Evaluating Fairness Regularization in Convolutional Neural Networks for Demographic Bias Reduction in Facial Image Classification

Kirat Kaur[1], **Marwa Mahmoud**[1]

Received January 26, 2026
Accepted May 5, 2026
Electronic access June 15, 2026

Facial image classifications have been widely deployed in security, commercial, and social applications, yet persistent demographic performance disparities raise concerns about algorithmic fairness. Prior work has shown that racial bias can remain even when models are trained on demographically balanced datasets, suggesting that dataset curation alone is insufficient to ensure equitable performance. This study evaluates whether fairness-aware regularization can reduce racial bias in convolutional neural networks under controlled, cross-dataset conditions. Three models—ResNet 50, ResNet 101, and a ResNet-based classifier incorporating Invariant Feature Regularization (INV-REG)—were trained and evaluated using the FairFace and UTKFace datasets. Model performance was assessed across racial subgroups using overall accuracy, max-min subgroup accuracy gaps, and population standard deviation to quantify demographic disparities. Cross-dataset evaluation was performed by training on FairFace and testing on UTKFace to assess robustness under distributional shift. The INV-REG model consistently reduced racial performance disparities relative to both ResNet baselines. On FairFace, INV-REG reduced the max-min accuracy gap from 6–8 percentage points to 3 points and lowered subgroup standard deviation by over 60%. Similar improvements were observed under cross-dataset evaluation on UTKFace, where INV-REG reduced the max-min gap from 15 to 6 points while improving overall accuracy. In contrast, increased model depth in ResNet 101 amplified cross-group variance despite comparable accuracy. These results support the notion that fairness-aware regularization improves consistent performance across demographic groups beyond dataset balancing alone and that architectural complexity can exacerbate bias without targeted intervention. This work provides initial evidence that both model design and training objectives jointly shape demographic bias in facial image classification systems.

Keywords: facial recognition, algorithmic fairness, demographic bias, convolutional neural networks, invariant feature regularization

Introduction

Facial image classification systems have become a central component of modern computer-vision applications, with widespread deployment in security, identity verification, and commercial platforms. Despite continued advances in convolutional neural network architectures, these systems often exhibit systematic performance disparities across demographic groups. Prior research has shown that classification accuracy, false positive rates, and error distributions can vary across demographic groups in facial analysis systems.[1,2] For example, Buolamwini and Gebru (2018) found that commercial facial image classification systems exhibited significantly higher error rates for darker-skinned individuals compared to lighter-skinned individuals, highlighting systemic bias in model performance.[3] These disparities raise concerns about fairness, reliability, and potential real-world harm, particularly for underrepresented populations.

The sources of demographic bias extend beyond dataset size and coverage. Even datasets designed to be diverse or balanced can contain hidden patterns in the data that models exploit during training. Recent studies demonstrate that demographic balancing alone does not ensure equitable performance, as convolutional neural networks continue to encode subtle appearance cues correlated with race, lighting conditions, and annotation artifacts.[2,4] Synthetic-face and feature-level bias studies further suggest that bias can emerge from both the structure of the dataset and the way models learn visual representations.[5,6] Research on broader computer-vision and face-recognition benchmarks also demonstrates that dataset design can strongly shape fairness outcomes. FACET, for example, provides a benchmark for evaluating fairness across computer-vision tasks, while studies on targeted dataset collection show that demographic representation can influence racial equity in face recognition.[7,8] Other work finds that appearance-related factors, including facial hairstyle and age-related labeling patterns, may create ad-

[1] *Cambridge Centre for International Research*

ditional sources of performance imbalance.[9,10] Bhatta et al. similarly show that deep CNN face matchers can learn unexpected visual shortcuts, reinforcing the need to evaluate model behavior beyond overall accuracy.[11] At the same time, demographic balancing alone does not guarantee fairness. Kolla and Savadamuthu show that racial distribution in training data can affect recognition bias, while Wu and Bowyer question what features should actually be balanced in a face-recognition dataset.[12,13] The FairFace Challenge and skewness-aware learning research further show that bias can persist even when researchers explicitly attempt to measure or mitigate demographic imbalance.[14,15] These findings suggest that dataset curation is important but insufficient on its own. Prior work has therefore explored several algorithmic approaches to bias reduction. Some studies focus on joint debiasing or demographic fairness transformers to reduce bias during model training.[16,17] Others investigate synthetic data as a tool for improving demographic diversity or reducing performance disparities in face-recognition systems.[18,19] Related work on fairness in facial-expression datasets similarly emphasizes that dataset bias and model bias must be evaluated together rather than treated as separate issues.[20]

In response to these limitations, algorithmic debiasing approaches have gained attention as a complement to dataset-level interventions. One such approach is invariant feature learning, which seeks to suppress demographic-specific information within learned representations while preserving task-relevant features. Related methods have attempted to reduce demographic bias by explicitly removing sensitive information from deep neural network embeddings, suppressing protected-attribute features, learning agnostic face representations, or normalizing face-recognition scores after comparison.[21–24] Invariant Feature Regularization (INV-REG) operationalizes this idea by penalizing variation in learned feature representations across learned demographic subgroups, without requiring explicit sensitive-attribute labels.[25] This framework provides a principled method for reducing demographic leakage at the representation level rather than relying solely on data rebalancing.

Despite growing interest in fairness-aware regularization, several gaps remain in the literature. Few studies directly compare standard convolutional models with fairness-regularized models under controlled, cross-dataset conditions. In addition, the relationship between model depth and demographic bias is not well understood. This is especially true when deeper models are trained on datasets that are already balanced across demographic groups.

The purpose of this study is to evaluate whether invariant feature regularization can meaningfully reduce racial performance disparities in facial recognition models and to assess how architectural depth influences demographic bias. Specifically, this work evaluates three models—ResNet 50, ResNet 101, and an INV-REG-enhanced ResNet—using the FairFace and UTKFace datasets. Model performance is analyzed using subgroup accuracy, max-min performance gaps, and population standard deviation to quantify demographic disparities and generalization behavior.

The scope of this study is limited to racial subgroup analysis within facial recognition tasks and does not examine other sensitive attributes such as gender or age. All evaluations are conducted using publicly available datasets and predefined fairness metrics, which constrain the analysis to dataset-specific demographic definitions and label quality.

The study uses a controlled experimental design in which baseline and fairness-regularized models are trained under the same conditions and evaluated both within and across datasets. This approach allows us to isolate the effects of model complexity and fairness-aware regularization on demographic bias while reducing the influence of other factors.

Methods

Research Design

This study employed a controlled experimental design to evaluate the effect of fairness-aware regularization on demographic bias in facial recognition models. Models were trained to classify demographic attributes (race) from facial images rather than perform identity recognition. A comparative analysis was conducted using three convolutional neural network configurations—ResNet 50, ResNet 101, and a fairness-regularized ResNet incorporating Invariant Feature Regularization (INV-REG). Models were trained and evaluated under identical conditions to isolate the effects of architectural depth and fairness regularization. Cross-dataset evaluation was performed to assess generalization under demographic distribution shift.

Datasets and Sample Characteristics

Two publicly available facial recognition datasets with differing demographic compositions and annotation characteristics were used.

FairFace contains approximately 108,000 images balanced across seven racial groups: White, Black, East Asian, Southeast Asian, Indian, Middle Eastern, and Latino. Due to its explicit demographic balancing, FairFace served as the primary dataset for training and in-distribution evaluation.[26]

UTKFace contains approximately 20,000 images annotated across five racial groups and exhibits known class imbalance and higher annotation noise. This annotation noise includes occasional mislabeled race categories and lower image quality (e.g., blurry or poorly lit faces), which can introduce variability in model performance and affect fairness measurements.

UTKFace was used exclusively for cross-dataset generalization evaluation.

Only images with valid race annotations were included. No human subjects were directly recruited, as all data were sourced from publicly available datasets.

Data Collection and Training

All images were obtained directly from the official dataset releases. Identical preprocessing procedures were applied across all experiments to ensure comparability between model configurations. Corrupted or unreadable images were removed, and race labels were verified and standardized across datasets. Images were resized to 224×224 pixels to match ResNet input specifications. It is important to note that FairFace and UTKFace use different race category definitions (seven and five groups, respectively). No direct mapping or merging of categories was performed. Instead, fairness metrics were computed separately within each dataset using its original labeling scheme. As a result, cross-dataset comparisons reflect overall trends in model behavior rather than one-to-one correspondence across subgroups.

Each dataset was split into 80% training and 20% testing subsets. Data augmentation techniques, including random horizontal flipping and minor spatial cropping, were applied during training only. Augmentation settings were held constant across all models to prevent training discrepancies.

Models were trained for 10 epochs using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. Weight decay was not explicitly applied. Early stopping was not used, and all models were trained for the full number of epochs. A fixed random seed was used to ensure consistent training conditions across experiments.

Model Architectures

Two standard convolutional neural network architectures were used as baselines: ResNet 50 and ResNet 101. ResNet 50 was selected due to its widespread use in fairness-related computer-vision studies, while ResNet 101 was included to evaluate the effect of increased architectural depth on demographic bias.

Both baseline models were trained using categorical cross-entropy loss and the Adam optimizer. No fairness constraints were applied during baseline training, allowing measurement of inherent demographic disparities under standard optimization objectives.

Fairness-Regularized Model: INV-REG

To evaluate algorithmic bias mitigation, an INV-REG-enhanced ResNet 50 model was implemented following Ma et al. (2023).[25] The total loss function was defined as:

$$L_{\text{total}} = L_{\text{CE}} + \lambda L_{\text{inv}} \tag{1}$$

where L_{CE} represents standard cross-entropy loss, L_{inv} represents the invariance regularization term penalizing demographic-specific variation in latent representations, and λ controls the strength of the regularization constraint. The regularization strength λ was set to 0.1 across all experiments. This value was chosen to balance model performance and fairness without extensive tuning.

In this approach, following Ma et al. (2023),[25] the standard loss is augmented with a regularization term that encourages the model to produce similar internal feature representations across groups. This reduces the model's reliance on demographic-related information while preserving task-relevant features. In this study, INV-REG was applied as a fairness-aware regularization technique within the standard training pipeline to compare model behavior against baseline models under identical conditions, rather than to introduce a new implementation of the partitioning procedure described in the original work.

INV-REG learns demographic partitions directly from data and suppresses variability across these partitions without relying on explicit sensitive-attribute labels. All training settings (e.g., learning rate and batch size) were held constant between the baseline ResNet 50 and INV-REG models to ensure a controlled comparison. Importantly, race labels were not used during training or to guide the regularization process. They were used only to evaluate fairness metrics after training. As a result, the regularization term operates without direct access to sensitive attribute labels during training.

Variables and Measurement

Model performance was evaluated using three metrics commonly employed in fairness literature:

- **Overall accuracy**, which measured standard classification performance.
- **Max-min accuracy gap**, defined as the difference between the highest- and lowest-performing racial subgroup:

$$\text{Gap} = \max(G_i) - \min(G_i) \tag{2}$$

  where G_i denotes subgroup accuracy.
- **Population standard deviation**, which measures variance in accuracy across racial subgroups and captures finer-grained demographic inequality. Population standard deviation provides a summary of how evenly performance is distributed across groups, with lower values indicating more consistent accuracy. Compared to simpler measures such as the max-min gap, it captures variation across all subgroups rather than only the extremes. However, it may be less intuitive to interpret and can be influenced by small subgroup sample sizes.

These metrics jointly quantify higher performance for certain groups and representational skew.

Data Analysis

Models were evaluated both within-dataset and across datasets. For cross-dataset analysis, all models were trained on FairFace and tested on UTKFace. This setup enabled the assessment of whether fairness-aware regularization improves robustness under demographic distribution shift.

Comparisons focused on differences in subgroup performance, disparity metrics, and overall accuracy across model configurations.

Ethical Considerations

This study exclusively used publicly available datasets that were released for research purposes. No new data were collected, and no identifiable personal information was introduced beyond existing dataset annotations. The analysis adhered to established ethical guidelines for fairness evaluation in machine learning research.

Results

Dataset Splits

Both the FairFace and UTKFace datasets were divided into 80% training and 20% testing subsets with proportional representation across racial groups. No additional sub-sampling was applied to preserve the natural distribution of each dataset. During training, data augmentation techniques, including random horizontal flips and color jitter, were applied to improve generalization. Table 1 summarizes the performance metrics for all evaluated models.

All UTKFace results reflect models trained on FairFace and evaluated on UTKFace for cross-dataset testing. As shown in Table 1, the INV-REG model consistently reduces subgroup performance gaps across both datasets while maintaining or improving overall accuracy. Detailed subgroup accuracy values for each model are provided in Table 2.

These results illustrate how performance differences across specific groups contribute to the overall fairness metrics.

Baseline Performance

ResNet 50 and ResNet 101 architectures were evaluated as baseline models on FairFace and UTKFace. Performance was measured using overall accuracy, the max-min subgroup accuracy gap, and population standard deviation.

On FairFace, ResNet 50 achieved an overall accuracy of 0.899, with a max-min gap of 0.06 and a population standard deviation of 0.0196. ResNet 101 achieved an overall accuracy of 0.895, with a larger max-min gap of 0.08 and a population standard deviation of 0.0250.

On UTKFace, ResNet 101 achieved an overall accuracy of 0.854, with a max-min gap of 0.15 and a population standard deviation of 0.0455.

Fairness-Aware INV-REG Performance

A ResNet 50 model incorporating invariance regularization (INV-REG) was evaluated on both datasets using the same metrics.

On FairFace, the INV-REG model achieved an overall accuracy of 0.906, with a max-min gap of 0.03 and a population standard deviation of 0.0075. On UTKFace, the INV-REG model achieved an overall accuracy of 0.884, with a max-min gap of 0.06 and a population standard deviation of 0.0180.

Cross-Dataset Evaluation

For cross-dataset evaluation, models trained on FairFace were evaluated on UTKFace. The ResNet 101 baseline achieved an overall accuracy of 0.854, with a max-min gap of 0.15 and a population standard deviation of 0.0455. Under the same evaluation setting, the INV-REG model achieved an overall accuracy of 0.884, with a max-min gap of 0.06 and a population standard deviation of 0.0180.

Discussion

This study evaluated the impact of fairness-aware regularization on racial performance disparities in facial recognition models across both in-distribution and cross-dataset settings. The results show that invariance regularization reduced performance gaps between demographic groups while preserving, and in some cases improving, overall classification accuracy. These findings suggest that this approach can reduce bias more effectively than dataset balancing or increasing model complexity alone.

Across both FairFace and UTKFace, baseline ResNet models showed clear differences in performance between racial groups. Deeper models increased this variation, even though overall accuracy stayed similar. This pattern matches prior work showing that increasing model complexity does not always improve fairness and can worsen performance differences for underrepresented groups. In contrast, the INV-REG model reduced max-min accuracy gaps and population standard deviation, showing more consistent performance across demographic groups.

Cross-dataset evaluation further demonstrated that fairness gains achieved through invariance regularization transferred under distributional shift. While the baseline ResNet 101 experienced a substantial increase in subgroup disparity when

Table 1 Model performance across datasets, including overall accuracy, max–min subgroup gap, and population standard deviation.

Model	Training Dataset	Testing Dataset	Overall Accuracy	Max-Min Gap	Population SD
ResNet 50	FairFace	FairFace	0.899	0.06	0.0196
ResNet 101	FairFace	FairFace	0.895	0.08	0.0250
INV-REG	FairFace	FairFace	0.906	0.03	0.0075
ResNet 101	FairFace	UTKFace	0.854	0.15	0.0455
INV-REG	FairFace	UTKFace	0.884	0.06	0.0180

Table 2 Per-subgroup accuracy across racial groups for each model and dataset.

FairFace (7 Groups)

Model	White	Black	Indian	East Asian	Southeast Asian	Middle Eastern	Latino
ResNet 50	0.93	0.89	0.90	0.92	0.90	0.88	0.87
ResNet 101	0.94	0.88	0.89	0.92	0.91	0.87	0.86
INV-REG	0.92	0.91	0.91	0.91	0.90	0.90	0.89

UTKFace (5 Groups)

Model	White	Black	Asian	Indian	Other
ResNet 101	0.93	0.85	0.87	0.84	0.78
INV-REG	0.91	0.89	0.89	0.88	0.85

evaluated on UTKFace, the INV-REG model maintained a significantly lower gap. This suggests an improved robustness to domain differences. This result supports the hypothesis that suppressing demographic-specific feature reliance encourages more generalizable representations. However, this conclusion is based on a single cross-dataset evaluation (FairFace to UTKFace). Additional experiments across a wider range of dataset shifts or qualitative analyses would be needed to more fully support claims about generalization.

Despite these improvements, residual disparities persisted, particularly for demographic groups with limited sample sizes. Small subgroup populations introduced variability in accuracy estimates and occasionally produced outliers, indicating that fairness-aware objectives alone may not fully resolve challenges associated with extreme data imbalance. This study also focused exclusively on racial attributes, and fairness behavior may differ for other protected characteristics. In addition, it does not examine intersectional effects, such as how race and gender interact, which may reveal more complex patterns of performance differences across groups. Fairness-aware regularization also introduces additional terms into the loss function, which can increase training complexity. In this study, the INV-REG model required similar training procedures to the baseline models, although tuning the regularization strength was important for achieving stable performance. While no major instability was observed, these methods may require additional tuning in larger or more complex settings.

A direct comparison between ResNet-50 and ResNet-101 under the same cross-dataset evaluation was not included. As a result, conclusions about the effect of model depth on fairness should be interpreted with caution. Including both models under identical evaluation conditions would provide a more direct assessment of the relationship between depth and demographic bias.

The choice of regularization strength may affect the trade-off between accuracy and fairness. This study did not include a detailed sensitivity analysis, and future work could explore how different values of λ influence model performance.

This study focuses on accuracy-based fairness metrics, such as subgroup accuracy gaps and standard deviation. Other measures, including false positive and false negative rates, may reveal different patterns of model behavior across groups. As a result, relying only on accuracy-based metrics may not fully capture all aspects of fairness.

The datasets were split once into training and testing sets, and all results are based on a single split. Using multiple splits or cross-validation could provide more reliable estimates of model performance and fairness. This was not explored in the current study and is an area for future work.

Subgroup sample sizes vary across datasets, particularly in UTKFace, which may introduce variability in fairness metrics. This study reports single-run results without confidence intervals or repeated trials across multiple random seeds. As a result, small differences in accuracy or fairness metrics may not reflect statistically robust improvements.

Dataset label quality may also affect the interpretation of

fairness metrics. In UTKFace, race labels may include noise or ambiguity, particularly for individuals whose identities do not fit clearly into predefined categories. Mislabeling or ambiguous cases could distort subgroup accuracy estimates and exaggerate or reduce observed gaps, especially when using max–min metrics.

A direct comparison between INV-REG and a baseline ResNet 50 model under cross-dataset evaluation was not included. Including this comparison would provide a clearer assessment of the effect of fairness regularization relative to its base architecture. This is an important direction for future work.

INV-REG represents one approach to bias mitigation, but other methods have also been proposed, including adversarial debiasing and output-based fairness constraints. Recent studies have also explored synthetic face datasets, real-versus-synthetic benchmarking, and curriculum domain adaptation as possible strategies for reducing demographic bias in face recognition systems.[27–29] These approaches aim to reduce bias either by removing demographic information from learned representations or by enforcing fairness at the prediction level. This study does not include direct comparisons with these alternative methods, and evaluating how INV-REG performs relative to other bias mitigation techniques is a valuable direction for future research.

While this study focuses on improving fairness in facial image classification systems, it is important to acknowledge broader ethical concerns surrounding their use. Prior auditing research has shown that publicly identifying biased commercial AI performance can influence accountability practices, demonstrating that fairness evaluation matters beyond technical accuracy alone.[30] Even if technical bias is reduced, such systems may still raise issues related to surveillance, privacy, and potential misuse. Therefore, improvements in fairness should be considered alongside ongoing discussions about the responsible and ethical deployment of these technologies.

Future work may explore combining invariance regularization with complementary approaches such as adversarial debiasing, reweighting strategies, or multi-attribute fairness objectives. Overall, the results demonstrate that targeted regularization techniques can meaningfully reduce demographic bias without sacrificing performance, contributing to the development of more equitable facial image classification systems.

However, it is important to note that the race categories used in FairFace and UTKFace are dataset-defined constructs and may not fully capture the complexity and fluidity of racial identity. Consequently, fairness evaluations based on these categories should be interpreted with caution, as they reflect the limitations of the underlying labeling schemes rather than definitive representations of real-world identity.

Acknowledgements

This study was conducted through the Cambridge Centre for International Research (CCIR). Gratitude is extended to Dr. Marwa Mahmoud (Department of Computer Science and Technology, University of Cambridge) for research supervision, guidance on fairness-aware machine learning methods, and feedback throughout the project. Additional thanks are extended to the Cambridge Centre for International Research for providing academic mentorship and research resources that supported this work.

References

1 A. Fan, X. Xiao and P. Washington, *Addressing Racial Bias in Facial Emotion Recognition*, 2023, arXiv:2308.04674.
2 I. Domínguez-Catena, D. Paternain and M. Galar, *Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition*, 2022, arXiv:2205.10049.
3 J. Buolamwini and T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 2018.
4 A. Pourramezan Fard, M. M. Hosseini, T. D. Sweeny and M. H. Mahoor, *AffectNet+: A Database for Enhancing Facial Expression Recognition with Soft-Labels*, 2024, arXiv:2410.22506.
5 R. Raina, M. Monares, M. Xu, S. Fabi, X. Xu, L. Li, W. Sumerfield, J. Gan and V. R. de Sa, *Exploring Biases in Facial Expression Analysis Using Synthetic Faces*, 2022.
6 T. Lian and O. Celiktutan, *A Feature-Level Bias Evaluation Framework for Facial Expression Recognition Models*, 2025, arXiv:2505.20512.
7 L. Gustafson *et al.*, *FACET: Fairness in Computer Vision Evaluation Benchmark*, 2023.
8 R. Hong *et al.*, *Evaluation of Targeted Dataset Collection on Racial Equity in Face Recognition*, 2023.
9 K. Ozturk *et al.*, *Can the Accuracy Bias by Facial Hairstyle Be Reduced Through Larger Training Data?*, 2024.
10 N. Panić *et al.*, *Addressing Demographic Bias in Age Estimation Models Using UTKFace and APPA-REAL*, 2024.
11 A. Bhatta *et al.*, *Our Deep CNN Face Matchers Have Developed Achromatopsia*, 2024.
12 M. Kolla and A. Savadamuthu, *The Impact of Racial Distribution in Training Data on Face Recognition Bias: A Closer Look*, 2023.
13 H. Wu and K. W. Bowyer, *What Should Be Balanced in a 'Balanced' Face Recognition Dataset?*, 2023, arXiv:2304.09818.
14 T. Sixta, J. C. S. Jacques Júnior, P. Buch-Cardona, N. M. Robertson, E. Vazquez and S. Escalera, *FairFace Challenge at ECCV 2020: Analyzing Bias in Face Recognition*, 2020.
15 M. Wang and W. Deng, *Mitigating Bias in Face Recognition Using Skewness-Aware Reinforcement Learning*, 2020.
16 S. Gong, X. Liu and A. K. Jain, *Jointly De-Biasing Face Recognition and Demographic Attribute Estimation*, 2019, arXiv:1911.08080.
17 K. Kotwal and S. Marcel, *Demographic Fairness Transformer for Bias Mitigation in Face Recognition*, 2024.
18 M. Huber, A. T. Luu, F. Boutros, A. Kuijper and N. Damer, *Bias and Diversity in Synthetic-Based Face Recognition*, 2024.
19 M. Yeung, T. Teramoto, S. Wu, T. Fujiwara, K. Suzuki and T. Kojima, *VariFace: Fair and Diverse Synthetic Dataset Generation for Face Recognition*, 2024, arXiv:2412.06235.
20 M. M. Hosseini, A. Pourramezan Fard and M. H. Mahoor, *Faces of Fairness: Examining Bias in Facial Expression Recognition Datasets and Models*, 2025, arXiv:2502.11049.

21 M. Alvi, A. Zisserman and C. Nellå ker, *Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings*, 2018.

22 P. Dhar, J. Gleason, A. Roy, C. D. Castillo and R. Chellappa, *PASS: Protected Attribute Suppression System for Mitigating Bias in Face Recognition*, 2021.

23 A. Morales, J. Fierrez, R. Vera-Rodriguez and R. Tolosana, *SensitiveNets: Learning Agnostic Representations with Application to Face Images*, 2021.

24 P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner and A. Kuijper, *Post-Comparison Mitigation of Demographic Bias in Face Recognition Using Fair Score Normalization*, 2020.

25 J. Ma, Z. Yue, T. Kagaya, T. Suzuki, J. Karlekar, S. Pranata and H. Zhang, *Invariant Feature Regularization for Fair Face Recognition*, 2023.

26 K. Kärkkäinen and J. Joo, *FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation*, 2021.

27 P. Melzi, R. Tolosana, R. Vera-Rodriguez and J. Fierrez, *Synthetic Data for the Mitigation of Demographic Biases in Face Recognition*, 2023.

28 P. Melzi, R. Tolosana, R. Vera-Rodriguez and J. Fierrez, *FRCSynonGoing: Benchmarking and Comprehensive Evaluation of Real and Synthetic Data to Improve Face Recognition Systems*, 2024.

29 F. Z. Ou *et al.*, *Troubleshooting Ethnic Quality Bias with Curriculum Domain Adaptation for Face Recognition*, 2023.

30 I. D. Raji and J. Buolamwini, *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products*, 2019.