On Improving the Generalization of Face Recognition in the Presence of Occlusions

Xiang Xu, Nikolaos Sarafianos, and Ioannis A. Kakadiaris
Computational Biomedicine Lab
University of Houston
{xxu21, nsarafianos, ioannisk}@uh.edu

Abstract

In this paper, we address a key limitation of existing 2D face recognition methods: robustness to occlusions. To accomplish this task, we systematically analyzed the impact of facial attributes on the performance of a state-of-the-art face recognition method and through extensive experimentation, quantitatively analyzed the performance degradation under different types of occlusion. Our proposed Occlusion-aware face REcOgnition (OREO) approach learned discriminative facial templates despite the presence of such occlusions. First, an attention mechanism was proposed that extracted local identity-related region. The local features were then aggregated with the global representations to form a single template. Second, a simple, yet effective, training strategy was introduced to balance the non-occluded and occluded facial images. Extensive experiments demonstrated that OREO improved the generalization ability of face recognition under occlusions by 10.17% *in a single-image-based setting and outperformed the baseline by approximately* 2% *in terms of rank-1 accuracy in an image-set-based scenario.*

1. Introduction

The goal of this paper is to present a face recognition method that is robust to facial occlusions originating from facial attributes. For example, given a facial image of an individual wearing sunglasses or a hat, we aspire to successfully match this probe image with the corresponding images in the gallery to obtain his/her identity. Note that there are other challenges (such as self-occlusions or extreme pose variations) that might affect the face recognition performance. External occlusions can be defined as those caused by facial accessories such as glasses, hats or different types of facial hair. Despite the recent success of face recognition methods [9, 33, 37], most existing research tends to focus solely on the pose challenge while failing to account for other factors such as occlusion, age, and facial expression variations, that can have a negative impact on the face recognition performance.

(a) (b) (c)

Figure 1: Depiction of incorrectly matched samples using ResNeXt-101 on the CFP-FF dataset: (a) false negative matches due to occlusions; (b) false negative matches due to age, pose, and facial expression variations; and (c) false positive matches due to similar appearance.

Aiming to gain a better understanding of the common failure cases, a ResNeXt-101 [14] model was trained on the VGGFace2 [3] dataset and was evaluated on the CFP dataset [34] using the frontal-to-frontal matching protocol (CFP-FF). This model was selected to serve as a baseline since its verification performance on the CFP dataset is almost state-of-the-art and at the same time, it is easy to train in all deep learning frameworks. Figure 1 presents the false positive and false negative pairs of images from the CFP dataset based on predictions of the ResNeXt-101 model. It is worth noting that in this protocol the faces have low variation in the yaw angle. The obtained results indicate that the sources of error for most false matching results originate

from factors such as occlusion and age difference. Similar results are observed in the same dataset using the frontal-to-profile matching protocol (CFP-FP). Based on these observations, we can confidently conclude that besides large pose variations, occlusion is a significant factor that greatly affects the recognition performance.

Why does face recognition performance degrade in the presence of occlusions? First, important identity-related information might be excluded when the face is occluded. Figure 1(a) depicts several samples that would be challenging for a human to decide whether the individuals with sunglasses belong to the same identity or not. Second, existing deep learning approaches are data-driven, which means that the generalization power of the model usually is limited by the training data. However, current datasets are collected by focusing mainly on the pose distribution. This introduces a large class imbalance in terms of occlusion since, in the majority of the selected samples, the entire face is visible. A limited number of approaches [6, 15, 56] have recently tried to address this problem. However, such methods require prior knowledge of occlusion presence [6, 56] to perform de-occlusion on synthetically generated occlusions, which is not practical in real-life applications.

Aiming to improve the generalization of face recognition methods, we first investigate which type of occlusion affects the most the final performance. Our experimental analysis indicated that the main reasons for the performance degradation in the presence of occlusions are: (i) identity signal degradation (*i.e.*, information related to the identity of the individual is lost in the presence of occlusion), and (ii) occlusion imbalance in the training set. To address the first challenge of identity signal degradation due to occlusions, an attention mechanism is introduced which is learned directly from the training data. Since the global facial template captures information learned from the whole facial image (regardless of whether occlusion occurs), the attention mechanism aims to disentangle the identity information using the global representation and extract local identity-related features from the non-occluded regions of the image. In this way, global and local features are jointly learned from the facial images and are then aggregated into a single template. To address the challenge of the occlusion imbalance in the training set, an occlusion-balanced sampling strategy is designed to train the model with batches that are equally balanced with non-occluded and occluded images. Based on this strategy, an additional learning objective is proposed that improves the discriminative ability of the embeddings learned from our algorithm. Thus, the proposed occlusion-aware feature generator results in facial embeddings that are robust to occlusions originating from visual attributes. Our results demonstrate that OREO significantly improves the face recognition performance without requiring any prior information or additional supervision. Through extensive experiments and ablation studies, we demonstrate that the proposed approach achieves comparable or better face recognition performance on non-occluded facial images and at the same time significantly improves the generalization ability of the facial embedding generator on facial images in which occlusion is present.

In summary, the contributions of this work are: (i) An analysis of the impact of attributes to the face recognition performance is conducted on the Celeb-A dataset and its key insights are presented. (ii) An attention mechanism is introduced that disentangles the features into global and local parts, resulting in more discriminative representations. In this way, the global features contain identity-related information while the local features learned through our attention mechanism are robust to occlusions caused by visual attributes. (iii) An occlusion-balanced sampling strategy along with a new loss function are proposed to alleviate the large class imbalance that is prevalent due to non-occluded images in existing datasets.

2. Related Work

Face Recognition under Occlusion: Face recognition techniques that generate 2D frontal images or facial embeddings from a single image have been proposed that: (i) use a 3D model [27, 49, 50], (ii) generative adversarial networks [1, 5, 8, 19, 38, 53, 57, 59], and (iii) various transformations [2, 63, 65]. Recent works have also focused on long-tail [54, 64] or noisy data [17] Additionally, multiple loss functions [9, 10, 23, 24, 25, 41, 44, 45, 58, 60, 62] have been developed to guide the network to learn more discriminative face representations, but usually, ignore facial occlusions. Early methods approached face recognition in the presence of occlusions by using variations of sparse coding [11, 39, 55, 61]. However, such techniques work well only with a limited number of identities, and with frontal facial images in a lab controlled environment. The works of He *et al*. [15] and Wang *et al*. [43] addressed this limitation by matching face patches under the assumption that the occlusion masks were known beforehand and that the occluded faces from the gallery/probe were also known, which is not realistic. Guo *et al*. [12] fit a 3D model on images in-the-wild to render black glasses and enlarge the training set. However, this method was designed to tackle a specific type of occlusion and cannot cover most occlusion cases that might appear in scenarios in the wild. Finally, Song *et al*. [36] proposed a mask-learning method to tackle occlusions in face recognition applications. A pairwise differential siamese network was introduced so that correspondences could be built between occluded facial blocks and corrupted feature elements.

Visual Attention: Several works have appeared recently that demonstrate the ability of visual attention to learn discriminative feature representations [16, 22, 32, 40, 66].

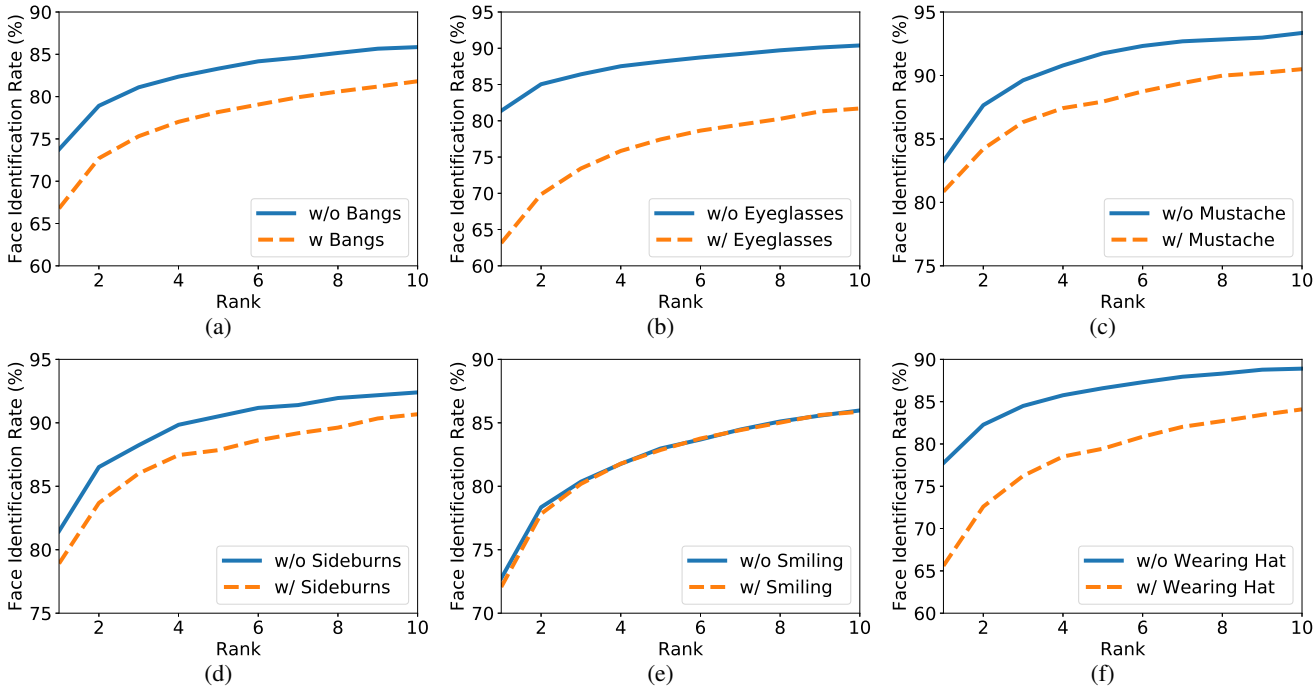

Figure 2: Depiction of CMCs demonstrating the impact of selected facial attributes on the face identification performance using a ResNeXt-101 as a backbone architecture that generates facial embeddings. The corresponding Celeb-A attributes are: (a) `Bangs`, (b) `Eyeglasses`, (c) `Mustache`, (d) `Sideburns`, (e) `Smiling`, and (f) `Wearing Hat`.

Most methods [7, 31, 42] extract saliency heatmaps at multiple-scales to build richer representations but fail to account for the correlation between the attention masks at different levels. Jetley *et al*. [20] proposed a self-attention mechanism that focuses on different regions to obtain local features for image classification under the hypothesis that the global representation contains the class information. However, in their architecture, the global features are not regularized by the identity loss function, which does not support their original hypothesis. Castanon *et al*. [4] used visual attention to quantify the discriminative region on facial images. Other methods [30, 46, 47, 51] apply attention mechanisms to weigh the representations from multiple images. Finally, Shi *et al*. [35] introduced a spatial transformer to find the discriminative region but their method required additional networks to obtain the local features.

In this paper, instead of simulating each case of facial occlusion, a new method is presented that directly learns from images in-the-wild that contain a plethora of occlusion variations. Our approach improves the generalization ability of the facial embedding generator, without having any prior knowledge of whether the occlusion is present.

3. Systematic Analysis on Impact of Facial Attributes on Face Recognition

Aiming to quantitatively analyze the impact of occlusion, a series of experiments are conducted on the Celeb-A dataset [26], which consists of 10,177 face identities and 40 facial attributes. Attributes that describe the subject (*e.g*., `Gender`) were ignored and only those that might impact the face recognition performance were selected: `Bangs`, `Eyeglasses`, `Mustache`, `Sideburns`, `Smiling`, and `Wearing Hat`. For each attribute, the images without this attribute were enrolled as the gallery and images w/ or w/o this attribute were enrolled as probes. In both the gallery and probe, each identity has only a single image. A ResNeXt-101 was deployed as the facial embedding generator and six face identification experiments were conducted. For each of the six attributes, Cumulative Match Curves (CMC) are provided in Fig. 2 w/ and w/o that attribute, respectively. Note that since there are different identities and a different number of images involved in each experiment, the horizontal comparison of the identification rate across attributes does not lead to meaningful conclusions.

In Fig. 2, the identification rates with and without each attribute are presented. Our results indicate that the face recognition performance decreases in the presence of the attributes `Bangs`, `Eyeglasses`, `Mustache`, `Sideburns`, and `Wearing Hat`. The attributes can be ranked according to the rank-1 identification rate degradation as follows: `Eyeglasses` (18.23%) > `Wearing Hat` (12.14%) > `Bangs` (6.97%) > `Sideburns` (2.56%) ∼ `Mustache` (2.41%). These results demonstrate that occlusion originating from facial accessories (*i.e*., eyeglasses,

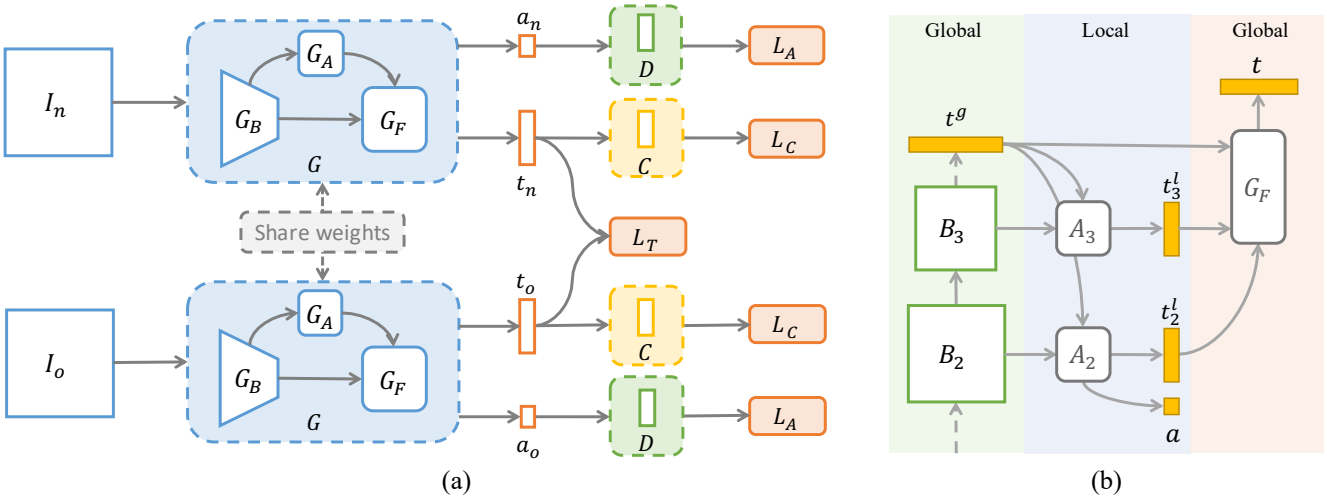

Figure 3: (a) Given a pair of non-occluded and occluded images (I_n, I_o), the template generator G learns the facial embeddings (t_n, t_o) and the attributes predictions a_n, a_o with the attributes classifiers (D) and identity classifier (C) using facial attribute classification loss L_A, identity classification loss L_C, and the proposed similarity triplet loss L_T. (b) Depiction of the generator in detail, which contains: (i) the output feature maps of the last two blocks (B_2, B_3) of the backbone architecture, (ii) the attention mechanism G_A consisting of masks (A_2, A_3) that learn the local features in two different ways, and (iii) G_F which aggregates the global and local features to the final embedding.

and hat) as well as facial hair (*i.e.*, mustache, bangs, and sideburns) is an important challenge that affects the performance of face recognition algorithms. Additionally, we observed that occlusion due to external accessories affects the performance more than occlusion originating from facial hair. Finally, note that the identification performance is almost the same regardless of whether the subject is smiling or not. The main reason for these results is that such datasets are collected from the web and thus, they usually cover a large range of head poses and contain enough images of smiling individuals. However, there is still a *high imbalance* in other factors such as occlusion, which reduces the robustness of face recognition methods. While testing the facial attribute predictor on the VGGFace2 dataset we observed class imbalance ranging from 19:1 (`Bangs`) to 6:1 (`Wearing Hat`) in the VGGFace2 dataset.

4. Improving the Generalization

The training process of OREO (depicted in Fig. 3 (a)) consists of (i) an occlusion-balanced sampling (OBS) to address the occlusion imbalance; (ii) an occlusion-aware attention network (OAN) to jointly learn the global and local features; and (iii) the objective functions that guide the training process. Aiming to balance the occluded and non-occluded images within the batch, random pairs of non-occluded and occluded images are sampled and provided as input to the network. Then, the proposed attention mechanism is plugged into the backbone architecture to generate the attention mask and aggregate the local with the global features to construct a single template. The final aggregated features are trained to learn occlusion-robust template guided by the softmax cross-entropy, sigmoid cross-entropy, and similarity triplet loss (STL) functions.

Occlusion-Balanced Sampling. In a hypothetical scenario in which the training data would be accompanied by occlusion ground-truth labels, the training set could easily be split into two groups of occluded and non-occluded images from which balanced batches could be sampled and fed to the neural network. However, this is not the case with existing face recognition training datasets since occlusion ground-truth labels are not provided. Aiming to generate occlusion labels, we focused on facial attributes that contain occlusion information. A state-of-the-art face attribute predictor [32] was trained on the Celeb-A dataset and it was then applied to the training set to generate pseudo-labels. Those attribute pseudo-labels can then be utilized to facilitate occlusion-balanced sampling during training. It is worth noting this approach is dataset-independent and can be generalized to any face dataset since we solely rely on the attribute predictions learned from the Celeb-A dataset. By using this strategy, the network G is feed-forwarded with pairs of randomly chosen non-occluded and occluded images denoted by $\{\{I_n, y_n\}, \{I_o, y_o\}\}_i, i = \{1, \ldots, N\}$, where y contains the identity and attributes of the facial images and N is the total number of pairs. Since OBS randomly generates pairs of occluded and non-occluded facial images in each epoch, their distribution within the batch is ensured to be balanced.

Occlusion-aware Attention Network. The template generator G consists of three components: (i) a backbone network G_B, (ii) an attention mechanism G_A, and (iii) a fea-

ture aggregation module G_F. Features are generated from two pathways as depicted in Figure 3 (b): a bottom-up for the global representations and a top-down for the local features. The bottom-up term describes the process of using representations of larger spatial dimensions to obtain a low-dimensional global embedding. The top-down term uses low-dimensional feature embeddings as input to the proposed attention masks applied in higher dimensions to generate local representations. In the bottom-up pathway, the process of the network to generate global features is described by $t^g = G_B(I)$. In the top-down pathway, since the global features include information from the occluded region of the face, an attention module G_A is proposed that distills the identity-related features from the global feature maps to the local representations t^l. Finally, a feature aggregation module G_F is employed that aggregates the global and local features into a single compact representation t. The objective of attention modules G_A is to help the model identify which areas of the original image contain important information based on the identity and attribute labels. Assuming the feature maps extracted from different blocks of the backbone network are denoted by $\{B_1, B_2, B_3, B_4\}$, then the two-step attention mechanism is designed as follows. In the first level (denoted by A_3 in Fig. 3 (b)), we focus on self-attention and thus, the feature map B_3 is first broadcasted and then added with the global representation t^g to generate the attention mask A_3. The objective of A_3 is to find the high-response region of the feature map by giving emphasis to the identity-related features and construct the local representation t^{l_3}. This process is described by the following equation:

$$t_3^l = A_3 * B_3 = h_3(t^g, B_3) * B_3 \ , \tag{1}$$

where h_3 is a sequence of convolution layers to reduce the channels and generate a single-channel attention map and the "$*$" operation corresponds to element-wise multiplication. The final global feature t^g is preserved as part of the final representation of the network so that t^g learns identity-related information. Thus, t^g guides the network to learn local attention maps on features from the previous block and distill the identity information from the most discriminative region to construct t^l. Our experiments indicated that the most discriminative region on the face from our first level attention mechanism is the eye region. While this is sufficient for most attributes, this is not the case when the individual is wearing glasses. To improve the generalization ability of the model in such cases, an additional attention mechanism is introduced on the feature map B_2 to force the network to focus on other regions of the face. In the second level (denoted by A_2 in Fig. 3 (b)), the attention map is guided by the facial attribute predictions in a weakly-supervised manner (*i.e.*, no attribute ground-truth information is utilized since we only have access to the visual attribute predictions). Thus, the local representations at this level are computed by:

$$t_2^l = (1 - A_2) * B_2 = (1 - h_2(t^g, B_2)) * B_2 \ , \tag{2}$$

where h_2 is an attention operation guided by both identity labels and attributes labels. Since the attention map A_2 is guided not only by the identity loss function but also by the attribute predictions, the network is capable of focusing on image regions related to both the identity and the visual attributes. The global and local features $\{t^g, t_2^l, t_3^l\}$ are concatenated into one single vector and are projected in a single feature template t, which enforces both global and local features to preserve semantic identity information.

Learning Objectives. Loss functions are employed for training: (i) the softmax cross-entropy loss L_C for identity classification, (ii) the sigmoid binary cross-entropy loss L_A for attribute prediction, and (iii) a new loss L_T designed for the occlusion-balanced sampling. The identity classification loss is defined as:

$$L_C = -\frac{1}{M}\sum_{i=0}^{M} log\frac{\exp\left(W_{y_i^c}t_i + b_{y_i^c}\right)}{\sum_{j=1}^{n}\exp\left(W_j t_i + b_j\right)}, \tag{3}$$

where t_i and y_i^c represent the features and the ground-truth identity labels of the i^{th} sample in the batch, W and b denote the weights and bias in the classifier, and M and n correspond to the batch size and the number of identities in the training set, respectively. Following that, the sigmoid binary cross-entropy loss L_A can be defined as

$$L_A = -\frac{1}{M}\sum_{i=0}^{M}\log\sigma(a_i)y_i^a + \log(1-\sigma(a_i))(1-y_i^a) \ , \tag{4}$$

where y^a corresponds to the attribute labels (pseudo-labels obtained by the attribute predictor) and $\sigma(\cdot)$ is the sigmoid activation applied on the attribute predictions a (i.e., predictions of the occlusion-aware face recognition network G).

In the matching stage, the cosine distance is used to compute the similarity between two feature embeddings. Since images with the same identity have a higher similarity score than those with a different identity, we introduce a similarity triplet loss (STL) as regularization to make the final facial embedding more discriminative. During training, each batch comprises pairs of non-occluded and occluded images with each pair having the same identity. Let t_n and t_o be the final feature representations of non-occluded images I_n and occluded images I_o respectively. The similarity matrix $S \in \mathbb{R}^{M\times M}$ within the batch is then computed, where M is the batch size. In the similarity matrix S, we aim to identify: (i) hard positives which are pairs of samples that originate from the same identity but have low similarity score $s^p(t_n, t_o)$ and (ii) hard negatives which are pairs of samples with different identities but with high similarity score

Method	Bangs		Eyeglasses		Mustache		Sideburns		Wearing Hat		ADP (%)
	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	
ResNeXt-101 [14]	73.76	66.79	81.38	63.14	83.26	80.85	81.47	78.91	77.74	65.60	8.46
OREO	**74.65***	**68.66***	**81.99**	**65.53***	**84.72***	**82.53***	**83.13***	**81.47***	**80.02***	**68.34***	**7.60**

Table 1: Comparison of rank-1 identification rate (%) on the Celeb-A dataset w/ and w/o the specified attribute. ADP corresponds to the average degradation percentage. The lower the value of ADP, the more robust the method is. Note that "$*$" denotes statistically significant improvements using the McNemar's statistical test.

$s^n(t_n, t_o)$. Then, the objective function is defined as:

$$L_T = \sum_{i=1}^{M} [s_i^n(t_n, t_o) - s_i^p(t_n, t_o) + m]_+ . \quad (5)$$

A margin $m \in \mathbb{R}^+$ is maintained to enforce that small angles (high similarity score) belong to the same identity and large angles (low similarity score) belong to different identities. Finally, the whole network is trained using the summation of the individual losses: $L = L_C + L_A + L_T$.

5. Experiments

Four evaluation datasets are used to extensively evaluate our algorithm under different scenarios. To generate the occlusion meta-data (*i.e.*, attributes) of the training set, a state-of-the-art face attribute predictor [32] is used to predict the occlusion-related attributes on the training set. The evaluation protocol of each dataset is strictly followed. For face verification, Identification Error Trade-off is reported with true acceptance rates (TAR) at different false accept rates (FAR). For face identification, CMC is reported. These evaluation metrics were computed using FaRE toolkit [48].

5.1. Comparison with state of the art

Celeb-A: In-the-wild Face Identification. We tested our algorithm on the Celeb-A dataset [26] under various types of occlusion. It is worth noting that the identity labels of the Celeb-A dataset were not utilized since OREO was trained solely on the VGGFace2 dataset. The rank-1 identification rate w/ and w/o each attribute is presented in Table 1. We observe that OREO outperforms ResNeXt-101 in all settings (w/ and w/o attributes), which suggests that our algorithm can learn robust discriminative feature representations regardless of occlusions. The improvement is statistically significant in all attributes according to the McNemar's test. In addition, OREO demonstrated a lower average degradation percentage than the baseline by 10.17% in terms of relative performance. This indicates that our algorithm improves the generalization ability of the facial embedding generator in the presence of occlusions.

CFP: In-the-wild Face Verification. The CFP [34] is used to evaluate face verification performance on images in-the-wild, which contain variations in pose, occlusion, and age. In the first experiment, we qualitatively visualize the region of frontal-face images that demonstrated the highest response, because such regions provide the most meaningful information to the final embedding. We chose a state-of-the-art algorithm [20] as our baseline to compare the attention mask learned from the data. Figure 4 depicts samples from the CFP dataset with different occlusion variations. Compared to the attention mask of Jetley *et al.* [20], our attention mechanism has the following advantages: (i) By combining the global with the local feature maps to generate the attention mask, the global representation demonstrates a higher response around the eye regions, which indicates that the eyes contain discriminative identity-related information. It also explains why the `Eyeglasses` attribute demonstrated the highest performance drop in Section 3; (ii) By learning the attention from local feature maps guided by the occlusion pseudo-labels generated by our method, we observe that the proposed attention mask is focusing more on the non-occluded region. In addition, it helps the embeddings aggregate information from the non-occluded facial region instead of solely relying on the eye regions as indicated by our first observation. Note that, the self-attention described in Eq. (1) is learned directly from the data without having access to ground-truth masks that would help our model identify better the most discriminative regions (hence the failure case depicted in the second to last image of the second row in Figure 4). This is why an additional attention map was introduced in Eq. (2) which is learned in a weakly-supervised manner from the attribute predictions and helps our algorithm to focus on the non-occluded region (right-most image of the second row).

In the second experiment, we used the CFP-FP protocol to quantitatively evaluate the face verification performance and the experimental results are presented in Table 3. We used the following metrics to evaluate the performance: (i) the verification accuracy, and (ii) the TAR at FAR equal to 10^{-3}, 10^{-2}, and 10^{-1}. Compared to all baselines, OREO achieves state-of-the-art results on this dataset in terms of accuracy and increases the TAR at low FARs. The moderately better accuracy results demonstrate that OREO can also improve the performance of general face recognition.

UHDB31: Lab-controlled Face Identification. The UHDB31 [21] dataset is used to evaluate the face identification performance under different pose variations. OREO

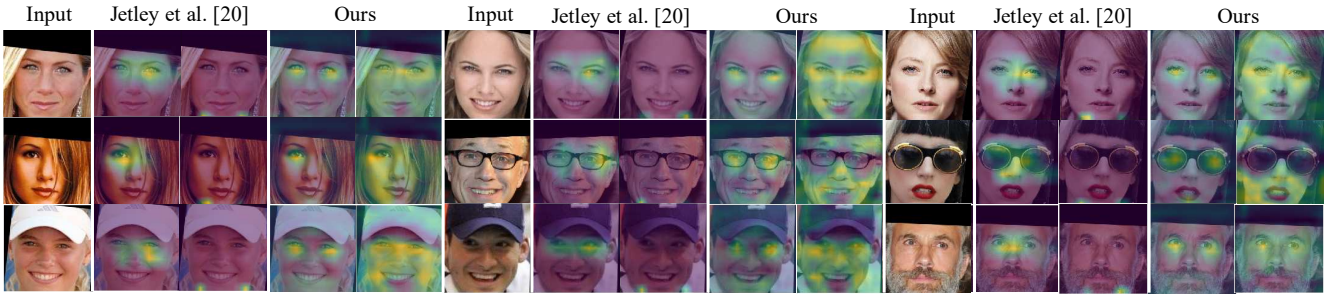

Figure 4: Visualization of the discriminative region using an attention mechanism. The first attention map of OREO focuses on the discriminative region (eyes), whereas the second attention map focuses on the non-occluded facial region.

Pitch \ Yaw	$-90°$	$-60°$	$-30°$	$0°$	$+30°$	$+60°$	$+90°$
$+30°$	58, 82, **96**	95, 99, **100**	100, 100, **100**	100, 100, **100**	99, 99, **100**	92, 99, **100**	60, 75, **94**
$0°$	84, 96, **98**	99, 100, **100**	100, 100, **100**	-	100, 100, **100**	99, 100, **100**	91, 96, **100**
$-30°$	44, 74, **86**	80, 97, **99**	99, 100, **100**	99, 100, **100**	97, 100, **100**	90, 96, **100**	35, 78, **95**

Table 2: Comparison of Rank-1 identification rate (%) of different methods on the `UHDB31.R128.I03` dataset. The results in each cell are ordered in the sequence of FaceNet [33], UR2D-E [49], and OREO.

Method	Acc. (%)	TAR (%) @ FAR=		
		10^{-3}	10^{-2}	10^{-1}
DR-GAN [38]	93.4 ± 1.2	-	-	-
MTL-CNN [52]	94.4 ± 1.2	-	-	-
ArcFace [9]	93.9 ± 0.8	80.2 ± 5.9	86.0 ± 2.8	94.3 ± 1.5
ResNeXt-101 [14]	97.1 ± 0.8	81.9 ± 11.4	92.3 ± 4.1	98.9 ± 0.8
OREO	$\mathbf{97.5 \pm 0.5}$	$\mathbf{85.5 \pm 5.3}$	$\mathbf{94.1 \pm 2.5}$	$\mathbf{99.2 \pm 0.7}$

Table 3: Comparison of the face verification performance with state-of-the-art face recognition techniques on the CFP dataset using CFP-FP protocol. The evaluation results are presented by the average $\pm$ standard deviation over 10 folds.

outperformed FaceNet [33] and UR2D-E [49] in all settings and especially in large pose variations (*e.g.*, yaw = -90^o, pitch = $+30^o$, which indicates that our algorithm is robust to pose variations (Table 2).

IJB-C: Set-based Face Identification and Verification. The IJB-C dataset [28] is a mixed media set-based dataset with open-set protocols comprising images with different occlusion variations. Two experiments are performed on this dataset following 1:1 mixed verification protocol and 1:N mixed identification protocol. To generate the facial embedding for the whole set, the corresponding images are fed to the neural networks and the average of the embeddings from all images within a set is computed. The evaluation metrics include the verification metric of ROC, the identification metric of retrieval rate, the true positive identification rate (TPIR) at different false positive identification rates (FPIR). From the obtained results in Table 4, we observe that OREO outperforms four out of six baselines in all metrics and comes second to ArcFace only in some cases under the mixed verification and identification protocols. ArcFace was trained on the MS1M [13] dataset which contains significantly more identities and data than the VGGFace2 dataset. OREO performs well across the board and outperforms ArcFace at high FARs in the verification protocol as well as in the identification retrieval protocol. When compared against the baseline, OREO significantly improves the performance in both verification and identification. For example, when FAR is equal to 10^{-7} the TAR improves from 28.72% to 51.97%. These results demonstrate that OREO can successfully learn features that are robust to occlusions.

5.2. Ablation Study

An ablation study was conducted to explore the impact of the proposed components: (i) the attention mechanism (OAN), (ii) the balanced training strategy (OBS), and (iii) the similarity triplet loss (STL). The results of the contributions of each component along with the backbone network [14] and an attention-based baseline [20] on the LFW [18] and CFP [34] datasets are depicted in Table 5. For faster turnaround, in this study we used only the first 500 subjects of VGGFace2 [3] as training set which is why the results between Table 3 and Table 5 are different. Compared to the backbone network, OREO increases the verification accuracy by at least 1.10%.

Occlusion-aware Attention Network. OAN helps the network focus on the local discriminative regions by pooling local features. From the results in Table 5, we observe that the backbone architecture that encodes only global representations achieves better results than the method of Jetley *et al.* [20], which leverages only local information. The

Method	1:1 Mixed Verification TAR (%) @ FAR=							1:N Mixed Identification TPIR (%) @ FPIR=			1:N Mixed Identification Retrieval Rate (%)		
	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{-3}	10^{-2}	10^{-1}	Rank-1	Rank-5	Rank-10
GOTS [28]	3.00	3.00	6.61	14.67	33.04	61.99	80.93	2.66	5.78	15.60	37.85	52.50	60.24
FaceNet [33]	15.00	20.95	33.30	48.69	66.45	81.76	92.45	20.58	32.40	50.98	69.22	79.00	81.36
VGGFace [29]	20.00	32.20	43.69	59.75	74.79	87.13	95.64	26.18	45.06	62.75	78.60	86.00	89.20
MN-vc [47]	-	-	-	**86.20**	**92.70**	**96.80**	**98.90**	-	-	-	-	-	-
ArcFace [9]	**60.50**	**73.56**	**81.70**	**87.90**	91.14	95.98	97.92	**70.90**	**81.98**	**87.63**	**92.25**	**94.31**	**95.30**
ResNeXt-101 [14]	28.72	58.09	71.19	81.76	90.70	95.75	98.86	53.66	71.50	82.47	91.88	95.51	97.29
OREO	**51.97**	**62.36**	**75.86**	85.19	**92.81**	**97.11**	**99.37**	**65.47**	**77.11**	**85.92**	**93.76**	**96.68**	**97.74**

Table 4: Comparison of the face verification and identification performance of different methods on the IJB-C dataset. Top performance is marked with **black** and second-best with **blue**. OREO significantly improves the face verification and identification performance compared to the baseline, and achieves state-of-the-art results in terms of retrieval rate.

Method	Module			CFP-FF				CFP-FP				LFW			
	OAN	OBS	STL	Acc. (%)	TAR (%) @ FAR=			Acc. (%)	TAR (%) @ FAR=			Acc. (%)	TAR (%) @ FAR=		
					10^{-3}	10^{-2}	10^{-1}		10^{-3}	10^{-2}	10^{-1}		10^{-3}	10^{-2}	10^{-1}
ResNeXt-101 [14]				94.77	43.71	82.15	97.03	88.66	32.71	50.20	85.17	94.77	64.77	84.50	96.31
Jetley *et al*. [20]	✓			93.47	40.48	78.07	94.75	83.71	22.23	39.07	71.31	92.70	55.63	77.23	93.61
OREO	✓			95.41	49.69	85.68	**97.56**	90.24	35.34	59.87	89.33	95.90	66.33	90.00	97.47
		✓		95.33	46.02	86.01	97.53	89.21	33.49	56.51	87.49	95.55	71.63	89.80	97.12
		✓	✓	95.66	48.39	87.76	97.74	89.49	36.97	59.56	88.31	96.17	**74.13**	90.63	97.63
	✓	✓	✓	**95.86**	**49.74**	**89.38**	97.48	**90.60**	**40.06**	**63.90**	**90.64**	**96.20**	68.13	**92.03**	**97.86**

Table 5: Impact of the individual proposed components under face verification protocols on the CFP and LFW datasets.

proposed attention mechanism outperforms both techniques since it includes both global and local features into the final embedding guided by the identity loss. Finally, compared to ResNeXt-101 which served as our baseline, OAN results in absolute improvements ranging from 0.64% to 1.58% in terms of verification accuracy.

Occlusion Balanced Sampling. OBS creates pairs of occluded and non-occluded images to alleviate the problem of occlusion imbalance during training. OBS results in absolute improvements ranging from 0.56% to 0.78% in terms of verification accuracy compared to the performance of the backbone network. These results indicate that OBS has a limited effect on these datasets because they contain a limited number of occluded samples.

Similarity Triplet Loss. STL works as regularization to the final loss, as it increases the similarity of two feature embeddings that belong to the same identity and penalizes the network when the similarity is high but the features are originating from different identities. STL does not force the representations to be learned in the same scale which is an advantage compared to other alternative similarity loss functions [9, 25, 41]. Since STL requires pairs of occluded and non-occluded images, OBS is a prerequisite for this loss, which is why an experimental result with only STL is not provided in Table 5. We observe that STL improves the performance by at least 0.83% in terms of verification accuracy. In addition, by comparing the TARs at low FARs on all three datasets, we observe that OBS along with STL are the components that affect the most performance compared to ResNeXt-101. The obtained results demonstrate that the features learned with STL are more discriminative than those learned using only a softmax cross-entropy loss.

6. Conclusion

In this paper, we systematically analyzed the impact of facial attributes to the performance of a state-of-the-art face recognition method and quantitatively evaluated the performance degradation under different types of occlusion caused by facial attributes. To address this degradation, we proposed OREO: an occlusion-aware approach that improves the generalization ability on facial images occluded. An attention mechanism was designed that extracts local identity-related information. In addition, a simple yet effective occlusion-balanced sampling strategy and a similarity-based triplet loss function were proposed to balance the non-occluded and occluded images and learn more discriminative representations. Through ablation studies and extensive experiments, we demonstrated that OREO achieves state-of-the-art results on several publicly available datasets and provided an effective way to better understand the representations learned by the proposed method.

Acknowledgment This material was supported by the U.S. Department of Homeland Security under Grant Award Number 2017-STBTI-0001-0201 with resources provided by the Core facility for Advanced Computing and Data Science at the University of Houston.

References

[1] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Towards open-set identity preserving face synthesis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 6713–6722, Salt Lake City, UT, Jun. 18–22 2018. 2

[2] K. Cao, Y. Rong, C. Li, X. Tang, and C. C. Loy. Pose-robust face recognition via deep residual equivariant mapping. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, Jun. 18–22 2018. 2

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proc. IEEE Conference on Automatic Face and Gesture Recognition*, Xi'an, China, May 15–19 2018. 1, 7

[4] G. Castañón and J. Byrne. Visualizing and quantifying discriminative features for face recognition. In *Proc. International Conference on Automatic Face and Gesture Recognition*, pages 16–23, Xi'an, China, May 15–19 2018. 3

[5] K. Chen, Y. Wu, H. Qin, D. Liang, X. Liu, and J. Yan. R3 adversarial network for cross model face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, Jun. 16–20 2019. 2

[6] L. Cheng, J. Wang, Y. Gong, and Q. Hou. Robust deep auto-encoder for occluded face recognition. In *Proc. ACM Multimedia Conference*, volume 1, pages 1099–1102, Queensland, Australia, 2015. 2

[7] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 5669–5678, Honolulu, HI, Jul. 21–26 2017. 3

[8] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou. UV-GAN: Adversarial facial UV map completion for pose-invariant face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7093–7102, Salt Lake City, UT, Jun. 19–21 2018. 2

[9] J. Deng, J. Guo, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, Jun. 16–20 2019. 1, 2, 7, 8

[10] Y. Duan, J. Lu, and J. Zhou. Uniformface: Learning deep equidistributed representation for face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, Jun. 16–20 2019. 2

[11] Y. Fu, X. Wu, Y. Wen, and Y. Xiang. Efficient locality-constrained occlusion coding for face recognition. *Neurocomputing*, 260:104–111, 2017. 2

[12] J. Guo, X. Zhu, Z. Lei, and S. Z. Li. Face synthesis for eyeglass-robust face recognition. *arXiv preprint arXiv:1806.01196*, pages 1–10, 2018. 2

[13] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Proc. European Conference on Computer Vision*, pages 87–102, Amsterdam, The Netherlands, 2016. 7

[14] K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. In *Proc. European Conference in Computer Vision*, pages 630–645, Amsterdam, Netherlands, Oct. 8–16 2016. 1, 6, 7, 8

[15] L. He, H. Li, Q. Zhang, and Z. Sun. Dynamic feature learning for partial face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7054–7063, Salt Lake City, UT, Jun. 18–22 2018. 2

[16] J. Hu, L. Shen, and G. Sun. Squeeze-and-Excitation Networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, Salt Lake City, UT, Jun. 18–22 2018. 2

[17] W. Hu, Y. Huang, F. Zhang, and R. Li. Noise-tolerant paradigm for training face recognition cnns. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 11887–11896, Long Beach, CA, Jun. 16–20 2019. 2

[18] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, MA, 2007. 7

[19] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In *Proc. IEEE International Conference on Computer Vision*, pages 2439–2448, Venice, Italy, Oct. 22–29 2017. 2

[20] S. Jetley, N. A. Lord, N. Lee, and P. H. S. Torr. Learn to pay attention. In *Proc. International Conference on Learning Representations*, pages 1–12, Vancouver, Canada, Apr. 30–May 3 2018. 3, 6, 7, 8

[21] H. A. Le and I. A. Kakadiaris. UHDB31: A dataset for better understanding face recognition across pose and illumination variation. In *Proc. IEEE International Conference on Computer Vision Workshops*, pages 2555–2563, Venice, Italy, Oct. 22–29 2017. 6

[22] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, Jun. 18–22 2018. 2

[23] B. Liu, W. Deng, Y. Zhong, M. Wang, J. Hu, X. Tao, and H. Yaohai. Fair loss: Margin-aware reinforcement learning for deep face recognition. In *Proc. International Conference on Computer Vision*, Seoul, South Korea, Oct. 27–Nov. 2 2019. 2

[24] H. Liu, X. Zhu, Z. Lei, and S. Z. Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, Jun. 16–20 2019. 2

[25] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. SphereFace: Deep hypersphere embedding for face recognition. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 6738–6746, Jul. 21–26 2017. 2, 8

[26] Z. Liu, P. Luo, X. Wang, and T. Xiaoou. Deep Learning Face Attributes in the Wild. In *Proc. International Conference on Computer Vision*, Dec. 11–18 2015. 3, 6

[27] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4838–4846, Las Vegas, NV, Jun. 26–Jul. 1 2016. 2

[28] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. IARPA Janus Benchmark-C: Face dataset

and protocol. In *Proc. IEEE International Conference on Biometrics*, pages 158–165, Queensland, Australia, Feb. 20–23 2018. 7, 8

[29] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. British Machine Vision Conference*, Swansea, United Kingdom, Sep. 7–10 2015. 8

[30] R. Ranjan, A. Bansal, H. Xu, S. Sankaranarayanan, J.-C. Chen, C. D. Castillo, and R. Chellappa. Crystal loss and quality pooling for unconstrained face verification and recognition. *arXiv preprint arXiv: 1804.01159*, 14(8):1–15, 2018. 3

[31] P. Rodríguez, J. M. Gonfaus, G. Cucurull, F. X. Roca, and J. Gonzàlez. Attend and Rectify: a Gated Attention Mechanism for Fine-Grained Recovery. In *Proc. European Conference on Computer Vision*, Munich, Germany, Sep. 8–14 2018. 3

[32] N. Sarafianos, X. Xu, and I. A. Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *Proc. European Conference on Computer Vision*, Munich, Germany, Sep. 8–14 2018. 2, 4, 6

[33] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, Boston, MA, Jun. 8–10 2015. 1, 7, 8

[34] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *Proc. IEEE Winter Conference on Applications of Computer Vision*, Lake Placid, NY, 2016. 1, 6, 7

[35] Y. Shi and A. K. Jain. Improving face recognition by exploring local features with visual attention. In *Proc. IEEE International Conference on Biometrics*, Gold Coast, Australia, Feb. 20–23 2018. 3

[36] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In *Proc. International Conference on Computer Vision*, Seoul, South Korea, Oct. 27–Nov. 2 2019. 2

[37] Y. Taigman, M. Y. Marc', A. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, Columbus, Ohio, Jun. 24–27 2014. 1

[38] L. Tran, X. Yin, and X. Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1424, Honolulu, Hawaii, Jul. 21–26 2017. 2, 7

[39] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):372–386, 2012. 2

[40] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017. 2

[41] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. CosFace: Large margin cosine loss for deep face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, Jun. 18–22 2018. 2, 8

[42] J. Wang, Y. Yuan, and G. Yu. Face attention network: An effective face detector for the occluded faces. *arXiv preprint arXiv:1711.07246*, 2017. 3

[43] R. Wang, J. Lu, and Y.-P. Tan. Robust point set matching for partial face recognition. *IEEE Transaction on Image Processing*, 25(3):1163–1176, 2016. 2

[44] Y. Wang, D. Gong, Z. Zhou, X. Ji, H. Wang, Z. Li, W. Liu, and T. Zhang. Orthogonal deep features decomposition for age-invariant face recognition. In *Proc. European Conference in Computer Vision*, Munich, Germany, Sep. 8–14 2018. 2

[45] Y. Wen, K. Zhang, and Zhifeng Li and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Proc. European Conference on Computer Vision*, pages 499–515, Amsterdam, Netherlands, Oct. 8–16 2016. 2

[46] W. Xie, L. Shen, and A. Zisserman. Comparator networks. In *Proc. European Conference on Computer Vision*, pages 782–797, Munich, Germany, Sep. 8–14 2018. 3

[47] W. Xie and A. Zisserman. Multicolumn networks for face recognition. In *Proc. British Machine Vision Conference*, pages 1–12, Northumbria University, United Kingdom, Sep. 3–6 2018. 3, 8

[48] X. Xu and I. A. Kakadiaris. FaRE: Open source face recognition performance evaluation package. In *Proc. IEEE International Conference on Image Processing*, pages 3272–3276, Sep. 22–25 2019. 6

[49] X. Xu, H. Le, P. Dou, Y. Wu, and I. A. Kakadiaris. Evaluation of 3D-aided pose invariant face recognition system. In *Proc. International Joint Conference on Biometrics*, pages 446–455, Denver, CO, Oct. 1–4 2017. 2, 7

[50] X. Xu, H. Le, and I. Kakadiaris. On the importance of feature aggregation for face reconstruction. In *Proc. IEEE Winter Conference on Applications of Computer Vision*, pages 922–931, Jan. 8–10 2019. 2

[51] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Honolulu, HI, Jul. 21–26 2017. 3

[52] X. Yin and X. Liu. Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Trans. on Image Processing*, 27(2):964–975, Feb. 2018. 7

[53] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *Proc. International Conference on Computer Vision*, pages 4010–4019, Venice, Italy, Oct. 22–29 2017. 2

[54] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 5704–5713, Long Beach, CA, Jun. 16–20 2019. 2

[55] Y. F. Yu, D. Q. Dai, C. X. Ren, and K. K. Huang. Discriminative multi-scale sparse coding for single-sample face recognition with occlusion. *Pattern Recognition*, 66:302–312, 2017. 2

[56] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan. Robust LSTM-autoencoders for face de-occlusion in the wild. *IEEE Transactions on Image Processing*, 27(2):778–790, 2018. 2

[57] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, S. Yan, and J. Feng. Towards pose invariant face recognition in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2207–2216, Salt Lake City, UT, Jun. 18–22 2018. 2

[58] J. Zhao, J. Li, X. Tu, F. Zhao, Y. Xin, J. Xing, H. Liu, S. Yan, and J. Feng. Multi-prototype networks for unconstrained set-based face recognition. In *Proc. International Joint Conference on Artificial Intelligence*, Macao, China, Aug. 10–26 2019. 2

[59] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng. 3D-Aided dual-agent GANs for unconstrained face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8828:1–14, 2018. 2

[60] K. Zhao, J. Xu, and M.-M. Cheng. Regularface: Deep face recognition via exclusive regularization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, Jun. 16–20 2019. 2

[61] S. Zhao and Z. P. Hu. A modular weighted sparse representation based on fisher discriminant and sparse residual for face recognition with occlusion. *Information Processing Letters*, 115(9):677–683, 2015. 2

[62] Y. Zheng, D. K. Pal, and M. Savvides. Ring loss: Convex feature normalization for face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 5089–5097, Salt Lake City, UT, Jun. 19–21 2018. 2

[63] Y. Zhong, J. Chen, and B. Huang. Toward end-to-end face recognition through alignment learning. *IEEE Signal Processing Letters*, 24(8):1213–1217, 2017. 2

[64] Y. Zhong, W. Deng, M. Wang, J. Hu, J. Peng, X. Tao, and Y. Huang. Unequal-training for deep face recognition with long-tailed noisy data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, Jun. 16–20 2019. 2

[65] E. Zhou, Z. Cao, and J. Sun. GridFace: Face rectification via learning local homography transformations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3–19, Salt Lake City, UT, Jun. 18–22 2018. 2

[66] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 5513–5522, Honolulu, HI, Jul. 21–26 2017. 2