Biased Feature Learning for Occlusion Invariant Face Recognition

Changbin Shao\textsuperscript{1,2}, Jing Huo\textsuperscript{1*}, Lei Qi\textsuperscript{1}, Zhen-Hua Feng\textsuperscript{3}
Wenbin Li\textsuperscript{1}, Chuanqi Dong\textsuperscript{1} and Yang Gao\textsuperscript{1}

\textsuperscript{1} State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
\textsuperscript{2} School of Computer, Jiangsu University of Science and Technology, Zhenjiang, China
\textsuperscript{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, Convolutional 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...
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.

2 Related Work

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

Linear regression. Assume that occlusion error is sparse relative to the standard (pixel) basis, SRC uses the $L_1$ 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 $L_1$ and nuclear norms. Low rank representation [Liu et al., 2013] extended the recovery of clean data from single sparse 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 produceduvation 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 $L_1$ and nuclear norms. Low rank representation [Liu et al., 2013] extended the recovery of clean data from single sparse 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 existing 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
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 $\mathbf{x} \in \mathbb{R}^{H \times H}$, in view of spatial continuity, we use a small image patch $\mathbf{u} \in \mathbb{R}^{h \times h}(h \ll H)$ as a basic occlusion unit. For the texture of $\mathbf{u}$, we first generate a random mean value $\mu (0 + \delta \leq \mu \leq 255 - \delta)$ for all the pixels, under a variance parameter $\delta$. And each pixel $u^i$ is set as $u^i = \mu + \delta \times v (v \sim N(0, 1))$, where $v$ is a random value following the normal distribution. For color images, the gray unit $u$ is extended to all the color channels.

Overall, for the occlusion unit $u$, we can generate $s$ different versions with different mean value $\mu$, and randomly select $s$ locations in the face $\mathbf{x}$ to form a unit occlusion set $\pi^u = \{r,c\mid 0 \leq r, c \leq H - h\}$, where $r$ and $c$ are the starting row and column coordinates in $\mathbf{x}$. As an integrated occlusion, the $\pi^u$ can be embedded into $\mathbf{x}$ to obtain an occluded face $\mathbf{x}^u$. Intuitively, the randomness or complexity of an occlusion can be approximated by multivariate cooperation of $s$, $\mu$, $\delta$ 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 $\overline{\pi}^u = \pi^{u_{h_1}} \cup \ldots \cup \pi^{u_{h_s}}$. 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 2: Some examples of synthetic faces with occlusions.](image)

![Figure 3: 2D visualization (t-SNE) of the features extracted by ResNet-Inception-V1.](image)
To implement it conveniently, we often design a label loss to
b}
version. For one sample
present a biased guidance strategy for model training.
tive learning strategy for model optimization. To this end, we
cluded data inevitably interferes with the feature extraction
jointly dominate the distribution of class centers. Here, oc-
hypotheses imply that features of face and occlusion sources
region from occlusion may lead a raw model to learn non-facial
Due to the strong fitting ability of a deep model, the spatial re-
ficult to optimize the model to adapt to all the sample types.
channel, when multiple types of samples are used to train it
3.3 Biased Guidance Strategy for Model Training
For a data-driven model relying on a single feature extraction
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 dif-
ficult to optimize the model to adapt to all the sample types.
Due to the strong fitting ability of a deep model, the spatial re-
region from occlusion may lead a raw model to learn non-facial

\[ \mathcal{L} = \mathcal{L}_{x, \in I_N} + \lambda \frac{n}{o} \mathcal{L}_{x, \in I_O} \]  

(4)

where \( \lambda \) is empirically associated with the non-occlusion per-
centage in \( x_n \). \( n \) and \( o \) are the numbers of normal and occlu-
sion samples, which aim at the cost sensitivities of categories.

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

\[ \frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}_{x, \in I_N}}{\partial \theta}, \quad \frac{\partial \mathcal{L}}{\partial w} = (1 + \lambda \frac{n}{o}) \frac{\partial \mathcal{L}_{x_m, \in I_N}}{\partial w} \]  

(5)

Clearly, the parameters \( w \) 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
\( \theta \) in \( C \) 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. Par-
ticularly, 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 sam-
ple to update parameters of classifier and use the whole loss
to optimize CNN. Overall, this approach is suitable for a va-
riety 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 mapp-
ing 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
In 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 (N-N), verification between normal and occluded faces (N-O) and occluded face pairs (O-O).

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.

**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 $3 \times 3$ convolution. Before training, all the images are resized to $128 \times 128$ for L09 and $112 \times 112$ 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 $\lambda$**

As stated in Section 3.3, the parameter $\lambda$ 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 $x_o$ ($0 \leq \lambda \leq 1$). So we conduct two experiments to investigate its sensitiveness. Since we set $n = o$, the loss function is simplified as $L = L_{x_o \in I_N} + \lambda L_{x_o \in I_O}$.

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 $\lambda$. (ii) For L09, as the increase of $\lambda$, 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 $\lambda$ 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 $\lambda$, its value is simply set as 0.5 for subsequent experiments.
work training of all the evaluated methods, we apply random
the proposed method to different loss functions. For net-
In this section, we examine the generalization capability of
5.3 Study of Generalization Capability
owing AdamW [Loshchilov and Hutter, 2019]. We set $lr=1e-4$
for R18 with the ArcFace loss to avoid non-convergence, and
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 capa-
bility to occluded faces but also maintains the good perfor-
mance for normal faces.

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

Acknowledgments
This work is supported by the Science and Technology Inno-
vation 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).

<table>
<thead>
<tr>
<th>Method</th>
<th>N-N</th>
<th>N-O</th>
<th>O-O</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline ($I_N$)</td>
<td>94.38</td>
<td>79.90</td>
<td>64.91</td>
<td>$I_N$</td>
</tr>
<tr>
<td>Dropout(0.5)</td>
<td>94.44</td>
<td>80.42</td>
<td>65.35</td>
<td>$I_N$</td>
</tr>
<tr>
<td>Crop(0.6-1)</td>
<td>95.59</td>
<td>80.32</td>
<td>64.20</td>
<td>$I_N$</td>
</tr>
<tr>
<td>Crop(0.8-1)</td>
<td>95.81</td>
<td>81.96</td>
<td>65.28</td>
<td>$I_N$</td>
</tr>
<tr>
<td><strong>Baseline ($I_N+I_O$)</strong></td>
<td><strong>94.10</strong></td>
<td><strong>84.48</strong></td>
<td><strong>71.55</strong></td>
<td><strong>$I_N+I_O$</strong></td>
</tr>
<tr>
<td><strong>BFL</strong></td>
<td><strong>95.28</strong></td>
<td><strong>85.72</strong></td>
<td><strong>72.42</strong></td>
<td><strong>$I_N+I_O$</strong></td>
</tr>
<tr>
<td><strong>BFL+Dropout(0.5)</strong></td>
<td><strong>95.16</strong></td>
<td><strong>86.04</strong></td>
<td><strong>73.08</strong></td>
<td><strong>$I_N+I_O$</strong></td>
</tr>
<tr>
<td><strong>BFL+Crop(0.6-1)</strong></td>
<td><strong>95.91</strong></td>
<td><strong>87.66</strong></td>
<td><strong>74.77</strong></td>
<td><strong>$I_N+I_O$</strong></td>
</tr>
<tr>
<td><strong>BFL+Crop(0.8-1)</strong></td>
<td><strong>96.13</strong></td>
<td><strong>88.12</strong></td>
<td><strong>75.10</strong></td>
<td><strong>$I_N+I_O$</strong></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>N-N</th>
<th>N-O</th>
<th>O-O</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline ($I_N$)</td>
<td>95.43</td>
<td>79.87</td>
<td>65.38</td>
<td>$I_N$</td>
</tr>
<tr>
<td>Dropout(0.5)</td>
<td>95.32</td>
<td>79.34</td>
<td>64.29</td>
<td>$I_N$</td>
</tr>
<tr>
<td>Crop(0.6-1)</td>
<td>96.96</td>
<td>80.32</td>
<td>64.58</td>
<td>$I_N$</td>
</tr>
<tr>
<td>Crop(0.8-1)</td>
<td>97.77</td>
<td>81.39</td>
<td>65.08</td>
<td>$I_N$</td>
</tr>
<tr>
<td><strong>Baseline ($I_N+I_O$)</strong></td>
<td><strong>95.08</strong></td>
<td><strong>83.11</strong></td>
<td><strong>68.70</strong></td>
<td><strong>$I_N+I_O$</strong></td>
</tr>
<tr>
<td><strong>BFL</strong></td>
<td><strong>95.55</strong></td>
<td><strong>83.41</strong></td>
<td><strong>70.43</strong></td>
<td><strong>$I_N+I_O$</strong></td>
</tr>
<tr>
<td><strong>BFL+Dropout(0.5)</strong></td>
<td><strong>96.12</strong></td>
<td><strong>84.01</strong></td>
<td><strong>70.05</strong></td>
<td><strong>$I_N+I_O$</strong></td>
</tr>
<tr>
<td><strong>BFL+Crop(0.6-1)</strong></td>
<td><strong>97.45</strong></td>
<td><strong>87.22</strong></td>
<td><strong>73.12</strong></td>
<td><strong>$I_N+I_O$</strong></td>
</tr>
<tr>
<td><strong>BFL+Crop(0.8-1)</strong></td>
<td><strong>97.35</strong></td>
<td><strong>85.98</strong></td>
<td><strong>70.78</strong></td>
<td><strong>$I_N+I_O$</strong></td>
</tr>
</tbody>
</table>

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

5.2 Comparison with Other Methods
In this subsection, we compare our 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 classifi-
cation layer. $\alpha-\beta$ in Crop($\alpha-\beta$) indicates the scale of random
cropping for faces. The results are reported with the aver-
age 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 $\lambda=0.5$.

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 im-
prove 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 effec-
tive 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 oc-
cclusion sources in $I_O$ are thoroughly different from the real-
istic 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 net-
work training of all the evaluated methods, we apply random
Crop (0.8-1.0) to the training images. Meanwhile, we use
the $L_2$ regularization with the weight decay of 0.01 follow-

ting AdamW [Loshchilov and Hutter, 2019]. We set $lr=1e-4$
for R18 with the ArcFace loss to avoid non-convergence, and
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 capa-
bility to occluded faces but also maintains the good perfor-
mance for normal faces.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>N-N</th>
<th>N-O</th>
<th>O-O</th>
<th>Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Baseline</td>
<td>95.85</td>
<td>81.29</td>
<td>65.63</td>
<td>Softmax</td>
</tr>
<tr>
<td>BFL</td>
<td>BFL</td>
<td>96.79</td>
<td>89.06</td>
<td>76.61</td>
<td>CosFace</td>
</tr>
<tr>
<td>L09</td>
<td>Baseline</td>
<td>98.63</td>
<td>83.14</td>
<td>68.36</td>
<td>ArcFace</td>
</tr>
<tr>
<td>BFL</td>
<td>BFL</td>
<td>98.52</td>
<td>90.81</td>
<td>76.98</td>
<td></td>
</tr>
<tr>
<td>R18</td>
<td>Baseline</td>
<td>98.32</td>
<td>85.14</td>
<td>68.36</td>
<td></td>
</tr>
<tr>
<td>BFL</td>
<td>BFL</td>
<td>98.68</td>
<td>91.36</td>
<td>77.54</td>
<td></td>
</tr>
<tr>
<td>Lv2</td>
<td>Baseline</td>
<td>97.06</td>
<td>79.29</td>
<td>63.29</td>
<td></td>
</tr>
<tr>
<td>BFL</td>
<td>BFL</td>
<td>98.30</td>
<td>86.18</td>
<td>70.07</td>
<td></td>
</tr>
<tr>
<td>Rv1</td>
<td>Baseline</td>
<td>98.89</td>
<td>82.70</td>
<td>65.89</td>
<td></td>
</tr>
<tr>
<td>BFL</td>
<td>BFL</td>
<td>98.80</td>
<td>88.14</td>
<td>72.47</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>98.17</td>
<td>75.12</td>
<td>64.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>98.67</td>
<td>83.91</td>
<td>70.27</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>99.57</td>
<td>89.83</td>
<td>72.77</td>
<td>CosFace</td>
</tr>
<tr>
<td></td>
<td></td>
<td>99.47</td>
<td>93.40</td>
<td>79.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>99.33</td>
<td>88.27</td>
<td>74.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>99.15</td>
<td>91.83</td>
<td>79.97</td>
<td></td>
</tr>
</tbody>
</table>

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


