Bring Clipart to Life
Supplementary Material

Nanxuan Zhao [1], Shengqi Dang [2], Hexun Lin [2], Yang Shi [2], Nan Cao [2]
[1]Adobe Research [2]Tongji University

1. More Results

1.1. Results on Different Frames of the Same Person

Our model can deal with different frames of the same person in the video well as shown in Fig. 1.

1.2. Results on Fine-grained Control

As shown in the Fig. 2, text-driven approaches like StyleCLIP fail to distinguish between round and rectangle glasses shapes even after tuning prompts using "black round glasses" and "black square thick-rimmed glasses", while our model can transfer the glasses successfully. Text can specify high-level description but may fail to deliver fine-grained control.

2. The Effect of Backbone

To demonstrate the effectiveness of taking CLIP [8] as our visual encoder, we compare with several baseline encoders including ViT [2], SWIN [5], SWIN-v2 [5], VGG19 [10], and DETR [1], covering both transformer-based and convolution-based backbones. As shown in Fig. 3, with CLIP visual encoders, the facial attributes can be transferred successfully, while the other encoders fail to do so.

Figure 1: Results on different frames of the same person. The clipart © from Open Peeps [11].

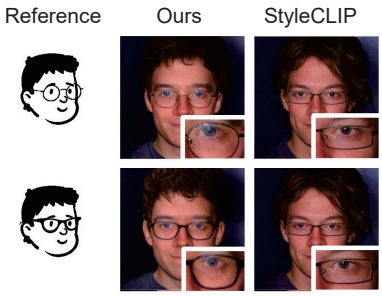

Figure 2: Results on fine-grained control. The clipart © from Open Peeps [11].

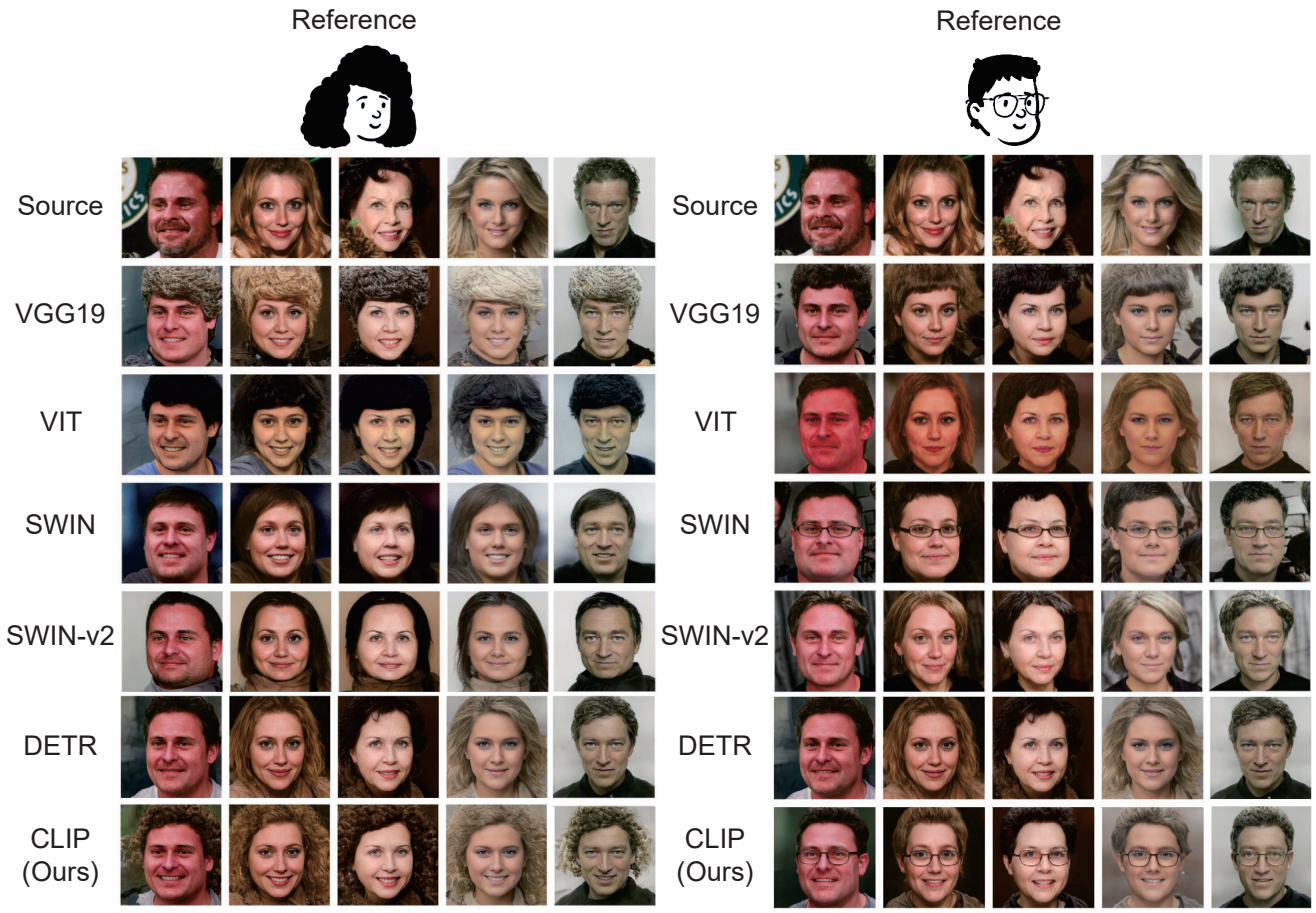

Figure 3: The effect of different visual backbones. The clipart © from Open Peeps [11].

3. More Comparisons

We also compare our method with two related works: Mind the GAP (**MTG**) [12] by shifting domain based on one shot, and StyleGAN-NADA (**SG-NADA**) [3] by leveraging the CLIP model for domain adaptation. The results are shown in Fig. 4 and Fig. 5. As can be seen, MTG is often bothered by the color distribution of reference images and fails to maintain the color and texture of original images. StyleGAN-NADA often generates more cartoon-like images, which are not realistic.

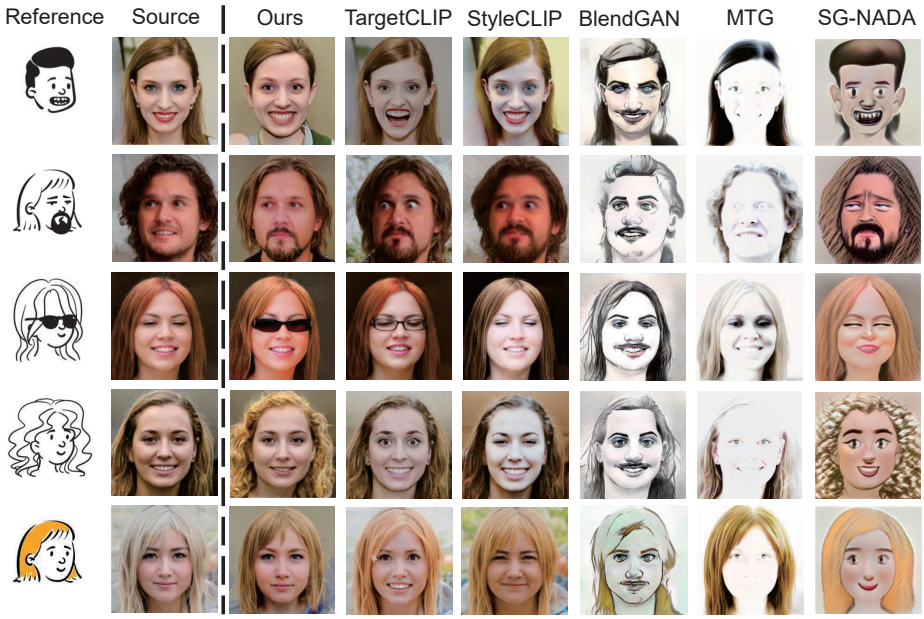

Figure 4: Comparison of different methods for clipart-driven face photo editing. The clipart © from Open Peeps [11].

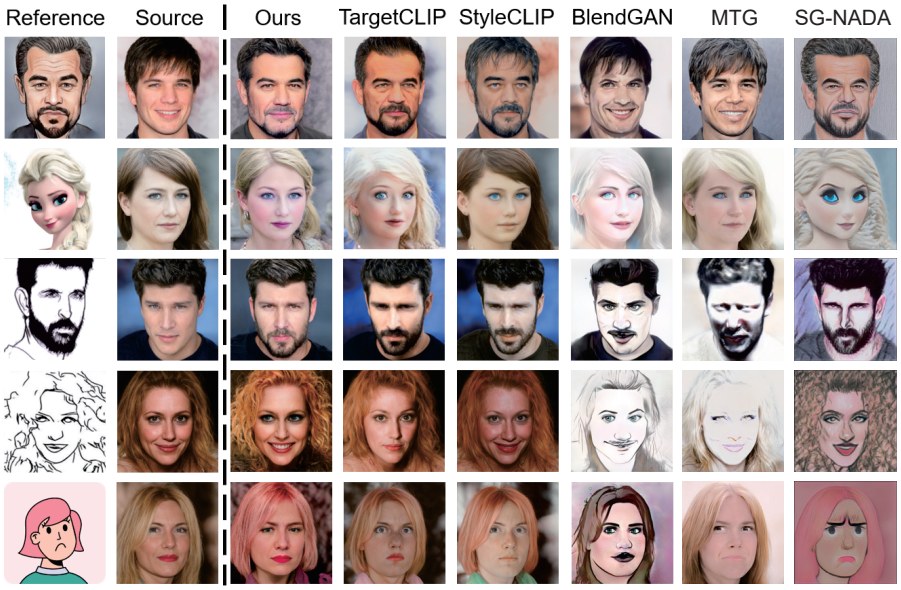

Figure 5: Comparison of different methods for clipart-driven face photo editing. The clipart © from Vue Color Avatar [4], cartoon from Disney Animation, and sketches from Toonify [7], PSP [9] and ArtLine [6].

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision (ECCV)*, pages 213–229, 2020.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[3] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022.

[4] Leo Ku. Vue color avatar. *https://github.com/Codennnn/vue-color-avatar*, 2021.

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[6] Vijish Madhavan. Artline. *https://github.com/vijishmadhavan/ArtLine*, 2020.

[7] Justin Pinkney. Toonify. *https://toonify.photos/*, 2020.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[9] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2287–2296, 2021.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[11] Pablo Stanley. Openpeeps dataset. *https://www.openpeeps.com/*, Accessed on 2023.

[12] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2022.