

Adaptive Content Feature Enhancement GAN for Multimodal Selfie to Anime Translation

BMVC 2021 Submission # 1231

7 Supplementary Material

In this supplementary material, we provide more additional results that are organized as below:

- Sec.7.1 Additional qualitative results compared with the state-of-the-art baselines.
- Sec.7.2 Additional qualitative results of our model test on selfie2anime dataset and CelebA-HQ dataset.
- Sec.7.3 Additional results in 256×256 resolution image.
- Sec.7.4 The illustration of network architecture used in the main paper.

7.1 Qualitative results compared with baselines

In this subsection, we show additional results compared with baselines [1, 2, 3, 4]. As mentioned above, we define **content** as the contours of the face and hair, some items (such as eyes and hats). And the **style** is defined as animation rendering, skin tone, hair color, eyes etc. Additional results further prove our model can improve the preservation of content feature of input images. The latent-guided translated images produced by each model are shown in Figure 1. The latent-guided translation is that uses latent code to generate vectors as style vectors to synthesis output images. In another word, it is translated input images with random style vectors. The reference-guided translated images produced by each model are shown in Figure 2.

7.2 Qualitative results from our model

In this subsection, we show additional translated results generated by our model. The latent-guided translated images which are test on selfie2anime dataset produced by our model are shown in Figure 3. Moreover, the model trained on selfie2anime dataset is used for testing on the CelebA-HQ dataset. The latent-guided and reference-guided translated images which are test on CelebA-HQ produced by our model are shown in Figures 4 and 5, respectively.

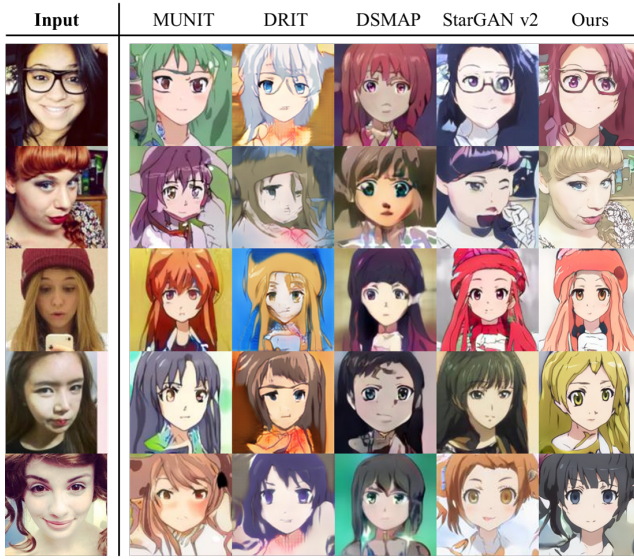


Figure 1: Qualitative comparison of latent-guided translation results. From left to right: input selfie, MUNIT, DRIT, DSMAP, StarGAN v2 and our model.

Method	StarGAN v2	DSMAP	Ours
FID	115.53	116.97	85.10
LPIPS	0.3552	0.4163	0.3431

Table 1: FID and LPIPS results compare to StarGAN v2 and DSMAP in 256×256 resolution. A low FID indicates high visual quality. A low LPIPS indicates that the translated images will not have structural changes depending on the different reference images.

7.3 Qualitative results and quantitative results compared with baselines in 256×256 resolution image.

In this subsection, we show additional results compared with StarGAN v2 and DSMAP [14, 21] in 256×256 resolution images. As shown in Figure 6, when using training image with a resolution of 256×256 , the quality of the image generated by StarGAN v2 is not as good as when the resolution is 128×128 . Obviously, the images generated by our model are of high quality, and the content feature is also preserved. As shown in Table 1, our model achieves the lowest FID and LPIPS compared to the baselines. In the FID results, better results than translated images in resolution of 128×128 were obtained.

7.4 Illustration of network architecture

We present details of architecture of subnetworks: (1) the content encoder (See Figure 7); (2) the decoder (See Figure 8); (3) the style encoder (See Figure 9); (4) the mapping network (See Figure 10); (5) the discriminator (See Figure 11).

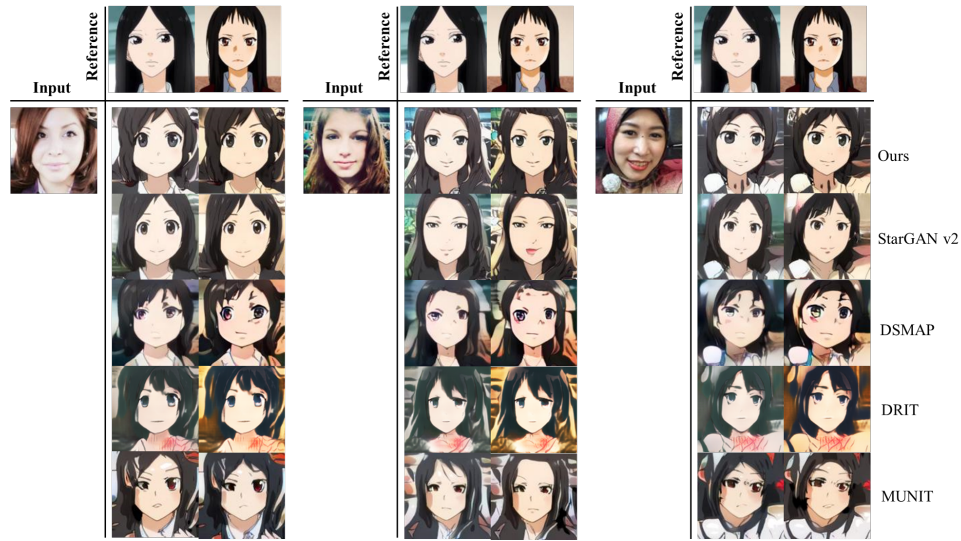


Figure 2: Qualitative comparison of reference-guided translation results. Each model translates the input selfie into anime domain and reflecting the styles of the references images.

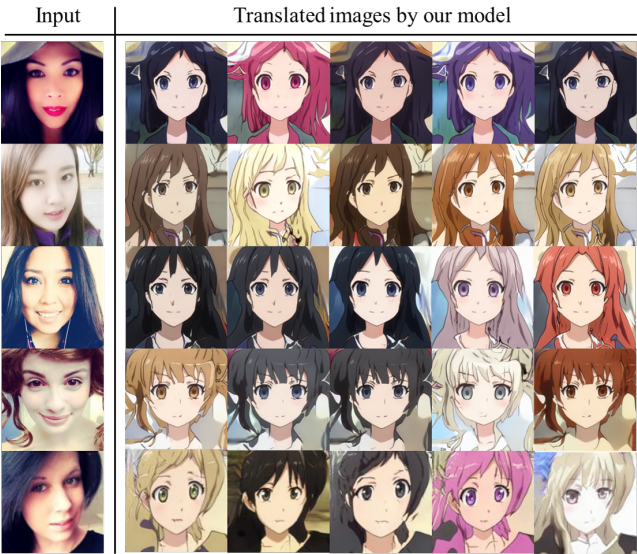


Figure 3: Additional latent-guided translation results on selfie2anime dataset from our model.

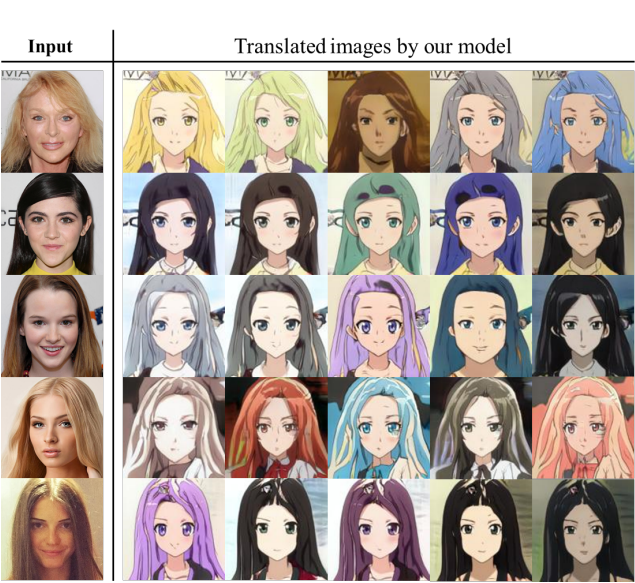


Figure 4: Latent-guided translation results on celebA-HQ dataset from our model which is trained on selfie2anime dataset.

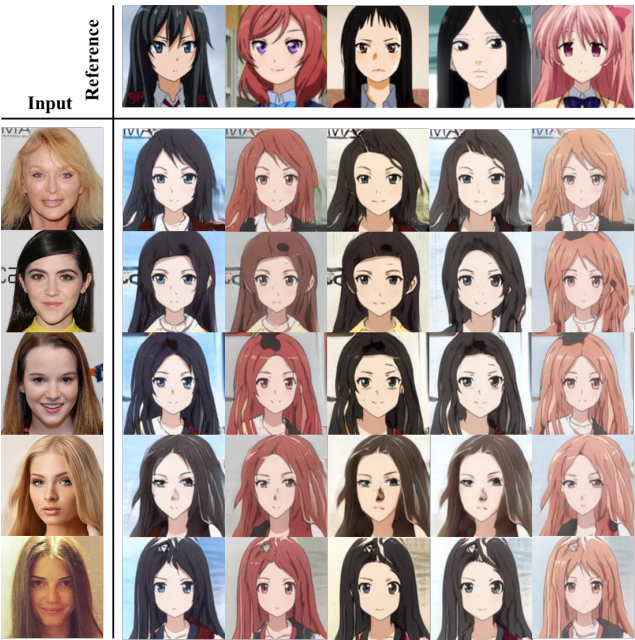


Figure 5: Reference-guided translation results on celebA-HQ dataset from our model which is trained on selfie2anime dataset.

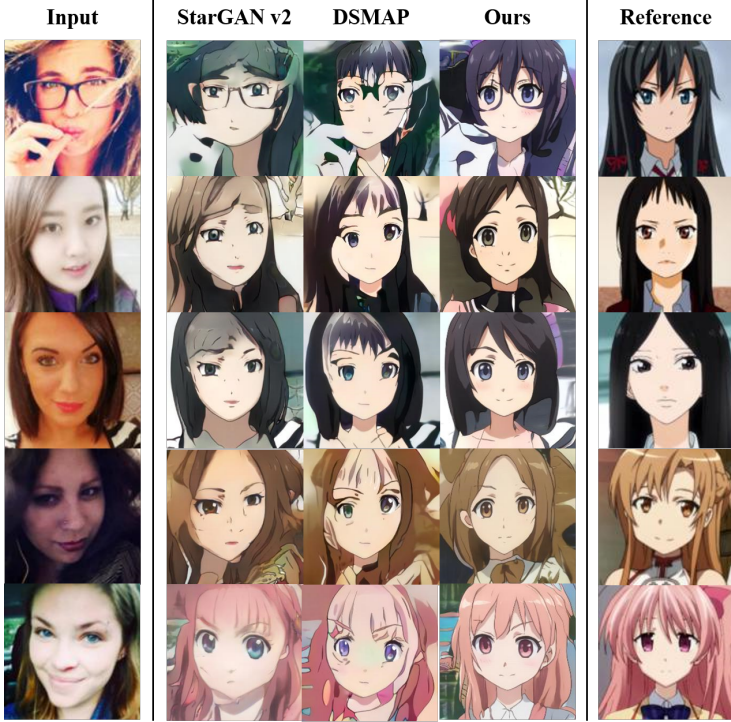


Figure 6: Reference-guided translation results. From left to right:input selfie, StarGAN v2, DSMAP, ours and reference image.

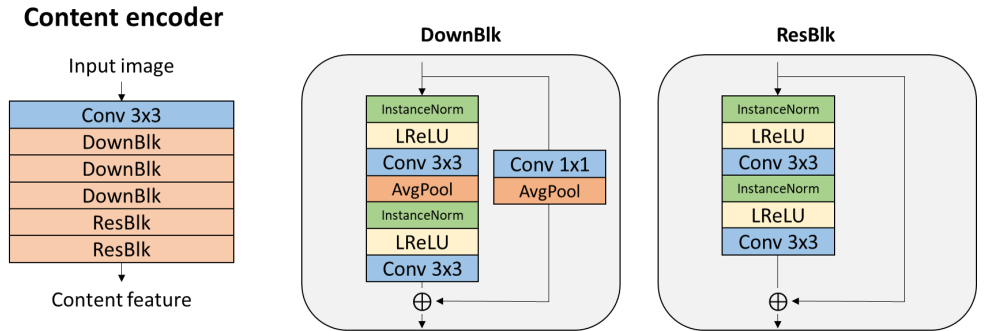


Figure 7: The architecture of the content encoder. An input image, i.e., $I_x \in R^{128 \times 128 \times 3}$, is converted to content feature with the output size in $R^{16 \times 16 \times 512}$.

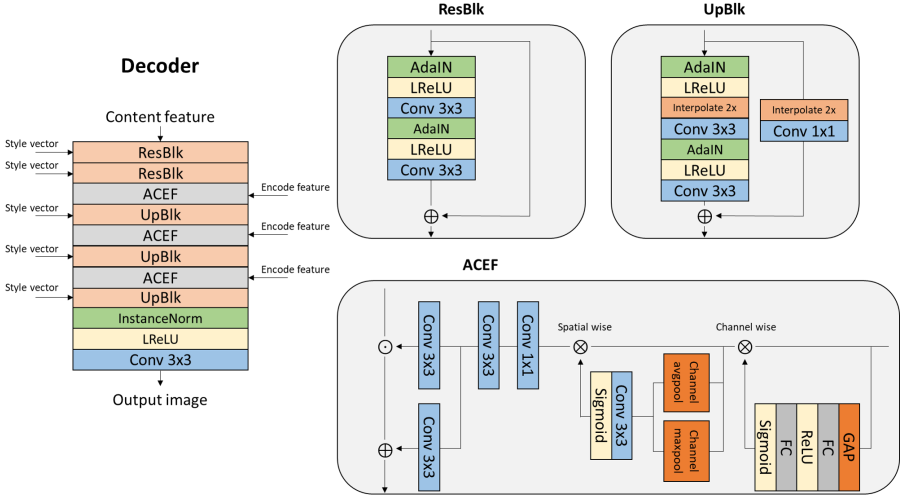


Figure 8: The architecture of the decoder. The content feature with the size $R^{16 \times 16 \times 512}$ is converted to output image with the output size in $R^{128 \times 128 \times 3}$. The style vectors are used to normalize the decode features by AdaIN [1]. The encode features are used to enhance the content feature by ACFE block.

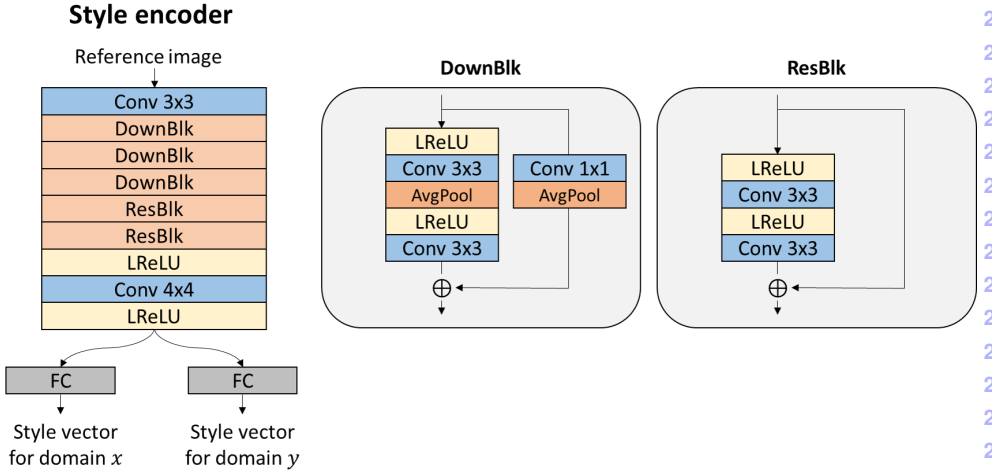


Figure 9: The architecture of the style encoder. An reference image, i.e., $I_y \in R^{128 \times 128 \times 3}$, is converted to style vector (also called style feature) with the output size in $R^{64 \times 1}$ for each domain.

Mapping network

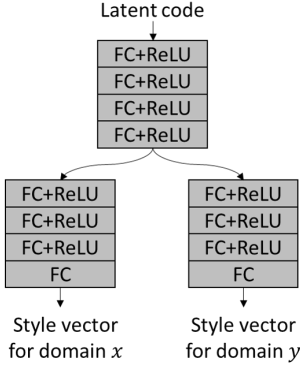


Figure 10: The architecture of the mapping network. The latent code with the size $R^{16 \times 1}$ is converted to style vector with the output size in $R^{64 \times 1}$ for each domain.

Discriminator

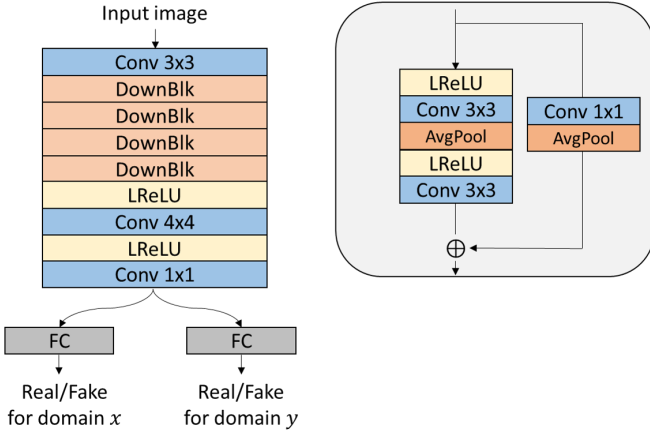


Figure 11: The architecture of the discriminator. An input image, i.e., $I_{x,y} \in R^{128 \times 128 \times 3}$, is converted to scalar for each domain.

References

[1] Hsin-Yu Chang, Zhixiang Wang, and Yung-Yu Chuang. Domain-specific mappings for generative adversarial style transfer. In *ECCV*, pages 573–589. Springer, 2020.

[2] 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.

[3] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017.

[4] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 172–189, 2018.

[5] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, pages 35–51, 2018.

322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367