FAIRVIT: FAIR VISION TRANSFORMER VIA ADAPTIVE MASKING

Bowei Tian[⋆]*, Ruijie Du*[†]*, Yanning Shen*[†*]

[⋆] Wuhan University, Hubei, China
[†] University of California, Irvine, CA, USA

ABSTRACT

Vision Transformer (ViT) has achieved excellent performance and demonstrated its promising potential on various computer vision tasks. The wide deployment of ViT in real-world tasks requires a thorough understanding of the societal impact of the model. However, most ViTs do not take fairness into account and existing fairness-aware algorithms designed for CNNs do not perform well on ViT. It is necessary to develop a new fair ViT framework. Moreover, previous works typically sacrifice accuracy for fairness. Therefore, we aim to develop an algorithm that improves fairness without sacrificing accuracy too much. To this end, we introduce a novel distance loss, and deploy adaptive fairness-aware masks on attention layers to improve fairness, which are updated with model parameters. Experimental results show the proposed methods achieve higher accuracy than alternatives, 6.72% higher than the best alternative while reaching a similar fairness result.

***Index Terms*—** Vision Transformer, Accuracy, Fairness, Adaptive Masking

1. INTRODUCTION

Vision transformer (ViT) [1, 2] has been widely adopted in various computer vision (CV) tasks, and is considered a viable alternative to the Convolutional Neural Network (CNN) [3]. Unlike CNN, ViT has a specialized structure that can extract global relationships via self-attention mechanism, leading to improved performance in various CV tasks, including image classification [2, 4], object detection [5] and instance segmentation [6]. Due to its excellent performance, the structure has formed the architectural backbone of many CV algorithms for real-world applications. However, wide deployment of CV algorithms highly relies how responsible they are [7, 8]. This motivates the investigation of the fairness of ViT.

Most of the existing debiasing methods for image classification tasks are specified for CNN or deep neural network (DNN) models [9, 10], and can not be directly applied to ViTs. However, several studies show that CV models make predictions by mixing sensitive features with input features [11, 10], and these sensitive features may capture biased relationships between the input features and the ground truth labels. For example, the sensitive feature 'gender' usually influences the accuracy of a face recognition task. In this case, it may lead to discriminatory results towards underrepresented groups, which causes severe social and ethical problems.

Fairness of ViT has been investigated in several recent works[8, 12]. Sudhakar et al. [8] propose TADeT, a targeted alignment technique that seeks to identify and eliminate bias from the query matrix in ViT. However, they directly manipulate the query matrix, which will sacrifice accuracy for fairness. Dehghani et al. [12] introduce a ViT-22B model containing 22 billion parameters, which leads to an improving tradeoff between fairness and accuracy. Yet they require an extremely high computing power, and when training on relatively small models such as ViT-base, fairness is still an issue.

To address the aforementioned challenges, we introduce a novel framework for fairness awareness. We introduce a novel adaptive masking framework and learn the group-specific mask weights to improve fairness. Meanwhile, a novel distance loss is also introduced to enhance the accuracy. Experiments were carried out on real datasets, and it is shown that we reach a 6.72% higher accuracy than the best alternative while achieving a similar fairness result.

2. PROBLEM STATEMENT

2.1. Classification Task

Given training dataset $T_t = \{(\mathbf{x}_i, s_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i$ is feature vector of the i_{th} data sample, s_i is the corresponding sensitive feature, and y_i denotes the ground truth label. the goal of a classification task is to find a model $f(\mathbf{x}; \boldsymbol{\theta})$ which maps the features $\mathbf{x}$ to and class labels $y \in \{0, 1, \cdots, C-1\}$, where C is the number of classes. The conventional classification framework entails the form of

$$\min_{\theta} L(f(\mathbf{x}, \boldsymbol{\theta}), y) \tag{1}$$

where $f(\mathbf{x}, \boldsymbol{\theta})$ is the learned model parameterized by $\boldsymbol{\theta}$, L is the loss function characterizing the discrepancy between the estimated label and the ground truth label. One typical choice for L is the cross-entropy loss [13]. However, cross-entropy loss is bias-oblivious and will lead to potential bias in the results [14]. Hence, the goal of the present paper is to develop a novel framework $f(\mathbf{x}, \boldsymbol{\theta})$ to mitigate bias.

* Corresponding author: yannings@uci.edu

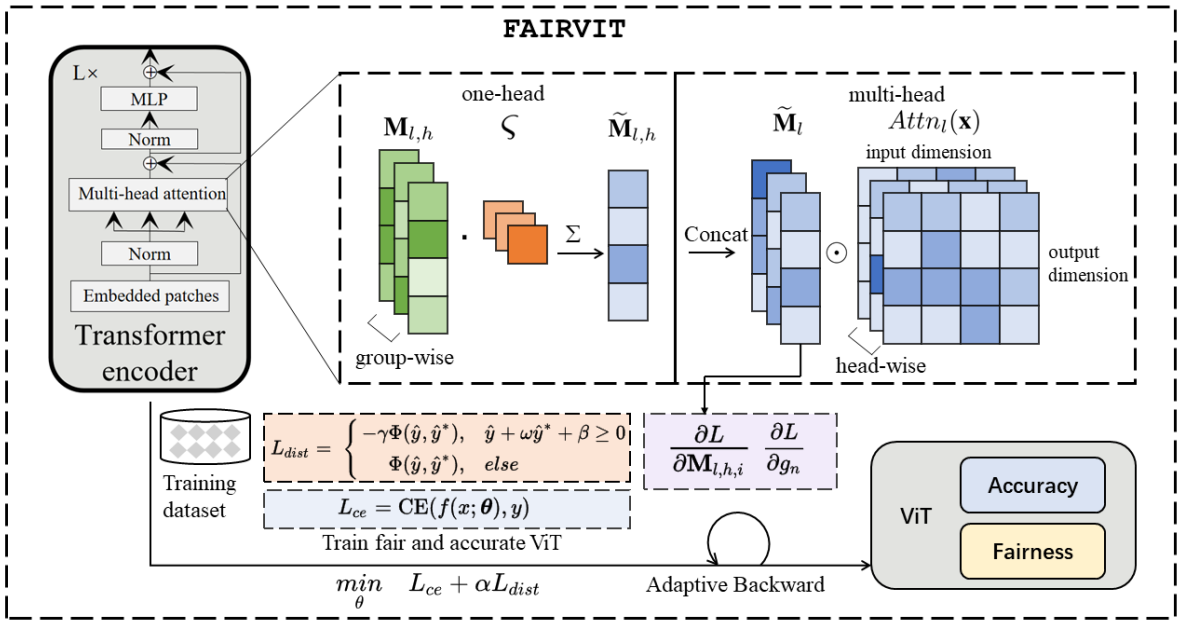

Fig. 1: Our purposed method of FairViT . We apply mask weight ς to group-wise masks $\mathbf{M}_{l,h}$, calculate their sum $\widetilde{\mathbf{M}}_{l,h}$, and concatenate each head to get $\widetilde{\mathbf{M}}_l$. When training, we consider a novel distance loss L_{dist} and also introduce an adaptive backward algorithm to optimize $\mathbf{M}_{l,h,i}$ and g_n.

2.2. Vision Transformer

Recently, the transformer has been adapted to computer vision by modeling relationships between different parts of an image using the self-attention mechanism. A widely-used transformer is ViT [1]. For the i_{th} sample, let $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, ..., x_{i,p}\}$ be a sequence of p input image patches, where each patch is a $v \times v \times c$-dimensional tensor, with $v \times v$ pixels and c channels. ViT first applies an embedding layer to each patch to convert it into a D-dimensional embedding vector $Embedding(\mathbf{x}) = \{e_1, e_2, ..., e_p\}$. Next, ViT applies a series of transformer encoder layers to the embeddings, and each encoder layer comprises two sub-layers: a Multi-Head Attention mechanism (MHA) and a position-wise FeedForward Network (FFN). The MHA layer models the interactions between the patch embeddings using self-attention, while the FFN layer applies a non-linear transformation to each patch embedding individually. The goal is to manipulate the encoder layer to address the accuracy and fairness issues.

3. FAIRNESS-AWARE VISION TRANSFORMER DESIGN

3.1. Adaptive Masking

Inspired by the wheelchair-accessible parking lots [15], where we create more opportunities for minority groups, we introduce similar ideas in the ViT structure to improve fairness and maintain the model's accuracy. Suppose there is head h [1](H heads in a transformer encoder layer) and the l-th layer. We split the dataset into G groups evenly, leaving the remainder as the last group, under the condition that 0 or 1 group contains samples with different sensitive features ($s = 0$ and $s = 1$), and attach each group to a mask. The mask $\mathbf{M}_{l,h,i}$ denotes the i_{th} mask ($i \in \{1, \cdots, G\}$) in layer l, head h, which we construct to control the information flow in forward propagation. To clarify, the one-Head Attention (HA) mechanism is:

$$\mathrm{Attn}_{l,h}(\mathbf{x}) = \mathrm{S}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{2}$$

$$\widetilde{\mathbf{M}}_{l,h} = \sum_{i=1}^{G} \varsigma_i \mathbf{M}_{l,h,i} \tag{3}$$

$$\mathrm{HA}(\mathbf{x}, \mathbf{M}_{l,h}) = \widetilde{\mathbf{M}}_{l,h} \odot \mathrm{Attn}_{l,h}(\mathbf{x}) \tag{4}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d is the dimension of the key vectors [1]. $\odot$ means element-wise product, $\mathrm{S}(\cdot)$ is softmax function, ς_i is the mask weight of $\mathbf{M}_{l,h,i}$, $\widetilde{\mathbf{M}}_{l,h}$ is the weighted sum of $\mathbf{M}_{l,h,i}$. We add a constraint $\mathrm{clamp}(\mathbf{M}_{l,h}, -1, 1)$ on $\mathbf{M}_{l,h}$, meaning any value in $\mathbf{M}_{l,h}$ larger than 1 are clamped to 1, and smaller than -1 are clamped to -1. We use -1 as the lowest constraint because the backward propagation of $\mathbf{M}_{l,h,i}$ may yield a negative value, which contains information. $\mathbf{M}_{l,h,i}$ was initialized with all 0. We get the Multi-Head Attention

(MHA) by the concatenate operation purposed by [1], formalized as below:

$$\begin{aligned}&\mathrm{MHA}(\mathbf{x}, \mathbf{M}_l) = \\ &\Lambda(\mathrm{HA}(\mathbf{x}, \mathbf{M}_{l,1}), \mathrm{HA}(\mathbf{x}, \mathbf{M}_{l,2}), \ldots, \mathrm{HA}(\mathbf{x}, \mathbf{M}_{l,H}))\end{aligned} \tag{5}$$

where Λ is the concatenate operation, and $\mathrm{HA}(\mathbf{x}, \mathbf{M}_{l,h}) \in \mathbb{R}^{p\times d}, \mathrm{MHA}(\mathbf{x}, \mathbf{M}_l) \in \mathbb{R}^{p\times(Hd)}$.

3.2. Mask and Weight Update

To get mask weight, we split each group into training-set and validation-set (0.9:0.1). At the end of every epoch, for all groups $i \in \{1, \ldots, G\}$, we get their index list ρ, sort ρ in ascending validation-set accuracy order, and set a disadvantage group ratio χ to determine the disadvantage group proportion from the beginning of ρ. We get a disadvantage group list I with length $r(\chi G)$, $r(\cdot)$ is the round operation. The remaining groups are loaded in list I_0. We further divide I into N sublists $(I_1, I_2, \cdots, I_N)$, leaving the last sublist as the remainder, and get ς_i by:

$$\varsigma_i = g_n, \quad i \in I_n, \quad n \in \{0, \cdots, N\} \tag{6}$$

where g_n is the mask weight determinant of I_n. We set the constraint of g_n by $\mathrm{clamp}(\varsigma, \epsilon, 1-\epsilon)$, and ϵ is a small value, set as $1e^{-8}$ in our experiment.

The $\mathbf{M}_{l,h,i}$ and g_n are updated through backward propagation. The gradient of masks can be obtained as:

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{M}_{l,h,i}} = \frac{\partial L}{\partial \mathrm{HA}} \odot \mathrm{Attn}_{l,h}(\mathbf{x}) \cdot g_n, \\ i \in I_n, \quad n \in \{0, \cdots, N\}\end{aligned} \tag{7}$$

To update g_n, we first obtain the computing map of g_n towards $\widetilde{\mathbf{M}}_{l,h}$ according to (3) and (6):

$$\widetilde{\mathbf{M}}_{l,h} = \sum_n (g_n \sum_{i\in I_n} \mathbf{M}_{l,h,i}) \tag{8}$$

then the gradient of g_n can be written as

$$\frac{\partial L}{\partial g_n} = \sum_p \sum_d (\frac{\partial L}{\partial \mathrm{HA}} \mathrm{Attn}_{l,h}(\mathbf{x}) \cdot (\sum_d \sum_{i\in I_n} \mathbf{M}_{l,h,i})). \tag{9}$$

Therefore, the parameters $\mathbf{M}_{l,h,i}$ and g_n can be updated through back propagation.

3.3. Distance Loss

Following some regularizer-based works on fairness [16, 14], we aim to design a regularizer to maintain accuracy. We define $\hat{y} = f_{\{y\}}(\mathbf{x}; \boldsymbol{\theta})$ as the predicted score of the ground truth label y, and $\hat{y}^* = \sum_{i\in\{topk\}/\{y\}} f_i(\mathbf{x}; \boldsymbol{\theta})$ as the sum over scores of the difference set between top k labels and the ground truth label, we set $k = 3$ by default. In each epoch, we split the data into training and validation stage. We learn a linear classifier using logistic regression in the validation stage, which maps $(\hat{y}, \hat{y}^*)$ pairs for each data sample to label $z = 0$ or 1, which indicates whether the sample was misclassified. Morspecifically, $z = \mathbb{1}(\hat{y} = \max(f(\mathbf{x}; \boldsymbol{\theta})))$. Suppose the learned linear classifier is with the form

$$\hat{y} + \omega\hat{y}^* + \beta = 0 \tag{10}$$

Then we introduce the following distance term

$$\Phi(\hat{y}, \hat{y}^*) = \frac{|\hat{y} + \omega\hat{y}^* + \beta|}{\sqrt{1+\omega^2}} \tag{11}$$

which indicates the distance between point $(\hat{y}, \hat{y}^*)$ and the hyperplane in (10). Hence, we introduce the distance loss as:

$$L_{dist} = \begin{cases} -\gamma\Phi(\hat{y}, \hat{y}^*), & \hat{y} + \omega\hat{y}^* + \beta \geq 0 \\ \Phi(\hat{y}, \hat{y}^*), & else \end{cases} \tag{12}$$

Algorithm 1 Adaptive Mask based FairViT

Require: Transformer model $f(\mathbf{x}; \boldsymbol{\theta})$, training data set T_t, validation data set T_v, threshold t, epoch E and learning rate lr.
Ensure: The accurate and fair transformer $f(\mathbf{x}; \boldsymbol{\theta})^*$.

$L = INF$, $i = 0$, $h = 0$.
Initialize the mask weight $\mathbf{M}_{l,h,i}$ with all values of 0.
while $h < E$ **and** $L > t$ **do**
  for all $\mathbf{x} \in T_t$ **do** {training stage}
    $L_{ce} = \mathrm{CE}(f(\mathbf{x}; \boldsymbol{\theta}), y)$.
    if i = 0 **then**
      $L = L_{ce}$.
      Obtain $\frac{\partial L}{\partial \boldsymbol{\theta}}$ through backward propagation.
    else
      Obtain L_{dist} by (12).
      $L = L_{ce} + \alpha \cdot L_{dist}$.
      Obtain $\frac{\partial L}{\partial \boldsymbol{\theta}}$ through backward propagation.
      Obtain $\frac{\partial L}{\partial \mathbf{M}_{l,h,i}}$, $\frac{\partial L}{\partial g_n}$ by (7-9).
      $\mathbf{M}_{l,h,i} = \mathbf{M}_{l,h,i} - lr \cdot \frac{\partial L}{\partial \mathbf{M}_{l,h,i}}$.
      $g_n = g_n - lr \cdot \frac{\partial L}{\partial g_n}$.
    end if
    $\boldsymbol{\theta} = \boldsymbol{\theta} - lr \cdot \frac{\partial L}{\partial \boldsymbol{\theta}}$.
  end for
  for all $\mathbf{x} \in T_v$ **do** {validation stage}
    $\hat{y} = f_{\{y\}}(\mathbf{x}; \boldsymbol{\theta})$.
    $\hat{y}^* = \sum_{i\in\{topk\}/\{y\}} f_i(\mathbf{x}; \boldsymbol{\theta})$.
  end for
  Update w and β in (10) by classifying $(\hat{y}, \hat{y}^*) \rightarrow \{0, 1\}$.
  $h = h + 1$.
end while
$f(\mathbf{x}; \boldsymbol{\theta})^* = f(\mathbf{x}; \boldsymbol{\theta})$.
return $f(\mathbf{x}; \boldsymbol{\theta})^*$.

where γ are non-negative hyper-parameters. The overall loss function is $L = L_{ce} + \alpha L_{dist}$. where $L_{ce} = \text{CE}(f(x;\boldsymbol{\theta}), y)$ is the cross-entropy loss. Because the hyperplane is meaningless at the beginning, in the first epoch, we train $\boldsymbol{\theta}$ with L_{ce}, and then we obtain and update the hyperplane in each epoch using $L_{ce} + \alpha L_{dist}$.

Distance loss can help improve accuracy. From the definition of $\hat{y}$ and $\hat{y}^*$, a higher $\hat{y}$ and a lower $\hat{y}^*$ implies a higher possibility for correct classification, so as the mapping from which we get (10). Therefore, ω should be negative, and the slope of the hyperplane should be positive. And the distance loss reduces $\Phi(\hat{y}, \hat{y}^*)$ when $\hat{y} + \omega\hat{y}^* + \beta < 0$, and increases $\Phi(\hat{y}, \hat{y}^*)$ if $\hat{y} + \omega\hat{y}^* + \beta \geq 0$, meaning that the point $(\hat{y}, \hat{y}^*)$ will keep going lower-right. In both cases, the distance loss promotes $\hat{y}^*$ to be smaller and $\hat{y}$ to be larger, which is beneficial to accuracy.

4. EVALUATION

4.1. Experimental Setup

We conduct experiments on the CelebA dataset [17]. All experiments are averaged over 5 independent runs, and the results are expressed as "mean $\pm$ standard deviation". We set $\gamma = 0.2$ and $\alpha = 0.1$ in our experiments. Our fairness metric are as follows:

Difference of Equalized Opportunity (ΔEO) [18, 14] evaluates the difference of true positive rates (TPR) on different sensitive attributes. Specifically, $\Delta\text{EO} = \sum_i^K \frac{R_i}{R} \Delta\text{EO}_i$ where $\Delta\text{EO}_i = |\text{TPR}_{y=i,s=1} - \text{TPR}_{y=i,s=0}|$, $\text{TPR}_{y=i,s=j} = P(\hat{y} = i, y = i, s = j)/P(y = i, s = j)$, $\hat{y}$ is the predicted label, y is the ground truth label, and s is a sensitive feature identifier, such as to identify male ($s = 1$) or female ($s = 0$), R_i is the number of samples in class i and R is the total number of samples.

Bias Amplification (BA) measures the amplification of unfairness on the test set compared with the training set [11]:

$$\text{BA} = \frac{\text{TPR}_{s=\mathbb{1}}}{\text{TPR}_{s=0} + \text{TPR}_{s=1}} - \frac{\text{TPR}'_{s=\mathbb{1}}}{\text{TPR}'_{s=0} + \text{TPR}'_{s=1}} \qquad (13)$$

where TPR' means the true positive rate in the training dataset, and TPR means the true positive rate in the test dataset. $\mathbb{1}$ is out of indicator function $\mathbb{1}\left(\frac{\text{TPR}'_{s=0}}{\text{TPR}'_{s=0}+\text{TPR}'_{s=1}} < \frac{1}{2}\right)$.

4.2. Experiment Results

Baselines: We select several baselines to compare with our work. Vallina [1] uses the rural cross-entropy loss to train ViT, Maximum Mean Discrepancy (**MMD**) [19] calculates the mean of penultimate layer feature activation values for each sensitive attribute settings and minimize their L_2 distance. Mitigating Bias in ViT via Target Alignment (**TADeT**) [8] uses a targeted alignment strategy to generate fair ViT that aims to identify and remove bias from their query matrix features. Fair Supervised Contrastive Loss (**FSCL**) [10] inherit the philosophy of supervised contrastive learning and encourages the representation of the same class to be closer than that of different classes, and **FSCL+** further introduces a group-wise normalization to improve fairness.

Table 1: Comparison of Vanilla [1], MMD [19], TADeT [8], FCSL [10], FSCL+ [10] and FairViT .

Baseline	**ACC**	**ΔEO**	**BA**
Vanilla	$\underline{79.49 \pm 0.6\%}$	0.193 ± 0.021	0.0670 ± 0.009
MMD	$75.01 \pm 1.8\%$	0.158 ± 0.025	0.0581 ± 0.008
TADeT	$73.34 \pm 1.2\%$	0.135 ± 0.012	0.0424 ± 0.009
FSCL	$76.79 \pm 0.6\%$	0.113 ± 0.004	0.0207 ± 0.005
FSCL+	$66.71 \pm 1.6\%$	$\mathbf{0.026 \pm 0.001}$	$\mathbf{0.0058 \pm 0.002}$
FairViT	$\mathbf{86.41 \pm 1.1\%}$	$\underline{0.061 \pm 0.014}$	$\underline{0.0180 \pm 0.004}$

Table 1 shows FairViT has a better fairness performance without sacrificing much accuracy. Compared to FSCL+, FairViT reaches a higher accuracy of 19.70. Moreover, while the average ΔEO in Vanilla, MMD, TADeT, and FSCL are 0.193, 0.158, 0.135, and 0.113 correspondingly, FairViT reach an average ΔEO of 0.061, which is a fairer result.

Table 2: Ablation Study of our methods.

Method	**ACC**	**ΔEO**	**BA**
L_{ce}	$79.49 \pm 0.6\%$	0.193 ± 0.021	0.067 ± 0.009
$L_{ce} + L_{dist}$	$81.30 \pm 0.5\%$	0.132 ± 0.011	0.048 ± 0.008
L_{ce} + adapt	$\underline{84.20 \pm 1.3\%}$	$\underline{0.106 \pm 0.008}$	$\underline{0.032 \pm 0.006}$
$L_{ce} + L_{dist}$ + adapt	$\mathbf{86.41 \pm 1.1\%}$	$\mathbf{0.061 \pm 0.014}$	$\mathbf{0.018 \pm 0.004}$

Method ablation study: We conduct an ablation study to examine the effectiveness of L_{dist} and mask weight. The results are in Table 2, and note that adapt is the adaptive mask method. We can see that both methods improve accuracy, and mask weight plays a more important role in enhancing accuracy and fairness. Moreover, applying both methods reaches the best result in accuracy and fairness, i.e. the accuracy of 86.41%, ΔEO of 0.061 and BA of 0.018, hence we consider it as the best solution.

5. CONCLUSION AND DISCUSSION

In this work, we propose a novel fair and accurate classification method on ViT. We apply the adaptive masks, use mask weights to mitigate bias and avoid losing much accuracy, and carefully design the distance loss to enhance accuracy. The experiments on baselines and ablation study demonstrate that we can improve accuracy while maintaining a similar fairness. In the future, we will consider further studying the reason why distance loss and mask weight work, and extend our methods into various neural networks.

6. REFERENCES

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[2] 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*, 2021, pp. 10012–10022.

[3] 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*, 2016, pp. 770–778.

[4] Mazda Moayeri, Phillip Pope, Yogesh Balaji, and Soheil Feizi, "A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19087–19097.

[5] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen, "Up-detr: Unsupervised pre-training for object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1601–1610.

[6] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia, "End-to-end video instance segmentation with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8741–8750.

[7] Yao Qiang, Chengyin Li, Prashant Khanduri, and Dongxiao Zhu, "Fairness-aware vision transformer via debiased self-attention," *arXiv preprint arXiv:2301.13803*, 2023.

[8] Sruthi Sudhakar, Viraj Prabhu, Arvindkumar Krishnakumar, and Judy Hoffman, "Mitigating bias in visual transformers via targeted alignment," *arXiv preprint arXiv:2302.04358*, 2023.

[9] Zhibo Wang, Xiaowei Dong, Henry Xue, Zhifei Zhang, Weifeng Chiu, Tao Wei, and Kui Ren, "Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10379–10388.

[10] 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*, 2022, pp. 10389–10398.

[11] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," *arXiv preprint arXiv:1707.09457*, 2017.

[12] Yuji Roh, Weili Nie, De-An Huang, Steven Euijong Whang, Arash Vahdat, and Anima Anandkumar, "Dr-fairness: Dynamic data ratio adjustment for fair training on real and generated data," *Transactions on Machine Learning Research*, 2023.

[13] Anqi Mao, Mehryar Mohri, and Yutao Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," *arXiv preprint arXiv:2304.07288*, 2023.

[14] Ruijie Du and Yanning Shen, "Fairness-aware user classification in power grids," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 1671–1675.

[15] Roland Graf, Sun Young Park, Emma Shpiz, and Hun Seok Kim, "igym: A wheelchair-accessible interactive floor projection system for co-located physical play," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–6.

[16] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 962–970.

[17] 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)*, December 2015.

[18] Moritz Hardt, Eric Price, and Nati Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[19] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.