When **95% Accuracy** Isn't Good Enough: Bridging the Gap Between Forensic Facial Recognition Benchmarks & Real-World Human Impact

UC Berkeley

Justin D. Norman · University of California, Berkeley · justin.norman@berkeley.edu

CVPR 2026 · HUMAN IN THE GENERATIVE AI ERA (HUG) WORKSHOP

CVPR·2026
DENVER · COLORADO
JUNE 11 – 15, 2026

Abstract

Facial recognition technology (FRT) is deployed in **high-stakes settings**, including airports, immigration, law enforcement, yet evaluated in *sanitized* laboratory conditions.

We **design and implement a forensic lineup style evaluation framework**, and show that reported **95%+ accuracies degrade to ~65%** under realistic conditions, then ask whether generative AI can be used to improve them. **We find it often makes things worse.**

We close with a governance framework for responsible biometric deployment.

>95%
Reported lab accuracy
(FaceNet · ArcFace)

~65%
Forensic-lineup accuracy under realistic degradations

> *117M Americans are in searchable face recognition databases without their knowledge or consent.* — Garvie et al., Georgetown Law (2016)

Why This Matters

- High-stakes deployments operate on **low-resolution, occluded, off-pose input images** — very different from lab conditions.
- Generative AI (enhancement & synthetic data) is being layered on top of FRT **without forensic validation**.
- No *public, model-agnostic* forensic evaluation framework previously existed.

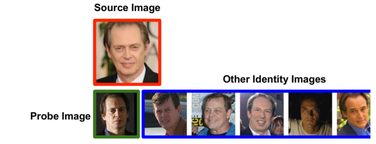
Forensic image degradations on a synthetic face. *Rows: blur kernel width, scale, additive noise, JPEG quality. We test each degradation across realistic forensic ranges to see how recognition holds up.*

(Fig. 2, Norman, Agarwal & Farid 2023)

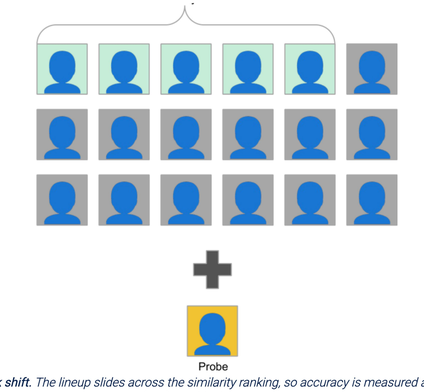
Methods and Materials

A Human-Centered Forensic Framework

Inspired by **eyewitness identification lineups**, the probe image is matched against **perceptually similar decoys** selected by embedding similarity — not random distractors.



Lineup creation. *Probe embedding (ArcFace / FaceNet) → similarity ranking → top-K most-similar identities become decoys. The probe must be distinguished from perceptually similar faces, not random ones.*



Similarity rank shift. *The lineup slides across the similarity ranking, so accuracy is measured at every difficulty level — not just an easy one.*

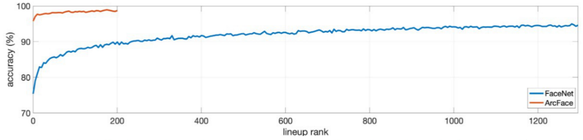
Datasets & Models

- **Synthesis AI** (~200K imgs · 8K identities; commercial procedural pipeline) + **CASIA-WebFace** (~494K imgs · ~10.5K identities).
- **FaceNet** & **ArcFace** embeddings.
- Stress-tested across degradation in **resolution, blur, noise, JPEG, gamma, head pose** and **occlusion**.

Results

HEADLINE

Lab-grade FRT collapses under forensic conditions.



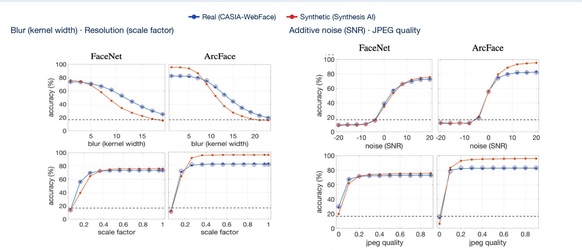
Accuracy vs. lineup difficulty. *As decoys become more perceptually similar to the probe (smaller rank), accuracy drops — sharply for FaceNet.*

73.1%
FaceNet · real faces
(forensic lineup)

82.7%
ArcFace · real faces
(forensic lineup)

63%
Worst case: occlusion + rotation + low light

Sensitivity to Degradation



Accuracy under degradation. *Even SOTA ArcFace collapses under moderate blur, low resolution, additive noise, and aggressive JPEG — failure modes invisible in lab benchmarks.*

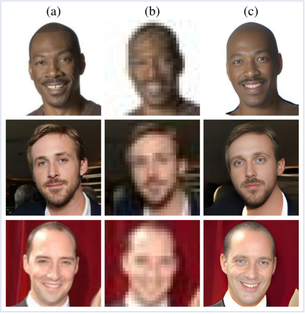
Demographic Gap (FaceNet, Baseline Accuracy)

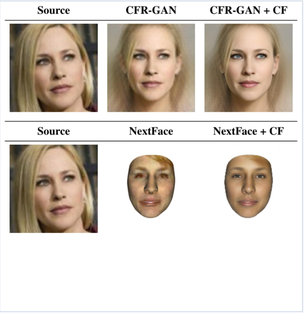
81.0%	76.0%	74.1%	75.6 / 75.9%
White (n=1,659)	Black (n=208)	All other (n=5,633)	Male / Female (no gap)

The lineup framework surfaces racial disparities that standard benchmarks obscure.

Discussion

Generative AI: Promise & Peril





(Left) *Super-resolution can hallucinate facial features — enhancement looks compelling but changes the identity (Fig. 4, Norman & Farid, CVPRW '24).* ***(Right)*** *Head-pose correction via CFR-GAN and NextFace produces frontalized faces whose biometric features no longer match the source (Fig. 5, Norman & Farid 2025).*

- **Image enhancement** (super-resolution, deblurring, head-pose correction) **often degrades** recognition. Naïve head-pose correction collapsed ArcFace from **89.5% → 43.4%**; **selective** restoration recovers to **92.8%**.
- **Synthetic training data** introduces two emergent harms: *diversity-washing* and *consent circumvention*. Microsoft FaceSynthetics generates 100K "diverse" images from only **511 base scans** (~30 Black men) — statistical ≠ representational diversity.

Conclusions

A four-pillar framework for responsible deployment:

- **Authority.** Decision-making power over deployment must be made explicit and contestable rather than left implicit in vendor contracts.
- **Limits.** The community must codify clear criteria for the forensic conditions under which FRT should *not* be deployed at all.
- **Audit.** Demographic auditing must be continuous and conducted under forensic conditions, not laboratory conditions, to surface real-world disparities.
- **Expertise.** Responsible deployment requires standing multidisciplinary teams that span legal, ethical, and technical expertise.

> *Understand the technical limits and sociotechnical harms **before** deployment — not after.*

CONTACT
Justin D. Norman
University of California, Berkeley
justin.norman@berkeley.edu

justintime.ai

SELECTED REFERENCES

1. Norman, Agarwal & Farid. *An Evaluation of Forensic Facial Recognition.* arXiv:2311.06145, 2023.
2. Norman & Farid. *An Investigation into the Impact of AI-Powered Image Enhancement on Forensic Facial Recognition.* IEEE/CVF CVPRW, pp. 4306–4314, 2024.
3. Norman & Farid. *Does Head Pose Correction Improve Biometric Facial Recognition?* arXiv:2512.03199, 2025.
4. Norman & Farid. *Detecting Deepfake Talking Heads from Facial Biometric Anomalies.* IEEE/CVF CVPRW, 2025.
5. Whitney & Norman. *Real Risks of Fake Data: Synthetic Data, Diversity-Washing & Consent Circumvention.* ACM FAccT, 2024.
6. Buolamwini & Gebru. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.* ACM FAccT, pp. 77–91, 2018.
7. Deng, Guo, Xue & Zafeiriou. *ArcFace: Additive Angular Margin Loss for Deep Face Recognition.* IEEE/CVF CVPR, pp. 4690–4699, 2019.
8. Schroff, Kalenichenko & Philbin. *FaceNet: A Unified Embedding for Face Recognition and Clustering.* IEEE/CVF CVPR, pp. 815–823, 2015.
9. Garvie, Bedoya & Frankle. *The Perpetual Line-Up: Unregulated Police Face Recognition in America.* Georgetown Law Center on Privacy & Technology, 2016.