Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/145/2025.21897

Fairness and Privacy Challenges in Face Recognition: A Deep Learning Perspective

Jiahui Lin[1,a,*]

[1]*School of Electrical Engineering and Artificial Intelligence Xiamen University Malaysia, Sepang, Malaysia*
a. inky1468@gmail.com
**corresponding author*

Abstract: Driven by deep learning, face recognition technology has become increasingly advanced and is now applied in various fields. However, this technology still faces challenges regarding fairness and privacy protection. Due to multiple factors, such as the imbalance of training data, there are significant discrepancies in the recognition accuracy of facial recognition systems across different populations, leading to legal issues. Privacy risks such as unauthorized data collection and identity theft have garnered widespread attention. This paper systematically analyzes the bias present in current mainstream face recognition models and discusses fairness optimization methods, such as data balancing, feature learning improvements, and multi-task learning. Simultaneously, the paper reviews the latest privacy protection technologies to mitigate data security risks, including federated learning, differential privacy, and homomorphic encryption. By exploring fairness and privacy-preserving technologies, this study aims to promote the development of face recognition technology toward a more equitable and secure direction.

Keywords: Facial recognition, deep learning, fairness, privacy protection, bias mitigation

1. Introduction

Since Woody Bledsoe, Helen Chan Wolf, and Charles first proposed face recognition in the 1960s, the technology has continuously advanced over the decades and is now widely used in social networks, healthcare, and security services [1]. From basic architectures such as CNNs and RNNs to complex architectures like AlexNet, GoogleNet, VGGNet, and ResNet, deep learning has significantly enhanced the accuracy of facial recognition [2]. However, despite these advancements, the technology still faces numerous challenges. The privacy issues associated with facial recognition technology remain unresolved. Individuals' facial data may be collected without their knowledge, infringing on personal privacy rights and being used in criminal activities such as fraud. Secondly, facial recognition suffers from dataset bias, performing well only in specific groups while lacking high accuracy in recognizing women and people of color. Therefore, this study aims to explore the issues present in the following facial recognition technologies, conducting a comparative analysis of the biases of different facial recognition models across various datasets, to understand how technological advancements improve fairness issues and discussing the application of the latest privacy protection technologies in facial recognition, providing methods to reduce privacy risks.

DOI: 10.54254/2755-2721/145/2025.21897

2. Fairness Issues in Facial Recognition

The modern face recognition system primarily relies on convolutional neural networks (CNN) and is trained using large-scale datasets such as LFW and MS-Celeb-1M. However, research has found significant performance disparities in these systems based on factors such as race, gender, and age. For instance, LFW is a well-known dataset composed of celebrity faces, with males accounting for as much as 77.5% and whites for 83.5% [3]. This bias not only demonstrates the unfairness of the model but may also pose ethical and legal risks. In 2018, the American Civil Liberties Union discovered that the software incorrectly identified 28 members of Congress as criminals. When photos of several prominent Black women, including Oprah and Michelle Obama, were scanned by Amazon's technology, the system mistakenly identified them as male. Such incidents have occurred frequently, leading Amazon, IBM, and Microsoft to suspend the provision of facial recognition technology to law enforcement agencies in 2020 [4].

2.1. Source of Bias

The recognition bias may be due to the following reasons. Firstly, the dataset used by the current model is unbalanced. Many mainstream datasets predominantly feature samples of white males, while the data for other groups, such as individuals with darker skin tones and females, is relatively scarce.

For instance, some popular datasets include LFW, VGGFace, and CASIA, and we define a racially unbalanced dataset as one that contains a difference of more than 5 percentage points between the most represented and the least represented races. Notably, even the FairFace dataset, which emphasizes racial equality, has a maximum racial difference of 8.3 percentage points, categorizing it as unbalanced. We find that each of these datasets is defined as racially unbalanced [5].

Secondly, it may be because the currently used loss function has a singular optimization objective. Existing face recognition systems focus on optimizing classification accuracy without considering the balance among different groups, resulting in poor generalization ability for minority groups. There may also be feature learning bias, where the model prioritizes learning features of certain races or genders, leading to decreased recognition ability for other groups.

For instance, according to a NIST survey, the team found that the recognition error rates for Asian and African American faces were higher than those for Caucasian images, with differences reaching 10 to 100 times [6].

2.2. Comparison of Mainstream Models

It can be observed from the above table that most models exhibit unfairness, with the recognition accuracy for white individuals being higher than that for people of color and the recognition accuracy for men being higher than that for women. However, over time, with algorithm optimization, this unfair disparity is gradually narrowing. SphereFace optimizes feature distribution by improving Softmax Loss, enhancing the feature distinguishability among different groups, while FairFace combines a fairness loss function (Fair Loss) to achieve a more balanced feature learning across different racial and gender groups.

DOI: 10.54254/2755-2721/145/2025.21897

Table 1: Comparison of Fairness in Mainstream Face Recognition Models

Model	Main architecture	Loss function	Racial fairness	Gender fairness	Major fairness issues
DeepFace (2014) [7]	CNN	Softmax	The main training data is white, and the racial bias is obvious	It performs better in males than females	The misidentification rate of dark-skinned groups was high
FaceNet (2015) [8]	CNN+Triplet Loss	Triplet Loss	The training data is more diverse, but still predominantly white	It performs better in males than females	The false rejection rate (FNR) is higher in the non-white group
VGGFace (2015) [9]	VGG-16	Softmax	Lack of ethnic diversity in training data	It performs better in males than females	Unstable performance towards minorities in the case of occlusion
SphereFace (2017) [10]	CNN+A-Softmax	A-Softmax	It is more balanced across multiple ethnic groups, but still has a greater bias towards darker skin groups	It performs better in males than females	The angular margin needs to be adjusted to optimize fairness
ArcFace (2018) [11[	ResNet+ArcFace Loss	ArcFace Loss	The generalization ability is strong, but the performance is still lower than that of whites in the dark-skinned group	The recognition rate for females is slightly lower than that of males	The imbalance in data leads to limited learning capacity for different populations
FairFace (2020) [12]	CNN	Softmax+Fair Loss	Reduce racial bias with balanced data sets	The gender ratio in the training data is more balanced	The overall accuracy is slightly lower than ArcFace, but the fairness optimization is better.

DOI: 10.54254/2755-2721/145/2025.21897

3. Future Optimization Directions for Fairness

Methods for optimizing fairness can be divided into data level, model level, and assessment level.

3.1. Data level optimization

Firstly, we can adopt a more balanced dataset. By utilizing Generative Adversarial Networks (GAN) and latent diffusion models to balance the data, we seek to find solutions to the issue of racial bias caused by uneven datasets [3].

Secondly, we can attempt adaptive resampling through the Learning Optimal Sample Weight (LOW) method, which dynamically adjusts the sampling weights of samples during the training process to ensure that the recognition accuracy of all categories tends to be balanced. LOW optimizes batch training by automatically estimating the weight of each sample in the loss function, allowing the model to focus more on relevant samples, effectively enhancing the model's learning and recognition capabilities for minority groups. It can effectively reduce the model's bias on imbalanced data, making its performance fairer across all groups [13].

3.2. Model-level optimization

Since Multi-Task Learning (MTL) can jointly train multiple tasks, it has shown superior performance compared to Single-Task Learning (STL) in many applications. However, the effectiveness of MTL largely depends on the reasonable allocation of task weights; imbalanced weights may lead to certain tasks dominating the training, resulting in uneven learning outcomes for different groups and subsequently affecting fairness.

Research indicates that weights can be dynamically adjusted based on the difficulty of the tasks to reduce performance disparities among different groups. Compared to manually setting fixed weights, this study proposes a dynamic weight multi-task learning framework based on deep CNN, employing a hard parameter sharing structure to share hidden layers among different tasks. The framework consists of two task branches, with Branch 1 responsible for face verification, extracting embedding features for matching, while Branch 2 calculates the probabilities of facial expressions through a softmax layer, achieving facial emotion recognition.

This model is based on the Inception-ResNet structure, containing approximately 20 hidden layers and 13 million parameters, which is significantly smaller than VGGFace (138 million parameters), thus reducing computational costs while ensuring high accuracy. The multi-task learning structure helps leverage pre-trained features from face recognition to enhance the effectiveness of facial expression recognition. The adaptive weight allocation method based on task difficulty can automatically optimize the learning proportions of different tasks during the training process, ensuring that the model does not solely focus on groups with larger data volumes but rather learns more equitably from features across different races, genders, and age groups [14].

Additionally, Sensitive Loss can also enhance the fairness and accuracy of face recognition. The role of Sensitive Loss in addressing fairness issues. Research has shown that there are significant biases between racial and gender groups in deep learning face recognition models, and the imbalance of training data can lead to much lower recognition accuracy for certain groups compared to others. For instance, popular pretrained models (VGG-Face, ResNet-50, and ArcFace) exhibit error discrepancies of up to 200% in recognition performance across different groups, with certain groups, such as Black women, being nearly twice as likely to be misidentified (false positives) compared to others [4].

Sensitive Loss aims to mitigate these biases and improve model fairness. It is based on Triplet Loss and optimizes the model's learning capability for minority groups by selecting sensitive triplets online. Unlike traditional debiasing methods, Sensitive Loss can be integrated as an additional

DOI: 10.54254/2755-2721/145/2025.21897

component to pre-trained networks. This allows models that perform well but lack fairness considerations, such as face recognition models, to achieve fairer outcomes without altering the backbone network structure.

Experimental results indicate that Sensitive Loss can effectively reduce identification errors between different groups and enhance overall fairness, with average accuracy and fairness metrics across multiple test databases (DiveFace, RFW, and BUPT-B) surpassing those of the baseline model [15].

3.3. Assessment Level Optimization

In addition to traditional accuracy and mean absolute error (MAE), Instance-Level Fairness can also be used to measure the performance of face recognition models across different groups. Research indicates that current face recognition systems exhibit significant performance disparities among different racial and gender groups, and this bias not only affects the overall recognition rate of the group but also leads to an imbalance in the false recognition rates among individuals. To address this issue, the Instance-Consistent Fair Face Recognition (IC-FFR) method proposes a fairness evaluation approach based on instance margin, ensuring that the false positive rate (FPR) and true positive rate (TPR) for all individuals remain consistent, thereby reducing individual-level unfairness.

IC-FFR has theoretically demonstrated that the imbalance between FPR and TPR can result in a higher recognition error rate for specific populations, and experiments conducted on multiple datasets (NFW, RFW, and BFW) show that this method can effectively reduce recognition errors between different racial and gender groups, enhancing the fairness of the face recognition system at the individual level [16]. Test across datasets is also required to ensure that fairness improvements are not specific to a single dataset.

4. Latest Privacy-preserving Technologies

With the widespread application of face recognition technology, ensuring privacy has become an important issue. Unauthorized data collection, misuse, and data breaches can pose privacy risks such as identity theft, surveillance abuse, and more. Therefore, the researchers proposed a series of privacy-preserving techniques to reduce the risk of privacy breaches.

4.1. Federated Learning

The concept of federated learning was proposed by Google and is a distributed machine learning technology that enables multiple participants to collaboratively train models without sharing raw data, preventing data leakage and effectively addressing the problem of data silos [17]. Google employs federated learning in Gboard (Google Input Method) and Android Messages to enhance the user input experience while ensuring data localization and minimizing the risk of user privacy breaches. Apple has also introduced federated learning in iOS 13 [18].

After mentioning federated learning, we will discuss three technologies that are closely related to federated learning and are also used for privacy protection.

4.2. Differential Privacy (DP)

It protects privacy by adding noise to the data or generalizing certain sensitive attributes so that third parties cannot distinguish between specific individuals. The relevant technologies we commonly use include k-anonymity, which ensures that each data point is indistinguishable from at least k other data points to prevent individual identification, and data diversification, which enhances privacy protection through generalization or perturbation of data. For example, Apple employs differential

DOI: 10.54254/2755-2721/145/2025.21897

privacy techniques in its Face ID training data to prevent user information leakage [19]. These methods often require data transmission to a third-party server, which still presents certain privacy risks and can affect accuracy to a certain extent.

4.3. Homomorphic Encryption (HE)

It allows for direct computation on encrypted data, enabling accurate results to be obtained without decrypting the original data, thereby ensuring data privacy [20].

Unlike differential privacy, the data and the model are not transmitted and cannot be inferred from each other's data. Therefore, the likelihood of a leak at the raw data level is very low, and even if a hacker steals the data, they will not be able to directly access the user's facial information. It is used to protect user data privacy, especially in machine learning and cloud computing scenarios. However, the computational overhead of HE is significant, particularly with fully homomorphic encryption, making it difficult to apply in real-time face recognition.

4.4. Secure Multi-Party Computation (SMC)

SMC allows multiple parties to work together to compute the result of a function without revealing their respective input data and ensures zero knowledge, where each party knows nothing but inputs and outputs. For example, the study used the SMC framework to train machine learning on two servers and under the semi-honest assumption to prevent privacy breaches [21]. However, due to the complexity of the calculation protocol, it may affect efficiency.

4.5. Generative Adversarial Networks (GANs)

By capturing the distribution of training data, GANs and their variants can produce more similar samples that are difficult to distinguish from the real ones without touching the original training data, relying only on the information retained by the discriminant model, confusing the line of sight, and effectively preventing the exposure of personal privacy [22].

However, there is still a risk of privacy compromise with this technology. If the distribution of training data is too centralized, an attacker can reconstruct the original training data by continuously sampling the images generated by the GAN. In addition, GANs are susceptible to some powerful attacks; for example, in the study of Hayes et al., it is possible to launch member inference attacks on GANs even without training additional models, thus speculating whether the GAN has seen a certain face [23].

4.6. Cancelable Biometrics

Cancelable Biometrics is a technology used to protect biometric data such as fingerprints, iris, faces, etc. The main idea is to transform (convert into a different format) the original biometric so that it cannot be directly restored, thus protecting privacy. For example, if a biometric template is leaked, it generates a new template but does not change the biometric itself [24].

There are also shortcomings to this technology. Conversion parameters need to be stored securely. Otherwise, an attacker may find a way to restore the original biometrics by analyzing the transformed data.

In the future, to solve these shortcomings, in addition to optimizing algorithms, improving computing efficiency, and hardware acceleration, an important research direction is to combine different privacy protection technologies.

For example, federated learning can be combined with SMC/HE, which can improve the efficiency of distributed computing. DP is combined with the GAN, and the gradient is DP-processed during

DOI: 10.54254/2755-2721/145/2025.21897

training, preventing the GAN from remembering the specific details of the training data. In the application of revocable biometrics, use HE to protect transform parameters from being guessed by attackers, or add DP to the template matching process to reduce the risk of biometric recovery to provide a more secure biometric system

5. Conclusion

With the advancement of deep learning, face recognition technology has made significant progress. However, issues of fairness and privacy continue to trouble us. The bias in datasets results in disparities in the model's recognition capabilities across different populations, leading to fairness concerns. To address this issue, researchers have proposed various optimization methods, including data balancing, fairness-based loss functions, and multi-task learning. Furthermore, facial recognition technology still faces challenges regarding privacy protection, and the risk of data leakage can be mitigated through technologies such as federated learning, homomorphic encryption, and differential privacy. Future research should focus on how to effectively integrate multiple fairness and privacy protection methods to create a more just and secure face recognition system.

References

[1] *Bledsoe, W. W. (1966). The model method in facial recognition. Panoramic Research Inc.*

[2] *S. Zhu, "Enhancing Facial Recognition: A Comprehensive Review of Deep Learning Approaches and Future Perspectives, " Appl. Comput. Eng., vol. 110, no. 1, pp. 137–145, Nov. 2024, doi: 10.54254/2755-2721/110/2024MELB0107.*

[3] *H. Han and A. K. Jain, "Age, gender and race estimation from unconstrained face images, " Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep. (MSU-CSE-14-5), 2014.*

[4] *Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*), 77–91.*

[5] *A. Sumsion, S. Torrie, D.-J. Lee, and Z. Sun, "Surveying Racial Bias in Facial Recognition: Balancing Datasets and Algorithmic Enhancements, " Electronics, vol. 13, no. 12, p. 2317, 2024, doi: 10.3390/electronics13122317.*

[6] *Grother, P., Ngan, M., & Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. National Institute of Standards and Technology (NIST). https://doi.org/10.6028/NIST.IR.8280*

[7] *Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification, " IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1701–1708.*

[8] *F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering, " IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.*

[9] *O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition, " British Machine Vision Conference (BMVC), 2015.*

[10] *W. Liu, Y. Wen, Z. Yu, and M. Li, "SphereFace: Deep hypersphere embedding for face recognition, " IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 212–220.*

[11] *J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition, " IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4690–4699.*

[12] *K. Krishnapriya, K. Albiero, M. C. King, and K. W. Bowyer, "Analysis of race and gender bias in deep face recognition, " IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 3, no. 1, pp. 42–52, 2021.*

[13] *C. Santiago, C. Barata, M. Sasdelli, G. Carneiro, and J. C. Nascimento, "LOW: Training deep neural networks by learning optimal sample weights, " Pattern Recognition, vol. 110, p. 107585, 2021. doi: 10.1016/j.patcog.2020.107585.*

[14] *Z. Ming, J. Xia, M. M. Luqman, J.-C. Burie, and K. Zhao, "Dynamic multi-task learning for face recognition with facial expression, " arXiv preprint arXiv:1911.03281, Nov. 2019. doi: 10.48550/arXiv.1911.03281.*

[15] *I. Serna, A. Morales, J. Fierrez, and N. Obradovich, "Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning, " Artificial Intelligence, vol. 305, p. 103682, 2022. doi: 10.1016/j.artint.2022.103682.*

[16] *Y. Li, Y. Sun, Z. Cui, P. Shen, and S. Shan, "Instance-Consistent Fair Face Recognition, " IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2025.3545781.*

[17] *Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augen stein, S., Eichner, H., Kiddon, C. and Ramage, D. Federated learning for mobile keyboard prediction. ArXiv Preprint arXiv:1811.03604. (2018)*

DOI: 10.54254/2755-2721/145/2025.21897

[18] *Apple Private Federated Learning (NeurIPS 2019 Expo Talk Abstract). (https://nips.cc/ExpoConferences/2019/schedule?talkid=40, 2019)*

[19] *Apple Inc., "Differential Privacy Overview, " Apple Inc., Cupertino, CA, USA, White Paper, 2017. [Online]. Available: https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf*

[20] *Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Giorgio Patrini, Guillaume Smith, and Brian Thorne. 2017. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. CoRR abs/1711.10677 (2017).*

[21] *Payman Mohassel and Yupeng Zhang. 2017. SecureML: A System for Scalable Privacy-Preserving Machine Learning. In IEEE Symposium on Security and Privacy. IEEE Computer Society, 19–38.*

[22] *I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks, " arXiv preprint arXiv:1406.2661, 2014. [Online]. Available: https://doi.org/10.48550/arXiv.1406.2661*

[23] *J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "LOGAN: Membership Inference Attacks Against Generative Models, " arXiv preprint arXiv:1705.07663, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1705.07663*

[24] *M. Rawat and N. Kumar, "Cancelable biometrics: A comprehensive survey, " Artif. Intell. Rev., vol. 53, pp. 3403–3446, Jun. 2020.*