Occlusion Robust Face Recognition Based on Mask Learning With Pairwise Differential Siamese Network

Lingxue Song[12], Dihong Gong[1], Zhifeng Li*[1], Changsong Liu[2] and Wei Liu[1]
[1]Tencent AI Lab [2]Tsinghua University
songlx15@mails.tsinghua.edu.cn gongdihong@gmail.com michaelzfli@tencent.com
lcs@tsinghua.edu.cn wl2223@columbia.edu

Abstract

Deep Convolutional Neural Networks (CNNs) have been pushing the frontier of face recognition over past years. However, existing general CNN face models generalize poorly for occlusions on variable facial areas. Inspired by the fact that the human visual system explicitly ignores the occlusion and only focuses on the non-occluded facial areas, we propose a mask learning strategy to find and discard corrupted feature elements from recognition. A mask dictionary is firstly established by exploiting the differences between the top conv features of occluded and occlusion-free face pairs using innovatively designed pairwise differential siamese network (PDSN). Each item of this dictionary captures the correspondence between occluded facial areas and corrupted feature elements, which is named Feature Discarding Mask (FDM). When dealing with a face image with random partial occlusions, we generate its FDM by combining relevant dictionary items and then multiply it with the original features to eliminate those corrupted feature elements from recognition. Comprehensive experiments on both synthesized and realistic occluded face datasets show that the proposed algorithm significantly outperforms the state-of-the-art systems.

1. Introduction

Deep Convolutional Neural Networks (CNNs) have recently made a remarkable improvement in unconstrained face recognition problem. Researchers are racing in ways to boost the performance using advanced network architectures [25, 3, 27, 6, 32] or designing new loss functions to facilitate discriminative feature learning [24, 33, 13, 31, 2, 42, 30]. Some of them even surpass human recognition ability on certain benchmark database [7].

*Corresponding author

Despite the huge success of deep learning models under general face recognition scenario, the deep features still show imperfect invariance to uncontrollable variations like pose, facial expression, illumination, and occlusion. Among all these factors, occlusion has been considered a highly challenging one. In real-life images or videos, facial occlusions can often be observed, *e.g*. facial accessories including sunglasses, scarves, and masks or other random objects like books and cups. As indicated in [17], without specifically trained with a large number of occluded face images, deep CNN-based models indeed cannot function well because of the larger intra-class variation and higher inter-class similarity that caused by occlusions.

One possible approach to improve the performance of CNN models under partial occlusions is to train the network with occluded faces. Daniel *et al*. [28] proposed to augment training data with synthetic occluded faces in a strategic manner and observed improved performance. However, it does not solve the problem intrinsically because it only ensures the features are more locally and equally extracted, as analyzed in [21]. The inconsistency between features of two faces with different occlusion situations still exists. For example, features of an occlusion-free face bear much more information in eyes area than that of a face wearing a pair of sunglasses unless the network is trained not to utilize the eyes area at all, which is unreasonable.

Inspired by the fact that the human visual system pays attention to the non-occluded facial areas for recognition (and ignores the occluded areas), we propose to discard feature elements that have been corrupted by occlusions. A core question would be: *given a face image with random partial occlusions, how to locate those corrupted feature elements?* It is not a big deal for traditional low-level features like LBP, HOG or SIFT because there is a clear correspondence between image pixels and final feature elements, but what about the deep CNN features? Therefore, the core of this work is to find corrupted feature elements under random partial occlusion and eliminate the response of these

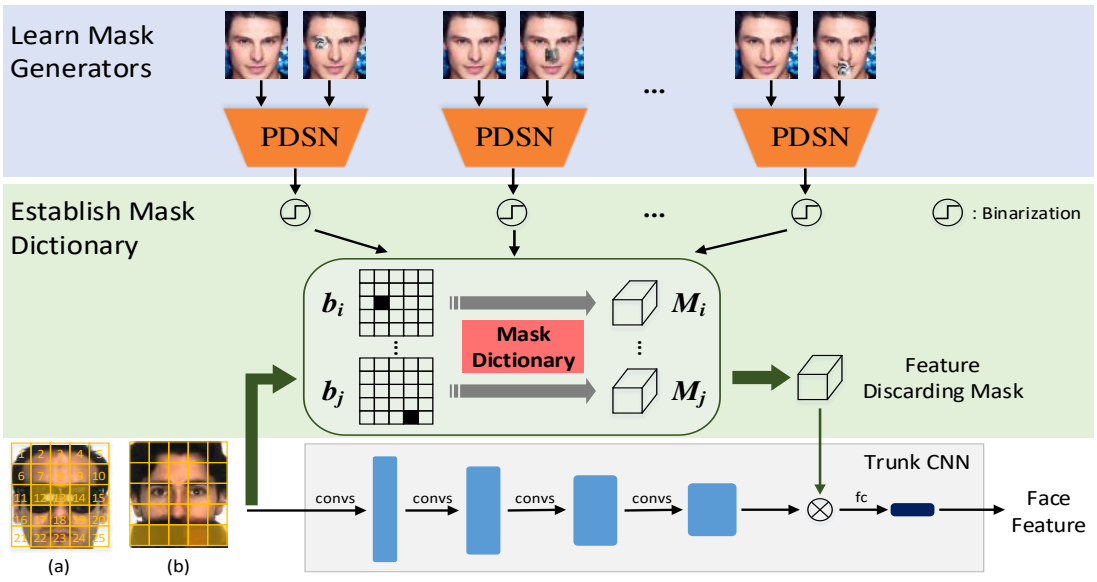

Figure 1. An overview of the proposed framework. Based on a trunk CNN model trained for face recognition, we propose the pairwise differential siamese network (PDSN) structure to learn the correspondence between occluded facial blocks and corrupted feature elements. Then a mask dictionary is established accordingly, which is used to composite the feature discarding mask (FDM) for a test face with random partial occlusions. Finally, we multiply the FDM with original face features to eliminate corrupted feature elements from recognition.

elements from the recognition process. It is worth stated that the facial occlusion detection problem is not the focus of this paper, thus we directly adopt a similar way as [23] to detect the occlusion location in image space.

To learn the correspondence between occluded facial regions and corrupted feature elements, we develop a novel pairwise differential siamese network (PDSN) structure with a mask generator module that takes pairwise images (one is a clean face and the other is an occluded one of the same identity) as input. The differential signal between the conv features of clean and corresponding occluded faces is fed into the mask generator module. It acts as a role of attention mechanism which encourages the model to focus on those feature elements that have deviated from its true values owing to partial occlusions. Moreover, we propose to learn the mask generator by minimizing a combination of two losses: a pairwise contrastive loss that penalizes the large differences between the masked conv features of clean and occluded faces, and a classification loss which ensures those feature elements that harm the recognition are masked out. With these two losses, our mask generator will identify those feature elements that are harmful for the recognition as well as far from its genuine values as corrupted ones. To handle the random partial occlusions, we first divide the aligned face into several predefined blocks and only learn PDSNs for these blocks, since severely performance dropping usually only occurs when critical facial components are missing. Then we construct a mask dictionary from these trained PDSNs by strategic binarization. Each item in this dictionary is a binary mask, named Feature Discarding Mask (FDM), which indicates the feature elements that should be set to zero when one facial block is occluded. In the testing phase, the FDM of a face with random partial occlusions is derived by element-wise logical "AND" of relevant dictionary items, which is then multiplied with the original face feature to discard those corrupted feature elements from the recognition. Figure 1 gives an overview of the proposed framework.

The main contributions of this paper are two-fold: (1) we propose a novel PDSN framework to explicitly find correspondence between occluded facial blocks and corrupted feature elements for deep CNN models, which is innovative and inspiring; (2) based on the PDSN, we develop a face recognition system that is robust for occlusions. Our system demonstrates superior performance on face datasets with both realistic and synthesized occlusions and generalizes very well on general face recognition tasks.

2. Prior Work

Partial occlusion is one of the major challenges in face recognition that has received much attention in the era of hand-crafted features. Before the emergence of deep CNNs, face recognition under partial occlusions has been typically handled using two types of algorithms, namely, (i) methods that extract local face descriptors only from the non-occluded facial areas or (ii) methods that recover clean faces from the occluded ones. The first type usually explicitly

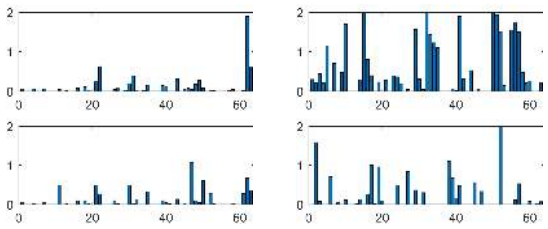

Figure 2. Neural response differences between two face images of different subjects with the same partial occlusion. Left: neural activation differences of the top conv layer. Right: neural activation differences of the top fc layer. We randomly sample 64 neurons for illustration here.

divides face images into several local regions. A support vector machine (SVM) is trained to identify which local regions are occluded and then they are discarded from recognition [20, 18, 22], with optional subspace methods [10, 12] to reduce feature dimension before the classification stage. However, the discriminative power of this kind of approach is limited in view of using shallow features like Local Gabor Binary Patterns (LGBP) [20]. Among the second type of methods, the sparse representation-based classification (SRC) [36] is considered to be the pioneering work on occlusion robust face recognition. This model reconstructs an occlusion-free face using a linear combination of images from the training set together with a sparse constraint term accounting for occlusions. Inspired by this model, researchers extended it by rethinking the distribution of the sparse constraint term [39, 5, 4] or characterizing the structure information of it [44, 9]. These approaches do not generalize well since they require test samples have identical subjects with the training samples.

Deep learning has been dominant in the field of face recognition for several years. As early as 2014, Sun *et al*. [26] have discovered that the feature learned by DeepID2+ show certain degree of robustness to image corruption in face verification task. Combining DeepID2+ features extracted from 25 face patches further improves the robustness. Cheng *et al*. [43] present an LSTM-Autoencoder to restore occluded facial areas in the wild and carried out recognition on the recovered face images. But there is no guarantee the recovered part indeed matches the identity of the individuals to be recognized especially under the open-set scenario. Daniel *et al*. [28] tackle the occlusion problem by augmenting training data with synthetic occluded faces that only specific regions where a CNN extracts the most discriminative features from are covered. In this way, features are more equally and locally extracted. Wan *et al*. [29] propose to add a MaskNet branch to the middle layer of CNN models which is expected to assign lower weights to hidden units activated by the occluded facial areas. But the middle conv layer is not discriminative enough and the MaskNet branch lacks additional supervision information to ensure the functionality.

In a word, the discriminative ability of traditional low-level feature-based methods is limited, and the existing few deep learning-based methods lack the awareness of how partial occlusions truly affect the CNN models. The inconsistency between features of two faces with different occlusion situations has not been carefully considered yet. The proposed method complements the missing piece of the puzzle and is able to explicitly locate corrupted feature elements for trained CNN models and discard them from the recognition, to ensure a fair comparison. Therefore, our approach is an intrinsic way with good generalization ability compared to the aforementioned studies.

3. Proposed Approach

The overall pipeline of the proposed approach is shown in Figure 1, which decomposes the problem of face recognition under random partial occlusions into three stages. *Stage* I: Learning mask generators using the proposed pairwise differential siamese network (PDSN) to capture the correspondence between occluded facial blocks and corrupted feature elements. *Stage* II: Establishing a mask dictionary from the learned mask generators. *Stage* III: In the testing phase, combining the feature discarding mask (FDM) of random partial occlusions from this dictionary, which is then multiplied with the original feature to eliminate the effect of partial occlusions from recognition.

3.1. Stage I: Learning Mask Generators

3.1.1 Problem Analysis

Face images fed into the CNN model are mostly well-aligned by detected facial keypoints, we divide the aligned face into non-overlapping $N \times N$ blocks, denoted as $\{b_j\}_{j=1}^{N*N}$, and aim to learn a mask generator for every b_j to find the corrupted feature elements when this block is occluded. In our implementation, we set $N = 5$ according to the input image size so that the facial components like eyes, nose tip and mouth are appropriately associated with a block. The face (a) in Figure 1 gives the division example.

Then we define our core problem in Stage I as: given the feature of a face image with block b_j occluded, denoted as $f(x_j)$, how to learn a mask generator $\mathscr{M}_\theta$ whose output is multiplied with the $f(x_j)$ to mask out those corrupted elements. Let the purified feature be denoted as $\tilde{f}(x_j)$, then $\tilde{f}(x_j) = \mathscr{M}_\theta(\cdot)f(x_j)$. There are two choices to be decided before running into the learning process:

The choice of f. For the CNN-based face recognition model, the face feature usually refers to the output of the final fully-connected (fc) layer before the classification layer. However, every neuron in the fc layer integrates information from all the output elements of the previous layer, so the oc-

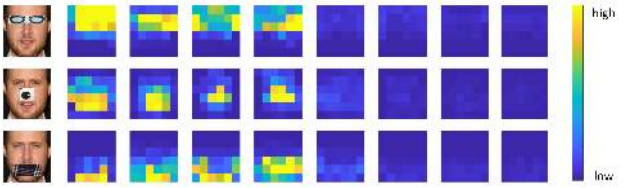

Figure 3. The median relative rate of change (MED) of neuron's activation values in the top conv layer under three types of occlusions. We select eight channels for illustration here.

cluded areas might be mixed up with the non-occluded areas in the final fc feature. From another perspective, neurons in the top fc layers are highly selective to identities [26]. Therefore even if different subjects are contaminated by the same occlusion, the positions of feature elements that changed by this occlusion will be highly dependant on face identity, as shown in the rightmost column in Figure 2. In contrast, we can see from the left column of Figure 2 that the positions of feature elements that changed by the same occlusion of different individuals are quite consistent for the top conv layer, and it still preserves local information, thus we choose the top conv feature as our f.

The output dimension of $\mathscr{M}_\theta$. In [29], they learned a 2D mask $M \in \mathbb{R}^{W\times H}$ for the 3D conv feature maps $U \in \mathbb{R}^{C\times W\times H}$. That is to say, the feature elements of all C channels in the same spatial location share the same weight from their learned mask. In other words, they assumed that feature elements of all the conv feature channels respond the same to the occlusion. With questions about the rationality of their hypothesis, we'd like to gain more insights into the true reaction of the top conv feature to partial occlusions. We use a criterion named *median relative rate of change* (denoted as MED) to capture the extent to which each feature element is away from its true value under partial occlusions. Given a pair of clean face image x_{clean} and its corresponding occluded face image x_{occ}, we first calculate the *relative rate of change* of neuron activation values of the top conv layer:

$$r_i = |\frac{f^i(x_{clean}) - f^i(x_{occ})}{f^i(x_{clean})}| \quad (1)$$

where r_i denotes the *relative rate of change* of the i^{th} feature element value of the top conv layer. We randomly select N images from the CASIA-WebFace [40] and add occlusions on the faces, then calculate the r_is for every face pair. The metric MED to approximately represents the altered degree of the i^{th} feature element under occlusions is obtained by calculating the median value of these r_is. If the MED of a feature element is high when an area of the input face is occluded, then it will likely bring unreasonable noise into the final feature.

In Figure 3 we show the MED values of feature elements in 8 channels of the top conv feature maps under three types of occlusions. Obviously, the feature values are altered in a different way for different channels, elements of some channels change very little while elements of some channels change drastically in the same spatial locations. This is interesting because in view of the receptive field, the same spatial location of different conv channels gather information from the same region of the input image, but they actually react quite differently to occlusions. Therefore, we believe the output dimension of $\mathscr{M}_\theta$ should be the same as the top conv feature maps, which is $C \times W \times H$.

3.1.2 Pairwise Differential Siamese Network

Given the analysis in Sec. 3.1.1, we propose the pairwise differential siamese network (PDSN) structure to learn the relations between occluded facial blocks and corrupted feature elements. As illustrated in Figure 4, it consists of a trunk CNN and a mask generator branch, forming a siamese architecture. The trunk CNN is responsible for extracting base face representation, which is shared by the clean and occluded face pairs and could be any CNN architecture. The core mask generator module $\mathscr{M}_\theta$ in our PDSN is expected to output a mask whose element is a real value in $[0, 1]$ and is multiplied with the input contaminated feature to diminish its corrupted elements: $\tilde{f}(x_j^i) = \mathscr{M}_\theta(\cdot)f(x_j^i)$, where $f(\cdot)$ is top conv feature and x_j^i denotes occluded face image of the i^{th} pair. The two faces inside an input pair belong to the same identity y^i and the only difference is that one of them has partial occlusion on the facial block b_j. The key requirement for learning the mask generator is that the remaining part of the feature $f(x_j^i)$ after masking should be as similar to its corresponding clean feature $f(x^i)$ as possible while guarantees a successful recognition.

To this end, we propose to learn $\mathscr{M}_\theta$ by minimizing a combination of two losses:

$$L_\theta = \sum_i \ell_{cls}(\theta; \tilde{f}(x_j^i), y^i) + \lambda \ell_{diff}(\theta; \tilde{f}(x_j^i), \tilde{f}(x^i)) \quad (2)$$

The first part of the cost, ℓ_{cls}, is accounting for evaluating the importance of each feature element for recognition, and the second part, ℓ_{diff}, assesses how far the feature of an occluded face is away from its true value. We will expand this formulation in the following part.

The classification loss ℓ_{cls}. To find the corrupted feature elements, an intuitive idea is that, these feature elements contribute little to identifying the input face and may instead cause higher classification loss. Therefore the most straightforward supervision signal is the identity information, that is, the occluded face should be correctly classified by the classifier of the trunk CNN after masking, which gives us the first loss item (softmax loss for example):

$$\ell_{cls}(\theta; \tilde{f}(x_j^i), y^i) = -log(p_{y^i}(F(\tilde{f}(x_j^i)))) \quad (3)$$

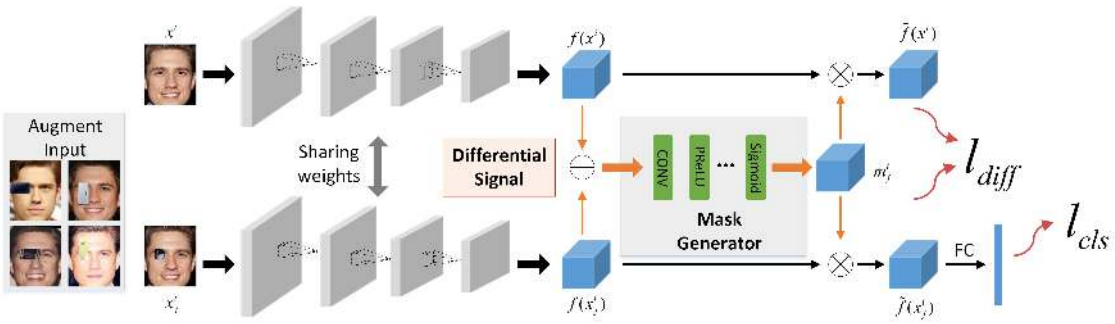

Figure 4. The illustration of the proposed Pairwise Differential Siamese Network.

The $\tilde{f}(x_j^i)$ is the top conv feature of an occluded face after masking, F is the fc layer(s) of the trunk CNN model next to the top conv layer, and it could also be the average pooling layer in models like [13].

The differential signal and pairwise loss ℓ_{diff}. The results shown in Figure 3 inspired us that the differential signal between the top conv activation values of an occluded face and its corresponding clean one could be a good indicator of which feature elements are potential corrupted ones. To put it another way, the differential input signal acts as a role of attention mechanism which encourages the mask generator to focus on those feature elements that have deviated from its true values owing to partial occlusion. Therefore we feed our mask generator module with the absolute difference between features of an occlusion-free face and its occluded counterpart.

To further make use of the supervision information of what this subject's occlusion-free feature looks like, we propose a pairwise contrastive loss that minimizes per-element differences between the masked features of the occluded and occlusion-free faces:

$$\ell_{diff}(\theta; \tilde{f}(x_j^i), \tilde{f}(x^i)) = \|\mathscr{M}_\theta(\cdot)f(x^i) - \mathscr{M}_\theta(\cdot)f(x_j^i)\|_1 \quad (4)$$

where $\mathscr{M}_\theta(\cdot) = \mathscr{M}_\theta(|f(x_j^i) - f(x^i)|)$, and $\|\cdot\|_1$ is the L1 norm. Obviously, this contrastive loss will punish those feature elements of the occluded face which are largely different from its occlusion-free one. Together with the classification loss, our mask generator will identify those feature elements that are harmful for the recognition as well as far from its genuine values as corrupted ones.

Thus, the overall object function in Eq. (2) used in our implementation is:

$$L_\theta = -\sum_i log(p_{y_i}(F(\mathscr{M}_\theta(\cdot)f(x_j^i)))) + \lambda\|\mathscr{M}_\theta(\cdot)f(x^i) - \mathscr{M}_\theta(\cdot)f(x_j^i)\|_1 \quad (5)$$

The λ is set to 10 to make the different components of the object function have the same scale in our experiments.

We implement $\mathscr{M}_\theta$ as a module with several conv blocks and learn the different θ for occlusions on different facial blocks. The different θ is accounting for the distinct property of different facial components. For example, the eyes bear much more significance than the cheek area, therefore the input distribution of the mask generator varies accordingly. When learning mask generator j, in addition to the faces that only the target block b_j is occluded, we augment samples with other blocks also occluded, which are the 4-neighbors of the target block b_j, to capture the dependency of adjacent blocks, as shown in Figure 4.

3.2. Stage II: Establishing the Mask Dictionary

In the testing phase, we don't have the paired images of a probe face and its occlusion location is random. Therefore, the trained PDSNs cannot be directly used to output the feature discarding mask(FDM) of a probe face. In Stage II, we would like to extract a fixed mask from every trained mask generator $\mathscr{M}_\theta$ and build a dictionary accordingly.

$$M_j[k] = \begin{cases} 0 & \text{if } \bar{m}_j[k] \in \{\tilde{m}_j[1], \ldots, \tilde{m}_j[\tau * K]\}, \\ 1 & \text{else.} \end{cases} \quad (6)$$

For a mask generator $\mathscr{M}_{\theta_j}$, we first feed the trained network with large amount of face pairs, one of which is occluded on the j^{th} facial block and obtain the output masks of this generator, forming a large set of $m_j^1, m_j^2, \ldots, m_j^P$, where P (about 200k in our experiment) is the number of the face pairs. After Min-Max normalizing each m_j^i, we calculate the element-wise mean value of these m_j^is and get a mean mask $\bar{m}_j$. It's possible to directly use this $\bar{m}_j$ as the FDM when the j^{th} block is occluded (referred to as *soft weight* schema). But this will reserve feature elements with very low mask values, which is inappropriate since the facial components inside this block have been totally lost. Therefore we believe setting those feature elements to zero to completely remove the noise is critical. We'll validate this in Sec. 4.2. The binarized FDM $M_j \in \mathbb{R}^{C\times W\times H}$ for this mask generator is derived by setting the feature loca-

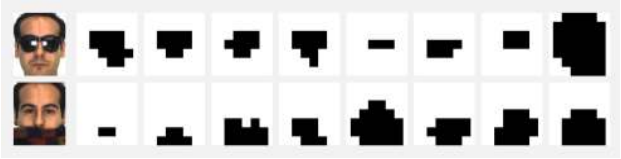

Figure 5. Examples of the feature discarding masks of two occlusion types combined from our mask dictionary.

Figure 6. Occlusion detection results of our FCN-8s segmentation network on the occluded Facesrub and AR test images.

tions with the smallest top $\tau * K$ mean values to zero:

$$M_j[k] = \begin{cases} 0 & \text{if } \bar{m}_j[k] \in \{\tilde{m}_j[1], \ldots, \tilde{m}_j[\tau * K]\}, \\ 1 & \text{else.} \end{cases} \quad (7)$$

where $k = 1, 2, \ldots, K$, $K = C \times W \times H$, k denotes the feature index, and $\{\tilde{m}_j[1], \ldots, \tilde{m}_j[\tau * K]\}$ is the sorted smallest $\tau * K$ values of $\bar{m}_j$. τ is the discarding threshold and it will be discussed later in Sec. 4.2. By this way, we construct a mask dictionary that every item is a binary mask which indicates whether to discard each feature element when one certain block of the aligned face is occluded.

3.3. Stage III: Occlusion Robust Recognition

With this mask dictionary, the FDM of a face with arbitrary partial occlusions could be derived by combining relevant dictionary items. By relevant we mean that if the occlusion area in a probe face has at least 0.5 IoU with a predefined facial block from the dictionary, we count this block as an occluded one for this face. For example, for the face (a) wearing sunglasses in Figure 1, its occlusion region covers block $\{b_j\}_{j=12}^{14}$, therefore its FDM is calculated by $M = M_{12} \wedge M_{13} \wedge M_{14}$, where $\wedge$ denotes the element-wise logical "AND" and the result M is still a binary mask. Figure 5 shows 8 channels of the FDM composited from the dictionary for sunglasses and scarf occlusions respectively.

4. Experiments

4.1. Implementation Details

Preprocessing. The standard MTCNN [41] is used to detect 5 face landmarks for all the images. After performing similarity transformation accordingly, we obtain the aligned face images and resize them to be 112×96 pixels.

Occlusion Detection. We train an FCN-8s [14] segmentation network to detect the occlusion location. The training data includes the synthetic occluded CASIA-WebFace dataset and images of 26 subjects (outside the test subjects) from the AR dataset. The vgg16 backbone is firstly trained with sufficient face images to provide a good initialization. Finally, our occlusion detection model works pretty well with a mean IU of 98.51 on our synthetic occluded Facescrub dataset [19]. Figure 6 shows some detection results.

Network Structure. We employ the refined ResNet50 model proposed in recently published ArcFace [2] as our trunk CNN model. The mask generator is simply implemented as a CONV-PReLU-BN structure with a sigmoid function to map the output into $[0, 1]$.

Training. The training procedure includes three stages. *Stage* 1: Train the trunk CNN on the CASIA-WebFace [40] dataset with the large margin cosine loss [31]. *Stage* 2: Fixing the model parameters of the trunk CNN, and train the mask generator modules with specifically designed face pairs as shown in Figure 4. We discovered that occlusions on the peripheral blocks of the faces barely affect the recognition accuracy(less than 0.1% drop), therefore we narrow down the number of needed mask generators from 25 to 9, which correspond to the central 3×3 blocks that cover the main facial components. *Stage* 3: After establishing our mask dictionary, we generate face samples with various random partial occlusions and calculate their corresponding FDMs using this dictionary. Then finetune the trunk CNN using these (face, mask) pairs with a small learning rate. This stage is designed for relieving the inconsistency between the real-value mask output by the mask generator and the final binarized version, so a few epochs are enough.

Testing. In the testing stage, the similarity score is computed by the cosine distance of the fc features of two faces. The nearest neighbor classifier and thresholding are used for face identification and verification respectively. Considering the fact that, when recognizing an occluded face, we have lost the information from the occluded part of this face. Therefore it is necessary to also exclude this portion from the other faces comparing with it, to ensure that the similarity scores are computed based on equivalent information.

Baseline Models. Two baseline models are considered. The first one is the state-of-the-art face recognition model trained on CASIA-WebFace dataset. We will refer to it as *Trunk CNN*. The second one has the same configuration with the first one but finetuned with synthetic occluded CASIA-WebFace dataset (average occluder area is 25% of the face images), which will be referred to as *Baseline*.

4.2. Ablation Study

The Effect of τ. We conduct exploratory experiment to investigate the effect of τ used in binarization. By varying τ from 0 to 0.45, we evaluate our method on the AR dataset.

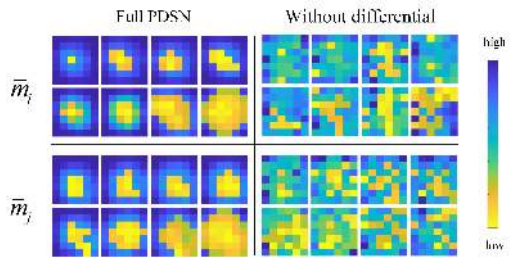

Figure 7. Illustration of the mean masks learned by our full PDSN and only by classification loss. $\bar{m}_i$ corresponds to occlusion on the left eye block and $\bar{m}_j$ corresponds to occlusion on the nose block.

τ	0	0.05	0.15	0.25	0.35	0.45
Acc.	95.84	97.29	97.36	**98.26**	97.98	97.92

Table 1. Rank-1 identification accuracy(%) comparison of different τ on AR dataset with sunglasses and scarf occlusions.

The probe set contains faces with sunglasses and scarf occlusions, and the gallery set contains 1 clean face for every subject. The rank-1 identification accuracy is given in Table 1. As τ being increased, the accuracy first rises up and then moves down as τ approaching 0.45. The best accuracy is achieved at $\tau = 0.25$ and the performance is not highly sensitive to this threshold.

Mask Type. To further explore the importance of binarization, we performed additional experiments with results shown in Table 2. First, by comparing the "Binary" and "Soft weight", we see that "Soft weight" noticeably decreases performance. We speculate that it's due to the excessive participation of features with very low mask values. Then we performed another experiment "Soft+Binary" to remove those features with mask values lower than the threshold (setting these mask values to 0) while keeping mask values above threshold unchanged (rather than setting them as 1). This version achieved comparable performance to the Binary version. Obviously, the importance of binarization is to completely eliminate the noise by setting a feature element with a very low mask value to zero. At the same time, the binary mask is highly efficient in terms of both calculation and storage.

The Differential Supervision. To investigate the importance of the differential input and pairwise loss. We set λ in the loss function in Eq. (5) to zero, and learn mask generators only from occluded face features. The mask dictionary is established with the same data and threshold τ. The performance comparison is shown in Table 3. The model trained with the pairwise supervision consistently outperforms the model that only trained with classification loss. In Figure 7 we visualize several channels of the mean masks of the left eye and nose blocks respectively under these two conditions. With our full PDSN, the mask elements with much lower weights (highlighted part in Figure 7) could reflect the occlusion location in image space to some extent, which is reasonable since the top conv layer still preserves spatial information. While the mean masks generated by classification loss only are in chaos. As discussed above, the differential input and contrastive loss help the model concentrate on the feature elements that have been altered a lot by partial occlusions, while the classification loss alone is likely to also diminish feature elements that are affected by some other factors unrelated to occlusion.

Mask type	Binary	Soft weight	Soft+Binary
Sunglasses	98.19	96.67	98.19
Scarf	98.33	97.22	99.03

Table 2. The rank-1 identification accuracy (%) in AR dataset Protocol 2. The results in the Protocol 1 have a similar conclusion.

Differential	AR sunglass	AR scarf	MF1 occ
No	95.97	97.92	54.80
Yes	98.19	98.33	56.34

Table 3. Rank-1 identification accuracy(%) of our method with and without differential supervision information. "MF1occ" refers to the occluded Facescrub probe set we synthesized.

4.3. Performance on LFW Benchmark

LFW [7] is a standard face verification benchmark dataset under unconstrained conditions. We evaluate our models strictly following the standard protocol of unrestricted with labeled outside data and report the mean accuracy on the 6,000 testing image pairs.

As shown in Table 4, the baseline model actually decreases the accuracy of the original trunk CNN by 0.52% when it is trained to gain more robustness to partial occlusions because most of the face images in the LFW dataset are not occluded. This phenomenon is consistent with [21], where performance encountered a drop when they tested a model that functions well for occluded objects on non-occluded objects. While our method can keep the performance of the trunk CNN since our design principle is just discarding those corrupted feature elements from comparison under partial occlusion condition, instead of forcing the trunk CNN to specifically accustom to partial occlusions.

4.4. Performance on MegaFace Challenge1

MegaFace Challenge [8] is a testing benchmark to evaluate the performance of face recognition algorithms at the million scale distractors. It contains a gallery set with more than 1 million face images. And the probe set consists of two datasets: Facescrub [19] and FGNet. In this study, we use the Facescrub dataset as our probe set. The training set is viewed as small if it is less than 0.5M. We evaluate the

Method	Training Data	#Models	Acc.
FaceNet [24]	200M	1	99.63
DeepID2+ [26]	2.6M	3	98.95
CenterFace [33]	0.7M	1	99.28
Baidu [11]	1.3M	1	99.13
SphereFace [13]	0.49M	1	99.42
CosFace [31]	5M	1	99.73
ArcFace [2]	0.49M	1	99.53
Trunk CNN	0.49M	1	99.20
Baseline	0.49M	1	98.68
Ours.	0.49M	1	**99.20**

Table 4. Face verification(%) on the LFW benchmark. "#Models" is the number of models used in the method for evaluation.

Methods	Protocol	MF1	MF1occ
SIAT_MMLAB	small	65.23	-
CenterFace [33]	small	65.49	-
DeepSense	small	70.98	-
SphereFace [13]	small	72.73	-
CosFace [31]	small	77.11	-
ArcFace [2]	small	77.50	-
FUDAN-CS_SDS	small	77.98	-
Trunk CNN	small	74.40	51.86
Baseline	small	68.81	53.03
Ours.	small	**74.40**	**56.34**

Table 5. Face identification accuracy(%) on MegaFace Challenge 1. "MF1occ" refers to the occluded Facescrub probe set.

basic trunk CNN, baseline model and our method under the small training set protocol on Challenge 1. The results are given in the "MF1" column of Table 5.

In order to test our method under partial occlusions, we synthesize the occluded Facescrub dataset. The occluding objects include sunglasses, mask, hand, eye mask, scarf, book, phone, cup, hat, fruit, microphone, hair, *etc*., all of which are common objects in real-life that may appear on the face, and each type of occluding objects has several different images that are distinct from those used in training phase. The left four images in Figure 6 show some examples. The results on this synthesized occluded Facescrub dataset are given in the "MF1occ" column of Table 5. Not surprisingly, a similar performance dropping on the original Facescrub probe set is observed for the baseline model. Compared to the baseline model, our method is superior on the occluded probe set without compromising the performance on the original probe set.

4.5. Performance on AR Dataset

We further evaluate our method through face identification experiments on the AR face database [15] with real-life occlusions. The AR database contains 4,000 face images with different facial expressions, illumination conditions and occlusions from 126 subjects. There are mainly two kinds of testing protocols in the existing literature. *Protocol* 1 refers to use more than 1 image per subject to form the gallery set (or training set). *Protocol* 2 refers to use only 1 image per subject to form the gallery set. Images of sunglasses and scarf occlusions are used for testing. We evaluate our method under both protocols and the results are given in Table 6. It is worth noting that the mask dictionary and the model are not finetuned with any AR face data at all, while other algorithms usually train with this dataset.

Table 6 shows that our method can significantly improve the performance of the trunk CNN model on faces with real-life sunglasses and scarf occlusions. The superior performance of our method than the baseline model indicates that simply shrink the range affected by occlusion is definitely not enough, it is essential to eliminate the corrupted portion from the comparison because it brings information inconsistency. And our mask dictionary captures the intrinsic feature structure of the trunk CNN model, which generalizes well to other face samples.

Methods	Protocol	Sunglass	Scarf
SRC[36]	1	87.00	59.50
NMR[37]	1	96.90	73.50
MLERPM[34]	1	98.00	97.00
SCF-PKR[38]	1	95.65	98.00
RPSM[35]	1	96.00	97.66
MaskNet [29]	1	90.90	96.70
Trunk CNN	1	98.19	99.72
Baseline	1	99.58	99.86
Ours.	1	**99.72**	**100.0**
RPSM[35]	2	84.84	90.16
Stringface[1]	2	82.00	92.00
LMA[16]	2	96.30	93.70
Trunk CNN	2	95.14	96.53
Baseline	2	96.67	96.39
Ours.	2	**98.19**	**98.33**

Table 6. Rank-1 face identification accuracy(%) on the AR dataset with natural occlusions.

5. Conclusions

In this paper, we propose an occlusion robust face recognition method based on the pairwise differential siamese network (PDSN) that explicitly builds correspondence between occluded facial blocks and corrupted feature elements. Competitive results on synthesized and realistic occluded face datasets demonstrate the superiority of the proposed method, especially the great generalization ability on general face recognition tasks.

References

[1] Weiping Chen and Yongsheng Gao. Recognizing partially occluded faces from a single sample per class using string-based matching. In *European Conference on Computer Vision*, pages 496–509, 2010.

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

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

[4] Ran He, Wei-Shi Zheng, and Bao-Gang Hu. Maximum correntropy criterion for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1561–1576, 2010.

[5] Ran He, Wei-Shi Zheng, Bao-Gang Hu, and Xiang-Wei Kong. A regularized correntropy framework for robust pattern recognition. *Neural computation*, 23(8):2074–2100, 2011.

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[7] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2008.

[8] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.

[9] Xiao-Xin Li, Dao-Qing Dai, Xiao-Fei Zhang, and Chuan-Xian Ren. Structured sparse error coding for face recognition with occlusion. *IEEE transactions on image processing*, 22(5):1889–1900, 2013.

[10] Zhifeng Li, Wei Liu, Dahua Lin, and Xiaoou Tang. Nonparametric subspace analysis for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–966, 2005.

[11] Jingtuo Liu, Yafeng Deng, Tao Bai, Zhengping Wei, and Chang Huang. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310*, 2015.

[12] Wei Liu, Zhifeng Li, and Xiaoou Tang. Spatio-temporal embedding for statistical face recognition from video. In *European Conference on Computer Vision*, pages 374–388, 2006.

[13] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 212–220, 2017.

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[15] Aleix M Martinez. The ar face database. *CVC Technical Report24*, 1998.

[16] Niall McLaughlin, Ji Ming, and Danny Crookes. Largest matching areas for illumination and occlusion robust face recognition. *IEEE transactions on cybernetics*, 47(3):796–808, 2016.

[17] Mostafa Mehdipour Ghazi and Hazim Kemal Ekenel. A comprehensive analysis of deep learning based representation for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–41, 2016.

[18] Rui Min, Abdenour Hadid, and Jean-Luc Dugelay. Improving the recognition of faces occluded by facial accessories. In *Face and Gesture 2011*, pages 442–447, 2011.

[19] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *IEEE International Conference on Image Processing (ICIP)*, pages 343–347, 2014.

[20] Hyun Jun Oh, Kyoung Mu Lee, and Sang Uk Lee. Occlusion invariant face recognition using selective local non-negative matrix factorization basis images. *Image and Vision Computing*, 26(11):1515 – 1523, 2008.

[21] Elad Osherov and Michael Lindenbaum. Increasing cnn robustness to occlusions by reducing filter support. In *IEEE International Conference on Computer Vision (ICCV)*, pages 550–561, 2017.

[22] Sohee Park, Hansung Lee, Jang Hee Yoo, Geonwoo Kim, and Soonja Kim. Partially occluded facial image retrieval based on a similarity measurement. *Mathematical Problems in Engineering*, 2015(1):1–11, 2015.

[23] Shunsuke Saito, Tianye Li, and Hao Li. Real-time facial segmentation and performance capture from rgb input. In *European Conference on Computer Vision*, pages 244–261, 2016.

[24] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[26] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2892–2900, 2015.

[27] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[28] Daniel Sáez Trigueros, Li Meng, and Margaret Hartnett. Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. *Image and Vision Computing*, 79:99–108, 2018.

[29] Weitao Wan and Jiansheng Chen. Occlusion robust face recognition based on mask learning. In *IEEE International Conference on Image Processing (ICIP)*, pages 3795–3799, 2017.

[30] Hao Wang, Dihong Gong, Zhifeng Li, and Wei Liu. Decorrelated adversarial learning for age-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3527–3536, 2019.

[31] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

[32] Yitong Wang, Dihong Gong, Zheng Zhou, Xing Ji, Hao Wang, Zhifeng Li, Wei Liu, and Tong Zhang. Orthogonal deep features decomposition for age-invariant face recognition. In *European Conference on Computer Vision*, pages 738–753, 2018.

[33] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515, 2016.

[34] Renliang Weng, Jiwen Lu, Junlin Hu, Gao Yang, and Yap-Peng Tan. Robust feature set matching for partial face recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 601–608, 2013.

[35] Renliang Weng, Jiwen Lu, and Yap-Peng Tan. Robust point set matching for partial face recognition. *IEEE Transactions on Image Processing*, 25(3):1163–1176, 2016.

[36] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[37] Jian Yang, Lei Luo, Jianjun Qian, Ying Tai, Fanlong Zhang, and Yong Xu. Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):156–171, 2017.

[38] Meng Yang, Lei Zhang, Simon Chi-Keung Shiu, and David Zhang. Robust kernel representation with statistical local features for face recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 24(6):900–912, 2013.

[39] Meng Yang, Lei Zhang, Jian Yang, and David Zhang. Robust sparse coding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 625–632, 2011.

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

[41] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[42] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5409–5418, 2017.

[43] Fang Zhao, Jiashi Feng, Jian Zhao, Wenhan Yang, and Shuicheng Yan. Robust lstm-autoencoders for face de-occlusion in the wild. *IEEE Transactions on Image Processing*, 27(2):778–790, 2018.

[44] Zihan Zhou, Andrew Wagner, Hossein Mobahi, John Wright, and Yi Ma. Face recognition with contiguous occlusion using markov random fields. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1050–1057, 2009.