Fair Facial Attribute Recognition via Group-Decoupled Vision Transformer with Mask-Guided Correlation Suppression

Huichang Huang[1*], **Kunchi Li**[1*], **Si Chen**[1†], **Da-Han Wang**[1]

[1] Fujian Key Laboratory of Pattern Recognition and Image Understanding, School of Computer and Information Engineering, Xiamen University of Technology
chensi@xmut.edu.cn

Abstract

Facial Attribute Recognition (FAR) holds significant potential for wide-ranging applications. However, traditionally trained FAR models exhibit unfairness, largely due to data bias—where certain sensitive attributes correlate statistically with target attributes. To address this, we propose a group-attention mechanism: first, each image is categorized into subgroups (e.g., Male/Female&short hair, Male/Female&long hair). Within the attention mechanism, distinct Query parameters are used for each group, with shared Key and Value parameters. As group-specific Query parameters are trained on subgrouped data, the noted bias is effectively mitigated. Consequently, integrating this Group-Attention into Vision Transformer (ViT) yields our novel Group-Decoupled ViT (GD-ViT) model. Moreover, to further attenuate the statistical correlation between sensitive and target attributes, we propose a Mask-Guided Correlation Suppression learning strategy. Specifically, in Stage 1, it first leverages a min-max dual-loss optimization strategy to train GD-ViT in capturing key regions related to sensitive attributes yet irrelevant to target attributes. Then, in Stage 2, it trains another GD-ViT by masking sensitive regions identified in Stage 1, fusing the masked output (as intermediate input) with the model's intermediate outputs. This weakens regions associated with sensitive attributes while enhancing others, suppressing the learning of key features related to sensitive attributes. Consequently, it encourages the model to focus more on intrinsic target attribute regions and balances the learning process between the sensitive attribute and the target attribute. Extensive experiments demonstrate that our method achieves superior performance across three benchmark datasets for fair facial attribute recognition.

Introduction

Facial Attribute Recognition (FAR) facilitates the automatic analysis of facial features (e.g., age, gender, hairstyle) and has been extensively employed across diverse scenarios (Meden et al. 2021; Zhang et al. 2021a; Shu et al. 2021a), including public security, human-computer interaction, and commercial services. However, traditional FAR models are afflicted by unfairness and numerous approaches have been developed to address this issue (Serna et al. 2022; Jung et al. 2021; Zhang et al. 2021b; Ma et al. 2023; Gong, Liu, and Jain 2020). For instance, TADeT (Sudhakar et al. 2023) mitigates bias by explicitly eliminating prominent group-specific information from query matrices. Another approach, the Debiased Self-Attention (DSA) method (Qiang et al. 2024), employs an adversarial learning framework that utilizes an auxiliary bias model to generate debiased data, thereby diminishing the salience of sensitive attribute information. However, these methods overlook demographic group heterogeneity and fail to explicitly mitigate the influence of sensitive attributes.

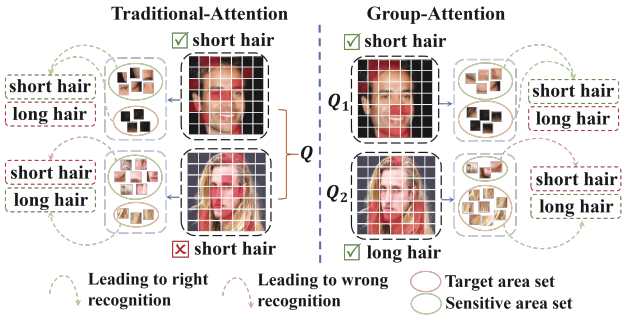

Figure 1: Toy illustration of Group-Attention versus the Traditional-Attention mechanisms.

The issue of unfairness primarily arises from data imbalance: the uneven distribution of samples across demographic groups causes models to learn spurious correlations between sensitive attributes (e.g., gender, ethnicity) and target attributes (Cui et al. 2024; Park et al. 2021; Quadrianto, Sharmanska, and Thomas 2019; Kong, Kim, and Kim 2025; Thong, Joniak, and Xiang 2023). For instance, as illustrated in the left subfigure of Fig. 1, due to the significant imbalance in general datasets—where images of males with short hair far outnumber those of males with long hair—models trained using the typical Vision Transformer (ViT) with a traditional attention mechanism tend to over-learn key regions associated with the sensitive attribute (gender: male) and erroneously correlate them with the target attribute (hair length), thereby impairing the performance in predicting the target attribute.

[*] These authors contributed equally.
[†] Corresponding author.

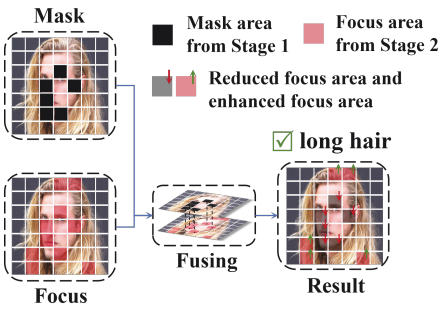

Figure 2: The strategy involves masking the output derived from the Stage 1 model and fusing it with the intermediate outputs of the Stage 2 model.

Based on the aforementioned analysis, we propose a Group-Attention mechanism (GA) with the objective of decoupling key features associated with sensitive attributes from target attributes. Specifically, each image is first categorized into subgroups (e.g., Male/Female & short hair, Male/Female & long hair). Within the attention mechanism, distinct Query parameters are employed for each group, while Key and Value parameters are shared across groups. Since group-specific Query parameters are trained on subgrouped data, the aforementioned bias is effectively alleviated (see the right subfigure of Fig. 1). Consequently, integrating this Group-Attention mechanism into ViT results in our novel Group-Decoupled ViT (GD-ViT) model.

Furthermore, given that mitigating spurious statistical associations between sensitive and target attributes is more conducive to alleviating the issue of unfairness, our objective is to suppress the learning of key regions that are critical to sensitive attributes yet irrelevant to target attributes. This enables a balanced learning process between the key regions of sensitive attributes and target attributes, thereby achieving enhanced performance. To this end, we propose a Mask-Guided Correlation Suppression learning strategy, which has two training stages. Specifically, it first trains a GD-ViT via a min-max dual-loss optimization strategy (Qiang et al. 2024): minimizing the sensitive attribute recognition loss and maximizing the target attribute recognition loss, which helps capture key regions specific to sensitive attributes yet unrelated to the target attribute.

Subsequently, in Stage 2, another GD-ViT is trained by masking the sensitive regions identified in Stage 1 and fusing the masked output—functioning as intermediate input—with the model's intermediate outputs (see Fig. 2). This process weakens regions associated with sensitive attributes while enhancing other regions, thereby suppressing the learning of key features pertaining to sensitive attributes. Consequently, it prompts the model to focus more intensively on intrinsic target attribute regions and balances the learning process between the two types of attributes. In summary, the highlights of this paper are as follows:

- We propose a Group-Attention (GA) mechanism that disentangles features correlated with sensitive attributes from those relevant to target attributes. By training group-specific Query parameters on subgrouped data, the model effectively suppresses bias. Incorporating GA into the Vision Transformer framework results in our proposed Group-Decoupled ViT (GD-ViT).
- We propose a Mask-Guided Correlation Suppression learning strategy that attenuates regions associated with sensitive attributes while amplifying others. This suppresses the acquisition of sensitive-related features, guiding the model to attend more to intrinsic target-relevant regions and promoting balanced learning between sensitive and target attributes.
- Extensive experiments across three benchmark datasets validate that our method (Fair-GDMS) achieves significant fairness improvements while preserving overall accuracy, outperforming state-of-the-art methods.

Related Work

Facial Attribute Recognition. Recent studies have addressed various challenges in facial attribute recognition (FAR). For instance, Kim et al. (Kim, Jain, and Liu 2022) introduced an adaptive loss function to handle recognition difficulties in low-quality face datasets. Another line of work focused on enhancing the robustness of FAR models, proposing an identity-sensitive conditional diffusion generation model with strong perturbation resistance to counter adversarial attacks (Liu et al. 2024). Additionally, Kollias et al. (Kollias, Sharmanska, and Zafeiriou 2024) develop a multi-task learning framework to jointly perform FAR and other recognition tasks, leveraging a distribution matching method to enable effective knowledge transfer. To mitigate annotation scarcity, Shu et al. (Shu et al. 2021b) proposed a spatial semantic block learning method to improve FAR performance under limited supervision. Despite these advancements, fairness remains a critical yet underexplored issue in FAR, particularly regarding biased representations and their impact on decision outcomes.

Fairness and Bias in FAR. To address fairness in facial attribute recognition (FAR), prior works have explored strategies such as re-weighted sampling(Park et al. 2022), metric learning(Park et al. 2022), image generation(Zhang et al. 2024), and multimodal fusion(Luo et al. 2024). While re-weighting highlights minority groups, it may harm representation quality. Metric learning enforces fairer feature distances but is noise-sensitive and supervision-dependent. Image generation improves group balance but risks bias from poor synthesis quality. Multimodal methods enhance semantics via auxiliary text, yet increase complexity and data requirements.

Recently, numerous ViT-based approaches have been proposed to mitigate unfairness. For instance, Fair-VPT (Park and Byun 2024) employs learnable prompts and contrastive loss to suppress sensitive attributes; Fair-ViT (Tian, Du, and Shen 2024) introduces adaptive group-specific masks and regularization to reduce bias at minimal cost. In contrast, DSA (Qiang et al. 2024) adopts a two-stage design, utilizing biased-model-generated masks to guide fair model training. However, these methods largely overlook demographic

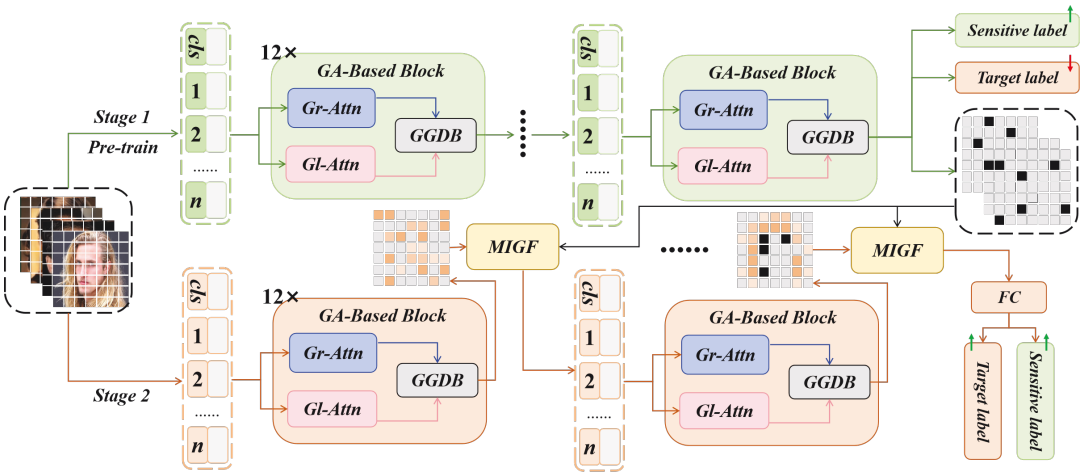

Figure 3: Illustration of the Fair-GDMS framework. Gr-Attn, Gl-Attn, and GGDB represent Group-Attention (Multi-head), Global-Attention (Multi-head), and Group-Global Dynamic Balance, respectively. These three modules make up the GA-Based Block (see Fig. 4). MIGF (**M**utual **I**nformation-**G**uided **F**usion, see Fig. 5) is the fusing strategy used in Stage 2.

group heterogeneity and fail to explicitly mitigate the influence of sensitive attributes. To address these limitations, we propose Fair-GDMS, which explicitly decouples group-specific representations and suppresses bias induced by sensitive features.

Methodology

Overview

As shown in Fig. 3, Fair-GDMS utilizes GD-ViT and incorporates Mask-Guided Correlation Suppression Learning through a two-stage training process: Stage 1 utilizes a min-max dual-loss optimization strategy to guide the model in identifying key regions associated with sensitive attributes yet irrelevant to the target attributes, while Stage 2 masks these identified regions and fuses the masked outputs with intermediate features, thereby suppressing sensitive signals and enhancing target-focused learning. In the following, we will first introduce the proposed Group-Attention Mechanism (GA), followed by a detailed explanation of the Mask-Guided Correlation Suppression Learning strategy.

Group-Attention Mechanism (GA)

As mentioned earlier, data bias often induces spurious correlations between a sample's sensitive and target attributes. To decouple key sensitive regions from target attributes, we propose a Group-Attention (GA) Mechanism. Specifically, GA first categorizes sensitive-target attribute combinations into four types: (1) TY-SY (Target-Yes, Sensitive-Yes; e.g., long hair&Female), (2) TN-SN (Target-No, Sensitive-No; e.g., short hair&Male), (3) TY-SN (Target-Yes, Sensitive-No; e.g., long hair&Male), and (4) TN-SY (Target-No, Sensitive-Yes; e.g., short hair&Female). A two-layer fully connected network is pre-trained to classify images into these groups.

However, group sample sizes are highly imbalanced—some have very few samples, others far more. For instance, long-haired females far outnumber long-haired males. Such underrepresentation compromises the

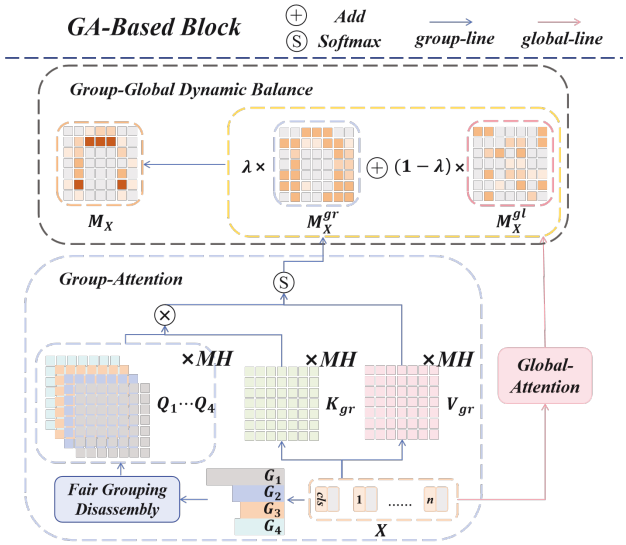

Figure 4: Illustration of the GA and GA-Based Block. MH denotes the number of heads in multi-head attention.

training of group-specific modules. To enrich minority groups, we reassign samples from majority to corresponding minority groups (sharing the same target attributes) with low probability (0.05 in this work)—for example, randomly reassigning long-haired female samples to the long-haired male group.

Next, we assign distinct group-specific query parameters while sharing key and value parameters across groups. Under this configuration, the group attention (GA) is computed as follows:

$$\text{Attn}_g = \text{Softmax}\left(\frac{Q_g K_{gr}^\top}{\sqrt{d_k}}\right), \tag{1}$$

$$M_X^{gr} = \text{Attn}_{g_x} \cdot V_{gr}, \tag{2}$$

where d_k is the dimension of key vectors, g_x means the group label of the sample X. M_X^{gr} denotes the Group-specific attention Map for X, and K_{gr}, V_{gr} are the shared key and value matrices used in the group attention. Through this grouping strategy, each group-specific query parameter is trained on sub-grouped data.

GA-Based Block. After obtaining the group-specific attention map, we combine it—as supplementary information to achieve balanced learning—with the global attention map derived from the traditional attention mechanism (referred to as "global attention" to distinguish it from group attention in this work) (see Fig. 4). Global attention Map for X can be denoted as M_X^{gl}, and it is expressed as follows:

$$M_X^{gl} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{3}$$

where Q, K, and V denote Query, Key, and Value parameters in the global attention.

Furthermore, to integrate both perspectives, we propose a Group-Global Dynamic Balancing module (GGDB, see Fig. 4) that adaptively fuses the two components, which is formulated as follows:

$$M_X = \lambda \cdot M_X^{gr} + (1-\lambda) \cdot M_X^{gl}, \tag{4}$$

where λ is a gating factor (which can be chosen between a learnable parameter gate init (gi) and a manually set parameter (msp)) that adaptively and dynamically balances group-specific and globally shared information, resulting in a fused feature M_X. Compared to using only group-specific or global features, M_X effectively captures both personalized group characteristics and shared global semantics.

In summary, the GA-based block (multi-head attention with 12 heads) generates group-specific attention maps and dynamically integrates them with global attention maps by GGDB. Though the group-specific maps are supplementary, their query parameters—trained on sub-grouped data—effectively alleviate bias. Combined with GGDB's dynamic integration, this achieves balanced training. Building on this, we incorporate the GA-based block into the Vision Transformer (ViT), resulting in the GD-ViT.

Mask-Guided Correlation Suppression Learning

Moreover, mitigating spurious correlations between sensitive and target attributes is critical for fairness; thus, our goal is to suppress learning key regions predictive of sensitive attributes but irrelevant to the target. Reducing the influence of such regions enables more balanced representation learning between sensitive and target attribute key areas, boosting overall model performance. To this end, we propose a Mask-Guided Correlation Suppression learning strategy, implemented in two stages.

Stage 1: Capturing and Masking Key Sensitive Regions. In Stage 1, we use a min-max dual-loss optimization strategy (Qiang et al. 2024) to train a GD-ViT, guiding the model to highlight the sensitive-attribute regions but irrelevant to the target. This strategy, via combined loss terms, enables the model to generate masks filtering out spurious, non-target information tied to sensitive attributes.

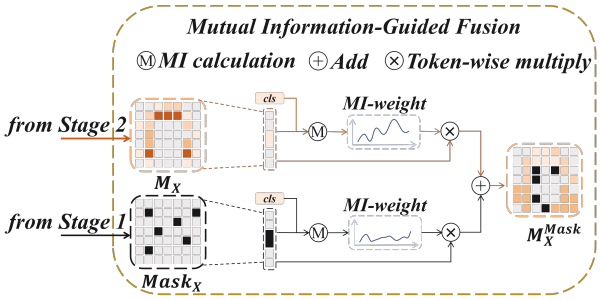

Figure 5: Illustration of the MIGF. M_X represents the intermediate output map of sample X in stage 2 and $Mask_X$ denotes the masked key region derived from stage 1, both cls tokens are identical and originate from Stage 2. The final map, M_X^{Mask} is obtained through MIGF.

Specifically, for an input sample, features F_X extracted through GD-ViT's 12-layer attention module are fed to the sensitive attribute prediction head f_s and target attribute prediction head f_t for classification. To focus on regions relevant to sensitive but not target attributes, we minimize binary cross-entropy loss for f_s and maximize it for f_t (denoted as $\mathcal{L}_\text{S}$ and $\mathcal{L}_\text{T}$, respectively). Accordingly, the overall loss in the min-max dual-loss optimization strategy, which is used in Stage 1 (denoted as $\mathcal{L}_{s1}$) is formulated as follows:

$$\mathcal{L}_{s1} = \mathcal{L}_\text{S}(f_s(F_X), s) - \mathcal{L}_\text{T}(f_t(F_X), y), \tag{5}$$

where s and y denote the ground-truth labels of the sensitive and target attributes, respectively. This objective encourages accurate sensitive attribute prediction (via minimizing $\mathcal{L}_\text{S}$) while discouraging correct target prediction (via maximizing $\mathcal{L}_\text{T}$), steering GD-ViT to focus on sensitive-but-non-target key regions.

Finally, we sort attention scores across GD-ViT layers, identify consistently high-attention regions as sensitive-but-non-target key regions, mask them, and pass the masked features—alongside original features—to Stage 2 for Correlation Suppression Learning. In the specific implementation, a threshold is set to filter key regions for mask processing.

Stage 2: Correlation Suppression Learning. In Stage 2, we train a second GD-ViT by fusing the model's intermediate outputs with the masked feature maps from Stage 1. For the fusing strategy, our goal is to direct the model to focus on regions informative for target attributes yet irrelevant to sensitive attributes. To this end, we propose a (**M**utual **I**nformation-**G**uided **F**usion (MIGF) strategy, which leverages mutual information to guide the fusing process.

Specifically, since classification inference relies on the cls token, MIGF computes the portion of the attention map that is more relevant to the cls token, derive a score from this portion, and use it to fuse M_X and $Mask_X$ (see Fig. 5). We use mutual information as this score, which is estimated following the setup of MINE (Ishmael Belghazi et al. 2018). First, M_X and $Mask_X$ are flattened, and then the two mu-

tual information values are computed as follows:

$$MI_X = MI\left(M_X; \text{CLS}_X\right) = T_\theta(M_X, \text{CLS}_X) - \log\left(\exp\left(T_\theta\left(M_X, \text{CLS}_{\pi(X)}\right)\right) + \epsilon\right), \quad (6)$$

$$MI_{Mask} = MI\left(Mask_X; \text{CLS}_X\right) = T_\theta(Mask_X, \text{CLS}_X) - \log\left(\exp\left(T_\theta\left(Mask_X, \text{CLS}_{\pi(X)}\right)\right) + \epsilon\right), \quad (7)$$

where T_θ denotes a trainable three-layer fully connected network, acting as the model for the mutual information scoring function, and ϵ is a small value employed to prevent numerical overflow. The first term calculates the mutual information score between the M_X/$Mask_X$ and the cls token CLS_X of X. The second term quantifies the mutual information between M_X/$Mask_X$ and the cls token of a sample randomly selected from the current mini-batch, denoted as $\text{CLS}_{\pi(X)}$ (which represents a cls token computed from a sample other than X). The subtraction operation enforces that the weighted fusion result maintains strong correlation with the correct cls token while minimizing correlation with the incorrect cls token.

Furthermore, within two groups sharing the same target attribute (e.g., long hair&Male vs. long hair&Female), the mutual information between M_X and their respective cls tokens CLS_X in these two groups should be as similar as possible, given that they share the same target attribute. To this end, we introduce a regularization term to realize this as follows:

$$R_X = \left|MI(M_X; CLS_X) - MI(M_{X'}; CLS_{X'})\right|. \quad (8)$$

In the specific implementation, for each sample X in the mini-batch, we randomly select a sample X' from another distinct group that shares the same target attribute, and use it to compute R_X. Subsequently, we combine MI_X with R_X, resulting in the final score as follows:

$$MIR_X = MI_X + \alpha \cdot R_X, \quad (9)$$

where α is a learnable parameter that balances informativeness and fairness. Consequently, fused features are generated as:

$$M_X^{Mask} = MIR_X \cdot M_X + MI_{Mask} \cdot Mask_X. \quad (10)$$

Next, the fused maps are fed to the sensitive attribute prediction head and the target attribute prediction head (still denoted as f_s and f_t) for classification. The loss used in Stage 2 ($\mathcal{L}_{s2}$) is defined as:

$$\mathcal{L}_{s2} = \mathcal{L}_\text{S}(f_s(M_X^{Mask}), s) + \mathcal{L}_\text{T}(f_t(M_X^{Mask}), y), \quad (11)$$

where $\mathcal{L}_\text{S}$ and $\mathcal{L}_\text{T}$ denote the binary cross-entropy losses for sensitive and target attribute predictions, respectively.

Experiment

In this section, we conduct a comprehensive comparison with recent models for fairness tasks on three datasets under the setting of various target-sensitive attribute groups, and on this basis demonstrate the SOTA performance of our model for the fair facial attribute recognition task.

Experimental Setup

Details. We selected three benchmark datasets: CelebA (Liu et al. 2015), UTKFace (Zhang, Song, and Qi 2017), and bFFHQ (Sagawa et al. 2019). Each contains facial images annotated with target and sensitive attributes. UTKFace and bFFHQ feature imbalanced training sets and balanced test sets to comprehensively evaluate model fairness and performance.

Similar to the classic ViT-B/16, the model accepts an image input size of 224×224, 1 classification token patch, 12 layers of Transformer encoder block. The GD-ViT used in both stages only sets two classification heads, the sensitive attribute target head and the target attribute target head. For the Male/Hair Color task, we use the standard enhancement used in Sia (Tao et al. 2023), with an initial learning rate of 5e-5, a learning rate decay factor of 0.65, and a weight decay factor of 0.65 to prevent overfitting. We also use a method of gradually increasing the learning rate and set the warmup-epoch with the highest learning rate to 25. The model are executed on a machine equipped with two NVIDIA GeForce RTX 3090 GPUs, running the Ubuntu 20.04 operating system. In addition to the accuracy Acc, the following fairness evaluation metrics are also used in the experiment:

Balanced Accuracy (BA) (Park et al. 2020) is an accuracy measure for the class imbalance problem. Its expression formula is as follows:

$$\text{BA} = \frac{1}{4}(\text{TPR}_a + \text{TNR}_a + \text{TPR}_{na} + \text{TNR}_{na}), \quad (12)$$

where TPR measures the true positive rate, while TNR measures the true negative rate in classification performance, a and na mean different sensitivity groups (e.g., male and female).

BA gives equal weight to positive and negative samples, which is especially suitable for scenarios with a large number of minority classes. It is more fair than traditional accuracy.

Equalized Odds (EOD) (Hardt, Price, and Srebro 2016) measures whether the positive decisions made by the model in different sensitive attribute groups are consistent. The expression formula is as follows:

$$\text{EOD} = \frac{1}{2}\left(|\text{TPR}_a - \text{TPR}_{na}| + |\text{FPR}_a - \text{FPR}_{na}|\right), \quad (13)$$

where FPR is the proportion of actual negatives incorrectly identified as positives. EOD reflects the fairness of the model's predictions for different groups. The smaller it is, the fairer it is.

Demographic Parity (DP) (Dwork et al. 2012) measures the difference in the probability of positive predictions (e.g., "attractive") between different demographic groups defined by a sensitive attribute A. The formal definition is:

$$\text{DP} = \mathbf{P}(\hat{Y} = 1 \mid A = 1) - \mathbf{P}(\hat{Y} = 1 \mid A = 0), \quad (14)$$

where $\hat{Y}$ denotes the model's predicted class (e.g., $\hat{Y} = 1$ for a positive prediction), and A represents a binary sensitive

Dataset	SA / TA	Method	Venue	Acc (%) ↑	BA (%) ↑	EOD ↓	DP ↓
CelebA	Male / Hair Color	DSA	[ECCV'24]	<u>90.9</u>	<u>82.9</u>	<u>0.259</u>	<u>0.2337</u>
		Fair-GDMS	[Ours]	**92.3**	**84.4**	**0.117**	**0.1023**
	Male / Attractive	ViT	[ICLR'20]	78.4	68.7	0.416	–
		FairCL	[ICLR'23]	75.3	–	0.168	–
		VPT+FSCL+	[CVPR'22]	69.3	66.5	0.206	–
		Fair-VPT	[CVPR'24]	78.6	76.3	<u>0.120</u>	–
		Fair-ViT	[ECCV'24]	**84.0**	<u>79.9</u>	–	<u>0.2837</u>
		Fair-GDMS	[Ours]	<u>83.2</u>	**80.5**	**0.070**	**0.1976**
	Male / Big Nose	ViT	[ICLR'20]	81.7	61.3	0.306	–
		FairCL	[ICLR'23]	80.0	–	<u>0.142</u>	–
		VPT+FSCL+	[CVPR'22]	<u>84.6</u>	63.6	0.251	–
		Fair-VPT	[CVPR'24]	79.9	<u>65.4</u>	0.159	–
		Fair-GDMS	[Ours]	**84.8**	**71.2**	**0.132**	–
UTK Face	Gender / Race	ViT	[ICLR'20]	–	88.4	0.134	–
		VPT+FSCL+	[CVPR'22]	–	89.0	0.099	–
		Fair-VPT	[CVPR'24]	–	<u>90.9</u>	<u>0.049</u>	–
		Fair-GDMS	[Ours]	–	**91.7**	**0.031**	–
bFFHQ	Gender / Age	ViT	[ICLR'20]	–	74.8	0.489	0.3410
		Fair-VPT	[CVPR'24]	–	<u>80.7</u>	0.371	–
		DSA	[ECCV'24]	–	78.8	<u>0.265</u>	0.2856
		Fair-GDMS	[Ours]	–	**80.9**	**0.215**	**0.2245**

Table 1: Experimental results on CelebA, UTK Face, and bFFHQ. TA and SA represent the target attribute and the sensitive attribute, respectively. The **bold** values show the best results and the second-best values are <u>underlined</u>.

attribute (e.g., gender or race). DP requires that the probability of receiving a positive prediction should be equal across groups, ensuring the model does not favor or discriminate against any group based on A.

Comparison on CelebA

Tab. 1 presents results on the CelebA dataset under multiple group settings and fairness metrics. In the Male/Hair Color task, our method consistently outperforms DSA in all metrics (ACC, BA, EOD, and DP), showing superior accuracy and fairness. For the Male/Attractive task, our method achieves comparable accuracy to Fair-ViT (83.2% vs. 84.0%) while surpassing the second-best results in fairness metrics, with an EOD of 0.05 (0.070 vs. 0.120) and a DP of 0.0861 (0.1976 vs. 0.2837). In the Male/Big Nose task, although VPT+FSCL slightly improves accuracy, it offers minimal fairness enhancement. Other methods improve EOD and BA but sacrifice accuracy. In contrast, our model achieves the best overall performance across all three metrics: Acc, EOD, and BA (84.8%, 71.2%, and 0.132).

Comparison on UTK Face

Following the protocol of previous works, we construct a completely balanced validation and test set and report the results in Tab. 1. Due to the balanced nature of these sets, the values of ACC and BA are identical; therefore, we focus our comparison on BA and EOD only. Among the baseline methods, Fair-VPT demonstrates the strongest overall performance across both metrics. However, our method outperforms it with improvements exceeding 0.8% in BA and 0.018 in EOD, highlighting the effectiveness of our approach even under perfectly balanced evaluation settings.

Comparison on bFFHQ

As shown in Tab. 1, we conduct experimental comparison on bFFHQ. Fair-GDMS achieves the highest balanced accuracy (80.9%) among all compared approaches. Compared to the baseline ViT and the fairness-enhancing methods Fair-VPT and DSA, our model significantly reduces both DP and EOD, reaching the lowest values of 0.2245 and 0.2147, respectively. These results show that our method improves recognition performance while achieving superior fairness by effectively mitigating sensitive attribute bias.

Ablation Study

We perform ablation studies to validate each component (see Tab. 2). Specifically, ViT denotes the vanilla Vision Transformer baseline; GD-ViT corresponds to our GD-ViT with GA; Fair-GDMS represents the entire framework.

To investigate the effect of the parameter α for R_i in MIGF, we report the results in Tab. 3. As α increases, the fairness metrics initially improve but degrade when α becomes excessively large, likely due to the suppression of discriminative features. The optimal trade-off between accuracy and fairness is observed when $\alpha = 0.5$.

We also compare fusion strategies within the GGDB module, evaluating a learnable gating mechanism against manually set fusion weights. Tab. 3 shows that the gating-based

Model	CelebA (Male / Hair Color)		
	Acc (%)↑	BA (%)↑	EOD↓
ViT	90.2	81.8	0.2763
GD-ViT	91.7	83.6	0.1881
Fair-GDMS	**92.3**	**84.4**	**0.1173**

Table 2: Ablation study of Fair-GDMS. Comparison across CelebA under different fairness metrics. The **bold** values show the best results.

Items	Male / Hair Color			
	Acc (%)↑	BA (%)↑	EOD↓	DP↓
$\alpha = 0.7$	<u>91.9</u>	<u>83.8</u>	0.1431	0.1286
$\alpha = 0.5$	**92.3**	**84.4**	<u>0.1173</u>	**0.1023**
$\alpha = 0.2$	91.8	83.7	**0.1121**	<u>0.1056</u>
$\alpha = 0.1$	91.6	83.5	0.1367	0.1304
$gi = 0.9$	**92.3**	**84.4**	**0.1173**	**0.1023**
$gi = 0.5$	<u>91.4</u>	<u>83.0</u>	0.1704	<u>0.1441</u>
$msp = 0.9$	90.9	82.6	<u>0.1653</u>	0.1447
$msp = 0.5$	90.6	82.3	0.2037	0.1901
$rp = 0.5$	81.3	73.6	0.4611	0.3078
$rp = 0.1$	<u>91.6</u>	<u>83.4</u>	0.1574	0.1496
$rp = 0.05$	**92.3**	**84.4**	**0.1173**	**0.1023**
$rp = 0.01$	91.5	83.0	<u>0.1383</u>	<u>0.1294</u>

Table 3: Impact of items in MIGF (α) and GGDB (learnable parameter gate init (gi) and manually set parameter (msp)) and random probability (rp) in GA. The **bold** values show the best results and the second-best values are <u>underlined</u>.

strategy yields more stable and balanced performance. Consequently, we adopt this approach in the final model, with the gating parameter initialized to $gi = 0.9$.

Then, we conduct experiments to determine the optimal probability for randomly assigning majority samples to minority groups in GA. As shown in Tab. 3, setting the probability too low fails to sufficiently enrich the minority group samples, leading to suboptimal performance. Conversely, an excessively high probability introduces numerous spurious features, which degrades the model's performance.

To further illustrate the effectiveness of our method, we conduct attention visualization experiments. As shown in Fig. 6, compared with traditional self-attention, Fair-GDMS focuses more on semantically relevant and non-sensitive target regions, while traditional attention tends to highlight sensitive attribute areas. This confirms that our group-decoupled mechanism effectively suppresses bias-related cues and promotes fairer representation learning.

Finally, Fig. 7 compares Stage 1 with and without the target-loss maximization (i.e., the second term of Eq. 5). When only the sensitive attribute loss is used, the model tends to mask regions corresponding to hair, which are indicative of hair color. However, after incorporating the target-loss maximization, the model reduces its focus on the hair and instead attends more to regions unrelated to hair.

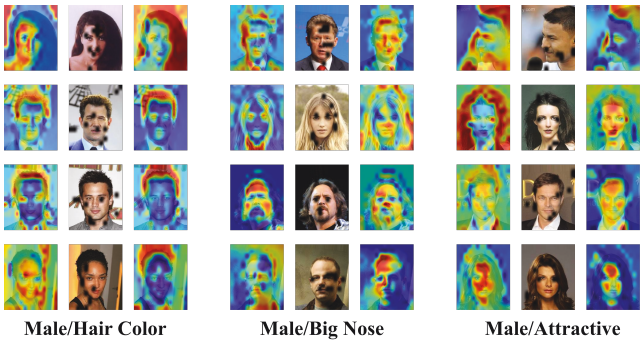

Figure 6: Visualizations for interpretability studies. For each task, we show the attention map of the traditional method, the key regions masked in Stage 1, and the attention map of our Fair-GDMS.

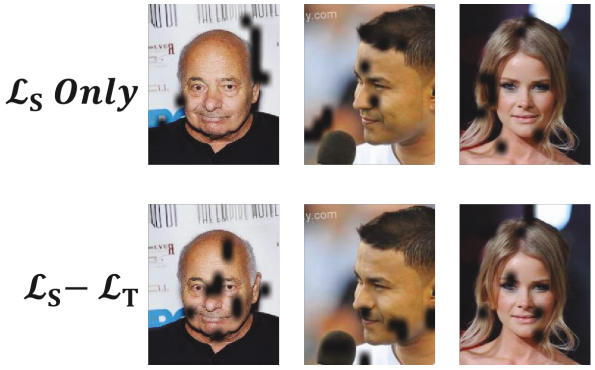

Figure 7: Visual explanation of whether to use $\mathcal{L}_T$ in Stage 1. The black regions indicate the masked portions of the sensitive key regions in Stage 1 for the GD-ViT.

Conclusion

In this work, we propose a Group-Attention (GA) mechanism to decouple key sensitive regions from target attributes. First, GA categorizes sensitive-target attribute combinations into subtypes and classifies images into these groups. It then employs distinct Query parameters for each group (with shared Key and Value parameters), and these group-specific Query parameters are trained on subgrouped data. Our experiments validate that this strategy effectively mitigates bias. We further integrate GA with the GGDB to form the GA-based block, whose dynamic integration via GGDB promotes balanced training. Building on this, we incorporate the GA-based block into ViT, yielding GD-ViT. Furthermore, we introduce a two-stage Mask-Guided Correlation Suppression Learning strategy. This strategy directs the model to focus more on intrinsic target attribute regions, balances learning between the two attribute types, and successfully mitigates spurious correlations between sensitive and target attributes. Building on these improvements, extensive experiments across three benchmark datasets show that our method achieves superior performance, while ablative studies validate the effectiveness of the proposed components.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62571466, 62372388, and 62576301, in part by the Natural Science Foundation of Xiamen under Grant 3502Z202573073, and in part by the Major Science and Technology Plan Project on the Future Industry Fields of Xiamen City under Grant 3502Z20241027.

References

Cui, J.; Zhu, B.; Wen, X.; Qi, X.; Yu, B.; and Zhang, H. 2024. Classes are not equal: An empirical study on image recognition fairness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23283–23292.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Gong, S.; Liu, X.; and Jain, A. K. 2020. Jointly debiasing face recognition and demographic attribute estimation. In *european conference on computer vision*, 330–347. Springer.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Ishmael Belghazi, M.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Devon Hjelm, R. 2018. MINE: mutual information neural estimation. *arXiv e-prints*, arXiv–1801.

Jung, S.; Lee, D.; Park, T.; and Moon, T. 2021. Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12115–12124.

Kim, M.; Jain, A. K.; and Liu, X. 2022. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18750–18759.

Kollias, D.; Sharmanska, V.; and Zafeiriou, S. 2024. Distribution matching for multi-task learning of classification tasks: a large-scale study on faces & beyond. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 2813–2821.

Kong, I.; Kim, K.; and Kim, Y. 2025. Fair representation learning for continuous sensitive attributes using expectation of integral probability metrics. *IEEE transactions on pattern analysis and machine intelligence*.

Liu, D.; Wang, X.; Peng, C.; Wang, N.; Hu, R.; and Gao, X. 2024. Adv-diffusion: imperceptible adversarial face identity attack via latent diffusion model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 3585–3593.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Luo, Y.; Shi, M.; Khan, M. O.; Afzal, M. M.; Huang, H.; Yuan, S.; Tian, Y.; Song, L.; Kouhana, A.; Elze, T.; et al. 2024. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12289–12301.

Ma, J.; Yue, Z.; Tomoyuki, K.; Tomoki, S.; Jayashree, K.; Pranata, S.; and Zhang, H. 2023. Invariant feature regularization for fair face recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 20861–20870.

Meden, B.; Rot, P.; Terhörst, P.; Damer, N.; Kuijper, A.; Scheirer, W. J.; Ross, A.; Peer, P.; and Štruc, V. 2021. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on information forensics and security*, 16: 4147–4183.

Park, S.; and Byun, H. 2024. Fair-vpt: Fair visual prompt tuning for image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12268–12278.

Park, S.; Hwang, S.; Kim, D.; and Byun, H. 2021. Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 2403–2411.

Park, S.; Kim, D.; Hwang, S.; and Byun, H. 2020. Readme: Representation learning by fairness-aware disentangling method. *arXiv preprint arXiv:2007.03775*.

Park, S.; Lee, J.; Lee, P.; Hwang, S.; Kim, D.; and Byun, H. 2022. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10389–10398.

Qiang, Y.; Li, C.; Khanduri, P.; and Zhu, D. 2024. Fairness-aware vision transformer via debiased self-attention. In *Proceedings of the european conference on computer vision*, 358–376. Springer.

Quadrianto, N.; Sharmanska, V.; and Thomas, O. 2019. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8227–8236.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Serna, I.; Morales, A.; Fierrez, J.; and Obradovich, N. 2022. Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *artificial intelligence*, 305: 103682.

Shu, Y.; Yan, Y.; Chen, S.; Xue, J.-H.; Shen, C.; and Wang, H. 2021a. Learning spatial-semantic relationship for facial attribute recognition with limited labeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11916–11925.

Shu, Y.; Yan, Y.; Chen, S.; Xue, J.-H.; Shen, C.; and Wang, H. 2021b. Learning spatial-semantic relationship for facial attribute recognition with limited labeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11916–11925.

Sudhakar, S.; Prabhu, V.; Krishnakumar, A.; and Hoffman, J. 2023. Mitigating bias in visual transformers via targeted alignment. *arXiv preprint arXiv:2302.04358*.

Tao, C.; Zhu, X.; Su, W.; Huang, G.; Li, B.; Zhou, J.; Qiao, Y.; Wang, X.; and Dai, J. 2023. Siamese image modeling for self-supervised vision representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2132–2141.

Thong, W.; Joniak, P.; and Xiang, A. 2023. Beyond skin tone: A multidimensional measure of apparent skin color. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4903–4913.

Tian, B.; Du, R.; and Shen, Y. 2024. FairViT: Fair Vision Transformer via Adaptive Masking. In *Proceedings of the european conference on computer vision*, 451–466. Springer.

Zhang, F.; He, Q.; Kuang, K.; Liu, J.; Chen, L.; Wu, C.; Xiao, J.; and Zhang, H. 2024. Distributionally generative augmentation for fair facial attribute classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22797–22808.

Zhang, W.; Ji, X.; Chen, K.; Ding, Y.; and Fan, C. 2021a. Learning a facial expression embedding disentangled from identity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6759–6768.

Zhang, Y.; Wei, X.-S.; Zhou, B.; and Wu, J. 2021b. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 3447–3455.

Zhang, Z.; Song, Y.; and Qi, H. 2017. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5810–5818.