Troubleshooting Ethnic Quality Bias with Curriculum Domain Adaptation for Face Image Quality Assessment

Fu-Zhao Ou[1] Baoliang Chen[1] Chongyi Li[2] Shiqi Wang[1] Sam Kwong[1,3]
[1]City University of Hong Kong, Hong Kong SAR, China
[2]Nankai University, Tianjin, China [3]Lingnan University, Hong Kong SAR, China
{fuzhao.ou, blchen6-c}@my.cityu.edu.hk lichongyi@nankai.edu.cn {shiqwang, cssamk}@cityu.edu.hk
https://github.com/oufuzhao/EQBM

Abstract

Face Image Quality Assessment (FIQA) lays the foundation for ensuring the stability and accuracy of face recognition systems. However, existing FIQA methods mainly formulate quality relationships within the training set to yield quality scores, ignoring the generalization problem caused by ethnic quality bias between the training and test sets. Domain adaptation presents a potential solution to mitigate the bias, but if FIQA is treated essentially as a regression task, it will be limited by the challenge of feature scaling in transfer learning. Additionally, how to guarantee source risk is also an issue due to the lack of ground-truth labels of the source domain for FIQA. This paper presents the first attempt in the field of FIQA to address these challenges with a novel Ethnic-Quality-Bias Mitigating (EQBM) framework. Specifically, to eliminate the restriction of scalar regression, we first compute the Likert-scale quality probability distributions as source domain annotations. Furthermore, we design an easy-to-hard training scheduler based on the inter-domain uncertainty and intra-domain quality margin as well as the ranking-based domain adversarial network to enhance the effectiveness of transfer learning and further reduce the source risk in domain adaptation. Extensive experiments demonstrate that the EQBM significantly mitigates the quality bias and improves the generalization capability of FIQA across races on different datasets.

1. Introduction

To maintain stable accuracy for face recognition systems in the wild deployment circumstance, numerous Face Image Quality Assessment (FIQA) techniques have been developed [38]. The existing FIQA schemes build their approaches on the assumption that the training and test data share similar distributions, and have shown promising outcomes on popular evaluation benchmarks [2, 21, 25, 33, 31, 41, 46, 52]. However, recently Babnik *et al*. [4] and Terhörst *et al*. [45] emphasize the potential problem of ethnic quality bias that leads to generalization challenges when FIQA models are deployed in real-world scenarios with different demographic distributions.

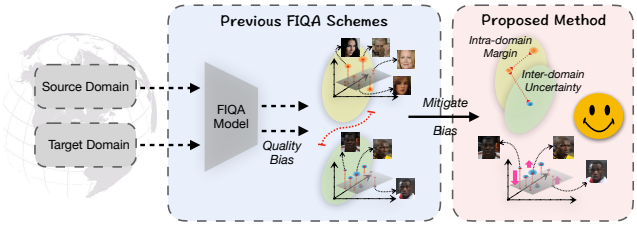

Figure 1. Illustration of the two different FIQA schemes. Previous FIQA schemes focus on formulating new quality relationships within the source domain, which may lead to ethnic quality bias due to the gap between the source and target domains. In contrast, in our proposed technique, the intra-domain margin and inter-domain uncertainty are involved in amending quality predictions of the target domain, which mitigates the ethnic quality bias.

On the one hand, the challenge stems from the unbalanced demographic attributes in datasets. For instance, mainstream FIQA training datasets, such as CASIA-Webface [53], MS-Celeb-1M [20], and VGGFace2 [34], are gathered from the Internet and contain people with unbalanced ethnic distributions. These datasets are overwhelmingly composed of people with Caucasian attributes, with over 74% of the samples with this demographic. Such a dramatic difference between the Caucasian (*i.e*. source domain) and other races (*i.e*., target domain) may lead to domain shifts in the deployment of FIQA models. Collecting and labeling more data along the underrepresented attributes for these million-scale training sets may be a solution but it is impractical. Moreover, the collection of face images would be involved in private legal issues across different districts. On the other hand, existing FIQA methods [2, 21, 25, 31, 46] focus on exploring new quality consistency relationships between quality scores and the per-

formance of face recognition models, largely escaping the research attention for amending quality scores against ethnic quality bias. Therefore, there appears to be a new challenging and meaningful problem: how to utilize unbalanced training data to reduce ethnic quality bias?

A potential solution is to adopt unsupervised domain adaptation to tackle the FIQA task via leveraging the source data to learn a robust quality regressor for the target domain [12, 14, 26]. Unfortunately, the majority of current domain adaptation schemes have not attained favorable advances solely for regression [14]. Moreover, how to guarantee the source risk of domain adaptation without ground truth labels in the FIQA task is also a challenge [28].

In this work, we build a novel Ethnic-Quality-Bias Mitigating (EQBM) framework based on the curriculum domain adaptation to address these issues. Herein, we describe the core idea of our framework in Fig. 1. Different from previous FIQA techniques [6, 33, 52] that focus on formulating new quality relationships within the source domain to yield quality scores, our EQBM framework explores the intra-domain quality margin and inter-domain uncertainty in the curriculum design to amend quality scores, showing significant advantages over previous works on evaluation benchmarks for mitigating the ethnic quality bias.

The main contributions of this work are threefold:

- We present the first attempt to address the challenging yet meaningful problem of ethnic quality bias via domain adaptation in the field of FIQA, which provides new insights into developing robust FIQA schemes.
- We devise a new framework to amend quality scores against the bias, in which we boost the potential of domain adaptation for migrating the ethnic quality bias via the Likert-scale quality probabilities and quality ranking.
- We propose a novel curriculum domain adaptation approach, a sequential adaptation strategy, which adapts from the easier samples to harder ones, in which the intra-domain quality margin and inter-domain uncertainty are coupled for the curriculum design.

2. Related Work

2.1. Face Image Quality Assessment

With the success of deep neural networks, there has been a surge of interest in developing deep learning-based approaches for FIQA. Existing methods can be broadly divided into two categories: quality regression-based [6, 7, 21, 52, 33] and recognition model-driven [2, 9, 31, 41, 46].

Quality Regression-based Technique. This scheme aims at calculating quality scores as annotations on a closed dataset of face recognition to train a FIQA regression model. For instance, Hernandez-Ortega *et al.* [21] computed the Euclidean distance of intra-class recognition embedding as the quality score for training and proposed FaceQnet to infer face image quality. Xie *et al.* [52] developed PCNet that relies on mated pairs to obtain annotations for FIQA network training. Best-Rowden and Jain [6] studied the relationship between face quality and annotations with partial or complete human efforts and evaluated the performance of trained FIQA regressors with these annotations. Ou *et al.* [33] proposed SDD-FIQA, in which the Wasserstein metric is utilized to generate quality annotations by computing the similarity distribution distance between positive and negative samples, and then a FaceQnet-like quality network is trained for quality prediction.

Recognition Model-driven Technique. This scheme proposes to learn or compute the embedding uncertainty as the quality score on face recognition models. In literature, Shi and Jain [41] proposed the Probabilistic Face Embedding (PFE) to introduce the uncertainty of samples to evaluate the quality. Subsequently, Chang *et al*. [9] improves the PFE via simultaneously learning the mean and uncertainty of Gaussian embedding distribution. Terhorst *et al*. [46] calculated the mean Euclidean distance of embeddings that are produced from a recognition model with the active dropout operator as the quality score. Meng *et al*. [31] proposed MagFace to reflect the quality of samples by adding an adaptive margin and regularization based on feature magnitude. Most currently, Žiga *et al*. [2] utilize the characteristics of face adversarial examples in the embedding space to estimate face quality by a quality aggregation function.

The aforementioned FIQA techniques focus on unveiling an improved association between quality scores and the performance of face recognition models. Nevertheless, there is rare research dedicated to resolving the generalization issue caused by quality bias in FIQA. To tackle this difficulty, instead of exploiting a new quality relationship, we concentrate on rectifying quality scores via the domain adaptation technique to diminish the FIQA quality bias for target domains, which is rarely touched in previous works.

2.2. Bias in Face Image Quality Assessment

In literature, face recognition bias has been demonstrated by exhibiting disparate performances of the recognition model for various demographic groups in terms of ethnicity and gender [19, 35, 50], which significantly affect the fairness of recognition models and thus promote numerous studies for mitigating the recognition bias in recent years.

For FIQA, Terhörst *et al*. [45] investigated the demographic bias of the previous FIQA methods by comparing the number of samples of each demographic attribute in a certain proportion of low-quality samples. More recently, Babnik *et al*. [4] conducted a more specific study to measure demographic biases of FIQA, including gender and ethnic-

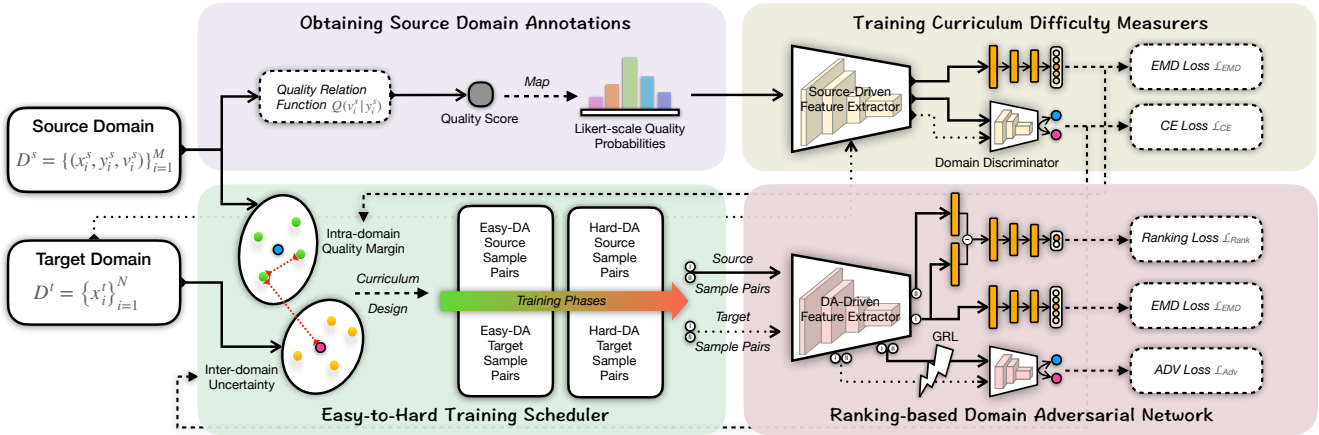

Figure 2. Overview of the proposed EQBM framework. Specifically, we first compute the quality score of the source domain by the FIQA quality relation function, and map it into the Likert-scale quality probabilities. Then, we train two curriculum difficulty measurers to obtain the mutual information between intra-domain quality margin and inter-domain uncertainty for building the curriculum-based easy-to-hard training scheduler. Finally, to amend the quality scores of the target domain, the ranking-based domain adversarial network is trained following the scheduler under the constraint of different losses and the Gradient Reversal Layer (GRL) for Domain Adaptation (DA).

ity, by comparing different FIQA evaluation results on each demographic group. Herein, this study revealed that the majority of the evaluated FIQA methods show a strong preference for Caucasian samples. Meanwhile, compared with ethnic bias, there was a less obvious trend in terms of gender bias for FIQA. Additionally, in the literature survey [38], for fairness and bias control, enhancing the robustness of FIQA or reducing bias on different demographic groups is expected to increase interest in future research.

These studies provide valuable prior guidance for developing FIQA methods to troubleshoot ethnic quality bias. Building upon these insights, we introduce the first endeavor to address FIQA bias by domain adaptation. Besides, we meticulously design the experimental settings based on the discoveries from these surveys to further explore the efficacy of FIQA on non-Caucasian groups under different deployed recognition models.

2.3. Domain Adaptation

Domain adaptation refers to the task of adapting a network from a labeled source domain to the target domain [17, 37, 43, 54]. Herein, the maximum mean discrepancy [27, 29, 48] and correlation alignment [10, 44, 49] are widely used to align marginal feature distributions of two domains. Meanwhile, another mainstream scheme is the adversarial training strategy [8, 13, 28, 47] that introduces a Gradient Reversal Layer (GRL) [18] to make the feature extractor independent of domain-specific information.

Furthermore, inspired by the curriculum learning strategy [5], several curriculum domain adaptation schemes have been developed in recent years. For instance, Yang *et al*. [56] proposed a curriculum-style domain adaptation approach with two stages of adaptation to infer numerous properties for the target domain. Chen *et al*. [12] developed a two-stage domain adaptation method to first split the target domain into confident and uncertain subdomains based on their prediction confidences, and then perform fine-level adaptation between the two subdomains. Shu *et al*. [42] proposed a curriculum learning-based strategy that leverages the loss of the network as weights to identify and eliminate unreliable source samples. PFAN [11] leveraged an easy-to-hard transfer strategy, which gradually selects target samples with higher cosine similarity to source prototypes on a per-category basis. In [36], curriculum graph co-teaching was proposed, which employs a dual classifier head, one of which is a graph convolutional network that aggregates features from similar samples across the domains.

These methods are not designed for FIQA tasks and thus do not take domain adaptation under the instability of source domain labels into account. Apart from previous domain adaptation methods, our method specially designs the domain adaptation for FIQA, which considers the mutual prior information between inter-domain uncertainty and intra-domain quality margin for curriculum design and involves the ranking-based domain adversarial network.

3. Methodology

In this section, we first describe the preliminaries of the problem formation. Then, the motivation and the details of each core component in our framework are described. The overview of the proposed framework is illustrated in Fig. 2.

3.1. Preliminaries

Given the face image set $\mathcal{X}^s$, identity label set $\mathcal{Y}^s$, and recognition embedding set $\mathcal{V}^s$ from the source domain $\mathcal{D} = \{(x_i^s, y_i^s, v_i^s)\}_{i=1}^M \subset \mathcal{X}^s \times \mathcal{Y}^s \times \mathcal{V}^s$, where M is the

number of samples. Commonly, there exists a quality relation function Q to yield quality scores. We can obtain the quality score q_i^s as the annotation of the i-th sample from the source domain by $q_i^s = Q(v_i^s|y_i^s)$. Thus, the source domain can be recomposed as the labeled training set by $\mathcal{D}^s = \{(x_i^s, q_i^s)\}_{i=1}^M$. Besides, the target domain is an unlabeled training set, namely $\mathcal{D}^t = \{x_i^t\}_{i=1}^N$. Meanwhile, the data distributions of the source domain P and the target one Q are different (*i.e.*, $P \neq Q$).

Our goal is to learn a deep-quality feature extractor denoted by $G_f \in \mathbb{R}^N$, where the feature representations with N dimensions are expected to be invariant between domains and improve the performance of target samples. Unfortunately, due to the fact that q_i^s is a scalar, it is still a challenge to train a robust G_f adapted to the target domain under the regression optimization: $\mathbb{R}^N \mapsto \mathbb{R}$, due to feature scaling problem in regression revealed by the study [14]. Specifically, the study found that there is a heightened vulnerability for G_f to experience performance degradation during the process of transferable representation learning for domain adaptation in regression tasks since regression performance is less robust to feature scaling. To this end, inspired by the Likert-scale quality probabilities distribution [40], we introduce the soft-mapping function to map q_i^s into a marginal probability distribution $\hat{q}_i^s$ associated with K-level anchors $\mathcal{C} = \{c^m\}_{m=1}^K$ as the label of the source domain during training G_f, which is given by

$$\hat{q}_i^{s,m} = \frac{\exp\left(-\beta \| q_i^s - c^m \|\right)}{\sum_{i=1}^K \exp\left(-\beta \| q_i^s - c^i \|\right)}, \tag{1}$$

where $\hat{q}_i^{s,m} \in \hat{q}_i^s$, β is a scaling constant, and $K = 5$ denotes the number of quality levels representing the level of "bad", "poor", "fair", "good", and "excellent" accordingly. Thus, the source domain can be reformulated as $\mathcal{D}^s = \{(x_i^s, \hat{q}_i^s)\}_{i=1}^M$. For quality prediction, given a face image x, the final predicted quality score is calculated by

$$\widetilde{q}(x) = \sum_{m=1}^{K} \phi_q\left(m | G_f(x)\right) \times c^m, \tag{2}$$

where $\phi_q(m|\cdot)$ is the marginal probability of m to be estimated by a quality classifier ϕ_q.

3.2. Curriculum Design for Domain Adaptation

With the quality probabilities distribution computed by Eq. (1) as the source-domain annotations, we can train a robust G_f for the target domain via domain adaptation. Unfortunately, for common domain adaptation methods [27, 28, 29, 47], all samples of source and target domains are fed to the network in the training phase, which may constrain the effect of transfer learning since the intra-domain learning difficulty of G_f is different among different samples. In order to explore the full potential for domain adaptation, we design a curriculum-style learning scheme with an easy-to-hard training pattern, in which the learning difficulty is determined by considering both intra-domain and inter-domain metrics. Concretely, we first take advantage of domain uncertainty by calculating the distance between the sample from one domain and the other domain in the same latent space to measure the inter-domain difficulty. For the intra-domain difficulty measure, we leverage the quality margin between the sample and others within the same domain, in order to introduce the pair-wise ranking under a certain margin to assist G_f in reducing the negative effect of q_i^s with low confidence and ensure the quality balance of the samples from the two domains in a batch during the transfer learning of domain adaptation.

3.2.1 Training Curriculum Difficulty Measurers

To obtain the intra-domain and inter-domain measures, we train two curriculum difficulty measures, including a naive source-driven quality network and a domain discriminator. For the naive source-driven quality network, both feature extractor G_f and quality classifier ϕ_q are trained by minimizing the constrained loss

$$\mathcal{L}_{EMD}\left(G_f, \phi_q\right) = \mathbb{E}_{\left(x_i^s, \hat{q}_i^s\right) \sim P^b} \mathsf{EMD}\left(\phi_q\left(G_f\left(x_i^s\right)\right), \hat{q}_i^s\right), \tag{3}$$

where $(x_i^s, \hat{q}_i^s) \sim P^b$ represents sampling a batch of b examples from source domain, and $\mathsf{EMD}(\cdot,\cdot)$ denotes the earth mover's distance. Then, the domain discriminator ϕ_{d_0} is trained by cross-entropy loss $\mathsf{CE}(\cdot,\cdot)$, which is given by

$$\begin{aligned} \mathcal{L}_{CE}\left(G_f, \phi_{d_0}\right) = & \mathbb{E}_{x_i^s \sim P^b} \mathsf{CE}\left(\phi_{d_0}\left(G_f\left(x_i^s\right)\right), 0\right) \\ & + \mathbb{E}_{x_i^t \sim Q^b} \mathsf{CE}\left(\phi_{d_0}\left(G_f\left(x_i^t\right)\right), 1\right), \end{aligned} \tag{4}$$

in which, 0 and 1 represent the domain labels of the source and target domains, respectively.

3.2.2 Easy-to-Hard Training Scheduler

To split the data into the easy-domain-adaptation subset and the hard one, we adopt the domain discriminator trained by Eq. (4) to compute the inter-domain uncertainty of each sample as the difficulty measure in terms of the inter-domain. The inter-domain uncertainty can be measured by the distance between the feature representation $F(\cdot|\phi_{d_0})$ yielded from ϕ_{d_0} and the learned class-specific center of the other domain. Specifically, the distances of the source and target domain samples are separately calculated by

$$\mu_i^s = \left|\left| F\left(x_i^s | \phi_{d_0}\right) - c_{|\phi_{d_0}}^1 \right|\right|_2^2, \tag{5}$$

and

$$\mu_i^t = \left|\left| F\left(x_i^t | \phi_{d_0}\right) - c_{|\phi_{d_0}}^0 \right|\right|_2^2, \tag{6}$$

where $c_{|\phi_{d_0}}^1$ and $c_{|\phi_{d_0}}^0$ indicate the learned class-specific center of the target domain and source domain of ϕ_{d_0}, re-

spectively. As a result, the data of the source domain is divided into the source easy-domain-adaptation training subset $\mathcal{D}^s_{easy} = \mathcal{D}^s_{|\leqslant\mathbb{E}[\mu^s_i]} \in \mathcal{D}^s$ and the source hard-domain-adaptation training subset $\mathcal{D}^s_{hard} = \mathcal{D}^s_{|>\mathbb{E}[\mu^s_i]} \in \mathcal{D}^s$, in which $\mathbb{E}[\mu^s_i]$ denotes the average of inter-domain uncertainties for the source domain. Similarly, the target domain is also divided into two training subsets, $\mathcal{D}^t_{easy}$ and $\mathcal{D}^t_{hard}$.

Furthermore, to compute the intra-domain quality margin, we utilize the quality difference $d(x_i, x_j)$ between the sample x_i and the other one x_j within the same domain via the source quality annotations or predictions by Eq. (2). The intra-domain quality margins for the source domain and the target domain are able to be respectively computed by

$$d^s\left(x^s_i, x^s_j\right) = |\delta_s(q^s_i) - \delta_s(q^s_j)|, \tag{7}$$

and

$$d^t\left(x^t_i, x^t_j\right) = |\delta_t\left(\widetilde{q}(x^t_i)\right) - \delta_t\left(\widetilde{q}(x^t_j)\right)|, \tag{8}$$

where $\delta(\cdot)$ represents the operator of the Z-score standardization in order to fix the variance σ of quality scores to one. To this end, for the easy training scheduler, we mine the sample pair $(x_i,\ x_j)$ under conditions subjected to $d(x_i, x_j) > 2\sigma$, $[x_i \in \mathcal{D}^s_{easy} \wedge x_j \in \mathcal{D}^s_{easy}]$ or $[x_i \in \mathcal{D}^t_{easy} \wedge x_j \in \mathcal{D}^t_{easy}]$. A similar operation is performed in the hard training scheduler under $d(x_i, x_j) > \sigma$.

3.3. Ranking-based Domain Adversarial Network

With the above easy-to-hard training scheduler, we can train a domain adversarial network to mitigate the quality bias for the target domain. Herein, guaranteeing the source risk during transfer learning is crucial for domain adaptation [57]. In mainstream adversarial domain adaptation pipelines [8, 28, 47], G_f is supervised by a classification loss in the source domain, which is based on the assumption that the source annotations are genuine ground-truth. However, genuine ground-truth quality labels are unavailable in FIQA. Therefore, to further reduce the negative effect caused by generated quality annotations q^s_i with low confidence during network training, we introduce a pairwise learning-to-rank based adversarial domain adaptation approach. Specifically, for the input sample pair (x^s_i, x^s_j), we adopt a binary decision strategy as our ranking loss to guide the network to learn which sample is of better quality, which is computed by

$$\begin{aligned}\mathcal{L}_{Rank}(G_f, \phi_r) = \mathbb{E}_{(x^s_i, x^s_j)\sim P^b} - [\eta_{i,j} \cdot \log(p^r\left(x^s_i, x^s_j\right)) \\ + (1 - \eta_{i,j}) \cdot \log(1 - p^r\left(x^s_i, x^s_j\right))],\end{aligned} \tag{9}$$

where

$$p^r\left(x^s_i, x^s_j\right) = \phi_r\left(G_f(x^s_i) - G_f(x^s_j)\right), \tag{10}$$

and

$$\eta_{i,j} = \begin{cases} 1, & \text{if } q^s_i > q^s_j \\ 0, & \text{otherwise} \end{cases}, \tag{11}$$

in which ϕ_r and $\eta_{i,j}$ denote the ranking clssifier and label respectively, and $G_f(x^s_i) - G_f(x^s_j)$ means the differences of the two corrponding features yielded by G_f. Meanwhile, according to Eq. (2), the ultimate goal of the network is to output a predicted quality score of the input sample. Thus, we also employ EMD loss to train the network by Eq. (3).

To mitigate quality bias for the target domain, the scheme of adversarial domain adaptation is employed to reduce the domain gap between the source and target domains. Herein, the GRL is utilized to train the domain discriminator ϕ_{d_1} across the source and target domains by the minimax game training strategy. We adopt the adversarial loss to achieve the optimization of ϕ_{d_1}, which is given by

$$\begin{aligned}\mathcal{L}_{Adv}(G_f, \phi_{d_1}) = -\mathbb{E}_{(x^s_i, x^s_j)\sim P^b}[\log\left(\phi_{d_1}\left(G_f\left(x^s_i\right)\right)\right) + \log \\ \left(\phi_{d_1}(G_f(x^s_j))\right)] - \mathbb{E}_{(x^t_i, x^t_j)\sim Q^b}[\log\left(1 - \phi_{d_1}(G_f(x^t_i))\right) \\ + \log\left(1 - \phi_{d_1}(G_f(x^t_j))\right)].\end{aligned} \tag{12}$$

To this end, our ranking-based domain adversarial network is trained by the minimax game

$$\min_{\phi_q,\, \phi_r} \max_{\phi_{d_1}} \mathcal{L}_{Rank} + \mathcal{L}_{EMD} - \lambda\mathcal{L}_{Adv}, \tag{13}$$

where λ is a hyper-parameter to trade off the two optimization objectives.

4. Experiments

4.1. Experimental Settings

Datasets. In accordance with [50], our bias study employs the BUPT-Transferface dataset for training and the RFW dataset for testing. BUPT-Transferface was proposed for the purpose of analyzing the ethnic bias in domain adaptation [51], and it contains 470K labeled images of 10K Caucasians as well as 50K unlabeled images of other races. The labeled images of Caucasians are utilized as the source domain, while the unlabeled images of non-Caucasians are used as the target domain. The RFW dataset consists of faces from four ethnic groups, namely Caucasian, Indian, Asian, and African. Each group contains 10K images of 3K identities suffering from various real-world quality degradation. All samples are cropped and resized to 112×112 via MTCNN [55], and they are normalized by subtracting the channel-wise mean value and dividing the variation.

Implementation. We train our framework on the BUPT-Transferface dataset using the MobilefaceNet [22] pre-trained on the MS-Celeb-1M [20] as the backbone of feature extractor based on its light-weight and high efficiency. Moreover, the architecture of the ϕ_q, ϕ_{d_0}, ϕ_{d_1}, and ϕ_r can be denoted by FC(256)-BN-FC(256)-BN-FC(n)-Softmax,

where FC and BN respectively means a fully connected layer and batch normalization layer, and n is the node number. Adam with $5e^{-4}$ weight decay is used as the optimizer. The training initial learning rates of all networks are set to $1e^{-4}$ that is divided twice by a factor of 10 when errors plateau with a batch size of 128. To normalize quality scores to the range of [0, 1], the q_i^s in Eq. (1) is normalized by min-max normalization, $\mathcal{C}$ = {0.1, 0.3, 0.5, 0.7, 0.9}, and β is set to 64. We train all networks with PyTorch on a machine equipped with a single NVIDIA GeForce RTX 3090 Ti GPU. In addition, the proposed framework is also implemented by MindSpore [15].

Evaluations. We compare the proposed EQBM with numerous state-of-the-art FIQA approaches, including FaceQnet [21], PFE [41], PCNet [52], SER-FIQ [46], SDD-FIQA [33], MagFace [31], and FaceQAN [2]. All competitors are trained on the MS-Celeb-1M dataset following their publicly available official implementations or code obtained directly from the authors. It is worth mentioning that our training set (BUPT-Transferface) is a subset of MS-Celeb-1M and its demographic distribution is approximately similar to MS-Celeb-1M. Moreover, our experiments are conducted on the cross-model setting for deployed face recognition models [2, 33].

Performance Criteria. In accordance with previous FIQA works [2, 21, 31, 33, 38, 46], the Error Versus Reject Characteristics (EVRC) curve is employed to measure the performance of FIQA approaches. The EVRC curve shows the False Non-Match Rate (FNMR) under different Ratios of Unconsidered Images (RUI) at a specific False Match Rate (FMR). As suggested in [2, 32, 33, 38], we report the Area Over Curve (AOC) results for quantitative comparisons among different EVRC curves [32, 33, 38]. The AOC is computed using the formula $\text{AOC} = 1 - \int_a^b g(\varphi)d\varphi$, where $g(\varphi)$ represents the FNMR at the RUI φ, and the lower and upper bounds of RUI, a and b, are fixed at 0 and 0.95, respectively. Furthermore, according to recommendations provided by [1, 39, 3], the partial Area Under Curve (pAUC) results are also reported in the part of our experiments. The pAUC aims at evaluating the FIQA performance at a lower discard rate, as it better aligns with the practical application scenario of FIQA.

4.2. Experimental Results

4.2.1 Results on Different Ethnic Groups

We first conduct evaluation experiments on different groups to examine the effectiveness of the proposed EQBM framework and compare it with the state of the arts on the RFW dataset. As done in [50], Caucasian is used as the source domain and other races as target domains to train our networks. In order to investigate the effect of quality bias on FIQA for different ethnic groups under different deployed face recognition models, three experimental groups are considered: 1) FIQA performance is evaluated under the deployed recognition model with bias, 2) FIQA methods are tested under the fair deployed recognition model, and 3) We compare the FIQA methods under the quality-aware deployed recognition model.

Table 1. AOC results under the FaceNet on the RFW dataset.

Type	FMR	Competitors					Ours	
		SER-FIQ	PCNet	MagFace	SDD-FIQA	FaceQAN	EQBM-M	EQBM-S
African	$1e^{-2}$	0.095	0.280	0.225	0.226	0.289	0.312	**0.330**
	$1e^{-3}$	0.083	0.191	0.186	0.172	0.206	0.225	**0.244**
	$1e^{-4}$	0.066	0.133	0.127	0.102	0.140	0.159	**0.168**
	EER	0.097	0.228	0.167	0.187	0.242	0.262	**0.280**
Indian	$1e^{-2}$	0.167	0.349	0.270	0.271	0.305	0.325	**0.397**
	$1e^{-3}$	0.104	0.265	0.247	0.240	0.264	0.274	**0.331**
	$1e^{-4}$	0.107	0.226	0.187	0.186	0.186	0.208	**0.245**
	EER	0.135	0.261	0.188	0.214	0.222	0.253	**0.314**
Asian	$1e^{-2}$	0.182	0.261	0.269	0.253	0.304	0.337	**0.353**
	$1e^{-3}$	0.102	0.184	0.225	0.192	0.246	0.245	**0.265**
	$1e^{-4}$	0.043	0.129	0.222	0.131	0.206	0.138	**0.169**
	EER	0.159	0.223	0.214	0.218	0.255	0.306	**0.311**

Table 2. AOC results under the ArcFace on the RFW dataset.

Type	FMR	Competitors					Ours	
		SER-FIQ	PCNet	MagFace	SDD-FIQA	FaceQAN	EQBM-M	EQBM-S
African	$1e^{-2}$	0.080	0.241	0.170	0.199	0.267	0.287	**0.294**
	$1e^{-3}$	0.081	0.215	0.174	0.165	0.226	0.241	**0.258**
	$1e^{-4}$	0.062	0.147	0.150	0.124	0.181	0.197	**0.211**
	EER	0.069	0.175	0.118	0.152	0.208	0.217	**0.229**
Indian	$1e^{-2}$	0.178	0.340	0.202	0.247	0.274	0.291	**0.366**
	$1e^{-3}$	0.102	0.272	0.187	0.209	0.240	0.247	**0.311**
	$1e^{-4}$	0.097	0.199	0.160	0.159	0.205	0.203	**0.242**
	EER	0.105	0.228	0.128	0.181	0.168	0.192	**0.252**
Asian	$1e^{-2}$	0.150	0.229	0.209	0.180	0.235	0.274	**0.289**
	$1e^{-3}$	0.097	0.189	0.157	0.118	0.157	0.189	**0.206**
	$1e^{-4}$	0.019	0.152	0.150	0.123	0.160	0.179	**0.191**
	EER	0.120	0.145	0.111	0.142	0.151	0.208	**0.217**

Table 3. AOC results under the GAC on the RFW dataset.

Type	FMR	Competitors					Ours	
		SER-FIQ	PCNet	MagFace	SDD-FIQA	FaceQAN	EQBM-M	EQBM-S
African	$1e^{-2}$	0.154	0.497	0.226	0.321	0.452	**0.479**	0.474
	$1e^{-3}$	0.244	0.420	0.237	0.305	0.419	0.442	**0.452**
	$1e^{-4}$	0.260	0.388	0.240	0.271	0.370	0.400	**0.403**
	EER	0.084	0.323	0.152	0.199	0.290	0.316	**0.323**
Indian	$1e^{-2}$	0.179	0.471	0.366	0.455	0.451	0.473	**0.529**
	$1e^{-3}$	0.173	0.404	0.288	0.382	0.385	0.401	**0.488**
	$1e^{-4}$	0.070	0.345	0.278	0.321	0.325	0.357	**0.431**
	EER	0.154	0.348	0.265	0.330	0.310	0.340	**0.399**
Asian	$1e^{-2}$	0.241	0.383	0.395	0.372	0.497	0.512	**0.536**
	$1e^{-3}$	0.286	0.372	0.378	0.293	0.425	0.414	**0.434**
	$1e^{-4}$	0.292	0.275	0.369	0.225	0.317	0.357	**0.368**
	EER	0.149	0.238	0.257	0.240	0.344	0.357	**0.373**

Evaluation under the Face Recognition Model with Bias. We use two mainstream open-source face recognition models as the deployed recognition models of FIQA, including FaceNet [34] and ArcFace [16] trained on CASIA-Webface [53] and VGGFace2 [34] datasets respectively, to test FIQA performance under the deployed recognition models with demographic bias. The EVRC curves of different races are shown in Fig. 3. Herein, Fig. 3 (a)-(c) and Fig. 3 (d)-(f) show the EVRC curves under the FaceNet and ArcFace, respectively. We can see that EQBM performs better than all competitors in most cases on different ethnic groups. Besides, for Indian and Asian groups, the gaps between EQBM-S and the best competitor (*i.e.* FaceQAN)

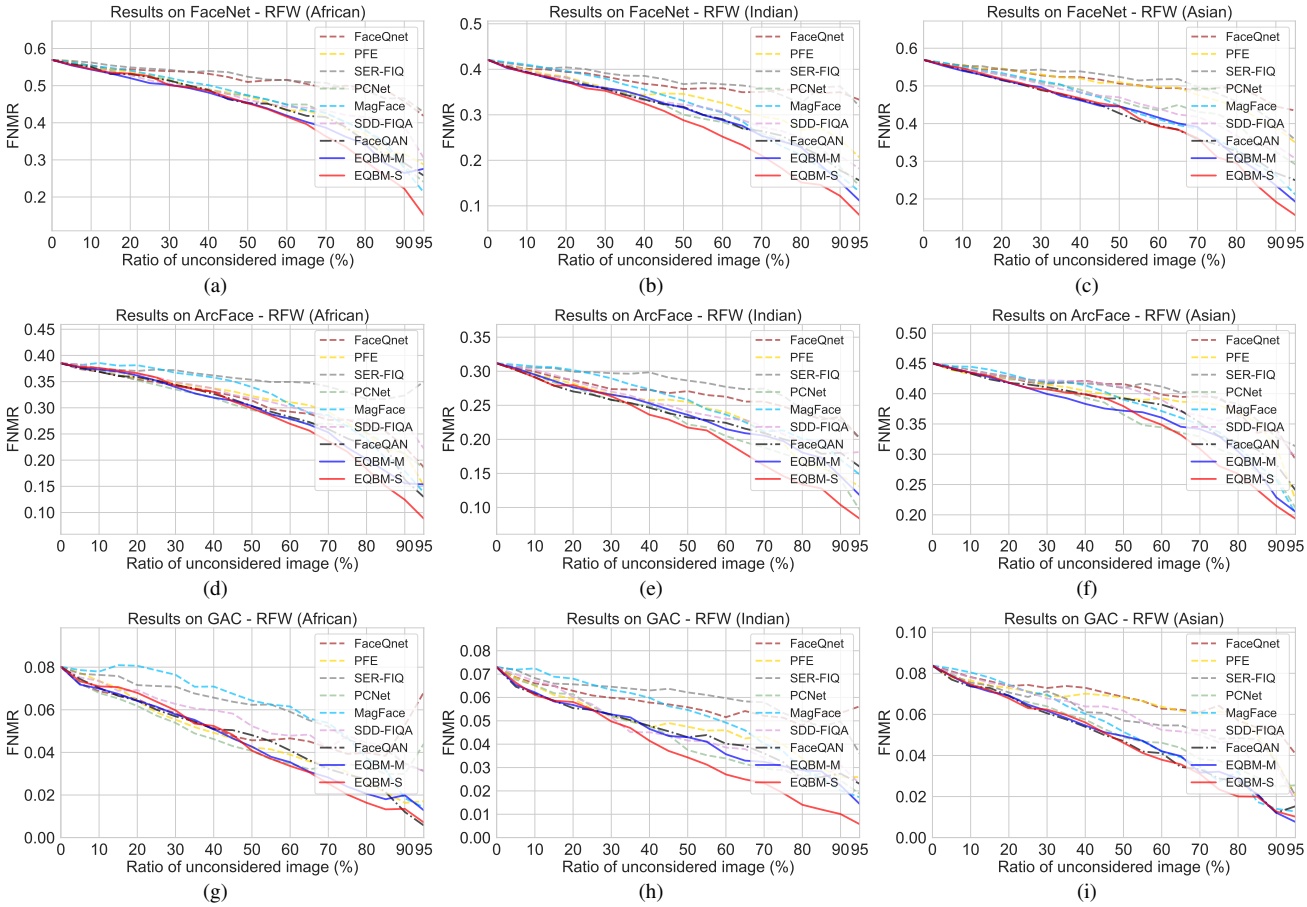

Figure 3. EVRC results at FMR=$1e^{-3}$ on the RFW dataset. Our framework (EQBM) is compared with seven FIQA methods under different deployed face recognition models "EQBM-S" represents the proposed method using the quality relation of SDD-FIQA to yield source quality annotations, while "-M" represents the one using MagFace.

are enlarged with the increased ratio of unconsidered images. Additionally, the AOC results at different FMRs are reported in Table 1 and Table 2 for quantitative analysis. For the AOC results under the FaceNet in Table 1, EQBM-S achieves over 13.2% on the African group, 41.4% on the Indian group, and 21.9% on the Asian group higher accuracy than the best competitor at FMR=EER. For the AOC results under the ArcFace in Table 2, EQBM-M and EQBM-S both surpass the FaceQAN by a non-trivial margin in most cases on different groups.

Evaluation under the Fair Face Recognition Model. To investigate the FIQA performance under the fair face recognition model, GAC [19] is adopted as the deployed recognition model of FIQA. The EVRC plots are shown in Fig. 3(g)-(i), and the AOC results are reported in Table 3. Intuitively, we can see that from Fig. 3(h), EQBM-S performs better obviously than other FIQA methods when RUI is greater than 30% on the Indian group. Meanwhile, we found that compared with results tested under the recognition model with bias, the curve of most FIQA methods declines faster under the fair recognition model, which suggests that they perform better in this circumstance. The results in Table 3 show that our EQBM-M achieves over 5.9% on the African group, and the EQBM-S outperforms all competitors on the other two groups.

Evaluaton under the Quality-aware Recognition Model. Here, we also employ AdaFace [24] recognition models trained on CASIA-Webface [53] (C) and WebFace4M [58] (W) datasets further to evaluate the FIQA performance under state-of-the-art quality-aware recognition models. The EVRC curves are presented in Fig. 5, where Fig. 5(a)-(c) and Fig. 5(d)-(f) plot for AdaFace (W) and AdaFace (C) models, respectively. Additionally, the corresponding pAUC (discard rate=0.3) results are presented in Table 4. Herin, the lower the pAUC value, the better the performance of the FIQA method. It is evident that EQBM-S outperforms all the other comparison methods.

The above analysis demonstrates that our framework is able to effectively mitigate the quality bias across different ethnic groups under all sorts of recognition models.

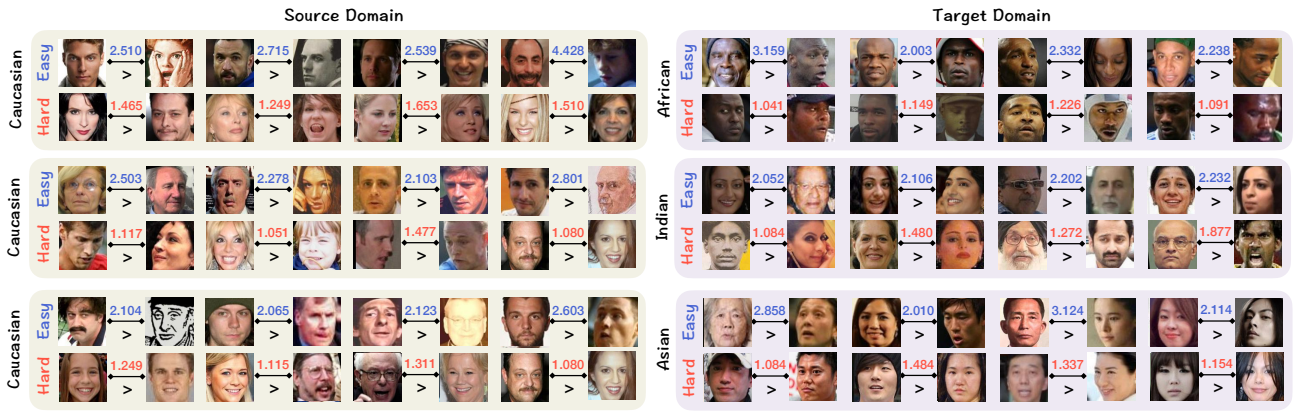

Figure 4. Illustration of the easy (blue) and hard (red) sample pairs and the corresponding intra-domain quality margin in our curriculum design. The quality margin is normalized by the Z-score standardization. This figure is best viewed in color.

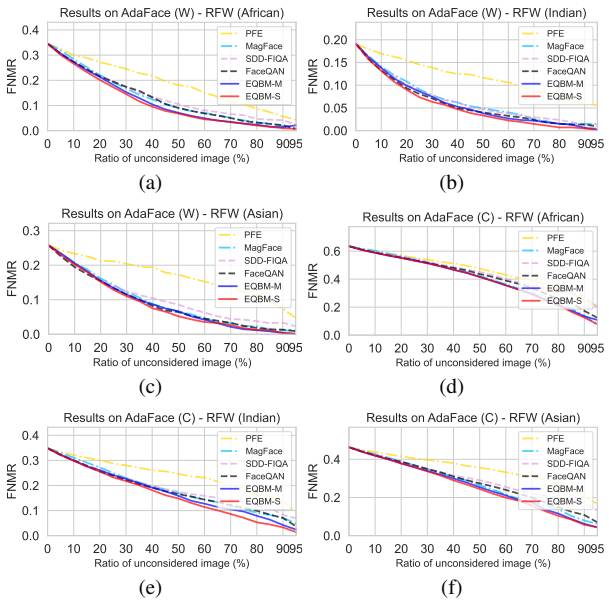

Figure 5. EVRC results at FMR=$1e^{-3}$ on the RFW dataset under the quality-aware recognition models.

4.2.2 Quality Analysis

Visulization of Easy and Hard Sample Pairs. To analyze the easy and hard samples in our curriculum-style training scheduler, we first show some sample pairs and their intra-domain quality margin in Fig. 4. Based on the illustration, we observe that: 1) intuitively, easy sample pairs are easier to distinguish the quality difference between face images than hard sample pairs, and 2) for the target domain, the pair-wise ranking correctness of sample pairs can also be ensured. This demonstrates that the intra-domain quality margin and the ranking-based domain adversarial network in our framework are able to reduce the source risk in domain adaptation.

Quality Visualization. We show some examples of face images and their corresponding quality scores before and after adaptation on different ethnic groups in Fig. 6. As the visual quality of the same line of faces improves from left to right, we can see that the rank of quality scores after adaptation is more reasonable.

Table 4. pAUC results at FMR=$1e^{-3}$ on the RFW dataset.

Method	AdaFace (W)				AdaFace (C)			
	African	Indian	Asian	$\widetilde{pAUC}(\downarrow)$	African	Indian	Asian	$\widetilde{pAUC}(\downarrow)$
PFE [41]	0.842	0.857	0.880	0.860	0.921	0.897	0.925	0.914
MagFace [31]	0.740	0.673	0.730	0.714	0.917	0.840	0.883	0.880
SDD-FIQA [33]	0.730	0.655	0.716	0.700	0.913	0.819	0.879	0.870
FaceQAN [2]	0.726	0.624	0.698	0.683	0.904	0.809	0.877	0.863
EQBM-M (Ours)	0.714	0.647	0.716	0.692	0.902	0.813	0.869	0.861
EQBM-S (Ours)	**0.693**	**0.608**	**0.703**	**0.668**	**0.900**	**0.802**	**0.865**	**0.856**

† $\widetilde{pAUC}$ denotes the average pAUC at a discard rate of 0.3 across three ethnic groups.

Quality Distribution. The quality distributions of different ethnic groups are shown in Fig 7. After the domain adaptation, the quality distribution of the three non-Caucasian ethnic groups dramatically changes, and the overall distribution tends to expand to the right. At the same time, we note that for the African group, the proportion of low-quality samples is significantly higher than the other two ethnic groups before adaptation, and then the quality distribution is similar to them after adaptation. This also provides evidence to some extent that our method is able to reduce the ethnic quality bias in FIQA effectively.

4.2.3 Results on Standard Benchmark Datasets

While our EQBM mitigates bias, we also wonder whether it can perform well on standard benchmarks. Hence, we evaluate our method on two standard benchmarks, including LFW [23] and IJB-C [30]. These two datasets exhibit imbalanced distribution in terms of ethnic demographics. For a fair comparison with other FIQA methods, we use the quality scores of different ethnic groups yielded by our EQBM to train a simple quality regression network following [21] for prediction. In Table 5, EQBM shows impressive performance on standard benchmark datasets.

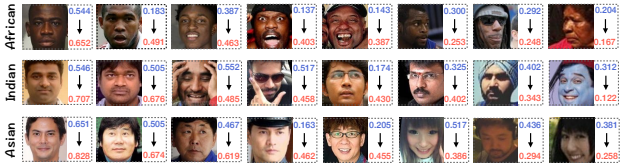

Figure 6. Illustration of face image examples and the corresponding quality scores before (blue) and after (red) adaptation. Zoom-in for a better view.

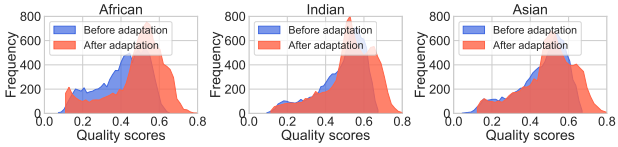

Figure 7. Illustrations of quality distributions before (blue) and after (red) adaptation on different ethnic groups.

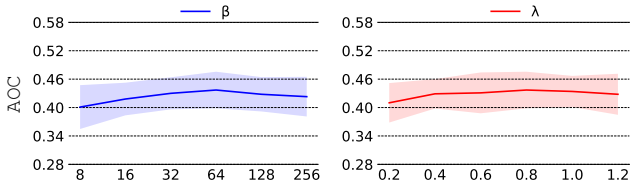

Figure 8. Parametric sensitivity of β (left) and λ (right).

4.2.4 Ablation Study

Parametric Sensitivity Analysis. In order to assess the sensitivity of the proposed EQBM to its hyper-parameters, we conduct experiments using the African group as the target domain and evaluate the AOC@FMR=$1e^{-3}$ with varying hyper-parameter values of β in Eq. (1) and λ in Eq. (13). Specifically, we systematically modified the target hyper-parameters while keeping the other ones fixed, and computed the AOC of the trained model. The results presented in Fig. 8 indicate that our method is generally robust to the changes in the hyper-parameters. We determined the optimal values of β and λ to be 64 and 0.8, respectively.

Contribution of each Key Component. To investigate the efficacy of the critical component of our proposed EQBM, we perform the ablation study on the RFW dataset using the fair deployed recognition model with and without (w/o) the implementation of the curriculum design and intra-domain quality ranking. The AOC results of this study are presented in Table 6, and we have the following observations: 1) without adaptation, SDD-FIQA shows impressive performance on the Caucasian, but is unable to attain satisfactory AOC results on other ethnic groups due to the existence of domain gaps, and 2) upon removing either the curriculum design or quality ranking, the FIQA model exhibits distinct performance degradation.

Comparison with other Domain Adaptation Methods. We also compare our proposed method with several established domain adaptation techniques, such as DAN [27], DANN [18], and RSD [14]. Herein, both DAN and DANN are mainstream unsupervised domain adaptation methods for classification, which require mapping the annotations from the source domain into the Likert-scale quality probabilities by Eq. (1). Meanwhile, to ensure a fair comparison, we maintained consistency in the feature extractor and training set with our setting, while adopting the remaining training parameters from their papers. As demonstrated in the upper part of Table 6, the EQBM surpasses these methods across various ethnic groups, suggesting the advantage of our domain adaptation for FIQA.

Table 5. AOC results at different FMR on LFW and IJB-C datasets.

Method	LFW				IJB-C			
	$1e^{-2}$	$1e^{-3}$	$1e^{-4}$	EER	$1e^{-2}$	$1e^{-3}$	$1e^{-4}$	EER
FaceQnet [21]	0.554	0.460	0.423	0.396	0.536	0.421	0.372	0.513
MagFace [31]	0.700	0.646	0.594	0.542	0.599	0.486	0.428	0.573
SDD-FIQA [33]	0.741	0.673	0.606	0.575	0.651	0.541	0.484	0.618
EQBM-M (Ours)	0.715	0.665	0.606	0.562	0.604	0.493	0.440	0.576
EQBM-S (Ours)	**0.750**	**0.703**	**0.627**	**0.576**	**0.657**	**0.546**	**0.490**	**0.632**

Table 6. Ablation study on the RFW dataset.

Method	Caucasian	African	Indian	Asian
SDD-FIQA [33]	0.493	0.301	0.374	0.287
DAN-S [27]	-	0.378	0.426	0.387
DANN-S [18]	-	0.393	0.445	0.406
RSD-S [14]	-	0.405	0.473	0.424
w/o Curriculum design	-	0.439	0.474	0.428
w/o Quality ranking	-	0.433	0.469	0.426
EQBM-S (Ours)	-	**0.452**	**0.488**	**0.434**

5. Conclusion

This paper has proposed a novel Ethnic-Quality-Bias Mitigation (EQBM) framework in Face Image Quality Assessment (FIQA) to address the generalization problem caused by quality bias in ethnic-based demographic distributions. Specifically, we first compute the Likert-scale quality probability distributions as quality annotations to remove the restriction of domain adaptation for regression. Meanwhile, to explore the full potential of mitigating FIQA quality bias in target domains, we design an easy-to-hard training scheduler based on the inter-domain uncertainty and intra-domain quality margin as well as the ranking-based domain adversarial network. Experimental results demonstrate that the EQBM significantly improves the generalization capability of FIQA methods under various ethnic-based demographic distributions.

Acknowledgment

This work is partially supported by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA), the Research Grant Council (RGC) of Hong Kong General Research Fund (GRF) under Grant 11203820 and Grant 11203220, and the National Natural Science Foundation of China under 62022002. We also gratefully acknowledge the support of MindSpore, CANN, and Ascend AI Processor used for this research.

References

[1] ISO/IEC DIS 29794-1 information technology, biometric sample quality, part 1: framework. 2022. 6

[2] Žiga Babnik, Peter Peer, and Vitomir Štruc. FaceQAN: Face image quality assessment through adversarial noise exploration. In *Proceedings of International Conference on Pattern Recognition (ICPR)*, pages 748–754, 2022. 1, 2, 6, 8

[3] Žiga Babnik, Peter Peer, and Vitomir Štruc. DifFIQA: Face image quality assessment using denoising diffusion probabilistic models. *arXiv preprint arXiv:2305.05768*, 2023. 6

[4] Ziga Babnik and Vitomir Štruc. Assessing bias in face image quality assessment. In *Proceedings of the 30th European Signal Processing Conference (EUSIPCO)*, pages 1037–1041, 2022. 1, 2

[5] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 41–48, 2009. 3

[6] Lacey Best-Rowden and Anil K Jain. Learning face image quality from human assessments. *IEEE Transactions on Information Forensics and Security*, 13(12):3064–3077, 2018. 2

[7] Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, and Naser Damer. CR-FIQA: face image quality assessment by learning sample relative classifiability. *arXiv preprint arXiv:2112.06592*, 2021. 2

[8] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Partial transfer learning with selective adversarial networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2724–2732, 2018. 3, 5

[9] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5710–5719, 2020. 2

[10] Chao Chen, Zhihong Chen, Boyuan Jiang, and Xinyu Jin. Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, pages 3296–3303, 2019. 3

[11] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 627–636, 2019. 3

[12] Pengfei Chen, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Unsupervised curriculum domain adaptation for no-reference video quality assessment. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 5178–5187, 2021. 2, 3

[13] Qingchao Chen, Yang Liu, Zhaowen Wang, Ian Wassell, and Kevin Chetty. Re-weighted adversarial adaptation network for unsupervised domain adaptation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7976–7985, 2018. 3

[14] Xinyang Chen, Sinan Wang, Jianmin Wang, and Mingsheng Long. Representation subspace distance for domain adaptation regression. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1749–1759, 2021. 2, 4, 9

[15] MindSpore Vision Contributors. MindSpore Computer Vision:mindspore computer vision toolbox and benchmark. https://github.com/mindspore-lab/mindcv/, 2022. 6

[16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 6

[17] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*, pages 877–894, 2021. 3

[18] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 3, 9

[19] Sixue Gong, Xiaoming Liu, and Anil K Jain. Mitigating face recognition bias via group adaptive classifier. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3414–3424, 2021. 2, 7

[20] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 87–102, 2016. 1, 5

[21] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In *Proceedings of International Conference on Biometrics (ICB)*, pages 1–8, 2019. 1, 2, 6, 8, 9

[22] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 5

[23] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2008. 8

[24] Minchul Kim, Anil K. Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18750–18759, 2022. 7

[25] Qiye Lian, Xiaohua Xie, Huicheng Zheng, and Yongdong Zhang. Variance of local contribution: an unsupervised image quality assessment for face recognition. In *Proceedings of International Conference on Pattern Recognition (ICPR)*, pages 4665–4670, 2022. 1

[26] Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X Yu, and Boqing Gong. Open compound domain adaptation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12406–12415, 2020. 2

[27] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 97–105, 2015. 3, 4, 9

[28] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018. 2, 3, 4, 5

[29] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 2208–2217, 2017. 3, 4

[30] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. Iarpa janus benchmark - c: Face dataset and protocol. In *Proceedings of International Conference on Biometrics (ICB)*, pages 158–165, 2018. 8

[31] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14225–14234, 2021. 1, 2, 6, 8, 9

[32] Martin Aastrup Olsen, Vladimír Šmida, and Christoph Busch. Finger image quality assessment features–definitions and evaluation. *IET Biometrics*, 5(2):47–64, 2016. 6

[33] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. SDD-FIQA: Unsupervised face image quality assessment with similarity distribution distance. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7670–7679, 2021. 1, 2, 6, 8, 9

[34] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. *British Machine Vision Association*, 2015. 1, 6

[35] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–10, 2020. 2

[36] Subhankar Roy, Evgeny Krivosheev, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Curriculum graph co-teaching for multi-target domain adaptation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5351–5360, 2021. 3

[37] Paolo Russo, Fabio M Carlucci, Tatiana Tommasi, and Barbara Caputo. From source to target and back: symmetric bi-directional adaptive gan. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8099–8108, 2018. 3

[38] Torsten Schlett, Christian Rathgeb, Olaf Henniger, Javier Galbally, Julian Fierrez, and Christoph Busch. Face image quality assessment: A literature survey. *ACM Computing Surveys (CSUR)*, 2022. 1, 3, 6

[39] Torsten Schlett, Christian Rathgeb, Juan Tapia, and Christoph Busch. Considerations on the evaluation of biometric quality assessment algorithms. *arXiv preprint arXiv:2303.13294*, 2023. 6

[40] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006. 4

[41] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 6902–6911, 2019. 1, 2, 6, 8

[42] Yang Shu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Transferable curriculum for weakly-supervised domain adaptation. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, pages 4951–4958, 2019. 3

[43] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. *Domain Adaptation in Computer Vision Applications*, pages 153–171, 2017. 3

[44] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Proceedings of European Conference on Computer Vision Workshops (ECCVW)*, pages 443–450. Springer, 2016. 3

[45] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Face quality estimation and its correlation to demographic and non-demographic bias in face recognition. In *Proceedings of IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11, 2020. 1, 2

[46] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5651–5660, 2020. 1, 2, 6

[47] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7167–7176, 2017. 3, 4, 5

[48] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 3

[49] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5018–5027, 2017. 3

[50] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 692–702, 2019. 2, 5, 6

[51] Mei Wang, Yaobin Zhang, and Weihong Deng. Meta balanced network for fair face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8433–8448, 2021. 5

[52] W. Xie, J. Byrne, and A. Zisserman. Inducing predictive uncertainty estimation for face recognition. *arXiv preprint arXiv:2009.00603*, 2020. 1, 2, 6

[53] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 1, 6, 7

[54] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2720–2729, 2019. 3

[55] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 5

[56] Yang Zhang, Philip David, Hassan Foroosh, and Boqing Gong. A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):1823–1841, 2019. 3

[57] Li Zhong, Zhen Fang, Feng Liu, Jie Lu, Bo Yuan, and Guangquan Zhang. How does the combined risk affect the performance of unsupervised domain adaptation approaches? In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, pages 11079–11087, 2021. 5

[58] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10492–10502, 2021. 7