Reproducibility Study of "FairViT: Fair Vision Transformer via Adaptive Masking"

Anonymous authors
Paper under double-blind review

Abstract

Recently, Vision Transformers (ViTs) have excelled in computer vision tasks but often struggle with fairness issues related to attributes like gender and hair colour. FairViT by Tian et al. (2024), aims to address this challenge, by introducing adaptive masking combined with a distance-based loss, to improve fairness and accuracy, while maintaining competitive computational efficiency compared to other baseline methods. In our reproducibility study, we evaluated FairViT on the CelebA dataset on tasks related to attractiveness and facial expression prediction, while considering specific sensitive attributes. We then compared FairViT against the Vanilla and Fair Supervised Contrastive Loss (FSCL) baseline models. Contrary to the original claim regarding the effectiveness of adaptive masking, we observed that its impact is negligible, in terms of both fairness and accuracy, a finding confirmed also on the UTKFace dataset. On the other hand, the distance-based loss demonstrated partial effectiveness, but mainly when tested in the context of different architectures. Finally, in terms of computational efficiency, FairViT required almost double training time per epoch compared to the Vanilla model and did not outperform FSCL, which had the lowest training time for the specified dataset size used by the authors. Overall, our findings highlight the potential effectiveness of the proposed distance loss. However, the adaptive masking method did not deliver the expected improvements while also increasing the computational cost. Our implementation is available at [1].

1 Introduction

Vision Transformers (ViT) have become a valuable asset for Computer Vision (CV) downstream applications (e.g., classification and detection), yielding excellent results when pre-trained on large amounts of data, compared to state-of-the-art (SOTA) Convolutional Neural Networks (CNNs) (Dosovitskiy et al., 2021). However, ViTs present risks related to fairness due to algorithmic and data biases (Sudhakar et al., 2023), stemming from datasets containing inaccurate or imbalanced demographic data, which can be transferred through the employed architecture. For instance, the CelebA dataset (Liu et al., 2015) introduces data bias into the model due to widespread inconsistencies and inaccuracies in its attribute labelling (Lingenfelter et al., 2022). When used in ViT models, this bias is propagated via the attention mechanism (Mandal et al., 2023), disproportionally emphasising certain features. As a result, predictions can be skewed, disadvantaging underrepresented groups, indicating the need for fairer CV architectures.

Existing methods to mitigate bias include pre-processing, in-processing, and post-processing dealing with fairness, in different stages of AI applications (Kamiran & Calders, 2011; Cruz & Hardt, 2024; Hardt et al., 2016). Pre-processing techniques, such as the use of counterfactual data augmentation (Brinkmann et al., 2023), refine the dataset. In-processing techniques manipulate the model's architecture for de-biasing. Such examples include the use of bi-level optimisation for adjusting the data sampling ratios (Roh et al., 2023), directly removing bias from the query matrix (Sudhakar et al., 2023) of a Vision Transformer, or using De-biased Self-Attention to adjust the ViT through adversarial training (Qiang et al., 2024). Lastly, post-processing techniques try to make a model fairer, by adjusting the classification thresholds after it is trained

[1] `https://anonymous.4open.science/r/FairViT-reproducibility-study-54B0/`

(Pleiss et al., 2017). Nevertheless, all these methods are susceptible to the trade-off between accuracy and fairness (Wang et al., 2021), highlighting the need for innovative approaches that balance these two properties, while maintaining computational efficiency.

An approach that addresses the fairness issues in ViT models employed by Tian et al. (2024), proposes the FairViT model, which uses adaptive masking combined with a distance loss to increase the accuracy of the Vision Transformer, while preserving competitive fairness performance. Inspired by this research, our contribution is to reproduce its experiments and offer valuable extensions to examine the robustness of the FairViT model.

In this work, we investigated the following:

> [**Reproducibility Study**]: **Reproduction of the original paper's experiments.** We tried to reproduce the main experiments presented in the original paper, in order to verify the authors' claims as presented in Section 2. Through our experiments we showed that we can only partially reproduce one of these claims.
>
> [**Extended Work**]: **Additional datasets.** As the original paper presents results only on the CelebA dataset, we further evaluated the proposed model on the UTKFace dataset (Zhang & Qi, 2017), examining the robustness of the described method in a different setting.
>
> [**Extended Work**]: **Ablation study.** We assessed the impact of the adaptive masking independently, without the combined usage of the proposed distance loss, as this ablation was not included in the original work.
>
> [**Extended Work**]: **Evaluation of the proposed loss function on additional architectures.** We investigated the extensibility of the proposed distance loss across different model architectures, with the goal of enhancing accuracy. Specifically, we experimented with a pretrained ResNet50 (He et al., 2015) and a Graph Neural Network (GNN) (Zhou et al., 2021) for classification tasks.
>
> [**Extended Work**]: **Codebase extension and enhancements.** We integrated additional input arguments into the code to facilitate easier ablation experiments. Additionally, we created scripts for data processing and included an implementation for visualising the attention heatmaps using Gradient Rollout (Abnar & Zuidema (2020)).

2 Scope of reproducibility

Our work reviews the reproducibility of FairViT: FairVision Transformer via Adaptive Masking of Tian et al. (2024), which proposes a novel approach to balance fairness and accuracy for Vision Transformers, improving accuracy while remaining competitive in terms of fairness when compared to its baseline competitors.

To evaluate these contributions, we examined the main claims of the authors:

- **Higher Accuracy and Competitive Fairness via Adaptive Masking.** FairViT introduces adaptive masks with learnable weights in its attention layers. These group-specific masks are dynamically updated during training, addressing biases inherent in the dataset and improving accuracy for underrepresented groups. As a result, FairViT outperforms SOTA ViT models including Vanilla ViT , TADeT-MMD and TADeT (Sudhakar et al., 2023), FSCL+ and FSCL (Park et al., 2022) — in terms of accuracy. Additionally, it achieves comparable or even superior fairness metrics, as demonstrated by evaluations of the authors on the CelebA dataset.
- **Further accuracy improvement via a novel distance loss function.** The authors propose a distance loss which further improves accuracy when combined with the standard cross-entropy loss. The loss acts as a regulariser, shifting misclassified points towards the desired side of the decision boundary, while maintaining correctly classified points on their original side.
- **Comparable computational efficiency.** FairViT maintains a computational cost comparable to the baseline models, striking a balance between performance with comparable fairness, and computational efficiency.

3 Methodology

To examine the contributions of FairViT, we attempted to reproduce the experimental results reported by Tian et al. (2024). We focused on comparing FairViT with two of the baselines established in the original paper, the Vanilla ViT and FSCL, reporting the models' performance in terms of accuracy, fairness, and computational efficiency. In this section, we provide a comprehensive overview of the used methodology, along with the detailed steps followed during the reproducibility process.

3.1 Model description

3.1.1 The Fair Vision Transformer

FairViT employs a Data-efficient Image Transformer (DeiT) (Touvron et al., 2021), which shares similarities with the original Vanilla architecture (Dosovitskiy et al., 2021), differing mainly on the way it is trained. The proposed model combines adaptive masking with a novel distance loss in order to improve accuracy while maintaining competitive fairness performance.

The objective of FairViT is to train a model f using samples $(\mathbf{x}, s, y)$, where $\mathbf{x}$ is the input, s is a sensitive attribute, and y is the target label. Let θ denote the learnable parameters of the model. Formally, the loss function is minimised:

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

During testing, the sensitive attribute s is treated as unknown and the adapative masking logic is not applied in the linear layers.

Adaptive Masking. To handle a binary sensitive attribute s, the training dataset is divided into G distinct parts. Two groups are defined, corresponding to the presence or absence of the sensitive attribute, and each group is allocated $\lfloor G/2 \rfloor$ parts, ensuring that the same number of parts is included in each group, since the sensitive attributes are (or treated as) binary. The training samples are distributed equally between all parts of each sensitive group, while parts belonging to two different sensitive groups may have a different number of samples. For instance, considering gender as the sensitive attribute, the parts allocated for male should consist of an equal number of samples, but this number can differ from the number of samples allocated to each part belonging to the female group. Subsequently, each of the G parts of the input space is attributed to a learnable mask $M_{l,h,i}$, where l denotes the layer of the transformer, h denotes the head of the multi-attention mechanism and i denotes the specific part of the input space.

In all cases presented below, $\odot$ indicates element-wise multiplication. $\tilde{M}_{l,h}$ denotes the mask matrix, which is derived as the summation of all groups' individual masks, weighted by a set of group-specific learnable coefficients ς_i, where i denotes the group index:

$$\tilde{M}_{l,h} \;=\; \sum_{i=1}^{G} \big(\varsigma_i \, M_{l,h,i}\big) \tag{2}$$

The described distribution of training data and masks in distinct groups with multiple parts ensures that each $M_{l,h,i}$ receives sufficient and balanced training data and is able to adapt on the specific characteristics of its allocated part during training. Thus, this setup enables the masks to capture group-specific characteristics and facilitates the learning of diverse features within each sensitive attribute group, as multiple masks and parts are allocated to each attribute.

By closer inspection of the code implementation, we concluded that the masking is involved in more than one parts of the overall architecture. We present the specific way it is implemented below, in order to provide a more informative view of the overall flow of the proposed method:

The masking method is applied to the weights of all linear layers included in the transformer block. Thus, it appears in four different parts of the transformer block, both in its attention module and its Multi-Layer Perceptron (MLP) module. Within the attention module the masking is applied firstly in the linear

projection of the input $\mathbf{x}$ which yields the queries, keys and values matrices. It is then involved again in the linear projection of the output of the attention operation ($\mathbf{Attn_{l,h}} = softmax(\frac{\mathbf{QK}^T}{\sqrt{d}})\mathbf{V}$). Finally, masking is applied twice within the transformer blocks' MLP module, as part of its two fully connected layers.

In order to obtain the queries, keys and values, as in the common transformer architecture, a linear layer is used, mapping an input $\mathbf{x} \in R^D$ to an output $\in R^{3D}$, which contains the query $\mathbf{q}$, key $\mathbf{k}$ and value $\mathbf{v}$ vectors, each of dimensionality D. Here, the mask matrix $\mathbf{M}$ belongs to the space R^H where H denotes the number of the distinct attention heads in multi-head attention. To obtain each of the vectors $\mathbf{q}$, $\mathbf{k}$ and $\mathbf{v}$, the masking value corresponding to each head is repeated D/H times and multiplied element-wise with the weights of the linear layer. It can be noted that this is equivalent to multiplying each group of D/H output features of a simple linear layer with their corresponding mask value (the mask responsible for these features).

In the case of the linear projection layer, applied to the self-attention outputs, the outputs $\mathbf{Attn_{l,h}}$ are re-weighted by the adaptive masking mechanism for each transformer layer l and head h. More specifically, the linear projection layer maps its input $\mathbf{x} \in R^D$ to an output of the same dimensionality, and the adaptive mask is a matrix $\mathbf{M} \in R^D$ multiplied element-wise with the linear layer's weight matrix of dimensionality $D \times D$ (after proper boradcasting). This is equivalent with multiplying each output feature of the linear layer with the corresponding element in the original mask. This leads to the resulting head attention (HA) which is also described in the original paper, and is defined as follows:

$$\mathrm{HA}(x, M_{l,h}) = Linear(\tilde{M}_{l,h} \odot \mathrm{Attn}_{l,h}(x)), \tag{3}$$

which matches the formulation the authors provide.

Subsequently, the single-head results are concatenated as usually done in the context of multi-head attention. Finally, the fully connected layers of the MLP module map an input $\mathbf{x} \in R^{D_1}$ to R^{D_2}. The masking operation in this case is a matrix $\mathbf{M} \in R^{D_2}$ that is multiplied element-wise with the output features of each layer. Equivalently, with proper broadcasting, the mask tensor is multiplied element-wise with the $D_1 \times D_2$ weights of each fully connected layer.

Distance Loss. In addition to the standard cross-entropy loss used for classification tasks, FairViT uses a distance loss, which encourages misclassified points to move closer to the correct side of the decision boundary, and ensures that the the correctly classified ones will retain. The proposed method fits a logistic regression model on the validation data using as features $\hat{y}$, which is the logit of the correct label, and $\hat{y}_k$, which represents the cumulative sum of the logits of the top-k incorrect labels. The goal is to predict the probability of correct classification of each sample by the ViT.

The decision boundary defined by this model is the following:

$$\hat{y} + \omega\, \hat{y}_k + \beta = 0, \tag{4}$$

Afterwards, the sigmoid function, $S(u) = \frac{1}{1+e^{-u}}$, is used to map logits into probabilities z. The distance loss $\mathcal{L}_{\text{dist}}$ is thus defined as:

$$\mathcal{L}_{\text{dist}} = \begin{cases} -\gamma\, \Phi(\hat{y}, \hat{y}_k), & \text{if } \hat{y} + \omega\, \hat{y}_k + \beta \geq 0, \\ \Phi(\hat{y}, \hat{y}_k), & \text{otherwise.} \end{cases} \tag{5}$$

where γ is a hyperparameter that regulates the influence of the distance Φ of a point $(\hat{y}, \hat{y}_k)$ to the decision boundary:

$$\Phi\big(\hat{y}, \hat{y}_k\big) = \frac{\big|\hat{y} + \omega\, \hat{y}_k + \beta\big|}{\sqrt{1 + \omega^2}}. \tag{6}$$

By penalising samples far away from the decision boundary, this loss encourages the ViT to adjust its weights in order to produce logits that are more likely to yield correct classification results. Combining the cross-entropy loss, $\mathcal{L}_{\text{ce}}$, with the distance loss $\mathcal{L}_{\text{dist}}$, the final training loss is the following:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \alpha\, \mathcal{L}_{\text{dist}} \tag{7}$$

where α is a hyperparameter.

3.2 Datasets

- **CelebA Dataset:** Due to its rich diversity and comprehensive annotations, many computer vision fairness studies rely on the CelebA dataset by Liu et al. (2015). The dataset consists of 202,599 images of 10,177 unique individuals, each of which is annotated for 40 binary attributes including gender, age-related traits and various physical features. These attributes allow for the rigorous evaluation of model biases across diverse demographic groups and are assigned with values ± 1. The dataset can be downloaded from its official website[2].

- **UTKFace Dataset:** The UTKFace dataset (Zhang & Qi, 2017) was used as a further extension of the original experiments to test the robustness of the proposed method. It provides a diverse set of over 20,000 facial images annotated with the attributes of age, gender, and ethnicity, making it suitable for evaluating the generalisability of the FairViT approach. The official dataset is available for download[3].

- **AIDS Dataset:** The AIDS dataset consists of 1999 graphs introduced in Riesen & Bunke (2008). It contains compounds checked for evidence of anti-HIV activity and is thus ideal for binary classification tasks.

- **PROTEINS Dataset:** PROTEINS is a dataset consisted of 1113 graphs introduced in Borgwardt et al. (2005). It is used for molecular property prediction and particularly for predicting whether molecules are enzymes or not.

3.3 Hyperparameters

In order to reproduce the results of the original paper, we used the default training hyperparameters as reported by the authors of FairViT. In the case of FSCL, we used the default hyperparameters set in its official GitHub repository, since we did not find specific settings for running this baseline in the FairViT paper. Regarding the ablation study of the distance loss in different architectures, for both the ResNet and the GNN we used the AdamW optimiser. For the former, we opted for a learning rate of 0.0005, whereas for the latter a learning rate of 0.01 with weight decay equal to 0.0005 was preferred. Overall, the required hyperparameters to validate our results are reported in our study's GitHub repository.

3.4 Experimental setup and code

For reproducing the experiments, we relied on the open-source implementation that the authors provide on GitHub [4]. We adapted the code in order to be able to also use a Vanilla ViT with the same pre-trained weights as those provided by default to FairViT (pretrained DeiT weights). For running the FSCL baseline we used the publicly available source code [5]. We integrated the provided model implementation in our final repository, performing only minor changes in order to calculate for FSCL's predictions the same fairness metrics used in FairViT. We note that it was not clear whether the authors pretrained the backbone ResNet used in FSCL, using the full training dataset or just the images of the first 80 individuals. For this reason, we decided to report the performance metrics of FSCL when pretrained on the full dataset to ensure fair comparison, as it is trained from scratch and does not use an already pretrained model, unlike ViT variants. Additionally, we trained FSCL using the small dataset to report its computational cost, alongside the version pretrained on the full dataset, as the authors mentioned they used the small dataset for pretraining. Moreover, regarding the FSCL baseline, we chose to run only one of the three experiments. The reason behind this is that we did not observe a better performance of FairViT compared to the Vanilla model, so further comparisons with even stronger baselines would be redundant given the computational and environmental cost of experiments.

For experimenting with the CelebA dataset, the authors' implementation does not load the official train and validation splits and instead performs a 90-10% split dynamically. We thus merged the train and validation

[2] https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[3] https://susanqq.github.io/UTKFace/
[4] https://github.com/abdd68/Fair-Vision-Transformer
[5] https://github.com/sungho-CoolG/FSCL

sets of CelebA and filtered the merged dataset to only keep the images of the first 80 celebrities for training and validation (resulting to approximately 1700 images), following the original paper. For testing we used the officially provided test set as a whole (19,962 images). We created custom scripts for performing the described data merging and filtering, as these were not included in the provided code.

As the UTKFace dataset is not officially split into train, validation and test parts, we performed a custom split into 10,000, 2,400, and 2,400 images for the training, validation, and test set respectively. In order to test for fairness, we introduced an artificial bias by setting the ratio of the male to female samples to 1:4 for the white race group (1,000 female and 4,000 male samples) and to 4:1 for the non-white group (4,000 female and 1,000 male samples). The test and validation sets were kept completely balanced.

Regarding the AIDS and PROTEINS datasets, we opted for a 0.8:0.1:0.1 split for training, validation and testing respectively. Both datasets were downloaded from the TUDataset repository (Morris et al., 2020).

Additionally, we adapted the code to include an implementation for calculating the Gradient Attention Rollout (Abnar & Zuidema, 2020) for chosen test images, as it was missing from the original paper's code. For this purpose, we adapted the open source implementation available at [6].

The exact method the authors used for selecting the model for reporting results was not clear. However, in the provided code, we observed that, for each metric, the best score across all epochs is saved and reported in the logs, independently from all other metrics. As we are not certain whether these are the metrics indeed reported in the paper and we believe this method would not provide a realistic insight into the models' performance, since it does not select one specific trained instance as common for all metrics, we selected the best model for each run based on the best accuracy achieved on the validation set across all epochs.

3.5 Evaluation Metrics

Evaluation Metrics. To evaluate the performance and fairness of the proposed approach, we used the same performance and fairness metrics as the authors did in the FairViT study. In particular, we evaluated performance using accuracy, and examined fairness using the following metrics:

- **Balanced Accuracy (BA)**: Adjusts for class imbalance by averaging the recall for each class, measuring how balanced performance is in terms of accuracy across groups:

$$\text{BA} = \frac{1}{4}\left(\text{TPR}_{s=0} + \text{TNR}_{s=0} + \text{TPR}_{s=1} + \text{TNR}_{s=1}\right) \tag{8}$$

- **Demographic Parity (DP)** (Hardt et al., 2016): Formally, it measures the extent to which the probability of a positive prediction is the same across all sensitive groups:

$$P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = b), \quad \forall a, b \in \text{Sensitive Groups} \tag{9}$$

- **Equalised Opportunity (EO)** (Hardt et al., 2016): Measures how "close" is the true positive rate (TPR) across all sensitive groups:

$$P(\hat{Y} = 1 \mid Y = 1, A = a) = P(\hat{Y} = 1 \mid Y = 1, A = b), \quad \forall a, b \in \text{Sensitive Groups} \tag{10}$$

3.6 Computational requirements

All experiments were performed on an NVIDIA A100 GPU. Training FairViT for a specific target attribute for 30 epochs took approximately 25 minutes for CelebA, whereas Vanilla ViT needed approximately 20 minutes. In total, our reproducibility study along with all the ablations and extensions added up to approximately 40 GPU hours. For the Carbon Intensity, we used the most recent value provided by the European Environment Agency, which states that generating one kWh of electricity in the Netherlands, in 2023, emitted an amount of greenhouse gases equivalent to 263 grams of CO2 [7]. All our experiments were ran using

[6] https://github.com/jacobgil/vit-explain.git

[7] https://www.eea.europa.eu/en/analysis/indicators/greenhouse-gas-emission-intensity-of-1?activeAccordion=

the Snellius infrastructure located in the Amsterdam Data Tower. The Snellius datacenter reports a Power Usage Effectiveness (PUE) of 1.19 [8]. Using the Machine Learning Emissions Calculator (Lacoste et al., 2019) we estimate that our experiments emitted approximately 3.2 kg of CO_2.

4 Results

As outlined in Section 2, the original paper presents three main claims. Our study partially validated the second claim regarding the effectiveness of the proposed distance loss. However, we could reproduce neither the first and most important claim nor the third claim related to FairVit's computational efficiency.

4.1 Results reproducing original paper

4.1.1 Higher Accuracy and Competitive Fairness via Adaptive Masking

To verify the first claim, we loaded the pretrained weights of the DeiT model into an instance of the FairViT model and fine-tuned it on the CelebA dataset using the proposed method. We then compared its performance with the Vanilla model, a Vision Transformer instance initialized with the same pretrained DeiT weights and fine-tuned solely using cross-entropy loss.

The fine-tuning of the models was performed by using images of the first 80 celebrities, as specified by the authors. The three tasks examined involve predicting a person's expression or attractiveness with gender as the sensitive attribute and predicting whether someone is attractive with hair colour as the sensitive attribute. The results presented in Table 1 correspond to evaluations carried out on the full test set replicating the experiments of the authors. Moreover, Table 2 presents the replication of the ablation study conducted by the authors, offering insights into the benefits of using the proposed adaptive masking mechanism in combination with the distance loss.

In contrast to the findings of the original paper, our results did not show significant improvements in accuracy or fairness metrics when the adaptive masks were employed. Specifically, Table 1 demonstrates comparable performance between the Vanilla and FairViT models, with each model performing better in certain metrics but no significant overall improvements observed for FairViT. Furthermore, upon inspecting the results in Table 2, no clear benefits can be attributed to adding a constant mask or subsequently making it trainable.

To further compare the Vanilla and FairViT models qualitatively, we also provide a visualisation of the Gradient Attention Rollout for randomly selected test images for the two models, in Figure 1. In contrast to the original paper, and while minor differences in the Gradient Attention Rollout may be seen for individual images, we do not observe a significant differences between the two models, in terms of the relevancy of the areas information is extracted from.

Table 1: Accuracy (ACC), Balanced accuracy (BA), Equal Opportunity (EO) and Demographic Parity (DP) for the three classification tasks on CelebA, averaged over three independent runs, for the Vanilla ViT, FCSL models. **Y** denotes the target attribute and **S** denotes the sensitive attribute.

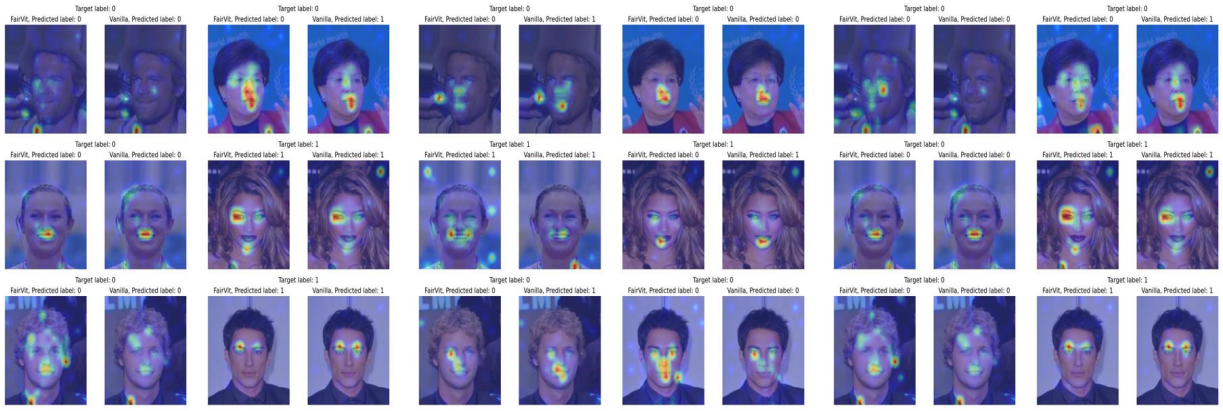
Model	**Y:** Attraction, **S:** Gender				**Y:** Expression, **S:** Gender				**Y:** Attraction, **S:** Hair Colour			
	ACC% ↑	BA% ↑	EO_{e-2} ↓	DP_{e-1} ↓	ACC% ↑	BA% ↑	EO_{e-2} ↓	DP_{e-1} ↓	ACC% ↑	BA% ↑	EO_{e-2} ↓	DP_{e-1} ↓
Vanilla	78.05	73.22	24.80	4.519	89.61	89.19	4.88	1.614	77.94	75.22	4.03	1.822
FSCL	80.353	73.33	40.12	4.904	-	-	-	-	-	-	-	-
FairViT	78.20	72.82	29.01	4.471	89.71	89.26	5.22	1.655	77.89	75.75	5.43	1.805

4.1.2 Further accuracy improvement via a novel distance loss function

To validate the second claim, we replicated the ablation study of the original paper and present its results in Table 2. This study focuses, among others, on isolating the impact of the distance loss from the adaptive

[8] `https://www.clouvider.com/amsterdam-data-tower-datacentre/`

Figure 1: Gradient Attention Rollout comparison for Vanilla and FairViT models. Across all three experiements no consistent differences are to be observe between the two models' Gradient Attention Rollouts.



Y: Attraction, **S:** Gender | **Y:** Expression, **S:** Gender | **Y:** Attraction, **S:** Hair

masking, allowing us to analyse the effectiveness of this specific component independently. By comparing the results of L_{ce} with $L_{ce}+L_{dist}$ in terms of accuracy, only a slight improvement is observed in two scenarios: (1) when the target attribute is expression and the sensitive attribute is gender, where an improvement of 0.21 percentage points is observed and (2) when the target attribute is attractiveness and the sensitive attribute is hair colour, where an improvement of 0.08 percentage points is observed. However, this improvement is minor and falls within the margin of statistical error, making it difficult to draw clear conclusions about the effectiveness of the distance loss. For this reason, we decided examine this module on a different architecture, as presented in Section 4.2.

Table 2: Accuracy (ACC), Balanced accuracy (BA), Equal Opportunity (EO) and Demographic Parity (DP), averaged over three independent runs, for different ablations of the FairViT architecture. **Y** and **S** denote the target and sensitive attribute, L_{ce} and L_{dist} denote the cross-entropy and distance loss and Θ, $\Delta\Theta$ denote the presence of adaptive masks, non-trainable or trainable respectively.

Model	**Y:** Attraction, **S:** Gender				**Y:** Expression, **S:** Gender				**Y:** Attraction, **S:** Hair Colour			
	ACC% ↑	BA% ↑	EO_{e-2} ↓	DP_{e-1} ↓	ACC% ↑	BA% ↑	EO_{e-2} ↓	DP_{e-1} ↓	ACC% ↑	BA% ↑	EO_{e-2} ↓	DP_{e-1} ↓
L_{ce}	78.05	73.22	24.80	4.519	89.61	89.19	4.88	1.614	77.94	75.22	**4.03**	1.822
$L_{ce} + \Delta\Theta$	77.92	72.53	28.68	4.543	89.86	89.27	6.91	1.812	77.62	75.44	5.85	1.846
$L_{ce} + L_{dist}$	78.02	73.38	**23.31**	4.528	89.82	89.42	**4.45**	**1.563**	78.02	75.22	4.71	1.901
$L_{ce} + L_{dist} + \Theta$	**78.24**	**73.39**	25.00	4.519	**90.19**	**89.71**	5.39	1.64	**78.24**	**75.81**	4.91	1.842
$L_{ce} + L_{dist} + \Delta\Theta$	78.20	72.82	29.01	**4.471**	89.71	89.26	5.22	1.655	77.89	75.75	5.43	**1.805**

4.1.3 Comparable computational efficiency

To investigate the third claim, we monitored the time cost per epoch for both the FairViT model and its baselines. The results, presented in Table 3, show that the FairViT model requires almost double time per epoch compared to the Vanilla model. Thus, its computational cost is not comparable to that of the Vanilla model, in contrast to the authors' claim. Moreover, we only trained the FSCL model with attractiveness as the target attribute and gender as the sensitive attribute, due to limited resources. When we used the same dataset as for FairViT (the 80 first celebrities of CelebA) for the contrastive pretraining, we observed a time cost per epoch smaller than that of FairViT, as reported in Table 3. We did notice a significant increase in the required time per epoch when performing contrastive pretraining on the full CelebA train set. However, this obviously happens as a result of the increased train set size and not because of the differences in architecture between the models.

Table 3: Training time per epoch (in seconds), averaged over three independent runs, for different ablations of each architecture. For FSCL, only one run was conducted. FSCL(80) denotes training with the images of the 80 first celebs and FSCL(Full) denotes training with the full dataset.

Model	**Y:** Attraction, **S:** Gender | time (sec)	**Y:** Expression, **S:** Gender | time (sec)	**Y:** Attraction, **S:** Hair Colour | time (sec)
Vanilla	**12.20**	**12.18**	**12.21**
FairViT	22.03	21.89	22.01
FSCL(80)	10	-	-
FSCL(Full)	1080	-	-

4.2 Results beyond original paper

4.2.1 Additional datasets

Motivated by our inability to reproduce results that clearly support the authors' claims, we decided to evaluate the proposed method and compare it to the Vanilla ViT on a different dataset, namely the UTKFace dataset, in order to verify our findings.

By intentionally creating an imbalanced training dataset, we investigated the models' accuracy and fairness in predicting gender, with race as the sensitive attribute. Since race consists of five distinct values in this dataset, we simplified the analysis by treating it as a binary attribute, grouping the input samples into "white" and "non-white" categories. Based on the obtained results, as demonstrated on Table 4, we did not observe improvements in any metrics when using the FairViT architecture, with the differences between the Vanilla and the FairViT architecture falling within the margin of statistical error.

Table 4: Accuracy (ACC), Balanced accuracy (BA), Equal Opportunity (EO) and Demographic Parity (DP) for one classification tasks on UTKFace, averaged over two independent runs, for the Vanilla and FairViT models. **Y** denotes the target attribute and **S** denotes the sensitive attribute.

Model	**Y**: Gender, **S**: Race			
	ACC% ↑	BA% ↑	EO_{e-2} ↓	DP_{e-1} ↓
Vanilla	**94.48**	**94.47**	**8.00**	**0.837**
FairViT	94.20	94.25	9.17	0.875

4.2.2 Evaluation of the proposed loss function on additional architectures

In order to determine the extensibility and effectiveness of the proposed distance loss in different settings and architectures, we decided to perform experiments by fine-tuning the pre-trained ResNet50 (He et al., 2015) on the CelebA dataset for the same classification tasks as the FairViT model. We also experimented with using the distance loss for graph classification using a simple Graph Neural Network (GNN) consisted of three Graph Convolutional Layers (GCN). The experiments were performed with and without using the distance loss on top of the cross-entropy loss traditionally used for classification tasks. Tables 5 and 6 present the obtained accuracy when fine-tuning the pre-trained ResNet model and when training the GNN model respectively, using only the standard cross-entropy loss and using a combination of the cross-entropy and the proposed distance loss. For interested readers, we additionally include a study of the effect of different values of α and γ for the two models in Appendix B.

We observe an improvement of 1 percentage point in accuracy for the task of predicting attractiveness and an improvement of 3 percentage points for predicting expression using the ResNet50 model. In the case of the graph classification task, we observe an improvement of 1.17 percentage points and 2.09 percentage points for the AIDS and PROTEINS datasets respectively, when the optimal configuration of α and γ was

used. These results validate to some extent the claim of the authors regarding the beneficial effects of the distance loss to overall accuracy.

Table 5: Accuracy results for training the ResNet50 model without and with the distance loss. Shown is the mean of 3 independent runs. Highlighted is the best result.

γ	**Y:** Attraction ACC% ↑	**Y:** Expression ACC% ↑
L_{ce}	76.44	87.41
$L_{ce} + L_{dist}$	**77.59**	**90.59**

Table 6: Accuracy results for training a GNN model without and with the distance loss, in the AIDS and PROTEINS datasets. Shown is the mean of 3 independent runs. Highlighted is the best result.

γ	**AIDS** ACC% ↑	**PROTEINS** ACC% ↑
L_{ce}	82.33	67.26
$L_{ce} + L_{dist}$	**83.50**	**69.35**

5 Discussion

In this work, we studied the reproducibility of "FairViT: Fair Vision Transformer via Adaptive Masking". Overall, we could not successfully reproduce the main claim of the authors and were only able to partially replicate one of the remaining claims.

The primary claim concerns the capability of adaptive masks to improve accuracy and maintain fairness for the examined classification tasks. The original paper reported a significantly higher accuracy (5-10 percentage points compared to the Vanilla model) and better fairness metrics for the FairViT model for all three presented classification tasks. However, we observed no clear improvements achieved by FairViT compared to its baselines, neither when the adaptive masks are combined with the proposed distance loss nor when they are used separately with only the standard cross-entropy loss (see Tables 1 and 2). The lack of performance improvement when using adaptive masks could likely be attributed to the fact that the group-specific masks may not be highly informative, as they are averaged before being applied and only receive small updates during training, thus leading to largely unaffected model performance.

Examining the proposed distance loss function, we managed to only partially validate the second claim of the authors. As presented in Table 2, our results do not show a consistent accuracy improvement when using the distance loss for the CelebA dataset in the context of FairViT. However, to further examine the validity of this claim, we studied the effect of the proposed loss function in two additional architectures, namely a ResNet and a GNN. In this case, the distance loss showed a greater benefit, achieving a performance increase of approximately 1-3 percentage points (see Tables 7, 8, 9 and 10). We thus conclude that the distance loss can indeed be effective in boosting accuracy, with its benefits being dependent on the employed architecture. Further studying its effect on different tasks and models would therefore be of great interest for future work.

Our results regarding the time cost of the FairViT model do not validate the claim of the authors that its computational complexity is comparable to the Vanilla ViT. In particular, in Table 3 we can see that its time cost is almost double per epoch. Additionally, when comparing the time taken per epoch for the training of FairViT with that of the FSCL model, we did not observe the difference described in the original paper, when training both models on the same data. In contrast, our experiments showed that the per-epoch training time for FSCL was not only comparable to, but in some cases even lower than that of FairViT. This stands in stark contrast to the original paper, where FSCL was reported to be approximately six times slower than FairViT. Given this significant discrepancy, we suspect that the authors may have used a different method

to measure FSCL's computational cost, trained the model on a larger subset of the dataset, or employed a different configuration.

5.1 What was easy

While trying to reproduce the experiments of the authors, we found some parts that were straightforward. The selected dataset was easy to download and process, the code was publicly available on GitHub and the hyperparameters with which FairViT was trained were reported in the original paper.

5.2 What was difficult

During our reproducibility study, we encountered several challenges and inconsistencies related to the environment, code, and documentation provided by the authors. More specifically, some packages specified in the environment setup were either unavailable or incompatible with the provided code. Additionally, the instructions provided in the GitHub repository for integrating the dataset with the code did not correspond to the provided data parsing code sections, while the structure the dataset should have to be correctly parsed by the code was not clearly presented neither in the paper nor in the code repository and comments.

We also noticed discrepancies between the default hyperparameter values in the code and the values mentioned in the paper. Some parts, such as the implementation of Gradient Attention Rollout, were missing from the available code and had to be integrated by us. In addition, the proposed adaptive masking mechanism was used in several parts of the architecture that were not described in the paper, making the understanding of the model's behaviour challenging.

Moreover, the ablation study in Table 2 of the original paper lacked information about the initialisation of the non-trainable masks, while code adaptations were needed in order to support running all ablations without hard-coding changes in the code. Finally, no details about how the Vanilla Vision Transformer was executed were provided, and the dataset size used for training the FSCL model was not specified in the original paper.

5.3 Communication with original authors

The correspondence with the authors was adequate. They willingly answered our questions regarding the structure and the split of the CelebA dataset, the way the baseline models were run, the mechanism of the adaptive masking, and the reported time cost.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

Karsten M. Borgwardt, Cheng Soon Ong, Stefan Schönauer, S. V. N. Vishwanathan, Alex J. Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(1):47–56, January 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti1007. URL https://doi.org/10.1093/bioinformatics/bti1007.

Jannik Brinkmann, Paul Swoboda, and Christian Bartelt. A multidimensional analysis of social biases in vision transformers, 2023. URL https://arxiv.org/abs/2308.01948.

André Cruz and Moritz Hardt. Unprocessing seven years of algorithmic fairness. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jr03SfWsBS.

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, 2021. URL https://arxiv.org/abs/2010.11929.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016. URL https://arxiv.org/abs/1610.02413.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Faisal Kamiran and Toon Calders. Data pre-processing techniques for classification without discrimination. *Knowledge and Information Systems*, 33, 10 2011. doi: 10.1007/s10115-011-0463-8.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning, 2019. URL https://arxiv.org/abs/1910.09700.

Bryson Lingenfelter, Sara R. Davis, and Emily M. Hand. A quantitative analysis of labeling issues in the celeba dataset. In George Bebis, Bo Li, Angela Yao, Yang Liu, Ye Duan, Manfred Lau, Rajiv Khadka, Ana Crisan, and Remco Chang (eds.), *Advances in Visual Computing*, pp. 129–141, Cham, 2022. Springer International Publishing. ISBN 978-3-031-20713-6.

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.

Abhishek Mandal, Susan Leavy, and Suzanne Little. Biased attention: Do vision transformers amplify gender bias more than convolutional neural networks?, 2023. URL https://arxiv.org/abs/2309.08760.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+ 2020)*, 2020. URL www.graphlearning.io.

Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification, 2022. URL https://arxiv.org/abs/2203.16209.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf.

Yao Qiang, Chengyin Li, Prashant Khanduri, and Dongxiao Zhu. Fairness-aware vision transformer via debiased self-attention, 2024. URL https://arxiv.org/abs/2301.13803.

Kaspar Riesen and Horst Bunke. Iam graph database repository for graph based pattern recognition and machine learning. In Niels da Vitoria Lobo, Takis Kasparis, Fabio Roli, James T. Kwok, Michael Georgiopoulos, Georgios C. Anagnostopoulos, and Marco Loog (eds.), *Structural, Syntactic, and Statistical Pattern Recognition*, pp. 287–297, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-89689-0.

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. ISSN 2835-8856. URL https://openreview.net/forum?id=TyBd56VK7z.

Sruthi Sudhakar, Viraj Prabhu, Arvindkumar Krishnakumar, and Judy Hoffman. Mitigating bias in visual transformers via targeted alignment, 2023. URL https://arxiv.org/abs/2302.04358.

Bowei Tian, Ruijie Du, and Yanning Shen. Fairvit: Fair vision transformer vianbsp;adaptive masking. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXV*, pp. 451–466, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73649-0. doi: 10.1007/978-3-031-73650-6_26. URL https://doi.org/10.1007/978-3-031-73650-6_26.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers distillation through attention, 2021. URL https://arxiv.org/abs/2012.12877.

Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H. Chi. Understanding and improving fairness-accuracy trade-offs in multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery amp; Data Mining*, KDD '21, pp. 1748–1757. ACM, August 2021. doi: 10.1145/3447548.3467326. URL http://dx.doi.org/10.1145/3447548.3467326.

Song Yang Zhang, Zhifei and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications, 2021. URL https://arxiv.org/abs/1812.08434.

A Grad Attention Rollout

Gradient Attention Rollout (GAR) (Abnar & Zuidema, 2020) is used for the interpretation of the attention mechanism in ViTs, through the generation of heat maps that measure the contribution of each input patch withing to the attention mechanism's final output. It is mathematically defined as:

$$A_l = \begin{cases} A_l(\mathbf{x}) \frac{\partial \hat{y}}{\partial A_l(\mathbf{x})} A_{l-1}, & \text{if } l > 0 \\ A_l(\mathbf{x}) \frac{\partial \hat{y}}{\partial A_l(\mathbf{x})}, & \text{if } l = 0 \end{cases} \tag{11}$$

where A_l represents the GAR at the l-th layer of the transformer, and $A_l(\mathbf{x})$ is the attention map at that layer for the input $\mathbf{x}$. The heat map is constructed by assigning the value $A_N^{0,i}$ to the i-th patch in the image, where A_N indicates the GAR of the final layer, and $A_N^{0,i}$ corresponds to the element in the first row and i-th column of the matrix A_N.

B The effect of hyperparameters α and γ on Distance Loss

Tables 7 and 9 illustrate the impact of varying the hyperparameter α, which regulates the contribution of the distance loss to the overall training loss, on the performance of the ResNet50 and GNN models. The

results indicate that the ResNet50 model tends to benefit more from higher values of α, while the GNN model performs better when the distance loss has a lower relative weight.

Tables 8 and 10 present the achieved accuracy for different values of the hyperparameter γ, for the model with the best performing α, for the ResNet50 and GNN models. A value of γ equal to 0.7 leads to the best performance for both tested experiments for the ResNet50 case. For the GNN model, results are dataset-dependent, with a larger value of 0.7 or 0.9 being preferable for the AIDS dataset and a smaller value of 0.1 leading to significantly higher accuracy for PROTEINS.

Table 7: Impact of α in the distance loss for ResNet50. Shown is the mean of 3 independent runs. Highlighted is the best result.

α	**Y:** Attraction ACC%	**Y:** Expression ACC%
No distance loss / 0	76.44	87.41
0.001	76.89	88.53
0.01	76.57	89.33
0.1	76.72	**90.13**
1	**77.33**	90.01

Table 8: Impact of γ in the distance loss. Shown is the mean of 3 independent runs. Highlighted is the best result.

γ	**Y:** Attraction ACC%	**Y:** Expression ACC%
No distance loss	76.44	87.41
0.1	75.91	89.66
0.3	76.23	88.88
0.5	77.33	90.01
0.7	**77.59**	**90.59**
0.9	77.47	89.66

Table 9: Impact of α in the distance loss for the GNN in the AIDS and PROTEINS datasets. Shown is the mean of 3 independent runs. Highlighted is the best result.

α	**AIDS** ACC%	**PROTEINS** ACC%
No distance loss / 0	82.33	67.26
0.001	**83.00**	66.97
0.01	81.17	**68.75**
0.1	82.00	66.67
1	80.50	66.67

Table 10: Impact of γ in the distance loss for the GNN in the AIDS and PROTEINS datasets. Shown is the mean of 3 independent runs. Highlighted is the best result.

γ	**AIDS** ACC%	**PROTEINS** ACC%
No distance loss	82.33	67.26
0.1	83.17	**69.35**
0.3	83.33	66.37
0.5	83.17	64.29
0.7	**83.50**	66.37
0.9	**83.50**	64.58