Adversarial Pose Regression Network for Pose-Invariant Face Recognitions

Pengyu Li[1], **Biao Wang**[1], **Lei Zhang**[1,2]
[1]Artificial Intelligence Center, DAMO Academy, Alibaba Group
[2]Department of Computing, The Hong Kong Polytechnic University
lipengyu007@gmail.com, wangbiao225@foxmail.com, cslzhang@comp.polyu.edu.hk

Abstract

Face recognition has achieved significant progress in recent years. However, the large pose variation between face images remains a challenge in face recognition. We observe that the pose variation in the hidden feature maps is one of the most critical factors to hinder the representations from being pose-invariant. Based on the observation, we propose an Adversarial Pose Regression Network (APRN) to extract pose-invariant identity representations by disentangling their pose variation in hidden feature maps. To model the pose discriminator in APRN as a regression task in its 3D space, we also propose an Adversarial Regression Loss Function and extend the adversarial learning from classification problems to regression problems in this paper. Our APRN is a plug-and-play structure that can be embedded in other state-of-the-art face recognition algorithms to improve their performance additionally. The experiments show that the proposed APRN consistently and significantly boosts the performance of baseline networks without extra computational costs in the inference phase. APRN achieves comparable or even superior to the state-of-the-art on CFP, Multi-PIE, IJB-A and MegaFace datasets. The code will be released[1], hoping to nourish our proposals to other computer vision fields.

Introduction

Face recognition is one of the most challenging topics in the computer vision field. Its performance has been significantly improved with the help of large-scale datasets (Guo et al. 2016) and the Convolutional Neural Network (Krizhevsky, Sutskever, and Hinton 2012). Even through top face recognition algorithms (Deng et al. 2019; Liu et al. 2017; Duan, Lu, and Zhou 2019; Kang et al. 2019; Liu et al. 2019a; Wang et al. 2019) have surpassed human performance in several evaluation datasets, Sengupta et al.(Sengupta et al. 2016) showed that the large intra-person variations caused by their poses degraded the performance of all state-of-the-art face recognition algorithms. The Pose-Invariant Face Recognition (PIFR) is far from being solved.

In this paper, we observe and prove that the hidden feature maps of the face recognition network are prone to pose changes. It makes identity representations fail to be pose-invariant. However, the existing algorithms for PIFR (Yin and Liu 2017; Tran, Yin, and Liu 2017; Shen et al. 2018; Peng et al. 2017; Cao et al. 2018a; Chen et al. 2016; Wang et al. 2019; Zhao et al. 2018; Deng et al. 2018) only focus on the final identity representations. Few of them have researched the hidden feature maps. Besides, most of the existing works (Tran, Yin, and Liu 2017; Shen et al. 2018; Deng et al. 2018) based on the adversarial learning defined pose discriminator as a classification task and need to generate synthetic faces. However, the pose estimation is contiguous in its 3D space objectively, and generating semantic identity-invariant synthetic faces is a challenging task that requiring lots of computational consumptions in the training phase.

Motivated by our observation, we propose a plug-and-play structure, which we name as Adversarial Pose Regression Network (APRN), to make the face recognition networks extract pose-invariant representations by reducing the pose variation in the hidden feature maps. The architecture of our proposal includes two modules, an APRN module and a face recognition network module. The pose discriminator in the APRN module employs the hidden feature maps of the face recognition network as inputs to estimate the facial poses, while the face recognition network reduces the pose variation to trick the discriminator and extracts representations for identification. The visualization of our architecture is in the supplementary material. Because the APRN module is removed in the inference phase, our APRN architecture has no extra computational costs for its inference. Besides, There is *NO GENERATOR* in our APRN, and it does not need to generate any identity-invariant synthetic faces. It makes our APRN much easier to be trained and needs less computational costs in each training iteration than other works (Tran, Yin, and Liu 2017; Shen et al. 2018; Deng et al. 2018). What's more, the APRN can be embedded in other existing general face recognition techniques (Deng et al. 2019; Kang et al. 2019; Wang et al. 2019) to improve their performance additionally.

To model the pose discriminator as a regression task in its 3D space, we propose a novel Adversarial Regression Loss Function and extend the adversarial learning from classification problems to regression problems.

This paper makes the following theoretical contributions:

- We observe and prove that feature maps from the hidden

[1] https://github.com/pengyuLPY/Adversarial-Pose-Regression-Network-for-Pose-Invariant-Face-Recognitions

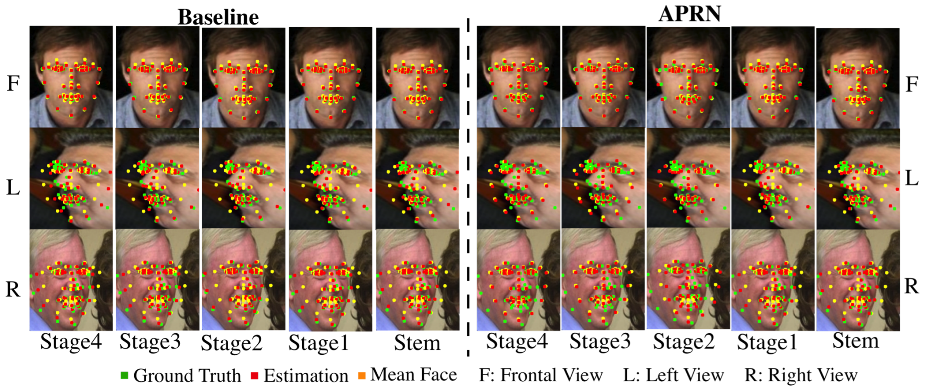

Figure 1: The estimation landmarks (red dots) are predicted precisely to the ground truth (green dots) based on the Baseline model's hidden feature maps, which proves its feature maps are pose-variant. APRN disentangles the pose variations in the hidden stages by transforming them into the frontal-view mean face ones (yellow dots).

layers are still prone to pose changes, which makes the identity representations fail to be pose-invariant.

- We propose a novel plug-and-play structure named as Adversarial Pose Regression Network (APRN) for PIFR. The APRN can be embedded in existing face recognition networks to make the representations pose-invariant without extra computational consumptions in the inference.
- We propose an Adversarial Regression Loss Function to extend the adversarial learning mechanism from classification task to regression task and to model pose discriminator as a contiguous regression network in the 3D space.

Because of the theoretical contributions above, given a face recognition network, the proposed APRN consistently and significantly boosts its performance. Our experiments show APRN achieves state-of-the-art performance on the CFP-FP, Multi-PIE and IJB-A evaluation protocol. As far as we know, we get the best single model without the private large-scale training datasets in MegaFace Challenge 1.

Related Work

Face Recognition. Face recognition is one of the most broadly researched topics in computer vision fields. Sun et al.(Sun, Wang, and Tang 2014) proposed a Convolutional Neural Networks for the face recognition problem. Besides, some novel loss functions (Deng et al. 2019; Schroff, Kaleni ko, and Philbin 2015; Wang et al. 2018b; Wen et al. 2016; Duan, Lu, and Zhou 2019) were proposed to make the learned representations compact in intra-identify and dispersive in inter-identify. However, Sengupta et al.(Sengupta et al. 2016) showed that most state-of-the-art algorithms degrade their performance dramatically from frontal-frontal to frontal-profile face verification. The Pose-Invariant Face Recognition is far from being solved.

Pose-Invariant Face Recognition (PIFR). Existing algorithms for PIFR are clustered into three groups: 1) Face synthesis on frontal view. Yin et al.(Yin et al. 2017a) proposed a 3DMM conditioned GAN to frontal the large-pose faces. Deng et al.(Deng et al. 2018) combined local and global adversarial CNNs to learn an identity-preserving facial UV completion model. 2) Pose-invariant feature extraction. Yin et al.(Yin and Liu 2017) learned different feature extractors with Multi-Task Convolutional Neural Network for different poses. Peng et al.(Peng et al. 2017) proposed a new feature reconstruction metric learning to disentangle identity explicitly and pose features. 3) An ensemble approach of the above two. Shen et al.(Shen et al. 2018) extended the Generative Adversarial Network to a three-player game, which helped it generate faces of arbitrary viewpoint and expression while preserving identity. Tran et al.(Tran, Yin, and Liu 2017) learned both a face synthesis generative and a discriminative representation based on the Generative Adversarial Network. However, most of the previous works only focused on the final representations. Few of them have researched pose variation in the hidden feature maps.

Adversarial Learning. Adversarial learning achieves a significant improvement in recent years. Especially the Generative Adversarial Network (GAN) has been widely used in many computer vision fields such as image synthesis (Radford, Metz, and Chintala 2015), image super-resolution (Ledig et al. 2017), and representation learning (Tran, Yin, and Liu 2017; Shen et al. 2018). In most previous works, Adversarial Learning was modeled as a binary classification task or N+1 categories classification task. Few of them have modeled adversarial learning as a regression task.

Inspired by the previous works, we propose an Adversarial Pose Regression Network and a novel Adversarial Regression Loss Function to disentangle the pose variations in face recognition networks.

The Proposal: Adversarial Pose Regression Network

In this section, we introduce our observation that the hidden feature maps of deep face recognition networks are prone to pose changes. Based on the observation, we formulate our Adversarial Pose Regression Network (APRN). Finally, we propose the Adversarial Regression Loss Function.

Observation

There are lots of works proved the facial poses can be estimated from a set of specific facial landmarks and the poses are by-products of facial landmarks (Gee and Cipolla 1994, 1996; Horprasert, Yacoob, and Davis 1996; Ruiz, Chong, and Rehg 2018). Based on their works, we infer that a feature map of a face recognition network is variant with pose changes if it can be used to precisely estimate the facial landmarks.

In this sub-section, A ResNet-101 (He et al. 2016) with the Softmax loss function is trained as the Baseline face recognition network. Besides, we train five facial landmarks regressive estimators respectively. The input of each estimator is the hidden feature map (stem, stage1, stage2, stage3, or stage4) of the Baseline face recognition network. We fix all the learnable parameters of the Baseline network when the estimators are training. The details of the network and the estimators are shown in the Table 1. The training dataset is MS-Celeb-1M (Guo et al. 2016). The ground truth of landmarks is annotated automatically (Hu et al. 2018a).

Face Recognition Network				APRN			
Module	Layer	Output Size		Inputs	Conv/BN/Relu 3×3, stride 2	Maxpool	FC
Layer Number		101	152				
Input Image		224*224		stem	[256,512, 1024,2048]	3*3 stride 2	2*M
Feature Extractor	stem	56*56					
	stage1	56*56		stage1	[512,1024,2048]		
	stage2	28*28		stage2	[1024,2048]		
	stage3	14*14		stage3	[2048]		
	stage4	7*7		stage4	[2048]		
	Identity Feature	256	512	Loss	pose estimators: L_1 Loss		
					pose discriminators: Adversarial Regression Loss		
Identification Classifier		N	N				

Table 1: The structure of the Pose-Invariant Face Recognition with Adversarial Pose Regression Network. N is the number of identities. M is the number of facial landmarks.

Table 2 shows that the Baseline model achieves excellent performance in all evaluation datasets, especially in the CFP-FP dataset which is a frontal-profile evaluation dataset. However, Table 3 shows the error range of its landmark estimators is from 1.18 (stem) to 1.50 (stage4). It proves the hidden feature maps of the good performance face recognition network have enough capacity to estimate facial landmarks precisely. The landmarks visualization in Figure 1 also supports the conclusion because the estimated landmarks (red dots) of the Baseline model are close to the ground truth (green dots) on all three views, even though the feature maps are extracted from the deepest layer (stage 4).

LFW	IJB-A (FAR=10^{-4})	CFP-FF	CFP-FP	Megaface unclean\clean
99.37	86.66	99.70	96.24	77.58\91.59

Table 2: Face recognition performance of the Baseline.

Furthermore, we do t-SNE analysis (Maaten and Hinton 2008) in Figure 2. Based on the figure, it can be inferred that the representations of the same identity are not compact because of the pose variation, even though the distribution of identities is almost discriminative. More discussion about t-SNE analysis is detailed in the ablation study.

Inputs	Stem	Stage1	Stage2	Stage3	Stage4
L_1 Distance	1.18	1.36	1.44	1.46	1.50

Table 3: Pose estimation performance of the Baseline.

Based on the experiments shown in Table 2, Table 3, Figure 1 and Figure 2, it is proved that even though the face recognition network gets good performance in evaluation datasets, its feature maps from hidden layers are still prone to pose changes.

Adversarial Pose Regression Network

Based on our observation, we propose the Adversarial Pose Regression Network (APRN) to extract pose-invariant identity representations by reducing the pose variation in the hidden feature maps.

The face recognition network with our APRN is formulated as Equation 1:

$$\mathbb{L} = \alpha V(D, R') + Lc(R) \tag{1}$$

The $\mathbb{L}$ is the loss function of our architecture. The D denotes the pose discriminator in APRN, which is modeled as a facial landmarks estimator. R denotes the face recognition network, and R' is a subset of R ($R' \in R$). The α is the loss weight to balance $V(D, R')$ and $Lc(R)$.

The $V(D, R')$ is an adversarial loss function. It can be formulated as (Goodfellow et al. 2014) or (Tran, Yin, and Liu 2017) if the pose discriminator is modeled as a discrete task. However, the pose estimation is contiguous, and it is better to be modeled as a regression task than a classification one. To extend the $V(D, R')$ as a contiguous regression task, we propose the Adversarial Regression Loss Function in the following section.

There is a crucial difference between our $V(D, R')$ and the existing works based on GAN (Cao et al. 2018a; Deng et al. 2018; Shen et al. 2018; Tran, Yin, and Liu 2017). There is *NO GENERATOR* in our APRN. The R' is not a generator but a part of the face recognition network. Our APRN does not need to generate any identity-invariant fake faces, which helps it easier to be trained and need much less computational consumptions in each training iteration than the previous works.

The $Lc(R)$ is an identification loss function. It can be set as Softmax loss function, Large Margin Cosine loss function (Wang et al. 2018b), Additive Angular Margin loss function (Deng et al. 2019) etc. The identification loss function is not the scope of this paper.

Based on Equation 1, the architecture with APRN consists of two modules, an APRN module and a face recognition network module. The pose discriminator in the APRN module employs the hidden feature maps of the face recognition network as inputs to regress the facial landmarks (optimized by the $V(D, R')$). The face recognition network reduces the pose variations to trick the discriminator (optimized by the

$V(D, R')$) and extracts the representations for identification (optimized by the $Lc(R)$). We visualize the architecture in the supplementary material. There are no extra computational costs in the inference phase because the APRN module is removed in the phase.

Adversarial Regression Loss Function

As the adversarial learning was treated as a binary classification or an N+1 categories classification in the previous works, the pose discriminator has to be modeled as a classification network in (Tran, Yin, and Liu 2017; Yin and Liu 2017), even though pose estimation is contiguous objectively. To model the pose discriminator as a regression network in adversarial learning, we propose an Adversarial Regression Loss Function and extend the adversarial learning from the classification task to the regression task.

As the face recognition algorithms have achieved an excellent performance in the frontal-view (Deng et al. 2019; Kang et al. 2019; Liu et al. 2019a; Wang et al. 2019), we disentangle the pose variations by transferring them to the same as the frontal-view mean face. Our Adversarial Regression Loss Function is formulated as the following:

$$\begin{aligned} L_{Adv\_Reg}(D, R') = \arg\min_{D} E[\|D(R'(x)) - l\|_1^1] \ + \\ \arg\min_{R'} E[\|D(R'(x)) - \tilde{l}\|_1^1] \end{aligned} \tag{2}$$

The x is the input face image. The l denotes the ground truth facial landmarks of x, and the $\tilde{l}$ denotes landmarks of the frontal-view mean face.

There are two parts in $L_{Adv\_Reg}(D, R')$. The D and R' are optimized respectively in different parts. In the first part, the discriminator (D) is optimized to estimate the facial landmarks (l).In the second part, the subset of face recognition network (R') is optimized to cheat the D by making its outputs close to the frontal-view mean face landmarks ($\tilde{l}$). The targets of the $D(R'(x))$ are adversarial in different parts of $L_{Adv\_Reg}(D, R')$. It makes our formulation be an adversarial learning mechanism.

Both parts in $L_{Adv\_Reg}(D, R')$ are essential in our proposal. The face recognition network would not be optimized to be pose-invariant if the second part is absent. Removing the first part and optimizing both D and R' in the second part leads D to output $\tilde{l}$ directly, even though the $R'(x)$ is generated by random noise. The influence of each part in Equation 2 is analyzed in the ablation study.

The $L_{Adv\_Reg}(D, R')$ is optimized iteratively with the Stochastic Gradient Descent (SGD). For the details on the optimization, we refer to (Goodfellow et al. 2014).

Experiment

Datasets

In this paper, MS-Celeb-1M (MS-1M) (Guo et al. 2016) and CASIA-WebFace(CASIA) (Yi et al. 2014) are used as training datasets *respectively* in different experiments. The ground truth of facial landmarks in both training datasets is annotated automatically with the algorithms proposed in (Hu et al. 2018a). In the **MS-Celeb-1M**, there are almost 100 thousand global celebrities and 10 million images released. We clean the dataset with the automatic method proposed in (Wu et al. 2018). The cleaned MS-Celeb-1M dataset in this paper contains 74,974 identities and 4.8 million images. The **CASIA-WebFace** consists of 494,414 near-frontal faces of 10,575 subjects from the internet.

Multi-PIE (Gross et al. 2010) dataset consists of 754, 200 images of 337 subjects. We follow the setting in (Zhao et al. 2018; Hu et al. 2018b; Yin et al. 2017b), 337 subjects with neutral expression, 13 poses within ±90° and 20 illuminations are used. The first 200 subjects are used for training. The rest 137 subjects are used for testing.

CFP (Sengupta et al. 2016), LFW (Huang et al. 2008), IJB-A (Klare et al. 2015) and MegaFace (Kemelmacher-Shlizerman et al. 2016) are used as evaluation datasets. The **CFP** consist of 10 folders, each folder contains 350 same-person pairs and 350 different-person pairs for both frontal-frontal (CFP-FF) and frontal-profile (CFP-FP) experiments. The **LFW** consists of 13,323 web photos of 5,749 celebrities which are divided into 6,000 face pairs in 10 splits. In this paper, we follow the standard protocols of LFW and CFP and report their mean accuracy and the standard error of the mean. The **IJB-A** dataset contains 5,397 images and 20,412 video frames split from 2,042 videos of 500 individuals. We evaluate the performance with its standard verification protocol. The true accepted rates (TAR) under varying false accepted rates (FAR) are reported. The **MegaFace** dataset includes the probe and gallery set. The probe set is the FaceScrub dataset (Ng and Winkler 2014), which contains 100,000 images of 530 identities, and the gallery set consists of about 1,027, 060 images from 690,572 different subjects. We report its rank1@10^6 accuracy[2]. We report both the performance tested on the original dataset and the cleaned one by Deng et al.(Deng et al. 2019)[3].

Implementation Details

For proving APRN can be embedded in different face recognition networks, three backbones are trained in this paper. They are CASIA-Net (Yi et al. 2014), ResNet-101 and ResNet-152 (He et al. 2016). ResNet-152 use the ArcFace loss (Deng et al. 2019) as $Lc(R)$ while CASIA-Net and ResNet-101 use the Softmax loss function simply. We train two models for every backbone, one is the Baseline model and the other one is our APRN. All three ARPN models use our proposed Adversarial Regression Loss Function as $V(D, R')$. Their pose discriminators employ the last convolutional layer of face recognition network as input except for the ablation study section. The training details are the same between Baseline models and APRN models, except the APRN models are pre-trained by the Baseline models and their learning rates are 10 times smaller than the Baselines. All the models are implemented on the publicly available PyTorch platform (Paszke et al. 2017).

Implementation details of CASIA-Net. For a fair comparison with other state-of-the-art works on the CFP dataset, we train the CASIA-Net with CASIA-WebFace. The imple-

[2] http://megaface.cs.washington.edu

[3] https://github.com/deepinsight/insightface

mentation details of our CAISA-Net is the same as other PIFR algorithms (Sengupta et al. 2016; Chen et al. 2016; Tran, Yin, and Liu 2017; Peng et al. 2017). We randomly sample 96×96 regions from the aligned 100×100 face images for data augmentation. Image intensities are linearly scaled to the range of $[-1, 1]$. The network is trained for 30 epochs. The learning rate of Baseline models is 0.1 and decays 10 times at the 20_{th}, 27_{th} and 29_{th} epoch. Momentum is 0.9, weight decay is 0.0005, and α in Equation 1 is 0.2. The pose discriminator is detailed in Table 4.

For fairly comparing with other works on the Multi-PIE dataset, we train another CASIA-Net with the same implementation details except includes the Multi-PIE as the training dataset. The protocol of Multi-PIE is the same as (Zhao et al. 2018; Hu et al. 2018b; Yin et al. 2017b).

Inputs	The last convolutional layer
Conv+BN+Relu	3×3×320,stride 1
	3×3×320,stride 1
	3×3×320,stride 1
Fully Connected (FC) Layer	M×2

Table 4: Structure of the pose discriminator for CASIA-Net.

Implementation details of ResNet-101 and ResNet-152. The details of the networks is introduced in the Table 1. We import the ArcFace loss function (Deng et al. 2019) as the $Lc(R)$ of the ResNet-152 to proves our APRN can be embedded in the state-of-the-art works to boosts their performance. Furthermore, the ResNet-101 with the Softmax loss function proves that the APRN works well in the complex backbone with a simple loss function as $Lc(R)$. The training dataset for both networks is the cleaned MS-Celeb-1M datasets. We also train a ResNet-101 with ArcFace loss function in the the supplementary material. We randomly sample 224×224 regions from the aligned 230×230 face images for data augmentation. Image intensities are scaled to [-1, 1]. The networks are optimized with SGD for 20 epochs. The learning rate of the Baseline models is 0.01 and decays ten times at the 16_{th}, 18_{th} and 19_{th} epoch. Momentum is 0.9, weight decay is 0.0005, and α is 0.2.

Comparison with the State-of-the-art

Comparison on CFP. Table 5 shows that our APRN improves the performance of the Baseline model from 90.46% to 94.61% and outperforms the state-of-the-art method (94.39%, Multi-task) in the Frontal-Profile protocol. Besides, in Frontal-Frontal protocol, the APRN also improves the performance of the Baseline model from 98.43% to 99.19% and surpasses the previous best model (98.83%, UV-GAN) trained in the small-scale dataset (CAISA). Furthermore, our APRN needs much less computational cost than previous methods (Tran, Yin, and Liu 2017; Yin and Liu 2017; Peng et al. 2017). It is because there are only three convolutional layers and a single fully-connected layer in our APRN. The extra module is much simpler than theirs. The computational consumption analysis is detailed in the supplementary material.

	Training Dataset	CFP-FF	CFP-FP
Sengupta et al.(Sengupta et al. 2016)	CASIA	96.40	84.91
Sankaran et al. (Sankaranarayanan et al. 2016)	CASIA	96.93	89.17
Chen et al.(Chen et al. 2016)	CASIA	98.67	91.97
Peng et al.(Peng et al. 2017)	CASIA	98.67	93.76
DR-GAN(Tran, Yin, and Liu 2017)	CASIA	97.84	93.41
Multi-task(Yin and Liu 2017)	CASIA	97.79	94.39
UV-GAN(Deng et al. 2018)	CASIA, UVDB	98.83	93.09
PIM(Zhao et al. 2018)	**MS-1M**	**99.44**	93.10
Co-Mining(Wang et al. 2019)	CASIA,	-	91.75
Ours: Baseline	CASIA	98.43	90.46
Ours: APRN	CASIA	**99.19**	**94.61**

Table 5: Performance on CFP (CASIA-Net).

Comparison on Multi-PIE. Table 6 shows that our APRN improves the Baseline model significantly. Especially in the large pose variations views, our APRN improves the performance from 88.8% to 92.8% in the ±75° view and from 65.6% to 77.1% in the ±90° view. The performance of APRN is superior to the state-of-the-art.

Methods	±0°	±30°	±60°	±75°	±90°
Zhu et al. (Zhu et al. 2013)	-	98.5	-	-	-
Zhu et al. (Zhu et al. 2014)	95.7	83.7	60.1	-	-
Yim et al. (Yim et al. 2015)	99.5	88.5	61.9	-	-
MvDN (Kan, Shan, and Chen 2016)	-	99.1	89.7	81.0	70.7
DR-GAN (Tran, Yin, and Liu 2017)	97.0	90.1	83.2	-	-
FF-GAN (Yin et al. 2017b)	95.7	92.5	85.2	77.2	61.2
CAPG-GAN (Hu et al. 2018b)	-	99.6	90.6	83.1	66.1
PIM (Zhao et al. 2018)	-	99.4	97.7	91.2	75.0
ours:Baseline	**100**	**99.9**	**98.8**	88.8	65.6
ours:ARPN	99.9	**99.9**	**98.8**	**92.8**	**77.1**

Table 6: Performance on the Multi-PIE (CASIA-Net).

Comparison on IJB-A. Table 7 shows that our APRN improves the performance of the Baseline from 93.1% to 95.1% when FAR is 10^{-3}. It is superior to the state-of-the-art which the AFRN gets. When the FAR is 10^{-1} or 10^{-2}, the APRN also improves the Baseline model and achieves comparable performance to the AFRN.

Comparison on MegaFace. Table 8 shows that our APRN improves the rank1@10^6 accuracy of Baseline ResNet-152 from 99.49% to 99.78%. In particular, $L_c(R)$ of the ResNet-152 is the same as the ArcFace (Deng et al. 2019) and the performance of Baseline has been already superior to the ArcFace (98.35%). It provides strong evidence for the effectiveness of APRN. Because the Baseline that exceeding state-of-the-art is much more difficult to be improved, but our APRN did it. The improvement also proves our APRN can be embedded in other state-of-the-art face recognition network (such as ArcFace) to improve it additionally. Besides, The APRN is superior to the 4_{th} performance (99.42%) on the MegaFace leaderboard[4]. As far as

[4] http://megaface.cs.washington.edu/results/facescrub.html

Methods	FAR=			
	10^{-1}	10^{-2}	10^{-3}	10^{-4}
DR-GAN(Tran, Yin, and Liu 2017)	-	77.4	53.9	-
Multi-task(Yin and Liu 2017)	-	77.5	53.9	-
FF-GAN(Yin et al. 2017b)	-	85.2	66.3	-
Contrastive CNN(Han et al. 2018)	95.3	84.0	63.9	-
DREAM(Cao et al. 2018a)	-	89.1	76.4	-
VGGFace2_ft(Cao et al. 2018b)	99.0	96.8	92.1	-
PRN^{+}(Kang, Kim, and Kim 2018)	98.8	96.5	91.9	-
FTL(Yin et al. 2019)	-	95.3	91.2	-
UniformFace(Duan, Lu, and Zhou 2019)	-	96.9	92.3	-
AFRN(Kang et al. 2019)	**99.8**	**98.5**	94.9	-
ours:Baseline	98.8	97.0	93.1	86.7
ours:APRN	99.1	98.2	**95.1**	**90.1**

Table 7: Performance on IJB-A (ResNet-101).

we know, our APRN is the best single model trained without private large-scale training datasets.[5]

	Methods	rank1@10^6 unclean/clean
Leaderboard	top1: Sogou AIGROUP-SFace*	-/99.94
	top2: SRC-Beijing-FR*	-/99.89
	top3: SenseTime PureFace*	-/99.80
	top4: EI Networks*	-/99.42
Publications	CosFace(Wang et al. 2018b)	84.26/97.91
	IMDB-Face(Wang et al. 2018a)	84.06/-
	FairLoss(Liu et al. 2019a)	77.45/-
	UniformFace (Duan, Lu, and Zhou 2019)	79.98/-
	Co-Mining(Wang et al. 2019)	-/87.37
	AdaptiveFace(Liu et al. 2019b)	-/95.02
	ArcFace(Deng et al. 2019)	-/98.35
Ours: ResNet-101	Baseline	77.58/91.59
	APRN	79.21/93.91
Ours: ResNet-152	Baseline	84.76/99.49
	APRN	**84.78/99.78**

Table 8: Performance on MegaFace Challenge 1. The methods with * means the private large-scale dataset was used.

Ablation Study

In this sub-section, we prove our APRN boosts the performance of Baseline models firstly.Secondly, we prove our APRN disentangles the pose variations in the hidden feature maps. Thirdly, we study the influence of the input layers for our APRN. Fourthly, we compare the performance of APRN trained with different α to study the influence of loss weight in Equation 1. Finally, we prove that both parts in our $L_{Adv\_Reg}(D, R')$ are indispensable. All the models in this sub-section are trained with MS-Celeb-1M. The architecture is based on the ResNet-101 shown in Table 1.

As a plug-and-play structure, our APRN boosts the performance of Baseline models consistently and significantly. We compare all Baseline models and the APRN models in Table 9. The table shows that our APRN always improves the performance of Baseline models on all evaluation datasets, no matter the Baseline model is a small-scale network as CASIA-Net, a medium-scale network as ResNet-101, or a large-scale one as ResNet-152.

Inference Backbone		LFW	IJB-A @10^{-4}	CFP-FF	CFP-FP	MegaFace unclean/clean
CASIA-Net	Baseline	97.80	61.85	98.43	90.46	60.79/68.94
	APRN	**98.80**	**64.72**	**99.19**	**94.61**	**61.01/70.12**
ResNet-101	Baseline	99.37	86.66	99.70	96.24	77.58/91.59
	APRN	**99.50**	**90.10**	**99.80**	**96.84**	**79.21/93.91**
ResNet-152	Baseline	99.59	91.29	99.93	97.75	84.76/99.49
	APRN	**99.62**	**93.62**	**99.94**	**97.89**	**84.78/99.78**

Table 9: As a plug-and-play structure, the APRN consistently and significantly boosts the performance of Baselines.

Our APRN disentangles the pose variations in the hidden feature maps. We do t-SNE analysis in Figure 2. We get three observations based on the figure and prove the observations quantitatively in the following: 1) The first column (A) in the figure shows that inter-identity representations (denoted as different colors) are discriminative both in the Baseline and our APRN. Their performance in Table 9 also proves the observation. 2) The second column (B) shows that the representations of intra-identity (denoted as the same colors) are much more compact in the APRN than in the Baseline. We calculate the cosine distance between every representation and the centroid of the identity. Mean±std of the distance in Baseline is **0.792±0.122**. The APRN's is **0.804±0.14**. The statistics result supports the observation in the second column. 3) The third column (C) shows that our APRN makes the representation of the same identity distributed uniformly in its discriminative local space regardless of their poses, but there is a margin in the Baseline to separate the frontal-view (denoted as dots) from the profile-view (denoted as circles). A linear Support Vector Machine (SVM) is trained to classify the frontal face and profile face to prove the observation. We sample 5000 images from the CFP dataset to be the training dataset and 2000 images to be the testing dataset. The pose classification accuracy on the Baseline Model is **94.95%/95.48%** (val/train). The one on the APRN is **85.20%/88.44%** (val/train). The SVM classifiers prove the APRN disentangles the pose variations.

The influence of the input hidden layer. We employ different hidden layers as the inputs of discriminators to study the influence of the input layers for our APRN. The pose discriminators for different hidden layers are shown in the Table 1. The face recognition accuracy is shown in Table 10 and the landmark estimation is illustrated in Figure 1.

Based on Table 10, we find:

1) Except for the *stem* layer, the APRN improves the performance of the Baseline network on all datasets no matter which layer is employed as the input.

2) The performance is better when the pose discriminator

[5] We also train a ResNet-101 with arcFace loss as $L_c(R)$. APRN improve its performance to 98.59% and outperformance the (Deng et al. 2019). More details are in the supplementary material.

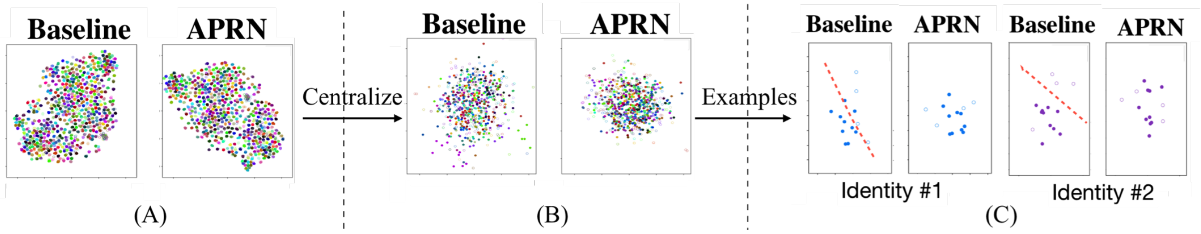

Figure 2: t-SNE analysis for our ResNet-101 in the CFP dataset. The different colors denote different subjects. The dots are the representations in frontal-view faces, and the circles are in the profile-view.

Inputs	LFW	JB-A @10^{-4}	CFP-FF	CFP-FP	MegaFace unclean/clean
None	99.37	86.66	99.70	96.24	77.58\91.59
stem	99.3	86.50	99.33	96.17	77.02\91.51
stage1	99.43	87.76	99.70	96.33	78.01\92.07
stage2	99.38	87.74	99.80	96.61	78.34\92.30
stage3	99.47	87.91	99.74	96.61	**79.29**\93.16
stage4	**99.50**	**90.10**	**99.80**	**96.84**	79.21\93.91

Table 10: The influence of the hidden layers to APRN. *None* is the Baseline model.

is equipped to the deeper layers. Especially on the CFP-FP and the MegaFace dataset, the performance is always improved by employing the deeper layer as the input of pose discriminator. It is because the pose variations are easier to be disentangled in deep layers than in the shallow ones.

3)The performances on CFP-FP, IJB-A, and MegaFace datasets are improved more than on LFW and CFP-FF. It is because the CFP-FP is a frontal-profile dataset, IJB-A contains extreme pose variations, and the MegaFace is collected in the wild with 1 million images.All of them have a stricter requirement for pose-invariance than the LFW and CFP-FF.

The influence of the loss weight. There is a loss weight α in Equation 1. We employ the stage4 layer as the input of pose discriminator and use different α to evaluate the influence of α. As Table 11 shows our APRN improves the performance of the Baseline model ($\alpha = 0.0$) when the α is set to be 0.1, 0.2, or 0.3. It improves most when the α equals to 0.2. However, the performance decreases when the α is 0.4. We think it is because a big α would lead the model to ignore the Lc(R) and lose some capacity of identification.

α	LFW	JB-A @10^{-4}	CFP-FF	CFP-FP	MegaFace unclean/clean
0.0	99.37	86.66	99.70	96.24	77.58/91.59
0.1	99.45	**90.42**	99.71	96.61	78.07/92.96
0.2	**99.50**	90.10	**99.80**	**96.84**	**79.21/93.91**
0.3	99.48	89.59	99.69	96.31	77.98/92.78
0.4	99.26	86.67	99.74	95.77	75.90/88.32

Table 11: The influence of α. $\alpha = 0.0$ is the Baseline model.

The influence of the two parts in $L_{Adv\_Reg}(D, R')$. To analyze the influence of the two parts in Equation 2, we reformulate the $L_{Adv\_Reg}(D, R')$ to be $L^1_{Reg}(D, R')$ and $L^2_{Reg}(D, R')$ as Equation 3 shows.

$$\begin{aligned} L^1_{Reg}(D, R') &= \arg\min_{D,R'} E[\|D(R'(x)) - l\|_1] \\ L^2_{Reg}(D, R') &= \arg\min_{D,R'} E[\|D(R'(x)) - \tilde{l}\|_1] \end{aligned} \tag{3}$$

$L^1_{Reg}(D, R')$ is the reformulation of the first part of $L_{Adv\_Reg}(D, R')$, and the $L^2_{Reg}(D, R')$ is the reformulation of the second part. The target of $L^1_{Reg}(D, R')$ is to estimate the ground truth landmarks by optimizing both D and R' simultaneously. $L^2_{Reg}(D, R')$ attempts to estimate the frontal-view mean facial landmarks.

	LFW	JB-A @10^{-4}	CFP-FF	CFP-FP	Megaface (unclean/clean)
None	99.37	86.66	99.70	96.24	77.58/91.59
L^1_{Reg}	99.43	88.30	99.79	96.14	78.53/92.58
L^2_{Reg}	99.37	87.81	99.79	96.46	78.64/92.42
L_{Adv_Reg}	**99.50**	**90.10**	**99.80**	**96.84**	**79.21/93.91**

Table 12: The influence of the two parts in $L_{Adv\_Reg}(D, R')$. *None* is the Baseline model.

Table 12 shows that both $L^1_{Reg}(D, R')$ and $L^2_{Reg}(D, R')$ have limited help to the face recognition. We hypothesize the reason why $L^2_{Adv\_Reg}(D, R')$ doesn't work well is that D ignores the inputs ($R'(x)$) and does not optimize the R' to be pose-invariant. To prove our hypothesis, we put one hundred images whose pixels are random noise within [-1, 1] as the inputs of the network. The L_1 Distance between the prediction and the frontal-view mean face is 0.83 ± 0.13 if the model is trained with $L^2_{Reg}(D, R')$. However, the distance is 4.12 ± 0.67 if the model is trained with $L_{Adv\_Reg}(D, R')$.

Conclusion

In this paper, we proposed an Adversarial Pose Regression Network (APRN) for Pose-Invariant Face Recognition. We also extend adversarial learning from classification task to regression task and presented the Adversarial Regression Loss Function to model the pose discriminator of APRN in its contiguous 3D space. As a plug-and-play structure, APRN consistently and significantly boosts the performance of state-of-the-art networks without any extra computational costs in the inference phase.

References

Cao, K.; Rong, Y.; Li, C.; Tang, X.; and Change Loy, C. 2018a. Pose-robust face recognition via deep residual equivariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5187–5196.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018b. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 67–74. IEEE.

Chen, J.-C.; Zheng, J.; Patel, V. M.; and Chellappa, R. 2016. Fisher vector encoded deep convolutional features for unconstrained face verification. In *2016 IEEE International Conference on Image Processing (ICIP)*, 2981–2985. IEEE.

Deng, J.; Cheng, S.; Xue, N.; Zhou, Y.; and Zafeiriou, S. 2018. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7093–7102.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4690–4699.

Duan, Y.; Lu, J.; and Zhou, J. 2019. UniformFace: Learning Deep Equidistributed Representation for Face Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3415–3424.

Gee, A.; and Cipolla, R. 1994. Determining the gaze of faces in images. *Image and Vision Computing* 12(10): 639–647.

Gee, A.; and Cipolla, R. 1996. Fast visual tracking by temporal consensus. *Image and Vision Computing* 14(2): 105–114.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

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

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 87–102. Springer.

Han, C.; Shan, S.; Kan, M.; Wu, S.; and Chen, X. 2018. Face recognition with contrastive convolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 118–134.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity mappings in deep residual networks. In *Proceedings of the European conference on computer vision (ECCV)*, 630–645. Springer.

Horprasert, T.; Yacoob, Y.; and Davis, L. S. 1996. Computing 3-d head orientation from a monocular image sequence. In *Proceedings of the second international conference on automatic face and gesture recognition*, 242–247. IEEE.

Hu, T.; Qi, H.; Xu, J.; and Huang, Q. 2018a. Facial landmarks detection by self-iterative regression based landmarks-attention network. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Hu, Y.; Wu, X.; Yu, B.; He, R.; and Sun, Z. 2018b. Pose-guided photorealistic face rotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8398–8406.

Huang, G. B.; Mattar, M.; Berg, T.; and Learned-Miller, E. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

Kan, M.; Shan, S.; and Chen, X. 2016. Multi-view deep network for cross-view classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4847–4855.

Kang, B.-N.; Kim, Y.; and Kim, D. 2018. Pairwise relational networks for face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 628–645.

Kang, G.-N.; Kim, Y.; Jun, B.; and Kim, D. 2019. Attentional Feature-Pair Relation Networks for Accurate Face Recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Kemelmacher-Shlizerman, I.; Seitz, S. M.; Miller, D.; and Brossard, E. 2016. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4873–4882.

Klare, B. F.; Klein, B.; Taborsky, E.; Blanton, A.; Cheney, J.; Allen, K.; Grother, P.; Mah, A.; and Jain, A. K. 2015. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1931–1939.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4681–4690.

Liu, B.; Deng, W.; Zhong, Y.; Wang, M.; and Hu, j. 2019a. Fair Loss: Margin-Aware Reinforcement Learning for Deep Face Recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Liu, H.; Zhu, X.; Lei, Z.; and Li, S. Z. 2019b. AdaptiveFace: Adaptive Margin and Sampling for Face Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11947–11956.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 212–220.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

Ng, H.-W.; and Winkler, S. 2014. A data-driven approach to cleaning large face datasets. In *Image Processing (ICIP), 2014 IEEE International Conference on*, 343–347. IEEE.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.

Peng, X.; Yu, X.; Sohn, K.; Metaxas, D. N.; and Chandraker, M. 2017. Reconstruction-based disentanglement for pose-invariant face recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1623–1632.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* .

Ruiz, N.; Chong, E.; and Rehg, J. M. 2018. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2074–2083.

Sankaranarayanan, S.; Alavi, A.; Castillo, C. D.; and Chellappa, R. 2016. Triplet probabilistic embedding for face verification and clustering. In *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*, 1–8. IEEE.

Schroff, F.; Kaleni ko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823.

Sengupta, S.; Chen, J.-C.; Castillo, C.; Patel, V. M.; Chellappa, R.; and Jacobs, D. W. 2016. Frontal to profile face verification in the wild. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, 1–9. IEEE.

Shen, Y.; Luo, P.; Yan, J.; Wang, X.; and Tang, X. 2018. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 821–830.

Sun, Y.; Wang, X.; and Tang, X. 2014. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1891–1898.

Tran, L.; Yin, X.; and Liu, X. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1415–1424.

Wang, F.; Chen, L.; Li, C.; Huang, S.; Chen, Y.; Qian, C.; and Change Loy, C. 2018a. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 765–780.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018b. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5265–5274.

Wang, X.; Wang, S.; Wang, J.; Shi, H.; and Mei, T. 2019. Co-Mining: Deep Face Recognition with Noisy Labels. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).*

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, 499–515. Springer.

Wu, X.; He, R.; Sun, Z.; and Tan, T. 2018. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security* 13(11): 2884–2896.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* .

Yim, J.; Jung, H.; Yoo, B.; Choi, C.; Park, D.; and Kim, J. 2015. Rotating your face using multi-task deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 676–684.

Yin, X.; and Liu, X. 2017. Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing (TIP)* 27(2): 964–975.

Yin, X.; Yu, X.; Sohn, K.; Liu, X.; and Chandraker, M. 2017a. Towards large-pose face frontalization in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3990–3999.

Yin, X.; Yu, X.; Sohn, K.; Liu, X.; and Chandraker, M. 2017b. Towards large-pose face frontalization in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3990–3999.

Yin, X.; Yu, X.; Sohn, K.; Liu, X.; and Chandraker, M. 2019. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5704–5713.

Zhao, J.; Cheng, Y.; Xu, Y.; Xiong, L.; Li, J.; Zhao, F.; Jayashree, K.; Pranata, S.; Shen, S.; Xing, J.; et al. 2018. Towards pose invariant face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2207–2216.

Zhu, Z.; Luo, P.; Wang, X.; and Tang, X. 2013. Deep learning identity-preserving face space. In *Proceedings of the IEEE International Conference on Computer Vision*, 113–120.

Zhu, Z.; Luo, P.; Wang, X.; and Tang, X. 2014. Multi-view perceptron: a deep model for learning face identity and view representations. In *Advances in Neural Information Processing Systems*, 217–225.