Towards Interpretable Face Recognition

Bangjie Yin[1*] Luan Tran[1*] Haoxiang Li[2†] Xiaohui Shen[3†] Xiaoming Liu[1]
[1]Michigan State University [2]Wormpex AI Research [3]ByteDance AI Lab

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 achieve the state-of-art face recognition accuracy with easily interpretable face representations.

1. Introduction

In the era of deep learning, one major focus in the research community has been on designing network architectures and objective functions towards discriminative feature learning [19, 20, 29, 34, 54, 59]. Meanwhile, given its superior even surpassing-human recognition accuracy [18, 36], there is a strong demand from both researchers and general audiences to interpret its successes and failures [15, 41], to understand, improve, and trust its decisions. Increased interests in visualizing CNNs lead to a set of useful tools to dissect their prediction paths to identify the important visual cues [41]. 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.

CNNs trained towards discriminative classification may learn filters with wide-spreading attentions – usually hard to interpret for human. Prior work even empirically demonstrate models and human attend to different image areas in visual understanding [11]. Without design to harness interpretability, even when filters are observed to actively respond to certain local 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 annotations to learn locally activated filters and construct a structured representation from bottom-up. However, in practice, this is rarely feasible. Manual annotations are expensive to collect, difficult to define in certain tasks, and sub-optimal compared with end-to-end learned filters.

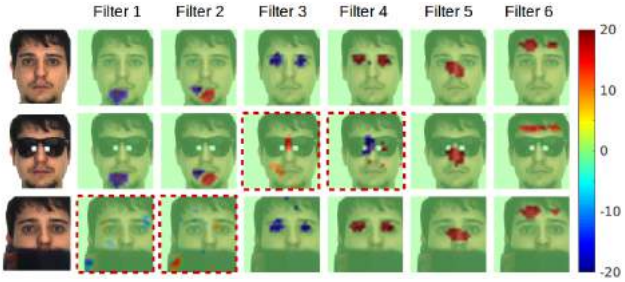

Figure 1. An example on the behaviors of an interpretable face recognition system: left most column is three faces of the same identity and right six columns are filter responses from six filters; each filter captures a clear and consistent semantic face part, e.g., eyes, nose, and jaw; heavy occlusions,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

A desirable solution would keep the end-to-end training pipeline intact and encourage the interpretability with a model-agnostic design. However, in the recent interpretable CNNs [68], 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 an interpretable face representation (Fig. 1). We define interpretability in this way that when each dimension of the representation is able to represent a face structure or a face part, the face representation is of higher interpretability. Although the concept of part-based representations has been

* Denotes equal contribution by the authors. † Dr. Li and Dr. Shen contributed to this work while employed by Adobe Inc. Project page is at http://cvlab.cse.msu.edu/project-interpret-FR

around [2, 13, 26, 28], 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. [68], 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 promotes the practical value of interpretability; 3) we demonstrate superior interpretability, while achieving improved or similar face recognition performance on three face recognition benchmarks, compared to base CNN architectures.

2. Related Work

Interpretable Representation Learning Understanding the visual recognition has a long history in computer vision [22, 37, 43, 48, 49]. In early days when most models use hand-craft features, a number of research focused on how to interpret the predictions. Back then visual cues include image patches [22], object colors [50], body parts [62], face parts [26], or middle-level representations [48] contingent on the tasks. For example, Vondrick et al. [57] develop the HOGgles to visualize HOG descriptors in object detection. Since features such as SIFT [35], LBP [1] are extracted from image patches and serve as building blocks in the recognition pipeline, it was intuitive to describe the process from the level of patches. With the more complicated CNNs, it demands new tools to dissect its prediction. Early works include direct visualization of the filters [66], deconvolutional networks to reconstruct inputs from different layers [67], gradient-based methods to generate novel inputs that maximize certain neurons [39], and etc. Recent efforts along this line include CAM [71] which leverages the global max pooling to visualize dimensions of the representation and Grad-CAM [44] which relaxes the constraints on the network with a general framework to visualize any convolution filters. While our method can be related to visualization of CNNs and we leverage tools to visualize our learned filters, 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. Attention model [61] has been used in image caption generation. By attention mechanism, their model can push the feature maps responding separately to each predicted caption word, which is seemingly close to our idea, but needs many labeled data for training. One recent work on learning a more meaningful representation is the interpretable CNNs [68], where two losses regularize the training of late-stage convolutional filters: one to encourage each filter to encode a distinctive object part and another to push it to respond to only one local region. AnchorNet [40] adopts the similar idea to encourage orthogonality of filters and filter responses to keep each filter activated by a local and consistent structure. Our method generally extends the ideas in AnchorNet with new aspects 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, e.g., image synthesis/editing [9, 46], face modeling [51–53] and recognition [30, 69]. 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.

Parts and Occlusion in Face Recognition As an extensively studied topic [8, 25, 42], early face recognition works constructing meaningful representations mostly aim to improve the recognition accuracy. Some representations are composed from face parts. The part-based models are either learned unsupervisedly from data [27] or specified by manually annotated landmarks [6]. Besides local parts, face attributes are also interesting elements to build up representations. Kumar et al. [24] encode a face image with scores from attribute classifiers and demonstrate improved performance before the deep learning era. In this paper, we propose to learn meaningful part-based face representations with a deep CNN, through the carefully designed losses. We demonstrate how to leverage the interpretable representation for occlusion-robust face recognition. Prior methods addressing face pose variations [6, 7, 27, 31, 32, 55, 64, 65] can be related since pose changes may lead to self-occlusions. However, this work is interested in more explicit situations when faces are occluded by hand, sunglasses, and other objects. Interestingly, this specific aspect is rarely studied with CNNs. Cheng et al. [10] propose to restore occluded faces with deep auto-encoder for improved recognition accuracy. Zhou et al. [72] argue that naively training a high capacity network with sufficient coverage in training data could achieve superior accuracy. In our experiment, we indeed observe improved recognition accuracy to occluded faces after augmenting training 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 interpretable representation.

Occlusion Handling with CNNs Different methods are

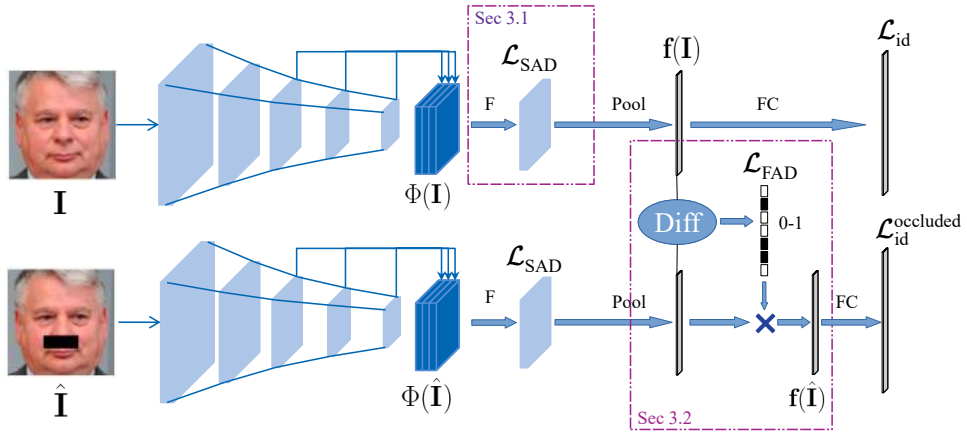

Figure 2. The overall network architecture of the proposed framework. While the Spatial Activation Diversity (SAD) loss promotes structured feature responses, Feature Activation Diversity (FAD) loss enforces them to be insensitive to local changes (occlusions).

proposed to handle occlusion with CNNs for robust object detection and recognition. Wang et al. [58] learn an object detector by generating an occlusion mask for each object, which synthesizes harder samples for the adversarial network. In [47], occlusion masks are utilized to enforce the network to pay attention to different parts of the objects. Ge et al. [14] detect faces with heavy occlusions by proposing a masked face dataset and applying it to their proposed LLE-CNNs. In contrast, our method enforces constraints for the spreadness of feature activations and guides the network to extract features from different face parts.

3. Proposed Method

Our 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 hypercolumn (HC) descriptor Φ, consisting of feature at multiple semantic levels. The proposed Spatial Activation Diversity (SAD) loss encourages the face representation to be structured with consistent semantic meaning. 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 to be insensitive to the occluded part, hence more robust to occlusion. At the same time, we mask out parts of the face representation sensitive to the occlusion and train to identify the input face solely based on the remaining elements. As a result, the filters respond to non-occluded parts are trained to capture more discriminative cues for identification.

3.1. Spatial Activation Diversity Loss

Novotny et al. [40] propose a diversity loss for semantic matching by penalizing correlations among filters weights and their responses. While their idea is general enough to extend to face representation learning, in practice, their design is not directly applicable due to the prohibitively large number of identities (classes) in face recognition. Their approach also suffers from degradation in recognition accuracy. 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. [40] propose 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 hypercolumn descriptor $\Phi(\mathbf{I})$, created by concatenating the filter responses of an image $\mathbf{I}$ at different convolutional layers [17]. 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* $\mathcal{L}_{\text{SAD}}^{\text{filter}}$ and $\mathcal{L}_{\text{SAD}}^{\text{response}}$, encouraging the orthogonality of the filters and of their responses, respectively. $\mathcal{L}_{\text{SAD}}^{\text{filter}}$ makes filters orthogonal by penalizing their correlations:

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

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 $\mathcal{L}_{\text{SAD}}^{\text{response}}$ is introduced to directly decorrelate the filters' *response maps* $\psi_i(\mathbf{I})$:

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

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 further apart by dilating their activations.

Our Proposed Modifications Novotny et al. [40] 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_i for the positive (or negative) class. In [40], the final feature representation $\mathbf{f}$ is obtained via global max-pooling operation on ψ. This design is not 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:

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

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 [40] where they observed minor performance degradation compared to the model without 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 the non-negative range $[0, +\infty)$. Hence it significantly limits the feasible learning space.

Most recent works choose to use global average pooling [55, 63]. However, when applying average-pooling to introduce interpretability, 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 aim to design a pooling operation that satisfies two objectives: 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 assign $d\%$ of elements with the smallest magnitude to 0. The size of the output remains the same. We apply $\mathcal{L}_{\text{SAD}}^{\text{response}}$ loss to 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 applied to $\text{LMF}(\psi(\mathbf{I}))$ to obtain the final representation $\mathbf{f}(\mathbf{I})$.

By removing small magnitude values from ψ_i, $\mathbf{f}$ 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.

3.2. Feature Activation Diversity Loss

One way to evaluate the effectiveness of the diversity loss is to compute the average location of the peaks within the kth response maps $\psi_i'(\mathbf{I})$ for an image set. If the average locations across K filters spread all over the face spatially, the diversity loss is well functioning and can associate each filer with a specific face area. With the SAD loss, we do observe the improved *spreadness* compared to the base CNN model trained without the SAD loss. Since we believe that more spreadness indicates *higher* interpretability, we hope to further boost the spreadness of the average peak locations across filters, i.e., 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 areas of a face image, and constrain on the difference between its feature response and that of the unoccluded original image. The second motivation for our proposal is to design an 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 experiments.

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:

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

By sharing the classifier's weights between two branches, this classifier is learned to be more robust to occlusion. It also leads to a better representation as filters respond to non-occluded parts need to be more discriminative.

3.3. Implementation Details

Our proposed method is model agnostic. To demonstrate this, we apply the SAD and FAD losses to two popu-

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	$3 \times 3/1$	$6 \times 6 \times K$
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	$3 \times 3/1$	$24 \times 24 \times K$
AvgPool	Ψ	$24 \times 24/1$	$1 \times 1 \times K$

lar network architectures: one inspired by the widely used CASIA-Net [56, 63], the other based on ResNet50 [19]. Tab. 1 shows the structure of the former. We add HC-descriptor-related blocks for our SAD loss learning. Conv33, conv44, conv54 layers are used to construct the HC descriptor via conv upsampling layers. We set the feature dimension $K = 320$. For ResNet50, we take the modified version in [12], where $K = 512$. We also construct the HC descriptor using 3 layers at different resolutions. To speed up training, we reuse the pretrained feature extraction networks shared by [56] and [12]. All new weights are randomly initialized using a truncated normal distribution with std of 0.02. The network is trained via SGD at an initial learning rate 10^{-3} and momentum 0.9. The learning rate is divided by 10 for twice when the training loss is stabled. We set the **LMF** rate ($d\%$) to $95.83\% = 1 - \frac{24}{24 \times 24}$, i.e., keeping 24 elements in a 24×24 feature response map. g_σ is a Gaussian kernel with $\sigma = 1.5$.

For FAD, the feature mask τ can be computed per image pair $\mathbf{I}$ and $\hat{\mathbf{I}}$. However, to obtain a more reliable mask, we opt to compute τ using multiple image pairs sharing the semantically equivalent occluded mask, i.e., $\tau_i(\{\mathbf{I}, \hat{\mathbf{I}}\}_{n=1}^N) = 1$ if $\frac{1}{N}\sum_{n=1}^N \left|\mathbf{f}_i(\mathbf{I}_n) - \mathbf{f}_i(\hat{\mathbf{I}}_n)\right| < t$, otherwise 0.

To mask the semantically equivalent local area of faces in a mini-batch regardless their poses, we first define a frontal face template with 142 triangles created by 68 landmarks. A 32×12 rectangle, randomly placed on the face, is selected as a normalized mask. Each of the rectangle's four vertices can be represented by the barycentric coordinate w.r.t. the triangle enclosing the vertex. For each image in a batch, corresponding four vertices of a quadrilateral can be found via the same barycentric coordinates. This quadrilateral denotes the location of a warped mask of that image (Fig. 3).

Figure 3. With barycentric coordinates, we warp the vertices of the template face mask to each image within the 64-image mini-batch.

Figure 4. Examples of (a) IJB-A, (b) IJB-C and (c) AR databases.

4. Experimental Results

Databases We train CASIA-Net with CASIA-WebFace [63], ReNet50 with MS-Celeb-1M [16], and test on IJB-A [23], IJB-C [3] and AR face [38] (Fig. 4). CASIA-WebFace contains $493,456$ images of $10,575$ subjects. MS-Celeb-1M includes 1M images of 100K subjects. Since it contains many labeling noise, we use a cleaned version of MS-Celeb-1M [16]. In our experiments, we evaluate IJB-A in three scenarios: original faces, synthetic occlusion and natural occlusion faces. For synthetic occlusion, we randomly generate a warped occluded area for each test image, as did in training. IJB-C extends IJB-A, also is a video-based face database with $3,134$ images and $117,542$ video frames of $3,531$ subjects. One unique property of IJB-C is its label on fine-grained occlusion area. Thus, we use IJB-C to evaluate occlusion-robust face recognition, using test images with at least one occluded area. AR face is another natural occlusion face database, with $\sim$4K faces of 126 subjects. We only use AR faces with natural occlusions, including wearing glasses and scarfs. Following the setting in [12], all training and test images are processed and resized to 112×112. Note all ablation and qualitative evaluations use CASIA-Net, while quantitative evaluations use both models.

4.1. Ablation Study

Different Thresholds We study the affect of the threshold of FAD loss. Here we use number-of-element-based thresholding. With an abuse of notation, t denotes the number of elements that the FAD loss encourages their similarity. We train different models with $t = 130, 260, 320$. The first three rows in Tab. 2 show the comparison of all three variants on IJB-A. When forcing all elements of $\mathbf{f}(\mathbf{I})$ and $\mathbf{f}(\hat{\mathbf{I}})$ to be the same ($t = K = 320$), the performance significantly drops on all three sets. In this case, the feature representation of the non-occluded face is negatively affected as being completely pushed toward a representation of the occluded one. While models with $t = 130$ and 260 perform similarly, we use $t = 260$ for the rest of the paper, given the observa-

Table 2. Ablation study on IJB-A database. 'BlaS': black mask with static sizes, 'GauD': Gaussian noise with dynamic sizes.

Method	IJB-A		Manual Occlusion		Natural Occlusion	
Metric (%)	@FAR=.01	@Rank-1	@FAR=.01	@Rank-1	@FAR=.01	@Rank-1
BlaS($t = 130$)	79.0 ± 1.6	89.5 ± 0.8	76.1 ± 1.7	88.0 ± 1.4	66.2 ± 4.0	73.0 ± 3.3
BlaS($t = 260$)	79.2 ± 1.8	89.4 ± 0.8	76.1 ± 1.4	88.0 ± 1.2	66.5 ± 6.4	72.3 ± 2.8
BlaS($t = 320$)	74.6 ± 2.4	88.9 ± 1.3	71.8 ± 3.1	87.5 ± 1.6	61.0 ± 6.5	71.6 ± 3.2
GauD($t = 260$)	$\mathbf{79.3} \pm 2.0$	$\mathbf{89.9} \pm 1.0$	$\mathbf{76.2} \pm 2.4$	$\mathbf{88.6} \pm 1.1$	$\mathbf{66.8} \pm 3.5$	$\mathbf{73.2} \pm 3.3$
SAD only	78.1 ± 1.8	88.1 ± 1.1	66.6 ± 5.6	81.2 ± 1.9	64.2 ± 6.9	71.0 ± 3.3
FAD only	76.7 ± 2.0	88.1 ± 1.1	75.2 ± 2.4	85.1 ± 1.2	66.5 ± 6.4	72.3 ± 2.8

tion that it makes occlusions affect less filters, pushes other filter responses away from any local occlusions, and subsequently enhances the spreadness of response locations.

Different Occlusions and Dynamic Window Size In FAD loss, we use the warped black window as the synthetic occlusion. It is important to introduce another type of occlusion to see its effects on face recognition. Thus, we use Gaussian noise to replace the black color in the window. Further, we employ a dynamic window size by randomly generating a value from $[12, 32]$ for both the window height and width. The face recognition results on IJB-A are shown in Tab. 2, where 'BlaS' means black window with static sizes, while 'GauD' means Gaussian noise window with dynamic sizes. It is interesting to find that the performance of 'GauD' is slightly better. Comparing to black window, Gaussian noise contains more diverse adversarial cues.

Spatial vs. Feature Diversity Loss Since we propose two different diversity losses, it is important to evaluate their respective effects on face recognition. As in Tab. 2, we train our models using either loss, or both of them. We observe that, while the SAD loss performs reasonably well on general IJB-A, it suffers from data with occlusions, being synthetic or natural. Alternatively, using only the FAD loss can improve the performance on the two occlusion datasets. Finally, using both losses, the row of 'BlaS($t = 260$)', improves upon both models with only one loss.

4.2. Qualitative Evaluation

Spreadness of Average Locations of Filter Response Given an input face, our model computes LMF($\psi(\mathbf{I})$), the 320 feature maps of size 24×24, where the average pooling of one map is one element of the final 320-d feature representation. Each feature map contains both the positive and negative response values, which are distributed at different spatial areas of the face. We select the locations of both the highest value for positive response and the lowest value for negative response as the *peak response locations*. To illustrate the spatial distribution of peak locations, we randomly select $1,000$ test images and calculate the weighted average location for each filter, with three notes. 1) there are two types of locations, for the highest (positive) and lowest (negative) responses respectively. 2) since the filters are responsive to semantic facial components, their 2D spatial locations may vary 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 w.r.t. the triangle enclosing it. Similar to Fig. 3, we use 68 estimated landmarks [21, 33] and control points on the image boundary to define the triangular mesh. 3) the weight of each image is determined by the magnitude of its peak response.

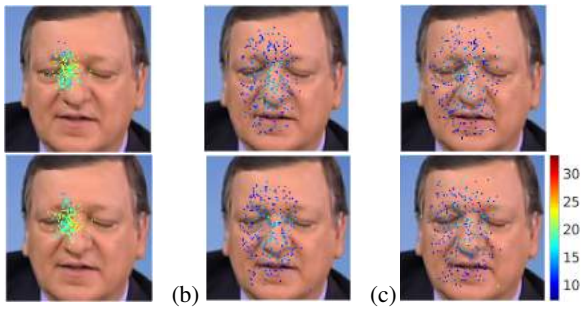

Figure 5. The average locations of 320 filters' peak responses (top: positive, bottom: negative responses) for three models: (a) base CNN ($\bar{d} = 6.9$), (b) our (SAD only, $\bar{d} = 17.1$), and (c) our model ($\bar{d} = 18.7$), where $\bar{d}$ quantifies the average locations spreadness. The color on each location denotes the standard deviation of peak locations. The face image size is 96×96.

Table 3. Compare standard deviations of peaks with varying d.

LMF (d%)	0	75.00	87.50	95.83
std (pos./neg.)	25.7/25.7	14.7/14.4	13.5/14.0	12.9/13.4

With that, the average locations for all feature maps are shown in Fig. 5. To compare the visualization results between our models and CNN base model, we compute $\bar{d} = \frac{1}{K}\sum_i^K \left\| c_i - \frac{1}{K}\sum_i^K c_i \right\|$ to quantify the average locations spreadness, where c_i denotes the (x, y) coordinates of the ith average location. For both the positive and negative peak response, we take the mean of their $\bar{d}$. As in Fig. 5, our model with SAD loss enlarges the spreadness of average locations. Further, our model with both losses continues 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, while in the base model all filters don't attach to specific facial part, results in average locations near the image center (Fig. 5 (a)). In addition, we compute the standard deviation for each filter's peak location. With much smaller standard deviations, our model can better concentrate on a local part than the base model.

In above analysis, we set the LMF rate d to 95.83%. It is worthy to ablate the impact of the rate d. We train models with $d = 0\%, 75\%, 87.5\%$ or 95.83%. Since before average pooling the feature map is of 24×24, the last 3 choices mean that we remove 24×18, 24×21 and 24×23 responses respectively and 0% denotes the base model. Tab. 3 compares the average of standard deviations of peak locations across 320 filters. Note the values of the best model $(12.9/13.4)$ equals to the average color of Fig. 5(c). When using a larger LMF rate, the model tends to be more concentrated onto a local facial part. For this reason, we set $d = 95.83\%$.

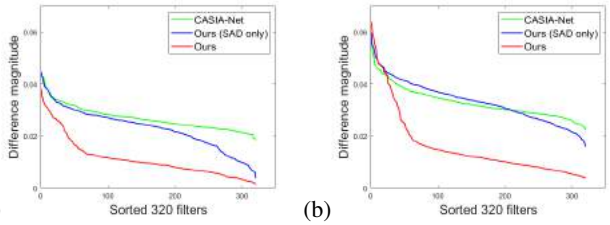

Figure 6. Mean of feature difference on two occluded parts: (a) eye part, and (b) nose part.

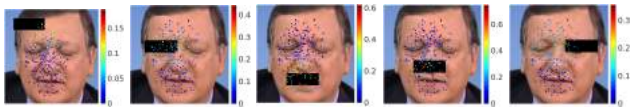

Figure 7. The correspondence between feature difference magnitude and occlusion locations. Best viewed electronically.

Mean Feature Difference Comparison Both of our losses promotes part-based feature learning, which leads to occlusion robustness. Especially, in FAD, we directly minimize the difference in a portion of representation of faces with and without occlusion. We now study the effect of our loss on faces with occlusion. Firstly, we randomly select $N = 1,000$ test faces in different poses and generate the synthetic occlusion. After that, for each filter, we calculate the mean of feature difference between the original and occluded faces $(\frac{1}{N}\sum_{n=1}^{N}|\mathbf{f}_i(\mathbf{I}_n) - \mathbf{f}_i(\hat{\mathbf{I}}_n)|)$ for $i = 1, 2, \ldots, K$. 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 difference. FDA loss further enhances robustness by only letting the occlusion modify a small portion of the representation, keeping the remaining elements invariant to the occluded part.

Visualization on Feature Difference Vectors Fig. 5 demonstrates that each of our filter spatially corresponds to a face location. Here we further study the relation of these average locations and semantic meaning on input images. In Fig. 7, we visualize the magnitude of each feature difference due to five different occlusions. We observe the locations of points with large feature difference are around the occluded face area, which means our learned filters are indeed sensitive to various facial areas. Further, the magnitude of the feature difference can vary with different occlusions. The maximum feature difference can be as high as 0.6 with occlusion in eye or mouth, meanwhile this number is only 0.15 in less critical area, e.g., forehead.

Filter Response Visualization Fig. 8 visualizes 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.

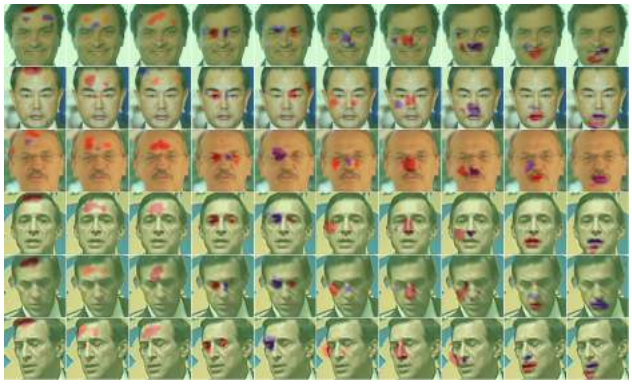

Figure 8. Visualization of filter response "heat maps" of 10 different filters on faces from different subjects (top 3 rows) and the same subject (bottom 3 rows). The positive and negative responses are shown as two colors within each image. Note the high consistency of response locations across subjects and across poses.

4.3. Quantitative Evaluation on Benchmarks

Our main objective is to show how we can improve the interpretability of face recognition while maintaining the recognition performance. Hence, the main comparison is between our proposed method and the base CNN model with the conventional softmax loss. Also, to show that our method is model agnostic, we use two different base CNN models, CASIA-Net and ResNet50. Our proposed method and the respective base model only differs in the loss functions. E.g., both our CASIA-Net-based model and base CASIA-Net model use the same network architecture as Tab. 1. Also, we perform data augmentation where the same synthetic faces that trained our models are fed to the training of base CASIA-Net model. We test on two types of datasets: the generic in-the-wild faces and occlusion faces.

Generic in-the-wild faces As shown in Tabs. 5, 6, when comparing to the base CASIA-Net model, our CASIA-Net-based model with two losses achieves the superior performance. The same superiority is demonstrated w.r.t. CASIA-Net with data augmentation, which shows that the gain is caused by the novel loss function design. For the deeper ResNet50 structure, our proposed model achieves similar performance as the base model, and both outperform the models with CASIA-Net as the base. Even comparing to state-of-the-art methods, the performance of our ResNet50-based model is still competitive. It is worthy note that this is the first time that *a reasonably interpretable representation is able to demonstrate competitive state-of-the-art recognition performance on a widely used benchmark*, e.g., IJB-A.

Occlusion faces We test our models and base models on multiple occlusion face datasets. The synthetic occlusion

Table 4. Comparison on three databases with occlusions.

Dataset	IJB-A synthetic occlusion				IJB-A natural occlusion				IJB-C natural occlusion			
Method ↓	Verification		Identification		Verification		Identification		Verification		Identification	
Metric (%) →	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
DR-GAN [56]	61.9 ± 4.7	35.8 ± 4.3	80.0 ± 1.1	91.4 ± 0.8	64.7 ± 4.1	41.8 ± 6.4	70.8 ± 3.6	81.7 ± 2.9	82.4	66.1	70.8	82.8
CASIA-Net	61.8 ± 5.5	39.1 ± 7.8	79.6 ± 2.1	91.4 ± 1.2	64.4 ± 6.1	40.7 ± 6.8	71.3 ± 3.5	81.6 ± 2.5	83.3	67.0	72.1	83.3
Ours (CASIA-Net)	76.2 ± 2.4	55.5 ± 5.7	88.6 ± 1.1	95.0 ± 0.7	66.8 ± 3.4	48.3 ± 5.5	73.2 ± 2.5	82.3 ± 3.3	83.8	69.3	74.5	83.6
ResNet50 [19]	93.0 ± 0.7	80.9 ± 4.7	92.8 ± 0.9	95.5 ± 0.8	$\mathbf{86.0} \pm 1.8$	64.3 ± 7.7	79.8 ± 4.2	84.9 ± 3.1	93.1	89.0	**87.5**	**91.0**
Ours (ResNet50)	$\mathbf{94.2} \pm 0.6$	$\mathbf{87.5} \pm 1.5$	$\mathbf{93.4} \pm 0.7$	$\mathbf{95.8} \pm 0.4$	$\mathbf{86.0} \pm 1.6$	$\mathbf{72.6} \pm 5.0$	$\mathbf{80.0} \pm 3.2$	$\mathbf{85.0} \pm 3.1$	**93.4**	**89.8**	87.4	90.7

Table 5. Comparison on IJB-A database.

Method ↓	Verification		Identification	
Metric (%) →	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
DR-GAN [56]	79.9 ± 1.6	56.2 ± 7.2	88.7 ± 1.1	95.0 ± 0.8
CASIA-Net	74.3 ± 2.8	49.0 ± 7.4	86.6 ± 2.0	94.2 ± 0.9
CASIA-Net agu.	78.9 ± 1.8	56.6 ± 4.8	88.5 ± 1.1	94.9 ± 0.8
Ours (CASIA-Net)	79.3 ± 2.0	60.2 ± 5.5	89.9 ± 1.0	95.6 ± 0.6
FaceID-GAN [45]	87.6 ± 1.1	69.2 ± 2.7	–	–
VGGFace2 [5]	93.9 ± 1.3	85.1 ± 3.0	$\mathbf{96.1} \pm 0.6$	$\mathbf{98.2} \pm 0.4$
PRFace [4]	94.4 ± 0.9	86.8 ± 1.5	92.4 ± 1.6	96.2 ± 1.0
ResNet50 [19]	$\mathbf{94.8} \pm 0.6$	86.0 ± 2.6	94.1 ± 0.8	96.1 ± 0.6
Ours (ResNet50)	94.6 ± 0.8	$\mathbf{87.9} \pm 1.0$	93.7 ± 0.9	96.0 ± 0.5

Table 6. Comparison on IJB-C database.

Method ↓	Verification		Identification	
Metric (%) →	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
DR-GAN [56]	88.2	73.6	74.0	84.2
CASIA-Net	87.1	72.9	74.1	83.5
Ours (CASIA-Net)	89.2	75.6	77.6	86.1
VGGFace2 [5]	95.0	90.0	89.8	**93.9**
Mn-v [60]	**96.5**	92.0	–	–
AIM [70]	96.2	**93.5**	–	–
ResNet50 [19]	95.9	93.2	**90.5**	93.2
Ours (ResNet50)	95.8	93.2	90.3	93.2

of IJB-A, the natural occlusion of IJB-A, and the natural occlusion of IJB-C have $500/25,795$, $466/12,703$, and $3,329/78,522$ subjects/images, respectively. As shown in Tab. 4, 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 AR faces, we select all 810 images with eyeglasses and scarfs occlusions, from which $6,000$ same-person and $6,000$ different-person pairs are randomly selected. We compute the representations of an image pair and its cosine distance. The Equal Error Rates of CASIA-Net, ours (CASIA-Net), ResNet50 and ours (ResNet50) are 21.6%, 16.2%, 4.2% and 3.9%, respectively.

4.4. Other Application

Despite the interpretable face recognition, another potential application of our method is partial face retrieval. Assuming we are interested in retrieving images with similar noses, we can define "nose filters" base on filters' average peak location with our models, as in Fig. 5. Then, the part-based feature for retrieving is constructed by masking out all elements in the identity feature $\mathbf{f}(\mathbf{I})$ except selected part-related filters. For demonstration, from IJB-A test set, we select one pair of images from a subset of 150 identities, to create a set of 300 images in total. Using different facial parts of each image as a query, our accuracies of retrieving the remaining image of the same subject as the rank-1 result are 71%, 58% and 69% for eyes, mouth, and nose respectively. Results are visualized in Fig. 9, we can retrieve facial parts that are not from the same identity but visually very similar to the query part.

Figure 9. Partial face retrieval with mouth (left), and nose (right).

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 where each dimension activates on a consistent semantic face part and captures its identity information. We empirically demonstrate the proposed method, enabled by our novel loss function design, can lead to more locally constrained individual filter responses and overall widely-spreading filters distribution, yet maintaining SOTA face recognition performance. A by-product of the harnessed interpretability is improved robustness to occlusions in face recognition.

Acknowledgement This work is partially sponsored by Adobe Inc. and Army Research Office under Grant Number W911NF-18-1-0330. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *TPAMI*, 2006. 2

[2] Thomas Berg and Peter N Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013. 2

[3] Jocelyn Adams Brianna Maze, Nathan Kalka James A. Duncan, Charles Otto Tim Miller, W. Tyler Niggel Anil K. Jain, Jordan Cheney Janet Anderson, and Patrick Grother. IARPA Janus Benchmark-C: Face dataset and protocol. In *ICB*, 2018. 5

[4] Kaidi Cao, Yu Rong, Cheng Li, Xiaoou Tang, and Chen Change Loy. Pose-robust face recognition via deep residual equivariant mapping. In *CVPR*, 2018. 8

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018. 8

[6] Zhimin Cao, Qi Yin, Xiaoou Tang, and Jian Sun. Face recognition with learning-based descriptor. In *CVPR*, 2010. 2

[7] Xiujuan Chai, Shiguang Shan, Xilin Chen, and Wen Gao. Locally linear regression for pose-invariant face recognition. *TIP*, 2007. 2

[8] Tsuhan Chen, Yufeng Jessie Hsu, Xiaoming Liu, and Wende Zhang. Principle component analysis and its variants for biometrics. In *ICIP*, 2002. 2

[9] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016. 2

[10] Lele Cheng, Jinjun Wang, Yihong Gong, and Qiqi Hou. Robust deep auto-encoder for occluded face recognition. In *ICM*, 2015. 2

[11] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *CVIU*, 2017. 1

[12] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 5

[13] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 2

[14] Shiming Ge, Jia Li, Qiting Ye, and Zhao Luo. Detecting masked faces in the wild with lle-cnns. In *CVPR*, 2017. 3

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1

[16] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 5

[17] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 1

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 5, 8

[20] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. In *ECCV*, 2014. 1

[21] Amin Jourabloo and Xiaoming Liu. Pose-invariant face alignment via CNN-based dense 3D model fitting. *IJCV*, 2017. 6

[22] Mayank Juneja, Andrea Vedaldi, CV Jawahar, and Andrew Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013. 2

[23] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark-A. In *CVPR*, 2015. 5

[24] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 2

[25] Erik Learned-Miller, Gary B Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua. Labeled faces in the wild: A survey. In *Advances in face detection and facial image analysis*. 2016. 2

[26] Haoxiang Li and Gang Hua. Probabilistic elastic part model: a pose-invariant representation for real-world face verification. *TPAMI*, 2017. 2

[27] Haoxiang Li, Gang Hua, Zhe Lin, Jonathan Brandt, and Jianchao Yang. Probabilistic elastic matching for pose variant face verification. In *CVPR*, 2013. 2

[28] Stan Z Li, Xin Wen Hou, Hong Jiang Zhang, and Qian Sheng Cheng. Learning spatially localized, parts-based representation. In *CVPR*, 2001. 2

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 1

[30] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Disentangling features in 3D face shapes for joint face reconstruction and recognition. In *CVPR*, 2018. 2

[31] Xiaoming Liu and Tsuhan Chen. Geometry-assisted statistical modeling for face mosaicing. In *ICIP*, 2003. 2

[32] Xiaoming Liu, Jens Rittscher, and Tsuhan Chen. Optimal pose for face recognition. In *CVPR*, 2006. 2

[33] Yaojie Liu, Amin Jourabloo, William Ren, and Xiaoming Liu. Dense face alignment. In *ICCV Workshop*, 2017. 6

[34] Yu Liu, Hongyang Li, and Xiaogang Wang. Learning deep features via congenerous cosine loss for person recognition. *arXiv:1702.06890*, 2017. 1

[35] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 2

[36] Chaochao Lu and Xiaoou Tang. Surpassing human-level face verification performance on LFW with gaussianface. In *AAAI*, 2015. 1

[37] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *IJCV*, 2016. 2

[38] Aleix M Martinez. The AR face database. *CVC Technical Report24*, 1998. 5

[39] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 2015. 2

[40] David Novotny, Diane Larlus, and Andrea Vedaldi. Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In *CVPR*, 2017. 2, 3, 4

[41] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018. 1

[42] Alice J. OToole, Carlos D. Castillo, Connor J. Parde, Matthew Q. Hill, and Rama Chellappa. Face space representations in deep convolutional neural networks. *Trends in cognitive sciences*, 2018. 2

[43] Devi Parikh and C Zitnick. Human-debugging of machines. *NIPS WCSSWC*, 2011. 2

[44] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 2

[45] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. Faceid-gan: Learning a symmetry three-player GAN for identity-preserving face synthesis. In *CVPR*, 2018. 8

[46] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017. 2

[47] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*. 3

[48] Saurabh Singh, Abhinav Gupta, and Alexei A Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*. 2012. 2

[49] Erik B Sudderth, Antonio Torralba, William T Freeman, and Alan S Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, 2005. 2

[50] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013. 2

[51] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3D face morphable model. In *CVPR*, June 2019. 2

[52] Luan Tran and Xiaoming Liu. Nonlinear 3D face morphable model. In *CVPR*, 2018. 2

[53] Luan Tran and Xiaoming Liu. On learning 3D face morphable model from in-the-wild images. *TPAMI*, 2019. 2

[54] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *CVPR*, 2017. 1

[55] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *CVPR*, 2017. 2, 4

[56] Luan Tran, Xi Yin, and Xiaoming Liu. Representation learning by rotating your faces. *TPAMI*, 2018. 5, 8

[57] Carl Vondrick, Aditya Khosla, Tomasz Malisiewicz, and Antonio Torralba. Hoggles: Visualizing object detection features. In *ICCV*, 2013. 2

[58] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *CVPR*, 2017. 3

[59] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 1

[60] Weidi Xie and Andrew Zisserman. Multicolumn networks for face recognition. In *BMVC*, 2018. 8

[61] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2

[62] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011. 2

[63] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014. 4, 5

[64] Xi Yin and Xiaoming Liu. Multi-task convolutional neural network for pose-invariant face recognition. *TIP*, 2018. 2

[65] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017. 2

[66] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 2

[67] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *ICCV*, 2011. 2

[68] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *CVPR*, 2018. 1, 2

[69] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Jian Wan, Nanxin Wang, and Xiaoming Liu. Gait recognition via disentangled representation learning. In *CVPR*, 2019. 2

[70] Jian Zhao, Yu Cheng, Yi Cheng, Yang Yang, Haochong Lan, Fang Zhao, Lin Xiong, Yan Xu, Jianshu Li, Sugiri Pranata, Shengmei Shen, Junliang Xing, Hengzhu Liu, Shuicheng Yan, and Jiashi Feng. Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. In *AAAI*, 2019. 8

[71] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2

[72] Erjin Zhou, Zhimin Cao, and Qi Yin. Naive-deep face recognition: Touching the limit of LFW benchmark or not? *arXiv:1501.04690*, 2015. 2