IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 February 2024, accepted 11 March 2024, date of publication 13 March 2024, date of current version 28 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3377220

RESEARCH ARTICLE

Integrating Pretrained Encoders for Generalized Face Frontalization

WONYOUNG CHOI[1], GI PYO NAM[2], JUNGHYUN CHO[2], (Member, IEEE), IG-JAE KIM[2], (Member, IEEE), AND HYEONG-SEOK KO[1]

[1]Seoul National University, Seoul 08826, Republic of Korea
[2]Korea Institute of Science and Technology, Seoul 02792, Republic of Korea

Corresponding author: Junghyun Cho (jhcho@kist.re.kr)

This work was supported in part by the Research and Development Program for Advanced Integrated-intelligence for Identification (AIID) through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT under Grant 2018M3E3A1057288; in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korean Government through MSIT (Development of 3D Object Identification Technology Robust to Viewpoint Changes) under Grant RS-2023-00227592; in part by Korea Institute of Science and Technology (KIST) Institutional Programs under Project 2E32301; and in part by the Automation and Systems Research Institute (ASRI), Seoul National University.

ABSTRACT In the field of face frontalization, the model obtained by training on a particular dataset often underperforms on other datasets. This paper presents the Pre-trained Feature Transformation GAN (PFT-GAN), which is designed to fully utilize diverse facial feature information available from pre-trained face recognition networks. For that purpose, we propose the use of the feature attention transformation (FAT) module that effectively transfers the low-level facial features to the facial generator. On the other hand, in the hope of reducing the pre-trained encoder dependency, we attempt a new FAT module organization that accommodates the features from all pre-trained face recognition networks employed. This paper attempts evaluating the proposed work using the ''independent critic'' as well as ''dependent critic'', which enables objective judgments. Experimental results show that the proposed method significantly improves the face frontalization performance and helps overcome the bias associated with each pre-trained face recognition network employed.

INDEX TERMS Face frontalization, face pose normalization, face recognition, generative modeling.

I. INTRODUCTION

Face frontalization is the process of transforming a given face image in non-frontal view to one in frontal view, which is needed from various computer vision tasks. For instance, when an individual has to be identified from images or videos which are often taken non-frontally, face frontalization can assist the duty or, in some cases, can be used to enhance the accuracy of the recognition programs.

Recently, improvements have been made continuously in face frontalization, for which deep learning techniques such as generative adversarial networks (GANs) [1] have been instrumental. They generate the frontal face image from the given profile face image using an encoder-decoder network, similarly to the image transformation technique of [2]. Typically, the face frontalization network is trained employing the supervised learning approaches [3], [4], [5], [6], [7], in which the model is successively updated to minimize the loss function (consisting of the pixel-wise and perceptual losses), ultimately enabling the generation of realistic frontal face images. The model obtained by training on a particular dataset often underperforms on other datasets [3]. It is because the encoder is not robustly trained to cover various identities, angles, and environments.

An alternative would be to use a pre-trained encoder. The face normalization model (FNM) [8] employs pre-trained face-expert networks such as LightCNN [9], VGGFace1 [10], and VGGFace2 [11] as the encoder to extract robust identity features that are invariant to poses, lighting condition, etc. However, with the identity-preserving loss function used in the FNM, which is the perceptual loss between the input profile and output generated face images, the frontalization

The associate editor coordinating the review of this manuscript and approving it for publication was Turgay Celik.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

performance is affected by the pre-trained face recognition network used as the encoder.

To address the above challenge, we propose the Pre-trained Feature Transformation GAN (PFT-GAN). It seeks to fully utilize rich and diverse facial feature information available from pre-trained face recognition networks. For that purpose PFT-GAN develops the feature attention transformation (FAT) module, which brings in the low-level features from the intermediary layers of the face recognition networks into the decoder. Those low-level features contain significant geometric attributes of the input profile face image, but can be difficult to process with a conventional decoder that is composed of convolutional neural networks (CNNs).

We note that the facial features extracted by pre-trained face recognition networks vary based on the network's structure, training dataset, and loss function [12]. For instance, ArcFace [13] has shown superior performance compared to FaceNet [14] in extracting facial features from individuals with wavy hair. In response to this observation, we expand the FAT module such that the decoder takes the result of integrating the features from multiple pre-trained face recognition networks. To our knowledge, this is a new attempt, which turns out effectively reduce the pre-trained encoder dependency in face frontalization.

So-called the "dependent critic" [15] evaluates the model with the same pre-trained face recognition network (say N) that was used during the training, thus it is difficult to discern the biases introduced by N. Since various studies on face frontalization use evaluation metrics based on their own recognition models, making objective comparisons between them can be tricky. To address the above issues, for evaluating the proposed work, we use the "independent critic" [15] as well, which uses a third-party recognition model that was not involved in the training process. The experimental results (in particular Table 4 in Section IV) show that the proposed frontalization model is *objectively* superior.

The main contributions of this study are summarized as follows: (1) we introduce a new framework that integrates multiple pre-trained face recognition networks effectively for the task of face frontalization, and (2) we suggest utilizing the independent critic for objective performance assessment of face frontalization.

II. RELATED WORK

A. GENERATIVE ADVERSARIAL NETWORK

The GAN, proposed in [1], involves a generator that produces synthetic data and a discriminator that evaluates the authenticity of the data. The generator and discriminator are trained adversarially to improve the quality of the generated data. Deep Convolutional GAN [16] that was proposed as an extension of the original GAN uses convolutional layers with strides and transposed convolutional layers with fractional strides. CycleGAN [2] enables the unsupervised image-to-image translation between two domains without requiring paired training data. To more reliably measure the difference between the real and synthetic data distributions, the concept of a WGAN was introduced in [17] and [18], proposing a gradient penalty term to ensure that the discriminator's gradient is bounded. As the GAN technology continues to evolve, face frontalization techniques are becoming increasingly sophisticated and useful.

B. DEEP LEARNING BASED FACE FRONTALIZATION

Deep learning-based face frontalization has seen significant progress in recent years, with many notable works proposed to enhance its performance and stability. TP-GAN [3] considers both global structures and local details for photorealistic frontal view synthesis. It effectively addresses the ill-posed nature of the problem through a combination of adversarial, symmetry, and identity-preserving losses, outperforming state-of-the-art methods in large-pose face recognition tasks. PIM [7], an advancement beyond TPGAN, has achieved enhanced generalizability and reduced overfitting through cross-domain adversarial training. HF-PIM [19] has successfully incorporated 3D face UV map and warping process into the GAN framework, resulting in the generation of high-quality face frontalization. DA-GAN [20] enhances face recognition performance through the integration of self-attention mechanisms in the generator and face-attention mechanisms in the discriminator. IPM [21] utilizes a contrastive loss function for the encoder to extract compact and relationship-preserving representations from input faces. A cross-domain rectification module is introduced to reduce representation discrepancies between recognition and reconstruction domains, enhancing the accuracy of reconstructed faces. It outperforms state-of-the-art methods in extensive experiments on benchmark datasets, demonstrating its effectiveness in handling images from uncontrolled scenes with high fidelity.

Unlike above researches, which are supervised approaches, there are also research efforts aiming to achieve generalization in face frontalization through unsupervised learning. FNM [8] utilizes a well-designed GAN with a face-expert network, pixel-wise loss, and face attention discriminators. This approach generates photorealistic, frontal, and neutral expression face images for face recognition, thereby improving recognition performance on both controlled and in-the-wild databases. Rotate-and-Render [22] introduced an unsupervised framework, which demonstrates the ability to synthesize photorealistic rotated faces from single-view images, effectively overcoming the challenges posed by the lack of high-quality paired training data in facial image processing. DRCycleGAN [23] employs disentangled representation learning and semantic-level cycle consistency loss for face frontalization without paired training data, achieving promising results and enhancing pose-invariant face recognition performance.

C. PRE-TRAINED FACE RECOGNITION NETWORK IN FACE FRONTALIZATION

In the field of face frontalization, extensive research has focused on utilizing pre-trained face recognition networks

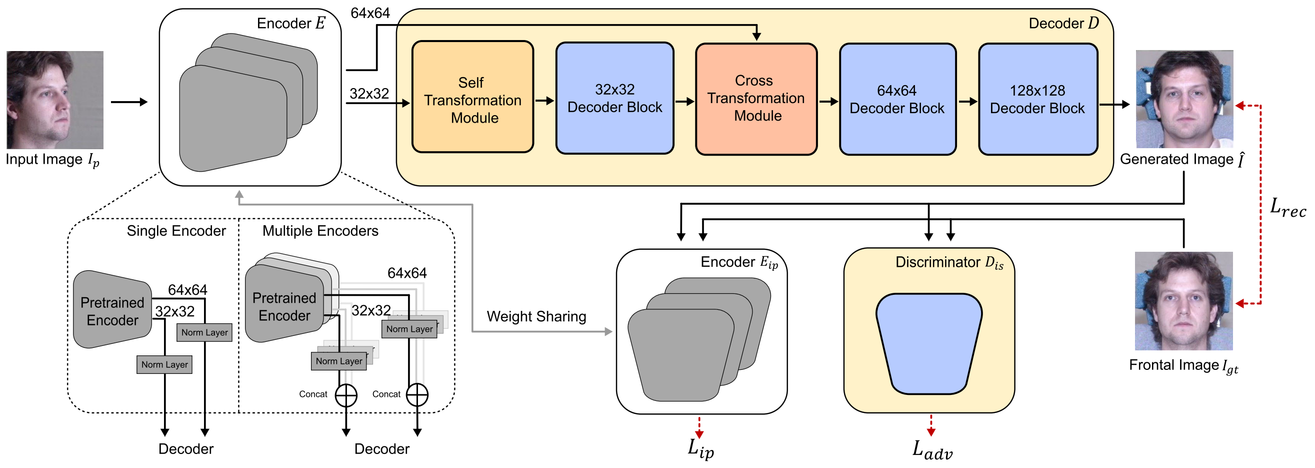


FIGURE 1. Overall network architecture of our proposed method, PFT-GAN. The main part of PFT-GAN can be broadly divided into two parts: the encoder *E* and the decoder *D*. The encoder *E* consists of one or several pre-trained face recognition networks, and the features output by each network are passed through a normalization layer before being concatenated. The features from the encoder *E* are then fed into the decoder *D* to reconstruct the frontal face. The dashed red lines represent the loss functions. L_{rec}, L_{adv}, and L_{ip} correspond to reconstruction loss, adversarial loss, and identity preserving loss, respectively. Detailed explanations of these losses are provided in Section III. The gray-colored networks are frozen during the training phase.

in the learning process. In TP-GAN, inspired by perceptual loss, a pre-trained LightCNN network [9] was integrated into the training process using identity-preserving loss. Subsequently, many face frontalization studies have applied feature information from pre-trained face recognition networks to the training loss function. This approach distills information from large-scale datasets into the frontalization generator. In FNM, for unsupervised learning, the frontalization generator employs an encoder based on a pre-trained face recognition network. Hence, although FNM produces qualitatively robust output images compared to previous methods, its quantitative performance is dependent on the performance of the pre-trained face recognition network used as the encoder. In PM-GAN [24], there is feature fusion between a trainable encoder and a pre-trained face recognition network was employed to enrich the diversity of the extracted features.

Previous methods have focused on implicitly distilling information from pre-trained face recognition networks in the form of loss or extracting the enrich features from the encoder. In this paper, we propose a method that directly involves low-level features as well as high-level features from pre-trained face recognition networks in the training of the frontalization generator to enhance the performance of face frontalization.

III. METHODOLOGY

In this section, we detail our novel approach model. The overall network architecture is depicted in Fig. 1. Initially, we present the pre-trained feature transformation module. Subsequently, we discuss the integration of multiple pre-trained face recognition networks. The details of the loss function and network will be introduced toward the end of this section.

A. NETWORK ARCHITECTURE OF PFT-GAN

PFT-GAN consists of four main components: an encoder $\boldsymbol{E}$ for extracting the key features of a face, a decoder $\boldsymbol{D}$ for reconstructing the frontal face from the key features of a face, a discriminator $\boldsymbol{D_{is}}$ for adversarial learning, and an identity feature encoder $\boldsymbol{E_{ip}}$ for identity preserving loss.

Firstly, the encoder $\boldsymbol{E}$ for extracting the key features of a face is composed of one or more pretrained face recognition encoders. The feature maps extracted from the encoder $\boldsymbol{E}$ pass through normalization layers before entering the decoder $\boldsymbol{D}$. If multiple pretrained face recognition encoders are used, the feature maps are combined before entering the decoder $\boldsymbol{D}$. We refer to the feature map that first enters the decoder $\boldsymbol{D}$ (e.g., the 32×32 feature map in Fig. 1) as the "seed feature" and the other feature maps (e.g., the 64×64 feature map in Fig. 1) as "non-seed features." Secondly, the decoder $\boldsymbol{D}$ for reconstructing the frontal face is comprised of a Feature Attention Transformation module (FAT module) and decoder blocks. The FAT module, a module we proposed, includes spatial attention and is effective in processing the feature maps coming from the encoder $\boldsymbol{E}$. More detailed workings and explanations are provided in the following sections. The decoder blocks consist of several ResNet blocks. Thirdly, the discriminator $\boldsymbol{D_{is}}$ for adversarial learning is a CNN network that takes a frontal face as input and outputs whether it's real or fake. We referenced the architecture from StarGAN v2 [25]. Lastly, the identity feature encoder $\boldsymbol{E_{ip}}$ is a network designed to ensure identity preservation by extracting identity features from both the generated frontal face and the real frontal face and using a distance loss function between them. The identity feature encoder $\boldsymbol{E_{ip}}$ uses the same encoder as the encoder $\boldsymbol{E}$ for extracting the key features.

B. FEATURE ATTENTION TRANSFORMATION MODULE

The feature transformation in our framework can be categorized into two types: the self-transformation and the cross-transformation.

1) SELF-TRANSFORMATION MODULE

The self-transformation module is responsible for transforming the seed feature map. For a given seed feature map $\boldsymbol{F}_s \in \mathbb{R}^{\mathbb{W}\times\mathbb{H}\times\mathbb{C}}$, to facilitate the smooth generation of a frontal face, the vertically flipped seed feature map $\boldsymbol{F}_s^{flip} \in \mathbb{R}^{\mathbb{W}\times\mathbb{H}\times\mathbb{C}}$, is concatenated and passed through a self-attention module. This module is a traditional self-attention component of a transformer. The self-transformation, which computes the transformed seed feature $\boldsymbol{F}_s^{self}$, can be summarized by (1):

$$\begin{aligned}\boldsymbol{F}_s^{self} &= \text{Attention}(\boldsymbol{F}_s^c\boldsymbol{W}_Q, \boldsymbol{F}_s^c\boldsymbol{W}_K, \boldsymbol{F}_s^c\boldsymbol{W}_V)\\ \boldsymbol{F}_s^c &= \text{Concat}(\boldsymbol{F}_s, \boldsymbol{F}_s^{flip})\end{aligned} \tag{1}$$

2) CROSS-TRANSFORMATION MODULE

The cross-transformation module supplements the feature restoring the frontal face with additional information derived from the low-level feature of the profile face, which are obtained from the pre-trained face recognition network. As illustrated in Fig. 2, this module, distinct from the aforementioned self-transformation module, treats features from the previous decoder block solely as a query, with the low-level features from the pre-trained face recognition network serving as both key and value.

$$\begin{aligned}\boldsymbol{F}_{i+1}^{cross} &= \text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})\\ \boldsymbol{Q} &= \text{UpsampleResblk}(\boldsymbol{F}_i^{cross}\boldsymbol{W}_Q),\\ \boldsymbol{K} &= (\boldsymbol{F}_{i+1}^{pre}\boldsymbol{W}_K),\\ \boldsymbol{V} &= (\boldsymbol{F}_{i+1}^{pre}\boldsymbol{W}_V)\end{aligned} \tag{2}$$

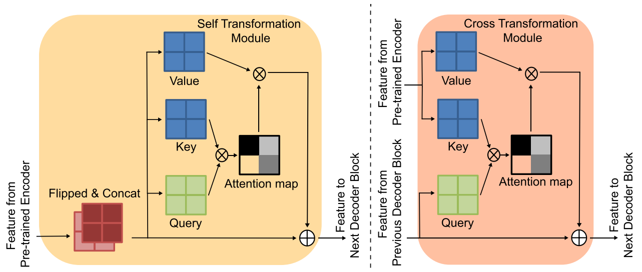

Self Transformation Module
Value
Key
Query
Attention map
Flipped & Concat
Feature from Pre-trained Encoder
Feature to Next Decoder Block
Feature from Pre-trained Encoder
Feature from Previous Decoder Block
Cross Transformation Module
Value
Key
Query
Attention map
Feature to Next Decoder Block

FIGURE 2. Details of FAT modules. The left figure presents a self-transformation module architecture and the right a cross-transformation module architecture.

3) CROSS-TRANSFORMATION MODULE FOR MULTIPLE ENCODERS

We connected multiple cross-transformation modules in parallel to integrate the features obtained from various pre-trained face recognition networks. The aligned features generated by each cross-transformation module were then combined using an element-wise sum and passed on to the subsequent decoder block. Fig. 3 illustrates how the cross-transformation modules function when two pre-trained face recognition networks are integrated into the architecture.

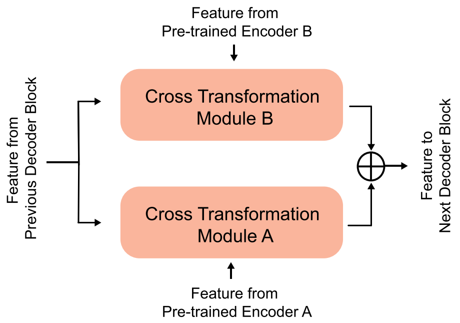

Feature from Pre-trained Encoder B
Cross Transformation Module B
Cross Transformation Module A
Feature from Pre-trained Encoder A
Feature from Previous Decoder Block
Feature to Next Decoder Block

FIGURE 3. Example of multi-cross-transformation module for integrating two pre-trained encoders.

C. LOSS FUNCTION

Similar to previous face frontalization methods, PFT-GAN was trained using a combination of three loss functions: reconstruction loss, adversarial loss, and identity-preserving loss.

1) RECONSTRUCTION LOSS

The reconstruction loss $\boldsymbol{L}_{rec}$ encourages the network to generate face images $\hat{\boldsymbol{I}}$ that are identical to the ground-truth images $\boldsymbol{I}_{gt}$ at the pixel level. It is defined as follows:

$$\boldsymbol{L}_{rec} = \left\|\hat{\boldsymbol{I}} - \boldsymbol{I}_{gt}\right\|_1 \tag{3}$$

where $\|\cdot\|_1$ denotes the L1 norm.

2) ADVERSARIAL LOSS

By incorporating the adversarial loss $\boldsymbol{L}_{adv}$, the decoder $\boldsymbol{D}$ is encouraged to produce more realistic frontalized face images that capture both the appearance and structural characteristics of frontal faces. The discriminator $\boldsymbol{D}_{is}$ provides feedback to the decoder $\boldsymbol{D}$, thereby enhancing the quality and realism of the generated frontalized images. The adversarial loss is denoted by the following expression:

$$\boldsymbol{L}_{adv} = \mathbb{E}_{I_{gt}}\left[logD_{is}(x)\right] + \mathbb{E}_{\hat{I}}\left[1 - logD_{is}(x)\right] \tag{4}$$

We employ the vanilla GAN loss function [1] to assess the performance of our model. Despite recent advancements in GANs, such as WGAN [17], [18], which show superior performance, using them for our purposes presents no issues.

3) IDENTITY PRESERVING LOSS

Identity preserving loss is a critical component in face frontalization, aiming to maintain the characteristics of the input face during the frontalization process. This loss function works by minimizing the distance between the features extracted from the generated face and those of the corresponding ground-truth face image using a pre-trained face recognition network. By reducing this distance, the model is encouraged to preserve the facial features and identity of the input face, thereby ensuring that the generated face image retains the essential characteristics of the original. Identity-preserving loss can be expressed as follows:

$$\boldsymbol{L}_{ip} = \left\|\phi_0(\hat{\boldsymbol{I}}) - \phi_0(\boldsymbol{I}_{gt})\right\|_2 + \left\|\phi_1(\hat{\boldsymbol{I}}) - \phi_1(\boldsymbol{I}_{gt})\right\|_2 \tag{5}$$

where, $\phi_0(\cdot)$ and $\phi_1(\cdot)$ represent features extracted from the last layer and the preceding layers, respectively, of a pre-trained face recognition network used as an encoder $\boldsymbol{E}$ and $\|\cdot\|_2$ denotes the L2 norm.

4) OVERALL LOSS

The final loss function was the weighted sum of the three aforementioned loss functions. This can be formulated as follows:

$$\boldsymbol{L}_{total} = \lambda_{rec}\boldsymbol{L}_{rec} + \lambda_{ip}\boldsymbol{L}_{ip} + \lambda_{adv}\boldsymbol{L}_{adv} \tag{6}$$

IV. EXPERIMENTAL RESULT

In this section, we present the dataset and implementation details.

A. EXPERIMENTAL SETTINGS

1) DATASETS

We used the CMU Multi-PIE face dataset [26] for our training and testing set. The Multi-PIE dataset comprises over 750,000 images of 337 individuals, captured under 15 viewpoints and 19 illumination conditions, showcasing a variety of facial expressions. This dataset is widely used for evaluating face synthesis and recognition in controlled settings. In line with previous face frontalization studies, we followed setting 2 for evaluating our model. Setting 2 involves using neutral expression images from all four sessions, encompassing 337 identities. For training, we used images of the first 200 identities across 11 poses. For testing, a frontal view image under normal illumination was selected as the gallery image for each of the remaining 137 identities, while the other images served as probes.

2) IMPLEMENTATION DETAILS

The face images selected for training and testing were aligned using the MTCNN face detector [27] and then cropped to a resolution of 128 × 128 pixels. In the decoder $\boldsymbol{D}$, each block comprises six ResNet blocks and an upsampled layer. Our network was implemented using PyTorch. For training, Adam optimizer was used. The hyperparameters employed for training included: $lr = 10^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.99$, $\lambda_{rec} = 1$, $\lambda_{ip} = 1$, $\lambda_{adv} = 0.1$.

B. EVALUATION METRICS

Akin to previous studies, we assessed the face frontalization performance using the rank-1 recognition rate. It is calculated by measuring the cosine distance between the feature vectors extracted from the generated frontal faces and the gallery images of the corresponding identities, using a pre-trained face recognition network. Typically, previous studies typically used the same face recognition network for both the training stage and computing the rank-1 recognition rate. However, using the same network for evaluation introduces subjectivity and biases the results toward that specific network. As demonstrated in Table 1, when LightCNN was used as the critic encoder during training, the module using it achieved the highest frontalization performance metric. In contrast, when VGGFace2 was used as the critic encoder, the module employing it exhibited superior performance in terms of frontalization metrics. Therefore, the selection of the evaluation encoder significantly impacts the performance metrics of the frontalization module.

TABLE 1. Rank-1 recognition rate(%) for frontalized face images from 90° profile face images using different pre-trained face recognition networks.

Train \ Critic	LightCNN	VGGFace1	VGGFace2
LightCNN	**72.76**	67.56	73.03
VGGFace1	58.65	65.01	63.52
VGGFace2	68.34	**69.32**	**77.38**

To objectively assess the performance metrics of the frontalization module, we distinguished between two existing measurement approaches: dependent and independent critics. In terms of the facial recognition networks used for training, we selected networks commonly utilized in facial frontalization research, including LightCNN, VGGFace1, and VGGFace2. The LightCNN used in our study is the LightCNN-9 layers network trained on the CASIA-WebFace [28] and MS-Celeb-1M [29] datasets. (Note that LightCNN⋆ and LightCNN† used in previous papers are the LightCNN-29 layers networks fine-tuned on Multi-PIE dataset after being trained on MS-Celeb-1M dataset). Additionally, we adapted the ArcFace network, which is IResNet50 architecture [30] and utilizes additive angular margin loss [13], trained with Glint360K [31] dataset. For the independent critic measurement, we employed the ArcFace⋆ network, which is same architecture with the ArcFace network, trained with MS-Celeb-1M dataset and the FaceNet [14] network. The baseline performances of these face recognition networks are detailed in Table 2.

TABLE 2. Rank-1 recognition rate (%) performance of pre-trained face recognition networks on MultiPIE setting 2. Details of each encoder are explained in Section IV-B.

Encoder	±0°	±15°	±30°	±45°	±60°	±75°	±90°
LightCNN⋆	-	98.59	97.38	92.13	62.09	24.18	5.51
LightCNN†	-	99.10	98.60	97.70	91.40	68.70	27.10
LightCNN	99.87	99.83	99.53	96.38	78.75	42.10	8.31
VGGFace1	98.78	98.52	96.51	92.33	80.70	60.59	30.73
VGGFace2	100	100	100	99.91	99.06	95.73	86.82
ArcFace	100	100	100	100	99.93	99.71	92.61
ArcFace⋆	100	100	100	99.97	99.27	95.89	53.03
FaceNet	93.98	92.33	87.37	75.00	52.24	31.53	13.85

C. THE QUANTITATIVE RESULTS OF VARIOUS PRE-TRAINED MODELS

We assessed the performance of modules constructed using various pre-trained face recognition networks in both single

TABLE 3. Rank-1 recognition rate (%) performance of dependent critic with various pre-trained face recognition networks on MultiPIE setting 2.

Method (Training Encoder)	Critic	±0°	±15°	±30°	±45°	±60°	±75°	±90°
TP-GAN (LightCNN*) [3]	LightCNN*	-	98.68	98.06	95.38	87.72	77.43	64.64
PIM (LightCNN†) [7]	LightCNN†	-	99.30	99.00	98.50	98.10	95.00	86.50
HF-PIM (LightCNN*) [19]	LightCNN*	-	***99.99***	99.98	99.88	99.14	96.40	92.32
DRCycleGAN (LightCNN*) [23]	LightCNN*	-	99.40	98.70	96.60	93.40	83.60	66.80
CAPG-GAN (LightCNN*) [6]	LightCNN*	-	99.82	99.56	97.33	90.63	83.05	66.05
DA-GAN (LightCNN*) [20]	LightCNN*	-	99.98	99.88	99.15	97.27	93.24	81.56
IPM (LightCNN*) [21]	LightCNN*	-	**100**	**100**	99.67	***99.50***	97.42	**93.83**
FNM (VGGFace2) [8]	VGGFace2	99.70	99.60	98.70	97.10	92.40	80.40	62.00
PFT-GAN (LightCNN)	LightCNN	99.92	99.55	98.85	96.61	91.37	84.62	72.76
PFT-GAN (VGGFace1)	VGGFace1	98.15	96.95	94.34	90.14	83.24	76.08	65.01
PFT-GAN (VGGFace2)	VGGFace2	***99.97***	99.94	99.62	97.59	92.18	87.07	77.38
PFT-GAN (ArcFace)	ArcFace	**100**	**100**	***99.99***	***99.98***	**99.56**	**98.27**	92.69
PFT-GAN (LightCNN, VGGFace1)	LightCNN	99.78	99.73	99.51	97.64	94.38	88.18	79.77
	VGGFace1	98.30	97.76	96.46	93.73	88.26	81.13	74.31
PFT-GAN (LightCNN, VGGFace2)	LightCNN	99.63	99.57	98.56	96.72	91.69	85.12	75.15
	VGGFace2	***99.97***	99.89	99.38	97.53	91.97	86.68	77.77
PFT-GAN (LightCNN, ArcFace)	LightCNN	99.80	99.84	99.76	99.28	96.58	92.08	84.99
	ArcFace	**100**	**100**	**100**	**99.99**	99.48	***98.04***	***92.83***

TABLE 4. Rank-1 recognition rate (%) performance of independent critic with various pre-trained face recognition networks on MultiPIE setting 2.

Method (Training Encoder)	Critic	±0°	±15°	±30°	±45°	±60°	±75°	±90°
FNM (VGGFace2) [8]	FaceNet	86.13	84.4	81.81	78.5	69.63	59.96	43.79
PFT-GAN (LightCNN)	FaceNet	92.45	90.25	87.46	81.60	74.88	67.75	57.17
PFT-GAN (VGGFace1)	FaceNet	90.73	89.87	86.02	80.08	71.84	62.59	52.82
PFT-GAN (VGGFace2)	FaceNet	91.95	91.97	90.43	85.02	79.95	74.22	60.72
PFT-GAN (ArcFace)	FaceNet	***93.13***	***93.15***	***91.84***	***88.93***	***83.73***	***77.78***	***67.35***
PFT-GAN (LightCNN, VGGFace1)	FaceNet	92.55	91.22	89.60	84.44	76.98	69.62	61.85
PFT-GAN (LightCNN, VGGFace2)	FaceNet	92.48	91.87	89.28	85.43	79.62	73.67	64.58
PFT-GAN (LightCNN, ArcFace)	FaceNet	**93.18**	**93.25**	**92.39**	**89.98**	**84.96**	**78.77**	**71.57**
FNM (VGGFace2) [8]	ArcFace*	98.50	97.94	96.30	92.37	82.92	66.88	44.76
PFT-GAN (LightCNN)	ArcFace*	***99.98***	99.85	99.05	96.73	89.35	81.64	67.09
PFT-GAN (VGGFace1)	ArcFace*	**100**	99.76	98.68	94.89	85.88	75.28	57.70
PFT-GAN (VGGFace2)	ArcFace*	**100**	99.93	99.62	97.52	92.10	85.18	70.47
PFT-GAN (ArcFace)	ArcFace*	**100**	***99.97***	**99.97**	***99.67***	***98.38***	***95.85***	***88.13***
PFT-GAN (LightCNN, VGGFace1)	ArcFace*	**100**	99.96	99.76	98.75	94.33	87.94	75.64
PFT-GAN (LightCNN, VGGFace2)	ArcFace*	***99.98***	99.93	99.64	98.53	94.46	88.70	76.32
PFT-GAN (LightCNN, ArcFace)	ArcFace*	**100**	**100**	***99.95***	**99.92**	**98.51**	**96.19**	**90.52**

and integrated configurations, based on the dependent critic. As depicted in Table 3, our proposed method exhibits high performance even at extreme angles such as 75° or 90°. Notably, when employing LightCNN or VGGFace1 with the PFT-GAN, the performance exceeded that of the original pre-trained face recognition networks. For instance, the performance of LightCNN at 90° improved from 8.31% to 72.76%. Among the configurations of our proposed approach, using ArcFace alone or integrating ArcFace and LightCNN into the module yielded the most effective results. The performance at 90° exceeded 92%, surpassing that of previous state-of-the-art face frontalization methods.

The performance evaluation based on the dependent critic, as mentioned earlier, may not provide an entirely objective assessment. Therefore, we also compared the performance using an independent critic, which evaluates all the models using the same third-party pre-trained face recognition network. As illustrated in Table 4, it is evident that integrating multiple pre-trained face recognition networks demonstrates a superior ability in accurately restoring frontal faces compared to using a single pre-trained network.

Similar to the dependent critic cases, the results show that frontalization significantly enhances facial recognition performance. For instance, our method using ArcFace improves

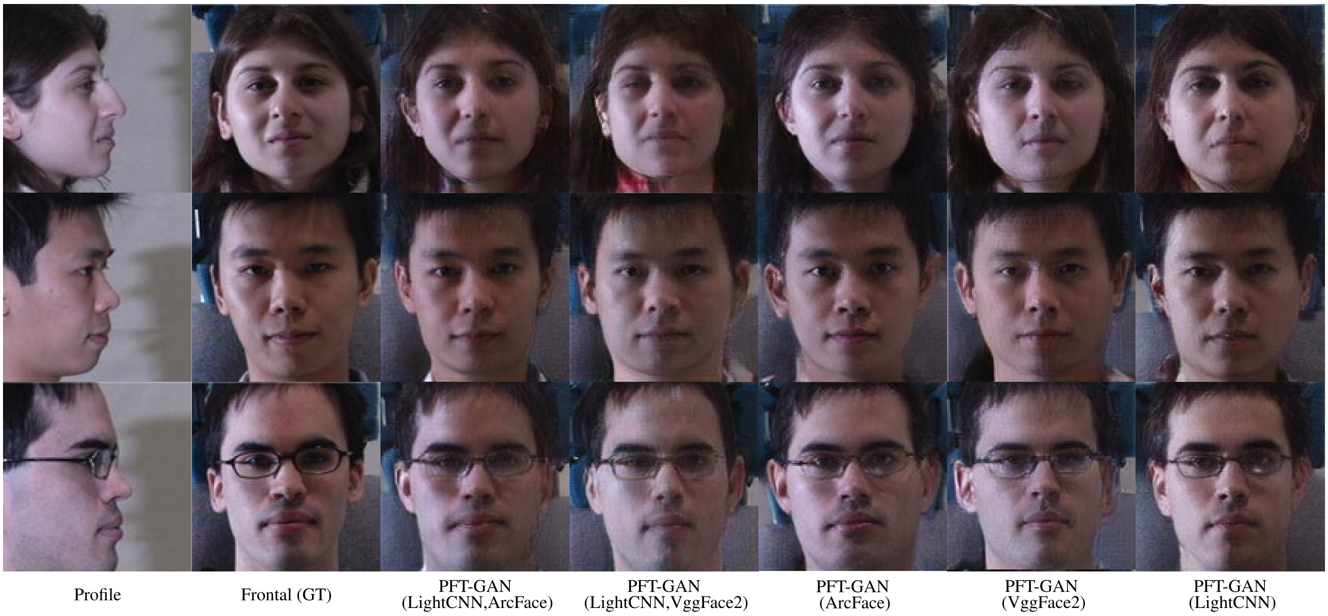


FIGURE 4. The frontalization results produced by variations of PFT-GAN on the extreme pose.

the performances of FaceNet and ArcFace* at 90° from 13.85% and 53.03% to 67.35% and 88.13% respectively. Furthermore, although the performance difference between the module using ArcFace alone and the module integrating ArcFace with LightCNN was minimal when evaluated by the dependent critic, a notable improvement was observed in the both independent critic evaluations (e.g., an increase from 67.35% to 71.57% at 90° with FaceNet critic and 88.13% to 90.52% at 90° with ArcFace⋆ critic).

In conclusion, the traditional approach in facial frontalization research using a dependent critic, which relies on various face recognition networks as benchmarks, poses challenges for objective comparisons. However, by assessing performance using the same face recognition network as the independent critic, we were able to confirm the significant contribution of our proposed approach in restoring frontal faces.

D. QUALITATIVE RESULTS

We demonstrate that our proposed method effectively maintains identities in the Multi-PIE test set, which is similar the training data environment in Fig. 4. Notably, when combining high-performance networks such as LightCNN and ArcFace, the results at an extreme angle over 75° bear a remarkable resemblance to real frontal photos.

We also compare our proposed PFT-GAN with several state-of-art methods, including IPM [21], DA-GAN [20], FNM [8], CAPG-GAN [6], and TP-GAN [3] on the Multi-PIE test set in Fig. 5. We note that recent researches on face frontalization often do not share their models or make their code publicly available. Therefore, we are evaluating our results by comparing them with the images provided in those papers. The comparison results clearly show that TP-GAN and CAPG-GAN struggle to preserve the identity of the input face with large facial angles, unlike PTF-GAN, which successfully maintains it. Additionally, PTF-GAN excels in retaining the local features of the input face for small facial angles, enabling the restoration of a frontal face that closely resembles the actual frontal face.

We further highlight out method's superior ability to consistently reconstruct frontal faces in the unconstrained LFW dataset [32], compared to two state-of-the-art methods, FNM and IPM, which have previously exhibited robust performance on the Multi-PIE test set in Fig. 5. The results clearly demonstrate that our approach more effectively retains the facial features (such as expressions and skin tone, etc) of the input face than FNM, which is trained using unsupervised methods. On the other hand, IPM tends to overly preserve the local features of the input face, leading to improperly generated frontal faces that appear distorted. Notably, our PFT-GAN, despite being trained on the constrained Multi-PIE dataset, exhibits robust results on other datasets. This indicates that our model does not overfit to the specific characteristics of the training dataset, showcasing its versatility and effectiveness in diverse scenarios.

E. ABLATION STUDIES

We conduct ablation studies to assess how different components of the proposed method affect the frontalization performance.

1) THE EFFECTS OF FAT MODULE

To evaluate the effectiveness of the FAT module, we constructed two types of networks: one incorporating the FAT module and the other without it. Networks that do not utilize the FAT module add non-seed features, which emerge from the middle of a pre-trained encoder, directly into the decoder after only passing through a normalization layer.

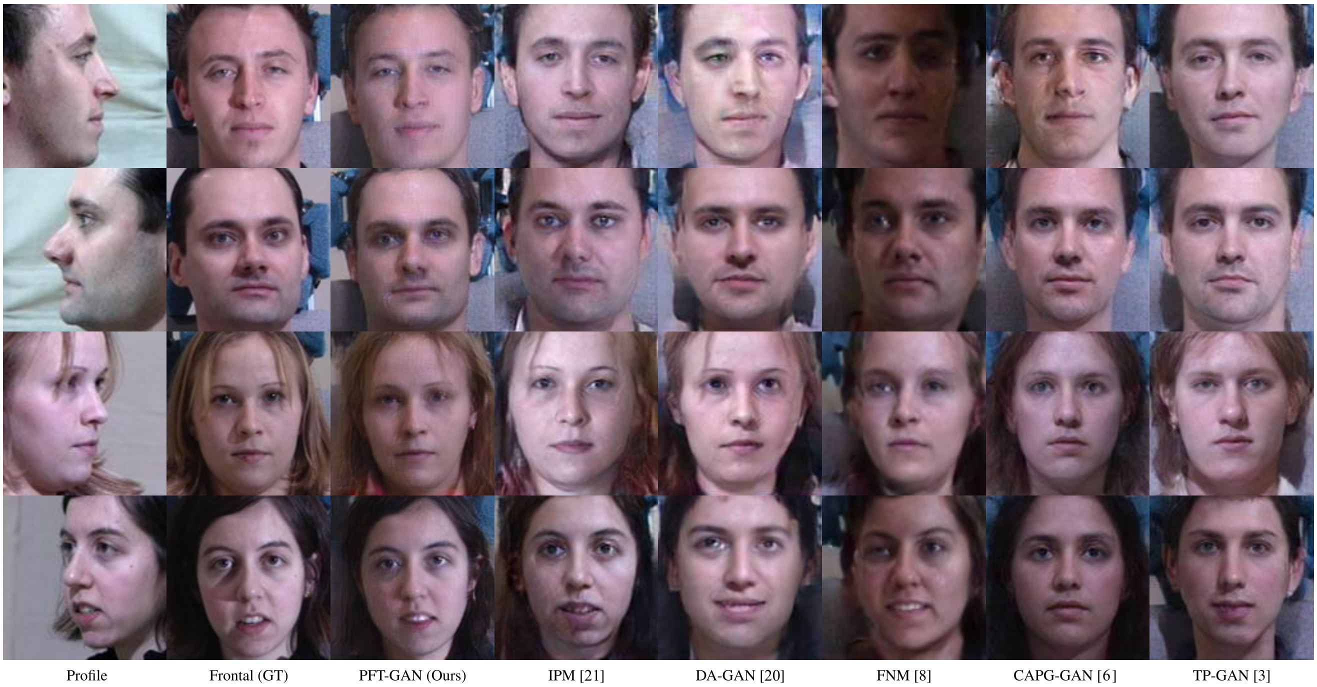
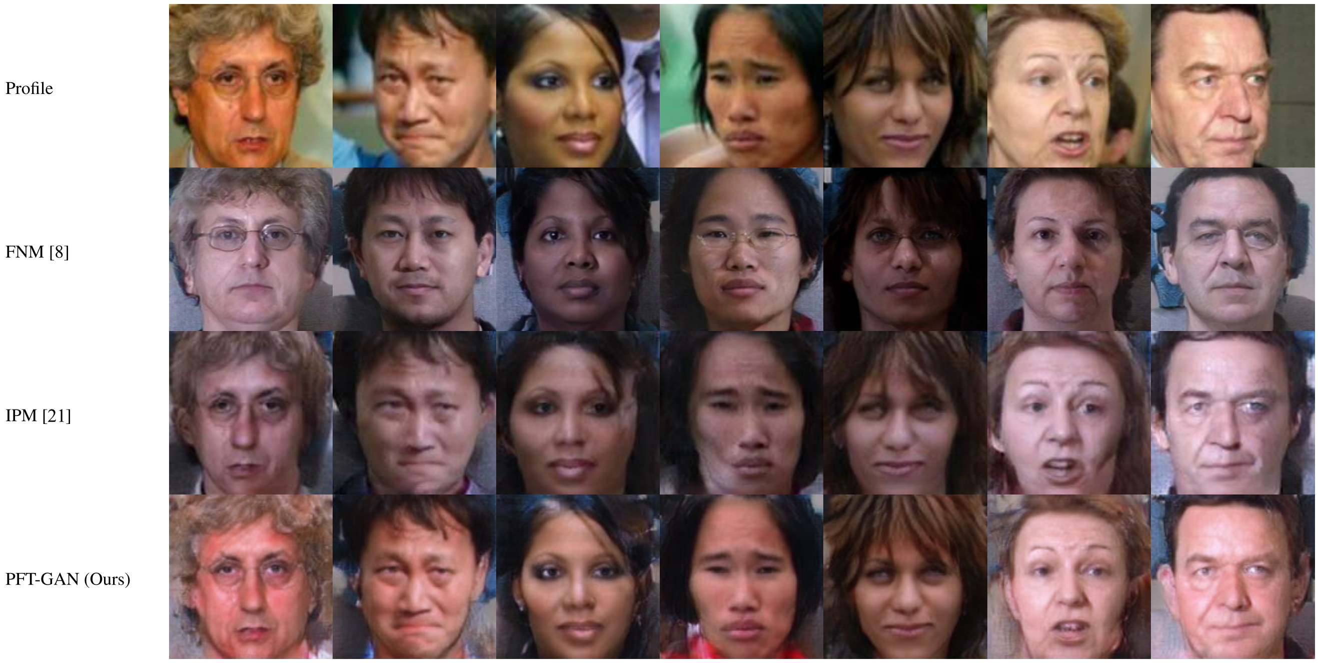


FIGURE 5. The frontalization results from various methods on the Multi-PIE dataset.

TABLE 5. The frontalization results from various methods on the LFW dataset.



Both networks were configured with an identical seed feature size of 32 × 32. The quantitative performance results for these two networks are presented in Table 6. Notably, there was a performance difference of 2.51% at the extreme angle of 90°, illustrating the impact of the FAT module.

The effectiveness of the cross-transformation module, which is a component of the FAT module, is visually presented in Fig. 6. The second row, showing the results from the network without the FAT module, reveals that the facial features were not adequately frontalized in extreme facial pose. In contrast, the third row, representing the results from the network with the FAT module, clearly demonstrates proper face frontalization. This indicates that the FAT module's effectiveness in transforming low-level

TABLE 6. Rank-1 recognition rate (%) performance of FAT module on Multi-PIE setting 2.(pre-trained encoder: LightCNN).

Module	±0°	±15°	±30°	±45°	±60°	±75°	±90°
w/o FAT	99.87	**99.80**	**99.48**	**97.66**	**91.37**	82.78	69.83
with FAT	**99.92**	99.55	98.85	96.61	**91.37**	**84.62**	**72.76**

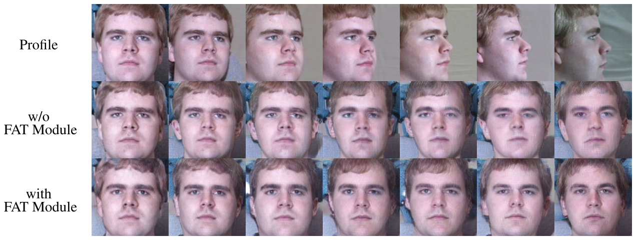


FIGURE 6. The frontalization results at different angles, comparing outcomes with and without FAT modules.

feature maps extracted from the pre-trained face recognition network. Overall, these results underscore the significant role of the FAT module in accurately frontalizing faces and effectively transforming low-level feature maps obtained from pre-trained face recognition networks.

2) THE EFFECTS OF SEED FEATURE SIZE

To determine the optimal seed feature size, modules with varying seed feature dimensions were conducted. For a seed feature size was 8 × 8, the self-transformation module was designed for the 8 × 8 seed feature, accompanied by three cross-transformation modules for 16 × 16, 32 × 32, and 64 × 64 seed features, totaling four transformation modules. Similarly, for seed feature sizes of 16 × 16 and 32 × 32, three and two transformation modules were constructed, respectively.

The findings of these experiments are summarized in Table 7. The results indicated that the performance of the frontalization module was most effective when the seed feature size was 32×32 across all facial angles. This observation led to the conclusion that as features from pre-trained face recognition networks become more high-level, they tend to exhibit bias toward the dataset on which the face recognition network was trained, resulting in information loss through compression. The 32 × 32 seed feature size emerged as the most suitable for facial frontalization training. Consequently, we fixed the seed feature size to 32 × 32 for all subsequent experiments.

TABLE 7. Rank-1 recognition rate (%) performance of seed feature size on Multi-PIE setting 2.(pre-trained encoder: LightCNN).

Seed Feature Map Size	±0°	±15°	±30°	±45°	±60°	±75°	±90°
8×8	36.90	34.44	31.04	27.46	23.09	20.13	16.20
16×16	59.93	57.61	54.26	47.66	40.55	35.18	29.85
32×32	**99.92**	**99.55**	**98.85**	**96.61**	**91.37**	**84.62**	**72.76**

V. CONCLUSION

In this paper, we proposed the PFT-GAN, a new framework for face frontalization that is built on pre-trained face recognition networks. The novelties associated with the PFT-CAN can be summarized as: (1) the use of the FAT module and (2) the integration of multiple pre-trained face recognition networks.

The FAT module is a bifurcation from the traditional autoencoder-based approaches. It adopts attention-based mechanisms to incorporate low-level feature maps from pre-trained face recognition networks, by which the proposed model can produce sharper frontalized results and preserve facial textures better compared to the conventional methods.

Although utilization of individual pre-trained face recognition network has been attempted as in [8], a scheme that can benefit from multiple pre-trained face recognition networks has not be explicitly proposed yet. The proposed integration, although it is a simple concatenation of each recognition network output, significantly enhances the frontalization performance when evaluated *objectively*, and addresses the overfitting problem commonly observed when conventionally employing a pre-trained face recognition network.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2014.

[2] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[3] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2458–2467.

[4] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 113–120.

[5] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Learning a high fidelity pose invariant model for high-resolution face frontalization," in *Advances in Neural Information Processing Systems*, vol. 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2018, pp. 2867–2877.

[6] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8398–8406.

[7] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, S. Yan, and J. Feng, "Towards pose invariant face recognition in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2207–2216.

[8] Y. Qian, W. Deng, and J. Hu, "Unsupervised face normalization with extreme pose and expression in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9843–9850.

[9] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[10] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–12.

[11] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[12] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. M. Moreno, J. Fierrez, and A. Kuijper, "A comprehensive study on face recognition biases beyond demographics," *IEEE Trans. Technol. Soc.*, vol. 3, no. 1, pp. 16–30, Mar. 2022.

[13] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[15] A. Razzhigaev, K. Kireev, E. Kaziakhmedov, N. Tursynbek, and A. Petiushko, "Black-box face recovery from identity features," in *Proc. Comput. Vis.–ECCV Workshops*. Glasgow, U.K., Cham, Switzerland: Springer, 2020, pp. 462–475.

[16] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds. San Juan, Puerto Rico, 2016.

[17] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, D. Precup and Y. W. Teh, Eds. 2017, pp. 214–223.

[18] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017.

[19] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Learning a high fidelity pose invariant model for high-resolution face frontalization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[20] Y. Yin, S. Jiang, J. P. Robinson, and Y. Fu, "Dual-attention GAN for large-pose face frontalization," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 249–256.

[21] J. Xin, Z. Wei, N. Wang, J. Li, X. Wang, and X. Gao, "Learning a high fidelity identity representation for face frontalization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6952–6964, Nov. 2023.

[22] H. Zhou, J. Liu, Z. Liu, Y. Liu, and X. Wang, "Rotate-and-render: Unsupervised photorealistic face rotation from single-view images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5910–5919.

[23] Y. Liu and J. Chen, "Unsupervised face frontalization using disentangled representation-learning CycleGAN," *Comput. Vis. Image Understand.*, vol. 222, Sep. 2022, Art. no. 103526.

[24] S. Cen, H. Luo, J. Huang, W. Shi, and X. Chen, "Pre-trained feature fusion and multidomain identification generative adversarial network for face frontalization," *IEEE Access*, vol. 10, pp. 77872–77882, 2022.

[25] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8185–8194.

[26] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[27] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[28] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*.

[29] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Comput. Vis.–ECCV 2016 14th Eur. Conf.*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 87–102.

[30] I. C. Duta, L. Liu, F. Zhu, and L. Shao, "Improved residual networks for image and video recognition," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 9415–9422.

[31] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, and Y. Fu, "Partial FC: Training 10 million identities on a single machine," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1445–1449.

[32] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. Workshop faces Real-Life Images, Detection, Alignment, Recognit.*, 2008.

WONYOUNG CHOI received the B.S. degree from Seoul National University, Seoul, South Korea, in 2012, where he is currently pursuing the combined master's and Ph.D. degree with the Department of Electrical and Computer Engineering. His research interests include computer graphics, real-time clothing rendering, and deep learning-based image generation.

GI PYO NAM received the B.S. degree in digital media technology from Sangmyung University, Seoul, South Korea, in 2009, and the Ph.D. degree in electronics and electrical engineering from Dongguk University, in 2014. He is currently a Senior Research Scientist with the Center for Artificial Intelligence, Artificial Intelligence and Robotics Institute, Korea Institute of Science and Technology (KIST), Seoul. His research interests include pattern recognition, biometrics, and image processing.

JUNGHYUN CHO (Member, IEEE) received the B.S. degree in industrial design and the M.S. degree in applied mathematics from KAIST, Daejeon, South Korea, in 2002 and 2004, respectively, and the Ph.D. degree in computer graphics from Seoul National University, Seoul, South Korea, in 2013. He has been a Principal Research Scientist with the Center for Artificial Intelligence, Korea Institute of Science and Technology (KIST), Seoul, since 2021. He is currently a Professor with Korea National University of Science and Technology (UST) and an Adjunct Professor with Yonsei University. His research interests include computer graphics, computer vision, and deep learning, especially for domain adaptation.

IG-JAE KIM (Member, IEEE) received the B.S. and M.S. degrees in EE from Yonsei University, Seoul, South Korea, in 1996 and 1998, respectively, and the Ph.D. degree in EECS from Seoul National University, in 2009. He was a Postdoctoral Researcher with the Massachusetts Institute of Technology (MIT) Media Laboratory, from 2009 to 2010. He is currently the Director of the Artificial Intelligence and Robotics Institute, Korea Institute of Science and Technology (KIST), Seoul. He is also an Associate Professor with Korea National University of Science and Technology (UST), a Guest Professor with Korea University, and an Adjunct Professor with Yonsei University. He has published over 100 fully-refereed papers in international journals and conferences, including *ACM Transaction on Graphics*, *Pattern Recognition*, CVPR, SIGGRAPH, and Eurographics. His research interests include pattern recognition, computer vision and graphics, deep learning, and computational photography.

HYEONG-SEOK KO received the B.S. and M.S. degrees in computer science from Seoul National University (SNU), South Korea, in 1985 and 1987, respectively, and the Ph.D. degree in computer science from the University of Pennsylvania, in 1994. Since 1996, he has been with the Department of Electrical and Computer Engineering, SNU, where he is currently a Professor. His research interests include clothing simulation and rendering.