FineFACE: Fair Facial Attribute Classification Leveraging Fine-grained Features*

Ayesha Manzoor and Ajita Rattani**

Dept. of Computer Science and Engineering,
University of North Texas at Denton, Texas, USA
{ayeshamanzoor}@my.unt.edu; {ajita.rattani}@unt.edu

Abstract. Published research highlights the presence of demographic bias in automated facial attribute classification algorithms, particularly impacting women and individuals with darker skin tones. Existing bias mitigation techniques typically require demographic annotations and often obtain a trade-off between fairness and accuracy, i.e., Pareto inefficiency. Facial attributes, whether common ones like gender or others such as "chubby" or "high cheekbones", exhibit high interclass similarity and intraclass variation across demographics leading to unequal accuracy. This requires the use of local and subtle cues using fine-grained analysis for differentiation. This paper proposes a novel approach to fair facial attribute classification by framing it as a fine-grained classification problem. Our approach effectively integrates both low-level local features (like edges and color) and high-level semantic features (like shapes and structures) through cross-layer mutual attention learning. Here, shallow to deep CNN layers function as experts, offering category predictions and attention regions. An exhaustive evaluation on facial attribute annotated datasets demonstrates that our FineFACE model improves accuracy by 1.32% to 1.74% and fairness by 67% to 83.6%, over the SOTA bias mitigation techniques. Importantly, our approach obtains a Pareto-efficient balance between accuracy and fairness between demographic groups. In addition, our approach does not require demographic annotations and is applicable to diverse downstream classification tasks. To facilitate reproducibility, the code and dataset information is available at https://github.com/VCBSL-Fairness/FineFACE.

Keywords: Fairness in AI · Facial Attribute Classification · Fine-grained Features.

1 Introduction

Automated facial analysis algorithms encompass face detection, face recognition, and facial attribute classification (such as gender, race, high cheekbones,

* Supported by National Science Foundation, USA
** Corresponding author.

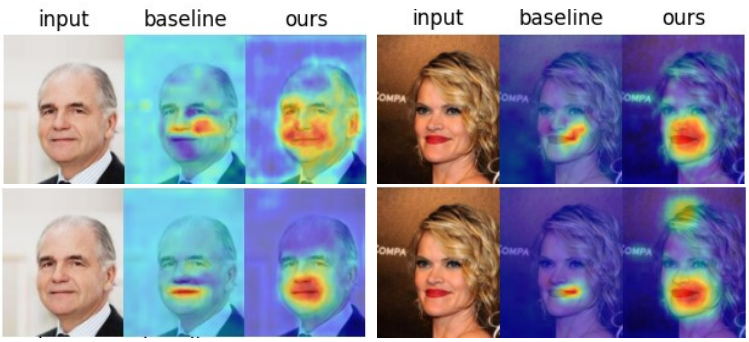

Fig. 1. Visualization of the attention map obtained by our proposed FineFACE over baseline (both using ResNet50 backbone) for facial attribute classification. The highly activated region is shown by the red zone on the map, followed by yellow, green, and blue zones. Top: "High Cheekbones" classifier. Bottom: "Smiling" classifier.

and attractiveness) [8,31,17]. These algorithms are deeply integrated into various sectors, such as surveillance and border control, retail and entertainment, healthcare, and education.

Numerous existing studies [1,12,26] investigating the *fairness of facial attribute classification* algorithms confirm the presence of *performance disparities between demographic groups, such as gender and race*. Thus, bias in these algorithms emerges as a significant societal issue that warrants immediate redress, particularly for the large-scale deployment of fair and trustworthy systems across demographics. In this direction, the vision community has proposed several bias mitigation techniques to address the performance disparities of facial attribute classifiers. Established bias mitigation techniques utilize regularization [13], attention mechanism [21], adversarial debiasing [33,3], GAN-based over-sampling [36,25], multi-task classification [5], and network pruning [15].

These existing bias mitigation techniques often require demographically annotated training sets and are limited in their generalizability. Importantly, these techniques often sacrifice overall classification accuracy in pursuit of improved fairness, making them *Pareto inefficient* [36,33]. It was demonstrated in [36] that fairness violations in vision models are largely driven by the variance component of bias-variance decomposition. Consequently, *one effective way to improve fairness is by decreasing the variance within each demographic subgroup by focusing on local and subtle cues*. This can be obtained through learning enhanced **feature representation** for each demographic subgroup, also supported by [4,12].

Following this line of thought, enhancing feature representation for each demographic subgroup is crucial in improving fairness without compromising overall performance. Traditional facial attribute classifiers [2,12,23,3] rely predominantly on high-level discriminative and semantically meaningful information often obtained from the final layers of the deep convolutional neural network (CNN). However, the lower layers of the deep learning model capture (low-level) essential features and patterns in faces vital for attribute classification, such as **(a)** facial contours and edges, including the outline of the face, jawline and

cheekbones, **(b)** texture of facial regions, such as skin and hair, **(c)** position and shape information, and **(d)** lighting condition and its effect on the appearance of facial features. Integrating low-level details from the lower layers of the model will capture local and detailed cues in the learned feature representation.

In our quest to identify these subtle and local cues for learning enhanced feature representation, we **aim** to leverage fine-grained analysis, integrating both high- and low-level features, toward fair facial attribute classification, FineFACE. This is facilitated through a cross-layer mutual attention learning technique that learns fine-grained features by considering the layers of a deep learning model from shallow to deep as independent 'experts' knowledgeable about low-level detailed to high-level semantic information, respectively. These experts are trained in leveraging mutual data augmentation to incorporate attention regions proposed by other experts. An ordinary deep learning model can be considered to use only the deepest expert (using high-level semantic information) for classification. In contrast, our method consolidates the prediction of the categorical label and the attention region of each expert for the final facial attribute classification task.

Fig. 1 shows the final CAM visualization obtained by our proposed FineFACE model based on the ResNet50 backbone, for facial attribute classification with "high cheekbone" and "smiling" as target variable using the CelebA dataset [17]. The highly activated region is shown by the red zone on the map, followed by yellow, green, and blue zones in the attention map. For cross-comparison, the visualization of the baseline ResNet50 is also shown for the same classification task. As illustrated in the maps, our FineFace model captures additional information, such as the contours of facial regions derived from the lower layers of the model, leading to enhanced feature representation and, hence, accurate and fair facial attribute classification.

Contributions. In summary, the contributions of our work are as follows: **(i)** We approach fair facial attribute classification from a novel perspective by reformulating it as a fine-grained classification task, **(ii)** We propose a novel approach based on cross-layer mutual attention learning where the prediction is consolidated from shallow (using low-level details) to deep layers (using high-level semantic details) regarded as an independent experts, **(iii)** Extensive evaluation on facial attribute annotated datasets namely, FairFace [11], UTKFace [34], LFWA+ [17], and CelebA [18], and **(iv)** Cross-comparison with the existing bias mitigation techniques, demonstrating the efficacy of our approach in terms of significant improvement in fairness as well as classification accuracy. Thus, obtaining state-of-the-art Pareto-efficient performance.

2 Related Work

In this section, we review the related academic literature.

Bias Mitigation of Facial Attribute Classification. Many studies have highlighted the systematic limitations of facial attribute classification (such as

gender, race, and age) between gender-racial groups [2,12,23]. Studies in [3,27] reported bias of facial attribute classification for attractive, smiling, and wavy hair as the target attributes across genders. Following this study, [36] reported bias of gender-independent target attributes, such as black hair, smiling, slightly open mouth, and eyeglasses, between genders. Consequently, numerous strategies have been proposed to mitigate bias [22]. [5] explored the joint classification of gender, age, and race by proposing a multi-task network. [33] included a variable for the group of interest and simultaneously learned a predictor and an adversary via adversarial debiasing. [13] leveraged the power of semantic preserving augmentations at the image level in a self-consistency setting for fair gender classification tasks. [24] introduced a framework that integrates fairness constraints directly into the loss function using Lagrangian multipliers for fair classification. [3] proposed "fair mixup," a data augmentation technique by interpolating data points that improve the generalization of the classifiers trained under group fairness constraints. [25] adopted structured learning techniques using deep-views of the training samples generated using GAN-based latent code editing to improve the fairness of the gender classifier. GAN-based SMOTE "g-SMOTE" was proposed by [36] to strategically enhance the training set for underrepresented subgroups to mitigate bias.

Fine-grained Visual Classification. Fine-grained classification is a challenging research task in computer vision, which captures the local discriminative features using attention learning [35,6] to distinguish different fine-grained categories. In addition to methods based on attention mechanisms, second-order pooling methods utilize the second-order statistics of deep features to compose powerful representations such as combined feature maps [14] and covariance among deep features [28] for fine-grained classification. Studies have also been proposed to use features or information learned from different layers within a CNN backbone for fine-grained classification. [32] proposed a multi-layered Deconvnet for gaining insight into the functions of intermediate feature layers. They discovered that shallow layers capture low-level details, whereas deep layers capture high-level information. [10] proposed the LayerCAM, which indicates the discriminative regions used by the different layers of a CNN to predict a specific category. Inspired by these two works, [16] proposed the CMAL-Net, which focuses on using attention regions predicted by layers of different depths to mark the cues they learned, letting layers of varying depths to learn from each other's knowledge to improve overall performance.

3 Proposed Methodology

In this section, we elaborate on our proposed FineFACE model based on learning features from different layers of the CNN using the attention mechanism and mutual learning, following the foundational works in [32,10,16].

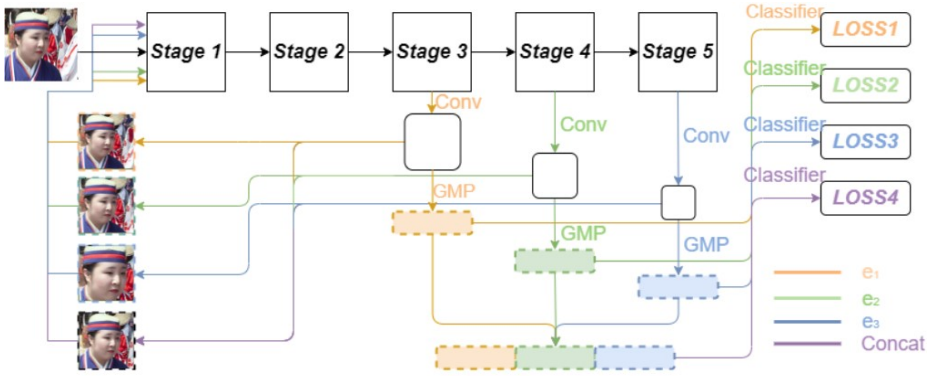

Fig. 2. FineFACE network structure. This figure illustrates this method by introducing three experts e_1, e_2, e_3, on a 5-stage backbone CNN (e.g., ResNet50). The working of each expert and the concatenation of experts are depicted in different colors. Each expert receives feature maps from a specific layer as input and generates a categorical prediction along with an attention region, which is used for data augmentation by other experts. This architecture is trained in multiple steps within each iteration. We start by training the deepest expert (e3), followed by the shallower experts. Finally, in the last step, we train the concatenation of experts to enhance overall performance.

3.1 Expert Construction: Using Shallow to Deep Layers

In this subsection, the construction of experts from shallow to deep layers. Any state-of-the-art CNNs, such as ResNet50, Res2NeXt50, etc. can be used as the backbone CNN denoted by β. β has M layers, and $\{l_1, l_2, ..., l_m, ..., l_M\}$ denote the layers of β from shallow to deep (except the fully connected layers). $\{e_1, e_2, ..., e_n, ..., e_N\}$ are N experts based on these M layers. Each expert encompasses layers from the first layer up to a certain layer such that - e_n consists of the layers from l_1 to l_{m_n}, and $1 \leq m_n \leq$ M. The experts $\{e_1, e_2, ..., e_n, ..., e_N\}$ progressively cover deeper layers of the backbone CNN, and e_N, the deepest expert, covers all layers from l_1 to l_M.

Let $\{x_1, x_2, ..., x_n, ..., x_N\}$ denote the intermediate feature maps produced by β for the experts $\{e_1, e_2, ..., e_n, ..., e_N\}$, respectively. $x_n \in R^{H_n \times W_n \times C_n}$ and H_n, W_n and C_n denote the height, width, and number of channels, respectively. A set of functions $\{F_1(.), F_2(.), ..., F_n(.), ..., F_N(.)\}$ are used to respectively compress $\{x_1, x_2,..., x_n, ..., x_N\}$ into 1D vectorial descriptors $\{v_1, v_2, ..., v_n, ..., v_N\}$, and $v_n \in R^{C_v}$. C_v denotes the length of the 1D vectorial descriptors, and these descriptors given by various experts are of the same length. The $F_n(.)$ for processing x_n is defined as:

$$v_n = F_n(x_n) = f^{GMP}(x_n^{''}), \tag{1}$$

$$x_n^{''} = f^{Elu}(f^{bn}(f^{conv}_{3\times3\times C_{v/2}\times C_v}(x_n^{'}))), \tag{2}$$

$$x_n^{'} = f^{Elu}(f^{bn}(f^{conv}_{1\times1\times C_n\times C_{v/2}}(x_n))), \tag{3}$$

where $f^{GMP}(.)$ denotes the Global Max Pooling. $f^{conv}(.)$ depicts the 2D convolution operation by its kernel size. $f^{bn}(.)$ and $f^{Elu}(.)$ denote batch normalization and Elu operations respectively. $x_n^{'}$ and $x_n^{''}$ are intermediate feature maps produced by e_n. Thereafter, $x_n^{''} \in R^{H_n\times W_n\times C_n}$ is used to generate the attention region of e_n as described in subsection 3.2. $\{p_1, p_2, ..., p_n, ..., p_N\}$ denote the prediction scores given by different experts, obtained as $p_n = f_n^{clf}(v_n)$, where $f_n^{clf}(.)$ denotes a fully connected layer-based classifier.

Apart from the prediction scores obtained by the various experts, an overall prediction score is also generated by combining the information from different experts. Specifically, $\{v_1, v_2, ..., v_n, ..., v_N\}$ are first concatenated for an overall descriptor v_{oval} as: $v_{oval} = f_{concat}(v_1, v_2, ..., v_n, ..., v_N)$, where $f_{concat}(.)$ denotes concatenation operation. Then v_{oval} is processed into an overall prediction score p_{oval} by a fully connected layer-based classifier as $p_{oval} = f_{oval}^{clf}(v_{oval})$

3.2 Attention Region Prediction

As mentioned above, $x_n^{''}$ denotes an intermediate feature map generated by the expert e_n. We move with an assumption that the classification problem is K-class, and $k_n \in 1, 2, ..., K$ is the category predicted by expert e_n and $x_n^{''} \in R^{H_n\times W_n\times C_n}$. The generation of the attention region proposed by e_n is initialized by producing the class activation map (CAM), which specifies the discriminative image region, for the category k_n based on $x_n^{''}$ specified. The CAM Ω_n ($\Omega_n \in R^{H_n\times W_n}$) produced by the expert e_n is defined as: $\Omega_n(\alpha,\beta) = \sum_{c=1}^{c_v} p_n^c x_n^{''c}(\alpha,\beta)$, where the coordinates (α, β) denote the spatial location of $x_n^{''}$ and $\Omega_n.p_n$ denotes the parameters of $f_n^{clf}(.)$ corresponding to the predicted category k_n. Then, after obtaining Ω, an attention map $\tilde{\Omega}_n \in R^{H_{in}\times W_{in}}$ (H_{in}, W_{in} are the height and width of the input image, respectively) is generated by upsampling Ω_n using a bilinear sampling kernel. Thereafter, $\tilde{\Omega}_n$ is applied with min-max normalization, and each spatial element of the normalized attention map $\tilde{\Omega}_n^{norm}$ is obtained by

$$\tilde{\Omega}_n^{norm}(\alpha, \beta) = \tilde{\Omega}_n(\alpha, \beta) \text{ - } \min(\tilde{\Omega}_n)/\max(\tilde{\Omega}_n) \text{ - } \min(\tilde{\Omega}_n). \tag{4}$$

The regions that the expert e_n considers discriminative can be found and cropped by generating a mask $\tilde{\Omega}_n^{mask}$ by setting the elements in $\tilde{\Omega}_n^{norm}$ to 1 for values greater than a threshold t ($t \in [0, 1]$) and 0 for the others. Then, a box that covers all the positive regions of $\tilde{\Omega}_n^{mask}$ is located and cropped from the input image. The cropped region is upsampled to the input image's size and the upsampled attention region A_n is considered as the attention region predicted by e_n and also as data augmentation for remaining experts. Apart from the attention regions proposed by various experts, an overall attention region A_{oval} is generated by summing up the attention information learned by different experts.

3.3 Multi-step Mutual Learning

The experts are trained using progressive multi-step strategy with cross-entropy loss. In the early steps, these experts are trained one by one, which allows them to "focus on" learning the clues of their own expertise without being diverted by other experts. In the last two steps, the experts get together to learn impactful information from the attention regions and the raw image, respectively. Specifically, every iteration of the training takes place in $N + 2$ steps, and in the first N steps, each expert is gradually trained from deep to shallow. In the first step, the deepest expert e_N is trained. Since the training of N involves the experts shallower than e_N, the attention regions proposed by all the experts and the overall attention region $\{A_1, A_2, ..., A_n, ..., A_N, A_{oval}\}$ are also generated at this step. These attention regions showcase the "specialized knowledge" of the experts by highlighting the basis on which each expert made its classification.

From step 2 to N, there is a progression to shallower experts by randomly selecting one input from a pool of images comprising of the raw input and the attention regions proposed by the other experts. The shallow experts rely on the attention regions proposed by deeper experts to learn semantic visual clues (e.g., eyes, nose, and mouth), while the deep experts take the help of shallow experts by learning low-level visual cues (e.g., facial contours like jawline, cheekbones, etc.) from their proposed attention regions. In step $N + 1$, all the experts and their concatenation are trained with the overall attention region A_{oval} in one pass. This step enforces all experts to work together and study the attention information they have combinedly gained for learning more fine-grained features. At step $N + 2$, the concatenation of all the experts is trained with the raw input to make sure the parameters of $f^{clf}_{oval}(.)$ fit the resolution of the original input. The **algorithm** for the multi-step mutual learning strategy is included in Section 2 of the supplementary material.

Inference Phase: Fig. 2 illustrates the inference stage of the proposed FineFACE model with $N + 1$ classifiers. For an input image during the inference, $N + 1$ prediction scores are produced by the proposed architecture. For each test image, the raw input and overall attention region are successively fed to the model obtaining $2 \times (N + 1)$ number of prediction scores. The final prediction score for the inference is the average of the $2 \times (N + 1)$ scores. This inference strategy maximizes the classification accuracy as well as fairness of the trained model due to obtaining two kinds of complementary information: (a) information from the prediction scores from various experts and the overall prediction score, and (b) the information from the raw input and overall attention region.

4 Experimental Details

We conducted two sets of **experiments (1)** a face-based gender classifier with gender as the target attribute and race and gender as the protected attributes following studies in [13,25].**(2)** 13 gender-independent facial attribute classifiers following studies in [36,27,3] with "bags under eyes", "bangs", "black hair", "blond hair", "brown hair", "chubby", "eyeglasses", "gray hair", "high cheekbones", "mouth

slightly open", "narrow eyes", "smiling", and "wearing hat" as the 13 gender independent target attributes and gender as the protected attribute. We used the mean scores of these 13 attribute classifiers, following studies in [36,27,3].

4.1 Datasets and Training Protocol

We used standard benchmark datasets widely adopted for evaluating fairness of facial attribute classifiers [13,36,27], namely, FairFace [11], UTKFace [34], LFWA+ [17], and CelebA [18]. In line with existing studies [13,25], a face-based gender classifier was trained on FairFace and evaluated on FairFace, UTKFace, LFWA+, and CelebA (40 attributes). Unlike UTKFace and LFWA+, CelebA does not have race annotations. Hence, we used only the gender attribute for CelebA. For the 13 gender-independent facial attribute classifiers, we used the CelebA (40 attributes) dataset for training and validation. For the fair comparison with the existing studies on fairness [36,27,3], we used only 13 gender-independent attributes from CelebA. Note that protected attribute annotation information is not used during the model training stage, but solely for the purpose of fairness evaluation of the facial attribute classifiers. Additional details on these datasets are given in Table 1

Dataset	**Images**	**Demographic groups**
FairFace	108K	White, Black, Indian, Asian, Southeast Asian, Middle Eastern, Latino Hispanic
UTKFace	20K	White, Black, Indian, Asian, Others
LFWA+	13K	White, Black, Indian, Asian, Undefined
CelebA	202K	Not Available (only gender information available)

Table 1. Dataset details including the number of images and demographic groups

4.2 Implementations Details

For a fair comparison with studies in [36,12,25], we utilized ResNet50 [9] as our method's backbone CNN architecture. The layers of ResNet50, excluding the fully connected layers, are grouped into 5 stages where each stage is a group of layers operating on feature maps of the same spatial size. We use these stages as building blocks for our experts: e_1 encompasses layers from stage 1 to stage 3, e_2 encompasses layers from stage 1 to stage 4, and e_3 encompasses layers from stage 1 to stage 5. In general, the number of stages in a model can be determined by grouping layers operating on the feature maps of the same spatial size, and accordingly, experts can be formed.

We trained all the models used in this study using Stochastic Gradient Descent (SGD) with the number of epochs determined using an early stopping mechanism, the momentum of 0.9, weight decay of 5×10^{-4}, and a mini-batch size of 16 determined using empirical evidence. The learning rate was set as 0.002 with cosine annealing [19]. We fixed the input image size as 448×448, following the common settings in existing fairness studies [7,20]. The threshold t, which is

used to generate a mask for the attention region, was set to 0.5 (see Section 3.2). We also conducted an *ablation study* of the related design choices (a) different pooling methods for building experts, and (b) the contribution of fusing the prediction scores. See **ablation studies** in Section 3 of the supplementary material for more details.

4.3 Evaluation Metrics

For the gender classifier, the standard evaluation metrics, namely, classification accuracy, Max-Min ratio (the ratio of the best and worst performing subgroups), and Degree of Bias (DoB) (standard deviation of accuracy) are used for fair comparison with the existing studies [13,25,33,5] on bias evaluation of gender classifier. As a fair model is supposed to have consistent accuracies across all subgroups, this implies that a fair model would have a max-min accuracy ratio closer to 1 and a DoB closer to 0.

For gender-independent facial attribute classifiers, following the studies in [36,27,3], we used classification accuracy, True Positive Rate (TPR), Difference of Equal Opportunity (DEO) and Difference in Equalized Odds (DEOdds) where Equal Opportunity (EO) requires a classifier to have equal TPRs on each subgroup and a violation of this equal opportunity is measured by the DEO. DEOdds measures the absolute difference in the probability of correctly predicting the positive class between the subgroups for each actual outcome, summed over all possible outcomes [36,3,27]. Furthermore, we also analyzed the maximum group accuracy and the minimum group accuracy associated with the best and worst performing demographic subgroups, respectively. A model that can maintain or improve accuracy and TPR while reducing DEO and DEOdds would be an ideal classifier in terms of enhancing accuracy as well as fairness.

5 Results

5.1 Face-based Gender Classification

In this section, we will discuss the performance and fairness of the face-based gender classifier across gender-racial groups.

Intra-Dataset Evaluation: Table 2 shows the performance of the baseline ResNet50 model and our proposed FineFACE model in gender classification when trained and tested on the FairFace dataset. As can be seen, our proposed FineFACE model reduced the Degree of Bias (DoB) and the Max-Min accuracy ratio by approximately 86% and 13%, respectively, over the baseline. At the same time, the overall classification accuracy improved by about 3% over the baseline ResNet50 model. Note, we also evaluated and compared performance of the baseline DenseNet architecture over FineFACE using DenseNet backbone for gender classification. The experimental results demonstrated the efficacy of FineFace in improving accuracy and fairness over the baseline DenseNet architecture. Thus, highlighting the importance of systematic construction of experts

from shallow to deep layers followed by attention region prediction and multi-stage learning by FineFACE over feature reuse between shallow to deep layers by the baseline DenseNet. More details on the experimental results are given in Section 1.1 of the supplementary material.

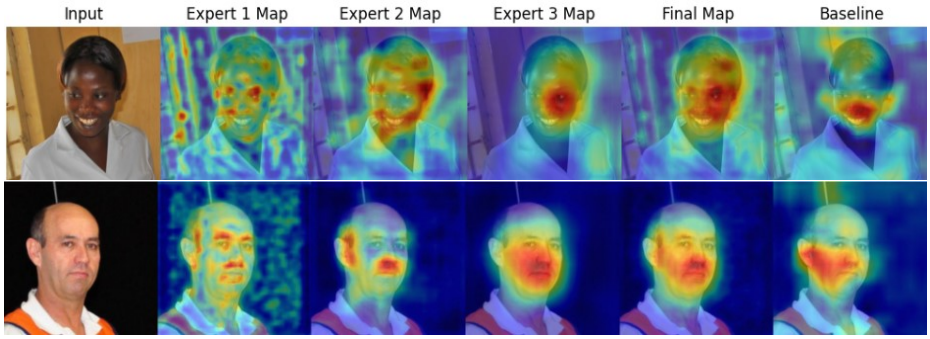

Fig. 3. Visualization Results of Gender Classifier. Left through right in each set of images are the input image from FairFace dataset, visualization results based on our FineFACE method's 3 experts ($\tilde{\Omega}_1^{norm}$, $\tilde{\Omega}_2^{norm}$, $\tilde{\Omega}_3^{norm}$), and our method's final visualization ($\tilde{\Omega}_{oval}^{norm}$), versus a basic ResNet50 architecture's final visualization ($\tilde{\Omega}_{ori}^{norm}$). Our FineFACE captures a more comprehensive feature representation of the image, thereby enhancing fairness as well as accuracy.

Cross-Dataset Evaluation: Table 3 shows the results of the baseline ResNet50 and our proposed FineFACE, based on ResNet50 backbone when trained on FairFace and evaluated on UTKFace and LFWA+ datasets. Our model significantly reduced the bias of the gender classifier by reducing DOB by approximately 55% and 77%, and Max-Min ratio by 18% and 17% over the baseline even on the cross-dataset evaluation, respectively. Overall, performance degradation of the classifiers is minimal on cross-dataset evaluation except for the UTKFace dataset due to poor quality samples majorly showing age progression. Table 4 shows the results on the CelebA test set. Our model reduced DOB by approximately 41% and Max-Min ratio by 2%. These results demonstrate the efficacy of our model in significantly reducing bias as well as improving accuracy even on the cross-dataset evaluation.

Fig. 3 shows the visualization of the attention map learned by the ResNet50-based FineFACE and the baseline ResNet50-based gender classifier. For proposed FineFACE, we generate 4 heatmaps for each image, i.e., $\tilde{\Omega}_1^{norm}$ from expert 1, $\tilde{\Omega}_2^{norm}$ from expert 2, $\tilde{\Omega}_3^{norm}$ from expert 3 and $\tilde{\Omega}_{oval}^{norm}$ which is the aggregation of the three experts' heatmaps and is used for the final prediction (refer to Section 3.2). The maps generated by Expert 1 show a focus on low-level features such as edges, evident from the scattered attention across the face, capturing details such as the outline of the face, eyes, nose, and mouth. The maps generated

by Expert 3 have more concentrated attention on key facial regions that are critical for gender classification, such as the central face area. Thus, there is a clear progression in the focus of attention from Expert 1 to Expert 3 and all the varying levels of attention are captured in the final concatenated map (Final Map). As the original ResNet50 has only a 1 classifier for prediction, we generated 1 heatmap $\tilde{\Omega}_{ori}^{norm}$ using the feature maps from the last convolutional layer. Note, $\tilde{\Omega}_{3}^{norm}$ and $\tilde{\Omega}_{ori}^{norm}$ are both generated based on the feature maps of the last convolutional layer of the ResNet50 backbone, but $\tilde{\Omega}_{3}^{norm}$ captures much more comprehensive and accurate information than $\tilde{\Omega}_{ori}^{norm}$. Further, the overall feature map from the proposed FineFACE model illustrates the efficacy of the fine-grained framework in capturing comprehensive and discriminant regions vital for gender classification over the baseline.

Race	White		Black		East Asian		SE Asian		Latino		Indian		Middle E				
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F	Max/ Min↓	Overall↑	DoB↓
Baseline	96.5	89.9	94.4	82.4	97.2	88.9	94.4	91.5	95.6	92.2	98.1	93.3	97.8	92.4	1.18	93.2	4.2
FineFACE	97.1	97	97.2	96.2	97	96.2	96.2	97	96	96.3	96.6	95.9	96.1	95.1	**1.02**	**96.4**	**0.6**

Table 2. Gender Classification Accuracy (%) on FairFace testset across different demographics using baseline ResNet50 and our proposed FineFACE. M stands for male and F stands for female. The top performance results are highlighted in bold.

	Race	White		Black		Asian		Indian		Others/ Undefined				
Dataset	Gender	M	F	M	F	M	F	M	F	M	F	Max/Min↓	Overall↑	DoB↓
UTKFace	Baseline	90.2	72.2	94.6	67.6	88.2	65.7	95.1	74.9	89.1	78.8	1.45	81.9	11.1
	FineFACE	91	86.5	93.9	79.7	91.1	81.1	95.1	86.3	90	88.3	**1.19**	**88.5**	**5**
LFWA+	Baseline	96.9	89.1	98.7	78.8	96.5	78.3	97.9	95	96.8	91.1	1.26	95.3	7.7
	FineFACE	99.1	98	98.9	95.3	98.6	94.8	100	100	98.4	96.2	**1.05**	**98.6**	**1.9**

Table 3. Cross-dataset evaluation - Gender Classification Accuracy (%) on UTKFace and LFWA+ test sets across different demographics for baseline ResNet50 and our proposed FineFACE.

Gender	M	F	Max/Min↓	Overall↑	DoB↓
Baseline	90.6	94.9	1.05	93.2	2.2
FineFACE	96.4	99	**1.03**	**98**	**1.3**

Table 4. Cross-dataset evaluation - Gender Classification Accuracy (%) on CelebA testset across gender for baseline Resnet50 and our proposed FineFACE.

Comparison with Published Work: Table 5 shows the performance of our proposed FineFACE method over published bias mitigation techniques based on multi-tasking [5], adversarial debiasing [33], deep generative views [25], and consistency regularization [13]. All these studies are reported for the ResNet50 based gender classifier trained and tested on the FairFace dataset.

As can be seen, our proposed FineFACE obtained the lowest DoB of 0.26 and the Max-Min accuracy ratio of 1.008 over all the existing published studies. Moreover, the overall accuracy was not only maintained but also increased by 1.74% compared to the second-best model (indicated as D in the Table) based on Deep generative views [25]. Therefore, our proposed method obtains state-of-the-art performance.

Further, existing bias mitigation techniques based on adversarial debiasing [33] and multitasking [5] need demographically annotated data during training. The generative techniques based on deep generative views [25] and consistency regularization [13] are computationally very expensive and obtain low generalizability. Compared to the existing methods, the proposed FineFACE offers significant advantages: it mitigates bias in the absence of protected attributes, offers high generalizability, and is application-agnostic. Importantly, our method significantly improves fairness along with overall classification accuracy, emphasizing the importance of fine-grained classification.

Method	Accuracy								DoB↓	Max/Min↓
	Black	East Asian	Indian	Latino Hispanic	Middle Eastern	Southeast Asian	White	Overall↑		
A	91.26	94.45	95.05	95.19	97.35	94.2	94.96	94.64	1.81	1.067
B	87.66	91.93	93.67	93.8	95.96	91.81	93.96	92.69	2.62	1.095
C	90.83	93.6	94.48	94.7	95.94	93.64	94.57	94	1.59	1.056
D	91.64	95.29	95.38	95.32	97.11	93.5	94.92	94.72	1.72	1.06
FineFACE	96.21	96.84	96.37	96.61	96.53	96.04	96.55	**96.46**	**0.26**	**1.008**

Table 5. Comparative Analysis with FineFACE. A: Multi-Tasking [5] , B: Adversarial debiasing [33], C: Consistency Regularization [13] D: Deep Generative Views based [25]. The top performance results are highlighted in bold.

5.2 Gender-independent Facial Attribute Classification

In this section, we will discuss the performance of our 13 gender-independent facial attribute classifiers. We report mean scores over the 13 labels [27] called gender-independent target attribute (refer to Section 4 for more details on the 13 target attributes) with gender as the protected attribute.

Table 6 shows the performance of the gender-independent facial attribute classifier using our proposed FineFACE over the baseline ResNet50 model. FineFACE improves overall accuracy, minimum group accuracy, and TPR, while significantly reducing bias by approximately 91% (DEO) and 92% (DEODD). There is a marginal reduction in maximum group accuracy by only 1.52%. Worth mentioning, facial attributes like "bags under eyes", "chubby", "high cheekbones", "narrow eyes", and "smiling" are more subtle in nature. Thus more detailed features from lower layers help in understanding the subtle cues differentiating a normal cheekbone from a high cheekbone, a narrow eye from a normal eye, and a natural curvature of lips versus a smile across gender. Thus, obtaining performance enhancement as well as bias reduction for these gender-independent facial attributes.

Method	Accuracy↑	Max. grp. Acc.	Min. grp. Acc.	TPR↑	Max. grp. TPR	Min. grp. TPR	DEO↓	DEODD↓
Baseline	92.47	**94.46**	90.14	67.9	73.88	61.34	12.54	16.54
FineFACE	**92.85**	92.94	**92.76**	**76.48**	**76.97**	**75.79**	**1.18**	**1.38**

Table 6. Facial Attribute Classification Accuracy (%) on the CelebA dataset for baseline and our proposed FineFACE - mean scores over the 13 attributes [27] called gender-independent are reported. The top performance results are highlighted in bold.

Comparison with Published Work: In this section, we compare the performance of our proposed FineFACE over bias mitigation techniques namely, domain-independent models [29] (Domain Indep.), regularization [24,30] (Regularized), FairMixup [3], GAN-based offline dataset debiasing [27] (GAN Debiasing), and adaptive sampling [36] (g-SMOTE + Adaptive Sampling), reported for the 13 gender-independent facial attribute classifiers.

We summarized the results in the Table 7. The proposed FineFACE has achieved improved performance in both overall accuracy and accuracy of the worst-performing group compared to the baseline classifier. Although there is a slight reduction in the accuracy of the best-performing group (by 1.52%), the performance of this group remains comparable to the overall and worst-performing groups which improve by approximately 0.4% and 2.5% respectively. Worth mentioning, among 6 existing bias mitigation techniques in Table 7, only g-SMOTE [27] and g-SMOTE adaptive sampling [36] are able to improve performance of the worst performing group, with respect to the baseline, along with our proposed FineFACE, thus obtaining **pareto-efficiency**. Furthermore, our method obtains the highest improvement in the TPR of the worst-performing groups, along with the second-best results for overall TPR and the TPR for the best-performing groups. The **key highlight** of our method is the substantial reduction in both DEO and DEOdds, 3× lower than the next best method i.e., Fair Mixup [3]. Comparison of the various fairness methods is visually represented in Fig. 1, Section 1.2 of the supplementary material

Method	Weight -ing*	Domain Indep.*	Baseline single task	GAN Debiasing	Regular -ized	g-SMOTE + Adap. Sampl.	g-SMOTE	Baseline FairMixup	Fair Mixup	FineFACE [ours]
Accuracy↑	91.45	91.24	92.47	92.12	91.05	92.56	92.64	92.74	88.46	**92.85**
Max. grp. Acc.	93.35	93.04	94.46	94.03	94.42	94.44	**94.59**	93.85	90.42	92.94
Min. grp. Acc	89.06	88.93	88.93	89.85	87.86	90.36	90.35	91.44	86.36	**92.76**
TPR↑	64.02	70.74	67.90	66.13	54.2	67.11	66.14	**79.13**	46.67	76.48
Max. grp. TPR	67.41	75.61	73.88	70.36	56.11	74.06	73.43	**80.89**	47.85	76.97
Min. grp. TPR	59.74	66.05	61.34	61.25	52.34	59.78	58.32	72.92	44.27	**75.79**
DEO↓	7.67	9.56	12.54	9.11	3.77	14.28	15.11	7.97	3.58	**1.18**
DEODD↓	9	13.29	16.54	12.04	5.06	19.3	19.32	10.06	4.29	**1.38**

Table 7. Fairness methods on the CelebA dataset - mean scores over the 13 labels [27] called gender independent.The top performance results are highlighted in bold.

We also reported the minimum group accuracy of the baseline ResNet50 model and our FineFACE for each of the 13 gender-independent attributes individually

in Table 8. Our model outperformed the baseline for all except 1 attributes ("Narrow Eyes" by the baseline is more accurate by approximately 4%). For the other 12 attributes, our model outperformed at least by 0.5% and at most by 7.9%. Thus, we also demonstrate the efficacy of our FineFACE in improving the minimum group accuracy for the majority (12 out of 13 attributes) of facial attributes on an individual basis.

Attribute Name	**Baseline**	**FineFACE**
Bags Under Eyes	73.24	80.73
Bangs	94.67	95.9
Black Hair	86.4	90.51
Blond Hair	91.96	94.41
Brown Hair	81.26	89.14
Chubby	89.29	95.64
Eyeglasses	99.23	99.73
Gray Hair	95.28	98.35
High Cheekbones	85.53	87.83
Mouth Slightly Open	93.46	94.23
Narrow Eyes	91.97	87.99
Smiling	91.64	93.17
WearingHat	98.21	99.08

Table 8. Minimum Group Accuracy for the 13 gender-independent individual attributes.

6 Conclusion

The task of facial attribute classification presents inherent complexities due to high inter-class similarity, significant intra-class variation, and demographic diversity, which often result in performance disparities across protected attributes.

To effectively tackle these challenges, it is essential to incorporate local and subtle cues into the classification process. In our research, we propose a novel fine-grained feature framework designed for demographically fair facial attribute classification. This framework integrates detailed low-level and semantic high-level information across shallow to deep layers of the model. Through extensive evaluation on widely used facial attribute datasets, our approach demonstrates significant effectiveness in learning fair representation, achieving up to a three-fold reduction in bias compared to state-of-the-art bias mitigation techniques. Importantly, our method achieves a Pareto-efficient balance between accuracy and fairness without requiring the presence of protected attribute labels during classifier training—a critical advantage given privacy concerns and regulatory constraints that often prohibit the collection of such sensitive data. Furthermore, our study marks the first benchmark evaluation of the fairness of facial attribute classifiers using fine-grained features compared to existing supervised bias mitigation techniques.

While the multi-step training strategy extends the training duration compared to the original backbone networks, the training is an offline process and the more significant concern in real-world applications is the inference cost which is affordable for our method. As part of future work, we will also explore other backbone architectures such as Transformer. In addition, we will further analyze biases across intersectional groups, such as gender + target attribute, following insights from recent studies [36,3].

Acknowledgements This work is supported by National Science Foundation (NSF) award no. 2129173.

References

1. Albiero, V., et al.: Analysis of gender inequality in face recognition accuracy. In: Proc. IEEE WACV Workshops 2020. pp. 81–89 (2020)
2. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: FAT. PMLR, vol. 81, pp. 77–91. PMLR (2018)
3. Chuang, C.Y., Mroueh, Y.: Fair mixup: Fairness via interpolation. arXiv preprint arXiv:2103.06503 (2021)
4. Cui, J., Zhu, B., Wen, X., Qi, X., Yu, B., Zhang, H.: Classes are not equal: An empirical study on image recognition fairness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23283–23292 (2024)
5. Das, A., Dantcheva, A., Brémond, F.: Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach. In: ECCV Workshops (1). vol. 11129, pp. 573–585 (2018)
6. Fu, J., Zheng, H., Mei, T.: Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proc. of IEEE Conf. on CVPR. pp. 4438–4446 (2017)
7. Gao, Y., Han, X., Wang, X., Huang, W., Scott, M.: Channel interaction networks for fine-grained image categorization. In: AAAI conference. pp. 10818–10825 (2020)
8. Guo, G., Zhang, N.: A survey on deep learning based face recognition. Comput. Vis. Image Underst. **189** (2019)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of the IEEE conf. on CVPR. pp. 770–778 (2016)
10. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: Layercam: Exploring hierarchical class activation maps for localization. IEEE Trans. on IP **30**, 5875–5888 (2021)
11. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proc. IEEE/CVF winter conference on applications of computer vision. pp. 1548–1558 (2021)
12. Krishnan, A., Almadan, A., Rattani, A.: Understanding fairness of gender classification algorithms across gender-race groups. In: Proc. 19th IEEE ICMLA. pp. 1028–1035 (2020)
13. Krishnan, A., Rattani, A.: A novel approach for bias mitigation of gender classification algorithms using consistency regularization. Image Vis. Comput. **137**, 104793 (2023)
14. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear convolutional neural networks for fine-grained visual recognition. IEEE trans. on PAMI **40**(6), 1309–1322 (2017)
15. Lin, X., Kim, S., Joo, J.: Fairgrape: Fairness-aware gradient pruning method for face attribute classification. In: ECCV (13). vol. 13673, pp. 414–432 (2022)

16. Liu, D., Zhao, L., Wang, Y., Kato, J.: Learn from each other to classify better: Cross-layer mutual attention learning for fine-grained visual classification. Pattern Recognition **140**, 109550 (2023)
17. 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)
18. Liu, Z., Luo, P., Wang, X., Tang, X.: Large-scale celebfaces attributes (celeba) dataset. Retrieved August **15**(2018), 11 (2018)
19. Loshchilov, I., Hutter, F.S.: Stochastic gradient descent with warm restarts. 2016
20. Luo, W., Zhang, H., Li, J., Wei, X.S.: Learning semantically enhanced feature for fine-grained image classification. IEEE SPL **27**, 1545–1549 (2020)
21. Majumdar, P., Singh, R., Vatsa, M.: Attention aware debiasing for unbiased model prediction. In: Proc. IEEE/CVF ICCVW 2021. pp. 4116–4124 (2021)
22. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6) (jul 2021)
23. Muthukumar, V.: Color-theoretic experiments to understand unequal gender classification accuracy from face images. In: Proc. IEEE CVPR Workshops 2019. pp. 2286–2295 (2019)
24. Padala, M., Gujar, S.: Fnnc: Achieving fairness through neural networks. In: Proc. IJCAI-20. pp. 2277–2283. ijcai.org (2020)
25. Ramachandran, S., Rattani, A.: Deep generative views to mitigate gender classification bias across gender-race groups. In: Proc. ICPR 2022 Int. Workshops and Challenges. vol. 13645, pp. 551–569. Springer (2022)
26. Ramachandran, S., Rattani, A.: A self-supervised learning pipeline for demographically fair facial attribute classification. In: The IEEE International Joint Conference on Biometrics (IJCB). IEEE (2024)
27. Ramaswamy, V.V., Kim, S.S.Y., Russakovsky, O.: Fair attribute classification through latent space de-biasing. In: CVPR. pp. 9301–9310 (2021)
28. Wang, Q., Xie, J., Zuo, W., Zhang, L., Li, P.: Deep cnns meet global covariance pooling: Better representation and generalization. IEEE trans. on PAMI **43**(8), 2582–2597 (2020)
29. Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., Russakovsky, O.: Towards fairness in visual recognition: Effective strategies for bias mitigation. In: Proc. IEEE/CVF CVPR 2020. pp. 8916–8925 (2020)
30. Wick, M.L., Panda, S., Tristan, J.: Unlocking fairness: a trade-off revisited. In: Adv. Neural Inf. Process. Syst. 32 (NeurIPS 2019). pp. 8780–8789 (2019)
31. Zafeiriou, S., Zhang, C., Zhang, Z.: A survey on face detection in the wild: Past, present and future. Comput. Vis. Image Underst. **138**, 1–24 (2015)
32. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV. pp. 818–833 (2014)
33. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proc. AAAI/ACM AIES 2018. pp. 335–340. ACM (2018)
34. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: Proc. IEEE Conf. on CVPR (July 2017)
35. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proc. of IEEE Conf. on CVPR. pp. 2921–2929 (2016)
36. Zietlow, D., Lohaus, M., Balakrishnan, G., Kleindessner, M., Locatello, F., Schölkopf, B., Russell, C.: Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In: Proc. IEEE/CVF Conf. on CVPR. pp. 10410–10421 (2022)