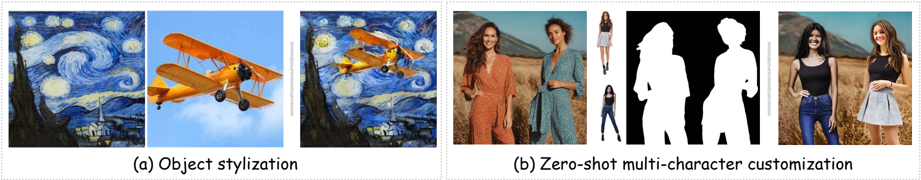
FreeCompose: Generic Zero-Shot Image Composition with Diffusion Prior (Supplementary Material)

Zhekai Chen[1*], Wen Wang[1,2*], Zhen Yang[1], Zeqing Yuan[1],
Hao Chen[1], and Chunhua Shen[1,2]

[1] Zhejiang University, China [2] Ant Group

Appendix

A More Applications



(a) Object stylization

(b) Zero-shot multi-character customization

Fig. A1: Other applications of FreeCompose, including (a) object stylization and (b) zero-shot multi-character customization.

Object stylization. During the image harmonization phase, the default prompts do not favor any particular style. However, if an object is composed onto a background that differs in style (for example, from a real plane to an oil-painting background as shown in Figure A1), these prompts can be used to transfer the object to match the style of the background.

Zero-shot multi-character customization. Animate Anyone [2] is a method that customizes images into videos by allowing zero-shot customization of a single character with a similar background. With the implementation of this method, it becomes possible to compose multiple customized characters together, thus enabling zero-shot multi-character customization.

B More Implementation Details

Optimization Steps. The best results in different cases are achieved through various optimization steps. Generally, we use 150 steps for object removal and 200 steps for image harmonization. However, for semantic image composition, the specific format of the conditions requires different numbers of steps. For instance, text requires 500 steps, while sketch and canny require 200 steps.

[*] Equal contribution. HC and CS are the corresponding authors. Part of this work was done when WW was an intern at Ant Group.

Timestep Choice. According to our observations, different timesteps have varying levels of influence on the optimization results. During the object removal phase, we use timesteps ranging from 50 to 400 to enhance efficiency. For the image harmonization phase, timesteps between 50 and 950 are employed to achieve a more balanced outcome. In the semantic image composition phase, timesteps between 50 and 100 are used specifically for the final fifty optimization steps to ensure smoothness in the resulting image..

T2I-Adapter Model. We utilize the T2I-Adapter, which was released by TencentARC[1], to apply the diffusion model to conditions in formats other than text. When it comes to image composition, sketch and canny are conditions more suitable than other formats, used for cases in our results.

Running Times. The running times depend on the optimization steps chosen for a specific task. In general, when using an RTX 3090 with a float 16 precision, the first 50 steps take approximately 30 seconds, including preparation time for each phase. Subsequent sets of 50 steps take around 25 seconds.

C More Results

C.1 Object Removal

We show some more object removal results in Figure A2. Our method can be widely applied to different types of objects and scenes, and can achieve good results in most cases.

Methods	LPIPS↓	SSIM↑	MSE↓
LaMa [6]	0.0133	0.9849	37.73
SD-inpainting [4]	0.1733	0.7639	372.01
Ours	0.1120	0.7882	293.46

Table A1: Object removal results.

In addition, **for object removal**, we follow Lama [6] to randomly sample 2100 images from the Places2 dataset and paste objects to produce fake images for object removal. As shown in Tab. A1, LaMa has superb results because it is trained on training split of Places. When compared with SD-inpainting [4], our method surpasses it in all scores.

C.2 Image Harmonization

We show some more image harmonization results in Figure A3. Our method automatically analyze the light and shadow of the environments and harmonize the object accordingly.

[1] https://huggingface.co/TencentARC

Methods	CLIP$_{fg}$ ↑	Dino$_{fg}$ ↑	CLIP$_{bg}$ ↑	Dino$_{bg}$ ↑	FID↓	QS↑
ObjectStitch [5]	72.13	67.14	79.05	87.99	34.71	35.82
ControlCom [8](har)	**73.35**	**68.59**	79.63	88.25	33.41	37.10
ControlCom(com)	70.45	66.37	**81.65**	**89.01**	**30.05**	**42.28**
Diff-harmonization [1]	71.96	68.20	75.31	87.82	38.13	20.82
Ours	72.52	67.93	78.40	87.31	33.42	37.08

Table A2: Composition results. The gray ones require training.

Moreover, **for image harmonization and semantic image composition** , since both image harmonization and semantic image composition focus on composing multiple images into a coherent one, we evaluate the two tasks using the same data and metrics, following Controlcom [8]. Specifically, we choose 80 backgrounds form COCO and 30 objects from DreamBooth dataset, composing 2400 cases for evaluation. We calculate CLIP Score and DINO Score with the foreground and the background images, FID with the background images, and QS with COCO2017 as provided by Paint-by-Example [7]. CLIP and DINO score assess the consistency between original images and composed images, FID and QS estimates the quality of images.

C.3 Semantic Image Composition

We show some more semantic image composition results in Figure A4. Our method enables the use of various conditions as guidance to guide the composition process. In cases where more intricate texture or structure is desired, canny edges can be employed as conditions to achieve superior outcomes, as demonstrated in the right column.

D Plug-and-Play On other Diffusion Models

Plug-and-Play on SDXL Model. We apply FreeCompose to a pre-trained SDXL model[2]., and the results are displayed in Figure A5. Thanks to the exceptional prior of the SDXL model, the results are particularly impressive, especially in terms of image harmonization. As shown in the right column, it can be observed that the bottle's reflection on the table in the first case and the object's shadow in the second case are well integrated with the background through our method.

E Algorithm

E.1 Object Removal

The pseudocode for our method in object removal phase is shown in Algorithm 1. The critical part is the calculation of the mask guided loss, which uses the mask for discarding semantic message during denoising of the target image.

[2] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

Algorithm 1 Object Removal

```
Require: Image $I$, Mask $M$, Diffusion model with parameter $\theta$, Optimization steps $S$. Get text embeddings $c_u, c_s, c_t$ from empty prompt and default prompts $P_s, P_t$. Given upper bound $t_{\max}$ and lower bound $t_{\min}$ for timesteps, guidance scale $w$, loss weight $\lambda_{\text{per}}$ and learning rate $\eta$.
$M' \leftarrow 1 - M$
$\hat{\mathbf{z}} \leftarrow \text{Encode}(I)$
$\mathbf{z} \leftarrow \hat{\mathbf{z}}$
for $s = 1, 2, \dots, S$ do
    $t \leftarrow \text{random}(t_{\min}, t_{\max})$
    $\alpha_t \leftarrow \text{scheduler}(t)$
    $\epsilon \leftarrow \mathcal{N}(0, \mathbf{I})$
    $\mathbf{z_t}, \hat{\mathbf{z_t}} \leftarrow \sqrt{\alpha_t}\mathbf{z} + \sqrt{1-\alpha_t}\epsilon, \sqrt{\alpha_t}\hat{\mathbf{z}} + \sqrt{1-\alpha_t}\epsilon$    ▷ random noise the latent
    $\epsilon_{us}, \epsilon_{cs} \leftarrow \epsilon_\theta(\hat{\mathbf{z}}, t, c_u), \epsilon_\theta(\hat{\mathbf{z}}, t, c_s)$
    $\epsilon_{ut}, \epsilon_{ct} \leftarrow \epsilon_\theta(\mathbf{z}, t, c_u, M), \epsilon_\theta(\mathbf{z}, t, c_s, M)$    ▷ mask guided calculation
    $\epsilon_s, \epsilon_t \leftarrow \epsilon_{us} + w(\epsilon_{cs} - \epsilon_{us}), \epsilon_{ut} + w(\epsilon_{ct} - \epsilon_{ut})$
    $\mathcal{L} \leftarrow ||\epsilon_s - \epsilon_t||_2^2 + \lambda_{\text{per}}\mathcal{L}_{\text{per}}(I \otimes M', \text{Decode}(\mathbf{z}) \otimes M')$
    $\mathbf{z} \leftarrow \mathbf{z} - \eta\nabla_{\mathbf{z}}\mathcal{L}$
end for
return The background image $\text{Decode}(\mathbf{z})$
```

E.2 Image Harmonization

The pseudocode for our method in image harmonization is presented in Algorithm 2. This section balances various losses to find a tradeoff between object identity, background features, and overall harmony.

E.3 Semantic Image Composition

The pseudocode for our method in semantic image composition is demonstrated in Algorithm 3. The key aspect is the utilization of condition features to guide the transformation and the replacement of the self-attention features of the target image, which forms the core of the semantic image composition phase.

F Discussion

F.1 Limitations

The first limitation concerns the object removal phase. Through the use of mask guided loss, the pipeline replaces the semantic information of the object with that of the background. However, if the mask is too large, the remaining background information may not be enough to accurately reconstruct the entire background, leading to the creation of artifacts. Additionally, it is important that the mask fully covers the object to be removed; otherwise, certain portions of the object may still be visible in the final result. In situations where there are similar objects

Algorithm 2 Image Harmonization

```
Require: Image $I$, Mask $M$, Diffusion model with parameter $\theta$, Optimization steps
    $S$. Get text embeddings $c_u, c_s, c_t$ from empty prompt and default prompts $P_s, P_t$.
    Given upper bound $t_{\max}$ and lower bound $t_{\min}$ for timesteps, guidance scale $w$, loss
    weight $\lambda_{\text{bak}}, \Lambda_{\text{for}}$ and learning rate $\eta$.
$M' \leftarrow 1 - M$
$\hat{\mathbf{z}} \leftarrow \text{Encode}(I)$
$\mathbf{z} \leftarrow \hat{\mathbf{z}}$
for $s = 1, 2, \ldots, S$ do
    $t \leftarrow \text{random}(t_{\min}, t_{\max})$
    $\alpha_t \leftarrow \text{scheduler}(t)$
    $\epsilon \leftarrow \mathcal{N}(0, \mathbf{I})$
    $\mathbf{z_t}, \hat{\mathbf{z_t}} \leftarrow \sqrt{\alpha_t}\mathbf{z} + \sqrt{1-\alpha_t}\epsilon, \sqrt{\alpha_t}\hat{\mathbf{z}} + \sqrt{1-\alpha_t}\epsilon$    ▷ random noise the latent
    $\epsilon_{us}, \epsilon_{cs} \leftarrow \epsilon_\theta(\hat{\mathbf{z}}, t, c_u), \epsilon_\theta(\hat{\mathbf{z}}, t, c_s)$
    $\epsilon_{ut}, \epsilon_{ct} \leftarrow \epsilon_\theta(\mathbf{z}, t, c_u), \epsilon_\theta(\mathbf{z}, t, c_s)$
    $\epsilon_s, \epsilon_t \leftarrow \epsilon_{us} + w(\epsilon_{cs} - \epsilon_{us}), \epsilon_{ut} + w(\epsilon_{ct} - \epsilon_{ut})$
    $\mathcal{L}_{\text{bak}} \leftarrow \mathcal{L}_{\text{per}}(I \otimes M', \text{Decode}(\mathbf{z}) \otimes M')$
    $\mathcal{L}_{\text{for}} \leftarrow \mathcal{L}_{\text{per}}(I \otimes M, \text{Decode}(\mathbf{z}) \otimes M)$
    $\mathcal{L} \leftarrow ||\epsilon_s - \epsilon_t||_2^2 + \lambda_{\text{bak}}\mathcal{L}_{\text{bak}} + \lambda_{\text{for}}\mathcal{L}_{\text{for}}$    ▷ compose all loss
    $\mathbf{z} \leftarrow \mathbf{z} - \eta\nabla_{\mathbf{z}}\mathcal{L}$
end for
return The harmonized image Decode($\mathbf{z}$)
```

present in the background, the pipeline may mistakenly replace the removed object with these similar objects, as they share a similar semantic message.

The second limitation pertains to the image harmonization phase. Although the pipeline achieves excellent results in terms of light and shadow, it struggles to strike a balance between the object's features and the overall naturalness when there is a significant contrast between the object and the background. For instance, when dealing with an object that has dark shadows against a bright background.

The third limitation relates to the semantic image composition phase. The quality of the output is partially influenced by the format and quality of the input conditions. When it comes to text prompts, the pipeline can only generate subtle variations. As for sketches, certain details are challenging to render realistically. Canny edges appear to be the most suitable format for conditions, but they are less accessible and more intricate.

F.2 Future Work

FreeCompose enables flexible composition among different objects and backgrounds by utilizing pre-trained diffusion models, without the need for additional training. In the future, we plan to expand our method to cover more composition tasks and further explore the potential of the pipeline. We also intend to investigate the feasibility of applying our method to video models and other generative

Algorithm 3 Semantic Image Composition

```
Require: Image $I$, Mask $M$, Diffusion model with parameter $\theta$, Optimization steps $S$
    and $\tau_s, \tau_l$ for restriction of step and layer to begin replacing. Get text embeddings
    $c_u, c_s, c_t$ from empty prompt and prompts $P_s, P_t$, and features $f_s, f_t$ from conditions
    $C_s, C_t$ through pre-trained T2I-Adapters. Given upper bound $t_{\max}$ and lower bound
    $t_{\min}$ for timesteps, guidance scale $w$ and learning rate $\eta$.
$M' \leftarrow 1 - M$
$\hat{\mathbf{z}} \leftarrow \text{Encode}(I)$
$\mathbf{z} \leftarrow \hat{\mathbf{z}}$
for $s = 1, 2, \ldots, S$ do
    $t \leftarrow \text{random}(t_{\min}, t_{\max})$
    $\alpha_t \leftarrow \text{scheduler}(t)$
    $\epsilon \leftarrow \mathcal{N}(0, \mathbf{I})$
    $\mathbf{z_t}, \hat{\mathbf{z_t}} \leftarrow \sqrt{\alpha_t}\mathbf{z} + \sqrt{1-\alpha_t}\epsilon, \sqrt{\alpha_t}\hat{\mathbf{z}} + \sqrt{1-\alpha_t}\epsilon$    ▷ random noise the latent
    $\epsilon_{us}, \{Q_{us}, K_{us}, V_{us}\} \leftarrow \epsilon_\theta(\hat{\mathbf{z}}, t, c_u; f_s)$    ▷ use condition features
    $\epsilon_{cs}, \{Q_{cs}, K_{cs}, V_{cs}\} \leftarrow \epsilon_\theta(\hat{\mathbf{z}}, t, c_s; f_s)$
    if $s > \tau_s$ then
    ▷ use condition features and replace self-attention features of layer index $l > \tau_l$
        $\epsilon_{ut} \leftarrow \epsilon_\theta(\mathbf{z}, t, c_u; f_t, \{Q_{us}, K_{us}, V_{us}\})$
        $\epsilon_{ct} \leftarrow \epsilon_\theta(\mathbf{z}, t, c_s; f_t, \{Q_{cs}, K_{cs}, V_{cs}\})$
    else
        $\epsilon_{ut}, \epsilon_{ct} \leftarrow \epsilon_\theta(\mathbf{z}, t, c_u; f_t), \epsilon_\theta(\mathbf{z}, t, c_s; f_t)$
    end if
    $\epsilon_s, \epsilon_t \leftarrow \epsilon_{us} + w(\epsilon_{cs} - \epsilon_{us}), \epsilon_{ut} + w(\epsilon_{ct} - \epsilon_{ut})$
    $\mathcal{L} \leftarrow ||\epsilon_s - \epsilon_t||_2^2$
    $\mathbf{z} \leftarrow \mathbf{z} - \eta \nabla_{\mathbf{z}} \mathcal{L}$
end for
return The composed image $\text{Decode}(\mathbf{z})$
```

models. Additionally, we will improve the user-friendliness and efficiency of the pipeline in future updates.

F.3 Negative Impact

Our FreeCompose aims to utilize the prior knowledge of pre-trained diffusion models and extend their use to tasks beyond their original purpose. However, it is important to acknowledge the potential for malicious applications of our method, such as generating deceptive images that composing real individuals with fabricated surroundings for the purpose of misinformation and disinformation. This is a common issue with generative models.

One possible way to address the negative impact is to adopt methods similar to that proposed by Pham et al. [3]. These methods leverage the capability of diffusion models to identify fake images and help prevent the abuse of our method. Furthermore, it is crucial to be mindful of employing unseen watermarks and other techniques to authenticate images in order to prevent the misuse of our method.

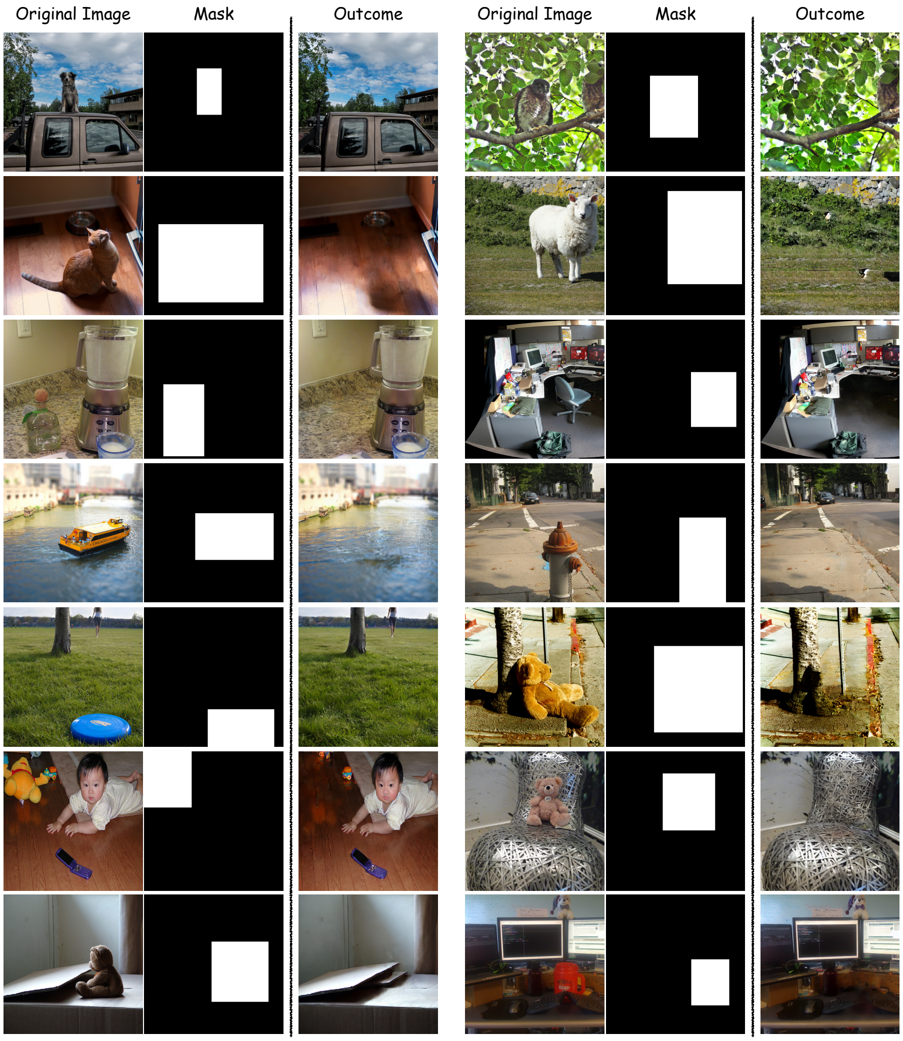


Fig. A2: More object removal results. We show more object removal results in this figure. The first column is the original image, the second column is the mask of the object to be removed, the third column is the result of the object removal.

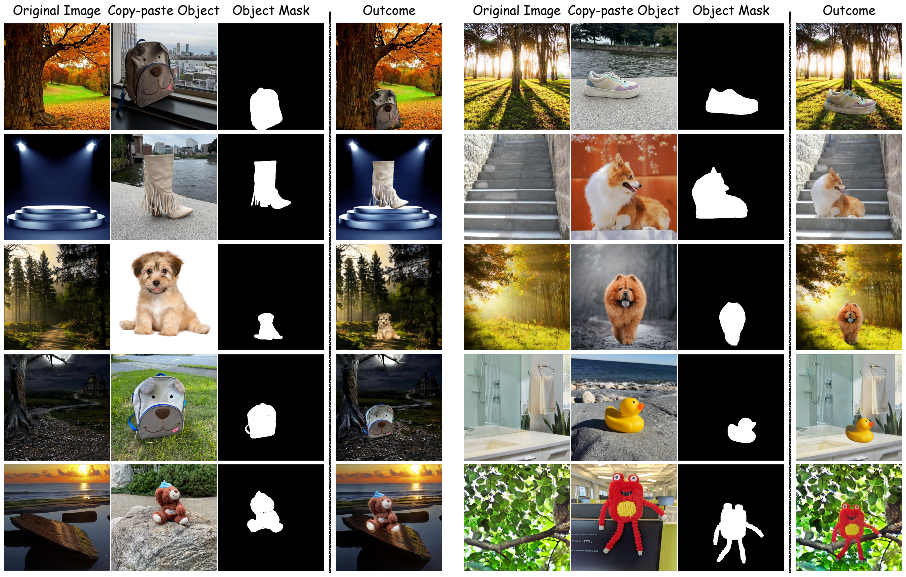


Fig. A3: More image harmonization results. We show more image harmonization results in this figure. The first column is the original background image, the second column is the object to be pasted and harmonized, the third column is the mask of the object after being pasted, and the fourth column is the result of the image harmonization.

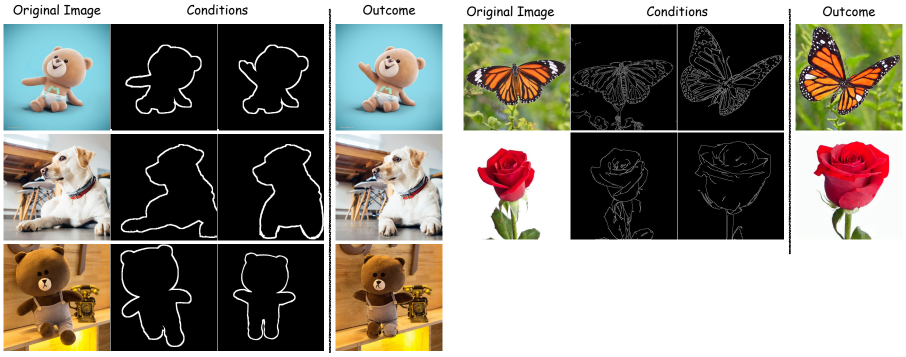


Fig. A4: More semantic image composition results. We show more semantic image composition results in this figure. The left is the cases using sketches as conditions and the right is the cases using canny edges as conditions. For each side, the first column is the original image, the second column and the third column are the corresponded condition for original image and the condition for target image, and the fourth column is the result of the semantic image composition.

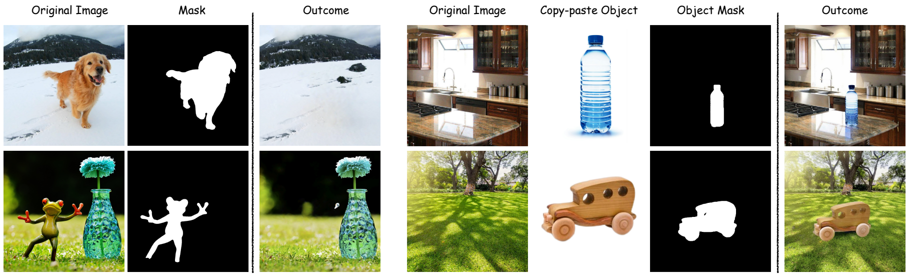


Fig. A5: FreeCompose on pre-trained SDXL. We apply our method on pre-trained SDXL model. The left column is the results of object removal and the right column is the results of image harmonization.

References

1. Chen, J., Zou, Z., Zhang, Y., Chen, K., Shi, Z.: Zero-shot image harmonization with generative model prior. arXiv preprint arXiv: 2307.08182 (2023)
2. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv: 2311.17117 (2023)
3. Pham, M., Marshall, K.O., Cohen, N., Mittal, G., Hegde, C.: Circumventing concept erasure methods for text-to-image generative models. In: The Twelfth International Conference on Learning Representations (2024)
4. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. Computer Vision and Pattern Recognition (2021). `https://doi.org/10.1109/CVPR52688.2022.01042`
5. Song, Y., Zhang, Z., Lin, Z., Cohen, S., Price, B., Zhang, J., Kim, S.Y., Aliaga, D.: Objectstitch: Object compositing with diffusion model. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
6. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. arXiv preprint arXiv: 2109.07161 (2021)
7. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
8. Zhang, B., Duan, Y., Lan, J., Hong, Y., Zhu, H., Wang, W., Niu, L.: Controlcom: Controllable image composition using diffusion model. arXiv preprint arXiv: 2308.10040 (2023)