Towards Pose Invariant Face Recognition in the Wild

Jian Zhao[1,2], Yu Cheng[3], Yan Xu[4], Lin Xiong[4], Jianshu Li[1], Fang Zhao[1], Karlekar Jayashree[4], Sugiri Pranata[4], Shengmei Shen[4], Junliang Xing[5], Shuicheng Yan[1,6], and Jiashi Feng[1]
[1]National University of Singapore, [2]National University of Defense Technology, [3]Nanyang Technological University
[4]Panasonic R&D Center Singapore, [5]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, [6]Qihoo 360 AI Institute
{zhaojian90, jianshu}@u.nus.edu, chengyu996@gmail.com, {yan.xu, lin.xiong, karlekar.jayashree, sugiri.pranata, shengmei.shen}@sg.panasonic.com
jlxing@nlpr.ia.ac.cn, {eleyans, elefjia}@nus.edu.sg

Abstract

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

Motivation

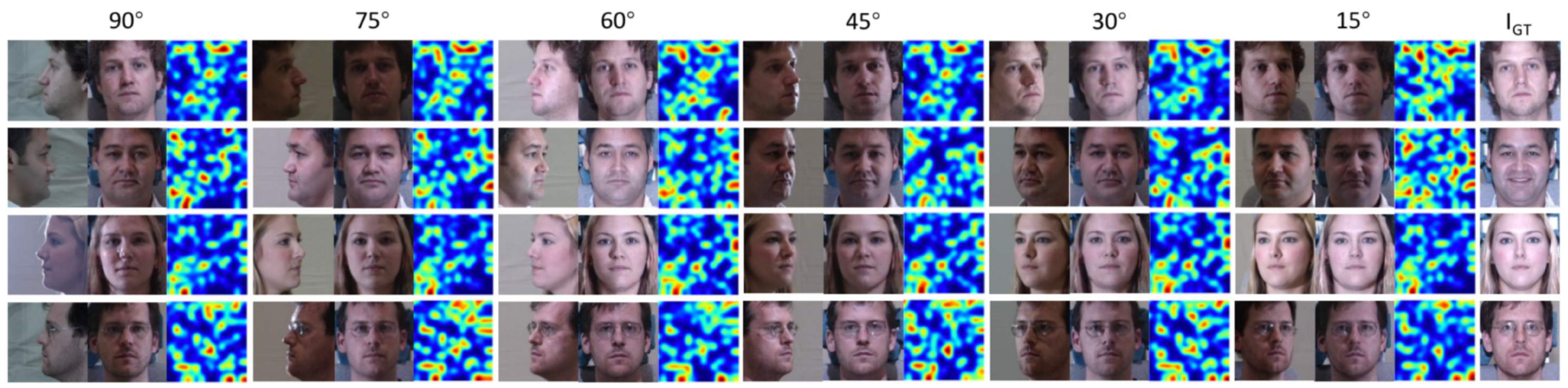


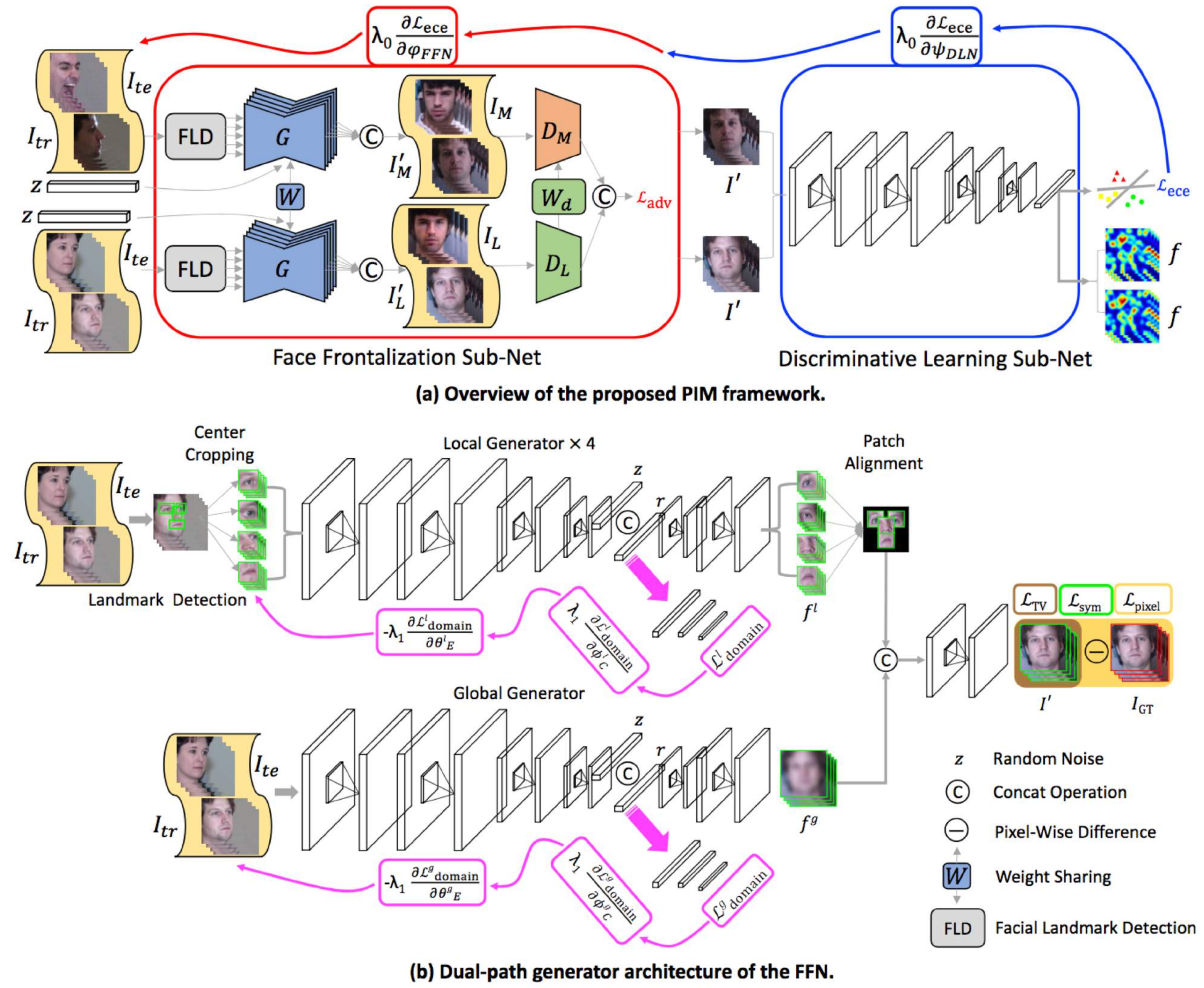
Fig. 1. Pose invariant face recognition in the wild. Each row shows a distinct identity under different poses along with other unconstrained factors (like expression, illumination, etc.), recovered frontal faces and learned facial representations (smoothed for better visualization, with blue indicting zero values) with our proposed PIM. The representations are extracted from the penultimate layer 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.

Contributions

1. We propose a **P**ose **I**nvariant **M**odel (PIM) for face recognition in the wild. PIM is a novel and unified deep neural network containing a **F**ace **F**rontalization sub-**N**et (FFN) and a **D**iscriminative **L**earning sub-**N**et (DLN) that jointly learn in an end-to-end way to allow them to mutually boost each other.
2. FFN is a carefully designed dual-path (i.e., simutaneously perceiving global structures and local details) **G**enerative **A**dversarial **N**etwork (GAN) incorporating unsupervised cross-domain adversarial training and a "learning to learn" strategy using siamese discriminator with dynamic convolution for high-fidelity and identity-preserving frontal view synthesis.
3. DLN is a generic **C**onvolutional **N**eural **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.
4. We develop effective and novel training strategies for FFN, DLN and the whole deep architecture, which generate powerful face representations.
5. 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 and LFW benchmark datasets.

Method



(a) Overview of the proposed PIM framework.

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

Fig. 2. **P**ose **I**nvariant **M**odel (PIM) for face recognition in the wild. The PIM contains an **F**ace **F**rontalization sub-**N**et (FFN) and a **D**iscriminative **L**earning sub-**N**et (DLN) that jointly learn end-to-end. FFN is a dual-path (i.e., simultaneously perceiving global structures and local details) GAN augmented by (1) unsupervised cross-domain (i.e., I_{tr} and I_{te}) adversarial training and (2) a siamese discriminator with a "learning to learn" strategy — convolutional parameters (i.e., W_d) dynamically predicted by the "learner" D_L of the discriminator and transferred to D_M. DLN is a generic **C**onvolutional **N**eural **N**etwork (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 Towards Pose Invariant Face Recognition in the Wild.

I. Face Frontalization Sub-Net

Domain Invariant Dual-Path Generator:

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

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

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

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

II. Face Frontalization Sub-Net

Dynamic Convolutional Discriminator:

$$x_{out} = W * x_{in}$$

$$x_{out} = U' * (W_d) *_{c_{in}} U * x_{in}$$

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

III. Discriminative Learning Sub-Net:

$$p_i = \frac{\exp[\tau_t\cdot(a_i^{\mathrm{T}}f)]}{\sum_j\exp[\tau_t\cdot(a_j^{\mathrm{T}}f)]}$$

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

Results

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

Tab. 1. Rank-1 recognition rates (%) across views, minor expressions and illuminations under Multi-PIE Setting-1.

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

Tab. 2. Rank-1 recognition rates (%) across views, illuminations and sessions under Multi-PIE Setting-2.

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

Tab. 3. Rank-1 recognition rates (%) across views, illuminations and sessions under Multi-PIE Setting-2.

Acknowledgement

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

Contact: Jian Zhao, Ph.D. candidate, Email: zhaojian90@u.nus.edu, Phone: (65) 9610 7176, Homepage: https://zhaoj9014.github.io/ | Lin Xiong, Research Engineer, Email: lin.xiong@sg.panasonic.com, Phone: (65) 83752875
Address: Vision and Machine Learning Lab, E4-#08-24, 4 Engineering Drive 3, National University of Singapore, Singapore 117583 | Panasonic R&D Center Singapore, Core Technology Group, 202 Bedok South Avenue 1 #02-11, Singapore 469332