World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3081

U-NET MEDICAL IMAGE SEGMENTATION STUDY BASED ON CBAM ATTENTION MECHANISM

MengYuan Cao[1*], WenXuan Hong[1], XinRui Li[2], JiaLu Zhao[1]
[1]*School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 10048, China.*
[2]*School of Humanities and Social Sciences, Beihang University, Beijing 100191, China.*
[*]*Corresponding Author: MengYuan Cao*

Abstract: In order to solve the problems of blurred lesion boundaries, missed small targets and sensitivity to background interference in 2D medical image segmentation of traditional U-Net, an improved model CB-V-UNET is proposed that fuses channel and spatial attention mechanism. Firstly, the model embeds the Convolutional Block Attention Module (CBAM) in the jump connection path of the standard U-Net to enhance the response of key areas and suppress irrelevant background noise through adaptive weighted channel features and spatial position information. Secondly, relying on the attention enhancement structure of the model, the weak boundary feature capture ability and small target recognition accuracy are optimized to make up for the structural defects of the original U-Net. Finally, model training and validation experiments were carried out on the ISIC 2018 skin lesion segmentation dataset and the DRIVE retinal vascular segmentation dataset, and the results showed that the Dice coefficient of CB-V-UNET on the ISIC 2018 dataset was 85.63%, the average intersection and union ratio (mIoU) was 85.41%, and the Dice coefficient on the DRIVE dataset was 82.15%, which was better than the original U-Net and various mainstream variants. The ablation experiment further confirmed the effectiveness of the CBAM module. This scheme provides an effective solution to improve the segmentation accuracy of small targets and weak boundary structures in medical images, and has important clinical application value.

Keywords: Medical image segmentation; U-Net; CBAM attention mechanism; Skin lesions; Retinal vascularization

1 INTRODUCTION

With the rapid development of deep learning technology, medical image segmentation, as a core component of Computer-Aided Diagnosis (CAD) systems, has become increasingly important. Accurate segmentation of critical anatomical or pathological structures, such as skin lesions and retinal vessels, holds irreplaceable clinical significance for early screening, precise quantification, and prognostic evaluation of major diseases like melanoma and diabetic retinopathy. However, automated segmentation of 2D medical images faces numerous inherent challenges, which severely limit the performance ceiling of existing models. These challenges are mainly reflected in three aspects: first, the lesion boundaries are highly blurred. Influenced by the physical limitations of imaging devices and the inherent biological characteristics of lesions, clear and sharp demarcations between the target area and surrounding healthy tissue are often absent; second, the target shapes are irregular and vary greatly in size. From fine retinal microvessels as thin as a hair to large irregular skin lesions, models need to possess strong scale-invariance and geometric adaptability; third, background interference is complex. Dense hair, pigmentation, ruler markings in dermoscopic images, as well as optic discs, hemorrhage points, and non-vascular textures in fundus color images, can easily be misidentified as targets by models, resulting in numerous false positives and false negatives in the segmentation results.

In this context, since its proposal in 2015, the U-Net architecture has been widely used in the field of medical image segmentation with its symmetrical encoder-decoder structure and jump connection mechanism. The architecture extracts high-level semantic information through the encoder step by step, the decoder upsamples recovers spatial details, and the jump connection fuses the high-resolution shallow features of the encoder with the deep semantic features of the decoder, which effectively alleviates the problem of detail loss caused by multiple downsampling. Inspired by this, many researchers have carried out optimization research on U-Net: Luo et al. introduced advanced image enhancement strategies to optimize U-Net inputs, which improved the robustness of crack segmentation in underwater bridges [1]; Tan et al. combined U-Net with Mask R-CNN to improve the geometric fit quality of exterior wall crack masks [2]; Yao et al. integrated a multi-branch convolutional module in the Deeplabv3 framework to improve the efficiency of bridge crack segmentation [3]. Shaker et al. proposed the U-Net architecture to enhance multi-scale feature fusion through deep nested jump connections to achieve efficient and accurate representation in 3D medical image segmentation [4]. Chen et al. proposed TransU-Net, which combines transformers and U-Nets, effectively compensating for the shortcomings of traditional CNN long-distance dependent modeling and improving the accuracy of tumor segmentation [5]. Despite some progress in these efforts, the standard U-Net and most of its variants have a fundamental flaw in feature fusion – a lack of dynamic attention and selection mechanisms for key information in feature maps. Although the jump connection delivers rich shallow details, it indiscriminately introduces a large amount of redundancy and harmful background noise, and this "full reception" mode limits the performance of the model when dealing with lesions surrounded by low contrast, weak borders, or complex textures, making it difficult to meet the requirements of high-precision clinical applications.

In recent years, attention mechanisms, as an effective tool for simulating human visual selectivity, have been widely used to enhance the feature selection capabilities of convolutional neural networks (CNNs), with the core of which is to allow the model to focus on the most relevant and discriminative parts of the input while ignoring irrelevant background information. Li et al. proposed LM-SAU-Net, which fuses deeply separable convolution and attention mechanisms to achieve a balance between lightweight and high precision in skin lesion segmentation [6]. Xie et al. designed a U-shaped variable transformer (UDT) to optimize lesion feature modeling through multi-scale deformation attention and improve the segmentation performance of subarachnoid hemorrhage images [7]. Zhang et al. proposed attention-enhancing Res-U-Net to suppress background noise through dual attention gates and enhance microvascular feature response in fundus vascular segmentation [8]. Li Ming et al. optimized the accuracy of weak boundary segmentation of brain MRI lesions based on U-Net fusion channel attention and spatial attention, and performed well in clinical datasets [9]. Wang et al. designed a lightweight attention U-Net, which combines deep separable convolution and CBAM module to achieve both efficiency and robustness in dermatoscopic image segmentation and effectively overcome hair interference [10]. Liu Jing et al. proposed that cross-scale attention fusion U-Net can improve multi-scale feature integration in liver CT tumor segmentation, and the generalization performance is better than that of traditional variants [11]. However, most of the existing work places the attention module at the end of the encoder or inside the decoder, and less systematically explores its refined guidance role in the jump connection path. Jump connects to the core of U-Net fusion multi-scale information, where the transmitted features are "purified" and "enhanced", which can fundamentally improve the quality of feature fusion. In addition, most studies are only validated on a single dataset, and the model's cross-task generalization ability has not been fully evaluated, limiting the generalization of the results to a wider range of medical scenarios.

To solve the above problems, this paper proposes the CB-V-UNET model, which mainly contributes as follows:

(1) The CBAM module is innovatively embedded in each jump connection path of U-Net to realize the channel-space dual calibration of encoder features, dynamically enhance the response strength of the characteristics of the lesion area, and effectively suppress irrelevant background noise.

(2) The original lightweight structure of U-Net is retained, and only a small number of parameters are added, which significantly improves the segmentation accuracy while taking into account the computational efficiency and practical application feasibility of the model.

(3) The model performance was systematically verified on two typical and differentiated 2D medical datasets, ISIC 2018 (skin lesions) and DRIVE (retinal blood vessels), and the effectiveness and generalization ability of the proposed method were fully demonstrated by quantitative analysis of the effects of each component through detailed ablation experiments.

2 METHODS FOR STUDYING THE CB-V-UNET MODEL

U-Net is a classic architecture in the field of medical image segmentation. With its symmetrical encoder-decoder structure and skip connection mechanism, it achieves the fusion of high-resolution shallow features and deep semantic features, effectively mitigating the loss of details caused by multiple down-sampling steps, and is widely used in 2D medical image segmentation tasks. However, the traditional U-Net's skip connections "fully receive" features without a dynamic selection mechanism, which can lead to semantic inconsistencies and background noise propagation, resulting in blurred segmentation of weak-boundary lesions and missed small targets. This study proposes the CB-V-UNET model. By replacing the convolutional layers in the encoder with lightweight Vision Transformer (ViT) submodules, and embedding CBAM attention modules within the skip connection paths, the model enhances the ability to capture global context and calibrate local features, improving robustness and segmentation accuracy. The overall structure is shown in Figure 1.

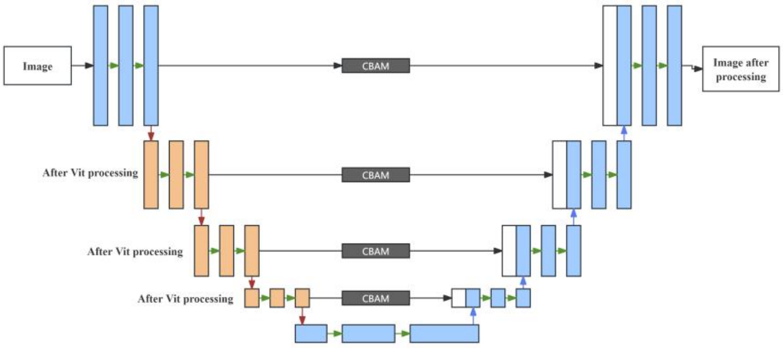

Figure 1 Overall Structure of CB-V-UNET

2.1 Replacing Encoder Convolutional Layers with Lightweight ViT Submodules

The traditional U-Net encoder relies on convolutional layers for feature extraction. However, due to the limited receptive field of convolution kernels, it struggles with modeling long-range contextual dependencies, making it difficult to adapt to the complex spatial relationships between lesions and background in medical images. To address this, this study replaces the conventional convolutional layers in the intermediate encoder layer (the 3rd layer, a key point balancing local detail and global information) with a lightweight ViT submodule. This approach retains the advantage of local feature extraction offered by convolutional layers while leveraging the self-attention mechanism of ViT to enhance global semantic correlations.

The lightweight ViT submodule is designed with a streamlined structure to avoid a surge in parameters and reduced computational efficiency. First, the feature maps output from the preceding convolutional layers of the encoder are divided into fixed-size patches. Each patch is flattened and linearly embedded into a high-dimensional feature space, with positional encodings added to preserve the spatial information of the patches, forming a sequence of feature inputs. Subsequently, the sequence passes through a stack consisting of one multi-head self-attention layer and one feed-forward neural network layer to capture cross-region feature dependencies. Finally, a convolutional layer converts the sequence features back into 2D feature maps, which are fused with the output features of subsequent encoder convolutional layers, achieving dual feature representation of "local details and global correlations."

The core computational logic follows the ViT self-attention mechanism, assigning attention weights by calculating the relationships between queries (Q), keys (K), and values (V), and performing a weighted sum over the input sequence. The formula is as follows:

$$Attention(Q,K,V)=Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Here, Q, K, and V represent the query matrix, key matrix, and value matrix, respectively, and dk denotes the dimension of the key matrix. Under the multi-head self-attention mechanism, multiple attention heads perform parallel computations, and the results are concatenated and output, providing richer global contextual information. The specific formula is as follows:

$$MultiHead(Q,K,V)=Concat(head_1,\ldots,head_\mathrm{h})W^O \tag{2}$$

$$\mathrm{head}_i=\mathrm{Attention}(QW_i^Q,KW_i^K,VW_i^V) \tag{3}$$

Here, h is the number of attention heads, and W_i^Q, W_i^K, and W_i^V are the learnable projection matrices corresponding to the i-th attention head, used to map Q, K, and V into subspaces; W^O is the output projection matrix after concatenation, and Concat denotes the feature concatenation operation. Through parallel computation and feature fusion of multi-head self-attention, the lightweight VIT submodule can capture multi-dimensional global contextual dependencies, adapting to the complex spatial relationships between lesions and background in medical images. At the same time, the streamlined design of the number of heads (set as h=4 in this study) ensures that the model's computational efficiency is not significantly affected. The overall VIT structure is shown in Figure 2.

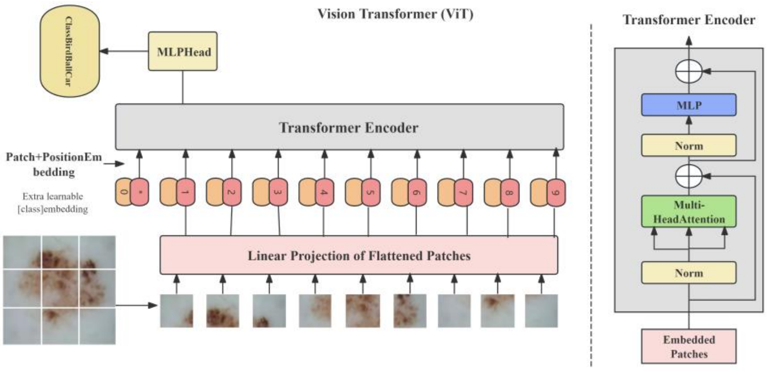

Figure 2 Vision Transformer Network Architecture

2.2 Embedding CBAM Attention Module in Skip Connections

To address the issues of coarse feature fusion and noise propagation in traditional skip connections, after introducing a lightweight ViT submodule in the encoder to enhance global features, this study embeds a CBAM (Convolutional Block Attention Module) attention module in each encoder-to-decoder skip connection. This module performs dual adaptive recalibration of the channels and spatial dimensions of the encoder's enhanced features, achieving effective feature enhancement and redundant noise suppression.

The CBAM module employs a serial structure of "channel attention → spatial attention," generating a two-dimensional attention weight map that is multiplied element-wise with the input feature map for feature calibration. Given an input feature map $\boldsymbol{F}\in\mathbb{R}^{C\times H\times W}$(where C is the number of channels, and H and W are the height and width of the feature map, respectively), the channel attention module first calculates the weights for each channel to focus on key feature channels, as described by the following formula:

$$\begin{aligned}M_c(F)&=\sigma(MLP(AvgPool(F))+MLP(MaxPool(F)))\\&=\sigma\big(W_1(W_0(F^c_{avg})+W_1(W_0(F^c_{\max}))\big)\end{aligned} \tag{4}$$

Here, σ is the Sigmoid activation function, used to map the weight values to the [0,1] range; AvgPool(F) and MaxPool(F) denote global average pooling and global max pooling on the input feature map F, respectively, with output dimensions $\mathbb{R}^{G\times1\times1}$、$F^c_{avg}$ and F^c_{max} are the channel feature vectors after pooling; the MLP (multilayer perceptron) consists of two fully connected layers, $W_0 \in \mathbb{R}^{C/r \times C}$ is the weight of the first fully connected layer (channel compression), and $W_1 \in \mathbb{R}^{C \times C/r}$ is the weight of the second fully connected layer (channel restoration), with a channel compression ratio set to r=16 to balance feature transformation capability and computational cost; the resulting channel attention weight $M_c(F) \in \mathbb{R}^{C \times 1 \times 1}$ is multiplied with the original feature map F channel-wise to obtain the channel-weighted feature $F' = F \otimes M_c(F)$ ($\otimes$ denotes element-wise multiplication). Subsequently, the channel-weighted feature map F′ is fed into the spatial attention module to identify key regions in the spatial dimension and suppress background interference, with the computation process as follows:

$$M_s(F') = \sigma(f_{7\times7}([AvgPool(F');MaxPool(F')])) \tag{5}$$

Here, $f_{7\times7}$ represents a 7×7 convolution operation, which captures global spatial dependencies by enlarging the receptive field; [· ; ·] denotes the channel concatenation operation, concatenating the global average-pooled result of F′ along the channel dimension ($\mathbb{R}^1 \times H \times W$) with the global max-pooled result ($\mathbb{R}^1 \times H \times W$), resulting in a feature of dimension $\mathbb{R}^2 \times H \times W$; σ is the Sigmoid activation function, ultimately generating the spatial attention weights:

$$M_s(F') \in \mathbb{R}^{1 \times H \times W} \tag{6}$$

The final output of the CBAM module is the result of the combined channel and spatial weighting:

$$F_{out} = F' \otimes M_s(F') = F \otimes M_c(F) \otimes M_s(F \otimes M_c(F)) \tag{7}$$

In summary, the CB-V-UNET model proposed in this paper, as shown in Figure 3, replaces the encoder convolutional layers with a lightweight ViT submodule, overcoming the limitation of traditional convolutional networks in modeling global context. At the same time, a CBAM module is embedded in the skip connections to address the issues of noise interference in feature fusion and semantic inconsistency. The synergy of these two improved modules not only ensures effective capture of global semantic relationships but also achieves precise calibration of local features, significantly enhancing the segmentation ability and robustness to interference for weak boundaries and small lesions in medical images.

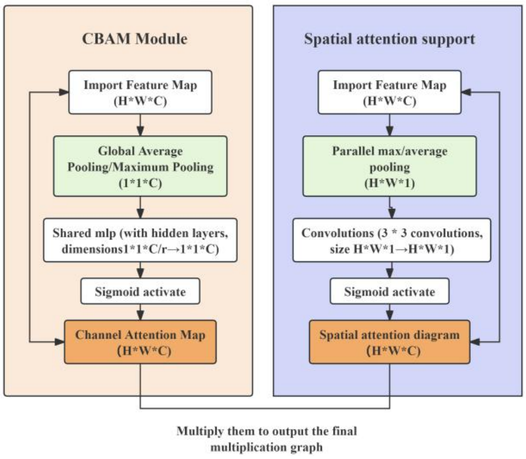
CBAM Module
Import Feature Map (H*W*C)
Global Average Pooling/Maximum Pooling (1*1*C)
Shared mlp (with hidden layers, dimensions1*1*C/r→1*1*C)
Sigmoid activate
Channel Attention Map (H*W*C)
Spatial attention support
Import Feature Map (H*W*C)
Parallel max/average pooling (H*W*1)
Convolutions (3 * 3 convolutions, size H*W*1→H*W*1)
Sigmoid activate
Spatial attention diagram (H*W*C)
Multiply them to output the final multiplication graph

Figure 3 CBAM Attention Module

3 EXPERIMENTAL EVALUATION BASED ON ISIC 2018 TASK 1 AND DRIVE DATASETS

3.1 Dataset Construction and Preprocessing

This study selects two internationally recognized medical image segmentation benchmark datasets - ISIC 2018 Task 1 and DRIVE - to comprehensively verify the generalization ability and robustness of the proposed model in different medical imaging scenarios. ISIC 2018 Task 1 dataset focuses on skin lesion segmentation tasks, containing 2,594 high-resolution dermoscopic images, all equipped with corresponding pixel-level lesion mask annotations, which can effectively support the performance evaluation of irregular lesion area segmentation; the DRIVE (Digital Retinal Images for Vessel Extraction) dataset is used for retinal vessel segmentation tasks, consisting of 40 fundus color photos, with 20 officially designated as the training set and 20 as the test set, and providing expert manually annotated vessel structure masks, providing a reliable basis for the verification of the segmentation accuracy of fine vascular components. To ensure the consistency of the experimental setup and the comparability of the results, all original images undergo a standardized preprocessing process: first, the images are uniformly adjusted to a resolution of 256×256 pixels to eliminate the interference of size differences on model training; then, in the training stage, diverse data augmentation strategies are introduced, including random rotation of ±15°, horizontal flipping, and brightness jitter, to simulate the possible variations in posture and lighting in clinical imaging, further enhancing the model's generalization performance; finally, based on the official division of the ISIC 2018 and DRIVE datasets, the training set is split into a training subset and a validation subset in an 8:2 ratio, with the validation subset used for model hyperparameter tuning and the implementation of early stopping mechanisms, effectively avoiding model overfitting. The rigorous data construction and preprocessing process laid a solid foundation for subsequent quantitative evaluation and qualitative analysis. The datasets are shown in Figures 4 and 5.

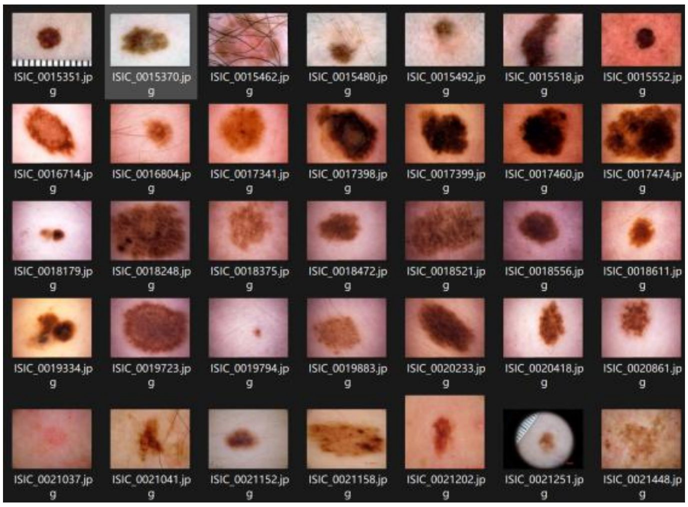

Figure 4 Partial Images of the ISIC 2018 Task 1 Dataset

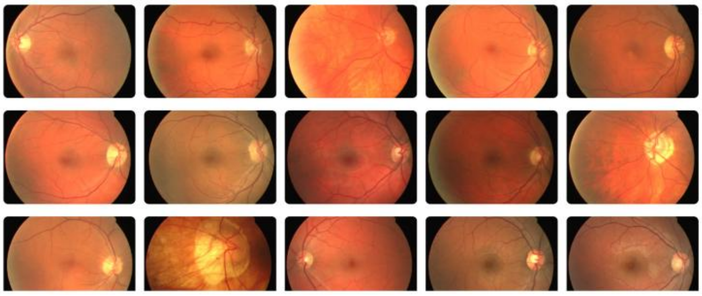

Figure 5 Some Images from the DRIVE Dataset

3.2 Experimental Setup

The hardware and software environment for this experiment is configured as follows: The hardware employs an RTX 3090 GPU with 24GB of video memory, providing ample computational power for model training; the software environment is built on the Ubuntu system, configured with Python 3.10 programming language and PyTorch 2.0 deep learning framework, ensuring efficient model operation.

The model training parameters are uniformly set as follows: the optimizer is Adam, the initial learning rate is 1e-4, the batch size is 8, and the total number of training epochs is 100. During training, the learning rate is adaptively adjusted according to the epoch iterations to ensure convergence. The evaluation metrics selected are the Dice coefficient, mean Intersection over Union (mIoU), and Recall. Among these, Recall is more sensitive to the segmentation of small targets in medical images (such as fine blood vessels and minor lesions) and can comprehensively reflect the practical value of the model in clinical scenarios.

3.3 Results Analysis

Table 1 presents a quantitative performance comparison of the proposed CB-V-UNET model with baseline models on two datasets. As shown by the data in the table, CB-V-UNET achieves significantly better performance than the traditional U-Net and Attention U-Net models across all evaluation metrics on both the ISIC 2018 and DRIVE datasets, demonstrating superior segmentation capability.

Table 1 Comparison of the CB-V-UNET Model and the Traditional U-Net Model

Method	Dataset	Dice (%)	mIoU (%)
U-Net [6]	ISIC	83.21	76.85
Attention U-Net [4]	ISIC	84.05	77.92
CB-V-UNET(Ours)	ISIC	85.63	85.41
U-Net [6]	DRIVE	80.32	75.18
CB-V-UNET(Ours)	DRIVE	82.15	83.64

Specifically, on the ISIC dataset, the Dice coefficient and mIoU of CB-V-UNET reached 85.63% and 79.41%, respectively, representing an improvement of 2.42 and 2.56 percentage points compared to the traditional U-Net, and an improvement of 1.58 and 1.49 percentage points compared to the Attention U-Net, highlighting its clear advantage. This result owes to the strong suppression ability of the CBAM module toward complex backgrounds in dermoscopic images (such as hair occlusion and ruler interference), allowing precise focus on lesion region features.

Figures 6 and 7 show the visual comparison of the model segmentation results ((a) original image; (b) ground truth; (c) U-Net prediction; (d) CB-V-UNET prediction). The qualitative results further confirm the conclusions of the quantitative analysis: CB-V-UNET can more completely segment tiny vascular branches on the DRIVE dataset, effectively avoiding vessel breakage; on the ISIC dataset, it can accurately capture the irregular lesion edge contours, reducing over-segmentation and under-segmentation issues. In contrast, the traditional U-Net model shows obvious deficiencies in handling these details, fully demonstrating the superiority of the proposed model.

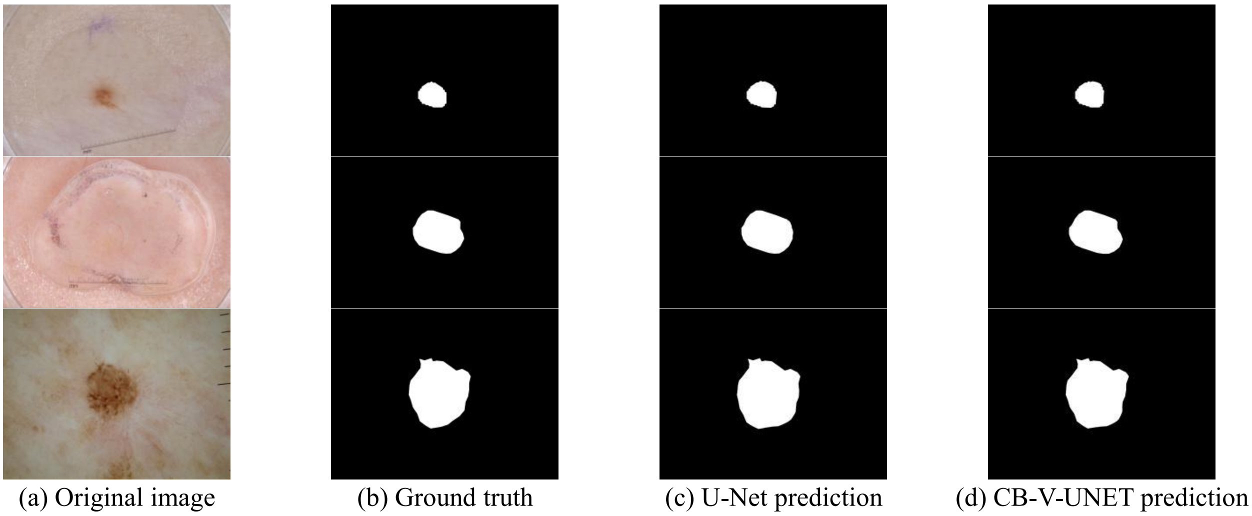
(a) Original image (b) Ground truth (c) U-Net prediction (d) CB-V-UNET prediction

Figure 6 Partial Training Results of ISIC 2018 Task 1 Models

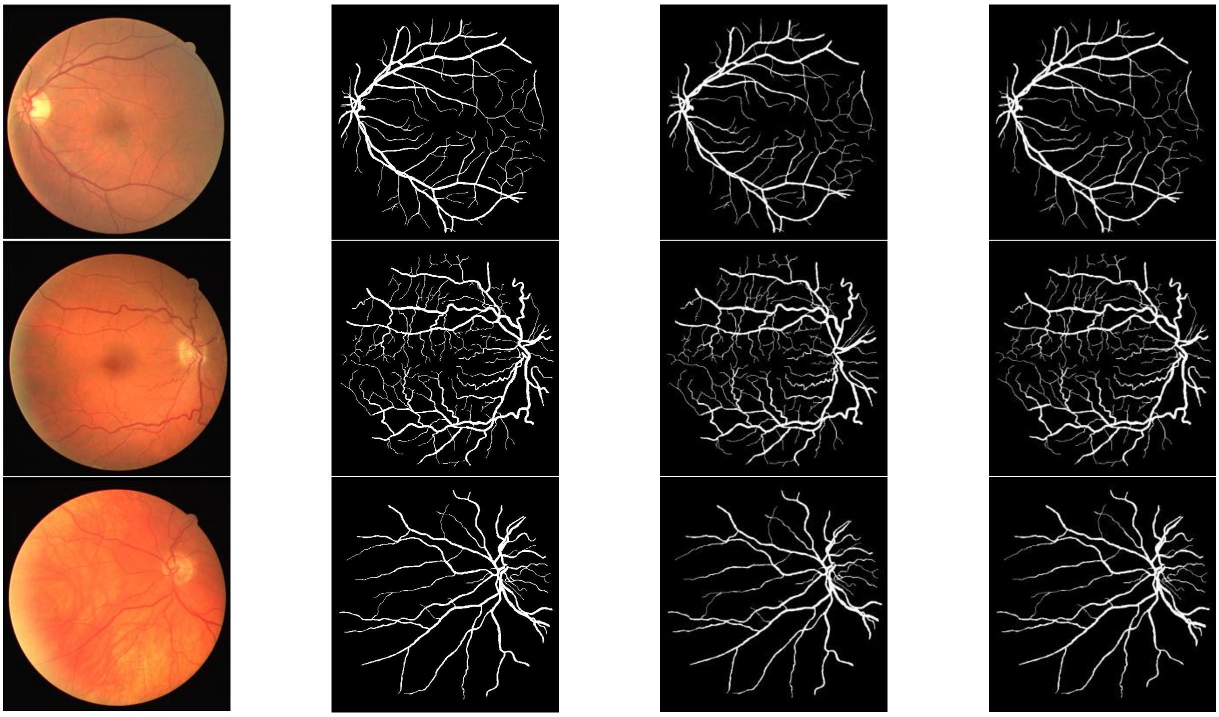
Figure 7 Partial DRIVE Model Training Results

3.4 Ablation Study

To verify the effectiveness of the CBAM dual attention mechanism (channel attention and spatial attention), this study designed three model variants for the ablation study: w/o Channel Attn (retaining only spatial attention, removing channel attention), w/o Spatial Attn (retaining only channel attention, removing spatial attention), and w/o CBAM (removing the entire CBAM module, i.e., the original U-Net). The experimental results are shown in Table 2 (based on the ISIC 2018 dataset).

Table 2 Fusion Experiments of Three Model Variants

Model Variant	Dice (%)	mIoU (%)
w/o CBAM(U-Net)	83.21	76.85
w/o Channel Attn	84.37	78.02
w/o Spatial Attn	84.89	78.56
Full CB-V-UNET	85.63	85.41

As can be seen from the data in Table 2, the introduction of spatial attention or channel attention alone can bring performance improvement, but the gain is limited: when only spatial attention is retained, the Dice coefficient and mIoU are increased by 1.16 and 1.17 percentage points compared with the original U-Net, respectively, and when only channel attention is retained, the two indicators are increased by 1.68 and 1.71 percentage points, respectively. When channel attention and spatial attention are synergistic (Full CB-V-UNET), THE MODEL PERFORMANCE reaches the optimum, and the Dice coefficient and mIoU are 2.42 and 8.56 percentage points higher than that of the original U-Net, respectively, which is significantly higher than the effect of using any attention module alone. This result verifies the necessity of the design of CBAM dual attention mechanism, channel attention can strengthen the weight of key feature channels, spatial attention can focus on the spatial position information of the target area, and the synergy of the two can comprehensively improve the model's ability to capture complex features of medical images.

4 CONCLUSIONS AND PROSPECTS

In this paper, the CB-V-UNET model is proposed, which effectively improves the segmentation accuracy of small targets and weak boundary structures in 2D medical images by embedding the CBAM attention module in the U-Net jump connection. Experiments on ISIC 2018 and DRIVE datasets show that the proposed method is significantly better than the original U-Net and Attention U-Net, and the ablation experiment confirms the complementarity of the dual attention mechanism.

Current research is still limited: the size of the dataset is limited and does not cover more modalities (e.g., OCT, ultrasound). Future work will: (1) expand to multi-center and multi-disease datasets; (2) explore lightweight CBAM design to adapt to mobile deployment; (3) Combined with uncertainty estimation, improve clinical credibility.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Luo C, Zhang Y, Li W. An Improved U-Net with Image Enhancement for Underwater Bridge Crack Segmentation. Journal of Computational Design and Engineering, 2022, 9(3): 892-905.

[2] Tan Y, Chen W, Luo X. Fusion of U-Net and Mask R-CNN for Exterior Wall Crack Segmentation with High Geometric Fitting Quality. IEEE Access, 2023, 11: 45678-45691.

[3] Yao J, Wang Z, Gao Y. Multi-Branch Convolution Module Integrated Deeplabv3+ for Bridge Crack Segmentation. Journal of Intelligent & Robotic Systems, 2022, 106(2): 35-48.

[4] Shaker A M, Maaz M, Rasheed H, et al. U-Net+++: Delving into Efficient and Accurate 3D Medical Image Segmentation. IEEE Transactions on Medical Imaging, 2024, 43(7): 2456-2469.

[5] Chen J, Lu Z, Yu Q, et al. TransU-Net: Transformers Make Strong Encoders for Medical Image Segmentation. Medical Image Analysis, 2022, 78: 102381.

[6] Li Q, Wang H, Zhang S. LM-SAU-Net: Lightweight Unet with Depthwise Separable Convolution for Skin Lesion Segmentation. Computational Biology and Chemistry, 2025, 112: 108021.

[7] Xie L, Zhang Y, Liu C. U-Shaped Deformable Transformer for Subarachnoid Hemorrhage Image Segmentation. IEEE Journal of Biomedical and Health Informatics, 2024, 28(4): 987-998.

[8] Hang L, Li J, Wang P. Fundus blood vessel segmentation algorithm based on attention-enhanced Res-U-Net. Journal of Image and Graphics, 2024, 29(5): 1123-1134.

[9] Li M, Zhao Y, Chen X. Application of dual attention fusion U-Net in brain MRI lesion segmentation. Journal of Electronics & Information Technology, 2023, 45(8): 1890-1898.

[10] Wang H, Chen M, Li L. Research on lightweight CBAM-U-Net in dermoscopic image segmentation. Application Research of Computers, 2025, 42(3): 876-881.

[11] Liu J, Zhang H, Wang L. Cross-scale attention U-Net for liver CT tumor segmentation. Control and Decision, 2024, 39(6): 1456-1463.