SGM-Net: Semantic Guided Matting Net

Qing Song · Wenfeng Sun · Donghan Yang · Mengjie Hu · Chun Liu

Abstract Human matting refers to extracting human parts from natural images with high quality, including human detail information such as hair, glasses, hat, etc. This technology plays an essential role in image synthesis and visual effects in the film industry. When the green screen is not available, the existing human matting methods need the help of additional inputs (such as trimap, background image, etc.), or the model with high computational cost and complex network structure, which brings great difficulties to the application of human matting in practice. To alleviate such problems, most existing methods (such as MODNet) use multi-branches to pave the way for matting through segmentation, but these methods do not make full use of the image features and only utilize the prediction results of the network as guidance information. Therefore, we propose a module to generate foreground probability map and add it to MODNet to obtain Semantic Guided Matting Net (SGM-Net). Under the condition of only one image, we can realize the human matting task. We verify
our method on the P3M-10k dataset. Compared with the benchmark, our method has significantly improved in various evaluation indicators.

**Keywords** Matting, Human Matting, Semantic Segmantation, Alpha Matte

1 Introduction

Semantic segmentation directly identifies the object category of each pixel, which belongs to rough semantics and is easy to blur the structural details. Human parsing is a fine-grained semantic segmentation task for human images, aiming to identify the components of human images at the pixel level [14, 41, 44]. Although human parsing enhances the processing of structure information, it is still essentially pixel-level coarse extraction. Different from them, image matting is the technique of extracting the foreground from a natural image by calculating its color and transparency. It can be used for background replacement, image synthesis, and visual effects, which has a wide application prospect in the film industry [20, 57].

Image matting needs to calculate the transparency of each pixel, which is more refined than human parsing. Specifically, for the input image $I \in \mathbb{R}^{H \times W \times 3}$, the extinction formula is decomposed into foreground $F \in \mathbb{R}^{H \times W \times 3}$, background $B \in \mathbb{R}^{H \times W \times 3}$, and alpha matte $\alpha \in \mathbb{R}^{H \times W}$ with a linear mixing assumption:

$$I = \alpha F + (1 - \alpha)B$$

(1)

where for color images, there are 7 unknown variables in the above expression and only 3 known variables, and thus, this decomposition is seriously limited [20].

Most of the existing matting methods need additional pictures as auxiliary inputs, such as additional background images [26, 33], and pre-defined trimaps [25, 37, 38]. However, taking additional background image as input requires that the two pictures must be aligned. Besides, the cost of pre-defined trimap is too high for us. Therefore, some latest work attempts to study the matting problem in a trimap-free setting. These researches include two directions: one is to study the alternatives to the trimap guidance and ease the requirements for manual input [11, 16, 28, 46]. For example, [11, 16] proposed technologies of automatic generation of trimap, while [46] proposed the progressive refinement network (PRN), which has good robustness to various types of masks as guidance. Another line of works try to get rid of any external guidance, hoping that the matting model can capture both semantics and details through end-to-end training on large-scale datasets [18, 27, 32, 47], and achieve the same level of video matting as video segmentation [24].

Compared with common objects, portraits have more abundant and complex details, such as hair, glasses, jewelry, etc. These factors make human matting more challenging. Therefore, in this work, we focus on the human matting and design a module to generate foreground probability map using the features in the segmentation network. This foreground probability map has the same size as the input image, which can be used as the guidance information to assist in the detail extraction of the human contour. By adding it to MODNet, we get a new human matting network SGM-Net, which uses a single RGB image as input to complete four tasks: extracting human
contour, generating foreground probability map, predicting detail information and fusing information, so as to predict accurate alpha matte.

We conduct a large number of experiments using P3M-10k dataset\cite{22} to evaluate the effectiveness of our method. Under the commonly used matting performance indicators, our method has excellent performance. For details, please refer to Sec\textsuperscript{4}. In addition, we capture a large number of natural human images from the Internet, which proves that our learning model can be extended to real-images.

![Fig. 1: Alpha matte results by MODNet and our SGM-Net from RGB input.](image)

2 Related Works

Image matting is to extract the desired target foreground from a given target image. Unlike binary mask output by image segmentation \cite{31,45} and human parsing \cite{40,42}, image/human matting needs to be given an alpha matte with accurate prediction probability of foreground for each pixel, which is represented by $\alpha$ in Eq\textsuperscript{1}. In this section, we will review the image matting methods related to our work.

2.1 Traditional Methods

Traditional image matting methods mostly predict alpha matte by sampling, propagation or using low-level features such as color \cite{1,10}. The sampling-based methods \cite{6,9,13} estimate foreground/background color statistics by sampling the pixels in the definite foreground/background regions, so as to solve the alpha matte of the unknown region. The propagation-based method \cite{3,12,19,21,35} estimate alpha matte by propagating the alpha value of the foreground/background pixels to the unknown region. However, the application effect of these methods in complex scenes is not ideal.
2.2 Trimap-based Methods

With the great progress of deep learning and the rise of computer vision technology, many methods based on convolutional neural network (CNN) are proposed and used in general image matting, which significantly improves the matting results. Cho et al. \[5\] and Shen et al. \[34\] introduced convolutional neural network into the traditional algorithm to reconstruct the alpha matte. Xu et al. \[38\] proposed an automatic encoder architecture, which takes RGB image and trimap as input and uses pure encoder-decoder network to directly predict alpha matte to achieve the state-of-the-art results. The trimap is a mask containing three regions: absolute foreground ($\alpha = 1$), absolute background ($\alpha = 0$) and unknown region ($\alpha = 0.5$). In this way, the matting algorithm only needs the prior information of two absolute regions to predict the masking probability of each pixel in the unknown region. Later, Lutz et al. \[30\] introduced a generative adversarial framework to improve the results. Tang et al. \[36\] proposed to combine the sampling-based method with deep learning. Hou et al. \[15\] proposed a two-encoder two-decoder structure for simultaneous estimation of foreground and alpha. With the development of the attention mechanism, \[39, 43\] introduced the attention mechanism into the human parsing task, which greatly improved the accuracy. Furthermore, research \[2, 29\] argued that the attention mechanism can effectively improve matting performance. \[25\] further improved the performance by introducing the contextual attention module.

2.3 Trimap-free Methods

Semantic estimation is needed to locate the approximate foreground before predicting the fine alpha matte, thus, it is difficult to denoise the image without the assistance of trimap.

Currently, trimap-free methods mostly focus on specific types of foreground objects, such as human and animals \[23, 47\] proposed a framework composed of a segmentation network and a fusion network, where the input is only a single RGB image. Then, \[28\] introduced a trimap-free framework, which consisting of mask prediction network, quality unification network and matting refinement network. Similarly, \[18\] proposed a framework composed of semantic estimation network, detail prediction network and semantic-detail fusion network, and introduced the attention mechanism — SE module \[17\]. As the representatives of another direction of trimap-free matting research, Chen et al. \[4\] realized foreground detail extraction by combining trimap generation network (T-Net) and matting network (M-Net). \[33\] introduced a framework that takes the background images along with other potential prior information (such as segmentation mask, motion cue, etc.) as additional inputs. \[46\] designed Progressive Refinement Network (PRN) to provide self-guidance for learning and progressively refines the uncertain matting areas by decoding. The model can use various types of mask guidance (such as trimap, rough binary segmentation mask or low-quality soft alpha matte) to obtain high-quality matting results and weaken the dependence of the model on trimap.
Our method directly learns the semantic information of the given images for the human foreground, generates foreground probability map while roughly extracting the human contour, extracts the fine details according to the foreground probability map and some semantic information, these are finally fused to generate accurate alpha mattes. Based on the original MODNet, a small amount of model operation cost is sacrificed to effectively improve the functioning of human matting.

3 Method

In this section, we will introduce our algorithm with more details.

3.1 Overview

Our SGM-Net is targeted to extract alpha mattes of specific semantic patterns – human. SGM-Net takes an image (usually 3 channels representing RGB) as the input and directly outputs a 1-channel alpha matte with the identical size of input, no auxiliary information (such as trimap or background image) is required. Fig. 2 shows its pipeline. Methods that are based on multiple models [4, 8, 34] have shown that regarding the trimap-free matting as a combination of trimap prediction (or segmentation) step and trimap-based matting step can obtain better performance, but the large amount of computation brought by multiple models is not as expected. We decompose the trimap-free matting task into multiple subtasks including semantic estimation, probability prediction, detail prediction and information fusion to expand and optimize this idea. Although our results are not able to surpass the trimap-based methods, they are better than the trimap-free methods based on multiple models.

The objective of SGM-Net is to obtain accurate alpha matte by generating coarse semantic classification information and fine foreground boundaries. As shown in Fig. 2, SGM-Net consists of three branches, which can learn different sub-objectives. Specifically, the low-resolution branch ($S_p$) of SGM-Net uses a segmentation network to estimate human semantics, and in this process, it does pixel-wise classification among foreground and background to generate foreground probability map. Based on it, the high-resolution branch ($D$) (supervised by the transition region ($\alpha \in (0, 1)$ in the ground truth matte) concatenates the input image (RGB) and foreground probability map to focus on portrait boundaries. Finally, at the end of the model, the fusion branch ($F$) (supervised by the ground truth matte) integrates rough human semantics and fine human boundary information to predict the final alpha matte. The whole networks are trained jointly in an end-to-end manner. We describe these branches in detail in the following sections.

3.2 Semantic Estimation and Foreground Probability Map Generation

Similar to existing trimap-free methods, the first step of SGM-Net is to locate the human in the input image $I$. we extract the high-level semantics of the image through an encoder (i.e., the low-resolution branch $S_p$ of SGM-Net), which makes the semantic
Fig. 2: Architecture of SGM-Net. Given an input image $I$, SGM-Net predicts human semantics $S_p(I)$, foreground probability map $F_p$, boundary details $d_p$, and final alpha matte $\alpha_p$. All branches are constrained by specify supervisions generated from the ground truth matte $\alpha_g$.

estimation more efficient. In addition, $S_p(I)$ is helpful for subsequent branches and joint optimization. Experiments in [18] show that some channels have more precise semantics than other channels, thus, the channel-wise attention mechanism can encourage using the right information and discourage those that are wrong. Therefore, we continue to use SE-block after $S_p$ to reweight the channels of $S_p(I)$ by extracting features of different scales. Further, in order to predict the rough semantic mask, we feed $S_p(I)$ into a convolution layer activated by the Sigmoid function to change its channel number to 1. Since the semantic mask is supposed to be smooth, we use L2 loss here, as shown below:

$$L_s = \frac{1}{2} \| S_{po} - G(\alpha_g) \|_2$$ (2)

where $G$ stands for Gaussian blur after $16 \times$ downsampling, which removes the detailed structures (such as hair) that are not important to human semantics.

According to Eq[1], the output of the matting network can be regarded as the probability prediction $\alpha\in(0, 1)$ that each pixel in the image is the foreground. Therefore, we extend the semantic segmentation process of this branch, add a probability prediction module that can fuse the feature information in the segmentation network and generate the probability prediction map of the foreground of the human. As shown in the Fig[3] The output of the module indicates the probability of each pixel in the input as the foreground, which can be used as guidance information to assist the high-resolution branch $D$ to complete the detail extraction of the human contour.
Fig. 3: Architecture of Foreground Probability Map Module. Given an image as input, the segmentation network extracts features of different scales, inputs all features into our module, uses encoder-decoder to recombine all features, and finally uses softmax to obtain the probability that each pixel of the original image is the foreground of a portrait. Visualize the probability map, the darker the color, the greater the probability that the pixel is the foreground of the person, and vice versa, it is more likely to be the background.

3.3 Detail Prediction

We process the unknown area around the foreground human by using a high-resolution branch $D$, which takes the input image $I$, foreground probability map $F_p$, and low-level features from $S_p$ as the input. Reusing low-level features can reduce the computational overheads of $D$ and improve the speed of the model. We take the concatenation of 3-channel image $I$ and the 3-channel foreground probability map $F_p$ as 6-channel input $I_t$ of $D$, as shown in Figure 3, the input after downsampling is concatenated with the low-level features in $S_p$ and transmitted to the encoder-decoder. In $D$, we use a skip link to reduce the impact of resolution change on prediction accuracy.

We use the dependency between sub-objectives, that is, foreground probability map and high-level human semantics $S_p(I)$ are a priori for detail prediction. We denote the output of $D$ as $D(I_t, S_p(I))$, calculate the portrait boundary details from it and learn it through L2 loss. The expression is as follows:

$$L_d = m_d \| d_p - \alpha_g \|_2$$

(3)

where $m_d$ is a binary mask that makes $L_d$ focus on the boundaries of the portrait. It comes from dilation and erosion on $\alpha_g$. The values are 1 if the pixels are inside the unknown region, otherwise, it is 0. Although $d_p$ may contain inaccurate values for the pixels with $m_d = 0$, the values in $d_p$ are highly accurate for the pixels with $m_d = 1$. 


3.4 Fusion Module

For the fusion branch, we use the concise CNN module to concatenate human semantics and boundary details. First, we upsample the output $S_p(I)$ of the semantic branch, match its shape with $D(I, S_p(I))$. We then concatenate them to predict the final alpha matte $\alpha_p$. The loss function of this process is as shown in Eq.4:

$$ L_\alpha = || \alpha_p - \alpha_g ||_1 + L_c^e $$

where $L_c$ is the compositional loss from [38]. It measures the absolute difference between the input image $I$ and the composite image, which is composed of prediction alpha matte $\alpha_p$, the ground truth foreground, and the ground truth background.

Our SGM-Net realizes end-to-end training by adjusting the weights of $L_s$, $L_d$ and $L_\alpha$, as:

$$ L = \lambda_s L_s + \lambda_d L_d + \lambda_\alpha L_\alpha $$

where $L_s$, $L_d$ and $L_\alpha$ are hyper-parameters that balances the three branch loss functions. Follow the settings in MODNet, we set $\lambda_s = \lambda_\alpha = 1$ and $\lambda_d = 10$.

4 Experiments

In this section, we compare SGM-Net with MODNet and other methods (such as DIM[38], AlphaGAN[30], SHM[4], etc.) on the P3M-10k [22] face-blurred images to verify the effectiveness of foreground probability map module. We also conduct further ablation experiments to evaluate the performance of SGM-Net in various aspects. Finally, we demonstrate the effectiveness of SGM-Net in adapting to real-world data.

4.1 Experiments Setting

We train all models on the same dataset and adopt the same training strategy, as follows:

**Dataset.** We evaluate all methods on the human matting dataset (P3M-10k face-blurred), which contains 9,421 training images and 500 testing images.

**Measurement.** Five metrics are used to evaluate the quality of predicted alpha matte: sum of absolute differences (SAD), mean squared error (MSE), mean absolute differences (MAD), gradient (Grad) and Connectivity (Conn). Among them, SAD, MSE and MAD measure and evaluate the pixel difference between the prediction and the ground truth alpha matte, while Grad and Conn measure clear details. In calculating all these metrics, we normalized the predicted alpha matte and ground truth to 0 to 1. Furthermore, all metrics are calculated over the entire images instead of only within the unknown regions and averaged by the number of pixels.

**Training Stage.** To make a fair comparison with MODNet, we use the pre-trained weight of ImageNet [7] to initialize the network. We train all networks on 8 NVIDIA Titan XP GPUs (input images are cropped to $512 \times 512$) with a batch size of 4. The
momentum of the SGD optimizer is set to 0.9 and the weight decay is $4.0\times10^{-5}$. The learning rate is initialized to 0.02, the training lasts 150 epochs, and is multiplied by 0.1 every 50 epochs.

4.2 Results on P3M-10k

In order to evaluate the effectiveness of our proposed method, we use ResNet-34 as the backbone of all trimap-free method and compare SGM-Net with other matting methods. As the objective and subjective results of different methods on P3M-10k shown in Table 1 and Fig. 4, our method is superior to MODNet and other trimap-free methods in all indicators, and even achieves competitive results with DIM based on trimap. The P3M-10k dataset contains two testing datasets, P3M-500-P and P3M-500-NP. P3M-500-P dataset blurs the identifiable faces, while P3M-500-NP does not, see the Image in Fig. 4 and Fig. 5. We also perform the same test on P3M-500-NP. The results are shown in Table 2 and Fig. 5.
Fig. 4: Subjective Results of Different Methods on P3M-500-P. We test several methods and show the results of some representative methods (DIM [38], MODNet [18] and ours) on P3M-500-P. Zoom it for the best visualization.

Table 3: Ablation of SGM-Net. FP-G: The module of foreground probability map generation. $S_p(I)$: Using the output of semantic branch as one of the inputs of the detail prediction branch or not. ‘↓’ means lower is better.

| PT-G | $S_p(I)$ | SAD ↓ | MSE ↓ | MAD ↓ | Grad ↓ | Conn ↓ |
|------|----------|-------|-------|-------|--------|--------|
| i    |          | 9.7328| 0.0031| 0.0056| 13.4755| 9.0085 |
| ii   | ✓        | 9.3406| 0.0029| 0.0054| 13.4160| 8.8537 |
| iii  | ✓        | 9.1552| 0.0027| 0.0053| 13.3744| 8.7354 |

4.3 Ablation Studies

We perform ablation study of SGM-Net on dataset P3M-500-P. It can be seen from Table 3 that compared with the MODNet, the network with the foreground probability map module can achieve better results. The results of Ex.i (MODNet) and Ex.ii show that the module we designed can effectively fuse the features of the segmented network, and generate a priori information to assist the detail prediction branch to complete the information extraction, e.g., 9.7328 SAD vs 9.34062 SAD, 9.0085 Conn vs 8.8537 Conn. The results of Ex.ii and Ex.iii (our method) show that the output of the segmented network ($S_p(I)$) will interfere with the branch extracted from the probability prediction map, and affect the actual effect of our module, e.g., 9.34062 SAD vs 9.15526 SAD of ours, 8.8537 Conn vs 8.7354 Conn of ours.
4.4 Results on Real-World Data

In order to study the ability of our model to extend to real-world data, we apply our model to a large number of real-world images for qualitative analysis. Fig. 6 shows some visual results. It can be seen that our method still has good performance even in a complex background. Note that the bouquet held by the female in the second image in Fig. 6 can also be well separated by our method. In addition, we can also separate the pets far away from the man’s body in the third image. These show the good performance of our method for human body extensions (small objects). The last column in Fig. 6 shows examples of synthesis with the help of automatic prediction of the foreground and the new background of the alpha matte, which have good visual quality.

5 Conclusion

In this paper, we focus on the problem of human matting. We design a foreground probability map generation module, add it to MODNet, and adjust the whole matting network accordingly to make the transition area smoother, so as to get SGM-Net. The use of green screen in human matting is avoided, and only RGB image is used as input to obtain a high-quality alpha matte. SGM-Net shows good performance on
The results of our method on the real-world data.

P3M-10k dataset and various real-world data, and is obviously better than MODNet. Although it is not as good as some matting methods based on trimap, the performance gap between them is greatly reduced.

References

1. Aksoy Y, Ozan Aydin T, Pollefeys M (2017) Designing effective inter-pixel information flow for natural image matting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 29–37
2. Chaudhari S, Mithal V, Polatkan G, Ramanath R (2021) An attentive survey of attention models. ACM Transactions on Intelligent Systems and Technology (TIST) 12(5):1–32
3. Chen Q, Li D, Tang CK (2013) Knn matting. IEEE transactions on pattern analysis and machine intelligence 35(9):2175–2188
4. Chen Q, Ge T, Xu Y, Zhang Z, Yang X, Gai K (2018) Semantic human matting. In: Proceedings of the 26th ACM international conference on Multimedia, pp 618–626
5. Cho D, Tai YW, Kweon I (2016) Natural image matting using deep convolutional neural networks. In: European Conference on Computer Vision, Springer, pp 626–643
6. Chuang YY, Curless B, Salesin DH, Szeliski R (2001) A bayesian approach to
digital matting. In: Proceedings of the 2001 IEEE Computer Society Conference
on Computer Vision and Pattern Recognition. CVPR 2001, IEEE, vol 2, pp II–II
7. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale
hierarchical image database. In: 2009 IEEE conference on computer vision and
pattern recognition, Ieee, pp 248–255
8. Deora R, Sharma R, Raj DSS (2021) Salient image matting. arXiv preprint
arXiv:210312337
9. Gastal ES, Oliveira MM (2010) Shared sampling for real-time alpha matting. In:
Computer Graphics Forum, Wiley Online Library, vol 29, pp 575–584
10. Grady L, Schiwietz T, Aharon S, Westermann R (2005) Random walks for inter-
active alpha-matting. In: Proceedings of VIIP, vol 2005, pp 423–429
11. Gupta V, Raman S (2016) Automatic trimap generation for image matting. In:
2016 International conference on signal and information processing (IConSIP),
Ieee, pp 1–5
12. He K, Sun J, Tang X (2010) Fast matting using large kernel matting laplacian
matrices. In: 2010 IEEE Computer Society Conference on Computer Vision and
Pattern Recognition, IEEE, pp 2165–2172
13. He K, Rhemann C, Rother C, Tang X, Sun J (2011) A global sampling method
for alpha matting. In: CVPR 2011, IEEE, pp 2049–2056
14. He Y, Yang L, Chen L (2017) Real-time fashion-guided clothing semantic parsing:
a lightweight multi-scale inception neural network and benchmark. In: AAAI
Workshops
15. Hou Q, Liu F (2019) Context-aware image matting for simultaneous foreground
and alpha estimation. In: Proceedings of the IEEE/CVF International Conference
on Computer Vision, pp 4130–4139
16. Hsieh CL, Lee MS (2013) Automatic trimap generation for digital image matting.
In: 2013 Asia-Pacific Signal and Information Processing Association Annual
Summit and Conference, IEEE, pp 1–5
17. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of
the IEEE conference on computer vision and pattern recognition, pp 7132–7141
18. Ke Z, Li K, Zhou Y, Wu Q, Mao X, Yan Q, Lau RW (2020) Is a green screen
really necessary for real-time portrait matting? arXiv preprint arXiv:201111961
19. Lee P, Wu Y (2011) Nonlocal matting. In: CVPR 2011, IEEE, pp 2193–2200
20. Levin A, Lischinski D, Weiss Y (2007) A closed-form solution to natural image
matting. IEEE transactions on pattern analysis and machine intelligence 30(2):228–
242
21. Levin A, Rav-Acha A, Lischinski D (2008) Spectral matting. IEEE transactions
on pattern analysis and machine intelligence 30(10):1699–1712
22. Li J, Ma S, Zhang J, Tao D (2021) Privacy-preserving portrait matting. In: Proceed-
ings of the 29th ACM International Conference on Multimedia, pp 3501–3509
23. Li J, Zhang J, Maybank SJ, Tao D (2022) Bridging composite and real: towards
end-to-end deep image matting. International Journal of Computer Vision pp 1–21
24. Li L, Zhou T, Wang W, Yang L, Li J, Yang Y (2022) Locality-aware inter-and intra-
video reconstruction for self-supervised correspondence learning. In: Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp
25. Li Y, Lu H (2020) Natural image matting via guided contextual attention. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 34, pp 11450–11457
26. Lin S, Ryabtsev A, Sengupta S, Curless B, Kemelmacher-Shlizerman I (2020) Real-time high-resolution background matting
27. Lin S, Yang L, Saleemi I, Sengupta S (2022) Robust high-resolution video matting with temporal guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 238–247
28. Liu J, Yao Y, Hou W, Cui M, Xie X, Zhang C, Hua X (2020) Boosting semantic human matting with coarse annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8563–8572
29. Lu H, Dai Y, Shen C, Xu S (2019) Indices matter: Learning to index for deep image matting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 3266–3275
30. Lutz S, Amplianitis K, Smolic A (2018) Alphagan: Generative adversarial networks for natural image matting. arXiv preprint arXiv:180710088
31. Minaee S, Boykov YY, Porikli F, Plaza AJ, Kehtarnavaz N, Terzopoulos D (2021) Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence
32. Qiao Y, Liu Y, Yang X, Zhou D, Xu M, Zhang Q, Wei X (2020) Attention-guided hierarchical structure aggregation for image matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 13676–13685
33. Sengupta S, Jayaram V, Curless B, Seitz SM, Kemelmacher-Shlizerman I (2020) Background matting: The world is your green screen. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2291–2300
34. Shen X, Tao X, Gao H, Zhou C, Jia J (2016) Deep automatic portrait matting. In: European conference on computer vision, Springer, pp 92–107
35. Sun J, Jia J, Tang CK, Shum HY (2004) Poisson matting. In: ACM SIGGRAPH 2004 Papers, pp 315–321
36. Tang J, Aksyoy Y, Oztireli C, Gross M, Aydin TO (2019) Learning-based sampling for natural image matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3055–3063
37. Wang J, Cohen MF (2008) Image and video matting: a survey
38. Xu N, Price B, Cohen S, Huang T (2017) Deep image matting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2970–2979
39. Yang L, Song Q, Wu Y, Hu M (2018) Attention inspiring receptive-fields network for learning invariant representations. IEEE transactions on neural networks and learning systems 30(6):1744–1755
40. Yang L, Song Q, Wang Z, Jiang M (2019) Parsing r-cnn for instance-level human analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 364–373
41. Yang L, Song Q, Wang Z, Hu M, Liu C, Xin X, Jia W, Xu S (2020) Renovating parsing r-cnn for accurate multiple human parsing. In: European Conference on
42. Yang L, Song Q, Wang Z, Liu Z, Xu S, Li Z (2021) Quality-aware network for human parsing. arXiv preprint arXiv:210305997
43. Yang L, Song Q, Wu Y (2021) Attacks on state-of-the-art face recognition using atten-tional adversarial attack generative network. Multimedia tools and applications 80(1):855–875
44. Yang L, Liu Z, Zhou T, Song Q (2022) Part decomposition and refinement network for human parsing. IEEE/CAA Journal of Automatica Sinica 9(6):1111–1114
45. Yu B, Yang L, Chen F (2018) Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11(9):3252–3261
46. Yu Q, Zhang J, Zhang H, Wang Y, Lin Z, Xu N, Bai Y, Yuille A (2021) Mask guided matting via progressive refinement network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1154–1163
47. Zhang Y, Gong L, Fan L, Ren P, Huang Q, Bao H, Xu W (2019) A late fusion cnn for digital matting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7469–7478