Received: 6 April 2024 | Accepted: 20 May 2024

DOI: 10.1111/srt.13783

Intelligent skin lesion segmentation using deformable attention Transformer U-Net with bidirectional attention mechanism in skin cancer images

Lili Cai[1] | Keke Hou[2] | Su Zhou[1]

[1]School of Biomedical Engineering, Guangzhou Xinhua University, Guangzhou, China

[2]School of Health Sciences, Guangzhou Xinhua University, Guangzhou, China

Correspondence
Keke Hou, School of Health Sciences, Guangzhou Xinhua University, Guangzhou, 510520, China.
Email: houkeke@xhsysu.edu.cn

Funding information
Youth Innovation Talent Project of Guangdong Province (Natural Science), Grant/Award Number: No. 2021KQNCX136; Teaching and Research Project of Guangzhou Xinhua University, Grant/Award Number: No. 2023J037; Education Science Planning Project of Guangdong Province (Higher Education Special Project), Grant/Award Number: No. 2023GXJK623; Youth Innovation Talent Project of Guangdong Province, Grant/Award Number: No.2022WQNCX099; Course Ideological and Political Demonstration Project of Guangzhou Xinhua University, Grant/Award Number: No. 2023KCSZ009; Teaching Quality and Teaching Reform Project of Guangdong Province, with Letter from Higher Education Office of Guangdong Provincial Department of Education [2023] No. 4; Teaching Quality and Reform Engineering Project of Guangzhou Xinhua University, Grant/Award Number: No. 2021TZ001; First-Class Undergraduate Major Construction Project of Guangzhou Xinhua University, Grant/Award Number: 2024YLZY018

Abstract

Background: In recent years, the increasing prevalence of skin cancers, particularly malignant melanoma, has become a major concern for public health. The development of accurate automated segmentation techniques for skin lesions holds immense potential in alleviating the burden on medical professionals. It is of substantial clinical importance for the early identification and intervention of skin cancer. Nevertheless, the irregular shape, uneven color, and noise interference of the skin lesions have presented significant challenges to the precise segmentation. Therefore, it is crucial to develop a high-precision and intelligent skin lesion segmentation framework for clinical treatment.

Methods: A precision-driven segmentation model for skin cancer images is proposed based on the Transformer U-Net, called BiADATU-Net, which integrates the deformable attention Transformer and bidirectional attention blocks into the U-Net. The encoder part utilizes deformable attention Transformer with dual attention block, allowing adaptive learning of global and local features. The decoder part incorporates specifically tailored scSE attention modules within skip connection layers to capture image-specific context information for strong feature fusion. Additionally, deformable convolution is aggregated into two different attention blocks to learn irregular lesion features for high-precision prediction.

Results: A series of experiments are conducted on four skin cancer image datasets (i.e., ISIC2016, ISIC2017, ISIC2018, and PH2). The findings show that our model exhibits satisfactory segmentation performance, all achieving an accuracy rate of over 96%.

Conclusion: Our experiment results validate the proposed BiADATU-Net achieves competitive performance supremacy compared to some state-of-the-art methods. It is potential and valuable in the field of skin lesion segmentation.

KEYWORDS
deformable attention Transformer, dual attention block, feature fusion, scSE attention block, skin lesion segmentation, U-Net

1 | INTRODUCTION

Skin diseases are among the widespread ailments characterized by a high incidence rate in the population. They are distributed in a wide range of ages. Many internal organ diseases are directly manifested on the skin. There are various types of skin diseases and melanoma is a malignant tumor that arises from normal melanocytes or pre-existing nevus cells in the lesional epidermis.[1] Malignant melanoma is characterized by rapid disease spread, poor healing effects, and an extremely high mortality rate. According to the statistics of the World Health Organization, there were an estimated 325 000 new cases of cutaneous melanoma worldwide in 2020, and 57 000 people died from the disease. The projection indicates a concerning surge in the occurrence of melanoma, with an anticipated escalation of more than 50% from the year 2020 to 2040.[2] However, if it is detected early, the 5-year survival rate can reach more than 90%.[3] The later the detection, the survival rate will be greatly lowered. Therefore, prompt diagnosis and early detection are essential for successful treatment of melanoma.

In early clinical practices, dermatologists usually identify melanoma by using the ABCDE rule based on the dermoscopic images.[4–6] The diagnostic results rely on visual observations manually. Nevertheless, the lesion at an early stage resembles the natural skin features. It is occasionally covered up by hair. Even experts might make mistakes in diagnosis. The advancement of computer-aided diagnostic system (CADS) provides convenient conditions for improving this situation. An automated CADS usually goes through four stages after image acquisition: image preprocessing, precise segmentation of the skin lump area, feature extraction, and lesion identification. Among them, the accurate lesion segmentation contributes to promoting the reliability of diagnosis in the follow-up process. Thus, achieving precise and automated lesion segmentation is a critical step in advancing the field of precision oncology.[7]

The process of segmenting skin lesions is fraught with numerous complexities, such as low contrast, artificial artifacts, and vague contour. These problems render the precise delineation exceedingly challenging. With the development of deep learning methods, a plethora of sophisticated artificial intelligence algorithms with high precision have been harnessed for the segmentation of skin tumors. Benefiting from convolutional neural network (CNN),[8] the U-shaped network (U-Net)[9] is distinguished by the incorporation of skip connections that bridge the two components. Nowadays, the U-Net and its derivatives have become prevalent in the field of skin lesion segmentation.[10–12] Despite obtaining some good results, there is still room for further optimization owing to the problems of coarse boundary location and diverse shapes in dermatological images. Hence, we focus on four public datasets of dermoscopic images that are dedicated to segmentation research. Drawing on the advantages of existing methods, a more competitive framework is proposed for the accurate skin lesion segmentation. The primary contributions are concluded as follows.

1. An efficacious medical image segmentation framework, namely BiADATU-Net, is designed in an end-to-end way in this paper. It integrates two different attention blocks into a deformable attention Transformer U-Net structure. We explore a promising design on the basis of the encoder-decoder architecture, expanding the flexibility of segmentation networks.
2. Deformable convolution is embedded into the dual attention and scSE attention block, designing two types of attention modules, called DAD-block and scSED-block, respectively. In the encoder part, DAD-block is placed at the front end of the deformable attention Transformer, while the scSED-block is positioned in the skip connection part before the decoder. Ablation experiments confirmed that this design enhanced feature extraction and fusion capabilities, facilitating promoting of segmentation performance.
3. We conducted comprehensive experiments across four skin cancer image datasets to evaluate the efficacy of BiADATU-Net, benchmarking it against a range of other advanced models. The experiment outcomes validate that our network has delivered superior performance, as evidenced by its accuracy, Dice coefficient, and Jaccard index scores.

The rest of this paper is outlined as follows: The “literature review” section provides a brief review of methods about skin lesion segmentation problems, discussing classical digital image processing techniques and CNN-based methods. Besides, a concise introduction of Transformer-based method and attention mechanism are presented for medical image segmentation. The “methods” section is dedicated to detailing our model. The “experiments and analysis” section elaborates our comparative experiment results analysis and ablation study. A concise discussion of experimental results is made in the “discussion” section. The “conclusion” part summarizes the paper.

2 | LITERATURE REVIEW

2.1 | Skin lesion segmentation

Traditional approaches to image segmentation predominantly utilize the establishment of various thresholds, capitalizing on features like grayscale levels, textural patterns, and color attributes to distinguish regions of interest.[13] By classifying each pixel into different regions, the entire image is segmented to separate different areas by categories. These techniques can be separated into three principal groups: thresholding approaches, region-based strategies, and edge-based methodologies. Glaister et al.[14] introduced a segmentation algorithm for skin lesions that is predicated on the concept of texture distinctiveness (TD). This algorithm utilizes TD as a core metric for discerning the sparse texture patterns present within the input images. The algorithm adeptly identifies and accentuates the disparities in texture distributions, thereby establishing an optimal threshold to effectively partition the image into normal skin and those indicative of lesions. Although this method is computationally simple, the selection of the optimal threshold can be a tedious task. Region-based methods partition an image into distinct regions by employing similarity criteria that define the coherence within each individual region. Abbas et al.[15] improved region-based active contour methods to segment multiple lesion tissues. However, when there are significant differences in texture and color between the skin and lesion tissues, region-based methods can lead to over-segmentation. Edge-based methods detect areas with abrupt changes in either grayscale or structure, segmenting the image by identifying pixels along the edges of the target region. Ali et al.[16] employed the Canny edge detector to locate skin lesion boundaries and evaluated the regularity. Nevertheless, this method is highly susceptible to the noise, leading to low segmentation accuracy. Oukil et al.[17] exploited the K-means algorithm to segment the regions of skin lesions. Utilizing the color and texture attributes of these lesions, they developed a superior feature extraction technique that facilitates the detection of melanoma.

Currently, deep learning techniques have been extensively utilized in the task of image segmentation. Akram et al.[18] combined CNN-based deep learning methods with edge detection techniques to achieve effective boundaries location of skin lesions. Then, they used the ResNet50 model to perform the recognition of the skin lesions. Taking the proposal of fully convolutional network (FCN) as an important node, the encoder-decoder structure was innovatively constructed to realize the dense prediction of image pixels. Multiple FCNs were leveraged to learn the subtle features of the lesion area.[19] The shortcoming of this approach is that the relationships between pixels are not fully considered. Besides, it is less sensitive to fine details in the image, resulting in under-segmentation results. The U-Net has garnered considerable interest among researchers due to its outstanding efficacy in the domain of medical image segmentation. The U-Net and its variants are still the mainstream research to date.[20] For instance, U-Net++,[10] residual U-Net,[21] recurrent residual U-Net(R2U-Net),[22] attention U-Net[11] are frequently applied in many semantic segmentation tasks. Oktay et al.[11] introduced attention mechanism into segmentation networks, enhancing the expressive power of CNN by adaptively learning feature weights. This enables important features to be assigned higher weights, facilitating faster learning of skin lesion characteristics. Alom et al.[22] investigated two novel structures based on residual U-Net. They have similar performance as the equivalent model in various segmentation tasks. Maurya et al.[23] harnessed the U-Net architecture to construct an automated telangiectasia detection model. The study established a pathway for the early detection of skin cancer by identifying key signs indicative of the disease. Jin et al.[24] devised a novel diffusion network to achieve classification and segmentation of skin lesions. The designed module incorporated the mechanism of multi-task learning to enhance segmentation performance. Nevertheless, the segmentation accuracy of U-Net is still challenging as a result of complex lesion shapes, blurred boundaries, and artifacts in skin cancer images.

2.2 | Transformer-based methods for medical image segmentation

Lack of capturing long-range dependencies imposes certain constraints on the segmentation accuracy for CNN. In recent years, Transformer[25] has revolutionized the field of computer vision, achieving remarkable success and setting new benchmarks in image analysis and processing. TransUNet[26] represents a pioneering effort to investigate the applicability of the Transformer architecture for medical image segmentation, breaking new ground in the integration of this technology within the field. Its overall architecture follows the design of U-Net, utilizing Transformer as an encoder to encode feature maps from the CNN into input sequences for extracting global context information. However, it overlooks the image-specific positional and channel information. Sun et al.[27] integrated dual attention mechanisms into the Transformer encoder, thereby significantly boosting the model's capacity for feature extraction. TEC-Net[28] integrated dynamic deformable convolution into the CNN and combined it with vision Transformer for skin lesion segmentation. Traditional Transformer encoder exhibits high computational and storage complexity when processing high-resolution feature maps. In response to these challenges, Zhu et al.[29] developed the deformable detection Transformer (DETR), a novel approach that effectively mitigates the issues of slow convergence and elevated complexity associated with the original Transformer framework. It combined the sparse spatial sampling of deformable convolution with the relationship modeling capabilities of Transformer, resulting in highly competitive detection performance.

2.3 | Attention mechanism

Attention mechanism is a computational model that mimics the way human attention is allocated, enabling models to focus on crucial parts when processing vast amounts of data. This mechanism enables models to dynamically adjust attention allocation based on the importance of input data, thereby enhancing the model's expressiveness and generalization capabilities.[25] As one of the key tools for enhancing performance of model, it has garnered widespread application across the realms of natural language processing and computer vision.[30] Squeeze-and-Excitation network (SENet)[31] leverages the channel attention mechanism to refine and invigorate input features. Convolutional block attention module (CBAM)[32] seamlessly amalgamates both channel and spatial attention mechanisms, culminating in a hybrid attention mechanism. Similarly, dual attention network (DANet)[33] adeptly fuses spatial and channel attention features to boost feature representation capability. Currently, attention mechanisms are widely utilized in medical-related semantic segmentation models. Azad et al.[34] incorporated channel attention into the Deeplabv3+ network to develop a segmentation model for skin diseases, resulting in an enhancement of segmentation accuracy. Li et al.[35] introduced a lightweight attention-based U-Net tailored for the segmentation of retinal vessel images, demonstrating performance metrics that surpass those of prevailing mainstream methodologies.

Inspired by the aforementioned studies, we propose BiADATU-Net for segmenting skin cancer images. It combines the deformable attention Transformer with the U-Net structure and introduces hybrid attention modules that include both dual attention and scSE attention mechanisms. This design effectively exploits the complementary advantages of the deformable attention Transformer and the U-Net,

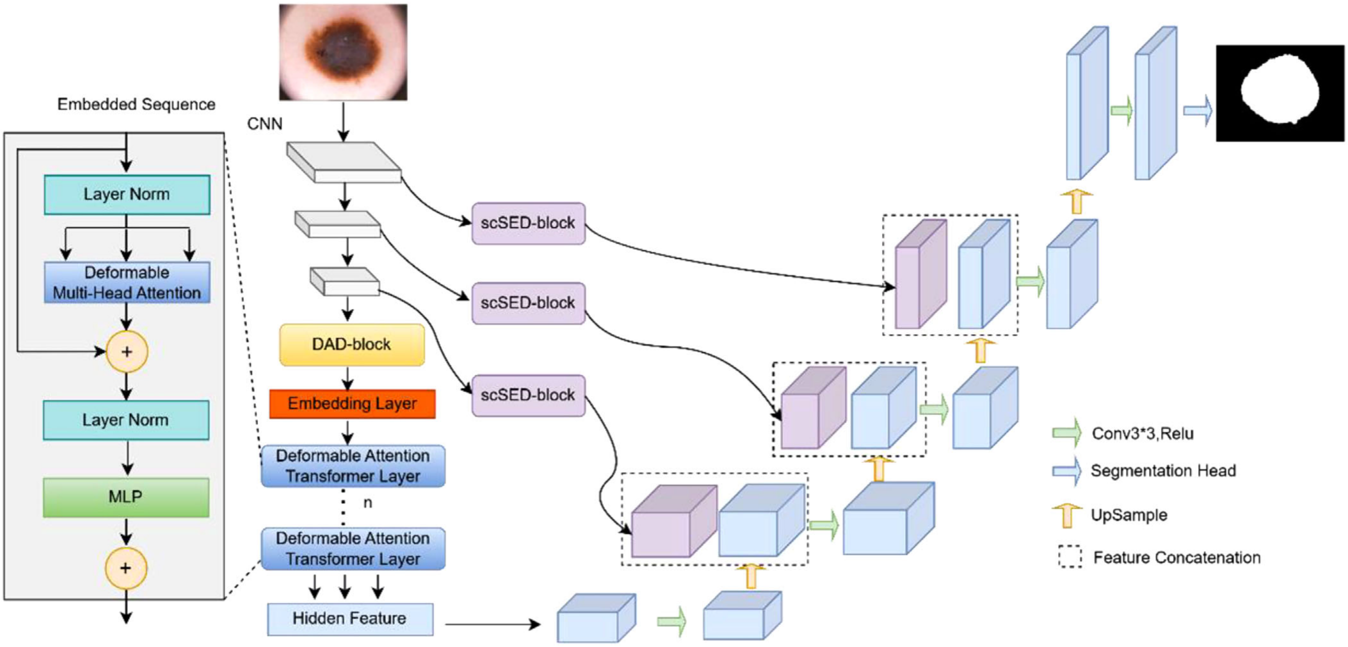

FIGURE 1 An overall structure diagram of BiADATU-Net.

enhancing the segmentation process. Moreover, the hybrid attention modules contribute to capturing densely global context information and image-specific position and channel features. The subsequent section delves into the architecture of BiADATU-Net.

3 | METHODS

In this section, we commence with a general introduction to the segmentation framework designed for skin lesions. Next, we describe the key components within the segmentation architecture, including the deformable attention Transformer, DAD-block, scSED-block, and the decoder.

3.1 | The overall segmentation framework

The BiADATU-Net is composed of an encoder with deformable attention Transformer and DAD-block, skip connection layers with scSED-block and decoder. The diagram is illustrated in Figure 1. The encoder part, as a key component for optimization, does not use the traditional transformer layer. Instead, it employs a deformable attention Transformer (DAT)[36] mechanism preceded by a specially designed DAD-block that integrates deformable convolution. For the skip connection layers, we also utilize the specifically tailored scSED-block placed in the middle of the entire network. The DAD-block and scSED-block constitute the bidirectional mixed attention modules, which help the model retain more valuable features and enhance the efficiency of feature extraction. As for the decoder part, we adopt the conventional convolutional module design.

To fully elucidate the rationality of the model we constructed, it is essential to first outline the constraints inherent in the traditional U-Net structure and Transformer with respect to feature extraction. Although traditional Transformers can capture long-distance feature dependencies through cascading self-attention modules, they tend to overlook the details of local features. Moreover, extracting features from high-resolution images requires substantial computation and memory resources. U-Net structures are adept at extracting local features for dense prediction, but still encounter certain limitations in capturing global features. To address these constraints, we have employed a more suitable DAT layer tailored for dense prediction tasks. This integration is complemented by the incorporation of a DAD-block preceding it, along with the integration of scSED-blocks within the skip connections. The benefits of these enhancements are evident: this approach merges the strengths of Transformer and U-Net in feature extraction while mitigating excessive computational overhead, yielding a more meaningful and robust feature representation. Meanwhile, the introduction of the DAD-block and scSED-block refines the features transmitted to the decoder, facilitating the learning of intricate image deformations and thereby enabling a more precise reconstruction of feature maps.

3.2 | Encoder with deformable attention Transformer and DAD-block

As illustrated in Figure 1, the encoder section is composed of four pivotal components: CNN blocks, DAD-block, embedding layer, and DAT layers. Following each stage of the CNN blocks, the feature map dimension is halved, while the channel count is doubled. The outputs from each level are preserved as inputs to the skip connections of U-Net. Then, the final output from the CNN blocks undergoes feature extraction and fusion at both position and channel levels via the DAD-block. This output is processed through an embedding layer encoding the

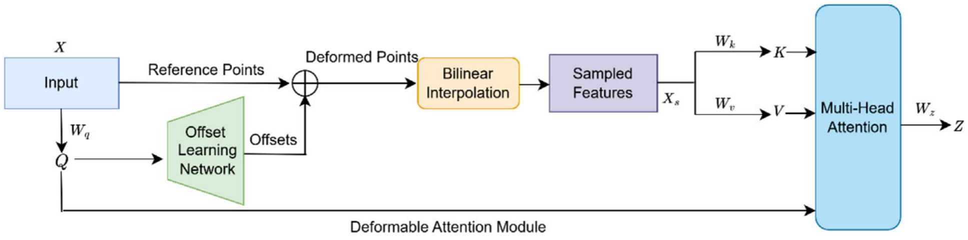

FIGURE 2 The structure diagram of deformable attention module.

feature map into a format suitable for the DAT layer. Through multiple layers of the DAT, features are refined and optimized. Ultimately, the fine-grained features are relayed to the decoder layer for feature reconstruction.

3.2.1 | Deformable attention Transformer

The traditional Transformer's self-attention mechanism is known for its substantial memory and computational demands. To cope with this problem, researchers have embarked on designing sparse self-attention mechanisms. Deformable attention Transformer is one of these innovations, achieving remarkable results in image classification.[36] Borrowed from this thought, the structure of DAT is depicted on the left side of Figure 1. This configuration primarily consists of the deformable attention mechanism, a feedforward neural network (FNN), and residual connections. A schematic of the deformable attention module is depicted in Figure 2.

The key and value vectors of the deformable attention mechanism are obtained by projecting sampled features from the original images. These sampled features are acquired through bilinear interpolation at sampling locations, which are dynamically determined by an offset learning network from the query vectors. Given an input feature map $X \in R^{H\times W\times C}$, the query vectors Q are generated initially by projecting the input. Simultaneously, according to the dimensions of the input feature map, a regular grid of reference points $P \in R^{\frac{H}{r}\times\frac{W}{r}\times 2}$ is produced as references, where r is a predefined parameter, called scaling factor. The reference points are linearly two-dimensional coordinates $\{(0,0),\dots,(\frac{H}{r}-1,\frac{W}{r}-1)\}$. To simplify computation, the coordinate values of these reference points are uniformly normalized to the interval $[-1,+1]$. The query vectors Q are then input into the $Offset(\cdot)$ network to learn the offset ΔP for each reference point. Then, the deformed points are derived by adding the reference points and the offset. Based on the coordinates of the deformed points, bilinear interpolation sampling is utilized to sample the input feature map, thereby obtaining the sampled features X_s. Subsequently, these sampled features are projected to generate deformed keys and values, namely K and V, respectively. The calculation formulas are listed below:

$$Q = XW_q,\quad K = X_s W_k,\quad V = X_s W_v \tag{1}$$

$$\Delta P = Offset\,(Q)\,,\ X_s = BI\,(X, P + \Delta P) \tag{2}$$

where W_q, W_k, $W_v \in R^{C\times C}$ are the projection matrices. $BI(\cdot)$ denotes bilinear interpolation sampling. Finally, the obtained Q, K and V are used to compute the output results through a multi-head self-attention (MSA) mechanism just as the traditional methods. The related formulas are:

$$Z^{(h)} = Softmax\left(\frac{Q^{(h)}{K^{(h)}}^T}{\sqrt{d}}\right)V^{(h)},$$

$$DMHA\ (Z) = Concat\left(Z^{(1)}, Z^{(2)}, \dots, Z^{(m)}\right)W_z \tag{3}$$

where $Z^{(h)}$ represents the attention output for the hth head, with m being the total number of heads, and $W_z \in R^{C\times C}$ is the output projection matrix for the deformable MSA. Ultimately, the output undergoes a Multi-Layer Perceptron (MLP) block and residual connections, forming the DAT. The computational formula of DAT can be represented as follows:

$$Z'_l = DMHA\,(LN\,(Z_l)) + Z_l \tag{4}$$

$$Z_{l+1} = MLP\left(LN\left(Z'_l\right)\right) + Z'_l \tag{5}$$

where $DMHA(\cdot)$ represents deformable MSA mechanism and $LN(\cdot)$ is layer normalization.

3.2.2 | Dual attention with deformable convolution block

As depicted in Figure 1, the dual attention with deformable convolution block is positioned preceding the DAT layers within the encoder, serving as a critical component for feature extraction. It aids in capturing image-based spatial and channel information. The rationale behind this setup is that while Transformer layers excel at extracting global features of an image, they are slightly less effective when it comes to capturing image-specific attributes. Within the DAD-block, the position attention module (PAM) and channel attention module (CAM)[33] effectively execute feature extraction pertaining to spatial and channel information of images. The structure diagram of DAD-block is illustrated in Figure 3. Considering the fixed size of traditional convolution kernels, their receptive field is inherently limited, rendering them less effective in perceiving the geometric shapes of target regions within an image. The incorporation of a deformable convolution layer renders the module's feature learning capabilities more flexible and adaptive, which is instrumental in augmenting the network's capacity

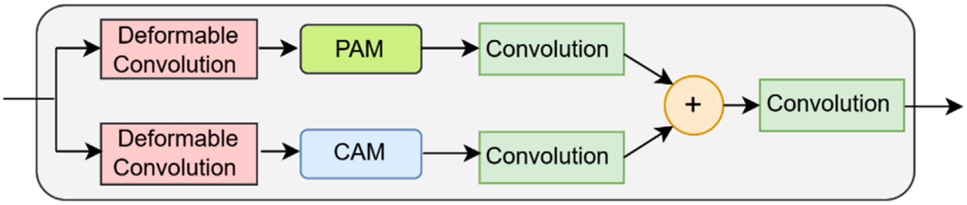

FIGURE 3 The structure diagram of DAD-block.

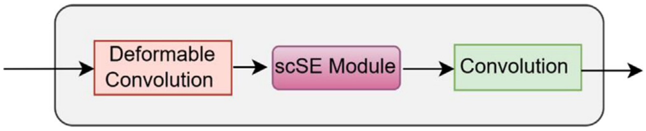

FIGURE 4 The structure diagram of scSED-block.

for learning geometric transformations. The outputs from the PAM and CAM, after being processed through a standard convolutional layer, are aggregated and subsequently subjected to another convolutional layer to achieve a deep fusion of features. This arrangement augments the model's capacity to discern and integrate pivotal image attributes, subsequently elevating the accuracy in handling complex data.

3.3 | Skip connection layers with scSED-block

The scSED-block is displayed in Figure 4. We consider incorporating a specially crafted scSED-block into all skip connection layers, replacing straightforward feature passing mechanism. The primary motivation behind this optimization stems from the tendency of traditional direct feature passing to propagate redundant features. Through scSED-block, features are selected and combined at both spatial and channel levels. This process aids in extracting more valuable features and improving the accuracy of feature map reconstruction.

The scSED-block encodes spatial and channel information separately, capturing pixel-level spatial information. It needs to go through a deformable convolution layer firstly, then obtains the fused features via scSE attention module.[37] By introducing cross-channel and cross-spatial information interaction, it enhances the network's perceptual capabilities. The scSED-block is designed for use in the skip connection layers, facilitating the strengthening of meaningful features and suppression of irrelevant ones. In this way, our model implements a bidirectional hybrid attention mechanism in the encoder part and the skip connection layers. As proven by experiments on four dedicated datasets of skin cancer images, it helps promote the model's overall performance and generalization ability.

3.4 | Decoder

The right half of Figure 1 illustrates the decoder component of the model, which receives the output from the deformable Transformer layer in the encoder section. This procedure entails a series of upsampling operations executed in a stepwise manner, each doubling the dimensions of the feature map. The expanded feature map is subsequently merged with the corresponding output from the skip connection layer, and the amalgamation is processed further through a convolutional layer for enhanced integration. These steps are iterated three times to reconstruct the feature map. Finally, the segmentation head layer produces the segmentation results for the target area in the image. By implementing these strategies in our segmentation model, we ensure that the decoder achieves high-quality feature reconstruction, yielding a high-precision segmentation framework focused on skin lesions.

4 | EXPERIMENTS AND ANALYSIS

4.1 | Dataset description and preprocessing

Four well-known skin cancer image datasets are utilized as benchmarks to construct an automated segmentation framework in the study. These datasets include ISIC2016,[38] ISIC2017,[39] ISIC2018,[40] and Pedro Hispano 2(PH2)[41] dataset. The first three datasets are sourced from the International Skin Imaging Collaboration (ISIC), which hosts international workshops and challenges based on biomedical imaging. They contain pigment lesion areas from different populations and have been annotated by experts for lesion contour and type. The ISIC2016 dataset consists of 900 training samples and 379 testing samples. The ISIC2017 dataset has 2000 training images, 150 validation images, and 600 testing images, while the ISIC2018 dataset contains 2594 images in total. According to the partition scheme suggested in Qin et al.,[42] it is divided into 1816 training images, 260 validation images, and 518 testing images.

The PH2 dataset is a dedicated dermoscopy image database. It is comprised of 200 skin lesion images with manually annotated segmentation by dermatologists. This dataset is specifically designed for clinical research and benchmark testing. We use the PH2 dataset as an additional test set for evaluating the segmentation model's generalization ability on the ISIC2018 dataset in conformity with the approach described in He et al.[7]

One of the challenges in skin lesion segmentation is the presence of hair and artifacts in the lesion area. A common approach to addressing this challenge is to employ classical digital image processing techniques to remove obstacles from the skin lesion images before segmentation. For example, Kasmi et al. developed the SharpRazor method to detect hair and artifacts noise and remove them from dermoscopic

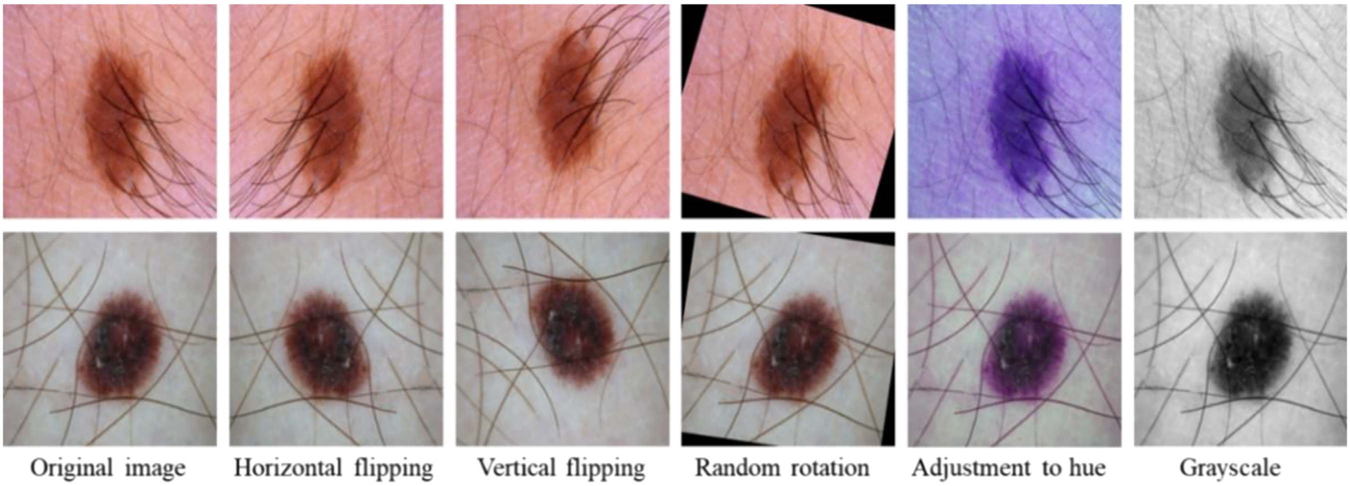

FIGURE 5 Image preprocessing.

images.[43] As for an automated and intelligent segmentation framework, it is essential to embody an end-to-end philosophy, minimizing multi-step processing to enhance segmentation efficiency. Therefore, the proposed BiADATU-Net circumvents the need for additional hair and artifacts noise processing methods, allowing direct input of skin images into the segmentation network. This operation demonstrates that our model can learn effective features from the lesion area, suppress interference from noise signals, and exhibit good robustness and anti-interference capability. Considering the limited sample size in existing skin image datasets, there is an increased risk of model overfitting during the training phase. To bolster the precision of lesion segmentation, this research employs data augmentation techniques to expand training samples. These operations include random horizontal flipping, vertical flipping, random rotation, grayscale conversion, and adjustments to image hue. Figure 5 demonstrates the processed data samples. Additionally, image resize and normalization are also performed. Image pixel resolution in the original database varies from 540 × 722 to 4499 × 6748.[44] Such large dimensions impose high hardware requirements on the training equipment. Therefore, we uniformly rescale the image size to 224 × 224 and perform normalization, mapping the pixel values to the range of 0 to 1.

4.2 | Implementation details

The experiments conducted in this paper are performed on a computer equipped with an Intel 16 vCPU Intel(R) Xeon(R) Platinum 8350C CPU @ 2.60 GHz and an NVIDIA GeForce RTX A5000 (24GB) GPU. The deep learning framework used is PyTorch, with Python version 3.8 and CUDA version 11.1. During the training process, we randomly initialize the network weights and update the parameters using the SGD optimizer. The sum of BCEWithLogitsLoss and Dice metric is used as loss function to guide the training. The learning rate is set to 0.0001 and a weight decay of 0.0001. The number of training iterations and batch size are set to 50 and 8, respectively. To improve computational efficiency, we fix the input image size to 224 × 224 × 3 uniformly. Taking the training on the ISIC2018 dataset as an example, the training and validation loss curves are depicted in Figure 6.

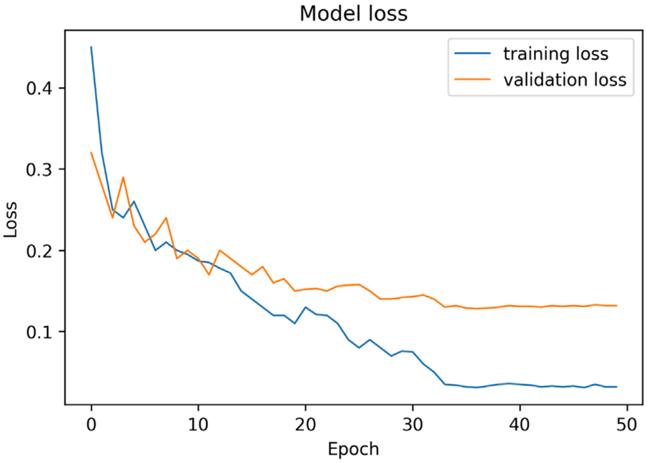

FIGURE 6 Loss curves on the ISIC2018 dataset.

4.3 | Evaluation metrics

We choose five commonly used indicators for image segmentation assessment, that is, Pixel Accuracy (Acc), Sensitivity (Sen), Precision (Pre), Dice Coefficient (DC), and Jaccard Index (JA). Among these metrics, DC and JA can be used as comprehensive measures to evaluate the similarity between the ground truth and the predicted region. The other three metrics are typically regarded as statistical measures of binary classification tasks. They are used as references in our task. Their calculation formulas are presented below:

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{6}$$

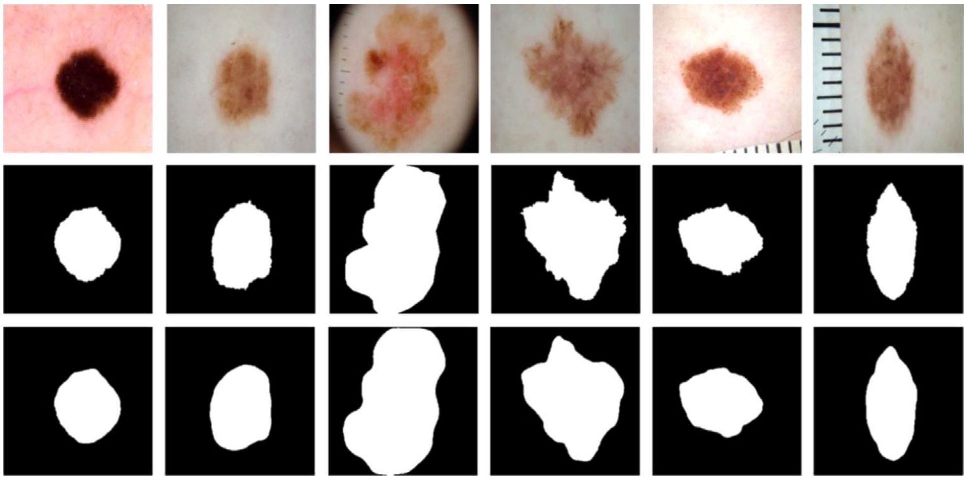

FIGURE 7 Visualization of segmentation results. From top to bottom: original images, ground truth, and predicted segmentation.

$$Sen = \frac{TP}{TP + FN} \tag{7}$$

$$Pre = \frac{TP}{TP + FP} \tag{8}$$

$$DC = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{9}$$

$$JA = \frac{TP}{TP + FN + FP} \tag{10}$$

where TP signifies the quantity of lesion pixels that have been accurately classified, while FP represents the count of non-lesion pixels erroneously categorized as lesions. Conversely, TN corresponds to the number of non-lesion pixels that have been correctly identified, and FN refers to the lesion pixels that have been misclassified as non-lesion. These metrics are quantified on a scale from 0 to 1, with values approaching 1 indicating superior model performance.

4.4 | Analysis of results

Figure 7 presents the comparison diagram between the true labels and predicted results of our model in the aforementioned datasets. Figure 7 demonstrates that our model has delivered satisfactory segmentation outcomes for some challenging samples, including those with low contrast or noise. Additionally, the model exhibits relatively accurate learning of irregular boundaries.

With the aim of confirming the segmentation performance difference between the proposed model and seven advanced models, for instance, U-Net,[9] R2U-Net,[22] Attention U-Net,[11] TransUNet,[26] CPF-Net,[45] FAT-Net[46] and DA-TransUNet,[27] we implemented experiments on the ISIC2016, ISIC2017, ISIC2018, and PH2 datasets. The segmentation metrics are exhibited from Table 1 to Table 4. It should be noted that we did not train the PH2 dataset. Instead, we used it as a supplementary test set to assess the generalization ability of the segmentation model established on the ISIC2018 dataset.

As shown in Table 1 to Table 4, it can be observed that the proposed model consistently achieves relatively better metrics. Table 1 delineates the prediction results of the ISIC2016 dataset. Compared with the classical U-Net model, our proposed model shows performance improvements on metrics such as Acc, Sen, Pre, DC, and JA, with increases of 2.98%, 4.35%, 2.76%, 2.11%, and 2.87%, respectively. When benchmarked against some state-of-the-art models like DA-TransUNet, FAT-Net, and CPF-Net, our model also exhibits a modest degree of performance enhancement. According to Table 2, our model achieves Acc, DC, and JA values of 0.9618, 0.8990, 0.8270, respectively, outperforming DA-TransUNet's scores of 0.9553, 0.8869, 0.8021 on the ISIC2017 dataset. The same conclusion can also be drawn regard to the experiments of the ISIC2018 dataset in Table 3. The results in Table 4 are more convincing as they are not trained on the PH2 dataset. The metrics are calculated directly using the training model on the ISIC2018 dataset. According to Table 4, the evaluation results reveal even higher scores of 0.9687, 0.9496, 0.9475, 0.9463, and 0.8998 for the five metrics, surpassing some advanced models such as CPF-Net, FAT-Net, and DA-TransUNet. All the above results indicate that the segmentation model developed in this study demonstrates superior generalization capabilities.

Figure 8 and Figure 9 present the visual comparison results of aforementioned models. These visualizations provide a comparative and intuitive reflection of the differences in segmentation performance among various models. Figure 8 depicts the comparison results on different models. By comparing the predicted results of each model, it is evident that our model exhibits a greater degree of segmentation accuracy in contrast with the other models. The extent of failure or missed segmentation is minimal, and the results closely approximate the manually annotated ground truth. Figure 9 illustrates the segmentation result details of different networks, with enhanced contrast in yellow and purple to distinguish between skin lesions and normal regions.

4.5 | Ablation experiment

To substantiate the efficacy of specifically designed modules incorporated in our proposed model, an ablation experiment was conducted

TABLE 1 Performance comparison of various networks on ISIC2016 dataset.

Citation	Method	Acc	Sen	Pre	DC	JA
Ronneberger et al. (2015)[9]	U-Net	0.9318	0.8771	0.9084	0.9003	0.8307
Alom et al. (2018)[22]	R2U-Net	0.9325	0.8798	0.9092	0.9012	0.8350
Oktay et al. (2018)[11]	Attention U-Net	0.9383	0.8829	0.9112	0.9027	0.8352
Feng et al. (2020)[45]	CPF-Net	0.9502	0.9078	0.9143	0.9134	0.8424
Chen et al. (2021)[26]	TransUNet	0.9455	0.9069	0.9070	0.9048	0.8373
Wu et al. (2022)[46]	FAT-Net	0.9526	0.9105	0.9135	0.9149	0.8449
Sun et al. (2023)[27]	DA-TransUNet	0.9538	0.9112	0.9152	0.9172	0.8491
Proposed	**BiADATU-Net**	**0.9616**	**0.9206**	**0.9360**	**0.9214**	**0.8594**

Abbreviations: Acc, accuracy; DC, dice coefficient; JA, Jaccard index; Pre, precision; Sen, sensitivity.

TABLE 2 Performance comparison of various networks on ISIC2017 dataset.

Citation	Method	Acc	Sen	Pre	DC	JA
Ronneberger et al. (2015)[9]	U-Net	0.9320	0.8434	0.9101	0.8499	0.7641
Alom et al. (2018)[22]	R2U-Net	0.9358	0.8542	0.9089	0.8506	0.7762
Oktay et al. (2018)[11]	Attention U-Net	0.9390	0.8502	0.9153	0.8602	0.7775
Feng et al. (2020)[45]	CPF-Net	0.9498	0.8623	0.9185	0.8799	0.7912
Chen et al. (2021)[26]	TransUNet	0.9452	0.8581	0.9172	0.8663	0.7845
Wu et al. (2022)[46]	FAT-Net	0.9512	0.8605	0.9207	0.8805	0.7918
Sun et al. (2023)[27]	DA-TransUNet	0.9553	0.8612	0.9265	0.8869	0.8021
Proposed	**BiADATU-Net**	**0.9618**	**0.8727**	**0.9479**	**0.8990**	**0.8270**

Abbreviations: Acc, accuracy; DC, dice coefficient; JA, Jaccard index; Pre, precision; Sen, sensitivity.

TABLE 3 Performance comparison of various networks on ISIC2018 dataset.

Citation	Method	Acc	Sen	Pre	DC	JA
Ronneberger et al. (2015)[9]	U-Net	0.9312	0.8521	0.9012	0.8579	0.7757
Alom et al. (2018)[22]	R2U-Net	0.9427	0.8635	0.9034	0.8637	0.7850
Oktay et al. (2018)[11]	Attention U-Net	0.9449	0.9036	0.8618	0.8673	0.7854
Feng et al. (2020)[45]	CPF-Net	0.9542	0.8757	0.9044	0.8886	0.8045
Chen et al. (2021)[26]	TransUNet	0.9532	0.8725	0.9015	0.8846	0.8014
Wu et al. (2022)[46]	FAT-Net	0.9556	0.8788	0.9052	0.8918	0.8092
Sun et al. (2023)[27]	DA-TransUNet	0.9561	0.8814	0.9048	0.8925	0.8186
Proposed	**BiADATU-Net**	**0.9626**	**0.9137**	**0.9168**	**0.9008**	**0.8343**

Abbreviations: Acc, accuracy; DC, dice coefficient; JA, Jaccard index; Pre, precision; Sen, sensitivity.

TABLE 4 Performance comparison of various networks on PH2 dataset.

Citation	Method	Acc	Sen	Pre	DC	JA
Ronneberger et al. (2015)[9]	U-Net	0.9237	0.9209	0.9014	0.8862	0.8300
Alom et al. (2018)[22]	R2U-Net	0.9321	0.9314	0.9071	0.8960	0.8368
Oktay et al. (2018)[11]	Attention U-Net	0.9325	0.9383	0.9124	0.9023	0.8416
Feng et al. (2020)[45]	CPF-Net	0.9412	0.9251	0.9285	0.9315	0.8729
Chen et al. (2021)[26]	TransUNet	0.9376	0.9344	0.9137	0.9205	0.8562
Wu et al. (2022)[46]	FAT-Net	0.9426	0.9409	0.9181	0.9342	0.8803
Sun et al. (2023)[27]	DA-TransUNet	0.9458	0.9432	0.9355	0.9358	0.8871
Proposed	**BiADATU-Net**	**0.9687**	**0.9496**	**0.9475**	**0.9463**	**0.8998**

Abbreviations: Acc, accuracy; DC, dice coefficient; JA, Jaccard index; Pre, precision; Sen, sensitivity.

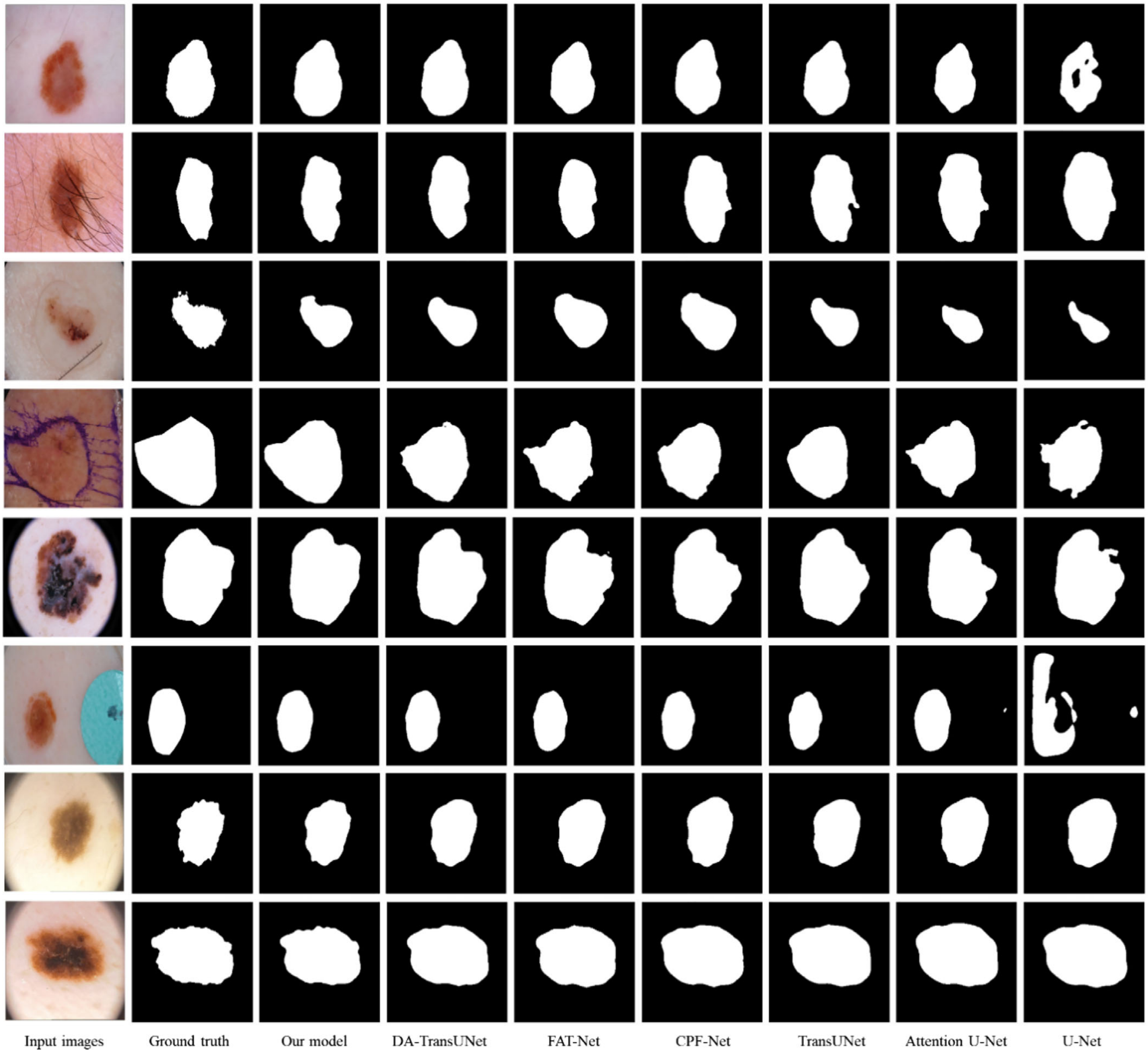

FIGURE 8 Visualization results of different models on the ISIC 2016, ISIC 2017, ISIC 2018, and PH2 test set.

using the ISIC2018 dataset as a case study. By sequentially adding DAT, DAD-block and scSED-block, we compare the DC and JA values on the ISIC2018 test set to determine if the inclusion of these modules has a positive impact. Table 5 presents the results of our experiments. As observed, the initial row corresponds to the baseline performance of the original model, which is evaluated without the incorporation of any additional modules. After individually adding the aforementioned modules, positive improvements are achieved in all cases, and the combination of the three modules engenders greater improvements. Compared to the model without adding any modules, the Dice coefficient and JA increase by 4.27%, 5.81%, respectively. Consequently, the incorporation of these modules has significantly contributed to the improved performance of the segmentation model.

5 | DISCUSSION

According to Figure 9, when images contain small regions of interest, models like U-Net, Attention U-Net, CPF-Net, and FAT-Net fail to effectively differentiate the true lesion area from other healthy skin. They tend to mis-segment skin lesion areas, resulting in multiple lesion blocks. However, our model successfully identifies these minor pseudo-lesion areas, reducing segmentation errors. When the skin lesion is surrounded by a significant amount of hair, we can observe from the area within the red box that the other six models display errors in segmentation due to interference from the hair. It indicates that their algorithm inadequately filters out hair noise from normal area, yielding inaccurate results. Nevertheless, our model is less affected compared

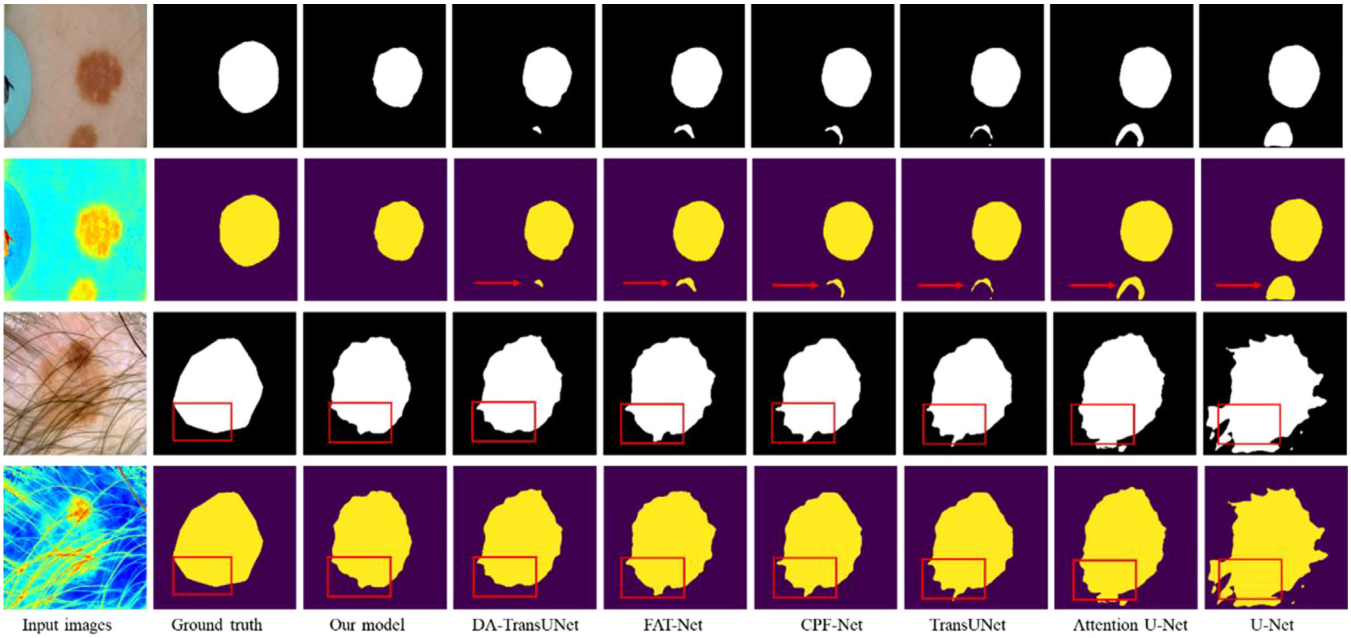

FIGURE 9 Visualization details of different models.

TABLE 5 Ablation experiment with different module combinations.

Number	DAT	DAD-block	scSED-block	DC	JA
1				0.8581	0.7762
2	✓			0.8745	0.7978
3		✓		0.8706	0.7934
4			✓	0.8752	0.7986
5	✓	✓		0.8901	0.8068
6	✓	✓	✓	**0.9008**	**0.8343**

Abbreviation: DAT, deformable attention Transformer; DC, dice coefficient; JA, Jaccard index.

to the other models overall, which further validates its effectiveness and superior robustness.

Based on the analysis above, our intelligent segmentation model demonstrates competitive performance in terms of skin lesion images. Integrating the strengths of deformable Transformer and U-Net, our model incorporates bidirectional attention modules that aid in complex feature extraction at both positional and channel attribute levels. This design facilitates feature reconstruction of segmentation tasks. Through experiments conducted on four dermatological image datasets, our model not only achieves enhanced segmentation accuracy but also manifests outstanding generalization capabilities.

Despite having certain advantages, our model currently exhibits some constraints. Firstly, while the incorporation of the DAD-block and scSED-block contributes to improved accuracy, it has also increased computational complexity, posing a constraint for applications requiring high real-time performance. Secondly, the decoder section still employs the traditional U-Net structure, leaving room for optimization. Particularly, as we consider the potential application of this model to a broader spectrum of medical segmentation tasks in the future, optimization and enhancement of the decoder part are necessary to analyze more complex data objects.

6 | CONCLUSION

This paper proposed a hybrid segmentation architecture, that is, BiADATU-Net, for skin cancer datasets. The model utilized deformable attention Transformer and bidirectional attention modules, that is, DAD-block and scSED-block, to enhance the learning capability of skin lesion features. Comprehensive experiments on the ISIC2016, ISIC2017, ISIC2018, and PH2 datasets indicated that our proposed model achieved commendable segmentation results in contrast with some sophisticated methods. The results on the PH2 dataset also confirmed the strong robustness and generalization capacity of our model. Besides, our model is adaptable for application to a variety of other medical segmentation tasks. In addition to high-precision skin lesion segmentation, future research will take into account the diagnosis and recognition of skin disease types to better assist in clinical treatment.

ACKNOWLEDGMENTS

The authors are highly grateful for the financial support provided by the College Youth Innovation Talent Project of Guangdong Province (Natural Science) (No. 2021KQNCX136), the Teaching and Research Project of Guangzhou Xinhua University (No. 2023J037), the Education Science Planning Project of Guangdong Province (Higher Education Special Project)(No. 2023GXJK623), the College Youth Innovation Talent Project of Guangdong Province (No.2022WQNCX099), the Course Ideological and Political Demonstration Project of Guangzhou

Xinhua University (No. 2023KCSZ009), the Teaching Quality and Teaching Reform Project of Guangdong Province, with Letter from Higher Education Office of Guangdong Provincial Department of Education [2023] No. 4, the Teaching Quality and Reform Engineering Project of Guangzhou Xinhua University (No. 2021TZ001), and the First-Class Undergraduate Major Construction Project of Guangzhou Xinhua University (No. 2024YLZY018).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The datasets will be available upon request to the corresponding author.

ORCID

Keke Hou https://orcid.org/0000-0001-8497-837X

REFERENCES

1. Sayan A, Plant R, Eccles B, Davies C, Ilankovan V. Recent advances in the management of cutaneous malignant melanoma: our case cohort. *Br J Oral Maxillofac Surg*. 2021;59(5):534-545.
2. Arnold M, Singh D, Laversanne M, et al. Global burden of cutaneous melanoma in 2020 and projections to 2040. *JAMA Dermatol*. 2022;158(5):495-503.
3. Ge Z, Demyanov S, Chakravorty R, Bowling A, Garnavi R. Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2017: 20th International Conference*. Springer; 2017:250-258.
4. Nachbar F, Stolz W, Merkle T, et al. The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *J Am Acad Dermatol*. 1994;30(4):551-559.
5. Binder M, Schwarz M, Winkler A, et al. Epiluminescence microscopy: a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists. *Arch Dermatol*. 1995;131(3):286-291.
6. Rigel DS, Friedman RJ, Kopf AW, Polsky D. ABCDE—an evolving concept in the early detection of melanoma. *Arch Dermatol*. 2005;141(8):1032-1034.
7. He Z, Li X, Chen Y, Lv N, Cai Y. Attention-based dual-path feature fusion network for automatic skin lesion segmentation. *BioData Min*. 2023;16(1):28-49.
8. Sahiner B, Chan HP, Petrick N, et al. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *IEEE Trans Med Imaging*. 1996;15(5):598-610.
9. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *2015 International Conference on Medical image computing and computer-assisted intervention*. Springer; 2015:234-241.
10. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: a nested U-Net architecture for medical image segmentation. *Deep Learn Med Image Anal Multimodal Learn Clin Decis Support*. 2018;11045:3-11.
11. Oktay O, Schlemper J, Folgoc LL, et al. Attention U-Net: learning where to look for the pancreas. In: *1st Conference on Medical Imaging with Deep Learning (MIDL 2018)*; 2018; Amsterdam, The Netherlands. arXiv preprint arXiv:1804.03999.
12. Xie J, Zhu R, Wu Z, Ouyang J. FFUNet: a novel feature fusion makes strong decoder for medical image segmentation. *IET Signal Process*. 2022;16(5):501-514.
13. Mete M, Sirakov NM. Lesion detection in demoscopy images with novel density-based and active contour approaches. *BMC Bioinformatics*. 2010;11(suppl 6):S23.
14. Glaister J, Wong A, Clausi DA. Segmentation of skin lesions from digital images using joint statistical texture distinctiveness. *IEEE Trans Biomed Eng*. 2014;61(4):1220-1230.
15. Abbas Q, Fondón I, Rashid M. Unsupervised skin lesions border detection via two-dimensional image analysis. *Comput Methods Programs Biomed*. 2011;104(3):e1-e15.
16. Ali AR, Li J, Yang G, O'Shea SJ. A machine learning approach to automatic detection of irregularity in skin lesion border using dermoscopic images. *PeerJ Comput Sci*. 2020;6:e268.
17. Oukil S, Kasmi R, Mokrani K, García-Zapirain B. Automatic segmentation and melanoma detection based on color and texture features in dermoscopic images. *Skin Res Technol*. 2022;28(2):203-211.
18. Akram A, Rashid J, Jaffar MA, Faheem M, Amin RU. Segmentation and classification of skin lesions using hybrid deep learning method in the Internet of Medical Things. *Skin Res Technol*. 2023;29(11):e13524.
19. Bi L, Kim J, Ahn E, Kumar A, Fulham M, Feng D. Dermoscopic image segmentation via multistage fully convolutional networks. *IEEE Trans Biomed Eng*. 2017;64(9):2065-2074.
20. Pennisi A, Bloisi DD, Suriani V, Nardi D, Facchiano A, Giampetruzzi AR. Skin lesion area segmentation using attention squeeze U-Net for embedded devices. *J Digit Imaging*. 2022;35(5):1217-1230.
21. Li D, Dharmawan DA, Ng BP, Rahardja S. Residual u-net for retinal vessel segmentation. In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE; 2019:1425-1429.
22. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK. Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation. Paper presented at University of Dayton. IEEE; 2018; Dayton, OH. arXiv preprint arXiv:1802.06955.
23. Maurya A, Stanley RJ, Lama N, et al. A deep learning approach to detect blood vessels in basal cell carcinoma. *Skin Res Technol*. 2022;28(4):571-576.
24. Jin Q, Cui H, Sun C, Meng Z, Su R. Cascade knowledge diffusion network for skin lesion diagnosis and segmentation. *Appl Soft Comput*. 2021; 99:106881-106893.
25. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *31st Conference on Neural Information Processing Systems (NIPS 2017)*. 2017; Long Beach, CA. arXiv preprint arXiv:1706.03762.
26. Chen J, Lu Y, Yu Q, et al. Transunet: transformers make strong encoders for medical image segmentation. 2021. arXiv preprint arXiv:2102.04306.
27. Sun G, Pan Y, Kong W, et al. Integrating spatial and channel dual attention with Transformer U-Net for medical image segmentation. 2023. arXiv preprint arXiv:2310.12570.
28. Sun R, Lei T, Zhang W, Wan Y, Xia Y, Nandi AK. TEC-Net: vision transformer embrace convolutional neural networks for medical image segmentation. 2023. arXiv preprint arXiv:2306.04086.
29. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: deformable transformers for end-to-end object detection. 2020. arXiv preprint arXiv:2010.04159.
30. Li R, Zheng S, Duan C, Su J, Zhang C. Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. *IEEE Geosci Remote Sens Lett*. 2021;19:1-5.
31. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE; 2018:7132-7141.
32. Woo S, Park J, Lee JY. CBAM: convolutional block attention module. In:*15th European Conference on Computer Vision (ECCV)*. Springer; 2018:3-19.
33. Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE; 2019:3146-3154.

34. Azad R, Asadi-Aghbolaghi M, Fathy M, Escalera S. Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation. *European conference on computer vision*. ECCV; 2020:251-266.
35. Li X, Jiang Y, Li M, Yin S. Lightweight attention convolutional neural network for retinal vessel image segmentation. *IEEE Trans Industr Inform*. 2020;17(3):1958-1967.
36. Xia Z, Pan X, Song S, Li LE, Huang G. Vision transformer with deformable attention. CVPR; 2022:4794-4803. arXiv preprint arXiv:2201.00520.
37. Roy AG, Navab N, Wachinger C. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. MICCAI; 2018:421-429. arXiv preprint arXiv:1803.02579.
38. Gutman D, Codella NC, Celebi ME, et al. Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). 2016. arXiv preprint arXiv:1605.01397.
39. Codella NC, Gutman D, Celebi ME, 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). In: *IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE; 2018. arXiv preprint arXiv:1710.05006.
40. Codella NC, Rotemberg V, Tschandl P, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). 2019. arXiv preprint arXiv:1902.03368.
41. Mendonça T, Ferreira PM, Marques JS, Marcal AR, Rozeira J. PH^2-A dermoscopic image database for research and benchmarking. *Annu Int Conf IEEE Eng Med Biol Soc*. 2013:5437-5440.
42. Qin C, Zheng B, Zeng J, et al. Dynamically aggregating MLPs and CNNs for skin lesion segmentation with geometry regularization. *Comput Methods Programs Biomed*. 2023;238:107601-107620.
43. Kasmi R, Hagerty J, Young R, et al. SharpRazor: automatic removal of hair and ruler marks from dermoscopy images. *Skin Res Technol*. 2023;29(4):e13203.
44. Kaur R, GholamHosseini H, Sinha R, Lindén M. Automatic lesion segmentation using atrous convolutional deep neural networks in dermoscopic skin cancer images. *BMC Med Imaging*. 2022;22(1): 103-115.
45. Feng S, Zhao H, Shi F, et al. CPF-Net: context pyramid fusion network for medical image segmentation. *IEEE Trans Med Imaging*. 2020;39(10):3008-3018.
46. Wu H, Chen S, Chen G, Wang W, Lei B, Wen Z. FAT-Net: feature adaptive transformers for automated skin lesion segmentation. *Med Image Anal*. 2022;76:102327-102340.

How to cite this article: Cai L, Hou K, Zhou S. Intelligent skin lesion segmentation using deformable attention Transformer U-Net with bidirectional attention mechanism in skin cancer images. *Skin Res Technol*. 2024;30:e13783. https://doi.org/10.1111/srt.13783