Context-Aware Image Matting for Simultaneous Foreground and Alpha Estimation

Qiqi Hou
Portland State University
qiqi2@pdx.edu

Feng Liu
Portland State University
fliu@cs.pdx.edu

Abstract

Natural image matting is an important problem in computer vision and graphics. It is an ill-posed problem when only an input image is available without any external information. While the recent deep learning approaches have shown promising results, they only estimate the alpha matte. This paper presents a context-aware natural image matting method for simultaneous foreground and alpha matte estimation. Our method employs two encoder networks to extract essential information for matting. Particularly, we use a matting encoder to learn local features and a context encoder to obtain more global context information. We concatenate the outputs from these two encoders and feed them into decoder networks to simultaneously estimate the foreground and alpha matte. To train this whole deep neural network, we employ both the standard Laplacian loss and the feature loss: the former helps to achieve high numerical performance while the latter leads to more perceptually plausible results. We also report several data augmentation strategies that greatly improve the network's generalization performance. Our qualitative and quantitative experiments show that our method enables high-quality matting for a single natural image.

1. Introduction

Natural image matting is the problem of estimating the foreground image and the corresponding alpha matte from an input image. It is a critical step of image composition, which is widely used in image and video production. Without any external information, matting is a seriously ill-posed problem. In practice, most existing matting methods take a trimap as input; however, matting is still underconstrained in the undefined area in the trimap.

Traditional methods solve the matting problem by inferring the alpha matte information in the undefined area from those in the defined areas [51]. For instance, the matte values in the undefined areas can be propagated from the known areas according to the spatial and appearance affinity between them [2, 6, 7, 20, 26, 27, 28, 46]. Alternatively, the undefined matte values can be computed by sampling the color or texture distribution of the known foreground and background areas and optimizing a carefully defined metric, such as the likelihood of the foreground, background, and alpha values [11, 18, 19, 49, 50]. While these methods provide promising results and some of them are incorporated into commercial tools, single natural image matting is still a challenging problem as these methods rely on the distinctive appearance of the foreground and background areas, such as their local or global color distribution.

Our research is inspired by the recent deep learning approaches to image matting. These deep matting approaches, such as [4, 33, 52] take an input image and the corresponding user-provided trimap as input and output an alpha map. They are shown robust for many challenging scenarios. These methods, however, only output the alpha map without the foreground or background image.

This paper presents a deep image matting method that simultaneously estimate the alpha map and the foreground image. Our method explores both local image and global context information for high-quality matting. This is inspired by the success of non-deep learning-based matting approaches that combines the global sampling and local propagation strategies [2, 6, 7, 20, 19]. Specifically, we designed a two-encoder-two-decoder fully convolutional neural network for context-aware simultaneous foreground image and alpha map estimation. The matting encoder learns to extract the local features while the context encoder learns more global features. We concatenate the features from these two encoders and feed them to an alpha decoder and a foreground decoder to estimate the alpha map and the corresponding foreground image simultaneously.

We explore a Laplacian loss and the feature loss to train our deep matting neural network. We found that the Laplacian loss enables our network to achieve the state-of-the-art numerical performance while the feature loss leads to more perceptually plausible matting results. We also found that some data augmentation methods are particularly helpful for our neural network to generalize to real-world images although our network is trained on a synthetic dataset pro-

Input image Trimap Our alpha map and composition results Results from Closed-form [27]

Figure 1. Real-world image matting. Our method is able to simultaneously estimate high-quality foreground images and alpha maps from real-world images although trained on a synthetic dataset. Our results keep final structures (the top example) while being free from the common color bleeding problem (the bottom example).

vided by Xu *et al*. [52].

To our best knowledge, this paper contributes the first deep matting method that enables simultaneous foreground and alpha estimation. Both our qualitative and quantitative experiments demonstrate that our method is able to generate state-of-the-art matting results on challenging real-world examples, as shown in Figure 1. We attribute the success of our method to 1) the integration of local visual features and global context information, 2) the combination of the Laplacian and feature loss, and 3) various effective data augmentation strategies that help generalizing our method to a wide variety of challenging real-world images.

2. Related Work

Image matting assumes that an image $\mathbf{I}$ is a linear composition of a foreground image $\mathbf{F}$ and a background image $\mathbf{B}$ according to an alpha map $\boldsymbol{\alpha}$ as follows [45].

$$\mathbf{I} = \boldsymbol{\alpha}\mathbf{F} + (1 - \boldsymbol{\alpha})\mathbf{B} \tag{1}$$

Given the input image $\mathbf{I}$, image matting aims to recover $\mathbf{F}$, $\mathbf{B}$ and $\boldsymbol{\alpha}$. Most of existing matting methods require a user-provided trimap that specifies known foreground and background areas, as well as an undefined area. In this way, matting is reduced to solving for the foreground, background, and alpha values in the undefined area. Given only the input image $\mathbf{I}$ and the trimap, matting is a seriously ill-posed problem. A rich literature exists for matting. These methods infer the matte information for the undefined area from the known foreground and background according to the trimap. They either propagate the matte information from the neighboring foreground or background areas to the unknown areas [2, 6, 7, 20, 26, 27, 28, 46], or more globally sample the appearance information of the known foreground and background and use them to optimize for the matting in the unknown area [11, 18, 19, 49, 50]. There are also methods that combine the local propagation strategy and the global sampling strategy to achieve more reliable results [2, 6, 7, 20, 19]. Wang and Cohen provided a good survey on these traditional image matting algorithms [51]. Our design of a double-encoder-double-decoder network to learn to estimate local and global context information is inspired by these hybrid methods.

Our work is most relevant to the recent deep learning approaches to image matting. Shen *et al*. trained a dedicated deep convolutional neural network for portrait matting [43]. Their method first employs a deep neural network to generate the trimap of a portrait image and then feeds it to an off-the-shelf matting method, namely the Closed-form Matting algorithm [27], to obtain the final matting result. Cho *et al*. developed a deep matting method that takes the matting results from the Closed-form Matting algorithm [27] and the KNN Matting algorithm [6] as input, and refine it using a deep neural network [8, 9]. Xu *et al*. developed a large-scale synthetic image matting dataset and used it to train a two-stage deep neural network for alpha matting. Their method produced high-quality matting results for both synthetic and real-world images [52]. Lutz *et al*. explores generative adversarial networks to achieve high-quality natural image matting [33]. In their recent work, Chen *et al*. addressed a difficult case of image matting, transparent object matting. By considering transparent object matting as a refractive flow estimation problem, they developed a two-stage neural network to estimate the refractive flow from only one input image for transparent object matting [4]. While these methods are able to estimate high-quality alpha maps, they do not generate the foreground component. Our work builds upon these deep learning methods and simultaneously estimate the foreground image and the alpha map, thus providing a

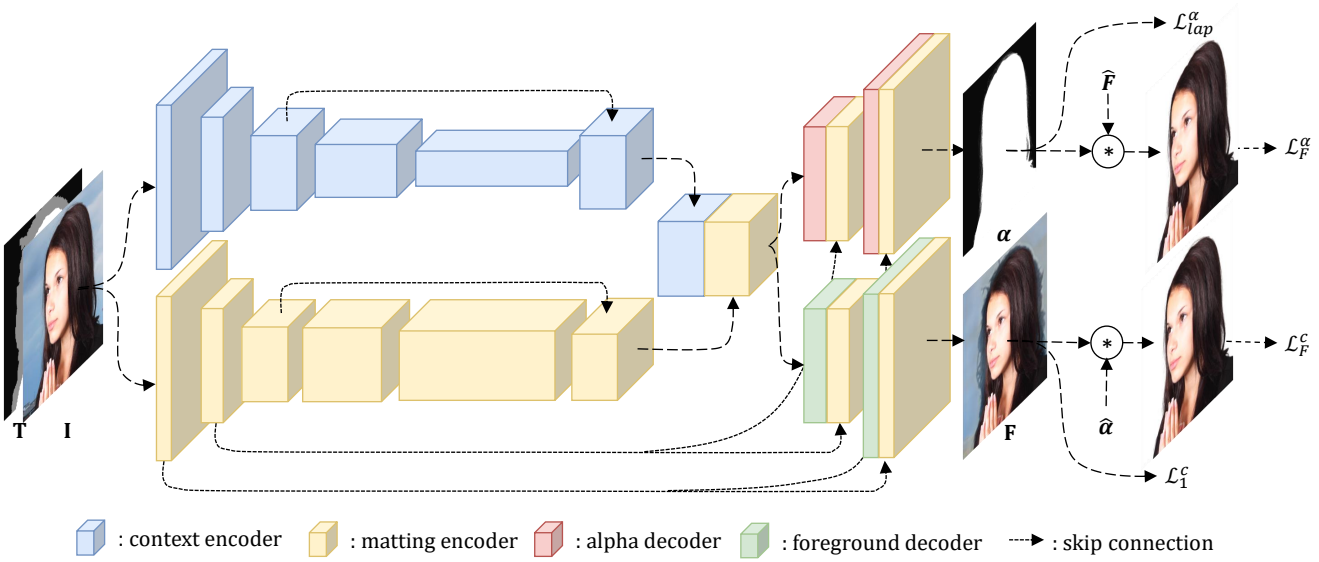


Figure 2. The architecture of our matting network. We design a two-encoder-two-decoder network. The matting encoder and the context encoder capture both visual features and more global context information. The features from these two encoders are concatenated and feed to the foreground and the alpha decoder to output the foreground image and the alpha map of the input image simultaneously.

complete solution to image matting. Our network learns to extract both local visual features and global context information to obtain high-quality image matting.

3. Context-Aware Image Matting

Our method takes an image $\mathbf{I}$ and a user-specified trimap $\mathbf{T}$ as input and aims to estimate the foreground $\mathbf{F}$ and the corresponding alpha map $\boldsymbol{\alpha}$, thus providing a full solution to matting. With the foreground and the alpha map, we can directly compute the background according to Equation 1.

We design a context-aware two-encoder-two-decoder deep neural network to simultaneously estimate the foreground and the alpha map, as shown in Figure 2. The outputs of the two encoders are concatenated and fed to the two decoder to generate the foreground and the alpha map, respectively. The two-encoder design of the network is inspired by the success of traditional matting algorithms that combine the local propagation and global sampling strategies for robust image matting [2, 6, 19, 20]. Specifically, the matting encoder is designed to learn to extract local features that are required to capture final image structures, such as hairs, while the context encoder learns to estimate more global context information that is helpful to disambiguate the foreground and background when they are similar to each other locally. Below we describe the encoders and decoders in more detail.

Matting encoder. We adopt the modified version of the Xception 65 architecture [10] from the deeplab v3+[5] and set the down-sampling factor as 4 by setting the `entroy flow`'s `block2` and `block3`'s stride as 1. This modification enables the `middle flow` to have a big spatial resolution. While traditional classification models [10, 21, 24, 41, 44] more aggressively compromise the spatial resolution to have a large valid receptive field, we use such a smaller down-sampling factor to retain sufficient spatial information that is important for the task of matting to capture fine image structures. Meanwhile, there is a trade-off between the computation/memory cost and spatial resolution. We empirically find that the down-sampling factor of 4 can get good matting results and cost a relatively small amount of computation and memory. We use skip connections to use features from the earlier layers as shown in Figure 2.

Context encoder. We also adopt the Xception 65 architecture [10] from [5]. Compared to the matting encoder, we use a much larger down-sampling factor of 16 to capture more global contextual information. We bilinearly upsample the final features by a factor of 4 so that the context features are of the same size as the local matting features from the matting encoder.

Alpha decoder and **foreground decoder** have the same network architecture. Specifically, we first bilinearly upsample the concatenated features from the encoders by a factor of 2 and then combine them with the intermediate features from the context encoder using a skip connection as shown in Figure 2. This is followed by two 3×3 convolutional layers with 64 channels. We repeat this process twice so that each decoder outputs the foreground image and the alpha map with the same size as the input image.

3.1. Loss functions

We compute the loss over both the alpha map and the foreground image. We explore a range of loss functions to

train our network. Below we describe them one by one.

We use a Laplacian loss [35] to measure the difference between the predicated alpha map $\boldsymbol{\alpha}$ and its ground truth $\hat{\boldsymbol{\alpha}}$.

$$\mathcal{L}_{lap}^{\alpha} = \sum_{i=1}^{5} 2^{i-1} \|L^i(\hat{\boldsymbol{\alpha}}) - L^i(\boldsymbol{\alpha})\|_1, \quad (2)$$

where $L^i(\boldsymbol{\alpha})$ indicates the i^{th} level of the Laplacian pyramid of the alpha map. This loss function measures the differences of two Laplacian pyramid representations and captures the local and global difference. We scale the contribution of a Laplacian level according to its spatial size.

We also use the feature loss to measure the perceptual quality of the alpha map. The feature loss, based on the differences between the high-level features extracted from a pre-trained convolutional neural network, has been shown effective in generating perceptually high-quality images in many image enhancement and synthesis tasks [14, 25, 35, 36, 40, 53, 55]. However, it is difficult to directly measure the perceptual quality of an alpha map. Our solution is to composite the ground-truth foreground image onto the black background using the alpha map and then measure the perceptual quality of the composition result as follows.

$$\mathcal{L}_{F}^{\alpha} = \sum_{layer} \|\phi_{layer}(\hat{\boldsymbol{\alpha}} * \hat{\mathbf{F}}) - \phi_{layer}(\boldsymbol{\alpha} * \hat{\mathbf{F}})\|_2^2, \quad (3)$$

where $\hat{\mathbf{F}}$ indicates the ground truth foreground and ϕ_{layer} indicates the features output by the $layer$ in a pre-trained VGG16 network [44]. Our method uses [`conv1_2`, `conv2_2`, `conv3_3`, `conv4_3`] to compute the features.

We follow the same setting to calculate the feature loss for the predicated foreground image. Here the feature loss $\mathcal{L}_F^c$ is computed on the composition result using the ground-truth alpha map with the foreground image as follows.

$$\mathcal{L}_{F}^{c} = \sum_{layer} \|\phi_{layer}(\hat{\boldsymbol{\alpha}} * \hat{\mathbf{F}}) - \phi_{layer}(\hat{\boldsymbol{\alpha}} * \mathbf{F})\|_2^2, \quad (4)$$

We also use the standard ℓ_1 loss for the predicted foreground $\mathbf{F}$. We only calculate the loss where the foreground is visible, in other words, the ground truth alpha matte is bigger than 0,

$$\mathcal{L}_{1}^{c} = \|\mathbb{1}(\hat{\boldsymbol{\alpha}} > 0) * (\hat{\mathbf{F}} - \mathbf{F})\|_1, \quad (5)$$

where $\mathbb{1}$ is an indicator function that takes 1 if the statement is true and 0 otherwise.

Finally, we apply the standard ℓ_2 regularization loss to all the convolutional layers. We will examine these loss functions in our experiments (Section 4).

3.2. Training

We initialize our neural network with pre-trained models from [5]. We use TensorFlow to train our neural network. Similar to [5], we use the "poly" learning rate policy

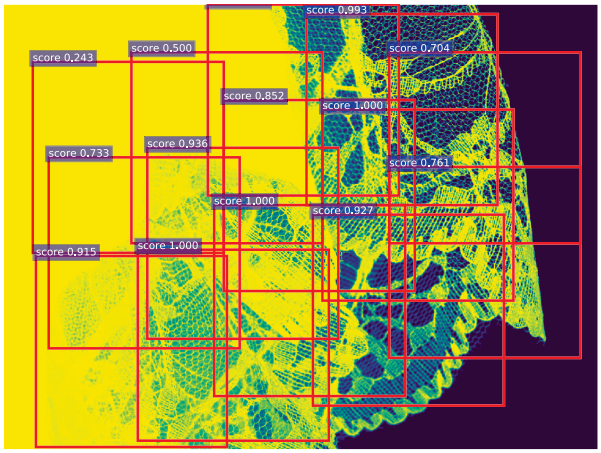
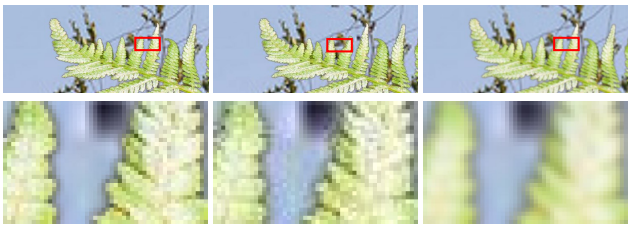


Figure 3. Image patch selection. The alpha map is illustrated using the color map, with yellow and blue indicating the foreground and background, respectively. The patches are selected to cover the unknown region but with relatively small overlaps among them.

to train our network, where $lr = lr_{init}(1 - \frac{iter}{max\_iter})^{power}$ with $lr_{init} = 7 \times 10^{-4}$ and $power = 0.9$. We use a mini-batch size of 6 and train the neural network for 1 million iterations for models (1-3) in Table 1. We fine-tune models (4-9) based on the pretrained model (3) with 10^5 iterations with $lr_{init} = 10^{-4}$.

Training dataset. We train our network using the matting dataset shared by Xu *et al*. [52]. This dataset contains 431 training images with the corresponding alpha maps and the foreground images. We create the training samples in a similar way to Xu *et al*. Specifically, we composite the foreground image onto a randomly selected background image from MS-COCO dataset [30]. We down-sample the foreground image gradually by a factor of 0.9 until the short side is 600 pixels. If the source image's short side is less than 600 pixels, we first scale it up to 780. In total, we generate 1957 scaled foreground image. Then we select image patches that contain unknown regions in the trimap. Specially, we slide windows of size 600×600 on the full image with a stride of 5 pixels to get a large amount of candidate windows and remove patches where less than 10% pixels are unknown. Furthermore, since many patches overlap with each other significantly, we employ non-maximum suppression(NMS) to remove overlapping patches. Specifically, we set the NMS threshold as 0.3 and only keep the top 30 image patches with the highest unknown pixel percentages in each image. Figure 3 shows an example of selected image patches. In total, we obtain 9,507 600×600 foreground image patches. Finally, we create training samples of size 225×225 by randomly cropping the composited image with the following data augmentation operators.

Data augmentation. Following Xu *et al*. [52], our training samples are obtained by compositing a foreground image and a background image using an alpha map. As reported in many papers [1, 13, 17, 22, 23, 31, 32, 37, 47], many sub-



(a) No augmentation (b) Re-JPEGing (c) Gaussian Blur

Figure 4. Data augmentation. In the composited image without any data augmentation (a), the foreground image contains some JPEG artifacts while the background is smooth, which produces a bias that will compromises the training of the network. Re-JPEGing introduces the artifacts to the foreground and the background to reduce the bias while Gaussian Blur does so by smoothing the high-frequency artifacts.

tle artifacts, such as misaligned JPEG blocks, compression quantization artifacts, and resampling artifacts, can sometimes affect their methods a lot despite that the images look plausible to the human eyes. Some splice detection methods [1, 22, 23, 32, 37, 47] even build their algorithms based on such an observation. Directly training the network on the composited images without special augmentation may suffer from a similar problem and thus compromises the generalization capability of the trained network.

Therefore, besides the resizing augmentation used in Xu *et al*. [52], we follow the post processing steps in the image splice detection methods [12, 23, 34] and use re-JPEGing and Gaussian blur to augment our training samples. These operators introduce subtle artifacts that are not visually noticeable but can make the network less bias to the small difference between the foreground and the background. As shown in Figure 4, the original background is smoother than the original foreground image. Therefore, it is possible that the network relies on this bias to differentiate the foreground from the background. Re-JPEGing and Gaussian blur can relieve this problem by introducing artifacts or remove these artifacts. For re-JPEGing, we keep 70% quality of the composited images. For Gaussian blur, we on-the-fly generate a Gaussian kernel with standard deviation in the range of [0, 3] and the kernel size in the range of [3, 5], and apply it to the composited image. We also randomly resize the composited image with a rate of between 0.5 and 1.

Besides, we also use some standard data augmentation operators. Specifically, we employ the gamma transforms to increase the color diversity. The gamma value is randomly selected from [0.2, 2]. We randomly flip the images horizontally. The trimap for each image is automatically generated by randomly dilating its corresponding ground truth alpha map in the range of [4, 25].

4. Experiments

We experiment with our methods on the synthetic Composition-1K dataset and a real-world matting image dataset, both of which are provided by Xu *et al*. [52]. As discussed in Section 3.2, our neural networks are all trained on the synthetic Composition-1K training set. We evaluate our models and compare to the state of the art methods on the Composition-1K testing set and the real-world matting image set. Specifically, the Composition-1K testing dataset contains 1000 composited images. They were generated by compositing 50 unique foreground images onto each of the 20 images from the PASCAL VOC 2012 dataset [15]. We used the code provided by Xu *et al*. [52] to generate these testing images. The real world image dataset contains 31 real world images pulled from the internet [52]. We conduct our user study on the real world images.

Since not all the methods produce both the foreground images and the alpha maps as the final matting results, we compare our methods to the state of the art on the alpha maps and the foreground images separately. Besides, we also report our user study and our ablation studies to more thoroughly evaluate our methods.

Table 1. Alpha map results on the Composition-1K testing set.

Methods	SAD	MSE(10^3)	Grad	Conn
Shared Matting[16]	128.9	91	126.5	135.3
Learning Based Matting [54]	113.9	48	91.6	122.2
Comprehensive Sampling [42]	143.8	71	102.2	142.7
Global Matting [19]	133.6	68	97.6	133.3
Closed-Form Matting [27]	168.1	91	126.9	167.9
KNN Matting [6]	175.4	103	124.1	176.4
DCNN Matting [8]	161.4	87	115.1	161.9
Three-layer Graph [29]	106.4	66	70.0	-
Deep Matting [52]	50.4	14	31.0	50.8
Information-flow Matting [2]	75.4	66	63.0	-
AlphaGan-Best[1] [33]	52.4	30	38.0	-
(1) ME + $\mathcal{L}_{deepmatting}$	49.1	13.4	26.7	49.8
(2) ME + $\mathcal{L}^{\alpha}_{lap}$	43.9	11.8	20.6	41.6
(3) ME + CE + $\mathcal{L}^{\alpha}_{lap}$	**35.8**	**8.2**	17.3	**33.2**
(4) ME + CE + $\mathcal{L}^{\alpha}_{lap}$ + $\mathcal{L}^{\alpha}_{F}$	38.8	9.0	19.0	36.0
(5) ME + CE + $\mathcal{L}^{\alpha}_{lap}$ + $\mathcal{L}^{\alpha}_{F}$ + DA	71.3	23.6	38.8	72.0
(6) ME + CE + $\mathcal{L}^{\alpha}_{lap}$ + $\mathcal{L}^{\alpha}_{F}$ + $\mathcal{L}^{c}_{1}$ + $\mathcal{L}^{c}_{F}$	38.0	8.8	**16.9**	35.4
(7) ME + CE + $\mathcal{L}^{\alpha}_{lap}$ + $\mathcal{L}^{\alpha}_{F}$ + $\mathcal{L}^{c}_{1}$ + $\mathcal{L}^{c}_{F}$ + DA	84.1	29.1	39.2	-
(8) ME + CE + $\mathcal{L}^{\alpha}_{lap}$ + $\mathcal{L}^{\alpha}_{F}$ + $\mathcal{L}^{c}_{1}$ + $\mathcal{L}^{c}_{F}$ + DA - ReJPEGing	55.1	15.5	24.6	54.7
(9) ME + CE + $\mathcal{L}^{\alpha}_{lap}$ + $\mathcal{L}^{\alpha}_{F}$ + $\mathcal{L}^{c}_{1}$ + $\mathcal{L}^{c}_{F}$ + DA - GaussianBlur	69.1	23.5	39.6	69.1

4.1. Evaluation on alpha maps

We compare our methods to both the state of the art non-deep learning methods, including Shared Matting [16], Learning Based Matting [54], Comprehensive Sampling [42], Global Matting [19], Closed-form Matting [27], KNN Matting [6], Three-layer Graph [29], Information-flow Matting [2], and recent deep learning matting approaches, including DCNN Matting [8], Deep Matting [52] and AlphaGan [33]. Table 1 reports the results on these methods as well as ours on the Composition-1K dataset. The results of the comparing methods are obtained

Table 2. The foreground result on the Composition-1k dataset.

Methods	SAD	MSE(10^3)
Global Matting [19]	220.39	36.29
Closed-Form Matting [27]	254.15	40.89
KNN Matting [6]	281.92	36.29
(6) ME + CE + $\mathcal{L}_{lap}^{\alpha} + \mathcal{L}_{F}^{\alpha} + \mathcal{L}_{1}^{c} + \mathcal{L}_{F}^{c}$	61.72	3.24
(7) ME + CE + $\mathcal{L}_{lap}^{\alpha} + \mathcal{L}_{F}^{\alpha} + \mathcal{L}_{1}^{c} + \mathcal{L}_{F}^{c}$ + DA	94.41	8.67
(8) ME + CE + $\mathcal{L}_{lap}^{\alpha} + \mathcal{L}_{F}^{\alpha} + \mathcal{L}_{1}^{c} + \mathcal{L}_{F}^{c}$ + DA - ReJPEGing	73.79	4.96
(9) ME + CE + $\mathcal{L}_{lap}^{\alpha} + \mathcal{L}_{F}^{\alpha} + \mathcal{L}_{1}^{c} + \mathcal{L}_{F}^{c}$ + DA - GaussianBlur	85.8	7.10

Table 3. Parameter numbers of our models and their performance on the Composition-1K dataset.

Methods	# of Parameters	SAD	MSE(10^3)	Grad	Conn
ME (model 2)	54.0 M	43.9	11.8	20.6	41.6
ME (deeper model 2)	117.0 M	43.7	11.0	21.2	42.6
ME + CE (model 3)	107.5 M	**35.8**	**8.2**	**17.3**	**33.2**

Table 4. Comparison of visual quality on the real-world dataset.

Methods	Mean score	Std
ME + CE + $\mathcal{L}_{lap}$	4.64	0.42
ME + CE + $\mathcal{L}_{lap} + \mathcal{L}_{F}$	4.69	0.40
ME + CE + $\mathcal{L}_{lap} + \mathcal{L}_{F}$ + DA	5.03	0.25

either from their papers or from the recent studies [33, 52].

To evaluate these methods, we use various metrics, including SAD, MSE, Gradient (Grad) and Connectivity (Conn) [39]. Note that the Conn metric fails on some results, which are denoted as "-". For the ablation analysis of our work, we reported our results on nine versions of our networks with different components. We use "ME", "CE", "DA" to indicate the matting encoder, the context encoder, and data augmentation, respectively.

As shown in Table 1, our two-encoder-two-decoder models (model (3-9)) generate matting results with significantly smaller errors than the state of the art methods. To understand what contributes to this improvement, we evaluated on a baseline method (model (2)) that removes the context encoder and found that this baseline model performs much worse according to all the four metrics. Therefore, the improvement can be mainly attribute to the use of our two encoders to capture both local visual features for fine structures and more global contextual information to disambiguate the locally similar foreground and background. Besides these numerical scores, our methods produce visually more plausible results as shown in Figure 6. For example, the last example has a strand of long hair. The results from existing methods either miss it entirely or the hair is broken into pieces while our methods better preserve it.

Number of parameters. We make model (2) deeper so that its number of parameters roughly match model (3). As shown in Table 3, while this deeper version of model (2) improves over the original one w.r.t SAD and MSE, it performs worse than our model (3) (ME + CE).

Sensitivity of trimap. Following the same process of the Deep Matting work [52], we examine our method's sensitivity to the trimap size by dilating the ground-truth to a range of sizes. As illustrated in Figure 5, our method is stable to the trimap sizes. Note, the scores of comparing methods were obtained from [52].

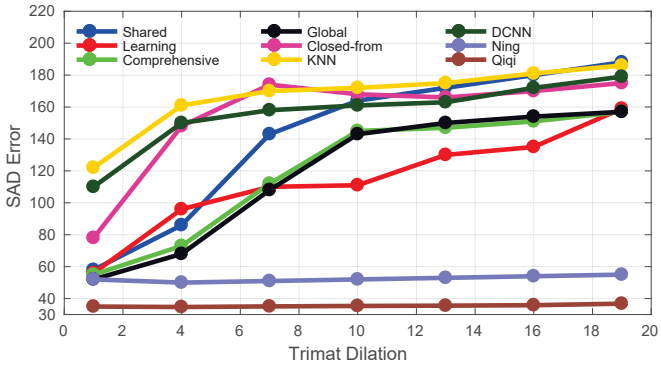


Figure 5. Sensitivity test with respect to trimap sizes.

4.2. Evaluation on foreground images

As existing deep learning methods only output alpha maps, we compare to three representative non-deep learning matting methods, namely Global Matting [19], Closed-Form Matting [27] and KNN Matting [6], on how well foreground images can be extracted from single input images on the Composition-1K dataset. We calculate the SAD and MSE of $\boldsymbol{\alpha} * \mathbf{F}$ following the previous work [38]. As shown in Table 2, our method reduces the error by a large margin.

4.3. Ablation study

As discussed in Section 4.1, our two-encoder structure brings in the major performance improvement. Besides, we found that proper loss functions and data augmentations are also important to obtain high-quality matting results and help generalizing to real-world images.

Loss functions. As shown in Table 1, our model (2) with the Laplacian loss $\mathcal{L}_{lap}^{\alpha}$ generates more numerically accurate results than our model (1) with the loss used in Deep Matting [52]. Our model (3) generates better result compared to the model (4) with both the Laplacian loss and the feature loss $\mathcal{L}_{F}^{\alpha}$. On the other hand, the feature loss enables our model (4) to generate visually better results that keep more final structures than our model (3), as shown in the last example in Figure 6. This is consistent with many other works on image synthesis tasks that the feature loss tends to produce perceptually better results (often at the expense of the numerical performance) [3, 14, 25, 35, 40, 53, 55].

When training our network with both the foreground decoder and the alpha decoder, color loss functions, namely $\mathcal{L}_{1}^{c}$ and $\mathcal{L}_{F}^{c}$, are naturally needed. By comparing models (4) and (6), (5) and (7) in Table 1, we can find that these color losses can improve the alpha map estimation slightly. This is in part because the color and the alpha decoders share

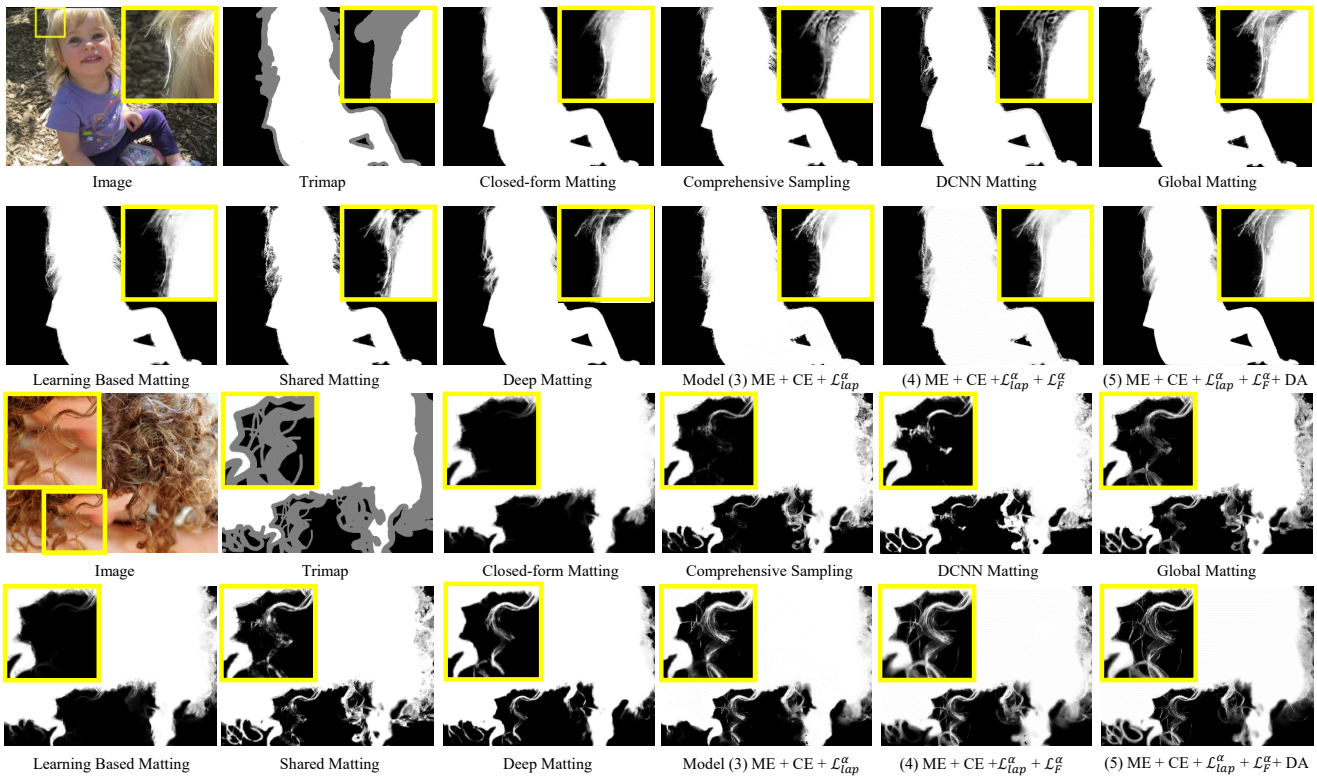


Figure 6. Comparison of the alpha matte on the real world images dataset [52].

the same learned features, and the tasks of foreground color estimation and alpha map estimation are relevant.

Data augmentation. As shown in Table 1, data augmentations, such as ReJPEGing and Gaussian blur, can greatly increases the errors of our methods on the Composition-1k testing dataset. On the other hand, we found that these data augmentations can greatly improve the generalization of our trained networks on the real world images. As shown in Figure 6, when trained with these data augmentation strategies, our models can maintain more fine details, such as hairs. Since these real world examples do not have ground truth, to obtain objective scores of these results, we evaluate the quality of composition results using our matting results. Specifically, we composite the foreground objects in source images onto some external background images using our matting results and then measure the visual quality of the composition results using the NIMA quality assessment algorithm [48]. As reported in Table 4, our data augmentation algorithms are helpful. We also test our methods on the Spectral Matting dataset [28] with the known ground truth. This dataset is generated by photographing dolls in front of a computer monitor displaying seven different background images. The trimap is generated by dilating the alpha map by 20 pixels by alpha map denoising. Our method with DA outperforms our method without DA significantly according to most of the metrics: 3.58 vs 4.28 (SAD), 6.64 vs 9.05 (MSE), and 2.57 vs 3.19 (Conn), and slightly reduces the performance according to Grad: 2.04 vs 1.92.

4.4. User study

To further evaluate the quality of our results, we conducted a user study. We compared our method (model (7)) with three representative methods, including Deep Matting [52] and two state-of-the-art non-deep learning methods, Close-form Matting [27] and Global matting [19].

Our study used all the 31 real-world images from Xu *et al*. [52]. We used a similar protocol to Xu *et al*. [52] to produce the results for the study. For the methods except Deep Matting, we composite the predicted foreground and alpha map onto a blank background image. We use the black background or the white background randomly with the exceptions that for certain foreground images, a particular background color is not appropriate. For example, it is meaningless to composite the black hair onto a black background image, so for such an example, we choose to use the white background. Since Deep Matting does not output the foreground image, we composite the input image using the estimated alpha map as suggested in their paper. Therefore, the comparison between our results with those from Deep Matting should be interpreted with a grain of salt.

Our user study recruited 42 students with different backgrounds. None of them have previous experience with the matting task. Therefore, we conducted a training session for each participant before the formal study. Specifically, each of them was shown two real-world images. For each image, we showed two matting results from different meth-

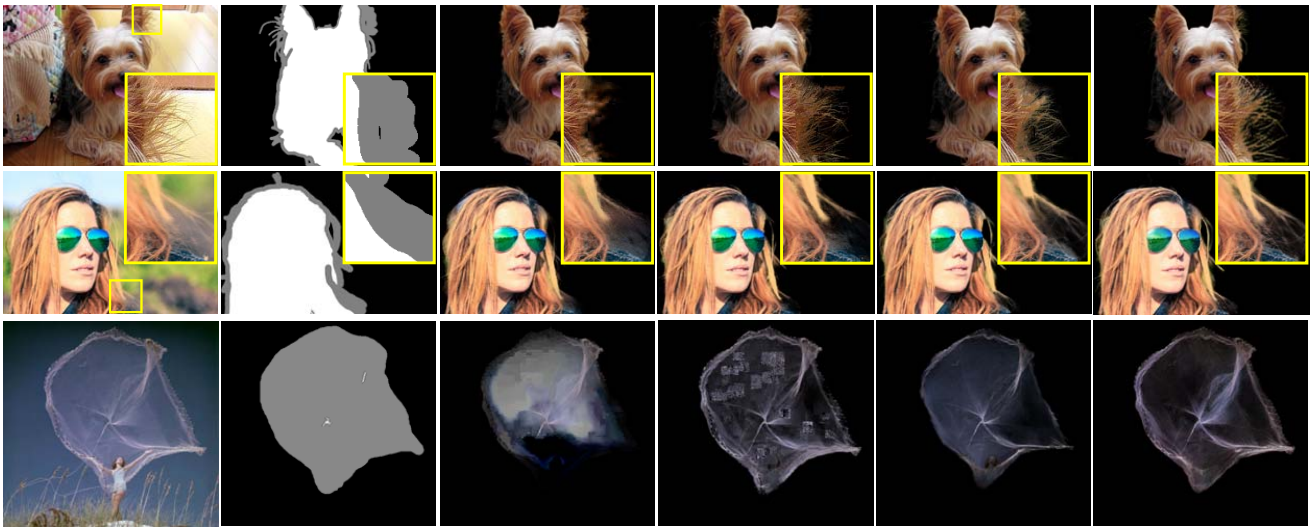


Figure 7. Comparison of the composite results on the real world image dataset[52].

ods without revealing which methods were used to generate these results. We then explained the differences between two results to the participant. This training session is helpful as the subtle difference in matting results was often difficult to spot for people with no prior matting experience.

In our study, we divided the 42 participants into three groups. Each group evaluated how our results compared to one of the three existing methods. In each trial, a participant was presented with a screen that only shows a source image and two corresponding matting results at a time. The participant could select which image to view by clicking the corresponding button or using the *left* or *right* key on the keyboard. In this way, the participant can flip between different images to examine the quality or compare the difference. In each trial, the participant was asked to choose a more accurate and realistic result between the two results. Each participant conducted 31 trials so that the results for all the 31 testing images are evaluated.

We calculated the percentage of the times that our results were preferred by the participants and then calculated the average and the standard deviation for each group. As reported in Table 5, more of our results are preferred by the participants than all the comparing methods. Figure 7 shows some examples in our study. They show that our method can better capture very fine structures like the hair in the first example even when the hair shares a similar color to the background. In the last example, our result not only keeps the delicate edge of the lace, which is lost in the other results, but also is free from the color bleeding problem where the blue background color contaminated the result.

Table 5. The user study in the real world image dataset [52].

Ours vs	Mean preference rate	Std
Global Matting [19]	85.48%	0.21
Closed-form Matting [27]	84.11%	0.19
Deep Matting [52]	77.67%	0.24

5. Conclusion

This paper presented a context-aware deep matting method for simultaneously estimating the foreground and the alpha map from a single natural image. We developed a two-encoder-two-decoder neural network for this task. The two encoders were designed to capture both the local fine structures and the more global context information to disambiguate the foreground and background with a similar appearance. The two decoders output the foreground and the alpha map respectively. Our experiments showed that using the feature loss helps to obtain visually more pleasant matting results while the Laplacian loss tends to optimize the numerical performance. Our experiments also showed that dedicated data augmentation methods, such as Re-JPEGING and Gaussian blurring, are helpful to generalize the neural network trained on a synthetic dataset to handle real-world challenging matting tasks.

Acknowledgments. The source images in Figure 1 are used under a Creative Commons license from Flickr users Robbie Sproule, MEGA PISTOLO and Jeff Latimer. The background images in Figure 1 are from the MS-COCO dataset [30]. Source images used in Figure 2, 3, 4, 6, and 7 are from the matting dataset shared by Xu *et al*. [52]. We thank Nvidia for their GPU donation and Google for their cloud credits.

References

[1] Shruti Agarwal and Hany Farid. Photo forensics from jpeg dimples. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2017.

[2] Yagiz Aksoy, Tunc Ozan Aydin, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image

matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 29–37, 2017.

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018.

[4] Guanying Chen, Kai Han, and Kwan-Yee K Wong. TOM-Net: Learning Transparent Object Matting from a Single Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9233–9241, 2018.

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[6] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. KNN matting. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2175–2188, 2013.

[7] Xiaowu Chen, Dongqing Zou, Steven Zhiying Zhou, Qinping Zhao, and Ping Tan. Image matting with local and nonlocal smooth priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1902–1907, 2013.

[8] Donghyeon Cho, Yu-Wing Tai, and Inso Kweon. Natural image matting using deep convolutional neural networks. In *European Conference on Computer Vision*, pages 626–643. Springer, 2016.

[9] Donghyeon Cho, Yu-Wing Tai, and In So Kweon. Deep convolutional neural network for natural image matting using initial alpha mattes. *IEEE Transactions on Image Processing*, 28(3):1054–1067, 2019.

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

[11] Yung-Yu Chuang, Brian Curless, David H Salesin, and Richard Szeliski. A bayesian approach to digital matting. In *null*, page 264. IEEE, 2001.

[12] Tiago José De Carvalho, Christian Riess, Elli Angelopoulou, Helio Pedrini, and Anderson de Rezende Rocha. Exposing digital image forgeries by illumination color classification. *IEEE Transactions on Information Forensics and Security*, 8(7):1182–1194, 2013.

[13] Virginia R de Sa. Learning classification with unlabeled data. In *Advances in neural information processing systems*, pages 112–119, 1994.

[14] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems*, pages 658–666, 2016.

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[16] Eduardo SL Gastal and Manuel M Oliveira. Shared sampling for real-time alpha matting. In *Computer Graphics Forum*, volume 29, pages 575–584. Wiley Online Library, 2010.

[17] Pallabi Ghosh, Vlad Morariu, Bor-Chun IS Larry Davis, et al. Detection of metadata tampering through discrepancy between image content and metadata using multi-task deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 60–68, 2017.

[18] Bei He, Guijin Wang, Chenbo Shi, Xuanwu Yin, Bo Liu, and Xinggang Lin. Iterative transductive learning for alpha matting. In *2013 IEEE International Conference on Image Processing*, pages 4282–4286. IEEE, 2013.

[19] Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In *CVPR 2011*, pages 2049–2056. IEEE, 2011.

[20] Kaiming He, Jian Sun, and Xiaoou Tang. Fast matting using large kernel matting laplacian matrices. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2165–2172. IEEE, 2010.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Fangjun Huang, Jiwu Huang, and Yun Qing Shi. Detecting double jpeg compression with the same quantization matrix. *IEEE Transactions on Information Forensics and Security*, 5(4):848–856, 2010.

[23] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 101–117, 2018.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[25] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[26] Philip Lee and Ying Wu. Nonlocal matting. In *CVPR 2011*, pages 2193–2200. IEEE, 2011.

[27] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):228–242, 2008.

[28] Anat Levin, Alex Rav-Acha, and Dani Lischinski. Spectral matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(10):1699–1712, 2008.

[29] Chao Li, Ping Wang, Xiangyu Zhu, and Huali Pi. Three-layer graph framework with the sumd feature for alpha matting. *Computer Vision and Image Understanding*, 162:34–45, 2017.

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In

European conference on computer vision, pages 740–755. Springer, 2014.

[31] Qingzhong Liu. Detection of misaligned cropping and recompression with the same quantization matrix and relevant forgery. In *Proceedings of the 3rd international ACM workshop on Multimedia in forensics and intelligence*, pages 25–30. ACM, 2011.

[32] Weiqi Luo, Jiwu Huang, and Guoping Qiu. Jpeg error analysis and its applications to digital image forensics. *IEEE Transactions on Information Forensics and Security*, 5(3):480–491, 2010.

[33] Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. Alphagan: Generative adversarial networks for natural image matting. *arXiv preprint arXiv:1807.10088*, 2018.

[34] Tian-Tsong Ng, Shih-Fu Chang, and Q Sun. A data set of authentic and spliced image blocks. *Columbia University, ADVENT Technical Report*, pages 203–2004, 2004.

[35] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1710, 2018.

[36] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–270, 2017.

[37] Alin C Popescu and Hany Farid. Exposing digital forgeries by detecting traces of resampling. *IEEE Transactions on signal processing*, 53(2):758–767, 2005.

[38] Brian L Price, Bryan S Morse, and Scott Cohen. Simultaneous foreground, background, and alpha estimation for image matting. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2157–2164. IEEE, 2010.

[39] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually motivated online benchmark for image matting. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1826–1833. IEEE, 2009.

[40] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.

[41] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018.

[42] Ehsan Shahrian, Deepu Rajan, Brian Price, and Scott Cohen. Improving image matting using comprehensive sampling sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 636–643, 2013.

[43] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *European Conference on Computer Vision*, pages 92–107. Springer, 2016.

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[45] Alvy Ray Smith and James F Blinn. Blue screen matting. In *SIGGRAPH*, volume 96, pages 259–268, 1996.

[46] Jian Sun, Jiaya Jia, Chi-Keung Tang, and Heung-Yeung Shum. Poisson matting. In *ACM Transactions on Graphics (ToG)*, volume 23, pages 315–321. ACM, 2004.

[47] Ashwin Swaminathan, Min Wu, and KJ Ray Liu. Digital image forensics via intrinsic fingerprints. *IEEE transactions on information forensics and security*, 3(1):101–117, 2008.

[48] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Trans. Image Process.*, 27(8):3998–4011, 2018.

[49] Jue Wang and Michael F Cohen. An iterative optimization approach for unified image segmentation and matting. In *IEEE International Conference on Computer Vision*, pages 936–943. IEEE, 2005.

[50] Jue Wang and Michael F Cohen. Optimized color sampling for robust matting. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[51] Jue Wang and Michael F Cohen. Image and video matting: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(2):97–175, 2008.

[52] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2970–2979, 2017.

[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[54] Yuanjie Zheng and Chandra Kambhamettu. Learning based digital matting. In *2009 IEEE 12th international conference on computer vision*, pages 889–896. IEEE, 2009.

[55] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.