Facial Feature Based Image-to-Image Translation Method

Shinjin Kang
School of Games, Hongik University
2639 Sejong-ro, Jochiwon, Sejong, Korea
[e-mail: directx@hongik.ac.kr]
Corresponding author: Shinjin Kang

Received October 6, 2020; revised November 23, 2020; accepted December 1, 2020; published December 31, 2020

Abstract

The recent expansion of the digital content market is increasing the technical demand for various facial image transformations within the virtual environment. The recent image translation technology enables changes between various domains. However, current image-to-image translation techniques do not provide stable performance through unsupervised learning, especially for shape learning in the face transition field. This is because the face is a highly sensitive feature, and the quality of the resulting image is significantly affected, especially if the transitions in the eyes, nose, and mouth are not effectively performed. We herein propose a new unsupervised method that can transform an in-wild face image into another face style through radical transformation. Specifically, the proposed method applies two face-specific feature loss functions for a generative adversarial network. The proposed technique shows that stable domain conversion to other domains is possible while maintaining the image characteristics in the eyes, nose, and mouth.

***Keywords*:** Game Character Generation, Character Customization, Virtual Character, Image Translation, Convolutional Neural Network

A preliminary version of this paper appeared in IEEE Conference of Games 2020. 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 supported by NCsoft.

http://doi.org/10.3837/tiis.2020.12.012 **ISSN : 1976-7277**

1. Introduction

Image-based deep learning has been developing at a rapid pace in recent years. In particular, research related to face field has proved its usefulness in the field of security and beauty, showing various practical success cases [1, 2]. As digital contents such as games, movies, VR, and interactive movies have gained great popularity, interest in stable facial transformation is increasing significantly. The algorithm that has received the most attention in this field is CycleGAN [3]. This algorithm has the advantage that a large number of images can be relatively easily used for deep learning network training using a translation method [4-6]. This translation technique has also been applied to the area of content creation and has been successfully utilized in cartoon creation [7] and image coloring [8]. However, the problem of translating in-wild face images to virtual characters has not been sufficiently addressed with conventional translation techniques. Numerous techniques have been suggested for dealing with style patterns in the target domain. However, style patterns could not be effectively learned during shape learning. As a result, there were several cases in which the CycleGAN based image-to-image translation resulted in distorted facial shapes. To overcome this problem, this study proposes a method of transforming real-world face images into different face areas, especially for game characters and cartoon domain, using radial shape transformation. We applied two face specific losses [9] to ensure that face shape learning works well in the generative adversarial network (GAN) framework. In addition, we attempted to increase the stability of learning by applying an attention map and adaptive layer-instance normalization (AdaLIN) to the network [10]. Our study demonstrated that the traditional GAN-based image-to-image translation technique can be effectively applied to games and cartoons, which require rapid shape changes.

2. Related Work

Image-to-image translation studies are derived from existing neural style transfer studies. In style transfer, the translated image retains the shape of the original or content image, and only the style is changed to the one that is desired. Gatys et al. proposed a network for artistic style that separates and recombines the style and content of images [11]. Their research showed that the content and style of images can be separated by using a convolutional neural network (CNN). They demonstrated that users can easily change image properties by separating the content and style. Risser et al. added an error term that compensated for the instability that occurs when synthesizing textures using style transfer [12]. They used histogram losses to improve the stability of texture synthesis. Their results showed that it is feasible to apply integrated local losses to a multiscale framework. Luan et al. proposed Matting Laplacian to limit the input-to-output transformation to be locally similar in the color space to avoid distortion [13]. Chen et al. Introduced StyleBank, which consists of multiple convolutional filter banks, each of which represents a single characteristic style. To transfer the image to a specific style, the filter bank is convolved with intermediate feature embeddings generated by a single auto encoder [14].

Dumoulin et al. indicated that with the introduction of a simple modification of the style transfer network (conditional instance normalization), multiple styles can be learned [15]. They demonstrated that this approach is flexible, yet comparable to single-purpose style transport networks in terms of qualitative and convergent properties. Huang et al. proposed an adaptive instance normalization layer [16]. They used the variance and mean of the content features with those of the style features. Their methodology showed rapid speed improvement.

without restrictions on a predefined set of styles. Johnson et al. introduced a pretrained loss network for image classification to define a perceptual loss function that measures the perceptual difference in content and style between images [17]. Li et al. suggested using feature transformation to match the content feature statistics directly to the style image statistics in the deep feature space [18]. By combining the functional transformation with a pretrained network, the transmission process can be implemented as a feedforward operation. Lu et al. introduced a novel framework for fast semantic style transfer [19]. Their method divides the semantic style transfer problem into a feature reconstruction part and a feature decoder part. The reconstruction part tactfully solves the problem of optimizing the loss of content and loss of style in the feature space, especially through reconstructed features. Selim et al. presented a method to convey a painting texture [20] that uses the concept of gain maps to impose new space constraints during image transfer. The gain map applies to the VGG function and captures the local spatial color distribution of the example picture. This conveys the texture of the painting style in a highly effective manner while preserving the structure of the face.

Sendik et al. proposed a texture synthesis method based on CNN and used statistics of pretrained functions [21]. They presented structural energy based on the correlation between self-similarities that characterize textures and deep features that capture regularity. Zhu et al. suggested a method to capture the special characteristics of an image collection and determined a method to convert those characteristics into another in an unsupervised manner [3]. In a recent gaming domain, Shi et al. proposed a method to automatically generate a game customization character using a face photo [9]. They addressed the problem of optimization on a set of physically meaningful face parameters to formulate the generation under the face similarity measurement and parameter search paradigm. Cao et al. propose the GAN model for unpaired translation to caricature [22]. They attempted to solve the complex cross-domain transfer problem by dividing it into two easy problems, using the geometry-to-geometry transformation module and the style transfer module, separately. Lui et al. proposed an unsupervised image-to-image conversion algorithms that work on a previously unseen target class specified at test time only by referencing a few examples [23]. They combined adversarial training plans to achieve these few-shot generation abilities. AlBahar et al. proposed a bi-directional transformation method [24] to address the guided translation problem of converting an input image to another image while respecting the constraints provided by the user-supplied guide image. Park et al. suggested an image conversion method by considering the content of that patch in the input regardless of the domain [25]. Their method attempts to maximize information between the two areas using contrast learning. Royer et al. proposed a transition method to learn the mapping between corpus level styles in each collection and preserve semantic content shared across two domains [26]. They utilized two adversarial autoencoders to generate the shared representation of a common domain. Kim et al. suggested a new method for image transformation to integrate a new attention module and a normalization function in an unsupervised manner [10]. The attention module guides the network to focus on the more important areas that distinguish the source and target domains using the auxiliary classifier. The study presented in this paper was motivated by segmentation losses in Shi et al. [9] and attention map and AdaLIN in Kim et al. [10]. The ideas presented in these studies were applied in the integrated GAN framework in the present study with five losses for face-oriented image-to-image translation.

3. Method

3.1 Generator

In this study, we selected a generator and a discriminator to effectively distinguish between the facial features. In particular, the generator was selected such that the facial features could be effectively detected and then reflected on the face image of the target domain. For conversion between two different face domains, we used the UGATIT network as the basic network [10]. UGATIT network is composed of two generators and one discriminator. The generator consists of an encoder, a decoder, and an auxiliary classifier. The generator includes several activation maps, via which the features in the image can be more effectively detected. In addition, it was characterized by applying the AdaLIN normalization technique. The network to be proposed must be highly effective in converting real world images into virtual world ones. Even subtle differences in faces revealed big differences in cognitive aspects. Even in the case of the same pixel, there is a risk that the resulting image will be an extremely unstable if the pixel is incorrectly generated by the generator at the eye or nose position. Owing to this problem, the network to be applied for face image translation must be able to correctly recognize changes in the eyes, nose, mouth, and eyebrows, which are the key features of a face. To this end, we added a specialized loss to the two faces obtained from a network that was trained in advance.

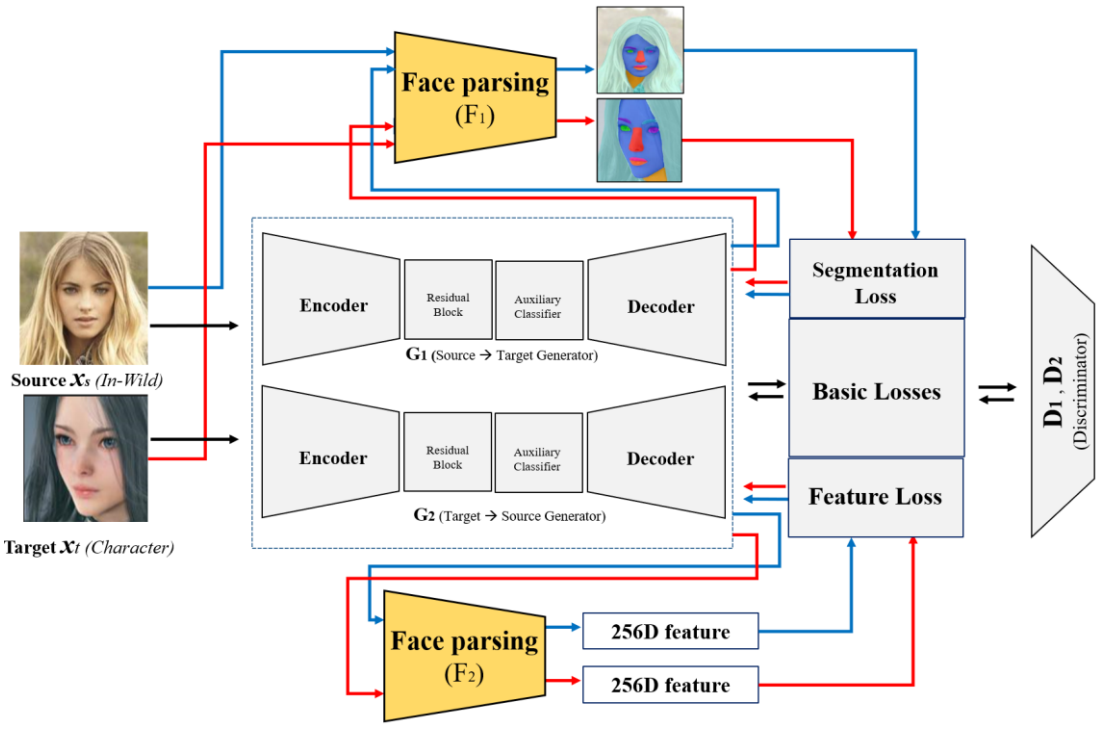

Fig. 1. The proposed network structure

Fig. 1 shows the proposed network architecture. The blue and red lines denote the information obtained from the source images and the flow of backpropagation from these images, respectively. Let $x_s \in X_s$ and $x_t \in X_t$ are samples from the source and target domains X_s, X_t. The proposed model consists of six sub networks. $G_1(x_s)$ and $G_2(x_t)$ are two generators for cycle consistency. $F_1(x_s$,

x_t) and $F_2(x_s, x_t)$ are two feature extractors, and the discriminators D_1 and D_2 determine whether the image created by the generator is a real or fake image. Here, F_1 and F_2 provide the two face-specific loss values, which enable stable shape transformation from the source to target domain. The G_1 aims to generate a real image with the target style image, and G_2 is used for the cycle loss. The final loss function L_{total} can be expressed as follows.

$$arg \min_{G_1,G_2} \max_{D_1,D_2} L_{total}(G_1, G_2, D_1, D_2, F_1, F_2) \tag{1}$$

L_{total} consists of the following loss terms: L_{lsgan}, L_{cycle}, $L_{identity}$, L_{cam}, and L_{face}. Here, the adversarial loss L_{lsgan} was adopted to make the distributions between the two image domains similar. The cycle loss L_{cycle} was used to maintain a cycle consistency between the two generators. The identity loss $L_{identity}$ was applied to match the color distributions. These three losses were calculated by G_1, G_2, D_1, and D_2 with the GAN framework. L_{cam} was calculated by the auxiliary classifiers to ensure that the feature differences are more exact [27]. A more detailed explanation of these terms is described in Sections [3] and [10]. The L_{face} is the summation of the two losses: feature loss $L_{feature}$ and segmentation loss L_{seg}. The feature and segmentation parsing networks calculated these two losses for the face-oriented image translation.

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

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

3.2 Segmentation Parsing Network

L_{face} loss was used to achieve image-to-image transition, especially with respect to faces. The L_{face} loss value was applied to reduce the feature difference between the image generated by the generator and the target image. L_{face} was used for the backpropagation of $G_{1,2}$. Siamese networks [28] were used for F_1 and F_2. Siamese networks comprise two CNNs that share weights. The CNNs convert the two images, i_1 and i_2, into vector representations, namely, F_1 (i_1) and F_2 (i_2). The proposed Siamese network was trained to generate segmented images from in-word images. Notably, the position and shape of the nose, mouth, and eyes are the key features of a face image. Therefore, we believed that if the main parts of the face are segmented and the pixel losses between them are used, the positions of the eyes, nose, and mouth can be accurately created. For this, we added the following L_{seg} term.

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

Instead of using an off-the-self model, we used the VGG [29] as the segmentation network, and for the dataset, we used CelebAMask-HQ [30]. The L1 loss between the generated segmented image and the ground truth image was used for training the network. These segmentation images could be normally generated by the generator at the beginning of learning. This is because the generator was not sufficiently trained to generate real images. To compensate for this, we used the difference in the pixel loss between the source image and the target image at the beginning of learning and used this difference between the generated image and the target image after learning to a certain extent. In the above-mentioned equation, the two terms reflect this situation.

3.3 Feature Parsing Network

We aim to transform in-wild facial images into characters in the virtual world. Segmentation loss aids in the stable generation of eyes, nose, and mouth, which are essential for the creation of faces. However, it does not guarantee the modification of various accessories or additional features attached to the face. To address this problem, in addition to the segmentation loss, a separate methodology capable of recognizing features in pixel units is needed. We used an additional feature detection network for this reason. We additionally obtained 256-dimensional feature vector values from the Light CNN-based facial feature network [31]. The difference was learned by calculating the cosine similarity. This additional loss was intended to compensate for the weakness of the segmentation loss. In the virtual world, in many cases, various types of additional decorations are attached to the face in order to reveal the character's personality. The feature loss makes the features of the virtual character stand out by expressing decorations as approximate as possible. The loss term $L_{feature}$ can be defined as follows.

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

3.4 Discriminator

$D_{1,2}$ had a similar structure to $G_{1,2}$. However, since D did not require a video decoder module, we used only the encoder and auxiliary network excluding the remaining network of G. Instead of a decoder, we added a classifier to classify real/fake image.

4. Experimental Results

We used 20,000 images from CelebA dataset for the source image. For the target image, we used 20,000 images from the MMORPG Black Desert [32]. This is a game that supports the latest game character customization function and is advantageous for deep learning because face rendering results are diverse, and the rendering quality is high. In particular, the Black Desert characters include a variety of hair styles, tattoos, beards, and hair ornaments. For this reason, if the image-to-image network cannot learn these features, it will generate very unstable images. We created a separate crawler to collect the Black Desert dataset. **Fig. 2** shows an example of the dataset. Images were scaled to 112x112 size and then used for training. The total iteration number was set to 100,000, and the learning rate was 0.001. The learning decay rate was 20% every 1/5. The batch size was 8. The Training: Validation: Test ratio was set to 8:1:1. The NVIDIA RTX Titan GPU took roughly 2 days to train. Face alignment and cropping was performed using the dlib library [33] before being fed into the network. **Fig. 3** shows a graph of the discriminator and the loss of the generator. Generator losses steadily decrease over time.

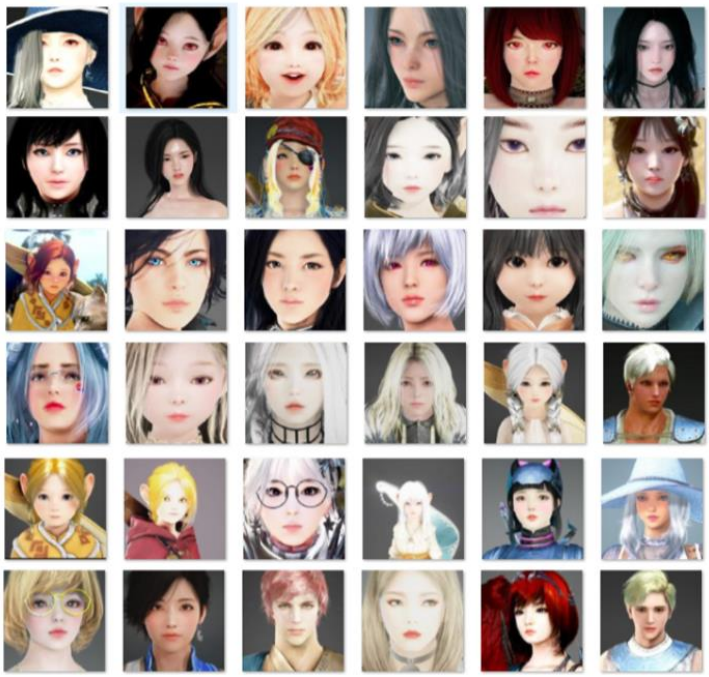

Fig. 2. Images from the Black Desert dataset

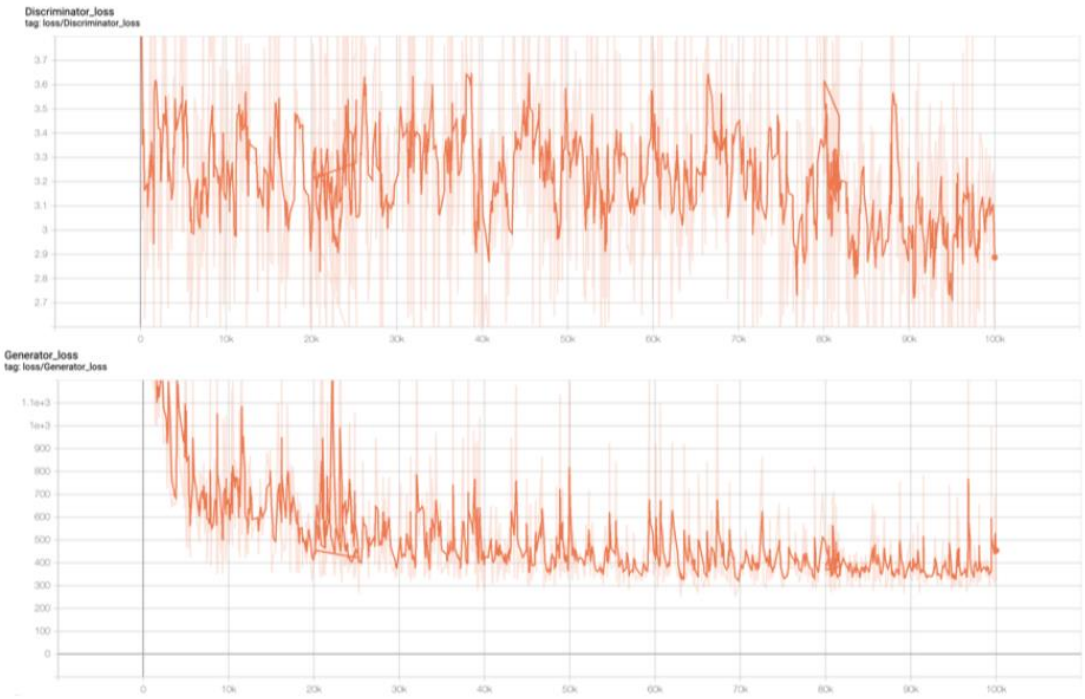

Fig. 3. Loss graph: (top) Discriminator loss (bottom) Generator loss

We first verified the result quality of the image generated by the methodology proposed in this study. We selected CycleGAN [3], UNIT [34], and UGATIT [10] for comparison. The comparison result is shown in **Fig. 4**. As can be seen in the figure, the nose, mouth, and eyes are more detailed, as compared to those in the images generated by other networks. Based on these results, it can be said that the proposed system enables the creation of the shapes of the nose, mouth, and eyes more specifically. CycleGAN and UNIT expressed similar facial shapes overall but could not express specific features. The UGATIT network responded to the angular change but could not accurately generate the resulting images. These results show that the proposed two feature loss enables the creation of a more detailed face as compared to the existing translation network. At the same time, it shows that it is possible to respond to changes at the angular level, to a certain extent, without a separate normalization process.

Fig. 5 shows the multiple generation results. The results show that our network is robust in terms of face scaling and rotation. Failure to attach special accessories to the face indicates that elements such as eyes, nose and mouth are relatively stable. This shows that the two loss values suggested in this study can function normally. The shader in the target black desert renders the overall skin tone lighter; therefore, the skin color changes to a lighter hue overall. The proposed network could not effectively restore the image with accessories such as hats and glasses. This is because they are often not included in the accessory or game customization datasets commonly used in the real world. In the proposed network, gender is usually not displayed. This is because there is an increased tendency for the network learned through it to convert to males and females as 90% of the collected Black Desert game datasets consist of data of females.

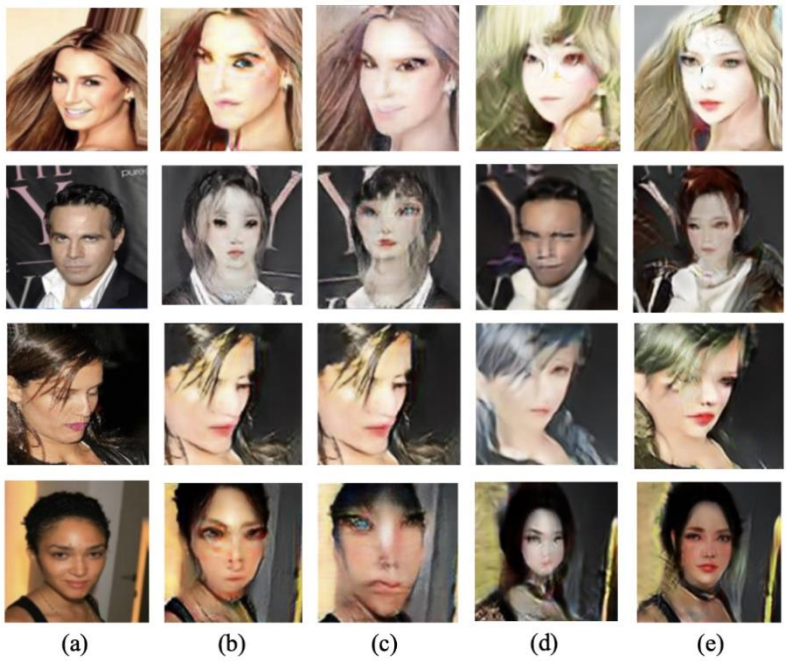

Fig. 4. Comparison of resulting images: (a) source images; images generated by (b) CycleGAN, (c) UNIT, (d) UGATIT, (e) the proposed method

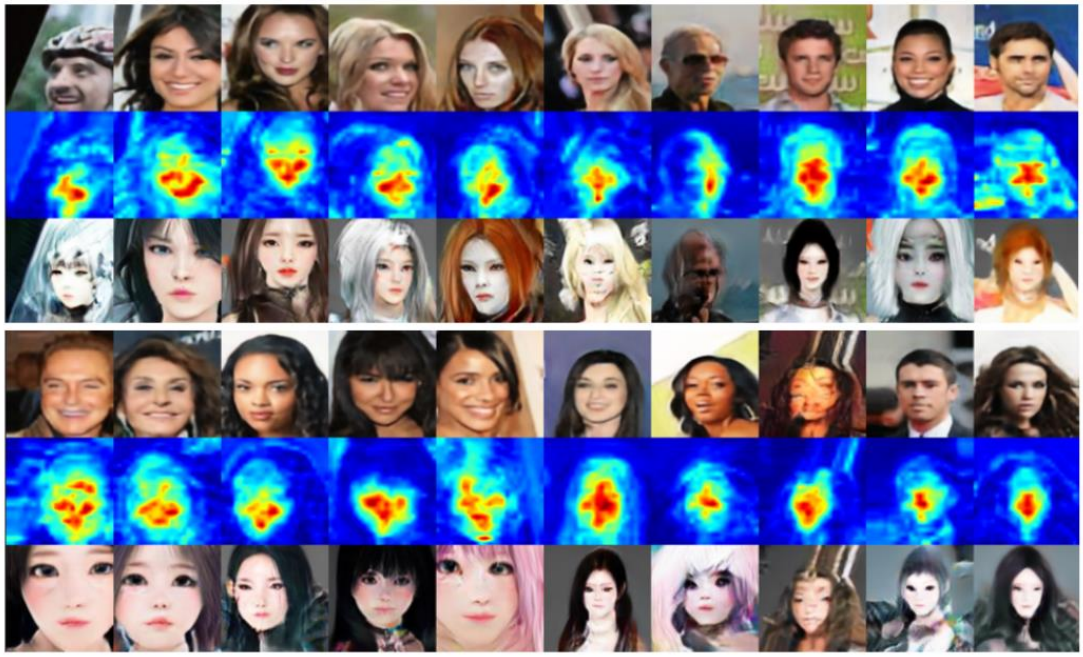

Fig. 5. Generated images from the Black Desert dataset

The performance of our network was also evaluated on another dataset (Anime-Face-Dataset) [35]. This dataset requires more radical changes than the Black Desert dataset. It is also abstracted at a high level, which makes it highly difficult to segment. **Fig. 6** shows the image generation results. The generated images show that the image-to-image translation was also performed normally. The Manga dataset shows the manner in which the facial expressions are reflected. In particular, the expression of the mouth and eyes was prominent. Compared to the Black Desert dataset results, the skin color was reflected normally, and the image was generated. This is because the Manga dataset has a relatively rich skin set. The rotation of the face was normally expressed, but the scale value was not accurately reflected. The manga dataset has a very small number of accessories attached to the face, which is why the accessories were not visible on the generated face.

Fig. 6. Generated images from the Manga dataset

We also evaluated the influence of the two loss effects. **Fig. 7** shows the results of applying each loss to the Black Desert dataset using the Class Activation Map (CAM) [36]. The leftmost rows are the results of applying the proposed basic three losses (L_{lsgan} + L_{cycle} + $L_{identity}$). Although approximate face shapes could be created, the facial details are not well expressed. The middle column presents the results of applying feature loss to the dataset. As can be seen, the details of the face are expressed with a great amount of detail. The rightmost column illustrates the result of applying segmentation loss to the dataset. As can be seen, the expressions of the nose, mouth and eyes are accurately generated in the images, and detailed expressions are achieved.

Fig. 8 shows the resulting image when the ratio of the two loss values was exchanged with that of the Manga dataset. As can be observed, the greater the amount of segmentation loss that was reflected, the more effectively the nose and mouth were learned; this led to an increase in the feature loss value that was reflected, and as a result, the overall face area was learned more effectively. However, the eyes, which form the core of the manga face, could not be learned when only the segmentation loss was applied. Through this, the two loss values had complementary properties. The more realistic the image was, the more useful was the segmentation loss, and the more abstract was the image, the more effective was the feature loss in understanding facial features.

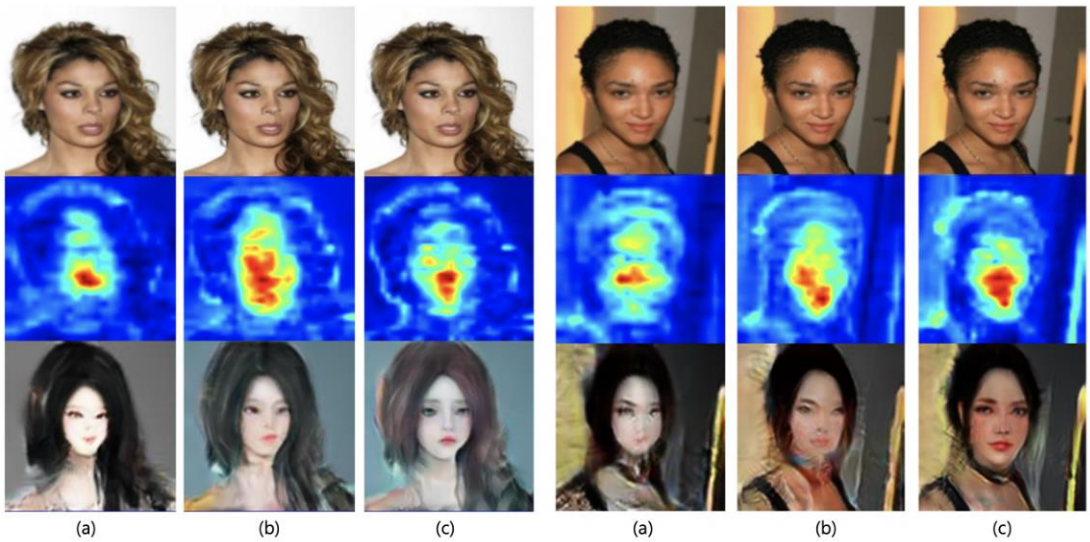

Fig. 7. Comparison of losses: (a) three basic losses (L_{lsgan} + L_{cycle} + $L_{identity}$) (b) basic losses with the feature loss $L_{feature}$, (c) basic losses with the feature loss $L_{feature}$ and segmentation loss L_{seg}

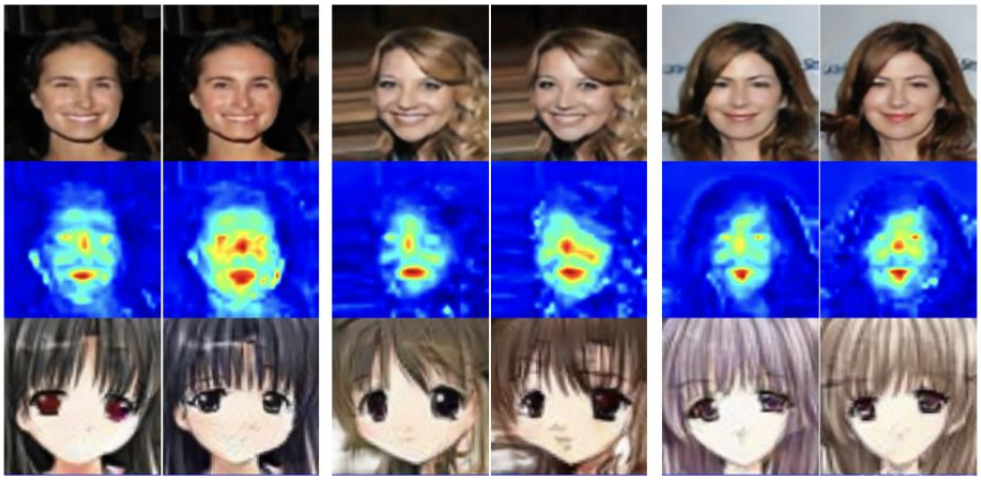

Fig. 8. Comparison of losses: (a) high segmentation loss weight (b) high feature loss weight

We calculated the Frechet inception distance (FID) [37] to quantitatively evaluate the quality of images generated by the GAN network. The average FID score calculated from the data of the Black Desert test set was 44.14, and the average FID score of the Manga Data Test set was 44.07. For each test image, we chose the image as a reference from the generator training set and the average FID calculated over the entire test set. The images created using our proposed method were 11.3% lower than the average FID of the UGATIT network. These results show that our network can create images that are more diverse and similar to the ground truth images.

Therefore, the effects of feature loss and segmentation loss on the proposed model were analyzed. A user survey was conducted to verify the visual quality of the translated images. A total of 50 participants were recruited from the Hongik University. They were requested to observe 100 images created using each model and then select the most realistic images. The survey results are presented in **Table 1** of the survey participants, 75.18% and 71.72% rated the images from the proposed method as the most realistic. The result for the Black Desert dataset was better than that for the Manga dataset. This is because segmentation loss had a positive effect when applied to the Black Desert dataset, thereby resulting in more realistic images.

Table 1. Subjective evaluation results of the various models

Models	**Selection Ratio *(Black Desert)***	**Selection Ratio *(Manga)***
CycleGAN	1.24%	2.44%
UNIT	3.74%	4.42%
UGATIT	19.84%	21.42%
Proposed	75.18%	71.72%

5. Discussion

This study proposed the use of two face-specialized losses—segmentation loss L_{seg} and feature loss $L_{feature}$—for training. Through this approach, image translation was attempted to minimize the two corresponding losses. Because the aforementioned face-specialized losses use two pre-trained networks trained using realistic face images, the quality of transition was better as the

target image was closer to the realistic image. If the shape of the target domain was very abstract, the two additional loss values were not calculated normally. To address this issue, a pre-trained network was developed to acquire segmentation maps and feature vectors by performing data labeling for target domain data separately.

6. Conclusions

In this paper, we introduced an image-to-image translation method for faces in games and Manga content. We proposed two feature losses specialized in facial transformation and confirmed their effectiveness. As a result, we demonstrated that stable translation from the source image to the target image is possible even with a comparative study method. This method is advantageous because it does not require data augmentation. The proposed image-to-image translation method can improve the quality of the result in proportion to the amount of image data; however, if additional attribute labeling is added, a more detailed translation becomes possible. In future research, more precise face translation will be attempted by the application of an automatically labelable network.

References

[1] Snapchat. [Online]. Available: https://play.google.com/store/apps/details?id=com.snapchat.android&hl=ko

[2] Timestamp Camera. [Online]. Available: https://play.google.com/store/apps/details?id=com.artifyapp.timestamp&hl=en_US

[3] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. of the IEEE international conference on computer vision*, pp. 2242-2251, 2017. Article (CorssRef Link)

[4] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. Courville, "Augmented cyclegan: Learning many-to-many mappings from unpaired data," *arXiv preprint arXiv:1802.10151*, 2018. Article (CorssRef Link)

[5] Y. Lu, Y. W. Tai, and C. K. Tang, "Conditional cyclegan for attribute guided face image generation," *arXiv preprint arXiv:1705.09966*, 2018. Article (CorssRef Link)

[6] S. Hong, S. Kim, and S. Kang, "Game sprite generator using a multi discriminator GAN," *KSII Transactions on Internet & Information Systems*, vol. 13, no. 8, 2019.

[7] H. Su, J. Niu, X. Liu, Q. Li, J. Cui, and J. Wan, "Unpaired photo-to-manga translation based on the methodology of manga drawing," *arXiv preprint arXiv:2004.10634*, 2020.

[8] C. Furusawa, K. Hiroshiba, K. Ogaki, and Y. Odagiri, "Comicolorization: semi-automatic manga colorization," *SIGGRAPH Asia 2017 Technical Briefs*, vol. 12, pp. 1-4, 2017. Article (CorssRef Link)

[9] T. Shi, Y. Yuan, C. Fan, Z. Zou, Z. Shi, and Y. Liu, "Face-to-parameter translation for game character auto-creation," in *Proc. of the IEEE International Conference on Computer Vision*, pp. 161-170, 2019. Article (CorssRef Link)

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

[11] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414-2423, 2016. Article (CorssRef Link)

[12] E. Risser, P. Wilmot, and C. Barnes, "Stable and controllable neural texture synthesis and style transfer using histogram losses," *arXiv preprint arXiv:1701.08893*, 2017.

[13] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 6997-7005, 2017. Article (CorssRef Link)
[14] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Stylebank: An explicit representation for neural image style transfer," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2770-2779, 2017. Article (CorssRef Link)
[15] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *arXiv preprint arXiv:1610.07629*, 2016.
[16] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. of the IEEE International Conference on Computer Vision,* pp. 1510-1519, 2017. Article (CorssRef Link)
[17] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. of European Conference on Computer Vision*, vol. 9906, pp. 694-711, 2016. Article (CorssRef Link)
[18] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M. H. Yang, "Universal style transfer via feature transforms," *Advances in Neural Information Processing Systems*, pp. 386-396, 2017.
[19] M. Lu, H. Zhao, A. Yao, F. Xu, Y. Chen, and L. Zhang, "Decoder network over lightweight reconstructed feature for fast semantic style transfer," in *Proc. of the IEEE International Conference on Computer Vision,* pp. 2488-2496, 2017. Article (CorssRef Link)
[20] A. Selim, M. Elgharib, and L. Doyle, "Painting style transfer for head portraits using convolutional neural networks," *ACM Transactions on Graphics (ToG)*, vol. 35, no. 4, pp. 1-18, 2016. Article (CorssRef Link)
[21] O. Sendik and D. Cohen-Or, "Deep correlations for texture synthesis," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 5, pp. 1-15, 2017. Article (CorssRef Link)
[22] K. Cao, J. Liao, and L. Yuan, "Carigans: Unpaired photo-to-caricature translation," *arXiv preprint arXiv:1811.00222,* 2018.
[23] M. Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, "Few-shot unsupervised image-to-image translation," in *Proc. of the IEEE International Conference on Computer Vision*, pp. 10551-10560, 2019.
[24] B. AlBahar and J. B. Huang, "Guided image-to-image translation with bi-directional feature transformation," in *Proc. of the IEEE International Conference on Computer Vision*, pp. 9016-9025, 2019.
[25] T. Park, A. A. Efros, R. Zhang, and J. Y. Zhu, "Contrastive learning for unpaired image-to-image translation," *arXiv preprint arXiv:2007.15651,* 2020. Article (CorssRef Link)
[26] A. Royer, K. Bousmalis, S. Gouws, F. Bertsch, I. Mosseri, F. Cole, and K. Murphy, "XGAN: Unsupervised image-to-image translation for many-to-many mappings," *Domain Adaptation for Visual Understanding*, pp. 33-49, 2020.
[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. of the IEEE International Conference on Computer Vision*, pp. 618-626, 2017. Article (CorssRef Link)
[28] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. of European Conference on Computer Vision*, pp. 850-865, 2016. Article (CorssRef Link)
[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[30] C. H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," *arXiv preprint arXiv:1907.11922*, 2019.
[31] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884-2896, 2018. Article (CorssRef Link)
[32] Black Desert and Perl Abyss. [Online]. Available: https://www.blackdesertonline.com/midseason
[33] Dlib. [Online]. Available: http://dlib.net/

[34] M. Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in Neural Information Processing Systems*, pp. 700-708, 2017.
[35] Anime-Face-Dataset. [Online]. Available: https://github.com/Mckinsey666/Anime-Face-Dataset
[36] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921-2929, 2016. Article (CorssRef Link)
[37] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," *Advances in Neural Information Processing Systems*, pp. 6626-6637, 2017.

Shinjin Kang received an MS degree at Korea University in 2003. After graduation, he worked at the Sony Computer Entertainment Korea (SCEK) and the NCsoft Korea. He received a PhD degree in Computer Science and Engineering at Korea University in 2011. He is currently a professor at the school of games in Hongik University.