Parsing-Conditioned Anime Translation: A New Dataset and Method – Supplementary Materials –

ZHANSHENG LI†, South China University of Technology, China and Singapore Management University, Singapore

YANGYANG XU*, The University of Hong Kong, China and South China University of Technology, China

NANXUAN ZHAO, Adobe Research, USA

YANG ZHOU, South China University of Technology, China and Singapore Management University, Singapore

YONGTUO LIU, University of Amsterdam, Netherlands and South China University of Technology, China

DAHUA LIN, The Chinese University of Hong Kong, China

SHENGFENG HE‡, Singapore Management University, Singapore


ACM Reference Format:

In this supplement, we provide more experimental details, comparisons, ablation studies, and applications of our method.

1 MORE EXPERIMENTAL RESULTS

1.1 Details of Compared Methods

We first detailledly introduce our compared methods. CycleGAN [Zhu et al. 2017] uses the cycle-consistency loss to constrain the translated image. The image should be translated back to be the same as the source image. UNIT [Liu et al. 2017] assumes that a paired image shares the same latent space. Based on the same idea, UI2I [Huang et al. 2021] further proposes the model transformation on the pre-trained StyleGAN2 generator. Note that UI2I has the same structure as Toonify [Pinkney and Adler 2020], thus we only show the results of UI2I here. MUNIT [Huang et al. 2018] and DRIT++ [Lee et al. 2020] are both proposed for the multi-modal image translation and assume that the shared latent space can be decomposed into the structure and appearance spaces. CouncilGAN [Nizan and Tal 2020] introduces the collaboration between multiple generators that avoids cycle-consistency constraints. ACLGAN [Zhao et al. 2020] introduces the adversarial-consistency loss that encourages the translated image to include more features from the sources. AnimeGANv2 [Chen et al. 2019] combines neural style transfer with GANs that unsupervised learns an image filter to produce cartoon-like effects. ReStyle [Alaluf et al. 2021] introduces a residual encoder to translate an image from one domain to another. By leveraging CLIP models [Radford et al. 2021], StyleGAN-NADA [Gal et al. 2021] presents a text-driven method that allows shifting a generative model to new domains. All compared methods are re-trained using the official codes on our proposed dataset for fair comparisons.

1.2 More Qualitative Comparisons

Other than the presented qualitative comparisons in the main paper, we also show more comparisons with UNIT [Liu et al. 2017], CouncilGAN [Nizan and Tal 2020], ACLGAN [Zhao et al. 2020], SG-NADA with “Anime”, “Photo” as text prompts (i.e., T-SG-NADA) and the original AnimeGANv2 [Chen et al. 2019] (i.e., without retraining) in Fig. 12, and Fig. 13 (the original AnimeGANv2 can be only applied to portrait-to-anime translation). We can see that none of them translate the visual-pleasure results, on the opposite, our method translates the most favorable results both on structure deformation and appearance consistency. The original AnimeGANv2 produces similar results to our re-trained ones, as anime data is only used to provide the unsupervised GAN loss for training their cartoon-like image filter.

1.3 User Study

Except for objective evaluations, we also introduce user studies in our evaluation. The user studies involve 50 participants that include 30 graduated students and 20 bachelors with majors in computer science. There are mainly two questions. The first one evaluates the content consistency between the source inputs and the translated ones. Given a source input, participants are asked to choose the most consistent one from the translated results acquired by different methods. Participants should consider the facial components, hair, expression, and colors. Another question cares about the perceptual quality of translated results, evaluating the visual quality solely on...
In this section, we provide more ablation studies on our StyleAnime.

2.1 The Effect of $\alpha$ Value in Latent Space Interpolation

The $\alpha$ value in Eq. (14) of the main paper controls the proportion of initial latent code $w^x$ and the anchor $w^x_{\text{anchor}}$. The interpolation results between two latent codes with different $\alpha$ values can be seen in Fig. 15. When $\alpha = 0.0$ in Eq. (14) of the main paper, the predicted animes are controlled by the initial latent code $w^x$ only, and when $\alpha = 1.0$, animes are controlled by the anchor code $w^x_{\text{anchor}}$. The anime determined by $w^x$ has a good structure consistent with the source input except for the unnoticeable artifacts and the pale

![Fig. 12. Portrait-to-Anime qualitative comparison of our method with UNIT [Liu et al. 2017], CouncilGAN [Nizan and Tal 2020], ACLGAN [Zhao et al. 2020], T-SG-NADA [Gal et al. 2021], and the original AnimeGANv2 [Chen et al. 2019].](image1)

![Fig. 13. Anime-to-Portrait qualitative comparison of our method with UNIT [Liu et al. 2017], CouncilGAN [Nizan and Tal 2020], ACLGAN [Zhao et al. 2020], and T-SG-NADA [Gal et al. 2021].](image2)

![Fig. 14. Results of the user study on portrait-to-anime task (a) and anime-to-portrait task (b) respectively. We show the percentage of preferred votes over different methods.](image3)
2.2 The Effect of Pose

We also present the visual results by translating the same portrait under different poses, as shown in Fig. 16. Our StyleAnime can handle the pose changes that are retained from the portraits. The translated animes share the same identity, showing the stability and consistency of our model.

2.3 Robustness on Predicted Parsing Maps

For analyzing the generalization ability of our model, we demonstrate the translation on real portraits and animes using predicted parsing maps. In case of no ground truth parsing map is available, we first predict the face parsing of input portraits or animes using the pre-trained DANet [Fu et al. 2019], then feeding the predicted parsing to our model instead of the ground truth parsing. We present the predicted parsing maps and translation results in Fig. 17. We can see that the predicted parsing maps are sufficiently accurate, and our model can synthesize plausible and consistent faces on these real-world cases, showing that our model has a strong generalization ability in practical scenarios.

2.4 Robustness on Large Rotation Profiles

We also present the translation results on large rotation profile portraits in Fig. 19. We can see that our model achieves plausible and consistent results on inputs with medium angles (e.g., 45°), but fails on inputs with large angles (e.g., 90°). This is largely due to the profile animes with large angles (e.g., 90°) are unavailable in our training dataset. See our discussion of possible solution on this problem in the limitation part of the main text.

2.5 Diversity on Different Kinds of Faces

To examine the tolerance of diverse input faces, we translate different kinds of faces into animes (i.e., South Asian, and East Asian people). The translated animes are shown in Fig. 18. We can see that the translated animes show consistent style and identity with the inputs, and present diverse results. Although our StyleAnime is trained on the FFHQ dataset [Karras et al. 2019] and most of the faces are Caucasians, our model can be generalized to other kinds of faces. Note that results from other methods have serious artifacts either on structure deformation or appearance consistency.

2.6 Full Cycle Translation

We proved the effectiveness of StyleAnime on Portrait-to-Anime and Anime-to-Portrait in the main paper. To further demonstrate the advantages of our method in maintaining identity consistency, we conducted a full cycle experiment of Portrait-Anime-Portrait. As shown in Fig. 20, our results maintain the identity and appearance of the original portraits in these full cycle examples.
3 MORE APPLICATIONS

In this section, we show more applications that are supported by our StyleAnime, including appearance-based anime editing and semantic-based anime editing. We also present the data processing of manga translation application.

3.1 Appearance-based Editing

In the main paper, we present the application of parsing-based editing. Not only editing on the structure, changing animes’ appearance (i.e., style) without influencing its original structure is also desired. Our model also supports appearance-only editing. To achieve this, we control the appearance code by feeding different appearance images into the appearance branch or our disentangled encoders. We show the editing results in Fig. 21. The structures of the resulting animes are fixed given the structure images shown in the first column. The appearances are controlled by images shown in the first row, and they can be either portraits or animes. Note how the appearances are adapted to the input appearance images, such as the color of hair.
Appearance

Structure

Expression

Pose

Fig. 21. Appearance editing on anime generation. The appearances of resulting animes can be controlled by the images shown in the 1st row, while the structures are maintained.

(a) Input  (b) Rec.  (c) Edited

Fig. 22. Real anime semantic editing results using SeFa [Shen and Zhou 2021]. (b) shows the reconstructed results.

3.2 Semantic-based Editing

Except for the parsing-based and appearance-based editing methods, our fine-tuned anime generator also supports the latent semantic editing [Härkönen et al. 2020; Shen et al. 2020; Shen and Zhou 2021]. In this case, image editing can be achieved by moving the latent code along a specific semantic direction. Here we first encode the real anime to the \( W^+ \) latent space of StyleGAN to obtain the latent code, then we add a direction on the latent code (‘expression’ and ‘pose’ directions in our case, and directions are obtained by SeFa [Shen and Zhou 2021]). The restructured and edited faces are shown in Fig. 22. We can see that the specific semantic attributes are modified with the corresponding latent direction, which indicates that the latent space of our generator is also structurally organized, interpretable, and editable.

3.3 Manga Translation Data Processing

As presented in the main paper, our StyleAnime supports manga translation successfully. Here we provide more data processing details of this application.

Compared with portraits and animes, a large scale of manga portraits is hard to obtain, hence we introduce a novel method for collecting the training data. For each of the anime in our Danbooru-Parsing dataset, we create its manga version using [Illvysvel 2017], and named the anime-manga dataset as Danbooru-Manga dataset. In this way, the parsing maps can be derived to manga directly from animes. Then we use those manga-parsing pairs to train a new parsing model (i.e., DANet [Fu et al. 2019] in this paper) to handle...
the real manga images. During the testing, the parsing model can
generate faithful parsing maps also on the real manga images, as
shown in Fig. 23.

3.4 Manga-to-Anime Translation
After obtaining the real manga’s parsing, we can translate the manga
to animes by our disentangled encoders. We feed manga parsing
into the parsing encoder, and the appearance code comes from a
reference image. Note here we did not re-train the encoders on
the manga data, and the manga-to-anime translation results are
shown in Fig. 24. We can find that the translated results preserve
the structure as the input mangas, and their appearances follow
the corresponding referenced animes. Though we didn’t retrain
the encoders on the manga images, our disentangled style-code
encoders trained on the anime and portrait can generalize well on
the manga data.

REFERENCES
Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. 2021. Restyle: A residual-based
stylegan encoder via iterative refinement. In ICCV. 6711–6720.
Stylegan-nada: Clip-guided domain adaptation of image generators. arXiv preprint
Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. Ganspace:
Jialu Huang, Sam Kwong, and Jing Liao. 2021. Unsupervised image-to-image translation
Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised
image-to-image translation. In ECCV. 172–189.
Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh,
and Ming-Hsuan Yang. 2020. Drit++: Diverse image-to-image translation via disen-
Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image
Ori Niraz and Ayellet Tal. 2020. Breaking the cycle-colleagues are all you need. In
CVPR. 7860–7869.
Justin NM Pinkney and Doron Adler. 2020. Resolution Dependent GAN Interpolation
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini
Learning transferable visual models from natural language supervision. In ICLR.
8748–8763.
Yujun Shen, Jinjin Gu, Xiaoniu Tang, and Bolei Zhou. 2020. Interpreting the latent space
of gans for semantic face editing. In CVPR. 9243–9252.
Yujun Shen and Bolei Zhou. 2021. Closed-form factorization of latent semantics in
Gans. In CVPR. 1532–1540.
Yihao Zhao, Ruihai Wu, and Hao Dong. 2020. Unpaired image-to-image translation
using adversarial consistency loss. In ECCV. 800–815.
Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-