Towards Pose Invariant Face Recognition in the Wild

Jian Zhao[1,2*†] Yu Cheng[3] Yan Xu[4] Lin Xiong[4] Jianshu Li[1] Fang Zhao[1]
Karlekar Jayashree[4] Sugiri Pranata[4] Shengmei Shen[4] Junliang Xing[5] Shuicheng Yan[1,6] Jiashi Feng[1]
[1]National University of Singapore [2]National University of Defense Technology [3]Nanyang Technological University
[4]Panasonic R&D Center Singapore [5]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[6]Qihoo 360 AI Institute

Abstract

Pose variation is one key challenge in face recognition. As opposed to current techniques for pose invariant face recognition, which either directly extract pose invariant features for recognition, or first normalize profile face images to frontal pose before feature extraction, we argue that it is more desirable to perform both tasks jointly to allow them to benefit from each other. To this end, we propose a Pose Invariant Model (PIM) for face recognition in the wild, with three distinct novelties. First, PIM is a novel and unified deep architecture, containing a Face Frontalization sub-Net (FFN) and a Discriminative Learning sub-Net (DLN), which are jointly learned from end to end. Second, FFN is a well-designed dual-path Generative Adversarial Network (GAN) which simultaneously perceives global structures and local details, incorporated with an unsupervised cross-domain adversarial training and a "learning to learn" strategy for high-fidelity and identity-preserving frontal view synthesis. Third, DLN is a generic Convolutional Neural Network (CNN) for face recognition with our enforced cross-entropy optimization strategy for learning discriminative yet generalized feature representation. Qualitative and quantitative experiments on both controlled and in-the-wild benchmarks demonstrate the superiority of the proposed model over the state-of-the-arts.

1. Introduction

Face recognition has been a key problem in computer vision for decades. Even though (near-) frontal[1] face recognition seems to be solved under constrained conditions, the more general problem of *face recognition in the wild* still needs more studies, desiderated by many practical applications. For example, in surveillance scenarios, free-walking people would not always keep their faces frontal to the cameras. Most face images captured in the wild are contaminated by unconstrained factors like extreme pose, bad illumination, large expression, *etc*. Among them, the one that harms face recognition performance arguably the most is pose variation. The performance of most face recognition models degrades by over 10% from frontal-frontal to frontal-profile verification [24]. In contrast, human can recognize faces with large pose variance without significant accuracy drop. In this work, we aim to mitigate such a gap between human performance and automatic models for recognizing unconstrained faces with large pose variations.

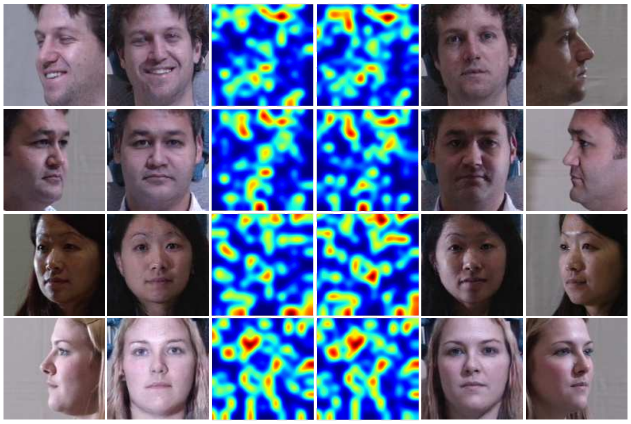


Figure 1: Pose invariant face recognition in the wild. *Col.* 1 & 6: distinct identities under different poses with other unconstrained factors (different expressions and lighting conditions); *Col.* 2 & 5: recovered frontal faces with our proposed PIM model; *Col.* 3 & 4: learned facial representations with our proposed PIM model. PIM can learn pose-invariant representations and recover photorealistic frontal faces effectively. The representations are extracted from the penultimate layer of PIM.

Recent studies [10, 19] discovered that human brain has a face-processing neural system that consists of several connected regions. The neurons in some of these regions perform face normalization (*i.e.*, profile to frontal) and others

[*] Homepage: https://zhaoj9014.github.io/.

[†] Work done in part during an internship at Panasonic R&D Center Singapore.

[1] "Near frontal" faces are almost equally visible for both sides and their yaw angles are within 10° from frontal view.

are tuned to identify the synthesized frontal faces, making face recognition robust to pose variation. This intriguing function of primate brain inspires us to develop a novel and unified deep neural network, termed as Pose Invariant Model (PIM), which jointly learns face frontalization and discriminative representation end-to-end that mutually boost each other to achieve pose-invariant face recognition. PIM takes face images of arbitrary poses with other potential distracting factors (*e.g.*, bad illuminations or different expressions) as inputs. It outputs facial representations invariant to pose variation and meanwhile preserves discriminativeness across different identities. As shown in Fig. 1, our proposed PIM can learn pose-invariant representations and effectively recover frontal faces.

In particular, PIM includes a Face Frontalization sub-Net (FFN) and a Discriminative Learning sub-Net (DLN) to learn the representations. The FFN contains a carefully designed dual-path Generative Adversarial Network (GAN) that simultaneously recovers global facial structures and local details. Besides, FFN introduces unsupervised cross-domain adversarial training and a "learning to learn" strategy with the siamese discriminator for achieving stronger generalizability and high-fidelity, identity-preserving frontal face generation. Cross-domain adversarial training is applied during training the generator to promote features that are indistinguishable *w.r.t.* the shift between source (training) and target (test) domains. In this way, the generalizability of FFN can be significantly improved even in case of only a few training samples from target domains. The discriminator in FFN introduces dynamic convolution to implement "learning to learn" for more efficient adaption and a siamese architecture featuring a pairwise training scheme to encourage the generator to produce photorealistic frontal faces without identify information loss. We introduce the other branch to the discriminator as the "learner", which predicts the dynamic convolutional parameters of the first branch from a single sample. DLN is a generic Convolutional Neural Network (CNN) for face recognition with our proposed enforced cross-entropy optimization strategy. Such a strategy reduces the intra-class distance while increasing the inter-class distance, so that the learned facial representations are discriminative yet generalizable.

We conduct extensive qualitative and quantitative experiments on various benchmarks, including both controlled and in-the-wild datasets. The results demonstrate the effectiveness of PIM on recognizing faces with extreme poses and the superiority over the state-of-the-arts consistently on all the benchmarks.

Our contributions are summarized as follows.

- We present a deep architecture unifying face frontalization and recognition in a mutual boosting way. It inherits the merits of existing pose invariant face recognition methods.
- We design a novel face frontalization network for photorealistic face frontalization that can generalize well across multiple domains and fast adapt to new application samples.
- We develop effective and novel training strategies for the frontalization network, the recognition network, and the whole deep architecture, which generate powerful face representations.
- Our deep architecture for pose invariant face recognition significantly outperforms the state-of-the-arts on three large benchmarks.

2. Related Work

Face Frontalization Face frontalization or normalization is a challenging task due to its ill-posed nature. Traditional methods address this problem through 2D/3D local texture warping [14, 38], statistical modeling [21], and deep learning based methods [18, 37]. For instance, Hassner *et al.* [14] used a single and unmodified 3D surface to approximate the shape of all the input faces, which is shown effective for face frontalization, but suffers big performance drop for profile and near-profile[2] faces due to severe texture loss and artifacts. Sagonas *et al.* [21] proposed to perform joint frontal view reconstruction and landmark detection by solving a constrained low-rank minimization problem. Kan *et al.* [18] used Stacked Progressive Auto-Encoders (SPAE) to rotate a profile face to frontal. Though with encouraging results, the synthesized faces lack fine details and tend to be blurry and unreal under a large pose. The quality of synthesized images with current methods is still far from satisfactory for recognizing faces with large pose variation.

Pose Invariant Representation Learning Conventional approaches often leverage robust local descriptors [8, 2, 7] and metric learning [4, 33] to tackle pose variance. In contrast, deep learning methods often handle pose variance through a single pose-agnostic or several pose-specific models with pooling operation and specific loss functions [6, 34]. For instance, the VGG-Face model [20] adopts the VGG architecture [27]. The DeepFace [30, 31] model uses a deep CNN coupled with 3D alignment. FaceNet [23] utilizes the inception architecture. The DeepID2+ [29] and DeepID3 [28] extend the FaceNet [23] model by including joint Bayesian metric learning and multi-task learning. However, such data-driven methods heavily rely on well annotated data. Collecting labeled data covering all variations is expensive and even impractical.

Our proposed PIM presents a similar idea with Two-Pathway GAN (TP-GAN) [17] and Disentangled Representation learning GAN (DR-GAN) [32]. TP-GAN considers photorealistic and identity preserving frontal view synthesis

[2] Faces with yaw angle greater than 60°.

and DR-GAN considers both face frontalization and representation learning in a unified network. Our proposed model differs from them in following aspects: 1) PIM aims to jointly learn face frontalization and pose invariant representations end-to-end to allow them to mutually boost each other for addressing large pose variance issue in unconstrained face recognition, whereas TP-GAN only tries to recover a frontal view from profile face images; 2) TP-GAN [17] and DR-GAN [32] suffer from poor generalizability and great optimization difficulties which limit their effectiveness in unconstrained face recognition, while our PIM architecture effectively overcomes these issues by introducing unsupervised cross-domain adversarial training, a "learning to learn" strategy using the siamese discriminator with dynamic convolution, and an enforced cross-entropy optimization strategy. Detailed experimental comparisons are provided in Sec. 4.

3. Pose Invariant Model

As shown in Fig. 2 (a), the proposed Pose Invariant Model (PIM) consists of a Face Frontalization sub-Net (FFN) and a Discriminative Learning sub-Net (DLN) that jointly normalize faces and learn face representation end-to-end. We now present each component in details.

3.1. Face Frontalization Sub-Net

3.1.1 Domain Invariant Dual-Path Generator

A photorealistic frontal face image is important for representing a face identity. A natural scheme is thus to generate this reference face from face images of arbitrary poses. Since the convolutional filters are usually shared across all the spatial locations, merely using a single-path generator cannot learn filters that are powerful enough for both sketching a rotated face structure and precisely recovering local textures. To address this issue, we propose a dual-path generator, as inspired by [17, 38], where one path aims to infer the global sketch and the other to attend to local facial details, as shown in Fig. 2 (b).

In particular, the global path generator G_{θ^g} (with learnable parameters θ^g) consists of a transition-down encoder $G_{\theta^g_E}$ and a transition-up decoder $G_{\theta^g_D}$. The local path generator G_{θ^l} also has an auto-encoder architecture, which contains four identical sub-networks that learn separately to frontalize the following four center-cropped local patches: left eye, right eye, nose and mouth. These patches are acquired by an off-the-shelf landmark detection model. Given an input face image I, to effectively integrate information from the global and local paths, we first align the feature maps f^l predicted by G_{θ^l} to a single feature map according to a pre-estimated landmark location template, which is further concatenated with the feature map f^g from the global path and then fed to following convolution layers to generate the final frontalized face image I'. We also concatenate a Gaussian random noise z at the bottleneck layer of the dual-path generator to model variations of other factors besides pose, which may also help recover invisible details.

Formally, let the input profile face image with four landmark patches be collectively denoted as I_{tr}. Then the predicted face is $I' = G_\theta(I_{tr})$. The key requirements for the FFN include two aspects. 1) The recovered frontal face image I' should visually resemble a real one and preserve the identity information as well as local textures. 2) It should be hardly possible for an algorithm to identify the domain of origin of the observation $I^{'}$ regardless of the underlying gap between source domain (with ample annotated data) and target domain (with rare annotated data).

To this end, we propose to learn the parameters $\{\theta^g, \theta^l_i\}$ (here i=$1,\ldots,4$ index the four local path models) by minimizing the following composite losses:

$$\mathcal{L}_{G_\theta} = -\mathcal{L}_{\text{adv}} + \lambda_0\mathcal{L}_{\text{ece}} - \lambda_1\mathcal{L}_{\text{domain}} + \lambda_2\mathcal{L}_{\text{pixel}} + \lambda_3\mathcal{L}_{\text{sym}} + \lambda_4\mathcal{L}_{\text{TV}}, \tag{1}$$

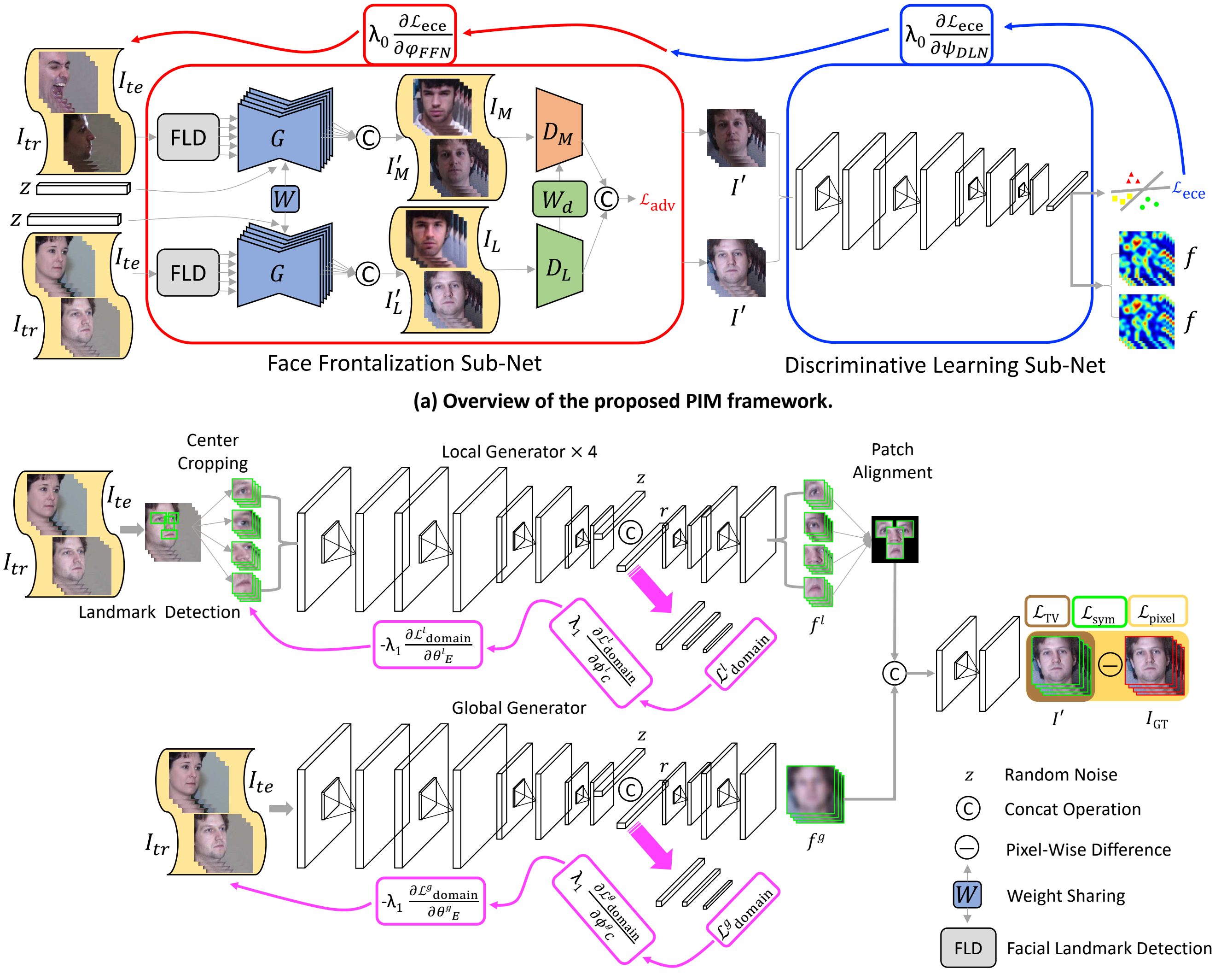
where $\mathcal{L}_{\text{adv}}$ is the adversarial loss for adding realism to the synthetic images and alleviating artifacts, $\mathcal{L}_{\text{ece}}$ is the enforced cross-entropy loss for preserving the identity information, $\mathcal{L}_{\text{domain}}$ is the cross-domain adversarial loss for domain adaption and generalization capacity enhancement, $\mathcal{L}_{\text{pixel}}$ is the pixel-wise ℓ_1 loss for encouraging multi-scale image content consistency, $\mathcal{L}_{\text{sym}}$ is the symmetry loss for alleviating self-occlusion issue, $\mathcal{L}_{\text{TV}}$ is the total variation loss for reducing spiky artifacts and $\{\lambda_k\}_{k=^4_0}$ are weighting parameters among different losses.

In order to enhance generalizability of the FFN and reduce over-fitting that hinders the practical application of most previous GAN-based models [17, 32], we adopt $\mathcal{L}_{\text{domain}}$ to promote the emergence of features encoded by G_{θ^g} and $G_{\theta^l_i}$ that are indistinguishable *w.r.t.* the shift between the source (training, I_{tr}) and target (testing, I_{te}) domains. Let I_i denote the images from both source and target domains, $y_i \in \{0, 1\}$ indicate which domain I_i is from, and $r_i = G_{\theta_E}(I_i)$ denote the representations. The cross-domain adversarial loss is defined as follows:

$$\mathcal{L}_{\text{domain}} = \frac{1}{N}\sum_i -y_i \log[C_\phi(r_i)] - (1-y_i)\log[1 - C_\phi(r_i)], \tag{2}$$

where ϕ denotes the learnable parameters for the domain classifier. Minimizing $\mathcal{L}_{\text{domain}}$ can reduce the domain discrepancy and help the generator achieve similar face frontalization performance across different domains, even if training samples from the target domain are limited. Such adapted representations are provided by augmenting the encoders of G_{θ^g} and $G_{\theta^l_i}$ with a few standard layers as the domain classifier C_ϕ, and a new gradient reversal layer to reverse the gradient during optimizing the encoders (*i.e.*, gradient update as in Fig. 2 (b)), as inspired by [11].

$\mathcal{L}_{\text{pixel}}$ is introduced to enforce the multi-scale content


$\lambda_0 \frac{\partial \mathcal{L}_{ece}}{\partial \varphi_{FFN}}$
$\lambda_0 \frac{\partial \mathcal{L}_{ece}}{\partial \psi_{DLN}}$
I_{te}
I_{tr}
FLD
G
W
z
I_M
I'_M
I_L
I'_L
D_M
W_d
D_L
$\mathcal{L}_{adv}$
I'
$\mathcal{L}_{ece}$
f
Face Frontalization Sub-Net
Discriminative Learning Sub-Net
Center Cropping
Landmark Detection
Local Generator × 4
Patch Alignment
r
f^l
$-\lambda_1 \frac{\partial \mathcal{L}^l_{domain}}{\partial \theta^l_E}$
$\lambda_1 \frac{\partial \mathcal{L}^l_{domain}}{\partial \phi^l_C}$
$\mathcal{L}^l$ domain
$\mathcal{L}_{TV}$
$\mathcal{L}_{sym}$
$\mathcal{L}_{pixel}$
I_{GT}
Global Generator
f^g
$-\lambda_1 \frac{\partial \mathcal{L}^g_{domain}}{\partial \theta^g_E}$
$\lambda_1 \frac{\partial \mathcal{L}^g_{domain}}{\partial \phi^g_C}$
$\mathcal{L}^g$ domain
Random Noise
Concat Operation
Pixel-Wise Difference
Weight Sharing
Facial Landmark Detection

(a) Overview of the proposed PIM framework.

(b) Dual-path generator architecture of the FFN.

Figure 2: Pose Invariant Model (PIM) for face recognition in the wild. The PIM contains an Face Frontalization sub-Net (FFN) and a Discriminative Learning sub-Net (DLN) that jointly learn end-to-end. FFN is a dual-path (*i.e.*, simultaneously perceiving global structures and local details) GAN augmented by (1) unsupervised cross-domain (*i.e.*, I_{tr} and I_{te}) adversarial training and (2) a siamese discriminator with a "learning to learn" strategy — convolutional parameters (*i.e.*, W_d) dynamically predicted by the "learner" D_L of the discriminator and transferred to D_M. DLN is a generic Convolutional Neural Network (CNN) for face recognition optimized by the proposed enforced cross-entropy optimization. It takes in the frontalized face images from FFN and outputs learned pose invariant facial representations.

consistency between the final frontalized face and corresponding ground truths, defined as $\mathcal{L}_{\text{pixel}} = \|I' - I_{GT}\|/|I_{GT}|$ where $|I_{GT}|$ is the size of I_{GT}. Since symmetry is an inherent feature of human faces, $\mathcal{L}_{\text{sym}}$ is introduced within the Laplacian space to exploit this prior information and impose the symmetry constraint on the recovered frontal view for alleviating self-occlusion issue:

$$\mathcal{L}_{\text{sym}} = \frac{1}{W/2 \times H} \sum_{i}^{W/2} \sum_{j}^{H} |I'_{i,j} - I'_{W-(i-1),j}|, \qquad (3)$$

where W, H denote the width and height of the final recovered frontal face image I', respectively.

The standard $\mathcal{L}_{\text{TV}}$ is introduced as a regularization term on the synthesized results to reduce spiky artifacts:

$$\mathcal{L}_{\text{TV}} = \sum_{i}^{W} \sum_{j}^{H} \sqrt{(I'_{i,j+1} - I'_{i,j})^2 + (I'_{i+1,j} - I'_{i,j})^2}. \qquad (4)$$

3.1.2 Dynamic Convolutional Discriminator

To increase realism of the synthesized images to benefit face recognition, we need to narrow the gap between the distributions of the synthetic and real images. Ideally, the generator should be able to generate images indistinguishable from real ones for a sufficiently powerful discriminator. Meanwhile, since the training sample size in this scenario is usually small, we need to develop a sample-efficient discriminator. To this end, we propose a "learning to learn" strategy using a siamese adversarial pixel-wise discriminator with dynamic convolution, as shown in Fig. 2 (a). This siamese architecture implements a pair-wise training scheme where each sample from the generator consists of two frontalized faces with the same identity and the corresponding real sample consists of two distinct frontal faces of the same person.

Different from conventional CNN based discriminators, we construct the second branch of the discriminator as the "learner" D_L that dynamically predicts the suitable convolutional parameters of the first branch D_M from a single sample. Formally, consider a particular convolutional layer in D_M. Given an input tensor (*i.e.*, feature maps from the previous layer) $x_{in} \in \mathbb{R}^{w \times h \times c_{\text{in}}}$ and kernel weights $W \in \mathbb{R}^{k \times k \times c_{\text{in}} \times c_{\text{out}}}$ where k is the kernel size, the output $x_{\text{out}} \in \mathbb{R}^{w' \times h' \times c_{\text{out}}}$ of the convolutional layer can be computed as $x_{\text{out}} = W * x_{\text{in}}$, where $*$ denotes the convolution operation.

Inspired by [3], we perform the following factorization, which is analogous to Singular Value Decomposition (SVD),

$$x_{\text{out}} = U' * (W_d) *_{c_{\text{in}}} U * x_{\text{in}}, \tag{5}$$

where $U \in \mathbb{R}^{1 \times 1 \times c_{\text{in}} \times c_{\text{in}}}$, $U' \in \mathbb{R}^{1 \times 1 \times c_{\text{in}} \times c_{\text{out}}}$, $W_d \in \mathbb{R}^{k \times k \times c_{\text{in}}}$ is the dynamic convolution kernel predicted by D_L and $*_{c_{\text{in}}}$ denotes independent filtering of c_{in} channels. Under the factorization of Eqn. (5), the number of parameters to learn by D_L is significantly decreased from $k \times k \times c_{\text{in}} \times c_{\text{out}}$ to $k \times k \times c_{\text{in}}$, allowing them to grow only linearly with the number of input feature map channels.

We leverage the same architecture of global-path encoder as D_M and D_L, learning separately without weight sharing, while two generator blocks in Fig. 2 (a) share their weights. The feature maps from D_M and D_L are further concatenated and fed into a fully connected bottleneck layer to compute $\mathcal{L}_{\text{adv}}$, which serves as a supervision to push the synthesized image to reside in the manifold of photorealistic frontal view images, prevent blur effect, and produce visually pleasing results. In particular, $\mathcal{L}_{\text{adv}}$ is defined as

$$\begin{aligned} \mathcal{L}_{\text{adv}} = \frac{1}{N} \sum_i & - y_i \log[D_{M \leftarrow L}(I_M, I_L)] \\ & - (1 - y_i) \log[1 - D_{M \leftarrow L}(I_{\text{M}}, I_L)], \end{aligned} \tag{6}$$

where $D_{M \leftarrow L}$ denotes the siamese discriminator with dynamic convolution, (I_M, I_L) denotes the pair of face images fed to $D_{M \leftarrow L}$ and y is the binary label indicating the pair is synthesized or real.

3.2. Discriminative Learning Sub-Net

The DLN is a generic CNN for face recognition trained by our proposed enforced cross-entropy optimization strategy for learning discriminative yet generalizable facial representations. This strategy reduces the intra-class distance while increasing the inter-class distance. Moreover, it helps improve the rubustness of the learned representations and address the potential over-fitting issue.

DLN takes the frontalized face images I' from the FFN as input, and outputs the learned pose invariant facial representations $f = M_\psi(I')$, which are further utilized for face verification and identification. Here M_ψ denotes the DLN model parameterized by ψ. We define every column vector of the weights of the last fully connected layer of DLN as an anchor vector a which represents the *center* of each identity in the feature space. Thus, the decision boundary can be derived when the feature vector has the same distance (cosine metric) to several anchor vectors (cluster centers), *i.e.*, $a_i^\top f = a_j^\top f$.

However, in such cases, the samples close to the decision boundary can be wrongly classified with a high confidence. A simple yet effective solution is to reduce the intra-class distance while increasing the inter-class distance of the feature vectors, through which the hard samples will be adjusted and re-allocated in the correct decision area. To achieve this goal, we propose to impose a selective attenuation factor as a regularization term to the confidence scores (predictions) of the genuine samples:

$$p_i = \frac{\exp[\tau_t \cdot (a_i^\top f)]}{\sum_j \exp[\tau_t \cdot (a_j^\top f)]}, \tag{7}$$

where p_i denotes the predicted confidence score *w.r.t.* the i^{th} identity, τ_t denotes the selective attenuation factor, a and f are ℓ_2 normalized to achieve boundary equilibrium during network training. In particular, τ_t in Eqn. (7) is updated by $\tau_{t+1} = \tau_t \left(1 - \frac{n}{B}\right)^\alpha$, where n denotes the batch index, B denotes the total batch number and α is the diversity ratio.

Selective attenuation on the confidence scores of genuine samples in turn increases the corresponding classification losses, which narrows the decision boundary and controls the intra-class affinity and inter-class distance.

The predictions of Eqn. (7) are used to compute the multi-class cross-entropy objective function for updating network parameters (*i.e.*, gradient update as in Fig. 2 (a)), which is an enforced optimization scheme:

$$\mathcal{L}_{\text{ece}} = \frac{1}{N} \sum_i -l_i \log(p) - (1 - l_i) \log(1 - p), \tag{8}$$

where l_i is the face identity ground truth.

4. Experiments

We evaluate PIM qualitatively and quantitatively under both controlled and in-the-wild settings for pose-invariant face recognition. For qualitative evaluation, we show visualized results of face frontalization on Multi-PIE [12] and LFW [16] benchmark datasets. For quantitative evaluation, we evaluate face recognition performance using the learned facial representations with a cosine distance metric on Multi-PIE [12] and CFP [24] benchmark datasets.

Implementation Details Throughout the experiments, the size of the RGB face images from training domain (I_{tr}), testing domain (I_{te}), and the FFN prediction (I') is fixed as 128×128; the sizes of the four RGB local patches (*i.e.*, left/right eye, nose and mouth) are fixed as 40×40, 40×40, 32×40 and 48×32, respectively; the dimensionality of the Gaussian random noise z is fixed as 100; the diversity ratio α and the constraint factors $\lambda_i, i \in \{0, 1^{3}, 2, 3, 4\}$ are empirically fixed as $0.9, 5\times10^{-3}, 0.1, 0.3, 5\times10^{-2}$ and 5×10^{-4}, respectively; the dropout ratio is fixed as 0.7; the weight decay, batch size and learning rate are fixed as 5×10^{-4}, 10 and 2×10^{-4}, respectively. We use off-the-shelf OpenPose [25] for landmark detection[4]. We initialize the DLN with ResNet-50 [15] and Light CNN-29 [35] architectures as our two baselines, which are pre-trained on MS-Celeb-1M [13] and fine-tuned on the target dataset. We initialize D_M and D_L with the same architecture as the global-path encoder and pre-train D_L on MS-Celeb-1M [13]. The proposed network is implemented based on the publicly available TensorFlow [1] platform, which is trained using Adam (β_1=0.5) on three NVIDIA GeForce GTX TITAN X GPUs with 12G memory.

4.1. Evaluations on the Multi-PIE Benchmark

The CMU Multi-PIE [12] dataset is the largest multi-view face recognition benchmark, which contains 754,204 images of 337 identities from 15 view points and 20 illumination conditions. We conduct experiments under two settings: **Setting-1** concentrates on pose, illumination and minor expression variations. It only uses the images in session one, which contains 250 identities. The images with 11 poses within ±90° and 20 illumination levels of the first 150 identities are used for training. For testing, one frontal view with neutral expression and illumination (*i.e.*, ID07) is used as the gallery image for each of the remaining 100 identities and other images are used as probes. **Setting-2** concentrates on pose, illumination and session variations. It uses the images with neutral expression from all four sessions, which contains 337 identities. The images with 11 poses within ±90° and 20 illumination levels of the first 200 identities are used for training. For testing, one frontal view with neural illumination is used as the gallery image for each of the remaining 137 identities and other images are used as probes.

Table 1: Component analysis: rank-1 recognition rates (%) under Multi-PIE [12] Setting-1. b1 and b2 denote ResNet-50 [15] and Light CNN-29 [35], respectively. PIM1 and PIM2 use ResNet-50 [15] and Light CNN-29 [35] as backbone architectures, respectively.

Method	$\pm90^\circ$	$\pm75^\circ$	$\pm60^\circ$	$\pm45^\circ$	$\pm30^\circ$	$\pm15^\circ$
b1	18.80	63.80	92.20	98.30	99.20	99.40
b2	33.00	76.10	95.20	97.90	99.20	99.80
w/o $\mathcal{L}_{\text{pixel}}$	60.60	82.30	89.60	93.70	98.50	98.60
w/o $G_{\theta^l_i}$	66.80	89.30	95.60	98.20	99.30	99.80
w/o D_φ	66.90	90.00	96.50	98.00	99.20	99.80
w/o dyn conv	69.80	90.70	96.80	98.10	99.40	99.80
w/o $\mathcal{L}_{\text{domain}}$	71.10	90.80	97.10	98.30	99.30	99.80
w/o $\mathcal{L}_{\text{sym}}$	72.30	90.40	96.80	98.20	99.30	99.80
PIM1	71.60	**92.50**	97.00	**98.60**	99.30	99.40
PIM2	**75.00**	91.20	**97.70**	98.30	**99.40**	**99.80**

4.1.1 Component Analysis

We first investigate different architectures and loss function combinations of PIM to see their respective roles in pose invariant face recognition. We compare eight variants of PIM, *i.e.*, different DLN architectures (ResNet-50 [15] vs. Light CNN-29 [35]), w/o $\mathcal{L}_{\text{pixel}}$, w/o local-path generator $G_{\theta^l_i}$, w/o siamese discriminator D_φ (D_L is removed), w/o dynamic convolution (siamese discriminator without sharing weights), w/o cross-domain adversarial training $\mathcal{L}_{\text{domain}}$ and w/o $\mathcal{L}_{\text{sym}}$, in each case.

Averaged rank-1 recognition rates are compared in Setting-1 in Tab. 1. The results on the profile images serve as our baselines (*i.e.*, b1 and b2). The results of the middle panel variations are all based on Light CNN-29 [35]. By comparing the results from the top and bottom panels, we observe that our PIM is not restricted to the DLN architecture used, since similar improvements (*e.g.* 52.80% *v.s.* 42.00% under ±90°) can be achieved with our joint face frontalization and discriminative representation learning framework. The pixel loss, dual-path generator and the "learning to learn" strategy using the siamese discriminator with dynamic convolution of the FFN contribute the most to improving the face recognition performance, especially for large pose cases. Although not apparent, the cross-domain adversarial training and symmetry loss also help improve the recognition performance. Cross-domain adversarial training is crucial for enhancing the generalization capacity of PIM on Multi-PIE [12] as well as other benchmark datasets. Fig. 3 illustrates the perceptual performance of these variants. As expected, the inference result without pixel loss, local-path generator or "learning to learn" strategy using the siamese discriminator with dynamic convolution deviates from the true appearance seriously. The synthesis without cross-domain adversarial training tends to present inferior generalizability while that without symmetry loss sometimes shows factitious asymmetrical effect.

4.1.2 Intermediate Results Visualization

Most previous works on face frontalization and pose invariant representation learning are dedicated to address prob-

[3] Cross-domain adversarial training is an option, if there is no need to do domain adaptation, simply set λ_1=0.

[4] For profile face images with large yaw angles, OpenPose [25] may fail to locate both eyes. In such cases, we use the detected eye after center cropping as the input left/right eye patch.

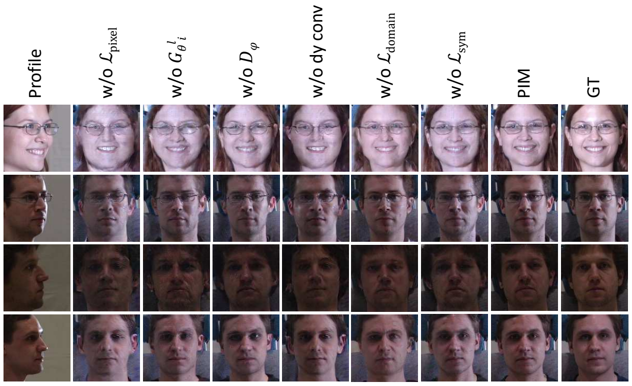

Profile
w/o $\mathcal{L}_{\text{pixel}}$
w/o $G_{\theta}{}^{l}{}_{i}$
w/o D_{φ}
w/o dy conv
w/o $\mathcal{L}_{\text{domain}}$
w/o $\mathcal{L}_{\text{sym}}$
PIM
GT

Figure 3: Component analysis. Synthesized results of PIM and its variants.

Table 2: Rank-1 recognition rates (%) across views, minor expressions and illuminations under Multi-PIE [12] Setting-1. "-" means the result is not reported. b1 and b2 denote ResNet-50 [15] and Light CNN-29 [35], respectively. PIM1 and PIM2 use ResNet-50 [15] and Light CNN-29 [35] as backbone architectures, respectively.

Method	±90°	±75°	±60°	±45°	±30°	±15°
b1	18.80	63.80	92.20	98.30	99.20	99.40
b2	33.00	76.10	95.20	97.90	99.20	99.80
CPF [37]	-	-	-	71.65	81.05	89.45
Hassner [14]	-	-	44.81	74.68	89.59	96.78
FV [26]	24.53	45.51	68.71	80.33	87.21	93.30
HPN [9]	29.82	47.57	61.24	72.77	78.26	84.23
FIP_40 [39]	31.37	49.10	69.75	85.54	92.98	96.30
c-CNN [36]	47.26	60.66	74.38	89.02	94.05	96.97
TP-GAN [17]	64.03	84.10	92.93	98.58	**99.85**	99.78
PIM1	71.60	**92.50**	97.00	**98.60**	99.30	99.40
PIM2	**75.00**	91.20	**97.70**	98.30	99.40	**99.80**

Table 3: Rank-1 recognition rates (%) across views, illuminations and sessions under Multi-PIE [12] Setting-2. "-" means the result is not reported. b1 and b2 denote ResNet-50 [15] and Light CNN-29 [35], respectively. PIM1 and PIM2 use ResNet-50 [15] and Light CNN-29 [35] as backbone architectures, respectively.

Method	±90°	±75°	±60°	±45°	±30°	±15°
b1	15.50	55.10	85.90	97.10	98.40	98.60
b2	27.10	68.70	91.40	97.70	98.60	99.10
FIP [39]	-	-	45.90	64.10	80.70	90.70
MVP [40]	-	-	60.10	72.90	83.70	92.80
CPF [37]	-	-	61.90	79.90	88.50	95.00
DR-GAN [32]	-	-	83.20	86.20	90.10	94.00
TP-GAN [17]	64.64	77.43	87.72	95.38	98.06	98.68
PIM1	81.30	92.70	96.60	97.30	98.40	98.80
PIM2	**86.50**	**95.00**	**98.10**	**98.50**	**99.00**	**99.30**

lems within a pose range of ±60°, since it is commonly believed with a pose larger than 60°, it is difficult for a model to generate faithful frontal images or learn discriminative yet generative facial representations. However, with enough training data and proper architecture and objective function design of the proposed PIM, it is in fact feasible to recover high-fidelity and identity-preserving frontal faces under very large poses and learn pose invariant representations for face recognition in the wild.

The intermediate results of recovered face images in the frontal view and learned facial representations are visualized in Fig. 1. We observe that the frontalized faces present compelling perceptual quality across poses larger than 60°, and the learned representations are discriminative and pose invariant.

4.1.3 Face Recognition Comparison

Tab. 2 shows the face recognition performance comparison of our PIM with two baselines and other state-of-the-arts in Setting-1. Regardless of the adopted DLN architecture, PIM consistently achieves the best performance across all poses (except comparable with TP-GAN [17] under ±30°), especially for large yaw angles. In particular, PIM (Light CNN-29 [35]) outperforms TP-GAN [17] and c-CNN Forest [36] by 10.97% and 27.74% under ±90°, respectively. Note that TP-GAN [17] adopts Light CNN-29 [35] as the feature extractor which has the same architecture as our DLN and c-CNN Forest [36] is an ensemble of three models, while our PIM has a more effective and efficient joint training scheme and a much simpler network architecture.

Tab. 3 shows the face recognition comparison of our PIM with two baselines and other state-of-the-arts in Setting-2. Similar to the observation under Setting-1, PIM consistently achieves the best performance across all poses. In particular, PIM (Light CNN-29 [35]) outperforms TP-GAN [17] by 21.86% under ±90°, and outperforms TP-GAN [17] and DR-GAN [32] by 10.38% and 14.90% under ±60°, respectively. This well verifies the superiority of our proposed cross-domain adversarial training, the "learning to learn" strategy using the siamese discriminator with dynamic convolution and the enforced cross-entropy optimization strategy in improving the overall recognition performance.

4.2. Evaluations on the CFP Benchmark

The CFP [24] dataset aims to evaluate the strength of face verification approaches across pose, more specifically, between frontal view (yaw angle<10°) and profile view (yaw angle>60°). CFP contains 7,000 images of 500 subjects, where each subject has 10 frontal and 4 profile face images. The data are randomly organized into 10 splits, each containing an equal number of frontal-frontal and frontal-profile pairs, with 350 genuine and 350 imposter ones, respectively. Evaluation systems report the mean and standard deviation of accuracy, Equal Error Rate (EER) and Area Under Curve (AUC) over the 10 splits for both frontal-frontal and frontal-profile face verification settings.

Tab. 4 compares the face recognition performance of our PIM (Light CNN-29 [35]) with other state-of-the-arts on the CFP [24] benchmark dataset. The results on the original images serve as our baseline. PIM achieves comparable performance as human under fontal-profile setting and outperforms human performance under frontal-frontal setting. In particular, for frontal-frontal cases, PIM gives stably similar saturated performance with b (Light CNN-29 [35]), both of which reduce the EER of human performance by around

Table 4: Face recognition performance (%) comparison on CFP [24]. The results are averaged over 10 testing splits.

Method	Frontal-Profile			Frontal-Frontal		
	Acc	EER	AUC	Acc	EER	AUC
FV+DML [24]	58.47±3.51	38.54±1.59	65.74±2.02	91.18±1.34	8.62±1.19	97.25±0.60
LBP+Sub-SML [24]	70.02±2.14	29.60±2.11	77.98±1.86	83.54±2.40	16.00±1.74	91.70±1.55
HoG+Sub-SML [24]	77.31±1.61	22.20±1.18	85.97±1.03	88.34±1.33	11.45±1.35	94.83±0.80
FV+Sub-SML [24]	80.63±2.12	19.28±1.60	88.53±1.58	91.30±0.85	8.85±0.74	96.87±0.39
Deep Features [24]	84.91±1.82	14.97±1.98	93.00±1.55	96.40±0.69	3.48±0.67	99.43±0.31
Triplet Embedding [22]	89.17±2.35	8.85±0.99	97.00±0.53	96.93±0.61	2.51±0.81	99.68±0.16
Chen *et al.* [5]	91.97±1.70	8.00±1.68	97.70±0.82	98.41±0.45	1.54±0.43	99.89±0.06
Light CNN-29 [35]	92.47±1.44	8.71±1.80	**97.77±0.76**	**99.64±0.32**	**0.57±0.40**	99.92±0.15
PIM (Light CNN-29 [35])	**93.10±1.01**	**7.69±1.29**	97.65±0.62	99.44±0.36	0.86±0.49	**99.92±0.10**
Human	94.57±1.10	5.02±1.07	98.92±0.46	96.24±0.67	5.34±1.79	98.19±1.13

5.00%. For more challenging frontal-profile cases, PIM consistently outperforms the baseline and other state-of-the-arts. In particular, PIM reduces the EER by 1.02% compared with b (Light CNN-29 [35]) and improves the accuracy by 1.13% over the 2^{nd}-best. This shows that the facial representations learned by PIM are discriminative and robust even at extreme pose variations.

4.3. Evaluations on the LFW Benchmark

LFW [16] contains 13,233 face images of 5,749 identities. The images were obtained by trawling the Internet followed by face centering, scaling and cropping based on bounding boxes provided by an automatic face locator. The LFW data have large in-the-wild variabilities, *e.g.*, in-plane rotations, non-frontal poses, low resolution, non-frontal illumination, varying expressions and imperfect localization.

As a demonstration of our model's superior generalizability to in-the-wild face images, we qualitatively compare the intermediate face frontalization results of our PIM (Light CNN-29 [35]) with TP-GAN [17], DR-GAN [32], and the approach from Hassner *et al.* [14], which are the state-of-the-arts aiming to generate photorealistic and identity preserving frontal view from profiles. As in Fig. 4, the predictions of TP-GAN [17] suffer severe texture loss and involved artifacts, and the predictions of DR-GAN [32] and the method by Hassner *et al.* [14] deviate from true appearance seriously, for both near-frontal (the top two rows) and profile (the bottom three rows) cases. Comparatively, PIM can faithfully recover high fidelity frontal view face images with finer local details and global face shapes. This well verifies that the unsupervised cross-domain adversarial training can effectively advance generalizability and reduce over-fitting, and that the "learning to learn" strategy using a siamese discriminator with dynamic convolution contributes to the synthesized perceptually natural and photorealistic results. Moreover, the joint learning scheme of face frontalization and discriminative representation also helps, since the two sub-nets leverage each other during end-to-end training to achieve a final win-win outcome.

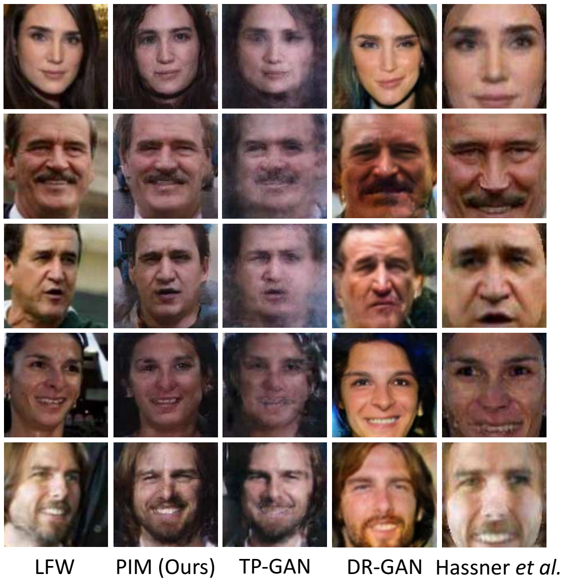


Figure 4: Comparison of face frontalization on LFW [16].

5. Conclusion

We proposed a novel Pose Invariant Model (PIM) to address the challenging face recognition with large pose variations. PIM unifies a Face Frontalization sub-Net (FFN) and a Discriminative Learning sub-Net (DLN) for pose invariant recognition in an end-to-end deep architecture. The FFN introduces unsupervised cross-domain adversarial training and a "learning to learn" strategy to provide high-fidelity frontal reference face image for effective learning face representation from DLN. Comprehensive experiments demonstrate the superiority of PIM over the state-of-the-arts. We plan to apply PIM for other domain adaption and transfer learning applications in the future.

Acknowledgement

The work of Jian Zhao was partially supported by **C**hina **S**cholarship **C**ouncil (CSC) grant 201503170248.

The work of Junliang Xing was partially supported by the National Science Foundation of Chian 61672519.

The work of Jiashi Feng was partially supported by NUS startup R-263-000-C08-133, MOE Tier-I R-263-000-C21-112, NUS IDS R-263-000-C67-646 and ECRA R-263-000-C87-133.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. 6

[2] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *TPAMI*, 28(12):2037–2041, 2006. 2

[3] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi. Learning feed-forward one-shot learners. In *NIPS*, pages 523–531, 2016. 5

[4] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *CVPR*, pages 3025–3032, 2013. 2

[5] J.-C. Chen, J. Zheng, V. M. Patel, and R. Chellappa. Fisher vector encoded deep convolutional features for unconstrained face verification. In *ICIP*, pages 2981–2985, 2016. 8

[6] W. Chen, T.-Y. Liu, Y. Lan, Z.-M. Ma, and H. Li. Ranking measures and loss functions in learning to rank. In *NIPS*, pages 315–323, 2009. 2

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 2

[8] J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *JOSA A*, 2(7):1160–1169, 1985. 2

[9] C. Ding and D. Tao. Pose-invariant face recognition with homography-based normalization. *PR*, 66:144–152, 2017. 7

[10] W. A. Freiwald and D. Y. Tsao. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, 330(6005):845–851, 2010. 1

[11] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(59):1–35, 2016. 3

[12] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *JIVC*, 28(5):807–813, 2010. 6, 7

[13] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, pages 87–102. Springer, 2016. 6

[14] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*, pages 4295–4304, 2015. 2, 7, 8

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6, 7

[16] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, pages 07-49, University of Massachusetts, Amherst, 2007. 6, 8

[17] 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. *arXiv preprint arXiv:1704.04086*, 2017. 2, 3, 7, 8

[18] M. Kan, S. Shan, H. Chang, and X. Chen. Stacked progressive auto-encoders (spae) for face recognition across poses. In *CVPR*, pages 1883–1890, 2014. 2

[19] S. Ohayon, W. A. Freiwald, and D. Y. Tsao. What makes a cell face selective? the importance of contrast. *JN*, 74(3):567–581, 2012. 1

[20] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. 2

[21] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Robust statistical face frontalization. In *ICCV*, pages 3871–3879, 2015. 2

[22] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa. Triplet probabilistic embedding for face verification and clustering. In *BTAS*, pages 1–8, 2016. 8

[23] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 2

[24] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *WACV*, pages 1–9, 2016. 1, 6, 7, 8

[25] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 6

[26] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *BMVC*, 2013. 7

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[28] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015. 2

[29] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *CVPR*, pages 2892–2900, 2015. 2

[30] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014. 2

[31] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. In *CVPR*, pages 2746–2754, 2015. 2

[32] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017. 2, 3, 7, 8

[33] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10(Feb):207–244, 2009. 2

[34] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515, 2016. 2

[35] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *arXiv preprint arXiv:1511.02683*, 2015. 6, 7, 8

[36] C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan, and T.-K. Kim. Conditional convolutional neural network for modality-aware face recognition. In *ICCV*, pages 3667–3675, 2015. 7

[37] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim. Rotating your face using multi-task deep neural network. In *CVPR*, pages 676–684, 2015. 2, 7

[38] 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 *CVPR*, pages 787–796, 2015. 2, 3

[39] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *ICCV*, pages 113–120, 2013. 7

[40] Z. Zhu, P. Luo, X. Wang, and X. Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In *NIPS*, pages 217–225, 2014. 7