Copyright ⓒ 2021 KSII

Eyeglass Remover Network based on a Synthetic Image Dataset

Shinjin Kang[1]* and Teasung Hahn[2]
[1] School of Games, Hongik University
2639 Sejong-ro, Jochiwon, Sejong, Korea
[e-mail: directx@hongik.ac.kr]
[2] NCsoft
12, Daewangpangyo-ro 644ben-gil, Bundang-gu, Seongnam-si, Gyeonggi-do, Korea
[e-mail: spinel@ncsoft.com]
*Corresponding author: Shinjin Kang

Received February 17, 2021; revised March 4, 2021; accepted March 12, 2021; published April 30, 2021

Abstract

The removal of accessories from the face is one of the essential pre-processing stages in the field of face recognition. However, despite its importance, a robust solution has not yet been provided. This paper proposes a network and dataset construction methodology to remove only the glasses from facial images effectively. To obtain an image with the glasses removed from an image with glasses by the supervised learning method, a network that converts them and a set of paired data for training is required. To this end, we created a large number of synthetic images of glasses being worn using facial attribute transformation networks. We adopted the conditional GAN (cGAN) frameworks for training. The trained network converts the in-the-wild face image with glasses into an image without glasses and operates stably even in situations wherein the faces are of diverse races and ages and having different styles of glasses.

***Keywords*:** Image-to-image Translation, Generative Adversarial Network, Eyeglass Remover

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 and Hongik University Research Fund.

http://doi.org/10.3837/tiis.2021.04.016 **ISSN : 1976-7277**

1. Introduction

Various generative adversarial network (GAN)-based networks have been proposed that convert in-the-wild face images for arbitrary purposes. These studies have been applied to perform diverse functions such as virtual makeup [1] and age/race-estimation [2, 3]. They have also been used in practical applications [4]. These studies have developed rapidly because the scale and quality of labeled datasets in the field of facial-attribute transformation have expanded greatly [5, 6]. The datasets are combined with various image-to-image translation technologies, and many facial feature transformation networks have emerged [7-9].

To apply a convolutional neural network (CNN)-based facial feature transformation network effectively to an in-the-wild image, the image must be pre-processed for learning purposes. In particular, accessories such as glasses, hats, and masks on or near the face interfere with the detection of facial feature regions such as the eyes, head, and mouth, which are the main facial features, thereby hindering the efficiency of CNN-based network learning. Therefore, face transition network researchers often regard images with these accessories as noise data and remove them from the dataset or perform related pre-processing manually. Although techniques such as face frontalization [10] have been proposed, there are still many difficulties in completely removing accessories such as glasses. These pre-processing costs have been an obstacle to improving network performance in the field of face transition research.

In this paper, we propose a network that can effectively remove only the glasses from among various accessories attached to the face. Glasses are the most frequently encountered among facial attachment accessories, but their effect may differs depending on race and gender; hence, it is surprisingly difficult to remove them using traditional machine learning methodologies. Our methodology has the advantage of improving the stability and accuracy of network learning by using the supervised learning technique; by using automated synthetic images to build the dataset, it incurs no data labeling cost. In addition, our methodology is easy to use in other applications because it uses a relatively easy-to-use dataset and pre-trained facial feature network.

There are few instances of studies that specialize in the field of eyeglass removal. Liang [11] used deep CNNs and Hu [12] used the unsupervised method; however, there are cases where the test set size is not large, or the resulting image quality is unstable. Our proposed method can be used in a pre-processing network for various facial attribute conversion techniques by recommending a data construction and network specialized for removing glasses.

2. Related Work

2.1 Facial Attribute Editing Networks

Facial attribute editing is a research field that creates various transformed images by changing the facial attributes within a facial image. Initial research on facial attribute editing was mainly based on autoencoder-based approaches such as Variational AutoEncoder (VAE)/GAN [13], IcGAN [14], and Fader-Network [15]. Recently, various GAN-based approaches have become mainstream research.

Choi et al. proposed StarGAN [16], a scalable approach that can perform image-to-image translations for multiple domains using only a single model. As StarGAN uses one neural network to convert images from one domain into many domains, it learns general knowledge

and creates images of higher quality. In addition, it is a very economical model if computing power is considered as a cost. He et al. proposed AttGAN [17], which improved the attribute editing performance by employing only the necessary adversarial loss, attribute classification loss, and reconstruction loss. They applied an attribute classification constraint to the generated image to guarantee the correct changes in the desired attributes. Liu et al. proposed STGAN [18], which attempts to address blurred image issues from a selective transfer perspective. Their model selectively takes the difference between the target and source attribute vectors as inputs. Selective transfer units are incorporated adaptively with an encoder--decoder to select and modify the encoder feature for enhanced attribute editing. Zhang et al. proposed that a spatial attention mechanism be introduced to the GAN framework (SaGAN) [19] to alter the attribute-specific region only and retain the rest unchanged. Their generator contains an attribute manipulation network (AMN) to edit the face image and a spatial attention network (SAN) to localize the attribute-specific region that restricts the alteration of the AMN within this region. Zhou1 et al. proposed GeneGAN [20], which can learn disentangled attribute subspaces from weakly labeled data by adversarial training. Their model can learn object transfiguration from two unpaired sets of images: one set containing images that "have" that kind of object and the other being the opposite, with the mild constraint that the objects are located approximately at the same place. Xiao et al. proposed DNAGAN [21], which attempts to disentangle different factors or attributes of images. The latent representations of images are DNA-like, in which each piece represents an independent factor of the variation. They also proposed ELEGANT [22], which receives two images with opposite attributes as inputs. Their model can transfer the same type of attributes from one image to another by exchanging certain parts of their encodings. Kim et al. proposed an unsupervised method of learning to transfer visual attributes [23]. They swapped attributes between two faces by exchanging attribute-relevant latent codes. Their method can learn the transfer function without any corresponding images. Yin et al. proposed GeoGAN [24] leveraged facial landmarks as geometric guidance to learn differentiable flows automatically, despite the existence of a large pose gap. They used geometry-aware flow, which serves as a well-suited representation for modeling the transformation between instance-level facial attributes. Chen et al. [25] proposed an end-to-end convolutional neural network that supports fast inference, edit-effect control, and quick partial model updates. They used Facelet-Bank [26] for each attribute to infer the feature deviation for attribute generation.

2.2 Eyeglasses Removal Networks

The removal of accessories from the face is a technology required during pre-processing in the field of face recognition. Because accessories that are attached to faces interfere with machine learning, pre-processing is applied to standardize the face with face frontalization technology. Early research in this field mainly used statistical learning techniques. Wu et al. [27] proposed a finding-and-replacing approach for removing glasses from a frontal face image. This method first finds the position of the glasses with an eye area detector and then replaces them with a composite glasses-free image. Park et al. [28] applied the recursive process of principal component analysis (PCA) reconstruction and error correction to create a face image without glasses. Because the temperatures of the glasses and the human face are different there is also an operation to remove the glasses using thermal imaging. Wong et al. proposed a nonlinear glass removal algorithm for thermal images based on kernel PCA [29]. This method performs PCA to transfer the visible reconstruction information from the visible feature space to the thermal feature space and then applies image reconstruction to remove the glasses from the thermal facial image. Unlike the aforementioned PCA-based methods, some researchers rely

on sparse coding and expectation-maximization to reconstruct the face. De Smet et al. [30] proposed a generalized expectation-maximization algorithm in which the estimation of the morphable model-related parameters is interleaved with visibility computations. Their method iteratively estimates the parameters of a three-dimensional (3D) morphable face model to approximate the appearance of a face in an image. They also suggested a visibility map that segments the image into visible and occluded regions. However, these PCA-based studies only used the front-of-face image and used the images taken in a controlled environment as learning data; hence, there were limitations in applying them to in-the-wild images including various lighting, face angles, and poses. To overcome these limitations, accessory removal technologies that can cope with in-the-wild images are being introduced. Hu et al. [31] proposed a unified eyeglass removal model called eyeglasses removal generative adversarial network (ERGAN). The proposed model learned to swap the eye area on two faces. The generation mechanism focuses on the eye area and evades the difficulty of generating a new face. Their method does not depend on the dense annotation of the eyeglasses' location, but benefits from large-scale face images with weak annotations. Lee et al. [32] proposed ByeGlassesGAN, an image-to-image GAN framework for spectacle removal. Their model consists of an encoder, a face decoder and a segmentation decoder. The segmentation decoder is used to predict the segmentation mask of the glasses. Zhao et al. [33] proposed a spectacle eyeglasses method based on attribute detection and an improved TV restoration model. Their method consists of several steps consisting of eyeglass position, eyeglass frame determination, color information determination, reflective area detection, eyeglass template extraction and eyeglass removal. Din et al [34] proposed a user-friendly method for face de-occlusion in facial images where the user has control of which object to remove. Their method could remove five commonly occurring occluding objects including hands, a medical mask, microphone, sunglasses, and eyeglasses.

Although promising results have been achieved, state-of-the-art methods still suffer from inaccurate, blurry, or incomplete images. To compensate for these shortcomings, we propose a stable supervised learning technique using synthetic images. Our methodology prevents overfitting for a specific network by mixing images generated from various networks and attempts to increase the quality of the resulting image by using a simple but reliable supervised learning technique.

3. Dataset for Training

3.1 Synthetic Glasses Image from Facial Attribute Editing Networks

The more detailed the attribute labeling is, the more precise the facial attribute editing can be in the facial attribute editing network. However, the cost of such labeling is very high and it is difficult for individual researchers to construct such a dataset. After acquisition of a large number of images with glasses being worn on the web using an image crawler, the images can be learned with the unsupervised image-to-image translation technique; however unstable image generation may occur owing to the limitations of unsupervised learning. For stable supervised learning, many image datasets that are paired and do not require additional attributed labeling are needed. For this purpose, we produced a large number of paired glasses attached to images using attGAN [17] and StarGAN [16]. To apply attGAN and StarGAN, an attributed dataset is required in advance. We created composite images with glasses using CelebA and Celeb-HQ images and their attribute data. For this experiment, we created 100,000 images with glasses using the StarGAN and attGAN. The ratio of StarGAN and AttGAN was

set to 3:7. 90% consisted of images with glasses -> images without glasses, and 10% consisted of images with glasses and images without glasses. This is to enable a response when an image without glasses enters the network. If an image without glasses was not learned, the entangled phenomenon of intentionally creating glasses or deforming other images of the face could occur when an image without glasses was input. **Fig. 1** shows the examples of generated synthetic images.

Fig. 1. Generated synthetic paired images with StarGAN [16] and attGAN [17]

3.2 Synthetic Glasses Image from Facial Attribute Editing Networks

If we can create a synthetic image from the attribute editing network, we can obtain a large number of images containing the desired attributes. However, if the network is trained with only a specific dataset, there is a possibility of overfitting to the dataset used for training. Therefore, it is necessary to increase the responsiveness of the network by using diverse datasets.

For this, we created a large number of composite glasses-wearing images using dlib [35]. When dlib receives a real image as an input value, it detects 68 facial landmark points. We can synthesize the glasses image to generate an image wearing glasses by using the landmark points. When the images generated in this way are frontal images, a relatively high-resolution image can be obtained. It is helpful to increase the distribution of learning data by generating a large number of similar glasses-wearing images. For the frontal face image dataset, we used the Pantone dataset [36] and other images from the web by using web crawler. We made 20 types of glasses images and overlaid them on frontal face images at random. In this way, 10,000 256 × 256 frontal face images were constructed with data augmentation and added to the training data. **Fig. 2** shows the examples of generated synthetic images.

Fig. 2. Synthetic images with Pantone dataset

3.3 Pseudo Labeling with Race/Gender/Age Estimation

As a result of observing the datasets used in this study, we found that there were significant differences in preferred glasses and attachment methods by race/gender/age. The type of glasses differed based on culture, the purpose of use (sunglasses, myopia, farsightedness, etc.) and the reasons for wearing them (driving a vehicle, reading, etc.). In particular, the CelebA and CelebHQ datasets, which are mainly used in the study, are composed of celebrities; hence, the ratio of sunglasses being worn is relatively high. When this difference in the images of wearing glasses by race/gender/age is reflected in the network, it can extract accurate feature information for removing the glasses. For this, if additional attachment information is input to the network, the learning efficiency can be improved. We also wanted to automate this labeling task. We used the FairFace [3] network to automatically extract additional eyeglass wearer information and input it into the network. FairFace extracts race/gender/age estimates as int values. We normalized these values and applied weights as the network input values. **Fig. 3** shows the examples of auto labeled images with FairFace.

Fig. 3. Auto labeled images with FairFace

4. Method

4.1 Generator

We used the cGAN network structures for supervised learning. By simply adding race/gender/age data, cGAN can provide generators and discriminators a condition that can assist learning. Using this model, images can be created according to the race/gender/age condition. The generator, encoder, residual block, auxiliary classifier, and decoder structures proposed in UGATIT [37] were used. Because the UGATIT network includes an auxiliary classifier, it has an advantage in learning the key features; it also includes an adaptive normalization function, so that the image-to-image translation is relatively stable. The proposed network structure is illustrated in **Fig. 4**. The data were input to the network after pairing images of faces wearing glasses and images of faces not wearing glasses. The two are entered into the FairFace networks and the estimated race/gender/age is used as the conditional value of cGAN. The three estimated conditions and images were concatenated and input to the generator and discriminator. The generator is trained to generate an image without glasses as realistically as possible to trick the discriminator.

Let $x_s \in X_s$ and $x_t \in X_t$ represent samples from the source and target domains. Furthermore, let $G_1(x_s)$ represents the translated source and target domains, respectively. Our model consists of one generators $G_1(x_s)$, one discriminators $D_1(G_1(x_s))$, and one feature extractor F_1. $G_1(x_s)$ creates an image that fits the target style based on the GAN framework.

The discriminators D_1 distinguish between real and fake translated images. The feature extractor F_1 provides the conditions under which the cGAN framework facilitates image transformation. The final loss function of our model can be written as the loss of L_{total}.

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

L_{total} comprises five loss terms: L_{lsgan}, $L_{identity}$, L_{pixel}, and L_{cam}. The adversarial loss L_{lsgan} is employed to match the distribution of the translated images to the target image distribution. The identity loss $L_{identity}$ is used to ensure that the color distributions of the input and output images are similar. L_{pixel} is used to check the L_1 pixel differences between the translated and target images. These losses are calculated using G and D with the cGAN frameworks. These terms are described in detail in [37]. L_{cam} uses information from auxiliary classifiers to determine the differences between the two domains [38].

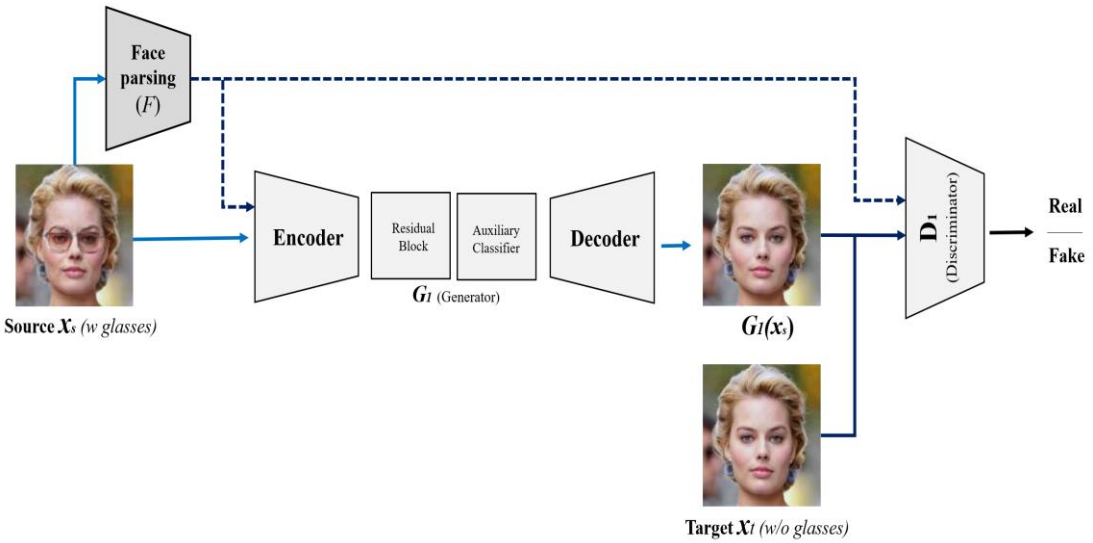


Fig. 4. Proposed network architecture

$$L_{total} = L_{lsgan}(G_1, D_1) + L_{identity}(G1, D_1) + L_{cam}(G_1, D_1) + L_{pixel}(G1, D_1) \tag{2}$$

4.2 Face Parsing Network

Our purpose is to convert a face image with glasses into one without glasses. Existing studies have attempted to remove the glasses from cropped image data by cropping only the eye position within the face. However, in this case, only partial image conversion occurs without having global information of the face. At this time, a side effect occurs where the partial images around the converted eyes are to inferior in continuity at the interface to the images of other areas in the face. To solve this problem, we attempted to reflect the features of the entire face in image generation rather than partially interpreting and transforming only the eye area to create a stable GAN image. In particular, these attempts enable the GAN framework, which generates unstable images, to respond to in-the-wild images taken from different angles and

lighting. We used the Fairface [3] network to input additional race/age/gender information into the network condition. This is based on the assumption that the style of wearing glasses varies according to race/age/gender. The results of our experiment confirmed that the networks learned in CelebA and CelebA-HQ could not respond normally to other face images crawled on the public web. This is to compensate for the disadvantage that data balancing is not complete because CelebA and CelebA-HQ mainly collect data of young celebrities. We divided the dataset into six race categories (white, black, Hispanic, East Asian, Southeast Asian, Indian, and Middle Eastern), seven age levels, and two gender categories. We then used one-hot encoding to input them to the network.

4.3 Discriminator

In this study, D has a structure similar to that of G. 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.

5. Experimental Results

For experiment, 100,000 source images were randomly selected from the synthetic glass wearing images. All images were resized to 256 × 256 pixels for training. For optimization, we set the maximum number of iterations to 1,000, the learning rate as 0.0002, and the decay rate as 20% per five iterations. The Adam optimizer, with a batch size of two, was used for training. The Intel i7-10700K CPU, 64M RAM, and NVIDIA RTX Titan GPU required approximately seven days for training. Pytorch was used as the deep learning library. On the loss graph, it stably converged from about 400 epochs, and then fine-tuned. The train:val:test dataset was set at a 9:1:1 ratio. The test data set consists of a synthetic image and an image searched as a face image on the Web.

First, we evaluated the quality of the proposed network. **Fig. 5** shows the multiple generation results. We can see that the proposed network produces an overall stable image with the eyeglasses removed. Glasses were removed at a rate of approximately 98% when the images with glasses were placed in the test set. In our network, because the eyes were visible behind commonplace vision correction glasses, these were stably removed. The cases where the glasses were not removed well were those with sunglasses. Because they cover the eye, it is difficult for the network to infer the eye in images with sunglasses. In particular, when the eyebrows are covered, the network is limited in inferring glasses and eyebrows simultaneously. The quality of the generated image shows that the generator can stably generate 256 × 256 resolution images. As a result of the experiment, Resent's skip-connection structure showed a higher level of quality compared to U-Net structures such as pix2pix.

It is noteworthy that the resulting image has clean skin than the ground truth image. These results indicate that the network interprets glasses as noise on the skin and learns to remove noise from the entire facial area. This is because the mustache and beard were partially included in the images with glasses used when creating the training data. This phenomenon seems to have occurred because the latent vector of the attGAN / StarGAN network we used was partially entangled. The glasses targeted by our network are very reliably removed, but owing to the limitations of the synthetic dataset, our latent vector is partially entangled with the skin texture.

We compared our method with the recent eyeglasses remover model, ERGAN [31]. Baseline method were implemented using the authors' codes. **Fig. 6** shows the comparison

results. The resulting image confirms that the result of our proposed method is of higher quality than the result generated by ERGAN. As ERGAN undergoes unsupervised training, labeling is unnecessary, thus it learns easily, but has the disadvantage of the resulting images' resolution not being high. An additional disadvantage is that the color continuity with other pixels in the face image may deteriorate. As our result is for not only the area of the glasses, but also when the whole face is learned, such regional discrepancies do not occur. As the ERGAN result used the author's pre-trained model, there is a possibility that the best learning was not achieved.

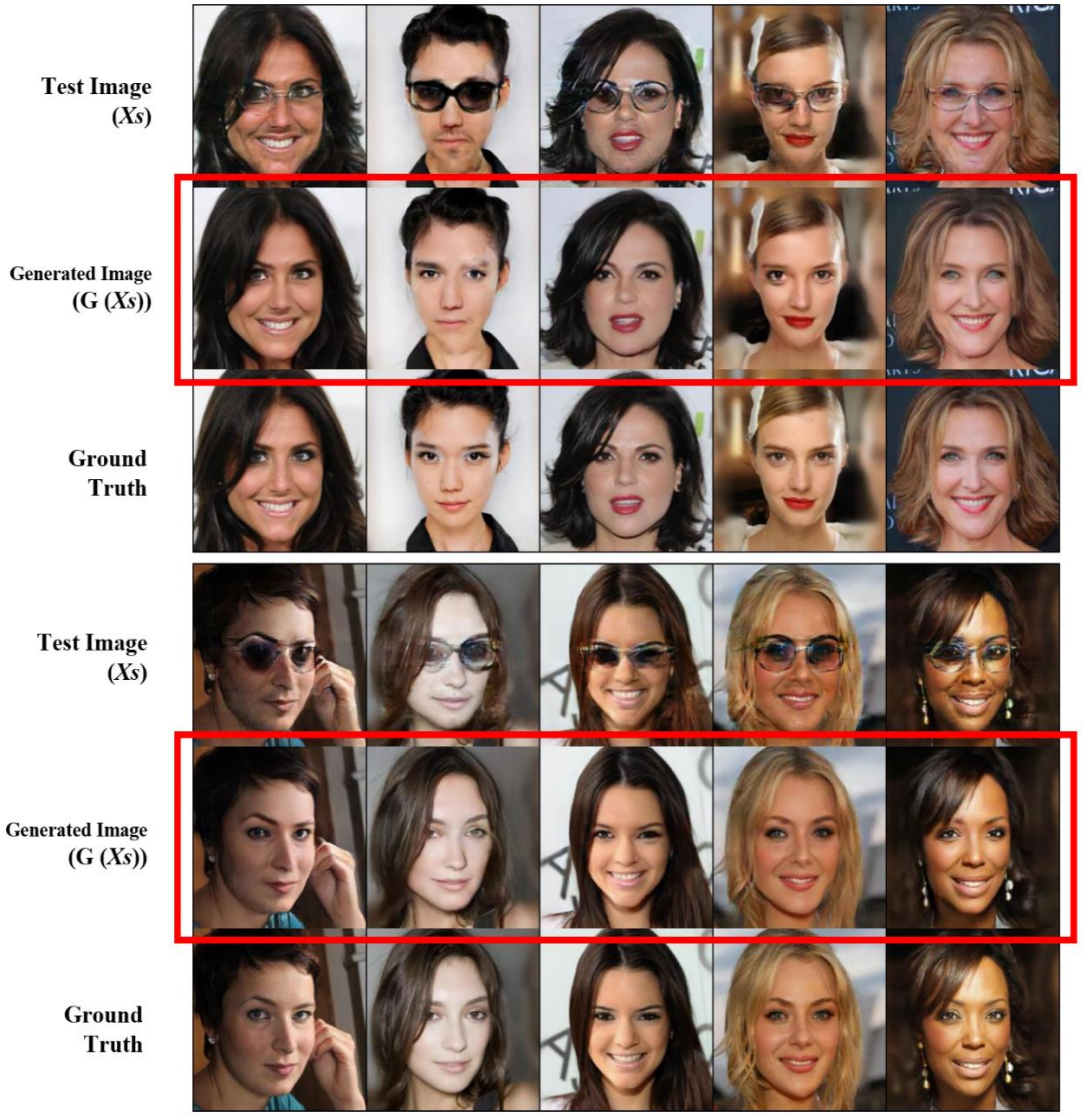


Fig. 5. Generation results

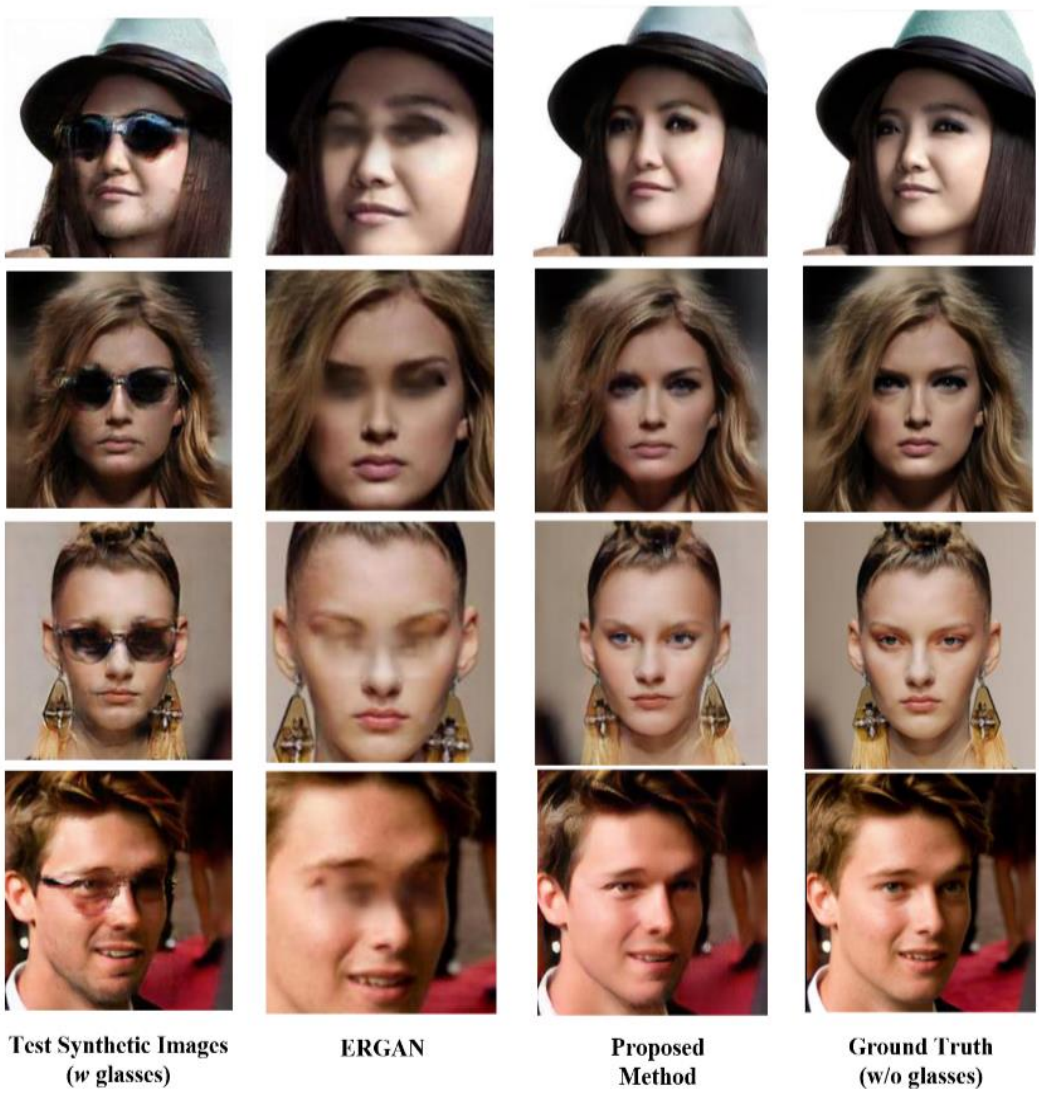


Fig. 6. Comparison between the generated faces of other glasses removal networks

Fig. 7 shows the results when we trained with cGAN using conditions for learning and when we used the general GAN. It shows that the performance of the glasses remover improves more when learning with cGAN than when learning with GAN. As seen from the resulting image, when cGAN is applied, the effect of removing the glasses is shown more clearly. There is a possibility that the dataset sampled from the CelebA and CelebA-HQ datasets may not be correctly balanced by race/age/gender. The condition added to the network is used as a correction value for these parameters. The performance of the generator proposed in this paper provides stable glasses removal. However, under unusual conditions (very old or young, ethnic minority, etc.), this additional information can supplement the stability of the network.

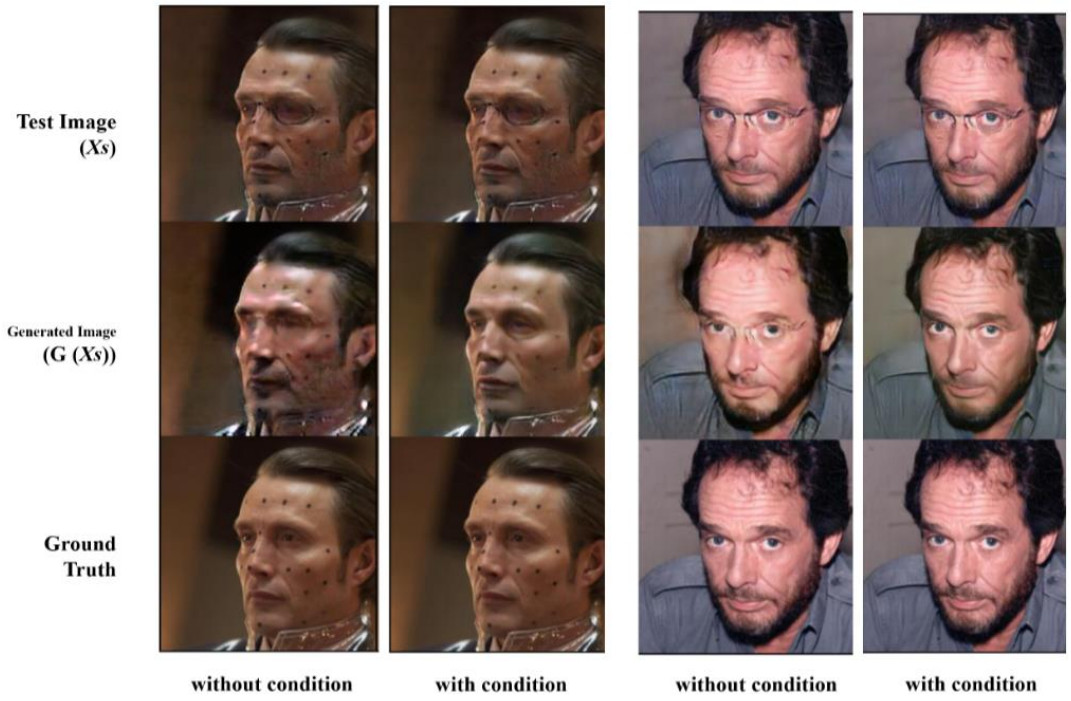


Fig. 7. Generated images with and without cGAN conditions

We determined the effect of the four additional losses based on GAN loss. **Fig. 8** shows the resulting images when applying each loss to the test dataset. The top image shows the test image entered into the trained network. The image at the bottom is the image with the target glasses removed. **Fig. 8(c)** shows the addition of pixel loss to the most basic GAN loss. The network creates an overall shape while removing the glasses to some extent. However, the disadvantage is that the resolution of the resulting image is reduced. **Fig. 8(b)** shows the identity loss, and cam loss. The image resolution was greatly improved, but the glasses were not clearly removed. **Fig. 8(a)** shows the results when all proposed losses are used. When the glasses were well removed, the image resolution was maintained. This experiment shows that pixel loss has the most direct influence on the removal of the glasses. The rest of the losses seem to contribute to improving the realism of the face in the image-to-image translation framework.

Fig. 9 Shows the difference between our supervised methodology and the unsupervised methodology. The original UGATIT network was used as the unsupervised learning network. Similar amounts of image data and training time were used. In most cases, our results reliably produce high-resolution eyeglass removal images. However, less than 40% of the UGATIT network learned by the unsupervised learning method succeeded in removing glasses. There are also sunglass images when glasses are not removed normally. This result confirms that the network trained with the stable supervised learning technique using a synthetic dataset demonstrates better results under the same data conditions.

To quantitatively evaluate the similarity between the generated and target face images, we used the pixel accuracy, mean accuracy, mean IU, and frequency-weighted IU values as our metrics. **Table 1** shows the overall performance of our network compared to UGATIT and CycleGAN with the test dataset. The pixels were considered to be the same if the color difference was within 20%. **Fig. 10** shows the differences in pixels. The proposed algorithm

exhibited an approximately 10% higher performance for all four indicators. Unlike normal objects, face images are very sensitive to changes around the eyes, nose, and mouth; hence, these numerical changes can be perceived more obviously. In the pixel difference image, we can see that UGATIT accurately restores the face shape at high resolution, so the pixel accuracy is higher; CycleGAN has less resilience than UGATIT. The proposed technique accurately restores the face shape and partially removes only the target glasses. Because the images created by GAN generate a different color for the hair from the original image, a considerable amount of pixel loss occurs in this area but is not perceived as much.

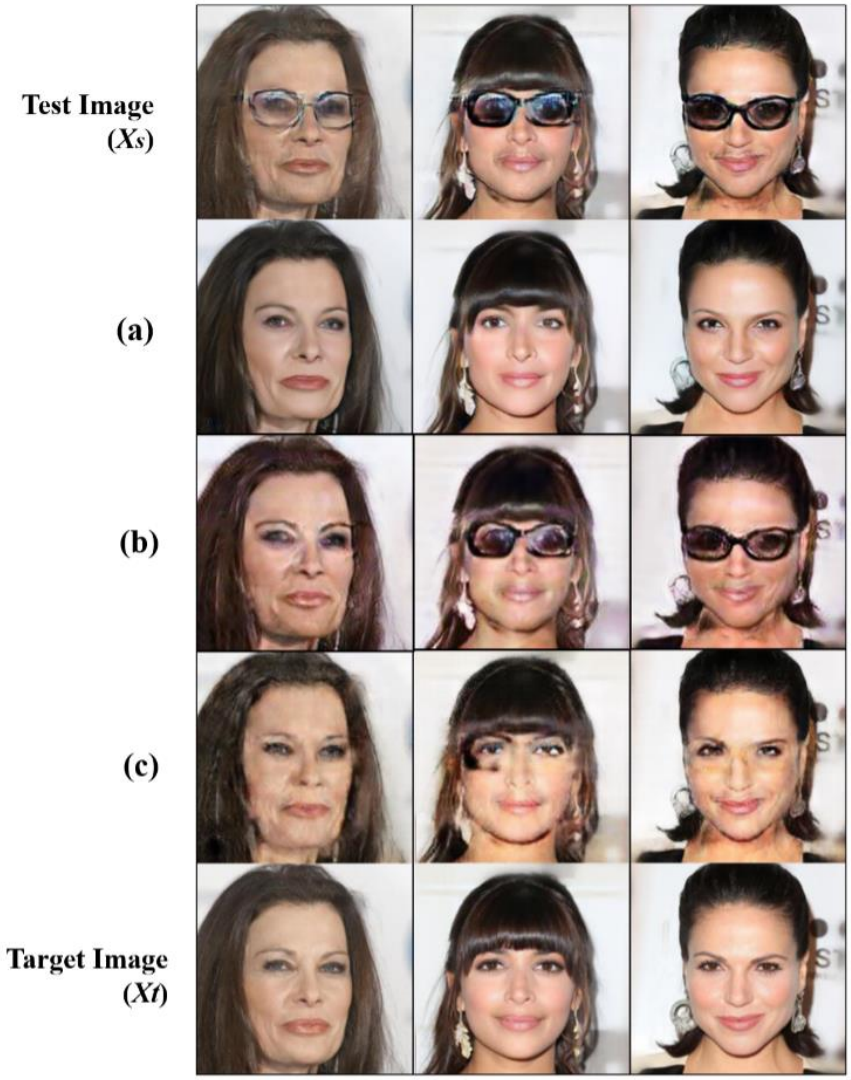


Fig. 8. Loss comparison : (a) identity loss + cam loss + pixel loss, (b) identity loss + cam loss, (c) pixel loss

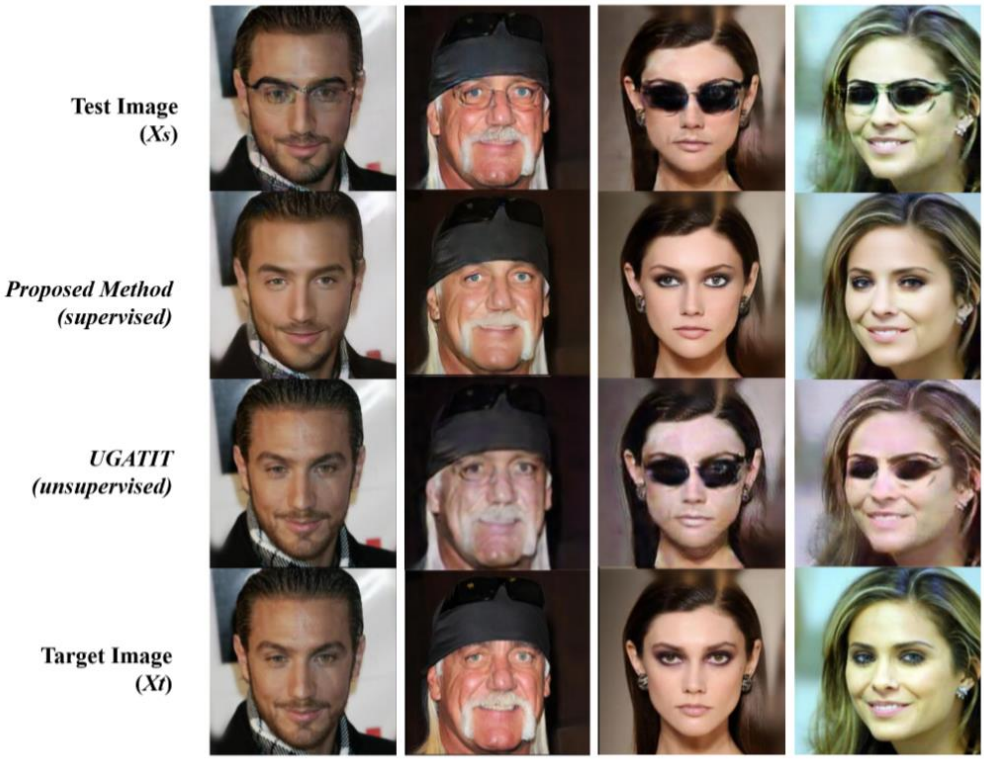


Fig. 9. Comparison with unsupervised learning method

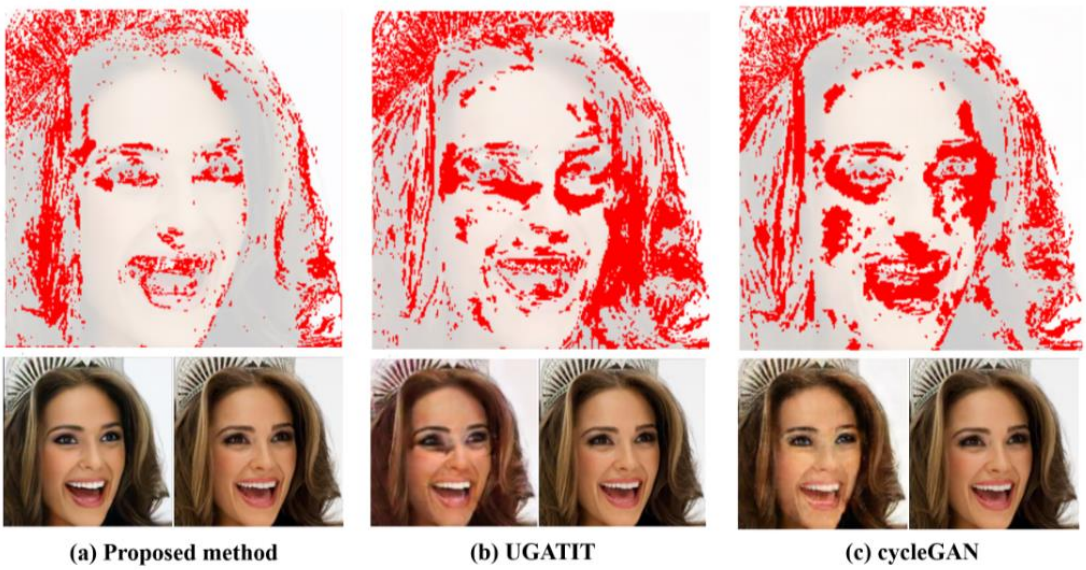


Fig. 10. Comparison with unsupervised learning method

Table 1. Evaluation metric of generated images

Models	Proposed method	UGATIT	CycleGAN
Pixel Accuracy	0.548	0.479	0.466
Mean Accuracy	0.513	0.464	0.410
Mean IU	0.358	0.290	0.275
Frequency Weighted IU	0.387	0.335	0.322

6. Conclusions

In this study, we proposed a supervised image-to-image translation technique using synthetic images to remove glasses from face images. Ours is a supervised learning technique, but it has the advantage that the labeling cost is very low because it uses the pseudo labeling method. The proposed network structure based on cGAN enables superior quality of the resulting image and the removal of glasses during image-to-image translation. Since we use the synthetic data set of our study, there is an advantage that a large amount of data can be easily acquired and the network can be trained. However, there was a limitation in responding perfectly to various exceptional situations occurring in the actual in-wild environment. Since the current network is learned mainly from the front image, it is difficult to cope with the angle change, and there are cases where it is not possible to completely remove it from the face image wearing sunglasses. Our research can be used for pre-processing to remove unnecessary facial accessories through various facial-related deep learning techniques. Because these pre-processing technologies are required in various facial image processing fields, it is expected that there will be numerous future research applications.

References

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

[2] W. Cao, V. Mirjalili, and S. Raschka, "Rank-Consistent Ordinal Regression for Neural Networks," *Pattern Recognition Letters*, pp. 325-331, Nov. 2020. Article (CrossRef Link)

[3] K. Kärkkäinen and J. Joo, "Fairface: Face Attribute Dataset for Balanced Race, Gender, and Age," *ArXiv preprint arXiv:1908.04913*, 2019. Article (CrossRef Link)

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

[5] Z. Liu, P. Luo, X. Wang, and X. Tang, "Large-scale Celebfaces Attributes (celeba) Dataset," Aug. 2018. Article (CrossRef Link)

[6] C. H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards Diverse and Interactive Facial Image Manipulation," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5549-5558, 2020. Article (CrossRef Link)

[7] 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. 2223-2232, 2017. Article (CrossRef Link)

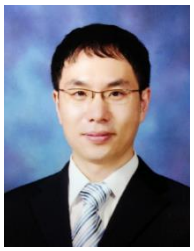
[8] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. Courville, "Augmented Cyclegan: Learning Many-to-many Mappings from Unpaired Data," *ArXiv preprint arXiv:1802.10151*, 2018. Article (CorssRef Link)

[9] Y. Lu, Y. W. Tai, and C. K. Tang, "Conditional Cyclegan for Attribute Guided Face Image generation," *ArXiv preprint arXiv:1705.09966*, 2017. Article (CrossRef Link)

[10] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in Style: a Stylegan Encoder for Image-to-image Translation," *ArXiv preprint arXiv:2008.00951*, 2020. Article (CorssRef link)

[11] M. Liang, Y. Xue, K. Xue, and A. Yang, "Deep Convolution Neural Networks for Automatic Eyeglasses Removal," *DEStech Transactions on Computer Science and Engineering*, 2017. Article (CrossRef Link)
[12] B. Hu, Z. Zheng, P. Liu, W. Yang, and M. Ren, "Unsupervised Eyeglasses Removal in the Wild," *IEEE Transactions on Cybernetics*, 2020. Article (CrossRef Link)
[13] A. B. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding Beyond Pixels Using a Learned Similarity Metric," in *Proc. of International Conference on Machine Learning*, pp. 1558-1566, 2016. Article (CrossRef Link)
[14] G. Perarnau, J. Weijer, J. Raducanu, and J. M. Álvarez, "Invertible Conditional Gans for Image Editing," *ArXiv preprint ArXiv:1611.06355*, 2016. Article (CrossRef Link)
[15] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. A. Ranzato, "Fader Networks: Manipulating Images by Sliding Attributes," in *Proc. of the 31st Conference on Neural Information Processing Systems*, pp. 5967-5976, 2018. Article (CrossRef Link)
[16] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "Stargan: Unified Generative Adversarial Networks for Multi-domain Image-to-image Translation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8789-8797, 2018. Article (CrossRef Link)
[17] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial Attribute Editing by Only Changing What You Want," *IEEE Transactions on Image Processing*, 2019. Article (CrossRef Link)
[18] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, "Stgan: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3668-3677, 2019. Article (CrossRef Link)
[19] G. Zhang, M. Kan, S. Shan, and X. Chen, "Generative Adversarial Network with Spatial Attention for Face Attribute Editing," in *Proc. of European Conference on Computer Vision*, vo1. 11210, pp. 422-437, 2018. Article (CrossRef Link)
[20] S. Zhou, T. Xiao, Y. Yang, D. Feng, Q. He, and W. He, "Genegan: Learning Object Transfiguration and Attribute Subspace from Unpaired Data," in *Proc. of British Machine Vision Conference*, 2017. Article (CrossRef Link)
[21] T. Xiao, J. Hong, and J. Ma, "Dna-gan: Learning Disentangled Representations from Multi-Attribute Images," in *Proc. of International Conference on Learning Representations Workshops*, 2018. Article (CrossRef Link)
[22] T. Xiao, J. Hong, and J. Ma, "Elegant: Exchanging Latent Encodings with Gan for Transferring Multiple Face Attributes," in *Proc. of European Conference on Computer Vision*, 2018. Article (CrossRef Link)
[23] T. Kim, B. Kim, M. Cha, and J. Kim, "Unsupervised Visual Attribute Transfer with Reconfigurable Generative Adversarial Networks," *ArXiv preprint ArXiv:1707.09798*, 2017. Article (CrossRef Link)
[24] W. Yin, Z. Liu, and C. C. Loy, "Instance Level Facial Attributes Transfer with Geometry-aware Flow," in *Proc. of AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, 2019. Article (CrossRef Link)
[25] Y. Chen, X. Shen, Z. Lin, X. Lu, I. Pao, and J. Jia, "Semantic Component Decomposition for Face Attribute Manipulation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9851-9859, 2019. Article (CrossRef Link)
[26] Y. Chen, H. Lin, M. Shu, R. Li, X. Tao, X. Shen, Y. Ye, and J. Jia, "Faceletbank for Fast Portrait Manipulation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3541-3549, 2018. Article (CrossRef Link)
[27] C. Wu, C. Liu, H. Y. Shum, Y. Q. Xy, and Z. Zhang, "Automatic Eyeglasses Removal from Face Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 3, pp. 322-336, 2004. Article (CrossRef Link)
[28] J. S. Park, Y. H. Oh, S. C. Ahn, and S. W. Lee, "Glasses Removal from Facial Image Using Recursive Error Compensation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 805-811, 2005. Article (CrossRef Link)
[29] W. K. Wong and H. Zhao, "Eyeglasses Removal of Thermal Image based on Visible Information," *Information Fusion*, vol. 14, no. 2, pp. 163-176, 2013. Article (CrossRef Link)

[30] M. Smet, M. Fransens, and L. Gool, "A Generalized EM Approach for 3D Model based Face Recognition under Occlusions," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1423-1430, 2006. Article (CrossRef Link)
[31] B. Hu, Z. Zheng, P. Liu, W. Yang, and M. Ren, "Unsupervised Eyeglasses Removal in the Wild," *IEEE Transactions on Cybernetics*, 2020. Article (CrossRef Link)
[32] Y. H. Lee and S. H. Lai, "ByeGlassesGAN: Identity Preserving Eyeglasses Removal for Face Images," in *Proc. of European Conference on Computer Vision*, vol. 12374, pp. 243-258, 2020. Article (CrossRef Link)
[33] M. Zhao, Z. Zhang, X. Zhang, L. Zhang, and B. Li, "Eyeglasses Removal Based on Attributes Detection and Improved TV Restoration Model," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2691-2712, 2021. Article (CrossRef Link)
[34] N. Din, K. Javed, S. Bae, and J. Yi, "Effective Removal of User-Selected Foreground Object from Facial Images Using a Novel GAN-Based Network," *IEEE Access*, vol. 8, pp. 109648-109661, 2020. Article (CrossRef Link)
[35] Dlib. [Online]. Aviailable: http://dlib.net/
[36] J. A. Buolamwini, "Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers," *Massachusetts Institute of Technology*, 2017. Article (CrossRef Link)
[37] 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*, 2020. Article (CrossRef Link)
[38] 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 (CrossRef Link)



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. And he is now a professor at the school of games in Hongik University.

Teasung Hahn received an BS degree at Sogang University in 2006. After graduation, he worked as a programmer in the game industry. And he is now a technical director at the NCsoft Korea.