LOGO

A Novel Skip-Connection Strategy by Fusing Spatial and Channel wise Features for Multi-Region Medical Image Segmentation

Dayu Tan, *Member, IEEE*, Rui Hao, Xiaoping Zhou, Junfeng Xia, Yansen Su and Chunhou Zheng

***Abstract*— Recent methods often introduce attention mechanisms into the skip connections of U-shaped networks to capture features. However, these methods usually overlook spatial information extraction in skip connections and exhibit inefficiency in capturing spatial and channel information. This issue prompts us to reevaluate the design of the skip-connection mechanism and propose a new deep-learning network called the Fusing Spatial and Channel Attention Network, abbreviated as FSCA-Net. FSCA-Net is a novel U-shaped network architecture that utilizes the Parallel Attention Transformer (PAT) to enhance the extraction of spatial and channel features in the skip-connection mechanism, further compensating for downsampling losses. We design the Cross-Attention Bridge Layer (CAB) to mitigate excessive feature and resolution loss when downsampling to the lowest level, ensuring meaningful information fusion during upsampling at the lowest level. Finally, we construct the Dual-Path Channel Attention (DPCA) module to guide channel and spatial information filtering for Transformer features, eliminating ambiguities with decoder features and better concatenating features with semantic inconsistencies between the Transformer and the U-Net decoder. FSCA-Net is designed explicitly for fine-grained segmentation tasks of multiple organs and regions. Our approach achieves over 48% reduction in FLOPs and over 32% reduction in parameters compared to the state-of-the-art method. Moreover, FSCA-Net outperforms existing segmentation methods on seven public datasets, demonstrating exceptional performance. The code has been made available on GitHub: https://github.com/Henry991115/FSCA-Net.**

***Index Terms*— Medical image segmentation, parallel attention transformer, dual-path channel attention, alzheimer's disease diagnosis.**

I. Introduction

MEDICAL imaging is an essential technology for assisting doctors in evaluating diseases and formulating treatment plans. In the field of medical image processing, segmentation is crucial as it involves outlining areas of interest within medical images, extracting pertinent image features, and supplying dependable information that assists doctors in diagnosis, decision-making, and devising treatment strategies. Accurate segmentation from medical images poses a fundamental and challenging issue, offering potential advantages for disease diagnosis and treatment planning [1]. In the field of medical image analysis applications, including disease diagnosis, treatment planning, and image-guided surgery, the segmentation of human organs is considered a crucial task [2]. Due to the significant time and labor costs involved in manual medical imaging processes, the significance of automatic segmentation methods is underscored for achieving efficient and precise segmentation results [3], [4].

U-Net [5] stands out as the predominant encoder-decoder network architecture in image processing tasks. It employs an encoder-decoder architecture to separately fulfill the functions of generating low-resolution features and upsampling to recover features. Utilizing skip connections in U-Net helps to recover spatial information that can be diminished through the pooling layers. This supports the encoder-decoder U-shaped network structure to restore comprehensive spatial information. The skip-connection mechanism employed in U-Net is simplistic but may present limitations. U-Net++ [6] introduces a deeper level of connectivity by incorporating a hierarchical connection mechanism called Nested U-Net. This approach aims to enhance segmentation accuracy and preserve fine-grained details through multi-scale feature fusion. Recent research introduces UCTransNet [7], which builds on the U-Net framework and integrates a Channel-based Cross Fusion Transformer (CCT) within its skip connections. This CCT mechanism enhances the integration of multi-scale contextual information from a channel standpoint, effectively bridging the semantic divide between features of high and low levels, consequently elevating the performance of segmentation tasks.

In summary, the aforementioned three methods mentioned above target the restoration of spatial information lost during encoder downsampling in the U-shaped network. Despite demonstrating good performance, these methods still have limitations in extracting image features. Specifically, U-Net and U-Net++ utilize a linear connection mechanism, while UCTransNet employs a non-linear skip connection mechanism. Although they extract features from the channel dimension,

This work was supported in part by the National Key Research and Development Program of China (2021YFE0102100), in part by National Natural Science Foundation of China (62303014, 62172002, 62322301), in part by the University Synergy Innovation Program of Anhui Province (GXXT-2022-035), in part by Anhui Provincial Natural Science Foundation (2108085QF267, 2308085QF225), in part by Education Department of Anhui Province (2023AH050061). (Corresponding author: Yansen Su)

Dayu Tan, Rui Hao, Xiaoping Zhou, Junfeng Xia, Yansen Su, and Chunhou Zheng are with the Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, Hefei 230601, China. (e-mail: suyansen@ahu.edu.cn)

reducing the semantic divide between features of high and low levels, their capability to offset the loss of spatial information could be much better. Driven by theoretical concerns, a problem arises: how to efficiently extract spatial and channel features from information at both elevated and lower levels, thereby compensating for downsampling loss and closing the semantic divide in encoder-decoder architecture within the U-shaped network structure? To address this problem, we re-evaluate the architecture of skip connection mechanisms and introduce a deep learning network named FSCA-Net.

It is a novel U-shaped network architecture and incorporates the Parallel Attention Transformer (PAT), Cross-Attention Bridge Layer (CAB), and Dual-Path Channel Attention (DPCA) modules. The PAT module connects U-Net and Transformer, facilitating collaboration between these networks and capturing features from both channel and spatial perspectives to achieve effective concatenation. The CAB module is a bridging layer to compensate for the loss of low-level features during encoder downsampling. Additionally, The DPCA module integrates the merged features from U-Net and Transformer alongside decoder features, tackling the semantic discrepancies among these features and reducing the impact of spatial information loss caused by encoder pooling layers. The DPCA module replaces the conventional skip connections found in traditional U-Net. The main contributions of this paper are as follows:

- We design a novel feature extraction module, PAT. Within the skip connection mechanism, PAT efficiently and comprehensively extracts features by integrating both the channel and spatial dimensions of the image. This effectively compensates for the semantic and resolution gaps and losses between low-level and high-level features.
- We construct a novel feature concatenation module, DPCA, which better integrates features with semantic inconsistencies between the Transformer and U-Net decoder. It guides channel and spatial information filtering for Transformer features and eliminates the semantic ambiguity between the Transformer and decoder features.
- This study presents a novel bridge layer for the network named CAB. It aims to address the loss of detailed feature information resulting from resolution reduction via downsampling. The features processed through the bridging layer contribute more effectively to the feature concatenation in the decoder stage.

The structure of this study is as follows: Section II discusses related work. The methodology of our proposed model is detailed in Section III. Section IV delves into the specific details of the experiments, followed by a comparison with state-of-the-art methods. In Section IV, a series of discussions regarding model training, model complexity, and clinical applications are presented. Finally, Section V presents the summarization of our research conclusions.

II. Related Work

This section begins with an examination of commonly used standard segmentation methods for medical images. Subsequently, we specifically introduces the medical image segmentation approach using U-shaped networks. Finally, we summarize the relevant research on the skip-connection mechanism of U-shaped networks.

A. Medical Image Segmentation

As the field of computer vision has evolved, conventional neural network approaches have progressively become outdated. These traditional neural network methods have certain limitations, such as a limited perceptual capability for the overall global information of the image and the increase in model parameter count due to higher image resolution, leading to high computational resource requirements. The Fully Convolutional Network (FCN) [8] is introduced to overcome these limitations. FCN eliminates constraints on sample image size, increases applicability, reduces redundant structures, and improves computational efficiency. U-Net stands out for its exceptional performance and succinct structure in the realm of image processing tasks. Furthermore, several attention mechanisms have been suggested with the objective of dampening irrelevant regions within images while accentuating noteworthy features in specific local areas. For instance, the Attention U-Net [9] introduces attention gate units, enabling the network to focus on essential regions selectively, thus improving segmentation precision. Another example is the squeeze-and-excitation attention mechanism [10], which excels in segmenting retinal blood vessels, a dataset that demands precise differentiation of regions and boundaries.

Recently, the Vision Transformer (ViT) [11] has demonstrated cutting-edge performance in ImageNet classification by implementing a Transformer with global self-attention on complete images. Following the success of Transformers in numerous computer vision domains, an innovative approach to medical image segmentation has emerged [12]–[19]. As a groundbreaking Transformer-based framework for medical image processing tasks, TransUNet took the lead. Based on the transformer architecture, there has been a proliferation of work in medical image analysis for specific regions. Similar to RTN [20], which directly utilizes multiple layers of transformer blocks to construct a reinforced transformer network for coronary CT angiography vessel-level image quality assessment. To overcome the scarcity of medical imaging data, Valanarasu et al. introduced the Gated Axial Attention model, known as Med_T [21]. Leveraging the innovations introduced by the Swin Transformer [22], SwinUnet [23] introduces an entirely Transformer-based U-shaped design, substituting the traditional convolutional blocks in U-Net with Swin Transformer modules. However, these methods primarily focus on addressing the limitations of convolutional operations rather than the U-Net architecture itself, potentially resulting in structural redundancy and high computational costs [24].

B. U-shaped Nets for Medical Image Segmentation

The framework of U-shaped Networks is predominantly derived from U-Net, comprising both an encoder and decoder structure. A key advantage of U-Net is its proficient feature extraction across various levels and the amalgamation of

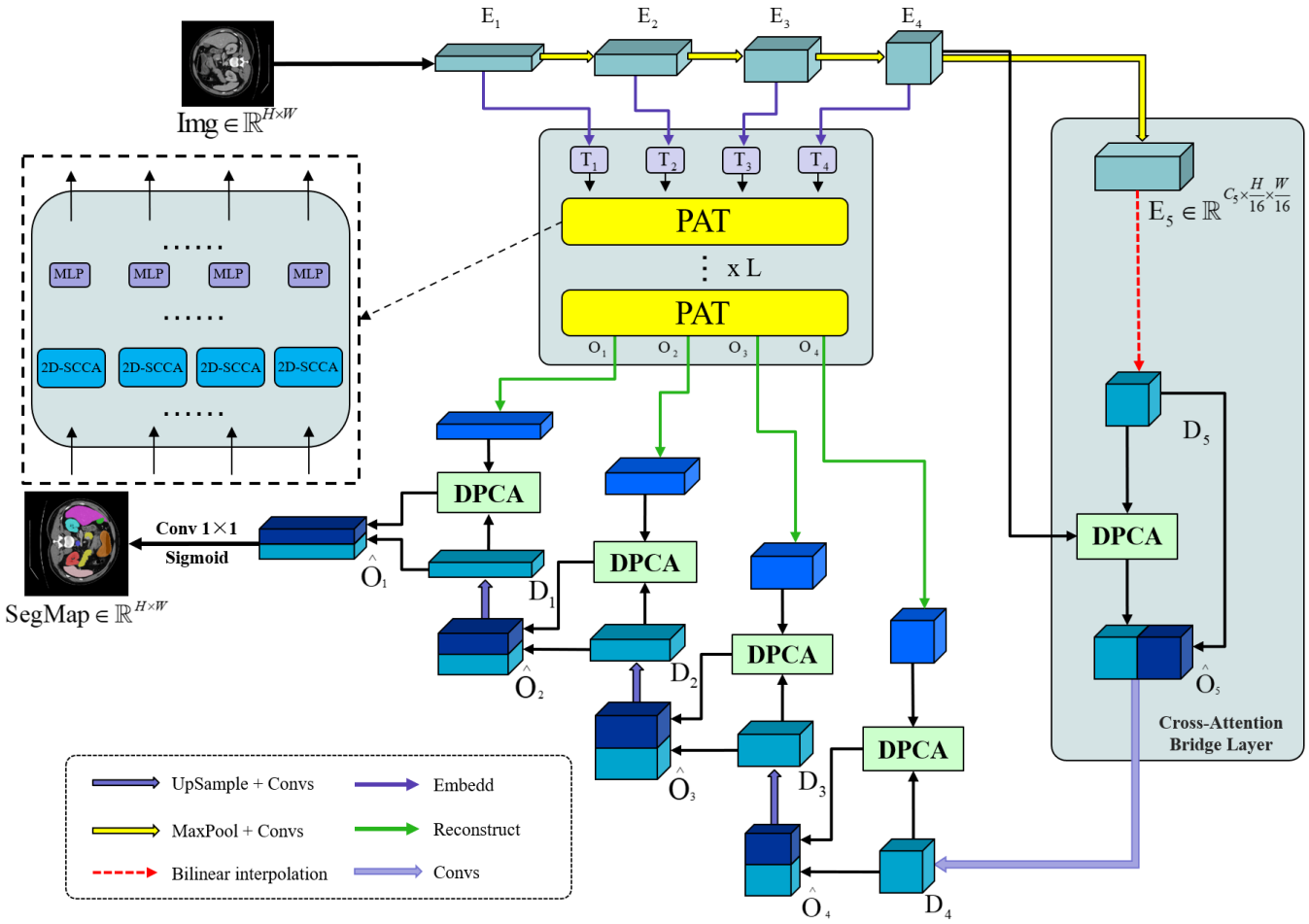


Fig. 1. Illustration of the proposed FSCA-Net. We primarily replace the original skip connections with Parallel Attention Transformer (PAT), use Cross-Attention Bridge Layer (CAB) at the bottommost level, and employ Dual-path Channel Attention (DPCA) in the upsampling part.

information via skip connections. This results in remarkable performance, particularly in tasks such as medical image segmentation. Consequently, U-Net has emerged as an indispensable asset in the realm of medical image processing tasks and has garnered substantial success in semantic segmentation tasks across diverse domains. As research in this field deepens, various improved methods based on U-Net continue to emerge.

UNet++, built upon U-Net, integrates the features from the four layers using feature concatenation, enabling the network to autonomously learn weights for features at different depths. With the introduction of attention mechanisms, many novel models have been proposed, such as Attention U-Net and Attention U-Net++ [25]. Attention U-Net introduces attention gate units, while Attention U-Net++ combines the feature pyramid structure of UNet++ with attention mechanisms. Attention mechanisms enable the network to focus better on critical regions, enhancing segmentation accuracy. Furthermore, there is a continuous influx of novel networks combining U-Net with other architectures. For instance, TransUNet combines U-Net and Transformer, SwinUnet adopts a purely Transformer approach inspired by the U-Net structure, Cascaded U-Net [26] employs a cascaded strategy utilizing multiple U-Net models, and UCTransNet operates channel-wise cross-attention for multiscale encoder feature fusion, achieving impressive segmentation results on the Glas [27] and MoNuSeg [28] datasets.

C. Skip Connections in U-shaped Nets

Initially introduced in U-Net, the skip connection mechanism is designed to mitigate semantic ambiguities due to scale issues within the encoder-decoder architecture and has shown significant effectiveness in recovering fine-grained details in target objects [29]. With the popularity of U-Net, various models have emerged that build upon and enhance its skip-connection mechanism. Some notable examples include U-Net++, U-Net 3+ [30], MultiResUNet [31], and UCTransNet.

Among the four models mentioned above, U-Net++ introduces a feature pyramid structure and employs a cascading approach to integrate features from different depths, allowing the model to perceive the importance of features across various layers. Utilizing full-scale skip connections and deep supervision, U-Net 3+ integrates semantic information from different levels. MultiResUNet aims to mitigate the semantic gap in encoder-decoder architecture by incorporating residual convolutional layers into the skip connections. UCTransNet takes

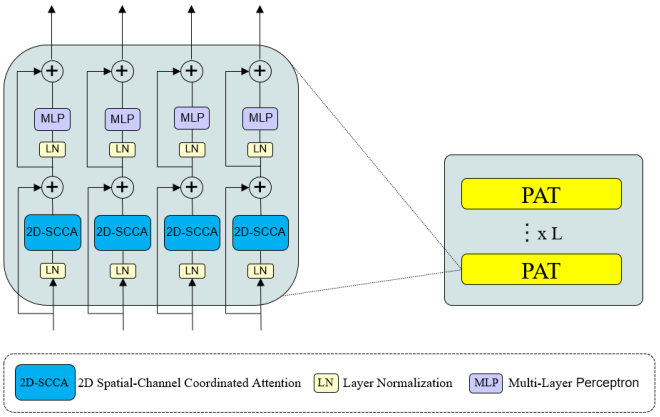


Fig. 2. The Parallel Attention Transformer (PAT) module is composed of linearly concatenated Transformer Blocks. The PAT consists of Layer Normalization (LN), 2D Spatial-Channel Coordinated Attention (2D-SCCA) and Multi-Layer Perceptron (MLP) components.

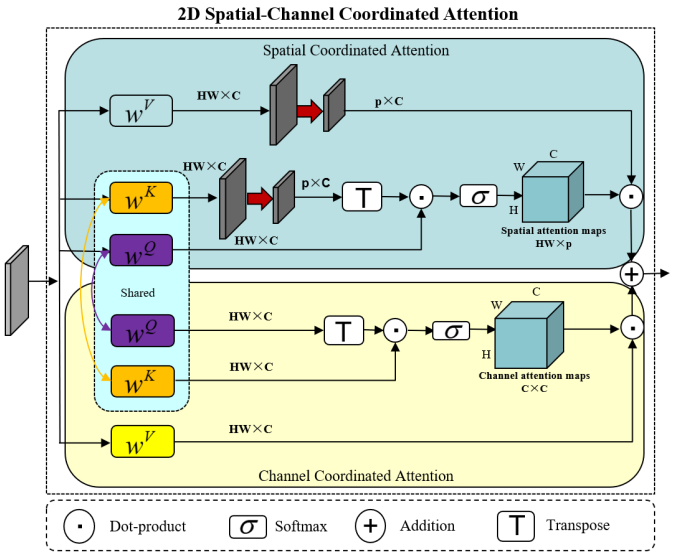


Fig. 3. 2D Spatial-Channel Coordinated Attention (2D-SCCA) is achieved by sharing the weights of Q and K layers between Channel Coordinated Attention (CCA) and Spatial Coordinated Attention (SCA).

a unique approach by introducing a Transformer branch into the U-Net architecture, replacing traditional skip-connections with a Channel Transformer (CTrans). Leveraging the Transformer Layer structure from TransUNet [32], it incorporates cross-channel attention, replacing multi-head self-attention, to integrate multi-scale channel features within the U-Net.

In TransUNet or TransFuse [17], Transformer modules are incorporated into the encoder or fused into two separate branches. However, these methods have shortcomings stemming from the limitations of the U-Net model, namely, skip connections. In the work by Wang et al. [7], they point out that there is inconsistency in the features in the encoder-decoder structure, highlighting that, in some instances, semantic variances between the shallow encoder and decoder features could result in a deficiency of semantic content within superficial features, which might restrict enhancements in performance. The conventional method of straightforward concatenation in traditional U-Net may not effectively reconcile these semantic disparities between encoder and decoder features under such circumstances.

III. Proposed Method

In this section, the proposed FSCA-Net consists of Encoder, the Bridge Layer, and Decoder. First, we illustrate the overall process of FSCA-Net, and then explain each part of FSCA-Net details.

A. FSCA-Net for Medical Image Segmentation

Fig. 1 illustrates the overall framework of the designed FSCA-Net. Expanding on the Transformer Layer structure, we integrate the Parallel Attention Transformer (PAT) module to comprehensively extract different level features across the images' channel and spatial dimensions. The Dual-Path Channel Attention (DPCA) mechanism is employed to seamlessly integrate skip-connection-generated features with those of the decoder. This approach mitigates the semantic information loss incurred during pooling processes and diminishes semantic gaps within the amalgamated features. Furthermore, we create the Cross-Attention Bridge Layer (CAB) to address the substantial loss of spatial information resulting from multiple downsampling operations.

The coordinated operation among the PAT, DPCA, and CAB modules is as follows: Specifically, the PAT module, inspired on the Transformer structure that ensures the extraction of multi-scale features across channel and spatial dimensions. Due to its characteristic of guiding channel information filtering, the DPCA module effectively integrates the feature information generated by the PAT module's Transformer structure and the decoder feature information of the U-shaped network. The CAB module ensures the capture of high-level feature information while also compensating for the loss of detail information. Moreover, through DPCA, it achieves channel dimension information filtering for deep-level features. CAB preprocesses the features for subsequent fusion in the decoder layers between the Transformer and U-shaped network, where semantic ambiguity exists.

B. PAT: Parallel Attention Transformer for Encoder Feature Transformation

We propose a Transformer-based approach called Parallel Attention Transformer (PAT), which incorporates both channel and spatial attention, as shown in Fig. 2. The PAT consists of feature embedding, Layer Normalization (LN), 2D Spatial-Channel Coordinated Attention (2D-SCCA), and Multi-Layer Perceptron (MLP) components.

To encode the block information, we utilize feature embedding. The feature embedding module is primarily responsible for preprocessing the input features from the four skip connections layers of the encoder before entering the PAT module.

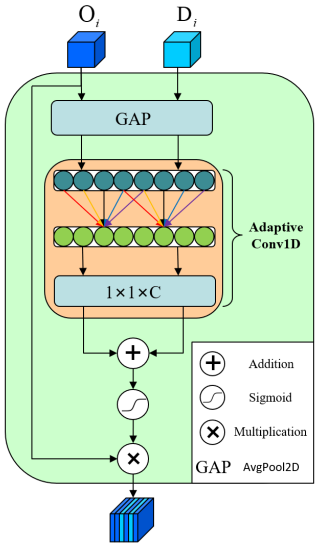


Fig. 4. Dual-Path Channel Attention (DPCA) is mainly composed of global pooling and convolutional blocks using adaptive convolutional kernels to better connect feature information between encoders and decoders.

The four skip connection layers are computed as follows:

$$E_i \in \mathbb{R}^{\frac{H}{2^{i-1}} \times \frac{W}{2^{i-1}} \times C_i} (i = 1, 2, 3, 4), \tag{1}$$

where H, W, and C correspond to the image's height, width, and channel count, respectively. We set $C_1 = 64$, $C_2 = 128$, $C_3 = 256$, and $C_4 = 512$ for our specific implementation, as this setup follows a common exponential growth pattern and is capable of enabling the encoder to capture sufficiently diverse features at every stage, ensuring that key information from the input image is preserved.

First, we reshape the four-layer features into flattened 2D block sequences of block sizes P, $\frac{P}{2}$, $\frac{P}{4}$, and $\frac{P}{8}$, respectively, where P represents the original block size. This reshaping facilitates the integration of feature channel and spatial dimensions within the PAT module.

As shown in Fig. 3, we design a coordinated attention mechanism that enables adequate global attention and captures rich spatial-channel feature representations. Specifically, it consists of the Spatial Coordinated Attention (SCA) and the Channel Coordinated Attention (CCA). SCA reduces the complexity of self-attention to linear by incorporating pairwise attention within a local neighborhood. Additionally, CCA efficiently learns the interdependencies among channel feature maps. The two attention modules are computed as follows:

$$\begin{cases} \hat{X}_c = CCA(Q_{channel}, K_{channel}, V_{channel}), \\ \hat{X}_s = SCA(Q_{spatial}, K_{spatial}, V_{spatial}), \end{cases} \tag{2}$$

$$\begin{cases} Q_{shared} = W_q X, \\ K_{shared} = W_k X, \end{cases} \tag{3}$$

where $\hat{X}_c$ and $\hat{X}_s$ are the channel and spatial attention maps, respectively. Then CCA represents the Channel Coordinated Attention module and SCA represents the Spatial Coordinated Attention module. The CCA and SCA formulas require their respective queries, keys, and value. Q_{shared}, K_{shared}, $V_{channel}$, $V_{spatial}$ are the matrices for shared queries, shared keys, channel value layer, and spatial value layer, respectively. W_q, W_k are learned weight matrices used to linearly transform input features to obtain the query vector and key vector. Furthermore, SCA and CCA share the Q_{shared} and K_{shared} to generate complementary and improved feature representations, so $Q_{channel} = Q_{shared} = Q_{spatial}$ and $K_{channel} = K_{shared} = K_{spatial}$.

The spatial attention is defined as follows:

$$Z_s = softmax(\frac{Q_{shared}\tilde{K}^T_{shared}}{\sqrt{d_k}}) \cdot \tilde{V}_{spatial}, \tag{4}$$

where $\tilde{K}^T_{shared}$ and $\tilde{V}_{spatial}$ denote projected shared keys and projected spatial value layer, respectively, and d_k is the size of each vector. The spatial attention module projects K_{shared} and $V_{spatial}$ layers from the shape of $HW \times C$ to a lower-dimensional matrix of shape $p \times C$. Then, the spatial attention map is computed by multiplying Q_{shared} layer with the transpose of the projected K_{shared} followed by applying $softmax$ to calculate the similarity between each feature and other features. Finally, the similarity is multiplied by the projected $V_{spatial}$ layer to generate the final spatial attention map of shape $HW \times C$.

The channel attention is defined as follows:

$$Z_c = V_{channel} \cdot softmax(\frac{Q^T_{shared}K_{shared}}{\sqrt{d_k}}). \tag{5}$$

Both attention mechanisms utilize the same Q_{shared} and K_{shared} layers. It calculates the channel attention map by multiplying the transpose of the Q_{shared} layer with the projected K_{shared} layer. The resulting similarity is obtained by applying $softmax$. This similarity is then multiplied by the projected $V_{channel}$ layer to generate the final channel attention map of shape $HW \times C$. Finally, the spatial and channel attention maps obtained from their respective modules are integrated to form a comprehensive attention map encompassing both spatial and dimensional aspects. This combined attention map is subsequently forwarded for additional processing within the FSCA-Net framework.

C. DPCA: Dual-Path Channel Attention for Feature Concatenation in Decoder

To better integrate the inconsistent semantic features between the U-Net decoder and the Transformer, we propose a Dual-Path Channel Attention (DPCA) module. As shown in Fig. 4, it guides the filtration of channel and spatial information from Transformer features.

Mathematically, we analyze the inputs to the DPCA module. We take $O_i \in \mathbb{R}^{C \times H \times W}$, where O_i is the i -th level PAT output, and $D_i \in \mathbb{R}^{C \times H \times W}$, where D_i is the i -th level decoder feature map, as the inputs of Dual-Path Channel Attention. In the DPCA module, we first apply global average pooling (GAP) to compress the spatial dimensions, producing vector $g(x)$ with its k^{th} channel

$$g(x) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{H} X^k(i, j), \tag{6}$$

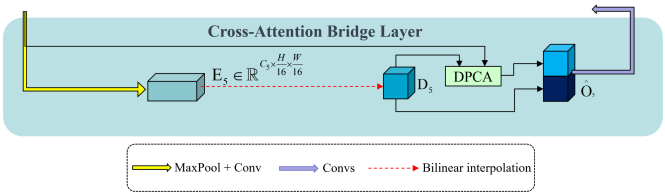


Fig. 5. Cross-Attention Bridge Layer (CAB) is designed to compensate for the loss of underlying features during downsampling.

where $g(x) \in \mathbb{R}^{C\times1\times1}$.

Next, we proceed with a one-dimensional convolution using adaptive kernel sizes to facilitate cross-channel information interaction. Unlike traditional attention mechanisms that use fully connected layers, we replace them with 1×1 adaptive convolutional layers to prevent the adverse effects of dimensionality reduction on attention mechanisms themselves.This approach not only prevents dimensionality reduction but also effectively captures inter-channel interactions. The following formula achieves the adaptive convolution kernel size:

$$k = \psi(C), \tag{7}$$

$$\psi(C) = |\frac{\log_2 C}{\gamma} + \frac{b}{\gamma}|, \tag{8}$$

where $\gamma = 2$, $b = 1$, k denotes the convolution kernel size, and C represents the channel count in the input features during the aforementioned convolution process. We use the following operation to generate the attention mask:

$$M_i = L_1 \cdot g(O_i) + L_2 \cdot g(D_i), \tag{9}$$

where $g(O_i)$ and $g(D_i)$ are the vectors obtained by applying global average pooling to O_i and D_i, respectively; L_1 and L_2 are weight parameters set for the two paths in channel attention; the feature information from the two paths is allocated after the Adaptive Conv1D layer, followed by the Addition operation. Here, we set $L_1 = L_2 = 1$.

Substantially, in our case, with the given two paths, the emphasis is on fusing the spatial and channel features from the PAT output O_i, where the dual paths are fused after one-dimensional convolution. The fused result is then multiplied channel-wise with the PAT output O_i, followed by applying the $sigmoid$ activation function, generating the final weighted feature map $\hat{O}_i$.

D. CAB: Cross-Attention Bridge Layer

In order to capture the advanced-level features and global context of an image, U-shaped segmentation networks undergo downsampling. However, the multiple downsampling steps in the feature encoding phase reduce the image resolution, leading to a loss of feature detail that cannot be compensated for merely by increasing the number of channels. Hence, we suggest employing the Cross Attention Bridge Layer (CAB), depicted in Fig. 5, as the bridge layer in the encoder-decoder framework. The feature map with the least difference in resolution from the base layer will be inputted into DPCA together with the base layer feature itself for channel dimension concatenation, ensuring the capture of high-level features while compensating for the loss of detailed information. Additionally, due to the characteristics of DPCA, the CAB can filter information on the channel for the two input feature maps.

Specifically, the E_5 obtained by downsampling E_4 is resized to match the same size as E_4. We use the bilinear interpolation for upsampling to ensure a slight scale and smooth processing of the image. The bilinear interpolation can be defined as:

$$\begin{aligned} f(P) =& (1-u)(1-v)f(Q_{11}) + u(1-v)f(Q_{21}) \\ &+ (1-u)vf(Q_{12}) + uvf(Q_{22}), \end{aligned} \tag{10}$$

where $Q_{ij}(i, j = 0, 1)$ are the four nearest pixels of pixel P and the coordinates of pixel p are (x, y); u and v are defined as $x-x_1 = u$ and $y-y_1 = v$, respectively. Next, the processed D_5 from E_5 and E_4 are inputted into a DPCA module, facilitating cross-channel information interaction between the two paths. The result is concatenated with D_5, and the output $\hat{O}_5$ is obtained through a convolutional layer.

IV. Experiments

This section begins with an introduction to the seven publicly available medical image datasets we utilized, along with the application of evaluation metrics during the experimental process. Subsequently, a detailed overview of the parameter configurations in our experiments is provided. Following that, we present the experimental results and visualizations related to attention. Finally, we outline the ablation experiments specifically designed for our model.

A. Datasets

We conduct experiments on the following seven datasets, focusing on the segmentation processing of fine-grained details in segmentation regions across all images. Additionally, apart from the CPCGEA dataset and MoNuSeg dataset, the other five datasets are multi-region segmentation datasets.

1) ***Synapse Multi-Organ Segmentation Dataset*** *[33]:* This dataset is derived from the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge. This dataset comprises abdominal CT scans from 30 patients, resulting in a 3D dataset. More specifically, it includes a total of 3,779 axial contrast-enhanced clinical CT images. Annotations for eight abdominal organs are provided. In the experimental section, we standardize the input images of the dataset to a resolution of 224×224. After random partitioning, the training set consists of 18 cases (encompassing 2212 axial slices), while the testing set comprises 12 cases.

2) ***ACDC*** *[34]:* The Automatic Cardiac Diagnosis Challenge (ACDC) collects nuclear magnetic resonance imaging results from different patients. Within the short-axis plane, pixel sizes vary between 0.83 and $1.75mm^2$. The dataset provides manually annotated images of three regions: the left ventricle (LV), right ventricle (RV), and myocardium (MYO), offering a reference standard for each patient's images. In the experimental section, we standardize the input images of the dataset to a resolution of 224×224. The dataset is randomly divided into 80 training cases (consisting of 1510 axial slices) and 20 testing cases.

3) ***COVID-19 Lung CT*** *[35]:* The dataset is compiled from the Corona Initiative and Radiopaedia, and the specific labeled images are obtained in the work by Ma et al. The dataset includes CT scan images from 20 patients diagnosed with the 2019 coronavirus disease. These scans were marked, segmented, and confirmed by radiology experts, providing mask images for the lungs and infected regions. The dataset includes annotations for two regions of the lung and the infected areas. In the experimental section, we standardize the input images of the dataset to a resolution of 512×512. The dataset is randomly divided into 16 training cases (consisting of 2837 axial slices) and 4 testing cases.

4) ***Hippocampus*** *[36]:* The data is sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, initiated by Michael W. Weiner, MD, in 2003. The project aims to monitor the progression of mild cognitive impairment and early-stage Alzheimer's disease through biomarkers, clinical assessments, and imaging techniques such as MRI and PET scans. For more information, visit www.adni-info.org.

The dataset consists of 30 brain hippocampal magnetic resonance imaging (MRI) scans from patients with Alzheimer's disease. The dataset includes original MRI images and annotated images for the left and right hippocampi. It specifically focuses on the hippocampal region. In the experimental section, we standardize the input images of the dataset to a resolution of 224×224. After random partitioning, the training set consists of 24 cases (consisting of 568 axial slices), while the testing set comprises 6 cases (consisting of 107 axial slices).

5) ***LiTS*** *[37]:* This dataset is specifically created for the research of automatic segmentation of the liver and its lesions (such as liver tumors). It originates from the LiTS Challenge at MICCAI 2017, comprising 131 training cases and 70 CT scan images for testing. The dataset provides manually annotated images of the liver and tumor. In the experimental section, due to computational resource and time constraints, we select 40 cases, standardize and process all images into 224×224 resolution images, and randomly divide the dataset into 32 training cases (4599 axial slices) and 8 testing cases.

6) ***CPCGEA****:* This dataset contains MRI data for prostate cancer patients, comprising two modalities: DWI and T2WI sequences. The images suffer from poor clarity and additional noise, possibly due to environmental conditions or image sensor issues. Artifacts like stripes and shadows may also be present, stemming from errors during image acquisition. It consists of 172 cases with manually annotated prostate cancer regions. In experiments, we standardize all images to 224×224 resolution and randomly split the dataset into 139 training cases (832 axial slices) and 33 testing cases.

7) ***MoNuSeg*** *[28]:* Careful annotations were applied to tissue images extracted from patients diagnosed with various organ tumors across multiple hospitals, forming this dataset. Formed through the acquisition of H&E (Hematoxylin and Eosin) stained tissue images at a 40x magnification, this dataset utilizes a common staining technique to enhance tissue slice contrast, commonly employed for tumor assessment (grading, staging, etc.). The dataset consists of 20 breast cancer samples and 10 prostate cancer samples. After random partitioning, 30 images are used for training, and 14 images are used for testing.

B. Evaluation Metrics

Due to variations in patterns and specific segmentation areas across the seven datasets, different evaluation metrics are employed for each dataset. For the Synapse, ACDC, COVID-19 Lung datasets, Hippocampus, and LiTS datasets, we primarily use mean DSC (Dice Similarity Coefficient) and mean HD95 (95th percentile of the Hausdorff distance) as performance evaluation metrics. For the CPCGEA dataset, which consists of single-label data, we use mean DSC, mean HD95, mean Recall, and mean Precision to comprehensively assess model performance. For the MoNuSeg dataset, we utilize mean DSC and mean IoU (Intersection over Union) to evaluate the segmentation performance of cell nuclei within the dataset.

C. Implementation Detail

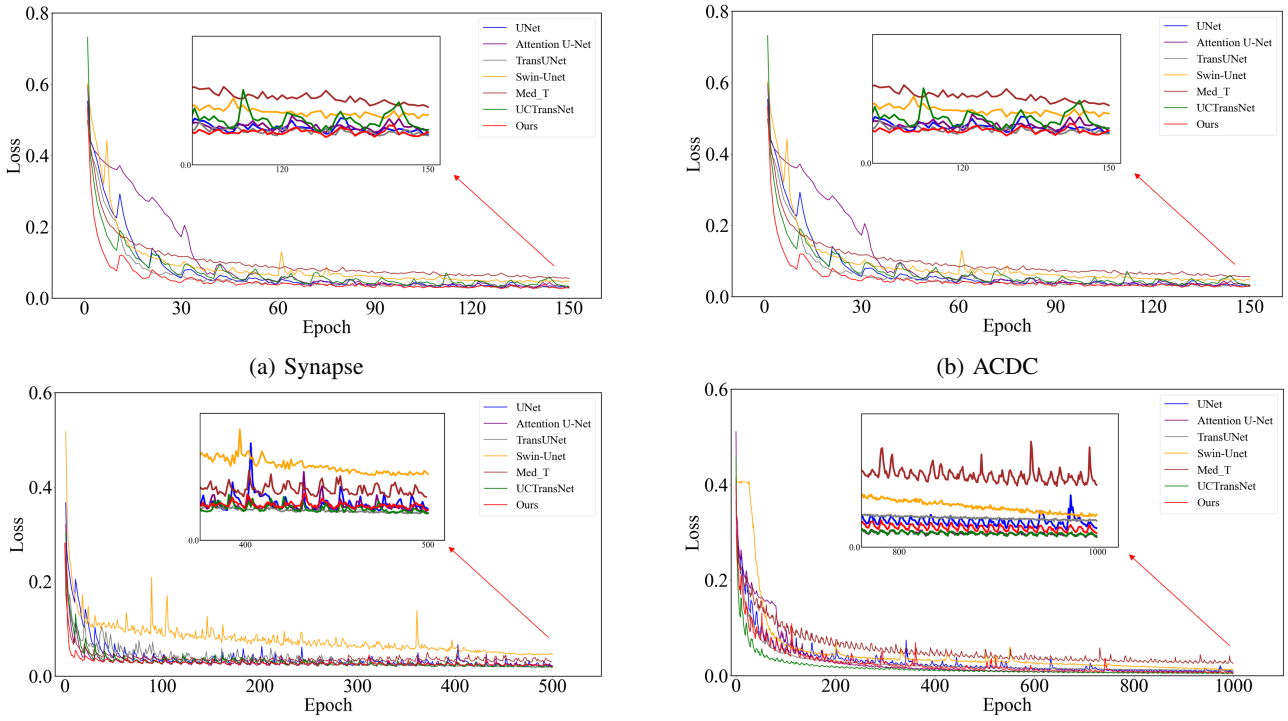
We implement our FSCA-Net model using PyTorch on an NVIDIA 3090 GPU with 24GB of memory. To prevent overfitting, a series of data augmentation operations are applied to the dataset before feeding the images into the model. Our FSCA-Net model is trained from scratch. For the ACDC and Synapse datasets, we set the batch size to 24 [23]. The batch size is set to 16 in the Hippocampus dataset and the LiTS dataset. For the CPCGEA dataset, the batch size is set to 8. The batch size for both the COVID-19 Lung CT and MoNuSeg datasets is established at 4 [15]. The patch size P for all seven datasets is set to 16. We utilize the Adam optimizer for model training, setting the learning rate to 0.001, with the loss function throughout the training process being a combination of cross-entropy loss and Dice loss, weighted as specified. It can be defined as:

$$\begin{aligned} L(Y,P) =& 1 - \sum_{i=1}^{I} (\lambda \frac{2 * \sum_{n=1}^{N} Y_{n,i} \cdot P_{n,i}}{\sum_{n=1}^{N} Y_{n,i}^2 + \sum_{n=1}^{N} P_{n,i}^2} \\ &+ \sum_{n=1}^{N} Y_{n,i} log P_{n,i}), \end{aligned} \tag{11}$$

where I represents the class count and N signifies the total voxel number; $Y_{n,i}$ and $P_{n,i}$ are the ground truths and output probabilities at voxel v for class i, respectively. We perform various experiments and present the average outcomes. Statistical analyses indicate that our method notably surpasses comparable approaches.

D. Comparison With Other Methods

We employ six state-of-the-art methods to validate the performance of FSCA-Net: U-Net, Attention U-Net, TransUNet, SwinUnet, Med_T, and UCTransNet. Among them, we draw loss maps for the training process of four datasets. In Fig. 6, the training loss trends of all methods are illustrated across the four datasets.



(a) Synapse (b) ACDC

(c) COVID-19 Lung (d) Hippocampus

Fig. 6. The training loss trends the proposed method and six comparative methods on the four datasets.

TABLE I
COMPARISONS WITH STATE-OF-THE-ART MODELS ON THE SYNAPSE MULTI-ORGAN SEGMENTATION DATASET. THE PERFORMANCES ON SEGMENTING THE EIGHT ORGANS ARE ALL REPORTED.

Method	DSC↑(%, mean)	HD95↓(mm, mean)	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
			DSC↑(%)							
U-Net	75.98	33.224	87.12	62.74	81.66	75.08	93.28	52.93	82.92	72.14
Attention U-Net	76.72	31.390	88.87	63.62	82.91	73.57	93.72	56.50	85.39	69.21
TransUNet	77.57	27.969	86.88	63.16	81.78	76.94	94.22	56.78	84.65	76.12
SwinUnet	78.46	24.607	85.17	65.73	82.17	78.67	94.00	57.00	89.03	75.90
Med_T	69.92	38.901	82.32	55.84	67.46	66.94	90.85	46.97	81.47	67.48
UCTransNet	77.45	36.213	88.67	61.21	81.13	76.01	93.70	59.16	87.76	71.93
Ours	**79.12**	29.100	87.90	**67.60**	**83.65**	**79.05**	93.80	**59.40**	86.85	74.70

TABLE II
COMPARISONS WITH STATE-OF-THE-ART MODELS ON THE ACDC DATASET.

Method	DSC↑ (%, mean)	HD95↓ (mm, mean)	RV	Myo	LV
			DSC↑(%)		
U-Net	89.42	2.907	87.42	87.60	93.23
Attention U-Net	88.49	1.511	85.33	86.85	93.29
TransUNet	90.12	1.660	89.22	88.11	93.05
SwinUnet	90.03	1.968	87.96	88.20	93.93
Med_T	84.96	3.430	80.63	83.80	90.46
UCTransNet	90.37	1.722	88.50	88.52	94.09
Ours	**91.44**	**1.107**	**89.32**	**90.15**	**94.87**

TABLE III
COMPARISONS WITH STATE-OF-THE-ART MODELS ON THE COVID-19 LUNG CT DATASET.

Method	DSC↑ (%, mean)	HD95↓ (mm, mean)	LL	RL	INF
			DSC↑(%)		
U-Net	70.32	112.736	85.38	79.51	46.08
Attention U-Net	73.89	91.177	88.40	88.43	44.84
TransUNet	82.69	42.425	95.31	96.13	56.64
SwinUnet	79.75	48.059	92.19	94.74	52.31
Med_T	80.55	33.122	94.24	94.78	52.62
UCTransNet	76.68	71.680	90.02	90.93	49.08
Ours	**83.63**	**15.124**	**96.27**	**96.81**	**57.80**

*1) **Experiments on the Synapse Multi-Organ Segmentation Dataset**:* Table I presents the results of the Synapse dataset. SwinUnet outperforms other existing methods, but our method achieves better segmentation performance for the aorta, gallbladder, kidneys, and pancreas compared to SwinUnet. The DSC of FSCA-Net reached 79.12%. Due to the large number of segmentation classes and the limitation of 2D models to learn information only from the 2D context of each slice, failing to fully utilize the continuity and context information of organs in the depth of the volume, the model cannot achieve optimal performance for segmenting the eight organs.

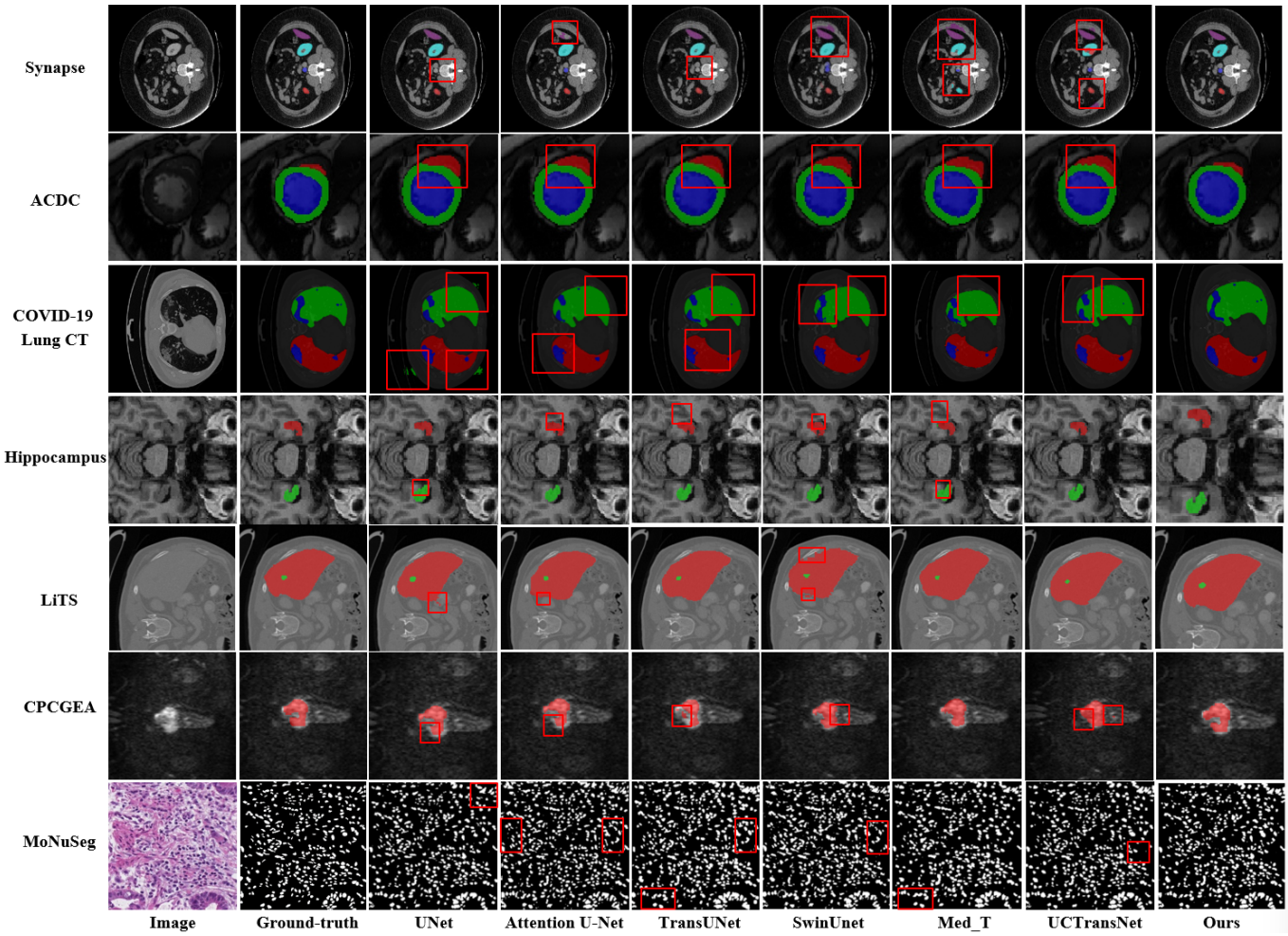


Fig. 7. The visualization results of FSCA-Net and six comparison methods for image segmentation on seven datasets, as well as the original image and GroundTruth. Incorrect segmentation areas are marked with red boxes.

The first row in Fig. 7 illustrates the visual comparison of the seven segmentation methods we have enumerated. Incorrectly segmented regions are marked with red boxes. U-Net and TransUNet fail to segment the aorta region, while Attention U-Net, SwinUnet, Med_T, and UCTransNet make errors in the liver and left kidney segmentation. We can observe that existing methods produce blurred boundaries for the liver and confuse the stomach, pancreas, spleen, and liver, whereas our method depicts the boundaries between them more smoothly.

*2) **Experiments on the ACDC dataset**:* Table II presents the results on the ACDC dataset. Based purely on CNN, the U-Net method achieves a DSC of 89.42%. SwinUnet and Med_T, based purely on Transformer, achieve DSCs of 90.03% and 84.96%, respectively. On this dataset, UCTransNet outperforms other existing methods. Our FSCA-Net achieves a DSC of 91.44%, surpassing UCTransNet.

The second row in Fig. 7 illustrates a qualitative comparison of FSCA-Net with existing segmentation methods on the ACDC dataset. Incorrectly segmented regions are marked with red boxes. Existing methods generally perform well in segmenting the Myo and the LV, but some methods often mistakenly classify parts of the background as the RV, with U-Net and UCTransNet showing significant misidentification of the RV area. In contrast, our proposed method improves the accuracy of RV segmentation.

In the examples shown, FSCA-Net segments all the organs more accurately and depicts the boundaries.

*3) **Experiments on the COVID-19 Lung CT dataset**:* Results on the COVID-19 Lung CT dataset are presented in Table III, demonstrating the significant segmentation performance of FSCA-Net. Segmenting the infected area is challenging as it is scattered and surrounded by lung regions, but FSCA-Net achieves superior performance in this region as well. Specifically, FSCA-Net improves the DSC score for the segmentation of infected areas by 1.16% compared to the best-performing existing method, TransUNet.

The third row in Fig. 7 shows visual comparisons of our method with other segmentation methods. Incorrectly segmented regions are marked with red boxes. The red and green regions represent the left and right lungs, respectively, while the blue indicates the infected area. Through visual observation of the images, we find that when using U-Net for image segmentation, the background is not correctly identified, leading to errors, and all existing methods have minor segmentation omissions in the infected areas of the lungs. Overall, FSCA-

TABLE IV
COMPARISONS WITH STATE-OF-THE-ART MODELS ON THE HIPPOCAMPUS DATASET.

Method	DSC↑ (%, mean)	HD95↓ (mm, mean)	LH DSC↑(%)	RH DSC↑(%)
U-Net	70.61	4.930	68.04	73.18
Attention U-Net	75.40	12.913	75.90	74.90
TransUNet	78.74	5.097	78.12	79.37
SwinUnet	79.66	3.381	80.00	79.33
Med_T	65.47	19.972	65.60	65.33
UCTransNet	75.65	8.155	76.40	74.90
Ours	**88.35**	**1.916**	**88.68**	**88.02**

TABLE V
COMPARISONS WITH STATE-OF-THE-ART MODELS ON LITS DATASET.

Method	DSC↑ (%, mean)	HD95↓ (mm, mean)	Liver DSC↑(%)	Tumor DSC↑(%)
U-Net	62.99	38.226	91.56	34.44
Attention U-Net	63.32	27.029	91.88	34.77
TransUNet	65.57	32.511	88.03	43.11
SwinUnet	64.72	34.302	91.69	37.74
Med_T	58.06	41.500	89.54	26.58
UCTransNet	64.95	24.496	92.90	36.99
Ours	**68.54**	**23.897**	**93.04**	**44.04**

TABLE VI
COMPARISONS WITH STATE-OF-THE-ART MODELS ON THE CPCGEA DATASET.

Method	DSC↑ (%, mean)	HD95↓ (mm, mean)	Recall↑ (%, mean)	Precison↑ (%, mean)
U-Net	60.79	9.508	65.32	66.50
Attention U-Net	61.92	7.752	64.75	65.37
TransUNet	62.57	7.895	65.14	66.69
SwinUnet	62.35	7.466	68.98	62.51
Med_T	60.47	7.284	64.82	66.19
UCTransNet	60.03	13.492	69.10	61.21
Ours	**64.27**	**7.047**	**69.92**	**66.93**

TABLE VII
COMPARISONS WITH STATE-OF-THE-ART MODELS ON THE MONUSEG DATASET.

Method	DSC↑(%, mean)	IoU↑(%, mean)
U-Net	78.46	65.71
Attention U-Net	78.48	65.61
TransUNet	78.04	64.06
SwinUnet	78.48	64.73
Med_T	78.48	64.70
UCTransNet	78.25	64.90
Ours	**79.82**	**66.60**

Net effectively handles the segmentation of small infected areas and achieves optimal performance in liver segmentation.

In summary, FSCA-Net accurately segments the gallbladder, kidneys, and pancreas, and also achieves good performance in the segmentation of aorta, liver, spleen, and stomach.

*4) **Experiments on the Hippocampus dataset**:* The results on the Hippocampus dataset are showcased in Table IV. Our approach significantly outperforms existing methods in segmenting the left and right hippocampus, achieving DSC scores of 88.68% and 88.02%, respectively.

The fourth row in Fig. 7 shows visual comparisons of FSCA-Net with other segmentation methods. Incorrectly segmented regions are marked with red boxes. Existing methods produce blurry boundaries for both hippocampus and missing segmentation regions. For example, TransUNet mistakenly identifies a small region inside the left hippocampus as the background.

*5) **Experiments on the LiTS dataset**:* According to Table V, TransUNet achieves an average DSC of 65.57%, demonstrating significant effectiveness in tumor segmentation. SwinUnet, on the other hand, balances liver and tumor segmentation with an average DSC of 64.72%. Our approach surpasses TransUNet by 1.52% in average DSC, with the highest DSC for tumor segmentation reaching 44.04%.

As shown in the visualization in Fig. 7, the comparison among all methods is presented. Incorrectly segmented regions are marked with red boxes. U-Net shows instances of notable mis-segmentation in liver segmentation, while Attention U-Net and SwinUnet show limited smoothness in liver edge handling. For tumor segmentation, SwinUnet produces noticeably inaccurate shapes.

In summary, our method demonstrates superior performance in both liver and tumor segmentation. However, due to resource constraints and inherent limitations of 2D models, there remains considerable potential for enhancing tumor segmentation.

*6) **Experiments on the CPCGEA dataset**:* To further validate the robustness of the model regarding variations in image quality, noise, and artifacts, we engage the experiments on this dataset. Firstly, as indicated in Table VI, it can be observed that TransUnet achieves the best DSC and Precision metrics among existing methods, reaching 62.57% and 66.69%, respectively. Med_T exhibits the optimal HD95 metric among existing methods at $7.284mm$. Additionally, UCTransNet achieves the highest Recall among existing methods, reaching 69.10%. Subsequently, our method surpasses all comparative methods across four metrics, notably achieving a DSC of 64.27%. Thus, our method demonstrates superior segmentation performance under variations in image quality, noise, and artifacts.

As depicted in the 6th row of Fig. 7, we visually compare FSCA-Net with six other comparative methods. Incorrectly segmented regions are marked with red boxes. U-Net, Attention U-Net, TransUNet, and SwinUnet exhibit significant areas of under-segmentation, while UCTransNet shows instances of notable mis-segmentation. Overall, FSCA-Net exhibits the most superior segmentation performance.

*7) **Experiments on the MoNuSeg dataset**:* Considering that the previous experiments focus on multi-organ/region segmentation, here we select a dataset for segmenting a single region with a smaller data size. Through this approach, we aim to validate that our model is also meaningful for single organ/region segmentation. Table VII presents the results of the MoNuSeg dataset. FSCA-Net achieves a DSC of 79.82% and an IoU of 66.60%, outperforming existing methods.

The seventh row in Fig. 7 visually compares the FSCA-Net with other segmentation methods. Regions that were incorrectly segmented are highlighted with red boxes. We can observe that U-Net does not segment local regions finely

TABLE VIII
ABLATION EXPERIMENTS ON THE SYNAPSE MULTI-ORGAN SEGMENTATION DATASET.

Method	DSC↑(%, mean)	HD95↓(mm, mean)	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
			DSC↑(%)							
Baseline	77.45	36.213	88.67	61.21	81.13	76.01	93.70	59.16	87.76	71.93
+PAT	77.88	31.888	87.75	67.50	81.80	76.60	93.70	58.15	83.80	73.70
+DPCA	77.57	31.068	87.58	64.81	80.98	76.34	93.69	59.18	86.18	71.78
+CBA	77.49	34.038	87.29	64.54	81.04	75.69	93.79	58.77	86.52	72.28
+PAT+DPCA	77.95	30.660	87.62	66.61	83.02	76.52	93.61	57.12	84.91	74.22
+PAT+CAB	78.27	30.621	87.05	66.60	82.60	78.93	93.19	57.50	86.74	73.56
+DPCA+CAB	78.20	29.244	87.38	66.89	82.62	78.69	93.55	58.24	85.58	72.68
+PAT+DPCA+CAB	**79.12**	**29.096**	87.91	**67.60**	**83.63**	**79.06**	**93.81**	**59.41**	86.89	**74.67**

TABLE IX
ABLATION EXPERIMENTS ON THE ACDC DATASET.

Method	DSC↑ (%, mean)	RV	Myo	LV
		DSC↑(%)		
Baseline	89.54	87.73	87.44	93.44
+PAT	90.51	88.65	88.69	94.17
+DPCA	90.65	88.88	88.99	94.10
+CAB	90.78	88.95	89.40	94.00
+PAT+DPCA	91.03	89.21	89.63	94.24
+PAT+CAB	91.06	88.40	90.01	94.78
+DPCA+CAB	90.87	88.77	89.63	94.21
+PAT+DPCA+CAB	**91.44**	**89.32**	**90.15**	**94.87**

TABLE X
ABLATION EXPERIMENTS ON THE MONUSEG DATASET.

Method	DSC↑(%, mean)	IoU↑(%, mean)
Baseline	77.95	64.50
+PAT	78.65	65.04
+DPCA	78.44	65.34
+CAB	78.53	65.66
+PAT+DPCA	79.24	65.84
+PAT+CAB	79.18	65.71
+DPCA+CAB	79.05	65.50
+PAT+DPCA+CAB	**79.82**	**66.60**

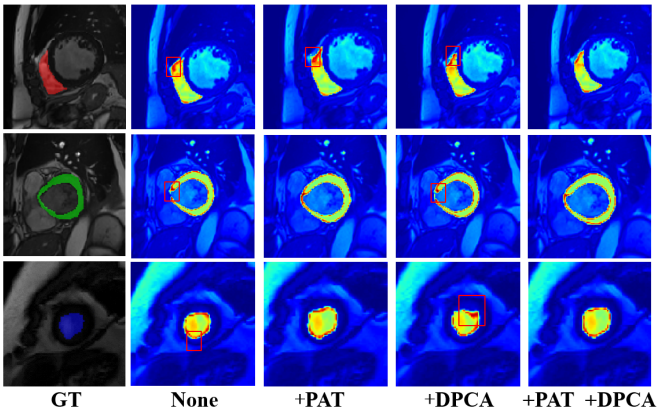


Fig. 8. The visualization results of two types of attention on the ACDC dataset.

enough, while some other comparative models may exhibit over-segmentation. It is noticeable that our FSCA-Net produced more accurate segmentation results, closely matching the ground-truth compared to the baseline models.

E. Ablation Studies

Our approach mainly introduces three modules: the Parallel Attention Transformer (PAT) module, the Cross-Attention Bridge Layer (CAB) module, and the Dual-path Channel Attention (DPCA) module. Our ablation experiments were conducted on three datasets.

Regarding the proposed modules displayed in Tables VI, VII, and VIII, "Baseline + PAT + DPCA + CAB" generally surpasses all other combinations on all datasets, demonstrating the effectiveness of this particular combination.

First, when we add any one of the PAT, DPCA, or CAB modules individually, the segmentation results for various organs in the ACDC and MoNuSeg datasets are significantly improved. On the Synapse dataset, the average segmentation results for eight organs are also enhanced. Next, we adopt a two-by-two combination approach of the three modules, resulting in further improved segmentation results on all three datasets. Finally, our segmentation performance reaches its best when we use all three modules: PAT, DPCA, and CAB. Overall, our approach achieves superior performance in all segmentation results on the ACDC and MoNuSeg datasets. Additionally, on the Synapse dataset, our method demonstrates improved segmentation results in six out of eight organs.

The experimental outcomes also underscore the significance of improving feature extraction capability within the encoder-decoder framework.

F. Attention Mechanism Visualization

In the designed model, we have utilized two attention-based modules. We employ attention heatmap visualization to visually demonstrate the roles of these modules in segmenting specific regions.

We select the ACDC dataset, specifically for segmenting the RV cavity, myocardium, and LV cavity, and conduct attention visualization analysis based on the two attention modules used in our model: PAT and DPCA.

The visualizations of Fig. 8 clearly demonstrate that including the PAT module within the skip connections significantly enhances the extraction of global contextual information. However, including PAT in the skip connections may lead to side effects, such as marking incorrect regions. On the other hand, the DPCA module is dedicated to filtering the feature

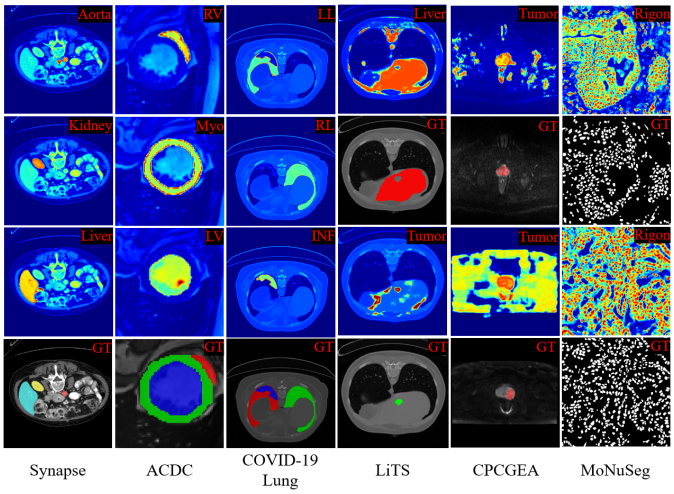


Fig. 9. The visualization images of attention weight heatmaps on multiple datasets.

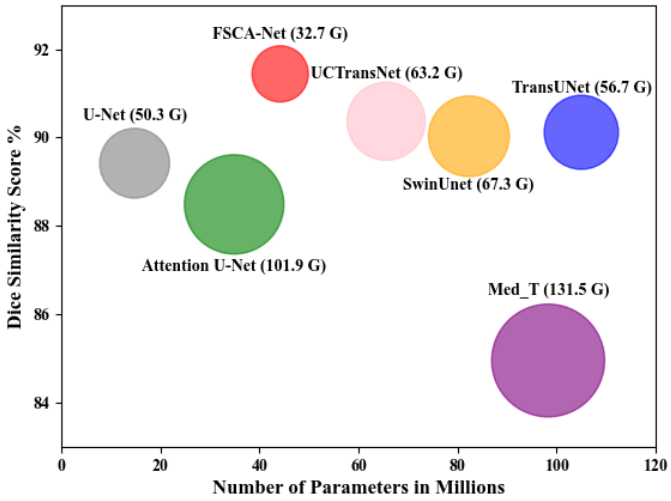


Fig. 10. Accuracy (Dice score) vs. model complexity (parameters and computational complexity) comparison on the ACDC dataset.

TABLE XI
EXPERIMENT ON DETERMINING THE NUMBER OF PATS ON THE ACDC DATASET.

Layer	DSC↑ (%, mean)	RV	Myo DSC↑(%)	LV
2	91.07	88.54	89.97	94.70
3	91.04	88.54	89.85	94.75
4	**91.44**	**89.32**	**90.15**	**94.87**
5	90.93	88.32	89.88	94.58
6	90.99	88.79	89.77	94.41
7	90.84	88.29	89.81	94.42
8	91.16	88.79	89.84	94.85

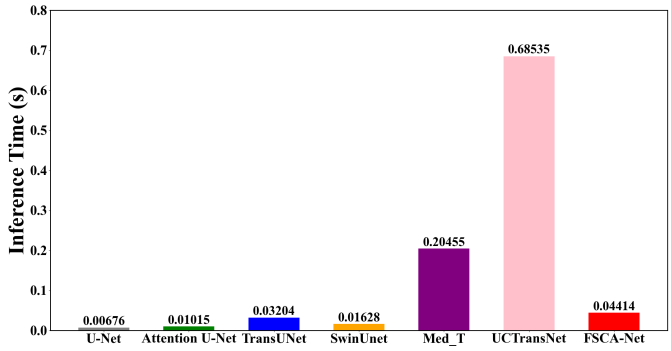


Fig. 11. The average inference time per image for existing methods and our method.

in the encoder-decoder framework, resulting in fine-grained optimization of the segmentation boundaries.

G. Discussion

To further validate the experiments and explore the practical clinical application, we engage in the following series of discussions:

1) Discussion on the the image characteristics: To visually demonstrate the performance disparities of the designed model across various datasets, we utilize attention weight heatmaps visualization. Based on Fig. 9, it's evident that on the Synapse dataset, the designed model accurately captures the regions of interest, resulting in an optimal segmentation performance. However, due to the dense distribution among the eight segmentation labels and minimal differences in image features, the segmentation performance is affected. On the ACDC and COVID-19 Lung datasets, distinct differences in segmentation area features and more dispersed regions contribute to the model's excellent segmentation performance. On the MoNuSeg dataset, despite dense segmentation areas, higher dataset quality and single segmentation label enable the designed model to achieve notable segmentation performance. On the LiTS dataset, the minimal feature differences between tumor and liver regions can cause attention weights to be dispersed to background areas, thereby limiting tumor segmentation capabilities. Fig. 9 shows that the designed model still effectively extracts the regions of interest, particularly the tumor regions. On the CPCGEA dataset, significant noise and poor image quality directly affect the final segmentation performance, but the designed model still achieves the best segmentation performance.

Through attention weight heatmap visualizations, we demonstrate the differences in the segmentation performance of FSCA-Net across different datasets, providing insights into its robustness and generalization capabilities.

2) Discussion on the number of PAT: In our skip-connection design, we propose the PAT module. Choosing the number of modules in this design becomes essential to extract spatial and channel-wise features from high and low-level layers. Due to limited computational resources, we experimentally selected the number of modules to be linearly varied from 2 to 8.

According to the experimental results presented in Table XI, our approach attains the most favorable segmentation performance for the three organs in the ACDC dataset when employing four layers of concatenation. It should be noted that our experiments on all seven public datasets are conducted with 4 PAT modules in series.

3) Discussion on the model complexity and inference time: In our study, we focus on balancing model performance,

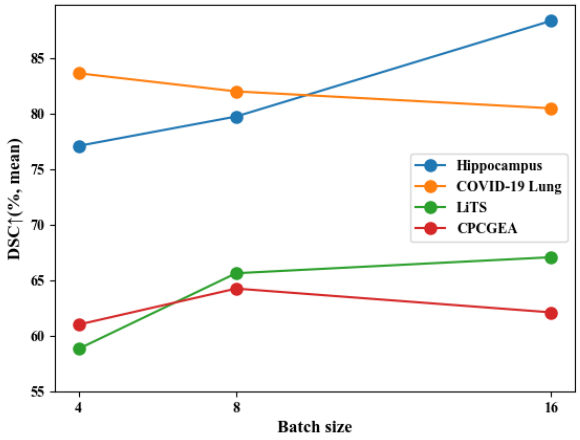


Fig. 12. The discussion on the selection of batch size.

model complexity, and inference time. Therefore, we engage a series of experiments where our proposed model is thoroughly compared against six comparison models.

The results are detailed in Fig. 10, which includes the DSC, model parameter count, and computational complexity of each model. Firstly, in terms of DSC scores, our model performs best in segmenting the three organs within the cardiac region, indicating significant effectiveness in enhancing segmentation accuracy. More importantly, not only does our model exhibit superior performance, but it also has the smallest computational complexity and fewer parameters.

Additionally, we calculate the inference time of the model, the results of which are shown in Fig. 11. From the figure, it can be seen that the inference time of FSCA-Net is significantly less than that of Med_T and UCTransNet, but is slightly inferior to some other U-shaped models.

Overall, our method optimizes the model architecture to effectively balance performance, complexity, and inference time. This demonstrates its immense potential for clinical practice.

4) Discussion on the batch size: To discuss the impact of batch size on our entire training process, we engage some experiments on the Hippocampus, COVID-19 Lung, LiTS, and CPCGEA datasets. Due to computational resource constraints, we only discuss the cases with batch sizes of 4, 8, and 16.

As depicted in Fig. 12, on the Hippocampus dataset, a batch size of 16 yields the optimal segmentation performance. On the COVID-19 Lung dataset, the best batch size is 4. Similarly, on the LiTS dataset, the best segmentation performance is achieved with a batch size of 16. On the other hand, for the CPCGEA dataset, a batch size of 8 results in the best segmentation performance. Furthermore, it is evident that the choice of batch size directly affects the model training process, thereby impacting segmentation performance.

5) Discussion on the actual clinical application: Considering the model's short inference time, low parameter count, and minimal computational requirements, its real-world clinical applications are both promising and diverse. These attributes enable the model to be effectively integrated into various clinical settings, improving efficiency and patient outcomes. Here are several ways to discuss its real-world clinical application:

Integration into Clinical Workflow: Given the model's short inference time, it can be seamlessly integrated into existing clinical workflows. For instance, it could be employed immediately after image acquisition to provide real-time segmentation analysis, thereby expediting the diagnostic process and enhancing workflow efficiency.

Usability for Medical Practitioners: With its low parameter count and minimal computational requirements, the model can be deployed on smart devices or cloud servers, ensuring easy accessibility for medical practitioners. This eliminates the need for complex hardware or software configurations, allowing medical personnel to swiftly perform image segmentation and obtain results.

Impact on Diagnostic Accuracy: Despite its low parameter count, the model has demonstrated proficient performance in image segmentation tasks. Therefore, its application in clinical practice holds the potential to improve diagnostic accuracy. Medical practitioners can use the model's segmentation results for more precise lesion detection and localization, improving diagnosis accuracy and reliability.

In conclusion, leveraging the model's characteristics, it can serve as a rapid, efficient, user-friendly, and accurate tool to provide better support and services in clinical medicine.

V. Conclusion

We propose a deep learning network model named FSCA-Net, designed for fine-grained segmentation of multiple regions. The proposed method improves semantic segmentation performance by fusing spatial and channel-wise features in the skip-connection mechanism. Our objective is to bridge the semantic and resolution gaps between features at different levels by using more effective feature concatenation methods. To achieve this, we introduce the dual-path channel attention module, which guides the channel and spatial dimensions of Transformer features and facilitates information filtering. This module serves as a novel feature concatenation component that replaces the traditional simple concatenation approach. Additionally, we propose the cross-attention bridge layer to compensate for the loss of low-level features. FSCA-Net achieves promising segmentation results on seven public datasets compared to existing methods. Not only have we enhanced semantic segmentation performance, but we have also achieved higher efficiency and compactness through the combination of U-Net and Transformer.

Although our current approach can achieve high-performance segmentation while ensuring a small number of parameters and computational cost, it lacks consideration for cross-dataset performance. Due to issues such as differences in data distribution, label formats, and limited data volume among different datasets, the designed model currently cannot efficiently perform cross-dataset validation. In future research, we aim to create a holistic medical image segmentation model for the entire human body, enhancing both diagnostic accuracy and treatment efficiency. We look forward to our research contributing to the advancement of the medical field.

ACKNOWLEDGMENT

The project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI), including support from the National Institutes of Health (NIH) and the Department of Defense. ADNI receives contributions from the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, various companies, and organizations. Funding also includes support from the Canadian Institutes of Health Research for Canadian sites. Private sector donations are facilitated through the Foundation for the National Institutes of Health. The research is granted by the Northern California Institute for Research and Education, coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California, and data distribution is managed by the Laboratory for Neuro Imaging at the University of Southern California.

REFERENCES

[1] A. He et al., "H2Former: An Efficient Hierarchical Hybrid Transformer for Medical Image Segmentation," *IEEE Transactions on Medical Imaging,* vol. 42, no. 9, pp. 2763–2775, Sep. 2023.

[2] H. Tang et al., "Clinically applicable deep learning framework for organs at risk delineation in CT images," *Nature Mach. Intell.*, vol. 1, no. 10, pp. 480-491, Sep. 2019.

[3] Y. Wang et al., "Abdominal multi-organ segmentation with organ-attention networks and statistical fusion," *Med. Image Anal.*, vol. 55, pp. 88-102, Jul. 2019.

[4] Q. F. Zhao et al., "Efficient Multi-Organ Segmentation From 3D Abdominal CT Images With Lightweight Network and Knowledge Distillation," *Med. Image Anal.*, vol. 42, pp. 2513-2523, Mar. 2023.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015,* pp. 234-241, 2015.

[6] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support,* pp. 3-11, 2018.

[7] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-wise Perspective with Transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 36, pp. 2441-2449, 2022.

[8] J. Long et al., "Fully Convolutional Networks for Semantic Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3431-3440, 2015.

[9] O. Oktay et al., "Attention U-Net: Learning Where to Look for the Pancreas," arXiv preprint arXiv:1804.03999, 2018.

[10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 7132-7141, 2018.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020.

[12] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers," arXiv preprint arXiv:2012.15840, 2020.

[13] Y. Zhang, R. Higashita, H. Fu, Y. Xu, Y. Zhang, H. Liu, J. Zhang, and J. Liu, "A Multi-Branch Hybrid Transformer Network for Corneal Endothelial Cell Segmentation," arXiv preprint arXiv:2106.07557, 2021.

[14] Z. Gao, B. Hong, X. Zhang, Y. Li, C. Jia, J. Wu, C. Wang, D. Meng, and C. Li, "Instance-Based Vision Transformer for Subtyping of Papillary Renal Cell Carcinoma in Histopathological Image," arXiv preprint arXiv:2106.12265, 2021.

[15] Y. Ji, R. Zhang, H. Wang, Z. Li, L. Wu, S. Zhang, and P. Luo, "Multi-Compound Transformer for Accurate Biomedical Image Segmentation," arXiv preprint arXiv:2106.14385, 2021.

[16] Y. Gao, M. Zhou, and D. Metaxas, "UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation," arXiv preprint arXiv:2107.00781, 2021.

[17] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation," arXiv preprint arXiv:2102.08005, 2021.

[18] A. Hatamizadeh et al., "UNETR: Transformers for 3D Medical Image Segmentation," arXiv preprint arXiv:2103.10504, 2021.

[19] D. Tan, Y. Su, X. Peng, H. Cheng, C-H. Zheng, X. Zhang, and W. Zhong, "Large-scale data-driven optimization in deep modeling with an intelligent decision-making mechanism," *IEEE Transactions on Cybernetics,* vol. 54, pp. 1-13, 2023.

[20] Yiting Lu, Jun Fu, Xin Li, Wei Zhou, Sen Liu, Xinxin Zhang, Wei Wu, Congfu Jia, Ying Liu, Zhibo Chen,"RTN: Reinforced Transformer Network for Coronary CT Angiography Vessel-level Image Quality Assessment", *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2022,* pp. 644-653, 2022.

[21] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical Transformer: Gated Axial-Attention for Medical Image Segmentation," arXiv preprint arXiv:2102.10662, 2021.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," arXiv preprint arXiv:2103.14030, 2021.

[23] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation," arXiv preprint arXiv:2105.05537, 2021.

[24] D. Tan, Z. Yao, X. Peng, H. Ma, Y. Dai*, Y. Su*, and W. Zhong, "Multi-Level Medical Image Segmentation Network Based on Multi-Scale and Context Information Fusion Strategy," *IEEE Transactions on Emerging Topics in Computational Intelligence,* 88(1), 474-487, 2024.

[25] C. Li et al., "Attention Unet++: A Nested Attention-Aware U-Net for Liver CT Image Segmentation," in *2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates,* pp. 345-349, 2020.

[26] H. Y. Liu, X. J. Shen, F. H. Shang and F. Wang, "CU-Net: Cascaded U-Net with Loss Weighted Sampling for Brain Tumor Segmentation," arXiv preprint arXiv:1907.07677, 2019.

[27] K. Sirinukunwattana, J. P. W. Pluim, H. Chen, X. Qi, P.-A. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, A. Böhm, O. Ronneberger, B. B. Cheikh, D. Racoceanu, P. Kainz, M. Pfeiffer, M. Urschler, D. R. J. Snead, and N. M. Rajpoot, "Gland Segmentation in Colon Histology Images: The GlaS Challenge Contest," arXiv preprint arXiv:1603.00275, 2016.

[28] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology," *IEEE Transactions on Medical Imaging,* vol. 36, no. 7, pp. 1550-1560, 2017.

[29] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, C. Pal, and Y. Bengio, "The importance of skip connections in biomedical image segmentation," *Deep learning and data labeling for medical applications,* pp. 179-187, Springer, 2016.

[30] H. Huang, Y. Wang, W. Lian, W. Zheng, and Q. Tian, "U-Net 3+: A full-scale connected U-net for medical image segmentation," *Pattern Recognition,* vol. 106, p. 107404, 2020.

[31] R. K. S. Tan, Y. Lu, and L. Zhang, "MultiResUNet: Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation," *Neural Networks,* vol. 121, pp. 74-87, 2020..

[32] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "TransUnet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv preprint arXiv:2102.04306, 2021.

[33] B. Landman, Z. Xu, J. E. Igelsias, M. Styner, T. Langerak, and A. Klein, "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proc. MICCAI: Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge,* 2015.

[34] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al., "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?," *IEEE Transactions on Medical Imaging,* vol. 37, no. 11, pp. 2514–2525, 2018.

[35] "Covid-19 ct lung and infection segmentation dataset," Zenodo [Online]. Available: https://zenodo.org/record/3757476#.X1nqY4vhWUn, Accessed on: August. 12, 2020.

[36] "Hippocampus," ADNI [Online]. Available: https://adni.loni.usc.edu/data-samples/access-data/, Accessed on: April. 16, 2023.

[37] P. Bilic, P. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C. Fu, X. Han, P. Heng, J. Hesser et al., "The liver tumor segmentation benchmark (lits). arxiv 2019," arXiv preprint arXiv:1901.04056, 2019.