Received: 14 August 2020 | Revised: 9 March 2021 | Accepted: 10 March 2021

IET Biometrics

DOI: 10.1049/bme2.12036

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Occlusion-invariant face recognition using simultaneous segmentation

Dan Zeng[1,2] | Raymond Veldhuis[2] | Luuk Spreeuwers[2] | Richard Arendsen[3]

[1]Southern University of Science and Technology, China

[2]University of Twente, The Netherlands

[3]20face BV, The Netherlands

Correspondence

Dan Zeng, Southern University of Science and Technology, China.
Email: zengd@sustech.edu.cn

Abstract

When using convolutional neural network (CNN) models to extract features of an occluded face, the occluded part will inevitably be embedded into the representation just as with other facial regions. Existing methods deal with occluded face recognition either by augmenting the training dataset with synthesized occluded faces or by segmenting occlusions first and subsequently recognize the face based on unoccluded facial regions. Instead, simultaneous occlusion segmentation and face recognition is developed to make the most of these correlated two tasks. This is inspired by the phenomenon that features corrupted by occlusion are traceable within a CNN trained to segment occluded parts in face images. Specifically, a simultaneous occlusion invariant deep network (SOIDN) is proposed that contains simultaneously operating face recognition and occlusion segmentation networks coupled with an occlusion mask adaptor module as their bridge to learn occlusion invariant features. The training of SOIDN is jointly supervised by classification and segmentation losses aiming to obtain (1) occlusion invariant features, (2) occlusion segmentation, and (3) an occlusion feature mask that weighs the reliability of features. Experiments on synthesized occluded dataset (e.g. LFW-occ) and real occluded face dataset (e.g. AR) demonstrate that SOIDN outperforms state of the art methods for face verification and identification.

1 | INTRODUCTION

Face recognition (FR), as one of the most ubiquitous biometric technologies, is widely used in such things as governmental, commercial, computer security, and voter verification applications. Deep convolutional neural networks (CNNs) have pushed the frontier of many computer vision applications, including unconstrained FR [1, 2]. FR systems tend to perform worse when encountering challenges such as large-pose variations, different facial expressions, heavy makeup, varying illumination, and occlusion. In particular, they suffer from significant accuracy degradation when challenged with occluded facial images [3, 4]. When using CNN models to extract features of an occluded face, the occluded part is inevitably embedded into the representation in latent space, just as with other facial regions [5]. Facial occlusion, such as scarves, glasses, face masks, and hats, can be anywhere and of any size or shape in a face image. As illustrated in Figure 1, facial appearance changes substantially with occlusion. Therefore, occluded FR continues to be considered one of the most intractable problems in the field.

Two essential factors related to occlusion challenges are (1) where is the occlusion (location) and (2) what is the occlusion (content) [7, 4]. If a dataset presents sufficient examples of occlusion location and content as they may occur in real-world applications [8], it goes without saying that training a CNN model with such occluded faces can render occlusion-robust features. However, no such dataset exists. Alternatively, some approaches augment the training dataset with synthesized occluded faces to ensure that features extracted more locally and equally can better handle occlusions [7]. It is worth mentioning that occlusion segmentation, that is, locating the occlusion region in a face image, has not been applied in these methods. Some methods explicitly detect/segment occlusion first and recognize the face sequentially based on unoccluded facial regions [9–12]. However, occlusion segmentation (OS) and FR constitute a sequential pipeline, resulting in FR that is dependent on OS

FIGURE 1 Examples of real-world occlusion from the MAsked FAces dataset [6]

results. As a consequence, the side effect of imperfect segmentation is unavoidably impaired FR [4].

In contrast, this paper develops simultaneous OR and FR for better information utilization (see Figure 2). We observe the phenomenon that features corrupted by occlusion are traceable within a CNN trained for segmenting occluded parts in face images [13] (see Figure 3). Specifically, pixel-wise occlusion is traced in feature maps of convolutional layers, thus ensuring that the occlusion location can be preserved through the segmentation CNN. This inspired us to leverage the deep occlusion response of OS to clean latent representations from occlusion artefacts. In a nutshell, we propose a simultaneous occlusion-invariant deep network (SOIDN) containing simultaneously operating FR and OR networks involving an occlusion mask adaptor (OMA) module as a bridge between their top convolutional layers to learn occlusion feature masks from the top convolutional layers of occlusion segmentation. Specifically, the OMA module intends to learn the correspondence between convolutional features of OS and the occlusion mask so that the channel-wise convolutional features of OS are correctly matched with their counterparts in FR. The proposed SOIDN enjoys several advantages: (1) two aspects of occlusion, location and content, are explicitly considered by the OS network and FR network, respectively, coherently combined, and optimized within a simultaneous architecture; (2) OS and FR can work together to obtain an occlusion-free face representation. If face representation extracted from the FR network is already unaffected by the actual occlusion, the occlusion feature mask plays a less important role in purifying representation. On the other hand, if face representation instead is affected by the occlusion and deteriorates discriminatively, we can exclude corrupted features by allowing for occlusion feature masks. To sum up, the proposed method is capable of recognizing faces under severe occlusion in a simplified yet well-motivated way.

The training of the proposed SOIDN is jointly supervised by minimizing two losses, namely, classification and segmentation. Classification loss ensures that the occlusion masks are

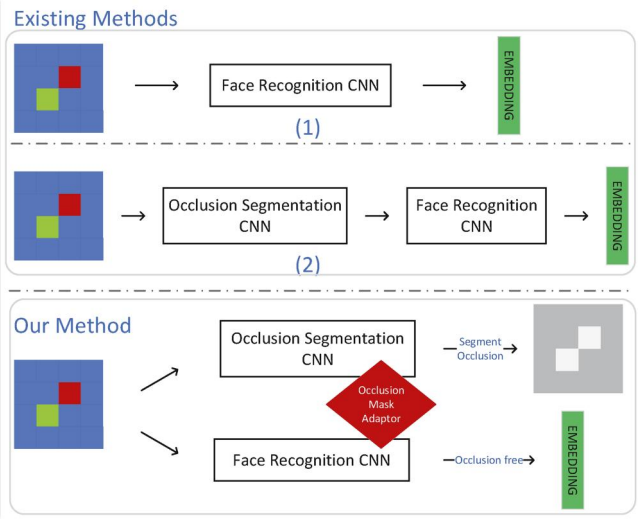

FIGURE 2 Existing methods either (1) rely on a deep CNN to extract occlusion-robust features (mainly using synthesized occluded faces for data augmentation) or (2) utilize occlusion segmentation and face recognition individually and sequentially to handle occlusion challenges. In contrast, our proposed method coherently combines both approaches and optimizes them within a simultaneous architecture to learn occlusion-invariant embedding features. CNN, convolutional neural network

optimized so that corrupted features are filtered out, and features that are not discriminative in terms of identity are penalized. Segmentation loss ensures that output maps segment face and non-face parts. With the supervision of two losses, we aim to obtain (1) occlusion-free features to preserve discriminability for occluded FR, (2) OR output involving facial parts that impair FR accuracy, and (3) an occlusion mask that weighs the reliability of features to minimize the considerable intraclass variations caused by occlusions. In the training phase, occlusion-free and synthetically occluded faces are used as the training dataset. In the testing phase, the proposed SOIDN can be applied to face images regardless of whether the occlusion is present. It explicitly masks out the occlusion

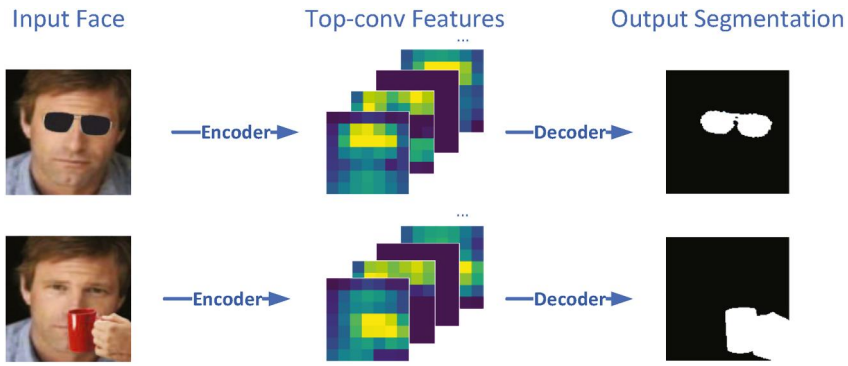

FIGURE 3 The deep response (top-convolutional features) generated by the occlusion segmentation network shows the trace of features corrupted by occlusion (sunglasses and cup). For simplicity, only the first four channels in top-convolutional features are shown

of the face and obtains occlusion-free features at the same time.

The three main contributions of the proposed method are as follows:

- We propose a novel SOIDN to coherently combine the FR and OR networks and optimize them within a simultaneous architecture to learn occlusion-invariant features.
- We design the occlusion mask adaptor (OMA) as a bridge in SOIDN, motivated by the phenomenon that features corrupted by occlusion are traceable within an OS network.
- We implement the proposed SOIDN with a combined loss function, including classification and segmentation losses, and achieve good performance.

The rest of this paper is organized as follows. Related work is shown in Section 2. The proposed method is described in Section 3. Experimental results are shown in Section 4. The conclusion is given in Section 5.

2 | RELATED WORK

Approaches to recognize faces under occlusions can be broadly classified into three categories: (i) occlusion robust feature extraction, (ii) occlusion-recovery-based FR, and (iii) occlusion-aware FR. In this section, we first briefly review the related work on occluded face recognition (OFR) before the emergence of deep-learning techniques. Then we elaborate on existing deep-learning methods that cope with FR under occlusion challenges and highlight the differences of the proposed method.

2.1 | Approaches not using deep learning for occluded FR

The first category, occlusion robust feature extraction, extracts handcrafted features or learns features from explicitly defined facial regions. Reference [14] introduces Kullback–Leibler divergence to measure the distance of local Gabor binary patterns descriptors [15] of the local region of test images and that of the unoccluded region of reference faces. A robust matching metric [16] is presented to match the difference of Gaussian filter descriptor of a facial part against its spatial neighbourhood in the other faces and select the minimal distance for FR. A random sampling patch-based method [17] is presented to treat all face patches equally and randomly select the patch to train the classifier. Subspace learning methods such as principal component analysis (PCA) and variants [18, 19] are developed to handle occlusion challenges. Independent component analysis [20] is used to find locally salient information from important facial parts. Statistical learning methods such as local Gaussian kernel-based features [21] or a simple Gaussian model [22] for feature probability estimation address occlusion occurrence as a probability problem. Mclaughlin et al. [23] propose the largest matching areas (LMAs) at each point on the face by assuming that the occluded test image region can be modelled by an unseen-data likelihood with a low posterior probability.

The second category, occlusion recovery-based FR, recovers a clean face from the occluded one for recognition. Sparse representation classifiers (SRCs) [24] and variants retain popularity and success in coping with occlusion challenges. The main idea of SRCs is to present a face using a linear combination of training samples and sparse constraint terms accounting for occlusions. SRC variants are developed by various aspects such as combining prior knowledge of pixel error distribution [25], using Gabor features instead of pixel values [26], applying downsampled SRCs [27] to locate occlusion at low computing complexity, importing mutual-incoherence regularization terms into the SRC scheme [28], exploiting the sparse error component with robust PCA [29], and introducing modular weight-based SRC [30]. Recently, Ref. [31] proposes a joint and collaborative representation with a local adaptive convolution feature containing local high-level features from local regular regions. Reference [32] proposes a hierarchical sparse and low-rank regression model using features based on image gradient direction. Robust point set matching (RPSM) [33] considers both geometric distribution consistency and textural similarity for simultaneous matching. Moreover, a constraint on the affine transformation is applied to prevent unrealistic face warping. However, these methods will fail if facial key points are unavailable because of occlusions, as facial alignment is required during preprocessing [31–33]. Moreover, the computation complexity is high, which slows the recognition process.

The third category, occlusion-aware FR, usually discards the occlusion part and performs FR based on the visible face

parts only. Ref. [9, 30] divide a face into multiple non-overlapping regions and train a support vector machine classifier to identify the occluded area. Reference [11] introduces a selective local non-negative matrix factorization (NMF) method to select features corresponding to occlusion-free regions for recognition. Another work [12] extends NMF to include occlusion estimation adaptively according to reconstruction errors. Finally, low-dimensional representations are learnt to ensure that features of the same class are close to the corresponding class centre.

2.2 | Deep-learning approaches for occluded FR

Face representation obtained by deep CNNs is vastly superior to traditional learning methods in the discriminative power that has pushed the frontier of deep FR [2]. Some methods [34, 35] take advantage of data augmentation to generate sufficient synthetically occluded faces for training a deep network. Lv et al. [34] synthesize occluded faces with various hairstyles and glasses to augment the training dataset. Specifically, 87 hairstyle templates with various bangs and 100 glasses templates are collected for augmentation so that the trained CNN model is robust to various hairstyles and glasses. In paper [35], instead of using synthetic occluded faces directly, the authors identify the importance of face regions based on their occlusion sensitivity and then train a CNN with identified facial regions covered to reduce model reliance on these regions. Specifically, training face images are augmented with occlusions located in high-effect regions (central part of the face) more frequently than in low-effect regions (outer parts of the face). In this way, the model is forced to learn more discriminative features from the outer part of the face, which results in less accuracy degradation when the central part of the face is occluded. Cen et al. [36] propose a deep dictionary representation-based classification scheme to alleviate the occlusion effect in FR, where the dictionary is used to code the deep convolutional features linearly.

Deep-learning techniques are also used for occluded face reconstruction. Reference [37] extends a stacked sparse denoising autoencoder to a double channel for facial occlusion removal. Zhao et al. [4] combine the long short-term memory (LSTM) and autoencoder architectures to address the face de-occlusion problem. The proposed robust LSTM-autoencoders consist of two LSTM components. One spatial LSTM network encodes face patches of different scales sequentially for robust occlusion encoding, and the other dual-channel LSTM network is used to decode the representation to reconstruct the face and detect the occlusion. In addition, the adversarial CNNs are introduced to enhance the discriminative information in the recovered faces. The generative adversarial network (GAN) [38] and variants retain popularity and succeed in synthesizing or generating new samples. Occlusion-aware GAN [39] is proposed to identify the corrupted image region with the associated corrupted region recovered by utilizing a GAN pretrained on occlusion-free faces. Reference [40] employs GAN for eyes-to-face synthesis with only eyes visible. The eyeglasses removal GAN [41] is proposed for eyeglasses removal in the wild via an unsupervised manner and is capable of rendering a competitive removal quality in terms of realism and diversity. In paper [42], the identity diversity GAN combines the CNN-based recognizer and GAN-based recognition to inpaint realism and identity-preserving faces with the recognizer treated as the third player to compete with the generator.

Deep-learning techniques are sometimes used to detect the occlusion and represent a face by excluding occlusion parts [5, 43, 44]. To cope with OFR with limited training samples, Reference [45] proposes a structural element feature extraction method to capture the local and contextual information inspired by human optic nerve characteristics for FR. In addition, an adaptive fusion method is proposed to use multiple features consisting of a structural element feature and a connected-granule labelling feature. To exploit the inherent multiscale features of a single CNN, FANet [46] introduced an agglomeration connection module to enhance context-aware features and augment low-level feature maps with a hierarchical structure so that it can cope with scale variations in face detection effectively. Refererence [43] predicts the occlusion probability of the predefined face components by training a multitask CNN. In paper [5], the authors propose adding the MaskNet module to the middle layer of CNN models, aiming to learn image features with high fidelity and ignore those distortions caused by occlusions. The MaskNet, a shallow convolutional network, assigns lower weights to hidden units activated by occluded facial areas. Song et al. [44] propose a pairwise differential siamese network (PDSN) to estimate a mask dictionary. They first detect the occlusion location in image space and then rely on mask dictionary learning (one is a clean face, and the other is an occluded face) to discard the corrupted features due to occlusion. However, this method performs OS and FR sequentially. Differently, the proposed SOIDN coherently combines both and optimizes them within a simultaneous architecture to learn occlusion-invariant features. To the best or our knowledge, this work is the first to carry out OS and FR simultaneously to make the most of the correlation relationship between them.

3 | PROPOSED APPROACH

3.1 | Problem statement

To address the OFR problem, extracting occlusion-invariant features is key. Generally, such features can be obtained by excluding occlusion regions in a given face image or distinguishing facial features from corrupted features in a feature representation. The former usually produces features of variable length due to varying occlusion shape and relies on feature comparison learning to search for the semantic correspondence between the partial face in the entire gallery face. The latter is capable of generating features with fixed length under different occlusions, and similarity among faces can be

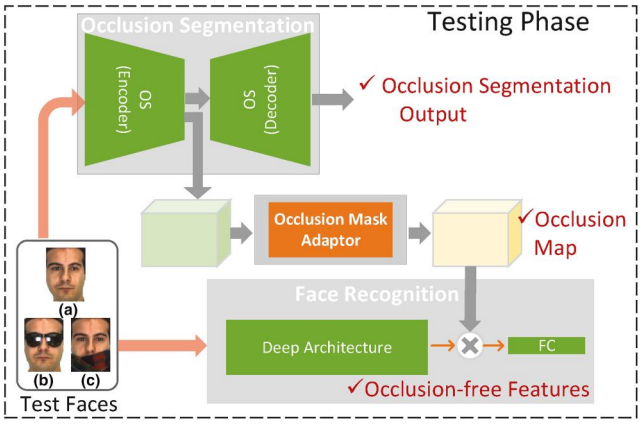

FIGURE 4 An overview of the proposed framework. It consists of an occlusion segmentation network $g(\cdot)$ and a face recognition network $f(\cdot)$ in parallel with the occlusion mask adaptor module $M(\cdot)$ as a bridge. For test faces, (a) indicates occlusion-free face image, and (b) and (c) indicate the images occluded by sunglasses and scarf, respectively

computed using distance metrics, that is, Euclidean or cosine, through the occlusion-invariant feature-embedding space. The overview of the proposed framework, which is within the latter group, is shown in Figure 4.

The formula definition of the proposed SOIDN is as follows: $x \in \mathbb{R}^{w \times h \times c}$ represents an input face image either with occlusion or occlusion-free. The final occlusion-invariant feature vector v with respect to input face image x can be denoted as

$$v = h(M(g(x)) * f(x)) \tag{1}$$

where $f(x) \in \mathbb{R}^{W \times H \times C}$ and $g(x) \in \mathbb{R}^{W \times H \times C'}$ represent top convolutional features from the FR network and OS network, respectively. Here $f(x)$ and $g(x)$ are required to have the same width and height. The OMA module $M(\cdot)$ takes OS features as inputs to generate the occlusion mask $M(g(x))$. We multiply each weight in the occlusion mask with FR features $f(x)$ at the same spatial location to mask out the corrupted features. In the FR CNN model, we often use the output of the final fully connected layers just before the classification layer as the face representation. Here $h(\cdot)$ represents the operation after the top-convolutional layer before the classification layer of the FR network. Finally, we can obtain occlusion-invariant feature representation v.

3.2 | Simultaneous occlusion-invariant deep network

We propose a novel SOIDN to simultaneously perform OS and FR for occlusion-invariant features extraction. The structure of the proposed method is shown in Figure 5. The deep architecture of FR can be arbitrary. Specifically, we adopt the widely used VGG16 [47] as an example of the FR network to illustrate how our method improves the embedded features for OFR. The OS network is responsible for detecting occlusion pixel-wise in a face image. For simplicity, we directly adopt FCN-8s [48] as an example for segmentation that can be substituted with other advanced semantic segmentation architectures. The OMA module is optimized to learn the correspondence between encoding OS features and the occlusion mask that can distinguish the corrupted elements in FR features. Occlusion mask generation encourages purified features (excluding corruption) to be as close as those extracted from the same identity yet as an occlusion-free face image constrained by the proposed classification loss. The FR network, if used alone, may extract features not significantly affected by occlusion. Luckily, with the presence of the OMA module and OS network, the FR network is capable of extracting occlusion-invariant features and functions well under occlusion.

To this end, we propose to learn occlusion-invariant features by minimizing a combination of two losses:

$$L = \sum_i l_{cls}(\theta; F(x_i), y_i^{cls}) + \lambda l_{seg}(\theta; G(x_i), y_i^{seg}) \tag{2}$$

The first classification loss l_{cls} ensures that the features after applying the occlusion mask are extracted discriminate and occlusion-invariant. The second segmentation loss l_{seg} guarantees segmentation of the occlusion part precisely in the image space. We use $F(\cdot)$ and $G(\cdot)$ to represent the FR and OS deep model. The coefficient of λ is used to balance these two tasks. The details are expanded in the following.

Classification loss l_{cls}: the FR network is trained to classify the identity of a face image. In addition, the OMA module is incorporated to ensure that the corrupted features are masked out and that only occlusion-free features qualify for FR. Lastly, we use softmax loss for the classification problem with the identity information being the supervision signal:

$$l_{cls}(\theta; F(x_i), y_i^{cls}) = -y_i^{cls} log(F(x_i)) \tag{3}$$

where y_i^{cls}, a one-shot vector of the ith face image x, is the target probability distribution. $F(x_i)$ is derived by forwarding the occlusion-free features in Equation (1) to the final fully connected layer (including the softmax operation), which is denoted as

$$F(x_i) = \frac{exp(v_i W_{y_i})}{\sum_{k=1}^{n} exp(v_i W_k)} \tag{4}$$

The last layer of the FR network is a softmax layer that outputs a probability distribution over the n identity classes y^{cls}, and the weights W are learnt.

OS loss l_{seg}: we use the supervision signal of segmentation to ensure that the OS network distinguishes the facial region from occlusion in the image space. In that case, we can trace the features corrupted by occlusion within the OS network and generate the occlusion mask in the end.

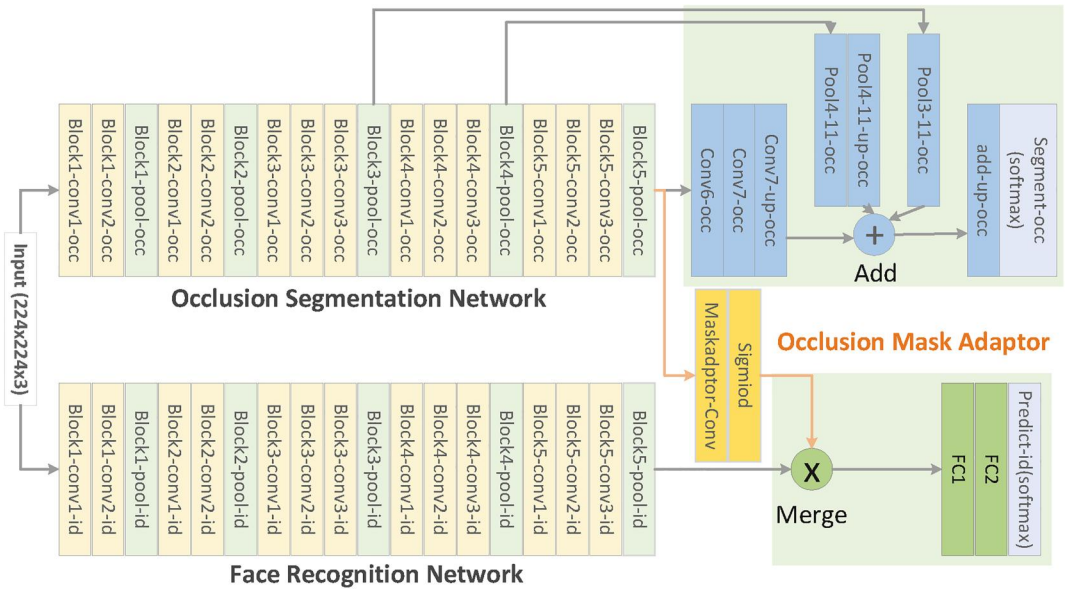

FIGURE 5 Structure of the proposed simultaneous occlusion-invariant deep network for occlusion-invariant feature extraction. VGG16 is taken as an example of the face recognition network

The most commonly used segmentation loss is a pixel-wise cross-entropy loss, which examines each pixel individually and compares the predicted class with the one-hot target segmentation. Pixel-loss is calculated as the log, which adds up over two classes, namely, *clean facial region* and *occluded facial region*, derived as

$$l_{seg}(\theta; G(x_i), y_i^{seg}) = -\sum_{clean} y_i^{seg} log(\hat{y}_i) - \sum_{occ} y_i^{seg} log(\hat{y}_i) \quad (5)$$

where the loss over the clean and occluded facial regions are summed to constitute the OS loss. We use $\hat{y}_i$ to represent the predicted pixel-wise class label. This scoring is repeated over the pixels and then averaged.

3.2.1 | Occlusion mask generation

One feasible way to generate the occlusion mask is to take pairwise images, including a clean face image and a corresponding occluded face image of the same identity as the input of a CNN to determine the differences between their features from which to learn the occlusion mask by using dictionary learning [44]. By contrast, we discover that features corrupted by occlusion are traceable within a CNN trained for OS. In view of this, we take advantage of traceable corrupted features to facilitate the occlusion mask generation. Furthermore, the requirements for pairwise face images are removed.

We use the OMA network to address the occlusion mask generation problem. Figure 6 shows the detailed architecture of the OMA network, which takes deep feature maps of $W \times H \times C$ as input and predicts an occlusion map of the same size. Herein, the sigmoid function is imposed to enforce

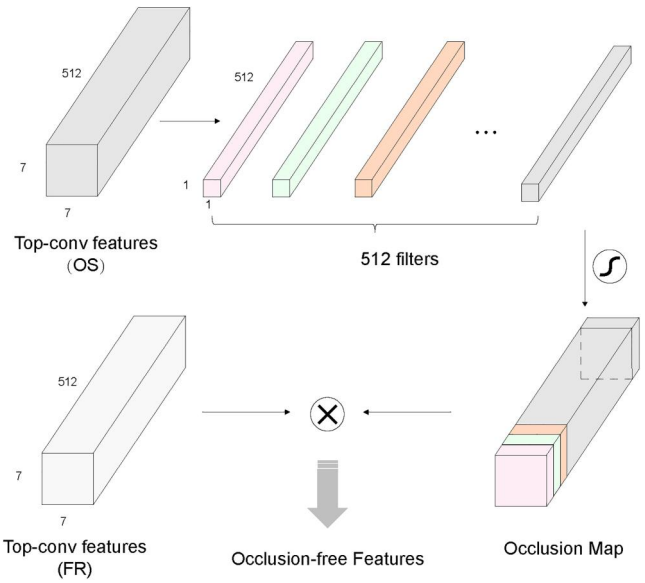

FIGURE 6 Up: the occlusion mask adaptor network consisting of 1×1 convolution layer and sigmoid. Down: the process of calculating occlusion-free features

the output values of the occlusion mask into the interval [0, 1]. The generated occlusion mask serves as an indicator for corrupted features, as it weighs the importance of features in terms of spatial locations and channels. As a result, the occlusion map ensures that the channel-wise OS convolutional features are correctly matched with the counterpart the FR features and that occlusion-free features are extracted. We continue to pass these features on to two fully connected layers to extract *occlusion-invariant features* for OFR, as indicated in Equation (1).

3.2.2 | Occlusion segmentation

The proposed SOIDN is capable of coping with the occlusion problem owing to use of the OS network. Specifically, FR and OS are simultaneously coupled with the OMA module as their bridge to learn occlusion-invariant features. Put simply, the segmentation output is not only affected by the OS loss but also implicitly adjusted by the classification loss. In view of this, the OS results can be considered a predictor of the robustness of the FR network in terms of occlusion. If the occlusion is accurately segmented from a face region, we can conclude that the FR network is sensitive to the occlusion because all the corrupted features that are masked out originate from the occlusion region in an image. In other words, the FR network performs better with the use of the occlusion mask, which also means the FR network is not very robust with the occlusion and vice versa. Furthermore, if an FR network is trained with sufficient occluded faces and can generate occlusion-invariant features independently, we find that the OS network fails to segment the occlusion accurately. This result is contrary to our expectations. The reason for this is that the training of proposed SOIDN is jointly supervised by minimizing a combination of classification and segmentation losses, with the former acting as the dominant loss.

4 | EXPERIMENTS

In this section, we first verify the effectiveness of the proposed SOIDN on synthesized occluded face dataset—e.g. Labeled Faces in the Wild (LFW)-occ—and real occluded face dataset (e.g. AR). Then we evaluate the performance of the proposed SOIDN and compare it with state-of-the-art methods.

4.1 | Datasets

The training dataset is composed of CASIA-WebFace [49] and synthetic occluded CASIA-WebFace. The occluded faces are randomly synthesized from an occlusion-free face using occlusion templates. In real-world applications, not all types of occlusions have the same probability of occurring; for example, a scarf-and-sunglasses often has a higher probability of occurrence than other occlusions. Hence, we collect occlusion templates to include typical occlusion examples. Figure 7 lists all occlusion templates used in the paper. Samples of training faces and corresponding occlusion labels are shown in Figure 8. To be sure the occluded faces do not dominate within-class variation, only subjects having more than 50 images are chosen for training, which results in 3459 involved individuals.

The LFW dataset [50] is a standard face verification benchmark dataset under unconstrained conditions. We synthesize the occluded LFW dataset to simulate real occlusions, namely LFW-occ. We apply the standard protocol of the LFW dataset to the LFW-occ and report the mean accuracy and equal error rate on the 6000 testing image pairs. Every image pair of the LFW-occ comprises a left face image from the LFW and the right image, which is synthesized to the occluded image in terms of the specific occlusion template. Examples of face pairs regarding the sunglasses occlusion for evaluation are shown in Figure 9(a).

FIGURE 7 The occlusion templates used to synthesize occluded faces, with the first two rows for eye-region based occlusions and the last two rows for occlusions around mouth and nose regions

The AR face database [51] is one of the very few benchmark datasets that contain real occlusions (see Figure 9(b)). It consists of over 4000 faces of 126 individuals—70 men and 56 women—taken in two sessions over a two-week interval. There are 13 images per individual in each session, and these images differ in terms of facial expression, illumination, and partial occlusion, with sunglasses and scarves becoming involved. Indexes 8 and 11 of each session indicate that the person is wearing sunglasses or a scarf, respectively. Indexes 9–10 and 11–12 combine the sunglasses or the scarf with illuminations, respectively.

4.2 | Experimental settings

In our experiments, all face images are preprocessed through face detection and face landmarking by using the standard multitask cascaded convolutional networks [52]. After applying affine transformation based on four landmarks, that is, left eye centre, right eye centre, nose tip, and mouth centre, the face images are aligned and resized to 224 × 224.

4.2.1 | Training phase

We employed the refined VGG16 model [47] as the FR network as well as the encoder part of the OS network. In practice, any advanced network can be alternatively used in the

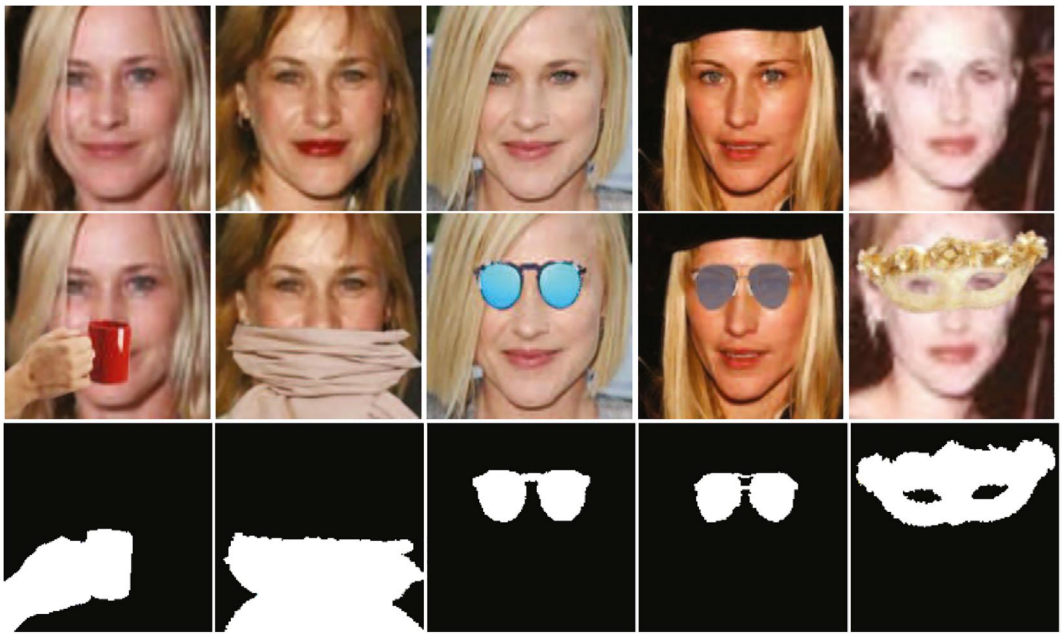

FIGURE 8 Training examples of occlusion-free faces (first row), synthesized occluded faces (second row), and occlusion labels for occlusion segmentation (third row). The occlusion labels for occlusion-free faces are omitted for simplicity

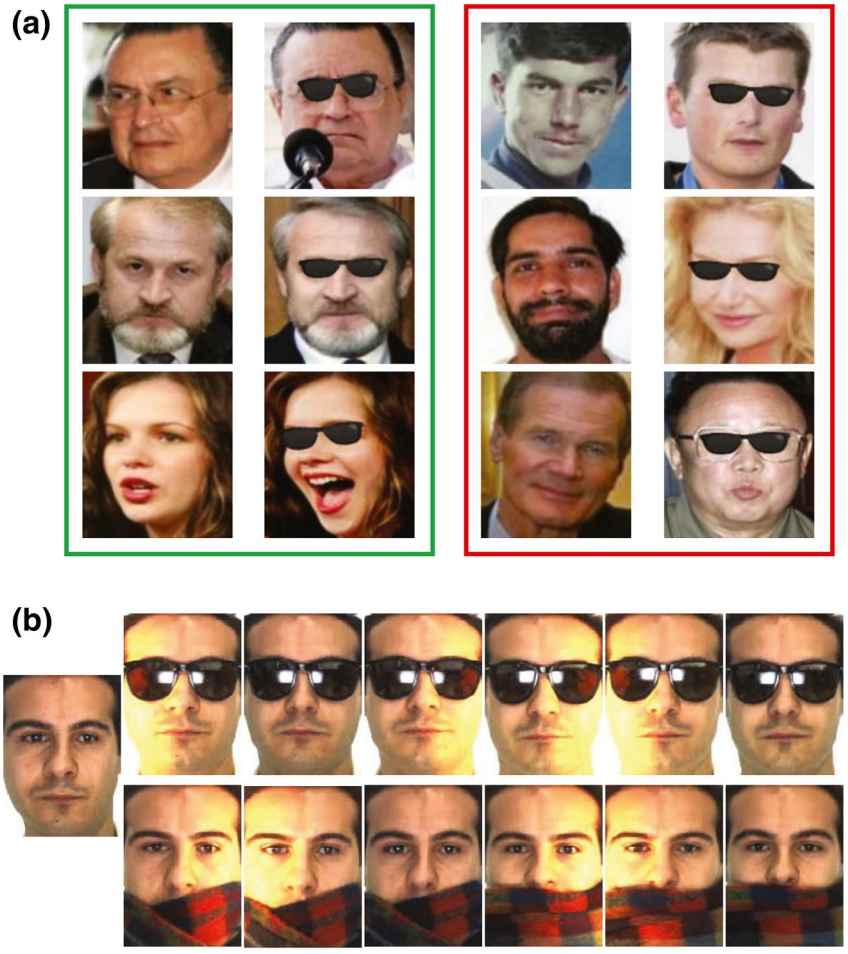

FIGURE 9 Samples are shown in (a) Labeled Faces in the Wild (LFW)-occ and (b) AR databases. In LFW-occ, three genuine pairs (green colour) and three impostor pairs (red colour) accounting for sunglasses are presented

TABLE 1 Ablative results on the AR dataset in terms of Rank-1 recognition accuracy (%)

Deep Model	OS Loss	Synthetic Occlusion	AR Sunglasses	AR Scarf
Trunk-CNN	No	No	65	95
Baseline	No	Yes	84	97
Proposed	Yes	Yes	**92**	**98**

Abbreviations: OS, Occlusion Segmentation
The best performances are typeset in bold

proposed SOIDN framework. The entire SOIDN is trained from end to end with a mixed occluded and occlusion-free face images by minimizing a combination of two losses (see Equation (2)). The hyperparameter λ is set to 1 by default. With the help of the OMA module, the SOIDN is easy to converge by around 20 epochs.

4.2.2 | Testing phase

First, the deep features of dimension 4096 from the fully connected layer are extracted. For distance measurement between two faces, the cosine metric is applied to obtain the similarity score. Finally, thresholding and the nearest neighbour classifier are used for face verification and identification, respectively.

4.2.3 | Baseline models

We take the VGGFace model [53] as our *trunk-CNN*, which is trained with VGG dataset of 2622 identities. Apart from that, the VGGFace model shares the same architecture with the VGG16 model except for the last softmax layer. The model trained with the same training data as the proposed SOIDN but without applying the OS module is regarded as the *baseline model*. Briefly, the hyperparameter λ in Equation (2) is set to 0. Data augmentation is involved in the baseline model to cope with occlusion implicitly and learn discriminative feature representations.

4.3 | Ablation study and analysis

4.3.1 | Contribution of different components

To explore the contributions of the deep OS supervision and data augmentation with synthetic occluded faces. If the hyperparameter λ in Equation (2) is set to 0, the objective is degraded to only include identity classification, and there is no OMA module applied to deep FR features (*baseline model*). We also investigate the importance of augmenting training data with synthetic occluded faces. It is worth mentioning that the proposed SOIDN requires to train with occlusion-free and synthetic occluded faces to ensure that the OS network branch functions well. Table 1 shows how each component contributes to the performance. As a result, training with augmented occluded faces improves the accuracy as our expectations. Remarkably, the model trained with OS supervision consistently outperforms the model that only trained with the classification loss.

TABLE 2 Rank-1 recognition accuracy (%) of the proposed approach on the AR dataset when different occlusion types are used in the training

Occlusion Type	AR Sunglasses	AR Scarf
Sunglasses	**95**	79
Sunglasses, cup, scarf	93	96
Sunglasses, cup, scarf, party mask	92	**98**

The best performances are typeset in bold

4.3.2 | The effect of synthetic occlusion for training

Since our method is trained with occlusion-free and synthetically occluded faces, we conduct exploratory experiments to investigate the effect of the occlusion type involved in the training. Table 2 shows how occlusion types affect performance. In short, the more occlusion types used to augment the training data, the more balanced the results that are achieved on different occlusions, because synthesized occluded faces ensure that the features are extracted more locally and equally. If there is only one occlusion type used for training, the performance suffers from a strong bias that results in accuracy degradation in an unseen occlusion type.

4.4 | Results of occlusion segmentation

The proposed SOIDN is capable of handling the occlusion problem, owing to the use of deep responses from the OS network. OS and FR are simultaneously performed to make use of their correlation relationship. Such modification would enhance the discriminative capability for FR at the expense of compromising the segmentation accuracy in some way. As a result, the occlusion detection model can work reasonably well with a mean IoU of 89.5 on the synthetically occluded faces. This mean IoU decreases compared with using the OS network only, with output IoU around 98.0, as it reflects the preservation of discriminative capability in the segmentation instead of a merely pixel-wise segmentation.

We show comparison results on OS by using the OS network and the proposed SOIDN in Figure 10. As the results demonstrate, the OS network renders more accurate

FIGURE 10 Examples of occlusion segmentation results on the Labeled Faces in the Wild-occlusions dataset. Each column of one subject shows, from left to right, an input image, the GT of the occlusion, segmentation results using the OS network, and the proposed SOIDN model. GT, ground truth; OS, occlusion segmentation; SOIDN, simultaneous occlusion-invariant deep network

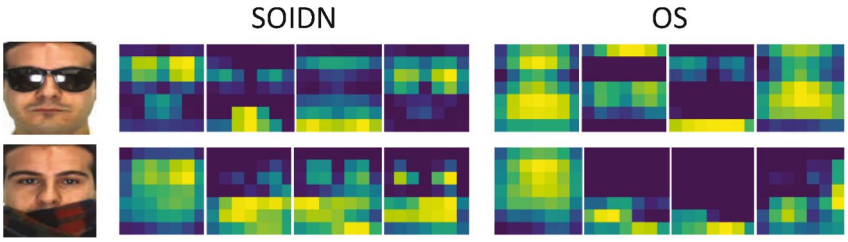

FIGURE 11 Illustration of the deep responses learnt from our proposed method and pure OS network. Our deep responses display discriminative facial regions to some extent. Only four channels of the deep response are shown for simplicity. OS, occlusion segmentation; SOIDN: simultaneous occlusion-invariant deep network

predication than SOIDN because of the mere use of a pixel-wise supervision signal. However, accurate occlusion introduced by pure OS network is redundant, as compact embedding is essential for FR. Based on observation, we find some tiny patches are segmented in our method if we take the party mask as an example. Similarly, the nasion in the sunglasses is also detected as a tiny patch. Such tiny patches instead of pixels contribute to masking the corrupted features due to occlusion in order to obtain occlusion-free features.

Apart from the OS demonstration, we also investigate the impact of classification supervision on the deep response of the OS network. Figure 11 illustrates the deep response (top-convolutional features) generated by the OS network. With our proposed method, the deep response is capable of locating the occlusion location in the image space to some extent, but it is not as good as the pure OS network does. Nevertheless, we observe that deep responses by our method show the potential ability to preserve the discriminative capability for FR. Specifically, the critical facial components such as eyes, nose, and mouth regions are displayed in the deep response. This is no surprise, with the incorporation of classification loss, such discriminative facial regions are emphasized and learnt to render compact feature embedding in the end.

4.5 | Results on Labeled Faces in the Wild occlusion dataset

We first compare the proposed SOIDN and baseline deep models under different occlusion categories in Table 3 to show there is a consistent improvement by using simultaneous segmentation. Specifically, up to 3% improvement has been achieved when occlusion occurs in the upper facial part (e.g. party mask). This is because compared with the lower facial part, the upper part in general contains more discriminative details. Superimposed occlusions such as party masks can heavily distort not only the discriminative information but also the global structure. In that case, getting rid of features corrupted by occlusion becomes essential. Utilizing occlusion-free features to recognize face is an effective way to solve this problem and can result in significant performance improvement (3% gain). Furthermore, the proposed SOIDN can obtain higher

TABLE 3 Face verification on the LFW-occ dataset regarding different occlusion categories

LFW-occ Dataset	Methods	Accuracy (%)
Sunglasses	Baseline	88.73
	SOIDN	**89.35**
Party mask	Baseline	86.37
	SOIDN	**89.50**
Scarf	Baseline	88.82
	SOIDN	**89.50**
Doctor mask	Baseline	88.82
	SOIDN	**89.37**
Cup	Baseline	89.30
	SOIDN	**89.90**

Abbreviations: LFW, labeled faces in the wild; SOIDN, simultaneous occlusion-invariant deep network. The best performances are typeset in bold.

accuracy and lower variance across different occlusions compared with the baseline method.

To further understand the embedded features learnt by the trunk-CNN model and SOIDN, we plot all images on the 2-D plane as a scatter plot. There are many dimension-reduction methods such as multidimensional scaling and PCA; we select t-SNE, as it can strongly reveal the dissimilar points and present the cluster clearly. For each face image, we use the embedding feature from the last layer of models as its t-SNE embedding.

Figure 12 shows the visualization of VGGFace and SOIDN. In this figure, the different subjects are encoded by colour, and the shape of each instance encodes the occlusion object. There are 60 images from five subjects presented in both of these views. For the projection of VGGFace, we find some images of the same subjects are loosely cluster together. But some images of different subjects are mixed with each other(Figure 12 A). Moreover, some images of the same subjects are evenly distributed into separated clusters (Figure 12 B and C). As for the projection of SOIDN, almost all the images of the same subjects are well grouped together. This indicates that the embedding features extracted by SOIDN are more robust to occlusion and can better present the image similarity. It is predictable that the proposed method outperforms deep models trained for general FR.

4.6 | Results on AR dataset

The AR face database, introduced in Sec 4.1, is one of the very few benchmark datasets that contain real occlusions. It consists of over 4000 faces of 126 individuals. Occlusions include sunglasses and scarfs and the faces show various expressions and variations of illumination. To explore how well the existing advanced deep models perform on the real occluded face dataset, we select several off-the-shelf deep models that are publicly available as a feature extractor for FR. We report Rank-1 recognition accuracy of SOIDN and the existing off-the-shelf deep models in Table 4. The results show that SOIDN consistently outperforms all deep models on both occlusions. This is remarkable as Inception-ResNet-V1 has a much deeper network architecture and was also trained with a larger scale training dataset (e.g. entire CASIA-WebFace, VGGFace2) compared with SOIDN, but their results on occluded face dataset perform worse. Simply utilizing deep models trained for unconstrained FR cannot handle the occlusion properly, which further confirms the effectiveness of the network architecture of SOIDN. It is worth noting that we use the single-sample-per-subject protocol for the experiments, which is the most challenging protocol as it requires only one image per subject for enrolment. Specifically, we enrol one occlusion-free face image, and the images of sunglasses and scarf occlusions are used for testing.

Table 5 reports a comparison of Rank-1 recognition accuracies with state-of-the-art OFR methods. The results show that the proposed SOIDN method is comparable to state-of-the-art methods. Specifically, SOIDN achieves a 98% accuracy on scarf occlusion, which is the same as the state of the art. In terms of sunglasses occlusion, SOIDN performs worse than PDSN but it is worth noting that the network architecture we used is very shallow compared with PDSN. Specifically, we utilize simply classic VGG16 and the other methods for example, PDSN is utilizing advanced CNN(e.g. ResNet50) as the network architecture. In addition, even though these methods follow the same protocols for testing, SOIDN is not tuned with any AR faces for training, while other methods—e.g. RPSM and LMA—are usually trained with this dataset. As for PDSN, it does not include AR faces to generate mask dictionary, while it incorporates AR faces to train OS. As for SOIDN, we employ the refined VGG16 model as the initial weights of SOIDN model and then trained with CASIA-WebFace in an end-to-end manner. It does not incorporate any AR faces during the entire training process; thus, the experimental settings are more stringent on our side. The reason why the proposed SOIDN outperforms the other methods is that the OS network and the FR network of SOIDN explicitly consider occlusion location and occlusion content and are coherently combined and optimized within a simultaneous architecture, which ensures robustness to occlusion variation. In addition, with the OMA block the OS task and the FR task can help each other to obtain occlusion-free face representation. While the other methods such as RPSM and LMA convert OFR into an image patch matching problem that cannot locate occlusion precisely and further degrades the recognition accuracy. The PDSN performs OS and FR sequentially and the imperfect segmentation will unavoidably impair FR.

5 | CONCLUSION

The FR results on synthesized and realistic face datasets obtained by the proposed SOIDN are promising. Herein, we propose addressing OFR in a simplified yet well-motivated way. Specifically, an OMA is designed as a bridge in SOIDN that is motivated by the phenomenon that corrupted features by occlusion are traceable within an OS network. We use the

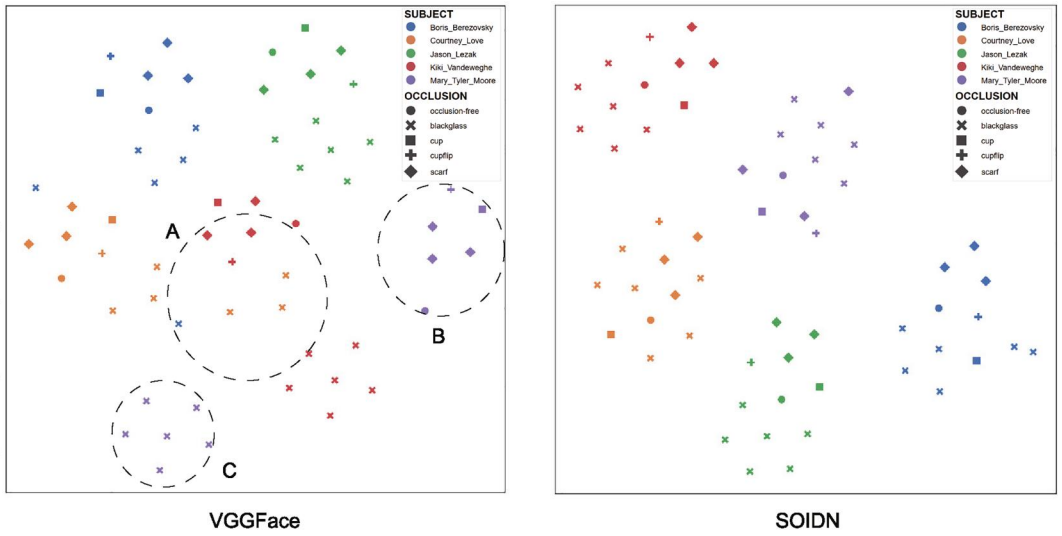

FIGURE 12 t-SNE projection by embedding features of VGGFace and simultaneous occlusion-invariant deep network. Please zoom in for better observation

Deep Model	Training Dataset	AR Sunglasses	AR Scarf
Inception-ResNet-V1 [54]	CASIA-WebFace	88	91
Inception-ResNet-V1 [54]	VGGFace2	80	81
MobileFace (arc loss) [55]	MS-Celeb-1M [56]	83	94
SOIDN	CASIA-WebFace*	**92**	**98**

Abbreviation: SOIDN, simultaneous occlusion-invariant deep network. The best performances are typeset in bold.

TABLE 4 Rank-1 recognition accuracy (%) of the proposed method and existing off-the-shelf deep models on AR dataset. CASIA-WebFace* indicates that synthetically occluded faces are generated from the CASIA-WebFace

TABLE 5 Rank-1 recognition accuracy (%) of the proposed simultaneous occlusion-invariant deep network approach and state-of-the-art methods on AR dataset

Deep Model	AR Sunglasses	AR Scarf
RPSM [33]	85	90
LMA [23]	96	94
PDSN [44]	**98**	**98**
Baseline	84	97
SOIDN	92	**98**

Abbreviations: LMA, largest matching area; PDSN, pairwise differential siamese network; RPSM, robust point set matching; SOIDN, simultaneous occlusion-invariant deep network. The best performances are typeset in bold.

classic VGG16 network as the FR network branch, but other advanced networks can be incorporated into the proposed framework for better performance. To the best of our knowledge, this work is the first to coherently combine FR and OS networks and optimize them within a simultaneous architecture rather than in a sequential pipeline. In the future, we will apply more advanced CNN architectures to the proposed framework and evaluate their performance.

ORCID

Dan Zeng https://orcid.org/0000-0002-9036-7791

REFERENCES

1. Best-Rowden, L., Jain, A.K.: Longitudinal study of automatic face recognition. IEEE Trans. Pattern. Anal. Mach. Intell. 40(1), 148–162 (2017)
2. Masi, I., et al.: Deep Face Recognition: A Survey. In: 31st SIBGRAPI Conference on Graphics, Patterns and Images, pp. 471–478. SIBGRAPI. Paraná, Brazil (2018)
3. Rasti, S., Yazdi, M., Masnadi-Shirazi, M.A.: Biologically inspired makeup detection system with application in face recognition. IET Biom. 7(6), 530–535 (2018)
4. Zhao, F., et al.: Robust LSTM-Autoencoders for face de-occlusion in the wild. IEEE Trans. Image Process. 27(2), 778–790 (2018)
5. Wan, W., Chen, J.: Occlusion robust face recognition based on mask learning. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 3795–3799. IEEE, Beijing (2017)
6. Ge, S., et al.: Detecting masked faces in the wild with lle-cnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2682–2690. IEEE, Venice, Italy (2017)
7. Osherov, E., Lindenbaum, M.: Increasing cnn Robustness to Occlusions by Reducing Filter Support. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 550–561. IEEE, Venice, Italy (2017)
8. Zhou, E., Cao, Z., Yin, Q.: Naive-Deep Face Recognition: Touching the Limit of LFW Benchmark or Not?. arXiv preprint arXiv:1501.04690 (2015)
9. Chen, Z., Xu, T., Han, Z.: Occluded face recognition based on the improved SVM and block weighted LBP. In: Proceedings of the International Conference on Image Analysis and Signal Processing, pp. 118–122. IEEE, Brussels, Belguim (2011)
10. Min, R., Hadid, A., Dugelay, J.L.: Improving the recognition of faces occluded by facial accessories. In: Proceedings of Face and Gesture, pp. 442–447. IEEE, Santa Barbara, CA (2011)

11. Oh, H.J., Lee, K.M., Lee, S.U.: Occlusion invariant face recognition using selective local non-negative matrix factorization basis images. Image Vis. Comput. 26(11), 1515–1523 (2008)
12. Neo, H.F., Teo, C.C., Teoh, A.B.: Development of partial face recognition framework. In 2010 Seventh International Conference on Computer Graphics, Imaging and Visualization, (pp. 142–146). IEEE, Sydney, NSW, Australia (2010)
13. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440. IEEE, Boston, Massachusetts (2015)
14. Zhang, W., et al.: Local gabor binary Patterns based on Kullback-Leibler divergence for partially occluded face recognition. IEEE Signal Process Lett. 14(11), 875–878 (2007)
15. Zhang, W., et al.: Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 786–791. IEEE, Santiago (2005)
16. Hua, G., Akbarzadeh, A.: A robust elastic and partial matching metric for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2082–2089. IEEE, Florida (2009)
17. Cheheb, I., et al.: Random sampling for patch-based face recognition. In: Proceedings of the 5th international Workshop on Biometrics and Forensics, pp. 1–5. Coventry, UK (2017)
18. Gottumukkal, R., Asari, V.K.: An improved face recognition technique based on modular PCA approach. Pattern Recogn Lett. 25(4), 429–436 (2004)
19. Leonardis, A., Bischof, H.: Robust recognition using eigenimages. Comput Vis Image Understand. 78(1), 99–118 (2000)
20. Kim, J., et al.: Effective representation using ICA for face recognition robust to local distortion and partial occlusion. IEEE Trans. Pattern. Anal. Mach. Intell. 27(12), 1977–1981 (2005)
21. Hotta, K.: Robust face recognition under partial occlusion based on support vector machine with local Gaussian summation kernel. Image Vis. Comput. 26(11), 1490–1498 (2008)
22. Seo, J., Park, H.: A robust face recognition through statistical learning of local features. In: Proceedings of International Conference on Neural Information Processing, pp. 335–341. Granada, Spain (2011)
23. McLaughlin, N., Ming, J., Crookes, D.: Largest matching areas for illumination and occlusion robust face recognition. IEEE Trans. Cybern. 47(3), 796–808 (2017)
24. Wright, J., et al.: Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell. 31(2), 210–227 (2009)
25. Zhou, Z., et al.: Face recognition with contiguous occlusion using markov random fields. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1050–1057 (2009)
26. Yang, M., et al.: Gabor feature based robust representation and classification for face recognition with Gabor occlusion dictionary. Pattern Recogn. 46(7), 1865–1878 (2013)
27. Li, Y., Feng, J.: Reconstruction based face occlusion elimination for recognition. Neurocomputing. 101, 68–72 (2013)
28. Ou, W., et al.: Robust face recognition via occlusion dictionary learning. Pattern Recogn. 47(4), 1559–1572 (2014)
29. Luan, X., et al.: Extracting sparse error of robust PCA for face recognition in the presence of varying illumination and occlusion. Pattern Recogn. 47(2), 495–508 (2014)
30. Zhao, S., Hu, Z.-p.: A modular weighted sparse representation based on Fisher discriminant and sparse residual for face recognition with occlusion. Inf, Process, Lett. 115(9), 677–683 (2015)
31. Yu, Y.-F., et al.: Discriminative multi-scale sparse coding for single-sample face recognition with occlusion. Pattern Recogn. 66, 302–312 (2017)
32. Wu, C.Y., Ding, J.J.: Occluded face recognition using low-rank regression with generalized gradient direction. Pattern Recogn. 80, 256–268 (2018)
33. Weng, R., Lu, J., Tan, Y.-P.: Robust point set matching for partial face recognition. IEEE Trans. Image Process. 25, 1163–1176 (2016)
34. Lv, J.-J., et al.: Data augmentation for face recognition. Neurocomputing. 230, 184–196 (2017)
35. Trigueros, D.S., Meng, L., Hartnett, M.: Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. Image Vis. Comput. 79, 99–108 (2018)
36. Cen, F., Wang, G.: Dictionary representation of deep features for occlusion-robust face recognition. IEEE Access. 7, 26595–26605 (2019)
37. Cheng, L., et al.: Robust deep auto-encoder for occluded face recognition. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 1099–1102. Brisbane, Australia (2015)
38. Goodfellow, I., et al.: Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems, pp. 2672–2680. Montréal, CANADA (2014)
39. Chen, Y.A., et al.: Occlusion-aware face inpainting via generative adversarial networks. In: Proceedings of the IEEE International Conference on Image Processing, pp. 1202–1206. ICIP, Beijing, China (2017)
40. Chen, X., et al.: From eyes to face synthesis: a new approach for human-centred smart surveillance. IEEE Access. 6, 14567–14575 (2018)
41. Hu, B., Yang, W., Ren, M.: Unsupervised Eyeglasses Removal in the Wild. IEEE Transactions on Cybernetics (Early Access. preprint arXiv: 1909.06989 1–13 (2020)
42. Ge, S., et al.: Occluded face recognition in the wild by identity-diversity inpainting. In: IEEE Transactions on Circuits and Systems for Video Technology (2020)
43. Xia, Y., Zhang, B., Coenen, F.: Face occlusion detection based on multi-task convolution neural network. In: Proceedings of Fuzzy Systems and Knowledge Discovery, pp. 375–379. Zhangjiajie, China (2015)
44. Song, L., et al.: Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In: Proceedings of the IEEE international conference on computer vision, pp. 773–782. Seoul, Korea (2019)
45. Zheng, W., Gou, C., Wang, F.-Y.: A novel approach inspired by optic nerve characteristics for few-shot occluded face recognition. Neurocomputing. 376, 25–41 (2020)
46. Zhang, J., et al.: Feature agglomeration networks for single stage face detection. Neurocomputing. 380, 180–189 (2020)
47. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556 (2014)
48. Saito, S., Li, T., Li, H.: Real-Time Facial Segmentation and Performance Capture From rgb Input, In: European Conference on Computer Vision, pp. 244–261. Amsterdam, the Netherlands (2016)
49. Yi, D., et al.: Learning Face Representation from Scratch. arXiv preprint arXiv:1411.7923 (2014)
50. Huang, G.B., et al.: Labelled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. In: Technical Report, pp. 07–49. University of Massachusetts, Amherst (2007)
51. Aleix, M., Robert, B.: The AR face database. CVC Tech. Rep. 24 (1998)
52. Zhang, K., et al.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett. 23(10), 1499–1503 (2016)
53. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep Face Recognition. In: Proceedings of the British Machine Vision Conference (BMVC), pp. 41.1–41.12. Swansea, UK, (2015)
54. Szegedy, C., et al.: Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 4278–4284. California (2017). preprint arXiv:1602.07261
55. Chen, S., et al.: Mobilefacenets: Efficient cnns for Accurate Real-Time Face Verification on Mobile Devices. In: Proceedings of Chinese Conference on Biometric Recognition, pp. 428–438. Urumchi, China (2018)
56. Guo, Y., et al.: Ms-celeb-1m: A Dataset and Benchmark for Large-Scale Face Recognition. In: European Conference on Computer Vision, pp. 87–102. Amsterdam, the Netherlands (2016)

How to cite this article: Zeng D, Veldhuis R, Spreeuwers L, Arendsen R. Occlusion-invariant face recognition using simultaneous segmentation. *IET Biometrics*. 2021;1−13. https://doi.org/10.1049/bme2.12036