# Mask Guided Matting via Progressive Refinement Network

Qihang Yu[1*] Jianming Zhang[2] He Zhang[2] Yilin Wang[2]
Zhe Lin[2] Ning Xu[2] Yutong Bai[1] Alan Yuille[1]
[1] The Johns Hopkins University [2] Adobe

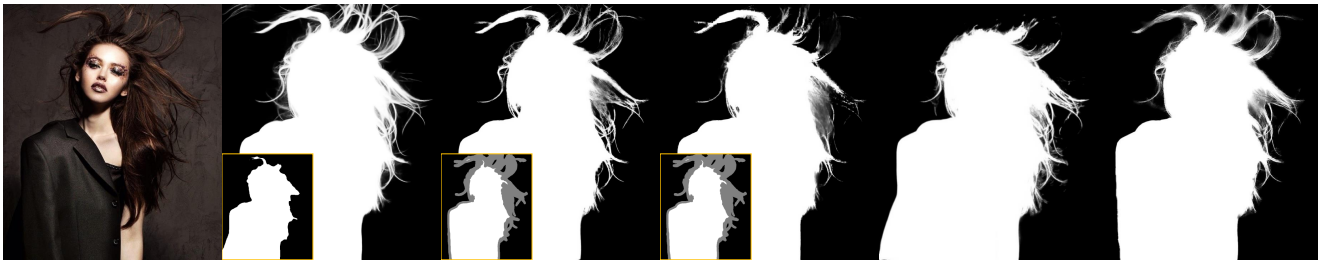


Figure 1: A visual comparison of MG and other matting methods including the commercial matting method in Photoshop. The guidance input (see Sec. 5 for details.) is located at the bottom-left of each image. Note that BSHM [27] has an internal segmentation prediction network thus does not take external mask. Best viewed zoomed in.

## Abstract

*We propose Mask Guided (MG) Matting, a robust matting framework that takes a general coarse mask as guidance. MG Matting leverages a network (PRN) design which encourages the matting model to provide self-guidance to progressively refine the uncertain regions through the decoding process. A series of guidance mask perturbation operations are also introduced in the training to further enhance its robustness to external guidance. We show that PRN can generalize to unseen types of guidance masks such as trimap and low-quality alpha matte, making it suitable for various application pipelines. In addition, we revisit the foreground color prediction problem for matting and propose a surprisingly simple improvement to address the dataset issue. Evaluation on real and synthetic benchmarks shows that MG Matting achieves state-of-the-art performance using various types of guidance inputs. Code and models are available at* https://github.com/yucornetto/MGMatting.

## 1. Introduction

Image matting is a fundamental computer vision problem which aims to predict an alpha matte to precisely cut out an image region. It has many applications in image and video editing [39, 41, 21]. Most previous matting methods require a well-annotated trimap as an auxiliary guidance input [39], which explicitly defines the regions of foreground and background as well as the unknown part for the matting methods to solve. Although such annotation makes the problem more tractable, it can be quite burdensome for users and limits the usefulness of these methods in many non-interactive applications.

Recently, researchers start to study the matting problem in a trimap-free setting. One direction is to get rid of any external guidance, and hope that the matting model can capture both semantics and details by end-to-end training on large-scale datasets [45, 31]. Nevertheless, these methods are faced with the generalization challenge due to the lack of semantic guidance when tested on complex real-world images. Another line of works investigate alternatives to the trimap guidance, easing the requirement for human input [27, 32, 19, 13]. For example, [19, 13] proposed techniques for automatic trimap generation, while [32] takes background images instead as extra inputs. However, these methods often require a very specific type of guidance they are trained with and thus become less appealing when the guidance inputs may have varied characteristics or forms.

In this work, we introduce a Mask Guided (MG) Matting method which takes a general coarse mask as guidance. MG Matting is very robust to the guidance input and can

*Work done during an internship at Adobe.

obtain high-quality matting results using various types of mask guidance such as a trimap, a rough binary segmentation mask or a low-quality soft alpha matte. To achieve such robustness to guidance input, we propose a Progressive Refinement Network (PRN) module, which learns to provide self-guidance to progressively refine the uncertain matting regions through the decoding process. To further enhance the robustness of our method to external guidance, we also develop a series of guidance mask perturbation operations including random binarization, random morphological operations, and also a stronger perturbation CutMask to simulate diverse guidance inputs during training.

In addition to alpha matting prediction, we also revisit the foreground color prediction problem for matting. Without accurately recovering the foreground color in the transparent region, the composited image will suffer from the fringing issue. We note that the foreground color labels in the widely-used dataset [41] are suboptimal for model training due to the labeling noise and limited diversity. As a simple yet effective solution, we propose Random Alpha Blending (RAB) to generate synthetic training data from random alpha mattes and images. We show that such simple method can improve the foreground color prediction accuracy without requiring additional manual annotations. As a result, combining with the proposed PRN, MG Matting is able to generate more visual plausible composition results.

Our contributions can be summarized as follows:

- We propose Mask Guided Matting, a general matting framework working with guidance masks in various qualities and even forms, and achieve a new state-of-the-art performance evaluated on both synthetic and real-world datasets.
- We introduce Progressive Refinement Network (PRN) along with a guidance perturbation training pipeline as a solution to learning a robust matting model.
- We study the problem of foreground color prediction for matting and propose a simple improvement using random alpha blending.

In addition, we collect and release a high-quality matting benchmark dataset of real images to evaluate the real-world performance of matting models.

## 2. Related Work

**Trimap-based Image Matting.** A majority of matting methods requires a trimap as additional input, which divides an image into foreground, background, and unknown regions. Traditional methods are often sampling-based or propagation-based. Sampling-based ones [11, 7, 15, 33, 38] estimate foreground/background color statistics through sampling pixels in the definite foreground/background regions to solve the alpha matte in the unknown region. The propagation-based methods [6, 20, 21, 22, 35, 16], also known as affinity-based methods, estimate alpha mattes by propagating the alpha value from foreground and background pixels to the unknown area.

Recently, deep learning approaches have been proved successful in many areas, including classification [17, 36, 25, 23], detection [14, 2, 3], and segmentation [5, 42]. It also have achieved great success in image matting. [41] created a matting dataset with annotated mattes composited to various background images, and trained a deep network on it. Later, [30] introduced a generative adversarial framework to improve the results. [37] proposed to combine the sampling-based method and deep learning. [29] introduced a new index-guided upsampling and unpooling operations to better keep details in the predictions. [18] proposed a two-encoder two-decoder architectures to simultaneous estimate foreground and alpha. [24] further boost the performance with a contextual attention module.

**Trimap-free Image Matting.** It is noticeable that there are also some trials [1, 34] to get rid of the trimap to predict alpha matte. [45] proposed a framework consisting of a segmentation network and a fusion network, where the input is only a single RGB image. Later, [27] introduced a trimap-free framework consisting of mask prediction network, quality unification network, and matting refinement network for human portrait matting. The trimap-free matting performance is further boosted with attention module [31]. However, these trimap-free methods still have some gap to trimap-based ones in terms of performance. Another direction is to use an alternative guidance to trimap. [32] introduced a framework taking background images along with other potential priors (*e.g.*, segmentation mask, motion cue) as additional inputs. It shows great potential and can obtain a comparable performance to state-of-the-art trimap-based methods.

**Foreground Color Decontamination.** Many conventional matting methods [11, 21] proposed to predict both alpha matte and foreground color for extracting foreground objects. However, it is only very recently [18] incorporated the foreground prediction into the deep learning framework. Later, [32] also predicts foreground color to reduces artifacts for a better composition result. Nevertheless, these methods mainly add a foreground decoder and directly learn from color label in [41], which only provides limited training samples and, more seriously, the color labels can be inaccurate and noisy(see Fig. 3). [10] proposes to use [21] to obtain a smoother color label.

Our method differs from algorithms mentioned above in the following folds: 1) Our model works in a more general setting where only an easy-to-obtain coarse mask, no matter user-defined or model-predicted, is needed as guidance. It could handle different qualities and even various types of guidance as input. Thus it could be used as either trimap-based or trimap-free model depending on what guidance is

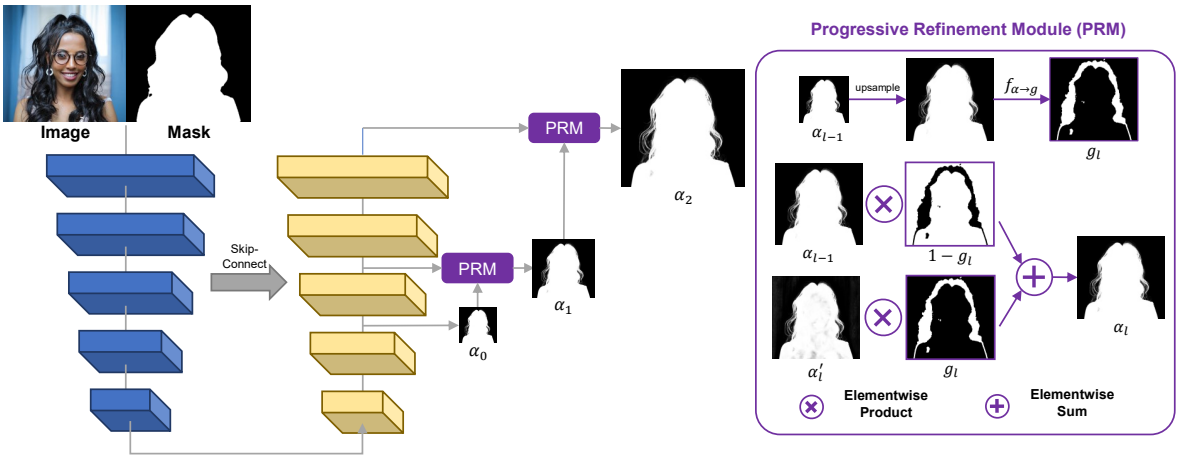


Figure 2: The proposed PRN. The network predicts alpha matte at multiple resolutions, while the one at lower-resolution provides guidance about uncertain region to be refined in the next prediction.

available. Our model could also leverage a stronger guidance to achieve even finer details. 2) Our methods could also predict the foreground color. Unlike [18], where the foreground prediction is directly learned from the color label, we note that the limited training data and inaccurate human label result in undesired results especially in the boundary regions. Instead, we propose to use Random Alpha Blending to avoid the bias in label, which not only introduces more diverse training samples but also avoid the inaccurate color label locating in boundary regions.

## 3. MG Matting

The problem of image matting can be formulated as:

$$\mathbf{I} = \alpha\mathbf{F} + (1-\alpha)\mathbf{B}, \alpha \in [0,1], \tag{1}$$

where $\mathbf{I}$, $\mathbf{F}$, $\mathbf{B}$, and $\alpha$ refer to the image color, foreground color, background color and alpha matte respectively. As only $\mathbf{I}$ is observed, this is a very ill-posed problem. To solve the matting problem, most methods require a trimap input, which labels the foreground region (*i.e.* $\alpha = 1$), the background region (*i.e.* $\alpha = 0$) and the unknown part. In practice, the trimap input can contain various levels of noise and errors, making the matting results inconsistent.

We relax the strong assumption of the trimap by proposing a Mask Guided Matting method. The mask guidance, such as a predicted segmentation mask or a rough manual selection, only provides a coarse spatial prior of the foreground region. Therefore, our MG Matting method needs more high-level semantic understanding of the input mask, so that it can detect the foreground/background region and the soft transparent part robustly. Meanwhile, our model has to capture image low-level patterns such as edge and texture to produce fine details of the target matte. Coordinating the high-level and the low-level feature learning is the key to the design of our MG Matting method.

To this end, we introduce Progressive Refinement Network (PRN), which provides a coarse-to-fine self-guidance to progressively refine the uncertain regions during the decoding process. In the following, we present the details of PRN, the training formulation and some data augmentation techniques to enhance the robustness of our model.

### 3.1. Progressive Refinement Network

An overview of the PRN is shown in Fig. 2. The structure of our PRN follows the popular encoder-decoder network with skip connections. Our network takes an image and a coarse mask as input and outputs a matte. During the decoding process, PRN has a side matting output at each feature level. The side outputs with deep supervision have been shown to improve the feature learning at different scales [40]. However, unlike [40], we find that linearly fusing the side outputs is not ideal for the matting problem (see Table 4 for details). This is because image region closer to the object boundary requires lower-level features to delineate the foreground, while identifying internal object regions needs higher-level guidance.

To address this problem, we introduce a Progressive Refinement Module (PRM) at each feature level to selectively fuse the matting outputs from the previous level and the current level. Specifically, for the current level $l$ we generate a self-guidance mask $g_l$ from the matting output $\alpha_{l-1}$ of the previous level using the following function:

$$f_{\alpha_{l-1} \to g_l}(x,y) = \begin{cases} 1 & \text{if } 0 < \alpha_{l-1}(x,y) < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The $\alpha_{l-1}$ is firstly upsampled to match the size of the raw matting output $\alpha_l'$ of the current level and then produces

resultant self-guidance mask $g_l$. The self-guidance mask defines the transparent region (*i.e.* $0 < \alpha < 1$) as unknown and replaces the unknown region of $\alpha_{l-1}$ with the current raw output $\alpha'_l$ to obtain an updated $\alpha_l$ of current level:

$$\alpha_l = \alpha'_l g_l + \alpha_{l-1}(1 - g_l). \quad (3)$$

In this way, confident regions predicted from the previous higher-level features are preserved and the current level only needs to focus on refining the uncertain region.

In practise, we obtain alpha matte side outputs at three feature levels of stride 8, 4, and 1 respectively (see Fig. 2) and slightly dilate the self-guidance masks for a more robust self-guidance. The initial base matte of $1/8$ image size will be progressively upsampled and refined, and the uncertain regions will also shrink gradually through the decoding process using the proposed PRM. The full network is trained end-to-end to auto-balance the refinement focus at multiple feature levels. Such self-guided refinement also makes model less reliant on the external mask guidance, leading to more robust matting performance.

**Training scheme.** For loss functions, we adopt the $l_1$ regression loss, composition loss [41], Laplacian loss [18] and denote them as $\mathcal{L}_{l1}$, $\mathcal{L}_{comp}$, $\mathcal{L}_{lap}$ respectively. We represent the ground truth alpha with $\hat{\alpha}$ and prediction alpha with $\alpha$. The overall loss functions is the summation of them:

$$\mathcal{L}(\hat{\alpha}, \alpha) = \mathcal{L}_{l1}(\hat{\alpha}, \alpha) + \mathcal{L}_{comp}(\hat{\alpha}, \alpha) + \mathcal{L}_{lap}(\hat{\alpha}, \alpha). \quad (4)$$

The loss is applied to each output head of the network. To make the training more focused on the unknown region, We further modulate the loss with $g_l$. The final loss function can be formulated as:

$$\mathcal{L}_{final} = \sum_l w_l \mathcal{L}(\hat{\alpha}_l \cdot g_l, \alpha_l \cdot g_l), \quad (5)$$

where $w_l$ is the loss weight assigning to the outputs of different levels. We use $w_0 : w_1 : w_2 = 1 : 2 : 3$ in our experiments. $g_l$ is generated from $\alpha_{l-1}$ by Eqn. 2, and $g_0$ is a mask filled with one so that the base level output can be supervised over the whole image to provide more holistic semantic guidance for the next level output.

For data augmentation, we follow the training protocol proposed in [24], including random composite two foreground object images, random resize images with random interpolation methods, random affine transformation, color jitters. We random crop $512 \times 512$ patches centered on an unknown region for training. Each patch is composited to a random background image from MS COCO dataset [26].

**Guidance Perturbation.** To ensure that our model can adapt to guidance masks from different sources and with different qualities, we propose a series of guidance perturbation to generate guidance masks from ground-truth alpha matte during training. Given a ground-truth alpha matte, we first binarize it with a random threshold uniformly sampled from 0 to 1. Then, the mask is dilated and/or eroded in random order with random kernel sizes from 1 to 30.

Figure 3: The color labels in the commonly used training data from [41] are noisy and inaccurate especially near the boundary part. Note that the hair near the ear falsely gets pinker. Best viewed in color and zoomed in.

Moreover, we provide a stronger guidance perturbation named CutMask to further improve the model robustness. Inspired by the successful natural image augmentation CutMix [43], we randomly select a patch size ranging from $1/4$ to $1/2$ image size. Then, two random patches of the guidance are selected and the content of one patch will overwrite another. This stronger perturbation provides additional localized guidance mask corruption, making the model more robust to semantic noises in external guidance masks.

Besides perturbing external guidance masks, we note that perturbing internal self-guidance mask is also very important to improve the robustness. Therefore, we randomly dilate the self-guidance masks to incorporate more variance. Particularly, during training, the self-guidance mask from output stride 8 is dilated by $K_1$ random sampled from $[1, 30]$ and the one from output stride 4 is dilated by $K_2$ from $[1, 15]$. For testing, we fix $K_1 = 15$ and $K_2 = 7$.

### 3.2. Foreground Color Estimation

As indicated in Eqn. 1, both alpha matte and foreground color need to be solved for foreground object extraction. Nevertheless, only a few matting methods learn to predict the foreground color [18, 32] and all of them used the popular Composition-1k dataset [41] for training.

However, there are a couple of issues in the Composition-1k dataset. First of all, this dataset only contains 431 foreground images with matting and foreground color ground truth, which is quite limited to train a foreground color model. Moreover, the foreground color labels, which were estimated using the color decontamination feature in Photoshop [41], are sometimes noisy and inaccurate near the boundary regions (see Fig. 3). This can introduce color spills and other artifacts into the images during data augmentation process, making the learning less stable. Besides, labels are only provided where the alpha value is greater than zero, so existing methods can only apply su-

| Methods | SAD | MSE ($10^{-3}$) | Grad | Conn |
|---|---|---|---|---|
| Learning Based Matting [46] | 113.9 | 48 | 91.6 | 122.2 |
| Closed-Form Matting [21] | 168.1 | 91 | 126.9 | 167.9 |
| KNN Matting [6] | 175.4 | 103 | 124.1 | 176.4 |
| Deep Image Matting [41] | 50.4 | 14 | 31.0 | 50.8 |
| IndexNet Matting [29] | 45.8 | 13 | 25.9 | 43.7 |
| AdaMatting [4] | 41.7 | 10.2 | 16.9 | - |
| Context-Aware Matting [18] | 35.8 | 8.2 | 17.3 | 33.2 |
| GCA Matting [24] | 35.3 | 9.1 | 16.9 | 32.5 |
| Ours$_{\text{TrimapFG}}$ | **31.5** | **6.8** | **13.5** | **27.3** |
| Ours$_{\text{Trimap}}$ | 32.1 | 7.0 | 14.0 | 27.9 |

Table 1: Results on Composition-1k test set. The subscripts denote the corresponding guidance inputs, *i.e*. TrimapFG, Trimap. The other evaluated methods all require a trimap as input.

| Methods | SAD | MSE ($10^{-3}$) | Grad | Conn |
|---|---|---|---|---|
| Learning Based Matting* [46] | 105.04 | 21 | 94.16 | 110.41 |
| Closed-Form Matting* [21] | 105.73 | 23 | 91.76 | 114.55 |
| KNN Matting* [6] | 116.68 | 25 | 103.15 | 121.45 |
| Deep Image Matting* [41] | 47.56 | 9 | 43.29 | 55.90 |
| HAttMatting* [31] | 48.98 | 9 | 41.57 | 49.93 |
| Deep Image Matting [41] | 48.73 | 11.2 | 42.60 | 49.55 |
| + Ours | **36.58** | **7.2** | **27.37** | **35.08** |
| IndexNet Matting [29] | 46.95 | 9.4 | 40.56 | 46.80 |
| + Ours | **35.82** | **5.8** | **25.75** | **34.23** |
| Context-Aware Matting [18] | 36.32 | 7.1 | 29.49 | 35.43 |
| + Ours | **35.04** | **5.4** | **24.55** | **33.35** |
| GCA Matting [24] | 39.64 | 8.2 | 32.16 | 38.77 |
| + Ours | **35.93** | **5.7** | **25.94** | **34.35** |

Table 2: Matting refinement results on Distinction-646 test set. Results with $*$ are from methods trained on Distinction-646 train set as reported in [31] for reference. Other results are only trained on composition-1k.

pervision to the foreground region [18], leading to unstable behaviors in the undefined part.

To address these issues, we propose a simple yet effective method, named Random Alpha Blending (RAB), to generate synthetic training data by blending a foreground image and a background image using a randomly selected alpha matte. Although the composited images may not be semantically meaningful, they can provide accurate and unbiased foreground color labels in the transparent region. The random alpha blending can also significantly make training data more diverse and improve the generalization of the foreground color prediction. Besides, we also note that RAB makes it possible to apply loss supervision over all image, leading to a much smoother prediction which is desired for robust compositing. (See Fig. 4)

For foreground estimation, we train a separate model using a basic encoder-decoder network, which takes an image and an alpha matte as input. The loss function is the summation of $l_1$ regression loss, compositing loss, and Laplacian loss. We note that although training a single model for both matte and foreground color prediction is possible, empirically this will degrade the matting performance [18], and the random alpha blending will destroy the semantic cue for the matting model. In addition, decoupling foreground color prediction from matting makes the color model transferable to the use cases where the matte is already given.

## 4. Experiments on Synthetic Datasets

In this section, we report the evaluation results of our method under the traditional synthetic data setting, where the test images are generated using foreground images with ground truth mattes and random background images.

**Evaluation Metrics.** We follow previous methods to evaluate the results by Sum of Absolute Differences (SAD), Mean Squared Error (MSE), Gradient (Grad) and Connectivity (Conn) errors using the official evaluation code [41].

**Network Architectures.** We adopt ResNet34-UNet proposed in [24] with an Atrous Spatial Pyramid Pooling (ASPP) [5] as the backbone for both PRN and color prediction. The first convolution layer is adjusted to take a 4-channel input consisting of a RGB image along with an external guidance input. Moreover, an alpha prediction head (Conv-BN-ReLU-Conv) is attached to the features at output stride 4 and 8 respectively to obtain side outputs.

**Training stage.** To fairly compare with previous deep image matting methods, we train our MG Matting model using the Composition-1k datasest [41] which contains 431 foreground objects and the corresponding ground-truth alpha mattes for training. The network is initialized with ImageNet [8] pre-trained weight. We use crop size 512, batch size of 40 in total on 4 GPUs, Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate is initialized to $1 \times 10^{-3}$. The training lasts for $100,000$ iterations with warm-up at the first $5,000$ iterations and cosine learning rate decay [28, 12]. We also apply a curriculum learning manner to help the PRN training. Particularly, for the first $5,000$ iterations, the predictions of output stride 4 and 1 will be guided by guidance mask generated from ground-truth alpha, and for the next $10,000$ iterations, the guidance will be evenly and randomly generated from self-prediction and ground-truth alpha. Afterwards, each alpha prediction should fully rely on its self-guidance. The foreground color prediction is trained under the exactly same settings except that the generated training samples are composited by random foreground and alpha matte. It is noticeable that with RAB, we can add foreground color supervision on the whole image instead of only foreground regions, which produces more smooth and stable results (see Fig. 4).

**Testing on Composition-1k.** The test set consists of 50 unique objects which are composited with 20 background

| Methods | SAD | MSE ($10^{-3}$) |
|---|---|---|
| Global Matting [15] | 220.39 | 36.29 |
| Closed-Form Matting [21] | 254.15 | 40.89 |
| KNN Matting [6] | 281.92 | 36.29 |
| Context-Aware Matting [18] | 61.72 | 3.24 |
| Ours | **49.80** | **2.48** |

Table 3: The foreground result ($\alpha \cdot F$) on the Composition-1k dataset.

| Methods | Whole Image | | Unknown Area | |
|---|---|---|---|---|
| | SAD | MSE ($10^{-3}$) | SAD | MSE ($10^{-3}$) |
| Baseline | 43.7 | 4.5 | 39.8 | 11.2 |
| Baseline + Deep Supervision | 37.8 | 3.7 | 36.3 | 9.5 |
| Baseline + Fusion Conv | 38.1 | 3.2 | 36.9 | 8.8 |
| PRN w/o CutMask | 33.9 | 2.9 | 32.8 | 7.5 |
| PRN | 32.3 | 2.5 | 32.1 | 7.0 |

Table 4: Ablation studies on Composition-1k dataset. Baselines: a ResNet34-UNet with ASPP; Deep supervision: adding side outputs and deep supervisions; Fusion Conv: use convolutions to combine different outputs.

images chosen from Pascal VOC [9], thus providing 1000 test samples in total. We note that since these synthetic datasets use PASCAL VOC images as background which may contain other salient objects, saliency/segmentation models may not be applicable to obtain a reasonable coarse mask. To best fairly compare MG Matting with other trimap-based methods, we test our model under two settings: 1) TrimapFG: We adopt the confident foreground regions in a trimap as a coarse guidance mask for our network; 2) Trimap: We normalize trimap to $[0, 1]$ with the unknown pixels being $0.5$ and use this soft mask as guidance. We follow the the evaluation setting in Composition-1k which only computes the evaluation on the unknown region.

We summarize the alpha results and foreground color results in Table 1 and Table 3 respectively. We note that although our model is not trained with trimap, it still shows great robustness and transferability on these unseen types of guidance. Our model surpasses previous state-of-the-art models by a large margin. It also performs consistently considering the gap between trimap and trimapFG. We also note that our foreground color prediction not only reduces the errors significantly, but also produces much smoother results (see Fig. 4), which is desired in complex real-world scenarios where alpha matte can be noisy.

**Testing on Distinction-646.** Distinction-646 [31] is a recent synthetic matting benchmark dataset, which improves the diversity of Composition-1k. It contains 1000 test samples obtained in a similar manner as Composition-1k. However, this dataset is released without official trimaps or other

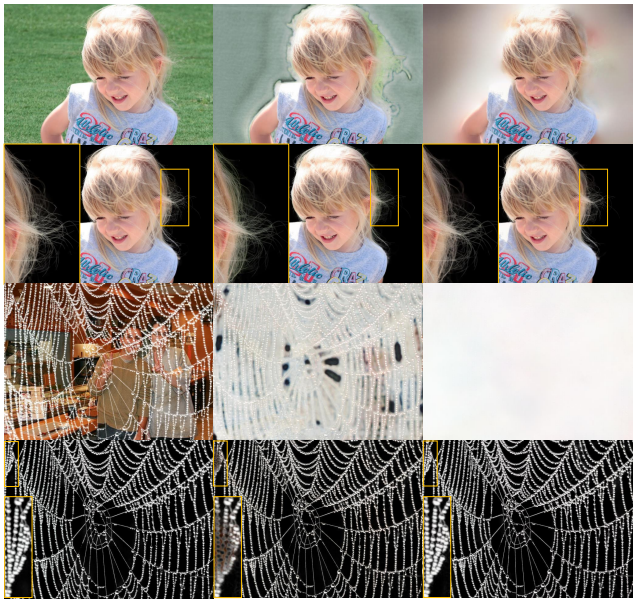

Figure 4: A visual comparison of foreground color decontamination. Each column from left to right: Input image and ground truth $\alpha \cdot F$, Foreground color prediction and $\alpha \cdot F$ of [18], predictions of our model with random alpha blending. Note that the background color is mixed into the prediction of [18], while our model can estimate a more smooth foreground color map and be more robust.

types of guidance, making it difficult to compare with previously reported results. Therefore, we use this benchmark mainly as a testbed to show how our method can refine a matte produced by another method.

We test a few state-of-the-art trimap-based baselines trained on Composition-1k. We firstly generate trimaps from ground-truth alpha mattes by thresholding and unknown region is dilated by kernel size 20. Then, we use these trimap-based methods to generate the matting results. Finally, we use these predicted alpha mattes as the guidance to our MG Matting method, and produce refined mattes.

As shown in Table 2, using the MG Matting as a refinement method consistently improves the results of other state-of-the-art methods. We also show the results reported by [31] in Table 2 for reference.

**Ablation Studies.** To validate the design of PRN and the introduced guidance perturbation, we conduct ablations studies as summarized in Table 4. Trimap is used as guidance masks in these experiments. However, we do not assume that the guidance type is known, so we purposefully do not use it to post-process the prediction by replacing the known foreground and background region. Instead, we report the two scores calculated over the whole image and the unknown region respectively for a more comprehensive evaluation of the robustness of our method.

We report ablations of different variants in Table 4. Base-

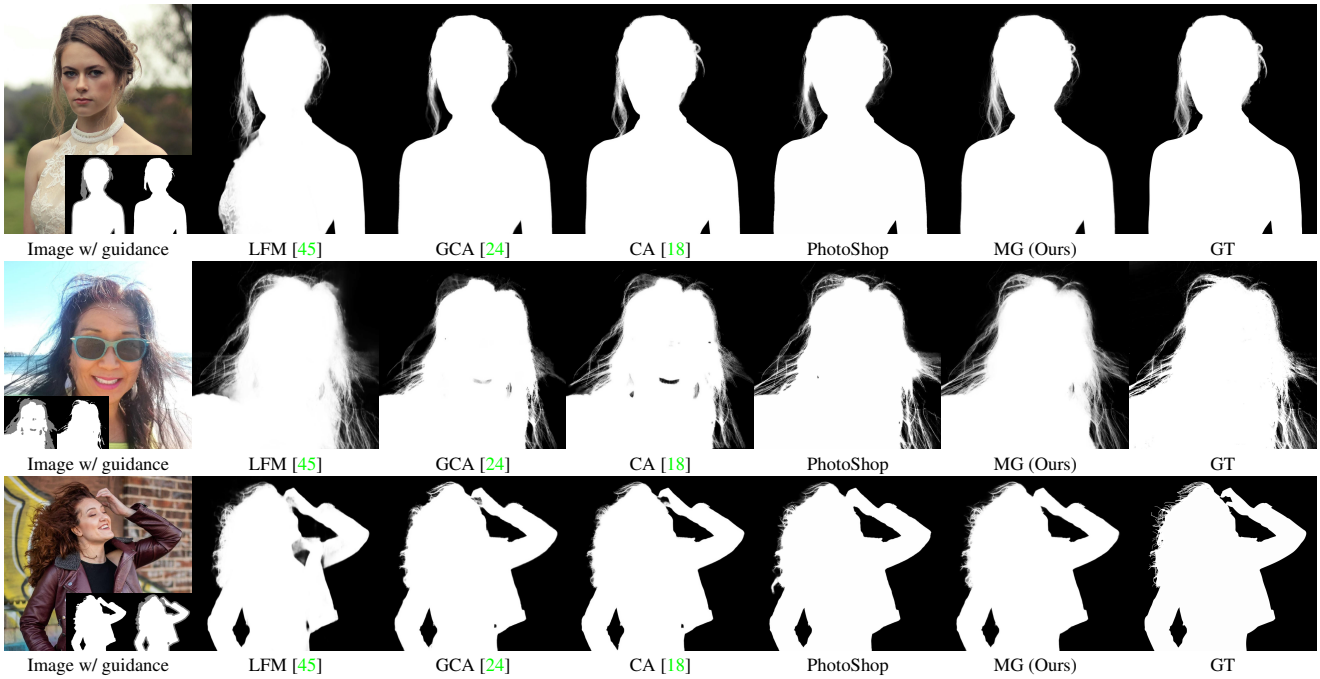


Figure 5: The visual comparison results among different methods on our portrait test set. We visualize representative examples with both high-quality studio-level portraits and selfies with strong noises. MG Mating performs well on different quality images and can maintain details. We note that our results, though only trained on composition-1k, are not only superior to previous state-of-the-art but also produces comparable or better results than commercial methods in PhotoShop.

| Methods | Whole Image | | Details | |
|---|---|---|---|---|
| | SAD | MSE ($10^{-3}$) | SAD | MSE ($10^{-3}$) |
| Deep Image Matting [41] | 28.5 | 11.7 | 19.1 | 74.6 |
| GCA Matting [24] | 29.2 | 12.7 | 19.7 | 82.3 |
| IndexNet Matting [29] | 28.5 | 11.5 | 18.8 | 72.7 |
| Context-Aware Matting [18] | 27.4 | 10.7 | 18.2 | 66.2 |
| Late Fusion Matting [45] | 78.6 | 39.8 | 24.2 | 88.3 |
| Ours | **26.8** | **9.3** | **17.4** | **55.1** |

Table 5: Results on Real-world Portrait test set.

line refers to a pure backbone without any add-ons. Adding side outputs and deep supervision to baseline improves the performance on both whole image or unknown area. We also try to use two convolution layers to fuse different outputs. However, linearly fusing the side outputs may not lead to better results. In contrast, the proposed PRN can better coordinate the semantic refinement and low-level detail refinement at different levels, thus obtaining a consistent improvement. We also show that the CutMask perturbation can further improve both the performance and robustness.

We also validate the effectiveness of RAB. We calculate the MSE and SAD of foreground color ($\mathbf{F}$) over foreground regions (*i.e.* $\alpha > 0$). The baseline achieves $\text{MSE} = 0.00623$ and $\text{SAD} = 82.30$, while with RAB, the performance is boosted to $\text{MSE} = 0.00321$ and $\text{SAD} = 62.01$.

## 5. Experiments on Real-world Portrait Dataset

We note that although the synthetic datasets are well-established benchmarks and provide sufficient data to train a good model, it remains an open question whether models trained on them are robust enough and can produce comparable results in real images. For example, [18] found that some easy data augmentations such as re-JPEGing and gaussian blur can avoid some shortcomings of the synthetic dataset and significantly improve the model's performance on real-world images, though at a cost of higher errors on the synthetic benchmark. This begs the question: *can the results on synthetic matting dataset reflect the performance on real images?*

Evaluation on real-world images is thus very crucial. However, due to the lack of high-quality matting benchmark datasets of real images, most previous models mainly compare their matting results visually or through user study. To better evaluate the matting methods in a real-world scenario, we collect a real-world image matting dataset consisting of 637 diverse and high-resolution images with matting annotation made by experts. The images in our dataset have various image quality and subjects of diverse poses. Moreover, since the dataset mainly contains solid objects where the main body can be easy to predicted, we also labeled detail masks covering the hair region and other soft tissues, which tells where the most important details of the

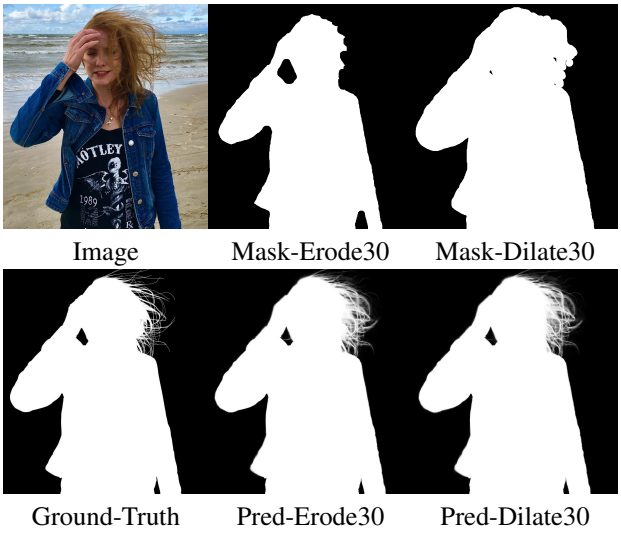


Figure 6: Our model is robust given different quality guidance masks and produces consistent alpha estimation.

image are located. By calculating errors in these regions, we can further compare the ability to capture object details for different models. We will release this dataset for better benchmarking matting methods on real images.

**Implementation Details.** We use the Composition-1k training set to train the model. Considering the semantic gap between the two datasets, we remove the transparent objects from the training data using the data list of [32]. Following [18], we also apply re-JEPGing, gaussian blur, and gaussian noises to the input image to make the model better adapt to real-world noises which are rarely seen in the synthetic dataset. Since these augmentations can change the color of the composited training image, thus the original color label may not be applicable. Therefore, we remove the composition loss from the supervision. Other training settings remain the same as in Sec. 4.

For trimap-based baselines, we follow [32] to generate trimaps from segmentation [44] automatically by labeling each pixel with foreground class probability $> 0.95$ as foreground, $< 0.05$ as background, and the rest as unknown, the unknown region is further dilated by $k = 20$ to ensure it will not miss the long hairs. For our model, we threshold the segmentation at prob $= 0.5$ to a binary mask.

**Results.** We compare the results with state-of-the-art trimap-based methods DIM [41], GCA [24], IndexNet [29], Context-Aware Matting [18], and trimap-free method Late Fusion Matting [45] which is trained on Composition-1k training set and an additional portrait dataset. The results of baselines are obtained through either the open-source inference demos or the provided pre-trained weights.

We summarize the results in Table 5 under two settings: Whole Image, where the errors are calculated across the whole image, which can measure the overall quality; Details, where the errors are calculated only in manual-labeled regions containing hair details or other soft areas.

Compared to other methods, our model achieves a superior performance, especially regarding to the detail part, which illustrates its ability to capture the boundary details. We also note that the trimap-free method LFM performs badly, which could be caused by the fact that their portrait training data is not diverse enough and thus limits the the generalizability of their model (see Fig. 5 for examples).

We compare our results with another trimap-free method BSHM [27]. We contacted the authors and obtained the test results on a 100 images subset of our portrait dataset. Since [27] can only deal with low-resolution images, we downsample images to longer-side 720, and the metrics are also computed on this scale. [27] achieves MSE $0.0155$ and SAD $10.66$ for whole image and MSE $0.0910$ and SAD $7.60$ for detail regions, while our MG Matting obtains a superior performance with MSE $0.0095$ and SAD $8.01$ for whole image and MSE $0.0637$ and SAD $5.94$ for details.

**Robustness to Guidance.** To verify how robust our model is to the external guidance mask, we conduct an experiments to feed the network with perturbed external guidance mask. Particularly, we erode/dilate the mask with kernel size $10$, $20$, $30$ respectively. We note that the model predict consistently given differently perturbed external guidance. The SAD error increases from $26.8$ to $27.1$, $27.2$, $27.4$ with mask eroded by $10$, $20$, and $30$ respectively. For dilation, the SAD error goes to $27.0$, $27.4$, $28.1$ with kernel $10$, $20$, $30$ respectively. A visual example is provided in Fig. 6.

## 6. Conclusion

In this paper, we present Mask Guided (MG) Matting, a general framework to resolve the natural image matting problem. Unlike previous methods, our method is not tailored to some specific guidance mask. Instead, it can handle versatile guidance masks such as a trimap, a rough segmentation mask, or a low-quality alpha matte. The key of the robustness of our model lies in the Progressive Refinement Network, which provides self-guidance and progressively refine the uncertain regions during the decoding process. Further, we also propose a simple yet effective method called Random Rendering to resolve the limitation of existing dataset and learn a better foreground color estimation model, which is important yet rarely studied before. Moreover, we release a new real-world matting dataset with high-quality label to better quantitatively evaluate matting models in a real-world scenario, which we hope could shed some light on the direction towards a real-life matting.

## References


[1] Yağiz Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. Semantic soft segmentation. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018.

[2] Yutong Bai, Qing Liu, Lingxi Xie, Weichao Qiu, Yan Zheng, and Alan L Yuille. Semantic part detection via matching: Learning to generalize to novel viewpoints from limited training data. In *ICCV*, pages 7535–7545, 2019.

[3] Yutong Bai, Angtian Wang, Adam Kortylewski, and Alan Yuille. Coke: Localized contrastive learning for robust keypoint detection. *arXiv preprint arXiv:2009.14115*, 2020.

[4] Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiaming Liu, Jiaying Liu, Jue Wang, and Jian Sun. Disentangled image matting. In *ICCV*, pages 8819–8828, 2019.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017.

[6] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *TPAMI*, 35(9):2175–2188, 2013.

[7] Yung-Yu Chuang, Brian Curless, David H Salesin, and Richard Szeliski. A bayesian approach to digital matting. In *CVPR*, volume 2, pages II–II. IEEE, 2001.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

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

[10] Marco Forte and François Pitié. $f$, $b$, alpha matting. *arXiv preprint arXiv:2003.07711*, 2020.

[11] 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.

[12] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[13] Vikas Gupta and Shanmuganathan Raman. Automatic trimap generation for image matting. In *2016 International Conference on Signal and Information Processing (IConSIP)*, pages 1–5. IEEE, 2016.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

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

[16] Kaiming He, Jian Sun, and Xiaoou Tang. Fast matting using large kernel matting laplacian matrices. In *CVPR*, pages 2165–2172. IEEE, 2010.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *ICCV*, pages 4130–4139, 2019.

[19] Chang-Lin Hsieh and Ming-Sui Lee. Automatic trimap generation for digital image matting. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–5. IEEE, 2013.

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

[21] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *TPAMI*, 30(2):228–242, 2007.

[22] Anat Levin, Alex Rav-Acha, and Dani Lischinski. Spectral matting. *TPAMI*, 30(10):1699–1712, 2008.

[23] Yingwei Li, Xiaojie Jin, Jieru Mei, Xiaochen Lian, Linjie Yang, Cihang Xie, Qihang Yu, Yuyin Zhou, Song Bai, and Alan L Yuille. Neural architecture search for lightweight non-local networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10297–10306, 2020.

[24] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *AAAI*, volume 34, pages 11450–11457, 2020.

[25] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and Cihang Xie. Shape-texture debiased neural network training. *ICLR*, 2021.

[26] 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 *ECCV*, pages 740–755. Springer, 2014.

[27] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-sheng Hua. Boosting semantic human matting with coarse annotations. In *CVPR*, pages 8563–8572, 2020.

[28] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *ICLR*, 2017.

[29] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *ICCV*, pages 3266–3275, 2019.

[30] Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. Alphagan: Generative adversarial networks for natural image matting. *BMVC*, 2018.

[31] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *CVPR*, pages 13676–13685, 2020.

[32] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *CVPR*, pages 2291–2300, 2020.

[33] Ehsan Shahrian, Deepu Rajan, Brian Price, and Scott Cohen. Improving image matting using comprehensive sampling sets. In *CVPR*, pages 636–643, 2013.

[34] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *ECCV*, pages 92–107. Springer, 2016.

[35] Jian Sun, Jiaya Jia, Chi-Keung Tang, and Heung-Yeung Shum. Poisson matting. In *SIGGRAPH*, pages 315–321. 2004.

[36] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019.

[37] Jingwei Tang, Yagiz Aksoy, Cengiz Oztireli, Markus Gross, and Tunc Ozan Aydin. Learning-based sampling for natural image matting. In *CVPR*, pages 3055–3063, 2019.

[38] Jue Wang and Michael F Cohen. Optimized color sampling for robust matting. In *CVPR*, pages 1–8. IEEE, 2007.

[39] Jue Wang and Michael F Cohen. *Image and video matting: a survey*. Now Publishers Inc, 2008.

[40] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, pages 1395–1403, 2015.

[41] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *CVPR*, pages 2970–2979, 2017.

[42] Qihang Yu, Dong Yang, Holger Roth, Yutong Bai, Yixiao Zhang, Alan L Yuille, and Daguang Xu. C2fnas: Coarse-to-fine neural architecture search for 3d medical image segmentation. In *CVPR*, pages 4126–4135, 2020.

[43] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019.

[44] He Zhang, Jianming Zhang, Federico Perazzi, Zhe Lin, and Vishal M Patel. Deep image compositing. *WACV*, 2021.

[45] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *CVPR*, pages 7469–7478, 2019.

[46] Yuanjie Zheng and Chandra Kambhamettu. Learning based digital matting. In *ICCV*, pages 889–896. IEEE, 2009.