Real-time communication with Virtual Background

Keyuan Zhang

Abstract

This paper will demonstrate a model to replace the background in real-time video communication by tackling portrait matting on mobile devices. The proposed model solves issues of previous models and achieves real-time inference speed while maintaining high visual performance. The proposed feature can attain over 25 FPS on 720P on iPhone11 in real-time video communication.

1 Introduction

Nowadays, there is an increasing need to use virtual backgrounds in real-time video communication. Replacing background in a video is a widely used technique that enables movie director or video creators to create more flexible content. Traditionally, replacing background is a two-fold process: first shots against a green screen then replaces the background while editing. The traditional method is time-consuming and can be employed in limited shooting environments, making real-time video communication impossible.

The key to applying virtual background in real-time video is to apply semantic human segmentation or matting first. With the success of convolution neural networks (CNN) in computer vision tasks, there has been an increasing number of works in this area [5, 16]. There are some approaches to matting specific object as foreground [3, 10, 15, 9], for example, portrait matting. Portrait matting is useful, especially in real-time video communication. Different from segmentation, it outputs more refined edges, which is more suitable for the application scene.

In this project, I propose a compact network for portrait matting, which can run on mobile devices in real-time. The proposed network is U-Net [7]like architecture with skip connection between encoder and decoder. U-Net is a well-known network used in biomedical segmentation; the skip connections restore low-level features during up-sampling, reducing the loss in down-sampling. Inspired by [12, 8], I have replaced traditional convolution with depthwise convolution, and applied linear bottleneck to reduce model size for real-time efficiency while maintaining its high performance.

The proposed virtual background feature is a part of the video customer service in several Chinese banks. Due to the confidentiality agreement with the company, I would not publish the entirety of related code or model file, but the overall performance is shown in the web demo

2 Related Work

2.1 Encoder-Decoder

Encoder-Decoder architecture is commonly applied in image segmentation task [7, 6, 2], the encoder is used to extract image features at a low level through sequential convolution and down-sampling. On the other hand, the decoder up-sample the feature map into the original image size while keeping the high-level information.

2.2 Depthwise Convolution

Depthwise convolution is a powerful operation used to reduce the computation cost while maintaining similar performance. This operation has been widely adopted in several works [4, 14, 13]. In this work, traditional convolution is been replaced by depthwise convolution to reduce computation cost.

2.3 Linear Bottleneck

Linear bottleneck [8] combines with depthwise convolution reduce the model size while maintaining high performance. According to [8], this module is particularly suitable for mobile design due to it reduces the memory footprint needed during inference. This module is used extensively in the encoder module to achieve real-time high-performance goals.

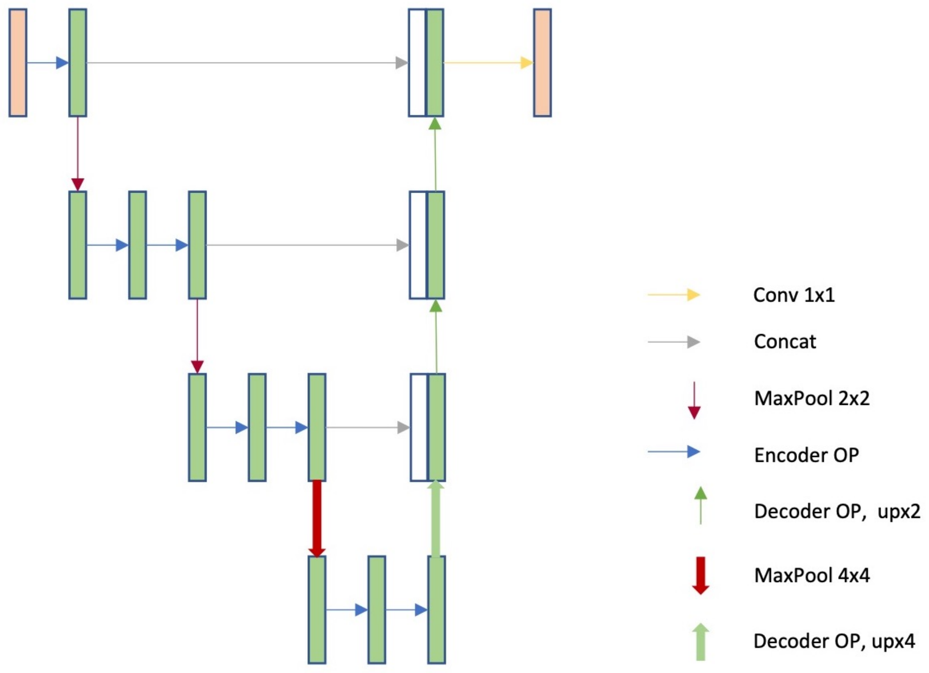

Figure 1: Model Architecture

2.4 Skip Connection

Skip connection [7] restore low-level features during up-sampling, which enables low-level features to restore in feature map, therefore enhance the accuracy. This connection has been used between encoder and decoder to ensure the low-level information has been stored during up-sampling.

3 Methods

3.1 Model Architecture

The overall model architecture follows a standard encoder-decoder structure, which is widely used in semantic segmentation [7, 6, 2]. Encoder reduces the size of inputs by summarizing the low-level information into high-level semantic information, which stores in a feature map. Decoder, on the other hand, recovers the spatial information and restores the original image size through upsampling. The model architecture is shown in Figure 1.

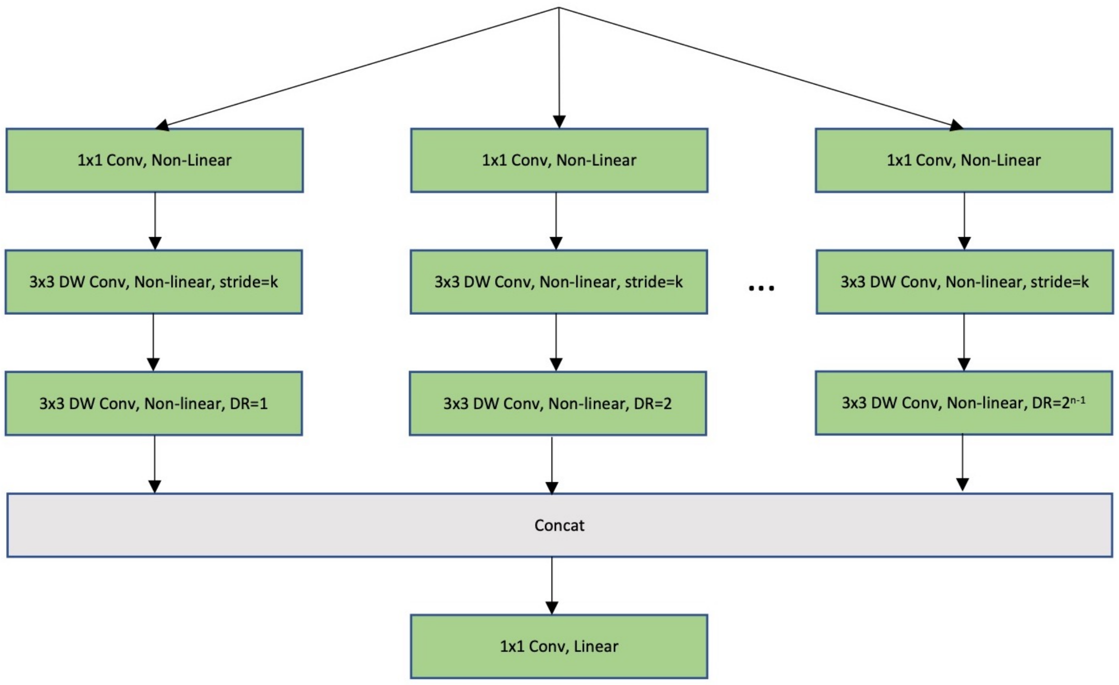
1x1 Conv, Non-Linear
1x1 Conv, Non-Linear
1x1 Conv, Non-Linear
3x3 DW Conv, Non-linear, stride=k
3x3 DW Conv, Non-linear, stride=k
...
3x3 DW Conv, Non-linear, stride=k
3x3 DW Conv, Non-linear, DR=1
3x3 DW Conv, Non-linear, DR=2
3x3 DW Conv, Non-linear, DR=2^{n-1}
Concat
1x1 Conv, Linear

Figure 2: Encoder Block

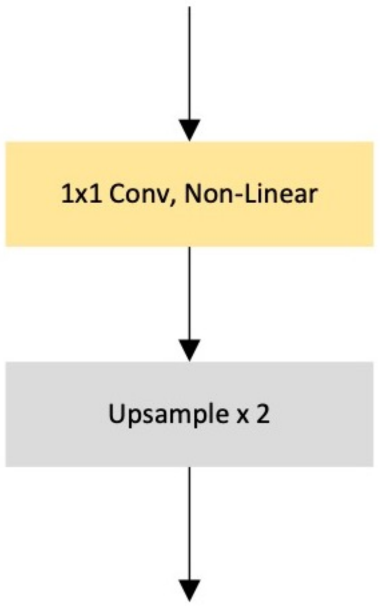

Figure 3: Decoder Block

3.1.1 Encoder

Encoder block follows a multi-branch dilated depthwise convolution with a linear bottleneck. The dilation rates are different for all branches to capture multi spatial information. The outputs of different branches are concatenated to a tensor, which contains multi-level information. After concatenation, the linear bottleneck is imposed where the output of intermediate layers is thinner than the input. The linear bottleneck decomposes traditional convolution, which connects two encoder blocks into a cheaper one by reducing channels. The encoder block is shown in Figure 2

3.1.2 Decoder

The decoder block performs upsampling to restore the original resolution. In order to restore the low-level information from the encoder, a skip connection is employed here to connect the encoder and decoder. The decoder block is illustrated in Figure 3

3.2 Loss Function

The loss function combines three different loss functions, which monitor the mean absolute difference (MAD) between the ground truth mask and the predicted mask, guide the model to diverge faster(not sure yet), and capture the fine-grained details in the edges.

The pixel-wise loss is frequently used in matting tasks. The pixel-wise loss L_1 measures the mean difference between ground truth mask and predicted one.

$$L_1 = \frac{1}{N} \sum_{i=1}^{N} |p_i - g_i^{gt}|$$

N denotes the total number of pixels, g is the ground truth, and p is vectorized output where each pixel value is indexed by subscript i.

The second loss is the softmax cross entropy:

$$L_2 = -\frac{1}{N} \sum_{i=1}^{N} (glogp + (1-g)log(1-p))$$

The last loss is the gradient loss [9]. Different from segmentation, matting requires finer edges, which needs additional loss function to monitor edge training. The gradient loss is shown as follows:

$$
\begin{aligned}
L_3 &= \frac{1}{N}\sum_{i=1}^{N} |G(A)_i - G(A^{gt})_i| \\
&= \frac{1}{2N}\sum_{i=1}^{N} \left(|(S * A - S * A^{gt})_i| + |(S^T * A - S^T * A^{gt})_i|\right)
\end{aligned}
$$

Where S is Sober-like filter S:

$$
S = \begin{bmatrix} -\frac{1}{8} & 0 & \frac{1}{8} \\ -\frac{2}{8} & 0 & \frac{2}{8} \\ -\frac{1}{8} & 0 & \frac{1}{8} \end{bmatrix}
$$

$G(A) = [S * A, S^T * A]$, where * is a convolution. $G(A)$ represents a two-channel output that contains gradient information along x-axis and y-axis of predicted mask, and $G(A^{gt})$ denotes the same information of ground truth mask.

The whole loss function used in this work is shown in below:

$$
L = L_1 + L_2 + L_3
$$

4 Evaluation

The performance metrics proposed in this work are mean absolute difference (MAD) and gradient error connectivity (Gradient) [17]. I used MAD to measure the pixel-wise accuracy of prediction, and Gradient to evaluate the edge fining performance. The MAD and Gradient error are defined as:

$$
\frac{1}{N}\sum_{i} |p_i - g_i|
$$

$$
\frac{1}{N}\sum_{i} \|\nabla p_i - \nabla g_i\|
$$

The training dataset comes from two different datasets, the first one is provided by Shen ty al.[11], which consists 2000 images of 600×800 resolution, and the second one obtains from [1], which consists 34427 images of the same resolution.

I split the dataset into training and testing with a ratio of $17 : 3$. To make the model more robust, I augmented images by scaling, rotation, and left-right flip. First, the image is descaled into input size 256×256 with a random scaling factor selected

Method	Gradient	MAD
Proposed Model	1.99	4.78
MDv3	2.06	5.49

Table 1: Evaluation result of proposed model and Mobile DeepLabv3

from 1 to 1.15. Then the rotation by $[-30^{\circ}, 30^{\circ}]$ is applied with a probability of 0.5, additional left-right flip is applied with the same probability. Finally, cropping is applied to ensure the size of the image match the input size of the model.

The training process consists two steps, I first trained the larger dataset around 500 thousand iterations to approach convergence, then fine tuning with the first dataset due to its high annotation quality. I trained the model with a batch size of 32 and a fixed learning rate of 10^{-4}.

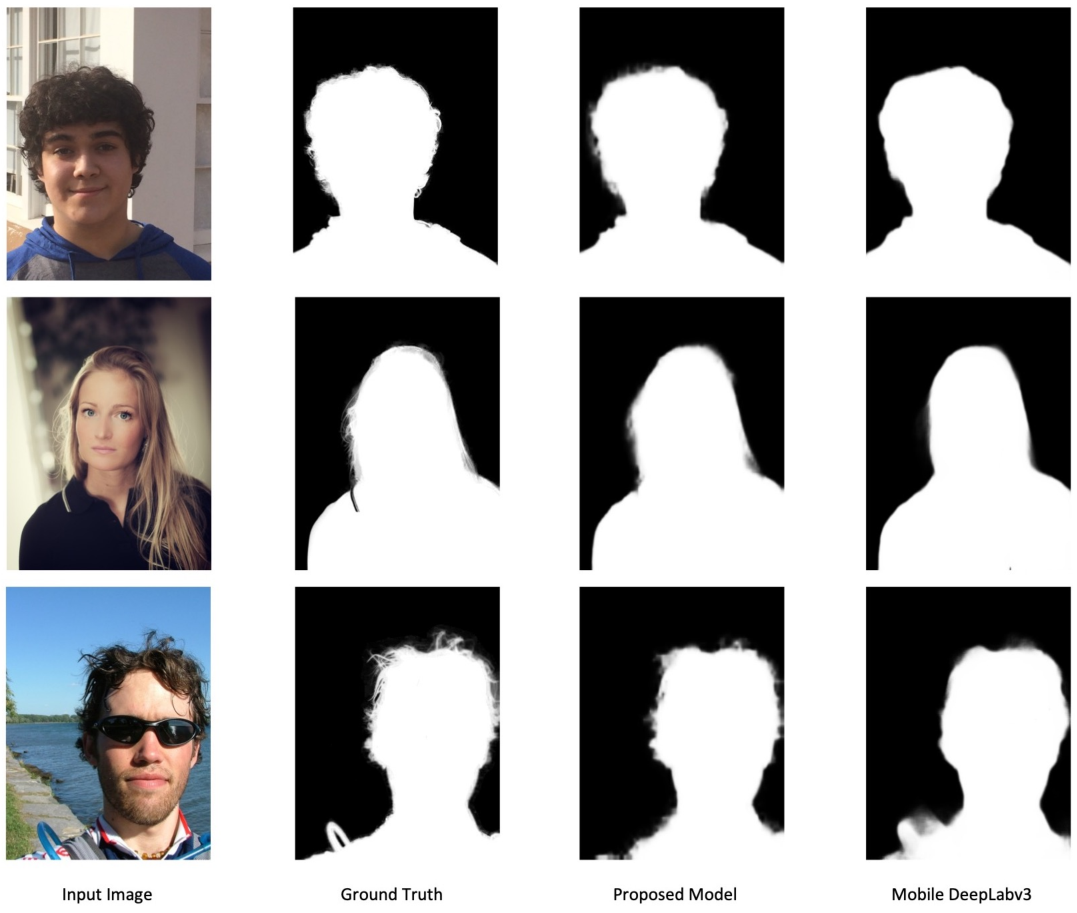

Figure 4: Visual performance of different models

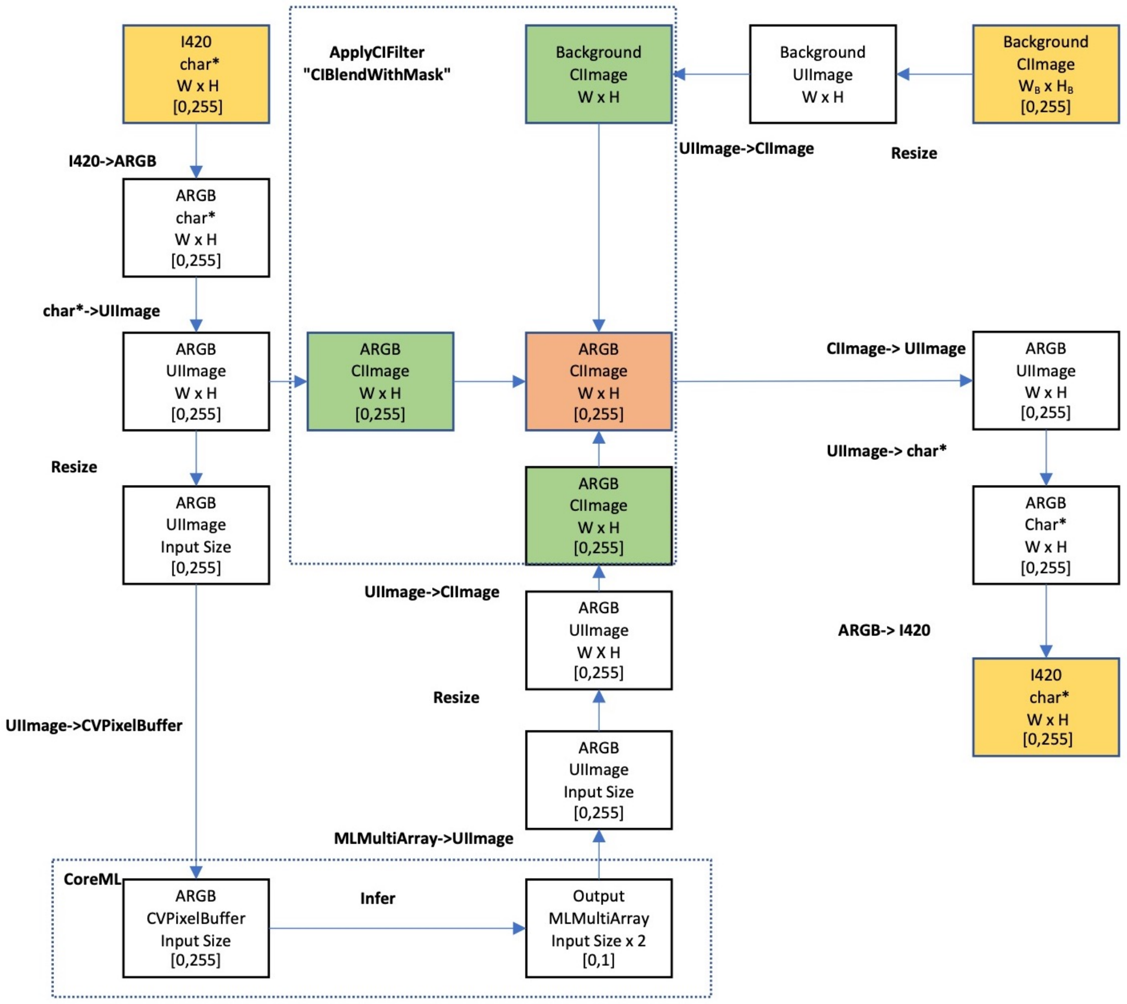

Figure 5: Application Flow Chart

5 Application

The final application platform is iOS for commercial purpose; with a brief review of deep learning inference architecture, I found three mainstream mobile inference architectures are promising: TensorFlow Lite, CoreML, and Caffe2.

Comparing with the other two architectures, CoreML is friendly for iOS devices; it seamlessly takes advantage of CPU and GPU. For A11 and newer bionic chip, CoreML can take benefit from bionic silicon to speed up inference efficiency. Specifically, I implemented CoreML as the mobile inference architecture and quantized the model by CoreML to further improve inference efficiency.

Mobile Device	360P	720P	1080P
iPhone7	19FPS	15FPS	8FPS
iPhoneX	20FPS	16FPS	10FPS
iPhone11	29FPS	25FPS	19FPS

Table 2: Running Speed in Real-time video communication. The speed performance includes resizing, inference, replacing background and codec

During the demo test, there are some losses in visual performance. The first one is the loss in the rotation because the training dataset did not contain images with camera rotation; when the user rotates the camera, the visual loss is heavy. This can be solved by a random rotate image $[-90^{\circ}, 90^{\circ}]$ during training.

The second one is caused by the image ratio of width and height. During the demo test, there are some losses in visual performance, because the ratio of input blob is $1:1$, and the main ratio of resolution in video communication is $16:9$. To overcome this, I modified the ratio of input blob into $16:9$. For example, with the input image size of 640×360, the input has to be resize to 256×144 to feed into network.

6 Conclusion

I have proposed an efficient model to perform real-time virtual background in video communication on mobile devices in this work. The proposed model can achieve over 30 FPS inference speed on iPhone 11. I have also shown the pipeline of implementing the model to the real-time application, including a literature review of related works, how to take advantage of a open-source dataset, design a compact model architecture, and necessary modification on model deployment. The proposed model enhances the user experience in video chatting, can be widely used in image and video editing, and could be used in human-focused super-resolution video chatting in the future.

References

[1] Aisgement. https://github.com/aisegmentcn/matting_human_datasets. Accessed: 2019-09-16. 4

[2] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2.1, 3.1

[3] Q. Chen, T. Ge, Y. Xu, Z. Zhang, X. Yang, and K. Gai. Semantic human matting. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 618–626, 2018. 1

[4] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 2.2

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[6] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2.1, 3.1

[7] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 2.1, 2.4, 3.1

[8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 1, 2.3

[9] S. Seo, S. Choi, M. Kersner, B. Shin, H. Yoon, H. Byun, and S. Ha. Towards real-time automatic portrait matting on mobile devices. *arXiv preprint arXiv:1904.03816*, 2019. 1, 3.2

[10] X. Shen, X. Tao, H. Gao, C. Zhou, and J. Jia. Deep automatic portrait matting. In *European conference on computer vision*, pages 92–107. Springer, 2016. 1

[11] X. Shen, X. Tao, H. Gao, C. Zhou, and J. Jia. Deep automatic portrait matting. In *European conference on computer vision*, pages 92–107. Springer, 2016. 4

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2.2

[14] M. Wang, B. Liu, and H. Foroosh. Design of efficient convolutional layers using single intra-channel convolution, topological subdivisioning and spatial" bottleneck" structure. *arXiv preprint arXiv:1608.04337*, 2016. 2.2

[15] N. Xu, B. Price, S. Cohen, and T. Huang. Deep image matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2970–2979, 2017. 1

[16] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1

[17] B. Zhu, Y. Chen, J. Wang, S. Liu, B. Zhang, and M. Tang. Fast deep matting for portrait animation on mobile phone. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 297–305, 2017. 4