2020 25th International Computer Conference, Computer Society of Iran (CSICC), Tehran, Iran

StarGAN Based Facial Expression Transfer for Anime Characters

Majid Mobini, Foad Ghaderi*
Human Computer Interaction Lab., Electrical & Computer Engineering Department, Tarbiat Modares University, Tehran, Iran
{m.mobini, fghaderi}@modares.ac.ir

***Abstract*—Human facial expression transfer has been well explored using Generative Adversarial Networks. Also, in case of anime style images, several successful attempts have been made to generate high-quality anime face images using GAN approach. However, the task of anime facial expression transfer is not well studied yet due to the lack of a clean labeled anime dataset. We address this issue from both data and model perspectives, by providing a clean labeled anime dataset and leveraging the use of the StarGAN image-to-image translation framework. Our collected dataset consists of about 5k high-quality anime face images including five major emotions collected from online image boards. We preprocessed our dataset by CARN super-resolution technique to improve quality of the images, and applied tuned StarGAN model to learn the mapping of an input anime image with arbitrary expression to the target expression. We evaluate our work by visually comparing the output translated results with the baseline model. Moreover, we provide a quantitative analysis of our proposed approach by computing the confusion matrix of expression transfer accuracy.**

Keywords—Facial Expression Transfer, Unpaired Image Translation, Generative Adversarial Network, Anime Generation

I. Introduction

The anime industry in east Asian countries especially Japan is growing very fast in a manner that the Hollywood Reporter published new record-breaking news[1] that Japan's anime industry hit a record revenue of $19.1 billion total in 2017. Considering the heavy workload required for designing anime characters for each anime series, creating and styling them with an automated approach could provide a significant cost reduction.

Generative Adversarial Networks (GANs), proposed by Goodfellow et al. [1], is an unsupervised learning technique that achieved surprisingly successful results in image generation tasks. The idea is to build not one, but two competing deep neural networks which are trained simultaneously in a mini-max game. Using GAN approach, recent researches provide several successful attempts to high-quality anime face generation [2, 3, 4], however, the task of anime facial expression transfer is still not well studied due to the lack of clean anime datasets that include emotion labels.

In this paper, we propose a model that is able to transfer anime character facial expression to the desired one with a promising rate of success. Overall, our contributions are as follows:

- We provide a clean labeled dataset, collected from Danbooru [2] and Getchu [3] online image boards, including five anime major emotions (i.e. happy, sad, crying, neutral, and surprised) with average one thousand images per each class. Besides that, we implement an open-source mobile application to facilitate process of labeling images and detecting false positives among them by experts.
- By proper use of StarGAN [5] framework as our base model for facial expression transfer, applied on preprocessed images using CARN super-resolution [6] model and also data augmentation techniques, we are able to transfer our input anime face images to the desired expression with promising success rate.

II. Related Works

Generative Adversarial Networks achieved promising results in a diverse range of computer vision tasks such as image generation, image translation, super-resolution imaging, and facial expression transfer. A typical GAN consists of a generator and a discriminator model which are trained simultaneously. The generator tries to fool the discriminator by generating fake images that the discriminator is unable to distinguish them from real ones, while the discriminator tries to distinguish the real images from the fake ones. Considering this mini-max game, GANs' power lay on the idea of an *adversarial loss*, that forces the generated images to be indistinguishable from real ones.

Several extensions of GANs were proposed in order to make control over generation process, such as CGAN [7] and ACGAN [8]. They generally take extra information (such as labels) as a part of the input to satisfy specific conditions using an auxiliary classifier.

Image-to-Image translation is to learn the mapping from a set of input images (input domain) to a set of output images (target domain), e.g. facial expression transfer. Depending on training data, domain transfer could be done in a supervised manner when we have access to paired training data, or unsupervised when aligned data is not available. Pix2Pix [9] is a popular supervised image-to-image translation framework which formulated paired image transfer as a general conditional GAN problem that not only learns the mapping between input image to output image, but also learns a loss function to train this mapping. To address the limitation of paired data, several methods tackle the unpaired setting. For instance, CycleGAN [10] learns the mapping from an

1 https://www.hollywoodreporter.com/news/2017-anime-industry-revenue-hits-a-record-19-billion-1167382
2 https://danbooru.donmai.us
3 http://getchu.com

978-1-7281-5937-9/20/$31.00 ©2020 IEEE

unpaired input domain to the output domain by combining adversarial loss with *cycle consistency loss*. The key idea is that generator network tries to reconstruct the original image from the fake one and compute the L1 norm as a cyclic loss. While both Pix2Pix and CycleGAN are only capable of learning the relations between two different domains at a time, StarGAN proposed a framework for multi-domain unpaired image-to-image translation, that learns the mappings between all available domains using only one generator. The idea is to learn to flexibly translate the input image into the corresponding domain, instead of learning a fixed translation.

Anime Face Generation using GAN approach was first explored by Mattya [11] and Rezoolab [12] following introduction of DCGAN [13]. By disentangling anime content and style, Xiang S and Li H [3] were able to generate anime portraits with a fixed content and a large variety of styles from different artists. PSGAN [4] generated full-body high-resolution anime character images by progressively increasing the resolution of both generated images and structural conditions during training. Jin Y et al. [2] proposed a conditional anime face generation framework based on DRAGAN [19] that was able to generate high quality anime faces. They provided a dataset for anime face data by crawling Getchu website, extracted 34 labels automatically using illustration2Vec [14], and postprocess result by SRGAN [15] for super-resolution imaging. However their labels covered only basic visual features like hair and eye colors, hat, glasses and etc. and were not rich enough for facial expression synthesis.

Unlike the above frameworks, we focused on collecting a clean dataset including facial expression of anime images. We used StarGAN framework as our baseline model for facial attribute transfer and CARN super-resolution framework as our preprocessing step for noise reduction.

III. Proposed Method

Before we describe our model for attribute transfer, we explain how we collect and label our anime face dataset. Then we demonstrate architecture and techniques used to make training more stable.

A. Data Preparation

Having a clean and balanced dataset is a key factor in successful training of GAN based models. There were several attempts for collecting high-quality anime face datasets scrapped from anime imageboards such as Danbooru (a free image hosting web service that users can upload anime pictures along with their tags). However, these datasets suffer from inter-image variance and noise [2]. To tackle this problem, Jin Y et al. [2] crawled standard images of anime games' characters provided by Getchu website. Images crawled from Getchu are more clean with higher quality but the tag data is not available for them. To overcome this issue, they used Illustration2Vec model to estimate the tags of the images automatically, however, output tags are not optimized for facial expression applications and is more related to visual features such as eye and hair color.

As mentioned above, none of existing anime face datasets have proper labels for facial expression. So at first step, we decided to create a well-suited anime face dataset for this purpose. We started to collect data by targeting Danbooru imageboard as images are already tagged there by the users. We've selected 'happy', 'sad', 'crying', and 'surprised' as our basic emotions based on popularity of published posts for each tag. Our dataset preparation approach is relatively similar to the method explained in [2]:

1. First we collected all images with mentioned tags from Danbooru website using the crawler tool gallery-dl[4]. We exclude 'manga' keyword from search result to limit our dataset to RGB images.
2. To detect faces, we used 'lbpcascade animeface' [16] pretrained cartoon face detector. We discard detected faces with confidence score less than 80%. As all faces in a picture are not necessarily corresponding to the image tags, we apply a heuristic and discard images that are detected to have more than 6 faces.
3. Using the result of the face detector, we know the location of eyes, mouth and nose. So we rotate the images such that the center of the eyes lie on the same horizontal line. We use length of eyes distance multiplied by 1.35 for selecting bounding box square around the face.
4. In the final step, we manually removed false positives because of errors in face detector or irrelevant tag for the cropped part (as we could have multiple faces in each image, the assigned tag may be valid only for some of them).

For facilitating time consuming process of removing false positives and make it possible to change the wrong tag (expression) to the correct one we implemented a mobile application, as shown in Figure 1. This dataset preparation software consists of a web service providing an API for specifying final label per each image or mark them as false positives, along with a mobile application frontend for ease of use and simplicity of working with the API by experts.

Also, we faced lack of data for some emotions like 'neutral' which there's no such tagged images in Danbooru website or 'sad' which number of tagged images are less than expected. So, we used same data preparation process for Getchu website to enrich our dataset, but considering the absence of tags for this image board, we tagged them manually using implemented mobile application labeling facilitator. Table I demonstrates summary of total number of extracted face images for each emotion after removing false positives.

[4] https://github.com/mikf/gallery-dl

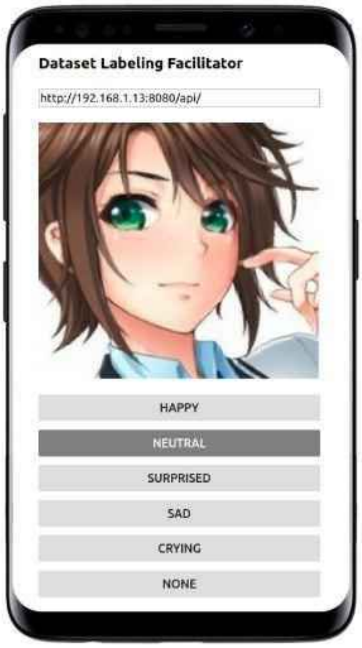


Fig. 1. Mobile application user interface implemented by us to facilitate removing false positives and correcting expression labels by experts.

TABLE I. DATA COLLECTION OUTCOME. THE MAIN SOURCES OF CRAWLING IMAGES FOR EACH EMOTION, KEYWORD USED FOR CRAWLING EACH EXPRESSION (SPECIFIC TO DANBOORU WEBSITE), AND THE TOTAL NUMBER OF EXTRACTED FACE IMAGES AFTER REMOVING FALSE POSITIVES.

Expression	Main Source	Keyword	Total Number [a]
happy	Danbooru	happy-manga	2041
sad	Danbooru	sad-manga	620
crying	Danbooru	crying-manga	686
neutral	Getchu	neutral-manga	1396
surprised	Getchu	surprised-manga	452

[a.] Total number of extracted face images after removing false positives.

B. Data Preprocessing & Augmentation

Almost seventy percent of our collected data from Getchu website are small square images that have a width/height of between 92px up to 112px. So we used the CARN super-resolution framework with 2x scale factor as a preprocessing step for increasing resolutions of images and also improving the quality of them by reducing noise. Finally, we scaled all of the images to 224 x 224 pixels. To compensate for the lack of data, we used a random flip followed by a random crop to extract a 192x192 face from each image. Using this data augmentation approach, we generate eight augmented images per each original sample.

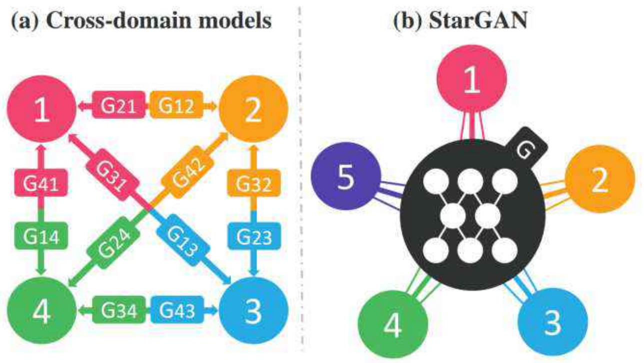
C. Model

Recent studies for image-to-image translation are designed to learn mapping between only two different domains. As a result, when dealing with multi-domain image translation, you need to train these models multiple times for every pair of two different domains. StarGAN solved this issue by adding an auxiliary domain classifier to the discriminator network. As shown in Figure 2, instead of training multiple times in cross-domain models, StarGAN efficiently learns the mapping between different domains using single generator network. We choose StarGAN as our baseline model, not only because of significant cost reduction in training process, but also it could use global information of other expressions to learn mappings of two selected ones.

Fig. 2. Training process for cross-domain models comparing with the StarGAN model. [5]

Adapted from CycleGAN, StarGAN has the generator network composed of two convolutional layer for down-sampling, six residual blocks as bottleneck layers, and two transposed convolutional layers for up-sampling. In the discriminator network, it uses a PatchGAN [9, 10, 17] based architecture which classifies local patches independent of faces being real or fake, and also determines the actual expression class that they belong to using an auxiliary domain classifier.

StarGAN objective functions to optimize discriminator and generator network are written in equations 1 and 2



respectively. In these equations, we can see three loss functions:

- Adversarial Loss (L_{adv}) is basic GAN loss function proposed by Goodfellow [1] and represents how much fake images are indistinguishable from the real ones. StarGAN extends this loss function to be conditioned not only on input image but also on the target domain label.
- Domain Classification Loss (L_{cls}) represents how much auxiliary domain classifier is successful in translating images to the target domain.
- Reconstruction Loss (L_{rec}) represents how much the generator network is successful in reconstructing input image from the translated one. This loss function applied to preserve content and identity of the input image while changing domain specific characteristics of them.

$$L_{Discriminator} = -L_{adv} + \lambda_{cls} L_{cls} \tag{1}$$

$$L_{Generator} = L_{adv} + \lambda_{cls} L_{cls} + \lambda_{rec} L_{rec} \tag{2}$$

While training StarGAN on our custom dataset with its default parameters for RaFD [18] dataset, we observe that classification loss for both discriminator and generator networks converge to zero after some epochs and model is not able to recover it. To stabilize training and overcome this issue, we tried to use a bigger generator network by increasing number of bottleneck layers from six to nine and number of up/down sampling layers from two to three. Loss functions per iteration charts in Figure 3 show that our networks are much more stable after this change.

Also we observe that generator network is only able to modify face images partially specially in mouth part. To make generator more flexible and strong to change different parts of the face, we increase generator network deceptive field and also decrease reconstruction loss coefficient (λ_{rec}) from ten to five.

D. Evaluation

We trained both baseline model and our proposed approach on our custom dataset for 250k iterations. Using a Nvidia 1080 TI GPU and the batch size equals to 16, it takes about one day and four hours to complete training process. For qualitive evaluation, we compare visual results of our proposed approach with the baseline model. As shown in Figure 4, we created higher quality outputs comparing with the baseline models. The higher quality can be for two reasons, first data augmentation (especially random face crop) helped a lot to increase the size of the dataset and handle spatial variance of face components locations and CARN-super resolution was useful to remove unwanted noise in training images and making training more stable. And the second reason is hyper parameter tuning done specially for the generator network to make it able to compete with the discriminator network which was described in detail in the last section. For quantitative analysis, we computed confusion matrix (Table II) of our expression transfer accuracy. We selected 100 random input images for test and asked three experts to manually specify labels of the translated output images in order to measure accuracy of domain transfer using majority vote.

Input | Crying Happy Neutral Sad Surprised

TABLE II. EXPRESSION TRANSFER CONFUSION MATRIX. THIS CONFUSION MATRIX REPRESENTS ACCURACY OF DOMAIN TRANSFER FOR 100 SAMPLE IMAGES LABELED BY THREE EXPERTS MANUALLY. WE ADD 'OTHER' LABEL TO THE OUTPUT COLUMNS FOR IMAGES THAT THEIR LABELS ARE NOT CATEGORIZED TO ANY OF OUR FIVE EMOTIONS OR NOT HAVING MINIMUM QUALITY.

expected \ output	happy	sad	neutral	crying	surprised	other
happy	89	2	5	0	1	3
sad	1	74	10	8	0	7
neutral	3	9	78	1	4	5
crying	0	14	5	72	0	9
surprised	0	1	5	0	91	3

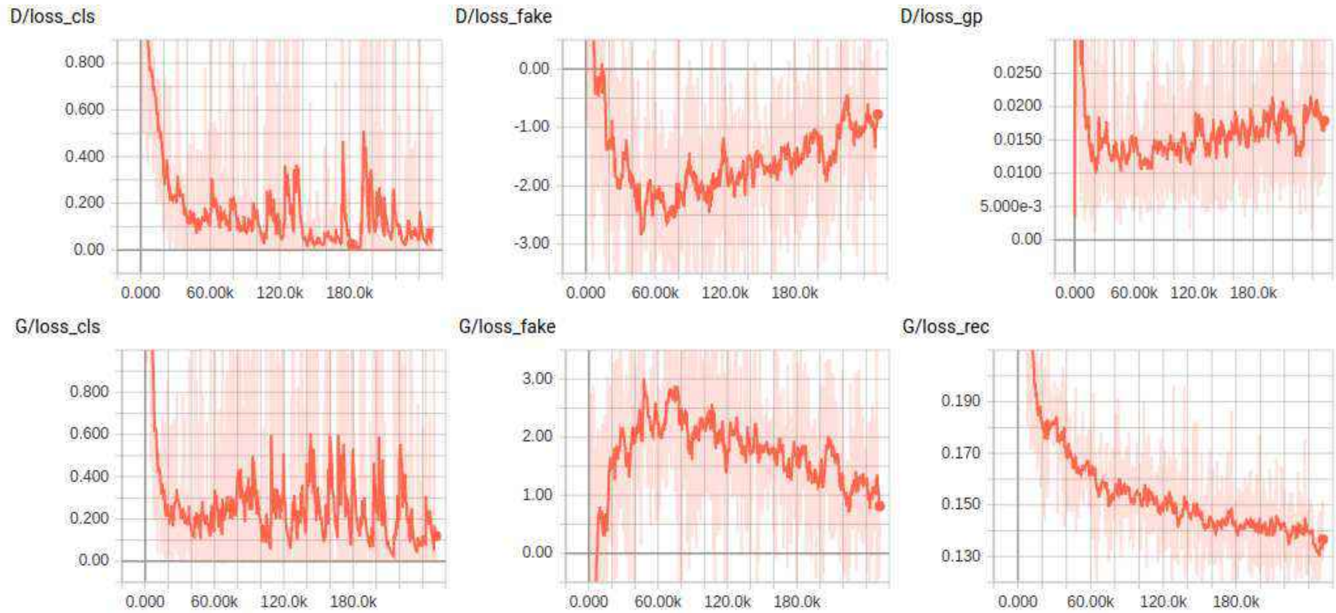


Fig. 3. Loss function per iteration for the proposed approach during the training process (outliers are ignored in chart scaling). The first row represents discriminator classification loss, adversarial loss and gradient loss per iteration. And the second row represents generator classification loss, adversarial loss and reconstruction loss per iteration.

rows are the
ther columns

Input | Crying Happy Neutral Sad Surprised

As shown in Table II, because of high correlation of face visual features between sad and crying expressions, and also between neutral and sad, we see more false positives for these facial expressions.

IV. CONCLUSION

By providing a clean dataset of anime images along with their emotion labels, and leveraging the use of the CARN super-resolution model for preprocessing, data augmentation techniques and StarGAN image-to-image translation framework, we've explored facial expression transfer for anime images which was not well studied before due to lack of a well-suited emotion-labeled dataset. The results confirm that our approach outperforms the original StarGAN in terms of quality of the translated anime images. Our future work will be towards adding more sample images to our dataset.

REFERENCES

[1] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In Advances in neural information processing systems 2014 (pp. 2672-2680).

[2] Jin Y, Zhang J, Li M, Tian Y, Zhu H. Towards the high-quality anime characters generation with generative adversarial networks. In Proceedings of the Machine Learning for Creativity and Design Workshop at NIPS 2017.

[3] Xiang S, Li H. Anime Style Space Exploration Using Metric Learning and Generative Adversarial Networks. arXiv preprint arXiv:1805.07997. 2018 May 21.

[4] Hamada K, Tachibana K, Li T, Honda H, Uchida Y. Full-body High-resolution Anime Generation with Progressive Structure-conditional Generative Adversarial Networks. In Proceedings of the European Conference on Computer Vision (ECCV) 2018 (pp. 0-0).

[5] Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018 (pp. 8789-8797).

[6] Ahn N, Kang B, Sohn KA. Fast, accurate, and lightweight super-resolution with cascading residual network. In Proceedings of the European Conference on Computer Vision (ECCV) 2018 (pp. 252-268).

[7] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784. 2014 Nov 6.

[8] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier gans. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 2017 Aug 6 (pp. 2642-2651). JMLR. Org.

[9] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 1125-1134).

[10] Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision 2017 (pp. 2223-2232).

[11] Mattya. chainer-dcgan. https://github.com/mattya/chainer-DCGAN, 2015.

[12] Rezoolab. Make illustration on computer with chainer. http://qiita.com/rezoolab/items/5cc96b6d31153e0c86bc, 2015.

[13] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434. 2015 Nov 19.

[14] Masaki Saito and Yusuke Matsui. Illustration2vec: a semantic vector representation of illustrations. In SIGGRAPH Asia 2015 Technical Briefs, page 5. ACM, 2015.

[15] Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, Shi W. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 4681-4690).

[16] Nagadomi. lbpcascade anime face detector. https://github.com/nagadomi/lbpcascade_animeface, 2014.

[17] Li C, Wand M. Precomputed real-time texture synthesis with markovian generative adversarial networks. InEuropean Conference on Computer Vision 2016 Oct 8 (pp. 702-716). Springer, Cham.

[18] Langner O, Dotsch R, Bijlstra G, Wigboldus DH, Hawk ST, Van Knippenberg AD. Presentation and validation of the Radboud Faces Database. Cognition and emotion. 2010 Dec 1;24(8):1377-88.

[19] Kodali N, Abernethy J, Hays J, Kira Z. On convergence and stability of gans. arXiv preprint arXiv:1705.07215. 2017 May 19.