algorithms

MDPI

Article

CCFNet: Collaborative Cross-Fusion Network for Medical Image Segmentation

Jialu Chen [1] and Baohua Yuan [1,2,*]

[1] The Aliyun School of Big Data, Changzhou University, Changzhou 213164, China; s21150812030@smail.cczu.edu.cn
[2] Jiangsu Engineering Research Center of Digital Twinning Technology for Key Equipment in Petrochemical Process, Changzhou University, Changzhou 213164, China
* Correspondence: yuanbaohua@cczu.edu.cn

Abstract: The Transformer architecture has gained widespread acceptance in image segmentation. However, it sacrifices local feature details and necessitates extensive data for training, posing challenges to its integration into computer-aided medical image segmentation. To address the above challenges, we introduce CCFNet, a collaborative cross-fusion network, which continuously fuses a CNN and Transformer interactively to exploit context dependencies. In particular, when integrating CNN features into Transformer, the correlations between local and global tokens are adaptively fused through collaborative self-attention fusion to minimize the semantic disparity between these two types of features. When integrating Transformer features into the CNN, it uses the spatial feature injector to reduce the spatial information gap between features due to the asymmetry of the extracted features. In addition, CCFNet implements the parallel operation of Transformer and the CNN and independently encodes hierarchical global and local representations when effectively aggregating different features, which can preserve global representations and local features. The experimental findings from two public medical image segmentation datasets reveal that our approach exhibits competitive performance in comparison to current state-of-the-art methods.

Keywords: Vision Transformer;CNN; medical image segmentation; collaborative cross-fusion

Citation: Chen, J.; Yuan, B. CCFNet: Collaborative Cross-Fusion Network for Medical Image Segmentation. *Algorithms* **2024**, *17*, 168. https://doi.org/10.3390/a17040168

Academic Editor: Frank Werner

Received: 21 March 2024
Revised: 13 April 2024
Accepted: 20 April 2024
Published: 21 April 2024

1. Introduction

Accurate image segmentation can more clearly identify changes in anatomic or pathologic structure in medical images [1], which is crucial in various computer-aided diagnostic applications, including lesion contour, surgical planning, and three-dimensional reconstruction. Medical image segmentation can detect and locate the boundaries of lesions in an image, thus helping to quickly identify the potential existence of tumors and cancerous regions, which may help clinicians save diagnosis time and improve the possibility of finding tumors [2]. Traditionally, the symmetric encoder–decoder structures has been the standard in medical image segmentation, where U-Net [3] has become the benchmark of choice among different variants with great success.

The U-Net model consists of convolutions, the fundamental operation of which is the convolution operator with two characteristics, weight sharing and local connection, which ensures the affine invariance of the model. Although these characteristics help to create effective and versatile medical imaging systems, they still require additional improvements to aid clinicians in early disease diagnosis [4]. Researchers have proposed various improved methods to add global context to convolutional neural networks (CNNs), among which influential methods include introducing an attention mechanism [5–8] and expanding convolution kernels [9–11] to extend their receptive fields. However, the locality of the convolutional layer's receptive fields leads to the inability of networks to utilize remote semantic dependence effectively, so their learning ability is limited to a relatively small area and they fail to fully explore object-level information, especially for organs in terms

of texture, shape, and size, typically yielding weaker properties and exhibiting sizable inter-patient variability.

Transformer [12] has shown high performance in language learning tasks, and the attention-based model has emerged as an appealing solution because of its ability to efficiently handle very long sequence dependencies, adapting to various vision tasks. Recent research has demonstrated the ability of Transformer modules to entirely take over the role of conventional convolutions by manipulating sequences of picture patches; the most representative of these is Vision Transformer (ViT) [13]. There have been many works proving that the ViT model can promote the development of many computer vision tasks, including semantic segmentation [14], object detection [15], and image classification [13], among others. The accomplishments of ViT in processing natural images have captivated the medical field. In response, researchers are delving into the capabilities of Transformer for medical image segmentation to address the intrinsic receptive field limitations of CNNs and make them suitable for medical imaging applications [16–19].

However, the performance of Transformer-based models largely depends on pre-training [20,21]. There are two problems with Transformer-based models' pre-training process. First, there is a scarcity of comprehensive and widely accepted large datasets for pre-training in the medical imaging field due to the extensive time professionals need to dedicate to annotating medical images (in contrast, ImageNet [22], a large dataset, is available for pre-training natural scene images). Secondly, pre-training consumes much time and many computing resources. Moreover, using large natural image datasets for pre-training medical image segmentation models is difficult because of the domain difference between medical and natural pictures. In addition, there are also some open challenges in different types of medical images. For example, when Swin UNETR [23] is pre-trained on a CT dataset and subsequently applied to different medical imaging modalities like MRI, its performance tends to decline. This is attributed to the significant differences in regional characteristics between CT and MRI images [24].

Fully exploiting CNNs' and Transformer's respective advantages to effectively integrate fine-grained and coarse-grained information in images, thereby boosting the precision and performance in deep learning models, has become a research direction researchers are actively working on. In Figure 1, we summarize various medical image segmentation methods that utilize a combined CNN and Transformer hybrid architecture. As shown in Figure 1a, researchers have incorporated Transformer into models with CNN as the backbone in different ways, either by adding them or replacing certain architectural components to create a network that combines Transformer and a CNN in a serial or embedded fashion. However, this strategy only uses stacking to fuse fine-grained and coarse-grained features, which may reduce the fusion effect and not fully leverage the synergistic capabilities of both network types. Figure 1b,c illustrate parallel frameworks of a CNN and Transformer, extracting distinct feature information from both structures and merging them multiple times before passing them to the decoder for decoding. In Figure 1b, additional branches are used to fuse the CNN and Transformer branches, but this introduces network overhead and lacks effective interaction. For example, TransFuse [25] uses a branch composed of BiFusion modules to simultaneously utilize the different characteristics of the CNN branch and the Transformer branch during the feature extraction process, alleviating the problem of ignoring intermediate features due to serial connections. However, its upsampling method struggles to effectively restore the mid-layer information, leading to a loss of detail.

In Figure 1c, each layer extracts features using the CNN and Transformer, respectively, and finally, adds and fuses them as the output. However, due to semantic differences between Transformer's and the CNN's features, this strategy limits the fusion's effectiveness. In Figure 1c, the input features are independently extracted by both the Transformer and the CNN, then added and merged as the input of the CNN or Transformer. However, due to the feature semantic differences between Transformer and CNNs, this strategy limits the fusion effect. For example, both HiFormer [26] and CiT-Net [27] use CNN and Transformer branches to extract features, respectively, but HiFormer only implements one-way feature

fusion from the CNN to Transformer, while CiT-Net only fuses the two features through a single fusion method and sends them to different branches. Both ignore the different characteristics of the two branches.

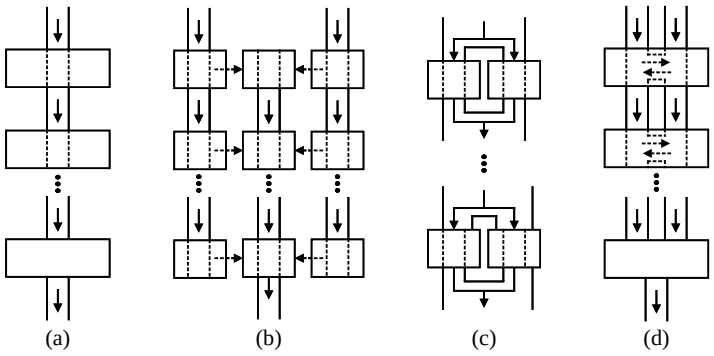


Figure 1. Comparison of different CNN and Transformer integration strategies. (**a**) Serial or embedded fusion strategy; (**b**) CNN and Transformer branch fusion strategy; (**c**) parallel fusion strategy of CNN and Transformer at each layer; (**d**) CCFNet fusion strategy. The arrows in the figure indicate the direction of data flow in the feature map.

In our study, we present the collaborative cross-fusion network (CCFNet), designed for medical image segmentation. CCFNet's encoder consists of two parts: first, in the shallow layers of the encoder, a CNN is employed to acquire convolutional depth features, compensating for detail lost in upsampling; second, in the deep encoder layers, a collaborative cross-fusion of Transformer and the CNN is applied, which can simultaneously enhance the image's local and global depictions. As shown in Figure 1d, compared with other combination strategies, CCFNet achieves close information interaction between the CNN and Transformer while continuously modeling local and global representations. By continuously aggregating the hierarchical representation and information interaction of global and local features, the information interaction of collaborative cross-fusion is closer and the feature fusion is more thorough. This makes CCFNet excellent in medical image segmentation tasks.

In CCFNet, high-resolution features extract fine-grained local information and perform depthwise convolutions. Since low-resolution features contain more global information (location and semantic information), feature prediction can fuse long-distance global information, and the self-attention mechanism facilitates the capture of deep information [28]. CCFNet processes low-resolution features through a parallel fusion of the CNN and Transformer within the collaborative cross-fusion module (CCFM). This method capitalizes on the self-attention mechanism's robust long-range dependency capabilities to ensure accurate medical image segmentation. Considering the complementarity of the two network features, the CCFM sequentially delivers global context from the Transformer branch to the feature maps through the spatial feature injector (SFI) block. This integration significantly boosts the global perceptual capabilities of the CNN branch. Likewise, the collaborative self-attention fusion (CSF) block progressively reintroduces the local features from the CNN branch back into the Transformer, enriching the local detail and creating a dynamic interplay of fused features. Finally, local–global feature complementarity can be achieved, and the network's feature-encoding capabilities can be enhanced. The experiments utilize Synapse, an open accessible dataset for multi-organ medical image segmentation. When the average Dice similarity coefficient (DSC) scores are compared with those from other hybrid models, our proposed CCFNet shows improved accuracy in organ segmentation. This paper's important contributions are summarized below:

- We propose CCFNet, a collaborative cross-fusion network, to integrate Transformer-based global representations with convolutional local features in a parallel interactive manner. Compared with other fusion methods, the collaborative cross-fusion module can not only encode the hierarchical local and global representations independently

but also aggregate the global and local representations efficiently, maximizing the capabilities of the CNN and Transformer.

- In the collaborative cross-fusion module, the CSF block is designed to adaptively fuse the correlation between the local tokens and the global tokens and reorganize the two features to introduce the convolution-specific inductive bias into the Transformer. The spatial feature injector block is designed to reduce the spatial information gap between local and global features, avoiding the asymmetry of extracted features and introducing the global information of the Transformer into the CNN.
- On two publicly accessible medical image segmentation datasets, CCFNet outperforms other competitive segmentation models, validating its effectiveness and superiority.

2. Related Work

2.1. CNN

The first widely known CNN model was U-Net [3], named because of its U-shaped network structure, which consisted of an encoder–decoder symmetrical network composed of convolution, downsampling, upsampling, and stitching operations. U-Net employed an encoder to capture global contextual information by downsampling feature representations. Conversely, the decoder restored these features to the original input resolution for semantic prediction through upsampling. Additionally, skip connections linked outputs from the encoder with corresponding decoder layers at the same resolution, helping to recover spatial details lost in downsampling and enhancing detail preservation. V-Net [29] was developed specifically for 3D image segmentation, focusing on direct, end-to-end training to facilitate volume segmentation in MRI scans of the prostate.

Most U-Net variants were based on the residual block [30], dense block [31], and attention mechanism [5–7] to improve the network's segmentation capabilities. The concepts behind dense and residual blocks can increase the rate of feature reuse and alleviate the problem of vanishing gradients. Dense-UNet [32] and Res-UNet [33], inspired by dense and residual connections, respectively, adapted the U-Net structure by replacing its sub-modules with versions that incorporate these connections, enhancing the network's architecture. UNet++ [34] added the dense block and convolutional layers between its decoder and encoder, aimed at narrowing the semantic differences in its processed features to improve segmentation accuracy. MultiResUNet [35] innovatively combined the MultiRes module and U-Net, in which the MultiRes module extends the concept of residual connections. In this configuration, feature maps produced by three 3×3 convolutions were spliced together as a combined feature map and then added to the outcome of a 1×1 convolution performed on the input feature map.

An attention mechanism in the network enables it to concentrate on vital sections while suppressing irrelevant input to improve processing efficiency. AttnUNet [8] introduced an attention mechanism into U-Net, adjusting encoder output features before they are concatenated with matching decoder features at each resolution level. This attention setup generates a gating signal that modulates feature importance at different spatial locations. FocusNet [36] employed the dual encoder–decoder architecture, leveraging attention gating to transfer relevant features from the decoder of one U-Net to the encoder of another. This mechanism enhances the propagation of features throughout the network.

On the improvement of a non-network structure, the authors of nnUNet [37] asserted that further enhancements could be achieved by gaining deeper insights into the data and training techniques tailored to medical data and implementing suitable preprocessing, and proposed a robust U-Net-based adaptive framework (including 2D and 3D frameworks) that automatically adjusted itself to preprocess data and selected the best network architecture for the task without human intervention.

2.2. Transformer

The groundbreaking Transformer architecture [12] sparked a paradigm shift in natural language processing, swiftly establishing itself as a commonly used foundational model for

visual comprehension tasks. Transformer has great visuals because of its ability to create distinct visual environments, but it also has an inherent drawback of not being able to exploit spatial environments in images as a CNN does. Recent works have evolved toward possible solutions to overcome this drawback, and most extant studies in medical image segmentation employ CNN–Transformer hybrid models for feature processing.

TransUNet [16] introduced the integration of the Transformer architecture into medical image segmentation frameworks. It introduced a novel approach by framing the segmentation task as a sequence-to-sequence prediction, incorporating self-attention mechanisms. This model adopts a hybrid CNN–Transformer design, strategically combining the spatial details extracted by CNN features with the Transformer's capabilities. Additionally, within the realm of medical image segmentation, several other hybrid models, built upon the U-Net architecture, have also seen enhancements. For example, UCTransNet [17] adopted the Channel Transformer (CTrans) module, leveraging the channel attention mechanism as a modification to the traditional U-Net skip connections. This implementation includes one sub-module designed for channel-wise cross-attention via Transformer technology and another for multi-scale channel fusion. These enhancements facilitate the efficient integration of multi-scale channel data with decoder features to enhance clarity and resolution. Like TransUNet, UNETR [18] employed a Transformer along with a convolutional decoder in its encoder architecture to build segmentation maps. MT-UNet [19] developed a U-shaped network, utilizing the mixed Transformer module (MTM) designed for both intra-sample and inter-sample learning, aimed at enhancing the precision of medical image segmentation. First, the MTM applied the LGG-SA module to assess self-similarity. It then explored inter-sample connections using an external attention mechanism. MSFusion-UNet [38] adopted a flexible and efficient multi-stream fusion encoder to facilitate multi-scale fusion of multiple imaging stream features through spatial attention. LeVit-UNet [39] used LeViT blocks to build a U-Net variant of the encoder, which enables it to learn long-range dependencies more efficiently. DCA module [40] used double cross-attention on the U-Net framework and enhanced skip connections to solve the semantic gap between encoder and decoder features. TransAttUnet [41] adopted an effective feature screening method by jointly designing multi-scale skip connections and multi-level guided attention. It learned non-local interactions between encoding features and passed key features to the decoder. From a macro perspective, these architectures integrated a CNN and Transformer in a serial combination manner.

In terms of parallel strategies, TransFuse [25] implemented the BiFusion module, which performed spatial attention on the CNN branch and channel attention on the Transformer branch. Following this, operations such as convolution, multiplication, concatenation, and residuals were executed to facilitate the merging of features from both branches. On this basis, FAFuse [42] improved the BiFusion module through a four-axis fusion module to improve the ability of representation learning. The HiFormer [26] algorithm uses CNN and Transformer branches to extract features. To provide positioning information and reuse features, it incorporates a skip connection that conveys local CNN features to the Transformer. CiT-Net [27] employs dynamically deformable convolutions within its CNN branch to enhance feature extraction capabilities, while it incorporates the compact convolution projection and the SW-ACAM module in the Transformer branch to more effectively capture long-term dependencies across dimensions. CSwin-PNet [43] connected a CNN and Swin Transformer [44] as the backbone for feature extraction, building a pyramidal network structure for feature encoding and decoding. CT-Net [45] utilized an asymmetric asynchronous branch parallel structure to efficiently extract local and global representations while reducing unnecessary computational costs. DPCTN [46] combined the dual-branch fusion of a CNN and Transformer. To reduce the information loss during the information pooling process, DPCTN specially adopted a three-branch transposed self-attention module to significantly improve the segmentation performance.

Several other studies also exist, ranging from developing hybrid CNN and Transformer models to improving the Transformer blocks themselves to handle the complexities

of medical imagery. Impacted by the general adoption of the Swin Transformer, Swin-Unet [28] proposed an innovative approach by substituting the convolutional blocks with Swin Transformer blocks for 2D medical image segmentation, and this marked the introduction of the first entirely Transformer-based U-shaped architecture. DS-TransUNet [47] expanded upon Swin-Unet by incorporating an encoder designed to handle multi-scale inputs and integrating a new multi-scale feature fusion module. This module uses the self-attention mechanism to effectively link global dependencies among features from various scales, enhancing the quality of segmentation across diverse medical images. MedT [48] improved upon the current system by incorporating a control mechanism within the self-attention module, specifically to tackle the scarcity of medical image segmentation datasets. Additionally, it introduced a local–global training strategy, optimizing the model's training process on medical images, which can further improve the performance.

3. Method

CCFNet follows a U-shaped structure featuring hierarchical decoder and encoder sections, in which skip connections facilitate the linkage between the decoder and encoder. It is essential to note that CCFNet is structured with two branches, which process information differently, as shown in Figure 2. The two branches preserve global contexts and local features through the parallel fusion layer composed of the CCFM. In this CCFM, the CSF block can adaptively fuse them according to the correlation between local and global tokens, thus introducing convolution-specific inductive bias into the Transformer. The SFI block can avoid asymmetry of extracted features and introduce global representations of Transformer into the CNN branch, which has extracted local semantic features through the detail feature extractor (DFE) block. Features from both parallel branches are successively fused to form features that are fused with each other, and finally, realize the complementarity of the two features. The proposed parallel branching approach has three main benefits: Firstly, the CNN branch gradually extracts low-level, high-resolution features to obtain detailed spatial information, which can help Transformer obtain rich features and accelerate its convergence. Second, the Transformer branch can capture global information while remaining sensitive to low-level contexts without building a deep network. Finally, during feature extraction, the proposed CCFM can leverage the different characteristics of Transformer and the CNN to the full extent, continuously aggregating hierarchical representations from global and local features.

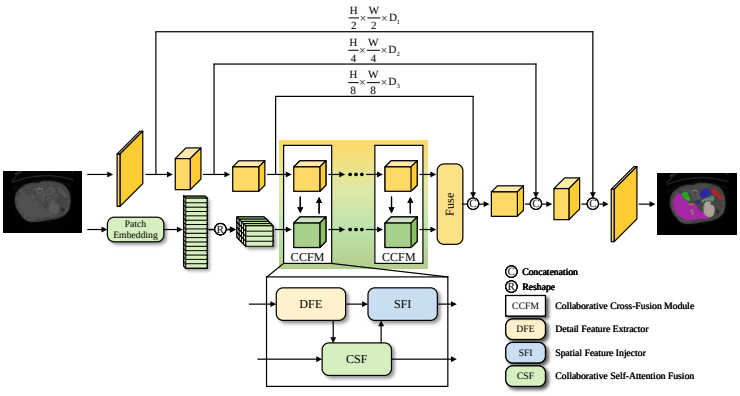


Figure 2. The overall architecture of the CCFNet model for medical image segmentation. The network follows a standard U-shaped structure and consists of four parts: CNN branch, Transformer branch, parallel fusion layer, and decoder. Among them, the CNN branch uses convolution to extract fine-grained features, the Transformer branch uses attention to capture global information, and the parallel fusion layer combines the CNN and Transformer in parallel to extract rich depth information.

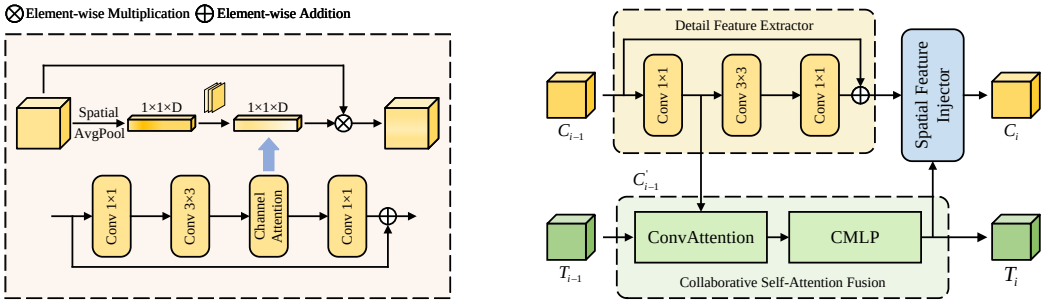
3.1. CNN Branch

The CNN branch adopts a feature pyramid structure. This is because while in the Transformer branch patch embedding is used to project image patches into vectors, which

results in the loss of local details, in the CNN the convolution kernels slide across overlapping feature maps, which provides the possibility of preserving fine local features. As a result, the CNN branch is able to supply local feature details to the Transformer branch. Specifically, as the network depth increases in the CNN branch, the resolution of feature maps gradually decreases, the number of channels gradually increases, the receptive field gradually increases, and the feature encoding changes from local to global. Given an input image $x \in R^{H\times W\times D_0}$, its spatial resolution is $H \times W$ and D_0 is the number of channels, the feature map generated by $F^{CNN}(\cdot)$ is represented as

$$\{f_l\}_{l=1}^{L} = F_l^{CNN}(x;\Theta) \in R^{\frac{H}{2^l}\times\frac{W}{2^l}\times D_l}, \tag{1}$$

where D represents the dimension of the feature map, Θ represents the parameters of the CNN branch, and L represents the quantity of feature layers. Specifically, the first block f_1 is made up of 2 convolutions (3×3) with strides 1 and 2, and each convolution block is followed by normalization and the GELU activation function to extract initial local features (such as edge and texture information). As shown in Figure 3a, f_2 and f_3 are stacked with SEConv blocks composed of three convolutional blocks and an SE module [6]. The number of SEConv blocks in f_2 and f_3 is 2 and 6, respectively. The efficient and lightweight SE module can be seamlessly integrated into the CNN architecture, which can help the CCFNet network to enhance local details, suppress irrelevant regions, correct channel features by modeling the relationship between channels, and improve the representational capacity of the neural network.



(a) SEConv (b) Collaborative Cross-Fusion Module

Figure 3. The architecture of the proposed CCFNet. (**a**) Shows the convolution process of SEConv; (**b**) shows the structure of the CCFM, there are two branches in the module whose inputs are C_{i-1} and T_{i-1}. The Transformer branch is composed of collaborative self-attention fusion blocks, and the CNN branch is composed of detail feature extractor blocks and spatial feature injector blocks.

The CNN branch of the parallel fusion layer consists of a six-layer stack of modules consisting of a DFE block and an SFI block. The feature map output C_i of each layer has the same resolution size ($\frac{H}{16}, \frac{W}{16}, D_4$), and the output of the i-th layer is expressed as

$$C_i = \mathrm{SFI}(\mathrm{DFE}(C_{i-1}), T_i), \tag{2}$$

where T_i is the i-th layer's CSF-block-coded image representation on the Transformer branch with the same resolution as C_{i-1}. The structures of the DFE block and SFI block are shown in Figure 3b. More detailed operations are described in the parallel fusion layer.

3.2. Transformer Branch

The CNN branch obtains rich local features under a limited receptive field through convolution operations, while the Transformer branch performs global self-attention through attention mechanisms. The Transformer branch has the same input image $x \in R^{H\times W\times D_0}$ as the CNN branch. Following [13,16], in the patch embedding we first divide x into an $N = \frac{H}{P} \times \frac{W}{P}$ sequence of patches, the size of each patch is $P \times P$, and the default setting is 16. After splitting the input images into small patches, the patches are flattened to a sequence

of 2D patches $\{x_p^i \in R^{p^2 \cdot D_0} | i = 1, \ldots, N\}$ and fed to a trainable linear layer, which converts the vectorized patches x_p into a sequence embedding space with an output dimension of D_4, and then, in order to facilitate the fusion with the CNN branch, the reshape operation is used to generate image $T_0 \in R^{\frac{H}{16} \times \frac{W}{16} \times D_4}$, which can be expressed as

$$T_0 = \text{Reshape}([x_p^1 E; x_p^2 E; \ldots; x_p^N E]), \tag{3}$$

where $E \in R^{(p^2 \cdot D_0) \times D_4}$ is the patch embedding projection. The Transformer branch in the parallel fusion layer is connected to six CSF blocks of attention operations, and the CSF block consists of a ConvAttention and a CMLP (convolution multi-layer perceptron). The feature map output T_i of each layer has the same resolution size ($\frac{H}{16}$, $\frac{W}{16}$, D_4). Therefore, the output of the i-th layer can be written as follows:

$$T_i' = \text{ConvAttention}(\text{Normal}(C_{i-1}'), \text{Normal}(T_{i-1})) + T_{i-1}, \tag{4}$$

$$T_i = \text{CMLP}(\text{Normal}(T_i')) + T_i', \tag{5}$$

where C_{i-1}' and T_{i-1} are the two inputs of the CSF block, C_{i-1}' is the intermediate output of the i-th layer DFE module on the CNN branch, which has the same resolution as T_{i-1}, and T_i is the encoded image representation. The structure of a CSF block is illustrated in Figure 3b. More detailed operations are described in the parallel fusion layer.

3.3. Parallel Fusion Layer

The parallel fusion layer has two branches, namely, the Transformer branch and the CNN branch, which process information in distinct ways. In the CNN branch, local features are collected hierarchically through a convolution operation, and local clues are also saved as feature maps. The parallel fusion layer fuses the feature representation of the CNN in a parallel manner through cascaded attention modules, which maximizes the preservation of local features and global representations. The parallel fusion layer is composed of six CCFMs superposed.

An image has two completely different representations: global features and local features. The former focuses on model object-level relationships between remote parts, while the latter aims at fine-grained details and is beneficial for pixel-level localization and tiny object detection. As shown in Figure 3b, a CCFM is used to efficiently combine these encoded features of the Transformer and CNN, which can interactively fuse convolution-based local features and Transformer-based global representations.

The CCFM has two inputs, C_{i-1} and T_{i-1}, where C_{i-1} is the input on the CNN branch with the same resolution as T_{i-1}, T_{i-1} is the input on the Transformer branch, and $C_{i-1}^{'}$ is the feature map formed after extracting features on the CNN branch with the same resolution and number of channels as T_{i-1}, which can be expressed as

$$C_{i-1}^{'} = \text{GELU}(\text{Normal}(\text{Conv2d}(C_{i-1}))). \tag{6}$$

The Transformer aggregates information between global tokens, but CNN only aggregates information in the limited local field of view of the convolution kernel, which leads to certain feature semantic differences between the Transformer and CNN. Therefore, by superimposing the feature maps of the CNN and Transformer, the CSF block adaptively fuses the self-attention weights with common information between them to calculate the mutual relationship between local tokens and global tokens.

As shown in Figure 3b, the CSF block consists of ConvAttention and CMLP. Like the traditional attention mechanism, the basic module of ConvAttention is multi-head self-attention (MHSA). As shown in Figure 4a, the difference is that ConvAttention has two inputs, adding T_{i-1} and $C_{i-1}^{'}$ to obtain feature maps F_i and T_{i-1} as its input. In addition, ConvAttention uses convolutional mapping. The specific operation is that T_{i-1} generates V_i through 3×3 convolutional mapping, and F_i generates Q_i and K_i through 3×3 convo-

lutional mapping. Subsequently, we use the flatten operation to project the patches into the d-dimensional embedding space as the input of the underlying module MHSA in the ConvAttention block.

$$F_i = C'_{i-1} + T_{i-1}, \tag{7}$$

$$Q_i/K_i = \text{Flatten}(\text{GELU}(\text{Normal}(\text{Conv2d}(F_i)))), \tag{8}$$

$$V_i = \text{Falatten}(\text{GELU}(\text{Normal}(\text{Conv2d}(T_{i-1})))). \tag{9}$$

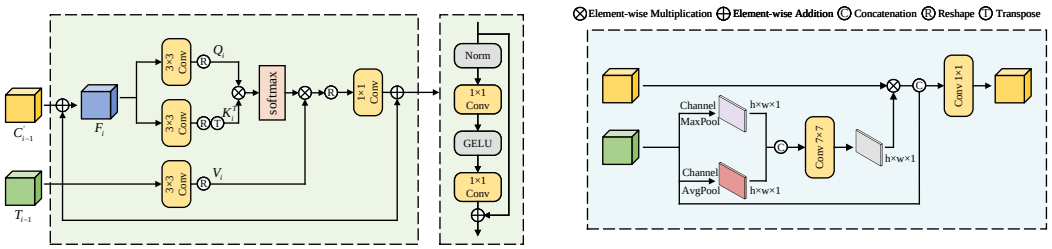
The MHSA is performed on the obtained Q_i, K_i, and V_i, an MHSA comprises h parallel self-attention heads. The calculation process is as follows:

$$\text{MHSA}(Q_i, K_i, V_i) = \text{Concat}(head_1, ..., head_h)W_{mhsa}, \tag{10}$$

where $W_{mhsa} \in R^{d \times d}$ represents the multi-headed trainable parameter weights. The self-attention of each head in MHSA is calculated as

$$\text{Attention}(Q, K, V) = \text{SoftMax}(\frac{QK^T}{\sqrt{d}} + B)V, \tag{11}$$

where Q, K, and $V \in R^{N_T \times d}$ are the query, key, and value matrices, which are obtained by convolution projection, $N_T = W_h \times W_w$ denotes the number of patch tokens. $\{W_h, W_w\}$ stand for the size of the feature F_i/T_i, and d is the query/key dimension. We follow [28,44] by including a relative position bias $B \in R^{N_T \times N_T}$. Since the relative position along each axis lies in the range $[-W_{h/w} + 1, W_{h/w} - 1]$, we parameterize a smaller deviation matrix $\hat{B} \in R^{(2W_h-1)\times(2W_w-1)}$; the value of B is taken from $\hat{B}$.



(a) Collaborative Self-Attention Fusion (b) Spatial Feature Injector

Figure 4. The architecture of the CCFM. (**a**) Structure of CSF block, which is composed of ConvAttention and CMLP; (**b**) the SFI block.

As shown in Figure 4a, a CMLP is then carried out, which consists of two convolution layers (1×1). The output T_{i+1} obtained after the CMLP is used as the input of the Transformer branch in the next fusion module, and at the same time, it is feature-fused with the feature map of the same resolution on the CNN branch.

Given the varying receptive fields of the CNN and Transformer, the features they extract exhibit asymmetry. At the same time, the information reflected by these features has a great gap in space. As shown in Figure 4b, when the Transformer branch is fused to the CNN branch, the SFI block uses the spatial attention weight map of the feature obtained on the Transformer branch. The calculation formula is as follows:

$$\begin{aligned} M(T) &= \sigma(\text{Conv2d}([\text{AvgPool}(T); \text{MaxPool}(T)])) \\ &= \sigma(\text{Conv2d}([T_{avg}; T_{max}])), \end{aligned} \tag{12}$$

where σ represents the sigmoid function, and T_{avg} and T_{max} represent the cross-channel average-pooled features and max-pooled features, respectively. The attention map is multiplied by the feature map on the CNN branch to achieve spatial information feature enhancement. Then, it is concatenated with the feature map on the Transformer branch, and the features are further fused by 1×1 convolution. The final output is used as the input of the CNN branch in the next fusion module. In the last layer of the parallel fusion

layer, the two features are finally used as the input of the decoder through fuse operation. Specifically, the outputs of the CNN branch and Transformer branch are added together and fused through a convolution.

3.4. Decoder

The decoder in CCFNet is a pure convolution module that consists of numerous upsampling steps to decode hidden features, with the ultimate output being the segmentation result. Firstly, bilinear interpolation is applied to the input feature map. The following operations are then repeated until the resolution of the original input is restored by concatenating the feature maps with the resolution improved by a factor of 2 with the feature maps on the corresponding jump joins, inputting them into successive convolution layers (3×3), and upsampling the output using bilinear interpolation. Finally, the feature maps with the restored original resolution are fed into a special convolution layer (segmentation head) to generate the pixel-level semantic prediction.

The encoder and decoder merge the semantic information of the encoder through skip connections and concatenation operations to obtain more contextual information. The outputs of the three layers of the CNN branch in the encoder are sequentially connected to the three layers of the decoder to regain local spatial information to improve finer details. The parallel fusion layer is a dual-stream fusion operation of the CNN and Transformer, which sends the fused feature output of the two features to the decoding layer.

3.5. Loss Function

In general segmentation tasks, Dice loss [29] and cross-entropy loss are both frequently used, with Dice loss being suitable for large-sized target objects and cross-entropy loss performing well for a uniform distribution of categories. Following the TransUNet [16] literature, the loss function used in CCFNet training also uses the combined form of Dice loss and binary cross-entropy, which is defined as

$$\begin{aligned} \mathcal{L}(G, Y) = 1 - \frac{2}{J} \sum_{j=1}^{J} \frac{\sum_{i=1}^{I} G_{i,j} Y_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} Y_{i,j}^2} \\ - \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} G_{i,j} log Y_{i,j} \end{aligned}, \tag{13}$$

where I and J are the number of voxels and classes, respectively; $Y_{i,j}$ and $G_{i,j}$, respectively, represent the predicted value and true value of class j at pixel i.

4. Experiments

4.1. Dataset

The effectiveness of the CCFNet model is demonstrated by experiments on two different public medical image datasets, such as the Synapse dataset [49] and the Automated Cardiac Diagnosis Challenge dataset [50] (ACDC).

The Synapse dataset, a public multi-organ segmentation dataset, contains 3779 axially enhanced abdominal clinical CT images from 30 abdominal CT scans. Following the partitioning method of the TransUNet [16] dataset, the dataset is divided into 18 cases for training and 12 cases for validation. The annotations for each image include eight abdominal organs (stomach, spleen, pancreas, liver, left kidney, right kidney, gallbladder, aorta); average HD (Hausdorff distance) and average DSC are used to evaluate CCFNet on this dataset.

ACDC is a publicly available dataset of cardiac magnetic resonance imaging (MRI) that contains 150 MRI 3D cases gathered from various individuals, with each instance covering the cardiac organ from the bottom to the top of the left ventricle. Following the setup in [16], only 100 well-annotated cases are used in the experiment, and labels for three key parts of the heart are chosen: the left ventricle (LV), right ventricle (RV), and myocardium (MYO).

Following the partitioning method of the dataset in TransUNet [16], the ACDC dataset is split at a ratio of 7:1:2, with training (1930 axial slices), validation, and test data, and the average DSC is used to evaluate the CCFNet approach on this dataset.

4.2. Evaluation Metrics

We use the Dice score and HD to evaluate the accuracy of segmentation in our experiments.

$$\text{Dice}(G, P) = \frac{2\sum_{i=1}^{I} G_i P_i}{\sum_{i=1}^{I} G_i + \sum_{i=1}^{I} P_i}, \tag{14}$$

$$\begin{aligned} \text{HD}(G^{'}, P^{'}) = max\{ & \max_{g^{'} \in G^{'}} \min_{p^{'} \in P^{'}} ||g^{'} - p^{'}||, \\ & \max_{p^{'} \in P^{'}} \min_{g^{'} \in G^{'}} ||p^{'} - g^{'}||\}. \end{aligned} \tag{15}$$

where P_i and G_i represent the predicted and actual values for voxel i, and $P^{'}$ and $G^{'}$ signify the sets of surface points for the prediction and ground truth, respectively.

4.3. Implementation Details

The CCFNet model is implemented based on PyTorch and trained on an Nvidia GeForce RTX 3090 GPU with 24 GB of memory. Unlike previous work (TransUNet [16], Swin-Unet [28]) in which models were initialized by pre-trained models on ImageNet [22], the CCFNet model is randomly initialized and trained from scratch, so the maximum number of training epochs is increased to 1000. The other variables are kept the same as the initial learning rate of 0.01, using a multi-learning rate strategy, batch size of 24, and using the SGD optimizer with momentum of 0.9 and weight decay of 1e-4. For all experiments, we apply simple data augmentation, such as random rotations and flips. We slice all samples layer by layer to analyze 3D datasets. Finally, in order to reconstruct the 3D prediction for assessment, we stack all of the prediction's 2D slices together.

4.4. Results

We evaluate the performance of the CCFNet model on two different types of dataset (Synapse and ACDC) and compare it with various state-of-the-art models.

As shown in Table 1, experiments are performed on Synapse using the same image size and preprocessing, and CCFNet is compared with various of the main Transformer- or CNN-based methods such as TransUNet, LeVit-UNet-384, MT-UNet, UCTransNet, TransFuse, and Swin-Unet. Meanwhile, to visually demonstrate the performance of the CCFNet model, some qualitative results of the CCFNet model on Synapse are visually contrasted with a variety of other approaches, such as Swin-Unet, TransUNet, and U-Net. As shown in Figure 5, the red boxes indicate areas where CCFNet outperforms the other methods. Specifically, CCFNet can outperform Swin-Unet by more than 7.08 mm and 2.46% on average HD and DSC, respectively. Among them, CCFNet has the highest DSC in five organs: the stomach, liver, kidney (left), kidney (right), and gallbladder. For some specific organs that are hard to segment, CCFNet can better capture remote dependence. In the first row of Figure 5, CCFNet can better segment the pancreas with long and narrow shapes than other models. In identifying large organs, CCFNet has better accuracy in recognizing and delineating stomach contours, as shown in the second row. The CCFNet segmentation results are primarily compatible with the ground truth labels. When it comes to identifying small organs, CCFNet has certain advantages. As shown in the third row, individual models may not fully identify the gallbladder. CCFNet can identify organ junctions more accurately, as shown in the fourth row, at the junction of the liver and stomach, while the other three models make some errors, which shows that CCFNet is effective. The visualization intuitively demonstrates the high segmentation accuracy of CCFNet, especially on some difficult-to-segment slices. The excellent performance is attributed to the CCFM in CCFNet, which can consider the local small organs while

focusing on large organs, showing the strong representation ability of CCFNet in learning low-level specifics and high-level semantic features that are critical in the segmentation of medical images.

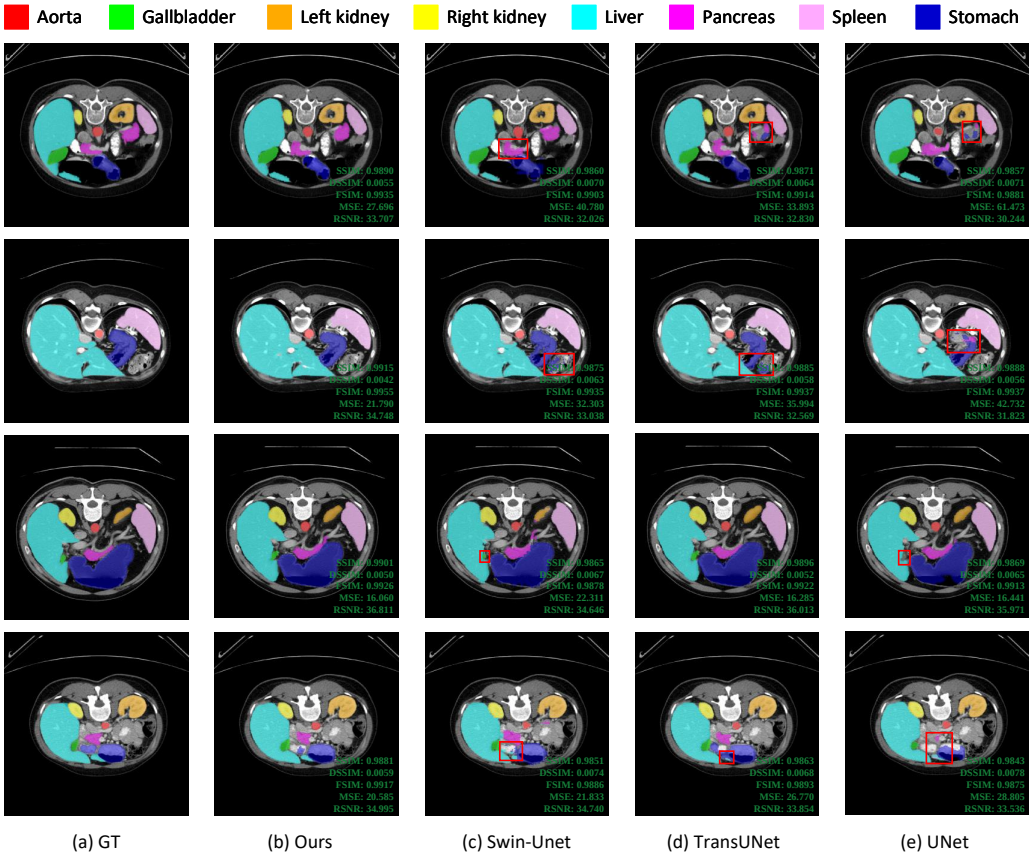


Figure 5. Qualitative visual comparison of CCFNet and other 2D methods on Synapse, and quantitative evaluation of images using SSIM, DSSIM, FSIM, MSE, and RSNR. (**a**) Ground truth; (**b**) CCFNet, our proposed method; (**c**) Swin-UNet, incorporating Swin Transformer blocks in both the encoder and decoder; (**d**) TransUNet, utilizing a ViT encoder on the ResNet-50 backbone and employing a UNet decoder; (**e**) UNet, featuring a U-shaped encoder-decoder architecture. The red box indicates areas where CCFNet outperforms the other methods.

Table 1. Comparison on Synapse (Dice score % for each organ, average Dice score %, and average Hausdorff distance in mm).

Method	Avg. DSC (%)	Avg. HD (mm)	Stomach	Spleen	Pancreas	Liver	Kidney (L)	Kidney (R)	Gallbladder	Aorta
V-Net [29]	68.81	-	56.98	80.56	40.05	87.84	77.10	80.75	51.87	75.34
DARR [51]	69.77	-	45.96	89.90	54.18	94.08	72.31	73.24	53.77	74.74
TransFuse [25]	77.42	-	73.69	87.03	57.06	94.22	80.57	78.58	63.06	85.15
U-Net [3]	76.85	39.70	75.58	86.67	53.98	93.43	77.77	68.60	69.72	89.07
R50 UNet [3]	74.68	36.87	74.16	85.87	56.90	93.74	80.60	78.19	63.66	87.74
R50 AttnUNet [8]	75.57	36.97	74.95	87.19	49.37	93.56	79.20	72.71	63.91	55.92
AttnUNet [8]	77.77	36.02	75.75	87.30	58.04	93.57	77.98	71.11	68.88	89.55
R50 ViT [13]	71.29	32.87	73.95	81.99	45.99	91.51	75.80	72.20	55.13	73.73
TransUNet [16]	77.48	31.69	75.62	85.08	55.86	94.08	81.87	77.02	63.13	87.23
LeVit-UNet-384 [39]	78.53	16.84	72.76	88.86	59.07	93.11	84.61	80.25	62.23	87.33
MT-UNet [19]	78.59	26.59	76.81	87.75	59.46	93.06	81.47	77.29	64.99	87.92
UCTransNet [17]	78.23	26.75	79.42	87.84	56.22	93.17	80.19	73.18	66.97	88.86
Swin-Unet [28]	79.13	21.55	76.60	90.66	56.58	94.29	83.28	79.61	66.53	85.47
Ours	81.59	14.47	80.47	88.19	56.89	95.37	87.42	83.50	72.49	88.35

Table 2 shows the experimental results of ACDC. CCFNet exceeds TransUNet by 1.35% and exceeds Swin-Unet by 1.07% in average DSC score. The excellent performance

is attributed to the SEConv and CCFM modules in CCFNet, which provide powerful representation capabilities in learning low-level specifics and high-level semantic features for CCFNet, which is crucial in medical image segmentation. This once again confirms that CCFNet using CNN–Transformer collaborative cross-fusion has a stronger ability to learn effective representations for medical image segmentation than other advanced methods, indicating that CCFNet has excellent generalization and robustness.

Table 2. Comparison on the ACDC dataset in DSC (%).

Method	Avg. DSC (%)	RV	Myo	LV
R50 UNet [3]	87.55	87.10	80.63	94.92
R50 AttnUNet [8]	86.75	87.58	79.20	93.47
R50 ViT [13]	87.57	86.07	81.88	94.75
UNETR [18]	88.61	85.29	86.52	94.02
TransUNet [16]	89.71	88.86	84.53	95.73
Swin-Unet [28]	90.00	88.55	85.62	95.83
MT-UNet [19]	90.43	86.64	89.04	95.62
UCTransNet [17]	89.69	87.92	85.43	95.71
LeViT-UNet-384 [39]	90.32	89.55	87.64	93.76
Ours	91.07	89.78	89.30	94.11

To comprehensively evaluate the performance of CCFNet, in Table 3 we compare the number of parameters and calculations between CCFNet and several mainstream network structures. As can be seen from the table, CCFNet has increased both the number of parameters and the amount of operations compared to Swin-Unet and TransUNet. This is because we make full use of the self-attention mechanism in CCFNet to obtain a more refined feature representation. However, it is important to emphasize that this modest increase in complexity resulted in significant improvements in segmentation accuracy. Its improvement in segmentation accuracy is attributed to the model's ability to effectively capture global information and multi-scale features, thereby more accurately segmenting structures in medical images.

Table 3. Comparison of number of parameters and FLOPs for 2D segmentation models in Synapse experiments.

Method	Params (M)	FLOPs (G)
U-Net [3]	31.13	55.84
Swin-Unet [28]	96.34	42.68
TransUNet [16]	105.32	38.52
MT-UNet [19]	79.07	44.72
UCTransNet [17]	65.6	63.2
Ours	137.36	76.18

4.5. Ablation Studies

We mainly perform ablation research using the Synapse database to assess the effectiveness of each basic component of CCFNet. All tests are executed with the same hyperparameters and are initiated from scratch to maintain fairness in comparison. Table 4 shows that incorporating the SEConv module into the baseline model results in a consistent gain in segmentation accuracy over the baseline. This is due to the critical role of this module in being able to compensate for the large amount of detail information lost during the upsampling process of the decoder. Adding the CCFM module also brings huge gains, because the CCFM module integrates global and local features to improve segmentation efficiency.

Table 4. Ablation studies on effects of different components of CCFNet.

Method	Avg. DSC (%)	Avg. HD (mm)
Baseline	78.91	24.08
Baseline + SEConv	80.20	19.33
Baseline + SEConv + CCFM	81.59	14.47

Table 5 displays the influences of evaluating various fusion techniques within CCFM modules by ablating different components in ablation studies on the Synapse dataset. Using only the DFE module is equivalent to removing the Transformer branch and only using the CNN branch. Using only the CSF module is equivalent to removing the CNN branch and only using the Transformer branch. The experimental results reveal that when removing different branches, the DSC scores on Synapse are almost the same. To some extent, this means that the global information represented by Transformer and the local information represented by the CNN play an equally important role in visual representation, indicating that both global features and local features are important for organ segmentation, and the fusion of the two can help the model to achieve more precise segmentation.

Table 5. Ablation studies of effects of different fusion methods on CCFM. The presence of a checkmark (✓) in the corresponding column indicates the utilization of the module, while the absence of such a mark signifies its non-utilization.

DFE	CSF	SFI	Avg. DSC (%)	Avg. HD (mm)
✓			80.20	19.33
	✓		79.93	18.44
	✓	✓	81.11	18.26
✓		✓	79.57	17.65
✓	✓		80.37	22.40
✓	✓	✓	81.59	14.47

When only removing the CSF block (using a splicing operation to fuse different features on the two branches), the DSC score on Synapse drops by 2.02%. This shows that the CSF block can reduce the feature semantic difference between two features, introduce convolution-specific inductive bias into the Transformer branch, and enrich the detailed features of the Transformer. When only the SFI block is removed (using splicing operations to fuse different features on the two branches), the DSC score on Synapse drops by 1.22%, and the performance on HD is far worse than the original model. This shows that the SFI block can reduce the spatial information gap between features caused by the asymmetry of extracted features and introduce the global information of the Transformer branch into the CNN branch. The results show that the operations of the CSF block and SFI block are conducive to the fusion between two different feature maps and can integrate the global encoding of the Transformer and the local encoding ability of CNN. This is because the cross-fusion between the CNN to Transformer and Transformer to the CNN can more thoroughly integrate the features between the two different methods than the general simple fusion.

4.6. 3D Implementation

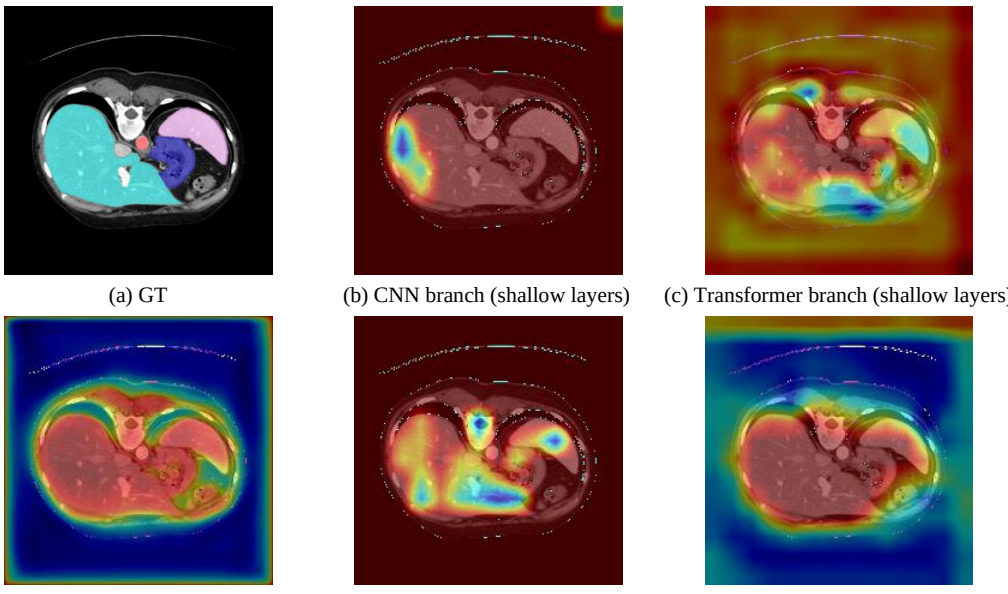
To further explore the performance of CCFNet, we implement a 3D implementation of CCFNet. The performance of 3D CCFNet is evaluated on the Synapse dataset and compared to a variety of state-of-the-art 3D models, and the final experimental results are shown in Table 6. Among them, nnFormer [52] is a Transformer model based on a cross structure that is mainly used to deal with the 3D image segmentation problem in medical image analysis. According to 3D implementation, the image input size is set to $64 \times 128 \times 128$ in the nnFormer. The 3D CCFNet shows a 0.31% improvement over nnFormer on average DSC, and 3D CCFNet achieves 8.78 mm in HD performance, meaning that 3D CCFNet can better delineate object boundaries.

Table 6. Comparison with 3D module on Synapse (Dice score % for each organ, average Dice score %, and average Hausdorff distance in mm).

Method	Avg. DSC (%)	Avg. HD (mm)	Stomach	Spleen	Pancreas	Liver	Kidney (L)	Kidney (R)	Gallbladder	Aorta
ViT [13]	67.86	36.11	70.44	81.75	42.00	91.32	74.70	67.40	45.10	70.19
R50 ViT [13]	71.29	32.87	73.95	81.99	45.99	91.51	75.80	72.20	55.13	73.73
UNETR [18]	79.56	22.97	73.99	87.81	59.25	94.46	85.66	84.80	60.56	89.99
nnFormer [52]	86.57	10.63	86.83	90.51	83.35	96.84	86.57	86.25	70.17	92.04
Ours	86.88	8.78	85.17	89.67	82.36	97.01	85.92	90.01	72.19	92.74

4.7. Analysis and Discussion

Feature Analysis. As shown in Figure 6, in order to verify our motivation and effects, the feature maps of Synapse are visualized using the Grad-CAM method [53]. Due to the limitations of convolution, the region of attention in Figure 6b tends to be locally informative and suffers from a lack of remote information capture. Because of the global features provided by the Transformer branch, Figure 6e learns to activate a larger region than the local region in Figure 6b, indicating that Figure 6e enhances the long-range feature dependence compared to Figure 6b. As the Transformer focuses on extracting global feature dependencies, the region of attention in Figure 6c is insufficient in local feature details. Because of the progressively finer local features captured by the CNN branch, critical local features are retained in Figure 6f. At the same time, the background is significantly suppressed, indicating that the learned feature representation in Figure 6f has a higher local perception ability than in Figure 6c.



(a) GT (b) CNN branch (shallow layers) (c) Transformer branch (shallow layers)

(d) Fusing CNN and Transformer (e) CNN branch (deep layers) (f) Transformer branch (deep layers)

Figure 6. Feature Analysis. (**a**) Ground truth; (**b**,**e**) attention maps of shallow and deep layers in the parallel fusion layer's CNN branch; (**c**,**f**) attention maps of shallow and deep layers in the parallel fusion layer's Transformer branch; and (**d**) the attention map finally integrating the CNN and Transformer (best color effect).

Defect analysis. In order to analyze the current shortcomings of CCFNet and improve them in future work, the ground truth of Synapse is compared with the prediction results of CCFNet separately. The red box underscores regions where the performance of CCFNet falls short, as shown in Figure 7. Although the masks of CCFNet-predicted organs are very close to the ground truth labels, there are still some problems. In organ segmentation, CCFNet can sometimes accurately segment organs but cannot accurately identify the corresponding organs. For example, in the first row, in the kidney segmentation, the middle part of the left kidney is identified as the right kidney. Similarly, in the segmentation of

the spleen in the two left images of the second row, part of the spleen is identified as the stomach. In addition, for specific organs, it is challenging to segment organs, such as the pancreas. In the two images on the right-hand side in the second row, CCFNet can segment the pancreas area but cannot accurately describe the pancreas's contour, which indicates that CCFNet still has room for improvement. Processing 2D pictures will lose much spatial information from the original 3D pictures. As shown in the third row, comparing the identification results of two adjacent liver slices. The former slice can accurately segment the liver, but the latter slice cannot segment the liver completely.

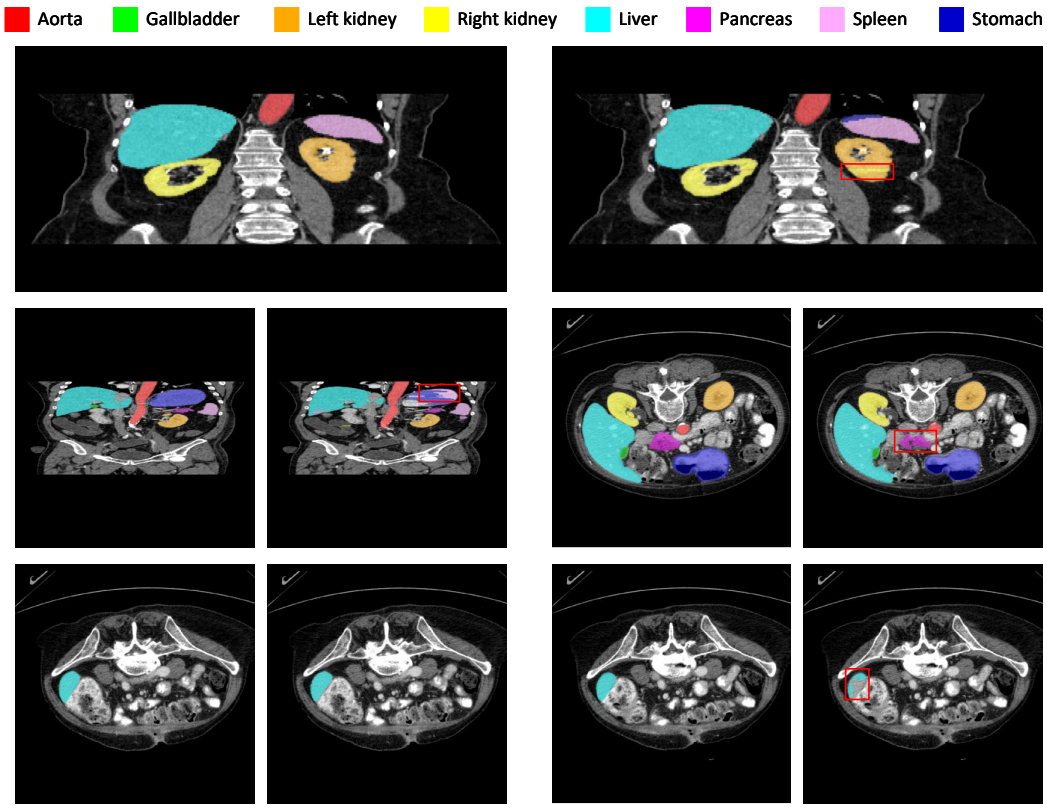


Figure 7. Qualitative visualization of CCFNet prediction results and ground truth on Synapse, each row of ground truth and model prediction results are placed in pairs (the left picture is ground truth, the right picture is the prediction result). The red box underscores regions where the performance of CCFNet falls short.

5. Conclusions

This paper first summarizes and discusses existing frameworks for medical image segmentation, and then, proposes a collaborative cross-fusion network to solve the existing problems. CCFNet utilizes the CNN's inductive bias in spatial correlation modeling and the Transformer's powerful capabilities in global relationship modeling to process the global features based on the Transformer and the local features extracted by convolutions in a parallel interactive manner. The different features extracted from the two branches are merged and exchanged employing the CCFM, which can not only encode the hierarchical local and global representations independently but also effectively aggregate the local and global representations to narrow the semantic gap between different network features. Experiments show that CCFNet displays a considerable advantage over previous Transformer-based models on various segmentation tasks, striking a balance in modeling long-term dependencies and preserving the details of underlying features, exploiting the CNN and Transformer to the fullest extent possible. We recognize that there are some limitations in our current CCFNet, particularly in terms of model performance and efficiency when dealing with highly complex datasets. In future work, we plan to design a more

lightweight and efficient parallel fusion network that can solve the problems currently found in CCFNet and test the model on more tasks. By developing a comprehensive fusion network, we expect to be able to overcome these limitations and further optimize the overall performance of the model.

Author Contributions: Conceptualization, J.C.; writing—original draft preparation, J.C. and B.Y.; writing—reviewing and editing, J.C. and B.Y.; methodology, J.C.; software, J.C.; investigation, B.Y.; funding acquisition, B.Y.; project administration, B.Y.; supervision, B.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Jiangsu Engineering Research Center of Digital Twinning Technology for Key Equipment in Petrochemical Process under Grant DTEC202003 and DTEC202102.

Data Availability Statement: The code for this paper and the data set used are available at https://github.com/cczu-CJL/CCFNet (accessed on 20 April 2022).

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

References

1. Wang, R.; Lei, T.; Cui, R.; Zhang, B.; Meng, H.; Nandi, A.K. Medical image segmentation using deep learning: A survey. *IET Image Process.* **2022**, *16*, 1243–1267. [CrossRef]
2. Jia, Y.; Kaul, C.; Lawton, T.; Murray-Smith, R.; Habli, I. Prediction of weaning from mechanical ventilation using convolutional neural networks. *Artif. Intell. Med.* **2021**, *117*, 102087. [CrossRef] [PubMed]
3. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer: Cham, Switzerland, 2015; pp. 234–241.
4. Tragakis, A.; Kaul, C.; Murray-Smith, R.; Husmeier, D. The Fully Convolutional Transformer for Medical Image Segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikola, HI, USA, 3–7 January 2023; pp. 3660–3669.
5. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
6. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
7. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2019; pp. 3146–3154.
8. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.J.; Heinrich, M.P.; Misawa, K.; Mori, K.; McDonagh, S.G.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999.
9. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11963–11975.
10. Liu, S.; Chen, T.; Chen, X.; Chen, X.; Xiao, Q.; Wu, B.; Pechenizkiy, M.; Mocanu, D.; Wang, Z. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *arXiv* **2022**, arXiv:2207.03620.
11. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large Kernel Matters – Improve Semantic Segmentation by Global Convolutional Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4353–4361.
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5999–6009.
13. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
14. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective With Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
15. Srinivas, A.; Lin, T.Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck Transformers for Visual Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 16519–16529.
16. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.

17. Wang, H.; Cao, P.; Wang, J.; Zaiane, O.R. Uctransnet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; pp. 2441–2449.
18. Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. Unetr: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 574–584.
19. Wang, H.; Xie, S.; Lin, L.; Iwamoto, Y.; Han, X.H.; Chen, Y.W.; Tong, R. Mixed transformer U-Net for medical image segmentation. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 2390–2394. [CrossRef]
20. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image Transformers distillation through attention. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual Event, 18–24 July 2021; pp. 10347–10357.
21. Bao, H.; Dong, L.; Wei, F. Beit: Bert pre-training of image Transformers. *arXiv* **2021**, arXiv:2106.08254.
22. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
23. Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H.R.; Xu, D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In Proceedings of the International MICCAI Brainlesion Workshop; Springer: Berlin/Heidelberg, Germany, 2022; pp. 272–284.
24. Matsoukas, C.; Haslum, J.F.; Söderberg, M.; Smith, K. Is it time to replace CNNs with Transformers for medical images? *arXiv* **2021**, arXiv:2108.09038.
25. Zhang, Y.; Liu, H.; Hu, Q. Transfuse: Fusing transformers and cnns for medical image segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021, Proceedings, Part I 24; Springer: Berlin/Heidelberg, Germany, 2021; pp. 14–24.
26. Heidari, M.; Kazerouni, A.; Soltany, M.; Azad, R.; Aghdam, E.K.; Cohen-Adad, J.; Merhof, D. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 6202–6212.
27. Lei, T.; Sun, R.; Wang, X.; Wang, Y.; He, X.; Nandi, A. CiT-Net: Convolutional Neural Networks Hand in Hand with Vision Transformers for Medical Image Segmentation. *arXiv* **2023**, arXiv:2306.03373.
28. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218.
29. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA , 25–28 October 2016; pp. 565–571. [CrossRef]
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
31. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
32. Zhang, Z.; Wu, C.; Coleman, S.; Kerr, D. DENSE-INception U-Net for medical image segmentation. *Comput. Methods Programs Biomed.* **2020**, *192*, 105395. [CrossRef] [PubMed]
33. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
34. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med Imaging* **2019**, *39*, 1856–1867. [CrossRef] [PubMed]
35. Ibtehaz, N.; Rahman, M.S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Netw.* **2020**, *121*, 74–87. [CrossRef] [PubMed]
36. Kaul, C.; Manandhar, S.; Pears, N. Focusnet: An attention-based fully convolutional network for medical image segmentation. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI), Venice, Italy, 8–11 April 2019; pp. 455–458. [CrossRef]
37. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
38. Jiang, M.; Yuan, B.; Kou, W.; Yan, W.; Marshall, H.; Yang, Q.; Syer, T.; Punwani, S.; Emberton, M.; Barratt, D.C.; et al. Prostate cancer segmentation from MRI by a multistream fusion encoder. *Med. Phys.* **2023**, *50*, 5489–5504. [CrossRef] [PubMed]
39. Xu, G.; Zhang, X.; He, X.; Wu, X. Levit-unet: Make faster encoders with transformer for medical image segmentation. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Xiamen, China, 13–15 October 2023; pp. 42–53.
40. Ates, G.C.; Mohan, P.; Celik, E. Dual cross-attention for medical image segmentation. *Eng. Appl. Artif. Intell.* **2023**, *126*, 107139. [CrossRef]
41. Chen, B.; Liu, Y.; Zhang, Z.; Lu, G.; Kong, A.W.K. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Trans. Emerg. Top. Comput. Intell.* **2023**, *8*, 55–68. [CrossRef]

42. Xu, S.; Xiao, D.; Yuan, B.; Liu, Y.; Wang, X.; Li, N.; Shi, L.; Chen, J.; Zhang, J.X.; Wang, Y.; et al. FAFuse: A Four-Axis Fusion framework of CNN and Transformer for medical image segmentation. *Comput. Biol. Med.* **2023**, *166*, 107567. [CrossRef]
43. Yang, H.; Yang, D. CSwin-PNet: A CNN-Swin Transformer combined pyramid network for breast lesion segmentation in ultrasound images. *Expert Syst. Appl.* **2023**, *213*, 119024. [CrossRef]
44. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ECCV), Virtual Event, 11–17 October 2021; pp. 10012–10022.
45. Zhang, N.; Yu, L.; Zhang, D.; Wu, W.; Tian, S.; Kang, X.; Li, M. CT-Net: Asymmetric compound branch Transformer for medical image segmentation. *Neural Netw.* **2024**, *170*, 298–311. [CrossRef] [PubMed]
46. Song, P.; Yang, Z.; Li, J.; Fan, H. DPCTN: Dual path context-aware transformer network for medical image segmentation. *Eng. Appl. Artif. Intell.* **2023**, *124*, 106634. [CrossRef]
47. Lin, A.; Chen, B.; Xu, J.; Zhang, Z.; Lu, G.; Zhang, D. DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 4005615. [CrossRef]
48. Valanarasu, J.M.J.; Oza, P.; Hacihaliloglu, I.; Patel, V.M. Medical transformer: Gated axial-attention for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Vancouver, BC, Canada, 8–12 October 2021; pp. 36–46.
49. Landman, B.; Xu, Z.; Igelsias, J.; Styner, M.; Langerak, T.; Klein, A. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In Proceedings of the MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge, Singapore, 8–12 September 2015; Volume 5, p. 12.
50. Bernard, O.; Lalande, A.; Zotti, C.; Cervenansky, F.; Yang, X.; Heng, P.A.; Cetin, I.; Lekadir, K.; Camara, O.; Gonzalez Ballester, M.A.; et al. Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? *IEEE Trans. Med Imaging* **2018**, *37*, 2514–2525. [CrossRef]
51. Fu, S.; Lu, Y.; Wang, Y.; Zhou, Y.; Shen, W.; Fishman, E.; Yuille, A. Domain adaptive relational reasoning for 3d multi-organ segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Lima, Peru, 4–8 October 2020, pp. 656–666.
52. Zhou, H.Y.; Guo, J.; Zhang, Y.; Yu, L.; Wang, L.; Yu, Y. nnformer: Interleaved transformer for volumetric segmentation. *arXiv* **2021**, arXiv:2109.03201.
53. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.