IET Research Journals

Submission Template for IET Research Journal Papers

IET Journals
The Institution of Engineering and Technology

Occlusion Invariant Face Recognition Using Simultaneous Segmentation

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Dan Zeng[1*], *Raymond Veldhuis*[2], *Luuk Spreeuwers*[3], *Richard Arendsen*[4]

[1] *Southern University of Science and Technology, China*
[1,2,3] *University of Twente, The Netherlands*
[4] *20face BV, The Netherlands*
* *E-mail: zengd@sustech.edu.cn*

Abstract: When using CNN models to extract features of an occluded face, the occluded part will inevitably be embedded into the representation in latent space, just as other facial regions. Existing methods deal with occluded face recognition either by augmenting the training dataset with synthesized occluded faces or by detecting/segmenting occlusions first and subsequently recognize the face based on unoccluded facial regions. Instead, we develop simultaneous occlusion segmentation and face recognition to make the most of the correlation relationship lie in 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, we propose a simultaneous occlusion invariant deep network (SOIDN), containing 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 proposed 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 datasets (e.g., LFW-occ) and real occluded face datasets (e.g., AR) demonstrate that the proposed approach outperforms state-of-the-art methods for face verification and identification when handling occlusion challenges.

1 Introduction

Face recognition, as one of the most ubiquitous biometric technologies, is widely used in applications such as governmental, commercial, computer security, voter verification and etc. Deep convolutional neural networks (CNNs) have pushed the frontier of many computer vision applications, including unconstrained face recognition [1, 2]. Face recognition 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, 37]. 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 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 Fig. 1, facial appearance changes substantially due to occlusion. Therefore, occluded face recognition is still considered one of the most intractable problems.

Two essential factors related to the occlusion challenge are: where is the occlusion (*location*) and what is the occlusion (*content*) [7, 37]. If there is a dataset that presents sufficient examples concerning occlusion location and content that may occur in real-world applications [6], 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 features extracted more locally and equally and can handle occlusions better [7]. It is worth mentioning that occlusion segmentation, i.e., 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 [29–32]. However, occlusion segmentation and face recognition constitute a sequential pipeline, resulted in face recognition dependent on occlusion segmentation results. As a consequence, the side effect of imperfect segmentation will unavoidably impair face recognition [37].

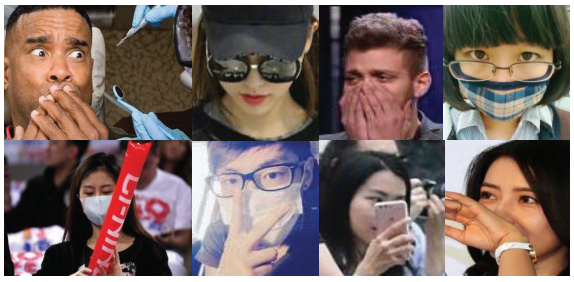

Fig. 1: Examples of real-world occlusion from the MAFA dataset [4]

In contrast, this paper develops simultaneous occlusion segmentation and face recognition for better information utilization (see Fig. 2). We observe the phenomenon that features corrupted by occlusion are traceable within a CNN trained for segmenting occluded parts in face images [8] (see Fig. 3). Specifically, pixel-wise occlusion is traced in feature maps of convolutional layer, which ensures the occlusion location can be preserved through the segmentation CNN. This inspired us to leverage such deep occlusion response of occlusion segmentation to clean that latent representation from occlusion artifacts. In a nutshell, we propose a simultaneous occlusion invariant deep network (SOIDN), containing simultaneously operating face recognition and occlusion segmentation networks, involving an occlusion mask adaptor module as a bridge between their top convolutional layers to learn occlusion feature masks from top-convolutional layers of occlusion segmentation. Specifically, the occlusion mask adaptor module intends to learn the correspondence between convolutional features of occlusion segmentation and the occlusion mask so that the channel-wise convolutional features of occlusion segmentation are correctly matched with their counterparts in face recognition. The proposed SOIDN enjoys several advantages: (1) Two aspects of occlusion including *location* and *content* explicitly considered by occlusion segmentation network and face recognition network, respectively,

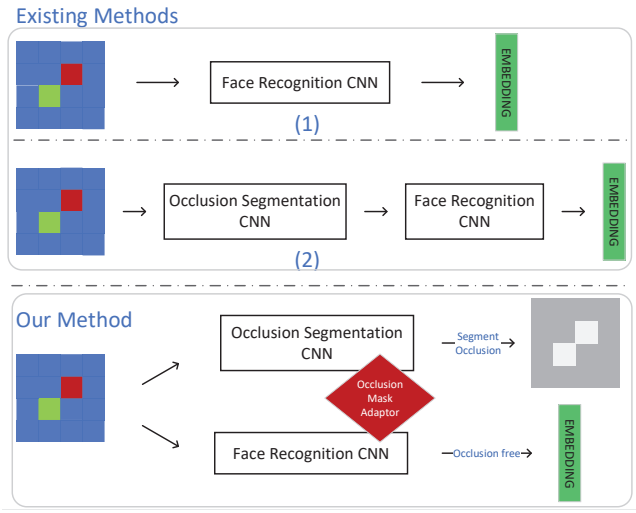

Fig. 2: Existing methods either (1) rely on a deep CNN to extract occlusion robust features (mainly use synthesized occluded faces for data augmentation) or (2) utilize occlusion segmentation and face recognition individually and sequentially to handle occlusion challenge. In contrast, our proposed method combines both coherently and optimizes in a simultaneous architecture to learn occlusion invariant embedding features.

are combined coherently and optimized in a simultaneous architecture; (2) Occlusion segmentation and face recognition can help each other to obtain an occlusion-free face representation. If face representation extracted from the face recognition network is already unaffected by the actual occlusion, then the occlusion feature mask plays a less important role in purifying representation. If face representation is rather affected by the occlusion and deteriorates the discriminatively, then we exclude corrupted features 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.

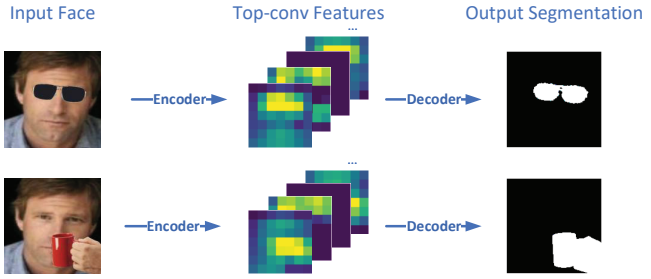

Fig. 3: The deep response (top-conv features) generated by occlusion segmentation network shows the trace of features corrupted by occlusion (sunglass and cup). Only the first four channels in Top-conv features are shown for simplicity.

The training of proposed SOIDN is jointly supervised by minimizing two losses, namely, classification loss and segmentation loss. Classification loss ensures the occlusion masks are 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 parts from nonface parts. With the supervision of two losses, we aim to obtain: (1) occlusion-free features to preserve discriminability for occluded face recognition, (2) occlusion segmentation output involving facial parts that impair face recognition accuracy, and (3) an occlusion mask that weighs the reliability of features to minimize the considerable intra-class 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 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 combine face recognition network and occlusion segmentation network coherently and optimize in a simultaneous architecture to learn occlusion invariant features.
- We design the occlusion mask adaptor as a bridge in SOIDN, being motivated by the phenomenon that features corrupted by occlusion are traceable within an occlusion segmentation network.
- We implement the proposed SOIDN with a combined loss function, including classification and segmentation losses, and can 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. Conclusion is given in Section 5.

2 Related Work

Approaches to recognize faces under occlusions can be broadly classified into three categories which are (i) occlusion robust feature extraction, (ii) occlusion recovery based face recognition, and (iii) occlusion aware face recognition. 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 face recognition under occlusion challenge and highlight the difference in the proposed method.

2.1 Approaches not using deep learning for OFR

The first category, occlusion robust feature extraction, extracts handcraft features or learns features from explicitly defined facial regions. Ref. [9] introduces Kullback-Leibler divergence to measure the distance of Local Gabor Binary Patterns (LGBP) descriptors [10] of the local region of test images and that of the unoccluded region of reference faces. Robust matching metric [11] is presented to match the difference of Gaussian (DoG) filter descriptor of facial part against its spatial neighborhood in the other faces and select the minimal distance for face recognition. A random sampling patch-based method [12] 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 [13, 14] are developed to handle occlusion challenge. Independent component analysis (ICA) [15] is used to find locally salient information from important facial parts. Statistic learning methods such as local Gaussian Kernel based features [16] or a simple Gaussian model [17] for feature probability estimation, addressing occlusion occurring as a probability problem. Mclaughlin et al. [18] propose the largest matching areas at each point on the face by assuming the occluded test image region can be modeled by an unseen-data likelihood with a low posterior probability.

The second category, occlusion recovery based face recognition, recovers a clean face from the occluded one for recognition. Sparse representation classifiers (SRC) [19] and variants retain popularity and success in coping with the occlusion challenge. The main idea of SRC 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 [20], using Gabor features instead of pixel values [21], applying downsampled SRC [22] to locate occlusion at a low computing complexity, importing mutual-incoherence regularization term in SRC scheme [23], exploiting the sparse error component with robust principal component analysis [24], and introducing modular weight-based SRC [25]. Recently, Ref. [26] proposes a joint and collaborative representation with local

adaptive convolution feature, containing local high-level features from local regular regions. Ref. [27] proposes a hierarchical sparse and low-rank regression model using features based on image gradient direction. Robust point set matching (RPSM) [28] 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 would fail to work if face keypoints are unavailable due to occlusions as facial alignment is required during preprocessing[26–28]. Moreover, the computation complexity is high, which makes the recognition process slow.

The third category, occlusion aware face recognition, usually discards the occlusion part and performs face recognition based on the visible face parts only. Ref. [29, 30] divide a face into multiple non-overlapping regions and train an SVM classifier to identify the occluded area. Ref. [31] introduces selective local non-negative matrix factorization (NMF) method to select features corresponding to occlusion-free regions for recognition. A work [32] extends NMF to include occlusions estimation adaptively according to the reconstruction errors. Finally, low-dimensional representations are learned to ensure that features of the same class close to that of the mean class center.

2.2 Deep learning Approaches for OFR

Face representation obtained by deep convolutional neural networks (DCNN) is vastly superior to traditional learning methods in discriminative power which has pushed the frontier of deep face recognition [2]. Some methods [33, 34] take the advantages of data augmentation to generate sufficient synthetically occluded faces for training a deep network. Lv et al. [33] synthesize occluded faces with various hairstyles and glasses to augment the training dataset. Specifically, 87 hairstyles 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 [34], instead of using synthetic occluded faces directly, they identify the importance of face regions based on their occlusion sensitivity and then train a CNN with identified facial regions covered to reduce the model's reliance on these regions. Specifically, training face images are augmented with occlusion 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 brings less accuracy degradation when the central face part is occluded. Cen et al. [35] propose a deep dictionary representation based classification (DDRC) scheme to alleviate the occlusion effect in face recognition, where the dictionary is used to code the deep convolutional features linearly.

Deep learning techniques also used for occluded face reconstruction. Ref. [36] extends stacked sparse denoising auto-encoder to double channel for facial occlusion removal. Zhao et al. [37] combine the LSTM and autoencoder architectures to address the face de-occlusion problem. The proposed robust LSTM-Autoencoders consists 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. Besides, the adversarial CNNs are introduced to enhance the discriminative information in the recovered faces. Generative adversarial network (GAN) [38] and variants retain popularity and succeed in synthesizing or generating new samples. Occlusionaware GAN [39] is proposed to identify the corrupted image region with associated corrupted region recovered by utilizing a GAN pre-trained on occlusion-free faces. Ref. [40] employs GAN for eyes-to-face synthesis with only eyes visible. Eyeglasses removal generative adversarial network (ERGAN) [41] is proposed for eyeglasses removal in the wild via an unsupervised manner, and capable of rendering a competitive removal quality in terms of realism and diversity. In paper [42], ID-GAN (identity diversity generative adversarial network) combines CNN-based recognizer and GANbased recognition to inpaint realism and identity-preserving faces with 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 occluded face recognition with limited training samples, Ref. [55] proposes a structural element feature extraction method to capture the local and contextual information inspired by the human optic nerve characteristics for face recognition. Besides, an adaptive fusion method is proposed to use multiple features consisting of a structural element feature, and a connected-granule labeling feature. To exploit inherent multi-scale features of a single convolutional neural network, FANet [56] introduced an Agglomeration Connection module to enhance the context-aware features and to augment low-level feature maps with a hierarchical structure so that it can cope with scale variations in face detection effectively. Ref. [43] predicts occlusion probability of the predefined face components by training a multitask CNN. In paper [5], they propose to add the MaskNet module to the middle layer of CNN models, aiming at learning image features with high fidelity and ignore those distorted caused by occlusions. The MaskNet, a shallow convolutional network, assigns lower weights to hidden units activated by the occluded facial areas. Recently, 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 occlusion segmentation and face recognition sequentially. Differently, the proposed SOIDN combines both coherently and optimizes in a simultaneous architecture to learn occlusion invariant features. To the best knowledge of us, this work is the first to carry out occlusion segmentation and face recognition simultaneously to make the most of correlation relationship lie in them.

3 Proposed Approach

3.1 Problem Statement

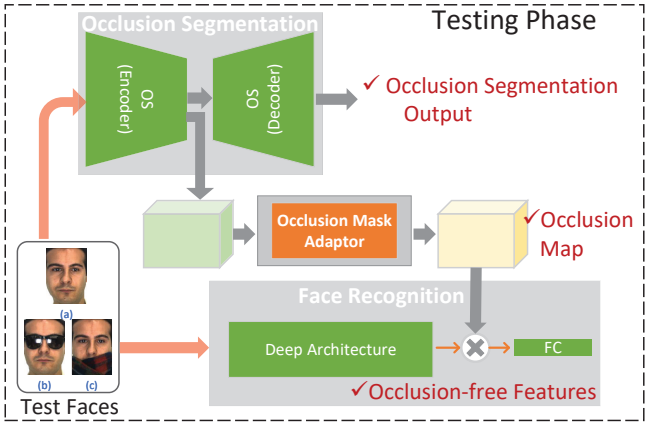

Fig. 4: An overview of the proposed framework. It consists of an occlusion segmentation network $g(\cdot)$ and a face recognition network $f(\cdot)$ in parallel, with the occlusion mask adaptor module $M(\cdot)$ as a bridge. For test faces, (a) indicates occlusion-free face image. (b) and (c) indicate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

To address the occluded face recognition problem, extracting occlusion invariant features is the 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 relying 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 the similarity among faces can be computed using distance metrics, i.e., Euclidean or Cosine, through the occlusion invariant feature embedding space. The overview of the proposed framework is shown in Fig. 4 which is under the latter group.

The formula definition of the proposed SOIDN is given 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 face recognition (FR) network and occlusion segmentation (OS) network, respectively. Here $f(x)$ and $g(x)$ required to have the same width and height. The occlusion mask adaptor (OMA) module $M(\cdot)$ takes OS features as input to generate occlusion mask $M(g(x))$. We multiply each weights in the occlusion mask with FR features $f(x)$ at the same spatial location to mask out the corrupted features. In face recognition 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 occlusion segmentation and face recognition for occlusion invariant features extraction. The structure of the proposed method is shown in Fig. 5. The deep architecture of face recognition can be arbitrary. More specifically, we adopt the widely used VGG16 [45] as an example of the FR network to illustrate how our method improves the embedded features for occluded face recognition. The OS network is responsible for detecting occlusion pixel-wise in a face image. For simplicity, we directly adopt fcn-8s [46] as an example for segmentation, which 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 the FR features. Occlusion mask generation encourages purified features (excluding corruption) to be as close as that extracted from the same identity yet occlusion-free face image which is constrained by the proposed classification loss. The FR network, if used alone, may extract features not significantly affected by occlusion to a certain extent. 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.

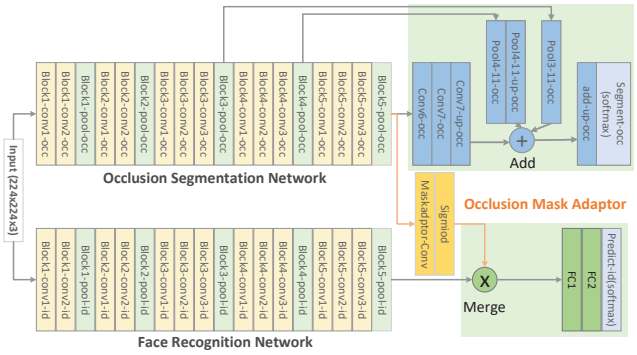

Fig. 5: Structure of the proposed SOIDN for occlusion invariant feature extraction. VGG16 is taken as an example of the FR network.

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 the features after applying the occlusion mask are extracted discriminate and occlusion invariant. The second segmentation loss l_{seg} guarantees to segment 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 the corrupted features are masked out, and only occlusion free features qualify for face recognition. 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 Eq. (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, which outputs a probability distribution over the n identity classes y^{cls}, and the weights W are learned.

Occlusion segmentation loss l_{seg}: We use the supervision signal of segmentation to ensure 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.

The most commonly used segmentation loss is a pixel-wise cross-entropy loss. It 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:

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

where the loss over the clean facial region and occluded facial region are added up to constitute the occlusion segmentation loss. We use $\hat{y}_i$ to represent the predicted pixel-wise class label. This scoring is repeated over 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 occlusion segmentation. 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 (Occlusion Mask Adaptor) network to address the occlusion mask generation problem. Fig. 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 the output

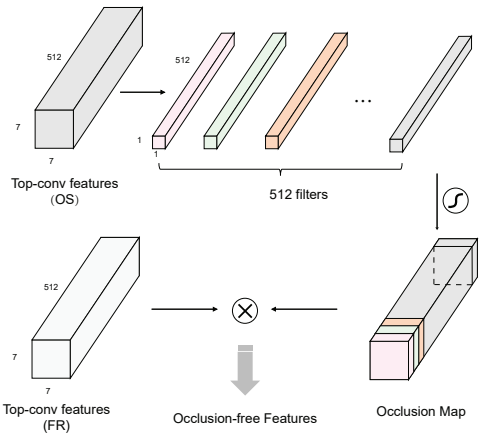

Fig. 6: Up: the OMA network consisting of 1×1 convolution layer and sigmoid. Down: the process of calculating occlusion-free features.

values of the occlusion mask into the interval of $[0, 1]$. 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 the channel-wise OS convolutional features are correctly matched with the counterpart the FR features and occlusion free features are extracted. We continue to pass these features on two fully connected layers to extract *occlusion invariant features* for occluded face recognition, which is indicated in Eq. (1).

3.2.2 Occlusion Segmentation: The proposed SOIDN is capable of coping with the occlusion problem, owing to the use of the OS network. Specifically, face recognition and occlusion segmentation 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 occlusion segmentation loss but also implicitly adjusted by the classification loss. In view of this, the occlusion segmentation results can be considered a predictor of the robustness of the FR network in terms of occlusion. If the occlusion is segmented from a face region accurately, we can conclude that the FR network is sensitive to the occlusion. Because the corrupted features that are masked out all originate from the occlusion region in an image. In other words, the FR network performs better with the use of occlusion mask, which also means 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 is that the training of proposed SOIDN is jointly supervised by minimizing a combination of classification and segmentation losses, with the former one 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., 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 [47] and synthetic occluded CASIA-Webface. The occluded faces are randomly synthesized from an occlusion-free face by 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 have a higher probability of occurrence compared with others. Hence, we collect occlusion templates to include typical occlusion examples. Fig. 7 lists all occlusion templates used in the paper. Samples of training faces and corresponding occlusion labels are shown in Fig. 8. To be sure the occluded faces do not dominate the within-class variation, only the subjects having more than 50 images are chosen for training, which results in 3459 individuals are involved.

Fig. 7: The occlusions 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.

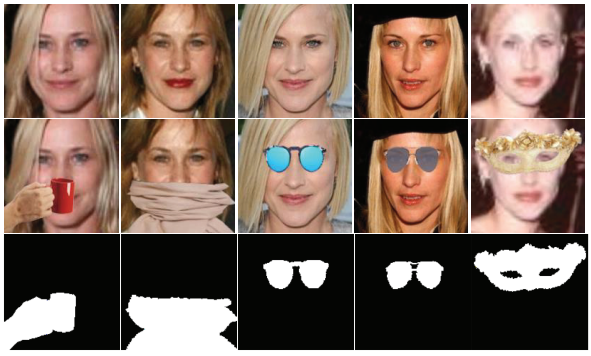

Fig. 8: Training examples of occlusion-free faces (first row), synthesized occluded faces (second row), and occlusion labels for occlusion segmentation (third row). The occlusion labels for occlusion-free faces are omitted for simplicity.

LFW dataset [48] 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 Fig. 9(a).

AR face database [49] is one of the very few benchmark datasets that contain real occlusions (see Fig. 9(b)). It consists of over 4000 faces of 126 individuals: 70 men and 56 women, taken in two sessions with a two-week interval. There are 13 images per individual in every session, and these images differ in terms of facial expression, illumination, and partial occlusion, getting sunglasses and scarves involved. Index 8 and 11 of each session indicates the person wearing sunglasses or a scarf, respectively. Index 9-10 and 11-12 combine the sunglasses or the scarf with illuminations, respectively.

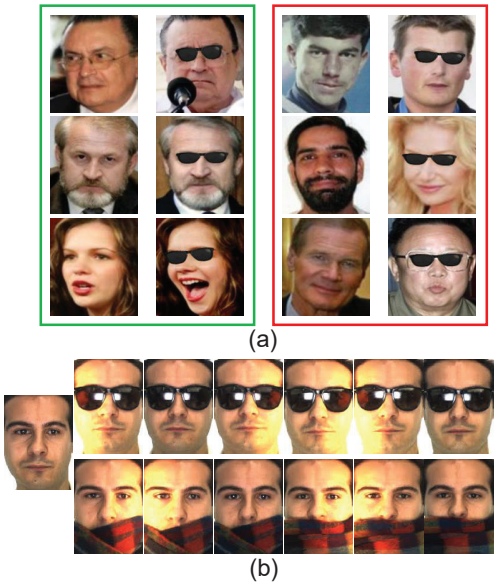

Fig. 9: Samples are shown in (a) LFW-occ and (b) AR databases. In LFW-occ, three genuine pairs (green color) and three impostor pairs (red color) accounting for sunglasses are presented.

4.2 Experimental Settings

In our experiments, all face images are preprocessed through face detection and face landmarking by using the standard MTCNN (Multitask Cascaded Convolutional Networks)[50]. After applying affine transformation based on four landmarks, i.e., left eye center, right eye center, nose tip, and mouth center, the face images are aligned and resized to 224×224.

Training phase: We employed the refined VGG16 model [45] 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 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 Eq. 2). The hyperparameter λ is set to 1 by default. With the help of the occlusion mask adaptor module, the SOIDN is easy to converge by around twenty epochs.

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 neighbor classifier are used for face verification and identification, respectively.

Baseline models: We take the VGGFace model [51] as our *trunk-CNN*, which is trained with VGG dataset of $2,622$ 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 of the proposed SOIDN, but without applying the occlusion segmentation module is regarded as the *baseline model*. To put it briefly, the hyperparameter λ in Eq. (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

Contribution of Different Components.: To explore the contributions of the deep occlusion segmentation supervision and data augmentation with synthetic occluded faces. If the hyperparameter λ in Eq. (2) is set to 0, the objective is degraded to only include identity classification, and there is no occlusion mask adaptor 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 occlusion segmentation 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 occlusion segmentation supervision consistently outperforms the model that only trained with the classification loss.

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 occlusion type involved in the training. Table 2 shows how occlusion types affect the performance. In short, the more occlusion types used to augment the training data, the more balanced results are achieved on different occlusions because synthesized occluded faces ensure 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, which 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. Occlusion segmentation and face recognition are simultaneously performed to make the correlation relationship lie in them. Such modification would enhance the discriminative capability for face recognition 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 occlusion segmentation by using the OS network and the proposed SOIDN in Fig. 10. As the results demonstrated, the OS network renders more accurate 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 face recognition. 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 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 occlusion segmentation demonstration, we also investigate the impact of the classification supervision on the deep response of the occlusion segmentation network. Fig. 11 illustrates

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**

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**

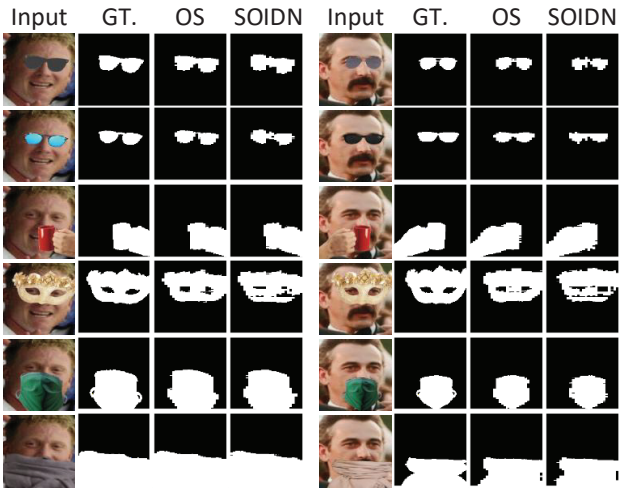

Fig. 10: Examples of occlusion segmentation results on the LFW-occ dataset. Each column of one subject shows, from left to right, an input image, the ground truth of the occlusion, segmentation results using the OS network, and the proposed SOIDN model.

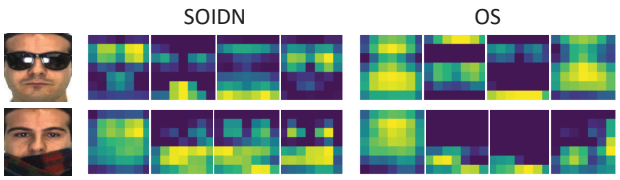

Fig. 11: Illustration of the deep responses learned from our proposed method and pure OS network. Our deep responses display discriminative facial regions to some extent. Only four channels of the deep response are shown for simplicity.

the deep response (top-conv features) that are generated by occlusion segmentation 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 good as the pure OS network does. Nevertheless, we observe that the deep responses by our method show the potential ability to preserve the discriminative capability for face recognition. 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 learned to render compact feature embedding in the end.

4.5 Results on LFW-occ 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 incurs in the upper facial part (e.g., party mask). The reason is that compared with the lower facial part, the upper part in general contains more discriminative details. The superimposed occlusion 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% gains). Furthermore, the proposed SOIDN can obtain higher accuracy and lower variance across different occlusions compared with the baseline method.

To further understand the embedded feature learned by the trunk-CNN model and SOIDN, we plot all images on the 2D plane as scatter plot. There are many dimension reduction methods such as

Table 3 Face verfication 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**

MDS 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.

Fig. 12 shows the visualization of VGGFace and SOIDN. In this figure, the different subjects are encoded by color, and the shape of each instance encodes the occlusion object. There are sixty images from five subjects presented in both of these two 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(Fig. 12 A). Moreover, some images of the same subjects are evenly distributed into separated clusters (Fig. 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 face recognition.

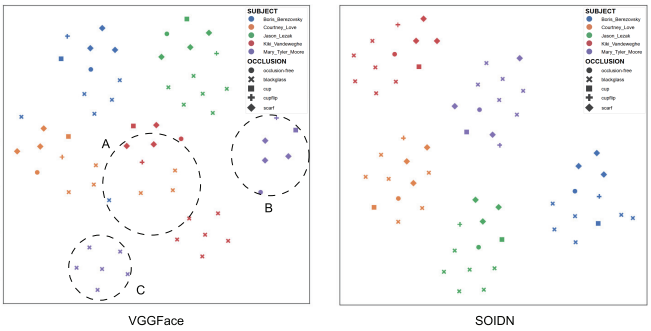

Fig. 12: t-SNE projection by embedding features of VGGFace and SOIDN. Please zoom in for better observation.

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 good 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 face recognition. 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

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

Deep Model	Training dataset	AR sunglasses	AR scarf
Inception-ResNet-V1 [52]	CASIA-Webface	88	91
Inception-ResNet-V1 [52]	VGGFace2	80	81
MobileFace (arcloss) [53]	MS-Celeb-1M [54]	83	94
SOIDN	CASIA-Webface*	**92**	**98**

Table 5 Rank-1 recognition accuracy (%) of the proposed approach SOIDN and state-of-the-art methods on AR dataset.

Deep Model	AR sunglasses	AR scarf
RPSM [28]	85	90
LMA [18]	96	94
PDSN [44]	**98**	**98**
Baseline	84	97
SOIDN	92	**98**

with SOIDN, but their results on occluded face dataset perform worse. Simply utilizing deep models trained for unconstrained face recognition cannot handle the occlusion properly, which further confirms the effectiveness of the network architecture of SOIDN. It is worth noting that we use the SSPP (single sample per subject) protocol for the experiments, which is the most challenging protocol as it requires only one image per subject for enrollment. Specifically, we enroll 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 occluded face recognition 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 occlusion segmentation. 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 occlusion segmentation network and the face recognition network of SOIDN explicitly consider occlusion location and occlusion content and are combined coherently and optimized in a simultaneous architecture, which ensures the robustness to occlusion variation. Besides, with the occlusion mask adaptor block the occlusion segmentation task and the face recognition task can help each other to obtain occlusion-free face representation. While the other methods such as RPSM and LMA convert occluded face recognition into an image patch matching problem that cannot locate occlusion precisely and further degrades the recognition accuracy. The PDSN performs occlusion segmentation and face recognition sequentially and the imperfect segmentation will unavoidably impair face recognition.

5 Conclusion

The face recognition results on synthesized and realistic face datasets obtained by proposed simultaneous occlusion invariant deep network (SOIDN) are promising. In this paper, we propose to address occluded face recognition in a simplified yet well-motivated way. Specifically, an occlusion mask adaptor is designed as a bridge in SOIDN, which is motivated by the phenomenon that corrupted features by occlusion are traceable within an occlusion segmentation network. We use classic VGG16 network as FR network branch, but any other advanced networks can be incorporated in the proposed framework for better performance. To our best knowledge, this work is the first to combine face recognition network and occlusion segmentation network coherently and optimize in 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.

6 References

1 Best-Rowden, L. and Jain, A.K.: 'Longitudinal study of automatic face recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **40**(1), pp. 148–162.

2 Masi, I., Wu, Y., Hassner, T. and Natarajan, P.: 'Deep face recognition: A survey' 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI), 2018, pp. 471–478.

3 Rasti, S., Yazdi, M. and Masnadi-Shirazi, M.A.: 'Biologically inspired makeup detection system with application in face recognition', *IET Biometrics*, 2018, **7**(6), pp. 530–535.

4 Ge, S., Li, J., Ye, Q. and Luo, Z.: 'Detecting masked faces in the wild with lle-cnns'. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2682–2690.

5 Wan, W. and Chen, J.: 'Occlusion robust face recognition based on mask learning'. Proceedings of the IEEE International Conference on Image Processing (ICIP), 2017, pp. 3795–3799.

6 Zhou, E., Cao, Z. and Yin, Q.: 'Naive-deep face recognition: Touching the limit of LFW benchmark or not?', arXiv preprint arXiv:1501.04690. (2015)

7 Osherov, E. and Lindenbaum, M.: 'Increasing cnn robustness to occlusions by reducing filter support' Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 550–561.

8 Long, J., Shelhamer, E. and Darrell, T.: 'Fully convolutional networks for semantic segmentation'. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

9 Zhang, W., Shan, S., Chen, X., et al.: 'Local Gabor binary patterns based on Kullback–Leibler divergence for partially occluded face recognition', *IEEE signal processing letters*, 2007, **14**(11), pp. 875–878.

10 Zhang, W., Shan, S., Gao, W., et al.: 'Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition'. Proceedings of the IEEE International Conference on Computer Vision, 2005, pp. 786–791.

11 Hua, G. and Akbarzadeh, A.: 'A robust elastic and partial matching metric for face recognition'. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2082–2089.

12 Cheheb, I., Al-Maadeed, N., Al-Madeed, S., et al.: 'Random sampling for patch-based face recognition'. Proc. 5th International Workshop on Biometrics and Forensics, 2017, pp. 1–5.

13 Gottumukkal, R. and Asari, V.K.: 'An improved face recognition technique based on modular PCA approach', *Pattern Recognition Letters*, 2004, **25**(4), pp. 429–436.

14 Leonardis, A. and Bischof, H.: 'Robust recognition using eigenimages', *Computer Vision and Image Understanding*, 2000, **78**(1), pp. 99–118.

15 Kim, J., Choi, J., Yi, J. and Turk, M.: 'Effective representation using ICA for face recognition robust to local distortion and partial occlusion', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(12), pp. 1977–1981.

16 Hotta, K.: 'Robust face recognition under partial occlusion based on support vector machine with local Gaussian summation kernel', *Image and Vision Computing*, 2008, **26**(11), pp. 1490–1498.

17 Seo, J. and Park, H.: 'A robust face recognition through statistical learning of local features'. Proceedings of International Conference on Neural Information Processing, 2011, pp. 335–341.

18 McLaughlin, N., Ming, J. and Crookes, D.: 'Largest matching areas for illumination and occlusion robust face recognition', *IEEE Transactions on Cybernetics*, 2017, **47**(3), pp. 796–808.

19 Wright, J., Yang, A.Y., Ganesh, A., et al.: 'Robust face recognition via sparse representation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(2), pp. 210–227.

20 Zhou, Z., Wagner, A., Mobahi, H., et al.: 'Face recognition with contiguous occlusion using markov random fields'. Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 1050–1057.

21 Yang, M., Zhang, L., Shiu, S.C., et al.: 'Gabor feature based robust representation and classification for face recognition with Gabor occlusion dictionary', *Pattern Recognition*, 2013, **46**(7), pp. 1865–1878.

22 Li, Y. and Feng, J.: 'Reconstruction based face occlusion elimination for recognition', *Neurocomputing*, 2013, **101**, pp. 68–72.

23 Ou, W., You, X., Tao, D., et al.: 'Robust face recognition via occlusion dictionary learning', *Pattern Recognition*, 2014, **47**(4), pp. 1559–1572.

24 Luan, X., Fang, B., Liu, L., et al.: 'Extracting sparse error of robust PCA for face recognition in the presence of varying illumination and occlusion', *Pattern Recognition*, 2014, **47**(2), pp. 495–508.

25 Zhao, S. and Hu, Z.P.: 'A modular weighted sparse representation based on fisher discriminant and sparse residual for face recognition with occlusion' *Information*

Processing Letters, 2015, **115**(9), pp. 677–683.
26 Yu, Y.F., Dai, D.Q., Ren, C.X., et al.: 'Discriminative multi-scale sparse coding for single-sample face recognition with occlusion', *Pattern Recognition*, 2017, **66**, pp. 302–312.
27 Wu, C.Y. and Ding, J.J.: 'Occluded face recognition using low-rank regression with generalized gradient direction', *Pattern Recognition*, 2018, **80**, pp. 256–268.
28 Weng, R., Lu, J., and Tan, Y. P.: 'Robust point set matching for partial face recognition', *IEEE Transactions on Image Processing*, 2016, **25**, pp. 1163–1176.
29 Chen, Z., Xu, T. and Han, Z.: 'Occluded face recognition based on the improved SVM and block weighted LBP'. Proc. International Conference on Image Analysis and Signal Processing, 2011, pp. 118–122.
30 Min, R., Hadid, A. and Dugelay, J.L.: 'Improving the recognition of faces occluded by facial accessories'. Proceedings of Face and Gesture, 2011, pp. 442–447.
31 Oh, H.J., Lee, K.M. and Lee, S.U.: 'Occlusion invariant face recognition using selective local non-negative matrix factorization basis images' *Image and Vision computing*, 2008, **26**(11), pp. 1515–1523.
32 Neo, H.F., Teo, C.C. and Teoh, A.B.: 'Development of partial face recognition framework' Proceedings of Computer Graphics, Imaging and Visualization, 2010, pp. 142–146.
33 Lv, J.J., Shao, X.H., Huang, J.S., Zhou, X.D., et al.: 'Data augmentation for face recognition', *Neurocomputing*, 2017, **230**, pp. 184–196.
34 Trigueros, D.S., Meng, L. and Hartnett, M.: 'Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss', *Image and Vision Computing*, 2018, **79**, pp. 99–108.
35 Cen, F. and Wang, G.: 'Dictionary representation of deep features for occlusion-robust face recognition', *IEEE Access*, 2019, **7**, pp. 26595–26605.
36 Cheng, L., Wang, J., Gong, Y., et al.: 'Robust deep auto-encoder for occluded face recognition'. Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 1099–1102.
37 Zhao, F., Feng, J., Zhao, J., et al.: 'Robust LSTM-Autoencoders for face de-occlusion in the wild' *IEEE Transactions on Image Processing*, 2018, **27**(2), pp. 778–790.
38 Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: 'Generative adversarial nets'. Proceedings of Advances in neural information processing systems, 2014, pp. 2672–2680.
39 Chen, Y.A., Chen, W.C., Wei, C.P., et al.: 'Occlusion-aware face inpainting via generative adversarial networks'. Proceedings of the IEEE International Conference on Image Processing (ICIP), 2017, pp. 1202–1206.
40 Chen, X., Qing, L., He, X., et al.: 'From eyes to face synthesis: a new approach for human-centered smart surveillance', *IEEE Access*, 2018, **6**, pp. 14567–14575.
41 Hu, B., Yang, W. and Ren, M.,: 'Unsupervised Eyeglasses Removal in the Wild', arXiv preprint arXiv:1909.06989. (2019)
42 Ge, S., Li, C., Zhao, S., et al.: 'Occluded Face Recognition in the Wild by Identity-Diversity Inpainting' *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.
43 Xia, Y., Zhang, B. and Coenen, F.: 'Face occlusion detection based on multi-task convolution neural network'. Proceedings of Fuzzy Systems and Knowledge Discovery, 2015, pp. 375–379.
44 Song, L., Gong, D., Li, Z., et al.: 'Occlusion Robust Face Recognition Based on Mask Learning with Pairwise Differential Siamese Network' Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 773–782.
45 Simonyan, K. and Zisserman, A.: 'Very deep convolutional networks for large-scale image recognition', arXiv preprint arXiv:1409.1556. (2014)
46 Saito, S., Li, T. and Li, H.: 'Real-time facial segmentation and performance capture from rgb input' European Conference on Computer Vision, 2016, pp. 244–261.
47 Yi, D., Lei, Z., Liao, S., et al.: 'Learning face representation from scratch', arXiv preprint arXiv:1411.7923. (2014)
48 Huang, G.B., Mattar, M., Berg, T., et al.: 'Labeled faces in the wild: A database for studying face recognition in unconstrained environments', 2008.
49 Aleix, M. and Robert, B.: 'The AR Face Database', CVC Tech. Rep, **24**, 1998.
50 Zhang, K., Zhang, Z., Li, Z., et al.: 'Joint face detection and alignment using multitask cascaded convolutional networks', *IEEE Signal Processing Letters*, 2016, **23**(10), pp. 1499–1503.
51 Parkhi, O.M., Vedaldi, A. and Zisserman, A.: 'Deep face recognition', 2015.
52 Szegedy, C., Ioffe, S., Vanhoucke, V., et al.: 'Inception-v4, inception-resnet and the impact of residual connections on learning'. arXiv preprint arXiv:1602.07261. (2016).
53 Chen, S., Liu, Y., Gao, X., et al.: 'Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices' Proceedings of Chinese Conference on Biometric Recognition, 2018, pp. 428–438.
54 Guo, Y., Zhang, L., Hu, Y., et al.: 'Ms-celeb-1m: A dataset and benchmark for large-scale face recognition' European Conference on Computer Vision, 2016, pp. 87–102.
55 Zheng, W., Gou, C. and Wang, F.Y.: 'A novel approach inspired by optic nerve characteristics for few-shot occluded face recognition' *Neurocomputing*, 2020, **376**, pp. 25–41.
56 Zhang, J., Wu, X., Hoi, S.C., et al.: 'Feature agglomeration networks for single stage face detection' *Neurocomputing*, 2020, **380**, pp. 180–189.