GT U-Net: A U-Net Like Group Transformer Network for Tooth Root Segmentation

Yunxiang Li[1], Shuai Wang[2], Jun Wang[3], Guodong Zeng[4], Wenjun Liu[1], Qianni Zhang[5], Qun Jin[6], and Yaqi Wang[7](✉)

[1] Microelectronics CAD Center, Hangzhou Dianzi University, Hangzhou, China
[2] School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, China
[3] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[4] sitem Center for Translational Medicine and Biomedical Entrepreneurship, University of Bern, Bern, Switzerland
[5] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK
[6] Department of Human Informatics and Cognitive Sciences, Faculty of Human Sciences, Waseda University, Tokyo, Japan
[7] College of Media Engineering, Communication University of Zhejiang, Hangzhou, China
wangyaqi@cuz.edu.cn

Abstract. To achieve an accurate assessment of root canal therapy, a fundamental step is to perform tooth root segmentation on oral X-ray images, in that the position of tooth root boundary is significant anatomy information in root canal therapy evaluation. However, the fuzzy boundary makes the tooth root segmentation very challenging. In this paper, we propose a novel end-to-end U-Net like Group Transformer Network (GT U-Net) for the tooth root segmentation. The proposed network retains the essential structure of U-Net but each of the encoders and decoders is replaced by a group Transformer, which significantly reduces the computational cost of traditional Transformer architectures by using the grouping structure and the bottleneck structure. In addition, the proposed GT U-Net is composed of a hybrid structure of convolution and Transformer, which makes it independent of pre-training weights. For optimization, we also propose a shape-sensitive Fourier Descriptor (FD) loss function to make use of shape prior knowledge. Experimental results show that our proposed network achieves the state-of-the-art performance on our collected tooth root segmentation dataset and the public retina dataset DRIVE. Code has been released at https://github.com/Kent0n-Li/GT-U-Net.

Keywords: Image segmentation · Shape-sensitive loss · Group transformer · Root canal therapy

© Springer Nature Switzerland AG 2021
C. Lian et al. (Eds.): MLMI 2021, LNCS 12966, pp. 386–395, 2021.
https://doi.org/10.1007/978-3-030-87589-3_40

1 Introduction

In worldwide, approximately 743 million people are affected by severe periodontitis, which is considered the sixth most common health disorder [1]. Nowadays, root canal therapy is a routine periodontal treatment for periodontitis and an incorrect evaluation of the treatment result will impede timely follow-up [2,3]. Since the tooth root boundary is the significant anatomy feature for carrying out the evaluation, tooth root segmentation becomes the most important step of root canal therapy automatic evaluation. Unfortunately, performing an accurate tooth root segmentation is a very challenging task due to the following reasons: 1) the tooth root boundaries are blurry and some tissues around the teeth have similar intensities to the teeth, as shown in Fig. 1 (a); 2) other bones and tissues may overlap with the tooth root in the oral X-ray images, as shown in Fig. 1 (b); 3) the quality of the X-ray image may be very poor such as overexposed or underexposed, as shown in Fig. 1 (c).

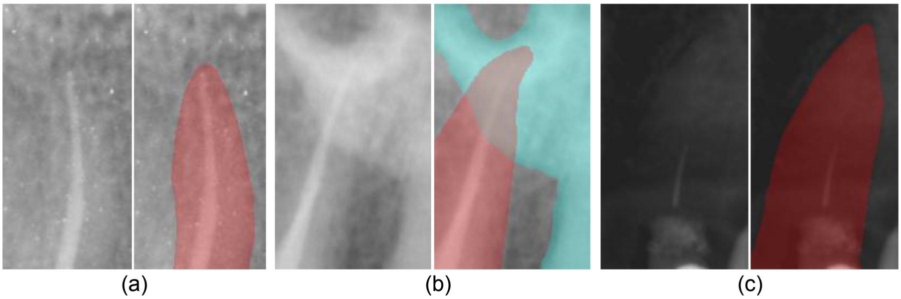

Fig. 1. Three examples of root canal therapy X-ray images. The left image in each group is the original image, and the right one depicts the tooth root and other tissues by red and cyan, respectively.

To address the problems mentioned above, Zhao et al. [4] provided a two-stage attention segmentation network, which can effectively alleviate the inhomogeneous intensity distribution problem by focusing on automatically catching the real tooth region. Lee et al. [5] adopted a fine-tuned mask R-CNN [6] algorithm to achieve the tooth segmentation. Nevertheless, these methods do not effectively solve the segmentation problem of fuzzy boundaries, and the performance improvement is mostly incremental. Chen et al. [7] proposed a novel MSLPNet with multi-scale structural similarity (MS-SSIM) loss, enhancing tooth segmentation that has fuzzy root boundaries. Cheng et al. [8] proposed U-Net+DFM to learn a direction field. It characterizes the directional relationship between pixels and implicitly restricts the shape of the segmentation result. Although these methods all achieve decent results in segmentation tasks, they are still limited by the intrinsic locality of convolutional neural networks (CNNs) and can not process global features very well. To relieve this problem, long-range dependencies via non-local operations are highly desired, and Transformer [9] provides a modeling pipeline to achieve that. Chen et al. proposed TransUNet [10], a Transformer-based encoder operating for segmentation, which adopts ViT [11] with 12 Transformer layers as the encoder. However, ViT relies on pre-trained weights obtained by a huge image corpus, which results in undesirable performance on insufficient datasets. To solve that, Aravind et al. [12] presented BoT-Net, an effective instance segmentation backbone, by combining transformer and convolution. Due to the high computational complexity of Transformer, BoTNet has only replaced a part of convolution in the last few layers of ResNet with Transformer.

In order to mitigate the problems existing in the present approaches, our network GT U-Net employs both the combination of convolution and Transformer without pre-training weights and the grouping structure and bottleneck structure that significantly reduces the amount of computation. Besides, FD loss also solves the problem of fuzzy boundary segmentation by making full use of shape prior knowledge. The main contributions of this paper are listed as follows:

(1) Our network retains the advantages of the general U-Net framework and introduces Transformer into the medical image segmentation application to solve the limitation of convolution.
(2) We design a grouping structure and a bottleneck structure, which greatly reduces the computation load of Transformer and makes it feasible in image segmentation.
(3) For the root segmentation task, we propose a shape-sensitive Fourier Descriptor loss function to deal with the problem of fuzzy boundary segmentation.

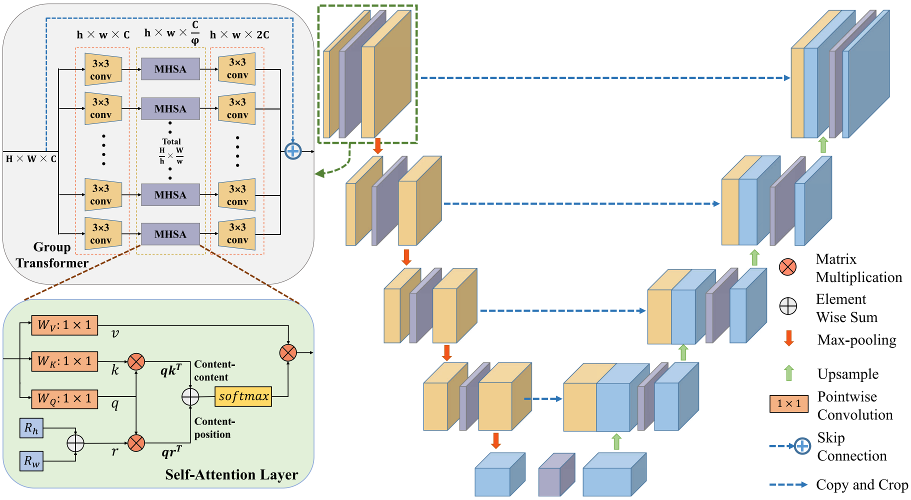

Fig. 2. The architecture of the proposed GT U-Net, which is composed of a U-shape general framework and group Transformer. Self-attention layer is the base structure of Multi-head self-attention (MHSA) where our MHSA has 4 heads, and we do not illustrate them on the figure for simplicity.

2 Method

GT U-Net follows the overall U-shaped structure, where both the encoders and decoders consist of group Transformer. It works with a shape-sensitive Fourier Description (FD) loss function for tooth root segmentation. The proposed method is described in full detail below.

2.1 U-Net Like Group Transformer Network

Overall Architecture: The architecture of GT U-Net is illustrated in Fig. 2. Specifically, our group Transformer is composed of a skip connection, a grouping module, 3 × 3 convolutions, Multi-head self-attention (MHSA) modules, and a merging module. Among them, the skip connection is used to solve the problem of gradient vanishing and keep the low-level information. Considering that MHSA requires $\mathrm{O}(n^2 d)$ memory and computation when performing globally across n entities, the grouping module and 3 × 3 convolutions are designed to reduce the amount of MHSA computation.

Grouping Structure and Bottleneck Structure: Transformer was first proposed for natural language processing (NLP) tasks and has become a hot topic in many computer vision tasks as a non-local operation with long-range dependence. However, it is difficult to apply Transformer directly to medical image tasks on account of the discrepancy between natural language and image. The number of words in natural language is limited, but the number of pixels increases quadratic with the increase of image size. Because of this, we designed group Transformer to solve the problem of too much computation due to the image characteristics in medical image segmentation. Assuming that the original feature block size is $H \times W \times C$, the computation of MHSA will be greatly reduced through the grouping structure and bottleneck structure. The calculation amount of MHSA before improvement [13] and that of our group Transformer (GT) are given in Eq. (1) and (2), respectively:

$$\mathit{\Omega}(MHSA) = 4HWC^2 + 2(HW)^2C \tag{1}$$

$$\mathit{\Omega}(GT) = 4hw(\frac{C}{\varphi})^2 + 2(hw)^2\frac{C}{\varphi} = \frac{4hw}{\varphi^2 HW}HWC^2 + \frac{2h^2w^2}{\varphi H^2W^2}(HW)^2C \tag{2}$$

where φ is the channel scaling factor of the bottleneck structure, and the size of each group Transformer unit is set to $h \times w$, which is determined by the input image size and task characteristics. The proposed method achieves local long-range dependencies at higher encoder and decoder layers. With the deepening of the network, the field of vision is gradually expanded to extract the junction features between different GT units and to achieve the global long-range dependencies.

Multi-head Self-attention: MHSA [9] is a type of attention mechanism which is more focused on the internal structure. The detailed architecture of MHSA is illustrated in Fig. 2, where the position encoding method is relative-distance-aware position encoding [14–16] with R_h and R_w for height and width. The attention logit is $qk^T + qr^T$, and q, k, v, r denote query, key, value, and position encodings, respectively.

Hybrid Structure of Convolution and Transformer: Transformer has the ability to extract global contexts, but it also has the limitation of lacking translation invariance while locality and translational equivariance are the basic property of convolution. Therefore, 3×3 convolution in our group Transformer is not only for the bottleneck structure, but it also helps construct a hybrid architecture where the convolution is responsible for feature extraction and Transformer is constructed to model the long-range dependencies.

2.2 Shape-Sensitive Fourier Descriptor Loss Function

From the anatomical prior knowledge, it is known that the tooth roots share a similar shape. Adding the shape information into the loss function helps better guide the model to segment tooth roots. (x_m, y_m) is one of the coordinates on a tooth root boundary that contains N pixels, and the boundary shape can be formed as a complex number $z(m) = x_m + jy_m$. The Fourier Descriptor [17] of this shape is defined as $DFT(z(m))$, that is Z(k) in Eq. (3):

$$Z(k) = DFT(z(m)) = \frac{1}{N}\sum_{m=0}^{N-1} z(m)e^{-j2\pi mk/N} \quad (k = 0, \cdots, N-1) \tag{3}$$

Fourier Descriptor is a quantitative representation of closed shapes independent of their starting point, scaling, location, and rotation. Therefore, the shape difference between the predicted boundary A and the manually labeled boundary B can be quantified, and it can be calculated based on Eq. (4). The original loss function is binary cross-entropy (BCE). In the proposed method, we add Fourier Descriptors to build a new shape-sensitive loss function. Considering their order of magnitude in the whole training process, we calculate the final Fourier Descriptor (FD) loss by Eq. (5):

$$\varDelta Z(k) = |Z_A(k) - Z_B(k)|. \tag{4}$$

$$FD\ loss = BCE(A, B) \times \frac{1}{1 + e^{-\beta \times \varDelta Z(k)}} \tag{5}$$

The novel FD loss pays equal attention to shape loss and BCE loss regardless of their order of magnitude. β is determined by the order of magnitude of $\varDelta Z(k)$, and it is set to 10 in this paper.

3 Experiment

3.1 Tooth Root Segmentation Dataset

We have built a new root canal therapy X-ray image dataset with patients' consent. The tooth root segmentation dataset contains 248 root canal therapy X-ray images from different patients in total. And three experienced stomatologists helped to complete the tooth root annotation. Specifically, to minimize inter-observer variability, the final annotation results needed the agreement of all three stomatologists.

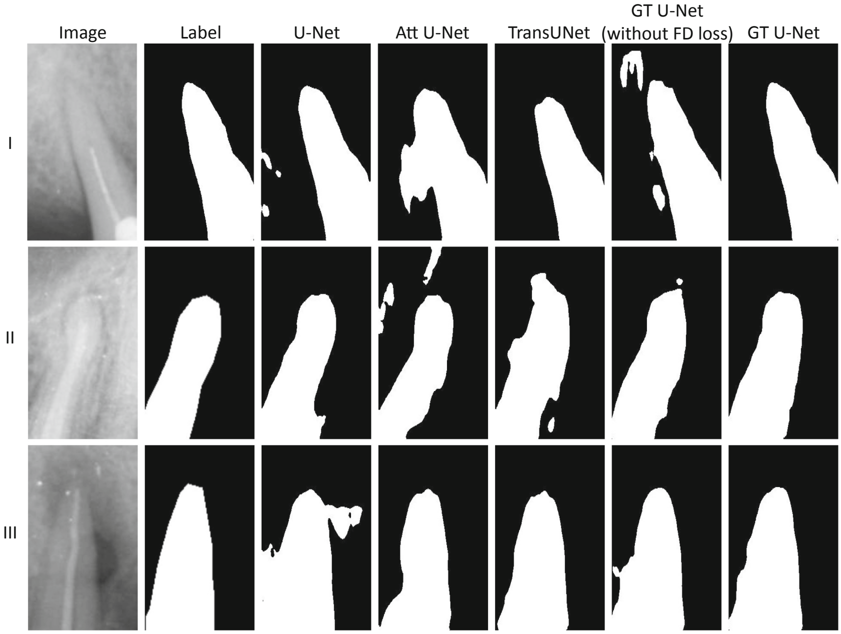

Fig. 3. Representative segmentation results achieved by different methods.

3.2 Implementation Details

In our implementation, three state-of-the-art methods are compared, U-Net [18], Attention U-Net [19], and TransUNet [10]. All compared methods and our method were implemented using PyTorch and trained on four RTX 2080Ti GPUs. Random center crop, random rotation, and axial flipping are adopted for the data augmentation to improve robustness and avoid overfitting. In the training, Adam optimizer is adopted and the initial learning rate and momentum are set to 2×10^{-4} and 0.5, respectively. The total number of training epochs is 200, and the batch size is 12. All the images are resized to 256×256 for input. $h \times w$ of the group Transformer unit is set to 8×8, and φ is set to 2.

To evaluate the segmentation performance, we compare the results with six metrics for evaluation, including Accuracy (ACC), Sensitivity (SE), Specificity (SP), Jaccard Similarity (JS), and Dice Coefficient (DICE). Besides, we adopt 3-fold cross-validation for testing the segmentation performance on the tooth root segmentation dataset.

Table 1. The evaluation performance of all methods on the Tooth Root Segmentation Dataset.

	U-Net	Att U-Net	TransUNet	**GT U-Net** (without FD loss)	**GT U-Net**
ACC (%)	92.81 ± 0.90	93.03 ± 0.56	93.41 ± 0.25	93.80 ± 0.32	**93.98 ± 0.11**
SE (%)	92.37 ± 0.68	89.36 ± 1.30	91.81 ± 0.74	**93.98 ± 0.38**	93.74 ± 1.18
SP (%)	93.74 ± 0.99	**95.60 ± 0.28**	95.05 ± 0.43	94.26 ± 0.35	94.79 ± 1.14
JS (%)	84.62 ± 1.30	83.85 ± 1.29	85.75 ± 0.65	86.63 ± 0.62	**86.87 ± 0.13**
DICE (%)	91.19 ± 1.01	91.15 ± 0.76	91.96 ± 0.40	92.32 ± 0.49	**92.54 ± 0.25**

3.3 Experimental Results

The average results with standard deviations of all methods are reported in Table 1. From this table, it can be summarized that our proposed GT U-Net achieves the best accuracy, sensitivity, Jaccard Similarity, and Dice Coefficient. In Fig. 3, we qualitatively compare the segmentation results on three oral X-ray images. It can be evidently seen that our method obtains the best segmentation results on the blurred boundaries. To verify the effectiveness of FD loss, we implement GT U-Net without FD loss. Based on the comparison in Table 1 and Fig. 3, FD loss can be effectively applied to the segmentation task with a fuzzy boundary but a similar shape by implicitly restricting the shape of the segmentation result with the anatomical prior knowledge.

Table 2. The segmentation performance achieved by all methods on the public DRIVE dataset.

	Year	ACC (%)	SE (%)	SP (%)	F1 (%)	AUC (%)
U-Net [18,20]	2015	95.55	78.22	98.08	81.74	97.52
MS-NFN [21]	2018	95.67	78.44	98.19	–	98.07
Residual U-Net [22,23]	2018	95.53	77.26	98.20	81.49	97.79
Recurrent U-Net [22,23]	2018	95.56	77.51	98.16	81.55	97.82
R2U-Net [22,23]	2018	95.56	77.92	98.13	81.71	97.84
DenseBlock-UNet [24]	2018	95.41	79.28	97.76	81.46	97.56
DUNet [20]	2019	95.58	78.63	98.05	81.90	97.78
ACE-Net [25]	2019	95.69	77.25	**98.42**	–	97.42
IterNet [23]	2020	95.73	77.35	98.38	82.05	**98.16**
GT U-Net	2021	**96.31**	**82.54**	98.24	**84.58**	98.02

3.4 Performance on the Public DRIVE Dataset

To further evaluate the performance of our GT U-Net, we applied it on the widely used retinal dataset DRIVE [26]. The dataset consists of 40 randomly selected color fundus retinal images of size 565 × 584. Officially, DRIVE is split into two equal groups for training and testing. The group Transformer unit parameters h and w are all set to 4, and φ is set to 2. Since FD loss is proposed for medical image segmentation with a similar shape, like tooth root segmentation, we did not implement it on the DRIVE dataset. To efficiently train our GT U-Net, we cropped image patches from the original images, and 64 × 64 is adopted as the sample patch size, which is widely adopted on this dataset.

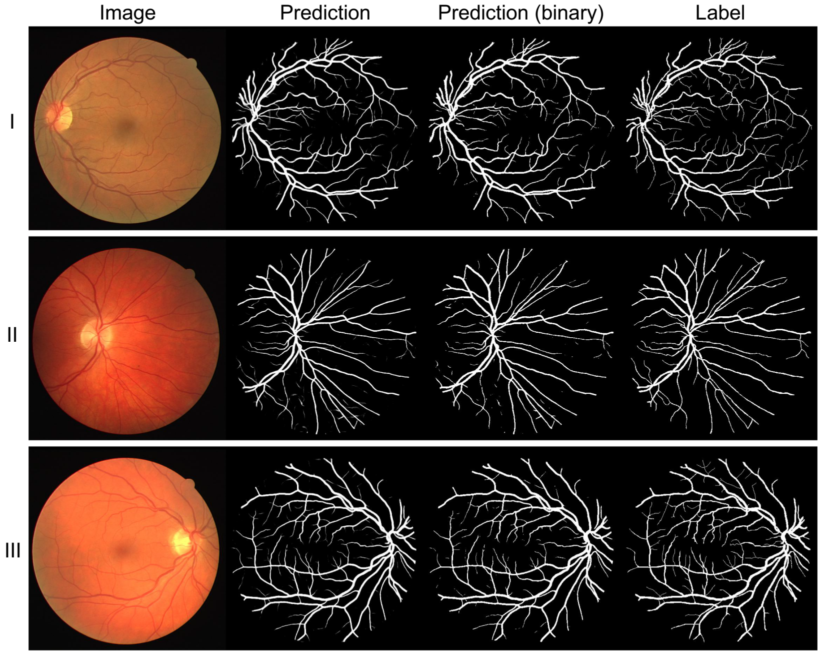

Fig. 4. Representative segmentation results of the public DRIVE dataset.

The comparison results are reported in Table 2. It can be observed that our GT U-Net outperforms other methods by at least 3.2% on SE, over 2.5% on F1, and over 0.5% on ACC. In terms of SP and AUC, our GT U-Net achieved competitive performance in comparison with the best methods, limiting the difference from them within 0.2%. Meanwhile, the visualization of some representative segmentation results is given in Fig. 4. It can be clearly observed that the vascular part of the retina can be segmented very well by our method, except for some very tiny areas.

4 Conclusion

In this paper, a novel end-to-end U-Net like group Transformer network (GT U-Net) for medical image segmentation is presented. GT U-Net uses the hybrid structure of convolution and Transformer to eliminate the need for pre-trained weights. Moreover, it significantly reduces the computational complexity of Transformer using a bottleneck structure and a grouping structure, and FD loss can be effectively applied to the segmentation task with a fuzzy boundary but a similar shape. Extensive experiments are conducted on tooth root and retinal vessel segmentation tasks, respectively, and the results show that the proposed GT U-Net has excellent segmentation performance and a great application perspective in a wide range of medical image segmentation tasks.

Acknowledgements. This work was supported by the National Key Research and Development Program of China (Grant No. 2019YFC0118404) and Public Projects of Zhejiang Province (Grant No. LGG20F020001).

References

1. Peres, M.A., et al.: Oral diseases: a global public health challenge. Lancet **394**(10194), 249–260 (2019)
2. Estrela, C., Holland, R., Estrela, C.R.D.A., Alencar, A.H.G., Sousa-Neto, M.D., Pécora, J.D.: Characterization of successful root canal treatment. Braz. Dent. J. **25**(1), 3–11 (2014)
3. Kaplan, T., Sezgin, G.P., Sönmez-Kaplan, S.: A survey study, Dental students' perception of difficulties concerning root canal therapy (2019)
4. Zhao, Y., et al.: TSASNet: tooth segmentation on dental panoramic x-ray images by two-stage attention segmentation network. Knowl.-Based Syst. **206**, 106338 (2020)
5. Lee, J.H., Han, S.S., Kim, Y.H., Lee, C., Kim, I.: Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs. Oral Surg. Oral Med. Oral Pathol. Oral Radiol. **129**(6), 635–642 (2020)
6. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
7. Chen, Q., et al.: MSLPNet: multi-scale location perception network for dental panoramic X-ray image segmentation. Neural Comput. Appl. **33**(16), 10277–10291 (2021). https://doi.org/10.1007/s00521-021-05790-5
8. Cheng, F., et al.: Learning directional feature maps for cardiac MRI segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 108–117. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_11
9. Vaswani, A., et al.: Attention is all you need. In: NIPS, pp. 6000–6010 (2017)
10. Chen, J.: TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
11. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

12. Srinivas, A., Lin, T.Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A.: Bottleneck transformers for visual recognition. arXiv preprint arXiv:2101.11605 (2021)
13. Liu, Z., et al. Swin transformer: hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
14. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 464–468 (2018)
15. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. arXiv preprint arXiv:1906.05909 (2019)
16. Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, Q.V.: Attention augmented convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019
17. Zahn, C.T., Roskies, R.Z.: Fourier descriptors for plane closed curves. IEEE Trans. Comput. **100**(3), 269–281 (1972)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
19. Oktay, O., et al.: Attention U-Net: learning where to look for the pancreas (2018)
20. Jin, Q., Meng, Z., Pham, T.D., Chen, Q., Wei, L., Su, R.: DUNet: a deformable network for retinal vessel segmentation. Knowl.-Based Syst. **178**, 149–162 (2019)
21. Wu, Y., Xia, Y., Song, Y., Zhang, Y., Cai, W.: Multiscale network followed network model for retinal vessel segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 119–126. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_14
22. Alom, M.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K.: Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation. arXiv preprint arXiv:1802.06955 (2018)
23. Li, L., Verma, M., Nakashima, Y., Nagahara, H., Kawasaki, R.: IterNet: retinal image segmentation utilizing structural redundancy in vessel networks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3656–3665 (2020)
24. Li, X., Chen, H., Qi, X., Dou, Q., Chi-Wing, F., Heng, P.-A.: H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans. Med. Imaging **37**(12), 2663–2674 (2018)
25. Zhu, Y., Chen, Z., Zhao, S., Xie, H., Guo, W., Zhang, Y.: ACE-Net: biomedical image segmentation with augmented contracting and expansive paths. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 712–720. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_79
26. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging **23**(4), 501–509 (2004)