# LGANET: LOCAL-GLOBAL AUGMENTATION NETWORK FOR SKIN LESION SEGMENTATION

*Qingqing Guo* [1], *Xianyong Fang* [1,*], *Linbo Wang* [1], *Enming Zhang* [2], *Zhengyi Liu* [1]

[1]School of Computer Science and Technology, Anhui University, Hefei, China
[2]Department of Clinical Science, Lund University Diabetes Centre, Malmö, Sweden

## ABSTRACT

Automatic segmentation of skin lesion is still challenging due to ambiguous boundary and noise interference of lesion regions. Recent exiting Transformer-based methods often directly apply Transformer to obtain long-range dependency to overcome these problems. However, they generally do not consider that patch partitioning strategy of Transformer could lead to the loss of local details around boundaries. Furthermore, dependencies across local windows only represent global information at a coarse level. Therefore, to overcome the limitations, two novel modules, Local Focus Module (LFM) and Global Augmentation Module (GAM) are proposed in this paper. LFM learns the local context around boundary regions to strengthen the discrimination between classes. And GAM learns the global context at a finer level to enhance global feature representation. Integrating LFM and GAM, a new Transformer encoder based framework, Local-Global Augmentation Network (LGANet), is proposed. LGANet is efficient in segmenting lesions with ambiguous boundary and with noise interference and its performances are demonstrated with extensive experiments on two public skin lesion segmentation datasets.

***Index Terms*—** Skin lesion segmentation, Transformer, Local detail information, Global dependency

## 1. INTRODUCTION

Melanoma is the most malignant cancer among skin cancer [1], which seriously threatens human health and life. Dermatologists usually identify lesions visually from dermoscopy images. But manual identification is a boring and heavy workload. Automated skin lesion segmentation can greatly improve the diagnostic efficiency and assist dermatologists for further analysis.

Skin lesions often have noise interferences and ambiguous boundaries (Fig. 1). It means global context and local information surrounding boundaries are both needed. Therefore, some Transformers-based methods applied Transformer directly to extract global context as supplementary to CNN branch [2, 3, 4]. And some other methods improve the Transformer based networks with extra boundary information to address ambiguous boundary problems [5, 6, 7].

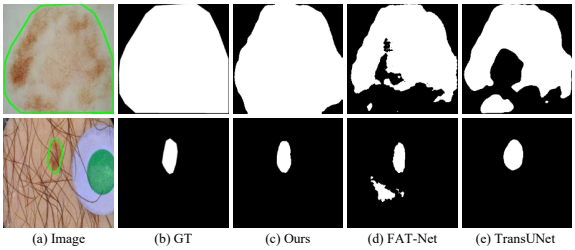


**Fig. 1**. Skin lesion segmentation results of two state-of-arts methods, TransUNet [2] and FAT-Net [3], and ours for the lesions with ambiguous boundaries and noise interferences.

However, these methods directly apply Transformer without considering that Transformer's patch partitioning strategy can destroy detailed information around boundaries when the partition lines cross the boundary areas. Especially for skin lesions in dermoscopic images, the contrast between background and lesion is very low, resulting in the indistinction between classes near the lesion boundaries. So the local inter-pixel correlations around lesion boundaries are still needed to be strengthened. Accordingly, Local Focus Module (LFM) is proposed to augment local detailed feature by performing self-attention within windows surrounding boundaries.

Besides, only focusing on information within windows ignores global dependencies. Many methods explore correlations between each local window after local window self-attention to capture global dependencies [8, 9]. But the global dependencies across each local window are coarse-grained. And different from the tasks in [8, 9], there are some noise interferences, such hair, and low contrasts between backgrounds and lesions in skin lesion images. Thus, fine-grained global information is more needed to distinguish between lesions and backgrounds. As is known, the pixel has more finer-grained information than local window. Accordingly, we proposed the Global Augmentation Module (GAM) to augment global context by capturing the correlations between local windows and global pixel representation. The global

*Corresponding author: Xianyong Fang (fangxianyong@ahu.edu.cn).

pixel representation can be adaptively learned by a linear layer.

The two modules, LFM and GAM, are integrated to Transformer encoder, leading to a new skin lesion segmentation network, Local-Global Augmentation Network (LGANet). Dense concatenations are adopted as decoder for final prediction. The main contributions can be summarized as follows:

- A local detailed information augmentation module, LFM, which learns local inter-pixel correlations surrounding boundaries to augment local context.
- A global context augmentation module, GAM, which learns global context at a finer level to further augment global dependencies.
- A deep skin lesion segmentation network, LGA, which integrates LFM and GAM into the Transformer encoder for more accurate segmentation of skin lesion images.

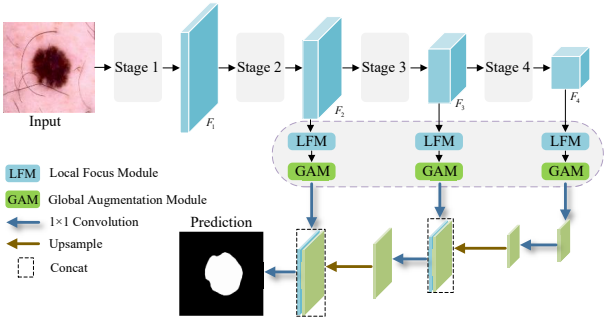


**Fig. 2**. The structure of the proposed LGANet. LFM and GAM are integrated into the Transformer encoder based framework to learn local detailed information around boundary and augment global context respectively, where dense concatenations are used for final pixel-level prediction.

## 2. METHOD

The architecture of the proposed method is shown in Fig. 2. Transformer is adopted as the encoder to extract robust features due to its stronger feature representation ability than convolution neural network (CNN) [10, 11]. Then the output feature maps of last three stages are input to LFM and GAM in sequence. Here, the output features of first stage are too coarse to consider. LFM can enhance local detailed feature representation around boundary regions to make up for the loss of local information caused by Transformer's patch partitioning strategy. GAM can further strengthen global dependency at a finer level. The combination of these two modules can improve the feature representation ability. Finally, the processed feature maps are fused by dense concatenation for pixel-level prediction.

### 2.1. Local Focus Module

Transformer's patch partitioning strategy leads to the destruction of detailed information around partition lines, especially for boundary regions. Therefore, we propose the LFM (Fig. 3), focusing on the local neighborhood of lesion edges to enhance local context around boundary regions.

The core of LFM is to conduct Self-Attention (SA) [12] within boundary windows. Different from the method in [13], the boundary windows are generated in a supervised manner in our method. Ground truth is divided to several non-overlapped windows. A binary matrix is generated according to the principle: The value is 1 whenever there are both backgrounds and lesions in windows; otherwise, the value is 0. This binary matrix is taken as supervision to generate the score map, which guides the network to choose the boundary windows.

The input feature map $F_i \in \mathbb{R}^{c\times h\times w}(i=2,3,4)$ is firstly reduced to be one channel by $1\times 1$ convolution $\mathcal{C}^{1\times 1}$. Then, we use adaptive average pooling operation to divide the feature map into $\frac{h}{s}\times\frac{w}{s}$ non-overlapped windows whose sizes are set to $s$ ($s$ is four in our experiment). Finally, the score map $M\in\mathbb{R}^{\frac{h}{s}\times\frac{w}{s}}$ of windows are predicted after the *Sigmoid* activation:

$$M=\mathcal{O}\left(\mathcal{P}^{\frac{h}{s}\times\frac{w}{s}}\left(\mathcal{C}^{1\times 1}\left(F_i\right)\right)\right), \tag{1}$$

where $\mathcal{O}$ and $\mathcal{P}$ indicate $Sigmoid$ function and adaptive average pooling operation respectively.

Finally, boundary windows can be located according to the score map. Windows with higher scores are more likely to be boundary windows. In this paper, we set the threshold to 0.5. Then Self-Attention is applied inside the windows with the scores of greater than the threshold.

### 2.2. Global Augmentation Module

Global context is also important. However, the coarse-grained global information captured by computing correlations across local windows after local window attention is insufficient for skin lesion images because there are some noise interferences in skin lesion images, *i.e.*, they need more finer-grained global information. To this end, GAM (Fig. 3) is designed to augment global context by capture the correlations between local windows and global pixel representation.

$F_l\in\mathbb{R}^{c\times h\times w}$ is the output feature map of LFM. $F_l$ is firstly flattened to $F_f\in\mathbb{R}^{c\times n}$, where $n=h\times w$. And at the same time, in order to obtain the information representation of the local window, tokens inside each window in $F_l$ are aggregated into the global token using the method in [9]. Assume that $T\in\mathbb{R}^{l\times c}$ being the aggregated tokens, where $l=\frac{h}{s}\times\frac{w}{s}$ is the number of tokens.

Individual pixels do not contain global information. Thus, we conduct information diffusion across pixels by a linear

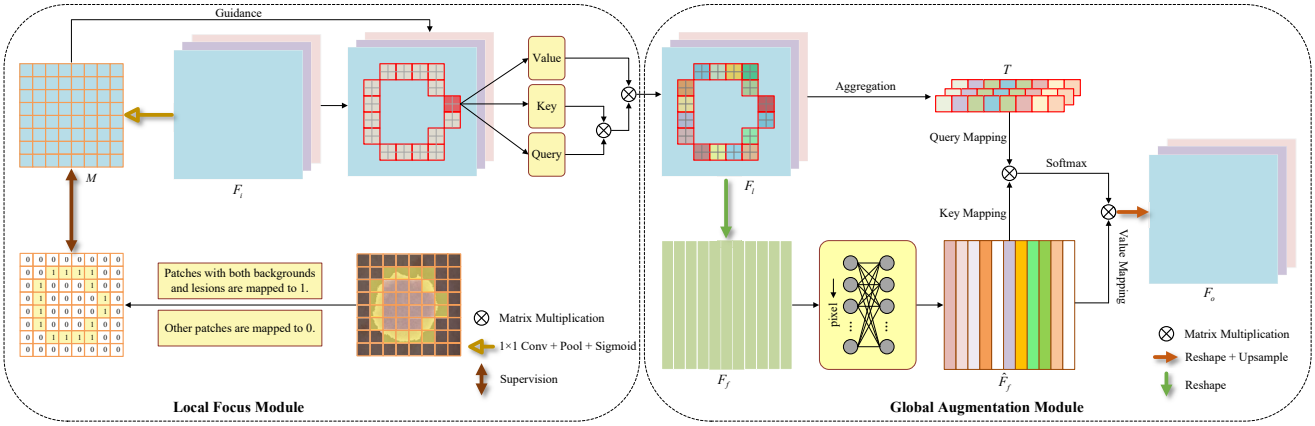


**Fig. 3**. The structure of LFM and GAM. The left is the structure of LFM and the right is the structure of GAM. The features are firstly processed by LFM to enhance the local context around boundary. Then, they are input to GAM to strengthen global context at a finer level.

layer $\mathcal{L}$ to obtain global pixel representation,

$$\hat{F}_f = \mathcal{L}\left(F_f\right). \tag{2}$$

Then, in order to enable the local token features to be more representative and informative, the cross-attention is conducted between the aggregated tokens and the global pixel representation,

$$\mathcal{A}\left(F_f, T\right) = \mathcal{Q}\left(T\right)\mathcal{K}\left(\hat{F}_f\right)^T, \tag{3}$$

$$T_o = \mathcal{S}\left(\mathcal{A}\left(\hat{F}_f, T\right)\right)\mathcal{V}\left(\hat{F}_f\right), \tag{4}$$

where $\mathcal{Q}, \mathcal{K}$ and $\mathcal{V}$ represent query, key and value mappings respectively. And $\mathcal{S}$ represents $Softmax$. In this way, the extracted global information is also finer. Finally, $T_o$ is reshaped and up-sampled to $F_o$ whose size is same as $F_l$.

### 2.3. Loss Function

To supervise the score map, the binary cross-entropy (BCE) loss is adopted. And for the supervision of final prediction, the combination of weighted binary cross-entropy (WBCE) loss and weighted Intersection over Union (WIoU) loss are also taken. The overall loss is set to be the weighted average of the losses from both predictions,

$$L_{all} = 0.7 \cdot L_p + 0.1 \cdot {L_s}^{(2)} + 0.1 \cdot {L_s}^{(3)} + 0.1 \cdot {L_s}^{(4)}, \tag{5}$$

where $L_p$ and $L_s^{(i)} (i = 2, 3, 4)$ are the losses for final prediction and the corresponding score map respectively.

## 3. EXPERIMENTS

### 3.1. Datasets and Evaluation Metrics

**Datasets** The propose method is evaluated on two public skin lesion segmentation datasets: ISIC 2016 [14] and ISIC 2018 [15]. The same dataset division policy as [3] is adopted for fair comparison. ISIC 2016 contains 1279 RGB skin lesions images, of which 900 are randomly chosen for training and 379 are used for testing. ISIC 2018 contains 2594 RGB skin lesions images. We randomly select 1815 samples for training set, 259 samples for validation set and 520 samples 191 for testing set.

**Evaluation Metrics** Six widely used metrics are employed to quantitatively evaluate the segmentation performances, including the Sensitivity (SE), Specificity (SP), Intersection over Union (IoU), Dice Similarity Coefficient (DSC), Accuracy (ACC) and Average Symmetric Surface Distance (ASSD).

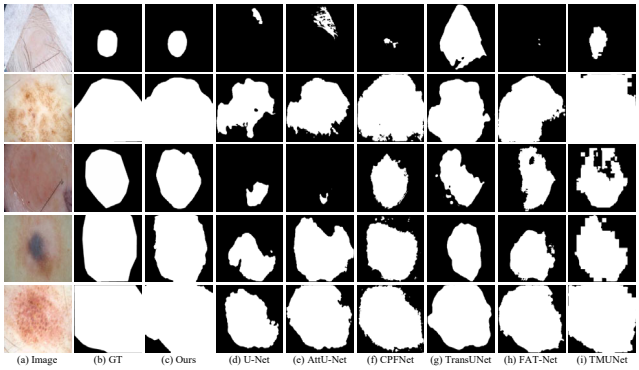


**Fig. 4**. Qualitative comparison of the segmentation results between different methods.

### 3.2. Implementation Details

Our framework is built by PyTorch with a single NVIDIA GeForce GTX 2080Ti GPU. The epoch is 100 and Adam is

**Table 1**. Quantitative comparisons of the segmentation results on two benchmarks. Best results are shown in bold.

| | Method | DSC↑ | IoU↑ | ACC↑ | SE↑ | SP↑ | ASSD↓ |
|---|---|---|---|---|---|---|---|
| ISIC 2016 | U-Net [16] | 88.84 | 81.84 | 94.66 | 90.16 | 96.56 | 7.54 |
| | AttU-Net [17] | 88.75 | 81.58 | 94.14 | 90.31 | 96.45 | 7.62 |
| | CPFNet [18] | 90.23 | 83.81 | 95.09 | 92.11 | 95.91 | 7.57 |
| | DAGAN [19] | 90.85 | 84.42 | 95.82 | 92.28 | 95.68 | – |
| | TransUNet [2] | 92.12 | 85.40 | 95.49 | 93.69 | 96.19 | 4.92 |
| | FAT-Net [3] | 91.59 | 85.30 | 96.04 | 92.59 | 96.02 | 5.01 |
| | TMUNet [7] | 92.20 | 85.54 | 95.60 | 92.32 | 96.89 | 5.37 |
| | C2FTrans [13] | 91.50 | 84.33 | 95.27 | 91.66 | 96.99 | – |
| | **Ours** | **93.88** | **88.47** | **96.55** | **94.09** | **97.51** | **4.18** |
| | Method | DSC↑ | IoU↑ | ACC↑ | SE↑ | SP↑ | ASSD↓ |
| ISIC 2018 | U-Net [16] | 85.45 | 77.33 | 94.04 | 88.00 | 96.97 | 7.71 |
| | AttU-Net [17] | 85.66 | 77.64 | 93.76 | 86.00 | **98.26** | 6.98 |
| | CPFNet [18] | 87.69 | 79.88 | 94.96 | 89.53 | 96.55 | 7.21 |
| | DAGAN [19] | 88.07 | 81.13 | 93.24 | 90.72 | 95.88 | – |
| | TransUNet [2] | 88.88 | 81.85 | 95.94 | 90.08 | 97.89 | 5.33 |
| | FAT-Net [3] | 89.03 | 82.02 | 95.78 | 91.00 | 96.99 | 5.06 |
| | TMUNet [7] | 90.59 | 82.80 | 96.03 | 90.38 | 97.46 | 6.02 |
| | C2FTrans [13] | 90.76 | **84.64** | **96.76** | **91.22** | 97.74 | – |
| | **Ours** | **91.64** | 84.56 | 96.42 | 89.96 | **98.22** | **4.58** |

the optimizer with an initial learning rate of $10^{-4}$. The batch size is set to 16 for all datasets. All images are re-sized to $256 \times 256$ as input with various data augmentations, including vertical, horizontal flip, and random rotation.

### 3.3. Comparison with other models

Some state-of-the-arts methods are compared, including four CNN-based models (U-Net [16], AttU-Net [17], CPFNet [18] and DAGAN [19]) and four Transformer-based models (TransUNet [2], FAT-Net [3], TMUNet [7] and C2FTrans [13]). The quantitative results of existing methods are reported in [3, 7, 13], except for TransUNet whose ASSDs are computed by the officially released codes.

**Quantitative results** Table 1 shows the quantitative comparison results. Obviously, our model achieves the highest scores in all metrics on ISIC 2016. Compared to C2FTrans, There are a slight decrease on ISIC 2018 in IoU, ACC and SE. And the highest scores in DSC and ASSD show that our method is more accurate for both regions and boundaries.

**Qualitative results** Fig. 4 shows that several examples of segmentation results from different methods. These images have ambiguous boundaries or noise interferences (see the first row in Fig. 4). It is observed that our result is more accurate and closer to the ground truth than others.

**Table 2**. Quantitative results for ablation study. Best results are shown in bold.

| Method | ISIC 2016 | | | ISIC 2018 | | |
|---|---|---|---|---|---|---|
| | DSC↑ | IoU↑ | ASSD↓ | DSC↑ | IoU↑ | ASSD↓ |
| Baseline | 93.67 | 88.09 | 4.28 | 91.39 | 84.15 | 4.99 |
| Baseline + LFM | 93.77 | 88.28 | 4.24 | 91.42 | 84.20 | 4.84 |
| Baseline + GAM | 93.68 | 88.11 | 4.24 | 91.47 | 84.28 | 4.77 |
| Baseline + LFM+ GAM | **93.88** | **88.47** | **4.18** | **91.64** | **84.56** | **4.58** |

**Table 3**. Quantitative results for different window sizes. Best results are shown in bold.

| Window size | ISIC 2016 | | | ISIC 2018 | | |
|---|---|---|---|---|---|---|
| | DSC↑ | IoU↑ | ASSD↓ | DSC↑ | IoU↑ | ASSD↓ |
| s = 2 | 93.65 | 88.05 | 4.21 | 91.47 | 84.29 | 4.77 |
| s = 4 | **93.88** | **88.47** | **4.18** | **91.64** | **84.56** | **4.58** |
| s = 8 | 93.53 | 86.67 | 4.50 | 91.34 | 84.05 | 4.77 |

### 3.4. Ablation studies

To evaluate the performances of each module in our proposed method, we compare our model with its three variants in Table 2. PVT v2 with additional dense concatenation is taken as the baseline. LFM and GAM are added to the baseline as different configurations. Table 2 shows that either LFM or GAM improves the performance of Baseline, demonstrating the effectiveness of each individual component. Our full model achieves the best performances, which demonstrates the necessity of taking both LEM and GAM.

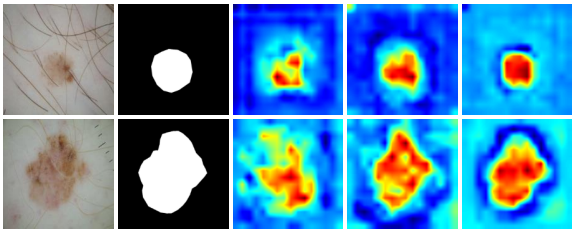


**Fig. 5**. The feature augmentation by our proposed modules. (a) Image; (b) GT; (c) Input represents the input feature map; (d) LFM denotes feature maps output by LFM; (e) GAM represents output feature map by GAM.

Fig. 5 shows the feature augmentation results after applying LFM and GAM. After LFM, the boundary contour is more clearer (Fig. 5(c)) and the responses of lesions features are more stronger after GAM (Fig. 5(d)).

We also study the influence of different window sizes in LFM (Table 3). As can been seen, $s = 4$ achieves the best performance on both ISIC 2016 and ISIC 2018. Therefore, $s$ is set to four in our model.

## 4. CONCLUSION

This paper proposes a novel framework, LGANet, for skin lesion segmentation. Particularly, two module, LFM and GAM are constructed. LFM aims at learning local inter-pixel correlations to augment local detailed information around boundary regions. While GAM aims at learning global context at a finer level to augment global information. Combing LFM and GAM makes LGANet more efficient in processing skin lesion images. Experimental results demonstrate the efficacy of the proposed method.

## 5. ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of Anhui Province (2108085MF210) and Key Natural Science Fund of Department of Education of Anhui Province (KJ2021A0042).

## 6. REFERENCES

[1] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal, "Cancer statistics, 2019," *CA: A cancer journal for clinicians*, vol. 69, no. 1, pp. 7–34, 2019.

[2] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[3] Huisi Wu, Shihuai Chen, Guilian Chen, Wei Wang, Baiying Lei, and Zhenkun Wen, "FAT-Net: Feature adaptive Transformers for automated skin lesion segmentation," *Medical Image Analysis*, vol. 76, pp. 102327, 2022.

[4] Yundong Zhang, Huiye Liu, and Qiang Hu, "TransFuse: Fusing Transformers and CNNs for medical image segmentation," in *MICCAI*, 2021, pp. 14–24.

[5] Jiacheng Wang, Fei Chen, Yuxi Ma, Liansheng Wang, Zhaodong Fei, Jianwei Shuai, Xiangdong Tang, Qichao Zhou, and Jing Qin, "XBound-Former: Toward cross-scale boundary modeling in Transformers," *arXiv preprint arXiv:2206.00806*, 2022.

[6] Jiacheng Wang, Lan Wei, Liansheng Wang, Qichao Zhou, Lei Zhu, and Jing Qin, "Boundary-aware Transformers for skin lesion segmentation," in *MICCAI*, 2021, pp. 206–216.

[7] Azad Reza, Heidari Moein, Wu Yuli, and Merhof Dorit, "Contextual attention network: Transformer meets U-Net," *arXiv preprint arXiv:2203.01932*, 2022.

[8] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang, "Swin-Unet: Unet-like pure Transformer for medical image segmentation," in *ECCVW*, 2022.

[9] Hongyi Wang, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xian-Hua Han, Yen-Wei Chen, and Ruofeng Tong, "Mixed Transformer U-Net for medical image segmentation," in *ICASSP*, 2022, pp. 2390–2394.

[10] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao, "Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions," in *ICCV*, 2021, pp. 568–578.

[11] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao, "PVT v2: Improved baselines with Pyramid Vision Transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.

[12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, "Dual attention network for scene segmentation," in *CVPR*, 2019, pp. 3146–3154.

[13] Xian Lin, Zengqiang Yan, Li Yu, and Kwang-Ting Cheng, "C2FTrans: Coarse-to-Fine Transformers for medical image segmentation," *arXiv preprint arXiv:2206.14409*, 2022.

[14] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern, "Skin lesion analysis toward melanoma detection: A challenge," *arXiv preprint arXiv:1605.01397*, 2016.

[15] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al., "Skin lesion analysis toward melanoma detection 2018: A challenge," *arXiv preprint arXiv:1902.03368*, 2019.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[17] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[18] Shuanglang Feng, Heming Zhao, Fei Shi, Xuena Cheng, Meng Wang, Yuhui Ma, Dehui Xiang, Weifang Zhu, and Xinjian Chen, "CPFNet: Context pyramid fusion network for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 10, pp. 3008–3018, 2020.

[19] Baiying Lei, Zaimin Xia, Feng Jiang, Xudong Jiang, Zongyuan Ge, Yanwu Xu, Jing Qin, Siping Chen, Tianfu Wang, and Shuqiang Wang, "Skin lesion segmentation via generative adversarial networks with dual discriminators," *Medical Image Analysis*, vol. 64, pp. 101716, 2020.