PriMera Scientific Engineering
Volume 7 Issue 6 December 2025
ISSN: 2834-2550

A Transformer and LLSKA-Based U-Shaped Network for Medical Image Segmentation — T-LLSKA UNet

Type: Research Article
Received: October 28, 2025
Published: November 26, 2025

Citation:
Lao Tei., et al. "A Transformer and LLSKA-Based U-Shaped Network for Medical Image Segmentation — T-LLSKA UNet". PriMera Scientific Engineering 7.6 (2025): 20-35.

Lao Tei[1]*, Deng Zengjie[1], Gui Hao[2], Li Guanxi[1] and Peng Lei[1]

[1]*South China Normal University, China*

[2]*Guangdong Shufu Information Technology Co., LTD, China*

***Corresponding Author:** Tao Lei, South China Normal University, 55 Zhongshan West Avenue, Guangzhou, Guangdong, China.

Abstract

With the rapid advancement of artificial intelligence, particularly the breakthrough progress of deep learning technologies, the field of medical image segmentation has achieved remarkable improvements in both accuracy and efficiency. These technological advancements have greatly fostered innovation in modern healthcare systems, playing a crucial role in enhancing diagnostic precision and treatment efficiency. This paper focuses on addressing the challenge in medical image segmentation decoders, which often struggle to balance global contextual understanding with local detail representation. To overcome this limitation, an improved Large Kernel Separable Attention (LSKA) module is proposed. By analyzing the quadratic parameter growth problem that occurs when the number of feature channels increases in LSKA, two enhanced modules are designed: the Sparse Full-channel LSKA (LSKA-SF) and the Local-channel Large Separable Kernel Attention (LLSKA). LSKA-SF introduces sparsity through group convolution to effectively reduce parameter complexity, while LLSKA enhances feature representation via local cross-channel interactions, successfully mitigating the quadratic parameter growth trend. Based on these improvements, this study constructs a hybrid U-shaped segmentation network named Transformer-LLSKA UNet, which integrates a Transformer encoder and an LLSKA-based decoder to efficiently model both global contextual information and local spatial details. Experimental results demonstrate that Transformer-LLSKA UNet achieves outstanding segmentation performance on multiple organ datasets, including the aorta, liver, and spleen. Specifically, it achieves an average Dice coefficient of 83.43% and an HD95 of 15.15, indicating significant improvements in segmentation accuracy and boundary precision. These results validate the superior generalization ability and practical value of the proposed model, highlighting its potential to advance intelligent medical image analysis and clinical decision-making applications.

Keywords: LSKA; LLSKA; Transformer; UNet

Introduction

Medical image segmentation, as one of the core tasks in medical image processing, plays a crucial role in clinical diagnosis, pathological analysis, and treatment planning. High-precision and robust segmentation techniques not only assist clinicians in better understanding lesion characteristics but also significantly enhance the performance of computer-aided diagnosis (CAD) systems, thereby improving the overall quality and efficiency of healthcare services. However, medical images often exhibit complex anatomical structures, blurred boundaries, and significant noise interference. Traditional segmentation methods based on handcrafted feature design struggle to handle these challenges effectively, making it difficult to achieve the level of accuracy required for clinical applications.

In recent years, the emergence of deep learning, particularly Convolutional Neural Networks (CNNs), has effectively addressed the limitations of traditional approaches and greatly improved the performance of medical image segmentation. CNNs possess a strong capability for local feature extraction through convolutional operations, enabling automatic learning of spatial structural information and enhancing both accuracy and efficiency. Among CNN-based methods, the UNet architecture has become a mainstream framework due to its elegant encoder-decoder design and skip connection mechanism, which facilitate precise localization and contextual information fusion. UNett [1] and its variants have been widely adopted in medical imaging tasks such as tumor detection, organ boundary delineation, and lesion segmentation.

Nevertheless, CNNs inherently rely on local convolutions, which restrict their ability to model long-range dependencies and capture global contextual relationships. Although several improvements, such as U-Net++ and Attention U-Net, have enhanced the representation of contextual information to some extent, their limitations remain evident in complex and fine-grained segmentation tasks.

Meanwhile, the Transformer [2, 3] architecture, originally developed for natural language processing, has demonstrated remarkable success in modeling long-range dependencies. Inspired by its success, researchers began exploring Vision Transformers (ViTs) for image analysis to overcome the limitations of CNNs. Dosovitskiy et al. [4] first proposed ViT, which applies self-attention mechanisms to explicitly capture global contextual information in images. Later, the Swin Transformer [5] introduced a hierarchical structure with shifted window attention, improving both computational efficiency and local feature modeling, thereby enabling Transformers to achieve outstanding results across various vision tasks.

In medical imaging, Transformer-based architectures have also attracted increasing attention. Chen et al. [3] proposed TransUNet, which effectively combines CNN and Transformer components. The CNN encoder extracts low-level local features, while the Transformer module captures high-level global semantic information, significantly improving segmentation performance. However, TransUNet still relies on CNN-based decoders, which remain limited in efficiently modeling long-range dependencies.

To further address these challenges, recent studies have explored Large Kernel Attention (LKA) [6] mechanisms, which leverage large receptive fields to capture long-range dependencies more effectively. Specifically, Guo et al. proposed the Visual Attention Network (VAN), which introduced LKA by combining depthwise and dilated convolutions, enabling efficient global information aggregation. LKA successfully integrates the local inductive bias of CNNs with the long-range modeling capability of Transformers. However, the use of large convolution kernels substantially increases computational cost and parameter count, posing challenges for practical deployment.

To mitigate these limitations, Lau et al. introduced the Large Separable Kernel Attention (LSKA) [7] mechanism, which decomposes large 2D kernels into cascaded 1 × k and k × 1 convolutions, reducing parameter complexity while preserving long-range modeling capacity. Nonetheless, the 1 × 1 convolutions in LSKA still lead to quadratic growth in parameter count with respect to channel number, increasing computational and memory demands, which is problematic for resource-constrained medical devices.

To further reduce decoder complexity, this study proposes the Sparse Full-channel Large Separable Kernel Attention (LSKA-SF) module, which introduces group convolutions into the LSKA structure to eliminate redundant computations. By processing channel groups independently, LSKA-SF effectively reduces computational cost with minimal performance degradation, improving efficiency

and enabling lightweight model deployment. Building upon LSKA-SF, we further propose the Local-channel Large Separable Kernel Attention (LLSKA) module, which restricts interactions to adjacent channels rather than full-channel communication. This design alleviates the quadratic parameter growth problem while maintaining comparable performance, achieving a balance between accuracy and efficiency.

Finally, to fully optimize medical image segmentation, we integrate the proposed LLSKA module into a U-shaped segmentation network and employ a Transformer-based encoder to construct a novel hybrid architecture, the Transformer-LLSKA UNet (T-LLSKA UNet). This model combines the strengths of both CNNs and Transformers: the Transformer encoder captures global semantic and long-range dependencies, while the LLSKA-based decoder efficiently reconstructs fine spatial details. The resulting architecture achieves a favorable trade-off between accuracy and computational cost, demonstrating superior performance across multiple segmentation tasks.

In summary, the proposed T-LLSKA UNet effectively integrates Transformer and LLSKA mechanisms to balance precision and efficiency in medical image segmentation. This study not only overcomes the inherent limitations of traditional CNN and pure Transformer architectures but also provides a novel research direction and technical foundation for future advancements in intelligent medical image analysis.

Materials and Methods

T-LLSKA UNet is a novel network architecture designed for medical image segmentation, as illustrated in Figure 1. Its design is inspired by the classical U-shaped UNet structure while incorporating the state-of-the-art Vision Transformer (MaxViT) and the Local-channel Large Separable Kernel Attention (LLSKA) module. By combining the strengths of Transformer-based global semantic modeling and CNN-based local feature extraction, the network achieves a complementary balance between global contextual understanding and local detail preservation. This integration enables T-LLSKA UNet to deliver significant improvements in both segmentation accuracy and computational efficiency within the field of medical image analysis.

As shown in Figure 1, the overall architecture of T-LLSKA UNet can be clearly divided into two main components: an encoder and a decoder. In the encoder, the network first employs a convolutional block for preliminary feature extraction. This block consists of two consecutive 3 × 3 convolutional layers, where the first convolution uses a stride of 2 to rapidly reduce the spatial resolution of the input image, thereby enabling efficient capture of low-level feature representations. Through this initial convolutional module, the spatial dimensions of the input image are reduced from $H \times W \times C$ to $\frac{H}{4} \times \frac{W}{4} \times C$, generating a foundational feature map that serves as the basis for subsequent high-level feature extraction.

The core of the encoder adopts the MaxViT architecture [8], a powerful and efficient model within the Vision Transformer family, capable of simultaneously capturing local spatial dependencies and global contextual relationships. MaxViT exhibits strong representational capacity and long-range dependency modeling, making it highly suitable for complex medical image segmentation tasks. In the encoder stage of T-LLSKA UNet, four consecutive MaxViT stages are employed, each consisting of two successive MaxViT blocks. After each stage, a down-sampling layer is applied to further reduce the spatial dimensions of the feature maps, thereby extracting increasingly abstract and semantically rich features while effectively capturing global contextual information. As the network depth increases, the spatial resolution is progressively reduced to $\frac{H}{32} \times \frac{W}{32}$, while the channel dimension is expanded to 8C, forming a comprehensive and high-level semantic representation of the input image.

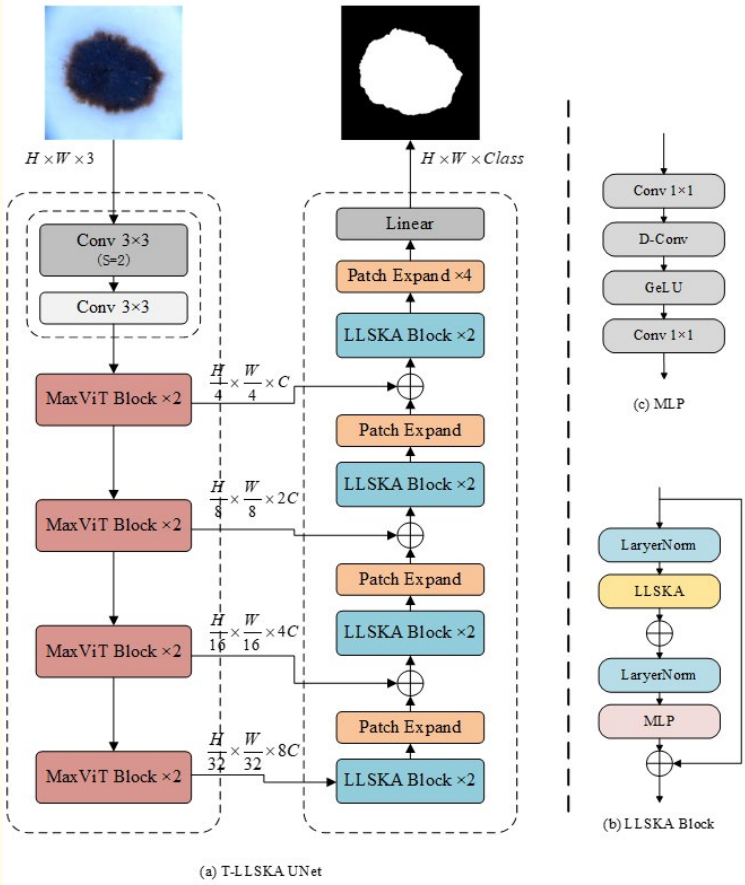

Figure 1: Overall architecture of the T-LLSKA UNet.

After the encoder extracts high-level semantic features, these representations are transmitted to the decoder through skip connections to facilitate the gradual recovery of spatial details and achieve refined segmentation. The decoder of T-LLSKA UNet is composed of four hierarchical stages, each containing two sequential Local-channel Large Separable Kernel Attention (LLSKA) modules. These modules ensure efficient spatial feature reconstruction and effective long-range dependency modeling. Specifically, the LLSKA module is designed to balance performance and efficiency by employing a convolutional mechanism that preserves local spatial feature extraction while introducing local channel interactions to significantly reduce parameter complexity. This design alleviates the computational and memory burdens commonly associated with traditional large-kernel attention mechanisms.

To progressively restore the spatial resolution of the feature maps, each LLSKA stage in the decoder is followed by a Patch Expand layer. The Patch Expand operation enlarges the spatial dimensions of the feature maps while appropriately reducing the number of channels, thereby reconstructing spatial details in a stepwise manner. Moreover, skip connections ensure that fine-grained local information from the encoder's feature maps is effectively propagated to the decoder, assisting in the high-quality recovery of spatial structure and boundary details.

At the final stage of the decoder—after the last Patch Expand operation—the network employs a Linear projection layer to map the reconstructed features into the final segmentation mask of size H × W × Class, thus achieving precise, pixel-level medical image

segmentation.

Each LLSKA block consists of three main components: Layer Normalization (LayerNorm), the Local-channel Large Separable Kernel Attention (LLSKA) module, and a Multi-Layer Perceptron (MLP). The LayerNorm layer normalizes feature statistics to enhance training stability, while the LLSKA module focuses on efficiently capturing long-range dependencies in both spatial and channel dimensions. The MLP further strengthens the network's feature representation and nonlinear transformation capabilities. In addition, residual connections are incorporated within each LLSKA block to facilitate effective feature propagation, ensuring stable and efficient training even at greater network depths.

Overall, the T-LLSKA UNet architecture effectively inherits the advantages of the traditional UNet in restoring spatial details while integrating the strengths of Vision Transformers and lightweight large-kernel attention mechanisms. Specifically, the MaxViT modules in the encoder provide powerful long-range dependency modeling and high-level semantic feature extraction, whereas the LLSKA modules in the decoder achieve precise spatial detail reconstruction and enhanced long-distance dependency refinement, leading to superior segmentation performance and computational efficiency.

Encoder

The encoder of the proposed T-LLSKA UNet framework is constructed based on the MaxViT architecture. MaxViT integrates the advantages of both convolutional operations and self-attention mechanisms, effectively combining local feature extraction with global contextual modeling. By introducing the Multi-Axis Attention module, MaxViT is capable of simultaneously capturing fine-grained local details and long-range global dependencies. This hybrid design enables the encoder to extract multi-scale feature representations more effectively, thereby enhancing the model's generalization capability and overall representational power.

Decoder

The decoder of the proposed framework is constructed by cascading multiple LLSKA (Local-channel Large Separable Kernel Attention) blocks. This design enables the decoder to capture long-range dependencies and global contextual information while preserving low-level spatial features, all with a reduced number of parameters.

In this section, we first employ a sparse full-channel interaction strategy to effectively reduce the model's parameter count and computational complexity. Subsequently, by comparing the performance of sparse full-channel interaction and local-channel interaction, we are motivated to propose the LLSKA module, which adopts an adaptive channel grouping strategy to determine the receptive range of group convolutions dynamically. Finally, a detailed analysis and comparison of the computational complexity between LSKA and LLSKA are presented to demonstrate the efficiency and scalability of the proposed approach.

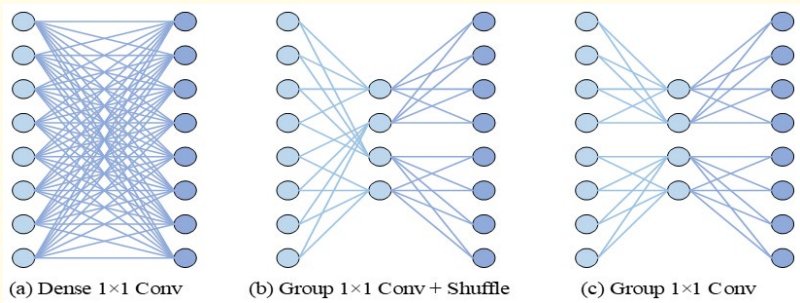

Figure 2: Connections of different 1 × 1 convolutional neural networks and their corresponding receptive fields.

Sparse Full-Channel Interaction

As illustrated in Figure 3a, the original LSKA module employs a 1 × 1 convolution (Conv) to capture full-channel information. In this configuration, the connections between output and input neurons are fully dense, as shown in Figure 2a, where each output neuron is directly connected to all input neurons, resulting in a channel receptive field of C.

In contrast, as shown in Figure 3b, the proposed design utilizes two group convolutions (GConv) followed by a channel shuffle operation to achieve full-channel interaction. As depicted in Figure 2b, each output neuron connects to all input neurons in an indirect yet efficient manner, maintaining the same effective channel receptive field of C.

This modified version of the LSKA module, which introduces sparsity in the full-channel interaction while preserving comprehensive feature communication, is referred to as the Sparse Full-channel Large Separable Kernel Attention (LSKA-SF) module.

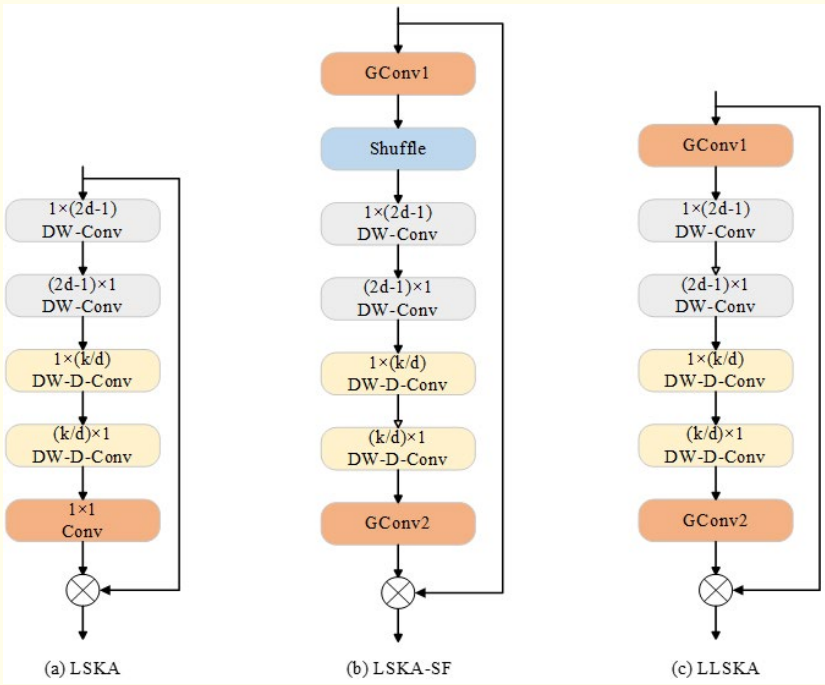

Figure 3: Comparison of different designs of large-kernel attention modules.

To ensure that the two cascaded Group Convolutions (GConvs) achieve the same receptive field as a standard 1 × 1 convolution while minimizing the number of parameters, we adopt the sparse decomposition strategy proposed in FalconNet [9].

As illustrated in Figure 4, let the input feature map be denoted as Feature $\in R^{H\times W\times \frac{C}{R}}$, the intermediate feature map as Feature $\in R^{H\times W\times \frac{C}{R}}$, and the output feature map as Output $\in R^{H\times W\times C}$, where H and W represent the spatial dimensions, and C denotes both the input and output channel numbers. The channel reduction ratio R (a hyperparameter) is set to 2 by default.

In the first group convolution (GConv1), the receptive field size is denoted as k, and the number of groups is $G_1 = \frac{C}{k}$. With C input channels and $\frac{C}{R}$ output channels, the total number of parameters in GConv1 is $\frac{Ck}{R}$.

To maintain an equivalent channel receptive field as that of the standard 1 × 1 convolution, the second group convolution (GConv2) employs $G_2 = \frac{C}{RG_1}$ groups, with $\frac{C}{R}$ input channels and C output channels. Consequently, the number of parameters in GConv2 is $\frac{C^2}{k}$. The total parameter count for both GConvs is therefore $\frac{Ck}{R} + \frac{C^2}{k}$. To minimize the total number of parameters, k should be set to $\sqrt{CR}$, which is dynamically adjusted based on C and R. The resulting total number of parameters for the two GConvs is $2C\sqrt{\frac{C}{R}}$.

By introducing channel sparsity while preserving the full-channel receptive field, the LSKA-SF module effectively reduces the parameter count, achieving a balance between computational efficiency and representational capacity.

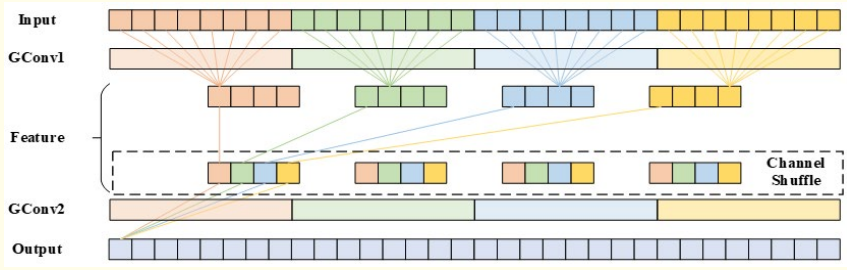

Figure 4: Channel receptive field of LSKA-SF.

GConv1 and GConv2 represent two group convolution operations, where different colors indicate distinct convolution groups. Input, Feature, and Output correspond to the input feature map, intermediate feature map, and output feature map, respectively. The number of rectangular blocks illustrates the number of feature channels, while the varying colors of these blocks denote the feature maps processed by different convolution groups.

Local Channel Interaction

As illustrated in Figure 3c, the local channel interaction is implemented using only two group convolutions (GConvs). As previously discussed, employing group convolution combined with channel shuffling can effectively reduce the number of parameters while maintaining full-channel interaction. However, the sparse connections introduced by shuffling decrease the amount of information each output neuron receives from input neurons, thereby weakening the network's representational capability.

To address this limitation, we remove the shuffle operation and employ only group convolutions, resulting in denser connections between output and input neurons. As shown in Figures 2b and 2c, when using group convolution with shuffling, a single output neuron connects to only four input neurons, whereas using group convolution alone allows each output neuron to connect to eight input neurons, effectively enhancing its representational capacity.

In the LLSKA module, the number of groups in the group convolution is set to be the same as in LSKA-SF, ensuring that both modules have an equivalent number of parameters. As illustrated in Figure 5, the channel receptive field of each output in LLSKA covers only a local subset of the input channels, rather than the entire input space, thereby focusing on localized channel interactions while maintaining computational efficiency.

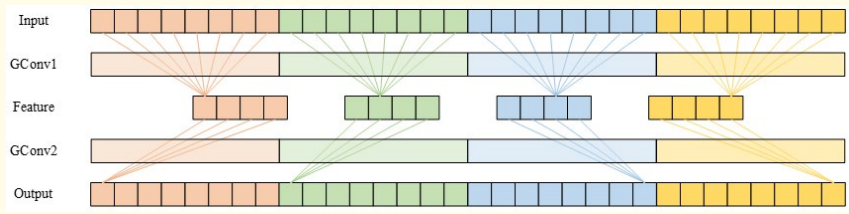

Figure 5: Channel receptive field of LLSKA.

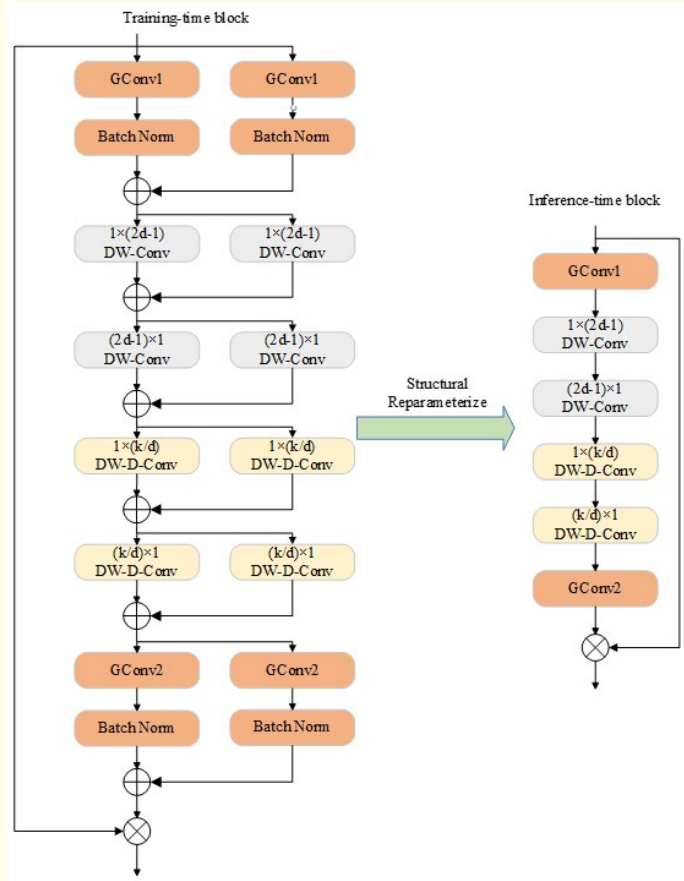

Figure 6: Structure diagram of LLSKA during training and inference.

Channel-Adaptive Strategy

Since the proposed module employs ***local channel interaction***, it is necessary to determine the receptive field size of channel interactions. The number of channels in feature maps varies across different stages of the network; therefore, the interaction range must be dynamically adjusted according to the channel dimension. To achieve this adaptability, the receptive field k is adjusted based on the

given number of channels C, which in turn determines the number of groups *G*. Hence, the grouping configuration can be adaptively established. The corresponding formulas for *G* and *k* are as follows:

$$G = \frac{C}{k} \tag{1}$$

$$k = \begin{cases} \frac{C}{\lfloor \sqrt{C}/\gamma \rfloor \times \gamma} \\ \sqrt{CR} \end{cases} \tag{2}$$

Here, $\lfloor t \rfloor$denotes the floor operation. Since the number of channels in feature maps within the network is typically a multiple of four, and the number of groups G must be divisible by C, we set the constant $\gamma = 4$. To prioritize minimizing the number of parameters in the module, "k" is first computed using $\sqrt{\mathrm{CR}}$. If $\sqrt{\mathrm{CR}}$ is not an integer, "k" is recalculated using the expression $\frac{C}{\lfloor \sqrt{C}/\gamma \rfloor \times \gamma}$.

Through this adaptive strategy, feature maps with a larger number of channels are assigned a wider receptive range, while those with fewer channels have a narrower one. This design not only accommodates varying feature dimensions across the network but also ensures parameter efficiency to maintain lightweight computation.

Parallel Branches

As illustrated in Figure 6, for the spatial convolution component, additional convolutional branches with different kernel sizes are introduced to enrich the network's spatial representation capability. For the channel convolution component, additional branches with identical convolution operations are incorporated to compensate for the parameter reduction and to enhance channel representation. Owing to the additive and homogeneous properties of convolutions, as well as parameter reparameterization techniques, these multiple branches can be merged into a single convolutional branch during inference, thereby introducing no additional inference cost while maintaining enhanced representational power during training.

Complexity Analysis of LSKA and LLSKA Algorithms

As shown in Figure 3a, for the LSKA structure, it is assumed that the input and output feature maps share the same spatial dimensions, denoted as H × W × C. We calculate the floating-point operations (FLOPs) and number of parameters (Params) for LSKA to quantitatively analyze its computational complexity. The formulas for computing the parameters and FLOPs of LSKA are as follows:

$$\text{Params} = (2d-1) \times C \times 2 + \left[\frac{k}{d}\right] \times C \times 2 + C \times C \tag{3}$$

$$FLOPs = \left((2d-1) \times C \times 2 + \left[\frac{k}{d}\right] \times C \times 2 + C \times C\right) \times H \times W \tag{4}$$

Here, k denotes the kernel size, and d represents the dilation rate.

The parameters and floating-point operations (FLOPs) for the LLSKA module are calculated as follows:

$$\text{Params} = (2d-1) \times \frac{C}{R} \times 2 + \left[\frac{k}{d}\right] \times \frac{C}{R} \times 2 + \frac{C^2}{RG_1} + CG_1 \tag{5}$$

$$FLOPS = \left((2d-1) \times \frac{C}{R} \times 2 + \left[\frac{k}{d}\right] \times \frac{C}{R} \times 2 + \frac{C^2}{RG_1} + CG_1\right) \times H \times W \tag{6}$$

Based on the above analysis, when C, R, d, k remain constant, to minimize Equations (5) and (6), the optimal condition is achieved when $\mathrm{G}_1 = \sqrt{\frac{\mathrm{C}}{\mathrm{R}}}$. Under this condition, Equations (7) and (8) can be derived as follows:

$$\text{Params} = (2d-1)\times\frac{C}{R}\times 2 + \left[\frac{k}{d}\right]\times\frac{C}{R}\times 2 + 2C\sqrt{\frac{C}{R}} \quad (7)$$

$$\text{FLOPS} = \left((2d-1)\times\frac{C}{R}\times 2 + \left[\frac{k}{d}\right]\times\frac{C}{R}\times 2 + 2C\sqrt{\frac{C}{R}}\right)\times H\times W \quad (8)$$

From the first and second terms of Equations (3) and (7), the number of parameters in the spatial convolution layer of LLSKA is $\frac{1}{R}$ of that in LSKA. Similarly, comparing the third term, it changes from C^2 to $C^{\frac{3}{2}}$. At the same time, it can be seen that LLSKA resolves the problem of quadratic growth—its number of parameters and FLOPs are both smaller than those of LSKA.

Results and Discussion

To verify the effectiveness of LLSKA, we conducted experiments on both image classification and medical image segmentation tasks. Models incorporating LLSKA and LSKA were compared on the skin lesion segmentation dataset and the Synapse dataset to evaluate their performance and validate the effectiveness of LLSKA in medical image analysis.

Datasets

Synapse Multi-organ Segmentation Dataset [10]: We evaluated the proposed method on the well-established Synapse multi-organ segmentation dataset to assess its performance. This dataset consists of 30 abdominal CT cases with a total of 3,779 axial slices. Each CT volume contains between 85 and 198 slices, with each slice having a spatial resolution of 512 × 512 pixels. The voxel spacing varies within the range of ([0.54-0.54] × [0.98-0.98] × [2.5-5.0]) mm^3. Our experimental setup strictly follows the protocols described in [3, 11].

Skin Lesion Segmentation: We further extended our experiments to the skin lesion segmentation task using benchmark dermoscopic datasets. Specifically, we employed the ISIC 2017 dataset [12], which includes 2,000 dermoscopic images for training, 150 for validation, and 600 for testing. In addition, we utilized the ISIC 2018 dataset [13], following the official data partitioning scheme described in prior works [14-17].

Experimental Environment and Parameter Settings

All experiments were implemented using the PyTorch framework and trained on a single NVIDIA RTX 3090 GPU. The batch size was set to 20, and training was performed using Stochastic Gradient Descent (SGD) with an initial learning rate of 0.05, momentum of 0.9, and a weight decay coefficient of 0.0001. The models were trained for 400 epochs in total.

The loss function was a weighted combination of the Dice loss and Cross-Entropy (CE) loss, defined as: $L_{total} = 0.6 \cdot L_{dice} + 0.4 \cdot L_{ce}$ Following [3], the same data augmentation techniques were applied to enhance the model's generalization capability.

Parameter	***Setting***
Learning rate	0.05
Momentum	0.9
Weight decay	0.0001
Batch size	20
Epoch	400

Table 1: Training parameter settings for the Synapse dataset.

Parameter	*Set up*
Learning rate	0.05
Momentum	0.9
Weight decay	0.0001
Batch size	4
Epoch	100

Table 2: Training parameter settings for the ISIC2017 and ISIC2018 datasets.

Analysis and visualization of experimental results for medical image segmentation tasks

Methods	*ISIC2017*				*ISIC2018*			
	DSC	*SE*	*SP*	*ACC*	*DSC*	*SE*	*SP*	*ACC*
UNet [1]	81.59	81.72	96.80	91.64	85.45	88.00	96.97	94.04
Att-UNet [18]	80.82	79.98	97.76	91.45	85.66	86.74	98.63	93.76
TransUNet [3]	81.23	82.63	95.77	92.07	84.99	85.78	96.53	94.52
Swin-Unet [2]	91.83	91.42	97.98	97.01	89.46	90.56	97.98	96.45
T-LLSKA UNet	91.19	88.26	99.00	97.09	91.22	89.12	98.42	96.56

Table 3: Comparison results of T-LLSKA UNet on the ISIC2017 and ISIC2018 datasets (blue text indicates the best performance, red indicates the second best).

In this chapter, the proposed network is compared with representative Transformer-based and CNN-based models on the ISIC 2017 and ISIC 2018 skin lesion segmentation datasets.

As shown in Table 3, on the ISIC 2017 dataset, T-LLSKA UNet demonstrates superior performance across four key evaluation metrics: Dice Similarity Coefficient (DSC), Sensitivity (SE), Specificity (SP), and Accuracy (ACC).

Specifically, the DSC of T-LLSKA UNet reaches 91.19%, ranking second only to Swin-Unet (91.83%), yet significantly higher than UNet (81.59%), Att-UNet (80.82%), and TransUNet (81.23%), indicating that T-LLSKA UNet achieves stronger segmentation performance and superior region reconstruction accuracy. The Sensitivity (SE) reaches 88.26%, which is notably higher than UNet (81.72%), Att-UNet (79.98%), and TransUNet (82.63%), suggesting that the proposed model is more effective in detecting lesion regions and reducing false negatives.

Furthermore, T-LLSKA UNet attains the highest Specificity (SP) of 99.00%, surpassing Swin-Unet (97.98%), demonstrating its robustness in minimizing false positive predictions. The Accuracy (ACC) reaches 97.09%, outperforming UNet (91.64%), Att-UNet (91.45%), and TransUNet (92.07%), and is comparable to Swin-Unet (97.01%), reflecting a high overall classification correctness.

These results collectively verify that T-LLSKA UNet effectively balances precision and sensitivity, achieving outstanding segmentation accuracy and robustness on complex skin lesion segmentation tasks.

On the ISIC 2018 dataset, T-LLSKA UNet continues to demonstrate outstanding performance across all four evaluation metrics—Dice Similarity Coefficient (DSC), Sensitivity (SE), Specificity (SP), and Accuracy (ACC).

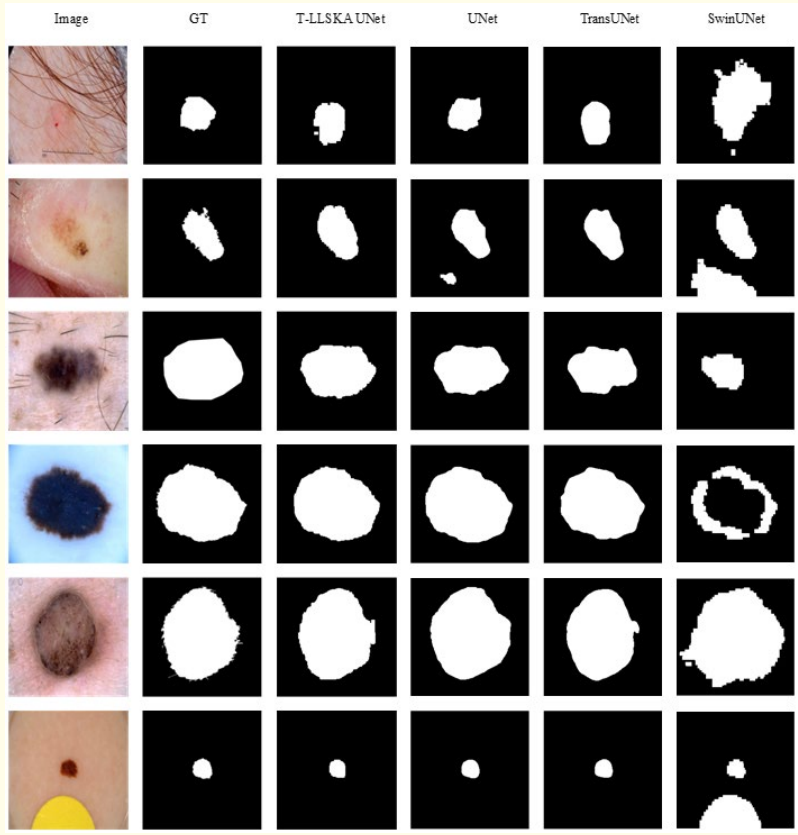

Figure 7: Visualization of T-LLSKA UNet on the ISIC2018 dataset.

Specifically, T-LLSKA UNet achieves a DSC of 91.22%, which is significantly higher than UNet (85.45%), Att-UNet (85.66%), TransUNet (84.99%), and Swin-Unet (89.46%), confirming its superior overall segmentation performance in medical image analysis. The Sensitivity (SE) reaches 89.12%, outperforming UNet (88.00%), Att-UNet (86.74%), and TransUNet (85.78%), and ranking just below Swin-Unet (90.56%). This indicates that T-LLSKA UNet achieves higher accuracy in lesion detection and effectively reduces false negatives.

The Specificity (SP) of 98.42% is slightly lower than that of Att-UNet (98.63%) but notably higher than the other models, suggesting that the proposed network exhibits superior capability in minimizing false positives. The Accuracy (ACC) reaches 96.56%, exceeding UNet (94.04%), Att-UNet (93.76%), and Swin-Unet (96.45%), and approaching TransUNet (94.52%), indicating an exceptionally high overall classification accuracy.

As shown in Figure 7, in the second row of qualitative comparisons, UNet and Swin-Unet are easily influenced by surrounding skin textures, leading to false detections. In contrast, T-LLSKA UNet and TransUNet, which both integrate Transformer and CNN architectures, exhibit stronger resistance to noise and can simultaneously capture long-range dependencies and local spatial details. In the fifth row, the segmentation results of T-LLSKA UNet are visually closer to the ground truth (GT) compared to TransUNet, demonstrating that the LLSKA module in the decoder enables the network to focus effectively on both local and global contextual information, thereby achieving more accurate and precise segmentation outcomes.

Methods	*Aor*	*Gal*	*LKid*	*RKid*	*Liv*	*Pan*	*Spl*	*Sto*	*Average*	
									DSC	*HD95*
TransUNet [3]	87.23	63.16	81.87	77.02	94.08	55.86	85.08	75.62	77.49	31.69
Swin-UNet [2]	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60	79.13	21.55
LeViT-UNet-384 [18]	87.33	62.23	84.61	80.25	93.11	59.07	88.86	72.76	78.53	16.84
MISSFormer [19]	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81	81.96	18.20
ScaleFormer [20]	88.73	74.97	86.36	83.31	95.12	64.85	89.40	80.14	82.86	16.81
HiFormer-B [21]	86.21	65.23	85.23	79.77	94.61	59.52	90.99	81.08	80.39	14.70
DAEFormer [22]	87.84	71.65	87.66	82.39	95.08	63.93	91.82	80.77	82.63	16.39
TransDeepLab [23]	86.04	69.16	84.08	79.88	93.53	61.19	89.00	78.40	80.16	21.25
PVT-CASCADE [24]	83.01	70.59	82.23	80.37	94.08	64.43	90.1	83.69	81.06	20.23
T-LLSKA UNet	88.99	72.01	84.73	83.97	95.13	67.56	91.92	83.18	83.43	15.15

Table 4: Comparison results of T-LLSKA UNet on the Synapse dataset (bold text indicates the best performance; all metric values are expressed as percentages).

T-LLSKA UNet employs MaxViT as its encoder and integrates the Local-channel Large Kernel Attention (LLSKA) mechanism, achieving a complementary balance between Convolutional Neural Networks (CNNs) and Transformers. Compared with pure CNN-based models, the MaxViT encoder introduces Transformer components to enhance global feature extraction, thereby alleviating the inherent limitation of CNNs' local receptive fields. Meanwhile, the LLSKA module, with its large receptive field, strengthens local feature extraction and compensates for the Transformer's relatively weak modeling of local spatial dependencies.

Compared with pure Transformer-based architectures, MaxViT effectively combines convolutional operations with self-attention mechanisms, ensuring robust global context modeling while improving the representation of fine-grained local features. The LLSKA module further expands the local receptive field with low computational cost, making the overall architecture more suitable for high-precision medical image segmentation.

The quantitative results on the Synapse multi-organ segmentation dataset (as shown in Table 4) demonstrate the superior performance of T-LLSKA UNet. The model achieves a high Dice Similarity Coefficient (DSC) and a low Hausdorff Distance (HD95), indicating excellent segmentation accuracy and robustness. Specifically, T-LLSKA UNet achieves the best results for multiple organs, including the aorta (Aor, 88.99%), right kidney (RKid, 83.97%), liver (Liv, 95.13%), pancreas (Pan, 67.56%), and spleen (Spl, 91.92%). The high segmentation accuracy for the liver and spleen demonstrates strong generalization capability for large organs, while the superior performance on the pancreas highlights the model's enhanced ability to capture fine details of small organs.

Compared with CNN-based models such as UNet and LeViT-UNet-384, T-LLSKA UNet significantly improves global contextual modeling, leading to higher segmentation accuracy for large organs (e.g., the liver and spleen). Additionally, the LLSKA mechanism enhances local feature extraction, thereby improving the segmentation of small organs such as the gallbladder and pancreas.

When compared with Transformer-based models (e.g., TransUNet and Swin-Unet), T-LLSKA UNet exhibits stronger local feature representation, resulting in higher boundary accuracy (i.e., lower HD95) and better segmentation quality for small, fine-grained anatomical structures.

Overall, T-LLSKA UNet, by synergistically integrating MaxViT and LLSKA, achieves an optimal balance between global semantic understanding and local spatial precision, demonstrating superior performance, robustness, and generalization across multi-organ medical image segmentation tasks.

Ablation Study on Medical Image Segmentation

Sparse Full-Channel Interaction and Local Channel Interaction

Methods	*ISIC2017*				*ISIC2018*			
	DSC	*SE*	*SP*	*ACC*	*DSC*	*SE*	*SP*	*ACC*
T-LLSKA UNet	91.19	88.26	99.00	97.09	91.22	89.12	98.42	96.56
T-LSKA-SF UNet	91.05	87.93	98.97	97.00	90.72	87.55	98.63	96.42

Table 5: Ablation experiments of T-LLSKA UNet and T-LSKA-SF UNet on the ISIC2017 and ISIC2018 datasets.

In this section, ablation experiments are conducted on medical image segmentation tasks to further demonstrate the effectiveness of each component within the LLSKA module. First, by comparing the performance of LSKA-SF and LLSKA on the ISIC 2017 and ISIC 2018 datasets, we evaluate the relative advantages of sparse full-channel interaction and local channel interaction in segmentation tasks. The experimental setup follows the same configuration described previously for these datasets. The results are summarized in Table 5, which shows that LLSKA consistently outperforms LSKA-SF across nearly all evaluation metrics, confirming that local channel interaction is more effective than sparse full-channel interaction for medical image segmentation.

Channel-Adaptive Strategy

According to Equations (7) and (8), the value of k in the LLSKA module determines the size of the channel receptive field. In this experiment, we set k to fixed values of 8 and 16, as well as to a dynamically adaptive strategy, and compare their performance. As shown in Table 6, the results indicate that the value of k has a noticeable impact on model performance. Models with a fixed k value fail to adapt to variations in the number of feature map channels, whereas the adaptive strategy dynamically adjusts the receptive field according to channel dimensions, resulting in improved segmentation performance and enhanced generalization capability.

Methods	*K*	*ISIC2017*				*ISIC2018*			
		DSC	*SE*	*SP*	*ACC*	*DSC*	*SE*	*SP*	*ACC*
T-LLSKA UNet	8	91.01	88.02	98.95	97.01	89.48	85.89	98.52	95.99
	16	91.74	90.50	98.40	97.00	90.37	87.64	98.64	96.44
	adapt	91.19	88.26	99.00	97.09	91.22	89.12	98.42	96.56

Table 6: Ablation experiments of the channel adaptation strategy in T-LLSKA UNet on the ISIC2017 and ISIC2018 datasets.

Parallel Branches

Table 7 presents the comparison of segmentation results between models trained with and without parallel branches. The results show that incorporating parallel branches consistently improves all evaluation metrics. In medical image segmentation tasks, introducing parallel branches into the LLSKA module enhances segmentation performance without increasing the model's inference parameters or inference time, thereby achieving better accuracy with no additional computational cost.

Methods	*Branch*	*ISIC2017*				*ISIC2018*			
		DSC	*SE*	*SP*	*ACC*	*DSC*	*SE*	*SP*	*ACC*
T-LLSKA UNet	N	90.78	87.76	98.94	96.95	90.87	88.61	98.27	96.34
	Y	91.19	88.26	99.00	97.09	91.22	89.12	98.42	96.56

Table 7: Ablation experiments of the parallel branches in T-LLSKA UNet on the ISIC2017 and ISIC2018 datasets.

Conclusion

This study addresses the challenge in medical image segmentation where decoders struggle to balance global contextual understanding with local detail preservation, and proposes a Transformer-LLSKA-based U-shaped segmentation network (T-LLSKA UNet). The proposed model employs a Transformer encoder for global feature modeling and integrates the LLSKA (Local-channel Large Separable Kernel Attention) module as the core of the decoder to achieve efficient feature fusion through local channel interaction. This design enhances feature representation capability while effectively reducing the number of parameters. Experimental results demonstrate that T-LLSKA UNet achieves a Dice Similarity Coefficient (DSC) of 83.43% on the Synapse multi-organ segmentation dataset, and exhibits excellent segmentation performance on both the ISIC 2017 and ISIC 2018 datasets. These results validate the model's high accuracy and strong generalization ability across multiple organs and diverse medical imaging scenarios. Overall, T-LLSKA UNet provides an effective and efficient new approach for intelligent medical image analysis.

Acknowledgements

This work is partially supported by a Teaching Reform Project of Guangdong Provincial Department of Education in Introduction to Artificial Intelligence (No.2023-4-511). Tao Lei is supported by a State Scholarship of China Scholarship Council (No.201806755032). We would like to thank Guangdong Shufu for specialized laboratories.

References

1. Ronneberger O, Fischer P and Brox T. "U-net: Convolutional networks for biomedical image segmentation". Proceedings of the Medical image computing and computer-assisted intervention-MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 (2015).
2. Cao H., et al. "Swin-unet: Unet-like pure transformer for medical image segmentation". Proceedings of the European conference on computer vision (2022).
3. Chen J., et al. "Transunet: Transformers make strong encoders for medical image segmentation". arXiv preprint arXiv:210204306 (2021).
4. Dosovitskiy A., et al. "An image is worth 16x16 words: Transformers for image recognition at scale". arXiv preprint arXiv:201011929 (2020).
5. Liu Z., et al. "Swin transformer: Hierarchical vision transformer using shifted windows". Proceedings of the IEEE/CVF international conference on computer vision (2021).
6. Guo M-H., et al. "Visual attention network". Computational visual media 9.4 (2023): 733-752.
7. Lau KW, Po L-M and Rehman YAU. "Large separable kernel attention: Rethinking the large kernel attention design in cnn". Expert Systems with Applications 236 (2024): 121352.
8. Tu Z., et al. "Maxvit: Multi-axis vision transformer". Proceedings of the European conference on computer vision (2022).
9. Cai Z and Shen Q. "Falconnet: Factorization for the light-weight convnets". Proceedings of the International Conference on Neural Information Processing (2023).
10. Miccai. Multi-Atlas Abdomen Labeling Challenge: Synapse multi-organ segmentation dataset 52 (2015).
11. Shaker AM., et al. "UNETR++: delving into efficient and accurate 3D medical image segmentation". IEEE Transactions on Medical Imaging (2024).
12. Codella NC., et al. "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)". Proceedings of the 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018) (2018).
13. Codella N., et al. "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)". arXiv preprint arXiv:190203368 (2019).
14. Aghdam EK., et al. "Attention swin u-net: Cross-contextual attention mechanism for skin lesion segmentation". Proceedings of the 2023 IEEE 20th international symposium on biomedical imaging (ISBI) (2023).

15. Asadi-Aghbolaghi M., et al. “Multi-level context gating of embedded collective knowledge for medical image segmentation”. arXiv preprint arXiv:200305056 (2020).
16. Azad R., et al. “Bi-directional ConvLSTM U-Net with densley connected convolutions”. Proceedings of the IEEE/CVF international conference on computer vision workshops (2019).
17. Eskandari S and Lumpp J. “Inter-scale dependency modeling for skin lesion segmentation with transformer-based networks”. arXiv preprint arXiv:231013727 (2023).
18. Xu G., et al. “Levit-unet: Make faster encoders with transformer for medical image segmentation”. Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV) (2023).
19. Huang X., et al. “Missformer: An effective transformer for 2d medical image segmentation”. IEEE transactions on medical imaging 42.5 (2022): 1484-1494.
20. Huang H., et al. “ScaleFormer: revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation”. arXiv preprint, ar-Xiv: 220714552 (2022).
21. Heidari M., et al. “Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation”. Proceedings of the IEEE/CVF winter conference on applications of computer vision (2023).
22. Azad R., et al. “Dae-former: Dual attention-guided efficient transformer for medical image segmentation”. Proceedings of the International workshop on predictive intelligence in medicine (2023).
23. Azad R., et al. “Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation”. Proceedings of the International Workshop on PRedictive Intelligence in MEdicine (2022).
24. Rahman MM and Marculescu R. “Medical image segmentation via cascaded attention decoding”. Proceedings of the IEEE/CVF winter conference on applications of computer vision (2023).