Feature Reconstruction Disentangling for Pose-invariant Face Recognition Supplementary Material

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
{xipeng.cs, dnm}@rutgers.edu, {xiangyu,ksohn,manu}@nec-labs.com

1. Summary of The Supplementary

This supplementary file includes two parts: **(a)** Additional implementation details are presented to improve the reproducibility; **(b)** More experimental results are presented to validate our approach in different aspects, which are not shown in the main submission due to the space limitation.

2. Additional Implementation Details

Pose-variant face generation We designed a network to predict 3DMM parameters from a single face image. The design is mainly based on VGG16 [4]. We use the same number of convolutional layers as VGG16 but replacing all max pooling layers with stride-2 convolutional operations. The fully connected (fc) layers are also different: we first use two fc layers, each of which has 1024 neurons, to connect with the convolutional modules; then, a fc layer of 30 neurons is used for identity parameters, a fc layer of 29 neurons is used for expression parameters, and a fc layer of 7 neurons is used for pose parameters. Different from [8] uses 199 parameters to represent the identity coefficients, we truncate the number of identity eigenvectors to 30 which preserves 90% of variations. This truncation leads to fast convergence and less overfitting. For texture, we only generate non-frontal faces from frontal ones, which significantly mitigate the hallucinating texture issue caused by self occlusion and guarantee high-fidelity reconstruction. We apply the Z-Buffer algorithm used in [8] to prevent ambiguous pixel intensities due to same image plane position but different depths.

Rich feature embedding The design of the rich embedding network is mainly based on the architecture of CASIA-net [6] since it is wildly used in former approach and achieves strong performance in face recognition. During training, CASIA+MultiPIE or CASIA+300WLP are used. As shown in Figure 3 of the main submission, after the convolutional layers of CASIA-net, we use a 512-d FC for the rich feature embedding, which is further branched into a 256-d identity feature and a 128-d non-identity feature. The 128-d non-identity feature is further connected with a 136-d landmark prediction and a 7-d pose prediction. Notice that in the face generation network, the number of pose parameters is 7 instead of 3 because we need to uniquely depict the projection matrix from the 3D model and the 2D face shape in image domain, which includes scale, pitch, yaw, roll, x translation, y translation, and z translations.

Disentanglement by feature reconstruction Once the rich embedding network is trained, we feed genius pair that share the same identity but different viewpoints into the network to obtain the corresponding rich embedding, identity and non-identity features. To disentangle the identity and pose factors, we concatenate the identity and non-identity features and roll though two 512-d fully connected layers to output a reconstructed rich embedding depicted by 512 neurons. Both self and cross reconstruction loss are designed to eventually push the two identity features close to each other. At the same time, a cross-entropy loss is applied on the near-frontal identity feature to maintain the discriminative power of the learned representation. The disentanglement of the identity and pose is finally achieved by the proposed feature reconstruction based metric learning.

3. Additional Experimental Results

In addition to the main submission, we present more experimental results in this section to further validate our approach in different aspects.

3.1. P1 and P2 protocol on MultiPIE

In the main submission, due to space considerations, we only report the mean accuracy over 10 random training and testing splits, on MultiPIE and 300WLP separately. In Table 1, we report the standard deviation of our method as a more complete comparison. From the results, the standard deviation of our method is also very small, which suggests that the performance is consistent across all the trials. We

Method	MultiPIE						
	15°	30°	45°	60°	75°	90°	Avg
SS	0.908(0.0088)	0.899(0.0088)	0.864(0.0072)	0.778(0.0084)	0.487(0.0119)	0.207(0.0156)	0.690(0.2600)
SS-FT	0.941(0.0067)	0.936(0.0090)	0.919(0.0105)	0.883(0.0113)	0.799(0.0108)	0.681(0.0130)	0.860(0.0940)
MSMT	0.965(0.0053)	0.955(0.0054)	0.945(0.0062)	0.914(0.0059)	0.827(0.0110)	0.689(0.0143)	0.882(0.0982)
MSMT+L2	0.972(0.0058)	0.965(0.0056)	0.954(0.0075)	0.923(0.0048)	0.849(0.0067)	0.739(0.0095)	0.900(0.0834)
MSMT+SR (ours)	**0.972(0.0060)**	**0.966(0.0069)**	**0.955(0.0068)**	**0.927(0.0068)**	**0.857(0.0066)**	**0.749(0.0105)**	**0.905(0.0797)**

Table 1. Rank-1 recognition accuracy comparisons on MultiPIE [1] under P1 testing protocol.

Method		MultiPIE						
		15°	30°	45°	60°	75°	90°	Avg
300WLP	MSMT (P1)	0.941(0.0051)	0.927(0.0059)	0.898(0.0073)	0.837(0.0106)	0.695(0.0135)	0.432(0.0110)	0.788(0.1794)
	Ours (P1)	**0.945(0.0067)**	**0.933(0.0068)**	**0.910(0.0073)**	**0.862(0.0082)**	**0.736(0.0096)**	**0.459(0.01359)**	**0.808(0.1709)**
300WLP	MSMT (P2)	1.00	1.00	0.992	0.943	0.797	0.488	0.870
	Ours (P2)	1.00	1.00	**0.993**	**0.964**	**0.838**	**0.511**	**0.884**

Table 2. Cross database evaluation under either P1 or P2 protocols. Training: CASIA [6] and 300WLP [8]. Testing: MultiPIE [1].

Method	MultiPIE						
	15°	30°	45°	60°	75°	90°	Avg
SS	1.00	0.998	0.985	0.892	0.563	0.250	0.781
SS-FT	0.999	0.993	0.981	0.951	0.874	0.753	0.925
MSMT	1.00	1.00	0.993	0.982	0.908	0.753	0.939
MSMT+L2	1.00	999	0.990	0.978	0.911	0.800	0.946
MSMT+SR (ours)	**1.00**	0.999	**0.995**	**0.982**	**0.931**	**0.817**	**0.954**

Table 3. Recognition accuracy of different baseline models.

also compare the cross database evaluation on both mean accuracy and standard deviation in Table 2. We show the models trained on 300WLP and tested on MultiPIE with both P1 and P2 protocol. Please note that with P2 protocol, our method still achieves better performance on MultiPIE than MvDN [3] with 0.7% gap. Further, across different testing protocols, the proposed method consistently outperforms the baseline method MSMT, which clearly shows the effectiveness of our proposed Siamese reconstruction based regularization for pose-invariant feature representation.

3.2. Control Experiments with P2 on MultiPIE

The P2 testing protocol utilizes all the 0° images as the gallery. The performance is expected to be better than that reported on P1 protocol in the main submission since more images are used for reference. There is no standard deviation in this experiment as the gallery is fixed by using all the frontal images. The results are shown in Table 3, which confirms the conclusion that the proposed feature reconstruction based regularization is effective in obtaining pose-invariant and highly discriminative feature representations for face recognition.

3.3. Recognition Accuracy on LFW

We also carried out additional experiments on LFW [2]. As we know, LFW contains mostly near-frontal faces. To better reveal the contribution of our method designed to regularize pose variations, we compare the performance with respect to statistics of pose range (correct pairs num. / total pairs num. in the range). Table 4 shows the results. Our approach outperforms VGG-Face especially in non-frontal settings (¿30), which demonstrates the effectiveness of the proposed method in handling pose variations.

3.4. Feature Embedding of MultiPIE

Figure 1 shows t-SNE visualization [5] of VGGFace [4] feature space and the proposed reconstruction-based disentangling feature space of MultiPIE [1]. For visualization clarity, we only visualize 10 randomly selected subjects from the test set with 0°, 30°, 60°, and 90° yaw angles. Figure 1 (a) shows that samples from VGGFace feature embedding have large overlap among different subjects. In contrast, Figure 1 (b) shows that our approach can tightly cluster samples of the same subject together which leads to little overlap of different subjects, since identity features have been disentangled from pose in this case.

3.5. Feature Embedding of 300WLP

Figure 2 shows t-SNE visualization [5] of VGGFace [4] feature space and the proposed reconstruction-based disentangling feature space, with 10 subjects from 300WLP [7]. Similar to the results of MultiPIE [1], the VGGFace feature embedding space shows entanglement between identity and the pose, i.e., the man with the phone in 45° view is overlapped with the frontal view image of other persons. In contrast, feature embeddings of our method are largely separated from one subject to another, while feature embeddings

of the same subject are clustered together even there are extensive pose variations.

3.6. Probe and Gallery Examples

In Figure 3, we show examples of gallery and probe images that are used in testing. Figure 3 (a) shows the gallery images in 0° from MultiPIE. Each subject only has one frontal image for reference. Figure 3 (b) shows probe images of various pose and expression from MultiPIE. Each subject presents all possible poses and expressions such as neutral, happy, surprise, etc. The illumination is controlled with plain front lighting. Figure 3 (c) shows the gallery images from 300WLP, with two near-frontal images of each subject randomly selected. Figure 3 (d) shows all poses of the same subject from 300WLP.

3.7. Failure cases in MultiPIE and 300WLP

In Figure 4, we show the typical failure cases generated by the proposed method on both MultiPIE and 300WLP. For MultiPIE, the most challenging cases come from exaggerated expression variations, e.g. Figure 4 (a), the second row. For 300WLP, the challenge mostly come from head pose variations and illumination variations. However, images in most failure pairs are visually similar.

References

[1] R. Gross, I. Matthew, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 2009. 2, 4, 5, 6

[2] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 2

[3] M. Kan, S. Shan, and X. Chen. Multi-view deep network for cross-view classification. In *CVPR*, 2016. 2

[4] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. 1, 2, 4

[5] L. van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 2014. 2

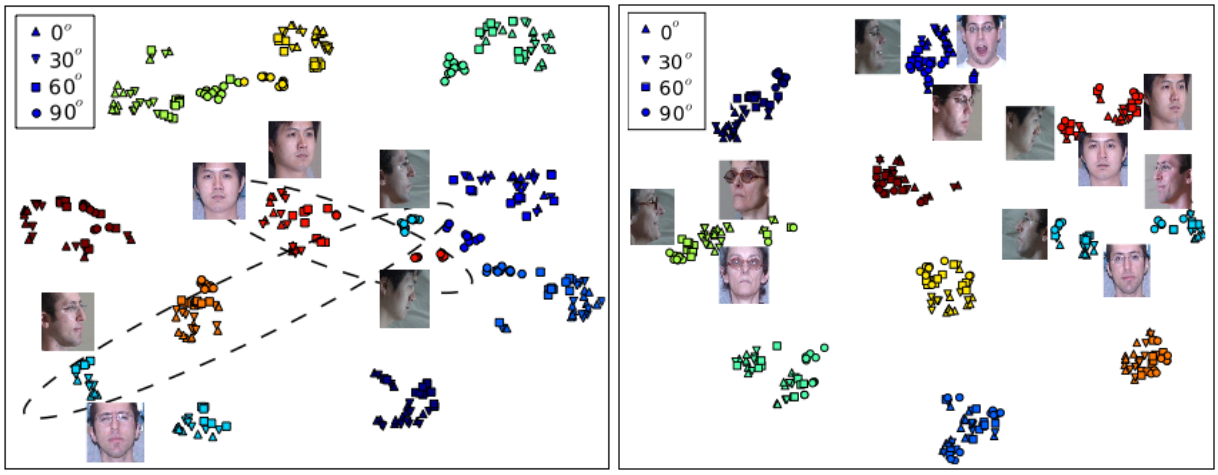
[6] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. In *CoRR*, 2014. 1, 2

[7] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016. 2, 4

[8] 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. 1, 2, 5, 6

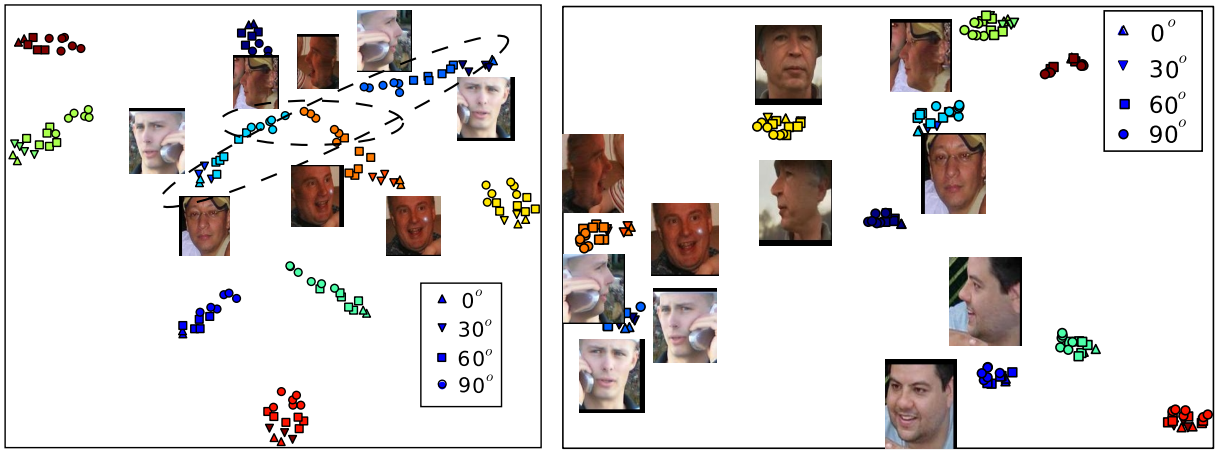
Method	LFW				
	$0-30^\circ$	$30-45^\circ$	$45-60^\circ$	$60-90^\circ$	$>30^\circ inavgerage$
VGG-Face	0.973 (5304/5524)	0.967 (410/424)	0.961 (49/51)	1.00 (1/1)	0.964
Ours	**0.986 (5445/5524)**	**0.981 (416/424)**	**1.00 (51/51)**	1.00 (1/1)	**0.983**

Table 4. Pose-wise recognition accuracy on LFW (correct pairs num. / total pairs num. in the range).



(a) VGGFace Feature Space (b) Reconstruction-based Disentangling Feature Space

Figure 1. t-SNE visualization of VGGFace [4] feature space (left) and the proposed reconstruction-based disentangling feature space (right), with 10 subjects from MultiPIE [1]. The same marker color indicates the same subject. Different marker shapes indicate different head poses. Our approach shows better results in disentangling pose factors from identity representations.

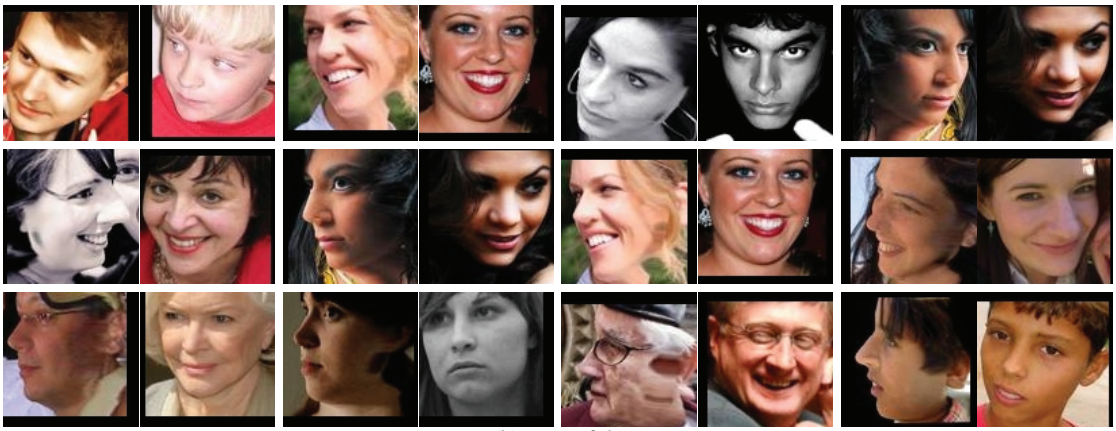


(a) VGGFace Feature Space (b) Reconstruction-based Disentangling Feature Space

Figure 2. t-SNE visualization of VGGFace [4] feature space (left) and the proposed reconstruction-based disentangling feature space (right), with 10 subjects from 300WLP [7]. The same marker color indicates the same subject. Different marker shapes indicate different head poses. Our approach shows better results in disentangling pose factors from identity representations.

Figure 3. The gallery and probe samples adopted in the testing from MultiPIE [1] and 300WLP [8]. (a) The gallery samples of MultiPIE. (b) The probe samples of MultiPIE. (c) The gallery samples of 300WLP. (d) The probe samples of 300WLP.

(a) MultiPIE failure cases



(b) 300WLP failure

Figure 4. Some failure cases in MultiPIE [1] and 300WLP [8]. Each case consists of a pair of images. The gallery image is on the left and the probe image is on the right. In both (a) and (b), the first row shows cases of 15° and 30°, the second row shows cases of 45° and 60°, and the third row shows cases of 75° and 90°. (b) follows the same layout as (a). In MultiPIE, most failures result from extensive expressions. In 300WLP, most failures results from the large pose and illumination changes. Images in most failure pairs are visually similar.