Pose Attention-Guided Profile-to-Frontal Face Recognition

Moktari Mostofa, Mohammad Saeed Ebrahimi Saadabadi, Sahar Rahimi Malakshan, and Nasser M. Nasrabadi
West Virginia University
{mm0251,me00018,sr00033}@mix.wvu.edu, nasser.nasrabadi@mail.wvu.edu

2022 IEEE International Joint Conference on Biometrics (IJCB) | 978-1-6654-6394-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCB54206.2022.10007935

Abstract

In recent years, face recognition systems have achieved exceptional success due to promising advances in deep learning architectures. However, they still fail to achieve expected accuracy when matching profile images against a gallery of frontal images. Current approaches either perform pose normalization (i.e., frontalization) or disentangle pose information for face recognition. We instead propose a new approach to utilize pose as an auxiliary information via an attention mechanism. In this paper, we hypothesize that pose attended information using an attention mechanism can guide contextual and distinctive feature extraction from profile faces, which further benefits a better representation learning in an embedded domain. To achieve this, first, we design a unified coupled profile-to-frontal face recognition network. It learns the mapping from faces to a compact embedding subspace via a class-specific contrastive loss. Second, we develop a novel pose attention block (PAB) to specially guide the pose-agnostic feature extraction from profile faces. To be more specific, PAB is designed to explicitly help the network to focus on important features along both "channel" and "spatial" dimension while learning discriminative yet pose-invariant features in an embedding subspace. To validate the effectiveness of our proposed method, we conduct experiments on both controlled and in-the-wild benchmarks including Multi-PIE, CFP, IJB-C, and show superiority over the state-of-the-arts.

1. Introduction

The advent of deep convolutional neural networks (CNNs) has led to promising achievement in unconstrained face recognition and verification techniques [36, 39]. It has even surpassed the human performance on several benchmark datasets [32]. However, a challenge that still remains to be solved, is that of extreme pose variations in profile faces. It degrade frontal-to-profile face verification accuracy by more than 10% compared to frontal-to-frontal matching accuracy [33]. The most prominent factors contributing to this performance degradation can be classified into three categories:

Figure 1: Frontal and Profile faces in the IJB-C Dataset under full pose variation, expression, and different imaging conditions.

- Facial appearance distortion: In comparison to controlled environment, real world profile faces have different imaging conditions besides pose such as expression, occlusion and illumination variations as shown in Figure 1. These variations cause substantial changes in facial appearance, which indicates loss of consistent information useful for face recognition.
- Missing semantic consistency: When a face view is changed from frontal to profile, the position and shape of facial texture varies nonlinearly. It inevitably introduces loss of semantic correspondence in 2D images along with confusion in interpersonal texture differences [10]. In consequence, extracted features from two images at different poses are no longer similar and cannot provide high matching accuracy as expected in conventional face recognition methods.
- Imbalance pose distributions: Deep learning based face recognition algorithms extensively rely on very large datasets, which usually suffer from uneven pose distributions. This data imbalance continues to force the model to lean towards frontal images than profile face image of a person. It results in poor matching accuracy during frontal-to-profile verification. In contrast, human can easily identify faces with extreme pose variations without significant drop in accuracy.

To address this performance gap between human and automatic models, traditional methods apply several local descriptors such as Gabor [9], Haar [42], and LBP [1] to measure local distortions and then adopt metric learning techniques [5, 43]. On the other hand, another research com-

munity emphasizes on frontal view synthesis across poses. They utilize 3D geometrical transformations [15, 39, 55] to reduce pose variations. Moreover, multiple novel architectures have been proposed [**?**, 28, 53] for face normalization, which aligns faces to a canonical pose. Although, they show impressive performance at normalizing small pose faces, their accuracy drops severely under extreme pose conditions. To handle this problem, some researchers opt on learning pose-robust features [2, 34, 50] for multi-view face images. Among them, Cao et al. [2] propose a lightweight DREAM block to perform "frontalization" in feature space, while others explore multi-task learning to perform pose-invariant face recognition (PIFR) [12, 50].

In this paper, we introduce a novel method to learn discriminative pose-invariant representation in a deep feature embedding subspace without performing profile face normalization (frontalization) or learning disentangled features. Instead, we explicitly deal with the pose variability by incorporating it as an "auxiliary information" via an attention mechanism to the feature extraction network. We hypothesize that learning with pose information allows for better generalization of the primary task by assisting it to focus on the current context and ignore unnecessary information. To this end, we first develop a deep coupled profile to frontal network using the contrastive loss, which is able to learn to map faces into a common compact 512-dimensional latent embedding subspace. Second, to incorporate pose as an auxiliary signal, we propose an easy-to-implement pose attention block (PAB), which automatically infers significant features from profile faces along channel and spatial axes in deeper layers of the network. In other words, PAB is designed to empirically guide to learn discriminative and pose-invariant features in an embedding subspace. Moreover, we also investigate the capability of these learned embedding features via a generative adversarial network (GAN) to synthesize a canonical (frontal) view. In a summary, this paper offers the following contributions:

- A novel coupled profile to frontal PIFR model utilizing pose as an auxiliary information (i.e., pose attention) is developed.
- A pose attention block (PAB) using a pretrained pose-estimation network is proposed to guide a discriminative and pose-invariant feature learning framework in an embedding subspace.
- Extensive experiments on different benchmark datasets and comparison to other state-of-the-art methods have been performed to validate the effectiveness of our proposed method.
- Capability of the embedding features learned in our proposed network is explored for frontal face synthesis via a GAN model, which indicates its usefulness in different face analysis tasks apart from face recognition.

2. Related Work

2.1. Face Frontalization

Face frontalization has become an extremely challenging task due to the self-occlusion that exists in 2D projections of the input face with large pose variations. To address this problem, traditional methods use 3D based models [15, 24, 55], statistical approaches [30], and deep learning based methods [21, 48, 49, 57] for face frontalization. Hassner et al. [15] used a 3D face model to generate frontal shape of all input faces. Although it is proved to be efficient in face frontalization task, it cannot achieve expected accuracy for profile and near profile faces, specifically faces with yaw angle greater than 60°. A statistical model is proposed in [30], which solves a constrained low-rank minimization problem to jointly perform frontal view reconstruction and landmark detection. Recently, deep learning based methods have shown outstanding performances in frontal face synthesis. In [48], a recurrent transform unit is proposed to reconstruct discrete 3D views. Yim et al. [49] applied a concatenated network structure to rotate a non-frontal face, where they regularize the output by image level reconstruction loss. With the emergence of GAN, researchers have concentrated more on GAN-based methods, which has advanced the performance of face frontalization methods. However, face frontalization is considered as an image-level pose-invariant representation, which can improve PIFR performance mostly for face images at near frontal or half profile.

2.2. Pose Invariant Representation Learning

Pose-invariant feature representation has been recently used as a mainstay of many face recognition tasks. Earlier works apply canonical correlation analysis (CCA) [14] to analyze the shared characteristics among pose-invariant samples. Recent deep learning based approaches focus on several aspects while training a network. To name, in [57], a deep neural network is trained to separate face identity from viewpoints. Kan et al. [22] propose feature pooling across different poses to allow a single network structure for inputs at multiple pose views. To disentangle poses in feature representation, several methods [50, 56] carefully factorize out the non-identity part. Authors in [6,25] mostly consider fusing information at the feature level or distance metric level. On the other hand, Cao et al. [2] propose a pose discrepancy corrector module. Recently many researchers [17, 41] followed them to empirically perform frontalization in feature space. Contrary to these approaches, we mostly concentrate to utilize pose as side information via an attention mechanism and guide the network to learn discriminative, and pose-invariant features in an embedding subspace.

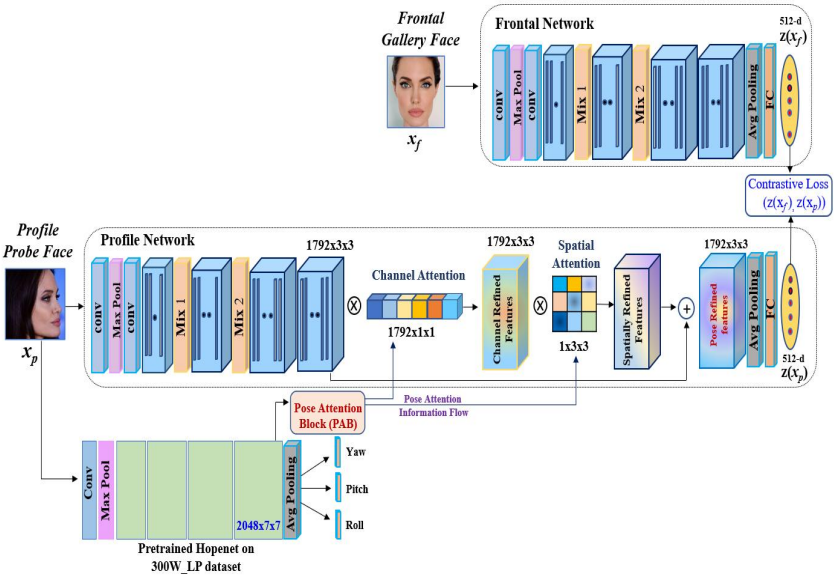

Figure 2: Block diagram of our proposed deep coupled profile-to-frontal PIFR network

3. Proposed Method

Here, we describe our proposed method which offers a new perspective of learning pose-invariant feature representation via incorporating pose specific auxiliary information into our deep profile to frontal face verification network. Inspired by the success of FaceNet face recognition system [32], we develop a deep coupled framework as shown in Figure 2. It learns mapping from both frontal and profile faces to a compact feature embedding subspace. Since profile faces have large pose variations, we use this angular knowledge to explore how auxiliary pose information can improve the embedding feature representation for profile faces. To implement this perspective, we propose the PAB module, which sequentially refine features in both channel and spatial dimension. In this section, first we discuss the implementation technique of our PAB module, and then we detail how we integrate the auxiliary pose information to our deep coupled network via a channel and spatial attention mechanism.

3.1. Pose Attention Block (PAB)

We adopt a robust pose estimation network, i.e., Hopenet [29], which has been trained on a large synthetically expanded dataset 300W-LP [54]. Hopenet uses ResNet50 as the backbone of their architecture and adds three fully connected layers to predict the intrinsic Euler angles (yaw, pitch and roll) directly from input off-angle face images as illustrated in Figure 2. To implement our proposed PAB module, we do not use these angles, instead we take the feature map of size $2048 \times 7 \times 7$ from the last convolutional layer of Hopenet. It already provides us with more complex abstract

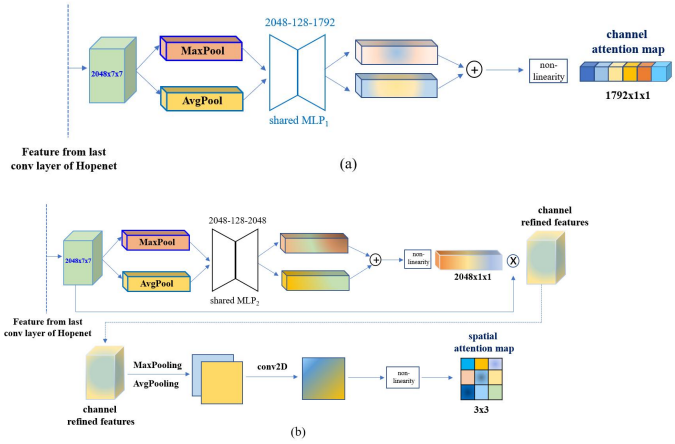

Figure 3: Block diagram of our PAB (a) Adaptive Channel attention module (ACAM) (b) Spatial Attention Module (SpAM)

features such as overall shapes, pose and texture of the input face.

Our proposed PAB module consists of two sequential attention modules: adaptive channel attention module (ACAM), and spatial attention module (SpAM), which aggregate pose information through inferring a 1D channel attention map and a 2D spatial attention map. Figure 3 illustrates the framework of the proposed PAB, which is integrated with our deep coupled learning framework in Figure 2. We now discuss each component in detail.

3.1.1 Adaptive Channel Attention Module (ACAM)

Given an input feature map, $x \in \mathbb{R}^{C \times H \times W}$ of size $2,048 \times 7 \times 7$, ACAM applies an average-pooling and a max-pooling

operation like in CBAM [45] to learn inter-channel dependencies and generates two different spatial context descriptors: x^c_{avg}, and x^c_{max}, respectively. To integrate spatial information, they are forwarded to a shared multilayer perceptron (MLP_1), which is typically a two layer fully connected network as shown in Figure 3(a). Since our ultimate goal is to distill features for the face recognition task along the channel dimension in order to highlight *"which of the feature maps are relevant to shape"*, we set hidden-layer size to 128 and output in a way that it can generate a 1D channel attention map of $1,792$ consistent with the feature map depth size in profile network. After that, we use element-wise summation to merge feature vectors obtained from the shared MLP_1 network. In a summary, the channel attention is computed as follows:

$$M_c(x) = \sigma(MLP_1(x^c_{avg}) + MLP_1(x^c_{max})), \quad (1)$$

where σ is the non-linear activation function.

3.1.2 Spatial Attention Module (SpAM)

Following CBAM [45], we design our spatial attention module (SpAM) to focus on the informative region in the spatial domain. In our SpAM, we compute a 2D spatial attention map of size 3×3 to be consistent with the feature map size of our FR network as follows:

$$F_c(x) = (\sigma(MLP_2(x^c_{avg}) + MLP_2(x^c_{max}))) \otimes x, \quad (2)$$

$$M_s(x) = \sigma(conv2D[F^s_c(x)_{avg}, F^s_c(x)_{max}]), \quad (3)$$

where $F_c(x)$ denotes channel refined features in Figure 3(b), which is obtained using another shared MLP_2. conv2D refers to the convolution on the concatenation of the average pooled feature map, $F^s_c(x)_{avg}$, and max pooled feature map, $F^s_c(x)_{max}$, along the channel axes, which ensures inter-spatial relationship of features.

3.2. Profile to Frontal Coupled Subspace Learning Network

Our goal is to learn a discriminative, pose-invariant feature representation from a pair of face images to a compact embedding subspace such that we can perform matching of profile face images with respect to a gallery of frontal face images in the embedded domain. Therefore, to learn a rich feature representation, we propose a coupled deep convolutional network guided by a pose attention module. Our method adopts a variant of InceptionResnet [37] architecture that is used in FaceNet [32] as the core element of our network. It is pretrained on VGGFace2 [3].

The model structure in Figure 2 illustrates that a pair of images goes through the coupled network consisting of two dedicated branches to extract features from both frontal and profile images. Since there exists pose variations in profile faces, we hypothesize that we can leverage pose as an auxiliary information to improve the ability of extracting highly discriminative features from these profile faces. To accomplish this, the profile image is also fed to a pretrained Hopenet pose estimation network. It provides a pose attended information via our PAB module to sequentially distills features along both channel and spatial dimension of our profile encoder.

Previous section explains the block design of our PAB module. It consists of two sub-modules: (1) ACAM, that generates a 1D channel attention map of $1,792$ to refine the feature maps ($1,792 \times 3 \times 3$) of our profile coupled network along the channel dimension, and (2) SpAM, which produces a 2D spatial attention map of size 3x3 to spatially attend the informative region in the feature maps of our profile network as in Figure 2. For more details note that, both 1D and 2D attention maps are multiplied with the feature maps of the profile network for adaptive feature refinement.

Such sharing of pose as an auxiliary information during feature extraction from profile faces results in informative and task relevant features, which otherwise would not have been attained from training only with a massive labelled faces. In addition, it also allows for better generalization of the PIFR task by looking at new interpretations of the features. Once the embeddings are established as feature vectors, we optimize the network via class-specific contrastive loss. It tries to minimize squared Euclidean distance between the features of positive pairs (i.e., when profile and frontal image share the same identity) and maximize it for negative pairs (i.e., when profile and frontal image comes from different identities).

4. Loss Function

Our goal is to learn a compact 512-D embedding subspace by coupling two mapping networks, one for frontal and another one for profile face image, via contrastive loss, L_{cont} [7]. We compute this loss metric, L_{cont} over a set of genuine (i.e., a profile face image of a subject with its corresponding frontal face image) and imposter (i.e., a profile face image of a subject and a frontal face image of a different subject) pairs such that images belonging to the same identity (genuine pair) are embedded as close as possible. Simultaneously, images of different identities are pushed away from each other in the common embedded subspace. The contrastive loss function is formulated as:

$$L_{cont}(z(x^i_p), z(x^j_f), Y) = \\ (1-Y)\frac{1}{2}(D_z)^2 + (Y)\frac{1}{2}(\max(0, m - D_z))^2, \quad (4)$$

where x^i_p and x^j_f denote the input profile and frontal face images, respectively. The variable Y is a binary label, which is equal to 0 if x^i_p and x^j_f belong to the same class

(i.e., genuine pair), and equal to 1 if x_p^i and x_f^j belong to the different class (i.e., impostor pair). $z(.)$ is used to denote the mapping function for x_p^i and x_f^j into a compact embedding subspace. To "tighten" the constraint, m is used as contrastive margin.

The Euclidean distance, D_z, between the embedding features, $z(x_p^i)$ and $z(x_f^j)$, is given by:

$$D_z = \left\| z(x_p^i) - z(x_f^j) \right\|_2 . \tag{5}$$

Therefore, if $Y = 0$ (i.e., genuine pair), then the contrastive loss function (L_{cont}) is given as:

$$L_{cont}(z(x_p^i), z(x_f^j), Y) = \frac{1}{2} \left\| z(x_p^i) - z(x_f^j) \right\|_2^2, \tag{6}$$

and if $Y = 1$ (i.e., impostor pair), then contrastive loss function (L_{cont}) is :

$$L_{cont}(z(x_p^i), z(x_f^j), Y) = \frac{1}{2} \max\left(0, m - \left\| z(x_p^i) - z(x_f^j) \right\|_2^2\right). \tag{7}$$

Thus, the total loss to optimize the entire network is denoted by L_{total} for coupling both the profile and frontal face in the embedded domain:

$$L_{total} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} L_{cont}(z(x_p^i), z(x_f^j), Y), \tag{8}$$

where *N* is the number of training samples. The main purpose of using the contrastive loss is to be able to use the class labels, and margin to ensure discriminative embedding subspace, which may not be obtained with some other metric such as the Euclidean distance. Finally, we use this pose attended discriminative embedding subspace for matching of the profile images with the frontal images.

5. Experiments

In this section, we describe our implementation details and the datasets that we have used to conduct our experiments. To evaluate the performance of our proposed PIFR network, we experiment under two settings: (1) face identification on controlled Multi-PIE [13] face dataset, and (2) face verification/identification on in-the-wild datasets including CFP [33] and IJB-C [26] with their official evaluation protocols. In addition, we also report face recognition accuracy compared to several state-of-the-art results on these datasets.

5.1. Datasets

Multi-PIE: The Multi-PIE dataset is the largest dataset released for multi-view face recognition with respect to controlled variations in illumination and expressions across different poses. It contains 7,54,204 images of 337 identities, captured at 15 view points ranging from $-90^\circ \sim +90^\circ$, over 20 illumination conditions. To evaluate our proposed method for identification task, we conduct experiments under two settings following protocol used in [53] for fair comparison. **Setting 1** includes images only from session 1 in the Multi-PIE dataset, which has 250 subjects. For training, we choose first 150 identities with 11 poses within $\pm 90^\circ$ and 20 illuminations. During testing, one frontal view with neutral expression and illumination (i.e., ID 07) is used in the gallery for each of the remaining 100 identities and other images are considered as probe images. **Setting 2** includes images only with neutral expression over all four sessions providing 337 identities. To train our network, we use first 200 identities, while rest of the 137 IDs have been used for testing. We maintain similar setup as setting 1 for our gallery and probe. **CFP**: The Celebrities in Frontal-Profile (CFP) dataset is introduced to handle large-pose variations. It contains identities of 500 celebrities, which have been collected under constrained (i.e., images at different pose, illumination and expression) and unconstrained (i.e., images collected from the Internet) settings. For each celebrity, it includes 10 frontal and 4 profile images. Following their standard 10-fold evaluation protocol [33], we split the dataset into 10 folds, each with 350 genuine and 350 impostor pairs to perform both frontal-to-frontal (FF), and frontal-to-profile (FP) verification task. **IJB-C**: The IARPA Janus Benchmark–C (IJB-C) [26] face dataset has been released to advance the unconstrained face recognition by modeling more practical face recognition use cases. It is an extension to the publicly available IJB-B [44] dataset, which contains $3,531$ subjects with extreme variations in expression, illumination, geographic origin, and more. In total, it has $31,334$ still images and $1,17,542$ video frames collected in unconstrained settings with different protocols. To evaluate our algorithm's ability, we perform both face verification (1:1), and identification (1:N) tasks following their protocol.

5.2. Implementation Details

To implement our proposed coupled learning framework, we have used InceptionResnet-v1 [37] pretrained on VGG-Face2 dataset. Since it is difficult to train the entire network from scratch, we freeze all the trained layers before average pooling for both frontal and profile mapping modules as shown in Figure 2. At the same time, our PAB module provides a pose attended 1D channel attention map, and a 2D spatial attention map to assist the profile network to use only the relevant features while extracting deep features from the profile faces. Therefore, the gradient also flows back through this PAB module to update its weights during optimization. Note that, since misleading pose information can misguide the training, we don't train Hopenet, which has been already trained on a very large dataset, and proved

Table 1: Rank-1 recognition rates (%) across poses and illuminations under Multi-PIE Setting-1.

Method	±90°	±75°	±60°	±45°	±30°	±15°
HPN [11]	29.8	47.5	61.2	72.7	78.2	84.2
c-CNN [47]	47.2	60.7	74.4	89.0	94.1	97.0
TP-GAN [18]	64.0	84.1	92.9	98.6	99.99	99.8
PIM [53]	75.0	91.2	97.7	98.3	99.4	99.8
CAPG-GAN [16]	77.1	87.4	93.7	98.3	99.4	99.9
FNM+VGG-Face [28]	41.1	67.3	83.6	93.6	97.2	99.0
FNM+Light CNN [28]	55.8	81.3	93.7	98.2	99.5	99.9
PF-cpGAN [38]	88.1	94.2	97.6	98.9	99.9	99.9
Backbone(without attention)	75.68	98.20	**100.0**	**100.0**	**100.0**	**100.0**
Ours	**89.5**	**98.7**	**100.0**	**100.0**	**100.0**	**100.0**

to be an efficient pose estimation model. The entire framework has been implemented in Pytorch. We used a batch size of 32 and the Adam optimizer [23] with first-order momentum of 0.5, and learning rate of 10^{-3}. For training, we generate same number of genuine, and impostor pairs from frontal, and profile images of the same/different subjects to avoid biasness towards positive pairs.

5.3. Evaluations on the Multi-PIE Benchmark

To show the effectiveness of our proposed method, first we evaluate our model on a controlled database, Multi-PIE for profile-to-frontal face recognition task under two different settings. We compare our method with several state-of-the-art PIFR algorithms including HPN [11], c-CNN [47], PIM [53], FNM [53], and competitive GAN-based methods : TP-GAN [18], CAPG-GAN [16], and PF-cpGAN [38].

Table 1 shows our rank-1 recognition accuracy compared to other approaches across full yaw variations and illuminations under setting-1. For this experimental setup, we consistently achieve 100% accuracy over yaw angles $< 75°$, while outperforming other baselines. Even under extreme pose (i.e., $\pm 75°$, and $\pm 90°$), when compared to CAPG-GAN, and PF-cpGAN, we significantly outperform them by achieving average 11.85%, and 3% higher accuracy, respectively. Compared to the performance of our backbone coupled network (without attention), we achieve **13.82%** more recognition accuracy for full profile face ($\pm 90°$). We also assess the performance of our proposed network on faces in Multi-PIE under setting-2, which consists of more challenging face identities than setting-1. Evaluation results, shown in Table 2 suggests that our proposed PAB module does assist the face recognition network to achieve 2.3%, and 2.7% increase over the best performing method, PIM [53] in the large pose variations; $\pm 90°$, and $\pm 75°$, respectively. Apart from this, our network achieves superior performance over the other baseline models [18,51,53] in all yaw angles. Similar to setting 1, we note **14.0%** improvement on backbone (without attention) for $\pm 90°$ profile faces as well. These improvements indicate the efficacy of our method for PIFR in a constrained environment.

Table 2: Rank-1 recognition rates (%) across poses and illuminations under Multi-PIE Setting-2.

Method	±90°	±75°	±60°	±45°	±30°	±15°
FF-GAN [51]	61.2	77.2	85.2	89.7	92.5	94.6
TP-GAN [18]	64.6	77.4	87.7	95.4	98.0	98.6
CAPG-GAN [16]	66.0	83.05	90.6	97.3	99.5	99.8
DA-GAN [52]	81.5	93.2	97.2	99.1	99.8	99.9
PIM [53]	86.5	95.0	98.1	98.3	98.5	99.0
Backbone (without attention)	74.8	96.8	**100.0**	**100.0**	**100.0**	**100.0**
Ours	**88.8**	**97.7**	**100.0**	**100.0**	**100.0**	**100.0**

Table 3: Performance comparison on CFP dataset. Mean Accuracy and equal error rate (EER) with standard deviation over 10 folds.

	Frontal-Profile (FP)		Frontal-Frontal (FF)	
Algorithm	Accuracy	EER	Accuracy	EER
FV+DML [33]	58.47(3.51)	38.54(1.59)	91.18(1.34)	8.62(1.19)
LBP+Sub-SML [33]	70.02(2.14)	29.60(2.11)	77.98(1.86)	16.00(1.74)
HoG+Sub-SML [33]	77.31(1.61)	22.20(1.18)	85.97(1.03)	11.45(1.35)
FV+Sub-SML [33]	80.63(2.12)	19.28(1.60)	88.53(1.58)	8.85(0.74)
Deep Features [33]	84.91(1.82)	14.97(1.98)	93.00(1.55)	3.48(0.67)
Triplet Embedding [31]	89.17(2.35)	8.85(0.99)	98.88(1.56)	2.51(0.81)
Light CNN-29 [46]	92.47(1.44)	8.71(1.80)	99.64(0.32)	0.57(0.40)
PIM (Light CNN-29) [46]	93.10(1.01)	7.69(1.29)	99.44(0.36)	0.86(0.49)
PR-REM [2]	93.25(2.23)	7.92(0.98)	98.10(2.19)	1.10(0.22)
PF-cpGAN [38]	93.78(2.46)	7.21(0.65)	98.88(1.56)	0.93(0.14)
Backbone (without attention)	92.57(1.10)	4.24(0.54)	97.10(0.11)	1.5(0.25)
Ours	**95.67(1.64)**	**2.02**(0.62)	**99.70(0.21)**	**0.55(0.35)**
Human [33]	94.57(1.10)	5.02(1.07)	96.24(0.67)	5.34(1.79)

5.4. Evaluations on the CFP Benchmark

We evaluate our proposed method on the Celebrities in Frontal-Profile (CFP) dataset to analysis face verification in unconstrained environment. To perform evaluation, we follow the standard 10-fold protocol like other approaches in the literature. We report the mean and standard deviation of accuracy, and Equal Error Rate (EER) over the 10 splits for both frontal-frontal (FF) and frontal-profile (FP) face verification settings.

Table 3 shows a comparison of our method with other state-of-the-art face recognition performance on the CFP benchmark datatset. For fair comparison, first, we consider three different hand-crafted feature extraction techniques: Hog [8], LBP [1], and Fisher Vector [35] along with metric learning techniques Sub-SML [4], and diagonal metric learning (DML) [4]. To compare against deep learning based approaches, we include Deep Features [6], Triplet Embedding [31], Light CNN-29 [46], and recently proposed GAN-based latent feature learning framework, PF-cpGAN [38].

From the results summarized in Table 3, we observe that our proposed method outperforms human performance for both FF and FP settings. It also makes substantial improvement over the conventional hand-crafted features by achieving average 18% higher accuracy with 24% decrease in EER for more challenging FP setting. In addition, when compared to best performing PF-cpGAN, our proposed method improves the accuracy by 1.89% and reduce EER significantly by 5% for FP verification. We also improve on the

Table 4: Performance evaluation on IJB-C benchmark. Symbol '-' indicates that the metric is not available for that protocol.

Method	1:1 Verification		1:N Identification	
	GAR@ FAR= 0.01	GAR@ FAR= 0.001	@ Rank-1	@ Rank-5
GOTS [26]	61.99	33.4	38.5	53.8
FaceNet [32]	81.76	66.45	69.22	78.7
VGGFace [27]	87.13	74.79	78.60	87.2
CFR-GAN [20]	86.46	74.81	-	-
FNM [28]	91.2	80.4	78.6	88.7
PR-REM [2]	90.6	80.2	77.1	87.6
Backbone(without attention)	89.1	79.9	71.8	81.2
Ours	**92.8**	**82.5**	**80.33**	**90.42**

performance of PR-REM [2] by 2.5% higher accuracy with approximately 6% lower EER. Moreover, we obtain 3.10% better accuracy than the backbone network where we do not apply any attention.

5.5. Evaluations on the IJB-C Benchmark

We further evaluate face recognition (i.e., verification and identification) on another challenging benchmark IJB-C, to validate the superiority of our proposed method in unconstrained environment. We compare with the recent state-of-the-art algorithms CFR-GAN [20], FNM [28], and PR-REM [2], along with prior works [27,32] in [26] for fair evaluation. As shown in Table 4, for profile to frontal verification, we improve the genuine accept rate (GAR) by approximately 7.69%, and 2.1% at the false accept rate (FAR) of 0.001 compared to recent works [20, 28]. Moreover, we also obtain outstanding performances on identification. Specifically, we achieve 1.73%, and 3.23% higher recognition accuracy for rank-1 in comparison to the FNM, and PR-REM, respectively. To show the significant contribution of our proposed PAB module, we compare our results with the backbone network (without attention). It shows our idea of incorporating pose information boosts FP verification performance by 2.60% at 0.001 FAR, and identification accuracy by 8.53% for rank-1.

5.6. Frontal Face Reconstruction from Pose-Invariant Features Learned in Deep Subspace

The purpose of our proposed PAB is to enhance the recognition performance of our coupled deep subspace learning framework via contributing in feature refinement. In addition, class-specific contrastive loss has been used to push the network achieve pose-invariance in the embedding feature domain. To validate our hypothesis, previous sections show comprehensive analysis on verification and identification task for both constrained and in the wild conditions. Apart from recognition task, there are many other scopes to utilize the feature vector learned in the deep subspace. For instance, if we could reconstruct frontal face from the deep features of its corresponding profile face, it can be used in many face analysis tasks including emotion detection, expression tracking etc. Moreover, it has broad applications in vision, graphics, and robotics. Therefore,

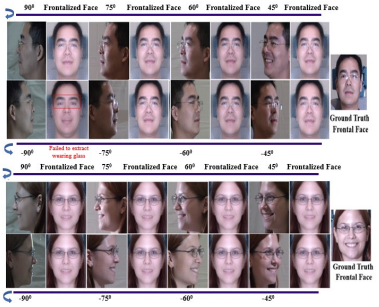

Figure 4: Reconstruction results via a GAN [40] model on Multi-PIE across different pose, illumination and expression using the compact 512-D embedding features learned using our proposed network.

Table 5: Rank-1 recognition rates (%) on Multi-PIE (Setting-1) for different embedding dimensions. We select 512-D for all experiments reported in this paper.

Dims	$\pm 90°$	$\pm 75°$
128	60.8	67.5
256	70.2	78.1
512	**89.5**	**98.7**

to demonstrate the usefulness of our proposed method, we adopt a GAN model [40] for reconstruction. To accomplish this task, we modify residual network used in the two-pathway encode-decoder architecture proposed by Tian et al. [40]. We consider their decoder module with the discriminator network for frontal face synthesis. For training/testing, we select profile and corresponding frontal images from setting 1 of Multi-PIE dataset. First, we extract 512-D embedding feature vector from our proposed PAB guided pose-invariant face recognition network for each of the profile image in the trainset. After that, these profile feature vectors are given as input to the decoder and corresponding frontal faces with no expression and neutral illumination are used as target, which force the network adversarially learn the image distribution of the frontal faces. To generate identity preserving, high visual quality frontal faces from its profile deep features, we incorporate pixel-wise L_1 reconstruction error, VGG-16 based Perceptual loss [19], and Light CNN-29 [46] network for identification loss along with adversarial loss. In Figure 4, we show some representative results on Multi-PIE test samples. Reconstruction results indicate that our proposed pose attention-guided coupled framework is able to provide robust, and discriminative features in the deep subspace for multiple use of profile to frontal matching in the embedded domain as well as high-fidelity frontal face synthesis.

6. Ablation Study

6.1. Embedding Dimensionality

To represent each face into a tightly compact embedding subspace, we explore different embedding dimensionalities:

Table 6: Rank-1 recognition rates (%) on Multi-PIE (Setting-1) for different approaches of spatial attention

Description	$\pm 90^{\circ}$	$\pm 75^{\circ}$
Channel Refined Features + 1×1 conv + Max Pool	87.8	94.5
Channel Refined Features + 3×3 conv + Stride-2	**89.5**	**98.7**

Table 7: Rank-1 recognition rates (%) on Multi-PIE (Setting-1) for different arrangements in attention mechanism

Description	$\pm 90^{\circ}$	$\pm 75^{\circ}$
InceptionResnet + spatial + channel	87.2	95.6
InceptionResnet + channel + spatial	**89.5**	**98.7**

128, 256, and 512. Experimental results reported in Table 5 illustrates that the network is able to extract features enriched with relevant information in 512 dimension.

6.2. Attention Map

In this section, we show the effective design approach of our proposed pose attention mechanism to efficiently guide the face recognition network. We first focus on computing different approaches of SpAM in the pose attention block, PAB. Finally, we observe different combination of channel and spatial attention in deep profile feature extraction. Each experiment has been explained in the following sections.

6.2.1 Spatial Attention

Given the channel-wise refined features, we explore two different approaches to generate a 2D spatial attention map: (1) first, we use average-and max-pooling across the channel axes, to generate two 2D descriptors, and then apply standard 1×1 convolution followed by a max pool layer. (2) second, we similarly generate two 2D descriptors, and apply 3×3 convolution with stride 2, which proves to be outperforming the first approach. We report the comparison of two methods in Table 6.

6.2.2 Arrangement of Spatial and Channel Attention

In this experiment, we apply channel and spatial attention in two different ways. From a spatial viewpoint, the channel attention works to attend global information whereas the spatial attention focuses on local neighbourhood. However, the network response can be different upon the sequential order of each attention mechanism. Table 7 summarizes the recognition performance on Multi-PIE for different attention sequences. The results show that the we achieve better performance when we use channel-spatial order rather than the vise versa.

6.3. Visualization

As shown in Figure 5, when compared to the backbone network (without attention), the similarity distributions of the genuine pairs and the imposter pairs in our proposed coupled PIFR network are more compact and distinct for

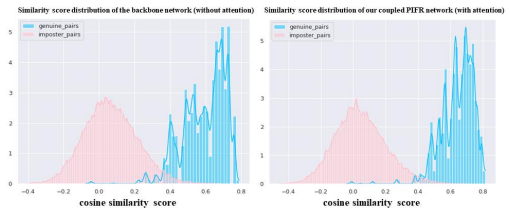

Figure 5: Comparing Cosine similarity distributions of the genuine pairs and imposter pairs for full profile faces ($\pm 90^{\circ}$) of Multi-Pie Setting 1 between the backbone network (without attention) and our coupled PIFR network (with attention)

full profile variations ($\pm 90^{\circ}$). Moreover, the area of similarity between genuine pairs spread more and overlap with the area of imposter pairs when we only train backbone network with constrastive loss without imposing attention on it. It further supports our proposed idea of pose refinement via PAB attention module.

7. Conclusion

In this paper, we propose a novel perspective of leveraging pose as auxiliary information to guide a coupled profile to frontal deep subspace learning framework for PIFR. A PAB module is designed to distill pose-specific useful features from profile faces in deep convolutional layers. To ensure discriminative, pose-invariant feature representation into a compact embedding subspace, we couple both profile and frontal face images via a contrastive loss, which maximizes the pair-wise similarity in the embedded domain. We perform a comprehensive experiments on several benchmark datasets both in controlled and uncontrolled environmental settings to evaluate the robustness of our model. The results indicate that our model remarkably outperform other state-of-the-art algorithms for profile-to-frontal pose-invariant face recognition. In addition, we conduct a quick experiment to explore the generative capability of the embedding features learned in deep subspace of our network. Moreover, we also investigate embedding dimensionality and attention mechanisms from different perspectives to offer an effective design choice of our proposed network.

8. Acknowledgements

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2022-21102100001. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041, 2006.

[2] K. Cao, Y. Rong, C. Li, X. Tang, and C. C. Loy. Pose-robust face recognition via deep residual equivariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5187–5196, 2018.

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[4] Q. Cao, Y. Ying, and P. Li. Similarity metric learning for face recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2408–2415, 2013.

[5] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3025–3032, 2013.

[6] J.-C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016.

[7] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

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

[10] C. Ding and D. Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on intelligent systems and technology (TIST)*, 7(3):1–42, 2016.

[11] C. Ding and D. Tao. Pose-invariant face recognition with homography-based normalization. *Pattern Recognition*, 66:144–152, 2017.

[12] C. Ding, C. Xu, and D. Tao. Multi-task pose-invariant face recognition. *IEEE Transactions on Image Processing*, 24(3):980–993, 2015.

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and vision computing*, 28(5):807–813, 2010.

[14] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[15] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4295–4304, 2015.

[16] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun. Pose-guided photorealistic face rotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8398–8406, 2018.

[17] J. Huang and C. Ding. Attention-guided progressive mapping for profile face recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.

[18] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE international conference on computer vision*, pages 2439–2448, 2017.

[19] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[20] Y.-J. Ju, G.-H. Lee, J.-H. Hong, and S.-W. Lee. Complete face recovery gan: Unsupervised joint face rotation and de-occlusion from a single-view image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3711–3721, 2022.

[21] M. Kan, S. Shan, H. Chang, and X. Chen. Stacked progressive auto-encoders (SPAE) for face recognition across poses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1883–1890, 2014.

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

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[24] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan. Morphable displacement field based image matching for face recognition across pose. In *European conference on computer vision*, pages 102–115. Springer, 2012.

[25] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4838–4846, June 2016.

[26] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, et al. Iarpa Janus benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE, 2018.

[27] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 2015.

[28] Y. Qian, W. Deng, and J. Hu. Unsupervised face normalization with extreme pose and expression in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9851–9858, 2019.

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

[30] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Robust statistical face frontalization. In *Proceedings of the IEEE international conference on computer vision*, pages 3871–3879, 2015.

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

[32] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[33] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016.

[34] Y. Shi, X. Yu, K. Sohn, M. Chandraker, and A. K. Jain. Towards universal representation learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6817–6826, 2020.

[35] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *BMVC*, volume 2, page 4, 2013.

[36] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1891–1898, 2014.

[37] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, Inception-Resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[38] F. Taherkhani, V. Talreja, J. Dawson, M. C. Valenti, and N. M. Nasrabadi. PF-cpGAN: Profile to frontal coupled GAN for face recognition in the wild. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020.

[39] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[40] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas. CR-GAN: learning complete representations for multi-view generation. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence Main track*, pages 942–948, 2018.

[41] E.-J. Tsai and W.-C. Yeh. PAM: Pose attention module for pose-invariant face recognition. *arXiv preprint arXiv:2111.11940*, 2021.

[42] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[43] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

[44] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, et al. Iarpa Janus benchmark-b face dataset. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 90–98, 2017.

[45] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. CBAM: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

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

[47] C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan, and T.-K. Kim. Conditional convolutional neural network for modality-aware face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3667–3675, 2015.

[48] J. Yang, S. E. Reed, M.-H. Yang, and H. Lee. Weakly-supervised disentangling with recurrent transformations for 3D view synthesis. In *Advances in neural information processing systems*, pages 1099–1107, 2015.

[49] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim. Rotating your face using multi-task deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 676–684, 2015.

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

[51] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3990–3999, 2017.

[52] Y. Yin, S. Jiang, J. P. Robinson, and Y. Fu. Dual-attention GAN for large-pose face frontalization. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 249–256. IEEE, 2020.

[53] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, et al. Towards pose invariant face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2207–2216, 2018.

[54] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3D solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.

[55] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 787–796, 2015.

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

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