

Reconstruction-Based Disentanglement for Pose-invariant Face Recognition

Xi Peng[†], Xiang Yu[‡], Kihyuk Sohn[‡], Dimitris N. Metaxas[†] and Manmohan Chandraker^{§‡}

[†]Rutgers, The State University of New Jersey

[§]University of California, San Diego

[‡] NEC Laboratories America

Fxi peng.cs, dnmG@rutgers.edu, Fxi angyu, ksohn, manuG@nec-labs.com

Abstract

*Deep neural networks (DNNs) trained on large-scale datasets have recently achieved impressive improvements in face recognition. But a persistent challenge remains to develop methods capable of handling large pose variations that are relatively under-represented in training data. This paper presents a method for learning a feature representation that is invariant to pose, without requiring extensive pose coverage in training data. We first propose to generate non-frontal views from a single frontal face, in order to increase the diversity of training data while preserving accurate facial details that are critical for identity discrimination. Our next contribution is to seek a rich embedding that encodes identity features, as well as non-identity ones such as pose and landmark locations. Finally, we propose a new feature reconstruction metric learning to explicitly disentangle identity and pose, by demanding alignment between the feature reconstructions through various combinations of identity and pose features, which is obtained from two images of the same subject. Experiments on both controlled and in-the-wild face datasets, such as MultiPIE, 300WLP and the profile view database CFP, show that our method consistently outperforms the state-of-the-art, especially on images with large head pose variations.*¹

1. Introduction

The human visual system is commendable at recognition across variations in pose, for which two theoretical constructs are preferred. The first postulates invariance based on familiarity where separate view-specific visual representations or templates are learned [6, 26]. The second suggests that structural descriptions are learned from images that specify relations among viewpoint-invariant primitives [10]. Analogously, pose-invariance for face recognition in

Figure 1. (a) Generic data-driven features for face recognition might confound images of the same identity under large poses with other identities, as shown two subjects (in different colors) from MultiPIE are mapped into the learned feature space of VGGFace [22]. (b) We propose a feature reconstruction metric learning to disentangle identity and pose information in the latent feature space. (c) The disentangled feature space encourages identity features of the same subject to be clustered together despite of the pose variation.

computer vision also falls into two such categories.

The use of powerful deep neural networks (DNNs) [15] has led to dramatic improvements in recognition accuracy. However, for objects such as faces where minute discrimination is required among a large number of identities, a straightforward implementation is still ineffective when faced with factors of variation such as pose changes [24]. Consider the feature space of the VGGFace [22] evaluated on MultiPIE [7] shown in Figure 1, where examples from the same identity class that differ in pose are mapped to distant regions of the feature space. An avenue to address this is by increasing the pose variation in training data. For instance, 4.4 million face images are used to train DeepFace [39] and 200 million labelled faces for FaceNet [32]. Another approach is to learn

This work was part of the Xi's internship at NEC Laboratories America.

¹Detail results and resource are referred to: <https://sites.google.com/site/xipengcshomepage/iccv2017>.

a mapping from different view-specific feature spaces to a common feature space through methods such as Canonical Correlation Analysis (CCA) [8]. Yet another direction is to ensemble over view-specific recognition modules that approximate the non-linear pose manifold with locally linear intervals [20, 12].

There are several drawbacks for the above class of approaches. First, conventional datasets including those sourced from the Internet have long-tailed pose distributions [19]. Thus, it is expensive to collect and label data that provides good coverage for all subjects. Second, there are applications for recognition across pose changes where the dataset does not contain such variations, for instance, recognizing an individual in surveillance videos against a dataset of photographs from identification documents. Third, the learned feature space does not provide insights since factors of variation such as identity and pose might still be entangled. Besides the above limitations, view-specific or multiview methods require extra pose information or images under multiple poses at test time, which may not be available.

In contrast, we propose to learn a novel reconstruction based feature representation that is invariant to pose and does not require extensive pose coverage in training data. A challenge with pose-invariant representations is that discrimination power of the learned feature is harder to preserve, which we overcome with our holistic approach. First, inspired by [50], Section 3.1 proposes to enhance the diversity of training data with images under various poses (along with pose labels), at no additional labeling expense, by designing a face generation network. But unlike [50] which frontalizes non-frontal faces, we *generate rich pose variations* from frontal examples, which leads to advantages in better preservation of details and enrichment rather than normalization of within-subject variations. Next, to achieve a rich feature embedding with good discrimination power, Section 3.2 presents a joint learning framework for identification, pose estimation and landmark localization. By jointly optimizing those three tasks, a *rich feature embedding* including both identity and non-identity information is learned. But this learned feature is still not guaranteed to be pose-invariant.

To achieve pose invariance, Section 3.3 proposes a feature reconstruction-based structure to explicitly *disentangle identity and non-identity* components of the learned feature. The network accepts a reference face image in frontal pose and another image under pose variation and extracts features corresponding to the rich embedding learned above. Then, it minimizes the error between two types of reconstructions in feature space. The first is *self-reconstruction*, where the reference sample’s identity feature is combined with its non-identity feature and the second is *cross-reconstruction*, where the reference sample’s non-identity feature is combined with the pose-variant sample’s identity feature. This encourages the network to regularize the pose-variant sample’s identity feature to be close to that of the reference sam-

ple. Thus, non-identity information is distilled away, leaving a disentangled identity representation for recognition at test.

Section 5 demonstrates the significant advantages of our approach on both controlled datasets and uncontrolled ones for recognition in-the-wild, especially on 90 cases. In particular, we achieve strong improvements over state-of-the-art methods on 300-WLP, MultiPIE, and CFP datasets. These improvements become increasingly significant as we consider performance under larger pose variations. We also present ablative studies to demonstrate the utility of each component in our framework, namely pose-variant face generation, rich feature embedding and disentanglement by feature reconstruction.

To summarize, our key contributions are:

- To the best of our knowledge, we are the first to propose a novel reconstruction-based feature learning that disentangles factors of variation such as identity and pose.
- A comprehensively designed framework cascading rich feature embedding with the feature reconstruction, achieving pose-invariance in face recognition.
- A generation approach to enrich the diversity of training data, without incurring the expense of labeling large datasets spanning pose variations.
- Strong performance on both controlled and uncontrolled datasets, especially for large pose variations up to 90°.

2. Related Work

While face recognition is an extensively studied area, we provide a brief overview of works most relevant to ours.

Face synthesization Blanz and Vetter pioneered 3D morphable models (3DMM) for high quality face reconstruction [2] and recently, blend shape-based techniques have achieved real-time rates [3]. For face recognition, such techniques are introduced in DeepFace [39], where face frontalization is used for enhancing face recognition performance. As an independent application, specific frontalization techniques have also been proposed [9]. Another line of work pertains to 3D face reconstruction from photo collections [29, 18, 42] or a single image [19, 50, 40], where the latter have been successfully used for face normalization prior to recognition. While most of the methods apply the framework of aligning 3DMM with the 2D face landmarks [47, 46, 25] and conduct further refinement. In contrast, our use of 3DMM for face synthesis is geared towards enriching the diversity of training data.

Deep face recognition Several frameworks have recently been proposed that use DNNs to achieve impressive performances [22, 32, 37, 38, 39, 43, 44]. DeepFace [39] achieved verification rates comparable to human labeling on large test datasets, with further improvements from works such as DeepID [38]. Collecting face images from the Internet, FaceNet [32] trains on 200 million images from 8 million

Figure 2. An overview of the proposed approach. (a) *Pose-variant face generation* utilizes a 3D facial model to synthesize new viewpoints from near-frontal faces. (b) *Rich feature embedding* is then achieved by jointly learning the identity and non-identity features using multi-source supervisions. (c) Finally, *Disentangling by reconstruction* is applied to distill the identity feature from the non-identity one for robust and pose-invariant representation.

subjects. The very deep network can only be well stimulated by the huge volume of training data. We also use DNNs, but adopt the contrasting approach of learning pose-invariant features, since large-scale datasets with pose variations are expensive to collect, or do not exist in several applications such as surveillance.

Pose-invariant face recognition Early works use Canonical Correlation Analysis (CCA) to analyze the commonality among different pose subspaces [8, 21]. Further works consider generalization across multiple viewpoints [34] and multiview inter and intra discriminant analysis [13]. With the introduction of DNNs, prior works aim to transfer information from pose variant inputs to a frontalized appearance [41, 45], which is then used for face recognition [51]. The frontal appearance reconstruction usually relies on large amount of training data and the pairing across poses is too strict to be practical. Stacked progressive autoencoders (SPAEC) [11] map face appearances under larger non-frontal poses to those under smaller ones in a continuous way by setting up hidden layers. The regression based mapping highly depends on training data and may lack generalization ability. Hierarchical-PEP [17] employs probabilistic elastic part (PEP) model to match facial parts from different yaw angles for unconstrained face recognition scenarios. The 3D face reconstruction method [50] synthesizes missing appearance due to large view points, which may introduce noise. Rather than compensating the missing information caused by severe pose variations at appearance level, we target learning a pose-invariant representation at feature level which preserves discrimination power through deep training.

Disentangle factors of variation Contractive discriminative analysis [28] learns disentangled representations in semi-supervised framework by regularizing representations to be

orthogonal to each other. Disentangling Boltzmann machine [27] regularizes representations to be specific to each target task via manifold interaction. These methods involve non-trivial training procedure, and the pose variation is limited to half-profile views ($\pm 45^\circ$). Inverse graphics network [16] learns an interpretable representation by learning and decoding graphics codes, each of which encodes different factors of variation, but has been demonstrated only on the database generated from 3D CAD models. Multi-View Perceptron [52] disentangles pose and identity factors by cross-reconstruction of images synthesized from deterministic identity neurons and random hidden neurons. But it does not account for factors such as illumination or expression that are also needed for image-level reconstruction. In contrast, we use carefully designed embeddings as reconstruction targets instead of pixel-level images, which reduces the burden of reconstructing irrelevant factors of variation.

3. Proposed Method

We propose a novel pose-invariant feature learning method for large pose face recognition. Figure 2 provides an overview of our approach. *Pose-variant face generation* utilizes a 3D facial model to augment the training data with faces of novel viewpoints, besides generating ground-truth pose and facial landmark annotations. *Rich feature embedding* is then achieved by jointly learning the identity and non-identity features using multi-source supervision. Finally, *disentanglement by feature reconstruction* is performed to distill the identity feature from the non-identity one for better discrimination ability and pose-invariance.

Figure 3. Pose-variant faces are used to finetune an off-the-shell recognition network r to learn the rich feature embedding e^r , which is explicitly branched into the identity feature e^i and the non-identity feature e^n . Multi-source supervisions, such as identity, pose and landmark, are applied for joint optimization.

3.1. Pose-variant Face Generation

The goal is to generate a series of pose-variant faces from a near-frontal image. This choice of generation approach is deliberate, since it can avoid hallucinating missing textures due to self-occlusion, which is a common problem with former approaches [9, 5] that rotate non-frontal faces to a normalized frontal view. More importantly, enriching instead of reducing intra-subject variations provides important training examples in learning pose-invariant features.

We reconstruct the 3D shape from a near-frontal face to generate new face images. Let \mathcal{I} be the set of frontal face images. A straightforward solution is to learn a nonlinear mapping $f(\cdot; \mathbf{s}) : \mathbb{R}^{3N}$ that maps an image x to the N coordinates of a 3D mesh. However, it is non-trivial to do so for a large number of vertices (15k), as required for a high-fidelity reconstruction.

Instead, we employ the 3D Morphable Model (3DMM) [2] to learn a nonlinear mapping $f(\cdot; \mathbf{s}) : \mathbb{R}^{235}$ that embeds x to a low-dimensional parameter space. The 3DMM parameters \mathbf{p} control the rigid affine transformation and non-rigid deformation from a 3D mean shape \bar{S} to the instance shape S . Please refer to Figure 2 for an illustration:

$$S(\mathbf{p}) = sR(\bar{S} + \mathbf{id} \mathbf{id} + \mathbf{exp} \mathbf{exp}) + \mathbf{T}, \quad (1)$$

where $\mathbf{p} = \{s, R, \mathbf{T}, \mathbf{id}, \mathbf{exp}\}$ including scale s , rotation R , translation \mathbf{T} , identity coefficient \mathbf{id} and expression coefficient \mathbf{exp} . The eigenbases \mathbf{id} and \mathbf{exp} are learned offline using 3D face scans to model the identity [23] and expression [3] subspaces, respectively.

Once the 3D shape is recovered, we rotate the near-frontal face by evenly manipulating the yaw angle in the range of $[-90^\circ, 90^\circ]$. We follow [50] to use a z-buffer for collecting texture information and render the background for high-quality recovery. The rendered face is then projected to 2D to generate new face images from novel viewpoints.

3.2. Rich Feature Embedding

Most existing face recognition algorithms [19, 20, 32, 43] learn face representation using only identity supervision. An

underlying assumption of their success is that deep networks can “implicitly” learn to suppress non-identity factors after seeing a large volume of images with identity labels [32, 39].

However, this assumption does not always hold when extensive non-identity variations exist. As shown in Figure 1 (a), the face representation and pose changes still present substantial correlations, even though this representation is learned through a very deep neural network (VGGFace [22]) with large-scale training data (2.6M).

This indicates that using only identity supervision might not suffice to achieve an invariant representation. Motivated by this observation, we propose to utilize multi-source supervision to learn a rich feature embedding e^r , which can be “explicitly” branched into an identity feature e^i and a non-identity feature e^n , respectively. As we will show in the next section, the two features can collaborate to effectively achieve an invariant representation.

More specifically, as illustrated in Figure 3, e^n can be further branched as e^p and e^l to represent pose and landmark cues. For our multi-source training data that are not generated, we apply the CASIA-WebFace database [44] and provide the supervision from an off-the-shelf pose estimator [48]. Therefore, we have:

$$\begin{aligned} e^i &= f(x; r, i), \quad e^n = f(x; r, n), \\ e^p &= h(e^n; w^p) = f(x; r, n, w^p), \\ e^l &= h(e^n; w^l) = f(x; r, n, w^l), \end{aligned}$$

where mapping $f(\cdot; w) : \mathbb{R}^d$ takes x and generates an embedding vector $f(x)$ and w denotes the mapping parameters. Here, r can be any off-the-shelf recognition network. $h(\cdot; \cdot)$ is used to bridge two embedding vectors. We jointly learn all embeddings by optimizing:

$$\begin{aligned} \argmin_{r, i, n, w^i, p, l} \quad & - \sum_{\text{image}} y^i \log \text{softmax}(w^{iT} e^i) \\ & + \lambda^p y^p - e^p \frac{\lambda^p}{2} + \lambda^l y^l - e^l \frac{\lambda^l}{2}, \quad (2) \end{aligned}$$

where y^i , y^p and y^l are identity, pose and landmark annotations and λ^i , λ^p and λ^l balance the weights between cross-entropy and l_2 loss.

By resorting to multi-source supervision, we can learn the rich feature embedding that “explicitly” encodes both identity and non-identity cues in e^i and e^n , respectively. The remaining challenge is to distill e^i by disentangling from e^n to achieve identity-only representation.

3.3. Disentanglement by Feature Reconstruction

The identity and non-identity features above are jointly learned under different supervision. However, there is no guarantee that the identity factor has been fully disentangled from the non-identity one since there is no supervision applied on the decoupling process. This fact motivates us to

Figure 4. A genuine pair $F \times_1, \times_2 G$ that share the same identity but different pose is fed into the recognition network r to obtain the rich embedding e_1^r and e_2^r . By regularizing the self and cross reconstruction, e_{11}^r and e_{21}^r , the identity and non-identity features are eventually disentangled to make the non-frontal peer e_2^i to be similar to its near-frontal reference e_1^i .

propose a novel reconstruction-based framework for effective identity and non-identity disentanglement.

Recall that we have generated a series of pose-variant faces for each training subject in Section 3.1. These images share the same identity but have different viewpoints. We categorize these images into two groups according to their absolute yaw angles: near-frontal faces ($< 5^\circ$) and non-frontal faces ($> 5^\circ$). The two groups are used to sample image pairs that follow a specially designed configuration: a reference image which is randomly selected from the near-frontal group and a peer image which is randomly picked from the non-frontal group.

The next step is to obtain the identity and non-identity embeddings of two faces that have the same identity but different viewpoints. As shown in Figure 4, a pair of images $\{x_k : k = 1, 2\}$ are fed into the network to output the corresponding identity and non-identity features:

$$\begin{aligned} e_k^i &= f(e_k^r; i) = f(x_k; r, i), \\ e_k^n &= f(e_k^r; n) = f(x_k; r, n). \end{aligned}$$

Note that i is not indexed by k as the network shares weights to process images of the same pair.

Our goal is to eventually push e_1^i and e_2^i close to each other to achieve a pose-invariant representation. A simple solution is to directly minimize the l_2 distance between the two features in the embedding subspace. However, this constraint only considers the identity branch, which might be entangled with non-identity, but completely ignores the non-identity factor, which provides strong supervision to purify the identity. Our experiments also indicate that a hard constraint would suffer from limited performance in large-pose conditions.

To address this issue, we propose to relax the constraint under a reconstruction-based framework. More specifically, we firstly introduce two reconstruction tasks:

$$e_{11}^r = g(e_1^i, e_1^n; c), \quad e_{21}^r = g(e_2^i, e_1^n; c),$$

where e_{11}^r denotes the *self reconstruction* of the near-frontal rich embedding; while e_{21}^r denotes the *cross reconstruction*

of the non-frontal rich embedding. Here, $g(\cdot, \cdot; c)$ is the reconstruction mapping with parameter c .

The identity and non-identity features can be rebalanced from the rich feature embedding by minimizing the self and cross reconstruction loss under the cross-entropy constraint:

$$\begin{aligned} \underset{i, n, c \text{ pair}}{\operatorname{argmin}} \quad & -i y_1^i \log \operatorname{softmax}(w^T e_1^i) \\ & + s \|e_{11}^r - e_1^r\|_2^2 + c \|e_{21}^r - e_1^r\|_2^2, \end{aligned} \quad (3)$$

where i , s and c weigh different constraints. Note that compared to (2), here we only finetune $\{i, n\}$ (as well as c) to rebalance the identity and non-identity features while keeping r fixed, which is an important strategy to maintain the previously learned rich embedding.

In (3), we regularize both self and cross reconstructions to be close to the near-frontal rich embedding e_1^r . Thus, portions of e_2^r to e_1^i and e_2^n are dynamically rebalanced to make the non-frontal peer e_2^i to be similar to the near-frontal reference e_1^i . In other words, we encourage the network to learn a normalized feature representation across pose variations, thereby disentangling pose information from identity.

The proposed feature-level reconstruction is significantly different from former methods [32, 9] that attempt to frontalize faces at the image level. It can be directly optimized for pose invariance without suffering from artifacts that are common issues in face frontalization. Besides, our approach is an end-to-end solution that does not rely on extensive preprocessing usually required for image-level face normalization.

Our approach is also distinct from existing methods [20, 19] that synthesize pose-variant faces for data augmentation. Instead of feeding the network with a large number of augmented faces and letting it automatically learn pose-invariant or pose-specific features, we utilize the reconstruction loss to supervise the feature decoupling procedure. Moreover, factors of variation other than pose are also present in training, even though we only use pose as the driver for disentanglement. The cross-entropy loss in (3) plays an important role in preserving the discriminative power of identity features across various factors.

4. Implementation Details

Pose-variant face generation A deep network is employed to predict 3DMM parameters of a near-frontal face as shown in Figure 2 (a). The network has a similar architecture as VGG16 [35]. We use pre-trained weights learned from ImageNet [15] to initialize the network instead of training from scratch. To further improve the performance, we make two important changes: (1) we use stride-2 convolution instead of max pooling to preserve the structure information when halving the feature maps; (2) the dimension of 3DMM parameters is changed to 66-d (30 identity, 29 expression and 7 pose) instead of 235-d used in [49]. We evenly sample new viewpoints in every 5° from near-frontal faces to

left/right profiles to cover the full range of pose variations.

Rich feature embedding The network is designed based on CASIA-net [44] with some improvements. As illustrated in Figure 3, we change the last fully connected layer to 512-d for the rich feature embedding, which is then branched into 256-d neurons for the identity feature and 128-d neurons for the non-identity feature. To utilize multi-source supervision, the non-identity feature is further forked into 7-d neurons for the pose embedding and 136-d neurons for the landmark coordinates. Three different datasets are used to train the network: CASIA-WebFace, 300WLP and MultiPIE. We use Adam [14] stochastic optimizer with an initial learning rate of 0.0003, which drops by a factor of 0.25 every 5 epochs until convergence. Note that we train the network from scratch on purpose, since a pre-trained recognition model usually has limited ability to re-encode non-identity features.

Disentanglement by reconstruction Once $\{r, i, n\}$ are learned in the rich feature embedding, we freeze r and finetune i and n to rebalance the identity and non-identity features as explained in Figure 4 and (3). The network takes the concatenation (384-d) of e^i and e^n and outputs the reconstructed embedding (512-d). The mapping is achieved by rolling through two fully connected layers and each of them has 512-d neurons. We have tried different network configurations but get similar performance. The initial learning rate is set to 0.0001 and the hyper-parameters i, s, c are determined via 5-fold cross-validation. We also find that it is import to do early stopping for effective reconstruction-based regularization. In (2) and (3), we use the cross-entropy loss to preserve the discriminative power of the identity feature. Other identity regularizations, *e.g.* triplet loss [32], can be easily applied in a plug-and-play manner.

5. Experiments

We evaluate our feature learning method on three main pose-variant databases, MultiPIE [7], 300WLP [49] and CFP [33]. We also compare with two top general face recognition frameworks, VGGFace [22] and N-pair loss face recognition [36], and three state-of-the-art pose-invariant face recognition methods, namely, MvDA [13], GMA [34] and MvDN [12]. Further, we present an ablation study to emphasize the significance of each module that we carefully designed and a cross-database validation demonstrates the good generalization ability of our method.

5.1. Evaluation on MultiPIE

MultiPIE [7] is composed of 754,200 images of 337 subjects with different factors of variation such as pose, illumination, and expression. There are 15 different head poses set up, where we only use images of 13 head poses with yaw angle changes from -90° to 90° , with 15° difference every consecutive pose bin in this experiment.

We split the data into train and test by subjects, of which the first 229 subjects are used for training and the remaining

Method	15	30	45	60	75	90	Avg
VGGFace [22]	0.972	0.961	0.926	0.847	0.628	0.342	0.780
N-pair [36]	0.990	0.983	0.971	0.944	0.811	0.468	0.861
MvDA [13] [†]	1.000	0.979	0.909	0.855	0.718	0.564	0.837
GMA [34] [†]	1.000	1.000	0.904	0.852	0.725	0.550	0.838
MvDN [12] [†]	1.000	0.991	0.921	0.897	0.810	0.706	0.887
Ours (P1)	0.972	0.966	0.956	0.927	0.857	0.749	0.905
Ours (P2)	1.000	1.000	0.995	0.982	0.931	0.817	0.954

Table 1. Rank-1 recognition accuracy on MultiPIE at different yaw angles. The numbers in the entry with [†] are obtained from [12]. We evaluate our method using gallery set composed of 2 frontal face images per subject (P1) as well as entire frontal face images (P2).

Method	15	30	45	60	75	90	Avg
VGGFace [22]	0.994	0.998	0.996	0.956	0.804	0.486	0.838
N-Pair [36]	1.000	0.996	0.993	0.962	0.845	0.542	0.859
Ours	1.000	0.999	0.995	0.994	0.978	0.940	0.980

Table 2. Recognition performance on 300WLP, the proposed method with two general state-of-the-art face recognition frameworks, *i.e.* VGG Face Recognition Network (VGGFace) and N-pair loss face recognition (N-pair).

108 are used for testing. This is similar to the experimental setting in [12], but we use entire data including both illumination and expression variations for training while excluding only those images taken with top-down views. Rank-1 recognition accuracy of non-frontal face images is reported. We take $\pm 15^\circ$ to $\pm 90^\circ$ as query and the frontal faces (0°) as gallery, while restricting illumination condition to be neutral.

To be consistent with the experimental setting of [12], we form a gallery set by randomly selecting 2 frontal face images per subject, of which there are a total of 216 images. We evaluate the recognition accuracy for all query examples, of which there are 619 images per pose. The procedure is done with 10 random selections of gallery sets and mean accuracy is reported.

Evaluation is shown in Table 1. The recognition accuracy at every 15° interval of yaw angle is reported while averaging its symmetric counterpart with respect to the 0-yaw axis. For the two general face recognition algorithms, VGGFace [22] and N-pair loss [36], we clearly observe more than 30% accuracy drop when the head pose approaches 90° from 75° . Our method significantly reduces the drop by more than 20%. The general methods are trained with very large databases leveraging across different poses, but our method has the additional benefit of explicitly aiming for a pose invariant feature representation.

The pose-invariant methods, GMA, MvDA, and MvDN demonstrate good performance within 30° yaw angles, but again the performance starts to degrade significantly when yaw angle is larger than 30° . When comparing the accuracy on extreme poses from 45° to 90° , our method achieves accuracy 3–4% better than the best reported. Besides the improved performance, our method has an advantage over

Method	MultiPIE							300WLP						
	15	30	45	60	75	90	Avg	15	30	45	60	75	90	Avg
SS	0.908	0.899	0.864	0.778	0.487	0.207	0.690	0.945	0.934	0.884	0.753	0.567	0.330	0.679
SS-FT	0.941	0.936	0.919	0.883	0.799	0.681	0.860	1.000	0.999	0.992	0.973	0.934	0.839	0.944
MSMT	0.965	0.955	0.945	0.914	0.827	0.689	0.882	1.000	0.993	0.993	0.986	0.968	0.922	0.971
MSMT+L2	0.972	0.965	0.954	0.923	0.849	0.739	0.900	1.000	0.997	0.996	0.991	0.973	0.933	0.977
MSMT+SR	0.972	0.966	0.956	0.927	0.857	0.749	0.905	1.000	0.999	0.995	0.994	0.978	0.940	0.980
MSMT [†]	0.993	0.989	0.982	0.959	0.903	0.734	0.927	1.000	0.998	0.997	0.994	0.981	0.922	0.977
MSMT [†] +SR	0.994	0.990	0.982	0.960	0.906	0.745	0.929	1.000	0.998	0.999	0.997	0.988	0.953	0.986

Table 3. Recognition performance of several baseline models, i.e., single source trained model on CASIA database (SS), single source model fine-tuned on the target database (SS-FT), multi-source multi-task models (MSMT), MSMT with direct identity feature + distance regularization (MSMT+L2), the proposed MSMT with Siamese reconstruction regularization models (MSMT+SR), MSMT with N-pair loss instead of cross entropy loss (MSMT[†]) and MSMT[†] with SR, evaluated on MultiPIE (P1) and 300WLP.

Method	Frontal-Frontal	Frontal-Profile
Sengupta et al. [33]	96.40	84.91
Sankarana et al. [31]	96.93	89.17
Chen et al. [4]	98.67	91.97
DR-GAN [41]	97.84	93.41
Human	96.24	94.57
Ours	98.67	93.76

Table 4. Verification accuracy comparison on CFP dataset.

MvDN, since it does not require pose information at test time. On the other hand, MvDN is composed of multiple sub-networks, each of which is specific to a certain pose variation and therefore requires additional information on head pose for recognition.

5.2. Evaluation on 300WLP

We further evaluate on a face-in-the-wild database, 300 Wild Large Pose [49] (300WLP). It is generated from 300W [30] face database by 3DDFA [49], in which it establishes a 3D morphable model and reconstruct the face appearance with varying head poses. It consists of overall 122,430 images from 3,837 subjects. Compared to MultiPIE, the overall volume is smaller, but the number of subjects is significantly larger. For each subject, images are with uniformly distributed continuously varying head poses in contrast to MultiPIE’s strictly controlled 15 head pose intervals. The lighting conditions as well as the background are almost identical. Thus, it is an ideal dataset to evaluate algorithms for pose variation.

We randomly split 500 subjects of 8014 images as testing data and the rest 3337 subjects of 106,402 images as the training data. Among the testing data, two 0 head pose images per subject form the gallery and the rest 7014 images serves as the probe. Table 2 shows the comparison with two state-of-the-art general face recognition methods, i.e. VGGFace [22] and N-pair loss face recognition [36]. To the best of our knowledge, we are the first to apply our pose-invariant face recognition framework on this dataset. Thus, we only compare our method with the two general

face recognition frameworks.

Since head poses in 300WLP continuously vary, we group the test samples into 6 pose intervals, (0, 15), (15, 30), (30, 45), (45, 60), (60, 75) and (75, 90). For short annotation, we mark each interval with the end point, e.g., 30 denotes the pose interval (15, 30). From Table 2, our method achieves consistently better accuracy especially when pose angle approaches 90, which is clearly contributed by our feature reconstruction based disentanglement.

5.3. Evaluation on CFP

The Celebrities in Frontal-Profile (CFP) database [33] focuses on extreme head pose face verification. It consists of 500 subjects, with 10 frontal images and 4 profile images for each, in a wild setting. The evaluation is conducted by averaging the performance of 10 randomly selected splits with 350 identical and 350 non-identical pairs. Our MSMT+SR finetuned on MultiPIE with N-pair loss is the model evaluated in this experiment. The reported human performance is 94.57% accuracy on the frontal-profile protocol and 96.24% on the frontal-frontal protocol, which shows the challenge of recognizing profile views.

Results in Table 4 suggest that our method achieves consistently better performance compared to state-of-the-art. We reach the same Frontal-Frontal accuracy as Chen et al. [4] while being significantly better on Frontal-Profile by 1.8%. We are slightly better than DR-GAN [41] on extreme pose evaluation and 0.8% better on frontal cases. DR-GAN is a recent generative method that seeks the identity preservation at the image level, which is not a direct optimization on the features. Our feature reconstruction method preserves identity even when presented with profile view faces. In particular, as opposed to prior methods, ours is the only one that obtains very high accuracy on both the evaluation protocols.

5.4. Control Experiments

We extensively evaluate recognition performance on various baselines to study the effectiveness of each module in our proposed framework. Specifically, we evaluate and compare the following models:

Method		MultiPIE							300WLP						
		15	30	45	60	75	90	Avg	15	30	45	60	75	90	Avg
MultiPIE	MSMT	0.965	0.955	0.945	0.914	0.827	0.689	0.882	1.000	0.996	0.988	0.953	0.889	0.720	0.904
	Ours	0.972	0.966	0.956	0.927	0.857	0.749	0.905	0.994	0.995	0.992	0.958	0.901	0.733	0.910
300WLP	MSMT	0.941	0.927	0.898	0.837	0.695	0.432	0.788	1.000	0.993	0.993	0.986	0.968	0.922	0.971
	Ours	0.945	0.933	0.910	0.862	0.736	0.459	0.808	1.000	0.999	0.995	0.994	0.978	0.940	0.980

Table 5. Cross database evaluation on MultiPIE and 300WLP. The top two rows show the model of MSMT and our method trained on CASIA and MultiPIE, while tested on both MultiPIE and 300WLP. The bottom two rows show the model of MSMT and our method trained on CASIA and 300WLP, while tested on both MultiPIE and 300WLP.

- SS: trained on a single source (e.g., CASIA-WebFace) using softmax loss only.
- SS-Ft: SS fine-tuned on a target dataset (e.g., MultiPIE or 300WLP) using softmax loss only.
- MSMT: trained on multiple data sources (e.g., CASIA + MultiPIE or 300WLP) using softmax loss for identity and L_2 loss for pose.
- MSMT+L2: fine-tuned on MSMT models using softmax loss and Euclidean loss on pairs.
- MSMT+SR: fine-tuned on MSMT models using softmax loss and Siamese reconstruction loss.
- MSMT[†]: trained on the same multiple data sources as MSMT, using N-pair [36] metric loss for identity and L_2 loss for pose.
- MSMT[†]+SR: finetuned on MSMT[†] models with N-pair loss and reconstruction loss.

The SS model serves as the weakest baseline. We observe that simultaneously training the network on multiple sources of CASIA and MultiPIE (or 300WLP) using multi-task objective (i.e., identification loss, pose or landmark estimation loss) is more effective than single-source training followed by fine-tuning. We believe that our MSMT learning can be viewed as a form of curriculum learning [1] since multiple objectives introduced by multi-source and multi-task learning are at different levels of difficulty (e.g., pose and landmark estimation or identification on MultiPIE and 300WLP are relatively easier than identification on CASIA-WebFace) and easier objectives allow to train faster and converge to better solution.

As an alternative to reconstruction regularization, one may consider reducing the distance between the identity-related features of the same subject under different pose directly (MSMT+L2). Learning to reduce the distance improves the performance over the MSMT model, but is not as effective as our proposed reconstruction regularization method, especially on face images with large pose variations.

Further, we observe that employing the N-pair loss [36] within our framework also boosts performance, which is shown by the improvements from MSMT to MSMT[†] and MSMT+SR to MSMT[†]+SR. We note that the MSMT[†] baseline is not explored in prior works on pose-invariant face recognition. It provides a different way to achieve similar goals as the proposed reconstruction method. Indeed, a collateral observation through the relative performances

of MSMT and MSMT[†] is that the softmax loss is not good at disentangling pose from identity, while metric learning excels at it. Indeed, our feature reconstruction metric might be seen as achieving a similar goal, thus, improvements over MSMT[†] are marginal, while those over MSMT are large.

5.5. Cross Database Evaluation

We evaluate our models, which are trained on CASIA with MultiPIE or 300WLP, on the cross test set 300WLP or MultiPIE, respectively. Results are shown in Table 5 to validate the generalization ability. There are obvious accuracy drops on both databases, for instance, a 7% drop on 300WLP and 10% drop on MultiPIE. However, such performance drops are expected since there exists a large gap in the distribution between MultiPIE and 300WLP.

Interestingly, we observe significant improvements when compared to VGGFace. These are fair comparisons since neither networks is trained on the training set of the target dataset. When evaluated on MultiPIE, our MSMT model trained on 300WLP and CASIA database improves 0.8% over VGGFace and the model with reconstruction regularization demonstrates stronger performance, showing 2.8% improvement over VGGFace. Similarly, we observe 6.6% and 7.2% improvements for MultiPIE and CASIA trained MSMT models and our proposed MSMT+SR, respectively, over VGGFace when evaluated on the 300WLP test set. This partially confirms that our performance is not an artifact of overfitting to a specific dataset, but is generalizable across different datasets of unseen images.

6. Conclusion

In the paper, we propose a new reconstruction loss to regularize identity feature learning for face recognition. We also introduce a data synthesization strategy to enrich the diversity of pose, requiring no additional training data. Rich embedding has already shown promising effects revealed by our control experiments, which is interpreted as curriculum learning. The self and cross reconstruction regularization achieves successful disentanglement of identity and pose, to show significant improvements on both MultiPIE, 300WLP and CFP with 2% to 12% gaps. Cross-database evaluation further verifies that our model generalizes well across databases. Future work will focus on closing the systematic gap among databases and further improve the generalization ability.

References

- [1] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009. **8**
- [2] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *TPAMI*, 25(9):1063–1074, 2003. **2, 4**
- [3] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. FaceWarehouse: a 3D facial expression database for visual computing. *TVCG*, 20(3):413–425, Mar. 2014. **2, 4**
- [4] J.-C. Chen, J. Zheng, V. Patel, and R. Chellappa. Fisher vector encoded deep convolutional features for unconstrained face verification. In *ICIP*, 2016. **7**
- [5] C. N. Duong, K. Luu, K. G. Quach, and T. D. Bui. Beyond principal components: Deep boltzmann machines for face modeling. In *CVPR*, 2015. **4**
- [6] S. Edelman and H. H. Bülthoff. Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, 32(12):2385–2400, 1992. **1**
- [7] R. Gross, I. Matthew, J. Cohn, T. Kanade, and S. Baker. MultiPie. *Image and Vision Computing*, 2009. **1, 6**
- [8] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.*, 16, 2004. **2, 3**
- [9] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained image. In *CVPR*, 2015. **2, 4, 5**
- [10] J. E. Hummel and I. Biederman. Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99(3):480–517, 1992. **1**
- [11] M. Kan, S. Shan, H. Chang, and X. Chen. Stacked progressive auto-encoders (spae) for face recognition across poses. In *CVPR*, 2014. **3**
- [12] M. Kan, S. Shan, and X. Chen. Multi-view deep network for cross-view classification. In *CVPR*, 2016. **2, 6**
- [13] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. In *ECCV*, 2012. **3, 6**
- [14] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint: 1412.6980*, 2014. **6**
- [15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. **1, 5**
- [16] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum. Deep convolutional inverse graphics network. In *NIPS*, 2015. **3**
- [17] H. Li and G. Hua. Hierarchical-pep model for real-world face recognition. In *CVPR*, 2015. **3**
- [18] S. Liang, L. Shapiro, and I. Kemelmacher-Shlizerman. Head reconstruction from internet photos. In *ECCV*, 2016. **2**
- [19] I. Masi, A. T. an Tr  n, T. Hassner, J. T. Leksut, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? In *ECCV*, 2016. **2, 4, 5**
- [20] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *CVPR*, 2016. **2, 4, 5**
- [21] A. Nielson. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE Trans. on Image Processing*, 11(3), 2002. **3**
- [22] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. **1, 2, 4, 6, 7**
- [23] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. In *AVSS*, 2009. **4**
- [24] X. Peng, J. Huang, Q. Hu, S. Zhang, A. Elgammal, and D. Metaxas. From circle to 3-sphere: Head pose estimation by instance parameterization. *Computer Vision and Image Understanding*, 136:92–102, 2015. **1**
- [25] X. Peng, S. Zhang, Y. Yu, and D. N. Metaxas. Toward personalized modeling: Incremental and ensemble alignment for sequential faces in the wild. *International Journal of Computer Vision*, pages 1–14, 2017. **2**
- [26] T. Poggio and S. Edelman. A network that learns to recognize 3-dimensional objects. *Nature*, 343(6255):263–266, 1990. **1**
- [27] S. Reed, K. Sohn, Y. Zhang, and H. Lee. Learning to disentangle factors of variation with manifold interaction. In *ICML*, 2014. **3**
- [28] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza. Disentangling factors of variation for facial expression recognition. In *ECCV*, 2012. **3**
- [29] J. Roth, Y. Tong, and X. Liu. Unconstrained 3d face reconstruction. In *CVPR*, 2015. **2**
- [30] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, 2013. **7**
- [31] S. Sankaranarayanan, A. Alavi, C. Castillo, and R. Chellappa. Triplet probabilistic embedding for face verification and clustering. In *arXiv preprint*, volume 1605.05396, 2016. **7**
- [32] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. **1, 2, 4, 5, 6**
- [33] S. Sengupta, J.-C. Chen, C. Castillo, V. Patel, R. Chellappa, and D. Jacobs. Frontal to profile face verification in the wild. In *WACV*, 2016. **6, 7**
- [34] A. Sharma, A. Kumar, H. D. III, and D. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, 2012. **3, 6**
- [35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *arXiv preprint*, 2014. **5**
- [36] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016. **6, 7, 8**
- [37] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS*, pages 1988–1996. 2014. **2**
- [38] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, 2014. **2**
- [39] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to Human-Level performance in face verification. In *CVPR*, 2014. **1, 2, 4**
- [40] A. T. Tran, T. Hassner, I. Masi, and G. G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. *CoRR*, abs/1612.04904, 2016. **2**
- [41] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017. **3, 7**
- [42] X. Wang, G. Guo, M. Merler, N. C. Codella, M. Rohith, J. R. Smith, and C. Kambhampettu. Leveraging multiple cues for

- recognizing family photos. *Image and Vision Computing*, 58:61–75, 2017. 2
- [43] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 2, 4
 - [44] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. In *CoRR*, 2014. 2, 4, 5
 - [45] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017. 3
 - [46] X. Yu, J. Huang, S. Zhang, and D. N. Metaxas. Face landmark fitting via optimized part mixtures and cascaded deformable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2212 – 2226, 2015. 2
 - [47] X. Yu, Z. Lin, J. Brandt, and D. N. Metaxas. Consensus of regression for occlusion-robust facial feature localization. In *ECCV*, 2014. 2
 - [48] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. In *ECCV*, 2016. 4
 - [49] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016. 5, 6, 7
 - [50] 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 *CVPR*, 2015. 2, 3, 4
 - [51] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *ICCV*, 2013. 3
 - [52] Z. Zhu, P. Luo, X. Wang, and X. Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In *NIPS*, 2014. 3