Journal of Korea Multimedia Society Vol. 25, No. 11, November 2022(pp. 1582-1590)
https://doi.org/10.9717/kmms.2022.25.11.1582

Face Recognition Research Based on Multi-Layers Residual Unit CNN Model

Zhang Ruyang[†], Lee Eung-Joo[††]

ABSTRACT

Due to the situation of the widespread of the coronavirus, which causes the problem of lack of face image data occluded by masks at recent time, in order to solve the related problems, this paper proposes a method to generate face images with masks using a combination of generative adversarial networks and spatial transformation networks based on CNN model. The system we proposed in this paper is based on the GAN, combined with multi-scale convolution kernels to extract features at different details of the human face images, and used Wasserstein divergence as the measure of the distance between real samples and synthetic samples in order to optimize Generator performance. Experiments show that the proposed method can effectively put masks on face images with high efficiency and fast reaction time and the synthesized human face images are pretty natural and real.

Key words: Face Recognition, CNN Model, Deep Learning, Generative Adversarial Network

1. INTRODUCTION

COVID-19 refers to pneumonia caused by a new type of coronavirus infection that began to spread in 2019, and is an acute respiratory infectious disease. The virus that causes this pneumonia can spread widely among people through respiratory droplets. In addition, objects that a person infected with the virus has come into contact with may also have residual virus, and people may become infected by touching these objects. Therefore, wearing a mask to travel and maintaining social distance in public places has become an important method to prevent the spread of the epidemic. At the same time, because the virus has the characteristics of contact transmission, the use of contact identification methods such as fingerprints or palm prints in public places also poses security risks. Facial recognition systems are much safer than other recognition methods because they avoid unnecessary contact. When masks become a necessity of life, it also poses challenges to existing face recognition systems [1]. The current deep learning- based face recognition method has achieved a good recognition rate in the face recognition of unobstructed objects, but it can no longer accurately identify the identity in the face of a large area of occlusion [2]. The main reason is that when training the face recognition neural network model, the face data wearing masks was not used for training [3]. Therefore, in order to improve the recognition rate of the face recognition system facing the mask occluded face, a mask-wearing face dataset with a large number of samples is needed. In the current lack of this type of data set, in order to better train the neural network to recognize the face wearing a mask, this paper solves this problem by wearing a mask to the face image in the existing face recognition data set.

※ Corresponding Author: Lee Eung-Joo, Address: (48520) Sinseon-ro, Nam-gu, Busan, Korea, TEL: ***-****-****
E-mail: ejlee@tu.ac.kr

Receipt date: May. 6, 2022, Revision date: Nov. 25, 2022
Approval date: Nov. 30, 2022

[†] Dept. of Information and Communication Engineering, Graduate School, Tongmyong University
(E-mail: dlzry@naver.com)
[††] Dept. of Information and Communication Engineering, Graduate School, Tongmyong University

2. RELATED TECHNOLOGIES

2.1 STN

The spatial transformer networks (STNs) module mainly consists of three parts: localization network, parameterized sampling grid, and image sampling [4]. The input of the localization network is the original image, and the output is a transformation parameter p, which maps the coordinate relationship between the input image and the ideal image. The parameter sampling network performs affine transformation on the feature image, and obtains the corresponding feature relationship by transforming the parameters and the coordinate position of the input feature map. Image sampling is to transform the original image through the feature relationship obtained by the first two networks to obtain the desired image. The main idea is to spatially transform the input image and output a transformed ideal image. This paper will use the spatial transformation network to transform the mask image to make it conform to the contour of the face, so as to obtain a realistic face image wearing a mask. One of the main reasons for using the spatial transformation network in this paper is that it is not necessary to label the control points of the mask image in advance, which improves the practicability of the algorithm. To achieve this goal, this paper will use a generative adversarial network to optimize the parameter p of the spatial transformation network.

2.2 Pyramid Convolutional Network

PyConv uses a pyramid-structured convolution, which contains convolution kernels of different depths and scales, and can extract features of different scales at the same time [5]. The structure of PyConv is shown in Fig. 1. It contains a pyramid composed of n layers of convolution kernels in different sizes, which can use multi-scale kernels to process the input without increasing the computational complexity and the number of parameters. The kernel contains different spatial scales, the larger the convolution kernel scale, the lower the depth. Since PyConv uses convolution kernels of different depths in different layers, it is necessary to divide the input features into different groups and perform convolution calculations independently, which is called group convolution.

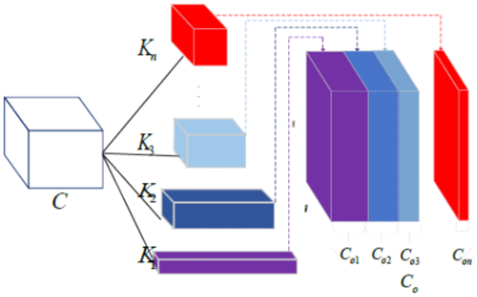


Fig. 1. The Model of the Pyramid Convolutional Network.

During the construction process, the number of channels in each layer of PyConv should be the same, which means that the number of input channels must be an exponential times of 2.

2.3 WGAN

Generative adversarial networks have always been difficult to train and unstable due to the need to achieve Nash equilibrium during training. The instability problem can also lead to mode collapse, resulting in a lack of diversity in sample generation, which is difficult to improve even with increased training time [6]. The Wasserstein GAN better solves the problem of unstable training, and no longer needs to carefully balance the training levels of the generator and the discriminator, ensuring the diversity of the generated results [7].

3. The combination of STN and WGAN- div generative adversarial network using Pyramid convolutional nerual network

The goal of this paper is to perform image synthesis given a normal face image I_F , a mask image I_{MF}, and a mask M. Perform spatial transformation on the mask image, correct its perspective, position, and orientation, so that the synthesized photo can be more natural.

Compared with the method of superimposing the labeled mask image on the recognized face through key point positioning, this paper proposes an image synthesis model that uses a pyramid convolution improved WGAN-div neural network combined with a spatial transformation network. (Py-WGAN-div), the model does not require any advance annotation of synthetic images during training. The model takes the generative adversarial network as the main body, and improves the neural network part of the generator and the discriminator respectively, and its structure is shown in Fig. 2.

In the generator part, the original generative adversarial network generates new images through a random noise. However, the directly generated images will have many problems, such as the low resolution of the generated face, and the wrong mask being synthesized on the face as the skin color. The purpose of this paper is to construct paired face data (including image pairs of faces without masks and wearing masks), rather than generating random faces. Therefore, the generator in this method generates a set of updated deformation parameters $\triangle_p$ (and the deformation parameters are continuously updated as the optimization progresses).

Compared with standard single convolution, multi-scale convolution can expand the receptive field of the convolution kernel without additional parameters, and obtain different spatial resolutions and depths due to the use of different sized convolution kernels [8]. The depth of the convolution kernel decreases as the size decreases, so that the convolution kernels of different sizes bring complementary information and help to obtain richer features.

In the structure design of the generator network, this paper adopts the network structure of PyConv. First, the input normal face image and mask image are superimposed by the number of channels, and then features are extracted through a large 7×7 convolution kernel. The purpose of using a large 7×7 convolution kernel is to preserve the information of the original image as much as possible and reduce the amount of computation. Then four PyConv volume base layers with batch normalization layers removed. The purpose of removing the batch normalization layer in the convolutional

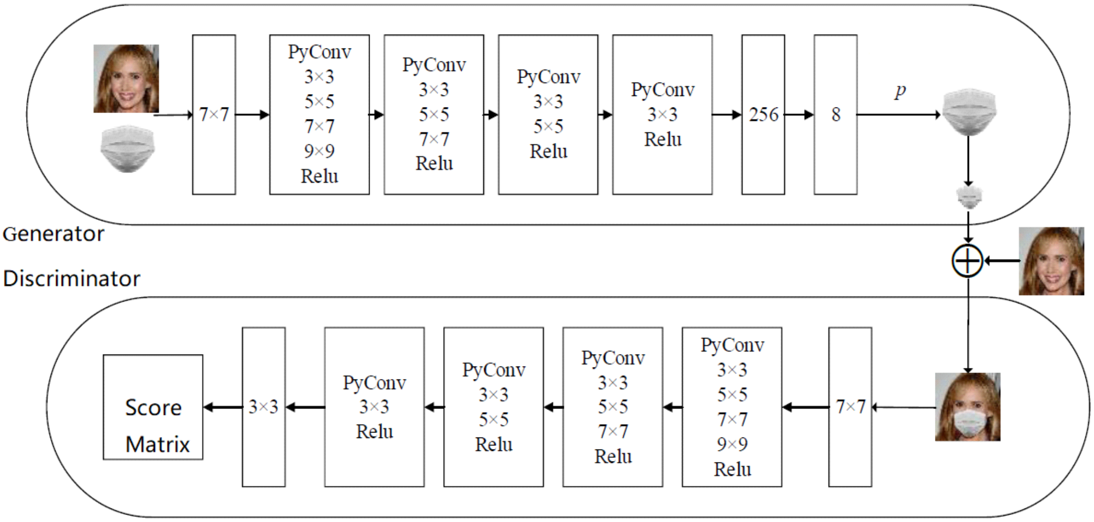


Fig. 2. The overall structure of the network.

layer in PyConv is to reduce the computational complexity and improve the training efficiency. The first PyConv layer in the generator uses four convolution kernels of different sizes (3×3, 5×5, 7×7, 9×9) to obtain features of different scales to enhance the feature extraction capability of the model. The output of the generator is an 8-dimensional vector that is used as a parameter for the spatial transformation network.

The input data of the discriminator is composed of real face images wearing masks and synthetic masks wearing images. The synthetic image is generated by the fusion of the deformed mask image and the unoccluded face image generated by the spatial transformation network, and the parameters of the spatial transformation network are generated by the generator. After the image is discriminated by the discriminator network, a score is output to represent the quality of image synthesis. The discriminator network is compared with the generator in that it does not use a fully connected layer, but passes a 3×3 convolution to obtain a 5×5×1 matrix to calculate the score.

This paper will use the objective function of WGAN-div as the optimization index. Although WGAN has been optimized to a large extent compared with GAN network, it still shows the problems of slow convergence and difficult training during the training process [9]. The main reason is that weight reduction is used directly when dealing with the constraints of the Lipschitz continuous condition, and most of the weights are within plus or minus 0.01 [10]. The discriminator hopes to widen the gap between real and fake samples as much as possible. Later, WGAN-GP and SNGAN proposed by other researchers achieved Lipschitz constraints through gradient penalty and spectral normalization methods, respectively. And WGAN cannot use momentum-based optimization algorithms. WGAN-div proposes the W divergence shown in formula (1) to really reduce the distance between the two distributions, and the settings of k and p are selected according to experience.

$$W_{k,p}(P_r,P_f)=\max_{D\in C^1} E_{x\sim P_r}[D(x)]-E_{x\sim P_f}[D(x)] - kE_{x\sim random}\left[\left(\|\nabla_x D(x)\|^p\right)\right] \quad (1)$$

4. EXPERIMENT RESULT

4.1 Experimental data

The image data used for training in this paper is selected from the face mask dataset like AR face

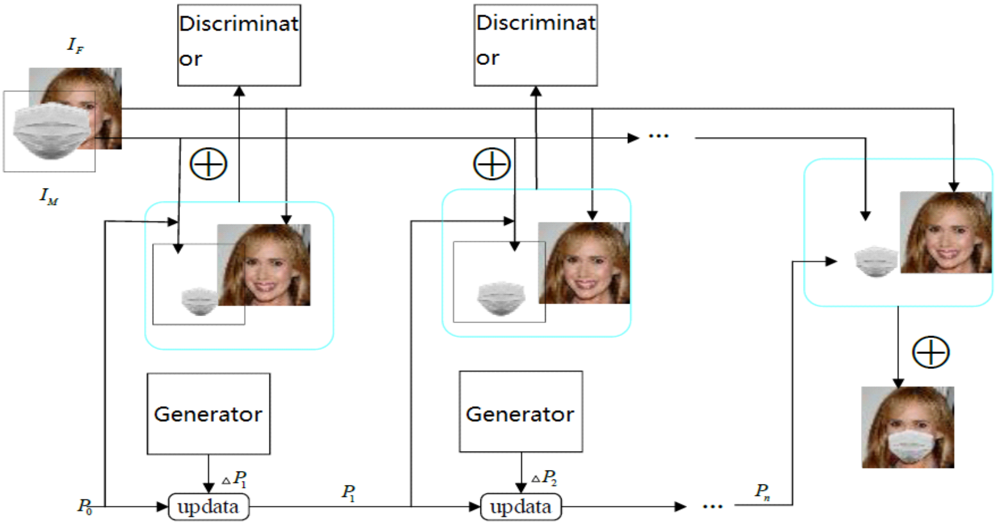


Fig. 3. Network training flow chart.

database and so on, the images captured from the Internet using web crawlers, and some images selected from the face mask dataset synthesized by other researchers [11]. After random translation, rotation, and scaling of the selected images, a total of 158,462 face images with masks were obtained as a dataset. Among them, 142,618 images (about 90%) are used as the training set to train the discriminator, and the rest are used as the test set. The image size in the dataset is uniformly scaled to 144 × 144 pixels [12]. And hand-made 20 masks of various types and colors. The size of the mask picture is also 144 × 144 pixels, and the mask is basically located in the center of the picture, as shown in Fig. 4.

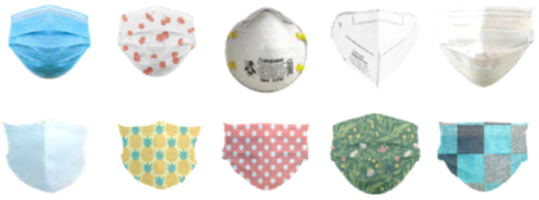


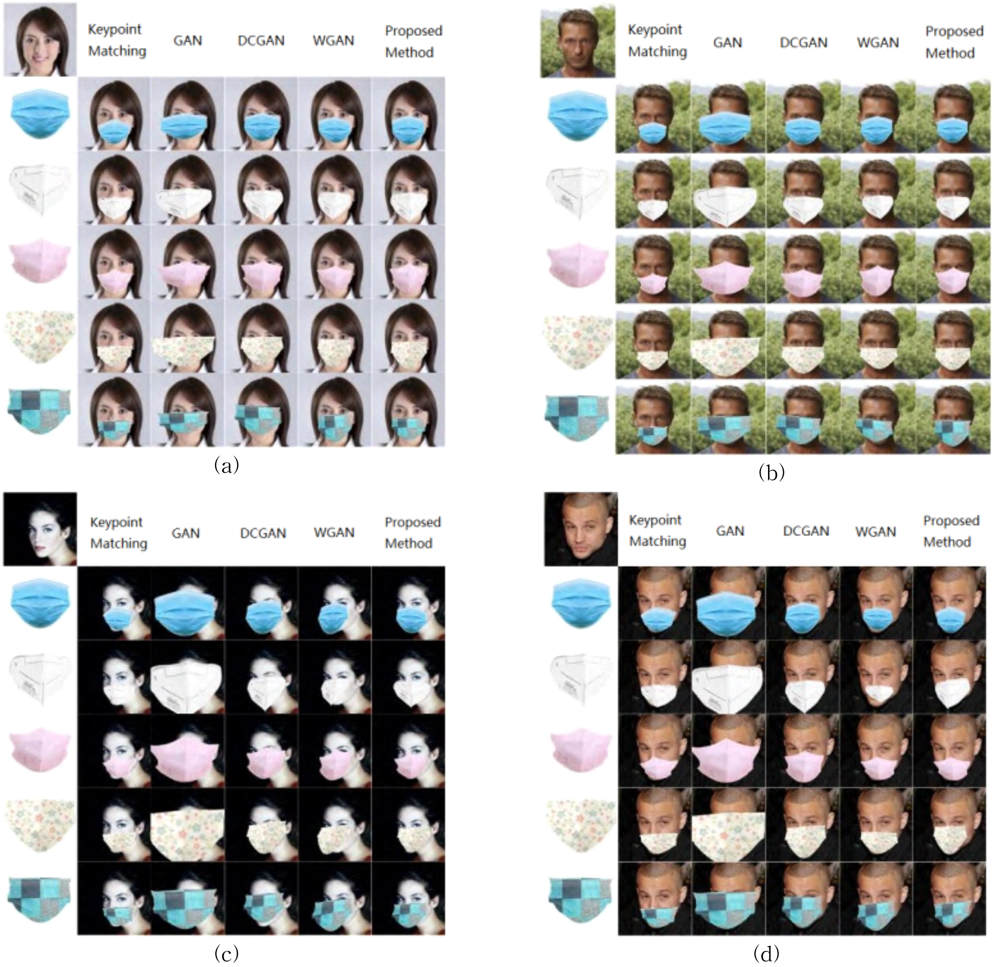
Fig. 4. Mask Images used in the experiment processing.

4.2 Analysis of results

The experiment uses the Py-WGAN-div-based generative adversarial network to train the training set, and randomly selects face pictures and mask pictures from the training set during training. The synthesized images become more natural after every 50,000 iterations and updating the transformation parameters of the mask when training with our method. It can be seen that the position of the mask gradually becomes more suitable for the face as the training progresses, and finally a more realistic image of the face wearing the mask is obtained.

When comparing algorithms, this paper selects the key point matching algorithm, algorithms based on GAN, DCGAN, and WGAN for comparison. According to the experiments in this paper, the original generative adversarial network algorithm has poor effect on face mask synthesis, and the mask can hardly cover the correct position. Therefore, this paper adds a space transformation network (STN) between the generators and discriminators of GAN, DCGAN, and WGAN in the comparison group algorithm for a more reasonable comparison. Fig. 5 is a comparison chart of the effect of different algorithms corresponding to different masks and face synthesis. In the comparison, 5 masks of different styles were selected, including the most common blue surgical masks, KN95 masks, pink, checkered, and speckled masks. The comparison experiment also selected four different skin tones and backgrounds, including various skin tones and background colors. Different faces, different masks, and different algorithms are compared, and the results are shown in Fig. 5. As can be seen in Fig. 5(a) and Fig. 5(b), when the face pose is relatively good, various algorithms can better synthesize the mask into the face image. Among them, the algorithm based on key point matching and the algorithm in this paper are the best, but the image produced by the algorithm in this paper is more natural and realistic.

It can be clearly seen in the figure that the effect of the pictures generated by the algorithms based on GAN and DCGAN is relatively poor, and the mask will cover the eyes or completely exceed the outline of the face; while the effect of the WGAN method is better than that of the algorithms based on GAN and DCGAN, However, synthetic masks cannot fit the contours of the human face well. The face in Fig. 5(c) is skewed to the right. In the results of other methods except this method, the mask can only fit the left half of the face well, and the mask on the right half of the face will be too large. At this time, although the algorithm in this paper is not very ideal, it can basically fit the contour of the face, which is relatively better. For the face with bowed head in Fig. 5(d), except for the algorithm based on key points and the algorithm in this paper, the face mask image obtained by oth-



(a) (b) (c) (d)

Fig. 5. Comparison of the results of different algorithms using 4 samples from (a) to (d).

er algorithms has a large distortion, which is manifested in that it cannot cover the chin, mouth and nose, or cover it. The area that should not be covered, in contrast, when the face pose in the image is not facing the camera, the face mask image obtained by the algorithm in this paper is still better. To sum up, it can be concluded that when the face pose in the image is not facing the camera, all the algorithms obtained face images with masks are lacking, but the algorithm in this paper is obviously better than other algorithms in the effect of face mask synthesis. It can basically fit the contours of the face, without blocking the undesired parts, and the details are more realistic.

In order to compare the synthesis effects of different algorithms more objectively, this paper uses IS Score (inception score), structural similarity (Structural SIMilarity), and depth feature metric image similarity (learned perceptual image patch similarity) three indicators to objectively evaluate each algorithm [13]. The effect of different GAN models on mask synthesis.

The IS evaluation method feeds the generated images into the trained Inception classification model. The output of the Inception classification model is a 1000-dimensional label, each dimension of which represents the probability that the input image belongs to a certain class. If the training results are good, the results will be more concentrated. The results are shown in Table 1. Although the masks generated by GAN and DCGAN networks are far from the desired results, their IS scores are higher than our algorithm. This phenomenon occurs because although IS can be used as an indicator of image synthesis quality, this indicator cannot really reflect the details in the synthesized image, such as whether the mask correctly covers the person's mouth and nose, and whether the area covered by the mask is too large. Loss of facial information, other parts of the face are enough to be retained and so on. Therefore, this paper also conducts manual evaluation on the generated images. The method of manual evaluation is to score 100 groups of different face images to 20 people. Each group of images includes the original image of the face, the type of mask worn, and the images synthesized by the two methods. Whether the image is real or not is judged, and the judgment results are shown in Table 1. It can be seen from the discrimination results that the images synthesized by this method are more realistic. In addition, this paper adopts two relatively objective evaluation methods, structural similarity (SSIM) and learned perceptual image patch similarity (LPIPS), to evaluate the generated images. SSIM is a reference image quality evaluation index, which compares and measures the similarity of images from three aspects: image brightness, image contrast and image structure [14]. The depth feature metric image similarity uses feature maps extracted by a pre-trained neural network to quantify the perceptual difference between two images. The more similar the two images, the closer the distance. The comparison results of SSIM and LPIPS indicators are shown in Table 2.

It can be seen from Table 2 that the proposed algorithm has a higher structural similarity than the comparison algorithm. The very small similarity of the depth feature metric image indicates that the distance between the synthetic mask wearing image and the real face wearing a mask is very close, which fully proves the effectiveness of the algorithm in this paper.

Table 1. Image authenticity recognition rate.

Method	Inception Score	Artifical(%)
GAN	2.641	0
DCGAN	2.492	30
WGAN	2.185	72
Proposed Method	2.297	79

Table 2. Recognition performance results.

Method	Structural similarity	Learned Percetual image patch similarity
GAN	0.625	0.181
DCGAN	0.793	0.110
WGAN	0.907	0.049
Proposed Method	0.941	0.011

5. CONCLUSION

This paper proposes a method of wearing masks to face images that combines generative adversarial network and spatial transformation network, and adopts the transformation parameters of spatial transformation network generated by generative adversarial network in the design, instead of directly generating face and mask fusion The special design of the post image. The multi-scale convolution method is used when designing the neural network, so that the generator can better extract features. During training, the Wasserstein divergence is used as a calculation to measure the distance between two different samples, which overcomes the problem of difficulty in training the generative adversarial network and prone to mode collapse. Compared with other methods, the method in this paper is more realistic on the synthetic image, and the mask fits the face better.

The experimental results show that the neural network model can learn the corresponding transformation parameters and synthesize high-quality images of faces wearing masks when neither the face nor the mask is marked. The experimental results confirm that the fused face image is not distorted and retains the facial features well, and also covers the mask to the correct position of the face. In the research process, it is also found that the method in this paper is not perfect when only half of the face is visible due to the angle problem. Therefore, how to synthesize masks without distortion on face pictures from any angle will be a further research direction; further, we will use the mask-wearing face dataset produced in this paper to conduct face recognition research on mask occlusion.

REFERENCE

[1] J. Wang and E.J. Lee, "Low Resolution Rate Face Recognition Based on Multi-scale CNN," *Journal of Journal of Korea Multimedia Society*, Vol. 21, No. 12, pp. 1467-1472, 2018.

[2] T. Yaniv, M. Yang, R.M. Aurelio, and L. Wolf, "Deepface: Closing the Gap to Human Level Performance in Face Verification," *Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701-1708, 2014.

[3] W.K. Xu and E.J. Lee, "Human-computer Catural User Interface Based on Hand Motion Detection and Tracking," *The Journal of Multimedia Information System*, Vol. 15, No. 4, pp. 501-507, 2012.

[4] S. Florian, K. Dmitry, and J. Philbin, "Facenet: A Unified Embedding for Face Recognition and Clustering," *Proceedings of 2015 IEEE Conference on Computer Vision and Patten Recognition*, pp. 815-823, 2015.

[5] A.R. Syafeeza, M. Khalil-Hani, S.S. Liew, and R. Bakhteri, "Convolutional Neural Network for Face Recognition with Pose and Illumination Variation," *Journal of IJET 2014*, Vol. 6, No. 1, pp. 44-57, 2014.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B Xu, David Warde-Farley, S Ozair, et al., "Generative adversarial Networks," *Journal of the Communications of the ACM*, Vol. 63, No. 11, pp. 139-144, 2020.

[7] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A.C. Courville, "Improved Training of Wasserstein GANs," *Proceedings of Advances in Nerual Information Processing Systems*, pp. 5767-5777, 2017.

[8] J. Wu, Z. Huang, J. Thoma, D Acharya, and L. Van Gool, "Wasserstein Divergence for GANs," *Proceedings of the European Conference on Computer Vision,* pp. 653-668, 2018.

[9] C. Szecedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., "Going Deeper with Convolutions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9, 2015.

[10] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," *International Conference on Learning Representations*, 2018.

[11] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730-3738, 2015.

[12] N. Pandey and A. Sacakis, "Poly-GAN: Multi-Conditioned GAN for Fashion Synthesis," *Journal of Neurocomputing*, Vol. 414, pp. 356-364, 2020.

[13] M. Jaderberg, K. Simonya, A. Zisserman, and K Kavukcuoglu, "Spatial Transformer Networks," *Proceedings of the Advances in Neural Information Processing Systems,* pp. 2017-2025, 2015.

[14] K.M. He, X.Y. Zhang, S.Q. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv Preprint*, arXiv:1512. 03385, 2015.

Ruyang Zhang

received his B.S. at Dalian Polytechnic University in China (2012-2016) and Master Degree in Tongmyong University(2016-2018). Currently he is studying in the Department of Information and Communication Engineering in Tongmyong University, Korea for his Doctor degree. His main research areas are image processing and face recognition.

Lee Eung-Joo

received his B. S., M. S. and Ph. D. in Electronic Engineering from Kyungpook National University, Korea, in 1990, 1992, and Aug. 1996, respectively. Since 1997 he has worked with the Department of Information & Communications Engineering, Tongmyong University, Korea, where he is currently a professor. From 2005 to July 2006, he was a visiting professor in the Department of Computer and Information Engineering, Dalian Polytechnic University, and from Dec 2018 he was appointed honorary professor of Dalian Polytechnic University, China. His main research interests include biometrics, image processing, and computer vision.