Image-to-Image Translation Method for Game-Character Face Generation

Shinjin Kang
School of Games
Hongik Univeristy
Sejong, Korea
directx@hongik.ac.kr

Yoonchan Ok
NCsoft
Seoul, Korea
okchany@ncsoft.com

Hwanhee Kim
NCsoft
Seoul, Korea
greentec@ncsoft.com

Teasung Hahn
NCsoft
Seoul, Korea
spinel@ncsoft.com

***Abstract*—Traditional image-to-image translation methods effectively change the style; however, these methods have several limitations in shape changing. Particularly, current image-to-image translation technology is not effective for changing a real-world face image to the face of a virtual character. To solve this problem, we propose a novel unsupervised image-to-image translation method that is specialized in facial changes accompanied by radical shape changes. We apply two feature loss functions specialized for faces in an image-to-image translation technique based on the generative adversarial network framework. The experimental results show that the proposed method is superior to other recent image-to-image algorithms in case of face deformations.**

***Index Terms*—Image-to-image Translation, Generative Adversarial Network, Game-Character Generation**

I. Introduction

Recent development in image-to-image translation technology has provided users with a variety of content that had not been experienced before. In particular, the practical application of face transformation techniques has been successful [Snapchat (2020)] and this success has created a market for mobile applications that can stylize user-provided images on demand [Timestamp (2020)]. With the expansion of the market for content such as games, movies, and cartoons, the inclination towards the transformation of faces for virtual characters is increasing. Recent advances in deep learning technology have garnered considerable interest in this field. In particular, the unsupervised cyclic-consistent generative adversarial network (CycleGAN) algorithm [Zhu et al. (2017)] has generated various derivation studies owing to the ease of data collection and processing [Almahairi et al. (2017)] [Lu et al. (2017)]. These image-to-image translation techniques have also been applied to the field of content creation, and have successfully been utilized in manga generation [Su et al. (2020)] and image colorization [Furusawa et al. (2017)]. However, the challenge of converting an in-wild face image into a virtual character has not been addressed properly with existing image-to-image translation technology. Till date, the techniques have been well-studied in terms of styles such as colors or regional patterns of the target domain, but have shown limitations in shape learning. In this study, we propose a technique that changes a real-world face image to another face domain (game character, cartoon, etc.) with radical deformation. We newly design two loss values to handle the radical facial changes, which are motivated by [Shi et al. (2019)], and attention map and adaptive layer-instance normalization (AdaLIN) [Kim et al. (2019)]. Further, we design the network structure by integrating two feature parsing networks generated into the generative adversarial network (GAN) framework.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2019R1A2C1002525). This work was also supported by NCsoft.

978-1-7281-4533-4/20/$31.00 ©2020 IEEE

II. Method

A. Generator

Our generator model is specialized for faces and constructed to learn and generate global and local facial features. For the baseline network in our generator, we used a network configuration similar to that of the unsupervised generative attentional networks with AdaLIN for image-to-image translation (UGATIT) [Kim et al. (2019)]. The UGATIT model features an auxiliary classifier and AdaLIN; these two features of the UGATIT network have a significant advantage in shape modification, compared with existing image-to-image models. The aim of the proposed method is to develop a network specialized in shape transformation in face images. The features of a face are considerably important, compared with other image transform domains. If proper learning is not applied, particularly to the positions of the eyes, nose, and mouth, a large error occurs in the cognitive aspects, regardless of how well the other areas are learned. Therefore, stable shape changes are necessary when face-specific information is used by the network. To achieve this, we incorporated two additional facial details: face segmentation and face feature. These details were obtained from two pre-learned networks, and they generated two losses for the CycleGAN training.

The proposed network architecture is shown in Fig. 1. Let $x_s \in X_s$ and $x_t \in X_t$ represent samples from the source and target domains, respectively. Let $G_1(x_s)$ and $G_2(x_t)$ represent the translated source and target domains, respectively. Our model consists of two generators ($G_1(x_s)$ and $G_2(x_t)$), two discriminators ($D_1(G_1(x_s))$ and $D_2(G_2(x_t))$), and two feature extractors ($F_1(x_s, x_t)$ and $F_2(x_s, x_t)$). $G_1(x_s)$ creates an image that fits the target style based on the GAN framework

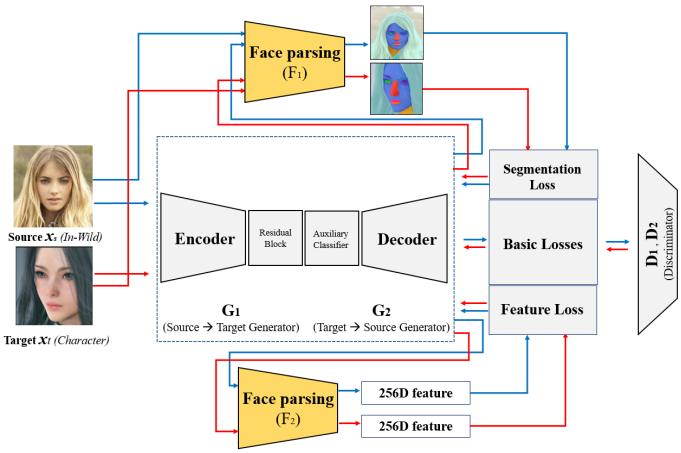

Fig. 1. Proposed network architecture. Blue line indicates the flow of information of source image x_s. Red line represents the flow of information of target image x_t.

and $G_2(x_t)$ is used for cycle consistency. The discriminators D_1 and D_2 distinguish between the real and fake translated images. The feature extractors F_1 and F_2 provide two loss values to the CycleGAN framework to facilitate shape transformation. The final loss function of our model can be written as a loss of L_{total}.

$$\underset{G_1,G_2}{argmin}\ \underset{D_1,D_2}{max}\ L_{total}(G_1, G_2, D_1, D_2, F_1, F_2) \tag{1}$$

L_{total} consists of five loss terms: L_{lsgan}, L_{cycle}, $L_{identity}$, L_{cam}, and L_{face}. The adversarial loss L_{lsgan} is employed to match the distribution of the translated images to the target image distribution. The cycle loss L_{cycle} is applied for a cycle consistency constraint to the generator. The identity loss $L_{identity}$ is used to ensure that the color distributions of the input and output images are similar. These three losses are calculated using G_1, G_2, D_1, and D_2 with the traditional GAN framework. These terms are described in detail in [Zhu et al. (2017)] and [Kim et al. (2019)]. L_{cam} uses the information from the auxiliary classifiers to determine the differences between two domains [Selvaraju et al. (2019)]. The additional face-feature loss L_{face} is the weighted summation of the segmentation loss L_{seg} and feature loss $L_{feature}$. These two losses are calculated using the segmentation and feature parsing networks.

$$\begin{aligned} L_{total} = {} & L_{lsgan}(G_1, G_2, D_1, D_2) + L_{cycle}(G_1, G_2, D_1, D_2) \\ & + L_{identity}(G1, G_2, D_1, D_2) + L_{cam}(G_1, G_2, D_1, D_2) \\ & + L_{face}(F_1, F_2, G_1, G_2) \end{aligned} \tag{2}$$

$$\begin{aligned} L_{face}(F_1, F_2) = {} & \alpha L_{seg}(F_1, G_1, G_2) \\ & + \beta L_{feature}(F_2, G_1, G_2) \end{aligned} \tag{3}$$

B. Segmentation Parsing Network

To specialize image-to-image translation on the face, we used the L_{face} loss. The purpose of the L_{face} loss is to calculate the difference between facial features in the image generated by GAN and the target x_t image on the GAN framework, and use it for the backpropagation of $G_{1,2}$. To achieve this, we used the Siamese networks [Bertinetto et al. (2016)] for F_1 and F_2. The Siamese network is composed of two convolutional neural networks (CNNs) sharing weights. The CNN converted two images i_1 and i_2 into vector representations of $F(i_1)$ and $F(i_2)$. To learn F, weights were trained in the direction of defining a loss function and creating a representation for distance. Using the face-segmentation image as an input image to the Siamese network, we considered this loss as a constraint on the shape and displacement of different face components in the two images, such as the eyes, mouth, and nose. Instead of using pre-trained models, we developed our facial segmentation model based on the VGG [Simonyan et al. (2014)], where we removed the fully connected layers. We trained this model on the semantic segmentation dataset from CelebAMask-HQ dataset [Lee et al. (2019)].

Generally, image segmentation generated in the early stage of GAN learning does not show a normal face shape; therefore, normal segmentation cannot be achieved. Because of this, it is impossible to calculate L_{face} until the eyes, nose, and mouth are constructed during training. However, as the overall learning rate is largely set during the early stage of learning and the size is reduced during the second half of learning, important face information cannot be reflected during the early stage of learning by L_{face}. To solve this problem, we added the static segmentation loss term at the beginning of learning. This term is calculated as the difference between the segmentation loss generated by inputting the source image x_s and target image x_t. The purpose of this term is to suggest the learning direction at the beginning of learning. The effect of this term is designed to continue until the segmentation loss generated by G_1 becomes less than a certain value as learning progresses, and then decrease linearly. The L_{face} loss function of the abovementioned process is defined as follows.

$$\begin{aligned} L_{seg}(x_s, x_t) = {} & \alpha_{decay} \| F_1((G_1(x_s))) - F_2((G_2(x_t))) \|_1 \\ & + \beta_{decay} \| F_1(x_s) - F_2(x_t) \|_1 \end{aligned} \tag{4}$$

In this formula, the first term is the run-time segmentation loss calculated during the learning phase and the second term is the static segmentation loss affecting the initial learning.

C. Feature Parsing Network

The objective of this study is to change the face images of people in the real world into character faces in the virtual world. The segmentation loss provides essential information for changing faces; however, some cartoons and game characters have various accessories (glasses, tattoos, ornaments, hats, beards etc.) that are different from the real-world faces. Consequently, there is a possibility that segmentation may not be performed correctly. To solve this problem, we used another Siamese network. The network was designed to generate a universal feature vector of the face. We used the pre-trained light-CNN face recognition model [Wu et al. (2018)] to extract

256-dimension facial embeddings of the two facial images and then compute the cosine distance between them as their similarity. This face recognition network was learned in a noisy label environment, and had the advantage of extracting a stable feature vector for images that had not been previously learned. It complements the weaknesses of the segmentation parsing network. If the target image type was considerably different from the normal image such that normal segmentation was impossible, the reflection weight of this network was increased. $L_{feature}$ consists of calculating cosine similarity as follows.

$$L_{feature}(x_s, x_t) = 1 - cos(F_1(G_1(x_s)), F_2(G_2(x_t))) \tag{5}$$

Unlike the segmentation loss, the initial static term is not used in this expression.

D. Discriminator

In this study, $D_{1,2}$ had a structure similar to $G_{1,2}$. However, because the image decoder module was unnecessary for D, only the encoder and auxiliary network parts were used, except for the residual network in G. A classifier that determines whether an image is real or fake was added and used instead of the decoder.

III. Experimental Results

For the source images, 20,000 images were randomly extracted from the CelebA data set. For the target images, we created a new game-character dataset. We developed a crawler in the web environment to collect the faces of characters created by customizing in the commercial MMORPG Black Desert [Black Desert (2014)]. These face images from the game were presumed to be suitable for testing in this experiment because of the latest character-face customization system and the ability to generate various face data. The total number of Black Desert face data collected was 20,000. The aim of our experiment was to compare the image-to-image translation of the CelebA dataset with that of the Black Desert dataset. These images showed various accessories, skin colors, and skin customization that would be difficult to observe in the real world, making the application of image-to-image translation techniques challenging.

All the images were resized to 112x112 pixels for training. For optimization, we set the maximum number of iterations as 100,000, learning rate as 0.001, and decay rate as 20% per 5 iterations. A stochastic gradient descent optimizer, with a batch size of eight, was used for training. The NVIDIA RTX Titan GPU required approximately two days for the training. The face alignment was performed using dlib library [Dlib (2020)] to align the input image before it was fed into the network.

First, we evaluated the quality of the proposed network. We compared our method with various models including CycleGAN [Zhu et al. (2017)], UNIT [Liu et al. (2017)], and UGATIT [Kim et al. (2019)]. All the baseline methods were implemented using the authors' code. Fig. 2 shows the comparison results. From the figure, we can observe that our result exhibits a more detailed expression of the eyes, nose, and mouth of the face compared with the face images generated by the other networks.

According to the obtained images, CycleGAN and UNIT attempted to mimic the entire face shape, but were unable to express specific details. Additionally, these models could not respond to the angle changes. UGATIT responded to the changes in facial expressions and angles to some extent, but did not accurately express them. Our model expressed the details of the face, and showed the shapes corresponding to the various face angles. The facial details and ability to respond to angle changes were observed as a result of the two losses proposed in this study, thereby verifying that the proposed method is effective.

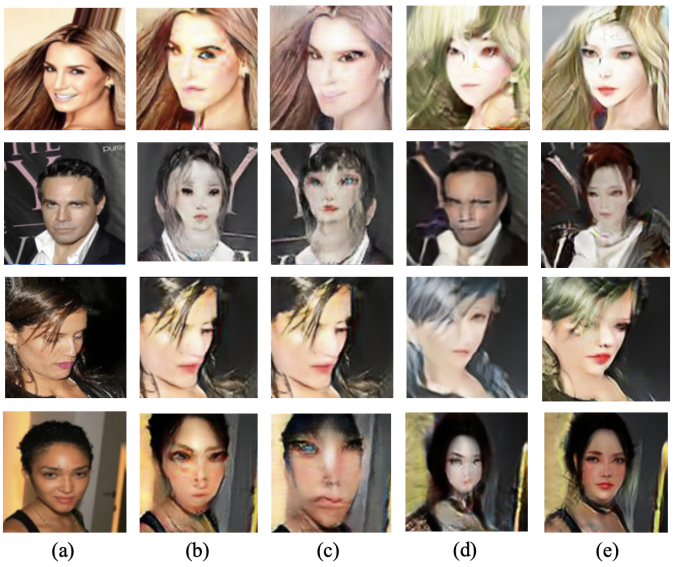

Fig. 2. Comparison between the generated faces: (a) source images; (b) CycleGAN; (c) UNIT; (d) UGATIT; (e) proposed method.

We determined the effect of the two additional losses. Fig. 3 shows the resulting images when applying each loss. The leftmost panel shows the results when the three losses (L_{lsgan} + L_{cycle} + $L_{identity}$) are used. Although it is possible to produce a result similar to the target images, it can be observed that detailed features such as the eyes, nose, and mouth are not well revealed. When feature loss is added, as shown in the images in the middle panel, it can be observed that the eyes, nose, and mouth are expressed more clearly, and color is also learned. The rightmost panel shows the results of adding the segmentation loss; all the feature details are accurately represented. Comparing the two losses, we can observe that the segmentation loss has a more substantial influence than the feature loss.

To quantitatively evaluate the similarity between the generated and in-wild face images, we used the Frechet inception distance (FID) [Heusel et al. (2017)] as our metrics. For each test image, we randomly selected an image from the generator training set as its reference and computed the average FID over

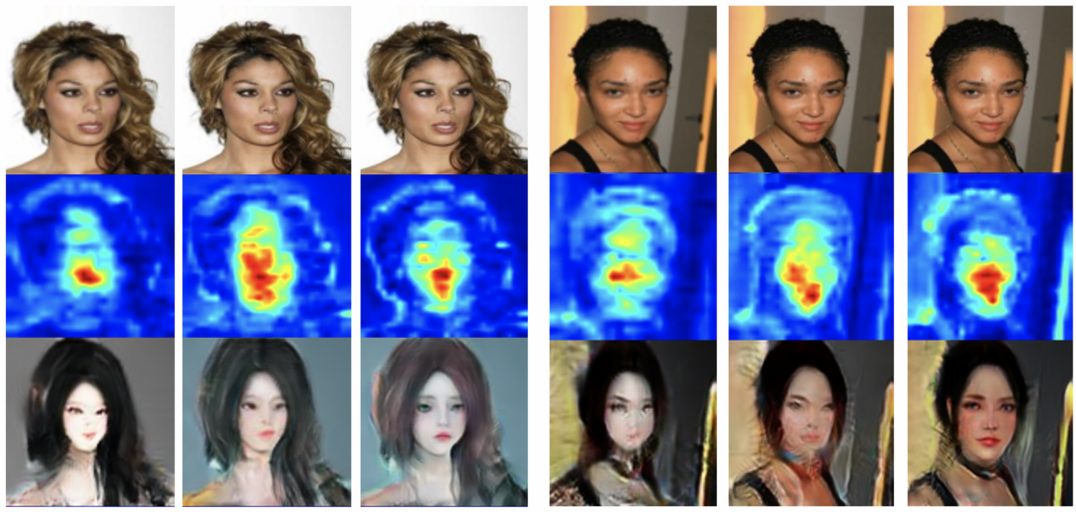

Fig. 3. Comparison with losses : (a) basic losses (L_{lsgan} + L_{cycle} + $L_{identity}$); (b) basic losses + feature $L_{feature}$; (c) basic losses + feature $L_{feature}$ + segmentation L_{seg}.

the entire test set. The FID score of our proposed model was 44.14. This score was approximately 11.36% lower than the FID score of the UGATIT network measured using the same method. This shows that the proposed model is more similar to ground truth, and produces a variety of results.

IV. Conclusions

In this study, we introduced an image-to-image translation technique between source and target images. The effectiveness of the technique was evaluated using two face-specific losses. The proposed method showed an advantage in face generation because of its unsupervised application without any additional data labeling. In particular, because it is possible to create an image similar to the target face image without separate face-related parameter data, it may be helpful for external researchers or users to create the desired image of specific content domain.

References

[Almahairi et al. (2017)] Almahairi, A., Rajeswar, S., Sordoni, A., Bachman, P., and Courville, A. (2018). Augmented cyclegan: Learning many-to-many mappings from unpaired data. arXiv preprint arXiv:1802.10151.

[Black Desert (2014)] Black Desert, Perl Abyss, https://www.blackdesertonline.com/midseason

[Bertinetto et al. (2016)] Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., and Torr, P. H. (2016, October). Fully-convolutional siamese networks for object tracking. In European conference on computer vision (pp. 850-865). Springer, Cham.

[Dlib (2020)] Dlib, http://dlib.net/

[Furusawa et al. (2017)] Furusawa, C., Hiroshiba, K., Ogaki, K., and Odagiri, Y. (2017). Comicolorization: semi-automatic manga colorization. In SIGGRAPH Asia 2017 Technical Briefs (pp. 1-4).

[Heusel et al. (2017)] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Advances in neural information processing systems (pp. 6626-6637).

[Kim et al. (2019)] Kim, J., Kim, M., Kang, H., and Lee, K. (2019). U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. arXiv preprint arXiv:1907.10830.

[Lee et al. (2019)] Lee, C. H., Liu, Z., Wu, L., and Luo, P. (2019). MaskGAN: towards diverse and interactive facial image manipulation. arXiv preprint arXiv:1907.11922.

[Liu et al. (2017)] Liu, M. Y., Breuel, T., and Kautz, J. (2017). Unsupervised image-to-image translation networks. In Advances in neural information processing systems (pp. 700-708).

[Lu et al. (2017)] Lu, Y., Tai, Y. W., and Tang, C. K. (2017). Conditional cyclegan for attribute guided face image generation. arXiv preprint arXiv:1705.09966.

[Selvaraju et al. (2019)] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[Shi et al. (2019)] Shi, T., Yuan, Y., Fan, C., Zou, Z., Shi, Z., and Liu, Y. (2019). Face-to-Parameter Translation for Game Character Auto-Creation. In Proceedings of the IEEE International Conference on Computer Vision (pp. 161-170).

[Simonyan et al. (2014)] Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[Snapchat (2020)] Snapchat https://play.google.com/store/apps/details?id=com.snapchat.android&hl=ko (accessed on 10 March 2019).

[Su et al. (2020)] Su, H., Niu, J., Liu, X., Li, Q., Cui, J., and Wan, J. (2020). Unpaired Photo-to-manga Translation Based on The Methodology of Manga Drawing. arXiv preprint arXiv:2004.10634.

[Timestamp (2020)] Timestamp Camera, Artify, https://play.google.com/store/apps/details?id=com.artifyapp.timestamp&hl=en_US

[Wu et al. (2018)] Wu, X., He, R., Sun, Z., and Tan, T. (2018). A light cnn for deep face representation with noisy labels. IEEE Transactions on Information Forensics and Security, 13(11), 2884-2896.

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