

CSSWin-UNet: A Transformer UNet Combined with Patch-Based and Cross-Shaped Windows for Medical Image Segmentation

Fufang Li^{1, a *}, Xinting Chen^{1, b}

¹ School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, Guangdong Province, 510006, China.

^a liff@gzhu.edu.cn, ^b xintingchen@e.gzhu.edu.cn

Abstract. In medical image segmentation, both Convolutional Neural Networks (CNNs) and self-attention mechanisms have shown success but also have limitations. CNNs excel at capturing local features but struggle with long-range dependencies, while self-attention effectively models global context but is computationally intensive and may lose fine local details. To address these challenges, we propose CSSWin-UNet, a novel U-shaped architecture that integrates two complementary self-attention mechanisms (Swin and CSWin Transformer) to achieve a balanced representation of both local features and global dependencies, while substantially reducing computational overhead. Experimental results on multiple public benchmark datasets demonstrate that CSSWin-UNet delivers superior segmentation performance with significantly lower model complexity, highlighting its potential for practical deployment in medical imaging tasks.

Keywords: Medical image segmentation; self-attention mechanisms; u-shaped architecture.

1. Introduction

Medical image segmentation is essential for accurate diagnosis, treatment planning, and patient monitoring. Recent advances in deep learning, especially CNNs and Transformer-based models[1, 2], have significantly improved segmentation accuracy and efficiency. Traditional manual methods are time-consuming and variable, while CNNs—particularly U-Net[1]—have automated segmentation by effectively preserving spatial information. However, CNNs struggle to capture long-range dependencies. Transformers[3, 4], leveraging self-attention, excel in modeling global context and have shown promising results in medical image segmentation, often surpassing CNN-based methods.

In this study, we propose a novel U-Net-inspired architecture that incorporates a newly designed Transformer-based module, referred to as CSSWin Transformer block, designed to leverage patch-based windows for local features and cross-shaped windows for global context. By combining Swin Transformer[5] and CSWin Transformer[6] mechanisms, the model captures local and global dependencies efficiently. Unlike hybrid CNN-Transformer models, our fully Transformer-based approach reduces feature degradation and enhances segmentation accuracy.

The contributions of this paper are summarized as follows:

- We design a novel module, CSSWin Transformer block, and integrate it into a U-Net framework to construct an innovative medical image segmentation model.
- The model combines Swin Transformer and CSWin Transformer, enabling effective capture of both local and global features while reducing computational complexity. In contrast to existing hybrid models that combine CNNs and Transformers, the proposed fully Transformer-based approach mitigates feature loss and enhances segmentation performance.
- The effectiveness of the proposed model is validated through extensive experiments on multiple medical image segmentation datasets.

The paper is organized as follows: Section 2 reviews CNN- and Transformer-based medical image segmentation. Section 3 details the proposed CSSWin module and U-shaped network. Section 4 presents experimental results, and Section 5 concludes the study.

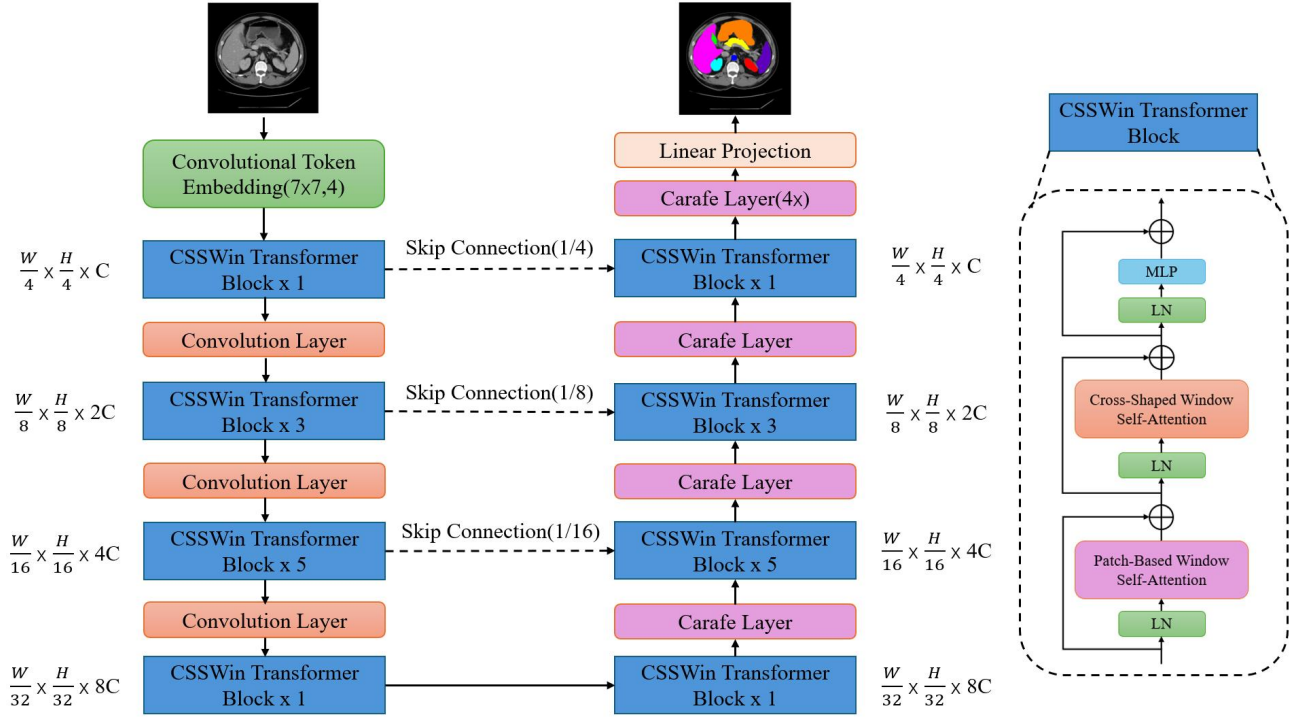


Fig. 1 Overview of the proposed CSSWin-UNet.

2. Related Works

2.1 CNN-Based Medical Image Segmentation

CNN-based methods have long been foundational in medical image segmentation. The U-Net architecture[1], known for its encoder-decoder design with skip connections, effectively captures multi-scale features and preserves contextual information. Various U-Net variants address specific challenges: 3D U-Net[7] processes volumetric data; ResUNet[8] uses residual connections for deeper networks; U-Net++[9] adds dense skip connections; Attention U-Net[10] integrates attention gates; and nnU-Net[11] automates pipeline configuration based on dataset and design fingerprints, achieving state-of-the-art results in numerous challenges. These models excel in tasks like tumor[12] and skin lesion segmentation[13]. However, CNNs are limited by their restricted receptive fields, hindering their ability to model long-range dependencies crucial for complex anatomical segmentation.

2.2 Transformer-Based Medical Image Segmentation

To address CNN limitations, Transformer-based architectures have gained attention for modeling long-range dependencies via self-attention. Vision Transformer (ViT)[14] demonstrated competitive performance in image classification, inspiring extensions to segmentation tasks. In medical image segmentation, various Transformer-based models have emerged. TransUNet[3] and UNETR[4] combine Transformer encoders with CNN decoders, balancing global and local features. TransFuse[15] uses a BiFusion module to combine low-level spatial and high-level semantic features, improving segmentation performance. Swin UNETR[16] and HiFormer[17] enhance efficiency and accuracy by integrating shifted window self-attention and multi-scale fusion. Pure Transformer models like Swin-Unet[18] and CSWin-UNet[19] utilize Swin Transformer[5] blocks and CSWin Transformer[6] blocks to improve receptive field modeling and spatial precision. Despite progress, balancing local-global feature extraction and computational cost remains challenging. Inspired by Swin and CSWin Transformers, we propose CSSWin-UNet, which integrates their strengths to improve accuracy while reducing complexity, offering a promising approach for medical image segmentation.

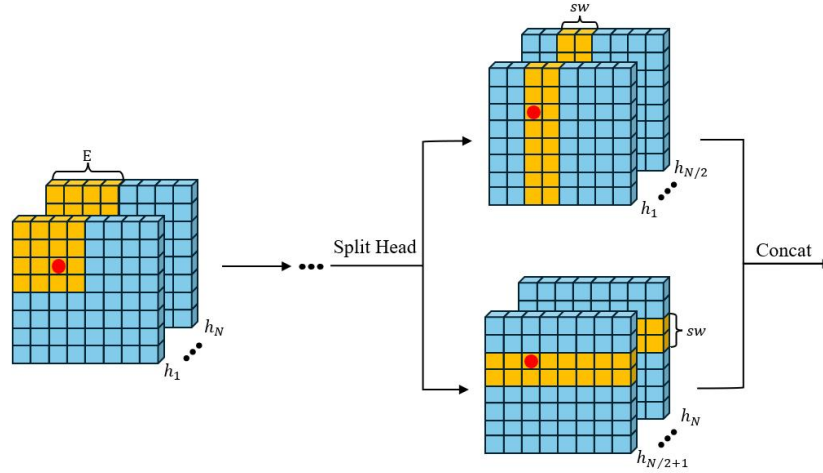


Fig. 2 Illustration of the window-based multi-head self attention (W-MSA) and CSWin self-attention mechanism.

3. Method

3.1 Architecture Overview

The CSSwin-Unet architecture (Fig. 1) consists of an encoder, a decoder, and skip connections, with CSSwin Transformer blocks as the core. The encoder uses a convolutional token embedding layer (7×7 kernel, stride 4) as in CvT[20] to generate $H/4 \times W/4 \times C$ tokens, followed by CSSwin Transformer blocks and convolutional layers for feature extraction and downsampling. The decoder mirrors the encoder with four symmetric stages, using skip connections for multi-scale feature fusion. Inspired by CSWin-UNet[19], the decoder incorporates CARAFE[21] for upsampling and channel reduction, alternating with CSSwin blocks. A final $4 \times$ CARAFE restores full resolution ($H \times W$), followed by a linear projection to generate the segmentation map.

3.2 CSSWin Transformer Block

Classical Transformers model long-range dependencies via global self-attention but face high computational cost and limited locality. Swin Transformer[5] mitigates this through shifted local windows and hierarchical design, enabling multi-scale representation. CSWin Transformer[6] further improves cross-region interaction and local detail preservation using cross-shaped self-attention and Local Enhanced Position Encoding (LePE). Motivated by their complementarity, we propose the CSSWin Transformer block (Fig. 1), which integrates patch-based window attention (W-MSA), cross-shaped attention (CSWin), LayerNorm, MLP, and skip connections. By combining W-MSA's local focus with CSWin's global modeling, our design enhances segmentation accuracy and efficiency. As illustrated in Fig. 2, W-MSA captures local dependencies, while CSWin models broader context with reduced complexity.

In the patch-based window self-attention module, the input feature $X \in \mathbb{R}^{H \times W \times C}$ is divided into $M = (H \times W)/E^2$ non-overlapping windows $[X_w^1, X_w^2, \dots, X_w^M]$, each of size $E \times E$. The window size E controls the trade-off between model capacity and efficiency. For the n -th attention head, with query, key, and value dimensions $d_n = C/N$ (where C is the number of channels and N is the number of heads), the window-based self-attention output W-Attention $_n$ is computed as follows:

$$\begin{aligned} X &= [X_w^1, X_w^2, \dots, X_w^M], \\ Y_n^i &= \text{Attention}(X_w^i W_n^Q, X_w^i W_n^K, X_w^i W_n^V), \\ \text{W-Attention}_n(X) &= [Y_n^1, Y_n^2, \dots, Y_n^M], \end{aligned}$$

where $i \in \{1, \dots, M\}$ denotes the index of the current window. The feature map of the i -th window is $X_w^i \in \mathbb{R}^{(E \times E) \times C}$. For the n -th attention head, $W_n^Q, W_n^K, W_n^V \in \mathbb{R}^{C \times d_n}$ are the query, key,

and value projection matrices, respectively. Y_n^i denotes the self-attention output for the i -th window in the n -th head. After independently computing self-attention in each head, the outputs from all heads are concatenated along the channel dimension, as illustrated below:

$$\begin{aligned} \text{head}_n &= \text{W-Attention}_n(X), \quad n = 1, \dots, N, \\ \text{W-MSA}(X) &= \text{Concat}(\text{head}_1, \dots, \text{head}_N)W^o, \end{aligned}$$

where head_n denotes the output of the n -th attention head, and $W^o \in \mathbb{R}^{C \times C}$ represents a learnable matrix that linearly transforms the concatenated heads. This transformation captures inter-head dependencies and integrates diverse attention into a unified representation.

In the horizontal stripe self-attention of the cross-shaped window module, the input feature $X \in \mathbb{R}^{H \times W \times C}$ is evenly divided into $M = H/sw$ non-overlapping stripes, each of size $sw \times W$. The stripe width sw is adjustable to balance model capacity and computational cost. For the n -th attention head, with projected queries, keys, and values of dimension $d_n = C/N$, the output of horizontal self-attention H-Attention_n is computed as follows:

$$\begin{aligned} X &= [X_h^1, X_h^2, \dots, X_h^M], \\ Y_n^i &= \text{Attention}(X_h^i W_n^Q, X_h^i W_n^K, X_h^i W_n^V), \\ \text{H-Attention}_n(X) &= [Y_n^1, Y_n^2, \dots, Y_n^M], \end{aligned}$$

where i denotes the index of the horizontal stripe, ranging from 1 to M , and $X_h^i \in \mathbb{R}^{sw \times W \times C}$ represents its corresponding feature map. The projection matrices for the query, key, and value in the n -th head are $W_n^Q, W_n^K, W_n^V \in \mathbb{R}^{C \times d_n}$, respectively. The self-attention output for the i -th stripe in the n -th head is Y_n^i . The vertical stripe attention is defined analogously, with the output of the n -th head denoted as $\text{V-Attention}_n(X)$.

We split the N attention heads evenly into two groups of $N/2$ heads each. One group computes self-attention over horizontal stripes, and the other over vertical stripes. The resulting feature maps are then concatenated along the channel dimension to produce the final output, as illustrated below:

$$\begin{aligned} \text{head}_n &= \begin{cases} \text{H-Attention}_n(X), & n = 1, \dots, N/2, \\ \text{V-Attention}_n(X), & n = N/2 + 1, \dots, N, \end{cases} \\ \text{CSWin-Attention}(X) &= \text{Concat}(\text{head}_1, \dots, \text{head}_N)W^o, \end{aligned}$$

where the output of the n -th attention head, head_n , is concatenated and linearly transformed by a learnable matrix $W^o \in \mathbb{R}^{C \times C}$, producing the final output. This process integrates horizontal and vertical context, enhancing the spatial understanding of the input image.

Building upon the aforementioned two self-attention mechanisms, the CSSWin Transformer block can be formally defined as follows:

$$\begin{aligned} \hat{X}_1^l &= \text{W-MSA}(\text{LN}(X^{l-1})) + X^{l-1}, \\ \hat{X}_2^l &= \text{CSWin-Attention}(\text{LN}(\hat{X}_1^l)) + \hat{X}_1^l, \\ X^l &= \text{MLP}(\text{LN}(\hat{X}_2^l)) + \hat{X}_2^l, \end{aligned}$$

where X^l denotes the output of the l -th CSSWin Transformer block or the preceding convolutional layer at each stage.

4. Experiments

4.1 Datasets

Synapse. The Synapse multi-organ segmentation dataset includes 30 abdominal CT scans from the MICCAI 2015 challenge, totaling 3,779 slices at 512×512 resolution. Following[3, 22], 18 scans are used for training and 12 for evaluation. Segmentation of eight abdominal organs is evaluated using average Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD).

ACDC. The 2017 ACDC challenge released a benchmark dataset of 100 multi-class 3D cardiac MRI scans with expert annotations for the right ventricle (RV), myocardium (MYO), and left

ventricle (LV)[23]. In this study, 70 scans were used for training, 10 for validation, and 20 for testing. Segmentation performance was assessed using the average Dice Similarity Coefficient (DSC).

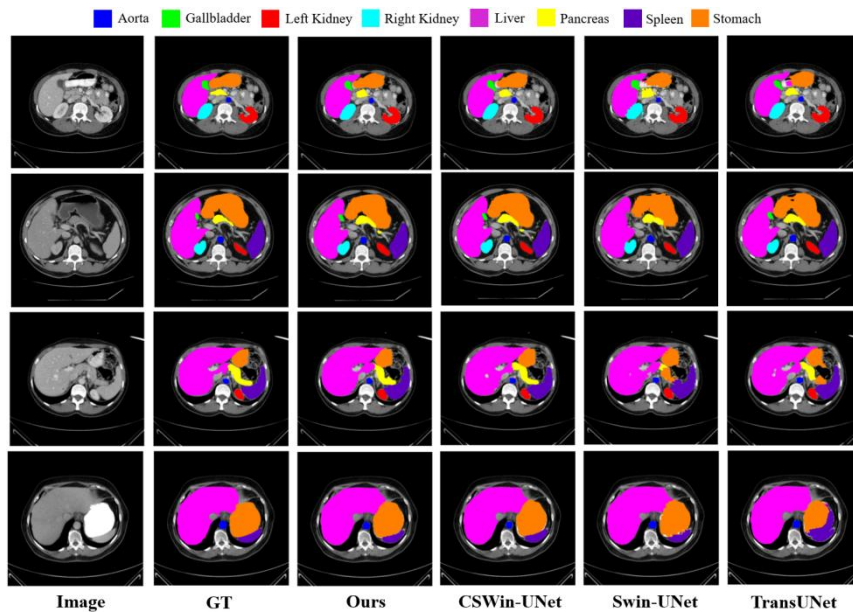


Fig. 3 Segmentation performance of different methods on eight organs from the Synapse dataset.

Table 1. Performance comparison of recent medical image segmentation methods on the Synapse dataset based on DSC and HD metrics. The best and second-best results are indicated in bold and underlined, respectively.

Methods	DSC↑	HD↓	Aorta	Gallblad der	Kidney (L)	Kidney (R)	Liver	Pancr eas	Spleen	Stomach
V-Net ^[25]	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR ^[23]	69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
U-Net ^[11]	76.85	39.70	<u>89.07</u>	69.72	77.77	68.60	93.43	53.98	86.67	75.58
Att-UNet ^[10]	77.77	36.02	89.55	<u>68.88</u>	77.98	71.11	93.57	58.04	87.30	75.75
TransUne _t ^[3]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
MPSHT ₆ ^[26]	79.76	21.55	87.74	65.36	80.55	<u>80.77</u>	93.64	59.84	89.46	81.13
TransDeepLab ^[27]	80.16	21.25	86.04	69.16	84.08	79.88	93.53	61.19	89.00	78.40
MT-UNE _t ^[28]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
Swin-UNet ^[18]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	<u>90.66</u>	76.60
CSWin-UNet ^[19]	<u>81.12</u>	18.86	87.13	67.85	<u>83.51</u>	78.53	95.23	65.94	89.05	<u>81.74</u>
Ours	82.00	<u>19.90</u>	86.52	66.38	85.77	82.02	<u>94.98</u>	<u>65.43</u>	91.31	83.63

Table 2. Quantitative comparison of recent medical image segmentation methods on the ACDC dataset using the DSC. The best and second-best results are highlighted in bold and underlined, respectively.

Methods	DSC↑	RV	MYO	LV
U-Net[1]	87.55	87.10	80.63	94.92
Att-UNet[10]	86.75	87.58	79.20	93.47
TransUnet[3]	89.71	88.86	84.53	95.73

MPSHT[26]	91.80	90.13	89.34	95.92
nnUNet[11]	90.91	89.21	90.20	93.35
UNetR[4]	88.61	85.29	86.52	94.02
MedFuseNet[29]	89.73	88.36	86.29	94.54
Swin-UNet[18]	90.00	88.55	85.62	95.83
CSWin-UNet[19]	91.46	89.68	88.94	95.76
Ours	92.63	93.35	88.17	96.35

Table 3. Comparison of computational complexity among different medical image segmentation methods on the Synapse dataset. The best and second-best results are highlighted in bold and with underlining, respectively.

Methods	Params(M)	FLOPs(G)	DSC↑	HD↓
TransUnet[3]	96.07	88.91	77.48	31.69
UNetR[4]	86.00	54.70	79.56	22.97
Swin-UNet[18]	27.17	6.16	79.13	21.55
CSWin-UNet[19]	23.57	4.72	81.12	18.86
Ours	22.79	4.95	82.00	19.90

4.2 Implementation Details

CSSWin-UNet was trained and evaluated on an NVIDIA® GeForce RTX™ 3090 GPU (24 GB VRAM) using Python and PyTorch. To accelerate convergence, Transformer blocks were initialized with ImageNet-pretrained weights[24]. Data augmentation (rotation, flipping) was applied to improve generalization. The model was trained with a batch size of 24 and an initial learning rate of 0.05, and used SGD optimizer with 0.9 momentum and 10^{-4} weight decay.

4.3 Experiment Results

Table 1 presents a performance comparison of our CSSWin-UNet with state-of-the-art methods on the Synapse dataset. Our model outperforms others in both average Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD), with DSC gains of 2.87% over Swin-UNet[18] and 0.88% over CSWin-UNet[19], and HD reductions of 6.69% and 1.65% compared to MT-UNet[28] and Swin-UNet[18], respectively. Significant improvements are observed in segmenting the kidneys, spleen, and stomach, demonstrating the model's robustness. To intuitively assess performance, Fig. 3 presents segmentation results on the Synapse dataset. Compared to CSWin-UNet[19], Swin-UNet[18], and TransUnet[3], which exhibit missegmentations in organs such as the stomach, pancreas, and spleen, our method provides more accurate outputs. The proposed CSSWin-UNet demonstrates superior robustness and boundary preservation, effectively handling complex anatomical structures under challenging conditions, as confirmed by both visual and quantitative evaluations.

Table 2 compares the performance of our CSSwin-UNet with state-of-the-art methods on the ACDC dataset. RV, MYO, and LV denote the right ventricle, myocardium, and left ventricle, respectively. Our method achieves an average Dice Similarity Coefficient (DSC) of 92.63%, outperforming Swin-UNet[18] and CSWin-UNet[19] by 2.63% and 1.17%. Notably, it shows significant improvement in segmenting both ventricles, highlighting its effectiveness and robustness.

While model performance is critical, computational efficiency is equally important. As shown in Table 3, we evaluate the computational efficiency of various methods using two metrics: the number of parameters (Params, in millions) and floating-point operations (FLOPs, in billions). We compare our method with state-of-the-art approaches on the Synapse dataset using DSC and HD metrics. Our model achieves competitive performance with notably fewer parameters, highlighting its superior efficiency.

5. Conclusion

In this study, we propose CSSWin-UNet, a U-shaped segmentation model featuring the CSSWin Transformer block, which combines Swin and CSWin self-attention to address limitations of CNNs and standard Transformers. This design balances local and global feature modeling with improved efficiency. Additionally, we incorporate CARAFE for detail-preserving upsampling. Experiments on public datasets show that CSSWin-UNet outperforms existing methods in accuracy and efficiency, demonstrating its effectiveness and potential for advancing attention-based medical image segmentation. We believe that CSSWin-UNet offers a novel and promising direction for extending attention mechanisms in medical image segmentation.

References

- [1] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, 9351: 234-241.
- [2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017, 30.
- [3] Chen J, Lu Y, Yu Q, et al. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint*, 2021, arXiv:2102.04306.
- [4] Hatamizadeh A, Tang Y, Nath V, et al. UNETR: Transformers for 3D medical image segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022: 574-584.
- [5] Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 10012-10022.
- [6] Dong X, Bao J, Chen D, et al. CSWin Transformer: A general vision transformer backbone with cross-shaped windows. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 12124-12134.
- [7] Çiçek Ö, Abdulkadir A, Lienkamp S S, et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, 2016, 9901: 424-432.
- [8] Diakogiannis F I, Waldner F, Caccetta P, et al. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020, 162: 94-114.
- [9] Zhou Z, Siddiquee M M R, Tajbakhsh N, et al. UNet++: A nested U-Net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support – DLMIA & ML-CDS 2018*, 2018, 11045: 3-11.
- [10] Oktay O, Schlemper J, Folgoc L L, et al. Attention U-Net: Learning where to look for the pancreas. *arXiv preprint*, 2018, arXiv:1804.03999.
- [11] Isensee F, Jaeger P F, Kohl S A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 2021, 18(2): 203-211.
- [12] Isensee F, Kickingereder P, Wick W, et al. No new-net. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries – BrainLes 2018, 2019*, 11383: 234-244.
- [13] Gessert N, Nielsen M, Shaikh M, et al. Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. *MethodsX*, 2020, 7: 100864.
- [14] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv preprint*, 2020, arXiv:2010.11929.
- [15] Zhang Y, Liu H, Hu Q. Transfuse: Fusing transformers and CNNs for medical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*, 2021, 12901: 14-24.
- [16] Hatamizadeh A, Nath V, Tang Y, et al. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. *International MICCAI BrainLesion Workshop*, 2021: 272-284.
- [17] Heidari M, Kazerouni A, Soltany M, et al. HiFormer: Hierarchical multi-scale representations using transformers for medical image segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023: 6202-6212.

- [18] Cao H, Wang Y, Chen J, et al. Swin-UNet: UNet-like pure transformer for medical image segmentation. European Conference on Computer Vision, 2022: 205-218.
- [19] Liu X, Gao P, Yu T, et al. CSWin-UNet: Transformer UNet with cross-shaped windows for medical image segmentation. Information Fusion, 2025, 113: 102634.
- [20] Wu H, Xiao B, Codella N, et al. CvT: Introducing convolutions to vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 22-31.
- [21] Wang J, Chen K, Xu R, et al. CARAFE: Content-aware reassembly of features. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 3007-3016.
- [22] Fu S, Lu Y, Wang Y, et al. Domain adaptive relational reasoning for 3D multi-organ segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020, 2020, 12261: 656-666.
- [23] Bernard O, Lalande A, Zotti C, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?. IEEE Transactions on Medical Imaging, 2018, 37(11): 2514-2525.
- [24] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115: 211-252.
- [25] Milletari F, Navab N, Ahmadi S A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D Vision (3DV), 2016: 565-571.
- [26] Zhao Y, Li J, Hua Z. MPSHT: Multiple progressive sampling hybrid model multi-organ segmentation. IEEE Journal of Translational Engineering in Health and Medicine, 2022, 10: 1-9.
- [27] Azad R, Heidari M, Shariatnia M, et al. TransDeepLab: Convolution-free transformer-based DeepLab v3+ for medical image segmentation. International Workshop on Predictive Intelligence in Medicine, 2022: 91-102.
- [28] Wang H, Xie S, Lin L, et al. Mixed Transformer U-Net for medical image segmentation. ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: 2390-2394.
- [29] Chen R, He S, Xie J, et al. MedFuseNet: Fusing local and global deep feature representations with hybrid attention mechanisms for medical image segmentation. Scientific Reports, 2025, 15(1): 5093.