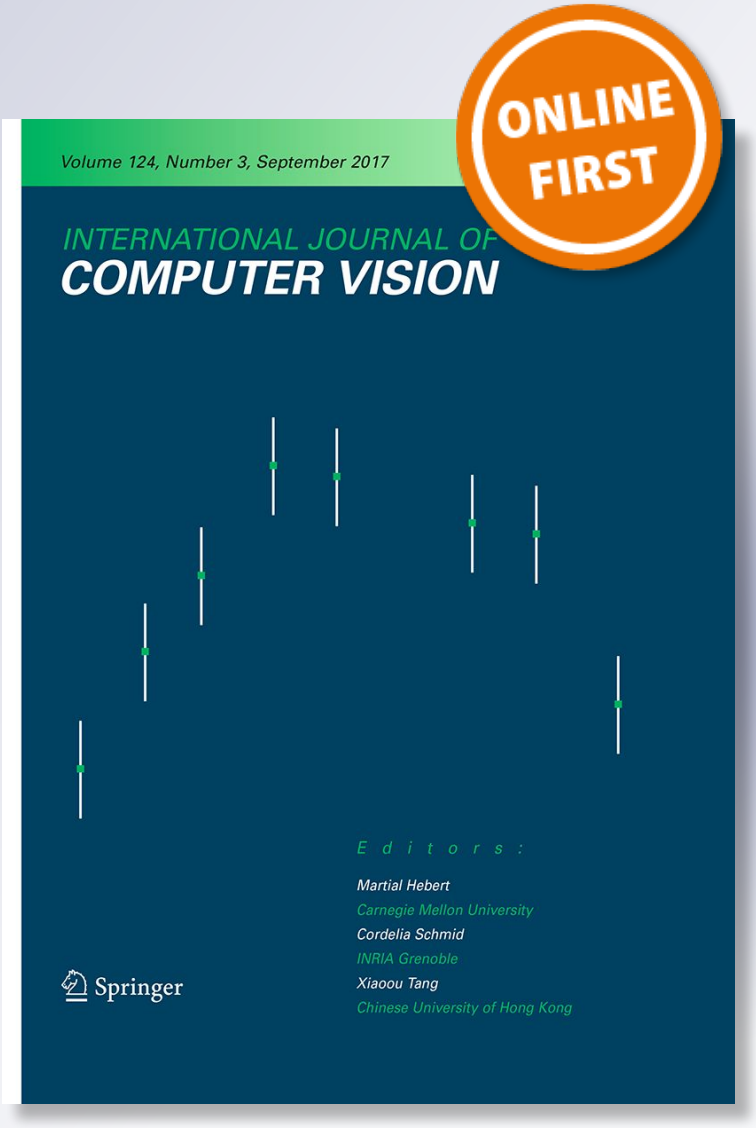
Recognizing Profile Faces by Imagining Frontal View

Jian Zhao, Junliang Xing, Lin Xiong, Shuicheng Yan & Jiashi Feng

International Journal of Computer Vision

ISSN 0920-5691

Int J Comput Vis
DOI 10.1007/s11263-019-01252-7



Springer

International Journal of Computer Vision
https://doi.org/10.1007/s11263-019-01252-7

Recognizing Profile Faces by Imagining Frontal View

Jian Zhao[1] · Junliang Xing[2] · Lin Xiong[3] · Shuicheng Yan[4,5] · Jiashi Feng[4]

Received: 8 December 2018 / Accepted: 9 October 2019

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 which simultaneously perceives global structures and local details, incorporating an unsupervised cross-domain adversarial training and a meta-learning ("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.

Keywords Pose-invariant face recognition · Face frontalization · Cross-domain adversarial learning · Meta-learning · Learning to learn · Enforced cross-entropy optimization · Generative adversarial networks

1 Introduction

Face recognition has been a key problem in computer vision. 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* (Wang et al. 2018a, b; Hao et al. 2017; Zhao et al. 2019). Among them, the one that harms face recognition performance arguably the most is pose variation. In fact, as demonstrated in Sengupta et al. (2016), the performance of most face recognition models degrades by over 10% from frontal-frontal to frontal-profile verification because discriminative descriptors suffer from misallignment issues. In contrast, humans can recognize faces in presence of large

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

Communicated by Tinne Tuytelaars.

✉ Junliang Xing
jlxing@nlpr.ia.ac.cn

Jian Zhao
zhaojian90@u.nus.edu
https://zhaoj9014.github.io

Lin Xiong
Lin.Xiong@jd.com

Shuicheng Yan
eleyans@nus.edu.sg

Jiashi Feng
elefjia@nus.edu.sg

1 Institute of North Electronic Equipment, Beijing, China

2 National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

3 JD Digits, Silicon Valley, USA

4 National University of Singapore, Singapore, Singapore

5 Yitu Technology, Beijing, China

Published online: 05 November 2019

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 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 (Chen et al. 2013; Kang and Kim 2013) or data-driven deep learning solutions from ample face data (Schroff et al. 2015; Masi et al. 2016), 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 (Zhu et al. 2013, 2014).

For the first category, conventional approaches often leverage robust local descriptors such as Gabor (Daugman 1985), Local Binary Pattern (LBP) (Ahonen et al. 2006), Histograms of Oriented Gradient (HOG) (Dalal and Triggs 2005) to tackle local distortions and then resort to metric learning algorithms (Chen et al. 2013; Weinberger and Saul 2009) 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 (Chen et al. 2009) or center loss (Wen et al. 2016) 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) (Hassner et al. 2015; Zhu et al. 2015). These methods are prominent at normalizing frontal or near-frontal faces, but their performance decreases for profile or near-profile[2] faces due to severe texture loss and involved artifacts. Recently, deep learning based methods (Zhao et al. 2018; Cao et al. 2018; Hu et al. 2018; Xiao et al. 2016a, b; Yim et al. 2015; Zhu et al. 2014; Zhao et al. 2019) are proposed for face frontalization. For instance, (Zhu et al. 2014) 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 (Freiwald and Tsao 2010; Ohayon et al. 2012) 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 meta-learning ("learning to learn")

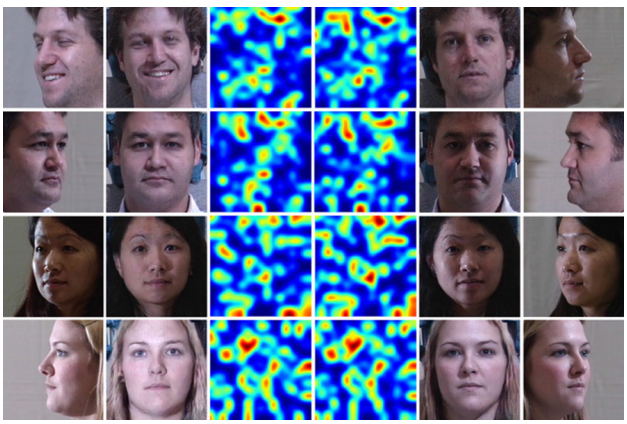


Fig. 1 Pose-invariant face recognition in the wild. Col. 1 and 6: distinct identities under different poses with other unconstrained factors (different expressions and lighting conditions); Col. 2 and 5: recovered frontal faces with our proposed Pose Invariant Model (PIM); Col. 3 and 4: learned facial 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 (deep level) of PIM

[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.

strategy using 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 meta-learning ("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 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.

A preliminary version of this work was accepted in IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018 (Zhao et al. 2018). We extend it in numerous ways:(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. (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 (Gross et al. 2010), CFP (Sengupta et al. 2016), IJB-A (Klare et al. 2015) and LFW (Huang et al. 2007) 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 to facilitate evaluation analysis across different settings. (6) We supplement an additional experiment on IJB-A 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., simutaneously perceving global structures and local details) Generative Adversarial Network (GAN) incorporating unsupervised cross-domain adversarial training and a meta-learning ("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 **N**etwork (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, CFP, IJB-A and LFW 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 (Zhao et al. 2003; Bowyer et al. 2006; Dave et al. 2018).

2.1 Generative Adversarial Networks

As one of the most significant advancements in the research of deep generative models (Kingma and Welling 2013; Rezende et al. 2014), GAN has drawn substantial attention

from the deep learning and computer vision community ever since it was first introduced by Goodfellow et al. (2014). 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 (2014) introduce the conditional version of GAN, conditioned on both the generator and discriminator for effective image tagging. Berthelot et al. (2017) 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 (Hassner et al. 2015; Zhu et al. 2015), statistical modelling (Sagonas et al. 2015), and deep learning based methods (Zhao et al. 2018; Cao et al. 2018; Hu et al. 2018; Kan et al. 2014; Yim et al. 2015; Zhao et al. 2019). For instance, (Hassner et al. 2015) 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. (2015) propose to perform joint frontal view reconstruction and landmark detection by solving a constrained low-rank minimization problem. Kan et al. (2014) use Stacked Progressive Auto-Encoders (SPAE) to rotate a profile face to frontal. Zhu et al. (2014) 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.

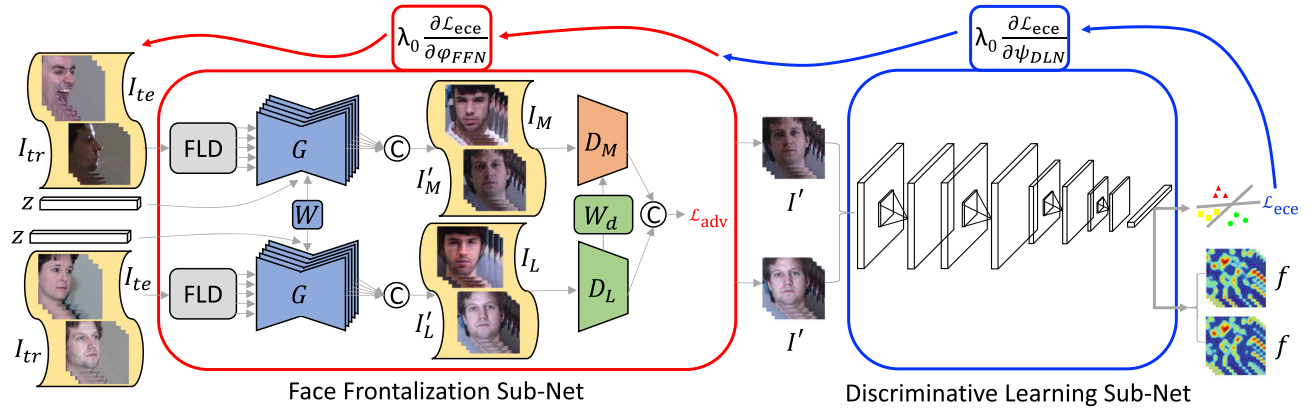
2.3 Pose-Invariant Representation Learning

Conventional approaches often leverage robust local descriptors such as Gabor (Daugman 1985), LBP (Ahonen et al. 2006), HOG (Dalal and Triggs 2005) to address local distortions and then resort to metric learning algorithms (Chen et al. 2013; Weinberger and Saul 2009) 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 (Chen et al. 2009; Wen et al. 2016). For instance, the VGG-Face model (Parkhi et al. 2015) adopts the VGG architecture (Simonyan and Zisserman 2014). The DeepFace (Taigman et al. 2014, 2015) model uses a deep CNN coupled with 3D alignment. FaceNet (Schroff et al. 2015) utilizes the inception architecture. The DeepID2+ (Sun et al. 2015a) and DeepID3 (Sun et al. 2015b) extend the FaceNet (Schroff et al. 2015) 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) (Huang et al. 2017) and Disentangled Represen-tation learning GAN (DR-GAN) (Tran et al. 2017). 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 and DR-GAN 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 meta-learning ("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 and DR-GAN are provided in Sect. 4.

3 Pose-Invariant Model

As shown in Fig. 2a, 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.



(a) Overview of the proposed PIM framework.

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

Fig. 2 Pose-**I**nvariant Model (PIM) for face recognition in the wild. The PIM contains a Face **F**rontalization 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 meta-learning ("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 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. Best viewed in color

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 Huang et al. (2017) and Zhu et al. (2015), where one path aims to infer the global sketch and the other to attend to local facial details, as shown in Fig. 2b.

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 (Simon et al. 2017). Given an input face image I, to effectively integrate informa-

tion 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'.

Face frontalization is indeed a one-to-many mapping problem. 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. Thus, there may exist slight intra-class variance of the frontalized faces during inference. To facilitate face frontalization with well-preserved identity information, we have carefully designed the network with an enforced cross-entropy loss to achieve discriminative and compact representations with a large intra-class affinity and inter-class separability.

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, ..., 4 index the four local path models) by minimizing the following composite losses:

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

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-**e**ntropy 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 **v**ariation loss for reducing spiky artifacts, and $\{\lambda_k\}_{k=0}^4$ 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 (Huang et al. 2017; Tran et al. 2017), 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:

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

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 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. 2b), as inspired by Ganin et al. (2016).

$\mathcal{L}_{\text{pixel}}$ 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, $\mathcal{L}_{\text{sym}}$ 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:

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

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:

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

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

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

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

$$E(\theta_E, \theta_D, \theta_C) = \frac{1}{n} \Sigma_{i=1}^{n} \mathcal{L}_E^i(\theta_E, \theta_D) - \lambda \left(\frac{1}{n} \Sigma_{i=1}^{n} \mathcal{L}_C^i(\theta_E, \theta_C) + \frac{1}{m} \Sigma_{j=1}^{m} \mathcal{L}_C^j(\theta_E, \theta_C) \right), \tag{6}$$

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

$$\begin{cases} (\hat{\theta}_E, \hat{\theta}_D) = argmin_{\theta_E, \theta_D} E(\theta_E, \theta_D, \hat{\theta}_C), \\ \hat{\theta}_C = argmin_{\theta_C} E(\hat{\theta}_E, \hat{\theta}_D, \theta_C). \end{cases} \tag{7}$$

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

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

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 meta-learning ("learning to learn") strategy using a siamese adversarial pixel-wise discriminator with dynamic convolution, as shown in Fig. 2a. 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 Bertinetto et al. (2016), we perform the following factorization, which is analogous to Singular Value **D**ecomposition (SVD):

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

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_{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. 2a 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 $\mathcal{L}_{\text{adv}}$, which serves as a supervision to push the synthesized image to reside in the manifold ofphotorealistic

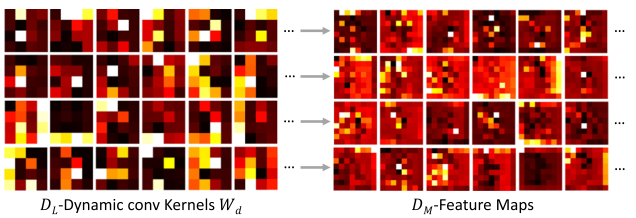


Fig. 3 Visualization of the dynamic convolution kernels predicted by D_L and corresponding feature maps of D_M. For better illustration, we vary the examples to D_L across the four rows while keeping the input of D_M unchanged. Best viewed in color

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, λ_0, λ_1, λ_2, λ_3, λ_4;

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

```
for e=1, ··· , nb_e do
    for s=1, ··· , nb_s 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
```

frontal view images, prevent blur effect, and produce visually pleasing results. In particular, $\mathcal{L}_{\text{adv}}$ is defined as

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

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[3] 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.

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 and safe 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:

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

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 $\tau_{t+1} = \tau_t \left(1 - \frac{n}{B}\right)^\alpha$, 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 for hard samples, narrows the decision boundary, controls the intra-class affinity and inter-class distance, and enforces the learned model to classify them better. The angular loss in SphereFace Liu et al. (2017) shares a similar goal but defines a margin based on angular distance.

We conduct a toy experiment on MNIST (LeCun 1997) and use t-SNE (Maaten and Hinton 2008) to visualize the learned representations in a two-dimensional space in Fig. 4. Compared with the representations learned by the standard cross-entropy with τ_t=1.0, our proposed enforced cross-entropy optimization strategy adaptively adjusts the selective attenuation factor τ_t, narrows the decision boundaries 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. 2a), which is an enforced optimization scheme. The detailed joint training procedures of our PIM are summarized in Algorithm 1.

3.3 Discussions

The paradigm of the proposed Pose Invariant Model (PIM) learning is different from those of multi-task learning (Ranjan et al. 2016; Yin and Liu 2017) and multi-stage learning (Kan et al. 2014; Huang et al. 2017; Zhao et al. 2017, 2018) for pose-invariant face recognition.

[3] The DLN is not restricted to a certain network architecture. Advanced networks can be deployed for high performance.

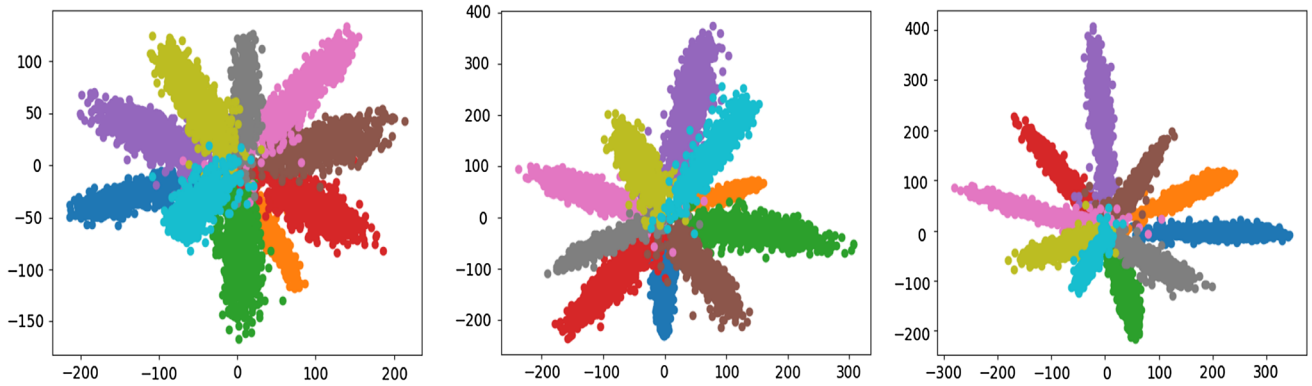


Fig. 4 Visualized comparison of the learned representations with the proposed enforced cross-entropy optimization scheme using varying selective attenuation factor τ_t on MNIST (LeCun 1997). The standard cross-entropy with $\tau_t = 1.0$ (left) leads to sparser representations with small intra-class 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

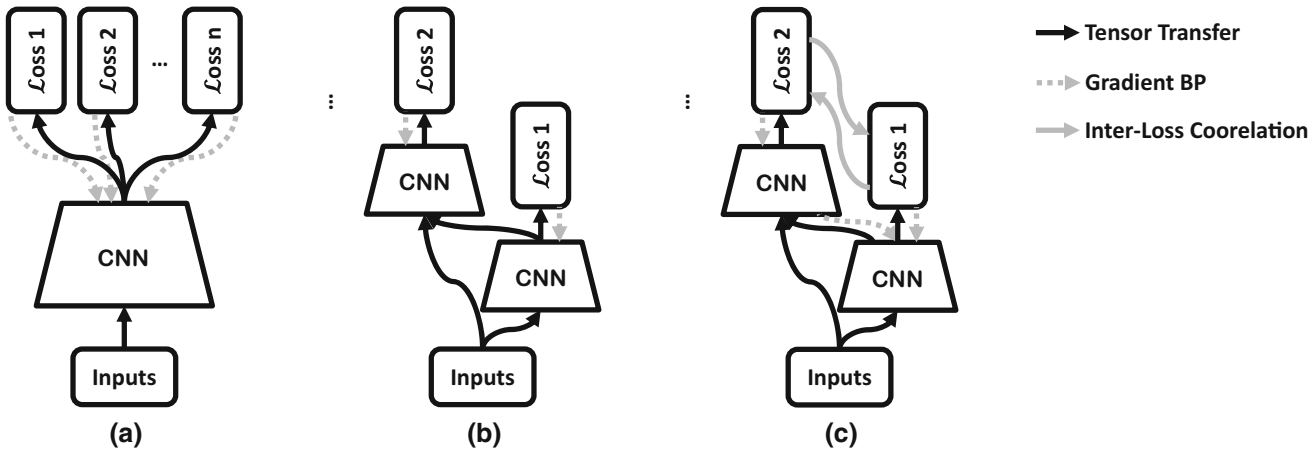


Fig. 5 Differences between the paradigms of multi-task learning (**a**), multi-stage learning (**b**) and the proposed Pose Invariant Model (PIM) learning (**c**). (a) treats face frontalization and discriminative representation learning as different tasks and the network is optimized with multiple losses. However, different losses target at different objectives, through which the mutual enhancing effect is not necessary promised. Whether local optimum can be achieved depends largely on the initialization of the network. Moreover, it encounters problems when the training data of different tasks are unbalanced. (b) 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 specic stage-wise loss. It is easily strucked at local optimum and constrained to sub-optimal solutions, since the solutions of earlier stages are fixed and unchanged anymore during the learning (optimization) of later stages. (c) is an implicit combination of (a) and (b), jointly learning face frontalization and pose-invariant facial representations in an end-to-end way to allow them to benefit each other (the solutions of earlier stages are fine-tuned during the learning of later stages in the collaborative optimization by different losses at different stages). In comparison, multi-task and multi-stage training are more difficult since different tasks and stages may have different convergence rates

As illustrated in Fig. 5a, the paradigm of 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 necessary promised. Whether local optimum can be achieved depends largely on the initialization of the network. Moreover, it encounters problems when the training data of different tasks are unbalanced.

As illustrated in Fig. 5b, the paradigm of 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 specic stage-wise loss. It is easily strucked at local optimum and constrained to sub-optimal solutions, since the solutions of earlier stages are fixed and unchanged anymore during the learning (optimization) of later stages.

Comparatively, as illustrated in Fig. 5c, our approach is an implicit combination of paradigm (a) and (b), jointly learning face frontalization and pose-invariant facial representations in an end-to-end way to allow them to benefit each other (the solutions of earlier stages are fine-tuned during the learning of later stages in the collaborative optimization by different losses at different stages). In comparison, multi-task and multi-stage training are more difcult 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 (Gross et al. 2010), CFP (Sengupta et al. 2016), IJB-A (Klare et al. 2015) and LFW (Huang et al. 2007) benchmark datasets. For quantitative evaluation, we evaluate face recognition performance using the learned facial representations with a cosine distance metric on Multi-PIE, CFP and IJB-A 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 (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 $\lambda_i, i \in \{0, 1^4, 2, 3, 4\}$ 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 (Simon et al. 2017) for landmark detection[5]. We initialize the DLN with the Light CNN-29 (Wu et al. 2015) architecture as our baseline, which is pre-trained on MS-Celeb-1M (Guo et al. 2016) 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. The proposed network is implemented based on the publicly available TensorFlow [63] platform, which is trained using Adam ($\beta_1 = 0.5$) on three NVIDIA GeForce GTX TITAN X GPUs with 12G memory.

[4] Cross-domain adversarial training is optional; if there is no need to do domain adaptation, simply set λ_1=0.

[5] For profile face images with large yaw angles, OpenPose 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 1 Network architecture of the global-path generator with domain classifier

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

The network architectures of the global- and local-path generators with corresponding domain classifiers are provided in Table 1 and Table 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 Table 3, where the top, middle and bottom panels show the structures of D_M, D_L and the real versus fake classifier, respectively, W_d is the dynamic convolution kernel predicted by D_L and transfered to D_M, and multiple input items at fc0 indicate concatenation of tensors.

Discussion on Choice on Trade-Off Hyperparmeter $\lambda\mathcal{L}_{\text{adv}}$ serves as a regularization term that penalizes higher-order inconsistencies between the synthetic images and corresponding ground truth for adding realism and alleviating artifacts. $\mathcal{L}_{\text{ece}}$ serves as an enforced cross-entropy supervision for achieving discriminative and compact representations with a large intra-class afnity and inter-class separability.

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	$3 \times 3/1$	$w \times h \times 3$

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

$\mathcal{L}_{\text{domain}}$ serves as a regularization term that reduces the domain discrepancy and helps the generator achieve similar face frontalization performance across different domains. $\mathcal{L}_{\text{pixel}}$ serves as a deep supervision for enforcing the multi-scale content consistency between the nal frontalized face and corresponding ground truths. $\mathcal{L}_{\text{sym}}$ serves as a regularization term within the Laplacian space to exploit facial symmetry prior information for alleviating self-occlusion issue. $\mathcal{L}_{\text{TV}}$ serves as a standard regularization term on the synthesized results for reducing spiky artifacts. We choose the trade-off hyperparameters $\lambda_i, i \in \{0, 1, 2, 3, 4\}$ based on the above intuition to balance the effects of different losses, such that the magnitudes of different losses are within the same range during the training process. As a common practice in tuning neural network parameters, these hyperparameters are first estimated in the range by calculating the magnitudes of the 6 terms during the training stage. Then they shifted to some neighboring values to check if the performance can be improved. We cannot assure the adopted parameter values obtains the best performance, though the produced results are already very competitive. A sensitivity analysis is performed in Sect. 4.2.1 to gain an insight into the respective influence of different choices of λ on the final performance.

4.2 Ablation Studies and Model Analyses

First we perform ablation studies and model analysis on the CMU Multi-PIE (Gross et al. 2010) 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 $\pm 90°$ 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 $\pm 90°$ 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.

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 seven variants of PIM, including baseline[6] (b: Light CNN-29 Wu et al. 2015), 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 $\mathcal{L}_{\text{sym}}$, in each case.

[6] The results on the profile (original) images serve as our baseline.

Table 4 Component analysis: rank-1 recognition rates (%) under Multi-PIE Setting-1

Method	±90°	±75°	±60°	±45°	±30°	±15°
b	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 $\mathcal{L}_{\text{sym}}$	72.30	90.40	96.80	98.20	99.30	99.80
λ_3=0.000	72.30	90.40	96.80	98.20	99.30	99.80
λ_3=0.025	74.50	91.00	97.50	98.30	99.40	99.80
λ_3=0.050	75.00	91.20	97.70	98.30	99.40	99.80
λ_3=0.075	75.10	91.40	97.80	98.30	99.40	99.80
λ_3=0.100	72.90	90.70	97.00	98.10	99.30	99.80
PIM	**75.00**	**91.20**	**97.70**	**98.30**	**99.40**	**99.80**

The best results are highlighted in bold

Averaged rank-1 recognition rates are compared under Setting-1 in Table 4. By comparing the results from the top and bottom panels, we observe that an improvement of 42.00% under ±90° can be achieved with our joint face frontalization and discriminative representation learning fra-mework. The pixel loss, dual-path generator and the meta-learning ("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 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 meta-learning ("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.

Sensitivity Analyses on Choice on Trade-Off Hyperparmeter λ We then perform a sensitivity analysis to gain an insight into the respective influence of different choices of λ on the final performance. Since the searching space is extremely large if we vary the values of $\lambda_i, i \in \{0, 1, 2, 3, 4\}$ individually and perform corresponding ablation study one by one, due to the limitation on computing resource, here we choose λ_3 of $\mathcal{L}_{\text{sym}}$ as an example, and evaluate its influence on the final performance by adjusting its value 5 times with a stride of 0.025, i.e., {0.000, 0.025, 0.050, 0.075, 0.100}. The

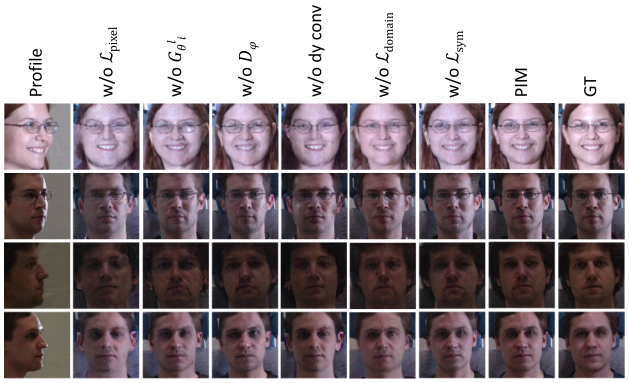


Fig. 6 Component analysis. Synthesized results of PIM and its variants

results are reported in the last-but-one panel of Table 4. By comparing the results of λ_3=0.000 and those of λ_3 >0.000, we observe that $\mathcal{L}_{\text{sym}}$ benefits the recognition performance by relieving self-occlusion for extreme poses. By comparing the results of $\lambda_3 \in \{0.025, 0.050, 0.075\}$, we find that our choice on the trade-off hyperparameter λ_3 is quite robust within a small range of perturbation, since the three groups of results are comparable even under large poses. The results of $\lambda_3 = 0.010$ drop significantly due to over-relying on $\mathcal{L}_{\text{sym}}$ and other components contributing less. Moreover, we also observe that our hyperparameter setting may not be optimal since the results of $\lambda_3 = 0.075$ are slightly improved under the poses larger than 60° compared with those of $\lambda_3 = 0.050$. We plan to solve the optimal hyperparameter searching in our future work.

Complexity The proposed PIM needs 35.1 GFLOPs given the input RGB image of 128 × 128 pixels and has 64.07M parameters. The training process of PIM takes about 16 hours on three NVIDIA GeForce GTX TITAN X GPUs with 12G memory. The inference time (NVIDIA GeForce GTX TITAN X GPU, Intel Core i7-4930K CPU@3.40GHZ) of PIM is about 72ms (60ms for FFN and 12ms for DLN, respectively) per input, which is applicable to real scenarios.

4.2.2 Intermediate Results Visualizations

Most previous works on face frontalization and pose-invariant representation learning address problems 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.

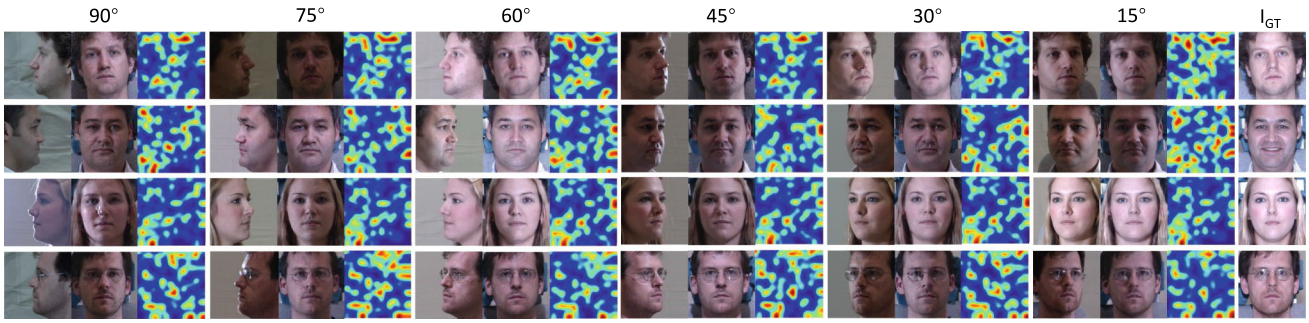


Fig. 7 Pose-invariant face recognition in the wild. Each row shows a distinct identity under different poses along with other unconstrained factors (like expression, illumination, etc.), recovered frontal faces and learned facial representations (smoothed for better visualization, with blue indicting zero values) with our proposed PIM. The representations are extracted from the penultimate layer (deep level) 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

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[7] 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 (Maaten and Hinton 2008) 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 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.

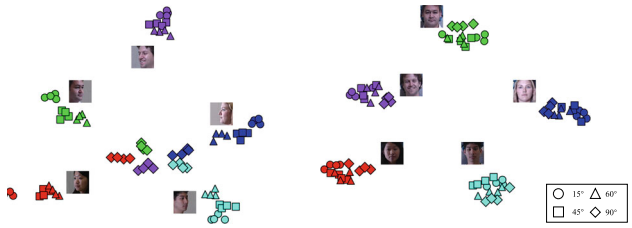


Fig. 8 Visualized comparison between the facial representations of the real profile faces (left) and the facial representations learned by PIM (right). Each visually colored cluster shows a distinct identity. Each shape represents a pose. One sampled face (left: real profile face; right: recovered frontal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

4.3 Comparisons with the State-of-the-Arts

4.3.1 Evaluations on the Multi-PIE Benchmark

Table 5 shows the face recognition performance comparison of our PIM with the baseline and other state-of-the-arts under Setting-1. PIM consistently achieves the best performance across all poses (except comparable with TP-GAN (Huang et al. 2017) under ±45° and ±30°), especially for large yaw angles. In particular, PIM outperforms TP-GAN and c-CNN Forest (Xiong et al. 2015) by 10.97% and 27.74% under ±90°, respectively. Note that TP-GAN adopts Light CNN-29 (Wu et al. 2015) as the feature extractor which has the same architecture as our DLN and c-CNN Forest is an ensemble of three models, while our PIM has a more effective and efficient joint training scheme and a much simpler network architecture.

Table 6 shows the face recognition comparison of our PIM with the baseline and other state-of-the-arts under Setting-2.

[7] Each pixel in the feature map has a very large receptive field, without one-to-one correspondence with the input RGB face image nor the recovered frontal-view face image. The backbone of DLN is initialized with Light CNN-29 (Wu et al. 2015) and pre-trained on MS-Celeb-1M (Guo et al. 2016) (large-scale dataset for face recognition) and finetuned on the target dataset. Thus, the network has achieved sufficient robustness against background variance, which will focus on the facial region during recognition.

Table 5 Rank-1 recognition rates (%) across views, minor expressions and illuminations under Multi-PIE Setting-1. "-" means the result is not reported

Method	±90°	±75°	±60°	±45°	±30°	±15°
b	33.00	76.10	95.20	97.90	99.20	99.80
CPF (Yim et al. 2015)	–	–	–	71.65	81.05	89.45
Hassner (Hassner et al. 2015)	–	–	44.81	74.68	89.59	96.78
FV (Simonyan et al. 2013)	24.53	45.51	68.71	80.33	87.21	93.30
HPN (Ding and Tao 2017)	29.82	47.57	61.24	72.77	78.26	84.23
FIP_40 (Zhu et al. 2013)	31.37	49.10	69.75	85.54	92.98	96.30
c-CNN (Xiong et al. 2015)	47.26	60.66	74.38	89.02	94.05	96.97
TP-GAN (Huang et al. 2017)	64.03	84.10	92.93	**98.58**	**99.85**	99.78
PIM	**75.00**	**91.20**	**97.70**	98.30	99.40	**99.80**

The best results are highlighted in bold

Table 6 Rank-1 recognition rates (%) across views, illumination and sessions under Multi-PIE Setting-2

Method	±90°	±75°	±60°	±45°	±30°	±15°
b	27.10	68.70	91.40	97.70	98.60	99.10
FIP (Zhu et al. 2013)	–	–	45.90	64.10	80.70	90.70
MVP (Zhu et al. 2014)	–	–	60.10	72.90	83.70	92.80
CPF (Yim et al. 2015)	–	–	61.90	79.90	88.50	95.00
DR-GAN (Tran et al. 2017)	–	–	83.20	86.20	90.10	94.00
TP-GAN (Huang et al. 2017)	64.64	77.43	87.72	95.38	98.06	98.68
PIM	**86.50**	**95.00**	**98.10**	**98.50**	**99.00**	**99.30**

The best results are highlighted in bold
"–" means the result is not reported

Similar to the observation under Setting-1, PIM consistently achieves the best performance across all poses. In particular, PIM outperforms TP-GAN by 21.86% under ±90°, and outperforms TP-GAN and DR-GAN (Tran et al. 2017) by 10.38% and 14.90% under ±60°, respectively. This well verifies the superiority of our proposed cross-domain adversarial training, the meta-learning ("learning to learn") strategy using the siamese discriminator with dynamic convolution and the enforced cross-entropy optimization strategy in improving the overall recognition performance.

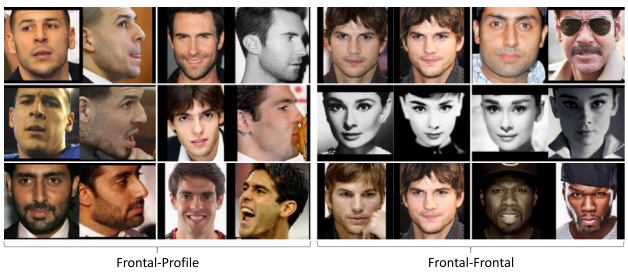


Fig. 9 Sample frontal-profile and frontal-frontal face pairs of CFP (Sengupta et al. 2016)

4.3.2 Evaluations on the CFP Benchmark

The CFP (Sengupta et al. 2016) dataset is aimed at evaluating the strength of face verification approaches across poses, more specifically, between frontal view (yaw angle$<$ 10°) and profile view (yaw angle$>$ 60°). 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 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, **E**qual **E**rror **R**ate (EER) and **A**rea **U**nder **C**urve (AUC) over the 10 splits for both frontal-frontal and frontal-profile face verification settings.

Table 7 compares the face recognition performance of our PIM with other state-of-the-arts on the CFP benchmark dataset. The results on the original images serve as our baseline. The corresponding ROC curves are provided in Fig. 10a 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 the baseline b (Light CNN-29 Wu et al. 2015), 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 the baseline and improves the accuracy by 1.13% over the 2nd-best. This shows that the facial rep-

Table 7 Face recognition performance (%) comparison on CFP (Sengupta et al. 2016)

Method	Frontal-Profile			Frontal-Frontal		
	Acc	EER	AUC	Acc	EER	AUC
FV+DML (Sengupta et al. 2016)	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 (Sengupta et al. 2016)	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 (Sengupta et al. 2016)	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 (Sengupta et al. 2016)	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 (Sengupta et al. 2016)	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 (Sankaranarayanan et al. 2016)	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. (2016)	91.97 ± 1.70	8.00 ± 1.68	97.70 ± 0.82	98.41 ± 0.45	1.54 ± 0.43	99.89 ± 0.06
b	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	**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

The best results are highlighted in bold
The results are averaged over 10 testing splits

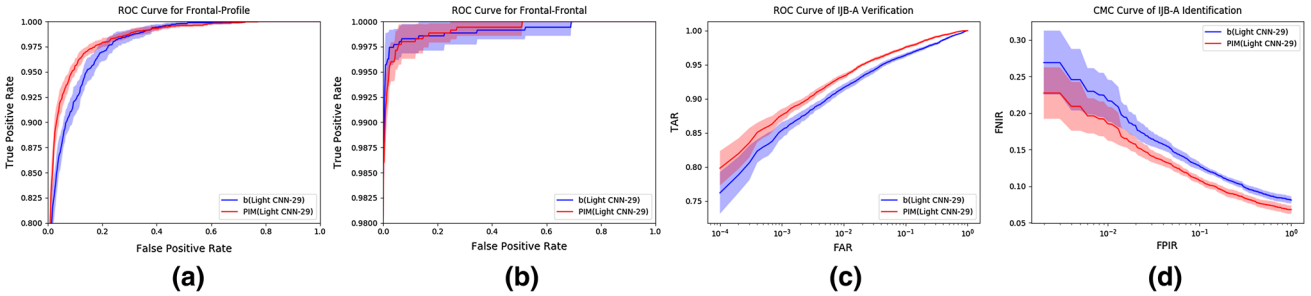


Fig. 10 Performance curve comparison between PIM and the baseline b on CFP (Sengupta et al. 2016) and IJB-A (Klare et al. 2015). **a** ROC curve for frontal-profile on CFP (Sengupta et al. 2016); **b** ROC curve for frontal-frontal on CFP (Sengupta et al. 2016); **c** ROC curve for verification on IJB-A (Klare et al. 2015); **d** CMC curve for identification on IJB-A (Klare et al. 2015). Best viewed in color

Fig. 11 Comparison of face frontalization results

resentations learned by PIM are discriminative and robust even at extreme pose variations. Visualized comparison of face frontalization results between PIM and DR-GAN (Tran et al. 2017) and TP-GAN (Huang et al. 2017) 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 (Klare et al. 2015) contains 5397 images and 2042 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

Table 8 Face recognition performance (%) comparison on IJB-A (Klare et al. 2015)

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 (Klare et al. 2015)	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 (Klare et al. 2015)	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 (Hassner et al. 2016)	0.631	0.309	–	–	–	0.846
LSFS (Wang et al. 2015)	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 (AbdAlmageed et al. 2016)	0.911	0.787	–	0.250	0.480	0.846
VGG-Face (Parkhi et al. 2015)	–	0.805 ± 0.030	–	0.33 ± 0.031	0.539 ± 0.077	0.913 ± 0.011
PAMs (Masi et al. 2016)	0.652 ± 0.037	0.826 ± 0.018	–	–	–	0.840 ± 0.012
Masi et al. (2016)	–	0.886	0.725	–	–	0.906
Triplet Embedding (Sankaranarayanan et al. 2016)	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 (Ranjan et al. 2016)	0.976 ± 0.004	0.922 ± 0.010	0.823 ± 0.020	0.113 ± 0.014	0.208 ± 0.020	$\mathbf{0.947 \pm 0.008}$
b	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	$\mathbf{0.976 \pm 0.005}$	$\mathbf{0.933 \pm 0.011}$	$\mathbf{0.875 \pm 0.018}$	$\mathbf{0.108 \pm 0.009}$	$\mathbf{0.185 \pm 0.023}$	0.944 ± 0.011

The best results are highlighted in bold

The results are averaged over 10 testing splits. "–" means the result is not reported. Standard deviation is not available for some methods

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 b (Light CNN-29 Wu et al. 2015) and other state-of-the-arts on IJB-A (Klare et al. 2015) unconstrained face verification and identification protocols are given in Table 8. The corresponding ROC and CMC curves are provided in Fig. 10c,d. With the joint learning of face frontalization and discriminative representations, our method outperforms the baseline 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 2^{nd}-best method All-In-One (Ranjan et al. 2016) 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 (Tran et al. 2017) and TP-GAN (Huang et al. 2017) 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 (Huang et al. 2007) contains 13,233 face images of 5749 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 with TP-GAN (Huang et al. 2017) and DR-GAN (Tran et al. 2017), 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 suffer severe texture loss and involved artifacts, and the predictions of DR-GAN 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 meta-learning ("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.

Fig. 12 Failure cases of face frontalization results by PIM

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

5 Conclusion

We proposed a novel **P**ose **I**nvariant **M**odel (PIM) to address the challenging face recognition with large pose variations. PIM unifies a **F**ace **F**rontalization sub-**N**et (FFN) and a **D**iscriminative **L**earning sub-**N**et (DLN) for pose-invariant recognition in an end-to-end deep architecture. The FFN adopts unsupervised cross-domain adversarial training and a meta-learning ("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.

Acknowledgements The work of Junliang Xing was partially supported by the National Science Foundation of China 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

AbdAlmageed, W., Wu, Y., Rawls, S., Harel, S., Hassner, T., Masi, I., et al. (2016). Face recognition using deep multi-pose representations. In *WACV* (pp. 1–9).

Ahonen, T., Hadid, A., & Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *28*(12), 2037–2041.

Berthelot, D., Schumm, T., & Metz, L. (2017) Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717.

Bertinetto, L., Henriques, J. F., Valmadre, J., Torr, P., & Vedaldi, A. (2016). Learning feed-forward one-shot learners. In *NeurIPS* (pp. 523–531).

Bowyer, K. W., Chang, K., & Flynn, P. (2006). A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition. *Computer Vision and Image Understanding*, *101*(1), 1–15.

Cao, J., Hu, Y., Zhang, H., He, R., & Sun, Z. (2018). Learning a high fidelity pose invariant model for high-resolution face frontalization. In *NeurIPS* (pp. 2867–2877).

Chen, D., Cao, X., Wen, F., & Sun, J. (2013). Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *CVPR* (pp. 3025–3032).

Chen, J.-C., Zheng, J., Patel, V. M., & Chellappa, R. (2016). Fisher vector encoded deep convolutional features for unconstrained face verification. In *ICIP* (pp. 2981–2985).

Chen, W., Liu, T.-Y., Lan, Y., Ma, Z.-M., & Li, H. (2009). Ranking measures and loss functions in learning to rank. In *NeurIPS* (pp. 315–323).

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR* (pp. 886–893).

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

Dave, R., Vyas, A., Mojidra, S., Desai, P., & Nikita, P. (2018). Face recognition techniques: A survey. arXiv preprint arXiv:1803.07288.

Ding, C., & Tao, D. (2017). Pose-invariant face recognition with homography-based normalization. *Pattern Recognition*, *66*, 144–152.

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

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, *17*(59), 1–35.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. in *NeurIPS*, 2672–2680.

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-pie. *Image and Vision Computing*, *28*(5), 807–813.

Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV* (pp. 87–102).

Hao, S., Wang, W., Ye, Y., Nie, T., & Bruzzone, L. (2017). Two-stream deep architecture for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *56*(4), 2349–2361.

Hassner, T., Harel, S., Paz, E., & Enbar, R. (2015). Effective face frontalization in unconstrained images. In *CVPR* (pp. 4295–4304).

Hassner, T., Masi, I., Kim, J., Choi, J., Harel, S., Natarajan, P., et al. (2016). Pooling faces: template based face recognition with pooled face images. in *CVPRW* (pp. 59–67).

Hu, Y., Wu, X., Yu, B., He, R., & Sun, Z. (2018). Pose-guided photorealistic face rotation. In *CVPR* (pp. 8398–8406).

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments (pp. 07–49). University of Massachusetts, Amherst, Tech. Rep.

Huang, R., Zhang, S., Li, T., & He, R. (2017). "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," arXiv preprint arXiv:1704.04086.

Kan, M., Shan, S., Chang, H., & Chen, X. (2014). Stacked progressive auto-encoders (spae) for face recognition across poses. in *CVPR*, 1883–1890.

Kang, B.-N., & Kim, D. (2013). Face identification using affine simulated dense local descriptors. In *URAI* (pp. 346–351).

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., et al. (2015). Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR* (pp. 1931–1939).

LeCun, Y. (1997). The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In *CVPR* (pp. 6738–6746).

Maaten, L. v d, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605.

Masi, I., Rawls, S., Medioni, G., & Natarajan, P. (2016). Pose-aware face recognition in the wild. In *CVPR* (pp. 4838–4846).

Masi, I., Tran, A. T., Leksut, J. T., Hassner, T., & Medioni, G. (2016). Do we really need to collect millions of faces for effective face recognition?. arXiv preprint arXiv:1603.07057.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Ohayon, S., Freiwald, W. A., & Tsao, D. Y. (2012). What makes a cell face selective? The importance of contrast. *Neuron*, *74*(3), 567–581.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *BMVC*.

Ranjan, R., Sankaranarayanan, S., Castillo, C. D., & Chellappa, R. (2016). An all-in-one convolutional neural network for face analysis. arXiv preprint arXiv:1611.00851.

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. arXiv preprint arXiv:1401.4082.

Sagonas, C., Panagakis, Y., Zafeiriou, S., & Pantic, M. (2015). Robust statistical face frontalization. In *ICCV* (pp. 3871–3879).

Sankaranarayanan, S., Alavi, A., Castillo, C. D., & Chellappa, R. (2016). Triplet probabilistic embedding for face verification and clustering. In *BTAS* (pp. 1–8).

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *CVPR* (pp. 815–823).

Sengupta, S., Chen, J.-C., Castillo, C., Patel, V. M., Chellappa, R., & Jacobs, D. W. (2016). Frontal to profile face verification in the wild. In *WACV* (pp. 1–9).

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*.

Simonyan, K., Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2013). Fisher vector faces in the wild. In *BMVC*.

Simonyan, K., & Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556.

Sun, Y., Liang, D., Wang, X., & Tang, X. (2015). Deepid3: Face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873.

Sun, Y., Wang, X., & Tang, X. (2015). Deeply learned face representations are sparse, selective, and robust. In *CVPR* (pp. 2892–2900).

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *CVPR* (pp. 1701–1708).

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2015). Web-scale training for face identification. In *CVPR* (pp. 2746–2754).
Tran, L., Yin, X., & Liu, X. (2017). Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*.
Wang, D., Otto, C., & Jain, A. K. (2015). Face search at scale: 80 million gallery. arXiv preprint arXiv:1507.07242.
Wang, W., Tulyakov, S., & Sebe, N. (2018). Recurrent convolutional shape regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(11), 2569–2582.
Wang, W., Yan, Y., Cui, Z., Feng, J., Yan, S., & Sebe, N. (2018). Recurrent face aging with hierarchical autoregressive memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(3), 654–668.
Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, *10*(Feb), 207–244.
Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *ECCV* (pp. 499–515).
Wu, X., He, R., Sun, Z., & Tan, T. (2015). A light cnn for deep face representation with noisy labels. arXiv preprint arXiv:1511.02683.
Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., & Kassim, A. (2016). Robust facial landmark detection via recurrent attentive-refinement networks. In *ECCV* (pp. 57–72).
Xiao, S., Liu, L., Nie, X., Feng, J., Kassim, A. A., & Yan, S. (2016). A live face swapper. In *ACM MM* (pp. 691–692).
Xiong, C., Zhao, X., Tang, D., Jayashree, K., Yan, S., & Kim, T.-K. (2015). Conditional convolutional neural network for modality-aware face recognition. in *ICCV*, 3667–3675.
Yim, J., Jung, H., Yoo, B., Choi, C., Park, D., & Kim, J. (2015). Rotating your face using multi-task deep neural network. In *CVPR* (pp. 676–684).
Yin, X., & Liu, X. (2017). Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing*, *27*, 964–975.
Zhao, J., Cheng, Y., Cheng, Y., Yang, Y., Zhao, F., Li, J., et al. (2019). Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. *AAAI*, *33*, 9251–9258.
Zhao, J., Cheng, Y., Xu, Y., Xiong, L., Li, J., Zhao, F., et al. (2018). Towards pose invariant face recognition in the wild. In *CVPR* (pp. 2207–2216).
Zhao, J., Li, J., Tu, X., Zhao, F., Xin, Y., Xing, J., Liu, H., Yan, S., Feng, J. (2019). Multi-prototype networks for unconstrained set-based face recognition. arXiv preprint arXiv:1902.04755.
Zhao, J., Xiong, L., Cheng, Y., Cheng, Y., Li, J., Zhou, L., et al. (2018). 3D-aided deep pose-invariant face recognition. In *IJCAI* (pp. 1184–1190).
Zhao, J., Xiong, L., Jayashree, P. K., Li, J., Zhao, F., Wang, Z., et al. (2017). Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *NeurIPS* (pp. 65–75).
Zhao, J., Xiong, L., Li, J., Xing, J., Yan, S., & Feng, J. (2018). "3d-aided dual-agent gans for unconstrained face recognition," *T-PAMI*.
Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys (CSUR)*, *35*(4), 399–458.
Zhu, X., Lei, Z., Yan, J., Yi, D., & Li, S. Z. (2015). High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR* (pp. 787–796).
Zhu, Z., Luo, P., Wang, X., & Tang, X. (2013). Deep learning identity-preserving face space. In *ICCV* (pp. 113–120).
Zhu, Z., Luo, P., Wang, X., & Tang, X. (2014). Multi-view perceptron: a deep model for learning face identity and view representations. in *NeurIPS*, 217–225.