ObjectStitch: Object Compositing with Diffusion Model

Yizhi Song[1*], Zhifei Zhang[2], Zhe Lin[2], Scott Cohen[2], Brian Price[2],
Jianming Zhang[2], Soo Ye Kim[2], Daniel Aliaga[1]
Purdue University[1], Adobe Research[2]
[1]{song630,aliaga}@purdue.edu, [2]{zzhang,zlin,scohen,bprice,jianmzha,sooyek}@adobe.com

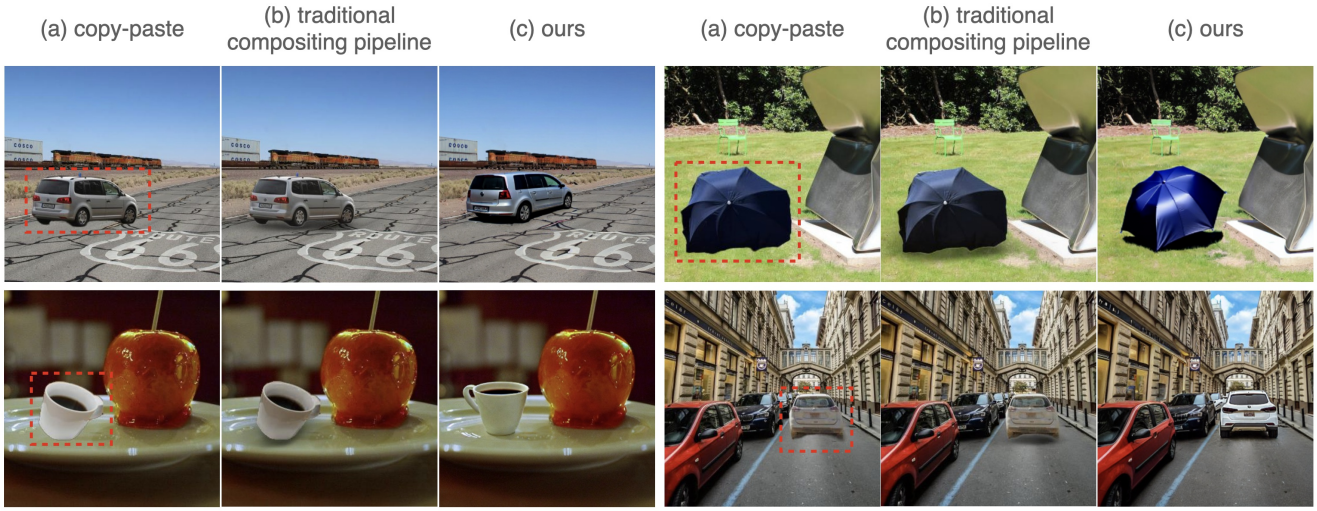

Figure 1. Example results of object compositing with (a) copy-and-paste scheme, (b) traditional compositing pipeline and (c) ours (ObjectStitch). Traditional compositing pipeline is done with best possible off-the-shelf models including foreground/background color harmonization [19, 47], poisson blending [34], and shadow synthesis [42]. ObjectStitch achieves more realistic results, and can address geometry correction, harmonization, shadow generation, and view synthesis all-in-one while preserving similar appearance to the reference object.

Abstract

Object compositing based on 2D images is a challenging problem since it typically involves multiple processing stages such as color harmonization, geometry correction and shadow generation to generate realistic results. Furthermore, annotating training data pairs for compositing requires substantial manual effort from professionals, and is hardly scalable. Thus, with the recent advances in generative models, in this work, we propose a self-supervised framework for object compositing by leveraging the power of conditional diffusion models. Our framework can hollistically address the object compositing task in a unified model, transforming the viewpoint, geometry, color and shadow of the generated object while requiring no manual labeling. To preserve the input object's characteristics, we introduce a content adaptor that helps to maintain categorical semantics and object appearance. A data augmentation method is further adopted to improve the fidelity of the generator. Our method outperforms relevant baselines in both realism and faithfulness of the synthesized result images in a user study on various real-world images.

1. Introduction

Image compositing is an essential task in image editing that aims to insert an object from a given image into another image in a realistic way. Conventionally, many sub-tasks are involved in compositing an object to a new scene, including color harmonization [6, 7, 19, 51], relighting [52], and shadow generation [16, 29, 43] in order to naturally blend the object into the new image. As shown in Tab. 1, most previous methods [6, 7, 16, 19, 28, 43] focus on a single sub-task required for image compositing. Consequently, they must be appropriately combined to obtain a composite image where the input object is re-synthesized to have the

* Work done during internship at Adobe.

Method	Geometry	Light	Shadow	View
ST-GAN [28]	✓	✗	✗	✗
SSH [19]	✗	✓	✗	✗
DCCF [51]	✗	✓	✗	✗
SSN [43]	✗	✗	✓	✗
SGRNet [16]	✗	✗	✓	✗
GCC-GAN [5]	✓	✓	✗	✗
Ours	✓	✓	✓	✓

Table 1. Prior works only focus on one or two aspects of object compositing, and they cannot synthesize novel views. In contrast, our model can address all perspectives as listed.

color, lighting and shadow that is consistent with the background scene. As shown in Fig. 1, results produced in this way still look unnatural, partly due to the viewpoint of the inserted object being different from the overall background.

Harmonizing the geometry and synthesizing novel views have often been overlooked in 2D image compositing, which require an accurate understanding of both the geometry of the object and the background scene from 2D images. Previous works [15, 21, 22] handle 3D object compositing with explicit background information such as lighting positions and depth. More recently, [28] utilize GANs to estimate the homography of the object. However, this method is limited to placing furniture in indoor scenes. In this paper, we propose a generic image compositing method that is able to harmonize the geometry of the input object along with color, lighting and shadow with the background image using a diffusion-based generative model.

In recent years, generative models such as GANs [4, 10, 17, 20] and diffusion models [1, 14, 30, 32, 33, 38, 39] have shown great potential in synthesizing realistic images. In particular, diffusion model-based frameworks are versatile and outperform various prior methods in image editing [1, 23, 32] and other applications [11, 31, 35]. However, most image editing diffusion models focus on using text inputs to manipulate images [1,3,9,23,33], which is insufficient for image compositing as verbal representations cannot fully capture the details or preserve the identity and appearance of a given object image. There have been recent works [23, 39] focusing on generating diverse contexts while preserving the key features of the object; however, these models are designed for a different task than object compositing. Furthermore, [23] requires fine-tuning the model for each input object and [39] also needs to be fine-tuned on multiple images of the same object. Therefore they are limited for general object compositing.

In this work, we leverage diffusion models to simultaneously handle multiple aspects of image compositing such as color harmonization, relighting, geometry correction and shadow generation. With image guidance rather than text guidance, we aim to preserve the identity and appearance of the original object in the generated composite image.

Specifically, our model synthesizes a composite image given (i) a source object, (ii) a target background image, and (iii) a bounding box specifying the location to insert the object. The proposed framework consists of a content adaptor and a generator module: the content adaptor is designed to extract a representation from the input object containing both high-level semantics and low-level details such as color and shape; the generator module preserves the background scene while improving the generation quality and versatility. Our framework is trained in a fully self-supervised manner and no task-specific labeling is required at any point during training. Moreover, various data augmentation techniques are applied to further improve the fidelity and realism of the output. We evaluate our proposed method on a real-world dataset closely simulating real use cases for image compositing.

Our contributions are summarized as follows:

- We present the first diffusion model-based framework for *generative object compositing* that can handle multiple aspects of compositing such as viewpoint, geometry, lighting and shadow.
- We propose a content adaptor module which learns a descriptive multi-modal embedding from images, enabling image guidance for diffusion models.
- Our framework is trained in a self-supervised manner without any task-specific annotations, employing data augmentation techniques to improve the fidelity of generation.
- We collect a high-resolution real-world dataset for object compositing with diverse images, containing manually annotated object scales and locations.

2. Related Work

2.1. Image Compositing

Image compositing is a challenging task in reference-guided image editing, where the object in a given foreground image is to be inserted into a background image. The generated composite image is expected to look realistic, with the appearance of the original object being preserved. Prior works often focus on an individual aspect of this problem, such as geometric correction [2, 28], image harmonization [6, 7, 51], image matting [50] and shadow generation [29].

Lin *et al*. [28] propose an iterative GAN-based framework to correct geometric inconsistencies between object and background images. In their model, STNs [18] are integrated into the generator network to predict a series of warp updates. Although their work can improve the realism of the scene geometry in the composite image, it is limited to inserting furniture into indoor scenes. Azadi *et al*. [2] focus on binary composition, using a composition-decomposition

network to capture interactions between a pair of objects.

Image harmonization aims at minimizing the inconsistency between the input object and the background image. Traditional methods [24, 46] usually concentrate on obtaining color statistics and then transfer this information between foreground and background. Recent works seek to solve this problem via deep neural networks. [6, 7] reformulate image harmonization as a domain adaptation problem, while [19] converts it to a representation fusing problem and utilizes self-supervised training.

Shadow synthesis is an effect that is often overlooked in previous image compositing methods although it is essential for generating realistic composite images. SGRNet [16] divides this task into a shadow mask generation stage and a shadow filling stage. SSN [43], focusing on soft shadows, predicts an ambient occlusion map as cue for shadow generation.

There are also works simultaneously addressing multiple sub-problems of image compositing. Chen *et al*. [5] develop a system including multiple network modules to produce plausible images with geometric, color and boundary consistency. However, the pose of the generated object is constrained by the mask input and the model cannot generalize to non-rigid objects such as animals. Our proposed model also handles the image composition problem in a unified manner, generating foreground objects which are harmonious and geometrically consistent with the background while synthesizing novel views and shadows.

2.2. Guided Image Synthesis

Recent years have witnessed great advances in diffusion models. Diffusion models are a family of deep generative models based on several predominant works [14, 44, 45], defined with two Markov chains. In the forward process, they gradually add noise to the data, and in the reverse process, they learn to recover the data from noise. Following the growth of research in this area, diffusion models have shown their potential in a great number of applications, such as text-to-image generation [9, 33, 38], image editing [1, 23, 32] and guided image synthesis [30].

Text-to-image synthesis has been a popular research topic over the past few years. Stable Diffusion [38] made a significant contribution to this task and it reduces the computational cost by applying the diffusion model in the latent space. Other works explore ways of more flexible and controllable text-driven image editing. Avrahami *et al*. [1] design an algorithm in their Blended Diffusion framework to fuse noised versions of the input with the local text-guided diffusion latent, generating smooth transition. Similarly, GLIDE [33] is also capable of handling text-guided inpainting after fine-tuning on this specific task. Some methods give users more straightforward control on images. SDEdit [32] allows stroke painting in images, blending user input into the image by adding noise to the input and denoising through a stochastic differential equation. SDG [30] injects Semantic Diffusion Guidance at each iteration of the generation and can provide multi-modal guidance.

While the versatility and generation quality of diffusion models have been repeatedly demonstrated, preserving object appearance remains a very challenging problem in image editing. Ruiz *et al*. [39] address this problem by optimizing their model using a reconstruction loss and a class-specific prior preservation loss. Though their approach can preserve the details of the object, multiple images of the same object are required to fine-tune their model on this target object. Kawar *et al*. [23] handle this issue by interpolating the initial text embedding and the optimized embedding for reconstruction, but their method is mainly designed for applying edits on objects, not for generating the same object in different contexts. In this paper, we propose the first generative object compositing framework based on diffusion models, which generates the harmonized and geometrically-corrected subject with a novel view and shadow. Furthermore, the characteristics of the original object is preserved in the synthesized composite image.

3. Proposed Method

We define the generative object compositing problem as: Given an input triplet (I_o, I_{bg}, M) that consists of an object image $I_o \in \mathbb{R}^{H_s \times W_s \times 3}$, a background image $I_{bg} \in \mathbb{R}^{H_t \times W_t \times 3}$, and its associated binary mask $M \in \mathbb{R}^{H_t \times W_t \times 1}$ with the desired object location set to 0 and the rest to 1, the goal is to composite the input object into the masked area $I_{bg} \bigotimes M$. M is considered as a soft constraint of the location and scale of the composited object. The output generated image should look realistic, while the appearance of the object is preserved. Our problem setting is different from text-guided image generation and inpainting in that the condition input is a reference object image rather than a text prompt. Inspired by the success of text-guided diffusion models [37, 38, 40], which inject text conditioning into the diffusion architecture, we design our method for generative compositing to leverage such pretrained models.

An overview of our framework is shown in Fig. 2. It consists of an object image encoder extracting semantic features from the object and a conditional diffusion generator. To leverage the power of pretrained text-to-image diffusion models, we introduce a content adaptor that can bridge the gap between the object encoder and conditional generator by transforming a sequence of visual tokens to a sequence of text tokens to overcome the domain gap between image and text. This design further helps to preserve the object appearance. We propose a two-stage training process: in the first stage, the content adaptor is trained on large image/text pairs to maintain high-level semantics of the object; in the second stage, it is trained in the context of diffusion

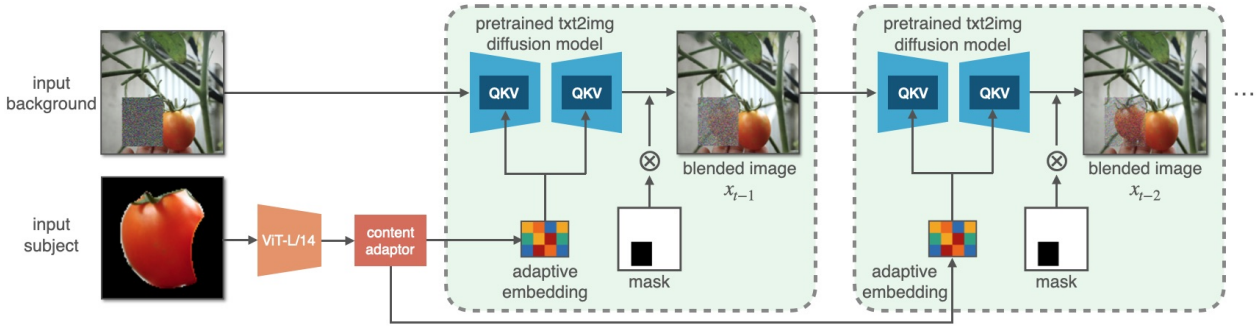

Figure 2. System pipeline. Our framework consists of a content adaptor and a generator (a pretrained text-to-image diffusion model). The input image I_o is fed into a ViT and the adaptor which produces a descriptive embedding. At the same time the background image I_{bg} is taken as input by the diffusion model. At each iteration during the denoising stage, we apply the mask M on the generated image I_{out}, so that the generator only denoises the masked area $I_{out} \bigotimes M$.

generator to encode key identity features of the object by encouraging the visual reconstruction of the object in the original image. Lastly, the generator module is fine-tuned on the embedding produced by the adaptor through cross attention blocks. All stages our trained in a self-supervised manner to avoid expensive annotation for obtaining object compositing training data.

3.1. Generator

As depicted in Fig. 2, we leverage a pretrained text-to-image diffusion model architecture and modify it for the compositing task by (i) introducing a mask in the input, and (ii) adjusting the U-Net input to contain the original background image outside the hole and noise inside the hole with mask blending. In order to condition the model on the guidance embedding E, an attention mechanism is applied as:

$$\text{Softmax}\left(\frac{(\boldsymbol{W}_Q E_x)(\boldsymbol{W}_K E)^T}{\sqrt{d}}\right)\boldsymbol{W}_V E = \boldsymbol{AV} \quad (1)$$

where E_x is an intermediate representation of the denoising autoencoder and $\boldsymbol{W}_Q \in \mathbb{R}^{d\times d_x}$, $\boldsymbol{W}_K \in \mathbb{R}^{d\times d_e}$ and $\boldsymbol{W}_V \in \mathbb{R}^{d\times d_e}$ are embedding matrices. The background outside the mask area should be perfectly preserved for generative object compositing. Thus, we use the input mask M for blending the background image I_{bg} with the generated image I_{out}. As a result, the generator only denoises the masked area $I_{out} \bigotimes M$. To use this model on our task, a straightforward way (a baseline) would be to apply image captioning on the object image and feeding the resulting caption directly to the diffusion model as the condition. However, text embedding cannot capture fine grained visual details. Therefore, in order to directly leverage the given object image, we introduce the content adaptor to transform the visual features from a pretrained visual encoder to text features (tokens) to use as conditioning for the generator.

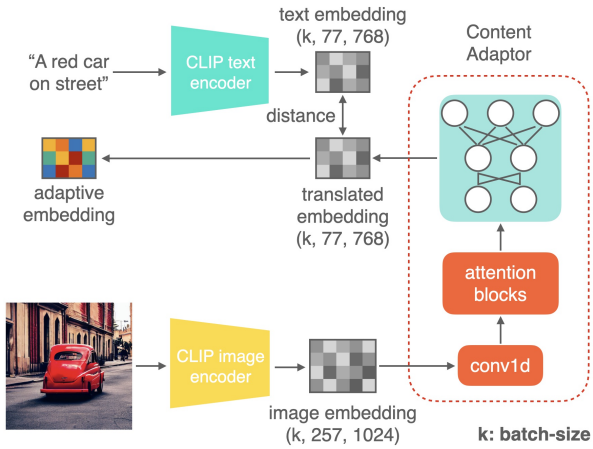

Figure 3. Structure of the Content Adaptor. In the first stage, it is trained on a large dataset of image-caption pairs to learn multi-modal sequential embeddings containing high-level semantics. In the second stage, it is fine-tuned under the diffusion framework to learn to encode identity features in *adaptive embedding*.

3.2. Content Adaptor

To prevent the loss of key identity information, we use an *image encoder* instead of a text encoder to produce the embedding from the input object image. However, the image embedding cannot be effectively utilized by the diffusion model for two reasons:

- The image embedding $\widetilde{E}$ and the text embedding E are from different domains. Since the diffusion model was trained on E, it cannot generate meaningful contents from image embedding sequence;
- A mismatch in the dimensions of $\widetilde{E} \in \mathbb{R}^{k\times 257\times 1024}$ and $E \in \mathbb{R}^{k\times 77\times 768}$, where k is batch-size.

Therefore, based on the above observations, we develop a sequence-to-sequence translator architecture as shown in Fig. 3. Given an image-caption pair as an input tuple (I, t), we employ two pretrained ViT-L/14 encoders C_t, C_i from CLIP [36] to produce text embedding $E = C_t(t)$ and image embedding $\widetilde{E} = C_i(I)$, respectively. The adaptor T

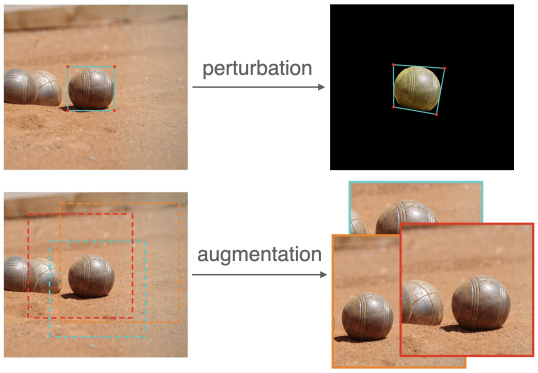

Figure 4. Illustration of our synthetic data generation and data augmentation scheme. The top row shows the data generation process including perspective warping, random rotation, and random color shifting. The original image is used as both the input background and ground truth, while the perturbed object is fed into the adaptor. The bottom row shows crop and shift augmentations, which help to improve the generation quality and preserve object details.

consists of three components: a 1D convolutional layer, attention blocks [48] and an MLP, where the 1D convolution modifies the length of the embedding from 257 to 77, the MLP maps the embedding dimension from 1024 to 768, and the attention blocks bridge the gap between text domain and image domain. We design a two-stage optimization method to train this module, which is explained in Sec. 3.3.2.

3.3. Self-supervised Framework

There is no publicly available image compositing training dataset with annotations which is sufficient for training a diffusion model, and it is extremely challenging to manually annotate such data. Therefore, we propose a self-supervised training scheme and a synthetic data generation approach that simulate real-world scenarios. We also introduce a data augmentation method to enrich the training data as well as improve the robustness of our model.

3.3.1 Data Generation and Augmentation

Training data generation. We collect our synthetic training data from Pixabay and use an object instance segmentation model [26] to predict panoptic segmentation and classification labels. We first filter the dataset by removing objects with very small or large sizes. Then, we apply spatial and color perturbations to simulate many real use-case scenarios where the input image and background image have different scene geometry and lighting conditions.

The top row of Fig. 4 illustrates this process. Inspired by [8], we randomly perturb the four points of the object bounding box to apply projective transformation, followed by a random rotation within the range $[-\theta, \theta]$ ($\theta = 20°$) and color perturbation. The segmentation mask (perturbed in the same way as the image) is used to extract the object.

This data synthesis pipeline is fully controllable and can also be employed as data augmentation during training. Another key advantage is that it is free of manual labeling, since the original image is used as the ground truth. We use the bounding box as the mask as it not only fully covers the object, but also extends to its neighboring area (providing room for shadow generation). We find that it is flexible enough for the model to apply spatial transformations, synthesize novel views and generate shadows and reflection.

Real-world evaluation data. Given that there is no existing dataset specifically for our task, we manually collect a new dataset, closely simulating real-world use cases, as an evaluation benchmark for object compositing. The dataset consists of 503 pairs of common objects (including both rigid and non-rigid objects such as vehicles and wildlife) and diverse background images (covering both indoor and outdoor scenes). It also contains challenging cases where there is a large mismatch of lighting conditions or viewpoints between foreground and background. The dataset images are collected from Pixabay. The labeling procedure closely simulate the real scenarios where the input object is placed at a target location in the background image and then scaled at the user's will. The compositing region is determined as a loose bounding box around the object.

Data augmentation. Inspired by [25], we introduce random shift and crop augmentations during training, while ensuring that the foreground object is always contained in the crop window. This process is illustrated in the bottom row of Fig. 4. Applying this augmentation method for both training and inference results in a notable improvement in the realism of the generated results. Quantitative results are provided in Sec. 4.4.

3.3.2 Training

Content adaptor pretraining. We first pretrain the content adaptor to keep the semantics of the object by mapping the image embedding to text embedding. At this first stage, we optimize the content adaptor on a sequence-to-sequence translation task, which learns to project the image embedding into a multi-modal space. During training, C_t, C_i are frozen, and the translator is trained on 3,203,338 image-caption pairs from a filtered LAION dataset [41].

Given the input image embedding $\widetilde{E}$, we use the text embedding E as target, the objective function of this translation task is defined as:

$$\mathcal{L}_{dist} = \|T(\tilde{E}) - E\|_1, \tag{2}$$

where $T(\cdot)$ is the content adaptor.

However, the multi-modal embedding obtained solely from this first stage optimization mostly carry high-level semantics of the input image without much of texture details. Hence, we further refine $\widetilde{E}$ to obtain $\hat{E}$, the adaptive embedding, for better appearance preservation.

Content adaptor fine-tuning. We further optimize the content adaptor to produce an adaptive embedding which maintains instance-level properties of the object. After pretraining, we insert the content adaptor to the *pretrained* diffusion framework by feeding the adaptive embedding as context to the attention blocks. Then, the diffusion model is frozen and the adaptor is trained using:

$$\mathcal{L}_{adapt} = \mathbb{E}_{T,\epsilon\sim\mathcal{N}(0,1)}[\|\epsilon - \epsilon_\theta(I_t \circ M, t, T(\widetilde{E}))\|_2^2], \quad (3)$$

where the content adaptor $T(\cdot)$ is optimized. It is trained on our synthetic dataset filtered from Pixabay, containing 467,280 foreground and background image pairs for training and 25,960 pairs for validation.

Generator fine-tuning. After the aforementioned two-stage training of the content adaptor is completed, we freeze the content adaptor and train the generator module after initializing the text-to-image diffusion model with pretrained weights. Crop and shift augmentations are applied in this process. Based on the Latent Diffusion model [38], the objective loss in our generator module is defined by:

$$\mathcal{L}_{gen} = \mathbb{E}_{\hat{E},\epsilon\sim\mathcal{N}(0,1)}[\|\epsilon - \epsilon_\theta(I_t \circ M, t, \hat{E})\|_2^2], \quad (4)$$

where I is the input image, I_t is a noisy version of I at timestep t, and ϵ_θ denotes the denoising model which is optimized. In order to adapt the text-guided generation task to the setting of image compositing, we apply the input mask on the image at every time-step.

4. Experiments

4.1. Training Details

The first training stage of the content adaptor takes 15 epochs with a learning rate of 10^{-4} and batch size of 2048. The image size processed by ViT is 224 $\times$ 224. To set up the pipeline, we employed the pretrained image encoder and text encoder from CLIP [36]. The second training stage of the content adaptor takes 13 epochs with the learning rate of 2×10^{-5} and batch size of 512. We resize the images to 512 $\times$ 512 and adapt the pretrained Stable Diffusion model [38] to our task. After the above process, the generator is trained for 20 epochs with the learning rate of 4×10^{-5} and batch size of 576. The input image size is the same as that in the second training stage. All training stages are conducted on 8 A100 GPUs with Adam optimizer.

4.2. Quantitative Evaluation

To obtain quantitative results on a real dataset that simulates real-world use cases, where users drag and drop an object to a background image, we conduct a user study to compare to three baselines, *i.e.*, copy-and-paste, pretrained stable diffusion model fine-tuned on BLIP [27], and a joint stable diffusion model integrating both BLIP and SDEdit [32]. Given the fact that there is no existing diffusion-based method addressing the same problem, we train the baselines on the same pretrained diffusion model using the same self-supervised training scheme and dataset as ours. To the best of our knowledge, after adapting to our setting, they are the closest baselines to our task. In the BLIP baseline, the content adaptor is replaced by BLIP, a state-of-the-art captioning model. The text embedding obtained from the BLIP caption is fed into the generator module. For the SDEdit baseline, we keep the BLIP model which will significantly improve the generation fidelity of a pretrained diffusion model. We set the noise strength to 0.65 in SDEdit, striking a balance between realism and faithfulness. Details are provided in the supplementary material.

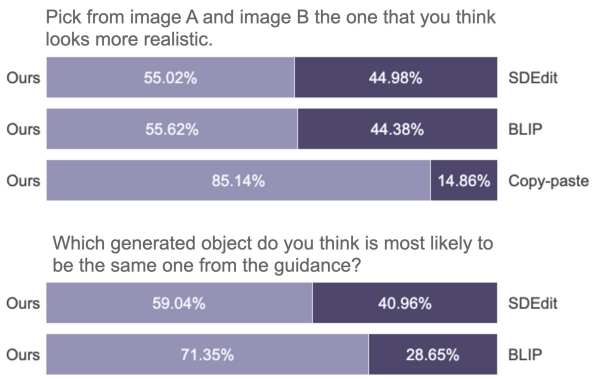

Figure 5. User study results. We conduct side-by-side comparisons between our method and one of baseline methods to quantify the generation quality in terms of realism and appearance preservation. The results show that our method outperforms baselines.

We perform a user study using our real testing dataset, which consists of 503 object-background pairs. For each example, we set up a side-by-side comparison between the results from one of the baselines and our method, respectively. The subject will be asked two questions: 1) which is more realistic and 2) which generated object is more close to the given object? We collected 1,494 votes from more than 170 users. As shown in Fig. 5, our method outperforms the other baselines for both questions. The higher preference rates demonstrate the effectiveness of the content adaptor and data augmentation method in improving generation fidelity. It also shows that our content adaptor is capable of maintaining details and attributes of the input objects, while not being constrained by the input when harmonizing objects to the background. In contrast, although SDEdit is better at preserving appearance than BLIP, it is constrained by the original pose, shape, and other attributes of the object, making the generated object less coherent with the background.

4.3. Qualitative Evaluation

In Fig. 6, we compare our method to various baselines. In the first two rows of Fig. 6, we place the same object

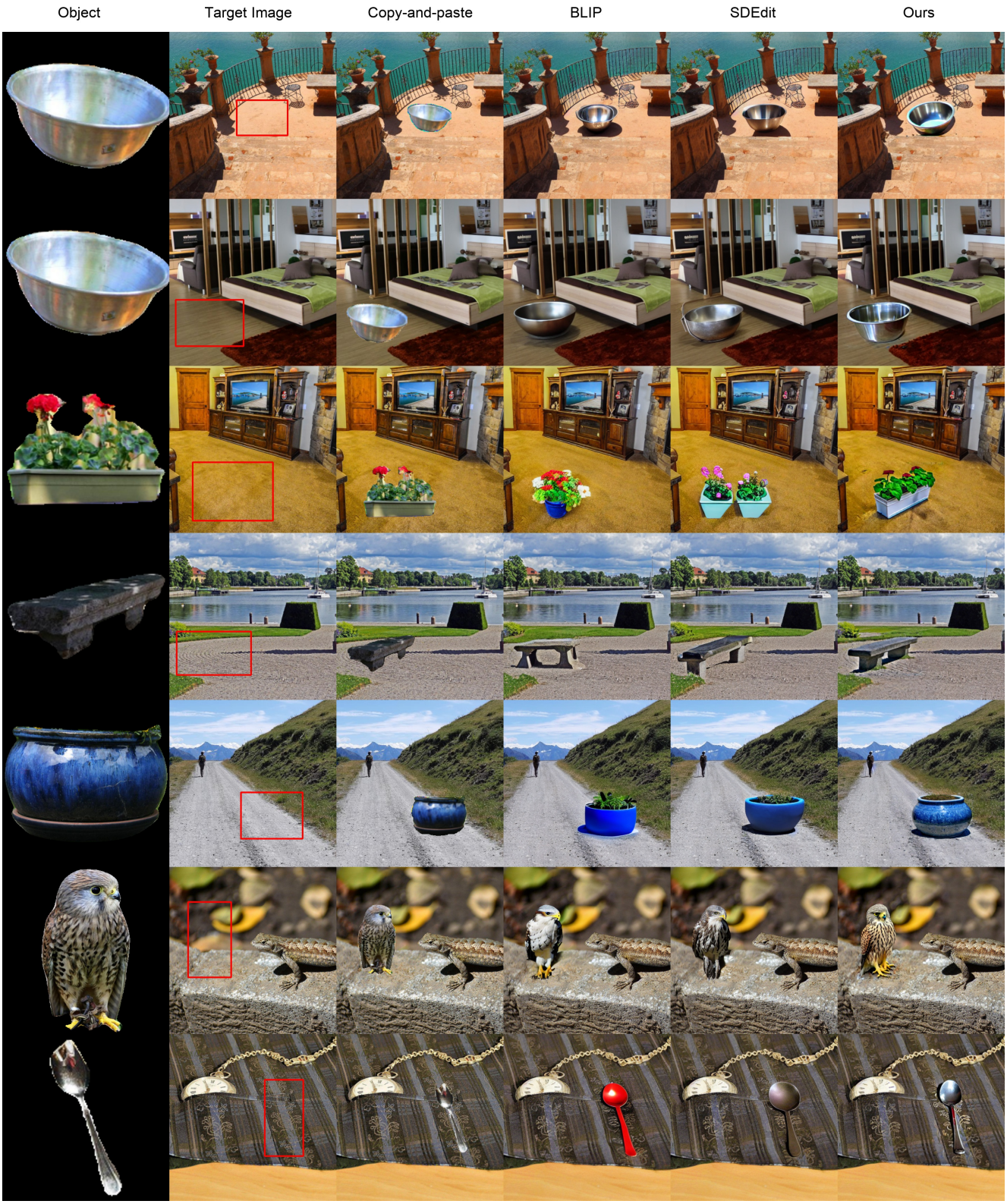

Figure 6. Qualitative comparison on the real-world evaluation dataset. Our method can appropriately correct the viewpoint and geometry while harmonizing the object with the new background, generating realistic composite results. Moreover, lighting and shadows are naturally adjusted in the composite image. Compared to baselines using text guidance, our method can better preserve the object appearance.

Adaptor	Optimization	Augmentation	BLIP [27]	FID ↓	CLIP text score ↑	CLIP image score ↑
✓	✓	✓	✗	**15.4331**	**29.8594**	**97.0625**
✗	✓	✓	✗	18.8254	29.8281	96.0000
✓	✗	✓	✗	16.1381	29.7969	96.8125
✓	✓	✗	✗	18.3698	29.7813	96.1875
✗	✗	✓	✓	17.9000	29.7188	95.8125

Table 2. Ablation study. We evaluate the effectiveness of the components listed in the first four columns: 1) whether the content adaptor is used; 2) whether the adaptor is optimized for appearance preservation; 3) whether crop augmentation is utilized in training; 4) if the adaptor module is replaced by BLIP to predict text embedding from an image. We use FID to assess the generation fidelity and modified CLIP score (explained in Sec. 4.4) to measure similarity of the guidance and the predicted image.

in two scenes with notable differences in the geometry. Results show that our model is capable of correcting geometric inconsistencies while preserving the characteristics of the object. The limitations of the baselines are also shown in Fig. 6: BLIP often fails to preserve the appearance; SDEdit preserves texture and pose better than the first baseline, but often cannot make appropriate geometry transformations.

4.4. Ablation Study

In Tab. 2, we demonstrate the effectiveness of each key component in our ObjectStitch framework by ablating each of them. We use FID [13] and a modified CLIP [12, 36] score, which measure the fidelity and semantic matching performance between given and generated objects. Instead of directly adopting the CLIP score, which semantically matches the prompt and the image, we modify it to *CLIP text score* and *CLIP image score* that are defined as follows:

$$\mathcal{C}_{txt} = E\left[s \cdot f(I_{pred}) \cdot g\left(B(I_{gt})\right)\right], \tag{5}$$

$$\mathcal{C}_{img} = E\left[s \cdot f(I_{pred}) \cdot f(I_{gt})\right], \tag{6}$$

where $B(\cdot)$ is pretrained BLIP [27], and s is a logit scale.

For the ablation models, we first remove the content adaptor to evaluate its effect. In order to keep the embedding dimension fed to the generator, we removed the attention blocks from the content adaptor. We also drop the bias and activation function of the last linear layer, such that in the adaptation process, the model will only learn a non-adaptive embedding dictionary. Without the attention layers, the model has difficulty bridging the domain gap between text and image, resulting in the degradation of generation fidelity as shown in the second row of Tab. 2.

The third row in Tab. 2 ablates the second training stage of the content adaptor by simply skipping it. Without this process, the model fails to learn descriptive embedding, which encodes instance-level features of the guidance object, leading to a drop in appearance preservation.

Data augmentation is another key element for better performance, *i.e.*, cropping and shift augmentation increases the mask area, thus improving the generation quality in detail. The fourth row in Tab. 2 illustrates that with the absence of data augmentation, there is a notable drop in realism, and the generation carries less accurate details.

To further assess the importance of the content adaptor module, we replace it with a pretrained BLIP model, which transfers the reference image to a text embedding. As shown in the last row in Tab. 2, this results in the lowest CLIP image score, indicating that high-level semantics represented by text are not sufficient for maintaining object identity, highlighting the importance of our content adaptor.

5. Conclusion and Future Work

We propose the first diffusion-based approach to tackle object compositing. Traditional methods require many steps including harmonization, geometry/view adjustment, and shadow generation. By contrast, our method directly achieves realistic compositing. We construct our model from a text-to-image generation network and introduce a novel content adaptor module. Furthermore, we present a fully self-supervised framework to train our model with synthetic data generation and augmentation. We show our superior performance over baseline methods through user studies on real-world examples.

Since we give the first attempt to handle this challenging task, our method has several limitations. It lacks control over the appearance preservation of the synthesized object. Potential solutions are training the visual encoder on pairs of images with the same object, or training it jointly with the UNet. Another limitation is caused by masking the output image, which prohibits our model from generating global effects, *i.e.*, the shadow can only be synthesized within the mask. To address this issue, we may need to synthesize complete background images and pair them with objects, and train with a conditional generation framework. We also observed artifacts around the object in the background. This can be improved by training the model to also predict an instance mask [49]. We will improve this in our future work.

Acknowledgements

This work is partially supported by NSF grants 1835739 and 2106717. We also thank Zhiquan Wang for discussion.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2, 3

[2] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. Compositional gan: Learning image-conditional binary composition. *International Journal of Computer Vision*, 128(10):2570–2585, 2020. 2

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[5] Bor-Chun Chen and Andrew Kae. Toward realistic image compositing with adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2019. 2, 3

[6] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Bargainnet: Background-guided domain translation for image harmonization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 1, 2, 3

[7] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8394–8403, 2020. 1, 2, 3

[8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016. 5

[9] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. *arXiv preprint arXiv:2210.15257*, 2022. 2, 3

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[11] Liu He, Yijuan Lu, John Corring, Dinei Florencio, and Cha Zhang. Diffusion-based document layout generation. *arXiv preprint arXiv:2303.10787*, 2023. 2

[12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 8

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 8

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3

[15] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. Deep outdoor illumination estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7312–7321, 2017. 2

[16] Yan Hong, Li Niu, and Jianfu Zhang. Shadow generation for composite image in real-world scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 914–922, 2022. 1, 2, 3

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 2

[19] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. Ssh: A self-supervised framework for image harmonization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4832–4841, 2021. 1, 2, 3

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[21] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (TOG)*, 30(6):1–12, 2011. 2

[22] Kevin Karsch, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, Hailin Jin, Rafael Fonte, Michael Sittig, and David Forsyth. Automatic scene inference for 3d object compositing. *ACM Transactions on Graphics (TOG)*, 33(3):1–15, 2014. 2

[23] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 2, 3

[24] Jean-Francois Lalonde and Alexei A Efros. Using color compatibility for assessing image realism. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 3

[25] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021. 5

[26] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13906–13915, 2020. 5

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 6, 8

[28] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative

adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9455–9464, 2018. 1, 2

[29] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8139–8148, 2020. 1, 2

[30] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. *arXiv preprint arXiv:2112.05744*, 2021. 2, 3

[31] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 2

[32] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2, 3, 6

[33] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 3

[34] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003. 1

[35] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 4, 6, 8

[37] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 6

[39] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 2, 3

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3

[41] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5

[42] Yichen Sheng, Yifan Liu, Jianming Zhang, Wei Yin, A Cengiz Oztireli, He Zhang, Zhe Lin, Eli Shechtman, and Bedrich Benes. Controllable shadow generation using pixel height maps. In *European Conference on Computer Vision*, pages 240–256. Springer, 2022. 1

[43] Yichen Sheng, Jianming Zhang, and Bedrich Benes. Ssn: Soft shadow network for image compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4380–4390, 2021. 1, 2, 3

[44] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[45] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[46] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. *ACM Transactions on Graphics (TOG)*, 29(4):1–10, 2010. 3

[47] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3789–3797, 2017. 1

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[49] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. *arXiv preprint arXiv:2212.05034*, 2022. 8

[50] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2970–2979, 2017. 2

[51] Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. *arXiv preprint arXiv:2207.04788*, 2022. 1, 2

[52] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7194–7202, 2019. 1