SCIENCE CHINA
Information Sciences

• MOOP •

February 2022, Vol. 65 124101:1–124101:3
https://doi.org/10.1007/s11432-021-3279-y

Automatic image matting and fusing for portrait synthesis

Zhike YI, Wenfeng SONG, Shuai LI* & Aimin HAO

State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191*, China*

Received 15 January 2021/Revised 16 April 2021/Accepted 5 May 2021/Published online 18 January 2022

Citation Yi Z K, Song W F, Li S, et al. Automatic image matting and fusing for portrait synthesis. Sci China Inf Sci, 2022, 65(2): 124101, https://doi.org/10.1007/s11432-021-3279-y

We propose an automatic image matting and fusing system for portrait synthesis in this study. We firstly use a face detection algorithm to determine if the input contains a face. Then, we use a semantic segmentation neural network to generate a trimap and feed the trimap and the portrait into the neural network to predict the alpha channel value. Finally, the input portrait's background is replaced with the given background via an image synthesis algorithm to obtain the synthesized portrait.

Trimap generation methods. As an important annotation, the trimap is used as an input for the matting algorithm. In addition to manual annotation, some algorithms are used for automatic trimap generation, including depth-assisted methods, binary segmentation-based methods, and the combination of image features and morphological dilation. In addition, some methods attempt to integrate trimap generation into the network structure; however, these methods require an estimated specified initial trimap. Inspired by these algorithms, we propose to automatically generate a trimap via a three-class semantic segmentation neural network, without requiring the original trimap.

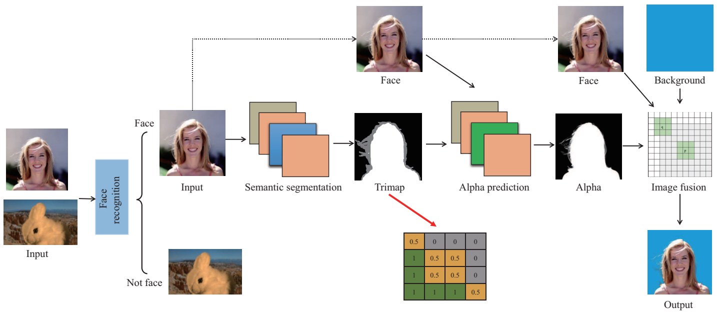
Alpha channel prediction methods. Existing alpha channel prediction algorithms can be partitioned into two categories: sampling-based methods and propagation-based methods. The basic principle of the sampling-based methods is to determine the foreground and background colors for a given pixel by sampling the pixel color. If the category (foreground or background) of a given pixel is determined, the alpha channel value of the pixel is calculated based on the actual color value of the pixel. The core assumption of the algorithm is that, in the vicinity of the boundary between the foreground and background, their color distributions should be consistent. Propagation-based methods can prevent alpha discontinuity problems with respect to sampling-based methods. Particularly, propagation-based methods use the constraints between adjacent pixels and propagate the opacity from the determined area to the unknown area to resolve the problems. Recently, deep learning has achieved remarkable success in many computer-vision tasks. Deep learning algorithms have also emerged for the prediction of opacity channels. Liu et al. [1] proposed the usage of coarse annotated data coupled with fine annotated data to boost end-to-end semantic human matting without trimaps as an extra input. In this study, we optimize the two parts separately instead of end-to-end training so that the sub-image generation model and the alpha channel prediction model can be particularly optimized alternatively in the training phase.

Fusion methods. Primary fusion studies are based on an alpha channel that obtains the fused images via a non-linear combination. This result is dependent on the accuracy of the alpha channel. Thereafter, gradient-domain studies are developed for detailed preservations, such as Poisson cloning and advanced Poisson cloning. These types of studies can significantly preserve color distributions. In this study, we propose a gradient domain-based detail-preserving fusion mechanism to synthesize portraits with a specified background that facilitates the prevention of sharp edge noise.

- **Pipeline.** From Figure 1 [1–4], our portrait synthesis involves replacing the input portrait's background. The main steps include face detection, trimap generation, alpha prediction, and image synthesis. To automate these steps, we initially use a face detection algorithm to determine if the input contains a face [5]. For the filtered portrait images, we use a semantic segmentation neural network to generate a trimap, and thereafter feed the automatically generated trimap and the portrait into the neural network to predict the alpha channel value. Herein, combining with the image synthesis algorithm, the input portrait's background is replaced with the given background, and the synthesized portrait is automatically obtained.

- **Portrait trimap generation.** To synthesize the portrait automatically, an important step is to determine the value of the opacity channel. The trimap partitions the image input into a foreground area (white, opacity value of 1), background area (black, opacity value of 0), and an unknown area (gray, opacity value is unknown) that enables the algorithm to focus only on the solution of the unknown region. However, it is difficult to automatically generate trimaps. Consequently, we use the semantic segmentation neural network based on DeepLab-v3+ and integrate with the three-point icon annotation for training, to generate the

* Corresponding author (email: lishuai@buaa.edu.cn)



(a)

Method	SAD	MSE	Grad	Conn
Shared matting	128.9	0.091	126.5	135.3
Global matting	133.6	0.068	97.6	133.3
Deep image matting	54.6	0.017	36.7	55.3
BSH matting	58.0	0.026	16.0	54.0
Ours	54.4	0.018	35.0	54.9

(b)

Figure 1 (Color online) (a) The pipeline of our method. We initially use a face detection algorithm to determine if the input contains a face. Thereafter, we use a semantic segmentation neural network to generate a trimap and feed the trimap and the portrait into the neural network to predict the alpha channel value. Subsequently, the input portrait's background is replaced with the given background via an image synthesis algorithm to obtain the synthesized portrait. (b) Performance comparison among different matting methods: shared matting [2], global matting [3], deep image matting [4], BSH matting [1], and ours.

trimap automatically, and enable the replacement of the portrait image background to be completed by the neural network.

Semantic segmentation distinguishes the objects in the image from the background and determines the target category of each pixel, which is a two-category problem. In contrast to traditional semantic segmentation, some pixel values at the boundaries of the background and portrait are considered as a single category, that is, the problem of generating tripartite maps is converted into a trimap (portrait, background, and boundary area) generating problem.

Particularly, for portrait image input, the image require pre-processing, mainly scaling and filling the picture to a fixed size. After importing semantic segmentation and the neural network to obtain the input, it also requires proportional scaling and cropping to restore it to its original input size. The trimap annotation performs the same deformation operation as the input, calculates the cross-entropy loss function with the prediction result of the network and returns it to the network for optimization. The loss is given by

$$\text{Loss} = -\frac{1}{n}\sum_{i}^{n}\sum_{j}^{T}\sum_{k}^{m} y_k \log(P_{jk}), \tag{1}$$

where n, T, and m represent the number of images, categories, and pixels respectively. In this study, the trimap is generated as a three-category problem. So T is taken as 3, and y_k indicates the category of the pixel k. P_{jk} indicates the probability that the pixel k belongs to category j. In practice, to improve the accuracy of prediction, we initially pre-train the semantic segmentation neural network on the VOC [6] (including 21 types of background objects), the human body segmentation (two classifications), and the trimap annotation to determine the final model. When training an opacity channel prediction network model, the parameters of the semantic segmentation neural network should be fixed so that the training process can focus on the optimization of the opacity channel.

- **Portrait alpha channel prediction.** The alpha channel prediction is also called image matting. The problem is used to predict the proportion of each pixel in the image belonging to the foreground. The matting algorithm aims to solve the following equation:

$$C_i = \alpha_i F_i + (1 - \alpha_i) B_i, \tag{2}$$

where C_i indicates the pixel value of the corresponding position of the fused image, and F_i and B_i are the foreground and background values of the corresponding pixel, respectively. α_i is the alpha value to be solved, which is between zero and one. In practice, we often only know the pixel values C_i of the fused image, and the remaining parameters are unknown. Therefore, it is significantly difficult to solve α_i from the above equation.

In contrast to previous studies [4], we propose a portrait-matting method. The corresponding gradient image is initially calculated using the Sobel operator. The RGB, trimap, and gradient channels are concatenated into a five-channel network input; then, the input passes through an encoder to obtain deep features. In this study, the network structure of VGG16 [7] is used as the encoder, which contained 14 convolutional layers and five max pooling layers. The encoder samples the original input to 1/32 of the original size to determine deep features that contain the boundary information of the objects in the image. Moreover, the decoder performs upsampling on the deep features to determine an output of the same size as the original image. Finally, through a sigmoid activation layer, it determines the alpha channel prediction. The network decoder is composed by five unpooling layers and six de-convolution layers.

The following loss function is used to optimize each pixel position in training to enable the network to output a reasonable alpha prediction value:

$$\text{Loss}_i = \lambda_1 \sqrt{(\hat{\alpha}_i - \alpha_i)^2 + \varepsilon^2} + \lambda_2 \sum_{k=1}^{3} \sqrt{(\hat{c}_{ik} - c_{ik})^2 + \varepsilon^2}, \tag{3}$$

where $\hat{\alpha}_i$ is the alpha channel value predicted by the network, α_i is the ground truth of the alpha channel value. k is the RGB channel, $\hat{c}_{ik}$ represents the pixel value of the channel k of the image synthesized by $\hat{\alpha}_i$, and c_{ik} represents the pixel value of the channel k of the synthesized image according to α_i. ε denotes the regular part of the loss function. In the experiment, $\varepsilon = 10^{-6}$, λ_1, and λ_2 are the weights of the two parts, and $\lambda_1 = \lambda_2 = 0.5$. The first part of the loss function shows the difference between the predicted and ground-truth alpha values, and the second part shows the difference between the predicted and actual alpha values after synthesizing the image. An adaptive moment estimation optimization algorithm is used to ensure that the training process is stable and convergent.

- **Portrait image fusion.** After obtaining the alpha channel of the image, an image fusion operation is performed for a specific background. To make the edges of the fused image smoother and significantly natural, we use a non-local mean filtering algorithm to perform the denoising operation after fusion. The specific operation is to perform pixel-wise denoising on the unknown area in the trimap.

$$u(p) = \frac{1}{Z(p)} \sum_{q \in N(p,r_1)} w(p,q) v(q), \tag{4}$$

where $u(p)$ is the pixel value after denoising at the position p. $v(q)$ represents the pixel value at the position q. $N(p, r_1)$ represents the adjacent area of p, an image patch centered on p with a side length of $2r_1 + 1$. $Z(p)$ is the normalization coefficient defined as

$$Z(p) = \sum_{q \in N(p,r_1)} w(p,q), \tag{5}$$

where $w(p, q)$ is the weight coefficient determined by the distance of the small area $N(p, r_2)$ located at p and the small area $N(q, r_2)$ located at q. The Euclidean distance $d(p, q)$ is defined as

$$d(p,q) = \frac{1}{2r_2 + 1} \sum_{j \in N(0,r_2)} (v(p+j) - v(q+j))^2. \tag{6}$$

From this Euclidean distance, $w(p, q)$ is defined as

$$w(p,q) = \mathrm{e}^{-\frac{d(p,q)^2}{10^2}}. \tag{7}$$

After the non-local mean filtering operation, the noise in the boundary of the fused image can be effectively removed, and the detailed information of the object boundary is significantly preserved. This improves the effect of the fused image.

• **Experiments and evaluations.** Our method is trained and evaluated on newly collected portraits as well as a public dataset [4]. This dataset contains 481 foreground images with fine alpha channel annotation values. The dataset is split into 431 images for training and 50 for testing. However, only 202 images for training and 11 for testing contain faces. Extending from the previous study [4], we replace 229 training images and 39 testing images with newly collected portraits. In the training dataset, each foreground is merged with 100 different VOC [6] background images. In the testing dataset, each foreground is combined with 50 different COCO [8] background images. The training is conducted on a GTX-2080Ti GPU. We use the metrics SAD, MSE, Grad, and Conn proposed in [9] to evaluate the results. The smaller the values of these indicators, the better is the performance. Performance comparisons are shown in Figure 1.

Training settings. We train the net with VGG16 as the backbone network, and use 6, 12, 24, and 48 epochs for training processing, with the batch size of 1, 2, 4 and 8, where the epoch 12 with batch size 1 gives the best result. We argue that during training, our network learns a universal matting extractor, which is then applicable to new images, specified via a few examples during testing.

Testing settings. In testing stage, the network for inference is the same with the training network.

As presented in Figure 1, our model achieves the optimal results in terms of all four metrics compared with most high-performance methods [1–4]. Note that the MSE is considerably worse than that of the method [4]. It states that the image gradient performs an important role in the training process, and our model can obtain a more accurate alpha value. In addition, BSH [1] uses a significantly large dataset, and our method achieves comparable performance with only 10% of the training dataset. We attribute our high performance to the whole pipeline. We integrate face detection, trimap generation, alpha channel generation, and image fusion to form a complete portrait synthesis framework. The trimap generation and alpha channel generation are implemented by the neural network framework, and a multi-picture parallel operation can be implemented on the GPU. Because face detection and image fusion cannot achieve multi-picture parallel, to reduce the duration of entire process and process multiple input images simultaneously, we implement face detection via a multi-thread mechanism, and use the CUDA kernel for parallel image fusion. For detailed results and performance comparisons, please refer to our complementary video.

Acknowledgements This work was supported in part by National Key R&D Program of China (Grant No. 2018YFB1700603), National Natural Science Foundation of China (Grant Nos. 61672077, 61532002), and Beijing Natural Science Foundation-Haidian Primitive Innovation Joint Fund (Grant No. L182016).

Supporting information Videos and other supplemental documents. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Liu J, Yao Y, Hou W, et al. Boosting semantic human matting with coarse annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 8563–8572

2 Gastal E S L, Oliveira M M. Shared sampling for real-time alpha matting. Comput Graphics Forum, 2010, 29: 575–584

3 He K, Rhemann C, Rother C, et al. A global sampling method for alpha matting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011

4 Xu N, Price B, Cohen S, et al. Deep image matting. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 311–320

5 Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett, 2016, 23: 1499–1503

6 Everingham M, van Gool L, Williams C K I, et al. The pascal visual object classes (VOC) challenge. Int J Comput Vis, 2010, 88: 303–338

7 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556

8 Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. In: Proceedings of European Conference on Computer Vision. Cham: Springer, 2014. 740–755

9 Rhemann C, Rother C, Wang J, et al. A perceptually motivated online benchmark for image matting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009