Attentional Feature-Pair Relation Networks for Accurate Face Recognition

Bong-Nam Kang[1,3], Yonghyun Kim[2,3], Bongjin Jun[1], Daijin Kim[3]
[1]StradVision, Inc. [2]Kakao Corp. [3]POSTECH
{bongnam.kang, bongjin.jun}@stradvision.com, aiden.kyh@kakaocorp.com, dkim@postech.ac.kr

Abstract

Human face recognition is one of the most important research areas in biometrics. However, the robust face recognition under a drastic change of the facial pose, expression, and illumination is a big challenging problem for its practical application. Such variations make face recognition more difficult. In this paper, we propose a novel face recognition method, called Attentional Feature-pair Relation Network (AFRN), which represents the face by the relevant pairs of local appearance block features with their attention scores. The AFRN represents the face by all possible pairs of the 9×9 local appearance block features, the importance of each pair is considered by the attention map that is obtained from the low-rank bilinear pooling, and each pair is weighted by its corresponding attention score. To increase the accuracy, we select top-K pairs of local appearance block features as relevant facial information and drop the remaining irrelevant. The weighted top-K pairs are propagated to extract the joint feature-pair relation by using bilinear attention network. In experiments, we show the effectiveness of the proposed AFRN and achieve the outstanding performance in the 1:1 face verification and 1:N face identification tasks compared to existing state-of-the-art methods on the challenging LFW, YTF, CALFW, CPLFW, CFP, AgeDB, IJB-A, IJB-B, and IJB-C datasets.

1. Introduction

Face recognition is one of the most important and interesting research areas in biometrics. However, the human appearances would be drastically changed under the unconstrained environment and the intra-person variations could overwhelm the inter-person variations, which make the face recognition difficult. Therefore, better face recognition requires for reducing the intra-person variations while enlarging the inter-person differences under the unconstrained environment.

Recent studies have targeted the same goal that minimizes the inter-person variations and maximizes the intra-person variations, either explicitly or implicitly. In deep learning-based face recognition methods, the deeply learned and embedded features are required to be not only separable but also discriminative to classify face images among different identities. This implies that the representation of a certain person *A* stays unchanged regardless of who he/she is compared with, and it has to be discriminative enough to distinguish *A* from all other persons. Chen *et al.* achieved good recognition performance [4] by extracting feature representations via the CNN. And then, those features are applied to learn metric matrix to project the feature vector into a low-dimensional space in order to maximize the between-class variation and minimize within-class variation via the joint Bayesian metric learning. Chowdhury *et al.* applied the bilinear CNN architecture [5] to the face identification task. Hassner *et al.* proposed the pooling faces [9] that aligned faces in the 3D and binned them according to head pose and image quality. Masi *et al.* proposed the pose-aware models (PAMs) [20] that handled pose variability by learning pose-aware models for frontal, half-profile, and full-profile poses to improve face recognition performance in an unconstrained environment. Sankaranarayanan *et al.* [27] proposed the triplet probabilistic embedding (TPE) that coupled a CNN-based approach with a low-dimensional discriminative embedding learned using triplet probability constraints. Crosswhite *et al.* proposed the template adaptation (TA) [6] that was a form of transfer learning to the set of media in a template, which obtained better performance than the TPE on the IJB-A dataset by combining the CNN features with template adaptation. Yang *et al.* proposed the neural aggregation network (NAN) [35] that produced a compact and fixed dimension feature representation. It adaptively aggregated the features to form a single feature inside the convex hull spanned by them and learned to advocate high-quality face images while repelling low-quality face images such as blurred, occluded and improperly exposed faces. Ranjan *et al.* [24] added an L2-constraint to the feature descriptors which restricted them to lie on a hypersphere of a fixed radius, where minimizing the softmax loss is equivalent to maximizing the cosine similarity for the positive pairs and minimizing it for the negative pairs. However, the above

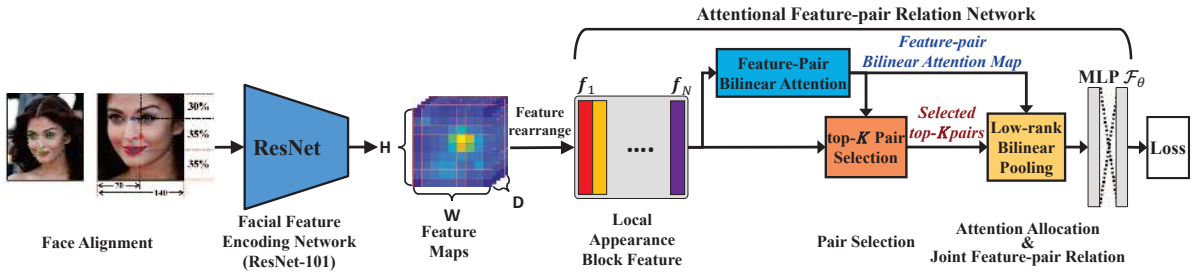


Figure 1. Working principle of the proposed Attentional Feature-pair Relation Network.

mentioned methods extracted the holistic features and did not designate what parts of the feature are meaningful and what parts of the features are separable and discriminative. Therefore, it is difficult to know what kind of features are used to discriminate the identities of face images clearly.

To overcome this disadvantage, some research efforts have been made regarding to the facial part-based representations for face recognition. In DeepID [30] and DeepID2 [29], a face region was divided into several of sub-regions using the detected facial landmark points at different scales and color channels, then these sub-regions were used for training different networks. Xie *et al.* proposed the comparator network [34] that used attention mechanism based on multiple discriminative local sub-regions, and compared local descriptors between pairs of faces. Han *et al.* [8] proposed the contrastive convolution which specifically focused on the distinct (contrastive) characteristics between two faces, where it tried to find the differences and put more attention for better discrimination of two faces. For example, the best contrastive feature for distinguishing two images of Stephen Fry and Brad Pitt might be "crooked nose". Kang *et al.* proposed the pairwise relational network (PRN) [14] that made all possible pairs of local appearance features, then each pair of local appearance features is used for capturing relational features. In addition, the PRN was constrained by the face identity state feature embedded from the LSTM-based sub-network to represent face identity. However, these methods largely were dependent on the accuracy of facial landmark detector and it did not use the importance of facial parts.

To overcome these demerits, we propose a novel face recognition method, called Attentional Feature-pair Relation Network (AFRN), which represents the face by the relevant pairs of local appearance block features with their attention scores: 1) the AFRN represents the face by all possible pairs of the 9×9 local appearance block features, 2) the importance of each pair is considered by the attention map that is obtained from the low-rank bilinear pooling, and each pair is weighted by its corresponding attention score, 3) we select top-K pairs of local appearance block features as relevant facial information and drop the remaining irrelevant, 4) The weighted top-K pairs are propagated to extract the joint feature-pair relation by using bilinear attention network. Figure 1 shows the working principle of the proposed AFRN.

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

- **Landmark free local appearance representation**: we propose a novel face recognition method using the attentional feature-pair relation network (AFRN) which represents the face by the relevant pairs of local appearance block features with their attention scores to captures the unique and discriminative feature-pair relations to classify face images among different identities.
- **Importance of pairs and removing irrelevant pairs**: to consider the importance of each pair, we compute the bilinear attention map by using the low-rank bilinear pooling, and each pair is weighted by its attention score, then we select top-K pairs of local appearance block features as relevant facial information and drop the remaining irrelevant. The weighted top-K pairs are propagated to extract the joint relational feature by using bilinear attention network.
- We show that the proposed AFRN improves effectively the accuracy of both face verification and face identification.
- To investigate the effectiveness of the AFRN, we present extensive experiments on the public available datasets such as LFW [11], YTF [33], Cross-Age LFW (CALFW), Cross-Pose LFW (CPLFW), Celebrities in Frontal-Profile in the Wild (CFP) [28], AgeDB [22], IARPA Janus Benchmark-A (IJB-A) [17], IARPA Janus Benchmark-B (IJB-B) [32], and IARPA Janus Benchmark-C (IJB-C) [21].

2. Proposed Methods

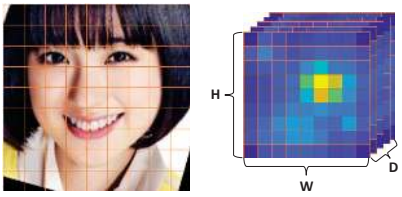
In this section, we describe the proposed methods in detail including a facial feature encoding network, attentional feature-pair relation network, top-K pairs selection and attention allocation.

2.1. Facial Feature Encoding Network

A facial feature encoding network is a backbone neural network which encodes a face image into deeply embed-

Table 1. The detailed configuration of the modified ResNet-101 for the facial feature encoding network.

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	$[(1 \times 1, 512), (3 \times 3, 512), (1 \times 1, 2048)] \times 3$



(a) (b)

Figure 2. Facial local blocks: (a) input face image. (b) facial local blocks on the feature maps.

ded features. We employ the ResNet-101 network [10] and modify it due to the differences of input resolutions, the size of convolution filters, and the size of output feature maps. A detailed architecture configuration of the modified ResNet-101 is summarized in Table 1. The non-linear activation outputs of the last convolution layer (*conv5_3*) are used as the feature maps of facial appearance representation.

2.2. Facial Local Feature Representation

The activation outputs of the convolution layer can be formulated as a tensor of the size $H \times W \times D$, where H and W denote the height and width of each feature map, and D denotes the number of channels in feature maps. Essentially, the convolution layer divides the input image into $H \times W$ sub-regions and uses D-dimensional feature maps to describe the facial part information within each sub-region. For clarity, since the activation outputs of the convolutional layer can be viewed as a 2-D array of D-dimensional features, we use each D-dimensional local appearance block feature $\boldsymbol{f}_i$ of the $H \times W$ sub-regions as the local feature representation of the i-th facial part. Based on the feature map in the *conv5_3* residual block, the face region is divided into 81 local blocks (9×9 resolution) (Figure 2), where each local block is used for the local appearance block feature of a facial part. Therefore, we extract totally 81 local appearance block features $\boldsymbol{A} = \{\boldsymbol{f}_i | i = 1, \cdots, 81\}$, where $\boldsymbol{f}_i \in \mathbb{R}^{2,048}$ in this work.

2.3. Attentional Feature-Pair Relation Network

The attentional feature-pair relation network (AFRN) is based on the low-rank bilinear pooling [15] which provides richer representations than linear models and finds attention distributions by considering every pair of features. The AFRN aims to represent a separable and discriminative

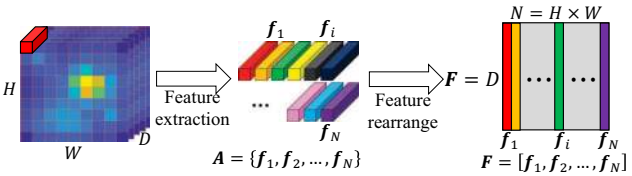


Figure 3. Facial feature rearrangement.

feature-pair relation which is pooled by feature-pair attention scores of feature-pair relations among all possible pairs of given local appearance block features. Thus, the AFRN exploits attentional feature-pair relations between all pairs of local appearance block features while extracts a joint feature-pair relation for pairs of local appearance block features.

Rearrange Local Appearance Block Features. To obtain a feature-pair bilinear attention map and a joint feature-pair relation for all of pairs of local appearance block features, we first rearrange a set of local appearance block features $\boldsymbol{A}$ into a matrix form $\boldsymbol{F}$ by stacking each local appearance block feature $\boldsymbol{f}_i$ in column direction, $\boldsymbol{F} = [\boldsymbol{f}_1, \cdots, \boldsymbol{f}_i, \cdots, \boldsymbol{f}_N] \in \mathbb{R}^{D \times N}$, where N ($= H \times W$) is the number of local appearance block features (Figure 3).

Feature-pair Bilinear Attention Map. An attention mechanism provides an efficient way to improve accuracy and reduce the number of input features at the same time by selectively utilizing given information. We adopt the feature-pair bilinear attention map $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N \times N}$. To obtain $\boldsymbol{\mathcal{A}}$, we compute a logit of the *softmax* for a pair $\boldsymbol{p}_{i,j}$ between local appearance block features $\boldsymbol{F}_i$ and $\boldsymbol{F}_j$ as:

$$\boldsymbol{\mathcal{A}}_{i,j} = \boldsymbol{p}^T \left(\sigma \left(\boldsymbol{U}'^T \boldsymbol{F}_i \right) \circ \sigma \left(\boldsymbol{V}'^T \boldsymbol{F}_j \right) \right), \quad (1)$$

where $\boldsymbol{\mathcal{A}}_{i,j}$ is the logit of the *softmax* for $\boldsymbol{p}_{i,j}$ and is the output of low-rank bilinear pooling. $\boldsymbol{U}' \in \mathbb{R}^{D \times L'}$, $\boldsymbol{V}' \in \mathbb{R}^{D \times L'}$, and $\boldsymbol{p} \in \mathbb{R}^{L'}$, where L' is the dimension of the reduced and pooled features by linear mapping $\boldsymbol{U}'$, $\boldsymbol{V}'$ and pooling $\boldsymbol{p}$ in the low-rank bilinear pooling. σ and $\circ$ denote the ReLU [23] non-linear activation function and Hadamard product (element-wise multiplication), respectively. To obtain $\boldsymbol{\mathcal{A}}$, the *softmax* function is applied element-wisely to each logit $\boldsymbol{\mathcal{A}}_{i,j}$. All above operations can be rewritten as a matrix form:

$$\boldsymbol{\mathcal{A}} = softmax \left(\left((\mathbb{1} \cdot \boldsymbol{p}^T) \circ \sigma \left(\boldsymbol{F}^T \boldsymbol{U}' \right) \right) \cdot \sigma \left(\boldsymbol{V}'^T \boldsymbol{F} \right) \right), \quad (2)$$

where $\mathbb{1} \in \mathbb{R}^N$. Figure 4 illustrates a process of the proposed feature-pair bilinear attention map.

Joint Feature-pair Relation. To extract a joint feature-

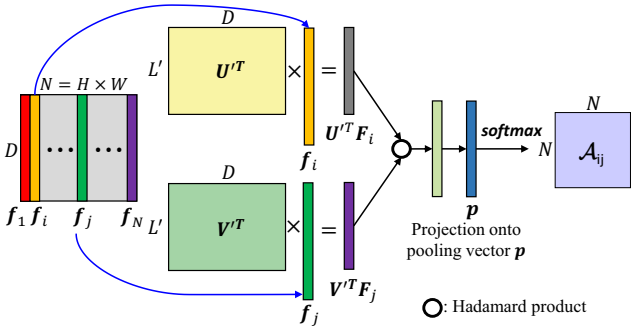


Figure 4. A process of the proposed feature-pair bilinear attention map.

pair relation for all of pairs of local appearance block features and reduce the number of pairs of local appearance block features, we use the low-rank bilinear pooling with the feature-pair bilinear attention map $\boldsymbol{\mathcal{A}}$ as:

$$\boldsymbol{r}_l^{'} = \sigma\left(\boldsymbol{F}^T\boldsymbol{U}\right)_l^T \cdot \boldsymbol{\mathcal{A}} \cdot \sigma\left(\boldsymbol{F}^T\boldsymbol{V}\right)_l, \tag{3}$$

where $\boldsymbol{U} \in \mathbb{R}^{D\times L}$ and $\boldsymbol{V} \in \mathbb{R}^{D\times L}$ are linear mappings. L is the dimension of the reduced and pooled features by pooling and linear mapping matrix $\boldsymbol{U}$ and $\boldsymbol{V}$ in the low-rank bilinear pooling for the feature-pair relation. $(\boldsymbol{F}^T\boldsymbol{U})_l \in \mathbb{R}^N$, $(\boldsymbol{F}^T\boldsymbol{V})_l \in \mathbb{R}^N$, and $\boldsymbol{r}_l^{'}$ denotes the l-th element of the intermediate feature-pair relation. The subscript l for the matrices indicates the index of column. σ denotes the ReLU [23] non-linear activation function. Eq. (3) can be viewd as a bilinear model for the pairs of local appearance block features where $\boldsymbol{\mathcal{A}}$ is a bilinear weight matrix (Figure 5). Therefore, we can rewrite Eq. (3) as:

$$\boldsymbol{r}_l^{'} = \sum_{i=1}^{N}\sum_{j=1}^{N} \boldsymbol{\mathcal{A}}_{i,j} \cdot \sigma\left(\boldsymbol{F}_i^T\boldsymbol{U}_l\right) \cdot \sigma\left(\boldsymbol{V}_l^T\boldsymbol{F}_j\right), \tag{4}$$

where $\boldsymbol{F}_i$ and $\boldsymbol{F}_j$ denote the i-th local appearance block feature and the j-the local appearance block features of input $\boldsymbol{F}$, respectively. $\boldsymbol{U}_l$ and $\boldsymbol{V}_l$ denote the l-th columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ matrices, respectively. $\boldsymbol{\mathcal{A}}_{i,j}$ denotes an element in the i-th row and j-th column of $\boldsymbol{\mathcal{A}}$.

Finally, the joint feature-pair relation $\tilde{\boldsymbol{r}}$ is obtained by projection $\boldsymbol{r}^{'}$ onto a learnable pooling matrix $\boldsymbol{P}$:

$$\tilde{\boldsymbol{r}} = \boldsymbol{P}^T\boldsymbol{r}', \tag{5}$$

where $\tilde{\boldsymbol{r}} \in \mathbb{R}^C$ and $\boldsymbol{P} \in \mathbb{R}^{L\times C}$. C is the dimension of the joint feature-pair relation by pooling $\boldsymbol{P}$ to obtain the final joint feature-pair relation $\tilde{\boldsymbol{r}}$.

2.4. Pair Selection and Attention Allocation

Only some facial part pairs are relevant to face recognition and irrelevant ones may cause over-fitting of the neural network. We need to select relevant pairs of local appearance block features, therefore we select them with top-K feature-pair bilinear attention scores as:

$$\boldsymbol{\Phi} = \left\{\boldsymbol{p}_{i,j} | \boldsymbol{\mathcal{A}}_{i,j} \text{ ranks top } K \text{ in } \boldsymbol{\mathcal{A}}\right\}, \tag{6}$$

where $\boldsymbol{p}_{i,j}$ is the selected pair of $\boldsymbol{F}_i$ and $\boldsymbol{F}_j$ with a top-K feature-pair attention score.

Different pairs of local appearance block features always have equal value scale, yet they offer different contributions on face recognition. So, we should rescale the pairs of local appearance block features to reflect their indeed influence. Mathematically, it is modeled as multiplying the corresponding feature-pair bilinear attention score. Therefore, we can substitute Eq. (4) as

$$\boldsymbol{r}_l^{'} = \sum_{k=1}^{K} \boldsymbol{\mathcal{A}}_{w_i(k),w_j(k))} \cdot \sigma\left(\boldsymbol{F}_{w_i(k)}^T\boldsymbol{U}_l\right) \cdot \sigma\left(\boldsymbol{V}_l^T\boldsymbol{F}_{w_j(k)}\right), \tag{7}$$

where $w_i(k)$ and $w_j(k)$ are i and j indexes of the k-th pair $\boldsymbol{p}_{i,j}$ in $\boldsymbol{\Phi}$. K denotes the number of the selected pairs by the pair selection layer.

Because Eq. (6) is not a differentiable function, it has no parameter to be updated and only conveys gradients from the latter layer to the former layer during back-propagation. The gradients of the selected pairs of local appearance block features will be copied from latter layer to the former layer and the gradients of the dropped pairs of local appearance block features will be discarded by setting the corresponding values to zero.

After the pair selection and attention allocation, the weighted pairs of local appearance block features are propagated the next step to extract the joint feature-pair relation. The joint feature-pair relation $\tilde{\boldsymbol{r}}$ is fed into two-layered multi-layer perceptron (MLP) $\mathcal{F}_{\boldsymbol{\theta}}$ followed by the loss function. We use the $1,024$ dimensional output vector of the last fully connected layer of $\mathcal{F}_{\boldsymbol{\theta}}$ as a final face representation.

3. Experiments

In this section, we describe the training dataset, validation set, and implementation details. We also demonstrate the effectiveness of the proposed AFRN on the LFW [11], YTF [33], IJB-A [17] and IJB-B [32] datasets.

3.1. Training Dataset

We use the VGGFace2 [2] dataset which has 3.2M face images from 8,631 unique persons. We detect face regions and their facial landmark points by using the multi-view face detector [36] and deep alignment networks (DAN) [18]. When detection is failed, we just discard that images and totally remove 24,160 face images from 6,561 subjects. Then, we have roughly 3.1M face images of 8,630 unique persons as the refined dataset. We divide this dataset into

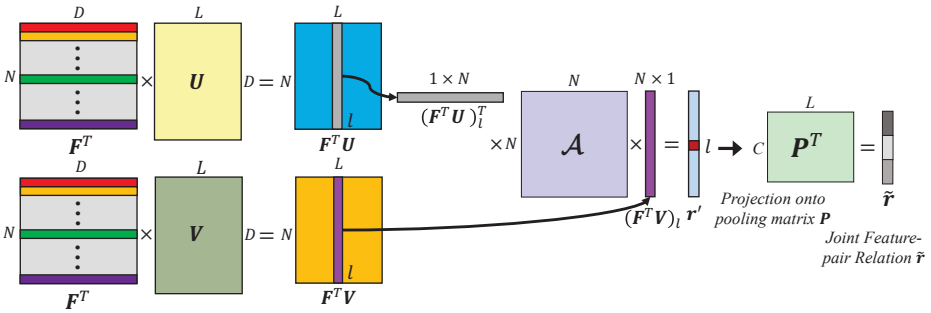


Figure 5. The joint feature-pair relation.

two sets: one for training set having roughly 2.8M face images, and another for validation set with 311,773 face images which are selected randomly about 10% from each subject. We use 68 facial landmark points for the face alignment. All of faces in both the training and validation sets are aligned to canonical faces by using the face alignment method in [14]. The faces with 140×140 resolutions are used and each pixel is normalized by dividing 255 to be in a range of $[0, 1]$.

3.2. Implementation Details

We extract 81 local appearance block features on the 9×9×2,048 feature maps in *conv5_3* residual block of the facial feature encoding network, and each local appearance block feature has 2,048 dimensions. Thus, the size of local appearance block features is $D = 2,048$ and the number of local appearance block features is $N = 81$. The size of the rearranged local appearance block features $\boldsymbol{F}$ is $\mathbb{R}^{2,048\times 81}$, the size C of the joint feature-pair relation is 1,024, which is equal to the rank L of the AFRN, and the rank $L^{'}$ of the feature-pair bilinear attention map is also $1,024$. Every linear mapping ($\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{U}^{'}$, $\boldsymbol{V}^{'}$, and $\boldsymbol{P}$) is regularized by the Weight Normalization [26]. We use the two-layered MLP consisting of $1,024$ units per layer with Batch Normalization (BN) [12] and ReLU [23] non-linear activation functions for $\mathcal{F}_{\boldsymbol{\theta}}$.

The proposed AFRN is optimized by jointly using the triplet ratio L_t, pairwise L_p, and identity preserving L_{id} loss functions proposed in [13] over the ground-truth identity labels. Adamax optimizer [16], a variant of Adam based on infinite norm, is used. The learning rate is $\min(i \times 10^{-3}, 4 \times 10^{-3})$ where i is the number of epochs starting from 1, then after 10 epochs, the learning rate is decayed by 0.25 for every 2 epochs up to 13 epochs, *i.e.* 1×10^{-3} for 11-th and 2.5×10^{-4} for 13-th epoch. We clip 2-norm of vectorized gradient to 0.25. We achieve the best results by setting the weight factors of loss functions as 1, 0.5, and 1 for L_t, L_p, and L_{id} by a grid search, respectively. We set the mini-batch size as 120 on four NVIDIA Titan X GPUs.

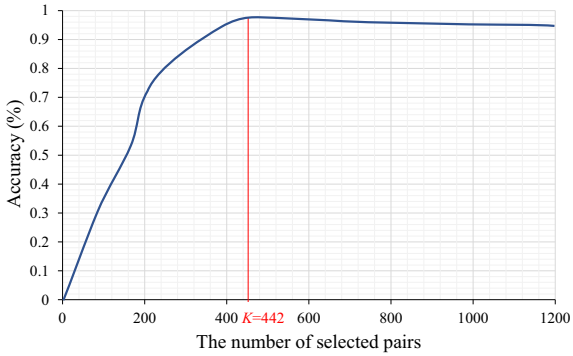


Figure 6. Accuracy plot with the different number K of feature-pair on the validation set.

3.3. Ablation Study

We conduct several experiments to analyze the proposed AFRN on the LFW [11] and YTF [33] datasets. Following the test protocol of *unrestricted with labeled outside data* [19], we test the proposed AFRN on the LFW and YTF by using a squared L_2 distance threshold to determine the classification of *same* and *different*, and report the results in Table 2 and 3, and then discuss results in detail.

Effects of Feature-pair Selection. In the feature-pair selection layer, we need to decide top-K local appearance pairs that we propagate to the next step. We perform an experiment to evaluate the effect of K. We train the AFRN model on the refined VGGFace2 training set with different value of K. The accuracy on validation set is reported in Figure 6. When K increases, the accuracy of our AFRN model increased until $K = 442$ (97.4%). After that, the accuracy of our model starts to drop. When K equals to 1,200, it is equivalent to not using the feature-pair selection layer in a face region. The performance in this case is 2.3% lower than the highest accuracy. This implies that it is important to reject irrelevant the pairs of local appearance block features.

Effects of Feature-pair Bilinear Attention. To evaluate the effects of the feature-pair bilinear attention in the proposed AFRN, we perform several experiments on the validation set, LFW and YTF datasets. We consider the atten-

Table 2. Effects of the feature-pair selection by the feature-pair bilinear attention on the validation set, LFW and YTF dataset.

Method	Val. set	LFW	YTF
(a) Baseline	94.2	99.60	95.1
(b) Feature-pair Attention w/o Pair Selection	95.1	99.71	96.1
(c) Feature-pair Attention w/ Pair Selection	**97.4**	**99.85**	**97.1**
(d) ArcFace [7]	-	99.78	-
(e) PRN [14]	-	99.76	96.3

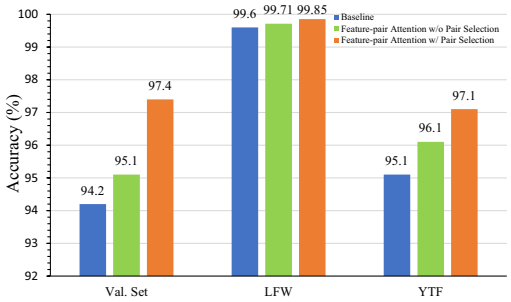


Figure 7. Effects of the feature-pair selection by the feature-pair bilinear attention on the validation set, LFW and YTF datasets.

tional feature-pair relation network without the feature-pair selection layer, which means that we use all pairs of local appearance block features for face recognition. We achieve 95.1% accuracy on the validation set, 99.71% accuracy on the LFW, and 96.1% accuracy on the YTF, respectively (Table 2 (b) and Figure 7). We use the normalized face image which include the background regions and is not cropped a face region tightly (see Figure 2). When not using pair selection, we observe that attention scores for pairs between background regions and face regions are not zero, and the accuracy is degraded in comparison with the baseline (Table 2 (a) and Figure 7). It indicates that all possible pairs are not necessarily for face recognition. Therefore, we need to remove irrelevant pairs of local appearance block features.

Then, we consider the attentional feature-pair relation network with the feature-pair selection layer of $K = 442$. We achieve 97.4% accuracy on the validation set, 99.85% accuracy on the LFW, and 97.1% accuracy on the YTF, respectively (Table 2 (c) and Figure 7). The experimental results show that the AFRN with top-K selection layer outperforms the current state-of-the-art accuracies as 99.78% (ArcFace [7]) on the LFW dataset and 96.3% (PRN [14]) on the YTF dataset.

Comparison with Other Attention Mechanisms. To compare with other attention mechanisms, we conduct ablation study with top-K pair selection ($K = 442$) for comparison with other attention mechanisms including the unitary attention [15] and co-attention [37] on the validation set, LFW, and YTF datasets. We achieve 97.4% accuracy on the validation set, 99.85% accuracy on the LFW, and 97.1% accuracy on the YTF, respectively (Table 3). It indicates that the proposed feature-pair bilinear attention shows better accuracy than the other attention mechanisms.

Table 3. Comparison results with other attention mechanisms.

Method	Val. set	LFW	YTF
(a) Unitary Attention [15]	95.3	99.53	95.3
(b) Co-attention [37]	96.1	99.63	95.8
(c) Feature-pair Bilinear Attention	**97.4**	**99.85**	**97.1**

3.4. Comparison with the State-of-the-art Methods

Detailed Settings in the 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**) using the triplet ratio, pairwise, and identity preserving loss functions [13] jointly over the ground-truth identity labels: **model A** is the facial feature encoding network model with only the global appearance feature (Table 1). **model B** is the AFRN model without the feature-pair selection layer. **model C** is the AFRN model with the feature-pair selection layer. All of convolution layers and fully connected layers used BN and ReLU as non-linear activation functions.

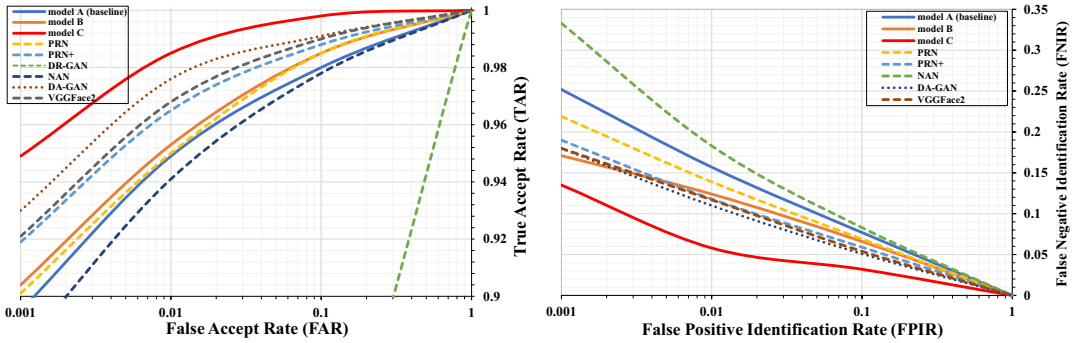
Experiments on the IJB-A dataset. We evaluate the proposed models on the IJB-A dataset [17] which contains face images and videos captured from the unconstrained environments. The IJB-A dataset is very challenging due to its full pose variation and wide variations in imaging conditions, and contains 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 the face detector [36] and the facial landmark points using DAN [18] landmark point detector, and then aligned the face image by using the alignment method in [14].

Three models (**model A**, **model B**, and **model C**) are trained on the roughly 2.8M refined VGGFace2 training set, 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 and Figure 8 (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) and Rank-N (Table 4 and Figure 8 (b)). We average all the $1,024$ dimensional output vectors of the last fully connected layer of $\mathcal{F}_{\boldsymbol{\theta}}$ for a media in the template, then we average these media-averaged features to get the final template feature as face representation. All performance evaluations are based on the squared L_2 distance threshold.

From the experimental results (Table 4 and Figure 8), we have the following observations. First, compared to **model A**, **model B** achieves a consistently superior accuracies (TAR and TPIR) by 0.4-0.9% for TAR at FAR=0.001-

Table 4. Comparison of performances of the proposed AFRN 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 [20]	0.652 ± 0.037	0.826 ± 0.018	-	-	-	0.840 ± 0.012	0.925 ± 0.008	0.946 ± 0.005
All-in-One [25]	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 [35]	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 [14]	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$^+$ [14]	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 [31]	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 [38]	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)	$\mathbf{0.895 \pm 0.015}$	$\mathbf{0.949 \pm 0.008}$	$\mathbf{0.980 \pm 0.005}$	$\mathbf{0.843 \pm 0.035}$	$\mathbf{0.923 \pm 0.005}$	$\mathbf{0.975 \pm 0.005}$	$\mathbf{0.992 \pm 0.004}$	$\mathbf{0.993 \pm 0.001}$
model B (AFRN w/o pair selection)	$\mathbf{0.904 \pm 0.013}$	$\mathbf{0.953 \pm 0.006}$	$\mathbf{0.985 \pm 0.002}$	$\mathbf{0.869 \pm 0.038}$	$\mathbf{0.935 \pm 0.004}$	$\mathbf{0.981 \pm 0.003}$	$\mathbf{0.993 \pm 0.003}$	$\mathbf{0.994 \pm 0.002}$
model C (AFRN w/ pair selection)	$\mathbf{0.949 \pm 0.013}$	$\mathbf{0.985 \pm 0.004}$	$\mathbf{0.998 \pm 0.002}$	$\mathbf{0.942 \pm 0.038}$	$\mathbf{0.968 \pm 0.004}$	$\mathbf{0.993 \pm 0.004}$	$\mathbf{0.995 \pm 0.001}$	$\mathbf{0.996 \pm 0.001}$



(a) ROC (b) DET

Figure 8. Comparison of three AFRN models with the *state-of-the-art* methods on the IJB-A dataset (average over 10 splits): (a) ROC (higher is better) and (b) DET (lower is better).

0.1 in verification task, 1.2-2.6% for TPIR at FPIR=0.01 and 0.1 in identification open set task, and 0.6% for Rank-1 in identification close set task. Second, **model C** shows a consistently higher accuracy than **model A** by the improvement of 1.8-5.4% TAR at FAR = 0.001-0.1 in the verification task, 4.5-9.9% TPIR at FPIR = 0.01-0.1 in the identification open set task, and 1.8% Rank-1 in the identification close set task. Third, **model C** shows a consistently higher accuracy than **model B** by the improvement of 1.3-4.5% TAR at FAR = 0.001-0.1 in the verification task, 3.3-7.3% TPIR at FPIR = 0.01-0.1 in the identification open set task, and 1.5% for rank-1 in the identification close set task. Last, although **model C** is trained from scratch, it outperformed the state-of-the-art method (DA-GAN [38]) by 0.7-1.9% TAR at FAR = 0.001-0.1 in the verification task, 2.2% for Rank-1 on identification close set task, and 5.2% for TPIR at FPIR = 0.01 in identification open set task on the IJB-A dataset. This validates the effectiveness of the proposed AFRN with the pair selection on the large-scale and challenging unconstrained face recognition.

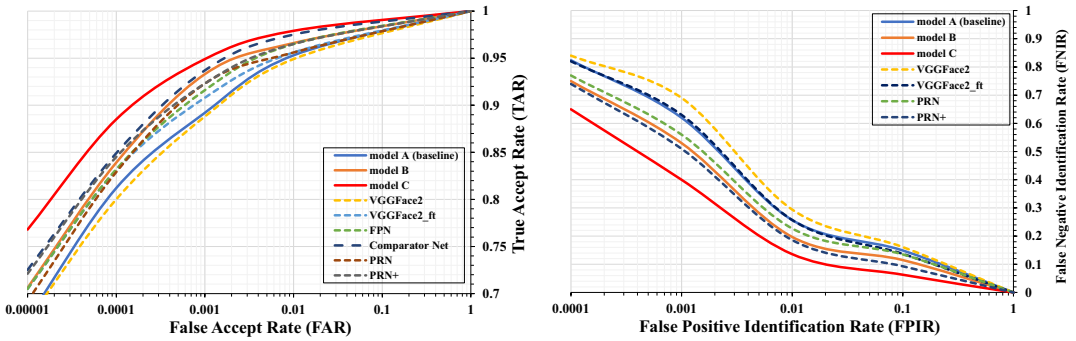
Experiments on the IJB-B dataset. We evaluate the proposed models on the IJB-B dataset [32] which contains face images and videos captured from the unconstrained environments. The IJB-B dataset is an extension of the IJB-A dataset, which contains 1,845 subjects with 21.8K still images (including 11,754 face and 10,044 non-face) and 55K frames from 7,011 videos, an average of 41 images per subject. Because images are labeled with ground truth bounding boxes, we only detect facial landmark points using DAN [18], and then aligned face images by using the face alignment method explained in [14].

Three models (**model A**, **model B**, and **modelC**) are trained on the roughly 2.8M refined VGGFace2 dataset, with no people overlapping with subjects in the IJB-B dataset. Unlike the IJB-A dataset, it does not contain any training splits. In particular, we use the 1:1 baseline verification protocol and 1:N mixed media identification protocol for the IJB-B dataset. For 1:1 face verification, we report the test results by using TAR *vs.* FAR (i.e. a ROC curve) (Table 5 and Figure 9 (a)). For 1:N face identification, we report the results by using TPIR *vs.* FPIR (equivalent to a DET curve) and Rank-N (Table 5 and Figure 9 (b)). We compare three proposed models with VGGFace2 [2], Face-PoseNet (FPN) [3], Comparator Net [34], and PRN [14]. Similarity to evaluation on the IJB-A, all performance evaluations are based on the squared L_2 distance threshold.

From the experimental results (Table 5 and Figure 9), we have the following observations. First, compared to

Table 5. Comparison of performances of the proposed AFRN 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 [34]	-	0.849	0.937	0.975	-	-	-	-	-
PRN [14]	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$^+$ [14]	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)	**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 (AFRN w/o pair selection)	**0.706**	**0.839**	**0.933**	**0.966**	$\mathbf{0.803 \pm 0.018}$	$\mathbf{0.885 \pm 0.018}$	$\mathbf{0.923 \pm 0.022}$	$\mathbf{0.962 \pm 0.010}$	$\mathbf{0.974 \pm 0.007}$
model C (AFRN w/ pair selection)	**0.771**	**0.885**	**0.949**	**0.979**	$\mathbf{0.864 \pm 0.017}$	$\mathbf{0.937 \pm 0.013}$	$\mathbf{0.973 \pm 0.015}$	$\mathbf{0.976 \pm 0.017}$	$\mathbf{0.977 \pm 0.007}$



(a) ROC (b) DET

Figure 9. Comparison of three AFRN models with the *state-of-the-art* methods on the IJB-B dataset: (a) ROC (higher is better) and (b) DET (lower is better).

model A, **model B** achieves a consistently superior accuracies (TAR and TPIR) by 1.3-4.1% for TAR at FAR = 0.00001-0.01 in the verification task, 3.4-6.0% for TPIR at FPIR = 0.01 and 0.1 in the identification open set task, and 1.2% for Rank-1 in the identification close set task. Second, **model C** shows a consistently higher accuracy than **model A** by the improvement of 2.6-9.8% TAR at FAR = 0.001-0.1 in the verification task, 8.6-12.1% TPIR at FPIR = 0.01-0.1 in the identification open set task, and 6.2% Rank-1 in the identification close set task. Third, **model C** shows a consistently higher accuracy than **model B** by the improvement of 1.3-6.5% TAR at FAR = 0.001-0.1 in the verification set task, 5.2-6.1% TPIR at FPIR = 0.01-0.1 in the identification open set task, and 5.0% for Rank-1 in the identification close set task. Last, although **model C** is trained from scratch, it outperformed the state-of-the-art method (Comparator Net [34]) by 0.4-3.6% at FAR = 0.0001-0.01 in verification task, another state-of the-art method (PRN$^+$ [14]) by 3.8% Rank-1 of identification close set task, and 5.0% TPIR at FPIR = 0.01 in the identification open set task on the IJB-B dataset. This validates the effectiveness of the proposed AFRN with the pair selection on the large-scale and challenging unconstrained face recognition.

More Experiments on the CALFW, CPLFW, CFP, AgeDB, and IJB-C datasets. Due to the limited space, we provide more experiments in Section A in the supplementary material.

4. Conclusion

We proposed the Attentional Feature-pair Relation Network (AFRN) which represented the face by the relevant pairs of local appearance block features with their weighted attention scores. The AFRN represented the face by all possible pairs of the 9×9 local appearance block features and the importance of each pair is weighted by the attention map that was obtained from adopting the low-rank bilinear pooling. We selected top-K block feature-pairs as relevant facial information, dropped the remaining irrelevant. The weighted pairs of local appearance block features were propagated to extract the joint feature-pair relation by using bilinear attention network. In experiments, we showed that the proposed AFRN achieved new state-of-the-art results in the 1:1 face verification and 1:N face identification tasks compared to current state-of-the-art methods on the challenging LFW, YTF, CALFW, CPLFW, CFP, AgeDB, IJB-A, IJB-B, and IJB-C datasets.

Acknowledgment. This research was supported by the MSIT(Ministry of Science, ICT), Korea, under the SW Starlab support program (IITP-2017-0-00897) supervised by the IITP (Institute for Information & communications Technology Promotion), IITP grant funded by MSIT (IITP-2018-0-01290), and also supported by StradVision, Inc.

References

[1] Kaidi Cao, Yu Rong, Cheng Li, Xiaoou Tang, and Chen Change Loy. Pose-robust face recognition via deep residual equivariant mapping. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, 2018.

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. *CoRR*, abs/1710.08092, 2017.

[3] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gérard Medioni. Faceposenet: Making a case for landmark-free face alignment. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1599–1608, Oct 2017.

[4] Jun-Cheng Chen, Vishal M. Patel, and Rama Chellappa. Unconstrained face verification using deep cnn features. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, March 2016.

[5] Aruni Roy Chowdhury, Tsung-Yu Lin, Subhransu Maji, and Erik Learned-Miller. One-to-many face recognition with bilinear cnns. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, March 2016.

[6] Nate Crosswhite, Jeffrey Byrne, Chris Stauffer, Omkar Parkhi, Qiong Cao, and Andrew Zisserman. Template adaptation for face verification and identification. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 1–8, May 2017.

[7] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *ArXiv e-prints*, Jan 2018.

[8] Chunrui Han, Shiguang Shan, Meina Kan, Shuzhe Wu, and Xilin Chen. Face recognition with contrastive convolution. In *European Conference on Computer Vision (ECCV 2018)*, September 2018.

[9] Tal Hassner, Iacopo Masi, Jungyeon Kim, Jongmoo Choi, Shai Harel, Prem Natarajan, and Gérard Medioni. Pooling faces: Template based face recognition with pooled face images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 127–135, June 2016.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[11] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 448–456, 2015.

[13] Bong-Nam Kang, Yonghyun Kim, and Daijin Kim. Deep convolutional neural network using triplets of faces, deep ensemble, and score-level fusion for face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 611–618, July 2017.

[14] Bong-Nam Kang, Yonghyun Kim, and Daijin Kim. Pairwise relational networks for face recognition. In *European Conference on Computer Vision (ECCV 2018)*, September 2018.

[15] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *CoRR*, abs/1610.04325, 2016.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *2015 International Conference on Learning Representation (ICLR 2015)*, 2015.

[17] Brendan F. Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, Mark Burge, and Anil K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939, June 2015.

[18] Mark Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2034–2043, July 2017.

[19] Gary B. Huang Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.

[20] Iacopo Masi, Stephen Rawls, Gérard Medioni, and Prem Natarajan. Pose-aware face recognition in the wild. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4838–4846, June 2016.

[21] Brianna Maze, Jocelyn Adams, James A. Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. Iarpa janus benchmark - c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165, Feb 2018.

[22] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1997–2005, July 2017.

[23] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, 2010.

[24] Rajeev Ranjan, Carlos D. Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *CoRR*, abs/1703.09507, 2017.

[25] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D. Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 17–24, May 2017.

[26] Tim Salimans and Diederik P Kingma. Weight normalization: A simple reparameterization to accelerate training of

deep neural networks. In *Advances in Neural Information Processing Systems 29*, pages 901–909. 2016.

[27] Swami Sankaranarayanan, Azadeh Alavi, Carlos Castillo, and Rama Chellappa. Triplet probabilistic embedding for face verification and clustering. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, Sept 2016.

[28] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M. Patel, Rama Chellappa, and David W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, March 2016.

[29] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. pages 1988–1996, 2014.

[30] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, June 2014.

[31] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recogntion (CVPR 2017)*, pages 1283–1292, 2017.

[32] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K. Jain, James A. Duncan, Kristen Allen, Jordan Cheney, and Patrick Grother. Iarpa janus benchmark-b face dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 592–600, 2017.

[33] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534, June 2011.

[34] Weidi Xie, Li Shen, and Andrew Zisserman. Comparator networks. In *European Conference on Computer Vision (ECCV 2018)*, September 2018.

[35] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5216–5225, July 2017.

[36] Jongmin Yoon and Daijin 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.

[37] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multi-modal factorized high-order pooling for visual question answering. *CoRR*, abs/1708.03619, 2017.

[38] Jian Zhao, Lin Xiong, Jianshu Li, Junliang Xing, Shuicheng Yan, and Jiashi Feng. 3d-aided dual-agent gans for unconstrained face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.