Racial Faces in-the-Wild: Reducing Racial Bias by Deep Unsupervised Domain Adaptation

Mei Wang[1], Weihong Deng[1]*, Jiani Hu[1], Jianteng Peng[2], 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]{pengjianteng, taoxunqiang, huangyaohai}@canon-ib.com.cn

Abstract

Despite of the progress achieved by deep learning in face recognition (FR), more and more people find that racial bias explicitly degrades the performance in realistic FR systems. Facing the fact that existing training and testing databases consist of almost Caucasian subjects, there are still no independent testing databases to evaluate racial bias and even no training databases and methods to reduce it. To facilitate the research towards conquering those unfair issues, this paper contributes a new dataset called Racial Faces in-the-Wild (RFW) database with two important uses, 1) racial bias testing: *four testing subsets, namely Caucasian, Asian, Indian and African, are constructed, and each contains about 3000 individuals with 6000 image pairs for face verification, 2)* racial bias reducing: *one labeled training subset with Caucasians and three unlabeled training subsets with Asians, Indians and Africans are offered to encourage FR algorithms to transfer recognition knowledge from Caucasians to other races. For we all know, RFW is the first database for measuring racial bias in FR algorithms. After proving the existence of domain gap among different races and the existence of racial bias in FR algorithms, we further propose a deep information maximization adaptation network (IMAN) to bridge the domain gap, and comprehensive experiments show that the racial bias could be narrowed-down by our algorithm.*

1. Introduction

The emergence of deep convolutional neural networks (CNN) [26, 35, 38, 21, 22] greatly advances the frontier of face recognition (FR) [43, 36, 34]. Through getting experience and knowledge from training data, deep networks simulate the perception of the human brain to perform FR and boost the performance to nearly 100% on the Labeled Faces in the Wild (LFW) dataset [23]. 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 explicitly degrades the performance in realistic FR systems [2, 10, 18, 7, 16, 32]. 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; according to [18], a year-long research investigation across 100 police departments revealed that African-American individuals are more likely to be stopped by law enforcement; Buolamwini et al. [10] found that the accuracies of 3 commercial gender classification algorithms drop largely on darker female faces. Based on these findings, MIT Technology Reviewer [2] suggested that racial bias in databases will reflect in algorithms, hence the performances of FR systems depend on the race. However, so little testing information available makes it hard to measure the racial bias in existing FR algorithms and there has yet to be a comprehensive study that investigates how deep FR algorithms are affected by it.

Model	LFW	RFW			
		Caucasian	Indian	Asian	African
Microsoft [5]	98.22	87.60	82.83	79.67	75.83
Face++ [4]	97.03	93.90	88.55	92.47	87.50
Baidu [3]	98.67	89.13	86.53	90.27	77.97
Amazon [1]	98.50	90.45	87.20	84.87	86.27
mean	98.11	90.27	86.28	86.82	81.89

Table 1. Racial bias in existing commercial recognition APIs. Face verification accuracies (%) on RFW database are given.

Aiming to facilitate the research towards this issue in deep FR, we construct a new Racial Faces in-the-Wild (RFW) database containing 625K images of 25K celebrities of different races as shown in Fig. 1 and Table 3. Different from existing datasets, RFW is collected to offer several new uses:

[*] Corresponding author

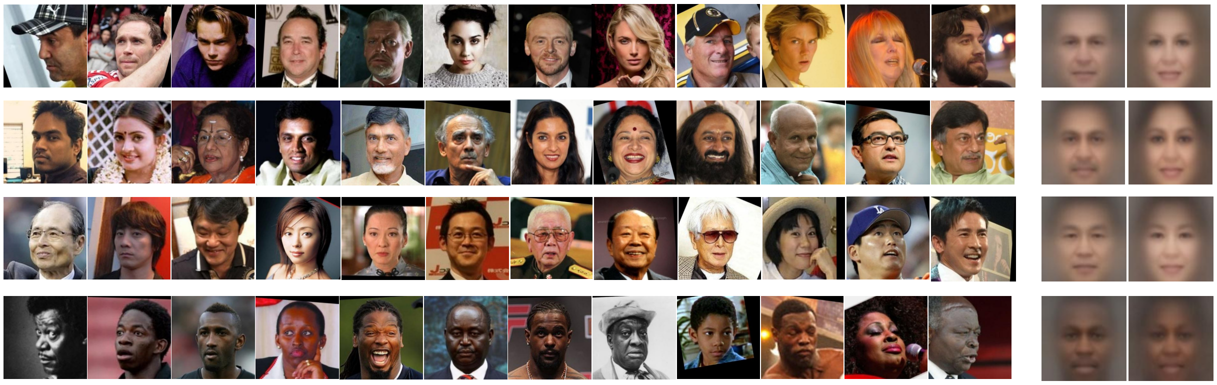


Figure 1. Examples and average faces of RFW database. In rows top to bottom: Caucasian, Indian, Asian, African. The last two columns present average faces of each race which consist of the average pixel values computed from all the aligned face images in a cohort. As the examples show, people of different races have certain differences in facial features and complexion.

1) Measure racial bias of FR algorithms. Four testing subsets, namely Caucasian, Asian, Indian and African, are constructed for face verification. Through setting a unified standard (e.g. similar distribution of pose, age and gender), four testing subsets of RFW ensure to exclude other factors except for race which can cause difference, and they can be used to fairly evaluate and compare the recognition ability of the algorithm on different races.

2) Reduce racial bias by transfer learning. A training set with four race-subsets is also released. The Caucasian subset consists of about 500K labeled images of 10k identities and other-race subsets contain 50K unlabeled images, respectively. We recommend to use transfer learning (TL) to transfer recognition knowledge among different races.

Based on our RFW, the first step have been done to verify the existence of racial bias in realistic FR systems through experiments. As shown in Table 1, existing commercial recognition APIs indeed work unequally well for different races, the maximum difference in error rate between the best and worst groups is 12%. This phenomenon has always been concealed in the previous papers, since the number of non-Caucasian people for test is also quite small. To reduce the racial bias, a domain adaptation (DA) approach specifically for FR is proposed, namely a deep information maximization adaptation network (IMAN). It identifies a feature space where data in the source and the target domains are similarly distributed, it also learns the feature space discriminatively, optimizing an mutual-information loss as an proxy to maximize the decision margin on the unlabeled target domain. Comprehensive experiments on our RFW show that the racial bias could be narrowed-down by IMAN.

Our contributions can be summarized into three aspects. 1) A new challenging RFW dataset is constructed and is released [1]. Compared with existing datasets, RFW can be used to fairly evaluate and study racial bias in FR algorithms. To the best of our knowledge, the RFW is the first benchmark that face pairs of different races are fairly integrated into the evaluation of the FR system. 2) Based on comprehensive experiments on RFW, we first prove that deep FR algorithms are also susceptible to ”other-race effect”. 3) An effective IMAN model is proposed, competitive results are delivered and show that racial bias can be well reduced by IMAN.

2. Related work

Database. In the last couple of years, various databases were continually developed to facilitate the FR research. In 2007, LFW dataset was introduced which marks the beginning of FR under unconstrained conditions. It contains 13,233 images of 5749 unique individuals, and provides 6000 pairs for face verification. CASIA-Webface [48] provided the first widely-used public training dataset for the deep model training purpose, which consists of 0.5M images of 10K celebrities, collected from the web. Currently, there have been more databases providing public available large-scale training data, especially three databases with over 1M images, namely MS-Celeb-1M [20], VGGface2 [12] and Megaface [24]. However, almost all databases contain significant racial bias, as shown in Table 2. Training on these databases may lead to unfair results on different races and testing on these databases will result in overlooking poor performance of non-Caucasian subjects.

Deep face recognition. Driven by graphics processing units (GPUs), massive annotated data and deep learning, deep FR [43] has made significant advances in recent years. More powerful loss functions are explored to learn deep features whose intra-class differences are small and inter-class differences are large. DeepFace [39] was the first to use a nine-layer CNN with softmax to perform FR. With 3D alignment for data processing, it reaches an accuracy of 97.35% on LFW. DeepID series [36, 46, 37] combined

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

Train/ Test	Database	Racial distribution (%)			
		Caucasian	Asian	Indian	African
train	CASIA-WebFace [48]	84.5	2.6	1.6	11.3
	VGGFace2 [12]	74.2	6.0	4.0	15.8
	MS-Celeb-1M [20]	76.3	6.6	2.6	14.5
test	LFW [23]	69.9	13.2	2.9	14.0
	IJB-A [25]	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

the softmax with contrastive loss to learn a discriminative representation. FaceNet [34] used a large private dataset to train a GoogleNet. It adopted a triplet loss function and achieved 99.63% on LFW. Wen et al. [45] proposed a center loss to reduce the intra-class features variations. Sphereface [27] and Arcface [14] are proposed to make learned features potentially separable with a larger angular distance that is equivalent to geodesic distance on a hypersphere manifold.

Deep unsupervised domain adaptation. Due to many factors, e.g., illumination, pose, and image quality, there is always a distribution change or domain shift between training and testing sets that can degrade the performance. Mimicking the human vision system, DA is a particular case of TL that utilizes labeled data in one or more relevant source domains to execute new tasks in a target domain [44]. Several methods have used the maximum mean discrepancy (MMD) loss for this purpose [42, 28, 29, 47, 30]. The deep domain confusion network (DDC) [42] used MMD in addition to the classification loss to reduce the distribution mismatch by one adaptation layer. Deep adaptation network (DAN) [28] matched the shift by adding multiple adaptation layers and exploring multiple kernels. Other methods have chosen an adversarial loss to minimize domain shift [17, 41, 40, 13]. The domain-adversarial neural network (DANN) [17] integrated a gradient reversal layer (GRL) into the standard architecture to ensure that the feature distributions over the two domains are made similar. However, due to lack of appropriate face databases, the research of DA is limited to digital classification and object classification. Considering that there is domain gap between faces of different races, our RFW promotes the development of DA in FR.

3. Racial Faces in-the-Wild: RFW

3.1. Creating RFW

In this section, we describe the dataset collection process, which is summarized in Fig. 2.

Databases selection. Instead of downloading images from websites and cleaning the images carefully, we collect the images of different races from existing databases. There are three principles guiding us to select candidate databases: 1) the candidates should be clean enough, 2) identities' number of the candidates should be large enough, 3) as few candidates as possible are selected in order to avoid inter-noise. After comparing all public available large-scale training datasets (over 1M images), MS-Celeb-1M [6] is found to be the best matchers.

Race detection. Face++ API [4] is used to estimate race of each celebrities in MS-Celeb-1M. However, API can not correctly detect race for every image and different images of one person may be distinguished as different races. Hence, for each identity, it will be accepted only if almost images are estimated as the same race, otherwise it will be abandoned.

Data re-cleaning. Because the automatical methods have already been used to clean MS-Celeb-1M, we only consider to re-clean the testing set of RFW manually. We randomly select 3000 people from each race, and pay extreme cautious to remove outlier faces for each identity as well as outlier identities for each race manually.

Testing set construction. We construct our testing set similar to LFW. 10 disjoint splits of image pairs are defined, and each contains 300 positive pairs and 300 negative pairs. Moreover, we also find that the random selection of face pairs in LFW may make the task easy and be away from reality. It is natural to impose certain constraints on face pairs. We apply cosine similarity measure of the well-established Arcface descriptor[2]. For each identity A, we randomly select one positive pair $\{A_i, A_j\}$ from 50% of pairs with smaller cosine similarity to enlarge intra-difference; and we randomly select one negative pair $\{A_i, B_j\}$ from 1% of pairs with larger cosine similarity to decrease demographic biases. Considering that this strategy will select noisy images more easily, we carefully clean these face pairs again.

Training set construction. We further collect a training set to facilitate the future research on racial bias. For source domain, about 500K images of 10k Caucasian people are randomly selected, and the labels are given as well. For target domains, we select 50K images of Indians, Asians and Africans respectively, and their labels are unavailable.

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 3. The number of identities and images in RFW

3.2. Statistics and analyses of RFW

As we know, the performance of FR is influenced by many factors, such as pose, age and gender. In order to

[2] https://github.com/deepinsight/insightface

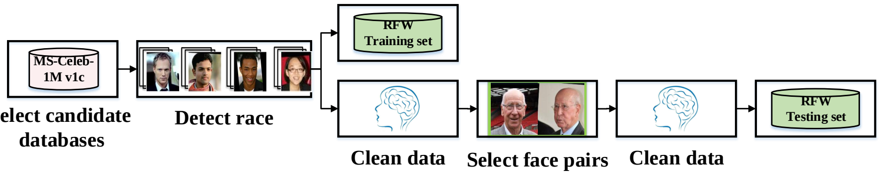
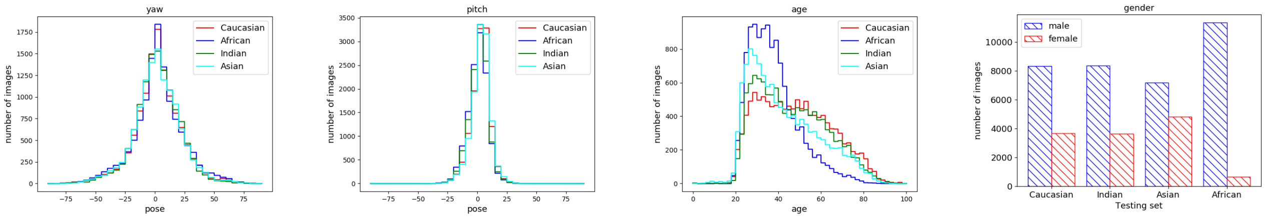


Figure 2. Overview of construction of RFW.



(a) yaw (b) pitch (c) age (d) gender

Figure 3. RFW statistics. We show the pose (yaw and pitch), age and gender distribution of four testing subsets.

ensure that our testing set can be used to fairly evaluate the recognition ability of the algorithm on different races, we compare these attributes, i.e. pose, age and gender, between four testing subsets to exclude other factors except for race which can cause difference. Face++ API is used to estimate pose, age and gender for each image and the distributions are given in Fig. 3. According to the figures, there are no large differences in pose, age and gender distribution between Caucasian, Indian and Asian testing sets, thus the recognition performance is only affected by different races. African set has a smaller age gap and the least balanced gender distribution which only contains 668 female images. Considering that recognition systems perform better on male with small age variation [10], the images of our African set are easier to be recognized compared with those of other sets. That is to say that recognition accuracy of African people may be even lower in reality.

Also, pose and age gap distribution of positive pairs in LFW and RFW are estimated, as illustrated in Fig. 5. Compared with LFW, the pose and age variations in our RFW are much larger. Further, the examples of difficult pairs selected by cosine-distance in RFW are presented in Fig. 4. The positive pairs in RFW contain obvious pose and age differences and the negative pairs are confusing even for human observers with careful inspection. This confirms the effectiveness of our constraints on selecting face pairs.

Then, based on our RFW, we examine four commercial recognition APIs, i.e. Face++, Baidu, Amazon, Microsoft to observe whether or not racial bias exists. Face verification accuracies are shown in Table 1 and ROC curves are presented in Fig. 9. First, we can find that FR accuracies are far from saturated when tested on our Caucasian testing subset. Microsoft achieves 98.22% on LFW and the accuracy drops about 10% on RFW. Second, FR systems indeed work unequally well for different races, the maximum difference in error rate between the best and worst groups is 12%. If we regard the mean of the four commercial recognition APIs as measurement, existing FR systems yield 90.27%, 86.28%, 86.82% and 81.86% average accuracies on Caucasian, Indian, Asian and African testing subset, respectively. Third, an interesting phenomenon is found from our experiments: APIs which are developed by East Asian companies perform better on Asian subjects than do APIs developed in the Western hemisphere.

4. Deep information maximization adaptation network

To reduce the racial bias, there is a strong incentive for transferring recognition knowledge from Caucasians (source) to other races (target) instead of training special models for each race. This is a typical unsupervised DA problem. Ben-David et al. [8] suggested that the expected DA loss for a target domain is bounded by three terms: 1) expected loss for the source domain; 2) domain divergence between source and target; and 3) the sum of the loss on the source and target domain. We call it three DA principles in our paper. Due to the absence of labeled target samples, most deep DA methods for object classification, such as MMD, only minimize the first two terms. Considering a larger number of identities of target domain in FR, only optimizing the two terms is not enough. Some methods [33, 50] try to minimize the third term utilizing pseudo target labels generated by maximum posterior probability of source classifier. However, pseudo target labels can not be obtained directly using source classifier in FR because there are no share classes between source and target domain. Inspired by [49, 19], we propose a deep Information maxi-

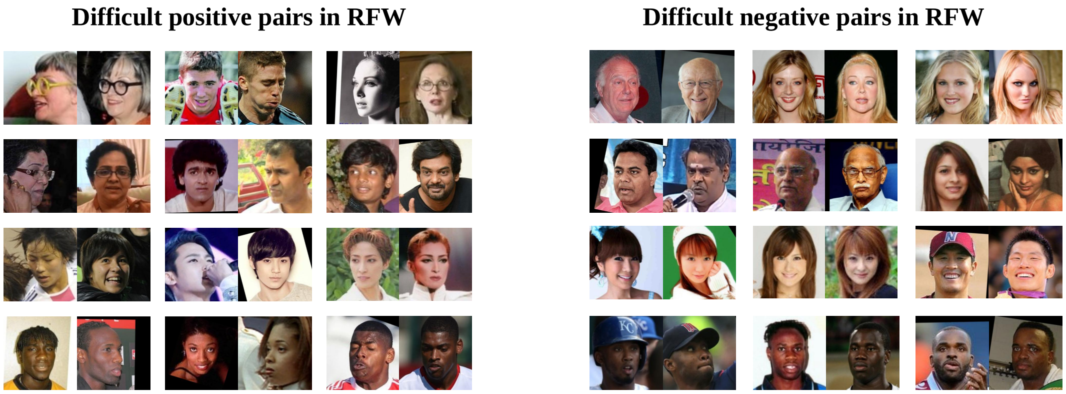
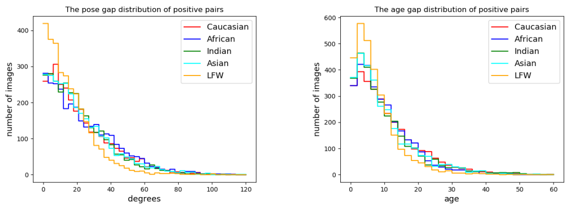

Difficult positive pairs in RFW
Difficult negative pairs in RFW

Figure 4. Examples of difficult positive and negative pairs in RFW dataset, which challenge the recognizer by the pose, age, expression and make-up variations of same people and the similar appearance of different people.



(a) pose gap

(b) age gap

Figure 5. Compared to the positive pairs in LFW, the pose and age differences of positive pairs in RFW are larger. This shows we successfully add variations to intra-class.

mization adaptation network (IMAN) which optimizes an mutual-information loss as an proxy to the expected error on the target domain. Combined with MMD, our IMAN simultaneously optimizes the three terms and narrows down the racial bias between domains effectively, as shown in Fig. 6.

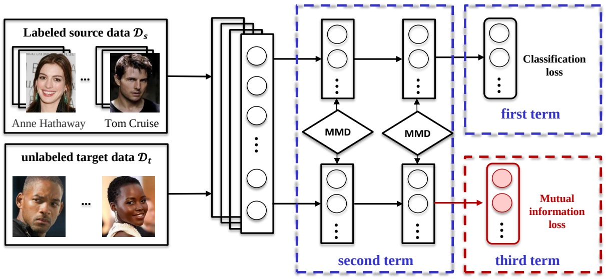


Figure 6. The IMAN architecture. The inputs of the upper network are source labeled images while the lower are target unlabeled data. Corresponding to three DA principles, classification loss supervises learning proceeds for source domain; MMD loss aims at minimizing the distribution discrepancy of two domains; our mutual-information loss aims to minimize the expected error and learn discriminative representations on the target domain.

In our case, source domain is a labeled training set with images of Caucasian subjects, namely $\mathcal{D}_s = \{x_i^s, y_i^s\}_{i=1}^M$, and target domain is an unlabeled training set with images of Asian, Indian or African subjects, namely $\mathcal{D}_t = \{x_i^t\}_{i=1}^N$.

4.1. Mutual-information loss

As the most widely used classification loss function, Softmax loss aims to maximize the conditional probability of the correct class by minimizing the cross-entropy for each training sample which can be presented as follows:

$$\mathcal{L}_{softmax} = \frac{1}{N}\sum_i \boldsymbol{y_i} log(\boldsymbol{p_i}) \qquad (1)$$

where $\boldsymbol{y_i}$ represents the ground truth labels (one-hot) of i-th samples. $\boldsymbol{p_i} = [p_{i1}, p_{i2}, ..., p_{ik}, ..., p_{ic}]$ means the probability assigned to the each class computed by $p_{ik} = \frac{e^{z_{ik}}}{\sum_{j=1}^{c} e^{z_{ij}}}$, z_{ik} is the k-th output of the network according to k-th class and c is the total number of classes. But when there are no labels for target data, how can we learn discriminative representations with the probability $\boldsymbol{p_i}$ only?

Based on the desideratum that an ideal posterior probability vector $\boldsymbol{p_i}$ should look like $[0, 0, ..., 1, ..., 0]$, we accordingly reduce amount of confusing labels by minimizing the entropy of $\boldsymbol{p_i}$. It actually coincides with the idea of Sphereface [27] and Arcface [14] which encourages large decision margin between classes. However, simply minimizing this entropy will cause that more decision boundaries are removed and most samples are assigned to the same class. In order to balance class separation and class balance, we simultaneously maximize the entropy of $\overline{\boldsymbol{p_0}}$ where $\overline{\boldsymbol{p_0}} = [\overline{p_1}, \overline{p_2}, ..., \overline{p_k}, ..., \overline{p_c}]$ is an estimate of the marginal distribution of $\boldsymbol{y}$ and is given by $\overline{p_k} = \frac{1}{N}\sum_i p_{ik}$. Therefore, our mutual-information loss can be presented by:

$$\mathcal{L}_M = \frac{1}{M}\sum_i H[\boldsymbol{p_i}] - \lambda H[\overline{\boldsymbol{p_0}}] = -I_t(\boldsymbol{X}; \widehat{\boldsymbol{y}}) \qquad (2)$$

where $H[\boldsymbol{p}]$ denotes the entropy of a probability vector $\boldsymbol{p}$, λ is the parameter for the trade-off between two entropies.

$I_t(\boldsymbol{X};\widehat{\boldsymbol{y}})$ means mutual-information between the empirical distribution on the inputs and the estimated label distribution $\widehat{\boldsymbol{y}}$ using $\boldsymbol{p_i}$. Therefore, our loss is equal to maximize this mutual-information.

Note that mutual-information loss is heavily dependent on initialization to generate a reliable probability vector $\boldsymbol{p_i}$ at first, we initialize the target classifier by pseudo target labels. These pseudo labels are obtained from clustering algorithm in Megaface [31] which is similar to spectral clustering. A graph is constructed where the nodes represent images and edges signify the two images have larger cosine-similarity. Then, each connected component with at least three nodes is saved as a cluster (identity). We take these pseudo labels to initialize our network for mutual-information loss.

4.2. Our adaptation network

In this paper, we embed the idea of mutual-information and MMD to deep network for learning transferable features. MMD has been widely adopted as a standard distribution distance metric to measure the discrepancy, it maps the extracted deep features to a reproducing kernel Hilbert space (RKHS) endowed by multiple kernels and compares the square distance between the empirical kernel mean embeddings. According to DAN [28], an empirical estimate of MMD is given as:

$$MMD^2(D_s,D_t)=\left\|\frac{1}{M}\sum_{i=1}^{M}\phi(\mathrm{x}_i^s)-\frac{1}{N}\sum_{j=1}^{N}\phi(\mathrm{x}_j^t)\right\|_H^2 \tag{3}$$

where ϕ represents the function that maps the original data to a RKHS. The kernel functions, which are associated with this mapping $k(x^s,x^t)=\langle\phi(x^s),\phi(x^t)\rangle$, is defined as the convex combination of m PSD kernels k_u, namely $\mathcal{K}=\{k=\sum_{u=1}^{m}\beta_u k_u:\sum_{u=1}^{m}\beta_u=1,\beta_u\geq 0,\forall u\}$, where β_u is the coefficients of u-th kernel.

Our IMAN pretrained with source data is 1) finetuned on labeled source samples, 2) optimized MMD loss to minimize the domain distribution discrepancy and learn domain-invariant representations, 3) finetuned by mutual-information loss to learn discriminative representations on the unlabeled target domain, simultaneously. The overall objective function for IMAN is given by:

$$\mathcal{L}=\mathcal{L}_C(X_s,y_s)+\alpha\sum_{l\in\mathcal{L}}MMD^2(D_s^l,D_t^l)+\beta\mathcal{L}_M(X_t) \tag{4}$$

where α and β are the parameters for the trade-off between three terms. $\mathcal{L}_M(X_t)$ is our mutual-information loss. If only unlabeled target samples are used for training, the network may learn more unreliable representations. Then, we should use source samples for training as well to ensure the accuracy. $\mathcal{L}_C(X_s,y_s)$ denotes 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 multi-kernel MMD between the source and target evaluated on the l-th layer representation. $\mathcal{L}_C(X_s,y_s)$, $\mathcal{L}_M(X_t)$ and $MMD^2(D_s^l,D_t^l)$ corresponds to the three DA principles, respectively.

Three important points that distinguish IMAN from relevant literature [49, 19] are: 1) **Deep DA.** We introduce mutual-information to deep DA. 2) **Initialize method.** To overcome the non-overlapping of source and target categories in FR, we propose to initialize target classifier using pseudo labels generated by clustering algorithms. 3) **Combining with MMD.** We combine mutual-information loss with MMD to further minimize the domain distribution discrepancy.

5. Experiments on RFW

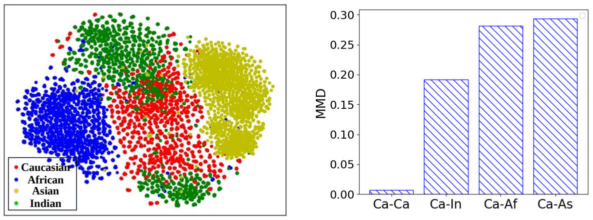
In this section, we conduct an exploratory experiments, namely ”other-race effect”, and then evaluate the proposed unsupervised DA method on our RFW dataset.

5.1. "Other-race effect" experiment

Psychological research indicates that humans recognize faces of their own race more accurately than faces of other races. The ”contact” hypothesis suggests that this ”other-race effect” occurs as a result of the greater experience we have with own-race versus other-race faces. Considering that Caucasian subjects are overwhelmingly dominant in numbers in training databases, so do deep FR algorithms inherit this ”other-race effect”? We conduct some experiments on four testing subsets of our RFW to go deep into this problem.

Existence of domain gap. We generate the average faces of four testing subsets and compare them by vision. As shown in Fig. 1, we can find that there indeed are certain discrepancies among different races in facial features and complexions, especially the African people.

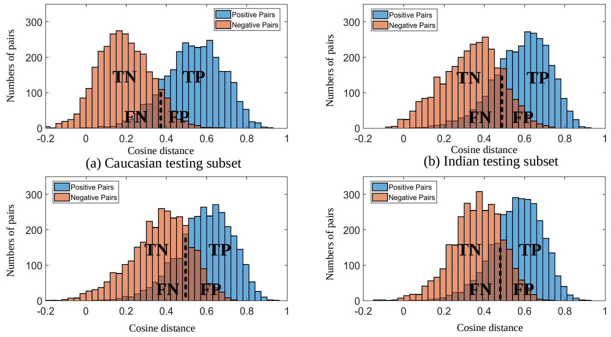
Then, the visualization and quantitative comparisons are conducted at feature level. To extract deep features, we train a deep model based on ResNet-34 by using CASIA-WebFace as the training data and Softmax as the loss function. Based on this model, the deep features of 3000 images of each testing subset are extracted and are visualized respectively using t-SNE embeddings [15], as shown in Fig. 7(a). The features almost completely separate according to race, but there is not a clear boundary between the features of Indians and Caucasians, which conforms our common sense that the faces of Indians are more westernized than the faces of Africans and Asians. Moreover, we use the MMD [9, 11] to compute distribution discrepancy between Caucasians and other races. The results are shown in Fig. 7(b), we make the same conclusions with the results of t-SNE: 1) the distribution discrepancies between Caucasians and



(a) T-SNE (b) MMD

Figure 7. (a) The feature space of four testing subsets. Each dot color represents a image belong to Caucasian, Indian, Asian or African. (b) The distribution discrepancy measured by MMD. 'Ca', 'As', 'In' and 'Af' represent Caucasian, Asian, Indian and African, respectively. 'Ca-As' represents the distribution discrepancy between Caucasian and Asian, and so on

other races are much larger than that between Caucasians themselves which means the existence of domain gap. 2) Africans and Asians have the larger domain discrepancies with Caucasians, followed by Indians.



(a) Caucasian testing subset (b) Indian testing subset (c) Asian testing subset (d) African testing subset

Figure 8. Distribution of cosine-distances of 6000 pairs on (a) Caucasian testing subset, (b) Indian testing subset, (c) Asian testing subset, and (d) African testing subset. In each subfigure, blue histogram presents distances distribution of 3000 positive pairs, and orange histogram presents that of 3000 negative pairs.

”Other-race effect”. After proving the existence of domain gap, some experiments are further conducted to investigate whether or not deep FR algorithms are susceptible to ”other-race effect”. Based on the features extracted by our trained Softmax model, we first compare the distribution of cosine-distances of 6000 pairs, as shown in Fig. 8. Considering that overlap between the histograms of positive and negative pairs means False Negative (FN) and False Positive (FP) in face verification, the degree of overlap in Caucasian set is much smaller than that in Indian, Asian and African set, which visually proves the recognition error of non-Caucasian subjects are much higher.

Then, we also examine some well-established methods, i.e. Center-loss [45], Sphereface [27], VGGFace2 [12] and ArcFace [14] on our RFW. We directly download the center-loss[3], Sphereface[4] and VGGFace2 (SeNet learned from scratch)[5] models from website. Note that ArcFace model[6] provided by author is trained on MS-Celeb-1M database which totally contains our RFW, we train an ArcFace model based on ResNet-34 with CASIA-Webface. The 10-fold cross-validation accuracies and ROC curves are given in Table 4 and Fig. 9. All the tested algorithms perform the best on Caucasian testing subset, followed by Indian, and the worst on Asian and African. For example, the accuracy of the ArcFace model on Caucasian testing subset reaches 92.15%, but its accuracy dramatically decreases to less than 83.98% on Asian subset. This is because that almost well-established models are predominantly trained on Caucasian faces. Therefore, deeply learned features tend to bias on distinguishing Caucasians rather than other races and the learned representations will discard information useful for discerning non-Caucasian faces. This phenomenon has always been concealed in the previous papers, since the number of non-Caucasian people for test is also quite small.

Model	LFW	RFW			
		Caucasian	Indian	Asian	African
Center-loss [45]	98.75	87.18	81.92	79.32	78.00
Sphereface [27]	99.27	90.80	87.02	82.95	82.28
Arcface[1][14]	99.40	92.15	88.00	83.98	84.93
VGGface2[2][12]	99.30	89.90	86.13	84.93	83.38
mean	99.18	90.01	85.77	82.80	82.15

[1] Different from the model provided by paper, Arcface here is a ResNet-34 model trained with CASIA-Webface.

[2] VGGFace2 here is a SeNet model trained with VGGFace2 database from scratch.

Table 4. Face Verification Accuracy (%) of some well-established models on RFW dataset

5.2. Domain adaptation experiment

To validate the proposed IMAN, we conduct experiments on our RFW dataset.

Implementation detail. For preprocessing, we share the uniform alignment methods as Arcface [14]. 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. The baseline model is ResNet-34 which is trained by using Caucasian training subset of RFW as the training data and Arcface as the loss function. The learning rate is started from 0.1 and decreased twice with a factor of 10 when errors plateau.

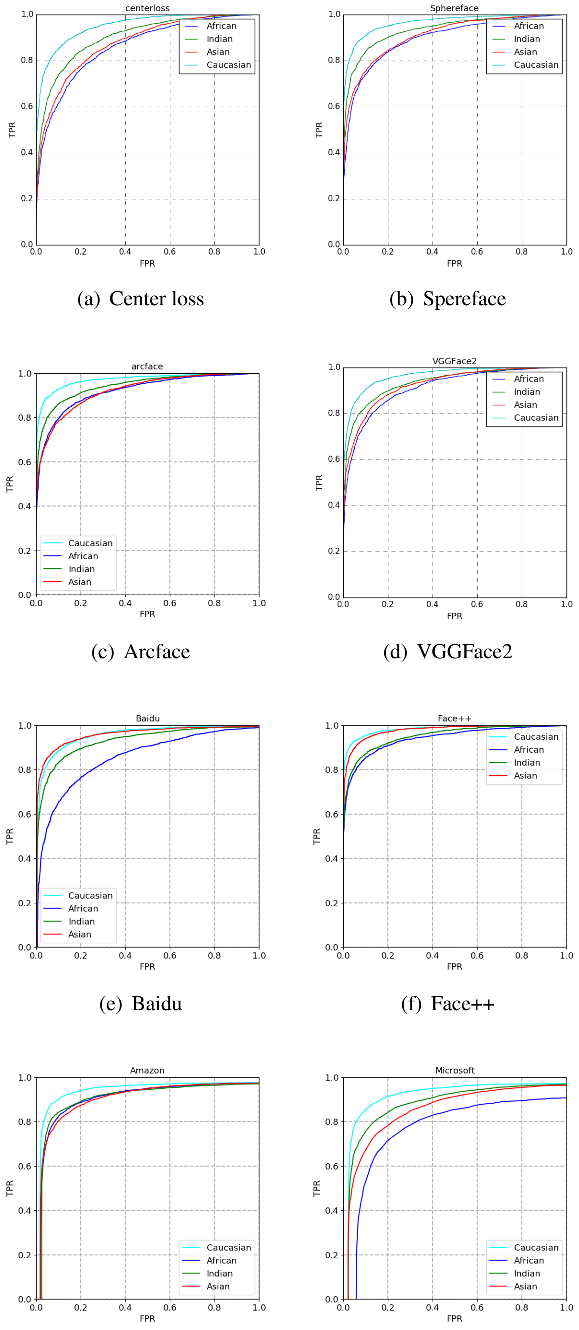
In IMAN, we train our method based on baseline network. We first initialize the target classifier with the pseudo-labeled target samples using learning rate of $5e-3$. Then

[3] https://github.com/walkoncross/caffe-face-centerloss

[4] https://github.com/wy1iu/sphereface

[5] https://github.com/ox-vgg/vgg_face2

[6] https://github.com/deepinsight/insightface



(a) Center loss

(b) Spereface

(c) Arcface

(d) VGGFace2

(e) Baidu

(f) Face++

(g) Amazon

(h) Microsoft

Figure 9. The ROC curves of (a) Center loss, (b) Spereface (c) Arcface, (d) VGGFace2, (e) Baidu, (f) Face++, (g) Amazon and (h) Microsoft on our four testing subsets.

using all labeled source and unlabeled target samples, we finetune the network by Equation (4) with learning rate of $1e-3$. We use Arcface as classification loss, the parameter α, β, λ are set to be 15, 5 and 0.2, respectively. For MMD, we follow the settings in DAN [28], and apply MMD to the last two fully-connected layers. In all experiments, we set the batch size, momentum, and weight decay as 200, 0.9 and $5e-4$, respectively.

Experimental result. Three DA tasks are performed, namely transferring knowledge from Caucasian to Indian, Asian and African. All learning algorithms are given about 500K labeled source examples and 50K unlabeled target examples. Then, we evaluate them on separate target testing subsets. Here are more details about the procedure used for each learning algorithms leading to the empirical results of Table 5:

Methods	Caucasian	Indian	Asian	African
Baseline	94.78	90.48	86.27	85.13
DDC [42]	-	91.63	87.55	86.28
DAN [28]	-	91.78	87.78	86.30
PL5 [31]	-	92.00	88.33	87.67
PL5+PL1	-	92.08	88.80	88.12
PL5+MMD	-	92.00	88.65	87.92
IMAN (ours)	-	93.55	89.87	88.88
IMAN* (ours)	-	**94.15**	**91.15**	**91.42**

Table 5. Verification Accuracy (%) on RFW dataset after DA

All these algorithms are based on baseline network, and finetune the network with source samples supervised by Arcface loss. The differences are,

• **DDC and DAN** simultaneously finetune the network by MMD. DDC applies MMD on one layer and DAN applies it on the last two layers.

• **PL5** initializes the target classifier and simultaneously finetunes the network by Softmax loss. The training data is pseudo-labeled target samples generated by Megaface' clustering algorithms [31]. And the learning rate is $5e-3$.

• **PL5+PL1** is based on PL5, it keeps on finetuning the network by Softmax loss with these pseudo-labels, but the learning rate is $1e-3$.

• **PL5+MMD** is based on PL5, it finetunes the network by MMD with learning rate of $1e-3$. The training data is all unlabeled target data.

• **IMAN** is based on PL5, and adds mutual-information loss to PL5+MMD. The learning rate is $1e-3$, training data is all unlabeled target data.

From Table 5, we can find that our IMAN dramatically outperforms all of the competing methods and achieves about 3% gains over baseline. Moreover, we have the following observations if we go deep into three DA principles described before. First, DDC [42] and DAN [28] only optimize the first two DA terms with help of MMD. But DAN is only superior to baseline by about 1%, DDC improves even less. This confirms our thought that only optimizing the two terms is not enough for FR. Second, PL5+PL1 is the method which finetunes the network by pseudo-labeled target samples, and it outperforms DAN and DDC. It shows that optimizing the third DA term by pseudo-labels is more effective than optimizing the second one. Finally, we compare the results among PL5+PL1, PL5+MMD and our IMAN. PL5+MMD is worse than PL5+PL1, but our IMAN outperforms PL5+PL1 by about 1% after adding mutual-information loss to PL5+MMD. This phenomenon proves the effectiveness of mutual-information loss. Why dose our IMAN using unlabeled data ourperform PL5+PL1 using pseudo-labeled data? The reason is that 1) the amount of pseudo-labeled data is much smaller than that of unlabeled data due to limitations of clustering algorithms, 2) we can not ensure the correctness of pseudo labels, 3) our mutual-information loss reduces the confusion possibilities and encourages large decision margin. Furthermore, we initialize the target classifier with Arcface loss instead of Softmax loss, our IMAN (denoted as IMAN*) is further improved, and obtains the best performances with 94.15%, 91.15% and 91.42% for Indian, Asian and African set.

6. Conclusion

An ultimate face recognition algorithm should perform perfectly and fairly on different demographic group. While the problem of racial bias is yet to be comprehensively studied, we have done the first step and create a benchmark for it. Our RFW database contains, 1) four testing subsets, namely Caucasian, Asian, Indian and African, to encourage FR algorithms to be fairly evaluated and compared on different races, 2) four training subsets to enable FR algorithms to transfer recognition knowledge from Caucasians to other races. Through experiments on our RFW, we first prove that there is domain gap among different races and the deep models trained on the current benchmarks do not perform well on non-Caucasian faces. Then, a deep information maximization adaptation network (IMAN) is introduced, it makes representations of source and the target similar and also learns the feature space discriminatively using unlabeled target data. 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] 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.

[8] S. Bendavid, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Z. Cao, M. Long, J. Wang, and M. I. Jordan. Partial transfer learning with selective adversarial networks. *arXiv preprint arXiv:1707.07901*, 2017.

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

[15] 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 *International conference on machine learning*, pages 647–655, 2014.

[16] 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.

[17] Y. Ganin. Unsupervised domain adaptation by backpropagation. In *International Conference on International Conference on Machine Learning*, pages 1180–1189, 2015.

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

[19] R. Gomes, A. Krause, and P. Perona. Discriminative clustering by regularized information maximization. In *International Conference on Neural Information Processing Systems*, pages 775–783, 2010.

[20] 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.

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

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

[23] 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.

[24] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, pages 4873–4882, 2016.

[25] 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.

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

[27] 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.

[28] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *International Conference on International Conference on Machine Learning*, pages 97–105, 2015.

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

[30] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016.

[31] A. Nech and I. Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *CVPR*, pages 3406–3415. IEEE, 2017.

[32] 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.

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

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

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

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

[37] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.

[38] 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.

[39] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014.

[40] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015.

[41] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4, 2017.

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

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

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

[45] 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.

[46] W.-S. T. WST. Deeply learned face representations are sparse, selective, and robust. *perception*, 31:411–438, 2008.

[47] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. *arXiv preprint arXiv:1705.00609*, 2017.

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

[49] S. Yuan and S. Fei. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *International Coference on International Conference on Machine Learning*, pages 1275–1282, 2012.

[50] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *The European Conference on Computer Vision (ECCV)*, September 2018.