3D-Aided Deep Pose-Invariant Face Recognition

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

Abstract

Learning from synthetic faces, though perhaps appealing for high data efficiency, may not bring satisfactory performance due to the distribution discrepancy of the synthetic and real face images. To mitigate this gap, we propose a **3D**-Aided Deep **P**ose-**I**nvariant Face Recognition **M**odel (3D-PIM), which automatically recovers realistic frontal faces from arbitrary poses through a 3D face model in a novel way. Specifically, 3D-PIM incorporates a simulator with the aid of a **3D M**orphable **M**odel (3D MM) to obtain shape and appearance prior for accelerating face normalization learning, requiring less training data. It further leverages a global-local **G**enerative **A**dversarial **N**etwork (GAN) with multiple critical improvements as a refiner to enhance the realism of both global structures and local details of the face simulator's output using unlabelled real data only, while preserving the identity information. Qualitative and quantitative experiments on both controlled and in-the-wild benchmarks clearly demonstrate superiority of the proposed model over state-of-the-arts.

1 Introduction

Even though (near-) frontal[1] face recognition seems to be solved under constrained conditions [Yim *et al.*, 2015; Schroff *et al.*, 2015], the more general problem of *face recognition in the wild* still needs more studies, desiderated by many practical applications. Among various challenging factors, the one that harms face recognition performance arguably the most is pose variation.

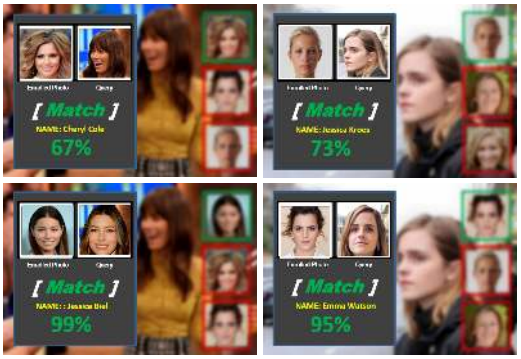

Figure 1: Face recognition with large pose variations. Top: Traditional face recognition system fails due to large pose variations. Bottom: The proposed 3D-PIM recovers natural frontal face images from arbitrary poses and recognizes the face correctly. Moreover, our 3D-PIM model potentially benefits face attribute recognition. Best viewed in color.

To address this challenge, some research attempts [Tran *et al.*, 2017; Huang *et al.*, 2017] have been made to employ synthetic frontal face images to learn pose-invariant models. However, naively learning from synthetic images can be problematic due to the distribution discrepancy between synthetic and real face images. The low-quality synthesized face images would cause the learned face recognition model to overfit to fake information only contained in synthetic images and fail to generalize well on real faces. Manually increasing the realism of the simulator is often expensive in terms of time cost and manpower, if possible.

In this work, we propose a novel and unified deep neural network, termed as **3D**-Aided Deep **P**ose-**I**nvariant Face Recognition **M**odel (3D-PIM), which automatically recovers natural frontal face images from arbitrary poses for pose-

*indicates equal contributions.

†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/. Author e-mails: {zhaojian90, jianshu}@u.nus.edu, {lin.xiong, yi.cheng, yan.xu, karlekar.jayashree, sugiri.pranata, shengmei.shen}@sg.panasonic.com, {chengyu996, zhouli2025}@gmail.com, jlxing@nlpr.ia.ac.cn, {elefjia, eleyans}@nus.edu.sg.

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

invariant face recognition. 3D-PIM effectively leverages a 3D face model and meanwhile overcomes the drawback of low realism essentially. The 3D-PIM takes faces of arbitrary poses with other potential distracting factors (*e.g.*, bad illumination or different expressions) as input. It recovers photo-realistic frontal faces with preserved discrimination across different identities, offering appealing robustness to pose variations, as illustrated in Fig. 1.

In particular, the 3D-PIM unifies a simulator for 3D face reconstruction and frontal view synthesis, and a refiner for realism refinement. The two components learn in a conjugated way. The simulator is aided by a **3D M**orphable **M**odel (3D MM) [Blanz and Vetter, 1999] to provide shape and appearance prior for accelerating face normalization learning, requiring less training data. The refiner is a global-local **G**enerative **A**dversarial **N**etwork (GAN) to improve the realism of both global structures and local details of the simulator's output using unlabeled real data, while preserving the identity information. Different from vanilla GANs, 3D-PIM introduces facial structure loss to address self-occlusion, identity perception loss to preserve identity information of the generated faces, and adversarial loss to avoid artifacts of both global facial structures and local details which are critical for face recognition. The refined synthetic frontal face images present photo-realistic quality with well preserved identity information, which facilitate pose-invariant model learning. For stabilizing the training process of 3D-PIM, we update the discriminator using a history of refined results.

We conduct extensive qualitative and quantitative experiments on various benchmarks, including both controlled and in-the-wild datasets. The results demonstrate the effectiveness of 3D-PIM on recognizing faces with extreme poses and also its superiority over the state-of-the-arts consistently on all the benchmarks.

2 Related Work

Traditional face frontalization methods rely on 2D/3D local texture warping [Zhu *et al.*, 2015], statistical modeling [Sagonas *et al.*, 2015], and deep learning based methods [Huang *et al.*, 2017; Tran *et al.*, 2017; Kan *et al.*, 2014; Yim *et al.*, 2015]. For instance, Kan *et al.* [Kan *et al.*, 2014] use **S**tacked **P**rogressive **A**uto-**E**ncoders (SPAE) to rotate a profile face to frontal. Despite encouraging results, the synthesized faces lack fine details and tend to be blurry and unreal under a large pose. The quality of synthesized images with current methods is still far from satisfactory for recognizing faces with large pose variation. Deep learning methods often handle pose variance through a single pose-agnostic or several pose-specific models with pooling operation and specific loss functions [Cheng *et al.*, 2017; Zhao *et al.*, 2017a; Li *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] model uses a deep CNN coupled with 3D alignment. FaceNet [Schroff *et al.*, 2015] utilizes the inception architecture. The DeepID2+ [Sun *et al.*, 2015b] and DeepID3 [Sun *et al.*, 2015a] extend the FaceNet [Schroff *et al.*, 2015] model by including joint Bayesian metric learning and multi-task learning. However, such data-driven methods heavily rely on well annotated data. Collecting labeled data covering all variations is expensive and even impractical.

Our proposed 3D-PIM is based on a similar idea with TP-GAN [Huang *et al.*, 2017] and DR-GAN [Tran *et al.*, 2017] that synthesize faces based on GAN framework, SimGAN [Shrivastava *et al.*, 2016] that learns from simulated and unsupervised images, and DA-GAN [Zhao *et al.*, 2017b] that considers incorporating 3D MM [Blanz and Vetter, 1999] as a prior during face syethesis. Our method differs from them in following aspects: 1) 3D-PIM aims to recover photo-realistic and identity-preserving frontal faces to address the large pose variance issue in unconstrained face recognition, whereas DA-GAN tries to synthesize profile faces for balancing pose distribution and SimGAN is designed for much simpler scenarios (*e.g.*, eye and hand image refinement); 2) TP-GAN suffers from severe over-fitting risk and DR-GAN suffers from identity information loss under large poses, which limit their effectiveness in face recognition.

3 3D-Aided Deep Pose-Invariant Model

3.1 Simulator

Large pose variation is the main challenge to unconstrained face recognition, and also the key obstacle for learning a well-performing pose-invariant model. To address this problem, we propose to impose a prior on the generation process, with the aid of a 3D MM [Blanz and Vetter, 1999]. This reduces the training complexity and leads to better empirical performance with limited data.

The 3D MM is one of the most successful methods that describe the 3D face space. Constructed by linear combinations of face scans in the PCA space, 3D MM can approximate an arbitrary face shape with considerable accuracy.

$$S = \bar{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp}, \tag{1}$$

where S is the 3D face, $\bar{S}$ is the mean shape, A_{id} denotes the principle axes trained on the 3D face scans with neutral expressions, α_{id} denotes the identity coefficient vector, A_{exp} denotes the principle axes trained on the offset between expression scans and neutral scans, and α_{exp} is the expression coefficient vector.

To fit 3D MM to a face image, we project the face model onto the image plane with the Weak Perspective Projection:

$$s = fPR(\alpha, \beta, \gamma)(S + t), \tag{2}$$

where s is the 2D positions of 3D points on the image plane, f is the scale factor, P is the the orthographic projection matrix $\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}$, $R(\alpha, \beta, \gamma)$ is the 3×3 rotation matrix constructed with α–pitch, β–yaw, γ–roll, and t is the translation vector.

The fitting process is to search the ground truth 2D coordinates s_{GT} of 3D points and estimate the 3D MM coefficients $\{\alpha_{id}, \alpha_{exp}, f, R(\alpha, \beta, \gamma), t\}$ by minimizing the Euclidean distance between s and s_{GT}:

$$\mathcal{L}_{\mathrm{sim}} = \frac{1}{2n}\sum_{i=1}^{n}\|s_i - s_{\mathrm{GT}i}\|_2^2, \tag{3}$$

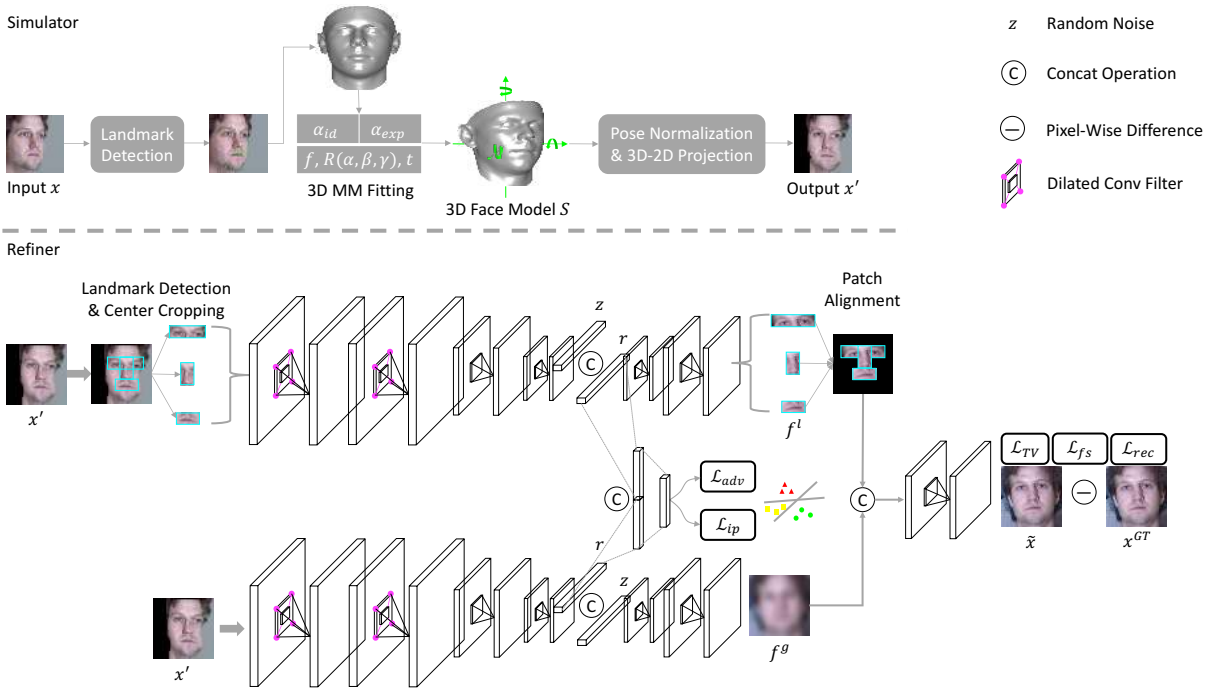

Figure 2: **3D**-Aided Deep **P**ose **I**nvariant **M**odel (3D-PIM) for pose-invariant face recognition. The 3D-PIM contains a simulator and a refiner, which learn in a conjugated way. The simulator is aided by a 3D MM. It localizes landmark points, estimates 3D MM coefficients, and produces synthesis faces with normalized poses, which are fed to the refiner for realism refinement. The refiner is a global-local GAN to improve the realism of both global structures and local details of the simulator's output using unlabelled real data, while preserving the identity information. 3D-PIM introduces facial structure loss ($\mathcal{L}_{\rm fs}$) to address self-occlusion, identity perception loss ($\mathcal{L}_{\rm ip}$) to preserve identity information of the generated faces, and adversarial loss ($\mathcal{L}_{\rm adv}$) to avoid artifacts. Best viewed in color.

where n is the number of the 2D facial landmarks.

Finally, we simulate frontal face images x' in canonical view from the input x under arbitrary poses, as shown in the upper panel of Fig. 2. However, the performance of the simulator decreases dramatically under large poses due to artifacts and severe texture loss caused by self-occlusion, causing the network to overfit to fake information only contained in synthetic images and fail to generalize well on real data.

3.2 Refiner

In order to generate photo-realistic and identity-preserved frontal face images which are truly beneficial for unconstrained face recognition, we further refine the above-mentioned simulated frontal face images with the proposed refiner network, which consists of a **G**lobal-**L**ocal **G**enerator (GLG) and a **G**lobal-**L**ocal **D**iscriminator (GLD), learning in a competitive way.

Global-Local Generator

Since 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 refining global face structures and local details. To address this issue, we propose a **G**lobal-**L**ocal **G**enerator (GLG), as inspired by [Huang *et al.*, 2017; Zhu *et al.*, 2015], where one path aims to refine the global sketch and the other to attend to refine local facial details, as shown in the bottom panel of Fig. 2.

In particular, the global-path generator G_{θ^g} (with learnable parameters θ^g) consists of a transition-down encoder and a transition-up decoder. The local path generator G_{θ^l} also has an auto-encoder architecture, containing three identical sub-networks that learn separately to refine the following three center-cropped local patches: eyes, nose, and mouth. These patches are acquired by an off-the-shelf landmark detection model. Given a simulated frontal face image x', 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 refined image $\tilde{x}$. We also concatenate a Gaussian random noise z at the bottleneck layer of the GLG to model variations of other factors besides pose, which may also help recover invisible details. Instead of using only the standard convolutional layers, we also employ a variant called dilated convolution in GLG. Dilated convolution uses kernels that are spread out, leading to a much larger input area for computing each output pixel with the same number of parameters and computational power. By using dilated convolutions, the model can effectively "see" a larger area of the input image when computing each output pixel than with standard convolutional layers. This is important for our refinement task, as the context information is critical for producing plausible hypothesis for the missing regions.

Formally, let the input simulated frontal face image with three landmark patches be collectively denoted as x'. Then the refined face is $\tilde{x} = G_\theta(x')$. The key requirement for the

GLG is that the refined image $\tilde{x}$ should visually resemble a real one and preserve the identity information as well as local textures.

To this end, we propose to learn the parameters $\{\theta^g, \theta_i^l\}$ (here i=1, ..., 3 index the three local-path models) by minimizing the following composite losses:

$$\mathcal{L}_{G_\theta} = -\mathcal{L}_{\text{adv}} + \lambda_0\mathcal{L}_{\text{ip}} + \lambda_1\mathcal{L}_{\text{rec}} + \lambda_2\mathcal{L}_{\text{fs}} + \lambda_3\mathcal{L}_{\text{TV}}, \quad (4)$$

where $\mathcal{L}_{\text{adv}}$ is the **adv**ersarial loss for adding realism to the synthetic images and alleviating artifacts, $\mathcal{L}_{\text{ip}}$ is the **i**dentity **p**erception loss for preserving the identity information, $\mathcal{L}_{\text{rec}}$ is the **rec**onstruction loss for encouraging multi-scale image content consistency, $\mathcal{L}_{\text{fs}}$ is the **f**acial **s**tructure loss for alleviating self-occlusion issue, $\mathcal{L}_{\text{TV}}$ is the **t**otal **v**ariation loss for reducing spiky artifacts, and $\{\lambda_k\}_{k=0}^{3}$ are weighting parameters among different losses.

$\mathcal{L}_{\text{rec}}$ is introduced to enforce the multi-scale content consistency between the final frontalized face and corresponding ground truths, defined as $\mathcal{L}_{\text{rec}} = \|\tilde{x} - \tilde{x}_{\text{GT}}\|/|\tilde{x}_{\text{GT}}|$, where $|\tilde{x}_{\text{GT}}|$ is the size of $\tilde{x}_{\text{GT}}$.

Since symmetry is an inherent feature of human facial structures, $\mathcal{L}_{\text{fs}}$ is introduced within the Laplacian space to exploit this prior information and impose the symmetry constraint on the recovered frontal view for alleviating self-occlusion issue:

$$\mathcal{L}_{\text{fs}} = \frac{1}{W/2 \times H}\sum_{i}^{W/2}\sum_{j}^{H}|\tilde{x}_{i,j} - \tilde{x}_{W-(i-1),j}|, \quad (5)$$

where W, H denote the width and height of the final recovered frontal face image $\tilde{x}$, respectively.

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

$$\mathcal{L}_{\text{TV}} = \sum_{i}^{W}\sum_{j}^{H}\sqrt{(\tilde{x}_{i,j+1} - \tilde{x}_{i,j})^2 + (\tilde{x}_{i+1,j} - \tilde{x}_{i,j})^2}. \quad (6)$$

Global-Local Discriminator

To add realism to the synthetic images to really benefit face recognition performance, we need to narrow the gap between the distributions of synthetic and real images. An ideal generator will make it impossible to classify a given image as real or refined with high confidence. Meanwhile, preserving the identity information is essential and critical for recognition. An ideal generator will generate the refined face images that have small intra-class distance and large inter-class distance in the feature space spanned by the deep neural networks for unconstrained face recognition. These motivate us to employ an adversarial **G**lobal-**L**ocal **D**iscriminator (GLD) which distinguishes real *v.s.* fake and identity information of global facial structures and local details simultaneously.

To facilitate this process, we simply leverage the same architecture respectively in the global-path encoder and the local-path encoder as the global-path discriminator D_{ϕ^g} (with learnable parameters ϕ^g) and the local-path discriminator D_{ϕ^l} (with learnable parameters ϕ^l), which learn separately without weight sharing, to avoid typical GAN tricks, as shown in the bottom panel of Fig. 2. The feature maps from D_{ϕ^g} and D_{ϕ^l} are further concatenated and fed into a fully connected layer to compute $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{ip}}$, which serve 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 while preserving the identity information. In particular, $\mathcal{L}_{\text{adv}}$ is defined as

$$\mathcal{L}_{\text{adv}} = \frac{1}{N}\sum_{i=1}^{N} -y_i\log[D_\phi(\tilde{x}_i)] - (1-y_i)\log[1 - D_\phi(x_i^{Real})], \quad (7)$$

where N is the number of training samples, x^{Real} denotes the real frontal face image, and y is the binary label indicating the image is synthesized or real.

An underlying problem of adversarial training is that the discriminator only focuses on the latest refined results. This lack of memory may cause 1) divergence of the adversarial training, and 2) re-introduced artifacts by the refiner which the discriminator has forgotten about. Any refined image generated by the refiner at any time during the entire training procedure is a "fake" image for the discriminator. Hence, the discriminator should be able to classify all these images as fake. Based on this observation, we further extend the GLD with an external buffer to improve the stability of adversarial training by updating the GLD using a history of refined results, rather than only with those in the current minibatch, as first introduced in [Shrivastava *et al.*, 2016].

In order to preserve the identity discriminability of refined face images, we define $\mathcal{L}_{\text{ip}}$ with the multi-class cross-entropy loss based on the output from the bottleneck layer of D_ϕ.

$$\begin{aligned}\mathcal{L}_{\text{ip}} = &\frac{1}{N}\sum_{j=1}^{N} -Y_j log[D_\phi(x_j^{Real})] - (1-Y_j)log[1 - D_\phi(x_j^{Real})] \\ &- \frac{1}{N}\sum_{i=1}^{N} Y_i log[D_\phi(\tilde{x}_i)] - (1-Y_i)log[1 - D_\phi(\tilde{x}_i)],\end{aligned} \quad (8)$$

where Y is the identity ground truth.

Thus, minimizing $\mathcal{L}_{\text{ip}}$ would encourage deep features of the refined face images with the same identity to be close to each other. If one visualizes the learned deep features in the high-dimensional space, these learned deep features form several compact clusters, and each cluster may be far away from others. Each cluster has a small variance. In this way, the refined face images are enforced with well preserved identity information.

Using $\mathcal{L}_{\text{ip}}$ alone makes the results prone to annoying artifacts, because the search for a local minimum of $\mathcal{L}_{\text{ip}}$ may go through a path that resides outside the manifold of natural face images. Thus, we combine $\mathcal{L}_{\text{ip}}$ with $\mathcal{L}_{\text{adv}}$ as the final objective function for D_ϕ to ensure that the search resides in that manifold and produces photo-realistic and identity-preserved face image:

$$\mathcal{L}_{\text{D}_\phi} = \mathcal{L}_{\text{adv}} + \lambda_4\mathcal{L}_{\text{ip}}, \quad (9)$$

where λ_4 is a weighting parameter between $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{ip}}$.

4 Experiments

We evaluate 3D-PIM qualitatively and quantitatively under both controlled and in-the-wild settings for pose-invariant

Table 1: Component analysis under Multi-PIE Setting-1.

Method	±90°	±75°	±60°	±45°	±30°	±15°
b1	18.80	63.80	92.20	98.30	99.20	99.40
b2	33.00	76.10	95.20	97.90	99.20	99.80
w/o refiner	28.81	46.52	61.26	78.41	89.87	98.74
w/o $\mathcal{L}_{\rm rec}$	65.86	85.73	96.29	97.06	98.04	98.83
w/o $\mathcal{L}_{\rm ip}$	66.70	86.53	97.10	98.44	98.93	99.15
w/o $\mathcal{L}_{\rm adv}$	67.13	87.97	97.51	98.75	99.14	99.40
w/o $G_{\theta_i^l}$	67.35	88.20	97.74	98.12	98.59	99.10
w/o $\mathcal{L}_{\rm fs}$	68.61	89.49	98.03	98.85	99.10	99.32
w/o DC	69.76	90.65	98.19	98.76	99.02	99.18
w/o $D_{\phi_i^l}$	70.77	91.67	98.21	99.03	99.17	99.37
w/o buffer	71.95	92.86	98.40	99.15	99.31	99.56
3D-PIM1	73.17	94.03	98.57	99.21	99.52	99.67
3D-PIM2	76.12	94.34	98.84	99.34	99.47	99.83

face recognition. For qualitative evaluation, we show visualized results on Multi-PIE [Gross *et al.*, 2010] benchmark dataset. For quantitative evaluation, we evaluate face recognition performance on Multi-PIE and IJB-A datasets.

Implementation Details Throughout the experiments, the size of the RGB images of the input profile face (x), the simulator synthesis (x'), and the GLG prediction ($\tilde{x}$) is fixed as 128×128; the sizes of the three RGB local patches (*i.e.*, eyes, nose and mouth) are fixed as 80×40, 32×40, and 48×32, respectively; the dimensionality of the Gaussian random noise z is fixed as 100; the constraint factors λ_0 to λ_4 are empirically fixed as 0.05, 1.0, 0.1, 5×10^{-4}, and 0.1, respectively; the batch size and learning rate are fixed as 16 and 3×10^{-5}, respectively; we use the RAR [Xiao *et al.*, 2016] framework for landmark detection; we merge two popular face models with Non-Rigid ICP [Amberg *et al.*, 2007] to construct our 3D MM; the identity basis (A_{id}) comes from the BFM [Paysan *et al.*, 2009] and the expression basis (A_{exp}) comes from the Face Warehouse [Cao *et al.*, 2014]; the proposed network is implemented based on the publicly available TensorFlow [Abadi *et al.*,] platform, which is trained using Adam (β_1=0.5) on a single NVIDIA TITAN X GPU with 12G memory.

4.1 Evaluations on Multi-PIE Benchmark

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

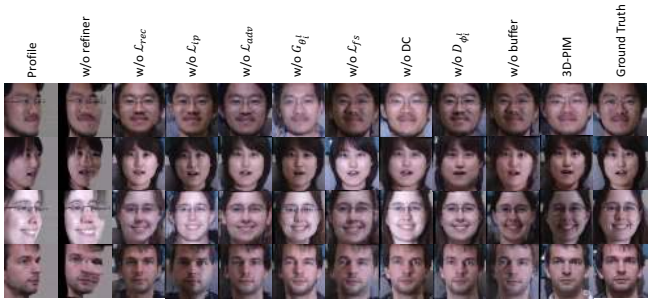

Figure 3: Synthesized results of 3D-PIM and its variants.

Component Analysis

We first investigate different architectures and loss function combinations of 3D-PIM to see their respective roles in pose-invariant face recognition. We compare 11 variants of 3D-PIM, *i.e.*, different face recognition network architectures (3D-PIM1: ResNet-50 [He *et al.*, 2016] vs. 3D-PIM2: Light CNN-29 [Wu *et al.*, 2015]), w/o refiner (recognition with simulator output), w/o local-path generator $G_{\theta_i^l}$, w/o **d**ilated **c**onvolution (DC), w/o local-path discriminator $D_{\phi_i^l}$, w/o buffer, w/o $\mathcal{L}_{\rm rec}$, w/o $\mathcal{L}_{\rm adv}$, w/o $\mathcal{L}_{\rm fs}$, and w/o $\mathcal{L}_{\rm ip}$, in each case.

Averaged Rank1 recognition rates are compared in Setting-1 in Tab. 1. The results on the profile images serve as our baselines (*i.e.*, b1 and b2). The results of the middle panel variations are all based on Light CNN-29. By comparing the results from the top and bottom panels, we observe that our 3D-PIM is not restricted to the face recognition backbone architecture used, since similar improvements (*e.g.* 54.37% *v.s.* 43.12% under ±90°) can be achieved with our joint 3D-aided frontal face reconstruction and global-local realism refinement framework. The refiner, reconstruction loss, and identity perception loss contribute the most to improving the face recognition performance, especially for large pose cases. The adversarial training, global-local refiner, facial structure loss, and DC collaboratively add realism to the simulator, for both global facial structures and local details, which are beneficial for improving the recognition performance. Although not apparent, the buffer also helps improve the recognition performance. It is especially useful for stabilizing the training process.

Fig. 3 illustrates the perceptual performance of these variants. As expected, the inference results without refiner, $\mathcal{L}_{\rm rec}$, and $\mathcal{L}_{\rm ip}$ deviate from the true appearance severely. The synthesis without adversarial training and buffer tends to present unnatural artifacts while that without facial structure loss sometimes shows factitious asymmetrical effect. The synthesis without the local-path generator and discriminator tends to lose local details while that without DC tends to lose high-frequency information.

Qualitative Results

Most previous works on face frontalization 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 generalizable facial representations. However, the proposed 3D-PIM is able to recover high-fidelity and identity-preserved frontal

Table 2: Rank1 recognition rates (%) under Multi-PIE Setting-1/Setting-2.

Method	$\pm 90^\circ$	$\pm 75^\circ$	$\pm 60^\circ$	$\pm 45^\circ$	$\pm 30^\circ$	$\pm 15^\circ$
CPF [Yim *et al.*, 2015]	- / -	- / -	- / 61.90	71.65 / 79.90	81.05 / 88.50	89.45/95.00
DR-GAN [Tran *et al.*, 2017]	- / -	- / -	- / 83.20	- / 86.20	- / 90.10	- / 94.00
TP-GAN [Huang *et al.*, 2017]	64.03 / 64.64	84.10 / 77.43	92.93 / 87.72	98.58 / 95.38	**99.85** / 98.06	99.78 / 98.68
3D-PIM	**76.12 / 86.73**	**94.34 / 95.21**	**98.84 / 98.37**	**99.34 / 98.81**	99.47 / **99.48**	**99.83 / 99.64**

Table 3: Face recognition performance comparison on IJB-A. The results are averaged over 10 testing splits.

Method	**Verification**			**Identification**	
	TAR@FAR=0.01	TAR@FAR=0.001	TAR@FAR=0.0001	FNIR@FPIR=0.01	Rank1
OpenBR [Klare *et al.*, 2015]	0.236±0.009	0.104±0.014	-	0.934±0.017	0.246±0.011
GOTS [Klare *et al.*, 2015]	0.406±0.014	0.198±0.008	-	0.953±0.024	0.433±0.021
Pooling faces [Hassner *et al.*, 2016]	0.309	-	-	-	0.846
Triplet Similarity [Sankaranarayanan *et al.*, 2016]	0.790±0.030	0.590±0.050	-	0.444±0.065	0.880±0.015
VGG-Face [Parkhi *et al.*, 2015]	0.805±0.030	-	-	0.539±0.077	0.913±0.011
PAMs [Masi *et al.*, 2016]	0.826±0.018	0.652±0.037	-	-	0.840±0.012
DR-GAN [Tran *et al.*, 2017]	0.831±0.017	0.699±0.029	-	-	0.901±0.014
Triplet Embedding [Sankaranarayanan *et al.*, 2016]	0.900±0.010	0.813±0.002	-	0.247±0.030	0.932±0.010
Template Adaptation [Crosswhite *et al.*, 2017]	0.939±0.013	0.836±0.027	-	0.226±0.049	0.928±0.001
NAN [Yang *et al.*, 2016]	0.941±0.008	0.881±0.011	-	0.183±0.041	0.958±0.005
DA-GAN [Zhao *et al.*, 2017b]	0.976±0.007	0.930±0.005	-	0.110±0.039	0.971±0.007
ℓ_2-softmax [Ranjan *et al.*, 2017]	0.970±0.004	0.943±0.005	0.909±0.007	0.085±0.041	0.973±0.005
3D-PIM	**0.989±0.002**	**0.977±0.004**	**0.953±0.012**	**0.064±0.045**	**0.990±0.002**

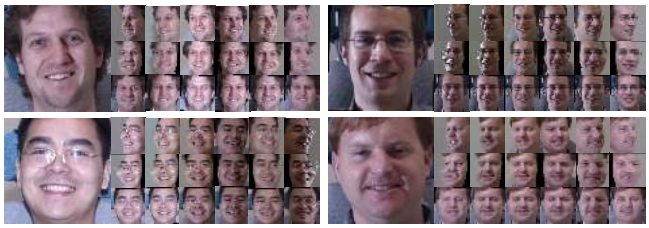

Figure 4: Comparison of face frontalization results. Each block shows a distinct identity under different poses along with other unconstrained factors (*e.g.*, expression, illumination) (*row.* 1), the raw simulated faces (*row.* 2), and the refined results by 3D-PIM (*row.* 3); for each identity, an enlarged ground truth frontal face image is provided for reference (*col.* 1).

faces under very large poses. The recovered face images in the frontal view are visualized in Fig. 4 (a). We observe that the 3D-PIM consistently recovers faces well for all cases, and the recovered frontal face images are intuitively beneficial for pose-invariant face recognition.

Quantitative Results

Tab. 2 shows the face recognition performance comparison of our 3D-PIM with other state-of-the-arts in Setting-1 and Setting-2. 3D-PIM consistently achieves the best performance across all poses, especially for large yaw angles. In particular, for Setting-1, 3D-PIM outperforms TP-GAN and c-CNN Forest by 12.09% and 28.86% under $\pm 90^\circ$, respectively. Note that TP-GAN has the same face recognition backbone as ours, and c-CNN Forest is an ensemble of three models which is much more complex than 3D-PIM. For Setting-2, 3D-PIM outperforms TP-GAN by 22.09% under $\pm 90^\circ$, and outperforms TP-GAN and DR-GAN by 10.65% and 15.17% under $\pm 60^\circ$, respectively.

4.2 Evaluations on IJB-A Benchmark

IJB-A contains $5{,}397$ images and $2{,}042$ videos from 500 subjects, captured from in-the-wild environment to avoid near frontal bias. For training and testing, 10 random splits are provided by each protocol, respectively. It contains two tasks, face verification and face identification. The performance is evaluated by TAR@FAR, FNIR@FPIR, and Rank metrics, respectively.

The performance comparison of 3D-PIM with other state-of-the-arts on IJB-A unconstrained face verification and identification protocols are given in Tab. 3. With the injection of photo-realistic and identity-preserved frontal faces generated by 3D-PIM without extra human annotation efforts, our "recognition via generation" method outperforms the 2^{nd}-best by 4.40% for TAR@FAR=0.0001 of verification, 1.70% for Rank1 of identification close set and 2.10% for FNIR@FPIR=0.01 of identification open set. This well shows the promising potential of recovered frontal view face images by 3D-PIM on large-scale and challenging unconstrained face recognition.

5 Conclusion

We proposed a novel **3D**-Aided Deep **P**ose-**I**nvariant Face Recognition **M**odel (3D-PIM) for photo-realistic and identity-preserved frontal face synthesis and a "recognition via generation" framework to address the challenging face recognition with large pose variations. 3D-PIM unifies a simulator for 3D face reconstruction and frontal view synthesis, and a refiner for realism refinement, which learn in a conjugated way. The simulator is aided by a 3D MM to provide shape and appearance prior for accelerating face normalization learning, requiring less training data, and the refiner is a global-local GAN to improve the realism of both global facial structures and local details of the simulator's output using unlabeled real data while preserving identity information. Comprehensive experiments demonstrate the superiority of 3D-PIM over state-of-the-arts. We plan to apply 3D-PIM to other transfer learning applications in future.

Acknowledgements

The work of Jian Zhao was partially supported by **C**hina **S**cholarship **C**ouncil (CSC) grant 201503170248. The work of Junliang Xing was partially supported by the National Science Foundation of Chian 61672519. The work of Jiashi Feng was partially supported by NUS startup R-263-000-C08-133, MOE Tier-I R-263-000-C21-112, NUS IDS R-263-000-C67-646 and ECRA R-263-000-C87-133.

References

[Abadi *et al.*,] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning.

[Amberg *et al.*, 2007] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *CVPR*, pages 1–8, 2007.

[Blanz and Vetter, 1999] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, pages 187–194, 1999.

[Cao *et al.*, 2014] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *TVCG*, 20(3):413–425, 2014.

[Cheng *et al.*, 2017] Yu Cheng, Jian Zhao, Zhecan Wang, Yan Xu, Karlekar Jayashree, Shengmei Shen, and Jiashi Feng. Know you at one glance: A compact vector representation for low-shot learning. In *ICCVW*, pages 1924–1932, 2017.

[Crosswhite *et al.*, 2017] Nate Crosswhite, Jeffrey Byrne, Chris Stauffer, Omkar Parkhi, Qiong Cao, and Andrew Zisserman. Template adaptation for face verification and identification. In *FG*, pages 1–8, 2017.

[Gross *et al.*, 2010] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *JIVC*, 28(5):807–813, 2010.

[Hassner *et al.*, 2016] Tal Hassner, Iacopo Masi, Jungyeon Kim, Jongmoo Choi, Shai Harel, Prem Natarajan, and Gerard Medioni. Pooling faces: template based face recognition with pooled face images. In *CVPRW*, pages 59–67, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Huang *et al.*, 2017] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *arXiv preprint arXiv:1704.04086*, 2017.

[Kan *et al.*, 2014] Meina Kan, Shiguang Shan, Hong Chang, and Xilin Chen. Stacked progressive auto-encoders (spae) for face recognition across poses. In *CVPR*, pages 1883–1890, 2014.

[Klare *et al.*, 2015] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR*, pages 1931–1939, 2015.

[Li *et al.*, 2016] Jianshu Li, Jian Zhao, Fang Zhao, Hao Liu, Jing Li, Shengmei Shen, Jiashi Feng, and Terence Sim. Robust face recognition with deep multi-view representation learning. In *ACM MM*, pages 1068–1072, 2016.

[Masi *et al.*, 2016] Iacopo Masi, Stephen Rawls, Gérard Medioni, and Prem Natarajan. Pose-aware face recognition in the wild. In *CVPR*, pages 4838–4846, 2016.

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

[Paysan *et al.*, 2009] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *AVSS*, pages 296–301, 2009.

[Ranjan *et al.*, 2017] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.

[Sagonas *et al.*, 2015] Christos Sagonas, Yannis Panagakis, Stefanos Zafeiriou, and Maja Pantic. Robust statistical face frontalization. In *ICCV*, pages 3871–3879, 2015.

[Sankaranarayanan *et al.*, 2016] Swami Sankaranarayanan, Azadeh Alavi, Carlos D Castillo, and Rama Chellappa. Triplet probabilistic embedding for face verification and clustering. In *BTAS*, pages 1–8, 2016.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[Shrivastava *et al.*, 2016] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training. *arXiv preprint arXiv:1612.07828*, 2016.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Sun *et al.*, 2015a] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.

[Sun *et al.*, 2015b] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *CVPR*, pages 2892–2900, 2015.

[Taigman *et al.*, 2014] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014.

[Tran *et al.*, 2017] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017.

[Wu *et al.*, 2015] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *arXiv preprint arXiv:1511.02683*, 2015.

[Xiao *et al.*, 2016] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *ECCV*, pages 57–72, 2016.

[Yang *et al.*, 2016] Jiaolong Yang, Peiran Ren, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. *arXiv preprint arXiv:1603.05474*, 2016.

[Yim *et al.*, 2015] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. Rotating your face using multi-task deep neural network. In *CVPR*, pages 676–684, 2015.

[Zhao *et al.*, 2017a] Jian Zhao, Jianshu Li, Fang Zhao, Shuicheng Yan, and Jiashi Feng. Marginalized cnn: Learning deep invariant representations. In *BMVC*, 2017.

[Zhao *et al.*, 2017b] Jian Zhao, Lin Xiong, Panasonic Karlekar Jayashree, Jianshu Li, Fang Zhao, Zhecan Wang, Panasonic Sugiri Pranata, Panasonic Shengmei Shen, Shuicheng Yan, and Jiashi Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *NIPS*, pages 65–75, 2017.

[Zhu *et al.*, 2015] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, pages 787–796, 2015.