Supplementary Materials for

"DCT-Net: Domain-Calibrated Translation for Portrait Stylization"

Paper ID 594

In this document we provide the following supplementary contents:

- Details of network architecture.
- Training details of content calibration network.
- Results of artistic portrait generation.
- Results of full-body image translation.
- Comparison with state-of-the-art methods.
- Limitations.

1. Network architecture

For the content calibration network, the generator and discriminator follow the StyleGAN2 config-f models and the detailed architecture can be found in [9]. For the texture translation network, we provide network structure details in Table 1.

Table 1: Details of generator architecture.

Operation	Output Size
input	$256 \times 256 \times 3$
Conv+LReLU	$256 \times 256 \times 32$
Conv+LReLU+Conv+LReLU	$128 \times 128 \times 64$
Conv+LReLU+Conv+LReLU	$64 \times 64 \times 128$
Residual Block	$64 \times 64 \times 128$
Residual Block	$64 \times 64 \times 128$
Residual Block	$64 \times 64 \times 128$
Residual Block	$64 \times 64 \times 128$
Conv+LReLU+Conv transpose+LReLU	$128 \times 128 \times 64$
Conv+LReLU+Conv transpose+LReLU	$256 \times 256 \times 32$
Conv+LReLU	$256 \times 256 \times 3$

Table 2: Details of discriminator architecture.

Operation	Output Size
input	$256 \times 256 \times 3$
DeConv+SN+LReLU	$128 \times 128 \times 32$
Conv+SN+LReLU	$128 \times 128 \times 32$
DeConv+SN+LReLU	$64 \times 64 \times 64$
Conv+SN+LReLU	$64 \times 64 \times 64$
DeConv+SN+LReLU	$32 \times 32 \times 128$
Conv+SN+LReLU	$32 \times 32 \times 128$
Conv+SN	$32 \times 32 \times 1$

Table 3: Details of regressor architecture.

Operation	Output Size
input	$256 \times 256 \times 3$
DeConv+SN+LReLU	$128 \times 128 \times 32$
Conv+SN+LReLU	$128 \times 128 \times 32$
DeConv+SN+LReLU	$64 \times 64 \times 64$
Conv+SN+LReLU	$64 \times 64 \times 64$
DeConv+SN+LReLU	$32 \times 32 \times 128$
Conv+SN+LReLU	$32 \times 32 \times 128$
Mean pooling	128
Fully connected layer	512
Fully connected layer	3

2. Training details of content calibration network.

Given a style dataset contains approximate 100 images for a similar style. We start from the pretrained StyleGAN2 model G_s trained on real faces (e.g., FFHQ dataset), and use a copy of G_s as our initialization model G_t. To adapt G_t generating images in the target domain, we fine-tune G_t with full loss function consisting of the original adversarial loss and an identity loss:

$$\mathcal{L}_{ccn} = \mathcal{L}_{adv} + \lambda_{id}\mathcal{L}_{id},$$

where λ_{id} denotes the weight of the identity loss. $\mathcal{L}_{id}$ is formulated as:

$$\mathcal{L}_{id} = 1 - cos\left(z_{id}(\hat{x}_t), z_{id}(\hat{x}_s)\right),$$

where $cos(\cdot,\cdot)$ represents the cosine similarity of two vectors and the id feature z_{id} is extracted from existing face recognition model [11]. $\hat{x}_t$ and $\hat{x}_s$ are outputs of fixed generator G_s and learnable generator G_t, respectively. The define of $\mathcal{L}_{adv}$ and more training parameters are the same with [9]. We set λ_{id}=0.1 and train G_t for around 1000 iterations.

3. Results of artistic portrait generation.

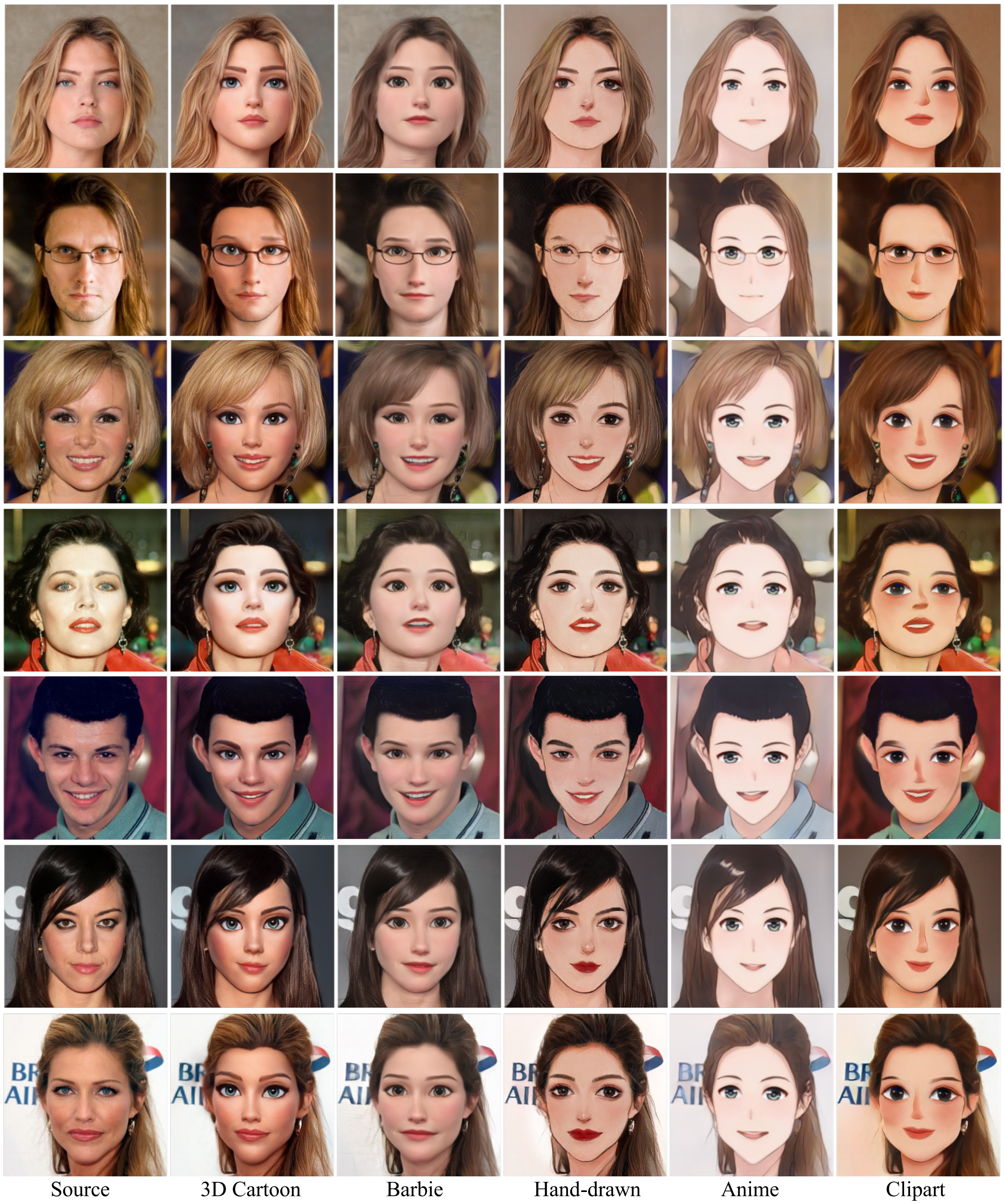


Figure 1: Results of synthesized portraits in various styles. Source image credits: CelebA [2].

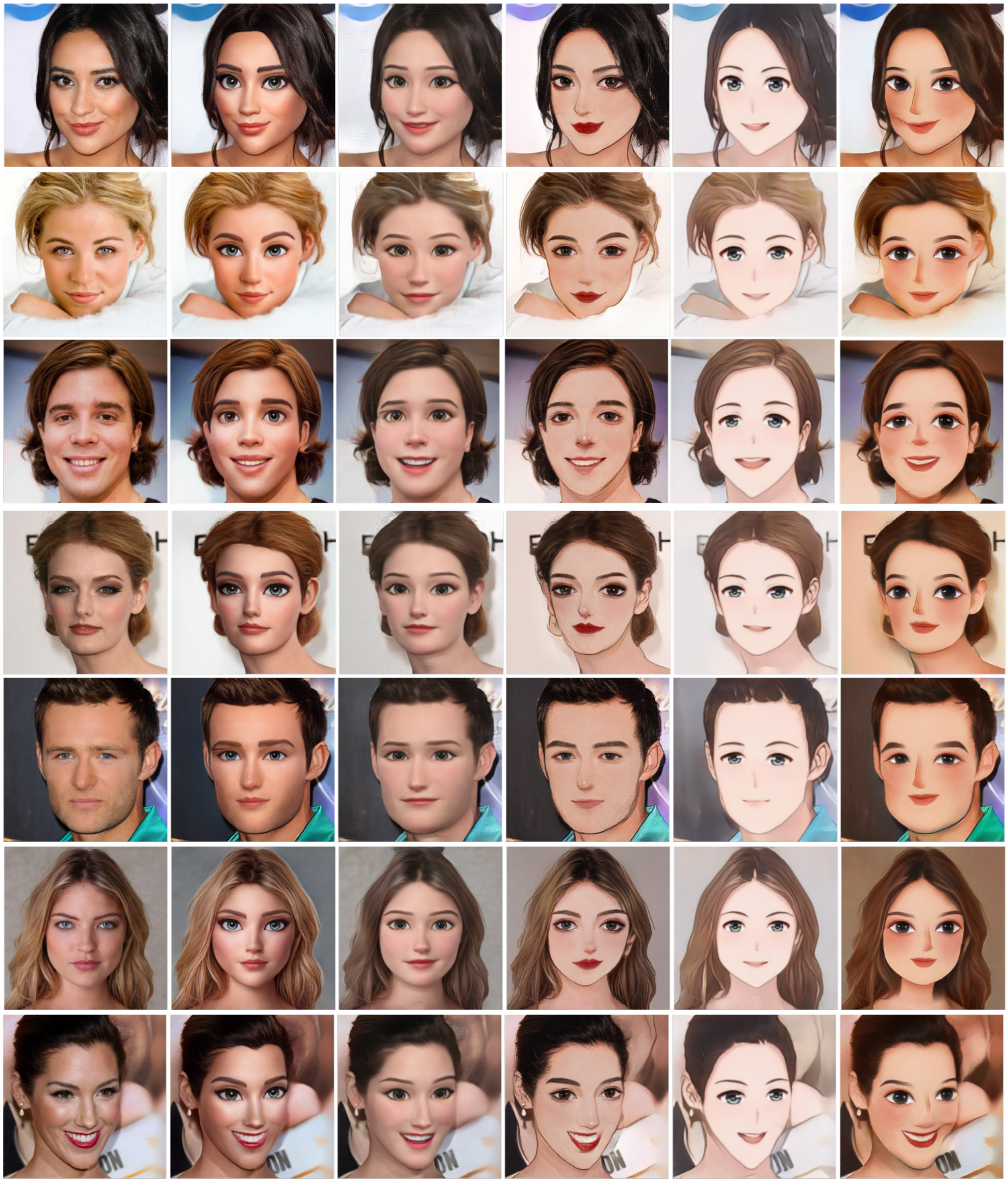


Figure 2: Results of synthesized portraits in various styles. Source image credits: CelebA [2].

4. Results of full-body image translation.

(a) 3D Cartoon

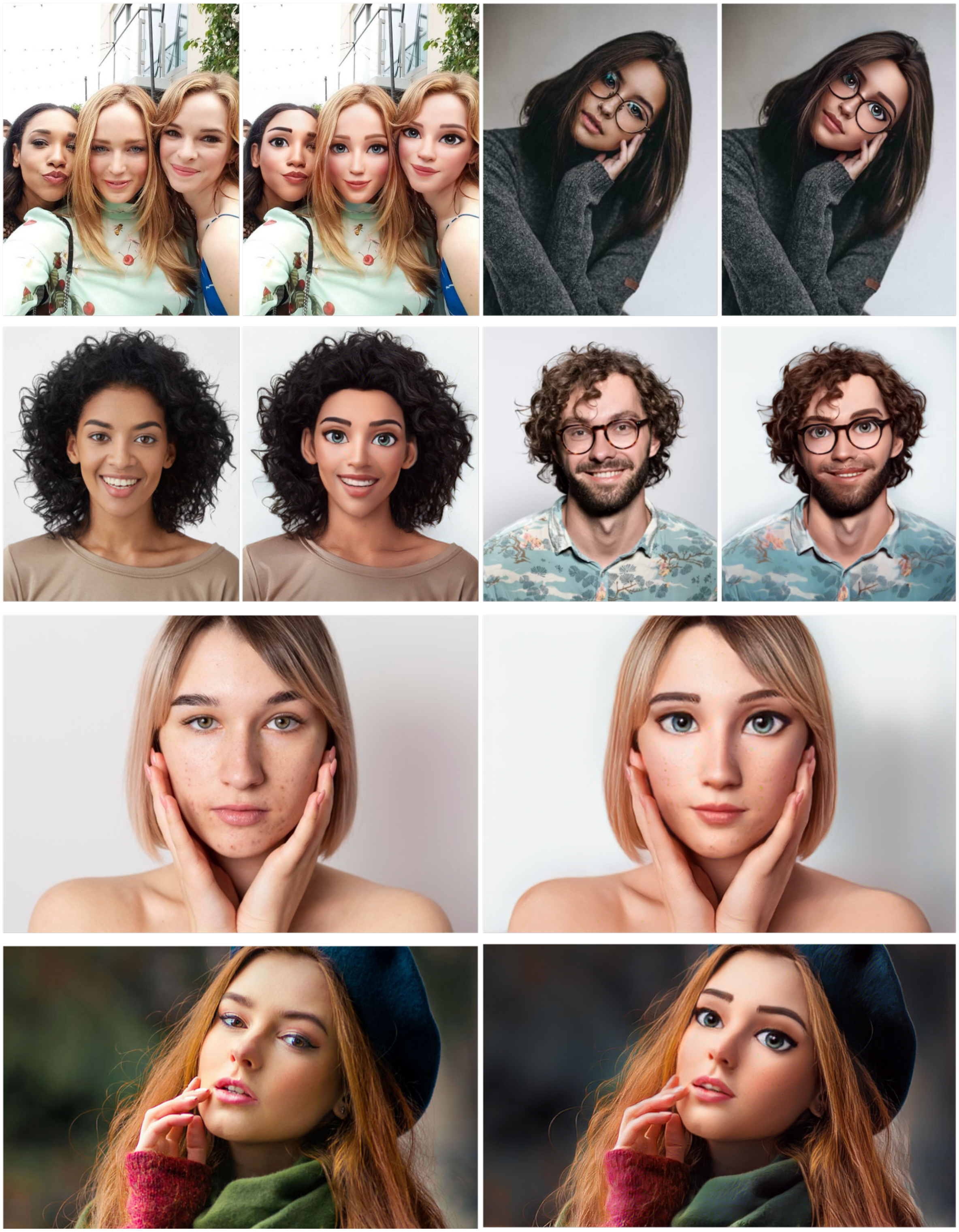


Figure 3: Results of stylized full images in 3D cartoon style. The source image in the left and the stylized result in the right. Source images: ©Unsplash[12], Google [1].

(b) Hand-drawn

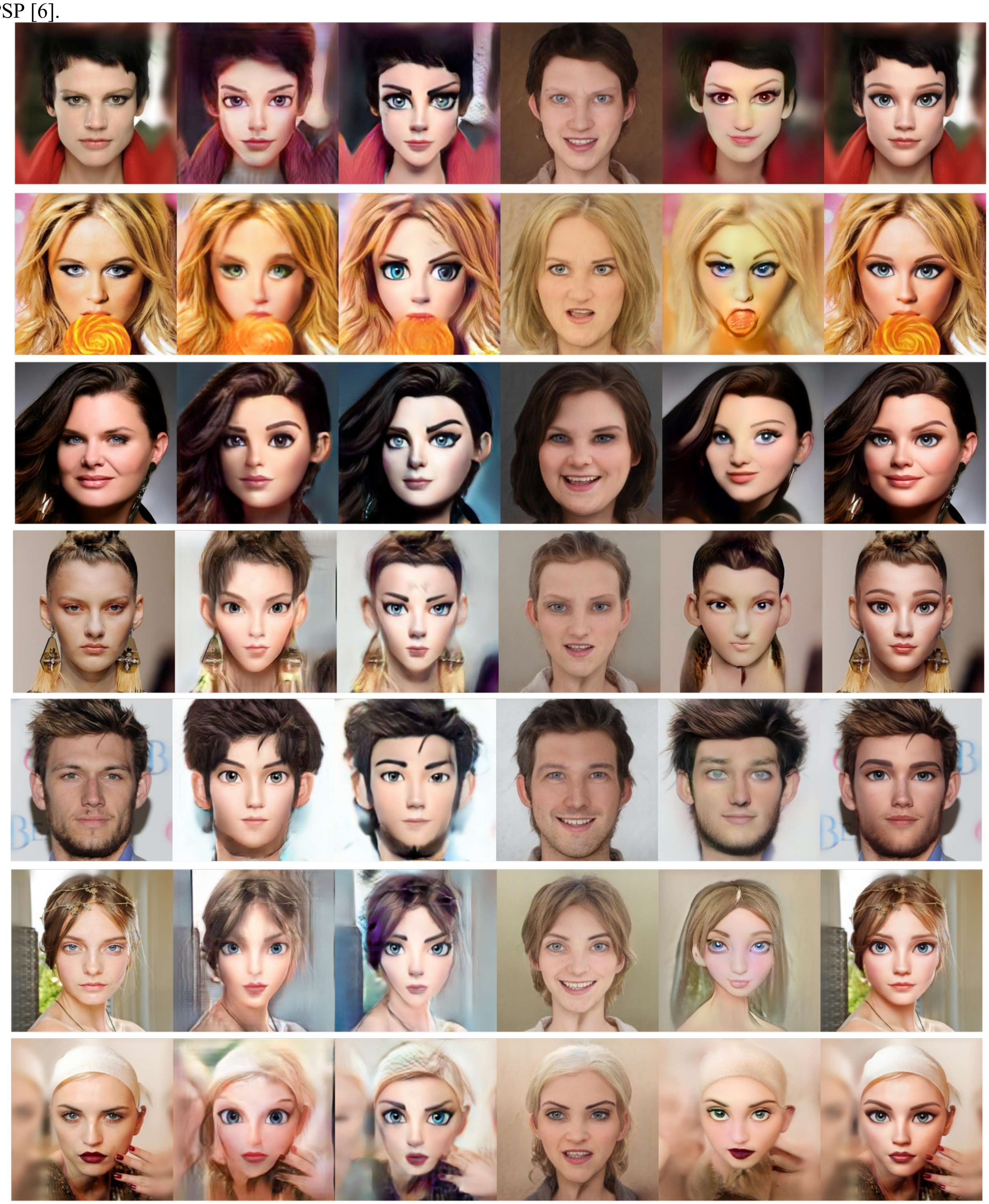
Figure 4: Results of stylized full images in hand-drawn style. The source image in the left and the stylized result in the right. Source images: ©Unsplash[12], Google [1].

(c) Anime

Figure 5: Results of stylized full images in anime style. The source image in the left and the stylized result in the right. Source images: ©Google [1].

5. Comparison with state-of-the-art methods.

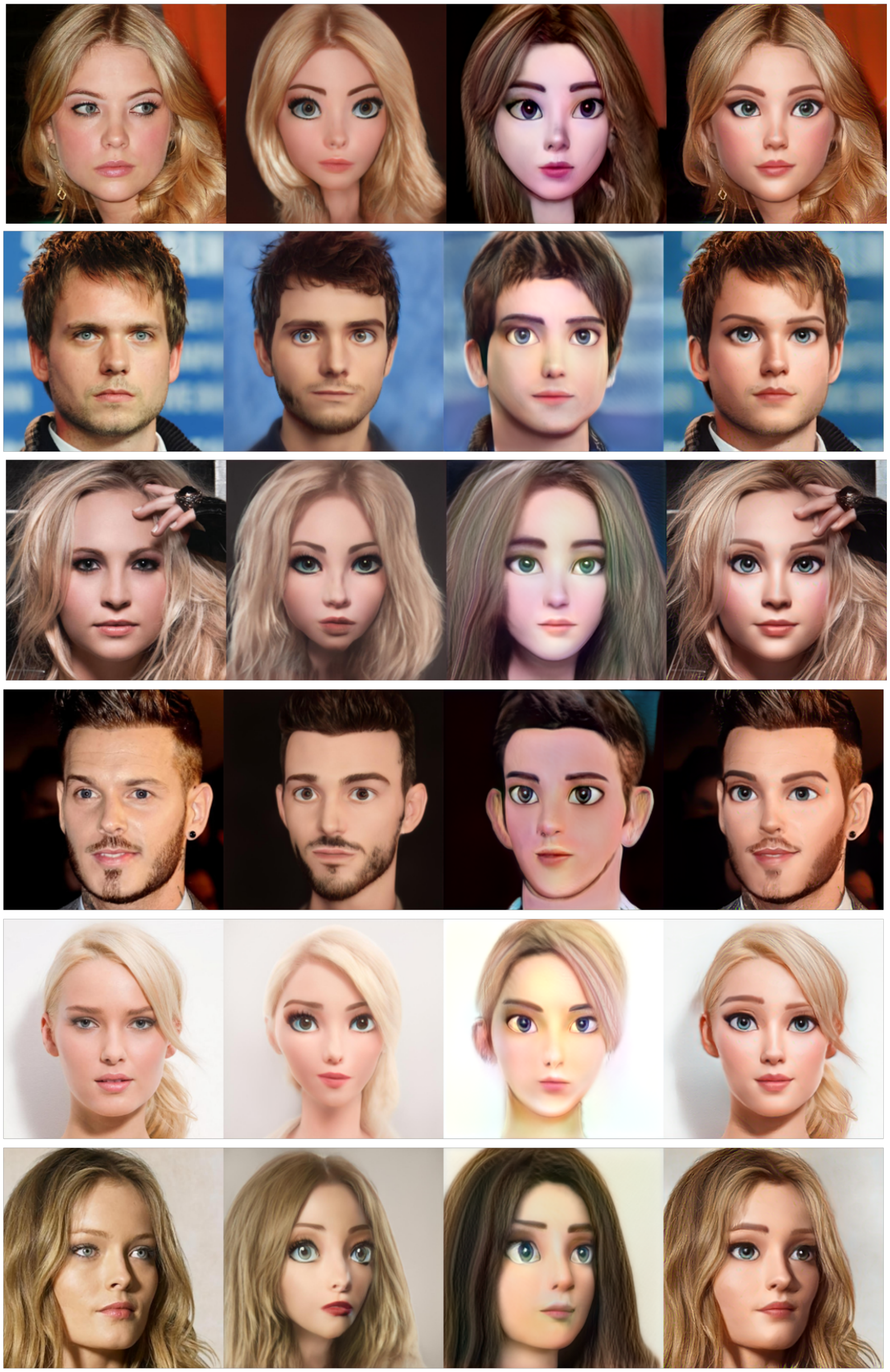
We provide more comparison results with four SOTA methods: CycleGAN [3], U-GAT-IT [4], Toonify [5], and PSP [6].



(a) Source (b) CycleGAN (c) U-GAT-IT (d) Toonify (e) PSP (f) Ours

Figure 6: Qualitative comparison with state-of-the-art methods. Source images: ©CelebA [2].

We provide more comparison results with two SOTA methods: AgileGAN [7] and Few-shot-Ada [10].



(a) Source (b) Agile-GAN (c) Few-shot-Ada (d) Ours

Figure 7: Qualitative comparison with state-of-the-art methods. Source images: ©AgileGAN [7].

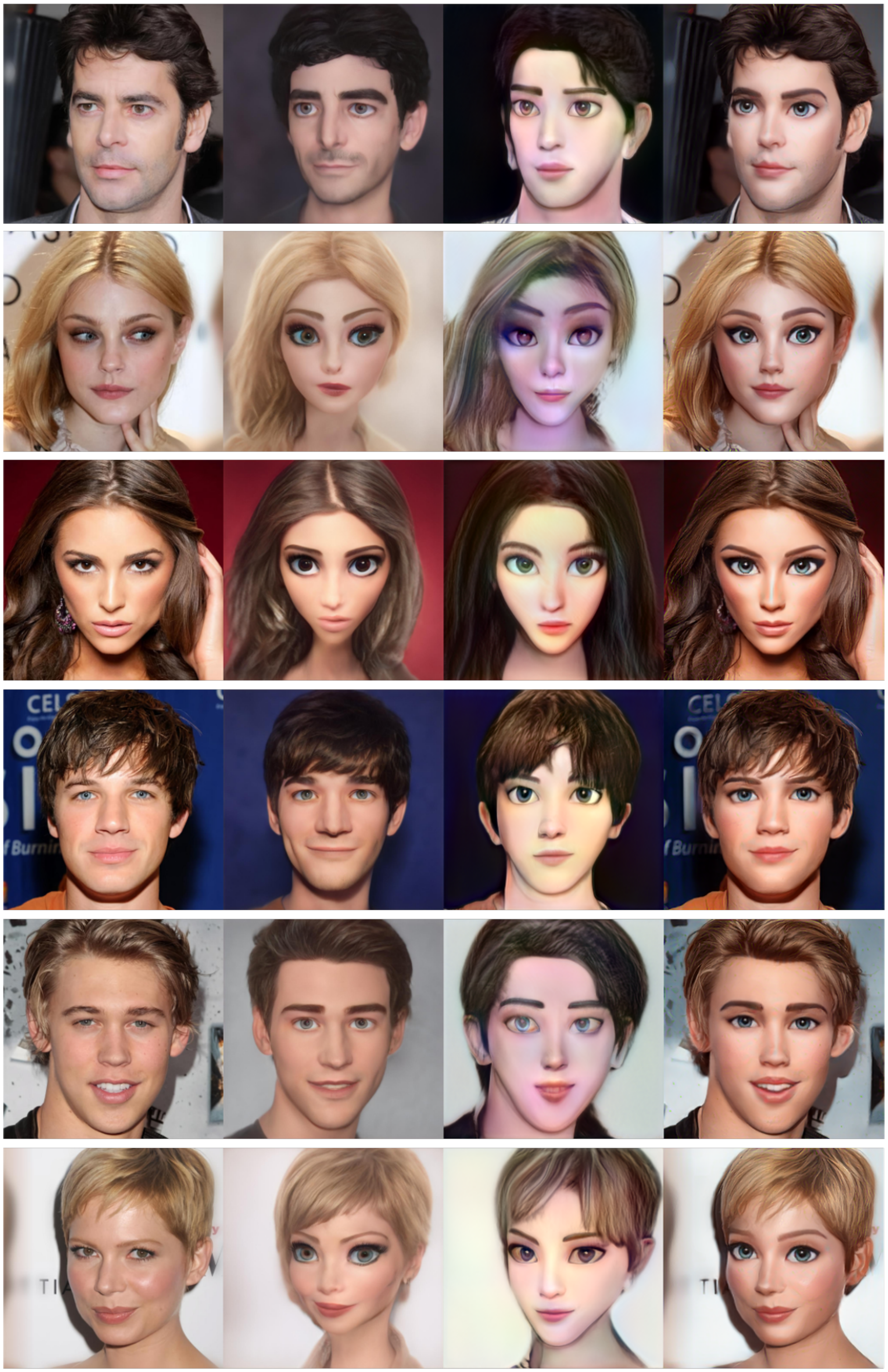
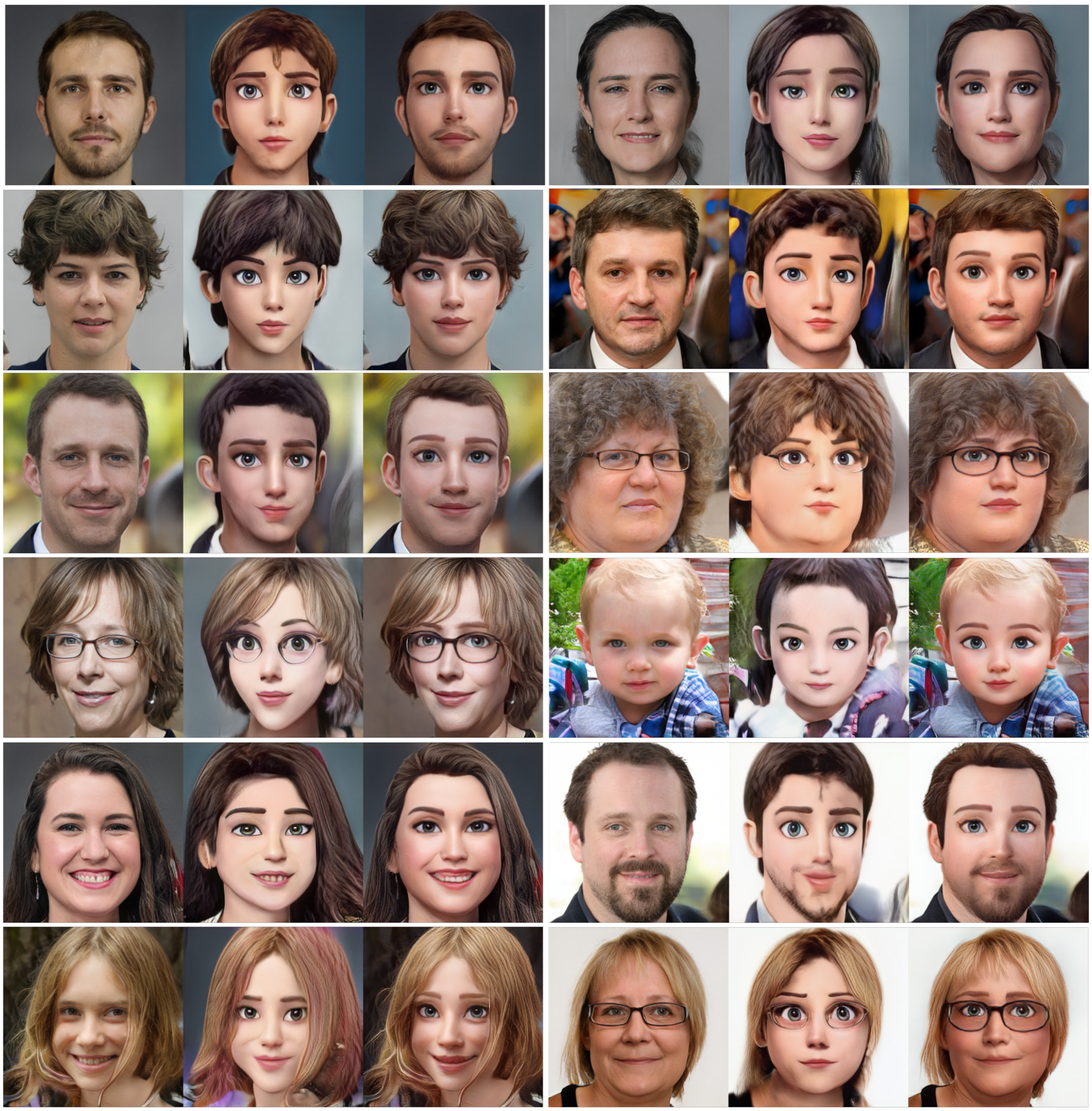


Figure 8: Qualitative comparison with state-of-the-art methods. Source images: ©AgileGAN [7].

Due to the nature of unconditional generative model, Few-shot-Ada performed worse for arbitrary photo transfer. So, we also evaluate this method in noise manner: random noises z are randomly sampled in its latent space, and we input z into both the source generator and the adapted style generator to produce photo images I_p and stylized results I_r, respectively. These aligned pairs (I_p, I_r) indicate the best capability of this adapted generation model for the transfer task (no inversion error is introduced). We use the synthesized photo image I_p as the input of our network to produce comparison results in Figure 9. As we can see, our method still outperforms this method with more content details preserved.



(a) Photo (z) (b) Few-shot-Ada (z) (c) Ours (a) Photo (z) (b) Few-shot-Ada (z) (c) Ours

Figure 9: Comparison with Few-shot-Ada [8] in noise manner. Source images: ©Agile-GAN [7].

6. Limitations.

Because of the inherent characteristics of some styles (i.e., hand-drawn and anime), our synthesized results might not be natural enough when there exist server lighting shadows in human faces, as shown in Figure 10. But some styles (i.e., 3D cartoon) can still be well handled owing to its specific nature.

(a) Source (b) 3D cartoon (c) Hand-drawn (d) Anime

Figure 10: Failure cases due to the disturbed illumination.

Reference

[1] Google. [EB/OL]. https://www.google.com/.

[2] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In IEEE Conference on Computer Vision and Pat- tern Recognition (CVPR), 2020.

[3] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A 1062 Efros. Unpaired image-to-image translation using cycle- 1063 consistent adversarial networks. In Proceedings of the IEEE 1064 international conference on computer vision, pages 2223–1065 2232, 2017.

[4] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In International Conference on Learning Representations, 2020.

[5] Justin NM Pinkney and Doron Adler. Resolution dependent gan interpolation for controllable image synthesis between domains. arXiv preprint arXiv:2010.05334, 2020.

[6] 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, pages 2287–2296, 2021.

[7] Guoxian Song, Linjie Luo, Jing Liu, Wan-Chun Ma, Chunpong Lai, Chuanxia Zheng, and Tat-Jen Cham. Agilegan: stylizing portraits by inversion-consistent transfer learning. ACM Transactions on Graphics (TOG), 40(4):1–13, 2021.

[8] Tero Karras, Samuli Laine, and Timo Aila. A style-based 947 generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4401–4410, 2019.

[9] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8110–8119, 2020

[10] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. 2021. Few-shot Image Generation via Cross-domain Correspondence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10743–10752.

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Addi- tive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4690–4699.

[12] Unsplash. [EB/OL]. https://unsplash.com/.