Mathematical Biosciences and Engineering

MBE, 20(1): 1058–1082.
DOI: 10.3934/mbe.2023049
Received: 30 August 2022
Revised: 12 October 2022
Accepted: 16 October 2022
Published: 24 October 2022

http://www.aimspress.com/journal/MBE

Research article

A robust and high-precision edge segmentation and refinement method for high-resolution images

Qiming Li* and Chengcheng Chen

College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

* **Correspondence:** Email: qmli@shmtu.edu.cn; Tel: +8602138282823.

Abstract: Limited by GPU memory, high-resolution image segmentation is a highly challenging task. To improve the accuracy of high-resolution segmentation, High-Resolution Refine Net (HRRNet) is proposed. The network is divided into a rough segmentation module and a refinement module. The former improves DeepLabV3+ by adding the shallow features of 1/2 original image size and the corresponding skip connection to obtain better rough segmentation results, the output of which is used as the input of the latter. In the refinement module, first, the global context information of the input image is obtained by a global process. Second, the high-resolution image is divided into patches, and each patch is processed separately to obtain local details in a local process. In both processes, multiple refine units (RU) are cascaded for refinement processing, and two cascaded residual convolutional units (RCU) are added to the different output paths of RU to improve the mIoU and the convergence speed of the network. Finally, according to the context information of the global process, the refined patches are pieced to obtain the refined segmentation result of the whole high-resolution image. In addition, the regional non-maximum suppression is introduced to improve the Sobel edge detection, and the Pascal VOC 2012 dataset is enhanced, which improves the segmentation accuracy and robust performance of the network. Compared with the state-of-the-art semantic segmentation models, the experimental results show that our model achieves the best performance in high-resolution image segmentation.

Keywords: high-resolution semantic segmentation; global process and local process; edge refinement; Sobel operator; cascading method; data augmentation

1. Introduction

With the rise of artificial intelligence and computer vision, research on processing visual media such as images and videos is also advancing. Among them, semantic segmentation is one of the important tasks. In recent years, with the development of image acquisition equipment, the pixel resolutions of the captured images are getting higher and higher, reaching 4K UHD (3840*2160) or even higher. Because it contains more information and features than low-resolution pictures, high-resolution pictures (over 1920*1080) are more useful and can obtain more accurate results. Therefore, the task of high-resolution semantic segmentation has received more and more attention and has been widely used in satellite remote sensing images [1], medical imaging [2,3], autonomous driving [4,5], security monitoring [6], microscopic imaging [2] and other fields that require high precision.

The goal of image segmentation is to understand the scene and content of an image. Most of the methods currently applied to semantic segmentation are based on the idea of encoder-decoder. After the image undergoes a series of feature extraction operations in the encoder part, a feature map with a lower resolution is generated; the image will contain contextual information, but lack information about edge details, which has an impact on the accuracy of segmentation edges. The decoder part is to take an up-sampling operation on the image and combine the features of the intermediate layer to generate a more accurate segmentation map. Because the information lost in the process of down-sampling and then up-sampling is irreversible. Therefore, for high-resolution image segmentation, determining how to reduce the loss of detail information to improve the segmentation refinement ability has become a challenge. However, semantic segmentation is a pixel-level classification, which requires high GPU computing power. Most semantic segmentation models cannot directly process 4K or even higher resolution images, but often use down-sampling or blocking [7] the input images. However, the former will cause the loss of detailed information, and the latter will lose contextual information, both of which will affect the segmentation accuracy.

At present, the methods of improving the segmentation accuracy mainly include: increasing the receptive field [8–10], such as pooling and dilated convolution [8,9] operations; enhancing the capability of edge detail extraction [3,5,7,9–11], such as adding skip connections [3,5,9,10]; and improving the labeling accuracy of the dataset, such as BIG [2], ISIC [3], DeepGlobe [12], etc. But these methods lose performance for 4K or even higher resolution images because no GPU can process such a high-resolution image one time now. Therefore, inspired by GLNet [7] and CascadePSP [2], the HRRNet is proposed in this paper. Most of the existing high-resolution segmentation methods adopt down-sampling to reduce the computational load, which will lose a lot of irretrievable information. HRRNet further adopts the cascade refinement method for the segmentation map. Compared with the previous work, this method can distinguish the categories of boundary pixels to the greatest extent, and achieve the effect of refined segmentation.

So as to make the network have better generalization ability, in terms of data sets, Pascal VOC 2012 is used for training. To improve the robust performance of the model, enhancement processing such as rain, fog, blur, and light intensity changes are added to the images in the dataset. The EBIG dataset (BIG dataset + pictures taken by mobile phones + pictures searched on the Internet) was used for testing with a total of 200 pictures, and the pixel resolution of all the images in this dataset is higher than 2K. The experimental results show that HRRNet can achieve 94.26% mean intersection over union (mIoU). Part of the experimental results is shown in Figure 1.

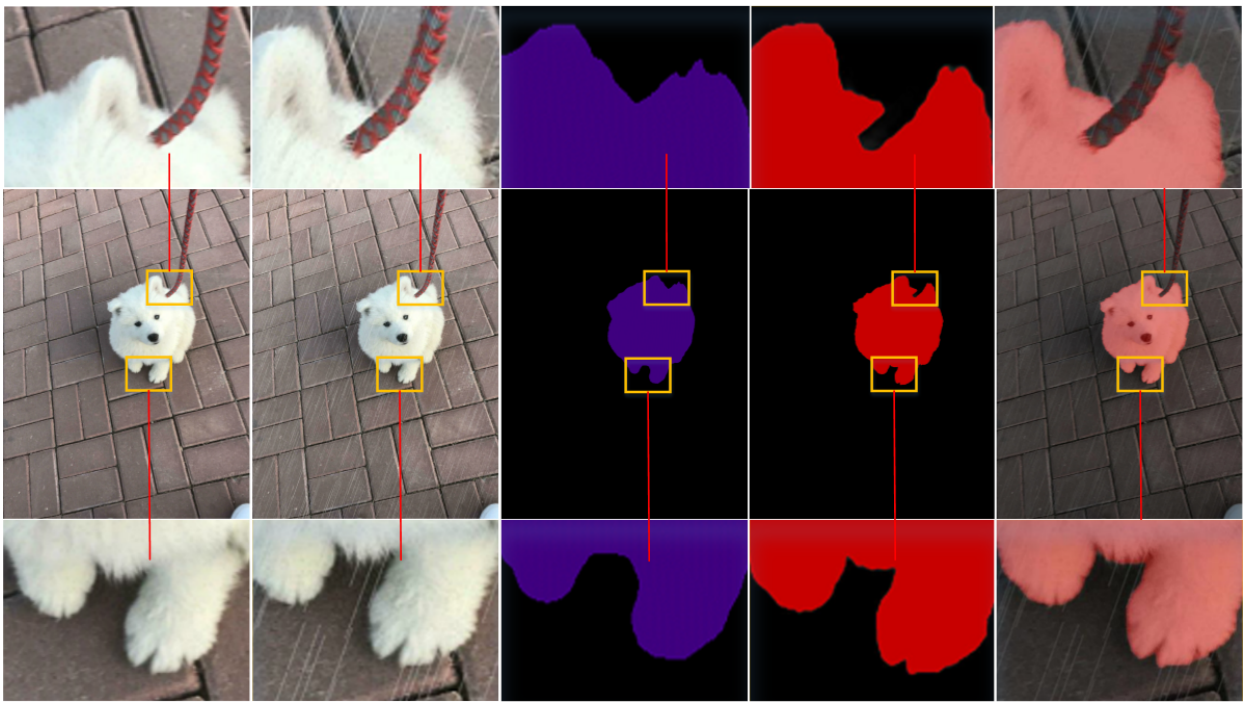
The main contributions of this paper are as follows:

1) A high-resolution edge refinement network HRRNet is proposed.

2) Improved DeepLabV3+ network by adding a shallow feature of the 1/2 size of the original image, which can better capture shallow detail feature information.

3) Add two cascaded RCUs to each output path after the Atrous Spatial Pyramid Pooling (ASPP) module in the RU to fine-tune the weights of the backbone network and speed up the network convergence.

4) The regional non-maximum suppression is added into the edge detection process of the Sobel operator to better refine the segmentation boundary.



(a) (b) (c) (d) (e)

Figure 1. High-resolution refined segmentation renderings. (a) the original image of 3024*4032 pixels taken by the mobile phone; (b) the result after rain enhancement; (c) the segmentation effect of the improved DeepLabV3+ in this paper; (d) the refined segmentation effect of HRRNet; (e) the fusion result of the HRRNet segmentation mask and the rain-enhance image.

2. Materials and methods

2.1. Related work

Image segmentation has always been a research hotspot in the field of image processing. In recent years, related algorithms have emerged. For high-resolution semantic segmentation tasks, most of the research focus has been on improving the contextual information connection, segmentation accuracy, and edge refinement capabilities in high-resolution images.

2.1.1. Contextual information task

Recently, FCN [4], U-Net [3], SegNet [5], PSPNet [6], BiSeNet [13,14], DeepLabV3 [8], DeepLabV3+ [9] and other good segmentation networks have been able to obtain excellent segmentation results in the segmentation tasks of medium and low-resolution (picture pixels below 1920*1080). FCN [4] is a pioneering work in the field of semantic segmentation, but the predicted feature resolution obtained by FCN is lower than that of DeepLabV3+ and other networks, lacking spatial consistency and context information (i.e., the connection between each pixel and surrounding pixels), the ability to extract features on high-resolution images is relatively weak. The proposal of U-Net [3] can obtain multi-level fusion features, but it has the process of border clipping [3], which causes the image to lose context information. To better obtain contextual information, SegNet adds a maximum pooling index [5] in the Decoder process, which can record the position information of pixels, so as to overcome the problem of inaccurate pixel position information in the process of up-sampling. However, these networks still suffer from insufficient contextual information when solving high-resolution tasks.

To obtain better contextual information [6,8,15], first, the researchers proposed a class of ideas [3,9,16] of using skip connections in the Encoder-Decoder [3,9] to fuse low-level and high-level feature information to obtain more comprehensive feature information. Second, researchers expanded the receptive field by deepening the network to obtain rich contextual information. In convolution, multiple small convolution kernels are used to replace large convolution kernels, or the method of atrous convolution is used to expand the receptive field, for example, atrous convolution in the DeepLab series. Furthermore, context information can also be obtained through multi-scale feature fusion methods, such as the classic ASPP [8], pyramid pooling module [15], and so on. Finally, getting better contextual information by embedding Attention Mechanism [17–22] is a more popular method recently, such as NLNet [19], DANet [17], CCNet [18], etc. Among them, NLNet captures long-distance dependencies, and its operation is to combine a relatively large search range and weight it. Through this method, a larger receptive field than 3*3 and 5*5 is obtained, thereby obtaining larger contextual information. Therefore, the Non-Local Operations method [19] is particularly important for pixel-level semantic segmentation. DANet proposes a location attention module and a channel attention module, which model the semantic interdependencies in spatial and channel dimensions [17] respectively, which can better capture the dependencies between each pixel in space and dimensions. CCNet proposes the Criss-Cross Attention (CCA) module [18] to obtain the contextual information of its surrounding pixels on the intersection path. CCNet obtains global contextual information by stacking two attention mechanism modules so that each pixel can finally capture the long-range dependencies [17–19] of all pixels. The above methods achieve a good level of capturing contextual information, but the division of edge information is not obvious enough, and the accuracy is slightly lower.

2.1.2. High resolution segmentation accuracy task

Effective measures to improve the accuracy of high-resolution segmentation mainly include fine-labeling the dataset, introducing the Transformer architecture [23–25], and embedding attention mechanisms in the network. Fundamentally speaking, the labeling problem of the dataset directly affects the accuracy of the results. Cheng et al. [2] re-annotated some images of the Pascal VOC 2012 dataset for training, which can improve the accuracy by nearly 2%. Second, the proposed Transformer architecture not only has a positive impact on NLP (Natural Language Processing) tasks

but also has a significant impact on vision tasks. Alexey Dosovitskiy et al. [23] introduced Transformer into vision tasks for the first time, splitting the image into small patches, and using the linear embedding sequence of these small patches as the input of Transformer, Image patches generated by Transformer are processed in the same way as Tokens in NLP applications, and the model is trained for image classification in a supervised manner. It has a lower performance on medium-scale datasets (sample size less than 14M) but can achieve excellent results on large datasets (14M-300M) images. To further study the impact of Transformer architecture [23,26] on vision tasks, Xie et al. [24] combined Transformer with segmentation tasks and designed a novel hierarchical Transformer encoder to output multi-scale features. It does not require positional encoding, which avoids interpolation of positional encoding, but results in performance degradation when the test resolution is different from the training resolution. Finally, the attention mechanism has been mentioned in Section 2.1.1, and will not be repeated here.

2.1.3. Edge refinement task

The mid- to low-level feature information play a key role in refining the segmentation results. When segmenting high-resolution images, the middle and low-level feature information in the network is fully utilized to obtain edge details. For example, in the field of image segmentation, most models combine post-processing CRF (Conditional Random Fields) [27] and use color contrast to refine the boundary information of the output result. However, the ideal segmentation effect cannot be achieved in areas where the color distinction is not obvious. Due to the slow training and inference speed of CRF and the lack of contextual information in deep networks, it is impossible to model the interactions between predicted pixels. V. Koltun et al. [28] proposed a fully connected conditional random field CRF, which enhanced the modeling ability of image structure and the utilization of connection information between pixels. For example, the fully connected paired conditional random field method is used in DeepLab-CRF [10] based on atrous convolution, which can shorten the inference speed and solve the problem of boundary refinement in the network.

Transposed convolution [5] uses different up-sampling methods to expand the receptive field and supplement the information of the feature map to achieve the effect of thinning the edge. But transposing the convolution can cause checkerboard artifacts [29] in the generated image, which affects the refinement task. RefineNet [11] refines low-resolution (rough-grained) semantic features in a recursive way [2,11], and rough-grained semantic features help to generate clear and detailed boundary information, Thus, the purpose of gradually recovering edge detail information is achieved by using the rough-grained features [17,30] of the intermediate layer. The cost of the method is to generate more parameters. CascadePSP adopts the method [2] of perturbing the Ground Truth (GT) image, generates rough segmentation results through simulation, and uses the Sobel edge detection algorithm [31] to strengthen the network's ability to judge the boundary. It has a good discriminative ability for the segmentation of edge pixels, but the ability for the classification of internal pixels of objects is not ideal. MagNet [32] introduces a multi-scale refinement framework, which uses the latter rough segmentation map to refine the previous segmentation map, but with the accumulation of different scale refinement modules, the network training and segmentation speed will be slowed down. FCtL [33] calculates the weight relationship of each pixel in the three branches to complement the information of different size blocks. Although this method is more accurate than MagNet, the number of parameters generated by the model will be higher, and it is not easy to deploy.

2.2. Methods

2.2.1. Model structure

As shown in Figure 2, the network model proposed in this paper can be divided into two main components: a rough segmentation module and a refinement module. In the rough segmentation module, the input image will be processed by the corresponding branch according to the size: the medium and low-resolution images will be directly processed by DeepLabV3+ to obtain the rough segmentation results, while the high-resolution images will first be downsampled, and then processed by DeepLabV3+ to obtain rough segmentation results of down-sampling size, and then upsampled to finally obtain the rough segmentation results of the original image size. In the refinement module, the rough segmentation result and the original image are taken together as input. First, the global context information of the image is extracted through a 3-RU-cascaded global process. Then, in the local process, the original image is divided into patches, and each patch is segmented and refined to extract the local details information of the image. Finally, the refined patches are pieced according to the GLNet strategy [7] to generate the final high-resolution refined segmentation result.

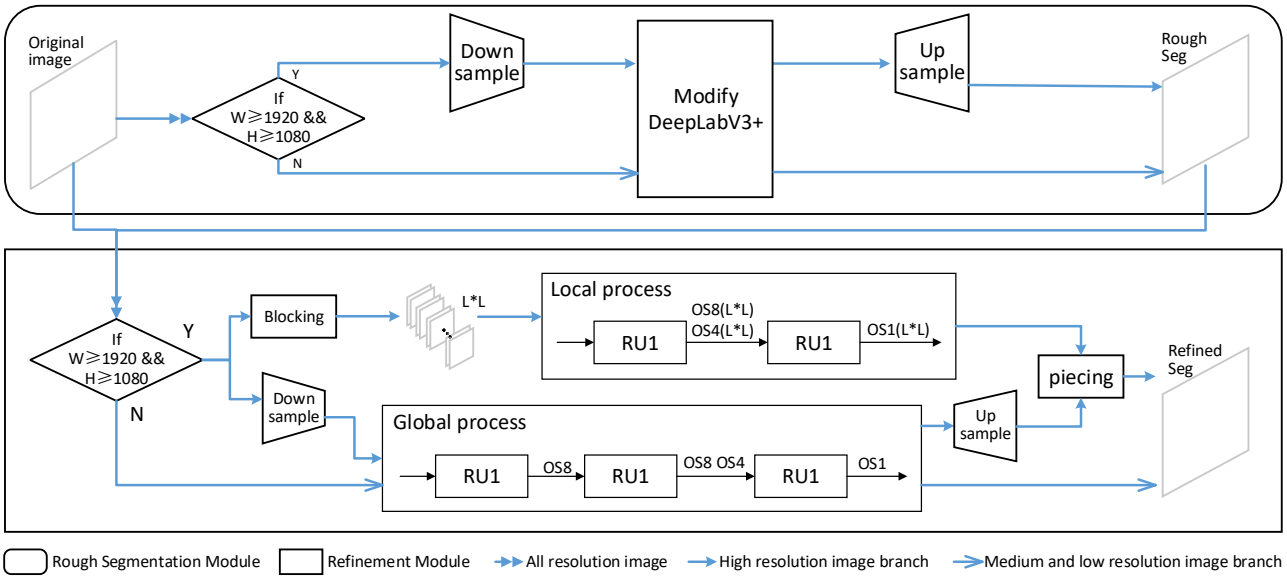


Figure 2. Structure of HRRNet.

2.2.2. Rough segmentation module

The Xception [9] structure of DeepLabV3+ uses convolution kernels of different scales for feature extraction, and the use of large convolution kernels will increase the computational of the network. To reduce the amount of model calculation, the rough segmentation module adopts the DeepLabV3+ segmentation model with Resnet50 as the feature extraction network. The input is the original image, and the output is a rough segmentation image of the same size as the original image. If the input image pixel is higher than 1920*1080, due to the limitation of GPU, it will undergo down-sampling and up-sampling operations during the rough segmentation process. In order to

obtain rich detailed information, this article improves the Encoder-Decoder with atrous convolution module [16] of DeepLabV3+: In the process of Decoder, 1/2 shallow feature information is added to restore more boundary details: Considering that the details are obvious in the shallow feature map of feature extraction, due to the influence of GPU computing power, the Output Stride (OS) of 8 is reserved in the Encoder (OS1 is the original size; OS4 is 1/4 side length, the actual is 1/16 of the size of the original image; OS8 is 1/8 of the side length, which is actually 1/64 of the size of the original image) feature map, The feature map is encoded by multi-scale atrous convolution in the spatial pyramid pooling [9] module to encode multi-scale context information, thereby outputting high-level semantic feature maps. The output feature map is upsampled by 4 times, and then spliced with the features in the same layer of encoder to obtain detailed features. To obtain more detailed features in the encoder, we fuse the 1/2 feature map of the Encoder with the features of the same layer of Decoder (the grey square in Figure 3), Then upsample the fused feature map by 2 times to obtain a rough segmentation map of the same size as the original image resolution. This improvement will improve the capture of edge details. Figure 3 is a structural diagram of a rough segmentation module adding a 1/2 shallow feature layer.

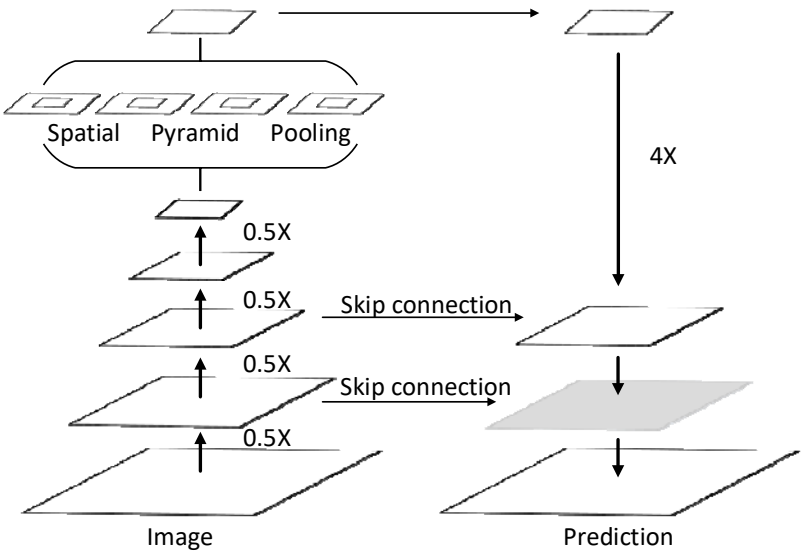


Figure 3. Add the shallow feature capture module with the size of 1/2 of the original figure.

We found that the deeper the network of the rough segmentation module, the weaker the segmentation ability of detailed information. As a result, the output accuracy of the rough segmentation module is low, which in turn affects the accuracy of refinement.

Among the methods of rough segmentation, the GT image is generally eroded and dilated to obtain roughly segmented images, but this will produce border jaggedness [34]. Moreover, this method perturbs the boundary, ignoring the problem of classification errors of internal pixels. To obtain more accurate segmentation, this paper generates a real rough segmentation map through the designed rough segmentation module, which reduces the misclassification of pixel categories within the region. An example is shown in Figure 4.

Figure 4. Comparison of real rough segmentation and perturbed GT methods. (a) original image, (b) real output by the rough segmentation module, (c) the result of perturbing the GT.

2.2.3. The refinement unit of the refinement module

The RU module in this paper uses the original image and three segmentation images of the same size at different scales as input to generate fine segmentation. The multi-scale input can make the model adaptively fuse the features of different scale feature maps to maximize the information obtained.

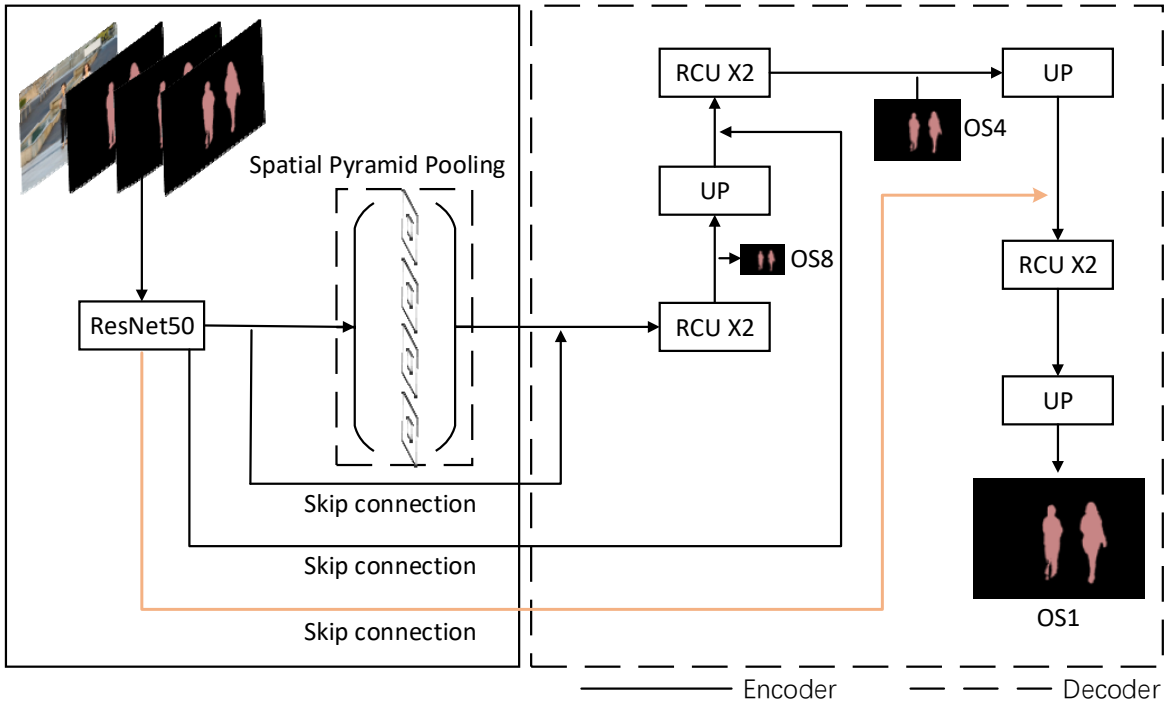


Figure 5. The network structure of the refine unit uses three-layer jump links to fuse feature information of different scales. The outermost jump links are links that add 1/2 shallow information in the rough segmentation module.

The RU module uses the improved Deeplabv3+ as the segmentation network, as shown in Figure 5: its input first passes through the Resnet50 [35] feature extraction network, and then passes through the spatial pyramid pooling module [6] to further obtain feature maps of different scales. Finally, the output of OS is 1 is obtained through feature fusion. Among them, there are also output

paths of OS4 and OS8 in the Decoder, so the RU module can obtain the segmentation results of OS8, 4, and 1, and these outputs will prepare the input for cascade refinement.

To make the RU module have better segmentation performance, two cascaded RCUs [11] are added to the output paths of feature maps of different scales in the Decoder. At the same time, a skip connection is used in the process of Encoder and Decoder, which can not only adjust the weight of the feature extraction network to obtain a higher mIoU, but also improve the convergence speed of the network. RCU mainly modifies the Residual Block [35] of Resnet and removes the BN (Batch-Normalization Layers) part, because the BN layer [11] will increase the amount of calculation and reduce the convergence speed of RU. The structure of RCU is shown in Figure 6.

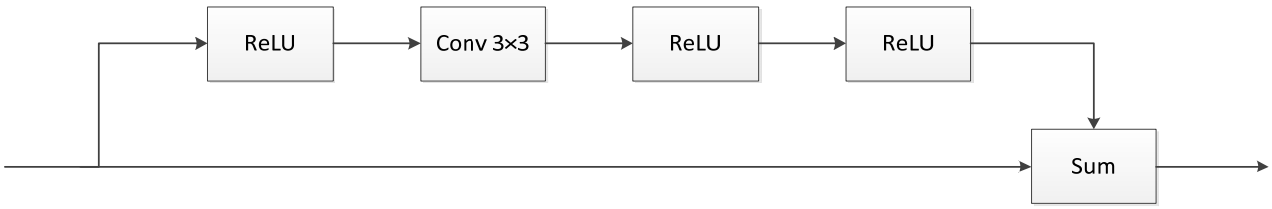


Figure 6. Structure of RCU.

2.2.4. Global process and local process

In this paper, inspired by GLNet [7], the refinement module adopts a combination of global and local processes to segment and refine high-resolution images. The global process aims to obtain global context information but ignores edge details. At present, most segmentation networks perform down-sampling operations on high-resolution images in the process of feature extraction to adapt to the input of the network, which will greatly lose detailed information; The local process can obtain fine edge detail information, but lose context information; Although the method of combining global and local modes in GLNet is trained by high-resolution images, due to the limitation of GPU, 4k or even higher resolution data cannot be input to the network for calculation at one time. So GLNet loses its performance for 4k and higher resolution data. Therefore, this paper draws on the idea of combining global and local and adopts the cascade method for refinement.

We find that the approach [2,11] of cascaded networks has been widely used in the field of computer vision recently, enabling models to efficiently extract both deep and shallow features in the input data. Here, we use the cascaded RU method for boundary refinement. This method will adaptively repair edge detail information in the network according to the features of different scales, so as to obtain a refined segmentation map.

As shown in Figure 2, the refinement module introduces a global and local mode. In the training process, due to the limitation of GPU, only the global mode is turned on at this time, and the data set of medium and low-resolution images is used for training, after that, the best RU module will be obtained. When predicting, the input high-resolution original image and the rough segmentation result are firstly divided into patches, and each image patch is segmented and refined after cascading two RUs. Then, the segmented refinement results are pieced according to the global context information output by the three-stage cascade, and finally, the refined complete high-resolution result map is output.

1) Global process

As shown in Figure 2, the refinement module takes the original image and the rough segmentation image of the original image size as input. When training and predicting through the global process, first determine whether the input resolution size is less than 2k, and if so, it will be directly input to the three-level cascaded RU, Otherwise, down-sampling is performed using the size compression Eqs (1) and (2), so as to adapt to the input of the network, and then input to the three-level cascaded RU module to obtain global information.

$$W(G) = W * \frac{L}{MAX(W,H)} \tag{1}$$

$$H(G) = H * \frac{L}{MAX(W,H)} \tag{2}$$

The structure of the three-stage cascade RU in Figure 2 is shown in Figure 7. The design purpose of this input is to better obtain features and detailed information on different scales to obtain more refined segmentation results. The input of the first layer of RU is four copies (one original image and three segmentation images, and the segmentation image is obtained by the rough segmentation module), and the output is a rough segmentation with an OS of 8. At this time, only the first RCU*2 is passed in the RU (as shown in Figure 5). Replace the last two rough segmentation inputs by the first layer with the segmentation map of OS8 output by the first layer, and upsample the segmentation map of OS8 to the input image size of the first layer as the input of the second layer. After the second layer of RU, the segmentation map of OS8 and OS4 is obtained. At this time, one and two RCU*2 will be passed in the RU respectively (as shown in Figure 5). Finally, upsample the newly generated segmentation maps of OS8 and OS4 in the second layer to the input image size of the first layer and replace the last two input images of the first layer, output a refined segmentation map of the same size as the input (OS is 1), at this time, the RU will go through two and three RCU*2 cascades respectively (as shown in Figure 5).

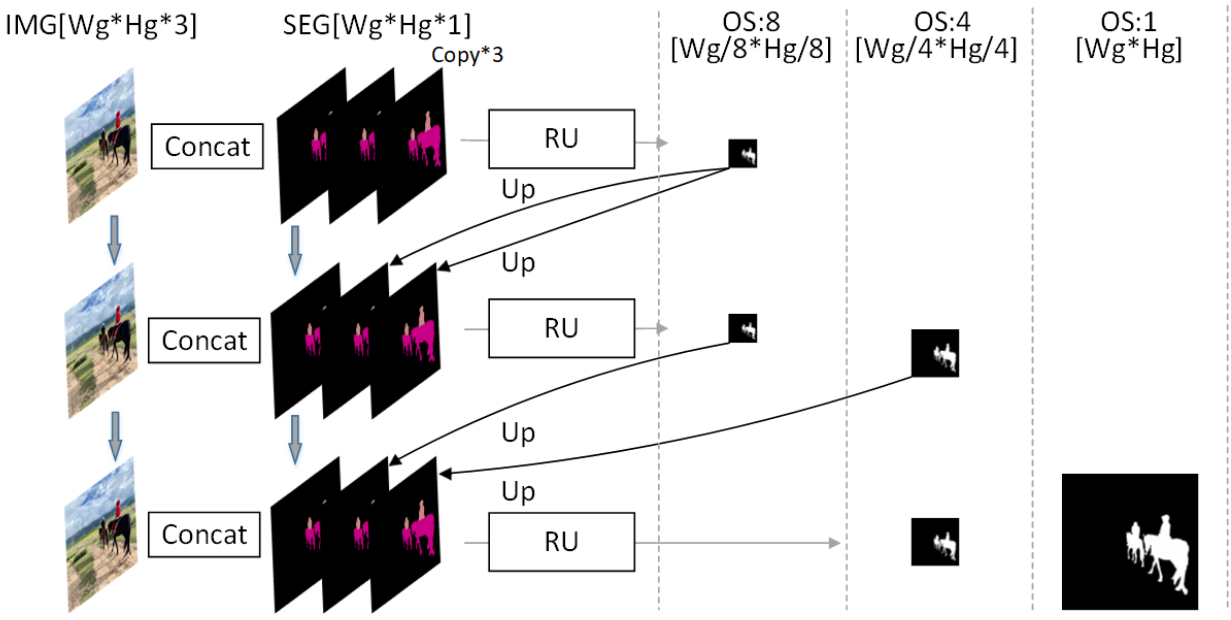


Figure 7. The global process in the refinement module. "Up" means up-sampling.

2) Local process

In the high-resolution prediction process, if the resolution of the input original image and rough segmentation image is greater than 2K, inspired by GLNet [7], the idea of combining global and local modes is adopted. According to Figure 2, the local mode is turned on in the global mode. First, the downsampled image obtains the global context information through three cascaded RUs. Then, the segmentation maps of OS4 and OS1 output by the global process are upsampled to the size of the original image and used together with the original image as the input of the local process. Then, according to the blocking idea of CascadePSP, divide the input picture with the size of L*L (700*700) [2]. In the local process, a two-level RU cascade structure is adopted to obtain L*L size refinement result patches, and the index of each patch is recorded. Finally, the GLNet method is used to concatenate [7] the refinement result patches of the local process according to the context information of the global process and obtain the result of high-resolution segmentation refinement. The network structure diagram of the local process is shown in Figure 8.

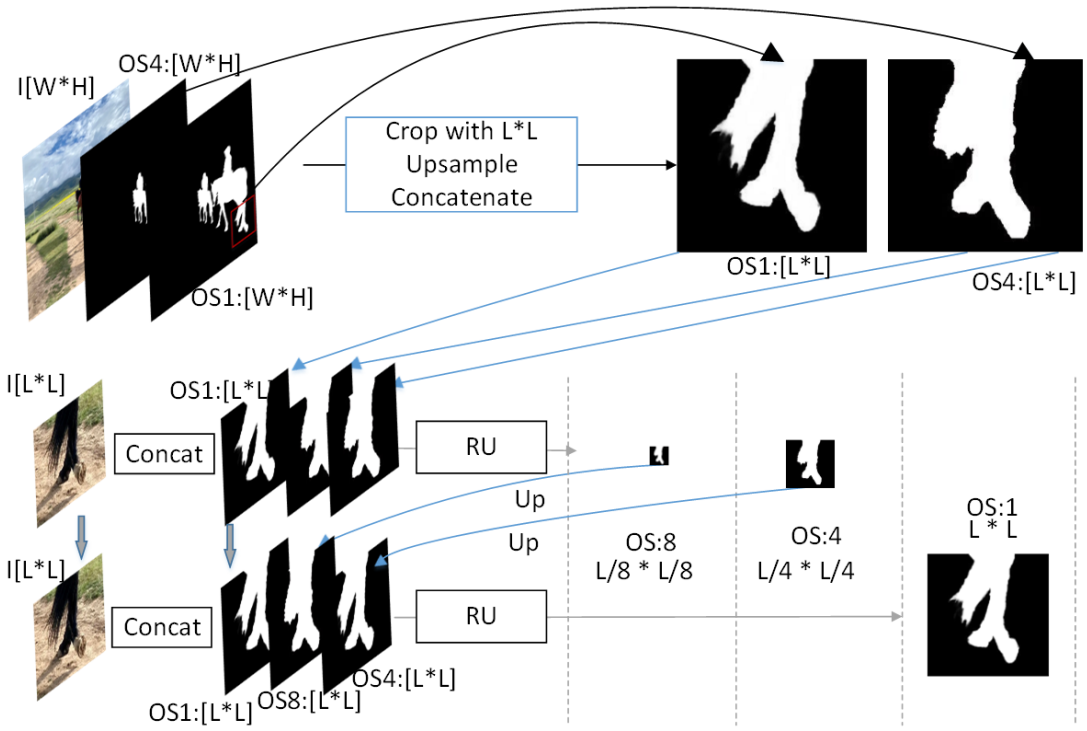


Figure 8. The local process in the refinement module. "Up" means up-sampling.

Start from the upper left corner of the original image and the split image and move in steps of L/2-32 for cropping. Input the cropped original and segmentation image patches (two copies of the OS4 segmentation patch and one OS1 segmentation patch) into the second-level cascade RU, After the first RU, the segmentation patches of OS8 and OS4 are obtained, respectively upsampled to L*L size, and combined with OS1 and image patches as the input of the second RU, and finally a refined segmentation patch of L*L size is generated.

2.2.5. Loss function

In the process of pixel-level classification tasks, the cross-entropy loss function can reduce the probability distribution difference between two pixels to make the predicted probability distribution as close to the real probability distribution as possible, which is beneficial to improve the convergence speed. The backpropagation is multiplicative, so the update of the entire weight matrix will be accelerated. So, after the first cascaded RCU block, the cross-entropy loss function is adopted, which is defined as Eq (3).

$$L = \sum_{c=1}^{M} y_c \log(p_c) \tag{3}$$

Among them, M represents the number of pixels, and y_c represents the real value of pixel c. p_c represents the predicted probability of pixel c. For different output steps, this paper uses different loss functions to construct the overall loss. For the output of the finer segmentation of Stride = 1, this paper adopts L1 + L2 loss as the loss after cascading RCU; For the loss of the intermediate step size of Stride = 4, the average value of L1 + L2 and cross entropy loss is used; For Stride = 8, the cross-entropy loss is used. At the same time, to better improve the segmentation accuracy, the Sobel operator has been improved. The regional non-maximum threshold suppression method is used to convolve the output OS1 map to calculate the grayscale difference (partial derivative) estimates of the two directions (horizontal direction G_x, vertical direction G_y). At the same time, to reduce the influence of gradient sparsity, the gradient loss is weighted. Equation (4) is the gradient loss function after weighting.

$$L_{grad} = \alpha \cdot \frac{1}{M} \sum_i \left\| \nabla(f_k(G_x)) - \nabla\left(f_k\left(G_y\right)\right) \right\|_1 \tag{4}$$

Among them, $f_k()$ represents a 3 × 3 matrix, ∇represents a gradient operator similar to Sobel, G_x and G_y are the estimated values of the grayscale difference (partial derivative) in two directions obtained by convolution of the graph, M represents the total number of pixels. α represents a weighting factor. Here we take α = 5.

Finally, the overall loss function of the network in this paper is constructed by adding the loss functions of different scales of the network, that is, Eq (5):

$$L = L_{CE}^{8_{RCU}} + \frac{1}{2}\left(L_{L1+L2}^{4_{RCU}} + L_{CE}^{4_{RCU}}\right) + L_{L1+L2}^{1_{RCU}} + L_{grad}^{1} \tag{5}$$

$L_{CE}^{8_{RCU}}$ represents the loss of a line with an output step size of 8 after passing through two RCUs; $L_{L1+L2}^{4_{RCU}}$ refers to the loss of a line with an output step size of 4 after passing through two RCUs; and L_{grad}^{1} uses a weighted gradient loss function on the line with an output step size of 1 to improve the boundary accuracy.

2.3. Data set

For high-resolution semantic segmentation networks, the current problems are:

1) The annotation cost of high-resolution datasets is high.

2) At present, there is no GPU computing tool that can directly complete the segmentation of 4K and above resolution images.

3) The current advanced semantic segmentation structures are not suitable for high-resolution semantic segmentation.

Therefore, in the experimental process of this paper, for the rough segmentation module, the data set Pascal VOC 2012 with medium and low pixels is used, and enhancement processing such as adding rain, fogging, blurring, and changing brightness is performed. This dataset is randomly generated in a 7:1:2 ratio of training, validation, and test sets. For the refinement module, MSRA-10K [36], DUT-OMRON [37], ECSSD [38], and FSS-1000 [39] are composed of a medium and low-resolution dataset (Total Data) with 36,572 images, a total of 1000+ categories, it is randomly generated in the ratio of 7:3 to train and validation sets for the refinement module training. The trained network model is tested on the EBIG dataset. In this paper, OpenCV is used to perform data enhancement on images, such as using uniform random numbers and thresholds to control the level of noise; using random noise, filters and other methods to superimpose rain and fog effects on images; using Gaussian motion blur to simulate real motion blur effects, etc. Several example images in Pascal VOC 2012 dataset after enhancement processing are shown in Figure 9.

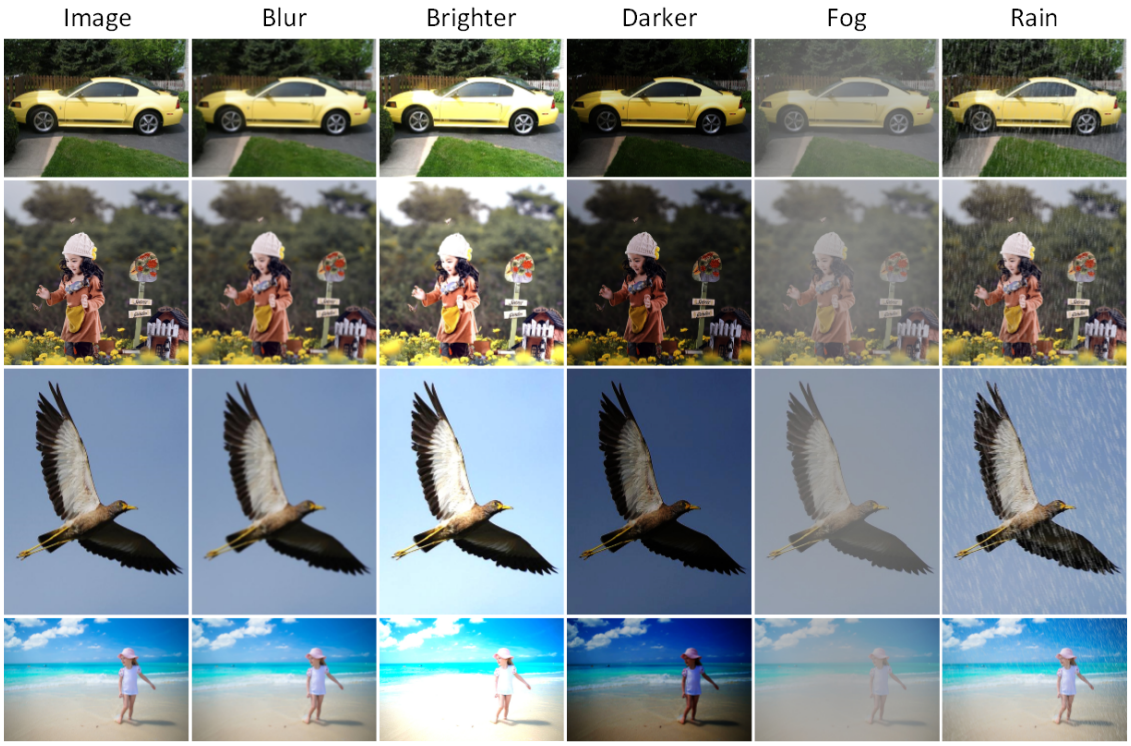


Figure 9. Example images in Pascal VOC 2012 dataset after enhancement processing.

3. Results and discussion

3.1. Performance of the rough segmentation module

HRRNet chooses the improved Deeplabv3+ as the semantic segmentation network of the rough segmentation module and uses Resnet50 as the feature extraction module. To verify the performance of the rough segmentation module, it is compared with semantic segmentation models such as U-Net [3], SegNet [5], DeepLabv3+ [9], ABMDRNet [40], etc. The dataset adopts Pascal

VOC 2012 enhanced by this paper. The feature extraction network selects VGG16 [41], Resnet34 [35], Resnet50 [35], Resnet101 [35] and MobileNet [42–44]; comprehensively compares Backbone, GFLOPs, Para, mIoU, Pix acc and other indicators.

In the training process of the rough segmentation module, the cross-entropy loss function is used, the batch size of training and verification is 4, and the method of automatically adjusting the learning rate is adopted to speed up the convergence of the model. 30K iterations on the augmented Pascal VOC 2012 dataset are performed.

The experiments in this paper are all completed under the equipment equipped with RTX3060. The results obtained on the validation set of the HRRNet rough segmentation module with the above parameters and configurations are shown in Table 1.

It can be seen from the results in Table 1 that the rough segmentation model proposed in this paper achieves the highest performance on the augmented dataset. To further facilitate visualization, the rough segmentation model in this paper and other segmentation networks with better performance are tested for pictures with different rain, fog, blur and light intensities. Some visualization results are shown in Figure 10.

Table 1. Experimental results of low-pixel segmentation network in enhancing PascalVOC 2012.

Method	BackBone	GFLOPs (M)	Para (M)	mIoU (%)	PixAcc (%)	AugData
U-Net [3]	VGG16	82.68	18.45	69	78.6	Yes
SegNet [5]	VGG16	74.74	14.03	62.2	75.9	Yes
DeepLabv3 [8]	Resnet34	58.43	10.32	72.4	80.4	Yes
DeepLabv3 [8]	Mobilenet	19.88	4.88	72.6	81.7	Yes
DeepLabv3 [8]	Resnet50	152.61	37.8	77.7	85.2	Yes
DeepLabv3+ [9]	Resnet101	255.06	55.91	79.7	87.6	Yes
DeepLabv3+ [9]	Mobilenet	28.4	4.98	76.8	84.1	Yes
DeepLabv3+ [9]	Resnet50	161.3	37.92	81.3	88.3	Yes
DeepLabv3+ [9]	Resnet101	233.75	56.03	82.1	89.2	Yes
ABMDRNet [40]	Resnet34	173.20	42.73	81.7	89.1	Yes
Ours	Resnet50	154.35	30.75	**82.5**	**91**	Yes

Through the comparative experiment of the rough segmentation module above, it can be seen that, the improved rough segmentation module in this paper is directly applied to the semantic segmentation of medium and low pixel images, has a good effect on detail extraction, and has the strongest generalization ability for different weather conditions and lighting environments. Although the segmentation speed of U-Net [3] and SegNet [5] model is fast, the segmentation network will lose edge details due to clipping and high-magnification down-sampling operation, which leads to weak segmentation refinement ability and low robustness. Deeplabv3 [8] and Deeplabv3+ [9] use the method of dilated convolution to increase the receptive field. With the deepening of the network, the network obtains better segmentation weight while the number of parameters increases, which improves the probability of correct pixel classification and the ability of context information acquisition, but still a lot of details are lost. ABMDRNet [40] uses the channel adaptive weighted fusion method for segmentation and proposes a multi-scale spatial context information module to better improve the segmentation ability, although multi-scale and multi-directional feature fusion

methods [45] can be used to extract richer features, the loss of information in low-resolution feature maps is ignored. In this paper, the problems affecting the segmentation performance are comprehensively considered, and the improved model is used for the extraction of rough segmentation features. It is found that the model achieves the best results in the validation set with mIoU of 82.5 and Pix acc of 91.0.

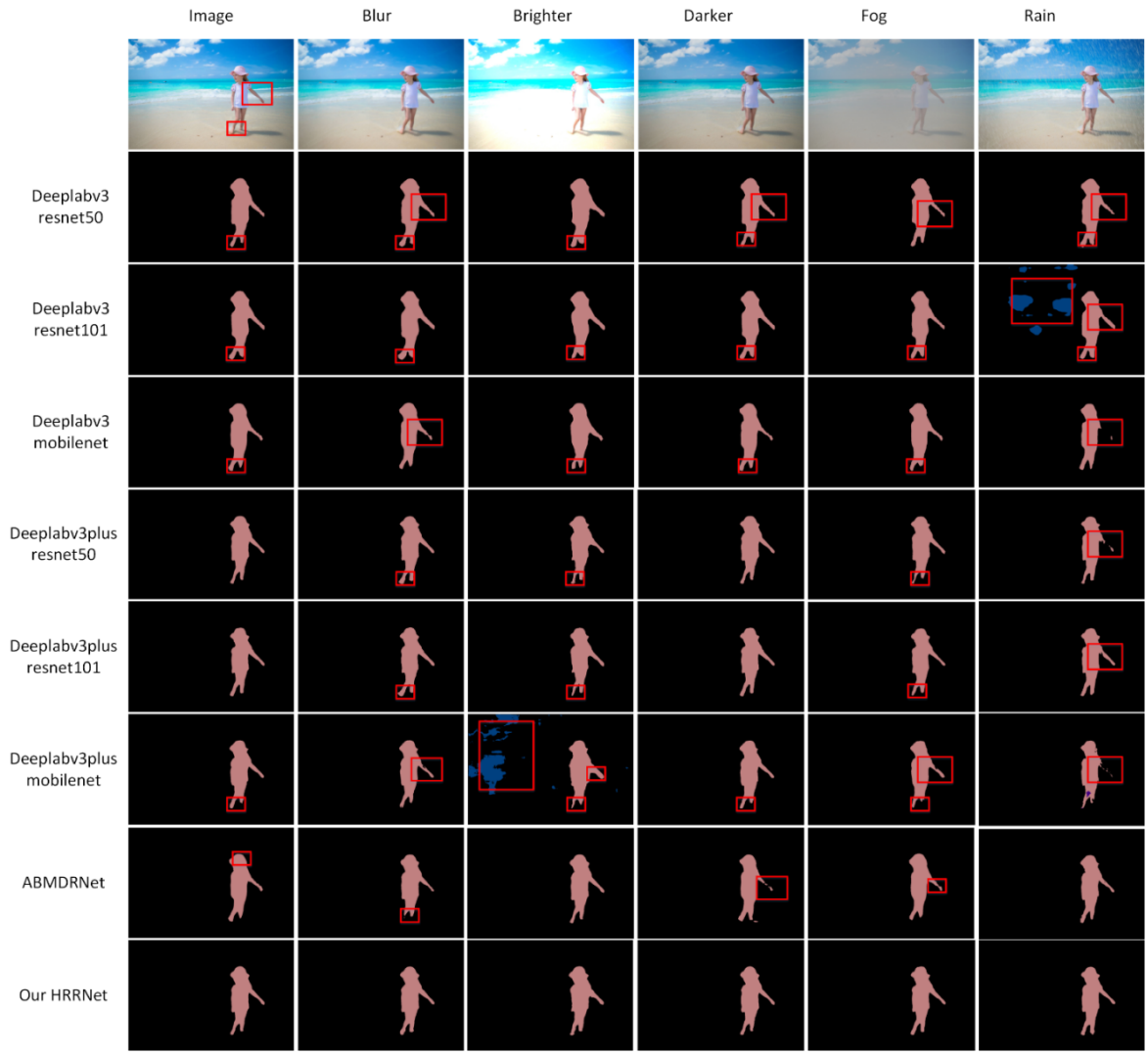


Figure 10. Segmentation performance comparison results of different segmentation models.

3.2. Performance of the refinement module

3.2.1. Improved edge detection operator

Take Img, Seg, GT as input to the refinement module, use the Total Data dataset for training, and use the improved edge detection operator Sobel [46,47] in the refinement module to improve the boundary accuracy.

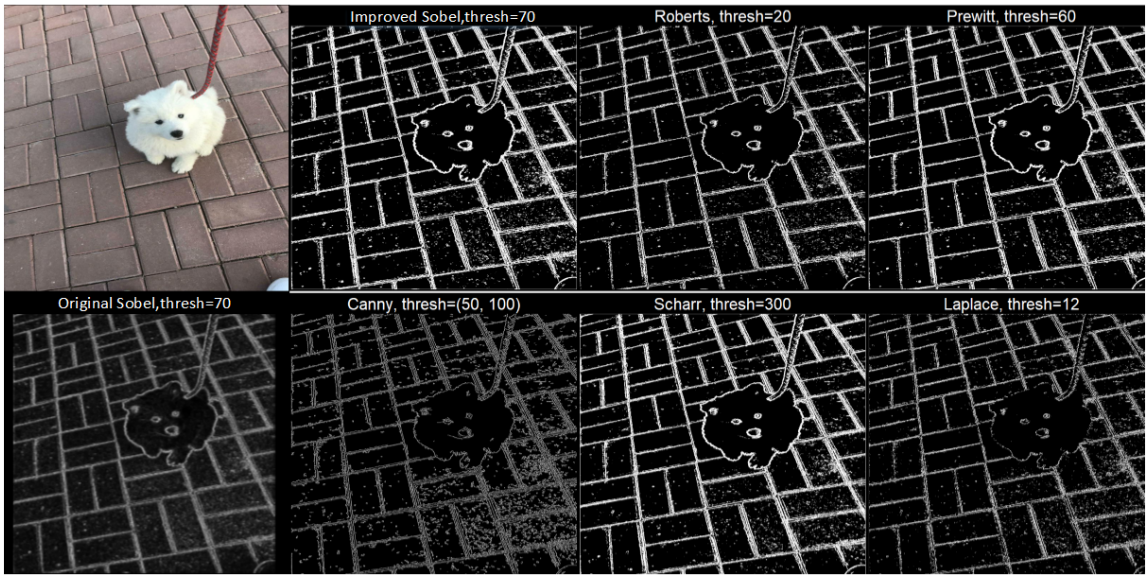


Figure 11. Performance comparison of the improved Sobel operator and other edge detection operators.

The standard Sobel operator has a fast detection speed, but during the detection process, the Sobel operator will determine whether each pixel is a boundary pixel, resulting in an increase in the detection error rate and ultimately affecting the detection effect of the Sobel operator. Therefore, we add the regional non-maximum threshold suppression method to the Sobel operator and set the regional threshold according to the gradient magnitude and parameters. Using the area threshold to determine whether the location of the area is a boundary area not only improves the detection efficiency but also improves the detection accuracy. We compare the improved Sobel operator with other edge detection algorithms and find that the improved Sobel operator has better boundary judgment ability. The result of the edge detection operator is shown in Figure 11. Other edge detection algorithms include Roberts, Prewitt, Canny, Scharr, and Laplace edge detection algorithms. Thresh is the optimal threshold.

3.2.2. Determination of the number of RCUs in the RU

To verify the impact of different numbers of RCUs [11] on the overall model performance, relevant experiments are carried out in this paper. The refinement module is trained, and only the global mode is turned on; the training data set adopts Total Data; the training iterates for 25,000 epochs. Finally, images with a resolution of 3024*4032 are randomly sampled from the EBIG dataset to verify the resulting model.

To adjust the weight of RU and obtain a finer segmentation effect, this paper gradually increases the number of RCU units in RU for comparative experiments. The results are shown in Table 2. Here, "Speed" is the time taken to process a 3024*4032 image.

Different numbers of RCUs have different effects on the experimental results. The model achieves the optimal mIoU when 2 RCUs are cascaded. However, as the number of added RCUs continues to increase, over-fitting and gradient disappearance will occur due to the lack of linear changes between layers in the network, resulting in a decrease in the mIoU of the model and an

increase in the amount of network parameters. Although RCU simplifies the residual unit, for cascade operations, the amount of calculation is greatly increased, resulting in slower processing of high-resolution images.

Table 2. Comparative experimental results of different RCU numbers in validation set.

RCU No.	GFLOPs (M)	Para (M)	Speed (s)	mIoU
0	171.3	40.78	4.2	83.58
1	234.2	45.36	6.8	84.97
2	323.8	48.84	9.1	85.96
3	440.3	59.44	26.4	84.27
4	610.4	78.25	34.5	79.73

3.2.3. Training and results in refinement module

It is determined that two RCUs are cascaded in the RU, the refinement module is finally trained. Since there is no GPU capable of processing 4K or even higher resolutions, in the process of training the high-resolution refinement module, the dataset still uses Total Data, and randomly select 10% of it as the validation set. During the training process, only the operation of the global process is used. After 36k iterations, the model reaches the convergence state and automatically stops training.

The training process is shown in Figure 12. It can be seen that the trained thinning module has produced a good thinning effect on rough segmentation and obtained good mIoU and Val Pix_Acc on the random verification set of Total Data.

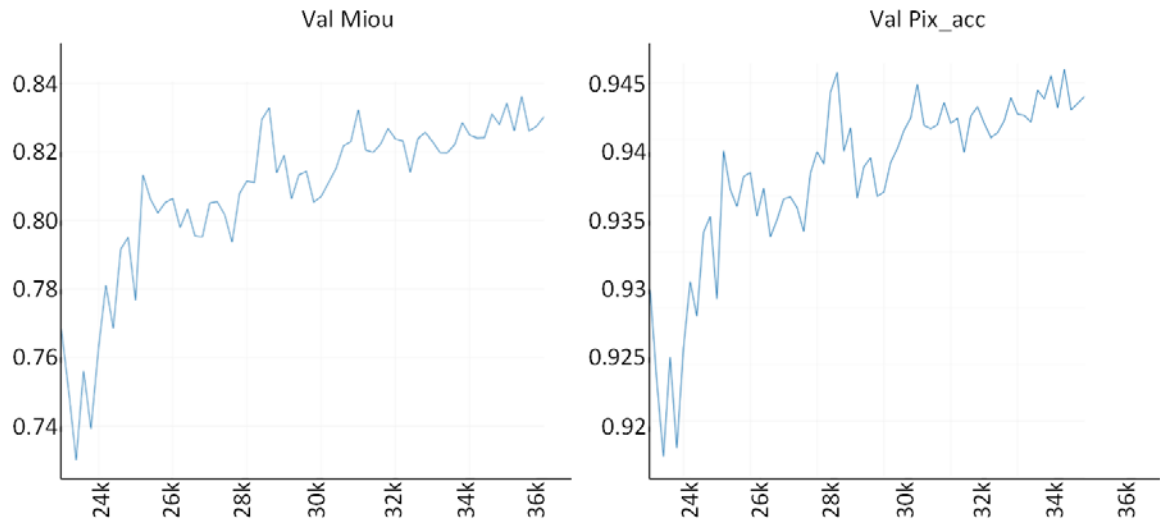
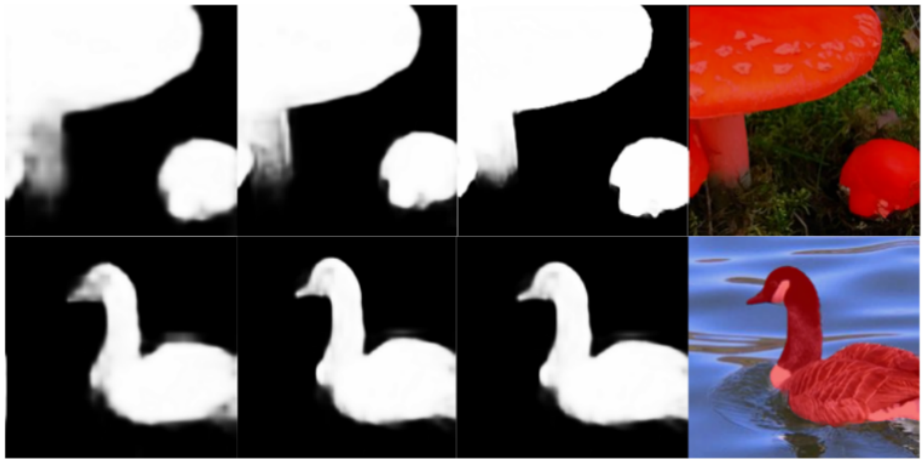


Figure 12. Visualization results of mIoU and Val Pix_Acc iteration accuracy.

During the training process, the global process of the high-resolution refinement module first takes the original image and the three copied Seg images as input, and outputs the segmentation images with OS 8, 4, and 1 by cascading the RU modules three times. At the same time, the network uses the Sigmoid function to binarize the roughly segmented image. In the refinement module, the

segmentation results of different OSs output by the global process are shown in Figure 13 (for the convenience of comparison, we adjust the segmentations of different OSs to the original image size).



(a) (b) (c) (d)

Figure 13. Segmentation refinement effect of cascaded RU modules. (a) rough segmentation map of OS8; (b) segmentation map of OS4; (c) refined segmentation map of OS1; (d) fusion result of the original image and the refined segmentation map of OS1.

3.3. Performance of HRRNet

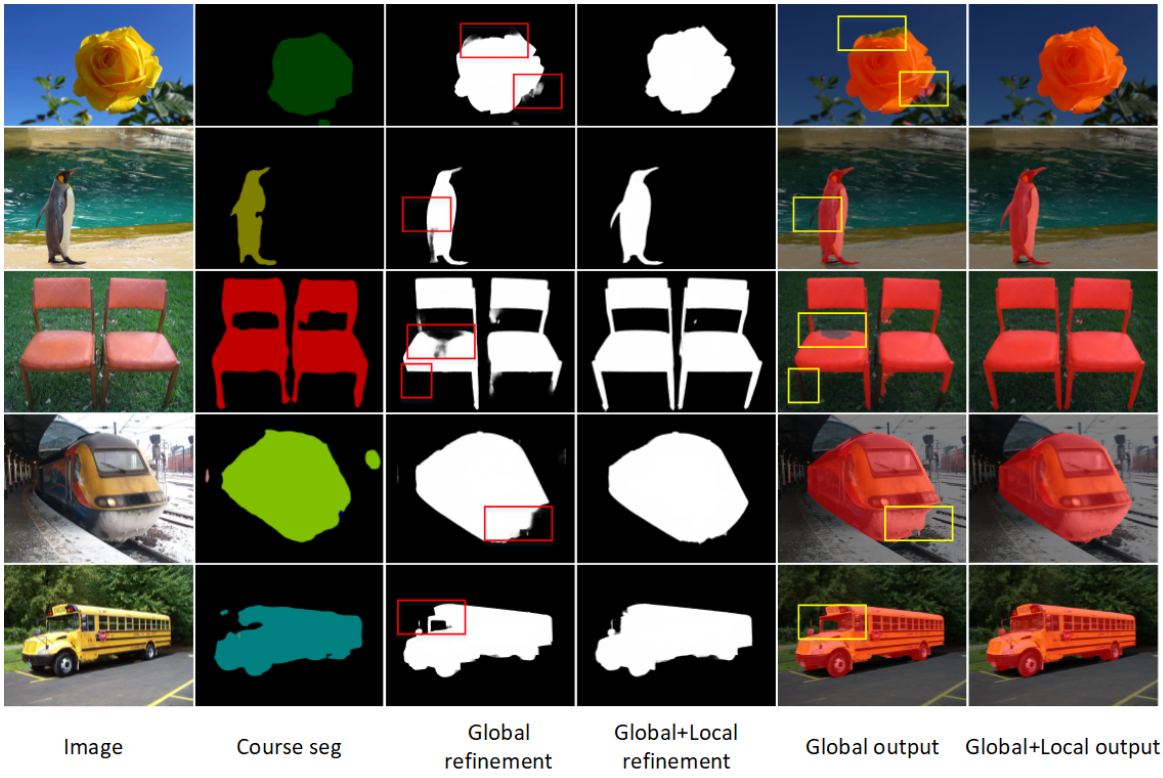


Figure 14. Step-by-step output results of segmentation based on the HRRNet model.

Table 3. Comparison of the results of different modes.

Number	Image size	Global	Local	IoU (%)
A	2848*2134	Y	N	93.26
A	2848*2134	Y	Y	96.39
B	3072*2304	Y	N	84.54
B	3072*2304	Y	Y	98.72
C	4608*3456	Y	N	88.31

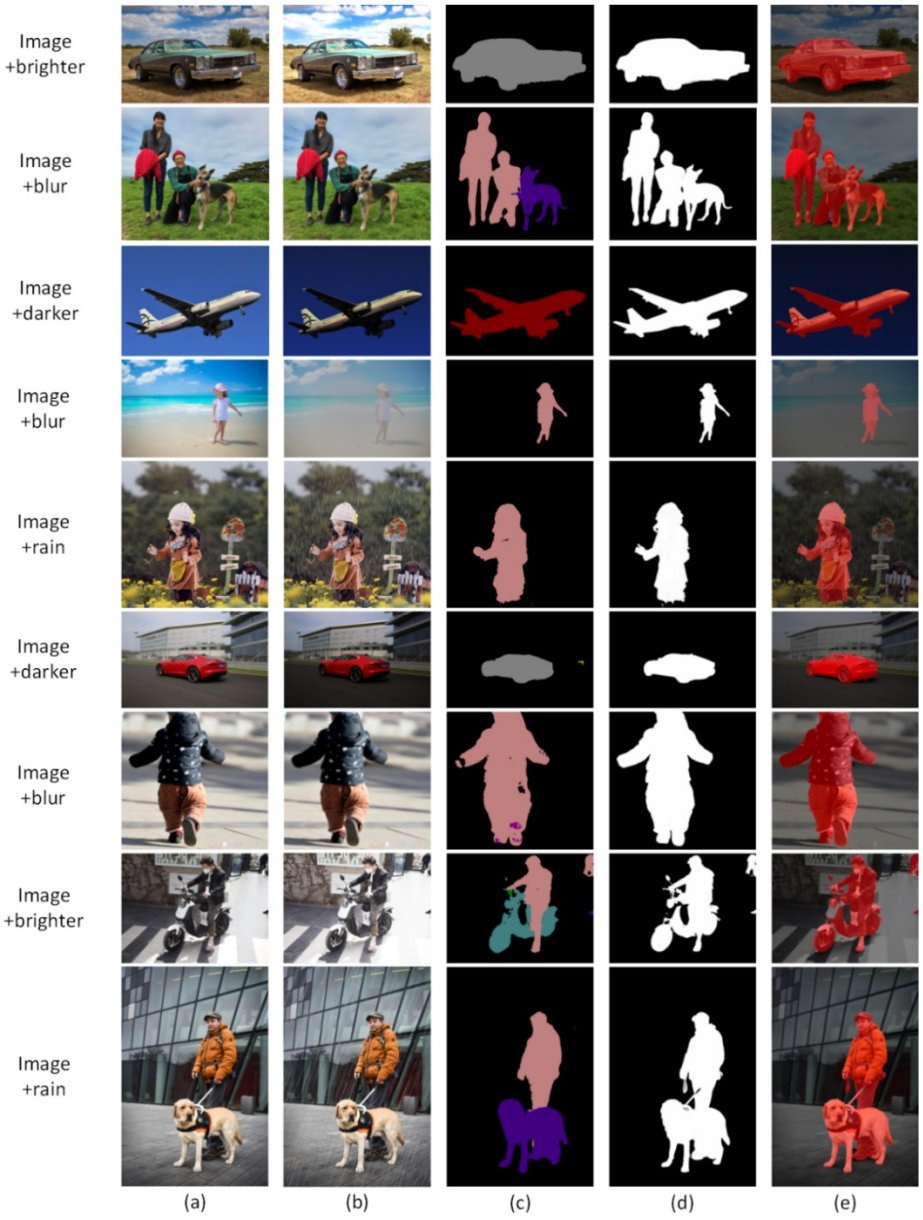


Figure 15. Refinement results of HRRNet on data-augmented high-resolution images. (a) Original image; (b) Data enhancement effect image; (c) The output result of the rough segmentation module in this paper; (d) The segmentation refinement result of HRRNet; (e) The fusion result of the original image and segmentation refinement.

Firstly, use the trained complete HRRNet model to perform segmentation and refinement tests on images with resolutions above 2K (1920*1080). To verify that the combination of global and local processes in the refinement module is helpful for the refinement and segmentation of high-resolution images, we randomly select images from the EBIG dataset for comparative experiments. The comparative data and partial visualization results of the global mode and the Global+Local mode are shown in Table 3 and Figure 14, respectively.

According to the experimental results in Table 3, for high-resolution images, if only the global mode is turned on, the network reduces the resolution of the feature map, making the edge segmentation ability of the segmentation map weak. If the method of combining global and local mode is adopted, that is, the image is blocked into patches after the global process, and the cascade RUs is performed on each patch, which further improves the ability of local segmentation and refinement.

Next, the pictures of the EBIG dataset are randomly sampled and processed using the data augmentation method in this paper, and the visual results are output in stages through HRRNet, as shown in Figure 15. HRRNet has good robustness for high-resolution images in different environments.

3.4. Comparison of the high-resolution segmentation models

Comparing the HRRNet model with RefineNet [11], DeeplabV3+ [9], MagNet [32], FCtL [33] and CascadePSP [2], the experimental dataset adopts the environment-enhanced EBIG dataset, and the experimental data are shown in Table 4. It can be seen that HRRNet can obtain 94.26% mIoU.

For high-resolution images, the idea of image blocking is more capable of extracting features than directly extracting features from the original image. The working principles of MagNet [32] and CascadePSP [2] are roughly the same. They both use the feature accumulation of the previous stage to enrich feature information and improve segmentation and refinement capabilities. FCtL [33] uses the correlation between local blocks of different scales to determine contextual information. In terms of feature fusion, this paper draws on the method [45] of multi-scale and multi-directional feature fusion to fuse features. Although the segmentation refinement performance of this method is high, it is less robust to the environment.

Table 4. Comparison of random sampling results between global mode and global + local mode in EBIG dataset.

Method	DataSet	E Aug	mIoU (%)
RefineNet [11]	EBIG	Y	80.40
DeeplabV3+ [9]	EBIG	Y	87.65
MagNet [32]	EBIG	Y	90.23
FCtL [33]	EBIG	Y	92.16
CascadePSP [2]	EBIG	Y	92.32
HRRNet	EBIG	Y	94.26

Randomly extract pictures from the EBIG dataset and perform data enhancement processing, and then compare the two segmentation network models with the best performance from Table 4. Part of the visualization results are shown in Figure 16. It can be seen that the HRRNet model in this paper can obtain better refinement ability on enhanced high-resolution images than the CascadePSP

model [2], and the rough segmentation module in this paper can better retain the feature information of high-resolution images. Moreover, the added RCU can better fine-tune the network weights, making the network boundary refinement ability more stable. The network has stronger generalization ability under different environmental conditions such as rain, fog, highlight, and darkness.

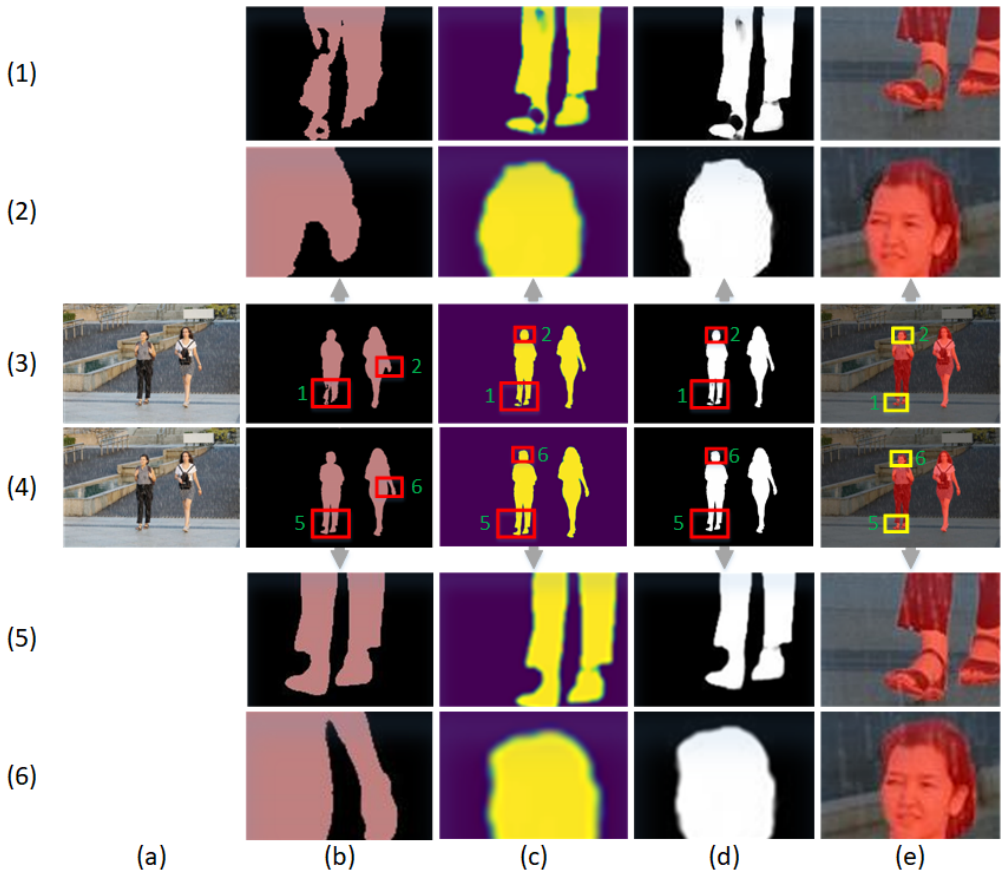


Figure 16. Comparison of segmentation performance between HRRNet and CascadePSP. Lines (1)–(3) are the experimental part of CascadePSP; lines (4)–(6) are the experimental part of HRRNet. Column (a) is the original image; (b) is the result of rough segmentation; (c) is the output of the refinement module; (d) is the result of binarization; (e) is the fusion result of the original image and segmentation refinement.

4. Conclusions

Aiming at the low accuracy of high-resolution image segmentation, the HRRNet model is proposed in this paper. The model is divided into two parts: a rough segmentation module and a refinement module. In the rough segmentation module, the DeepLabV3+ [9] network is improved, a low-level feature layer is added, and the rough segmentation results outputted are used as the input of the refinement module, so that the network reduces the probability of misclassifying the pixels inside the object. In the refinement module, RU is proposed to improve the segmentation accuracy by cascading multiple RUs to construct global and local processes. At the same time, two RCU units are cascaded on the output paths of different scales after the ASPP module of RU, which effectively improves the segmentation accuracy and the convergence speed of the network. In addition, multiple

training datasets such as Pascal VOC 2012 have been enhanced by adding rain, fog, light and shade, and blur to improve the robustness of the model. Finally, the HRRNet is compared with the state-of-the-art high-resolution segmentation network on the enhanced EBIG dataset. The experimental results show that the HRRNet model proposed in this paper achieved 94.26% mIoU and obtains the optimal refinement ability. However, it was also found in the experiment that HRRNet did not perform well in the segmentation of areas with similar colors and small target objects in high-resolution images, which is also an important part of our future research.

Acknowledgments

This research was funded by the National Natural Science Foundation of China, grant number 41701523 and Natural Science Foundation of Shanghai, China, grant number 14ZR1419700.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. Z. Zheng, Y. Zhong, J. Wang, A. Ma, Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 4095–4104. https://doi.org/10.1109/CVPR42600.2020.00415
2. H. K. Cheng, J. Chung, Y. Tai, C. Tang, CascadePSP: Toward class-agnostic and very high-resolution segmentation via global and local refinement, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 8887–8896. https://doi.org/10.1109/CVPR42600.2020.00891
3. O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer-assisted Intervention*, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
4. E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39** (2015), 640–651. https://doi.org/10.1109/TPAMI.2016.2572683
5. V. Badrinarayanan, K. Kendall, R. Cipolla, SegNet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE Trans. Pattern Anal. Mach. Intell*., **39** (2017), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615
6. H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 6230–6239. https://doi.org/10.1109/CVPR.2017.660
7. W. Chen, Z. Jiang, Z. Wang, K. Cui, X. Qian, Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **6** (2019), 8916–8925. https://doi.org/10.1109/CVPR.2019.00913

8. L. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, preprint, arXiv:1706.05587.
9. L. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, *Lect. Notes Comput. Sci.*, **11211** (2018), 833–851. https://doi.org/10.1007/978-3-030-01234-2_49
10. L. Chen, G. Papandreou, K. Murphy, A. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, *IEEE Trans. Pattern Anal. Mach. Intell.*, **40** (2018), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184
11. G. Lin, A. Milan, C. Shen, I. Reid, RefineNet: Multi-path refinement networks for high-resolution semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, **6** (2017), 5168–5177. https://doi.org/10.1109/CVPR.2017.549
12. I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, et al., DeepGlobe 2018: A challenge to parse the earth through satellite images, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, **6** (2018), 172–181. https://doi.org/10.1109/CVPRW.2018.00031
13. C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, BiSeNet: Bilateral segmentation network for real-time semantic segmentation, *Lect. Notes Comput. Sci.* **11217** (2018), 334–349. https://doi.org/10.1007/978-3-030-01261-8_20
14. C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, N. Sang, BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation, *Int. J. Comput. Vision*, **129** (2021), 3051–3068. https://doi.org/10.1007/s11263-021-01515-2
15. X. Li, T. Lai, S. Wang, Q. Chen, C. Yang, R. Chen, et al., Weighted feature pyramid networks for object detection, in *2019 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking*, (2019), 1500–1504. https://doi.org/10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00217
16. M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, C. Pal, The importance of skip connections in biomedical image segmentation, *Lect. Notes Comput. Sci.*, **10008** (2016), 179–187. https://doi.org/10.1007/978-3-319-46976-8_19
17. J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, et al., Dual attention network for scene segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **6** (2019), 3141–3149. https://doi.org/10.1109/CVPR.2019.00326
18. Z. Huang, X. Wang, Y. Wei, L. Huang, H. Shi, W. Liu, et al., CCNet: Criss-cross attention for semantic segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **10** (2019), 603–612. https://doi.org/10.1109/ICCV.2019.00069
19. X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 7794–7803. https://doi.org/10.1109/CVPR.2018.00813
20. S. Woo, J. Park, J. Lee, I. S. Kweon, CBAM: Convolutional block attention module, *Lect. Notes Comput. Sci.*, **11211** (2018), 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
21. C. Zhang, G. Lin, F. Liu, R. Yao, C. Shen, CANet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **6** (2019), 5212–5221. https://doi.org/10.1109/CVPR.2019.00536

22. H. Zhou, J. Du, Y. Zhang, Q. Wang, Q. Liu, C. Lee, Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition, *IEEE/ACM Trans. Audio Speech Lang. Process.*, **29** (2021), 2617–2629. https://doi.org/10.1109/TASLP.2021.3096037
23. R. Ranftl, A. Bochkovskiy, V. Koltun, Vision transformers for dense prediction, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 12159–12168. https://doi.org/10.1109/ICCV48922.2021.01196
24. R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: Transformer for semantic segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 7242–7252. https://doi.org/10.1109/ICCV48922.2021.00717
25. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, *Adv. Neural Inf. Process. Syst.*, **15** (2021), 12077–12090. https://doi.org/10.48550/arXiv.2105.15203
26. S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 6877–6886. https://doi.org/10.1109/CVPR46437.2021.00681
27. K. Khan, N. Ahmad, K. Ullah, I. Din, Multiclass semantic segmentation of faces using CRFs, *Turkish J. Electr. Eng. Comput. Sci.*, **25** (2017), 3164–3174. https://doi.org/10.3906/elk-1607-332
28. M. T. T. Teichmann, R. Cipolla, Convolutional CRFs for semantic segmentation, preprint, arXiv:1805.04777.
29. Y. Kinoshita, H. Kiya, Fixed smooth convolutional layer for avoiding checkerboard artifacts in CNNs, in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, **5** (2020), 3712–3716. https://doi.org/10.1109/ICASSP40776.2020.9054096
30. L. Wang, D. Li, Y. Zhu, L. Tian, Y. Shan, Dual Super-resolution learning for semantic segmentation, in P*roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 3773–3782. https://doi.org/10.1109/CVPR42600.2020.00383
31. T. Hu, Y. Wang, Y. Chen, P. Lu, H. Wang, G. Wang, Sobel heuristic kernel for aerial semantic segmentation, in *2018 25th IEEE International Conference on Image Processing (ICIP)*, (2018), 3074–3078. https://doi.org/10.1007/CVPR42870.2018.00670
32. C. Huynh, A. T. Tran, K. Luu, M. Hoai, Progressive semantic segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **1** (2021), 16750–16759. https://doi.org/10.1109/CVPR46437.2021.01648
33. Q. Li, W. Yang, W. Liu, Y. Yu. S. He, From contexts to locality: Ultra-high resolution image segmentation via locality-aware contextual correlation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 7232–7241. https://doi.org/ICCV48922.2021.00716
34. X. Qi, M. Fei, H. Hu, H. Wang, A novel 3D expansion and corrosion method for human detection based on depth information, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **761** (2017), 556–565. https://doi.org/10.1007/978-981-10-6370-1
35. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **12** (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90

36. M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, S. Hu, Global contrast based salient region detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, **37** (2015), 569–582. https://doi.org/10.1109/TPAMI.2014.2345401
37. C. Yang, L. Zhang, H. Lu, X. Ruan, M. Yang, Saliency detection via graph-based manifold ranking, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2013), 3166–3173. https://doi.org/10.1109/CVPR.2013.407
38. J. Shi, Q. Yan, L. Xu, J. Jia, Hierarchical image saliency detection on extended CSSD, *IEEE Trans. Pattern Anal. Mach. Intell.*, **38** (2016), 717–729. https://doi.org/10.1109/TPAMI.2015.2465960
39. X. Li, T. Wei, Y. Chen, Y. Tai, C. Tang, FSS-1000: A 1000-class dataset for few-shot segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 2866–2875. https://doi.org/10.1109/CVPR42600.2020.00294
40. Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, J. Han, ABMDRNet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 2633–2642. https://doi.org/CVPR46437.2021.00266
41. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556.
42. A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, et al., Searching for mobileNetV3, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, **10** (2019), 1314–1324. https://doi.org/10.1109/ICCV.2019.00140
43. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 4510–4520. https://doi.org/10.1109/CVPR.2018.00474
44. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, MobileNets: Efficient convolutional neural networks for mobile vision applications, preprint, arXiv:1704.04861.
45. J. Sun, W. Lin, A target recognition algorithm of multi source remote sensing image based on visual internet of things, *Mob. Networks Appl.*, **27** (2022), 784–793. https://doi.org/10.1007/s11036-021-01907-1
46. W. Dong, D. Peng, X. Liu, T. Wang, J. Long, Eight direction improved Sobel algorithm based on morphological processing in 5G smart grid, in *2021 2nd International Conference on Computing, Networks and Internet of Things*, (2021), 1–5. https://doi.org/10.1145/3468691.3468721
47. Y. Ma, H. Ma, P. Chu, Demonstration of quantum image edge extration enhancement through improved Sobel operator, *IEEE Access*, **8** (2020), 210277–210285. https://doi.org/10.1109/ACCESS.2020.3038891

AIMS Press