Exploring the Janus Face of Synthetic Images: From Privacy-secure Biometrics to Universal and Robust Deepfake Detection

Tanusree Ghosh

PhD Supervisor: Dr. Ruchira Naskar,
Dept. of Information Technology, Indian Institute of Engineering Science and Technology, Shibpur, India
2021itp001.tanusree@students.iiests.ac.in

Abstract

The rise of generative AI presents a profound duality. On one hand, it offers a powerful solution to data scarcity and privacy challenges in biometrics. On the other, it is weaponized to create deepfakes that threaten digital integrity. Existing detectors for these deepfakes are brittle, failing against real-world transformations and novel generative models. This dissertation confronts this duality head-on. First, I establish the viability of synthetic data for building fair and private biometric systems. Second, to counter the malicious use of this technology, this dissertation develops deepfake detectors designed to be robust, generalizable, and efficient by construction. My work introduces novel, lightweight feature sets on different cues (e.g., colour cue-based Relative Chrominance Difference, Gradient features, Depth cues, etc.) that are inherently resilient to OSN transformations and improve generalisation to unseen forgeries. Whereas, accomplished results confirm state-of-the-art performance, achieving high accuracy in challenging real-world scenarios with a significant reduction in model complexity, my current and future work focuses on achieving superior generalisation while being OSN manipulation resistant.

Introduction and Motivation

The advent of generative AI has produced synthetic images of such remarkable realism that they are often indistinguishable from authentic photographs. This technological leap presents a classic "Janus Face": it simultaneously unlocks unprecedented opportunities for innovation while introducing formidable threats to societal trust and security.

The positive face lies in its potential to solve critical data challenges. Training effective and fair biometric models requires vast, diverse datasets, but collecting real facial images is fraught with privacy and ethical challenges. My dissertation explores the hypothesis that high-quality synthetic data can serve as a privacy-preserving substitute.

Conversely, the negative face is its malicious deployment in the wild. Adversaries use these same models to create "Deepfakes" for misinformation campaigns on Online Social Networks (OSNs). However, current detectors suffer from key weaknesses: vulnerability to OSN transformations, poor generalization to new generative models, and high computational complexity. To safely harness the benefits of generative AI, we must build strong defenses against its misuse.

Dissertation Thrusts and Progress

My research is organised into two primary thrusts that mirror the dual complementary theme of the dissertation.

Thrust 1: Synthetic Data as a Privacy-Enhancing Tool for Biometrics (Accomplished Work)

This research thrust confronts the inherent tension between the need for large, diverse datasets in biometrics and the critical importance of individual privacy. While prior studies have explored using synthetic data for applications like face recognition, these efforts have often been hindered by the fact that the data generated with GANs has been found to inherit and even amplify the demographic biases present in the real data used to train them, thereby limiting their effectiveness in building fair systems.

In response, my work (Ghosh et al. 2025) leverages the high-fidelity and precise control of modern text-to-image diffusion models (DM). It is driven by the **substitutability hypothesis**: *Can DM generated synthetic data serve as a viable substitute for real-world data in training robust and fair biometric models?* To investigate this, we developed *DiffSynFace*, a large-scale dataset of over 40,000 demographically diverse faces generated using seven state-of-the-art diffusion models, covering a wide range of ethnicities, ages, and genders [1].

The primary outcome of this work is the validation that synthetic data is a powerful alternative to real facial data. Our experiments show that state-of-the-art biometric classifiers (age, gender and ethnicity) trained on *DiffSynFace* can match, and in key cross-dataset scenarios, even exceed the generalisation performance of models trained on real data. This thrust pioneers a privacy-by-design approach for developing the next generation of secure and fair biometric systems.

[1] https://github.com/tanusreeg/DiffSynFace

Thrust 2: Robust and Efficient Synthetic Media Detection (Accomplished Work)

The foundational work of my thesis has focused on creating detectors that are resilient to the degradations common in OSN environments. I have explored different image-representation domains to solve this problem.

Robust Feature Engineering for OSN Contexts: The philosophy of such work is that generative models, optimised for photorealism in the pixel domain, are less constrained in other domains. My contributions lie in identifying and exploiting these domains.

- **RCD-Net:** In (Ghosh and Naskar 2024), I worked on the **Relative Chrominance Distance (RCD)** feature. The insight is that generators faithfully replicate luminance but not the subtle relationships between color channels. This minimalist feature allows a lightweight CNN to outperform heavy-parameterised models, achieving 98.86% accuracy on detecting synthetic face images with less than 1% of the parameters of competitors.
- **Gradient Features:** To counter pixel-level perturbations from OSNs, my work in (Ghosh and Naskar 2023a) uses features derived from image gradients. By learning from the gradient magnitude and direction, the detector focuses on stable image structure rather than volatile pixel values, achieving an average accuracy of 96.54% against common post-processing attacks.
- **Exploiting Spatial Artifacts:** My other works in this thrust, **STN-Net** (Ghosh and Naskar 2023b) and **LPQ-Net** (Kundu, Ghosh, and Naskar 2024), focus on amplifying high-frequency artifacts and leveraging blur-invariant phase information, respectively, to ensure high performance on heavily compressed images. LPQ-Net maintains over 96% accuracy on images downloaded directly from OSNs like Facebook and Instagram.

Multi-Modal and Generalizable Architectures: I have also worked to develop architectures for efficiency and explainability in video deepfake detection. My work on **DepthFakeNet** (under review) uses 3D geometric inconsistencies for more interpretable detection. The **Multi-Level Feature Fusion (MLFF)** framework (Ghosh and Naskar 2025) sets a new state-of-the-art performance (98.99% AUC on the FF++ dataset) while reducing model parameters by over 50% through efficient multi-scale feature fusion.

To broaden generalization beyond faces, my work demonstrates that augmenting training data with distortions like blur and noise significantly improves performance on unseen DM images (Das et al. 2023). To facilitate robust, real-world analysis, we also developed a dataset variant of DiffSynFace by processing images through major OSNs to capture the effects of platform-specific compression (Ghosh et al. 2024).

Current Progress (as of Sept 30, 2025)— A Leap to Universal Generalization: My current work targets a universal fingerprint of the image synthesis process (beyond faces). Under the guidance of my supervisor, I have developed a simple, threshold-based feature extraction method that exploits artefacts inherent to the upsampling layers found in nearly all generative models. In an extreme generalization test—training on a limited ProGAN dataset(4 classes) and testing on 38 unseen models—this feature boosts the performance of standard detectors (e.g., ResNet-50, ViT etc.) by up to 42% over raw RGB input, surpassing the state-of-the-art under the same training protocol. This work is currently under review. Its primary limitation is reduced performance on heavily JPEG-compressed images, which directly motivates my future research on robustness.

Future Work: Unifying Generalization, Robustness, and Ethical Application: My future work will focus on creating an unified synthetic image detection framework that is robust to real-world distortions, especially high-level JPEG compression. I am planning to achieve this by mathematically modelling JPEG artefacts to design a novel feature extractor. Fusing this with existing/ novel artefact-based detectors will create a single, **unified framework** invariant to both the generative source and post-generation compression. Concurrently, I am planning to explore the broader application of synthetic media in biometrics, such as **emotion recognition**. This involves creating a new synthetic dataset to analyze and proactively **mitigate algorithmic bias**, advancing media forensics while contributing to the ethical application of generative AI.

References

Das, S.; Dutta, D.; Ghosh, T.; and Naskar, R. 2023. Universal detection and source attribution of diffusion model generated images with high generalization and robustness. In *International Conference on Pattern Recognition and Machine Intelligence*, 441–448. Springer.

Ghosh, T.; and Naskar, R. 2023a. Leveraging Image Gradients for Robust GAN-Generated Image Detection in OSN context. In *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 1–5. IEEE.

Ghosh, T.; and Naskar, R. 2023b. Stn-net: A robust gan-generated face detector. In *International Conference on Information Systems Security*, 141–158. Springer.

Ghosh, T.; and Naskar, R. 2024. Less is more: A minimalist approach to robust GAN-generated face detection. *Pattern Recognition Letters*, 179: 185–191.

Ghosh, T.; and Naskar, R. 2025. Multi-level feature fusion for generalized face forgery detection. *Neurocomputing*, 653: 131235.

Ghosh, T.; Saha, T.; Naskar, R.; et al. 2024. Identifying Diffusion Model Generated Synthetic Faces Covering OSN Context and Ethnic Diversity. In *2024 IEEE 21st India Council International Conference (INDICON)*, 1–6. IEEE.

Ghosh, T.; Seth, B.; Kar, S.; and Naskar, R. 2025. Evaluating the substitutability of generative AI-generated faces in biometric applications: From a lens of age, gender, ethnicity detection. *Pattern Recognition Letters*.

Kundu, S.; Ghosh, T.; and Naskar, R. 2024. Using Local Phase Quantization to Identify Fake Faces in Online Social Networks. In *TENCON 2024-2024 IEEE Region 10 Conference (TENCON)*, 323–326. IEEE.