Noname manuscript No.
(will be inserted by the editor)

Bridge Composite and Real: Towards End-to-end Deep Image Matting

Jizhizi Li[1], **Jing Zhang**[1], **Stephen J. Maybank**[2], **Dacheng Tao**[1],

Received: date / Accepted: date

Abstract Extracting accurate foregrounds from natural images benefits many downstream applications such as film production and augmented reality. However, the furry characteristics and various appearance of the foregrounds, e.g., animal and portrait, challenge existing matting methods, which usually require extra user inputs such as trimap or scribbles. To resolve these problems, we study the distinct roles of semantics and details for image matting and decompose the task into two parallel sub-tasks: high-level semantic segmentation and low-level details matting. Specifically, we propose a novel Glance and Focus Matting network (GFM), which employs a shared encoder and two separate decoders to learn both tasks in a collaborative manner for end-to-end natural image matting. Besides, due to the limitation of available natural images in the matting task, previous methods typically adopt composite images for training and evaluation, which result in limited generalization ability on real-world images. In this paper, we investigate the domain gap issue between composite images and real-world images systematically by conducting comprehensive analyses of various discrepancies between foreground and background images. We find that a carefully designed composition route RSSN that aims to reduce the discrepancies can lead to a better model with remarkable generalization ability. Furthermore, we provide a benchmark containing 2,000 high-resolution real-world animal images and 10,000 portrait images along with their manually labeled alpha mattes to serve as a test bed for evaluating matting model's generalization ability on real-world images. Comprehensive empirical studies have demonstrated that GFM outperforms state-of-the-art methods and effectively reduces the generalization error. The code and the dataset will be released.

✉ Jing Zhang (jing.zhang1@sydney.edu.au)
✉ Dacheng Tao (dacheng.tao@sydney.edu.au)
[1] The University of Sydney, Sydney, Australia
[2] Birkbeck College, London, U.K.

1 Introduction

Image matting refers to extracting the foreground alpha matte from an input image, requiring both hard labels for the explicit foreground or background and soft labels for the transition areas, which plays an important role in many applications, e.g., virtual reality, augmented reality, entertainment, etc. Typical foregrounds in the task of image matting have furry details and diverse appearance, e.g., animal and portrait, which lay a great burden on image matting methods. How to recognize the semantic foreground or background as well as extracting the fine detail for trimap-free natural image matting remains challenging in the image matting community.

For image matting, an image $\mathbf{I}$ is assumed to be a linear combination of foreground $\mathbf{F}$ and background $\mathbf{B}$ via a soft alpha matte $\alpha \in [0, 1]$, *i.e.*,

$$\mathbf{I}_i = \alpha_i \mathbf{F}_i + (1 - \alpha_i)\, \mathbf{B}_i, \tag{1}$$

where i denotes the pixel index. It is a typical ill-posed problem to estimate $\mathbf{F}$, $\mathbf{B}$, and α given $\mathbf{I}$ from Eq. (1) due to the under-determined nature. To relieve the burden, previous matting methods adopt extra user input such as trimap (Xu et al., 2017) and scribbles (Levin et al., 2007) as priors to decrease the degree of unknown. Based on sampling neighboring known pixels (Wang and Cohen, 2007; Ruzon and Tomasi, 2000; Wang and Cohen, 2005) or defining an affinity matrix (Zheng et al.,

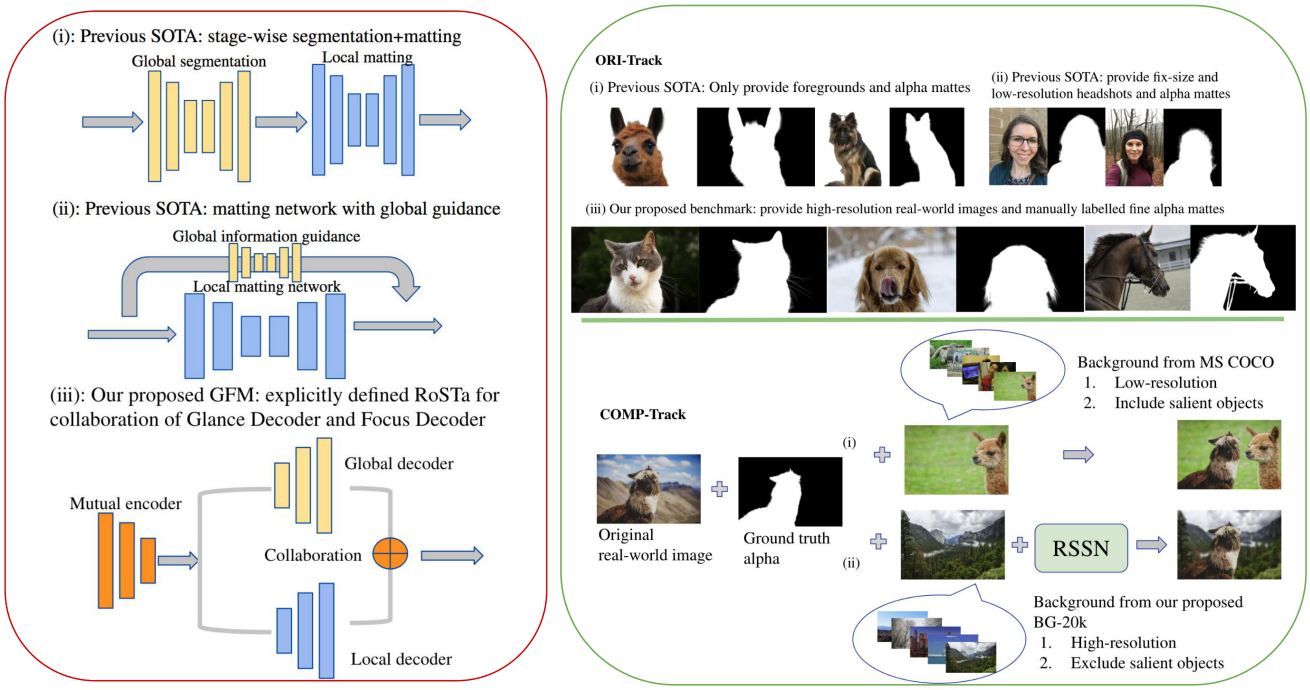
2008), the known alpha values (*i.e.*, foreground or background) are propagated to the unknown pixels. Usually, some edge-aware smoothness constraints are used to make the problem tractable (Levin et al., 2007). However, either the sampling or calculating affinity matrix is based on low-level color or structural features, which is not so discriminative at indistinct transition areas or fine edges. Consequently, their performance is sensitive to the size of unknown areas and may suffer from fuzzy boundaries and color blending. To address this issue, deep convolutional neural network (CNN)-based matting methods have been proposed to leverage its strong representative ability and the learned discriminative features (Xu et al., 2017; Chen et al., 2018; Zhang et al., 2019; Qiao et al., 2020; Liu et al., 2020; Yu et al., 2021). Although CNN-based methods can achieve good matting results, the prerequisite trimap or scribbles make them unlikely to be used in automatic applications such as the augmented reality of live streaming and film production.

To address this issue, end-to-end matting methods have been proposed (Chen et al., 2018; Zhang et al., 2019; Shen et al., 2016; Qiao et al., 2020; Liu et al., 2020) in recent years. Most of them can be categorized into two types. The first type shown in (i) of Figure 1(a) is a straightforward solution which is to perform global segmentation (Aksoy et al., 2018) and local matting sequentially. The former aims at trimap generation (Chen et al., 2018; Shen et al., 2016) or foreground/background generation (Zhang et al., 2019) while the latter is image matting based on the trimap or other priors generated from the previous stage. The shortage of such a pipeline attributes to its sequential nature, since they may generate an erroneous semantic error which could not be corrected by the subsequent matting step. Besides, the separate training scheme in two stages may lead to a sub-optimal solution due to the mismatch between them. The second type is shown in (ii) of Figure 1 (a), global information is provided as guidance while performing local matting. For example, coarse alpha matte is generated and used in the matting networks in (Liu et al., 2020) and in (Qiao et al., 2020), spatial- and channel-wise attention is adopted to provide global appearance filtration to the matting network. Such methods avoid the problem of state-wise modeling and training but bring in new problems. Although global guidance is provided in an implicit way, it is challenging to generating alpha matte for both foreground/background areas and transition areas simultaneously in a single network due to their distinct appearance and semantics.

To solve the above problems, we study the distinct roles of semantics and details for natural image matting and explore the idea of decomposing the task into two parallel sub-tasks, semantic segmentation and details matting. Specifically, we propose a novel end-to-end matting model named Glance and Focus Matting network (GFM). It consists of a shared encoder and two separate decoders to learn both tasks in a collaborative manner for natural image matting, which is trained end-to-end in a single stage. Moreover, we also explore different data representation formats in the global decoder and gain useful empirical insights into the semantic-transition representation. As shown in Figure 1(a)(iii), compared with previous methods, GFM is a unified model that models both sub-tasks explicitly and collaboratively in a single network.

Another challenge for image matting is the limitation of available matting dataset. As shown in Figure 1(b) ORI-Track, due to the laborious and costly labeling process, existing public matting datasets only have tens or hundreds of high-quality annotations (Rhemann et al., 2009; Shen et al., 2016; Xu et al., 2017; Zhang et al., 2019; Qiao et al., 2020). They either only provide foregrounds and alpha mattes (Xu et al., 2017; Qiao et al., 2020) as in (i) of Figure 1(b) ORI-Track, or provide fix-size and low-resolution (800×600) portrait images with inaccurate alpha mattes (Shen et al., 2016) generated by an ensemble of existing matting algorithms as in (ii) of Figure 1(b) ORI-Track. Due to the unavailability of real-world original images, as shown in (i) of Figure 1(b) COMP-Track, a common practice for data augmentation in matting is to composite one foreground with various background images by alpha blending according to Eq. (1) to generate large-scale composite data. The background images are usually choosing from existing benchmarks for image classification and detection, such as MS COCO (Lin et al., 2014) and PASCAL VOC (Everingham et al., 2010). However, these background images are in low-resolution and may contain salient objects. In this paper, we point out that the training images following the above route have a significant domain gap with those natural images due to the *composition artifacts*, attributing to the resolution, sharpness, noise, and illumination discrepancies between foreground and background images. The artifacts serve as cheap features to distinguish foreground from background and will mislead the models during training, resulting in overfitted models with poor generalization on natural images.

In this paper, we investigate the domain gap systematically and carry out comprehensive empirical analyses of the composition pipeline in image matting. We identify several kinds of discrepancies that lead to the domain gap and point out possible solutions to them. We then design a novel composition route named RSSN



(a) (b)

Fig. 1 (a) The comparison between representative end-to-end-matting methods in (i) and (ii) and our GFM in (iii). (b) The comparison between existing matting datasets and our proposed benchmark as well as the comparison between existing composition methods and our RSSN.

that can significantly reduce the domain gap arisen from the discrepancies of resolution, sharpness, noise, etc. Along with this, as shown in (ii) of Figure 1(b) COMP-Track, we propose a large-scale high-resolution clean background dataset (BG-20k) without salient foreground objects, which can be used in generating high-resolution composite images. Extensive experiments show that the proposed composition route along with BG-20k can reduce the generalization error by 60% and achieve comparable performance as the model trained using original natural images. It opens an avenue for composition-based image matting since obtaining foreground images and alpha mattes are much easier than those from original natural images by leveraging chroma keying.

To fairly evaluate matting models' generalization ability on real-world images, we make the first attempt to establish a large-scale benchmark consists of 2000 high-resolution real-world animal images and 10,000 real-world portrait images along with manually carefully labeled fine alpha mattes. Comparing with previous datasets (Xu et al., 2017; Qiao et al., 2020; Shen et al., 2016) as in (i) and (ii) of Figure 1(b) ORI-Track which only provide foreground images or low-resolution inaccurate alpha mattes, our benchmark includes all the high-resolution real-world original images and high-quality alpha mattes (more than 1080 pixels in the shorter side), which are beneficial to train models with better generalization on real-world images, and also suggests several new research problems which will be discussed later.

The contributions of this paper are four-fold:

- We propose a novel model named GFM for end-to-end image matting, which simultaneously generates global semantic segmentation and local alpha matte without any priors as input but a single image.
- We design a novel composition route RSSN to reduce various kinds of discrepancies and propose a large-scale high-resolution background dataset BG-20k to serve as better candidates for generating high-quality composite images.
- We construct a large-scale real-world images benchmark to benefit training a better model with good generalization by its large scale, diverse categories, and high-quality annotations.
- Extensive experiments on the benchmark demonstrate that GFM outperforms state-of-the-art (SOTA) matting models and can be a strong baseline for future research. Moreover, the proposed composition route RSSN demonstrates its value by reducing the generalization error by a large margin.

2 Related Work

2.1 Image Matting

Most classical image matting methods are using auxiliary input like trimaps (Li et al., 2017; Sun et al., 2004; Levin et al., 2008; Chen et al., 2013; Levin et al., 2007). They sample or propagate foreground and background labels to the unknown areas based on local smoothness assumptions. Recently, CNN-based methods improve them by learning discriminative features rather than relying on hand-crafted low-level color features (Xu et al., 2017; Lu et al., 2019; Hou and Liu, 2019; Cai et al., 2019; Tang et al., 2019). Deep Matting (Xu et al., 2017) employed an encoder-decoder structure to extract high-level contextual features. IndexNet (Lu et al., 2019) focused on boundary recovery by learning the activation indices during down-sampling. However, trimap-based methods require user interaction, so are not likely to be deployed in automatic applications. Recently, Chen et al. (Chen et al., 2018) proposed an end-to-end model that first predicted the trimap then carried out matting. Zhang et al. (Zhang et al., 2019) also devised a two-stage model that first segmented the foreground or the background and then refined them with a fusion net. Both methods separate the process of segmentation and matting in different stages, which may generate erroneous segmentation results that mislead the subsequent matting step. Qiao et al. (Qiao et al., 2020) employed spatial and channel-wise attention to integrate appearance cues and pyramidal features while predicting, however, the distinct appearance and semantics of foreground/background areas and transition areas bring a lot of burden to a single-stage network and limit the quality of alpha matte prediction. Liu et al. (Liu et al., 2020) proposed a network to perform human matting by predicting the coarse mask first, then adopted a refinement network to predict a more detailed one. Despite the necessity of stage-wise training and testing, a coarse mask is not enough for guiding the network to refine the detail since the transition areas are not defined explicitly.

In contrast to previous methods, we devise a novel end-to-end matting model via multi-task learning, which addresses the segmentation and matting tasks simultaneously. It can learn both high-level semantic features and low-level structural features in a shared encoder, benefiting the subsequent segmentation and matting decoders collaboratively. One close related work with ours is AdaMatting (Cai et al., 2019), which also has a structure of a shared encoder and two decoders. There are several significant differences: 1) AdaMatting requires a coarse trimap as an extra input while our GFM model only takes a single image as input without any priors; 2) the trimap branch in AdaMatting aims to refine the input trimap, which is much easier than generating a global representation in our case because the initial trimap actually serves as an attention mask for learning semantical features; 3) both the encoder and decoder structures of GFM are specifically designed for end-to-end matting, which differs from AdaMatting; and 4) we systematically investigate the semantic-transition representations in the global decoder and gain useful empirical insights.

2.2 Matting Dataset

Existing matting datasets (Rhemann et al., 2009; Shen et al., 2016; Xu et al., 2017; Zhang et al., 2019; Qiao et al., 2020) only contain foregrounds and a small number of annotated alpha mattes, *e.g.*, 27 training images and 8 test images in alphamatting (Rhemann et al., 2009), 431 training images and 50 test images in Comp-1k (Xu et al., 2017), and 596 training images and 50 test images in HAttMatting (Qiao et al., 2020). DAPM (Shen et al., 2016) proposes 2,000 real-world portrait images but at fix-size and low-resolution, together with limited quality alpha mattes generated by an ensemble of existing matting models. In contrast to them, we propose a high-quality benchmark consists of 10,000 high-resolution real-world portrait images and 2,000 animal images and manually annotated alpha matte for each image. We empirically demonstrate that the model trained on our benchmark has a better generalization ability on real-world images than the one trained on composite images.

2.3 Image Composition

As the inverse problem of image matting and the typical way of generating synthetic dataset, image composition plays an important role in image editing. Researchers have been dedicated to improve the reality of composite images from the perspective of color, lighting, texture compatibility, and geometric consistency in the past years (Xue et al., 2012; Tsai et al., 2017; Chen and Kae, 2019; Cong et al., 2020). Xue et al. (Xue et al., 2012) conducted experiments to evaluate how the image statistical measure including luminance, color temperature, saturation, local contrast, and hue determine the realism of a composite. Tsai et al. (Tsai et al., 2017) proposed an end-to-end deep convolutional neural network to adjust the appearance of the foreground and background to be more compatible. Chen

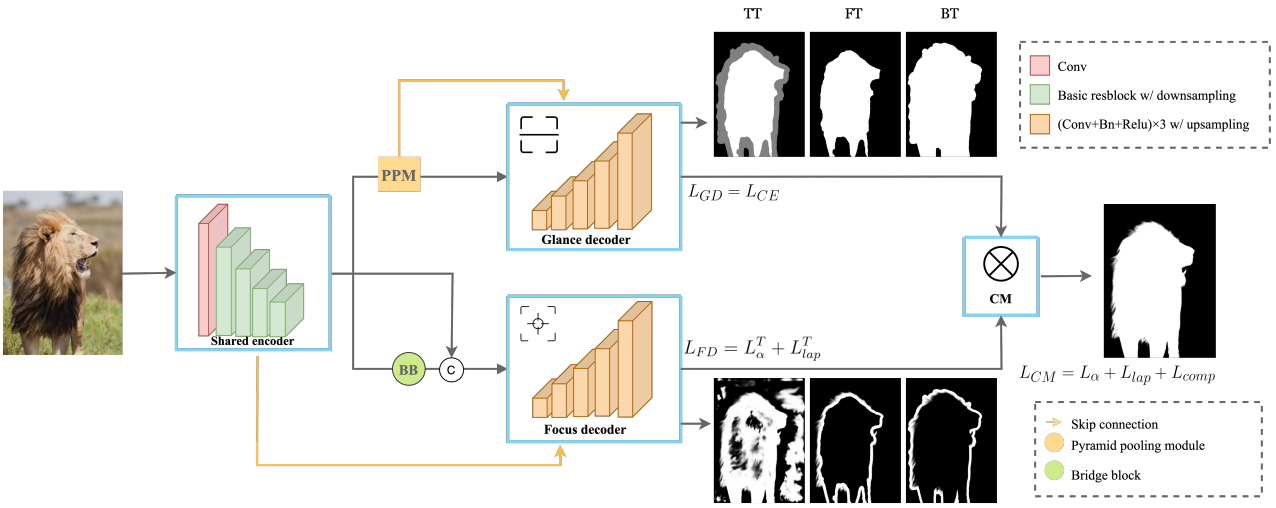


Fig. 2 Diagram of the proposed Glance and Focus Matting (GFM) network, which consists of a shared encoder and two separate decoders responsible for rough segmentation of the whole image and details matting in the transition area.

et al. (Chen and Kae, 2019) proposed a generative adversarial network(GAN) architecture to learn geometrically and color consistent in the composites. Cong et al. (Cong et al., 2020) contributed a large-scale image harmonization dataset and a network using a novel domain verification discriminator to reduce the inconsistency of foreground and background. Although they did a good job in harmonizing the composites to be more realistic, the domain gap still exists when fitting the synthesis data into the matting model, the reason is a subjective agreed standard of harmonization by a human is not equivalent to a good training candidate for a machine learning model. Besides, such procedures may modify the boundary of the foreground and result in inaccuracy of the ground truth alpha matte. In this paper, we alternatively focus on generating composite images that can be used to reduce the generalization error of matting models on natural images.

3 GFM: Glance and Focus Matting Network

When tackling the image matting problem, we humans first glance at the image to quickly recognize the salient rough foreground or background areas and then focus on the transition areas to distinguish details from the background. It can be formulated as a rough segmentation stage and a matting stage roughly. Note that these two stages may be intertwined that there will be feedback from the second stage to correct the erroneous decision at the first stage, for example, in some ambiguous areas due to the protective coloration of animals or occlusions. To mimic the human experience and empower the matting model with proper abilities at both stages, it is reasonable to integrate them into a single model and explicitly model the collaboration. To this end, we propose a novel Glance and Focus Matting network for end-to-end image matting as shown in Figure 2.

3.1 Shared Encoder

GFM has an encoder-decoder structure, where the encoder is shared by two subsequent decoders. As shown in Figure 2, the encoder takes a single image as input and processes it through five blocks $E_0 \sim E_4$, where each reduces the resolution by half. We adopt the ResNet-34 (He et al., 2016) or DenseNet-121 (Huang et al., 2017) pre-trained on the ImageNet training set as our backbone encoder. Specifically, for DenseNet-121, we add a convolution layer to reduce the output feature channels to 512.

3.2 Glance Decoder (GD)

The glance decoder aims to recognize the easy semantic parts and leave the others as unknown areas. To this end, the decoder should have a large receptive field to learn high-level semantics. As shown in Figure 2, we symmetrically stack five blocks $D_4^G \sim D_0^G$ as the decoder, each of which consists of three sequential 3×3 convolutional layers and an upsampling layer. To enlarge the receptive field further, we add a *pyramid pooling module (PPM)* (Zhao et al., 2017; Liu et al., 2019) after E_4 to extract global context, which is then connected to each decoder block D_i^G via element-wise sum.

Loss Function The training loss for the glance decoder is a cross-entropy loss L_{CE} defined as follows:

$$L_{CE} = -\sum_{c=1}^{C} G_g^c log\left(G_p^c\right), \tag{2}$$

where $G_p^c \in [0,1]$ is the predicted probability for cth class, $G_g^c \in \{0,1\}$ is the ground truth label. The output of GD is a two- or three-channel ($C = 2$ or 3) class probability map depends on the semantic-transition representation, which will be detailed in Section 3.4.

3.3 Focus Decoder (FD)

As shown in Figure 2, FD has the same basic structure as GD, *i.e.*, symmetrically stacked five blocks $D_4^F \sim D_0^F$. Different from GD, which aims to do roughly semantic segmentation, FD aims to extract details in the transition areas where low-level structural features are very useful. Therefore, we use a *bridge block (BB)* (Qin et al., 2019) instead of the PPM after E_4 to leverage local context in different receptive fields. Specifically, it consists of three dilated convolutional layers. The features from both E_4 and BB are concatenated and fed into D_4^F. We follow the U-net (Ronneberger et al., 2015) style and add a shortcut between each encoder block E_i and the decoder block D_i^F to preserve fine details.

Loss Function The training loss for FD (L_{FD}) is composed of an alpha-prediction loss L_α^T and a Laplacian loss L_{lap}^T in the unknown transition areas (Xu et al., 2017), *i.e.*,

$$L_{FD} = L_\alpha^T + L_{lap}^T. \tag{3}$$

Following (Xu et al., 2017), the alpha loss L_α^T is calculated as absolute difference between ground truth α and predicted alpha matte α^F in the unknown transition region. It is defined as follows:

$$L_\alpha^T = \frac{\sum_i \sqrt{\left(\left(\alpha_i - \alpha_i^F\right) \times W_i^T\right)^2 + \varepsilon^2}}{\sum_i W_i^T}, \tag{4}$$

where i denotes pixel index, $W_i^T \in \{0,1\}$ denotes whether pixel i belongs to the transition region or not. We add $\varepsilon = 10^{-6}$ for computational stability. Following (Hou and Liu, 2019), the Laplacian loss L_{lap}^T is defined as the $L1$ distance between the Laplacian pyramid of ground truth and that of prediction.

$$L_{lap}^T = \sum_i W_i^T \sum_{k=1}^{5} \left\| (Lap^k(\alpha_i) - Lap^k(\alpha_i^F) \right\|_1, \tag{5}$$

where Lap^k denotes the kth level of the Laplacian pyramid. We use five levels in the Laplacian pyramid.

3.4 RoSTa: Representation of Semantic and Transition Areas

To investigate the impact of the representation format of the supervisory signal in our GFM, we adopt three kinds of **R**epresentations **o**f **S**emantic and **T**ransition **a**reas (RoSTa) as the bridge to link GD and FD.

- **GFM-TT** We use the classical 3-class trimap T as the supervisory signal for GD, which is generated by dilation and erosion from ground truth alpha matte with a kernel size of 25. We use the ground truth alpha matte α in the unknown transition areas as the supervisory signal for FD.
- **GFM-FT** We use the 2-class foreground segmentation mask F as the supervisory signal for GD, which is generated by the erosion of ground truth alpha matte with a kernel size of 50 to ensure the left foreground part is correctly labeled. In this case, the area of $\mathcal{I}(\alpha > 0) - F$ is treated as the transition area, where $\mathcal{I}(\cdot)$ denotes the indicator function. We use the ground truth alpha matte α in the transition area as the supervisory signal for FD.
- **GFM-BT** We use the 2-class background segmentation mask B as the supervisory signal for glance decoder, which is generated by dilation of ground truth alpha matte with kernel size as 50 to ensure the left background part is correctly labeled. In this case, the area of $B - \mathcal{I}(\alpha > 0)$ is treated as the transition area. We use the ground truth alpha matte α in the transition area as the supervisory signal for FD.

3.5 Collaborative Matting (CM)

As shown in Figure 2, CM merges the predictions from GD and FD to generate the final alpha prediction. Specifically, CM follows different rules when using different RoSTa as described in Section 3.4. In GFM-TT, CM replaces the transition area of the prediction of GD with the prediction of FD. In GFM-FT, CM adds the predictions from GD and FD to generate the final alpha matte. In GFM-BT, CM subtracts the prediction of FD from the prediction of GD as the final alpha matte. In this way, GD takes charge of recognizing rough foreground and background by learning global semantic features, and FD is responsible for matting details in the unknown areas by learning local structural features. Such task decomposition and specifically designed parallel decoders make the model simpler than the two-stage ones in (Chen et al., 2018; Zhang et al., 2019). Besides, both decoders are trained simultaneously that the loss can be back-propagated to each of them via the

CM module. In this way, our model enables interaction between both decoders that the erroneous prediction can be corrected in time by the responsible branch. Obviously, it is expected to be more effective than the two-stage framework, where the erroneous segmentation in the first stage could not be corrected by the subsequent one and thus mislead it.

Loss Function The training loss for collaborative matting (L_{CM}) consists of an alpha-prediction loss L_α, a Laplacian loss L_{lap}, and a composition loss L_{comp}, *i.e.*,

$$L_{CM} = L_\alpha + L_{lap} + L_{comp}. \tag{6}$$

Here L_α and L_{lap} are calculated according to Eq. (4) and Eq. (5) but in the whole alpha matte. Following (Xu et al., 2017), the composition loss (L_{comp}) is calculated as the absolute difference between the composite images based on the ground truth alpha and the predicted alpha matte by referring to (Levin et al., 2007). It can be defined as follows:

$$L_{comp} = \frac{\sum_i \sqrt{\left(C(\alpha_i) - C(\alpha_i^{CM})\right)^2 + \varepsilon^2}}{N}, \tag{7}$$

where $C(\cdot)$ denotes the composited image, α^{CM} is the predicted alpha matte by CM, and N denotes the number of pixels in the alpha matte.

To sum up, the final loss used during training is calculated as the sum of L_{CE}, L_{FD} and L_{CM}, *i.e.*,

$$L = L_{CE} + L_{FD} + L_{CM}. \tag{8}$$

4 RSSN: A Novel Composition Route

Since labeling alpha matte of real-world natural images is very laborious and costly, a common practice is to generate large-scale composition images from a few of foreground images and the paired alpha mattes (Xu et al., 2017). The prevalent matting composition route is to paste one foreground with various background images by alpha blending according to Eq. (1). However, since the foreground and background images are usually sampled from different distributions, there will be a lot of *composition artifacts* in the composite images, which lead to a large domain gap between the composition images and natural ones. The composition artifacts may mislead the model by serving as cheap features, resulting in overfitting on the composite images and producing large generalizing errors on natural images. In this section, we systematically analyze the factors that cause the *composition artifacts* including **R**esolution discrepancy, **S**emantic ambiguity, **S**harpness discrepancy, and **N**oise discrepancy. To address these issues, we propose a new composition route named RSSN and a large-scale high-resolution background dataset named BG-20k.

4.1 Resolution Discrepancy and Semantic Ambiguity

In the literature of image matting, the background images used for composition are usually chosen from existing benchmarks for image classification and detection, such as MS COCO (Lin et al., 2014) and PASCAL VOC (Everingham et al., 2010). However, these background images are in low-resolution and may contain salient objects, causing the following two types of discrepancies.

1. **Resolution Discrepancy**: A typical image in MS COCO (Lin et al., 2014) or Pascal VOC (Everingham et al., 2010) has a resolution about 389×466, which is much smaller compared to the high-resolution foreground images in matting dataset such as Compsition-1k (Xu et al., 2017). The resolution discrepancy between foreground and background images will result in obvious artifacts as shown in Figure 3(b).
2. **Semantic ambiguity**: Images in MS COCO (Lin et al., 2014) and Pascal VOC (Everingham et al., 2010) are collected for classification and object detection tasks, which usually contain salient objects from different categories, including various animals, human, objects. Directly pasting the foreground image with such background images will result in semantic ambiguity for end-to-end image matting. For example, as shown in Figure 3(b), there is a dog in the background which is beside the leopard in the composite image. Training with such images will mislead the model to ignore the background animal, *i.e.*, probably learning few about semantics but more about discrepancies.

To address these issues, we collect a large-scale high-resolution dataset named BG-20k to serve as good background candidates for composition. We only selected those images whose shortest side has at least 1080 pixels to reduce the resolution discrepancy. Moreover, we removed those images containing salient objects to eliminate semantic ambiguity. The details of constructing BG-20k are presented as follows.

1. We collected 50k high-resolution (HD) images using the keywords such as *HD background, HD view, HD scene, HD wallpaper, abstract painting, interior design, art, landscape, nature, street, city, mountain, sea, urban, suburb* from websites with open licenses[1], removed those images whose shorter side

[1] `https://unsplash.com/` and `https://www.pexels.com/`

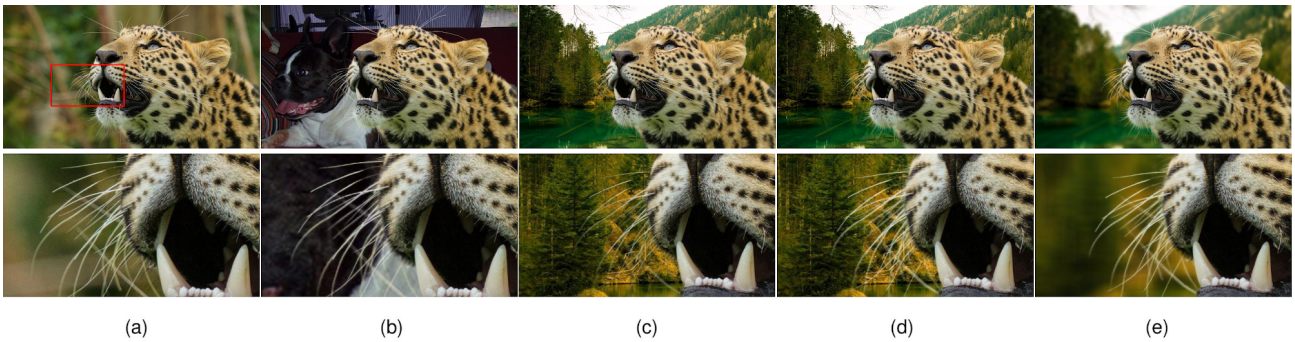


Fig. 3 Comparison of different image composition methods. (a) Original natural image. (b) Composite with background from MS COCO (Lin et al., 2014) with foreground computed by (Levin et al., 2007). (c) Composite with background from our proposed BG-20k by alpha blending of original image directly. (d) Composite with background from our proposed BG-20k with foreground computed by (Levin et al., 2007). (e) Composite with large-aperture effect.

Fig. 4 Some examples from our BG-20k dataset.

has less than 1080 pixels and resized the left images to have 1080 pixels at the shorter side while keeping the original aspect ratio. The average resolution of images in BG-20k is 1180×1539;
2. We removed duplicate images by a deep matching model (Krizhevsky et al., 2012). We adopted YOLO-v3 (Redmon and Farhadi, 2018) to detect salient objects and then manually double-checked to make sure each image has no salient objects. In this way, we built BG-20k containing 20,000 high-resolution clean images;
3. We split BG-20k into a disjoint training set (15k) and validation set (5k).

An composition example using the background image from BG-20k is shown in Figure 3(c) and Figure 3(d). In (c), we use the foreground image computed by multiplying ground truth alpha matte with the original image for alpha blending, in (d), we use the foreground image computed by referring to the method in (Levin et al., 2007) for alpha blending. As can be seen, there are obvious color artifacts in (c) that blends both colors of foreground and background in the fine details. The composite image in (d) is much more realistic than that in (c). Therefore, we adopt the method in (Levin et al., 2007) for computing foreground images in our composition route. More examples of BG-20k are presented in Figure 4 and the supplementary video.

4.2 Sharpness Discrepancy

In photography, it is usually to use a large aperture and focal length to capture a sharp and salient foreground image within the shallow depth-of-field, thus highlighting it from the background context, which is usually blurred due to the out-of-focus effect. An example is shown in Figure 3(a), where the leopard is the center of interest and the background is blurred. Previous composition methods dismiss this effect, producing a domain gap of sharpness discrepancy between the composite images and natural photos. Since we target the image matting task, where the foregrounds are usually salient in the images, thereby we investigate this effect in our composition route. Specifically, we simulate it by adopting the averaging filter in OpenCV with a kernel size chosen from $20, 30, 40, 50, 60$ randomly to blur the background images. Since some natural photos may not have blurred backgrounds, we only use this technique in our composition route with a probability of 0.5. An example is shown in Figure 3(e), where the background is chosen from BG-20k and blurred using the averaging filter. As can be seen, it has a similar style to the original image in (a).

4.3 Noise Discrepancy

Since the foreground and background come from different image sources, they may contain different noise distributions. This is another type of discrepancy, which will mislead the model to search noise cues during training, resulting in overfitting. To address this discrepancy, we adopt BM3D (Dabov et al., 2009) to remove noise in both foreground and background images in RSSN.

Furthermore, we add Gaussian noise with a standard deviation of 10 to the composite image such that the noise distributions in both foreground and background areas are the same. We find that it is effective in improving the generalization ability of trained models.

4.4 The RSSN Composition Route

We summarize the proposed composition route RSSN in Pipeline 1. The input of the pipeline is the matting dataset, e.g., AM-2k and PM-10k as will be introduced in Section 5.1, DIM (Xu et al., 2017), or DAPM (Shen et al., 2016), and the proposed background image set BG-20k. If the matting dataset provides original images, e.g., AM-2K and DAPM (Shen et al., 2016), we compute the foreground from the original image given the alpha matte by referring to (Levin et al., 2007). We random sample K background candidates from BG-20k for each foreground for data augmentation. We set $K = 5$ in our experiments. For each foreground image and background image, we carried out the denoising step with a probability of 0.5. To simulate the effect of large-aperture, we carried out the blur step on the background image with a probability of 0.5, where the blur kernel size was randomly sampled from $\{20, 30, 40, 50, 60\}$. We then generated the composite image according to the alpha-blending equation Eq. (1). Finally, with a probability of 0.5, we added Gaussian noise to the composite image to ensure the foreground and background areas have the same noise distribution. To this end, we generate a composite image set that has reduced many kinds of discrepancies, thereby narrowing the domain gap with natural images.

Pipeline 1: The Proposed Composition Route: RSSN

Input: The matting dataset M containing $|M|$ images and the background image set BG-20k
Output: The composite image set C

for each $i \in [1, |M|]$ **do**
  if there are original images in M, e.g. AM-2k, PM-10k **then**
    Sample an original image $I_i \in M$
    Sample the paired alpha matte $\alpha_i \in M$
    Compute the foreground F_i given (I_i, α_i) (Levin et al., 2007)
  else
    Sample a foreground image $F_i \in M$
    Sample the paired alpha matte $\alpha_i \in M$
  end if
  for each $k \in [1, K]$ **do**
    Sample a background candidate $B_{ik} \in$ BG-20k
    if $random() < 0.5$ **then**
      $F_i = Denoise(F_i)$ //denoising by BM3D (Dabov et al., 2009)
      $B_{ik} = Denoise(B_{ik})$
    end if
    if $random() < 0.5$ **then**
      Sample a blur kernel size $r \in \{20, 30, 40, 50, 60\}$
      $B_{ik} = Blur(B_{ik}, r)$ // the averaging filter
    end if
    Alpha blending: $C_{ik} = F_i \times \alpha_i + B_{ik} \times (1 - \alpha_i)$
    if $random() < 0.5$ **then**
      $C_{ik} = AddGaussianNoise(C_{ik})$
    end if
  end for
end for

5 Empirical Studies

5.1 Benchmark for Real-world Image Matting

Due to the tedious process for generating manually labeled high-quality alpha mattes, the amount of real-world matting dataset is very limited, most previous methods adopted composite dataset such as Comp-1k (Xu et al., 2017), HATT-646 (Qiao et al., 2020) and LF (Zhang et al., 2019) for data augmentation. However, as discussed in Section. 4.4, the *composition artifacts* caused by such convention would result in large domain gap when adapting to real-world images. To fill this gap, we propose two large-scale high-resolution real-world image matting datasets AM-2k and PM-10k, consists of 2,000 animal images and 10,000 portrait images respectively, along with the high-quality manually labeled alpha mattes to serve as an appropriate training and testing bed for real-world image matting. We also setup two evaluation tracks for different purposes. Details are presented as follows.

5.1.1 AM-2k

AM-2k (Animal Matting 2,000 Dataset) consists of 2,000 high-resolution images collected and carefully selected from websites with open licenses. AM-2k contains 20 categories of animals including: *alpaca, antelope, bear, camel, cat, cattle, deer, dog, elephant, giraffe, horse, kangaroo, leopard, lion, monkey, rabbit, rhinoceros, sheep, tiger, zebra*, each with 100 real-world images of various appearance and diverse backgrounds. We ensure the shorter side of the images is more than 1080 pixels. We then manually annotate the alpha mattes using open-source image editing software, e.g., Adobe Photoshop, GIMP, etc. We randomly select 1,800 out of 2,000 to form the training set and the rest 200 as the validation set. Some examples and their ground truth are shown in Figure 5. Please note that AM-2k has no privacy or license issue and is ready to release to public.

Fig. 5 Some examples from our AM-2k dataset. The alpha matte is displayed on the right of the original image.

5.1.2 PM-10k

PM-10k (Portrait Matting 10,000 Dataset) consists of 10,000 high-resolution images collected and carefully selected from websites with open licenses. We ensure PM-10k includes images with multiple postures and diverse backgrounds. We process the images and generate the ground truth alpha mattes as in AM-2k. We then split 9,500 of 10,000 to serve as training set and 500 as validation set.

5.1.3 Benchmark Tracks

To benchmark the performance of matting models 1) trained and tested both on real-world images; and 2) trained on composite images and tested on real-world images, we setup the following two evaluation tracks.

ORI-Track (Original Images Based Track) is set to perform end-to-end matting tasks on the original real-world images. The ORI-Track is the primary benchmark track.

COMP-Track (Composite Images Based Track) is set to investigate the influence of domain gap in image matting. As discussed before, the composite images have a large domain gap with natural images due to the composition artifacts. If we can reduce the domain gap and learn a domain-invariant feature representation, we can obtain a model with better generalization. To this end, we set up this track by making the first attempt towards this research direction. Specifically, we construct the composite training set by alpha-blending each foreground with five background images from the COCO dataset (Lin et al., 2014) (denoted as COMP-COCO) and our BG-20k dataset (denoted as COMP-BG20K), or adopting the composition route RSSN proposed in Section 4.4 based our BG-20K (denoted as COMP-RSSN). Moreover, unlike previous benchmarks that evaluate matting methods on composite images (Xu et al., 2017; Zhang et al., 2019; Qiao et al., 2020), we evaluate matting methods on real-world images in the validation set same as the ORI-Track to validate their generalization ability.

Experiments were carried out on two tracks of the AM-2k and PM-10k datasets: 1) to compare the proposed GFM with SOTA methods, where we trained and evaluated them on the ORI-Track; and 2) to evaluated the side effect of domain gap caused by previous composition method and the proposed composition route, where we trained and evaluated GFM and SOTA methods on the COMP-Track, *i.e.*, COMP-COCO, COMP-BG20k, and COMP-RSSN, respectively.

5.2 Evaluation Metrics and Implementation Details

5.2.1 Evaluation Metrics

Following the common practice in (Rhemann et al., 2009; Zhang et al., 2019; Xu et al., 2017), we used the mean squared error (MSE), the sum of absolute differences (SAD), gradient (Grad.), and connectivity (Conn.) as the major metrics to evaluate the quality of alpha matte predictions. Note that the MSE and SAD metrics evaluate the pixel-wise difference between the prediction and ground truth alpha matte, while the gradient and connectivity metrics favor clear details. Besides, we also use some auxiliary metrics such as Mean Absolute Difference (MAD), SAD-TRAN (SAD in the transition areas), SAD-FG (SAD in the foreground areas), and SAD-BG (SAD in the background areas) to comprehensively evaluate the quality of alpha matte predictions. While MAD evaluates the average quantitative difference regardless of the image size, SAD-TRAN, SAD-FG, and SAD-BG evaluate SAD in different semantic areas, respectively. In addition, we also compared the model complexity of different methods in terms of the number of parameters, computational complexity, and inference time.

5.2.2 Implementation Details

During training, we used multi-scale augmentation similar to (Xu et al., 2017). Specifically, we cropped each of the selected images with size from $\{640 \times 640, 960 \times 960, 1280 \times 1280\}$ randomly, resized the cropped image to 320×320, and randomly flipped it with a probability of 0.5. The encoder of GFM was initialized with the ResNet-34 (He et al., 2016) or DenseNet-121 (Huang et al., 2017) pre-trained on the ImageNet dataset. GFM was trained on two NVIDIA Tesla V100 GPUs. The batch size was 4 for DenseNet-121 (Huang et al., 2017) and 32 for ResNet-34 (He et al., 2016). For the COMP-Track, we composite five training images by using five

Table 1 Results on the ORI-Track and COMP-Track of AM-2k. (d) stands for DenseNet-121 Huang et al. (2017) backbone, (r) stands for ResNet-34 He et al. (2016) backbone. Representations of TT, FT and BT can refer to Section 3.4.

Dataset	**AM-2k**											
Track	ORI											
Method	SHM	LF	SSS	HATT	SHMC	GFM-TT(d)	GFM-FT(d)	GFM-BT(d)	GFM-TT(r)	GFM-FT(r)	GFM-BT(r)	
SAD	17.81	36.12	552.88	28.01	61.50	**10.27**	12.74	12.74	10.89	12.58	12.61	
MSE	0.0068	0.0116	0.2742	0.0055	0.0270	**0.0027**	0.0038	0.0030	0.0029	0.0037	0.0028	
MAD.	0.0102	0.0210	0.3225	0.0161	0.0356	**0.0060**	0.0075	0.0075	0.0064	0.0073	0.0074	
Grad.	12.54	21.06	60.81	18.29	37.00	**8.80**	9.98	9.13	10.00	10.33	9.27	
Conn.	17.02	33.62	555.97	17.76	60.94	**9.37**	11.78	10.07	9.99	11.65	9.77	
SAD-TRAN	10.26	19.68	88.23	13.36	35.23	**8.45**	9.66	8.67	9.15	9.34	8.77	
SAD-FG	0.60	3.79	401.66	1.36	10.93	**0.57**	1.47	3.07	0.77	1.31	2.84	
SAD-BG	6.95	12.55	62.99	13.29	15.34	1.26	1.61	1.00	**0.96**	1.93	1.00	
Track	COMP-COCO			COMP-BG20K			COMP-RSSN					
Method	SHM	GFM-TT(d)	GFM-TT(r)	SHM	GFM-TT(d)	GFM-TT(r)	SHM	GFM-TT(d)	GFM-FT(d)	GFM-BT(d)	GFM-TT(r)	
SAD	182.70	46.16	30.05	52.36	25.19	16.44	23.94	19.19	20.07	22.82	**15.88**	
MSE	0.1017	0.0223	0.0129	0.02680	0.0104	0.0053	0.0099	0.0069	0.0072	0.0078	**0.0049**	
MAD	0.1061	0.0273	0.0176	0.03054	0.0146	0.0096	0.0137	0.0112	0.0118	0.0133	**0.0092**	
Grad.	64.74	20.75	17.22	22.87	15.04	14.64	17.66	13.37	12.53	**12.49**	14.04	
Conn.	182.05	45.39	29.19	51.76	24.31	15.57	23.29	18.31	19.08	19.96	**15.02**	
SAD-TRAN	25.01	17.10	15.21	15.32	13.35	12.36	12.63	12.10	12.12	12.06	**12.03**	
SAD-FG	23.26	8.71	4.74	3.52	3.79	1.46	4.56	4.37	3.47	5.20	**1.15**	
SAD-BG	134.43	20.36	10.1	33.52	8.05	**2.62**	6.74	2.72	4.48	5.56	2.71	
Dataset	**PM-10k**											
Track	ORI											
Method	SHM	LF	SSS	HATT	SHMC	GFM-TT(d)	GFM-FT(d)	GFM-BT(d)	GFM-TT(r)	GFM-FT(r)	GFM-BT(r)	
SAD	16.64	37.51	687.16	22.66	57.85	11.89	12.76	13.45	**11.52**	12.10	13.34	
MSE	0.0069	0.0152	0.3158	0.0038	0.0291	0.0041	0.0044	0.0039	0.0038	0.0037	**0.0036**	
MAD.	0.0097	0.0152	0.3958	0.0131	0.0340	0.0069	0.0074	0.0078	**0.0067**	0.0070	0.0078	
Grad.	14.54	21.82	69.72	15.16	37.28	12.9	12.61	13.22	**13.07**	14.68	13.09	
Conn.	16.13	36.92	691.08	11.95	57.86	11.24	12.15	11.37	10.83	11.38	**10.54**	
SAD-TRAN	8.53	16.36	63.64	9.32	23.04	**7.80**	7.81	**7.80**	8.00	8.82	8.02	
SAD-FG	0.74	11.63	240.02	0.79	13.08	1.65	**0.69**	3.98	0.97	0.93	3.88	
SAD-BG	7.37	9.52	383.49	12.54	21.72	2.44	4.26	1.67	2.54	2.35	**1.44**	
Track	COMP-COCO			COMP-BG20K			COMP-RSSN					
Method	SHM	GFM-TT(d)	GFM-TT(r)	SHM	GFM-TT(d)	GFM-TT(r)	SHM	GFM-TT(d)	GFM-FT(d)	GFM-BT(d)	GFM-TT(r)	
SAD	168.75	61.69	34.58	34.06	21.54	20.29	22.02	18.15	19.68	21.80	**13.84**	
MSE	0.0926	0.0309	0.0165	0.0160	0.0088	0.0086	0.0094	0.0071	0.0078	0.0075	**0.0049**	
MAD	0.0960	0.0355	0.0198	0.0194	0.0125	0.0118	0.0126	0.0106	0.0114	0.0126	**0.0080**	
Grad.	53.83	32.00	18.73	23.02	19.21	16.85	18.65	18.12	17.50	16.97	**14.44**	
Conn	167.28	61.26	33.96	33.7	20.97	19.67	21.61	17.57	19.25	19.09	**13.15**	
SAD-TRAN	23.86	17.69	11.21	12.85	11.82	10.26	10.6	10.78	10.61	10.38	**9.00**	
SAD-FG	21.27	17.42	13.99	9.66	5.31	7.56	5.24	2.66	3.58	5.74	**1.32**	
SAD-BG	123.62	26.59	9.37	11.55	4.41	**2.46**	6.19	4.70	5.48	5.68	3.52	

different backgrounds for each foreground on-the-fly during training. It took about two days to train GFM for 500 epochs on ORI-Track and 100 epochs for COMP-Track. The learning rate was fixed to 1×10^{-4} for the ORI-Track and 1×10^{-5} for the COMP-Track.

For baseline methods **LF** (Zhang et al., 2019) and **SSS** (Aksoy et al., 2018), we used the official codes released by authors. For **SHM** (Chen et al., 2018), **HATT** (Qiao et al., 2020) and **SHMC** (Liu et al., 2020) with no public codes, we re-implemented them according to the papers. For SHMC (Liu et al., 2020) which does not specify the backbone network, we used ResNet-34 (He et al., 2016) for a fair comparison. These models were trained using the training set on the ORI-Track or COMP-Track.

5.3 Quantitative and Subjective Evaluation

5.3.1 Results on the ORI-Track

We benchmarked several SOTA methods (Chen et al., 2018; Zhang et al., 2019; Aksoy et al., 2018; Qiao et al., 2020; Liu et al., 2020) on the ORI-Track of AM-2k and PM-10k. The results are summarized in the top rows of Table 1. GFM-TT, GFM-FT, and GFM-BT denote the proposed GFM model with different RoSTa as described in Section 3.4. (d) and (r) stand for using DenseNet-121 (Huang et al., 2017) and ResNet-34 (He et al., 2016) as the backbone encoder, respectively. There are several empirical findings from Table 1.

First, SSS (Aksoy et al., 2018) achieved the worst performance with a large foreground/background SAD error of 401.66 or 383.49 compare with others, the reason is twofold. 1) They adopt the pre-trained Deeplab-ResNet-101 (Chen et al., 2017) model as the semantic feature extractor to calculate affinities. The pre-trained model may generate limited representative features on the high-resolution matting dataset which degrade the performance. 2) This method aims to extract all the semantic regions in the image while other matting methods are trained to extract only the salient animal or portrait foreground. **Second**, SHMC using global guidance (Liu et al., 2020) and stage-wise method LF (Zhang et al., 2019) perform better than SSS (Aksoy et al., 2018) in all evaluation metrics. However, the SAD errors in the transition area dominate the total errors, which is 35.23 and 19.68, 23.04 and 16.36 for AM-2k and PM-10k, respectively. The reason is that both of them have not explicitly define the transition area, thereby

the matting network has limited ability to distinguish the details in the transition area when it needs segment foreground and background using the same network at the same time. It can also be confirmed by the scores of Grad. and Conn. **Third**, HATT(Qiao et al., 2020) performs better than SHMC (Liu et al., 2020) and LF (Zhang et al., 2019) in terms of SAD error in the transition area and foreground area because the attention module it adopted can provide better global appearance filtration. However, using a single network to model both the foreground and background areas and the transition areas of plentiful details makes it hard to powerful representative features for both areas, resulting in large SAD errors especially in the background areas as well as large Grad. and Conn. errors.

Fourth, SHM (Chen et al., 2018) performs the best among all the SOTA methods. It reduces the SAD error in the transition area from 13.36 to 10.26 for AM-2k, 9.32 to 8.53 for PM-10k, and the SAD error in the background area from 13.29 to 6.95 for AM-2k, 12.54 to 7.37 for PM-10k compared with HATT (Qiao et al., 2020). We believe the improvement credits to the explicit definition of RoSTa (*i.e.*, the trimap) and the PSPNet (Zhao et al., 2017) used in the first stage which has a good semantic segmentation capability. However, SHM (Chen et al., 2018) still has large error in the background area due to its stage-wise pipeline, which will accumulate the segmentation error into the matting network. **Last**, compare with all the SOTA methods, our GFM outperforms them in all evaluation metrics, achieving the best performance by simultaneously segmenting the foreground and background and matting on the transition areas, no matter which kind of RoSTa it uses. For example, it achieves the lowest SAD error in different areas, *i.e.* 8.45 v.s. 10.26 for AM-2k, 7.80 v.s. 8.53 for PM-10k in the transition area, 0.57 v.s. 0.60 for AM-2k, 0.69 v.s. 0.74 for PM-10k in the foreground area, and 0.96 v.s. 6.95 for AM-2k, 1.44 v.s. 7.37 for PM-10k in the background area compared with the previous best method SHM (Chen et al., 2018). The results of using different RoSTa are comparable, especially for FT and BT, since they both define two classes in the image for segmentation by the Glance Decoder. GFM using TT as the RoSTa performs the best due to its explicit definition of the transition area as well as the foreground and background areas. We also tried two different backbone networks, ResNet-34 (He et al., 2016) and DenseNet-121 (Huang et al., 2017). Both of them achieve better performance compared with other SOTA methods.

The reason of GFM's superiority over other methods can be explained as follows. First, compared with stage-wise methods, e.g., SHM (Chen et al., 2018), SSS (Aksoy et al., 2018), and LF (Zhang et al., 2019), GFM can be trained in a single stage and the collaboration module acts as an effective gateway to back-propagate matting errors to the responsible branch adaptively. Second, compared with methods that adopt global guidance, e.g., HATT (Qiao et al., 2020) and SHMC (Liu et al., 2020), GFM explicitly model the end-to-end matting task into two separate but collaborate sub-tasks by two distinct decoders. Moreover, it uses a collaboration module to merge the predictions according to the definition of RoSTa, which explicitly defines the role of each decoder.

From Figure 6, we can find similar observations. SHM (Chen et al., 2018), LF (Zhang et al., 2019), and SSS (Aksoy et al., 2018) fail to segment some foreground parts, implying inferiority of its stage-wise network structure, since they do not distinguish the foreground/background and the transition areas explicitly in the model. It is hard to balance the role of semantic segmentation for the former and matting details for the latter, which requires global semantic and local structural features, respectively. HATT (Qiao et al., 2020) and SHMC (Liu et al., 2020) struggle to obtain clear details in the transition areas since the global guidance is helpful for recognizing the semantic areas while being less useful for matting of details. Compared to them, our GFM achieves the best results owing to the advantage of a unified model, which deals with the foreground/background and transition areas using separate decoders and optimizes them in a collaborative manner. More results of GFM can be found in the supplementary video.

5.3.2 Results on the COMP-Track

We evaluated SHM (Chen et al., 2018), the best performed SOTA method and our GFM with two different backbones on the COMP-Track of AM-2k and PM-10k including COMP-COCO, COMP-BG20K, and COMP-RSSN. The results are summarized in the bottom rows of Table 1, from which we have several empirical findings. **First**, when training matting models using images from MS COCO dataset (Lin et al., 2014) as backgrounds, GFM performs much better than SHM (Chen et al., 2018), *i.e*, 46.16 and 30.05 v.s. 182.70 for AM-2k, 61.69 and 34.58 v.s. 168.75 for PM-10k in terms of whole image SAD, confirming the superiority of the proposed model over the two-stage one for generalization. **Second**, GFM using ResNet-34 (He et al., 2016) performs better than using DenseNet-121 (Huang et al., 2017), perhaps due to the more robust representation ability of residual structure in ResNet-34 (He et al., 2016). **Third**, when training matting models using background

images from the proposed BG-20k dataset, the errors of all the methods are significantly reduced, especially for SHM (Chen et al., 2018), *i.e.*, from 182.70 to 52.36 for AM-2k, or 168.75 to 34.06 for PM-10k, which mainly attributes to the reduction of SAD error in the background area, *i.e.*, from 134.43 to 33.52 for AM-2k and 123.62 to 11.55 for PM-10k. There is the same trend for GFM(d) and GFM(r). These results confirm the value of our BG-20k, which helps to reduce resolution discrepancy and eliminate semantic ambiguity in the background area.

Fourth, when using the proposed RSSN for training, the errors can be reduced further for SHM (Chen et al., 2018), *i.e.*, from 52.36 to 23.94 for AM-2k and 34.06 to 22.02 for PM-10k, from 25.19 to 19.19 and 21.54 to 18.15 for GFM(d), and from 16.44 to 15.88 and 20.19 to 13.84 for GFM(r). The improvement is attributed to the composition techniques in RSSN: 1) we simulate the large-aperture effect to reduce sharpness discrepancy; and 2) we remove the noise of foreground/background and add noise to the composite image to reduce noise discrepancy. Note that the SAD error of SHM (Chen et al., 2018) has dramatically reduced about 87% from 182.70 to 23.93 for AM-2k or 168.75 to 22.02 for PM-10k when using RSSN compared with the traditional composition method based on MS COCO dataset, which is even comparable with the one obtained by training using original images, *i.e.*, 17.81 for AM-2k and 16.64 for PM-10k. It demonstrates that the proposed composition route RSSN can significantly *narrow* the domain gap and help to learn down-invariant features. **Last**, We also conducted experiments by using different RoSTa in GFM(d) on COMP-RSSN, their results have a similar trend to those on the ORI-TRACK.

5.4 Model Ensemble and Hybrid-resolution Test

Table 2 Model Ensemble and Hybrid-resolution Test on AM-2k.

Track		Ensemble				
Method		SAD	MSE	MAD	Grad.	Conn.
GFM-ENS(d)		**9.21**	**0.0021**	**0.0054**	**8.00**	**8.16**
GFM-ENS(r)		9.92	0.0024	0.0058	8.82	8.92
Track		Hybrid				
d_1	d_2	SAD	MSE	MAD	Grad.	Conn.
1/2	1/2	12.57	0.0041	0.0074	9.26	11.75
1/3	**1/2**	**10.27**	**0.0027**	**0.0060**	**8.80**	**9.37**
1/3	1/3	11.58	0.0028	0.0067	11.67	10.67
1/4	1/2	13.23	0.0045	0.0078	10.00	12.26
1/4	1/3	14.65	0.0047	0.0086	12.46	13.68
1/4	1/4	17.29	0.0055	0.0102	16.50	16.28

Model Ensemble Since we propose three different RoSTa for GFM, it is interesting to investigate their complementary. To this end, we calculated the result by a model ensemble which takes the median of the alpha predictions from three models as the final prediction. As shown in Table 2, the result of the model ensemble is better than any single one, *i.e.*, 9.21 v.s. 10.27 for GFM-TT(d), 9.92 v.s. 10.89 for GFM-TT(r), confirming the complementary between different RoSTa.

Hybrid-resolution Test For our GFM, we also proposed a hybrid-resolution test strategy to balance GD and FD. Specifically, we first fed a down-sampled image to GFM to get an initial result. Then, we used the full resolution image as input and only used the predicted alpha matte from FD to replace the initial prediction in the transition areas. For simplicity, we denote the down-sampling ratio at each step as d_1 and d_2, which are subject to $d_1 \in \{1/2, 1/3, 1/4\}$, $d_2 \in \{1/2, 1/3, 1/4\}$, and $d_1 \leq d_2$. Their results are listed in Table 2. A smaller d_1 increases the receptive field and benefits the Glance Decoder, while a larger d_2 benefits the Focus Decoder with clear details in high-resolution images. Finally, we set $d_1 = 1/3$ and $d_2 = 1/2$ for a trade-off.

5.5 Ablation Study

Table 3 Ablation study of GFM on AM-2k.

Track	ORI				
Method	SAD	MSE	MAD	Grad	Conn
GFM-TT(d)	10.27	**0.0027**	**0.0060**	8.80	9.37
GFM-TT(r)	10.89	0.0029	0.0064	10.00	9.99
GFM-TT(r2b)	**10.24**	0.0028	**0.0060**	**8.65**	**9.33**
GFM-TT-SINGLE(d)	13.79	0.0040	0.0081	13.45	13.04
GFM-TT-SINGLE(r)	15.50	0.0040	0.0091	14.21	13.15
GFM-TT(d) excl. PPM	10.86	0.0030	0.0064	9.91	9.92
GFM-TT(d) excl. BB	11.27	0.0035	0.0067	9.33	10.40
GFM-TT(r) excl. PPM	11.90	0.0035	0.0070	10.50	11.07
GFM-TT(r) excl. BB	11.29	0.0032	0.0066	9.59	10.43
Track	COMP-RSSN				
GFM-TT(d)	25.19	0.0104	0.0146	15.04	24.31
GFM-TT(d) w/ blur	21.37	0.0081	0.0124	14.31	20.50
GFM-TT(d) w/ denoise	22.95	0.0090	0.0134	14.37	22.10
GFM-TT(d) w/ noise	19.87	0.0075	0.0116	**13.22**	18.97
GFM-TT(d) w/ RSSN	**19.19**	**0.0069**	**0.0112**	13.37	**18.31**

5.5.1 Results on the ORI-Track

To further verify the benefit of the designed structure of GFM, we conducted ablation studies on several variants of GFM on the ORI-Track of AM-2k, including 1) motivated by Qin et.al (Qin et al., 2019), in GFM encoder when using ResNet-34 (He et al., 2016) as the backbone, we modified the convolution kernel of E_0 from

Fig. 6 Qualitative comparisons on the AM-2k and PM-10k ORI-Track.

7×7 with stride 2 to 3×3 with stride 1, removed the first max pooling layer in E_0, and added two more encoder layers E_5 and E_6 after E_4, each of which had a max pooling layer with stride 2 and three basic res-blocks with 512 filters, denoting "r2b"; 2) using a single decoder to replace both FD and GD in GFM, denoting "SINGLE"; 3) excluding the pyramid pooling module (PPM) in GD, and 4) excluding the bridge block(BB) in FD. The results are summarized in the top rows of Table 3. **First**, when using $r2b$ structure, all the metrics have been improved compared with GFM-TT(r), which is attributed to the larger feature maps at the early stage of the encoder part. However, it has more parameters and computations than GFM-TT(r), which will be discussed later. **Second**, using a single decoder results in worse performance, *i.e.*, SAD increases from 10.27 to 13.79 for GFM-TT(d) and 10.89 to 15.50 for GFM-TT(r), which confirms the value of decomposing the end-to-end image matting task into two collaborative sub-tasks. **Third**, without PPM, SAD increases from 10.86 to 10.27 for GFM-TT(d) and 11.90 to 10.89 for GFM-TT(r), demonstrating that the global context features by PPM due to its larger receptive field are beneficial for semantic segmentation in GD. **Fourth**, without BB, SAD increases from 10.27 to 11.27 for GFM-TT(d) and 10.89 to 11.29 for GFM-TT(r), demonstrating that the learned local structural features by BB due to its

Table 4 Comparison of model parameters, computational complexity, and inference time. (d) and (r) stand for DenseNet-121 (Huang et al., 2017) and ResNet-34 (He et al., 2016).

Method	Parameters (M)	Complexity (GMac)	Inference time (s)
SHM (Chen et al., 2018)	79.27	870.16	0.3346
LF (Zhang et al., 2019)	37.91	2821.14	0.3623
HATT (Qiao et al., 2020)	106.96	1502.46	0.5176
SHMC (Liu et al., 2020)	78.23	139.55	0.4863
GFM-TT(d)	46.96	244.0	0.2085
GFM-TT(r)	55.29	132.28	0.1734
GFM-TT(r2b)	126.85	1526.79	0.2268

dilated convolutional layers are beneficial for matting in FD.

5.5.2 Results on the COMP-Track

In order to verify the different techniques in the proposed composition route RSSN, we conducted ablation studies on several variants of RSSN, including 1) only using the simulation of large-aperture effect, denoting "w/ blur"; 2) only removing foreground and background noise, denoting "w/ denoise"; 3) only adding noise on the composite images, denoting "w/ noise"; and 4) using all the techniques in RSSN, denoting "w/ RSSN". We used the BG-20k for sampling background images in these experiments. The results are summarized in the bottom rows of Table 3. **First**, compared with the baseline model listed in the first row, which was trained using the composite images by alpha-blending, each technique in the proposed composition route is helpful to improve the matting performance in terms of all the metrics. **Second**, simulation of large-aperture effect and adding noise on the composite images are more effective than denoising. **Third**, different techniques are complementary to each other that they contribute to the best performance achieved by RSSN collaboratively.

5.6 Model Complexity Analysis

We compared the number of model parameters (million, denoting "M"), computational complexity (denoting "GMac"), and inference time (seconds, denoting "s") of each method on an image resized to 800×800. All methods are performed on a server with an Intel Xeon CPU (2.30GHz) and an NIVDIA Tesla V100 GPU (16GB memory). As shown in Table 4, GFM using either DenseNet-121 (Huang et al., 2017) or ResNet-34 (He et al., 2016) as the backbone surpasses SHM (Chen et al., 2018), LF (Zhang et al., 2019), HATT (Qiao et al., 2020), and SHMC (Liu et al., 2020) in running speed, *i.e.*, taking about 0.2085s and 0.1734s to process an image. In terms of parameters, GFM has fewer parameters than all the SOTA methods except for LF (Zhang et al., 2019). For computational complexity, GFM has fewer computations than all the SOTA methods when adopting ResNet-34 (He et al., 2016) as the backbone, *i.e.*, 132.28 GMacs. When adopting DenseNet-121 (Huang et al., 2017), it only has more computations than SHMC (Liu et al., 2020) while being smaller. As for GFM(r2b), it has more parameters and computations. Although it can achieve better results, a trade-off between performance and complexity should be made for practical applications. Generally, GFM is light-weight and computationally efficient.

6 Conclusion and Future Works

In this paper, we propose a novel deep matting model for end-to-end natural image matting. It addresses two challenges in the matting task: 1) recognizing various foregrounds with diverse shapes, sizes, and textures from different categories; and 2) extracting details from ambiguous context background. Specifically, a Glance Decoder is devised for the first task and a Focus Decoder is devised for the latter one, while they share an encoder and are trained jointly. Therefore, they collaboratively accomplish the matting task and achieve superior performance than state-of-the-art matting methods. Besides, we also investigate the domain discrepancy issue between composite images and natural ones, which suggests that the common practice for data augmentation may not be suitable for training end-to-end matting models. To remedy this issue, we establish two large-scale real-world image matting datasets AM-2k and PM-10k, which contains 2,000 high-resolution animal images from 20 categories and 10,000 high-resolution

portrait images along with the manually labeled alpha mattes. Furthermore, we systematically analyze the factors affecting composition quality including resolution, sharpness, semantic, and noise, and propose a novel composition route together with a large-scale background dataset containing 20,000 high-resolution images without salient objects, which can effectively address the domain discrepancy issue. Extensive experiments validate the superiority of the proposed methods over state-of-the-art methods. We believe the proposed matting method and composition route will benefit the research for both trimap-based and end-to-end image matting. Moreover, the proposed dataset can provide a test bed to study the matting problem regarding the domain discrepancy issue.

There are several interesting research directions that can be explored in future work. First, after taking a detailed analysis of the error source as evidenced by SAD-TRAN, SAD-FG, and SAD-BG, the error in the transition areas is larger than in that in the foreground and background areas, *i.e.*, 8.45 v.s. 1.83 or 7.80 v.s.4.09, even if the size of transition areas is usually much smaller than that of foreground and background areas. It tells that the performance could be further enhanced by devising a more effective Focus Decoder as well as leveraging some structure-aware and perceptual losses. Second, there is still room to improve the composite-based models to match those trained using original images, since the cost required to generate a composite dataset is much easier than constructing a natural images-based one. Given that alpha matting and alpha blending are inverse problems, it is interesting to see whether or not these two tasks benefit each other if we model them in a single framework. Third, it is interesting to investigate the impact of synthesizing appearance-like composite images to reduce the domain gap and augmentation techniques to avoid overfitting, as well as their differences in improving the generalization ability of matting models.

References

Aksoy Y, Oh TH, Paris S, Pollefeys M, Matusik W (2018) Semantic soft segmentation. ACM Transactions on Graphics 37(4):1–13

Cai S, Zhang X, Fan H, Huang H, Liu J, Liu J, Liu J, Wang J, Sun J (2019) Disentangled image matting. In: Proceedings of the IEEE International Conference on Computer Vision, pp 8819–8828

Chen BC, Kae A (2019) Toward realistic image compositing with adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 8415–8424

Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(4):834–848

Chen Q, Li D, Tang CK (2013) Knn matting. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(9):2175–2188

Chen Q, Ge T, Xu Y, Zhang Z, Yang X, Gai K (2018) Semantic human matting. In: Proceedings of the ACM International Conference on Multimedia, pp 618–626

Cong W, Zhang J, Niu L, Liu L, Ling Z, Li W, Zhang L (2020) Dovenet: Deep image harmonization via domain verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 8394–8403

Dabov K, Foi A, Katkovnik V, Egiazarian K (2009) Bm3d image denoising with shape-adaptive principal component analysis. In: SPARS'09-Signal Processing with Adaptive Sparse Structured Representations

Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88(2):303–338

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778

Hou Q, Liu F (2019) Context-aware image matting for simultaneous foreground and alpha estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 4130–4139

Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4700–4708

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp 1097–1105

Levin A, Lischinski D, Weiss Y (2007) A closed-form solution to natural image matting. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(2):228–242

Levin A, Rav-Acha A, Lischinski D (2008) Spectral matting. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(10):1699–1712

Li X, Liu K, Dong Y, Tao D (2017) Patch alignment manifold matting. IEEE transactions on neural networks and learning systems 29(7):3214–3226

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco:

Common objects in context. In: Proceedings of the European Conference on Computer Vision, pp 740–755

Liu J, Yao Y, Hou W, Cui M, Xie X, Zhang C, Hua Xs (2020) Boosting semantic human matting with coarse annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 8563–8572

Liu JJ, Hou Q, Cheng MM, Feng J, Jiang J (2019) A simple pooling-based design for real-time salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Lu H, Dai Y, Shen C, Xu S (2019) Indices matter: Learning to index for deep image matting. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3266–3275

Qiao Y, Liu Y, Yang X, Zhou D, Xu M, Zhang Q, Wei X (2020) Attention-guided hierarchical structure aggregation for image matting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Qin X, Zhang Z, Huang C, Gao C, Dehghan M, Jagersand M (2019) Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv preprint arXiv:180402767

Rhemann C, Rother C, Wang J, Gelautz M, Kohli P, Rott P (2009) A perceptually motivated online benchmark for image matting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1826–1833

Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International Conference on MICCAI, pp 234–241

Ruzon MA, Tomasi C (2000) Alpha estimation in natural images. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition, pp 18–25

Shen X, Tao X, Gao H, Zhou C, Jia J (2016) Deep automatic portrait matting. In: Proceedings of the European Conference on Computer Vision, pp 92–107

Sun J, Jia J, Tang CK, Shum HY (2004) Poisson matting. ACM Transactions on Graphics 23(3):315–321

Tang J, Aksoy Y, Oztireli C, Gross M, Aydin TO (2019) Learning-based sampling for natural image matting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3055–3063

Tsai YH, Shen X, Lin Z, Sunkavalli K, Lu X, Yang MH (2017) Deep image harmonization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3789–3797

Wang J, Cohen MF (2005) An iterative optimization approach for unified image segmentation and matting. In: Proceedings of the IEEE International Conference on Computer Vision, pp 936–943

Wang J, Cohen MF (2007) Optimized color sampling for robust matting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1–8

Xu N, Price B, Cohen S, Huang T (2017) Deep image matting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2970–2979

Xue S, Agarwala A, Dorsey J, Rushmeier H (2012) Understanding and improving the realism of image composites. ACM Transactions on Graphics 31(4):1–10

Yu Q, Zhang J, Zhang H, Wang Y, Lin Z, Xu N, Bai Y, Yuille A (2021) Mask guided matting via progressive refinement network. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition

Zhang Y, Gong L, Fan L, Ren P, Huang Q, Bao H, Xu W (2019) A late fusion cnn for digital matting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7469–7478

Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2881–2890

Zheng Y, Kambhamettu C, Yu J, Bauer T, Steiner K (2008) Fuzzymatte: A computationally efficient scheme for interactive matting. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition, pp 1–8