Bi-DLKA Unet: Merging Bi-level Routing Attention and Deformable Large Kernel Attention for Medical Image Segmentation

Chunfei Liu, Baoshan Sun

***Abstract*—In the field of medical image segmentation, deep learning-based methods have gained widespread recognition for their efficiency, with Transformer-based architectures proving particularly effective. However, these architectures are typically associated with high computational complexity and substantial memory requirements due to the self-attention mechanisms, which compute relationships between all tokens. In recent years, numerous studies have aimed to address this issue by introducing sparse attention mechanisms. These methods rely on artificial, content-agnostic sparse attention but still struggle to accurately capture long-range dependencies. In this study, we propose a novel network architecture, Bi-DLKA Unet, which incorporates dynamic sparse attention through bi-level routing Attention, thereby optimizing the allocation of computational resources to local feature maps that contribute most to the predictions. Within the Encoder-Decoder module, we integrate BiFormer, designed to enhance semantic information extraction and feature map resolution restoration. Additionally, in the Skip Connection segment, we introduce the Deformable Large Kernel Attention module, which combines the strengths of large convolutional kernels and deformable convolutions, allowing the model to better extract image features without incurring the high computational cost typical of traditional attention mechanisms. We rigorously evaluated the proposed Bi-DLKA Unet using the publicly available Synapse multi-organ segmentation dataset. Notably, our method showed statistically significant improvements in Dice coefficients, surpassing state-of-the-art algorithms such as MISSFormer by 0.51% in Dice coefficients.**

***Index Terms*—Medical Image Segmentation, Semantic Segmentation, Deep Learning, Deformable Large Kernel Attention.**

I. INTRODUCTION

Medical image segmentation is a crucial task within medical image analysis, aimed at extracting regions of interest from various types of medical images, including CT scans, MRI scans, and ultrasound images. This segmentation process facilitates the identification of important structures, such as organs, tumors, and blood vessels. Accurate and reliable medical image segmentation plays an indispensable role in computer-aided diagnosis and image-guided clinical surgeries. Moreover, it is significantly important for treatment planning, disease monitoring, and prognostic predictions for patients. Traditional medical image segmentation methods primarily depend on manual annotations by physicians or the manual design of digital image techniques. However, manually designed algorithms often struggle to meet the efficiency and accuracy requirements when processing large volumes of medical images, and the results of manual annotations are frequently influenced by subjective factors, such as the annotator's knowledge and experience. To overcome these challenges, deep learning-based methods for medical image segmentation have emerged, enhancing both accuracy and efficiency [1,2,3,4].

Transformer[5]-based network models have proven to be effective tools in the field of medical image segmentation, with their superior performance well-documented. However, these models face specific challenges: (1) while retaining the inherent feature extraction capabilities of Transformers, it is essential to mitigate the significant memory consumption resulting from the quadratic computational complexity of self-attention mechanisms; (2) effectively retaining spatial texture information within the context and facilitating the transfer of strongly correlated data to relevant modules also presents additional challenges. To tackle the first issue, researchers have proposed various sparse attention mechanisms [6,7,8], enabling each query to focus on a limited number of key-value pairs rather than the entire set. Nevertheless, existing methods either rely on manually designed static patterns or share sampled key-value pairs across all queries. Concerning the second challenge, some studies advocate incorporating attention mechanisms into the skip connection components to enhance the transmission of contextual and spatial texture features. In summary, although Transformer-based models exhibit considerable promise in medical image segmentation, optimizing memory usage and retaining spatial information remain critical areas of research.

In this paper, we propose a deep learning network architecture named Bi-DLKA Unet for conducting 2D medical image segmentation tasks. The Bi-DLKA Unet comprises the Biformer [9] module, the RA [10] module, and the deformable large-kernel convolution [11] module. The encoder, composed of Biformer , is tasked with extracting semantic information from the input feature map, while the decoder, constructed from Biformer and RA, utilizes the features extracted by the encoder to restore the resolution of the feature map. D-LKA effectively conveys contextual information by integrating deformable convolution with large-kernel convolution for feature extraction. It compensates for the loss of spatial texture information during the encoder's computation, bridging the semantic gap between the encoder and decoder and facilitating feature fusion in the decoder. Our contributions can be summarized as follows:

1) We integrated a dynamic, query-aware sparse attention

Manuscript received December 15, 2024

Chunfei Liu, School of Computer Science and Technology, Tiangong University, Tianjin, China
Baoshan Sun, School of Computer Science and Technology, Tiangong University, Tianjin, China

mechanism into a symmetric U-shaped neural network architecture. The encoder employs the Bi-level Routing Attention mechanism and Reverse Attention to extract semantic information from the input image. Subsequently, in the decoder, the extracted features are upsampled to match the input resolution, enabling accurate pixel-level segmentation predictions.

2) We introduced the D-LKA module, which leverages the advantages of large convolution kernels and deformable convolutions to more effectively capture complex spatial relationships. The D-LKA module incorporates dynamic attention over extensive spatial regions, adapting to variations in object scales.

3) Numerous experiments have validated the effectiveness of our proposed Bi-DLKA Unet in enhancing segmentation accuracy for CT and MRI medical images. The skip fusion module successfully integrates multi-scale feature information, bridging the semantic gap between the encoder and decoder.

II. Related Work

A. Using Attention Mechanisms in U-Net

U-Net [12], a deep learning architecture originally developed for biomedical image segmentation, has become a standard approach due to its capacity to effectively preserve spatial information through skip connections. These connections enable the decoder to utilize feature maps from the encoder, thereby enhancing segmentation accuracy. Recent studies have introduced attention mechanisms within the skip connections of U-Net to improve model performance by concentrating on significant regions of the image and suppressing irrelevant information.

Among the most commonly employed techniques is the channel attention mechanism, which prioritizes informative channels while diminishing less relevant ones. This mechanism is exemplified in models such as Attention U-Net [13], which integrates channel-wise attention into its skip connections. By applying this mechanism, Attention U-Net enhances focus on vital features, particularly in challenging tasks like tumor or organ segmentation. However, while channel attention proves beneficial, it often struggles to effectively capture spatial dependencies, especially in medical images with complex anatomical variations.

The spatial attention mechanism, on the other hand, directs focus to different regions within an image, emphasizing critical areas for the segmentation task. The Convolutional Block Attention Module (CBAM) [14] integrates both channel and spatial attention, allowing the model to dynamically prioritize the most relevant regions. The spatial attention map is generated via a 2D convolution applied to the feature map, enabling the network to attend to high-relevance areas. Despite improvements in segmentation across various settings, CBAM still faces challenges in addressing global contextual dependencies, particularly in large images with varying object sizes and shapes.

Another significant advancement is the Squeeze-and-Excitation (SE) block [15], which adaptively recalibrates feature maps by modeling interdependencies between channels. The SE block enhances U-Net′ s capability to emphasize important feature channels, thus improving segmentation accuracy. Nevertheless, this block primarily operates at the channel level and may not fully capture the complex spatial relationships essential in medical imaging.

Additionally, Polarized Self-Attention [16], which introduces polarization to self-attention mechanisms to prioritize relevant spatial and channel features, has been utilized in recent studies, including YOLOV10. While self-attention effectively captures global context and long-range dependencies, it incurs high computational costs and memory requirements, especially with large images and dense feature maps.

Despite the achievements of these attention mechanisms, they exhibit inherent limitations. While channel attention mechanisms effectively emphasize critical feature channels, they do not comprehensively address spatial dependencies. Conversely, spatial attention mechanisms enhance focus on regions of interest but may still struggle to capture long-range dependencies and global context. Integrating these mechanisms into U-Net increases computational complexity, notably in the case of self-attention and multi-scale attention approaches, which demand substantial memory and computational resources.

To address these challenges, we propose introducing Deformable Large Kernel Attention (DLKA), which aims to combine the advantages of large convolution kernels and deformable convolutions to capture complex spatial relationships more effectively. DLKA introduces dynamic attention over extensive spatial regions while adapting to the variable nature of object scales, a common challenge in medical image segmentation. This mechanism allows the model to concentrate on both local and global contexts, thereby improving its ability to segment intricate anatomical structures. By mitigating the computational overhead associated with traditional self-attention mechanisms, DLKA offers a more efficient alternative capable of effectively handling large and diverse medical images.

B. Reconstructing U-Net with Optimized Transformer Modules

In recent years, the integration of Transformer modules into the U-Net architecture has emerged as a promising strategy for improving the performance of medical image segmentation tasks. Transformer-based models have the advantage of capturing long-range dependencies and contextual information, which are crucial for accurately segmenting complex medical images. Several optimized Transformer models have been developed and successfully incorporated into the U-Net framework, leading to significant advancements in segmentation accuracy.

One prominent example is Swin UNet [17], which uses the Swin Transformer as its backbone. By employing a hierarchical structure with shifted-window-based self-attention, Swin U-Net effectively captures both local and global features in medical images. This enables the model to preserve fine spatial details while also benefiting from global context, making it particularly effective for complex segmentation tasks, such as segmenting tumors or organs with intricate boundaries.

SegFormer [18] proposes a lightweight and efficient framework for semantic segmentation. By reducing the data volume prior to self-attention computation, it decreases computational load while maximizing the retention of the network's feature extraction capabilities. The framework effectively captures multiscale features and delivers robust performance. Furth ermore, it eliminates complex decoder architectures, thereby enhancing efficiency. However, it does not resolve the quadratic complexity issue associated with the self-attention mechanism, which may still result in significant computational overhead when tackling tasks of substantial scale.

HiFormer [19] features two key innovations: first, it integrates CNN and transformer modules at shallow network levels, facilitating efficient fusion of local and global features; second, the Double-Level Fusion (DLF) module enhances feature reusability and consistency. This design enables HiFormer to achieve outstanding performance in medical image segmentation, particularly in balancing fine details and long-range dependencies. However, the model exhibits limitations in low-contrast images (e.g., skin lesions), indicating areas for further improvement.

Furthermore, Swin UNETR [20] is a state-of-the-art 3D brain tumor segmentation model that combines Swin Transformer and U-Net architecture. It utilizes hierarchical shifted windows for long-range dependency modeling, enhancing feature extraction across multiple resolutions. Despite its performance, Swin UNETR’s main limitation is the high computational cost and memory usage due to the transformer’s complexity, making it challenging for resource-constrained environments.

Despite the successes of these methods, challenges remain in efficiently capturing sparse features and addressing the computational overhead associated with traditional self-attention mechanisms. To overcome these limitations, we propose Biformer, a Transformer-based model that utilizes a dynamic sparse attention mechanism. Biformer aims to reduce the computational burden by selectively attending to the most relevant features, offering a more efficient and scalable solution for complex medical image segmentation tasks.

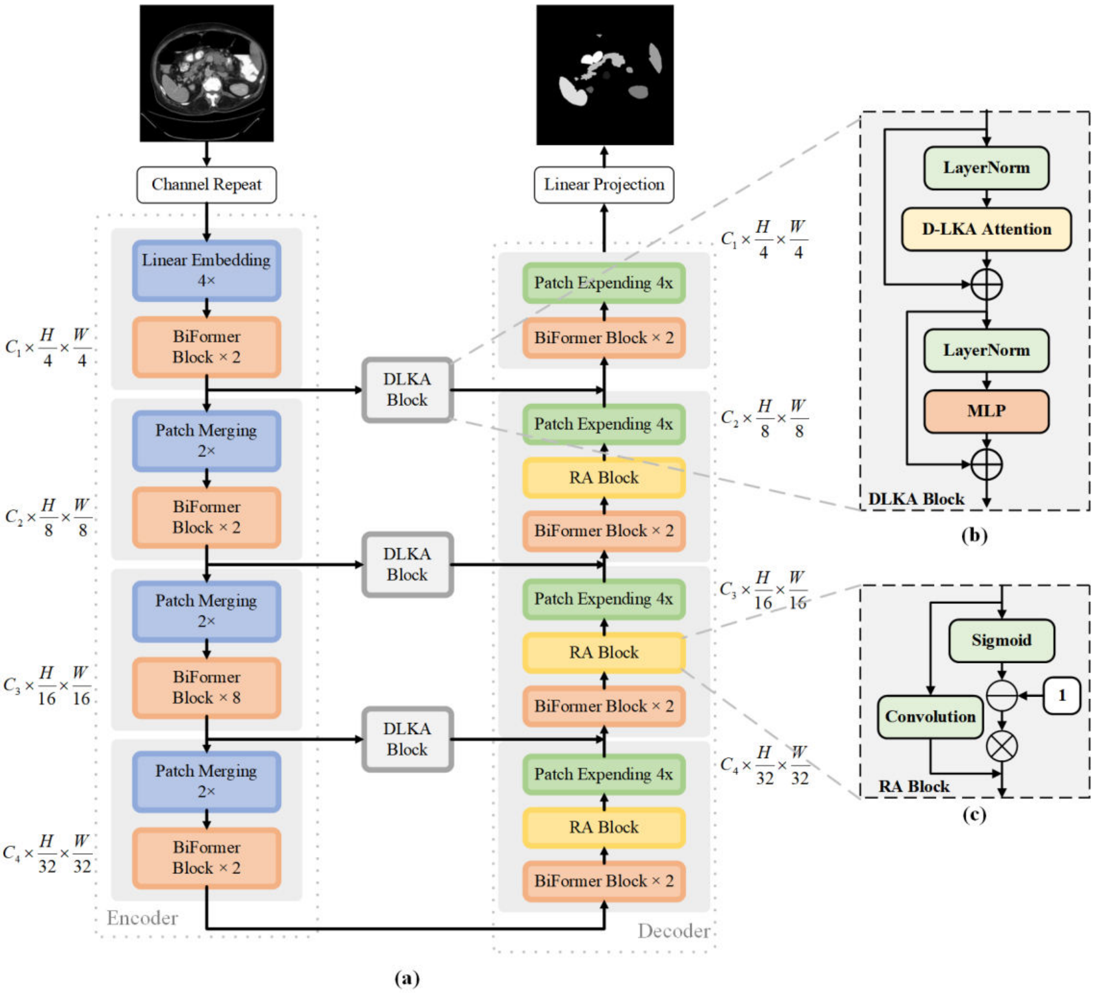

Fig.1 Overview of Bi-DLKA Unet

III. Methods

A. Overview of the Bi-DLKA Unet

The architecture of the proposed Bi-DLKA Unet is illustrated in Fig.1(a). The core component of the network architecture is an Encoder-Decoder structure based on BiFormer modules. In the Decoder section, to enhance the model's ability to segment image edges, a Reverse Attention mechanism module is added after each Decoder module. Additionally, in the skip connection section, D-LKA is incorporated to strengthen its capability to extract and enhance spatial texture information.

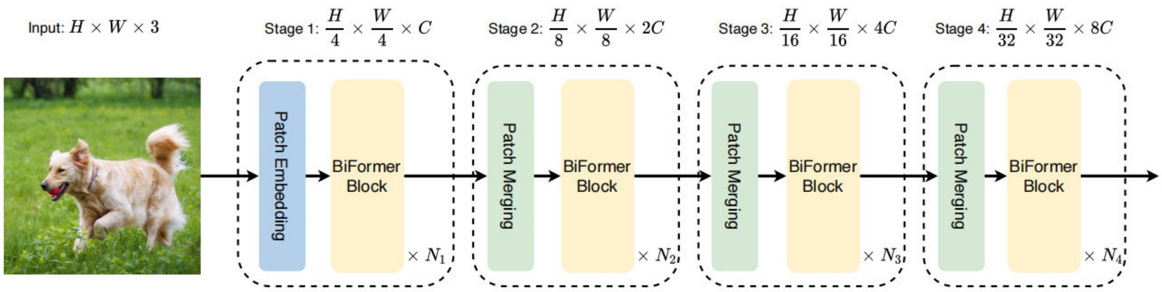

Fig.2 Structure of Biformer Encoder

B. *Biformer*

The computational complexity associated with conventional self-attention mechanisms poses significant scalability challenges when dealing with large datasets. To address this issue, various studies have introduced different sparse attention mechanisms that focus on a limited number of key-value pairs for each query, rather than attending to all of them. Nonetheless, many of these approaches either rely on manually designed static patterns or utilize a shared set of sampled key-value pairs across all queries.

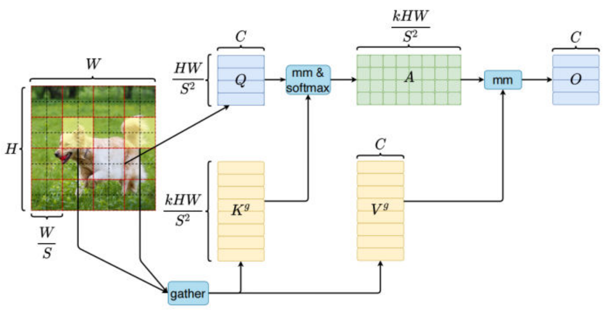

Fig.3 Bi-level Routing Attention (BRA)

Biformer represents a vision transformer model that features an innovative attention mechanism termed bi-level routing attention (BRA).The framework of Biformer Encoder is illustrated in Fig.2 This mechanism is dynamic and query-aware, allowing for sparse attention. The operational framework of BRA is illustrated in Fig.3. The BRA algorithm consists of two main stages: initially, it eliminates the most irrelevant key-value pairs at a coarse region level, preserving only a minimal subset of routed regions. Subsequently, a detailed token-to-token attention mechanism is executed within the union of these routed areas. The BRA can be mathematically expressed when given a 2D input feature map $X \in \mathbb{R}^{H\times W\times C}$:

$$
\begin{aligned}
&X^r = \mathrm{Partition}(X),\\
&Q = X^r W^q, K = X^r W^k, V = X^r W^v,\\
&Q^r = RegionAverage(Q),\\
&K^r = RegionAverage(K),\\
&A^r = Q^r (K^r)^T,\\
&I^r = TopkIndex(A^r),\\
&K^g = gather(K, I^r), V^g = gather(V, I^r),\\
&O = Attention(Q, K^g, V^g) + LCE(V).
\end{aligned}
$$

where X^r is $S \times S$ pathes reshape from X. W^q,W^k,$W^v \in \mathbb{R}^{C\times C}$ are projection weights for the query, key, value respectively. $A^r \in \mathbb{R}^{S^2\times S^2}$represents adjacency matrix of region-to-region affinity graph derived from $Q^r$and K^r. $I^r \in \mathbb{R}^{S^2\times k}$ represents routing index matrix. $K^g, V^g \in \mathbb{R}^{S^2\times\frac{kHW}{S^2}\times C}$ are gathered key and value tensor. Further more, a local context enhancement term LCE(·), parametrized with a 5×5depth-wise convolution, is introduced into BiFormer.

BiFormer operates by concentrating exclusively on a limited set of relevant tokens during query processing, thereby avoiding interactions with unrelated tokens. This characteristic makes it especially well-suited for dense prediction tasks, as it can adeptly retrieve semantic information from the original image. Additionally, BiFormer demonstrates commendable performance along with high computational efficiency.

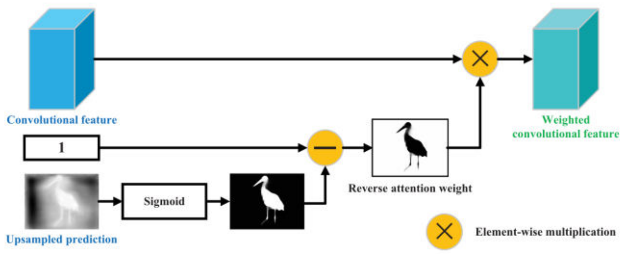

Fig.4 Reverse Attention

C. *Self - Reverse Attention*

The primary goal of the reverse attention mechanism is to direct the network's focus towards areas and object details that have not been sufficiently captured during the learning process. This is accomplished by dynamically adjusting the intermediate features within the network. The process begins with feature extraction from multiple layers, with particular focus on shallow features, which retain fine-grained details but may lack comprehensive semantic context. Before initiating side-output residual learning, the mechanism applies an elimination process to the predicted salient regions, reducing the influence of high-scoring areas in the saliency map derived from the shallow features. This crucial step "clears out" already detected regions, allowing the network to refocus on unrecognized areas and finer details. Through this strategic elimination, the reverse attention block facilitates the transmission of guidance from higher to lower layers, helping the network concentrate on previously overlooked sections of the image. As a result, the network gains a more nuanced understanding of the scene, improving both resolution and accuracy in the final saliency map. Overall, the reverse attention mechanism plays a pivotal role in side-output residual learning, providing r efined guidance that significantly enhances the quality and efficiency of salient object detection. The schematic diagram of the reverse attention mechanism is shown in the Fig.4.

The initial reverse attention mechanism operates across modules at various semantic levels, whereas the proposed self-reverse attention mechanism specifically targets overlooked regions within the same semantic level of the image, thereby enhancing segmentation cues derived from the image itself. The schematic diagram of the self-reverse attention mechanism is shown in the Fig.1(c).

D. *Deformable Large Kernel Attention*

Deformable Large Kernel Attention (D-LKA) is a novel attention mechanism designed to efficiently capture both local and global contextual information in medical image segmentation. The core idea behind D-LKA is to combine the benefits of large convolution kernels and deformable convolutions, enabling the model to better represent volumetric data without incurring the high computational costs typically associated with traditional attention mechanisms. D-LKA works by dynamically adjusting the sampling grid via deformable convolutions, allowing the model to flexibly adapt its receptive field to different data patterns, especially those with irregular structures such as organs or lesions in medical images. The structural diagram of D-LKA is shown in Fig.1(b) and Fig.5.

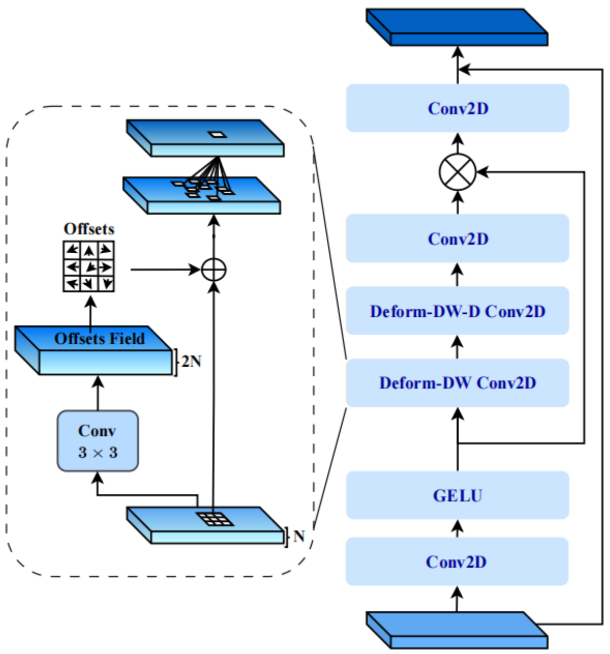

Fig.5 Architecture of the deformable LKA module

In the 2D version of D-LKA, deformable convolutions replace standard convolutions to capture shape variations, which is particularly crucial for medical image segmentation where objects often have complex and irregular forms. The D-LKA mechanism avoids conventional normalization functions, such as sigmoid or softmax, which can lead to the loss of high-frequency details, thereby preserving important fine-grained information.

By combining large kernel sizes with deformable sampling grids, D-LKA achieves a balance between computational efficiency and the ability to capture rich contextual information, making it highly effective for tasks such as medical image segmentation where both local and global dependencies are essential.

Incorporating Deformable Large Kernel Attention (D-LKA) into the skip connections of the U-Net architecture significantly enhances the model's ability to transfer contextual and spatial texture information across different layers. Traditionally, U-Net utilizes skip connections to directly propagate high-resolution feature maps from the encoder to the decoder, facilitating the recovery of details lost during downsampling. By replacing conventional convolution operations in the skip connections with D-LKA, the model gains the capability to adaptively focus on relevant spatial regions while capturing long-range dependencies within the feature maps. D-LKA's deformable sampling grid allows the network to dynamically adjust the receptive fields, effectively addressing the variability in spatial structures present in medical images. This adaptive attention mechanism not only preserves fine-grained texture details but also enriches the contextual understanding of the segmented regions, leading to better delineation of complex anatomical structures and improved segmentation performance overall.

E. *Loss Fuction*

To measure the discrepancy between the ground truth and model predictions, we utilize a hybrid loss function that combines Dice Loss [21] and Cross-Entropy Loss, following the design of Swin-Unet. The combined loss function is formulated as:

$$\mathrm{Loss} = \alpha \mathrm{Loss}_{\mathrm{Dice}} + (1-\alpha)\mathrm{Loss}_{\mathrm{CE}}$$

where α is a weighting hyperparameter, set to 0.6. The Dice Loss between two binary volumes is defined as:

$$Loss_{Dice} = 1 - \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i + \sum_i^N g_i}$$

where, the summation runs over all N voxels, with p_i representing the predicted binary segmentation volume and g_i denoting the ground truth binary volume.

The Cross-Entropy Loss for two binary volumes is expressed as:

$$Loss_{CE} = -\frac{1}{N} \times \sum_i \sum_{c=1}^{M} y_{ic} log(p_{ic})$$

where M is the total number of categories. y_{ic} is the ground truth indicator function, which is 1 if the true class of sample i is c and 0 otherwise. p_{ic} is the predicted probability of sample i belonging to class c.

In contrast to the conventional U-Net architecture, which uses a bottleneck structure between the Encoder and Decoder to preserve semantic information, our approach eliminates this module. Instead, we directly forward the final output of the Encoder to the Decoder.

IV. EXPERIMENT

A. *DataSet*

Experiments were performed utilizing the Synapse multi-organ segmentation dataset, which comprises 30 abdominal CT scans containing a total of 3,779 axial abdominal clinical images. Aligning with the methodology established in Swin-Unet, the dataset was randomly partitioned into 18 scans desi gnated for training and 12 for testing. Our evaluation metrics included the average Dice-Similarity Coefficient (DSC) and the average Hausdorff Distance (HD), focusing on eight specific abdominal organs: the aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, and stomach.

B. *Result*

We compare our proposed Bi-DLKA Unet with state-of-the-art methods on the Synapse multi-organ CT dataset (see Table 1). Classic networks, such as Unet, TransUnet, and Swin Unet, are used as benchmarks. Our method outperforms others in terms of average Dice Similarity Coefficient (DSC) and average Hausdorff Distance (HD). Specifically, our final results achieve 82.47% in DSC and 19.65 in HD. Compared to MISSFormer, our method shows improvements of 0.51% in DSC. Compared to Swin Unet, our method shows improvements of 1.9 in HD. Furthermore, Bi-DLKA Unet consistently surpasses previous best results across three distinct organs, with the most significant enhancement observed in Pancreas segmentation (2.5%). Our experiment validates the effectiveness of bridging the semantic gap between the Encoder and Decoder by incorporating variable large kernel convolutions within the Skip Connection structure, as well as enhancing edge segmentation capability in the Decoder. Fig.6 demonstrates our superior classification accuracy and enhanced ability to address overfitting and boundary segmentation issues, achieving significant improvements compared to previous architectures.

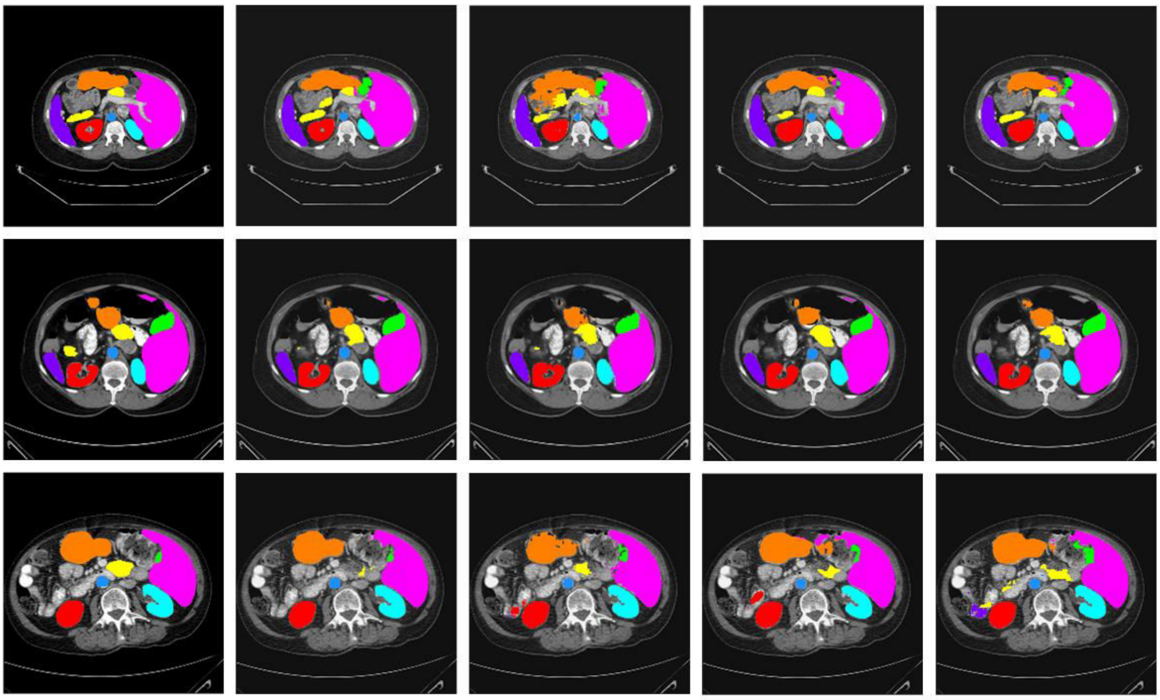

Fig.6 Result of our method on Synapse Dataset

Table 1 Result of our method on Synapse Dataset

Method	DSC	HD	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
U-net	76.85	39.70	89.07	**69.72**	77.77	68.60	93.43	53.98	86.67	75.58
V-Net	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
Att-UNet	77.77	36.02	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	75.75
TransUnet	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
UCTransNet	78.99	30.29	-	-	-	-	-	-	-	-
Swin Unet	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
MISSFormer	81.96	18.20	86.99	68.65	**85.21**	82	94.41	65.67	**91.92**	**80.81**
Our Method	**82.47**	**19.65**	87.65	69.64	85.09	**82.63**	**94.76**	**68.17**	91.29	80.54

C. *Ablation study*

To evaluate the effectiveness of BiFormer, RA, and D-LKA in enhancing network performance, we sequentially incorporated these three modules into the Swin Unet architecture and trained the network on the Synapse dataset to assess its segmentation performance. To control for variables, we removed the bottleneck structure from the original Swin Unet. As shown in Table 2, the experimental results demonstrate that the addition of the BiFormer, RA, and D-LKA modules led to incremental improvements in the network's segmentation performance, as measured by the DSC metric, compared to the original Swin Unet. These findings further substantiate that the proposed modules effectively enhance U-Net-based architectures.

Table 2 Result of our method on Synapse Dataset

Method	DSC(%)
Swin-Unet	79.13
Bi-Unet	81.83
Bi-Unet + RA	82.06
Bi-Unet + RA + D-LKA	82.47

V. Conclusion

In our study, we propose a novel network architecture nam ed Bi-DLKA Unet. This architecture integrates dynamic spar se attention mechanisms and bi-level routing mechanisms to enhance feature extraction capabilities, adaptability, and cont ext-aware attention mechanisms based on deformable convol ution and large-kernel convolution. Within the encoder-deco der module, we introduce BiFormer and Reverse Attention t o improve semantic information extraction and feature map r esolution restoration. Additionally, in the skip connection se gment, we include the D-LKA module, which highlights cont extual and spatial texture information from the original input feature maps. Our research is of significant importance for i mproving the efficiency and accuracy of medical image seg mentation, with potential applications in computer-aided dia gnosis and image-guided clinical surgery. However, challeng es still exist. In our future research, we will explore lower-co mplexity methods to enhance the segmentation performance of the network while reducing its computational complexity. The key to this method lies in addressing the quadratic comp utational complexity issue of self-attention mechanisms, utili zing algorithms with linear time complexity, such as Mamba, to extract contextual information.

Acknowledgment

This work is partically supported by Natural Science Foun dation of China Grants(61972456, 61173032)and Tianjin Nat ural Science Foundation (20JCYBJC00140).

References

[1] R. Azad, A. Kazerouni, M. Heidari, E. K. Aghdam, A. Molaei, Y. Jia, et al., "Advances in medical image analysis with vision Transformers: A comprehensive review," *Medical Image Analysis,* vol. 91, 2024, pp. 103000.

[2] Yao, W., Bai, J., Liao, W. et al,"From CNN to Transformer: A Review of Medical Image Segmentation Models," *Journal of Imaging Informatics in Medicine*, 2024, pp.1529-1547.

[3] Khan, Asifullah et al, "A Recent Survey of Vision Transformers for Medical Image Segmentation," ArXiv, 2023, abs/2312.00634.

[4] Anusha Aswath, Ahmad Alsahaf, Ben N.G. Giepmans, et al. "Segmentation in large-scale cellular electron microscopy with deep learning: A literature survey,"*Medical Image Analysis*, 2023, vol.89, pp.102920

[5] Vaswani A , Shazeer N , Parmar N ,et al., "Attention Is All You Need,"*Proceedings of the 31st International Conference on Neural Information Processing Systems 2017*, 2017, pp.6000-6010

[6] Ze Liu, Yutong Lin, Yue Cao, et al., "Swin transformer: Hierarchical vision transformer using shifted windows,"*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp.10012-10022

[7] Zhengzhong Tu, Hossein Talebi, Han Zhang,et al., "Maxvit: Multi-axis vision transformer,"*ECCV*, 2022.

[8] Wenxiao Wang, Lu Yao, Long Chen, et al., "Crossformer: A versatile vision transformer hinging on cross-scale attention,"*International Conference on Learning Representations, ICLR*, 2022.

[9] Lei Zhu, Xinjiang Wang, Zhanghan Ke, etal., "BiFormer: Vision Transformer with Bi-Level Routing Attention,"*2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp.10323-10333.

[10] Huang Q , Xia C , Wu C ,et al., "Semantic Segmentation with Reverse Attention,"*BMVC 2017*, 2017.

[11] Reza Azad, Leon Niggemeier, Michael Huttemann, et al., "Beyond Self-Attention: Deformable Large Kernel Attention for Medical Image Segmentation,"*2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp.1276-1286.

[12] Olaf Ronneberger, Philipp Fischer, Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation,"*Medical Image Computing and Computer-Assisted Intervention(MICCAI 2015)*, 2015, vol.9351, pp.234-241

[13] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, et al., "Attention U-Net: Learning Where to Look for the Pancreas,"*Arxiv*, 2018.

[14] Sanghyun Woo, Jongchan Park, Joon-Young Lee,et al., "CBAM: Convolutional Block Attention Module,"*ECCV 2018*, 2018, vol.11211, pp.3-19

[15] Jie Hu, Li Shen, Gang Sun, "Squeeze-and-Excitation Networks,"*M2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp.7132-7141.

[16] Huajun Liu, Fuqiang Liu, Xinyi Fan, et al.,"Polarized Self-Attention: Towards High-quality Pixel-wise Regression,"*Neurocomputing*, 2022, vol.506, pp.158-167.

[17] Hu Cao, Yueyue Wang, Joy Chen, et al., "Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation,"*Computer Vision-ECCV 2022 Workshops Cham 2023*, 2023, vol.13803, pp.205-218

[18] Enze Xie, Wenhai Wang, Zhiding Yu, et al., "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers,"*NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2024, No.924, pp.12077-12090.

[19] Xixin Wu, Hui Lu, Kun Li, et al., "Hiformer: Sequence Modeling Networks With Hierarchical Attention Mechanisms,"*IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023, vol.31, pp.3993-4003.

[20] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, et al., "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images,"*BrainLes 2021*, 2021,vol.12962, pp.272-284

[21] F. Milletarì, N. Navab, Seyed-Ahmad Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation,"*2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp.565-571.

Chunfei Liu, Graduate student. He is currently pursuing a master's degree in computer science and technology at Tiangong University, Tianjin, China.His research interests include computer vision, medical image segmentation, nuclear magnetic resonance images, deep learning, and semantic segmentation. His current focus is on semi-supervised medical image segmentation of cardiac MRI images.

Baoshan Sun, associate professor, holds a doctoral degree in engineering, and serves as a master's supervisor. Visiting scholar sent by the state to study abroad in the UK and member of the CCF China Computer Society. Selected for the "Outstanding Young Teacher Funding Program" in Tianjin universities, awarded the title of Excellent Guidance Teacher for Engineering Majors in Tianjin, and awarded the title of Excellent Class Mentor at Tianjin University of Technology. In recent years, leading a research team has achieved a series of teaching and research achievements in the field of teaching and research. He has published over 30 academic papers in SCI journals, EI journals, and international conferences. Hosted and participated in 3 National Natural Science Foundation projects and 6 provincial and ministerial level teaching and research projects. Guided undergraduate and graduate students to win over 20 national and provincial awards in subject competitions. Mainly engaged in scientific research work on artificial intelligence, multi-objective optimization algorithms, cloud computing and big data analysis, computational intelligence algorithms and applications.