MSML: Enhancing Occlusion-Robustness by Multi-Scale Segmentation-Based Mask Learning for Face Recognition

Ge Yuan, Huicheng Zheng*, Jiayu Dong
School of Computer Science and Engineering, Sun Yat-sen University
Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
Guangdong Key Laboratory of Information Security Technology
zhenghch@mail.sysu.edu.cn

Abstract

In unconstrained scenarios, face recognition remains challenging, particularly when faces are occluded. Existing methods generalize poorly due to the distribution distortion induced by unpredictable occlusions. To tackle this problem, we propose a hierarchical segmentation-based mask learning strategy for face recognition, enhancing occlusion-robustness by integrating segmentation representations of occlusion into face recognition in the latent space. We present a novel multi-scale segmentation-based mask learning (MSML) network, which consists of a face recognition branch (FRB), an occlusion segmentation branch (OSB), and hierarchical elaborate feature masking (FM) operators. With the guidance of hierarchical segmentation representations of occlusion learned by the OSB, the FM operators can generate multi-scale latent masks to eliminate mistaken responses introduced by occlusions and purify the contaminated facial features at multiple layers. In this way, the proposed MSML network can effectively identify and remove the occlusions from feature representations at multiple levels and aggregate features from visible facial areas. Experiments on face verification and recognition under synthetic or realistic occlusions demonstrate the effectiveness of our method compared to state-of-the-art methods.

Introduction

Deep convolutional networks have achieved great success in extracting discriminative features for face recognition. Many related architectures (Parkhi, Vedaldi, and Zisserman 2015; Schroff, Kalenichenko, and Philbin 2015; Li et al. 2020; Ding et al. 2020; Yu et al. 2020), loss functions (Wang et al. 2018; Deng et al. 2019; Wen et al. 2016), and datasets (Martínez and Benavente 1998; Huang et al. 2007; Yi et al. 2014; Kemelmacher-Shlizerman et al. 2016; Liu et al. 2015; Ding et al. 2020; Geng et al. 2020; Chen et al. 2020) have been proposed. However, in realistic unconstrained scenarios where faces could be occluded, most existing deep models are not sufficiently robust.

Previously, straightforward solutions have been proposed in (Liu et al. 2016; Trigueros, Meng, and Hartnett 2018; Zhong et al. 2020) to train deep models with occluded images to improve the robustness. Osherov et al. (Osherov and Lindenbaum 2017) proposed to constrain the filter support of deep networks to improve robustness. Although these solutions can recover some performance on occluded samples, the discriminative ability of deep models on non-occluded samples is suffered. In essence, these methods did not appropriately handle the distribution distortion between occluded and non-occluded samples in the embedding space.

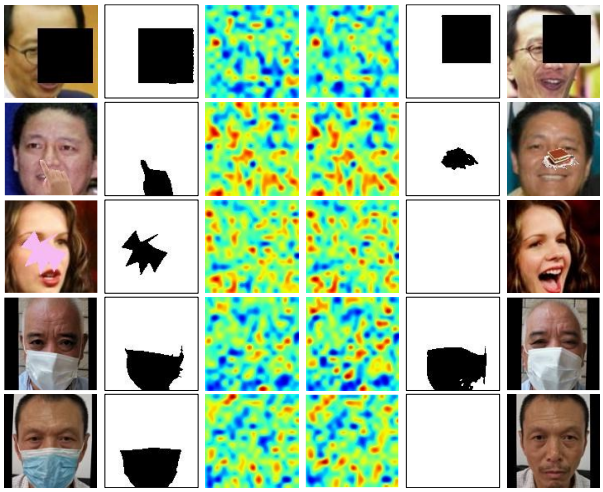


Figure 1: Occlusion-robust face representations through multi-scale segmentation-based mask learning. Column 1 & 6: Input faces with various possible occlusions. Column 2 & 5: Occlusions detected by the proposed occlusion segmentation branch (OSB) in our method. Column 3 & 4: Occlusion-robust face representations learned by our MSML network. MSML can handle various occlusion cases, *i.e.* none occlusion, geometric shapes, synthetically added objects, realistic face masks. (Best viewed in color.)

Distribution distortion comprises missing responses and mistaken responses. Missing responses mean that feature extractors cannot capture valid facial parts from occluded faces. Mistaken responses mean that unexpected responses induced by occlusions occur in the output space. These two error responses increase intra-class distance and decrease inter-class distance. To tackle the missing response prob-

*Corresponding author.

lem, some GAN-based methods have been proposed to reconstruct the occluded parts (Li et al. 2017; Xie et al. 2018; Ren et al. 2019). Despite their performance in recovering visual patterns, the recognition accuracy is limited due to the challenge of preserving identity (Mathai, Masi, and AbdAlmageed 2019). How to eliminate the mistaken responses becomes the focus problem for occluded face recognition.

Learning masks is a key idea to eliminate the mistaken responses and improve the robustness to occlusion. Such a strategy reduces the negative effect of occlusions in feature extraction (Wan and Chen 2017; Song et al. 2019). Once the mask of occlusion is obtained, pixels that may lead to mistaken responses can be excluded from feature extraction and a more reasonable embedding space can be obtained. However, previous methods only adopt mask learning in a middle layer. Due to challenges in real-world applications, the generated mask may still contain erroneous responses, which further result in mistaken responses in the subsequent layers. As the features flow deeper, the mistaken responses may accumulate in the latent space, leading to more severe distribution distortion in the embedding space. In addition, the shallower layers without mask learning still suffer from distribution distortion, with mistaken responses primarily distributed around the locations of occlusions due to the small receptive fields. In the deep layers, the features get entangled. The simple masks that lack deep semantics cannot block mistaken responses effectively. To this end, for the facial features at different layers, we generate latent masks of different scales.

In this work, we propose a multi-scale segmentation-based mask learning (MSML) face recognition network which can alleviate the distribution distortion hierarchically and boost the performance of occluded face recognition. The proposed MSML network consists of a face recognition branch (FRB), an occlusion segmentation branch (OSB), and hierarchical feature masking (FM) operators. The OSB consists of an encoding stage and a decoding stage, which aims to predict the precise occluded pixels, as shown in Figure 1. Hierarchical segmentation representations of occlusion are generated in all decoding stages of the OSB. After fed to the hierarchical FM operators, these segmentation representations of occlusion will be transformed to the multi-scale latent masks. Finally, the hierarchical FM operators use the latent masks to purify the contaminated facial features in all corresponding layers. To realize optimal transforming and purifying process, we experimentally explore various architectures of the FM operators. Consequently, the purifying process can effectively eliminate the mistaken responses induced by occlusions in the embedding space. Figure 1 visualizes the extracted 512-D features in a 2D space, which shows that our method can extract occlusion-robust face representations under various occlusion cases. The major contributions of our work are three-fold:

1) We propose a deep occlusion-robust face recognition framework with multi-scale segmentation-based mask learning[1]. The end-to-end framework does not require extra manual annotations or substantial samples of occlusion.

2) We present the hierarchical feature masking (FM) operators to transform the hierarchical segmentation representations into the latent masks and purify the contaminated facial features at the corresponding layers effectively.

3) The proposed method achieves state-of-the-art robustness to both synthetic and realistic occlusions on various benchmarks (Huang et al. 2007; Martínez and Benavente 1998; Kemelmacher-Shlizerman et al. 2016; Li et al. 2020).

[1]The code is available at: https://github.com/ygtxr1997/MSML.

Related Works

Traditional methods. Various traditional machine learning methods have been developed to tackle occlusion encountered in face recognition. Based on sparse representation, Wright et al. (Wright et al. 2008) reconstructed clean faces through a sparse linear combination of gallery images. With sparsity properly harnessed, the proposed framework can be robust to occluded faces. Stringface (Chen and Gao 2010) matched two faces through a string-to-string scheme to find the most discriminative substrings. Some works aimed to recognize partial faces based on feature set matching (Weng et al. 2013; Weng, Lu, and Tan 2016). Following the idea of sparse representation (Wright et al. 2008), McLaughlin et al. (McLaughlin, Ming, and Crookes 2016) proposed to find the largest matching area (LMA) in testing images that can be represented by training images. Yang et al. (Yang et al. 2016) converted the rank minimization problem into the nuclear norm minimization problem for optimization. Laplacian-uniform mixture-driven iterative robust coding (LUMIRC) (Zheng et al. 2020) modeled the distribution of the reconstruction residuals with a Laplacian-uniform mixture function. Although their theoretical contributions are sound, the practical values are limited by the complexity in real-world scenarios.

Deep learning-based methods. Recently, deep learning has been widely used in occlusion-robust face recognition. Wan et al. (Wan and Chen 2017) proposed MaskNet for learning different weights for spatial locations of the feature maps in the medial layer of a deep face network. Song et al. (Song et al. 2019) proposed learning masks by comparing feature maps extracted from an occluded face and its counterpart through the pairwise differential siamese network (PDSN). A Light CNN framework was developed in (Wu et al. 2018) to learn a robust face representation on noisily labeled datasets by introducing Max-Feature-Map (MFM), which is able to separate noisy and informative signals. For the face de-occlusion task, Zhao et al. (Zhao et al. 2017) proposed a LSTM-autoencoder model to detect occlusions and restore natural faces. Ding et al. (Ding et al. 2020) collected two datasets named MFV and MFI for evaluating masked face recognition models and proposed a latent part detection (LPD) model to locate the latent facial part. Li et al. (Li et al. 2020) proposed a de-occlusion distillation framework showing the efficacy of the amodal completion mechanism. Geng et al. (Geng et al. 2020) introduced an identity aware mask GAN (IAMGAN) to obtain sufficient training data in masked face recognition and proposed a domain constrained ranking (DCR) loss to tackle the large intra-class variation between masked faces and full faces. Previous single-layer mask learning methods (Wan and Chen 2017; Song et al.

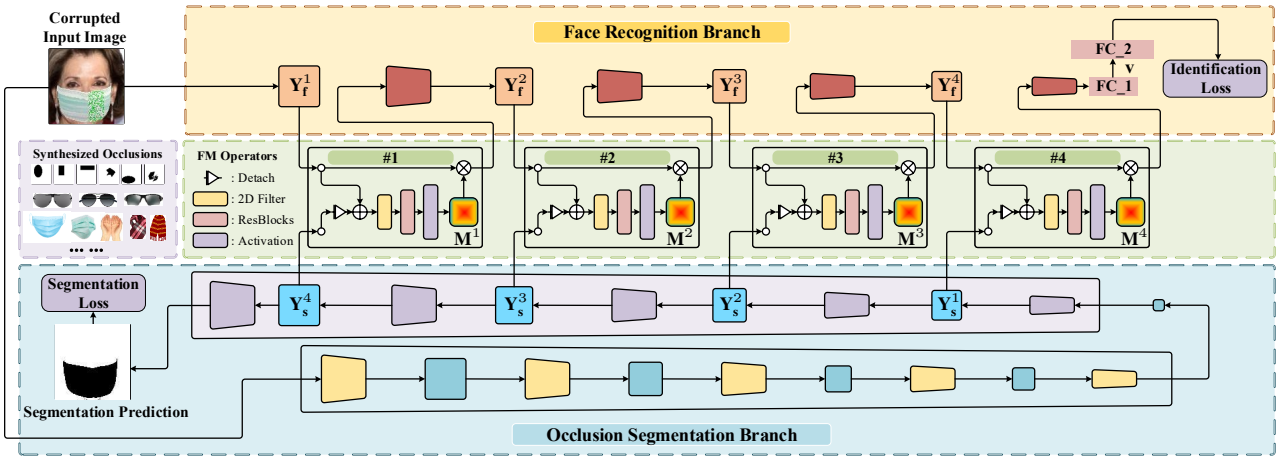


Figure 2: Hierarchical feature masking (FM) operators bridge the FRB and the OSB. Sharing the similar architecture, each FM operator receives the facial features and occlusion segmentation representations at the corresponding layer, and outputs purified facial features.

2019) and attention-based method LPD (Ding et al. 2020) only eliminate noises in the first processing layer or a middle layer, which cannot sufficiently remove the noise and suffer from the distribution distortion in the embedding space. In contrast, we present multi-scale segmentation-based mask learning to generate optimal latent masks and eliminate mistaken responses at multiple levels to correct facial semantic representations.

Proposed Method

Network Architecture

Figure 2 shows the overall framework of the proposed multi-scale segmentation-based mask learning (MSML) network, where the FRB and the OSB are bridged by k FM operators. In our method, k is set to 4, which is consistent with the number of the stages (excluding the stem stage) in the FRB.

FRB is a regular face recognition network supervised by an identification loss. An embedded facial feature vector $\mathbf{v}$ is extracted by the first fully-connected layer (FC_1) of the FRB. For the OSB, we adopt an encoder-decoder structure (Long, Shelhamer, and Darrell 2015) to generate hierarchical occlusion segmentation representations. The OSB is supervised by a semantic segmentation loss. Each FM operator transforms the segmentation representations into the latent mask and purifies the contaminated facial features at the corresponding layer.

Previous mask learning methods (Wan and Chen 2017; Song et al. 2019) only adopt a single layer of segmentation representation to generate a single scale of mask. But in real-world challenging situations, facial features at different layers need masks of different scales. Embracing this hierarchical principle, our method eliminates mistaken responses in multiple layers with multi-scale latent masks guided by segmentation representations.

In the training stage, one of the various types of occlusions as shown in Figure 2 is synthetically added to the original input face. Supervised by the binary mask of the synthesized occlusion, the decoder of the OSB can generate hierarchical occlusion segmentation representations $\mathbb{Y}_\mathbf{s} = \{\mathbf{Y}_\mathbf{s}^1, \mathbf{Y}_\mathbf{s}^2, ..., \mathbf{Y}_\mathbf{s}^k\}$, where $\mathbf{Y}_\mathbf{s}^j$ is the output of the j-th transpose convolutional layer. Similarly, the multi-layer facial features generated by the FRB can be denoted as $\mathbb{Y}_\mathbf{f} = \{\mathbf{Y}_\mathbf{f}^1, \mathbf{Y}_\mathbf{f}^2, ..., \mathbf{Y}_\mathbf{f}^k\}$, where $\mathbf{Y}_\mathbf{f}^i$ is output at the i-th convolution stage FRB_i (excluding the stem stage). After fed to the hierarchical FM operators, the segmentation representations are converted to the multi-scale latent masks. Subsequently, the generated latent masks can purify the contaminated facial features, alleviating the mistaken response problem at multiple layers. In our method, $\mathbf{Y}_\mathbf{f}^i$ and $\mathbf{Y}_\mathbf{s}^j$ share the same height h and width w, where $i = 1, 2, ..., k$ and $j = k + 1 - i$.

Occlusion Segmentation Branch

The OSB is a fully convolutional network with an encoder-decoder structure (Long, Shelhamer, and Darrell 2015). Supervised by the occlusion segmentation loss, the OSB can learn to predict pixel-wise occlusions. The hierarchical occlusion segmentation representations $\{\mathbf{Y}_\mathbf{s}^1, \mathbf{Y}_\mathbf{s}^2, ..., \mathbf{Y}_\mathbf{s}^k\}$ are generated by the decoder of the OSB. $\mathbf{Y}_\mathbf{s}^i$ output by the deeper transpose convolutional layers contains more precise location information about the occlusion. In $\mathbf{Y}_\mathbf{f}^i$ extracted from a shallower layer, the mistaken responses primarily distribute near the location of the occlusion due to the smaller receptive field. In addition, $\mathbf{Y}_\mathbf{f}^i$ at a shallower layer lacks semantic information. $\mathbf{Y}_\mathbf{s}^i$ at a deeper layer complements $\mathbf{Y}_\mathbf{f}^i$ at a shallower layer due to larger receptive field and more semantic information. This means it is more reasonable to use deeper occlusion representations to generate latent masks for purifying shallower facial features. So we bridge the FRB

and OSB in a reverse fashion. Without requiring extra labor-consuming annotations, we adopt the binary masks of synthetically added occlusions as the training labels of the OSB.

Compared to the PDSN (Song et al. 2019) which uses an independently trained fully convolutional network, our method uses an end-to-end multitask learning framework where the OSB and the FRB are trained simultaneously. Moreover, the single-layer feature masks in MaskNet (Wan and Chen 2017) and PDSN are generated from manually predefined coarse-grained $N \times N (N < 10)$ grids, which leads to poor generalization on diverse unseen occlusions. In contrast, with the pixel-wise predicting ability of the OSB, the FM operators can generate fine-grained latent masks which adapt to the shape of occlusions. Also, the FM operators can adapt the levels of generated latent masks to those of $\{\mathbf{Y}_\mathbf{s}^1, \mathbf{Y}_\mathbf{s}^2, ..., \mathbf{Y}_\mathbf{s}^k\}$, which contain different levels of semantic information.

Feature Masking Operator

$\mathbb{Y}_\mathbf{s}$ are fed to the hierarchical FM operators which can generate multi-scale latent masks $\mathbb{M} = \{\mathbf{M}^1, \mathbf{M}^2, ..., \mathbf{M}^k\}$ and output purified facial features $\mathbb{Z}_\mathbf{f} = \{\mathbf{Z}_\mathbf{f}^1, \mathbf{Z}_\mathbf{f}^2, ..., \mathbf{Z}_\mathbf{f}^k\}$. Let FM_i denote the i-th FM operator in the MSML network. The facial feature $\mathbf{Y}_\mathbf{f}^i$ in each layer is extracted from the purified output $\mathbf{Z}_\mathbf{f}^{i-1}$ of FM_{i-1} as shown in Figure 2.

The k FM operators share the similar architecture and generate multi-scale $\mathbb{M}$. Specifically, the generation of $\mathbf{M}^i$ can be formulated as:

$$\mathbf{M}^i = F(\Phi_s(\mathbf{Y}_\mathbf{s}^j; \mathbf{W}_i)), \tag{1}$$

where $\Phi_s(\cdot; \mathbf{W}_i)$ is a 2D filter with the weight matrix $\mathbf{W}_i$ and kernel size $s \times s$, $i = 1, 2, ..., k$, and $j = k + 1 - i$, $F(\cdot)$ indicates the mask scheme function (binarization, sigmod, or tanh). Then the purified facial feature $\mathbf{Z}_\mathbf{f}^i$ is generated as:

$$\mathbf{Z}_\mathbf{f}^i = \mathbf{M}^i \circ \mathbf{Y}_\mathbf{f}^i, \tag{2}$$

where $\circ$ denotes Hadamard product, $i = 1, 2, ..., k$.

Since $\mathbf{Y}_\mathbf{f}^i$ provides auxiliary facial patterns which can help the generation of the latent masks, we concatenate $\mathbf{Y}_s^j$ and $\mathbf{Y}_f^i$ before transmitted to the 2D filter Φ_s. We also insert a residual learning module Θ_r which consists of r residual blocks (He et al. 2016) after the filter to adjust the complexity of FM operator. In this way, Equation 1 for the mask generation can be reformulated as:

$$\mathbf{M}^i = F(\Theta_r(\Phi_s([\mathbf{Y}_\mathbf{s}^j, \mathbf{Y}_\mathbf{f}^i]; \mathbf{W}_i))), \tag{3}$$

where $[\cdot]$ denotes the concatenation.

The detailed architecture of FM_i is shown in Figure 2. Considering the pixel labels of synthetic occlusions are sufficient to supervise the OSB in the training stage, we add a detach link before the concatenation to avoid the gradients of the FRB impacting the optimization of the OSB. In some generative adversarial networks (GANs) (Goodfellow et al. 2014), the detach link is used to avoid the gradients of the discriminator to be propagated to the generator.

Optimization

The proposed network can be trained through the joint optimization. Two losses are used for training: occlusion segmentation loss $\mathcal{L}_{occ}$, and face classification loss $\mathcal{L}_{cls}$. The total loss $\mathcal{L}_{total}$ can be formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{occ} \tag{4}$$

where λ is a weighing factor. Experimentally, we found that the training process of the proposed model is not sensitive to the value of λ. Therefore, λ is set to 1 for convenience. Cross-entropy loss or other SOTA face recognition losses (Wang et al. 2018; Deng et al. 2019) can be selected as $\mathcal{L}_{cls}$. Considering the continuation of realistic occlusions, we choose a consensus segmentation loss (Masi, Mathai, and AbdAlmageed 2020) as $\mathcal{L}_{occ}$.

Experiments

Implementation Details

Models. We select two state-of-the-art face recognition networks as the baselines: LightCNN-29 (L29) (Wu et al. 2018) and ArcFace-18 (A18) (Deng et al. 2019). The lightweight model L29 replaces the ReLU activation function with Max-out activation function and drops the Batch Normalization layers. A18 shares the similar structure with ResNet18 (He et al. 2016) and replaces the cross entropy loss with an Additive Angular Margin Loss. The MSML network using L29 (or A18) as the FRB is denoted as MSML(L29) (or MSML(A18)). Following the architecture of U-Net (Ronneberger, Fischer, and Brox 2015), the skip connections bridge the encoder and the decoder of the OSB. The encoder adopts ResNet18 as the backbone. The subsequent decoder comprises 5 transpose convolutional layers.

Face preprocessing. All faces are detected, aligned and cropped based on 5 landmarks with MTCNN (Zhang et al. 2016). The aligned faces are resized to 128×128 pixels for L29 and 112×112 for A18. Three types of occlusions (random connected geometric shapes, realistic objects collected from the web, and synthetic face masks rendered through 3D scheme (Zhu et al. 2016, 2015)), as shown in Figure 3, are synthetically added to the faces during training. All the baseline models and our models use the same augmentation scheme if there is no other specific instruction.

We use these occlusions for the following four reasons. First, a realistic occlusion is often simply connected (Masi, Mathai, and AbdAlmageed 2020). Simply connected geometric shapes contain this basic characteristic. Second, the selected real objects are common in real life, such as sunglasses, scarves, hands, fruits, and cups. Third, masked face recognition has received much attention in recent years (Li et al. 2020; Geng et al. 2020; Ding et al. 2020). We perform 3D synthetic face mask augmentation to make the models generalize well to the faces wearing masks. Finally, the various occlusions avoid the over-fitting of the OSB.

Training. L29 is pretrained on MS-Celeb-1M (Guo et al. 2016) and trained on CASIA-WebFace (Yi et al. 2014), while A18 is trained on MS1MV2 (Deng et al. 2019) from scratch. We employ stochastic gradient descent (SGD) as the

Figure 3: The training inputs and the occlusions detected by the OSB.

Figure 4: In 1:1 verification, the compared pair might contain block occlusions of different positions and sizes.

optimizer. The weight decay is set to 10^{-5} and the momentum is set to 0.9. For MSML(L29), the initial learning rate of FRB and OSB are 0.001 and 0.01 respectively and are divided by 3 every 15 epochs. For MSML(A18), the initial learning rate of FRB and OSB are 0.1 and 0.01 respectively and are divided by 10 at 11, 16, 21 epochs. We set the batch size of 64 for MSML(L29) and 512 for MSML(A18). The embedded feature vectors are 128-D for MSML(L29) and 512-D for MSML(A18).

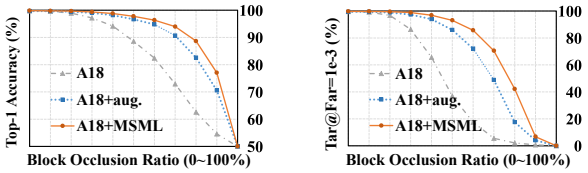
Testing. We conduct the experiments on multiple commonly used datasets as follows:

- The LFW dataset (Huang et al. 2007) contains 13,233 images from 5,749 identities and provides 3,000 matched image pairs and 3,000 mismatched image pairs for 1:1 verification testing.
- The MegaFace dataset (Kemelmacher-Shlizerman et al. 2016) includes over 1 million face images and provides a common testing benchmark consisting of a probe set called Facescrub and 1 million face distractors.
- The AR dataset (Martínez and Benavente 1998) contains over 4,000 face images of 126 subjects with variations in expression, illumination, and occlusion.
- The MFV dataset (Ding et al. 2020) contains over 6,000 pairs of realistic faces wearing masks and non-occluded faces. The face images vary in pose, illumination, and background.

Considering the randomness of synthetic occlusions in the evaluation on the LFW and MegaFace datasets, we repeat the experiments for 10 times and show the average results.

Experiments on the LFW Dataset

Following the testing protocol of LFW (Huang et al. 2007), we performed a 1:1 face verification evaluation. We occlude the testing images with random black blocks of 0% to 100% input size forming LFW_{block}. The random blocks vary in positions and sizes, as shown in Figure 4. The accuracy and the



(a) Accuracy (b) TAR(@FAR=1e-3)

Figure 5: The 1:1 verification results (%) on LFW_{block}. The proposed MSML(A18) achieves prominent occlusion-robustness.

Method	Data	LFW	MF1	MF1_{occ}
CenterFace (2016)	0.7M	99.28	65.49	-
CosFace (2018)	3.9M	99.73	77.11	-
TrunkCNN (2019)	0.5M	99.20	74.40	51.86
PDSN (2019)	0.5M	99.20	74.40	56.34
L29 (2018)	0.5M	99.33	71.31	54.63
L29+MSML	0.5M	99.40	75.35	62.53
A18 (2019)	5.8M	99.77	76.92	64.50
A18+MSML	5.8M	**99.83**	**79.63**	**68.33**

Table 1: The 1:1 verification results (%) on the LFW dataset and the 1:N identification results (%) on MF1 and MF1_{occ}.

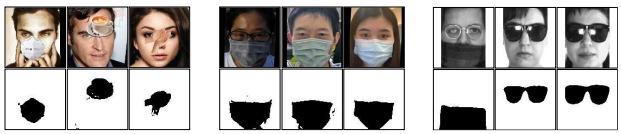
true acceptance rate when the false acceptance rate is below 1e-3 (TAR(@FAR=1e-3)) are reported in Figure 5 to compare the performance of the proposed MSML(A18), A18 using same augmentation scheme, and A18 without augmentation scheme. MSML(A18) achieves prominent occlusion-robustness. The results in Table 1 show that the proposed method does not compromise its performance on regular non-occluded face recognition tasks.

Experiments on the MegaFace Dataset

In the 1:N face identification experiments on the MegaFace dataset (Kemelmacher-Shlizerman et al. 2016), the Facescrub dataset merges with over 1,000,000 face distractors forming the large probe set (MF1). We follow (Song et al. 2019) and add realistic object occlusions to the probe set of MegaFace (MF1_{occ}). These realistic object occlusions differ from those used in the training stage. Figure 6 (a) shows some examples of the testing images and the predicted occlusions. As shown in Table 1, the performance of L29 and A18 drops after we add the occlusions. The proposed model shows strong robustness to synthetically occluded samples compared to PDSN.

Experiments on the AR Dataset

We verify the robustness of the proposed method to realistic disguise by conducting 1:N face identification experiments on the AR dataset (Martínez and Benavente 1998). Protocol 1 uses multiple images per subject in the gallery for identification. Protocol 2 uses only one gallery image per subject for identification, which is more challenging. All gallery images are holistic frontal face images. The probe set comprises the



(a) (b) (c)

Figure 6: The input images and the predicted occlusions on (a) MegaFace, (b) MFV, and (c) AR datasets.

Method	Sunglass	Scarf
NMR (2016)	96.90	73.50
LUMIRC (2020)	97.35	96.70
MLERPM (2013)	98.00	97.00
SCF-PKR (2013)	95.65	98.00
RPSM (2016)	96.00	97.66
MaskNet (2017)	90.90	96.70
Trunk CNN (2019)	98.19	99.72
PDSN (2019)	99.72 +1.53	100.00 +0.28
L29	98.02	98.78
L29+MSML	**99.84** +1.82	**100.00** +1.22

(a) Protocol 1

Method	Sunglass	Scarf
VGGFace2 (2018)	88.30	78.20
ArcFace (2019)	85.50	76.40
RPSM (2016)	84.84	90.16
Stringface (2010)	82.00	92.00
LMA (2016)	96.30	93.70
DDF (2020)	98.00	94.10
Trunk CNN (2019)	95.14	96.53
PDSN (2019)	98.19 +3.05	98.33 +1.80
L29	96.44	96.76
L29+MSML	**98.80** +2.36	**99.37** +2.61

(b) Protocol 2

Table 2: Face identification performance on the AR dataset. Sunglass and scarf denote the probe faces are occluded by realistic sunglasses and scarves, respectively.

faces occluded by scarves or sunglasses. The experimental results are shown in Table 2. The proposed method achieves the best robustness with scarf and sunglass occlusions under two testing protocols. Whether using one or multiple gallery images of each subject, our method improves the performance beyond that of the baseline model. Compared to PDSN, the proposed method yields better performance, even if our baseline L29 underperforms that (Trunk CNN) of PDSN.

Experiments on the MFV Dataset

We conduct the 1:1 face verification experiment on the MFV dataset to compare the proposed MSML network and the state-of-the-art models, as shown in Table 3. To calculate TAR(@FAR=1e-3), we use a larger number of testing pairs compared to LPD (Ding et al. 2020) (6,000 vs. 400). The results show MSML is robust to unseen realistic occlusions and achieves impressive performance. Compared with A18, the proposed MSML boosts the accuracy by 3.5% and

Method	#Pairs	Acc.	TAR
CosFace (2018)	400	86.86	-
PDSN (2019)	400	87.40	-
R50 (2020)	400	95.37	-
LPD (2020)	400	97.94	-
A18 (2019)	6000	95.40	81.10
A18+MSML	6000	**98.90**	**91.93**

Table 3: The 1:1 verification results (%) on the MFV dataset.

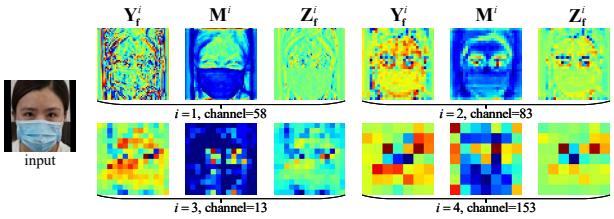


Figure 7: The visualized intermediate feature maps.

TAR(@FAR=1e-3) by 10.83%, achieving the highest performance among all the methods.

Visualization

To verify our analysis about distribution distortion and show the efficacy of our method, we conduct a series of visualization experiments.

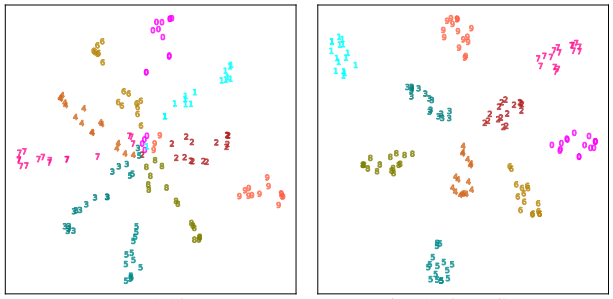
Embedded feature vector. After fed with a face under one of various occlusion cases, the MSML network generates an embedded feature vector. As shown in Figure 1, we visualize the 512-D embedded feature vectors of our MSML(A18) by normalizing the values into a heat map. The visualization results show that our method can extract occlusion-invariant embedded feature vectors.

Feature maps in the latent space. We visualize the feature maps at certain channels of $\mathbf{Y}_\mathbf{f}^i$, $\mathbf{M}^i$, and $\mathbf{Z}^i$ in Figure 7. The hierarchical FM operators generate multi-scale latent masks $\mathbf{M}^i$ to purify the contaminated facial features $\mathbf{Y}_\mathbf{f}^i$ and obtain the cleaned facial features $\mathbf{Z}^i$. The FM operators substantially eliminate the mistaken responses induced by the occlusion.

Distribution of the embedding space. As shown in Figure 8, we map the 512-D embedding space of 10 identities under various occlusion cases onto the 2D space through t-SNE (van der Maaten and Hinton. 2008). Each point denotes an embedded feature vector extracted from a face under one of the various occlusions. The points of the same color share the same identity. Compared to the baseline A18, the proposed MSML(A18) extracts embedded features with smaller intra-class distance and bigger inter-class distance. The t-SNE results demonstrate the distribution distortion is alleviated with our method.

Ablation Study

Variations of FM operators. We compare the 1:1 face verification accuracy using the FM operators with different ar-



(a) A18 (b) A18+MSML

Figure 8: Distributions of the embedding space mapped onto 2D space. (Best viewed in color.)

Model	Layers	Accuracy(%)
v0: L29 (2018)	1,2,3,4	88.37
v1: (Seg, 1, 0, Bin)	1,2,3,4	93.00
v2: (Seg, 1, 0, Tan)	1,2,3,4	93.88
v3: (Seg, 1, 0, Sig)	1,2,3,4	94.43
v4: (Seg, 3, 0, Sig)	1,2,3,4	94.13
v5: (Seg, 1, 1, Sig)	1,2,3,4	94.96
v6: (Seg, 1, 2, Sig)	1,2,3,4	95.07
v7: (Cat, 1, 0, Sig)	1,2,3,4	95.27
v8: (Cat, 3, 0, Sig)	1,2,3,4	95.87
v9: (Cat, 3, 1, Sig)	1,2,3,4	96.20
v10: (Cat, 3, 2, Sig)	1,2,3,4	**96.30**
v11: (Cat, 3, 2, Sig)	2,3,4	91.73
v12: (Cat, 3, 2, Sig)	1,3,4	95.40
v13: (Cat, 3, 2, Sig)	1,2,4	95.35
v14: (Cat, 3, 2, Sig)	1,2,3	90.27

Table 4: Variations of the hierarchical FM operators.

chitectures on the LFW dataset. In the evaluation, we occlude the samples by random blocks ranging from 1% to 40% of the input size. The architecture of the FM operators is denoted as (T, S, R, F). Specifically, $T = Seg$ indicates $\mathbf{Y}_\mathbf{s}^j$ is not concatenated with $\mathbf{Y}_\mathbf{f}^i$, while $T = Cat$ indicates the opposite case. S indicates the kernel size of Φ_s. R indicates the number of the ResBlocks in Θ_r inserted after the filter. $F \in \{Bin, Sig, Tan\}$ indicates the activation function, where Bin, Sig and Tan denotes the binarization, sigmoid, and tanh function respectively.

As shown in Table 4, compared to the occlusion-free evaluation result (99.33%) in Table 1, the baseline model v0 exhibits a considerable performance drop (-10.96%). But the accuracies of the models v1 $\sim$ v10 are not less than 93%. The results of models v1, v2 and v3 show that the sigmoid activation function yields better results than binarization and tanh. The sigmoid activation function generates a soft mask rather than a hard one. The improvement of the models v3 $\sim$ v6 highlights the efficacy of Θ_r. The models with concatenation (v7 $\sim$ v10) show higher accuracy compared to those without concatenation (v3 $\sim$ v6). The model v10 with 3×3 2D filter kernel size and two residual blocks achieves the highest accuracy.

Mask learning on different layers. The results of models v10 $\sim$ v14 in Table 4 verify the effectiveness of the hierarchical architecture of MSML. Erasing any one of the FM operators causes a performance degradation.

Model	MFV	LFW_{poly}		MF1_{occ}	
	Acc.	IOU	Acc.	IOU	Acc.
A18 (w/o aug.)	91.77	-	99.47	-	59.60
A18	95.40	-	99.53	-	64.50
+MSML_{O18}	**98.90**	97.67	**99.82**	93.47	68.33
+MSML_{O34}	98.43	**97.90**	99.72	**93.80**	**69.04**

Table 5: Variations of the OSB. Whether using R-18 or R-34 in OSB improves the performance under various occlusions.

Module	#Params	GFLOPs
ArcFace-R18	24.02M	2.60
OSB-R18	11.43M	0.94
FM@Layer1	0.07M	0.23
FM@Layer2	0.28M	0.22
FM@Layer3	1.06M	0.21
FM@Layer4	3.00M	0.15

Table 6: The #Params and GFLOPs of various modules of MSML(A18).

Variations of the OSB. Table 5 compares the performance of the models with different encoders in the OSB. We adopt MFV, LFW occluded by random polygons (LFW_{poly}), and MF1_{occ} for evaluation. The OSB using ResNet-34 (O34) shows a higher IOU than the OSB using ResNet-18 (O18). Adopting stronger backbone as the encoder of the OSB can further improve the performance on MF1_{occ} for 1:N identification, which is harder than 1:1 verification. Overall, no matter which backbone is used as the encoder of the OSB, our methods improve the recognition performance.

Model complexity. We provide the #params and FLOPs of the modules of MSML(A18) in Table 6. The model size and FLOPs are largely determined by the FRB and OSB.

Conclusion

In this paper, we propose a multi-scale segmentation-based mask learning (MSML) face recognition network, which tackles the mistaken response problem for occluded face recognition. With the guidance of hierarchical occlusion segmentation representations generated by the occlusion segmentation branch (OSB), the feature masking (FM) operators can generate multi-scale latent masks to purify the contaminated facial features in the face recognition branch (FRB). The purifying process can substantially eliminate mistaken responses induced by occlusions. In this way, the MSML network can effectively identify and remove the occlusions from feature representations and aggregate features from visible facial areas. Experimental results show that our method outperforms other methods under various occlusion scenarios while achieving competitive performance on regular face recognition tasks. The visualized results demonstrate the alleviation of distribution distortion with our method.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 61976231, U1611461, 61573387, 61172141) and the Guangdong Basic and Applied Basic Research Foundation (No. 2019A1515011869).

References

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. VGGFace2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face & Gesture Recognition*, 67–74.

Chen, T.; Pu, T.; Wu, H.; Xie, Y.; Liu, L.; and Lin, L. 2020. Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning. *arXiv preprint arXiv:2008.00923*.

Chen, W.; and Gao, Y. 2010. Recognizing partially occluded faces from a single sample per class using string-based matching. In *European Conference on Computer Vision*, 496–509.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Ding, F.; Peng, P.; Huang, Y.; Geng, M.; and Tian, Y. 2020. Masked face recognition with latent part detection. In *ACM International Conference on Multimedia*, 2281–2289.

Geng, M.; Peng, P.; Huang, Y.; and Tian, Y. 2020. Masked face recognition with generative data augmentation and domain constrained ranking. In *ACM International Conference on Multimedia*, 2246–2254.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2672–2680.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, 87–102.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Kemelmacher-Shlizerman, I.; Seitz, S. M.; Miller, D.; and Brossard, E. 2016. The MegaFace benchmark: 1 million faces for recognition at scale. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4873–4882.

Li, C.; Ge, S.; Zhang, D.; and Li, J. 2020. Look through masks: Towards masked face recognition with de-occlusion distillation. In *ACM International Conference on Multimedia*, 3016–3024.

Li, Y.; Liu, S.; Yang, J.; and Yang, M.-H. 2017. Generative face completion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3911–3919.

Liu, H.; Duan, H.; Cui, H.; and Yin, Y. 2016. Face recognition using training data with artificial occlusions. In *IEEE Visual Communications and Image Processing Conference*, 1–4.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, 3730–3738.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Martínez, A.; and Benavente, R. 1998. The AR face database. Technical Report 24, Computer Vision Center at the U.A.B.

Masi, I.; Mathai, J.; and AbdAlmageed, W. 2020. Towards learning structure via consensus for face segmentation and parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5507–5517.

Mathai, J.; Masi, I.; and AbdAlmageed, W. 2019. Does generative face completion help face recognition? In *International Conference on Biometrics*, 1–8.

McLaughlin, N.; Ming, J.; and Crookes, D. 2016. Largest matching areas for illumination and occlusion robust face recognition. *IEEE Transactions on Cybernetics*, 47(3): 796–808.

Osherov, E.; and Lindenbaum, M. 2017. Increasing CNN robustness to occlusions by reducing filter support. In *IEEE International Conference on Computer Vision*, 550–561.

Parkhi, O. M.; Vedaldi, A.; and Zisserman, A. 2015. Deep face recognition. In *British Machine Vision Conference*, 1–12.

Ren, Y.; Yu, X.; Zhang, R.; Li, T. H.; Liu, S.; and Li, G. 2019. Structureflow: Image inpainting via structure-aware appearance flow. In *IEEE International Conference on Computer Vision*, 181–190.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 815–823.

Song, L.; Gong, D.; Li, Z.; Liu, C.; and Liu, W. 2019. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In *IEEE International Conference on Computer Vision*, 773–782.

Trigueros, D. S.; Meng, L.; and Hartnett, M. 2018. Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. *Image and Vision Computing*, 79: 99–108.

van der Maaten, L.; and Hinton., G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11): 2579–2605.

Wan, W.; and Chen, J. 2017. Occlusion robust face recognition based on mask learning. In *IEEE International Conference on Image Processing*, 3795–3799.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. CosFace: Large margin cosine loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5265–5274.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, 499–515.

Weng, R.; Lu, J.; Hu, J.; Yang, G.; and Tan, Y.-P. 2013. Robust feature set matching for partial face recognition. In *IEEE International Conference on Computer Vision*, 601–608.

Weng, R.; Lu, J.; and Tan, Y.-P. 2016. Robust point set matching for partial face recognition. *IEEE Transactions on Image Processing*, 25(3): 1163–1176.

Wright, J.; Yang, A. Y.; Ganesh, A.; Sastry, S. S.; and Ma, Y. 2008. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2): 210–227.

Wu, X.; He, R.; Sun, Z.; and Tan, T. 2018. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11): 2884–2896.

Xie, J.; Lu, Y.; Gao, R.; and Wu, Y. N. 2018. Cooperative learning of energy-based model and latent variable model via mcmc teaching. In *AAAI Conference on Artificial Intelligence*, 4292–4301.

Yang, J.; Luo, L.; Qian, J.; Tai, Y.; Zhang, F.; and Xu, Y. 2016. Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1): 156–171.

Yang, M.; Zhang, L.; Shiu, S. C.-K.; and Zhang, D. 2013. Robust kernel representation with statistical local features for face recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 24(6): 900–912.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*.

Yu, M.; Zheng, H.; Peng, Z.; Dong, J.; and Du, H. 2020. Facial expression recognition based on a multi-task global-local network. *Pattern Recognition Letters*, 131: 166–171.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10): 1499–1503.

Zhao, F.; Feng, J.; Zhao, J.; Yang, W.; and Yan, S. 2017. Robust LSTM-autoencoders for face de-occlusion in the wild. *IEEE Transactions on Image Processing*, 27(2): 778–790.

Zheng, H.; Lin, D.; Lian, L.; Dong, J.; and Zhang, P. 2020. Laplacian-uniform mixture-Driven iterative robust coding with applications to face recognition against dense errors. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9): 3620–3633.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *AAAI Conference on Artificial Intelligence*, 13001–13008.

Zhu, X.; Lei, Z.; Liu, X.; Shi, H.; and Li, S. Z. 2016. Face alignment across large poses: A 3D solution. In *IEEE Conference on Computer Vision and Pattern Recognition*, 146–155.

Zhu, X.; Lei, Z.; Yan, J.; Yi, D.; and Li, S. Z. 2015. High-fidelity pose and expression normalization for face recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 787–196.