IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 June 2023, accepted 19 July 2023, date of publication 25 July 2023, date of current version 9 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3298826

RESEARCH ARTICLE

A VAN-Based Multi-Scale Cross-Attention Mechanism for Skin Lesion Segmentation Network

SHUANG LIU[1], ZENG ZHUANG[1], YANFENG ZHENG[1], AND SIMON KOLMANIČ[2]

[1]School of Computer Science and Engineering, Dalian Minzu University, Dalian, Liaoning 116600, China
[2]Faculty of Electrical Engineering and Computer Science, University of Maribor, 2000 Maribor, Slovenia

Corresponding author: Shuang Liu (dlnuliushuang@qq.com)

This work was supported in part by the Chinese–Slovenian Scientific and Technological 2021 Cooperation Project of the Ministry of Science and Technology under Grant 13-20, and in part by the Liaoning Province Economic and Social Development Research Project 2023 of Provincial Social Science Association under Grant 2023lslybkt-039.

ABSTRACT With the rise of deep learning technology, the field of medical image segmentation has undergone rapid development. In recent years, convolutional neural networks (CNNs) have brought many achievements and become the consensus in medical image segmentation tasks. Although many neural networks based on U-shaped structures and methods, such as skip connections have achieved excellent results in medical image segmentation tasks, the properties of convolutional operations limit their ability to effectively learn local and global features. To address this problem, the Transformer from the field of natural language processing (NLP) was introduced to the image segmentation field. Various Transformer-based networks have shown significant performance advantages over mainstream neural networks in different visual tasks, demonstrating the huge potential of Transformers in the field of image segmentation. However, Transformers were originally designed for NLP and ignore the multidimensional nature of images. In the process of operation, they may destroy the 2D structure of the image and cannot effectively capture low-level features. Therefore, we propose a new multi-scale cross-attention method called M-VAN Unet, which is designed based on the Visual Attention Network (VAN) and can effectively learn local and global features. We propose two attention mechanisms, namely MSC-Attention and LKA-Cross-Attention, for capturing low-level features and promoting global information interaction. MSC-Attention is designed for multi-scale channel attention, while LKA-Cross-Attention is a cross-attention mechanism based on the large kernel attention (LKA). Extensive experiments show that our method outperforms current mainstream methods in evaluation metrics such as Dice coefficient and Hausdorff 95 coefficient.

INDEX TERMS CNNs, deep learning, medical image processing, NLP, semantic segmentation.

I. INTRODUCTION

Skin cancer is one of the most common and deadliest forms of cancer, with its primary risk factor being the production of melanin by melanocytes in the epidermis at an abnormally high rate due to ultraviolet (UV) radiation. The lethal form of skin cancer is melanoma, and outdoor activities and exposure to sunlight have been important contributing factors to the increasing incidence of melanoma in the past 70 years. Due to the increased exposure to UV radiation, melanoma has rapidly increased in the white population in the past few decades, with an annual increase of about 3-7% [1]. According to the American Cancer Society (ACS) forecast, the number of new melanoma cases in the United States is expected to reach 97,610 in 2023, with a high mortality rate of 7,990. However, if melanoma is detected early, the five-year survival rate of patients is over 94% [2]. In the United States, 90% of the cost of treating melanoma is related to late detection. Therefore, detecting melanoma at an earlier stage can significantly reduce healthcare costs and is also crucial

The associate editor coordinating the review of this manuscript and approving it for publication was Badri Narayan Subudhi.

for saving patients' lives. The current diagnostic approach for melanoma mainly relies on the use of thermoscopic images. However, due to varying levels of expertise among clinicians, the sensitivity and specificity rates for diagnosing melanoma using these images range from 69% to 92% and from 94% to 99% respectively [3]. Therefore, the use of image segmentation techniques can undoubtedly minimize reliance on clinical experience and‘ reduce misdiagnosis rates. Early melanoma lesions exhibit varying shades of color, fuzzy edges, and are often hidden in hair follicles, making them difficult to detect. Furthermore, the presence of bubbles, ruler marks, lighting changes, and low contrast in actual thermoscopic images pose significant challenges to image segmentation. As a result, the segmentation of skin lesions is inherently challenging.

Currently, medical image segmentation is primarily based on fully convolutional networks (FCNN) using encoder-decoder structures [4], [5], such as U-Net [4], which is almost the most widely used model in current segmentation projects. U-Net is a fully symmetric U-shaped model consisting of an encoder and a decoder that captures fine-grained features and fuses high- and low-level semantic features through skip connections, enabling it to achieve high accuracy with less training data. U-Net's simple structure and high performance have spurred further research into this technology and the development of a series of variant networks, such as Res-UNet [6], U-Net++ [7], U2-Net [8], BCDU-Net [9], which have also been widely used in skin lesion image segmentation. Based on the FCNN method, excellent results have proved the strong feature segmentation ability of convolutional neural network (CNN) in image segmentation. However, due to the limited convolution receptive field, it has certain limitations, which restrict its performance in image segmentation [10]. To address these issues, Attention U-Net proposed the Attention Gate block [11], which extracts locally important features through attention mechanisms and dynamically learns the weights of attention through forward feedback. However, these methods still have certain limitations in modeling long-term dependencies. Recently, the success of Transformers in the natural language processing (NLP) [12] field has attracted attention. Researchers have successfully developed Vision Transformer (ViT) [13] for the visual domain. Through multi-head attention mechanisms (MSA), Transformers can effectively establish long-term dependencies between global contexts, and their performance on large datasets is comparable to CNNs. TransUnet [14] which developed based on ViT, was the first to introduce Transformers into the field of image segmentation, demonstrating that Transformer-based methods yield better results than CNN-based methods. Through further design, TransNorm [15] attempted to integrate Transformers into skip connections and achieved good results. However, ViT requires training on large datasets and is affected by quadratic complexity. Moreover, TransUnet still relies on CNN for hierarchical feature extraction. In DeiT [16], the authors proposed a data-efficient image transformer that enables training Transformers on medium-scale datasets. In Swin Transformer [17], the authors proposed a hierarchical transformer that applies window-based computing to reduce the computational complexity of ViT, achieving tremendous success. Subsequently, Swin-Unet [18] improved U-Net by using Swin Transformer as the backbone network without requiring convolutional operations, further enhancing segmentation performance. The emergence of networks such as CrossViT [19], Attention Swin-Unet [20], and HiFormer [21] demonstrated the powerful performance of multiscale feature representation on ViT. Multiscale feature representation can effectively help model the remote relationship of feature information and fuse low-level and high-level features to further capture fine-grained features.

Although Transformer-based visual approaches are effective in modeling global contextual information, they treat images as one-dimensional sequences and ignore their two-dimensional structure. This forced change in the image dimensions can lead to loss of accuracy in local and global features [23], which ultimately results in suboptimal segmentation performance.

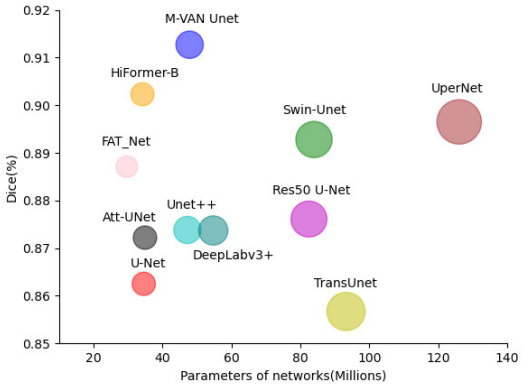


FIGURE 1. Model parameter comparison chart. our approach is compared to other methods in terms of parameter quantity and Dice results obtained on the ISIC2018 dataset [22], and it outperforms them.

In this paper, we drew inspiration from Swin-Unet [18] and made innovative improvements to the U-Net architecture. Similar to Swin-Unet, our network structure consists of four parts: encoder, bottleneck, decoder, and skip connections. In the construction of the segmentation network, we employed the VAN block as the new backbone network. The encoder and decoder were constructed using DownSample block and the VAN block, respectively. The DownSample block facilitated downsampling and dimension expansion to learn deeper feature representations. The decoder performed upsampling through deconvolutional operations and fused the output features from the residual connections and skip connections, while further extracting features using the VAN block. To enhance the expressive power of the model, we introduced the LKA-Cross Attention block at the skip connection level to fuse the low-level features from the encoder with the high-level features obtained from the

decoder's upsampling. This cross-level fusion enabled better capture of the correlation between features at different levels, thereby enhancing the model's expressive power. At the bottleneck position, we integrated the ideas from Inception [24] and SENet [25] and designed the MSC Attention block. This block utilizes the PSA block and VAN block with different receptive field sizes to extract multiscale information and employs a multi-channel attention mechanism to learn the relationships and semantic features between different channels. This design effectively captures important multiscale contextual information in medical images, particularly in cases with significant variations in target sizes and shapes. As shown in FIGURE 1, our proposed method not only outperforms other models in terms of Dice coefficient but also has fewer parameters compared to most models. Through multiple ablation experiments, we have demonstrated the effectiveness of the proposed structural design. Our contributions are as follows:

- Based on the VAN block, we have redesigned the segmentation network, referred to as the M-VAN Unet model.
- We have designed the LKA-Cross Attention block, a novel cross-attention mechanism that enhances the fusion of low-level and high-level features in skip connections.
- We have designed the MSC Attention block, a novel multi-scale channel attention mechanism, for enhancing the bottleneck structure.

II. RELATED WORKS

A. CNN-BASED SEGMENTATION NETWORKS

Currently, CNN has become the de facto standard for medical image segmentation tasks. In particular, with the emergence of U-Net, researchers have started to focus on improving the U-shaped encoder and decoder structure, which has the advantages of simplicity, excellent performance, and modular design. For example, H-DenseU-Net [26] replaces the original U-Net encoder with a residual network and dense skip connections to extract more complex features. U-Net++ [7] inherits the structure of U-Net and also borrows the dense connection method of DenseNet [27]. It redesigns the skip connection structure between the encoder and decoder with dense connections between each layer to bridge the semantic gap of feature maps between the encoder and decoder. This method is more efficient than increasing the feature map resolution to capture more feature information. In addition, many researchers have also borrowed and improved this structure [28], [29]. Meanwhile, Oktay et al. [11] also proposed skip connection structures using Attention Gate block, which extract locally important features through attention mechanisms, allowing the network to focus on specific important objects while ignoring redundant regions. However, the receptive field of convolutional operations is limited by the size of the convolution kernel. Therefore, these CNN-based methods share a common drawback, which is the limitation in capturing long-range dependencies. In fact, the locality and weight-sharing properties of convolutional operations make them unable to understand global context. In CNN, the problem of limited receptive fields is a common challenge. In the past few years, many scholars have proposed different solutions. Among them, Yu et al. [45] proposed a new convolution method, which introduces dilated convolution into the convolutional kernel to increase the receptive field of the CNN, thereby improving its performance. Huang et al. [31] proposed an attention-based method based on self-attention, called Criss-Cross Attention (Ccnet), which can simultaneously consider contextual information in different directions, better extracting image features and improving the accuracy of image segmentation. Although self-attention was originally proposed to solve machine translation problems, it has been successfully applied to image segmentation tasks due to its spatial adaptability. In addition, Meng et al. [23] proposed a VAN model that combines the advantages of CNN and self-attention, with local structural information, long-term dependencies, and adaptability, and is considered a method that surpasses other backbone networks based on CNN and Transformer. In the next section, we will discuss in detail the characteristics and advantages of VAN.

B. TSRANSFORMERS

In recent years, the success of Transformer in the field of NLP has received widespread attention, making it an important milestone in the history of deep learning. Compared to traditional recurrent neural network (RNN) and CNN, Transformer models the dependencies between any two positions in a sequence through self-attention mechanism, which can better handle long sequences and capture long-range dependencies. In the field of computer vision, more and more Transformer-based methods are emerging. Among them, ViT was the first method applied to the visual domain, which divides the input image into multiple patches and inputs them into a Transformer encoder, and then feeds the output to an MLP layer for classification. Subsequently, TransUnet [14] based on ViT was the first to introduce Transformer into the field of image segmentation, and its performance on large datasets was comparable to that of CNN. In order to solve the problem that ViT needs large datasets to show its advantages, a series of improved methods have been proposed. Swin Transformer [17] proposed a hierarchical transformer that applies a sliding window to reduce the computational complexity of ViT, and based on this, Swin-Unet [18] was proposed for segmentation. DeiT [16] proposed a data-efficient image transformer that enables the method to show advantages on small and medium-sized datasets. Networks such as CrossViT [19], HiFormer [21], and Cat [30] further improve network performance by extracting features at multiple scales using cross-attention. Although Transformer-based visual approaches can effectively model global contextual information and overcome the limitations of large datasets, they treat images as one-dimensional sequences and ignore their two-dimensional structure. Forcibly changing the dimension of the image may lead to loss of accuracy in

local and global features, ultimately affecting segmentation performance. Therefore, it is necessary to further explore how to preserve the two-dimensional structural information of images to improve segmentation performance.

C. MULTI-SCALE CHANNEL ATTENTION MECHANISM

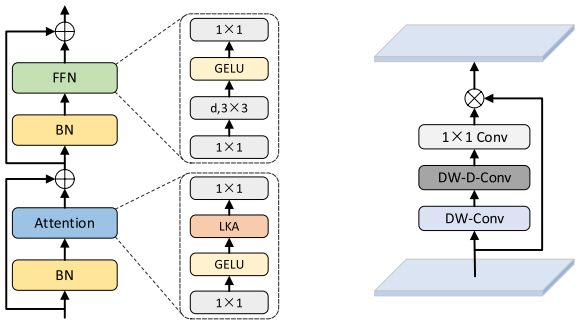
Several studies have shown [32], [33], [34], [35] that embedding a multi-channel attention block into an existing CNN network can significantly improve its performance. The SENet [25] network learns the dependencies between channels in the form of channel attention, and automatically learns the importance of each channel in the feature map. This improves the channels of the feature map that are useful for the current task and suppresses the feature channels that are not useful for the current task. Wang [34] believe that the dimensionality reduction in SENet has side effects on the channel attention mechanism, and capturing all the dependencies between channels is inefficient and unnecessary. They made some improvements to SENet and used an efficient channel attention (ECA) block directly after the global average pooling layer, replacing the fully connected layer with a 1×1 convolution (Conv) layer to avoid dimensionality reduction and efficiently capture feature information between channels. EPSANet [36] introduced the idea of multi-scale and PSA block based on multi-channel attention. It replaced the bottleneck structure of ResNet [37] with a new block efficient pyramid split attention (EPSA), which can be used as a ''plug-and-play'' block for existing backbone networks and significantly improve their performance.

III. METHODS

In this study, we propose an end-to-end segmentation network model called M-VAN Unet, as shown in FIGURE 3. The proposed model consists of an encoder, decoder, LKA-Cross attention block, and MSC-Attention block. M-VAN Unet is designed based on a U-shaped structure and embedded with the VAN backbone network. It relies on the large kernel attention (LKA) mechanism to enhance contextual feature capturing and address long-range dependency and adaptability issues. We proposed the LKA-Cross Attention block to enhance the skip connections between the encoder and decoder. It is a two-level cross-attention mechanism designed to enhance the semantic fusion of low-level and high-level features. We also designed the MSC-Attention block at the bottleneck, which is a multi-scale attention mechanism used to extract critical multi-scale and multi-channel contextual information.

A. ENCODER

Our encoder draws inspiration from the conversion approach of Swin-Unet [18]. We replaced the Swin Transformer block with the VAN block [23]. Moreover, we substituted the Linear Embedding layer and Patch Merging layer with DownSample block to enable linear dimension expansion and downsampling. The spatial resolution of the encoder and bottleneck structure forms a sequence of four stages, denoted by



(a) VAN (b) LKA

FIGURE 2. VAN block and LKA block diagram. d means depth wise convolution.

$\frac{H}{2} \times \frac{W}{2} \times C$, $\frac{H}{4} \times \frac{W}{4} \times 4C$, $\frac{H}{8} \times \frac{W}{8} \times 8C$ and $\frac{H}{16} \times \frac{W}{16} \times 16C$. In this context, H and W respectively represent the height and width of the input image, while $\boldsymbol{C}$ refers to the number of channels. The features of the first three stages' resolutions are subjected to feature learning by multiple consecutive VAN block, while the features of the last stage's resolution are fed into a bottleneck structure for multi-scale channel learning. The feature information is downsampled at each stage through DownSample block and then undergoes a 1×1 Conv and a VAN block for learning before entering the next stage. Here, the DownSample block consists of a 3×3 Conv layer, a Batch Normalization layer (BN), and a ReLU layer. All other layers in each stage maintain the same input dimensions, i.e., spatial resolution and channel number. Due to the loss of feature information in deep networks, the outputs of the first three VAN block stages, containing features of different scales, are effectively fused with the high-level feature information of the corresponding stage of the encoder through LKA-Cross attention block using skip connections.

B. VAN BLOCK

The VAN block is built on the LKA mechanism [23] and comprises two main components: the Attention block based on LKA, and the Feedforward Neural Network (FFN) block, as shown in FIGURE 2. Prior to each block, a BN is applied for normalization, and the output is obtained by adding L copies of the feature maps to the outputs of the Attention and FFN block through skip connections. In the Attention block, the feature maps are first passed through a 1×1 Conv layer, GELU activation function, LKA, and another 1×1 Conv layer before the final output is obtained. Similarly, in the FFN block, the feature maps are passed through a 1×1 Conv layer, 3×3 Depthwise Conv, GELU activation function, and another 1×1 Conv layer before the final output is obtained. The computation formula for the VAN block is described as follows:

$$\boldsymbol{F^l} = BN(Attention(\boldsymbol{F^{l-1}})) + \boldsymbol{F^{l-1}}, \quad (1)$$

$$\boldsymbol{W} = BN(FFN(\boldsymbol{F^l})) + \boldsymbol{F^l}. \quad (2)$$

$\boldsymbol{F^{l-1}}$ represents the input features of VAN, while $\boldsymbol{F^l}$ and $\boldsymbol{W}$ represent the output features of the attention and FFN

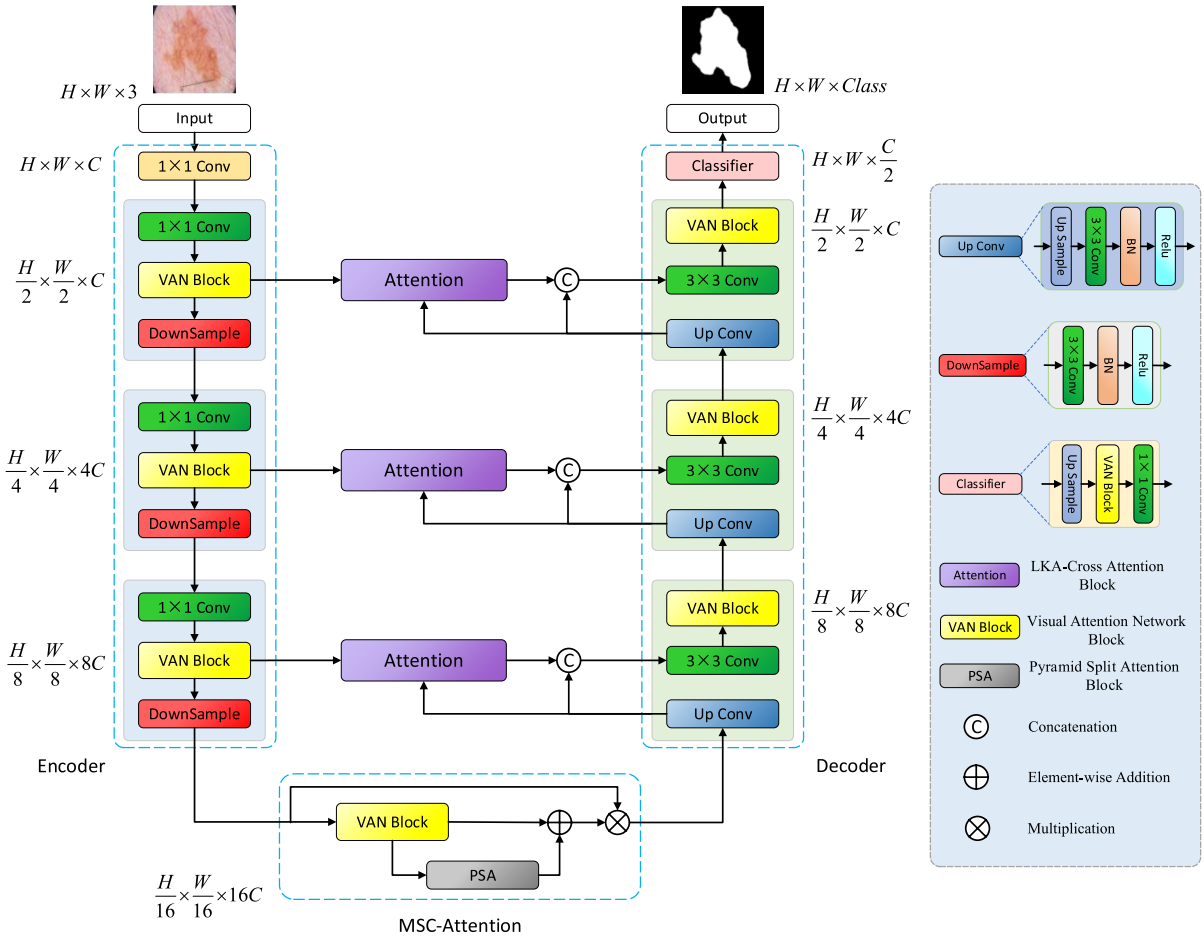


FIGURE 3. Model architecture of M-VAN Unet. The M-VAN Unet consists of an encoder, LKA-Cross attention block, MSC-Attention block, and a decoder. The encoder and decoder are constructed using the VAN block. The LKA-Cross attention block component is built using the LKA technique, while the MSC-Attention block component combines the VAN and PSA techniques.

components. In this context, $\boldsymbol{F^{l-1}}, \boldsymbol{F^{l}}, \boldsymbol{W} \in \mathbb{R}^{C\times H\times W}$. The attention calculation formula is as follows:

$$LKA = Conv_{1\times 1}(DW\_D\_Conv(DW\_Conv(\boldsymbol{F^{l}}))), \quad (3)$$

$$Attention = LKA \otimes \boldsymbol{F}. \quad (4)$$

In this context, $\boldsymbol{F}, \boldsymbol{Attention} \in \mathbb{R}^{C\times H\times W}$, $\otimes$ means element-wise product. LKA refers to large kernel convolution, which is decomposed into three parts: spatial local convolution (depthwise convolution), spatial long-range convolution (depthwise dilated convolution), and channel convolution (1×1 Conv). LKA can capture long-term relationships with relatively small computational costs and parameters. Unlike common attention methods, LKA does not require additional normalization functions such as Sigmoid and Softmax. Researchers [23] have found that the key feature of the LKA attention mechanism is the adaptive adjustment of the output based on the input features, rather than the normalized attention map.

C. DECODER

As the model follows the symmetry of the U-shaped architecture, three VAN block are used as decoders corresponding to the three stages of the encoder. In order to maintain consistent feature dimensions across each stage, we used an UpConv layer to reduce the feature dimensionality and gradually increase the spatial resolution through upsampling. The UpConv layer consists of an Upsample function, a 3×3 Conv, a BN layer, and a ReLU activation function. The resolution is increased by a factor of two by controlling the stride and padding. Since upsampling does not bring additional information and can also impact image quality, the high-level feature information obtained through upsampling is effectively fused with the low-level feature information from the corresponding stage encoder through LKA-Cross attention block. The fused features are then concatenated with the upsampling feature information and sent to the VAN block for feature learning. Finally, the features are passed through a classifier, where the resolution is restored to that of the input image by upsampling. The feature channel number C is transformed into the final segmentation class output through a VAN block and a 1×1 Conv.

D. LKA-CROSS ATTENTION BLOCK

The purpose of the skip connections in U-Net is to provide fine-grained feature information to the decoder and

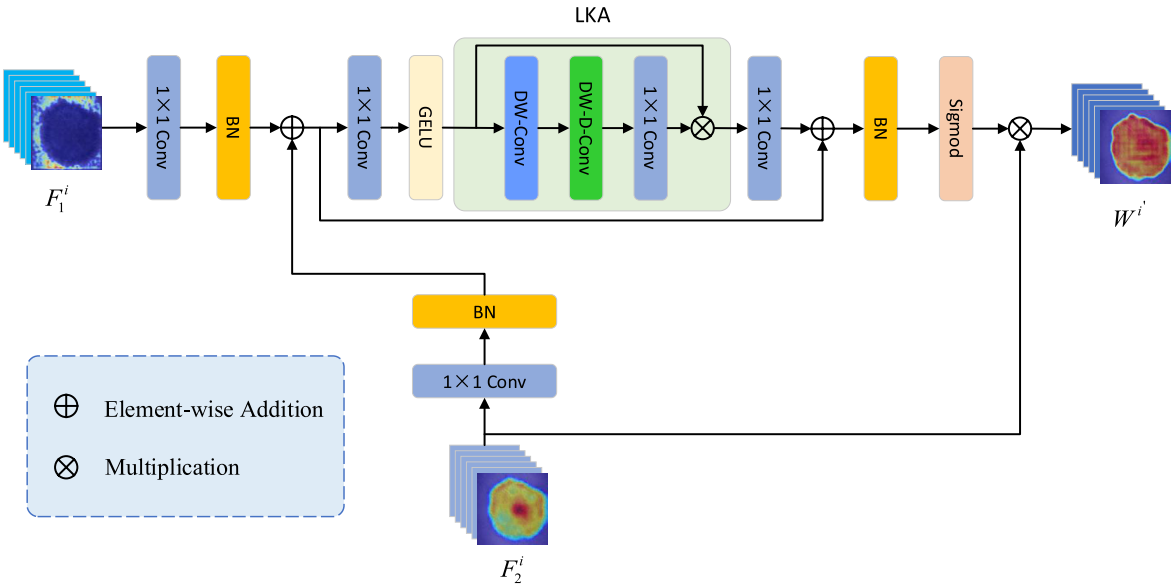


FIGURE 4. LKA-Cross attention block. The low-level features and high-level features are separately input into a LKA block for learning and fusion. After being processed through batch normalization and sigmoid activation, they are fused again through a secondary attention learning process with a copy of the high-level features. Finally, the output results are obtained.

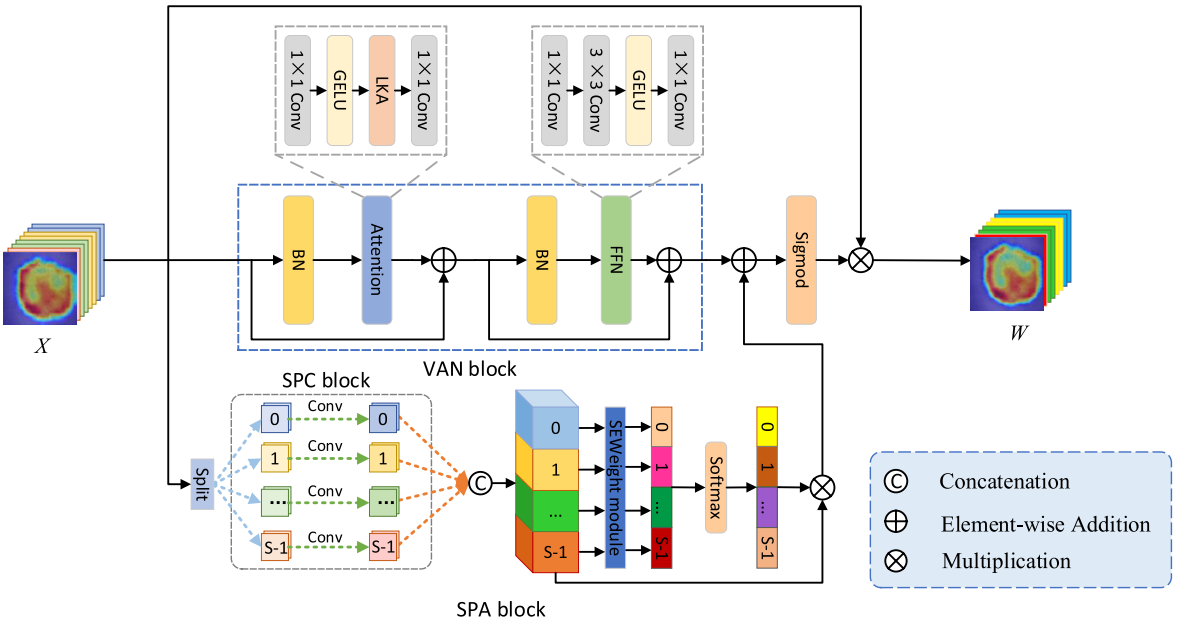


FIGURE 5. MSC-Attention block. This structure comprises of three channels. The first channel includes skip connection attention, the second channel includes a VAN block, and the third channel includes a PSA block.

to address the problem of feature information loss caused by deep networks and upsampling. This is crucial for feature learning in the process of image segmentation. Some researchers have redesigned the skip connection part in several extended U-Net architectures [7], [39], which has yielded promising results. This highlights the necessity of research on the skip connection part. In this work, we also designed a cross-attention mechanism with multi-level attention called LKA-Cross Attention to enhance feature fusion. Our design is illustrated in FIGURE 4. In the same layer, the feature information copy ($\boldsymbol{F_1^i}$) obtained from the VAN block in the encoder is processed with a 1 × 1 Conv and Batch Normalization, and then fused with the feature information ($\boldsymbol{F_2^i}$) that undergoes the same processing using the add method. Next, the fused feature enters the first-level LKA block for feature learning. After undergoing BN, Sigmoid activation, and other steps, the feature is fused with a copy of input feature $\boldsymbol{F_2^i}$ using second-level attention fusion, and the resulting feature is denoted as $\boldsymbol{W^{i'}}$. In the decoder, we adopt a method similar to the original skip-connection in U-Net for the fusion of the LKA-Cross attention block output feature $\boldsymbol{W^{i'}}$ and the upsampled deep feature $\boldsymbol{F_2^i}$. This approach aims to reduce

spatial information loss caused by excessive downsampling and overly deep networks. Subsequently, a linear layer is used to adjust the dimensionality of the connected features to match that of the upsampled feature, and the resulting feature is sent for feature learning in the VAN block. Finally, the i-th decoder block outputs feature W^i. The entire process can be represented by the following equation:

$$F^i(F_1^i, F_2^i) = Conv_{1\times1}(BN(F_1^i)) + Conv_{1\times1}(BN(F_2^i)), \tag{5}$$

$$W^{i'} = Sig \bmod (BN(F^i(F_1^i, F_2^i))) \otimes F_2^i, \tag{6}$$

$$W^i = VAN(\mathrm{LN}(Cat(W^{i'}, F_2^i))). \tag{7}$$

The variable i in the formula denotes the cross-attention fusion that takes place at the i-th layer. $i \in \{1, 2, 3\}$. F_1^i, F_2^i, $W^{i'}$, $W^i \in \mathbb{R}^{H\times W\times C}$.

In FIGURE 4, we visualized the activation heatmaps of features F_1^i, F_2^i and W^i using the Grad-CAM [40] technique. The figure shows that the feature attention of the low-level feature F_1^i is too scattered and not entirely focused on the target. On the other hand, the high-level feature F_2^i is overly concentrated on the center of the target, causing the edges to gradually lose focus. After the fusion of LKA-Cross attention block, the activation heatmap shows a more comprehensive attention to the target lesion.

E. MSC-ATTENTION BLOCK

In the bottleneck structure of a network, as the image resolution decreases and the feature dimension increases, there is a problem in medical images where the size and shape of the object vary greatly, making it difficult to capture features effectively. Research on networks such as SENet [25] and EPSANet [36] has shown that learning through different channels can effectively capture spatial information of feature maps at different scales, enriching the feature space scale. Furthermore, learning through multiple channels can strengthen the required features while suppressing the unnecessary ones.

In this network, we have designed a two-level multi-scale channel attention mechanism called MSC-Attention, as shown in FIGURE 5, that uses a residual connection channel and two different feature extraction channels within the block. The first residual connection channel is designed to address feature loss problems. The second channel is a continuation of the fourth stage of the encoder and is consistent with the Swin-Unet [18] bottleneck structure. The third channel incorporates the PSA block [36], which is an efficient multi-scale channel attention mechanism. The multi-scale feature extraction in the PSA block is achieved through the use of the split and concat (SPC) block [36]. The SPC block divides the input feature map (X) into S parts $(X_0, X_1, \ldots X_{S-1})$, each with $C' = \frac{C}{S}$ channels, where $X_i \in \mathbb{R}^{C'\times H\times W}$, $i \in \{0, 1, 2, \ldots S-1\}$. The next step is to extract spatial information from feature maps of different scales using multi-scale convolution. In this case, multi-scale convolution and group convolution are mainly used. The purpose of using group convolution is to reduce the number of parameters, and the size of the Group (G) is adjusted based on the size of the convolution kernel. Therefore, the formula for the multi-scale feature extraction process of the SPC block is as follows:

$$F_i = Conv(K_i \times K_i, G_i)(X_i), \tag{8}$$

$$K_i = 2 \times (i-1) + 1, G_i = 2^{\frac{K_i-1}{2}}, \tag{9}$$

$$F = Cat([F_0, F_1, F_2, F_3]. \tag{10}$$

The variable F represents the output of the SPC block. $i \in \{0, 1, 2, \ldots S-1\}$, $F_i \in \mathbb{R}^{C'\times H\times W}$. After extracting multi-scale features from the SPC block, the channel attention weights are computed for different scales of the feature F_i. The weight vector for the entire multi-scale channel attention is then integrated, and the calculation formula is shown below:

$$Z_i = SEWeight(F_i), i \in 0, 1, 2, \ldots, S-1, \tag{11}$$

$$Z = Z_0 \oplus Z_1 \oplus Z_2 \oplus Z_3. \tag{12}$$

To establish a long-term channel attention dependency and facilitate information interaction among multiple scales of channel attention, we utilize Softmax to re-weight the channel attention information. The corresponding feature map F_i at each scale is then multiplied at a Channel-Wise level with the attention vector (Att_i) that has been re-weighted. Finally, the resulting feature maps, which have been weighted by multi-scale channel attention, are concatenated dimensionally to form the feature map W'. After fusion with the feature information from the second channel and sigmoid activation, the resulting feature map W is multiplied with X to generate a more enriched feature map W. The formula for this calculation is shown below:

$$Att_i = Soft\max(Z_i) = \frac{\exp(Z_i)}{\sum_{i=0}^{S-1}(Z_i)}, \tag{13}$$

$$W' = Cat\left(\left[F_i \cdot Att_i, \ldots, F_{S-1} \cdot Att_{S-1}\right]\right) \tag{14}$$

$$i \in 0, 1, 2, \ldots S-1, \tag{15}$$

$$W = Sig \bmod (W' + VAN(X)) \times (X). \tag{16}$$

IV. EXPERIMENTS

To validate the effectiveness of our approach, we conducted relevant experiments on two skin lesion datasets. In this section, we introduce the datasets, the detailed implementation details of the experiments, and the evaluation metrics. We also provide comparisons of the experimental results and segmentation effect diagrams, and discuss and analyze the experimental results in detail.

A. DATASET

1) ISIC 2018 DATASET

The ISIC2018 skin lesion dataset [22] is a widely used dataset of skin diagnostic images provided by the ISIC Foundation. The dataset includes 2594 training set images

TABLE 1. Comparison results of methods on ISIC 2018 dataset and PH^2 dataset.

Methods	ISIC2018						PH^2					
	Dice	HD95	SE	SP	JS	ACC	Dice	HD95	SE	SP	JS	ACC
U-Net [4]	0.8625	3.0400	0.8722	0.9582	0.7908	0.9260	0.9073	1.5621	0.9305	0.9548	0.8372	0.9339
Res50-Unet [6]	0.8761	2.0614	0.9364	0.9340	0.7940	0.9344	0.8974	2.2206	0.9168	0.9401	0.8215	0.9311
U-Net++ [7]	0.8738	2.5327	0.8588	0.9737	0.7773	0.9484	0.9113	1.5545	0.9095	0.9653	0.8420	0.9382
TransUnet [14]	0.8567	3.6225	0.8449	0.9713	0.7524	0.9427	0.9191	1.5910	0.9498	0.9199	0.8550	0.9418
DeepLab v3+ [42]	0.8737	2.4399	0.8923	0.9171	0.7980	0.9339	0.9136	1.8424	0.9391	0.9636	0.8447	0.9386
Upernet [43]	0.8965	2.4742	0.9014	0.9525	0.8295	0.9461	0.9393	1.2076	0.9511	0.9710	0.9049	0.9707
FAT-Net [44]	0.8871	2.4201	0.8937	0.9123	0.8096	0.9314	0.9233	1.9921	0.9313	0.9684	0.8626	0.9448
HiFormer-B [21]	0.9023	1.7413	0.9228	0.9577	0.8322	0.9473	0.9446	1.1428	0.9425	0.9729	0.8863	0.9624
Attention Unet [11]	0.8722	2.5187	0.9170	0.9394	0.7921	0.9335	0.9103	1.6840	0.9198	0.9602	0.8362	0.9455
Swin-Unet [18]	0.8928	2.4692	0.9102	0.9628	0.8301	0.9435	0.9434	1.1899	0.9418	0.9720	0.8861	0.9638
M-VAN Unet	0.9127	1.7041	0.9258	0.9796	0.8417	0.9643	0.9508	1.1199	0.9480	0.9770	0.9071	0.9729

(20.0% melanoma, 72.0% nevus, 8.0% seborrheic keratosis), 100 validation set images, and 1000 test set images, with pixel ranges varying from 0.5 to 29 million pixels. As the resolution of each photo is different, we have uniformly changed it to 224 × 224 resolution. A unique feature of the ISIC2018 dataset is that it not only provides the images themselves but also provides pixel-level annotations of these images, accurately segmenting the lesions in the images. This makes the dataset useful for developing and evaluating skin lesion segmentation algorithms.

2) PH^2 DATASET

PH^2 dataset [41] is an image dataset for skin lesion classification and segmentation, developed with funding from the Portuguese Science and Technology Foundation (FCT). The dataset includes 200 skin lesion images captured by a handheld digital camera, of which 80 are benign melanoma images, 80 are malignant melanoma images, and 40 are normal skin images. Each input image has a resolution of 768 × 560 pixels. For ease of training, we resized the images to a uniform resolution of 224 × 224, and split the dataset into a training set and a test set in an 8:2 ratio.

B. IMPLEMENTATION DETAILS

Our model algorithm was implemented using Python 3.9 and PyTorch 1.11.0. The experiments were conducted on a system running Windows 11, with an AMD R5 5600X CPU and 32GB of memory. Training was performed on an RTX3090 GPU with 24GB of memory. To increase the diversity of training data, improve the generalization performance and robustness of the model, we applied data augmentation using the Albumentations library [38], including random flips, rotations, scaling, occlusions, and Gaussian blurring. In order to ensure a fair comparison with other methods, we standardized the input image resolution to 224 × 224 and set the batch size to 16. Specifically, after applying data augmentation to the images, we utilized the Resize function from the Albumentations library to scale the resolution of the images to 224 × 224. Our method employed the SGD optimizer for backpropagation with a momentum of 0.9 and weight decay of 1e-4. The learning rate was adjusted using the StepLR mechanism, with a decay factor of 0.1 every 20 steps, starting from an initial learning rate of 0.01. We performed a total of 200 iterations during the training process. CrossEntropyLoss and DiceLoss were combined as the loss functions for training.

C. COMPARISON RESULTS

In our experiments on two skin lesion segmentation datasets, we used different evaluation metrics to provide a comprehensive evaluation of our method. The evaluation metrics used were the Dice coefficient (Dice), 95% Harsdorf distance (HD95), sensitivity (SE), specificity (SP), Jaccard index (JS), and Accuracy (ACC).

1) ISIC2018 SEGMENTATION

We conducted comparative experiments between our proposed method and other methods based on CNN and Transformer on the ISIC 2018 dataset. To ensure consistency, all training and testing were conducted on the same device. As shown in TABLE 1, the results demonstrate that our proposed method outperforms other methods in terms of evaluation metrics on the ISIC 2018 dataset. Specifically, compared to the pure convolutional scheme, our method increased the Dice coefficient by 3.89% and 1.62% compared to Unet++ and Upernet, respectively. Compared to the Transformer scheme, our method increased the Dice coefficient by 5.6% and 1.99% compared to TransUnet and Swin-Unet, respectively. Compared to the hybrid scheme, our method increased the Dice coefficient by 1.04% compared to HiFormer-B. This shows that our method has excellent segmentation performance. We show the visual results of skin lesion segmentation in FIGURE 6, where the green line

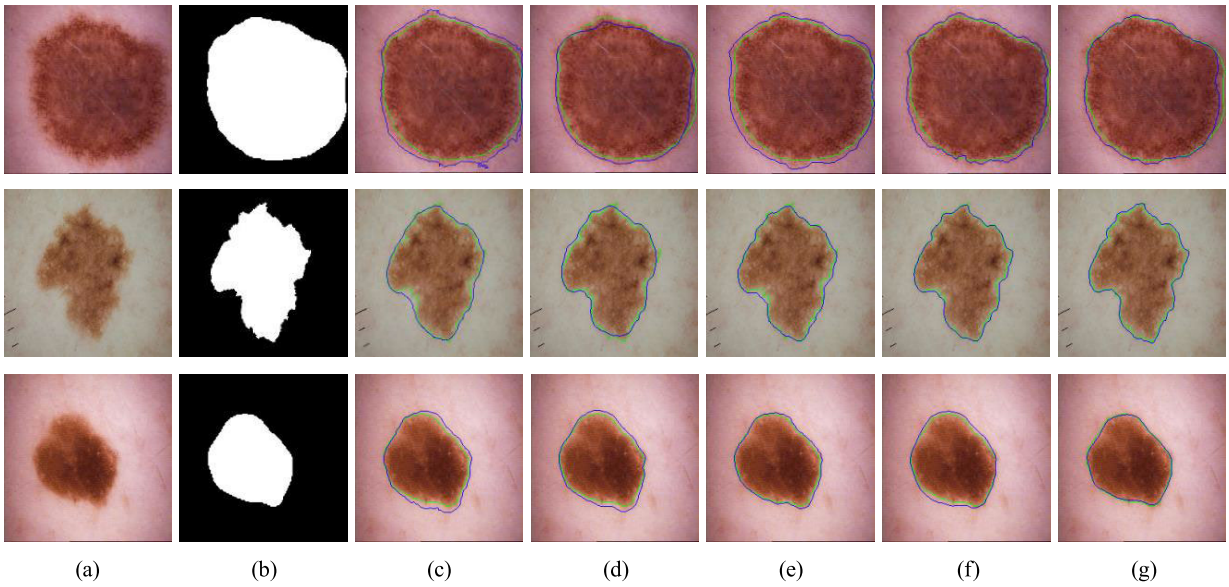


FIGURE 6. Vishual comparison of different methods on the ISIC 2018 skin lesion segmentation dataset. The true boundary is shown in green, and the predicted boundary is shown in blue. (a) Input Image. (b) Ground Truth. (c) U-Net. (d) Attention Unet. (e) FAT-Net. (f) Swin-Unet. (g) Our method.

represents the true lesion boundary and the blue line represents the predicted boundary. Compared to the segmentation results of other methods, our method generates smoother and more accurate edge contours that are closer to the true boundary. This indicates that our method can capture more detailed structures and generate more accurate edge contours.

2) PH^2 SEGMENTATION

In order to conduct comparative experiments on the PH^2 dataset, we used the pre-trained weights from ISIC2018, and the training and testing methods were the same as those used on the ISIC2018 dataset. As shown in TABLE 1, our method achieved better evaluation metrics than other methods, which is similar to the results obtained on the ISIC2018 dataset. Specifically, compared to the pure convolutional scheme, our method improved the Dice coefficient by 3.95% and 1.15% compared to Unet++ and Upernet, respectively. Compared to the Transformer scheme, our method improved the Dice coefficient by 3.17% and 0.74% compared to TransUnet and Swin-Unet, respectively. Compared to the hybrid scheme, our method improved the Dice coefficient by 0.62%. Overall, our method demonstrated superior learning ability in terms of the HD95, SE, SP, JS, and ACC metrics. These comparative experiments on the two datasets demonstrate that our method has satisfactory generalization ability across different datasets.

D. ABLATION STUDY

To better investigate the influence of different factors on the effectiveness of our method, we conducted ablation experiments on the ISIC 2018 dataset to verify the effectiveness of the VAN block, LKA-Cross attention block, and MSC-Attention block. Additionally, we investigated the influence of different input resolutions on the segmentation results. These ablation experiments aim to thoroughly evaluate the importance of different components of our method and provide us with more detailed conclusions to better understand the role of these blocks in improving our method. Through the ablation experiments, we can clearly see that the VAN block, LKA-Cross attention block, and MSC-Attention block significantly improve our method. Moreover, different input resolutions were found to have varying effects on the segmentation results. This further confirms the effectiveness and practicality of our method's innovations in practical applications.

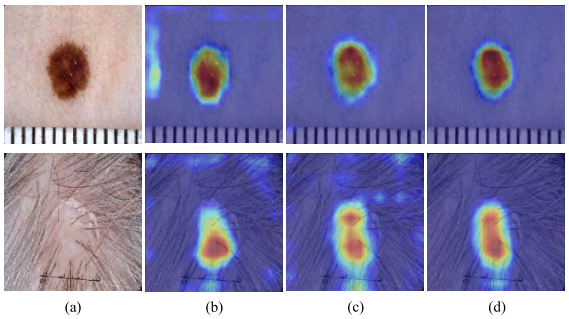


FIGURE 7. The visualization of bottleneck structure ablation experiments on the ISIC2018 skin lesion segmentation dataset was performed using Grad-CAM. (a) Represents the input image. (b) Displays the feature map of the bottleneck structure without MSC-Attention. (c) Illustrates the feature map of the bottleneck structure before MSC-Attention. (d) Presents the feature map of the bottleneck structure after MSC-Attention.

1) IMPACT OF THE VAN BLOCK

To investigate the performance improvement of the VAN block in our base network, we conducted ablation

experiments. To avoid interference, we used the basic U-Net model as a Baseline. To ensure fairness in the experiment, we did not use the LKA-Cross attention block and MSC-Attention block proposed in our model in this comparative experiment. Baseline+ResNet50 is a model that replaces the feature extraction block of the Baseline model with the ResNet-layers of ResNet50. Baseline+VAN model is a model that replaces the feature extraction block of the Baseline model with the VAN block.

The ablation experiment results on the ISIC2018 dataset are shown in TABLE 2, indicating that the segmentation performance of the VAN block is higher than that of the Baseline and ResNet50. Specifically, the Baseline+VAN model improved the evaluation metrics of Dice and ACC by 2.74% and 2.09%, respectively, compared to the Baseline model. Moreover, it also increased by 1.45% and 0.96% compared to the Baseline+ResNet50 model, respectively. This suggests that the VAN block is indeed effective in improving the segmentation performance of the network, which is consistent with our previous hypothesis that the VAN block can effectively capture global features. We believe that the reason why the VAN block can effectively enhance segmentation performance is due to the existence of large-kernel convolutional attention. This is achieved through a combination of deep convolution, dilated convolution, and channel convolution, which increases the receptive field without significantly increasing the number of parameters. The expansion of the receptive field implies a better ability to capture global features, thereby improving the model's performance.

TABLE 2. Ablation experiments of the van block on ISIC 2018.

Model	Dice	ACC
Baseline	0.8639	0.9212
Baseline+ResNet50	0.8768	0.9325
Baseline+VAN	**0.8913**	**0.9421**

2) IMPACT OF THE LKA-CROSS ATTENTION BLOCK

To explore the impact of the LKA-Cross attention block on model performance, we conducted multiple experiments. As presented in TABLE 3, when the LKA-Cross attention block was added to the model, the segmentation performance improved compared to the baseline VAN Unet. Specifically, we conducted ablation tests on the ISIC2018 dataset and found that the LKA-Cross attention block improved the Dice coefficient and ACC by 1.23% and 1.83%, respectively. This indicates the effectiveness of fusing low-level and high-level features at the jump connection. Furthermore, by visualizing the heat map of feature activations in FIGURE 4, we found that integrating the LKA-Cross attention block led to a more comprehensive focus on the target lesion, illustrating the validity of the LKA-Cross attention block in the medical image segmentation task. We further analyzed the factors contributing to the performance improvement of the model through the LKA-Cross attention block. In the medical image segmentation task, interactions among different feature maps can help the model better fuse the information of these maps, thus improving the segmentation accuracy. Specifically, the LKA-Cross attention block allows the model to pay more attention to those features that are relevant to the task, while suppressing those that are unrelated to the task, making the model's judgment more accurate.

TABLE 3. Ablation experiments of the lka-cross attention block on ISIC 2018.

M-VAN Unet	LKA-Cross attention	Dice	ACC
✔	✘	0.8913	0.9421
✔	✔	0.9036	0.9604

3) IMPACT OF THE MSC-ATTENTION BLOCK

Analysis of the experimental results in TABLE 4 indicates that the segmentation accuracy is significantly improved by using the MSC-Attention block compared to the baseline model. Specifically, in the ablation experiments conducted on the ISIC2018 dataset, the MSC-Attention block improved the Dice coefficient and ACC by 1.14% and 2.03%, respectively. This result confirms the effectiveness of the MSC-Attention block, and indirectly confirms the effectiveness of the PSA block. FIGURE 7 presents a visual comparison between the baseline VAN Unet model and the model with the multiscale attention bottleneck structure using the Grad-CAM tool. It is noteworthy that attentional distraction is more prominent in the baseline model (FIGURE 7 (b) compared to FIGURE 7 (c)), whereas after processing the features using the bottleneck structure of the MSC-Attention block, the feature activation heat map in FIGURE 7 (d) demonstrates more accurate attention than in FIGURE 7 (b) and FIGURE 7 (c). This result further confirms the importance of the MSC-Attention block in improving segmentation accuracy, and provides intuitive visualization support for a deeper understanding of its effectiveness.

4) INPUT RESOLUTION INFLUENCE

In the aforementioned comparative experiments, we used an input resolution of 224×224 pixels for the skin lesion dataset. To investigate the impact of input resolution on segmentation results [46], we conducted additional ablation experiments with low-resolution (112×112) and high-resolution (384×384) inputs, as shown in TABLE 5, to assess the influence of this factor on model performance.

The results demonstrate that as the input resolution increases, the segmentation results become more accurate. The most significant improvement is observed when transitioning from 112×112 to 224×224 input resolution. High-resolution input samples provide finer details, leading to better segmentation results. Although high-resolution inputs yield higher Dice and ACC scores, indicating the model's segmentation capability, they also introduce

TABLE 4. Ablation experiments of the msc-attention block on ISIC 2018.

M-VAN Unet	Multi-attention	Dice	ACC
✔	✘	0.8913	0.9421
✔	✔	0.9027	0.9624

TABLE 5. Ablation experiments of the Input the resolution on ISIC 2018.

Input Resolution size	Dice	ACC
112 × 112	0.8898	0.9562
224 × 224	0.9127	0.9643
384 × 384	**0.9201**	**0.9698**

higher computational complexity and additional computational costs. On the other hand, low-resolution inputs reduce computational complexity but result in subpar segmentation performance. Therefore, we have chosen 224 × 224 pixels as the input resolution for our model.

V. DISCUSSION AND CONCLUSION

After exploration and experimentation in this paper, we proposed a skin lesion segmentation model based on the VAN with a multi-scale cross attention mechanism. This model significantly improves the current performance of skin lesion segmentation and has the following main contributions and scientific achievements:

Firstly, we redesigned the basic segmentation network by borrowing the Swin-Unet model and introducing the VAN block, which improved the ability to capture global features and led to better segmentation results. Secondly, we proposed the LKA-Cross attention block between the encoder and decoder, which can effectively promote the fusion of low-level and high-level features, thereby improving the feature fusion and segmentation performance of the model. Next, we designed an MSC-Attention block at the bottleneck position, which can effectively capture features at different scales and obtain richer multi-scale feature information, further improving the segmentation performance and robustness of the model. Finally, we further validated the effectiveness and improvement of our method through ablation experiments and Grad-CAM feature visualization. We conducted comprehensive experimental validation of the proposed method, and the results showed that our method outperformed current mainstream methods in evaluation metrics such as Dice, HD95, and ACC.

We believe that the superiority and innovation of the skin lesion segmentation model proposed in this paper will provide strong support for clinical applications and patient treatment.

REFERENCES

[1] U. Leiter, U. Keim, and C. Garbe, "Epidemiology of skin cancer: Update 2019," *Adv. Exp. Med. Biol.*, vol. 1268, pp. 123–139, Sep. 2020, doi: 10.1007/978-3-030-46227-7_6.

[2] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA, A Cancer J. Clinicians*, vol. 73, no. 1, pp. 17–48, Jan. 2023, doi: 10.3322/caac.21763.

[3] D. S. Rigel, J. Russak, and R. Friedman, "The evolution of melanoma diagnosis: 25 years beyond the ABCDs," *CA, A Cancer J. Clinicians*, vol. 60, no. 5, pp. 301–316, Sep. 2010, doi: 10.3322/caac.20074.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[5] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and K. H. Maier-Hein, "NnU-Net: Self-adapting framework for U-Net-based medical image segmentation," 2018, *arXiv:1809.10486*.

[6] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-UNet for high-quality retina vessel segmentation," in *Proc. 9th Int. Conf. Inf. Technol. Med. Educ. (ITME)*, Oct. 2018, pp. 327–331, doi: 10.1109/ITME.2018.00080.

[7] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020, doi: 10.1109/TMI.2019.2959609.

[8] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U^2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404, doi: 10.1016/j.patcog.2020.107404.

[9] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional ConvLSTM U-Net with Densley connected convolutions," 2019, *arXiv:1909.00166*.

[10] R. Azad, A. R. Fayjie, C. Kauffmann, I. B. Ayed, M. Pedersoli, and J. Dolz, "On the texture bias for few-shot CNN segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2673–2682.

[11] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and Ł. Kaiser, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[14] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[15] R. Azad, M. T. Al-Antary, M. Heidari, and D. Merhof, "TransNorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model," *IEEE Access*, vol. 10, pp. 108205–108215, 2022, doi: 10.1109/ACCESS.2022.3211501.

[16] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers distillation through attention," *presented at the 38th Int. Conf. Mach. Learn.*, 2021, pp. 1–11. [Online]. Available: https://proceedings.mlr.press/v139/touvron21a.html

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[18] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino, Eds. Cham, Switzerland: Springer, 2023, pp. 205–218, doi: 10.1007/978-3-031-25066-8_9.

[19] C. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 347–356.

[20] E. Khodapanah Aghdam, R. Azad, M. Zarvani, and D. Merhof, "Attention Swin U-Net: Cross-contextual attention mechanism for skin lesion segmentation," 2022, *arXiv:2210.16898*.

[21] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, "HiFormer: Hierarchical multi-scale representations using transformers for medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, Hawaii, Jan. 2023, pp. 6191–6201.

[22] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.
[23] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," 2022, *arXiv:2202.09741*.
[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, 2017, pp. 4278–4284, doi: 10.1609/aaai.v31i1.11231.
[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[26] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018, doi: 10.1109/TMI.2018.2845918.
[27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[28] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, Jan. 2020, doi: 10.1016/j.neunet.2019.08.025.
[29] Z. Dong, Y. He, X. Qi, Y. Chen, H. Shu, J.-L. Coatrieux, G. Yang, and S. Li, "MNet: Rethinking 2D/3D networks for anisotropic medical image segmentation," 2022, *arXiv:2205.04846*.
[30] H. Lin, X. Cheng, X. Wu, and D. Shen, "CAT: Cross attention in vision transformer," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6, doi: 10.1109/ICME52920.2022.9859720.
[31] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.
[32] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 2018, pp. 3–19.
[33] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1971–1980.
[34] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.
[35] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," 2020, *arXiv:2012.11879*.
[36] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, "EPSANet: An efficient pyramid squeeze attention block on convolutional neural network," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2022, pp. 1161–1177.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, Nevada, Jun. 2016, pp. 770–778.
[38] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020, doi: 10.3390/info11020125.
[39] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 3, pp. 2441–2449, doi: 10.1609/aaai.v36i3.20144.
[40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 618–626.
[41] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, "PH^2—A dermoscopic image database for research and benchmarking," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 5437–5440, doi: 10.1109/EMBC.2013.6610779.
[42] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, vol. 2018, pp. 801–818.
[43] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 418–434.
[44] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, "FAT-Net: Feature adaptive transformers for automated skin lesion segmentation," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102327, doi: 10.1016/j.media.2021.102327.
[45] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
[46] O. Rukundo, "Effects of image size on deep learning," *Electronics*, vol. 12, no. 4, p. 985, Feb. 2023, doi: 10.3390/electronics12040985.

SHUANG LIU was born in Jinzhou, Liaoning, China. She received the Ph.D. degree in traffic information engineering and control from Dalian Maritime University, in 2006.

She finished the postdoctoral research in computer science and technology with the Dalian University of Technology, in April 2015. She is currently a Professor with the School of Computer Science and Engineering, Dalian Minzu University, Dalian, China. Her research interests include machine learning, object detection, knowledge graphs, and scientific visualization. Her academic papers are published both national and international journals and conferences, such as IEEE Access, in 2022, IJCIS, in 2021, and Information, in 2020.

ZENG ZHUANG was born in Xuzhou, Jiangsu, China, in 1996. He received the B.E. degree in software engineering from Beihua University, China. He is currently pursuing the master's degree in computer science and technology with Dalian Minzu University. His primary research interests include machine learning and medical image segmentation.

YANFENG ZHENG was born in Liaoning, China, in 1998. He received the B.E. degree in information management and information system from the Dongbei University of Finance and Economics. He is currently pursuing the master's degree in computer science and technology with Dalian Minzu University. His research interests include computer vision and salient object detection.

SIMON KOLMANIČ was born in Ormož, Slovenia, in 1972. He received the Ph.D. degree in computer science and informatics from the School of Electrical Engineering and Computer Science, Maribor University, Slovenia, in 2006. He is currently teaching with the School of Electrical Engineering and Computer Science, University of Maribor. He also teaches subjects of algorithmic fundamentals, computer graphics and animation, directing subjects in cinema 4D, and Blender: Introduction to geometric modeling and computer graphics, media computer animation, computer graphics, and image processing.