Exploiting Transferable Knowledge for Fairness-aware Image Classification

Sunhee Hwang*, Sungho Park*, Pilhyeon Lee*, Seogkyu Jeon,
Dohyung Kim, and Hyeran Byun†

Department of Computer Science, Yonsei University, Seoul, Republic of Korea
{sunny16, qkrtjdgh18, lph1114, jone9312, dohkim02, hrbyun}@yonsei.ac.kr

Abstract. Recent studies have revealed the importance of fairness in machine learning and computer vision systems, in accordance with the concerns about the unintended social discrimination produced by the systems. In this work, we aim to tackle the fairness-aware image classification problem, whose goal is to classify a target attribute (*e.g.*, attractiveness) in a fair manner regarding protected attributes (*e.g.*, gender, age, race). To this end, existing methods mainly rely on protected attribute labels for training, which are costly and sometimes unavailable for real-world scenarios. To alleviate the restriction and enlarge the scalability of fair models, we introduce a new framework where a fair classification model can be trained on datasets without protected attribute labels (*i.e.*, target datasets) by exploiting knowledge from pre-built benchmarks (*i.e.*, source datasets). Specifically, when training a target attribute encoder, we encourage its representations to be independent of the features from the pre-trained encoder on a source dataset. Moreover, we design a Group-wise Fair loss to minimize the gap in error rates between different protected attribute groups. To the best of our knowledge, this work is the first attempt to train the fairness-aware image classification model on a target dataset without protected attribute annotations. To verify the effectiveness of our approach, we conduct experiments on CelebA and UTK datasets with two settings: the conventional and the transfer settings. In the both settings, our model shows the fairest results when compared to the existing methods.

1 Introduction

Artificial Intelligence (AI) systems have been widely used for decision making such as visual recognition [1], criminal justice [2], or employment [3]. Although AI systems are proved to be effective, they have raised concerns due to their biased results against some human characteristics such as gender, age, or race, which are referred to as protected attributes. As pointed out in the literature [4–6], AI models are highly dependent on training datasets, thus they tend to learn unfair thoughts or biases from the datasets and produce discriminatory outputs.

*Equal contributions
†Corresponding author

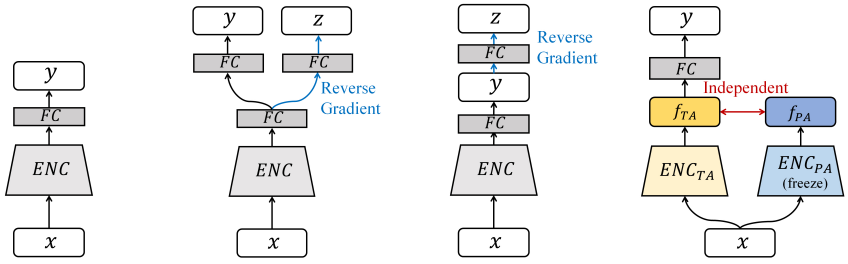


Fig. 1. Illustrations of four different classification models: (a) conventional classification model, (b) Protected Attribute Adversarial Learning (PAAL) [5, 15] with domain adversarial training of neural network [16], (c) Adversarial De-biasing (AdvDe) [14], and (d) Ours. x and y denote an input image and its target attribute label respectively, while z represents its protected attribute label. Unlike the previous fairness-aware methods, our model is trained without z by leveraging a pre-trained protected attribute encoder on another dataset.

For example, image captioning models may generate biased captions against gender, *i.e.*, while captions for images with women are biased towards shopping or cooking, those for images with men are inclined towards driving or shooting [7]. In addition, face recognition systems usually perform well on Caucasians, yet they often fail to identify faces of other races [8, 9]. To prevent socially negative impacts, researchers have paid their attention to developing fair AI models that produce unbiased results with regard to protected attributes [10–14].

To build fair models, previous methods mainly focus on excluding information related to protected attributes from the feature representation by utilizing a domain adaptation technique [5, 15], adversarial de-biasing [14], or disentangled representation learning [17, 10]. Despite their impressive improvements in the perspective of fairness, they still have limitations in that protected attribute labels for new datasets are inevitably required when they are supposed to be deployed in a new circumstance. Acquiring additional annotations of protected attributes is time-consuming and may be even infeasible in real-world situations, which limits the scalability of the fair models.

Therefore, we set two goals in this paper: 1) predicting target attributes (*e.g.*, attractiveness) in a fair way with respect to protected attributes (*e.g.*, gender, age), namely fairness-aware image classification, and 2) transferring knowledge about protected attributes from a pre-built dataset (*i.e.*, source dataset) to another dataset without protected attribute labels (*i.e.*, target dataset).

To this end, we introduce a new framework with two encoders, where one encoder pre-trained on a source dataset provides guidance on protected attributes and the other encoder fairly predicts target attributes according to the guidance. In Fig. 1, we compare our method to the existing approaches with the simple dia-

gram illustrations. Given an input image x, the conventional classification model (Fig. 1a) is trained to predict the target label y without considering protected attributes. Upon the conventional model, the previous fairness-aware approaches (Fig. 1b and Fig. 1c) adopt gradient reversal layers to make their models unable to predict the protected attribute label z. On the other hand, our method (Fig. 1d) does not exploit the label z. Instead, it encourages the feature representation f_{TA} to be independent of the representation f_{PA} from the pre-trained protected attribute encoder. By doing so, our model can learn the fair representation with respect to the protected attribute without using z.

Technically, we introduce a Feature Independency Triplet loss to promote the independency of the features from two different encoders. In specific, we first select three samples in a target dataset and encode them with the pre-trained protected attribute encoder. Next, we choose one of them as an anchor and calculate the feature distances between the anchor and the other samples. The sample with a larger distance is set to a positive sample and the other to a negative one. Then, in the feature space of the target attribute encoder, we pull the positive sample to the anchor while pushing the negative sample away from the anchor. This encourages the independency between the representations from the target and protected attribute encoders. In addition, we propose a Group-wise Fair loss to minimize the difference in error rates between protected attribute groups. Firstly, we cluster the features from the protected attribute encoder into k groups. Then, we train the target attribute encoder to equalize the error rates of target attribute classification among the groups by minimizing their Wasserstein distance. These two proposed losses work complementarily to each other and enable the target attribute encoder to learn fair representation in terms of the protected attribute without explicit labels.

Our key contributions are summarized as follows:

- We propose a new framework for fairness-aware image classification, where fair representations can be learned without protected attribute labels by exploiting knowledge from external pre-built benchmarks.
- We design a Feature Independency Triplet loss to reduce the dependency between representations of two encoders for protected attributes and target attributes.
- To further improve fairness, we introduce a Group-wise Fair loss to minimize the gap between the error rates of different protected attribute groups.
- We compare our method with existing fairness-aware classification approaches on two most popular benchmarks: CelebA and UTK datasets. The results validate the effectiveness of our model, achieving the fairest results regarding *Equality of Opportunity* in two experimental settings, namely the conventional and the transfer settings. Notably, our method in the transfer setting shows fairer results than the existing methods in the conventional setting, which confirms the efficacy of our transfer learning approach.

2 Related Work

2.1 Fairness-aware Image Classification

Recently, fairness studies in computer vision mainly attempt to solve the societal bias problems on image classification task [5, 15, 18, 19, 13]. There are two mainstreams of fairness-aware approaches for mitigating biases related to protected attributes: pre-processing and in-processing.

Pre-processing methods aim at constructing a new de-biased dataset from an original biased dataset with respect to protected attributes. Quadrianto *et al.* [13] propose a data-to-data translation method that maps a biased dataset into a fair dataset. Meanwhile, Sattigeri *et al.* [19] introduce a method generating a de-biased dataset based on Generative Adversarial Networks (GANs) [20] with two fairness-aware constraints, *Demographic Parity* [21] and *Equality of Opportunity* [12].

In-processing approaches devise new algorithms to eliminate discriminatory factors in models. They mainly focus on learning invariant feature representations against protected attributes (*e.g.*, gender, age, race). For example, inspired by the domain adversarial training of neural network [16], Wang *et al.* [5] train an image classification model to misclassify the protected attribute label with a Gradient Reversal Layer (GRL) [16] but to correctly classify target labels at the same time. Similarly, Kim *et al.* [15] utilize an adversarial strategy [20, 22] and the gradient reversal technique [16] to eliminate unwanted biases by minimizing the mutual information between the feature embedding and the protected attribute. Besides, Adversarial De-biasing [14] is introduced to make the prediction for the target label which is not predictive for the protected attribute label.

Overall, prior methods achieve fair results in pre-processing and in-processing ways, but they still suffer from the essentially required cost for obtaining additional protected attribute labels for new datasets. On the contrary, our method leverages information from the pre-built source dataset to perform fair classification on the target dataset without protected attribute labels.

2.2 Transfer Learning in Computer Vision

Transfer learning is an important research topic that focuses on exploiting knowledge from a problem to tackle a different but related problem [23]. Since transfer learning allows utilizing pre-trained models for various tasks with time-saving, it draws much attention from researchers in the computer vision domain. For instance, transfer learning is actively investigated in image captioning [24], classification [25], generation [26], and object detection [27].

However, transfer learning has not been investigated for fairness-aware image classification. Herein, we are the pioneering work to obtain knowledge for protected attributes from the source dataset and transfer it into the target dataset without protected attribute labels to tackle fairness-aware image classification.

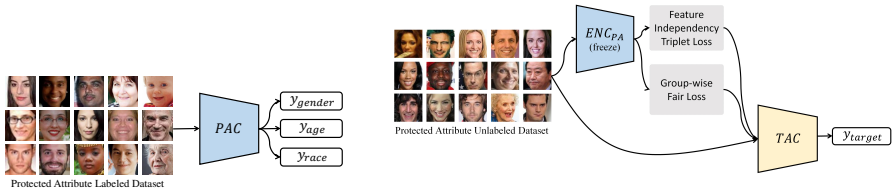


Fig. 2. Overview of the proposed framework for transfer learning. (a) In the first stage, we train the Protected Attribute Classifier (PAC) to predict three protected attributes on the source dataset in a multi-task way. (b) Then, we train the Target Attribute Classifier (TAC) on the target dataset to classify target attributes in a fair way regarding the protected attributes. To make the training without protected attribute labels feasible, we propose to utilize the encoder of the pre-trained PAC with Feature Independency Triplet Loss and Group-wise Fair Loss.

3 Fairness Definition

Fairness of the AI system can be defined as the ability to produce fair decisions with regard to protected attributes such as gender. The most widely used definition is *Equality of Opportunity* [12], which is based on the principal that individuals should be provided equal opportunities for desired results. Formally, all protected attribute groups should have the same true positive rates for the target attribute as follows:

$$\mathcal{P}(\hat{Y} = 1|p = 0, Y = 1) = \mathcal{P}(\hat{Y} = 1|p = 1, Y = 1), \tag{1}$$

where p, Y, $\hat{Y} \in \{0, 1\}$ denote the protected attribute, the target attribute, and the prediction respectively. In this work, we focus on improving fairness in terms of *Equality of Opportunity*.

4 Approach

Our main goal is to train a fair classification model on the target dataset without protected attribute labels. To this end, we devise a two-step strategy for transfer learning as illustrated in Fig. 2. In the first step, we train a Protected Attribute Classifier (PAC) using the source dataset with protected attribute labels (Fig. 2a). Then, we leverage the representation from the encoder of the PAC to transfer knowledge about protected attributes to a Target Attribute Classifier (TAC) (Fig. 2b). By utilizing the obtained knowledge in the first stage, the TAC is able to learn fair representations without explicit protected attribute labels. Specifically, to transfer knowledge of the PAC into the TAC, we introduce a Feature Independency Triplet loss and a Group-wise Fair loss, which will be detailed in this section.

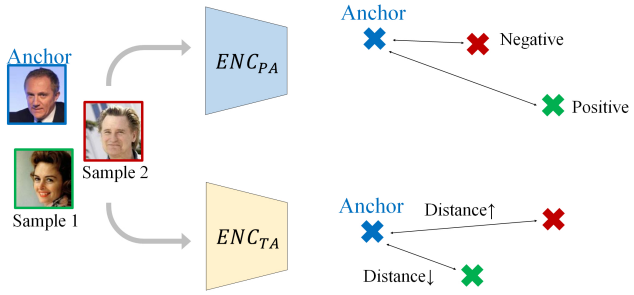


Fig. 3. Schematic visualization of the Feature Independency Triplet loss. In the feature space of the encoder of the protected attribute classifier (ENC_{PA}), we select three samples and choose one of them as an anchor. Then, we set the sample with a larger feature distance from the anchor to be a positive sample and the other to be a negative sample. Afterwards, in the space of the encoder of the target attribute classifier (ENC_{TA}), the Feature Independency Triplet loss minimizes the feature distance between the anchor and the positive sample, while maximizing the distance between the negative pair.

4.1 Protected Attribute Classifier

We first train the Protected Attribute Classifier (PAC) on the source dataset to encode representations with respect to multiple protected attributes. The PAC consists of a feature encoder with several convolutional layers (ENC_{PA}) and fully connected layers. Given a set of training data $X_s = \{x_1, x_2, ..., x_n\}$, gender labels $Y_g = \{g_1, g_2, ..., g_n\}$, age labels $Y_a = \{a_1, a_2, ..., a_n\}$, and race labels $Y_r = \{r_1, r_2, ..., r_n\}$, we optimize the PAC by minimizing three cross-entropy loss functions simultaneously:

$$\mathcal{L}_{PAC} = -\sum_{i=1}^{n} g_i log(\hat{g_i}) - \sum_{i=1}^{n} a_i log(\hat{a_i}) - \sum_{i=1}^{n} r_i log(\hat{r_i}), \tag{2}$$

where $\hat{g}$, $\hat{a}$, $\hat{r}$, and n denote the prediction of three different classifiers and the number of samples in the source dataset respectively.

4.2 Target Attribute Classifier

The Target Attribute Classifier (TAC) is composed of a convolutional feature encoder (ENC_{TA}) and fully connected layers. Given a set of training images $X_t = \{x_1, x_2, ..., x_m\}$ and corresponding labels $Y_t = \{t_1, t_2, ..., t_m\}$, we train the TAC with the following cross-entropy loss function:

$$\mathcal{L}_{target} = -\sum_{i=1}^{m} t_i log(\hat{t_i}), \tag{3}$$

where $\hat{t_i}$ and m denote the prediction of the target attribute classifier and the number of samples in the target dataset respectively.

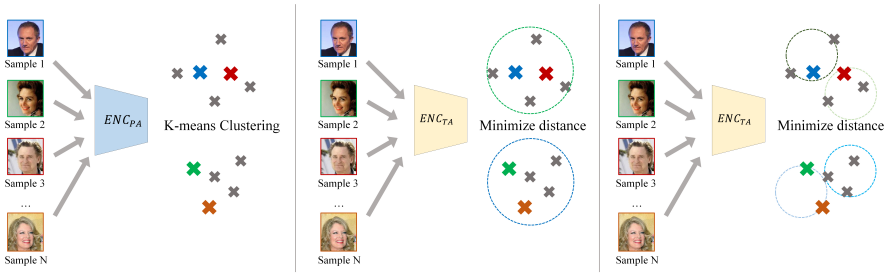


Fig. 4. The process of the Group-wise Fair loss. We group input images by k-means clustering based on their protected attribute features from the ENC_{PA}. Afterwards, during training the target attribute classifier, we aim to minimize the error rate discrepancy between different groups. The same process is performed in the subgroups as well.

4.3 Feature Independency Triplet loss

We propose a Feature Independency Triplet loss to encourage target attribute features from ENC_{TA} to be independent of the protected attributes. Fig. 3 shows the schematic visualization for the Feature Independency Triplet loss. Firstly, we randomly select two samples x_i and x_j for each anchor sample x_a in the mini-batch $X = [x_1, x_2, ..., x_k]$ of the target dataset, where k is the batch size. The anchor and the selected samples are encoded by the pre-trained ENC_{PA} into f_a,f_i, and f_j respectively. Based on the anchor feature f_a, we calculate the euclidean distance $d(f_a, f_i)$ and $d(f_a, f_j)$. We assign the sample which is more distant from x_a to a positive sample x_p and the other to a negative sample x_n. Thereafter, we construct k tuples $[(x_a^1, x_p^1, x_n^1), (x_a^2, x_p^2, x_n^2), ..., (x_a^k, x_p^k, x_n^k)]$, where the negative sample x_n^i is more similar to the anchor sample x_a^i in terms of the protected attributes. The Feature Independency Triplet loss is defined as:

$$\mathcal{L}_{triplet} = \sum_{i=1}^{N} \max(d(h_a^i, h_p^i) - d(h_a^i, h_n^i) + \alpha, 0), \tag{4}$$

where h_a^i, h_p^i, and h_n^i are encoded features of x_a^i, x_p^i, and x_n^i from ENC_{TA} respectively.

4.4 Group-wise Fair Loss

We introduce a Group-wise Fair loss to further reinforce the fairness of our model with respect to the protected attributes (See Fig. 4). Inspired by the prior work [28], we aim to minimize the discrepancy on misclassification rate between different protected attribute groups as follows:

$$minimize|P(\hat{y} \neq y|G_1) - P(\hat{y} \neq y|G_2)|, \tag{5}$$

where $\hat{y}$ and y denote the prediction and the target label respectively. G_1 and G_2 indicate the different protected attribute groups respectively.

However, since protected attribute labels are unavailable for the target dataset, we exploit the transferred knowledge of protected attributes from ENC_{PA} in order to separate the groups in terms of protected attributes. Specifically, we extract features $F = [f_1, f_2, ..., f_m]$ of input images from the target dataset $X_t = [x_1, x_2, ..., x_m]$ using the pre-trained ENC_{PA}. Then, we cluster X_t into two groups G_1 and G_2 with the k-means clustering algorithm based on F. Subsequently, to satisfy the equation (5), we minimize the Wasserstein distance between the two groups G_1 and G_2 in terms of the last fully connected features H_g from ENC_{TA}. The Wasserstein distance between two groups G_1 and G_2 are as follows:

$$\mathcal{W}(H_{G_1}, H_{G_2}) = \inf_{\gamma \in \prod(H_{G_1}, H_{G_2})} \mathbb{E}_{(z_1,z_2)\sim\gamma}[\|z_1 - z_2\|]. \tag{6}$$

where $\prod(H_{G_1}, H_{G_2})$ is the set of all joint distributions $\gamma(z_1, z_2)$ whose marginals are respectively H_{G_1} and H_{G_2}.

Although Group-wise Fair loss improves fairness between the protected attribute groups, the bias in terms of the target attribute still exists. Therefore, we propose a loss to minimize the Wasserstein distance between output features of subgroups as follows:

$$\begin{aligned} \mathcal{W}(H_{G_1^+}, H_{G_1^-}) &= \inf_{\gamma \in \prod(H_{G_1^+}, H_{G_1^-})} \mathbb{E}_{(z_1,z_2)\sim\gamma}[\|z_1 - z_2\|]. \\ \mathcal{W}(H_{G_2^+}, H_{G_2^-}) &= \inf_{\gamma \in \prod(H_{G_2^+}, H_{G_2^-})} \mathbb{E}_{(z_1,z_2)\sim\gamma}[\|z_1 - z_2\|], \end{aligned} \tag{7}$$

where G^+ and G^- respectively denote the group of positive samples and negative samples in terms of the target attribute respectively, while G_1 and G_2 are distinguished in terms of the protected attribute.

4.5 Full Objective Function for TAC

The final objective function for training the TAC is defined as:

$$\mathcal{L}_{TAC} = \lambda_1 \mathcal{L}_{target} + \lambda_2 \mathcal{L}_{triplet} + \lambda_3 \mathcal{L}_{group}, \tag{8}$$

where λ_* are the hyper-parameter for balancing the losses.

5 Experiment

We conduct experiments on two classification tasks: (1) attractiveness classification on CelebA dataset [29] and (2) race classification on UTK Face dataset [30].

Table 1. Dataset bias of the attractiveness attribute towards Male, Young, and Pale Skin attributes on CelebA dataset [29].

		Male			Young			Pale Skin		
		Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
TA=1	PA=1	19,014	2,651	1,914	78,239	9,404	9,116	5,021	595	610
	PA=0	64,589	7,681	7,984	48,549	5,428	5,998	78,582	9,737	9,288
TA=0	PA=1	49,247	5,807	5,801	5,364	928	782	1,984	261	230
	PA=0	29,920	3,728	4,263	30,618	4,107	4,066	77,183	9,274	9,834

For the quantitative evaluation in terms of fairness, we measure the *Equality of Opportunity* (*Eq.Opp.*) [12], which is defined as:

$$Eq.Opp. = |TPR_{p=0} - TPR_{p=1}|, \tag{9}$$

where TPR and p denote True Positive Rate (TPR) and a binary protected attribute label respectively.

5.1 Experimental Settings

- **Attractiveness Classification**: For the attractiveness classification task, we train the PAC on UTK Face dataset with three protected attributes, gender, age, and race. We set 19,708, 2,000, and 2,000 images of UTK dataset for training, validation, and test set, respectively. Then, we train the TAC on CelebA dataset to classify the attractiveness attribute. CelebA dataset is composed of train, validation, and test set with 162,770, 19,867, and 19,962 images respectively. The bias in the CelebA dataset is demonstrated in Table. 1. Since age and race attributes do not exist in CelebA dataset, we substitute them with young and pale skin attributes. We pre-process all the images by randomly cropping (178×178) and resizing to 64×64 in CelebA dataset.
- **Race Classification**: For the race attribute classification task (Caucasian or others), we train the PAC on CelebA dataset with only the gender attribute since it does not contain age and race attributes. Then, the TAC is trained to classify the race attribute on UTK Face dataset. In this experiment, we manually compose UTK Face dataset to be biased in terms of the protected attribute as follows: 4,000 Caucasian male, 1,000 other male, 1,000 Caucasian female, and 4,000 other female images for the training set, 1,000 images of each group for the test set, and others for the validation set. We use the cropped images of UTK Face dataset and resize it into 64×64.

5.2 Implementation Details

For the TAC and PAC, we use ResNet-18 [1] as our backbone network. We implement our networks in the Pytorch framework [31] and use the Adam optimizer [32] with $\beta_1 = 0.5$, $\beta_2 = 0.999$, learning rate $= 10^{-4}$, and the batch size

Table 2. Protected attribute classification results on UTK Face and CelebA dataset. Top three rows indicate the result of UTK Face dataset, and last row denotes the result of CelebA dataset.

Dataset	Attribute [Labels]	Validation	Test
UTK Face	Gender [Male, Female]	0.87	0.86
	Race [White, Black, Asian, Indian, Others]	0.75	0.73
	Age [0 9, 10 19, ..., 50+]	0.59	0.57
CelebA	Gender [Male, Female]	0.97	0.96

of 256. We early-stop training when the network converges in the validation set. All the networks are trained from scratch to prevent the model learning any unwanted biases from the datasets used for pre-training. For all experiments, we set λ_1, λ_2 and λ_3 as 1, 1e-3, and 1e-3, respectively. In the classification phase, we use the same threshold of 0.5 to fairly compare the methods.

5.3 Comparative Methods

To validate the effectiveness of our method, we compare it to the following approaches:

- **ResNet-18**: we adopt ResNet-18 [1] as the conventional classification model. The last fully connected layer is replaced with three fully connected layers with batch normalization and ReLU activation following [16].
- **Protected attribute adversarial learning**: we compare our model to a protected attribute adversarial learning approach (PAAL) [5, 15]. We adopt ResNet-18 as a backbone network and add two parallel branches on top of that for the domain adversarial training of neural network [16]: a classifier with three fully connected layers for the target label and Gradient Reversal Layers (GRL) composed of three fully connected layers for mitigating bias to the protected attribute.
- **Adversarial de-biasing**: We also compare our model with the Adversarial De-biasing (AdvDe) approach [14]. On top of the ResNet-18, we add one fully connected layer for adversarially training the model not to predict the protected attribute.

5.4 Protected Attribute Classification

Table. 2 shows the protected attribute classification accuracies on CelebA and UTK Face datasets. Our model achieves the top-1 classification accuracy of 86%, 73%, and 57% for gender, race, and age attributes on CelebA dataset respectively, and the accuracy of 96% for the gender attribute on UTK Face dataset.

Table 3. Equality of Opportunity (*Eq.Opp.*) for attractiveness classification on CelebA dataset [29] with respect to the two protected attributes: Young (age related) and Male. The lower is better. ResNet-18 is trained without any protected attribute labels. While Adversarial de-biasing and Protected Attribute Adversarial Learning models are trained with protected attribute labels of the target dataset, our method utilizes those of the source dataset.

	True Positive Rate		*Eq.Opp.*	True Positive Rate		*Eq.Opp.*
	Young=1	Young=0		Male=1	Male=0	
ResNet-18 [1]	86.21	65.60	20.61	63.85	89.55	25.70
AdvDe [14]	83.95	68.41	15.54	67.35	85.10	17.75
PAAL [15, 5]	91.87	80.31	11.56	81.35	94.24	12.89
Ours (All)	95.95	90.28	**5.61**	87.25	97.48	**10.23**

Table 4. Equality of Opportunity (*Eq.Opp.*) for attractiveness classification on CelebA dataset [29] regarding the skin color related protected attribute (Pale Skin).

	True Positive Rate		*Eq.Opp.*
	Pale Skin=1	Pale Skin=0	
ResNet-18 [1]	92.46	84.07	8.39
AdvDe [14]	91.80	83.93	7.85
PAAL [15, 5]	94.26	90.04	4.22
Ours	99.18	95.26	**3.92**

5.5 Attractiveness Classification

We conduct comparison on the attractiveness classification results as shown in Table. 3 and Table. 4. As described in Sec. 3, the objective of our model is to ensure *Equality of Opportunity* on different protected attributes such as gender, age, and race. In these experiments, the results demonstrate that our model achieves the fairest results on CelebA dataset with respect to the Young (5.61), Male (10.23), and Pale Skin (3.92) attributes.

In addition, we verify the contributions of our proposed loss functions through an ablation study as shown in Table. 5. We evaluate the results of models only with the Feature Independency Triplet loss or the Group-wise Fair loss, and the full model. Table. 5 shows that our loss function improves fairness step by step.

Furthermore, we validate that the improvement of our model is not caused by an additional usage of the source dataset through the experimental results in two different settings, as shown in Table. 6. In the first setting (*i.e.*, conventional setting), we conduct comparison with the comparative models by setting both the source and the target datasets to CelebA dataset (denoted with asterisk (*)). For this setting, we only consider gender as the protected attribute. In the second setting (*i.e.*, transfer setting), we compare fairness of all the models trained with protected attribute labels in UTK Face dataset and target attribute labels in CelebA dataset (denoted with dagger (†)). In both setting, our method shows the fairest results, verifying its effectiveness.

Table 5. Ablation Study on CelebA dataset for attractiveness classification. *Eq.Opp.* denotes Equality of Opportunity.

	True Positive Rate		*Eq.Opp.*	True Positive Rate		*Eq.Opp.*
	Young=1	Young=0		Male=1	Male=0	
Triplet Loss	93.66	83.63	10.03	80.36	95.87	15.51
Group-wise Fair Loss	93.90	84.27	9.63	82.08	95.79	13.71
All	95.95	90.28	5.61	87.25	97.48	10.23

Table 6. Equality of Opportunity (*Eq.Opp.*) for attractiveness classification on CelebA dataset [29] in terms of gender attribute. ResNet-18 (first row) is trained without any protected attribute. Asterisk (*) (2-5 rows) and dagger (†) (6-8 rows) denote the results of the conventional and the transfer settings, respectively.

	True Positive Rate		*Eq.Opp.*
	Male=1	Male=0	
ResNet-18 [1]	63.85	89.55	25.70
AdvDe* [14]	81.35	94.24	12.89
PAAL* [15, 5]	67.35	85.10	17.75
Ours*	88.51	97.77	**9.26**
AdvDe† [14]	68.90	89.58	20.98
PAAL† [15, 5]	73.46	94.40	20.94
Ours†	87.25	97.48	**10.23**

5.6 Race Classification

We also compare the race classification results of our model to baseline [1], AdvDe [14], and PAAL [15, 5]. For this experiment, we set CelebA dataset as the source dataset and UTK Face dataset as the target dataset. As shown in Table. 7, our model achieves the fairest *Eq.Opp.* of 1.7.

Moreover, to see how our model works under varying levels of the bias in the training dataset, we change the composition of training samples among four groups: Caucasian Males, Other Males, Caucasian Females, and Other Females. In Table. 8, our method performs better in terms of both accuracy and fairness (*Eq.Opp.*) in less imbalanced setting, as expected. In contrast, the gap of *Eq.Opp.* between the baseline and our model is larger in extremely imbalanced settings. This indicates that our model works well in the challenging situations.

5.7 t-SNE Visualization

To deeply analyze the effectiveness of our model, we visualize the representations from the TAC and other models using t-SNE method [33]. We first train them to fairly classify attractiveness attributes in terms of gender attributes on CelebA dataset and conduct visualization on 1,000 male and 1,000 female images randomly sampled in the test set. The visualization results are shown in

Table 7. Equality of Opportunity (*Eq.Opp.*) for race classification on UTK Face dataset [30], where we set the protected attribute as gender.

	True Positive Rate		Eq.Opp.
	Male=1	Male=0	
ResNet-18 [1]	86.4	67.2	19.2
AdvDe [14]	86.6	68.3	18.3
PAAL [15, 5]	76.8	73.8	3.0
Ours	76.1	74.4	**1.7**

Table 8. Equality of Opportunity (*Eq.Opp.*) for race classification on UTK Face dataset [30] under different statistics. We present the ratio of the number of training images in the following four cases: (Male,Caucasian):(Female,Others):(Male,Others):(Female,Caucasian).

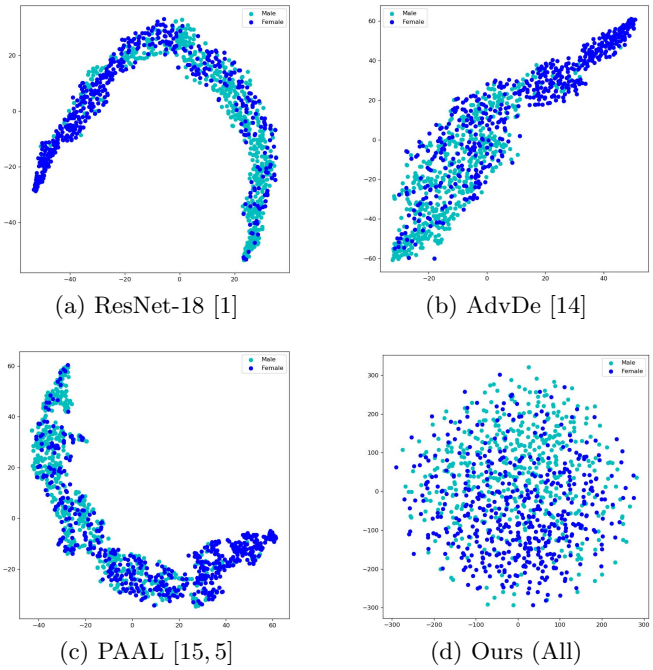
	ResNet-18 [1] (without de-biasing)			Ours		
	True Positive Rate		*Eq.Opp.*	True Positive Rate		*Eq.Opp.*
ratio	Male=1	Male=0		Male=1	Male=0	
1:1:1:1	86.8	84.8	2.0	86.1	87.5	1.4
1.5:1.5:1:1	79.6	87.0	7.3	84.7	86.5	1.8
4:4:1:1	67.2	86.4	19.2	74.4	76.1	1.7
9:9:1:1	57.3	81.8	24.5	61.0	64.8	3.8

Fig. 5, where the dark and light blue color denote female and male samples, respectively. We observe that the representations of female and male samples are separately grouped in other methods, indicating the bias towards the gender attribute. In contrast, the representations of our method are more scattered with respect to the gender attribute. This demonstrates that our model successfully learns fair representations with respect to the protected attribute.

6 Conclusion

In this paper, we tackled the problem of the biased results of AI systems in terms of sensitive characteristics, such as gender, age, or race. Since various real-world datasets do not have annotations for protected attributes, we proposed a framework for fairness-aware image classification, which can be trained on a dataset without protected attribute labels (*i.e.*, target dataset) by transferring knowledge from another dataset with protected attribute labels (*i.e.*, source dataset).

To leverage the knowledge, we introduced the Feature Independency Triplet loss which encourages the representation for target attributes to be independent of protected attributes. Moreover, we designed the Group-wise Fair loss to minimize the discrepancy on the misclassification rates among protected attribute groups. To validate the effectiveness of our method, we conducted experiments of facial attribute classification on CelebA and UTK Face datasets. Our experiments demonstrate that the proposed method achieved the fairest performance



(a) ResNet-18 [1] (b) AdvDe [14]

(c) PAAL [15, 5] (d) Ours (All)

Fig. 5. t-SNE Visualization [33] on CelebA dataset [29]. We visualize the representations of the attractiveness classifiers. Each of dark/light blue group denotes randomly selected female and male samples respectively. Less clustered are better.

in terms of *Equality of Opportunity*. In addition, through the t-SNE visualization, we showed that our representations are invariant to protected attributes.

To summary, we present a knowledge transfer method which works between two datasets with similar domains (*i.e.*, face images). However, the transfer between different domains is not investigated in this work. Adopting domain adaptation techniques would be interesting for future work.

7 Acknowledgement

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017M3C4A7069370) and Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (Development of framework for analyzing, detecting, mitigating of bias in AI model and training data) under Grant 2019-0-01396 and (Artificial Intelligence Graduate School Program (YONSEI UNIVERSITY)) under Grant 2020-0-01361.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
2. Barenstein, M.: Propublica's compas data revisited. arXiv preprint arXiv:1906.04711 (2019)
3. Farnadi, G., Babaki, B., Getoor, L.: Fairness in relational domains. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. AIES '18, New York, NY, USA, Association for Computing Machinery (2018) 108–114
4. Brandao, M.: Age and gender bias in pedestrian detection algorithms. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. (2019)
5. Wang, T., Zhao, J., Yatskar, M., Chang, K.W., Ordonez, V.: Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: The IEEE International Conference on Computer Vision (ICCV). (2019)
6. Hwang, S., Byun, H.: Unsupervised image-to-image translation via fair representation of gender bias. In: The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2020) 1953–1957
7. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, Association for Computational Linguistics (2017) 2979–2989
8. Wang, M., Deng, W., Hu, J., Tao, X., Huang, Y.: Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 692–702
9. Garcia, M.: Racist in the machine: The disturbing implications of algorithmic bias. World Policy Journal **33** (2016) 111–117
10. Park, S., Kim, D., Hwang, S., Byun, H.: Readme: Representation learning by fairness-aware disentangling method. arXiv preprint arXiv:2007.03775 (2020)
11. Hutchinson, B., Mitchell, M.: 50 years of test (un)fairness: Lessons for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19, New York, NY, USA, Association for Computing Machinery (2019) 49–58
12. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. NIPS'16, Red Hook, NY, USA, Curran Associates Inc. (2016) 3323–3331
13. Quadrianto, N., Sharmanska, V., Thomas, O.: Discovering fair representations in the data domain. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019)
14. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. AIES '18, New York, NY, USA, Association for Computing Machinery (2018) 335–340
15. Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: Learning not to learn: Training deep neural networks with biased data. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019)
16. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17** (2016) 2096–2030

17. Creager, E., Madras, D., Jacobsen, J.H., Weis, M., Swersky, K., Pitassi, T., Zemel, R.: Flexibly fair representation learning by disentanglement. In Chaudhuri, K., Salakhutdinov, R., eds.: Proceedings of the 36th International Conference on Machine Learning. Volume 97 of Proceedings of Machine Learning Research., Long Beach, California, USA, PMLR (2019) 1436–1445
18. Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., Rohrbach, A.: Women also snowboard: Overcoming bias in captioning models. In: European Conference on Computer Vision, Springer (2018) 793–811
19. Sattigeri, P., Hoffman, S.C., Chenthamarakshan, V., Varshney, K.R.: Fairness gan: Generating datasets with fairness properties using a generative adversarial network. IBM Journal of Research and Development **63** (2019) 3–1
20. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680
21. Edwards, H., Storkey, A.: Censoring representations with an adversary. International Conference on Learning Representations (2016)
22. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in neural information processing systems. (2016) 2172–2180
23. West, J., Ventura, D., Warnick, S.: Spring research presentation: A theoretical foundation for inductive transfer. Brigham Young University, College of Physical and Mathematical Sciences **1** (2007)
24. Wang, C., Yang, H., Meinel, C.: Image captioning with deep bidirectional lstms and multi-task learning. ACM Trans. Multimedia Comput. Commun. Appl. **14** (2018)
25. Lee, K.H., He, X., Zhang, L., Yang, L.: Cleannet: Transfer learning for scalable image classifier training with label noise. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
26. Gamrian, S., Goldberg, Y.: Transfer learning for related reinforcement learning tasks via image-to-image translation. In Chaudhuri, K., Salakhutdinov, R., eds.: Proceedings of the 36th International Conference on Machine Learning. Volume 97 of Proceedings of Machine Learning Research., Long Beach, California, USA, PMLR (2019) 2063–2072
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. (2015) 91–99
28. Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th international conference on world wide web. (2017) 1171–1180
29. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV). (2015)
30. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2017)
31. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In Wallach,

H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., eds.: Advances in Neural Information Processing Systems 32. Curran Associates, Inc. (2019) 8024–8035
32. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
33. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9** (2008) 2579–2605