sensors

MDPI

Article

MDEU-Net: Medical Image Segmentation Network Based on Multi-Head Multi-Scale Cross-Axis

Shengxian Yan, Yuyang Lei, Jing Zhang, Xiao Gao, Xiang Li, Penghui Wang and Hui Cao *

Shaanxi Key Laboratory of Ultrasonics, School of Physics and Information Technology, Shanxi Normal University, Xi'an 710062, China; 2023303971@snnu.edu.cn (S.Y.); lyy_weirong@snnu.edu.cn (Y.L.); zhanggale@snnu.edu.cn (J.Z.); gao0515@snnu.edu.cn (X.G.); lixiangideal@snnu.edu.cn (X.L.); wang_penghui@snnu.edu.cn (P.W.)
* Correspondence: caohui@snnu.edu.cn

Abstract: Significant advances have been made in the application of attention mechanisms to medical image segmentation, and these advances are notably driven by the development of the cross-axis attention mechanism. However, challenges remain in handling complex images, particularly in multi-scale feature extraction and fine-detail capture. To address these limitations, this paper presents a novel network architecture, multi-head multi-scale cross-axis attention MDEU-Net, that leverages a multi-head attention mechanism processing input features in parallel. The proposed architecture enables the model to focus on both local and global information while capturing features at various spatial scales. Additionally, a gated attention mechanism facilitates efficient feature fusion by selectively emphasizing key features rather than relying on simple concatenation and improves the model's ability to capture critical details at multiple scales. Furthermore, the incorporation of residual connections further mitigates the gradient vanishing problem by enhancing the model's capacity to capture complex structures and fine details. This approach accelerates computation and enhances processing efficiency, while experimental results demonstrate that the proposed network outperforms traditional architectures in terms of performance.

Keywords: medical image segmentation; cross-axis attention; multi-scale features; multinomial attention; feature fusion

Academic Editor: Francesco Mercaldo

Received: 1 April 2025
Revised: 28 April 2025
Accepted: 3 May 2025
Published: 5 May 2025

Citation: Yan, S.; Lei, Y.; Zhang, J.; Gao, X.; Li, X.; Wang, P.; Cao, H. MDEU-Net: Medical Image Segmentation Network Based on Multi-Head Multi-Scale Cross-Axis. *Sensors* **2025**, *25*, 2917. https://doi.org/10.3390/s25092917

1. Introduction

Medical image segmentation is a fundamental task in medical image processing with broad applications in areas such as disease diagnosis [1,2], surgical planning, and lesion monitoring [3–6]. Lesions are precisely identified and located using this technique that enables more accurate diagnoses. The development of precise surgical plans to minimize risks and complications is also supported while measurements of the size and volume of organs, tissues, or lesions are provided. Additionally, it offers quantitative metrics for assessing the condition and monitoring treatment outcomes.

Currently, two-dimensional segmentation models based on convolutional neural networks (CNNs) [7,8] have become widely adopted in medical image segmentation [9,10]. One of the most prominent models is U-Net [11], which features a unique U-shaped encoder–decoder architecture and improves segmentation accuracy through the introduction of skip connections. This is particularly useful for medical image processing tasks. So, the success of U-Net has inspired the development of various network architectures, including several U-Net variants. U-Net++ [12] enhances model performance by replacing traditional connectivity with nested and dense skip connections [13,14]. Additionally,

ResU-Net [15] models incorporate residual connections into U-Net to mitigate the gradient vanishing problem [16] during training and demonstrate strong performance in medical image segmentation tasks.

With the continuous development and refinement of network models. TBConvL-Net [17] combines the advantages of Transformer global modeling and Convolutional Neural Network (CNN) local feature extraction to address the instability problem in medical image segmentation. However, it still has shortcomings in extracting detailed and boundary features. Cluster Center Transformer [18] and LCAUnet [19] can effectively focus on the local edges around the region by using the attention mechanism, but they still have limitations in generalization ability and computational resource requirements. QRMFO [20] is an optimization algorithm simulating moth–flame behavior with strong global search capability and an efficient parallel mechanism. However, it may prematurely converge to local optima, failing to attain the global optimum. MCANET's [21] attention mechanisms mitigate traditional CNNs' long-range dependency capture limitations; however, its single-head design fails to reliably model global context, leading to performance issues.

Conventional convolutional neural networks, comprising feedforward neural networks (FNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNN), alongside transformer neural networks frequently utilized in prevalent large-scale models, possess the capability to extract both local and global features within medical image segmentation applications. However, when dealing with complex and variable morphologies, it is challenging to achieve precise segmentation, especially in cases involving pathological regions or intricate organ morphology details, where its performance becomes limited. In medical image segmentation, the accuracy of segmentation is significantly affected by factors that contribute to challenges in achieving precise results, such as morphology, scale, and boundary ambiguity. As a result, developing methods to effectively capture multi-scale information in medical images [22,23] and improving model robustness and accuracy through flexible network architectures have become key research focuses in the field.

An improved method for medical image segmentation integrating cross-axis attention [21,24,25] and multi-scale feature fusion to enhance segmentation accuracy is presented in this paper. First, the introduction of the cross-axis attention mechanism enables the model to effectively capture global information from multiple directions within the image, thereby boosting the boundary recognition capabilities of complex lesion regions. Simultaneously, multi-scale feature extraction refines the feature representation of lesion regions at different scales and enhances the model's ability to process lesions of varying sizes. Next, efficient feature fusion [26–28] is implemented through a gating mechanism that selectively emphasizes key features and mitigates the gradient vanishing problem via residual connections. This enhances the model's ability to capture complex structures and details. Finally, the multi-head attention mechanism learns attention patterns across multiple subspaces in parallel to enhance the model's robustness and segmentation accuracy for medical images with complex morphologies and ambiguous boundaries.

In summary, our main contributions can be summarized as follows:

1. Cross-axis attention mechanisms enable cross-exchange computation between x-axis and y-axis information. This achieves a more comprehensive capture of critical visual features within images.

2. Multi-scale feature extraction through hierarchical scale-based feature capture effectively accommodates structural variations in size and morphology within medical imaging data. This approach demonstrates particular advantage for segmentation of minute pathological entities requiring discriminative feature discrimination across spatial granularities.

3. Multi-head attention mechanisms perform parallel computation across independent attention heads, enabling information capture across diverse subspaces. Each head learns distinct attention patterns, and their fusion into a unified representation significantly enhances model performance.

4. Efficient feature fusion via gating mechanisms selectively attends to critical regions of input feature maps instead of mere concatenation, enabling enhanced multi-scale feature capture. Residual connections alleviate vanishing gradients by directly transmitting inputs to deeper layers.

2. Network Architecture

The MDEU-Net model is proposed in this paper to address the challenges of multi-scale information extraction and detail capture in complex images, and it incorporates the MDMSC Multihead Attention [29] module and the EF Efficient Feature Fusion module.

2.1. General Organization

The architecture of the MDEU-Net model enhances the U-Net encoder–decoder structure, as shown in Figure 1. The encoder is composed of four convolutional modules that each consist of two convolutional layers [30] followed by a ReLU activation function [31] and a max pooling [32] operation. Within the module, the feature map undergoes a 3×3 convolution and ReLU activation. Then, the 'same' padding is employed to ensure the spatial dimensions of the feature map remain unchanged. Subsequently, a downsampling operation is performed, which reduces the size of the feature map by half, while the number of channels is adjusted according to the labeling shown in Figure 1. Furthermore, a multi-head, multi-scale cross-axis attention mechanism is incorporated to enhance feature extraction. The decoder also consists of four convolutional modules that each contain an upsampling layer, an efficient feature fusion module, and two convolutional layers. The key innovations of the model include the incorporation of a multi-head, multi-scale cross-axis attention mechanism at each downsampling stage, which enhances multi-scale feature modeling. The model also adopts an efficient feature fusion module that replaces the traditional U-Net skip connections and improves feature fusion and information transfer efficiency.

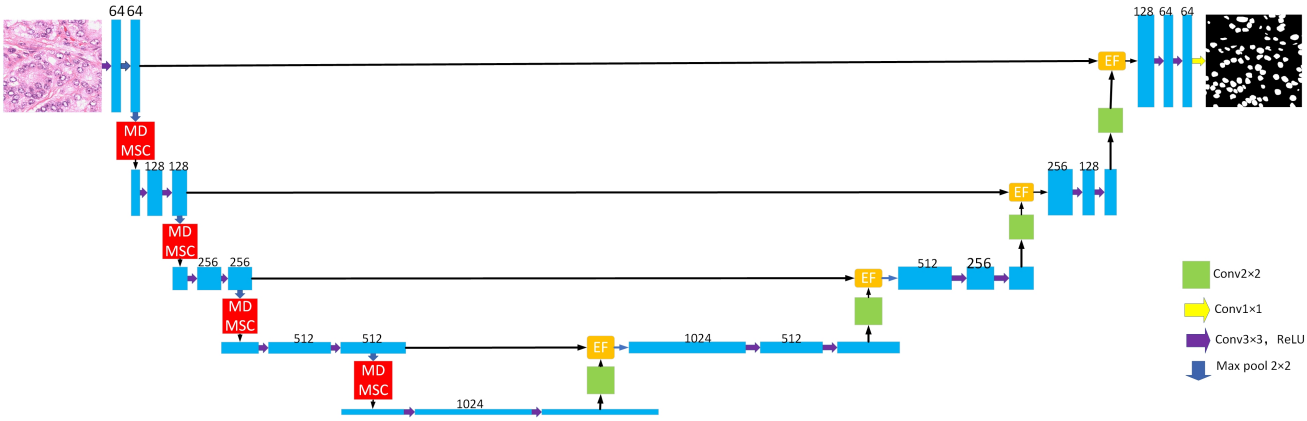

Figure 1. MDEU-Net network overall architecture.

2.2. Multi-Head Multi-Scale Cross-Axis Attention Module

As shown in Figure 2a, the proposed architecture consists of eight parallel multi-scale channel (MSC) modules that extract multi-scale features from the image through downsampling and distribute the channels evenly to eight single-head attention modules. Each single-head attention module follows the MSC principle by extracting multi-scale features and then sequentially stacking the output channels of all modules. A residual

connection (RC) is introduced that facilitates effective information transfer to mitigate the vanishing gradient problem and reduce information loss.

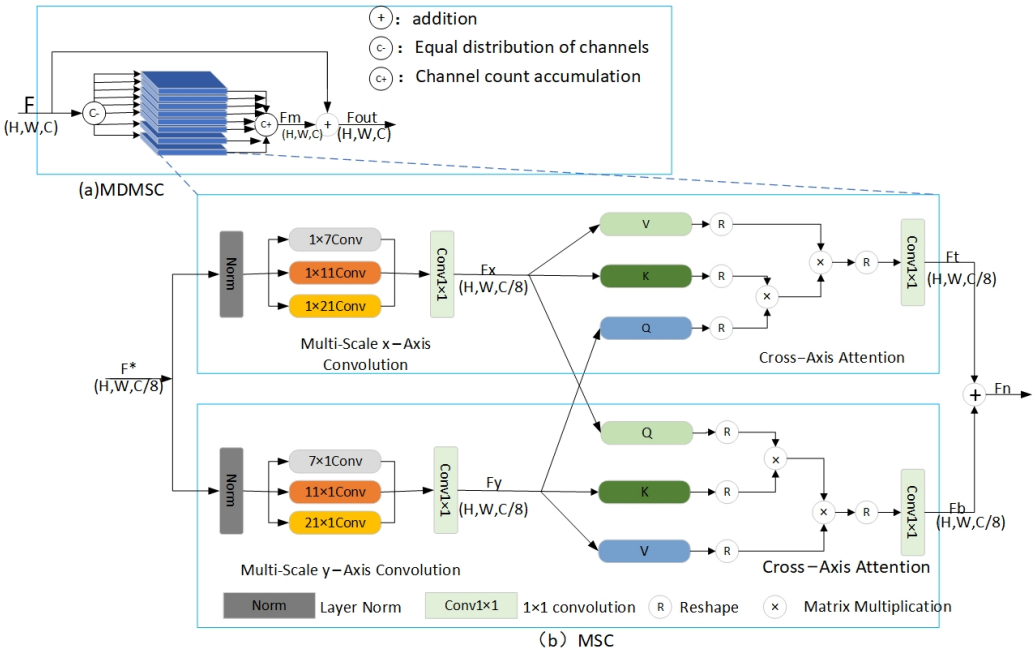

Figure 2. MDMSC network structure.

The multi-scale cross-axis attention structure illustrated in Figure 2b features a network design that consists of two parallel branches, each of which is responsible for extracting the horizontal and vertical features of the image, respectively. Each branch employs a distinct 1D convolutional kernel size to capture multi-scale contextual information along a single spatial dimension. These features are integrated and enhanced in the orthogonal spatial dimension through a cross-axis attention mechanism that facilitates the capture of both horizontal and vertical information. The core idea of this approach is to leverage the multi-layer perceptron (MLP) capabilities of convolution to construct fine-grained hierarchical feature representations and to enable comprehensive and detailed extraction of image features.

Taking the topmost branch in Figure 2b as an example, the input is $F^* \in R^{H \times W \times \frac{C}{8}}$ and the F_x output [21] is

$$F_x = Conv1 \times 1\left(\sum_{i=0}^{2} Conv1D_i^x(Norm(F^*))\right) \quad (1)$$

where it denotes the one-dimensional convolution along the x-axis that we set to 1×7, 1×11, and 1×21 according to SegNeXt [33]. $Norm(\cdot)$ denoting layer normalization, and $Conv1 \times 1$ denoting a 1×1 convolution. Similarly, for the bottom branch, the F_y output can be expressed as [21]

$$F_y = Conv1 \times 1\left(\sum_{i=0}^{2} Conv1D_i^y(Norm(F^*))\right) \quad (2)$$

Inspired by the self-attention mechanism of the Transformer, we have come to understand the crucial roles of the Key Matrix, the Value Matrix, and the Query Matrix in the attention mechanism. Specifically, the primary function of the Key Matrix is to extract significant feature information from the input data, essentially labeling each data point with a unique feature identifier. In contrast, the Value Matrix carries the actual content to be output or the weight information, integrating these features through weighted aggregation based on their relevance to the query. Lastly, the Query Matrix represents the current input information or the focal

point of the model's attention. We propose to compute the cross-attention between F_x and F_y, which aims to better utilize the multi-scale convolutional features in both spatial directions. F_x will be used as the key matrix [34] and the value matrix [35], and F_y will be used as the query matrix [36]. The computation is as follows [21]:

$$F_t = MHCA_y(F_y, F_x, F_x) \tag{3}$$

where $MHCA_y(\cdot,\cdot,\cdot)$ denotes the cross attention along the x-axis, and F_t represents the output obtained after performing attention calculation along the x-axis. For the bottom branch, feature extraction is performed along the y-axis in the same way as denoted below [21]:

$$F_b = MHCA_y(F_x, F_y, F_y) \tag{4}$$

$MHCA_x(\cdot,\cdot,\cdot)$ denotes cross attention along the y-axis, and F_b represents the output obtained after performing attention calculation along the x-axis.

For the obtained F_t and F_b, the output F_n of the proposed multi-scale cross-axis attention can be expressed as [21]

$$F_n = Conv1 \times 1(F_t) + Conv1 \times 1(F_b) \tag{5}$$

The single-head attention works as above, and the overall MDMSC output F_{out} is expressed as

$$Fout = F + F_m \tag{6}$$

where F is the downsampled output. F_m represents the feature output after the fusion of eight single-head attention mechanisms.

2.3. Efficient Feature Fusion Module

As illustrated in Figure 3, the EF module consists of two sub-modules: EA and EC. The EA module (gated attention mechanism) fuses low-level and high-level semantic features through skip connections and refines them using residual connections. The fused features are then passed to the EC module for further feature extraction.

The structure of the EA module is depicted in Figure 3c. Initially, group convolution is employed for intra-group feature fusion that contrasts with standard convolution and reduces computational complexity significantly. Following, the ReLU activation function is applied after convolving the input features. The low-level semantic features are then concatenated through residual connections that are transmitted via skip connections. Finally, the residual connection helps mitigate the interference of high-level semantic features on low-level features when the correlation between two input features is weak, and this helps prevent degradation in overall model performance. The computational expression is as follows [37]:

$$EA(g,x) = x \times \left(1 + Sigmoid\left(Conv1 \times 1\left(Relu\left(W_g + W_x\right)\right)\right)\right) \tag{7}$$

$$W_g = Relu(BN(GroupConv32(g))) \tag{8}$$

$$W_x = Relu(BN(GroupConv32(x))) \tag{9}$$

where sigmoid and ReLU are the activation functions, and BN is the Batch Normalization operation. GroupConv32 is the 32-component group convolution, and Conv1 × 1 is the regular convolution with a convolution kernel size of 1 × 1. In this model, g is the semantic feature obtained by up-sampling, and x is the low-level semantic feature passed by the jump connection.

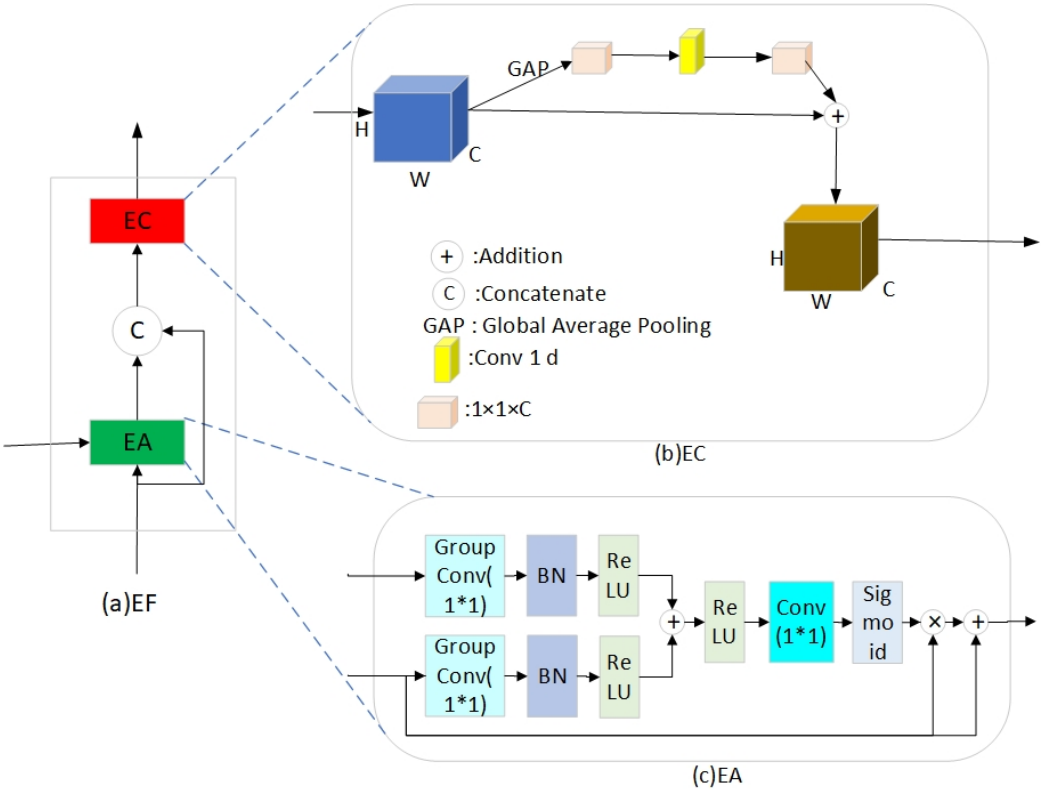

Figure 3. Structure of EF efficient feature fusion module.

The structure of the EC module is depicted in Figure 3b, and it incorporates a residual network based on the ECA from S-Unet [38]. The EC module eliminates the need for additional weight parameters through the global average pooling (GAP) layer, which consists of a GAP layer followed by a 1 × 1 convolutional layer. This significantly reduces the model's parameter count and mitigates the risk of overfitting while improving computational efficiency. However, the GAP operation compresses the spatial dimensions of the feature map into a single global average, which results in the loss of spatial structure information and potentially causes the loss of local spatial details. The limitation is effectively addressed by introducing the residual network that enhances the model's ability to recover spatial information.

3. Experiments and Discussion of Results

3.1. Experimental Preparation

3.1.1. Experimental Data

This study utilizes the PanNuKe dataset for cell nucleus segmentation [39], which consists of 300 accurately labeled microscopic images. For model training and evaluation, the dataset is split into 260 for training and 40 for testing. The image annotations in the PanNuKe dataset were manually performed by skilled technicians who covered the precise contours and locations of the cell nuclei to ensure that the model could learn accurate segmentation information. For the retinal blood vessel segmentation task, we utilized the Retinal Blood Vessels in the Eye (RITE) dataset from the DRIVE database, which consists of 40 high-quality retinal blood vessel images. The dataset was split into 30 training and 10 testing groups for model training and evaluation. Each image is meticulously labeled with the exact contours and structural details of the blood vessels that aid the model in learning the intricate features of retinal blood vessels. Data augmentation techniques were applied during training to enhance model performance and fully utilize the dataset. These techniques expand the dataset in a way that increases sample diversity, reduces the risk of

overfitting, and improves the model's generalization ability and robustness. The images in the training set were subjected to rotation, translation, scaling, and color transformations. Additionally, data augmentation helps mitigate noise, uneven illumination, and biases in biomedical images, which in turn improves segmentation accuracy.

3.1.2. Experimental Methods

All experiments were conducted using the PyTorch 2.0.0 framework on a computing platform with Python 3.8 (Ubuntu 20.04 operating system) and an Nvidia RTX 4090 GPU (24 GB graphics memory). The MDEU-Net model was trained with the Adam optimizer, a learning rate of 0.0001, momentum of 0.9, and weight decay of 1×10^{-8} . The model was trained for 200 epochs on the DSB2018 dataset with a batch size of 4 and for 40 epochs on the RITE dataset with a batch size of 2. After training, After model training, the loss function curve shown in Figure 4 indicates a significant decrease in the cell nucleus segmentation task between epochs 0 and 100, stabilizing at approximately 0.04. Similarly, the ocular vascular segmentation task showed a notable reduction between epochs 0 and 30 that ultimately converged to around 0.1. The effectiveness of the MDEU-Net model was demonstrated during the training process when its performance gradually improved over time.

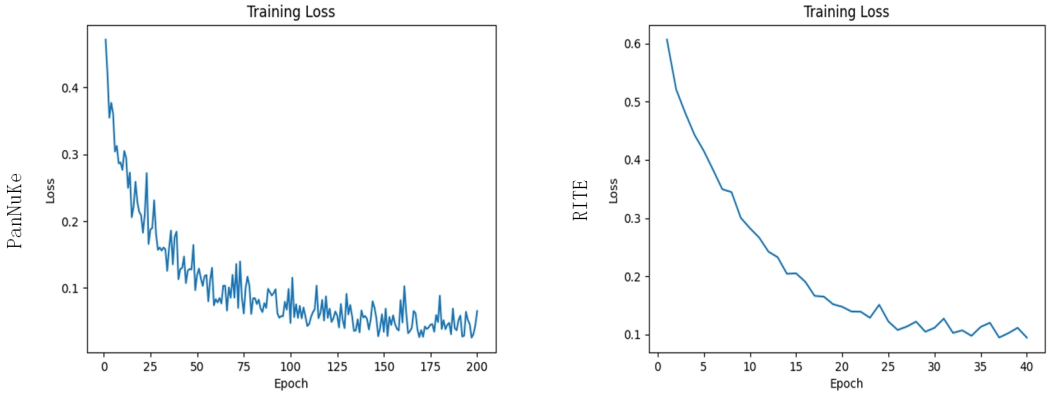

Figure 4. Training 200 rounds of nuclei and 40 rounds of retinal vessel loss function images.

3.2. Experimental Evaluation Indicators and Results

3.2.1. Evaluation Indicators

Semantic segmentation is a pixel-level task in image segmentation. Commonly used evaluation metrics include recall, mean intersection over union (mIoU), and accuracy (Acc). All of these metrics are computed from the confusion matrix. The definitions of true positive (*TP*), false positive (*FP*), false negative (*FN*), and true negative (*TN*) are provided in Table 1:

TP: indicates that a sample is predicted to be a positive example and the true label is a positive example.

FN: indicates that a sample is predicted to be a counterexample and the true label is a positive example.

FP: indicates that a sample is predicted to be a positive example and the true label is a negative example.

TN: indicates that a sample is predicted to be a counterexample and the true label is a counterexample.

The binary semantic segmentation evaluation metric can be expressed as

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

$$mIoU = \frac{\frac{TP}{TP+FP+FN} + \frac{TN}{TN+FN+FP}}{2} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

Table 1. Confusion matrix for classification results.

The Real Situation	Projected Results	
	Standard Practice	**Counter-Example**
standard practice	TP (True Positive)	FN (False Negative)
counter-example	FP (False Positive)	TN (True Negative)

3.2.2. Experimental Results and Analysis

Table 2 presents the performance comparison of seven deep learning models for cell nucleus and retinal blood vessel segmentation tasks. As shown in Table 2, the MDEU-Net model significantly outperforms other medical image segmentation models in both cell nucleus and retinal blood vessel segmentation. Compared with the original U-Net model in the task of cell nucleus segmentation, MDEU-Net improves the mIoU, accuracy (Acc), and recall by 11.52% , 3.51%, and 6.19%, respectively. Improvements of 8.33%, 0.6%, and 5.59% in mIoU, accuracy, and recall are observed for retinal blood vessel segmentation, respectively. MDEU-Net achieves a 0.47% improvement in mIoU for the cell nucleus segmentation task, demonstrating its effectiveness compared with the more advanced MCANET model. For the retinal vessel segmentation task, it demonstrates improvements of 1.02% in mIoU, 0.13% in accuracy (Acc), and 2.40% in recall, respectively. These results demonstrate that MDEU-Net achieves superior segmentation performance, which is particularly evident in the recall rate. It also maintains a relatively small model parameter size and shows notable improvements in both accuracy and intersection over union (IoU).

Table 2. Comparison of the results of cell nuclei and retinal vasculature in the MDEU-NET model and other segmentation model metrics.

	Params	**Flops**		**RITE**			**PanNuKe**	
Method	**M**	**G**	**mIoU**	**Acc (%)**	**Recall (%)**	**mIoU**	**Acc (%)**	**Recall (%)**
U-Net	1.56	4.08	83.77	98.17	87.81	80.01	92.67	87.23
Res-Unet	4.11	2.56	80.67	97.98	82.04	82.53	93.07	88.31
UNet++	13.41	31.13	87.40	98.67	88.37	85.42	94.19	91.06
Att-Unet	13.75	32.23	85.21	98.43	87.07	85.94	94.41	90.97
SUnet	23.01	6.31	86.74	98.53	90.24	88.00	95.30	92.82
MCANET	5.56	16.44	90.08	99.04	91.80	91.16	-	-
MDEU-Net	17.18	44.11	92.32	99.19	93.45	91.53	96.18	93.42

As shown in Figure 5, the MDEU-Net model demonstrates exceptional segmentation performance in both retinal blood vessel and cell nucleus segmentation tasks while maintaining low computational complexity. This indicates that the model not only enhances segmentation accuracy but also reduces the demand for hardware resources.

As shown in Figure 6, it compares the segmentation results of the MDEU-Net model with those of other networks for cell nucleus and retinal blood vessel segmentation. MDEU-Net efficiently captures global features and accurately recognizes object boundaries and contours. Furthermore, the model excels in local feature extraction that is particularly effective in handling complex backgrounds and fine structures.

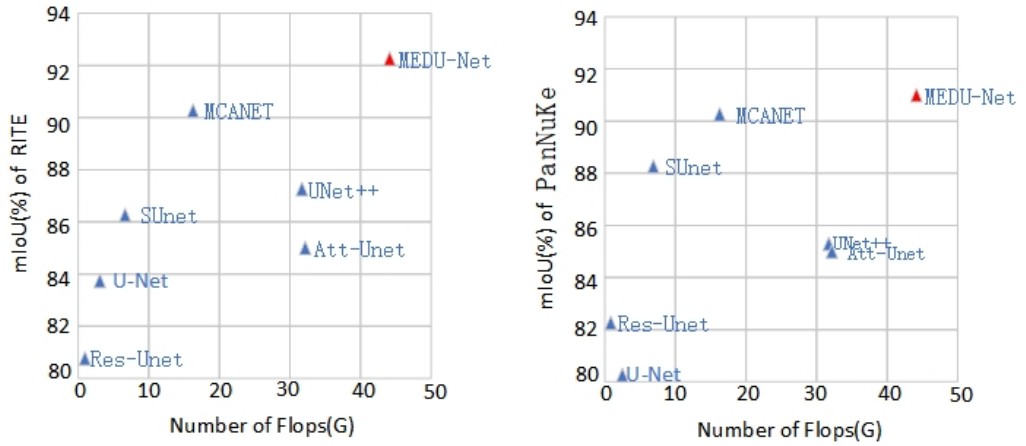

Figure 5. Comparison of segmentation model parameters and segmentation results.

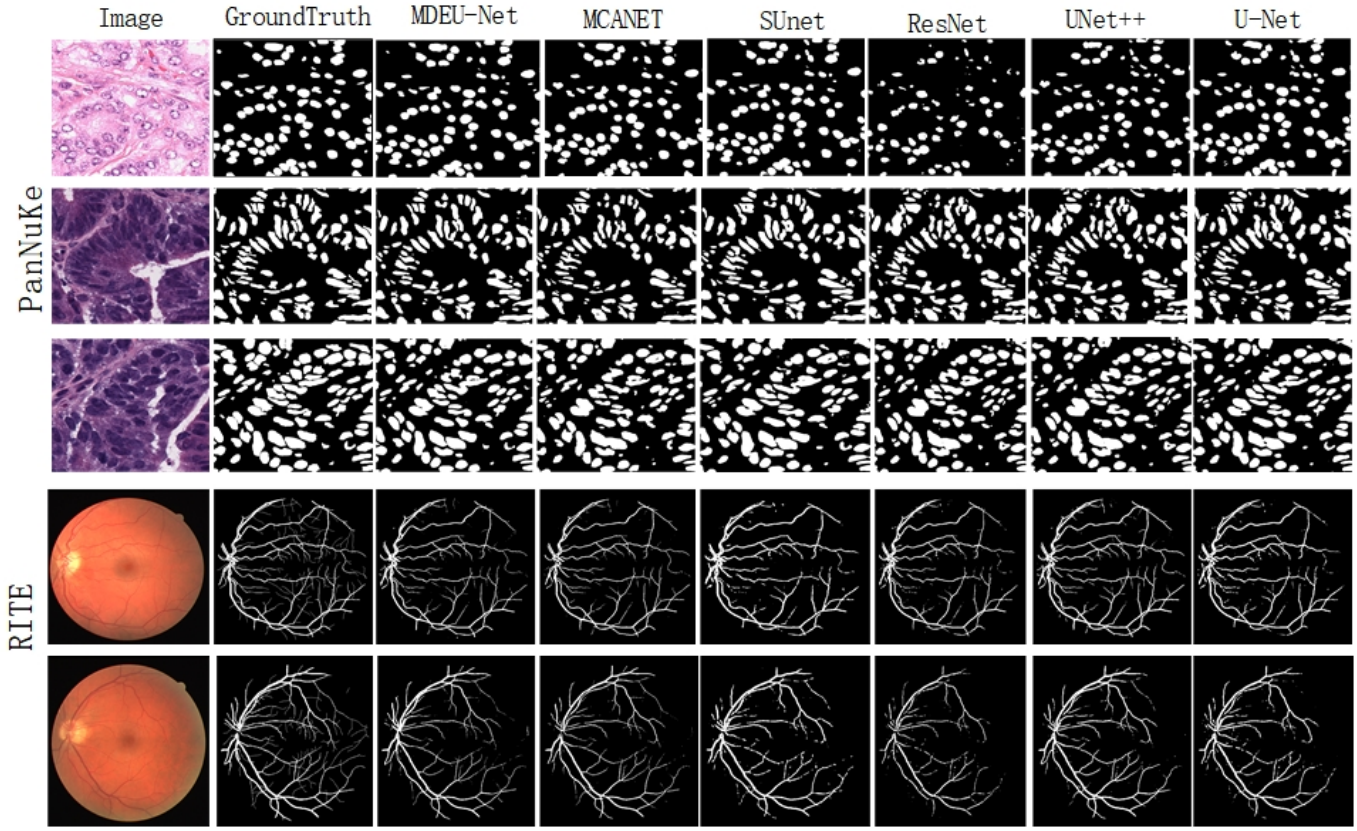

Figure 6. Visualization of the segmentation results of different methods for cell nuclei and retinal vessels.

As shown in Figure 7, a comparison of the magnified prediction maps for different segmentation models in specific regions containing labeled maps that involve MDEU-Net, S-Unet, and Res-Unet is presented. The MDEU-Net model demonstrates superior contour accuracy in the nucleus segmentation magnification map, which shows a better fit with the labeled map than the other models. In retinal vascular segmentation magnification maps where MDEU-Net outperforms other models in terms of vascular continuity and detail restoration, it places particular emphasis on the precise reconstruction of capillaries and vessel termini that closely align with the labeled maps.

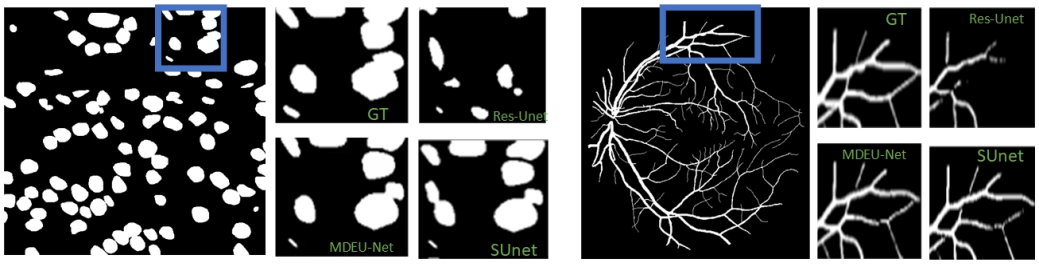

Figure 7. Enlarged view of different model visualizations.

The multi-head attention mechanism is a critical component in modern neural networks. Unlike single-head attention, which shares query, key, and value matrices, multi-head attention assigns independent matrices to each head and enables the model to focus on different parts of the input simultaneously. The optimal number of heads is not universally fixed and varies depending on the specific task and model architecture. To evaluate the impact of the multi-head attention mechanism on segmentation performance, Table 3 presents segmentation metrics for different head configurations across the nucleus and retinal vessel datasets. It is important to note that the number of output channels in the first downsampling layer of the MDEU-NET model is 64, and that means the number of heads must be a divisor of 64 to ensure proper model operation. As shown in the table, the configurations MDEU2-NET, MDEU4-NET, MDEU8-NET, MDEU16-NET, and MDEU32-NET correspond to two, four, eight, sixteen, and thirty-two heads, respectively. Experimental results indicate that the model with eight heads achieves superior segmentation performance in both the nucleus and retinal vessel tasks compared with other configurations.

Table 3. Comparison of the results of cell nuclei and retinal vasculature in the MDEU-NET model and other segmentation model metrics.

Method	Params M	Flops G	RITE mIoU	RITE Acc (%)	RITE Recall (%)	PanNuKe mIoU	PanNuKe Acc (%)	PanNuKe Recall (%)
MDEU2-NET	17.62	44.11	91.87	99.15	93.12	90.91	95.97	93.51
MDEU4-NET	17.52	44.11	91.04	99.05	92.72	91.46	95.13	93.78
MDEU8-NET	17.18	44.11	92.32	99.19	93.45	91.53	96.18	93.42
MDEU16-NET	17.46	44.11	91.59	99.11	93.25	90.86	95.62	93.04
MDEU32-NET	17.44	44.11	90.71	99.03	91.72	89.49	95.86	92.31

The MDEU-NET model innovatively introduces the MDMSC multi-head attention module and the EF efficient feature fusion module on top of the U-Net architecture. To evaluate the contribution of these two modules to the model's segmentation performance, we designed an ablation experiment. Specifically, we removed each module individually and conducted segmentation experiments on the PanNuKe dataset and the RITE dataset. Each experiment employed the Adam optimizer with a learning rate of 0.0001, momentum of 0.9, and weight decay set to 1×10^{-8} . The results are presented in Tables 4 and 5, demonstrating the impact of each module on the overall segmentation performance.

Table 4. Impact of ablation experiment on segmentation performance of RITE dataset.

MDMSC	EF	mIoU	Acc (%)	Recall (%)
√	√	92.32	99.19	93.45
×	√	91.12	99.07	92.15
√	×	91.05	99.06	92.17

Table 5. Impact of ablation experiment on segmentation performance of PanNuKe dataset.

MDMSC	EF	mIoU	Acc (%)	Recall (%)
√	√	91.53	96.18	93.42
×	√	87.90	95.26	92.47
√	×	88.59	95.53	93.12

Tables 4 and 5 show the impact of the ablation experiment on the segmentation performance of the RITE dataset and the PanNuKe dataset, respectively. As can be seen from the segmentation performance results presented in the tables, the removal of either the MDMSC multi-head attention module or the EF efficient feature fusion module has an effect on the segmentation performance. To further verify the importance of these two modules, we conducted a visual analysis of the segmentation performance of the model after removing each module. These visual results reinforce the experimental evidence and also provide a more intuitive perspective for understanding the specific contribution of each component in the segmentation task. The segmentation results are shown in Figure 8.

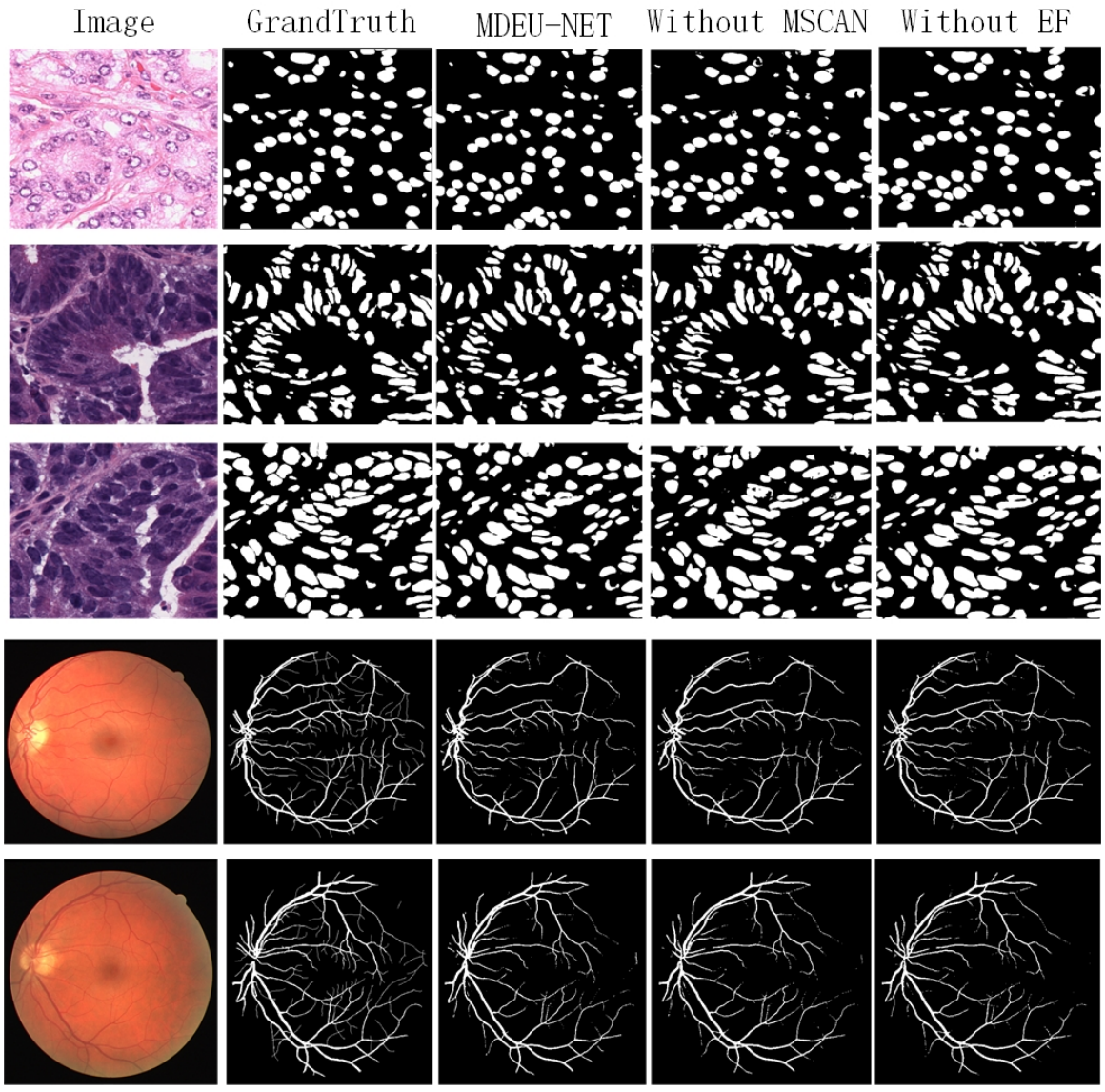

Figure 8. Visualization of the segmentation performance of different models after removing MDMSC and EF.

4. Conclusions

This paper presents MDEU-NET, a novel medical image segmentation network based on the U-Net architecture, which incorporates several optimizations and enhancements. First, a multi-head attention mechanism is employed, enabling the model to capture information from multiple perspectives simultaneously. This significantly improves the model's representational capacity and performance compared with single-head attention. Second, a multi-scale cross-axis attention mechanism is designed to address the variability of organs and lesion sites in medical images, enabling each attention head to focus on features across different scales and axes, thereby enhancing the segmentation process. Finally, the Efficient Feature Fusion (EF) module is introduced as a superior alternative to traditional feature fusion methods. This module integrates the attention mechanism with a gating module, enabling detailed information to be captured at multiple levels, thereby enhancing the emphasis on critical features and improving both feature extraction accuracy and segmentation performance. Experimental results demonstrate that the MDEU-NET model excels in segmentation tasks, with a significant improvement in recall rate and notable gains in accuracy and intersection ratio.

Author Contributions: Conceptualization, S.Y. and Y.L.; methodology, S.Y. and J.Z.; software, S.Y.; validation, S.Y., Y.L. and X.L.; formal analysis, S.Y. and J.Z.; investigation, S.Y. and P.W.; resources, S.Y. and X.G.; data curation, S.Y., P.W., X.G. and X.L.; writing—original draft preparation, S.Y.; writing—review and editing, H.C.; visualization, S.Y.; supervision, H.C.; project administration, H.C.; funding acquisition, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 12374440.

Data Availability Statement: The data is available upon request. Interested parties can contact the corresponding author at [2023303971@snnu.edu.cn] to obtain the dataset.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kaloumenou, M.; Skotadis, E.; Lagopati, N.; Efstathopoulos, E.; Tsoukalas, D. Breath analysis: A promising tool for disease diagnosis—The role of sensors. *Sensors* **2022**, *22*, 1238. [CrossRef]
2. Wu, J.; Fang, H.; Shang, F.; Yang, D.; Wang, Z.; Gao, J.; Yang, Y.; Xu, Y. SeATrans: Learning segmentation-assisted diagnosis model via transformer. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Marrakesh, Morocco, 7–11 October 2024; Springer: Berlin/Heidelberg, Germany, 2022; pp. 677–687.
3. Jiang, H.; Diao, Z.; Shi, T.; Zhou, Y.; Wang, F.; Hu, W.; Zhu, X.; Luo, S.; Tong, G.; Yao, Y.D. A review of deep learning-based multiple-lesion recognition from medical images: Classification, detection and segmentation. *Comput. Biol. Med.* **2023**, *157*, 106726. [CrossRef]
4. Gu, Y.; Wu, Q.; Tang, H.; Mai, X.; Shu, H.; Li, B.; Chen, Y. Lesam: Adapt segment anything model for medical lesion segmentation. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 6031–6041 [CrossRef]
5. Bagley, J.C.; Phillips, A.K.; Buchanon, S.; O'Neil, P.E.; Huff, E.S. Incidence and effects of anomalies and hybridization on Alabama freshwater fish index of biotic integrity results. *Environ. Monit. Assess.* **2025**, *197*, 50. [CrossRef]
6. Cohen, A.B.; Diamant, I.; Klang, E.; Amitai, M.; Greenspan, H. Automatic detection and segmentation of liver metastatic lesions on serial CT examinations. In *Medical Imaging 2014: Computer-Aided Diagnosis, Proceedings of the SPIE Medical Imaging, San Diego, CA, USA, 15–20 February 2014*; SPIE: San Diego, CA, USA, 2014; Volume 9035, pp. 327–334.
7. Kattenborn, T.; Leitloff, J.; Schiefer, F.; Hinz, S. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 24–49. [CrossRef]
8. Rajaraman, S.; Antani, S.K.; Poostchi, M.; Silamut, K.; Hossain, M.A.; Maude, R.J.; Jaeger, S.; Thoma, G.R. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ* **2018**, *6*, e4568. [CrossRef]
9. Antonelli, M.; Reinke, A.; Bakas, S.; Farahani, K.; Kopp-Schneider, A.; Landman, B.A.; Litjens, G.; Menze, B.; Ronneberger, O.; Summers, R.M.; et al. The medical segmentation decathlon. *Nat. Commun.* **2022**, *13*, 4128. [CrossRef]

10. Mazurowski, M.A.; Dong, H.; Gu, H.; Yang, J.; Konz, N.; Zhang, Y. Segment anything model for medical image analysis: An experimental study. *Med. Image Anal.* **2023**, *89*, 102918. [CrossRef]
11. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention, Proceedings of the MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015;* Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
12. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Proceedings of the 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018*; Proceedings 4; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
13. Liu, F.; Ren, X.; Zhang, Z.; Sun, X.; Zou, Y. Rethinking skip connection with layer normalization in transformers and resnets. *arXiv* **2021**, arXiv:2105.07205.
14. Oyedotun, O.K.; Shabayek, A.E.R.; Aouada, D.; Ottersten, B. Going deeper with neural networks without skip connections. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Online, 25–28 October 2020; pp. 1756–1760.
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
16. Ding, Z.; Jiang, S.; Zhao, J. Take a close look at mode collapse and vanishing gradient in GAN. In Proceedings of the 2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI), Online, 27–29 May 2022; pp. 597–602.
17. Iqbal, S.; Khan, T.M.; Naqvi, S.S.; Naveed, A.; Meijering, E. TBConvL-Net: A hybrid deep learning architecture for robust medical image segmentation. *Pattern Recognit.* **2025**, *158*, 111028. [CrossRef]
18. Song, W.; Wang, X.; Guo, Y.; Li, S.; Xia, B.; Hao, A. Centerformer: A novel cluster center enhanced transformer for unconstrained dental plaque segmentation. *IEEE Trans. Multimed.* **2024**, *26*, 10965–10978. [CrossRef]
19. Wang, G.; Ma, Q.; Li, Y.; Mao, K.; Xu, L.; Zhao, Y. A skin lesion segmentation network with edge and body fusion. *Appl. Soft Comput.* **2025**, *170*, 112683. [CrossRef]
20. Xia, J.; Cai, Z.; Heidari, A.A.; Ye, Y.; Chen, H.; Pan, Z. Enhanced moth-flame optimizer with quasi-reflection and refraction learning with application to image segmentation and medical diagnosis. *Curr. Bioinform.* **2023**, *18*, 109–142.
21. Shao, H.; Zeng, Q.; Hou, Q.; Yang, J. Mcanet: Medical image segmentation with multi-scale cross-axis attention. *arXiv* **2023**, arXiv:2312.08866.
22. Sinha, A.; Dolz, J. Multi-scale self-guided attention for medical image segmentation. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 121–130. [CrossRef]
23. Feng, Y.; Cong, Y.; Xing, S.; Wang, H.; Ren, Z.; Zhang, X. GCFormer: Multi-scale feature plays a crucial role in medical images segmentation. *Knowl. Based Syst.* **2024**, *300*, 112170. [CrossRef]
24. Lu, S.; Liu, M.; Yin, L.; Yin, Z.; Liu, X.; Zheng, W. The multi-modal fusion in visual question answering: A review of attention mechanisms. *PeerJ Comput. Sci.* **2023**, *9*, e1400. [CrossRef]
25. Biswas, S.; Gogoi, A.; Biswas, M. Aspect Ratio Approximation for Simultaneous Minimization of Cross Axis Sensitivity Along Off-Axes for High-Performance Non-invasive Inertial MEMS. In Proceedings of the International Conference on Micro/Nanoelectronics Devices, Circuits and Systems, Silchar, India, 29–31 January 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 463–469.
26. Ding, Y.; Zhang, Z.; Zhao, X.; Hong, D.; Cai, W.; Yu, C.; Yang, N.; Cai, W. Multi-feature fusion: Graph neural network and CNN combining for hyperspectral image classification. *Neurocomputing* **2022**, *501*, 246–257. [CrossRef]
27. Zhao, H.h.; Liu, H. Multiple classifiers fusion and CNN feature extraction for handwritten digits recognition. *Granul. Comput.* **2020**, *5*, 411–418. [CrossRef]
28. Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; Barnard, K. Attentional feature fusion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Online, 5–9 January 2021; pp. 3560–3569.
29. Cordonnier, J.B.; Loukas, A.; Jaggi, M. Multi-head attention: Collaborate instead of concatenate. *arXiv* **2020**, arXiv:2006.16362.
30. Ilina, O.; Ziyadinov, V.; Klenov, N.; Tereshonok, M. A survey on symmetrical neural network architectures and applications. *Symmetry* **2022**, *14*, 1391. [CrossRef]
31. Shen, K.; Guo, J.; Tan, X.; Tang, S.; Wang, R.; Bian, J. A study on relu and softmax in transformer. *arXiv* **2023**, arXiv:2302.06461.
32. Syed, A.S.; Sierra-Sosa, D.; Kumar, A.; Elmaghraby, A. A hierarchical approach to activity recognition and fall detection using wavelets and adaptive pooling. *Sensors* **2021**, *21*, 6653. [CrossRef]
33. Guo, M.H.; Lu, C.Z.; Hou, Q.; Liu, Z.; Cheng, M.M.; Hu, S.M. Segnext: Rethinking convolutional attention design for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 1140–1156.
34. Rehman, M.U.; Nizami, I.F.; Ullah, F.; Hussain, I. IQA Vision Transformed: A Survey of Transformer Architectures in Perceptual Image Quality Assessment. *IEEE Access* **2024**, *12*, 183369–183393. [CrossRef]

35. Gluth, S.; Kern, N.; Kortmann, M.; Vitali, C.L. Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nat. Hum. Behav.* **2020**, *4*, 634–645. [CrossRef]
36. Huang, H.; Zhou, P.; Li, Y.; Sun, F. A lightweight attention-based CNN model for efficient gait recognition with wearable IMU sensors. *Sensors* **2021**, *21*, 2866. [CrossRef]
37. Ruan, J.; Xie, M.; Gao, J.; Liu, T.; Fu, Y. Ege-unet: An efficient group enhanced unet for skin lesion segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Marrakesh, Morocco, 7–11 October 2024; Springer: Berlin/Heidelberg, Germany, 2023; pp. 481–490.
38. Li, X.; Qin, X.; Huang, C.; Lu, Y.; Cheng, J.; Wang, L.; Liu, O.; Shuai, J.; Yuan, C.A. SUnet: A multi-organ segmentation network based on multiple attention. *Comput. Biol. Med.* **2023**, *167*, 107596. [CrossRef]
39. Jia, Y.; Chen, G.; Chi, H. Retinal fundus image super-resolution based on generative adversarial network guided with vascular structure prior. *Sci. Rep.* **2024**, *14*, 22786. [CrossRef]