sensors

MDPI

Article

ELA-Net: An Efficient Lightweight Attention Network for Skin Lesion Segmentation

Tianyu Nie [1], Yishi Zhao [2,3] and Shihong Yao [1,*]

[1] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China; niety@cug.edu.cn
[2] School of Computer Science, China University of Geosciences, Wuhan 430074, China; zhaoyishi@cug.edu.cn
[3] Engineering Research Center of Natural Resource Information Management and Digital Twin Engineering Software, Ministry of Education, Wuhan 430074, China
* Correspondence: yaoshh@cug.edu.cn

Abstract: In clinical conditions limited by equipment, attaining lightweight skin lesion segmentation is pivotal as it facilitates the integration of the model into diverse medical devices, thereby enhancing operational efficiency. However, the lightweight design of the model may face accuracy degradation, especially when dealing with complex images such as skin lesion images with irregular regions, blurred boundaries, and oversized boundaries. To address these challenges, we propose an efficient lightweight attention network (ELANet) for the skin lesion segmentation task. In ELANet, two different attention mechanisms of the bilateral residual module (BRM) can achieve complementary information, which enhances the sensitivity to features in spatial and channel dimensions, respectively, and then multiple BRMs are stacked for efficient feature extraction of the input information. In addition, the network acquires global information and improves segmentation accuracy by putting feature maps of different scales through multi-scale attention fusion (MAF) operations. Finally, we evaluate the performance of ELANet on three publicly available datasets, ISIC2016, ISIC2017, and ISIC2018, and the experimental results show that our algorithm can achieve 89.87%, 81.85%, and 82.87% of the mIoU on the three datasets with a parametric of 0.459 M, which is an excellent balance between accuracy and lightness and is superior to many existing segmentation methods.

Keywords: skin lesion segmentation; lightweight network; attention mechanism; deep learning; medical image processing

check for updates

Citation: Nie, T.; Zhao, Y.; Yao, S. ELA-Net: An Efficient Lightweight Attention Network for Skin Lesion Segmentation. *Sensors* **2024**, *24*, 4302. https://doi.org/10.3390/s24134302

Academic Editor: Paweł Pławiak

Received: 4 June 2024
Revised: 25 June 2024
Accepted: 27 June 2024
Published: 2 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Skin cancer is a serious malignant tumor and ranks among the most prevalent forms of cancer worldwide [1]. The primary types of skin cancer include basal cell carcinoma, squamous cell carcinoma, and melanoma. Melanoma, while accounting for less than 5% of all skin cancers, has the highest malignancy and mortality rates [2]. An important consensus has appeared that early detection and treatment during screening can significantly increase the survival rate for melanoma patients, with rates reaching up to 90% [3]. Dermoscopy, a non-invasive technique, plays a pivotal role in examining and assessing skin lesions [4]. It aids in the diagnosis and treatment of skin disorders by magnifying images of the skin's surface and providing physicians with a detailed view of its fine structure. Research has demonstrated that dermoscopy can substantially enhance diagnostic accuracy [5,6]. However, as skin lesions become increasingly intricate, manual examination of dermoscopic images becomes a laborious and time-intensive endeavor. It heavily relies on the operator's expertise and dexterity, and any lapses in the process can lead to misdiagnoses or omissions. Consequently, computer-aided diagnosis (CAD) systems have been developed to elevate diagnostic accuracy and efficiency.

CAD systems have been widely used in the medical field and have become important tools in many hospitals for detecting skin diseases. Within the workflow of CAD systems,

precise segmentation of lesion regions in skin images stands as a critical element. Early segmentation techniques primarily include thresholding [7], edge detection [8], and clustering [9]. Nonetheless, these methods are highly sensitive to initial thresholds and noise, resulting in the instability of segmentation outcomes. Furthermore, their computational efficiency is relatively subpar.

With the rapid development of deep learning, current segmentation methods based on Convolutional Neural Networks (CNNs) facilitate the progress of biomedical segmentation. Most of the methods attempt to construct models with high complexity to obtain multi-scale information, thereby achieving excellent segmentation accuracy. For instance, Deeplabv3+ [10] introduces atrous convolution with different dilation rates to expand the sensory field. PSPNet [11] proposes a pyramid pooling module to capture contextual information. EncNet [12] adopts a combination of atrous convolution and multi-scale strategy to achieve excellent performance on multiple datasets. In addition, U-Net [13]-based segmentation methods, including Attention U-Net [14], U-Net++ [15], V-Net [16], and U-Net v2 [17], are developed to enhance feature propagation.

In addition, the detection of skin lesions by whole-body system images is another research idea, which is characterized by the possibility of obtaining global-scale prediction data and is suitable for large-scale screening and disease surveillance, but since whole-body system images involve skin images from different body parts and under different lighting conditions, this diversity may pose a challenge for lesion segmentation. In current research practice, it is common practice to apply machine learning methods to whole-body system images, and CNNs are widely used to process whole-body system image data [18–20]. However, researchers have found that this method can be less effective in predicting smaller lesion regions than larger lesion regions [21], so methods that utilize whole-body image detection lack fine-grained prediction ability, and in practical application, appropriate methods should be selected according to specific needs and research purposes as a way to enhance work efficiency.

The current hospitals may face limitations in terms of economic budget and hardware resources, potentially constraining their ability and willingness to deploy the high-complexity models that require substantial computational power. Therefore, it is necessary to design a lightweight segmentation model [22,23] so that it can achieve skin disease segmentation in resource-limited situations. However, due to the relatively simple structure and limited design of general lightweight models, when dealing with complex images of skin diseases (with large lesion boundaries (Figure 1a), irregular lesion areas (Figure 1b), low image contrast, making it difficult to define the lesion area and surrounding tissue boundaries (Figure 1c) and other issues), it is difficult to capture subtle edge features, thereby impacting the accuracy of segmentation tasks and resulting in potential issues like misjudgments and missed detections. Consequently, there is a critical need to design a segmentation model that is both lightweight and high-accuracy, aiming to offer more dependable support for the analysis and diagnosis of medical images in clinical settings.

In this paper, we present an efficient lightweight attention network (ELANet) for skin lesion segmentation. The core component of this network is the bilateral residual module (BRM), which forms the backbone architecture of the network by connecting a number of BRMs in series through three downsampling modules, and this structure enables efficient feature extraction from the input image with very low computational complexity. Particularly, the semantic and spatial information of the input stream is particularly important during network processing. For this reason, two attention mechanisms, channel and spatial, are embedded in the two branches of the BRM to fully capture the feature information of the image so as to enhance the segmentation accuracy of the model. In addition, we unite feature maps of different scales and adopt the strategy of multiscale fusion to achieve the multiplexing of feature maps, which can enrich the receptive domain of the network and expand the receptive field. Finally, at the end of the network, feature maps with sufficient information are processed using the pyramid pooling module (PPM),

which helps the network learn more comprehensive and detailed features and achieves improved segmentation accuracy with low computational overhead.

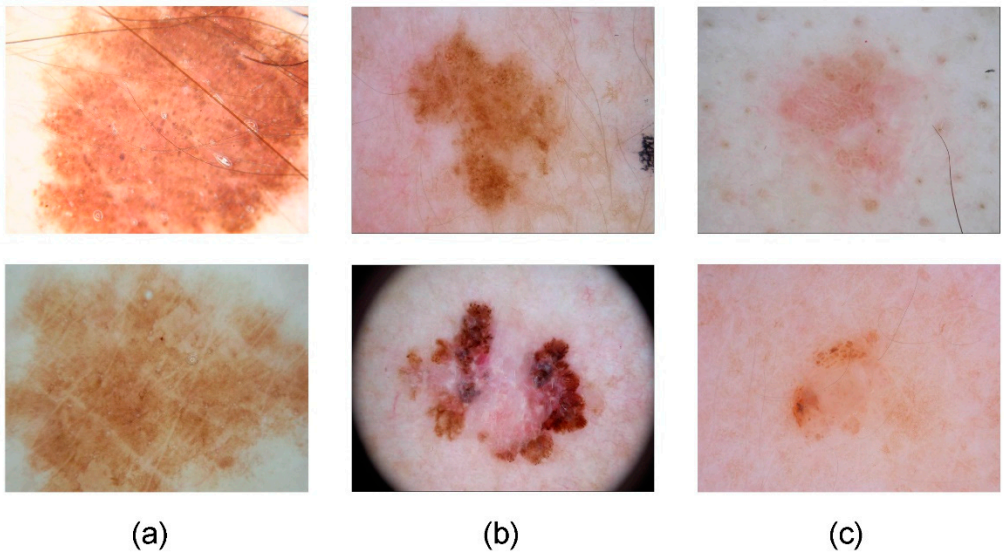

Figure 1. Some images of the public skin lesion dataset, ISIC2018. (**a**) Image of a lesion with oversized boundaries, (**b**) Image of a lesion with irregular lesion areas, (**c**) Image of a lesion with blurred boundaries.

In summary, the innovation of our work lies in the design of a novel module (BRM) for extracting skin lesion image features and combining it with a multi-scale strategy to construct a segmentation model, ELANet, which can help physicians diagnose skin lesions more accurately and improve the efficiency and accuracy of treatment.

The main contributions of this paper can be summarized as follows:

1. We propose a lightweight segmentation model, ELANet, which achieves efficient skin lesion segmentation with an extremely low parameter, facilitating its deployment in resource-constrained clinical devices.
2. We design the BRM for seamless integration into the model. Its internal atrous convolution and attention mechanisms comprehensively capture feature information, thereby enhancing the model's segmentation accuracy.
3. We construct a multi-scale attention fusion (MAF) module that combines output streams of various scales, effectively solving the information loss caused by down sampling operations and obtaining global context information.
4. We conduct extensive experiments on three public datasets, ISIC2016, ISIC2017, and ISIC2018, achieving state-of-the-art performance and attaining an excellent balance between accuracy and lightness.

2. Related Work

In this section, we provide an overview of the challenges faced in the field of skin lesion segmentation as well as related research advances. The main focus is on three key aspects: methods for skin lesion segmentation, lightweight networks, and attention mechanisms. Related work in these three areas is described below.

2.1. Skin Lesion Segmentation

Currently, CNN-based models have been widely used in the field of dermatological segmentation. Inspired by the fully convolutional network (FCN) [24], Yu et al. [25] first devised the fully convolutional residual network (FCRN), enhancing segmentation performance through the integration of multi-scale contextual information. Bi et al. [26] then proposed an automatic skin damage segmentation method based on FCN architecture and

introduced deep class-specific learning to overcome the feature ambiguity. Subsequently, Esfahani et al. [27] developed DPFCN, featuring a novel dense pooling layer that achieves high-accuracy segmentation of skin images. Tang et al. [28] proposed a method for skin lesion segmentation (SLS) utilizing the separable-UNet architecture with stochastic weighted averaging, aiming to improve pixel-level discriminative representation. Arora et al. [29] developed a custom U-Net architecture with Attention Gates-Based design, incorporating atrous convolution to expand the receptive field. Wu et al. [30] introduced a network based on a dual encoder architecture, incorporating the adaptive dual attention module (ADAM), for automated skin lesion segmentation from dermatoscopic images. Abhishek et al. [31] designed a deep semantic segmentation framework for dermatoscopic images by augmenting the RGB dermatoscopic image with additional color bands and intrinsic, greyscale, and shadow-attenuated images.

However, the above approaches focus on segmentation accuracy while ignoring the variations in parameter quantity, resulting in large parameters in the model, which is detrimental in clinical conditions where lightweight deployment is required. Hence, our algorithms focus on lightweight designs that reduce the number of model parameters to enable deployment in low-memory practical applications.

2.2. Lightweight Networks

The utilization of complex model architectures tends to put a considerable load on computing devices. To tackle this challenge, researchers have initiated investigations into the design and optimization of lightweight networks. Enet [32], introduced by Adam et al., emerged as one of the pioneering networks tailored for lightweight real-time segmentation. Its compact structure is well-suited for deployment on embedded devices. Building upon Enet, Romera et al. [33] proposed ErfNet, which incorporates residual concatenation and factorization convolution, yielding more precise outputs. Recently, Jeya et al. [34] proposed a lightweight network called UNext, which combines the MLP and UNet structures to reduce the number of model parameters while ensuring accuracy. In addition, MALUNet [35], designed by Ruan et al., is also based on the U-shaped structure and introduces various attention mechanisms to reduce the model size, achieving excellent performance in the skin lesion segmentation task. They also proposed another lightweight model, EGEUNet [36], which uses the grouping idea to capture information from different perspectives, and the model solves the skin lesion segmentation task with only 50 KB parameters.

The common approach to lightweight modeling involves reducing the network's depth or the number of neurons per layer. However, such simplifications can hinder the model's capacity to fully capture essential information, thereby affecting the model's predictive ability. To address this issue, we introduce an attention mechanism that enhances the model's ability to perceive features without increasing additional computational overhead, thereby improving overall performance. It meets the dual requirements of accuracy and lightness in practical application environments.

2.3. Attention Mechanisms

Attention mechanisms have proven to be useful tools for neural networks, facilitating the learning of essential information and enhancing a model's predictive capabilities. Wang et al. [37] introduced the residual attention network, which analyzes image features by stacking attention blocks. Jaderberg et al. [38] devised the spatial converter, enabling the network to operate on feature maps in the spatial domain. On the other hand, Hu et al. [39] focused on channel relations and proposed Squeeze and Excitation (SE) blocks to re-model the interdependence between channels. However, both spatial and channel attention omit some feature information; therefore, Sanghyun et al. [40] combined these two dimensions and proposed CBAM, which sequentially infers the attention graph along two independent dimensions (channel and spatial) to achieve a more comprehensive resource allocation. In addition, Kaul et al. [41] also designed FocusNet through a hybrid attention mechanism, which achieved good results in the fields of skin cancer and lung lesion segmentation.

Across various levels of neural networks, there are differences in the spatial and semantic information contained in feature maps, which provide the spatial position and category information of the target, respectively. However, embedding only one type of attention mechanism will ignore another type of feature information, which cannot meet global needs. Therefore, in this paper, we design a hybrid attention mechanism that concurrently incorporates both spatial and channel attention mechanisms. This innovative approach enables the model to pinpoint the target's position and capture its category features simultaneously, thereby enhancing the model's performance in skin lesion segmentation tasks.

3. Methods

In this section, we will first describe the overall architecture of ELANet and its three core components, including the downsampling module (DSM), the pyramid pooling module (PPM), and the feature fusion module (FFM). Then the multi-scale attention fusion operations used by the network are described, after which the structure of the core module BRM will be given.

3.1. ELANet Architecture

The network's overall structure is depicted in Figure 2, and it is divided into three primary layers. Each of these layers conducts feature processing and aggregation through a sequence of modules. The processing details for each layer are elaborated upon below:

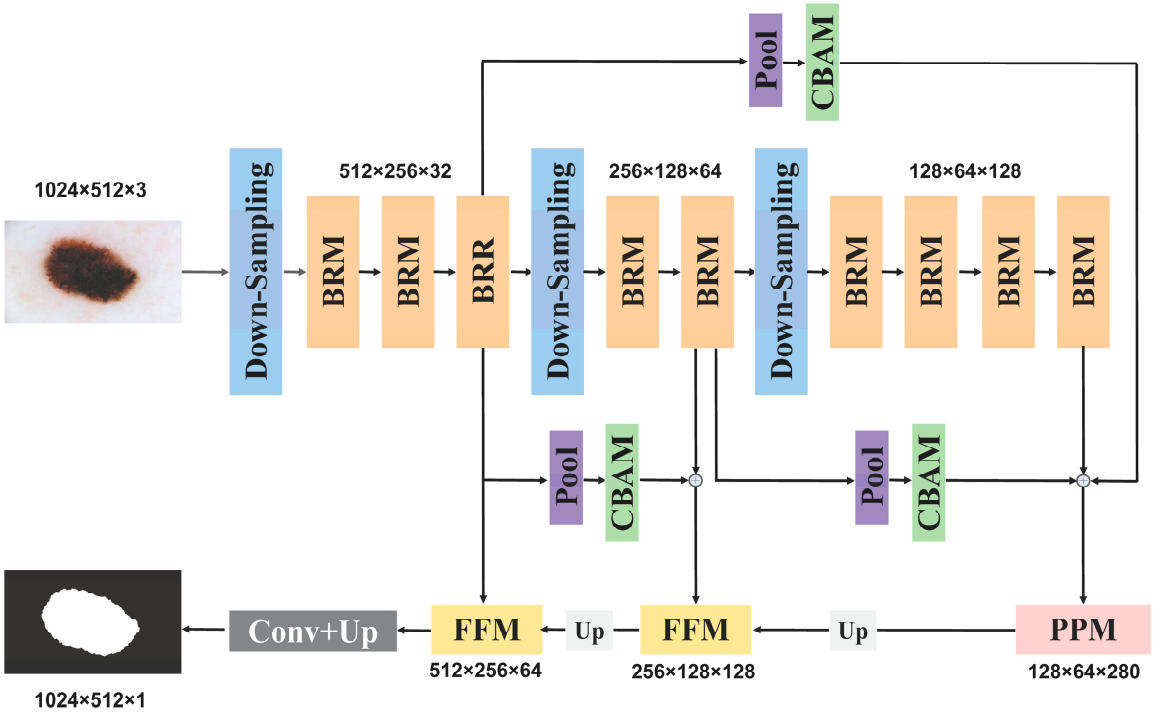

Figure 2. The overall architecture of ELANet.

First Level: At this level, the input image undergoes processing through a downsampling module, reducing its size by half and increasing the channel count to 32. This results in the formation of a feature map with dimensions of 512 × 256 × 32. Then, after three bilateral residual modules, the integration of features is achieved. During the integration process, the size of the feature map and the number of channels are kept constant.

Second Level: The feature map is further downsampled to half its previous size, simultaneously doubling the number of channels. This results in the creation of a feature map with dimensions of 256 × 128 × 64.

Third Layer: The third layer features a dimension of 128 × 64 × 128 for its feature map. Each of these three differently sized feature maps contains ample global contextual

information. To maximize the utility of this information, we employ attention mechanisms to aggregate it, ensuring their effective guidance of feature information during the decoding process.

Finally, aggregate the outputs from the three levels and feed them into the pyramid pooling module to further expand the sensory field. While restoring the feature map size, the feature aggregation module is used, which can merge high-level semantic and spatial information together. Finally, after specific processing, the network outputs the prediction results.

3.1.1. Downsampling Module

The downsampling module plays a crucial role in resizing the input graph while augmenting its channel count. This process aims to enhance feature representation and decrease computational complexity, as shown in Figure 3a. For the input feature graph $F \in R^{C_{in} \times H \times W}$, is split into two branches. One branch employs maximum pooling to downsample the size to half, resulting in $F' \in R^{C_{in} \times H/2 \times W/2}$. The other branch employs 3×3 convolution with a stride of two while controlling the number of output channels as $(C_{out} - C_{in})$, producing the output $F'' \in R^{(C_{out} - C_{in}) \times H/2 \times W/2}$ (where C_{in} and C_{out} represent the number of input and output channels, H and W represent the width and height of the feature map). Afterward, they are merged to get the output using the activation function. Compared with directly using the pooling layer to perform the downsampling operation, this can supplement the feature information and does not add too much computational burden. The specific formula is as follows:

$$DSM(F) = \rho(Concat(C_{3\times3}(F), MaxP(F))), \quad (1)$$

where $DSM(F)$ denotes the output of the feature map F after the downsampling module, ρ stands for the BN+ReLU operation, BN is the batch normalization, and ReLU is the nonlinear activation function, while $C_{3\times3}$ denotes the 3×3 convolution, and $MaxP$ is the maximum pooling.

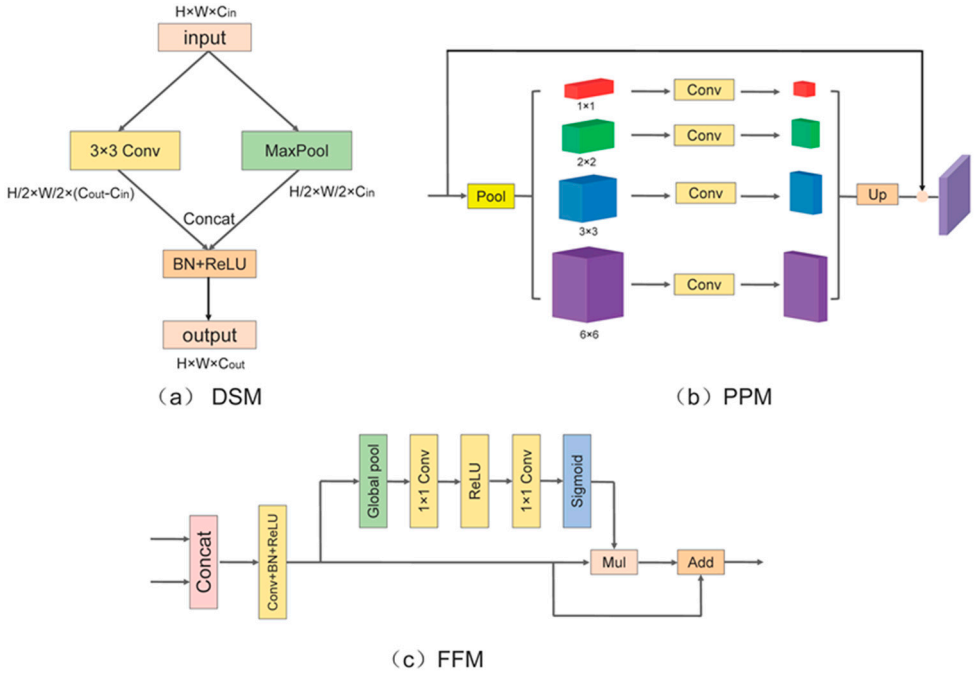

Figure 3. Three modules used in ELANet. (**a**) downsampling module (DSM), (**b**) pyramid pooling module (PPM), (**c**) feature fusion module (FFM).

3.1.2. Pyramid Pooling Module

The PPM serves as the core component of PSPNet [11]. It operates by decomposing the input information into four feature maps with sizes of 1×1, 2×2, 3×3, and 6×6 following an adaptive pooling operation. Subsequently, they are respectively adjusted to the original $1/4$ by a single 1×1 convolution, after which they are up-sampled to the initial sizes by bilinear interpolation, summed up, and finally summed up with the residuals to obtain the final aggregated features. The PPM is able to acquire the feature information of different scales and regions and fuse them together, which helps to enhance the model's understanding of the global contextual information, thus improving the performance of image segmentation, as shown in Figure 3b.

3.1.3. Feature Fusion Module

Feature fusion is a common processing technique, but to enhance the depth of interaction between the two input features, we employ a more comprehensive and effective feature aggregation method [42]. This approach leverages global information to guide the fused module in suppressing irrelevant feature details while improving its capacity to effectively utilize information compared with traditional methods. In Figure 3c, it combines the two given inputs and adjusts the channel count to the desired number using a 1×1 convolution. Subsequently, global pooling is applied to incorporate global information, enhancing the network's sensitivity to features. Following the concept of squeeze and excitation [39], this result is then dot-multiplied with the inputs to recalibrate the response information. Finally, it is summed with the initial signal to produce the aggregated output.

3.2. Multiscale Attention Fusion

Multi-scale fusion is increasingly prevalent in detection and segmentation tasks. For lightweight networks with fewer layers, effectively utilizing feature maps at various scales is crucial for improving segmentation performance. However, simple summation might not facilitate efficient information interaction. To convey deep features optimally, we incorporate the lightweight attention mechanism CBAM [40] into the fusion process. This modification allows us to emphasize or suppress specific feature information without introducing additional computational overhead. The module comprises spatial attention and channel attention mechanisms, as outlined below:

Channel Attention Module: This module is designed to guide the network's attention to specific channel-wise feature information. It takes the input feature map and performs average pooling and maximum pooling operations to reduce the feature map's spatial dimensions. The maximum pooling operation identifies the most significant features in each channel, while the average pooling operation computes the average feature value for each channel. These compressed feature maps are then passed through a multilayer perceptron (MLP) for further processing, allowing the network to learn complex relationships among input features. Finally, the outputs from these operations are fused and passed through an activation function. This module amplifies the model's response to task-critical features, as described below:

$$CA(F) = \sigma(\mathrm{MLP}(AvgP(F)) + \mathrm{MLP}(MaxP(F))) \tag{2}$$

where $CA(F)$ denotes the output of input F through the channel attention module, $AvgP$ and $MaxP$ stand for average and maximum pooling, respectively, and σ denotes the sigmoid function.

Spatial Attention Module: This module is designed to assist the model in adapting to variations in feature importance across different spatial locations. Unlike the channel attention module, it sequentially performs average pooling and maximum pooling operations along the channel axis. Subsequently, the module applies a single 7×7 convolutional layer, which is capable of learning to identify feature distinctions in various spatial locations, generating attention feature maps. Finally, the results are obtained through an activa-

tion function. This module enhances the network's emphasis on critical regions within the image, thereby enhancing the model's performance and robustness. The formula is as follows:

$$SA(F) = \sigma(C_{7\times7}([AvgP(F); MaxP(F)])), \tag{3}$$

where $SA(F)$ denotes the output of input F after the spatial attention module, and $C_{7\times7}$ denotes the 7×7 convolution.

In this paper, the CBAM used in the multiscale attention aggregation section computes the attention map sequentially along both channel and spatial dimensions, which to some extent solves the problem of balancing the model between perceptual details and global context. The specific formula is as follows:

$$\begin{aligned} F' &= CA(F) \otimes F, \\ F'' &= SA(F') \otimes F', \end{aligned} \tag{4}$$

where $\otimes$ denotes element-by-element multiplication, CA and SA denote channel and spatial attention, respectively, F is the initial input feature map, F' is the output feature map of the first step after channel attention, and F'' is the final output result after CBAM.

3.3. Bilateral Residual Module

Lightweight networks must deliver efficient performance with a limited number of layers; an excessively deep backbone structure is not suitable. Therefore, the key to building such a model lies in designing efficient and lightweight components. The split-shuffle-non-bottleneck (SS-nbt) [43] lacks sufficient bilateral information interactions and has limitations in capturing information from various locations in the image. On the other hand, the efficient asymmetric residual (EAR) module [44] and bottleneck residual unit (BRU) [45] incorporate channel attention mechanisms and convolutional blocks at multiple locations, resulting in a more complex overall structure. Additionally, the repetitive use of a single attention mechanism introduces redundancy as the network deepens, potentially affecting feature selection and concentration. To address these concerns, we propose the bilateral residual module (Figure 4) to capture image features more effectively.

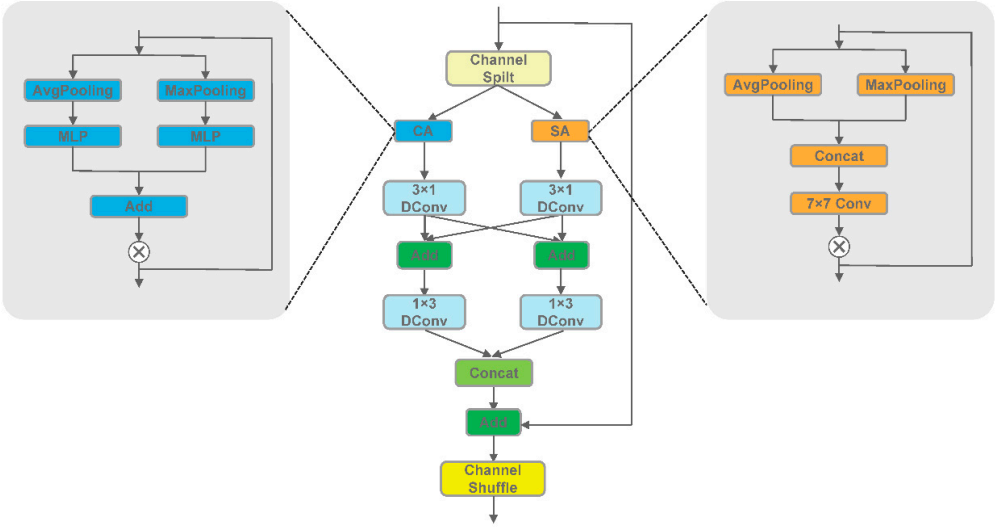

Figure 4. Bilateral Residual Module (BRM).

Specifically, when an input feature map $F_{in} \in R^{C\times H\times W}$ is passed into the bilateral residual module, it undergoes an initial channel separation operation. This operation,

designed to be more computationally efficient and memory-saving compared with standard convolution-based channel separation, is expressed as follows:

$$x_1, x_2 = SP(F_{in}), \tag{5}$$

where x_1, and x_2 represent the two results output after channel separation, and SP represents the channel split operation.

In the next process, each of the two branches will pass through the attention mechanism once. It should be noted that the selection of the attention module is not fixed, but two different attention mechanisms are assigned to the left and right branches, i.e., channel attention and spatial attention. The channel attention mechanism enhances fine-grained feature selection by effectively capturing the correlations and importance between different channels within the image. Meanwhile, the spatial attention mechanism aids the network in concentrating on information from various locations in the image, allowing it to better capture spatial, structural, and contextual information. This combination of dual attention mechanisms empowers the model to gain a more comprehensive and accurate understanding of the image content, avoiding the redundant utilization of a single module at different levels. Such a design strategy can focus on the important features in the image more accurately, thus improving segmentation performance. The specific formula is as follows:

$$y_{att1} = CA(x_1), \tag{6}$$

$$y_{att2} = SA(x_2), \tag{7}$$

where y_{att1} and y_{att2} represent the outputs of the left and right branches after the attention mechanism, CA represents the channel attention mechanism, and SA represents the spatial attention mechanism.

Then the sensory field is expanded with null convolution to obtain feature information at long distances. To achieve a more powerful feature extraction capability while maintaining the number of model parameters, an asymmetric convolution strategy is used as an alternative to the traditional 3×3 convolution. However, since two branches used in different network layers may suffer from insufficient information sharing, an information-sharing operation is performed between the decomposed convolutional layers to achieve complementarity of features from different branches. This approach is effective in inducing features from different branches to merge to some extent, thus achieving a more integrated feature representation. After the previous steps, we combine the results of the two branches, which are residually summed with the original input. Finally, a channel shuffle operation is performed to obtain the output of the bilateral residual module. The specific formula for the above operation is as follows:

$$y_1 = DC_{1\times3,l}(DC_{3\times1,l}(y_{att1}) + DC_{3\times1,r}(y_{att2})), \tag{8}$$

$$y_2 = DC_{1\times3,r}(DC_{3\times1,r}(y_{att2}) + DC_{3\times1,l}(y_{att1})), \tag{9}$$

$$F_{out} = SH(Concat(y_1, y_2) + F_{in}), \tag{10}$$

where y_1, y_2 represent the outputs of the two branches, $DC_{m\times n,l}$ and $DC_{m\times n,r}$ refer to the $m \times n$ atrous convolution in the left and right branches, respectively. F_{out} represents the final output feature maps of the bilateral residual module, SH stands for the channel shuffle operation, and $Concat$ is the stagewise concatenation operation.

4. Experiment

4.1. Dataset

We select the public skin lesion datasets ISIC2016 [46], ISIC2017 [47], and ISIC2018 [48,49] published by the International Skin Imaging Collaboration (ISIC) to evaluate the model, and the thyroid nodule segmentation dataset TN3K for evaluating the model generalization capability. The specific details of each dataset are shown in Table 1.

Table 1. Detailed description of the three skin lesion datasets ISIC2016, ISIC2017, and ISIC2018, and the thyroid nodule dataset TN3K.

Dataset	Training	Validation	Testing
ISIC2016	900	/	379
ISIC2017	2000	150	600
ISIC2018	2594	100	1000
TN3K	2879	/	614

4.2. Evaluation Metrics

We choose two common metrics to assess the model's performance: mean intersection-over-union (mIoU) and accuracy (Acc). mIoU quantifies the similarity between the predicted results and the ground truth, while Acc is computed as the ratio of correctly predicted samples to the total predicted samples. The formulas for these metrics are as follows:

$$\text{mIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{TP+FP+FN}, \tag{11}$$

$$\text{Acc} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{12}$$

where TP stands for the number of true positives, FP stands for the number of false positives, FN stands for the number of false negatives, and TN stands for the number of true negatives, all of which are parameters in the confusion matrix. k is the number of the foreground.

4.3. Implementation and Configuration

In the experimental process, we employ the BCE (Binary Cross-Entropy) loss [50] function for training. The optimization is performed using stochastic gradient descent (SGD) as the optimizer. We initialize the learning rate at 0.01 with a momentum parameter of 0.9. To mitigate overfitting and enhance generalization, a weight decay strategy with a decay coefficient of 0.0001 is employed. Additionally, a cosine annealing technique for learning rate reduction is utilized to guide the training process toward optimal convergence. The training epoch is set equal to 1000, and the batch size is 8.

To further improve the model's generalization performance and reduce the risk of overfitting, we introduce various data augmentation techniques, including random flipping, Gaussian blurring, and image rotation. Moreover, to maintain a consistent image size and prevent distortion caused by augmentation operations, images are padded with a gray border and resized to a uniform input size of 1024 × 512 pixels. It is worth noting that all experiments are conducted within the PyTorch framework and are trained using an NVIDIA GeForce RTX 3090 Ti graphics card with 24 GB of memory (NVIDIA, Santa Clara, CA, USA).

4.4. Comparisons with State-of-the-Art Methods

To verify the performance of ELANet, we conducted several comparative experiments on the ISIC2016, ISIC2017, and ISIC2018 datasets, respectively, and the specific experimental data are displayed in Table 2. The comparison experiments include some representative medical segmentation algorithms, such as the most classical UNet [13], which is characterized by a compact and symmetric encoder-decoder structure and guarantees the transfer of feature information through jump connections; TransUNet [51] and SwinUNet [52], which combine the ideas of Transformer and CNN in order to better understand the medical image's semantic information; and the lightweight medical segmentation models MALUNet [35] and EGEUNet [36], which improve the accuracy of the models in understanding and analyzing images by combining the attention mechanism as well as the grouping idea.

Table 2. Comparison with the state-of-the-art methods on three skin lesion datasets.

Method	ISIC2016		ISIC2017		ISIC2018		Params (M)	FLOPs (G)
	mIoU (%)	Acc (%)	mIoU (%)	Acc (%)	mIoU (%)	Acc (%)		
UNet [13]	85.56	95.03	80.86	91.40	82.54	91.95	24.891	225
PSPNet [11]	86.25	95.62	72.51	87.85	76.89	89.41	46.602	357
DeeplabV3+ [10]	85.76	95.52	76.02	89.48	74.82	88.20	41.216	353
CGNet [53]	86.25	95.30	82.02	91.86	76.88	89.17	0.492	6.896
BiseNetV2 [54]	87.45	96.08	79.08	90.69	79.68	90.84	3.341	24.571
Fastscnn [55]	88.61	96.36	80.41	91.32	80.25	90.94	1.398	1.853
ICNet [56]	88.62	96.27	80.16	91.27	77.92	89.52	47.528	30.745
ErfNet [33]	89.76	96.71	81.82	91.96	81.25	91.42	2.082	29.062
HRNet [57]	89.60	96.69	79.43	90.94	78.98	90.45	9.636	37.157
MobileNetV3 [58]	**90.12**	**96.86**	79.80	91.96	81.20	91.55	3.282	8.687
Segmenter [59]	82.60	94.00	76.67	89.32	75.52	87.80	6.685	4.722
Segformer [60]	87.40	95.80	79.11	90.79	80.74	91.05	3.716	3.682
TransUNet [51]	81.45	93.57	78.43	90.23	77.36	89.65	66.815	32.63
SwinUNet [52]	85.29	94.95	81.84	91.88	82.86	**92.21**	27.145	5.91
UNext [34]	83.77	94.44	**82.28**	**92.15**	81.54	91.45	1.471	0.439
MALUNet [35]	83.30	94.11	80.93	91.49	82.65	91.96	0.175	0.083
EGEUNet [36]	81.33	93.21	81.02	91.46	81.51	91.55	**0.053**	**0.072**
ELANet	89.87	96.78	81.85	91.99	**82.87**	92.19	0.459	8.430

Note: bold indicates best performance.

Although MobileNetV3 and UNext have the highest accuracy on the ISIC2016 and ISIC2017 datasets, respectively, when combining the three datasets, they are less accurate on the other datasets, and the models are not stable enough, whereas ELANet excels on all three datasets, with a better average accuracy than the other models. In addition, our model parameter is 0.45 M, which puts very little storage pressure on the device in real deployment. Through the visual analysis in Figure 5, we find that most models have higher parameter counts than ELANet, and the accuracy of models with fewer parameters is lower than that of ELANet; thus, our approach strikes an excellent balance between accuracy and lightweightness.

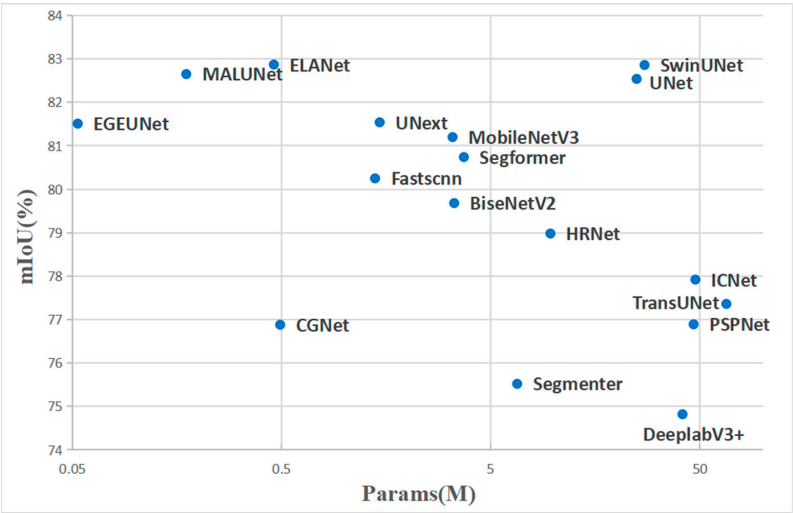

Figure 5. Accuracy-Params comparisons on the skin lesion datasets ISIC2018.

In Figure 6, the final results of the above segmentation models on the three datasets are visualized. Looking to see that the models' prediction accuracy for ISIC2016 is generally higher than that of ISIC2017 and ISIC2018, we speculate that it may be because there is less

training data in the ISIC2016 dataset, which leads to easier fitting of the models. Whereas in ISIC2017 and ISIC2018, the accuracy of each model is closer, it may be that when the data features are increased to a certain level, the prediction performance of the model also tends to stabilize.

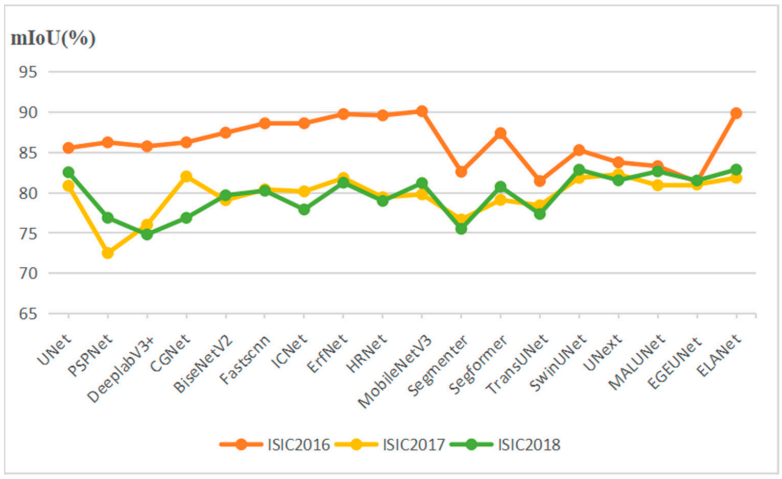

Figure 6. The line charts show the segmentation performance of the different models on the three skin lesion datasets.

In addition to the numerical comparisons, we show in Figure 7 a visualization image of the prediction results for each model on the ISIC2018 dataset. The first two columns display the input images and their corresponding ground truth labels, while the subsequent columns showcase the predicted images from different models. These visualizations highlight ELANet's exceptional segmentation capabilities, particularly for skin lesion images with oversized lesion boundaries, blurred boundaries, and irregular shapes. It is clear that ELANet not only identifies more lesion areas compared with other models but also excels at handling edge details. These results further demonstrate the effectiveness of ELANet in the task of skin lesion segmentation, and its superior performance relies on the interplay of the various modules in the ELANet structure, which helps the model to cope with various complex tasks.

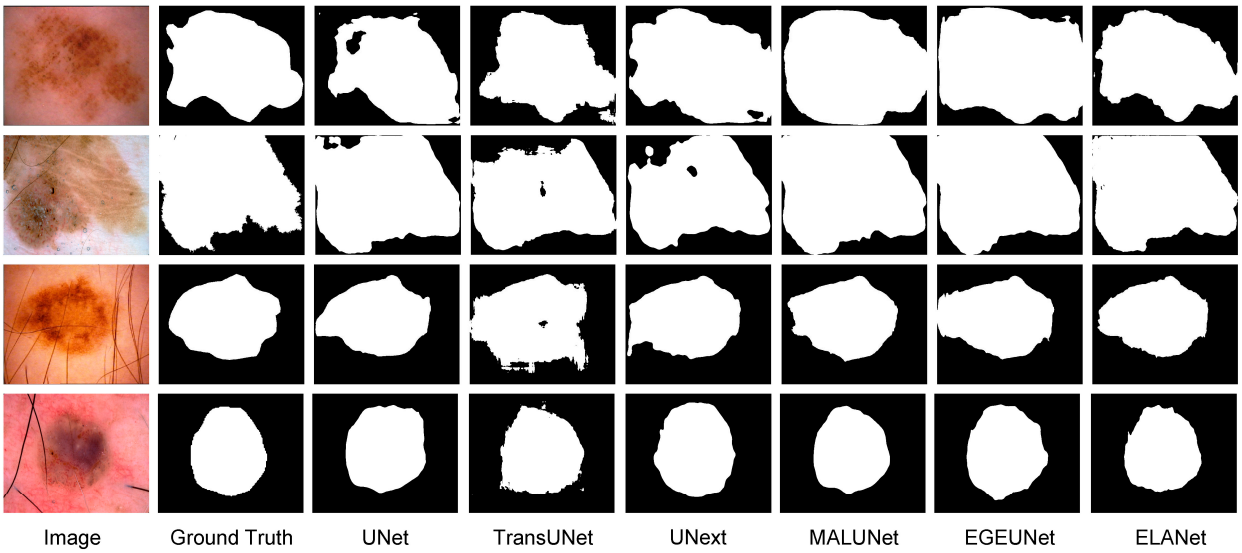

Figure 7. Visual comparisons with different state-of-the-art methods on ISIC2018.

4.5. Ablation Experiments

To verify that the modules used in the model are valid, we conduct ablation experiments on the ISIC2016 dataset, which are divided into three parts: choice of attentional mechanism in the bilateral residual module, multi-scale attention fusion, and pyramid pooling module. The experimental environments and dataset distributions used by these networks are identical to ensure fairness in the experiments.

Attention Mechanisms: Three ways of equipping all spatial attention, all channel attention, and one attention mechanism each for left and right in both branches of the BRM. As seen in Model 1-Model 3 of Table 3, the segmentation accuracy of only configuring the spatial attention mechanism is the lowest, followed by the way of only equipping the channel attention, and the highest is the method of the hybrid attention mechanism, with the mIoU reaching 89.18%, which is improved by 3.55% compared with the lowest way. Importantly, this enhanced performance is achieved with minimal impact on the number of model parameters and computational requirements, highlighting the reasonability and effectiveness of this configuration.

Table 3. Ablation experiments on the ISIC2016 dataset.

Method	MAF	PPM	Attention			mIoU (%)	Acc (%)	Params (M)	FLOPs (G)	Time (S/Image)
			SA	CA	HA					
Model 1	×	×	√	×	×	85.63	95.10	**0.270**	6.754	0.0331
Model 2	×	×	×	√	×	87.84	96.06	0.273	**6.679**	**0.0328**
Model 3	×	×	×	×	√	89.18	96.51	0.271	6.716	0.0337
Model 4	√	×	×	×	√	89.24	96.56	0.289	7.260	0.0362
Model 5	×	√	×	×	√	89.29	96.56	0.329	7.158	0.0354
Model 6 (ELANet)	√	√	×	×	√	**89.87**	**96.78**	0.459	8.430	0.0417

Note: MAF: Multi-scale Attention Fusion, PPM: Pyramid Pooling Module, Attention: Attention Mechanisms for Bilateral Residual Modules, SA: Spatial Attention, CA: Channel Attention, HA: Hybrid Attention, M: Million, G: Giga-Billion, bold indicates best performance.

Multiscale Attention Fusion: After determining the choice of hybrid attention, we add a multiscale attention fusion method to the model. As shown in Model 4 in Table 3, the segmentation performance of the model has a small improvement, and the number of parameters is only 0.018 M more, which is an acceptable range.

Pyramid Pooling Module: As shown in Model 5 in Table 3, after we introduce the PPM. The segmentation precision and accuracy are improved, while the number of parameters and the computation volume are only increased by 0.058 M and 0.442 G.

Finally, in our study, the introduction of MAF and PPM, combined with the hybrid attention mechanism (Model 6 shown in Table 3), led to a notable improvement in segmentation accuracy, achieving a mIoU of 89.87% and an accuracy of 96.78%. Also, this improvement is accomplished without a substantial increase in the number of model parameters or computational requirements, thus meeting the accuracy and lightness criteria for the model. This underscores the effectiveness of selecting the hybrid attention mechanism, which allows for the acquisition of more feature information without significantly adding to the computational burden. Additionally, the introduction of the MAF and PPM, while resulting in a modest increase in parameters and computation, brought substantial performance enhancements to the model. These modules have proven to be effective and viable additions, significantly enhancing the skin lesion segmentation task.

In Figure 8, visualizations of each model in the ablation experiments are presented. It is evident that in Model 1, when the spatial attention mechanism is selected for the bilateral residual module, the segmentation performance is poor, especially in recognizing target irregularities and in darker environments. In Model 2, after switching to the channel attention mechanism, the prediction ability is slightly improved, but it still struggles with challenging tasks in average lighting conditions.

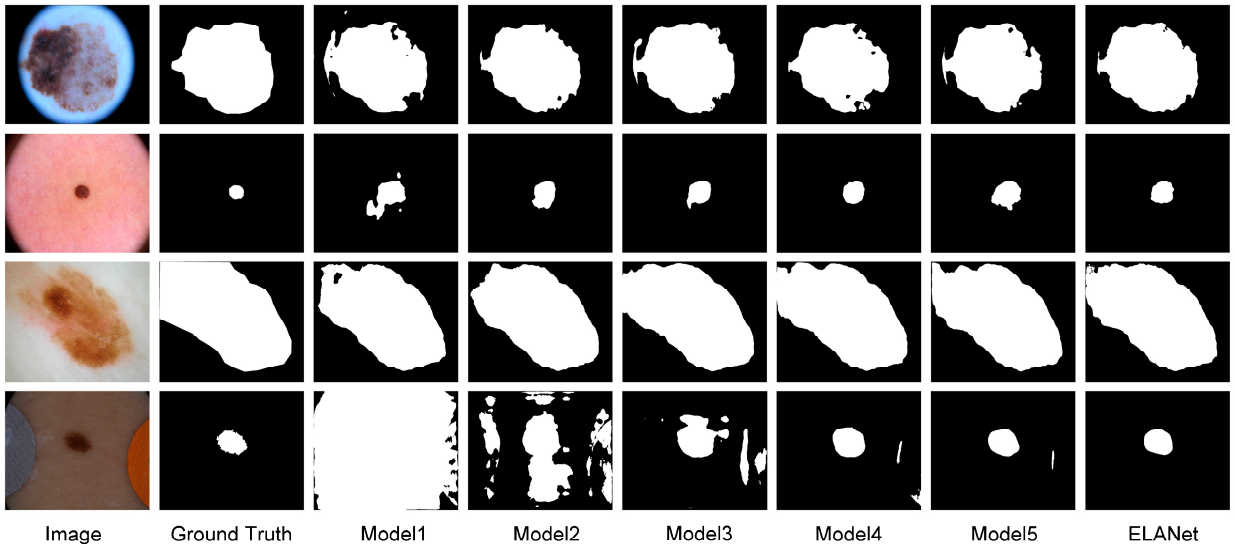

Figure 8. Visualisation comparing ablation experiments in the ISIC2016 dataset.

Model 3, equipped with the hybrid attention mechanism where both types of attention mechanisms work together, enhances the model's ability to locate the spatial position of the target compared with the method using only one attention module, but there's still room for improvement in capturing detailed information. By further adding the MAF and PPM in Models 4 and 5, respectively, the model's ability to process edge detail information is expanded, and it is observed that both methods have a slight improvement in performance.

The final Model 6 (ELANet), which adds both MAF and PPM, exhibits excellent segmentation performance, even for irregular images with blurred edges, and recognition in dark environments. This indicates that our approach effectively addresses these complexities and achieves outstanding performance.

4.6. Generalization Experiment

To assess the generalization ability of the proposed method, we finally performed generalization experiments on the TN3K (thyroid nodule segmentation dataset), the results of which are shown in Table 4. From the experimental data, we can see that ELANet does not achieve the highest accuracy. We analyze the reason for this, which is due to the fact that the skin lesion dataset and the thyroid nodule dataset have different data features and distributions, and there are large morphological differences between the lesion regions of the images, thus leading to differences in the performance of the models on the different datasets. But it performs better compared with most models. This proves that ELANet has excellent generalization ability when dealing with different types of medical image tasks, showing its good potential and stability in the field of medical imaging.

Table 4. Generalization experiments on the thyroid nodule dataset.

Method	mIoU (%)	Acc (%)	Params (M)	FLOPs (G)
UNet [13]	84.94	96.21	24.891	225
CGNet [53]	83.17	93.63	0.492	6.896
MobileNetV3 [58]	82.87	95.31	3.282	8.687
Segformer [60]	79.10	84.79	3.716	3.682
TransUNet [51]	**86.86**	96.47	66.815	32.63
SwinUNet [52]	80.57	94.96	27.145	5.91
UNext [34]	86.64	**96.63**	1.471	0.439
MALUNet [35]	82.99	95.65	0.175	0.083

Table 4. *Cont.*

Method	mIoU (%)	Acc (%)	Params (M)	FLOPs (G)
EGEUNet [36]	83.14	95.75	**0.053**	**0.072**
ELANet	85.42	96.30	0.459	8.430

Note: bold indicates best performance.

5. Conclusions

In this paper, we propose an efficient, lightweight attention network for skin lesion segmentation called ELANet. The BRM used in the network can fully extract the image features, and by embedding two different attentional mechanisms into the BRM, two different kinds of detailed information can be generated to complement each other in order to enhance the segmentation performance of the model. Additionally, ELANet employs a MAF strategy to fuse features of various scales, resulting in a global contextual understanding. The use of asymmetric and atrous convolutions helps to reduce parameters while expanding the receptive field, making ELANet suitable for clinical environments with limited computational resources. Finally, the comparison experiments on three public skin lesion datasets show that ELANet can achieve the most excellent segmentation performance at very low parameter counts, outperforming other common segmentation models. ELANet also demonstrates competence in handling complex segmentation tasks, such as those involving irregular regions, blurred boundaries, and oversized boundaries, with great potential to be a good aid to healthcare professionals for accurate diagnosis.

In the future, optimizing the structure of ELANet and improving the generalization ability of the model so that it can be applied to other medical image processing fields are the main challenges.

Author Contributions: T.N.: Conceptualization, Methodology, Software, Validation, Writing—original draft, Visualization. Y.Z.: Funding acquisition. S.Y.: Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: The research work reported in this paper is supported by the National Natural Science Foundation of China under Grant No. 62072342.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2019. *CA Cancer J. Clin.* **2019**, *69*, 7–34. [CrossRef] [PubMed]
2. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
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*; Springer: Cham, Switzerland, 2017; Volume 20, pp. 250–258.
4. Kharazmi, P.; AlJasser, M.I.; Lui, H.; Wang, Z.J.; Lee, T.K. Automated detection and segmentation of vascular structures of skin lesions seen in Dermoscopy, with an application to basal cell carcinoma classification. *IEEE J. Biomed. Health Inform.* **2016**, *21*, 1675–1684. [CrossRef] [PubMed]
5. Siegel, R.L.; Miller, K.D.; Goding Sauer, A.; Fedewa, S.A.; Butterly, L.F.; Anderson, J.C.; Cercek, A.; Smith, R.A.; Jemal, A. Colorectal cancer statistics, 2020. *CA A Cancer J. Clin.* **2020**, *70*, 145–164. [CrossRef] [PubMed]
6. Vestergaard, M.E.; Macaskill, P.H.P.M.; Holt, P.E.; Menzies, S.W. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: A meta-analysis of studies performed in a clinical setting. *Br. J. Dermatol.* **2008**, *159*, 669–676. [CrossRef] [PubMed]
7. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **1979**, *9*, 62–66. [CrossRef]

8. Khan, J.F.; Bhuiyan, S.M.A.; Adhami, R.R. Image segmentation and shape analysis for road-sign detection. *IEEE Trans. Intell. Transp. Syst.* **2010**, *12*, 83–96. [CrossRef]
9. Sheikh, Y.A.; Khan, E.A.; Kanade, T. Mode-seeking by medoid shifts. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
10. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
11. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
12. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7151–7160.
13. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer: Cham, Switzerland, 2015; Volume 18, pp. 234–241.
14. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
15. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 1856–1867. [CrossRef]
16. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
17. Peng, Y.; Sonka, M.; Chen, D.Z. U-Net v2: Rethinking the Skip Connections of U-Net for Medical Image Segmentation. *arXiv* **2023**, arXiv:2311.17791.
18. Birkenfeld, J.S.; Tucker-Schwartz, J.M.; Soenksen, L.R.; Avilés-Izquierdo, J.A.; Marti-Fuster, B. Computer-aided classification of suspicious pigmented lesions using wide-field images. *Comput. Methods Programs Biomed.* **2020**, *195*, 105631. [CrossRef] [PubMed]
19. Soenksen, L.R.; Kassis, T.; Conover, S.T.; Marti-Fuster, B.; Birkenfeld, J.S.; Tucker-Schwartz, J.; Naseem, A.; Stavert, R.R.; Kim, C.C.; Senna, M.M.; et al. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Sci. Transl. Med.* **2020**, *13*, 581.
20. Strzelecki, M.H.; Strąkowska, M.; Kozłowski, M.; Urbańczyk, T.; Wielowieyska-Szybińska, D.; Kociołek, M. Skin lesion detection algorithms in whole body images. *Sensors* **2021**, *21*, 6639. [CrossRef] [PubMed]
21. Betz-Stablein, B.; D'alessandro, B.; Koh, U.; Plasmeijer, E.; Janda, M.; Menzies, S.W.; Hofmann-Wellenhof, R.; Green, A.C.; Soyer, H.P. Reproducible naevus counts using 3D total body photography and convolutional neural networks. *Dermatology* **2022**, *238*, 4–11. [CrossRef]
22. Dai, D.; Dong, C.; Xu, S.; Yan, Q.; Li, Z.; Zhang, C.; Luo, N. Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation. *Med. Image Anal.* **2022**, *75*, 102293. [CrossRef]
23. Xu, Q.; Ma, Z.; Na, H.E.; Duan, W. DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation. *Comput. Biol. Med.* **2023**, *154*, 106626. [CrossRef]
24. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
25. Yu, L.; Chen, H.; Dou, Q.; Qin, J.; Heng, P.A. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Trans. Med. Imaging* **2016**, *36*, 994–1004. [CrossRef]
26. Bi, L.; Kim, J.; Ahn, E.; Kumar, A.; Feng, D.; Fulham, M. Step-wise integration of deep class-specific learning for dermoscopic image segmentation. *Pattern Recognit.* **2019**, *85*, 78–89. [CrossRef]
27. Nasr-Esfahani, E.; Rafiei, S.; Jafari, M.H.; Karimi, N.; Wrobel, J.S.; Samavi, S.; Soroushmehr, S.M.R. Dense pooling layers in fully convolutional network for skin lesion segmentation. *Comput. Med. Imaging Graph.* **2019**, *78*, 101658. [CrossRef]
28. Tang, P.; Liang, Q.; Yan, X.; Xiang, S.; Sun, W.; Zhang, D.; Coppola, G. Efficient skin lesion segmentation using separable-Unet with stochastic weight averaging. *Comput. Methods Programs Biomed.* **2019**, *178*, 289–301. [CrossRef]
29. Arora, R.; Raman, B.; Nayyar, K.; Awasthi, R. Automated skin lesion segmentation using attention-based deep convolutional neural network. *Biomed. Signal Process. Control* **2021**, *65*, 102358. [CrossRef]
30. Wu, H.; Pan, J.; Li, Z.; Wen, Z.; Qin, J. Automated skin lesion segmentation via an adaptive dual attention module. *IEEE Trans. Med. Imaging* **2020**, *40*, 357–370. [CrossRef] [PubMed]
31. Kumar, A.; Hamarneh, G.; Drew, M.S. Illumination-based transformations improve skin lesion segmentation in dermoscopic images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 728–729.
32. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. ENet: A deep neural network architecture for real-time semantic segmentation. *arXiv* **2016**, arXiv:1606.02147.
33. Romera, E.; Alvarez, J.M.; Bergasa, L.M.; Arroyo, R. ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation. *IEEE Trans. Intell. Transp. Syst.* **2017**, *19*, 263–272. [CrossRef]

34. Valanarasu, J.M.J.; Patel, V.M. Unext: Mlp-based rapid medical image segmentation network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2022; pp. 23–33.
35. Ruan, J.; Xiang, S.; Xie, M.; Liu, T.; Fu, Y. MALUNet: A Multi-Attention and Light-weight UNet for Skin Lesion Segmentation. In Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 6–8 December 2022; pp. 1150–1156.
36. Ruan, J.; Xie, M.; Gao, J.; Liu, T.; Fu, Y. EGE-UNet: An Efficient Group Enhanced UNet for Skin Lesion Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2023; Volume 154, pp. 481–490.
37. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
38. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
39. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
40. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
41. Kaul, C.; Manandhar, S.; Pears, N. FocusNet: An attention-based fully convolutional network for medical image segmentation. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 455–458.
42. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018; pp. 325–341.
43. Wang, Y.; Zhou, Q.; Liu, J.; Xiong, J.; Gao, G.; Wu, X.; Latecki, L.J. LEDNet: A lightweight encoder-decoder network for real-time semantic segmentation. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1860–1864.
44. Gao, G.; Xu, G.; Yu, Y.; He, J.; Yang, J.; Yue, D. MSCFNet: A lightweight network with multi-scale context fusion for real-time semantic segmentation. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 25489–25499. [CrossRef]
45. Gao, G.; Xu, G.; Li, J.; Yu, Y.; Lu, H.; Yang, J. FBSNet: A fast bilateral symmetrical network for real-time semantic segmentation. *IEEE Trans. Multimed.* **2022**, *25*, 3273–3283. [CrossRef]
46. Gutman, D.; Codella, N.C.; Celebi, E.; Helba, B.; Marchetti, M.; Mishra, N.; Halpern, A. 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). *arXiv* **2016**, arXiv:1605.01397.
47. Codella, N.C.; Gutman, D.; Celebi, M.E.; Helba, B.; Marchetti, M.A.; Dusza, S.W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; 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 Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 168–172.
48. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **2018**, *5*, 180161. [CrossRef]
49. Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M.E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv* **2019**, arXiv:1902.03368.
50. Creswell, A.; Arulkumaran, K.; Bharath, A.A. On denois-ing autoencoders trained to minimise binary cross-entropy. *arXiv* **2017**, arXiv:1708.08487.
51. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
52. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 205–218.
53. Wu, T.; Tang, S.; Zhang, R.; Cao, J.; Zhang, Y. CGNet: A Light-weight Context Guided Network for Semantic Segmentation. *IEEE Trans. Image Process.* **2020**, *30*, 1169–1179. [CrossRef] [PubMed]
54. Yu, C.; Gao, C.; Wang, Y.; Yu, G.; Shen, C.; Sang, N. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-time Semantic Segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]
55. Poudel, R.P.; Liwicki, S.; Cipolla, R. Fast-SCNN: Fast semantic segmentation network. *arXiv* **2019**, arXiv:1902.04502.
56. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8-14 September 2018; pp. 405–420.
57. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5693–5703.

58. Howard, A.; Sandler, M.; Chen, B.; Wang, W.; Chen, L.-C.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R.; et al. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
59. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 7262–7272.
60. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.