EGE-UNet: an Efficient Group Enhanced UNet for skin lesion segmentation

Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu (✉)⋆

Shanghai Jiao Tong University, China
{jackchenruan, xiemingye, gaojingsheng, louisa_liu, yzfu}@sjtu.edu.cn

Abstract. Transformer and its variants have been widely used for medical image segmentation. However, the large number of parameter and computational load of these models make them unsuitable for mobile health applications. To address this issue, we propose a more efficient approach, the Efficient Group Enhanced UNet (**EGE-UNet**). We incorporate a Group multi-axis Hadamard Product Attention module (GHPA) and a Group Aggregation Bridge module (GAB) in a lightweight manner. The GHPA groups input features and performs Hadamard Product Attention mechanism (HPA) on different axes to extract pathological information from diverse perspectives. The GAB effectively fuses multi-scale information by grouping low-level features, high-level features, and a mask generated by the decoder at each stage. Comprehensive experiments on the ISIC2017 and ISIC2018 datasets demonstrate that EGE-UNet outperforms existing state-of-the-art methods. In short, compared to the TransFuse, our model achieves superior segmentation performance while reducing parameter and computation costs by **494x** and **160x**, respectively. Moreover, to our best knowledge, this is the first model with a parameter count limited to just **50KB**. Our code is available at https://github.com/JCruan519/EGE-UNet.

Keywords: Medical image segmentation · Light-weight model · mobile health.

1 Introduction

Malignant melanoma is one of the most rapidly growing cancers in the world. As estimated by the American Cancer Society, there were approximately 100,350 new cases and over 6,500 deaths in 2020 [14]. Thus, an automated skin lesion segmentation system is imperative, as it can assist medical professionals in swiftly identifying lesion areas and facilitating subsequent treatment processes. To enhance the segmentation performance, recent studies tend to employ modules with larger parameter and computational complexity, such as incorporating self-attention mechanisms of Vision Transformer (ViT) [7]. For example, Swin-UNet [4], based on the Swin Transformer [11], leverages the feature extraction ability

⋆ Yuzhuo Fu is the corresponding author and this work was partially supported by the National Natural Science Foundation of China (Grant No. 61977045).

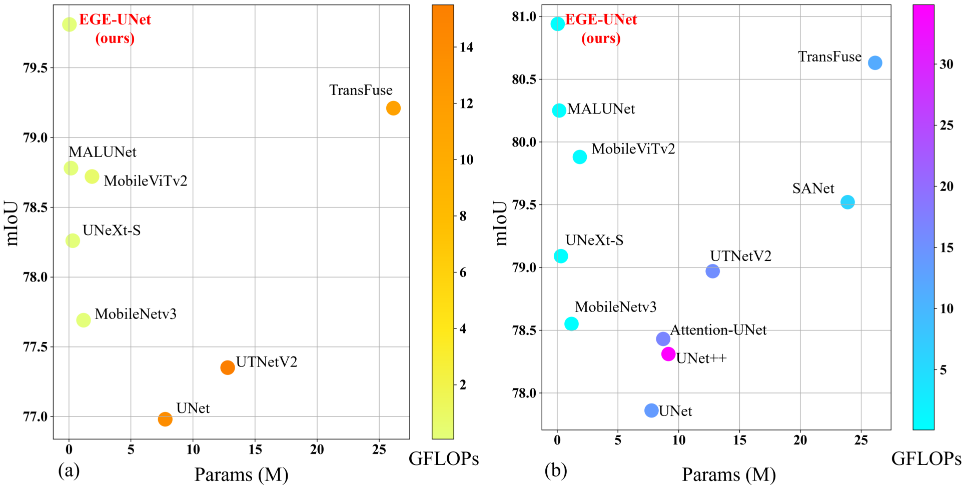

Fig. 1. (a) and (b) respectively show the visualization of comparative experimental results on the ISIC2017 and ISIC2018 datasets. The X-axis represents the number of parameters (lower is better), while Y-axis represents mIoU (higher is better). The color depth represents computational complexity (GFLOPs, lighter is better).

of self-attention mechanisms to improve segmentation performance. TransUNet [5] has pioneered a serial fusion of CNN and ViT for medical image segmentation. TransFuse [26] employs a dual-path structure, utilizing CNN and ViT to capture local and global information, respectively. UTNetV2 [8] utilizes a hybrid hierarchical architecture, efficient bidirectional attention, and semantic maps to achieve global multi-scale feature fusion, combining the strengths of CNN and ViT. TransBTS [23] introduces self-attention into brain tumor segmentation tasks and uses it to aggregate high-level information.

Prior works have enhanced performance by introducing intricate modules, but neglected the constraint of computational resources in real medical settings. Hence, there is an urgent need to design a low-parameter and low-computational load model for segmentation tasks in mobile healthcare. Recently, UNeXt [22] has combined UNet [18] and MLP [21] to develop a lightweight model that attains superior performance, while diminishing parameter and computation. Furthermore, MALUNet [19] has reduced the model size by declining the number of model channels and introducing multiple attention modules, resulting in better performance for skin lesion segmentation than UNeXt. However, while MALUNet greatly reduces the number of parameter and computation, its segmentation performance is still lower than some large models, such as TransFuse. Therefore, in this study, we propose EGE-UNet, a lightweight skin lesion segmentation model that achieves state-of-the-art while significantly reducing parameter and computation costs. Additionally, to our best knowledge, this is the first work to reduce parameter to approximately **50KB**.

To be specific, EGE-UNet leverages two key modules: the Group multi-axis Hadamard Product Attention module (GHPA) and Group Aggregation Bridge module (GAB). On the one hand, recent models based on ViT [7] have shown promise, owing to the multi-head self-attention mechanism (MHSA). MHSA divides the input into multiple heads and calculates self-attention in each head, which allows the model to obtain information from diverse perspectives, integrate different knowledge, and improve performance. Nonetheless, the quadratic complexity of MHSA enormously increases the model's size. Therefore, we present the Hadamard Product Attention mechanism (HPA) with linear complexity. HPA employs a learnable weight and performs a hadamard product operation with the input to obtain the output. Subsequently, inspired by the multi-head mode in MHSA, we propose GHPA, which divides the input into different groups and performs HPA in each group. However, it is worth noting that we perform HPA on different axes in different groups, which helps to further obtain information from diverse perspectives. On the other hand, for GAB, since the size and shape of segmentation targets in medical images are inconsistent, it is essential to obtain multi-scale information [19]. Therefore, GAB integrates high-level and low-level features with different sizes based on group aggregation, and additionally introduce mask information to assist feature fusion. Via combining the above two modules with UNet, we propose EGE-UNet, which achieves excellent segmentation performance with extremely low parameter and computation. Unlike previous approaches that focus solely on improving performance, our model also prioritizes usability in real-world environments. A clear comparison of EGE-UNet with others is shown in Figure 1.

In summary, our contributions are threefold: (1) GHPA and GAB are proposed, with the former efficiently acquiring and integrating multi-perspective information and the latter accepting features at different scales, along with an auxiliary mask for efficient multi-scale feature fusion. (2) We propose EGE-UNet, an extremely lightweight model designed for skin lesion segmentation. (3) We conduct extensive experiments, which demonstrate the effectiveness of our methods in achieving state-of-the-art performance with significantly lower resource requirements.

2 EGE-UNet

The overall architecture. EGE-UNet is illustrated in Figure 2, which is built upon the U-Shape architecture consisting of symmetric encoder-decoder parts. We take encoder part as an example. The encoder is composed of six stages, each with channel numbers of $\{8, 16, 24, 32, 48, 64\}$. While the first three stages employ plain convolutions with a kernel size of 3, the last three stages utilize the proposed GHPA to extract representation information from diverse perspectives. In contrast to the simple skip connections in UNet, EGE-UNet incorporates GAB for each stage between the encoder and decoder. Furthermore, our model leverages deep supervision [27] to generate mask predictions of varying scales, which are utilized for loss function and serve as one of the inputs to

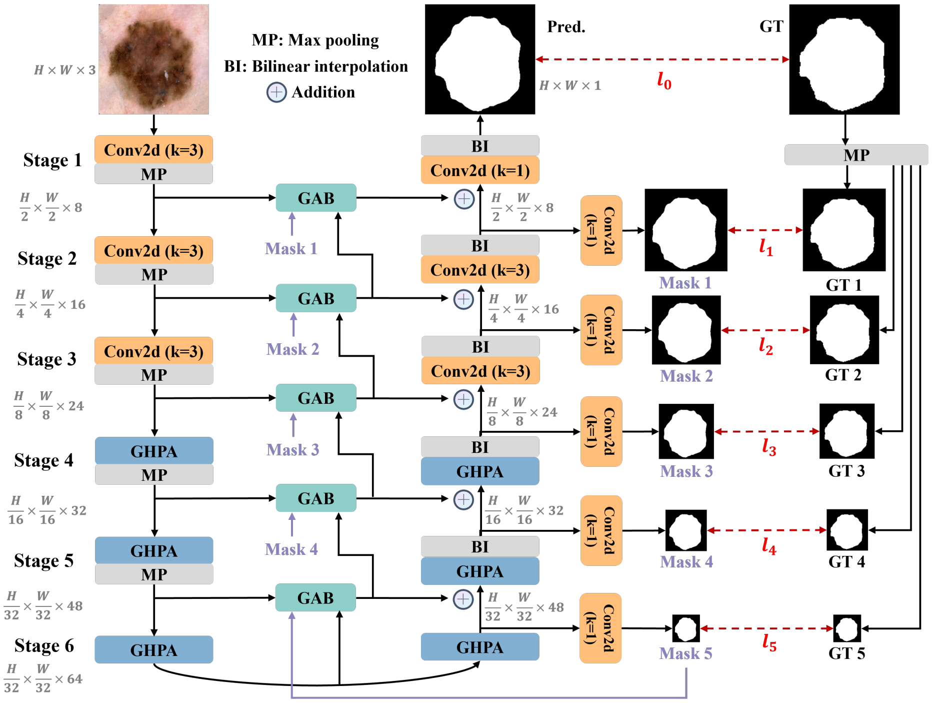
MP: Max pooling
BI: Bilinear interpolation
⊕ Addition
H × W × 3
Pred.
GT
H × W × 1
l_0
Stage 1
Conv2d (k=3)
MP
$\frac{H}{2}\times\frac{W}{2}\times 8$
Stage 2
$\frac{H}{4}\times\frac{W}{4}\times 16$
Stage 3
$\frac{H}{8}\times\frac{W}{8}\times 24$
Stage 4
GHPA
$\frac{H}{16}\times\frac{W}{16}\times 32$
Stage 5
$\frac{H}{32}\times\frac{W}{32}\times 48$
Stage 6
$\frac{H}{32}\times\frac{W}{32}\times 64$
GAB
Mask 1
Mask 2
Mask 3
Mask 4
Mask 5
BI
Conv2d (k=1)
Conv2d (k=3)
GT 1
GT 2
GT 3
GT 4
GT 5
l_1
l_2
l_3
l_4
l_5

Fig. 2. The overview of EGE-UNet.

Algorithm 1: The Pytorch-style Pseudo-code for GHPA

```
# Input: X, the feature map with shape [B, C, H, W]
# Output: Out, the feature map with shape [B, C, H, W]
# Params: a, the hyperparameter and default by 8 in this paper
          b, the hyperparameter and default by 8 in this paper
          $P_{xy}$ , the randomly initialized tensor with shape [1, C//4, a, b]
          $P_{zx}$ , the randomly initialized tensor with shape [1, 1, C//4, a]
          $P_{zy}$ , the randomly initialized tensor with shape [1, 1, C//4, b]
# Operator: DW, Depthwise Separable Convolution
            LN, LayerNorm          BI, Bilinear interpolation

x1, x2, x3, x4 = torch.chunk(LN(X), 4, dim=1)
x1, x4 = x1 * DW(BI($P_{xy}$)), DW(x4)
x2 = (x2.permute(0,3,1,2) * DW(BI($P_{zx}$))).permute(0,2,3,1)
x3 = (x3.permute(0,2,1,3) * DW(BI($P_{zy}$))).permute(0,2,1,3)
Out = DW(LN(torch.cat([x1,x2,x3,x4], dim=1)))
```

GAB. Via the integration of these advanced modules, EGE-UNet significantly reduces the parameter and computational load while enhancing the segmentation performance compared to prior approaches.

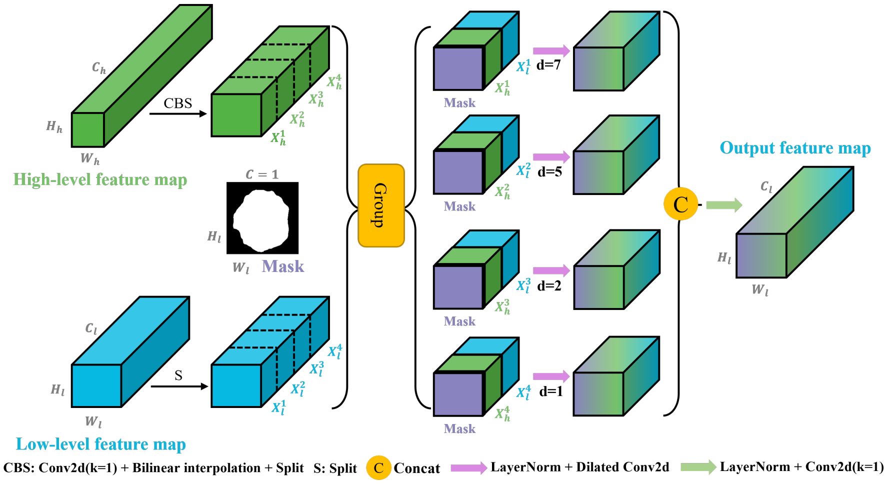

Fig. 3. The architecture of Group Aggregation Bridge module (GAB).

Group multi-axis Hadamard Product Attention module. To overcome the quadratic complexity issue posed by MHSA, we propose HPA with linear complexity. Given an input x and a randomly initialized learnable tensor p, bilinear interpolation is first utilized to resize p to match the size of x. Then, we employ depth-wise separable convolution (DW) [10][20] on p, followed by a hadamard product operation between x and p to obtain the output. However, utilizing simple HPA alone is insufficient to extract information from multiple perspectives, resulting in unsatisfactory results. Motivated by the multi-head mode in MHSA, we introduce GHPA based on HPA, as illustrated in Algorithm 1. We divide the input into four groups equally along the channel dimension and perform HPA on the height-width, channel-height, and channel-width axes for the first three groups, respectively. For the last group, we only use DW on the feature map. Finally, we concatenate the four groups along the channel dimension and apply another DW to integrate the information from different perspectives. Note that all kernel size employed in DW are 3.

Group Aggregation Bridge module. The acquisition of multi-scale information is deemed pivotal for dense prediction tasks, such as medical image segmentation. Hence, as shown in Figure 3, we introduce GAB, which takes three inputs: low-level features, high-level features, and a mask. Firstly, depthwise separable convolution (DW) and bilinear interpolation are employed to adjust the size of high-level features, so as to match the size of low-level features. Secondly, we partition both feature maps into four groups along the channel dimension, and concatenate one group from the low-level features with one from the high-level features to obtain four groups of fused features. For each group of fused features, the mask is concatenated. Next, dilated convolutions [25] with kernel size of 3 and different dilated rates of {1, 2, 5, 7} are applied to the different groups, in order to extract information at different scales. Finally, the four groups are concatenated along the channel dimension, followed by the application of a plain convolution with the kernel size of 1 to enable interaction among features at different scales.

Loss function. In this study, since different GAB require different scales of mask information, deep supervision [27] is employed to calculate the loss function for different stages, in order to generate more accurate mask information. Our loss function can be expressed as equation (1) and (2).

$$l_i = Bce(y, \hat{y}) + Dice(y, \hat{y}) \tag{1}$$

$$\mathcal{L} = \sum_{i=0}^{5} \lambda_i \times l_i \tag{2}$$

where Bce and $Dice$ represent binary cross entropy and dice loss. λ_i is the weight for different stage. In this paper, we set λ_i to 1, 0.5, 0.4, 0.3, 0.2, 0.1 from $i = 0$ to $i = 5$ by default.

3 Experiments

Datasets and Implementation details. To assess the efficacy of our model, we select two public skin lesion segmentation datasets, namely ISIC2017 [1][3] and ISIC2018 [2][6], containing 2150 and 2694 dermoscopy images, respectively. Consistent with prior research [19], we randomly partition the datasets into training and testing sets at a 7:3 ratio.

EGE-UNet is developed by Pytorch [17] framework. All experiments are performed on a single NVIDIA RTX A6000 GPU. The images are normalized and resized to 256×256. We apply various data augmentation, including horizontal flipping, vertical flipping, and random rotation. AdamW [13] is utilized as the optimizer, initialized with a learning rate of 0.001 and the CosineAnnealingLR [12] is employed as the scheduler with a maximum number of iterations of 50 and a minimum learning rate of 1e-5. A total of 300 epochs are trained with a batch size of 8. To evaluate our method, we employ Mean Intersection over Union (mIoU), Dice similarity score (DSC) as metrics, and we conduct 5 times and report the mean and standard deviation of the results for each dataset.

Comparative results. The comparative experimental results presented in Table 1 reveal that our EGE-UNet exhibits a comprehensive state-of-the-art performance on the **ISIC2017** dataset. Specifically, in contrast to larger models, such as TransFuse, our model not only demonstrates superior performance, but also significantly curtails the number of parameter and computation by 494x and 160x, respectively. In comparison to other lightweight models, EGE-UNet surpasses UNeXt-S with a mIoU improvement of 1.55% and a DSC improvement of

Table 1. Comparative experimental results on the ISIC2017 and ISIC2018 dataset.

Dataset	**Model**	**Params(M)↓**	**GFLOPs↓**	**mIoU(%)↑**	**DSC(%)↑**
ISIC2017	UNet [18]	7.77	13.76	76.98	86.99
	UTNetV2 [8]	12.80	15.50	77.35	87.23
	TransFuse [26]	26.16	11.50	79.21	88.40
	MobileViTv2 [15]	1.87	0.70	78.72	88.09
	MobileNetv3 [9]	1.19	0.10	77.69	87.44
	UNeXt-S [22]	0.32	0.10	78.26	87.80
	MALUNet [19]	0.177	0.085	78.78	88.13
	EGE-UNet (Ours)	**0.053**	**0.072**	**79.81±0.10**	**88.77±0.06**
ISIC2018	UNet [18]	7.77	13.76	77.86	87.55
	UNet++ [27]	9.16	34.86	78.31	87.83
	Att-UNet [16]	8.73	16.71	78.43	87.91
	UTNetV2 [8]	12.80	15.50	78.97	88.25
	SANet [24]	23.90	5.96	79.52	88.59
	TransFuse [26]	26.16	11.50	80.63	89.27
	MobileViTv2 [15]	1.87	0.70	79.88	88.81
	MobileNetv3 [9]	1.19	0.10	78.55	87.98
	UNeXt-S [22]	0.32	0.10	79.09	88.33
	MALUNet [19]	0.177	0.085	80.25	89.04
	EGE-UNet (Ours)	**0.053**	**0.072**	**80.94±0.11**	**89.46±0.07**

Table 2. Ablation studies on the ISIC2017 dataset. (a) the macro ablation on two modules. (b) the micro ablation on GHPA. (c) the micro ablation on GAB.

Type	**Model**	**Params(M)↓**	**GFLOPs↓**	**mIoU(%)↑**	**DSC(%)↑**
(a)	Baseline	0.107	0.076	76.30	86.56
	Baseline + GHPA	0.034	0.058	78.82	88.16
	Baseline + GAB	0.126	0.086	78.78	88.13
(b)	w/o multi-axis grouping	0.074	0.074	79.13	88.35
	w/o DW for initialized tensor	0.050	0.072	79.03	88.29
(c)	w/o mask information	0.052	0.070	78.97	88.25
	w/o dilation rate of Conv2d	0.053	0.072	79.11	88.34

0.97%, while exhibiting parameter and computation reductions of 17% and 72% of UNeXt-S. Furthermore, EGE-UNet outperforms MALUNet with a mIoU improvement of 1.03% and a DSC improvement of 0.64%, while reducing parameter and computation to 30% and 85% of MALUNet. For the **ISIC2018** dataset, the performance of our model also outperforms that of the best-performing model. Besides, it is noteworthy that EGE-UNet is the first lightweight model reducing parameter to about 50KB with excellent segmentation performance. Figure 1 presents a more clear visualization of the experimental findings and Figure 4 shows some segmentation results.

Ablation results. We conduct extensive ablation experiments to demonstrate the effectiveness of our proposed modules. The baseline utilized in our work is referenced from MALUNet [19], which employs a six-stage U-shaped architecture

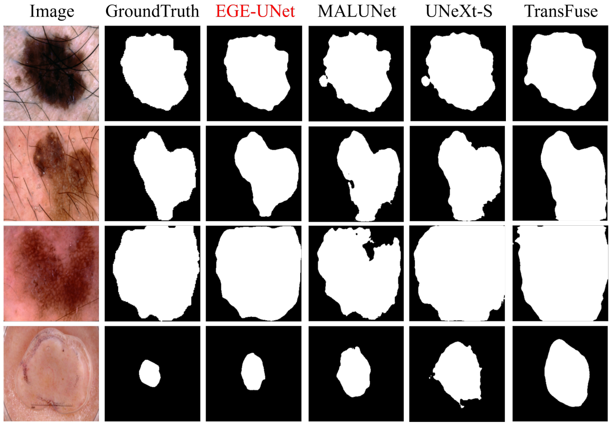

Fig. 4. Qualitative comparisons on the ISIC2018 dataset.

with symmetric encoder and decoder components. Each stage includes a plain convolution operation with a kernel size of 3, and the number of channels at each stage is set to {8, 16, 24, 32, 48, 64}. In Table 2(a), we conduct macro ablations on GHPA and GAB. Firstly, we replace the plain convolutions in the last three layers of baseline with GHPA. Due to the efficient multi-perspective feature acquisition of GHPA, it not only outperforms the baseline, but also greatly reduces the parameter and computation. Secondly, we substitute the skip-connection operation in baseline with GAB, resulting in further improved performance. Table 2(b) presents the ablations for GHPA. We replace the multi-axis grouping with single-branch and initialize the learnable tensors with only random values. It is evident that the removal of these two key designs leads to a marked drop. Table 2(c) illustrates the ablations for GAB. Initially, we omit the mask information, and mIoU metric even drops below 79%, thereby confirming once again the critical role of mask information in guiding feature fusion. Furthermore, we substitute the dilated convolutions in GAB with plain convolutions, which also leads to a reduction in performance.

4 Conclusions and Future Works

In this paper, we propose two advanced modules. Our GHPA uses a novel HPA mechanism to simplify the quadratic complexity of the self-attention to linear complexity. It also leverages grouping to fully capture information from different perspectives. Our GAB fuses low-level and high-level features and introduces a mask to integrate multi-scale information. Based on these modules, we propose EGE-UNet for skin lesion segmentation tasks. Experimental results demonstrate the effectiveness of our approach in achieving state-of-the-art performance with significantly lower resource requirements. We hope that our work can inspire further research on lightweight models for the medical image community.

Regarding limitations and future works, on the one hand, we mainly focus on how to greatly reduce the parameter and computation complexity while improving performance in this paper. Thus, we plan to deploy EGE-UNet in a real-world environment in the future work. On the other hand, EGE-UNet is currently designed only for the skin lesion segmentation task. Therefore, we will extend our lightweight design to other tasks.

References

1. https://challenge.isic-archive.com/data/#2017
2. https://challenge.isic-archive.com/data/#2018
3. Berseth, M.: Isic 2017-skin lesion analysis towards melanoma detection. arXiv preprint arXiv:1703.00523 (2017)
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
5. 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 preprint arXiv:2102.04306 (2021)
6. 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 preprint arXiv:1902.03368 (2019)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Gao, Y., Zhou, M., Liu, D., Metaxas, D.: A multi-scale transformer for medical image segmentation: Architectures, model efficiency, and benchmarks. arXiv preprint arXiv:2203.00131 (2022)
9. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)
10. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
11. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
12. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
14. Mathur, P., Sathishkumar, K., Chaturvedi, M., Das, P., Sudarshan, K.L., Santhappan, S., Nallasamy, V., John, A., Narasimhan, S., Roselind, F.S., et al.: Cancer statistics, 2020: report from national cancer registry programme, india. JCO global oncology **6**, 1063–1075 (2020)

15. Mehta, S., Rastegari, M.: Separable self-attention for mobile vision transformers. arXiv preprint arXiv:2206.02680 (2022)
16. 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 preprint arXiv:1804.03999 (2018)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
19. Ruan, J., Xiang, S., Xie, M., Liu, T., Fu, Y.: Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1150–1156. IEEE (2022)
20. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
21. Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al.: Mlp-mixer: An all-mlp architecture for vision. Advances in Neural Information Processing Systems **34**, 24261–24272 (2021)
22. Valanarasu, J.M.J., Patel, V.M.: Unext: Mlp-based rapid medical image segmentation network. arXiv preprint arXiv:2203.04967 (2022)
23. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 109–119. Springer (2021)
24. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 699–708. Springer (2021)
25. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
26. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 14–24. Springer (2021)
27. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 3–11. Springer (2018)