



# Towards Diverse Anime Face Generation: Active Label Completion and Style Feature Network

H. Li<sup>†</sup>  and T. Han 

AI Lab, ZhongAn Information Technology Service Co., Ltd., Shanghai, 200002, China



**Figure 1:** Illustration of diverse anime faces generated. (a) The generated male and female faces, and their difference respectively from the left to right. (b) The left two columns are created with the simple GA-GAN. The right two are generated with the SGA-GAN on various style features. (c) The predicted quality of the synthetic face images is High, Median, and Low respectively. (d) Comparison with face images generated using ACGAN.

## Abstract

It is interesting to use an anime face as personal virtual image to replace the traditional sequence code. To generate diverse anime faces, this paper proposes a style-gender based anime GAN (SGA-GAN), where the gender is directly conditioned to ensure the gender differentiation, and style features serve as a condition to guarantee the style diversity. To extract style features, we train a style feature network (SFN) as a multi-task classifier to simultaneously fulfill gender classification, style classification, and image quality estimation. To make full use of available data, partly labeled or unlabeled, during the SFN training, we propose a label completion method to actively complete the missing gender or style labels. The active label completion is essentially a weakly-supervised learning process through ensembling three distinct classifiers to improve the generalization capability. Experiments verify that the active label completion can improve the model accuracy and the style feature as a condition can make better the diversity of generated anime faces.

## 1. Introduction

With the rapid development of Internet applications, personal virtual assets, such as personal images, tokens, etc., have attracted more and more attention of Internet enterprises. Such virtual assets are usually bound to a unique sequence code, which seems tedious and difficult for people to remember and distinguish due to the lack

of visual and intuitive feelings. Therefore, in virtual asset related applications, it is more interesting and practical to generate a personal virtual image instead of sequence code for each user. One possible way of meeting this application requirement is to create diverse anime faces as personal virtual images.

Due to the powerful ability of generative models, generative adversarial network (GAN) has achieved great success in many tasks. GAN is a system of two neural networks contesting with each other

<sup>†</sup> lihongyu@zhongan.io

in unsupervised machine learning [GPAM\*14], usually requiring a huge amount of data for training.

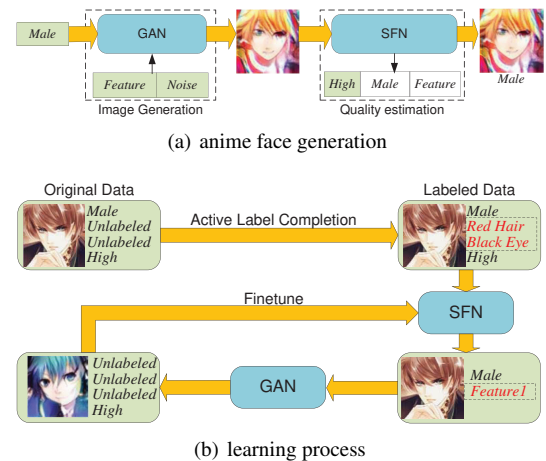
For the purpose of the anime face generation, three conditions must be satisfied that, i) the generated female and male faces must be differentiable; ii) each generated face must be characteristic and diverse; iii) the generated faces must be of high quality, i.e. clear enough. Unfortunately, however, either gender or style labels are generally in shortage in available datasets. As a consequence, it is impossible to directly apply such data to training the expected generative models. In addition, even if trained with careful design and clean dataset, the existing GAN models still often generate unclear or low-quality anime faces.

In this paper, we attempt to fill in the gap between the requirement of creating personal virtual images and the unavailability of effective generative models to some extent and our contributions are twofold. On one hand, to solve the issue involving label shortage, this paper proposes an active label completion method before style feature learning. In essence, the active label completion is a weakly-supervised learning process through aggregating three distinct learners, deep neural networks, to improve the generalization capability. In this way, missing parts in both style and gender labels of collected data can be predicted and completed with a higher accuracy than only using a single neural network. With complete labels, a style feature network (SFN) can be well trained to extract style features used in the following GAN. On the other hand, towards diverse anime face generation, we design a style-gender based anime GAN, for short SGA-GAN, by conditioning the gender attribute and style features so as to ensure the gender differentiation and style diversity. In addition, although the original training data are of high quality, the generated anime faces are still uneven in quality. As a result, we need to filter out those of low quality to meet the requirements in real applications. The face image quality is estimated still with the SFN finetuned with generated face images of various quality.

## 2. Method

The proposed anime face generation is divided into two stages, image generation and quality estimation, as shown in Fig. 2(a). The generative adversarial network (GAN) model is used to create anime face images in the image generation stage, where the gender is required to be specified and the style feature and noise are randomly selected. The style feature network (SFN) model estimates the quality of the generated image in the quality estimation stage, where the image must be regenerated once its quality is unsatisfied.

In order to learn the above-mentioned SFN and GAN models, we need to collect enough training data with the corresponding labels. Due to the lack of sufficient labels in available datasets, we propose to actively complete them in a way of ensembling three weakly-supervised classifiers. With the complete labels, the SFN is trained to extract the style feature, i.e., the last fully-connected vector in the network. The gender label and the style feature are combined together to form a characteristic data, which is the unique to an anime face and can thus be applied to training a conditional GAN. To make it have the ability of estimating the quality of the generated images, the SFN is needed to be finetuned with the newly generated



**Figure 2:** (a) The image generation stage randomly creates an anime face image, the quality of which is evaluated by the SFN in the quality estimation stage. (b) The whole process is mainly composed of active label completion, SFN training, anime GAN training and SFN finetuning with image quality.

images and their annotated quality. In sum, as presented in Fig. 2(b), the whole learning process is mainly composed of active label completion, SFN training, anime GAN training and SFN finetuning with image quality. Due to the limit of space, please refer to the supplementary file for the pseudo codes involving the anime face generation and model learning process.

### 2.1. Active Label Completion

In our anime face synthesis, there exist three integrant tasks which require three different ground-truth labels during training. One is the gender label  $y^g$  used to distinguish the male and female faces, as the generated face must be same as the specified one. Second, the style labels  $y^s$ , such as *Eye Color*, *Hair Color* and etc., are needed to train a good generator that can create diverse faces. Furthermore, to estimate the face image quality, it is also necessary to provide the quality label  $y^q$  in our work.

Unfortunately, the currently available datasets have only partial labels. In particular, the gender and quality labels are not generally provided and the style attributes are often either incomplete or missing. To solve the existing problems in image labels, we propose an ensemble learning method to actively complete the missing gender and style labels, called *active label completion*. Specifically, we first manually annotate a fraction of gender labels and then use them together with the incomplete style labels to train multiple separate classifiers in a weakly-supervised learning way. In this study, three classifiers with different backbone networks, respectively ResNet [HZRS16], DenseNet [HLvdMW17] and SeNet [HSS18], are trained and ensembled through averaging their output probabilities. The missing labels are finally completed with the classification results obtained with the ensemble classifier, as shown in the top of Fig.3. Through avoiding the overfitting of a single classifier, the ensemble way can more correctly infer those missing labels for either partly labeled or unlabeled images.

To make each classifier  $f$  achieve our goals, we define a new loss function as follows,

$$\mathcal{L}(f) = w^g \mathcal{L}^g(y^g, f) + w^s \mathcal{L}^s(y^s, f) + w^q \mathcal{L}^q(y^q, f) \quad (1)$$

where weights  $w^g$ ,  $w^s$  and  $w^q$  are used to balance the losses of three tasks,  $\mathcal{L}^g$ ,  $\mathcal{L}^s$  and  $\mathcal{L}^q$ . To make full use of the available label information, the loss function is deliberately designed to only penalize the labeled attributes. That is, during the loss computation, for the  $i$ -th image, the weights  $w_i^g$  and  $w_i^s$  are set to 0 for unlabeled attributes, 1 for labeled attributes.

As the quality of original anime face images are all high and there do not exist low-quality data, it is impossible to directly train the model for quality estimation. As a result, we do not use the quality attribute during training and thus set the weight  $w^q$  to 0 at this phase. We will discuss how to use the quality attribute to finetune our network model in more detail in subsection 2.4.

## 2.2. Style Feature Network

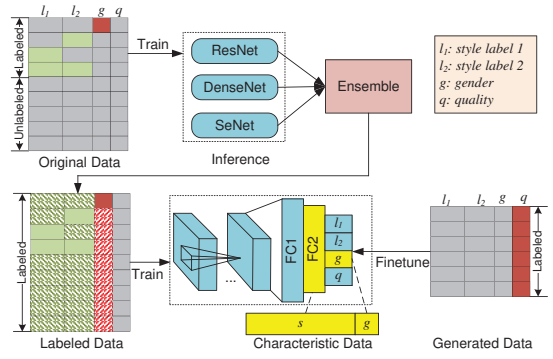
Although both the gender and style labels of all images are estimated after active label completion, the style attributes are still not suitable for directly training the GAN model. Firstly, the obtained style labels are noisy to some extent and may consequently make the GAN model drifted away from the ideal one. Secondly, the inner correlation among style attributes, e.g., *Pink Hair* is closer to *Red Hair* than *Blue Hair*, is hard to be described during the anime GAN training. In addition, our empirical studies have shown that the conditional GAN training is more inclined to fail when an input condition has many discrete assigned values and is thus high-dimensional. Based on such considerations, we propose a style feature network (SFN) to extract more robust style features of face images in place of style labels.

The SFN is a multi-task classifier to simultaneously fulfill gender classification, style classification, and image quality estimation, as demonstrated in the bottom of Fig. 3. In the SFN, ResNet [HZRS16] is employed as the backbone, where the final FC layer is of 512 dimensions and produces non-negative values after the Rectified Linear Unit (ReLU) activation. Since the high-dimensional and non-negative values are unsuitable as the condition of the anime GAN, we add an extra FC layer of only 30 dimensions after the original FC layer. The 30-dimensional FC vectors will be extracted as style features to form the characteristic data with the gender attribute as the condition of the anime GAN.

The loss function is still defined as in Eq. 1 during the SFN training. Because the original labels are more trustworthy than the estimated ones, they are supposed to be endowed with more confidence. For this purpose, if the corresponding label is actively completed, we set the weight  $w_i^g$  or  $w_i^s$  to 0.2 in the loss function, otherwise keep it as before. The quality estimation task will be done later through finetuning.

## 2.3. Anime GAN

Motivated by the idea of the conditional GAN in [OOS17], we propose a style-gender based anime GAN, for short SGA-GAN, which directly conditions the characteristic data to train the generator and



**Figure 3:** Top: Ensemble learning with incomplete labels for label completion. Bottom: SFN training with complete labels to extract style features for characteristic data and finetuning with generated images of different quality.

discriminator. Since the characteristic data comprise of style feature and gender attribute, the style diversity and gender distinction can be guaranteed for generated images.

The SRResNet [LWS\*17] is employed as the backbone in the SGA-GAN. The generator  $G$  aims to create a fake image  $X_{fake}$  with random noise on the conditions of gender  $g$  and style feature  $s$ . Therefore, the total loss  $\mathcal{L}(G)$  for the generator is composed of three parts, the adversarial loss  $\mathcal{L}_a(G)$ , the gender loss  $\mathcal{L}_g(G)$ , and the style loss  $\mathcal{L}_s(G)$ . It can be formulated as,

$$\mathcal{L}(G) = w_a \mathcal{L}_a(G) + \mathcal{L}_g(G) + w_s \mathcal{L}_s(G), \quad (2)$$

where

$$\mathcal{L}_a(G) = E[\log P(S = fake | X_{fake})],$$

$$\mathcal{L}_g(G) = -E[\log P(C = g | X_{fake})],$$

$$\mathcal{L}_s(G) = -E[\text{MSE}(F, s | X_{fake})].$$

Here the Euclidean loss  $\text{MSE}(\cdot, \cdot)$  is used to measure the distance between the input and output style features. The discriminator has a similar definition on the loss function, except that the condition is a real image  $X_{real}$  and the adversarial part changes into,

$$\mathcal{L}_a(D) = -E[\log P(S = real | X_{real})] - E[\log P(S = fake | X_{fake})].$$

During the SGA-GAN training,  $w_a$  and  $w_s$  are set to 2 and 0.3 respectively to balance the effect of different distributions of gender and style features. The noise and style feature are randomly generated under truncated normal distribution  $[-1, 1]$ . If  $w_s$  becomes 0, the SGA-GAN will turn into a simpler gender based anime GAN (GA-GAN), which only conditions the gender class to train the networks.

## 2.4. SFN Finetuning

The anime faces generated with the SGA-GAN are uneven in image quality. To filter out low-quality images, we continue to finetune the SFN model with the generated data. At this time, the weight  $w^q$  in the loss function Eq. 1 is relaxed and adjustable. Several thousands of generated images are picked out and manually annotated with

**Table 1:** Gender classification

Method	ResNet	DenseNet	SeNet	Ensemble	SFN
Acc.(%)	84.01	85.90	84.75	<b>86.89</b>	86.80

different quality, *Low*, *Median* or *High*. In practice, the quality estimation task is casted as a regression problem and three levels of quality, *Low*, *Median*, and *High*, are respectively qualified to be 0, 0.5, and 1. In our quality definition, images of *High* quality are supposed to be clear, good-looking in perception, and with regular facial configuration. The *Median* quality means that the image should have regular facial features, but seems either blurry or bad-looking. The *Low*-quality image is considered to be distorted or even not a facial image.

### 3. Experiments

In our experiments, we used two datasets from [JZL\*17] and [Pav] as the training set. In the first dataset comprising 32897 images, 17395 images are offered incomplete style labels and the total number of male images is quite small. So more male faces were fetched from 6232 images in the second dataset where the style labels are not provided. Here we picked out over 2K male and about 4K female images and annotated them with the gender attribute. Meanwhile, two labels, *Hair Color* and *Eye Color*, were selected as style attributes. In addition, all face images were detected, cropped, and resized to  $96 \times 96$  pixels before training. Over 1K images with well balanced gender attribute were held out for validating the performance of different models on gender classification.

**Gender Classification:** As discussed previously, annotated style attributes are probably inaccurate. Therefore, in estimating active label completion, we chose gender classification on different models for comparative analysis. A fraction of data was first used to train and ensemble three independent models, ResNet, DenseNet, and SeNet. With the ensemble classifier, all data can be completed with missing labels. The SFN was finally trained on all data and its accuracy of gender classification achieved 86.80%, close to the ensemble result 86.89% and over 2% higher than 84.01% originally obtained with the ResNet, as illustrated in Table 1. This indicates that active label completion can effectively improve the generalization capability of the SFN and result in better representative style features for characteristic data.

**Diversity Analysis:** To observe the distinction between generated female and male face images, we first conducted studies with the simple GA-GAN where only the gender attribute was changed and the gender-unrelated information, i.e., the noise, was fixed. Fig. 1(a) presents some images generated in this way, where the first two columns are respectively the generated male and female face images and the third is the difference between the first two columns. The image difference shows that the gender is distinguishable in appearance although with the similar style, in accordance with our visual perception that the male images usually have large nose but small eyes.

To evaluate the style diversity, we compared the results obtained with GA-GAN and SGA-GAN respectively, and can perceptually

feel that the style with SGA-GAN is more diverse and the face gender is more distinguishable from Fig. 1(b). For the purpose of quantitative evaluation, we computed the Fréchet Inception Distance (FID) score as the diversity indicator [HRU\*17]. The results show that the SGA-GAN performs better than the GA-GAN since its FID scores, 58.8 (*male*) and 61.9 (*female*), are much lower than those of the GA-GAN, 77.7 (*male*) and 77.8 (*female*).

**Image Quality Evaluation:** Fig. 1(c) presents some generated face images with different quality. It is observed that the *High*-quality images (left column) are usually well looking, and in the *Median*-quality images (middle column) there are obvious facial characteristics like eye, nose, mouth, but the detail or the facial configuration is not good enough. The *Low* quality (right column) generally has unexpected artifacts, such as distortion or eye disappearing, which will be filtered out for real application. This indicates the SFN after finetuning can well distinguish the quality of generated anime face images. As a comparison, Fig. 1(d) presents some face images generated using ACGAN [JZL\*17], where only the minority of produced faces were male and either low-quality or unexpected images occurred more frequently, e.g., the left-most column.

### 4. Conclusions

In this paper, we propose an active label completion method to predict missing parts in both style and gender labels through aggregating distinct classifiers. Furthermore, the complete labels make it possible to train a style feature network and learn style features for anime face generation. We also design a style-gender based anime GAN to create personal anime face images, where style features serve as a condition to guarantee the style diversity. To prevent the unclear or distorted face images from occurring, we achieve image quality estimation through finetuning the style feature network.

### References

- [GPAM\*14] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial nets. In *NIPS* (2014), pp. 2672–2680. 2
- [HLvdMW17] HUANG G., LIU Z., VAN DER MAATEN L., WEINBERGER K. Q.: Densely connected convolutional networks. In *CVPR* (2017), pp. 2261–2269. 2
- [HRU\*17] HEUSEL M., RAMSAUER H., UNTERTHINER T., NESSLER B., HOCHREITER S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS* (2017), pp. 6629–6640. 4
- [HSS18] HU J., SHEN L., SUN G.: Squeeze-and-excitation networks. In *CVPR* (2018), pp. 7132–7141. 2
- [HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *CVPR* (2016), pp. 770–778. 2, 3
- [JZL\*17] JIN Y., ZHANG J., LI M., TIAN Y., ZHU H., FANG Z.: Towards the automatic anime characters creation with generative adversarial networks. *arXiv preprint arXiv:1708.05509* (2017). 4
- [LWS\*17] LEDIG C., WANG Z., SHI W., THEIS L., HUSZAR F., CABALLERO J., CUNNINGHAM A., ACOSTA A., AITKEN A., TEJANI A.: Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR* (2017), pp. 105–114. 3
- [OOS17] ODENA A., OLAH C., SHLENS J.: Conditional image synthesis with auxiliary classifier gans. In *ICML* (2017), pp. 2642–2651. 3
- [Pav] PAVITRAKUMAR78: Anime-face-gan-keras. <https://github.com/pavitrakumar78/Anime-Face-GAN-Keras>. 2017. 4