DiffuMatting: Synthesizing Arbitrary Objects with Matting-level Annotation

Xiaobin Hu[1⋆], Xu Peng[1,2⋆], Donghao Luo[1†], Xiaozhong Ji[1], Jinlong Peng[1], Zhengkai Jiang[1], Jiangning Zhang[1], Taisong Jin[2†], Chengjie Wang[1], and Rongrong Ji[2]

[1] Tencent Youtu Lab
[2] Key Laboratory of Multimedia Trusted Perception and Efficient Computing. Ministry of Education of China, Xiamen University. 361005, PR. China.

Abstract. Due to the difficulty and labor-consuming nature of getting highly accurate or matting annotations, there only exists a limited amount of highly accurate labels available to the public. To tackle this challenge, we propose a DiffuMatting which inherits the strong Everything generation ability of diffusion and endows the power of 'matting anything'. Our DiffuMatting can *1).* act as an anything matting factory with high accurate annotations *2).* be well-compatible with community LoRAs or various conditional control approaches to achieve the community-friendly art design and controllable generation. Specifically, inspired by green-screen-matting, we aim to teach the diffusion model to paint on a fixed green screen canvas. To this end, a large-scale green-screen dataset (Green100K) is collected as a training dataset for DiffuMatting. Secondly, a green background control loss is proposed to keep the drawing board as a pure green color to distinguish the foreground and background. To ensure the synthesized object has more edge details, a detailed-enhancement of transition boundary loss is proposed as a guideline to generate objects with more complicated edge structures. Aiming to simultaneously generate the object and its matting annotation, we build a matting head to make a green-color removal in the latent space of the VAE decoder. Our DiffuMatting shows several potential applications (*e.g.,* matting-data generator, community-friendly art design and controllable generation). As a matting-data generator, DiffuMatting synthesizes general object and portrait matting sets, effectively reducing the relative MSE error by 15.4% in General Object Matting. The dataset is released in our project page at https://diffumatting.github.io.

Keywords: Matting generator · Diffusion · Controllable generation

⋆ ⋆ these authors contributed equally to this work.
† indicates equally corresponding authors.

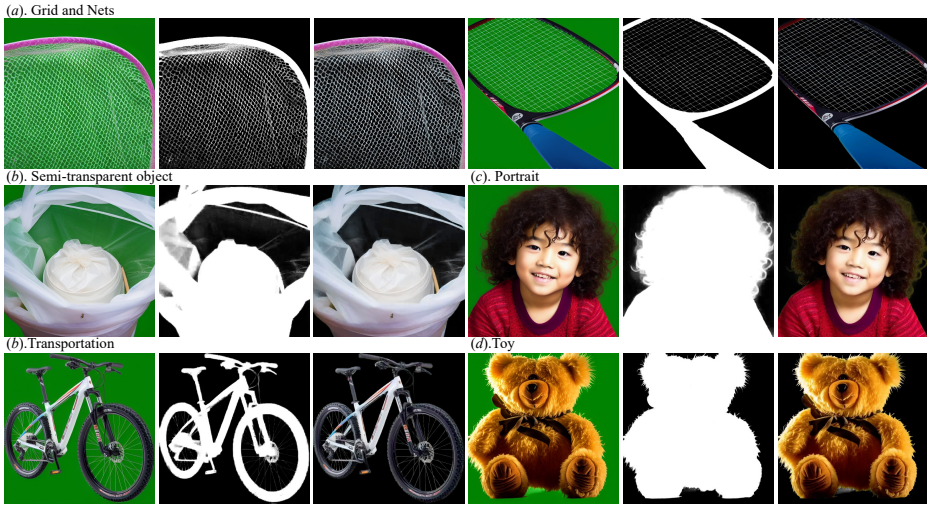


Fig. 1: Green-screen objects with matting-level annotations generation by DiffuMatting, including nets, grid and semitransparent tough objects and extended to almost any class (*e.g.,* Transportation, Architecture, Toy).

1 Introduction

High-accurate annotation (including segmentation of complex topological structures or matting objects) aims to provide highly delicate labels for objects. With the development of network architecture, the performance of accurate annotation has been boosted to some extent, and it provides more geometrically descriptions or transparency of the objects needed in many applications, augmented reality (AR) [32], image editing [8], 3D reconstruction [25], image composition [48]. Besides, considering the laborious labeling process and unacceptable cost, there only exists limited training pairs, *e.g.,* Adobe matting dataset [44] with only 493 foreground matting objects and Dinstinctions-646 [30] composed of 646 foreground images with manually annotated alpha mattes. It remains an open question of how to construct a highly accurate and efficient data-factory to simultaneously synthesize the delicate objects and their matting-level annotations, which also facilitating the composition and creation of content.

Recently, the emergence of diffusion models has shown its milestones in the exceptional generative ability to synthesize high-quality and diverse images. It unavoidably raises the consideration if the generative capability benefits the downstream tasks, especially on matting, segmentation, and image composition. There exists some work [3,10,36] on exploring the synthetic images to effectively improves the classification accuracy. As a more challenging task, segmentation requires both the synthetic images and its pixel-level aligned annotations. The segmentation data-generator can be categorized as 'conditional' and 'unconditional' manners. Conditional generation [42,45] requires semantic masks as the conditional inputs, and the generation scope is restricted within the semantic layouts. It fails to generate arbitrary objects ('Anything') but is only limited in the close domain with the training set. In contrast, unconditional genera-

tions [38, 43] can largely extend the generation space not limited to the same domain of semantic layouts. But it is still facing several challenges: *1).* It still does not fully exploit the Anything Generation of denoising diffusion probabilistic models, not beyond the class-priors in the training setting. *2).* It fails to generate the matting-level annotations.

To tackle the first challenge of Anything Generation, inspired by the green-screen matting, we ingeniously teach the diffusion model to paint on a fixed pure green screen canvas, which easier to distinguish the foreground and matting-level annotations. In other words, such a simple way embeds the object and high-accurate mask into the same green-screen image. This setting not only inherits the generic knowledge of the large-scale diffusion model by avoiding converging to class-prior training domain but also endows the diffusion model to naturally learn the background (green canvas) and foreground objects. Given that current SOTA diffusion models, (*e.g.,* DALLE3, SDXL [29], Midjourney), struggle to consistently generate arbitrary ultra-detailed objects on the green canvas as depicted in Fig. 2, it underscores the absence of a stable, robust diffusion model tailored for green-canvas-based scenarios. Thus, a generic green-canvas-based dataset Green100k is collected including 100,000 high-resolution images paired with high-accurate annotations, which makes it possible to train our DiffuMatting diffusion models. Additionally, a control loss with a green background is proposed to leverage cross-attention features associated with the 'green' color, ensuring the preservation of a pure and stable canvas background. *To address the second challenge of achieving high-accuracy matting annotations*, we propose a detailed-enhancement of transition boundary loss. This approach aims to generate objects with higher detail, akin to those in matting tasks, while also preventing transition boundary crashes. The mask generation is solely based on the latent space, allowing for the synthesis of well-aligned image-mask pairs without compromising the ability to generate anything. This approach circumvents the need for involvement in coarse-level, noise-based synthesis processes. In such a setting, the synthesis of sub-pixel well-aligned annotations can be simplified as a matting-header operation in the decoder, followed by matting-level post-processing. Our contributions are summarized as:

- We propose two novel loss functions: the green-background control loss and the transition boundary loss. The former leverages 'green' cross-attention features to ensure a stable canvas background, while the latter focuses on enhancing boundary detail and avoiding transition boundary crashes.
- To produce annotations at the matting level, Green100K is collected for training and a matting head is built to make a green-color removal on the latent space of VAE decoder which avoids being involved in the coarse-level noise-based synthesis processes.
- As a role of matting-data factor, we synthesize the general object and portrait matting set respectively and demonstrate that the inclusion of synthesized data leads to a reduction of 15.4% relative MSE error in General Object Matting.



Fig. 2: Visual performance of our DiffuMatting on green-screen object generation in comparison with SOTA Midjourney and SD-XL models, and these models have difficulties in consistently generating objects on the pure green-screen.

– DiffuMatting is well-compatible with various community LoRAs and existing control models (*e.g.,* ControlNet) without additional training or allows users to customize specific styles images with matting annotations.

2 Related Work

2.1 Text-to-image Diffusion Models

Diffusion models have recently demonstrated remarkable success in the field of image generation, leading to advancements in various applications and disciplines. Their impressive performance has greatly boosted the progress of text-guided image synthesis. Concretely, the development of large-scale text-to-image diffusion models, trained on extensive datasets of text-image pairs, has established new benchmarks in this domain. Prominent examples of such models include Stable diffusion [33], Imagen [34], DALL-E3, and other variants [9, 11, 14, 18, 28]. However, these SOTA diffusion models fail to steadily generate arbitrary ultra-detailed objects on the green canvas via a text-based prompt control. Our work inherits the generic knowledge of large-scale models by avoiding the domain restriction of image-mask training pairs. Additionally, by introducing a fixed green screen canvas, it enables the diffusion model to inherently learn both the background (green canvas) and foreground.

2.2 Synthetic Data Generation

Previous studies [7, 13] on dataset synthesis primarily employ 3D scene graphs (graphics engines) to generate images along with their corresponding labels. Nevertheless, these datasets often demonstrate a domain gap compared to real-world datasets, encompassing discrepancies in both appearance and content. A generative adversarial network leverages image-to-image translation to mind this gap in appearance and content. To generate infinite synthetic images and mask correspondingly, DatasetGAN [47] employed a limited number of labeled real images to train a segmentation mask decoder. Building upon DatasetGAN, BigDatasetGAN [19] expanded the class diversity to the scale of ImageNet, generating 1k classes with manually annotated 5 images per class. With the emergence of diffusion models, there have been some initial attempts to apply them to generate synthesis images for downstream tasks, *e.g.,* classification task [2, 10], synthetic ImageNet [35], altering the color [39], segmentation [4, 27, 38, 41, 43]. FreeMask [45]

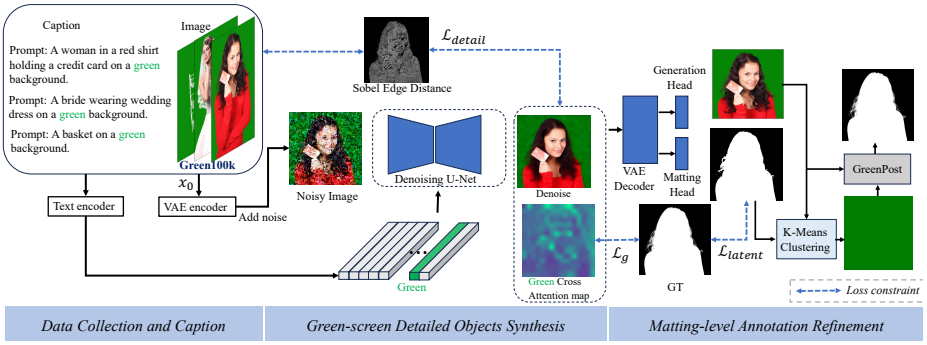


Fig. 3: An overview of Our DiffuMatting Network. Our DiffuMatting mainly consists of Green100k data collection and caption, green-screen detailed objects synthesis assisted by the green-background control loss $\mathcal{L}_g$ and the detailed-enhancement loss of transition boundary $\mathcal{L}_{detail}$, and matting-level annotation refinement via a matting-head in VAE latent space constrained by $\mathcal{L}_{latent}$ and GreenPost.

synthesizes training images on the scene understanding with the conditions of semantic mask provided by realistic dataset. Also in these works [1, 4, 17, 45], they also require the condition of the semantic layout, leading to tricky generate user-specified subjects. Ideally, the text-conditioned generation model is capable of relaxing the strict constraints, and synthesizing arbitrary objects via text phrase, which makes it potential to generate 'Anything'. DiffuMask [43] exploits the potential of the cross-attention map guided by text phrases to synthesize the predefined classes in the training set. Different from the above methods, our DiffuMatting can generate matting-level sub-pixel annotations which are much more expensive and labor-consuming than pixel-level masks. In addition, our DiffuMatting is not restricted by the training set but be largely extended to generate the most objects in real and virtual life.

2.3 Matting-level Dataset

Image matting refers to the precise estimation of the foreground object in images and videos. This technique holds significant importance in image and video editing applications, especially in the realm of film production where it is utilized to create captivating visual effects. However, considering the significant labor on matting-level annotations, Composition-1k [44], the most widely used dataset, only contains 431 foregrounds for training and 20 foregrounds for testing. To improve the versatility and robustness of the matting, Distinctions-646 dataset [30] is proposed with only 646 distinct FG images, AIM-500 [22] comprising 500 high-quality real natural images, P3M-500 [20] with 500 high-fidelity portraits and AM-2k [21] with 2,000 animal images. In summary, the amount of matting-level annotation is extremely rare compared with other pixel-level segmentations 11 million well-annotated images in SAM [16]. It is urgent to investigate a much cheaper and time-saving manner to synthesize abundant and almost unlimited images and matting-level annotations for any object.

3 Methodology

Motivation and Applications. Despite the widespread use of large-scale models for image generation, there is little research focus on matting-level transparent image generation. Incorporating transparency is extensively used in the majority of visual content editing software and workflows to facilitate the composition and creation of content, which is significantly needed by the commercial market. In addition, the scarcity of matting-level data hinders the development of matting algorithms towards high accuracy and strong robustness. Driven by these demands, we propose DiffuMatting to generate the 4-channels images with the high-accurate matting channels for the following core-meaning applications: *1).* Data factory. DiffuMatting can act as a data factory for downstream matting tasks in Fig. 6. *2).* Community-friendly art design and controllable generation. DiffuMatting exhibits excellent compatibility with community LoRAs and various conditional control approaches in Fig. 9 and Fig. 10. It also supports the image composition in desired scenario in Fig. 7.

3.1 Data Collection of Green100k Dataset

To address the dataset issue, we build a matting-level Green100k Dataset with 10,0000 green-screen-background objects and its high-accurate annotations which are mainly collected from public matting-level datasets and self-shoot green-screen portrait videos. Specifically, for self-collected portrait videos, we totally collected 20 portrait videos on the green-screen background. We annotate the above green-screen images in a matting manner with Adobe Premiere Pro Chromakey and Adobe Photoshop. To avoid some consecutive video frames, we sample one frame every 20 consecutive frames and the amount of self-collected images arrives at 1,1963. Considering that DiffuMatting is a general object generation model inheriting everything generation ability, we aim to increase the diversity of the collected data set and then include a small part of the high-accurate salient object dataset. Specifically, we collect the public available matting-level or high-accurate dataset including Adobe matting [44] (454 images), Distinction646 [30] (590 images), P3M [20] (481 images), PPM [15] (98 images), Video240k matting-set [24] (5,0000 images), AM2K [21] (1942 images), DIS5K [31] (5314 images), HRSOD [46] (1918 images), MSRA10k [12] (9884 images), DUTS [40] (10426 images) and Thinobject5k [23] (5594 images) dataset. The objects of the public matting-level dataset are synthesized into a green-screen background based on the corresponding accurate annotations. For the training of the stable diffusion model, we use the instruct-blip algorithm [6] to caption our Green100k dataset. To avoid the misalignment between the content and captions, a great labor of manual correction is implemented on all data.

3.2 Preliminaries of Cross-attention Mechanism

Stable diffusion is a probabilistic model specifically designed to estimate a data distribution p(x) by iteratively reducing noise in a normally distributed variable. Given an input image x_0, the noise estimation process is defined as:

$$\mathcal{L}_{noise} = \mathbb{E}_{z\sim\mathcal{E}(x),C,\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t, t, C)||_2^2\right], \tag{1}$$

where a Variational AutoEncoder $\mathcal{E}$ compress input image x into low-dimension latent space z. A conditional U-Net denoiser ϵ_θ aims to estimate noise θ in latent space with the aid of timestep t, t-th noisy latent representation z_t, and other text-prompt conditions C extracted by text-encoder. The cross-attention mechanism is implemented to fuse the visual and textual embedding via:

$$\mathcal{A} = \text{Softmax}\left(\frac{\ell_Q(\varphi(z_t))\ell_K(\tau_\theta(\mathcal{P}))^T}{\sqrt{d}}\right), \tag{2}$$

where $\ell_Q(\varphi(z_t)$ is the operation to flatten and linearly project deep spatial features of noisy image $\varphi(z_t)$ to Query vector, and $\ell_K(\tau_\theta(\mathcal{P}))$ is to lineraly project prompt $\mathcal{P}$ into textural embedding $\tau_\theta(\mathcal{P})$ as a Key matrix (K). d is latent projection dimension, ℓ_K and ℓ_Q are linearly learnable projection matrices.

3.3 Green-background Control

The overview structure of DiffuMatting is shown in Fig. 3. Given Green100k dataset, we need strong priors to restrict fore- and background to ensure the clean green screen of background zone. Intuitively, we can directly tie the main object of prompts with cross-attention mask to split fore- and background. However, to empower model of Everything Generation ability and largely enhance the generalization, we cannot get the class-priors of objects previously. Instead, the text token of 'green' is an available and stable prompt to derive background without class-priors. For a 'green' (j-th) text token, the corresponding weight is $\mathcal{A}_j \in \mathbb{R}^{H\times W}$. Our proposed green-background control mechanism is :

$$\mathcal{L}_g = \frac{1}{u}\sum_{l=1}^{u}|\mathcal{A}_j^l - (1-M)|_{mean}, \tag{3}$$

where M is GT mask of main object normalized to [0,1]. $\mathcal{A}_j^l$ represents cross-attention map corresponding to the j-th at the l-th cross-attention layer. u is total cross-attention layer. We supervise cross-attention map A_j to be close to the background segmentation mask $(1-M)$. $mean$ is the pixel-level averaging.

3.4 Detailed-enhancement of Transition Boundary

To keep the detailed-object generation and avoid the crash of the transition boundary, we extract the high-frequency information (*e.g.,* edge) [5] to enhance the details generation around the transition boundary.

$$H = (I \otimes S_x + I \otimes S_y) \odot I \odot M, \tag{4}$$

where S_x, S_y means the horizontal and vertical Sobel kernels operation to filter the high-frequency priors. I and M are the input gray image and mask, respectively. $\otimes$ refer to convolution operation and $\odot$ is the Haramard product on the pixel-level value. To guide the synthesis image to mimic the detailed objects, we restrict the synthesis image to high-frequency features closer to the GT images.

$$\mathcal{L}_{detail} = H_{\hat{z_0}} - H_{z_0}, \tag{5}$$

where z_0 is the latent representation of an image x_0, and $\hat{z_0}$ is the estimated latent representation directly from t-th noising representation z_t and the predicted noise $\epsilon_\theta(z_t, t, C)$ as follows:

$$\hat{z_0} = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta}{\sqrt{\bar{\alpha}_t}}, \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s, \tag{6}$$

where t means the t-th noise injection step, and $\bar{\alpha}_t$ is derived by predefined sequence of coefficients (α_t), controlling the variance schedule.

3.5 Mask Generation and Refinement

We aim to get the corresponding matting-level annotations as well as the stable and pure green-screen-based objects. To this end, we firstly get the coarse mask $\tilde{M}_{cp}$ in the latent representation space via adding a matting header. The higher dimension (128) of the latent representation feature ($f_{\mathbf{decoder}}$) is used to get the one-channel matting map in the VAE decoder as follows:

$$\tilde{M}_{cp} = \mathbf{ConR}(f_{\mathbf{decoder}}), \tag{7}$$

where **ConR** is the matting-header with two stack of 2d convolution layer and SiLU activation function. Afterwards, we tie $\tilde{M}_{cp}$ with the GT (M) based on the dice and $L1$ loss as follows:

$$\mathcal{L}_{latent} = \frac{1}{N}\sum_{j=1}^{N} |\tilde{M}_{cp_j} - M_j| + 1 - \frac{2|\tilde{M}_{cp} \cap M|}{|\tilde{M}_{cp}| + |M|}, \tag{8}$$

where N is total pixel amount, and $\cap$ is intersection of $\tilde{M}_{cp}$ and M. $||$ means pixel-level sum of a matrix. Lastly we adopts a post-processing GreenPost, using green-background priors to refine the pixel-level mask $\tilde{M}_{cp}$. The GreenPost introduces the BackgroundMattingV2 [24] as the post-processing after using the K-Means to cluster the background color calculated by the pixel-level mask $\tilde{M}_{cp}$ to fill the foreground pixel values. Then original image and accurate background are imported into the GreenPost to obtain the matting-level annotation $\tilde{M}_p$.

$$\tilde{M}_p = \text{GreenPost}(\tilde{M}_{cp}), \tag{9}$$

3.6 Objective Function

To control the pure and clean green-background, $\mathcal{L}_g$ is proposed to optimize the cross-attention feature for accurate fore- and background learning via a bidirectional manner. This setting is beneficial for the downstream GreenPost operation to get matting level annotations. In addition, to encourage the object with more details, a detailed-enhancement loss $\mathcal{L}_{detail}$ of transition boundary is used to facilitate the boundary details via high-frequency information (*e.g.,* edge) alignment. The $\tilde{M}_{cp}$ is derived by the VAE decoder matting head, and $\mathcal{L}_{latent}$ are applied to mind the gap with GT denoted as M.

$$\mathcal{L}_{total} = \mathcal{L}_{noise} + \mathcal{L}_g + \mathcal{L}_{detail} + \mathcal{L}_{latent}. \tag{10}$$

Table 1: Green-screen generation quality evaluation.

Metric	LoRA	Dreambooth	Ours
GSG ↓	134.17	133.34	**98.98**
A_s ↑	4.78	4.85	**5.26**

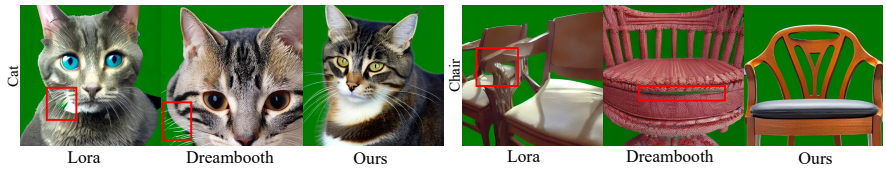


Fig. 4: Visual performance of our DiffuMatting on green-screen-based object generation in comparison with LoRA and Dreambooth fine-tuning in our Green100K.

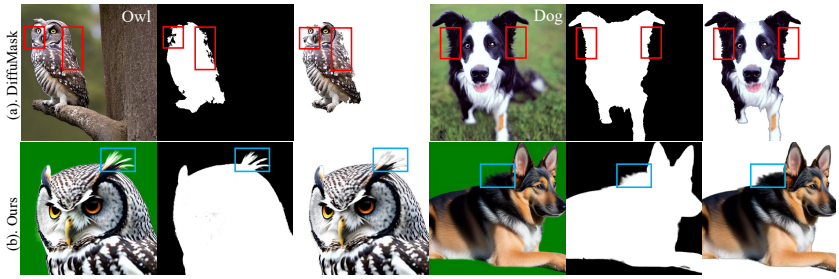


Fig. 5: Matting-level annotation analysis. Our matting-level annotation (blue box) *vs.* pixel-level mask generated by DiffuMask [43] (red box) in the class of dog and Owl (bird). Each class-object generation by DiffuMask requires fine-tuning AffineityNet for this specific class to get post-processing annotation results.

4 Experiments

4.1 Experimental Setups

Training Details. Stable Diffusion V1.5 is adopted as the pretrained model. Our model undertakes two tasks including the object generation on the pure green background and matting-level annotation tasks. First, we only train a DiffuMatting generation model without matting parameters to ensure the pure and stable green-screen generation with a learning rate of 2e-6 and batch size of 2 on the two NVIDIA V100 GPUs for 1.5 days. Then, we use the well-trained parameters in the last step and jointly train the generation and matting stage on the two NVIDIA V100 GPUs for 1 day. This setting can avoid the misalignment of green-screen generation at the beginning of training, which leads to the deterioration of matting parameters training.

Evaluation Metric. Considering no existing available criteria to evaluate the quality of stable and pure green-screen background, we propose a green-screen generation quality (GSG) that uses the K-Means to cluster the dominant color of synthesis images and calculate the distance with the green-screen color.

$$\mathbf{GSG} = |(\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} ||\mathbf{x} - \mu_i||^2)_{\text{top}} - P_g|_{Eu}, \tag{11}$$

Ratio	SAD↓	MSE ↓	GRAD ↓	CONN ↓
0	75.6	0.033	45.1	88.2
1/3	50.2	0.018	27.6	45.3
1/2	46.3	0.013	26.2	41.2
2/3	44.7	**0.011**	25.3	40.6
4/5	**43.3**	0.012	**24.8**	**39.6**
1	45.8	0.013	25.9	43.7

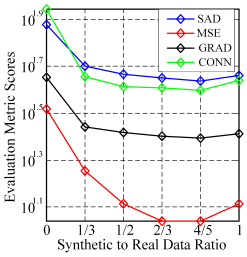


Fig. 6: The general matting results using our synthesis 10K General-matting set based on the Indexnet [26].

where x is a set of observations divided into k sets. μ_i is the mean of points in S_i. $()_{\text{top}}$ denotes the top set S, the dominant color of an image. P_g is the GT geometric coordinates of pure green-screen color and $||_{Eu}$ is the Euclidean distance between the estimated dominant color and GT coordinates. Following previous works, the sum of absolute differences (SAD), mean squared error (MSE), gradient (Grad.), and Connectivity (Conn.) are used.

4.2 Consistency Green-Screen Generation Capability

To verify the stability, consistency, and pure green-screen generation, we compare our proposed method with LoRA and Dreambooth models fine-tuning in the same training set Green100K with the default setting. Specifically, we randomly generate 10 class images (*e.g.,* dog, cat, woman, man, bike, chair, bridge, laptop, desk, car) and each class contains 10 images. The aesthetic-score A_s [37] is to analyze the aesthetic score of the green-screen images. The quantitative results are listed in Table 1. Our training setting powered by the green-background control is much better than LoRA and Dreambooth models. Some qualitative results are also shown in Fig. 4. DiffuMatting shows a strong generation ability on the details of transition area (cat whiskers) and the purity of green-screen.

4.3 Matting-level Annotation Analysis

As a recent-related work [43], DiffuMask is a diffusion model to synthesize pixel-level annotations. DiffuMask fails to provide a matting-level annotation and suffers from the misalignment of synthesized images and annotations to a certain extent. Each class-object generation requires fine-tuning AffineityNet for this specific class to get post-processing annotation. DiffuMask released two class AffineityNet weights (*i.e.*, dog and bird). We compare matting-level annotations of DiffuMatting with the pixel-level masks in Fig. 5. It shows that our DiffuMatting achieves hair-level annotation (*e.g.,* fur, hair) compared with DiffuMask.

4.4 Applications

Downstream Matting Task. To verify the benefits of our matting-level annotation to the down-streaming matting-level task, we analyze our synthesized images on General-object matting. *1).* General Object Matting: We synthesize the

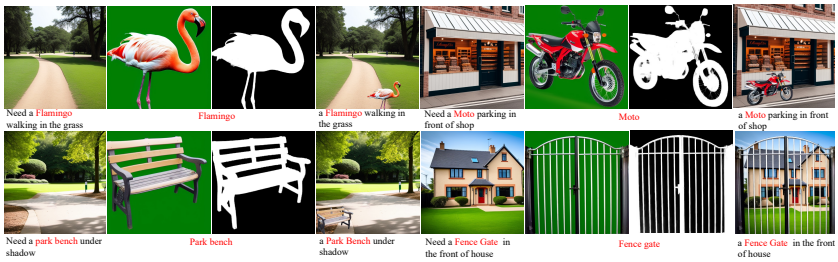


Fig. 7: Image Composition. DiffuMatting generates green-screen objects, copies the object with the matting-level annotation, and then pastes it into desired scenario.

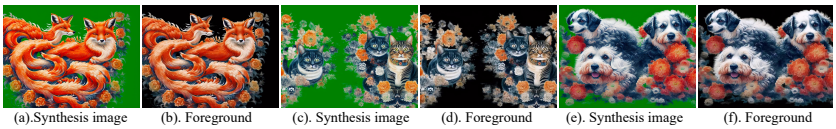


Fig. 8: Art design: Style of blue and white porcelain texture with alpha-channel generated by DiffuMatting.

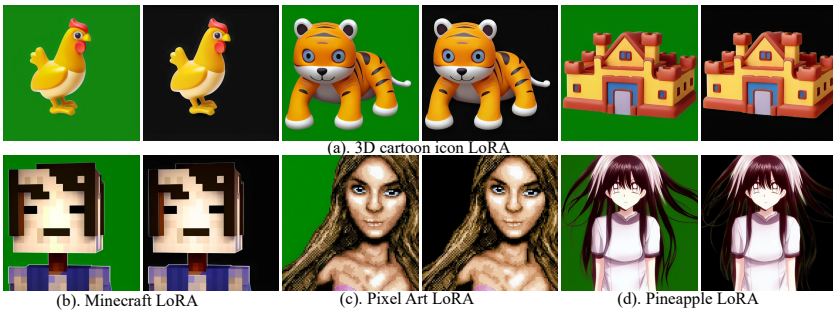


Fig. 9: Applying to Community LoRAs: we select some typical style LoRAs (*i.e.,* 3D cartoon icon LoRA, Minecraft LoRA, Pixel Art LoRA, Pineapple LoRA) for testing.

10K general-object matting set (GOM) based on the around 500 object classes guided by the detailed captions of Green100K dataset. We train a general matting model based on the commonly-used Indexnet [26] under the same default setting on various training sets, consisting of images from the general-object matting set (GOM) and Deep Image Matting (DIM) dataset. The synthetic to real data ratio γ denotes that γ ratio of images from the DIM dataset and $(1-\gamma)$ ratio of images from the synthesized GOM. As shown in Fig. 6, the combined dataset (highlighted in bold) used for matting gets better results (15.4% relative error lower on MSE, 9.4% on CONN) than only using the DIM dataset (Ratio γ=1), indicating that synthesized images by our DiffuMatting is useful for downstream general object matting even though it stills exists domain difference between the real and synthesized dataset.



Fig. 10: Controllable text-to-image generation combing with the ControlNet. Our Diffumatting can be directly combined with the ControlNet for sketch-guided and pose-guided text-to-image generation.

Table 2: Ablation study. The quality evaluation of green-screen synthesis. GB is denoted as the green-background control loss, and DE is detailed-enhancement loss.

Metric	*w/o* GB	*w/o* DE	Ours
GSG ↓	119.8	112.5	**98.98**
A_s ↑	5.02	4.95	**5.26**

Image Composition: Generate, Copy and Paste. Image composition usually suffers from the coarse mask that does not describe the edges and holes of objects. The coarse mask can be generated by the segmentation algorithms (*e.g.,* SAM) in a low-cost manner. In contrast, matting-level annotations require much larger labor-consuming cost which is not cheaply available for image composition. But our DiffuMatting greatly decreases the cost of matting annotations benefiting from its great generalization that supports generating 'anything' matting-level annotations. Thus, our DiffuMatting can generate green-screen objects, copy the object with the matting-level annotation, and paste the object into composition scenarios in a closed loop. Some image composition results can be seen in the Fig. 7, adding different elements to a target scene.

Community-friendly Art Design and Controllable Generation. DiffuMatting is well compatible with LoRA models, and it provides the potential for users to customize the specific-style RGBA 4-channels images. DiffuMatting acts as the base model and a style-based LoRA can be trained by the users' style images. The self-training style-LoRA results of blue and white porcelain texture in 4-channels are shown in Fig. 8. Besides, our DiffuMatting can be ap-



Fig. 11: Visual quality of ablation study with different configurations.

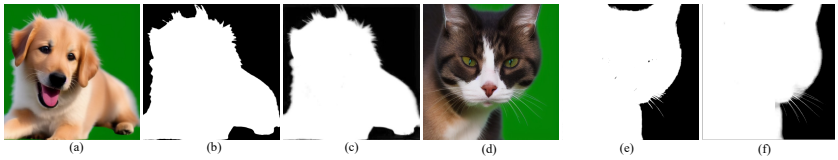
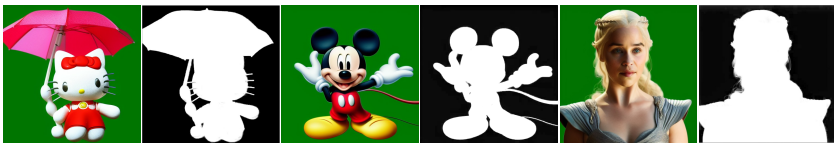


Fig. 12: The matting refinement refines the pixel-level mask to the matting-level annotation. (a) and (d) green-screen synthesized images, (b) and (e) pixel-level masks from the matting-head of latent space, (c) and (f) the outputs of the GreenPost matting refinement process.

plied to the various community LoRAs models without any additional training process. For example, Minecraft LoRA, 3D cartoon icon LoRA, pixel art LoRA, and Pineapple LoRA are used to get diverse results in Fig. 9. We see that applying the existing Community LoRA can still keep the style quality, and get a good transparency channel to extract the foreground objects. The existing control models (*e.g.,* ControlNet) can be directly applied in DiffuMatting to enrich the controllable image editing functionality in Fig. 10. For sketch-guided text-to-image generation, our model can generate editable style images (*e.g.,* Western and Asian style castle, or the specified color of hair) according to the ControlNet signal. Given human pose knowledge, our model can also edit the portrait gesture guided by the ControlNet.

4.5 Ablation Study

Green-background Control. Fig. 11 shows that the setting without green background control leads to the color bleeding around the transition area of fore- and background, *e.g.,* the white hair color of elder affecting the green-screen background and the transition hair area of woman is mixed by the green and hair color. Also in Table 2, the setting without green background control gets a much worse GSG score, which indicates that the green-screen background is not satisfactory. It also deteriorates the aesthetic-score in such a setting.
Detailed-enhancement of transition boundary. As shown in Fig. 11, the setting without the detailed-enhancement of transition boundary suffers from the detail loss of the boundary, especially on the hair area. After adding both green-background control and detailed-enhancement of transition boundary, the detailed-preserving of fur or hair and distinction of background and foreground



(a) Hello Kitty (b). Mickey Mouse (c). Daenerys Targaryen

Fig. 13: Anything Generation Ability. The object generation (*e.g.,* cartoon and celebrity) beyond the Green100K.

are improved to steadily generate high-quality synthesis image and matting annotation. From Table 2, we find that the detailed-enhancement loss has a greater impact on the aesthetic score compared to green-background loss.

Mask Generation and Refinement. Our matting head can acquire a pixels-level mask which does not satisfy our matting-level annotation. Then the matting refinement process guided by the background-prior is proposed to refine the pixels-level mask. As shown in Fig. 12, after the GreenPost matting refinement process, the matting output shows more transparent information.

Strong Generalization beyond Green100K. Green100K is a teacher dataset that teaches the diffusion model to paint on a fixed green-screen canvas. After learning this principle, our model can also inherit the Anything Generation Ability of pretrained large-scale diffusion model. As shown in Fig. 13, the images and matting-level annotations largely beyond Green100K (*e.g.,* Hello Kitty, Mickey Mouse, or even Daenerys Targaryen) can be easily and steadily generated.

5 Conclusion

We propose a DiffuMatting taught by our Green100k to paint on a green-screen canvas, which ingeniously separates the foreground. With the aid of the background, transition boundary control, and matting refinement, DiffuMatting combines the robust generative capabilities of diffusion and the added functionality of 'matting anything'. As such, our DiffuMatting serves as a highly accurate 'anything matting' factory which facilitates the down-streaming matting task, while also achieving the composition and creation of content or community-friendly art design and controllable generation.

Limitation and Social Impacts. Our work is primarily centered around the generation on the matting-level foreground of 'Anything'. It has limitations with simultaneously synthesizing non-green screen images and matting-level annotations without any assistance of image inpainting technique. DiffuMatting technology can be used for content creation and carry risks in illicit industries. To mitigate potential misuse, explicit markings will be applied to generated images.

Acknowledgments. This work was supported by National Key R&D Program of China (No. 2022ZD0118202), in part by the National Natural Science Foundation of China (No. 62072386), in part by Yunnan Provincial Major Science and Technology Special Plan Project (No. 202402AD080001), in part by by Henan Province key research and development project (No. 231111212000) and the Open Foundation of Henan Key Laboratory of General Aviation Technology (No. ZHKF-230212).

References

1. Azadi, S., Tschannen, M., Tzeng, E., Gelly, S., Darrell, T., Lucic, M.: Semantic bottleneck scene generation. arXiv preprint arXiv:1911.11357 (2019) 5
2. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. arXiv preprint arXiv:2304.08466 (2023) 4
3. Bansal, H., Grover, A.: Leaving reality to imagination: Robust classification via generated datasets. arXiv preprint arXiv:2302.02503 (2023) 2
4. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. ICLR (2022) 4, 5
5. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023) 7
6. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023) 6
7. Devaranjan, J., Fidler, S., Kar, A.: Unsupervised learning of scene structure for synthetic data generation (Sep 9 2021), uS Patent App. 17/117,425 4
8. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. IEEE transactions on pattern analysis and machine intelligence **34**(10), 1915–1926 (2011) 2
9. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706 (2022) 4
10. He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., Qi, X.: Is synthetic data from generative models ready for image recognition? ICLR (2022) 2, 4
11. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) 4
12. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.: Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3203–3212 (2017) 6
13. Kar, A., Prakash, A., Liu, M.Y., Cameracci, E., Yuan, J., Rusiniak, M., Acuna, D., Torralba, A., Fidler, S.: Meta-sim: Learning to generate synthetic datasets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4551–4560 (2019) 4
14. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in Neural Information Processing Systems **35**, 26565–26577 (2022) 4
15. Ke, Z., Sun, J., Li, K., Yan, Q., Lau, R.W.: Modnet: Real-time trimap-free portrait matting via objective decomposition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1140–1147 (2022) 6
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 5
17. Le Moing, G., Vu, T.H., Jain, H., Pérez, P., Cord, M.: Semantic palette: Guiding scene generation with class proportions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9342–9350 (2021) 5
18. Li, B., Xue, K., Liu, B., Lai, Y.K.: Bbdm: Image-to-image translation with brownian bridge diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1952–1961 (2023) 4

19. Li, D., Ling, H., Kim, S.W., Kreis, K., Fidler, S., Torralba, A.: Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21330–21340 (2022) 4
20. Li, J., Ma, S., Zhang, J., Tao, D.: Privacy-preserving portrait matting. In: Proceedings of the 29th ACM international conference on multimedia. pp. 3501–3509 (2021) 5, 6
21. Li, J., Zhang, J., Maybank, S.J., Tao, D.: Bridging composite and real: towards end-to-end deep image matting. International Journal of Computer Vision **130**(2), 246–266 (2022) 5, 6
22. Li, J., Zhang, J., Tao, D.: Deep automatic natural image matting. arXiv preprint arXiv:2107.07235 (2021) 5
23. Liew, J.H., Cohen, S., Price, B., Mai, L., Feng, J.: Deep interactive thin object selection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 305–314 (2021) 6
24. Lin, S., Ryabtsev, A., Sengupta, S., Curless, B.L., Seitz, S.M., Kemelmacher-Shlizerman, I.: Real-time high-resolution background matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8762–8771 (2021) 6, 8
25. Liu, F., Tran, L., Liu, X.: Fully understanding generic objects: Modeling, segmentation, and reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7423–7433 (2021) 2
26. Lu, H., Dai, Y., Shen, C., Xu, S.: Indices matter: Learning to index for deep image matting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3266–3275 (2019) 10, 11
27. Park, M., Yun, J., Choi, S., Choo, J.: Learning to generate semantic layouts for higher text-image correspondence in text-to-image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7591–7600 (2023) 4
28. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023) 4
29. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 3
30. Qiao, Y., Liu, Y., Yang, X., Zhou, D., Xu, M., Zhang, Q., Wei, X.: Attention-guided hierarchical structure aggregation for image matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13676–13685 (2020) 2, 5, 6
31. Qin, X., Dai, H., Hu, X., Fan, D.P., Shao, L., Van Gool, L.: Highly accurate dichotomous image segmentation. In: European Conference on Computer Vision. pp. 38–56. Springer (2022) 6
32. Qin, X., Fan, D.P., Huang, C., Diagne, C., Zhang, Z., Sant'Anna, A.C., Suarez, A., Jagersand, M., Shao, L.: Boundary-aware segmentation network for mobile and web applications. arXiv preprint arXiv:2101.04704 (2021) 2
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 4
34. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-

to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022) 4
35. Sariyildiz, M.B., Alahari, K., Larlus, D., Kalantidis, Y.: Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In: CVPR 2023–IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 4
36. Sauer, A., Schwarz, K., Geiger, A.: Stylegan-xl: Scaling stylegan to large diverse datasets. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–10 (2022) 2
37. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022) 10
38. Shenoda, M., Kim, E.: Diffugen: Adaptable approach for generating labeled image datasets using stable diffusion models. arXiv preprint arXiv:2309.00248 (2023) 3, 4
39. Trabucco, B., Doherty, K., Gurinas, M., Salakhutdinov, R.: Effective data augmentation with diffusion models. arXiv preprint arXiv:2302.07944 (2023) 4
40. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 136–145 (2017) 6
41. Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., Li, H.: Semantic image synthesis via diffusion models. arXiv preprint arXiv:2207.00050 (2022) 4
42. Wang, Y., Qi, L., Chen, Y.C., Zhang, X., Jia, J.: Image synthesis via semantic composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13749–13758 (2021) 2
43. Wu, W., Zhao, Y., Shou, M.Z., Zhou, H., Shen, C.: Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. International Conference on Computer Vision (2023) 3, 4, 5, 9, 10
44. Xu, N., Price, B., Cohen, S., Huang, T.: Deep image matting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2970–2979 (2017) 2, 5, 6
45. Yang, L., Xu, X., Kang, B., Shi, Y., Zhao, H.: Freemask: Synthetic images with dense annotations make stronger segmentation models. arXiv preprint arXiv:2310.15160 (2023) 2, 4, 5
46. Zeng, Y., Zhang, P., Zhang, J., Lin, Z., Lu, H.: Towards high-resolution salient object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7234–7243 (2019) 6
47. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.F., Barriuso, A., Torralba, A., Fidler, S.: Datasetgan: Efficient labeled data factory with minimal human effort. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10145–10155 (2021) 4
48. Zhu, P., Abdal, R., Femiani, J., Wonka, P.: Barbershop: Gan-based image compositing using segmentation masks. arXiv preprint arXiv:2106.01505 (2021) 2