

RCLD-UNET: AN ASYMMETRIC U-SHAPED ENCODER-DECODER BASED ON RECURRENT CRISS-CROSS ATTENTION AND LINEAR DEFORMABLE CONVOLUTION FOR MEDICAL IMAGE SEGMENTATION

Bing SU, Mengyu ZHAO*, Yifeng LIN, Peng JIN

Changzhou University

School of Computer Science and Artificial Intelligence

No. 2468, Yanzheng West Avenue, Wujin District

Changzhou City, Jiangsu Province, China

e-mail: subing@cczu.edu.cn, S23150812069@smail.cczu.edu.cn,

linyifeng_cczu@cczu.edu.cn, S22150812024@smail.cczu.edu.cn

Abstract. With the continuous advancement of medical technology, medical image segmentation plays a pivotal role in various aspects of disease diagnosis, treatment planning, and therapeutic evaluation. In recent years, methods based on U-shaped Convolutional Neural Networks (CNNs), U-shaped Transformer architectures, and U-shaped Mamba models have achieved significant progress in this domain, greatly promoting research and application development. Although CNNs excel at local feature extraction, they are inherently limited in modeling global dependencies due to the locality of convolution operations. Transformers address this limitation through self-attention mechanisms that capture long-range dependencies effectively, but their quadratic computational complexity with respect to sequence length severely impacts efficiency. While the Mamba architecture demonstrates promising performance in long-sequence modeling, its generalization to other tasks remains limited. To address these challenges, this paper proposes RCLD-UNet, an efficient asymmetric U-shaped encoder-decoder network designed for 2D medical image segmentation. Built upon the U-Net backbone, RCLD-UNet incorporates the Recurrent Criss-Cross Attention (RCCA) module in the encoder, which leverages sparse connections and recurrent operations to efficiently capture global contextual information, thereby overcoming the limitations of standard convolutions in modeling global dependencies. In the decoder, a Linear Deformable Convolution

* Corresponding author

(LDCConv) module is introduced to enhance feature extraction through irregular convolutions, while ensuring linear parameter growth to balance performance and computational cost. Extensive experimental results on three public datasets indicate that RCLD-UNet achieves superior performance compared to existing methods. Specifically, the proposed model achieves average Dice Similarity Coefficients (DSC) of 84.14 %, 92.02 %, and 80.84 % on the Synapse, ACDC, and MoNuSeg datasets, respectively.

Keywords: RCLD-UNet, medical image segmentation, convolutional neural networks, recurrent criss-cross attention, linear deformable convolution

1 INTRODUCTION

With the continuous advancement of deep learning technologies, computer vision has seen widespread adoption in medical image analysis. As a key task within this field, medical image segmentation plays an essential role in a variety of clinical and research applications, such as vessel segmentation in retinal images, cell segmentation in electron microscopy (EM) images, tumor segmentation in computed tomography (CT) and magnetic resonance imaging (MRI) scans, liver segmentation, and multi-organ segmentation. These advanced segmentation techniques have not only enhanced the understanding of medical images but also significantly improved the accuracy and efficiency of computer-aided diagnosis.

Convolutional Neural Networks (CNNs) have been widely adopted in the field of medical image segmentation, particularly with the introduction of Fully Convolutional Networks (FCNs) [1] and their variants such as U-Net [2]. U-Net, with its symmetric encoder-decoder architecture and skip connections, has effectively combined local and spatial features, achieving notable success in medical image segmentation. Subsequently, numerous U-Net-based variants have been proposed to further enhance segmentation performance. For example, U-Net++ [3] improves feature fusion through nested and dense skip connections; UNet3+ [4] introduces full-scale skip connections and deep supervision to better capture both fine-grained and coarse-grained features; and Attention U-Net [5] incorporates attention gating modules in the decoder to focus on salient regions, thereby enhancing the extraction of relevant local features and improving segmentation accuracy. Despite the strength of convolutional operations in local feature extraction, CNNs are inherently limited by their receptive fields, making it difficult to model global contextual dependencies. Although techniques such as dilated convolutions [6], image pyramids [7], and attention mechanisms [8] have been introduced to mitigate this issue, CNNs still struggle to capture long-range dependencies within images, which restricts their ability to meet the high-precision and global perception requirements of medical image segmentation tasks.

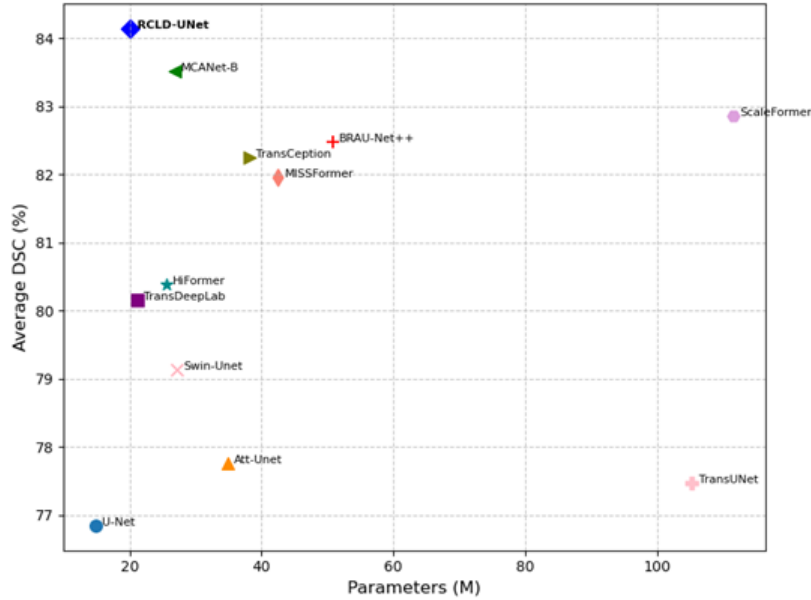


Figure 1. Comparison of parameter and Dice accuracy trade-offs on the Synapse dataset

Dataset	Num of Class	Images	Input Size	Modality	Train	Valid	Test
Synapse	9	30	224 × 224	CT	18 (2 211 slices)	—	12
ACDC	4	100	224 × 224	MRI	70 (1 304 slices)	10 (182 slices)	20 (40 slices)
MoNuSeg	2	44	224 × 224	Microscopy Images	24	6	14

Table 1. Detailed introduction of the dataset we used

The successful application of Transformer architectures in natural language processing (NLP) has inspired extensive exploration of their potential in visual tasks. As a pioneering work, Vision Transformer (ViT) [9] introduced the Transformer architecture into computer vision by dividing 2D images into sequences of position-encoded patches, demonstrating performance comparable to or even surpassing traditional CNN-based methods on several large-scale visual benchmarks. This validated the Transformer’s strong capability in modeling long-range dependencies. Building on this strength, researchers have further investigated the integration of CNNs and Transformers. In 2021, TransUNet [10] was the first to propose a hybrid framework that effectively combines high-resolution local features extracted by CNNs with global contextual information encoded by Transformers, enabling

precise segmentation of complex structures and lesion regions in medical images. Subsequently, a series of 3D medical image segmentation models such as TransBTS [11], CoTr [12], and UNETR [13], as well as 2D models including MTU-Net [14], TMU [15], UCTransNet [16], ScaleFormer [17], MISSFormer [18], and MedT [19], have been developed, further validating the effectiveness of CNN-Transformer hybrid architectures in medical image segmentation. In parallel, pure Transformer-based models such as SegFormer [20] and Swin-Unet [21] have also emerged, exploring the potential of fully Transformer-driven frameworks in this domain. Despite the remarkable performance of both Transformer-based and hybrid CNN-Transformer models, their high computational complexity and resource consumption remain major limitations, particularly when applied to high-resolution medical image segmentation tasks.

In recent years, the Mamba architecture, based on state space models (SSMs), has garnered significant attention in medical image segmentation tasks, leading to the development of variants such as U-Mamba [22], VM-UNet [23], and HC-Mamba [24]. Compared with CNNs and Transformers, Mamba demonstrates superior performance in capturing long-range dependencies while maintaining linear computational complexity, thereby improving efficiency. However, Mamba is primarily suited for long sequence modeling, and the advantages of SSM mechanisms are limited in other tasks, resulting in constrained performance improvements.

In summary, existing segmentation methods based on CNNs, Transformers, and Mamba architectures each exhibit certain limitations. To address these challenges, we propose an asymmetric U-shaped encoder-decoder network, named RCLD-UNet. It builds upon the simple yet effective U-Net architecture and comprises an encoder, a bottleneck, a decoder, and skip connections. Specifically, the encoder incorporates the Recurrent Criss-Cross Attention (RCCA) module from CCNet [25] to enhance global context modeling and compensate for the limitations of conventional convolutions in capturing spatial dependencies. In the decoder, the Linear Deformable Convolution (LDConv) module from LDConv [26] is adopted to enable flexible feature reconstruction while maintaining linear parameter growth. Extensive experiments conducted on three public datasets—Synapse, ACDC, and MoNuSeg—demonstrate that RCLD-UNet achieves strong segmentation performance and good generalization with low parameter overhead. The main contributions of this work are summarized as follows:

1. RCCA modules are introduced at each stage of the U-Net encoder and the bottleneck. These modules adopt a dual recurrent design with parameter sharing based on the Criss-Cross Attention (CCA) mechanism, effectively avoiding parameter growth during recurrence. This enhances the ability to model contextual information and compensates for the limitations of convolution in capturing global dependencies.
2. To address the need for deep feature modeling in the decoder, LDConv modules are introduced in its deeper layers to replace the conventional double convolution

structure. These modules leverage a flexible sampling mechanism and a linear parameter design, which not only reduce computational and memory overhead but also enhance the feature modeling capability of the decoder.

- 3. Experimental results on the Synapse, ACDC, and MoNuSeg datasets demonstrate that the proposed model outperforms various state-of-the-art methods with only a modest increase in parameter count, highlighting its potential and advantages in medical image segmentation. Table 1 provides detailed information on the three datasets, and Figure 1 illustrates that RCLD-UNet achieves high Dice Similarity Coefficient (DSC) scores despite a modest increase in parameter count.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details the overall architecture of the proposed model and its key modules. Section 4 describes the datasets and experimental settings. Section 5 concludes the paper with a summary of the main findings and contributions.

2 RELATED WORK

2.1 U-Net and Its Variants

CNN-based U-shaped architectures, with U-Net as the classical representative, form the foundation of a modern medical image segmentation. Subsequent variants like U-Net++ [3], ResUNet [27], Attention U-Net [5], and CE-Net [28] introduced dense connections, residual units, attention gates, and context modules, respectively, to enhance feature fusion and representation. In recent years, Transformer-based models such as MaxViT-UNet [29], EMCAD [30], ScaleFormer [17], MISSFormer [18], and VM-UNet [23] have emerged, incorporating multi-axis attention, hybrid attention, or even state space modeling to push the limits of performance. Table 2 summarizes representative U-Net variants and their architectural innovations.

Model	Year	Key Improvements
U-Net [2]	2015	Encoder-decoder architecture with skip connections
UNet++ [3]	2018	Nested skip connections with deep supervision
Attention U-Net [5]	2018	Attention gates in skip connections
ResUNet [27]	2018	Use of ResBlocks instead of standard convolutions
CE-Net [28]	2019	Context encoder module between encoder and decoder
MISSFormer [18]	2021	Transformer-based encoder with context bridge
ScaleFormer [17]	2022	Cross-scale interaction and dual-axis attention
MaxViT-UNet [29]	2023	Multi-axis attention (MaxViT blocks) in decoder
VM-UNet [23]	2024	State space module (VSS)
EMCAD [30]	2024	Multi-scale convolution and hybrid attention

Table 2. Summary of U-Net and its representative variants with key architectural improvements

2.2 Evolution of Attention Mechanisms

Attention mechanisms have been extensively applied in vision tasks to enhance spatial, channel, and contextual modeling. Classic methods such as Squeeze-and-Excitation Network (SENet) [31], Convolutional Block Attention Module (CBAM) [32], and CCNet [25] focus on channel recalibration, combined attention, and criss-cross contextual reasoning, respectively. With the rise of Transformers, attention-based segmentation models have rapidly evolved. Swin-Unet [21] and CSWin-UNet [33] adopt window-based and cross-shaped attention to balance efficiency and global context modeling. TransUNet [10] and UCTransNet [16] introduce multi-head and channel attention into encoder and skip connections, improving semantic representation. Efficient Attention [34] and Visual Attention Network (VAN) [35] further reduce complexity while preserving long-range modeling capacity. Table 3 summarizes representative attention mechanisms, their key innovations, and integration points.

2.3 Development in Convolutional Operations

The design of convolution operations has continued to evolve to improve receptive field modeling, deformation adaptability, and computational efficiency. Classic methods such as dilated convolution [6], deformable convolution [36], and depth-wise separable convolution [37] optimize the structure through sparse sampling, learnable offsets, and factorized operations, respectively. Ghost convolution [38] reduces redundant computation via linear feature generation, while involution [39] enhances spatial modeling by adopting a "spatial-specific, channel-shared" strategy. ConvNeXt [40] modernizes conventional convolutional architectures by integrating key design elements from Transformers, leading to improved performance on vision tasks. Recent approaches include LDConv [26] for linear deformable modeling, FADC [41] for frequency-adaptive dilation, PConv [42] with a pinwheel-shaped pattern for small object detection, and FDConv [43] combining frequency and spatial modulation to enhance frequency-domain representation. Table 4 summarizes the core ideas, advantages, and typical applications of these convolutional operations.

3 METHOD

This section provides a systematic introduction to the proposed RCLD-UNet model. We first present an overview of its overall architecture and processing flow, followed by an in-depth analysis of the structural design, implementation, and functional roles of each integrated module within the network.

3.1 Model Architecture

Figure 2a) illustrates the overall architecture of the proposed RCLD-UNet. The model adopts a typical U-shaped structure composed of an encoder, a decoder,

Method	Year	Attention Type	Application Location	Advantages	Model Type
SENet [31]	2018	Channel Attention	Middle Layer	Enhances channel-wise feature recalibration	CNN
CBAM [32]	2018	Spatial + Channel Attention	Middle Layer	Enhances feature attention in both channel and spatial dimensions	CNN
CCNet [25]	2020	Recurrent Criss-Cross Attention	Middle Layer	Efficient dense context modeling	CNN
Efficient Attention [34]	2021	Linear Self-Attention	Middle Layer	Linear complexity, efficient global modeling for high-res images	CNN
TransUNet [10]	2021	Multi-head Self-Attention (ViT)	Encoder	Models long-range dependencies via Transformer encoder	CNN-Transformer
Swin-UNet [21]	2022	Window-based Self-Attention	Encoder & Decoder	Efficient hierarchical global context modeling	Transformer
UCTransNet [16]	2022	Channel-wise Transformer Attention	Skip Connections	Solves semantic gap in skip connections	CNN-Transformer
VAN [35]	2022	Large Kernel Attention	Middle Layer	Efficient long-range spatial modeling with low complexity	CNN
CSWin-UNet [33]	2025	Cross-shaped Window Self-Attention	Encoder & Decoder	Efficient horizontal & vertical stripe attention	Transformer

Table 3. Summary of representative attention mechanisms, their types, application locations, advantages, and model categories

Convolution Type	Year	Core Idea	Advantages & Limitations	Application
Dilated Convolution [6]	2016	Expands receptive field via sparse sampling	Efficient multi-scale context; may cause grid artifacts	Classification, detection, segmentation
Deformable Convolution [36]	2017	Adds learnable offsets	Flexible geometry modeling with low overhead	Detection, segmentation
Depthwise Separable Convolution [37]	2017	Factorizes convolution into channel and spatial parts	Reduces parameters; limited feature interaction	Classification, detection
Ghost Convolution [38]	2020	Generates redundant features via linear operations	Low cost; retains feature quality	Classification, detection
Involution [39]	2021	Spatial-specific, channel-shared kernels	Enhances spatial modeling with low cost	Classification, detection, segmentation
ConvNeXt [40]	2022	Fuses Transformer ideas into convolution	Combines CNN simplicity with Transformer-level performance	Classification, detection, segmentation
Linear Deformable Convolution (LDCConv) [26]	2024	Enables deformable modeling with linear growth	Balances deformability and efficiency	Detection, segmentation
Frequency-Adaptive Dilated Convolution (FADC) [41]	2024	Adjusts dilation rate based on frequency	Boosts receptive field and frequency response	Detection, segmentation
Pinwheel-Shaped Convolution (PConv) [42]	2024	Pinwheel sampling for small object modeling	Enhances small target detection with low overhead	Small target detection
Frequency Dynamic Convolution (FDCConv) [43]	2025	Frequency-guided dynamic kernel generation	Improves diversity with minimal cost	Classification, detection, segmentation

Table 4. Summary of convolution operations, their core ideas, advantages, and applications

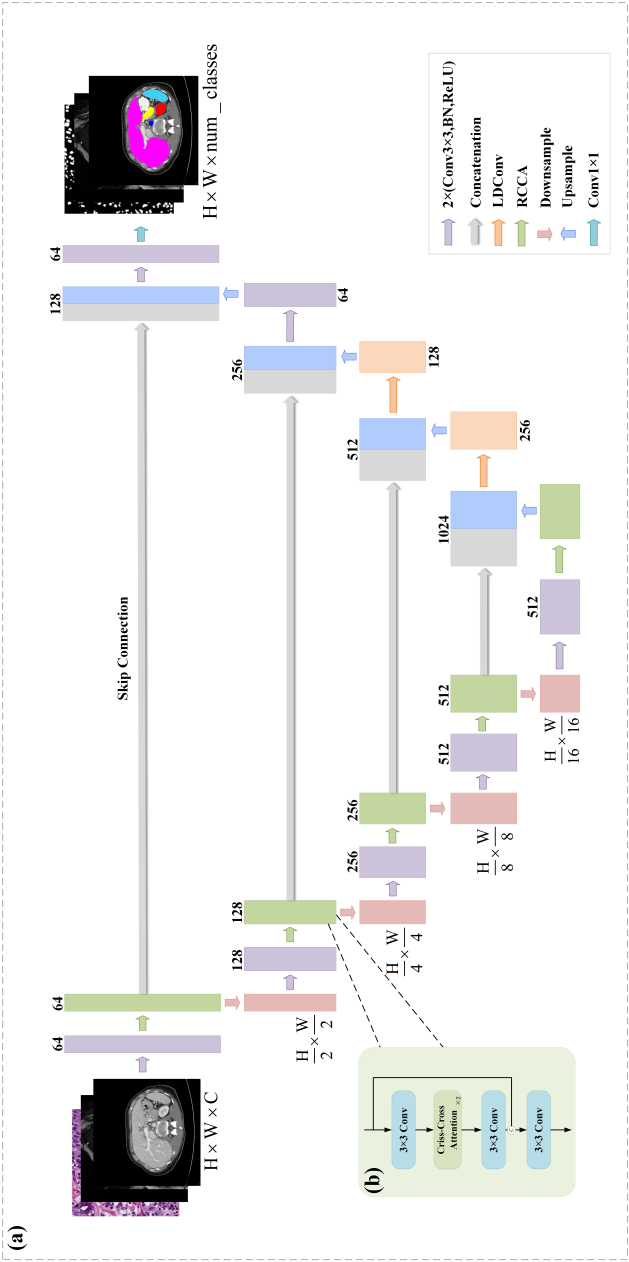


Figure 2. a) The overall architecture of RCLD-UNet. b) Recurrent Criss-Cross Attention (RCCA) module: A 3×3 convolution is used for channel compression, followed by recursive Criss-Cross Attention (CCA) to enhance contextual features. The output is then fused with the original input via channel-wise concatenation to introduce residual information.

a bottleneck, and skip connections, enabling progressive extraction and fusion of multi-scale features.

In the encoder, to enhance the model's ability to capture global contextual information, an RCCA module is introduced after each convolutional block. The feature maps sequentially pass through four encoding stages, each consisting of a double convolutional block followed by an RCCA module. Max-pooling is applied at each stage to progressively reduce spatial resolution and extract high-level semantic features. The layer-wise integration of RCCA effectively improves the global perception capability of local features.

The bottleneck serves as a crucial bridge between the encoder and decoder and plays an important role in semantic representation. Therefore, an RCCA module is also integrated into this stage to further enhance the global representation of features. The encoded features first pass through a double convolutional block, followed by RCCA processing, yielding globally enhanced semantic representations that provide the decoder with richer contextual information.

In the decoder, to strike a balance between parameter efficiency and feature reconstruction capability, LDConv modules are introduced in the two deeper decoding stages, replacing traditional double convolution blocks. This design reduces the number of model parameters while improving segmentation performance. The bottleneck output is first processed by two deep decoder stages (each consisting of an LDConv module and bilinear upsampling), followed by two shallow decoder stages (each consisting of a double convolutional block and bilinear upsampling), progressively restoring spatial resolution to achieve fine-grained segmentation output.

3.2 Criss-Cross Attention (CCA)

The CCNet introduces the CCA module, as shown in Figure 3. This module can extract information from the horizontal and vertical directions for each position, thereby simulating global dependencies in the image with minimal computational cost and memory usage. Assuming the output feature of the double convolution is $F \in R^{C \times W \times H}$, where W and H are the width and height of the feature map, and C is the number of channels. First, we pass the feature F through two 1×1 convolutional layers to obtain two feature maps, Q and K , respectively, where $\{Q, K\} \in R^{C' \times W \times H}$. Here, C' represents the reduced number of channels after convolution. Subsequently, we generate an attention map $A \in R^{(H \times W - 1) \times (W \times H)}$ from Q and K through the *Affinity* operation. This attention map A reflects the correlation between the features of Q and K . The specific operation is that for each position u in the feature map Q , its corresponding vector $Q_u \in R^{C'}$ can be obtained. Based on this positional information, the set of all feature vectors in K that lie in the same row and column as position u can be obtained as $\Omega_u \in R^{(H+W-1) \times C'}$. The *Affinity* operation is described as follows:

$$d_{i,u} = Q_u \Omega_{i,u}^T, \quad (1)$$

where $d_{i,u} \in D$ represents the degree of correlation between the features Q_u and $\Omega_{i,u}$, $i = [1, \dots, H+W-1]$, $D \in R^{(H+W-1) \times (W \times H)}$. Applying the softmax operation along the channel dimension of D yields the attention map A .

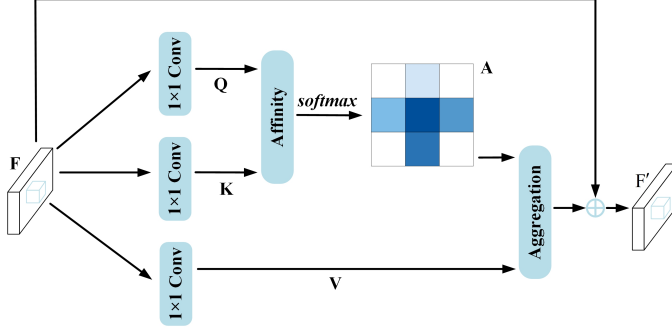


Figure 3. The specific process of the Criss-Cross Attention (CCA) module

Then, the feature map F is passed through another 1×1 convolutional layer to generate a feature map $V \in R^{C \times H \times W}$, which is used for feature adaptation. For each position u in the feature map V , we can obtain a corresponding vector $V_u \in R^C$ and a set $\Phi_u \in R^{(H+W-1) \times C}$. This set Φ_u is the collection of feature vectors that are either in the same row or in the same column as position u . The *Aggregation* operation, as defined by Equation (2), is used to gather contextual information.

$$F'_u = \sum_{i=0}^{H+W-1} A_{i,u} \Phi_{i,u} + F_u. \quad (2)$$

F'_u represents the feature vector at position u in $F' \in R^{C \times W \times H}$. $A_{i,u}$ represents the scalar value at position u with channel index i in A . The contextual information is added to the local features F to enhance the pixel-wise representation capabilities. Therefore, the CCA module possesses powerful context extraction capabilities, enabling it to aggregate contextual information purposefully based on the attention map, making it more robust in semantic segmentation tasks.

3.3 Recurrent Criss-Cross Attention (RCCA)

Due to the limitation that a single CCA operation can only capture contextual information along horizontal and vertical directions, it fails to model the dependencies between pixels located on non-intersecting paths in the feature map. To address this issue, CCNet introduces a recurrent mechanism based on the original CCA module to enhance global modeling of non-local pixel relationships. Specifically, the RCCA module stacks two CCA operations in a recurrent manner to form an R-shaped structure, thereby enabling both direct and indirect information exchange between any pair of pixels.

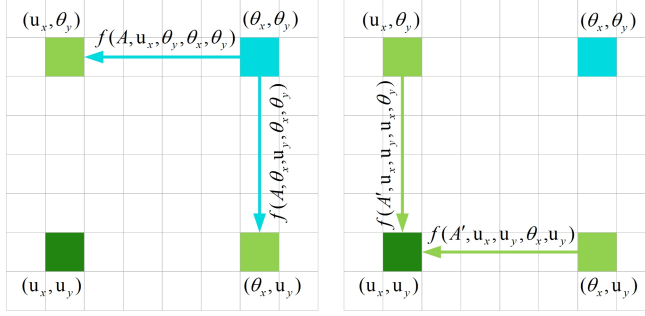


Figure 4. Spatial information propagation in the RCCA module. The first attention step establishes direct dependencies along horizontal and vertical directions, while the second enables indirect connections via intermediate pixels, allowing each position to aggregate contextual information from the entire feature map.

Algorithm 1 Pseudocode of the Recurrent Criss-Cross Attention (RCCA) module

Require: Input feature map $x \in \mathbb{R}^{B \times C \times H \times W}$, recurrence steps r

Ensure: Refined feature map y

- 1: $x_1 \leftarrow \text{Conv}_{3 \times 3}(x) \rightarrow \text{BN} \rightarrow \text{Activation}$ ▷ Channel reduction
 - 2: **for** $i = 1$ to r **do**
 - 3: Compute $Q, K, V \leftarrow \text{Conv}_{1 \times 1}(x_1)$
 - 4: Reshape Q, K along height and width axes:
 $Q_{\text{row}}, K_{\text{row}} \leftarrow \text{reshape}(Q, K, \text{axis} = H)$
 $Q_{\text{col}}, K_{\text{col}} \leftarrow \text{reshape}(Q, K, \text{axis} = W)$
 - 5: Compute attention energies:
 $E_{\text{row}} \leftarrow Q_{\text{row}} K_{\text{row}}^\top + \text{INF_mask}$
 $E_{\text{col}} \leftarrow Q_{\text{col}} K_{\text{col}}^\top$
 - 6: Apply softmax:
 $A_{\text{row}}, A_{\text{col}} \leftarrow \text{softmax}([E_{\text{row}}, E_{\text{col}}])$
 - 7: Compute attended values:
 $V_{\text{row}}, V_{\text{col}} \leftarrow \text{reshape}(V)$
 $out_{\text{row}} \leftarrow V_{\text{row}} A_{\text{row}}^\top$
 $out_{\text{col}} \leftarrow V_{\text{col}} A_{\text{col}}^\top$
 - 8: Aggregate with residual connection:
 $x_1 \leftarrow \gamma \cdot (out_{\text{row}} + out_{\text{col}}) + x_1$
 - 9: **end for**
 - 10: $x_2 \leftarrow \text{Conv}_{3 \times 3}(x_1) \rightarrow \text{BN} \rightarrow \text{Activation}$
 - 11: $x_{\text{cat}} \leftarrow \text{Concat}(x, x_2)$ ▷ Channel-wise fusion
 - 12: $y \leftarrow \text{Conv}_{3 \times 3}(x_{\text{cat}}) \rightarrow \text{BN} \rightarrow \text{Dropout}$
 - 13: **return** y
-

As illustrated in Figure 2 b), the RCCA module first applies a 3×3 convolution to compress the channel dimension of the input feature map. It then successively performs two CCA operations to model spatial dependencies. The resulting feature map is further processed by another 3×3 convolution, concatenated with the original input along the channel dimension, and passed through a final 3×3 convolution to match the target output dimension.

As shown in Figure 4, the RCCA propagation mechanism allows contextual information to be transferred between any two spatial positions. Even if two pixels are not located in the same row or column, indirect connections can be established through intermediate positions, thereby approximating global receptive field modeling.

With this dual-recurrent design, the RCCA module effectively captures dense and global contextual representations, producing feature maps with enhanced semantic expressiveness. The pseudocode of the RCCA module is presented in Algorithm 1.

3.4 Linear Deformable Convolution (LDConv)

LDConv, proposed in [26], aims to overcome the limitations of standard and deformable convolutions in terms of sampling flexibility and parameter efficiency. Traditional convolution operations rely on fixed, regular sampling grids that primarily focus on local regions, making it difficult to effectively capture global contextual information. Deformable convolution introduces learnable offsets to adjust the sampling positions of the original convolution kernel to a certain extent, partially alleviating the receptive field limitations of standard convolution. However, both methods share a common drawback: their sampling grids remain regular in structure, and as the kernel size increases, the number of parameters grows quadratically, leading to significant computational and memory overhead.

To address these issues, this paper incorporates the LDConv module into the decoder to replace the conventional double convolution structure. By leveraging a more flexible sampling strategy and a parameter structure with linear growth, the LDConv module enables more efficient feature modeling. According to the original paper, when the number of sampling points is set to $N = 5$, LDConv can effectively substitute the standard 3×3 convolution, offering improved computational efficiency and modeling flexibility. Therefore, this paper adopts $N = 5$ as a representative example, and Figure 5 illustrates the overall pipeline of the LDConv module. The feature extraction process of the module mainly consists of the following three steps:

1. **Initial Sampling Coordinates Generation.** Based on the number of sampling points N , initial sampling offsets P_n are generated using the sampling algorithm proposed in the original LDConv paper. These offsets are then added to the central reference coordinate P_0 , yielding the base sampling coordinates $P_0 + P_n$. This step constructs a set of sampling locations corresponding to a convolutional kernel with N sampling points on the feature map.

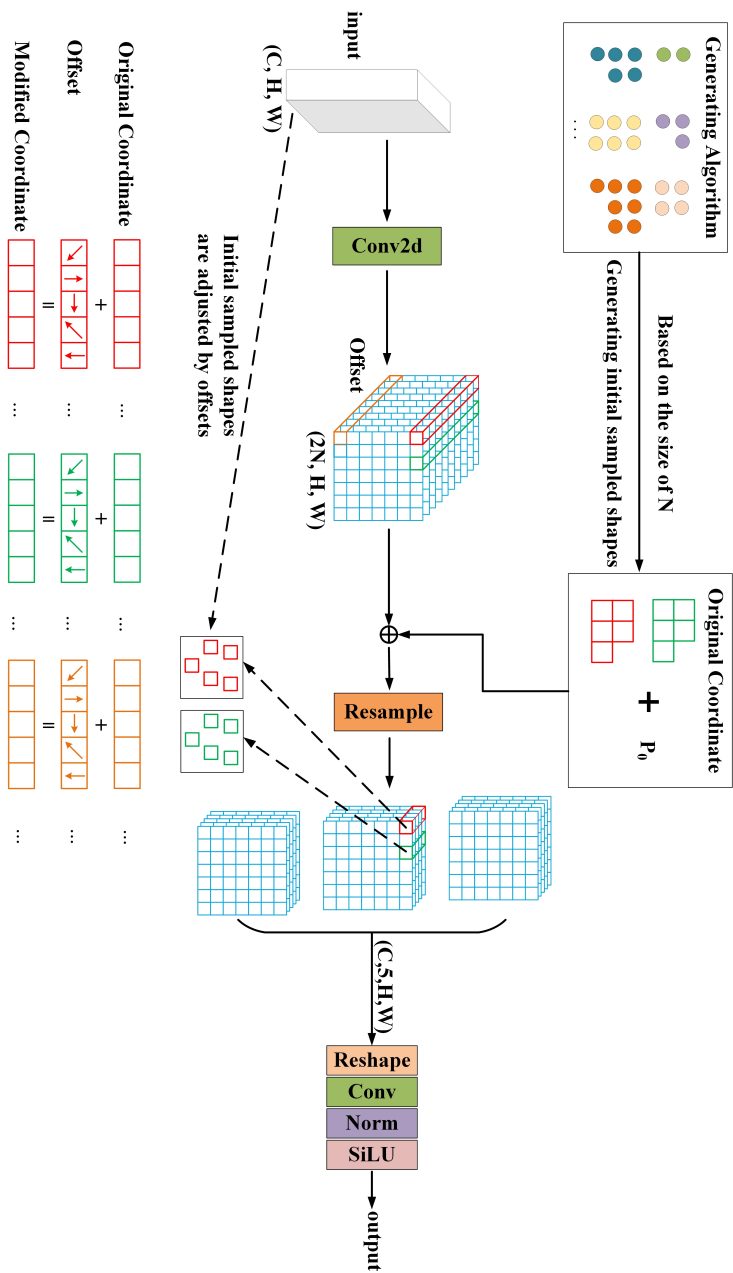


Figure 5. LDConv Module. It first assigns initial sampling coordinates to convolutions of arbitrary sizes and adjusts the sample shape with learnable offsets. Compared to the original coordinates, resampling changes the sampling shape at each position.

Algorithm 2 Pseudocode of the Linear Deformable Convolution (LDConv) module**Require:** Feature map $x \in \mathbb{R}^{B \times C \times H \times W}$, number of sampling points n_p , stride s **Ensure:** Output feature map y

- 1: $offset \leftarrow \text{Conv}_{3 \times 3}(x)$ ▷ Learn $2n_p$ offsets (x and y directions)
- 2: $p_0 \leftarrow$ generate regular grid positions based on H, W, s
- 3: $p_n \leftarrow$ generate static kernel shape offsets based on n_p
- 4: $p \leftarrow p_0 + p_n + offset$ ▷ Final sampling positions
- 5: **for** each sampling position p_i in p **do**
- 6: Compute 4 neighboring grid points: $q_{lt}, q_{rb}, q_{lb}, q_{rt}$
- 7: Compute bilinear interpolation weights: $g_{lt}, g_{rb}, g_{lb}, g_{rt}$
- 8: Sample features from x : $x_{q_{lt}}, x_{q_{rb}}, x_{q_{lb}}, x_{q_{rt}}$
- 9: Compute interpolated feature:

$$x_{offset_i} \leftarrow g_{lt} \cdot x_{q_{lt}} + g_{rb} \cdot x_{q_{rb}} + g_{lb} \cdot x_{q_{lb}} + g_{rt} \cdot x_{q_{rt}}$$
- 10: **end for**
- 11: Reshape x_{offset} along height dimension for grouped 1D convolution
- 12: $y \leftarrow \text{Conv1D} + \text{BN} + \text{SiLU}(x_{offset})$
- 13: **return** y

2. **Learnable Offset Adjustment.** A convolutional operation is employed to generate an offset tensor with dimensions (B, 2N, H, W), representing learnable 2D coordinate shifts for each of the N sampling points at every spatial location. These offsets are added to the initial sampling coordinates to obtain the final sampling locations, which are spatially adaptive and dynamically configured for each position in the feature map, enabling deformable kernel shapes.
3. **Feature Resampling and Output Generation.** Based on the adjusted sampling coordinates, features at the corresponding locations are obtained via interpolation and resampling. These sampled features are then processed through a sequence of convolution, normalization, and activation operations to produce the final output feature map.

Through the above three steps, the LDConv module enables the construction of convolutional kernels with an arbitrary number of sampling points, thereby achieving a flexible and efficient feature extraction mechanism. Algorithm 2 presents the pseudocode detailing the specific operations of the LDConv module.

4 EXPERIMENT

4.1 Datasets

To comprehensively evaluate the performance of the proposed model, we conducted experiments on three datasets: Synapse, ACDC, and MoNuSeg.

Synapse: The Synapse dataset consists of 30 abdominal CT scans, comprising a total of 3 779 axial clinical abdominal CT images. Each CT volume contains approximately $85 \sim 198$ slices with a resolution of 512×512 pixels and a voxel spacing of $([0.54 \sim 0.54] \times [0.98 \sim 0.98] \times [2.5 \sim 5.0]) \text{ mm}^3$. Following the protocol of TransUNet, we split the dataset into 18 training and 12 testing cases. We report the average DSC and HD for eight abdominal organs: aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, and stomach.

ACDC: ACDC is one of the challenge tasks from MICCAI 2017, aiming to segment the end-diastolic (ED) and end-systolic (ES) frames in cardiac cine-MRI. The challenge collected MRI scans from various patients, with each scan manually annotated with three anatomical labels: left ventricle (LV), right ventricle (RV), and myocardium (MYO). Following the processing strategy used in Perspective + UNet, we divided the ACDC dataset into a training set (70 cases), a validation set (10 cases), and a test set (20 cases), and evaluated the model's segmentation performance using the average DSC.

MoNuSeg: The MoNuSeg dataset is an important component of the multi-organ nuclei segmentation challenge held at MICCAI 2018. We selected 30 images from the dataset for training and 14 images for testing. Following the methodology of UCTransNet, we employed the DSC and IoU as evaluation metrics to assess the segmentation performance of our model.

4.2 Evaluation Metrics

We employed three evaluation metrics to evaluate the performance of our model: DSC, HD, and IoU. The prediction results can be categorized into true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), and DSC and IoU are calculated based on these values. The corresponding formulas are given below:

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN}, \quad (3)$$

$$IoU = \frac{TP}{TP + FP + FN}. \quad (4)$$

HD can be described as:

$$HD(Y, \hat{Y}) = \max \left\{ \max_{y \in Y} \min_{\hat{y} \in \hat{Y}} d(y, \hat{y}), \max_{\hat{y} \in \hat{Y}} \min_{y \in Y} d(y, \hat{y}) \right\}, \quad (5)$$

where Y and $\hat{Y}$ represent the ground truth and the predicted segmentation result respectively. $d(y, \hat{y})$ represents the Euclidean distance between Y and $\hat{Y}$.

4.3 Implementation Details

All experiments in this study were conducted on an NVIDIA RTX 3090 GPU equipped with 24GB of memory. To mitigate the risk of overfitting, a variety of data augmentation techniques were employed during training, including image flipping, Gaussian noise addition, random rotation and scaling, among others, to enhance the model's robustness and generalization capability. For the Synapse and ACDC datasets, input image resolution was uniformly set to 224×224 , with 150 training epochs and a batch size of 8. The model was optimized using stochastic gradient descent (SGD) with a learning rate of 0.05 and a weight decay coefficient of 0.0001. For the MoNuSeg dataset, we adopted the training strategy proposed in UTransNet, maintaining the same input resolution of 224×224 , with a maximum of 2000 training epochs and an early stopping mechanism (patience = 50). This task utilized the Adam optimizer with an initial learning rate of 0.001 and a batch size of 4.

4.4 Loss Function

For the Synapse and ACDC datasets, we employ a combined loss function of multi-class cross-entropy and Dice loss, while for the MoNuSeg dataset, we use a combination of binary cross-entropy and Dice loss. The cross-entropy loss function, which is suitable for both binary and multi-class classification tasks, measures the discrepancy between predicted values and true labels. It can be succinctly expressed as:

$$Loss_{CE} = - \sum_{i=1}^N y_i \log(\hat{y}_i). \quad (6)$$

Dice loss is a metric used to evaluate the similarity between two samples, particularly suited for image segmentation tasks due to its ability to handle imbalanced foreground and background pixels. A higher Dice coefficient indicates greater similarity between the samples. For each pixel, the Dice coefficient is calculated as per Equation (7), and the Dice loss is derived as per Equation (8). Both metrics effectively quantify the performance of the segmentation model.

$$Dice = \frac{2 \cdot \sum_{i=1}^N p_i \cdot g_i}{\sum_{i=1}^N p_i + \sum_{i=1}^N g_i}, \quad (7)$$

$$Loss_{Dice} = 1 - \frac{2 \cdot \sum_{i=1}^N p_i \cdot g_i}{\sum_{i=1}^N p_i + \sum_{i=1}^N g_i}. \quad (8)$$

Here, p_i represents the predicted probability for the i^{th} pixel, g_i denotes the true label for the i^{th} pixel, and N is the total number of pixels in the image. Thus, the loss function used for the Synapse and ACDC datasets is calculated as per Equation (9),

while for the MoNuSeg dataset, it is computed using Equation (10).

$$Loss = 0.4 \cdot Loss_{CE} + 0.6 \cdot Loss_{Dice} \quad (9)$$

$$Loss = 0.5 \cdot Loss_{CE} + 0.5 \cdot Loss_{Dice} \quad (10)$$

For Equation (9), the 0.4 and 0.6 represent the weights for the cross-entropy loss and dice loss, respectively. For Equation (10), we employ equal weights of 0.5 for both the binary cross-entropy loss and dice loss.

4.5 Results and Discussion

We evaluate the proposed RCLD-UNet model on three benchmark medical image segmentation datasets: Synapse, ACDC, and MoNuSeg. The evaluation includes comparisons with a wide range of representative methods covering various architectural paradigms, including CNN-based models (e.g., U-Net, Att-UNet), Transformer-based models (e.g., Swin-UNet, TransCepion), CNN-Transformer hybrid models (e.g., TransUNet, BRAU-Net++), Mamba-based architectures (e.g., VM-UNet), and CNN-Mamba hybrids (e.g., HC-Mamba).

4.5.1 Results on the Synapse Dataset

As shown in Table 5, we report the overall DSC, HD95, organ-wise Dice scores, and parameter counts. The proposed RCLD-UNet achieves the highest DSC of 84.14% while reducing HD95 to 18.79, indicating superior segmentation accuracy and boundary consistency. The visual comparisons in Figure 6a) further demonstrate the effectiveness of our model. In the first row, RCLD-UNet effectively reduces the over-segmentation of the stomach observed in U-Net; compared to other methods, our model exhibits less over-segmentation in the stomach region. In the second row, our method corrects the liver mis-segmentation issues observed in U-Net, Swin-Unet, and TransCepion. Moreover, it achieves slightly better segmentation performance for the pancreas and stomach compared to other methods. The third row shows that, compared to TransUNet, our method yields more precise boundary delineation for both the left and right kidneys. Although minor under-segmentation is observed in the stomach for some cases, our model achieves relatively complete gallbladder segmentation with only minimal over-segmentation. Overall, RCLD-UNet exhibits superior robustness and generalization across multiple critical organs.

4.5.2 Results on the ACDC Dataset

For the ACDC dataset, we evaluate Dice scores for the left ventricle, right ventricle, and myocardium, as well as the average DSC, as summarized in Table 6. Our model achieves the best performance across all categories, with an average DSC of

Methods	Params (M)	Avg. DSC↑	Avg. HD↓	Aor	Gal	LKid	RKid	Liv	Pan	Spl	Sto
U-Net [21]	14.80	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
R50 Atf-Unet [10]	–	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
Att-Unet [5]	34.88	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
R50 ViT [10]	–	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUNet [10]	105.28	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Swin-Unet [21]	27.17	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
TransDeepLab [44]	21.14	80.16	21.25	86.04	69.16	84.08	79.88	93.53	61.19	89.00	78.40
HiFormer [45]	25.51	80.39	14.70	86.21	65.69	85.23	79.77	94.61	59.52	90.99	81.08
MISSFormer [45]	42.46	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
ScaleFormer [44]	111.60	82.86	16.81	88.73	74.97	86.36	83.31	95.12	64.85	89.40	80.14
MCANet-B [46]	26.80	83.52	15.82	88.35	71.47	87.30	84.53	95.19	65.18	91.75	84.42
TransCeption [47]	38.14	82.24	20.89	87.60	71.82	86.23	80.29	95.01	65.27	91.68	80.02
BRAU-Net++ [45]	50.76	82.47	19.07	87.95	69.10	87.13	81.53	94.71	65.17	91.89	82.26
VM-U-Net [23]	–	81.08	19.21	86.40	69.41	86.16	82.76	94.17	58.80	89.51	81.40
Hc-Mamba [24]	–	79.58	26.34	89.93	67.65	84.57	78.27	95.38	52.08	89.49	79.84
RCLD-U-Net	20.12	84.14	18.79	90.24	68.60	86.47	84.90	95.16	70.32	93.01	84.40

Table 5. Comparison of segmentation performance of different methods on the Synapse dataset, with the best results highlighted in bold. The DSC metrics correspond to the aorta (Aor), gallbladder (Gal), left kidney (LKid), right kidney (RKid), liver (Liv), pancreas (Pan), spleen (Spl), and stomach (Sto), respectively.

Methods	Avg. DSC (%)	RV	Myo	LV
R50-U-Net [10]	87.55	87.10	80.63	94.92
R50 Att-UNet [10]	86.75	87.58	79.20	93.47
R50 ViT [21]	87.57	86.07	81.88	94.75
TransUNet [10]	89.71	88.86	84.53	95.73
Swin-UNet [21]	90.00	88.55	85.62	95.83
ScaleFormer [17]	90.17	87.33	88.16	95.04
UNETR [48]	88.61	85.29	86.52	94.02
UCTransNet [49]	89.69	87.92	85.43	95.71
MCTE [48]	91.31	89.14	89.51	95.27
MISSFormer [18]	91.19	89.85	88.38	95.34
UDTransNet [49]	90.11	89.06	86.10	95.18
RCLD-Unet	92.02	90.10	89.99	95.97

Table 6. Comparison of segmentation performance of different methods on the ACDC dataset, with the best results highlighted in bold

92.02%. As shown in Figure 6b), our method achieves more complete segmentation of the left ventricle and myocardium compared to other approaches, closely approximating the ground truth, while slight differences remain in the segmentation of the right ventricle. In the first row, both our method and the other methods exhibit some degree of under-segmentation in the right ventricle. In the second row, our method demonstrates relatively more complete segmentation of the right ventricle compared to R50-UNet and MISSFormer. In the third row, unlike R50-UNet, which fails to accurately capture the right ventricle, our method successfully segments it with high precision, further confirming its reliability in delineating cardiac structures.

Methods	DSC (%)	IoU (%)
U-Net [2]	75.48	62.27
UNet++ [3]	77.13	64.13
R34-Unet [2, 50]	78.45	65.30
Att-Unet [5]	76.06	62.08
Swin-UNet [21]	79.03	65.74
MISSFormer [18]	77.64	63.74
UCTransNet [16]	80.08	66.98
RCLD-Unet	80.84	68.03

Table 7. Comparison of segmentation results from different methods on the MoNuSeg dataset. The best results are highlighted in bold.

4.5.3 Results on the MoNuSeg Dataset

Table 7 presents quantitative comparisons on the MoNuSeg dataset, where our model achieves the highest scores in both DSC and IoU. The qualitative results

in Figure 6c) further support this finding. In the first row, our method significantly reduces background false positives compared to U-Net, though minor over-segmentation remains. In the second row, our method achieves more complete segmentation of the target nuclei, while other methods (such as U-Net, Swin-Unet, MISSFormer, and UCTransNet) exhibit varying degrees of over-segmentation or false positive regions. These results demonstrate the superiority of our method in modeling nuclear boundaries and improving segmentation precision in complex histopathological images.

4.6 Ablation Studies

To evaluate the impact of each module on model performance, we conducted a series of systematic ablation studies on the Synapse dataset, including investigations on the effectiveness of the introduced modules, the placement of RCCA modules, and the integration of LDConv modules.

Methods	Avg. DSC (%)	Avg. HD (mm)
Baseline (U-Net)	76.85	39.70
Baseline+RCCA	83.76	20.59
Baseline+LDConv	82.68	27.09
Baseline+LDConv+RCCA	84.14	18.79

Table 8. Evaluation of the effectiveness of progressively introducing RCCA and LDConv modules

4.6.1 Ablation Study on the Effectiveness of Individual Modules

As shown in Table 8, introducing the RCCA module into the U-Net encoder significantly improves model performance, with the DSC increasing to 83.76% and the HD decreasing to 20.59mm, despite an increase in the number of parameters. Furthermore, integrating the LDConv module into the decoder leads to a slight reduction in parameter count while also enhancing performance, achieving a DSC of 82.38% and reducing the HD to 28.54mm. When both RCCA and LDConv modules are incorporated simultaneously, the parameter count slightly decreases compared to using only the RCCA module, and the model achieves its best performance, with the DSC increasing to 84.14% and the HD further decreasing to 18.79mm.

4.6.2 Ablation Study on the Placement of RCCA Modules

As shown in Table 9, we evaluated the impact of inserting RCCA modules at different layers of the U-Net encoder and its bottleneck. In this experiment, RCCA modules were progressively added layer by layer, starting from the first encoder layer and extending to the entire encoder and bottleneck. The results indicate that as

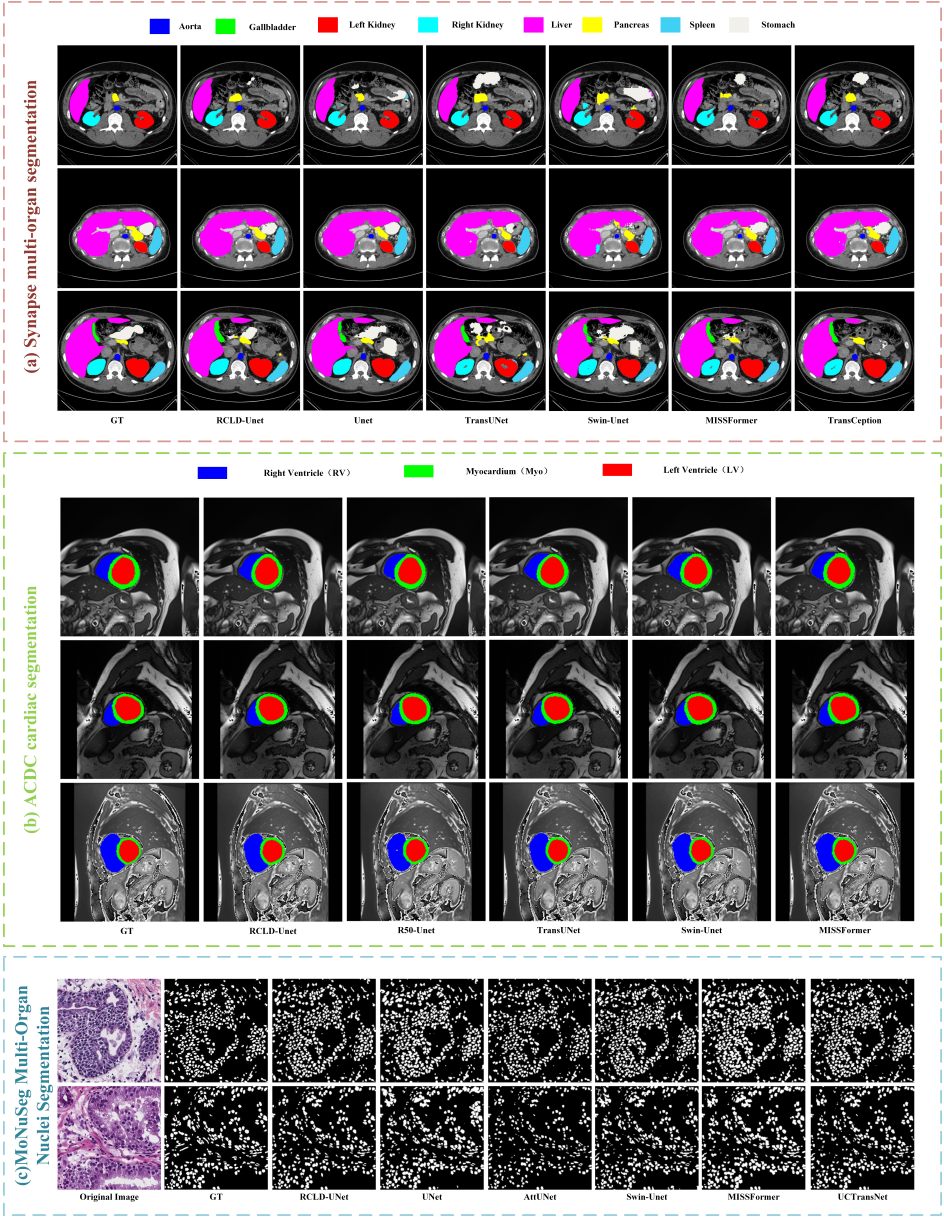


Figure 6. Comparison of visualization results of different methods on three datasets, with RCLD-UNet showing the best segmentation results on all of them: a) Synapse; b) ACDC; c) MoNuSeg.

Methods	Eco1	Eco2	Eco3	Eco4	Bottle	Avg. DSC (%)	Avg. HD (mm)	Params (M)
UNet + RCCA	✗	✗	✗	✗	✗	76.85	39.70	14.80
	✓	✗	✗	✗	✗	83.84	27.60	13.45
	✓	✓	✗	✗	✗	82.77	30.39	13.68
	✓	✓	✓	✗	✗	82.14	36.13	14.61
	✓	✓	✓	✓	✗	82.24	31.50	18.32
	✓	✓	✓	✓	✓	83.76	20.59	22.03

Table 9. Evaluation of RCCA placement in the encoder and bottleneck

more RCCA modules were introduced, the overall number of parameters gradually increased. When the RCCA module was applied only to the first encoder layer, the model achieved the highest DSC of 83.84 %. However, when RCCA modules were incorporated into all encoder layers and the bottleneck, the DSC slightly decreased to 83.76 %, while the HD was significantly reduced to its lowest value of 20.59 mm. Considering the trade-off between DSC and HD, we ultimately adopted the configuration with RCCA modules applied to all encoder layers and the bottleneck in the final model design.

Methods	Params (M)	Avg. DSC (%)	Avg. HD (mm)	Aor	Gal	LKid	RKid	Liv	Pan	Spl	Sto
U-Net	14.80	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
U-Dec1.LD	11.85	83.02	27.70	89.44	68.20	85.74	83.14	95.23	70.19	90.93	81.25
U-Dec12.LD	11.48	82.68	27.09	89.53	69.24	87.10	80.99	94.19	69.05	90.80	80.57
U-Dec4.LD	13.34	82.76	28.67	90.35	71.03	82.25	81.11	94.81	71.10	90.88	80.54

Table 10. Evaluation of LDConv integration into the decoder

4.6.3 Ablation Study on the Integration of LDConv Modules

As shown in Table 10, we evaluated the effect of integrating LDConv modules into different layers of the U-Net decoder. In this study, we define the deepest decoder layer (closest to the bottleneck) as Layer 1 and the shallowest layer (closest to the output) as Layer 4. When the LDConv module was applied to Layer 1 (U-Dec1.LD), the model achieved the highest DSC of 83.02 %. Upon further integration into Layer 2 (U-Dec12.LD), the DSC slightly decreased to 82.68 %, while the HD95 was further optimized to 27.09 mm. Due to computational resource limitations, we also tested the effect of applying the LDConv module only in Layer 4 (U-Dec4.LD), which yielded a DSC of 82.76 %, but with a less favorable HD95 compared to the deeper configurations. Considering both performance and model complexity, we ultimately selected the configuration with LDConv modules applied to Layers 1 and 2 as the final model design.

4.7 Limitations

Although the proposed model demonstrates strong performance in experimental evaluations, several aspects still warrant further improvement. First, while a favorable balance between accuracy and parameter count has been achieved, the current architecture has not yet been extended to incorporate more advanced paradigms such as Transformer-based frameworks or the emerging Mamba family of large-scale models. Second, the model has thus far been validated exclusively on 2D medical image segmentation tasks, without exploration in 3D imaging or multi-modal scenarios, which limits its applicability in real-world clinical settings. Furthermore, despite outperforming most baseline methods in terms of parameter efficiency, the model has yet to reach a truly lightweight level, and its inference speed and computational efficiency remain areas with significant potential for optimization.

5 CONCLUSION

This paper proposes a novel asymmetric U-shaped encoder-decoder architecture that integrates RCCA and LDConv to enhance feature modeling and segmentation accuracy in 2D medical image segmentation tasks. Built upon the U-Net framework, the model incorporates RCCA modules into the encoder to strengthen global contextual representation, while LDConv is employed in the decoder to replace conventional convolutions, alleviating spatial information loss during upsampling and improving the expression of complex structures and fine-grained details. Experimental results demonstrate that the proposed method achieves excellent segmentation performance and strong generalization across Synapse, ACDC, and MoNuSeg datasets. Notably, it shows outstanding performance in the segmentation of organs such as the spleen, pancreas, and heart, highlighting its potential in clinical applications, including abdominal organ analysis and cardiac structure delineation, thereby providing valuable support for medical image-based diagnosis and treatment planning. Future work may explore integrating RCCA modules into the decoder or skip connections to enhance long-range dependency modeling and mitigate the semantic gap between the encoder and decoder. In parallel, introducing LDConv into the encoder or bottleneck blocks could further improve the extraction of local structures and high-level semantic features, thereby optimizing overall model performance and adaptability.

REFERENCES

- [1] LONG, J.—SHELHAMER, E.—DARRELL, T.: Fully Convolutional Networks for Semantic Segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

- [2] RONNEBERGER, O.—FISCHER, P.—BROX, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.): *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer, Cham, Lecture Notes in Computer Science, Vol. 9351, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
- [3] ZHOU, Z.—RAHMAN SIDDIQUEE, M.M.—TAJBAKHS, N.—LIANG, J.: UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In: Stoyanov, D. et al. (Eds.): *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA 2018, ML-CDS 2018)*. Springer, Cham, Lecture Notes in Computer Science, Vol. 11045, 2018, pp. 3–11, doi: 10.1007/978-3-030-00889-5_1.
- [4] HUANG, H.—LIN, L.—TONG, R.—HU, H.—ZHANG, Q.—IWAMOTO, Y.—HAN, X.—CHEN, Y.W.—WU, J.: UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1055–1059, doi: 10.1109/ICASSP40776.2020.9053405.
- [5] OKTAY, O.—SCHLEMPER, J.—LE FOLGOC, L.—LEE, M.—HEINRICH, M.—MISAWA, K.—MORI, K.—MCDONAGH, S.—HAMMERLA, N.Y.—KAINZ, B.—GLOCKER, B.—RUECKERT, D.: Attention U-Net: Learning Where to Look for the Pancreas. *CoRR*, 2018, doi: 10.48550/arXiv.1804.03999.
- [6] CHEN, L.C.—PAPANDREOU, G.—KOKKINOS, I.—MURPHY, K.—YUILLE, A.L.: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, 2018, No. 4, pp. 834–848, doi: 10.1109/TPAMI.2017.2699184.
- [7] ZHAO, H.—SHI, J.—QI, X.—WANG, X.—JIA, J.: Pyramid Scene Parsing Network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6230–6239, doi: 10.1109/CVPR.2017.660.
- [8] SCHLEMPER, J.—OKTAY, O.—SCHAAP, M.—HEINRICH, M.—KAINZ, B.—GLOCKER, B.—RUECKERT, D.: Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images. *Medical Image Analysis*, Vol. 53, 2019, pp. 197–207, doi: 10.1016/j.media.2019.01.012.
- [9] DOSOVITSKIY, A.—BEYER, L.—KOLESNIKOV, A.—WEISSENBORN, D.—ZHAI, X.—UNTERTHINER, T.—DEHGHANI, M.—MINDERER, M.—HEIGOLD, G.—GELLY, S.—USZKOREIT, J.—HOULSBY, N.: An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*, 2020, doi: 10.48550/arXiv.2010.11929.
- [10] CHEN, J.—LU, Y.—YU, Q.—LUO, X.—ADELI, E.—WANG, Y.—LU, L.—YUILLE, A.L.—ZHOU, Y.: TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *CoRR*, 2021, doi: 10.48550/arXiv.2102.04306.
- [11] WANG, W.—CHEN, C.—DING, M.—YU, H.—ZHA, S.—LI, J.: TransBTS: Multimodal Brain Tumor Segmentation Using Transformer. In: de Bruijne, M., Catkin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.): *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Springer, Cham, Lecture Notes in Computer Science, Vol. 12901, 2021, pp. 109–119, doi:

- 10.1007/978-3-030-87193-2-11.
- [12] XIE, Y.—ZHANG, J.—SHEN, C.—XIA, Y.: CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. In: de Bruijne, M., Cattin, P. C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.): Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Springer, Cham, Lecture Notes in Computer Science, Vol. 12903, 2021, pp. 171–180, doi: 10.1007/978-3-030-87199-4-16.
 - [13] HATAMIZADEH, A.—TANG, Y.—NATH, V.—YANG, D.—MYRONENKO, A.—LANDMAN, B.—ROTH, H. R.—XU, D.: UNETR: Transformers for 3D Medical Image Segmentation. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1748–1758, doi: 10.1109/WACV51458.2022.001816.
 - [14] WANG, H.—XIE, S.—LIN, L.—IWAMOTO, Y.—HAN, X. H.—CHEN, Y. W.—TONG, R.: Mixed Transformer U-Net for Medical Image Segmentation. ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 2390–2394, doi: 10.1109/ICASSP43922.2022.9746172.
 - [15] AZAD, R.—HEIDARI, M.—WU, Y.—MERHOF, D.: Contextual Attention Network: Transformer Meets U-Net. In: Lian, C., Cao, X., Rekik, I., Xu, X., Cui, Z. (Eds.): Machine Learning in Medical Imaging (MLMI 2022). Springer, Cham, Lecture Notes in Computer Science, Vol. 13583, 2022, pp. 377–386, doi: 10.1007/978-3-031-21014-3-39.
 - [16] WANG, H.—CAO, P.—WANG, J.—ZAIANE, O. R.: UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-Wise Perspective with Transformer. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, No. 3, pp. 2441–2449, doi: 10.1609/aaai.v36i3.20144.
 - [17] HUANG, H.—XIE, S.—LIN, L.—IWAMOTO, Y.—HAN, X. H.—CHEN, Y. W.—TONG, R.: ScaleFormer: Revisiting the Transformer-Based Backbones from a Scale-Wise Perspective for Medical Image Segmentation. In: De Raedt, L. (Ed.): Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22). International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 964–971, doi: 10.24963/ijcai.2022/135.
 - [18] HUANG, X.—DENG, Z.—LI, D.—YUAN, X.: MISSFormer: An Effective Medical Image Segmentation Transformer. CoRR, 2021, doi: 10.48550/arXiv.2109.07162.
 - [19] VALANARASU, J. M. J.—OZA, P.—HACIHALILOGLU, I.—PATEL, V. M.: Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. In: de Bruijne, M., Cattin, P. C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.): Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Springer, Cham, Lecture Notes in Computer Science, Vol. 12901, 2021, pp. 36–46, doi: 10.1007/978-3-030-87193-2-4.
 - [20] XIE, E.—WANG, W.—YU, Z.—ANANDKUMAR, A.—ALVAREZ, J. M.—LUO, P.: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., Wortman Vaughan, J. (Eds.): Advances in Neural Information Processing Systems 34 (NeurIPS 2021). Curran Associates, Inc., 2021, pp. 12077–12090, https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf.

- [21] CAO, H.—WANG, Y.—CHEN, J.—JIANG, D.—ZHANG, X.—TIAN, Q.—WANG, M.: Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (Eds.): *Computer Vision – ECCV 2022 Workshops*. Springer, Cham, *Lecture Notes in Computer Science*, Vol. 13803, 2022, pp. 205–218, doi: 10.1007/978-3-031-25066-8_9.
- [22] MA, J.—LI, F.—WANG, B.: U-Mamba: Enhancing Long-Range Dependency for Biomedical Image Segmentation. *CoRR*, 2024, doi: 10.48550/arXiv.2401.04722.
- [23] RUAN, J.—LI, J.—XIANG, S.: VM-UNet: Vision Mamba UNet for Medical Image Segmentation. *CoRR*, 2024, doi: 10.48550/arXiv.2402.02491.
- [24] XU, J.: HC-Mamba: Vision MAMBA with Hybrid Convolutional Techniques for Medical Image Segmentation. *CoRR*, 2024, doi: 10.48550/arXiv.2405.05007.
- [25] HUANG, Z.—WANG, X.—HUANG, L.—HUANG, C.—WEI, Y.—LIU, W.: CC-Net: Criss-Cross Attention for Semantic Segmentation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 603–612, doi: 10.1109/ICCV.2019.00069.
- [26] ZHANG, X.—SONG, Y.—SONG, T.—YANG, D.—YE, Y.—ZHOU, J.—ZHANG, L.: LDConv: Linear Deformable Convolution for Improving Convolutional Neural Networks. *Image and Vision Computing*, Vol. 149, 2024, Art.No. 105190, doi: 10.1016/j.imavis.2024.105190.
- [27] ZHANG, Z.—LIU, Q.—WANG, Y.: Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, Vol. 15, 2018, No. 5, pp. 749–753, doi: 10.1109/LGRS.2018.2802944.
- [28] GU, Z.—CHENG, J.—FU, H.—ZHOU, K.—HAO, H.—ZHAO, Y.—ZHANG, T.—GAO, S.—LIU, J.: CE-Net: Context Encoder Network for 2D Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, Vol. 38, 2019, No. 10, pp. 2281–2292, doi: 10.1109/TMI.2019.2903562.
- [29] KHAN, A. R.—KHAN, A.: MaxViT-UNet: Multi-Axis Attention for Medical Image Segmentation. *CoRR*, 2023, doi: 10.48550/arXiv.2305.08396.
- [30] RAHMAN, M. M.—MUNIR, M.—MARCULESCU, R.: EMCAD: Efficient Multi-Scale Convolutional Attention Decoding for Medical Image Segmentation. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11769–11779, doi: 10.1109/CVPR52733.2024.01118.
- [31] HU, J.—SHEN, L.—SUN, G.: Squeeze-and-Excitation Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.
- [32] WOO, S.—PARK, J.—LEE, J. Y.—KWEON, I. S.: CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): *Computer Vision – ECCV 2018*. Springer, Cham, *Lecture Notes in Computer Science*, Vol. 11211, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.
- [33] LIU, X.—GAO, P.—YU, T.—WANG, F.—YUAN, R. Y.: CSWin-UNet: Transformer UNet with Cross-Shaped Windows for Medical Image Segmentation. *Information Fusion*, Vol. 113, 2025, Art.No. 102634, doi: 10.1016/j.inffus.2024.102634.
- [34] SHEN, Z.—ZHANG, M.—ZHAO, H.—YI, S.—LI, H.: Efficient Attention: Attention with Linear Complexities. 2021 IEEE Winter Conference on Applications of Com-

- puter Vision (WACV), 2021, pp. 3530–3538, doi: 10.1109/WACV48630.2021.00357.
- [35] GUO, M. H.—LU, C. Z.—LIU, Z. N.—CHENG, M. M.—HU, S. M.: Visual Attention Network. *Computational Visual Media*, Vol. 9, 2023, No. 4, pp. 733–752, doi: 10.1007/s41095-023-0364-2.
- [36] DAI, J.—QI, H.—XIONG, Y.—LI, Y.—ZHANG, G.—HU, H.—WEI, Y.: Deformable Convolutional Networks. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 764–773, doi: 10.1109/ICCV.2017.89.
- [37] HOWARD, A. G.—ZHU, M.—CHEN, B.—KALENICHENKO, D.—WANG, W.—WEYAND, T.—ANDREETTO, M.—ADAM, H.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR*, 2017, doi: 10.48550/arXiv.1704.04861.
- [38] HAN, K.—WANG, Y.—TIAN, Q.—GUO, J.—XU, C.—XU, C.: GhostNet: More Features from Cheap Operations. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1577–1586, doi: 10.1109/CVPR42600.2020.00165.
- [39] LI, D.—HU, J.—WANG, C.—LI, X.—SHE, Q.—ZHU, L.—ZHANG, T.—CHEN, Q.: Involution: Inverting the Inference of Convolution for Visual Recognition. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12316–12325, doi: 10.1109/CVPR46437.2021.01214.
- [40] LIU, Z.—MAO, H.—WU, C. Y.—FEICHTENHOFER, C.—DARRELL, T.—XIE, S.: A ConvNet for the 2020s. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11966–11976, doi: 10.1109/CVPR52688.2022.01167.
- [41] CHEN, L.—GU, L.—ZHENG, D.—FU, Y.: Frequency-Adaptive Dilated Convolution for Semantic Segmentation. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3414–3425, doi: 10.1109/CVPR52733.2024.00328.
- [42] YANG, J.—LIU, S.—WU, J.—SU, X.—HAI, N.—HUANG, X.: Pinwheel-Shaped Convolution and Scale-Based Dynamic Loss for Infrared Small Target Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39, 2025, No. 9, pp. 9202–9210, doi: 10.1609/aaai.v39i9.32996.
- [43] CHEN, L.—GU, L.—LI, L.—YAN, C.—FU, Y.: Frequency Dynamic Convolution for Dense Image Prediction. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 30178–30188, doi: 10.1109/CVPR52734.2025.02809.
- [44] AZAD, R.—NIGGEMEIER, L.—HÜTTEMANN, M.—KAZEROONI, A.—AGHDAM, E. K.—VELICHKO, Y.—BAGCI, U.—MERHOF, D.: Beyond Self-Attention: Deformable Large Kernel Attention for Medical Image Segmentation. 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1287–1297, doi: 10.1109/WACV57701.2024.00132.
- [45] LAN, L.—CAI, P.—JIANG, L.—LIU, X.—LI, Y.—ZHANG, Y.: BRAU-Net++: U-Shaped Hybrid CNN-Transformer Network for Medical Image Segmentation. *CoRR*, 2024, doi: 10.48550/arXiv.2401.00722.
- [46] SHAO, H.—ZENG, Q.—HOU, Q.—YANG, J.: MCANet: Medical Image Segmenta-

- tion with Multi-Scale Cross-Axis Attention. *Machine Intelligence Research*, Vol. 22, 2025, No. 3, pp. 437–451, doi: 10.1007/s11633-025-1552-6.
- [47] AZAD, R.—JIA, Y.—AGHDAM, E. K.—COHEN-ADAD, J.—MERHOF, D.: Enhancing Medical Image Segmentation with TransCeption: A Multi-Scale Feature Fusion Approach. *CoRR*, 2023, doi: 10.48550/arXiv.2301.10847.
- [48] HU, J.—CHEN, S.—PAN, Z.—ZENG, S.—YANG, W.: Perspective+ Unet: Enhancing Segmentation with Bi-Path Fusion and Efficient Non-Local Attention for Superior Receptive Fields. Vol. 15009, 2024, pp. 499–509, doi: 10.1007/978-3-031-72114-4_48.
- [49] WANG, H.—CAO, P.—YANG, J.—ZAIANE, O.: Narrowing the Semantic Gaps in U-Net with Learnable Skip Connections: The Case of Medical Image Segmentation. *Neural Networks*, Vol. 178, 2024, Art. No. 106546, doi: 10.1016/j.neunet.2024.106546.
- [50] HE, K.—ZHANG, X.—REN, S.—SUN, J.: Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.



Bing Su received his B.Sc. and his Ph.D. degrees from the Nanjing University of Aeronautics and Astronautics (NUAA), China. He is currently Associate Professor with the Department of Computer Science, School of Information and Mathematics, Changzhou University. His current research interests include network security, wireless sensor networks, the Internet of Things, routing protocols, and cloud computing.



Mengyu ZHAO received her Bachelor of Engineering degree in Computer Science and Technology from the Changzhou University, China, in 2022, and is currently pursuing her Master of Engineering degree in the same discipline at the same university. Her research focuses on medical image segmentation in the field of computer vision.



Yifeng LIN received his B.Sc. degree in software engineering from the Jilin University, Jilin, China, in 2005, his M.Sc. degree in software engineering from the Jilin University, Jilin, China, in 2007, and his Ph.D. degree in computer application technology from the Jilin University, Jilin, China, in 2012. His current research interest is deep learning.



Peng JIN received his B.Sc. degree in computer science and technology from the Wuzhou University, Wuzhou, China, in 2021, and now he is pursuing his M.Sc. degree in computer science and technology from the Changzhou University, Changzhou, China. His current research interests include semantic segmentation and time series model.