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

UC Berkeley

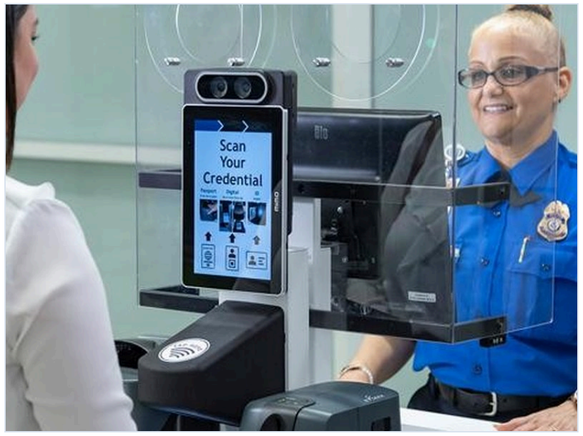
Why This Matters

117M

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

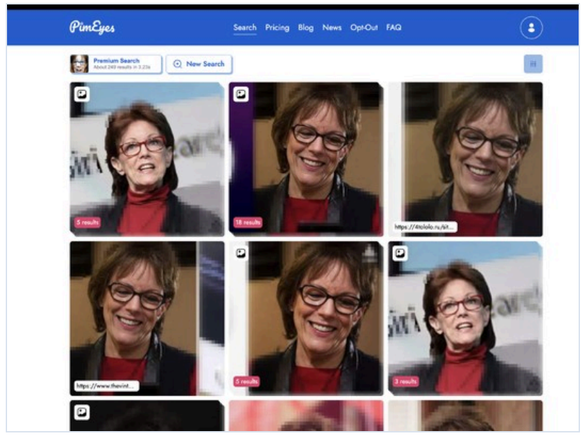
Garvie, Bedoya & Frankle, Georgetown Law CPT (2016)

- 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.

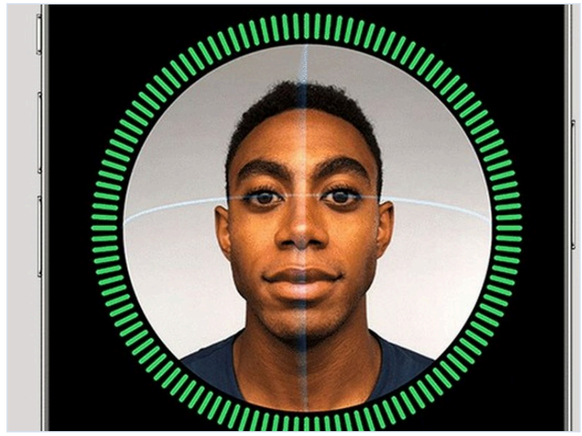


AIRPORTS

LAW ENFORCEMENT



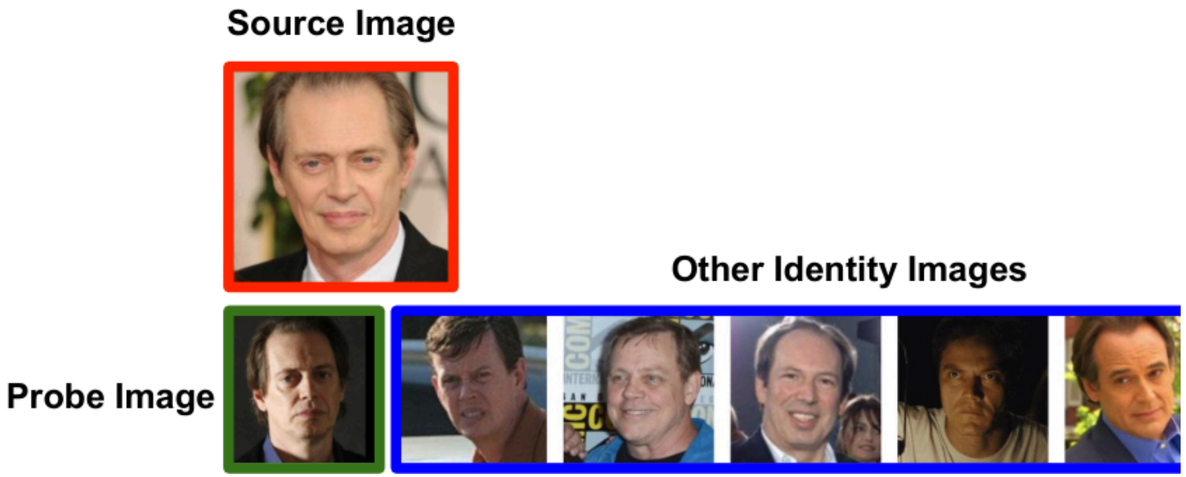
CONSUMER SEARCH



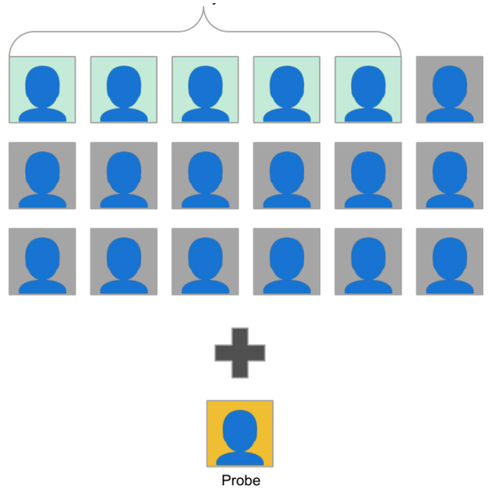
PERSONAL DEVICES

TL;DR No **model-agnostic, publicly accessible** evaluation framework exists for FRT under forensic conditions.

A Forensic Evaluation for Biometric FRT



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



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

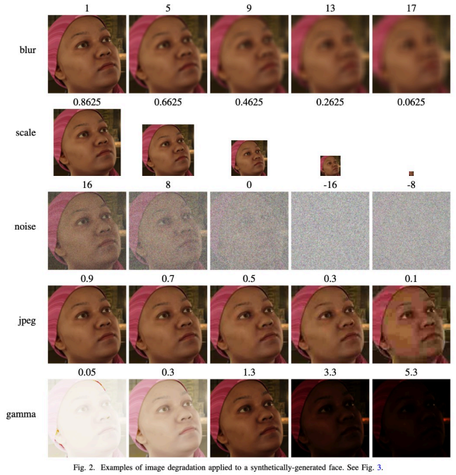


Fig. 2. Examples of image degradation applied to a synthetically-generated face. See Fig. 3.

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%

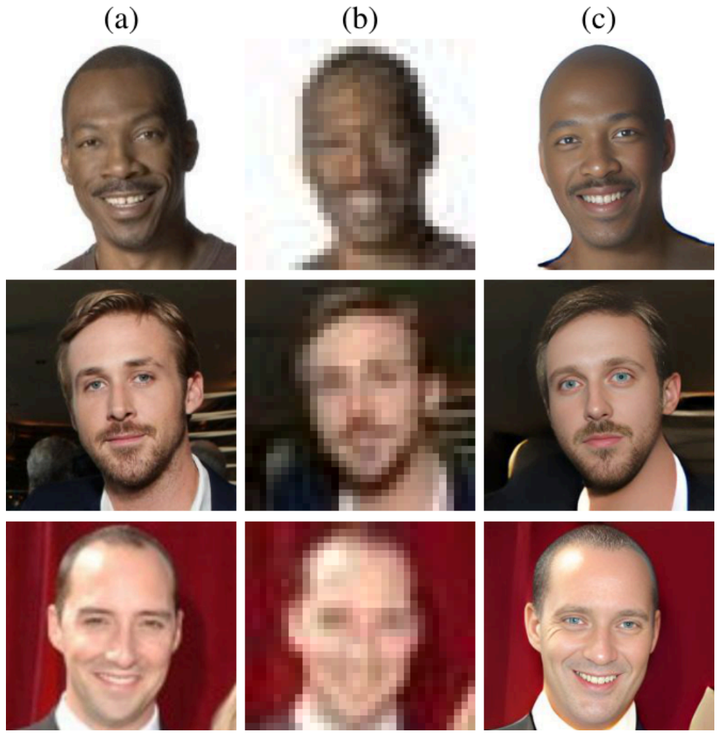
Forensic-lineup accuracy

WHAT THAT AVERAGE HIDES:

82.7%	73.1%	63%
ArcFace	**FaceNet**	**Worst case**
real faces · forensic lineup	real faces · forensic lineup	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.

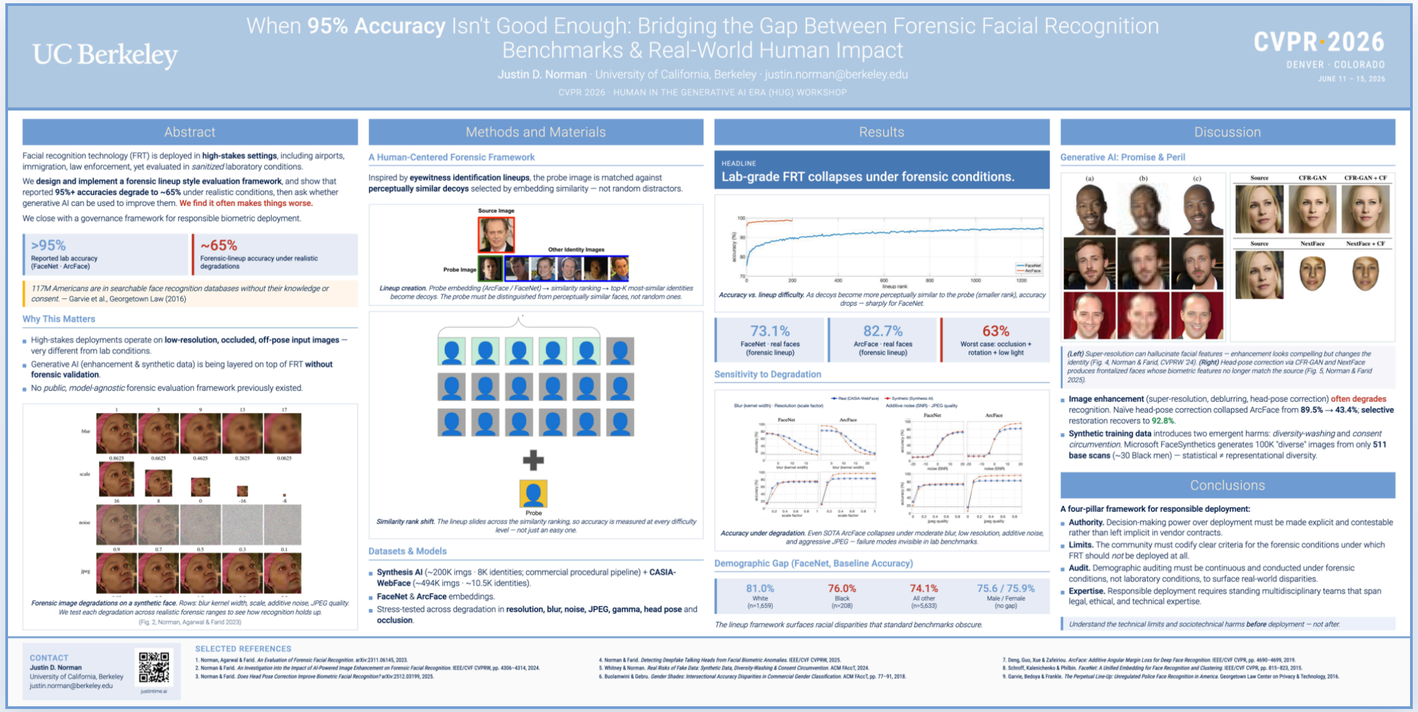
NAÏVE HEAD-POSE CORRECTION

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**.



justintime.ai

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