel fusion strategy to adequately utilize the multi-scale features extracted from CNN and PVT. We also add PVT in the decoder to better restore the segmentation map. Experiments on two public segmentation datasets demonstrate the improved performance of the proposed CPT-Unet over the comparison methods. The source code is available at https://github.com/ShengYue007/CPT-Unet.git.

1. Introduction

Medical image segmentation is an important topic in the field of medical image analysis [1]. Among them, such as lesion segmentation [2], organ segmentation [3] and heart segmentation [4], aim to localize multi-type regions by assigning class label to each pixel of the image. Early medical image segmentation methods, including edge detection [5], template matching technology [6], statistical shape model [7], active contour [8], and machine learning [9], are inefficient for modeling feature representation. In recent years, various deep learning-based methods have emerged as the mainstream in the medical image segmentation field. The most popular way is to adopt the convolutional neural network (CNN) such as fully convolutional networks(FCN) [10], U-shape network and its variants [11–16], which are all encoder–decoder architecture. Although CNN-based methods can well model local features of the image, due to the limited receptive field of convolution operation, CNNs are hard to capture the global dependencies of features, which is specially important information in semantic segmentation. The existing methods expand the receptive field by gradually upsampling and increasing the number of convolutions, which lead to the problems such as the loss of low-level features and the local details.

To address the above challenge, Transformer-based methods have been introduced into the field of medical image segmentation [17–19]. Transformer [14] which is originally used for sequence to sequence prediction task in NLP field, has attracted intensive research interests in image segmentation and classification application field. By modeling the links between feature tokens and adopting a self-attention mechanism, the Transformer-based methods show the advantage in capturing global contextual information. In recent several years, Transformer has been used for the image recognition and object detection tasks in the computer vision field and has been proved of good performance. Vision Transformer (VIT) [20] is regarded as the first-line framework for visual processing tasks, especially successfully used in image classification tasks. However, due to low resolution, high computational complexity and other reasons, it is not suitable for dense pixel-level prediction tasks such as segmentation and target detection. A general framework

[*] Corresponding author.
E-mail addresses: zhujinghua@hlju.edu.cn (J. Zhu), xiheran@hlju.edu.cn (H. Xi).

https://doi.org/10.1016/j.bspc.2023.105361
Received 18 January 2023; Received in revised form 21 July 2023; Accepted 5 August 2023
Available online 17 August 2023

Pyramid Vision Transformer (PVT) for intensive prediction [21] is proposed as the backbone of downstream intensive prediction task. By designing a gradual reduction pyramid and hole reduction attention layer, increased scale and higher resolution outputs can be obtained under limited resources. Although these efforts have got satisfactory results, pure Transformer-based models still have limitations in capturing local details due to the lack of spatial generalization bias, resulting in reduced discrimination between background and foreground, which significantly impacts the segmentation performance.

To enjoy the benefits of Transformer-based and CNN-based methods simultaneously, in recent years, many researchers are committed to combining CNN and Transformer to form a new hybrid model, such as TransUnet [22], TransFuse [23] and TransClaw U-net [24]. As the pioneer of this series of work, TransUnet obtains low-level features through Resnet50 and captures long-range correlations by encoding these low-level features through Transformer. By combining CNN and Transformer in a stacked manner, TransUnet has achieved good segmentation performance. However, previous works mainly focused on stacking features sequentially obtained by Transformer and CNN, ignoring the fact that features of different sizes in each layer of CNN contain different global information. In this paper, we propose a hybrid framework named Cross Pyramid Transformer U-net (CPT-Unet) to further leverage Pyramid Transformer's advantages in medical image segmentation. Specifically, we design a parallel dual branch encoder in CPT-Unet that consists of a CNN and a PVT. The input image is fed into these two branches simultaneously to capture the global dependency and low-level spatial details in a much shallower manner. Moreover, to adequately utilize the multi-scale features from CNN and PVT, we design a novel fusion block as the core component of our CPT-Unet. Each module fuses the low-level features extracted by CNN and the high-level semantic context features obtained by PVT through the fusion block to obtain richer features. The fused feature is delivered to the decoder to obtain the segmentation prediction. This fusion method enables us to obtain global context information from CNN features of different scales, thus obtaining excellent feature representation capability.

In summary, the contributions of this paper are four-fold:

(1) The primary contribution of this paper is that we propose a novel hybrid parallel architecture CPT-Unet for multi-organ segmentation. To the best of our knowledge, CPT-Unet is the first parallel style hybrid architecture that synthesizes the advantage of CNN and PVT. This architecture is different from existing methods that mostly focus on extracting and fusing semantic features from single scale semantic features. To achieve this target, we design three modules including a hybrid parallel encoder, a feature fusion block, and a hybrid parallel decoder, leading to the following three technical contributions.

(2) We present a parallel dual branch encoder in CPT-Unet, which integrates CNN and PVT into the U-shaped network. The input image is fed into these two branches simultaneously to cooperatively learn the feature representation. In this way, we can explore the advantage of CNN and Transformer to the full and thus enhance the feature extract ability of the encoder.

(3) We propose a novel feature fusion block that ensembles the multi-scale features extracted from the CNN encoder and the Transformer encoder to obtain accurate feature representation, including both high-level semantics and low-level details.

(4) We redesign the decoder by substituting the traditional Upsampling with PVT and add skip-connections at each scale to integrate low-level features. In this way, CPT-Unet can better restore the segmentation map and thus further improve the segmentation performance. We conduct experiments to compare CPT-Unet with a wide range of baseline segmentation approaches. The experiment results show that the proposed CPT-Unet surpasses the state-of-the-art methods by over 1% in the task of multi-organ segmentation on Synapse. For the automated cardiac diagnosis task on ACDC dataset, CPT-Unet outperforms TransUnet by nearly 2.44% in average.

The rest of this paper is organized as follows: In Section 2, we give a brief survey of the related work of segmentation methods. Then we overview CPT-Unet and explain each component of it in Section 3. We introduce the experiment datasets and implementation details of the performance evaluation of CPT-Unet in Section 4. In Section 5, we conclude the technical contributions and findings of this paper. We also analyze the shortcomings and prospects of this paper in Section 6.

2. Related work

In this section, we first introduce the traditional methods for multi-organ segmentation, then we summarize the most typical CNN-based methods used in medical image segmentation. Next, we review vision Transformers in the field of medical image segmentation. Finally, we make an overview of the CNN and Transformer hybrid methods and compare these methods with our proposed method.

2.1. Traditional multi-organ segmentation

Conventional statistical approaches predict organ labels by establishing statistical models of the shape and location distribution of different organs, which involves co-registering images [7,25–27]. The multi-atlas label fusion methods make new segmentations by combining the propagated reference segmentations estimated on the register images in training dataset [25,28–31]. The image registration accuracy has significant effect on both statistical methods and multi-atlas label fusion methods, because organs' shape, size, and appearance usually demonstrate significant differences among patients with different diseases and treatments. To solve the challenges of registration methods, registration-free methods are proposed to train the segmentation models on unregistration images. Hand-crafted organ features are used by some methods [32,33]. The well-selected image features are proved useful by many recent multi-organ segmentation approaches [34,35]. With the rapid advance of deep learning, a large number of new methods have sprung up which can be broadly classified into three categories: CNN-based methods, Transformer-based methods and hybrid methods. We will introduce these methods in the following subsections in details.

2.2. CNN-based methods

CNN, a fundamental framework for deep learning, has powerful image processing and analyzing ability. At present, traditional machine methods have been substituted by CNN in medical image segmentation field. Since 2015, researchers have proposed Unet [14] and achieved competitive results in the subsequent studies. Subsequently, a large number of researchers devoted to the improvement of Unet, which led to the emergence of several new models based on the U-shaped structure, for example, Res-Unet [36], Attention Unet [37], Unet++[12] and so on. To further improve segmentation performance, some improved convolution operations are introduced, such as deformation convolution [38] and depth convolution [39]. BCU-net [40] seamlessly bridged U-net and ConvNeXt, not only mitigated the class imbalance problem, but also took full advantage of the complementary local and global pathological semantics. BCU-net combined semantic information from a deeper, low-resolution layer with local information from a shallower, higher-resolution layer through skip connections. In recent years, a large number of studies focused on designing various skip connection operations which is an important technique of U-shaped network [41,42]. The bottleneck between the encoder and decoder is used to force the model to learn a compressed representation of the input image, which only contains the important and useful information to restore the segmentation map in the decoder. To this end, various modules are designed to recalibrate and highlight the most discriminant features [43–45]. Because of their powerful learning ability, CNN and its variants have become the essential models for medical image segmentation tasks. However, their receptive fields are limited, which makes them inefficient in capturing global contexts.

2.3. Transformer-based methods

Vaswani et al. [14] propose Transformer for the first time for machine translation in NLP. Because of its strong ability of global context modeling, a large number of researchers began to study how to use Transformer to improve the performance of various computer vision tasks. Vision Transformer (ViT) proposed by Dosovitskiy et al. [20] was the first Transformer used for computer vision tasks. ViT is directly applied to full-size images through a Transformer with global self-attention. Gao et al. [46] applied ViT on both 2D and 3D CT scans to diagnose COVID-19. They proposed to construct an image sub-volume by extracting a fixed number of slices thereby normalizing image sequences with a varying number of slices. They also proved that the performance of ViT is better than that of Densenet, which is a competitive CNN model. Shome et al. [47] proposed ViT based model for COVID-19 diagnosis by training the model on a self-collected large COVID-19 chest X-ray image dataset. However, the image resolution of ViT output is low and the memory consumption is high. Wang et al. [47] proposed PVT to perform intensive partition training on images to achieve the effect of high output resolution. PVT can also reduce the computation load of large feature maps through gradually shrinking pyramids, thus improving the performance of processing multiple downstream tasks. PVT has achieved the best performance in target detection, semantic/instance segmentation, and other tasks. Dong et al. [48] proposed a new image polyp segmentation framework, named Polyp-PVT, which utilized PVT as the backbone in the encoder to explicitly extract more powerful and robust features. ResPVT [49] also used PVT as a backbone and far surpassed previous networks not only in accuracy but also with a much higher frame rate which may drastically reduce costs in both real time prediction and offline analysis.

2.4. Hybrid methods

Recent research has proved that the CNN and Transformer hybrid architecture is helpful in exploring their advantages to the full.TransUnet [22] is the first to prove that Transformer can promote the feature representation capability of the encoder. It encodes global context by Transformer and preserves the low-level CNN features using skip connections in a U-shaped hybrid architecture. The hybrid feature extraction network adopts lightweight processing to adapt to actual application scenarios [50]. MT-Unet [51] is another hybrid medical image segmentation model that learns the inter- and intra-affinities simultaneously through a Transformer module. TransFuse [23] combines Transformer and CNN into a segmentation network in a similar style. In this way, TransFuse can capture both the long-range relations and the low-level features. Zhu et al. [52] propose to improve the segmentation training by introducing shift patch tokenization strategy based on improved Swin-Transformer, and design an edge detection module based on CNN to extract edge features from input MRI scans to combine deep semantics and edge information in multi-modal MRI. DSTransUnet [53] uses two scales Swin Transformer in the encoder–decoder framework to simultaneously learn the coarse-grained and fine-grained feature representation. Li et al. [54] designed a dual encoder strategy that uses CNN and Transformer to form a dual feature extraction encoder to obtain both local and global features in parallel. These two kinds of features enable remote connection of dual-stream context information by bypassing the connection benign interaction.

Though the above studies have achieved big success in medical image segmentation, they all ignore the fact that the global semantic context differs in different scales feature maps. Different from the above hybrid architectures, we combine CNN and PVT in parallel style in the encoder to extract local feature and long-range relation. Furthermore, to better model the global relationship, we put forward a strategy to fuse multi-scale semantic context from the feature map output by CNN and PVT. We also apply Transformer to the decoder of the hybrid U-shaped network to further improve the segmentation performance.

3. Proposed method

In this section, we first describe the overall framework of CPT-Unet, and then introduce each component of it in details.

3.1. Framework of CPT-Unet

Fig. 1 shows the framework of the proposed CPT-Unet which is based on the classic U-shape encoder–decoder architecture. We design a hybrid encoder for CPT-Unet to enhance the feature representation ability. Specifically, we design the encoder to be CNN-Transformer hybrid model which contains CNN branch and Transformer branch. Unlike the existing hybrid models that simply stack CNN and Transformer or train them separately, our CPT-Unet explores a parallel and interactive way to train CNN and Transformer. The CNN branch takes the input medical image and down-samples it through multiple CNN blocks to extract different scales features from the input medical image. The Transformer branch is based on PVT which divides the input medical image into a number of P×P size patches and tokenizes them into a tokenized sequence, which is then used as the input of PVT. We design a fusion block to fuse the CNN feature and Transformer feature. The output features of each Transformer group are fused with the same level output feature from CNN, and then the fused features are passed to the next CNN block and Transformer block. For the Transformer block, it is necessary to rearrange the features to the form of tokens first. Then we reshape the hidden feature output from the last fusion block and use PVT as the decoder to recover the pixel-level full-resolution prediction map through four PVT blocks and skip-connections. In the following subsections, we will introduce each of the component in details.

3.2. CNN branch

The traditional organ segmentation method is to obtain the global context of features through CNN with hundreds of layers, such a deep model leads to a great consumption of resources. Inspired by TransUnet [22], we use Resnet50 as the CNN backbone, and only keep three CNN modules, which not only provides us with a shallower model but also keeps richer low-level spatial details. We downsample the input image layer by layer to get ($F^1 \in R^{H\times W\times C_1}$), ($F^2 \in R^{\frac{H}{2}\times\frac{W}{2}\times C_2}$), ($F^3 \in R^{\frac{H}{4}\times\frac{W}{4}\times C_3}$) respectively, where C, H, W in brackets denote the channel number, height and width of image respectively. F^i represents the features obtained through the ith CNN module.

3.3. Transformer branch

We adopt PVT as the Transformer branch for the following reasons. First, PVT has fewer parameters compared with ViT, it can train a larger model with the same computing resources and improve the performance of the model. Second, PVT is faster in processing images and can complete reasoning tasks faster. Third, PVT introduces a pyramid structure which can extract features at different scales, thus better capturing the multi-scale feature of images and improving the performance of the model. As shown in Fig. 1, the first step of Transformer is image serialization. We divide the input image x according to its size and get $N = HW/P^2$ patches, each patch is of size $P \times P$. N is the number of patches, and its size is determined by the length of the sequence. Subsequently, each patch is mapped into a one-dimensional vector by linear mapping and then labeled (Eq. (1)) to reconstruct the input image x to 2D patches flattened sequence $\{ x_p^{i,j} \in R^{p^2\cdot C} \mid i = 1,\dots,3; j = 1,\dots,N\}$,a matrix of vector, the size of patches changes with the change of the dimensionality of the matrix–vector.

$$T_0^i = \left[x_p^{i,1}E_i + x_p^{i,2}E_i + \cdots + x_p^{i,N}E_i\right] + E_{pos} \tag{1}$$

where $E_i \in R^{(P^2\cdot C_i)\times D}$ is the projection of patch embedding, $E_{pos} \in R^{N\times D}$ is the position embedding, D is the dimension of features.

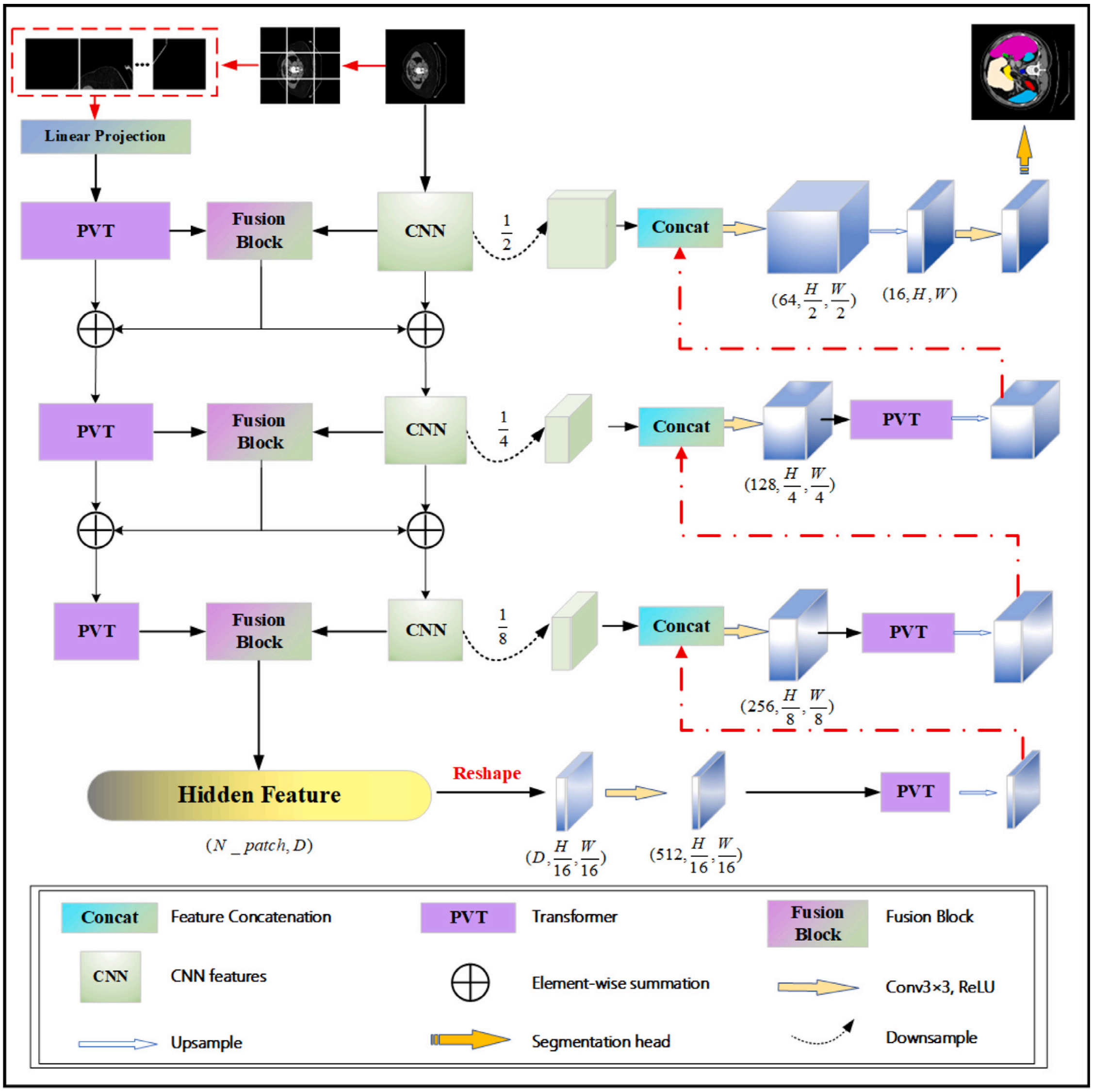


Fig. 1. The framework of the proposed CPT-Unet model.

As shown in Fig. 3, the PVT module of our CPT-Unet has L layers each of which contains spatial reduction attention (SRA) and multi-layer perceptron(MLP), the output of the lth layer can be written as follows:

$$\tilde{T}_l^i = SRA\left(LN\left(T_{l-1}^i\right)\right) + T_{l-1}^i \tag{2}$$

$$T_l^i = MLP\left(LN\left(\tilde{T}_l^i\right)\right) + \tilde{T}_l^i \tag{3}$$

where $LN(\cdot)$ is the layer normalization operator, T_l^i is the encoded feature representation of the ith scale.

In PVT, we set the patch size to 16 × 16 and the sequence length to 196. This setting allows for the segmentation of input images into smaller blocks, further improving the model's receptive field and ability to capture fine details. At the same time, setting the sequence length to 196 enables the model to better learn the relationship between local and global features, thus improving its performance.

3.4. Fusion module

The existing CNN-Transformer hybrid model usually stacks CNN and Transformer, or trains them separately, alone in parallel, and finally fuses the features learned from them. Such strategy limits the extraction of features at different scales, and makes it difficult to optimize the segmentation performance due to the differences in the features and computing mechanisms of the two branches. As shown in Fig. 1, we extract several intermediate features at different scales in the CNN branch and Transformer branch, fuse them through the fusion block and stream the fusion results back to the original two branches. This method can make up the performance difference between different branches to a certain extent and avoid losing information of different scales.

As shown in Fig. 2, we first rearrange the Transformer tokens into the form of $C \times H \times W$ and concatenate them with CNN features. Then by a Channel-wise fusion operation (consisting of 1 × 1 convolution, ReLU function, and residual connection), we get the fusion feature M_i and split M_i into $M_i^T \in R^{H \times W \times C}$ and $M_i^F \in R^{H \times W \times C}$ along channel dimension. M_i^T and M_i^F are passed into MLP block and convolution block respectively. Each fused feature is passed back to the CNN and Transformer branches and added to the original input features T_i and F_i respectively.

$$M_i^T, M_i^F = Split\left(CWF(Cat(P(T_i), F_i))\right) \tag{4}$$

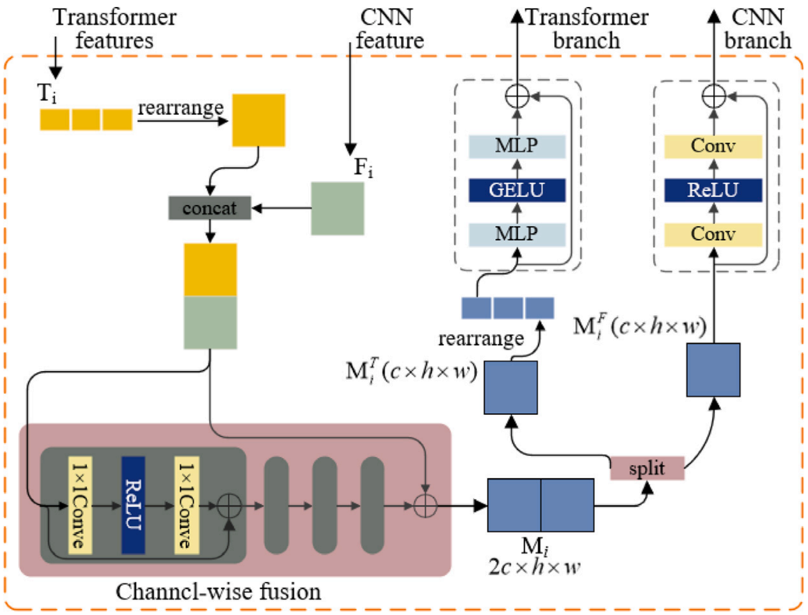


Fig. 2. The architecture of the proposed fusion block.

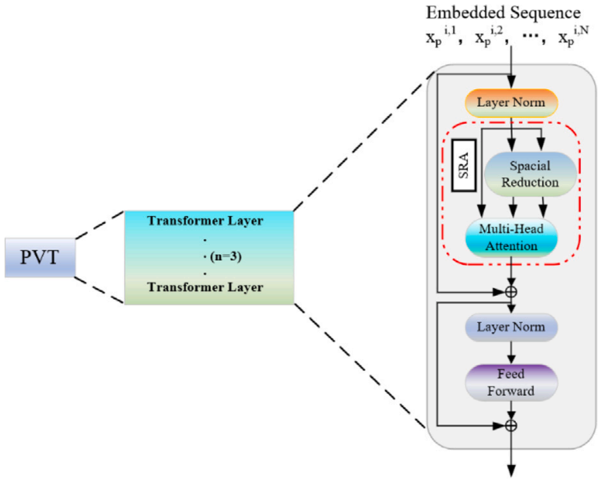


Fig. 3. The architecture of the PVT block.

$$\tilde{M}_i^F = CNN\_Branch(M_i^F) \tag{5}$$

$$\tilde{M}_i^T = Trans\_Branch\left(P\left(M_i^T\right)\right) \tag{6}$$

$$F_{i+1} = F_i + \tilde{M}_i^F \tag{7}$$

$$T_{i+1} = T_i + \tilde{M}_i^T \tag{8}$$

Where $P(\cdot)$ is reordering of dimensions, $Cat(\cdot)$ is stitching of features, $CWF(\cdot)$ is channel-wise fusion operation, and $Split(\cdot)$ is segmentation of features. F_i and T_i are the input features of CNN and Transformer of layer i respectively. $\tilde{M}_i^T$ and $\tilde{M}_i^F$ are the fusion features that flow back to CNN and Transformer branch, respectively.

CPT-Unet is a hybrid model, where the central role of CNN is to acquire local features, and Transformer is used to capture global contextual information. Then the two are fused and fed back to the original branch. We design the architecture in this way for the following considerations: (1) We make full use of feature information under different dimensions by fusing features of different scales of CNN and PVT in multiple branches, which significantly improves the positioning ability and boundary segmentation accuracy; (2) By fusing the low-level features extracted by CNN and high-level semantic features extracted by PVT, more comprehensive features can be obtained. If this feature is fed back to the original branch, the feature information of the original branch can be further improved.

3.5. Decoder

In the pixel recovery process, we combine the two decoding results by introducing cascaded up-sampling and skip-connections connections to keep the detailed information of the image. In addition, we also introduce PVT modules in the decoder to further explore long-range contextual information during the upsampling process. As shown in the right part of Fig. 1, The decoder consists of three stages. Unlike U-net and its variants, each stage of our model includes not only nearest neighbor upsampling and skip connections, but also PVT blocks. Specifically the sequence of hidden features output by the encoder is reshaped and passed to the cascaded up-sampler. The reconstructed features are first fed into the PVT module. We choose this design because: (1) PVT modules can further improve the performance and generalization ability of the model through techniques such as multi-head attention and residual connections; (2) they can build long-range dependencies and global context interactions in the decoder for better decoding performance. To reduce the loss of local details such as the boundaries and shapes of organs due to reduced image resolution, we use multiple cascaded up-sampling blocks including 2 × up-sampling operators, 3 × 3 convolutional layer, and ReLU layer to recover the original resolution. To solve the gradient disappearance, during the up-sampling procedure, we use three skip-connection operations to fuse the up-sampled features obtained in the three stages with the output of the 3rd/2nd/1st CNN feature. The skip-connections make our model obtain a more accurate prediction image.

4. Experiments and analysis

We evaluate the segmentation performance of CPT-Unet through a series of experiments on several medical image datasets. We first briefly describe the datasets and evaluation metrics used in the experiments. Then CTP-Unet is experimentally compared with other SOTA

Table 1
Segmentation performance comparisons on Synapse dataset (average dice score and dice score for each organ). The best performance are highlighted in bold.

Method	DSC (%)↑	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
V-net [15]	68.81	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR [16]	69.77	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50 U-net [11]	74.68	87.74	63.66	80.60	78.19	93.74	50.90	85.87	74.16
Unet [11]	76.85	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
R50 AttnUnet [17]	75.57	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
AttnUnet [37]	77.7	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
Contexnet [56]	71.17	79.92	51.17	77.58	72.04	91.74	43.78	86.65	66.51
DABnet [57]	74.91	85.01	56.89	77.84	72.45	93.05	54.39	88.23	71.45
EDAnet [58]	75.43	84.35	62.31	76.16	71.65	93.20	53.19	85.47	77.12
Enet [59]	77.63	85.13	64.91	81.10	77.26	93.37	57.83	87.03	74.41
TransUnet [22]	77.48	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
FSSnet [60]	74.59	82.87	64.06	78.03	69.63	92.52	53.10	85.65	70.86
TransClaw U-net [24]	78.09	85.87	61.38	84.83	79.36	94.28	57.65	87.74	73.55
CoTra [61]	74.59	82.87	64.06	78.03	69.63	92.52	53.10	85.65	70.86
Swin-Unet [62]	79.13	85.47	66.53	83.28	79.61	94.29	56.58	**90.66**	76.60
MT-Unet [51]	78.08	85.87	61.38	84.83	79.36	94.28	57.65	87.74	73.55
LeVit-Unet [63]	78.53	87.33	62.23	84.61	80.25	93.11	59.07	88.86	72.76
WAD [64]	80.30	87.73	69.93	83.95	79.78	93.95	61.02	88.86	77.16
MORSE [65]	80.85	88.92	67.53	84.83	81.68	**96.83**	59.70	87.73	**79.58**
CPT-Unet	**81.93**	**89.68**	**70.37**	**85.09**	**81.99**	94.75	**66.38**	90.21	77.96

methods. To further demonstrate the effectiveness of our model, we conduct several ablation experiments under different settings. We also train and test our model on other medical image dataset to evaluate the generalization ability of CPT-Unet. The results of the qualitative and quantitative comparison with different methods and the ablation experiments are given and analyzed.

4.1. Dataset and metrics

TransUnet [22] and other hybrid models use Synapse and ACDC to evaluate their model performance, to be fair for comparing with these models, we also use the same datasets in our experiments.

Synapse is a public multi-organ segmentation dataset (Synapse)[1]: 30 abdominal clinical CT-enhanced scan cases from MICCAI 2015 multi-atlas abdominal marker challenge are used. The dataset contains 3779 axially contrast-enhanced abdominal clinical CT images. The CT images followed the settings in [22]. 18 (2212 axial slices) cases were randomly selected as training data, and 12 (1567 axial slices) cases were used as the test data sets. We use the same performance metrics as in [55], the average Dice Similarity Coefficient (DSC) and the average Hausdorff-95 distance (HD95) are used to evaluate the segmentation performance, We evaluate the performance of eight organs such as aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach in CT images. DSC is used to measure the similarity of the two sets and takes the value range of [0,1], with larger values indicating that the two sets' DSC is sensitive to the internal filling of the mask. HD95 is sensitive to the segmented boundary.

Automated Cardiac Diagnostic Challenge (ACDC)[2]: is collected on an MRI scanner for different patient exams. Cine MR images acquired in the breath-hold state were obtained. The operation and setup were the same as [22], and each patient scan was manually annotated with basic myocardial (MYO), left ventricular (LV), and right ventricular (RV) conditions. A total of 100 samples were included in the dataset, of which 70 were used for training (containing a total of 1930 axial slices), 10 were used for validation, and the remaining 20 were used for testing.

4.2. Implementation details

All experiments were conducted on NVIDIA GeForce RTX 2060 GPU. The input images were adjusted to 224 × 224. We did simple data enhancements, such as random rotation and random flip. In addition, stochastic gradient descent (SGD) was used with an initial learning rate of 1 $\times 10^{-2}$, a momentum of 0.9, and a weight decay of 1 $\times 10^{-4}$. We set the batch size to 4 and trained 200 epochs. The above parameters are all the optimal parameters that obtained through exhaustive methods. In order to make the experiment results reliable and stable, and reduce the errors caused by randomness, all experimental results were obtained based on five times five-fold cross-validation. The result with the best performance on the validation set is selected as the final model of training.

4.3. Comparison with existing methods

We compare our model with several current state-of-the-art models, including (1) V-net [15]; (2) DARR [16]; (3) U-net [14]; (4) AttnUnet [37]; (5) TransU-net [22]; (6) TransclawU-net [24]; (7) Swin-Unet [62]; (8) MT-Unet [51]; (9) LeVit-Unet [63]; (10) WAD [64]; (11) MORSE [65]. We give the compare results in Table 1 and find that CPT-Unet achieves the best results in terms of evaluation metrics with DSC value of 81.83%. The DSC values of all the other models are below 81%. Compared with the state-of-the-art MORSE model, our model achieves an accuracy improvement of 1.08% in DSC. The above results demonstrate that our model has better segmentation performance. We can also see that CPT-Unet has the best segmentation results in 7 out of 8 organs, which is an evidence of the effectiveness of our model in discriminating classes.

4.4. Ablation experiment

4.4.1. The impact of fusion strategy

The introduction of Transformer in the encoder enables better capture of global contextual information. To evaluate our proposed CPT-Unet model, we compare our implementation with three different fusion strategies, namely CNN features and Transformer at different scales without fusion (CPT-Unet-LRD), CNN features and Transformer fused and then passed into the Transformer module (CPT-Unet-LD) and CNN features and Transformer are fused and then passed into the

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

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

Table 2
Ablation study on fusion strategy on Synapse dataset.

Method	DSC (%)↑	HD95 (mm) ↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
CPT-Unet-LRD	78.86	30.38	88.38	62.43	82.26	79.32	94.40	59.74	88.07	76.28
CPT-Unet-LD	79.29	29.17	88.23	65.05	80.01	76.58	94.15	62.69	90.80	76.84
CPT-Unet-RD	79.20	28.69	88.69	66.02	82.36	80.81	94.28	61.31	88.53	71.61
CPT-Unet-D	81.48	**16.96**	88.29	**72.00**	84.92	81.35	94.71	63.65	**91.43**	75.46
CPT-Unet	**81.93**	19.46	**89.68**	70.37	**85.09**	**81.99**	**94.75**	**65.38**	90.21	**77.96**

Table 3
Ablation study on the number of fusion block on Synapase dataset.

Number	DSC (%)↑	HD95 (mm) ↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
1	78.20	33.43	86.76	64.24	79.34	75.34	94.05	61.24	89.09	75.52
2	80.59	23.22	87.35	67.17	84.63	80.41	94.89	62.92	**91.73**	75.65
3	**81.93**	**19.46**	**89.68**	**70.37**	**85.09**	**81.99**	94.75	**65.38**	90.21	77.96
4	81.13	20.40	88.49	67.46	83.89	81.46	**95.31**	62.55	90.85	**79.03**

Table 4
Ablation study on the number of Transformer layer on Synapase dataset.

Number	DSC (%)↑	HD95 (mm) ↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
1	78.96	38.17	84.42	66.46	79.63	79.61	92.31	62.55	89.21	77.47
2	80.05	28.59	86.49	67.36	81.59	81.46	94.97	61.35	90.03	77.15
3	**81.93**	**19.46**	**89.68**	**70.37**	**85.09**	**81.99**	94.75	**65.38**	90.21	77.96
4	80.59	24.61	87.39	68.51	82.57	80.89	**95.14**	61.46	**90.63**	**78.14**

CNN module (CPT-Unet-RD). Table 2 shows that the detection accuracy and localization accuracy results are poor when the intermediate features are not fused. We observe that the performance improves when the intermediate features are fused in one direction. The DSC index growth is relatively large when the intermediate features are passed into the CNN module, indicating that the Transformer makes up for the lack of CNN's ability to capture global contextual information. When intermediate features are passed into the Transformer module, the HD95 is small, which indicates that CNN makes up for the lack of the Transformer's ability to capture detailed information. The experimental results also show that the best performance is obtained when the intermediate features are fused in both directions, and the edge detection accuracy and localization accuracy are substantially increased. In order to explore the influence of PVT on the decoder, we conduct ablation experiments on the decoder. As shown in Table 2 (CPT-Unet-D), adding PVT to the decoder can effectively improve the segmentation performance, which is 0.45% higher than that without adding PVT to the decoder. Through the above analysis, we can get the conclusion that the two branches in CPT-Unet can extract the missing information of the other respectively and the intermediate bidirectional fusion strategy is helpful for obtaining satisfactory convergence results, and the PVT in the decoder can effectively improve the segmentation performance.

4.4.2. The effect of the number of fusion block

As mentioned above, the fusion block enables the model to capture long-range dependencies while obtaining local spatial detail information by fusing the CNN and Transformer modules. In order to further evaluate the change of overall performance, different numbers of fusion blocks are added to the encoder module. We set the number of fusion blocks to 1/2/3/4, and observe from Table 3 that when the number of fusion modules is 3, the result is the most accurate. This is because the more PVT layers, the more difficult it will be to train the model, and the more prone to over-fitting.

We further conduct ablation experiments to explore the influence of PVT layers on the model. As shown in Table 4, when the number of layers of each PVT module is 3, the segmentation details of the model are the best, and when the number of layers of PVT exceeds 3, the segmentation effect decreases, which further proves the effectiveness of our model.

These ablation experiments confirm our intuition of building multi-scale joint fusion in CNN-Transformer hybrid encoder to obtain accurate advanced semantic learning.

4.5. Qualitative results

We present the results of qualitative comparisons on the Synapse dataset. Fig. 4 shows that (1) TransUnet suffers from over-segmentation or inaccurate segmentation (e.g., the first row shows that the liver is over-segmented. The third row shows that stomach is under-segmented), which is because serially connected CNN and Transformer cannot fully exploit the capability of both. (2) MT-Unet has a false-positive problem (in the first row, the right kidney is misjudged; in the second row, the left kidney part is detected incorrectly), which indicates that CPT-Unet is more effective than other methods in suppressing the noise prediction. (3) Our method shows the segmentation of the gallbladder and right kidney with an overwhelming advantage. (e.g., in the first and third rows, the gallbladder and right kidney are not detected or incorrectly detected by other models.) We also far outperform other models in the ability to segment the left kidney and spleen due to the balanced perception of the local and global environment, which has the consistent results with that in Tables 1 and 4 . We can observe that the predictions of TransUnet and MT-Unet tend to have rougher boundaries and shapes than the results predicted by CPT-Unet (e.g., pancreas in the third row).

We also visualize the ablation study results as shown in Figs. 5 and 6. From Fig. 5, we find that the segmentation performance is the worst when the number of fusion modules is 1, and the segmentation performance is the best when the number of fusion modules is 3; When the number of fusion modules is 4, there is a false positive. Fig. 6 shows that CPT-Unet-LR had false positive and incomplete segmentation in the first

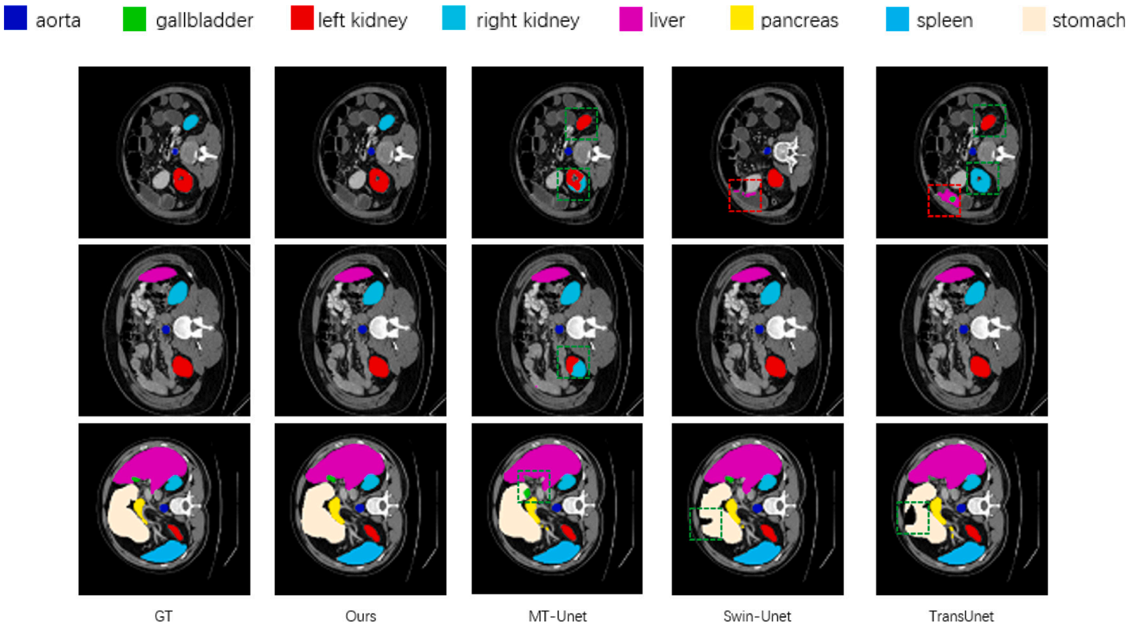


Fig. 4. Qualitative results of comparison methods of different approaches.

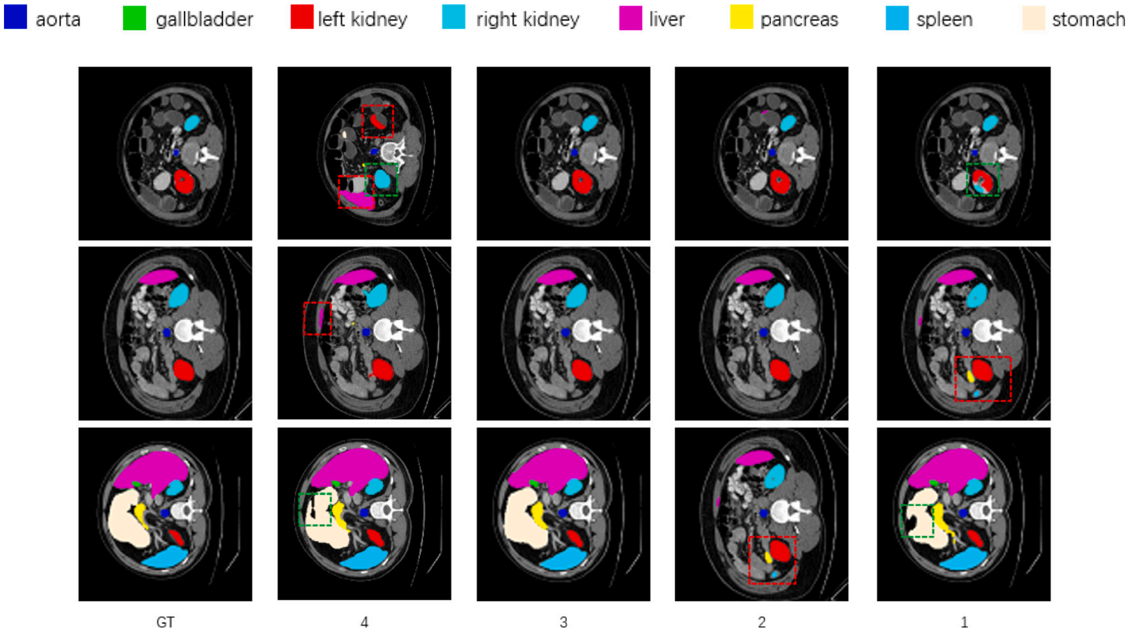


Fig. 5. Qualitative comparison of different number of fusion blocks.

two cases. In the last case, the gastric segmentation was incomplete, and the boundary was not smooth. It can be seen intuitively from Figs. 5 and 6 that our method has the best visual effect of segmentation.

4.6. Generalization to other dataset

We conduct experiments on the MR dataset ACDC with automatic heart segmentation to verify the generalization ability of CPT-Unet. We observe from Table 5 that the performance of CPT-Unet is better than the pure CNN-based methods (Unet and Attnuet) and other CNN and Transformer combined methods (TransUnet, MT-Unet and Swin-Unet). To thoroughly evaluate our CPT-Unet model, we conduct the same ablation experiments in Section 4.4 on the ACDC dataset(see Table 6,Table 7 and Table 8). DSC and HD95 metrics and the most accurate test on all three different annotations. Table 7 shows the effect of the number of fusion blocks on the performance, the detection accuracy of our model detection increases as the number of fusion blocks increases. The experiments in this Section further verify that our method can be easily generalized to other datasets and tasks.

5. Conclusion

We propose a new medical image segmentation network named CPT-Unet which fully utilizes CNN's ability to extract local features

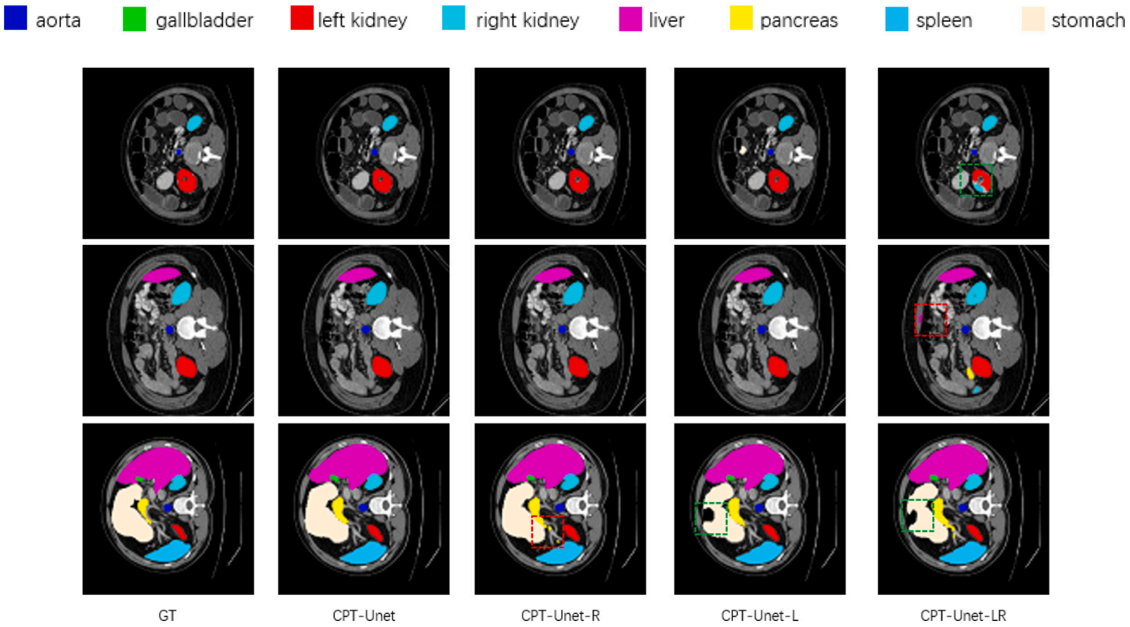


Fig. 6. Qualitative comparison of ablation study of fusion strategy.

Table 5
Comparison of model performance on the ACDC dataset (average dice score).

Method	DSC (%)↑	RV	Myo	LV
Unet [14]	87.60	84.62	84.52	93.68
AttnUnet [37]	86.90	83.27	84.33	93.53
R50 Unet [11]	87.55	87.10	80.63	94.92
R50 AttnUnet [17]	86.75	87.58	79.20	93.47
nnUnet [66]	91.59	90.25	89.10	95.41
nnformer [67]	91.78	90.22	89.53	95.59
TransUnet [22]	89.71	86.67	87.27	95.18
Swin Unet [62]	88.07	85.77	84.42	94.03
MT-Unet [51]	90.43	86.64	89.04	95.62
LeVit-Unet [63]	90.32	89.55	87.64	93.76
CS-Unet [68]	91.37	89.20	89.47	95.42
CPT-Unet	**92.15**	**90.75**	**89.83**	**95.87**

Table 6
Ablation study on fusion strategy on ACDC dataset.

Method	DSC (%)↑	HD95 (mm) ↓	RV	Myo	LV
CPT-Unet-LRD	91.27	2.02	88.44	89.44	95.94
CPT-Unet-LD	91.53	1.77	89.20	89.65	95.75
CPT-Unet-RD	91.32	1.54	88.96	89.18	95.81
CPT-Unet-D	91.76	**1.15**	**89.50**	**89.72**	**96.07**
CPT-Unet	**92.15**	1.63	**90.75**	**89.83**	95.87

Table 7
Ablation study on the number of fusion block on ACDC dataset.

Number	DSC (%)↑	HD95 (mm) ↓	RV	Myo	LV
1	90.95	2.41	88.49	89.24	95.14
2	91.37	1.94	89.35	89.15	95.61
3	**92.15**	**1.63**	**90.75**	89.83	**95.87**
4	91.62	1.82	89.52	**89.92**	95.44

Table 8
Ablation study on the number of Transformer layer on ACDC dataset.

Number	DSC (%)↑	HD95 (mm) ↓	RV	Myo	LV
1	90.34	3.62	88.17	88.31	94.54
2	90.93	2.73	88.55	89.12	95.13
3	**92.15**	**1.63**	**90.75**	**89.83**	**95.87**
4	91.42	1.74	89.43	89.21	95.64

and the Transformer's powerful ability to capture global context. CPT-Unet is a hybrid architecture that combines the Transformer and CNN in parallel style through the fusion blocks. In this way, CPT-Unet integrates multi-scale low-level detail features and high-level semantic features layer-by-layer to enhance the network learning ability. We appraise a succession of tests on single and multiple organ datasets to evaluate our CPT-Unet. The results demonstrate that our CPT-Unet has good generalization and segmentation performance compared with the SOTA methods.

6. Shortcomings and prospects

The Transformer can model long range relations on the prerequisite of much more training data compared with the CNN model. However, obtaining sufficient training data for medical image segmentation could be extremely challenging due to the need for pixel-level annotation and medical experts, which are time-consuming and labor-intensive. In such cases, semi-supervised deep learning approaches become particularly important. Up to now, CNN and Transformer hybrid models for semi-supervised medical image segmentation remain under-explored. Thus, we will focus on studying semi-supervised hybrid segmentation model in our future work. In this paper, we only fit our model to the abdominal multi-organ segmentation and cardiac diagnostic tasks, we will focus on generalizing CPT-Unet to other medical tasks like vessel segmentation and landmark detection.

CRediT authorship contribution statement

Jinghua Zhu: Supervision, Conceptualization. **Yue Sheng:** Writing – original draft, Methodology. **Hui Cui:** Translate articles. **Jiquan Ma:** Writing, Revision. **Jijian Wang:** Data curation, Visualization. **Heran Xi:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgments

This work was supported by the Natural Science Foundation of Heilongjiang Province of China [grant numbers LH2022F045].

References

[1] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, Paul Kennedy, Deep learning techniques for medical image segmentation: achievements and challenges, J. Digit. Imaging 32 (4) (2019) 582–596.

[2] Yuyin Zhou, Lingxi Xie, Wei Shen, Yan Wang, Elliot K. Fishman, Alan L. Yuille, A fixed-point model for pancreas segmentation in abdominal CT scans, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 693–701.

[3] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, Pheng-Ann Heng, H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes, IEEE Trans. Med. Imaging 37 (12) (2018) 2663–2674.

[4] Lequan Yu, Jie-Zhi Cheng, Qi Dou, Xin Yang, Hao Chen, Jing Qin, Pheng-Ann Heng, Automatic 3D cardiovascular MR segmentation with densely-connected volumetric convnets, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 287–295.

[5] Djemel Ziou, Salvatore Tabbone, et al., Edge detection techniques-an overview, Pattern Recognit. Image Anal. C/C Raspoznavaniye Obrazov I Analiz Izobrazhenii 8 (1998) 537–559.

[6] S.S. Adagale, S.S. Pawar, Image segmentation using PCNN and template matching for blood cell counting, in: 2013 IEEE International Conference on Computational Intelligence and Computing Research, IEEE, 2013, pp. 1–5.

[7] Tobias Heimann, Hans-Peter Meinzer, Statistical shape models for 3D medical image segmentation: a review, Med. Image Anal. 13 (4) (2009) 543–563.

[8] Yanshan Zhang, Yuru Tian, A new active contour medical image segmentation method based on fractional varying-order differential, Mathematics 10 (2) (2022) 206.

[9] Hyunseok Seo, Masoud Badiei Khuzani, Varun Vasudevan, Charles Huang, Hongyi Ren, Ruoxiu Xiao, Xiao Jia, Lei Xing, Machine learning techniques for biomedical image segmentation: an overview of technical aspects and introduction to state-of-art applications, Med. Phys. 47 (5) (2020) e148–e167.

[10] Jonathan Long, Evan Shelhamer, Trevor Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[11] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[12] Z. Zhou, Mmr Siddiquee, N. Tajbakhsh, J. Liang, UNet++: A nested U-net architecture for medical image segmentation, 2018.

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

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, arXiv.

[15] F. Milletari, N. Navab, S. A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision, 3DV, 2016.

[16] Shuhao Fu, Yongyi Lu, Yan Wang, Yuyin Zhou, Wei Shen, Elliot Fishman, Alan Yuille, Domain adaptive relational reasoning for 3d multi-organ segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 656–666.

[17] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, Daniel Rueckert, Attention gated networks: Learning to leverage salient regions in medical images, Med. Image Anal. 53 (2019) 197–207.

[18] Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

[19] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.

[21] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, Ling Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 568–578.

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

[23] Yundong Zhang, Huiye Liu, Qiang Hu, Transfuse: Fusing transformers and cnns for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 14–24.

[24] Yao Chang, Hu Menghan, Zhai Guangtao, Zhang Xiao-Ping, Transclaw u-net: Claw u-net with transformers for medical image segmentation, 2021, arXiv preprint arXiv:2107.05188.

[25] Toshiyuki Okada, Marius George Linguraru, Masatoshi Hori, Ronald M. Summers, Noriyuki Tomiyama, Yoshinobu Sato, Abdominal multi-organ segmentation from CT images using conditional shape–location and unsupervised intensity priors, Med. Image Anal. 26 (1) (2015) 1–18.

[26] Timothy F. Cootes, Gareth J. Edwards, Christopher J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.

[27] Juan J. Cerrolaza, Mauricio Reyes, Ronald M. Summers, Miguel Ángel González-Ballester, Marius George Linguraru, Automatic multi-resolution shape modeling of multi-organ structures, Med. Image Anal. 25 (1) (2015) 11–21.

[28] Akinobu Shimizu, Rena Ohno, Takaya Ikegami, Hidefumi Kobatake, Shigeru Nawano, Daniel Smutek, Segmentation of multiple organs in non-contrast 3D abdominal CT images, Int. J. Comput. Assist. Radiol. Surg. 2 (3) (2007) 135–142.

[29] Zhoubing Xu, Ryan P. Burke, Christopher P. Lee, Rebeccah B. Baucom, Benjamin K. Poulose, Richard G. Abramson, Bennett A. Landman, Efficient multi-atlas abdominal segmentation on clinically acquired CT with SIMPLE context learning, Med. Image Anal. 24 (1) (2015) 18–27.

[30] Tong Tong, Robin Wolz, Zehan Wang, Qinquan Gao, Kazunari Misawa, Michitaka Fujiwara, Kensaku Mori, Joseph V. Hajnal, Daniel Rueckert, Discriminative dictionary learning for abdominal multi-organ segmentation, Med. Image Anal. 23 (1) (2015) 92–104.

[31] Miyuki Suzuki, Marius George Linguraru, Kazunori Okada, Multi-organ segmentation with missing organs in abdominal CT images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2012, pp. 418–425.

[32] Elena Casiraghi, Paola Campadelli, Stella Pratissoli, Gabriele Lombardi, Automatic abdominal organ segmentation from CT images, ELCVIA Electron. Lett. Comput. Vis. Image Anal. 8 (1) (2009) 1–14.

[33] Sanjay Saxena, Neeraj Sharma, Shiru Sharma, S. Singh, Ashish Verma, An automated system for atlas based multiple organ segmentation of abdominal CT images, BJMCS 12 (2016) 1–14.

[34] Herve Lombaert, Darko Zikic, Antonio Criminisi, Nicholas Ayache, Laplacian forests: Semantic image segmentation by guided bagging, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2014, pp. 496–504.

[35] Baochun He, Cheng Huang, Fucang Jia, Fully automatic multi-organ segmentation based on multi-boost learning and statistical shape model search., in: VISCERAL Challenge@ ISBI, 2015, pp. 18–21.

[36] Xiao Xiao, Shen Lian, Zhiming Luo, Shaozi Li, Weighted res-unet for high-quality retina vessel segmentation, in: 2018 9th International Conference on Information Technology in Medicine and Education, ITME, IEEE, 2018, pp. 327–331.

[37] A. Jo Schlemper, B. Ozan Oktay A, B. Michiel Schaap, C. Mattias Heinrich, A. Bernhard Kainz, A. Ben Glocker, A. Daniel Rueckert, Attention gated networks: Learning to leverage salient regions in medical images, Med. Image Anal. 53 (2019) 197–207.

[38] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, IEEE (2016).

[39] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable ConvNets v2: More deformable, better results, 2018, arXiv.

[40] Hongbin Zhang, Xiang Zhong, Guangli Li, Wei Liu, Jiawei Liu, Donghong Ji, Xiong Li, Jianguo Wu, BCU-net: Bridging ConvNeXt and U-net for medical image segmentation, Comput. Biol. Med. 159 (2023) 106960.

[41] Qiangguo Jin, Zhaopeng Meng, Changming Sun, Hui Cui, Ran Su, RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans, Front. Bioeng. Biotechnol. 8 (2020) 1471.

[42] Dmitrii Lachinov, Philipp Seeböck, Julia Mai, Felix Goldbach, Ursula Schmidt-Erfurth, Hrvoje Bogunovic, Projective skip-connections for segmentation along a subset of dimensions in retinal OCT, 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. 431–441.

[43] Changlu Guo, Márton Szemenyei, Yugen Yi, Wenle Wang, Buer Chen, Changqi Fan, Sa-unet: Spatial attention u-net for retinal vessel segmentation, in: 2020 25th International Conference on Pattern Recognition, ICPR, IEEE, 2021, pp. 1236–1242.

[44] Reza Azad, Afshin Bozorgpour, Maryam Asadi-Aghbolaghi, Dorit Merhof, Sergio Escalera, Deep frequency re-calibration u-net for medical image segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3274–3283.

[45] Reza Azad, Nika Khosravi, Dorit Merhof, SMU-net: Style matching U-net for brain tumor segmentation with missing modalities, in: International Conference on Medical Imaging with Deep Learning, PMLR, 2022, pp. 48–62.

[46] Xiaohong Gao, Yu Qian, Alice Gao, COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models, 2021, arXiv preprint arXiv:2107.01682.

[47] Debaditya Shome, T. Kar, Sachi Nandan Mohanty, Prayag Tiwari, Khan Muhammad, Abdullah AlTameem, Yazhou Zhang, Abdul Khader Jilani Saudagar, Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare, Int. J. Environ. Res. Public Health 18 (21) (2021) 11086.

[48] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, Ling Shao, Polyp-pvt: Polyp segmentation with pyramid vision transformers, 2021, arXiv preprint arXiv:2108.06932.

[49] Roi Nachmani, Issa Nidal, Dror Robinson, Mustafa Yassin, David Abookasis, Segmentation of polyps based on pyramid vision transformers and residual block for real-time endoscopy imaging, J. Pathol. Inform. 14 (2023) 100197.

[50] Yang Xu, Xianyu He, Guofeng Xu, Guanqiu Qi, Kun Yu, Li Yin, Pan Yang, Yuehui Yin, Hao Chen, A medical image segmentation method based on multi-dimensional statistical features, Front. Neurosci. 16 (2022).

[51] Hongyi Wang, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xian-Hua Han, Yen-Wei Chen, Ruofeng Tong, Mixed transformer u-net for medical image segmentation, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 2390–2394.

[52] Zhiqin Zhu, Xianyu He, Guanqiu Qi, Yuanyuan Li, Baisen Cong, Yu Liu, Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI, Inf. Fusion 91 (2023) 376–387.

[53] Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, Guangming Lu, David Zhang, Ds-transunet: Dual swin transformer u-net for medical image segmentation, IEEE Trans. Instrum. Meas. (2022).

[54] Yuanyuan Li, Ziyu Wang, Li Yin, Zhiqin Zhu, Guanqiu Qi, Yu Liu, X-Net: a dual encoding–decoding method in medical image segmentation, Vis. Comput. (2021) 1–11.

[55] Shuhao Fu, Yongyi Lu, Yan Wang, Yuyin Zhou, Wei Shen, Elliot Fishman, Alan Yuille, Domain adaptive relational reasoning for 3d multi-organ segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 656–666.

[56] Rudra P.K. Poudel, Ujwal Bonde, Stephan Liwicki, Christopher Zach, Contextnet: Exploring context and detail for semantic segmentation in real-time, 2018, arXiv preprint arXiv:1805.04554.

[57] Gen Li, Inyoung Yun, Jonghyun Kim, Joongkyu Kim, Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation, 2019, arXiv preprint arXiv:1907.11357.

[58] Shao-Yuan Lo, Hsueh-Ming Hang, Sheng-Wei Chan, Jing-Jhih Lin, Efficient dense modules of asymmetric convolution for real-time semantic segmentation, in: Proceedings of the ACM Multimedia Asia, 2019, pp. 1–6.

[59] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, Eugenio Culurciello, Enet: A deep neural network architecture for real-time semantic segmentation, 2016, arXiv preprint arXiv:1606.02147.

[60] Xuetao Zhang, Zhenxue Chen, QM Jonathan Wu, Lei Cai, Dan Lu, Xianming Li, Fast semantic segmentation for scene perception, IEEE Trans. Ind. Inform. 15 (2) (2018) 1183–1192.

[61] Yutong Xie, Jianpeng Zhang, Chunhua Shen, Yong Xia, Cotr: Efficiently bridging cnn and transformer for 3d 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 III 24, Springer, 2021, pp. 171–180.

[62] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, Manning Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III, Springer, 2023, pp. 205–218.

[63] Guoping Xu, Xingrong Wu, Xuan Zhang, Xinwei He, Levit-unet: Make faster encoders with transformer for medical image segmentation, 2021, arXiv preprint arXiv:2107.08623.

[64] Yijiang Li, Wentian Cai, Ying Gao, Chengming Li, Xiping Hu, More than encoder: Introducing transformer decoder to upsample, in: 2022 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2022, pp. 1597–1602.

[65] Chenyu You, Weicheng Dai, Yifei Min, Lawrence Staib, James S. Duncan, Implicit anatomical rendering for medical image segmentation with stochastic experts, 2023, arXiv preprint arXiv:2304.03209.

[66] Fabian Isensee, Paul F. Jäger, Simon A.A. Kohl, Jens Petersen, Klaus H. Maier-Hein, Automated design of deep learning methods for biomedical image segmentation, 2019, arXiv preprint arXiv:1904.08128.

[67] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, Yizhou Yu, Nnformer: Interleaved transformer for volumetric segmentation, 2021, arXiv preprint arXiv:2109.03201.

[68] Qianying Liu, Chaitanya Kaul, Jun Wang, Christos Anagnostopoulos, Roderick Murray-Smith, Fani Deligianni, Optimizing vision transformers for medical image segmentation, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.