Paint by Example: Exemplar-based Image Editing with Diffusion Models

A. Additional results

In this section, we provide additional results of our method in different application scenarios and results from an ablation study on mask shape augmentation. Fig. 1 demonstrates the ability of our method in editing any region of real images. Our method is able to understand the objects in the reference images and generate corresponding objects in the edited region. The generated objects are highly in harmony with the source images. In Fig. 2, we show results of the same object with different source images, which demonstrates the robustness of our method. The results in Fig. 3 show that our model can produce plausible outputs given arbitrary mask shapes. We also conduct an ablation study on mask shape augmentation. From the results in Fig. 4, we can see that the model fails to generalize to arbitrary masks without this technique. More visual results are shown in Fig. 5.

Figure 1. Our method enables the user to edit different regions in the same source images.

Figure 2. Results of the same object with different source images. Our method is robust for different objects or different source images, even for some complicated objects, like 'Big Ben'.

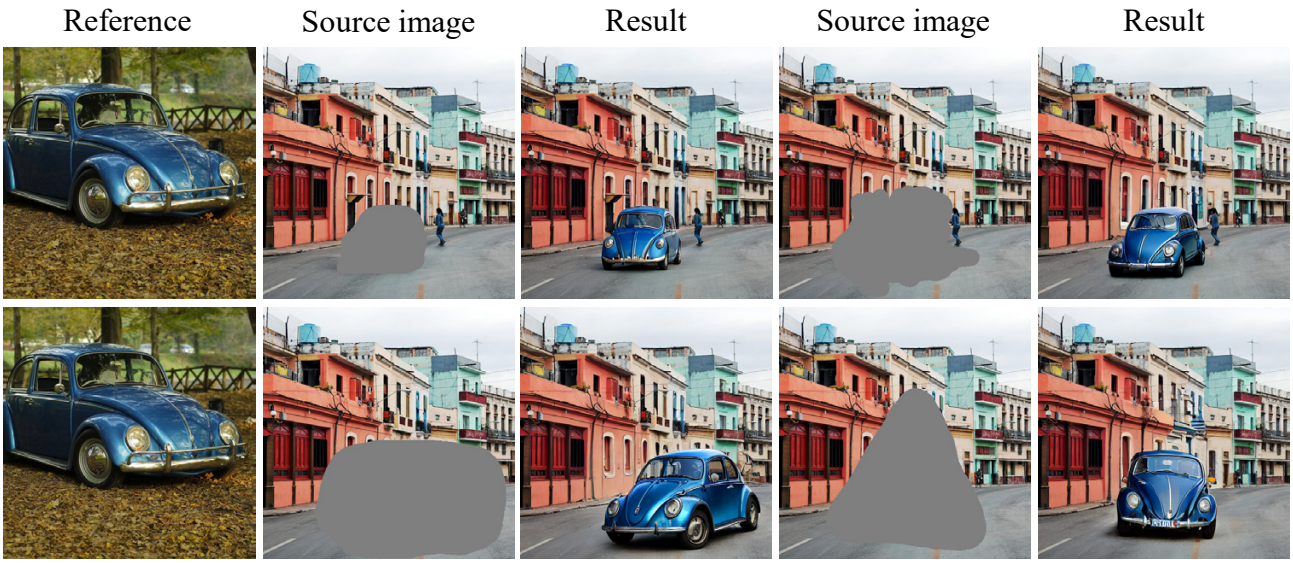

Figure 3. The impact of mask shape on the final results.

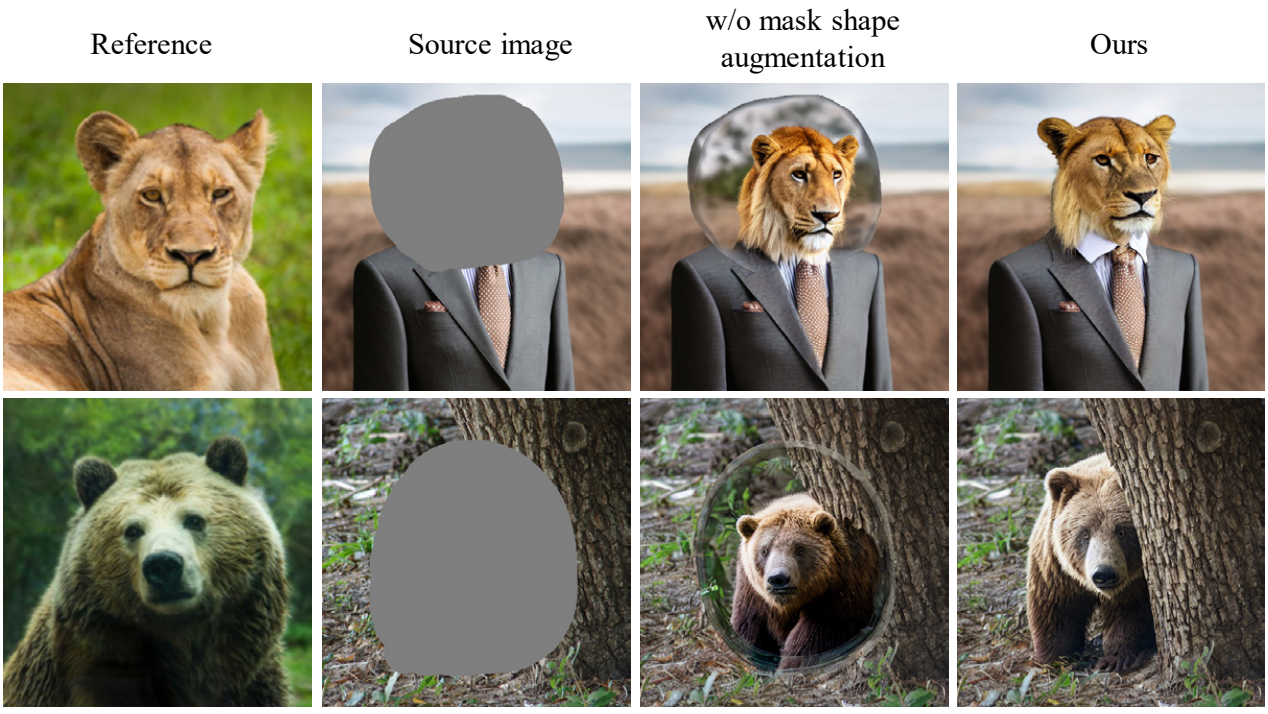

Figure 4. Importance of mask shape augmentation.

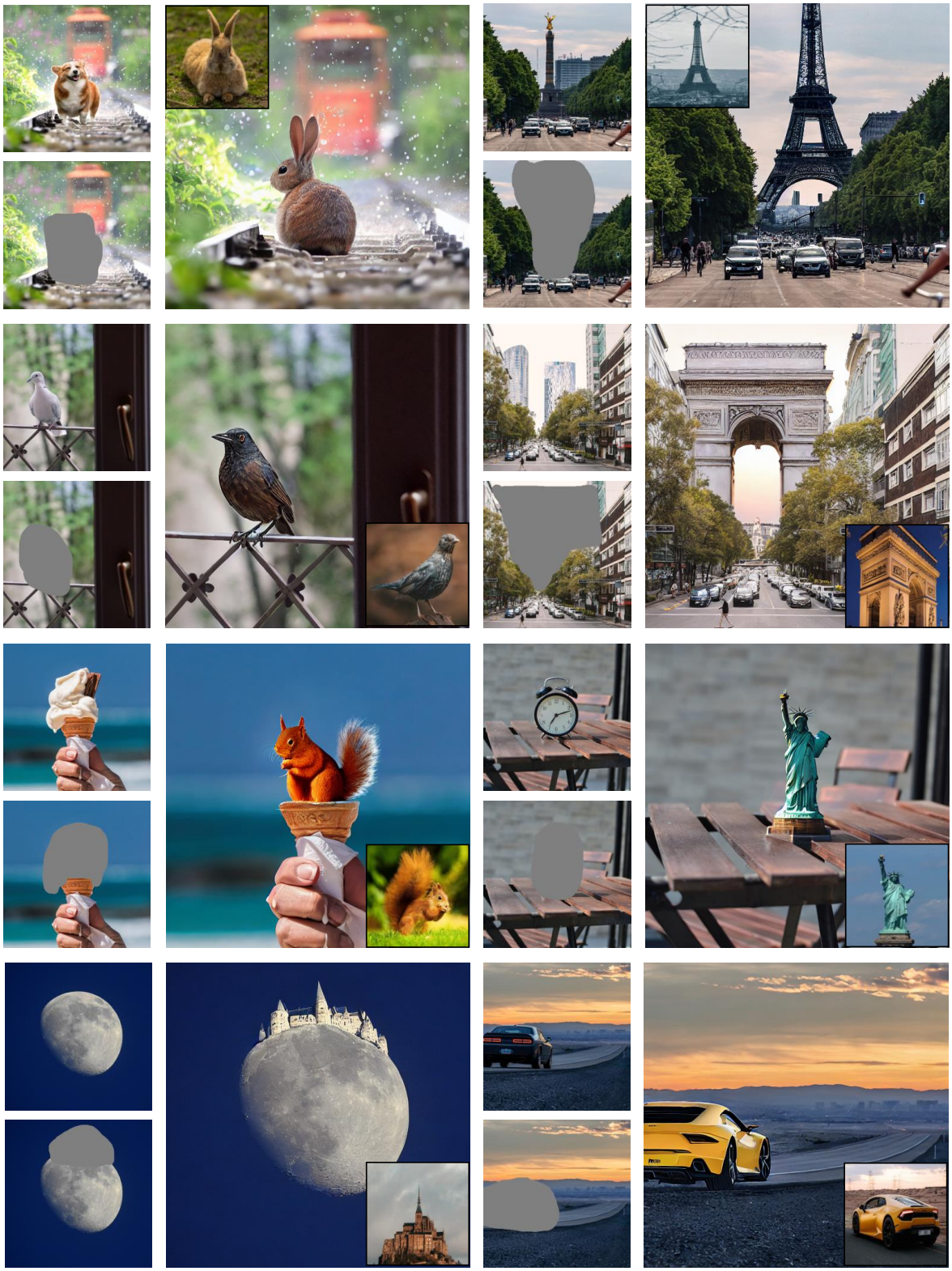

Figure 5. Additional visual results.

B. Limitation and hard-case analysis

Our algorithm achieves an impressive performance on in-the-wild images. Such as the result in Fig. 6(a), we manage to put a car on the head, showing great power to synthesize rare images. However, it struggles with some hard cases. Specifically, b) we tried the reference image with multiple instances, but sometimes the output may have a mismatched count of elements. c) When the reference image is not a close-up of a singe object, our method can still produce meaning result by properly compositioning the concepts. d) The result is robust to the super-large mask with arbitrary shape, though the model tends to fill the mask area and leads to oversized effect. e) For some rarer objects like dinosaur, our method can hardly understand them well. f) Because the majority of the training data is natural photos, our method does not perform well with some artificial images, such as oil paintings.

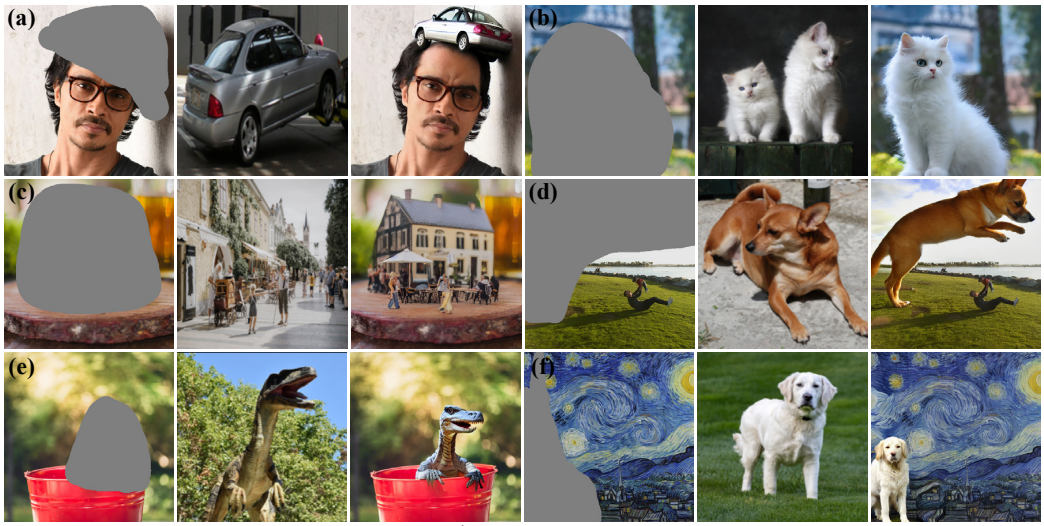

Figure 6. Some failure cases.

C. Implementation details

We adopt the Stable Diffusion [4] as our baseline model and choose their publicly released v1-4 model for text-driven image generation as initialization. First, We modify it to a text-driven image inpainting model by expanding 5 additional channels of the first convolution layer in the U-net (4 represents the encoded masked-image and 1 for the mask region). The new added weights are zero-initialized. We choose CLIP [3] pretrained model (ViT-L) as our image encoder and choose its feature from the last hidden state as condition. We utilize 15 fully-connected (FC) layers to decode the feature from pretrained encoder and inject it into the diffusion process through cross attention. We train the model using exponential moving average of weights and AdamW [2] optimizer with a constant learning rate of $1e-5$. We use the HorizontalFlip ($p = 0.5$), Rotate ($limit = 20$), Blur ($p = 0.3$) and ElasticTransform ($p = 0.3$) from Albumentations [1] for image augmentation. To quantitatively compare different settings, we adopt the CLIP image encoder (ViT-B) as feature extractor for FID, QS and CLIP Score. For the comparison with Stable Diffusion [4], we utilize their officially released code and pretrained text-driven inpainting model for testing (v1-5-inpainting). For Blended Diffusion (image), we utilize CLIP (ViT-B) image encoder for encoding the reference images, which corresponds to its text encoder.

References

[1] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. 5

[2] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[3] 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. 5

[4] 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. 5