

MALUNet: A Multi-Attention and Light-weight UNet for Skin Lesion Segmentation

1st Jiacheng Ruan

Shanghai Jiao Tong University Shanghai Jiao Tong University

Shanghai, China

jackchenruan@sjtu.edu.cn

2nd Suncheng Xiang*

Shanghai, China

xiangsuncheng17@sjtu.edu.cn

3rd Mingye Xie

Shanghai, China

xiemingye@sjtu.edu.cn

4rd Ting Liu

Shanghai, China

louisia_liu@sjtu.edu.cn

5th Yuzhuo Fu*

Shanghai Jiao Tong University

Shanghai, China

yzfu@sjtu.edu.cn

Abstract—Recently, some pioneering works have preferred applying more complex modules to improve segmentation performances. However, it is not friendly for actual clinical environments due to limited computing resources. To address this challenge, we propose a light-weight model to achieve competitive performances for skin lesion segmentation at the lowest cost of parameters and computational complexity so far. Briefly, we propose four modules: (1) *DGA* consists of dilated convolution and gated attention mechanisms to extract global and local feature information; (2) *IEA*, which is based on external attention to characterize the overall datasets and enhance the connection between samples; (3) *CAB* is composed of 1D convolution and fully connected layers to perform a global and local fusion of multi-stage features to generate attention maps at channel axis; (4) *SAB*, which operates on multi-stage features by a shared 2D convolution to generate attention maps at spatial axis. We combine four modules with our U-shape architecture and obtain a light-weight medical image segmentation model dubbed as *MALUNet*. Compared with UNet, our model improves the mIoU and DSC metrics by 2.39% and 1.49%, respectively, with a 44x and 166x reduction in the number of parameters and computational complexity. In addition, we conduct comparison experiments on two skin lesion segmentation datasets (ISIC2017 and ISIC2018). Experimental results show that our model achieves state-of-the-art in balancing the number of parameters, computational complexity and segmentation performances. Code is available at <https://github.com/JCruan519/MALUNet>.

Index Terms—Light-weight model, Medical image segmentation, Attention mechanism, Mobile health

I. INTRODUCTION

Medical image segmentation models can assist doctors in diagnosis and improve the efficiency of hospital operations. Most existing methods are based on UNet [1], an Encoder-Decoder model with an U-shape architecture. Due to its simplicity and scalability advantages, plenty of improved models are proposed, such as UNet++ [2], AttentionUNet [3], 3D-UNet [4], V-Net [5], UNet3+ [6] and so on. Recently, researchers have introduced the self-attention mechanism (SA) of ViT [7] into UNet. Some essential extensions like TransUNet [8],

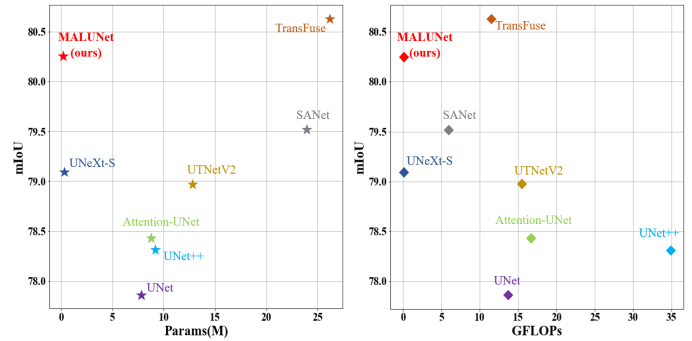


Fig. 1. Visualization of comparison results on the ISIC2018 dataset. X-axis corresponds to the number of parameters and GFLOPs, respectively (lower the better). Y-axis represents the mIoU (higher the better).

TransBTS [9], Swin-UNet [10], DS-TransUNet [11], TransFuse [12], UTNet [13] and UTNetV2 [14] have been proposed, further improving segmentation performances.

However, the previous works still have the following problems. Firstly, previous studies tend to introduce more complex modules into UNet in exchange for improving performances. However, due to the limited memory of mobile medical devices, many models with a large amount of parameters cannot be employed well in real-world clinical scenarios. Secondly, medical image segmentation belongs to the layout-specific task [15]. Namely, the variance between samples is slight, but within samples is significant in the medical dataset. The modules in recent works are often improved based on the SA module or its variants, while SA is not expert in modeling the relationship between samples [16]. Finally, for medical image segmentation, it is imperative to fully use multi-stage and multi-scale information due to varying target sizes. However, many models ignore the importance of that.

In this paper, we propose four modules to solve the problems: (1) Dilated Gated Attention Block (DGA), which helps our model pay more attention to the target region by obtaining global and local information. At the same time, the depthwise separable convolution [17] further reduces the number of

* Suncheng Xiang and Yuzhuo Fu are the co-corresponding authors. This work was partially supported by the National Natural Science Foundation of China (Grant No. 61977045).

parameters and computational complexity. (2) An efficient external attention mechanism for the light-weight medical segmentation model is presented, called Inverted External Attention Block (IEA), which can enhance the information association between samples and obtain characteristics of the overall dataset. (3) For the acquisition of multi-stage and multi-scale information, attention bridge modules are proposed for channel and spatial axes, dubbed as Channel Attention Bridge Block (CAB) and Spatial Attention Bridge Block (SAB), which can generate corresponding attention maps. With the help of the modules mentioned above, we could drastically reduce the number of channels in our model and maintain a competitive segmentation effect with the minimum number of parameters and lowest complexity.

The contributions lie in three folds:

- Four attention modules are proposed, which could obtain global and local feature information respectively, depict the sample characteristics of the whole dataset, and fuse multi-stage information to generate corresponding attention maps at the channel and spatial levels.
- We present MALUNet based on our U-shape architecture and four attention modules, which has the minimum amount of parameters and the lowest computational complexity.
- Extensive experiments are performed on ISIC2017 [18] and ISIC2018 [19] datasets. Results show that our model achieves the state-of-the-art in terms of the balance between parameters, computational complexity and segmentation performances.

II. RELATED WORKS

FCN [20] is the pioneer in image segmentation, which utilizes fully convolution to extract feature information. After that, UNet [1], an Encoder-Decoder model based on the fully convolution, is developed for medical image segmentation and achieves terrific results.

Most medical image segmentation models are improved based on UNet, such as UNet++ [2], Attention-UNet [3], Res-UNet [21] [22], Dense-UNet [23], etc. UNet++ replaces the cropping and concatenating operation in the skip-connection part of UNet with the convolution operation in a dense way to obtain better feature information and compensate for the information loss caused by sampling. Attention-UNet utilizes attention gates to control the importance of features, making it pay more attention to targets. In order to further decrease the loss of information and improve performances, Res-UNet and Dense-UNet replace plain convolutions with Res-block in ResNet [24] and Dense-block in DenseNet [25], respectively.

Google transplanted the SA in natural language processing [26] to computer vision, and proposed ViT [7] as the backbone. Because of its powerful feature extraction ability, researchers recently focused on how to combine ViT and its variants with UNet to obtain better results [27] [28]. For example, Swin-UNet [10] combines Swin Transformer [29] with UNet and receives a better segmentation effect. DS-TransUNet [11] utilizes patches of different scales as input, and two

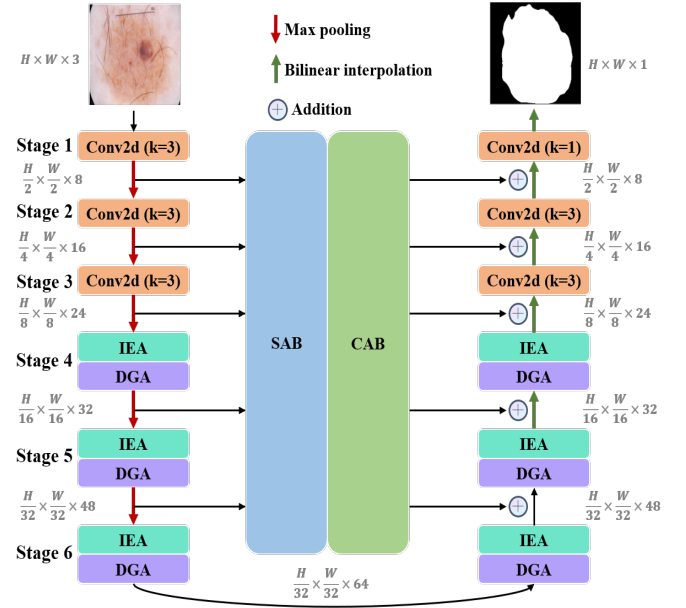


Fig. 2. The illustration of MALUNet architecture.

parallel Swin Transformers are used as encoders to attain more abundant information. TransFuse [12] applies the parallel encoder structure of CNN and Transformer to obtain local and global information simultaneously. UTNetV2 [14] improves the UTNet [13], which is a hybrid multi-level structure. Local modeling is introduced to reduce the dependence on large-scale data via using depthwise separable convolution as the projection and feed-forward network in the Transformer block.

The above networks are designed to improve performances while ignoring the shortcomings of heavy parameters and computational complexity, making it difficult to be applied in the real medical environment. Recently, Jeya et al. proposed UNeXt [30] based on combining MLP [31] with UNet, which significantly reduces the number of parameters on the premise of ensuring the segmentation performances. A light-weight model is necessary for practical applications and mobile health. Therefore, based on previous studies, this paper proposes a light-weight model for medical image segmentation, and introduces a variety of attention modules to ensure performances.

III. METHODS

In this section, we first introduce the four modules proposed in this paper: DGA, IEA, CAB and SAB. Afterward, we elaborate on the proposed light-weight model, MALUNet, as shown in Fig. 2.

A. Dilated Gated Attention Block

Medical image segmentation belongs to dense prediction tasks, and the simultaneous acquisition of global and local information is crucial to improving performances. Global information is helpful for the model to understand the overall lesion structure and its relationship with the background, making it locate the lesion area more accurately. Moreover,

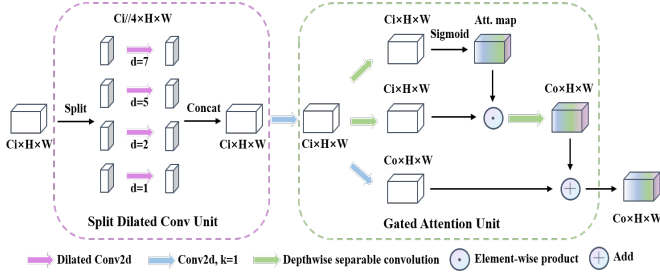


Fig. 3. Overview of the proposed DGA architecture. Note that unless otherwise specified, the convolution kernel size is 3×3 .

local information helps to obtain the details of edges and corners of the lesion area, making predictions more complete.

Therefore, DGA is proposed in this paper. As shown in Fig. 3, this module is composed of two sub-units: Split Dilated Conv Unit (SDC) and Gated Attention Unit (GA). SDC splits the feature map along the channel dimension into four parts, and obtains the global (convolutions with dilated rates of 5 and 7) and local (convolutions with dilated rates of 1 and 2) feature information through depthwise separable convolution with different dilated rates [17] [32]. Then, the concatenating operation is carried out in the channel dimension to restore the size of the feature map, followed by a convolution operation to interact with the global and local information. Next, for GA, the attention map with the same shape as the input feature is generated via depthwise separable convolution to suppress the unimportant region of feature information transmitted by SDC, so that model pays more attention to the vital information. Finally, a residual connection operation is applied to obtain the output. The above process can be expressed by formulas (1) to (5).

$$x_1, x_2, x_3, x_4 = \text{Chunk}_4(X) \quad (1)$$

$$x'_1, x'_2, x'_3, x'_4 = W_1(x_1), W_2(x_2), W_5(x_3), W_7(x_4) \quad (2)$$

$$X' = W(\text{Concat}(x'_1, x'_2, x'_3, x'_4)) \quad (3)$$

$$\text{Att} = \sigma(\text{DW}(X')) \quad (4)$$

$$\text{Out} = \text{DW}(\text{DW}(X') \odot \text{Att}) + W(X') \quad (5)$$

Where Chunk_4 indicates that the input feature map is divided into four parts along the channel dimension, W_i represents depthwise separable convolution with dilated rates of i , Concat denotes the concatenating operation, W represents the plain convolution operation, σ is the sigmoid function, DW indicates the depthwise separable convolution, and \odot represents the element-wise multiplication.

B. Inverted External Attention Block

EAMLP [16] proposes a new attention mechanism, called external attention (EA), which employs two memory units (two Conv1D operations that share parameters) to characterize the feature information of the entire dataset. Besides, inspired by the inverted residual module in MobileNetV2 [33], a

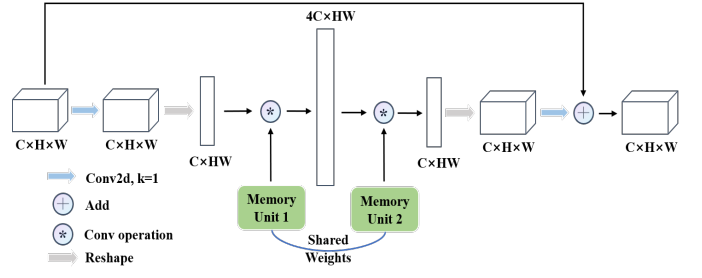


Fig. 4. The illustration of IEA architecture.

general light-weight backbone in computer vision, we propose an inverted external attention block (IEA) for light-weight medical image segmentation model as shown in Fig. 4.

Given an input $X \in \mathcal{R}^{C \times H \times W}$, after 1×1 convolution and reshaping operation, the size is changed to $X \in \mathcal{R}^{C \times HW}$. We use Memory Unit 1 to expand the feature map fourfold to obtain $\text{Att} \in \mathcal{R}^{4C \times HW}$, and then Memory Unit 2 is applied to restore the dimension, followed by reshaping operation to restore to the original feature map size. Finally, the output feature map can be obtained through a 1×1 convolution and added with the residual information. It is worth noting that in IEA, we no longer fix the number of channels between two memory units as 64 as EA, but apply four times expansion. The advantage is that the memory unit maps the input to a higher dimensional space, which makes the memory unit describe the overall feature information of the dataset more comprehensively.

C. Channel Attention Bridge Block

The acquisition of multi-stage and multi-scale information plays an essential role in segmenting targets of different sizes, and the fusion of multi-stage and multi-scale information has been proved to be the key to improve the performance. Therefore, we propose a bridge attention module for channel level, called CAB. It is utilized to generate channel attention map by concatenating the features of different stages at channel axis to better integrate information. CAB can be expressed by formulas (6) to (10).

$$t'_i = \text{GAP}(t_i) \quad (6)$$

$$T = \text{Concat}(t'_1, t'_2, \dots, t'_{s-1}) \quad (7)$$

$$T' = \text{Conv1D}(T) \quad (8)$$

$$\text{Att}_i = \sigma(\text{FC}_i(T')) \quad (9)$$

$$\text{Out}_i = t_i + t_i \odot \text{Att}_i \quad (10)$$

Where GAP refers to global average pooling, t_i represents the feature map of different stages obtained from the encoder, Concat indicates concatenating operation in the channel dimension, s denotes the total number of stages, Conv1D represents 1D convolution operation, FC_i is the fully connected layer for stage i , σ is the sigmoid function, and \odot is the element-wise multiplication.

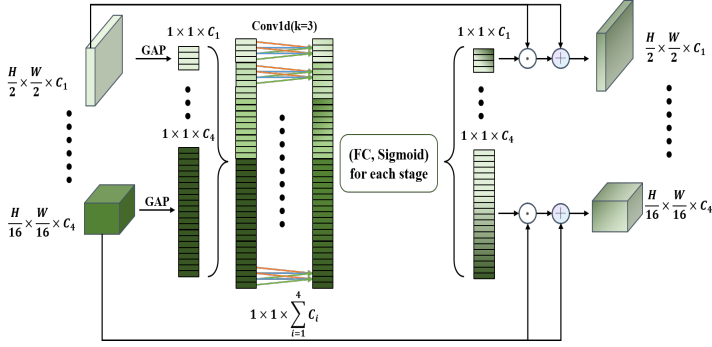


Fig. 5. Overview of the proposed CAB architecture.

As shown in Fig. 5, CAB is visualized by taking the five-stage model ($s = 5$) as an example. Note that CAB subdivides the multi-stage and multi-scale information fusion into local information fusion (1D convolution operations) and global information fusion (different fully connected layers for each stage), to provide a more informative attention feature map.

D. Spatial Attention Bridge Block

CBAM [34] believes that both channel and spatial attention play an essential role in suppressing irrelevant information. Therefore, inspired by CBAM, we introduced another bridge attention module, SAB, as shown in Fig. 6. Also, five-stage SAB is visualized as an example. This module fuses the multi-stage and multi-scale information at spatial axis to generate the attention map for each stage.

First, for each stage's feature map, we employ average-pooling and max-pooling operations in the channel dimension, and concatenate them to get the feature map with two channels, while the height and width remain unchanged. Then, a shared dilated convolution operation is utilized (the dilated rate is 3 and kernel size is 7), followed by a sigmoid function to generate a spatial attention map for each stage. Finally, the generated spatial attention map is element-wise multiplied by the original map and added with the residual information.

E. MALUNet

After combining our attention modules and U-shape architecture, our model is obtained, dubbed as MALUNet, as shown in Fig. 2.

In MALUNet, a six-stage U-shape architecture is applied, and the number of channels in each stage is $\{8, 16, 24, 32, 48, 64\}$. Based on UNet, we add one stage, and the plain convolution operation is reduced from two to one in Stage 1-3. In Stage 4-6, we use IEA and DGA, which are arranged in series. The former depicts the overall features of the dataset and the relationship between samples, and the latter acquires the global and local feature information in samples. They complement each other and make full use of the advantages of different attention mechanisms. In addition, SAB-first order is employed for the sequential arrangement of two bridge attention modules. The two bridge attention modules can fuse the multi-stage and multi-scale features of

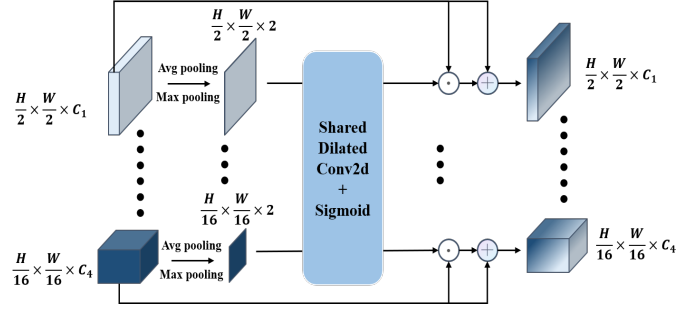


Fig. 6. Overview of the proposed SAB architecture.

Stages 1-5 to generate the attention maps in the spatial and channel dimension. And then, we add features obtained by bridge attention modules with features of the decoder part to reduce the feature semantic difference between the encoder and decoder while alleviating the information loss caused by the sampling process.

IV. EXPERIMENTS

A. Datasets

In this section, we conduct extensive experiments on two public datasets of skin lesion segmentation, the International Skin Imaging Collaboration challenge dataset (ISIC2017 and ISIC2018) [18] [19], to train and evaluate the proposed model. ISIC2017 and ISIC2018 have 2150 and 2694 dermoscopy images with segmentation mask labels, respectively. We randomly divide datasets in a ratio of 7 : 3 as our experiments' training and testing sets. For ISIC2017, there are 1500 images in training sets and 650 images in testing sets. For ISIC2018, there are 1886 images in training sets and 808 images in testing sets. Note that we perform comparative experiments on ISIC2017 and ISIC2018 and conduct ablation experiments on ISIC2018.

B. Implementation details

All experiments are implemented on a single NVIDIA GeForce RTX3080 GPU. Empirically, all images are normalized and resized to 256×256 , and we employ data augmentation including vertical flip, horizontal flip and random rotation. The loss function is BceDice loss, which can be expressed by formula (11). We utilize AdamW [35] as the optimizer with an initial learning rate of 0.001. A cosine annealing learning rate scheduler is used with a maximum number of iterations of 50 and a minimum learning rate of 0.00001. The training epoch is set equal to 300, and the batch size is 8.

$$L_{Bce} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

$$L_{Dice} = 1 - \frac{2|X \cap Y|}{|X| + |Y|}$$

$$L_{BceDice} = \lambda_1 L_{Bce} + \lambda_2 L_{Dice}$$
(11)

Where N is the total number of samples, y_i is the real label, p_i is the prediction. $|X|$ and $|Y|$ represent ground truth and

prediction, respectively. λ_1 and λ_2 refer to the weight of two loss functions. In this paper, both weights are taken as 1 by default.

C. Evaluation Metrics

Five metrics including Mean Intersection over Union (**mIoU**), Dice similarity score (**DSC**), Accuracy (**Acc**), Sensitivity (**Sen**), Specificity (**Spe**) are used to measure segmentation performances. In addition, **Params** is utilized to indicate the number of parameters, and the unit is Million (M). The computational complexity is calculated regarding the number of floating point operators (**GFLOPs**). Note that the **Params** and **GFLOPs** of models are measured with 256×256 input size.

$$\begin{cases} \mathbf{mIoU} = \frac{TP}{TP+FP+FN} \\ \mathbf{DSC} = \frac{2TP}{2TP+FP+FN} \\ \mathbf{Acc} = \frac{TP+TN}{TP+FP+FN} \\ \mathbf{Sen} = \frac{TP}{TP+FN} \\ \mathbf{Spe} = \frac{TN}{TN+FP} \end{cases} \quad (12)$$

Where TP , FP , FN , TN represent true positive, false positive, false negative, and true negative.

D. Comparison with state-of-the-arts

This section compares MALUNet with state-of-the-arts that employ different ideas to improve UNet [1] in recent years on ISIC2017 and ISIC2018 datasets. The experimental results are shown in Table I and Table II. Note that UNeXt [30], a light-weight medical image segmentation model, is based on the five-stage U-shape architecture, and the number of channels in each stage is $\{32, 64, 128, 160, 256\}$ in the original paper. In order to compare more obviously with our model, we employ UNeXt-S, and its number of channels is $\{8, 16, 32, 64, 128\}$ in our experiments, which is the default setting used in its official open source codes. Therefore, the parameters and computational complexity of UNeXt-S and our model are the same sizes, which is convenient to compare the performance.

According to the results, compared with UNet, MALUNet still improves the performance in an all-around way. In addition, the parameters are reduced by 44x, and the computational complexity is decreased by 166x. For ISIC2018 dataset, compared with TransFuse [12], although our model is 0.38% and 0.23% lower in mIoU and DSC, it can not be ignored that MALUNet is 150x and 139x lower than TransFuse in terms of parameters and computational complexity. Compared with UNeXt-S in the same starting point as ours, the performance of MALUNet on the two datasets is better than UNeXt-S, and further reduces the number of parameters and computational complexity by 0.125M and 0.017GFLOPs. As shown in Fig. 1, it can be seen more clearly that MALUNet achieves state-of-the-art in terms of the balance between parameters, computational complexity and segmentation performances. Besides, we visualize some results, as shown in Fig. 7, and it is evident that our model achieves better performances with more precise edges and no hollows.

TABLE I
COMPARATIVE EXPERIMENTAL RESULTS ON ISIC2017 DATASET. (**BOLD** INDICATES THE BEST AND UNDERLINE INDICATES THE SECOND BEST.)

Model	Params ↓	GFLOPs ↓	mIoU ↑	DSC ↑	Acc ↑	Spe ↑	Sen ↑
UNet [1]	7.77	13.78	76.98	86.99	95.65	97.43	86.82
TransFuse [12]	26.27	11.53	79.21	88.40	<u>96.17</u>	97.98	87.14
UTNetV2 [14]	12.80	15.50	77.35	87.23	95.84	<u>98.05</u>	84.85
UNeXt-S [30]	<u>0.30</u>	<u>0.10</u>	78.26	87.80	95.95	97.74	<u>87.04</u>
MALUNet (ours)	0.175	0.083	<u>78.78</u>	<u>88.13</u>	96.18	98.47	84.78

TABLE II
COMPARATIVE EXPERIMENTAL RESULTS ON ISIC2018 DATASET. (**BOLD** INDICATES THE BEST AND UNDERLINE INDICATES THE SECOND BEST.)

Model	Params ↓	GFLOPs ↓	mIoU ↑	DSC ↑	Acc ↑	Spe ↑	Sen ↑
UNet [1]	7.77	13.78	77.86	87.55	94.05	<u>96.69</u>	85.86
UNet++ [2]	9.16	34.90	78.31	87.83	94.02	95.75	88.65
Attention-UNet [3]	8.73	16.74	78.43	87.91	94.13	96.23	87.60
TransFuse [12]	26.27	11.53	80.63	89.27	94.66	95.74	91.28
UTNetV2 [14]	12.80	15.50	78.97	88.25	94.32	96.48	87.60
SANet [36]	23.90	5.99	79.52	88.59	94.39	95.97	89.46
UNeXt-S [30]	<u>0.30</u>	<u>0.10</u>	79.09	88.33	94.39	96.72	87.15
MALUNet (ours)	0.175	0.083	<u>80.25</u>	<u>89.04</u>	<u>94.62</u>	96.19	<u>89.74</u>

TABLE III
ABLATION PERFORMANCES OF UNET UNDER DIFFERENT SETTINGS.

#Channels	Params ↓	GFLOPs ↓	mIoU ↑	DSC ↑	Acc ↑	Spe ↑	Sen ↑
{32,64,128,256,512}	7.77	13.76	77.86	87.55	94.05	96.69	85.86
{16,32,64,128,160,256}	1.30	0.42	78.59	88.01	94.15	96.07	88.16
{8,16,32,64,128,160}	0.57	0.14	78.50	87.96	94.24	96.75	86.42
{8,16,32,48,64,96}	0.21	0.10	78.44	87.92	94.11	96.08	87.98
{8,16,24,32,48,64}	0.11	0.07	78.22	87.77	94.09	96.32	87.15

E. Ablation Studies

a) *Determination of baseline*: UNet [1] is a five-stage U-shape architecture model. The number of channels in each stage is $\{32, 64, 128, 256, 512\}$, and the number of plain convolutions in each stage is two. In this paper, we change the five-stage U-shape architecture to six stages, and the number of plain convolutions in each stage is one. Table III shows the effect of baseline under different settings. It is noted that the UNet in five-stage uses concatenating operation between encoder and decoder, and the six-stage model is element-wise addition. Under the premise of considering the performance and the size of model comprehensively, we utilize the six-stage UNet with the number of channels of $\{8, 16, 24, 32, 48, 64\}$, dubbed as BasicUNet (BU), to carry out the subsequent ablation experiments.

b) *Ablation study on the single module*: The ablation experiment of the single module is based on BasicUNet. We train and test each proposed module on ISIC2018 dataset to prove its effectiveness. Moreover, to illustrate the advantages of our IEA, we also employ EA [16] for comparison. Note that (1) the number of intermediate channels in EA used here is the same as the original paper, which is set to 64; (2) DGA replaces the convolution operation in BasicUNet in Stage 4-6; (3) EA and

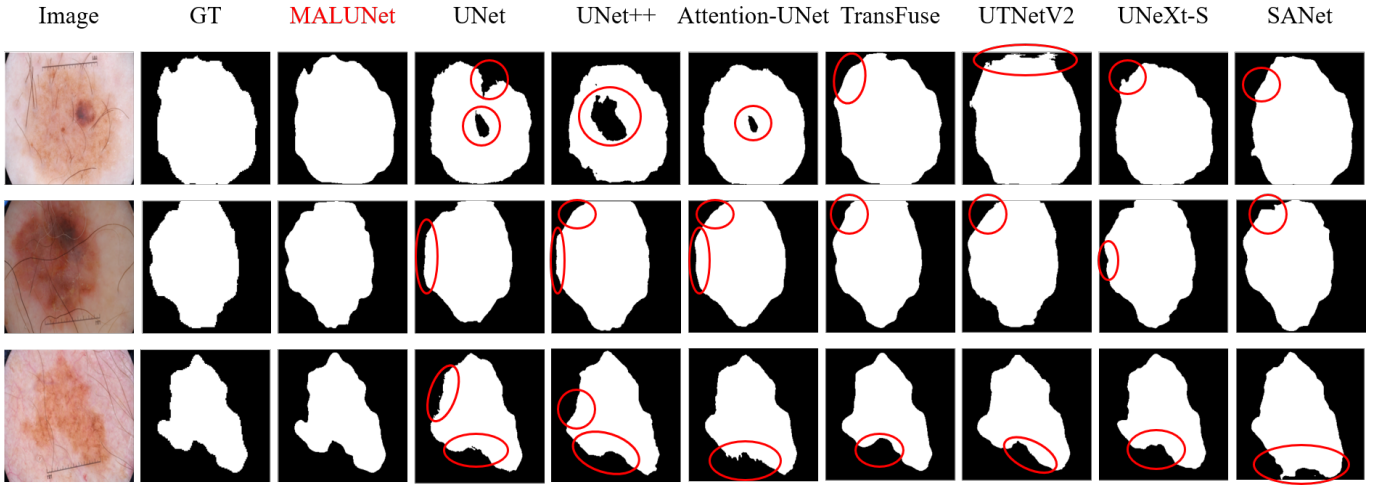


Fig. 7. Visualization of some predictions. The red mark circles defect parts of the predictions.

TABLE IV
ABLATION PERFORMANCES OF THE SINGLE MODULE.

Module	Params ↓	GFLOPs↓	mIoU↑	DSC↑	Acc↑	Spe↑	Sen↑
BasicUNet (BU)	0.11	0.07	78.22	87.77	94.09	96.32	87.15
BU + DGA	0.08	0.06	79.12	88.43	94.33	96.29	88.25
BU + EA	0.15	0.08	79.31	88.46	94.44	96.65	87.57
BU + IEA	0.18	0.09	79.77	88.75	94.58	96.77	87.76
BU + CAB	0.12	0.07	78.98	88.26	94.36	96.69	87.12
BU + SAB	0.11	0.07	78.61	88.02	94.24	96.60	86.92

TABLE V
ABLATION PERFORMANCES OF FOUR MODULES. (A) REPRESENTS THE CAB-FIRST ORDER AND (B) REPRESENTS THE SAB-FIRST ORDER.

Model	Params ↓	GFLOPs↓	mIoU↑	DSC↑	Acc↑	Spe↑	Sen↑
BU+DGA+IEA+(A)	0.247	0.097	80.35	89.10	94.71	96.57	88.93
BU+DGA+IEA+(B)	0.247	0.097	80.23	89.14	94.84	97.34	87.00
BU+IEA+DGA+(A)	0.175	0.083	79.80	88.77	94.42	95.65	90.58
BU+IEA+DGA+(B)	0.175	0.083	80.25	89.04	94.62	96.19	89.74

IEA are utilized in Stage 4-6 before the plain convolution; (4) CAB and SAB fuse features from the encoder of Stage 1-5. As shown in Table IV, DGA can improve the performance while reducing the amount of parameters and computational complexity at the same time. It can be shown that obtaining global and local feature information simultaneously can make the model “see” more comprehensively. In addition, the other three modules can also improve the performance while only a few additional parameters are introduced.

c) *Ablation study on four modules*: Four modules hybrid ablation experiment on ISIC2018 is utilized to determine the best form of MALUNet. The ablation experiment results are shown in Table V. After combining the four modules with BasicUNet (BU), although the performance of BU+DGA+IEA+(B) is the best, the starting point of this paper is to design a light-weight medical image segmentation model by balancing the relationship between the parameters, computational complexity and performances. Therefore, BU+IEA+DGA+(B) is finally selected as MALUNet in this paper.

V. CONCLUSION

In this study, we propose four attention modules in order to: 1) Obtain global and local feature information; 2) Depict the sample feature of the whole dataset so that the relationship between samples can be established; 3) Generate a complete

channel attention map by fusing the multi-stage and multi-scale features locally and globally at channel axis; 4) Fuse the multi-stage and multi-scale feature information, and obtain the corresponding spatial attention map at spatial axis. MALUNet is presented by combining our proposed modules and six-stage U-shape architecture, which is a light-weight medical image segmentation model. The comparative experiments and visualization results on ISIC2017 and ISIC2018 datasets can prove that we realize state-of-the-art in terms of the balance between parameters, computational complexity and performances. It is believed that our work provides new enlightenment for the subsequent development of light-weight medical image segmentation models. However, the limitation is that MALUNet is proposed only for skin lesion segmentation. Therefore, we will improve performances through pruning and neural architecture search methods and make our model applicable to other diseases in future work.

REFERENCES

- [1] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [2] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: A nested u-net architecture for medical image segmentation,” in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2018, pp. 3–11.
- [3] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, “Atten-

- tion u-net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.
- [4] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3d u-net: learning dense volumetric segmentation from sparse annotation,” in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.
 - [5] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.
 - [6] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, “Unet 3+: A full-scale connected unet for medical image segmentation,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1055–1059.
 - [7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
 - [8] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “Transunet: Transformers make strong encoders for medical image segmentation,” *arXiv preprint arXiv:2102.04306*, 2021.
 - [9] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, “Transbts: Multimodal brain tumor segmentation using transformer,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 109–119.
 - [10] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, “Swin-unet: Unet-like pure transformer for medical image segmentation,” *arXiv preprint arXiv:2105.05537*, 2021.
 - [11] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, “Ds-transunet: Dual swin transformer u-net for medical image segmentation,” *IEEE Transactions on Instrumentation and Measurement*, 2022.
 - [12] Y. Zhang, H. Liu, and Q. Hu, “Transfuse: Fusing transformers and cnns for medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 14–24.
 - [13] Y. Gao, M. Zhou, and D. N. Metaxas, “Utunet: a hybrid transformer architecture for medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 61–71.
 - [14] Y. Gao, M. Zhou, D. Liu, and D. Metaxas, “A multi-scale transformer for medical image segmentation: Architectures, model efficiency, and benchmarks,” *arXiv preprint arXiv:2203.00131*, 2022.
 - [15] J. Chen, T. He, W. Zhuo, L. Ma, S. Ha, and S.-H. G. Chan, “Tvconv: Efficient translation variant convolution for layout-aware visual processing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 548–12 558.
 - [16] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, “Beyond self-attention: External attention using two linear layers for visual tasks,” *arXiv preprint arXiv:2105.02358*, 2021.
 - [17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.
 - [18] M. Berseth, “Isic 2017-skin lesion analysis towards melanoma detection,” *arXiv preprint arXiv:1703.00523*, 2017.
 - [19] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic),” *arXiv preprint arXiv:1902.03368*, 2019.
 - [20] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
 - [21] X. Xiao, S. Lian, Z. Luo, and S. Li, “Weighted res-unet for high-quality retina vessel segmentation,” in *2018 9th international conference on information technology in medicine and education (ITME)*. IEEE, 2018, pp. 327–331.
 - [22] A. Khanna, N. D. Londhe, S. Gupta, and A. Semwal, “A deep residual u-net convolutional neural network for automated lung segmentation in computed tomography images,” *Biocybernetics and Biomedical Engineering*, vol. 40, no. 3, pp. 1314–1327, 2020.
 - [23] S. Cai, Y. Tian, H. Lui, H. Zeng, Y. Wu, and G. Chen, “Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network,” *Quantitative imaging in medicine and surgery*, vol. 10, no. 6, p. 1275, 2020.
 - [24] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
 - [25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
 - [26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
 - [27] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, “Transformers in medical imaging: A survey,” *arXiv preprint arXiv:2201.09873*, 2022.
 - [28] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, “Transformers in medical image analysis: A review,” *arXiv preprint arXiv:2202.12165*, 2022.
 - [29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.
 - [30] J. M. J. Valanarasu and V. M. Patel, “Unext: Mlp-based rapid medical image segmentation network,” *arXiv preprint arXiv:2203.04967*, 2022.
 - [31] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit *et al.*, “Mlp-mixer: An all-mlp architecture for vision,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 261–24 272, 2021.
 - [32] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” *arXiv preprint arXiv:1511.07122*, 2015.
 - [33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
 - [34] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
 - [35] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
 - [36] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, “Shallow attention network for polyp segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 699–708.