Fair-VPT: Fair Visual Prompt Tuning for Image Classification

Sungho Park*
Yonsei University
Republic of Korea
qkrtjdgh18@yonsei.ac.kr

Hyeran Byun*
Yonsei University
Republic of Korea
hrbyun@yonsei.ac.kr

Abstract

Despite the remarkable success of Vision Transformers (ViT) across diverse fields in computer vision, they have a clear drawback of expensive adaption cost for downstream tasks due to the increased scale. To address this, Visual Prompt Tuning (VPT) incorporates learnable parameters in the input space of ViT. While freezing the ViT backbone and tuning only the prompts, it exhibits superior performances to full fine-tuning. However, despite the outstanding advantage, we point out that VPT may lead to serious unfairness in downstream classification. Initially, we investigate the causes of unfairness in VPT, identifying the biasedly pre-trained ViT as a principal factor. Motivated by this observation, we propose a Fair Visual Prompt Tuning (Fair-VPT) which removes biased information in the pre-trained ViT while adapting it to downstream classification tasks. To this end, we categorize prompts into "cleaner prompts" and "target prompts". Based on this, we encode the class token in two different ways by either masking or not masking the target prompts in the self-attention process. These encoded tokens are trained with distinct objective functions, resulting in the inclusion of different information in the target and cleaner prompts. Moreover, we introduce a disentanglement loss based on contrastive learning to further decorrelate them. In experiments across diverse benchmarks, the proposed method demonstrates the most superior performance in terms of balanced classification accuracy and fairness.

1. Introduction

Pre-trained language models [5, 6, 13, 44, 54] based on self-attention mechanisms have achieved immense success in the field of Natural Language Processing (NLP), attributed to their remarkable efficiency and capability to handle large-scale datasets. In light of the accomplishments in NLP, many studies [14, 17, 36, 53, 57] have tried to alternative conventional Convolutional Neural Networks (CNN) [23] to self-attention-based architectures (*e.g.,* Transformer [54] and BERT [13]) across diverse tasks of computer vision. Notably, Vision Transformer (ViT) [14] has demonstrated outstanding performance and versatility, leading to its adoption in various fields, including image classification [47], semantic segmentation [33], and image captioning [10].

Method	TA	SA		Acc. (↑)	EO (↓)
		M	*F*		
VPT [24]	*A*	52.7	93.1	**81.7**	**32.1**
	NA	89.1	65.2		
VPT [24]-Head+NCM [39]	*A*	68.1	84.6	76.1	47.4
	NA	99.4	21.0		
ViT [14]+NCM [39]	*A*	20.6	92.3	69.4	82.3
	NA	99.6	6.7		

Table 1. **Exploring primary factors of unfairness.** This reports classification accuracy (Acc.) and equalized odds (EO) [21] on CelebA [35]. The target attribute (TA) corresponds to whether inputs are attractive (*A*) or not (*NA*) and the sensitive attribute (SA) is set to *Gender* (abbreviated as *M* for male and *F* for female). "Head" and "NCM" denote the classification head and Nearst Class Mean classifier [39], respectively. The scores reported in the third and fourth columns represent the classification accuracy of the respective groups. The results indicate that the pre-trained ViT is the key factor contributing to the unfairness in VPT.

Nonetheless, a notable limitation of ViT is the expensive adaptation cost for downstream tasks. Due to the increased scale compared to conventional CNNs, full fine-tuning for each task becomes cost-prohibitive and occasionally inefficient [24]. To address this, Visual Prompt Tuning (VPT) was proposed [24], which is an efficient approach for adapting the pre-trained transformer models to downstream computer vision tasks. It prepends a small portion of learnable parameters (*i.e.,* prompt) to the input space for each task, resulting in effective adaptation while keeping the ViT backbones frozen. In many cases, it has demonstrated superior performance compared to full fine-tuning. However, we point out that VPT may cause unfairness issues in classification tasks.

As deep learning models deployed in real-world applica-

*Corresponding authors with equal contribution.

tions have raised controversial ethical problems, the ethics of AI have been considered as crucial as their performances [4, 7, 56]. For instance, COMPAS system, utilized in the US court, judged black defendants to be more likely to recidivate than white defendants [1], and Google Photos incorrectly identified some black people as gorillas [15]. To deal with the issues, researchers have defined various notions of fairness [16, 21, 28] and tried to ensure it in terms of sensitive attributes which are characteristics that should not be discriminated against, such as gender, ethnicity, and region [25, 29, 42, 46].

In this context, we initially evaluate the fairness of VPT on the benchmark dataset (i.*e.,* CelebA [35]) and analyze the underlying causes of observed unfairness. In Table 1, we compare three different methods by measuring classification accuracy (Acc.) for the target attribute *Attractive* and fairness with equalized odds (EO) [21] for the sensitive attribute *Gender*. It is devised to investigate the extent of unfairness arising from three major components of VPT: a pre-trained ViT, learnable prompts, and a classification head. First, VPT exhibits significantly unfair performances, particularly favoring majority groups (*i.e., Attractive-Female* and *Not Attractive-Male*). Subsequently, we replace the classification head with Nearest Class Mean (NCM) classifiers [39]. The deteriorated fairness implies that the classification head partially alleviates unfairness. Lastly, ViT+NCM, without any training for the downstream task, shows the most unfair results. This indicates that biased information pertaining to the sensitive attribute within the pre-trained ViT stands out as a primary factor of the unfairness arising from VPT.

Therefore, we propose a novel method, namely Fair Visual Prompt Tuning (Fair-VPT), which removes biased information related to sensitive attributes in the pre-trained ViT model while adapting it to downstream classification tasks. Inheriting the philosophy of prompt tuning, we first add learnable parameters, known as prompts, in the input space. The key idea is to select certain prompts as "cleaner prompts", which are encoded to contain biased information from the pre-trained model. In contrast, the remaining prompts, referred to as "target prompts", are encoded to learn only target features that are not correlated with the sensitive attribute. To this end, we encode the class token in dual parallel manners. One method involves encoding the class token by masking the target prompts in the self-attention process, while the other method encodes it with all the prompts and image patches following the original mechanism [24]. Subsequently, the two types of encoded class tokens are respectively encouraged to predict sensitive and target attributes through classification heads. Consequently, most of the information related to the sensitive attribute is included within the cleaner prompts through the masked self-attention process, while the rest information pertaining to the target attribute is learned by the target prompts. Moreover, we introduce a disentanglement loss based on contrastive loss to explicitly mitigate the correlation between the prompts. During inference, fairness can be ameliorated by excluding the cleaner prompts for target classification.

In the experiment section, we conduct extensive validation on several benchmark datasets, *i.e.,* CelebA [35], UTK Face [60], bFFHQ [30], and Waterbirds [48]. In all the experiments, the proposed method markedly enhances the fairness of ViT and achieves the most superior generalized performance and fairness. Moreover, through an ablation study, we demonstrate the effectiveness of each proposed component and justify the design of the proposed method. We summarize the main contributions of this paper as follows:

- To the best of our knowledge, we investigate the unfairness stemming from Visual Prompt Tuning (VPT) and its underlying causes for the first time.
- We propose Fair Visual Prompt Tuning (Fair-VPT) that efficiently adapts the pre-trained ViT model to downstream classification tasks while eliminating biased information related to sensitive attributes
- Through extensive experiments on benchmark datasets, we demonstrate that the proposed method efficiently enhances the fairness of ViT in various scenarios.

2. Related Work

2.1. Fairness-aware Classification

Numerous studies [12, 41, 42, 45, 46, 49, 50, 55, 58, 59] have tried to ensure fairness with respect to sensitive attributes in image classification tasks. Some approaches [42, 45, 55, 59] enhanced fairness by preventing the encoder networks from learning biased information related to sensitive attributes. Other approaches [12, 41, 49] disentangled the feature space into subspaces for target and sensitive attributes. In downstream tasks, these methods excluded the subspaces for sensitive attributes to ameliorate fairness. Additionally, certain methods [37, 48] mitigated imbalance between demographic groups by minimizing the worst-case training loss across all the groups. On the other hand, several approaches [46, 50, 58] endeavored to improve fairness by generating an unbiased dataset by utilizing Generative Adversarial Networks (GANs) [19]. Recently, Sudhakar et al. [52] and Qiang et al. [43] tried to address the unfairness problem caused by ViT models.

Notably, some studies [8, 9, 29, 30, 40, 61] have addressed model fairness under limited supervision of sensitive attributes. Several approaches [30, 32, 40] capitalized on the observation that the bias inducing significantly biased results, *i.e.,* malignant bias, is more easily learned than the target attribute. Consequently, these approaches estimated the bias using deliberately biased networks and sub-

sequently eliminated it. Similarly, other approaches [51, 61] utilized the proxies with high correlation to sensitive attributes, such as feature representation from biased networks, to estimate sensitive attributes. On the other hand, some research [29, 34] enhanced fairness without demographics by up-weighting the misclassified samples. DRO [22] demonstrated that minimizing the worst-case risk over all appropriate distributions can enhance fairness concerning potential sensitive attributes.

2.2. Transformers in Computer Vision

Inspired by the remarkable achievement of transformer models in the field of Natural Language Processing (NLP), a variety of research has attempted to introduce architectures based on self-attention into the various fields of computer vision. Vision Transformer [14] is the initial work that utilized the transformer architecture for image processing. It introduced an approach of tokenizing an image into patches, which are then utilized as inputs for the transformer. It achieved strong performances for various tasks, including image classification, detection, and segmentation. Subsequently, Touvron et al. [53] proposed a new distillation strategy based on distillation tokens to enhance the data efficiency in training, and Liu et al. [36] significantly reduced the computation cost by introducing hierarchical feature maps. On the other hand, some works [2, 3, 17, 57] utilized the ViT-based backbones to encode the frame-level features in video understanding tasks, such as video action recognition. In addition, other studies [11, 31, 38] exploited the vision-language pre-training models based on transformer architectures for multi-modal tasks (*e.g.,* visual question answering (VQA)).

From a different perspective, Jia et al. [24] proposed Visual Prompt Tuning (VPT), which efficiently adapts the pre-trained transformer model on large-scale datasets into various downstream tasks. They added a small amount of learnable parameters, *i.e.,* prompts, in the input space and only trained the prompts for transfer learning. It achieved comparable performances with full fine-tuning while keeping the transformer backbone frozen. Nevertheless, the analysis from the perspective of fairness was overlooked. In this paper, we identify unfairness in VPT and analyze its underlying causes, proposing a novel method that not only adapts the pre-trained ViT model to downstream tasks but also effectively enhances fairness.

3. Proposed Method

The proposed method is designed to ensure fairness with respect to sensitive attributes while adapting the pre-trained ViT to downstream classification tasks. The overall framework, which is illustrated in Figure 1, comprises the key components: categorized prompts, masked self-attention, and disentanglement loss.

3.1. Fairness Definition

Fairness notions have been diversely defined in the literature [7, 16, 18, 21, 56]. In this section, we first introduce the widely used notions of fairness, *i.e.,* demographic parity [16], equal opportunity [21], and equalized odds [21], and define fairness in this paper. Demographic parity means that a model should ensure the same ratio of positive outcomes across sensitive groups. Since it pursues the equality of outcome, it has the drawback of overlooking the real data distributions. Equal opportunity mitigates the shortcoming by ensuring the equality of True Positive Rates (TPR) across sensitive groups. Nevertheless, it is still limited in that it does not address the imbalance of negative outcomes. To equally consider positive and negative outcomes, we adopt the equalized odds (EO) to define fairness, which is measured as follows:

$$EO = \frac{|TPR_{s=0} - TPR_{s=1}| + |FPR_{s=0} - FPR_{s=1}|}{2} \tag{1}$$

where s and FPR denote the sensitive attribute and false positive rates.

3.2. Preliminaries

Let training samples (x,y,s) be given, where $x \in X$ is the input image, $y \in Y$ is the target label, and $s \in S$ is the sensitive attribute. For Vision Transformer (ViT) [14], the input image x is reshaped into N patches $x_p \in \mathbb{R}^{N \times 3 \times w_p \times h_p}$. Here, w_p and h_p are the fixed width and height of the patches. The patches are flattened with the linear protection and added with the positional embeddings as follows:

$$z_0 = [x^{cls}, E(x_p^{(1)}), E(x_p^{(2)}), ..., E(x_p^{(N)})], \tag{2}$$

where $E(x_p) \in \mathbb{R}^{N \times D}$ represents the embedded patches and x^{cls} denotes the class token. Subsequently, they are encoded by the transformer T as follows:

$$z_l = T_l(z_{l-1}), l = 1, 2, ...L. \tag{3}$$

Here, the transformer layer T_l consists of Multi-headed Self-Attention (MSA) and Feed-Forward Networks (FFN), where LayerNorm (LN) and residual connections are respectively deployed before and after each block. Finally, the encoded class token x_L^{cls} are fed into the classification head $C(\cdot)$ as follows:

$$y' = C(x_L^{cls}). \tag{4}$$

When a pre-trained model $\hat{T}$ is given, Visual Prompt Tuning (VPT) [24] adapts it to downstream classification tasks. The prompts $P \in \mathbb{R}^{N \times D}$, which are M learnable parameters, are added in the input space as follows:

$$\hat{z}_0 = [x^{cls}, P^{(1)}, ..., P^{(M)}, E(x_p^{(1)}), ..., E(x_p^{(N)})]. \tag{5}$$

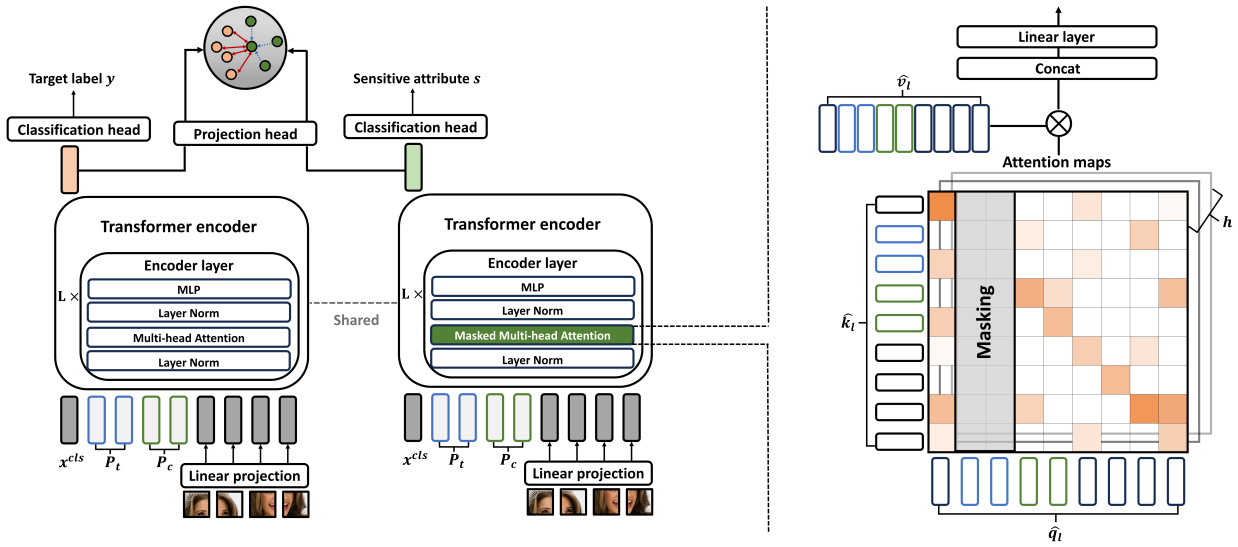

Figure 1. *Illustration of the proposed method*. The transformer encoder and linear projection are frozen, with only the prompts, classification head, and projection head being trained. The prompts are categorized into the target prompts (*i.e.,* P_t) and cleaner prompts (*i.e.,* P_c). These are concurrently encoded using Multi-head Attention (left) and Masked Multi-head Attention (right), which involves masking the target prompts. The differently encoded tokens (*i.e.,* orange and green) are trained to predict the target label and sensitive attribute, respectively.

The input sequence is encoded by the pre-trained model $\hat{T}$ as follows:

$$\hat{z}_l = \hat{T}_l(\hat{z}_{l-1}), l = 1, ...L, \tag{6}$$

While maintaining $\hat{T}$ frozen, only the prompts and classification head $\hat{C}(\cdot)$ for downstream tasks are trained through the classification loss $L_d = l(\hat{C}(\hat{x}_L^{cls}), y)$, where l represents the cross entropy loss.

3.3. Fair Visual Prompt Tuning

3.3.1 Categorizing Prompts

As aforementioned, the key idea is to designate certain prompts as "cleaner prompts", intended for the absorption of biased information originating from the pre-trained ViT. To this end, we reformulate the input sequence $\hat{z}_0$ using target prompts P_t and cleaner prompts P_c as follows:

$$\hat{z}_0 = [x_{cls}, P_t^{(1)}, ..., P_t^{(\alpha)}, P_c^{(1)}, ..., P_c^{(M-\alpha)}, E(x_p)], \tag{7}$$

where $\alpha \in \mathbb{N}, 1 \leq \alpha \leq M$. In addition, we will represent each vector in $\hat{z}$ as $\hat{z}^i$ (*e.g.,* $\hat{z}_0^0 = x_0^{cls}$).

3.3.2 Masked MSA and Encoding Prompts

Let denote query, key, and value representations of $\hat{z}_l$ as $\hat{k}_l, \hat{q}_l, \hat{v}_l \in \mathbb{R}^{(1+N+M)d_k}$. We calculate standard self-attention $SA(\hat{z}_l)$ and masked self-attention $SA^*(\hat{z}_l)$ in each layer as follows:

$$SA(\hat{z}_l) = softmax(\frac{\hat{q}_l \hat{k}_l^T}{\sqrt{d_k}})\hat{v}_l, \tag{8}$$

$$SA^*(\hat{z}_l) = softmax(\frac{\hat{q}_l \hat{k}_l^T + Mask}{\sqrt{d_k}})\hat{v}_l. \tag{9}$$

Here, the mask is defined as follows:

$$Mask_{i,j} = \begin{cases} -inf & \text{if} \quad 1 \leq j \leq \alpha \\ 0 & \text{else} \end{cases}, \tag{10}$$

where $i, j \in \{0, 1, ..., M + N\}$. Following the previous works [14, 24], self-attention is extended to Multihead Self-Attention (MSA) with h attention layers. The transformer layers encode the input sequence in two parallel ways, utilizing standard and masked MSA as follows:

$$\hat{z}_l = \hat{T}_l(\hat{z}_{l-1}), l = 1, ...L, \tag{11}$$

$$\hat{z^*}_l = \hat{T}_l^*(\hat{z^*}_{l-1}), l = 1, ...L, \tag{12}$$

where $\hat{T}_l^*$ is the transformer layers incorporating masked MSA. The resulting class tokens are input into the classification heads for the target label y and sensitive attribute s, represented by $\hat{C}(\hat{z}_L^{(0)})$ and $\tilde{C}(\hat{z}_L^{*(0)})$, respectively. The classification loss L_{cls} is formulated as:

$$L_{cls} = l(\hat{C}(\hat{z}_L^{(0)}), y) + l(\tilde{C}(\hat{z}_L^{*(0)}), s). \tag{13}$$

Since the target prompts are masked in $\hat{z^*}_l$, predicting the sensitive attribute using $\hat{z^*}_l$ encourages information related to the sensitive attribute to be included in the cleaner prompt. Simultaneously, by predicting the target label with $\hat{z}_l$, the rest information about the target label is contained in the target prompts.

3.3.3 Contrastive Loss for Disentanglement

However, a limitation of the aforementioned approach is that L_{cls} cannot ensure the exclusion of information related to the target label in the cleaner prompts, which leads to a loss of useful information for the target task. Besides, redundant information regarding sensitive attributes may be partially included in the target prompts. To address these issues, we design a disentanglement loss based on supervised contrastive learning [26] to explicitly decorrelate the prompts. Let I examples $\{x(i)\}_{i=1,...,I}$ be randomly sampled into a mini-batch. We embed $r(i) = g(\hat{z}_l(i))$ and $r^*(i) = g(\hat{z}^*_l(i))$ with the shared projection network $g(\cdot) \in \mathbb{R}^{d_k d_p}$. They are normalized to lie on the unit hypersphere and share a lower dimensional latent space for contrastive learning. When an embedded sample $r^*(i)$ is selected as an anchor, we define the positive set $P(i)$ and negative set $N(i)$ as follows:

$$P(i) = \{r^*(j) | y(j) = y(i), s(j) = s(i)\}, \tag{14}$$

$$N(i) = \{r(k) | y(k) = y(i), s(k) = s(i)\}. \tag{15}$$

While both $P(i)$ and $N(i)$ have the same target label and sensitive attribute with the anchor, it is worth noting that they are encoded in different manners, *i.e.,* Masked MSA or MSA. The disentanglement loss is defined as follows:

$$L^{dis} = -\sum_{\forall r^*(i)} \frac{1}{|P(i)|} \sum_{r^*(j) \in P(i)} \log \frac{exp(r^*(j) \cdot r^*(i)/\tau)}{\sum_{r(k) \in N(i)} exp(r(k) \cdot r^*(i)/\tau)}, \tag{16}$$

where $\frac{1}{|P(i)|}$ is the normalization term, exp denotes the exponential function, and $\cdot$ represents the inner product. τ is a temperature parameter. The loss encourages the target and cleaner prompts to include distinct information to each other by diminishing the similarity between r and r^*.

The overall loss is formally defined as:

$$L = L_{cls} + \lambda L_{dis}, \tag{17}$$

where λ is a hyper-parameter.

3.3.4 Training Downstream Classifier

After the prompt tuning, it is essential to exclude the cleaner prompts to ensure fairness with respect to the sensitive attribute. Therefore, we introduce a mask for the cleaner prompts in the self-attention process, defined as:

$$\overline{Mask}_{i,j} = \begin{cases} -inf & \text{if} \quad \alpha < j \leq M \\ 0 & \text{else} \end{cases}. \tag{18}$$

With keeping the transformer model and prompts frozen, the encoded class token $\overline{z}^{(0)}_L$ is achieved through the masked MSA, which is denoted as $\overline{SA}(\hat{z}_l) = softmax(\frac{\hat{q}_l \hat{k}_l^T + \overline{Mask}}{\sqrt{d_k}})\hat{v}_l$. Subsequently, the final classifier C_f is trained to predict the target label utilizing $\overline{z}^{(0)}_L$ while maintaining the other frozen components.

4. Experiment

4.1. Dataset

For reliable evaluation, we conduct extensive experiments on various benchmark datasets as follows.

- **CelebA** [35]: comprising about 200k facial images of celebrities with 40 facial attributes. To evaluate the fairness, we set *Male* as the sensitive attribute following the previous works [25, 42]. Based on the Pearson correlation [20] with it, we choose two target attributes, *Attractive* and *Big Nose*.
- **UTK Face** [60]: containing about 20k facial images with annotation for *Gender*, *Race*, and *Age*. Since the data split is not explicitly specified in the original version, we determine it based on the setting in [42]. Specifically, we set *Male* as the sensitive attribute and *Race* as the target label. We deliberately construct the imbalanced training set, where the majority groups (*i.e.,* Male-White and Female-Black) include four times as many samples as the minority groups (*i.e.,* Female-White and Male-Black). Additionally, we configure the validation and test sets as fully balanced datasets.
- **bFFHQ** [30]: containing about 21k images from the Flickr-Faces-HQ (FFHQ) dataset with a high resolution of 1024x1024. The target label is set to *Age* and the sensitive attribute to *Gender*, which are highly correlated to each other in the training set. To be specific, the minority groups (*i.e.,* Young-Male and Old-Female) account for only 0.5% of the training samples. The test set is composed as a fully unbiased dataset.
- **Waterbirds** [48]: including about 5k images which combine bird photos from CUB and backgrounds from the Place. The target label y ={*Waterbird*, *Landbird*} has a significant correlation with the background s={*Water*, *Land*}. Specifically, the minority groups (*i.e.,* Waterbirds-Land and Landbirds-Water) constitute only 0.5% of the training set. The validation and test sets are configured to have equal proportions of backgrounds for each target class.

4.2. Implementation Detail

For all comparable methods, we utilize the ViT-B/16 backbone [14] pre-trained on ImageNet-21k [47]. For ViT, we train a single fully connected layer for downstream classification without fine-tuning the backbone. For VPT-based methods, we exploit VPT-shallow following the original paper [24]. The length and dimensions of prompts (*i.e.,* M and

Method	Target label	Sensitive attribute		Accuracy (↑)	Balanced Accuracy (↑)	Equalized Odds (↓)
		s=0	s=1			
ViT [14]	t=0	69.1	96.4	78.4	68.7	41.6
	t=1	82.7	26.8			
VPT [24]	t=0	65.2	93.1	**81.7**	75.0	32.1
	t=1	89.1	52.7			
VPT [24]+AT [45]	t=0	38.9	63.7	67.6	63.2	24.0
	t=1	86.9	63.5			
VPT [24]+FSCL+ [42]	t=0	30.7	60.1	69.3	66.5	20.6
	t=1	93.6	81.8			
Fair-VPT (Ours)	t=0	73.8	85.8	78.6	**76.3**	**12.0**
	t=1	78.8	66.7			

Table 2. **Experimental results for *Attractive* on CelebA**. We set *Male* to the sensitive attribute and evaluate the results through three metrics. The third and fourth columns represent the group-wise classification accuracy. The bold represents the best scores in each metric.

Method	Target label	Sensitive attribute		Accuracy (↑)	Balanced Accuracy (↑)	Equalized Odds (↓)
		s=0	s=1			
ViT [14]	t=0	98.1	79.1	81.7	61.3	30.6
	t=1	12.8	55.1			
VPT [24]	t=0	98.3	81.6	82.7	62.8	28.5
	t=1	15.4	55.8			
VPT [24]+AT [45]	t=0	99.4	86.3	81.2	57.3	23.7
	t=1	4.5	38.8			
VPT [24]+FSCL+ [42]	t=0	99.3	89.9	**84.6**	63.6	25.1
	t=1	12.2	53.2			
Fair-VPT (Ours)	t=0	92.7	79.1	79.9	**65.4**	**15.9**
	t=1	35.6	53.9			

Table 3. **Experimental results for *Big Nose* on CelebA**. The sensitive attribute is set to *Male*. The proposed method achieves the best scores in terms of Balanced Accuracy and Equalized Odds.

D) are set to 50 and 768, respectively. In addition, the classification heads are designed as a single fully connected layer. For VPT+AT, we incorporate adversarial training [45] into VPT. To be specific, we introduce an auxiliary classification head that predicts the sensitive attribute with $\hat{x}_L^{cls}$, followed by a Gradient Reversal Layer (GRL). This framework encourages the prompts to exclude information about the sensitive attribute. For VPT+FSCL+, we employ the fair supervised contrastive loss [42] for prompt tuning. Subsequently, we train the downstream classifier based on the cross entropy loss. For ours, we determined hyper-parameters τ and α to be 0.1 and 25. Specifically, 25 cleaner prompts are randomly selected from the entire set of prompts. The other hyper-parameter λ is varied in the range of 0.1 to 1, depending on the dataset. All methods are trained using the Adam optimizer [27] with weight decay, and the initial learning rates are determined in the range of 0.1 to 0.01, depending on the datasets. We report the average score after three independent trials for all experiments and select the best model based on the validation set for 10 epochs. More details are provided in Appendix.

4.3. Evaluation metric

In our experiments, multiple metrics are exploited for reliable evaluation. Classification accuracy (Acc.) measures model performance within the data distribution of the test set, while balanced accuracy (BAcc.), calculated by averaging group-wise accuracy, assesses the generalization performance for classification. We adopt equalized odds (EO) [21] as the primary metric for evaluating fairness, supplementing it with equal opportunity (Eopp) [21] and demographic parity (DP) [16]. The results measured by Eopp and DP are provided in Appendix.

4.4. Comparison on CelebA

In Table 2 and 3, we show the comparison results on CelebA [35]. For both target attributes, *i.e., Attractive* and *Big Nose*, ViT [14] demonstrates the most unfair performances in

Method	TL	SA		BAcc. (↑)	EO (↓)
		s=0	s=1		
ViT [14]	t=0	96.0	80.3	88.4	13.4
	t=1	83.1	94.4		
VPT [24]	t=0	95.3	82.3	89.0	12.6
	t=1	83.6	94.9		
VPT [24]+AT [45]	t=0	95.5	81.5	88.9	11.6
	t=1	84.8	94.1		
VPT [24]+FSCL+ [42]	t=0	96.1	85.8	89.0	9.9
	t=1	82.3	91.9		
Fair-VPT (Ours)	t=0	95.1	89.3	**90.9**	**4.9**
	t=1	87.5	91.6		

Table 4. **Experimental results on UTK.** The target label and sensitive attribute are respectively set to *Race* and *Gender*. TL, SA, BAcc., and EO denote the target label, sensitive attribute, balanced accuracy, and equalized odds, respectively. Following the previous work [42], we compose the biased training set and balanced test set. Since accuracy and balanced accuracy are equal in this setting, accuracy is not reported.

terms of equalized odds (EO) and the lowest worst-group accuracy, which stands at 25.8 and 12.8, respectively. VPT [24] significantly enhances classification performance over ViT, achieving the highest accuracy; however, the balanced accuracy (BAcc.) is lower than ours due to its biased performances. Moreover, it demonstrates poor results in terms of fairness. For *Attractive*, VPT+AT [45] and VPT+FSCL+ [42] significantly enhances the fairness, while largely degrading the classification performances, where both accuracy and balanced accuracy are notably inferior to VPT. For *Big Nose*, while they ameliorate fairness over VPT, it rather aggravates the worst-group accuracy. For both settings, the proposed method achieves the most superior performances in terms of fairness. Despite a slight trade-off between fairness and accuracy, the proposed method exhibits improved generalization performances, resulting in the highest balanced accuracy. Particularly, the worst-group accuracy of ours is improved by 39.9% compared to ViT for *Attractive*.

4.5. Comparison on UTK Face

We reported the comparison results on UTK Face [60] in Table 4. We set the target label as *Race* and the sensitive attribute as *Gender*. Due to the imbalanced training set, ViT and VPT demonstrate unfair results, which stand at 13.4% and 12.6% in terms of equalized odds, respectively. VPT+AT and VPT+FSCL+ enhances fairness while maintaining balanced accuracy over VPT, but the improvement is not significant. When considering the results on CelebA together, they suggest that the straightforward integration of conventional fairness methods based on CNNs (*e.g., Adversarial Training [45]* into VPT framework may result in a significant information loss for the target label or only marginal improvements in fairness.

Method	TL	SA		BAcc. (↑)	EO (↓)
		s=0	s=1		
ViT [14]	t=0	99.1	54.3	74.8	48.9
	t=1	46.3	99.5		
VPT [24]	t=0	98.9	48.3	76.0	46.3
	t=1	57.5	99.5		
VPT [24]+AT [45]	t=0	99.5	58.7	77.5	43.1
	t=1	53.1	98.7		
Fair-VPT (Ours)	t=0	99.1	62.3	**80.7**	**37.1**
	t=1	61.9	99.5		

Table 5. **Experimental results on bFFHQ.** We set *Age* to the target label and *Gender* to the sensitive attribute. Since the minority groups have very few samples, comparable methods show poor performances for the groups. The proposed method markedly improves the worst group accuracy, resulting in an improvement in fairness.

4.6. Comparison on bFFHQ

To validate the proposed method under extremely biased scenarios, we conduct additional experiments on bFFHQ [30], where the minority groups contain only 0.5% of the entire sample. In Table 5, ViT and VPT correctly classify almost all samples from the majority groups, *i.e.,* {t=0/s=0} and {t=1/s=1}, while they demonstrate poor classification performance in the minority groups. Although VPT+AT significantly improves the worst-group accuracy and fairness, the proposed method achieves the highest worst-group accuracy and the lowest equalized odds, which stands at 61.9 and 37.1, respectively. Moreover, it notably improves balanced accuracy over VPT.

4.7. Ablation Study

Through an ablation study, we demonstrate the effectiveness of each proposed component and justify the design of the proposed method. In Table 6, we reported the results for *Attractive* on CelebA and for *Race* on UTK Face. The first row corresponds to our baseline, *i.e.,* VPT, which is the only one not employing *categorized Prompt*; categorizing prompts into the target prompts and cleaner prompts. In addition, it encodes the class token only using standard MSA into $\hat{z}_L^{(0)}$. In the second row, the prompts are categorized, and the class token is parallelly encoded using standard MSA into $\hat{z}_L^{(0)}$ and masked MSA for target prompts into $\hat{z}_L^{*(0)}$. Precisely, it corresponds to the proposed method without L_{dis} and significantly improves fairness. It demonstrates that the proposed framework based on *categorized Prompt* and masked MSA is effective. The full model (*i.e.,* the third row) further enhances fairness, achieving the best balanced accuracy and equalized odds in all the results.

Categorized Prompts	L_{cls}			L_{dis}	CelebA			UTK Face	
	$\hat{z}_L^{(0)}$	$\hat{z}_L^{*(0)}$	$\overline{z}_L^{(0)}$		Acc. (↑)	BAcc. (↑)	EO (↓)	BAcc. (↑)	EO (↓)
	✓				**81.7**	75.0	32.1	89.0	12.6
✓	✓	✓			77.9	75.9	15.0	89.4	8.1
✓	✓	✓		✓	78.6	**76.3**	**12.0**	**90.9**	**4.9**
✓		✓		✓	78.0	74.0	25.2	88.0	10.9
✓		✓	✓		77.3	72.9	29.1	89.2	12.2
✓		✓	✓	✓	78.4	73.9	24.4	89.6	9.4

Table 6. **Ablation study on CelebA and UTK Face.** We set *Attractive* to the target label on CelebA. *Catergotized Prompts* represents whether the prompts are categorized into the target and cleaner prompts. In addition, the second to fourth columns indicate whether the encoded tokens are utilized for visual prompt tuning. We set VPT as the baseline (the first row) and the third row indicates the full proposed model.

Method	TL	SA		Acc.	BAcc.	EO
		s=0	s=1			
ViT [14]	t=0	99.7	77.8	85.1	80.5	31.3
	t=1	52.0	92.6			
VPT [24]	t=0	99.6	82.9	**86.8**	81.2	29.2
	t=1	50.3	92.0			
VPT +AT [45]	t=0	98.7	81.3	86.3	81.6	27.0
	t=1	54.8	91.5			
Fair-VPT (Ours)	t=0	93.9	70.9	83.3	**84.3**	**18.7**
	t=1	78.9	93.6			

Table 7. **Experimental results on Waterbirds.** This experiment is conducted to validate the effectiveness of the proposed method in mitigating the background bias. Since the smallest group, *i.e.,* {t=1/s=0} includes only 56 training samples, the worst-group accuracy is notably low in the baselines. Nevertheless, the proposed method significantly elevates the worst-group accuracy, reaching 78.9%.

When excluding the loss term $l(\hat{C}(\hat{z}_L^{(0)}), y)$ in the proposed method (*i.e.,*, the third row), classification performance is declined, as anticipated. However, fairness is still enhanced compared to the baseline. We believe that it may be attributed to L_{cls}, which encourages the exclusion of information related to sensitive attributes from the target prompts. In the fifth row, $\overline{z}_L^{(0)}$ and $\hat{z}_L^{*(0)}$ are encoded by masking the sensitive and target prompts, respectively. Since the tokens are encoded independently, it fails to substantially enhance fairness over the baseline. Lastly, when incorporating L_{cls} into it, fairness is notably ameliorated by excluding sensitive attribute information from the target prompts.

4.8. Effectiveness in Addressing General Bias

To validate the effectiveness of the proposed method for a more general bias, we conduct comparison experiments on Waterbirds [48], including the background bias. Since the smallest group contains only 56 training samples, both ViT and VPT demonstrate significantly inferior worst-group accuracy, and VPT+AT exhibits only marginal improvement in fairness. Meanwhile, the proposed method notably increases classification accuracy in the smallest group to 78.9%, resulting in the most superior balanced accuracy and equalized odds. These results indicate that the proposed method can be utilized not only to address facial attributes but also to mitigate more general biases such as the background.

5. Conclusion

In this paper, we initially investigated the unfairness arising from Visual Prompt Tuning (VPT) and its underlying causes. To address the unfairness problem in VPT, we proposed a Fair Visual Prompt Tuning (Fair-VPT) which eliminates biased information related to sensitive attributes in the pre-trained ViT model while adapting it to downstream classification tasks. Specifically, we categorized the prompts into the target and cleaner prompts, and encoded the class token in dual parallel manners. Consequently, it encouraged information related to the sensitive attribute and target label to be included in the cleaner prompts and target prompts, respectively. Moreover, we introduced a disentanglement loss based on contrastive learning to further decorrelate them. In the inference time, we excluded the cleaner prompts to ameliorate fairness. To validate the proposed method, we conducted extensive experiments on multiple benchmarks. In various scenarios, the proposed method achieved the most superior generalization performance and fairness.

Acknowledge

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2B5B02001467).

References

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. ProPublica, 2016. 2

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6816–6826, 2021. 3

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, pages 813–824. PMLR, 2021. 3

[4] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc. 2

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. 1

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. 1

[7] Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 2, 3

[8] Junyi Chai and Xiaoqian Wang. Self-supervised fair representation learning without demographics. In *Advances in Neural Information Processing Systems*, 2022. 2

[9] Junyi Chai, Taeuk Jang, and Xiaoqian Wang. Fairness without demographics through knowledge distillation. In *Advances in Neural Information Processing Systems*, 2022. 2

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 1

[11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, page 104–120, Berlin, Heidelberg, 2020. Springer-Verlag. 3

[12] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1436–1445, Long Beach, California, USA, 2019. PMLR. 2

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 1

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 3, 4, 5, 6, 7, 8

[15] Conor Dougherty. Google photos mistakenly labels black people gorillas. Twitter, 2015. 2

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery. 2, 3, 6

[17] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3

[18] Sixue Gong, Xiaoming Liu, and Anil Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *In Proceeding of European Conference on Computer Vision*, Virtual, 2020. 3

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[20] Emily M. Hand and Rama Chellappa. Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, page 4068–4074. AAAI Press, 2017. 5

[21] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc. 1, 2, 3, 6

[22] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. 3

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[24] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[25] Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12115–12124, 2021. 2, 5

[26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, pages 18661–18673. Curran Associates, Inc., 2020. 5

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[28] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30*, pages 4066–4076. Curran Associates, Inc., 2017. 2

[29] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without demographics through adversarially reweighted learning. Red Hook, NY, USA, 2020. Curran Associates Inc. 2, 3

[30] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. In *Advances in Neural Information Processing Systems*, 2021. 2, 5, 7

[31] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*, 2020. 3

[32] Jongin Lim, Youngdong Kim, Byungjai Kim, Chanho Ahn, Jinwoo Shin, Eunho Yang, and Seungju Han. Biasadv: Bias-adversarial augmentation for model debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3832–3841, 2023. 2

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 1

[34] Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. *CoRR*, abs/2107.09044, 2021. 3

[35] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 1, 2, 5, 6

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 3

[37] Vishnu Suresh Lokhande, Kihyuk Sohn, Jinsung Yoon, Madeleine Udell, Chen-Yu Lee, and Tomas Pfister. Towards group robustness in the presence of partial group labels. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022. 2

[38] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23, 2019. 3

[39] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2624–2637, 2013. 1, 2

[40] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*, pages 20673–20684. Curran Associates, Inc., 2020. 2

[41] Sungho Park, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Learning disentangled representation for fair facial attribute classification via fairness-awareunbiased information alignment. *Proceedings of AAAI-2021*, 2021. 2

[42] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10389–10398, 2022. 2, 5, 6, 7

[43] Yao Qiang, Chengyin Li, Prashant Khanduri, and Dongxiao Zhu. Fairness-aware vision transformer via debiased self-attention, 2023. 2

[44] Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019. 1

[45] Edward Raff and Jared Sylvester. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 189–198, 2018. 2, 6, 7, 8

[46] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9301–9310, 2021. 2

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,

Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 1, 5

[48] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 2, 5, 8

[49] Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, pages 746–761. Springer, 2020. 2

[50] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3:1–3:9, 2019. 2

[51] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised learning of debiased representations with pseudo-attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16742–16751, 2022. 3

[52] Sruthi Sudhakar, Viraj Prabhu, Arvindkumar Krishnakumar, and Judy Hoffman. Mitigating bias in visual transformers via targeted alignment. In *British Machine Vision Conference*, 2023. 2

[53] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers; distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 1, 3

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1

[55] Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, 2018. 2

[56] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5309–5318, 2019. 2, 3

[57] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 3

[58] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575, 2018. 2

[59] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, page 335–340, New York, NY, USA, 2018. Association for Computing Machinery. 2

[60] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 5, 7

[61] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, page 1433–1442, New York, NY, USA, 2022. Association for Computing Machinery. 2, 3