Boundary-Aware Transformers for Skin Lesion Segmentation

Jiacheng Wang[1], Lan Wei[2], Liansheng Wang[1(✉)], Qichao Zhou[3(✉)], Lei Zhu[4], and Jing Qin[5]

[1] Department of Computer Science at School of Informatics, Xiamen University, Xiamen, China
jiachengw@stu.xmu.edu.cn, lswang@xmu.edu.cn

[2] School of Electrical and Computer Engineering, Xiamen University Malaysia, Bandar Sunsuria, Malaysia

[3] Manteia Technologies Co., Ltd., Xiamen, China
zhouqc@manteiatech.com

[4] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
lzhu@cse.cuhk.edu.hk

[5] Center for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China
harry.qin@polyu.edu.hk

Abstract. Skin lesion segmentation from dermoscopy images is of great importance for improving the quantitative analysis of skin cancer. However, the automatic segmentation of melanoma is a very challenging task owing to the large variation of melanoma and ambiguous boundaries of lesion areas. While convolutional neutral networks (CNNs) have achieved remarkable progress in this task, most of existing solutions are still incapable of effectively capturing global dependencies to counteract the inductive bias caused by limited receptive fields. Recently, transformers have been proposed as a promising tool for global context modeling by employing a powerful global attention mechanism, but one of their main shortcomings when applied to segmentation tasks is that they cannot effectively extract sufficient local details to tackle ambiguous boundaries. We propose a novel boundary-aware transformer (BAT) to comprehensively address the challenges of automatic skin lesion segmentation. Specifically, we integrate a new boundary-wise attention gate (BAG) into transformers to enable the whole network to not only effectively model global long-range dependencies via transformers but also, simultaneously, capture more local details by making full use of boundary-wise prior knowledge. Particularly, the auxiliary supervision of BAG is capable of assisting transformers to learn position embedding as it provides much spatial information. We conducted extensive experiments to evaluate the proposed BAT and experiments corroborate its effectiveness, consistently outperforming state-of-the-art methods in two famous datasets (Code is available at https://github.com/jcwang123/BA-Transformer).

J. Wang and L. Wei—Contributed equally.

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12901, pp. 206–216, 2021.
https://doi.org/10.1007/978-3-030-87193-2_20

Keywords: Transformer · Medical image segmentation · Deep learning

1 Introduction

Melanoma is one of the most rapidly increasing cancers all over the world. According to the American Cancer Society's estimation, there are about 100,350 new cases and over 65,00 deaths in 2020 [13]. Segmenting skin lesions from dermoscopy images is a key step in skin cancer diagnosis and treatment planning. In current clinical practice, dermatologists usually need to manually delineate skin lesions for further analysis. However, manual delineation is usually tedious, time-consuming, and error-prone. To the end, automated segmentation methods are highly demanded in clinical practice to improve the segmentation efficiency and accuracy.

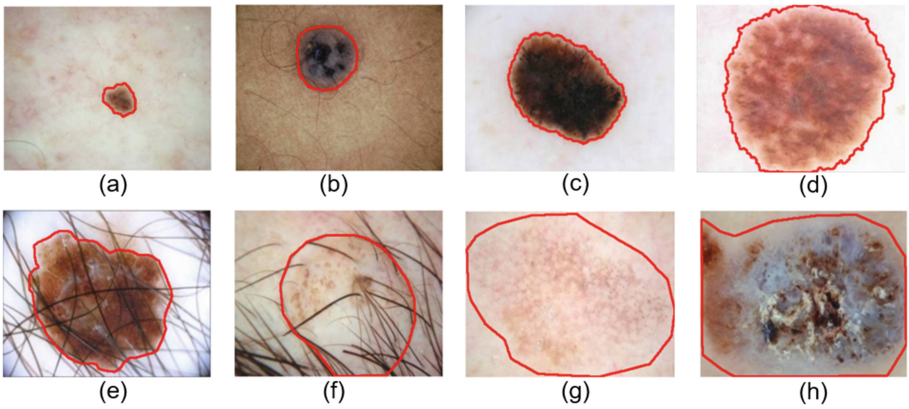

Fig. 1. The challenges of automatic skin lesion segmentation from dermoscopy images: (a)–(d) large skin lesion variations in size, shape, and color, (f)(g) partial occlusion by hair, and (f)–(h) ambiguous boundaries.

It remains, however, a very challenging task because (1) skin lesions have large variations in size, shape, and color (see Fig. 1 (a–d)), (2) present of hair will partially cover the lesions destroying local context, (3) the contrast between some lesions to normal skin are relatively low, resulting in ambiguous boundaries (see Fig. 1 (e–h)), and (4) the limited training data make the task even harder. A lot of effort has been dedicated to overcoming these challenges. Traditional methods based on various hand-crafted features are usually not stable and robust, leading to poor segmentation performance when facing lesions with large variations [15]. The main reason is that these hand-crafted features are incapable of capturing distinctive representations of skin lesions. To solve the problem, deep learning models based on convolutional neural networks (CNN) have been proposed and achieved remarkable performance gains compared with

traditional methods such as some advanced version of the fully convolutional network (FCN) [20,21]. However, these models are still insufficient to tackle the challenges of skin lesion segmentation due to the inductive bias caused by the lack of global context. With regard to this, researchers propose various approaches to enlarging the receptive fields inspired by the advancement of dilated convolution [18,19]. Lee *et al.* [10] extensively incorporate the dilated attention module with boundary prior so that the network predict boundary key-points maps to guide the attention module.

Nevertheless, most existing solutions are still incapable of effectively capturing sufficient global context to deal with above mentioned challenges. Recently, *transformers* have been proposed to regard an image as a sequence of patches and aggregate feature in global context by self-attention mechanisms [4,14]. For example, TransUNet [5], a hybrid architecture of CNN and transformer, performs well on Synapse multi-organ segmentation. Yet, it is difficult for transformer based framework to achieve the same success on skin lesion segmentation, which usually has only thousands of data not the same as what they have done in the COCO 2017 Challenge [4] containing 118k training images and 5k validation images. Limited images make it difficult to encode position embedding, and hence will not always be able to accurately and effectively model long-range interactions. Moreover, regions of lesion cover a relatively small area compared to normal tissues and generally has ambiguous boundary not as human organs, which interferes with segmentation performance by a large margin.

In this paper, we propose boundary-aware transformer (BAT) to ably handle aforementioned problems, by holistically leveraging the advancement of boundary-wise prior knowledge and transformer-based network. In fact, this design is based on the intuitions for human beings to perceive lesions in vision, i.e. considering global context to coarsely locate lesion area and paying special attention to ambiguous area to specify the exact boundary. Concretely, we propose a boundary-wise attention gate (BAG) in transformer architecture to make full use of boundary-wise prior knowledge. Firstly, BAG would learn which patches in the sequence belong to ambiguous boundary, thus providing a patch-wise attention map to guide this attention gate. Secondly, a novel key-patch map generation algorithm is introduced for adeptly giving the ground-truth label that can best represent the ambiguous boundary of target lesion. Thirdly, the auxiliary supervision of BAG provides feedback to train transformers that can let it efficiently learn position embedding on a relatively small dataset. We evaluate our model on different publicly available databases. One is the ISBI 2016 and PH2 dataset following the experimental setting in most recent work [10], the other one is the latest ISIC 2018 dataset consisting of 2594 labeled images in total. All the experiment results demonstrate the significant performance gains of our proposed framework.

2 Method

An overview of our proposed model is illustrated in Fig. 2. We will first introduce our basic transformer network that leverages the intrinsic local locality of CNN

and innate long-range dependency to complement the skin lesion segmentation in Sect. 2.1. Then, the overall framework of our boundary-aware transformer for segmenting regions of ambiguous boundary will be elaborated in Sect. 2.2. In the end, Sect. 2.3 introduces the hybrid objective function to efficiently train our boundary-aware transformer.

2.1 Basic Transformer for Segmentation

Our basic transformer (BT) blends typical CNN's and transformer's architectures to perform segmentation task, by three procedures: image sequentialization, sequence transformation, and atrous prediction. Image sequentialization is typically the primary step at which an image is converted into 1D sequential embedding fitting the input format of subsequent sequence transformation. Atrous prediction is designed to effectively remedy the local feature representation which has been ignored in sequence transformation and robustly segment lesions at multiple scales.

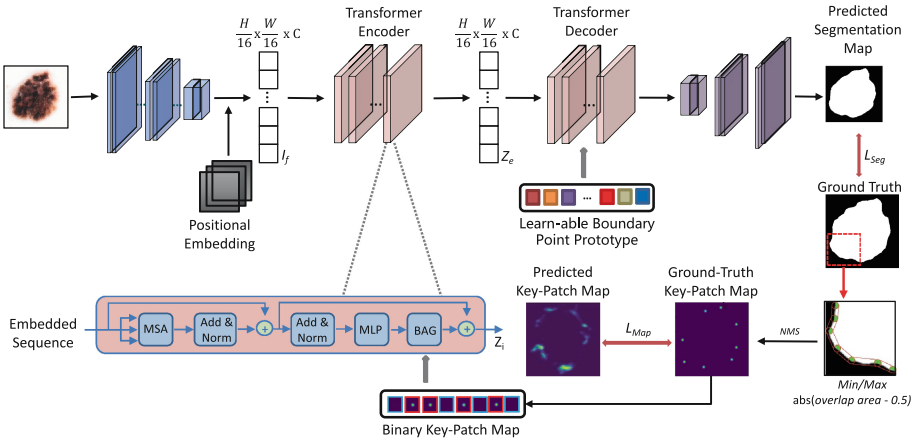

Fig. 2. An overview of the proposed Boundary-Aware Transformer framework.

Image Sequentialization. Current literature processing sequentialization by typical CNN encoder [4,11] or linear projection [7,22]. Despite its successful advancement achieved on some natural vision tasks, linear projection remains the shortcoming that is the high dependency on the amount of dataset, leading to inferior segmentation performance in medical domain [5]. To this end, we perform sequentialization through the former. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$ with width W and height H, the CNN backbone (ResNet50 as default in this work) produces the corresponding image feature map $I_f \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C}$, that the size of each patch is 16×16. The feature map is then flattened into 1D patch embedding and added by a learnable positial embedding [8] which is randomly initialized to

compensate spatial information destroyed by sequentialization, resulting in final sequential embedding as $E \in \mathbb{R}^{L \times C}, L = \frac{HW}{256}$.

Sequence Transformation. Transformer encoder composed of n stacked encoder layers is applied to capture long-range context in a whole dermoscopic image. Each layer in the encoder consists of a multi-head self-attention module (MSA) and a Multilayer Perceptron (MLP) following typical design [17]. Assumed that the input of i-th layer is Z^{i-1} (specially, $Z^0 \leftarrow E$), the output can be written as follows:

$$Z^i = MSA(Z^{i-1}) \oplus MLP(MSA(Z^{i-1})), \tag{1}$$

Eventually, transformed feature of the last layer Z^n will be reshaped to 2D format as $Z \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$, for dense prediction in the next.

Atrous Prediction. For segmenting lesions at multiple scales, this module takes transformed feature Z after self-attention mechanism as input and aims to produce a dense prediction. Aiming to enhance the local feature representation and handle the multi-scale lesion context, an atrous prediction module is designed as follows:

$$\hat{S}_{pred} = \delta(d_1^1([d_1^3(Z), d_3^3(Z), d_6^3(Z)])). \tag{2}$$

Here, $d_r^s(\cdot)$ denotes dilated convolution function with a dilation rate r and filter size of $s \times s$. δ is a sigmoid function. The enhanced feature maps $d_r^s(Z)$ with various receptive fields are concatenated across channel-wise and projected into segmentation map space.

2.2 Boundary-Aware Transformer

Efforts to incorporate structural boundary information to CNNs have been made a lot these years, but there is little literature investigating the effectiveness on transformer. We argue that the equipment of boundary information can also let transformer obtain more power in addressing lesions with ambiguous boundary. To this end, we devise the boundary-aware transformer (BAT), in which a boundary-wise attention gate (BAG) is added at end of each transformer encoder layer to refine transformed feature. BAG's architecture is similar to conventional spatial attention gate including (1) a key-patch map generator which takes the transformed feature as input and output a binary patch-wise attention map $\hat{M}_{pred} = \delta(d_1^1(Z)) \in \mathbb{R}^{L \times 1}$, where value 1 indicates that the corresponding patch is at ambiguous boundary. (2) and a residual attention scheme for preserving boundary-wise information. Hence, the boundary-aware transformed feature can be re-written as:

$$\begin{aligned} V^{i-1} &= MSA(Z^{i-1}) \oplus MLP(MSA(Z^{i-1})), \\ Z^i &= V^{i-1} \oplus (V^{i-1} \otimes \hat{M}^{i-1}), \end{aligned} \tag{3}$$

where $\oplus$ and $\otimes$ denote element-wise addition and channel-wise multiplication, respectively.

In addition to BAGs in transformer encoder layers, a query embedding based BAG is applied after encoder to refine the feature Z^n. It plays the same role in boundary-wise attention but comes true by a totally different way. Here, instead of learning the linear projection as classifier, we refer a learnable embedding Q_b as context prototype for regions among ambiguous boundary. It will be compared with all patch embedding (Z) after aforementioned blocks, to produce a similarity map M^n. Those patches with high similarity will be the regions of ambiguous boundary. Similar to other BAGs, a residual attention scheme is also applied here as: $Z^{n+1} = Z^n \oplus (Z^n \otimes \hat{M}^n)$.

By this design, BAT learns robust feature representation of ambiguous boundary in a variety of ways, which is of great significance to handle segmentation of lesions with ambiguous boundary. Following our basic design, feature Z^{n+1} is fed into atrous prediction module to produce the segmentation map $\hat{S_{pred}}$.

Boundary-Supervised Generator. As the generator doesn't necessarily know on its own which patches can best represent structural boundary of target lesion, we introduce a novel algorithm to produce ground-truth key-patch map to train the generator with full supervision. Besides the enhancement of boundary features, this design can also help in accelerating training transformer thanks to the auxiliary constraints.

Specifically, boundary points set is produced using conventional edge detection algorithm at first. For each point in this set, we draw a circle of radius r (set to 10 as default) and calculate the proportion p of lesion area in this circle. Larger or smaller proportion indicates that boundary is not smooth in this cricle. Thus we score each point as $|p - 0.5|$, representing the assistance in segmenting ambiguous parts. Non-maximum suppression is then utilized to filter points with larger proportion than neighbour k (set to 30 as default) points. Next, filtered points' 2D location (x, y) is mapped into 1D location as $\lfloor x/16 \rfloor * 16 + \lfloor y/16 \rfloor$, and patch labels at these location are set to 1 and others are set to 0, leading to final ground-truth M_{GT}.

2.3 Objective Function

To train the segmentation network including the proposed BAGs, we emply two types of loss functions. The first one is a Dice loss function to minimize the difference between the groud-truth segmentation map and the predicted segmentation map as L_{Seg}. The second one is a Cross-Entropy loss to reduce the predicted key-patch map and its ground-truth as L_{Map}. Total loss is defined as:

$$\begin{aligned} L_{Total} =& L_{Seg} + \sum_{i=1}^{n+1} L_{Map}^i, \\ L_{Seg} = \phi_{DICE}(S_{GT}, \hat{S}_{Pred}), & L_{Map}^i = \phi_{CE}(M_{GT}, \hat{M}_{Pred}^i), \end{aligned} \tag{4}$$

where M_{Pred}^i denotes the predicted key-patch map at i-th transformer encoder layer. ϕ_{DICE}, ϕ_{CE} denote Dice loss function and Cross-Entropy loss function,

respectively. n denotes the number of transformer encoder layers and is set to 4 as default.

3 Experimental Results

3.1 Datasets

We conduct extensive experiments on the skin lesion segmentation datasets from International Symposium on Biomedical Imaging (ISBI) of the years 2016 and 2018. The datasets are collected from a variety of different treatment centers, archived by the International Skin Imaging Collaboration (ISIC), which hosted a challenge named skin lesion analysis toward melanoma detection to boost the performance of melanoma diagnosis. ISIC 2016 contains a total number of 900 samples for training and a total number of 379 dermoscopy images for testing. We follow up the same experimental protocols in the most recent work [10], in which we train our model on the training set of ISIC 2016 and extensively evaluate it on PH2 dataset. ISIC 2018 contains 2594 training samples in total and annotation of its public test set is missing, therefore we perform five-fold cross-validation on its training set for fair comparison.

3.2 Implementation Details

Our network is implemented on a single NVIDIA RTX 3090Ti. All images are empirically resized to (512×512) considering the efficiency, and we do data augmentation including vertical flip, horizontal flip and random scale change (limited 0.9–1.1). Each mini-batch includes 24 images and we utilize Adam with an initial learning rate of 0.001 to optimize the network. Learning rate decrease in half when loss on the validation set has not dropped by 10 epochs. The encoder of each network has been pre-trained on ImageNet and all parameters are then fine-tuned for 500 epochs in total.

Table 1. Experimental results on different datasets.

Model	ISIC 2016 + PH2		Model	ISIC 2018	
	Dice ↑	*IoU* ↑		*Dice* ↑	*IoU* ↑
SSLS[1]	0.783	0.681	DeepLabv3 [6]	0.884	0.806
MSCA[3]	0.815	0.723	U-Net++ [23]	0.879	0.805
FCN [12]	0.894	0.821	CE-Net [9]	0.891	0.816
Bi *et al.* [2]	0.906	0.839	MedT [16]	0.859	0.778
Lee *et al.* [10]	0.918	0.843	TransUNet [5]	0.894	0.822
BAT	0.921	0.858	BAT	0.912	0.843

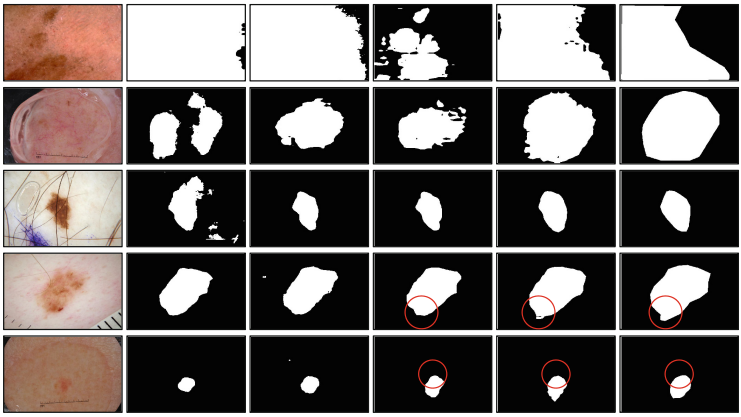

Fig. 3. Visual comparison of lesion segmentation results produced by different methods.

3.3 Comparison with State-of-the-Arts

For baseline comparisons, we run experiments on both convolutional and transformer-based methods. With regard to the evaluation metrics, we employ a Dice coefficient (Dice), and a Intersection over Union (IoU). Table 1 displayed the comparative study of our proposed boundary-aware transformer (BAT) with other methods on different datasets. It's obviously shown that our model achieves the best segmentation performance.

On the *ISIC 2016 + PH2* dataset, we compare our method with five state-of-the-art methods. Among them, Lee *et al.* [10] is a 2D attention-based model with use of boundary-prior knowledge, achieving best segmentation performance on skin lesion segmentation recently. As seen in the Table 1, our BAT achieves 0.920 in Dice and 0.858 in IoU, outperform Lee *et al.* by 0.2% and 1.5% in Dice and IoU, respectively. We extensively conduct experiments with other SOTA segmentation networks on the *ISIC 2018* dataset, including three famous convolutional models for segmentation (DeepLabv3 [6], UNet++ [23], CE-Net [9]) and two transformer-based network to address medical image segmentation (TransUNet [5], MedT [16]). Even compared with other state-of-the-art segmentation

Table 2. Experimental results on different datasets.

Trans.	BAG	ISIC 2016 + PH2		ISIC 2018	
		Dice ↑	*IoU* ↑	*Dice* ↑	*IoU* ↑
		0.884	0.805	0.879	0.810
✓		0.900	0.827	0.890	0.821
✓	✓	0.921	0.858	0.912	0.843

models, our BAT still achieves the consistent and significant improvement on both metrics. It's noteworthy that transformer-based network has superior performance than conventional CNNs, indicating the effectiveness of utilizing global context to detect skin lesion. In addition, compared with TransUNet, our method leveraging the boundary-prior knowledge significantly improves the segmentation performance (1.8% on Dice and 2.1% on IoU), proving that the combination of boundary information and transformer architecture is indeed helpful to segment target lesion.

Figure 3 visualizes five typical challenging cases of lesion segmentation results. It is observed that our results are closest to the ground truth, when compared with our competitors. The first three rows represent cases with various color, size and shape, and our BAT outperforms others with most stable segmentation performance, indicating the robust advancement of global context. The last two rows highlight some small regions of ambiguous boundary and it's shown that our BAT is capable of tacking such problems, due to the use of boundary-wise prior knowledge.

3.4 Ablation Study

We further conduct ablation studies to demonstrate the effectiveness of three major components in BAT: (1) the transformer-based self-attention mechanism (Trans.), (2) boundary-wise attention gate (BAG). As shown in the Table 2, by the incorporation of self-attention mechanism, the IoU increases by a large margin on both datasets. This result indicate that it's essential to integrate global context to improve the skin lesion detection. On the other hand, applying BAGs to guide transformer further improves the performance significantly, confirming the effectiveness of boundary-wise prior knowledge to tackling challenging cases, such as lesions with ambiguous boundary.

4 Conclusion

We present a novel and efficient context-aware network, namely boundary-aware transformer (BAT) network, for accurate segmentation of skin lesion from dermoscopy images. Extensive experiments on two public datasets confirm the effectiveness of our proposed BAT, to help yield much better segmentation results for skin lesions. Our full model outperforms state-of-the-art models by a large margin in segmentation accuracy and the intuitive visualization shows that our BAT has most satisfactory performance on skin lesions with ambiguous boundary.

References

1. Ahn, E., et al.: Automated saliency-based lesion segmentation in dermoscopic images. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3009–3012. IEEE (2015)

2. Bi, L., Kim, J., Ahn, E., Kumar, A., Fulham, M., Feng, D.: Dermoscopic image segmentation via multistage fully convolutional networks. IEEE Trans. Biomed. Eng. **64**(9), 2065–2074 (2017). https://doi.org/10.1109/TBME.2017.2712771
3. Bi, L., Kim, J., Ahn, E., Feng, D., Fulham, M.: Automated skin lesion segmentation via image-wise supervised learning and multi-scale superpixel based cellular automata. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pp. 1059–1062. IEEE (2016)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision, ECCV 2020. Lecture Notes in Computer Science, vol. 12346. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
5. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation (2021)
6. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: International Conference on Machine Learning, pp. 1243–1252. PMLR (2017)
9. Gu, Z.: Ce-net: Context encoder network for 2d medical image segmentation. IEEE Trans. Med. Imaging **38**(10), 2281–2292 (2019)
10. Lee, H.J., Kim, J.U., Lee, S., Kim, H.G., Ro, Y.M.: Structure boundary preserving segmentation for medical image with ambiguous boundary. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4816–4825 (2020). https://doi.org/10.1109/CVPR42600.2020.00487
11. Li, Z., Liu, X., Creighton, F.X., Taylor, R.H., Unberath, M.: Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. arXiv preprint arXiv:2011.02910 (2020)
12. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
13. Mathur, P., et al.: Cancer statistics, 2020: report from national cancer registry programme, India. JCO Glob. Oncol. **6**, 1063–1075 (2020)
14. Prangemeier, T., Reich, C., Koeppl, H.: Attention-based transformers for instance segmentation of cells in microstructures. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 700–707. IEEE (2020)
15. Tu, Z., Bai, X.: Auto-context and its application to high-level vision tasks and 3d brain image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **32**(10), 1744–1757 (2010). https://doi.org/10.1109/TPAMI.2009.186
16. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: gated axial-attention for medical image segmentation. arXiv preprint arXiv:2102.10662 (2021)
17. Vaswani, A., et al.: Attention is all you need (2017)
18. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arxiv 2015. arXiv preprint arXiv:1511.07122 615 (2019)
19. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 472–480 (2017)

20. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.: Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans. Med. Imaging **36**(4), 994–1004 (2017). https://doi.org/10.1109/TMI.2016.2642839
21. Yuan, Y., Chao, M., Lo, Y.C.: Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance. IEEE Trans. Med. Imaging **36**(9), 1876–1886 (2017)
22. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. arXiv preprint arXiv:2012.15840 (2020)
23. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1