Fair Face Recognition

ZHOU ZHOU* and YUNQING ZHU*, New York University, USA

Face recognition is a popular technology that allows people to unlock their phones and obtain passwords. Law enforcement, security departments, and many mobile apps use this technology to verify an individual's identity. However, some papers and researchers indicate that some face recognition systems perform better for certain demographic groups, for example, young and light-skinned males. In this case, we explore two different methods to enhance fairness in face recognition tasks. The first approach employs Masked Autoencoders (MAE), leveraging their ability to learn robust and unbiased feature representations through partial image reconstruction. The second method utilizes a self-supervised learning framework. We chose the UTKFace dataset to evaluate fairness performance and focus on mitigating biases related to age, gender and race. Experimental results demonstrate that both techniques effectively reduce bias. Each method offers unique strengths in fairness enhancement. Our findings highlight these methods in fostering equitable outcomes in face recognition systems.

CCS Concepts: • **Security and privacy** → **Formal security models**.

Additional Key Words and Phrases: Computer Vision, Fairness Face Model, Self-supervised Learning, Deep Learning

ACM Reference Format:
Zhou Zhou and Yunqing Zhu. 2024. Fair Face Recognition. 1, 1 (December 2024), 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Face recognition is a popular technology that allows people to unlock their phones and obtain passwords. Law enforcement, security departments, and many mobile apps use this technology to verify an individual's identity. However, some papers and researchers indicate that some face recognition systems perform better for certain demographic groups, for example, young and light-skinned males.

In this paper, we did two research about enhancing fairness in face recognition tasks. The first one is proving the effect of self-supervised on face recognition tasks. Self-supervised learning is an approach where a model is trained to predict parts of its input data based on other parts, enabling it to learn representations without explicit labels. This technique has shown great potential in improving model robustness and fairness. The improvement of self-supervised learning on fairness has been shown in previous research and we did more work to show its performance related to image classification and face recognition tasks. After showing the effect of self-supervised learning on image classification tasks, we demonstrate a novel architecture of the Masked Autoencoders (VAE) model based on a dynamic masking method called the sensitivity-awareness masking method.

*Both authors contributed equally to this research.

Authors' Contact Information: Zhou Zhou, zz3382@nyu.edu; Yunqing Zhu, yz9661@nyu.edu, New York University, New York, New York, USA.

ACM XXXX-XXXX/2024/12-ART
https://doi.org/XXXXXXX.XXXXXXX

This method aims to adjust the masking strategy of the VAE model to enhance fairness without significantly increasing computational complexity. We also provide suggestions on utilizing pre-trained models in our methods, which can further improve the fairness of the final model.

To evaluate our approaches, we chose the UTKFace dataset[10] to evaluate fairness performance. Experimental results demonstrate that both techniques effectively reduce bias. Each method offers unique strengths in fairness enhancement. Our findings demonstrate the effectiveness of these methods in reducing bias while maintaining or improving classification performance. This study provides new insights into the development of fairer face recognition systems, paving the way for broader applications in bias mitigation and ethical AI development.

2 Related Work

Face recognition applications are already involved in daily life. However, many people still do not realize model bias across demographic groups. Some early approaches try to solve this issue based on dataset balancing. For example, oversampling underrepresented groups or augmenting the dataset with synthetic examples. However, these methods cannot effectively generalize to unseen datasets. Many researchers offer some models to improve the fairness in computer vision tasks. Felix, et al[3] offer a Fair Diffusion to attenuate biases after the deployment of generative text-to-image models. Moreno, et al[2] use Vision-Language Driven Image Augmentation to improve fairness. Furthermore, Kimmo, et al[7] construct a novel face image dataset that contains 108,501 images with an emphasis on balanced race composition in the dataset.

The original purpose of self-supervised learning (SSL) is to learn from unlabeled data, especially to solve small datasets without enough training samples. Self-supervised learning shows its advantages in efficiency in utilizing all available data and reducing fine-tuning time for a specific task. A notable advancement in SSL is the Masked Autoencoder (MAE) [4], which utilizes masking methods to learn representations of an image. Building on this foundation, AutoMAE [1] uses an adversarially trained mask generator and a mask-guided image modeling process to improve the efficiency of the model through a different masking method. However recent research is developing the potential of self-supervised learning in aspects other than efficiency and accuracy. Minimizing humans' role in training could make deep learning models more human-like, improving the generalization of the models. Self-supervised learning has emerged as a promising paradigm for improving fairness. Dan, et al[9] indicate that SSL can significantly improve model fairness while maintaining performance on par with supervised methods in some datasets. When extra data is not available, self-supervised often generates lower accuracy compared with traditional methods. However, research shows that there is a trade-off between accuracy and fairness. The performance of self-supervised learning can be better than only measuring accuracy.

3 Methods

3.1 Residual Networks in Face Recognition

Kaiming He et al.[5] introduced residual networks. It has revolutionized deep learning by addressing the degradation problem in training deep neural networks. The ResNet architecture uses skip connections to pass input features directly to deeper layers. It allows the network to learn residual mappings rather than directly approximating the underlying functions. In this case, we select ResNet-50 as the baseline model to evaluate fairness and accuracy on the UTKFace dataset. Figure 1 illustrates the backbone structure of the ResNet model. Then, we construct the ResNet-50 model and train it on the UTKFace dataset to analyze its performance.

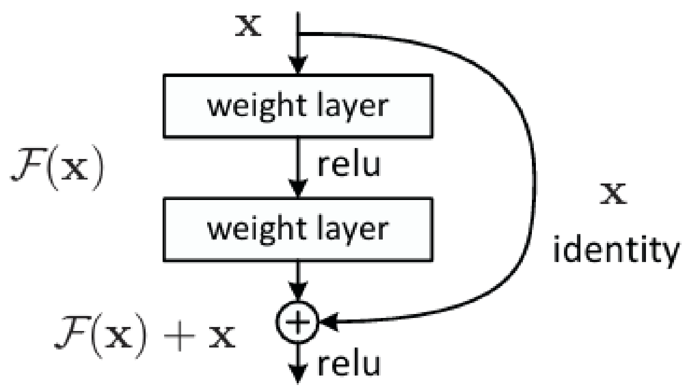


Fig. 1. Residual learning: a building block [5].

3.2 Self-supervised Learning model

We use a TFC model[9] to improve fairness in face recognition. This approach leverages all representations of face images to learn robust and unbiased embeddings. The TFC model consists of three main components. First, it uses a convolutional neural network to extract spatial features from input images. Then, this model incorporates two Transformer Encoder modules, one for temporal-space encoding and the other for frequency-space encoding. Each encoder can capture contextual dependencies within the input. Finally, we use fully connected layer to learn lower-dimensional space. We also use a simple multitask target classifier to implement downstream tasks. It will find the age, gender and race of images. Besides, we also follow Figure 2 process to implement fine-tuning steps to achieve the best model performance.

3.3 Sensitivity-Awareness Masking Method

We demonstrated that self-supervised learning methods contribute to improving model fairness, supported by related research. Building on this foundation, we propose an enhanced approach leveraging Masked Autoencoders (MAE)[4]. Recent studies have explored dynamic strategies for masking input images. For instance, AutoMAE[1] employs an adversarial network to predict masking patterns, allowing the MAE model to prioritize more challenging regions of the image. Inspired by this, we introduce a novel sensitivity-awareness masking method, which detects the sensitive features from comprehensive inputs like images and allows model trainers to select the way to adjust masking methods based on these features. This method also eliminates the need for training an adversarial network during the pre-training process. This approach identifies sensitive regions within images and dynamically adjusts the masking strategy during MAE pre-training based on the sensitivity of the detected regions.

3.4 Heatmaps evaluating

We utilize heatmap-based analysis in Vision Transformers (ViTs) to detect sensitive parts of the image. Visualizing attention heatmaps is a popular method to explain the attention mechanism in transformer-based models. By extracting and analyzing attention maps from self-attention layers,

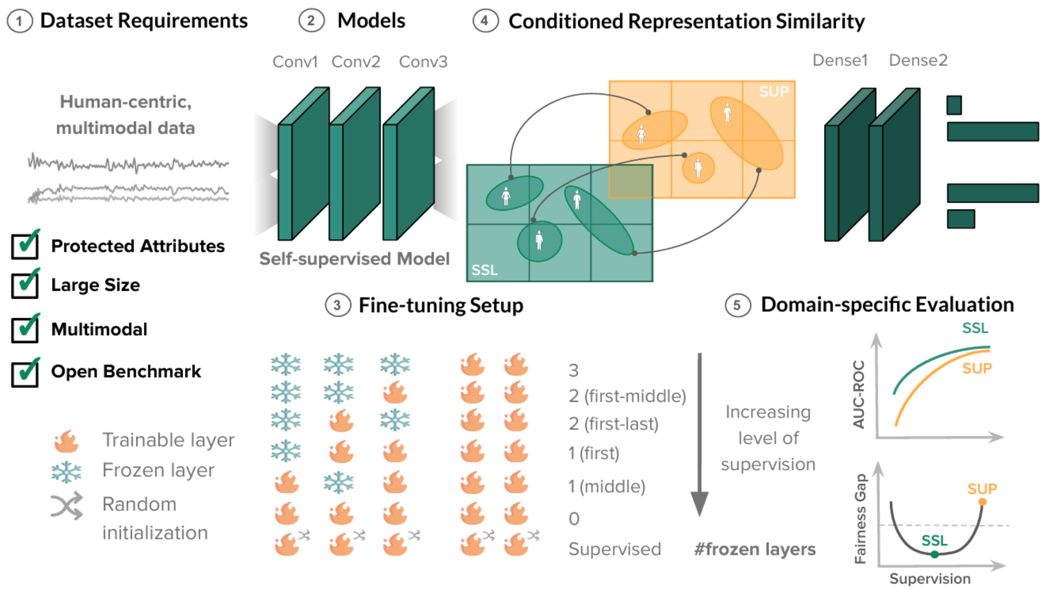


Fig. 2. Overview of the proposed fairness assessment framework for SSL [9].

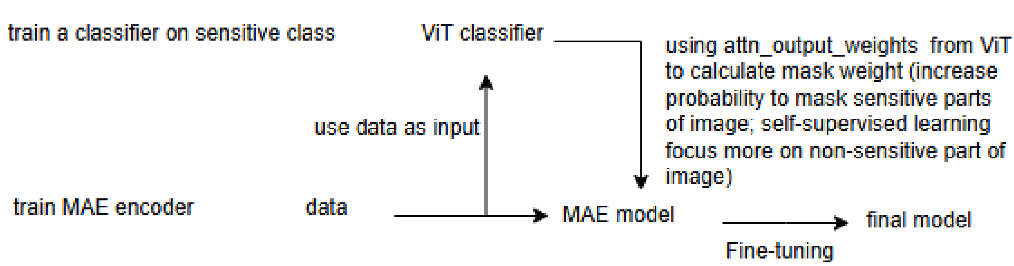


Fig. 3. The pipeline of the sensitivity-awareness masking method

attention heatmaps can reveal the importance of specific image regions during the prediction process. We try to develop a method utilizing attention heatmaps during the training process instead of only as an explanation tool. Firstly, we trained a sensitive classifier on sensitive features. We try to keep the architecture of this classifier similar to the main MAE models. We collected the attention heatmap from this model when predicting sensitive features. The attention map is calculated by averaging attention maps from all attention layers and all batches. By analyzing this attention map, we can calculate the weight of every patch when predicting the sensitive feature. We use the average weight of all patches because, in the training process of the MAE model, all images in a mini-batch will share the same mask to increase efficiency. Also, using the average weight masking method still contains randomness. As a result, we only adjust the probability of each patch learned by the MAE model instead of discarding a part of the images. From the view of the model, the patch with a higher score means that this patch is more informative when classifying sensitive features. In other words, this patch of image is more sensitive.

3.5 Applying the heatmap in the training process

After extracting the sensitive indices of the image parts, we sorted the weight from the heatmap and applied those to the MAE model. While the original MAE paper employs random masking, we propose two masking methods utilizing the weight extracted from the attention heatmap: (1) masking more sensitive parts and (2) masking non-sensitive parts.

3.5.1 Masking sensitive parts. Masking sensitive parts of the image can help the model learn more about non-sensitive parts of the image. This suggests the model utilizes features other than sensitive features to finish the classification. In our age prediction task, although the face is the most important feature containing both sensitive information and the target information, we can still predict one person's age based on other information like clothing. This masking method indicates that the model can learn to rely more frequently on non-sensitive features during prediction, potentially enhancing fairness.

3.5.2 masking non-sensitive parts. Masking non-sensitive parts of the images can help the model spend more time learning sensitive parts of the image. This is similar to the ideas from AutoMAE, focusing on a more informative part of the image. Instead of attempting to ignore sensitive features, this masking method encourages the model to analyze sensitive features in greater depth, which may improve performance for tasks that heavily rely on these features.

Theoretically, both methods can affect the fairness and the overall accuracy of the model. We think the actual performance of the model can highly depend on the dataset. For example, whether sensitive features and target features overlap can affect the accuracy and fairness of the model. In this situation, we can hardly remove sensitive features from the data without affecting accuracy a lot. The quantity and quality of sensitive and non-sensitive features can also affect the result. In some human face datasets, the facial detail is dominant in the image as a headshot, while others contain full-body shots, which is more informative and the facial detail is not dominant. In those cases, predicting ages based on body size and poses can be a more efficient way as younger people usually have more muscles and better weight management. In our practical case, classifying races and ages are both mainly based on face features. The selection of the masking method will affect our model, and we will analyze this more in the discussion section. In summary, we suggest to select masking methods based on the actual situation.

3.6 Applying pre-trained model

As using a pre-trained model becomes an efficient and accurate method now in deep learning, we also provide ideas about how to use pre-train weights in our method. Using pre-trained models can significantly improve the accuracy as well as fairness metrics.

3.6.1 Applying masking on input data. During the fine-tuning process, as we already know which patch is more sensitive, we can simply mask those patches, by simply changing their value to 0. This is a simple method of adjusting the fairness of the model without altering the pre-trained model's architecture or weights. Applying masking to sensitive features in input data is a quick way to block sensitive information. In ViT-based models, the hidden features correspond to the patches in the original image, which makes it possible to just the input image based on attention heatmaps.

3.6.2 Using pre-trained model when generating patch weights. Since in our method, the model generating attention heatmap is a separate model from the VAE model, we can apply a pre-trained model in the sensitive weight-generating process and train the VAE model normally based on the target dataset. The sensitive classifier does not need to have the same hyperparameters and architecture as the VAE model as long as it is a ViT-based model that can extract heatmap. Using pre-trained models can generate a sensitive classifier with a much higher accuracy especially when the training dataset is relatively small. Using a more accurate sensitivity score to train VAE can further improve the fairness metrics of the VAE model by our methods.

4 Evaluation

4.1 ResNet-50 Model&TFC Model Performance

The UTKFace dataset consists of over 20,000 face images captured in the wild. Each image contains a single face. The label of each image is embedded in the file name with format. Image filename contains [age]_[gender]_[race]_[date&time]. In here, [age] is an integer from 0 to 116. [gender] is either 0 (male) or 1 (female). [race] is an integer from 0 to 4, denoting White, Black, Asian, Indian, and Others (like Hispanic, Latino, Middle Eastern). To evaluate the model's performance, we use mean squared error (MSE) loss to measure the age prediction error. Furthermore, we choose cross-entropy loss which is applied to assess the classification accuracy for gender and race. Finally, we analyze the false positive rate (FPR) and false negative rate (FNR) across demographic groups to explore model fairness performance.

4.2 Sensitive-Awareness Masking Method

We evaluate the performance of the sensitivity-awareness masking method by training a classification model to predict ages. The sensitive feature identified for this study is race, while the classification task involves predicting age categories grouped into seven classes: 0−10, 10−20, 20−30, 30−40, 40−50, 50−60, and 70+ years. Each image is resized to 224x224 pixels and we apply basic data augmentation techniques, including random horizontal flips with a 50% probability and random rotations within a 10-degree range. In the sensitive feature extractor part, we use a ViT-Tiny [8] model to predict the race of each image. In the MAE training, we employ a 75% masking ratio during pre-training following the original MAE paper [4]. We evaluate the methods in three ways, (1) Random Masking as the benchmark, (2) Sensitive Patch Masking that covers sensitive features and encourages the model to emphasize non-sensitive regions, and (3) Non-Sensitive Patch Masking that covers non-sensitive features and encourages the model to emphasize sensitive regions. The effectiveness of the masking strategies is evaluated across five racial groups: White, Black, Asian, Indian, and Others. Four metrics are used to assess the performance of every group: Accuracy, Precision, Recall, and F1-Score. Metrics are calculated individually for each racial group and compared in a table to evaluate the performance and fairness of various masking strategies. This systematic evaluation highlights the strengths and weaknesses of each approach in promoting equitable outcomes in face recognition systems.

5 Results

5.1 ResNet-50 Model&TFC Model Performance

We use ResNet-50 as a baseline to measure model ACC performance and fairness. Table 1 shows ResNet-50 model and TFC model performance result. Table 2 and table 3 indicates FNR and FPR results by their race.

5.2 Sensitive-awareness masking method

When training the sensitive feature classifier, we obtained an accuracy of 59.76%. Table 4 shows the detailed result of MAE and our methods.

Table 1. ResNet-50 Model&TFC Model ACC Performance

Performance Result	ResNet-50	TFC
Age Mean Absolute Error (MAE)	4.84	6.2
Gender Accuracy	0.91	0.88
Race Accuracy	0.81	0.76

Table 2. ResNet-50 Model&TFC Model Gender Fairness Performance

Performance Result	ResNet-50	TFC
White Gender FPR	0.14	0.08
White Gender FNR	0.11	0.07
Black Gender FPR	0.13	0.09
Black Gender FNR	0.10	0.08
Asian Gender FPR	0.13	0.11
Asian Gender FNR	0.15	0.09
Indian Gender FPR	0.06	0.05
Indian Gender FNR	0.10	0.06
Others Gender FPR	0.12	0.07
Others Gender FNR	0.11	0.08

Table 3. ResNet-50 Model&TFC Model Age Fairness Performance

Performance Result	ResNet-50	TFC
White Age FPR	0.11	0.09
White Age FNR	0.12	0.08
Black Age FPR	0.12	0.10
Black Age FNR	0.31	0.27
Asian Age FPR	0.02	0.03
Asian Age FNR	0.32	0.28
Indian Age FPR	0.18	0.16
Indian Age FNR	0.29	0.23
Others Age FPR	0.10	0.08
Others Age FNR	0.31	0.23

6 Discussion

6.1 ResNet-50&TFC Model Result

We can find the ResNet-50 model achieves higher accuracy for gender 91% and race classification 81% compared to the TFC model (88% and 76%). However, TFC demonstrates better fairness metrics across most groups. TFC model reduces both False Positive Rates (FPR) and False Negative Rates (FNR) for gender and age across all racial groups. Furthermore, we can find some racial groups show

Original MAE				
Group	Accuracy	Precision	Recall	F1-Score
0 (White)	0.339934	0.281561	0.339934	0.279494
1 (Black)	0.381679	0.257497	0.381679	0.300772
2 (Asian)	0.601852	0.520882	0.601852	0.537369
3 (Indian)	0.473896	0.382388	0.473896	0.406210
4 (Others)	0.561404	0.489923	0.561404	0.508305
Mean	0.471753	0.386450	0.471753	0.406430
Std	0.112313	0.118770	0.112313	0.117047
Std/Mean	0.238076	0.307336	0.238076	0.287988
Over All Accuracy: 42.71%				

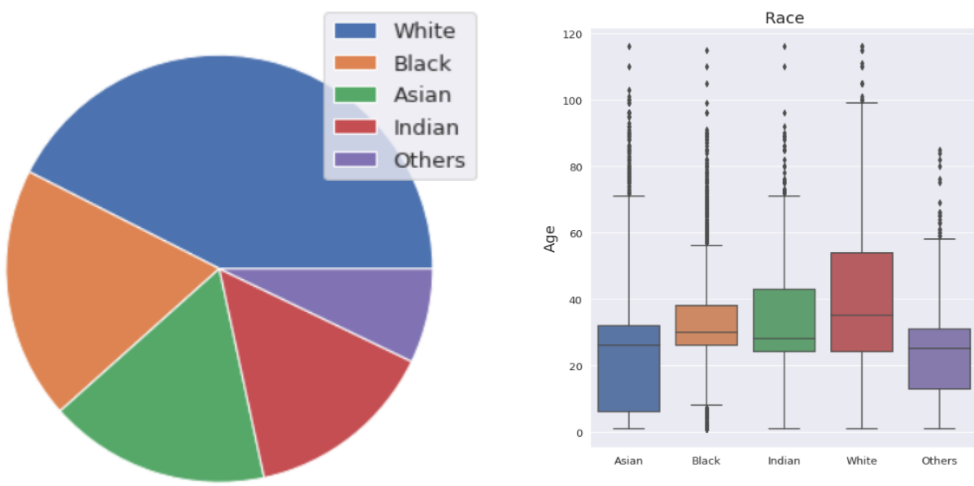
Masking Sensitive Patches				
Group	Accuracy	Precision	Recall	F1-Score
0 (White)	0.331683	0.342217	0.331683	0.280070
1 (Black)	0.416031	0.326819	0.416031	0.359027
2 (Asian)	0.560185	0.483469	0.560185	0.496557
3 (Indian)	0.457831	0.381124	0.457831	0.398946
4 (Others)	0.543860	0.488945	0.543860	0.500341
Mean	0.461918	0.404515	0.461918	0.406988
Std	0.094147	0.077180	0.094147	0.093822
Std/Mean	0.203818	0.190796	0.203818	0.230528
Over All Accuracy: 41.95%				

Masking Non-Sensitive Patches				
Group	Accuracy	Precision	Recall	F1-Score
0 (White)	0.308581	0.236189	0.308581	0.231068
1 (Black)	0.427481	0.282191	0.427481	0.315129
2 (Asian)	0.587963	0.505701	0.587963	0.517804
3 (Indian)	0.449799	0.379063	0.449799	0.366213
4 (Others)	0.500000	0.367835	0.500000	0.415316
Mean	0.454765	0.354196	0.454765	0.369106
Std	0.102380	0.103524	0.102380	0.107497
Std/Mean	0.225127	0.292279	0.225127	0.291236
Over All Accuracy: 41.12%				

Table 4. Results for Original MAE, Masking Sensitive Patches, and Masking Non-Sensitive Patches, including Mean, Std, and Std/Mean.

larger reductions in fairness metrics. For example, "Others" and "Asian" groups exhibit significant improvements in age-related fairness metrics. It indicates that the TFC model can improve face recognition fairness in the minority groups. The results demonstrate that it is possible to mitigate fairness issues without drastically sacrificing model performance.

6.2 Sensitive-awareness masking method



(a) Race distribution in UTKFace

(b) Age distribution in UTKFace

Fig. 4. UTKFace dataset exploration, adapted from [6]

In the original UTKFace dataset, the race and age distribution are not equally distributed. The distributions are shown in Fig. 4. The "white" category has a mean age of 37.98, significantly higher than the "Asian" and "Others" categories, which have mean ages of 25.87 and 23.22, respectively. This inequality can affect the accuracy of age prediction among different groups. For example, distinguishing people with ages 0–10 and 10–20 is significantly more difficult than distinguishing people with ages 10–20 and 20–30. Because of this, we are not going to analyze the difference between groups as it is highly dependent on the dataset property.

While the original Masked Autoencoder (MAE) achieves the highest overall accuracy, it displays significant fairness disparities, as indicated by larger performance gaps across demographic groups. Masking sensitive patches, though slightly reducing accuracy, provides the most balanced outcomes among groups, showing its ability in fairness-critical scenarios. Conversely, masking non-sensitive patches can focus on sensitive features, potentially benefiting minority group classifications. But its performance is below masking sensitive patches. We conclude that the reason for this performance is the risk of reinforcing biases when sensitive and target features overlap, as seen in face images from UTKFace.

However, UTKFace is a relatively small dataset for self-supervised learning especially for MAE. UTKFace contains only 20K+ images. In comparison, the size of a common dataset for self-supervised learning models pretraining is significantly larger. For example, ImageNet-1k includes 1,281,167 training images, 50,000 validation images, and 100,000 test images, while another dataset, ImageNet-21k, contains 14 million images. The smaller size of datasets like UTKFace poses challenges for applying self-supervised learning models like MAE. Limited data restricts the model's ability to develop robust, generalized representations, which can negatively affect both accuracy and fairness.

7 Conclusions

We explore and evaluate face recognition model fairness. We use ResNet-50 as baseline and use TFC model to improve its fairness. Furthermore, we choose FPR and FNR as fairness metrics to measure models performance on the UTKFace dataset. The table 1, table 2 and table 3 indicates that the ResNet-50 model achieves higher classification accuracy than the TFC model. But the TFC model demonstrates superior fairness by reducing disparities in FPR and FNR across all demographic groups. It is a trade-off between accuracy and fairness in the face recognition system. There are some limitations in these models. First, we only evaluated fairness on the UTKFace dataset which may not generalize to other datasets. Second, there still exists a trade-off between accuracy and fairness although there is not a big gap between them. We also do not know other specific application trade-off distribution. The masking methods in the MAE model are still a potential improvement. There are many potential ways to mask the image like our method and adversarial methods. Trying different masks can improve not only the accuracy of the MAE model but also other metrics like fairness. In the future, we plan to extend this model to other datasets and tasks, such as text or video classification. Moreover, we also need to solve the trade-off between accuracy and fairness.

References

[1] Haijian Chen, Wendong Zhang, Yunbo Wang, and Xiaokang Yang. 2024. Improving Masked Autoencoders by Learning Where to Mask. arXiv:2303.06583 [cs.CV] https://arxiv.org/abs/2303.06583

[2] Moreno D'Incà, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. 2023. Improving Fairness using Vision-Language Driven Image Augmentation. arXiv:2311.01573 [cs.CV] https://arxiv.org/abs/2311.01573

[3] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. 2023. Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness. arXiv:2302.10893 [cs.LG] https://arxiv.org/abs/2302.10893

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked Autoencoders Are Scalable Vision Learners. arXiv:2111.06377 [cs.CV] https://arxiv.org/abs/2111.06377

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV] https://arxiv.org/abs/1512.03385

[6] Sven Knoblauch. 2021. UTKFace Data Exploration. https://www.kaggle.com/code/svenknoblauch/utkface-data-exploration Accessed: 2024-12-12.

[7] Kimmo Kärkkäinen and Jungseock Joo. 2019. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. arXiv:1908.04913 [cs.CV] https://arxiv.org/abs/1908.04913

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

[9] Sofia Yfantidou, Dimitris Spathis, Marios Constantinides, Athena Vakali, Daniele Quercia, and Fahim Kawsar. 2024. Using Self-supervised Learning Can Improve Model Fairness. arXiv:2406.02361 [cs.LG] https://arxiv.org/abs/2406.02361

[10] Song Yang Zhang, Zhifei and Hairong Qi. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Received N/A; revised N/A; accepted N/A