Article:
Zhang, Z., Liang, R., Chen, X. et al. (4 more authors) (2020) Semi-Supervised Face Frontalization in the Wild. Information Forensics and Security, IEEE Transactions on. ISSN 1556-6013

https://doi.org/10.1109/TIFS.2020.3025412

Semi-Supervised Face Frontalization in the Wild

Zhihong Zhang, Ruiyang Liang, Xu Chen, Xuexin Xu, Guosheng Hu, Wangmeng Zuo, Edwin R. Hancock, *Fellow, IEEE*

***Abstract*—Synthesizing a frontal view face from a single nonfrontal image, i.e. face frontalization, is a task of practical importance in a wide range of facial image analysis applications. However, to train the frontalization model in a supervised manner, most existing face frontalization methods rely on the availability of nonfrontal-frontal face pairs (typically from the Multi-PIE dataset) captured in a constrained environment. Such approaches, in return, limit the generalizability of their application to unconstrained scenarios. Unfortunately, although a large amount of in-the-wild face datasets are available, they cannot easily be utilized for face frontalization training since the nonfrontal and frontal facial images are not paired. To train a frontalization network which generalizes well to both constrained and unconstrained environments, we propose a *semi-supervised* learning framework which effectively uses both (labeled) indoor and (unlabeled) outdoor faces. Specifically, to achieve this goal, this paper presents a Cycle-Consistent Face Frontalization Generative Adversarial Network (CCFF-GAN) which consists of both (1) the supervised and (2) the unsupervised components. For (1), we use the indoor paired (labeled) data to learn a roughly accurate frontalization network which may not generalize well to outdoor (in-the-wild) scenarios. For (2), to cope with the generalization issue, the unsupervised part uses the unpaired (unlabeled) images under the perceptual cycle consistency constraint in the semantic feature space to generalize the network from controlled (indoor) to uncontrolled (outdoor) environment. Extensive experiments demonstrate the effectiveness of the proposed method in comparison with the state-of-the-art face frontalization methods, especially under the in-the-wild scenarios.**

***Index Terms*—Face frontalization, face synthesis, face recognition**

I. Introduction

DEEP learning (DL) has proven to be a powerful tool in a wide range of face analysis and recognition tasks [1], [2]. However, accurately recognizing faces in unconstrained environments is still challenging for several reasons. Specifically, large pose variations are one of the major factors that significantly reduce the performance of face recognition algorithms [3], [4]. *Pose Invariant Face Recognition* (PIFR) has therefore attracted significant attention recently.

To tackle the pose variation problem in face recognition, a variety of studies have been conducted [5], [6], [7], [8], [9]. Those methods can be roughly categorized into two groups. The first of these seeks pose-invariant features to represent the face so that the recognition can be performed using such features to avoid the impact of pose variations [5], [6], [7], [8], [9]. The second group of methods aim to eliminate pose variations by first synthesizing a frontal view of the face from a given nonfrontal image, i.e., face frontalization (FF). Face recognition is then performed using the synthesized frontal face [10], [11], [12], [13], [14], [15], [16], [13], [17], [18], [19]. Compared with invariant feature based methods, the face frontalization methods are intrinsically easier to interpret since they are able to generate a high quality frontal face image from its nonfrontal counterpart. This capability is of particular relevance and practical importance in many applications where the transparency of the decision process is important, such as law enforcement and visually identifying suspects.

Recently, deep learning methods have achieved promising results in frontal face image synthesis [10], [11], [12], [13], [14], [15], [16], [13], [17], [18], [19]. These methods often take nonfrontal-frontal facial image pairs as the ground-truth to learn the projection from a nonfrontal view to a frontal view in a data-driven manner. However, most existing deep learning based facial frontalization methods [10], [11], [12], [13], [14], [16] only work well on images taken in a constrained environment while cannot generalize well to unconstrained scenarios. This results from the fact that most existing methods use the *constrained* images from the Multi-PIE [20] database for training. It is known that Multi-PIE (contains facial images from 337 subjects with 13 poses and 20 illuminations) is the largest database which has explicit pose annotation and can be used to facilitate the construction of deep models (e.g. GANs) to learn the mapping from different poses to the frontal one. Clearly, Multi-PIE based training has two drawbacks, namely (1) The training data lacks diversity (337 subjects only) so that the trained model can not capture sufficient amount of interpersonal variation. (2) The images are all captured in the same constrained environment. Thus those models trained on Multi-PIE cannot generalize well to faces image in the wild. A natural solution to problems (1) and (2) is to train models using a large unconstrained database (e.g. CASIA WebFace [21], MegaFace [22], MS-Celeb-1M [23]) which has a large number of unconstrained training images. So far only a handful of methods (e.g. FF-GAN[13], DR-GAN[12]) do have the ability to leverage outdoor unconstrained face images for training. However, those methods can not utilize both paired and unpaired data in a unified training framework. Specially, face normalization method FNM[24] is able to combine both paired and unpaired data into a unified framework. But the restored images suffer from color bias issue.

In Fig. 1, we summarise some results obtained with the previous TPGAN method (i.e trained on Multi-PIE) along

Zhihong Zhang, Ruiyang Liang, Xu Chen and Xuexin Xu are with the School of Informatics, Xiamen University, Xiamen, 361005, China.

Guosheng Hu is with the Anyvision group, Belfast, BT3 9AD, UK.

Wangmeng Zuo is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China.

Edwin R. Hancock is with the the Department of Computer Science, The University of York, York, YO10 5GH, UK.

* Corresponding author: Xu Chen (E-mail: chenxu31@gmail.com)

This work is supported by the Research Funds of State Grid Shaanxi Electric Power Company and State Grid Shaanxi Information and Telecommunication Company (contract no.SGSNXT00GCJS1900134).

Fig. 1. Each column consists of 4 subcolumns, which represent the input face, the face restored by TP-GAN [14], the face restored by FNM [24] and the face restored by CCFF-GAN, respectively.

with our method for comparison. We observe that there is an obvious color bias between the synthetic frontal face obtained by TP-GAN method and the corresponding non-frontal input. In some cases, the synthetic faces are even incomplete and miss a lot of fine detail around facial features. FNM method also suffers from color bias issue but with better facial details compared to TP-GAN. More results can be referred to in Fig. 7. Our method, on the other hand, does not suffer from these drawbacks. As we know, it is nontrivial to use unconstrained images for facial frontalization training because of the lack of *paired* data. Specifically, the nonfrontal images from an in-the-wild database usually do not have their frontal counterparts (from the same subjects with exactly frontal faces under the same conditions, such as illumination, expression, etc) to establish the nonfrontal-frontal pairs necessary for learning frontalization in a supervised manner.

To achieve promising face frontalization (FF) performance in both constrained and unconstrained environments, in this paper, we propose a semi-supervised learning method using both constrained and unconstrained face images. Clearly, it is very challenging to use unpaired faces to facilitate the facial frontalization training. Motivated by the recent success of unpaired image-to-image translation [25], we develop the *Cycle-Consistent Face Frontalization Generative Adversarial Network*, termed as **CCFF-GAN** to make use of unconstrained data for training. Specifically, CCFF-GAN is based on two generators, namely $G_{N\to F}$ and $G_{F\to N}$ that respectively learn the nonfrontal-to-frontal and frontal-to-nonfrontal translations. Note that the generator $G_{N\to F}$ is required to be the inverse mapping of $G_{F\to N}$, i.e., $G_{N\to F}(G_{F\to N}(x)) \approx x$, and vice versa. In this way, unpaired data can successfully be applied to learn $G_{N\to F}$ and $G_{F\to N}$. To train the model, paired and unpaired data are treated separately. Specifically, the unpaired non-frontal faces are sequentially fed to $G_{N\to F}$ and $G_{F\to N}$ to reconstruct themselves; Similarly, $G_{F\to N}$ and $G_{N\to F}$ process the (near) frontal faces. This 'cycle' process is unsupervised learning via the aforementioned self-reconstruction. On the other hand, the paired (labeled) faces can be fed to either $G_{F\to N}$ and $G_{N\to F}$ using direct supervision (labeled poses) rather than self-reconstruction. This clear supervision information reduces training difficulty and alleviates the intrinsic ambiguity brought by Cycle constraint. In addition, unlike the original cycle constraint which measures the pixel-wise difference using ℓ_1 loss, we instead propose to use the perceptual loss [26] to measure the semantic similarity between the cycle-reconstructed image and the original one. Since the unpaired data can be used for training (unsupervised learning), it is natural to incorporate the paired constrained images (supervised learning) to construct a semi-supervised learning framework.

The main contributions of this paper are summarized as follows.

- We present a semi-supervised learning framework, *i.e.* CCFF-GAN, which exploits both the paired indoor face images with a limited number of subjects and the unpaired in-the-wild faces with much more inter-personal variations to train the face frontalization network. Thus, our CCFF-GAN can generate high quality frontalization result and generalize well to unconstrained face images.
- For effectively leveraging paired and unpaired images in training, pixel-level fidelity and perceptual cycle consistency are respectively proposed to learn the face frontalization network. Adversarial loss and identity preserving loss are further introduced to enchance the visual quality and recognition performance of the frontalized image.
- Extensive experiments are conducted on Multi-PIE [20], LFW [27], IJB-A [28] and CFP [3] datasets. The results show that CCFF-GAN can achieve very promising face frontalization performance on in-the-wild faces. In addition, our method can effectively improve the pose-invariant face recognition performance.

The remainder of this paper is organized as follows. Sec. II briefly surveys the related work. Sec. III presents our CCFF-GAN and Sec. IV reports the experimental results. Finally, Sec. V ends this work with some remarks.

II. Related Work

A. Face Frontalization

Face frontalization aims to generate a frontal view from a given face with arbitrary nonfrontal view. Early efforts in face frontalization usually explicitly use a 3D model, typically a 3D Morphable Model (3DMM), to reconstruct a 3D face by fitting the given 2D nonfrontal face image to the 3D model. Then, a frontal face image can be generated by rotating and rendering the 3D model and then projecting the image of the face back onto the appropriate 2D plane. The pose transformation can thus be handled intrinsically in a 3D space. For example, Ferrari *et al* [29] present an effective face frontalization algorithm for frontal view rendering of a face image based on fitting a 3DMM. Zhu *et al.* [30] propose a High-fidelity *Pose and Expression Normalization* (HPEN) method, aiming at automatically generating a frontal face with neutral expression under a landmark matching assumption. Hassner *et al.* [31] proposed to use a single, unmodified 3D reference as an approximation to all query faces for producing frontalized views. Recently, deep learning technique has demonstrated its effectiveness in many computer vision tasks, including face frontalization. Deep learning based face frontalization methods often utilize a Convolutional Neural Network (CNN), typically with a encoder-decoder structure, to learn the mapping from a nonfrontal view to a frontal view which requires hundreds of thousands of paired nonfrontal-frontal face images for training [10], [11], [12], [13], [14], [15], [16], [13], [17], [18], [19]. For

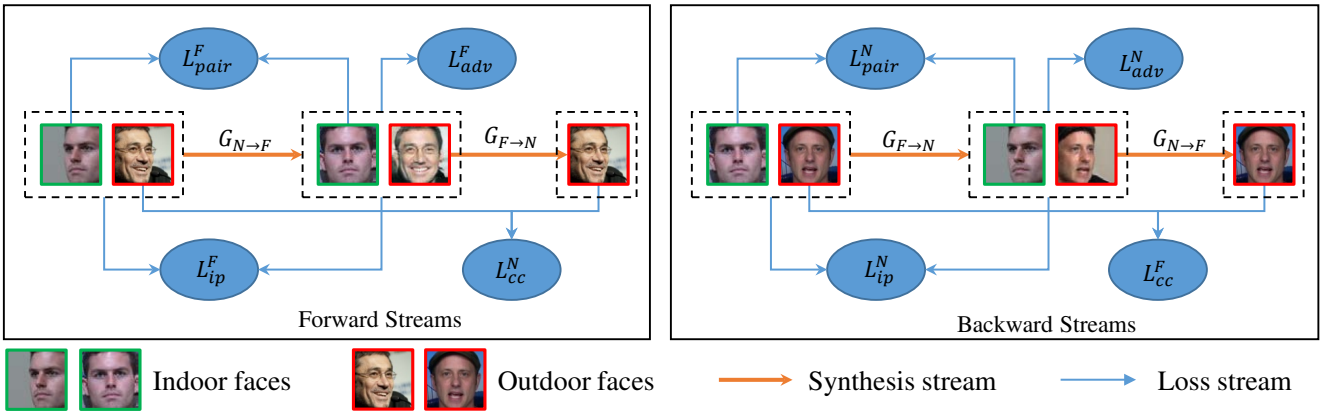


Fig. 2. The framework of the proposed CCFF-GAN. Generators $G_{N\to F}$ and $G_{F\to N}$ learn the nonfrontal-to-frontal and frontal-to-nonfrontal translations, respectively. For outdoor face images, the synthetic frontal/nonfrontal faces are brought back to nonfrontal/frontal view to compare with the original faces in high-level semantic feature space in forward/backward streams, respectively. L^F_{pair} and L^N_{pair} are for paired data which measure pixel-wise differences between the generated results and the ground truth. Their unpaired counterparts L^F_{cc} and L^N_{cc} leverage perceptual cycle consistency constraint to guide the training process. L^F_{adv} and L^N_{adv} are adversarial losses. L^F_{ip} and L^N_{ip} are identity preserving loss.

example, Yin *et al.* [13] present a novel deep 3D Morphable Model (3DMM) conditioned Face Frontalization Generative Adversarial Network (FF-GAN) to frontalize faces by utilizing shape and appearance priors from the 3DMM. Huang *et al.* [14] process the global and local transformations separately through a *Two-Pathway Generative Adversarial Network* (TP-GAN) to better preserve the facial texture details. Zhang *et al.* [15] reconstruct the frontal facial view by explictly 'moving' pixels from the nonfrontal facial view, instead of 'synthesising' them. This prevents the generated results from being blurry. Hu *et al.* [16] have proposed a novel Coupled-Agent Pose-Guided Generative Adversarial Network (CAPG-GAN) to generate both neutral and profile head pose face images by utilizing facial landmark heatmaps to guide the training. Tran *et al.* [19] propose to learn a representation that both for frontal face image synthesis and pose-invariant face recognition. Qian *et al.* [24] proposed Unsupervised Face Normalization with Extreme Pose and Expression in the Wild (FNM). Their model first encode images by utilizing a pre-trained face expert network and then tried to recover photorealistic images from the extracted feature. Recently, Rong *et al.* [32] proposed FI-GAN, aiming at improving the recognition performance under large face poses via a Feature-Mapping Block which maps the features of profile space to the frontal space. In most face frontalization methods, a large number of paired nonfrontal-frontal face images, typically from the MultiPIE dataset, are required to train the model in a fully supervised manner. Although these deep learning based face frontalization methods show promising results on indoor face images, their generalizability on outdoor face images is still questionable, since the training samples are captured in a controlled environment. Moreover, considering that the training samples (from MultiPIE dataset) are captured from only a few subjects, which further limits the generalizability of face frontalization models in practical applications. In this paper, we propose to utilize both indoor and outdoor face images to train the face frontalization model in a semi-supervised manner by making use of both indoor and outdoor face images, which can effectively improve the generalizability of the face frontalization model. It is worth noting that the FNM [24] method also utilizes both indoor and in-the-wild data for training. But different from the proposed method, the FNM is a face normalization method that conducts both face frontalization and expression normalization, and it cannot rotate the frontal face to a nonfrontal one like our proposed method. We also compare the results of our proposed method and FNM later in experiment section.

B. Adversarial Image Synthesis

Goodfellow's *Generative Adversarial Network* (GAN) [33] has stimulated intense interest and consistently demonstrated its effectiveness in a wide range of tasks, especially in image synthesis. Typically, a GAN is composed of a generator and a discriminator. The generator is trained to synthesize fake images to fool the discriminator, while the discriminator learns to differentiate the fake images from the real ones. The generator and discriminator are trained in turn in an adversarial manner. GANs have been proven to be powerful tools for image synthesis since the generator is trained to synthesize realistic images that accurately match the detailed data distribution of their real counterparts. Recently, Zhu *et al.* [25] address the image-to-image translation problem by introducing a cycle-consistent adversarial network (CycleGAN), which has received significant attention. Perhaps the most attractive characteristic of CycleGAN is that it does not require paired images for training, i.e., it can be trained in an entirely unsupervised manner. Motivated by this work, in this paper we aim to jointly learn both the nonfrontal-to-frontal and frontal-to-nonfrontal translations in a cycle consistent manner. Unlike the original CycleGAN which uses only unpaired images, in this paper we make use of both paired (captured in the controlled environment) and unpaired (captured in the uncontrolled environment) nonfrontal-frontal face images to

train the model, with the aim of effectively reducing the intrinsic ambiguity encountered with CycleGAN.

III. Methodology

In this section, we introduce our novel method in detail. We first present an overview of the proposed method in Sec. III-A. Then we introduce the perceptual cycle consistency constraint and the associated network architecture in Sec. III-B and Sec. III-C, respectively. Finally, the loss functions are detailed in Sec. III-D.

A. Overview

In this paper, we explain how to make use of both paired (captured in a constrained environment) and unpaired (captured in an unconstrained environment) nonfrontal-frontal face images to address face frontalization problem in a semi-supervised manner. The indoor labeled face pairs can only be used to learn frontalizations with poor generalizability to faces imaged in-the-wild. Outdoor unlabeled faces, although they cannot be used to learn stable frontalizations, can on the other hand be used to learn the characteristics of in-the-wild faces. Intuitively, utilizing in-the-wild faces to assist in the construction of the face frontalization model would contribute to improving the generalizability. The reason is that these faces span much greater variance in the inter-personal variations (identities/subjects) and intra-personal variations (poses, lightings, expressions, etc). In this paper, we propose to make use of both indoor and outdoor faces in a semi-supervised manner, with the aim of achieving stable in-the-wild face frontalization. The framework is illustrated in Fig. 2.

As illustrated, the proposed *Cycle-Consistent Face Frontalization Generative Adversarial Network* (CCFF-GAN) contains two generators (i.e., $G_{N\rightarrow F}$ and $G_{F\rightarrow N}$) which respectively learn the nonfrontal-to-frontal and frontal-to-nonfrontal mappings. Note that in the frontal-to-nonfrontal translation, a pose code is required to specify the target pose. In this work, the pose code is a one-hot vector specifying the pose of the desired face. Although such a pose code is unnecessary in the nonfrontal-to-frontal translation, we still use it for the sake of uniformity. In this case, the value of pose code is set to 0.

Both indoor and outdoor face images are jointly used to learn the nonfrontal-frontal translations in unconstrained environments. The indoor data (from the Multi-PIE [20] dataset) provides paired nonfrontal-frontal face images to serve as the direct supervision in the learning or training of the network. While for outdoor data, the *cycle consistency* constraint is applied to regularize the translations, by requiring that the two generators be the inverse mappings of one-another in both the forward (nonfrontal-frontal-nonfrontal) and backward (frontal-nonfrontal-frontal) processing streams as illustrated in Fig. 2. However, instead of the ℓ_1 loss that performs at the pixel level, we use instead the perceptual loss to apply the cycle consistency constraint in the high-level semantic feature space, which will be detailed later on in Sec. III-B.

Unlike the original CycleGAN [25] that uses two discriminators, in this paper, we use a single conditional discriminator D for both nonfrontal and frontal face images. This takes an image and a pose code as inputs to determine whether the given image is real or synthetic.

Our model makes use of both paired and unpaired nonfrontal-frontal face images to learn the face translation. The paired images provide both direct supervision and also serve as anchors during the training process. This effectively alleviates the intrinsic ambiguity inherent in CycleGAN. The unpaired data can effectively improve the generalizability of the face frontalization model by learning from the unconstrained face images. The loss functions involved in the learning process will be elaborated in Sec. III-D.

B. Perceptual Cycle Consistency

CycleGAN [25] addresses the image-to-image translation problem by learning mappings between the source and target domains using unpaired images only. Since no aligned image pairs are available to provide direct supervision, the cycle consistency constraint is used instead furnishing a source of indirect supervision to guide the training. To be specific, for each image x from the source domain, two generators G and F are required to satisfy forward cycle consistency relation: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, where G and F learn the source-to-target and target-to-source mappings, respectively. Likewise, for each image y from the target domain, a similar backward cycle consistency relation holds: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$.

The original CycleGAN adopts the ℓ_1 loss to measure the similarity between the cycle-reconstructed image and the original one at the pixel-level, i.e., $\|F(G(x)) - x\|_1 + \|G(F(y)) - y\|_1$. However, we argue that such pixel-wise regularization is not suitable for our nonfrontal-frontal face translation case, especially for those faces captured in the unconstrained environment. This is because the semantic structures presented in face images can change considerably during the rotation of head. A face image captured in an unconstrained environment often contains a certain proportion of complicated natural background structure. In general, the larger the facial pose, the bigger the background area. When transferring a nonfrontal face to a frontal one, a part of background area will be covered by the frontalized face. However, such an occluded background area can not be accurately restored when this process is reversed, i.e. when transferring the frontalized face back to the nonfrontal one.

Thus we argue that the pixels should not be treated equally in the cycle consistency constraint in nonfrontal-frontal face translations. Intuitively, the desired regularization in nonfrontal-frontal translation should focus on the face area and ignore the background.

To achieve this goal, in this paper we use the perceptual loss [26] instead of pixel-wise ℓ_1 loss for cycle consistency constraint. The perceptual loss measures the high-level semantic feature differences rather than pixel-wise differences between two images. In this work, we use a pre-trained face recognition network (e.g., ResNet, Light CNN) to extract the feature representations from the face images. Since such network is trained to recognize faces, the trained model should focus on extracting facial features and ignore the background area. Based on such features, *perceptual cycle consistency regularization* is more robust than the original pixel based version

especially for in-the-wild face images. The corresponding loss function is detailed later in Sec. III-D.

C. Network Architecture

The architectures of the generators and the discriminator are illustrated in Fig. 3. Here the generator is a convolutional neural network with an encoder-decoder structure. It takes a face image and a target pose code as inputs. The face image is fed into the encoder which commences with a convolutional layer, followed by 4 encoder blocks. Each encoder block consists of a residual block and a convolutional layer with stride 2 to downsample the feature map. The numbers of convolutional filters in these encoder blocks are 64,128,256 and 512, respectively. Three residual blocks [34] are appended to improve the nonlinear modeling ability of the generators, followed by a bottleneck layer (i.e., the fully connected layer) in the middle to transform the feature map to a vector with dimension 512. This vector is then concatenated with the input pose code. The decoder is symmetrical with the encoder, which consists of 4 decoder blocks and a convolutional output layer. Each decoder block contains a deconvolutional layer [35] to upsample the feature map, followed by a residual block. Finally, an additional convolutional layer is used to generate the output, i.e., the synthesized face image. In addition, the Instance Normalization (IN) [36] and ReLU nonlinearity are applied after each convolutional and deconvolutional layer, with an exception that the $tanh$ activation is used to normalize the output of the generator.

The discriminator commences with 4 convolutional layers with stride 2 to gradually downsample the feature map, followed by 2 residual blocks. The number of filters for these convolutional layers are 64, 128, 256 and 512, respectively. Then, a fully connected layer is applied to generate a 512-dimension feature vector. This feature vector is concatenated with the input pose code, and followed by 2 fully connected layers with output dimensions 128 and 1, respectively. The Instance Normalization (IN) [36] is also applied after each convolutional layer. For activation, we adopt the Leaky ReLU [37] with slope 0.2 after each convolutional and fully connected layers as suggested in DCGAN [38] except the last layer.

D. Loss Functions

In this section, we describe the loss functions used in this work.

1) Pixel-wise Fidelity Loss for Paired Data: The proposed CCFF-GAN makes use of both paired and unpaired nonfrontal-frontal face images to train the model. The paired data can be directly used to guide the training by minimizing the following loss functions:

$$\begin{aligned} L_{pair}^{N} =& \mathbb{E}_{(x,y)\sim(\mathcal{X}_p,\mathcal{Y}_p)}(\|G_{F\to N}(y,c_x)-x\|_1) \\ L_{pair}^{F} =& \mathbb{E}_{(x,y)\sim(\mathcal{X}_p,\mathcal{Y}_p)}(\|G_{N\to F}(x,c_y)-y\|_1) \end{aligned} \tag{1}$$

where $x \in \mathcal{X}_p$ and $y \in \mathcal{Y}_p$ represent a pair of nonfrontal-frontal face images, (c_x, c_y) denote the pose codes of (x, y), respectively.

2) Perceptual Cycle Consistency Loss for Unpaired Data: As introduced in Sec. III-B, we use the perceptual cycle consistency constraint to guide the training of the unpaired data. The loss functions are defined as follows:

$$\begin{aligned} L_{cc}^{N} =& \mathbb{E}_{(x,y)\sim(\mathcal{X}_u,\mathcal{Y}_u)}\|\phi_{3,4}(G_{F\to N}(G_{N\to F}(x,c_y),c_x))- \\ & \phi_{3,4}(x)\|_2 \\ L_{cc}^{F} =& \mathbb{E}_{(x,y)\sim(\mathcal{X}_u,\mathcal{Y}_u)}\|\phi_{3,4}(G_{N\to F}(G_{F\to N}(y,c_x),c_y))- \\ & \phi_{3,4}(y)\|_2 \end{aligned} \tag{2}$$

where $x \in \mathcal{X}_u$ and $y \in \mathcal{Y}_u$ represent the unpaired nonfrontal-frontal face images, (c_x, c_y) respectively denote the pose codes of (x, y), and $\phi_{i,j}$ indicates the feature map obtained by j-th convolution (after the activation) in i-th block of the pre-trained face recognition network.

3) Discrimination and Adversarial Loss: The discriminator distinguishes a real face image from a synthetic one, while encouraging the generator to synthesize realistic face images. The generators and discriminator are trained alternatingly in an adversarial manner. To be specific, in the discrimination stage, the discriminator is trained to determine whether the given image is real or synthesized by minimizing the following loss functions:

$$\begin{aligned} L_{dis}^{N} =& \mathbb{E}_{(x,y)\sim(\mathcal{X},\mathcal{Y})}((D(x,c_x)-1)^2+D(G_{F\to N}(y,c_x),c_x))^2 \\ L_{dis}^{F} =& \mathbb{E}_{(x,y)\sim(\mathcal{X},\mathcal{Y})}((D(y,c_y)-1)^2+D(G_{N\to F}(x,c_y),c_y))^2 \end{aligned} \tag{3}$$

where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ respectively represent the nonfrontal and frontal face images, (c_x, c_y) denotes the pose codes of (x, y). In the generation stage, on the other hand, the generators are encouraged to synthesize realistic face images to fool the discriminator. This involves the following adversarial loss functions:

$$\begin{aligned} L_{adv}^{N} =& \mathbb{E}_{(x,y)\sim(\mathcal{X},\mathcal{Y})}(D(G_{F\to N}(y,c_x),c_x)-1)^2 \\ L_{adv}^{F} =& \mathbb{E}_{(x,y)\sim(\mathcal{X},\mathcal{Y})}(D(G_{N\to F}(x,c_y),c_y)-1)^2 \end{aligned} \tag{4}$$

Note that in this paper, the LSGAN [39] is adopted instead of the original GAN [33] to mitigate the training instability issue.

4) Identity Preserving Loss: Finally, to preserve the identity information in nonfrontal-frontal translations, we adopt the identity preserving loss as proposed in [14]. To be specific, we extract the high-level representations from the inputted face and the synthetic one via the pre-trained face recognition network, and require these two representations to be the same. The loss functions are defined as follows:

$$\begin{aligned} L_{ip}^{N} =& \mathbb{E}_{(x,y)\sim(\mathcal{X},\mathcal{Y})}\|\phi_{-2}(G_{F\to N}(y,c_x))-\phi_{-2}(y)\|_2 \\ L_{ip}^{F} =& \mathbb{E}_{(x,y)\sim(\mathcal{X},\mathcal{Y})}\|\phi_{-2}(G_{N\to F}(x,c_y))-\phi_{-2}(x)\|_2 \end{aligned} \tag{5}$$

where ϕ_{-2} indicates the feature map extracted by the last but one layer of the pre-trained face recognition network. Note that we utilize the most abstract features extracted from the last but one layer of the pre-trained face recognition model to calculate the identity preserving loss. Such feature is extracted from a fully connection layer that typically behind a average global pooling operation. So although the original face and synthesized face may have drastically different facial yaw, their spacial geometry structures information will be removed by the

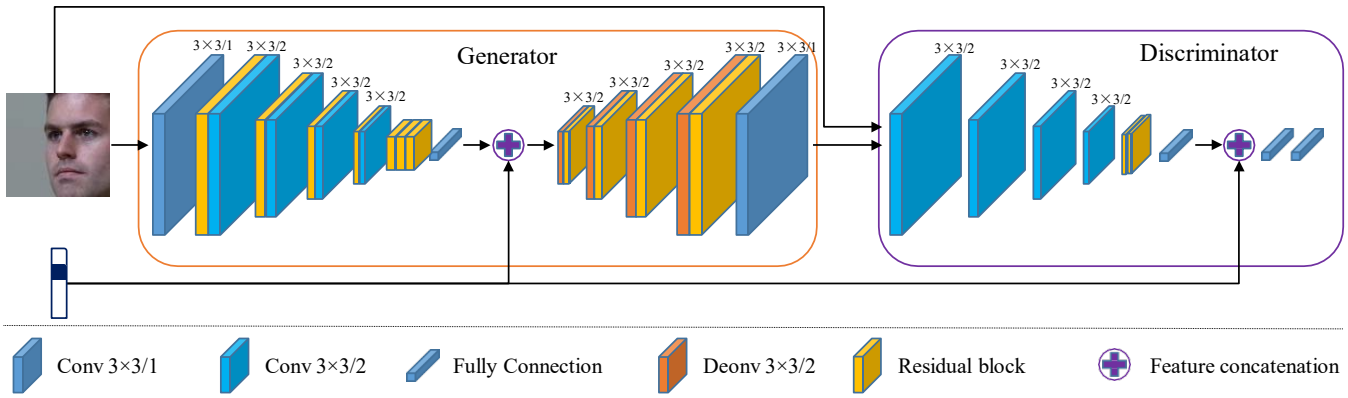


Fig. 3. The network architectures.

global pooling operation, leaving only the abstract features that do not contain facial yaw information.

5) Total Loss: The final loss function for the generators is a weighted sum of individual loss functions described above, which is defined as follow:

$$L^F_{pair} + L^N_{pair} + \lambda_1(L^F_{cc} + L^N_{cc}) + \lambda_2(L^F_{adv} + L^N_{adv}) + \lambda_3(L^F_{ip} + L^N_{ip}), \quad (6)$$

where λ_1, λ_2, and λ_3 are the tradeoff parameters.

E. Training Details

We train the two generators and the shared discriminator in turn at each iteration. To be specific, we first train the discriminator with frontal real and synthetic face images while the generators are frozen. Then, the nonfrontal-to-frontal generator is trained with the other generator (i.e., the frontal-to-nonfrontal generator) and the discriminator fixed. After that, the discriminator is trained again with nonfrontal real and synthetic face images with the generators fixed. Finally, we train the frontal-to-nonfrontal generator while the other generator (i.e., the nonfrontal-to-frontal generator) and the discriminator are frozen. The training strategy is detailed in Algorithm 1.

Input: nonfrontal face dataset $\mathcal{X}$, frontal face dataset $\mathcal{Y}$

for $t = 1, 2, \ldots, T$ **do**
  train D with loss: L^F_{dis}
  train $G_{N\to F}$ with loss: $L^F_{pair} + \lambda_1 L^N_{cc} + \lambda_2 L^F_{adv} + \lambda_3 L^F_{ip}$
  train D with loss: L^N_{dis}
  train $G_{F\to N}$ with loss: $L^N_{pair} + \lambda_1 L^F_{cc} + \lambda_2 L^N_{adv} + \lambda_3 L^N_{ip}$
end

Algorithm 1: Training Strategy

We adopt **Adam** [40] as the optimizer to train the network with a learning rate of 10^{-4} for 40,000 iterations The batch size is 4 with each mini-batch consists of 2 paired and 2 unpaired nonfrontal-frontal face images. Other hyper-parameters are empirically set as: $\lambda_1 = 5$, $\lambda_2 = 0.05$ and $\lambda_3 = 0.01$.

IV. Experiments

To evaluate the effectiveness of the proposed CCFF-GAN, we evaluate it both qualitatively and quantitatively in comparison with the state-of-the-art face frontalization methods. In this section, we first introduce the experimental settings in Sec. IV-A, including a description of the training and test datasets, as well as the preprocessing procedure. Then we present some representative visual results for face frontalization in Sec. IV-B, and subsequently report the quantitative results and analyses for face recognition in Sec. IV-C. Finally, we present ablation study in Sec. IV-D to investigate the role of different losses respectively.

A. Experimental Settings

The proposed CCFF-GAN is trained in a semi-supervised manner by exploiting both paired and unpaired nonfrontal-frontal face images. The paired face images were from the Multi-PIE [20] dataset which contains 750,000+ face images captured from 337 subjects in a constrained environment, with 13 poses from $-90°$ to $90°$ and 20 illuminations. Following the settings in [14], we used the first 200 subjects to train the model, leaving the remaining 137 subjects for testing.

In addition to the paired data, we also collected unpaired nonfrontal-frontal face images to train our model. Such unpaired data was from the MS-Celeb-1M [23] dataset, which consists of about 10 million face images harvested from nearly 100,000 subjects, and most of these images were captured in an unconstrained environment. However, note that the majority of the face images in the MS-Celeb-1M dataset are frontal faces, while the number of nonfrontal faces is very limited. To avoid the pose imbalance problem, we selected only a subset of the face images in the MS-Celeb-1M dataset for each pose range. To be specific, we first calculate the poses of the face images in the MS-Celeb-1M dataset using the 3DDFA algorithm [41] and categorized the images into different pose groups. Then we randomly selected a fixed size subset of the face images at every pose group to avoid pose imbalance problem. Summary satatistics for the training data are presented in Table I. Each of the face images used in our experiments is aligned and cropped to the size of $96 \times 96 \times 3$,

TABLE I
NUMBER OF TRAINING SAMPLES ACROSS DIFFERENT POSE.

Pose	$-90°$	$-75°$	$-60°$	$-45°$	$-30°$	$-15°$	$0°$	$15°$	$30°$	$45°$	$60°$	$75°$	$90°$
MS-Celeb-1M[23]	10	243	2876	15000	15000	15000	15000	15000	15000	15000	1919	174	3
Multi-PIE [20]	12420	12420	12420	12420	12420	12420	12420	12420	12420	12420	12420	12420	12420

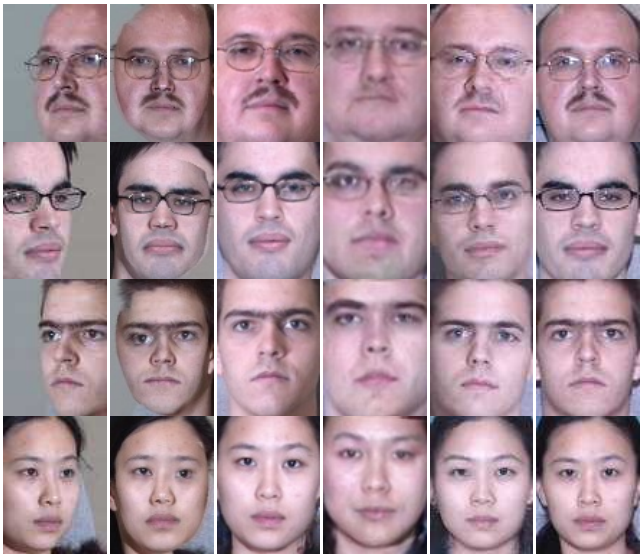
and then the pixel intensity values are linearly scaled into the interval $[-1, 1]$.

B. Qualitative Evaluation

Most existing face frontalization methods only use indoor face images that are captured in a constrained environment to train the model. This limits their generalizabilities in an unconstrained environment since such data cover too few inter-personal variations (identities/subjects) and intra-personal variations (poses, lightings, expressions, etc) to learn a robust frontalization. To address this issue, in this paper we also make use of in-the-wild face images (which cover much more inter-personal and intra-personal variations) to assist in training the face frontalization model via cycle-consistent image synthesis. To demonstrate the effectiveness of the proposed method, in this section we qualitatively evaluate the proposed CCFF-GAN by presenting some of the representative synthesis results obtained, and compare them with those produced by state-of-the-art face frontalization methods. In addition, we trained our nonfrontal-to-frontal generator $G_{N\to F}$ solely using only paired training data (*i.e.*, from the Multi-PIE[20] dataset). This model serves as a **Baseline** for the purposes of comparison and demonstrating the benefits of using in-the-wild (unpaired) nonfrontal-frontal face images.

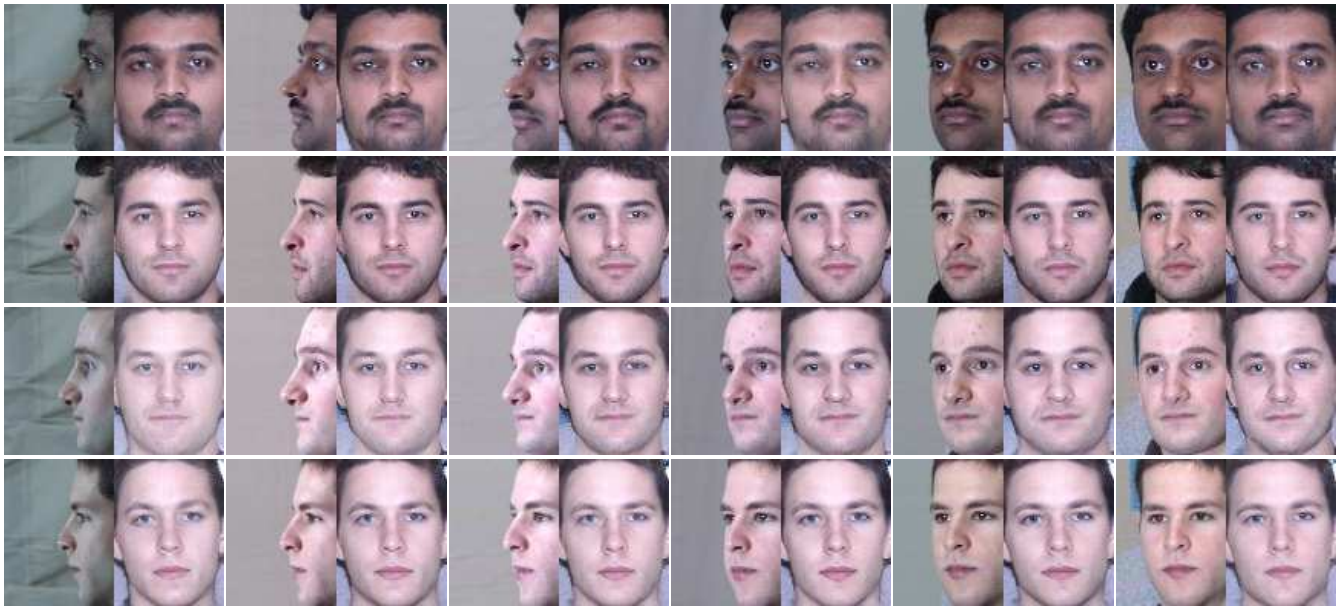
Fig. 4 shows the results of the proposed CCFF-GAN and several state-of-the-art face frontalization methods. In the figure the columns respectively represent (a) the input nonfrontal face, and the synthesized frontal face obtained using (b) the HPEN [30] method, (c) the TP-GAN [14] method, (d) the FF-GAN [13] method, (e) the CCFF-GAN method, and (f) the ground truth. It is clear that the traditional face frontalization methods (e.g., HPEN [30]) struggle to reconstruct the shape of the face. On the other hand, the results of the deep learning based face frontalization methods (e.g., FF-GAN [13]) tend to lack high-quality facial details. This implies poor generalizability since the models were trained on a limited set of subjects. By contrast, the proposed CCFF-GAN generates more realistic results. This can mainly be attributed to the proposed semi-supervised learning framework that learns the data distribution of frontal faces from both indoor and in-the-wild face images. It is worth noting that although TP-GAN [14] also generates high-quality results with fine facial details, it requires additional facial landmarks to assist in the face frontalization. This limits its generalizability in unconstrained environments, as we will analyze in greater detail later. More qualitative visual results for the proposed method based on Multi-PIE are presented in Fig. 5. Here we demonstrate the effectiveness of the proposed CCFF-GAN on different poses.

The major difference between the proposed CCFF-GAN and the competing methods is that we use both indoor and in-the-wild face images to train the model. Thus, it is reasonable that the proposed method generalizes well to in-the-wild face images captured in an unconstrained environment. To demonstrate this, we have evaluated the trained model on LFW [27] and IJB-A [28] datasets, and compared the results with those obtained with TP-GAN [14] method. Besides, we also compared the results with that of FNM [24], which also uses both indoor and in-the-wild data for training. It is worth noting that the FNM method is a face normalization method. That is, it not only performs face frontalizatoin, but also removes expression (synthesizes neural expression) from given input. Some representative results are shown in Fig. 7 , where the leftmost two columns and the rightmost two columns are from LFW and IJB-A, respectively. The sub-columns in each column respectively represent a) the input nonfrontal image, b) the results of the TP-GAN method c) the results of the FNM method and d) the results of the proposed CCFF-GAN. From these results, it is clear that that there is an obvious color bias between the synthetic frontal face obtained by TP-GAN method and the corresponding nonfrontal input. To be specific, the facial skin colors of the synthetic faces obtained by TP-GAN often differs from that of the input faces. This is because the TP-GAN model was trained on indoor face images which lack facial texture variations since these images were captured in a constrained environment and were from a very limited number of subjects (i.e., 200 subjects). Moreover, TP-GAN failed to synthesize the frontal face in some cases (e.g., 2nd row in column (b) and last row in column (c)) where the detected facial landmarks are inaccurate. These



(a) Input (b) [30] (c) [14] (d) [13] (e) Ours (f) GT

Fig. 4. Comparison with state-of-the-art face frontalization methods.



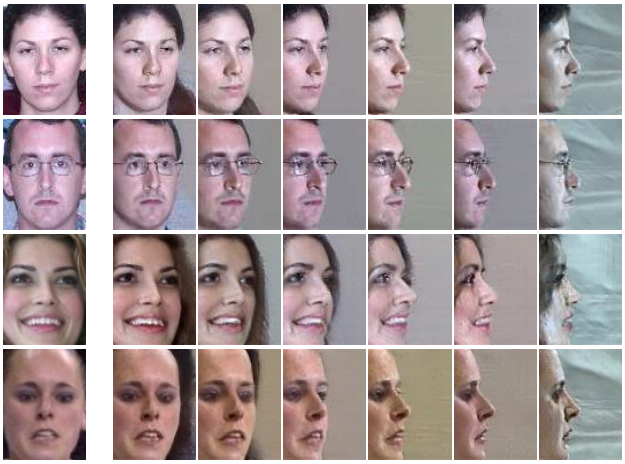
(a) 90° (b) 75° (c) 60° (d) 45° (e) 30° (f) 15°

(g) −15° (h) −30° (i) −45° (j) −60° (k) −75° (l) −90°

Fig. 5. Face frontalization from arbitrary poses in a constrained environment on Multi-PIE. Each column consists of 2 subcolumns, which represent the input face,and the face restored by our method respectively.

results indicate that the performance of TP-GAN relies critically on the landmark detection accuracy, which limits its generalization ability in unconstrained environments. On the other hand, although also incorporating the in-the-wild data into training, we can observe that the FNM method also suffers from the color bias problem. Obviously, the training of FNM was dominated by the indoor data, since the synthetic faces have very similar illuminations, backgrounds and skin colors with MultiPIE faces. By contrast, the proposed CCFF-GAN is able to synthesize more realistic frontal faces from in-the-wild nonfrontal faces as well as better preserving the facial details. This implies good generalizability to an unconstrained environment, a feature mainly attributable to the proposed semi-supervised learning framework basaed on both indoor and in-the-wild face images.

In addition, some representative frontal-to-nonfrontal synthetic results are shown in Fig. 6, where the first two rows were sampled from indoor data and the last two rows were from in-the-wild data. From the results, we can found that the nonfrontal-to-frontal generator produces good results in most cases, while also produces some artifacts when facing extreme pose (e.g., 90°), especially for in-the-wild faces. This is due to the fact that the number of profile face images is very rare in outdoor training set (see Table I), which hampers the generator in learning a reliable frontal-to-nonfrontal mapping.



(a) Input (b) 15° (c) 30° (d) 45° (e) 60° (f) 75° (g) 90°

Fig. 6. Face rotation to arbitrary poses. 1-2 rows sampled from indoor data and 3-4 rows sampled from outdoor data.

C. Face Recognition

To further demonstrate the effectiveness of the proposed CCFF-GAN, we have also quantitatively evaluated it on face recognition. In this work, we employed ResNet50 [1] and Light CNN [2] as the face recognition models, which are further refined using the training set of paired and unpaired data. The trained model are also used to compute the identity preserving loss as stated in Section III-D. We first conducted face recognition on the LFW [27], IJB-A [28] and CFP [3] datasets, where most face images were captured in an unconstrained environment. The results are shown in Table II, Table III and Table V, respectively.

Specifically, to demonstrate the superiority of using images collected from unconstrained environment, we conduct our training process using two different configurations of dataset. We first train our model using only Multi-PIE [20] dataset, which, as expected, leads to performance drop in face recognition test. Then the preprocessed MS-Celeb-1M dataset is included to jointly train the model along with Multi-PIE dataset. The second configuration which is the complete version of our CCFF-GAN prevails. Not surprisingly, the proposed CCFF-GAN outperforms other competing face frontalization methods on the LFW, IJB-A and CFP datasets in most cases, thus demonstrating its superiority in the unconstrained environment. In the IJB-A verification protocol, FNM outperforms all the other method including our method because FNM conducts both face frontalization and expression normlization, while other methods including ours do not conduct expression normalization. The good generalization ability of the CCFF-GAN can be attributed to the underlying semi-supervised learning framework which learns face frontalization from both indoor and in-the-wild face images. By contrast, most existing face frontalization methods are only trained on indoor face images (typically from Multi-PIE[20] dataset), which limits their generalization abilities in the unconstrained environment.

[1]The code is publicly available at https://github.com/auroua/InsightFace_TF

[2]The code is publicly available at https://github.com/AlfredXiangWu/LightCNN

TABLE II
FACE VERIFICATION RESULTS ON LFW.

Methods	ACC(%)	AUC(%)
TP-GAN [14]	91.17 ± 1.44	92.78 ± 0.02
Hassner et al. [31]	93.62 ± 1.17	98.38 ± 0.06
HPEN [30]	96.25 ± 0.76	99.39 ± 0.02
LightCNN [42]	98.87 ± 0.61	99.69 ± 0.17
ResNet50 [43]	98.98 ± 0.52	99.79 ± 0.14
FF-GAN [13]	96.42 ± 0.89	99.45 ± 0.03
A3FCNN [15]	96.63 ± 0.99	99.29 ± 0.42
FI-GAN [32]	$98.30 \pm -$	$99.60 \pm -$
Ours(Multi-PIE only) + LightCNN	98.83 ± 0.50	99.63 ± 0.24
Ours + LightCNN	98.93 ± 0.56	99.67 ± 0.21
Ours(Multi-PIE only) + ResNet50	98.08 ± 0.47	99.78 ± 0.21
Ours + ResNet50	$\mathbf{99.20} \pm 0.49$	$\mathbf{99.83} \pm 0.13$

We also conducted the face recognition experiment on the Multi-PIE dataset following the settings used with TP-GAN [14]. Specifically, we selected a single frontal face image for each subject in the test dataset and treated the selected face images as the gallery set, leaving the remaining face images as the probe or test set. Then, we synthesized the frontal view for each nonfrontal face image in probe set using the trained CCFF-GAN. We then extracted the deep features using the pre-trained recognition network. The rank-1 recognition accuracy is evaluated by comparing the features from the frontalized faces in the probe set and those from the real frontal faces in the gallery set. The comparison was performed using the cosine distance metric. The evaluation results are given in Table IV and compared with the competing methods.

Our method achieves very competitive performance. Note that the methods compared in Table IV are obtained using the models training and tested only on Multi-PIE, They thus tend to overfit the test set. In comparison, our method is designed for in-the-wild faces (161,460 images from Multi-PIE, 110,225 images from MS-Celeb-1M in training set), the strong performance of our method on Multi-PIE means our methods can generalize well to constrained environment.

D. Ablation Study

In this section, we conduct an ablation study on several variations of the proposed CCFF-GAN by dropping each of the 4 loss functions in turn. This gives us insights into the individual roles of the different loss functions in the training process. All the presented results are generated using LFW [27] and IJB-A [28] datasets. Note that, our model is trained on MS-Celeb-1M [23] and Multi-PIE [20] databases only. We present visual results of 4 different CCFF-GAN variants obtained using the partial or curtailed loss function are shown in Fig. 8 along with the input profile images and the outputs the originql CCFF-GAN with the full loss function, while the corresponding quantitative results are shown in Table VI.

1) Remove L_{pair} loss: When trained without the L_{pair} loss, the generated face frontalization results suffer from model collapse problem. The generated faces lack of diversity with only minor differences in facial features and its quantitative performance has significantly deteriorated compared with other CCFF-GAN variants. As mentioned above, the L_{pair} loss is used to supervise learning from paired data and gives

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

Fig. 7. Face frontalizatoin from arbitrary poses in the wild on LFW (Columns 1-2) and IJB-A (Columns 3-4). Each column consists of 4 subcolumns, which represent the input face, the face restored by TP-GAN [14], the face restored by FNM [24] and the face restored by CCFF-GAN, respectively.

TABLE III
PERFORMANCE COMPARISON ON IJB-A DATABASE.

Methods	Verification		Identification	
	FAR=0.01	FAR=0.001	Rank1	Rank5
OpenBR [28]	23.6 ± 0.9	10.4 ± 1.4	24.6 ± 1.1	37.5 ± 0.8
TP-GAN [14]	31.5 ± 1.8	9.2 ± 1.1	48.6 ± 5.0	59.3 ± 5.6
GOTS [28]	40.6 ± 1.4	19.8 ± 0.8	44.3 ± 2.1	59.5 ± 2.0
Wang [44]	72.9 ± 3.5	51.0 ± 6.1	82.2 ± 2.3	93.1 ± 1.4
PAM [45]	73.3 ± 1.8	55.2 ± 3.2	77.1 ± 1.6	88.7 ± 0.9
LightCNN [42]	81.7 ± 2.8	69.7 ± 4.8	97.7 ± 1.1	98.5 ± 0.8
ResNet50 [43]	81.2 ± 2.3	67.2 ± 4.8	97.7 ± 1.3	98.6 ± 0.9
DR-GAN [12]	77.4 ± 2.7	53.9 ± 4.3	85.5 ± 1.5	94.7 ± 1.1
DR-GAN$_{AM}$ [19]	87.2 ± 1.4	78.1 ± 3.5	92.0 ± 1.3	96.1 ± 0.7
FF-GAN [13]	85.2 ± 1.0	66.3 ± 3.3	90.2 ± 0.6	95.4 ± 0.5
FNM [24]	**93.4** ± 0.9	**83.8** ± 2.6	96.0 ± 0.5	98.6 ± 0.3
A3FCNN [15]	80.4 ± 3.3	60.0 ± 8.6	92.2 ± 2.3	97.4 ± 0.9
Ours(Multi-PIE only) + LightCNN	82.9 ± 4.2	69.9 ± 6.4	97.8 ± 1.1	98.4 ± 0.9
Ours + LightCNN	82.8 ± 4.2	69.9 ± 5.8	97.8 ± 1.3	98.7 ± 0.8
Ours(Multi-PIE only) + ResNet50	83.4 ± 2.5	71.4 ± 4.7	97.9 ± 1.6	**99.1** ± 0.7
Ours + ResNet50	84.1 ± 2.6	72.3 ± 4.6	**98.1** ± 1.4	98.9 ± 0.7

strong supervision over the entire training process. Due to the complexity of unpaired data, it is not trivial for the network to learn a mapping from unpaired profile image to unpaired frontal image. On the other hand, the paired data has fewer variations in pose, illumination, hue, etc, which makes it much easier to learn the transformation. We argue that the training process require both paired data and the L_{pair} loss to prevent the training from being dominated by model collapse and producing unsatisfying results.

2) Remove L_{adv} loss: As shown in the third row of Fig. 8, the network becomes degenerate, producing an output image which is identical to the input image. This results from the absence of pose constrain normally introduced by the L_{adv} loss. In other words there is no penalty for not rotating the face. Although the restored images have a whitening color tone, the images still preserve a lot of identity information, which coincide with its second best performance in quantitative results. Those results indicate that the L_{adv} loss is indispensable in forcing the generated faces to rotate pose.

3) Remove L_{ip} loss: In this scenario, the network is finally able to produce visually pleasing synthesized images. The reason for introducing identity preserving loss [14] is its capacity to enforce the constraint that the generated images preserve identity information present in the original input images. Without this loss, the synthesized faces tend to be blurred and have deformations around the face contouring. The quantitative results also suffers from the poor quality of synthesized images causing performance loss of different extents in different tests.The cause of these effects is the loss of identity information. The results coincide with our intuition.

4) Remove L_{cc} loss: The L_{cc} loss is the unpaired counterpart of the L_{pair} loss. The synthesized results are expected to be deteriorate without using it. We observe the synthesized faces are in frontal pose but with significant distortions and deformations on the face contouring, eyes, nose, etc, which are far from plausible. Its quantitative results are also inferior to

TABLE IV
COMPARISON OF STATE-OF-THE-ART METHODS IN TERMS OF RECOGNITION ACCURACY (%) ON MULTI-PIE DATABASE. AVG1 AND AVG2 ARE THE AVERAGE ACCURACY IN 15° - 60° AND 15° - 90° RESPECTIVELY.

Methods	±15°	±30°	±45°	±60°	±75°	±90°	avg1	avg2
Zhu et al. [7]	90.7	80.7	64.1	45.9	-	-	70.4	-
Zhu et al. [5]	92.8	83.7	72.9	60.1	-	-	77.4	-
CPF [6]	95.0	88.5	79.9	61.9	-	-	81.3	-
DR-GAN [12]	94.0	90.1	86.2	83.2	-	-	88.4	-
DR-GAN$_{AM}$ [19]	95.0	91.3	88.0	85.8	-	-	90.0	-
A3FCNN [15]	98.7	98.9	95.8	92.7	-	-	96.5	-
LightCNN [42]	98.6	97.4	92.1	62.1	24.2	5.5	87.5	63.3
ResNet50 [43]	100.0	99.8	99.1	**95.3**	88.1	73.1	98.5	92.6
FF-GAN [13]	94.8	93.4	91.0	87.0	82.7	71.7	91.6	86.8
TP-GAN [14]	98.7	98.1	95.4	87.7	77.4	64.6	95.0	87.0
CAPG-GAN [16]	99.8	99.6	97.3	90.6	83.1	66.1	96.8	89.4
FNM [24]	98.9	98.1	96.8	92.7	80.6	63.8	96.6	88.5
GSP-GAN [46]	99.4	99.2	98.1	93.9	82.9	65.6	97.7	89.9
FI-GAN [32]	98.8	98.5	97.4	96.2	88.2	**77.0**	97.7	92.7
Ours(Multi-PIE only) + LightCNN	98.7	97.4	95.1	89.6	78.4	62.5	94.5	87.0
Ours + LightCNN	99.2	98.5	96.5	91.8	81.8	66.1	96.5	89.0
Ours(Multi-PIE only) + ResNet50	100.0	99.8	99.1	94.7	87.7	73.4	98.4	92.5
Ours + ResNet50	**100.0**	**99.8**	**99.2**	94.9	**88.3**	73.9	**98.5**	**92.7**

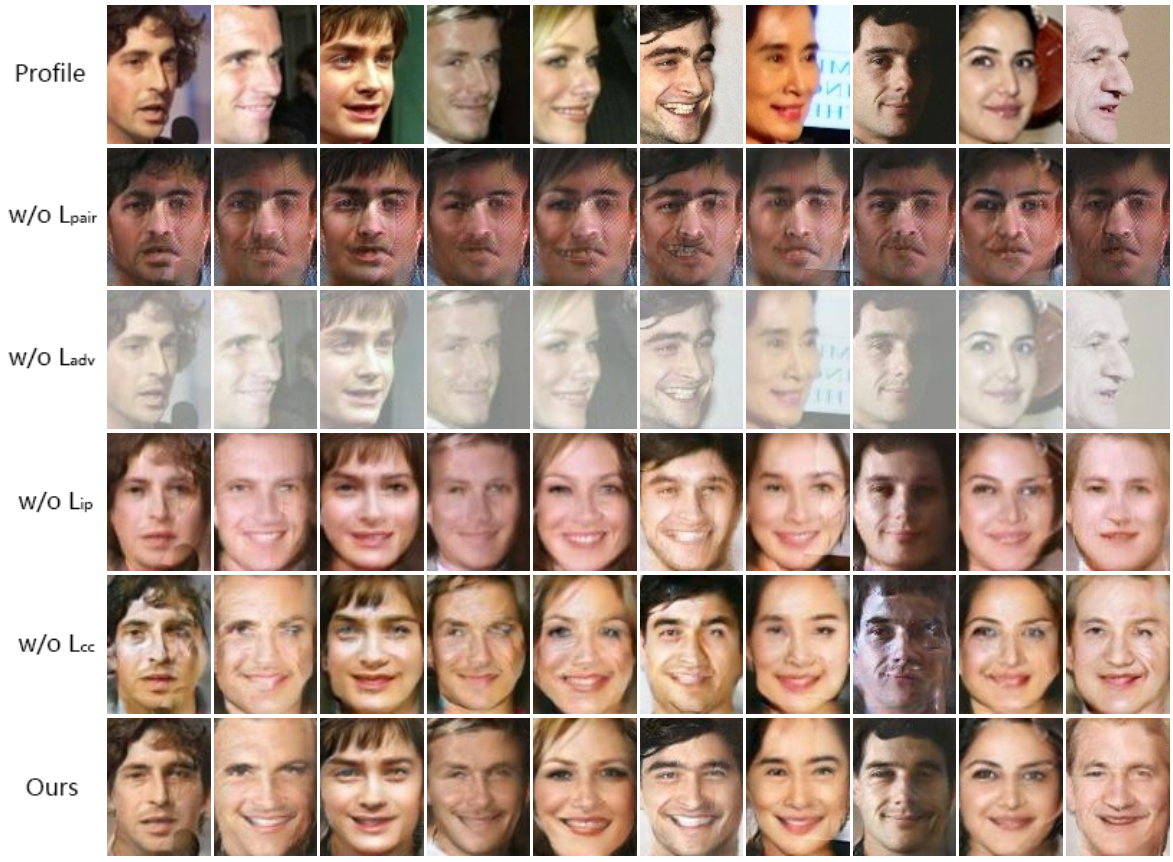


Fig. 8. Model Comparsion: synthesis results of CCFF-GAN and its variants on LFW(Columns 1-5) and IJB-A(Columns 6-10).

the complete model but close to the performance of $w/o\ L_{cc}$ model. These might be the result of not fully utilizing the unpaired dataset since we have dropped the L_{cc} loss in this setting. As a result the network mostly learns the easier paired transformation, which generalizes poorly to the faces in the wild. We also notice that the synthesized results are expected to deteriorate without using L_{cc}. However, there is no color bias between the input and synthetic face images. This is because even without L_{cc}, the unpaired in-the-wild face images still contribute to the adversarial learning and identity preserving constraint, encouraging the generators to better learn the data distribution of frontal faces from both indoor and in-the-wild face images.

TABLE V
PERFORMANCE (ACCURACY) COMPARISON ON CFP.

Method	Frontal-Frontal	Frontal-Profile
Sengupta et al. [3]	96.40 ± 0.69	84.91 ± 1.82
Sankarana et al. [47]	96.93 ± 0.61	89.17 ± 2.35
LightCNN [42]	99.37 ± 0.30	91.56 ± 1.89
ResNet50 [43]	99.54 ± 0.31	94.25 ± 1.33
DR-GAN [12]	97.84 ± 0.79	93.41 ± 1.17
DR-GAN$_{AM}$ [19]	98.36 ± 0.75	93.89 ± 1.39
Chen et al. [48]	98.67 ± 0.36	91.97 ± 1.70
PIM [17]	99.44 ± 0.36	93.10 ± 1.01
Peng et al. [8]	$98.67 \pm -$	$93.76 \pm -$
FI-GAN [32]	$98.90 \pm -$	$94.20 \pm -$
Ours(Multi-PIE only) + LightCNN	99.20 ± 0.33	91.17 ± 1.63
Ours + LightCNN	99.27 ± 0.39	91.87 ± 1.42
Ours(Multi-PIE only) + ResNet50	99.55 ± 0.30	93.39 ± 1.44
Ours + ResNet50	**99.61** ± 0.23	**94.30** ± 1.26

TABLE VI
ABLATION STUDY IN TERMS OF PERFORMANCE ON IJB-A DATABASE.

Methods	Verification		Identification	
	FAR=0.01	FAR=0.001	Rank1	Rank5
$w/o\ L_{pair}$	19.33±3.02	8.18 ± 2.10	37.61±6.10	51.42±8.72
$w/o\ L_{adv}$	82.20±2.72	70.55±4.17	96.52±1.72	96.97±1.34
$w/o\ L_{ip}$	68.47±3.81	52.22±5.02	94.20±2.16	96.26±1.87
$w/o\ L_{cc}$	64.37±4.15	45.84±5.11	93.05±2.91	95.45±1.34
Ours	**84.07**±2.60	**72.29**±4.62	**98.13**±1.41	**98.93**±0.67

V. DISCUSSIONS AND FUTURE WORK

To summarize, various practical methods have been proposed to address the face frontalization problem (*i.e.*, synthesizing a frontal view of face from a nonfrontal one) and these have achieved promising performance. However, most of the existing face frontalization methods rely on paired nonfrontal-frontal face images to train the model in a fully supervised manner. This limits their generalization ability in the unconstrained environment since such paired images (typically from the Multi-PIE dataset) were captured not only in a constrained environment but also from a very limited set of subjects. To address this problem, in this paper we have proposed a semi-supervised face frontalization framework, which learns mappings between the nonfrontal and frontal face images by utilizing both indoor constrained and outdoor unconstrained face images. To regularize the nonfrontal-frontal translation on unpaired outdoor nonfrontal-frontal face images, we have adopted a variant of the cycle consistency constraint. In doing so, we perform regularization in a high-level semantic feature space rather than the visual image space. Our experimental results demonstrate the effectiveness of the proposed method compared with previous face frontalization methods, especially for face images captured in an unconstrained environment.

1) Strengths: In face frontalization, there are very few methods that have utilized the in-the-wild face database for training. Only a handful of them (e.g, FNM [24]) use both paired and unpaired nonfrontal-frontal face images. However, results show that artefacts such as the skin color bias persists and they failed to reproduce realistic variation in facial appearance (i.e, the generated frontal face appears very similar to its exemplar from the Multi-PIE database). Our proposed method does not suffer from those drawbacks. This can be attributed to the use of the proposed semi-supervised framework. This leverages both inter-personal and intra-personal variations from the in-the-wild face images.

2) Weaknesses: In our method, we have incorporated a pose code to guide the training process. Those pose codes were obtained by 3DDFA [41], which is far from accurate when compared with handcrafted pose code. Such bias may impair the training process. The reason for this is that the pose information given is not precise and may sometimes even be rather noisy. On the other hand, those pose codes are encoded in a discrete way which means a loss of information.

3) Future Work: To further improve our method, more investigation should be made into the field of obtaining accurate face pose information and how to encode pose information so that it is easier for the network to learn and produce more satisfying result.

REFERENCES

[1] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[2] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[3] S. Sengupta, J.-C. Chen, C. D. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *Workshop on Applications of Computer Vision*, 2016, pp. 1–9.

[4] M. Kan, S. Shan, and X. Chen, "Multi-view deep network for cross-view classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4847–4855.

[5] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-view perceptron: a deep model for learning face identity and view representations," in *Advances in Neural Information Processing Systems*, 2014, pp. 217–225.

[6] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 676–684.

[7] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 113–120.

[8] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker, "Reconstruction-based disentanglement for pose-invariant face recognition," in *IEEE International Conference on Computer Vision*, 2017, pp. 1632–1641.

[9] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajan *et al.*, "Face recognition using deep multi-pose representations," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–9.

[10] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked progressive auto-encoders (spae) for face recognition across poses," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1883–1890, 2014.

[11] J. Yang, S. Reed, M. H. Yang, and H. Lee, "Weakly-supervised disentangling with recurrent transformations for 3d view synthesis," in *International Conference on Neural Information Processing Systems*, 2015, pp. 1099–1107.

[12] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 4, 2017, p. 7.

[13] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3990–3999.

[14] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2439–2448.

[15] Z. Zhang, X. Chen, B. Wang, G. Hu, W. Zuo, and E. R. Hancock, "Face frontalization using an appearance-flow-based convolutional neural network," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2187–2199, 2018.

[16] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8398–8406.

[17] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing *et al.*, "Towards pose invariant face recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2207–2216.

[18] 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*, 2018, pp. 2867–2877.

[19] L. Tran, X. Yin, and X. Liu, "Representation learning by rotating your faces." *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 12, pp. 3007–3021, 2019.

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

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

[22] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4873–4882.

[23] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world," *Electronic imaging*, vol. 2016, no. 11, pp. 1–6, 2016.

[24] Y. Qian, W. Deng, and J. Hu, "Unsupervised face normalization with extreme pose and expression in the wild," 2019.

[25] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[26] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[27] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," *Month*, 2007.

[28] B. F. Klare, B. Klein, E. Taborsky, and A. Blanton, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1931–1939.

[29] C. Ferrari, G. Lisanti, S. Berretti, and A. D. Bimbo, "Effective 3d based frontalization for unconstrained face recognition," in *International Conference on Pattern Recognition*, 2017.

[30] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 787–796.

[31] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4295–4304.

[32] C. Rong, X. Zhang, and Y. Lin, "Feature-improving generative adversarial network for face frontalization," *IEEE Access*, vol. 8, pp. 68 842–68 851, 2020.

[33] 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*, 2014, pp. 2672–2680.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 630–645.

[35] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 2018–2025.

[36] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[37] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *International Conference on Machine Learning*, vol. 30, 2013, p. 1.

[38] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *Computer Science*, 2015.

[39] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," *arXiv:1611.04076*, 2016.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[41] X. Zhu, Z. Lei, S. Z. Li *et al.*, "Face alignment in full pose range: A 3d total solution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[42] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[44] D. Wang, C. Otto, and A. K. Jain, "Face search at scale." *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2016.

[45] I. Masi, S. Rawls, G. G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4838–4846.

[46] X. Luan, H. Geng, L. Liu, W. Li, Y. Zhao, and M. Ren, "Geometry structure preserving based gan for multi-pose face frontalization and recognition," *IEEE Access*, 2020.

[47] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*. IEEE, 2016, pp. 1–8.

[48] J.-C. Chen, J. Zheng, V. M. Patel, and R. Chellappa, "Fisher vector encoded deep convolutional features for unconstrained face verification," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 2981–2985.