FREE LUNCH FOR CROSS-DOMAIN OCCLUDED FACE RECOGNITION WITHOUT SOURCE DATA

Taoshan Zhang, Youjun Xiang, Xianfeng Li, Zichun Weng, Zhen Chen, Yuli Fu*

School of Electronic and Information Engineering, South China University of Technology
Guangzhou 510641, China

ABSTRACT

Most recognizing occluded faces methods focus on synthetic-occluded faces for training due to the lack of real-occluded data. However, the performance may suffer from degradation since the synthetic-occluded and real-occluded face images are under different distributions. Hence, it draws our eyes to transfer the model from the synthetic to the real-world domain. In this paper, we propose a source data-free domain adaptive occluded face recognition framework to optimize the network in the target domain via redefining it as a pseudo labels denoising problem. To obtain reliable pseudo labels, we train synthetic-occluded and non-occluded images via distribution alignment to extract occlusion-robust features. Nonetheless, completely correct labels are still unattainable. Then, a denoising strategy is proposed to optimize pseudo labels by centroid-based feature clustering. Experiments show that the proposed approach can effectively recognize the real-occluded face; it also reminds the occluded faces recognition community about the feasibility of domain adaptation in existing tasks.

Index Terms— occluded face recognition, adversarial learning, domain adaptation, source data-free

1. INTRODUCTION

Face recognition techniques have made remarkable improvements in unconstrained problems [1, 2]. However, the performance may be harmed in real-life applications because of uncontrollable variations like pose, facial expression, illumination, and occlusion. Among all these variations, occlusion has been considered a highly challenging one due to the massive loss of identity information.

Many approaches have been proposed for solving the occlusion problem. Traditional methods use hand-crafted features [3, 4] to extract local face descriptors from the non-occluded facial areas. Furthermore, deep convolutional neural networks are applied to improve deep discriminative features. Wan et al. [5] propose a branch named MaskNet to assign higher weights to the hidden units activated by the non-occluded facial parts. [6, 7] propose to use the mask learning strategy to find and discard corrupted feature elements from recognition. However, the above and other data-driven [8, 9] approaches rely on training synthetic occluded face images. Although they can achieve good performance in the intra-database scenario, they generalize unsatisfactorily to unseen occlusion because of the huge domain gap between synthetic-occluded (source domain) and real-occluded (target domain) face images, as shown in Fig. 1.

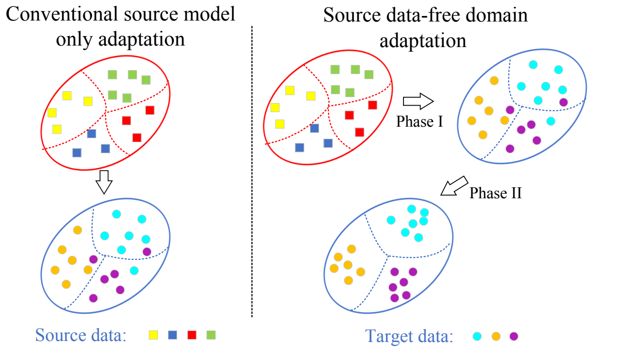
Conventional source model only adaptation
Source data-free domain adaptation
Phase I
Phase II
Source data:
Target data:

Fig. 1. Comparison between conventional occluded face recognition approach and our source data-free domain adaptive occluded face recognition.

It is natural to think that *Domain adaptation* (DA) [10–12] can be used to alleviate the domain gap for occluded face recognition. DA has been widely used in classification tasks and made good progress. Further, considering that source and target data are not freely available and can't be trained together in practical applications, some researchers proposed *source data-free domain adaptation* (SFDA) for classification [13, 14], which made significant strides. Inspired by [15, 16], this paper redefines such a cross-domain occluded face recognition problem as a pseudo labels denoising problem.

To our knowledge, this is the first time to apply DA (especially SFDA) for occluded face recognition. This paper proposes an effective training approach, source data-free domain adaptive occluded face recognition, which involves two phases, as shown in Fig. 2. We claim that a better-

This work has been supported in part by the National Natural Science Foundation of China under Grant 62001171, in part by the Natural Science Foundation of Guangdong Province under Grant 2018A030313869, 2019A1515010861 and 2021A11515011966.

*Corresponding author: Youjun Xiang

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

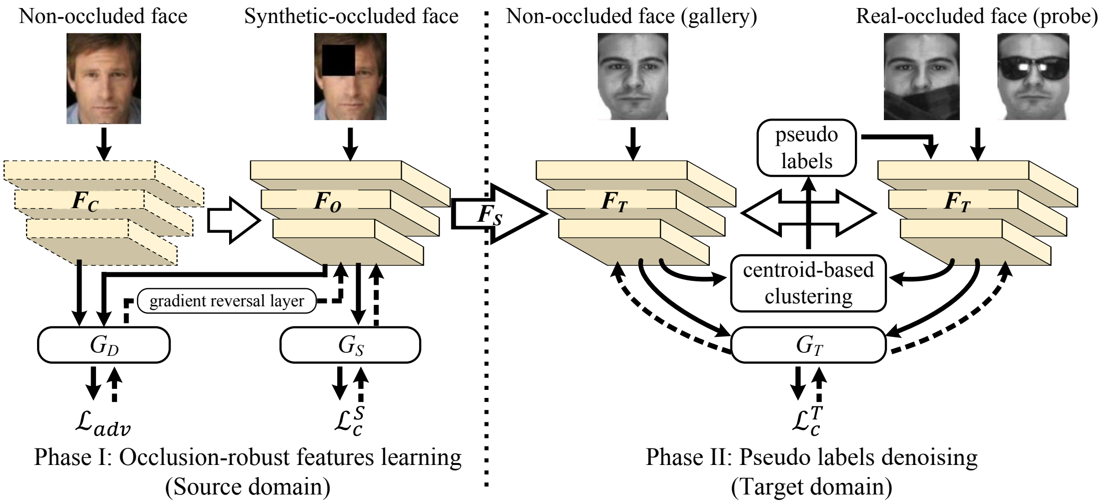
Non-occluded face
Synthetic-occluded face
Non-occluded face (gallery)
Real-occluded face (probe)
F_C
F_O
F_S
F_T
F_T
pseudo labels
centroid-based clustering
gradient reversal layer
G_D
G_S
G_T
$\mathcal{L}_{adv}$
$\mathcal{L}_c^S$
$\mathcal{L}_c^T$
Phase I: Occlusion-robust features learning (Source domain)
Phase II: Pseudo labels denoising (Target domain)

Fig. 2. Detailed overview of the proposed source data-free domain adaptive occluded face recognition approach including occlusion-robust feature extracting stage and learning with noisy labels stage.

trained source model brings a better final result; in this way, occlusion-robust feature extracting is designed. Different from conventional methods masking out the occlusion-corresponding features directly, we train both synthetic-occluded and corresponding non-occluded face images from the source domain to align their representation distributions by adversarial aligning in Phase I. To connect the source domain with the target domain, cleaning the ”bridge” (i.e. pseudo labels) is the key issue. Undoubtedly, completely clean pseudo labels are still unattainable. Considering noisy labels will not only blur the discriminative category boundary but also mislead the model, we propose a pseudo label cleaning strategy via feature centroid-based clustering in Phase II. Trained with denoised pseudo labels, features far from centroid will move toward their corresponding category centroids in the target domain.

The main contributions of this paper are three-folds. (1) We propose an adversarial-based alignment to learn occlusion-robust features for occluded face recognition. (2) We innovatively redefine the occluded face recognition with different occlusions into a problem of denoising pseudo labels and make it solvable. (3) We successfully explore the feasibility of domain adaptation in occluded faces recognition.

2. SOURCE DATA-FREE DOMAIN ADAPTIVE OCCLUDED FACE RECOGNITION

The assumption of the proposed Source Data-Free Domain Adaptive Occluded Face Recognition (SFOFR) is given source data D_s (including non-occluded face images $(x_c, y_c) \in X_c \times Y_c$ and synthetic-occluded face images $(x_o, y_o) \in X_o \times Y_o$) and target data D_t (including gallery $(x_g, y_g) \in X_g \times Y_g$ and probe $x_p \in X_p$). Fig. 2 displays the framework of the proposed method.

2.1. Phase I: Occlusion-robust feature learning

In phase I, we leverage both non-occluded face data (x_c, y_c) and occluded face data (x_o, y_o) to extract occlusion robust features from x_o. For the sake of occlusion robustness, two properties of features should be improved: occlusion invariance and identity discrimination. For this goal, we propose an adversarial-based network and train it with adversarial loss $\mathcal{L}_{adv}$ and classification loss $\mathcal{L}_c^S$ to improve these two properties, respectively.

The Adversarial Loss $\mathcal{L}_{adv}$: Occlusion invariance means that the occluded face feature $f_o = F_O(x_o)$ should be consistent with non-occluded face feature $f_c = F_C(x_c)$. To achieve this, we propose an adversarial-based alignment strategy to align f_o to have similar distribution with f_c. Here, we use a well-trained general face recognition network F_C to extract features from x_c and initialize the occluded face feature extractor F_O. To fix the distribution of f_c, we freeze the weight of F_C. Then a discriminator G_D is introduced to distinguish the f_c and f_o in the embedding space, and plays a min-max game with F_o, which can be formulated as:

$$\min_{G_D} \max_{F_O} \mathcal{L}_{adv} \tag{1}$$

where the adversarial loss $\mathcal{L}_{adv}$ is a standard GAN loss, i.e.,:

$$\begin{aligned}\mathcal{L}_{adv} = -\mathbb{E}_{(x_c)\sim(X_c)}[\log G_D\left(F_C(x_c)\right)] \\ -\mathbb{E}_{(x_o)\sim(X_o)}[\log(1 - G_D\left(F_O(x_o)\right))]\end{aligned} \tag{2}$$

F_O and G_D are optimized in the opposite direction so that we introduce a gradient reversal layer [17] between F_O and G_D.

The Classification Loss $\mathcal{L}_c^S$: Feature discrimination is essential for classification tasks. To retain the discrimination for f_o, we also train the classifier G_S and the feature extractor F_O together by penalising the angles between the deep face features and their corresponding weights in a multiplicative way [2], which reduces the intra-class variation and increases the inter-class variation. The loss can be formulated as:

$$\mathcal{L}_c^S = -\frac{1}{N_{bs}}\sum_{i=1}^{N_{bs}} \log \frac{e^{s(\cos(\theta_{y_o^i}+m))}}{e^{s(\cos(\theta_{y_o^i}+m))}+\sum\limits_{j=1,j\neq y_o^i}^{n_c} e^{s(\cos(\theta_j))}} \tag{3}$$

where N_{bs} and n_c represent the batch size and the class number, respectively. y_o^i is the label of i-th sample x_o^i, θ_j is the angle between the weight W_j and the feature of x_o^i. m and s are the margin parameter and feature scale, respectively.

In phase I, adversarial loss $\mathcal{L}_{adv}$ and classification loss $\mathcal{L}_c^S$ will be used in training simultaneously, and the total loss function can be represented as:

$$\mathcal{L}_1 = \mathcal{L}_{adv} + \lambda \mathcal{L}_c^S \tag{4}$$

where λ is a hyperparameter balancing the two losses. The feature extractor F_O trained in this phase will be used as source model F_S in the next phase.

2.2. Phase II: Pseudo label cleaning

In order to further adapt the model to target data and obtain more precise labels for probe, we propose a pseudo label denoising strategy. To train a preliminary target classifier G_T, we initialize target model F_T with source model F_S. Then train the classifier with labeled gallery (X_g, Y_g) using standard cross-entropy loss:

$$\mathcal{L}_c^T = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \log\delta_k\left(G_T\left(F_T\left(x_i\right)\right)\right) \tag{5}$$

where $\delta_k(a) = \frac{\exp(a_k)}{\sum_i \exp(a_k)}$ denotes the k-th element in softmax output of K-dimensional vector a, and K is the number of categories of target data. N represents the number of training samples. The labels of probe can be obtained as follow:

$$\hat{y}_p^{(0)} = \arg\max_k \delta_k\left(G_T\left(F_T\left(x_p\right)\right)\right) \tag{6}$$

Although we obtain the pseudo labels of probe, we have to admit that the labels are still noisy. Fortunately, label denoising techniques can be applied to clean the labels, which can effectively improve the performance of occluded face recognition. Thus, we propose a pseudo label denoising strategy via centroid-based clustering. As shown in Fig 3, the strategy attempts to obtain pseudo labels utilizing feature centroid and gradually move features towards their corresponding centroid during training with these pseudo labels. The key is how to obtain rational feature centroids and reliable pseudo labels. Here, in the n-th epoch, we calculate the initial feature centroids $c'^{(n)}_k$ for each category k and initial pseudo labels $\hat{y}'^{(n)}_p$ first. Then update them once to obtain the final feature centroids $c_k^{(n)}$ and final pseudo labels $\hat{y}_p^{(n)}$ of the n-th epoch.

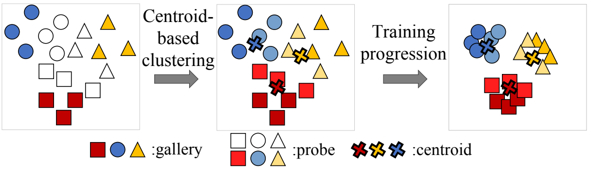

Fig. 3. The illustration of the proposed pseudo label cleaning. Different colors represent different labels.

Initial feature centroids $c'^{(n)}_k$ are calculated using features of gallery and probe as follows:

$$c'^{(n)}_k = \frac{\alpha\sum_{X_g,Y_g}\mathbb{1}(y_g=k)F_T(x_g)+\sum_{X_p}p_k(x_p)F_T(x_p)}{\alpha\sum_{X_g,Y_g}\mathbb{1}(y_g=k)+\sum_{X_p}p_k(x_p)} \tag{7}$$

where $\alpha = \frac{N_p}{N_g}$ is used to balance the difference between the number of probe N_p and gallery N_g. $\mathbb{1}(\cdot)$ is an indicator function. Probability $p_k(x) = \delta_k\left(G_T\left(F_T\left(x\right)\right)\right)$ is used as a weight to calculate the centroid of probe, which can obtain more robust and rational centroid than using pseudo labels directly. Then initial pseudo label $\hat{y}'^{(n)}_p$ is obtained by minimizing the distance between features of x_p and feature centroids:

$$\hat{y}'^{(n)}_p = \arg\min_k D(c'^{(n)}_k, F_T(x_p)) \tag{8}$$

where $D(a,b)$ measures the cosine distance between a and b. The final feature centroids $c_k^{(n)}$ and pseudo labels $\hat{y}_p^{(n)}$ can be obtained leveraging initial pseudo labels $\hat{y}'^{(n)}_p$ as follows:

$$c_k^{(n)} = \frac{\alpha\sum_{X_g,Y_g}\mathbb{1}(y_g=k)F_T(x_g)+\sum_{X_p,\hat{Y}_p}\mathbb{1}(\hat{y}'^{(n)}_p=k)F_T(x_p)}{\alpha\sum_{X_g,Y_g}\mathbb{1}(y_g=k)+\sum_{X_p,\hat{Y}_p}\mathbb{1}(\hat{y}'^{(n)}_p=k)} \tag{9}$$

$$\hat{y}_p^{(n)} = \arg\min_k D(c_k^{(n)}, F_T(x_p)) \tag{10}$$

So that, we can train the target model F_T and classifier G_T simultaneously with pseudo labels $\hat{y}_p^{(n)}$ using loss in Eq.(5).

3. EXPERIMENT

3.1. Experimental Setup

We use CASIA WebFace dataset [18] as source domain data, which consists of 494,414 images of 10,575 identities, and generate random partial occlusion on it as synthesize-occluded face data. We use the AR face dataset [19] with real-life occlusion as target domain data. The AR dataset contains 2,600 face images from 100 subjects with different facial expressions, illumination conditions, and occlusions. For each subject, there are 14 non-occluded face images, 6 face images with sunglasses occlusion and 6 face images with scarf occlusion. We employ the refined ResNet50 model proposed in ArcFace [2] as the backbone. In addition, following [6], we train ArcFace on CASIA WebFace as our baseline.

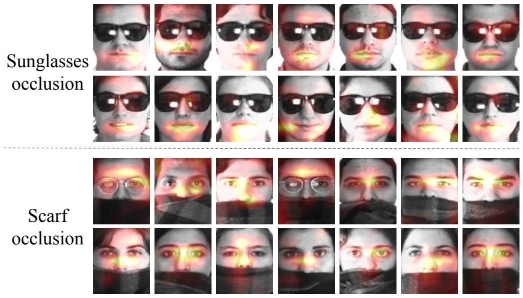

Fig. 4. Class Activation Mapping (CAM) visualizations results for Phase I.

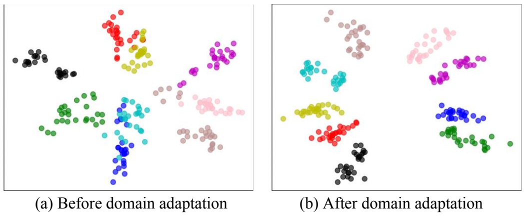
(a) Before domain adaptation
(b) After domain adaptation

Fig. 5. t-SNE visualizations for face features. Different colors represent different classes.

3.2. Visualization Result

Fig. 4 visualizes the contribution of each part of the face image to the prediction result. The heat map illustrates that the proposed method focuses on the non-occluded recognizable face regions. For instance, the occlusion-robust feature learning model pays more attention to the mouth and eyes of the face images with sunglasses and scarf occlusion, respectively. These face regions can be used to effectively be identified. In other words, the model pays more attention to the non-occluded face part rather than the occluded part utilizing adversarial-based alignment strategy in Phase I. In addition, Fig. 5 visualizes the features distribution of target data before and after domain adaptation. Comparing Fig. 5 (a) and Fig. 5 (b), we can see that after feature clustering in Phase II, feature becomes more compact within classes, and the distribution of different classes is more disperse with discriminative margins.

3.3. Comparison Result

There are mainly two kinds of testing protocols in the existing literature. Protocol 1 refers to use more than one image per subject to form the gallery set. Protocol 2 refers to use only one image per subject to form the gallery set. Images of sunglasses and scarf occlusions are used as probe for testing. Since PDSN-Soft [6] is not tested in protocol 1, we also reimplement it for comparison.

We compare our method with some existing occluded face recognition approaches. As shown in Table 1 and Table 2, the proposed model only trained with source data in Phase I obviously improves the performance in both protocols compared to baseline and other occluded face recognition approaches. It indicates that the proposed adversarial-based alignment strategy can effectively learn occlusion robust features for face recognition. Moreover, we notice that the accuracy increases greatly after adding pseudo label denoising. As shown in Table 1, face recognition accuracy even reach 100% under two different types of real-life occlusion in Protocol 1. In Protocol 2, which is obviously a more difficult task, our approach still performs well and achieves an accuracy close to 100% as shown in Table 2. It demonstrates that we effectively solve the problem of occluded face recognition in real life with both enough and scarce gallery for each subject.

Table 1. Face recognition accuracy(%) on AR face dataset with natural occlusions under Protocol 1.

Methods	References	Sunglasses	Scarf
SRC [20]	TPAMI	87.00	59.50
NMR [21]	TPAMI	96.90	73.50
MLERPM [22]	ICCV	98.00	97.00
SCF-PKR [23]	TNNLS	98.65	98.00
RPSM [24]	TIP	96.00	97.66
MaskNet [5]	ICIP	90.90	96.70
PDSN-Soft [6]	ICCV	99.33	99.67
Baseline (Source only)	-	87.17	92.17
SFOFR (Source only)	Ours	99.83	99.67
SFOFR	Ours	**100.00**	**100.00**

Table 2. Face recognition accuracy(%) on AR face dataset with natural occlusions under Protocol 2.

Methods	References	Sunglasses	Scarf
RPSM [24]	TIP	84.84	90.16
Stringface [25]	ECCV	82.00	92.00
LMA [26]	TCYBERN	96.30	93.70
PDSN-Soft [6]	ICCV	96.76	97.22
Baseline (Source only)	-	85.50	85.50
SFOFR (Source only)	Ours	98.33	97.50
SFOFR	Ours	**100.00**	**99.67**

In general, the above indicates that our method is a competitive and effective way to improve the generalization ability of real-life occluded face recognition.

4. CONCLUSION

In this paper, we propose a source data-free domain adaptive occluded face recognition approach that can adapt the source model trained on synthetic-occluded face to the target domain with unseen occlusion only using target data. Robust pre-trained models and reliable pseudo labels are both the key to promote performance. Experiment results on the real-life occluded face show that the proposed method can achieve promising recognition capability. It also reminds the occluded faces recognition community about the feasibility of domain adaptation in existing tasks.

5. REFERENCES

[1] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu, "Cosface: Large margin cosine loss for deep face recognition," in *CVPR*, 2018, pp. 5265–5274.

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019, pp. 4685–4694.

[3] Rui Min, Abdenour Hadid, and Jean-Luc Dugelay, "Improving the recognition of faces occluded by facial accessories," in *2011 IEEE International Conference on Automatic Face Gesture Recognition*, 2011, pp. 442–447.

[4] Hyun Jun Oh, Kyoung Mu Lee, and Sang Uk Lee, "Occlusion invariant face recognition using selective local non-negative matrix factorization basis images," *Image and Vision Computing*, vol. 26, no. 11, pp. 1515–1523, 2008.

[5] Weitao Wan and Jiansheng Chen, "Occlusion robust face recognition based on mask learning," in *ICIP*, 2017, pp. 3795–3799.

[6] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu, "Occlusion robust face recognition based on mask learning with pairwise differential siamese network," in *ICCV*, 2019, pp. 773–782.

[7] Haibo Qiu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao, "End2end occluded face recognition by masking corrupted features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[8] Daniel Sáez Trigueros, Li Meng, and Margaret Hartnett, "Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss," *Image and Vision Computing*, vol. 79, pp. 99–108, 2018.

[9] Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, and Xiaoming Liu, "Towards interpretable face recognition," in *ICCV*, 2019, pp. 9347–9356.

[10] Jonathan Munro and Dima Damen, "Multi-modal domain adaptation for fine-grained action recognition," in *CVPR*, 2020, pp. 119–129.

[11] Rameswar Panda, Amran Bhuiyan, Vittorio Murino, and Amit K. Roy-Chowdhury, "Unsupervised adaptive re-identification in open world dynamic camera networks," in *CVPR*, 2017, pp. 1377–1386.

[12] Seungmin Lee, Dongwan Kim, Namil Kim, and Seong-Gyun Jeong, "Drop to adapt: Learning discriminative features for unsupervised domain adaptation," in *ICCV*, 2019, pp. 91–100.

[13] Y. Kim, S. Hong, and D. Cho, "Domain adaptation without source data," in *arXiv:2007.01524*, 2020.

[14] R. Li, Q. Jiao, W. Cao, H.-S. Wong, and S. Wu, "Model adaptation: Unsupervised domain adaptation without source data," in *CVPR*, 2020.

[15] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang, "A free lunch for unsupervised domain adaptive object detection without source data," in *AAAI*, 2021.

[16] Lingling Lv, Youjun Xiang, Xianfeng Li, Hanye Huang, Rongju Ruan, Xiaoyan Xu, and Yuli Fu, "Combining dynamic image and prediction ensemble for cross-domain face anti-spoofing," in *ICASSP*, 2021.

[17] Y. Ganin and V. Lempitsky, "Universal source-free domain adaptation," in *ICML*, 2015.

[18] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, , no. 11, 2014.

[19] Aleix M. Martinez, "The ar face database," 1998.

[20] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma, "Robust face recognition via sparse representation," *TPAMI*, vol. 31, no. 2, pp. 210–227, 2009.

[21] Jian Yang, Lei Luo, Jianjun Qian, Ying Tai, Fanlong Zhang, and Yong Xu, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," *TPAMI*, vol. 39, no. 1, pp. 156–171, 2017.

[22] Renliang Weng, Jiwen Lu, Junlin Hu, Gao Yang, and Yap-Peng Tan, "Robust feature set matching for partial face recognition," in *ICCV*, 2013, pp. 601–608.

[23] Meng Yang, Lei Zhang, Simon Chi-Keung Shiu, and David Zhang, "Robust kernel representation with statistical local features for face recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 6, pp. 900–912, 2013.

[24] Renliang Weng, Jiwen Lu, and Yap-Peng Tan, "Robust point set matching for partial face recognition," *TIP*, vol. 25, no. 3, pp. 1163–1176, 2016.

[25] Weiping Chen and Yongsheng Gao, "Recognizing partially occluded faces from a single sample per class using string-based matching," in *ECCV*, 2010.

[26] Niall McLaughlin, Ji Ming, and Danny Crookes, "Largest matching areas for illumination and occlusion robust face recognition," *IEEE Transactions on Cybernetics*, vol. 47, no. 3, pp. 796–808, 2017.