

# When 95% Accuracy Isn't Good Enough.

Bridging the gap between biometric facial recognition benchmarks and real-world human impact.

**Justin D. Norman**

University of California, Berkeley

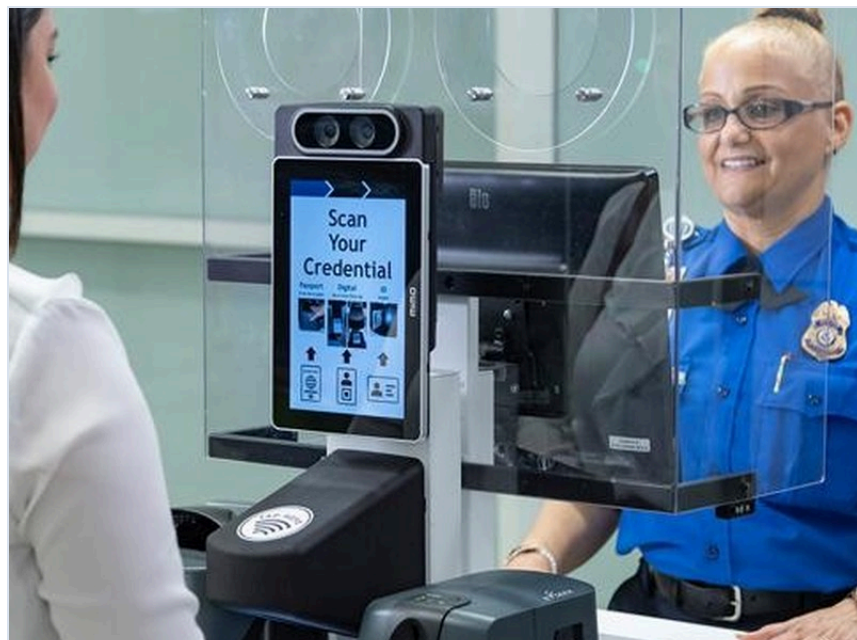
CVPR 2026 · HuG Workshop

# 117M

Americans in searchable facial-recognition databases *without their knowledge or consent*

Garvie, Bedoya & Frankle, Georgetown Law CPT (2016)

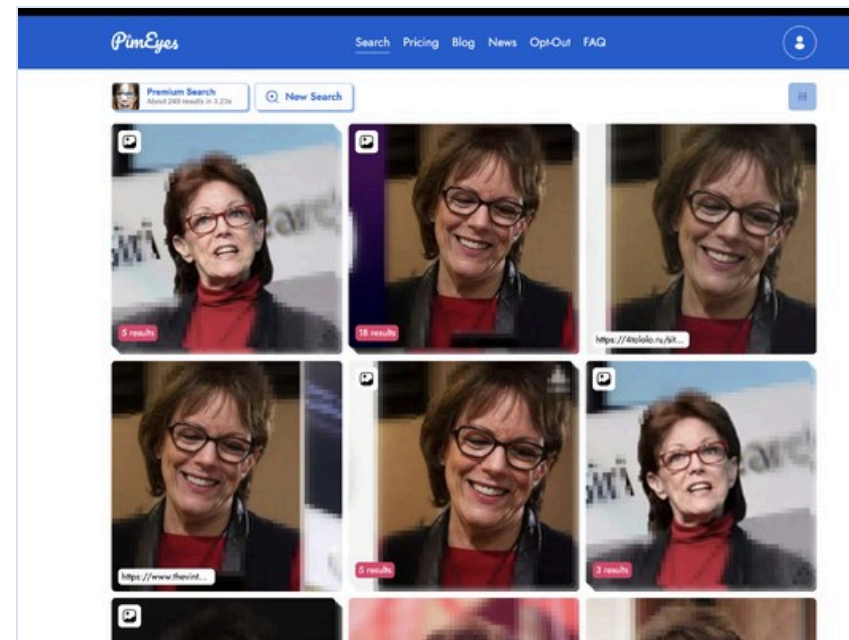
- FRT now operates across **airports, immigration, law enforcement, and consumer devices**, despite documented failures and significant public opposition.
- Current evaluations **fail to capture real-world, high-stakes scenarios** and rarely differentiate between use cases.
- Calls for abolition, while important, *don't immediately impact* people **already in the system** or already harmed by it.
- A massive **information & expertise gap** sits between government, FRT vendors, and the public.



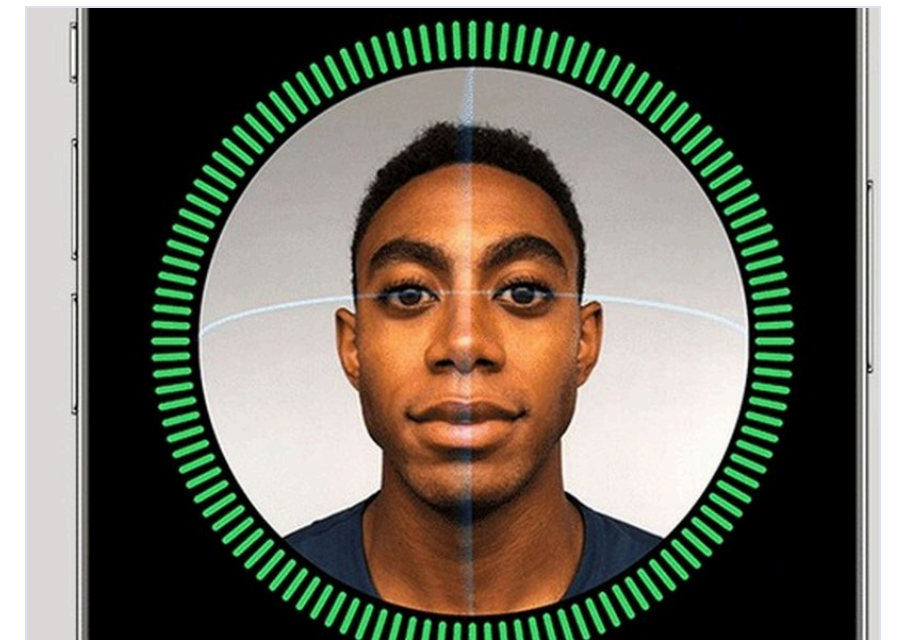
AIRPORTS



LAW ENFORCEMENT



CONSUMER SEARCH



PERSONAL DEVICES

# A Forensic Evaluation for Biometric FRT

Source Image



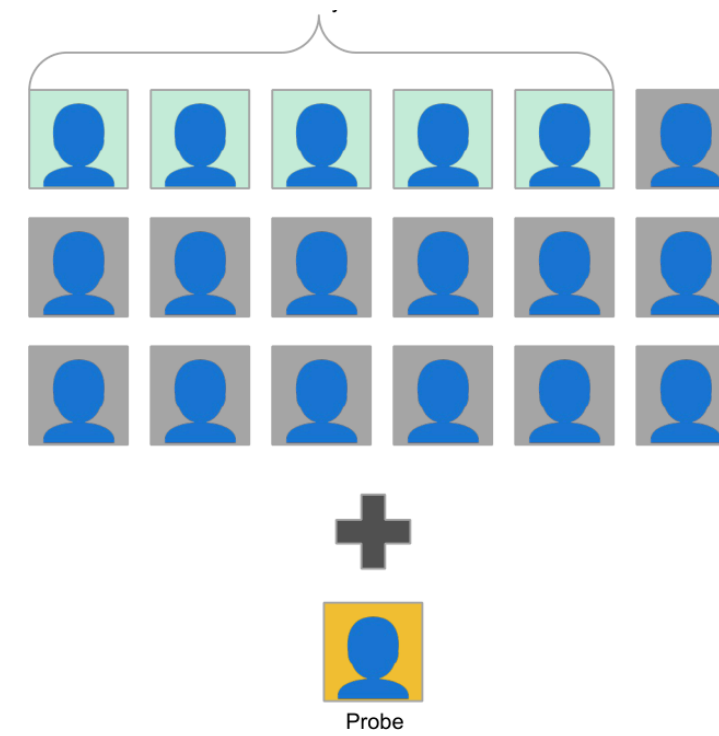
Other Identity Images



Probe Image



*Probe vs. perceptually similar decoys, not random distractors.*



*Lineups slide across the similarity ranking, so accuracy is measured at every difficulty.*

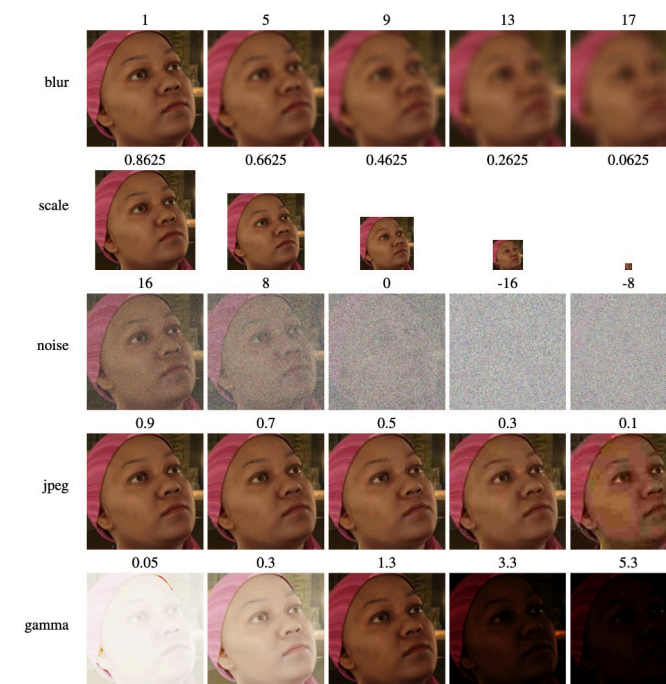


Fig. 2. Examples of image degradation applied to a synthetically-generated face. See Fig. 3.

*Parametric image degradation: blur, scale, noise, JPEG, gamma at varied levels.*

Inspired by eyewitness identification: a probe is matched against five **perceptually similar** decoys, and the lineup is shifted across the similarity ranking so the test reflects **biometric difficulty**, not laboratory difficulty.

# Lab-Grade FRT Collapses Under Forensic Conditions

>95%

Reported lab accuracy



-30 PP ON AVERAGE

~65%

Forensic-lineup accuracy

WHAT THAT AVERAGE HIDES:

82.7%

ArcFace

real faces · forensic lineup

73.1%

FaceNet

real faces · forensic lineup

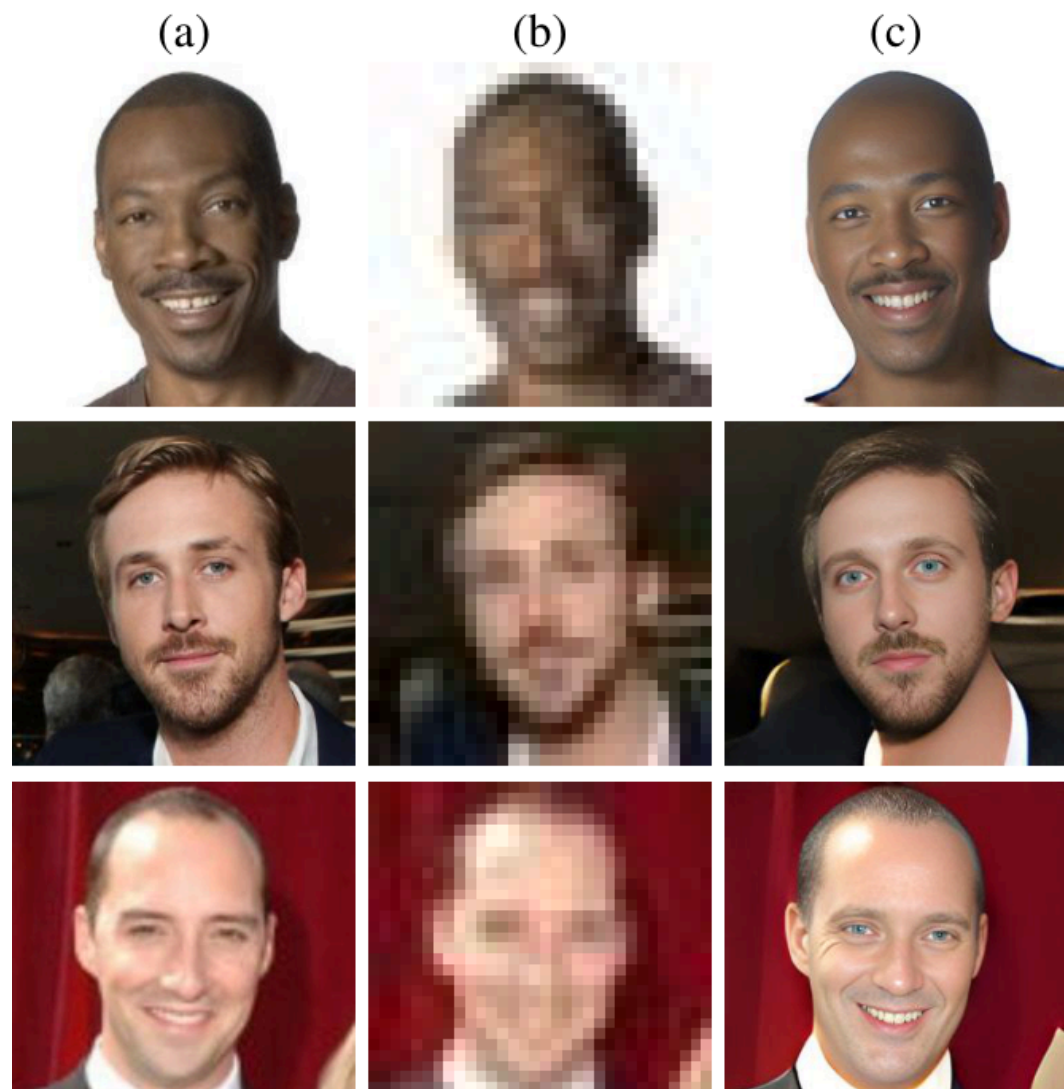
63%

Worst case

occlusion + rotation + low light

Demographic gap (FaceNet baseline): White 81% Black 76% Other 74% ~7 pp gap on a single model.

# And Generative AI Often Makes It Worse



(a) original · (b) low-resolution · (c) AI-enhanced. Visually compelling, but identity has changed.

**Enhancement** can hallucinate features with no obvious tell.

NAÏVE HEAD-POSE CORRECTION

**ArcFace 89.5% → 43.4%**

Recognition accuracy nearly halved after a routine pre-processing step.

**Synthetic training data** introduces emergent harms.

*Diversity-washing* and *consent circumvention*. Microsoft FaceSynthetics generated **100K "diverse" images from just 511 base scans**.

Abstract

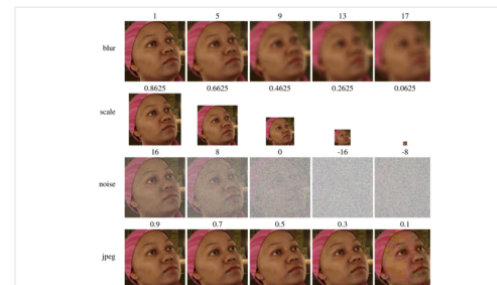
Facial recognition technology (FRT) is deployed in **high-stakes settings**, including airports, immigration, law enforcement, yet evaluated in *sanitized* laboratory conditions. We **design and implement a forensic lineup style evaluation framework**, and show that reported **95%+ accuracies degrade 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.



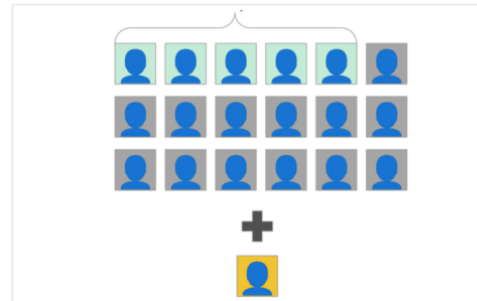
Forensic image degradations on a synthetic face. Rows: blur kernel width, scale, additive noise, JPEG quality. We test each degradation across realistic forensic ranges to see how recognition holds up. (Fig. 2, Norman, Agarwal & Fand 2023)

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

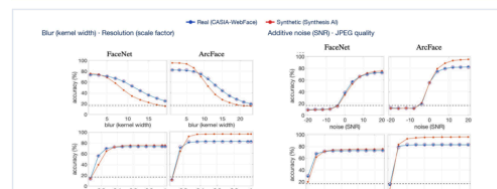
HEADLINE  
Lab-grade FRT collapses under forensic conditions.



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.

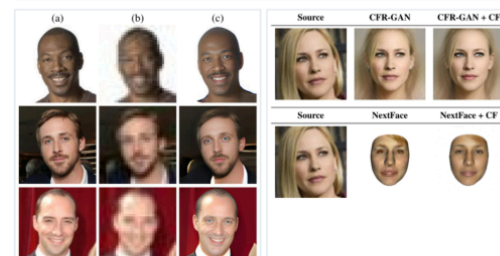
Demographic Gap (FaceNet, Baseline Accuracy)

**81.0%** White (n=1,659) | **76.0%** Black (n=208) | **74.1%** All other (n=5,633) | **75.6 / 75.9%** 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 & Fand, CVPR'24). (Right) Head-pose correction via CFR-GAN and NextFace produces frontalized faces whose biometric features no longer match the source (Fig. 5, Norman & Fand 2025).

- Image enhancement** (super-resolution, deblurring, head-pose correction) **often degrades** recognition. Naive 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.



justintime.ai

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

SELECTED REFERENCES

1. Norman, Agarwal & Fand. An Evaluation of Forensic Facial Recognition. arXiv:2311.06145, 2023.  
2. Norman & Fand. An Investigation into the Impact of AI-Powered Image Enhancement on Forensic Facial Recognition. IEEE/CV CVPR, pp. 4306-4314, 2024.  
3. Norman & Fand. Does Head-Pose Correction Improve Biometric Facial Recognition? arXiv:2312.09196, 2023.  
4. Norman & Fand. Detecting Deepfake Talking Heads from Facial Biometric Anomalies. IEEE/CV CVPR, 2025.  
5. Whitney & Norman. Real Biases of Fake Data: Synthetic Data, Diversity-Washing & Consent Circumvention. ACM Race1, 2024.  
6. Bouammi & Gidycz. Gender-Shed: Intersectional Accuracy Disparities in Commercial Gender Classification. ACM Race1, pp. 17-41, 2018.  
7. Deng, Guo, Xue & Zafeeris. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. IEEE/CV CVPR, pp. 4690-4699, 2019.  
8. Schuff, Kalamchenko & Pihlan. FaceNet: A Unified Embedding for Face Recognition and Clustering. IEEE/CV CVPR, pp. 811-823, 2015.  
9. Garvie, Reddy & Franklin. The Repeatable Line-Up: Unregulated Police Face Recognition in America. Georgetown Law Center on Privacy & Technology, 2016.

Framework, full results, generative-AI failure cases, and a four-pillar governance proposal. See you at my poster!