# Pose-Invariant Face Recognition via Adaptive Angular Distillation

**Zhenduo Zhang, Yongru Chen, Wenming Yang*, Guijin Wang, Qingmin Liao**

Department of Electronic Engineering/Shenzhen International Graduate School, Tsinghua University, China
{zhangzd18@mails, yongru-c19@mails, yang.wenming@sz, wangguijin@, liaoqm@}.tsinghua.edu.cn

## Abstract

Pose-invariant face recognition is a practically useful but challenging task. This paper introduces a novel method to learn pose-invariant feature representation without normalizing profile faces to frontal ones or learning disentangled features. We first design a novel strategy to learn pose-invariant feature embeddings by distilling the angular knowledge of frontal faces extracted by teacher network to student network, which enables the handling of faces with large pose variations. In this way, the features of faces across variant poses can cluster compactly for the same person to create a pose-invariant face representation. Secondly, we propose a Pose-Adaptive Angular Distillation loss to mitigate the negative effect of uneven distribution of face poses in the training dataset to pay more attention to the samples with large pose variations. Extensive experiments on two challenging benchmarks (IJB-A and CFP-FP) show that our approach consistently outperforms the existing methods.

## Introduction

Pose-Invariant Face Recognition(PIFR) has drawn great attention under either controlled lab environment or unrestricted environment due to the increasing demand for face recognition (FR) systems. Deep learning methods have recently achieved great success in FR, but these approaches heavily rely on sufficient data. However, as it is impractical to collect massive images containing faces across all the pose variations, which will lead deep networks to be biased towards distinguishing frontal faces.

To compensate for the pose, appearance-level face alignment is first conducted as the standard preprocessing. Appearance level alignment usually warps the face to the frontal or designated pose. Consequently, the FR systems only need to compare faces under the same pose. Although these approaches have achieved the state-of-the-art performance in pose-invariant face recognition, they inevitably bring an extra processing burden and may introduce some artifacts, harming face recognition (Cao et al. 2018b). To avoid these limitations, feature level alignment methods explore a discriminative identity feature space, regardless of pose variations (Kan et al. 2016, 2014). Thanks to deep learning development, disentangled representation learning has become one of the most effective approaches to learn discriminative identity features. However, in order to supervise the extraction of disentangled features, the auxiliary information is usually needed (Peng et al. 2017), which inevitably leads to additional estimation errors or domain bias.

Our proposed approach tries to overcome the weaknesses of the above methods. Firstly, to tackle the limitation of appearance-level face alignment, we follow the line of feature level alignment and propose to learn a pose-invariant discriminant identity feature. To tackle the limitation of feature disentanglement methods, we try to learn the pose-invariant feature without manual label annotation of poses, which is needed in feature disentanglement.

In this paper, to assure the pose-invariance and representation capacity of features, we propose to recover the feature of the frontal face from the features extracted from faces of arbitrary poses with the same identity. However, the main difficulty is the full information of the frontal face cannot be inferred from a profile face, leading to an ill-posed task. We utilize the frontal face as a supervision signal and encourage the student network to mimic the feature of the frontal face when dealing with the non-frontal face.

To implement this perspective, we propose a Pose-Adaptive Angular Distillation (PAD) loss to distill the angular knowledge of frontal faces from teacher network to student network. In this way, the features of faces across variant poses can cluster compactly for the same person, to create a pose-invariant face representation with small intra-class distance and large inter-class distance. Besides, the PAD loss endows each sample with different weights according to its pose and hard level so as to overcome the limitation of the uneven distribution of pose.

We conduct extensive qualitative and quantitative experiments on two challenging benchmarks: CFP (Sengupta et al. 2016) and IJB-A (Klare et al. 2015). The results illustrate the effectiveness of our method of recognizing faces with extreme poses and the superiority over the exsiting methods on CFP and IJB-A.

- We propose a new perspective to learn pose-invariant feature embeddings: recovering the feature of the frontal face from the features extracted from faces of arbitrary poses with the same identity.
- A Pose-Adaptive Angular Distillation(PAD) loss that

*Corresponding author: Wenming Yang

overcomes the limitation of uneven distribution of pose and further improves the performance of PIFR on the challenging benchmarks.

- Experiments on two challenging benchmarks (CFP and IJB-A) demonstrate that our approach favorably performs against the existing methods of pose-invariant face recognition.

## Related Work

**Deep-learning Approaches of PIFR** Deep learning has dominated the field of pose-invariant face recognition. Recently, there are two main categories of methods that achieve quite a competitive performance in this field: Face Frontalization and Feature Disentangle. Face frontalization mainly normalizes faces with different poses to frontal faces of the same identity in image space before recognition (Dovgard and Basri 2004; Ferrari et al. 2016; Hassner 2013; Hassner et al. 2015; Zhu et al. 2015). Such methods are mainly based on the Generative Adversarial Network(GAN) (Goodfellow et al. 2014). For instance, DR-GAN (Tran, Yin, and Liu 2017), TP-GAN (Huang et al. 2017), CAPG-GAN (Hu et al. 2018), CR-GAN (Tian et al. 2018), and PIM (Zhao et al. 2018a) treat the normalization process as a 2D image-to-image translation problem. In order to take the intrinsic 3D properties of the human face into consideration, several methods, including FF-GAN (Yin et al. 2017), 3D-PIM (Zhao et al. 2018b), UV-GAN (Deng et al. 2018) and HF-PIM (Cao et al. 2018a) make an attempt to introduce prior knowledge of 3D face, such as the 3DMM (Blanz and Vetter 1999) coefficients to supervise the frontalization process. On the other hand, feature disentangle methods have emerged to learn features invariant to some factors in both face recognition and person re-identification (Peng et al. 2017; Huang et al. 2020; Chanho and Bumsub 2019). For example, the work in (Peng et al. 2017) explicitly disentangles identity and pose features through aligning the features reconstructed across various combinations of identity and pose features, which are extracted from two faces with the same identity but different poses.

### Knowledge Distillation

In general, the main usage of knowledge distillation is to transfer the knowledge from several networks to another one (Hinton, Dean, and Vinyals 2014; Feng et al. 2020; David et al. 2016). Our work is closely related to the following very recent knowledge distillation work for recognition. Oki (Feng et al. 2020) proposes to introduce Triplet loss to knowledge distillation. The Triplet Distillation aims to reduce the distance between similar samples and increase the gap between dissimilar samples. Truong (Truong et al. 2020) proposes the method of Angular Distillation, which aims to minimize the angular distance between the outputs of the student network and teacher network. Porrello (Angelo, Luca, and Simone 2020) proposes the View Knowledge Distillation to transfer the view knowledge of persons from the teacher network to the student network, learning a robust feature representation for person re-identification.

In our work, motivated by (Angelo, Luca, and Simone 2020), we aim to transfer the knowledge of frontal face representation to the student network when dealing with profile faces. The knowledge in the teacher network is a complete and compact representation without the distortion caused by pose changes.

## Proposed Approach

### Overall Distillation Framework

As is illustrated in Fig. 1, our framework contains a teacher network $\mathcal{F}_{\theta_T}$ and a student network $\mathcal{F}_{\theta_S}$. $\theta_S$ and $\theta_T$ are the parameters of the teacher network and student network. $\mathcal{F}_{\theta_S}$, $\mathcal{F}_{\theta_T}$:$R^{W\times H\times 3} \mapsto R^D$ map each face image to a fixed-size representation. We use the average of all frontal face features of the same identity within a batch, called **frontal center**, to provide the knowledge of frontal faces. We leverage the teacher network to generate the frontal center of every identity during training. Especially, **Positive Frontal Center (PFC)** is defined for each identity: the frontal center of the same identity class. The positive frontal center can represent features extracted from different frontal samples of the same class approximately. Similarly, we define the **Negative Frontal Center (NFC)** for every identity: the frontal centers of other identity classes.

As for the student network, we feed a batch of faces with different pose variations into the student network. Each batch contains $N$ classes the same as the teacher network and each class contains $M$ faces. We denote the batch of faces as $\mathcal{X}_{N,M}$.

During the training stage, we sample a batch of near-frontal faces, including $N$ identity classes and $C$ samples per class. This batch is fed into the teacher network and $N$ frontal centers can be obtained to represent the $N$ identity classes. We denote the batch of near-frontal faces as $\mathcal{X}^F_{N\times C}$ and the frontal center can be calculated as:

$$f^C_n = \frac{1}{C}\sum_{c=1}^{C} \mathcal{F}_{\theta_T}\left(\mathcal{X}^F_{n,c}\right), n = 1, .., N \tag{1}$$

where $n$ indicates the $n_{th}$ class in the batch. For the $n_{th}$ identity class, the PFC is defined as: $f^P_n = f^C_n$, which is the frontal center with the same identity. The set of NFC of the $n_{th}$ class is defined as: $F^N_n = \{f^C_k\}_{k\neq n}$. For clarity, we use the lowercase $f$ to represent the single feature and use the uppercase $F$ to represent the feature set.

Hilton (Hinton, Dean, and Vinyals 2014) suggests that we can transfer the knowledge to the student network by training it with a soft target distribution produced from the teacher network. In our work, the PFC can be fed into the pretrained classier to obtain the soft target distribution. The Kullback-Leibler loss can be used to ask the student to mimic the PFC of the same identity.

$$\mathcal{L}_{KL} = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\left[KL\left(\widehat{y}^P_n, \widehat{y}_{n,m}\right)\right] \tag{2}$$

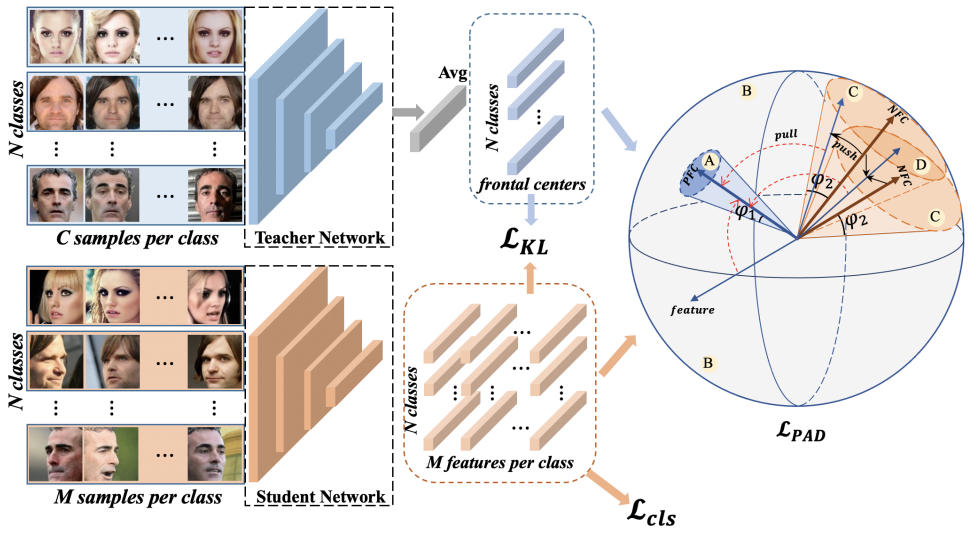


Figure 1: An overview of the proposed approach. a) The input of the teacher is a batch with $N \times C$ samples and the input of students is a batch with $N \times M$ samples. b) The supervision losses contain three parts: the classification loss $\mathcal{L}_{cls}$, $\mathcal{L}_{KL}$ and $\mathcal{L}_{PAD}$. $\mathcal{L}_{PAD}$ minimizes the angular distance between a feature and the positive frontal center(PFC) and simultaneously maximizes the angular distance between a feature and the negative frontal center(NFC).

$$\widehat{y}_n^P = SoftMax\left(\frac{h_n^P}{\tau}\right), \widehat{y}_{n,m} = SoftMax\left(\frac{h_{n,m}}{\tau}\right) \tag{3}$$

where $h_{n,m}$ and $h_n^P$ are the classifier outputs of face $\mathcal{X}_{n,m} \in \mathcal{X}_{N,M}$ and the corresponding frontal center. $\widehat{y}_n^P$ and $\widehat{y}_{n,m}$ are the smooth labels of $f_n^P$ and $f_{n,m}$, which are both smoothed by a temperature $\tau = 10$ (Angelo, Luca, and Simone 2020), respectively. $\mathcal{L}_{cls}$ is cross-entropy loss with ArcFace head (Deng et al. 2019), which supervises the student network to learn the identity feature.

$$\mathcal{L}_{cls} = -\frac{1}{NM} \cdot \sum_{n=1}^{N}\sum_{m=1}^{M} log\left(\frac{e^{s\cdot cos(\theta_{y_{n,m}}+m)}}{e^{s\cdot cos(\theta_{y_{n,m}}+m)} + \sum_{j=1,j\neq y_{n,m}}^{N} e^{s\cdot cos\theta_j}}\right) \tag{4}$$

where $y_{n,m}$ is the actual label of $f_{n,m}$ and $\theta_{y_{n,m}}$ is the angle between $f_{n,m}$ and the corresponding weight in the hyperspace. $s$ and $m$ are the feature scale and margin, respectively.

We notice that both $\mathcal{L}_{KL}$ and $\mathcal{L}_{cls}$ perform the feature alignment on the label level. To further guarantee the student network to mimic the teacher network, we further deploy a metric-learning-based loss function to supervise the knowledge transduction between the teacher and the student, which is named as Pose-Adaptive Angular Distillation Loss. In this way, the features of faces across variant poses can cluster compactly for the same person, to create a pose-invariant face representation with small intra-class distance and large inter-class distance.

## Pose-Adaptive Angular Distillation Loss

The distribution of face poses is uneven in the training dataset and most faces have small pose variations. The student network will be biased towards samples with small pose variations and cannot perform well on samples with extreme poses when trained on such a dataset.

Inspired by the Proxy-Anchor (Kim et al. 2020) loss, we propose a Pose-Adaptive Angular Distillation (PAD) loss to address this issue. The PAD loss distills the angular knowledge of frontal faces extracted by teacher network to student network, which deals with faces with different pose variations. Hence, the features of faces across variant poses cluster compactly to create a pose-invariant face representation.

### Definition

$$\mathcal{L}_{PAD} = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\{\phi\left[\alpha_{n,m} max\{d_{cos}\left(f_{n,m}, f_n^P\right) - \delta_1(n), 0\}\right] + \frac{1}{|S_n^N|}\sum_{f_n^N \in S_n^N}\phi\left[\alpha_{n,m} max\{\delta_2(n) - d_{cos}\left(f_{n,m}, f_n^N\right), 0\}\right]\} \tag{5}$$

where $\phi(x) = log\left(1 + e^x\right)$ is the Softplus function and $\alpha_{n,m}$ is the weight coefficient which is adaptive to the pose of face. $\delta_1(n)$ and $\delta_2(n)$ are the margins which are adaptive to the identity class. $S_n^N$ is the subset of the NFC set $F_n^N$, which is selected according to the distance between NFCs and the PFC. Specifically, if $|S_n^N| = K$, for the $n_{th}$ identity class, we will select $K$ NFCs from $F_n^N$, which are the first $K$ closest with the PFC $f_n^P$. Minimizing $\mathcal{L}_{PAD}$ is equivalent to minimizing $d_{cos}\left(f_{n,m}, f_n^P\right)$ when $d_{cos}\left(f_{n,m}, f_n^P\right) > \delta_1(n)$ and maximizing $d_{cos}\left(f_{n,m}, f_n^N\right)$ when $d_{cos}\left(f_{n,m}, f_n^N\right) < \delta_2(n)$.

**Geometric Interpretation** The PAD loss also has a clear geometric interpretation from the perspectives of angular and hypersphere as shown in Fig. 1. The whole space of hypersphere can also be divided into four parts. $\varphi_1(n)$ and $\varphi_2(n)$ (shown as $\varphi_1$ and $\varphi_2$ in Fig. 1) can be derived from $\delta_1(n)$ and $\delta_2(n)$ directly, where $\varphi_1(n) = arccos(1-\delta_1(n))$ and $\varphi_2(n) = arccos(1-\delta_2(n))$. This figure shows a simple case of our loss, in which $|S_n^N| = 2$.

**Part A**: The feature lies inside the cone region with $f_n^P$ as the center and $\varphi_1(n)$ as the angle range, where $d_{cos}\left(f_{n,m}, f_n^P\right) < \delta_1(n)$. Since $max\{d_{cos}\left(f_{n,m}, f_n^P\right) - \delta_1(n), 0\} = 0$ is a constant, the gradient of $\mathcal{L}_{PAD}$ at $d_{cos}\left(f_{n,m}, f_n^P\right)$ will be zero. Hence, the loss will not pull the feature embedding towards the PFC. This means if a feature $f_{n,m}$ is very close to $f_n^P$, this corresponding sample will not make contribution to the total gradient, which can reduce the weight of samples with a small pose variation.

**Part B**: In this region, $d_{cos}\left(f_{n,m}, f_n^P\right) > \delta_1(n)$ and $d_{cos}\left(f_{n,m}, f_n^N\right) > \delta_2(n)$. The feature $f_{n,m}$ will be only pulled by the PFC $f_n^P$ and will not be pushed by the NFC $f_n^N$ since $f_{n,m}$ is far enough from $f_n^N$.

**Part C**: In this region, $d_{cos}\left(f_{n,m}, f_n^P\right) > \delta_1(n)$ and $d_{cos}\left(f_{n,m}, f_n^N\right) < \delta_2(n)$. The feature only lies inside the pushing scope of one NFC. Thus, the feature $f_{n,m}$ will be pulled by the PFC $f_n^P$ and pushed by only one NFC.

**Part D**: The case is similar to Part C, but the feature lies inside the intersection of pushing scopes of the two NFCs. The feature $f_{n,m}$ will be pulled by the PFC $f_n^P$ and pushed by the two NFCs at the same time.

**Adaptive to large-pose variation and hard samples** The PAD loss is adaptive to samples with large-pose variations and hard samples. This characteristic is demonstrated by the gradient of our loss with respect to $d_{cos}\left(f_{n,m}, f_n^P\right)$ and $d_{cos}\left(f_{n,m}, f_n^N\right)$, which is given by (when $d_{cos}\left(f_{n,m}, f_n^P\right) - \delta_1(n) > 0$ and $d_{cos}\left(f_{n,m}, f_n^N\right) - \delta_2(n) < 0$)

$$\frac{\partial L_{PAD}}{\partial\left(d_{cos}\left(f_{n,m}, f_n^P\right)\right)} = \alpha_{n,m}\left(1 - \frac{1}{1 + h_+\left(f_{n,m}\right)}\right) \tag{6}$$

$$\frac{\partial L_{PAD}}{\partial\left(d_{cos}\left(f_{n,m}, f_n^N\right)\right)} = -\frac{\alpha_{n,m}}{|S_n^N|}\left(1 - \frac{1}{1 + h_-\left(f_{n,m}\right)}\right) \tag{7}$$

$$\begin{aligned} h_+\left(f_{n,m}\right) &= exp\left[\alpha_{n,m}\left(d_{cos}\left(f_{n,m}, f_n^P\right) - \delta_1(n)\right)\right] \\ h_-\left(f_{n,m}\right) &= exp\left[\alpha_{n,m}\left(\delta_2(n) - d_{cos}\left(f_{n,m}, f_n^N\right)\right)\right] \end{aligned} \tag{8}$$

$$\alpha_{n,m} = \sigma\left(\frac{4}{\pi} yaw(\mathcal{X}_{n,m}) - 1\right) \tag{9}$$

The yaw angle in radian unit can be predicted using FDN (Zhang et al. 2020) and the Equation 9 is quoted from (Cao et al. 2018b), which is a non-linear function of the yaw angle. $\sigma(.)$ is the sigmoid function and the sample with a larger yaw angle will get a larger $\alpha_{n,m}$. Hence, in Equation 6 and Equation 7, the absolute value of gradient with respect to $d_{cos}\left(f_{n,m}, f_n^P\right)$ and $d_{cos}\left(f_{n,m}, f_n^N\right)$ will become larger. This demonstrates the property of pose adaptation and the loss will pay more attention to faces with large yaw angles. Furthermore, we exploit the adaptation of hard samples. For a hard sample with the larger distance from PFC and a smaller distance from NFCs, $h_+\left(f_{n,m}\right)$ and $h_-\left(f_{n,m}\right)$ will become larger simultaneously. Hence, the absolute value of gradient respect to $d_{cos}\left(f_{n,m}, f_n^P\right)$ and $d_{cos}\left(f_{n,m}, f_n^N\right)$ will become larger at the same time, which makes the network focus on the hard samples.

**Adaptive to identity class** Furthermore, we set the margin $\delta_1(n)$ and $\delta_2(n)$ be adaptive to identity class.

$$\begin{aligned} \delta_1(n) &= \frac{\mu_1}{\sigma_P^2(n)}\pi_P(n) \\ \delta_2(n) &= \frac{\mu_2}{\pi_N(n)} \\ \pi_P(n) &= \frac{1}{M}\sum_{m=1}^{M} d_{cos}(f_{n,m}, f_n^P) \\ \pi_N(n) &= \frac{1}{|S_n^N|}\sum_{f_n^N \in S_n^N} d_{cos}(f_n^N, f_n^P) \\ \sigma_P^2(n) &= \frac{1}{M}\sum_{m=1}^{M}\left(d_{cos}(f_{n,m}, f_n^P) - \pi_P(n)\right)^2 \end{aligned} \tag{10}$$

where $\mu_1$ and $\mu_2$ are the hyper-parameters. $\pi_P(n)$ is the mean distance and $\sigma_P^2(n)$ is the variance of distance between feature embeddings and the PFC $f_n^P$ of the $n_{th}$ class in each batch. $\pi_N(n)$ is the mean distance between the PFC $f_n^P$ and the NFC $f_n^N$ in $S_n^N$.

For $\delta_1(n)$, if the variance $\sigma_P^2(n)$ is larger, the features scatter more loosely and $\delta_1(n)$ should be turned smaller to pull more features towards PFC. If the mean value of distance $\pi_P(n)$ is smaller, the features have a relatively smaller intra-class distance. The $\delta_1(n)$ should be smaller to guarantee more features to be pulled towards the PFC. For $\delta_2(n)$, if the $\pi_N(n)$ is small, it means the PFC of the $n_{th}$ class is close to the NFC $f_n^N$ in the $S_n^N$. The $\delta_2(n)$ should be larger to guarantee more features to be pushed away from the NFC.

Combined with the PAD loss, the overall loss can be written as:

$$\mathcal{L}\left(\theta_S\right) = \mathcal{L}_{cls} + \lambda_1\mathcal{L}_{KL} + \lambda_2\mathcal{L}_{PAD} \tag{11}$$

where $\lambda_1$ and $\lambda_2$ are the weights of $\mathcal{L}_{KL}$ and $\mathcal{L}_{PAD}$.

# Experiments

## Experimental Settings

We use the popular dataset MS-Celeb-1M (Guo et al. 2016) for training both teacher network and the student network. We cleaned the dataset by selecting the identity containing at least 10 near-frontal faces ($yaw\ angle \leq 10°$) and 5 non-frontal faces ($yaw\ angle > 10°$). It results in $4.6M$ images and $50.3K$ identities. For evaluation, we adopt two benchmarks for pose-invariant face recognition: CFP-FP (Sengupta et al. 2016) and IJB-A (Klare et al. 2015) datasets with

Table 1: Face verification performance (%) comparison on CFP-FP. The results are averaged over 10 testing splits.

| Methods | Acc on CFP-FP |
|---|---|
| Human | 94.57 |
| DA-GAN (Zhao et al. 2019) | 95.96 |
| PF-cpGAN (Taherkhani et al. 2020) | 93.78 |
| DR-GAN (Tran, Yin, and Liu 2017) | 93.41 |
| DREAM (Cao et al. 2018b) | 93.98 |
| PIM (Zhao et al. 2018a) | 93.10 |
| HF-PIM (Cao et al. 2018a) | 95.42 |
| TAL (Zhang et al. 2021) | 97.21 |
| Ours | **97.78** |

official evaluation protocols (Sengupta et al. 2016; Klare et al. 2015). For data pre-processing, we first resize the aligned face images to $112 \times 112$ and then linearly normalize the pixel values of RGB images to $[-1, 1]$ (Deng et al. 2019). The initial learning rate is $0.001$ and the default hyper-parameters of our method are $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, $\mu_1 = 0.01$ and $\mu_2 = 0.4$. We set $N = 20$, $C = 5$ and $M = 8$. For all the models during inference stage, we extract the 512-D feature embeddings and use cosine distance as the metric. We use 4 GeForce GTX 1080 GPUs for training and we select ResNet50, ResNet34 and ResNet18 as backbones due to the limitation of computation capacity.

## Comparison with the State-of-the-Art Methods on CFP-FP and IJB-A BenchMark

We conduct evaluations on IJB-A and CFP-FP benchmarks and make comparison with the state-of-the-art methods in the field of PIFR. We first evaluate our method on a challenging benchmark IJB-A that covers full pose variation. The dataset contains 500 subjects with of $5.7K$ images and $20.4K$ frames extracted from videos. Following the standard protocol in (Klare et al. 2015), we evaluate our method on both verification $(1 : 1)$ and identification tasks $(1 : N)$. CFP-FP is a challenging dataset created to examine the problem of frontal to profile face verification in the wild. The dataset contains 500 celebrities, each of which has 10 frontal and 4 profile face images. Evaluation systems report the mean of accuracy over the 10 splits for the frontal-profile face verification settings.

As is illustrated in Table. 2, we compare the face verification and identification performance of our proposed method with other PIFR methods on the IJB-A dataset. From Table. 2, our result with the setting of **ResNet50 + KL + PAD** significantly outperforms the performance of other PIFR approaches. In addition, we also make comparison of the face verification performance between our proposed methods and other approaches on the CFP-FP benchmark. Table. 1 demonstrates that our result with the setting **ResNet50 + KL + PAD** also has a better verification performance on the CFP-FP benchmark than other PIFR methods. Therefore, both Table. 2 and Table. 1 verify the effectiveness and superiority of our method.

## Ablation Study of the KL loss and PAD loss

As is shown in Table. 3, we conduct our ablation study using ResNet18, ResNet34 and ResNet50. We use the setting **ResNet + ArcFace** as the baseline. Firstly, we verify the validation of our teacher-student network with the Kullback-Leibler loss. In this experiment, the teacher network is introduced to obtain the frontal center for each identity and we add the *KL* loss to transfer the knowledge of frontal center to the student network. When *KL* loss is added to different backbone networks, the verification accuracy of the CFP-FP dataset is improved by $0.68\%$, $0.54\%$ and $0.14\%$ respectively. Then, we verify the validation of PAD loss by adding it to the teacher-student network. When PAD loss is added to the networks with different backbones, the verification accuracy of CFP-FP dataset is improved by $1.7\%$, $1.37\%$ and $0.96\%$ respectively.

As is shown in Table. 4, there are similar analyses on the IJB-A benchmark, and we can see that both the teacher-student network with the KL loss and the Pose-adaptive Angular Distillation loss make remarkable improvements in the performance of face verification and identification. This indicates it will help enhance the performance on profile faces by minimizing the distance between the feature embedding and the PFC and maximizing the distance between the feature embedding and the NFCs.

## Visualization of Feature Space and Statistical Analysis

We sample 9 identities from the IJB-A dataset, and faces are across different poses. The t-SNE visualization result of their feature embeddings is shown in Fig. 2. We can see that the features extracted under the setting of PAD loss cluster more compactly compared with the ones without PAD loss. It can be concluded that compared with the Baseline, the teacher-student paradigms, together with the PAD loss, can significantly reduce the intra-class distance and enlarge the inter-class distance between different identities, and this is consistent with the geometric meaning of our proposed loss.

In addition to the qualitative visual illustration of the inter-class distance and intra-class distance, we also make a quantitative statistical analysis to show the effectiveness of our teacher-student paradigm together with the PAD loss. We use the mean distance between the feature embeddings and their corresponding PFCs to measure the intra-class distance, which is defined as $\frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M} d_{cos}(f_{n,m}, f_n^P)$. $N$ is the number of sampled identity classes, and $M$ is the number of samples of each identity. We use the mean distance between different frontal centers of all sampled identity classes to measure the inter-class distance, which is defined as $\frac{1}{N(N-1)}\sum_{n=1}^{N}\sum_{m=1,m\neq n}^{N} d_{cos}(f_n^C, f_m^C)$. $d_{cos}(f_n^C, f_m^C)$ is the distance between the frontal center of the identity $n$ and the frontal center of identity $m$. As is shown in Fig. 3, after we introduce the teacher-student paradigm and add the PAD loss, the mean inter-class distance becomes larger, and the mean intra-class distance becomes smaller.

Table 2: Verification and Identification performance analysis on IJB-A benchmark. Results reported are the "average ± standard deviation" over the 10 folds specified in the IJB-A protocol

| Methods↓ | Verification | | Identification | |
|---|---|---|---|---|
| Metrics→ | TAR @ FAR=0.01 | TAR @ FAR=0.001 | Rec @ Rank-1 | Rec @ Rank-5 |
| DR-GAN (Tran, Yin, and Liu 2017) | $77.4 \pm 2.7$ | $53.9 \pm 4.3$ | $85.5 \pm 1.5$ | $94.7 \pm 1.1$ |
| FF-GAN (Yin et al. 2017) | $85.2 \pm 1.0$ | $66.3 \pm 1.3$ | $90.2 \pm 0.6$ | $95.4 \pm 0.5$ |
| DREAM (Cao et al. 2018b) | $89.1 \pm 1.6$ | $76.4 \pm 3.1$ | $94.6 \pm 1.1$ | $96.8 \pm 1.0$ |
| PIM (Zhao et al. 2018a) | $93.1 \pm 1.1$ | $87.5 \pm 1.8$ | $94.1 \pm 1.1$ | — |
| HF-PIM (Cao et al. 2018a) | $95.2 \pm 0.7$ | $89.7 \pm 1.4$ | $96.1 \pm 0.5$ | $97.9 \pm 0.2$ |
| PF-cpGAN (Taherkhani et al. 2020) | $95.8 \pm 0.8$ | $91.2 \pm 1.3$ | — | — |
| TAL (Zhang et al. 2021) | $95.8 \pm 1.2$ | $90.2 \pm 1.9$ | $96.5 \pm 1.2$ | $98.0 \pm 0.7$ |
| Ours | $\mathbf{96.4 \pm 1.1}$ | $\mathbf{91.5 \pm 1.4}$ | $\mathbf{96.9 \pm 1.3}$ | $\mathbf{98.2 \pm 0.8}$ |

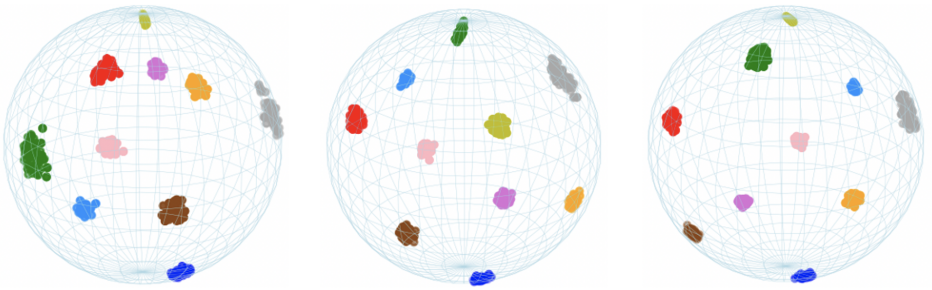


Figure 2: Feature distribution visualization of different settings

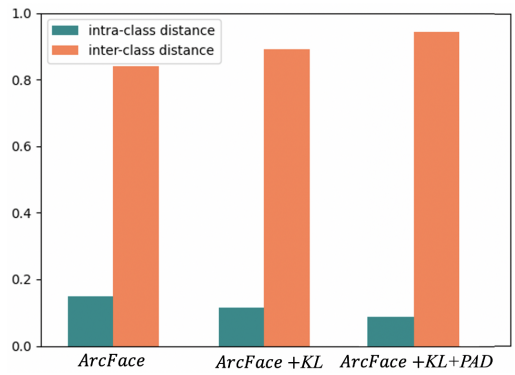


Figure 3: Statistical Analysis of intra-class distance and inter-class distance.

Table 3: Face verification performance (%) comparison on CFP-FP. The results are averaged over 10 testing splits.

| Methods | Acc on CFP-FP |
|---|---|
| ResNet-18 + ArcFace | 93.88 |
| ResNet-18 + ArcFace + *KL* | 94.52 |
| ResNet-18 + ArcFace + *KL* + *PAD* | 96.14 |
| ResNet-34 + ArcFace | 94.42 |
| ResNet-34 + ArcFace + *KL* | 94.93 |
| ResNet-34 + ArcFace + *KL* + *PAD* | 96.23 |
| ResNet-50 + ArcFace | 96.74 |
| ResNet-50 + ArcFace + *KL* | 96.85 |
| ResNet-50 + ArcFace + *KL* + *PAD* | 97.78 |

## Evaluation on CFP-FF and LFW

We conduct experiments on the CFP-FF (Sengupta et al. 2016) and LFW (Huang et al. 2008) datasets. LFW contains $13.2K$ face images of $5.7K$ identities. As shown in Table. 5, our approach does not reduce the performance on CFP-FF and LFW. In particular, the knowledge from the positive frontal center hardly affects the performance of non-profile face recognition since the frontal center encodes more compact and complete identity information, which prevents the network from losing the discriminative information. This shows that our approach does not have the loss of discriminative information, which is mentioned in (Huang et al. 2020).

## Analysis on Influences of $|S_n^N|$

From Table. 6, we notice that by increasing $|S_n^N|$ within a relatively small range, the performance can be improved. This is because more NFCs will push the feature embeddings to enlarge the inter-class distance. However, if the $|S_n^N|$ is too large, the performance will turn worse. From the geometric perspective, we think it is because too many NFCs are involved in every training step, and the influence of different NFCs may be canceled out.

Table 4: Verification and Identification performance analysis on IJB-A benchmark. Results reported are the "average ± standard deviation" over the 10 folds specified in the IJB-A protocol

| Methods↓ | Verification | | Identification | |
|---|---|---|---|---|
| Metrics→ | TAR @ FAR=0.01 | TAR @ FAR=0.001 | Rec @ Rank-1 | Rec @ Rank-5 |
| ResNet-18 + ArcFace | $93.2 \pm 2.1$ | $83.5 \pm 1.3$ | $93.8 \pm 1.6$ | $95.9 \pm 1.1$ |
| ResNet-18 + ArcFace + *KL* | $93.6 \pm 1.2$ | $86.2 \pm 1.7$ | $94.3 \pm 0.9$ | $96.8 \pm 1.6$ |
| ResNet-18 + ArcFace + *KL* + *PAD* | $95.3 \pm 1.3$ | $89.8 \pm 1.2$ | $95.5 \pm 1.6$ | $97.4 \pm 0.7$ |
| ResNet-34 + ArcFace | $93.9 \pm 0.8$ | $86.1 \pm 1.7$ | $94.2 \pm 1.8$ | $96.6 \pm 0.7$ |
| ResNet-34 + ArcFace + *KL* | $94.3 \pm 1.2$ | $87.9 \pm 1.8$ | $94.8 \pm 0.9$ | $97.0 \pm 0.6$ |
| ResNet-34 + ArcFace + *KL* + *PAD* | $95.6 \pm 0.8$ | $90.8 \pm 1.5$ | $95.9 \pm 1.1$ | $97.7 \pm 0.9$ |
| ResNet-50 + ArcFace | $94.4 \pm 0.6$ | $88.7 \pm 1.5$ | $95.7 \pm 0.8$ | $97.5 \pm 0.3$ |
| ResNet-50 + ArcFace + *KL* | $95.1 \pm 1.2$ | $89.2 \pm 1.8$ | $96.2 \pm 0.9$ | $97.8 \pm 0.6$ |
| ResNet-50 + ArcFace + *KL* + *PAD* | $96.4 \pm 1.1$ | $91.5 \pm 1.4$ | $96.9 \pm 1.3$ | $98.2 \pm 0.8$ |

Table 5: Comparative performance analysis on CFP-FF and LFW. The backbone is ResNet-50.

| settings | Acc on CFP-FF | Acc on LFW |
|---|---|---|
| ArcFace | 99.72 | 99.75 |
| ArcFace + *KL* | 99.73 | 99.75 |
| ArcFace + *KL* + *PAD* | 99.75 | 99.77 |

Table 6: Comparative analysis of different $|S_n^N|$ settings. Evaluation is conducted on CFP-FP and IJB-A. Results are reported as Acc and TAR @ FAR=0.001.

| $\|S_n^N\|$ setting | Acc on CFP-FP | TAR on IJB-A |
|---|---|---|
| $\|S_n^N\| = 1$ | 97.61 | $91.1 \pm 1.2$ |
| $\|S_n^N\| = 3$ | 97.69 | $91.4 \pm 1.1$ |
| $\|S_n^N\| = 5$ | 97.78 | $91.5 \pm 1.4$ |
| $\|S_n^N\| = 10$ | 97.66 | $91.1 \pm 1.2$ |
| $\|S_n^N\| = 15$ | 97.58 | $90.6 \pm 1.7$ |

### Analysis on Influences of $\alpha_{n,m}$

We conduct the ablation experiment on different settings of $\alpha_{n,m}$, shown in Table. 7. From Table. 7, we can know that the performance with Equation 9 is better than the one with $\alpha_{n,m} = 1$. This result shows that we can improve the performance by assigning different weights to different samples with different poses since this will eliminate the influence of the uneven distribution of poses.

### Analysis on Influences of the adaptive setting of margin $\delta_1(n)$ and $\delta_2(n)$

We conduct the ablation experiment on different margin settings, where the margin is the constant value or adaptive to the different identity classes. The adaptive setting is defined in Equation 10 and Equation 11. The comparison of the two settings is shown in Table. 8. By making the margin $\delta_1(n)$ and $\delta_2(n)$ adaptive to the identity class, the face recognition performance on CFP-FP and IJB-A benchmark can be improved. Hence, Table. 8 shows the superiority of the identity adaptive setting.

Table 7: Comparative analysis of $\alpha_{n,m}$ setting. Evaluation is conducted on CFP-FP and IJB-A. Results are reported as Acc and TAR @ FAR=0.001.

| $\alpha_{n,m}$ setting | Acc on CFP-FP | TAR on IJB-A |
|---|---|---|
| $\alpha_{n,m} = 1$ | 97.44 | $90.14 \pm 1.3$ |
| Nonlinear setting | 97.78 | $91.5 \pm 1.4$ |

Table 8: Comparative analysis of $\delta_1(n)$ and $\delta_2(n)$ setting. Evaluation is conducted on CFP-FP and IJB-A. Results are reported as Acc and TAR @ FAR=0.001.

| $\delta_1(n)$, $\delta_2(n)$ setting | Acc on CFP-FP | TAR on IJB-A |
|---|---|---|
| $\delta_1 = 0.1$, $\delta_2 = 0.4$ | 97.66 | $90.62 \pm 1.3$ |
| Adaptive Setting | 97.78 | $91.5 \pm 1.4$ |

## Conclusion

In summary, our first contribution is to propose a new perspective to learn pose-invariant feature embeddings: recovering the feature of the frontal face from the features extracted from faces of arbitrary poses with the same identity by distilling the angular knowledge of frontal faces extracted by teacher network to student network. Directed by this perspective, we effectively make features cluster more compactly around the PFC, which can create a pose-invariant and complete feature representation to enhance the performance on faces with large variations. Our second contribution is the Pose-Adaptive Angular Distillation loss. The PAD loss treats each sample with different weights according to its pose and hard level so as to overcome the limitation of the uneven distribution of pose. We achieve competitiver performance on both CFP-FP and IJB-A benchmarks, which has shown the effectiveness of our approach.

## Acknowledge

This work was partly supported by the Natural Science Foundation of China (No.62171251), the Natural Science Foundation of Guangdong Province (No.2020A1515010711), the Special Foundation for the Development of Strategic Emerging Industries

of Shenzhen (Nos. JCYJ20170817161845824 and JCYJ20200109143035495).

## References

Angelo, P.; Luca, B.; and Simone, C. 2020. Robust Re-Identification by Multiple Views Knowledge Distillation. In *European Conference on Computer Vision*.

Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces.

Cao, J.; Hu, Y.; Zhang, H.; He, R.; and Sun, Z. 2018a. Learning a High Fidelity Pose Invariant Model for High-resolution Face Frontalization. In *International Conference on Neural Information Processing Systems*.

Cao, K.; Rong, Y.; Li, C.; Tang, X.; and Loy, C. C. 2018b. Pose-Robust Face Recognition via Deep Residual Equivariant Mapping. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5187–5196.

Chanho, E.; and Bumsub, H. 2019. Learning Disentangled Representation for Robust Person Re-identification. In *International Conference on Neural Information Processing Systems*.

David, L.-P.; Léon, B.; Bernhard, S.; and Vladimir, V. 2016. Unifying distillation and privileged information. In *Proceedings of International Conference on Learning Representation*.

Deng, J.; Cheng, S.; Xue, N.; Zhou, Y.; and Zafeiriou, S. 2018. UV-GAN: Adversarial Facial UV Map Completion for Pose-Invariant Face Recognition. 7093–7102.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4685–4694.

Dovgard, R.; and Basri, R. 2004. Statistical symmetric shape from shading for 3D structure recovery of faces. In *European Conference on Computer Vision*.

Feng, Y.; Wang, H.; Hu, H. R.; Yu, L.; Wang, W.; and Wang, S. 2020. Triplet Distillation For Deep Face Recognition. In *2020 IEEE International Conference on Image Processing (ICIP)*, 808–812.

Ferrari, C.; Lisanti, G.; Berretti, S.; and Del Bimbo, A. 2016. Effective 3D based frontalization for unconstrained face recognition. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 1047–1052.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *International Conference on Neural Information Processing Systems*.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. MS-Celeb-1M: challenge of recognizing one million celebrities in the real world. *Electronic Imaging, (11):1–6*.

Hassner, T. 2013. Viewing Real-World Faces in 3D. In *2013 IEEE International Conference on Computer Vision*, 3607–3614.

Hassner, T.; Harel, S.; Paz, E.; and Enbar, R. 2015. Effective face frontalization in unconstrained images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4295–4304.

Hinton, G.; Dean, J.; and Vinyals, O. 2014. Distilling the Knowledge in a Neural Network. In *Proceedings of NIPS Deep Learning and Representation Learning Workshop*.

Hu, Y.; Wu, X.; Yu, B.; He, R.; and Sun, Z. 2018. Pose-Guided Photorealistic Face Rotation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8398–8406.

Huang, G.; Mattar, M.; Berg, T.; and Learned-Miller, E. 2008. Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. *Tech. rep.*

Huang, R.; Zhang, S.; Li, T.; and He, R. 2017. Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2458–2467.

Huang, Y.; Zha, Z.-J.; Fu, X.; Hong, R.; and Li, L. 2020. Real-World Person Re-Identification via Degradation Invariance Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14072–14082.

Kan, M.; Shan, S.; Chang, H.; and Chen, X. 2014. Stacked Progressive Auto-Encoders (SPAE) for Face Recognition Across Poses. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1883–1890.

Kan, M.; Shan, S.; Zhang, H.; Lao, S.; and Chen, X. 2016. Multi-View Discriminant Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1): 188–194.

Kim, S.; Kim, D.; Cho, M.; and Kwak, S. 2020. Proxy Anchor Loss for Deep Metric Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3235–3244.

Klare, B. F.; Klein, B.; Taborsky, E.; Blanton, A.; Cheney, J.; Allen, K.; Grother, P.; Mah, A.; Burge, M.; and Jain, A. K. 2015. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1931–1939.

Peng, X.; Yu, X.; Sohn, K.; Metaxas, D. N.; and Chandraker, M. 2017. Reconstruction-Based Disentanglement for Pose-Invariant Face Recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 1632–1641.

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 *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–9.

Taherkhani, F.; Talreja, V.; Dawson, J.; Valenti, M. C.; and Nasrabadi, N. M. 2020. PF -cpGAN: Profile to Frontal Coupled GAN for Face Recognition in the Wild. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 1–10.

Tian, Y.; Peng, X.; Zhao, L.; Zhang, S.; and Metaxas, D. N. 2018. CR-GAN: Learning complete representations for multi-view generation. In *International Joint Conferences on Artificial Intelligence*.

Tran, L.; Yin, X.; and Liu, X. 2017. Disentangled Representation Learning GAN for Pose-Invariant Face Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1283–1292.

Truong, T.-D.; Duong, C. N.; Quach, K. G.; Nguyen, D.; Le, N.; Luu, K.; and Bui, T. D. 2020. Beyond Disentangled Representations: An Attentive Angular Distillation Approach to Large-scale Lightweight Age-Invariant Face Recognition.

Yin, X.; Yu, X.; Sohn, K.; Liu, X.; and Chandraker, M. 2017. Towards Large-Pose Face Frontalization in the Wild. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 4010–4019.

Zhang, H.; Wang, M.; Liu, Y.; and Yuan, Y. 2020. FDN: Feature Decoupling Network for Head Pose Estimation.

Zhang, Z.; Chen, Y.; Yang, W.; Wang, G.; and Liao, Q. 2021. Triplet Angular Loss for Pose-Robust Face Recognition. In *2021 IEEE International Joint Conference on Neural Network(IJCNN)*.

Zhao, J.; Cheng, Y.; Xu, Y.; Xiong, L.; Li, J.; Zhao, F.; Jayashree, K.; Pranata, S.; Shen, S.; Xing, J.; Yan, S.; and Feng, J. 2018a. Towards Pose Invariant Face Recognition in the Wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2207–2216.

Zhao, J.; Xiong, L.; Cheng, Y.; Cheng, Y.; Li, J.; Zhou, L.; Xu, Y.; Karlekar, J.; Pranata, S.; Shen, S.; Xing, J.; Yan, S.; and Feng, J. 2018b. 3d-aided deep pose-invariant face recognition. In *International Joint Conferences on Artificial Intelligence*.

Zhao, J.; Xiong, L.; Li, J.; Xing, J.; Yan, S.; and Feng, J. 2019. 3D-Aided Dual-Agent GANs for Unconstrained Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10): 2380–2394.

Zhu, X.; Lei, Z.; Yan, J.; Yi, D.; and Li, S. Z. 2015. High-fidelity Pose and Expression Normalization for face recognition in the wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 787–796.