Locality-Aware Channel-Wise Dropout for Occluded Face Recognition

Mingjie He, Jie Zhang, *Member, IEEE*, Shiguang Shan, *Fellow, IEEE*, Xiao Liu, Zhongqin Wu, and Xilin Chen, *Fellow, IEEE*

***Abstract*— Face recognition remains a challenging task in unconstrained scenarios, especially when faces are partially occluded. To improve the robustness against occlusion, augmenting the training images with artificial occlusions has been proved as a useful approach. However, these artificial occlusions are commonly generated by adding a black rectangle or several object templates including sunglasses, scarfs and phones, which cannot well simulate the realistic occlusions. In this paper, based on the argument that the occlusion essentially damages a group of neurons, we propose a novel and elegant occlusion-simulation method via dropping the activations of a group of neurons in some elaborately selected channel. Specifically, we first employ a spatial regularization to encourage each feature channel to respond to local and different face regions. Then, the locality-aware channel-wise dropout (LCD) is designed to simulate occlusions by dropping out a few feature channels. The proposed LCD can encourage its succeeding layers to minimize the intra-class feature variance caused by occlusions, thus leading to improved robustness against occlusion. In addition, we design an auxiliary spatial attention module by learning a channel-wise attention vector to reweight the feature channels, which improves the contributions of non-occluded regions. Extensive experiments on various benchmarks show that the proposed method outperforms state-of-the-art methods with a remarkable improvement.**

***Index Terms*— Occluded face recognition, locality-aware channel-wise dropout, spatial attention module.**

I. Introduction

With the huge success of deep learning, a remarkable improvement has been achieved for face recognition under controlled settings (i.e., occlusion-free images, near-frontal poses, neutral expressions, normal illuminations, etc.). However, in realistic unconstrained scenarios, face recognition remains a challenging task due to various factors including very large poses, very low resolution and occlusions. Among these factors, the occlusion is an intractable problem which leads to a severe degeneration in recognition accuracy.

The occlusions always bring about two primary issues, i.e., the missing of facial information and the noise from occlusion. To improve the robustness against occlusion, many efforts [1]–[6] have been made to recover the occluded faces. The work in [2] uses a multi-scale spatial long short-term memory (LSTM) encoder to encode occluded face patches, and then another LSTM is employed to reconstruct the occlusion-free face image. Based on the Generative Adversarial Network (GAN) [7], another work [5] proposes a face completion model to generate visually plausible contents for the occluded face regions. Although a huge progress has been made, the performance of occlusion removal is still far from satisfactory. The main reason is that these methods are commonly trained with artificial occluded images. For instance, the images are manually generated by randomly putting a black rectangle or several object templates including sunglasses, scarfs, phones, and cups on them, which differ significantly from real-life occlusions. Therefore these methods always suffer from poor generalizations under realistic scenarios. Besides, how to well recover the occluded face regions as well as preserve the identity information is another challenge for these methods.

Another type of methods focus on suppressing the noise caused by occlusions. They attempt to discard the corrupted feature elements which are extracted from the occluded regions [8]–[13]. PDSN [12] builds a mask dictionary in advance by comparing the features extracted from normal faces and those from occluded faces. During the recognition process, the occluded facial regions are detected by a segmentation network and then the noise is removed by discarding the corrupted feature elements retrieved from the mask dictionary. To some extent, the occlusion discarding method relieves the influence of occlusions. However, it may be non-trivial to precisely detect the real-world occlusions which usually have various shapes and textures since the occlusion detection modules are also trained with artificial occlusions. Even if the corrupted feature elements have been located perfectly, directly zeroing out them will incorporate a peculiar pattern into the final feature representation. Moreover, due to the uncertain size of the occluded region, the final feature will have an unfixed number of valid elements. Thus, the traditional metrics designed for the fixed-length vectors may fail for evaluations.

Manuscript received April 12, 2021; revised November 22, 2021; accepted November 23, 2021. Date of publication December 10, 2021; date of current version January 3, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700800, in part by the National Natural Science Foundation of China under Grant 61806188 and Grant 61976219, and in part by the Shanghai Municipal Science and Technology Major Project under Grant 2017SHZDZX01. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ran He. *(Corresponding author: Shiguang Shan.)*

Mingjie He, Jie Zhang, and Xilin Chen are with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing 100190, China, and also with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: hemingjie@ict.ac.cn; zhangjie@ict.ac.cn; xlchen@ict.ac.cn).

Shiguang Shan is with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing 100190, China, also with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China, also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai 200031, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: sgshan@ict.ac.cn).

Xiao Liu and Zhongqin Wu are with Tomorrow Advancing Life Education Group (TAL), Beijing 100080, China (e-mail: liuxiao15@tal.com; wuzhongqin@tal.com).

Digital Object Identifier 10.1109/TIP.2021.3132827

One simple way of tackling occluded face recognition problem is to leverage a massive number of occluded faces in the real world to directly train deep neural networks, which are forced to learn occlusion-robust face features. However, it is hard to collect such a training set. As an alternative, augmenting the training images with artificially synthesized occlusions has been studied and significant improvements have been witnessed in [14]–[17]. However, the image-level occlusion simulation is still not an elegant solution since the artificial occluded images are also generated by using limited hand-craft occlusion templates, which cannot fully represent the arbitrary patterns in realistic occlusions.

The common issue of all the aforementioned methods is that the artificially synthesized occlusion cannot precisely represent the real-world occlusions, leading to poor generalizations under realistic scenarios. In this paper, we propose a novel and elegant method which can better simulate realistic occlusions. Here, we hold the opinion that the occlusion essentially damages a group of neurons. To synthesize different occlusions, a natural approach is to drop out the activations of various neurons. However, conventional dropout operation cannot simulate the real-life occlusion. The reason is that real-life occlusion affects a contiguous region in an activation map, while the conventional dropout operation discards discrete activations. To better simulate the feature damaged by occlusions, we propose the revised dropout method, namely the locality-aware channel-wise dropout (LCD) to drop a group of activations which are affected by the same facial occlusion. For conventional neural networks, partial occlusion usually affects activations across a large number of feature channels, which makes it difficult to simulate occlusions with a few channels. Inspired by [18], we encourage each feature channel to only respond to local and different face regions via the spatial regularization. As shown in Fig. 1, the heat maps of 3 different feature channels are visualized and it can be seen that these channels respond to various face regions. Then, LCD can simulate the occlusion at local regions by dropping out a few feature channels.

Setting our LCD at a middle depth in the neural network can encourage its succeeding layers to minimize the intra-class feature variance caused by occlusions, leading to improved robustness against occlusion. Furthermore, we design an auxiliary spatial attention module which learns a channel-wise attention vector to reweight the channels during the feature extraction process. After jointly trained with our LCD, the deep network is optimized to focus more on the channels which activated on the non-occluded regions and suppress others which are affected by the occluded regions.

Compared with previous works, our method has three major advantages: 1) our method does not require artificially synthesized occlusions. Instead, it gracefully simulates the realistic occlusions in intermediate features. 2) Different from previous works which use additional module to detect or recover the occluded region, our method imposes minor increase in model complexity during the inference phase. 3) Our method is a more practical approach which can be seamlessly integrated with any existing face recognition method for improving robustness to occlusions.

The main contributions of this paper are summarized as follows:

- We propose a novel method to better simulate realistic occlusions by dropping a group of activations in intermediate features. It significantly improves the robustness to occlusions by encouraging the neural network to emphasize on learning discriminative features from the non-occluded face regions.
- An auxiliary spatial attention module is designed to improve the contributions of non-occluded regions by adaptively reweighting the feature channels.
- Our method significantly outperforms the state-of-the-art methods on IJB-C, LFW and MegaFace benchmarks, especially on the IJB-C dataset with large-scale real-occluded face images.

The rest of the paper is organized as follows. Section II gives a brief overview of the related works. Section III introduces the detailed formulation of the proposed method, followed by a discussion with other state-of-the-art methods in Section IV. Section V presents the ablation study and the experimental results on three databases. Finally, the conclusion is summarized in Section VI.

II. Related Works

The existing occlusion robust face recognition methods can be broadly grouped into the following three categories: occluded face completion methods, occlusion-aware discarding methods and occlusion-robust feature extraction methods. In this section, we provide a brief overview of the recent works which are most relevant to this paper.

A. Occluded Face Completion Methods

The occluded face completion methods are pixel-level approaches which aim to recover the occluded face regions. Considering the low-rank property of non-occluded images, early works attempt to solve the problem by using robust principal component analysis (PCA) to reconstruct the corrupted low-rank face images [19]–[23]. The work in [21] proposes the robust principal component analysis (RPCA) which improves the performance of removing shadows from face images. In [22], an important extension of RPCA, namely the low-rank representation (LRR) is presented to extend the recovery of corrupted images from a single subspace to multiple subspaces. Both RPCA and LRR assume that the occluded pixels are sparse, while real-world face images usually contain dense occlusions, which make the matrix non-sparse. To this end, [23] presents the double nuclear norm-based matrix decomposition to remove the dense occlusions.

Recently, more works resort to deep learning for improving the occluded face completion [1]–[6], [24]. In [1], the difference between the activation values of two stacked sparse denoising auto-encoders (SSDAs) is used to indicate occluded and un-occluded face regions. Then, the final occlusion-free image is reconstructed by transferring the encoding activations of the un-occluded region to the occluded region. In [3], the context encoder combines the auto-encoder architecture with context information of the occluded part to produce visually pleasing images. To further improve the context encoder, [4]

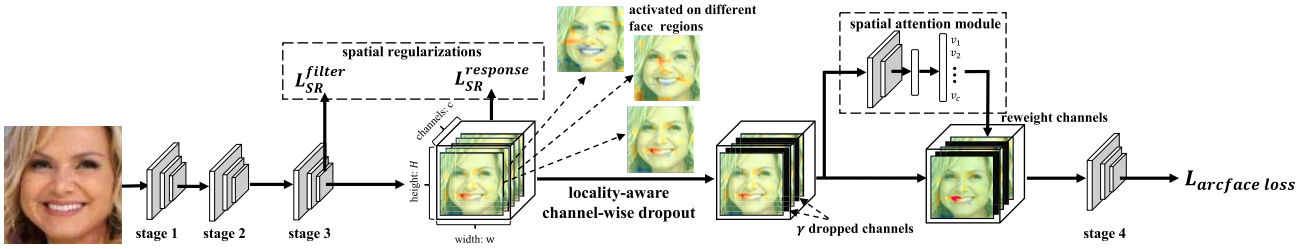

Fig. 1. The overall architecture of the proposed locality-aware channel-wise dropout (LCD). Two spatial regularization losses are employed to encourage each feature channel to respond to local and different face regions. Then, our LCD achieves a feature-level occlusion simulation by randomly dropping out a few feature channels. Furthermore, the auxiliary spatial attention module learns a channel-wise attention vector to reweight the feature channels, which improves the contributions of non-occluded regions.

introduces both global and local discriminators. Specifically, the global discriminator pursues the global consistency of the overall image and the local discriminator looks at a small area centered at the reconstructed region to judge the quality in details. To ensure the new generated contents more photo-realistic, a semantic parsing loss is developed in [5]. To refine local face textures, a 3D morphable model (3DMM) is utilized in [6] to further assist the learning of the local discriminator. In the unsupervised face normalization method (FNM) [24], multiple local discriminators are integrated into a novel unsupervised framework. It generates impressive high quality faces which have dispelled various face variations including occlusions.

Overall, the above-mentioned face completion methods have shown promising results of transforming occluded faces to un-occluded ones. However, two major issues still remain. Firstly, except for a few methods (e.g., the FNM [24]), most previous methods require pair-wise training data (i.e., one occluded face and one un-occluded face of the same person). Unfortunately, such training sets containing a mass of pair-wise faces with natural occlusion are extremely rare. An alternative and commonly used approach is using synthetically occluded faces. However, these synthetically occluded faces, e.g., using manually designed occlusion templates or a black/white rectangle cannot fully represent natural occluded faces. Secondly, the faces generated by GAN-based methods are usually visually pleasing. However, how to remove the occlusion while preserving the identity information still remains a challenging problem.

B. Occlusion-Aware Discarding Methods

These approaches aims to remove the noise caused by occlusions with two pipelines, which either discards the occluded pixels before the face feature extraction, or discards the corrupted feature elements during the feature extraction.

Following the former pipeline, some works detect the occlusions first and then extract a feature representation from the non-occluded regions only. The early works [8]–[10], [25] usually employ a nearest neighbor classifier (NNC) or a support vector machine (SVM) to classify the occluded face regions. Since these methods are designed based on the traditional feature descriptors, it is non-trivial for them to obtain discriminative ability for face recognition in complex scenarios. Recently, the LPD [13] designs a neural network to locate the latent facial parts which are less affected by a specific occlusion (i.e., the respirator), and then extracts discriminative features from the selected latent part.

Following the spirit of the latter pipeline, the mask leaning methods [11], [12] locate and discard the corrupted feature elements rather than the occluded pixels. In [11], the MaskNet adaptively learns feature masks for occluded face images and automatically assigns lower weight to the hidden units activated by the occluded face regions. The PDSN [12] establishes a mask dictionary to represent the correspondence between the occluded facial block and the corrupted feature elements. During the testing phase, a segmentation network is employed to detect the occluded facial blocks, and then the corrupted feature elements are set to zero by retrieving the relevant dictionary items.

Although above-mentioned occlusion-aware discarding methods can alleviate the occlusion issue, completely discarding the occluded regions still takes the risk of reducing the system reliability. Firstly, precise and fine-grain occlusion detection is an essential prerequisite for these methods. However, such occlusion detection is non-trivial to obtain and a coarse detection will increase the risk of losing discriminative information or inducing unreliable information. Furthermore, even if the occluded face regions or the corrupted feature elements have been located perfectly, directly zeroing out the corrupted elements still incorporates a peculiar pattern into the final feature, which may harm the recognition accuracy. Moreover, due to the arbitrary occlusion, the feature vectors for occluded faces have unfixed number of valid elements. Thus, the conventional metrics for fixed-length vectors are not fully applicable. Although works on the instance-to-class distance [26] and the reconstruction-based similarity measurement [27] are proposed to tackle the above problem, the commonly used window sliding makes them relatively time-consuming. On the other side, directly ignoring occluded elements will potentially break the global cues of face images such as chin contours, which is also harmful to face recognition system.

C. Occlusion-Robust Feature Extraction Methods

Directly learning occlusion-robust feature is the most straightforward and effective way to handle occlusion face recognition. For this purpose, many efforts are devoted to seeking a feature space that is less affected by occlusion and

meanwhile preserves the discriminative capability of distinguishing identities. Many works seek such feature space via the sparse representation classification (SRC) [28], in which the occluded image is represented by a linear combination of training samples plus a sparse constraint term accounting for occlusions. The LR-LUM [29] combines both the robust sparsity constraint and the low-rank constraint which outperforms previous methods in handing structured occlusion such as sunglasses and scarfs. Other works [30]–[32] extend the sparse representation by combining more discriminative feature descriptors. In [32], the JCR-ACF proposes a joint and collaborative representation with local adaptive convolution feature, which can improve the recognition accuracy by employing information from different local face regions. Although these SRC-based methods have made considerable progresses, they do not generalize well in practical scenarios since they requires that the test samples have identical identities with a pre-defined close-set.

Owing to leveraging massive training sets, recent deep learning methods reveal significant superiority on face recognition. Enlarging the training datasets with sufficient occluded faces may be an effective way to improve the occlusion robustness of face embedding. Unfortunately, such training sets of a mass of identities are extremely rare. As an alternative solution, synthetic image augmentation method has been studied by previous works [14]–[17]. The work in [14] enriches the training set by synthesizing occluded faces with various pre-defined hairstyles and glasses templates. Although an accuracy gain has been witnessed, the diversity of the manually designed occlusion templates needs to be further improved. More recently, BFL [17] proposes an enhanced augmentation schema to randomly generate multi-scale spatially occluded samples and then modifies the loss to balance the impact of normal and occluded samples for training. To some extent, these augmentation methods improve the robustness of neural network, but the discrepancy between the synthetic and real occluded faces still limits the further improvements of the robustness.

Another line of researches [33]–[35] focus on the attention mechanism for robust feature extraction. The state-of-the-art method named InterpretFR [35] employs a Siamese network to compare the feature elements from a normal face with and without synthetic occlusion. Then it encourages the neural network to identify the input face solely based on the feature elements which are less sensitive to occlusions. However, it still needs synthesizing occluded faces by using artificial occlusion templates and fails to generalize well on unseen occlusions. In contrast, our method can realistically simulate arbitrary occlusions via dropping out a random group of filter responses, leading to an improved performance under real world scenarios.

III. Methodology

A. Overview

The proposed method attempts to learn occlusion robust face features by simulating occlusions during the training process. Different from previous methods which augment face images with synthesized occlusions, we propose to directly simulate the influence of arbitrary occlusions on intermediate features. As the occlusion essentially damages a group of neurons, we propose a revised dropout method, namely the locality-aware channel-wise dropout (LCD) to simulate occlusions by dropping a group of feature channels. Considering that, for the conventional neural networks, the occlusion usually affects the activations from most channels which is even impossible to simulate the occlusions by dropping several channels. Therefore, we first employ a spatial regularization to encourage each feature channel to respond to local and different face regions. Then, our LCD simulates the occlusions by dropping out a few feature channels, in which sense we name this method as locality-aware channel-wise dropout. Moreover, to improve the contributions of non-occluded regions for learning occlusion robust features, we design an auxiliary spatial attention module to reweight the feature channels. The whole framework shown in Fig. 1 is end-to-end trainable and can be easily applied on any existing convolution neural network.

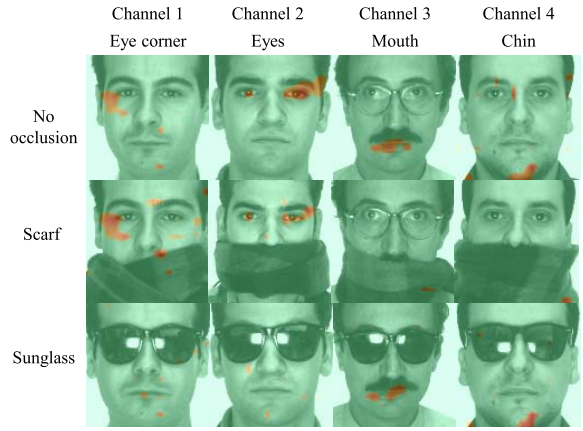

Fig. 2. The visualization of four feature channels regularized by the spatial regularization. As seen, each feature channel activates on relatively consistent face regions and the occlusion on that region makes the corresponding feature channel less activated.

B. Locality-Aware Channel-Wise Dropout

1) Spatial Regularization of Feature Channels: For the purpose of locality-awareness of the filters (i.e., channels), we need to enforce each feature channel to respond to different local face regions. We noticed that the spatial activation diversity loss exploited in [35] (modified from [18]) can actually achieve this effect. Therefore, we borrow the two diversity losses L_{SR}^{filter} and $L_{SR}^{response}$, in [35], respectively encouraging the filters and their responses to be orthogonal. Here, L_{SR}^{filter} is designed to constrain the filters orthogonal by penalizing their correlations:

$$L_{SR}^{filter} = \sum_{i \neq j} \left| \sum_{p} \frac{\left\langle \mathbf{w}_i^p, \mathbf{w}_j^p \right\rangle}{\left\| \mathbf{w}_i^p \right\|_F \left\| \mathbf{w}_j^p \right\|_F} \right|, \tag{1}$$

where $\mathbf{w}_i^p$ denotes the column of filter $\mathbf{w}_i$ at spatial location p. The second term $L_{SR}^{response}$ is exploited to further decorrelate

the filters' response maps:

$$L_{SR}^{response} = \sum_{i \neq j} \left\| \frac{\langle \mathbf{f}_i, \mathbf{f}_j \rangle}{\|\mathbf{f}_i\|_F \|\mathbf{f}_j\|_F} \right\|^2, \tag{2}$$

where $\mathbf{f}_i$ denotes the response of i-th filter (i.e., the i-th channel of features).

We visualize four feature channels regularized by the spatial regularization in Fig. 2, in which each column shows the same feature channel for different images. As seen, each feature channel activates on relatively consistent face parts, e.g., eye corner, eyes, mouth and chin. Moreover, if there is a natural occlusion, e.g., eyeglass or scarf, the corresponding feature channel is less activated, while other feature channels activated on non-occluded regions are not affected.

2) Simulating Occlusions via Channel-Wise Dropout: Given an input feature map $\mathbf{F} \in \mathbb{R}^{c \times h \times w}$, we first generate an all-one mask matrix $\mathbf{M} \in \mathbb{R}^{c \times h \times w}$ with the same size as F, where c, h, w denotes the channel number, the height, the width of the feature map, respectively. Second, we randomly sample γ distinct channel indexes $\{r_1, r_2, .., r_\gamma\}$ from the c channels. Then, the mask values for these channels are set to zero:

$$M_{i,j,k} = \begin{cases} 0 & i \in \{r_1, r_2, .., r_\gamma\} \\ 1 & \text{others} \end{cases}. \tag{3}$$

Finally, the output of the LCD is obtained by the product of the mask matrix and the input:

$$\mathbf{F}_{drop} = \mathbf{F} \circ \mathbf{M}, \tag{4}$$

where $\circ$ denotes Hadamard product. As shown in Fig. 1, all the $\gamma \times h \times w$ feature elements within the γ feature maps will be dropped out to zero. At the training stage, our LCD performs as an effective occlusion simulation and encourages the network to identify the input face solely based on the remaining features, which makes it more robust to occlusions. It should be mentioned that similar to the conventional dropout, the LCD is not employed during the inference process.

In our method, γ is a crucial parameter which controls how many feature channels will be dropped out. In other words, it determines how many face regions will be discarded during the training process. A larger γ simulates a more severe occlusion where more regions in the face are occluded. To simulate the complex pattern of realistic occlusions, a dynamic γ for each training samples is required. With this in mind, the γ is designed as a stochastic variable satisfying uniform distribution within $[\gamma_{\min}, \gamma_{\max}]$. A larger $\gamma_{\max}$ (e.g., $\gamma_{\max} = 0.6 * c$) is recommended to improve the robustness to severer occlusions.

Since the activations of convolutions layers are commonly normalized by batch-normalization (BN) layers, how to make the LCD compatible with the BN layers is worth exploring. In a conventional BN layer, the feature elements sharing the same channel index will be normalized together. However, this process will have some problems when applying the LCD before the BN layer. Specifically, for a mini-batch with n samples, let $x_{t,i,j,k}$ denote its t, i, j, k-th element in the $n \times c \times h \times w$ feature tensor. The conventional BN layer computes the mean for i-th channel:

$$u_i = \frac{1}{nhw} \sum_{t}^{n} \sum_{j}^{h} \sum_{k}^{w} x_{t,i,j,k}. \tag{5}$$

However, since features of a sub-set channels are set to zero, the number of valid samples for the i-th channel is no longer equal to n. To resolve this problem, the calculation of u_i must be modified to:

$$u_i = \frac{1}{(n - \eta_i)hw} \sum_{t}^{n} \sum_{j}^{h} \sum_{k}^{w} x_{t,i,j,k}, \tag{6}$$

where η_i denotes the number of training samples which have zero values in the i-th channel. Besides, the calculation of the channel variance also requires similar modification. To be free from these modifications, always setting the LCD after the conventional BN layer is an alternative solution. Moreover, this simple setting is even more favorable to obtain stable BN parameters as all the training samples are involved in the computation of mean and variance.

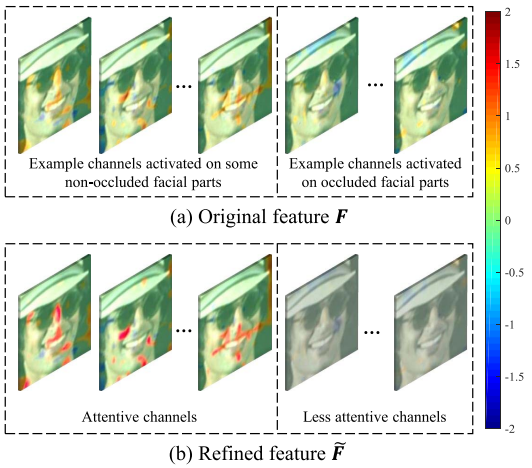

Fig. 3. (a) Due to the occlusion by sunglass and hat, the feature channels corresponding to the forehead and the eyes are less activated. (b) The spatial attention module attempts to reweight the features, making the network focus on the channels which activated on the non-occluded facial parts.

C. Spatial Attention Module

By randomly dropping out feature channels which respond to local and different face regions, the channel-wise dropout with the spatial regularization achieves occlusion-invariant feature learning by implicitly simulating various occlusions in the training stage. To further attentively emphasize the non-occluded facial features of the current input face image during the inference stage, we design an attention module to explicitly reweight the feature channels. As shown in Fig. 3, it makes the network focus on the feature channels activated on some non-occluded facial parts. It is worth noting that, by utilizing the aforementioned spatial regularization, each feature channel is encouraged to respond to local and different face regions. In this sense, reweighting these feature channels actually performs as a spatial-wise attention approach, which is named as the spatial attention module (SAM).

For the input feature map $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_c]$, our goal is to learn an attention vector $\mathbf{v} = [v_1, v_2, \ldots, v_c]$ which controls the weight of each feature channel. As shown in Fig. 1, a light-weight module is designed to learn the attention vector $\mathbf{v}$. Specifically, we first employ a 1×1 convolution layer and a fully-connected layer to obtain a global view of all input feature maps. Then, another fully-connected layer is utilized to extract the channel-wise attention vector. It is worth noting that previous attention block in SENet [36] utilizes a global average pooling as a global descriptor, while in our method the global pooling of dropped zero value feature maps will suffer from local minimum. To this end, the 1×1 convolution layer in our proposed module is a necessary and effective information aggregation strategy to achieve attentions.

The refined feature map $\widetilde{\mathbf{F}}$ is obtained by channel-wise multiplication between each feature map $\mathbf{f}_x$ and its corresponding attention scalar v_x:

$$\widetilde{\mathbf{F}} = \mathbf{F} \cdot \mathbf{v} = [\mathbf{f}_1 \cdot v_1, \mathbf{f}_2 \cdot v_2, \ldots, \mathbf{f}_c \cdot v_c]. \tag{7}$$

D. Joint Loss Function

In our framework, we employ ArcFace [37] as the face identification loss. Suppose that we have a training batch with n images, y_i denotes the class label of i-th training image. θ_j is defined as the angle between the feature vector and the j-th class center. The ArcFace loss is formulated as:

$$L_{arcface} = -\frac{1}{n}\sum_{i=1}^{n} \frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))} + \sum_{j \neq y_i} e^{s\,cos(\theta_j)}}, \tag{8}$$

where m denotes the angular margin and s is the feature scaling parameter.

The joint loss function is a weighted sum of the face identification loss and the aforementioned spatial regularization losses:

$$L_{total} = L_{arcface} + \alpha L_{SR}^{filter} + \beta L_{SR}^{response}, \tag{9}$$

where α and β are the loss weights for the filter orthogonal loss and the response orthogonal loss, respectively.

During the training process, we first conduct the forward propagation of the backbone' shallow stages (1, 2, 3) and obtain the intermediate features. Then, as shown in Fig. 1, we calculate the filter orthogonal loss L_{SR}^{filter} and the response orthogonal loss $L_{SR}^{response}$. Whereafter, for each training sample i, we get a random drop out rate γ_i and randomly drop out γ_i channels in its intermediate features. Then, the spatial attention module learns the attention vector $\mathbf{v}_i$ and the refined feature $\widetilde{\mathbf{F}}_i$ is calculated by Eq. 7. Taking $\widetilde{\mathbf{F}}_i$ as input, the succeeding layers in the backbone will extract the final face embedding and the face identification loss will be calculated. After obtaining the joint loss via Eq. 9, the whole network will be updated by backward propagation. Formally, the Algorithm 1 summarizes the training process.

Algorithm 1 Training With the Locality-Aware Channel-Wise Dropout and the Spatial Attention

Input: a mini-batch with n training images and their labels, $\gamma_{\min}$, $\gamma_{\max}$, the loss weights α and β.
while *not converged* **do**
  Forward propagation to get the intermediate features: $\mathbf{F}_i \in \mathbb{R}^{c \times h \times w}, i \in [1, n]$;
  Compute the filter orthogonal loss L_{SR}^{filter};
  Compute the response orthogonal loss $L_{SR}^{response}$;
  for *each* $i \in [1, n]$ **do**
    Randomly select the γ_i : $\gamma_i \sim Uniform(\gamma_{\min}, \gamma_{\max})$;
    Randomly sample γ_i channels from the $\mathbf{F}_i$;
    Set all the values in these γ_i channels to zero;
    Compute the attention vector $\mathbf{v}_i$;
    Compute the refined feature: $\widetilde{\mathbf{F}}_i = \mathbf{F}_i \cdot \mathbf{v}_i$
  end
  Compute the output of the succeeding layers;
  Compute the identification loss $L_{arcface}$;
  Compute the total loss L_{total}: $L_{total} = L_{arcface} + \alpha L_{SR}^{filter} + \beta L_{SR}^{response}$;
  Backward propagation;
  Update the weights of the neural network;
end
Output: the trained neural network.

IV. Discussion

The proposed LCD method can be categorized as a form of structured dropout. Conventionally, other structured dropout methods are designed for the purposes of alleviating the over-fitting issue and improving the generalization ability. In contrast, our LCD is designed for simulating facial occlusions and achieving occlusion robustness. In this section, we discuss the comparisons of our LCD with two representative structured dropout methods, i.e., the DropBlock [38] and the weighted channel dropout (WCD) [39]. Furthermore, we also provide a comparison with the occluded face recognition method InterpretFR [35] which is most relevant to our method.

A. Differences From DropBlock [38]

Both our method and DropBlock drop some activations in the intermediate feature layers, but they differ in two aspects. 1) The Dropblock drops partial regions within each feature channel for enhancing feature learning while our method employs spatial regularization to enforce each feature channel to respond to local face regions and further drop several feature channels to simulate partial occlusions. 2) Our method further proposes spatial attention module to improve the contributions of undamaged neurons to further promote the robustness to occlusion. Experimental results on IJB-C datasets show our method significantly outperforms Dropblock for occluded face recognition.

B. Differences From the Weighted Channel Dropout [39]

Both our method and the weighted channel dropout (WCD) drop channels but for different purposes and in different ways. Firstly, in terms of designing purpose, the WCD is for alleviating the over-fitting issue when fine-tuning neural

network on small datasets, while our method aims to improve the robustness to partial occlusion for face recognition. Secondly, in terms of what channels to drop, the WCD drops out the feature channels which have relatively lower activation magnitudes, while in our method each feature channel is enforced to respond to local facial regions and thus can be dropped to simulate the feature corruption due to the occlusion of that region. Comparison experiments on IJB-C datasets show the superiority of our method in handling occluded face recognition problem.

C. Differences From InterpretFR [35]

Both our method and InterpretFR resort to occlusion robust feature learning for tackling face recognition under occlusions. However, we have two major differences. 1) InterpretFR synthesizes occlusions by putting black boxes on faces, which is unrealistic and the performance may severely degenerate under real world scenarios. Instead of simply utilizing fixed templates to synthesize occlusion, we propose the locality-aware channel-wise dropout to simulate more realistic occlusions, leading to performance improvement for occluded face recognition. 2) During training, InterpretFR masks out the elements of the final face representation sensitive to the occlusion, which may be sub-optimal as it only leverages remaining features for recognition without considering information recovering. Differently, we perform channel-wise dropout on the stage 3 of ResNet and design spatial attention module in the stage 4 to implicitly recover the absent information as verified by experiment in Section V-C, which is more favorable to tackle face recognition under occlusions.

V. Experiments

A. Datasets

The training images are collected from the MS-Celeb-1M dataset [40]. Since this dataset contains many labeling noise, we manually clean it and finally collect 3.7 Million images from 50K identities. The revised dataset is used as our training set for the experiments. We extensively test our method on three popular benchmarks, including IJB-C [41], MegaFace [42] and LFW [43].

1) IJB-C: As an extension of previous IJB-A [44] and IJB-B [45], the IJB-C [41] is a large-scale dataset which contains 117, 542 video frames and 31, 334 images. As 57% of the face images in IJB-C are natural occluded, this dataset is a commonly used benchmark for occlusion-robust face recognition. In this work, we employ two evaluation settings. First, we conduct comparisons on the holistic IJB-C dataset (including both occluded faces and non-occluded faces) for general evaluations. Second, to further verify the effectiveness of tackling occlusions, the occlusion subset of IJB-C (the occluded face images only) is employed for evaluations. All the two settings follow the standard IJB-C testing protocol. The true accept rate (TAR) and the false accept rate (FAR) are used as the evaluation metrics.

2) MegaFace: The MegaFace challenge 1 (MF1) benchmark [42] evaluates how the face recognition method performs with a huge scale of distracters. Specifically, the gallery set in MF1 contains one million face distractors, and the probe set Facescrub [46] contains 106,863 face images of 530 identities. In the testing pipeline, each Facescrub image will be added into the galley set and the remaining images of the same identity are exploited as probes. The rank-1 identification accuracy is used as the measurement of the face recognition performance. It should be noted that the un-cleaned version of MegaFace datasets are employed in the evaluation for the fair comparison with the state-of-the-art methods.

3) LFW: The LFW [43] is a well-known unconstrained face verification benchmark. It contains 13,233 images form 5,749 identities. In our work, we follow the standard 10-fold cross validation protocol to report the mean accuracy on the 6,000 testing image pairs.

TABLE I

Performance on IJB-C Occlusion Subset With Channel-Wise Dropout Applied at Different Depths of the Network

Methods	@FAR=.0001	@FAR=.001	@FAR=.01
baseline	39.69	88.52	95.14
channel-wise dropout (stage 1)	38.77	88.18	95.04
channel-wise dropout (stage 2)	53.57	90.55	95.73
channel-wise dropout (stage 3)	57.88	90.80	95.48
channel-wise dropout (stage 4)	37.59	88.96	95.48
channel-wise dropout (stage 2+3)	53.03	90.31	95.50
channel-wise dropout (stage 1+2+3)	51.55	90.12	95.46

B. Implementation Details

In our experiments, the face images are aligned based on five facial landmarks detected by [47] and normalized to a size of 112×112. We employ the ResNet-50 [48] as the baseline network and implement our method on the Tensorflow [49] platform. The Arcface loss [37] with the margin of 0.5 and the scale of 64 is utilized as the identification loss in our experiments. All the models in our experiments are trained on four NVIDIA TITAN XP GPUs by SGD. The loss weight of the ArcFace loss is set to 1 and the loss weights of the filter orthogonal regularization and the response orthogonal regularization are 100 and 1, respectively.

C. Where to Deploy the Channel-Wise Dropout

Here, we investigate the best stage to deploy the channel-wise dropout. We first integrate the channel-wise dropout into four different stages of the plain ResNet-50, respectively. Then, we apply the channel-wise dropout after more stages (stage $2 + 3$ or stage $1 + 2 + 3$). Specifically, the channel-wise dropout is conducted on the outputs of the last 3×3 convolution layer in the stages specific to each setting.

Table I summarizes the results on the IJB-C occlusion subset of the three different settings. As seen, the channel-wise dropout deployed only after stage 3 achieves the best results among all the models. It outperforms the baseline by an improvement up to 18.19% in terms of TAR when FAR $= 0.0001$. This setting works better than the channel-wise dropout conducted in the shallow stages. The reason behind it is that features in stage 3 have larger receptive field than those in stage 2 and stage 1, which can well characterize facial components, leading to better simulation of partial occlusions on faces. Besides, conducting channel-wise dropout in the stage 4 performs severely worse than

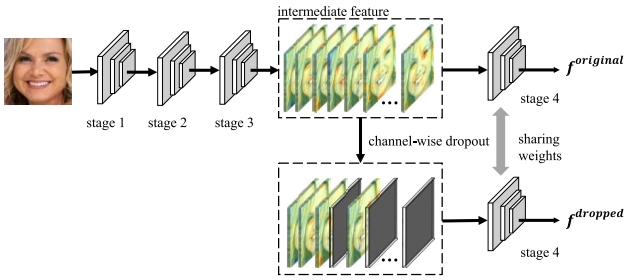

Fig. 4. The experiment designed to explore why the channel-wise dropout can improve the robustness against occlusions.

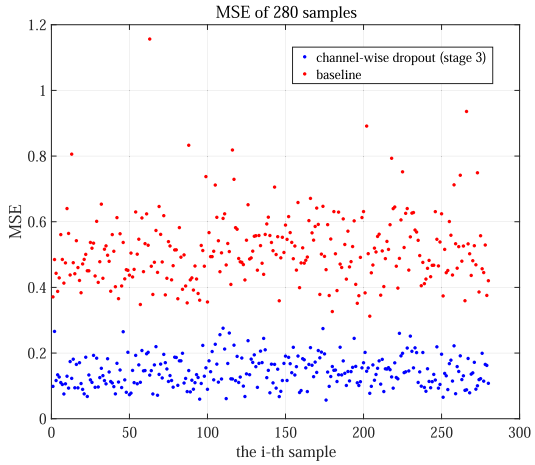

Fig. 5. The experiment designed to verify the roles of the channel-wise dropout and the neural layers behind it.

that in the stage 3, or even worse than the baseline model. We argue that employing several succeeding neural layers to compensate the damaged activations plays an important role for face recognition under occlusions. We conduct a further experiment to verify this analysis as bellow.

For a face recognition model, we compare the similarity of output features from stage 4 with and without conducting channel-wise dropout in stage 3, as shown in Fig. 4. The features similarity is evaluated by mean square error (MSE) as shown in Eq. 10:

$$MSE = \frac{1}{L}\sum_{i=1}^{L}\left(f_i^{original} - f_i^{dropped}\right)^2, \tag{10}$$

where the L denotes the length of the final feature vector and f_i denotes the i-th elements within the feature vector.

We compare the baseline mode and our model trained with the channel-wise dropout integrated in the stage 3. The results are shown in Fig. 5. As seen, our model trained with channel-wise dropout integrated in the stage 3 achieves notably lower MSE than baseline, which means the final features of our method on "occluded face" are much closer to those on clean face, leading to occlusion-robust face recognition model. The result proves that after integrating the channel-wise dropout in the stage 3, the succeeding layers in the following stage 4 play a crucial role in compensating the absent of facial information caused by occlusions. This experimental results also explain why conducting the channel-wise dropout in the stage 4 leads to poor results in Table I. If the occlusion is simulated in the last stage, no succeeding layers are enforced to learn features

TABLE II

THE ABLATION STUDY FOR EACH COMPONENT OF OUR METHOD. **CD**: THE CHANNEL-WISE DROPOUT. **SR**: THE TWO SPATIAL REGULARIZATION LOSSES. **SAM**: THE SPATIAL ATTENTION MODULE

baseline	CD	SR	SAM	@FAR=.0001	@FAR=.001
✓				39.69	88.52
✓			✓	49.42	89.65
✓	✓			57.88	90.80
✓	✓	✓		64.71	91.50
✓	✓	✓	✓	76.27	91.83

TABLE III

PERFORMANCE ON IJB-C DATASET

Methods	@FAR=.0001	@FAR=.001	@FAR=.01
DR-GAN [50]	-	73.6	88.2
PFE [51]	-	89.6	93.3
DUL [52]	-	90.2	94.2
CASIA-Net InterpretFR [35]	-	89.2	75.6
ResNet-50 InterpretFR [35]	-	93.2	95.8
Cutout [53]	74.7	94.6	97.7
DropBlock [38]	77.3	94.5	97.6
WCD [39]	87.5	93.3	96.1
LCD (ours)	89.8	95.4	97.5

TABLE IV

PERFORMANCE ON IJB-C OCCLUSION SUBSET

Methods	@FAR=.0001	@FAR=.001	@FAR=.01
DR-GAN [50]	-	66.1	82.4
CASIA-Net InterpretFR [35]	-	69.3	83.8
ResNet-50 InterpretFR [35]	-	89.8	93.4
Cutout [53]	47.9	90.5	96.0
DropBlock [38]	51.0	90.4	95.7
WCD [39]	74.6	88.5	93.1
LCD (ours)	76.3	91.8	95.4

robust to occlusions. Based on this analysis, the channel-wise dropout integrated into the stage 3 of the ResNet-50 are treated as the best setting for all the following experiments.

D. Ablation Study

We conduct ablation studies to evaluate the effectiveness of each component in our method. The results on the IJB-C occlusion subset are shown in Table II. As seen, all the three components are beneficial to the performance improvement. Specifically, when only the channel-wise dropout (CD) is conducted, the TAR when FAR $= 0.0001$ is improved by 18.19%. Moreover, by jointly training with the spatial regularization (SR), a further improvement of 6.83% is witnessed. Besides, the spatial attention module (SAM) alone can promote the baseline a lot, but its marriage with CD and SR leads to a greater improvement up to 36.58% in terms of TAR when FAR $= 0.0001$. We attribute this improvement to the complement of SAM to CD. Specifically, CD with SR achieves occlusion-invariant feature learning by implicitly simulating various occlusions in the training stage, while SAM can further attentively emphasize the non-occluded facial features of the current input face image during the inference stage. For simplicity, our method consisting of all the three components is abbreviated as LCD in the following experiments.

E. Comparisons With State-of-the-Art Methods

1) Evaluations on the IJB-C Benchmark: We firstly evaluate our LCD on the IJB-C dataset and compare it with

the state-of-the-art occlusion robust method named InterpretFR [35]. The results are shown in Table III, where the result of InterpretFR is directly quoted from [35]. As seen, our LCD outperforms InterpretFR with an improvement up to 2.2% of TAR when FAR = 0.001, which demonstrates the effectiveness of our locality-aware channel-wise dropout and spatial attention module. Besides, we compare with the image augmentation method Cutout [53], which randomly puts black box on faces to simulate occlusions. Attributed to simulating more realistic occlusion by LCD, our method significantly surpasses Cutout with 15.1% improvement in terms of TAR when FAR = 0.0001. Since DropBlock [38] follows the similar spirit by randomly zeroing out continuous activations of all channels to enhance the feature representations, we conduct further comparison with DropBlock and our method also significantly outperforms it, demonstrating the superiority of proposing LCD for occlusion synthesis. Detailed analysis between our method and DropBlock are illustrated in Section IV.

To further verify the effectiveness of our LCD for tackling face recognition under occlusions, we make more challenging experiments on the IJB-C occlusion subset, which only consists of occluded faces. Similar conclusion can be achieved that our LCD outperforms InterpretFR with an improvement up to 2.0% in terms of TAR when FAR = 0.001, demonstrating the superiority of our method again. It is worth mentioning that our LCD significantly surpasses Cutout and DropBlock with improvements up to 28.4% and 25.3% in terms of TAR when FAR = 0.0001, respectively. Our LCD can simulate more realistic occlusions than both Cutout and DropBlock, which markedly improves the robustness to occlusions for face recognition. Furthermore, comparing to the weighted channel-wise dropout (WCD), our method achieves an improvement up to 3.3% of TAR when FAR = 0.001, demonstrating that our method is more superior in dealing with occluded face recognition. Detailed analysis between our method and WCD are illustrated in Section IV.

2) Evaluations on the LFW Benchmark: Table V summarizes the accuracy results on the LFW dataset. As seen, our LCD performs better than the occlusion discarding method PDSN [12]. Furthermore, when comparing with stronger competitor CurricularFace [54] which utilizes a larger backbone of ResNet100 and much more training images, our LCD still achieves a comparable result of 99.78%. It is worth noting that the LFW is a general face recognition benchmark which mainly consists of non-occluded faces. Therefore, these results indicate that our method also generalizes well under non-occluded scenarios.

3) Evaluations on the MegaFace Benchmark: Finally, we evaluate our method on the MegaFace which is a more challenging benchmark for general scenarios. Table VI shows the rank-1 accuracies of recent methods on this benchmark. By leveraging only 3.8 million face images for training, our method with a light backbone of ResNet-50 surpasses both ArcFace [37] and CurricularFace [54] which use ResNet-100 as a backbone and are trained with 5.8 million face images. Attributed to the LCD which encourages the network to learn more comprehensive features from faces, our method not only improves the robustness to occlusions but also enhances the feature representation learning for general face recognition.

TABLE V

VERIFICATION PERFORMANCE (%) OF VARIOUS METHODS ON LFW DATASET. #IMAGE IS THE NUMBER OF IMAGES USED FOR TRAINING

Methods	#Image	Accuracy
DeepID [55]	0.1M	99.47
Deep Face [56]	4.4M	97.35
VGG Face [57]	2.6M	98.95
FaceNet [58]	200M	99.63
Baidu [59]	1.3M	99.13
Center Loss [60]	0.7M	99.28
Range Loss [61]	5M	99.52
Marginal Loss [62]	3.8M	99.48
SphereFace [63]	0.5M	99.42
SpherFace+ [64]	0.5M	99.47
PDSN [12]	0.5M	99.20
CosFace [65]	5M	99.73
ArcFace, R100 [37]	5.8M	99.83
MV-Arc-Softmax [66]	3.28M	99.78
CurricularFace, R100 [54]	5.8M	99.80
LCD (ours)	3.8M	99.78

TABLE VI

RANK-1 IDENTIFICATION ACCURACY (%) ON MEGAFACE CHALLENGE 1. #IMAGE IS THE NUMBER OF IMAGES USED FOR TRAINING

Methods	#Image	MF1
RegularFace [67]	3.1M	75.61
UniformFace [68]	3.8M	79.98
CosFace [65]	5.0M	82.72
ArcFace, R100 [37]	5.8M	81.03
PFE [51]	4.4M	78.95
CurricularFace, R100 [54]	5.8M	81.26
LCD (ours)	3.8M	83.57

VI. CONCLUSION AND FUTURE WORK

Different from previous methods which augment face images with synthesized occlusions, we propose a novel method to better simulate realistic occlusions by dropping a group of activations in intermediate features. We first employ a spatial regularization to encourage each feature channel to respond to different face regions. Then, the locality-aware channel-wise dropout is proposed to simulate occlusions by dropping out a few feature channels. In addition, we design an auxiliary spatial attention module to reweight the feature channels, which can further emphasize the contributions of non-occluded regions. By directly simulating the influence of arbitrary occlusion on intermediate features, the proposed method improves the robustness against occlusion by encouraging the neural network to capture more discriminative information from the non-occluded face regions. Extensive experiments on various benchmarks have shown that the proposed method is a practical and effective approach which outperforms state-of-the-art methods with a remarkable improvement. From the practice in this work, we can conclude that the well-known dropout strategy is not only effective for improving the generalizability but also good for achieving occlusion robustness after simple modification. Our work also shows that the simulation of occlusion in feature-level rather than image-level can be a good direction to further study.

As a possible future work, we would like to try the methods as in [69] or [70] to automatically optimize the dropout rate in Eq. 3 during the training process.

REFERENCES

[1] L. Cheng, J. Wang, Y. Gong, and Q. Hou, "Robust deep auto-encoder for occluded face recognition," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1099–1102.

[2] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan, "Robust LSTM-autoencoders for face de-occlusion in the wild," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 778–790, Feb. 2018.

[3] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[4] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, 2017.

[5] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3911–3919.

[6] X. Yuan and I. K. Park, "Face de-occlusion using 3D morphable model and generative adversarial network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10062–10071.

[7] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[8] R. Min, A. Hadid, and J.-L. Dugelay, "Improving the recognition of faces occluded by facial accessories," in *Proc. IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*, Mar. 2011, pp. 442–447.

[9] Z. Chen, T. Xu, and Z. Han, "Occluded face recognition based on the improved SVM and block weighted LBP," in *Proc. Int. Conf. Image Anal. Signal Process.*, Oct. 2011, pp. 118–122.

[10] S. Park, H. Lee, J.-H. Yoo, G. Kim, and S. Kim, "Partially occluded facial image retrieval based on a similarity measurement," *Math. Problems Eng.*, vol. 2015, pp. 1–11, Jan. 2015.

[11] W. Wan and J. Chen, "Occlusion robust face recognition based on mask learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3795–3799.

[12] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential Siamese network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 773–782.

[13] F. Ding, P. Peng, Y. Huang, M. Geng, and Y. Tian, "Masked face recognition with latent part detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2281–2289.

[14] J.-J. Lv, X.-H. Shao, J.-S. Huang, X.-D. Zhou, and X. Zhou, "Data augmentation for face recognition," *Neurocomputing*, vol. 230, pp. 184–196, Mar. 2017.

[15] E. Osherov and M. Lindenbaum, "Increasing CNN robustness to occlusions by reducing filter support," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 550–561.

[16] D. S. Trigueros, L. Meng, and M. Hartnett, "Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss," *Image Vis. Comput.*, vol. 79, pp. 99–108, Nov. 2018.

[17] C. Shao *et al.*, "Biased feature learning for occlusion invariant face recognition," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2020, pp. 666–672.

[18] D. Novotny, D. Larlus, and A. Vedaldi, "AnchorNet: A weakly supervised network to learn geometry-sensitive features for semantic matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5277–5286.

[19] J.-S. Park, Y. H. Oh, S. C. Ahn, and S.-W. Lee, "Glasses removal from facial image using recursive error compensation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 805–811, May 2005.

[20] S. Fidler, D. Skocaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 337–350, Mar. 2006.

[21] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis," *J. ACM*, vol. 58, no. 3, pp. 1–37, May 2011.

[22] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[23] F. Zhang, J. Yang, Y. Tai, and J. Tang, "Double nuclear norm-based matrix decomposition for occluded image recovery and background modeling," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1956–1966, Jun. 2015.

[24] Y. Qian, W. Deng, and J. Hu, "Unsupervised face normalization with extreme pose and expression in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9851–9858.

[25] H. J. Oh, K. M. Lee, and S. U. Lee, "Occlusion invariant face recognition using selective local non-negative matrix factorization basis images," *Image Vis. Comput.*, vol. 26, no. 11, pp. 1515–1523, 2008.

[26] J. Hu, J. Lu, and Y.-P. Tan, "Robust partial face recognition using instance-to-class distance," in *Proc. Vis. Commun. Image Process. (VCIP)*, Nov. 2013, pp. 1–6.

[27] L. He, H. Li, Q. Zhang, and Z. Sun, "Dynamic feature matching for partial face recognition," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 791–802, Feb. 2019.

[28] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[29] J. Dong, H. Zheng, and L. Lian, "Low-rank laplacian-uniform mixed model for robust face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11897–11906.

[30] M. Yang, L. Zhang, S. C. K. Shiu, and D. Zhang, "Gabor feature based robust representation and classification for face recognition with Gabor occlusion dictionary," *Pattern Recognit.*, vol. 46, no. 7, pp. 1865–1878, Jul. 2013.

[31] K. Jia, T.-H. Chan, and Y. Ma, "Robust and practical face recognition via structured sparsity," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 331–344.

[32] M. Yang, X. Wang, G. Zeng, and L. Shen, "Joint and collaborative representation with local adaptive convolution feature for face recognition with single sample per person," *Pattern Recognit.*, vol. 66, pp. 117–128, Jun. 2017.

[33] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2285–2294.

[34] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.

[35] B. Yin, L. Tran, H. Li, X. Shen, and X. Liu, "Towards interpretable face recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9348–9357.

[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[37] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[38] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "DropBlock: A regularization method for convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10727–10737.

[39] S. Hou and Z. Wang, "Weighted channel dropout for regularization of deep convolutional neural network," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8425–8432.

[40] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 87–102.

[41] B. Maze *et al.*, "IARPA Janus benchmark—C: Face dataset and protocol," in *Proc. Int. Conf. Biometrics (ICB)*, Feb. 2018, pp. 158–165.

[42] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The MegaFace benchmark: 1 million faces for recognition at scale," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4873–4882.

[43] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.

[44] B. F. Klare *et al.*, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1931–1939.

[45] C. Whitelam *et al.*, "IARPA Janus benchmark-B face dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 592–600.

[46] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 343–347.

[47] Z. He, M. Kan, J. Zhang, X. Chen, and S. Shan, "A fully end-to-end cascaded CNN for facial landmark detection," in *Proc. 12th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*, May 2017, pp. 200–207.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 630–645.

[49] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. USENIX Symp. Oper. Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.
[50] L. Tran, X. Yin, and X. Liu, "Representation learning by rotating your faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 3007–3021, Dec. 2019.
[51] Y. Shi and A. Jain, "Probabilistic face embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6902–6911.
[52] J. Chang, Z. Lan, C. Cheng, and Y. Wei, "Data uncertainty learning in face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5710–5719.
[53] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.
[54] Y. Huang *et al.*, "CurricularFace: Adaptive curriculum learning loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5901–5910.
[55] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1988–1996.
[56] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1701–1708.
[57] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Assoc. (BMVC)*, 2015, pp. 1–12.
[58] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[59] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang, "Targeting ultimate accuracy: Face recognition via deep embedding," 2015, *arXiv:1506.07310*.
[60] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 499–515.
[61] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5409–5418.
[62] J. Deng, Y. Zhou, and S. Zafeiriou, "Marginal loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 60–68.
[63] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 212–220.
[64] W. Liu *et al.*, "Learning towards minimum hyperspherical energy," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 6222–6233.
[65] H. Wang *et al.*, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5265–5274.
[66] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei, "Mis-classified vector guided softmax loss for face recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12241–12248.
[67] K. Zhao, J. Xu, and M.-M. Cheng, "RegularFace: Deep face recognition via exclusive regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1136–1144.
[68] Y. Duan, J. Lu, and J. Zhou, "UniformFace: Learning deep equidistributed representation for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3415–3424.
[69] S.-I. Maeda, "A Bayesian encourages dropout," 2014, *arXiv:1412.7003*.
[70] Y. Liu, W. Dong, L. Zhang, D. Gong, and Q. Shi, "Variational Bayesian dropout with a hierarchical prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7124–7133.

Mingjie He received the M.S. degree from the University of Science and Technology of China, Hefei, China, in 2014. He is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences and an Engineer with the Institute of Computing Technology, Chinese Academy of Sciences (CAS). His research interests cover computer vision and machine learning.

Jie Zhang (Member, IEEE) received the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China. He is currently an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS). His research interests cover computer vision, pattern recognition, machine learning, particularly face recognition, image segmentation, weakly/semi-supervised learning, and domain generalization.

Shiguang Shan (Fellow, IEEE) received the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2004. He has been a Full Professor of this institute since 2010 and is currently the Deputy Director of the CAS Key Laboratory of Intelligent Information Processing. He has published more than 300 papers, with totally more than 25,000 Google Scholar citations. His research interests cover computer vision, pattern recognition, and machine learning. He was a recipient of China's State Natural Science Award in 2015, and China's State S&T Progress Award in 2005 for his research work. He has served as the Area Chair (or Senior PC) for many international conferences, including ICCV11, ICPR12/14/20, ACCV12/16/18, FG13/18, ICASSP14, BTAS18, AAAI20/21, IJCAI21, and CVPR19/20/21. He was/is an Associate Editor of several journals, including IEEE TRANSACTIONS ON IMAGE PROCESSING, *Neurocomputing*, *CVIU*, and *PRL*.

Xiao Liu received the Ph.D. degree in computer science from Zhejiang University in 2015. He worked at Baidu from 2015 to 2019. He is currently a Researcher with Tomorrow Advancing Life Education Group (TAL). His research interests include the applied AI, such as intelligent multimedia processing, computer vision, and learning systems. His research results have expounded in more than 40 publications at journals and conferences, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, CVPR, ICCV, ECCV, AAAI, and MM. As a Key Team Member, he achieved the best performance in various competitions, such as the ActivityNet challenges, NTIRE super resolution challenge, and EmotioNet facial expression recognition challenge.

Zhongqin Wu received the master's degree in computer science from Fudan University, China. He worked with Baidu and lead the Department of Computer Vision and the Augmented Reality Laboratory. He is currently a Scientist with Tomorrow Advancing Life Education Group (TAL) and is in charge of the AI laboratory.

Xilin Chen (Fellow, IEEE) is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS). He has authored one book and more than 200 papers in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. He is a fellow of ACM, IAPR, and CCF. He was a recipient of several awards, including China's State Natural Science Award in 2015, and China's State S&T Progress Award in 2000, 2003, 2005, and 2012 for his research work. He served as an Organizing Committee member for many conferences, including the General Co-Chair for FG13/FG18, the Local Chair for ICME07, ACM MM09, and ICIP17, and the Finance Chair for ISCAS13. He is/was the Area Chair of CVPR 2017/2019/2020 and ICCV 2019. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, and *Journal of Visual Communication and Image Representation*, a Leading Editor of the *Journal of Computer Science and Technology*, and an Associate Editor-in-Chief of the *Chinese Journal of Computers*, and *Chinese Journal of Pattern Recognition and Artificial Intelligence*.