COMP4471 Milestone
ANIMEGAN: Face Manipulation on Anime Illustrations

Huang Jiaxin
HKUST
jhuangbo@connect.ust.hk

Shao Yuheng
HKUST
yshaoam@connect.ust.hk

Zhao Yizhe
HKUST
yzhaocj@connect.ust.hk

Figure 1: Random samples from our dataset (above) and existing anime dataset (below).

Abstract

Image attribute editing is well studied after GAN came out. Nice models exist on human facial editing works, but they are not suitable for anime illustrations' attribute editing. In the work so far, we have built a high-quality anime face dataset and applied some existing models on it to find issues through the evaluation of results. The next step is to improve the network architecture and training process to get a stable and high-quality model demonstrated by qualitative and quantitative result evaluation.

1. Introduction

Anime fans often imagine how their favorite characters would change at their will. As deep learning has succeeded in human face manipulation, it's time to introduce a way by which everyone can enjoy "dressing up" their ideal anime girls.

1.1. Problem

Generating user-defined features from any anime girl illustration with decent accuracy and quality.

1.2. Overall Plan for Approaching the Problem

We will first build our own high quality and labeled anime face dataset. Then we will modify and train Generative Adversarial Network on our dataset to implement a stable and high-quality model learning from the empirical result of IcGan[1], FaderNet[2], AttGan[3], StarGan[4] and STGan[5].

2. Problem Statement

2.1. Dataset

It is noticed that a well-organized anime face dataset is currently unavailable. Hence, we build our own dataset as follows:

Setting Image Source. Images on commercial game websites are typically of higher quality than those uploaded by individual anime fans. Higher quality is evaluated as: 1. Stable illustration standard. 2. Uniform white background. 3. Diversified illustration style. A comparison is shown in Figure 1.

Image Scraping. We scripted and downloaded over 47,000 images from over 12,000 individual websites.

Image Preprocessing. lbpcascade_animeface and OpenCV is used to identify character faces and crop them into 128 × 128 px head portraits.

Filtering and Labeling. Illustration2vec is used to label the dataset. We also move on to utilize the same tool for filtering. About 37,000 images are left after this step. random samples from our final dataset shown in Figure 1.

2.2. Expected Results and Evaluation

We expect our network to outperform previous GAN structures on various evaluation metrics.

Qualitative Evaluation. By putting together our results with results generated by comparable networks, observable improvements can be concluded.

Quantitative Evaluation. PSNR(Peak Signal to Noise Ratio)/SSIM(Structural Similarity Index) is used to evaluate reconstruction quality. We expect the scores to outperform these networks with a relatively large margin

after fine-tuning the network structure as well as hyperparameters.

We also plan to train a high-accuracy classifier using the same training set and a subset of labels as classes. We will check the accuracy of the generated feature using this classifier.

3. Technical Approach

After building the dataset (details in section 2.1), we will train existing models such as IcGan, FaderNet, AttGan, StarGan and STGan on it to discover their shortcomings. From the empirical results, we will try to improve the training process to make the models suitable on our dataset. If possible, we also want to improve the network architecture including generator architecture and discriminator architecture. Both qualitative evaluation and quantitative evaluation will then be applied on the generated results to evaluate the performance (details in section 2.2).

4. Intermediate/Preliminary Results

This section will analyze the preliminary performance of ANIMEGAN from the perspective of reconstruction, attribute editing capability and future extension. Comparison with similar networks is also detailed.

4.1. Reconstruction

As being a widely adopted metric to evaluate the visual quality of generated images, we list PSNR/SSIM results in Table 1 and compare it with IcGan, FaderNet, AttGan, StarGan and the original STGan. Benefited from the idea of difference-attribute-vector, ANIMEGAN exhibits higher reconstruction capability than IcGan, AttGan and StarGan and is comparable to STGan. This is consistent with visual intuition. However, note that illustration pictures are much less detailed than photo-realistic images, on which the above-mentioned networks are trained. Consequently, we expect our PSNR/SSIM score to outperform these networks with a large margin after fine-tuning the network structure as well as hyperparameters.

4.2. Attribute Editing

As shown in Figure 2, the preliminary ANIMEGAN is able to capture large facial attributes, e.g. hair, and successfully apply a change of color to it. Several interesting properties of the network are revealed from the first row. Firstly, even when its hair style differs from other samples significantly, the network is still able to accurately capture its region, suggesting that the network has high capability of spatial cognition. Further, although having a single eye may seem peculiar to a naïve classifier, the image goes through ANIMEGAN properly without being attempted to add the missing eye on. This shows that neither the discriminator overfits nor the generator simply memorizes samples seen, even when our dataset is relatively small. Having this said, we conclude that the network in general is highly extendable and allows further scaling. However, as is exhibited in the fourth row, when Purple Hair is applied, surprisingly, a blue hair is resulted. The reason is unclear but one certain thing is that the feature of hair color is not learned simply via RGB values, which, in this case, suffices and digging into more intrinsic property even does harm to it.

	PSNR	SSIM
IcGAN	15.28	0.430
FaderNet	30.62	0.908
AttGAN	24.07	0.841
StarGAN	22.80	0.819
STGAN	31.67	0.948
ANIMEGAN	28.18	0.88

Table 1: Reconstruction evaluation of GAN variants.

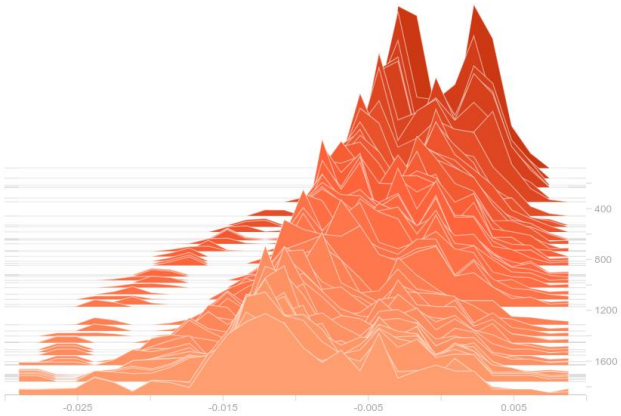

Figure 3: Generator weights distribution during training. X-axis: weight parameters. Y-axis: number of iterations.

4.3. Future Extension

As shown in Figure 2, Unfortunately, some subtle attributes fail to be extracted by ANIMEGAN. They either possess a small region in image (e.g. eye color) or are relatively abstract (e.g. smile). Thus, we plan to augment ANIMEGAN by providing incentives to the network to focus on these subtle areas, as is demonstrated by SC-FEGAN, where a special pretraining procedure is enforced. To be more concrete, random masks with sizes and positions similar to the eye and mouths are applied on the sample before feeding them into the generator.

Figure 2: Preliminary results of ANIMEGAN for facial attribute editing.

Secondly, noticing that the generator weights exhibits instability during training (Figure 3), we intend to mitigate this problem by trying a combination of some state-of-the-art techniques, e.g. ReZero proposed by Bachlechner, et al[6]. This may also help to extract subtle attributes, where learning is not efficient due to gradient vanishing.

References

[1] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M Alvarez. Invertible conditional gans for image editing. arXiv preprint *arXiv:1611.06355*, 2016.

[2] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5967–5976, 2017.

[3] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Arbitrary facial attribute editing: Only change what you want. arXiv preprint *arXiv:1711.10678*, 2017.

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[5] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo and Shilei Wen. STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing. arXiv preprint *arXiv:1904.09709*, 2019.

[6] Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, and Garrison W. Cottrell, Julian McAuley. ReZero is All You Need: Fast Convergence at Large Depth. arXiv preprint *arXiv:2003.04887*, 2020.