

Automatically Extract Semi-Transparent Motion-Blurred Hand From a Single Image

Xiaomei Zhao¹ and Yihong Wu

Abstract—When we use video chat, video game, or other video applications, motion-blurred hands often appear. Accurately extracting these hands is very useful for video editing and behavior analysis. However, existing motion-blurred object extraction methods either need user interactions, such as user supplied trimaps and scribbles, or need additional information, such as background images. In this letter, a novel method which can automatically extract the semi-transparent motion-blurred hand just according to the original RGB image is proposed. The proposed method separates the extraction task into two subtasks: alpha matte prediction and foreground prediction. These two subtasks are implemented by Xception based encoder-decoder networks. The images of extracted motion-blurred hands are calculated by multiplying the predicted alpha mattes and foreground images. Experiments on synthetic and real datasets show that the proposed method has promising performance.

Index Terms—Motion-blurred hand, semi-transparent, alpha matte prediction, foreground prediction, automatically.

I. INTRODUCTION

HAND language is one of the most important human gesture languages. Poor hand extraction results can greatly reduce the performance of video editing and behavior analysis. Extracting hands can be implemented by hand segmentation methods [1], [2]. However these methods can't deal with motion-blurred hands, which are very common in practical applications. Traditional methods [3]–[5], which were designed to predict the alpha mattes or foreground images of motion-blurred objects, generally needed user interactions [3], [4] or short-exposure frames [5]. Zhao *et al.* [6] proposed a deep learning network to predict the alpha mattes of motion-blurred hands, and then extracted motion-blurred hands by subtracting background components from the original images. A simple flow chart of this method is shown in Fig. 1(a). An obvious drawback of this method is that it needs background images, which are inconvenient to obtain. In this letter, we propose a method which can automatically extract semi-transparent motion-blurred hands just according

Manuscript received July 5, 2019; revised August 28, 2019; accepted August 29, 2019. Date of publication September 5, 2019; date of current version September 19, 2019. This work was supported by the National Natural Science Foundation of China under Grant 61836015, Grant 61421004, and Grant 61572499. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Kai Liu. (*Corresponding author:* Yihong Wu.)

The authors are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zhaoxiaomei14@mails.ucas.ac.cn; yhwu@nlpr.ia.ac.cn).

This letter has supplementary downloadable material available at <http://ieeexplore.ieee.org>, provided by the authors.

Digital Object Identifier 10.1109/LSP.2019.2939754

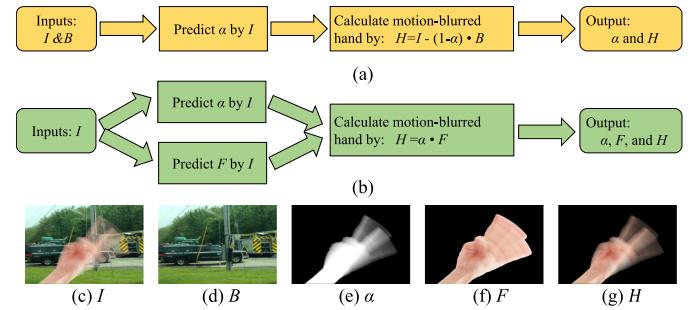


Fig. 1. Flow charts of different automatic motion-blurred hand extraction methods: (a) the method proposed in [6]; (b) the method proposed in this letter; (c-g) show examples of different kinds of images, where I denotes original image, B denotes background image, α denotes alpha matte, F denotes foreground image, H denotes the extracted motion-blurred hand.

to the original RGB images, without requiring any additional information.

An image I , which contains a motion-blurred hand, is made up by combining foreground hand F and background B : $I = \alpha \cdot F + (1 - \alpha) \cdot B$, $\alpha \in [0, 1]$, where α is called alpha matte, demonstrating the transparency of motion-blurred hand; F demonstrates the color of motion-blurred hand. Therefore, both α and F are related to motion-blurred hands. The proposed method separates the task of extracting semi-transparent motion-blurred hands into two subtasks: alpha matte prediction and foreground prediction. The extracted hands are calculated by multiplying the predicted α and F . A simple flow chart of the proposed method is shown in Fig. 1(b).

Alpha mattes can be calculated by matting methods [7]–[11]. However, most of matting methods need additional information, such as user supplied trimaps and scribbles [7]–[11]. In order to avoid the need of user interactions, several matting methods [12]–[14] employ CNN networks to predict trimaps explicitly or implicitly. The above matting methods focus on static objects, rather than motion-blurred objects. Zhao *et al.* [6] proposed a motion-blurred object matting network which only uses RGB images as inputs and directly outputs the predicted alpha mattes. In this letter, our alpha matte prediction network is developed from the matting network in [6] by adding a perceptual loss [15].

Up to now, most of matting methods concentrate on alpha matte prediction. Very few methods can predict foreground images. A recently proposed sampling- and learning-based matting method [11] can estimate the foreground color of unknown regions. However, this method needs user supplied trimaps to annotate the background, foreground, and unknown regions. Besides, this method focuses on static objects, rather than

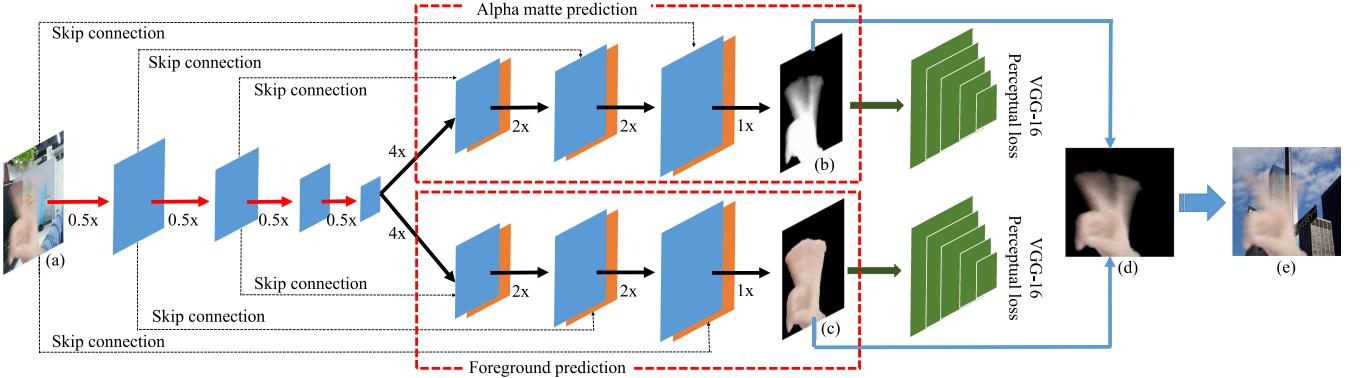


Fig. 2. The network architecture of the proposed method, where the alpha matte prediction network and foreground prediction network share the same Xception [16] based encoder and have different decoders. In this figure, (a) is the original image; (b) is the predicted alpha matte; (c) is the predicted foreground; (d) is the extracted motion-blurred hand which is calculated by multiplying (b) and (c); (e) is the new image with new background. In this figure, $0.5 \times$, $4 \times$, $2 \times$, $1 \times$ denote the downsampling and upsampling ratios.

motion-blurred objects. In this letter, a network which can automatically predict the foreground images of motion-blurred hands is proposed. The foreground prediction network is a Xception [16] based encoder-decoder network. It only uses the original RGB images as inputs and directly outputs the predicted foreground images. During training, L1-loss and perceptual loss [15] are employed. As shown in Fig. 2, the proposed foreground prediction network shares the same encoder with alpha matte prediction network. But each prediction network has its own independent decoder.

Human matting methods [12], [13] extracted human just by $\alpha \cdot I$, rather than $\alpha \cdot F$. A human image I can also be formulated as $I = \alpha \cdot F + (1 - \alpha) \cdot B$. Therefore, in semi-transparent areas where $0 < \alpha < 1$, $\alpha \cdot I$ still contains background information. However, in human images, almost all of the pixels, whose alpha values are between 0 and 1, are located at isolated hairs. Isolated hairs are very thin and have low transparency. Thus, in human images without motion blur, the background information contained in $\alpha \cdot I$ is not obvious and $\alpha \cdot I \approx \alpha \cdot F$. In contrast, in motion-blurred hand images, the areas with high transparency are large. If we extract motion-blurred hands by $\alpha \cdot I$, the extracted hand images will contain obvious background information, which can greatly reduce the sense of reality when changing background. Thus, motion-blurred hands should be calculated by $\alpha \cdot F$, rather than $\alpha \cdot I$.

In summary, the main contributions in this letter are:

- 1) A novel framework which can automatically extract the semi-transparent motion-blurred hand from a single image is proposed. This framework consists of two main parts: alpha matte prediction and foreground prediction. The images of extracted hands are calculated by multiplying the predicted alpha mattes and foreground images.
- 2) Alpha matte prediction and foreground prediction are implemented by Xception-based encoder-decoder networks, and L1-loss and perceptual loss are used for training. As far as we know, this is the first method that successfully extract semi-transparent motion-blurred object from a single image.
- 3) We enlarge the existing synthetic motion-blurred hand dataset to train the proposed model. Then we use the trained model to process real videos. Experimental

results show that the proposed method has promising performance.

II. METHOD

The architecture of the proposed method is shown in Fig. 2. As shown in this figure, the CNN network outputs two kinds of prediction results: alpha mattes and foreground images. The images of extracted hands are calculated by multiplying these two kinds of results.

A. Network Architecture

Encoder-decoder networks have demonstrated their great performance on many pixel-to-pixel prediction tasks, such as segmentation [17]–[19], depth prediction [20], matting [9], [21], and so on. Encoder-decoder networks usually employ pre-trained image recognition networks, such as VGG [22], ResNet [23], and Xception [16], as the backbone of encoder, and stack several upsampling blocks as decoder. Previous article [17] has shown that encoder-decoder networks based on Xception have better performance and faster speed. Therefore, we employ Xception based encoder-decoder networks for alpha matte prediction and foreground prediction. As shown in Fig. 2, the proposed two prediction networks share the same encoder, which contains 4 downsampling steps. In each step, the downsampling ratio equals to 0.5. Each of these two prediction networks has its own decoder, which contains 3 upsampling steps. The up-sampling ratios equal to 4, 2, 2 respectively. In each upsampling step, skip connection is used to recover spatial information. The decoders for alpha matte prediction and foreground prediction have similar structure. But the output of alpha prediction decoder has only one channel, while the output of foreground prediction decoder has three channels, which are red, green, and blue channels respectively.

B. Loss Function

For most of pixel-to-pixel prediction networks, including matting networks [9]–[11], [13], [14], pixel-wise losses, such as pixel-wise L1-loss and L2-loss, are generally used. However, pixel-wise losses ignore the correlation among pixels. A

solution is employing Conditional Random Fields (CRF) [24]–[26]. However CRF runs slowly. Another solution is employing perceptual loss. Perceptual loss has been successfully used for style transfer and super-resolution [15], [27]. It calculates the differences between high-level features extracted from predicted images and groundtruth images, and minimizes these differences by backpropagation. High-level features can be extracted by pre-trained convolutional networks.

In this letter, the overall loss is $L^o = L^\alpha + L^f$, where L^α denotes alpha prediction loss, L^f denotes foreground prediction loss. Both L^α and L^f are made up by combining L1 losses and perceptual losses.

1) *Alpha Prediction Loss*: The alpha prediction loss contains three parts: alpha absolute loss l_{ab}^α , alpha compositional loss l_c^α , and alpha perceptual loss l_p^α . l_{ab}^α is the L1 loss between predicted alpha mattes and groundtruth alpha mattes. l_c^α is the L1 loss between predicted compositional images and groundtruth compositional images. Compositional images are generated by $I_c = \alpha_* \cdot F + (1 - \alpha_*) \cdot B$, where F and B are given foreground images and background images, α_* denotes predicted alpha mattes or groundtruth alpha mattes.

The alpha perceptual loss l_p^α calculates the L2 loss between high-level features extracted from predicted alpha mattes and groundtruth alpha mattes. VGG-16 [22], which is pre-trained for image recognition and contains 5 convolutional blocks, is used as the feature extractor. All of the 5 level feature maps extracted by these 5 convolutional blocks are used to calculate l_p^α . It should be mentioned that VGG-16 is just employed as a feature extractor, which is only used during training.

The overall alpha prediction loss is $L^\alpha = \lambda_{ab}^\alpha l_{ab}^\alpha + \lambda_c^\alpha l_c^\alpha + \lambda_p^\alpha l_p^\alpha$, where λ_{ab}^α , λ_c^α , and λ_p^α are the loss weights. In our experiments, λ_{ab}^α and λ_c^α are set to 0.5. λ_p^α is set to 0.001.

2) *Foreground Prediction Loss*: The foreground prediction loss contains two parts: foreground absolute loss l_{ab}^f and foreground perceptual loss l_p^f . l_{ab}^f is the L1 loss between predicted foreground and groundtruth foreground. l_p^f is the L2 loss between high-level features extracted from predicted foreground and groundtruth foreground. Similar to l_p^α , l_p^f also uses VGG-16 as the feature extractor and all of the 5 levels of feature maps are employed. The overall foreground prediction loss is $L^f = \lambda_{ab}^f l_{ab}^f + \lambda_p^f l_p^f$, where λ_{ab}^f and λ_p^f are loss weights. In our experiment, they are set to 1 and 0.001 respectively.

III. EXPERIMENT

In motion-blurred object extraction task, it is very difficult to obtain groundtruth alpha mattes and groundtruth foreground images for real dataset. This is because it is almost impossible for human to assign accurate alpha values and foreground colors to image pixels in semi-transparent blurred areas. However, a large amount of motion-blurred hand images and their groundtruth are needed to train our models. To solve this problem, we employ the synthetic motion-blurred hand datasets provided by [6]. To increase the diversity of skin colors, we also enlarge these datasets. In our experiment, the training, validation, and testing synthetic datasets contain 30279, 8283, and 10140 cases respectively. These datasets contain synthetic motion-blurred hand images, groundtruth alpha mattes, and groundtruth foreground images. All of our models are trained on the synthetic

TABLE I
EVALUATION SCORES ON SYNTHETIC TESTING DATASET. SAD IS SHORT FOR SUM OF ABSOLUTE DIFFERENCES. MSE IS SHORT FOR MEAN SQUARED ERROR. PL IS SHORT FOR PERCEPTUAL LOSS. SE IS SHORT FOR SHARING ENCODER

Alpha prediction				
Methods	SE	PL	SAD($\times 10^3$)	MSE($\times 10^{-3}$)
Model 1	no	no	2.57	1.29
Model 2	yes	no	2.47	1.20
Model 3	no	yes	2.08	0.73
Model 4	yes	yes	2.02	0.71
Foreground prediction				
Methods	SE	PL	SAD($\times 10^3$)	MSE($\times 10^{-3}$)
Model 1	no	no	8.84	2.30
Model 2	yes	no	8.73	2.06
Model 3	no	yes	8.08	1.71
Model 4	yes	yes	7.94	1.58

training dataset. The synthetic validation dataset is used to monitor training processes. The synthetic testing dataset is used to evaluate the performances of different models. According to the evaluation results, the best performing model is chosen to process real videos. Our experiments are implemented under tensorflow with one Nvidia RTX 2080ti GPU and one Intel Core i7 9700 k CPU.

A. Evaluation on Synthetic Testing Dataset

In this letter, four different models are trained. They are called as Model 1, Model 2, Model 3, and Model 4. As shown in Table I, these four models differ in whether sharing encoder (SE) or whether employing perceptual loss (PL). It is worth to mention that the alpha matte prediction network of Model 1 is the same matting network proposed in [6]. Table I also shows the evaluation scores on the synthetic testing dataset. As shown in this table, Model 4 has the best performance. Therefore, Model 4 is chosen to process real videos in next subsection. In our experiment, during training Model 4, the learning rate, momentum, and weight decay were set to 3.5×10^{-3} , 0.9, and 4×10^{-5} respectively.

As shown in Table I, sharing encoder can slightly improve the prediction performance, while adding perceptual loss can obviously improve the prediction performance. In order to show the effectiveness of perceptual loss qualitatively, several prediction examples of Model 2 and Model 4 are shown in Fig. 3. Fig. 3 shows that the results, which are predicted by the model trained with perceptual loss, have obvious more accurate and reasonable shapes and textures.

B. Experiments on Real Videos

As described in last subsection, the best-performing model, Model 4, is chosen to predict the alpha mattes and foreground images of motion-blurred hands in real videos. The extracted hand images are calculated by multiplying the predicted alpha mattes and foreground images. Because it is very difficult to obtain the groundtruth of real datasets, in this subsection, we just show and compare the motion-blurred hand extraction performances qualitatively.

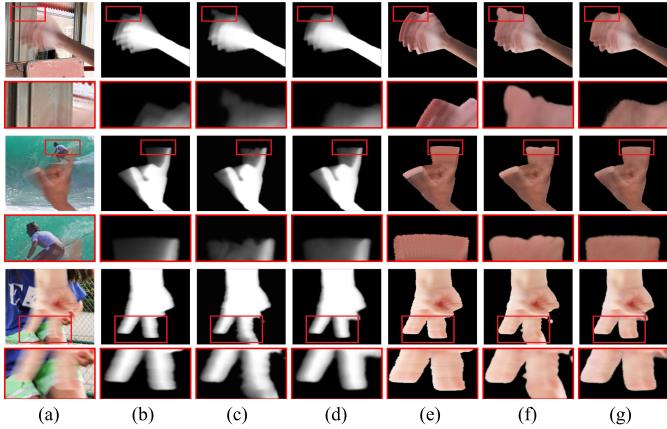


Fig. 3. Comparison of results predicted by the networks trained without or with perceptual loss: (a) original images; (b) groundtruth alpha mattes; (c) alpha mattes predicted by Model 2; (d) alpha mattes predicted by Model 4; (e) groundtruth foreground images; (f) foreground images predicted by Model 2; (g) foreground images predicted by Model 4. Model 2 is trained without perceptual loss. Model 4 is trained with perceptual loss.

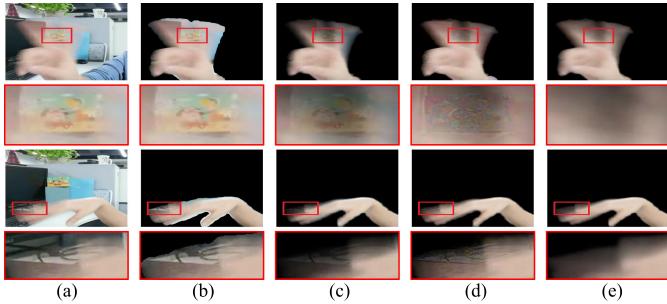


Fig. 4. Comparison of hand extraction performance with different methods. (a) Original RGB frames. (b) Hand images extracted by hand segmentation method [2]. The segmentation network has been finetuned on our dataset. (c) Hand images extracted by multiplying the predicted alpha mattes and original frames. This strategy is popular used in human matting methods [12], [13]. (d) Hand images extracted by subtracting background components from original frames [6]. (e) Hand images extracted by the method proposed in this letter.

1) Comparisons With Other Methods: The proposed motion-blurred hand extraction method is compared with three methods. For convenience, these three methods are called as Method 1, Method 2, and Method 3 respectively. Method 1 is a hand segmentation method [2], which predicts binary hand masks and extracts hands by multiplying the predicted masks with original frames. Method 2 extracts motion-blurred hands by multiplying the predicted alpha mattes and original frames. This strategy is popularly used in human matting methods [12], [13]. Method 3 extracts motion-blurred hands by subtracting background components from original frames [6].

As shown in Fig. 4, the hand images extracted by Method 1 and Method 2 contain obvious background information. This is because these two methods extract motion-blurred hands by multiplying the predicted binary masks or alpha mattes with original frames. Fig. 4 also shows that the hand images extracted by Method 3 contain obvious color distortions. Method 3 extracts motion-blurred hands by $I - (1 - \alpha) \cdot B$, as shown in Fig. 1(a). If the gap between the predicted alpha matte and groundtruth alpha matte is α_{gap} , then the gap between the

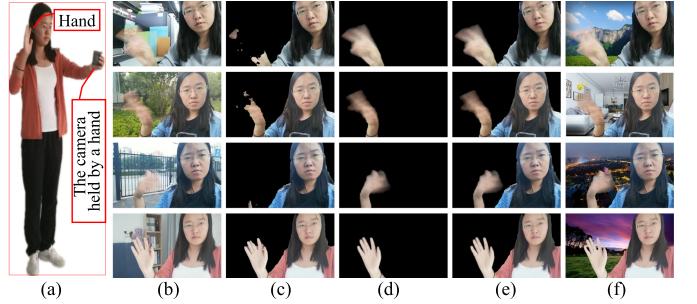


Fig. 5. Some examples of human soft segmentation: (a) a picture shows how the videos are captured; (b) original frames; (c) human segmentation results predicted by state-of-the-art image segmentation method Deeplab v3+ [17]; (d) the motion-blurred hand images extracted by the proposed method; (e) the modified human soft segmentation results; (f) frames with new background.

extracted motion-blurred hand and groundtruth motion-blurred hand is $\alpha_{gap} \cdot B$. In most of cases, $\alpha_{gap} \neq 0$. Therefore, if background B is colorful and contains complex textures, $\alpha_{gap} \cdot B$ will cause color distortions in semi-transparent blurred areas. In contrast, the proposed method extracts motion-blurred hands by multiplying the predicted α and predicted F , whose smoothness is easily to be guaranteed. Thus, for the proposed method, the extracted motion-blurred hands can be smooth and natural. In addition, Method 3 needs background images, which are not convenient to obtain, while our method doesn't need any additional information. In summary, our method has the best performance and can be applied more conveniently.

2) Modifying Human Segmentation Results: The proposed motion-blurred hand extraction method is useful in practical applications. It can be used to modify human segmentation results. The method in [6] did the same work. But it needed background images. Therefore, in [6], videos were captured by static cameras in order to obtain background images from the nearby frames. In contrast, the method proposed in this letter doesn't need background images and can be used in videos captured by moving cameras. The four examples shown in Fig. 5 are captured at different places by a hand-held camera. As shown in Fig. 5(a), we hold camera by one hand and make hand gestures by the other hand. These examples demonstrate that state-of-the-art segmentation method can't deal with motion-blurred hands, and our motion-blurred hand extraction method successfully modifies human segmentation results and has good performance to change background.

IV. CONCLUSION AND FUTURE WORK

In this letter, a novel method which can automatically extract the semi-transparent motion-blurred hand from a single image is proposed. It neither needs user interactions, nor needs additional information. This novel method contains two main parts: alpha matte prediction and foreground prediction. The images of extracted hands are calculated by multiplying the predicted alpha mattes and foreground images.

Further experiments on general motion-blurred objects show that the proposed method also works well to process other objects which just have one single color. In the future, we will extend our method to extract complete motion-blurred objects, such as motion-blurred human.

REFERENCES

- [1] S. Bambach, S. Lee, D. J. Crandall, and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 1949–1957. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7410583>
- [2] A. Urooj and A. Borji, "Analysis of hand segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 4710–4719. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8578593>
- [3] H. T. Lin, Y.-W. Tai, and M. S. Brown, "Motion regularization for matting motion blurred objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2329–2336, Nov. 2011, doi: [10.1109/TPAMI.2011.93](https://doi.org/10.1109/TPAMI.2011.93).
- [4] R. Kohler, M. Hirsch, B. Scholkopf, and S. Harmeling, "Improving alpha matting and motion blurred foreground estimation," in *Proc. IEEE Int. Conf. Image Process.*, Melbourne, VIC, Australia, 2013, pp. 3446–3450. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6738711>
- [5] H. Myeong, S. Lin, and K. M. Lee, "Alpha matting of motion-blurred objects in bracket sequence images," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, 2014, pp. 125–139. [Online]. Available: https://link.springer.com/content/pdf/10.1007%2F978-3-319-10578-9_9.pdf
- [6] X. Zhao and Y. Wu, "Automatic motion-blurred hand matting for human soft segmentation in videos," in *Proc. IEEE Int. Conf. Image Process.*, Taipei, Taiwan, 2019, pp. 1450–1454. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8803053>
- [7] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 228–242, Feb. 2008, doi: [10.1109/TPAMI.2007.1177](https://doi.org/10.1109/TPAMI.2007.1177).
- [8] Q. Chen, D. Li, and C.-K. Tang, "KNN matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2175–2188, Sep. 2013, doi: [10.1109/TPAMI.2013.18](https://doi.org/10.1109/TPAMI.2013.18).
- [9] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 311–320. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8099524>
- [10] D. Cho, Y.-W. Tai, and I. S. Kweon, "Deep convolutional neural network for natural image matting using initial alpha mattes," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1054–1067, March 2019, doi: [10.1109/TIP.2018.2872925](https://doi.org/10.1109/TIP.2018.2872925).
- [11] J. Tang, Y. Aksoy, C. Oztireli, M. Gross, and T. O. Aydin, "Learning-based sampling for natural image matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 3055–3063. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/papers/Tang_Learning-Based_Sampling_for_Natural_Image_Matting_CVPR_2019_paper.pdf
- [12] X. Shen, X. Tao, H. Gao, C. Zhou, and J. Jia, "Deep automatic portrait matting," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 92–107. [Online]. Available: https://link.springer.com/content/pdf/10.1007%2F978-3-319-46448-0_6.pdf
- [13] Q. Chen, T. Ge, Y. Xu, Z. Zhang, X. Yang, and K. Gai, "Semantic human matting," in *Proc. ACM Multimedia Conf. Multimedia Conf.*, Seoul, South Korea, 2018, pp. 618–626. [Online]. Available: <http://delivery.acm.org/10.1145/3250000/3240610/p618-chen.pdf?ip=159.226.182.77&id=3240610>
- [14] Y. Zhang *et al.*, "A late fusion CNN for digital matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 7469–7478. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/papers/Zhang_A_Late_Fusion_CNN_for_Digital_Matting_CVPR_2019_paper.pdf
- [15] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 694–711. [Online]. Available: https://link.springer.com/content/pdf/10.1007%2F978-3-319-46475-6_43.pdf
- [16] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 1800–1807. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8099678>
- [17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 833–851. [Online]. Available: https://link.springer.com/content/pdf/10.1007%2F978-3-030-01234-2_49.pdf
- [18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, Munich, Germany, 2015, pp. 234–241. [Online]. Available: https://link.springer.com/content/pdf/10.1007%2F978-3-319-24574-4_28.pdf
- [19] B. Vijay, K. Alex, and C. Roberto, "SegNet: A deep convolutional encoder-decoder architecture for scene segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 1, 2017, doi: [10.1109/TPAMI.2016.2644615](https://doi.org/10.1109/TPAMI.2016.2644615).
- [20] L. He, G. Wang, and Z. Hu, "Learning depth from single images with deep neural network embedding focal length," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4676–4689, Sep. 2018, doi: [10.1109/TIP.2018.2832296](https://doi.org/10.1109/TIP.2018.2832296).
- [21] G. Chen, K. Han, and K.-Y. K. Wong, "TOM-Net: Learning transparent object matting from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 9233–9241. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8579060>
- [22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representation*, 2015. [Online]. Available: <https://iclr.cc/archive/www/doku.php%3Fid=iclr2015:accepted-main.html>
- [23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7780459>
- [24] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, Granada, Spain, 2011, pp. 109–117. [Online]. Available: <http://papers.nips.cc/paper/4296-efficient-inference-in-fully-connected-crf-with-gaussian-edge-potentials.pdf>
- [25] S. Zheng *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 1949–1957. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7410536>
- [26] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: [10.1109/TPAMI.2017.2699184](https://doi.org/10.1109/TPAMI.2017.2699184).
- [27] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 4681–4690. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8099502>