Supplementary Materials for CaricatureBooth: Data-Free Interactive Caricature Generation in a Photo Booth

Zhiyu Qu[1] Yunqi Miao[2] Zhensong Zhang[3] Jifei Song[1] Jiankang Deng[4] Yi-Zhe Song[1]

[1]SketchX, CVSSP, University of Surrey [2]University of Warwick
[3]Independent Researcher [4]Imperial College London

{z.qu, y.song}@surrey.ac.uk Yunqi.Miao.1@warwick.ac.uk j.deng16@imperial.ac.uk

1. The necessity of using face data instead of caricature data

In our main paper, we highlight that our model sidesteps the need for caricature data by synthesising training images through Thin Plate Spline (TPS) deformation of standard face landmarks. Here we show more visualisations to demonstrate its necessity. In generative tasks, diffusion models [2, 7] significantly surpass GANs [1] in terms of diversity and creativity. However, diffusion models also have a much higher demand for data compared to GANs. For caricature generation tasks, collecting large-scale caricature data is particularly challenging, whereas numerous large-scale face datasets [6, 11] are already available. More importantly, large-scale image-landmark pairs are required for layout conditional generation [10] tasks. The technology for face landmark prediction is well-established, allowing us to easily obtain a large number of face-landmark pairs. In contrast, predicting landmarks for caricature faces remains an unsolved problem. Therefore, previous studies [3–5] *manually* annotate thousands of caricature images to obtain their landmarks, which is *impractical* given the vast amount of data required by diffusion models. We apply the Face of Art (FoA) [9] and ArtFacePoints (AFP) [8] designed for predicting face landmarks in artworks to caricature data and illustrate the results in the Fig. 1. For caricature images that closely resemble human faces, the predictions are relatively accurate, though certain facial regions still lack precision (see the last two columns). However, for highly exaggerated caricature styles, the landmark predictions are effectively invalid (see the first two columns).

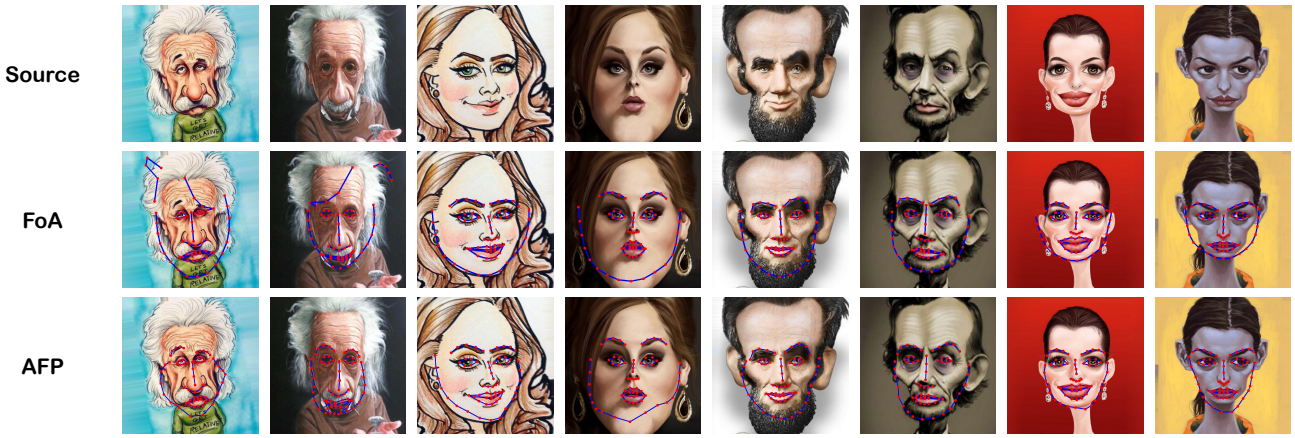

Figure 1. **Landmark prediction for caricature data.** We apply the Face of Art (FoA) [9] and ArtFacePoints (AFP) [8] to caricature data. Even though face landmark prediction is a straightforward task, landmark prediction for caricatures remains an unresolved challenge.

2. Illustration of the interactive process

In Fig. 2, we illustrate the interactive process by dragging different facial regions of the landmark. We perform successive operations such as "stretch the chin", "narrow the eyes" and "open the mouth". By changing only the landmark condition and leaving the other conditions unchanged (*e.g.*, identity, random seed, text prompt, guidance scale, *etc.*), we produce results that achieve an effect similar to that of image editing.

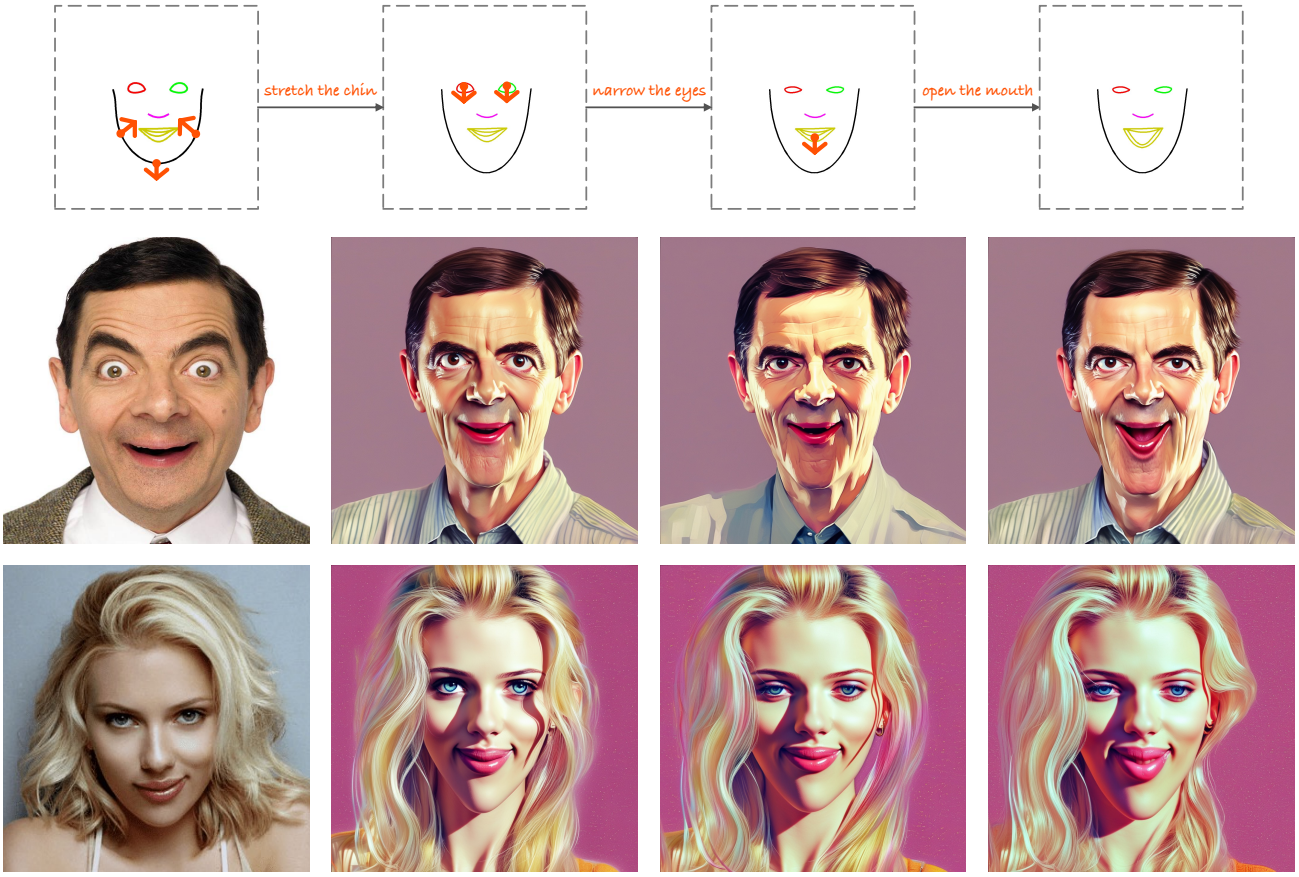

Figure 2. **Multiple-round interactive editing based on one identity.** We show the caricature results generated through a few dragging actions on different regions of the landmark.

3. Ablation study

We vary the scale of the face embedding condition by adjusting the attention coefficient of the IdentityNet during the inference stage, which results in Fig. 3.

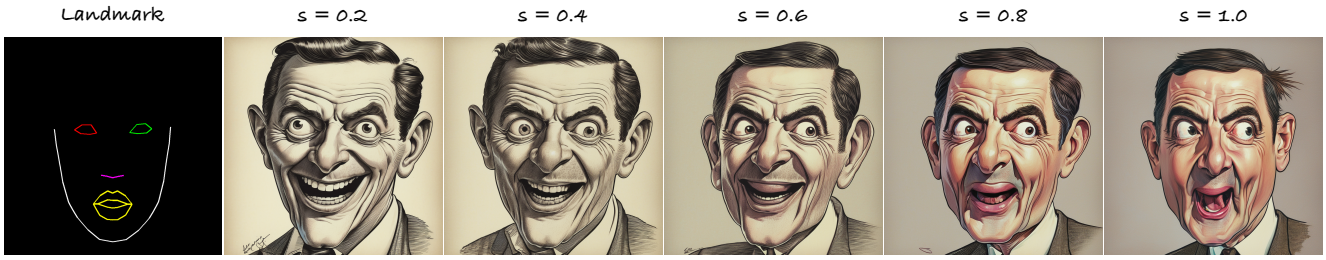

Figure 3. Results under different scale of the ID condition of Mr. Bean. As the condition scale gradually increases, the similarity between the generated results and the source identity also progressively improves.

4. Additional visualisations

In Fig. 4, we show generated results under extremely exaggerated landmark conditions. The TPS algorithm applies varying degrees of deformation to different facial regions, enabling *CaricatureBooth* to effectively handle various deformation conditions and adjust expressions such as slimming of the face, widening of the face, enlargement of the mouth and hearty laughter, among others.

Figure 4. **Generated caricatures under exaggerated landmark conditions.** Note that the different painting styles shown in the results are changed by the text prompts.

References

[1] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NeurIPS*, 2014. 1

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1

[3] Xin Huang, Dong Liang, Hongrui Cai, Yunfeng Bai, Juyong Zhang, Feng Tian, and Jinyuan Jia. Double reference guided interactive 2d and 3d caricature generation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 1

[4] Xin Huang, Dong Liang, Hongrui Cai, Juyong Zhang, and Jinyuan Jia. Caripainter: Sketch guided interactive caricature generation. In *ACM Multimedia*, 2022.

[5] Jing Huo, Wenbin Li, Yinghuan Shi, Yang Gao, and Hujun Yin. WebCaricature: a benchmark for caricature recognition. In *BMVC*, 2018. 1

[6] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[8] Aline Sindel, Andreas Maier, and Vincent Christlein. Artfacepoints: High-resolution facial landmark detection in paintings and prints. In *ECCV*, 2022. 1

[9] Jordan Yaniv, Yael Newman, and Ariel Shamir. The face of art: landmark detection and geometric style in portraits. *ACM Transactions on graphics*, 2019. 1

[10] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1

[11] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *CVPR*, 2022. 1