

Pose Invariant Face Recognition in the Wild

Jian Zhao, *Student Member, IEEE*, Yu Cheng, Lin Xiong, Junliang Xing, *Senior Member, IEEE*,
Shuicheng Yan, *Fellow, IEEE*, and Jiashi Feng, *Member, IEEE*

Abstract—Extreme pose variation is one of the key obstacles to accurate face recognition in practice. Compared with current techniques for pose invariant face recognition, which either expect pose invariance from hand-crafted features or data-driven deep learning solutions, 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, incorporating an unsupervised cross-domain adversarial training and a “learning to learn” strategy using siamese discriminator with dynamic convolution 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 representations with large intra-class affinity and inter-class separability. Qualitative and quantitative experiments on both controlled and in-the-wild benchmark datasets demonstrate the superiority of the proposed model over the state-of-the-arts. The complete source code, trained models and online demo of this work will be released to facilitate future research on pose invariant face recognition in the wild.

Index Terms—Pose Invariant Face Recognition, Face Frontalization, Cross-Domain Adversarial Learning, Domain Adaption, Transfer Learning, Learning to Learn, Dynamic Convolution, Enforced Cross-Entropy Optimization, Generative Adversarial Networks.

1 INTRODUCTION

FACE recognition has been a key problem in computer vision. Even though (near-) frontal¹ face recognition seems to be solved under constrained conditions, the more general problem of *face recognition in the wild* still needs more studies, desired 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. In fact, as demonstrated in [1], the performance of most face recognition models degrades by over 10% from frontal-frontal to frontal-profile verification because discriminative descriptors suffer from misalignment issues. In contrast, humans can recognize faces in presence of 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.

Some research efforts have been made to address the pose

- Jian Zhao and Yu Cheng were interns at Panasonic R&D Center Singapore during this work. Jian Zhao is the corresponding author. Homepage: <https://zhaoj9014.github.io/>.
- Jian Zhao, Shuicheng Yan, and Jiashi Feng are with Department of Electrical and Computer Engineering, National University of Singapore, Singapore. Jian Zhao is also with School of Computer, National University of Defense Technology, China. E-mail: zhaojian90@u.nus.edu, {elefjia, eleyans}@nus.edu.sg.
- Yu Cheng is with Nanyang Technological University, Singapore. E-mail: chen0974@e.ntu.edu.sg.
- Lin Xiong is with Core Technology Group, Learning & Vision, Panasonic R&D Center Singapore, Singapore. E-mail: lin.xiong@sg.panasonic.com.
- Junliang Xing is with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China. E-mail: jlxing@nlpr.ia.ac.cn.

1. “Near frontal” faces are almost equally visible for both sides and their yaw angles are within 10° from the frontal view.

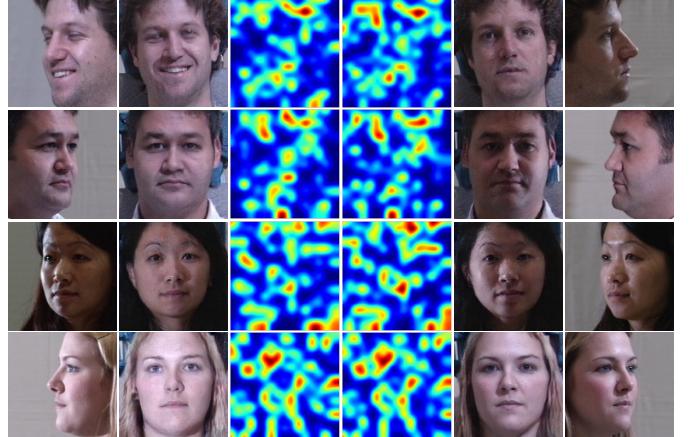


Fig. 1: Pose invariant face recognition in the wild. Col. 1 & 6: distinct identities under different poses with other unconstrained factors (different expressions and lighting conditions); Col. 2 & 5: recovered frontal faces with our proposed **Pose Invariant Model** (PIM); Col. 3 & 4: learned facial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

variation challenge, which can be roughly classified into two categories, *i.e.*, the discriminative methods and the generative ones. The discriminative category tries to adopt hand-crafted pose-invariant features [2], [3] or data-driven deep learning solutions from ample face data [4], [5], while the other category resorts to generative techniques to recover a frontal view from a profile face image and then use the recovered face images for face recognition [6], [7].

For the first category, conventional approaches often leverage robust local descriptors such as Gabor [8], Local Binary Pattern

(LBP) [9], **Histograms of Oriented Gradient (HOG)** [10] to tackle local distortions and then resort to metric learning algorithms [2], [11] to achieve pose invariance. However, due to the tradeoff between invariance and discriminability, hand-crafted features usually cost huge human-engineering effort and fail to deal with extreme pose cases effectively. In contrast, deep learning methods often handle pose variability issues with a single pose-agnostic model or several pose-specific models with pooling operation and employ ranking loss [12] or center loss [13] for optimization to ensure large intra-class affinity and inter-class separability. However, such data-driven methods are too computationally complex for practical application. Moreover, massive labelled training data covering all underlying variations are usually expensive and unavailable.

For the second category, previous efforts often utilize 3D geometrical transformations to render a frontal view by first aligning the 2D face image with either a general or an identity-specific **3D Morphable Model (3D MM)** [14], [15]. These methods are prominent at normalizing frontal or near-frontal faces, but their performance decreases for profile or near-profile² faces due to severe texture loss and involved artifacts. Recently, deep learning based methods [7], [16], [17], [18] are proposed for face frontalization. For instance, Zhu *et al.* [7] propose a **Multi-View Perceptron (MVP)** model, which can untangle the identity and view features, and meanwhile infer a full spectrum of multi-view images, given a single 2D face image. Although their results are encouraging, the synthesized images sometimes lack fine details and tend to be blurry and unreal under a large pose. Thus, they only leverage the intermediate features for face recognition. Among current generative methods, the quality of synthesized images is still far from satisfactory for performing practical facial analysis tasks, such as face verification (*i.e.*, 1:1 compare) and identification (*i.e.*, 1:N search).

Research [19], [20] has shown that the human brain has a face-processing neural system consisting of several connected regions. The neurons in some of these regions perform face normalization (*i.e.*, profile to frontal) and others are tuned to identify the synthesized frontal faces, making face recognition robust to pose variation. This intriguing function of the primate brain inspires us to develop a novel and unified deep neural network, which we call **Pose Invariant Model (PIM)**. The PIM jointly learns face frontalization and discriminative representations end-to-end that mutually boost each other to achieve pose-invariant face recognition. It takes as input the face images with arbitrary poses and other potential distracting factors (*e.g.*, bad illumination or different expressions), and outputs facial representations that are invariant to pose variation and meanwhile preserve discriminativeness across different identities. As shown in Fig. 1, the PIM can learn pose-invariant representations and effectively recover frontal faces.

In particular, PIM includes a **Face Frontalization sub-Net (FFN)** to normalize the profile faces 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 adopts the unsupervised cross-domain adversarial training and a “learning to learn” strategy using

² We define a “near profile” pose as one that obscures many features, specifically the second eye. This roughly corresponds to the yaw angle greater than 60 degrees.

siamese discriminator with dynamic convolution for achieving stronger generalizability and high-fidelity, identity-preserving frontal face generation. Cross-domain adversarial training is inspired by the theory of domain adaption, and it 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 pair-wise training scheme to encourage the generator to produce photorealistic frontal faces without identify information loss. We use the other branch in 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.

Compared with a preliminary version of this work, we make following extensions. 1) We visually analyze the dynamic convolution kernels predicted by learner branch of the discriminator and corresponding feature maps from main branch of the discriminator to explain the effectiveness of the “learning to learning” strategy in a more transparent way. 2) We add many details on gradient update with cross-domain adversarial learning, discuss differences between our proposed PIM and multi-task learning as well as multi-stage learning, including parameter setting, network architecture and training procedure. Our source code, trained models and online demo will be released upon acceptance. We will also release the explored datasets of synthesized frontal face images by the PIM. 3) We visualize detailed pose invariant face recognition results by our proposed PIM across a wide range of poses with other unconstrained factors (*e.g.*, bad illumination, large expression, *etc.*) on Multi-PIE [21], CFP [1], IJB-A [22] and LFW [23] to show the effectiveness and potential of our approach for real-world application. 4) We perform feature space analysis to gain an insight into the superiority of the proposed joint face frontalization and discriminative representation learning scheme of our PIM over existing discriminative solutions. 5) We replenish the performance curve comparison between PIM and baseline method under CFP [1] to facilitate evaluation analysis across different settings. 6) We supplement an additional experiment on IJB-A [22] to further verify the effectiveness and generalizability of PIM for pose invariant face recognition in the wild. 7) We study and discuss the failure cases of PIM to analyze its current limitation and possible future work on improvement.

Our contributions are summarized as follows.

- We propose a **Pose Invariant Model (PIM)** for face recognition in the wild. PIM is a novel and unified deep neural network containing a **Face Frontalization sub-Net (FFN)** and a **Discriminative Learning sub-Net (DLN)** that jointly learn in an end-to-end way to allow them to mutually boost each other.

- FFN is a carefully designed dual-path (*i.e.*, simultaneously perceiving global structures and local details) Generative Adversarial Network (GAN) incorporating unsupervised cross-domain adversarial training and a “learning to learn” strategy using siamese discriminator with dynamic convolution for high-fidelity and identity-preserving frontal view synthesis.
- DLN is a generic Convolutional Neural Network (CNN) for face recognition with our proposed enforced cross-entropy optimization strategy for learning discriminative yet generalized feature representations with large intra-class affinity and inter-class separability.
- We develop effective and novel training strategies for FFN, DLN and the whole deep architecture, which generate powerful face representations.
- As a by-product, the recovered frontal face images by PIM can also be utilized by conventional descriptors and learning algorithms so as to eliminate the negative effects from unconstrained conditions.

Based on the above model innovations and technical contributions, we present a high-performance pose invariant face recognition system. It achieves state-of-the-art performance on Multi-PIE [21], CFP [1], IJB-A [22] and LFW [23] benchmark datasets. The source code, trained models and online demo of our deep architecture will be made available to the community.

2 RELATED WORK

We review some recent studies which are most related to this work. For a thorough review on face recognition, please refer to the good surveys in this field [24], [25], [26].

2.1 Generative Adversarial Networks

As one of the most significant advancements in the research of deep generative models [27], [28], GAN has drawn substantial attention from the deep learning and computer vision community ever since it was first introduced by Goodfellow *et al.* [29]. The GAN framework learns a generator network and a discriminator network with competing loss. This min-max two-player game provides a simple yet powerful way to estimate target distribution and to generate novel image samples. Mirza and Osindero [30] introduce the conditional version of GAN, conditioned on both the generator and discriminator for effective image tagging. Berthelot *et al.* [31] propose a new **Boundary Equilibrium GAN** (BE-GAN) framework paired with a loss derived from the Wasserstein distance for training GAN, making the trade-off controllable between image diversity and visual quality. These successful applications of GAN motivate us to develop FFN based on GAN.

2.2 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], [15], statistical modelling [32], and deep learning based methods [18], [33]. For instance, Hassner *et al.* [14] use a single and unmodified 3D surface to approximate the shape of all input faces, which is shown effective for face frontalization but suffers big performance drop for profile and near-profile faces due to severe texture loss and artifacts. Sagonas *et al.* [32] propose to perform joint frontal view reconstruction

and landmark detection by solving a constrained low-rank minimization problem. Kan *et al.* [33] use Stacked Progressive Auto-Encoders (SPAЕ) to rotate a profile face to frontal. Zhu *et al.* [7] propose a **Multi-View Perceptron** (MVP) model, which is able to untangle the identity and view features and meanwhile infer a full spectrum of multi-view images, given a single 2D face image. Despite its encouraging results, the synthesized faces lack fine details and tend to be blurry and unreal under a large pose. Thus, they only leverage the intermediate features for face recognition. Among current generative methods, the quality of synthesized images is still far from satisfactory for practical facial analysis like face verification and identification.

2.3 Pose Invariant Representation Learning

Conventional approaches often leverage robust local descriptors such as Gabor [8], LBP [9], HOG [10] to address local distortions and then resort to metric learning algorithms [2], [11] to achieve pose invariance. However, due to the tradeoff between invariance and discriminability, hand-crafted features usually cost huge human-engineering effort and fail to deal with extreme pose cases effectively. In contrast, deep learning methods often handle pose variance through a single pose-agnostic model or several pose-specific models with pooling operation and specific loss functions [12], [13]. For instance, the VGG-Face model [34] adopts the VGG architecture [35]. The DeepFace [36], [37] model uses a deep CNN coupled with 3D alignment. FaceNet [4] utilizes the inception architecture. The DeepID2+ [38] and DeepID3 [39] extend the FaceNet [4] model by including joint Bayesian metric learning and multi-task learning. However, such data-driven methods are too computationally complex for practical usage. Moreover, massive labelled training data covering all underlying variations are expensive and unavailable.

Our proposed PIM presents a similar idea with **Two-Pathway GAN** (TP-GAN) [40] and **Disentangled Representation learning GAN** (DR-GAN) [41]. TP-GAN considers photorealistic and identity-preserving frontal view synthesis 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 [40] and DR-GAN [41] 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 using 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 between PIM, TP-GAN [40] and DR-GAN [41] 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 facial representations end-to-end. We now present each component in details.

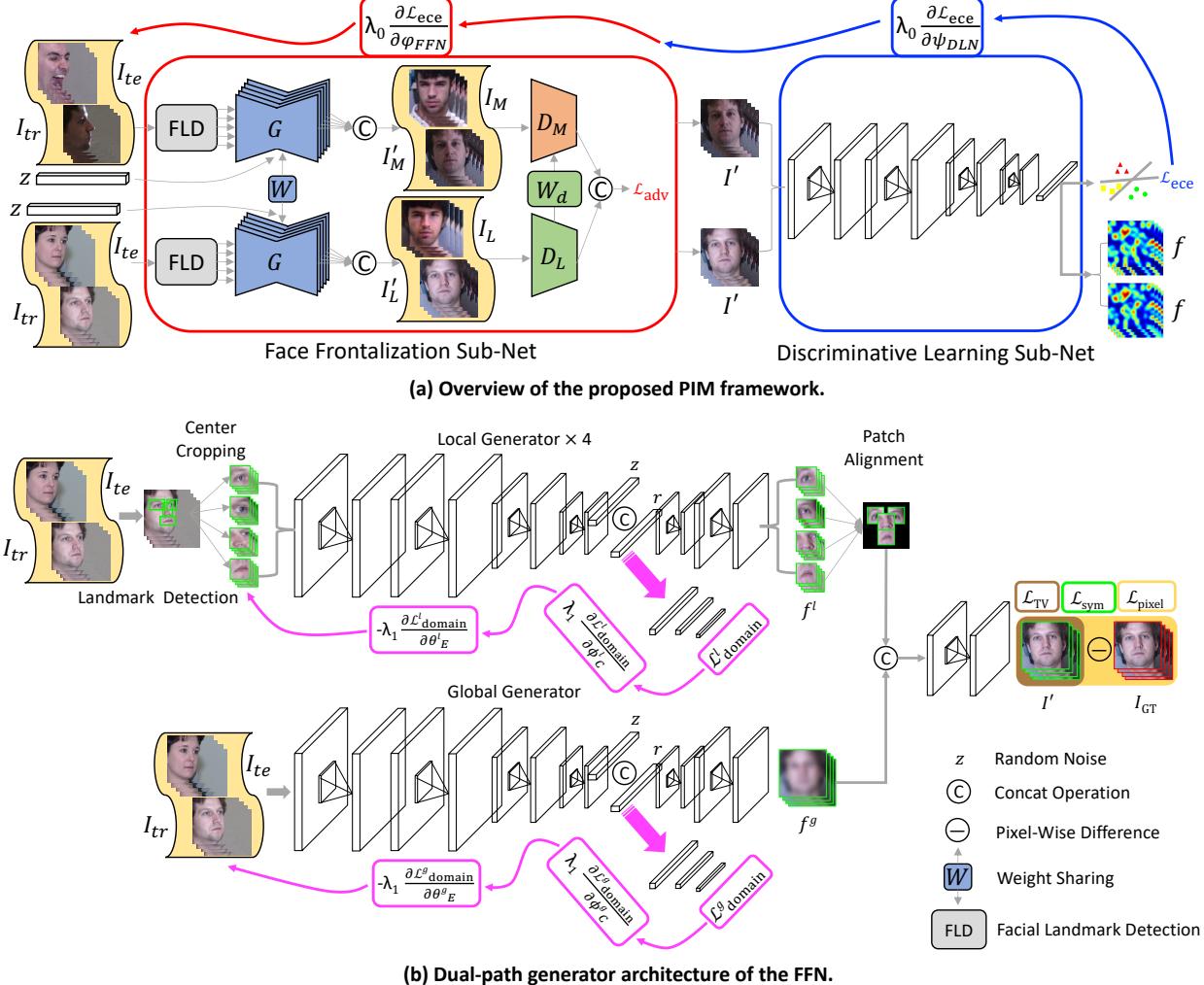


Fig. 2: Pose Invariant Model (PIM) for face recognition in the wild. The PIM contains a Face Frontalization sub-Net (FFN) and a Discriminative Learning sub-Net (DLN) that jointly learn end-to-end. FFN is a dual-path (*i.e.*, simultaneously perceiving global structures and local details) GAN augmented 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.

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 [15], [40], 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 profile 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 $\{\theta^g, \theta_i^l\}$ (here $i=1, \dots, 4$ index the four local path models) by minimizing the following composite losses:

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

where \mathcal{L}_{adv} is the **adversarial** loss for adding realism to the synthetic images and alleviating artifacts, \mathcal{L}_{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}_{sym} is the **symmetry** loss for alleviating self-occlusion issue, \mathcal{L}_{TV} is the total variation loss for reducing spiky artifacts, and $\{\lambda_k\}_{k=4}$ 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 [40], [41], we adopt $\mathcal{L}_{\text{domain}}$ to promote the emergence of features encoded by G_{θ^g} and $G_{\theta_i^l}$ that are indistinguishable w.r.t. the shift between the source (training, I_{tr}) and target (testing, I_{te}) domain. Let I_i denotes the images from both source and target domains, $y_i \in \{0, 1\}$ indicates which domain I_i is from, and $r_i = G_{\theta_E}(I_i)$ denotes the representations. The cross-domain adversarial loss is defined as follows:

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

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 when training samples from the target domain are limited. Such adapted representations are provided by augmenting the encoders of G_{θ^g} and $G_{\theta_i^l}$ 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 [42].

$\mathcal{L}_{\text{pixel}}$ is introduced to enforce the multi-scale content 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}_{sym} is introduced within the Laplacian space to exploit such prior information and impose the symmetry constraint on the recovered frontal view for alleviating self-occlusion issue:

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

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

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

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

Gradient Update with Cross-Domain Adversarial Training

For completeness, we further analyze the gradient update with cross-domain adversarial training. To make the presentation succinct, we use a notation that is slightly different from the context. Let $G_E(\cdot; \theta_E)$ be the encoder of the global- and local-path generator to transform the input from RGB space to feature space, with parameters θ_E ; let $G_D(\cdot; \theta_D)$ be the decoder of the

global- and local-path generator for frontal view recovery, with parameters θ_D ; let $G_C(\cdot; \theta_C)$ be the domain classifier of the global- and local-path generator for computation of the domain prediction, with parameters θ_C . Thus, the loss functions for the dual-path generator and cross-domain adversarial training can be respectively expressed by

$$\begin{cases} \mathcal{L}_E^i(\theta_E, \theta_D) = \mathcal{L}_E(G_D(G_E(x_i; \theta_E), \theta_D), y_i), \\ \mathcal{L}_C^i(\theta_E, \theta_C) = \mathcal{L}_C(G_C(G_E(x_i; \theta_E), \theta_C), d_i), \end{cases} \quad (5)$$

where y_i denotes the ground truths for the dual-path generator and d_i denotes the binary domain indicator.

Essentially, cross-domain adversarial training needs to optimize

$$\begin{aligned} E(\theta_E, \theta_D, \theta_C) = & \frac{1}{n} \sum_{i=1}^n \mathcal{L}_E^i(\theta_E, \theta_D) - \lambda \left(\frac{1}{n} \sum_{i=1}^n \mathcal{L}_C^i(\theta_E, \theta_C) \right. \\ & \left. + \frac{1}{m} \sum_{j=1}^m \mathcal{L}_C^j(\theta_E, \theta_C) \right), \end{aligned} \quad (6)$$

where n denotes the number of samples from the source (training) domain, and m denotes that from the target (testing) domain, by finding the saddle points $\hat{\theta}_E, \hat{\theta}_D, \hat{\theta}_C$ such that

$$\begin{cases} (\hat{\theta}_E, \hat{\theta}_D) = \operatorname{argmin}_{\theta_E, \theta_D} E(\theta_E, \theta_D, \hat{\theta}_C), \\ \hat{\theta}_C = \operatorname{argmin}_{\theta_C} E(\hat{\theta}_E, \hat{\theta}_D, \theta_C). \end{cases} \quad (7)$$

The saddle points defined by Eqn. (7) are the stationary points found via the following gradient updates:

$$\begin{cases} \theta_E \leftarrow \theta_E - \mu \left(\frac{\partial \mathcal{L}_D^i}{\partial \theta_E} - \lambda \frac{\partial \mathcal{L}_C^i}{\partial \theta_E} \right), \\ \theta_D \leftarrow \theta_D - \mu \frac{\partial \mathcal{L}_D^i}{\partial \theta_D}, \\ \theta_C \leftarrow \theta_C - \mu \lambda \frac{\partial \mathcal{L}_C^i}{\partial \theta_C}, \end{cases} \quad (8)$$

where μ denotes the learning rate. Clearly, the parameters of the dual-path generator with cross-domain adversarial training can be updated with **BackPropogation** (BP) and **Stochastic Gradient Descent** (SGD) algorithm.

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) $x_{in} \in \mathbb{R}^{w \times h \times c_{in}}$ and kernel weights $W \in \mathbb{R}^{k \times k \times c_{in} \times c_{out}}$ where k is the kernel size, the output $x_{out} \in \mathbb{R}^{w' \times h' \times c_{out}}$ of the convolutional layer can be computed as $x_{out} = W * x_{in}$, where $*$ denotes the convolution operation.

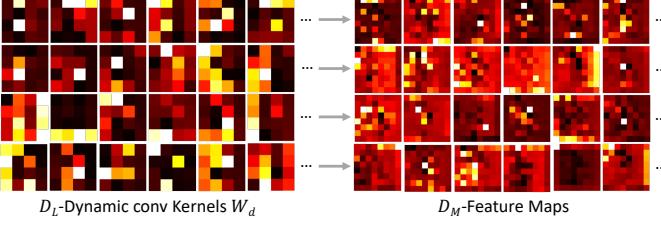


Fig. 3: Visualization of the dynamic convolution kernels predicted by D_L and corresponding feature maps of D_M . For better illustration, we vary the examples to D_L across the four rows while keeping the input of D_M unchanged. Best viewed in color.

Inspired by [43], we perform the following factorization, which is analogous to **Singular Value Decomposition (SVD)**:

$$x_{\text{out}} = U' * (W_d) *_{{c}_{\text{in}}} U * x_{\text{in}}, \quad (9)$$

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 ${c}_{\text{in}}$ channels. Under the factorization of Eqn. (9), 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 it to grow only linearly with the number of input feature map channels.

Analyses of the Learned Dynamic Convolution Kernels

We visualize the dynamic convolution kernels predicted by D_L and corresponding feature maps of D_M in Fig. 3. Different input (examples) of D_L defines different convolution kernel W_d . Applying such a dynamic convolution kernel to the same input of D_M yields different responses. In this manner, the discriminator is enhanced with more captured information for pushing the recovered frontal view face images to reside in the manifold of real images and produce visually pleasing results.

We leverage the same architecture of global-path encoder as D_M and D_L , learned separately without weight sharing, while two generator blocks in Fig. 2 (a) with their weights shared. The feature maps from D_M and D_L are further concatenated and fed into a fully connected bottleneck layer to compute \mathcal{L}_{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}_{adv} is defined as

$$\begin{aligned} \mathcal{L}_{\text{adv}} = & \frac{1}{N} \sum_i -y_i \log[D_{M \leftarrow L}(I_M, I_L)] \\ & - (1 - y_i) \log[1 - D_{M \leftarrow L}(I_M, I_L)], \end{aligned} \quad (10)$$

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 whether 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 robustness of the learned representations and address the potential over-fitting issue.

Algorithm 1 Joint learning algorithm of Pose Invariant Model (PIM).

Input: Pairs of face images under arbitrary poses from the source (training) domain I_{tr} , the corresponding frontal view ground truth I_{GT} , and the associated identity labels l_i , pairs of real frontal face images I_{real} , pairs of face images under arbitrary poses from the target (testing) domain I_{te} , pairs of Gaussian random noise z , max number of epoches (nb_e), batch size (b), number of network updates per step (nb_s), input size, landmark location template, weight decay, learning rate (lr), optimizer, keep probability of dropout, α , $\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4$;

Output: PIM FFN generator G_θ , domain classifier C_ϕ , discriminator D_φ , and DLN M_ψ ;

for $e=1, \dots, \text{nb_e}$ **do**

1. Optimize FFN D_φ ;

2. Optimize FFN G_θ ;

3. Optimize FFN C_ϕ ;

4. Optimize DLN M_ψ ;

5. Update τ_t ;

6. Measure network convergence;

7. Visualize intermediate results I' , f ;

end for

Archive G_θ , C_ϕ , D_φ , and M_ψ models for each training epoch;

end for

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 may 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:

$$p_i = \frac{\exp[\tau_t \cdot (a_i^\top f)]}{\sum_j \exp[\tau_t \cdot (a_j^\top f)]}, \quad (11)$$

where p_i denotes the predicted confidence score w.r.t. the i^{th} identity, τ_t denotes the selective attenuation factor, and a, f are ℓ_2 normalized to achieve boundary equilibrium during network training. In particular, τ_t in Eqn. (11) is updated by $\tau_{t+1} = \tau_t (1 - \frac{n}{B})^\alpha$, 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.

We conduct a toy experiment on MNIST [44] and use t-SNE [45] to visualize the learned representations in a two-dimensional space in Fig. 4. Compared with the representations learned by the standard cross-entropy with $\tau_t=1.0$, our proposed enforced cross-entropy optimization strategy adaptively adjusts the selective attenuation factor τ_t , narrows the decision boundaries

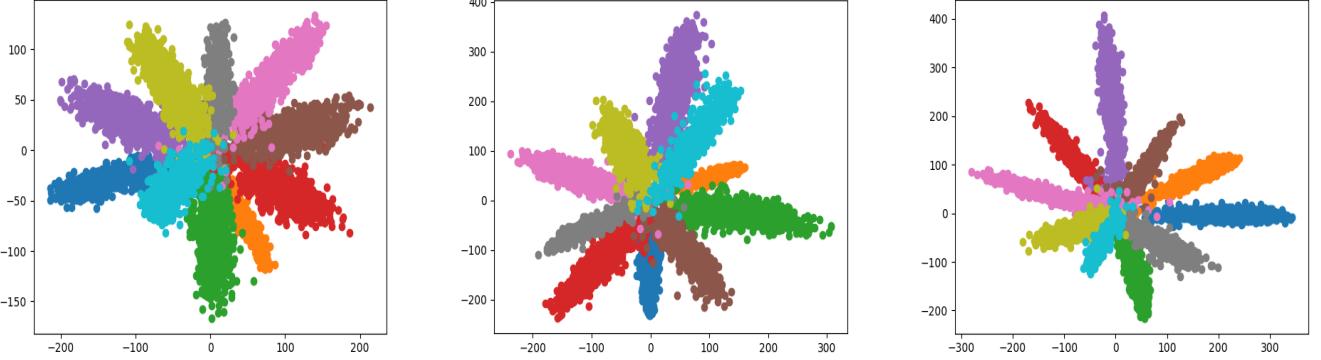


Fig. 4: Visualized comparison of the learned representations with the proposed enforced cross-entropy optimization scheme using varying selective attenuation factor τ_t on MNIST [44]. The standard cross-entropy with $\tau_t = 1.0$ (left) leads to sparser representations with small intra-class affinity and inter-class separability compared with those with $\tau_t = 0.9$ (middle) and $\tau_t = 0.7$ (right). With the adaptive decrease of the attenuation factor, the representations become more discriminative and compact with a large intra-class affinity and inter-class separability. Best viewed in color.

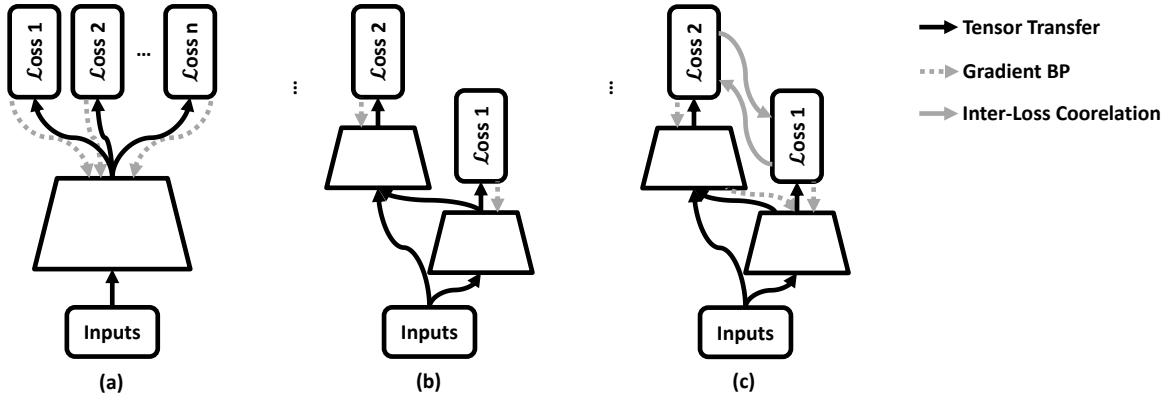


Fig. 5: Differences between the multi-task learning (a), the multi-stage learning (b) and the proposed Pose Invariant Model (PIM) (c).

and achieves discriminative and compact representations with a large intra-class affinity and inter-class separability.

The predictions of Eqn. (11) 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:

$$\mathcal{L}_{\text{ece}} = \frac{1}{N} \sum_i -l_i \log(p) - (1 - l_i) \log(1 - p), \quad (12)$$

where l_i is the face identity ground truth.

The detailed joint training procedures of our PIM are summarized in Alg. 1.

3.3 Discussions

Our model, which only has a single task and stage, is different from multi-task learning [46], [47] and multi-stage learning [33], [40], [48] for pose invariant face recognition.

As illustrated in Fig. 5 (a), multi-task learning treats face frontalization and discriminative representation learning as different tasks and the network is optimized with multiple losses. However, different losses target at different objectives, through which the mutual enhancing effect is not necessarily promised. Moreover, it encounters problems when the training data of different tasks are unbalanced.

As illustrated in Fig. 5 (b), multi-stage learning treats face frontalization and discriminative representation learning as different and separate stages, where the outputs of early stages may serve as inputs to later stages. The network of each stage is solely optimized by the specific stage-wise loss. It is easily strucked at local optimum and constrained to sub-optimal solutions.

Comparatively, as illustrated in Fig. 5 (c), our approach jointly learns face frontalization and pose invariant facial representations in an end-to-end way to allow them to benefit each other. Moreover, multi-task and multi-stage training are more difficult since different tasks and stages may have different convergence rates.

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 [21], CFP [1], IJB-A [22] and LFW [23] benchmark datasets. For quantitative evaluation, we evaluate face recognition performance using the learned facial representations with a cosine distance metric on Multi-PIE [21], CFP [1] and IJB-A [22] datasets.

4.1 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

TABLE 1: Network architecture of the global-path generator with domain classifier.

Layer	Input	Filter/Stride	Output Size
conv0	I	$7 \times 7/1$	$128 \times 128 \times 64$
conv1	conv0	$5 \times 5/2$	$64 \times 64 \times 64$
conv2	conv1	$3 \times 3/2$	$32 \times 32 \times 128$
conv3	conv2	$3 \times 3/2$	$16 \times 16 \times 256$
conv4	conv3	$3 \times 3/2$	$8 \times 8 \times 512$
[flatten, fc0]	conv4	-	512
fc1	fc0	-	256
gr0	fc1	-	256
fc2	fc1	-	256
fc3	fc2	-	2
[fc4, reshape]	fc1, z	-	$8 \times 8 \times 128$
deconv0	fc4	$3 \times 3/2$	$16 \times 16 \times 64$
deconv1	deconv0	$3 \times 3/2$	$32 \times 32 \times 32$
deconv2	deconv1	$3 \times 3/2$	$64 \times 64 \times 16$
deconv3	fc4	$3 \times 3/2$	$16 \times 16 \times 512$
deconv4	deconv3, deconv0	$3 \times 3/2$	$32 \times 32 \times 256$
deconv5	deconv4, deconv1	$3 \times 3/2$	$64 \times 64 \times 128$
deconv6	deconv5, deconv2	$3 \times 3/2$	$128 \times 128 \times 64$
conv6	deconv6, f^l	$5 \times 5/1$	$128 \times 128 \times 64$
conv7	conv6	$5 \times 5/1$	$128 \times 128 \times 32$
conv8	conv7	$5 \times 5/1$	$128 \times 128 \times 3$

(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 $0.9, 5 \times 10^{-3}, 0.1, 0.3, 5 \times 10^{-2}$ 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 [49] for landmark detection⁴. We initialize the DLN with ResNet-50 [50] and Light CNN-29 [51] architectures as our two baselines, which are pre-trained on MS-Celeb-1M [52] 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 [52]. The proposed network is implemented based on the publicly available TensorFlow [53] platform, which is trained using Adam ($\beta_1=0.5$) on three NVIDIA GeForce GTX TITAN X GPUs with 12G memory.

The network architectures of the global- and local-path generators with corresponding domain classifiers are provided in Tab. 1 and Tab. 2, respectively, where the top, middle and bottom panels show the structures of encoder, domain classifier and decoder, gr denotes the gradient reversal layer with identical forward output and reversed backward gradient for cross-domain adversarial training, f^l denotes the feature maps from local-path generator after patch alignment, and multiple input items at [fc4, reshape], deconv4, deconv5, deconv6 and conv6 indicate concatenation of tensors.

The network architecture of the siamese discriminator with dynamic convolution (D_M and D_L) is provided in Tab. 3, where the top, middle and bottom panels show the structures of D_M , D_L and the real *v.s.* fake classifier, respectively, W_d is the dynamic

3. Cross-domain adversarial training is optional; if there is no need to do domain adaptation, simply set $\lambda_1=0$.

4. For profile face images with large yaw angles, OpenPose [49] may fail to locate both eyes. In such cases, we use the detected eye after center cropping as the input left/right eye patch.

TABLE 2: Network architecture of the local-path generator with domain classifier.

Layer	Input	Filter/Stride	Output Size
conv0	I	$3 \times 3/1$	$w \times h \times 64$
conv1	conv0	$3 \times 3/2$	$w/2 \times h/2 \times 128$
conv2	conv1	$3 \times 3/2$	$w/4 \times h/4 \times 256$
conv3	conv2	$3 \times 3/2$	$w/8 \times h/8 \times 512$
[flatten, fc0]	conv3	-	512
fc1	fc0	-	256
gr0	fc1	-	256
fc2	fc1	-	256
fc3	fc2	-	2
[fc4, reshape]	fc1, z	-	$w/8 \times h/8 \times 256$
deconv0	fc4	$3 \times 3/2$	$w/4 \times h/4 \times 256$
deconv1	deconv0	$3 \times 3/2$	$w/2 \times h/2 \times 128$
deconv2	deconv1	$3 \times 3/2$	$w \times h \times 64$
conv4	deconv2	$3 \times 3/1$	$w \times h \times 64$
conv5	conv4	$3 \times 3/1$	$w \times h \times 3$

convolution kernel predicted by D_L and transferred to D_M , and multiple input items at fc0 indicate concatenation of tensors.

4.2 Ablation Studies and Model Analyses

First we perform ablation studies and model analysis on the CMU Multi-PIE [21] benchmark dataset, which is the largest multi-view face recognition benchmark and 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 $\pm 90^\circ$ 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 $\pm 90^\circ$ 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 3: Network architecture of the siamese discriminator with dynamic convolution (D_M and D_L).

Layer	Input	Filter/Stride	Output Size
conv0	I'_M or I_M	$7 \times 7/1$	$128 \times 128 \times 64$
conv1	conv0	$5 \times 5/2$	$64 \times 64 \times 64$
conv2	conv1	$3 \times 3/2$	$32 \times 32 \times 128$
conv3	conv2	$3 \times 3/2$	$16 \times 16 \times 256$
conv4	conv3	$1 \times 1/1$	$16 \times 16 \times 256$
conv5	conv4, max pool0 (W_d)	$4 \times 4/2$	$8 \times 8 \times 512$
conv6	conv5	$1 \times 1/1$	$8 \times 8 \times 512$
conv0_1	I'_L or I_L	$7 \times 7/1$	$128 \times 128 \times 64$
conv1_1	conv0_1	$5 \times 5/2$	$64 \times 64 \times 64$
conv2_1	conv1_1	$3 \times 3/2$	$32 \times 32 \times 128$
conv3_1	conv2_1	$3 \times 3/2$	$16 \times 16 \times 256$
conv4_1	conv3_1	$3 \times 3/2$	$8 \times 8 \times 512$
max pool0 (W_d)	conv4_1	$3 \times 3/2$	$4 \times 4 \times 512$
[flatten, fc0]	conv6, conv4_1	-	256
fc1	fc0	-	2

TABLE 4: Component analysis: rank-1 recognition rates (%) under Multi-PIE [21] Setting-1. b1 and b2 denote ResNet-50 [50] and Light CNN-29 [51], respectively. PIM1 and PIM2 use ResNet-50 [50] and Light CNN-29 [51] as backbone architectures, respectively.

Method	$\pm 90^\circ$	$\pm 75^\circ$	$\pm 60^\circ$	$\pm 45^\circ$	$\pm 30^\circ$	$\pm 15^\circ$
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 $G_{\theta_i^l}$	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}_{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

4.2.1 Component Level Model Evaluations

We first investigate different architectures and loss function combinations of PIM to gain an insight into their respective roles in pose invariant face recognition. We compare eight variants of PIM, including different DLN architectures (ResNet-50 [50] vs. Light CNN-29 [51]), w/o $\mathcal{L}_{\text{pixel}}$, w/o local-path generator $G_{\theta_i^l}$, 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}_{sym} , in each case.

Averaged rank-1 recognition rates are compared under Setting-1 in Tab. 4. 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 [51]. 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 $\pm 90^\circ$) 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 [21] as well as other benchmark datasets. Fig. 6 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.

4.2.2 Intermediate Results Visualizations

Most previous works on face frontalization and pose invariant representation learning address problems within a pose range of $\pm 60^\circ$, 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

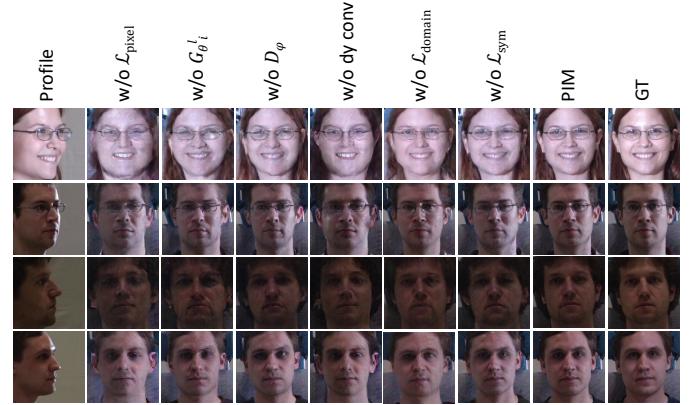


Fig. 6: Component analysis. Synthesized results of PIM and its variants.

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.

The detailed results across a wide range of poses from 15° to 90° are visualized in Fig. 7. Here, we show qualitative results of PIM under large pose variations as it is the main target, but PIM also handles large variations in expression, self-/non-self occlusion and illumination, as illustrated in Fig. 11 left-top. Our proposed PIM consistently provides faithfully high-fidelity recovered frontal view face images, and discriminative and pose invariant representations for all cases. This well verifies that the joint learning scheme of face frontalization and pose invariant representations is effective, and both intermediate results of frontalized face images and learned facial representations are beneficial to face recognition in the wild.

We further use t-SNE [45] to visualize the facial representations in a two-dimensional space in Fig. 8. The left-half illustrates the deep features of the original profile face images extracted by DLN (Light CNN-29 [51]) and the right-half illustrates the facial representations learned by PIM. It is clear that face images with a large pose are not separable in the deep feature space spanned by the DLN, revealing that even though the DLN is pre-trained on millions of images, large pose variation is still the main obstacle to pose invariant face recognition. However, with the proposed joint face frontalization and discriminative representaiton learning scheme of our PIM, the facial representations of different identities can be easily separated into corresponding clusters.

4.3 Comparisons with the State-of-the-Arts

4.3.1 Evaluations on the Multi-PIE Benchmark

Tab. 5 shows the face recognition performance comparison of our PIM with two baselines and other state-of-the-arts under Setting-1. Regardless of the adopted DLN architecture, PIM consistently achieves the best performance across all poses (except comparable with TP-GAN [40] under $\pm 30^\circ$), especially for large yaw angles. In particular, PIM (Light CNN-29 [51]) outperforms TP-GAN [40] and c-CNN Forest [56] by 10.97% and 27.74% under $\pm 90^\circ$, respectively. Note that TP-GAN [40] adopts Light CNN-29 [51] as the feature extractor which has the same architecture as

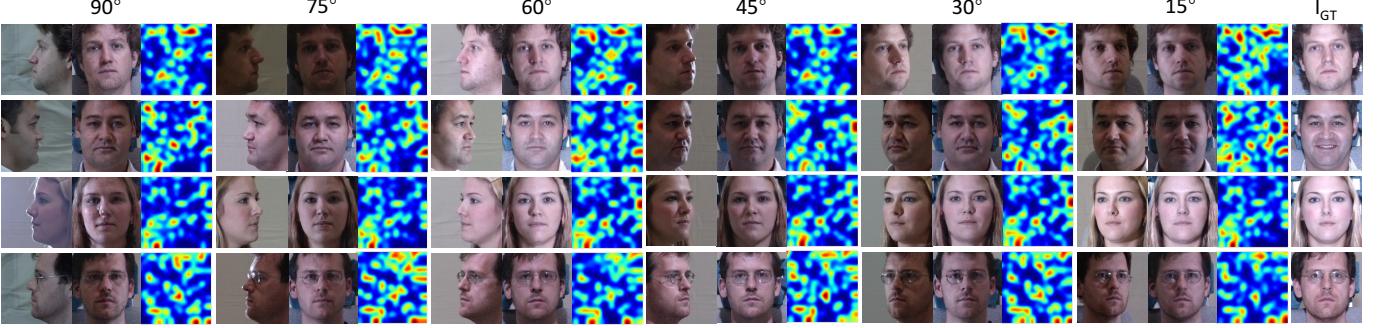


Fig. 7: Pose invariant face recognition in the wild. Each row shows a distinct identity under different poses along with other unconstrained factors (like expression, illumination, *etc.*), recovered frontal faces and learned facial representations (smoothed for better visualization, with blue indicating zero values) with our proposed PIM. The representations are extracted from the penultimate layer of PIM. The ground truth frontal face images are provided in the right-most column. These examples indicate that the facial representations learned by PIM are robust to pose variance, and the recovered frontal face images retain the intrinsic global structures and local details. Best viewed in color.

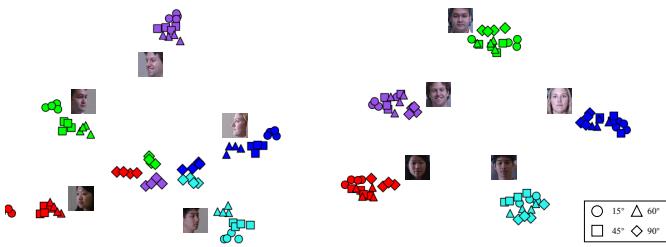


Fig. 8: Visualized comparison between the facial representations of the real profile faces (left) and the facial representations learned by PIM (right). Each visually colored cluster shows a distinct identity. Each shape represents a pose. One sampled face (left: real profile face; right: recovered frontal face) within each class is visualized for better illustration. Best viewed in color.



Fig. 9: Sample frontal-profile and frontal-frontal face pairs of CFP [1].

our DLN and c-CNN Forest [56] 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. 6 shows the face recognition comparison of our PIM with two baselines and other state-of-the-arts under Setting-2. Similar to the observation under Setting-1, PIM consistently achieves the best performance across all poses. In particular, PIM (Light CNN-29 [51]) outperforms TP-GAN [40] by 21.86% under $\pm 90^\circ$, and outperforms TP-GAN [40] and DR-GAN [41] by 10.38% and 14.90% under $\pm 60^\circ$, 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.

TABLE 5: Rank-1 recognition rates (%) across views, minor expressions and illuminations under Multi-PIE [21] Setting-1. “-” means the result is not reported. b1 and b2 denote ResNet-50 [50] and Light CNN-29 [51], respectively. PIM1 and PIM2 use ResNet-50 [50] and Light CNN-29 [51] as backbone architectures, respectively.

Method	$\pm 90^\circ$	$\pm 75^\circ$	$\pm 60^\circ$	$\pm 45^\circ$	$\pm 30^\circ$	$\pm 15^\circ$
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
CPF [18]	-	-	-	71.65	81.05	89.45
Hassner [14]	-	-	44.81	74.68	89.59	96.78
FV [54]	24.53	45.51	68.71	80.33	87.21	93.30
HPN [55]	29.82	47.57	61.24	72.77	78.26	84.23
FIP_40 [6]	31.37	49.10	69.75	85.54	92.98	96.30
c-CNN [56]	47.26	60.66	74.38	89.02	94.05	96.97
TP-GAN [40]	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 6: Rank-1 recognition rates (%) across views, illumination and sessions under Multi-PIE [21] Setting-2. “-” means the result is not reported. b1 and b2 denote ResNet-50 [50] and Light CNN-29 [51], respectively. PIM1 and PIM2 use ResNet-50 [50] and Light CNN-29 [51] as backbone architectures, respectively.

Method	$\pm 90^\circ$	$\pm 75^\circ$	$\pm 60^\circ$	$\pm 45^\circ$	$\pm 30^\circ$	$\pm 15^\circ$
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 [6]	-	-	45.90	64.10	80.70	90.70
MVP [7]	-	-	60.10	72.90	83.70	92.80
CPF [18]	-	-	61.90	79.90	88.50	95.00
DR-GAN [41]	-	-	83.20	86.20	90.10	94.00
TP-GAN [40]	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

4.3.2 Evaluations on the CFP Benchmark

The CFP [1] dataset is aimed at evaluating the strength of face verification approaches across poses, more specifically, between frontal view (yaw angle $< 10^\circ$) and profile view (yaw angle $> 60^\circ$). Sample face pairs are shown in Fig. 9. CFP contains 7,000 images of 500 subjects, where each subject has 10 frontal and 4 profile 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)

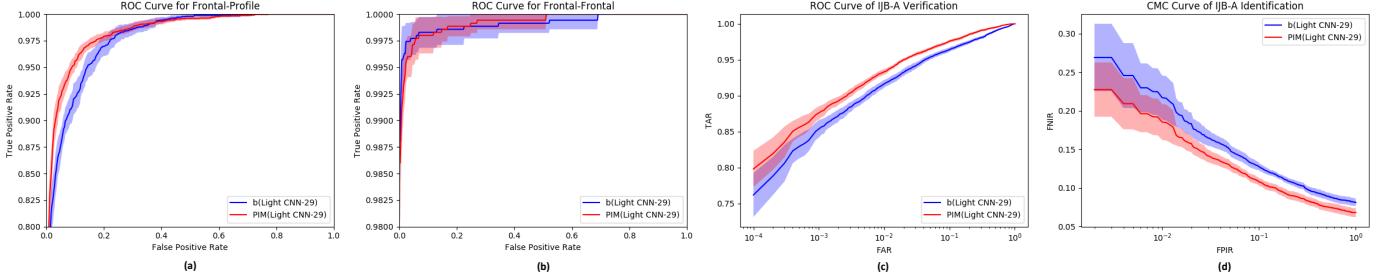


Fig. 10: Performance curve comparison between PIM (Light CNN-29 [51]) and the baseline b (Light CNN-29 [51]) on CFP [1] and IJB-A [22]. (a) ROC curve for frontal-profile on CFP [1]; (b) ROC curve for frontal-frontal on CFP [1]; (c) ROC curve for verification on IJB-A [22]; (d) CMC curve for identification on IJB-A [22]. Best viewed in color.

TABLE 7: Face recognition performance (%) comparison on CFP [1]. The results are averaged over 10 testing splits.

Method	Frontal-Profile			Frontal-Frontal		
	Acc	EER	AUC	Acc	EER	AUC
FV+DML [1]	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 [1]	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 [1]	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 [1]	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 [1]	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 [57]	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 <i>et al.</i> [58]	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 [51]	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 [51])	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

over the 10 splits for both frontal-frontal and frontal-profile face verification settings.



Fig. 11: Comparison of face frontalization results.

Tab. 7 compares the face recognition performance of our PIM (Light CNN-29 [51]) with other state-of-the-arts on the CFP [1] benchmark dataset. The results on the original images serve as our baseline. The corresponding ROC curves are provided in Fig. 10 (a) and (b). PIM achieves comparable performance as the human under frontal-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 [51]), both of which reduce the EER of human performance by around 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 [51]) and improves the accuracy by 1.13% over the 2nd-best. This shows that the facial representations learned by PIM are discriminative and robust even at extreme pose variations. Visualized comparison of face

frontalization results between PIM and DR-GAN [41] and TP-GAN [40] is provided in Fig. 11 right-top to further perceptually verify the superiority of our method.

4.3.3 Evaluations on the IJB-A Benchmark

IJB-A [22] contains 5,397 images and 2,042 videos from 500 subjects, which are split into 20,412 frames, 11.4 images and 4.2 videos per subject, captured from in-the-wild environments to avoid the near frontal bias, along with protocols for evaluation of both verification and identification. For training and testing, 10 random splits are provided by each protocol, respectively. The verification task requires the evaluation system to determine whether two input face sets are of the same subject. At a given threshold, the evaluation system measures the **True Accept Rate** (TAR), which is the fraction of genuine comparisons that correctly exceed the threshold, and the **False Accept Rate** (FAR), which is the fraction of impostor comparisons that incorrectly exceed the threshold. For identification, the evaluation system needs to determine the subject matching a probe identity from a closed set or an open set. For a closed set, the evaluation system measures the percentage of probe searches returning probe gallery mates within a given rank. For an open set, at a given threshold, the evaluation system measures the **False Positive Identification Rate** (FPIR), which is the fraction of comparisons between probe sets and non-mate gallery sets that corresponds to a match score exceeding the threshold, and the **False Negative Identification Rate** (FNIR), which is the fraction of probe searches that fail to match a mated gallery set above a score of the threshold.

The performance comparison of PIM with the baseline (Light CNN-29 [51], tested on the original images) and other state-of-the-arts on IJB-A [22] unconstrained face verification and identification protocols are given in Tab. 8. The corresponding ROC and CMC curves are provided in Fig. 10 (c) and (d). With the joint learning of face frontalization and discriminative representations, our method outperforms the baseline

TABLE 8: Face recognition performance (%) comparison on IJB-A [22]. The results are averaged over 10 testing splits. “-” means the result is not reported. Standard deviation is not available for some methods.

Method	Verification			Identification		
	TAR @ FAR=0.10	TAR @ FAR=0.01	TAR @ FAR=0.001	FNIR @ FPIR=0.10	FNIR @ FPIR=0.01	Rank-1
OpenBR [22]	0.433±0.006	0.236±0.009	0.104±0.014	0.851±0.028	0.934±0.017	0.246±0.011
GOTS [22]	0.627±0.012	0.406±0.014	0.198±0.008	0.765±0.033	0.953±0.024	0.433±0.021
Pooling faces [59]	0.631	0.309	-	-	-	0.846
LSFS [60]	0.895±0.013	0.733±0.034	0.514±0.060	0.387±0.032	0.617±0.063	0.820±0.024
Deep Multi-pose [61]	0.911	0.787	-	0.250	0.480	0.846
VGG-Face [34]	-	0.805±0.030	-	0.33±0.031	0.539±0.077	0.913±0.011
PAMs [5]	0.652±0.037	0.826±0.018	-	-	-	0.840±0.012
Masi <i>et al.</i> [62]	-	0.886	0.725	-	-	0.906
Triplet Embedding [57]	0.964±0.005	0.900±0.010	0.813±0.020	0.137±0.014	0.247±0.030	0.932±0.010
All-In-One [46]	0.976±0.004	0.922±0.010	0.823±0.020	0.113±0.014	0.208±0.020	0.947± 0.008
Light CNN-29 [51]	0.964±0.006	0.915±0.010	0.843±0.024	0.128±0.009	0.216±0.028	0.930±0.010
PIM (Light CNN-29 [51])	0.976±0.005	0.933±0.011	0.875±0.018	0.108±0.009	0.185±0.023	0.944±0.011

by 3.20% for TAR@FAR=0.001 of verification and 3.10% for FNIR@FPIR=0.01, 1.40% for rank-1 of identification. It also outperforms the 2nd-best method All-In-One [46] by 5.20% for TAR@FAR=0.001 of verification and 2.30% for FNIR@FPIR=0.01 of identification. This well shows the promising potential of recovered frontal view face images and pose invariant representations learned by PIM on large-scale and challenging unconstrained face recognition. Visualized comparison of face frontalization results between PIM and DR-GAN [41] and TP-GAN [40] is provided in Fig. 11 left-bottom to further perceptually verify the superiority of our method.

4.3.4 Evaluations on the LFW Benchmark

LFW [23] contains 13,233 face images of 5,749 identities 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 [51]) with TP-GAN [40] and DR-GAN [41], which are the state-of-the-arts aiming to generate photorealistic and identity-preserving frontal view from profiles. As in Fig. 11 right-bottom, the predictions of TP-GAN [40] suffer severe texture loss and involved artifacts, and the predictions of DR-GAN [41] deviate from true appearance seriously. 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 representations also helps, since the two sub-nets leverage each other during end-to-end training to achieve a final win-win outcome.

4.4 Failure Case Study and Discussion

As shown in Fig. 12, we observe failure cases of face frontalization by PIM when input profile faces present extremely low resolution



Fig. 12: Failure cases of face frontalization results by PIM.

(*e.g.*, 1st-panel col. 2, 3, 5, 6, 2nd-panel col. 1, 2, 3, 4, 6), heavy occlusion (*e.g.*, 1st-panel col. 2, 4, 2nd-panel col. 7), atrocious illumination (*e.g.*, 2nd-panel col. 3, 4), large expressions (*e.g.*, 1st-panel col. 7, 2nd-panel col. 6, 7) and even fragmentary elements (*e.g.*, 1st-panel col. 1, 2nd-panel col. 5). Such extreme scenarios can fail landmark detection. We plan to solve it in future.

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 adopts unsupervised cross-domain adversarial training and a “learning to learn” strategy to provide high-fidelity and identity-preserving frontal reference face images for effectively learning face representations from DLN. Comprehensive experiments demonstrate the superiority of PIM over the state-of-the-arts. We plan to apply PIM to other domain adaption and transfer learning applications in the future.

ACKNOWLEDGMENTS

The work of Jian Zhao was partially supported by China Scholarship Council (CSC) grant 201503170248. The work of Junliang Xing was partially supported by the National Science Foundation of China 61672519. The work of Jia Shi 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] 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*, 2016, pp. 1–9.
- [2] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *CVPR*, 2013, pp. 3025–3032.
- [3] B.-N. Kang and D. Kim, "Face identification using affine simulated dense local descriptors," in *URAI*. IEEE, 2013, pp. 346–351.
- [4] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.
- [5] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *CVPR*, 2016, pp. 4838–4846.
- [6] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *ICCV*, 2013, pp. 113–120.
- [7] ——, "Multi-view perceptron: a deep model for learning face identity and view representations," in *NIPS*, 2014, pp. 217–225.
- [8] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *JOSA A*, vol. 2, no. 7, pp. 1160–1169, 1985.
- [9] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *TPAMI*, vol. 28, no. 12, pp. 2037–2041, 2006.
- [10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, pp. 886–893.
- [11] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *JMLR*, vol. 10, no. Feb, pp. 207–244, 2009.
- [12] W. Chen, T.-Y. Liu, Y. Lan, Z.-M. Ma, and H. Li, "Ranking measures and loss functions in learning to rank," in *NIPS*, 2009, pp. 315–323.
- [13] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016, pp. 499–515.
- [14] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *CVPR*, 2015, pp. 4295–4304.
- [15] 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*, 2015, pp. 787–796.
- [16] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *ECCV*. Springer, 2016, pp. 57–72.
- [17] S. Xiao, L. Liu, X. Nie, J. Feng, A. A. Kassim, and S. Yan, "A live face swapper," in *ACM MM*, 2016, pp. 691–692.
- [18] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *CVPR*, 2015, pp. 676–684.
- [19] W. A. Freiwald and D. Y. Tsao, "Functional compartmentalization and viewpoint generalization within the macaque face-processing system," *Science*, vol. 330, no. 6005, pp. 845–851, 2010.
- [20] S. Ohayon, W. A. Freiwald, and D. Y. Tsao, "What makes a cell face selective? the importance of contrast," *JN*, vol. 74, no. 3, pp. 567–581, 2012.
- [21] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *JIVC*, vol. 28, no. 5, pp. 807–813, 2010.
- [22] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *CVPR*, 2015, pp. 1931–1939.
- [23] 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," pages 07–49, University of Massachusetts, Amherst, Tech. Rep., 2007.
- [24] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *CSUR*, vol. 35, no. 4, pp. 399–458, 2003.
- [25] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition," *CVIU*, vol. 101, no. 1, pp. 1–15, 2006.
- [26] R. Dave, A. Vyas, S. Mojidra, P. Desai, and P. Nikita, "Face recognition techniques: A survey," *arXiv preprint arXiv:1803.07288*, 2018.
- [27] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [28] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *arXiv preprint arXiv:1401.4082*, 2014.
- [29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.
- [30] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.
- [31] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017.
- [32] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical face frontalization," in *ICCV*, 2015, pp. 3871–3879.
- [33] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked progressive auto-encoders (spae) for face recognition across poses," in *CVPR*, 2014, pp. 1883–1890.
- [34] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.
- [35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [36] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014, pp. 1701–1708.
- [37] ——, "Web-scale training for face identification," in *CVPR*, 2015, pp. 2746–2754.
- [38] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *CVPR*, 2015, pp. 2892–2900.
- [39] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015.
- [40] 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.
- [41] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *CVPR*, 2017.
- [42] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, vol. 17, no. 59, pp. 1–35, 2016.
- [43] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *NIPS*, 2016, pp. 523–531.
- [44] Y. LeCun, "The mnist database of handwritten digits," <http://yann.lecun.com/exdb/mnist/>, 1998.
- [45] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *JMLR*, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [46] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," *arXiv preprint arXiv:1611.00851*, 2016.
- [47] X. Yin and X. Liu, "Multi-task convolutional neural network for pose-invariant face recognition," *TIP*, 2017.
- [48] J. Zhao, L. Xiong, P. K. Jayashree, J. Li, F. Zhao, Z. Wang, P. S. Pranata, P. S. Shen, S. Yan, and J. Feng, "Dual-agent gans for photorealistic and identity preserving profile face synthesis," in *NIPS*, 2017, pp. 65–75.
- [49] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR*, 2017.
- [50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
- [51] 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.
- [52] 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*, 2016, pp. 87–102.
- [53] 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."
- [54] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *BMVC*, 2013.
- [55] C. Ding and D. Tao, "Pose-invariant face recognition with homography-based normalization," *PR*, vol. 66, pp. 144–152, 2017.
- [56] 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*, 2015, pp. 3667–3675.
- [57] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in *BTAS*, 2016, pp. 1–8.
- [58] J.-C. Chen, J. Zheng, V. M. Patel, and R. Chellappa, "Fisher vector encoded deep convolutional features for unconstrained face verification," in *ICIP*, 2016, pp. 2981–2985.
- [59] T. Hassner, I. Masi, J. Kim, J. Choi, S. Harel, P. Natarajan, and G. Medioni, "Pooling faces: template based face recognition with pooled face images," in *CVPRW*, 2016, pp. 59–67.
- [60] D. Wang, C. Otto, and A. K. Jain, "Face search at scale: 80 million gallery," *arXiv preprint arXiv:1507.07242*, 2015.
- [61] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajan *et al.*, "Face recognition using deep multi-pose representations," in *WACV*, 2016, pp. 1–9.

- [62] I. Masi, A. T. Tran, J. T. Leksut, T. Hassner, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" *arXiv preprint arXiv:1603.07057*, 2016.



Jian Zhao received the Bachelors degree from Beihang University in 2012, and the Masters degree from the National University of Defense Technology in 2014. He is currently pursuing the Ph.D. degree with the Electrical and Computer Engineering Department, National University of Singapore. His research is focused on developing deep neural network models and algorithms for human-centric image understanding, applied to face recognition, image generation, and human parsing. He has published several cutting-edge projects on unconstrained/large-scale/low-shot face verification/identification and human parsing. He has won the top-3 awards several times on world-wide competitions on face recognition, human parsing and pose estimation.



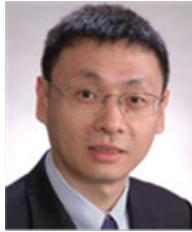
Yu Cheng is currently the undergraduate student in Nanyang Technological University, Singapore. He is currently a research intern of Learning & Vision, Core Technology Group, Panasonic R&D Center Singapore, Singapore. His research is focused on utilizing deep learning models and algorithms for facial recognition, object detection, and image generation.



Lin Xiong received the Bachelors degree from Shanxi University of Science & Technology in 2003, and the Ph.D. degree with School of Electronic Engineering, Xidian University, China, in 2014. He is currently a research engineer of Learning & Vision, Core Technology Group, Panasonic R&D Center Singapore, Singapore. His research interests include unconstrained/large-scale face recognition, person re-identification, deep learning architecture engineering, transfer learning, Riemannian manifold optimization, sparse and low-rank matrix factorization.



Xing Junliang received his dual B.S. degrees in computer science and mathematics from Xi'an Jiaotong University, Shaanxi, China, in 2007, and the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2012. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Dr. Xing was the recipient of Google Ph.D. Fellowship 2011, the Excellent Student Scholarships at Xi'an Jiaotong University from 2004 to 2007 and at Tsinghua University from 2009 to 2011. He has published more than 70 papers on international journals and conferences. His current research interests mainly focus on computer vision problems related to faces and humans.



Shuicheng Yan is currently the Vice-President and the Chief Scientist with Qihoo 360 Technology Company Ltd., and the Head of the 360 Artificial Intelligence Institute. He is also a tenured Associate Professor with the National University of Singapore. He has authored/co-authored over 500 high quality technical papers, with Google Scholar citation over 25 000 times and an h-index 70. His research areas include computer vision, machine learning, and multimedia analysis. He is an IAPR Fellow and the ACM Distinguished Scientist. His team received seven times winner or honorable-mention prizes in five years over PASCAL, VOC, and ILSVRC competitions, which are core competitions in the field of computer vision, along with over ten times the Best (student) Paper Awards and especially a Grand Slam with the ACM MM, the top conference in the field of multimedia, including the Best Paper Award, the Best Student Paper Award, and the Best Demo Award. He is a TR Highly Cited Researcher of 2014, 2015, and 2016.



Jiashi Feng received the Ph.D. degree from the National University of Singapore (NUS) in 2014. He was a Post-Doctoral Research Fellow with the University of California, Berkeley. He joined NUS as a Faculty Member, where he is currently an Assistant Professor with the Department of Electrical and Computer Engineering. His research areas include computer vision, machine learning, object recognition, detection, segmentation, robust learning, and deep learning.