Pseudo Facial Generation with Extreme Poses for Face Recognition

Guoli Wang[1], Jiaqi Ma[2], Qian Zhang[3], Jiwen Lu[1,*], Jie Zhou[1]
[1] Tsinghua University [2] Wuhan University [3] Horizon Robotics
guoli.wang@mail.tsinghua.edu.cn, jiaqima@whu.edu.cn
qian01.zhang@horizon.ai, {lujiwen,jzhou}@tsinghua.edu.cn

Abstract

Face recognition has achieved a great success in recent years, it is still challenging to recognize those facial images with extreme poses. Traditional methods consider it as a domain gap problem. Many of them settle it by generating fake frontal faces from extreme ones, whereas they are tough to maintain the identity information with high computational consumption and uncontrolled disturbances. Our experimental analysis shows a dramatic precision drop with extreme poses. Meanwhile, those extreme poses just exist minor visual differences after small rotations. Derived from this insight, we attempt to relieve such a huge precision drop by making minor changes to the input images without modifying existing discriminators. A novel lightweight pseudo facial generation is proposed to relieve the problem of extreme poses without generating any frontal facial image. It can depict the facial contour information and make appropriate modifications to preserve the critical identity information. Specifically, the proposed method reconstructs pseudo profile faces by minimizing the pixel-wise differences with original profile faces and maintaining the identity consistent information from their corresponding frontal faces simultaneously. The proposed framework can improve existing discriminators and obtain a great promotion on several benchmark datasets.

1. Introduction

Face recognition aims to figure out the ground-truth identity for an unknown facial image. Inspired by the development of deep learning, the performance of face recognition has been improved rapidly [38, 40, 59, 60, 12, 10, 23]. LightCNN [47] is proposed as a superb model for face recognition recently and reaches new heights on several benchmark datasets. However, pose variations result in an unsolved problem in desired real-world applications. As is shown in Table 1, LightCNN recognizes faces well on the Multi-PIE dataset [13] in the range of $\pm 15^\circ$ and $\pm 45^\circ$, and suffers a slight drop in $\pm 60^\circ$ and $\pm 75^\circ$. But for extreme poses such as $\pm 90^\circ$, the recognition rate declines dramatically, where the decrease is up to 35.09%.

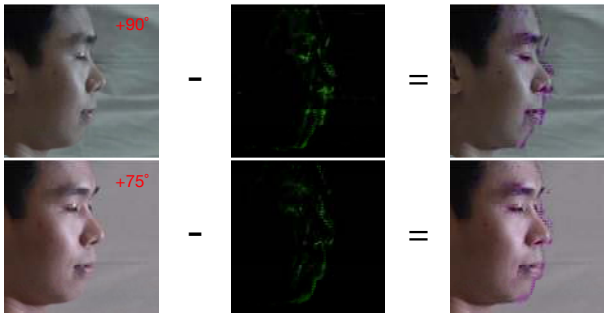

Figure 1. An illustration of the proposed pseudo facial generation. From left to right, the input faces, residual faces and pseudo faces are visualized in sequence. $+90^\circ$ and $+75^\circ$ faces are respectively displayed in the first and second lines.

Why does the recognition rate drop heavily when it comes to the extreme profile faces? Deep learning methods are data-driven and usually extract features with a propensity [3]. Generally speaking, the frontal and profile facial images lie in diverse domains. Models prefer to learn discriminative features from the dominant domain, and may fail in the subordinate domain. Both the imbalance of data distribution and domain gap devote to this aforementioned phenomenon, and researchers have tried several different paths to settle it [9]. The existing methods are mainly divided into two categories. One kind of them is to extract pose invariant embeddings from original faces to maintain the invariance [4, 30, 42]. The others rotate facial images to the frontal pose and recognize them directly [43, 52].

For those embedding based methods, metric learning and multi-view learning are usually applied to obtain pose invariant embedding features. Florian *et al*. [36] train a CNN to optimize the desired embeddings directly rather than an intermediate bottleneck layer. Kan *et al*. [21] consider the imbalance of data distribution as large discrepancies between views, and devise several adaptive sub-nets to release

*Jiwen Lu is the corresponding author.

the discrepancies. Wu *et al*. [48] propose a couple deep learning approach to discover a shared feature subspace, and the heterogeneous recognition problem can be approximately considered as a homogeneous face matching. Considering the imbalance of poses which may cause a long tail distribution problem, the performance of the obtained pose invariant embeddings are often unsatisfied.

For those rotation based methods, the first attempt can date back to 1990s by Roberto *et al*. [1] and Alex *et al*. [31]. Huang *et al*. [18] propose a Two-Pathway Generative Adversarial Network (TP-GAN) to generate realistic frontal faces. It is restricted by the huge computational consumption and the intention only for frontal images. Hu *et al*. [16] upgrade the TP-GAN to an arbitrary pose rotation as well as decreasing training and inference time. But a GAN-based frontal generation method needs to learn plenty of parameters for approaching the ground-truth images. Yin *et al*. [52] propose a multi-task problem to devise three sub-tasks: pose, illumination and emotion, then adjust weights between the main recognition task and three sub-tasks adaptively. Although multi-task is a shortcut to achieve better performance, it needs hand-craft for adjusting weights of sub-task networks. Cao *et al*. [3] improve the profile pose recognition rates by learning deep frontal residual mappings. Compared with our proposed method, this method mainly focuses on the transform of learned residual features, whereas our method concentrates on generating pseudo profile facial images from original inputs. Recently, a flow-based method named FFWM [43] also frontalizes facial images and reaches excellent scores on several benchmark datasets. FFWM contains a Waro Attention Module (WAM) and an Illumination Preserving Module (IPM), that can synthesize realistic and illumination preserved frontal faces. Unlike those former pose invariant models, FFWM generates frontal faces by estimating flows and can well recognize facial images with extreme poses.

Different from the above face frontalization based methods, we try to make minor pixel-wise changes to input facial images. This innovation derives from the observation that there exists a dramatic precision drop between $\pm 75^\circ$ and $\pm 90^\circ$ in Table 1, but $+75^\circ$ and $+90^\circ$ facial images in Fig. 1 may not be visually distinguished apparently. Compared with those traditional GAN-based methods, we reduce the number of parameters and flops by designing a novel lightweight generator. As is shown in Fig. 1, residual images mainly describe facial contour information and the generated pseudo facial images can make appropriate modifications to preserve the critical identity information. To illustrate its efficiency and expandable ability, the lightweight generator is applied to LightCNN-29-v2 and relieve the aforementioned phenomenon successfully. Our proposed framework can help any existing discriminator obtain a great promotion on several benchmark datasets without carrying too much burden. Quantitative and qualitative experiments demonstrate the efficiency and effectiveness of our proposed method. Comparing with the baseline LightCNN-29-v2, our method shows its appealing charm by its superb performance, and can be further added to any other high quality facial discriminator for a promising promotion. To conclude, our contributions can be summarized as the following points:

Pose	$\pm 15^\circ$	$\pm 30^\circ$	$\pm 45^\circ$	$\pm 60^\circ$	$\pm 75^\circ$	$\pm 90^\circ$
Rank-1	100.00	100.00	99.94	98.83	92.91	57.82

Table 1. Rank-1 recognition rates (%) on the Multi-PIE dataset under **Setting2** by the pretrained LightCNN-29-v2.

- We provide a novel, straight-forward and simple method to relieve the dramatic precision drop for extreme poses by generating pseudo profile facial images under minor pixel-wise modifications rather than generating fake frontalized faces.
- A lightweight pseudo profile facial generator is proposed as the front-end input of any existing facial discriminator. The inherent identity information can be well preserved by the generator at a low computational consumption.
- Quantitative and qualitative experimental results confirm that the proposed framework can perform better than the pre-trained discriminator. Specifically, it can achieve a surprising recognition rate of 93.68% for $\pm 90^\circ$ on the Multi-PIE dataset under **Setting2**.

2. Related Work

2.1. Face Frontalization

Face frontalization is a challenging synthesis problem, especially for extreme poses. Traditionally, we can divide this topic into the following categories: 2D/3D local texture warping [15, 63], statistic modeling [33] and methods based on deep learning [7, 18, 20, 51, 53, 32].

Zhao *et al*. [56, 57] propose the Pose Invariant Model (PIM) for face recognition in the wild, which contains both the face frontalization sub-net and the discriminative learning sub-net. In [58], 3D morphable model is utilized to assist a PIM-based model with prior knowledge. Cao *et al*. [2] discover a connection between 2D and 3D surface spaces and combine it with GAN.

Huang *et al*. [18] leverage the synthesized facial images which preserve identity information for face recognition and attribution estimation tasks. Hu *et al*. [16] propose CAPG-GAN to apply facial landmark heatmaps to the face rotation model for guidance in both training and inference

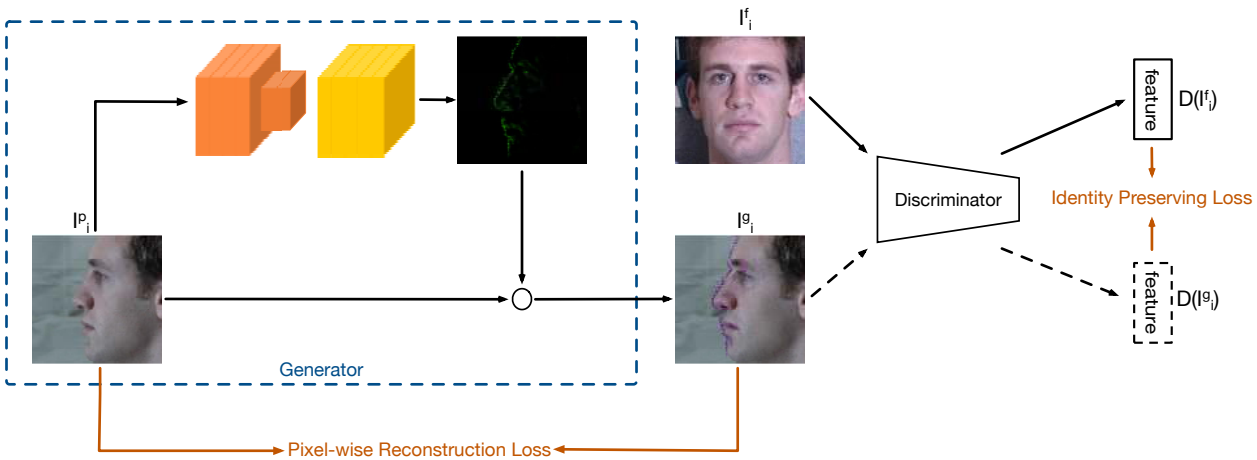

Figure 2. The pipeline of our proposed framework. It contains a pseudo profile facial generator and a discriminator. During the training stage, parameters of the discriminator are frozen. And the generator is supervised by the identity preserving loss and the pixel-wise reconstruction loss simultaneously.

stages. It represents a type of GAN-based models which need relative high computational costs for learning and generating fake facial images. Wei *et al*. [43] devise an Illumination Preserving Module to learn features from illumination inconsistent image pairs. As depicted in Table 6, the proposed FFWM achieves high Rank-1 recognition rates in those extreme poses on the Multi-PIE dataset, but fails to reach a comparable performance with other state-of-the-art methods under small poses. That is to say, those synthesized facial images may confront an information loss and will damage the identity preserving. To prevent from this kind of information loss and realize the reduction of computational consumption, a lightweight pseudo profile facial generator is proposed in this work.

2.2. Facial Feature Mapping

Facial feature mapping extracts meaningful feature maps to capture the geometric variations for recognition and detection. A series of feature map based methods rely on Spatial Transformer Network (STN) [19]. By aligning facial landmarks, Dong *et al*. [5] design a supervised transformer layer to frontalize facial images and improve the performance of face detection. Zhong *et al*. [61] also utilize the STN module and learn the optimized transform parameters by facial feature maps. Zhou *et al*. [62] propose a face rectification module called GridFace and perform an end-to-end training with a recognition module. However, GridFace has two problems. First, it optimizes both the face rectification and recognition module simultaneously, which is at high computational consumption. Second, similar to face frontalization, it faces a challenge to maintain identity information and is insignificant to rectify those extreme poses. Cao *et al*. [3] propose another feature mapping method for face recognition. It originates from the features learning from recognition models and projects the profile facial features to the frontal facial feature space.

3. Proposed Method

In this section, based on the observed phenomenon, we first give a definition of the profile face reconstruction problem and define the symbols used in our methodology in Sec. 3.1. Secondly, we illustrate the network architecture of our proposed lightweight **P**seudo **P**rofile **F**acial **G**enerator (**PPFG**) in Sec. 3.2. Finally, we elaborate the pipeline of the whole framework and formulate the overall loss functions in Sec. 3.3.

3.1. Problem Formulation

One typical existing method for face recognition is LightCNN, which is proposed by Wu *et al*. [47]. As is shown in Table 1, the LightCNN can get a good result on the Multi-PIE dataset when the facial poses is in the range of $0°$ to $\pm 75°$. But for profile facial images under $\pm 90°$, the Rank-1 recognition rate drops heavily to 57.82% on Multi-PIE. To make a further analysis, we investigate on large amounts of faces with extreme poses to compare their differences from an intuitive view. Fig. 1 shows two examples under $+90°$ and $+75°$, respectively. It can be clearly found that the pair between $+90°$ and $+75°$ exists just a little visual disparity, but their overall Rank-1 recognition rate has a 35.09% drop.

To reduce this dramatic drop, traditional methods generally design generators to get fake frontal faces, and extract embedding features of those unrealistic faces. However, retraining a face generator costs a lot, and considering its huge flops, this kind of method is hard to be deployed

	Input Resolution	Operation	Channel	Stride
Encoder	128×128×3	Conv3×3 Block	24	1
	128×128×24	Inverted Residual Block	24	1
	128×128×24	Inverted Residual Block	24	1
	128×128×24	Inverted Residual Block	48	2
	64×64×48	Inverted Residual Block	48	1
	64×64×48	Inverted Residual Block	48	1
Decoder	64×64×48	Deconv2×2 Block	24	2
	128×128×24	Inverted Residual Block	12	1
	128×128×12	Inverted Residual Block	12	1
	128×128×12	Inverted Residual Block	12	1
	128×128×12	Conv3×3 Block	12	1
	128×128×12	Conv2D, 1×1	3	1
	128×128×3	Sigmoid	3	1
Processor	128×128×3	Subtraction	3	1

Table 2. The architecture of the proposed Pseudo Profile Facial Generator G.

in real-world scenes. We handle this tough question in an innovative way. What if we reconstruct pseudo profile faces by minimizing the pixel-wise differences with original profile faces and maintaining the identity consistent information from their corresponding frontal faces simultaneously? The input facial image will be revised by minor changes with a little computational consumption and the precision drop phenomenon will be relieved finally.

Let $\mathbb{P}_{data}$ be a dataset which contains facial images within both frontal and profile poses. We assume that a facial image with a 0°pose of the identity i is a frontal image I_i^f . And a face with other poses of the identity i is defined as a profile image I_i^p . Given a pair of frontal and profile facial images of the identity i sampled from $\mathbb{P}_{data}$ — $\{I_i^f, I_i^p\}$, our goal is to train a generator G to synthesize the corresponding pseudo profile facial image $I_i^g = G(I_i^p)$ from its original input I_i^p. The generated I_i^g is expected to own a similar semantic embedding feature with a frontal face I_i^f and tends to be as close as to the original profile face I_i^p at the same time.

To reach this goal, we propose a lightweight pseudo profile facial generator G as the front-end input. Any face recognition discriminator D can be applied in our framework as a back-end and achieve better face recognition performance. The pipeline of our proposed framework is illustrated in Fig. 2.

3.2. Lightweight Pseudo Profile Facial Generator

Traditional face recognition methods [18, 32, 53] usually utilize those profile poses to reconstruct frontal poses. GAN is used to establish the pixel-level connections between the profile and frontal faces [42, 54]. Those methods fill in the domain gap by synthesizing frontal facial images. Their computational costs are high and those synthetic images may lose some important identity information.

According to the above reasons, we devise a novel lightweight pseudo profile facial generator to actually settle this problem. Following the rule of 'simple is the best', the architecture of the generator is designed as efficient as possible. It mainly consists of several stacked inverted residual blocks introduced in MobileNetV2 [34]. The expand ratio is set as 1. The whole architecture of the proposed generator network G is illustrated in Table 2. Conv3x3 Block means a 3×3 kernel-sized convolutional layer with a leaky ReLU and Deconv2x2 Block means a 2×2 kernel-sized deconvolutional layer with a leaky ReLU. Conv2D, 1x1 denotes a 1×1 kernel-sized convolutional layer. Channel represents the output channel size and stride means the sliding step. The channel size is at most 48 with a relative shallow depth. As depicted in Table 3, compared with CAPG-GAN [16] which has 79.59M parameters and 62G FLOPs, our proposed generator has only 16.10K parameters with 103.56M FLOPs.

Methods \ Metrics	Parameters	FLOPs
CAPG-GAN [16]	79.59M	62G
LightCNN-29-v2 [47]	10.4M	3.6G
Our Generator G	16.10K	103.56M

Table 3. Parameters and FLOPs comparison between existing methods and the proposed Pseudo Profile Facial Generator G.

The generator is divided into three parts: Encoder, Decoder and Pixel-wise Processor. By directly subtracting those residual images from original profile facial images, we generate the final pseudo profile faces as Fig. 1. From Fig. 3, it is clear that residual images can describe partial outline information of facial images. The pseudo profile facial images are generated from them. The architecture of the generator is represented in the blue box of Fig. 2.

If we apply the frontal poses as reconstruct supervision information, the proposed PPFG will try to generate pseudo frontal facial images. It will belong to the face frontalization based method as well as causing dramatic geometry and pixel-wise changes, which increases the difficulty to train a good generator.

3.3. Pipeline Description

Fig. 2 depicts the whole pipeline of our proposed framework. Similar to GAN, our proposed method includes a generator and a discriminator. However, it must be emphasized that the discriminator is frozen during the training stage, and the parameters of the proposed generator are updated by the pixel-wise reconstruction loss and identity preserving loss.

3.3.1 Pixel-wise Reconstruction Loss

In the training stage, one original profile facial image I_i^p is firstly fed into the generator $G(.)$ to obtain a pseudo profile facial image as:

$$I_i^g = G(I_i^p). \tag{1}$$

We employ a pixel-wise reconstruction loss to hold the content consistency between the original profile image and its corresponding generated pseudo image as:

$$\mathcal{L}_{Recon} = \|I_i^g - I_i^p\|_1. \tag{2}$$

3.3.2 Identity Preserving Loss

To preserve the identity consistent information, we minimize the embedding distances between the generated pseudo ones and their corresponding frontal facial images. Suppose $D(.)$ denotes a embedding feature from a facial discriminator, the identity preserving loss is formulated as:

$$\mathcal{L}_{Id} = \|D(I_i^g) - D(I_i^f)\|_2^2. \tag{3}$$

It forces the embedding features of pseudo profile facial images to be as close as their frontal images. In this paper, LightCNN-29-v2 [47] is chosen as the discriminator due to its good trade-off between recognition accuracy and FLOPs. It is also noted that any other discriminators can be well combined with the proposed generator for different requirements.

3.3.3 Overall Loss

Based on the above losses, we reformulate the overall loss function as follows:

$$\begin{aligned}\mathcal{L}_{overall} &= \mathcal{L}_{Recon} + \gamma\mathcal{L}_{Id} \\ &= \|I_i^g - I_i^p\|_1 + \gamma\|D(I_i^g) - D(I_i^f)\|_2^2,\end{aligned} \tag{4}$$

where γ denotes a trade-off parameter between the pixel-wise reconstruction loss and the identity preserving loss. In Sec. 4.2, we discuss the influence of γ and set it according to the experimental results on Multi-PIE.

From Table 3, it can be found that both the parameters and FLOPs of our proposed generator are far less than LighCNN-29-v2. Therefore, it is possible to apply our proposed generator in practical situations.

4. Experiments

We evaluate the proposed framework qualitatively and quantitatively on several benchmark datasets. In the following parts, datasets for our experiments are introduced firstly and followed by some implementation details in the training stage. Secondly, we choose the hyper-parameter γ by some experimental evaluation. Finally, we demonstrate the superiority of our method on both qualitative and quantitative results and reveal the benefits of the proposed PPFG.

4.1. Datasets and Experimental Settings

Dataset Details: Six benchmark datasets: Multi-PIE [13], LFW [17], CFP [37], IJB-B [45], IJB-C [27] and MegaFace [22] are adopted to evaluate our method.

γ \ Poses	±15°	±30°	±45°	±60°	±75°	±90°	AVG
25	**100.00**	**100.00**	99.94	99.20	95.63	79.62	95.73
100	**100.00**	**100.00**	99.85	99.13	**96.07**	85.91	96.83
1000	**100.00**	**100.00**	**99.98**	**99.41**	96.04	87.35	**97.13**
10000	**100.00**	**100.00**	99.94	99.17	95.61	**87.37**	97.02

Table 4. Rank-1 recognition rates (%) with different γ settings on the Multi-PIE dataset. The best results are highlighted in **bold**.

Multi-PIE [13] is utilized as our training and testing set. It consists of multiple Pose, Illumination and Expression (PIE) variants in a controlled setting. Therefore, it is widely applied to evaluate the performance on PIE invariant face recognition and synthesis. The whole dataset includes 754,204 facial images of 337 identities from 20 illuminations and 15 poses. To compare with other methods in a fair condition, we only select 13 poses in the range of $\pm 90°$ with the neutral expression and carry out the same setting as [43]. In particular, we employ two kinds of different settings denoted as $\mathbf{Setting1}$ and $\mathbf{Setting2}$, respectively. For $\mathbf{Setting1}$, only images in Session 1 are exploited. All the first 150 identities are used to compose the training set. The testing set consists of a gallery set and a probe set. The remaining 99 identities in frontal pose and normal illumination are selected as the gallery set and the others are selected as the probe set. For $\mathbf{Setting2}$, we select images which are in a neutral expression from all four sessions. The first 200 identities are chosen for training and the remaining 137 identities are for testing. The composition of the testing set is similar to $\mathbf{Setting1}$.

The Celebrities in Frontal-Profile (CFP) dataset [37] contains 500 identities in the wild. Following the standard setting as [3], we evaluate our method on CFP by a 10-fold protocol. Both the Frontal-Frontal and Frontal-Profile experimental evaluations are conducted on the proposed and other comparable methods. LFW [17] consists of 13,233 facial images in the wild. It can be used to evaluate the frontalization performance in an uncontrolled setting. Same as [47], 6,000 facial image pairs are applied to evaluate the methods. IJB-B [45] contains 1,845 subjects with 21,798 still images and 55,026 frames from 7,011 videos. IJB-C [27] adds 1,661 new subjects to IJB-B. It contains 31,334 still images and 117,542 frames from 11,779 videos. As in [8], we report TAR (@FAR=1e-4) results for the 1:1 verification protocol. MegaFace [22] includes one million photos of more than 690K individuals as the gallery set and 100K photos of 530 individuals from FaceScrub [28] as the probe.

Note that CFP, LFW, IJB-B, IJB-C and MegaFace are just considered as the testing sets and we use the MS-Celeb-1M dataset [14] to train our proposed PPFG.

Implementation Details: All the inputs are cropped and resized to a fixed size of 128×128. LightCNN-29-v2 [47] is chosen as our facial discriminator and is pretrained on

Method	±15°	±30°	±45°	±60°	±75°	±90°	AVG
TP-GAN [18]	99.78	99.85	98.58	92.93	84.10	64.03	89.88
CAPG-GAN [16]	99.95	99.37	98.28	93.74	87.40	77.10	92.64
PIM [56]	99.80	99.40	98.30	97.70	91.20	75.00	93.57
3D-PIM [58]	99.83	99.47	99.34	98.84	94.34	76.12	94.66
FNM [32]	99.90	99.50	98.20	93.70	81.30	55.80	88.07
FFWM [43]	**100.00**	**100.00**	**100.00**	98.86	96.54	88.55	97.33
LightCNN-29-v2 [47]	**100.00**	**100.00**	99.97	99.44	95.25	62.40	92.84
Ours	**100.00**	**100.00**	99.95	99.49	97.98	88.74	97.69
LightCNN-29-v2* [47]	**100.00**	**100.00**	**100.00**	99.85	**99.04**	92.20	98.52
Ours*	**100.00**	**100.00**	**100.00**	**99.87**	98.94	**94.07**	**98.81**

Table 5. Rank-1 recognition rates (%) on Multi-PIE under **Setting1**. The best results are highlighted in **bold**. The symbol of $*$ represents that LightCNN-29-V2 is finetuned on Multi-PIE.

Method	±15°	±30°	±45°	±60°	±75°	±90°	AVG
DR-GAN [42]	94.00	90.10	86.20	83.20	-	-	88.38
FF-GAN [53]	94.60	92.50	89.70	85.20	77.20	61.20	83.40
TP-GAN [18]	98.68	98.06	95.38	87.72	77.43	64.64	86.99
CAPG-GAN [16]	99.82	99.56	97.33	90.63	83.05	66.05	89.41
PIM [56]	99.30	99.00	98.50	98.10	95.00	86.50	96.07
3D-PIM [58]	99.64	99.48	98.81	98.37	95.21	86.73	96.37
HF-PIM [2]	99.99	99.98	99.98	99.14	96.40	92.32	97.97
DA-GAN [54]	99.98	99.88	99.15	97.27	93.24	81.56	95.18
FFWM [43]	99.86	99.80	99.37	98.85	97.20	93.17	98.04
MVF-HQ [11]	99.9	99.9	99.4	98.7	96.3	87.4	96.93
LightCNN-29-v2 [47]	**100.00**	**100.00**	99.94	98.83	92.91	57.82	91.58
Ours	**100.00**	**100.00**	**99.98**	**99.41**	96.04	87.35	97.13
LightCNN-29-v2* [47]	99.94	99.91	99.80	98.92	96.72	91.97	97.88
Ours*	99.96	99.92	99.83	99.39	**97.60**	**93.68**	**98.40**

Table 6. Rank-1 recognition rates (%) on Multi-PIE under **Setting2**. The best results are highlighted in **bold**. The symbol of $*$ represents that LightCNN-29-V2 is finetuned on Multi-PIE.

Method	Frontal-Frontal		Frontal-Profile	
	ACC(%)	EER	ACC(%)	EER
Sengupta *et al.* [37]	96.40 ± 0.69	3.48 ± 0.67	84.91 ± 1.82	14.97 ± 1.98
Sankarana *et al.* [35]	96.93 ± 0.61	2.51 ± 0.81	89.17 ± 2.35	8.85 ± 0.99
Chen *et al.* [6]	98.67 ± 0.36	1.40 ± 0.37	91.97 ± 1.70	8.00 ± 1.68
DR-GAN [42]	97.84 ± 0.79	-	93.41 ± 1.17	-
DREAM [3]	-	-	-	6.02
Human	96.24±0.67	5.34± 1.79	94.57 ± 1.10	5.02 ± 1.07
LightCNN-29-v2 [47]	99.51 ± 0.44	0.31 ± 0.35	92.94 ± 2.00	6.06 ± 1.35
Ours	**99.60 ±0.43**	**0.23 ± 0.25**	**94.10 ± 2.30**	**5.70 ± 1.60**

Table 7. Face recognition accuracy (ACC) and equal-error-rate (EER) results on CFP. The best results are highlighted in **bold**. The perceptional results of human beings are highlighted in blue.

the MS-Celeb-1M dataset [14]. During the training stage, all parameters of the discriminator are frozen. The learning rate is initialized as 0.0002 with 512 batch size. The momentum and the weight decay are set to 0.9 and 0.0001, respectively. We train the generator with 100 epochs. According to Sec. 4.2, we empirically set γ as 1,000 for better recognition accuracy.

For Multi-PIE, facial images under 0°are chosen as the frontal faces and the others are considered as the profile faces during the training stage. For MS-Celeb-1M, the location of nose landmark decides whether the facial image belongs to a frontal one or not. In particular, we empirically select frontal facial images whose x-axis value ranging from 65 to 79 as well as y-axis value ranging from 79 to 86.

To further analyze the effectiveness of our method, we also finetune LightCNN-29-v2 on the training set of Multi-PIE. The basic learning rate is set as 0.01 with 1,024 batch size and the total epoch number is 120.

4.2. Parameter Selection

The hyper-parameter γ is used to balance the trade-off between the pixel-wise reconstruction loss and identity preserving loss. We conduct an experiment on the Multi-PIE dataset under **Setting2** to choose the best γ for the following training. γ ranges from 25 to 10,000. For a fair comparison, all PPFG models are respectively trained with the pretrained LightCNN-29-v2 and selected from the 100-th epoch and evaluated on the testing set. According to Table 4, with the increase of γ, the average Rank-1 recognition rate rises rapidly. However, there is a slight decrease when γ is set as 10,000. As a whole, γ is set to 1,000 and the following experiments will apply this setting.

4.3. Quantitative Evaluation

Table 5 illustrates the Rank-1 recognition rates on the Multi-PIE dataset under **Setting1**. It can be found that FFWM [43] outperforms the previous frontalization methods [16, 18, 32, 56, 58] and achieves a 88.55% Rank-1 recognition rate under ±90°. As a comparison, the pretrained LightCNN-29-v2 only gets a poor recognition rate at 62.40%. However, once combining the pre-trained LightCNN-29-v2 with our proposed PPFG, the recognition rate can go up to 88.74%. After finetuning LightCNN-29-v2 on Multi-PIE, our proposed method can even achieve a surprising 94.07% recognition rate under ±90°, which outperforms FFWM with a 5.52% recognition rate promotion. The average Rank-1 recognition rate of our proposed method also ascends to the best at 98.81%.

Table 6 further lists the experimental results on Multi-PIE under **Setting2**. It can be seen that LightCNN-29-v2 without finetuning on Multie-PIE gets a poor result, our proposed method can still get a tolerable performance, whose average recognition rate is just lower than FFWM [43] and HF-PIM [2]. After finetuning, our proposed method achieves the best average Rank-1 recognition rate at 98.40%.

Considering the experimental results in Table 5 and Table 6 simultaneously, we can find several interesting phenomena. First, compared with a single LightCNN-29-v2 discriminator, a discriminator combined with our proposed generator can acquire better average recognition rates. In particular, our method can get a comparable or even slight better performance than LightCNN-29-v2 ranging from −45° to +45°. When it comes to extreme poses such as ±75° and ±90°, our proposed method brings in a great promotion. It reveals the effectiveness of our proposed method. Second, compared with all the frontalization methods, our

Method	ACC (%)
DeepFace [40]	97.35
VGGFace [29]	99.13
FaceNet [36]	99.63
DeepID2+ [39]	99.47
WST Fusion [41]	98.37
SphereFace [26]	99.42
RangeLoss [55]	99.52
HiReSR-9+ [46]	99.03
FF-GAN [53]	96.42
CAPG-GAN [16]	99.37
DA-GAN [54]	99.56
R100, ArcFace [43]	**99.83**
LightCNN-29-v2 [47]	99.60
Ours	99.62 (+0.02)

Table 8. Face recognition accuracy (ACC) results on LFW. The best results are highlighted in **bold**. Compared with the baseline, the improvement of our proposed method is highlighted in blue.

Method	IJB-B (%)	IJB-C (%)
ResNet50 [4]	78.4	82.5
SENet50 [4]	80.0	84.0
ResNet50+SENet50 [4]	80.0	84.1
MN-v [50]	81.8	85.2
MN-vc [50]	83.1	86.2
ResNet50+DCN(Kpts) [49]	85.0	86.7
ResNet50+DCN(Divs) [49]	84.1	88.0
SENet50+DCN(Kpts) [49]	84.6	87.4
SENet50+DCN(Divs) [49]	84.9	88.5
VGG2,R50,ArcFace [8]	89.8	92.1
MS1MV2,R100,ArcFace [8]	**94.2**	**95.6**
LightCNN-v2-29	85.93	87.99
Ours (LightCNN-v2-29)	87.00 (+1.07)	89.02 (+1.03)
ResNet50-SE	89.45	91.35
Ours (ResNet50-SE)	89.91 (+0.46)	92.03 (+0.68)

Table 9. 1:1 verification TAR (@FAR=1e-4) on IJB-B and IJB-C. The best results are highlighted in **bold**. Our improvements over the pre-trained discriminators are highlighted in blue.

Methods	Data Refinement	
	No (%)	Yes (%)
Softmax [25]	54.85	-
Contrastive Loss [25]	65.21	-
Triplet [25]	64.79	-
Center Loss [44]	65.49	-
SphereFace [25]	72.73	-
CosFace [8]	**82.72**	-
AM-Softmax [8]	72.47	-
SphereFace+ [24]	73.03	-
FaceNet [36]	70.49	-
CASIA, R50, ArcFace [8]	77.50	92.34
MS1MV2, R100, ArcFace [8]	81.03	**98.35**
LightCNN-v2-29	72.30	86.49
Ours (LightCNN-v2-29)	73.15 (+0.85)	87.39 (+0.9)
ResNet50-SE	76.73	92.92
Ours (ResNet50-SE)	77.46 (+0.73)	93.70 (+0.78)

Table 10. The rank-1 face identification accuracy on MegaFace Challenge 1. Data refinement removes noises in MegaFace [8]. The best results are highlighted in **bold**. Our improvements over the pre-trained discriminators are highlighted in blue.

method achieves better performance on these two settings. It implies that minor pixel-wise modifications for input images may be a better choice for face recognition with extreme poses rather than generating fake frontalized faces.

For further evaluation, Table 7 refers to the face recognition performance (ACC and EER) of our method with other state-of-the-art methods and even human beings perception on CFP. For ACC, a higher score means a better performance. On contrast, a lower EER is better. Under the frontal-frontal setting, our proposed method performs the best in all metrics, and even can recognize more accurate than human beings. The proposed method decreases more than 5% for the EER rate and gains a accuracy increase more than 3%. Under the frontal-profile setting, our method beats all other state-of-the-art methods, but still exists a tiny gap with Human beings.

Table 8 lists the face recognition accuracy on the LFW dataset while Table 9 refers to TAR (@FAR=1e-4) on IJB-B and IJB-C for the 1:1 verification protocol. The rank-1 face identification accuracy results on MegaFace Challenge 1 are listed in Table 10. Although the frontal facial images dominate these datasets, which is inconsistent with the initial desired scene of our proposed method. Our proposed method can still achieve competitive results on these datasets. Note that ResNet100 with ArcFace [8] has much more parameters than our methods based on LightCNN-v2-29 and ResNet50-SE. Our proposed method with ResNet50-SE obtains the second highest results on IJB-B and MegaFace with data refinement. Furthermore, the experimental results of both LightCNN-v2-29 and ResNet50-SE on IJB-B, IJB-C and MegaFace confirm that any existing face recognition discriminator can be integrated to the front-end generator and achieve better performance.

4.4. Qualitative Evaluation

In this subsection, we qualitatively compare our original facial images, residual images and corresponding generated images. The selected images are all from the 201-th identity in Multi-PIE and we make a comparison in Fig. 3. In order to better visualize residual images, we remap its color space to generate more clear images in the second row.

Note that for those extreme poses, the generated pseudo facial images are modified by some minor changes according to the residual images. It can be observed that residual images mainly describe the contour information of faces. As is shown in Fig. 3, the residual images under -90° approximately depict the shapes of eye, nose and mouth. After subtracting residual images from the original inputs, the generated pseudo facial images can make appropriate modifications to preserve the critical identity information. Therefore, our method gains a remarkable promotion for extreme

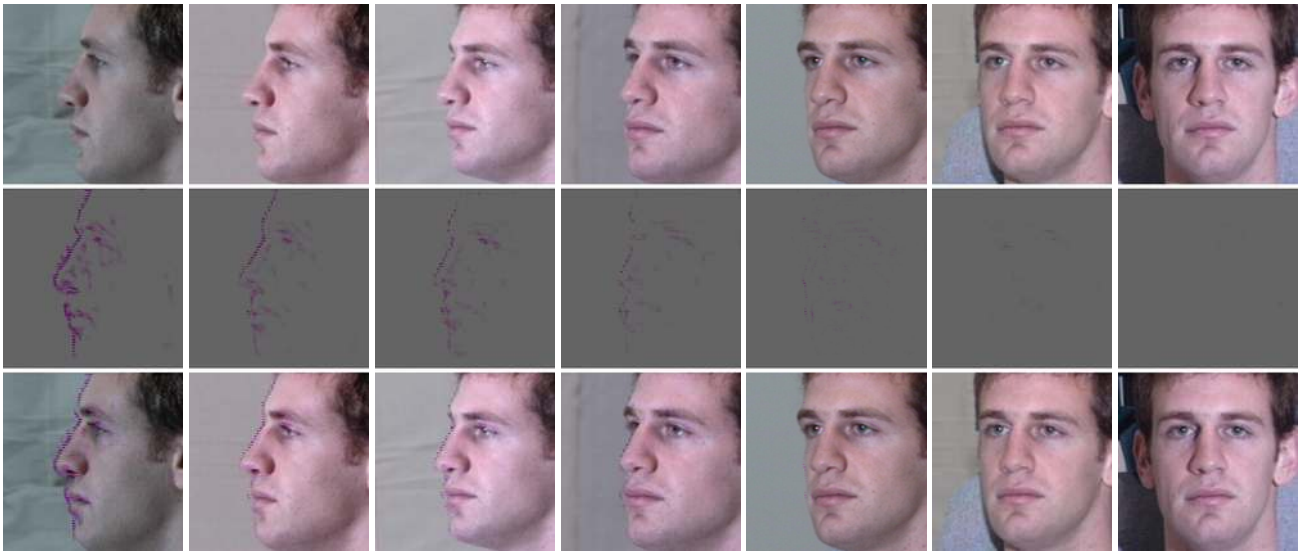

Figure 3. From top to bottom, it is a comparison of original facial images, visualized residual images and the generated pseudo facial images. From left to right, the faces turn from -90° to 0° at an interval of 15°.

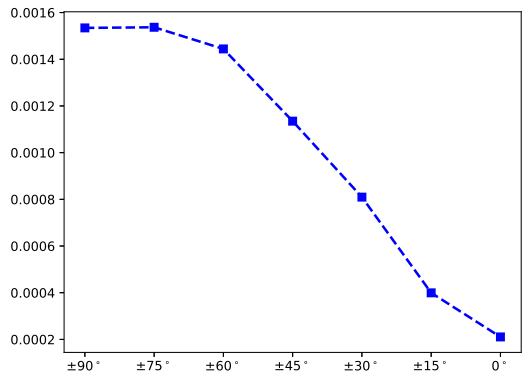

Figure 4. The mean values of residual images with different poses on Multi-PIE.

poses. Although the generated pseudo facial images are not friendly to perception of human beings, they can be well understood by facial discriminators.

It can be seen that from left to right, the residual images contain gradually reduced information from -90° to 0°. We can not even distinguish the visual differences between the generated pseudo frontal face and original one in Fig. 3. The phenomenon can be further confirmed by Fig. 4. We calculate the mean values of residual images with different poses, and find that the mean values are closely related to poses. Input images under $\pm 60^\circ$, $\pm 75^\circ$ and $\pm 90^\circ$ get higher mean values. When faces tend to be frontal, their corresponding residual images have smaller mean values. It reflects that our proposed method tries to maintain the original features of those frontal faces. As a result, the generated pseudo frontal faces just achieve a slight promotion or keep a comparable performance with input frontal faces.

5. Conclusion

In this paper, we propose a lightweight pseudo profile facial generator to reconstruct facial images with extreme poses, and apply it to a pre-trained face recognition discriminator. Compared with other GAN-based methods, our proposed method needs less computational consumption and reaches higher accuracy. Different from those face frontalization models, we just make some minor changes to the original inputs and generate pseudo profile faces. As is pointed out in this paper, any existing face recognition discriminator can be integrated to the proposed PPFG frontend for better results. Experimental results on six benchmark datasets sufficiently confirm the effectiveness of our proposed method on the face recognition task. It is noteworthy that our method is not limited to face recognition with extreme poses, and those tasks suffer from the domain gap can also benefit from it, such as the occlusion and makeup problems. We will further explore these problems in the future.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant Grant U1813218, 61822603, Grant U1713214, in part by Beijing Academy of Artificial Intelligence (BAAI), and in part by a grant from the Institute for Guo Qiang, Tsinghua University.

References

[1] Roberto Brunelli and Tomaso A. Poggio. Face recognition: Features versus templates. *TPAMI*, 1993. 2

[2] Jie Cao, Yibo Hu, Hongwen Zhang, Ran He, and Zhenan Sun. Learning a high fidelity pose invariant model for high-resolution face frontalization. In *NeurIPS*, 2018. 2, 6

[3] Kaidi Cao, Yu Rong, Cheng Li, Xiaoou Tang, and Chen Change Loy. Pose-robust face recognition via deep residual equivariant mapping. In *CVPR*, 2018. 1, 2, 3, 5, 6

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018. 1, 7

[5] Dong Chen, Gang Hua, Fang Wen, and Jian Sun. Supervised transformer network for efficient face detection. In *ECCV*, 2016. 3

[6] Jun-Cheng Chen, Jingxiao Zheng, Vishal M. Patel, and Rama Chellappa. Fisher vector encoded deep convolutional features for unconstrained face verification. In *ICIP*, 2016. 6

[7] Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T. Freeman. Synthesizing normalized faces from facial identity features. In *CVPR*, 2017. 2

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 5, 7

[9] Changxing Ding and Dacheng Tao. A comprehensive survey on pose-invariant face recognition. *ACM TIST*, 2016. 1

[10] Boyan Duan, Chaoyou Fu, Yi Li, Xingguang Song, and Ran He. Cross-spectral face hallucination via disentangling independent factors. In *CVPR*, pages 7927–7935, 2020. 1

[11] Chaoyou Fu, Yibo Hu, Xiang Wu, Guoli Wang, Qian Zhang, and Ran He. High-fidelity face manipulation with extreme poses and expressions. *TIFS*, 16:2218–2231, 2021. 6

[12] Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, and Ran He. Dual variational generation for low shot heterogeneous face recognition. In *NeurIPS*, pages 2670–2679, 2019. 1

[13] Ralph Gross, Iain A. Matthews, Jeffrey F. Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image Vis. Comput.*, 2010. 1, 5

[14] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 5, 6

[15] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2015. 2

[16] Yibo Hu, Xiang Wu, Bing Yu, Ran He, and Zhenan Sun. Pose-guided photorealistic face rotation. In *CVPR*, 2018. 2, 4, 6, 7

[17] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007. 5

[18] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In *ICCV*, 2017. 2, 4, 6

[19] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NeurIPS*, 2015. 3

[20] Meina Kan, Shiguang Shan, Hong Chang, and Xilin Chen. Stacked progressive auto-encoders (SPAE) for face recognition across poses. In *CVPR*, 2014. 2

[21] Meina Kan, Shiguang Shan, and Xilin Chen. Multi-view deep network for cross-view classification. In *CVPR*, 2016. 1

[22] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, pages 4873–4882, 2016. 5

[23] Jianshu Li, Pan Zhou, Yunpeng Chen, Jian Zhao, Sujoy Roy, Shuicheng Yan, Jiashi Feng, and Terence Sim. Task relation networks. In *WACV*, pages 932–940, 2019. 1

[24] Weiyang Liu, Rongmei Lin, Zhen Liu, Lixin Liu, Zhiding Yu, Bo Dai, and Le Song. Learning towards minimum hyperspherical energy. In *NeurIPS*, pages 6225–6236, 2018. 7

[25] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, pages 6738–6746, 2017. 7

[26] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017. 7

[27] Brianna Maze, Jocelyn C. Adams, James A. Duncan, Nathan D. Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. IARPA janus benchmark - C: face dataset and protocol. In *ICB*, pages 158–165, 2018. 5

[28] Hongwei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *ICIP*, pages 343–347, 2014. 5

[29] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, 2015. 7

[30] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N. Metaxas, and Manmohan Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *ICCV*, 2017. 1

[31] Alex Pentland, Baback Moghaddam, and Thad Starner. View-based and modular eigenspaces for face recognition. In *CVPR*, 1994. 2

[32] Yichen Qian, Weihong Deng, and Jiani Hu. Unsupervised face normalization with extreme pose and expression in the wild. In *CVPR*, 2019. 2, 4, 6

[33] Christos Sagonas, Yannis Panagakis, Stefanos Zafeiriou, and Maja Pantic. Robust statistical frontalization of human and animal faces. *IJCV*, 2017. 2

[34] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 4

[35] Swami Sankaranarayanan, Azadeh Alavi, Carlos Domingo Castillo, and Rama Chellappa. Triplet probabilistic embedding for face verification and clustering. In *BTAS*, 2016. 6

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1, 7

[37] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Domingo Castillo, Vishal M. Patel, Rama Chellappa, and David W. Jacobs. Frontal to profile face verification in the wild. In *WACV*, 2016. 5, 6

[38] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *NeurIPS*, 2014. 1

[39] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *CVPR*, 2015. 7

[40] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1, 7

[41] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Web-scale training for face identification. In *CVPR*, 2015. 7

[42] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *CVPR*, 2017. 1, 4, 6

[43] Yuxiang Wei, Ming Liu, Haolin Wang, Ruifeng Zhu, Guosheng Hu, and Wangmeng Zuo. Learning flow-based feature warping for face frontalization with illumination inconsistent supervision. In *ECCV*, 2020. 1, 2, 3, 5, 6, 7

[44] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, volume 9911, pages 499–515, 2016. 7

[45] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn C. Adams, Tim Miller, Nathan D. Kalka, Anil K. Jain, James A. Duncan, Kristen Allen, Jordan Cheney, and Patrick Grother. IARPA janus benchmark-b face dataset. In *CVPR*, pages 592–600, 2017. 5

[46] Wanglong Wu, Meina Kan, Xin Liu, Yi Yang, Shiguang Shan, and Xilin Chen. Recursive spatial transformer (rest) for alignment-free face recognition. In *ICCV*, 2017. 7

[47] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light CNN for deep face representation with noisy labels. *TIFS*, 2018. 1, 3, 4, 5, 6, 7

[48] Xiang Wu, Lingxiao Song, Ran He, and Tieniu Tan. Coupled deep learning for heterogeneous face recognition. In *AAAI*, 2018. 2

[49] Weidi Xie, Li Shen, and Andrew Zisserman. Comparator networks. In *ECCV*, volume 11215, pages 811–826, 2018. 7

[50] Weidi Xie and Andrew Zisserman. Multicolumn networks for face recognition. In *BMVC*, page 111, 2018. 7

[51] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Du-Sik Park, and Junmo Kim. Rotating your face using multi-task deep neural network. In *CVPR*, 2015. 2

[52] Xi Yin and Xiaoming Liu. Multi-task convolutional neural network for pose-invariant face recognition. *TIP*, 2018. 1, 2

[53] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017. 2, 4, 6, 7

[54] Yu Yin, Songyao Jiang, Joseph P. Robinson, and Yun Fu. Dual-attention GAN for large-pose face frontalization. In *FG*, pages 249–256, 2020. 4, 6, 7

[55] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *ICCV*, 2017. 7

[56] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Jayashree Karlekar, Sugiri Pranata, Shengmei Shen, Junliang Xing, Shuicheng Yan, and Jiashi Feng. Towards pose invariant face recognition in the wild. In *CVPR*, 2018. 2, 6

[57] Jian Zhao, Junliang Xing, Lin Xiong, Shuicheng Yan, and Jiashi Feng. Recognizing profile faces by imagining frontal view. *IJCV*, 128(2):460–478, 2020. 2

[58] Jian Zhao, Lin Xiong, Yu Cheng, Yi Cheng, Jianshu Li, Li Zhou, Yan Xu, Jayashree Karlekar, Sugiri Pranata, Shengmei Shen, Junliang Xing, Shuicheng Yan, and Jiashi Feng. 3d-aided deep pose-invariant face recognition. In *IJCAI*, 2018. 2, 6

[59] Jian Zhao, Lin Xiong, Jayashree Karlekar, Jianshu Li, Fang Zhao, Zhecan Wang, Sugiri Pranata, Shengmei Shen, Shuicheng Yan, and Jiashi Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *NeurIPS*, pages 66–76, 2017. 1

[60] Jian Zhao, Lin Xiong, Jianshu Li, Junliang Xing, Shuicheng Yan, and Jiashi Feng. 3d-aided dual-agent gans for unconstrained face recognition. *TPAMI*, 41(10):2380–2394, 2019. 1

[61] Yuanyi Zhong, Jiansheng Chen, and Bo Huang. Toward end-to-end face recognition through alignment learning. *SPL*, 2017. 3

[62] Erjin Zhou, Zhimin Cao, and Jian Sun. Gridface: Face rectification via learning local homography transformations. In *ECCV*, 2018. 3

[63] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, 2015. 2