Spatial Constrained Mask R-CNN for Optic Disc and Cup Segmentation

Yating Zhou, Yanfeng Zhou, Yaoru Luo, Jia He,
Yudong Zhang, Ge Yang, and Yuanhao Guo

National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences
yuanhao.guo@ia.ac.cn

Abstract. In this work, we develop a deep learning-based pipeline for the task of optic disc/cup segmentation in retinal fundus images. In addition, based on the segmentation results, our method automatically diagnoses glaucoma and localizes the fovea center. In our method, we first propose a spatial constrained Mask R-CNN to segment the optic disc and cup. Considering the spatial constraint that the optic cup always locates inside the optic disc, we devise the detection branch of Mask R-CNN to detect the region of optic disc and we devise the mask branch to segment both the optic disc and cup within the detected optic disc region. Then, according to the centroid of the detected optic disc, we crop multiple square regions enclosing the optic disc which are used to train classification models for recognizing glaucoma. In the end, for the fovea localization, we estimate an approximate location of the fovea according to the positional relationship between the optic disc center and the fovea center. We crop the region of interest surrounding the fovea and adapt the DenseNet to perform a regression for finetuning an accurate location of fovea center. From validation, the dice coefficient of optic disc and cup segmentation from our method has achieved 96.4% and 87.4% respectively; The AUC of glaucoma classification achieves 96.2%; And the mean squared error of fovea localization is 11.7-pixel.

Keywords: Glaucoma Screening, Optic Disc/Cup Segmentation, Fovea Localization, Spatial Constrained Mask-RCNN.

1 Introduction

Glaucoma is a common ocular disease that may result in blindness if early clinical treatment is not introduced [1]. This disease is caused by optic neuropathy/apoptosis and optic disc atrophy. The early symptoms of glaucoma are relatively difficult to recognize and detect. Irreversible visual damage often occurs before an effective treatment. Therefore, early diagnosis and treatment for glaucoma are very important. At present, there are several possible methods to detect glaucoma. The first method refers to an expert clinical diagnosis which measures intraocular pressure or assesses the optic nerve head (ONH). The second method is an expert automated system based on machine learning which predicts the possibility of the occurrence of glaucoma only using the retinal fundus images. In this report, we focus on the latter method and propose a deep learning-based pipeline to automatically segment the regions of optic disc/cup

(OD & OC), recognize the glaucoma, and localize the fovea center. Specifically, we propose an improved version of Mask R-CNN [2] which takes advantage of the special constraints between OD and OC. Our spatial constrained Mask R-CNN only detects the region of OD as we find that OC is always located within OD. We then devise the mask branch of Mask-RCNN to segment both OD and OC in the detected local region of OD. Due to this spatial constraint, the detection accuracy of OD is improved and the segmentation of implicit OC region becomes easier accordingly. Based on the detected region, we train classification models to recognize glaucoma according to the appearance and shape variation of OD and OC. We also use the detected OD center to approximate the location of fovea. For an accurate estimation of the fovea center, we adapt a deep neural network as a regression model to finetune the fovea localization given an image patch around the fovea.

2 Methodology

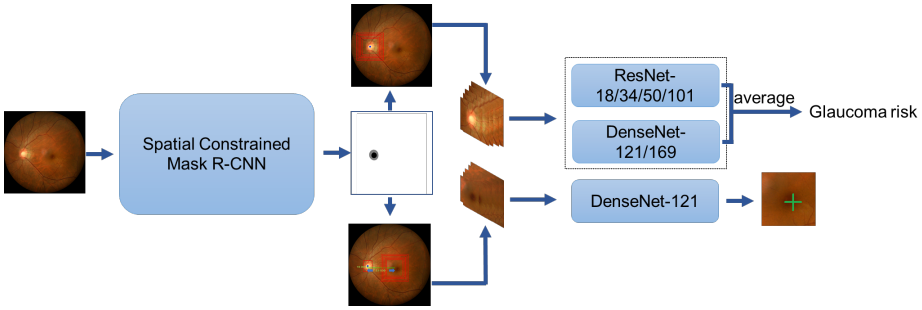

Fig. 1. The proposed pipeline for OD&OC segmentation, glaucoma classification and fovea localization.

Fig. 1 shows the whole pipeline of our method. First, we use the spatial constrained Mask R-CNN to segment OD and OC. Second, according to center of the detected region of OD, we crop multiple squared regions enclosing OD. We use the ResNet [3] and DenseNet [4] with various depths for glaucoma classification. The final prediction result is obtained by an ensemble the results of these models. Third, we use the positional relationship between OD center and the fovea to estimate the location of fovea, and then we crop multiple regions of interest surrounding the estimated fovea and adapt DenseNet-121 to perform a regression for an accurate localization of fovea center. We use fovea ROI to denote these regions of interest surrounding the fovea in the remaining sections.

2.1 OD and OC segmentation

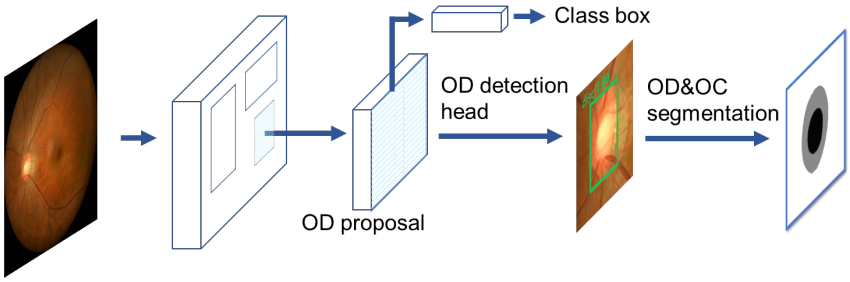

Fig. 2. Spatial Constrained Mask R-CNN Model Architecture.

In the segmentation of OD and OC, a generic segmentation model like UNet [19] is challenged, because (1) the OD and OC region occupies a very limited region of the whole image and (2) the intensity of OC region is usually quite high which results in unclear boundaries to distinguish OC. The former prevents an effective localization of the region of interest, which result in an inefficient OD and OC segmentation; the latter makes it difficult to accurately segment the OC.

As a result, a straightforward pipeline may consist of a detection and a segmentation. However, this structure is redundant because both tasks share many common features of the object. Instead, we can use an instance segmentation, like the Mask-RCNN, for an efficient implementation of both detection and segmentation. The Mask-RCNN is a typical instance segmentation method which simultaneously localizes and segments the targets of multiple-class. It is a two-stage model. In the first stage, Region Proposal Network (RPN) [6] is used to generate a series of region proposals each of which may contain an object candidate regardless its specific category. In the second stage, the detection branch predicts a fine-grained class of the object in each region proposal and simultaneously finetunes its bounding-box; the mask branch predicts a binary-class mask which segments the foreground (the object).

However, in some retinal funds images with glaucoma, the cup-to-disc ratio (CDR) is relatively large. This results in a high overlap between OC and OD and leads the Mask R-CNN to easily mis-identify OC as OD, yielding incorrect detection of the OD. Such localization result will produce incorrect segmentation of OD and OC, because parts of the real region of OD is excluded and the segmentation of OC will be totally incorrect because its label is assigned as OD in the detection branch.

To solve these problems, as shown in Fig.2, we propose a spatial constrained Mask-RCNN. We have observed that the OC is always located inside the region OD. Based on this spatial constraint, we devise the detection branch of Mask R-CNN to only detect the region of OD and the mask branch to segment both OD and OC. This architecture eliminates the influence of some inaccurate detection bounding box on OC segmentation, because the detected region is always used as OD and the segmentation of OD and OC is performed within this region. The segmentation of OC is improved because we avoid a naive ignoring of this region in previous methods.

2.2 Glaucoma classification

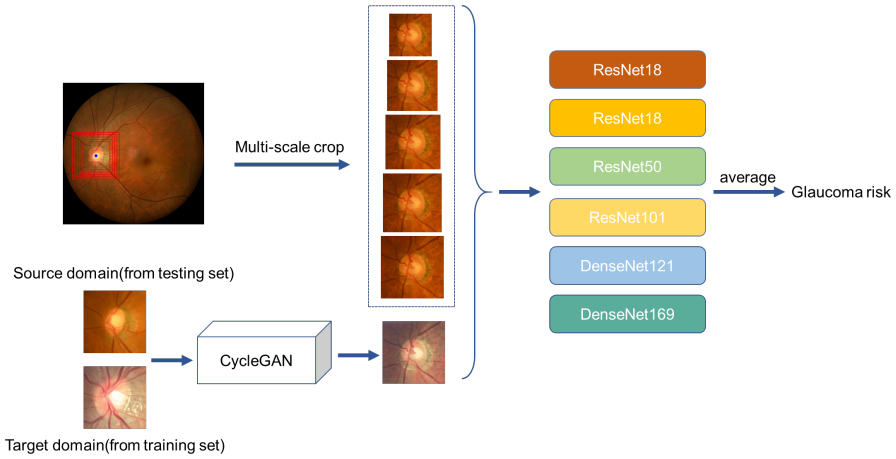

Fig. 3. The Pipeline of Glaucoma Classification.

The machine learning-based methods for glaucoma screening using images include the following two categories. The first type is using global features of the whole image as clues and the second is only using the features from the ONH region. Studies have shown that using local ONH region for glaucoma screening can achieve better performance, because ONH is the one that is mostly affected by glaucoma and focusing on the ONH region is conducive to better use of model parameters [5]. In our method, we choose to use ONH region to classify glaucoma.

We show the classification pipeline in Fig. 3. According to the estimated center of OD, we crop multiple regions with various scales enclosing the OD (384x384, 416x416, 448x448, 480x480, 512x512 pixels). We do this to solve the problem of inconsistent object size caused from inconsistent image size from the training and validation set. It should be noted that there exists a domain shift from our training data to the validation (testing) data. This will deteriorate the performance of our classification model due to the inconsistent color distribution of the region of interest. Therefore, we use a generative model, like the CycleGAN [7] to map the source image (validation data) to the target domain (training data). This method transfers the style of training set to the validation set (including testing set) which aligns the distribution of training and validation data. As for the classification, we devise an ensemble strategy to train the ResNet18, ResNer34, ResNet50, ResNet101, DenseNet121 and DenseNet169 separately and finally average the results of the models which obtain the top performances on the validation set. This ensemble strategy takes the advantage of data generalization of diverse models and accordingly improves the classification accuracy.

2.3 Fovea localization

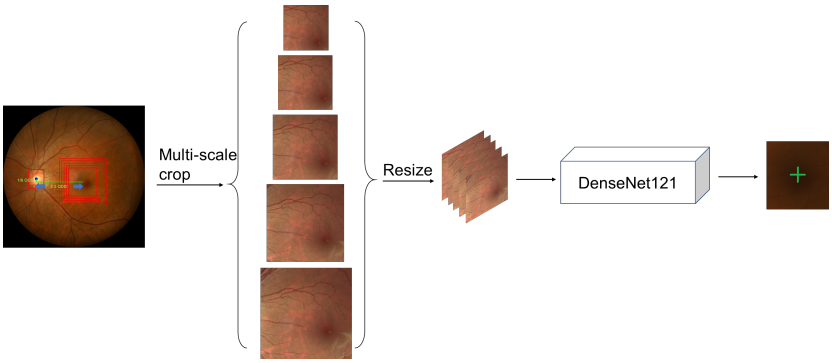

Fig. 4. Pipeline of Fovea Localization Using Regression.

Fovea is defined as the center of the macula. The whole retinal fundus image includes many details that does not contribute to the fovea localization, so it is difficult to estimate the accurate location of the fovea center directly through the full-size image. We note that the anatomical relationship between optic disc and fovea has been proved to be informative for a coarse fovea localization in many studies [8][9]. In our method, we also use this anatomical relationship [10]. Fig. 4 shows the pipeline of our fovea localization method. First, we obtain the center of OD based on the detection result. Second, starting from the OD center, we approximate the fovea region by searching the region directing down of 1/6 OD diameter (ODD) and right/left of 2.5 ODD (right eye and left eye can be distinguished.). Finally, we crop multi-scale squared fovea ROI (384x384, 448x448, 480x480, 512x512, 576x576, 640x640 pixels), and resize them into 512x512 pixels. We use all these image patches that include the fovea center at various locations to train the deep learning network for finetuned localization. We use multi-scale crop considering the diverse macular size. As for the localization model, we adjust DenseNet121 to perform a fovea coordinate regression. Specifically, we replace the last Softmax layer of the original network with a two-dimensional (2D) fully connected layer to predict the accurate 2D coordinates of the fovea center. During testing, we also crop multiple fovea ROI with various scales and take the average of prediction results.

3 Experiments

3.1 Dataset

In the challenge of REFUGE2, the training set consists of three splits: training split, validation split, and testing split. We use 18train, 18val, and 18test to denote these splits respectively. Each split separately includes 400 images. The image size of the 18train is 2124x2056 pixels, and the image size of 18val/test is 1634x1634 pixels. In order to improve the performance of our method, we use extra available training data. For segmentation task, we use 551 images from the RIGA dataset [11]. For fovea localization

task, we use 223 images from the IDRiD dataset [12]. As for the validation set of the REFUGE2, there are 400 images without any annotations. We use 20val to denote this dataset in the remaining sections. In order to align the image size of the training set, we resize all the images in the training set and our extra training images into 1634x1634 pixels, with zero padding on the shorter side to remain its original ratio.

3.2 Data pre-processing and augmentation

In order to augment the diversity of the data set and alleviate the imbalance between glaucoma and non-glaucoma data, we use a variety of data augmentation strategies. We use geometric transformation methods including horizontal flip, vertical flip, random rotation and scaling. We also use color transformation including noise, blur, sharpening, RGB shift, HUE saturation conversion, CLAHE [13], and brightness contrast conversion.

For classification, the domain shift from training data to validation data may deteriorate the generalization of the model. So, we introduce the unsupervised model, the CycleGAN to process the validation (testing) images as the style of training images because this method is suitable for un-paired training data. In Fig. 5, we show the results of this data pre-processing. In top row, we show the original image examples from 20val and in bottom row, we show the corresponding results processed by our style transfer model. We can see that the generated images of 20val are apparently close to the training data. This effect will benefit the classification model.

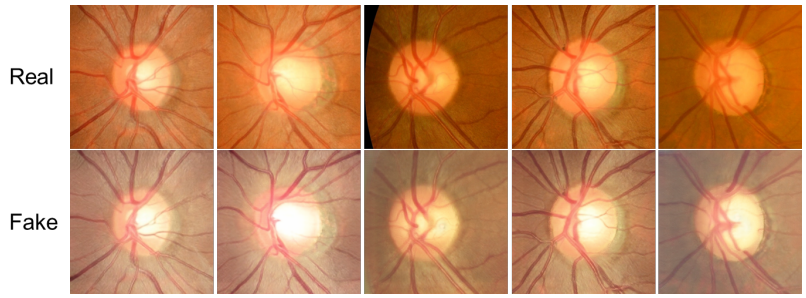

Fig. 5. Image examples of 20val before and after style transfer. W use the CycleGAN to map the 20val image to the domain of 18test. Top row: original images. Bottom row: style transferred image.

3.3 Implementation details

Training implementation: We use python and PyTorch to implement our whole method. All models are optimized using stochastic gradient descent (SGD). Our firmware to train the models consists of 4 NVIDIA 2080Ti GPUs.

For segmentation task, we implement our model based on the open-source library MMDetection [14]. We train multiple models with different backbone networks including ResNet50, ResNet101, and ResNeXt101 [17]. The model is pre-trained on ImageNet [15]. We finetune the model by setting batch size as 2. We set the learning rate as 0.01 which is divided by 10 at epoch 8 and 11.

For classification task, we use standard Pytorch implementation [20]. The models are also pre-trained on ImageNet, and we set batch size as 32. We use BCE loss to train the model. We set the maximum training epochs as 10.
For fovea localization task, we train the DenseNet121 from scratch. We set batch size as 60 and use wing loss [16] as the loss function. We set the initial learning rate as 0.01 and decrease it to 80 percent of the original value for every 4 epoch until reaching maximum 60 epochs.
Testing implementation: For segmentation task, we use the same data augmentation strategies as used in training which results in 40 images for a single image. Considering the 3 different models, we finally obtain in total 120 predictions for each testing image. The final prediction result is obtained by voting of the total predictions.
For classification task, we first obtain the center of OD from the segmentation result from last step. We then crop 5 squared regions with diverse scales surrounding the OD. We next use the CycleGAN to transfer the style of these images. We input these images into the 6 classification networks mentioned in previous sections. We average the prediction results to obtain the final prediction probability of glaucoma.
For localization task, we also obtain the OD center based on the segmentation result of the spatial constrained Mask R-CNN. We then use the positional relationship between OD and the fovea center to crop the fovea ROI of 6 scales. After resizing these images to 512x512 pixels, we use our trained regression model to predict the coordinates of the fovea center. We finally average all the prediction coordinates for an accurate estimation of fovea localization.

3.4 Results

For OD&OC segmentation, we use dice coefficient and CDR to evaluate the performance of our method. We compare the performance of the proposed spatial constrained Mask R-CNN (SC-Mask-RCNN) with the original Mask R-CNN and a generic segmentation model, the Deeplabv3+ [18]. Table 1 shows the results on 20val. When using the same backbone network, our spatial constrained Mask R-CNN achives the best performance measured using the three metrics. OC segmentation has been improved from 86.01% to 86.87% which gains an increase of 0.86%. This shows that our model solves the mis-classification problem of OC in previous instance segmentation methods. The OD dice and CDR of our method are slightly improved compared to Mask R-CNN, because the OD segmentation is relatively easy considering its clear appearance to the background. Moreover, comparing the results of spatial constrained Mask R-CNN with different backbone networks, we can see that a deeper model may result in better performance in this segmentation task. It should be noted that the ensemble of the results of the three models obtains an obvious improvement compared each of the model. This is because different model is equipped with various generalization ability on different data. Finally, the dice coefficient of the ensembled spatial constrained Mask R-CNN has achieved 96.4% and 87.4% for OD and OC segmentation. So, we will use this ensembled model in final competition.

Table 1. Segmentation Results Comparison on 20val.

Network	Backbone	OD dice	OC dice	CDR
DeepLabV3+	ResNet50	0.9336	0.8476	0.05296
Mask R-CNN	ResNet50	0.9620	0.8601	0.04256
SC-Mask R-CNN	ResNet50	0.9624	0.8687	0.04243
SC-Mask R-CNN	ResNet101	0.9624	0.8693	0.04193
SC-Mask R-CNN	ResNeXt101	**0.9626**	**0.8710**	**0.04152**
Ensembled SC- Mask R-CNN	--	**0.9644**	**0.8743**	**0.03906**

Table 2. Classification Results on 20val.

Network	20val-512	20val-multi-scale	20val-CycleGAN
ResNet18	93.87	93.45	93.34
ResNet34	**95.68**	**94.93**	**94.16**
ResNet50	**94.84**	**94.58**	**94.14**
ResNet101	94.24	93.96	93.63
DenseNet121	93.04	93.36	93.28
DenseNet169	92.93	93.17	92.56

In Table 2, we show the experimental results of our classification models on 20val. We use 20val-512 to represent the result obtained only using a 512x512 pixels local patch enclosing the OD for a testing image. We use 20val-multi-scale to represent the result obtained using multiple image patches with 5-scale regions enclosing the OD (384x384, 416x416, 448x448, 480x480, 512x512 pixels) for a testing image. We then obtain the final prediction by averaging the prediction results of these 5-scale image patches. We use 20val-CycleGAN to represents the result that obtained on the testing images generated by CycleGAN only using 512x512 image patch. Here, we have multiple classification models and multiple input image patches, so we ensemble the results of these models and data. Finally, we can obtain the AUC of 96.2% for glaucoma classification. We can see that this final ensembled prediction result outperforms the performance of each model, which shows that such ensemble strategy integrates the generalization ability of each model on different data.

In fovea localization task, we conduct an ablation study. Table 3 shows the experimental results on val20. We have the following observations. (1) The fovea localization model trained using the full-size image shows a high prediction error. This is because the fovea occupies only a small region of the image and the full-size image introduces too much noise for the estimation of fovea center. (2) The image processing of CLAHE shows its advantage in this fovea localization task. (3) The fovea localization using image patches which only contains the whole region of macula obtains a large margin performance improvement. (4) One can find that the multi-scale crop further improves the model performance. This is because the initial estimation of the fovea center is approximated by the positional relationship between the OD center and fovea center, which may introduce very large variations. Such multi-scale crop enhances the

generalization ability of the model on this data diversity. (5) When using extra traing data, our method achieves the best performance which obtains 11.7 pixels mean squared error (MSE) for fovea localization.

Table 3. Ablation Study of Fovea Localization on 20val.

Training Data source	Fovea ROI extracted	Multi-scale crop	CLAHE	MSE
REFUGE2 training set	×	---	×	93.1325
REFUGE2 training set	×	---	√	80.1022
REFUGE2 training set	√	×	√	25.1341
REFUGE2 training set	√	√	√	**17.9296**
REFUGE2 training set+IDRiD	√	√	√	**11.6992**

4 Conclusions

In this report, we propose a spatial constrained Mask R-CNN, which uses the spatial constraints between OD and OC to simultaneously segment OD and OC according to the OD detection results. Our method improves the segmentation accuracy of OC without affecting the segmentation accuracy of OD, thereby improving the accuracy of the CDR. For the classification and localization tasks, we have adapted effective data augmentation methods according to the characteristics of the task and used suitable model ensemble strategies to obtain accurate prediction results. We use standard deep neural network architectures for the classification and localization tasks, though. For future work, we will study more detailed pathological knowledge related to glaucoma, making full use of the correlation from fovea position, OD/OC shape to glaucoma screening.

References

1. Khalil T, Khalid S, Syed A M. Review of machine learning techniques for glaucoma detection and prediction. In: IEEE Science and Information Conference. 2014, pp. 438-442.
2. He K, Gkioxari G, Dollár P, et al. Mask R-CNN. In: IEEE international conference on computer vision. 2017, pp. 2961-2969.
3. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition. 2016, pp. 770-778.
4. Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. In: EEE conference on computer vision and pattern recognition. 2017, pp. 4700-4708.
5. Orlando J I, Fu H, Breda J B, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical image analysis, 2020, 59: 101570.
6. Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. 2015, pp. 91-99.

7. Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE international conference on computer vision. 2017, pp. 2223-2232.
8. Son, J., et al. Classification of findings with localized lesions in fundoscopic im-ages using a regionally guided CNN. Computational Pathology and Ophthalmic Medical Image Analysis. 2018: 176-184.
9. de Moura Garcia P, Kreutz D L, Welfer D. A novel method for detecting the fovea in fundus images of the eye. In: IEEE Conference on Graphics, Patterns and Images. 2016, pp. 104-111.
10. Huang Y, Zhong Z, Yuan J, et al. Efficient and robust OD detection and fovea localization using region proposal network and cascaded network. Biomedical Signal Processing and Control, 2020, 60: 101939.
11. Almazroa A, Alodhayb S, Osman E, et al. Retinal fundus images for glaucoma analysis: the RIGA dataset. Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications. 2018, 10579: 105790B.
12. Porwal P, Pachade S, Kamble R, et al. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. Data, 2018, 3(3): 25.
13. Setiawan A W, Mengko T R, Santoso O S, et al. Color retinal image enhancement using CLAHE. In: International Conference on ICT for Smart Society. 2013, pp. 1-3.
14. Chen K, Wang J, Pang J, et al. Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155, 2019.
15. Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition. 2009, pp. 248-255.
16. Feng Z H, Kittler J, Awais M, et al. Wing loss for robust facial landmark localisation with convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 2235-2245.
17. Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks. In: IEEE conference on computer vision and pattern recognition. 2017, pp. 1492-1500.
18. Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.
19. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.
20. Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems. 2019: 8026-8037.