Biased Feature Learning for Occlusion Invariant Face Recognition

Changbin Shao[1,2], **Jing Huo**[1*], **Lei Qi**[1], **Zhen-Hua Feng**[3]
Wenbin Li[1], **Chuanqi Dong**[1] and **Yang Gao**[1]
[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2] School of Computer, Jiangsu University of Science and Technology, Zhenjiang, China
[3] Department of Computer Science, and the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK
shaochangbin@163.com, huojing@nju.edu.cn, qilei.cs@gmail.com, z.feng@surrey.ac.uk
liwenbin@nju.edu.cn, dongchuanqi@smail.nju.edu.cn, gaoy@nju.edu.cn

Abstract

To address the challenges posed by unknown occlusions, we propose a Biased Feature Learning (BFL) framework for occlusion-invariant face recognition. We first construct an extended dataset using a multi-scale data augmentation method. For model training, we modify the label loss to adjust the impact of normal and occluded samples. Further, we propose a biased guidance strategy to manipulate the optimization of a network so that the feature embedding space is dominated by non-occluded faces. BFL not only enhances the robustness of a network to unknown occlusions but also maintains or even improves its performance for normal faces. Experimental results demonstrate its superiority as well as the generalization capability with different network architectures and loss functions.

1 Introduction

As an important authentication technique, Face Recognition (FR) has been widely used in many practical applications. A high-performance FR system relies on discriminative feature extraction that is robust to appearance variations, *e.g.*, in pose, expression, illumination, and occlusion. Occlusion, as an intractable covariate of face variations, is very challenging for the FR community. Occlusion-invariant FR aims to learn a model with good generalization capability such that it can be readily adapted to occluded faces, not only normal faces

Recently, Convolution Neural Networks (CNNs) have been proven to be able to extract robust features for unconstrained FR thus the performance of modern FR systems has been significantly improved [Taigman *et al.*, 2014; Schroff *et al.*, 2015]. However, the performance of a FR model often degrades in the presence of unknown occlusions or disguises [Singh *et al.*, 2019]. There are only few studies that focus on generalized feature learning for occlusion-invariant FR. To close this gap, we aim to improve the generalization capability of a model for unknown occlusions.

For closed set protocols, several linear methods have been proposed to mitigate the difficulties posed by occlusions. These methods can be divided into representation-based and image completion methods. Driven by the hypothesis that the data from the same source lies in the same subspace, Sparse-Representation-based Classification (SRC) [Wright *et al.*, 2009b] performs regression-based identification with sparse constrains. Image completion methods attempt to recover clean data using low-rank and sparse constrains [Wright *et al.*, 2009a]. However, due to the limitation of linear operations, most traditional methods only perform well under constrained scenarios. Recently, with the development of Generative Adversarial Networks (GAN) [Goodfellow *et al.*, 2014], deep face completion methods have demonstrated promising results for realistic content generation of occluded facial parts. But this roundabout way is also limited to closed set protocols due to the difficulties in identity preservation.

Recently, several studies focusing open set occlusion-invariant FR have been proposed. [Saeztrigueros *et al.*, 2018] attempts to boost occlusion-invariant FR using auxiliary occlusion samples. [Song *et al.*, 2019] focuses on mask designs for final measure, which applies masks to deep CNN feature maps so that only non-occluded features are highlighted. However, it may only work well for frontal faces.

Overall, there are two main challenges for occlusion-invariant FR under open set scenarios. The first one is the lack of data. For a data-driven model, a big training data is crucial. However, to the best of our knowledge, existing datasets only contain few occluded faces. Moreover, there is no standard benchmark or evaluation protocol for model test. Almost all the existing studies perform model evaluation on their own synthetic datasets. The second challenge is model training. For network training, the occlusion data could improve the feature representation ability of a model to occluded faces, but may bring negative effects to the conventional feature distribution of normal faces. Therefore, it is important to design a good learning scheme that can avoid the negative effects of occlusion sources for feature embedding.

As we all know, due to the powerful fitting ability of deep model and its non-convex characteristic, there may be many convergence points of CNN parameters to produce discriminative features. Regardless of network architectures, we assume that there are some specific parameters sets adapted to normal and occlusion samples. To enhance the generalization capability of a model against occlusions, it is necessary

*Corresponding Author

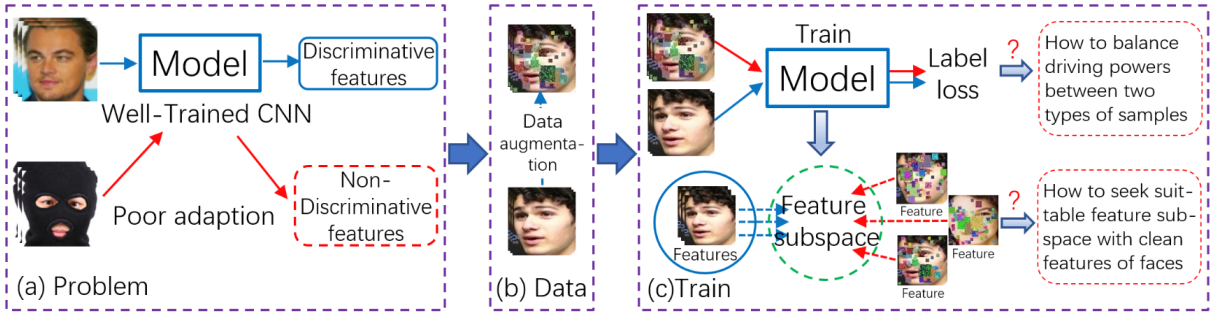

Figure 1: The proposed BFL framework. (a) A well trained CNN model may poorly adapt to occluded faces. (b) We synthesize occluded faces by adding random spatially continuous noises to clean faces. (c) The proposed BFL method focuses on both clean and occluded faces. Moreover, BFL pays attention to the feature learning process so that the learned CNN features are only associated with clean faces.

to seek a better model learning strategy.

In this paper, we propose a novel Biased Feature Learning (BFL) framework for occlusion-invariant FR, as shown in Fig. 1. To boost the generalization ability of a FR model for unknown occlusions, we first use an enhanced data augmentation scheme to randomly generate multi-scale spatially continuous noises for training samples. Second, we modify the label loss to balance the impact of normal and occluded samples for network training. Further, we manipulate the optimization process of the network so that it can produce suitable features only associated with clean face features. Last, to highlight unknown face occlusions, we use realistic occlusion samples to perform model evaluation rather than simply adding random patches to faces.

The main contributions of this paper include: 1) A biased guidance strategy for feature learning. It can skillfully prompt embedding learning to focus on clean facial features without using detection technologies. Besides, It can be easily extended to other tasks that need to deal with polluted or adversarial samples. 2) An enhanced data augmentation method for model training. It can effectively simulate disordered distributions of real occlusions, and can be simply embedded into a deep framework to generate interferences to normal faces. 3) An objective benchmark for model evaluation. We refine the traditional protocol into three indicators, as shown in Fig. 5. Meanwhile, we collect realistic occlusions and apply them to LFW to create a reasonable benchmarking dataset.

2 Related Work

In this section, we introduce the related work by dividing them into three categories.

Linear regression. Assume that occlusion error is sparse relative to the standard (pixel) basis, SRC uses the $\mathcal{L}_1$ regularization to code a query sample as a linear combination of atoms and assigns the label to the class with the minimum reconstruction error. To enhance the discrimination of coding, structured sparse coding [Li *et al.*, 2013] and non-negative dictionary learning [Ou *et al.*, 2018] were proposed. To address the small-sample-size problem, extended dictionaries [Deng *et al.*, 2012; Shao *et al.*, 2017] with intra-class face variations posed by occlusions were developed. Due to the low-rank characteristic of occlusion in comparison to face size, [Iliadis *et al.*, 2017; Wu and Ding, 2018] appended low-rank constrains to occlusion error. To characterize the 2D structural information of occlusions, [Yang *et al.*, 2017] used the nuclear norm to deal with occlusion and illumination variations. However, all these linear methods are limited to frontal faces under closed set scenarios.

Face completion. Given the sparsity of noise and low rank of clean data, robust PCA [Candès *et al.*, 2011] can be used to recover corrupted low-rank matrix by minimizing its weighted $\mathcal{L}_1$ and nuclear norms. Low rank representation [Liu *et al.*, 2013] extended the recovery of clean data from single subspace to a union of multiple subspaces. These methods provide effective face completion techniques. For example, [Zhang *et al.*, 2015] presented the double nuclear norm based matrix decomposition for occluded face recovery. In recent years, deep generation has been widely used for face completion. [Li *et al.*, 2017] generated semantic contents for missing values by a combination of reconstruction, adversarial and semantic parsing losses. [Zhao *et al.*, 2017] used multi-scale spatial LSTM to perform face completion. Recently, [Yuan and Park, 2019] used a 3D morphable model and GAN to perform face de-occlusion. In contrast to linear completion methods, deep generation methods can overcome unconstrained face variations. But, the generated contents can not able to preserve the original identity thus they are rarely used for open set FR.

Deep feature learning. To perform occlusion-invariant FR, [Saeztrigueros *et al.*, 2018] detected the sensitivity of a model to different occlusion regions and forced a model to focus on the whole face region equally via center-focused occlusion samples. For a pre-trained teacher model, [Song *et al.*, 2019] used the differences of feature maps between frontal faces and their occluded versions to learn generators that produced masks for face. Then a mask dictionary was established to produce mask scores for query samples. Although the use of occlusion samples is simple and the process of mask method is complicated, they are suitable for FR in open set scenarios. However, these methods are usually limited to frontal faces and simple occlusions, and the synthetic datasets of these exiting methods are not public for future studies. In this paper, to achieve occlusion-invariant FR under unconstrained scenarios, we focus on generalized feature learning regardless of

Figure 2: Some examples of synthetic faces with occlusions. We set δ=20 and use three different occlusion units, 40×40, 20×20 and 10×10. The numbers of different types are 4, 8 and 32. The overlap of occlusions leads to the occlusion area less than 60% of an image.

network architectures and loss functions, and present a benchmark dataset for open set scenarios.

3 The Biased Feature Learning Framework

This section first discusses the sensitivity of a FR model to occlusions. Then we introduce an enhanced data augmentation method for model training. Last, we present the biased feature learning strategy and give a brief discussion.

3.1 Sensitivity of FR Models to Occlusions

CNNs have been widely used in multiple computer vision tasks. In FR, most models rely on global features via a discriminative loss design [Wen *et al.*, 2016; Wang *et al.*, 2018; Deng *et al.*, 2019]. In this case, a pre-trained model is often sensitive to face occlusions. A pre-trained model on clean data often ineluctably captures large area of occluded pixels, which brings uncertainty to decision making. In fact, occlusions may lead to larger intra-class variations and higher inter-class similarities. For example, sunglasses may be viewed as face features for different persons.

For the generalization capability of a model, we argue that occluded faces are essentially viewed as polluted data or outliers. The disordered data distributions produced by unknown pollution sources lead to poor adaptation of a high-performance FR model. Fig. 3 shows the disordered feature deviations between occluded faces and their original ones.

3.2 Training Data Augmentation

As spatially continuous or extreme noises, occlusions can be presented in arbitrary ways. In practical scenarios, besides some simple occlusions posed by sunglasses and scarfs, an unknown occlusion can be presented in any shape, texture and size, even appears as a hand, leaf, stocking and stains of a camera. However, for model training, existing datasets usually only contain some simple occlusions. Further, as opposed to a large number of normal samples, the scarce occlusion samples are insufficient for model training.

In practice, it is difficult to collect occluded faces jointly with their clean versions. In addition, it is almost impossible to collect all possible occlusion types. Instead of constructing a comprehensive dataset, we propose a synthetic method to conveniently simulate occlusions for exiting labeled datasets. The principle is to consider randomness in size, shape and texture on the basis of spatial continuity.

Given a gray face image $\mathbf{x} \in R^{H \times H}$, in view of spatial continuity, we use a small image patch $\mathbf{u} \in R^{h \times h} (h \ll H)$ as a basic occlusion unit. For the texture of $\mathbf{u}$, we first generate a random mean value μ $(0 + \delta \leq \mu \leq 255 - \delta)$ for all the pixels, under a variance parameter δ. And each pixel $\mathbf{u}^i$ is set as $\mathbf{u}^i = \mu + \delta \times v$ $(v \sim N(0, 1))$, where v is a random value following the normal distribution. For color images, the gray unit $\mathbf{u}$ is extended to all the color channels.

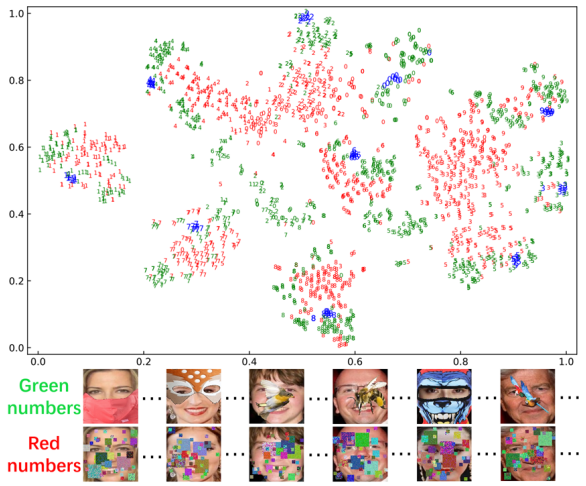

Figure 3: 2D visualization (t-SNE) of the features extracted by ResNet-Inception-V1. We visualize 10 subjects: blue for normal faces, red for synthetic occlusions and green for realistic occlusions.

Overall, for the occlusion unit $\mathbf{u}$, we can generate s different versions with different mean value μ, and randomly select s locations in the face $\mathbf{x}$ to form a unit occlusion set $\overline{u}^s = \{\mathbf{u}^{r,c} | 0 \leq r, c \leq H - h\}$, where r and c are the starting row and column coordinates in $\mathbf{x}$. As an integrated occlusion, the $\overline{u}^s$ can be embedded into $\mathbf{x}$ to obtain an occluded face $\overline{\mathbf{x}}$. Intuitively, the randomness or complexity of an occlusion can be approximated by multivariate cooperation of s, μ, δ and (r, c), hence the final synthetic occlusion is a combination of multiple occlusion units with random textures.

For a more convenient implementation, multiple occlusion units with different sizes h can be used to form a multi-scale occlusion set $\overline{U} = \overline{u}_{h_1}^{s_1} \cup ... \cup \overline{u}_{h_n}^{s_n}$. Fig. 2 shows some examples synthesized by the proposed method.

Data augmentation is an important approach to boost the performance of a model. Existing augmentation methods usually appeal to various presentations of features, such as flipping, rotation, local warping and cropping. In contrast to these methods, the main motivation of the proposed method is to simulate the disturbance of occlusions to face features. As shown in Fig. 3, compared to realistic occluded faces (green), the distributions of synthetic occluded faces (red) reflect similar disordered deviations from their original ones (blue).

In fact, this approach can be considered as an enhanced version of additional continuous noises (such as simple square noises). One advantage of the proposed method is that it is designed for occlusions in open set. There is no overlap between synthetic and realistic occlusions. So it is suitable to check the generalization ability of a model. Second, the synthetic occlusions can simulate similar or more extreme pollution for face features. And there are identity labels and type labels (clean v.s. occluded) for both synthetic and existing samples. This is convenient to manipulate model training.

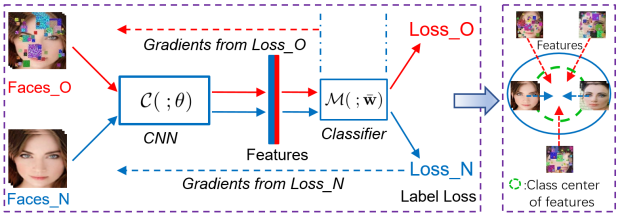

Figure 4: A brief schematic diagram of biased guidance strategy and its motivation. Motivation (right): As a goal of feature learning, class center is only dominated by face features, and used as a guidance of occluded samples. Method (left): The parameters of a model are selectively updated according to dash lines.

3.3 Biased Guidance Strategy for Model Training

For a data-driven model relying on a single feature extraction channel, when multiple types of samples are used to train it to produce discriminative feature embedding space, it is difficult to optimize the model to adapt to all the sample types. Due to the strong fitting ability of a deep model, the spatial region from occlusion may lead a raw model to learn non-facial features in some dimensions for classification, or push the model to ignore the same region of normal samples. These hypotheses imply that features of face and occlusion sources jointly dominate the distribution of class centers. Here, occluded data inevitably interferes with the feature extraction of normal data. Therefore, it is necessary to design an effective learning strategy for model optimization. To this end, we present a biased guidance strategy for model training.

Given an extended training dataset $I = I_N \cup I_O$, where I_N and I_O are separately the normal face set and its occluded version. For one sample $\mathbf{x}_i \in I$ with label $y_i \in Y$, after the forward pass of a CNN model, the feature is denoted as $\mathbf{f}_i = \mathcal{C}(\mathbf{x}_i, \theta)$, where $\mathcal{C}$ is the feature extraction function defined by a CNN and θ is the model parameters.

For the conventional model training with $\mathbf{x}_i \in I$, to enhance discrimination of $\mathbf{f}_i$, an efficient classifier usually uses a label mapping layer $\mathcal{M}(\mathbf{f}_i, \mathbf{w}, \mathbf{b})$ to obtain posterior probability $p(y_i|\mathbf{f}_i)$, and estimates the mapping matrix $\mathbf{w}$ and bias vector $\mathbf{b}$ via its maximum likelihood as:

$$(\mathbf{w}, \mathbf{b})^* = \arg\max_{\mathbf{w},\mathbf{b}} \prod_i p(y_i|\mathbf{f}_i; \mathbf{w}, \mathbf{b}) \tag{1}$$

To implement it conveniently, we often design a label loss to minimize its negative log-likelihood:

$$\mathcal{L} = \sum_i -\log(\mathcal{M}(\mathbf{f}_i, \bar{\mathbf{w}}); y_i) \tag{2}$$

where $\bar{\mathbf{w}} = (\mathbf{w}, \mathbf{b})$. Here, the key is that $\bar{\mathbf{w}}$ controls the data distribution of feature $\mathbf{f}_i$. Take the commonly used softmax loss for an example, the predicted label $\hat{y}_i$ from $\mathcal{M}$ can be denoted as:

$$\hat{y}_i = \mathcal{M}(\mathbf{f}_i, \bar{\mathbf{w}}) = \text{softmax}(\bar{\mathbf{w}}^T \mathbf{f}_i) \tag{3}$$

For one-hot coding, matrix $\bar{\mathbf{w}}$ maps feature $\mathbf{f}_i$ into one softmax space, in which the label vector l_i is represented by $l_i^{k(k=y_i)} = 1$ and $l_i^{k(k\neq y_i)} = 0$. To ignore constant operation of softmax, we can observe that the hyper-plane $\bar{\mathbf{w}}_{y_i}^T$ dominates the feature center of the y_ith class.

Now, for the two types of samples $\mathbf{x}_n \in I_N$ and $\mathbf{x}_o \in I_O$, instead of the rough training with Eq. (2), there are two main aspects should be considered: (i) how to balance the driving power between I_N and I_O for the learning of model parameters θ; and (ii) how to manipulate the feature subspace or class centers with face features rather than hybrid features containing occlusions. To balance these two drivers, we remodify the label loss $\mathcal{L}$ as:

$$\mathcal{L} = \mathcal{L}_{\mathbf{x}_n \in I_N} + \lambda \frac{n}{o} \mathcal{L}_{\mathbf{x}_o \in I_O} \tag{4}$$

where λ is empirically associated with the non-occlusion percentage in $\mathbf{x}_o$. n and o are the numbers of normal and occlusion samples, which aim at the cost sensitivities of categories.

For the data distribution of each class center, given non-face pixels in $\mathbf{x}_o$, its feature $\mathbf{f}_o$ may contain non-face information extracted by the raw $\mathcal{C}$, so the posterior probability $p(y_o|\mathbf{f}_o)$ should no longer participate in the learning of $\bar{\mathbf{w}}$. To dominate the class center only with non-occluded face features, we can make one modification for parameter update. The gradients of all parameters $\{\theta, \bar{\mathbf{w}}\}$ are reformulated as:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}_{\mathbf{x}_i \in I}}{\partial \theta}, \quad \frac{\partial \mathcal{L}}{\partial \bar{\mathbf{w}}} = (1 + \lambda \frac{n}{o}) \frac{\partial \mathcal{L}_{\mathbf{x}_n \in I_N}}{\partial \bar{\mathbf{w}}} \tag{5}$$

Clearly, the parameters $\bar{\mathbf{w}}$ with respect to class centers will be only updated by the branch loss of normal samples. This biased class center drags or guides the parameter learning of θ in $\mathcal{C}$ via overall driving of Eq. (4). By this biased guidance strategy, the model can produce discriminative features with a minimum deviation from conventional face features.

Fig. 4 briefly demonstrates the proposed BFL method. Particularly, this approach is simple to implement, in which we decompose a loss for two types of samples with Eq. (4). For feature learning, we only use the loss branch of normal samples to update parameters of classifier and use the whole loss to optimize CNN. Overall, this approach is suitable for a variety of loss functions attached to a label mapping layer.

3.4 A Brief Discussion on BFL

Essentially, occlusion is a data pollution problem caused by unknown spatially continuous or extreme noises. As Fig.3 implied, random occlusions drag samples away from their original data distribution, and produce disordered outliers. It is different from the cross-domain problem, where we can seek or construct one or multiple mappings to obtain one common subspace for different domains. Here, disordered outliers are not lied in one or several domains. Therefore, for classification, the best way is to build a many-to-one mapping from outliers to clean samples so that outliers converge to their true class centers at the feature level. This is also the main motivation of the proposed BFL method. The right sub-figure in Fig. 4 briefly demonstrates this concept.

4 Evaluation Dataset and Protocol

To evaluate the performance of a model, a public dataset with reasonable evaluation protocol is necessary. For example, the well-known LFW dataset has been a widely used benchmark for normal face verification. However, there is no a public face dataset specially designed for occlusion-invariant FR. In

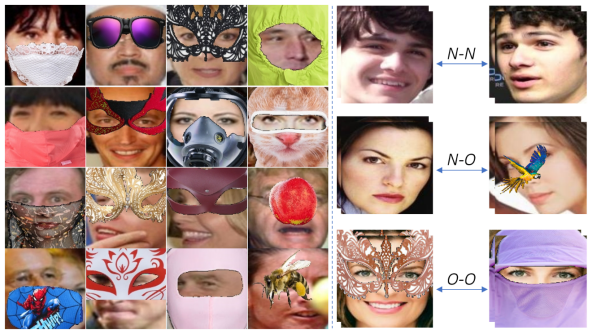

Figure 5: Left: Some typical synthetic examples for frontal, lateral and random occluded face in O-LFW. Right: Specification of face verification in occlusion scene. We refine the evaluation as three types: normal face pairs (N-N), pairs between normal and occluded faces (N-O) and occluded face pairs (O-O).

this section, we modify LFW and extend its evaluation protocol for occlusion-invariant FR.

4.1 Extended Evaluation Protocol

For face verification with occlusions, there are two types of faces on a dataset, normal and occluded faces. A general verification protocol is not sufficient for model evaluation. Therefore, it is necessary to redefine the evaluation as three different types: normal face pairs verification (N-N), verification between normal and occluded faces (N-O) and occluded face pairs (O-O), as shown in Fig. 5.

4.2 The Occluded LFW Dataset (O-LFW)

There are many existing references for LFW in normal face verification and the standard protocol is suitable for open set evaluation. To facilitate the research in occlusion-invariant FR, we collect many realistic occlusion sources and apply them to LFW for a new benchmark (O-LFW). For occlusion sources set, there are 200 types of occlusions in total, including 90 upper half occlusions, 70 lower half occlusions, 30 random occlusions and 10 large area occlusions.

For O-LFW, the original $6K$ face pairs arranged in left and right of LFW are directly used for N-N verification. For the verification of N-O pairs, all faces on the right are replaced with their own synthetic versions. For O-O verification, we first synthesize $6K$ occluded versions for the faces on the left side. To avoid abundant overlaps of the same occlusions, all the right occluded faces are synthesized with random occlusion sources. Some examples are shown in Fig. 5.

Therefore, there are three settings in O-LFW, each with $6K$ pairs, for evaluation. For each setting, all the faces are in the order of the standard pairs list of LFW. It will be released for further studies in occlusion-invariant FR.

5 Experiments

Dataset. We use CASIA-WebFace (10575 classes with 0.49M samples) as the training set of I_N, and synthesize the same number of virtual samples as its occluded version I_O (we simply set $n = o$). The parameters for synthetic occlusions are the same as that in Fig. 2. For testing, we use the O-LFW dataset with the cosine similarity measure.

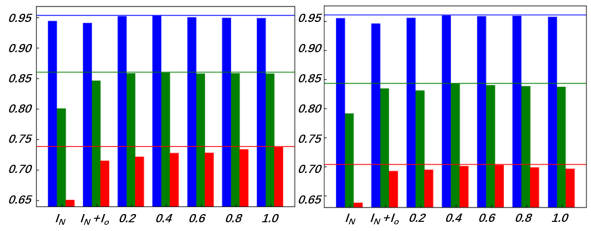

Figure 6: Performance exhibitions of different schemes on L09 (left) and R18 (right). Subscripts I_N and I_N+I_O separately denote conventional training with only normal samples and hybrid samples; Latter 5 decimals denote the λ values with the proposed biased guidance strategy. The blue, green and red bars respectively reflect the accuracy of N-N, N-O and O-O.

Model. We select two typical networks with advanced residual structure for comparisons: LightCNN-9 (L09) and Resnet-18 (R18). L09 is with the Maxout activation function and without Batch Normalization (BN) layers [Wu *et al.*, 2018]. R18 contains conventional Relu activation and BN [He *et al.*, 2016]. For R18, we replace the first layer with 3×3 convolution. Before training, all the images are resized to 128×128 for L09 and 112×112 for R18. The output feature vector is uniformly set to 256 for both models.

Network training. To verify the superiority of the proposed method fairly, all the models are trained with the Adam optimizer in PyTorch. The batch size is separately set as 128 for original I_N and 256 for hybrid I_N+I_O. For all the experiments, random horizontal flip is applied to the training images. The softmax loss is used and the learning rate (lr) is set to $5e-4$ in subsection 5.1 and 5.2.

5.1 Sensitiveness Analysis on λ

As stated in Section 3.3, the parameter λ in Eq. (4) dominates the driving power of I_O in the overall loss, which was empirically given one number associated with non-occlusion percentage in $\mathbf{x}_o$ ($0 \leq \lambda \leq 1$). So we conduct two experiments to investigate its sensitiveness. Since we set $n = o$, the loss function is simplified as $\mathcal{L} = \mathcal{L}_{\mathbf{x}_n \in I_N} + \lambda \mathcal{L}_{\mathbf{x}_o \in I_O}$.

In this part, we only conduct 20 training epochs for network training. Given the performance tends to saturation after 11 epochs, we report the average accuracy between $11th$ and $20th$ epochs to avoid the randomness of results. We display the performance trends in Fig. 6.

As shown in Fig. 6, we can find: (i) The classical training method achieves substantial improvements for N-O and O-O under the auxiliary occlusion set I_O, but it is accompanied by an obvious performance drop on N-N. In contrast, the proposed BFL method obtains remarkable improvements for all the three settings with different λ. (ii) For L09, as the increase of λ, the N-N performance is dropping along with the rising of O-O. For R18, N-N maintains stable improvements but with slight fluctuation of N-O and O-O.

Overall, we argue that the best λ should not be fixed for different I_O. It is best to match the percentage of non-occlusion face and separately set for different models. Besides, since the two models perform differently in terms of λ, its value is simply set as 0.5 for subsequent experiments.

Method	N-N	N-O	O-O	Dataset
Baseline (I_N)	94.38	79.90	64.91	I_N
Dropout(0.5)	94.44	80.42	65.35	I_N
Crop(0.6-1)	95.59	80.32	64.20	I_N
Crop(0.8-1)	95.81	81.96	65.28	I_N
Baseline (I_N+I_O)	**94.10**	**84.48**	**71.55**	I_N+I_O
BFL	**95.28**	**85.72**	**72.42**	I_N+I_O
BFL+Dropout(0.5)	**95.16**	**86.04**	**73.08**	I_N+I_O
BFL+Crop(0.6-1)	**95.91**	**87.66**	**74.77**	I_N+I_O
BFL+Crop(0.8-1)	**96.13**	**88.12**	**75.10**	I_N+I_O

Table 1: Verification results (%) of L09 on O-LFW datasets.

Method	N-N	N-O	O-O	Dataset
Baseline(I_N)	95.43	79.87	65.38	I_N
Dropout(0.5)	95.32	79.34	64.29	I_N
Crop(0.6-1)	96.96	80.32	64.58	I_N
Crop(0.8-1)	96.77	81.39	65.08	I_N
Baseline (I_N+I_O)	**95.08**	**83.11**	**68.70**	I_N+I_O
BFL	**95.55**	**83.41**	**70.43**	I_N+I_O
BFL+Dropout(0.5)	**96.12**	**84.01**	**70.05**	I_N+I_O
BFL+Crop(0.6-1)	**97.45**	**87.22**	**73.12**	I_N+I_O
BFL+Crop(0.8-1)	**97.35**	**85.98**	**70.78**	I_N+I_O

Table 2: Verification results (%) of R18 on O-LFW datasets.

5.2 Comparison with Other Methods

In this subsection, we compare totality and sub-components of BFL framework with some existing methods.This is also equivalent to the ablation study for Section 5.3. For dropout schemes, the ratio is set as 0.5 and applied before the classification layer. α-β in Crop(α-β) indicates the scale of random cropping for faces. The results are reported with the average value between the $16th$ and $20th$ epochs in Table 1 and Table 2, in which Baseline denotes the conventional training way and BFL is the proposed method with λ=0.5.

According to Table 1 and Table 2, we can conclude that:

1) Overall, other data augmentation schemes are effective for the traditional N-N setting but do not perform well for N-O and O-O. For the proposed I_O, it can significantly improve N-O and O-O but bring negative effects for N-N. For the proposed BFL method, it not only significantly improves the performance of all the cases but also obtains more effective improvements united with other methods.

2) For the two models, L09 is more effective than R18 to mitigate the occlusion issue. The main reason is that the occlusion sources in I_O are thoroughly different from the realistic occlusions in the O-LFW test set, which brings different covariate shifts for the intermediate CNN feature maps so that it is unfavorable for the application of BN in R18.

5.3 Study of Generalization Capability

In this section, we examine the generalization capability of the proposed method to different loss functions. For network training of all the evaluated methods, we apply random Crop (0.8-1.0) to the training images. Meanwhile, we use the $\mathcal{L}_2$ regularization with the weight decay of 0.01 following AdamW [Loshchilov and Hutter, 2019]. We set lr=$1e$-4 for R18 with the ArcFace loss to avoid non-convergence, and lr=$5e$-4 for all the others. After 100 epochs, average results of the last 5 epochs are reported.

Model	Method	N-N	N-O	O-O	Loss
L09	Baseline	95.85	81.29	65.63	Softmax
	BFL	**96.79**	**89.06**	**76.61**	
	Baseline	98.63	85.14	68.36	CosFace
	BFL	**98.52**	**90.81**	**76.98**	
	Baseline	98.52	85.14	68.36	ArcFace
	BFL	**98.68**	**91.36**	**77.54**	
R18	Baseline	97.06	79.29	63.29	Softmax
	BFL	**98.30**	**86.18**	**70.07**	
	Baseline	98.69	82.70	65.89	CosFace
	BFL	**98.80**	**88.14**	**72.47**	
	Baseline	98.17	75.12	61.95	ArcFace
	BFL	**98.67**	**83.91**	**70.27**	
Lv2	Baseline	99.57	89.83	72.77	CosFace
	BFL	**99.47**	**93.40**	**79.03**	
Rv1	Baseline	99.33	88.27	74.65	CosFace
	BFL	**99.15**	**91.83**	**79.97**	

Table 3: Verification results (%) of 4 models on different losses.

Besides, we provide comparisons on LightCNN-V2 (Lv2) and ResNet-Inception-v1 (Rv1) with the same setting. We separately report the results after 10 and 20 epochs of fine-tuning. All the results are reported in Table 3.

Overall, the improvements of BFL method on N-N, N-O and O-O are in the interval of (-0.18, 1.14), (3.56, 7.77) and (5.32, 10.98) in terms of accuracy. We can conclude that, for different models with different loss functions, the proposed BFL framework not only improves the generalization capability to occluded faces but also maintains the good performance for normal faces.

6 Conclusion

To address the challenges posed by unknown occlusions, we presented a reasonable model evaluation protocol and benchmarking dataset for occlusion-invariant face recognition. In addition, we proposed a novel biased feature learning framework for deep network training. The proposed BFL framework uses a biased guidance strategy to promote the feature learning of a face recognition model. The experimental results demonstrated the merits of our BFL method as well as its generalization capability with different network architectures and loss functions.

Acknowledgments

This work is supported by the Science and Technology Innovation 2030 – "New Generation Artificial Intelligence" Major Project (No. 2018AAA0100900), National Science Foundation of China (No. 61806092, No. 61902153) and Jiangsu Natural Science Foundation (No. BK20180326).

References

[Candès *et al.*, 2011] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.

[Deng *et al.*, 2012] Weihong Deng, Jiani Hu, and Jun Guo. Extended src: Undersampled face recognition via intraclass variant dictionary. *TPAMI*, 34(9):1864–1870, 2012.

[Deng *et al.*, 2019] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4685–4694, 2019.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pougetabadie, Mehdi Mirza, Bing Xu, David Wardefarley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Iliadis *et al.*, 2017] Michael Iliadis, Haohong Wang, Rafael Molina, and Aggelos K Katsaggelos. Robust and low-rank representation for fast face identification with occlusions. *TIP*, 26(5):2203–2218, 2017.

[Li *et al.*, 2013] Xiaoxin Li, Daoqing Dai, Xiaofei Zhang, and Chuanxian Ren. Structured sparse error coding for face recognition with occlusion. *TIP*, 22(5):1889–1900, 2013.

[Li *et al.*, 2017] Yijun Li, Sifei Liu, Jimei Yang, and Minghsuan Yang. Generative face completion. In *CVPR*, pages 5892–5900, 2017.

[Liu *et al.*, 2013] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *TPAMI*, 35(1):171–184, 2013.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[Ou *et al.*, 2018] Weihua Ou, Xiao Luan, Jianping Gou, Quan Zhou, Wenjun Xiao, Xiangguang Xiong, and Wu Zeng. Robust discriminative nonnegative dictionary learning for occluded face recognition. *Pattern Recognition Letters*, 107:41–49, 2018.

[Saeztrigueros *et al.*, 2018] Daniel Saeztrigueros, 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.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[Shao *et al.*, 2017] Changbin Shao, Xiaoning Song, Zhenhua Feng, Xiaojun Wu, and Yuhui Zheng. Dynamic dictionary optimization for sparse-representation-based face classification using local difference images. *Information Sciences*, 393:1–14, 2017.

[Singh *et al.*, 2019] Maneet Singh, Richa Singh, Mayank Vatsa, Nalini K Ratha, and Rama Chellappa. Recognizing disguised faces in the wild. *TBIOM*, 1(2):97–108, 2019.

[Song *et al.*, 2019] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In *ICCV*, pages 773–782, 2019.

[Taigman *et al.*, 2014] Yaniv Taigman, Ming Yang, Marcaurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014.

[Wang *et al.*, 2018] 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 *CVPR*, pages 5265–5274, 2018.

[Wen *et al.*, 2016] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515. Springer, 2016.

[Wright *et al.*, 2009a] John Wright, Arvind Ganesh, Shankar R Rao, Yigang Peng, and Yi Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *NIPS*, pages 2080–2088, 2009.

[Wright *et al.*, 2009b] John Wright, Allen Y Yang, Arvind Ganesh, Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *TPAMI*, 31(2):210–227, 2009.

[Wu and Ding, 2018] Choying Wu and Jianjiun Ding. Occluded face recognition using low-rank regression with generalized gradient direction. *Pattern Recognition*, 80:256–268, 2018.

[Wu *et al.*, 2018] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *TIFS*, 13(11):2884–2896, 2018.

[Yang *et al.*, 2017] 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. *TPAMI*, 39(1):156–171, 2017.

[Yuan and Park, 2019] Xiaowei Yuan and Inkyu Park. Face de-occlusion using 3d morphable model and generative adversarial network. In *ICCV*, pages 10062–10071, 2019.

[Zhang *et al.*, 2015] Fanlong Zhang, Jian Yang, Ying Tai, and Jinhui Tang. Double nuclear norm-based matrix decomposition for occluded image recovery and background modeling. *TIP*, 24(6):1956–1966, 2015.

[Zhao *et al.*, 2017] Fang Zhao, Jiashi Feng, Jian Zhao, Wenhan Yang, and Shuicheng Yan. Robust lstm-autoencoders for face de-occlusion in the wild. *TIP*, 27(2):778–790, 2017.