SPG-CDENet: Spatial Prior-Guided Cross Dual Encoder Network for Multi-Organ Segmentation

Xizhi Tian, Changjun Zhou*, and Yulin. Yang*

***Abstract*— Multi-organ segmentation is a critical task in computer-aided diagnosis. While recent deep learning methods have achieved remarkable success in image segmentation, huge variations in organ size and shape challenge their effectiveness in multi-organ segmentation. To address these challenges, we propose a Spatial Prior-Guided Cross Dual Encoder Network (SPG-CDENet), a novel two-stage segmentation paradigm designed to improve multi-organ segmentation accuracy. Our SPG-CDENet consists of two key components: an spatial prior network and a cross dual encoder network. The prior network generates coarse localization maps that delineate the approximate ROI, serving as spatial guidance for the dual encoder network. The cross dual encoder network comprises four essential components: a global encoder, an local encoder, a symmetric cross-attention module, and a flow-based decoder. The global encoder captures global semantic features from the entire image, while the local encoder focuses on features from the prior network. To enhance the interaction between the global and local encoders, a symmetric cross-attention module is proposed across all layers of the encoders to fuse and refine features. Furthermore, the flow-based decoder directly propagates high-level semantic features from the final encoder layer to all decoder layers, maximizing feature preservation and utilization. Extensive qualitative and quantitative experiments on two public datasets—the Automated Cardiac Diagnosis Challenge and Synapse Multi-Organ CT—demonstrate the superior performance of SPG-CDENet compared to existing segmentation methods. Specifically, SPG-CDENet achieves a 95% Hausdorff Distance of 12.75(Synapse) and a Dice Similarity Coefficient of 94.25%, 85.97%. Ablation studies further validate the effectiveness of the proposed modules in improving segmentation accuracy.**

***Index Terms*— Multi-organ Segmentation; Spatial Prior Network; Cross Dual Encoder Network; Symmetric Cross-Attention Module;**

I. Introduction

MULTI-ORGAN segmentation is a foundational task in medical image analysis, focused on precisely delineating organ boundaries in medical images. It plays a critical role in computer-aided diagnosis, supporting applications such as disease detection, treatment planning, and surgical navigation. With the progress of deep learning, numerous models [8], [12], [17], [18], [21], [26], [28] have emerged to improve the accuracy and efficiency of multi-organ segmentation.

U-Net-based image segmentation models have garnered significant attention due to their exceptional performance across various tasks. These models can be broadly classified into two primary categories based on the approach used to extract semantic features: CNN-based methods (Fig. 1(a)) and Transformer-based methods (Fig. 1(b)). Classic CNN-based architectures [15], [19], [27] have demonstrated remarkable success in diverse image segmentation tasks. However, CNNs are inherently limited by their local receptive fields, which restrict their ability to capture long-range dependencies, often leading to suboptimal segmentation accuracy, particularly in complex or contextually rich images. To overcome these limitations, Transformer-based methods [12], [21], [24], [35] have been introduced for image segmentation. These models utilize self-attention mechanisms to effectively model global contextual information, thereby improving segmentation performance. For instance, MissFormer [12] incorporates a multi-scale information integration module to better capture features across various spatial scales and levels of detail. Despite their strengths in global context modeling, Transformer-based models often struggle with capturing fine-grained local details, which can lead to blurred object boundaries, particularly in segmentation tasks requiring high spatial precision [13].

Although U-Net and its variants have achieved convincing performance across a wide range of image segmentation tasks, their effectiveness in multi-organ segmentation remains limited due to the inherent complexity of abdominal medical imaging. First, unlike natural images, abdominal scans are typically grayscale with low contrast, leading to ambiguous boundaries between organs and the surrounding background, as well as between adjacent organs with similar intensity and texture. These visual ambiguities significantly hinder the model's ability to achieve precise boundary delineation. Second, abdominal organs exhibit substantial inter-patient variability in appearance, including differences in shape, size, and texture. The substantial inter-patient variability in organ appearance significantly impairs model generalization, may resulting in prediction inconsistency, and reduced robustness on anatomically atypical cases. To address the first challenge, some studies [16], [28] have designed more complex network architectures to increase the receptive field, enabling the capture of finer-grained se-

Xizhi Tian is with the School of Computer Science and Technology, Zhejiang Normal University, Jinhua, China. e-mail: tianxizhi0@gmail.com.

Changjun Zhou is with the School of Computer Science and Technology, Zhejiang Normal University, Jinhua, China. e-mail: zhouchangjun@zjnu.edu.cn.

Yulin Yang is with Metaverse and Artificial Intelligence Institute of Wenzhou University, Wenzhou, China. e-mail: yangyulin1991@gmail.com.

*Corresponding author.

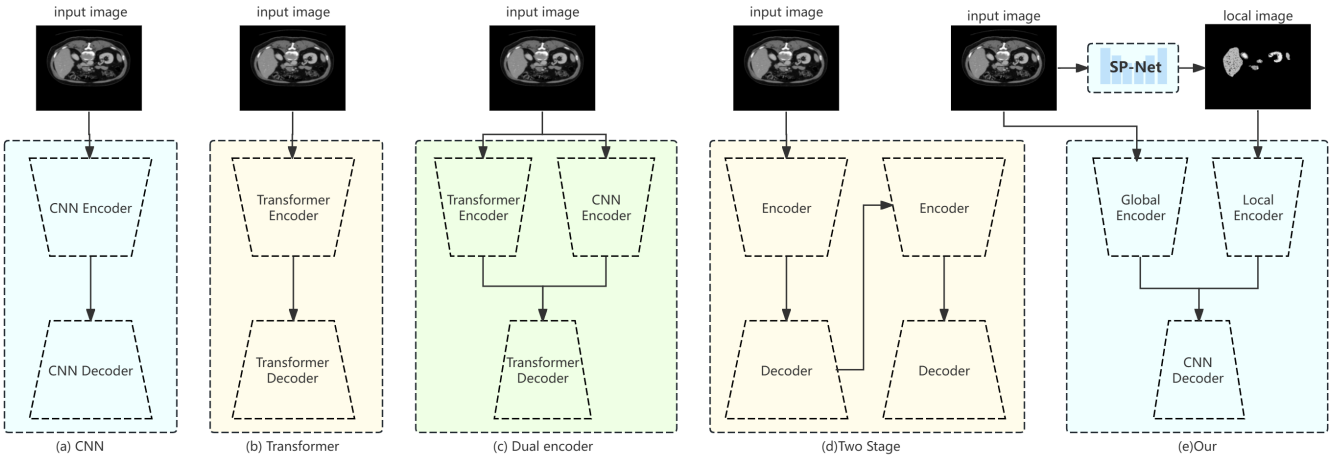

Fig. 1. Conceptual comparison of the three most popular models used for medical image segmentation, where (a) classical U-Net [26]; (b) Swin U-Net [21]; (c) ParaTransCNN encoder [28]; (d)Hierarchical FCN [9]; (e)Our SPG-CDENet

mantic features and thereby improving boundary segmentation accuracy. ParaTransCNN [28] incorporates an additional encoder path to improve the segmentation precision of complex anatomical boundaries, as shown in Fig.1 (C). Furthermore, to tackle the second challenge, several approaches [5], [11], [14] have introduced prior knowledge networks to localize organ regions in multi-organ segmentation tasks, achieving more precise segmentation results. For instance, PG-SAM [14] introduces fine-grained semantic priors generated by a medical large language model to guide the Segment Anything Model toward more accurate multi-organ segmentation in medical images. Meanwhile, DCL-Net [5] proposes a Dual-Contrastive Learning framework wherein Stage-1 produces a mask prior to guide Stage-2's organ-aware local refinement, as shown in Fig.1 (D).

Different from previous studies, we propose an Prior-Guided Cross Dual Encoder Network (SPG-CDENet) for more accurate multi-organ segmentation. As illustrated in Fig.1(e), SPG-CDENet includes two parts: a spatial prior network (SP-Net) and a crossing dual encoder network (CDE-Net). SP-Net employs a pre-trained multi-organ segmentation model followed by a post-processing module to localize region of interest(ROI) in the input image. These localized regions serve as spatial priors that guide the subsequent segmentation stage. To make fuller use of above localized regions, we device a crossing dual encoder network. CDE-Net consists of four parts: a global encoder, a local encoder,a symmetric cross-attention module and a flow-based decoder. The global encoder is responsible for extracting holistic features, while the local encoder focuses on the local features produced by the prior network. The symmetric cross-attention module integrates global and local features. The flow-based decoder propagates high-level semantic information to each layer of the decoder. Our contributions are summarized as follows:

- We propose a Prior-Guided Cross Dual Encoder Network, which consists of a spatial prior network and a crossing dual-encoder architecture. It can address the low boundary accuracy caused by background-interest similarity and the limited generalization due to large inter-interest variability. The effectiveness of our proposed model is validated through both qualitative and quantitative experiments.
- We design a spatial prior network that utilizes a pre-trained model to obtain the prior location information. Experimental results demonstrate that our model exhibits strong robustness and generalization across different pre-trained models.
- To fully exploit the location prior map, we design a crossing dual-encoder architecture consisting of four modules: a global encoder, a local encoder, a symmetric cross-attention module and a flow-based decoder. Ablation studies demonstrate that the proposed crossing dual-encoder design significantly enhances the segmentation performance of the model. Furthermore, we validate the effectiveness of the proposed symmetric cross-attention module.

II. Relation Work

A. Multi-Organ Segmentation Models

Multi-organ segmentation is challenging due to anatomical variability in organ shape, size, and position. Recent models based on CNNs, Transformers, and their hybrids have shown promising results. CNN-based methods [15], [19], [26], [27] excel at local feature extraction. For example, Alom et al. [15] proposed R2U-Net, which innovatively incorporates residual connections to alleviate the vanishing gradient problem in deep networks and enhance feature propagation. However, CNN-based models still face inherent limitations in modeling long-range dependencies, while transformers are effective for long-range context modeling. For example, Liu et al. [20] proposed the Swin Transformer, which combines local window-based self-attention to significantly reduce computational complexity while preserving strong global modeling capabilities, and

MISSFormer [12] enhances local detail with multi-scale fusion. However, Transformer-based models often struggle to capture fine-grained local details, which are critical for precise boundary delineation in medical image segmentation. Hybrid models combine the strengths of CNNs and Transformers. TransUNet [24] integrates CNN features with Transformer context. ATTransUNet [3] further improves representation via adaptive attention. Swin-Unet [21] extends the Swin Transformer into a U-shaped architecture tailored for medical image segmentation, enabling hierarchical feature extraction with enhanced global-local contextual understanding.

B. Architecture-Driven Segmentation

Although U-Net variants have demonstrated persuasive performance in medical image segmentation, recent studies [4], [7], [28] suggest that more sophisticated network architectures can better capture complex semantic features to improve the accuracy of segmentation. For example, Zhang et al. [7] introduced G-CASCADE, which enhances the decoder with cascaded graph convolutional layers to improve the model's boundary modeling capabilities. Sun et al. [28] introduced ParaTransCNN, a parallel CNN-Transformer architecture that adds an auxiliary encoder for capturing fine-grained details, achieving more accurate segmentation in complex anatomical regions with subtle boundary variations. Xie et al. [4] proposed CoTr, a dual-encoder framework that integrates a CNN encoder with a deformable Transformer encoder. By introducing a DeTrans module, the model focuses attention on key spatial regions, effectively capturing global context for improved multi-organ segmentation. These studies collectively demonstrate that more sophisticated network architectures can effectively enhance model generalization and improve the accuracy of organ boundary segmentation. Despite architectural innovations, existing methods generally lack explicit modeling of organ spatial priors, limiting their ability to exploit the predictable anatomical layout of organs. To overcome this limitation, we propose a crossing dual-encoder architecture that incorporates a prior map generated in the first stage to guide the local encoder, thereby enhancing feature extraction and improving segmentation accuracy.

C. Anatomical Prior-Guided Segmentation

Compared to natural images, organs in medical images often follow predictable anatomical patterns, such as consistent shape and spatial arrangement. Consequently, some studies have explored the use of prior knowledge to capture the spatial distribution of target structures, aiming to improve segmentation accuracy. For example, Huang et al. [11] proposed Deep Atlas Prior, which for the first time integrates anatomical priors into multi-organ segmentation via a dedicated prior-aware loss function. Their approach improves boundary precision and spatial consistency across organs. Wen et al. [5] introduced DCL-Net, a dual-stage contrastive learning framework where the first stage generates a coarse mask that serves as a spatial prior to guide the second stage's organ-aware feature refinement. This design significantly enhances structure-specific feature learning and improves segmentation accuracy, particularly in low-contrast regions. Zhong et al. [14] proposed PG-SAM, a framework that integrates fine-grained semantic priors extracted by a medical-domain large language model to enhance the Segment Anything Model (SAM) for multi-organ segmentation. PG-SAM demonstrates strong generalization across unseen anatomical structures and imaging modalities, achieving state-of-the-art performance on multiple benchmarks. Jiang et al. [25] proposed a two-level cascading U-Net, which uses a U-Net variant for the coarse segmentation in the first stage, then concatenates the result with the original input to utilize automatic context for fine-tuning in the second stage. These studies highlight the promising potential of incorporating prior knowledge into multi-organ segmentation frameworks. Inspired by these advancements, we propose an SP-Net, which introduces a Plug-and-Play pre-trained segmentation network to produce the interest location priors.

III. Method

Given a grayscale medical image $x \in \mathbb{R}^{H\times W\times 1}$, where H and W denote the image height and width, our proposed SPG-CDENet enhances multi-organ segmentation by explicitly incorporating prior anatomical cues. As shown in Fig. 2, SPG-CDENet consists of two components: a Spatial Prior Network (SP-Net) and a Crossing Dual-Encoder Network (CDE-Net). SP-Net takes x as input and generates a coarse spatial prior $\hat{M}_p$ to highlight potential the ROI. CDE-Net then uses both x and $\hat{M}_p$ to predict the final segmentation mask. CDE-Net includes two parallel encoders: a global encoder for capturing semantic features from the entire image and a local encoder focused on region-specific features guided by $\hat{M}_p$. To facilitate feature interaction between the two branches, we introduce a symmetric cross-attention mechanism, enabling effective fusion of global context and local details. The following sections detail the architecture of SPG-CDENet, including SP-Net, CDE-Net, and the proposed symmetric cross-attention strategy.

A. Spatial Prior Network

Multi-organ segmentation remains a challenging task due to the complex anatomical structures of the human body. Organs exhibit significant variability in size and shape across different organs. Furthermore, the low contrast between adjacent anatomical structures—including both organs and surrounding tissues—introduces ambiguity, especially in the segmentation of small or poorly defined organs. These factors often result in inaccurate localization and blurred boundaries in conventional segmentation models. To overcome these limitations, we propose SP-Net, a novel approach that generates a coarse segmentation mask to represent prior spatial information of the ROI. As shown in Fig. 2, SP-Net comprises two main components: a pre-trained multi-organ segmentation model and a post-processing module. The segmentation model takes the entire image x as input and generates a multi-class segmentation mask, where each label corresponds to a specific organ. The post-processing module then binarizes this mask to distinguish the ROI from the background. The resulting binary mask

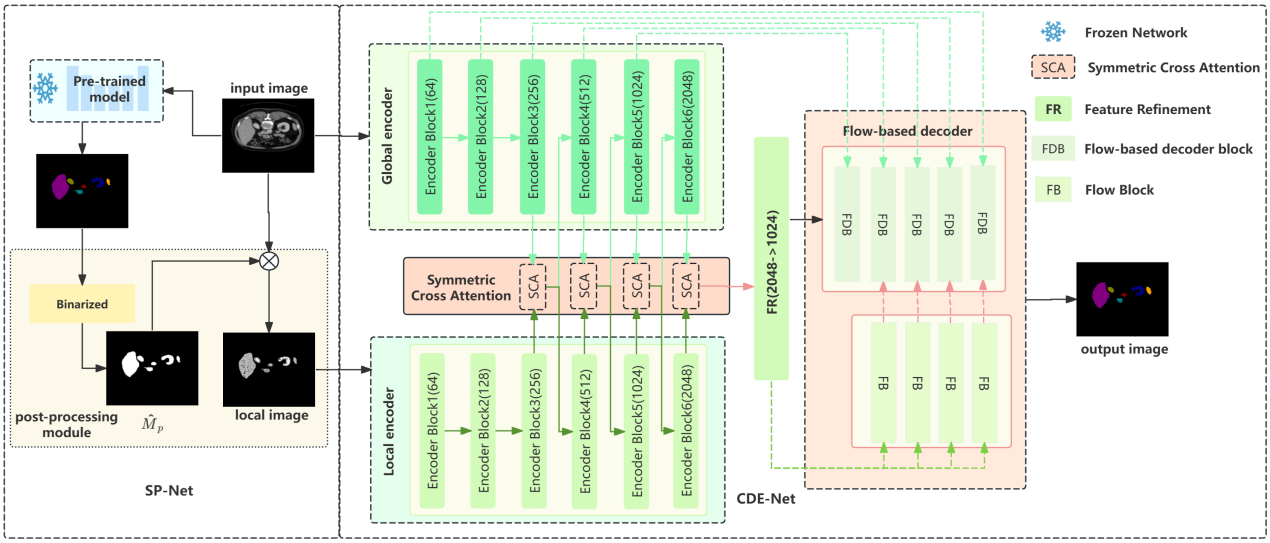

Fig. 2. Overview of the SPG-CDENet

serves as a spatial prior for the subsequent segmentation stage, guiding the model to focus on anatomically relevant regions and mitigating interference caused by inter-class ambiguity. This process can be formulated as follows:

$$x_l = x \times \mathbb{B}(PTM(x) > \tau) \tag{1}$$

where τ is a fixed threshold set to 0. $PTM(\cdot)$ denotes the pre-trained segmentation model, and $\mathbb{B}(\cdot)$ represents the binarization operation. x_l is the output of the SP-Net, which provides spatial guidance for the downstream segmentation model.

B. Crossing Dual Encoder-based Segmentor

As illustrated in the Fig. 2, the proposed CDE-Net takes the input and local images as input and try to predict a mask about different organ. Proposed consists of three parts: dual encoder network, a flow-based decoder, and a symmetric crossing attention module. To be specific, we device a dual-encoder, the global encoder takes the original grayscale image as input, while the local encoder processes local image generated in the SP-Net. To strengthen the feature interaction between the two encoders, we proposed the symmetric cross-attention mechanism to integrate the features between the two encoders. For better global semantic guidance during decoding, we proposed the flow-based decoder, which propagate the final-layer features from the encoder to each layer of the decoder. This design allows the decoder to access high-level contextual information at all stages, thereby improving the segmentation of small or ambiguous structures. In the following, we provide a detailed description of the three key components of our CDE-Net.

1) Dual Encoder Network: As shown in Fig. 2, we employ two parallel networks as encoders: one processes the entire image to extract global semantic features, while the other focuses on the local image to extract fine-grained local features. Each encoder follows the layered ResNet-50 structure, producing hierarchical features with progressively decreasing spatial resolution and increasing channel dimensions. The two encoders share an identical architecture. Taking the global encoder as an example, the output of each layer can be represented as :

$$F_{Enc}^{GE(i)} = \begin{cases} \mathrm{Res}_i(x), & i = 0 \\ \mathrm{Res}_i\left(F_{Enc}^{GE(i-1)}\right), & i = 1,2,3,4,5 \end{cases} \tag{2}$$

Where $Res_i(\cdot)$ represents the i-th layer of Resnet-50 Block. The local encoder shares the same architecture as the global encoder, with the only difference being that it takes the local image as input to the first layer. The feature maps from the final layers of the dual encoders are concatenated and then passed through a feature refinement (FR) module, which performs feature compression and channel adjustment. This process can be formalized as:

$$F_{global} = FR(Cat[F_{Enc}^{GE(5)}, F_{Enc}^{LE(5)}]) \tag{3}$$

FR module consists of two consecutive Conv Blocks with identical structures, as illustrated in Fig. 3. Each Conv Block is composed of a 3×3 convolution followed by a BatchNorm2d layer and a ReLU activation. We denote the output of the FR module as $F_{global} \in \mathbb{R}^{H \times W \times C}$, which serves as the high-level semantic representation for the decoder.

2) Symmetric Cross-Attention Module: As shown in Fig. 2, we apply the Symmetric Cross-Attention (SCA) module between the third and sixth layers of the global and local encoders to facilitate effective communication between them. The architecture of proposed SCA module, depicted in Fig. 3, contains two branches: global attention branch and local attention branch. In the global attention branch, a multi-head attention is applied to allow the local encoder to be guided by global context. Specifically, semantic features from the global encoder are used as queries (Q), while the semantic features from the local encoder serve as keys (K) and values (V). The output of the multi-head attention is then combined with

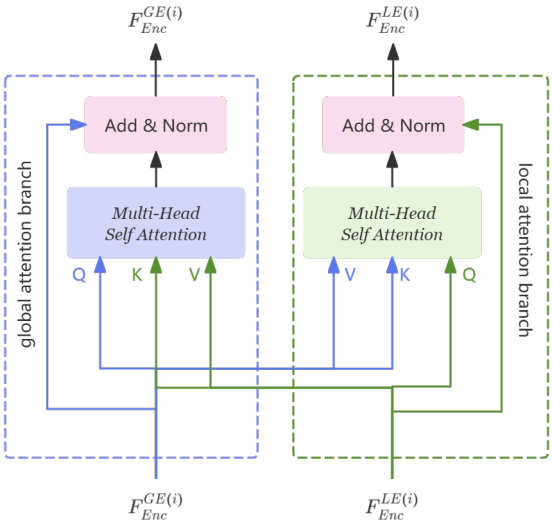

Fig. 3. Symmetric Cross-Attention Module

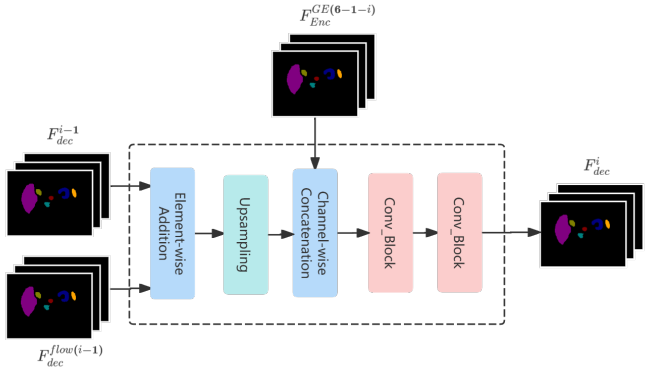

Fig. 4. Overview of the Flow-based Decoder. The decoder takes three inputs — semantic features from the previous decoder layer $\boldsymbol{F_{Dec}^{i-1}}$, the encoder $\boldsymbol{F_{Enc}^{GE(6-1-i)}}$, and the Flow Block layer $\boldsymbol{F_{Dec}^{flow(i-1)}}$, respectively.

the original semantic features through a residual connection, yielding the final result. Taking the i-th layer as an example, this process can be expressed as:

$$Q = F_{Enc}^{GE(i)} W^{QG}, K = F_{Enc}^{LE(i)} W^{KG}, V = F_{Enc}^{LE(i)} W^{VG}, \tag{4}$$

$$F_{Enc}^{GE(i+1)} = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right) V + F_{Enc}^{GE(i)} \tag{5}$$

where W^{QG}, W^{KG}, W^{VG} are learnable parameters in the multi-head attention mechanism of the global attention branch.

The local attention branch follows same architecture as the global attention branch. However, in the local attention, semantic features captured from the local encoder are used as queries (Q), while semantic features captured from the global encoder serve as keys (K) and values (V). This process can be formally expressed as:

$$Q = F_{Enc}^{LE(i)} W^{QL}, K = F_{Enc}^{GE(i)} W^{KL}, V = F_{Enc}^{GE(i)} W^{VL}, \tag{6}$$

$$F_{Enc}^{LE(i+1)} = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right) V + F_{Enc}^{LE(i)} \tag{7}$$

where W^{QL}, W^{KL}, W^{VL} are learnable parameters in local-guided attention branch.

3) Flow-Based Decoder: To enhance the global semantic guidance during decoding, we design a flow-based decoder as shown in the Fig. 2, which comprises a backbone decoder and a flow module. The flow module delivers multi-scale high-level semantic features from F_{global} to each decoding layer, with a upsampling layer and two consecutive Conv Blocks, thereby reinforcing semantic consistency during reconstruction. The detailed implementation of flow module is described as follows:

$$\{F_{Dec}^{flow(i)}\}_{i=1}^{4} = FB_{s_i}(F_{global}, c_i) \tag{8}$$

where $FB_{s_i}(\cdot)$ represents the flow module at the i-th level. c_i and s_i represent the number of output channels and the upsampling factor at that level, respectively. Specifically, the configuration is defined as $(c_i, s_i) \in \{(1024, 2), (512, 4), (256, 8), (128, 16)\}$.

After generating the multi-scale flow features, we incorporate them into the decoding pathway to construct the complete flow-based decoder. Except for the first layer, as shown in the Fig.4, each decoding layer F_{Dec}^{i} integrates three types of input features: the output from the previous decoding layer F_{Dec}^{i-1}, the skip connection feature from the global encoder $F_{Enc}^{GE(6-1-i)}$, and the flow feature at the corresponding scale $F_{Dec}^{flow(i-1)}$. The overall decoding process can be formulated as follows:

$$F_{Dec}^{i} = \begin{cases} \text{FDB}(F_{global}, F_{Enc}^{GE(6-1-i)}), & i = 1 \\ \text{FDB}(F_{Dec}^{i-1}, F_{Enc}^{GE(6-1-i)} + F_{Dec}^{flow(i-1)}), & i = 2, 3, 4, 5 \end{cases} \tag{9}$$

where $FDB(\cdot)$ represents flow-based decoder block. Semantic features F_{Dec}^{5} is fed into an output layer to generate the final segmentation mask. The output layer consists of two Conv Blocks followed by a sigmoid activation function. This process can be formulated as

$$\hat{M} = \sigma(OL(F_{Dec}^{5}))), \tag{10}$$

where OL is the output layer. $\hat{M}$ denotes the final predicted mask generated by our model.

C. Loss Function

For training our SPG-CDENet network, we employ a combination of Dice loss and Cross Entropy as the loss function, defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{dice}}(M, \hat{M}) + \lambda_2 \mathcal{L}_{\text{ce}}(M, \hat{M}) \tag{11}$$

here $\mathcal{L}$ is the total loss, with λ_1 and λ_2 as the weight coefficients, with respective set to 0.4 and 0.6 based on experimental results. And M represent the ground-truth segmentation mask provided by expert annotations.

TABLE I
COMPARISON ON THE SYNAPSE MULTI-ORGAN CT DATASET (AVERAGE DSC SCORE % AND AVERAGE HAUSDORFF DISTANCE IN MM, WITH HIGHER DSC AND LOWER HD BEING PREFERRED, AND THE SPATIAL PRIOR NETWORK OF SPG-CDENET (OURS) IS U-NET)

Methods	DSC(%)↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
V-Net [29]	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR [15]	69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50 U-Net [24]	74.68	36.87	87.47	66.36	80.60	78.19	93.74	56.90	85.87	74.16
U-Net [26]	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
R50 Att-UNet [24]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
Att-UNet [2]	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
R50 ViT [24]	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUNet [24]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
TransNorm [30]	78.40	30.25	86.23	65.10	82.18	78.63	94.22	55.34	89.50	76.01
Swin U-Net [21]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
TransDeepLab [23]	80.16	21.25	86.04	69.16	84.08	79.88	93.53	61.19	89.00	78.40
HiFormer [10]	80.39	14.70	86.21	65.69	85.23	79.77	94.61	59.52	90.99	81.08
MISSFormer [12]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
TransCeption [6]	82.24	20.89	87.60	71.82	86.23	80.29	95.01	65.27	91.68	80.02
DAE-Former [32]	82.63	16.39	87.84	71.65	87.66	82.39	95.08	63.93	91.82	80.77
FCT [22]	83.53	-	**89.85**	72.73	88.45	**86.60**	**95.62**	66.25	89.77	79.42
ParaTransCNN [28]	83.86	15.86	88.12	68.97	87.99	83.84	95.01	69.79	**92.71**	84.43
AHGNN [33]	84.03	13.26	89.27	74.53	86.99	83.49	95.03	69.89	92.38	84.86
SPG-CDENet (Ours)	**85.97**	**12.75**	87.03	**75.36**	**89.77**	86.03	94.72	**72.78**	92.69	**89.40**

IV. EXPERIMENTAL

A. Experimental Settings

1) Dataset: We evaluate our proposed method on two widely-used benchmarks, the Synapse multi-organ CT dataset from the MICCAI 2015 Multi-Atlas Labeling Beyond the Cranial Vault Challenge (BTCV) and the ACDC cardiac MRI dataset [31], followed by a detailed description of their data splits and evaluation protocols. **The Synapse multi-organ segmentation dataset** comprises 30 abdominal CT scans with a total of 3,779 axial slices. Following the widely adopted experimental protocol, we utilize 18 cases for training and 12 cases for testing, consistenting with the previous method [12], [21], [24]. The evaluation is conducted on eight abdominal organs, including the aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, and stomach. **The ACDC dataset** consists of 100 short-axis cardiac MR images collected from different patients under breath-hold conditions. Each image is manually annotated for three cardiac structures: the left ventricle (LV), right ventricle (RV), and myocardium (MYO). In line with previous works [12], [21], [24], we use 70 images for training, 10 for validation, and 20 for testing.

2) Evaluation Metrics: To quantitatively assess the segmentation performance of our method, we adopt the average Dice Similarity Coefficient (DSC) and the average Hausdorff Distance (HD) as evaluation metrics. The DSC measures the overlap ratio between the predicted segmentation and the ground truth, while the HD evaluates the boundary agreement between them. Higher DSC and lower HD indicate better segmentation accuracy and boundary consistency.

3) Implement Details: All experiments were conducted on a single NVIDIA RTX 4090 GPU, with network parameters optimized using the Stochastic Gradient Descent (SGD) optimizer. The input image and local image size was set to 224×224, the initial learning rate is 5e-3, the momentum 0.9, and the weight decay 1e-4. In all experiments, we applied basic data augmentation techniques, including random rotations and flips. It is worth noting that, we apply consistent data augmentation to both global and local images for enhancing the model's robustness to local image variations.

4) Comparison of experimental models: We compare SPG-CDENet with CNN-based segmentation networks like R50 U-Net [24], U-Net [26], Att-UNet [2], pure Transformer segmentation network systems such as Swin U-Net [21], TransDeepLap [23], MISSFormer [12], DAE-Former [32], the mixed CNN and Transformer models like TransUNet [24], HiFormer [10], FCT [22], TransCeption [6], and dual encoder models like ParaTransCNN [28].

B. Comparison with state-of-the-art methods

We compare SPG-CDENet with the state-of-the-art models on the Synapse dataset and ACDC dataset, quantitative results are shown in Fig. 5 and Fig. 6, while qualitative results are presented in Tables 1 and 2.

1) Results on Synapse Multi-Organ Segmentation: In Table 1, our SPG-CDENet uses UNet as the SP-Net. Our SPG-CDENet achieves the best overall performance, attaining a DSC of 85.97%, which surpasses the second-best method by 1.94%, and ranking first in terms of the HD with a score of 12.75, which surpasses the second-best method by 0.51. This demonstrates the effectiveness of our approach. Our model demonstrates significant improvements in organs such as the gallbladder, kidney(L), pancreas, and stomach, with increases of 0.83%, 1.32%, 2.89%, and 4.54%, respectively, compared to the second-best model. Additionally, the visualization of segmentation results is shown in Fig. 5. Our model achieves superior overall performance and more precise local segmentation on the Synapse dataset. In particular, it produces accurate and complete delineations of small organs, such as the pancreas and stomach, which further confirms the consistency between qualitative and quantitative assessments. Furthermore, compared to the U-Net model, our approach achieves improvements in every organ except for the Aorta, which enable the model to capture both organ-specific positional priors and image-driven appearance features. This result demonstrates the

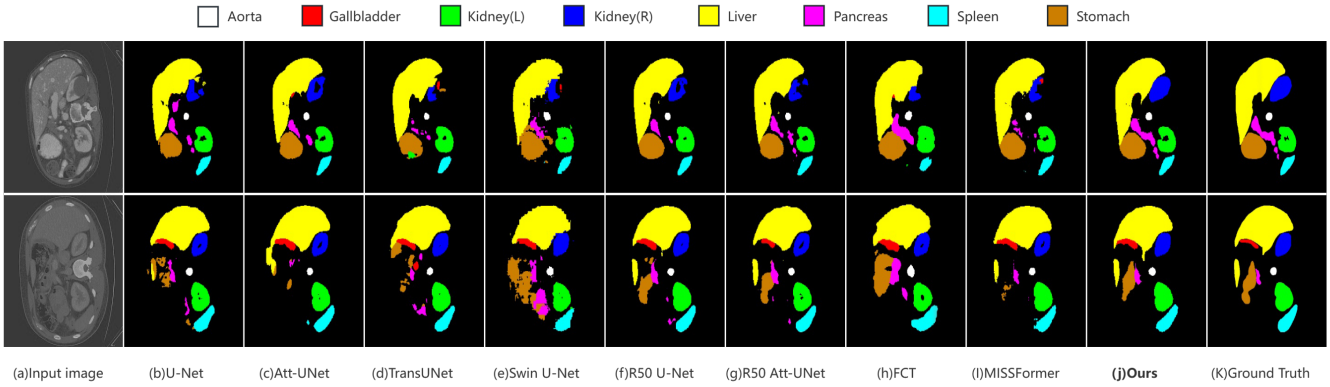

Fig. 5. Qualitative results of different models on the Synapse dataset, from left to right: Input image; U-Net [26]; Att-UNet [2]; TransUNet [24]; Swin U-Net [21]; R50 U-Net [24]; R50 Att-UNet [24]; FCT [22]; MISSFormer [12]; SPG-CDENet (Ours); Ground Truth.

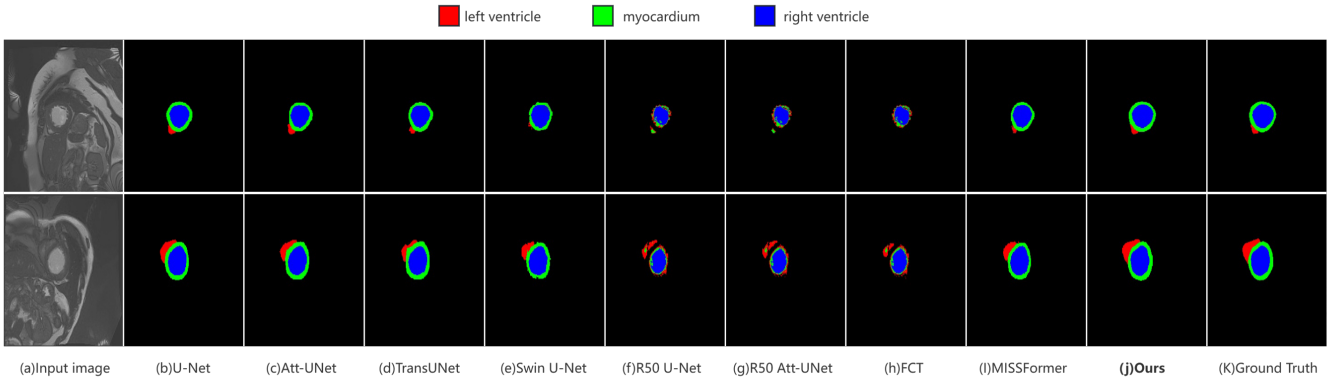

Fig. 6. Qualitative results of different models on the ACDC dataset, from left to right: Input image; U-Net [26]; Att-UNet [2]; TransUNet [24]; Swin U-Net [21]; R50 U-Net [24]; R50 Att-UNet [24]; FCT [22]; MISSFormer [12]; SPG-CDENet (Ours); Ground Truth.

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE ACDC DATASET (AVERAGE DSC SCORE % , WITH HIGHER DSC BEING PREFERRED, AND THE SPATIAL PRIOR NETWORK OF SPG-CDENET (OURS) IS U-NET)

Methods	DSC(%)↑	RV	Myo	LV
R50-U-Net [24]	87.55	87.10	80.63	94.92
U-Net [26]	87.79	85.48	82.90	94.97
R50-AttnUNet [24]	86.75	87.58	79.20	93.47
Att-UNet [2]	89.58	87.89	85.56	95.29
TransUNet [24]	89.71	88.86	84.53	95.73
Swin U-Net [21]	90.00	88.55	85.62	95.83
MISSFormer [12]	90.86	89.55	88.04	94.99
ParaTransCNN [28]	91.31	**92.76**	85.84	95.34
AHGNN [33]	92.02	-	-	-
FCT [22]	92.84	92.02	90.61	95.89
SPG-CDENet (Ours)	**94.25**	92.61	**91.74**	**98.40**

effectiveness of our method by integrating SP-Net with a cross dual-encoder architecture.

2) Results on ACDC Segmentation: We also evaluated our method on the ACDC dataset, and as shown in the Table II, SPG-CDENet achieves the highest segmentation accuracy with a DSC of 94.25%, surpassing the second-best model by 1.41%. Compared with the U-Net (SP-Net), our model achieves consistent improvements across all cardiac substructures on the ACDC dataset, demonstrating the overall effectiveness and strong generalization ability of the proposed SPG-CDENet framework. The visual results, as shown in Fig. 6, demonstrate that SPG-CDENet produces the smoothest boundaries and the most complete segmentations, highlighting the model's strong generalizability and robustness.

C. Ablation Studies On Components

To investigate the impact of the symmetric cross-attention module, the dual-encoder module and the spatial prior network on the performance of our model, we conducted structural ablation experiments on Synapse dataset and ACDC dataset. Based on the full SPG-CDENet model, We designed three ablation variants: Model-1 serves as the baseline, consisting of a ResNet-50 encoder combined with a flow-based decoder. Model-2 incorporates the spatial prior network and a local encoder dedicated to processing the prior map, where the features from the two encoders remain independent without interaction. Finally, SPG-CDENet enhances this design by integrating a symmetric cross-attention module to fuse the features extracted from the two encoders at each layer.

TABLE III
ABLATION STUDIES ON COMPONENTS(SYNAPSE). SP-NET INDICATES THE SPATIAL PRIOR NETWORK, IMPLEMENTED USING U-NET. SCA INDICATES THE SYMMETRIC CROSS-ATTENTION MODULE. ✓INDICATES THE COMPONENT IS INCLUDED.

Model	SP-Net	SCA	DSC(%)↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
Model-1			80.09	21.11	85.74	58.73	81.53	77.36	94.16	64.07	92.55	86.56
Model-2	✓		82.47	21.80	81.59	76.90	83.14	79.94	94.20	65.04	92.22	86.74
SPG-CDENet	✓	✓	**85.97**	**12.75**	**87.03**	**75.36**	**89.77**	**86.03**	**94.72**	**72.78**	**92.69**	**89.40**

TABLE IV
FUSION TYPE ABLATION STUDY ON SYNAPSE (THE SPATIAL PRIOR NETWORK IS U-NET)

Fusion Type	DSC(%)↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
None	82.47	21.80	81.59	76.90	83.14	79.94	94.20	65.04	92.22	86.74
Concat	82.92	27.78	80.11	77.38	79.84	79.63	93.59	75.03	90.57	87.18
Cross-attention	83.17	18.39	86.34	72.88	85.62	79.33	95.06	69.47	92.39	84.26
SPG-CDENet	**85.97**	**12.75**	**87.03**	**75.36**	**89.77**	**86.03**	**94.72**	**72.78**	**92.69**	**89.40**

TABLE V
ABLATION STUDIES ON COMPONENTS(ACDC). SP-NET INDICATES THE SPATIAL PRIOR NETWORK, IMPLEMENTED USING U-NET. SCA INDICATES THE SYMMETRIC CROSS-ATTENTION MODULE. ✓INDICATES THE COMPONENT IS INCLUDED.

Model	SP-Net	SCA	DSC(%)↑	RV	Myo	LV
Model-1			90.37	89.02	85.51	96.58
Model-2	✓		91.04	90.01	86.65	96.47
SPG-CDENet	✓	✓	**94.25**	**92.61**	**91.74**	**98.40**

TABLE VI
FUSION TYPE ABLATION STUDY ON ACDC (THE SPATIAL PRIOR NETWORK IS U-NET)

Fusion Type	DSC(%)↑	RV	Myo	LV
None	91.04	90.01	86.65	96.47
Concat	91.64	90.46	87.57	96.89
Cross-attention	92.53	91.77	88.70	97.13
SPG-CDENet	**94.25**	**92.61**	**91.74**	**98.40**

As presented in Table III and Table V, on both the Synapse and ACDC datasets, Model-1 achieved DSC scores of 80.09% and 90.37%, surpassing the performance of most CNN-based models. Adding the SP-Net and Local Encoder substantially improves the segmentation performance, confirming the effectiveness of anatomical prior guidance. Further integrating the two encoders through cross-attention fusion enables the full SPG-CDENet to fully exploit complementary features, yielding an additional 3.5% DSC improvement and a 9.05 HD reduction on the Synapse dataset, along with consistent gains on the ACDC dataset. Notably, while SP-Net and the Local Encoder provide limited performance boost when used independently, but combining them with the SCA module results in a substantial performance improvement. This demonstrates the critical role of SCA module in fusing both global and local features effectively.

D. Ablation Study On Fusion Type

To explore how different interaction methods between the global and local encoders affect model performance, we conducted an ablation study on various fusion types. We evaluate four variants, including **None**, **Concat**, **Cross Attention [34]**, and **SCA modual**. When the fusion type is set to **None**, the two encoder streams remain independent without any feature interaction. This configuration corresponds to Model-2 in Section IV.C (Ablation Studies on Components) of the experimental results. For the **Concat** variant, the feature maps are concatenated along the channel dimension and subsequently compressed by a 1×1 convolution to form the fused representation. Under the **Cross Attention** setting, two encoder uses cross-attention to fuse features from two inputs, enhancing one feature map by attending to the other. SPG-CDENet refers to our full model that utilizes the SCA module to facilitate interaction between the two encoder streams.

The results are shown in Table IV and Table VI. It can be observed that under the **None** setting, the two encoder streams remain completely independent, resulting in the lowest average DSC and a relatively high HD, indicating that the lack of feature interaction limits model performance. Both concatenation and cross-attention fusion improve the overall segmentation performance, with the latter achieving more consistent and significant gains across all organs. For this phenomenon, we hypothesize that the possible reason is that concatenating features from the two encoders introduces redundant and misaligned information due to their differing segmentation scopes. Without effective interaction modeling, this leads to unstable boundary predictions and consequently a higher HD. Our proposed Symmetric Cross-Attention achieves the best segmentation results in both DSC and HD. It improves the performance for every organ, and also outperforms other fusion types, demonstrating the effectiveness of Symmetric Cross-Attention. Moreover, it has demonstrated strong performance on both the Synapse and ACDC datasets, further proving the robustness of Symmetric Cross-Attention.

E. Ablation Study On Spatial Prior Network

To investigate the impact of different SP-Net backbone models on overall network performance, we replaced the Spatial Prior Network in SPG-CDENet with three representative segmentation models: UNet (a CNN-based architecture), AttU-Net (a CNN with attention mechanisms), and TransUNet (a hybrid CNN and Transformer architecture). This experiment was conducted using a pre-trained U-Net model as the initial backbone for training. Then, we directly swapped the pre-trained U-Net with AttU-Net and TransUNet without further

TABLE VII
ABLATION STUDY ON SPATIAL PRIOR NETWORK(SYNAPSE)

Model	DSC(%)↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
U-Net [26]	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
our + U-Net	85.97	12.75	87.03	75.36	89.77	86.03	94.72	72.78	92.69	89.40
	(↑9.12)	(↓26.95)		(↑5.64)	(↑12.00)	(↑17.43)	(↑1.29)	(↑18.80)	(↑6.02)	(↑13.82)
Att-UNet [2]	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
our + AttU-Net	85.91	13.23	86.70	75.24	89.84	85.88	94.72	73.03	92.59	89.27
	(↑8.14)	(↓22.79)		(↑6.36)	(↑11.86)	(↑14.77)	(↑1.15)	(↑14.99)	(↑5.29)	(↑13.52)
TransUNet [24]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
our + TransUNet	86.02	12.63	86.75	75.53	89.76	86.12	94.80	72.70	92.89	89.63
	(↑8.54)	(↓19.06)		(↑12.40)	(↑7.89)	(↑9.10)	(↑0.72)	(↑16.84)	(↑7.81)	(↑14.01)

TABLE VIII
ABLATION STUDY ON SPATIAL PRIOR NETWORK(ACDC)

Methods	DSC(%)↑	RV	Myo	LV
U-Net [26]	87.79	85.48	82.90	94.97
our + U-Net	**94.25**	92.61	**91.74**	**98.40**
	(↑6.46)	(↑7.13)	(↑8.84)	(↑3.43)
Att-UNet [2]	89.58	87.89	85.56	95.29
our + AttU-Net	93.63	91.09	91.59	98.20
	(↑4.05)	(↑3.20)	(↑6.03)	(↑2.91)
TransUNet [24]	89.71	88.86	84.53	95.73
our + TransUNet	93.86	91.83	91.66	98.11
	(↑4.15)	(↑2.97)	(↑7.13)	(↑2.38)

training, and the results were still impressive. The results are shown in Table VII and Table VIII.

On the Synapse dataset, our proposed model achieves substantial improvements when using ROI prior information generated by the U-Net model. Specifically, the average DSC improves by 9.12%, while the average HD decreases by 26.95. This demonstrates that incorporating ROI prior information significantly enhances segmentation performance. Furthermore, when the pre-trained model was replaced with AttU-Net and TransUNet—without further training—the performance remained excellent, the average DSC improves by 8.14% and 8.54%, and the average HD decreases by 22.79 and 19.06, highlighting the robustness and plug-and-play capability of our approach.

On the ACDC dataset, consistent improvements are also observed across all three SP-Net variants. The average DSC increases by 6.46%, 4.05%, and 4.15%, respectively. This further validates that our model is highly effective and stable across different datasets. Overall, the proposed method consistently outperforms baselines across diverse SP-Nets and datasets, confirming its effectiveness, adaptability, and reliability.

V. CONCLUSION

In this work, we propose a novel SPG-CDENet that integrates a Spatial Prior Network and a crossing dual-encoder network. Experiments on the Synapse and ACDC datasets demonstrate that our model achieves superior qualitative and quantitative performance. Furthermore, we analyze the impact of prior networks with different architectures and performance levels on the final segmentation results. Experimental results demonstrate that our prior network shows strong generalization and plug-and-play capability, enabling easy replacement of different prior models without performance degradation.To further validate our design, we conduct an ablation study on various interaction strategies between the global and local encoders. The results show that the proposed SCA module achieves the best performance, confirming its effectiveness in leveraging prior information for segmentation. In future work, we plan to extend our method to 3D medical image segmentation tasks to further enhance its generalization and applicability.

REFERENCES

[1] Zhao F, Tang Y, Lu L, et al. "A Curvature-Guided Coarse-to-Fine Framework for Enhanced Whole Brain Segmentation[C]," International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, pp.13-22, 2024.

[2] Oktay O, Schlemper J, Folgoc L L, et al. "Attention u-net: Learning where to look for the pancreas[J]," arXiv preprint arXiv:1804.03999, 2018.

[3] Li X, Pang S, Zhang R, et al. "Attransunet: An enhanced hybrid transformer architecture for ultrasound and histopathology image segmentation[J]," Computers in Biology and Medicine, vol. 152, pp. 106365, 2023.

[4] Xie Y, Zhang J, Shen C, et al. "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation[C]," International conference on medical image computing and computer-assisted intervention. Cham: Springer International Publishing, pp.171-180, 2021.

[5] Wen L, Feng Z, Hou Y, et al. "Dcl-net: Dual contrastive learning network for semi-supervised multi-organ segmentation[C]," ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp.1876-1880, 2024.

[6] Azad R, Jia Y, Aghdam E K, et al. "Enhancing medical image segmentation with transception: A multi-scale feature fusion approach[J]," arXiv preprint arXiv:2301.10847, 2023.

[7] Rahman M M, Marculescu R. "G-cascade: Efficient cascaded graph convolutional decoding for 2d medical image segmentation[C]," Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp.7728-7737, 2024.

[8] Song J, Chen X, Zhu Q, et al. "Global and local feature reconstruction for medical image segmentation[J]. IEEE Transactions on Medical Imaging," vol.41, no.9, pp.2273-2284, 2022.

[9] Roth H R, Oda H, Hayashi Y, et al. "Hierarchical 3D fully convolutional networks for multi-organ segmentation[J]," arXiv preprint arXiv:1704.06382, 2017.

[10] Heidari M, Kazerouni A, Soltany M, et al. "Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation[C]," Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp.6202-6212, 2023.

[11] Huang H, Zheng H, Lin L, et al. "Medical image segmentation with deep atlas prior[J]," IEEE Transactions on Medical Imaging, vol.40, no.12, pp.3519-3530, 2021.

[12] Huang X, Deng Z, Li D, et al. "Missformer: An effective transformer for 2d medical image segmentation[J]," IEEE transactions on medical imaging, vol.42, no.5, pp.1484-1494, 2022.

[13] Yu Q, Zhao X, Pang Y, et al. "Multi-view aggregation network for dichotomous image segmentation[C]," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.3921-3930, 2024.

[14] Zhong Y, Luo Z, Liu C, et al. “PG-SAM: Prior-Guided SAM with Medical for Multi-organ Segmentation[J],” arXiv preprint arXiv:2503.18227, 2025.
[15] Alom M Z, Yakopcic C, Hasan M, et al. “Recurrent residual U-Net for medical image segmentation[J],” Journal of medical imaging, vol.6, no.1, pp.014006-014006, 2019.
[16] Zhang L, Gao W, Yi J, et al.“ SACNet: A Spatially Adaptive Convolution Network for 2D Multi-organ Medical Segmentation[C],” 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp.1748-1751, 2024.
[17] Ai Y, Liu J, Li Y, et al. “SAMA: a self-and-mutual attention network for accurate recurrence prediction of non-small cell lung cancer using genetic and CT data[J],” IEEE Journal of Biomedical and Health Informatics, 2024.
[18] Yang Y, Chen Q, Li Y, et al. “Segmentation guided crossing dual decoding generative adversarial network for synthesizing contrast-enhanced computed tomography images[J],” IEEE Journal of Biomedical and Health Informatics, vol.28, no.8, pp.4737-4750, 2024.
[19] Li X, Qin X, Huang C, et al. “SUnet: A multi-organ segmentation network based on multiple attention[J],” Computers in Biology and Medicine, vol.167, pp.107596, 2023.
[20] Liu Z, Lin Y, Cao Y, et al. “Swin transformer: Hierarchical vision transformer using shifted windows[C],” Proceedings of the IEEE/CVF international conference on computer vision, pp.10012-10022, 2021.
[21] Cao H, Wang Y, Chen J, et al. “Swin-unet: Unet-like pure transformer for medical image segmentation[C],” European conference on computer vision. Cham: Springer Nature Switzerland, pp.205-218, 2022.
[22] Tragakis A, Kaul C, Murray-Smith R, et al. “The fully convolutional transformer for medical image segmentation[C],” Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp.3660-3669, 2023.
[23] Azad R, Heidari M, Shariatnia M, et al. “Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation[C],” International Workshop on PRedictive Intelligence In MEdicine. Cham: Springer Nature Switzerland, pp.91-102, 2022.
[24] Chen J, Lu Y, Yu Q, et al. “Transunet: Transformers make strong encoders for medical image segmentation[J],” arXiv preprint arXiv:2102.04306, 2021.
[25] Jiang Z, Ding C, Liu M, et al. “Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task[C],” International MICCAI brainlesion workshop. Cham: Springer International Publishing, pp.231-241, 2019.
[26] Ronneberger O, Fischer P, Brox T. “U-net: Convolutional networks for biomedical image segmentation[C],” International Conference on Medical image computing and computer-assisted intervention. Cham: Springer international publishing, pp.234-241, 2015.
[27] Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, et al. “Unet++: A nested u-net architecture for medical image segmentation[C],” International workshop on deep learning in medical image analysis. Cham: Springer International Publishing, pp.3-11, 2018.
[28] Sun H, Xu J, Duan Y. “Paratranscnn: Parallelized transcnn encoder for medical image segmentation[J],” arXiv preprint arXiv:2401.15307, 2024.
[29] Milletari F, Navab N, Ahmadi S A. “V-net: Fully convolutional neural networks for volumetric medical image segmentation[C],” 2016 fourth international conference on 3D vision (3DV). Ieee, pp.565-571, 2016.
[30] Azad R, Al-Antary M T, Heidari M, et al. “Transnorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model[J],” IEEE access, vol.10, pp.108205-108215, 2022.
[31] Bernard O, Lalande A, Zotti C, et al. “Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?[J],” IEEE transactions on medical imaging, vol.37, no.11, pp.2514-2525, 2018.
[32] Azad R, Arimond R, Aghdam E K, et al. “Dae-former: Dual attention-guided efficient transformer for medical image segmentation[C],” International workshop on predictive intelligence in medicine. Cham: Springer Nature Switzerland, pp.83-95, 2023.
[33] Chai S, Jain R K, Mo S, et al. “A novel adaptive hypergraph neural network for enhancing medical image segmentation[C],” International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, pp.23-33, 2024.
[34] Vaswani A, Shazeer N, Parmar N, et al. “Attention is all you need[J],” Advances in neural information processing systems, vol.30, 2017.
[35] Wang H, Xie S, Lin L, et al. “Mixed transformer u-net for medical image segmentation[C],” ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp.2390-2394, 2022.