Racial Faces in-the-Wild: Reducing Racial Bias by Information Maximization Adaptation Network

Mei Wang[1], Weihong Deng[1*], Jiani Hu[1], Xunqiang Tao[2], Yaohai Huang[2]
[1]Beijing University of Posts and Telecommunications, [2]Canon Information Technology (Beijing) Co., Ltd
[1]{wangmei1, whdeng, jnhu}@bupt.edu.cn, [2]{taoxunqiang, huangyaohai}@canon-ib.com.cn

Abstract

Racial bias is an important issue in biometric, but has not been thoroughly studied in deep face recognition. In this paper, we first contribute a dedicated dataset called Racial Faces in-the-Wild (RFW) database, on which we firmly validated the racial bias of four commercial APIs and four state-of-the-art (SOTA) algorithms. Then, we further present the solution using deep unsupervised domain adaptation and propose a deep information maximization adaptation network (IMAN) to alleviate this bias by using Caucasian as source domain and other races as target domains. This unsupervised method simultaneously aligns global distribution to decrease race gap at domain-level, and learns the discriminative target representations at cluster level. A novel mutual information loss is proposed to further enhance the discriminative ability of network output without label information. Extensive experiments on RFW, GBU, and IJB-A databases show that IMAN successfully learns features that generalize well across different races and across different databases.

1. Introduction

The emergence of deep convolutional neural networks (CNN) [38, 55, 59, 31, 32] greatly advances the frontier of face recognition (FR) [63, 58, 54]. However, more and more people find that a problematic issue, namely racial bias, has always been concealed in the previous studies due to biased benchmarks but it explicitly degrades the performance in realistic FR systems [2, 13, 25, 8]. For example, Amazon's Rekognition Tool incorrectly matched the photos of 28 U.S. congressmen with the faces of criminals, especially the error rate was up to 39% for non-Caucasian people. Although several studies [49, 29, 23, 50, 36] have uncovered racial bias in non-deep FR algorithms, this field still remains to be vacant in deep learning era because so little testing information available makes it hard to measure the racial bias.

To facilitate the research towards this issue, in this work we construct a new Racial Faces in-the-Wild (RFW) database, as shown in Fig. 1 and Table 4, to fairly measure racial bias in deep FR. Based on experiments on RFW, we find that both commercial APIs and SOTA algorithms indeed suffer from racial bias: the error rates on African faces are about two times of Caucasians, as shown in Table 1. To investigate the biases caused by training data, we also collect a race-balanced training database, and validate that racial bias comes on both data and algorithm aspects. Some specific races are inherently more difficult to recognize even trained on the race-balanced training data. Further research efforts on algorithms are requested to eliminate racial bias.

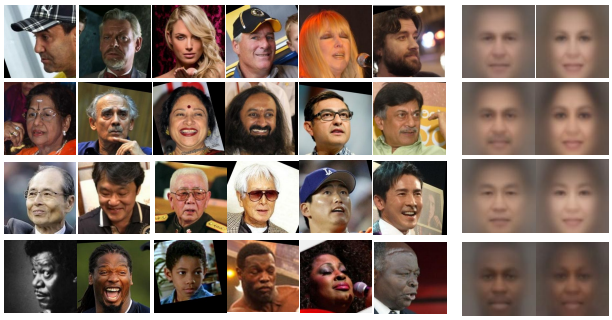


Figure 1. Examples and average faces of RFW database. In rows top to bottom: Caucasian, Indian, Asian, African.

Unsupervised domain adaptation (UDA) [64] is one of the promising methodologies to address algorithm biases, which can map two domains into a domain-invariant feature space and improve target performances in an unsupervised manner [61, 40, 60, 24]. Unfortunately, most UDA methods for object recognition are not applicable for FR because of two unique challenges. First, face identities (classes) of two domains are non-overlapping in FR, so that many skills in state-of-the-art (SOTA) methods based on sharing classes are inapplicable. Second, popular methods by the global alignment of source and target domain are insufficient to acquire the discriminating power for classification in FR. How to meet these two challenges is meaningful but few works have been proposed in this community.

	Model	RFW			
		Caucasian	Indian	Asian	African
commercial API	Microsoft [5]	87.60	82.83	79.67	75.83
	Face++ [4]	93.90	88.55	92.47	87.50
	Baidu [3]	89.13	86.53	90.27	77.97
	Amazon [1]	90.45	87.20	84.87	86.27
	mean	90.27	86.28	86.82	81.89
SOTA algorithm	Center-loss [65]	87.18	81.92	79.32	78.00
	Sphereface [39]	90.80	87.02	82.95	82.28
	Arcface[1][21]	92.15	88.00	83.98	84.93
	VGGface2 [15]	89.90	86.13	84.93	83.38
	mean	90.01	85.77	82.80	82.15

[1] Arcface here is trained on CASIA-Webface using ResNet-34.

Table 1. Racial bias in deep FR systems. Verification accuracies (%) evaluated on 6000 difficult pairs of RFW database are given.

In this paper, we propose a new information maximization adaptation network (IMAN) to mitigate racial bias, which matches global distribution at domain-level, at the meantime, learns discriminative target distribution at cluster-level. To circumvent the non-overlapping classes between two domains, IMAN applies a spectral clustering algorithm to generate pseudo-labels, by which the network is pre-adapted with Softmax and the target performance is enhanced preliminarily. This clustering scheme of IMAN is fundamentally different from other UDA methods [51, 69, 16, 18] that are inapplicable to FR. Besides pseudo label based pre-adaptation, a novel mutual information (MI) based adaptation is proposed to further enhance the discriminative ability of the network output, which learns larger decision margins in an unsupervised way. Different from the common supervised losses and supervised MI methods [56, 34], MI loss takes advantage of all unlabeled target data, no matter whether they are successfully assigned pseudo-labels or not, in virtue of its unsupervised property.

Extensive experimental results show that IMAN conducted to transfer recognition knowledge from Caucasian (source) domain to other-race (target) domains. Its performance is much better than other UDA methods. Ablation study shows that MI loss has unique effect on reducing racial bias. In addition, IMAN is also helpful in adapting general deep model to a specific database, and achieved improved performance on GBU [48] and IJB-A [37] databases. The contributions of this work are three aspects. 1) A new RFW dataset is constructed and is released [1] for the study on racial bias. 2) Comprehensive experiments on RFW validate the existence and cause of racial bias in deep FR algorithms. 3) A novel IMAN solution is introduced to address racial bias.

[1] http://www.whdeng.cn/RFW/index.html

2. Related work

Racial bias in face recognition. Several studies [49, 29, 23, 50, 36] have uncovered racial bias in non-deep face recognition algorithms. The FRVT 2002 [49] showed that recognition accuracies depend on demographic cohort. Phillips et al. [50] evaluated FR algorithms on the images of FRVT 2006 [11] and found that algorithms performed better on natives. Klare et al. [36] collected mug shot face images of White, Black and Hispanic from the Pinellas County Sheriff's Office (PCSO) and concluded that the Black cohorts are more difficult to recognize. In deep learning era, existing racial bias databases are no longer suitable for deep FR algorithms due to their small scale and constrained conditions; commonly-used testing databases of deep FR, e.g. LFW [33], IJB-A [37], don't include significant racial diversity, as shown in Table 2. Although some studies, e.g. unequal-training [9] and suppressing attributes [8, 43, 44, 42], have made effort to mitigate racial and gender bias in several computer vision tasks, this study remains to be vacant in FR. Thus, we construct a new RFW database to facilitate the research towards this issue.

Train/ Test	Database	Racial distribution (%)			
		Caucasian	Asian	Indian	African
train	CASIA-WebFace [67]	84.5	2.6	1.6	11.3
	VGGFace2 [15]	74.2	6.0	4.0	15.8
	MS-Celeb-1M [30]	76.3	6.6	2.6	14.5
test	LFW [33]	69.9	13.2	2.9	14.0
	IJB-A [37]	66.0	9.8	7.2	17.0
	RFW	**25.0**	**25.0**	**25.0**	**25.0**

Table 2. The percentage of different race in commonly-used training and testing databases

Deep unsupervised domain adaptation. UDA [64] utilizes labeled data in relevant source domains to execute new tasks in a target domain [61, 40, 41, 24, 60]. However, the research of UDA is limited to object classification, very few studies have focused on UDA for FR task. Luo et al. [70] integrated the maximum mean discrepancies (MMD) estimator to CNN to decrease domain discrepancy. Sohn et al. [57] synthesized video frames from images by a set of transformations and applied a domain adversarial discriminator to align feature space of image and video domains. Kan et al. [35] utilized the sparse representation constraint to ensure that source domain shares similar distribution as target domain. In this paper, inspired by Inception Score [52, 10] used in Generative Adversarial Nets (GAN), we introduce MI as a regularization term to domain adaptation and propose a novel IMAN method to address this unique challenge of FR in an unsupervised way.

3. Racial Faces in-the-Wild: RFW

Instead of downloading images from websites, we collect them from MS-Celeb-1M [6]. We use the "Nationality" attribute of FreeBase celebrities [27] to directly select Asians and Indians. For Caucasians and Africans, Face++ API [4] is used to estimate race. An identity will be accepted only if its most images are estimated as the same race, otherwise it will be abandoned. To avoid the negative effects caused by the biased Face++ tool, we manually check some images with low confidence scores from Face++.

Then we construct our RFW database with four testing subsets, namely Caucasian, Asian, Indian and African. Each subset contains about 10K images of 3K individuals for face verification. All of these images have been carefully and manually cleaned. Besides, in order to exclude overlapping identities between RFW and commonly-used training datasets, we further remove the overlapping subjects by manual inspection, when the subject and its nearest neighbor in CASIA-Webface and VGGFace2 (based on Arcface [21] feature) are found to be of the same identity.

For the performance evaluation, we recommend to use both the biometric receiver operating characteristic (ROC) curve and LFW-like protocol. Specifically, ROC curve, which aims to report a comprehensive performance, evaluates algorithms on all pairs of 3K identities (about 14K positive vs. 50M negative pairs). In contrast, LFW-like protocol facilitates easy and fast comparison between algorithms with 6K pairs of images. Further, inspired by the ugly subset of GBU database [48], we have selected the "difficult" pairs (in term of cosine similarity) to avoid the saturated performance to be easily reported [2].

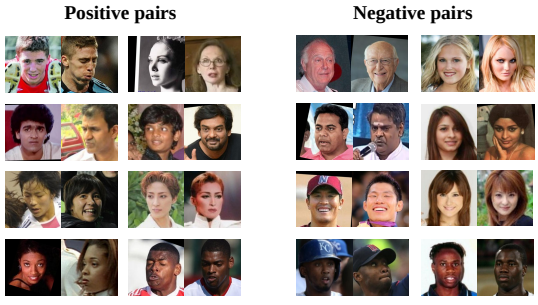


Figure 2. Examples of pairs in RFW database. We select 6K difficult pairs according to cosine similarity to avoid saturated performance, these images challenge the recognizer by variations of same people and the similar appearance of different people.

In RFW, the images of each race are randomly collected from MS-Celeb-1M without any preference, and thus they are suitable to fairly measure racial bias. We have validated that, across varying races, their distributions of pose, age, and gender are similar. As evidence, the detailed distributions measured by Face++ API are show in Fig. 3(a)-3(d). One can see from the figures that there is no significant difference between different races.

Moreover, the pose and age gap distributions of 3K difficult positive pairs are show in Fig. 3(e) and 3(f), which indicates that the selected difficult pairs are also fair across different races and contain larger intra-person variations. And Fig. 2 presents some examples of the 6K selected pairs, and one can see from the figure that some pairs are very challenging even for human.

4. Information maximization adaptation network

In our study, source domain is a labeled training set, namely $\mathcal{D}_s = \{x_i^s, y_i^s\}_{i=1}^M$ where x_i^s is the i-th source sample, y_i^s is its category label, and M is the number of source images. Target domain is an unlabeled training set, namely $\mathcal{D}_t = \{x_i^t\}_{i=1}^N$ where x_i^t is the i-th target sample and N is the number of target images. The data distributions of two domains are different, $P(X_s, Y_s) \neq P(X_t, Y_t)$. Our goal is to learn deep features invariant between domains and improve the performance of target images (faces of colored skin in our study) in an unsupervised manner. In the face recognition task, the identities (class) of two domains are non-overlapping, which poses a unique challenge different from other tasks.

4.1. Clustering-based pseudo labels for pre-adaptation

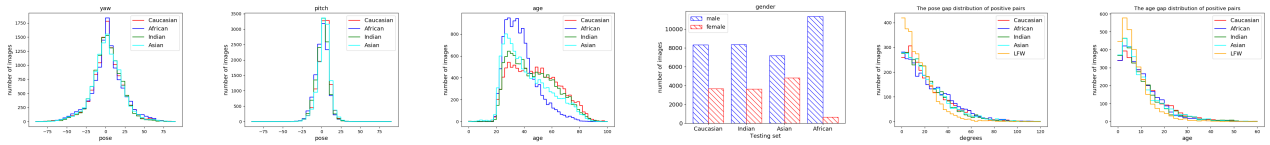
Previous UDA methods apply the source classifier to predict pseudo-labels in the target domain, by which the network can be fine-tuned using supervised losses [51, 69, 16, 18, 66]. Unfortunately, these well-established approaches are inapplicable in face recognition due to the non-overlapping identities between two domains. Therefore, we introduce a clustering algorithm into UDA to generate pseudo-labels for pre-adaptation training. The detailed steps of our clustering algorithm are given as following:

First, we feed unlabeled target data X_t into network and extract deep features $\mathcal{F}(X_t)$. Then, with these deep presentations, we construct a $N \times N$ adjacency matrix, where N is the number of faces in target domain and entry at (i, j), i.e. $s(i, j)$, is the cosine similarity between target face x_i^t and x_j^t.

Second, we can build a clustering graph $\mathcal{G}(n, e)$ according to adjacency matrix, where the node n_i represents i-th target image and edge $e(n_i, n_j)$ signifies that two target images have larger cosine-similarity than the parameter λ:

$$e(n_i, n_j) = \begin{cases} 1, & if\ s(i,j) > \lambda \\ 0, & otherwise \end{cases} \quad (1)$$

[2] All data and baseline code for evaluating will be publicly available for the research purpose.



(a) yaw (b) pitch (c) age (d) gender (e) pose gap (f) age gap

Figure 3. RFW statistics. We show the (a) yaw pose, (b) pitch pose, (c) age and (d) gender distribution of 3000 identities in RFW, as well as (e) Pose gap distribution and (f) age gap distribution of positive pairs in LFW and RFW.

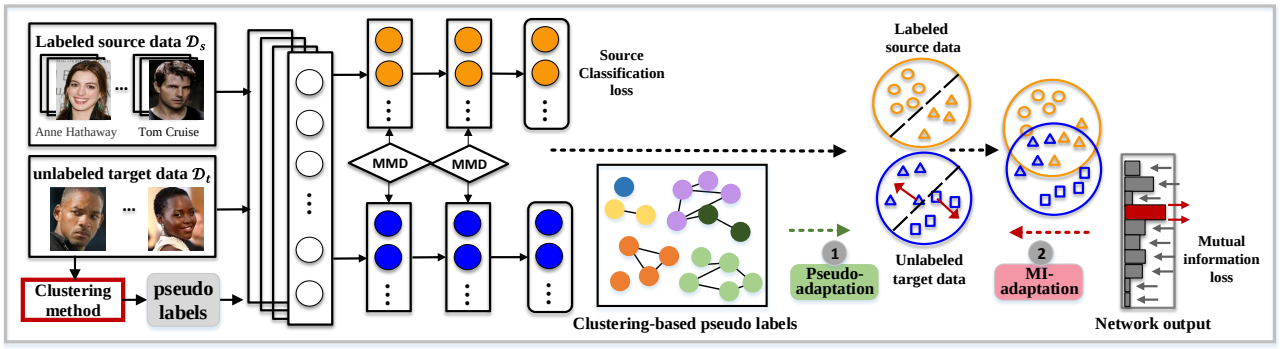


Figure 4. Overview of IMAN architecture. **Step-1: Pseudo-adaptation.** Pseudo-labels of target images are generated by clustering algorithm and then are utilized to pre-adapt the network with supervision of Softmax to obtain preliminary improvement of target domain. **Step-2: MI-adaptation.** With mutual information loss, the distribution of target classifier's output is further optimized and larger decision margins are learned without any label information.

Then, we simply save each connected component with at least p nodes as a cluster (identity) and obtain pseudo-labels of these target images; the remaining images will be abandoned. So, we only obtain pseudo-labels of partial images with higher confidence to alleviate negative influence caused by falsely-labeled samples. After that, we pre-adapt the network with the standard Softmax loss.

4.2. Mutual information loss for discriminant adaptation

Although pre-adaptation has derived preliminary prediction of the target images, it is insufficient to boost the performance in target domain due to the imperfection of pseudo-labels. How can we take full advantage of the full set of target images and learn more discriminative representations? Based on the preliminary prediction, we propose to further optimize the distribution of classifier's output without any label information. Our idea is to learn large decision margins in feature space through enlarging the classifier's output of one class while suppressing those of other classes in an unsupervised way. Different from supervised mutual information [56, 19, 45, 34], our MI loss maximizes mutual information between unlabeled target data $\mathbf{X_t}$ and classifier's prediction $\mathbf{O_t}$ inspired by [68, 26].

Based on the desideratum that an ideal conditional distribution of classifier's prediction $p(\mathbf{O_t}|x_i^t)$ should look like $[0, 0, ..., 1, ..., 0]$, it's better to classify samples with large margin. Grandvalet [28] proved that a entropy term $\frac{1}{N}\sum_{i=1}^{N} H(\mathbf{O_t}|x_i^t)$ very effectively meets this requirement, because it is maximized when the distribution of classifier's prediction is uniform and vice versa. However, in the case of fully unsupervised learning, simply minimizing this entropy will cause that more decision boundaries are removed and most samples are assigned to the same class. Therefore, we prefer to uniform distribution of category. An estimate of the marginal distribution of classifier's prediction $p(\mathbf{O_t})$ is given as follows:

$$p(\mathbf{O_t}) = \int p(x_i^t) p(\mathbf{O_t}|x_i^t) dx_i^t = \frac{1}{N}\sum_{i=1}^{N} p\left(\mathbf{O_t}|x_i^t\right) \quad (2)$$

we suggest that maximizing the entropy of $\mathbf{O_t}$ can make samples assigned evenly across the categories of dataset.

In information theory, mutual information between X and Y , i.e. $I(X;Y)$, can be expressed as the difference of two entropy terms:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (3)$$

If X and Y are related by a deterministic, invertible function, then maximal mutual information is attained. In our case, we combine the two entropy terms and obtain mutual

information between data $\mathbf{X_t}$ and prediction $\mathbf{O_t}$:

$$\begin{aligned}\mathcal{L}_M &= \tfrac{1}{N}\sum_{i=1}^{N} H(\mathbf{O_t}|x_i^t) - \gamma H(\mathbf{O_t}) \\ &= \tfrac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_C} p(o_j^t|x_i^t) log p(o_j^t|x_i^t) - \gamma \sum_{j=1}^{N_C} p(o_j^t) log p(o_j^t) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N_C} p(x_i^t) p(o_j^t|x_i^t) log p(o_j^t|x_i^t) - \gamma \sum_{j=1}^{N_C} p(o_j^t) log p(o_j^t) \\ &= H\left[\mathbf{O_t}|\mathbf{X_t}\right] - \gamma H\left[\mathbf{O_t}\right] \approx -I(\mathbf{X_t};\mathbf{O_t})\end{aligned} \tag{4}$$

where the first term is the entropy of conditional distribution of $\mathbf{O_t}$ which can enlarge the classifier's output of one class while suppressing those of other classes; and the second term is the entropy of marginal distribution of $\mathbf{O_t}$ which can avoid most samples being assigned to the same class. N is the number of target images, and N_C is the number of target categories. But without groundturth labels, how can we obtain N_C and guarantee the accuracy of classifier's prediction? Benefiting from clustering-based pseudo labels, we utilize the number of clusters to substitute for N_C, and obtain preliminary prediction through pre-adaptation to guarantee accuracy for mutual information loss.

4.3. Adaptation network

As shown in Fig. 4, the architecture of IMAN consists of a source and target CNN, with shared weights. Maximum mean discrepancy (MMD) estimator [61, 40, 12, 14], which is a standard distribution distance metric to measure domain discrepancy, is adopted on higher layers of network which are called adaptation layers. We simply use a fork at the top of the network after the adaptation layer. The inputs of source CNN are source labeled images while those of target CNN are target unlabeled data. The goal of training is to minimize the following loss:

$$\mathcal{L} = \mathcal{L}_C(X_s, Y_s) + \alpha \sum_{l\in\mathcal{L}} MMD^2(D_s^l, D_t^l) + \beta \mathcal{L}_M(X_t) \tag{5}$$

where α and β are the parameters for the trade-off between three terms. $\mathcal{L}_M(X_t)$ is our mutual-information loss on unlabeled target data X_t. $\mathcal{L}_C(X_s, Y_s)$ denotes source classification loss on the source data X_s and the source labels Y_s. $\mathcal{D}_*^l$ is the l-th layer hidden representation for the source and target examples, and $MMD^2(D_s^l, D_t^l)$ is the MMD between the source and target evaluated on the l-th layer representation. The empirical estimate of MMD between two domains is defined as $MMD^2(D_s, D_t) = \left\| \frac{1}{M}\sum_{i=1}^{M}\phi(\mathrm{x}_i^s) - \frac{1}{N}\sum_{j=1}^{N}\phi(\mathrm{x}_j^t) \right\|_H^2$, where ϕ represents the function that maps the original data to a reproducing kernel Hilbert space.

The entire procedure of IMAN is depicted in Algorithm 1. Source classification loss supervises learning proceeds for source domain. MMD minimizes the domain discrepancy to learn domain-invariant representations. Additionally, in the pre-training stage, MMD provides more reliable underlying target representations for clustering leading to higher quality of pseudo-labels. Clustering-based pseudo-labels can improve the performance of target domain preliminarily and guarantee the accuracy of network's prediction for unsupervised MI loss. MI loss can further take full advantage of all target data, no matter whether they are successfully clustered or not, to learn larger decision margins and enhance the discrimination ability of network for target domain.

Algorithm 1 Information Maximization Adaptation Network (IMAN).

Input:
Source domain labeled samples $\{x_i^s, y_i^s\}_{i=1}^M$, and target domain unlabeled samples $\{x_i^t\}_{i=1}^N$.

Output:
Network layer parameters Θ.

1: ***Stage-1: // Pre-training:***
2: Pre-train network by MMD [61] and source classification loss to minimize domain discrepancy and provide more reliable target representations for clustering;
3: **Repeat:**
4: ***Stage-2: // Pre-adaptation:***
5: Adopt clustering algorithms to generate pseudo-labels of partial target images according to Eqn. (1); Pre-adapt the network on them with supervision of Softmax to obtain preliminary improvement of target domain;
6: ***Stage-3: // MI-adaptation:***
7: Adapt the network with mutual information loss according to Eqn. (5) to further enhance the discrimination ability of network output;
8: **Until convergence**

5. Experiments on RFW

5.1. Racial bias experiment

Experimental Settings. We use the similar ResNet-34 architecture described in [21]. It is trained with the guidance of Arcface loss [21] on the CAISA-Webface [67], and is called Arcface(CASIA) model. CASIA-Webface consists of 0.5M images of 10K celebrities in which 85% of the photos are Caucasians. For preprocessing, we use five facial landmarks for similarity transformation, then crop and resize the faces to 112×112. Each pixel ([0, 255]) in RGB images is normalized by subtracting 127.5 and then being

divided by 128. We set the batch size, momentum, and weight decay as 200, 0.9 and $5e-4$, respectively. The learning rate is started from 0.1 and decreased twice with a factor of 10 when errors plateau.

Existence of racial bias. We extract features of 6000 pairs in RFW by our Arcface(CASIA) model and compare the distribution of cosine-distances, as shown in Fig. 5(c). The distribution of Caucasian has a more distinct margin than that of other races, which visually proves the recognition errors of non-Caucasian subjects are much higher. Then, we also examine some SOTA algorithms, i.e. Centerloss [65], Sphereface [39], VGGFace2 [15] and ArcFace [21], as well as four commercial recognition APIs, i.e. Face++, Baidu, Amazon, Microsoft on our RFW. The biometric ROC curves evaluated on all pairs are presented in Fig. 6; the accuracies in LFW-like protocol are given in Table 1 and its ROC curves are given in the Supplementary Material. First, all SOTA algorithms and APIs perform the best on Caucasian testing subset, followed by Indian, and the worst on Asian and African. This is because that the learned representations predominantly trained on Caucasians will discard information useful for discerning non-Caucasian faces. Second, a phenomenon is found coincident with [11]: APIs which are developed by East Asian companies perform better on Asians, while APIs developed in the Western hemisphere perform better on Caucasians.

Existence of domain gap. The visualization and quantitative comparisons are conducted at feature level. The deep features of 1.2K images are extracted by our Arcface(CASIA) model and are visualized respectively using t-SNE embeddings [22], as shown in Fig. 5(a). The features almost completely separate according to race. Moreover, we use the MMD to compute distribution discrepancy between the images of Caucasians and other races in Fig. 5(b). From the figures, we make the same conclusions: the distribution discrepancies between Caucasians and other races are much larger than that between Caucasians themselves, which conforms that there is domain gap between races.

Cause of racial bias. We download more images of non-Caucasians from Website according to FreeBase celebrities [27], and construct an Equalizedface dataset. It contains 590K images from 14K celebrities which has the similar scale with CASIA-Webface database but is approximately race-balanced with 3.5K identities per race. Using Equalizedface as training data, we train an Arcface(Equal) model in the same way as Arcface(CASIA) model and compare their performances on 6000 difficult paris of RFW, as shown in Table 3. Compared with Arcface(CASIA) model, Arcface(Equal) model trained equally on all races performs much better on non-Caucasians which proves that racial bias in databases will reflect in FR algorithm. However, even with balanced training, we see that non-Caucasians still perform poorly than Caucasians. The reason may be that faces of colored skin are more difficult to extract and preprocess feature information, especially in dark situations. Moreover, we also train specific models on 7K identities of the same race, its performance is a bit lower compared to balanced training (3.5K people for each race). We believe there exists cooperative relationships among different races due to similar low-level features so that this mixture of races would improve the recognition ability.

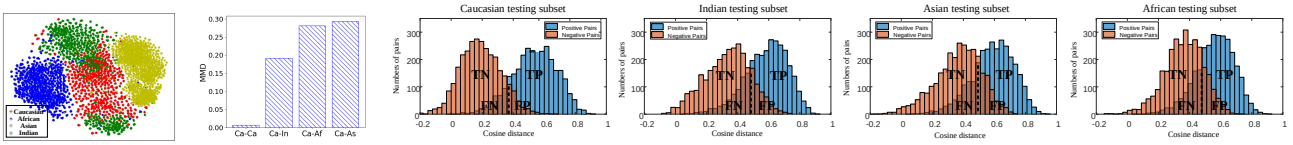
5.2. Domain adaptation experiment

Datasets. A training set with four race-subsets is also constructed according to RFW. One training subset consists of about 500K labeled images of 10k Caucasians and three other subsets contain 50K unlabeled images of non-Caucasians, respectively, as shown in Table 4. We use Caucasian as source domain and other races as target domains, and evaluate algorithms on 6000 pairs and all pairs of RFW.

Implementation detail. For preprocessing, we share the uniform alignment methods as Arcface(CASIA) model as mentioned above. For MMD, we follow the settings in DAN [40], and apply MMD to the last two fully-connected layers. In all experiments, we use ResNet-34 as backbone and set the batch size, momentum, and weight decay as 200, 0.9 and $5e-4$, respectively. In pre-training stage, the learning rate is started from 0.1 and decreased twice with a factor of 10 when errors plateau. In pre-adaptation stage, we pre-adapt network on pseudo-labeled target samples and source samples using learning rate of $5e-3$. In MI-adaptation stage, we adapt the network with learning rate of $1e-3$ using all source and target data. In IMAN-A(Arcface), Arcface [21] is used as source classification loss and the parameter α, β and γ are set to be 10, 5 and 0.2, respectively. In IMAN-S(Softmax), Softmax is used as source classification loss and the parameter α, β and γ are set to be 2, 5 and 0.2.

Experimental result. Three UDA tasks are performed, namely transferring knowledge from Caucasian to Indian, Asian and African. Due to the particularity of task, very few studies have focused on UDA in FR task. The latest work is performed by Luo et al. [70] who utilizes MMD-based method, i.e. DDC [61] and DAN [40], to perform scene adaptation. Therefore, we also compare our IMAN with these two UDA methods. DDC adopts single-kernel MMD on the last fully-connected layers; DAN adopts multi-kernel MMD on the last two fully-connected layers.

From Table 5 and Fig. 7, we have the following observations. First, without adaptation, Arcface, which published in CVPR'19 and reported SOTA performance on the LFW and MegaFace challenges, can not obtain perfect performance on non-Caucasians due to race gap. Second, MMD-based methods, i.e. DDC and DAN, obtain limited improvement compared with Softmax and Arcface model, which confirms our thought that the popular methods by the global alignment of source and target domain are insufficient for

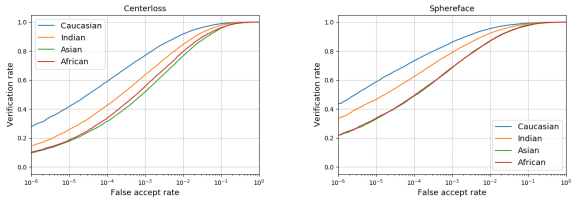


(a) T-SNE (b) MMD (c) Distribution of cosine-distances of 6000 pairs

Figure 5. (a) The feature space of four testing subsets. Each color dot represents a image belong to Caucasian, Indian, Asian or African. (b) The distribution discrepancy between Caucasians and other races measured by MMD. 'Ca', 'As', 'In' and 'Af' represent Caucasian, Asian, Indian and African, respectively. (c) Distribution of cosine-distances of 6000 pairs on Caucasian, Indian, Asian and African subset.

Training Databases	LFW	CFP-FP	AgeDB-30	Caucasian	Indian	Asian	African
CASIA-WebFace [67]	99.40	**93.91**	93.35	92.15	88.00	83.98	84.93
Equalizedface (ours)	**99.55**	92.74	**95.15**	**93.92**	**92.98**	90.60	**90.98**
Caucasian-7000	99.20	88.00	94.61	93.68	-	-	-
Indian-7000	98.53	90.80	86.47	-	90.37	-	-
Asian-7000	98.05	87.71	86.05	-	-	**91.27**	-
African-7000	98.45	86.44	89.62	-	-	-	90.88

Table 3. Verification accuracy (%) of ResNet-34 models trained with different training datasets.



(a) Center loss (b) Spereface

(c) Arcface (d) VGGFace2

Figure 6. The ROC curves of (e) Center loss, (f) Spereface (g) Arcface, (h) VGGFace2 evaluated on all pairs.

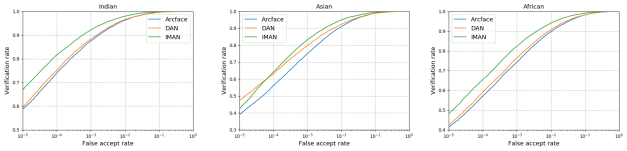
Subsets	Train		Test	
	# Subjects	# Images	# Subjects	# Images
Caucasian	10000	468139	2959	10196
Indian	-	52285	2984	10308
Asian	-	54188	2492	9688
African	-	50588	2995	10415

Table 4. Statistic of training and testing dataset.

face recognition. Third, we can find that our IMAN-S and IMAN-A both dramatically outperform all of the compared methods and IMAN-A achieves about 3% gains over Arcface model. Furthermore, when pre-adapting network with supervision of Arcface loss instead of Softmax loss in

Methods	Caucasian	Indian	Asian	African
Softmax	94.12	88.33	84.60	83.47
DDC-S [61]	-	90.53	86.32	84.95
DAN-S [40]	-	89.98	85.53	84.10
IMAN-S (ours)	-	**91.08**	**89.88**	**89.13**
Arcface [21]	94.78	90.48	86.27	85.13
DDC-A [61]	-	91.63	87.55	86.28
DAN-A [40]	-	91.78	87.78	86.30
IMAN-A (ours)	-	93.55	89.87	88.88
IMAN*-A (ours)	-	**94.15**	**91.15**	**91.42**

Table 5. Verification accuracy (%) on 6000 pairs of RFW dataset. "-S" represents the methods using Softmax as source classification loss; while "-A" represents the ones using Arcface.



(a) Indian set (b) Asian set (c) African set

Figure 7. The ROC curves of Arcface, DAN-A, and IMAN-A models evaluated on all pairs of (a) Indian, (b) Asian and (c) African set.

the second stage, our IMAN-A (denoted as IMAN*-A) is further improved, and obtains the best performances with 94.15%, 91.15% and 91.42% for Indian, Asian and African set. Especially, we further optimize IMAN*-A by performing pre-adaptation and MI-adaptation alternatively and iteratively in task Caucasian→African, and show the accuracy at each iteration in Fig. 8. The performance gradually increases until convergence.

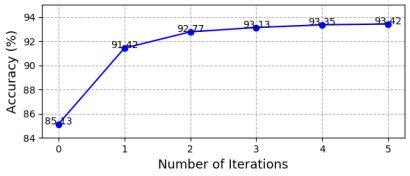


Figure 8. Verification accuracy of IMAN*-A at each iteration when performing pre-adaptation and MI-adaptation alternatively in task Caucasian→African. The value at the 0-th iteration means accuracy of Arcface tested on 6K pairs of African set.

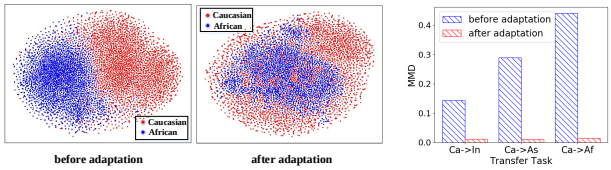
Ablation Study. IMAN consists of two main contributions comparing with existing UDA methods, i.e. pseudo-adaptation and MI-adaptation. To evaluate their effectiveness, we perform ablation study using Arcface loss as source classification loss. In Table 6, the results of IMAN w/o pseudo-labels are unsatisfactory because MI loss depends on pseudo-adaptation to guarantee the accuracy of classifier and only performing MI-adaptation with a randomly-initialized classifier is meaningless. To get a fair comparison, as we can see from the results of IMAN w/o MI, pseudo-adaptation is superior to baseline by about 2.3% on average, and our IMAN outperforms pseudo-adaptation by about 1.1% benefiting from MI-adaptation. It shows that each component has unique effect on reducing racial bias.

Methods	Indian	Asian	African
w/o pseudo-labels	91.02	86.88	85.52
w/o MI	92.08	88.80	88.12
IMAN-A (ours)	**93.55**	**89.87**	**88.88**

Table 6. Ablation study on 6000 pairs of RFW dataset.

Visualization. To demonstrate the transferability of the IMAN learned features, the visualization comparisons are conducted at feature level. First, we randomly extract the deep features of 10K source and target images in task Caucasian→African with Arcface model and IMAN-A model, respectively. The features are visualized using t-SNE, as shown in Fig. 9(a). After adaptation, more source and target data begin to mix in feature space so that there is no boundary between them. Second, we compute domain discrepancy between source and target domain using Arcface and IMAN-A activations respectively. Fig. 9(b) shows that discrepancy using IMAN-A features is much smaller than that using Arcface features. Therefore, we conclude that our IMAN does help to minimize domain discrepancy and align feature space between two domains benefited from MMD.

Additional experiments on IJB-A and GBU. Besides race gap, there are other domain gaps which make the learnt model degenerate in target domain, e.g. different lighting condition, pose and image quality. To validate our IMAN method, we further adopt it to reduce these domain gaps by



(a) Feature visualization (b) Domain discrepancy

Figure 9. (a) Feature visualization in task Caucasian→African. (b) Distribution discrepancy of source and target domain.

Method	Ugly	Bad	Good
LRPCA-face [48]	7.00	24.00	64.00
Fusion [47]	15.00	80.00	98.00
VGG [47]	26.00	52.00	85.00
Arcface(CASIA) [21]	75.00	90.32	96.21
DAN-A [40]	80.77	93.66	97.60
IMAN-A (ours)	**85.38**	**96.00**	**98.88**

Table 7. VR at FAR of 0.001 for GBU partitions.

Method	IJB-A: Verif. TAR@FAR's of			IJB-A: Identif.	
	0.001	0.01	0.1	Rank1	Rank10
Bilinear-CNN [20]	-	-	-	58.80	-
Face-Search [62]	-	73.30	-	82.00	-
Deep-Multipose [7]	-	78.70	-	84.60	94.70
Triplet-Similarity [53]	-	79.00	94.50	88.01	97.38
Joint Bayesian [17]	-	83.80	-	90.30	97.70
VGG [46]	64.19	84.02	96.09	91.11	**98.25**
Arcface(CASIA) [21]	74.19	87.11	94.87	90.68	96.07
DAN-A [40]	80.64	90.87	96.22	92.78	97.01
IMAN-A (ours)	**84.19**	**91.88**	**97.05**	**94.05**	98.04

Table 8. Verification performance (%) of IJB-A. "Verif" represents the 1:1 verification and "Identif." denotes 1:N identification.

using CASIA-Webface as source domain and using GBU [48] or IJB-A [37] as target domain. The images in CASIA-Webface are collected from Internet under unconstrained environment and most of the figures are celebrities taken in ambient lighting. GBU is split into three partitions with face pairs of different recognition difficulty, i.e. Good, Bad and Ugly. Each partition consists of a target set and a query set, and both them contain 1085 images of 437 distinct people. The images are frontal and are taken outdoors or indoors in atriums and hallways with digital camera. IJB-A contains 5,397 images and 2,042 videos of 500 subjects, and covers large pose variations and contains many blurry video frames. The results on GBU and IJB-A databases are shown in Table 7 and 8. After adaptation, our IMAN-A surpasses other compared methods, even better than Arcface(CASIA) model. In particular, it outperforms the SOTA counterparts by a large margin on the GBU, although it is only based on the unsupervised adaptation.

6. Conclusion

An ultimate face recognition algorithm should perform fairly on different races. We have done the first step and create a benchmark, i.e. RFW, to fairly evaluate racial bias. Through experiments on our RFW, we first verify the existence of racial bias. Then, we address it in the viewpoint of domain adaptation and design a novel IMAN method to bridge the domain gap and transfer knowledge between races. The comprehensive experiments prove the potential and effectiveness of our IMAN to reduce racial bias.

References

[1] Amazon's reignition tool. https://aws.amazon.com/rekognition/.

[2] Are face recognition systems accurate? depends on your race. https://www.technologyreview.com/s/601786.

[3] Baidu cloud vision api. http://ai.baidu.com.

[4] Face++ research toolkit. www.faceplusplus.com.

[5] Microsoft azure. https://www.azure.cn.

[6] Ms-celeb-1m challenge 3: Face feature test/trillion pairs. http://trillionpairs.deepglint.com/.

[7] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajan, et al. Face recognition using deep multi-pose representations. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016.

[8] M. Alvi, A. Zisserman, and C. Nellaker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. *arXiv preprint arXiv:1809.02169*, 2018.

[9] A. Amini, A. Soleimany, W. Schwarting, S. Bhatia, and D. Rus. Uncovering and mitigating algorithmic bias through learned latent structure. *AIES*, 2019.

[10] S. Barratt and R. Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

[11] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, and Y. M. Lui. Focus on quality, predicting frvt 2006 performance. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–8. IEEE, 2008.

[12] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[13] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, pages 77–91, 2018.

[14] R. Cafiero, A. Gabrielli, M. A. Mu&Ntilde, and oz. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[15] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. *arXiv preprint arXiv:1710.08092*, 2017.

[16] C. Chen, W. Xie, T. Xu, W. Huang, Y. Rong, X. Ding, Y. Huang, and J. Huang. Progressive feature alignment for unsupervised domain adaptation. *arXiv preprint arXiv:1811.08585*, 2018.

[17] J.-C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016.

[18] M. Chen, K. Q. Weinberger, and J. Blitzer. Co-training for domain adaptation. In *Advances in neural information processing systems*, pages 2456–2464, 2011.

[19] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[20] A. R. Chowdhury, T.-Y. Lin, S. Maji, and E. Learned-Miller. One-to-many face recognition with bilinear cnns. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[21] J. Deng, J. Guo, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018.

[22] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, pages 647–655, 2014.

[23] N. Furl, P. J. Phillips, and A. J. O'Toole. Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. *Cognitive Science*, 26(6):797–815, 2002.

[24] Y. Ganin. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015.

[25] C. Garvie. *The perpetual line-up: Unregulated police face recognition in america*. Georgetown Law, Center on Privacy & Technology, 2016.

[26] R. Gomes, A. Krause, and P. Perona. Discriminative clustering by regularized information maximization. In *NIPS*, pages 775–783, 2010.

[27] Google. Freebase data dumps. https://developers.google.com/freebase/data, 2015.

[28] Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005.

[29] P. J. Grother, G. W. Quinn, and P. J. Phillips. Report on the evaluation of 2d still-image face recognition algorithms. *NIST interagency report*, 7709:106, 2010.

[30] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, pages 87–102. Springer, 2016.

[31] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[32] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 2017.

[33] 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, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[34] B. Jun, T. Kim, and D. Kim. A compact local binary pattern using maximization of mutual information for face analysis. *Pattern Recognition*, 44(3):532–543, 2011.

[35] M. Kan, S. Shan, and X. Chen. Bi-shifting auto-encoder for unsupervised domain adaptation. In *ICCV*, pages 3846–3854, 2015.

[36] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

[37] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR*, pages 1931–1939, 2015.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[39] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, volume 1, 2017.

[40] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015.

[41] M. Long, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. *arXiv preprint arXiv:1605.06636*, 2016.

[42] V. Mirjalili, S. Raschka, A. Namboodiri, and A. Ross. Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images. In *2018 International Conference on Biometrics (ICB)*, pages 82–89. IEEE, 2018.

[43] V. Mirjalili, S. Raschka, and A. Ross. Gender privacy: An ensemble of semi adversarial networks for confounding arbitrary gender classifiers. *arXiv preprint arXiv:1807.11936*, 2018.

[44] A. Othman and A. Ross. Privacy of facial soft biometrics: Suppressing gender but retaining identity. In *European Conference on Computer Vision*, pages 682–696. Springer, 2014.

[45] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, pages 751–760. ACM, 2010.

[46] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[47] P. J. Phillips. A cross benchmark assessment of a deep convolutional neural network for face recognition. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 705–710. IEEE, 2017.

[48] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer. The good, the bad, and the ugly face challenge problem. *Image & Vision Computing*, 30(3):177–185, 2012.

[49] P. J. Phillips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone. Face recognition vendor test 2002. In *Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on*, page 44. IEEE, 2003.

[50] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2):14, 2011.

[51] K. Saito, Y. Ushiku, and T. Harada. Asymmetric tri-training for unsupervised domain adaptation. *arXiv preprint arXiv:1702.08400*, 2017.

[52] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[53] S. Sankaranarayanan, A. Alavi, and R. Chellappa. Triplet similarity embedding for face verification. *arXiv preprint arXiv:1602.03418*, 2016.

[54] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[55] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[56] M. Singh, S. Nagpal, M. Vatsa, R. Singh, and A. Noore. Supervised cosmos autoencoder: Learning beyond the euclidean loss! *arXiv preprint arXiv:1810.06221*, 2018.

[57] K. Sohn, S. Liu, G. Zhong, X. Yu, M.-H. Yang, and M. Chandraker. Unsupervised domain adaptation for face recognition in unlabeled videos. *arXiv preprint arXiv:1708.02191*, 2017.

[58] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS*, pages 1988–1996, 2014.

[59] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, et al. Going deeper with convolutions. CVPR, 2015.

[60] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *CVPR*, volume 1, page 4, 2017.

[61] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *Computer Science*, 2014.

[62] D. Wang, C. Otto, and A. K. Jain. Face search at scale: 80 million gallery. *arXiv preprint arXiv:1507.07242*, 2015.

[63] M. Wang and W. Deng. Deep face recognition: A survey. *arXiv preprint arXiv:1804.06655*, 2018.

[64] M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135 – 153, 2018.

[65] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515. Springer, 2016.

[66] S. Xie, Z. Zheng, L. Chen, and C. Chen. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 5419–5428, 2018.

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

[68] S. Yuan and S. Fei. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *ICML*, pages 1275–1282, 2012.

[69] W. Zhang, W. Ouyang, W. Li, and D. Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3801–3809, 2018.

[70] W. D. H. S. Zimeng Luo, Jiani Hu. Deep unsupervised domain adaptation for face recognition. In *FG*, pages 453–457. IEEE, 2018.