Subjective and objective evaluation of deepfake videos

Pavel Korshunov

Sébastien Marcel

Idiap Research Institute, Switzerland

April 21, 2021

Deepfakes detection: humans vs. machines

What are deepfakes?

- Some examples
- Datasets

Were you spoofed by deepfakes?

- Results and analysis of subjective test

How machines are doing?

- Evaluate on the same deepfakes
- Cross-db generalization problem

Deepfakes: the good, the bad and the ugly

The good 1

Salvador Dalí

https://www.youtube.com/watch?v=MZ2X-fSIPSU

The bad

Vladimir Putin

https://www.youtube.com/watch?v=sbFHhpYU15w

The ugly

Donald Trump

https://www.youtube.com/watch?v=8o0iOm-2sLw

These were funny but what about science

Databases of deepfakes

- DeepfakeTIMIT (Idiap): the first but the smallest
- FaceForensics++ (includes a set from Google)
- Celeb-DF: Youtube videos
- From Facebook: the largest to date (100′000 videos)

Is it fake or not?

Demo video 2 (With zoomed-in face)

What about this one?

Demo video 4 (With zoomed-in face)

Can humans spot deepfakes?

Can humans spot deepfakes?

Subjective evaluation

- Crowdsourcing scenario (uncontrolled environment)
- Pre-selected 120 videos from Facebook dataset
- 60 deepfakes in five categories of difficulty
- 57 subjects with about 20 subjects per video

Different deepfake categories

(f) Very easy (g) Easy (h) Moderate (i) Difficult (j) Very difficult

Different deepfake categories

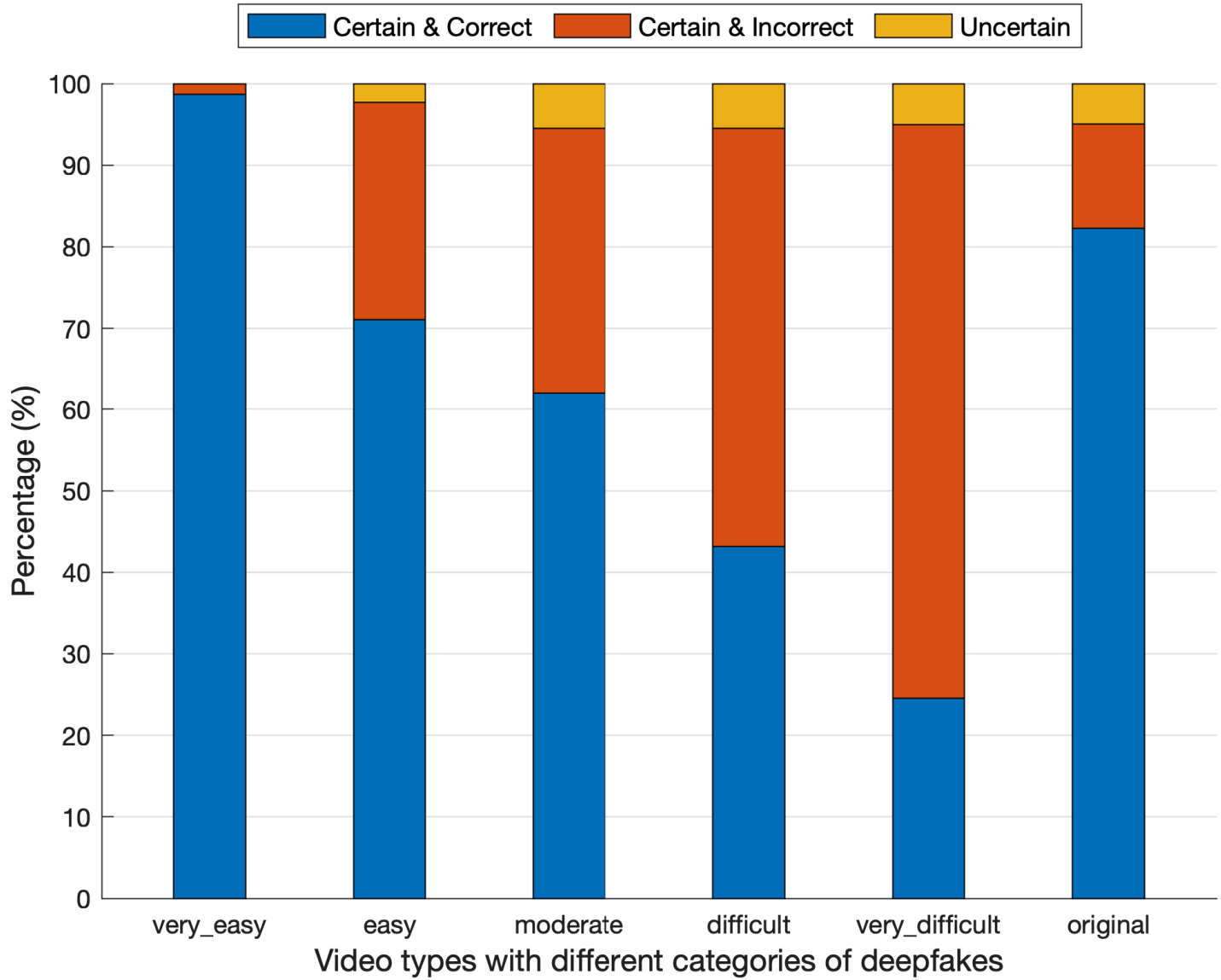
(f) Very easy (g) Easy (h) Moderate (i) Difficult (j) Very difficult

Subjective evaluation

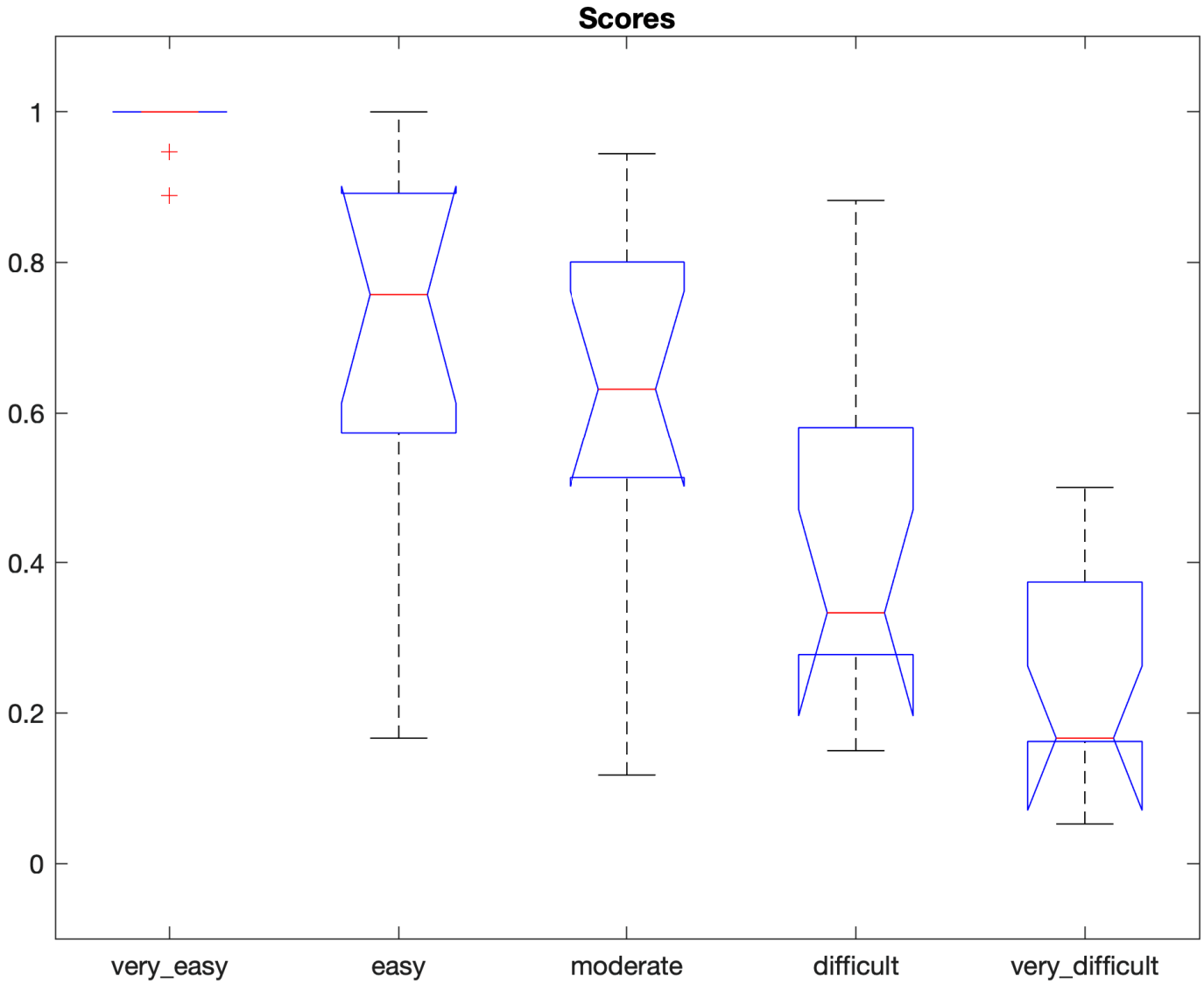
Nuts and bolts

- Test took 16 minutes on average
- 25 seconds per video (longer for originals)
- First two videos were dummies to remove bias
- Two people deemed unreliable (wrong honeypot question)
- A few people took long breaks (kept their scores)
- The score is the percentage of correct answers for the video
- ANOVA test: deepfake categories are significantly different

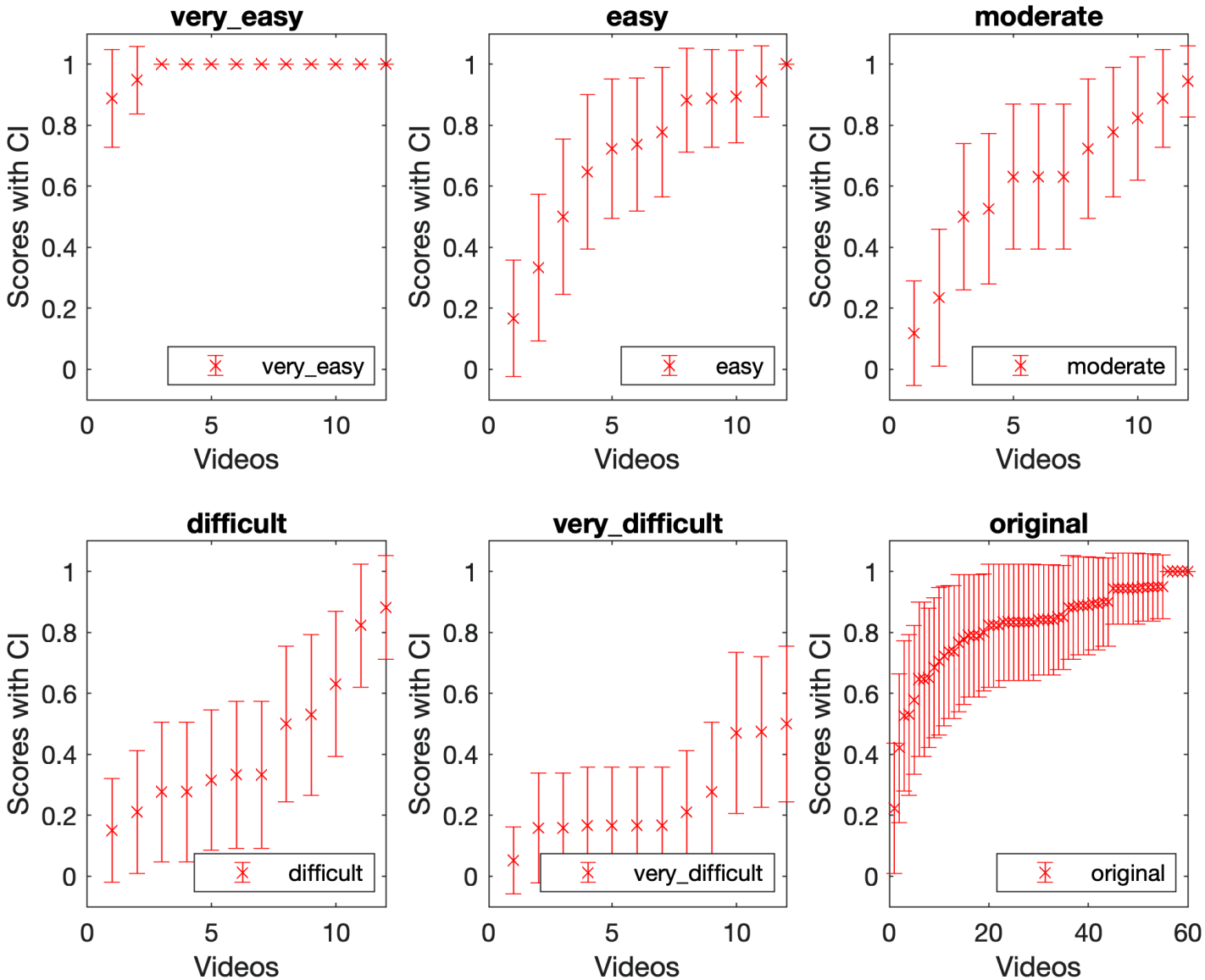
Main subjective results



Median scores with confidence intervals



For each video and category



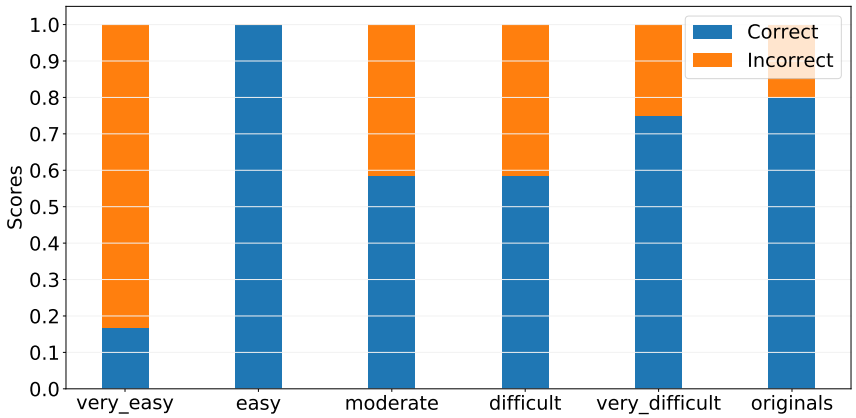
Can machines spot deepfakes?

Can machines spot deepfakes?

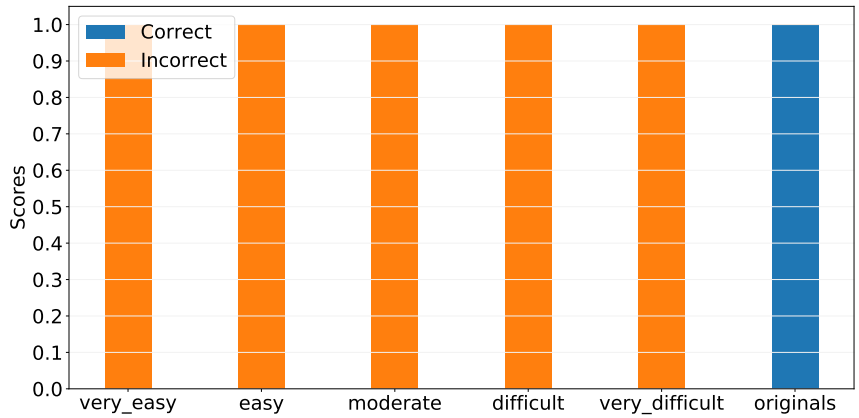
Objective evaluation

- Xception and EfficientNet-B4 networks
- Pre-trained on Google and Celeb-DF databases
- Area under the curve (AUC) is 100% on the same databases
- Same 120 videos as in subjective evaluation
- Threshold at false accept rate (FAR) of 10% on Dev set

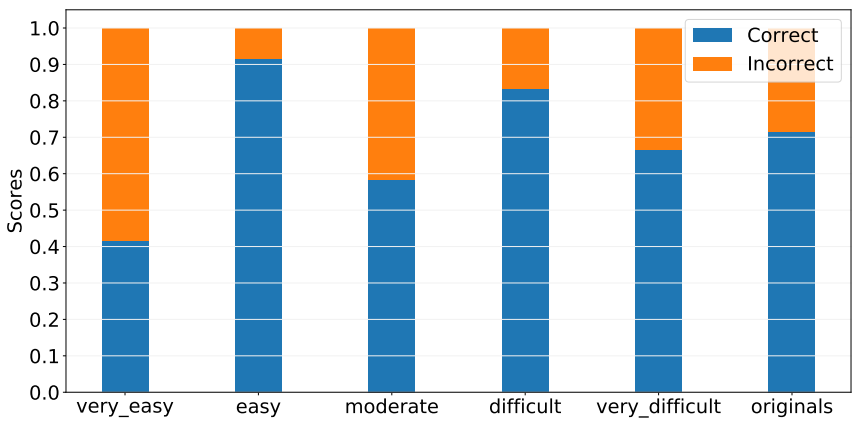
Main objective results (FAR is 10%)



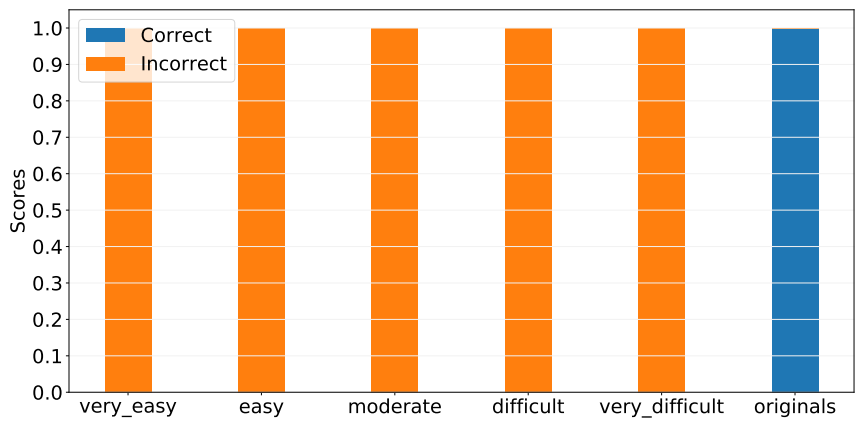
(k) EfficientNet trained on Google



(l) EfficientNet trained on Celeb-DF

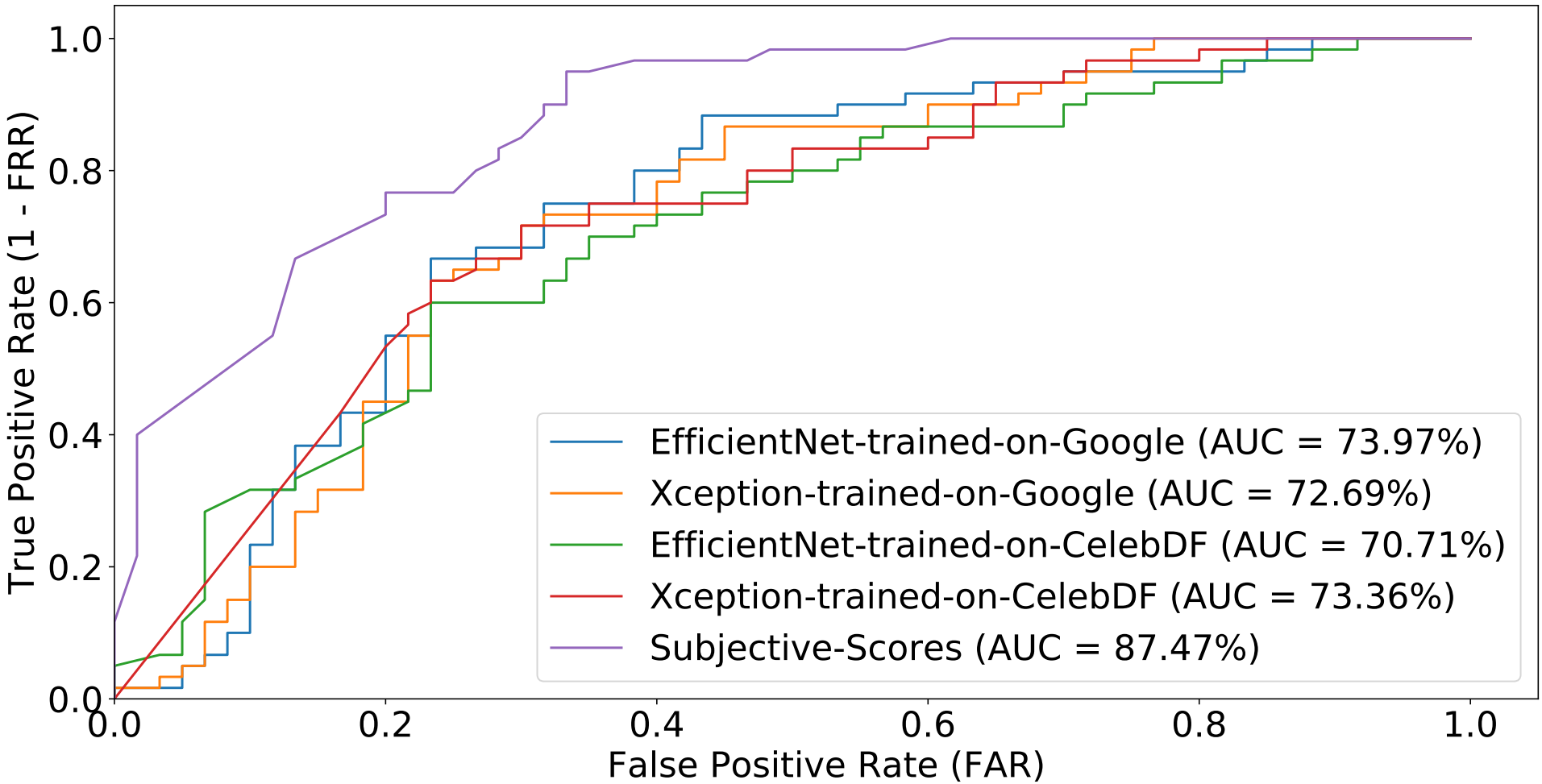


(m) Xception trained on Google



(n) Xception trained on Celeb-DF

ROC curves for humans and algorithms



Conclusions

Can deepfakes spoof human vision?

- YES

Can deepfakes spoof machine vision?

- YES

Both are bad for different reasons

- Do not anthropomorphize machine vision