Is *Your* Face Fake?

Social Impacts of Algorithmic ‘Fake’ Determination

(DFDC Risk-a-thon & CVPR medial forensics workshop 2020)

Ameya Prabhu

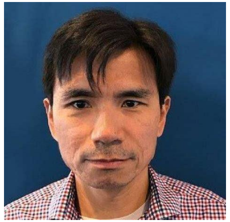
Joanna Materzyńska

Puneet K. Dokania

Philip H. S. Torr

University of Oxford

Ser-Nam Lim

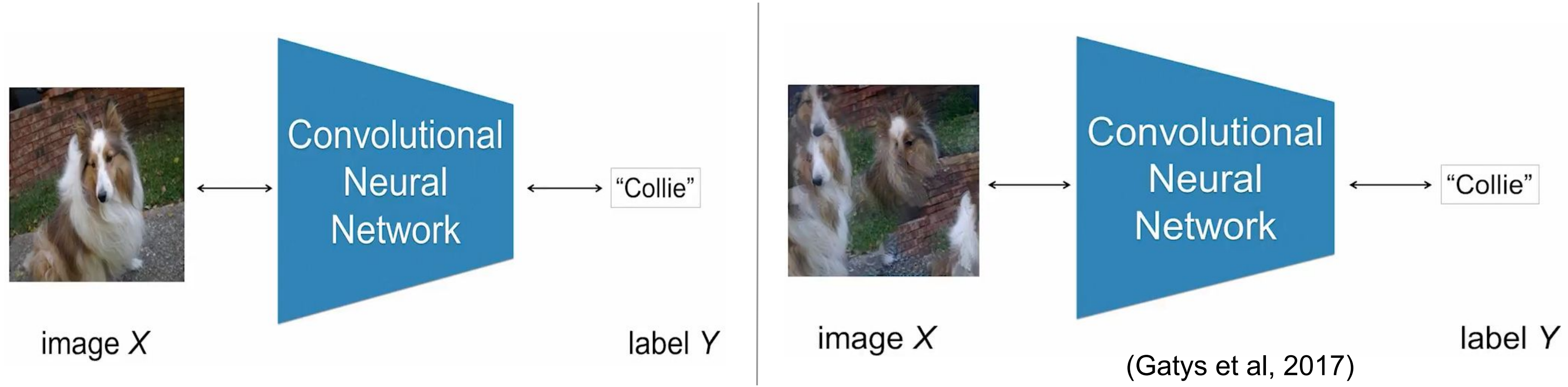


Facebook AI
(Sponsor)

The Big Challenge: Deepfake Detection

☑ Automatic Fake Generation ⟷ Automatic Detection

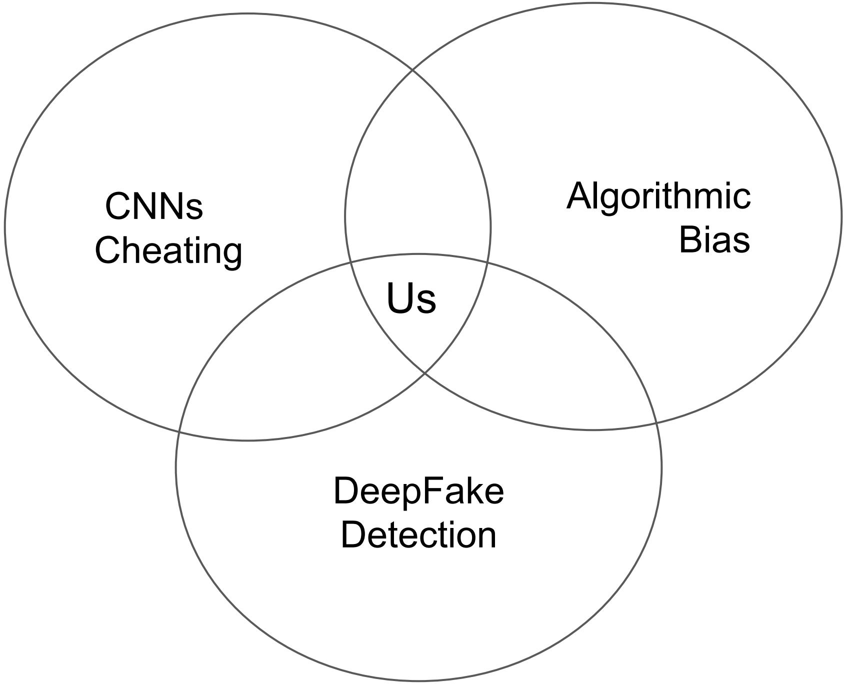
☑ Deep CNNs + large & diverse datasets → ‘Fake’ & ‘Real’ classes



(Gatys et al, 2017)

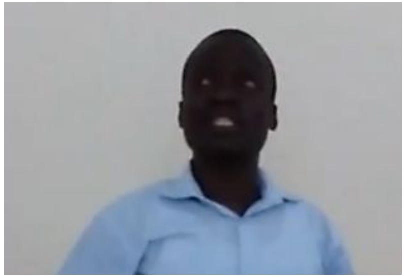
Source: Efros “Making Computers Study Harder”

Our Direction

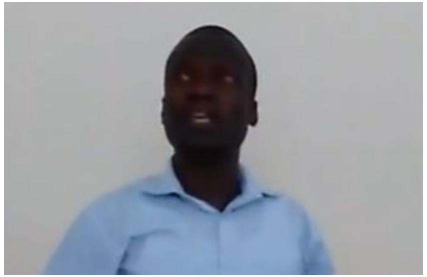


Cheating detection of blending artifacts; which affect already people marginalized by face

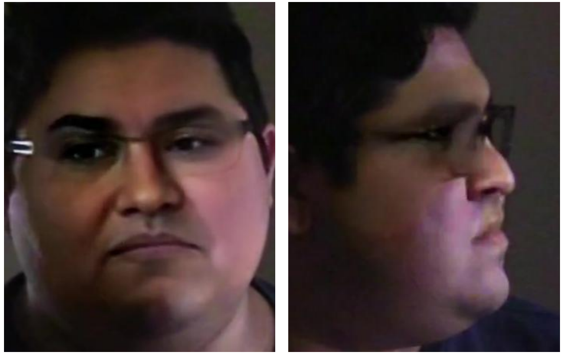
#1: Blending Artifacts → Colour Gradients



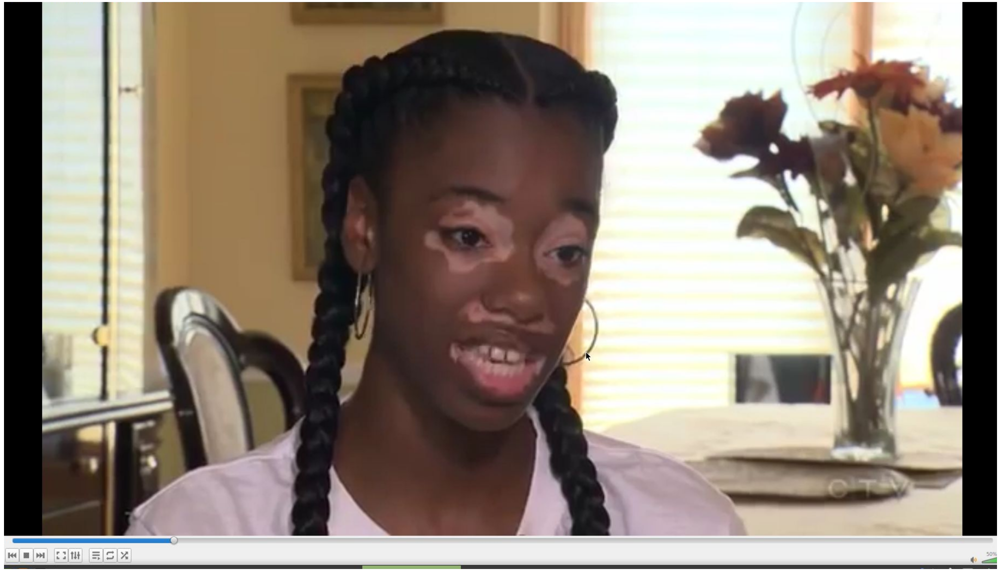
Original



Fake

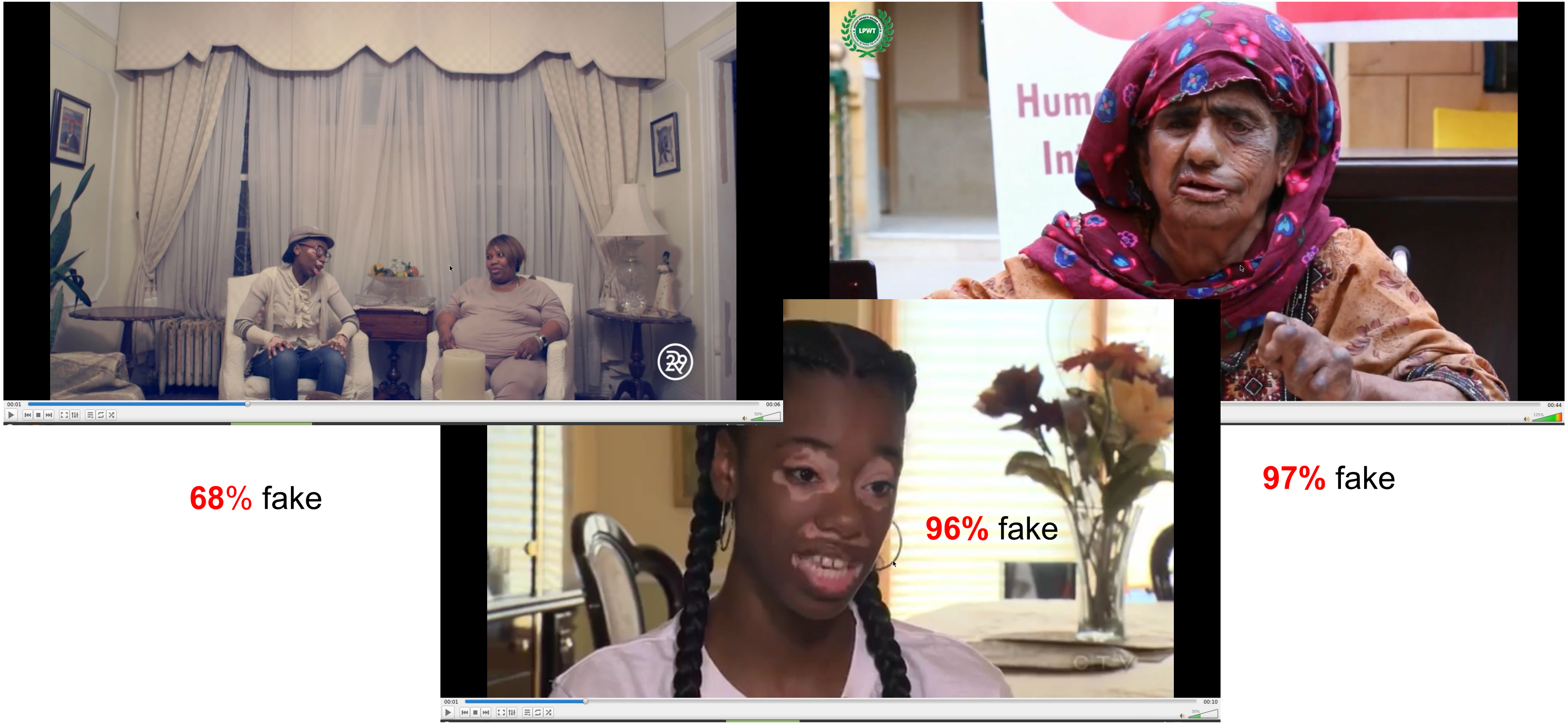


More clear examples



Vitiligo Victims

People with Leprosy & Vitiligo

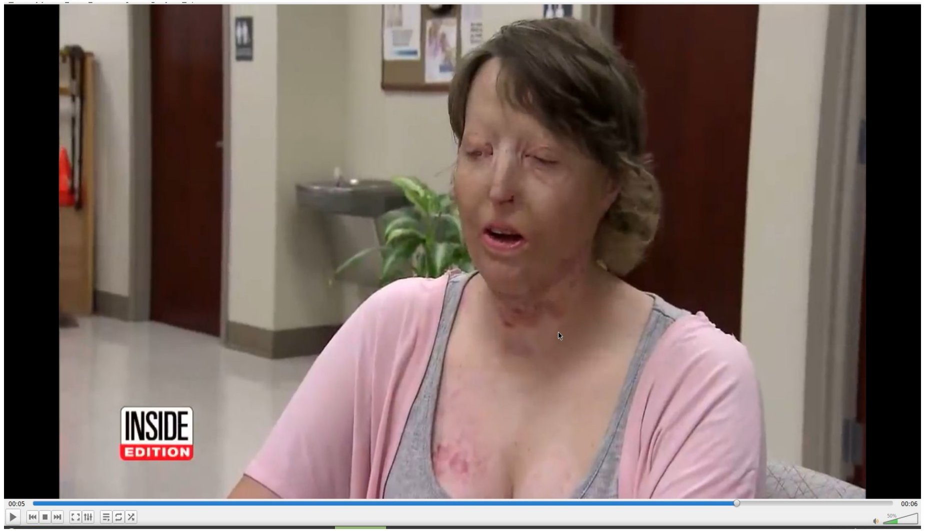


68% fake

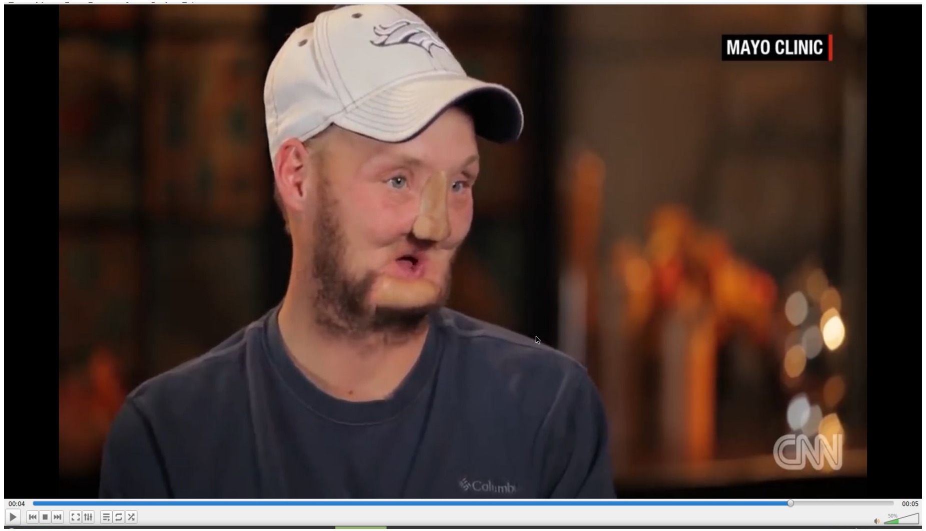
96% fake

97% fake

Burn victims

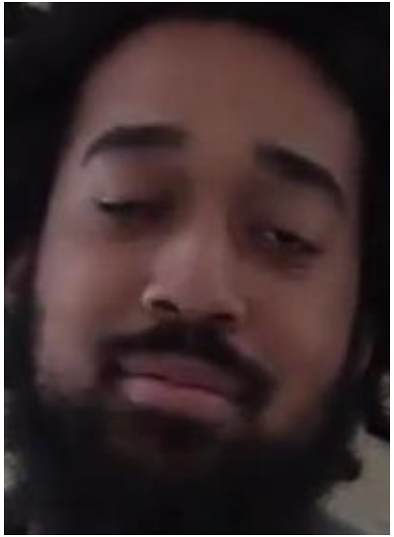


81% fake

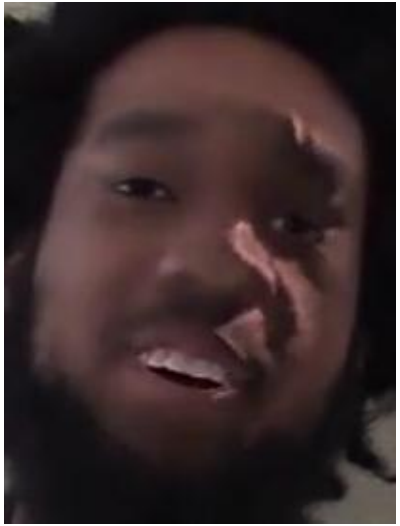


96% fake

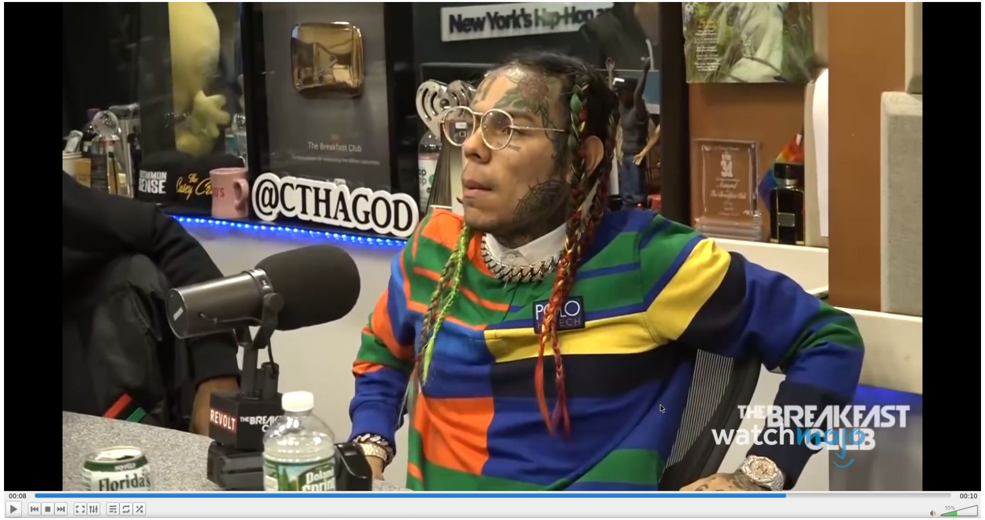
#2: Blending Artifacts → Jaggedness



Original

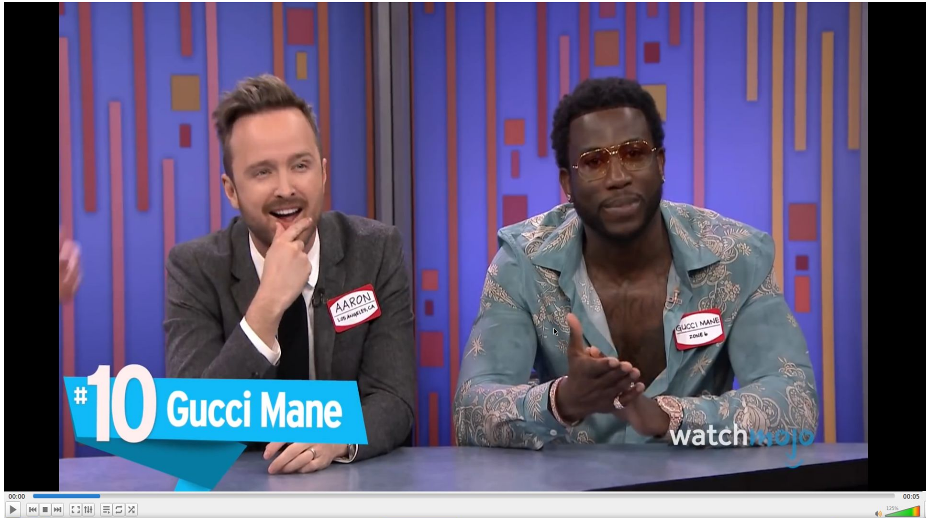


Fake

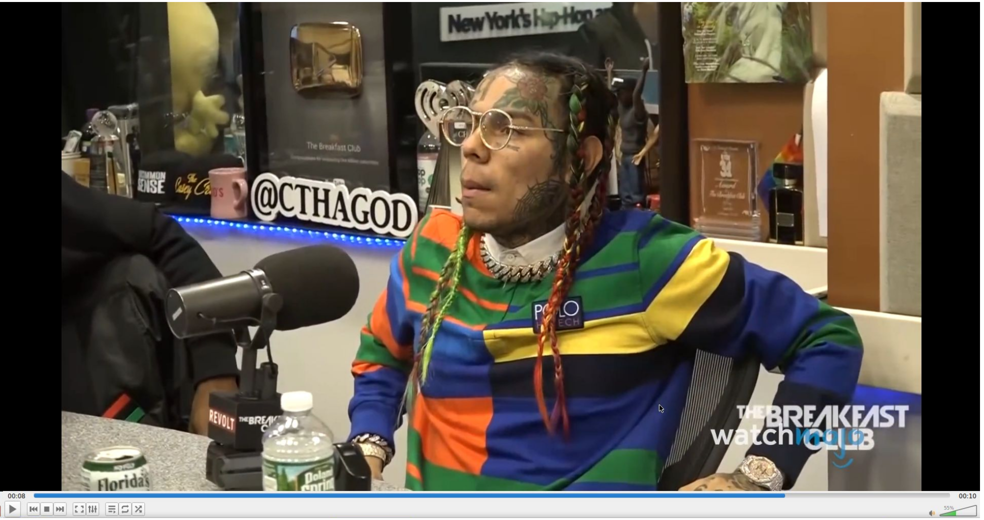


Face Tattoos

(c) Face Tattoos

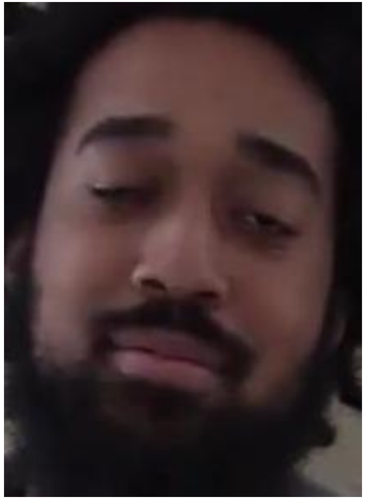


62% Fake

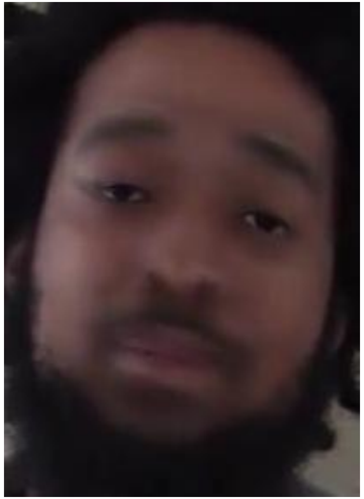


79% Fake

#3: Blending Artifacts → Smooth Faces

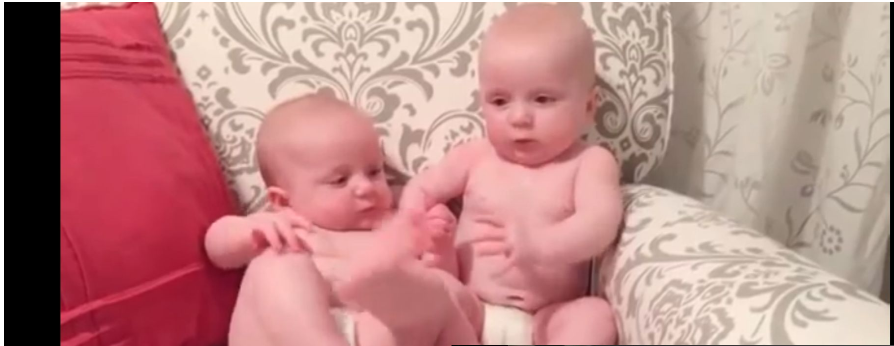


Original

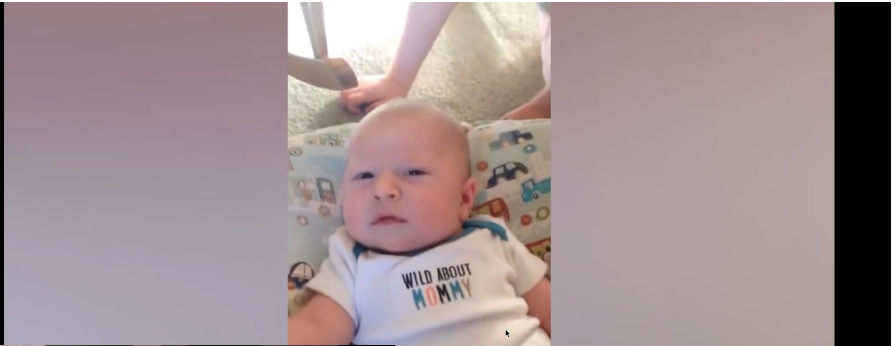


Fake

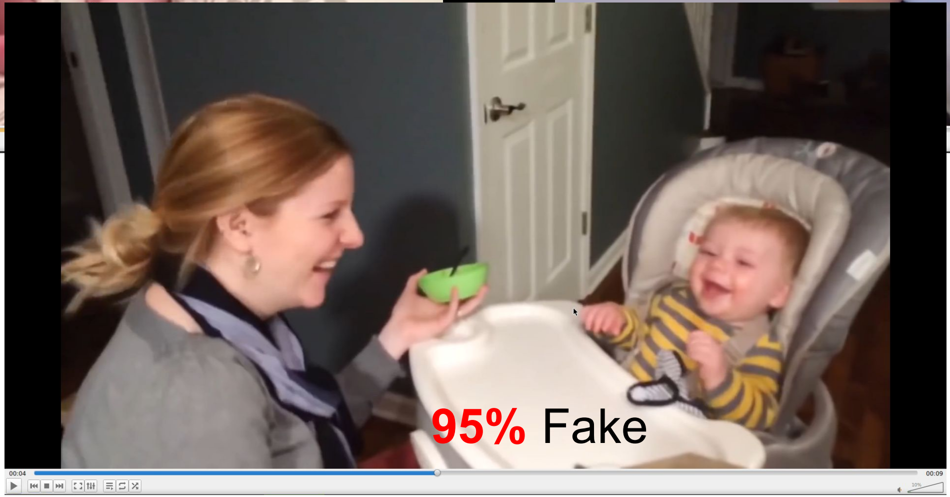
Babies



85% Fake



93% Fake



95% Fake

Results-at-a-Glance and Take Aways

Quantitative Results

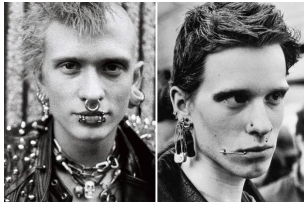
- ○ Babies (Rate: 14/20 babies classified fake)
- ○ Burn victims (Rate: 2/5 people with burns classified fake)
- ○ Leprosy/vitiligo victims (Rate: 1/5 and 2/5 people classified fake)
- ○ Face Tattoos (Rate: 2/10 rappers classified fake)

No Consistent Patterns

- Need detailed further to identify any systemic pattern currently
- The models are brittle to a variety of artifacts even on high-res videos, with no changes

Need to prevent a lot more varieties of o.o.d false positives...

- Genuine editing: Instagram filters
- Heavy makeup, piercings, etc. should not be classified ‘fake’



Using "BriePearls" Instagram filter
by Christian Venables (@csavenables)
wow!
filters

Thank You!