Journal of Multimedia Information System VOL. 8, NO. 2, June 2021 (pp. 85-92): ISSN 2383-7632 (Online)
http://doi.org/10.33851/JMIS.2021.8.2.85

Frontal Face Generation Algorithm from Multi-view Images Based on Generative Adversarial Network

Young- Jin Heo[1], Byung-Gyu Kim[1*], Partha Pratim Roy[2]

Abstract

In a face, there is much information of person's identity. Because of this property, various tasks such as expression recognition, identity recognition and deepfake have been actively conducted. Most of them use the exact frontal view of the given face. However, various directions of the face can be observed rather than the exact frontal image in real situation. The profile (side view) lacks information when comparing with the frontal view image. Therefore, if we can generate the frontal face from other directions, we can obtain more information on the given face. In this paper, we propose a combined style model based the conditional generative adversarial network (cGAN) for generating the frontal face from multi-view images that consist of characteristics that not only includes the style around the face (hair and beard) but also detailed areas (eye, nose, and mouth).

Key Words: GAN, StyleGAN, cGAN, Deep learning, Classification, Frontal face.

I. INTRODUCTION

Recently, the deep learning is triggering the development of image analysis of various data types. Deep learning keeps the state of the art (SOTA) model in almost every field of image analysis by finding and learning the features, shapes and patterns of images. In the classification of ImageNet [1] and CIFAR10 [2] datasets, models such as InceptionResnet [3], and Big transfer (BiT) [4] utilizing ResNet [5] have been selected as SOTAs and many researchers are actively studying to achieve higher accuracy [18], [19], [20], [21], [22], [23], [24]. Based on these models, face expression recognition (FER), Face Recognition and Face Generation are being investigated. Most of face datasets are often taken in a laboratory or controlled in front. However, in real life, there are more pictures of different directions than the exact frontal one. There is a research [6] that present problems and analyzed only profile face data. If we can make a generation as good quality about the front one from the side view image, we can utilize it in various applications.

The main models in a field of image generation are Ganerative Adversarial Network (GAN) [8] and Variational AutoEncoder (VAE) [9]. In particular, many different GAN models focus on producing high image quality which the style is transferred. Concretely, there are some networks that generate images by extracting the charac-teristics of the image and synthesizing the disentangled attributes such as StarGAN [10], InterfaceGAN [7], CycleGAN [11], DiscoGAN [12], and StyleGAN [13].

Shen et al. extracted the latent space through the trained StyleGAN and proposed a methodology for various style edits [7]. As shown in Fig. 1, the more we change to the profile from frontal face, the more certain people lose their identity.

In this paper, we suggest a new method only using deep learning without image editing techniques for generating the frontal face from images of other directions. Conditional generative adversarial network (cGAN) generated an image corresponding to X according to stationary post-conditions Y as posterior X|Y [14]. Using this method, we generate a frontal face while Y is style condition, but we get style vector by our models instead of fixed value Y.

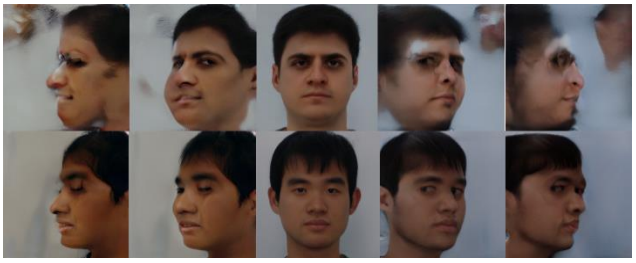


Fig. 1. Pose generation example of interfaceGAN [7].

Manuscript received June 12, 2021; Accepted June 17, 2021. (ID No. JMIS-21M-06-020)
Corresponding Author (*): Byung-Gyu Kim, Dept. of IT Engineering, Sookmyung Women's University, Seoul, Korea, + 82-2-2077-7293, bg.kim@sookmyung.ac.kr.
[1]Dept. of IT Engineering, Sookmyung Women's University, Seoul, Korea, yj.heo@ivpl.sookmyung.ac.kr
[2]IIT Roorkee, Haridwar Highway, Roorkee, Uttarakhand 247667, India, proy.fcs@iitr.ac.in

The four models that each model extracts diverse style of face are proposed. First, style-encoder model has a structure similar to discriminator, obtaining style vectors by image, and it's called StyleEncoder.

The second model is advanced version of first model which has an attention mechanism by 1×1 convolution, defining A-StyleEncoder. Also, without training model which finds the style, use the classification model trained by another dataset to generate a frontal view. One of them is InceptionResNet trained by VGGFace2 and we compare the results each of models in experiments. A-StyleEncoder model gets the properties which are coarse feature of face like hair, beardand outline. The pre-trained Inception Resnet model extracts details of face such as eye, nose and mouth type. To combine these features together, we merge the model's output and it could be considered not only specific area of face but also overall style. After, a style vector as input on generator, generates frontal face.

In next section, we will introduce other models and methodologies that define styles from image related to the suggested models. We propose a method for training each network's parameters and the four types of style extraction model in approach in Section 3. In Section 4, we compare the output image of each model and PSNR result with the original image. Also, we will visualize the vector from the StyleEncoder model to check whether the style was extracted properly and specify the parametric structure of generator, discriminator and styleEncoder model. Finally, we explain limitation of the research, future work and important effect in Section 5.

II. RELATED WORK

2.1. cGAN

The conditional-GAN [14] trains the model by conditioning a class label on the generator and the discriminator in order to generate an image of a given type that meets the desired conditions. The learning method is the addition of the condition y from the existing GAN structure as the following:

$$\min_{\mathrm{G}} \max_{D} V(D,G) = \mathbb{E}_{x\sim P_{data(x)}}[\log D(x|y)] + \mathbb{E}_{z\sim P_{z(z)}}\left[\log\left(1 - D\left(G(z|y)\right)\right)\right], \quad (1)$$

where D and G represent the discriminator and generator, respectively. x is a real image, z is a noise input for generating fake image. Conditional Generative Adversarial Network overcomes the limitations of the generated images with random Gaussian sample. In this paper, image-dependent vector by style network replaces class label y. The generator model is responsible for generating identically corresponding frontal face by the designed style encoder model.

2.2. StyleGAN, InterfaceGAN

A Style-Based Generator Architecture (StyleGAN) for GANs by NVIDIA, presents an advanced model which generates high-quality image [10]. The StyleGAN generates the image gradually, starting from a very low resolution to a high resolution. It modifies the central feature corresponding to each resolution for each level separately. Resolution of up to 82 affects pose, general hair style, face shape, and other levels affect more micro features. Also, the StyleGAN employed AdaIN (adaptive instance normalization) module that editing each channel by information vector w. This mechanism produces state of the art results with high resolution images, allowing for a better understanding of GAN outputs.

The InterfaceGAN proposes a novel approach, interpreting the latent space of GANs for semantic face editing [7]. Specifically, this research made latent space of image for finding semantic subspaces and using trained face synthesis model. InterfaceGAN is capable of changing several semantic elements (pose, gender, glasses, etc.) of application controllable trained StyleGAN model, but pose generation loses the meaning of determining the same identity in Figure 1. In this paper, we introduce how to produce an image without losing identity with a simple GAN structure.

2.3. Classification Model

As a backbone model, the residual network is widely taken into account for SOTAs. InceptionResnet [3] is Inception style networks that utilize residual connections instead of filter concatenation. One of models, Inception Resnet-v1, is a hybrid Inception version with significantly improved recognition performance. We select the InceptionResnet-v1 pre-trained by VGGFace2.

III. PROPOSED APPROACH

3.1. Formulation

We point to the fact that the cGAN has fatal drawback property that class label y should be a fixed vector, passing on generator. If we want to produce an image explained by more complex style which unable to express in simple class, we cannot generate the desired image. On the other hand, a style-encoder alternates particular label, representing a different style of every image. Therefore, the formulation for training network as the following:

$$\theta_G^*, \theta_{\mathrm{D}}^*, \theta_S^* = \min_{\theta_G} \max_{\theta_D} f(\theta_G, \theta_D, \theta_S). \quad (2)$$

Each of $\theta_G, \theta_D, \theta_S$ defines parameters of the generator, discriminator, style-encoder network and θ^* is the optimal

Journal of Multimedia Information System VOL. 8, NO. 2, June 2021 (pp. 85-92): ISSN 2383-7632 (Online)
http://doi.org/10.33851/JMIS.2021.8.2.85

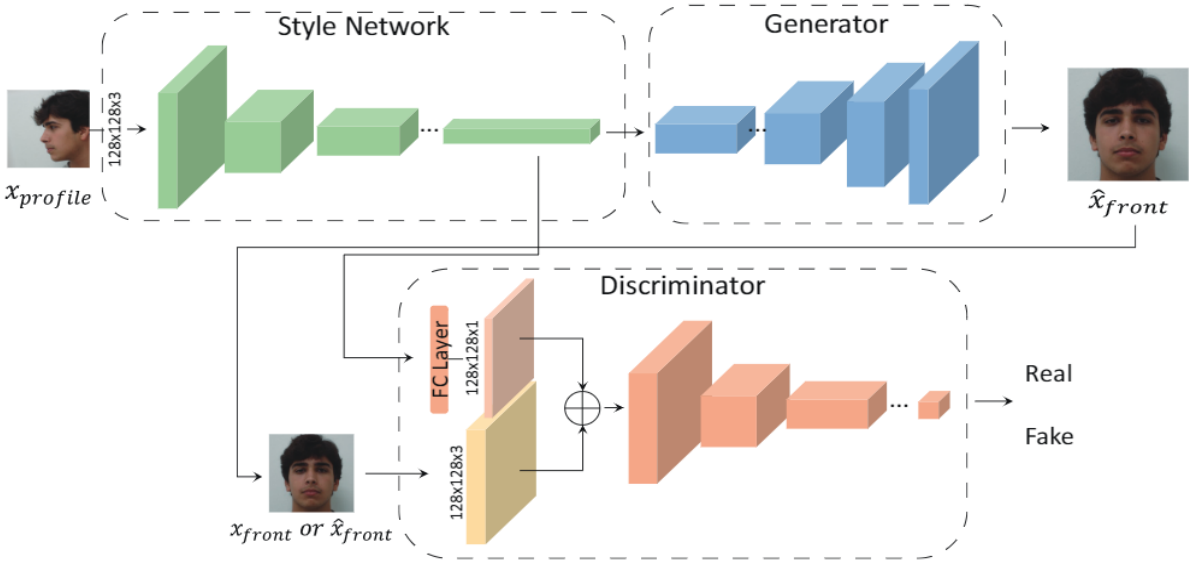


Fig. 2. Overall structure of the model: $\hat{x}_{front}$ is the predicted front and x_{front} means the real front face.

parameters. Model process is termed function f, then the optimal parameters of G and D are computed by:

$$\theta_G^* = \underset{\theta_G}{\operatorname{argmin}} f\left(\theta_G(\theta_S), \theta_D^*(\theta_G(\theta_S))\right), \quad (3)$$

$$\theta_D^* = \underset{\theta_D}{\operatorname{argmax}} f(\theta_G^*(\theta_S), \theta_D). \quad (4)$$

Finally, we get f on $x \in R^{C\times W\times H}$ (C: channel, W, H: image of width and height.) in that $x_{frontal}$ is the front face of x and $x_{profile}$ is side one. Therefore, we can express as:

$$\begin{aligned} & f(\theta_G, \theta_D, \theta_S) \\ & = \mathbb{E}_{x\sim p_{data}}[\log (D(x_{frontal}|S(x_{profile}; \theta_S); \theta_D))] \\ & + \mathbb{E}_{x\sim p_{data}}[\log (1 - D(G(S(x_{profile}; \theta_S); \theta_G) \\ & \qquad |S(x_{profile}; \theta_S); \theta_D))]. \end{aligned} \quad (5)$$

By this formulation, we can get the frontal face image by multi-view images.

3.2. Overall Network

The overall structure of the designed model is shown in Figure 2. If we insert the side face as the input, we extract the style vector through the style network. Extracted style vector is used in both generator and discriminator. The generator generates the front real image with the entered style vector. Discriminator receives the actual front or predicted front and side profile style vectors as inputs and finally learns whether the frontal face of the person is real or fake.

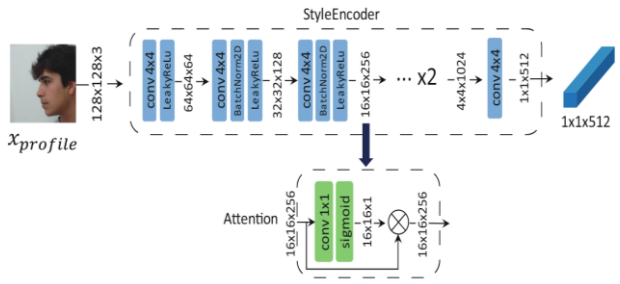
3.3. Style Network Architecture

To give the generator using various style vectors, we design four types of models that can extract styles.

3.3.1. StyleEncoder (SE)

The StyleEncoder model consists of architecture similar to discriminator in that encodes an image. For an image, the image features are represented through a simple Convolutional Neural Network (CNN) structure of five layers. The activation function uses LeakyRelu and proceeds Batch Normalization (BN). The final size of the Style Vector is 1×1×512. If there is no attention module in Figure 3 of (a), it is a simple style-encoder model.

3.3.2. A-StyleEncoder (A-SE)



(a) A-StyleEncoder model structure.

(b) A-StyleEncoder + InceptionResnet model structure.

Fig. 3. Primary style network architectures. Merged version of A-StyleEncoder and InceptionResnet (A-SE+IR).

A-StyleEncoder is an advanced model of style-encoder as adding attention module to enhance features. The structure

is Fig. 3 (a) and the feature selected in the middle stage of the style-encoder passes through a 1x1 convolution, sigmoid function and again multiplies itself. The output x_l of middle layer can be given as:

$$LeakyRelu\left(BatchNorm\left(Conv(x)\right)\right) = \mathcal{F}(x), \quad (6)$$

$$x_l = \mathcal{F}(x_{l-1}) \otimes \sigma(C(\mathcal{F}(x_{l-1}))), \quad (7)$$

where $\otimes$ denote element-wise multiplication, C(·) is point-wise convolution operation by 1x1 filter, and l is the output layer of attention module.

3.3.3. InceptionResnet (IR)

InceptionResnet [3] outperforms in area of classi-fication and is widely used as a backbone network. We consider a pre-trained InceptionResnet-v1 by vggface2 dataset. This method recognizes our dataset image as 100 percent accuracy. As with other style-encoder models, the face style is drawn in size 1×1×512 with the input image.

3.3.4. A-StyleEncoder + InceptionResnet (A-SE + IR)

Finally, we concatenate the outputs of A-StyleEncoder and InceptionResnet to make use of each property which determines type of style. A-StyleEncoder defines a style of around face (hair, beard, face shape) and InceptionResnet decides a feature to recognize people as eye, nose and mouth. Thus, to generate image of all types, we concatenate features together as Eq (8):

$$style = S_{A-StyleEncoder}(x) \oplus S_{InceptionResnet}(x). \quad (8)$$

By concatenating these outputs as shown in Fig. 3(b), it is feasible to generate the frontal face by profile face.

IV. EXPERIMENTS

4.1. Dataset and Training Detail.

We used the FEI Face dataset [15] with 11 directions (1 front and 10 sides) for each of 200 individuals. We preprocess an image with Multi-task Cascaded Convolutional Networks (MTCNN) [16] to detect face and crop. The size of final cropped image is 3×128×128 (C×W×H). The input is ten profiles turned 18 degrees from -90 to 90 degrees. Both generator and discriminator were trained alternately, and the style model was trained with the generator. We set 100 epochs and took 20 hours with Geforce GTX1080Ti. We used the Adam optimizer and learning rate 0.0002.

4.2. Features Visualization.

Before entering the generator and discriminator, we visualize the output features of A-StyleEncoder and InceptionResnet by concatenating them. The size of the total output feature was 1×1×1024. We used the principal component analysis (PCA) technique to project it in two dimension. The output of each model for visualization was aggregated as follows:

$$PCA(S_{A-StyleEncoder}(x) \oplus \alpha \times S_{InceptionResnet}(x)), \quad (9)$$

where α is set to 2, and the result of visualizing in two dimension through the principal component analysis (PCA) function as shown in Fig. 4. It can be seen that people of similar styles are grouped together. This means that the 'A-SE + IR' is effective.

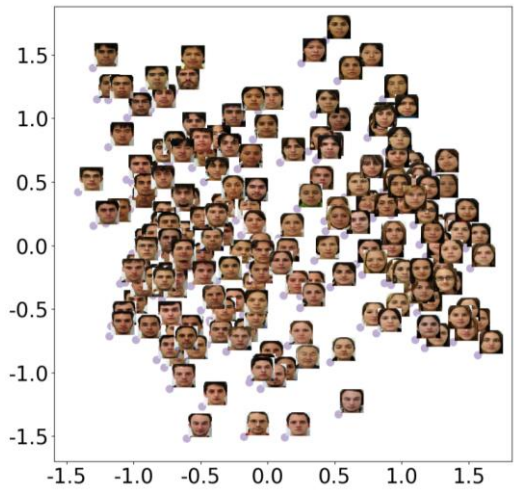


Fig. 4. Visualization of style output of 'A-SE + IR' with all people of dataset.

4.3. Detailed Network Architecture.

Specific architectures of the generator and discriminator are shown in Table 1 and Table 2. Each model was referred to the architecture of DCGAN [17]. The generator produces images with a style vector, and the discriminator concatenates the input image and style to find out the real or false image.

4.4. Performance Analysis.

The result of generating a frontal view from each person's profile compares with the models of Style Network Architecture and the InterfaceGAN using the pre-trained StyleGAN on Flickr-Faces-HQ Dataset (FF-HQ). As we can see in Fig. 5, the merged version (A-SE + IR) of A-StyleEncoder and InceptionResnet outperforms than other style networks. As mentioned, A-StyleEncoder (A-SE) can extract type of hair, face shape (around of face) and select features such as eyes, nose and mouth better than other models on InceptionResnet (IR). The results of the 'A-SE + IR' model for others are shown in Fig. 6. It can be seen that hair styles and facial features are applied independently compared to other generator models. Hence, we consider the characteristics of A-Style Encoder and

InceptionResnet at the same time and get the style of the whole face. Finally, we compared a Peak Signal-to-noise ratio (PSNR) of output results in Table 3. The proposed 'A-SE +IR' model obtained the highest PSNR value.

Table 1. Generator Architecture.

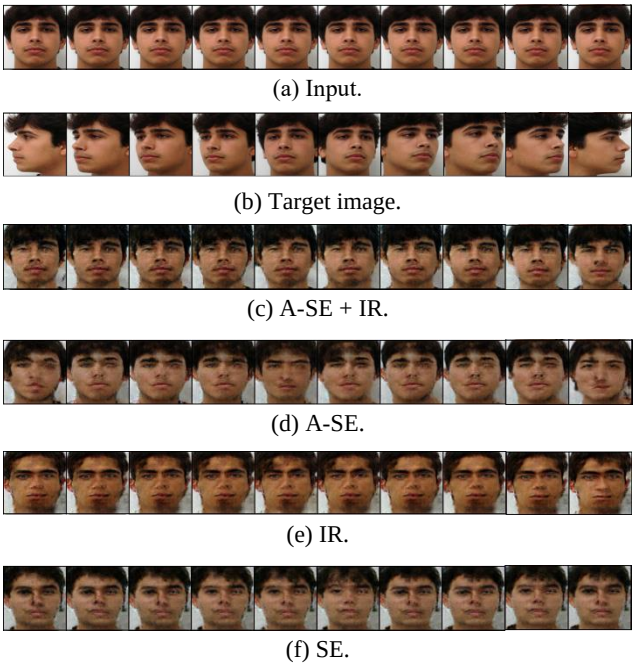
Layer	Norm	Activation	Output Shape
Style s	-	-	1×1×1024
ConvT	BatchNorm	ReLU	4×4×1024
ConvT	BatchNorm	ReLU	8×8×512
ConvT	BatchNorm	ReLU	16×16×256
ConvT	BatchNorm	ReLU	32×32×128
ConvT	BatchNorm	ReLU	64×64×64
ConvT	BatchNorm	Tanh	128×128×3

Table 2. Discriminator Architecture.

Layer	Norm	Activation	Output Shape
Style s	-	-	1×1×1024
FC	-	-	128×128×1
Image x ⊕ s	-	-	128×128×4
Conv	BatchNorm	LeakyReLU	64×64×64
Conv	BatchNorm	LeakyReLU	32×32×128
Conv	BatchNorm	LeakyReLU	16×16×256
Conv	BatchNorm	LeakyReLU	8×8×512
Conv	BatchNorm	LeakyReLU	4×4×1024
Conv	-	-	1×1×1

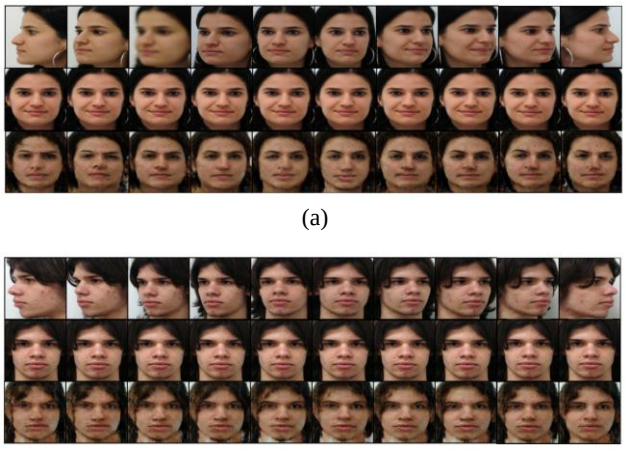
V. CONCLUSION

We have investigated some style extraction models and proposed a style extraction model called 'A-SE + IR' by concatenating the results of the attention style encoder model and the InceptionResnet for giving the condition of the generator and discriminator to generate the front face from the side. Also, we developed a frontal face generation module that would extract complex features by applying a conditional generator. This model not only extracts styles around the face such ascertain people's hair styles, but also confirms that the facial features are well drawn. We verified the possibility of generating a frontal face with reliable quality, from side view images.



(a) Input.
(b) Target image.
(c) A-SE + IR.
(d) A-SE.
(e) IR.
(f) SE.

Fig. 5. Results of all models mentioned in Style Network Architecture.



(a)
(b)

Fig. 6. Results of generating frontal face from multi-viewer image by our best models.

REFERENCES

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in Proceeding of *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

[2] CIFAR10 dataset of Laboratory of Toronto, "https://www.cs.toronto.edu/ kriz/cifar.html".

[3] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint* arXiv:1602.07261, 2016.

[4] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby, "Big transfer (bit): General visual representation learning," *arXiv preprint* arXiv:1912.11370, vol. 6, 2019.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and JianSun, "Deep residual learning for image recognition," in Proceedings of *the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.

[6] Maja Pantic and Ioannis Patras, "Dynamics of facial expression: recognition of facial actions and their

temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 433–449, 2006.

[7] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou, "Interfacegan: Interpreting the disentangled face representation learned by gans," *arXiv preprint* arXiv:2005.09635, 2020.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in Proceeding of *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[9] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint* arXiv:1312.6114, 2013.

[10] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-WooHa, Sunghun Kim, and Jaegul Choo, "Stargan: Unified generative adversarial networks for multi-domainimage-to-image translation," in Proceedings of *the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, 2018.

[11] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei AEfros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proceedings of *the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[12] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim, "Learning to discover cross-domain relations with generative adversarial networks," *arXiv preprint* arXiv:1703.05192, 2017.

[13] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in Proceedings of *the IEEE conference on computer vision and pattern recognition*, pp.4401-4410, 2019.

[14] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint* arXiv: 1411.1784, 2014.

[15] The FEI face database of Laboratory of FEI, "http://fei.edu.br/ cet/facedatabase.html".

[16] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503,2016.

[17] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint* arXiv: 1511.06434, 2015.

[18] Rohit Srivastava, Ravi Tomar, Ashutosh Sharma, Gaurav Dhiman, Naveen Chilamkurti, and Byung-Gyu Kim, "Real-Time Multimodal Biometric Authentication of Human Using Face Feature Analysis," *Computers, Materials & Continua*, vol. 49, no.1, pp. 1-19 (DOI:10.32604/cmc.2021.015466), 2021.

[19] Dami Jeong, Byung-Gyu Kim, and Suh-Yeon Dong, "Deep Joint Spatiotemporal Network (DJSTN) for Efficient Facial Expression Recognition," *Sensors*, vol. 2020, no. 20, p. 1963 (https://doi.org/10.3390/s20071936), 2020.

[20] Ji-Hae Kim, Gwang-Soo Hong, Byung-Gyu Kim, and Debi P. Dogra, "deepGesture: Deep learning-based gesture recognition scheme using motion sensors," *Displays*, vol. 55, pp. 34-45 (https://doi.org/10.1016/j.displa.2018.08.001), 2018.

[21] Ji-Hae Kim, Byung-Gyu Kim, Partha Pratim Roy, and Da-Mi Jeong, "Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure," *IEEE Access*, vol. 7, pp. 41273-41285, 2019.

[22] Dong-hyeon Kim, Dong-seok Lee, and Soon-kak Kwon, "Fall Situation Recognition by Body Centerline Detection using Deep Learning," *Journal of Multimedia Information System*, vol. 7, no. 4, pp. 257-262, 2020.

[23] Woon-Ha Yeo, Young-Jin Heo, Young-Ju Choi, and Byung-Gyu Kim, "Place Classification Algorithm Based on Semantic Segmented Objects," Applied Sciences, vol. 2020, no. 10, p. 9069 (https://doi.org/10.3390/app10249069), Dec. 2020.

[24] S. Mukherjee, S. Ghosh, S. Ghosh, P. Kumar, and P. P. Roy, "Predicting Video-frames Using Encoder-convlstm Combination," in Proceeding of *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2027-2031 (doi: 10.1109/ICASSP.2019.8682158), 2019.

Authors

Young-Jin Heo received her BS degrees in the Department of IT Engineering from Sookmyung Womens University, Korea, in 2019. In 2019, she joined the Department of Computer Engineering for pursuing her MS degree at Sookmyung Womens University.

Her research interests include image generation (GAN), Deepfake Detection (DFDC) and Deep learning.

Byung-Gyu Kim has received his BS degree from Pusan National University, Korea, in 1996 and an MS degree from Korea Advanced Institute of Science and Technology (KAIST) in 1998. In 2004, he received a PhD degree in the Department of Electrical Engineering and Computer Science from Korea Advanced Institute of Science and Technology (KAIST). In March 2004, he joined in the real-time multimedia research team at the Electronics and Telecommunications Research Institute (ETRI), Korea where he was a senior researcher. In ETRI, he developed so many real-time video signal processing algorithms and patents and received the Best Paper Award in 2007.

From February 2009 to February 2016, he was associate professor in the Division of Computer Science and Engineering at SunMoon University, Korea. In March 2016, he joined the Department of Information Technology (IT) Engineering at Sookmyung Women's University, Korea where he is currently an associate professor.

In 2007, he served as an editorial board member of the International Journal of Soft Computing, Recent Patents on Signal Processing, Research Journal of Information Technology, Journal of Convergence Information Technology, and Journal of Engineering and Applied Sciences. Also, he is serving as an associate editor of Circuits, Systems and Signal Processing (Springer), The Journal of Supercomputing (Springer), The Journal of Real-Time Image Processing (Springer), Helyion Journal (Elsevier), and International Journal of Image Processing and Visual Communication (IJIPVC). From 2018, he is serving as the Editor-in-Chief (EiC) of the Journal of Multimedia Information System. He also served as Organizing Committee of CSIP 2011 and Program Committee Members of many international conferences. He has received the Special Merit Award for Outstanding Paper from the IEEE Consumer Electronics Society, at IEEE ICCE 2012, Certification Appreciation Award from the SPIE Optical Engineering in 2013, and the Best Academic Award from the CIS in 2014. He has been honored as an IEEE Senior member in 2015.

He has published over 250 international journal and conference papers, patents in his field. His research interests include software-based image and video object segmentation for the content-based image coding, video coding techniques, 3D video signal processing, wireless multimedia sensor network, embedded multimedia communication, and intelligent information system for image signal processing. He is a senior member of IEEE and a professional member of ACM, and IEICE.

Dr. Partha Pratim Roy received his Ph.D. degree in computer science in 2010 from Universitat Autònoma de Barcelona, (Spain). He worked as postdoctoral research fellow in the Computer Science Laboratory (LI, RFAI group), France and in Synchromedia Lab, Canada. Presently, Dr. Roy is working as Assistant Professor at Indian Institute of Technology (IIT), Roorkee. His main research area is Pattern Recognition.