Towards Interpretable Face Recognition

Bangjie Yin[1], Luan Tran[1], Haoxiang Li[3], Xiaohui Shen[2], and Xiaoming Liu[1]

[1] Michigan State University
[2] Adobe Inc.
[3] Aibee

{yinbangj, tranluan, liuxm}@msu.edu, xshen@adobe.com, lhxustcer@gmail.com

Abstract. Deep CNNs have been pushing the frontier of visual recognition over past years. Besides recognition accuracy, strong demands in understanding deep CNNs in the research community motivate developments of tools to dissect pre-trained models to visualize how they make predictions. Recent works further push the interpretability in the network learning stage to learn more meaningful representations. In this work, focusing on a specific area of visual recognition, we report our efforts towards interpretable face recognition. We propose a spatial activation diversity loss to learn more structured face representations. By leveraging the structure, we further design a feature activation diversity loss to push the interpretable representations to be discriminative and robust to occlusions. We demonstrate on three face recognition benchmarks that our proposed method is able to improve face recognition accuracy with easily interpretable face representations.

Keywords: Interpretable CNNs, Face Representation

1 Introduction

In the era of deep learning, one major focus in the research community has been on designing neural network architectures and objective functions towards discriminative feature learning over the past years [1,2,3,4,5]. Meanwhile, given its superior even human-level surpassing visual recognition accuracy [6,7], there is a strong demand from both researchers and general audiences to interpret its successes and failures [8,9], to understand, improve, and trust its decisions. Increased interests in visualizing the CNNs lead to a set of useful tools to dissect their prediction paths to identify the important visual cues [9]. While it is interesting to see the visual evidences for predictions from pre-trained models, what's more interesting is to guide the learning towards better interpretability.

Deep CNNs trained towards discriminative classification may learn filters with wide-spreading attentions, which are usually hard to interpret for human. Prior work even empirically demonstrate models and human attend to different image areas in visual understanding [10]. Without design to harness interpretability, even when filters are observed to actively respond to certain local

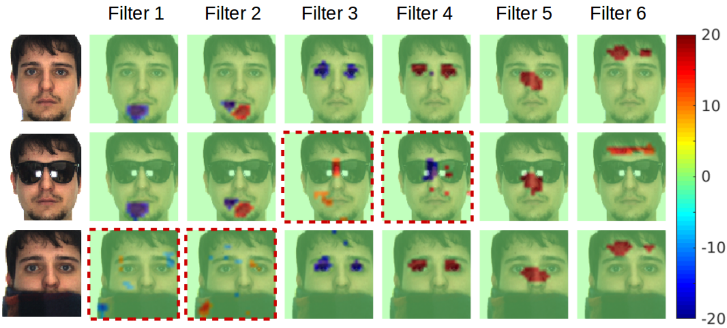


Fig. 1: An example on the behaviors of an interpretable face recognition system: left most column is three faces of the same identity and right six columns are filter responses from six filters; each filter captures a clear and consistent semantic face part, e.g., eyes, nose, and jaw; heavy occlusions, eyeglass or scarf, alternate responses of corresponding filters and make the responses being more scattered, as shown in red bounding boxes.

structure across several images, there is nothing preventing them to simultaneously capture a different structure; and the same structure may activate other filters too. One potential solution to address this issue is to provide detailed annotations to learn locally activated filters and construct a structured representation from bottom-up. However, in practice, this is rarely feasible. Manual annotations are both expensive to collect, difficult to define in certain tasks, and sub-optimal compared with end-to-end learned filters.

A desirable solution would keep the end-to-end training pipeline intact and encourage the interpretability with a model-agnostic design. However, in the recently proposed interpretable CNNs [11], where filters are trained to represent object parts to make the network representation interpretable, they observe degraded recognition accuracy after introducing interpretability. While the work is seminal and inspiring, this drawback largely limits its practical applicability.

In this paper, we study face recognition and strive to learn a interpretable face representation, as illustrated in Figure 1. We define the interpretability in the face representation that each dimension in the face representation aims to represent a human face structure or a face part. Although the concept of part-based representations has been around for years [12,13,14,15], prior methods are not easily applicable to deep CNNs. Especially in face recognition, as far as we know, this problem is rarely addressed in the literature.

In our method, the filters are learned end-to-end from data and constrained to be locally activated with the proposed spatial activation diversity loss. We further introduce a feature activation diversity loss to better align filter responses across faces and encourage filters to capture more discriminative visual cues for face recognition, especially occluded face recognition. Compared with the interpretable CNNs from Zhang et al. [11], our final face representation does not compromise recognition accuracy, instead it achieves improved performance as well as enhanced robustness to occlusion. We empirically evaluate our method on three face recognition benchmarks with detailed ablation studies on the proposed objective functions.

To summarize, our contributions in this paper are in three-fold: 1) we propose a spatial activation diversity loss to encourage learning interpretable face representations; 2) we introduce a feature activation diversity loss to enhance discrimination and robustness to occlusions, which promote the practical value of interpretability; 3) we further improve the state-of-the-art face recognition performance on three face recognition benchmarks.

2 Prior Work

2.1 Interpretable Representation Learning

Understanding the visual recognition has a long history in computer vision [16,17,18,19,20
In early days when most sophisticated visual recognition models use hand-craft features, a number of research focused on how to interpret the predictions as well. Back then visual cues include image patches [18], body parts [21], face parts [15], or middle-level representations [19] contingent on the tasks. For example, Vondrick et al. [22] develop the HOGgles to visualize HOG descriptors used in object detection. Singh et al. [19] mine discriminative middle-level patches for improved recognition accuracy. Since features such as SIFT [23], LBP [24] are generally extracted from image patches and serve as basic building blocks in the full visual recognition pipeline, it was intuitive to describe the process from the level of image patches. With the more complicated CNNs, it demands new tools to dissect its prediction. Early works include direct visualization of the filters [25], deconvolutional network to reconstruct inputs from different layers [26], gradient-based methods to generate novel inputs which maximize certain neurons [27], and etc. Recent efforts along this line include CAM [28] which leverages the global max pooling layer in some network architecture to visualize dimensions of the final representation and Grad-CAM [29] which relaxed the constraints on the network architecture with a more general framework to visualize any convolution filters in the network. While our method can be related to visualization of CNNs and we leverage the tools to visualize the learned filters in our network, it is not the focus of this paper.

Visualization of CNNs is a good way to interpret the network but by itself it does not make the network more interpretable. One recent work on learning a more meaningful representation is the interpretable CNNs from Zhang et al. [11]. In their method, they design two losses to regularize the training of late-stage convolutional filters including one to encourage each filter to encode a distinctive object part and one to keep it respond to only one local region. AnchorNet from Novotny et al. [30] adopt similar idea to encourage orthogonality of the filters and filter responses to keep each filter activated by a local and consistent structure. In our method, we generally follow the losses in AnchorNet with some tweaks for face recognition in designing our spatial activation diversity loss. Another line of research in learning interpretable representations is also referred to as feature disentangling, such as InfoGAN [31], face editing [32], 3D face recognition [33], and face modeling [34]. They intend to factorize the latent representation to describe the inputs from different aspects, of which the direction is largely diverged from our goal in this paper.

2.2 Face Recognition

Face Recognition is extensively studied in computer vision [35]. We selectively discuss literature related to interpretability in face recognition in this section. Interestingly, early works constructing meaningful face representations for face recognition are mostly intended to improve the recognition accuracy. Some face representations are composed from face parts. The part-based models are either learned unsupervisedly from data [36] or specified by manually annotated facial landmarks [37]. Besides local face parts, different face attributes are also interesting elements to build up face representations. Kumar et al. [38] proposed to encode a face image with scores from attribute classifiers and demonstrate improved face verification performance before the deep learning era. In this paper, we propose to learn meaningful part-based face representations with a deep CNN and the face part filters are learned with the carefully designed diversity losses.

We demonstrate how we leverage the interpretable face representation for occlusion robust face recognition in our experiments. Prior methods addressing pose variations in face recognition [36,37,39,40,41,42] can be related since pose changes may lead to self-occlusions. However, in this work, we are more interested in more explicit situations when faces are occluded by hand, sunglasses, and other objects. Interestingly, this specific problem is rarely studied with deep CNNs. Cheng et al. [43] propose to restore occluded faces with deep auto-encoder for improve recognition accuracy. Zhou et al. [44] argue that naively training a high capacity network with sufficient coverage in training data could achieve superior recognition performance. In our experiment, we indeed observed improved recognition accuracy to occluded faces after augmenting training data with synthetic occluded faces. However, with the proposed method, we can further improve robustness to occlusion without increasing network capacity, which highlights the merits of learning interpretable face representation.

3 Proposed Method

3.1 Overall Network Architecture

Our full network architecture in training is shown in Fig. 2. From a high-level perspective, we construct a Siamese network with two branches sharing weights to learn face representations from two faces: one with synthetic occlusion and one without. We would like to learn a set of diverse filter $\mathbf{F}$, which applies on a hyper-column descriptor $\varPhi$, consisting of feature at multiple semantic levels. The proposed Spatial Activation Diversity Loss (SAD) loss encourages the face representation to be structured with per-dimensional consistent semantic. Softmax loss helps encode the identity information. The input to the lower network branch is a synthetic occluded version of the above input. The proposed Feature Activation Diversity (FAD) loss requires filters insensitive to the occluded part

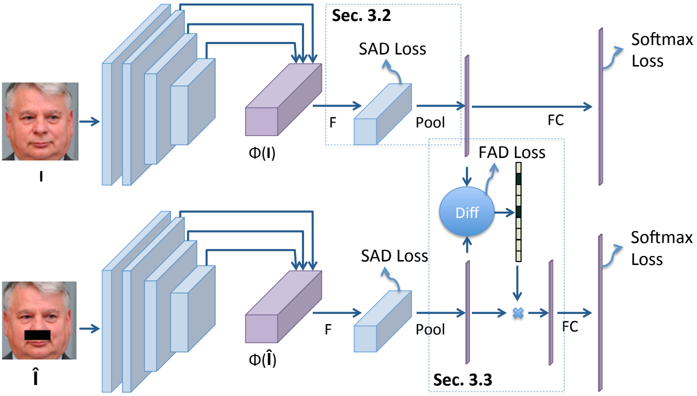


Fig. 2: The overall network architecture of the proposed method.

to be more robust to the occlusion. At the same time, we mask out parts sensitive to the occlusion in the face representation and train to identify the input face solely based on the remaining ones. As a result, the filters respond to non-occluded parts are trained to capture more discriminative cues for identification. The details of loss functions are described in the following sections.

3.2 Spatial Activation Diversity Loss

Novotny et al. [30] proposed a diversity loss for the semantic matching task by penalizing correlations among filters weights and their responses. While their idea is general enough to extend to face recognition feature learning, in practice, their design is not directly applicable due to the prohibitively large number of identities (classes) in face classifier training. Novotny et al. [30]'s approach also suffers from degradation in the recognition accuracy. In this section we first introduce their diversity loss and then describe our proposed modifications tailored to face recognition.

Spatial Activation Diversity Loss For each of K class in the training set, Novotny et al. [30] proposed to learn a set of diverse filters with discriminative power to distinguish an object of the category and background images. The filers $\mathbf{F}$ apply on a hypercolumns descriptor $\Phi(\mathbf{I})$, created by concatenating the filter responses of an image $\mathbf{I}$ at different convolutional layers [45]. This helps $\mathbf{F}$ to aggregate features at different semantic levels. The response map of this operation is denoted as $\psi(\mathbf{I}) = \mathbf{F} * \Phi(\mathbf{I})$.

The diversity constraint is implemented by two *diversity losses* $\mathcal{L}^{\text{filter}}_{\text{SAD}}$ and $\mathcal{L}^{\text{response}}_{\text{SAD}}$, encouraging the orthogonality of the filters and of their responses, respectively. $\mathcal{L}^{\text{filter}}_{\text{SAD}}$ makes filters orthogonal by penalizing their correlations:

$$\mathcal{L}^{\text{filter}}_{\text{SAD}}(\mathbf{F}) = \sum_{i \neq j} \left| \sum_{p} \frac{\langle \mathbf{F}^p_i, \mathbf{F}^p_j \rangle}{\|\mathbf{F}^p_i\|_F \ \|\mathbf{F}^p_j\|_F} \right|, \tag{1}$$

where $\mathbf{F}_i^p$ is the column of filter $\mathbf{F}_i$ at the spatial location p. Note that orthogonal filters are likely to respond to different image structures, but this is not necessarily the case. Thus, the second term $\mathcal{L}_{\text{Div}}^{\text{response}}$ is introduced to directly decorrelates the filters' *response maps* $\psi_k(\mathbf{I})$:

$$\mathcal{L}_{\text{SAD}}^{\text{response}}(\mathbf{I}; \Phi, \mathbf{F}) = \sum_{i \neq j} \left\| \frac{\langle \psi_i, \psi_j \rangle}{\|\psi_i\|_F \|\psi_j\|_F} \right\|^2 . \tag{2}$$

This term is further regularized by using the smoothed response maps $\psi'(\mathbf{I}) \doteq g_\sigma * (\psi(\mathbf{I}))$ in place of $\psi(\mathbf{I})$ in $\mathcal{L}_{\text{SAD}}^{\text{response}}$ loss computing. Here the channel-wise Gaussian kernel g_σ is applied to encourage filter responses to spread farther apart by dilating their activations.

Our Proposed Modifications Novotny et al. [30] learn K sets of filters, one for each of K categories. The discrimination of the features are maintained by K binary classification losses for each category vs. background images. The discriminative loss is proposed to enhance (or suppress) the maximum value in the response maps ψ_k for the positive (or negative) class. In [30], the final feature representation $\mathbf{f}$ is obtained via global max-pooling operation on ψ. This design is not direct applicable for face classification CNN as the number of identities K are usually prohibitively large (usually in the order of ten thousands or above).

Here, to make the feature discriminative, we only learn **one** set of filters and connect the representation $\mathbf{f}(\mathbf{I})$ directly to a K-way softmax classification:

$$\mathcal{L}_{id} = -\log(P_c(\mathbf{f}(\mathbf{I}))). \tag{3}$$

Here we minimize the negative log-likelihood of feature $\mathbf{f}(\mathbf{I})$ being classified to its ground-truth identity c.

Furthermore, global max-pooling could lead to unsatisfied recognition performance, as shown in [30] where they observed minor performance degradation compared to the model without their diversity loss. One empirical explanation of this performance degradation is that max-pooling has similar effect to ReLU activation which makes the response distribution biased to non-negative range $[0, +\infty)$. Hence it significantly limits the feasible learning space.

Most recent works choose to use global average pooling [46,39]. However, the problem of average-pooling when applied to introduce interpretability is that it does not promote desired spatially peaky distribution. Empirically, we found the learned feature response maps of average pooling failed to have strong activation in small local regions.

Here we propose to design a pooling operation that satisfies two conditions: i) promote peaky distribution to be well-cooperated with the spatial activation diversity loss; ii) maintain the statistics of the feature responses for the global average-pooling to achieve good recognition performance. Based on these considerations, we propose the operation termed **Large magnitude filtering** (LMF), as follows:

For each channel in the feature response map, we make $d\%$ of elements with smallest magnitude to be 0. The size of the output remains the same. The $\mathcal{L}_{\text{SAD}}^{\text{response}}$ loss is applied on the modified response map $\psi'(\mathbf{I}) \doteq g_\sigma * (\text{LMF}(\psi(\mathbf{I})))$ in place of $\psi(\mathbf{I})$ in Eqn. 2.

Then, the conventional global average pooling is be applied to $\text{LMF}(\psi(\mathbf{I}))$ to get the final representation $\mathbf{f}(\mathbf{I})$. By removing small magnitude values from ψ_k, $\mathbf{f}$ won't be affected much after global average pooling, which favors discriminative feature learning. On the other hand, the peaks of the response maps are still well maintained, which leads to more reliable computation of the diversity loss.

3.3 Feature Activation Diversity Loss

One way to evaluate whether the diversity loss is effective is to compute the average location of the peaks within the kth response maps $\psi'_k(\mathbf{I})$ for a large set of images. If the average locations across K filters are spread all over the face spatially, we consider the diversity loss is well functioning and able to associate each filer with a specific area of the face. With the spatial activation diversity loss, we do observe the improved *spreadness* compared to the base CNN model trained without the SAD loss. However, with the ultimate goal of interpretable face recognition, we hope to further boost the spreadness of the average peak locations across filters (or elements of the learnt representation).

Motivated by the goal of learning part-based face representations, it is desirable to encourage that any local face area only affects a small subset of the filter responses. To fulfill this desire, we propose to create synthetic occlusion on local face areas of a face image, and constrain on the difference between its feature response and that of the unoccluded original face image. The second motivation for our proposal is to design a occlusion-robust face recognition algorithm, which, in our view, should be a natural by-product or benefit of the part-based face representation.

With this in mind, we propose a Feature Activation Diversity (FAD) Loss to encourage the network to learn filters robust to occlusions. That is, occlusion in a local region should only affect a small subset of elements within the representation. Specifically, leveraging pairs of face images $\mathbf{I}$, $\hat{\mathbf{I}}$, where $\hat{\mathbf{I}}$ is a version of $\mathbf{I}$ with a synthetically occluded region, we enforce the majority of two feature representations, $\mathbf{f}(\mathbf{I})$ and $\mathbf{f}(\hat{\mathbf{I}})$, to be similar:

$$\mathcal{L}_{\text{FAD}}(\mathbf{I}, \hat{\mathbf{I}}) = \sum_i \left| \tau_i(\mathbf{I}, \hat{\mathbf{I}}) \left[\mathbf{f}_i(\mathbf{I}) - \mathbf{f}_i(\hat{\mathbf{I}}) \right] \right|, \tag{4}$$

where the feature selection mask $\tau(\mathbf{I}, \hat{\mathbf{I}})$ is defined with threshold t: $\tau_i(\mathbf{I}, \hat{\mathbf{I}}) = 1$ if $\left| \mathbf{f}_i(\mathbf{I}) - \mathbf{f}_i(\hat{\mathbf{I}}) \right| < t$, otherwise $\tau_i(\mathbf{I}, \hat{\mathbf{I}}) = 0$. There are multiple design choices for the threshold: number of elements based or value based. We evaluate and discuss these choices in the experiment section and supplementary material.

We also would like to correctly classify occluded images using just subset of feature elements, which is insensitive to occlusion. Hence, the softmax identity loss in the occlusion branch is applied to the masked feature:

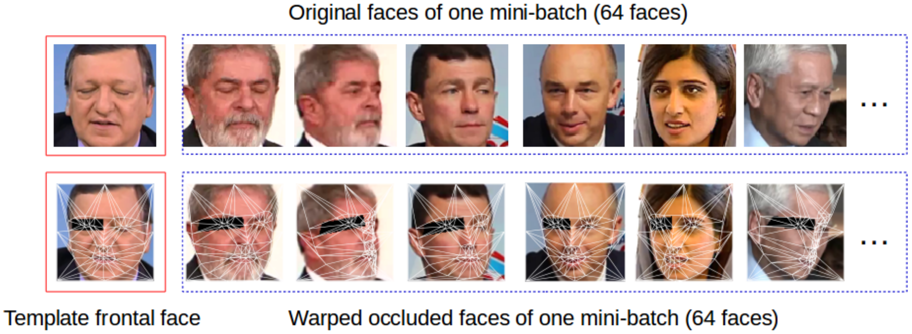

Original faces of one mini-batch (64 faces)
Template frontal face
Warped occluded faces of one mini-batch (64 faces)

Fig. 3: The mask warping process. The white lines in the second row define 142 triangles. The barycentric coordinates of triangles help to warp the vertices of the mask from the template frontal face to each image within the mini-batch.

$$\mathcal{L}_{id}^{occluded} = -\log(P_c(\tau(\mathbf{I}, \hat{\mathbf{I}}) \odot \mathbf{f}(\hat{\mathbf{I}}))). \tag{5}$$

By sharing the weight of the classifier between two branches, this classifier is learned to be more robust to occlusion. This also leads to a better presentation as filters respond to non-occluded parts need to be more discriminative.

3.4 Implementation Details

The detailed network structure is presented in Tab. 1, which is inspired by the widely used CASIA-Net [46,47]. We added HC-descriptor-related blocks for our spatial diversity loss learning.

To speed up the training process, we reuse the feature extraction network from a pretrained model shared by Tran et al. [47], and conv32, conv42, conv52 layers are used to construct the HC descriptor via conv upsampling layers. All new weights are randomly initialized using a truncated normal distribution with std of 0.02. The entire network will then be jointly trained using Adam optimizer with a learning rate of 0.001.

For FAD, the feature mask τ can be computed per image pair $\mathbf{I}$ and $\hat{\mathbf{I}}$. However, to obtain a more reliable feature masking, we decide to compute τ using multiple image pairs, even the whole mini-batch, with the same *physical* occluded mask. $\tau(\{\mathbf{I}_i\}, \{\hat{\mathbf{I}}_i\}) = 1$ if $\left|\bar{\mathbf{f}}_i(\mathbf{I}) - \bar{\mathbf{f}}_i(\hat{\mathbf{I}})\right| < t$, otherwise 0.

To provide the same *physical* mask to images in a batch regardless their poses, we first define a frontal face template with 142 triangles created by 68 facial landmarks. A random 32×12 a rectangle is selected as a normalized mask. Each one of the rectangle's four vertices can be represented by the barycentric coordinate w.r.t. the triangle enclosing the vertex. On each image of a mini-batch, corresponding four vertices of a quadrilateral can be found via the same coordinates. This quadrilateral denotes the location of a warped mask of that input image. An example of this mask warping process is shown in Fig. 3.

Table 1: The structures of our network architecture.

Layer	Input	Filter/Stride	Output Size
conv11	Image	$3 \times 3/1$	$96 \times 96 \times 32$
conv12	conv11	$3 \times 3/1$	$96 \times 96 \times 64$
conv21	conv12	$3 \times 3/2$	$48 \times 48 \times 64$
conv22	conv21	$3 \times 3/1$	$48 \times 48 \times 64$
conv23	conv22	$3 \times 3/1$	$48 \times 48 \times 128$
conv31	conv23	$3 \times 3/2$	$24 \times 24 \times 128$
conv32	conv32	$3 \times 3/1$	$24 \times 24 \times 96$
conv33	conv32	$3 \times 3/1$	$24 \times 24 \times 192$
conv41	conv33	$3 \times 3/2$	$12 \times 12 \times 192$
conv42	conv41	$3 \times 3/1$	$12 \times 12 \times 128$
conv43	conv42	$3 \times 3/1$	$12 \times 12 \times 256$
conv51	conv43	$3 \times 3/2$	$6 \times 6 \times 256$
conv52	conv51	$3 \times 3/1$	$6 \times 6 \times 160$
conv53	conv52	$3 \times 3/1$	$6 \times 6 \times N^f$
conv43-U	conv43	upsampling	$24 \times 24 \times 256$
conv44	conv43-U	$1 \times 1/1$	$24 \times 24 \times 192$
conv53-U	conv53	upsampling	$24 \times 24 \times 320$
conv54	conv53-U	$1 \times 1/1$	$24 \times 24 \times 192$
Φ (HC)	conv33,44,54	$3 \times 3/1$	$24 \times 24 \times 576$
Ψ	Φ	$3 \times 3/1$	$24 \times 24 \times N^f$
AvgPool	Ψ	$24 \times 24/1$	$1 \times 1 \times N^f$

4 Experimental Results

This section provides ablation studied, qualitative and quantitative evaluation. Firstly, to further analyze the influence of parameters setting in our model, ablation studies are conducted. We set the different thresholds in feature activation diversity loss to explore the face recognition performance, and turn off one of the diversity losses also help us to understand the effect of the proposed two loss functions. Secondly, to better illustrate the results of our method, we present qualitative visualization of the learnt representations, response maps, etc. Lastly, we compare the face recognition performance on benchmark datasets: IJB-A [48], IJB-C [49] and AR face [50].

4.1 Experimental Settings

Databases For training purpose, we use CASIA-WebFace [46] databases. For testing, we use IJB-A [48], IJB-C [49] and AR face database [50]. CASIA-WebFace contains $493,456$ images of $10,575$ different subjects. IJB-A [48] is a video-based face recognition database containing around $20,412$ videos and $5,396$ facial images. For verification and identification evaluation, it provides a protocol for templates-to-template matching. Different poses, illumination conditions and image qualities in IJB-A make it very challenging for face recognition. More importantly, this dataset also provides annotations on natural occlusions, which help us evaluate on occluded face recognition. In our experimental setting, IJB-A will be evaluated in three different scenarios, including original faces, synthetic occlusion faces and natural occlusion faces. For synthetic occlusion, we randomly generate a warped occluded area, a black quadrilateral, for each testing image in IJB-A, which is the same as what we did in training stage.

IJB-C extends IJB-A, also is a video-based face database with $3,134$ still images and $117,542$ frames from natural scene videos of $3,531$ different subjects. The IJB-C protocols can evaluate face detection, verification, identification and clustering. For verification, IJB-C gives $19,557$ genuine comparisons and $15,638,932$ impostor comparisons. For identification, IJB-C provides two

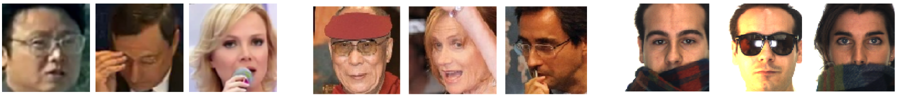


Fig. 4: Example images of IJB-A, IJB-C and AR face database. The occlusions include scarf, eyeglass, strong illumination, hands, etc.

.

Table 2: Performance comparison on IJB-A database.

Method	IJB-A		Manual Occlusion		Natural Occlusion	
Metric	@FAR=.01	@Rank-1	@FAR=.01	@Rank-1	@FAR=.01	@Rank-1
$t = 130$	79.0 ± 1.6	$\mathbf{89.5} \pm 0.8$	$\mathbf{76.1} \pm 1.7$	$\mathbf{88.0} \pm 1.4$	66.2 ± 4.0	$\mathbf{73.0} \pm 3.3$
$t = 260$	$\mathbf{79.2} \pm 1.8$	89.4 ± 0.8	$\mathbf{76.1} \pm 1.4$	$\mathbf{88.0} \pm 1.2$	$\mathbf{66.5} \pm 6.4$	72.3 ± 2.8
$t = 320$	74.6 ± 2.4	88.9 ± 1.3	71.8 ± 3.1	87.5 ± 1.6	61.0 ± 6.5	71.6 ± 3.2

galleries, and our evaluation is based the average performance of these two galleries. One unique property of IJB-C is its label on fine-grained occlusion area. Therefore, we use IJB-C to evaluate occlusion-robust face recognition, by including testing images with at least one occluded face area.

AR face database is another natural occluded face database, containing almost 4K faces of 126 subjects. Each subject has images under different conditions, including illumination, expressions and natural occlusions. In the evaluation of AR, we only take faces with natural occlusions, including wearing glass and scarfs. Some examples faces in IJB-A, IJB-C and AR databases are shown in Fig. 4. Following the setting in [39], all training images are processed and resized to 110 × 110. A random crop of size 96 × 96 will be used as input to feed into the network. All test images are resized to 96 × 96.

4.2 Ablation Study

Different Thresholds As mentioned in previous sections, the threshold we select does affect the face recognition performance. This is an important parameter to study in our experiments, therefore we evaluate the effect of different thresholds on face recognition performance on IJB-A. Here we explore an option to define threshold on the remaining number of elements in the feature representation $\mathbf{f}$. We train different models with $t = 130, 260, 320$, in t denotes the number of elements in two $320d$ features that the FAD loss encourages their similarity. Table 2 shows the result comparison on all three variants of IJB-A dataset. When forcing all elements of $\mathbf{f}(\mathbf{I})$ and $\mathbf{f}(\hat{\mathbf{I}})$ to be the same ($t = 320$), the performance significantly drops on all three sets. In this case, the feature representation of the non-occluded face will be negatively affected as being completely pushed toward a representation of the occluded one. While the model with $t = 130$ has similar performance with model $t = 260$, we will use the latter model for the rest of the paper due to the observation that the latter model affects less filters, push other filter responses away from any local occlusions, and subsequently helps to enhance the spreadness of the average response locations.

Table 3: Performance comparison on IJB-A database.

Method ↓	Verification		Identification	
Metric (%) →	@FAR=0.01	@FAR=0.001	@Rank-1	@Rank-5
Base CNN	78.3 ± 2.4	51.9 ± 6.7	88.7 ± 1.0	94.8 ± 0.7
Base CNN agu.	78.9 ± 1.8	56.6 ± 4.8	88.5 ± 1.1	94.9 ± 0.8
Ours (SAD only)	78.1 ± 1.8	56.6 ± 4.4	88.1 ± 0.9	95.0 ± 1.0
Ours (FAD only)	76.7 ± 2.0	56.1 ± 5.3	88.1 ± 1.1	**95.5** ± 0.7
Ours	**79.2** ± 1.8	**60.0** ± 3.1	**89.4** ± 1.5	95.3 ± 0.5

Table 4: Performance comparison on synthetic occlusion faces of IJB-A.

Method ↓	Verification		Identification	
Metric (%) →	@FAR=0.01	@FAR=0.001	@Rank-1	@Rank-5
Base CNN	61.8 ± 5.5	39.1 ± 7.8	79.6 ± 2.1	91.4 ± 1.2
Base CNN agu.	75.1 ± 2.6	50.7 ± 5.5	85.7 ± 1.4	93.8 ± 1.2
Ours (SAD only)	66.6 ± 5.6	42.1 ± 7.3	81.2 ± 1.9	91.7 ± 1.4
Ours (FAD only)	75.2 ± 2.4	51.1 ± 4.9	85.1 ± 1.2	93.6 ± 1.2
Ours	**76.1** ± 1.4	**56.1** ± 4.4	**88.0** ± 1.7	**94.8** ± 0.6

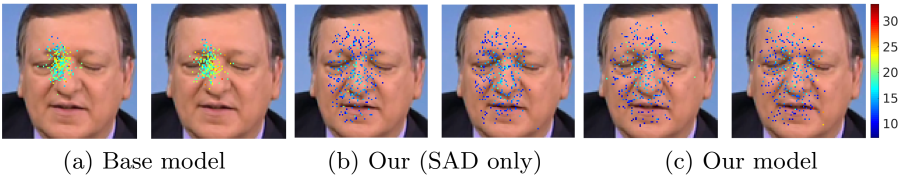
Table 5: Performance comparison on natural occlusion faces of IJB-A.

Method ↓	Verification		Identification	
Metric (%) →	@FAR=0.01	@FAR=0.001	@Rank-1	@Rank-5
DR-GAN [47]	64.7 ± 4.1	41.8 ± 6.4	70.8 ± 3.6	81.7 ± 2.9
Base CNN	64.4 ± 6.1	40.7 ± 6.8	71.3 ± 3.5	81.6 ± 2.5
Base CNN agu.	65.7 ± 4.9	41.4 ± 6.2	71.1 ± 3.0	81.3 ± 1.1
Ours (SAD only)	64.2 ± 6.9	40.1 ± 6.5	71.0 ± 3.3	81.6 ± 2.2
Ours (FAD only)	64.2 ± 5.5	42.2 ± 6.9	71.3 ± 3.8	81.3 ± 2.5
Ours	**66.5** ± 6.4	**47.1** ± 5.9	**72.3** ± 2.8	**82.3** ± 1.9

Spatial vs. Feature Diversity Loss Since we propose two different diversity losses in our model learning, it is important to evaluate the effects of two losses on face recognition performance respectively. As shown in Tables 3, 4, 5, we can train our models using either one of the two diversity losses, or both of them. From the results, we can observe that, while the SAD loss performs reasonably well on general IJB-A dataset, it suffers for data with occlusions, being synthetic or natural. Alternatively, using only the FAD loss can improve the performance on the two datasets with occlusions. Finally, using both losses achieves the best performance on all three datasets.

4.3 Qualitative Evaluation

Spreadness of Average Locations of Filter Response Given an input face image, our model computes $\psi'(\mathbf{I})$, the 320 feature maps of size 24×24, where the average pooling of each map will be one element of the final feature representation. Each feature map has the highest and the lowest response values, which are located in different spatial areas of the face. To better illustrate the spatial distribution of the locations of the peaks, we randomly select $1,000$ testing face

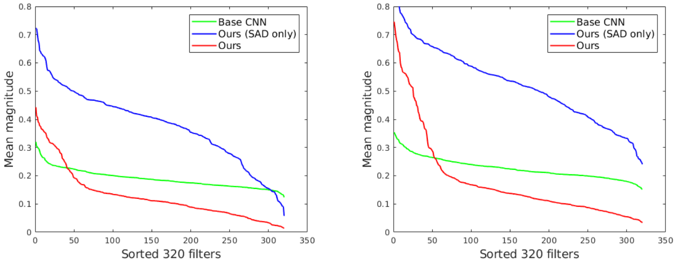


(a) Base model (b) Our (SAD only) (c) Our model

Fig. 5: The average locations of peak responses of 320 filters for three models, where each shows average locations of positive (left) and negative (right) peak responses. The color denotes the standard deviation of peak locations. The face size is 96 × 96.

images and calculate the weighted average location for each filter, with three notes. One is that there are two locations, for the highest and lowest responses respectively. The other is that, since the filters are responsive to semantic facial components, their spatial locations on the images may change with pose. To compensate that, we warp the peak location in an arbitrary-view face to a canonical frontal-view face, by its barycentric coordinates with respect to the triangle that enclosing it. Similar to Fig. 3, we use 68 estimated facial landmarks [51,52] and control points on the image boundary to define the triangular mesh. Finally, the weight of each location is determined by the magnitude of its peak response. With that, the average locations for all feature maps (or filters) are shown in Fig. 5. Comparing the visualization results between our models and CNN base model, for both positive and negative response, our model with SAD loss enlarge the spreadness of the average locations. Furthermore, our model with both losses continue to push the filter responses apart from each other. This demonstrates that indeed our model is able to push filters to attach to diverse face areas, when compared to the CNN base model where all filters are concentrated on the left eye corner. In addition, we compute the standard deviation for each filter's peak location from $1,000$ images. From Fig. 5, we can observe that base model has larger standard deviations than SAD only model and our model, which means our model can be better concentrated on a local face part than the base model.

Mean Feature Difference Comparison Both of our losses promotes part-based feature learning, which could results in occlusion robustness. Especially, in FAD, we directly minimize the difference in a portion of representation of face with and without occlusion. In this section, we study the effect of our loss on face images with occlusion. Firstly, we randomly select $1,000$ different face images in different poses and then generate the synthetic occlusion. After that, we calculate the mean of feature difference of each filter on both original and occluded faces based on different models. Fig. 6 (a) and (b) illustrates the sorted feature difference of three models at two different occlusion parts, eye and nose, respectively. Compare to the base CNN (trained with $\mathcal{L}_{id}$), both of our losses have smaller magnitude of differences. Diversity properties of SAD loss could help to reduce the feature change on occlusion, even without directly minimizing this



(a) Eye part occlusion (b) Nose part occlusion

Fig. 6: Comparison of mean feature difference on two occlusion parts.

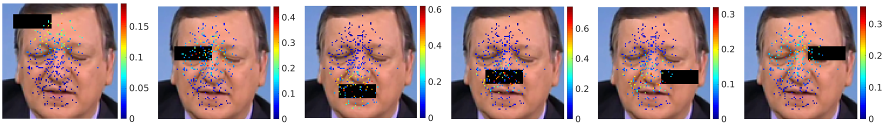


Fig. 7: Visualization of the correspondence between feature difference magnitude and occlusion locations. Best viewed electronically.

.

difference. FDA loss further enhances robustness by only letting the occlusion alternate a small portion of the representation, keeping the remaining element insensitive to the occluded part.

Visualization on Feature Difference Vectors As demonstrated in Fig. 5, each of our filter can correspond to a spatial face location. Here we further study the relation of these average locations of response and physical meaning on input images. In Fig. 7, we visualize the magnitude of changes of each filter response due to six different occlusion locations. We observe the locations of points with large feature difference are consistent with the occluded area in faces, which means our learned filters indeed sensitive to various facial areas.

Another interesting observation from this figure is that the magnitude of the feature difference can be very different with different occlusion. The maximum feature difference can be as high as 0.7 in with occlusion in eye or mouth area, meanwhile this number is only 0.3 in less critical area (forehead, cheek, etc).

Filter Response Visualization As shown in Fig. 8, we visualize the feature responses of some filters on different subjects' faces. From the heat maps, we can see how each filter is attached to a specific *semantic* location on the faces, independent to either identities or poses. This is especially impressive for faces with varying poses, in that despite no pose prior is used in training, the filter can always respond to the semantically equivalent local part.

4.4 Quantitative Evaluation

Standard Deviation of Peak Locations of Filter Responses Figure 5 shows the average of the peak locations of 320 filter responses, across 1,000

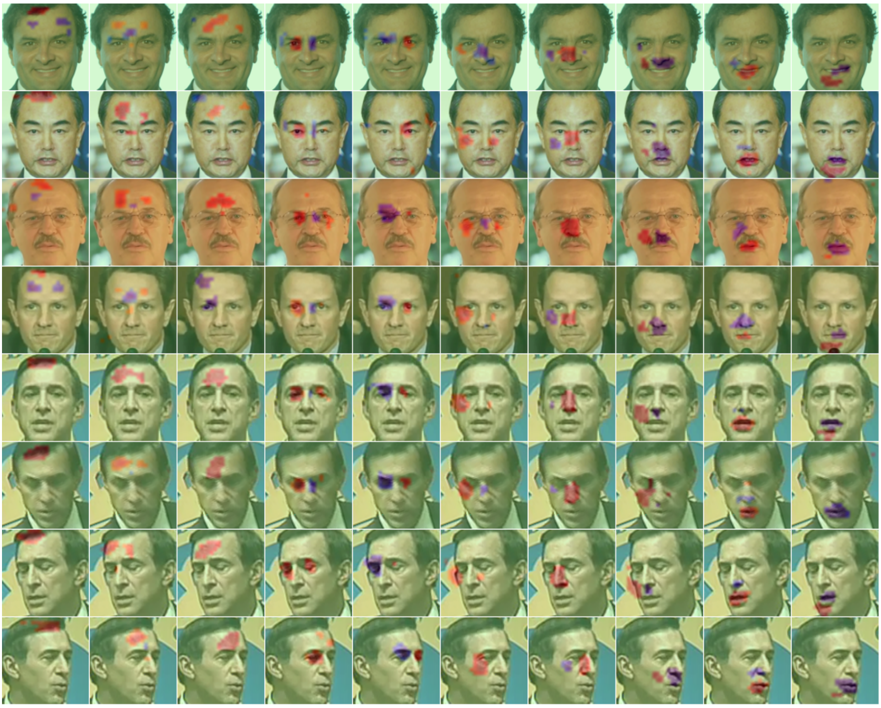


Fig. 8: Visualization of filter response “heat maps” of 10 different filters on faces from different subjects (top 4 rows) and the same subject (bottom 4 rows). The positive and negative responses are shown as two colors within each image. Note the high consistency of response locations across subjects and across poses.

images. We can also compute the standard deviation of those peak locations, for both the positive responses and negative responses, respectively. As shown in Fig. 9, our model can generate filter responses whose locations have *smaller* standard deviations than CNN base model, i.e., our filter can be more concentrated on one specific location of the face. Note that the SAD loss only model also reduces the standard deviation over the CNN base model , our model can have a slightly smaller standard deviations than SAD loss only model.

Evaluation on Benchmark Datasets We test our models on IJB-A dataset to compare the face recognition performance. Note that our main objective is to show how we can improve the interpretability of face recognition while maintaining face recognition performance. Hence, the baseline model for comparison is a conventional CNN model with softmax loss, rather than advanced CNN models with many add-on elements. For fair comparison, the CNN base model uses the same network architecture as Tab.1. Also, we perform data augmentation for CNN base model where the same synthetic faces used for our model training are supplemented into the training data.

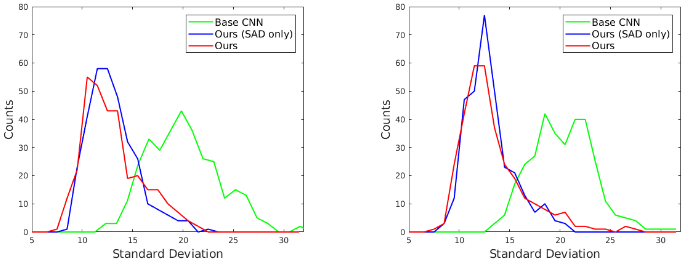


Fig. 9: Histograms of standard deviations of peak locations for positive (left) and negative (right) responses.

.

Table 6: Performance comparison on natural occlusion faces of IJB-C.

Method ↓	Verification		Identification	
Metric (%) →	@FAR=0.01	@FAR=0.001	@Rank-1	@Rank-5
DR-GAN [47]	82.4	66.1	70.8	82.8
Base CNN	83.3	67.0	72.1	83.3
Base CNN agu.	83.7	68.9	73.6	83.4
Ours	**83.8**	**69.3**	**74.5**	**83.6**

As shown in Tables 3, 4 and 5, when comparing to the CNN base models, our model achieves the best performance. It is worthy note that this is the first time that a reasonably interpretable representation is able to improve the recognition performance. Furthermore, the performance improvement on the occlusion datasets are more substantial than the generic IJB-A database, which shows the advantage of interpretable representations in handling occlusions.

For IJB-C, we select faces with at least one occluded area based on the labels. The natural occluded face recognition performance on IJB-C is shown in Tab. 6. Similar to the results in Tab. 5, on this fine-grained occlusion dataset, our proposed method also achieves the superior performance compared to the base CNN models.

For testing face verification on AR face database, we select all 810 face images with eyeglasses and scarfs occlusions. From this set, we randomly sample 6,000 same person pairs and 6,000 different person pairs from the occluded faces. We compute the representations of an image pair and their cosine distance. The experimental results are shown in Fig. 10, we observe that our model achieves the superior performance comparing to the other two models. The Equal Error Rate (EER) of our model, the base CNN model and the base CNN model with data augmentation is 16%, 22%, and 19%, respectively.

5 Conclusions

In this paper, we present our efforts towards interpretable face recognition. Our grand goal is to learn from data a structured face representation in which each dimension activates on a consistent semantic face part and captures its identity

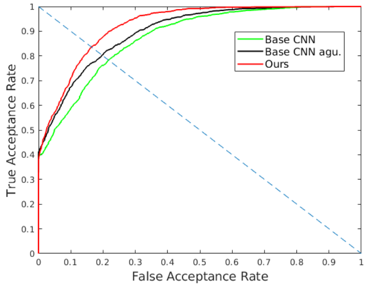


Fig. 10: ROC curves of different models on AR database.

.

information. We propose two novel losses to encourage both spatial activation diversity and feature activation diversity in the final stage convolutional filters and the face representation. We empirically demonstrate the proposed method can lead to more locally constrained individual filter responses and overall widely-spreading filters distribution. A by-product from the harnessed interpretability leads to improved robustness to occlusions in face recognition.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016)
2. Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., Keutzer, K.: Densenet: Implementing efficient convnet descriptor pyramids. arXiv preprint arXiv:1404.1869 (2014)
3. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. arXiv preprint arXiv:1708.02002 (2017)
4. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: ECCV. (2016)
5. Liu, Y., Li, H., Wang, X.: Learning deep features via congenerous cosine loss for person recognition. arXiv preprint arXiv:1702.06890 (2017)
6. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: ICCV. (2015)
7. Lu, C., Tang, X.: Surpassing human-level face verification performance on LFW with gaussianface. In: AAAI. (2015)
8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
9. Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., Mordvintsev, A.: The building blocks of interpretability. Distill (2018)
10. Das, A., Agrawal, H., Zitnick, L., Parikh, D., Batra, D.: Human attention in visual question answering: Do humans and deep networks look at the same regions? CVIU (2017)
11. Zhang, Q., Wu, Y.N., Zhu, S.C.: Interpretable convolutional neural networks. arXiv preprint arXiv:1710.00935 (2017)
12. Li, S.Z., Hou, X.W., Zhang, H.J., Cheng, Q.S.: Learning spatially localized, parts-based representation. In: CVPR. (2001)
13. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR. (2008)

14. Berg, T., Belhumeur, P.N.: Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: CVPR. (2013)
15. Li, H., Hua, G.: Probabilistic elastic part model: a pose-invariant representation for real-world face verification. TPAMI (2017)
16. Mahendran, A., Vedaldi, A.: Visualizing deep convolutional neural networks using natural pre-images. IJCV (2016)
17. Sudderth, E.B., Torralba, A., Freeman, W.T., Willsky, A.S.: Learning hierarchical models of scenes, objects, and parts. In: ICCV. (2005)
18. Juneja, M., Vedaldi, A., Jawahar, C., Zisserman, A.: Blocks that shout: Distinctive parts for scene classification. In: CVPR. (2013)
19. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: ECCV. (2012)
20. Parikh, D., Zitnick, C.: Human-debugging of machines. NIPS WCSSWC (2011)
21. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: ICCV. (2011)
22. Vondrick, C., Khosla, A., Malisiewicz, T., Torralba, A.: Hoggles: Visualizing object detection features. In: ICCV. (2013)
23. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
24. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. TPAMI (2006)
25. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV. (2014)
26. Zeiler, M.D., Taylor, G.W., Fergus, R.: Adaptive deconvolutional networks for mid and high level feature learning. In: ICCV. (2011)
27. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: CVPR. (2015)
28. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. (2016)
29. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. https://arxiv. org/abs/1610.02391 v3 (2016)
30. Novotny, D., Larlus, D., Vedaldi, A.: Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In: CVPR. (2017)
31. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: NIPS. (2016)
32. Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., Samaras, D.: Neural face editing with intrinsic image disentangling. arXiv preprint arXiv:1704.04131 (2017)
33. Liu, F., Zeng, D., Zhao, Q., Liu, X.: Disentangling features in 3D face shapes for joint face reconstruction and recognition. In: CVPR. (2018)
34. Tran, L., Liu, X.: Nonlinear 3D face morphable model. In: CVPR. (2018)
35. Learned-Miller, E., Huang, G.B., RoyChowdhury, A., Li, H., Hua, G.: Labeled faces in the wild: A survey. In: Advances in face detection and facial image analysis. (2016)
36. Li, H., Hua, G., Lin, Z., Brandt, J., Yang, J.: Probabilistic elastic matching for pose variant face verification. In: CVPR. (2013)
37. Cao, Z., Yin, Q., Tang, X., Sun, J.: Face recognition with learning-based descriptor. In: CVPR. (2010)
38. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV. (2009)

39. Tran, L., Yin, X., Liu, X.: Disentangled representation learning GAN for pose-invariant face recognition. In: CVPR. (2017)
40. Chai, X., Shan, S., Chen, X., Gao, W.: Locally linear regression for pose-invariant face recognition. TIP (2007)
41. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Towards large-pose face frontalization in the wild. In: ICCV. (2017)
42. Yin, X., Liu, X.: Multi-task convolutional neural network for pose-invariant face recognition. IEEE Transactions on Image Processing (2018)
43. Cheng, L., Wang, J., Gong, Y., Hou, Q.: Robust deep auto-encoder for occluded face recognition. In: ICM. (2015)
44. Zhou, E., Cao, Z., Yin, Q.: Naive-deep face recognition: Touching the limit of LFW benchmark or not? arXiv preprint arXiv:1501.04690 (2015)
45. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: CVPR. (2015)
46. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
47. Tran, L., Yin, X., Liu, X.: Representation learning by rotating your faces. arXiv preprint arXiv:1705.11136 (2017)
48. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark-A. In: CVPR. (2015)
49. Brianna Maze, J.A., James A. Duncan, N.K., Tim Miller, C.O., Anil K. Jain, W.T.N., Janet Anderson, J.C., Grother, P.: IARPA Janus Benchmark-C: Face dataset and protocol. In: ICB. (2018)
50. Martinez, A.M.: The AR face database. CVC Technical Report24 (1998)
51. Liu, Y., Jourabloo, A., Ren, W., Liu, X.: Dense face alignment. In: ICCV Workshop. (2017)
52. Jourabloo, A., Liu, X.: Pose-invariant face alignment via CNN-based dense 3D model fitting. IJCV (2017)