Hierarchical Feature-Pair Relation Networks for Face Recognition

Bong-Nam Kang[1,2], Yonghyun Kim[1], Bongjin Jun[2], and Daijin Kim[1]
[1]Department of Computer Science and Engineering, POSTECH, Korea
[2]Stradvision, Inc., Korea
{bnkang, gkyh0805, dkim}@postech.ac.kr, bongjin.jun@stradvision.com

Abstract

We propose a novel face recognition method using a Hierarchical Feature Relational Network (HFRN) which extracts facial part representations around facial landmark points, and predicts hierarchical latent relations between facial part representations. These hierarchical latent relations should be unique relations within the same identity and discriminative relations among different identities for face recognition task. To do this, the HFRN extracts appearance features as facial parts representations around facial landmark points on the feature maps, globally pool these extracted appearance features onto single feature vectors, and captures the relations for the pairs of appearance features. The HFRN captures the locally detailed relations in the low-level layers and the locally abstracted global relations in the high-level layers for the pairs of appearance features extracted around facial landmark points projected on each layer, respectively. These relations from low-level layers to high-level layers are concatenated into a single hierarchical relation feature. To further improve the accuracy of face recognition, we combine the global appearance feature with the hierarchical relation feature. In experiments, the proposed method achieves the comparable performance in the 1:1 face verification and 1:N face identification tasks compared to existing state-of-the-art methods on the challenging IARPA Janus Benchmark A (IJB-A) and IARPA Janus Benchmark B (IJB-B) datasets.

1. Introduction

Face recognition in unconstrained environments is a challenging problem in computer vision society. Faces of the same identity can look very different when presented in different illuminations, facial poses, facial expressions, and occlusions. Such variations within the same identity could overwhelm the variations due to identity differences and make face recognition challenging. To overcome these problems, many deep learning-based approaches have been proposed as the feature learning and achieved high accuracies of face recognition such as DeepFace [26], DeepID series [24, 23, 35], FaceNet [22], PIMNet [10], SphereFace [17], ArcFace [4], and PRN [11].

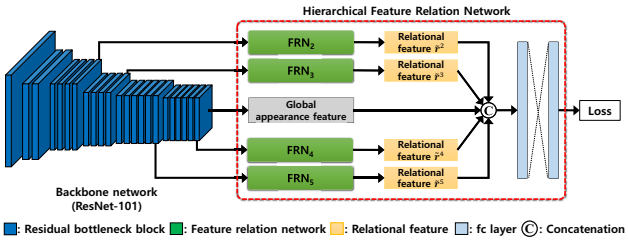


Figure 1. Overview of the proposed Hierarchical Feature Relational Network.

The deep learned features need to be not only separable but also discriminative to classify face images among different identities. This means that the representation of a certain person A for face recognition stays unchanged regardless of who it is compared with, and this representation has to be discriminative enough to distinguish A from all other persons. However, these features are learned implicitly for separable and distinct representations to classify among different identities without what part of the feature is meaningful, and what part of the features is separable and discriminative. To do this, some research efforts have been made regarding facial part-based representations for face recognition. In DeepID [24] and DeepID2 [23], a face region is divided into several of sub-regions based on the detected facial landmark points at different scales and color channels, then these regions are used for training different networks. The comparator network [32] used attention mechanism based on multiple discriminative local regions (landmark points based), and comparing local descriptors between pairs of faces. In [5], they proposed contrastive convolution which specifically focuses on the distinct characteristics between the two faces to compare, i.e., those contrastive characteristics. By contrast, when humans compare two faces, we are trying to find the differences and putting more attention of them for better distinguishing of the two faces.

In this paper, we propose the hierarchical feature relational network (HFRN) to represent unique and discriminative representations from locally detailed information to high-level abstracted global information for face recognition. To do this, the HFRN predicts latent relations for the pairs of appearance features on each feature map, and is trained to capture unique relations within the same identity and discriminative relations among different identities. These relations are captured from the low-level appearance features on the low-level feature map to the high-level appearance features on the high-level feature map. In low-level layers, the feature maps represent locally detailed information. The HFRN captures the locally detailed relations for pairs of appearance features extracted around facial landmark points on the low-level feature maps. In contrast, the feature maps in high-level layers represent abstract and global information. Therefore, the HFRN captures the more abstracted and global relations for pairs of appearance features extracted around facial landmark points on the high-level feature maps. These relations from low-level layers to high-level layers are concatenated into a single hierarchical relation feature. To further improve accuracy of face recognition, we combine the global appearance feature with the hierarchical relation feature (Figure 1).

The main contributions of this paper can be summarized as follows:

- We propose a novel face recognition method using the hierarchical feature relational network (HFRN) which captures the unique and discriminative pairwise relations from low-level appearance features to high-level appearance feature to classify face images among different identities. These relational features represent locally detailed relations in low-level layers and locally high-level abstracted relational features in high-level layers.

- To further increase accuracy, we combine hierarchical relational features with the global appearance features which are extracted from the last convolutional layer and contain global information of a given face.

- We show that the proposed HFRN is very useful to increase the accuracy of both face verification and face identification.

- To investigate the effectiveness of the HFRN, we present extensive experiments on the public available datasets such as Labeled Faces in the Wild (LFW) [8], YouTube Faces (YTF) [31], IARPA Janus Benchmark-A (IJB-A) [12], and IARPA Janus Benchmark-B (IJB-B) [30].

2. Related Work

Part-based face recognition. Several previous researches proposed to use part-based representation for a face image. In [15], the face image is densely divided into overlapping patch regions at multiple scales, and each region of patches is represented by local features such as Local Binary Pattern (LBP) or SIFT, then represented as a bag of spatial appearance features by clustering. In DeepId [24], ten different regions in a given face image are defined by five large regions at fixed positions and five small regions around each facial landmark point, and these regions are then cropped, respectively. For each region, RGB and gray-scale patch regions of five different scales were generated and each trained with a single convolutional neural network to output a feature vector of 160 dimensions. The features are then concatenated and the dimensionality was reduced with additional training on a validation set. In DeepID2 [23], 400 patch regions were cropped at different positions, scales, color channels and horizontal flipping, and used for training 200 different convolutional networks. After feature selection, 25 patches were selected to extract a 4,000-dimensional feature vector, which was finally reduced to 180-dimensional vector by PCA. The authors showed that combining these features from different regions substantially improved the accuracy of face recognition. In this paper, unlike the DeepID methods, we crop facial part regions and extract appearance features on the feature maps from low-level layers to high-level layers within the projected ROIs around facial landmark points, and capture hierarchical feature relations between facial parts.

Relation learning. Google's DeepMind proposed the relational network to perform spatial relational reasoning by modeling the relations between features at every spatial location and the features at every other location [21]. To model the co-occurrence statistics of features, a bilinear CNN [16] was proposed for fine-grained classification problems. The bilinear CNN makes the descriptor of an image by the outer product of the feature maps. As for few-shot learning, [28] was proposed to learn a local similarity metric with a deep neural network. As its extension, [25] was proposed, and experiments with models with more capacity, where the feature maps of images (from a support set and test set) are concatenated and fed into a relational network module for similarity learning. In this paper, we modify the relation network with appearance features from low-level layers to high-level layers and the global appearance feature to represent hierarchical feature relations. It shows that the relation learning helps to improve the accuracy of the face verification and identification tasks.

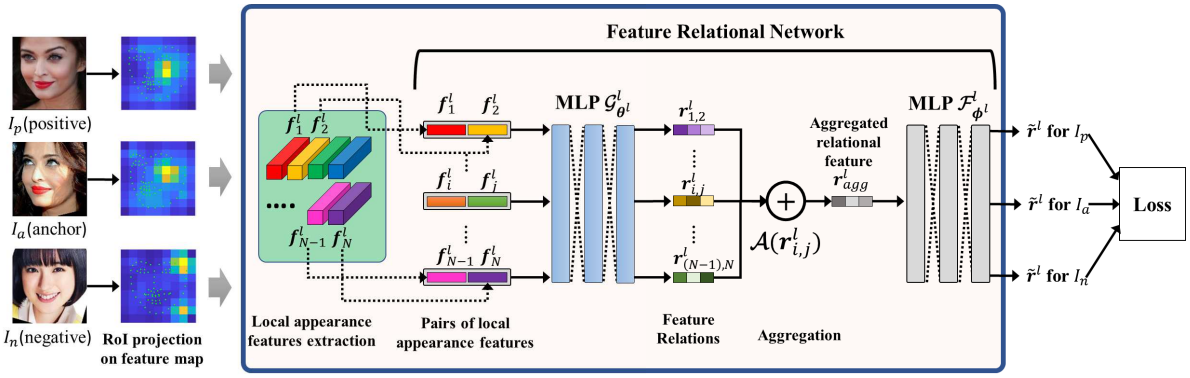


Figure 2. Details of the Feature Relational Network in the proposed hierarchical feature relational network.

3. Hierarchical Feature Relational Network

The Hierarchical Feature-pair Relation Network (HFRN) captures the latent feature-pair relations for pairs of appearance features on each feature maps and consists of the facial part representations for feature extraction, the feature-pair relation network for capturing feature-pair relations, and the hierarchical feature-pair relation network for obtaining the hierarchical feature-pair relation.

3.1. Facial Part Representations

To capture the feature-pair relations between facial parts, we first extract the appearance features as the facial part representation around each i-th landmark point. Each $m \times m$ region of interest (RoI) corresponding to each i-th facial landmark point in the input image is projected onto $m^{'} \times m^{'}$ region on the feature map of l-th layer. Within $m^{'} \times m^{'}$ RoIs, we extract appearance features, and each of them is then pooled into a single appearance feature by global average pooling (GAP) (Figure 3):

$$\boldsymbol{f}_i^l = GAP(\boldsymbol{A}_i^l), \tag{1}$$

where $\boldsymbol{A}_i^l$ is i-th $m^l \times m^l \times c^l$ appearance feature corresponding to i-th facial landmark point on the feature map in l-th layer. c^l denotes the dimension of channels for feature map in l-th layer. Because $\boldsymbol{A}_i^l$ is extracted by the RoI projection, the location of the projected RoI are continuous coordinates which mean that coordinates of RoI are floating-numbers. To extract the exact value of each sampling point in the $m^{'} \times m^{'}$ region, we compute directly through bilinear interpolation from the nearby grid points on the feature map likely RoI-Align [6]. After appearance feature extraction, we apply the GAP to each $\boldsymbol{A}_i^l$, and then we obtain the $1\times1\times c^l$ dimensional feature $\boldsymbol{f}_i^l$. We totally extract N $\boldsymbol{f}_i^l$s ($\boldsymbol{F}^l = \{\boldsymbol{f}_1^l, \cdots, \boldsymbol{f}_i^l, \cdots, \boldsymbol{f}_N^l\}$) in each l-th layer. Figure 3 illustrates the process of the appearance feature extraction. With this appearance feature set $\boldsymbol{F}^l$, we make all possible pairs $\boldsymbol{P}^l = \{\boldsymbol{p}_{1,2}^l, \cdots, \boldsymbol{p}_{i,j}^l, \cdots, \boldsymbol{p}_{N-1,N}^l\}$ in the l-th layer. Each $\boldsymbol{p}_{i,j}^l$ is a pair of two appearance feature $\boldsymbol{f}_i^l$ and $\boldsymbol{f}_j^l$ which are i-th and j-th appearance features corresponding to each facial landmark point in l-th layer, respectively. Using these pairs of appearance features $\boldsymbol{P}^l$, we capture the feature-pair relations for face recognition.

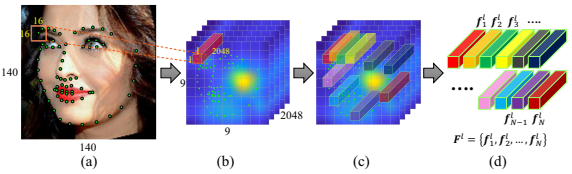


Figure 3. Example of appearance feature extraction in *conv5_3* residual bottleneck block. 16×16 RoI around i-th facial landmark point is projected into 1×1 RoI on the feature map in *conv5_3*. In this RoI, we extract appearance feature followed by global averaging pooling, we obtain $1 \times 1 \times 2,048$ dimensional appearance feature. We extract 68 appearance features on each feature map in each l-th layer.

3.2. Feature Relation Network

The feature relation network (FRN) captures the unique and discriminative relations of pairs of appearance features extracted from projected RoI on the feature map to classify identities. The relation feature $\boldsymbol{r}_{i,j}^l$ in l-th layer represents a latent relation of a pair of two appearance features $\boldsymbol{f}_i^l$ and $\boldsymbol{f}_j^l$, and can be written as follows:

$$\boldsymbol{r}_{i,j}^l = \mathcal{G}_{\boldsymbol{\theta}^l}^l(\boldsymbol{p}_{i,j}^l), \tag{2}$$

where $\mathcal{G}_{\boldsymbol{\theta}^l}^l$ is a multi-layer perceptron (MLP) and its parameters $\boldsymbol{\theta}^l$ are learnable weights for relations of l-th layer. $\mathcal{G}_{\boldsymbol{\theta}^l}^l$ takes $\boldsymbol{p}_{i,j}^l$ as input and outputs a relation $\boldsymbol{r}_{i,j}^l$ between $\boldsymbol{f}_i^l$ and $\boldsymbol{f}_j^l$. The same MLP operates on all possible parings $\boldsymbol{P}^l$ of appearance features $\boldsymbol{F}^l$.

In pairing of appearance features, the permutation order of appearance features is a critical to capture unique and discriminative relations. Without this permutation order invariance, we would have to learn to operate on all possible permuted pairs of appearance features in $\boldsymbol{F}^l$. To incor-

porate this permutation order invariance, we constrain the FRN with an aggregation function:

$$\boldsymbol{r}^l_{agg} = \mathcal{A}(\boldsymbol{R}^l) = \sum_{\boldsymbol{r}^l_{i,j} \in \boldsymbol{R}^l} \boldsymbol{r}^l_{i,j}, \tag{3}$$

where $\mathcal{A}$ is a summation as the aggregation function which summates all possible relations $\boldsymbol{R}^l$, where $\boldsymbol{R}^l$ is $\{\boldsymbol{r}^l_{1,2}, \cdots, \boldsymbol{r}^l_{i,j}, \cdots, \boldsymbol{r}^l_{N-1,N}\}$, among all possible pairs $\boldsymbol{P}^l$ of appearance features $\boldsymbol{F}^l$ in the l-th layer. After aggregation, we can obtain an aggregated relational feature $\boldsymbol{r}^l_{agg}$ in l-th layer. Finally, a prediction $\tilde{\boldsymbol{r}}^l$ of relational feature of the FRN in l-th layer can be performed with:

$$\tilde{\boldsymbol{r}}^l = \mathcal{F}^l_{\boldsymbol{\phi}^l}(\boldsymbol{r}^l_{agg}), \tag{4}$$

where $\mathcal{F}^l_{\boldsymbol{\phi}^l}$ is a function with its learnable parameters $\boldsymbol{\phi}^l$, and is implemented by the MLP. Therefore, the final form of the FRN in each l-th layer is a composite function as follows:

$$FRN_l(\boldsymbol{P}^l) = \tilde{\boldsymbol{r}}^l = \mathcal{F}^l_{\boldsymbol{\phi}^l}(\mathcal{A}(\mathcal{G}^l_{\boldsymbol{\theta}^l}(\boldsymbol{p}^l_{i,j}))). \tag{5}$$

Figure 2 shows the details of the feature relational network in the proposed HFRN.

3.3. Hierarchical Feature Relations

All of operations above mentioned are applied to the feature maps in each l-th layer. Therefore, we can obtain L predictions of relational features from all L feature maps. After the last convolution layer, we concatenate all predictions of relational features into a single relational feature $\boldsymbol{r}_c$:

$$\boldsymbol{r}_c = \|^L_{l=l_0} \tilde{\boldsymbol{r}}^l, \tag{6}$$

where $\|$ denotes the concatenation of vectors from $\tilde{\boldsymbol{r}}^{l_0}$ in l_0-th low-level layer to $\tilde{\boldsymbol{r}}^L$ in L-th high-level layer. To obtain the hierarchical relational feature $\boldsymbol{r}_h$, this concatenated relational feature is fed into the fully connected layer with $1,024$ units followed by cross-entropy loss with softmax.

$$HFRN = \boldsymbol{r}_h = \mathcal{H}_{\boldsymbol{\psi}}(\boldsymbol{r}_c), \tag{7}$$

where $\mathcal{H}_{\boldsymbol{\psi}}$ is the two-layered MLP with its learnable parameters $\boldsymbol{\psi}$. All of operation above mentioned including the concatenation operation in the HFRN can be summarized as Algorithm 1.

3.4. Loss function

To learn the each FRN$_l$, we use jointly the triplet ratio loss L_t, pairwise loss L_p, and identity preserving (*softmax*) loss L_{id} [10] to minimize distances between faces that have the same identity and to maximize distances between faces that are of different identity.

$$L = \lambda_1 L_t + \lambda_2 L_p + \lambda_3 L_{id}. \tag{8}$$

Algorithm 1: Procedure of the proposed HFRN

Result: hierarchial relational feature $\boldsymbol{r}_h$
Input: N facial landmark points $\boldsymbol{s}$
Input: l_0: start layer (or layer block) index for HFRN
Input: L: the number of layers
Input: N: the number of facial landmark points
for $l \leftarrow l_0$, L **do**
  for $i \leftarrow 1$, N **do**
    Extract $\boldsymbol{A}^l_i$ corresponding to i-th facial landmark point;
    Obtain $\boldsymbol{f}^l_i$ for $\boldsymbol{A}^l_i$ using GAP (Eq. (1));
  end
  Make all possible pairs $\boldsymbol{P}^l$ between $\boldsymbol{f}^l_i$ and $\boldsymbol{f}^l_j$ in $\boldsymbol{F}^l$;
  foreach $\boldsymbol{p}^l_{i,j} \in \boldsymbol{P}^l$ **do**
    Compute the relation $\boldsymbol{r}^l_{i,j}$ of $\boldsymbol{p}^l_{i,j}$ using Eq. (2);
  end
  Aggregate all relations $\boldsymbol{R}^l$ into $\boldsymbol{r}^l_{agg}$ using Eq. (3);
  Predict relational fature $\tilde{\boldsymbol{r}}^l$ using Eq. (4);
end
Concatenate all $\tilde{\boldsymbol{r}}^l$ into $\boldsymbol{r}_c$ using Eq. (6);
Obtain the hierarchical relational feature $\boldsymbol{r}_h$ using Eq. (7);

Triplet ratio loss L_t is defined to maximize the ratio of distances between positive and negative pairs in the triplets of faces. To maximize L_t, the Euclidean distances of positive pairs should be minimized and those of negative pairs should be maximized. Let $F(I) \in \mathbb{R}^d$, where I is an input image, denote the output of a network (in the FRN, the output of $\mathcal{F}_{\boldsymbol{\phi}}$), the L_t is defined as follows:

$$L_t = \sum_{\forall T} \max\left(0, 1 - \frac{\|F(I_a) - F(I_n)\|_2}{\|F(I_a) - F(I_p)\|_2 + m}\right), \tag{9}$$

where $F(I_a)$ is the network output of an anchor face I_a, $F(I_p)$ is the network output of a positive face image I_p, and $F(I_n)$ is the network output of a negative face I_n in the triplets of faces T, respectively. m is a minimum ratio margin in Euclidean space. From recent work [10] by Kang *et al.*, they reported that an unbalanced range of distances measured between the pairs of data using only L_t; this result means that although the ratio of the distances is bounded in a certain range of values, the range of the absolute distances is not. To solve this problem, the pairwise loss function L_p is added to constrain L_t. Pairwise loss L_p is defined to minimize the sum of the squared Euclidean distances between $F(I_a)$ and $F(I_p)$. These pairs of I_a and I_p are in the triplets of faces T.

$$L_p = \sum_{(I_a, I_p) \in T} \|F(I_a) - F(I_p)\|^2_2. \tag{10}$$

The joint training with L_t and L_p minimizes the absolute Euclidean distance between face images of a given pair in

Table 1. The architecture of the backbone network in the proposed HFRN.

Layer name	Output size	Filter (kernel, #, stride)
conv1	140×140	5×5, 64, 1
pool	70×70	3×3 max pool, -, 2
conv2_x	70×70	$[(1 \times 1, 64), (3 \times 3, 64), (1 \times 1, 256)] \times 3$
conv3_x	35×35	$[(1 \times 1, 128), (3 \times 3, 128), (1 \times 1, 512)] \times 4$
conv4_x	18×18	$[(1 \times 1, 256), (3 \times 3, 256), (1 \times 1, 1024)] \times 23$
conv5_x	9×9	$[(1 \times 1, 512), (3 \times 3, 512), (1 \times 1, 2048)] \times 3$
	1×1	global average pool, 8630-d fc, softmax

the triplets of faces T. We also use these loss functions with the identity preserving loss (*softmax* loss) L_{id} jointly.

4. Experiments

4.1. Training Dataset

We use the VGGFace2 [2] dataset which has 3.2M face images for training set from 8,631 unique persons. We detect face regions and their facial landmark points by using the multi-view face detector [36] and cascade facial landmark point detector [13]. When detection of face regions or facial landmark detector is failed, we discard that images. Thus, we remove 24,160 face images from 6,561 subjects. After removing them, we have roughly 3.1M face images of 8,630 unique persons. We divide this refined dataset into two splits: one for training set having roughly 2.8M face images, and one for validation set with 311,773 face images which are selected randomly about 10% from each subject in refined dataset. We use 68 facial landmark points for the face alignment and extraction of appearance features. All of faces in both the training and validation sets are aligned to canonical faces by using the face alignment method [11]. The faces with 140×140 resolutions are used and each pixel is normalized by dividing 255 to be in a range of $[0, 1]$.

4.2. Implementation details

Backbone network. We use the modified ResNet-101 [7] as a backbone network. Our backbone network takes the RGB values of the aligned face image with 140×140 resolution as its input, and has 64 5×5 convolution filters with a stride of 1 in the first layer. After 3×3 max pooling with a stride of 2, it has several 3-layer residual bottleneck blocks. In the last layer, we use the global average pooling with 9×9 filter in each channel and use the fully connected layer. The output of the fully connected layer are fed into *softmax* loss layer (Table 1).

Detailed settings in HFRN. We first extract a set of appearance features $\boldsymbol{A}^l = \{\boldsymbol{A}_1^l, \cdots, \boldsymbol{A}_i^l, \cdots, \boldsymbol{A}_{68}^l\}$ within each projected local region around 68 facial landmark points by RoI projection on each feature map in l-th residual bottleneck block. We set RoI to 16×16 resolution around each landmark point in input facial image space, and project each RoI onto the $70 \times 70 \times 256$ (*conv2_3*), $35 \times 35 \times 512$ (*conv3_4*), $18 \times 18 \times 1,024$ (*conv4_23*), and $9 \times 9 \times 2,048$ (*conv5_3*) feature maps, respectively (Table 1). Projected RoIs on each residual bottleneck block have 8×8, 4×4, 2×2, and 1×1 resolutions, respectively. Therefore, we obtain $\boldsymbol{A}_i^2 \in \mathbb{R}^{8 \times 8 \times 256}$, $\boldsymbol{A}_i^3 \in \mathbb{R}^{4 \times 4 \times 512}$, $\boldsymbol{A}_i^4 \in \mathbb{R}^{2 \times 2 \times 1,024}$, and $\boldsymbol{A}_i^5 \in \mathbb{R}^{1 \times 1 \times 2,048}$ from each l-th residual bottleneck block. These appearance features $\boldsymbol{A}^l$ are then pooled into $1 \times 1 \times c^l$ dimensional features $\boldsymbol{F}^l = \{\boldsymbol{f}_1^l, \cdots, \boldsymbol{f}_i^l, \cdots, \boldsymbol{f}_{68}^l\}$ per each $\boldsymbol{A}^l$ by GAP. We make $2,278$ ($= {}^{68}C_2$) possible pairs of appearance features per each $\boldsymbol{F}^l$. Then we use three-layered MLP consisting of $1,000$ units per layer with batch normalization (BN) [9] and rectified linear unit (ReLU) [19] non-linear activation functions for each $\mathcal{G}_{\boldsymbol{\theta}^l}^l$, and three-layered MLP consisting of $1,000$ units per layer with BN and ReLU non-linear activation functions for each $\mathcal{F}_{\boldsymbol{\phi}^l}^l$. To aggregate all of relations from $\mathcal{G}_{\boldsymbol{\theta}^l}^l$, we use summation as an aggregation function.

Each FRN$_l$ for l-th residual bottleneck block in HFRN is optimized by triplet ratio L_t, pairwise L_p, and identity preserving L_{id} loss functions [11] over the ground-truth identity labels using stochastic gradient descent (SGD) optimization method with learning rate 0.1. For weights of loss functions, we set $\lambda_1 = 1$, $\lambda_2 = 0.5$, and $\lambda_3 = 1$ by a grid search, and achieve the best results.

To obtain the hierarchical relational feature $\boldsymbol{r}_h$, we concatenate all predictions of relational features $\tilde{\boldsymbol{r}}^l$ into a single feature $\boldsymbol{r}_c \in \mathbb{R}^{4,000}$, then $\boldsymbol{r}_c$ is fed into two-layered MLP with 1,204 units per layer followed by cross-entropy loss with softmax. We used mini-batch size of 128 on four NVIDIA Titan X GPUs. During training, we froze the backbone CNN model to only update weights of the HFRN.

4.3. Ablation Study

We conduct a number of ablation experiments to analyze the proposed HFRN on the LFW [8] and YTF [31]. Following the test protocol of *unrestricted with labeled outside data* [14], we test on the LFW and YTF by using a squared L_2 distance threshold to determine classification of *same* and *different*, and report the results (Table 2), and discussed in detail next.

Architecture. Table 2 shows various configurations of HFRN models. HFRN$_{x:y}$ denotes the configuration of concatenation with predictions of relational features from x-th layer to y-th layer, HFRN$_x$ denotes the configuration with only predictions of relational feature of x-th layer, and $^+$ superscription denotes a combined model with global appearance feature $\boldsymbol{f}^g$ (the output of GAP in *conv5_3*) and predictions of relational features.

Effects of feature relations. The HFRN uses feature maps from the low-level layer to the high-level layer to cap-

Table 2. Effects of hierarchical feature relations on the LFW and YTF datasets.

	$\tilde{\boldsymbol{r}}^2$	$\tilde{\boldsymbol{r}}^3$	$\tilde{\boldsymbol{r}}^4$	$\tilde{\boldsymbol{r}}^5$	$\boldsymbol{f}^g$	LFW	YTF
(1) HFRN_g (baseline)					✓	99.60	95.1
(2) HFRN_2	✓					95.33	92.8
(3) HFRN_3		✓				96.75	93.6
(4) HFRN_4			✓			98.33	94.8
(5) HFRN_5				✓		99.61	95.3
(6) $\text{HFRN}_{4:5}$			✓	✓		99.71	96.0
(7) $\text{HFRN}_{3:5}$		✓	✓	✓		99.77	96.6
(8) $\text{HFRN}_{2:5}$	✓	✓	✓	✓		99.80	96.7
(9) HFRN_5^+				✓	✓	99.65	95.7
(10) $\text{HFRN}_{4:5}^+$			✓	✓	✓	99.75	96.4
(11) $\text{HFRN}_{3:5}^+$		✓	✓	✓	✓	99.81	96.7
(12) $\text{HFRN}_{2:5}^+$	✓	✓	✓	✓	✓	99.83	96.9

ture unique and discriminative relation features among different identities. We evaluate each prediction of relational features on the LFW and YTF datasets in terms of accuracy of verification. We define four different types of models such as HFRN_2, HFRN_3, HFRN_4, and HFRN_5. HFRN_2 uses only $\tilde{\boldsymbol{r}}^2$, HFRN_3 uses only $\tilde{\boldsymbol{r}}^3$, HFRN_4 uses only $\tilde{\boldsymbol{r}}^4$, and HFRN_5 uses only $\tilde{\boldsymbol{r}}^5$ to verify face images *same* and *different*. HFRN_2, HFRN_3, HFRN_4, and HFRN_5 achieve 95.33%, 96.75%, 98.33%, and 99.61% accuracies on the LFW, and achieve 92.8%, 93.6%, 94.8%, and 95.3% accuracies on the YTF, respectively (Table 2 (2)-(5) and Figure 4). From the experimental results (Table 2 (1)-(5) and Figure 4), HFRN_5 achieves slightly better accuracy of verification than the baseline HFRN_g (99.61% vs. 99.60% on the LFW, and 95.3% vs. 95.1% on the YTF).

Effects of hierarchical feature relations. To investigate the effectiveness of the proposed hierarchical feature relations, we perform experiments on the LFW and YTF datasets in terms of accuracy of verification. To do this, we define three different types of models such as $\text{HFRN}_{4:5}$, $\text{HFRN}_{3:5}$, and $\text{HFRN}_{2:5}$. An evaluation of our proposed hierarchical feature relations is shown in Table 2 (6)-(8). $\text{HFRN}_{4:5}$, $\text{HFRN}_{3:5}$, and $\text{HFRN}_{2:5}$ achieve 99.71%, 99.77%, and 99.80% accuracies on the LFW, and achieve 96.0%, 96.6%, and 96.7% accuracies on the YTF, respectively (Table 2 and Figure 4 (6)-(8)). From the experimental results (Table 2 (6)-(8) and Figure 4), we observe that the accuracy of verification increases steadily with the combination of relational features from the higher level relational feature to the lower level relational feature. $\text{HFRN}_{2:5}$ achieves better accuracy of verification than the baseline HFRN_g (99.80% vs. 99.60% on the LFW, and 96.7% vs. 95.1% on the YTF).

Fusion of HFRN and global appearance feature. To further increase the accuracy of face recognition, we combine the proposed HFRN with the global appearance feature $\boldsymbol{f}^g$. We define three different types of models such as $\text{HFRN}_{4:5}^+$, $\text{HFRN}_{3:5}^+$, and $\text{HFRN}_{2:5}^+$. $\text{HFRN}_{4:5}^+$, $\text{HFRN}_{3:5}^+$, and $\text{HFRN}_{2:5}^+$ achieve 99.75%, 99.81%, and 99.83% accuracies on the LFW, and achieve 96.4%, 96.7%, and 96.9% accuracies on the YTF, respectively (Table 2 (8)-(12) and Figure 4). From the experimental results (Table 2 and Figure 4 (9)-(12)), we observe that the combination of HFRN and the global appearance feature $\boldsymbol{f}^g$ increases the accuracy of verification. $\text{HFRN}_{2:5}^+$ achieves the comparable results with the existing *state-of-the-art* (99.83% *vs.* 99.83% (ArcFace [4]) on the LFW; 96.9% *vs.* 98.02% ArcFace [4]) on the YTF). Table 3 shows the comparison of performances of the proposed HFRN with the *state-of-the-art* methods on the LFW and YTF.

Table 3. Comparison of the number of training images, the architecture, and the accuracy of the proposed method with the *state-of-the-art* methods on the LFW and YTF.

Method	Images	Architecture	Acc. on LFW(%)	Acc. on YTF(%)
DeepFace [26]	4M	AlexNet	97.25	91.4
DeepID [24]	202,599	AlexNet	97.45	-
DeepID3 [35]	300,000	VGGNet-10	99.52	-
FaceNet [22]	200M	GoogleNet-24	99.63	95.1
CenterFace [29]	0.7M	LeNet+-7	99.28	94.9
$\text{PIMNet}_{\text{fusion}}$ [10]	198,018	GoogleNet-24	99.08	-
SphereFace [17]	494,414	ResNet-64	99.42	95.0
PRN [11]	2.8M	ResNet-100	99.76	96.3
ArcFace [4]	3.8M	ResNet-100	99.83	98.02
model A (baseline, only $\boldsymbol{f}^g$)	2.8M	ResNet-100	**99.60**	**95.1**
model B ($\text{HFRN}_{2:5}$)	2.8M	ResNet-100	**99.80**	**96.7**
model C ($\text{HFRN}_{2:5}^+$)	2.8M	ResNet-100	**99.83**	**96.9**

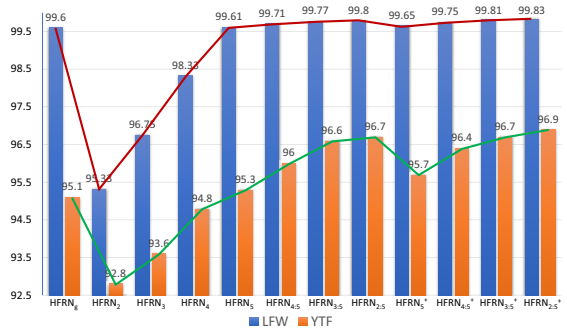


Figure 4. Effects of hierarchical feature relations on the LFW and YTF datasets.

4.4. Comparison of the State-of-the-art Methods

Detailed settings in models. For fair comparison in terms of the effects of each network module, we train three kinds of models (**model A**, **model B**, and **model C**) under the supervision of cross-entropy loss with softmax: **model A** is the backbone network model with only the global appearance feature $\boldsymbol{f}^g$ (Table 1). **model B** is the $\text{HFRN}_{2:5}$, and uses the output of $\mathcal{H}_{\psi}$. The output is the $1 \times 1 \times 1,024$ dimensional feature. **model C** is the combined model $\text{HFRN}_{2:5}^+$ which concatenates the output $\boldsymbol{f}^g$ of **model A** and the concatenated relational feature $\boldsymbol{r}_c$. $\boldsymbol{f}^g$ is the feature of size $1 \times 1 \times 2,048$ from each face image. $\boldsymbol{r}_c$ is the feature of size $1 \times 1 \times 4,000$. This combined feature is fed into $\mathcal{H}_{\psi}$.

Table 4. Comparison of performances of the proposed HFRN method with the *state-of-the-art* on the IJB-A dataset. For verification, TAR *vs.* FAR are reported. For identification, TPIR *vs.* FPIR and the Rank-N accuracies are presented.

Method	1:1 Verification TAR			1:N Identification TPIR				
	FAR=0.001	FAR=0.01	FAR=0.1	FPIR=0.01	FPIR=0.1	Rank-1	Rank-5	Rank-10
Pose-Aware Models [18]	0.652 ± 0.037	0.826 ± 0.018	-	-	-	0.840 ± 0.012	0.925 ± 0.008	0.946 ± 0.005
All-in-One [20]	0.823 ± 0.02	0.922 ± 0.01	0.976 ± 0.004	0.792 ± 0.02	0.887 ± 0.014	0.947 ± 0.008	0.988 ± 0.003	0.986 ± 0.003
NAN [33]	0.881 ± 0.011	0.941 ± 0.008	0.978 ± 0.003	0.817 ± 0.041	0.917 ± 0.009	0.958 ± 0.005	0.980 ± 0.005	0.986 ± 0.003
VGGFace2 [2]	0.904 ± 0.020	0.958 ± 0.004	0.985 ± 0.002	0.847 ± 0.051	0.930 ± 0.007	0.981 ± 0.003	0.994 ± 0.002	0.996 ± 0.001
VGGFace2_ft [2]	0.921 ± 0.014	0.968 ± 0.006	0.990 ± 0.002	0.883 ± 0.038	0.946 ± 0.004	0.982 ± 0.004	0.993 ± 0.002	0.994 ± 0.001
PRN [11]	0.901 ± 0.014	0.950 ± 0.006	0.985 ± 0.002	0.861 ± 0.038	0.931 ± 0.004	0.976 ± 0.003	0.992 ± 0.003	0.994 ± 0.003
PRN$^+$ [11]	0.919 ± 0.013	0.965 ± 0.004	0.988 ± 0.002	0.882 ± 0.038	0.941 ± 0.004	0.982 ± 0.004	0.992 ± 0.002	0.995 ± 0.001
DR-GAN [27]	0.539 ± 0.043	0.774 ± 0.027	-	-	-	0.855 ± 0.015	0.947 ± 0.011	-
DREAM [1]	0.868 ± 0.015	0.944 ± 0.009	-	-	-	0.946 ± 0.011	0.968 ± 0.010	-
DA-GAN [37]	0.930 ± 0.005	0.976 ± 0.007	0.991 ± 0.003	0.890 ± 0.039	0.949 ± 0.009	0.971 ± 0.007	0.989 ± 0.003	-
model A (baseline, only f^g)	$\mathbf{0.895 \pm 0.015}$	$\mathbf{0.949 \pm 0.008}$	$\mathbf{0.980 \pm 0.005}$	$\mathbf{0.843 \pm 0.035}$	$\mathbf{0.923 \pm 0.005}$	$\mathbf{0.975 \pm 0.005}$	$\mathbf{0.992 \pm 0.004}$	$\mathbf{0.993 \pm 0.001}$
model B (HFRN$_{2:5}$)	$\mathbf{0.923 \pm 0.013}$	$\mathbf{0.971 \pm 0.006}$	$\mathbf{0.993 \pm 0.002}$	$\mathbf{0.896 \pm 0.038}$	$\mathbf{0.953 \pm 0.004}$	$\mathbf{0.988 \pm 0.003}$	$\mathbf{0.994 \pm 0.003}$	$\mathbf{0.996 \pm 0.003}$
model C (HFRN$^+_{2:5}$)	$\mathbf{0.929 \pm 0.013}$	$\mathbf{0.975 \pm 0.004}$	$\mathbf{0.998 \pm 0.002}$	$\mathbf{0.902 \pm 0.038}$	$\mathbf{0.958 \pm 0.004}$	$\mathbf{0.992 \pm 0.004}$	$\mathbf{0.994 \pm 0.001}$	$\mathbf{0.996 \pm 0.001}$

Table 5. Comparison of performances of the proposed HFRN method with the *state-of-the-art* on the IJB-B dataset. For verification, TAR *vs.* FAR are reported. For identification, TPIR *vs.* FPIR and the Rank-N accuracies are presented.

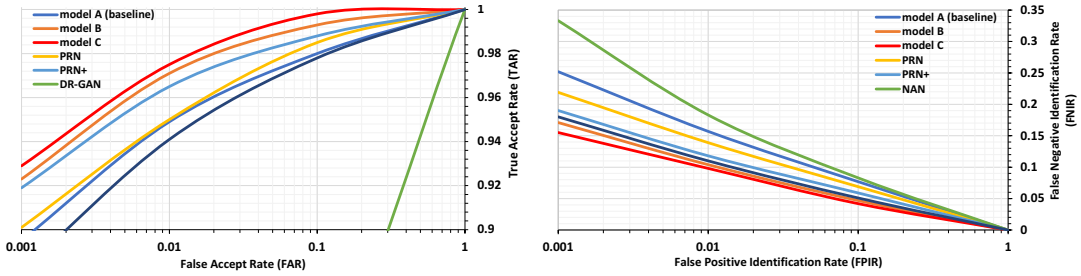
Method	1:1 Verification TAR				1:N Identification TPIR				
	FAR=0.00001	FAR=0.0001	FAR=0.001	FAR=0.01	FPIR=0.01	FPIR=0.1	Rank-1	Rank-5	Rank-10
VGGFace2 [2]	0.671	0.800	0.888	0.949	0.706 ± 0.047	0.839 ± 0.035	0.901 ± 0.030	0.945 ± 0.016	0.958 ± 0.010
VGGFace2_ft [2]	0.705	0.831	0.908	0.956	0.743 ± 0.037	0.863 ± 0.032	0.902 ± 0.036	0.946 ± 0.022	0.959 ± 0.015
FPN [3]	-	0.832	0.916	0.965	-	-	0.911	0.953	0.975
Comparator Net [32]	-	0.849	0.937	0.975	-	-	-	-	-
PRN [11]	0.692	0.829	0.910	0.956	0.773 ± 0.018	0.865 ± 0.018	0.913 ± 0.022	0.954 ± 0.010	0.965 ± 0.013
PRN$^+$ [11]	0.721	0.845	0.923	0.965	0.814 ± 0.017	0.907 ± 0.013	0.935 ± 0.015	0.965 ± 0.017	0.975 ± 0.007
model A (baseline, only f^g)	**0.673**	**0.812**	**0.892**	**0.953**	$\mathbf{0.743 \pm 0.019}$	$\mathbf{0.851 \pm 0.017}$	$\mathbf{0.911 \pm 0.017}$	$\mathbf{0.950 \pm 0.013}$	$\mathbf{0.961 \pm 0.010}$
model B (HFRN$_{2:5}$)	**0.741**	**0.869**	**0.930**	**0.966**	$\mathbf{0.833 \pm 0.018}$	$\mathbf{0.925 \pm 0.018}$	$\mathbf{0.953 \pm 0.022}$	$\mathbf{0.974 \pm 0.010}$	$\mathbf{0.975 \pm 0.007}$
model C (HFRN$^+_{2:5}$)	**0.748**	**0.875**	**0.943**	**0.975**	$\mathbf{0.844 \pm 0.017}$	$\mathbf{0.927 \pm 0.013}$	$\mathbf{0.965 \pm 0.015}$	$\mathbf{0.975 \pm 0.017}$	$\mathbf{0.976 \pm 0.007}$

All of convolution layers and fully connected layers use BN and ReLU as nonlinear activation functions.

Experiments on the IJB-A. We evaluated the proposed method on the IJB-A dataset [12] which contains face images and videos captured from unconstrained environments. It features full pose variation and wide variations in imaging conditions thus is very challenging. It contains 500 subjects with 5,397 images and 2,042 videos in total, and 11.4 images and 4.2 videos per subject on average. We detect the face regions using face detector [36] and facial landmark points using DAN landmark point detector [13], and then align the face image by using the alignment method [11].

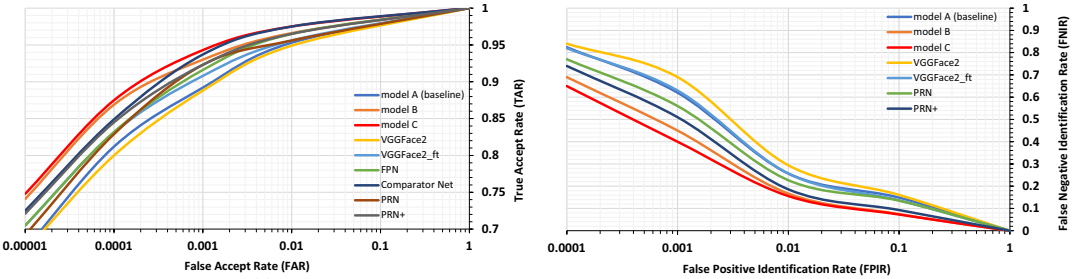
Our models such as **model A**, **model B**, and **model C** are trained on the roughly 2.8M refined VGGFace2, with no people overlapping with subjects in the IJB-A dataset. The IJB-A dataset provides 10 split evaluations with two protocols (1:1 face verification and 1:N face identification). For 1:1 face verification, we report the test results by using true accept rate (TAR) *vs.* false accept rate (FAR) (i.e. receiver operating characteristics (ROC) curve) (Table 4, Figure 5 (a)). For 1:N face identification, we report the results by using the true positive identification rate (TPIR) *vs.* false positive identification rate (FPIR) (equivalent to a decision error trade-off (DET) curve), Rank-N (Table 4, Figure 5 (b)). All measurements are based on a squared L_2 distance threshold. From the experimental results (Table 4, Figure 5), we have the following observations. First, compared to **model A**, **model B** achieves a consistently superior accuracy (TAR and TPIR) by 1.3-2.8% for TAR at FAR=0.001-0.1 in verification, 3.0-5.3% for TPIR at FPIR=0.01 and 0.1 in identification open set, and 1.3% for Rank-1 in identification close set. Second, compared to **model A**, **model C** achieves also a consistently superior accuracy (TAR and TPIR) by 1.8-3.4% for TAR at FAR=0.001-0.1 in verification, 3.5-5.9% for TPIR at FPIR=0.01 and 0.1 in identification open set, and 1.7% for Rank-1 in identification close set. Third, compared to **model B**, **model C** achieves also a consistently superior accuracy (TAR and TPIR) by 0.4-0.6% for TAR at FAR=0.001-0.1 in verification, 0.5-0.6% for TPIR at FPIR=0.01 and 0.1 in identification open set, and 0.4% for Rank-1 in identification close set. Last, more importantly, **model C** is trained from scratch, achieves comparable results compared to the state-of-the-art (DA-GAN [37]) in verification, and outperforms DA-GAN by 2.2% for Rank-1 on identification close set and 1.2% for TPIR at FPIR=0.01 in identification open set on the IJB-A. This well shows the effectiveness of the HFRN on large-scale and challenging unconstrained face recognition.

Experiments on the IJB-B. We evaluate the proposed method on the IJB-B dataset [30] which contains face images and videos captured from unconstrained environments. The IJB-B dataset is an extension of the IJB-A, having 1,845 subjects with 21.8K still images (including 11,754 face and 10,044 non-face) and 55K frames from 7,011



(a) (b)

Figure 5. Results of the IJB-A dataset (average over 10 splits). (a) ROC (higher is better); (b) DET (lower is better).



(a) (b)

Figure 6. Results of the IJB-B dataset. (a) ROC (higher is better); (b) DET (lower is better).

videos, an average of 41 images per subject. Because images in this dataset are labeled with ground truth bounding boxes, we only detect facial landmark points using DAN [13], and then align face images by using the face alignment method [11].

Our models such as **model A**, **model B**, and **model C** are trained on the roughly 2.8M refined VGGFace2 dataset, with no people overlapping with subjects in the IJB-B dataset. In particular, we use the 1:1 Baseline Verification protocol and 1:N Mixed Media Identification protocol for the IJB-B. For face verification, we report the test results by using TAR *vs.* FAR (i.e. ROC curve) (Table 5, Figure 6 (a)). For face identification, we report the results by using TPIR *vs.* FPIR (equivalent to DET curve) and Rank-N (Table 5, Figure 6 (b)). We compare our proposed methods with VGGFace2 [2], FacePoseNet (FPN) [3], and PRN [11]. All measurements are based on a squared L_2 distance threshold. From the experimental results (Table 5, Figure 6), we have the following observations. First, compared to **model A**, **model B** achieves a consistently superior accuracy (TAR and TPIR) by 1.3-6.8% for TAR at FAR=0.00001-0.01 in verification, 7.4-9.0% for TPIR at FPIR=0.01 and 0.1 in identification open set, and 4.2% for Rank-1 in identification close set. Second, compared to **model A**, **model C** achieves also a consistently superior accuracy (TAR and TPIR) by 2.2-7.5% for TAR at FAR=0.00001-0.01 in verification, 7.6-10.1% for TPIR at FPIR=0.01 and 0.1 in identification open set, and 5.4% for Rank-1 in identification close set. Third, compared to **model B**, **model C** achieves also a consistently superior accuracy (TAR and TPIR) by 0.6-1.3% for TAR at FAR=0.001-0.1 in verification, 0.2-1.1% for TPIR at FPIR=0.01 and 0.1 in identification open set, and 1.2% for Rank-1 in identification close set. Last, more importantly, **model C** is trained from scratch, outperforms the current state-of-the-art (Comparator Net [32]) by 7.4% at FAR=0.0001 in verification, and PRN$^+$ [11] by 3.0% for Rank-1 of identification close set and FPIR=0.01 in identification open set on the IJB-B. This well shows the effectiveness of the HFRN on large-scale and challenging unconstrained face recognition.

5. Conclusion

We proposed the Hierarchical Feature Relational Network (HFRN), which captured the locally detailed relations in the low-level layers and the locally abstracted global relations in the high-level layers for the pairs of appearance features extracted around facial landmark points, respectively. These relations were concatenated into a single hierarchical relation feature, then it was fed into a classification network. The proposed HFRN achieved comparable performance in both 1:1 face verification and 1:N face identification tasks compared to *state-of-the-art* methods on the IJB-A and IJB-B datasets.

Acknowledgment

This research was supported by the MSIT, Korea, under the SW Starlab support program (IITP-2017-0-00897) and the StradVision, Inc., Korea.

References

[1] K. Cao, Y. Rong, C. Li, X. Tang, and C. Change Loy. Pose-robust face recognition via deep residual equivariant mapping. In *CVPR 2018*, 2018. 7

[2] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. *CoRR*, abs/1710.08092, 2017. 5, 7, 8

[3] F. J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni. Faceposenet: Making a case for landmark-free face alignment. In *ICCVW 2017*, 2017. 7, 8

[4] J. Deng, J. Guo, and S. Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *ArXiv e-prints*, 2018. 1, 6

[5] C. Han, S. Shan, M. Kan, S. Wu, and X. Chen. Face recognition with contrastive convolution. In *ECCV 2018*, 2018. 1

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV 2017*, 2017. 3

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR 2016*, 2016. 5

[8] G. B. 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, 5

[9] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML 2015*, 2015. 5

[10] B.-N. Kang, Y. Kim, and D. Kim. Deep convolutional neural network using triplets of faces, deep ensemble, and score-level fusion for face recognition. In *CVPRW 2017*, 2017. 1, 4, 6

[11] B.-N. Kang, Y. Kim, and D. Kim. Pairwise relational networks for face recognition. In *ECCV 2018*, September 2018. 1, 5, 6, 7, 8

[12] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR 2015*, 2015. 2, 7

[13] M. Kowalski, J. Naruniec, and T. Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *CVPRW 2017*, 2017. 5, 7, 8

[14] G. B. H. E. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014. 5

[15] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *CVPR 2013*, pages 3499–3506, 2013. 2

[16] T. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *ICCV 2015*, 2015. 2

[17] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR 2017*, 2017. 1, 6

[18] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *CVPR 2016*, 2016. 7

[19] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML 2010*, 2010. 5

[20] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *FG 2017*, 2017. 7

[21] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. P. Lillicrap. A simple neural network module for relational reasoning. *CoRR*, abs/1706.01427, 2017. 2

[22] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR 2015*, 2015. 1, 6

[23] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS 2014*, 2014. 1, 2

[24] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR 2014*, 2014. 1, 2, 6

[25] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[26] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR 2014*, 2014. 1, 6

[27] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR 2017*, pages 1283–1292, 2017. 7

[28] O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, and D. Wierstra. Matching networks for one shot learning. In *NIPS 2016*, 2016. 2

[29] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV 2016*, 2016. 6

[30] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother. Iarpa janus benchmark-b face dataset. In *CVPRW 2017*, 2017. 2, 7

[31] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, 2011. 2, 5

[32] W. Xie, L. Shen, and A. Zisserman. Comparator networks. In *ECCV 2018*, September 2018. 1, 7, 8

[33] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. In *CVPR 2017*, 2017. 7

[34] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014.

[35] X. W. Yi Sun, Ding Liang and X. Tang. Deepid3: Face recognition with very deep neural networks. *CoRR*, abs/1502.00873, 2015. 1, 6

[36] J. Yoon and D. Kim. An accurate and real-time multi-view face detector using orfs and doubly domain-partitioning classifier. *Journal of Real-Time Image Processing*, Feb 2018. 5, 7

[37] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng. 3d-aided dual-agent gans for unconstrained face recognition.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018. 7