Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Automatic Hair Segmentation in Complex Background

Huijun Zhu
School of computer science and engineering
Nanjing University of Science and Technology
Nanjing, Jiangsu Province, China
446145005@qq.com

Yufei Liu
Nanjing Foreign Language School
Nanjing, Jiangsu Province, China
lyfnfls@163.com

Yong Liu*
School of computer science and engineering
Nanjing University of Science and Technology
Nanjing, Jiangsu Province, China
Liuy1602@njust.edu.cn

***Abstract*—Fine hair area segmentation, which allows the individual processing of a person's hair, has important applications in digital entertainment simulation. This paper proposes a method to automatically and precisely extract hair contour under complex background. Our method divides hair semantic segmentation into two phases. In the first stage, an improved BiseNet network was used to segment the rough hair area automatically. In the second stage, we generated corresponding trimaps through the output of the previous stage and input them into a small convolutional network together with the original input to refine the rough segmentation and get more accurate edges. Applying this method to a large number of examples, the experiment results show that this method can not only automatically extract hair region effectively, but also have good effect on the processing of hair edge under complex background.**

I. INTRODUCTION

Hair is an important element of human appearance. In recent years, hair collection, design and simulation have attracted more and more interest. Hair segmentation can be used for many applications such as makeup changes and portrait editing, as well as for facial recognition and gender classification. Therefore, hair segmentation and hair details are very important. However, due to the very big difference in the appearance of the hair for different ages, genders and races in an unconstrained environment, it can be very difficult to distinguish hair from the complex environment. As a result, the automatic hair segmentation task is often quite challenging, even the most advanced algorithms have difficulties in dealing with different hair colors, hairstyles, head postures and confusing background colors.

Previous studies on hair segmentation can be divided into three groups.

The first group asked the user to provide rough parts manually, and then to refine them automatically0. For example, the method of Wang et al.[2] first needs to obtain the seed point pixels identified as hair, and then transmit them to a pre-trained classifier to perform hair segmentation. [3] enables the user to draw a small amount of strokes in the hair area, and then uses the characteristics of these strokes to guide the overall image segmentation. Wang et al. [4] first used fixed active segmentation to find a good candidate hair region, and then used graph-cut to refine the region. Next, the bayesian method was used to find hair seeds in the image, and the support vector machine (SVM) classifier and graph-cut were used to generate the final segmentation result. However, these methods need to find some seed pixels that are identified in the hair area, and since the method is based on color and direction information, the performance will decrease when the background and hair color are very similar.

The second group does not require the user to provide the seed pixel of the hair, but is based on the matching and optimization of the rough estimation of the hair mask. Wang et al. [5] first built a rough hair probability map based on the training image (with ground truth manually marked) and the prior information of the color and location of the test image. Then, according to the relationship between the rough HPM of the test image and the training image, the final segmentation result is obtained by further adjustment. In [6] , a hybrid model based on the color and position information of hair is established to obtain the rough estimation of face, hair and background area. Then, graph clipping or cyclic confidence dissemination algorithms are used to optimize Markov networks for hair segmentation tasks. Wang et al. [7] generated a partition-based model for the input image and the Markov random field was used to acquire the final segmentation result. However, these methods are very sensitive to the accuracy of face detection, and the hair model is based on the empirically defined head area, so the different posture of the head and the shielding of the face cannot be well handled.

The third group is the neural network method. This method attempts to obtain hair segmentation results by using machine learning technology [8]by training a great quantity of portrait images with artificial annotation of the hair area. These methods follow the general segmentation process, such as pyramid scenario parse network PSPNet[9].

In our work, the BiseNet[10] with big and powerful function was employed. We used the CelebAMask-HQ[11] large face image data set to train the BiseNet model, and the output of the network was the prediction of these pixels. BiseNet was trained end-to-end. We automatically generate trimaps from BiseNet's output via morphological operations. However, the mask from the above work failed to capture some of the finer details at the edges of the hair area. The reason mainly has two aspects: One is that due to BiseNet uses a light-weight backbone network that cannot learn deeper image semantic information. lightweight design makes it in the pursuit of segmentation speed at the same time give up

* Corresponding author: Yong Liu is with School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (E-mail: liuy1602@njust.edu.cn).

This work was supported in part by National Natural Science Fund of China [Grant No. 61473155], Assembly Pre-research Sharing Technology Project [Grant No. 41412040102], Six talent peaks project in Jiangsu Province [Grant No. GDZB-039], Primary Research & Development Plan of Jiangsu Province [Grant No. BE2017301].

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE
978-1-7281-6321-5/19/$31.00 ©2019 IEEE

some degree of precision, the other one is that a part of the spatial information is lost in the down-sampling process of BiseNet (even though it has supplemented the spatial information by fusing the feature learned by Spatial Path ,some of the information can not be recovered). Therefore, in view of the above defects of BiseNet in hair region segmentation, we propose a two stage method. The specific method is that the first stage, we use the improved BiseNet network hair area segmentation. It can be more subtle than original BiseNet output. However, the accuracy of such image edit we want the original intention is not enough, so we put forward the second phase——refinement stage. We carried out subsequent processing on the above segmentation results, specifically by generating the corresponding trimap through morphological operation. Based on the input image and its corresponding trimap, we also designed a subsequent small convolutional network to deal with the edge part of the above segmentation results more finely.

In summary, our contribution includes three aspects: (1) We improved BiseNet and trained with celebAMask-HQ dataset and obtained more accurate segmentation results. (2) Based on the output of semantic segmentation, a trimap was automatically generated by morphological operation, and combing the original image and the trimap, the result of semantic segmentation was fine tuned by a subsequent convolution network.

II. PROPOSED METHOD AND ALGORITHM

Our proposed framework is shown in figure 1. The input image(figure 1. (a)) is fed to the improved BiseNet network to generate a pixel-level prediction map(figure 1.(b)). Next, the prediction of the previous step generates trimaps(figure 1.(c)) automatically through morphological operations. Finally, the input images and corresponding trimaps are fed into a refinement stage to gain more accurate segmentation results(figure 1.(d)).

The following describes the different system components in detail.

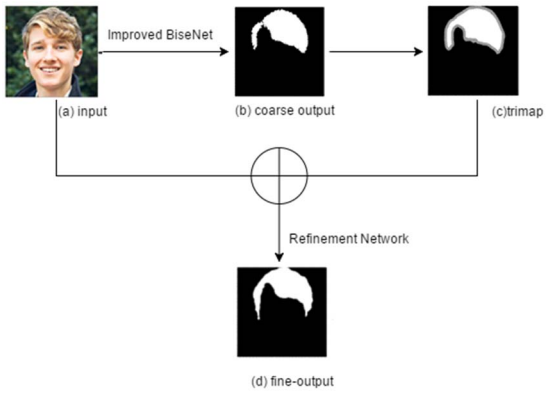

Figure. 1. Our framework

A. Semantic Segmentation Section

To reliably separate hair area from a chaotic background, local and global context information should be considered simultaneously. The improvement of existing semantic segmentation network is mainly in two aspects: the first is to retain more spatial information; the second is to add receptive field. Convolutional neural network (CNN) uses consecutive down-sampling to encode high-level semantic information. However, in the semantic segmentation task, the spatial information of the image, which is very important for predicting the detailed output, is lost due to continuous down sampling operations. The original FCN[12] network up-sampled the images to the original size through deconvolution. In order to make up for the loss of accuracy in the process of up-sampling, it adopted a skip-connected network structure to encode different level features so as to optimize the output. U-net [13] introduces a useful skipped connection network structure for this task, the global convolutional network [14], which combines the u-shape structure with the "big kernel". LRR[15]used Laplacian pyramid to reconstruct the network. RefineNet [16] increased the multipath refining structure (multi - path refinement) to refine predictions. In order to achieve the feature selection, DFN[17] designed channel attention block. However, in these structures, some of the lost spatial information is still not easy to recover.

Therefore, the widely successful BiseNet algorithm is used. Compared with the previous network used for segmentation, this algorithm can availably make up the loss of spatial information without shrinking the receptive field, and can also be used to solve other intensive prediction issues.

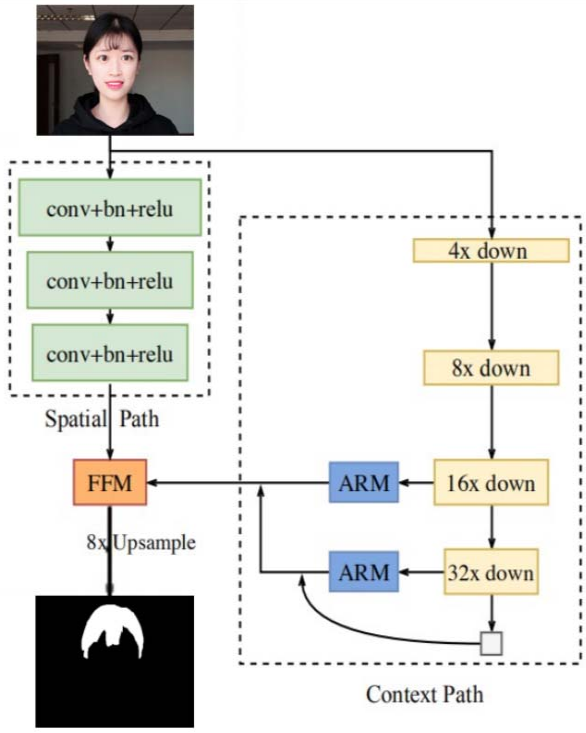

Figure. 2. The structure of the original BiseNet

1) The structure of BiseNet

BiseNet network consists of two parts: Spatial Path (SP) and Context Path (CP).

The Spatial Path consists of three convolution layers with stride of 2, followed by batch normalization and ReLU function. The output feature map extracted by this path is 1/8 of the original image. Because of the large space size of feature graph, rich spatial information can be encoded. The Context Path adopts the lightweight model X-ception[18] as the backbone network to quickly sample the feature map and obtain a larger receptive field. Then, a global average pool is added to the tail of the lightweight model as an attention refinement module(ARM), which can provide global context

information for the maximum receptive field. Finally, a special feature fusion module(FFM) is adopted to merge the output characteristics of the two paths. Specifically, the output characteristics of the two paths are concatenate. Then, we used batch normalization to balance the size of features, then calculated a weight vector, such as SENet[19]. This weight summarized the series features into feature vectors and vector can re-weight features, whic is equivalent to feature selection and combination, The architecture of the original BiseNet is shown in Figure 2.

2) Improvements to BiseNet

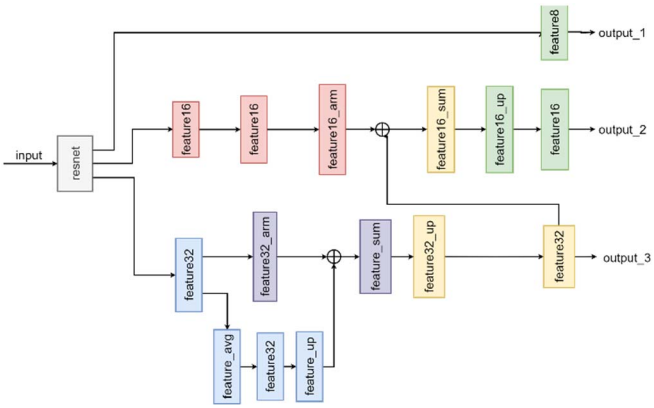

Figure. 3. The architecture of improved architecture

Cause the BiseNet uses Xception as its backbone network, it actually gives up part of the accuracy while ensuring the real-time performance. For this, we replace the simple convolutional layers with the residual blocks[20] to expand the depth of the Context Path of BiseNet, and apply FFM to the Context Path to optimize the prediction results. The improved architecture is shown in Figure 3. First, the images are fed into the residual network to get three feature maps of different scales, and then the three maps are sampled and fused. The specific rules are shown in Figure 3,where the feature maps named after "avg" represents the output from the average pool, the feature maps named after "arm" represents the output from the "Attention Refinement Module" which has been described in [10], the feature maps named after "sum" represents the output of the concatenate operation, and the feature maps named "up" represents the output from deconvolution operation. Finally, we can get three characteristic images of two scales, one eighth and two sixteenth of the original image size. After this improvement, the depth of the network is deeper, and the feature obtained by this path is much more enriched to learn more global information and semantic information. For Spatial Path, we simply use the output_1(as shown in Figure 3) of the improved Context_Path to replace the original method of [10], cause it has the same size as which of the feature map obtained from the original Spatial Path, and also can satisfy the intention of retaining more spatial information. The FFM module is used to integrates Context Path and Spatial Path.

3) Loss

The loss function is the weighted sum of the concatenated loss and the loss of each stage in the Context Path (where the overall loss and the loss of each stage in the Context Path are all Softmax loss, as Equation 1 shows) as Equation 2 shows.

$$\text{loss} = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -\log\left(\frac{e^{p_j}}{\sum_j e^{p_j}}\right) \qquad (1)$$

$$\mathrm{L(X;W)} = l_p(X;W) + \alpha \sum_{i=1}^{K} l_i(X_i;W) \qquad (2)$$

Where l_p is the concatenated loss. X_i is the output feature of the stage i of the Context_Path model. l_i is the loss of the i-th output feature map in the context path. In our architecture, K is equal to 2. The parameter α is used to balance the weight of the main loss and the auxiliary loss. In the training, α is 1 and L is the joint loss function.

B. Refinement stage

Although the improved BiseNet can improve the location of most semantic segmentation algorithms and get more precise segmentation results, its accuracy is still not enough for human image editing. In order to overcome this lack of accuracy, we use a four-layer convolutional network to further refine the output of the previous stage.

The idea of segmentation from coarse to fine is reflected in many methods, such as the coarse-to-fine landmark localization[21], which is used to solve the problem of facial feature point location, which improves the accuracy of location, such as the strategy of extinction from coarse to fine used in deep image matting[22]. Inspired by them, we designed our refinement stage. Although the structure is simple, we find it works well.

1) Network structure

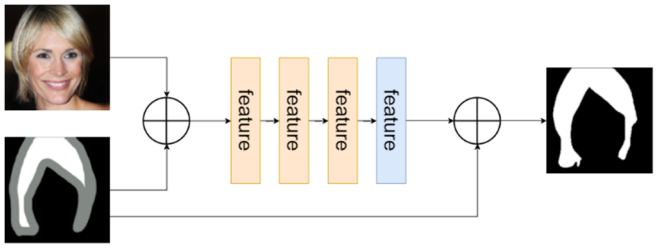

Figure. 4. The architecture of the refinement network

In this stage, we design a four-layer convolutional network. First of all, we build trimaps according to the output of the first phase. We stacked the original images and the trimaps(scaled between 0 and 255) together, forming a 4 channel input. In the convolution. The output is the corresponding ground truth labels. Considering that the purpose of this phase is to fine the edge of the output of the first stage, it needs abundant spatial information, we don't use the down-sampling in this stage. Since the layer is not very deep, the calculation of this phase will not be very large. We use batch normalization in the first three layers to adjust the network and accelerate its convergence. The specific network mechanism is shown in figure 4.

2) Loss

We use cross entropy loss function and restrict the loss function more strictly to guide the network to modify the edge of the previous stage. Specifically, we set a weight of 0.1 for the loss of the background region and foreground region in trimap, and 0.8 for the unknown region in trimap. In this way, the network will pay more attention to the prediction of the unknown region (i.e. the edge of the hair region).

3) Application

Next, we use the trained network to refine the output of the semantic segmentation part which is trained using an improved BiseNet. Specifically, we applied morphological operations to the output segmentation mask to obtain a trimap representing foreground (hair), background, and unknown pixels. Noting that the input at this stage was automatically generated from the output of the semantic segmentation section. By putting the original images and the corresponding trimaps into the refinement network, we can get finer and sharper segmentation results.

C. Application

The algorithms we use for hair area segmentation can be used for various hair operations or for applications that create hair databases for further processing.

As an example application, we implemented a simple tool that can manipulate hair using the mask generated by our algorithm. Figure 4 shows an example of how we can change hair color using the proposed method. We use the extinction algorithm to improve the edge of the result of hair color change As shown in Figure 5.

Figure. 5. Use our method to change hair color

III. EXPERIMENTS

A. Experimental details

The images used to train our semantic segmentation network were taken from the CelebAMask-HQ large face image dataset, which contained 30000 images with hyperpixel labels. And we only use the labels of hair and label the rest of the classes as background. The size of each image in the training set is 1024×1024 pixels, which is much more than BiseNet receptive field size. Therefore, we resized each training image by cropping them and added a copy of each horizontally flipped image for data enhancement.

In this part, we used small batch random gradient descent (SGD) . During the training, the batch size was 20, the momentum was 0.9, and the weight attenuation was 1e-4. We adopted a "poly" learning rate strategy, in which the initial rate of each iteration was multiplied by $\left(1-\frac{iter}{\max\_iter}\right)^{power}$, and power was 0.9. The initial learning rate is 2.5e-2.

After the semantic segmentation part converges, we fix its parameters and began training refinement network. The network structure adopted in this stage is very simple, and we only use categories to predict losses. We use Adam to update the two phases. During the training, set the learning rate to 10e-5.

All of our experiments were conducted on the public platform TensorFlow on a PC equipped with two TiTan 1080 Ti graphics cards.

B. Experimental results

We show the performance of FCN, original BiseNet, our improved BiseNet without the refinement process and our whole method respectively in table 1.

Table. 1. Overall IOU and recall rate of CelebAMask-HQ dataset compared to other benchmarks

Methods	IOU	recall
FCN	0. 6776	0.7994
Original BiseNet	0.7890	0.8995
Our method(without the refinement stage)	0.8972	0.9551
Our method(with the refinement stage)	0.9181	0.9676

We compared the hair segmentation results of improved BiseNet and the final results (with refinement process) as shown in Figure 6. It can be seen obviously that after the second stage, the edge of hair region is more accurately segmented. In the first column, we make the edges of the hair finer, in the second and the third column, some incorrectly segmented area is fixed by the refinement network.

Figure. 6. The result of the first stage and the second stage of our method. The first row shows the original images, the second row shows the results of the improved BiseNet, and the last row shows the results after refinement.

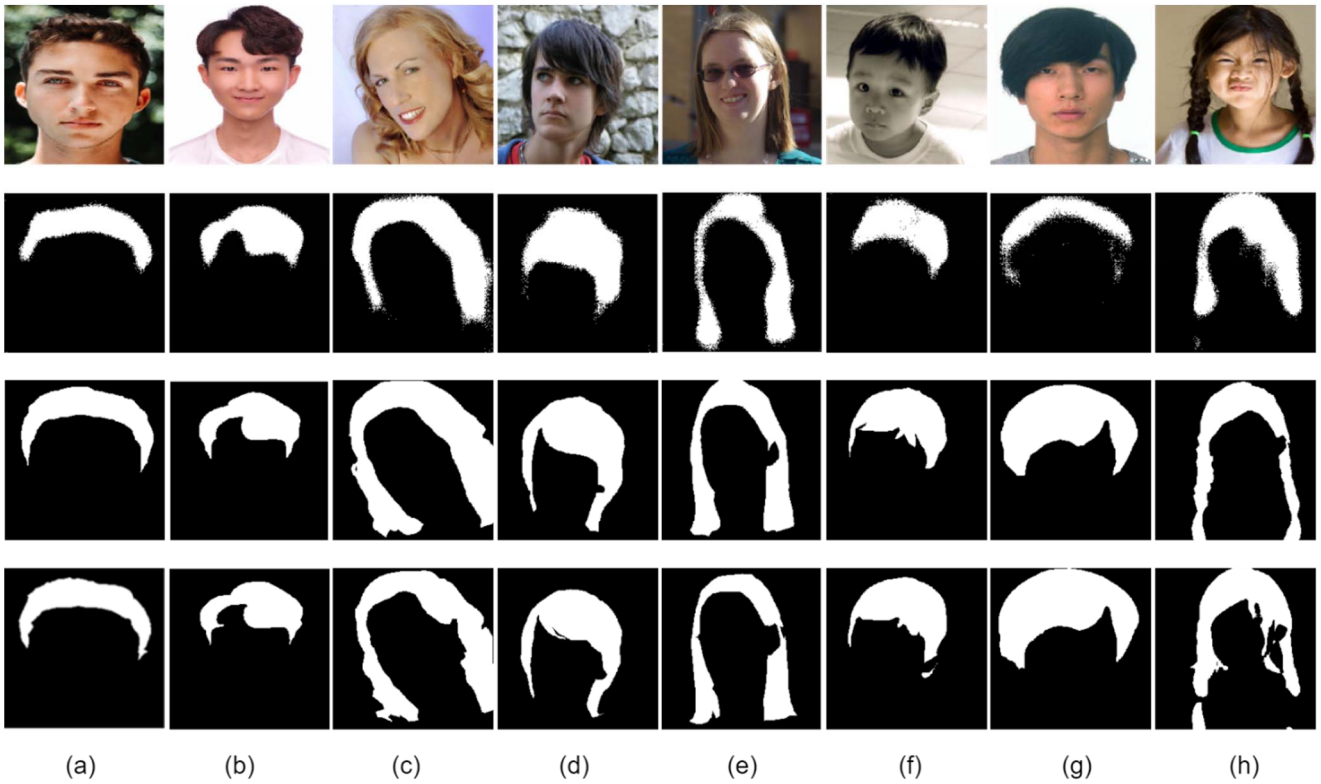

Figure. 7. The results of FCN, BiseNet, and our method. Row 1 shows the original input, row 2 shows the results of FCN, row 3 shows the results of BiseNet, and the last row shows the results of our method.

Figure 7 shows the results of FCN, BiseNet and our method. Where (a),(d) and (e) are in the complex background, (c) shows the complicated hairstyle, (h)shows facial occlusive. It can be seen that the last row shows more details than the other two rows. However in (h),we can see our method lose some details in the lower right corner and make some wrong segmentation due to the complexity of the facial hair. We think it's because of the lack of training data. But in terms of overall effect, our methods can not only make more precise results in the edge area, but also perform better in complex background, as shown in the fourth and fifth column.

Figure 8 shows where we want to zoom in to show the effect of our approach.

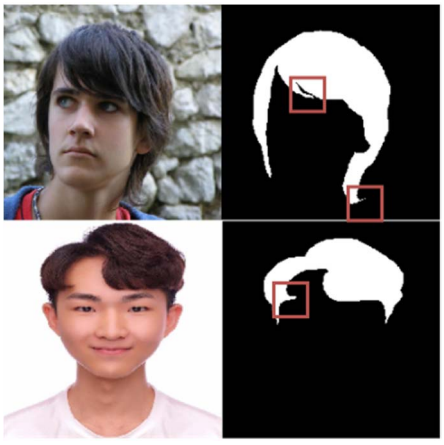

Figure.8. Segmentation results

IV. Conclusions and Future Work

We propose a precise hair parting system for photo editing applications. We implemented this using a two-phase approach. In the first stage, we used the improved BiseNet network to generate the segmentation of the hair region, specifically, we replace the common convolutional layers in context path with residual block, which can increase the depth of the network, make the network be capable of learning more semantic features of higher level. Also, we fuse the feature of different scales learned in this path to improve the segmentation results. We then generated the corresponding trimap from the output of this stage automatically. In the second stage, a refinement stage stacks the original image and the corresponding trimap into a four-channel input, which was sent into a small four-layer convolution to get more refined segmentation results. In this step, we also use the technology of skip connection. Experimental results showed that the proposed method achieves fine performance in hair segmentation work. In the future, we will explore the application of this method in other semantic segmentation tasks.

References

[1] Hu, LW, Ma, CY, Luo, LJ, Wei, LY, & Li, H.: Capturing braided hairstyles. Acm Transactions on Graphics, 33(6), 1-9(2014)

[2] Wang, D. , Shan, S. , Zeng, W. , Zhang, H. , & Chen, X. : A novel two-tier Bayesian based method for hair segmentation. Proceedings of the International Conference on Image Processing, ICIP 2009, 7-10 November 2009, Cairo, Egypt. IEEE (2009)

[3] Chai, Menglei, Wang, Lvdi, Weng, Yanlin, Jin, Xiaogang, & Zhou, Kun.:Dynamic hair manipulation in images and videos. Acm Transactions on Graphics, 32(4), 1-8(2013)

[4] Wang, D. , Chai, X. , Zhang, H. , Chang, H. , Zeng, W. , & Shan, S. : A novel coarse-to-fine hair segmentation method. Journal of Software, 24(10), 233-238(2011)

[5] Wang, D. , Shan, S. , Zhang, H. , Zeng, W. , & Chen, X. . (2014). Data-driven hair segmentation with isomorphic manifold inference. Image and Vision Computing, 32(10), 739-750(2014)

[6] Lee, K. C. , Anguelov, D. , Sumengen, B. , & Salih Burak Göktürk. :Markov random field models for hair and face segmentation. IEEE International Conference on Automatic Face and Gesture Recognition(2008)

[7] N. Wang, H. Ai, and F. Tang.:What are good parts for hair shape modeling? In: IEEE Conference on Computer Vision and Pattern Recognition(2012)

[8] Chai, M. , Shao, T. , Wu, H. , Weng, Y. , & Zhou, K. : Autohair: fully automatic hair modeling from a single image. Acm Transactions on Graphics, 35(4), 116-128(2016)

[9] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition(2017)

[10] Yu, Changqian, Wang, Jingbo, Peng, Chao, Gao, Changxin, Yu, Gang, & Sang, Nong. : Bisenet: bilateral segmentation network for real-time semantic segmentation. arXiv(2018)

[11] Lee, C. H. , Liu, Z. , Wu, L. , & Luo, P.. Maskgan: towards diverse and interactive facial image manipulation. arXiv (2019)

[12] Shelhamer, E. , Long, J. , & Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)

[13] Olaf Ronneberger, Philipp Fischer, & Thomas Brox.:U-Net: Convolutional Networks for Biomedical Image Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention(2015)

[14] Peng, C. , Zhang, X. , Yu, G. , Luo, G. , & Sun, J.: Large kernel matters -- improve semantic segmentation by global convolutional network. IEEE Conference on Computer Vision and Pattern Recognition (2017)

[15] Ghiasi, G. , & Fowlkes, C. C.: Laplacian pyramid reconstruction and refinement for semantic segmentation.In: European Conference on Computer Vision (2016)

[16] Guosheng Lin, Anton Milan, Chunhua Shen, & Ian Reid. RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)

[17] Yu, C. , Wang, J. , Peng, C. , Gao, C. , & Sang, N.: Learning a Discriminative Feature Network for Semantic Segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)

[18] Francois Chollet. :Xception: Deep Learning with Depthwise Separable Convolutions. IEEE Conference on Computer Vision and Pattern Recognition (2017)

[19] Hu, Jie, Shen, Li, Albanie, Samuel, Sun, Gang, & Wu, Enhua. . Squeeze-and-excitation networks. arXiv (2017)

[20] K. He, X. Zhang, S. Ren, & J. Sun.:Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (2015)

[21] Zhou, E. , Fan, H. , Cao, Z. , Jiang, Y. , & Yin, Q. :Extensive Facial Landmark Localization with Coarse-to-Fine Convolutional Network Cascade.In:IEEE International Conference on Computer Vision Workshops(2013)

[22] Xu, N., Price, Brian, Cohen, Scott, & Huang, Thomas. Deep image matting.arXiv:1703.03872(2017)