CNN-GCN Aggregation Enabled Boundary Regression for Biomedical Image Segmentation

Anonymous Author

Anonymous Institute

Abstract. Accurate segmentation of anatomic structure is an essential task for biomedical image analysis. Recent popular object contours regression based segmentation methods attained researchers' attentions. They took a new starting point to tackle segmentation tasks instead of commonly used dense pixels classification methods. However, because of the nature of CNN based network (lack of spatial information) and the difficulty of this methodology itself (need more spatial information), these methods needed extra process to maintain more spatial features, which may cause longer inference time or tedious design and inference process. To address the issue, this paper proposes a simple, intuitive deep learning based approach. We develop a novel multi-level aggregated network to directly regress the coordinates of the contour of instances in an end-to-end manner. The proposed network seamlessly links convolution neural network (CNN) with Attention Refinement Module (ARM) and Graph Convolution Network (GCN). By iteratively and hierarchically fusing the features over different layers of the CNN, our approach obtains sufficient semantic information from the input image and pays distinct attention to the objects' contours with the help of ARM and GCN. Primarily, thanks to the proposed aggregation GCN and vertices sampling method, our model benefits from direct feature learning of the instances' boundary locations from sparse to dense and the spatial information propagation across the whole image. Experiments on the segmentation of fetal head (FH) in ultrasound images and of the optic disc (OD) and optic cup (OC) in colour fundus images demonstrate that our method outperforms state-of-the-art methods in terms of effectiveness and efficiency.

Keywords: Regression · Segmentation · GCN · Attention · Aggregation.

1 Introduction

The accurate assessment of anatomic structures in biomedical images plays an essential role in the management of many medical conditions or diseases. For instance, FH circumference in ultrasound images can be used to estimate the gestational age and to monitor the growth of the fetus [14]. Likewise, the size of the OD and OC in colour fundus images is of great importance for the diagnosis of glaucoma [19]. However, manual annotation is time-consuming and costly, so an automatic segmentation method is necessary.

The (biomedical) image segmentation task is a fundamental problem in the field of computer vision. The previous deep learning based methods [7,6,22,27] regarded segmentation as a pixel-wise classification problem and classified each pixel of an image into a class. Benefit from CNN's excellent ability to extract high-level semantic features, they attained good results in segmentation tasks. For example, M-Net [6] obtained state-of-the-art performance in OD & OC segmentation task, but it needed additional process, e.g. multi-scale input, ellipse fitting. In order to maintain enough semantic and spatial information from biomedical images, U-Net++ [27] proposed an aggregated CNN to fuse and reuse multi-level features across different layers, but it may result in unnecessary use of information flow. As some low-level features are unnecessarily over-extracted while object boundaries are simultaneously under-sampled. To address this issue, we apply the ARM working as a filter between CNN encoder and GCN decoder, which cooperates with the GCN to gain more useful and representative semantic and spatial features, especially the boundary location information from CNN. Instead of dense pixels classification, recent works [24,3,25] exploited object contours for efficient segmentation and achieved comparable performance with pixel-based segmentation methods. Specifically, DARNet [3] exploited the combination of Fully Convolution Networks (FCNs) [18] and Active Contour Models (ACMs) [16]. Nevertheless, it needs to initialize object contour and iteratively optimize to predict objects' contour during inference, which requires a relatively long running time. Other methods [24,25] represented object boundary with polar space coordinate, then regressed the distances between the centre point and the boundary points with CNN. Besides, they found that CNN cannot regress the Euclidean space coordinate representation of the boundary well, as some noise may be added, and the CNN may not maintain enough spatial information [24,25]. Our proposed aggregation GCN can handle this issue well, and our experiment results prove that. Besides, those methods' performance suffers from the low-quality of the centre point, so, Xie *et al.* [24] utilized centre sampling methods to predict and select high-quality centre points to improve the segmentation result. In contrast, our boundary representation method is not sensitive to the centre point as the boundary does not have too many correlations with the centre point.

Although the above methods achieved state-of-the-art performance on many segmentation tasks, it would inevitably necessitate sophisticated network design and tedious inference process. Conversely, inspired by the manner that clinicians annotate images, our method takes a simple but effective way to directly regress the boundary location without any further bells and whistles (multi-scale input, ellipse fitting, iterative optimization inference, centre selection process). In summary, this work has the following contributions:

- We take a simple and intuitive approach to (biomedical) image semantic segmentation and regard it as a vertex-wise boundary regression problem in an end-to-end fashion.
- We propose aggregated mechanisms on both CNN and GCN (with vertices sampling methods), to enable them to reuse and fuse the contextual and

spatial information. The additional attention mechanism helps the GCN decoder to gain more useful location information from the CNN encoder.
– We propose a new loss function suitable for object boundary localization, which helps prevent taking a large update step when approaching a small range of errors in the late training stage.

2   Method

2.1   Data Representation

The object boundaries are extracted from the binary mask image and uniformly divided into N vertices with the same interval $\Delta \theta$ (e.g. N $= 360$, $\Delta \theta = 1°$) . The geometric centre of the boundary is defined as the centre vertex. We represent the object boundary with vertices and edges as $B = (V, E)$, where V has $N + 1$ vertices in the Euclidean space, $V \in \mathbb{R}^{N \times 2}$, and $E \in \{0, 1\}^{(N+1) \times (N+1)}$ is a sparse adjacency matrix, representing the edge connections between vertices, where $E_{i,j} = 1$ means vertices V_i and V_j are connected by an edge, and $E_{i,j} = 0$ otherwise. In our work, every two consecutive vertices on the boundary and the centre vertex are connected to form a triangle. For the OD & OC segmentation, the OD and OC are divided separately while the centre of the OC is shared as the centre vertex. More details are shown in Fig. 1.
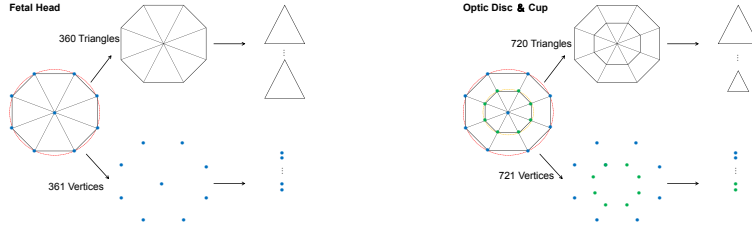

Fig. 1: Illustration of how object contours are represented to make it compatible for GCN, left: Fetal Head, right: Optic Disc & Cup. The boundary is represented by equally sampled vertices along it and its geometric center is defined as the center vertex. Each triangle consists of three vertices and three edges where two vertices are from the boundary and the other is the center vertex. Then, the vertices locations and their geometric relationships defined by an adjacency matrix from the triangulations can be used by GCN. For the optic disc (OD) and optic cup (OC) segmentation, the centre of the OC is shared as the centre vertex. However, triangulations are made for both the OD and OC.

2.2   Graph Convolution and Vertices Sampling

Following [4], the graph convolution used in this work is defined in Fourier space by formulating spectral filtering with a kernel g_θ using a recursive Chebyshev polynomial [4]. Furthermore, the polynomial order is set as 3.
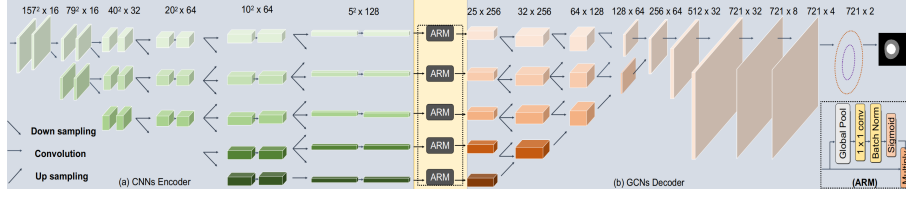
Fig. 2: Overview of our proposed network structure. The feature maps size of the CNN encoder and vertex maps of the GCN decoder for each stage (columns) are shown. In the CNN encoder, the horizontal arrow represents a CNN Residual Block [13] with kernel size 3 x 3, stride 1, followed by a Batch Normalization (BN) layer [15] and Leaky ReLU as the activation function. The down-sampling is conducted by setting stride size as 2. The lower level feature is bi-linearly up-sampled by a factor 2. In the GCN decoder, down-sampling and up-sampling are conducted by graph vertices sampling, which is described in Section 2.1. The horizontal arrow represents residual graph convolution (ResGCN) blocks [17] with polynomial order 3. In this figure, the example is for OD & OC segmentation, and for FH segmentation, the convolution operations are the same.

To achieve multi-scale aggregated graph convolutions on different vertex resolutions, insipred by [20], we employ the permutation matrix $Q_d \in \{0,1\}^{m \times n}$ to down-sample m vertices, m = 360 or 720 in our work. Q_d is obtained by iteratively contracting vertex pairs, which uses a quadratic matrix to maintain surface error approximations [9]. The down-sampling can be seen as a pre-processing, and the discarded vertices are recorded with barycentric coordinates so that the up-sampling can map the discarded vertices back with the same barycentric location information. The up-sampled vertices V_u can be obtained by a sparse matrix multiplication, i.e., $V_u = Q_u V_d$, where $Q_u \in \mathbb{R}^{m \times n}$ is another transformation matrix, V_d are down-sampled vertices. The up-sampling is applied during learning, and it operates convolution transformations on retained vertices.

2.3 Proposed Aggregation Network

Semantic Encoder: Fig. 2 (a) shows the structure of our image context encoder, which aims to lessen the location information loss and extract a wider spectrum of semantic features through different receptive fields. The encoder takes input images of shape 314×314×3 (Fundus OD & OC images) or 140×140×1 (Ultrasound FH images), with operations of up-sampling and down-sampling. Our encoder has six output features, for each, the shape is 5×5×128, and then those intermediate features will be as input to the attention refinement module. **Attention Module:** We propose an Attention Refinement Module (ARM) to refine the output features of each level (row) in the encoder, which easily integrates the global context information. As Fig. 2 (b) shows, ARM contains five attention blocks, and each block employs global channels average pooling to capture global context through the different channels, and computes an attention vector to guide the feature learning through a convolution layer followed by a

BN layer and sigmoid as the activation function. For the filter, the kernel size is 1×1, and the stride is 1.

Spatial Decoder: The decoder takes refined multi-paths outputs from the attention module, then decodes with ResGCN blocks [17] through different levels. Benefits from the graph sampling, our decoder can regress the location of the object boundary from sparse to dense. For each ResGCN Block, it consists of 4 graph convolution layers, followed by a BN layer [15] and Leaky ReLU as the activation function. After going through ResGCN blocks and graph vertices up-samplings, the number of vertices is up-sampled from 25 to 721, and each vertex is represented by a vector of length 32. At last, three graph convolution layers are added to generate 2D object contour vertices, which reduces the vertex feature map channels to 2, as each contour vertex has two dimensions: x and y.

2.4   Loss Function

We regard segmentation as a vertices location regression problem. L2 and L1 loss have been used in regression tasks by CNN based networks [11,12]. However, it is difficult for the L1 loss to continuously converge and find the global minimization in the late training stage without careful tuning of the learning rate. It is commonly known that the L2 loss is sensitive to outliers which may lead to unstable training in the early training stage.

Inspired by Wing-loss [5] and Smooth-L1 loss [10], we propose a new loss function (figures are shown in supplementary material) that can prevent the model from taking large update steps when reaching small range errors in the late training stage and can recover quickly when dealing with large errors during the early training stage. Our loss function is defined as:

$$L(x) = \begin{cases} W[e^{(|x|/\epsilon)} - 1] & if \quad |x| < W \\ |x| - C & otherwise \end{cases} \tag{1}$$

Where W should be non-negative and limit the range of the non-linear part, ϵ decides the curvature between $(-W, W)$ and $C = W - W[e^{(|w|/\epsilon)} - 1]$ connects the linear and non-linear parts. For the OD & OC segmentation tasks, we integrate a weight mask and assign more weights to the vertices that belong to the OC, to improve the OC segmentation performance, as the OC is usually difficult to segment due to the image quality or poor color contrast.

3   Experiments

3.1   Datasets

We evaluate our approach with two major types of biomedical images on two segmentation tasks respectively: fundus images of retinal for OD & OC segmentation, and ultrasound images of the fetus for FH segmentation.

Fudus OD & OC images: 2068 images from five datasets (Refuge [19], Drishti-GS [23], ORIGA [26], RIGA [1], RIM-ONE [8]) are merged together. 190 fundus

images are randomly selected as the retina test dataset, the rest 1878 fundus images are used for the training .Considering the negative influence of non-target areas in fundus retina images, we first localize the disc centers by detector [21] and crop to 314×314 pixels and then transmit into our network.

Ultrasound FH images: The HC18-Challenge dataset [1] [14], contains 999 two-dimensional (2D) ultrasound images with size of 800×540 pixels. We zero-padding each image to shape of 840×840, and then resize into 140×140 as the input image, then we randomly select 94 images as the test dataset, and the model is trained on the rest 905 images.

3.2   Implementation Details

To augment our dataset, we perturb the input image of training dataset by randomly rotating images for both segmentation tasks. Specifically, the rotation ranges from -15 to 15 degree. 10 % of training dataset are randomly selected as the validation dataset. We use stochastic gradient descent with a momentum of 0.9 to optimize our loss function. The number of graph vertices for OC & OD is sampled to 721, 512, 256, 128, 64, 32, 25 crosses seven stages with graph vertices sampling method. We trained our model 300 epochs for all datasets, with a learning rate of 1e-2 and decay rate of 0.997 every epoch. The batch size is set as 48. All training processes are conducted on a server with 8 TESLA V100 and 4 TESLA P100, and all test experiments are performed on a local machine Geforce RTX 2080Ti.

4   Results

In this section, we show our qualitative (Fig. 3) and quantitative results (Tab. 1) on the OD & OC segmentation and FH segmentation task. We compare our model with other state-of-the-art methods, including U-Net [22], PolarMask [24], M-Net [6], U-Net++ [27], DANet [7], DARNet [3], DeepLabv3+ [2] through running their open public source code. Dice score and Area Under the Curve (AUC) are used as the segmentation accuracy metrics.

Optic Disc & Cup Segmentation: We perform evaluation experiments on the retina test dataset, which is merged with five different fundus OD & OC images datasets. Fig. 3 and Tab. 1 show qualitative and quantitative results. We achieve 96.88 % and 92.46 % Dice score on OD & OC segmentation separately without any bells and whistles (multi-scale training, ellipse fitting, longer training epochs, etc.) As for the inference speed, our model achieves faster result with 66.6 milliseconds (ms) per image than PolarMask [24] (72.1 ms) and DARNet [3] (1239.2 ms). In the supplementary material, we also show some 'failed' cases compared with ground truth. According to the comments from an expert at anonymous accredited ophthalmology reading center, our model produces more accurate results than the ground truth.

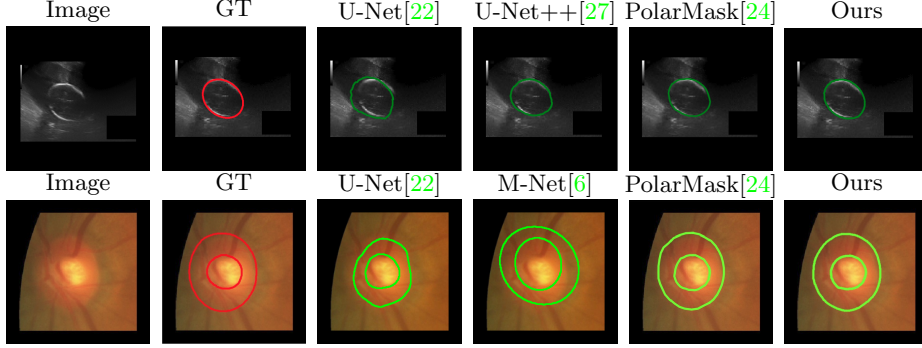[1] https://hc18.grand-challenge.org/

Fig. 3: Qualitative results of segmentation on retina test dataset and HC18-Challenge [14]. Top row is the ultrasound FH segmentation results, and the bottom row is the fundus OD & OC segmentation results. More qualitative results will be shown in supplementary material.

Tasks \ Methods	OC		OD		FH	
	Dice Score	AUC	Dice Score	AUC	Dice Score	AUC
U-Net [22]	0.9016	0.9186	0.9522	0.9648	0.9625	0.9688
M-Net[6]	**0.9335**	**0.9417**	0.9230	0.9332	-	-
U-Net++ [27]	0.9198	0.9285	0.9626	0.9777	0.9701	0.9789
DANet [7]	0.9232	0.9327	0.9654	0.9726	0.9719	0.9786
DARNet [3]	0.9235	0.9339	0.9617	0.9684	0.9719	0.9790
PolarMask [24]	0.9238	0.9366	0.9670	0.9782	0.9723	0.9780
Our method	0.9246	0.9376	**0.9688**	**0.9784**	**0.9738**	**0.9796**

Table 1: Segmentation results on retina test dataset for OD & OC and on HC18-Challenge [14] for FH. The performance is reported as Dice score (%) and AUC (%). The best result in each category are highlighted in bold.

Fetal Head Segmentation: Tab. 1 and Fig. 3 shows the quantitative and qualitative results respectively, our model achieves 0.9738 % Dice score and 0.9796 % AUC, which outperforms U-Net [22] by 1.2 % in terms of Dice score. Our model (60.2 ms) is faster than PolarMask [24] (65.5 ms) and DARNet [3] (1011.9 ms) for per image inference.

5   Ablation Study

We conduct several experiments on ablation studies to investigate the effectiveness of our model step by step. The ablation results on the parameter settings for proposed loss function and different loss function are shown in Tab. 2 and Fig. 4. Due to the page limitation, other (the network structure components, the interval angle for vertices sampling) ablation experiment results will be shown in supplementary material.

Parameters of Loss Function: We perform Experiments to evaluate the effect of parameter settings of our proposed loss function (Fig. 4). When w = 6, $\epsilon = 5$, our model achieve the best performance on OD & OC segmentation test dataset, and w = 8, $\epsilon = 5$, for FH segmentation test dataset.
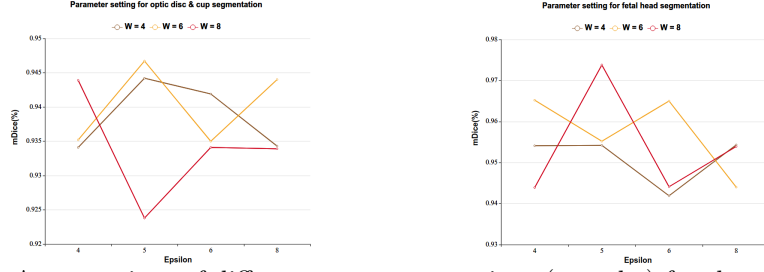
Fig. 4: A comparison of different parameter settings (w and ϵ) for the proposed loss function, measured with the mean Dice score (%) on retina test dataset for OD & OC. With w = 6, ϵ = 5, our model achieves the best performance (92.46 % & 96.88 %). On HC18-Challenge test dataset [14] for FH segmentation, with w = 8, ϵ = 5, our model gains the best results (97.38%). It shows that our network is not sensitive to these parameters as no significantly different results are found.

Tasks	OC		OD		FH	
Loss Function	Dice Score	AUC	Dice Score	AUC	Dice Score	AUC
L1	0.9108	0.9256	0.9543	0.9636	0.9503	0.9684
L2	0.9103	0.9208	0.9553	0.9668	0.9442	0.9571
Smooth-L1 [10]	0.9086	0.9112	0.9521	0.9652	0.9395	0.9452
Our proposed Loss						
weight mask = 0	0.9183	0.9218	0.9616	0.9738		
weight mask = 3	0.9223	0.9338	0.9646	0.9768		
weight mask = 5	**0.9246**	**0.9376**	**0.9688**	**0.9784**	**0.9738**	**0.9796**
weight mask = 7	0.9173	0.9238	0.9623	0.9718		
weight mask = 9	0.9109	0.9215	0.9603	0.9708		

Table 2: Performance comparisons (%) of different loss functions and weight mask settings on the OD & OC and FH segmentation respectively. For weight mask = 5, our model achieves best performance on the OD & OC segmentation.

Loss Function: We conduct experiments to evaluate the effectiveness of our proposed loss function. We compare with L1, L2, Smooth-L1 [10] loss functions. Tab. 2 shows the quantitative results on OD & OC and FH segmentation tasks respectively. As illustrated, our loss function achieves a mean Dice score that is 1.5 % relatively better than that of L1 loss function on OD & OC and 2.5 % relatively better than L1 loss function on FH segmentation. It also shows comparing with no-weight mask loss function, our proposed weight mask helps to improve OD & OC segmentation results by 0.7 % when weight mask = 5.

6   Conclusion

We propose a simple and intuitive regression methodology to tackle segmentation tasks by directly regressing the boundary of the instances instead of pixel-wise dense predictions. We have demonstrated its effectiveness and efficiency over other state-of-the-art methods, and its potentials on the segmentation problems of the fetal head and optic disc & cup. It is anticipated that our approach can be widely applicable to real world biomedical applications.

References

1. Almazroa, A., Alodhayb, S., Osman, E., Ramadan, E., Hummadi, M., Dlaim, M., Alkatee, M., Raahemifar, K., Lakshminarayanan, V.: Retinal fundus images for glaucoma analysis: the riga dataset. In: Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications. vol. 10579, p. 105790B. International Society for Optics and Photonics (2018)
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
3. Cheng, D., Liao, R., Fidler, S., Urtasun, R.: Darnet: Deep active ray network for building segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7431–7439 (2019)
4. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: Advances in Neural Information Processing Systems. pp. 3844–3852 (2016)
5. Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2235–2245 (2018)
6. Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., Cao, X.: Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE Transactions on Medical Imaging 37(7), 1597–1605 (2018)
7. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3146–3154 (2019)
8. Fumero, F., Alayón, S., Sanchez, J.L., Sigut, J., Gonzalez-Hernandez, M.: Rim-one: An open retinal image database for optic nerve evaluation. In: 24th International Symposium on Computer-based Medical Systems (CBMS). pp. 1–6. IEEE (2011)
9. Garland, M., Heckbert, P.S.: Surface simplification using quadric error metrics. In: Proceedings of the 24th Annual Conference on Computer graphics and interactive techniques. pp. 209–216. ACM Press/Addison-Wesley Publishing Co. (1997)
10. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1440–1448 (2015)
11. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 580–587 (2014)
12. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(9), 1904–1916 (2015)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
14. van den Heuvel, T.L., de Bruijn, D., de Korte, C.L., van Ginneken, B.: Automated measurement of fetal head circumference using 2D ultrasound images. PloS one 13(8) (2018)
15. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
16. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. International Journal of Computer Vision 1(4), 321–331 (1988)

17. Li, G., Müller, M., Thabet, A., Ghanem, B.: Can GCNs go as deep as CNNs? arXiv preprint arXiv:1904.03751 (2019)
18. 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)
19. Orlando, J.I., Fu, H., Breda, J.B., van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical Image Analysis **59**, 101570 (2020)
20. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3d faces using convolutional mesh autoencoders. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 704–720 (2018)
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. pp. 91–99 (2015)
22. 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)
23. Sivaswamy, J., Krishnadas, S., Joshi, G.D., Jain, M., Tabish, A.U.S.: Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In: 2014 IEEE 11th international symposium on biomedical imaging (ISBI). pp. 53–56. IEEE (2014)
24. Xie, E., Sun, P., Song, X., Wang, W., Liu, X., Liang, D., Shen, C., Luo, P.: Polarmask: Single shot instance segmentation with polar representation. arXiv preprint arXiv:1909.13226 (2019)
25. Xu, W., Wang, H., Qi, F., Lu, C.: Explicit shape encoding for real-time instance segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5168–5177 (2019)
26. Zhang, Z., Yin, F.S., Liu, J., Wong, W.K., Tan, N.M., Lee, B.H., Cheng, J., Wong, T.Y.: Origa-light: An online retinal fundus image database for glaucoma analysis and research. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. pp. 3065–3068. IEEE (2010)
27. Zhou, Z., Siddiquee, M.M.R., 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)