Matting by Generation

Zhixiang Wang
The University of Tokyo
Tokyo, Japan
wangzx@nii.ac.jp

Baiang Li
Hefei University of Technology
Hefei, China
ztmotalee@gmail.com

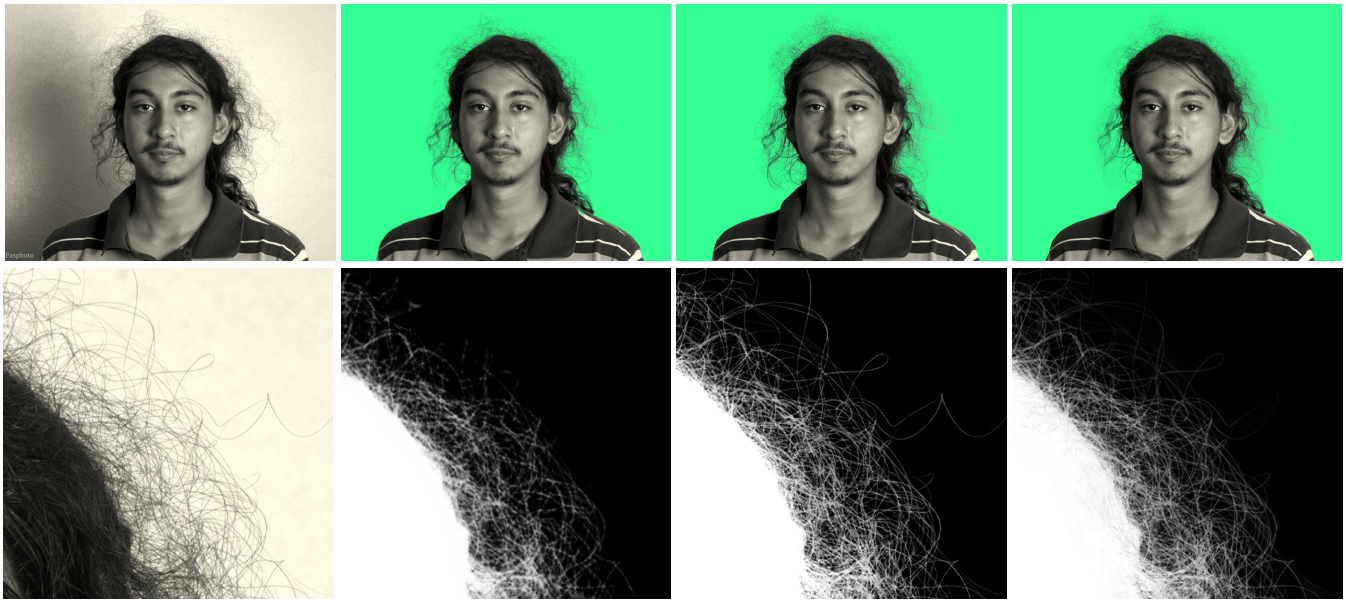
Jian Wang†
Snap Research
New York, USA
jwang4@snap.com

Yu-Lun Liu
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
yulunliu@cs.nycu.edu.tw

Jinwei Gu
The Chinese University of Hong Kong
Hong Kong SAR
jwgu@cuhk.edu.hk

Yung-Yu Chuang
National Taiwan University
Taipei, Taiwan
cyy@csie.ntu.edu.tw

Shin'ichi Satoh
National Institute of Informatics
Tokyo, Japan
satoh@nii.ac.jp



Input | ViTAE-S [Ma et al. 2023] | Ours | Human annotation

Figure 1: Matting by Generation. We crack the trimap-free matting problem in a conditional generative way as opposed to the previous regression-based fashion. With only an image as input, our method automatically extracts the foreground (*e.g.*, person) and generates high-quality boundary details benefiting from the rich generative prior, leading to photorealistic compositions. Compared with the human annotation, our results provide crisper details and greater fidelity to the input image in this example.

ABSTRACT

This paper introduces an innovative approach for image matting that redefines the traditional regression-based task as a generative modeling challenge. Our method harnesses the capabilities of latent diffusion models, enriched with extensive pre-trained knowledge, to regularize the matting process. We present novel architectural innovations that empower our model to produce mattes with superior resolution and detail. The proposed method is versatile and can perform both guidance-free and guidance-based image matting, accommodating a variety of additional cues. Our comprehensive evaluation across three benchmark datasets demonstrates the superior performance of our approach, both quantitatively and qualitatively. The results not only reflect our method's robust effectiveness but also highlight its ability to generate visually compelling mattes that approach photorealistic quality. The code for this paper is available at https://github.com/lightChaserX/alphaLDM.

†Corresponding author

SIGGRAPH Conference Papers '24, July 27–August 01, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0525-0/24/07
https://doi.org/10.1145/3641519.3657519

CCS CONCEPTS

• **Computing methodologies** → **Image manipulation**; **Machine learning approaches**;

KEYWORDS

Diffusion models, image matting

ACM Reference Format:
Zhixiang Wang, Baiang Li, Jian Wang, Yu-Lun Liu, Jinwei Gu, Yung-Yu Chuang, and Shin'ichi Satoh. 2024. Matting by Generation . In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27–August 01, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3641519.3657519

1 INTRODUCTION

Image matting as a fundamental problem in computer vision has been investigated for decades [Li et al. 2023d]. It enables many real applications, such as visual effects synthesis [Li et al. 2022a], image editing [Kawar et al. 2023], *etc.* Its goal is to predict the foreground and the alpha matte from an input image. This is a highly ill-posed inverse problem with only the input being known. The forward model is the composition equation [Porter and Duff 1984] given by:

$$C = \alpha F + (1 - \alpha)\, B\,, \tag{1}$$

where C is the input, F is the foreground, B is the background, and $\alpha \in [0, 1]$ is the linear combination coefficient. The main challenge lies in the ill-posedness, which is a mixed difficulty — to find where the foreground is and what the opacity value is in the boundary.

Existing methods, regardless of traditional or learning-based, leverage additional inputs to reduce the ill-posedness. For example, one could mitigate unknown parameter B by capturing another background image [Lin et al. 2021; Sengupta et al. 2020], or could add priors for α through user annotated trimaps.[1] Besides using additional input provided by humans, methods [Li et al. 2023a; Yu et al. 2021] employing rough masks from other algorithms, such as Segment Anything (SAM) [Kirillov et al. 2023a], aim to alleviate the training burden in segmentation and enhance focus on boundary matte quality. However, these approaches are not entirely satisfactory, primarily due to their reliance on the segmentation network's accuracy. Imprecise initial segmentation can significantly compromise the quality of matting results, particularly at the boundaries. This dependency raises concerns about the efficacy of solely relying on rough segmentation masks for achieving high-quality matting.

Recent advancements in end-to-end matting methods [Ke et al. 2022; Li et al. 2021] have sought to address these limitations by eliminating the need for these additional inputs, thereby reducing the reliance on human-generated data. Nevertheless, developing an effective end-to-end matting algorithm from scratch poses significant challenges due to the task's inherent complexity. These methods typically employ strategies such as constraining the application domain to portrait images [Li et al. 2021; Ma et al. 2023] and imposing implicit segmentation prior [Ke et al. 2022]. While these approaches reduce ambiguity between segmentation and matting and encourage the model to capture boundary details more effectively, achieving high-quality boundary mattes remains challenging, as shown in Figure 1. The prevailing issue with existing matting approaches lies in their handling of boundary regions, which are often challenging due to factors such as *low visibility* (contrast, image quality) and *imperfect* human annotations[2]. These limitations can result in unnatural compositions, highlighting the need for more sophisticated solutions.

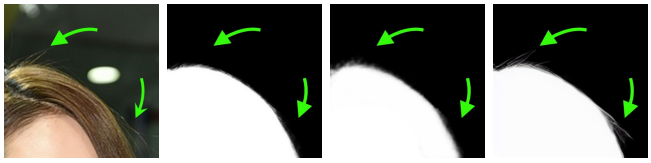


Figure 2: Imperfect human annotation. The training data are usually either blurry or lacking in some details. Therefore, the regression-based model would overfit the imperfect ground truth.

In this paper, we propose a simple yet effective technique of matting by generation. We transform the traditional *regression* problem into a conditional *generative* modeling problem, leveraging a diffusion model enriched with pre-trained knowledge about image semantics and matte details. There are several key advantages to this approach. Firstly, the generative model is adept at handling the uncertainties inherent in data, enabling it to learn the matte distribution more effectively than regression models. It also allows us to mitigate the negative impact of imperfect labels, such as ground-truth (GT) mattes generated by either humans or machines. These GT mattes, often derived from low-level image features, tend to contain imperfections, as exemplified in Figure 1. Utilizing such flawed mattes to train a regression-based model can lead to overfitting and suboptimal outcomes, as demonstrated in Figure 2. In contrast, our generative prior empowers our method to identify semantically correct boundaries and even generate results surpassing the GT mattes' quality. Secondly, our pre-trained diffusion model, with its vast database of billions of images, captures a more comprehensive image distribution. This broader understanding aids in regularizing the training process, offering more detailed and low-level property insights. Thus, the generative capabilities of our model shine in scenarios where image visibility is limited. Finally, our method offers versatility, accommodating both guidance-free and guidance-based matting. In most instances, it can perform accurate matting without additional hints. Nevertheless, in cases where the foreground is ambiguous, users can provide supplementary guidance to extract the desired matte.

List of Contributions. Our research makes the following three significant contributions:

- We convert the regression problem into a generative modeling problem, utilizing generative diffusion prior to regularize training effectively.

[1] Even with a known background image B, the problem is still ill-posed (3 unknowns for foreground color, plus unknown alpha, for 3 equations across R, G, B channels.)

[2] Although there are solutions for capturing ground-truth alpha matte [Smirnov et al. 2023], they often involve hardware requirements and are hard to scale up.

- We develop a model capable of processing high-resolution inputs efficiently and effectively.
- Our model is versatile and capable of handling scenarios with a variety of hints, including trimaps, masks, texts, and no hints at all.

2 RELATED WORK

Guidance-based Matting. Given only a single image, matting is an ill-posed inverse problem. Therefore, some matting methods require additional guidance, such as trimaps [Chuang et al. 2001], scribles [Levin et al. 2008], and clicks [Wei et al. 2021]. Methods of this category are typically referred to as guidance-based or trimap-based methods. Conventional trimap-based matting methods can be roughly divided into two categories: sampling-based methods [Chuang et al. 2001; Feng et al. 2016; Gastal and Oliveira 2010; He et al. 2011; Shahrian et al. 2013; Wang and Cohen 2007; Yang et al. 2018] and propagation-based methods [Aksoy et al. 2017; Chen et al. 2013; Grady et al. 2005; Levin et al. 2008; Sun et al. 2004; Wang and Cohen 2007]. Sampling-based methods usually resolve matting on a pixel-by-pixel basis by collecting color samples and forming a probabilistic distribution for each pixel's neighborhood. In contrast, propagation-based methods aim to obtain the matte for the entire image at once by establishing pixel affinities and solving an equation. Complex scenes often pose a challenge to these methods. In recent years, deep learning has been introduced to solve the matting problem and gained success [Cho et al. 2019; Liu et al. 2021; Lu et al. 2019a,b; Sun et al. 2021; Xu et al. 2017]. For example, Mask-guided matting [Yu et al. 2021] takes a general coarse mask as guidance, and proposes a Progressive Refinement Network module to achieve robust guidance. Matting Anything [Li et al. 2023a] leverages the recent Segment Anything Model (SAM), and further proposes a model that can estimate the alpha matte of any target instance with prompt-based user guidance in an image.

Guidance-free Matting. Given the considerable expense associated with acquiring additional guidance, efforts have been made to conduct matting without them, especially for specific foreground scenarios like portraits. These approaches are commonly referred to as guidance-free or trimap-free methods. Example methods of this type include SHM [Chen et al. 2018], SHMC [Liu et al. 2020], HATT [Qiao et al. 2020] and GFM [Li et al. 2022b]. MODNet [Ke et al. 2022] performs portrait matting by optimizing a series of sub-objectives simultaneously via explicit constraints. DugMatting [Wu et al. 2023] explores the explicitly decomposed uncertainties to efficiently and effectively improve matting. P3M-Net [Ma et al. 2023] specifically models the interactions between encoders and decoders to perform privacy-preserving portrait matting better.

Diffusion Models. Our approach builds upon the diffusion model [Ho et al. 2020; Song et al. 2021], a generative model that has garnered significant attention owing to its exceptional generative capabilities [Rombach et al. 2022]. Diffusion models have also demonstrated remarkable results in text-based image editing tasks, including InstructPix2Pix [Brooks et al. 2023], Imagic [Kawar et al. 2023], and SINE [Zhang et al. 2023]. In addition, it has been successfully used for various tasks [Fei et al. 2023; Xia et al. 2023] such as super-resolution [Yue et al. 2023], inpainting [Tang et al. 2023], segmentation [Burgert et al. 2023; Xu et al. 2023b]. Our work represents a pioneering effort in applying diffusion models to image matting. Compared with the recent concurrent unpublished diffusion-based methods for image matting [Hu et al. 2023; Xu et al. 2023a], the main difference with our work is the setting. Those methods are trimap-based, while our method facilitates both trimap-free and guidance-based matting. In addition, they are based on the pixel diffusion model, whereas we employ the latent diffusion model (LDM) [Rombach et al. 2022]. The LDM pre-trained on billions of images offers powerful prior. Furthermore, the latent mechanism helps mitigate the impact of potentially imperfect training data, as shown in Figure 1 and Figure 2.

3 METHOD

We solve the matting problem in a conditional generation manner by training a diffusion model to jointly model the distribution of alpha matte $p(\boldsymbol{\alpha})$ and draw an alpha matte $\boldsymbol{\alpha}$ from the distribution conditioned on the input image $\mathbf{x}$. Thanks to its generative ability and pre-trained rich image knowledge, our model can find the foreground and generate alpha matte with fine boundary details without guidance (Section 3.2). Our tailored high-resolution inference enables the process of arbitrary-resolution images (Section 3.3). Besides guidance-free matting, we can seamlessly integrate additional guidance to our trained model, such as a trimap, coarse mask, scribbles, and texts, to alleviate ambiguity in matting (Section 3.4).

3.1 Generative Formulation

We model the distribution of alpha matte $p(\boldsymbol{\alpha})$ with a pre-trained latent diffusion model [Rombach et al. 2022]. Given an alpha matte $\boldsymbol{\alpha} \sim p(\boldsymbol{\alpha})$, we encode it with the pre-trained encoder $\mathcal{E}$ to get its latent representation $\mathbf{z}^{(\boldsymbol{\alpha})} = \mathcal{E}(\boldsymbol{\alpha})$. We then apply the diffusion process to the latent representation. Let $\mathbf{z}_0 \coloneqq \mathbf{z}^{(\boldsymbol{\alpha})}$, the *forward* process gradually adds a small amount of Gaussian noises to the latent of alpha matte $\mathbf{z}_0$ in T steps. Therefore, a discrete Markov chain $\{\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_T\}$ is constructed such that

$$\mathbf{z}_t = \sqrt{1-\beta_t}\mathbf{z}_{t-1}^{(\boldsymbol{\alpha})} + \sqrt{\beta_t}\boldsymbol{\epsilon}_{t-1} = \sqrt{\sigma_t}\mathbf{z}_0 + \sqrt{1-\sigma_t}\boldsymbol{\epsilon}\,, \tag{2}$$

where the step $t \in \{1, \ldots T\}$, $\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are Gaussian noises and $\sigma_t \coloneqq \prod_{s=1}^{t} \beta_s$. The variance schedule $\{\beta_1, \ldots, \beta_T\}$ enables multiple scales of Gaussian noises added to $\mathbf{z}_0$.

To model the distribution of $\mathbf{z}^{(\boldsymbol{\alpha})}$, the *backward* process trains a score-based model ϵ_θ to predict the noise introduced to the noisy sample $\mathbf{z}_t$ at step t. The objective of training is to minimize

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t,\mathbf{z}_0}\left[\|\epsilon_t - \epsilon_\theta(\mathbf{z}_t, t)\|_2^2\right]. \tag{3}$$

Training the model on a set of alpha mattes $\{\boldsymbol{\alpha}_i\}_{i=1}^{N} \sim p(\boldsymbol{\alpha})$ enables modeling their distribution $p(\boldsymbol{\alpha})$. After training, we can perform ancestral sampling [Song et al. 2020] to generate a sample $\mathbf{z}_0$ from a normally distributed variable $\mathbf{z}_T \in \mathcal{N}(\mathbf{0}, \mathbf{1})$. Subsequently, by passing $\mathbf{z}_0$ through the decoder $\mathcal{D}$, we obtain a matte $\hat{\boldsymbol{\alpha}} \sim p(\boldsymbol{\alpha})$.

3.2 Conditional Generation with a Single Input Image

The matting task aims to produce the alpha matte corresponding to a given input image $\mathbf{x}$, rather than generating a random alpha matte.

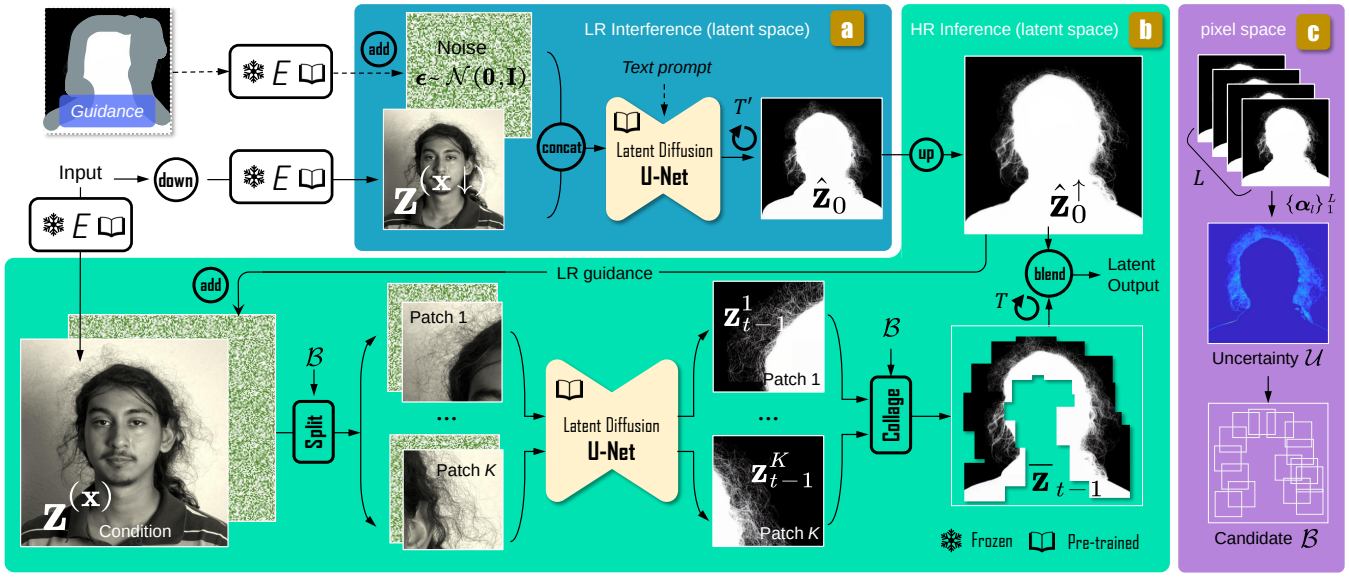


Figure 3: Method. (a) The low-resolution inference path can be used alone if we do not need very high-quality mattes or have a limited computational budget. The input is the low-resolution latent feature $\mathbf{z}^{(\mathbf{x}\downarrow)}$ of the down-sampled image $\mathbf{x}\downarrow$ and the sampled noise ϵ_t. If there is spatial guidance $c_\mathcal{S}$ present, we will combine it with the sampled noise as the noisy sample. If a text prompt $c_\mathcal{T}$ is provided, we will deliver it to the U-Net. The output of this path is the denoised latent feature $\hat{\mathbf{z}}_0$. This path requires a few steps $T' \sim 10$. (c) We run this step multiple times with different random seeds to get L predictions in the pixel space. With them, we estimate the uncertainty map $\mathcal{U}$, and the set of candidate regions $\mathcal{B} = \{b_i\}_1^B$. (b) The high-resolution path. We first add the up-sampled latent feature to the sampled noise. Then, we split the high-resolution latent input and noise into overlapped patches according to $\mathcal{B}$. These patches are respectively fed into the diffusion denoising network. Finally, we merge all denoised patches to get a collage. We perform "split" and "collage" during every denoising step $t \in \{1, \ldots, T\}$. We will use a specific text prompt: "enhance details" if there is a text prompt used in the LR path.

As a result, we condition the generation process on the input image $\mathbf{x}$. Specifically, we concatenate the latent $\mathbf{z}^{(\mathbf{x})} := \mathcal{E}(\mathbf{x})$ of the input image $\mathbf{x}$ with the noisy sample $\mathbf{z}_t$ and then feed the concatenated tensor to the model. To teach the model to generate alpha matte $\boldsymbol{\alpha}$ conditioned on the input image $\mathbf{x}$, we train it with paired data $\{(\mathbf{x}_i, \boldsymbol{\alpha}_i)\}_{i=1}^N \in p_{\text{data}}$ by minimizing:

$$\mathcal{L} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t,\mathbf{z}_0,\mathbf{z}^{(\mathbf{x})}}\left[\|\epsilon_t - \epsilon_\theta(\mathbf{z}_t, \mathbf{z}^{(\mathbf{x})}, t)\|_2^2\right]. \tag{4}$$

We initialize model ϵ_θ with pre-trained weights from Stable Diffusion (SD) [Rombach et al. 2022]. The weights learned on billion-level natural images [Schuhmann et al. 2022] possess extensive knowledge of image semantics and details. To adapt the denoising score-based model ϵ_θ for alpha matte generation, we extend its architecture by duplicating its input layers. The weights of these newly added layers are initialized to 0. Following this modification, we proceed to fine-tune the denoising score-based model. Upon completing the training process, we can draw a sample $\hat{\boldsymbol{\alpha}}$ from $p(\boldsymbol{\alpha})$ conditioned on the input image $\mathbf{x}$ with the ancestral sampling and decoding.

3.3 HR Inference with LR Guidance

The current image resolution is typically high, often exceeding 2K. Applying the diffusion model to such high-resolution (HR) inputs requires computational resources that are not readily available. Inference with low-resolution (LR) images is sub-optimal for generating mattes with detailed boundaries. To address this issue, we propose an HR inference method leveraging patch-based inference. However, applying patch-based inference, like MultiDiffusion [Bar-Tal et al. 2023], to high-resolution matting presents two major challenges: the lack of context and redundant computations. To overcome the issue of limited context, we use a predicted low-resolution matte to guide the process. For reducing computational load, we take advantage of the sparsity inherent in alpha mattes.

Patch Sampling. The diffusion model produces stochastic alpha mattes under different random seeds. However, stochastic results generally occur at the boundary regions, where the matte quality is inadequate, while other regions are deterministic and their matte quality is good enough. We pay attention to these boundary regions, which represent small portions of the input. Other regions can be directly determined through up-sampling the matte from LR inference. Taking advantage of the sparsity of fractional alpha values in the matte, we can reduce computations while maintaining good quality. First, we downsample the HR image and pass it through the diffusion model, yielding a low-resolution matte prediction $\hat{\boldsymbol{\alpha}}_i$. We perform the low-resolution inference L times using varying random seeds, resulting in L low-resolution predictions

$\mathcal{A} = \{\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2, \ldots, \hat{\boldsymbol{\alpha}}_L\}$. Their standard deviation is calculated to approximate the uncertainty map $\mathcal{U} = \sqrt{\mathbb{E}(\mathcal{A} - \mathbb{E}(\mathcal{A}))}$. We identify regions on the uncertainty map $\mathcal{U}$ with high information entropy as candidate patches $\mathcal{B} = \{b_i\}_{i=1}^{B}$ that require refinement. As depicted in Figure 3, high uncertainty is often observed around complex regions, such as hair boundaries. Thus, based on this uncertainty map, we select candidate regions for further processing.

Patch-Based Inference. We perform inference on the selected patches $\mathcal{B}$. The noise latent for each patch is not independently sampled. This strategy ensures consistent prediction for different patches, especially for overlapped areas. We sample a noise $\mathbf{z}_T \in \mathcal{N}(\mathbf{0}, \mathbf{1})$ with the same size as the input image. Then, we crop patches $\{\mathbf{z}_T^1, \mathbf{z}_T^2, \ldots, \mathbf{z}_T^K\}$ from the sampled noise $\mathbf{z}_T$, where $\mathbf{z}_T^k = F(\mathbf{z}^T | b_i)$ and F denotes the cropping operator. The image condition used to condition the model is cropped from the input image's latent similarly. We feed the noise and image latent of the patch to the diffusion model. During the ancestral sampling, each step t will produce latent samples $\{\mathbf{z}_t^1, \mathbf{z}_t^2, \ldots, \mathbf{z}_t^K\}$ for patches $\{b_1, b_2, \ldots, b_K\}$. Before passing them to the next step $t-1$, we merge them by

$$\bar{\mathbf{z}}_t = \sum_{k=1}^{K} F^{-1}(\mathbf{z}_t^k | b_k), \tag{5}$$

where F^{-1} is the uncropping operator which puts the latent patch $\mathbf{z}_t^k$ back to the patch location b_k where it was cropped from the input image. Then, we get the latent patches for the next step from $\bar{\mathbf{z}}_t$. After the ancestral sampling, we will obtain the denoised latent $\bar{\mathbf{z}}_0$. It is finally merged with the up-sampled LR matte on latent space to get the final alpha matte. The coarse-to-fine strategy is similar to previous high-resolution matting methods [Lin et al. 2021]; however, the main difference is the proposed guidance mechanism for the diffusion model.

Guidance Mechanism. Performing the model on cropped patches often produces flawed results because the model cannot perceive the context information of the whole image and could be misled by local patches. To address this issue, we propose to use the predicted LR matte as guidance. Although the matte predicted from LR input has imperfect boundary details, it has sufficiently accurate predictions for other regions. Thus, instead of starting from pure noise $\boldsymbol{\epsilon} \in \mathcal{N}(\mathbf{0}, \mathbf{1})$, we start the backward process from

$$\mathbf{z}_T = \sqrt{(1-\sigma_T)/\sigma_T}\,\boldsymbol{\epsilon} + \hat{\mathbf{z}}_0^{\uparrow}, \tag{6}$$

where $\hat{\mathbf{z}}_0^{\uparrow}$ denotes the upsampled latent corresponding to one of the predicted LR mattes $\{\hat{\boldsymbol{\alpha}}_l\}_1^L$ This strategy is simple but effective. During training, $\mathbf{z}_T$ is a summation of the ground truth alpha matte latent $\mathbf{z}_0$ and Gaussian noise $\boldsymbol{\epsilon}$. $\mathbf{z}_0$ contains low-frequency (foreground and background) and high-frequency (boundary) information while ϵ_T is high-frequency. The noise will flood the high-frequency information in $\mathbf{z}_0$. In other words, $\mathbf{z}_T$ is approximately the combination of the low frequency of $\mathbf{z}_0$ and the high frequency of $\boldsymbol{\epsilon}$. The model learns to extract low-frequency data from noisy samples. During inference, the given $\hat{\mathbf{z}}_0^{\uparrow}$ also contains both low-frequency (foreground and background) and high-frequency (boundary) information. The high frequency, which could be inaccurate, is flooded, and the model can extract the correct low frequency. This strategy can also facilitate the incorporation of users' guidance in the next section.

3.4 Additional Guidance

Matting without *any* guidance could lead to ambiguity. For example, when there are multiple people in an image, it could be difficult to determine which one to extract. Additional guidance, such as a human-annotated trimap, a coarse mask derived from semantic segmentation, scribbles, clicks, and a text prompt, would be helpful in this case. Our method can incorporate additional guidance if present.

Text Guidance. Adding text guidance is relatively easy since we use the text-to-image generative diffusion model. We annotate the text description of training images with BLIP2 [Li et al. 2023b]. Each annotation describes the target foreground in the training image. Given the CLIP feature $c_{\mathcal{T}}$ of the text prompt $\mathcal{T}$, we use the cross-attention mechanism to inject the control into the denoising model. We train the denoising model with the annotated prompt by minimizing

$$\mathcal{L} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t,\mathbf{z}_0,\mathbf{z}^{(\mathbf{x})},c_{\mathcal{T}}}\left[\|\epsilon_t - \epsilon_\theta(\mathbf{z}_t, \mathbf{z}^{(\mathbf{x})}, c_{\mathcal{T}}, t)\|_2^2\right]. \tag{7}$$

During training, for small patches, we use a specific prompt "enhance details" instead of avoiding confusing the model.

Spatial Guidance. Spatial guidance like a trimap, coarse mask [Yu et al. 2021], and scribbles are more popular than text prompts for image matting. Inspired by the guidance mechanism described in Section 3.3, we use a similar method for injecting spatial guidance $\mathcal{S}$. We extract this guidance's latent representation $c_{\mathcal{S}}$ and a mask indicating unknown regions m_{unknown}. For coarse mask, $m_{\text{unknown}} = \mathbf{I}$ and for scribble m_{unknown} represents regions without scribbles. At inference, we perform ancestral sampling from

$$\mathbf{z}_T = \sqrt{(1-\sigma_T)/\sigma_T}\,\boldsymbol{\epsilon} + (1 - m_{\text{unknown}})\, c_{\mathcal{S}}\,. \tag{8}$$

We can apply various kinds of guidance[3] directly at the inference time without training with them.

4 EXPERIMENTS

4.1 Protocol

Dataset. We conduct experiments on real-world datasets rather than synthetic ones. We train our model on the training set of P3M-10K [Li et al. 2021], a dataset containing 9,421 high-resolution real-world face-blurred portrait images and human-annotated alpha mattes that are not perfectly accurate. We evaluate the performance on three benchmarks: P3M-P dataset containing 500 face blurred images. Each image has a corresponding trimap and alpha matte. They are validation sets from P3M-10K that share a similar distribution of alpha matte with the training set. PPM-100 [Ke et al. 2022] is a dataset with 100 high-resolution images with corresponding fine annotations. RVP [Yu et al. 2021] consists of 636 portraits with alpha mattes and coarse segmentation masks.

[3]The domain of the guide image should match that of the alpha matte.

Table 1: Quantitative results of trimap-free portrait matting. We compare our method with trimap-free portrait matting methods. †For the trimap-based method DiffMat, we provide a mask with all pixels labeled unknown. *We removed ambiguous samples from the dataset, 5 out of 100 from PPM and 20 out of 500 from P3M-P, which will be elaborated on in the supplementary.

	PPM*				P3M-P*			
	MSE ↓	MAD ↓	SAD ↓	Conn ↓	MSE ↓	MAD ↓	SAD ↓	Conn ↓
DiffMat† [Xu et al. 2023a]	522.1	594.9	5681.3	5623.9	510.2	582.7	999.1	989.0
MODNet [Ke et al. 2022]	4.5	10.1	96.0	81.1	11.3	17.4	29.9	26.6
P3M [Li et al. 2021]	5.8	9.6	93.3	96.1	2.7	5.1	8.8	8.3
ViTAE-S [Ma et al. 2023]	3.4	6.5	62.6	59.3	1.8	4.3	7.4	7.2
Ours	**2.5**	**6.3**	**56.9**	**54.0**	**1.6**	**4.1**	**7.1**	**6.8**

Table 2: Quantitative results of guidance-based portrait matting. We compare our method with guidance-based portrait matting methods. The guidance is a mask in the top portion of the table, while in the bottom portion, it is a trimap. †P3M-P does not provide the segmentation mask; therefore, we use coarse masks extracted from trimaps m as guidance ($m[m >= 0.5] = 1$). Although our scores are worse than some methods since the label is imperfect, our visual results are better. Note that P3M has the same distribution as the training set; therefore, it could not reflect overfitting the imperfect labels.

	RVP				P3M-P†			
	MSE ↓	MAD ↓	SAD ↓	Conn ↓	MSE ↓	MAD ↓	SAD ↓	Conn ↓
MAM [Li et al. 2023a]	20.7	36.3	48.5	44.6	7.9	13.4	23.1	20.5
MG-Mat [Yu et al. 2021]	**9.4**	20.7	29.2	**25.5**	5.7	12.8	22.0	18.4
DiffMat [Xu et al. 2023a]	16.6	32.5	44.4	41.2	45.0	49.5	84.4	84.5
Ours - mask	11.6	**19.6**	**26.7**	26.1	1.6	4.2	7.2	6.2
IndexNet [Lu et al. 2019b]	–	–	–	–	1.2	4.2	7.0	6.0
MatteFormer [Park et al. 2022]	–	–	–	–	1.4	4.1	7.1	6.5
DiffMat [Xu et al. 2023a]	–	–	–	–	**1.0**	**3.6**	**6.2**	**5.2**
Ours - trimap	–	–	–	–	1.6	4.0	6.9	6.0

Compared methods. We compare our approach against several state-of-the-art matting methods. **Guidance-free:** MODNet [Ke et al. 2022], P3M [Li et al. 2021], ViTAE-S [Ma et al. 2023]. These methods, except for MODNet[4], are trained on P3M-10K. **Trimap-based:** IndexNet [Lu et al. 2019b], MatteFormer [Park et al. 2022], and DiffMat [Xu et al. 2023a], which is a concurrent method using diffusion models for trimap-based matting. **Mask Guided Matting:** MAM [Li et al. 2023a], which incorporates SAM [Kirillov et al. 2023b] as its backbone, and MG-Mat [Yu et al. 2021].

Metrics. Evaluation metrics include the Sum of Absolute Differences (SAD), Mean Squared Error (MSE), Mean Absolute Difference (MAD), and Connectivity (Conn.) [Rhemann et al. 2009]. We apply all metrics on whole images. We scale the MAD and MSE values by a factor of 10^3.

[4]We utilize the publicly available checkpoints released by MODNet. It was trained on proprietary datasets. We attempted to align MODNet's training data with ours. However, this reproduction resulted in unsatisfactory outcomes.

4.2 Trimap-free Matting

Table 1 presents the quantitative results for trimap-free setting. Our method achieves the best scores in all metrics. It consistently delivers good results on three benchmarks, showcasing its robustness and versatility across diverse scenarios. Our approach outperforms established methods like MODNet, P3M, ViTAE-S, and MAM, particularly regarding accuracy and boundary detail handling. This is evident in the lower SAD, MSE, MAD, and improved Connectivity scores compared to the competing methods. These results highlight the effectiveness of our generative modeling approach and the use of pre-trained diffusion models in addressing the complexities of trimap-free matting, especially in challenging cases involving intricate details and varying image qualities.

Figure 4 presents the qualitative comparisons. The input shows a complex scene with a person that includes intricate details like hair and shoelaces, which are challenging for matting algorithms. As highlighted by the insets, MODNet and P3M outputs lack fine detail, particularly in the hair and feet regions. In contrast, the results from ViTAE-S, while quantitatively close to our method, visually lack the nuanced details that our approach captures. Our result is more similar to the ground truth, including a clear and precise boundary matte, which faithfully reproduces the fine details of the subject, such as individual hair strands and the shoe's silhouette. This is evident even in cases where quantitative scores are similar, showcasing the added benefit our approach brings in creating high-fidelity human boundary mattes.

Figure 9 showcases the visual results on the RVP dataset, focusing on a challenging scenario involving complex hair details against a sunset backdrop. The input image presents significant matting difficulty due to the intricate hair strands silhouetted against the varying tones of the sky. MODNet and P3M results display notable artifacts and fail to capture the finer hair details, as evidenced in the zoomed insets. ViTAE-S, although quantitatively competitive, visually lacks fidelity in reproducing the hair's fine structure, as the comparison with the ground truth reveals a less accurate matte. Our method, on the other hand, shows a remarkable capture of detail, closely mirroring the ground truth. The insets highlight our approach's capability to preserve the delicate strands of hair and the subtle nuances in the silhouette, which are critical for a realistic matting outcome. Despite their close quantitative scores, this visual comparison underscores the qualitative edge of our method over ViTAE-S, illustrating our approach's advanced ability to generate detailed human boundary mattes that are distinct and more aligned with the actual scene.

4.3 Guidance-based Matting

When the foreground is ambiguous, it is inherently challenging for trimap-free matting. In contrast, our method can also use guidance such as text prompt and coarse mask to reduce ambiguity (Figure 5).

Table 2 shows the quantitative comparisons for the guidance-based setting. Although the trimap-based method—DiffMat [Xu et al. 2023a]—performs better than our method in the trimap-based setting, it relies on the high quality of the trimaps. They would fail when using coarse guidance, such as masks from semantic segmentation. The mask-based method MAM [Li et al. 2023a] refines the mask generated from SAM [Kirillov et al. 2023a]. Benefiting

Figure 4: Visual results of trimap-free matting on PPM-100 [Ke et al. 2022]. Our method achieves more accurate matting results, especially around thin and detailed structures, compared to prior work. We extracted the foreground using the technique proposed by Germer *et al.* [Germer et al. 2021] and composited it onto a new background sampled from a public background database [Lin et al. 2021].

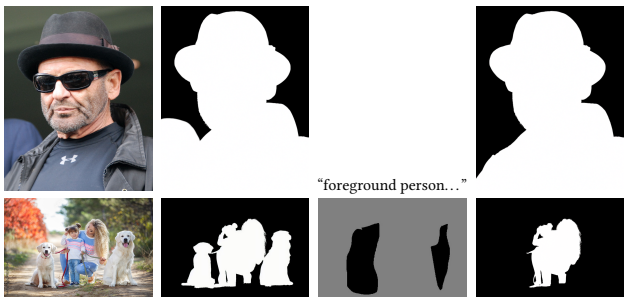


Figure 5: Use of guidance. With various guidance, we can reduce ambiguity.

Table 3: Ablation study. We implement four variants of our method and conduct the ablation study on PPM-100: (1) Our model without using the pre-trained SD weights; (2) Training with the same prompt for all cropped patches from an image; (3) Our model trained with resized full image rather than patches with different scales; (4) Adding pixel losses to our training phase.

	MSE	MAD
Ours	**2.5**	**6.3**
- (1) *w/o* denoising prior	63.6	71.1
- (2) *w/o* specific patch prompt	38.6	46.8
- (3) *w/o* multi-scale data	8.6	15.1
+ (4) pixel loss	58.0	65.1

from SAM, it can predict alpha mattes for "any" foreground objects. However, its matting performance is sub-optimal.

Figure 11 shows the qualitative comparison under the trimap-based setting. The scores of DiffMat [Xu et al. 2023a] are better, but we notice their visual results are worse than ours. At the same time, the imperfection of the ground-truth mattes is also observed. We suppose the worse results of DiffMat compared to us are because they are based on the pixel diffusion model, which overfits the training labels.

5 ANALYSES AND DISCUSSIONS

Ablation studies. Table 3 demonstrates the importance of the pre-trained generative prior, the prompting and the multi-scale training strategy (will be elaborated on in the supplementary), and operating

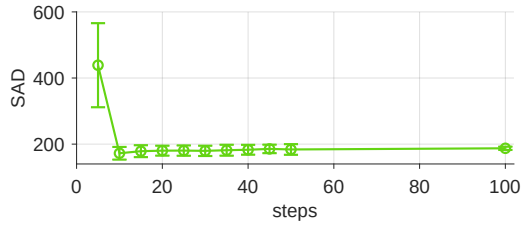


Figure 6: Randomness test. We use 5 different random seeds to test the model on selected images. With the increase of diffusion steps, the mean and std. of SAD error decrease.

Figure 7: HR inference with LR guidance.

in latent space. Stable Diffusion (SD) with vast pre-trained knowledge significantly induces semantic information. Table 3 indicates, without this, training the diffusion model is prone to lose semantic information, *e.g.* resulting in an incomplete person. Besides semantic information, this generative prior enables us to hallucinate details, *e.g.*, hair boundary. Without it, we could converge to the existing methods. Besides, operating within the latent space adeptly preserves essential details, crucial for producing high-quality mattes with emphasis on boundary regions. This is an advantage of our method over concurrent works built on pixel-based diffusion models. Figure 7 shows the effectiveness of HR inference with LR guidance.

Effect of randomness. Figure 6 depicts that with the increases of steps, the randomness decreases. Besides, we notice that infer with larger patches will also reduce the randomness.

Soft matte. Our model can produce soft mattes for out-of-focus blur, as shown in Figure 10, even though the training dataset does not contain annotations for such blur.

Limitations. Firstly, while our method reduces inference time for HR images compared to naive approaches, it is important to clarify that the inherent limitations of the diffusion model make it less efficient than prior regression-based methods. Processing a 512×512 images with 50 steps requires about 5s on a NVIDIA V100 GPU card. However, it is worth noting that some ongoing research efforts are focused on enhancing sampling efficiency, which could help mitigate this limitation [Li et al. 2023c; Song et al. 2023]. Secondly, our model trained on portrait datasets shows potential for adaptation to other domains, such as animal matting (see Figure 12), but is unsuitable for matting scenarios with markedly different characteristics, such as subjects like fire. Thirdly, our method is designed for image matting and cannot guarantee temporal consistency for videos (see Figure 8). Enhancing temporal coherence remains a subject for future research.

Figure 8: Video inference. By individually processing downsampled frames, our method produces temporal inconsistency in videos. While employing high-resolution frames mitigates this issue, it still suffers from problems similar to regression-based methods.

6 CONCLUSIONS

Our approach presents a straightforward yet highly efficient technique for matting. It can perform both trimap-free and guidance-based image matting tasks. By reframing the problem as a generative task and leveraging diffusion models enriched with pre-trained knowledge for regularization, we have devised innovative designs that empower our model to produce high-resolution and high-quality results. Our experimental results on three benchmark datasets not only demonstrate the efficacy of our method in quantitative terms but also showcase its exceptional visual performance, making it a promising solution for the field of matting.

ACKNOWLEDGMENTS

We thank anonymous reviewers for their constructive feedback. We also thank Yutong Dai and Zhanghan Ke for their help. This research was supported by NTU112L9009, NTU113L894003, NSTC112-2634-F-002-006, MOST110-2221-E-002-124-MY3, JSPS KAKENHI Grant Number JP23K24876, JST ASPIRE Program Grant Number JPMJAP2303, and the Value Exchange Engineering, a joint research project between Mercari, Inc. and RIISE.

REFERENCES

Yagiz Aksoy, Tunç Ozan Aydin, and Marc Pollefeys. 2017. Designing effective inter-pixel information flow for natural image matting. In *CVPR*.
Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. In *ICML*.
Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*.
Ryan Burgert, Kanchana Ranasinghe, Xiang Li, and Michael S Ryoo. 2023. Peekaboo: Text to image diffusion models are zero-shot segmentors. In *CVPRW*.
Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. 2018. Semantic human matting. In *ACM MM*.
Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. 2013. KNN matting. *IEEE TPAMI* 35, 9 (2013), 2175–2188.
Donghyeon Cho, Yu-Wing Tai, and In So Kweon. 2019. Deep Convolutional Neural Network for Natural Image Matting Using Initial Alpha Mattes. *IEEE TIP* 28, 3 (2019), 1054–1067.
Yung-Yu Chuang, Brian Curless, David H. Salesin, and Richard Szeliski. 2001. A Bayesian Approach to Digital Matting. In *CVPR*.
Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. 2023. Generative Diffusion Prior for Unified Image Restoration and Enhancement. In *CVPR*.
Xiaoxue Feng, Xiaohui Liang, and Zili Zhang. 2016. A Cluster Sampling Method for Image Matting via Sparse Coding. In *ECCV*.
Eduardo S. L. Gastal and Manuel M. Oliveira. 2010. Shared Sampling for Real-Time Alpha Matting. In *Eurographics*.

Thomas Germer, Tobias Uelwer, Stefan Conrad, and Stefan Harmeling. 2021. Fast multi-level foreground estimation. In *ICPR*.

Leo Grady, Thomas Schiwietz, Shmuel Aharon, and Rüdiger Westermann. 2005. Random walks for interactive alpha-matting. In *Proceedings of the IASTED International Conference on Visualization, Imaging and Image Processing*.

Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. 2011. A global sampling method for alpha matting. In *CVPR*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*.

Yihan Hu, Yiheng Lin, Wei Wang, Yao Zhao, Yunchao Wei, and Humphrey Shi. 2023. Diffusion for Natural Image Matting. *arXiv preprint arXiv:2312.05915* (2023).

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-Based Real Image Editing with Diffusion Models. In *CVPR*.

Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. 2022. MODNet: Real-time trimap-free portrait matting via objective decomposition. In *AAAI*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. 2023a. Segment Anything. In *ICCV*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023b. Segment Anything. In *ICCV*.

Anat Levin, Dani Lischinski, and Yair Weiss. 2008. A Closed-Form Solution to Natural Image Matting. *IEEE TPAMI* 30, 2 (2008), 228–242.

Jiachen Li, Jitesh Jain, and Humphrey Shi. 2023a. Matting Anything. *arXiv: 2306.05399* (2023).

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. 2021. Privacy-Preserving Portrait Matting. In *ACM MM*.

Jizhizi Li, Jing Zhang, Stephen J. Maybank, and Dacheng Tao. 2022b. Bridging composite and real: towards end-to-end deep image matting. *IJCV* 130, 2 (2022), 246–266.

Jizhizi Li, Jing Zhang, and Dacheng Tao. 2023d. Deep Image Matting: A Comprehensive Survey. *arXiv preprint arXiv:2304.04672* (2023).

Peizhuo Li, Kfir Aberman, Zihan Zhang, Rana Hanocka, and Olga Sorkine-Hornung. 2022a. GANimator: Neural Motion Synthesis from a Single Sequence. *ACM TOG* 41, 4 (2022), 138.

Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. 2023c. SnapFusion: Text-to-Image Diffusion Model on Mobile Devices within Two Seconds. In *NeurIPS*.

Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2021. Real-time high-resolution background matting. In *CVPR*.

Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian sheng Hua. 2020. Boosting semantic human matting with coarse annotations. In *CVPR*.

Yuhao Liu, Jiake Xie, Xiao Shi, Yu Qiao, Yujie Huang, Yong Tang, and Xin Yang. 2021. Tripartite information mining and integration for image matting. In *ICCV*.

Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. 2019a. Context-Aware Image Matting for Simultaneous Foreground and Alpha Estimation. In *ICCV*.

Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. 2019b. Indices matter: Learning to index for deep image matting. In *ICCV*.

Sihan Ma, Jizhizi Li, Jing Zhang, He Zhang, and Dacheng Tao. 2023. Rethinking Portrait Matting with Pirvacy Preserving. *IJCV* 131, 8 (2023), 2172–2197.

GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, and Nojun Kwak. 2022. Matteformer: Transformer-based image matting via prior-tokens. In *CVPR*.

Thomas Porter and Tom Duff. 1984. Compositing Digital Images. In *SIGGRAPH*.

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

Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. 2009. A perceptually motivated online benchmark for image matting. In *CVPR*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS* (2022).

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

Ehsan Shahrian, Deepu Rajan, Brian Price, and Scott Cohen. 2013. Improving image matting using comprehensive sampling sets. In *CVPR*.

Dmitriy Smirnov, Chloe LeGendre, Xueming Yu, and Paul Debevec. 2023. Magenta Green Screen: Spectrally Multiplexed Alpha Matting with Deep Colorization. In *Proceedings of the Digital Production Symposium*.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *ICML*.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency models. In *ICML*.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICML*.

Jian Sun, Jiaya Jia, Chi-Keung Tang, and Heung-Yeung Shum. 2004. Poisson matting. *ACM TOG* 23, 3 (2004), 315–321.

Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. 2021. Semantic image matting. In *CVPR*.

Luming Tang, Nataniel Ruiz, Chu Qinghao, Yuanzhen Li, Aleksander Holynski, David E Jacobs, Bharath Hariharan, Yael Pritch, Neal Wadhwa, Kfir Aberman, and Michael Rubinstein. 2023. RealFill: Reference-Driven Generation for Authentic Image Completion. *arXiv preprint arXiv:2309.16668* (2023).

Jue Wang and Michael F. Cohen. 2007. Optimized Color Sampling for Robust Matting. In *CVPR*.

Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Hanqing Zhao, Weiming Zhang, and Nenghai Yu. 2021. Improved Image Matting via Real-time User Clicks and Uncertainty Estimation. In *CVPR*.

Jiawei Wu, Changqing Zhang, Zuoyong Li, Huazhu Fu, Xi Peng, and Joey Tianyi Zhou. 2023. dugMatting: decomposed-uncertainty-guided matting. In *ICML*.

Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. 2023. DiffIR: Efficient Diffusion Model for Image Restoration. In *ICCV*.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. 2023b. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*.

Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. 2017. Designing effective inter-pixel information flow for natural image matting. In *CVPR*.

Yangyang Xu, Shengfeng He, Wenqi Shao, Kwan-Yee K Wong, Yu Qiao, and Ping Luo. 2023a. DiffusionMat: Alpha Matting as Sequential Refinement Learning. *arXiv preprint arXiv:2311.13535* (2023).

Xin Yang, Ke Xu, Shaozhe Chen, Shengfeng He, Baocai Yin Yin, and Rynson Lau. 2018. Active Matting. In *NeurIPS*.

Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. 2021. Mask guided matting via progressive refinement network. In *CVPR*.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. 2023. ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting. In *NeurIPS*.

Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris Metaxas, and Jian Ren. 2023. SINE: SINgle Image Editing with Text-to-Image Diffusion Models. In *CVPR*.

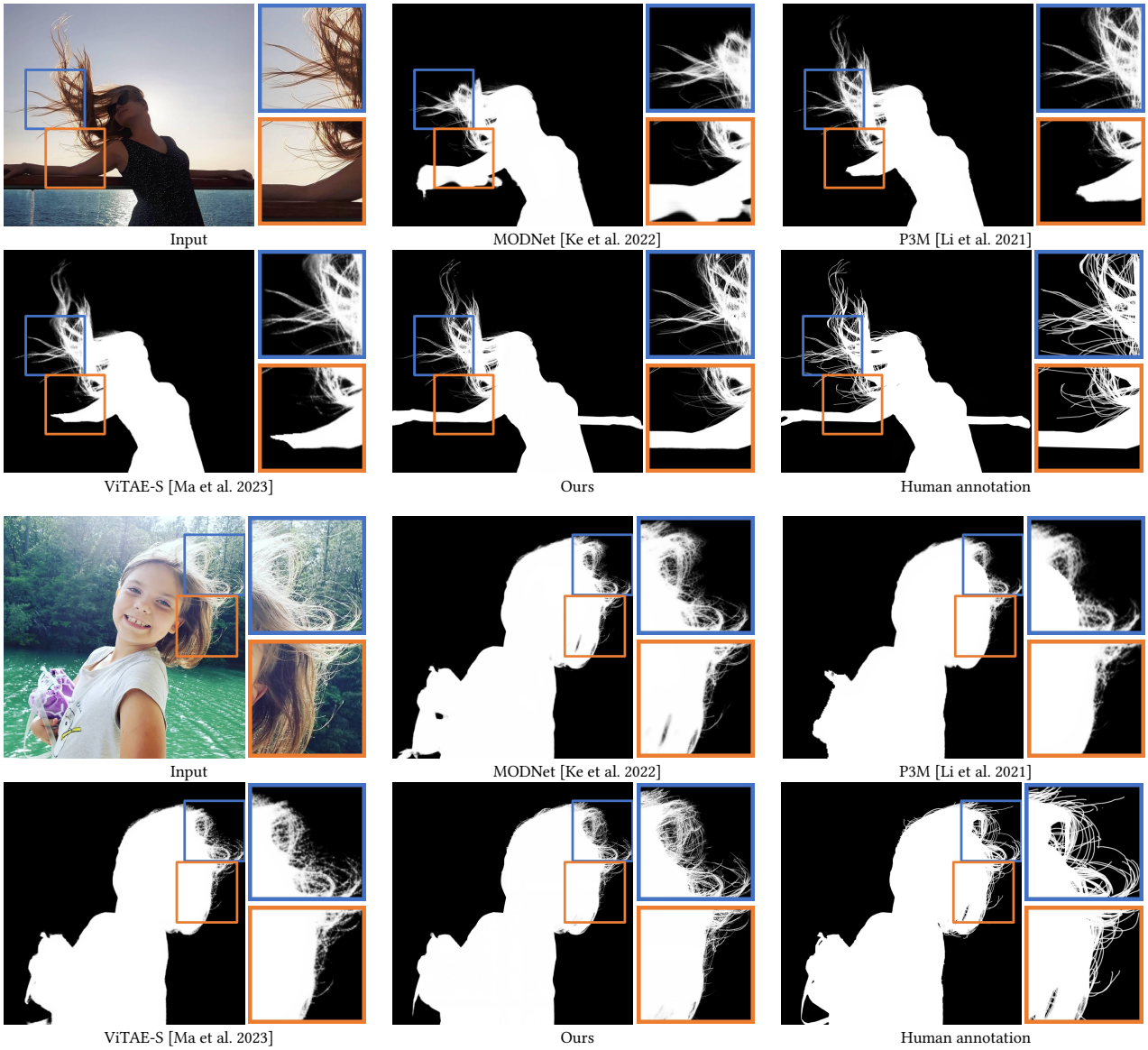


Figure 9: Visual results of guidance-free matting on RVP [Yu et al. 2021] dataset.

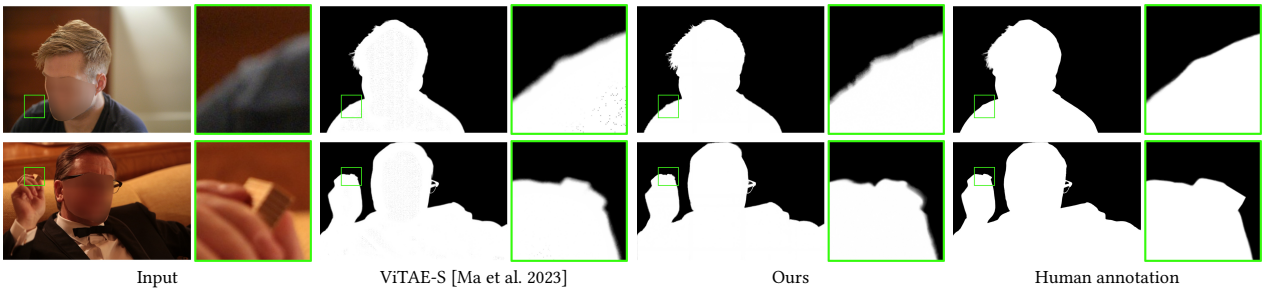


Figure 10: Matting with out-of-focus blur. Compared to the hard label in the out-of-focus regions of the human annotations, we generate soft mattes.

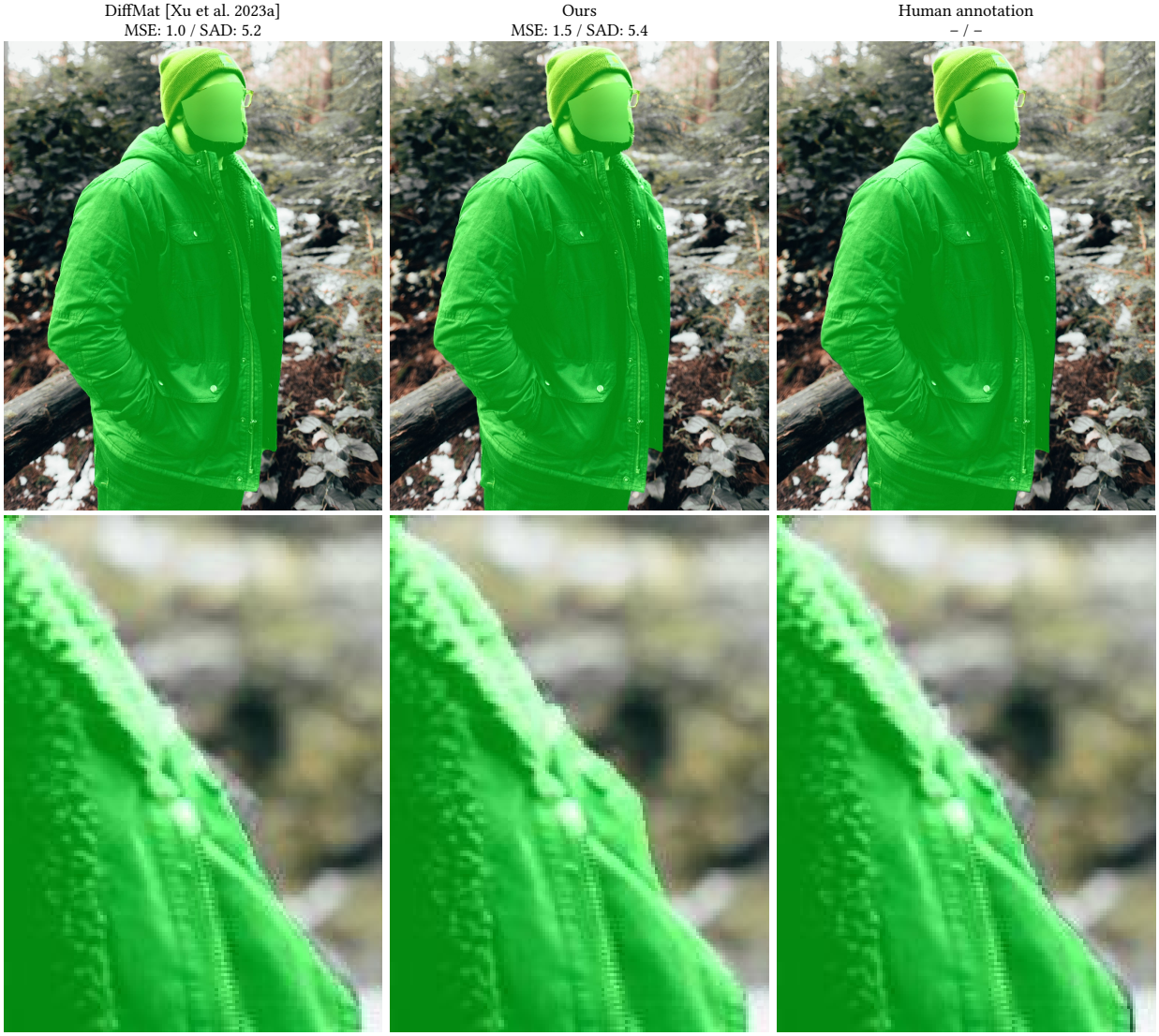


Figure 11: Results of trimap-based matting. Our visual results look better, but our evaluation score is worse than DiffMat, mainly because of the imperfect human annotation.

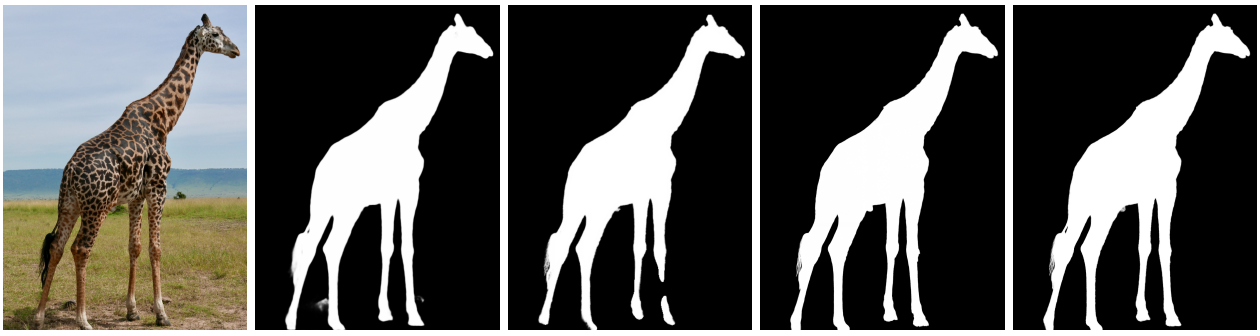


Figure 12: Matting beyond portraits. Based on SAM, MAM can generate a semantically correct alpha matte for the giraffe image but sacrifice some detail. ViTAE-S, on the other hand, fails to produce a semantically correct result and loses details. Our result closely matches the human annotation.