Fast Portrait Matting Using Spatial Detail-Preserving Network

Shaofan Cai[1], Biao Leng[1(✉)], Guanglu Song[1], and Zheng Ge[2]

[1] School of Computer Science and Engineering, Beihang University, Beijing 100191, China
lengbiao@buaa.edu.cn

[2] Graduate School of Information, Production and Systems, Waseda University, Kitakyushu 169-8050, Japan

Abstract. Image matting plays an important role in both computer vision and graphics applications. Natural image matting has recently made significant progress with the assistance of powerful Convolutional Neural Networks (CNN). However, it is often time-consuming for pixel-wise label inference. To get higher quality matting in an efficient way, we propose a well-designed SDPNet, which consists of two parallel branches—Semantic Segmentation Branch for half image resolution and Detail-Preserving Branch for full resolution, capturing both the semantic information and image details, respectively. Higher quality alpha matte can be generated while largely reducing the portion of computation. In addition, Spatial Attention Module and Boundary Refinement Module are proposed to extract distinguishable boundary features. Extensive Experiments show that SDPNet provides higher quality results on Portrait Matting benchmark, while obtaining 5x to 20x faster than previous methods.

Keywords: Portrait · Fast matting · Detail-preserving · Deep learning

1 Introduction

Matting refers to the problem of accurate foreground estimation in images and videos. It is one of the key techniques in many image editing and film production applications. Mathematically, the input image can be modeled as a convex combination of a foreground and background colors as follows [7]:

$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i \tag{1}$$

where I_i, F_i, B_i and α_i denote the natural RGB image, foreground, background color and alpha matte at pixel i respectively. Thus, for a three-channel color image, at each pixel, there are 7 unknown values but only 3 known values.

Given an input image I, finding F, B, and α simultaneously without any user interaction makes natural matting problem highly ill-posed. Image matting

L. Cheng et al. (Eds.): ICONIP 2018, LNCS 11306, pp. 329–339, 2018.
https://doi.org/10.1007/978-3-030-04224-0_28

Fig. 1. (a) Image from our Synthetic dataset(1280 × 960 pixels). (b) Trimap. Red color stands for definite foreground and blue color stands for definite background. The rest of the region stands for unknown. (c) Result of Our SDPNet. The running time is 40ms on GPU. (d) Our labeled groundtruth.

techniques [5,11] require a trimap (or strokes) indicating definite foreground, definite background and unknown region. Traditional matting algorithms can be divided into two classes, color sampling based methods and matting affinity based methods. The limitation of these methods is that the distinguishing feature largely rely on color. When the color distributions overlay between the foreground and background, it is really tough for such approaches to generate clear alpha matte without low-frequency "smearing" or high-frequency "chunky" artifacts. To overcome this problem, recently deep learning based methods are proposed for image matting. Instead of relying primarily on color information, CNN also extracts structure and semantic information, which helps to produce high quality alpha matte (Fig. 1(c)).

Although CNN [10] provides powerful assistance for image matting, amount of the huge parameter and calculation make it expensive for multi-megapixel images produced by digital cameras. Shen et al. [14] proposed an automatic matting with the help of semantic segmentation [12]. But their approach has a high computational complexity. Zhu et al. [20] designed a fast and effective method for portrait matting. It can realize real-time matting on the mobile phone for a low-resolution image. However, their approaches fail to distinguish tiny details in the hair areas because they downsample the input size of image to 128×128. When the resolution of input image get higher, the speed of inference will be largely limited and it is not detail-preserving.

In this paper, we focus on fast portrait matting techniques with decent prediction accuracy. To achieve our goal, we propose a network, named Spatial Detail-Preserving Network(SDPNet). Different from previous single branch matting network [14,20], our SDPNet uses two branch to utilize processing efficiency of low-resolution images and high inference quality of high-resolution ones. The idea is that low-resolution images can go through the full semantic segmentation network first for a coarse score map. The second branch is used to capture details structure to refine the coarse semantic map. Then the output of two branches

will be aggregated to generate a high quality alpha matte. We also consider the impact of different pixels in full-resolution feature map to improve matting performance. Our contributions in this work are as follows:

- A Spatial Detail-Preserving network (SDPNet) is proposed, which utilizes semantic and structure information in lower-resolution branch along with details from higher-resolution branch efficiently.
- Further more, we present Spatial Attention Module to improve the quality of feature map via spatial embedding. Boundary Refinement Module is adopted to refine the boundary of feature map produced by Semantic Segmentation Branch.
- Experiments show that our proposed method achieves 5x+ speed of inference. SDPNet can run at resolution 800×600 in speed of 40 fps while accomplishing high-quality portrait alpha matte.

2 Related Work

2.1 Natural Image Matting

Natural image matting is crucial for image and video editing, but it remains challenging because it is a severely underconstrained problem. Interactive image matting aims to predict alpha matte in unknown regions. [7] tried to apply Gaussian mixture models on both background and foreground. To infer the alpha matte in the unknown regions, closed-form matting [11] uses a matting Laplaian matrix, under a color line assumption. Large-Kernel Laplacian [9] helps accelerating matting Laplacian computation. Shared matting [8] was the first real-time matting algorithm running on modern GPUs by shared sampling. Inter-Pixel Information Flow Matting [1] proposed a purely affinity-based natural image matting method.

Recently, deep-learning based methods have shown great potential on solving computer vision tasks. DCNN [6] is the first attempt to apply deep learning on image matting problem. They used a relatively shallow neural network to deal with patches of images, with the result of closed-form and KNN matting as extra input. Xu [18] released a large matting dataset with high-quality foreground and alpha matte. Then they trained an encoder-decoder structure network on this dataset.

2.2 Semantic Segmentation

Traditional semantic segmentation methods adopt hand-craft feature to learn the representation. Recently, CNN based methods largely improve the performance. FCN [12] is the pioneer work to use fully convolution layers in semantic segmentation task. Encoder-decoder structures [2] can restore the feature map from higher layers with spatial information from lower layers. ICNet [19] incorporates multi-resolution branches under label guidance to achieve realtime inference without significantly reducing performance.

3 Proposed Algorithm

As shown in Fig. 2, the proposed SDPNet consists of two branches, Semantic Segmentation Branch, and Detail-Preserving Branch, which respectively captures the structure and details components of the input image. Specially, the input size of Detail-Preserving branch is full resolution ($h \times w$), and the input size of Semantic Segmentation Branch is lower resolution ($e.g.$ $\frac{h}{2} \times \frac{w}{2}$), with input image height h and width w. Given a high-resolution image and trimap, each branch has different functionalities. The Semantic Segmentation Branch provides the roughly boundary and semantic information of the image from lower resolution. The Detail-Preserving Branch captures the detail information, such as points, lines or edges, from full resolution. Finally, the feature maps from two branches are fused together, resulting in a high quality alpha matte.

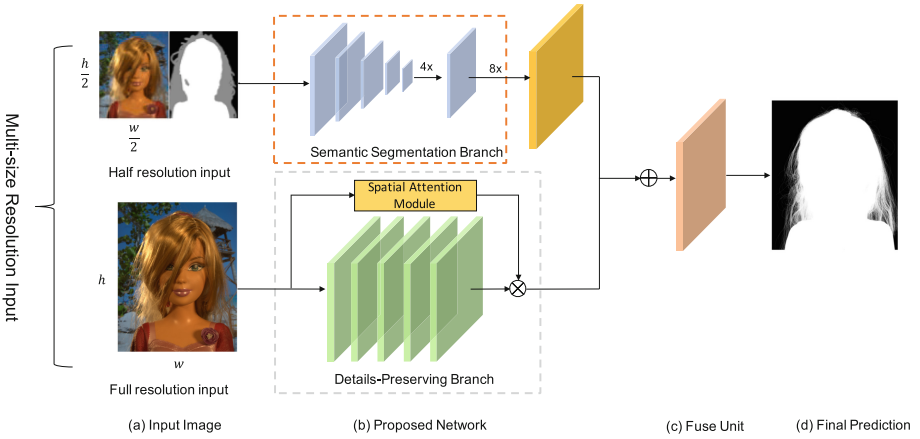


Fig. 2. Overall architecture of SDPNet. It contains two branches, Semantic Segmentation Branch and Detail-Preserving Branch and a feature fusing module. The Semantic Segmentation Branch (Sect. 3.1) generates a rough boundary mask from half resolution and the Detail-Preserving Branch (Sect. 3.2) captures details and structures from full resolution. Detail-Preserving Branch contains a Spatial Attention Module. Finally SPDNet fuses results from two branches by Feature Fusing Unit. The whole SDPNet is end-to-end trainable.

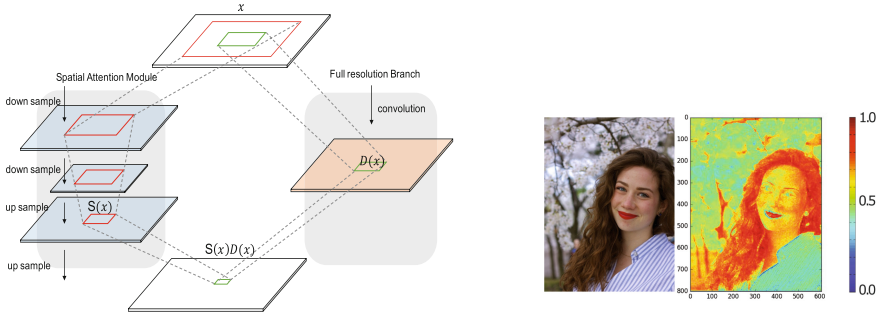
3.1 Semantic Segmentation Branch

Image resolution is the most critical factor that affect speed, since above analysis shows a half-resolution image only uses nearly quarter time compared to the full-resolution one. A naive approach is to directly use small-resolution image as input. We downsample images with ratios 1/2 and feed the resulting images into our Semantic Segmentation Branch. The detail structure of Semantic Segmentation Branch is shown in Table. This sub-network consists of an Encoder and a Decoder. Similar to Unet [3], we employ skip connections in encoder-decoder

network. The encoder network consists of one convolutional layer and 11 resnext [17] blocks. The decoder network uses deconvolution as upsampling module. At each stage after upsampling, the feature maps are fed to Boundary Refinement Modules, which will be illustrated later.

3.2 Detail-Preserving Branch

The tiny structures and details components of image will be destroyed during downsampling operations, such as max pooling or convolution with stride 2. Hence, we design a Detail-Preserving Branch to capture low-level features that are missing in the half-resolution branch. We can limit the number of convolutional layers since half-resolution branch already catches most semantically information. Here we use only three convolutional layers with kernel stride size 3×3 and stride 1 to extract low-level features. The Details structure of this branch is shown in Fig. 3.



(a) Spatial Attention Module (b) Score map of full resolution

Fig. 3. (a) Details of structures in spatial attention module. (b) The score map generated by spatial attention module.

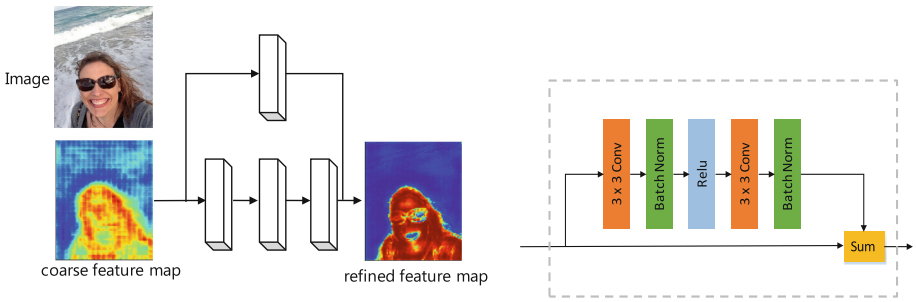
3.3 Spatial Attention Module

Spatial Attention Module aims at improving Detail-Preserving Branch features. Following previous attention mechanism in [15], we apply Spatial Attention Module in Detail-Preserving Branch. The module's target is to output scores for each pixel of feature maps. Given the input image I and trimap T with height h and width w, max pooling are performed several times to increase the receptive field rapidly after a small number of convolution layers. After reaching the lowest resolution, the global information is then expanded by a symmetrical upsample operations. We use linear interpolation up sample the output after one 1×1 convolution layer with stride 1. The number of upsampling module is the same as max pooling to keep the output score map size the same as the input feature map. Then we use a sigmoid layer to normalize the output score maps range to $[0, 1]$. The full module is illustrated in Fig. 3(a). It also shows that the consecutive

up-sample and down-sample operations can expand receptive filed. Experiment in is conducted to verify this.

3.4 Boundary Refinement Module

We propose a Boundary Refinement Block, schematically depicted in Fig. 4(b). The feature maps after upsampling go through the Boundary Refinement Block, which is designed to model the boundary alignment as a residual structure. More specially, we use $\widetilde{S}$ denote refined score map: $\widetilde{S} = S + \mathcal{R}(S)$, where S is the coarse score map, and $\mathcal{R}(\cdot)$ is residual branch. After refinement, the boundary information is embeded in its output feature map, as show in Fig. 4(a).



(a) Feature map Refinement.

(b) BR: Boundary Refinement Module.

Fig. 4. (a) Refine feature map from coarse to fine. (b) Components of the boundary refinement blocks (BR)

4 Experiments

In this section, we evaluate the performance of SDPNet on publicly available 2K Portrait Matting Dataset [14] and our Synthetic Portrait Matting Dataset.

4.1 Datasets

2K Portrait Matting Dataset: We choose the primary dataset from [14], which is collected from Flickr. We evaluate the proposed method on the benchmark dataset. This dataset collects 2000 portrait image with labeled alpha matte as ground truth. These images are split into the training and testing sets with 1700, 300 images respectively.

Synthetic Portrait Matting Dataset: We further evaluate our method using real-world examples. We download some pictures, whose background color is pure, from Internet and manually label the trimap. With the selected portrait images, we labele alpha values with intensive user interaction tools provided by

[16] to make sure they are with high quality. The alpha matte is calculated while labeling. After this labeling process, we collect 200 image with high-quality alpha mattes. These images are randomly split into the training and testing sets with 150 and 50 images respectively. We also download some background pictures in real scenes. We randomly sample N background images in them and composite the portrait foreground onto those background images. Finally we got 20,000 (N = 100) training portrait images and 50 (N = 1) test images.

4.2 Implementation Details

Inspired by the work [4], we use the "poly" learning rate policy in which current learning rate is defined as the base one multiplying $(1 - \frac{iter}{max_{iter}})^{power}$. We set base learning rate to 0.001 and power to 0.9. Momentum and weight decay are set to 0.9 and 0.0001 respectively. The proposed network is trained on the training set above. To avoid overfitting, we randomly crop a 480 × 480 patch and this patch can cover the unknown region in the trimap. In order to generate trimaps for training, we randomly dilate the alpha matte by random size to make our network more robust to different quality of trimap. For data augmentation, we adopt random flip and random resize between 0.75 and 1.5 for all images, and additionally add random rotation between −45 and 45°. We also apply random Gamma transforms to increase color variation.

4.3 Accuracy Measure

We select the gradient error and mean squared error to measure matting quality, which can be expressed as:

$$G(\alpha^p, \alpha^{gt}) = \frac{1}{T} \sum_i \parallel \bigtriangledown \alpha_i^p - \bigtriangledown \alpha_i^{gt} \parallel \tag{2}$$

$$MSE(\alpha^p, \alpha^{gt}) = \frac{1}{T} \sum_i (\alpha_i^p - \alpha_i^{gt})^2 \tag{3}$$

where α^p is the predicted alpha matte and α^{gt} is corresponding ground truth. T is the number of pixels in unknown region of given trimap. $\bigtriangledown$ is the operator to compute gradients. Specially, alphamatting [13] points out that the correlation of SAD and MSE with the percption of average human observer is rather low, Gradient Error, which is more reliable, outperforms both of other two metric with a higher correlation.

4.4 Ablation Study on Synthetic Portrait Matting Datasets

In this subsection, we will step-wise decompose our approach to revel the effect of each component. In the following experiments, we evaluate all comparisons on Synthetic Portrait Matting dataset.

Ablation for Boundary Refinement Module: To refine the coarse feature scores after upsampling, we use our Boundary Refinement Module to refine score map. As show in Table 1, this module further improves the performance on two metrics – – MSE and gradient error. It reduces gradient error from 24.70 to 22.92 and MSE from 0.0133 to 0.0113.

Ablation for Detail-Preserving Branch: By contrast, the proposed SDP-Net is motivated by the decomposition of a image signal into structure and details. For fair comparison, we keep the same amount of calculations of the single Semantic Segmentation Branch's and Two Branch Network, show as Table 1. Especially, gradient error decreases dramatically from 24.7 to 20.45, which is an obvious improvement.

Ablation for Spatial Attention Module: We evaluate the effectiveness of spatial attention learning mechanism. As show in Table 1, the network trained using spatial attention module consistently outperform the networks without it, which proves the effectiveness of our method.

Table 1. The quantitative comparisons of proposed SDPNet on the Synthetic Portrait testing dataset. **SS:** Semantic Segmentation. **BR:** Boundary Refinement Module. **SA:** Spatial Attention Module. **DP:** Detail-Preserving Branch.

Method	Grad Error	MSE
SS Branch	24.70	0.0133
SS Branch + BR	22.92	0.0113
SS Branch + DP Branch	20.45	0.0126
SS Branch + DP Branch + BR	20.12	0.0113
SS Branch + DP Branch + BR +SA	**19.63**	**0.0107**

4.5 Comparison with State-of-the-Art Methods on 2k Portrait Matting Dataset

To further confirm the performance of our method, we also compare our methods with others. We visually and quantitatively evaluate our methods in 2k-Portrait Matting Dataset [14].

Quantitative Analysis. In experiments, we quantitatively evaluate the SDP-Net on 2k Portrait Matting Dataset [14] and compared it with DAPM [14] and LDN+FB [20]. We also use FCN [12] to generate trimap, then using closed-form [11] to calculate alpha matte. As show in Table 2, our method achieves lower gradient error than other two deep learning based methods.

Running Time. We evaluate our method and state-of-the-art methods on the same PC with an Intel(R) Core i7 CPU and a Nvidia Titan X GPU. Table 3 shows speed comparison between our method and other methods. Running time

Table 2. Results on 2k-Portrait Matting of [14]. DAPM means the approach of Deep Automatic Portrait Matting in [14]. LDN +FB means the approach in [20].

Method	Trimap-FCN [12] + Closed-form [11]	DAPM [14]	LDN+FB [20]	Ours
Grad ($\times 10^{-3}$)	4.14	3.03	7.40	**2.48**

Table 3. Speed comparison with other methods. Running time for a 800 × 600 image. All the method run by their publicly available scripts except for DIM [18], which we implement as its paper. **G:**GPU. **C:**CPU

Method	Closed-form [11]	Shared [8]	Info [1]	DIM [18](G)	Ours(C)	Ours(G)
Time (sec)	9.88	63.65	9.15	0.23	**1.76**	**0.024**

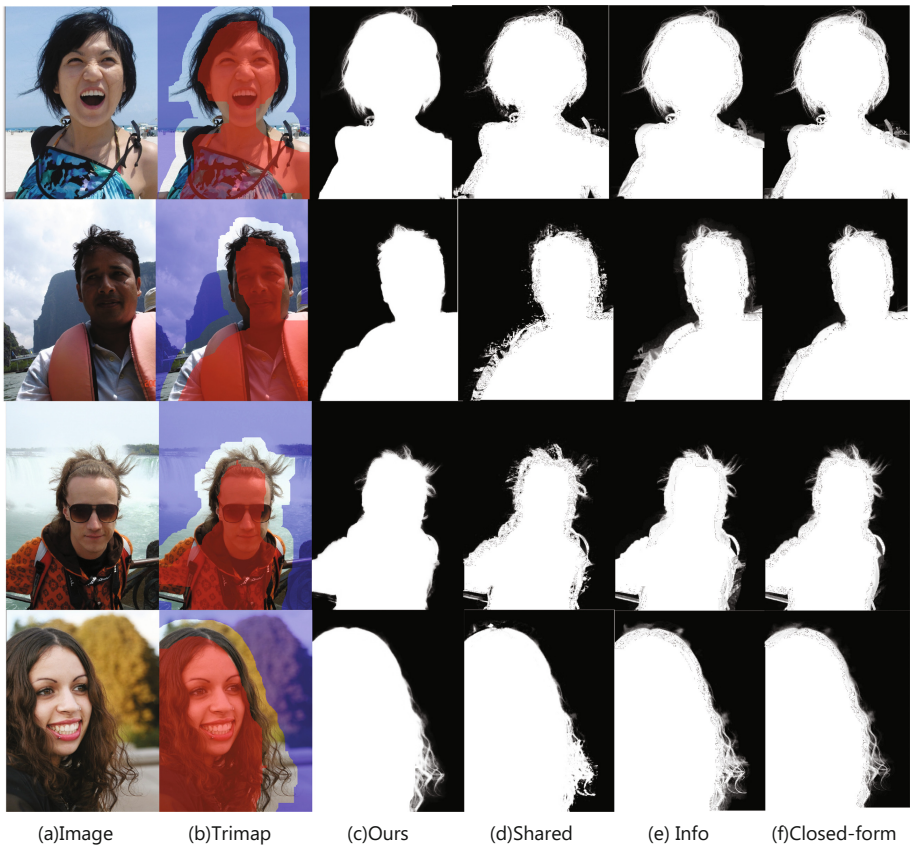


Fig. 5. Visual comparisons on 2k portrait matting dataset. (a) Image (b) Trimap (c) Ours (d) Shared-Matting [8] (e) Information-flow [1] (e) Closed-form [11]

for a 800 × 600 image, our SDPNet is nearly 5.6 times faster than closed-form [11], 36.11 times faster than shared-matting [8] and 5.19 times faster than Information-flow [1] matting on CPU. DIM [18] achieves state of the art performance in public available test set, but it is very time-consuming for a large resolution input. SDPNet is almost 10 times faster than DIM, while still generate alpha matte with fine details. Visually Comparison is showed in Fig. 5.

5 Conclusion

This paper proposes the Spatial Detail-Preserving Network (SDPNet) for fast portrait matting. SDPNet can simultaneously capture semantic structure and low-level details by its network design, which contains two branches: Semantic Segmentation Branch for lower resolution and Detail-Preserving Branch for full resolution. With the spatial attention mechanism and stage-wise refinement, our approach can capture the discriminative features for portrait matting. Our experimental results show that the proposed approach indeed takes less time for inference. Besides, SDPNet can also improve the quality of alpha matte, which shows our approach is comparable with the state-of-the-art matting methods.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (No. 61472023) and Beijing Municipal Natural Science Foundation (No. 4182034).

References

1. Aksoy, Y., Aydın, T.O., Pollefeys, M., Zürich, E.: Designing effective inter-pixel information flow for natural image matting. In: Computer Vision and Pattern Recognition (CVPR) (2017)
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
3. Barkau, R.L.: UNET: one-dimensional unsteady flow through a full network of open channels. Technical report, Hydrologic Engineering Center Davis CA (1996)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. arXiv preprint arXiv:1606.00915 (2016)
5. Chen, Q., Li, D., Tang, C.K.: KNN matting. IEEE Trans. Pattern Anal. Mach. Intell. **35**(9), 2175–2188 (2013)
6. Cho, D., Tai, Y.-W., Kweon, I.: Natural image matting using deep convolutional neural networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 626–643. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_39
7. Chuang, Y.Y., Curless, B., Salesin, D.H., Szeliski, R.: A bayesian approach to digital matting. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 2, p. 2. IEEE (2001)

8. Gastal, E.S., Oliveira, M.M.: Shared sampling for real-time alpha matting. In: Computer Graphics Forum, vol. 29, pp. 575–584. Wiley Online Library (2010)
9. He, K., Sun, J., Tang, X.: Fast matting using large kernel matting laplacian matrices. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2165–2172. IEEE (2010)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
11. Levin, A., Lischinski, D., Weiss, Y.: A closed-form solution to natural image matting. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 228–242 (2008)
12. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
13. Rhemann, C., Rother, C., Wang, J., Gelautz, M., Kohli, P., Rott, P.: A perceptually motivated online benchmark for image matting. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1826–1833. IEEE (2009)
14. Shen, X., Tao, X., Gao, H., Zhou, C., Jia, J.: Deep automatic portrait matting. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 92–107. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_6
15. Wang, F., et al.: Residual attention network for image classification. arXiv preprint arXiv:1704.06904 (2017)
16. Wang, J., Agrawala, M., Cohen, M.F.: Soft scissors: an interactive tool for realtime high quality matting. ACM Trans. Graph. (TOG) **26**(3), 9 (2007)
17. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5987–5995. IEEE (2017)
18. Xu, N., Price, B., Cohen, S., Huang, T.: Deep image matting. In: Computer Vision and Pattern Recognition (CVPR) (2017)
19. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: ICNet for real-time semantic segmentation on high-resolution images. arXiv preprint arXiv:1704.08545 (2017)
20. Zhu, B., Chen, Y., Wang, J., Liu, S., Zhang, B., Tang, M.: Fast deep matting for portrait animation on mobile phone. In: Proceedings of the 2017 ACM on Multimedia Conference, pp. 297–305. ACM (2017)