Can Pretrained Face Verification Models Distinguish True Identity from Deepfakes?

Pai Chet Ng[1]
paichet.ng@singaporetech.edu.sg
Kostas N. Plataniotis[1]
kostas@ece.utoronto.ca

[1] Infocomm Technology Cluster,
Singapore Institute of Technology
Singapore

[2] The Edward S. Rogers Sr. Department
of Electrical & Computer Engineering
University of Toronto,
Canada

Abstract

This work aims to assess the reliability of pretrained face verification models in differentiating between authentic identities and their deepfake equivalents. With deepfake technology increasingly able to produce highly realistic and deceptive visual content, this work focuses on evaluating the effectiveness of widely used pretrained models—VGG-Face, DeepFace, and Facenet—in recognizing such falsified identities. We investigate three critical research questions: (1) How proficient are these pretrained models at distinguishing between real and deepfake images? (2) What enhancements can be implemented to improve their detection capabilities? (3) What are the implications of these advancements for ensuring user security on digital platforms? By conducting comprehensive evaluations, this research aims to uncover the current limitations of these models and propose potential improvements, contributing to more robust face verification systems that can effectively counteract advanced identity spoofing technologies.

1 Introduction

Imagine a scenario where a digitally synthesized replica of your face is used in an attempt to impersonate you on social media. This emerging threat challenges the reliability of pretrained face verification models, which are developed to authenticate identities through advanced facial recognition techniques [5, 14, 16]. Our work addresses this critical issue by examining the ability of these models to differentiate authentic users from their artificially generated counterparts. We focus on scenarios where deepfakes closely mimic the visual characteristics of legitimate users, a more subtle and sophisticated threat than general deepfake anomalies [1, 8, 15]. The significance of addressing this challenge is crucial—it underpins the safeguarding of individual identities and the protection of digital ecosystems against advanced frauds.

As the capabilities of deep learning have significantly advanced, face verification models such as VGG-Face [9], DeepFace [13], and FaceNet [12], have become fundamental to security frameworks across various platforms. These models are integral for tasks ranging from

© 2024. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

unlocking smartphones to authenticating identities in financial services and securing social media interactions. Despite their proven effectiveness in conventional scenarios, the emergence of deepfakes presents a unique challenge. Deepfakes can skillfully mimic genuine users, raising critical questions about the reliability of these pretrained models: Can they effectively differentiate between a real user and a deepfake replica of the same individual? This dilemma indicates a crucial need to refine security measures in our increasingly digital world, where synthetic media is growing more prevalent.

Recent research has begun to address the threats that deepfakes pose to face verification systems. Works like those by Rusia and Singh [10] and Firc et al. [3] explore the landscape of identity threats and the specific challenges deepfakes present to biometric systems. However, there is a noticeable gap in the literature: no work has yet compared different pretrained face verification models to evaluate their performance against deepfakes that closely resemble the genuine user. Our work aims to fill this gap by methodically assessing the robustness of various state-of-the-art face verification models against deepfakes designed to replicate the facial features of registered users. In addressing this, our research is structured around three pivotal questions:

1. How do pretrained face verification models perform when tasked with distinguishing between genuine users and their deepfake counterparts?

2. What enhancements can be applied to these models to reliably reject deepfakes while affirming true user identities?

3. What are the security implications of these findings for digital identity verification systems, and how do they affect user privacy?

The contributions of our research are significant in advancing the field of digital identity security, linking directly to the posed questions. We provide a comprehensive evaluation of leading pretrained models, discussing potential enhancements and adaptations in model architecture that could bolster their resilience against deepfakes. Additionally, by exploring the broader security implications, our findings contribute to an enhanced understanding of how improved deepfake detection can not only secure digital identities but also protect user privacy in an era dominated by advanced digital impersonation technologies.

2 Overview of the Face Verification Problem

Face verification is a critical process in identity authentication [11], where a query face image, denoted as x_q, is compared against a template image, x_t, that has been previously registered and stored in a database. This template is authenticated during a controlled enrollment phase, where the user's identity is verified through reliable means such as physical identification or biometric data. The verification process begins by detecting and normalizing the face from the input image to ensure uniformity regardless of the original image's conditions. The normalized face is then transformed into a high-dimensional embedding by a deep neural network $f(x) : R^{H\times W\times C} \rightarrow R^d$, where H, W, and C represent the height, width, and number of channels of the image, respectively, and d represents the dimensionality of the embedding.

The core of the verification process involves calculating the distance between two embeddings. Let $D(\cdot)$ be the distance function, if Euclidean distance is chosen, then the distance

between the embeddings of the query and the template images can be described as follows:

$$D(x_q, x_t) = \|f(x_q) - f(x_t)\|_2. \tag{1}$$

Let τ be the decision threshold, then the system verifies the identity if $D(x_q, x_t) \leq \tau$, affirming that the query image matches the authentic template, and rejects it otherwise.

State-of-the-Art in Face Verification

The field of face verification has seen substantial advancements with the introduction of deep learning. State-of-the-art models like VGG-Face [9], DeepFace [13], and FaceNet [12] have transformed the landscape by utilizing deep neural networks to extract discriminative and robust features from face images. These models have demonstrated high accuracy across diverse scenarios, including variations in lighting, pose, and facial expressions.

- VGG-Face [9]: Developed by the Visual Geometry Group at the University of Oxford, VGG-Face is an adaptation of the VGG-16 model that was initially designed for image classification tasks. By training on a large dataset of facial images, VGG-Face achieves remarkable accuracy in face recognition tasks, including a 97.78% accuracy rate on the Labeled Faces in the Wild (LFW) dataset [6]. This model leverages convolutional neural networks (CNNs) with deep layers, enabling it to capture complex facial features at various levels of abstraction.

- Facebook DeepFace [13]: DeepFace, developed by Facebook, utilizes a nine-layer neural network that includes over 120 million connection weights. It was trained on a massive dataset of four million images uploaded by Facebook users, enabling the model to achieve an accuracy of 97.35% on the LFW dataset. DeepFace is particularly noted for its robust performance across diverse demographic groups and complex real-world scenarios.

- Google FaceNet [12]: Google's FaceNet represents a significant advancement in the field by introducing a novel training approach using the triplet loss function. This method involves learning a high-quality embedding of each face into a 128-dimensional Euclidean space, where distances directly correspond to a measure of face similarity. Achieving an unprecedented 99.63% accuracy on the LFW dataset, FaceNet's architecture is designed to optimize performance by closely aligning facial features in a normalized embedding space.

The state-of-the-art performance of these pretrained models is attributed to their ability to learn rich, discriminative representations of facial features. While these models excel in traditional verification scenarios, their performance against deepfakes that are perfect replicas of genuine users has not been thoroughly verified. This work aims to explore whether these pretrained models can detect subtle anomalies indicative of deepfakes, thereby enhancing security measures in digital identity verification systems and ensuring robust protection against emerging threats. In doing so, we seek to answer a pivotal question: Can these state-of-the-art face verification models reliably distinguish between real users and their deepfake counterparts even when the artificial images bear an uncanny resemblance to the genuine identities?

Figure 1: Samples of face images generated using StarGAN [2]: the first row displays the original images, the second row shows images transformed by gender alteration, and the third row presents images with age alteration.

3 Experiments Evaluation

3.1 Deepfake Generation

In this work, we utilize StarGAN [2] to generate deepfake images, specifically focusing on transforming the gender and the age of individuals in images. StarGAN is a versatile generative adversarial network (GAN) [4] known for its ability to perform image-to-image translations across multiple domains within a single model. Unlike traditional GANs that require separate models for each transformation, StarGAN efficiently handles multiple transformations by using a single unified architecture. This is achieved by inputting both the image and the target domain label, which instructs the model on the desired transformation. Our goal is to generate male faces expressing sadness from original faces, which serves as a test case for exploring the potential misuse of deepfake technology in face verification systems.

Using the publicly available CelebA dataset [7], we apply StarGAN only to the testing set as defined in the original StarGAN paper. This ensures that our transformations are based on a consistent and standardized subset of data, providing clear benchmarks for performance and evaluation. The deepfake images generated by StarGAN, although visually convincing, should not be used for authenticating or verifying user identity. In our experiments, we would like to examine if these generated images would result in a significant distance metric when compared to the original images in a face verification system, despite having originated from the same individual. This large distance metric is crucial for ensuring the security of face verification systems against potential misuse. By intentionally generating deepfake images that could be mistaken for genuine users, we can test and enhance the robustness of these systems.

3.2 Image Pair Sampling Strategy

We generate the image pair with two sampling strategies to simulate realistic scenarios that might occur in a face verification system. The template image is sampled from the training set of the CelebA dataset. This image simulates the genuine image registered in the user database during initial system setup. These strategies are designed to test the system's ability to differentiate between genuine and deepfake images from the same and different users.

- **Same faces verification**: The first data sampling approach focuses on verifying the identity of a user against deepfake images generated from the same user. The query

image is sampled from the test set with a random choice mechanism: With a 50% probability, the query image is a genuine image of the same user, sourced from the real images in the test set. With another 50% probability, the query image is a deepfake image generated from the same user's real image using StarGAN, simulating an attempt to use a deepfake image for verification.

- **Different faces verification**: The second type of data sampling strategy extends the complexity of the verification scenario by including images from different users. The query image is sampled from the test set with a randomized mechanism: With a 50% probability, the query image is a deepfake generated from the same user, With another 50% probability, the query image is a deepfake image generated from a different user's real image. This tests the system's ability to reject verification attempts using deepfake images from the same user as well as users other than the registered individual.

Model	Accuracy of Positive Pair		Accuracy of Negative Pair		EER (%)
VGG-Face [9]	0.707 0.717	0.712	0.599 0.581	0.590	66.49 65.70
FaceNet [12]	0.526 0.548	0.537	0.933 0.821	0.877	74.30 68.40
DeepFace [13]	0.550 0.543	0.547	0.940 0.848	0.894	75.45 69.65

Table 1: Accuracy scores for positive and negative pairs, where negative pairs are sampled from the same user but are deepfake-generated. Each model undergoes testing against deepfakes created through gender and age alterations. For each model, the top row displays the results against gender alteration deepfakes, and the bottom row for age alteration deepfakes. The number following these two results represents the average accuracy for both positive and negative pairs.

3.3 Evaluation Results

Same faces verification The results summarized in Table 1 provide a detailed comparison of the performance of three pretrained face verification models: VGG-Face, FaceNet, and DeepFace. A notable observation is the relatively higher accuracy of positive pairs observed with VGG-Face compared to FaceNet and DeepFace. For VGG-Face, the average accuracy across both gender and age alterations for positive pairs is approximately 0.712, indicating a strong ability to correctly verify the true identities of users against their genuine images. In contrast, FaceNet and DeepFace show lower average accuracies for positive pairs, at 0.537 and 0.547, respectively. FaceNet and DeepFace exhibit significantly higher accuracies for negative pairs, with averages of 0.877 and 0.894, respectively, compared to 0.590 for VGG-Face. This suggests that while FaceNet and DeepFace may struggle more with positive pair verification, they excel in identifying and rejecting deepfakes. This could be due to their training processes and optimization strategies, which might be more aligned towards distinguishing subtle discrepancies introduced by deepfake technologies than VGG-Face. The discrepancy in performance between positive and negative pair verifications highlights

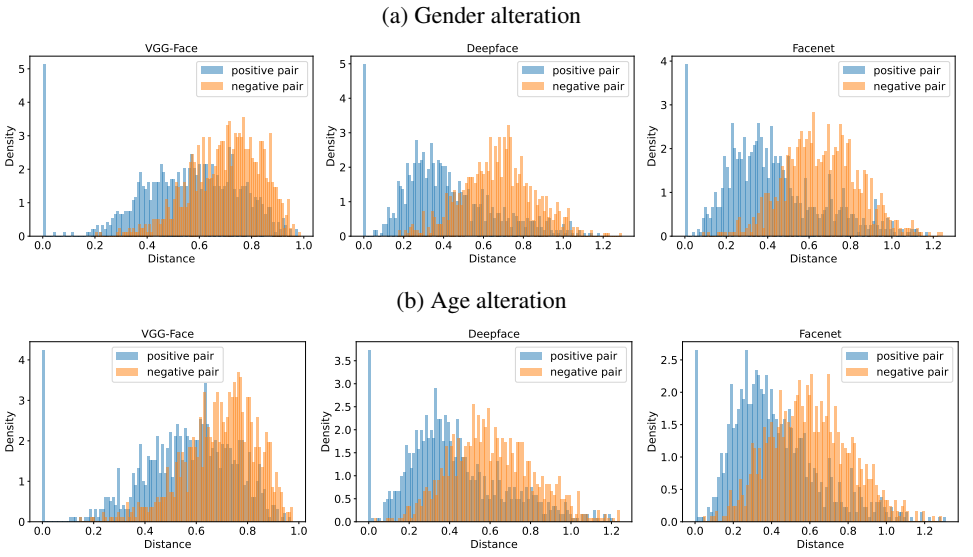

Figure 2: The density distribution of distance scores for both positive and negative pairs, where the negative pairs consist of deepfake images generated from the same user with (a) gender alteration and (b) age alteration.

an interesting trade-off in model design and training. Models like VGG-Face that are highly tuned for positive verification might sacrifice some sensitivity to the anomalies introduced by deepfake techniques, whereas models such as FaceNet and DeepFace, possibly due to their reliance on different learning principles such as triplet loss (FaceNet) and extensive negative sampling, might be better at detecting deepfakes but at the cost of lower positive pair accuracy.

Model	Accuracy of Negative Pair (same users)		Accuracy of Negative Pair (different users)	
VGG-Face [9]	0.633 0.609	0.621	0.995 0.994	0.995
FaceNet [12]	0.929 0.818	0.874	1.00 0.99	0.99
DeepFace [13]	0.768 0.761	0.765	0.947 0.952	0.950

Table 2: Accuracy scores for negative pairs, where the first type consists of negative pairs sampled from the same users but deepfake-generated, and the second type consists of negative pairs sampled from different users. Each model is tested against deepfakes created through gender and age alterations. For each model, the top row displays the results for gender alteration deepfakes, while the bottom row shows the results for age alteration deepfakes. The number following these results represents the average accuracy across both types of negative pairs.

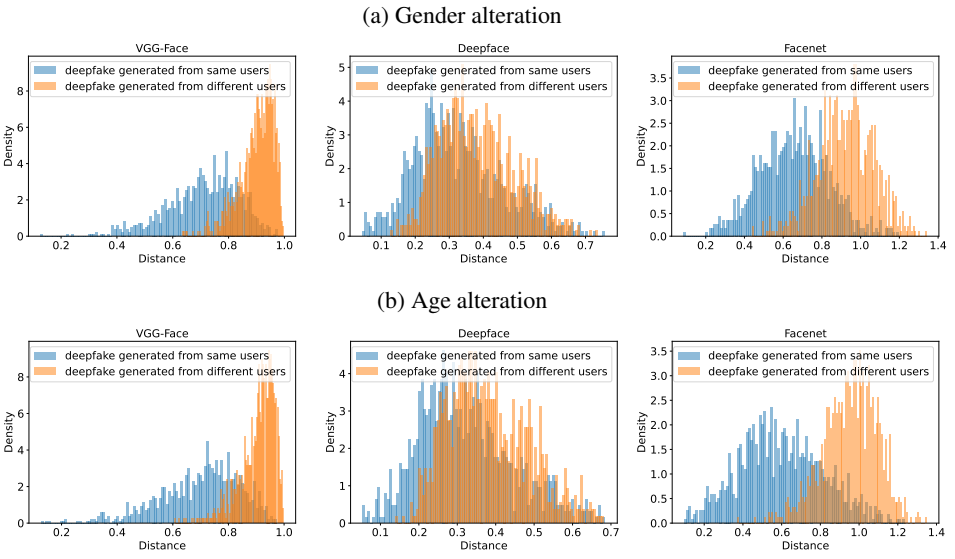

Figure 3: The density distribution of distance scores for negative pairs, categorized into two types: the first type consists of deepfake images generated from the same user, and the second type includes deepfake images generated from different users. Additionally, each type involves two kinds of deepfake manipulations: (a) gender alteration and (b) age alteration.

Different faces verification The results in Table 2 illustrate a significant variation in the performance of pretrained face verification models when distinguishing between negative pairs generated from the same users and those from different users. Models like VGG-Face, FaceNet, and DeepFace demonstrate a notably higher accuracy in identifying deepfake images when they originate from different users. For instance, VGG-Face achieves an accuracy close to 0.995, FaceNet reaches perfect or near-perfect accuracy of 1.00 or 0.99, and DeepFace also performs well with accuracies around 0.947 to 0.952. These high accuracy rates confirm the models' effectiveness in recognizing and rejecting deepfakes that do not match the biometric data of the registered user. However, the challenge intensifies when the deepfakes are generated from the same users. The accuracy for such scenarios drops significantly: VGG-Face averages around 0.621, FaceNet at 0.874, and DeepFace at 0.765. This decline highlights a critical vulnerability; the models struggle to detect subtle manipulations when deepfakes mimic the genuine user's facial features closely. This difficulty arises because the deepfakes retain many biometric markers that are identical to the original user, making it challenging for the models to discern the subtle inconsistencies typically used to identify forgeries.

Discussion The disparity in performance between recognizing genuine users and detecting deepfakes, as revealed in the results, highlight the need for tailored enhancements to current face verification models. To improve the reliability of these models in rejecting deepfakes while affirming true user identities, one enhancement could involve integrating advanced anomaly detection algorithms that focus on finer deviations in facial features that are characteristic of deepfakes but absent in genuine images. Machine learning techniques such as unsupervised learning or semi-supervised learning could be utilized to better understand the

boundary between genuine and manipulated images without relying solely on labeled training data. Additionally, the incorporation of adversarial training, where models are routinely challenged with new types of deepfakes during their training phase, could significantly improve their resilience.

From a security standpoint, these findings illuminate crucial vulnerabilities within digital identity verification systems, particularly in their ability to cope with advanced deepfake technologies. The difficulty models face in distinguishing between real users and their deepfake counterparts from the same user poses a serious threat to user privacy. If malicious actors can easily bypass such systems using deepfakes, it could lead to unauthorized access to sensitive personal information, financial fraud, or misrepresentation in digital media. Strengthening these systems is imperative not only to protect individual user privacy but also to uphold the integrity and trustworthiness of digital platforms across various sectors, including banking, social media, and national security. Proactively addressing these challenges will be essential as the technology behind deepfakes continues to evolve and become more accessible.

4 Conclusion

Our experimental results of pretrained face verification models against both genuine and deepfake images has revealed significant challenges in accurately distinguishing between authentic users and advanced deepfakes, especially when the forgeries closely mimic the facial features of genuine users. This research underscores the necessity for targeted enhancements to bolster the robustness of these systems, recommending the integration of advanced anomaly detection algorithms that focus on subtle deviations typical of deepfakes and the adoption of adversarial training methods to enhance the models' discriminative capabilities. These improvements are essential not only for maintaining the accuracy of face verification processes but also for ensuring the security of digital identity verification systems across various applications. The profound security implications of these findings highlight the current vulnerability of these models to deepfake attacks, posing significant risks to user privacy and security, and underlining the potential for misuse in critical sectors such as access control, financial services, and social media. Strengthening these systems against the growing sophistication of deepfake technologies is imperative, thereby safeguarding individual privacy and maintaining the integrity of digital interactions across platforms. As we navigate the evolving landscape of digital threats, developing more resilient face verification technologies becomes a crucial priority for researchers and developers.

Acknowledgment

This research is supported by the Ministry of Education, Singapore, under its Academic Research Tier 1 (Grant number: GMS 956).

References

[1] Sim Wei Xiang Calvert and Pai Chet Ng. Leveraging transfer learning for region-specific deepfake detection. In *Proceedings of the 2024 IEEE Region 10 Conference*

(TENCON) - Special Session on Ethical and Technical Challenges of Deepfakes and Disinformation. IEEE, 2024.

[2] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, June 2018. doi: 10.1109/CVPR.2018.00916.

[3] Anton Firc, Kamil Malinka, and Petr Hanáček. Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. *Heliyon*, 9(4), 2023.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[5] Md Rezwan Hasan, Richard Guest, and Farzin Deravi. Presentation-level privacy protection techniques for automated face recognition—a survey. *ACM Computing Surveys*, 55(13s):1–27, 2023.

[6] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[7] 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.

[8] Asad Malik, Minoru Kuribayashi, Sani M Abdullahi, and Ahmad Neyaz Khan. Deepfake detection for human face images and videos: A survey. *Ieee Access*, 10:18757–18775, 2022.

[9] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[10] Mayank Kumar Rusia and Dushyant Kumar Singh. A comprehensive survey on techniques to handle face identity threats: challenges and opportunities. *Multimedia Tools and Applications*, 82(2):1669–1748, 2023.

[11] Anil Kumar Sao and B. Yegnanarayana. Face verification using template matching. *IEEE Transactions on Information Forensics and Security*, 2(3):636–641, Sep. 2007. ISSN 1556-6021. doi: 10.1109/TIFS.2007.902920.

[12] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, Los Alamitos, CA, USA, jun 2015. IEEE Computer Society. doi: 10.1109/CVPR.2015.7298682. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298682.

[13] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, June 2014. doi: 10.1109/CVPR.2014.220.

[14] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.

[15] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22412–22423, 2023.

[16] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003.