Supplementary Materials for

"Unpaired Cartoon Image Synthesis via Gated Cycle Mapping"

[1]Yifang Men, [1]Yuan Yao, [1]Miaomiao Cui, [2]Zhouhui Lian, [1]Xuansong Xie, [1]Xian-Sheng Hua

[1]DAMO Academy, Alibaba Group
[2]Peking University, China

In this document we provide the following supplementary contents:

- More results.
- Applications of generating cartoon images in different styles.
- Comparison with state-of-the-art methods.
- Details of network architecture.
- Model parameters of different solutions.
- The extended data of cartoon portraits.
- Limitations.
- Discussions.

1. Results of cartoon style interpolation

Our model constructs a complex manifold that is constituted of various cartoon images in different contents and diverse styles. We can travel along this manifold to synthesize an animation from one cartoon style to another, thus visualizing the encoded low dimensional space. We show the visualization results in the supplemental video ('6357_video.mp4'), which proves the effectiveness of style-controllable cartoon image synthesis and the continuity of our constructed manifold. Results of some representative frames are also depicted in Figure 1.

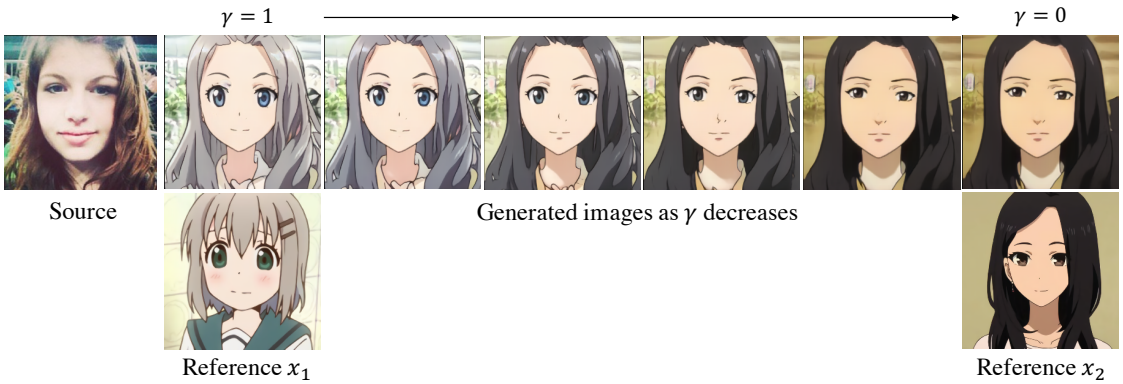


Figure 1: Style interpolation results of some representative frames as γ decreases. The mixed style code z_{mix} is computed by $z_{mix} = \gamma z_1 + (1-\gamma) z_2$, where z_1 and z_2 are extracted from reference image x_1 and x_2 respectively. Source images: ©selfie2anime [4].

2. Results in different scene cases

We roughly divide all real scenarios into portraits and scenes and use ‘scenes’ to represent all non-portrait scenarios. Due to the limitation of space, only example pairs covering limited cases with small resolution are presented in the main paper. Here we show high-resolution cartoonized results for more use cases, such as animals, foods, city views and other objects.

Source Results

Figure 2: Cartoonized results (right) and the corresponding source photos (left) for different scene cases. Results are generated with random styles. Source images: ©Google [10].

Figure 3: Cartoonized results (right) and the corresponding source photos (left) for different scene cases. Results are generated with random styles. Source images: ©Google [10].

3. Results of generating cartoon portraits with the extended method

Given an image from extended dataset as the reference and an arbitrary person portrait collected from the Internet as the source, our model can generate high-quality results with content details highly preserved and facial expressions precisely controlled (see Figure 4). It is recommended to see more results of cartoon video synthesis in the supplemental video.

Figure 4: Results generated by the extended method. Source images: ©Google [10].

4. Applications

4.1. Cartoon portrait synthesis in arbitrary styles

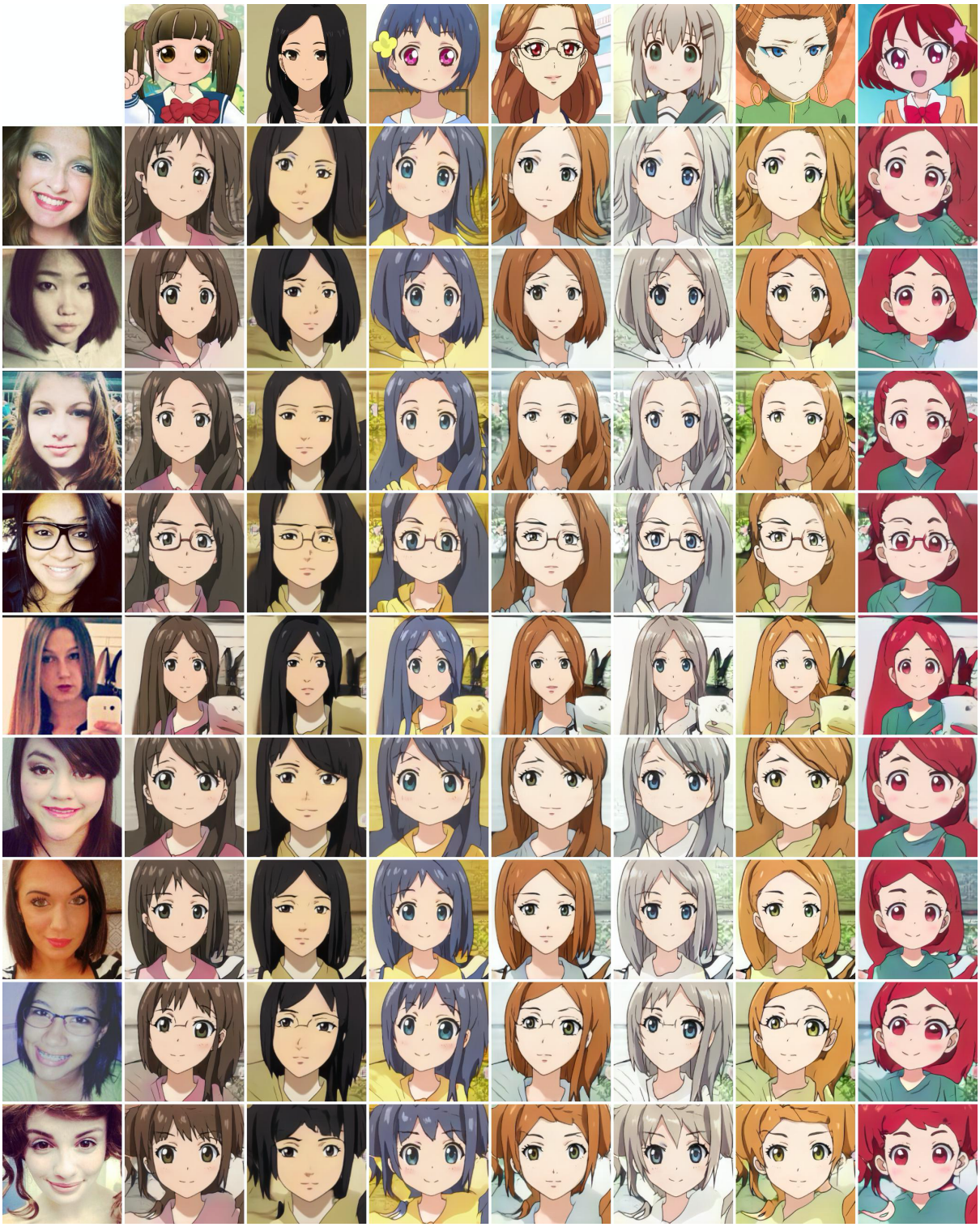


Figure 5: Results of transferring arbitrary styles to photo portraits **in the test set**. Source images: ©selfie2anime [4].

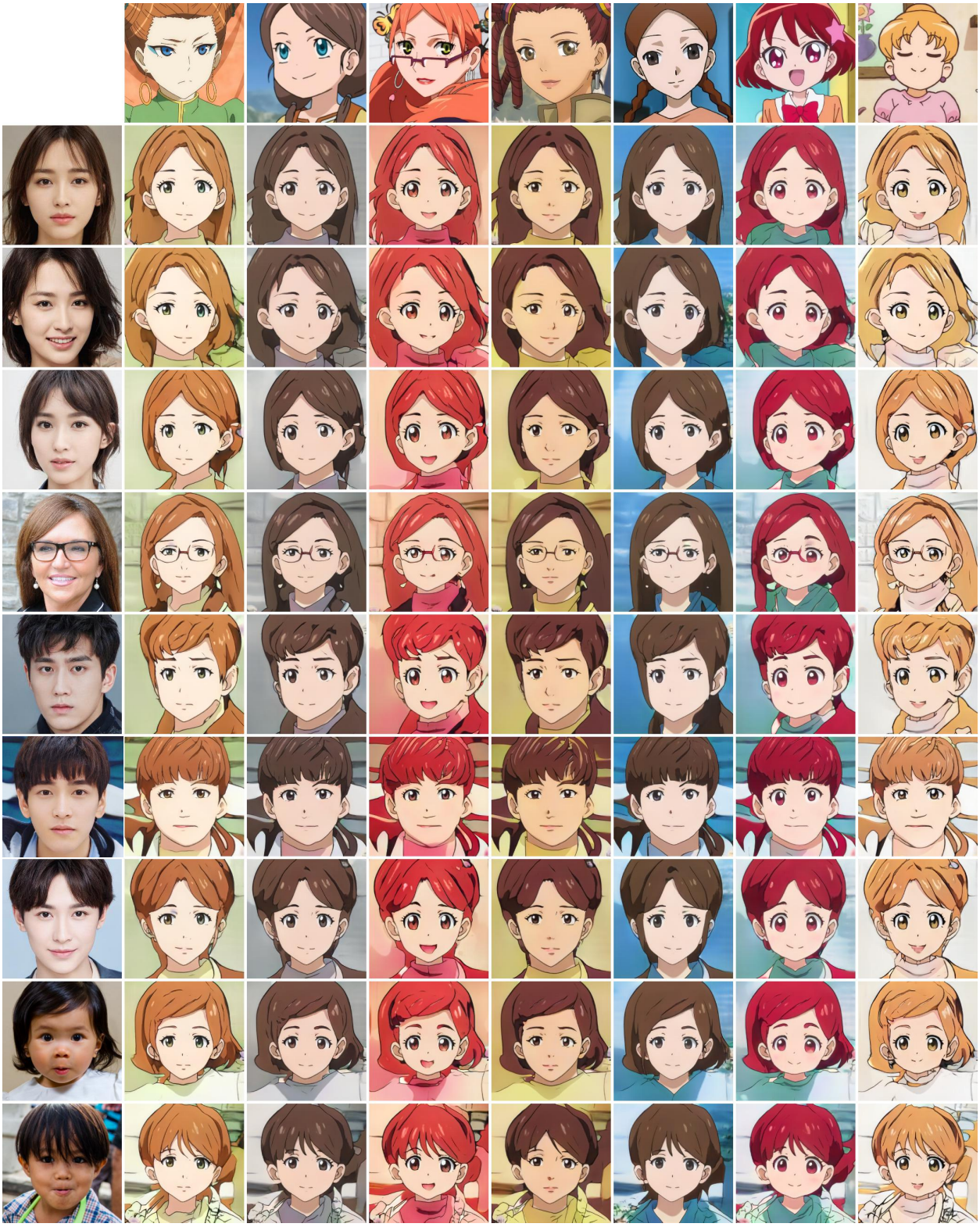


Figure 6: Results of transferring arbitrary styles to photo portraits **in the wild**. Source images: ©FFHQ [11], ©Google [10].

4.2. Cartoon scene synthesis in arbitrary styles

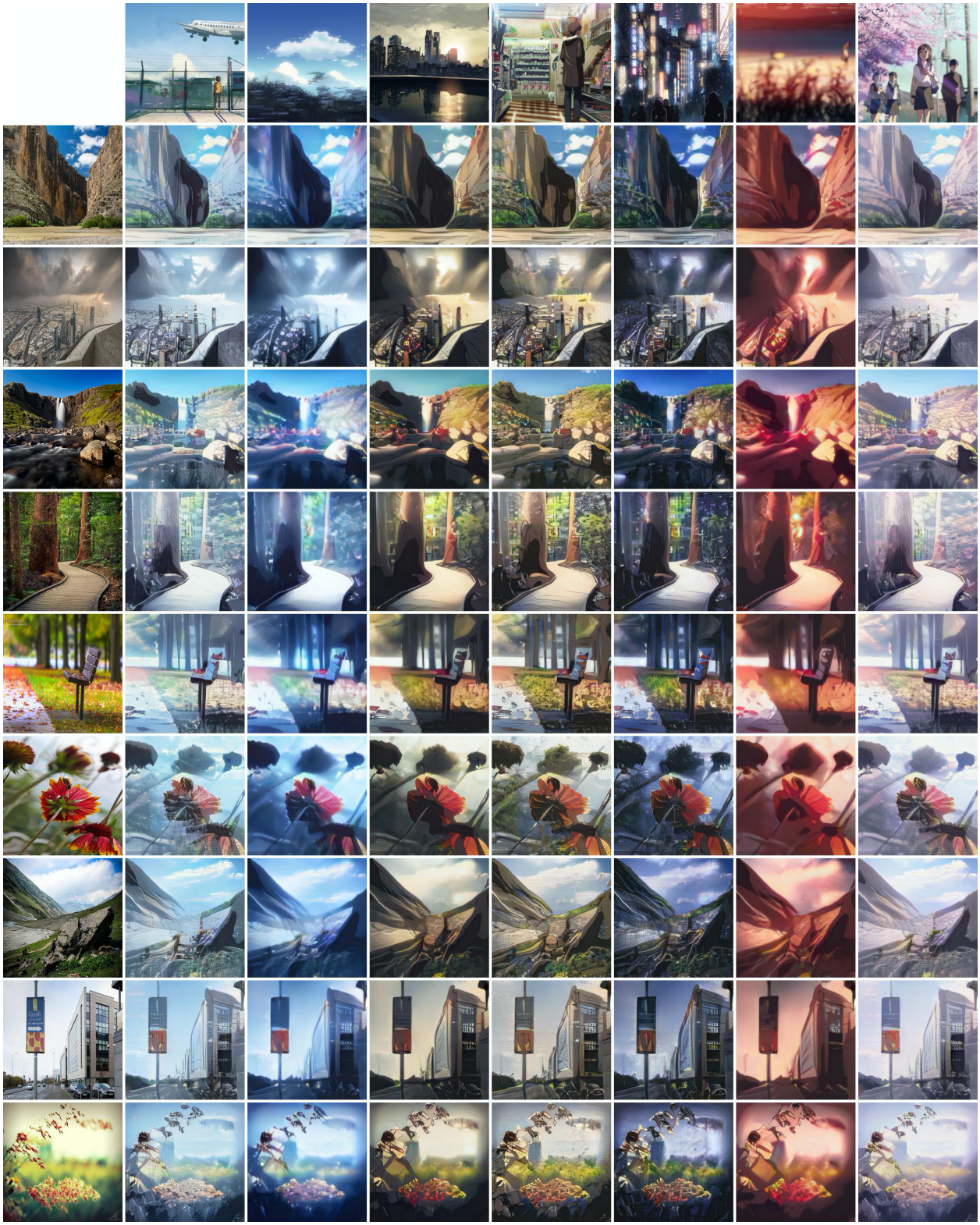


Figure 7: Results of transferring arbitrary styles to photo scenes **in the test set**. Source images: ©White-box [7].

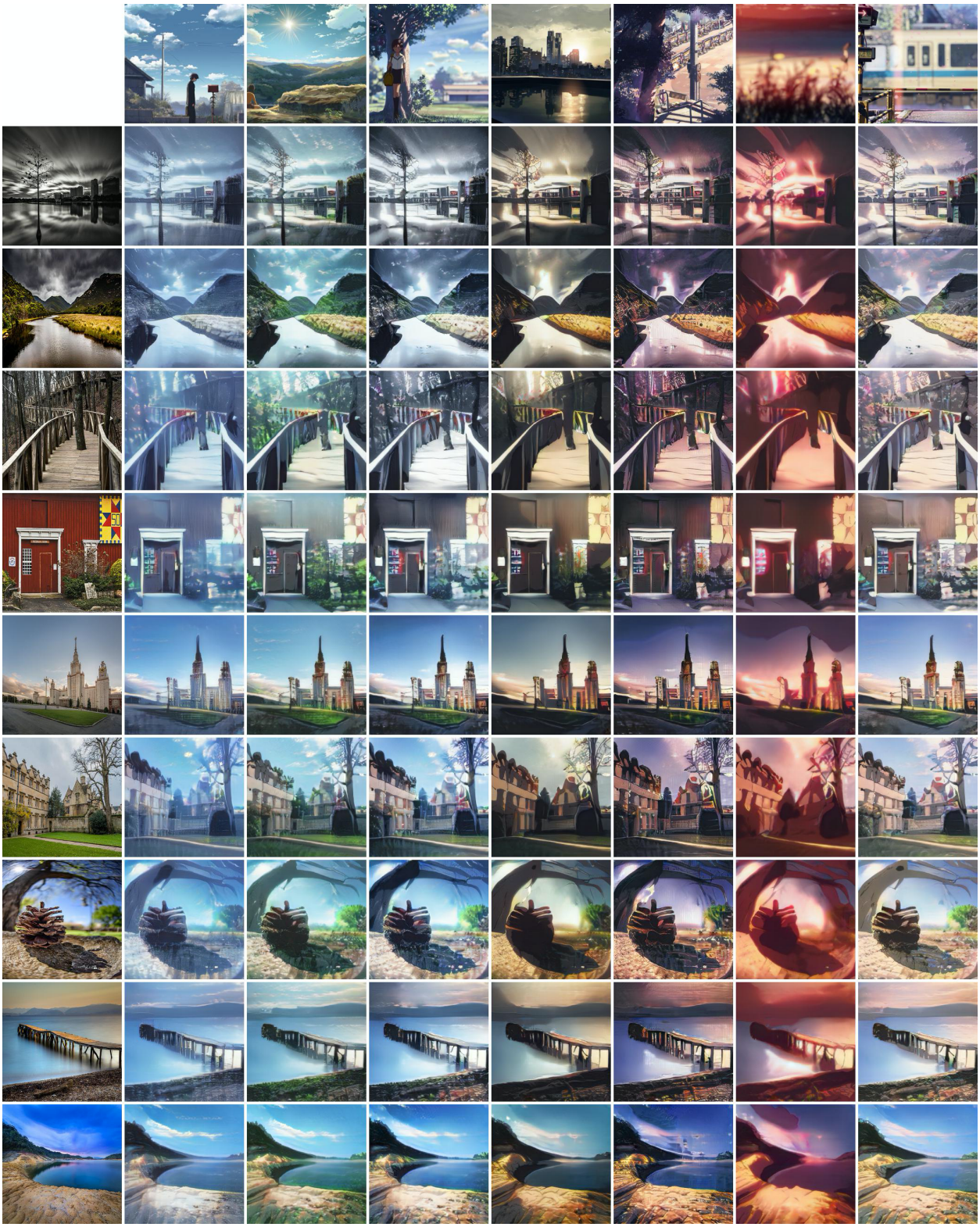


Figure 8: Results of transferring arbitrary styles to photo scenes **in the wild**. Source images: ©White-box [7].

5. Comparisons with state-of-the-art (SOTA) methods

To compare with SOTA methods as much as possible, we show only one style of our generated images and more can be found in Section 4.

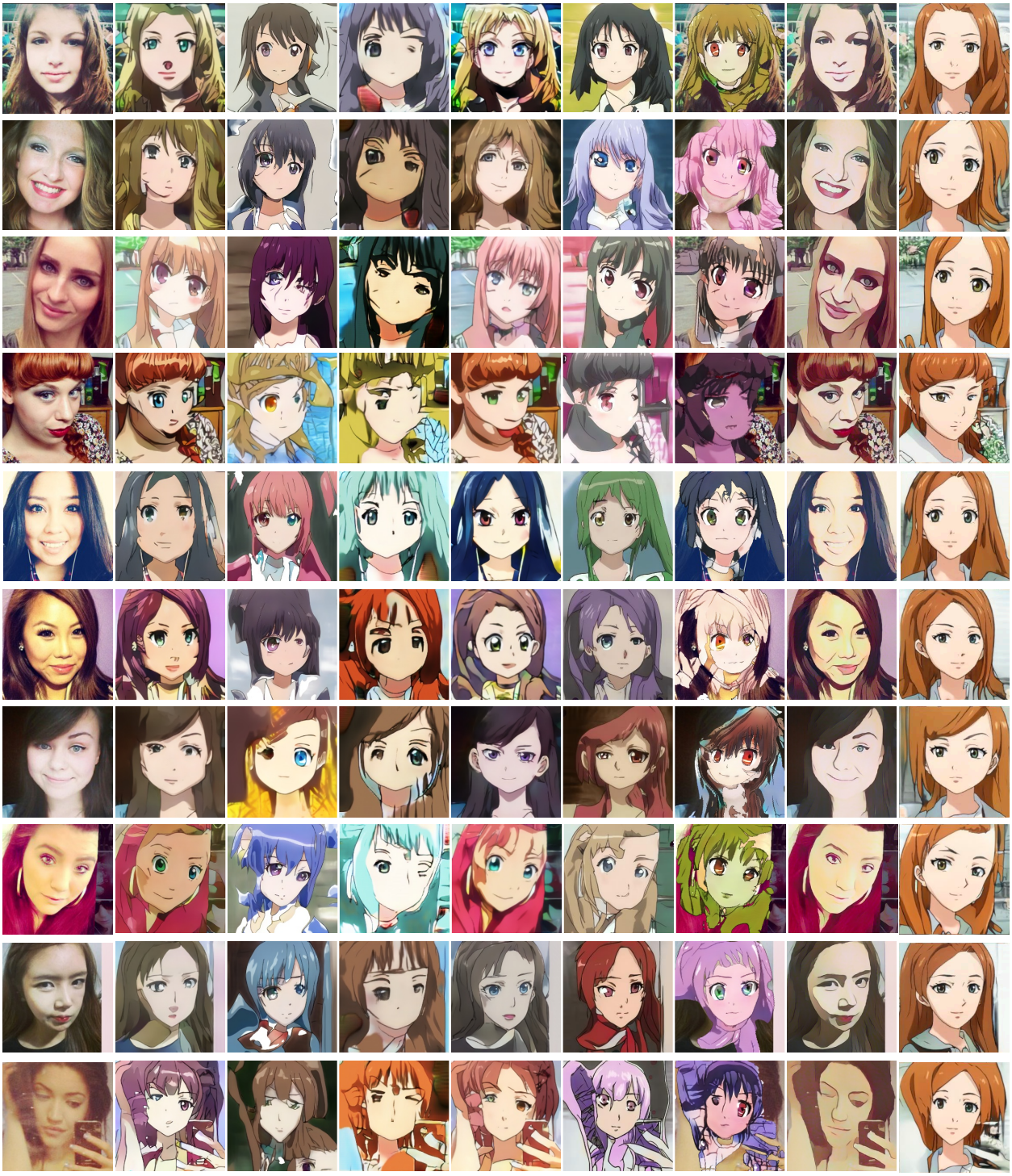


Figure 9: Comparison with state-of-the-art portrait cartoonization methods. Source images: ©selfie2anime [4].

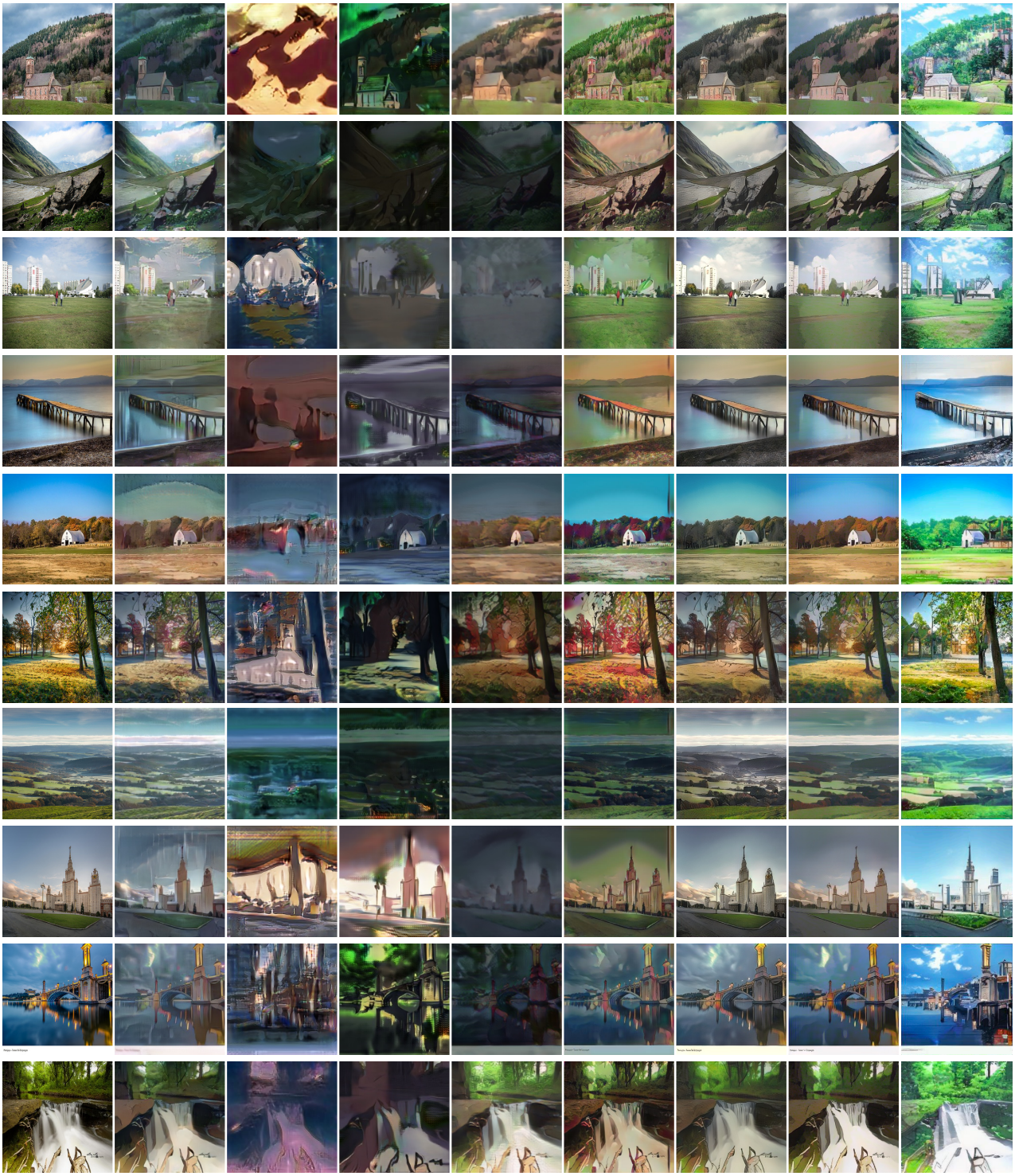


Figure 10: Comparison with state-of-the-art scene cartoonization methods. Source images: ©White-box [7].

We also compare our method with different solutions, such as example-based neural style transfer (ENST) [12], HiSD [13] and StarGAN-v2 [14]. ENST methods can transfer texture patterns from a style exemplar to content images. However, they only introduce style features in pixel level and fail to produce exaggerated geometry deformation for specific components (e.g., delicate big eyes and simplified mouths).

Most existing multi-domain models aims to handle tasks (e.g., face editing, cat2dog) where all domain images belonging to the similar species, and no semantic gap is considered. When scene and portraits are simultaneously learned, HiSD and StarGAN2 suffer from mode collapse and fail to produce meaningful results. So, we trained their models with only face data and provide comparison results in Figure 11.

(a) Input (b) ENST (c) StarGAN-v2 (d) HiSD (e) Ours

Figure 11: Comparison with ENST, HiSD and StarGAN-v2 methods. Source images: ©selfie2anime [4].

6. Details of network architecture

we provide detailed configurations of the proposed framework, which consists of three modules: a generator, a gated style encoder and a gated discriminator described below.

Type	Layer	Norm	Activation	Output Size
Encoder Down-sampling	ConvBlock	-	-	$256 \times 256 \times 64$
	DownResBlock	IN	LReLU	$128 \times 128 \times 128$
	DownResBlock	IN	LReLU	$64 \times 64 \times 256$
	DownResBlock	IN	LReLU	$32 \times 32 \times 512$
	DownResBlock	IN	LReLU	$16 \times 16 \times 512$
Encoder Bottleneck	ResBlock	IN	LReLU	$16 \times 16 \times 512$
	ResBlock	IN	LReLU	$16 \times 16 \times 512$
Decoder Bottleneck	ResBlock	AdaIN	LReLU	$16 \times 16 \times 512$
	ResBlock	AdaIN	LReLU	$16 \times 16 \times 512$
Decoder Up-sampling	UpResBlock	AdaIN	LReLU	$32 \times 32 \times 512$
	UpResBlock	AdaIN	LReLU	$64 \times 64 \times 256$
	UpResBlock	AdaIN	LReLU	$128 \times 128 \times 128$
	UpResBlock	AdaIN	LReLU	$256 \times 256 \times 64$
	ConvBlock	IN	LReLU	$256 \times 256 \times 3$

Table 1: Details of generator architecture. The extracted style code is injected into the generator via AdaIN paramet

Type	Layer	Norm	Activation	Output Size
Encoder Down-sampling	ConvBlock	-	-	$256 \times 256 \times 64$
	DownResBlock	-	LReLU	$128 \times 128 \times 128$
	DownResBlock	-	LReLU	$64 \times 64 \times 256$
	DownResBlock	-	LReLU	$32 \times 32 \times 512$
	DownResBlock	-	LReLU	$16 \times 16 \times 512$
	DownResBlock	-	LReLU	$8 \times 8 \times 512$
	DownResBlock	-	LReLU	$4 \times 4 \times 512$
	ConvBlock	-	LReLU	$1 \times 1 \times 512$
	Reshape	-	-	512
Gated Mapping Module	DSLayer * 2	-	-	512
	GSLayer * 2	-	-	64

Table 2: Details of gated style encoder architecture. DSLayer and GSLayer denote the domain-specific layer and group-specific layer, respectively. Both of them are constructed with fully-connected layer.

Type	Layer	Norm	Activation	Output Size
Discriminator Down-sampling	ConvBlock	-	-	$256 \times 256 \times 64$
	DownResBlock	-	LReLU	$128 \times 128 \times 128$
	DownResBlock	-	LReLU	$64 \times 64 \times 256$
	DownResBlock	-	LReLU	$32 \times 32 \times 512$
	DownResBlock	-	LReLU	$16 \times 16 \times 512$
	DownResBlock	-	LReLU	$8 \times 8 \times 512$
	DownResBlock	-	LReLU	$4 \times 4 \times 512$
	ConvBlock	-	LReLU	$1 \times 1 \times 512$
	Reshape	-	-	512
Gated Mapping Module	DSLayer * 2	-	-	512
	GSLayer * 2	-	-	1

Table 3: Details of gated discriminator architecture.

7. Model parameters of different solutions.

The design of a common translator for face and scene synthesis can significantly reduce model parameters by using a common G, and gated mapping unit in style encoder E_s is extremely lightweight with few parameters introduced. We show quantitative results of each module in our network, our solution (S1) and two-model solution (S2) in Table 1. E_s provides a style guidance for synthesis and it is necessary for deformable translation.

	G	E_s_4	E_s_2	S1(G+E_s_4)	S2(2G+2E_s_2)
Params(M)	33.88	20.98	20.92	54.86	109.60

Table 4. Parameter comparison of different solutions. E_s_k means E_s for k categories.

8. The extended data of cartoon portraits

To make our model widely applicable in the real world, we further extend the proposed method to video synthesis of cartoon portraits with the data expansion. Based on selfie2anime dataset in [4], we build an extended dataset by introducing a set of cartoon portraits containing diverse content characteristics and different facial expressions (e.g., open/closed eyes/mouth) in a similar cartoon style as a new class, which is added to the original selfie2anime dataset, making it possible to synthesize dynamic facial expressions. This new set consists of 1000 cartoon portraits at 256 × 256 resolution, which are designed by a single artist with a similar cartoon style. We show some example images in Figure 11.

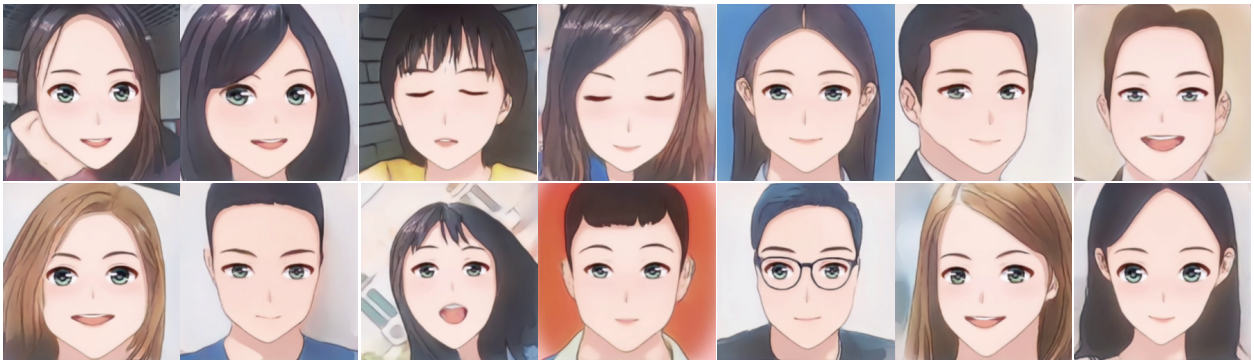
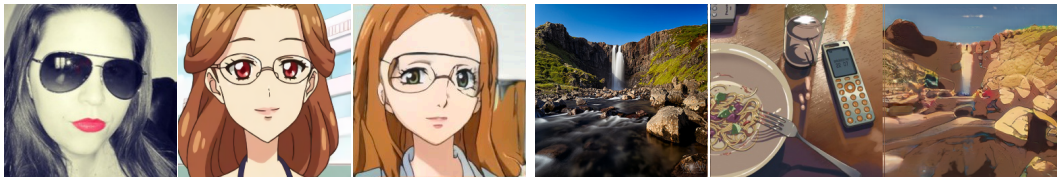


Figure 12: Examples from our extended cartoon portrait data.

9. Limitations

For portrait cartoonization, due to the absence of anime faces wearing sunglasses in the training data, our method fails to synthesize unseen accessories in cartoon styles, as shown in Figure 12(a). For scenery cartoonization, when there exists drastically different semantics between the source and the reference, the style-guided translation model may produce reasonable but meaningless results (e.g., unrealistic color transfer in Figure 12 (b)).



(a) Source Reference Result (b) Source Reference Result

Figure 13: Failure cases for cartoon portraits and sceneries. Source image credits: (a) selfie2anime [4], (b) White-box [7].

10. Discussions.

Negative Societal Impact. The aim of this paper is to automatically synthesize anime faces/scenes for source photos, which has no negative societal impact.

Use of Personal Data and Human Subjects. Our training phase is based on the existing selfie2anime dataset [4] and no self-collected real faces are used. For the testing phase, besides test images in [4], we evaluate our methods with arbitrary faces collected from the Internet (all copyrights are commented). If this is not allowed, we will remove or replace these cases.

Attribution of Data Assets. We comment the copyrights of used source portraits in the caption of the corresponding figures.

Data Contribution. No data contribution. We use a new cartoon portrait dataset for the video cartoonization task, but this dataset is not claimed as a contribution of this paper and we respect the copyright of the original author.

Reference

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

[2] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 172–189, 2018.

[3] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In Proceedings of the European conference on computer vision (ECCV), pages 35–51, 2018.

[4] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. arXiv preprint arXiv:1907.10830, 2019.

[5] Ori Nizan and Ayellet Tal. Breaking the cycle-colleagues are all you need. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7860– 7869, 2020.

[6] Yihao Zhao, Ruihai Wu, and Hao Dong. Unpaired image-to- image translation using adversarial consistency loss. arXiv preprint arXiv:2003.04858, 2020.

[7] Xinrui Wang and Jinze Yu. Learning to cartoonize using white-box cartoon representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8090–8099, 2020.

[8] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. Cartoongan: Generative adversarial networks for photo cartoonization. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 9465–9474, 2018.

[9] Jie Chen, Gang Liu, and Xin Chen. Animegan: A novel lightweight gan for photo animation. In International Symposium on Intelligence Computation and Applications, pages 242–256. Springer, 2019.

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

[11] 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.

[12] Kalischek, Nikolai, Jan D. Wegner, and Konrad Schindler. "In the light of feature distributions: moment matching for Neural Style Transfer." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[13] Xinyang Li, Shengchuan Zhang, Jie Hu and Liujuan Cao et al. Image-to-image translation via hierarchical style disentanglement[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8639-8648.

[14] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8188–8197, 2020.