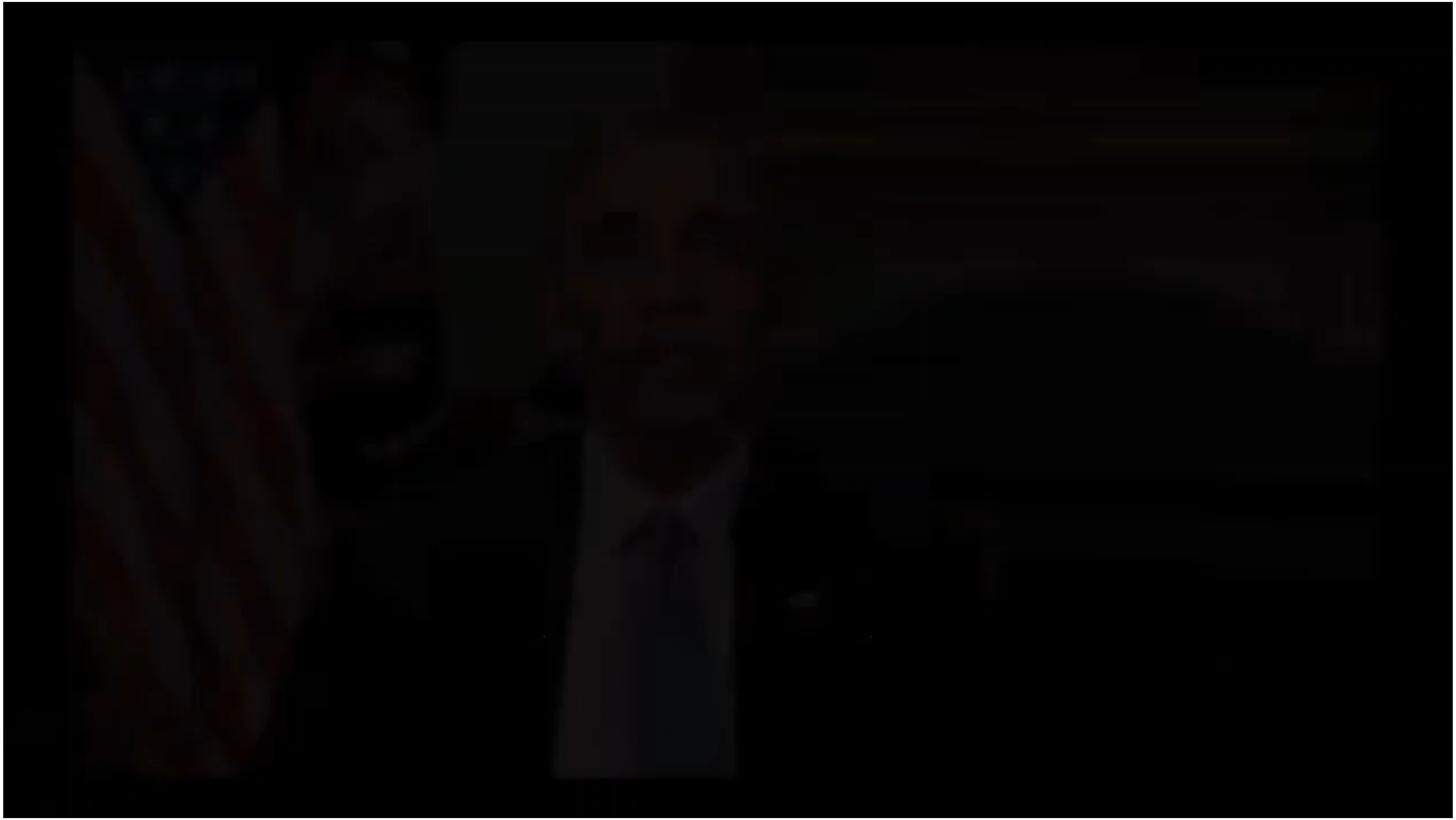
# What’s wrong with this video?
# Comparing Explainers for Deepfake Detection

Samuele Pino
samuele.pino@mail.polimi.it
CSE Track

POLITECNICO MILANO 1863

HONOURS PROGRAMME HP-SR in Information Technology



https://www.youtube.com/watch?v=cQ54GDm1eL0

# Goals

Automatic classifiers can already detect if a video is real or fake.

But can we understand the reason why a video is detected as fake?

Video

Manipulation detector

REAL

FAKE

**WHY?**

In this work we develop, extend and compare explanation techniques for deepfake detection.

# Overview

# Overview

- Introduction

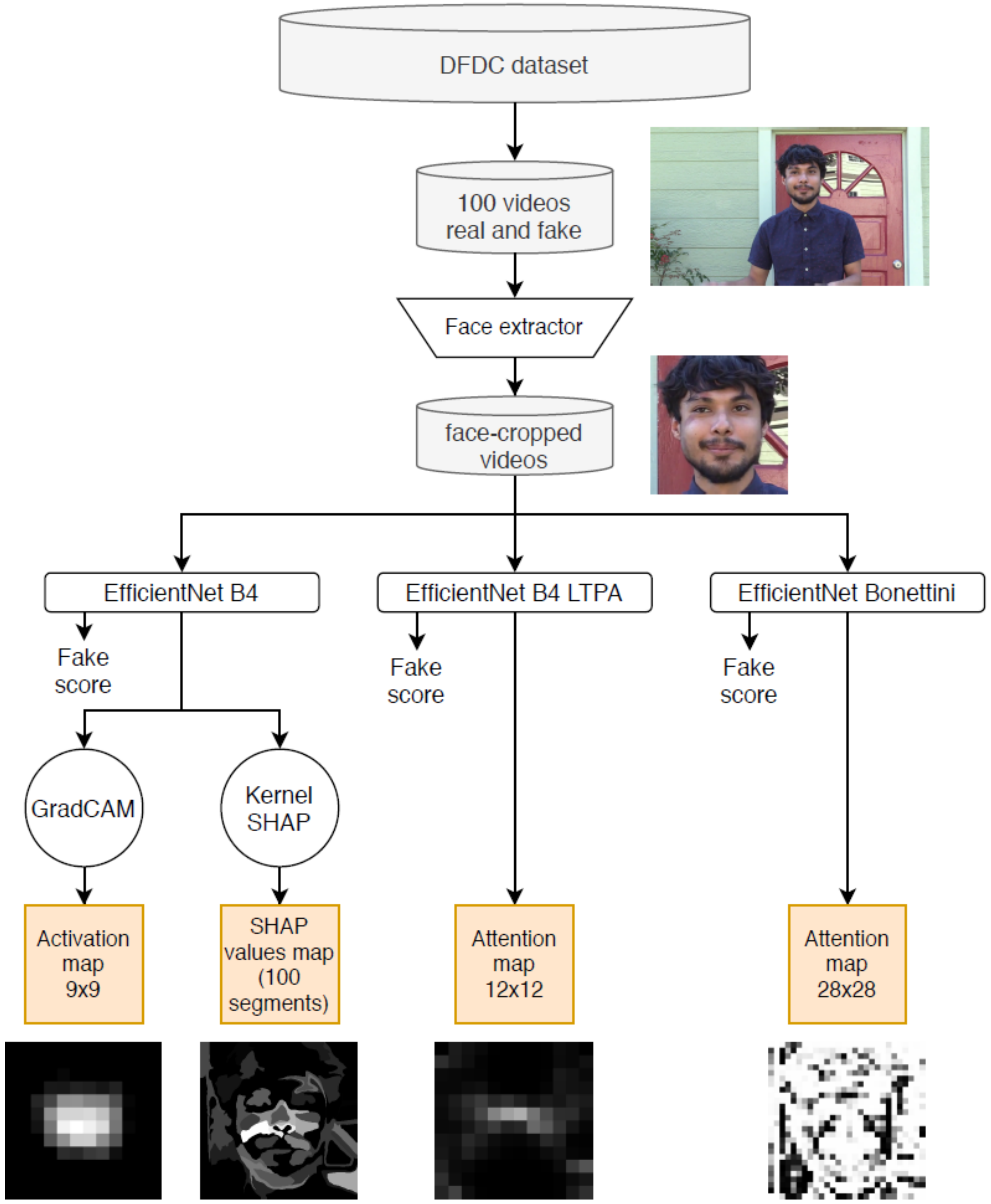
**Original** showing Alison Brie

**Deepfake** showing Jim Carrey instead of Brie

[Jan Kietzmann, Linda W. Lee, Ian P. McCarthy, Tim C. Kietzmann, Deepfakes: Trick or treat?, 2020]

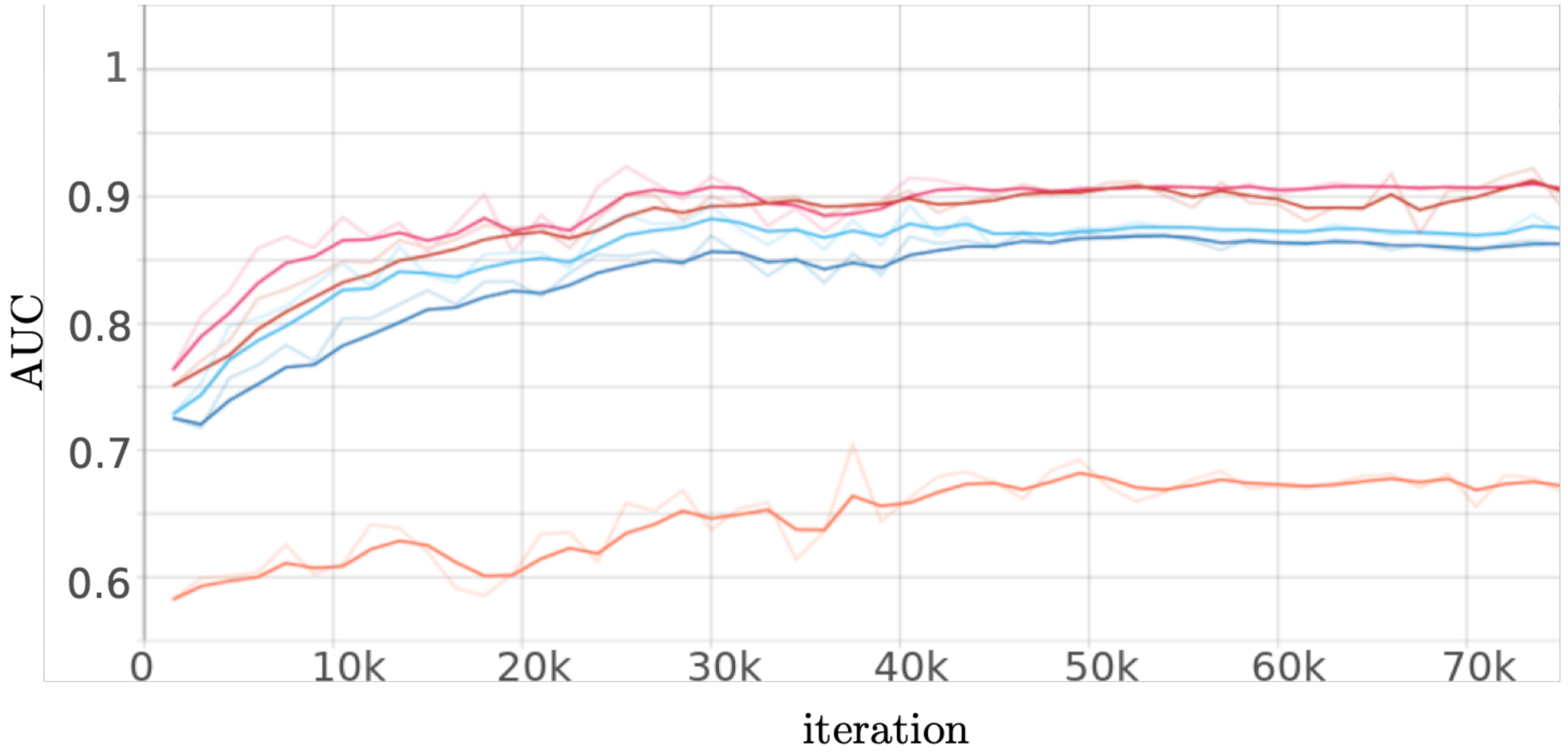
# Overview

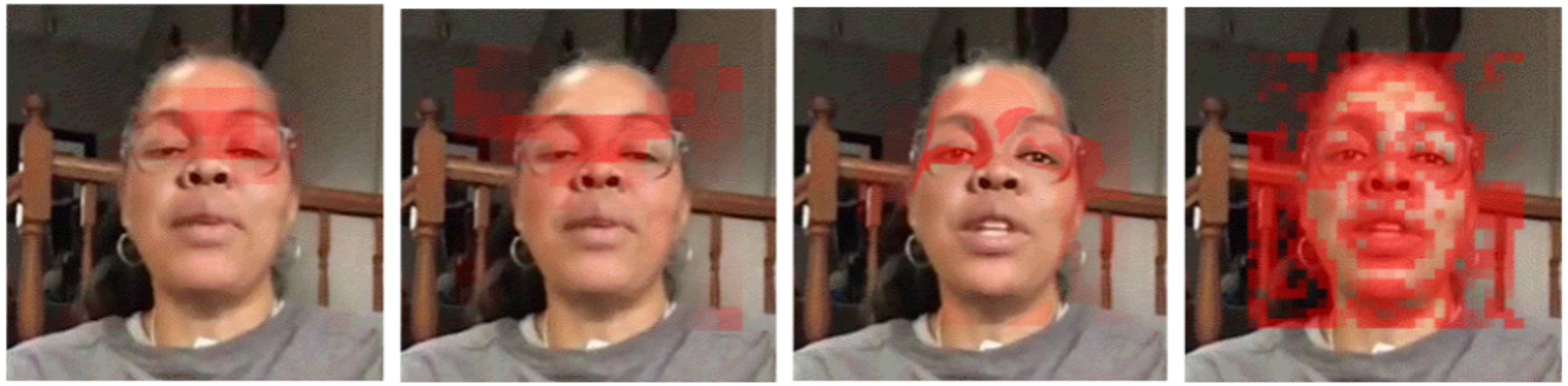
- Introduction
- Approach



# Overview

- Introduction
- Approach
- Experiments

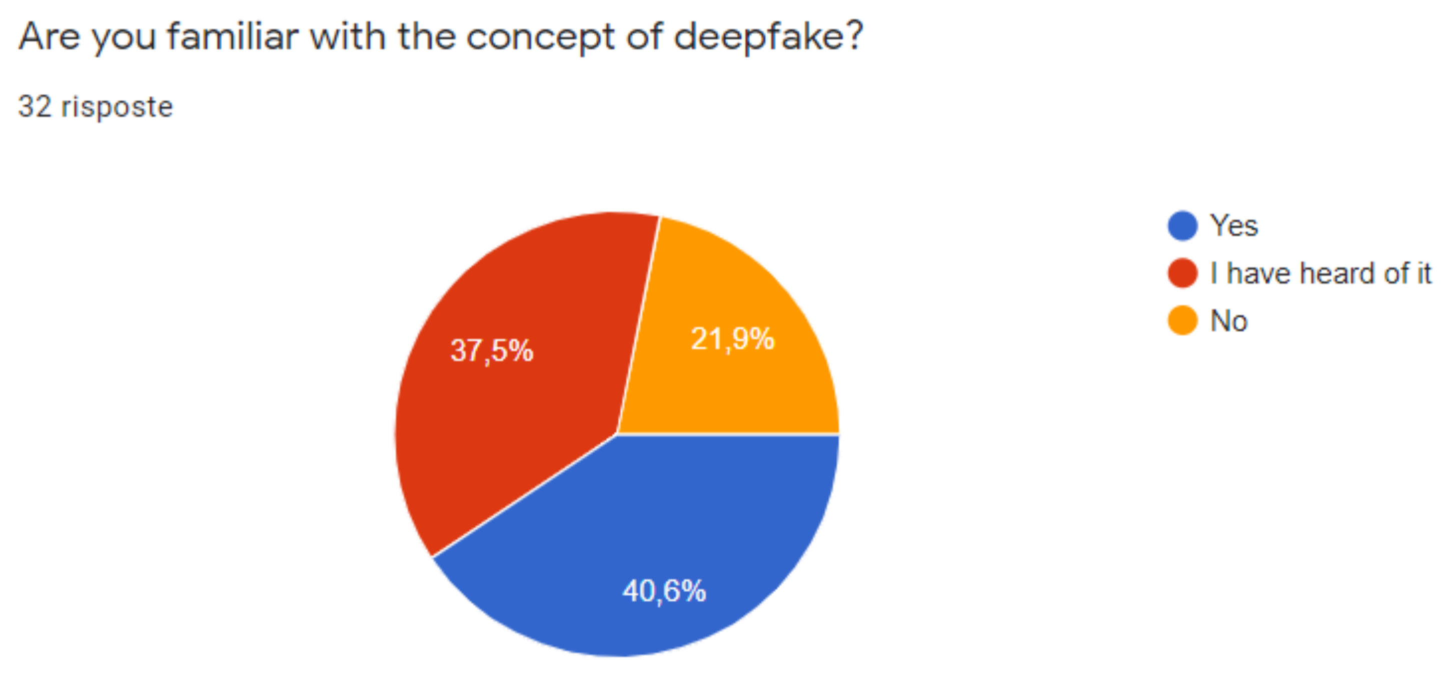


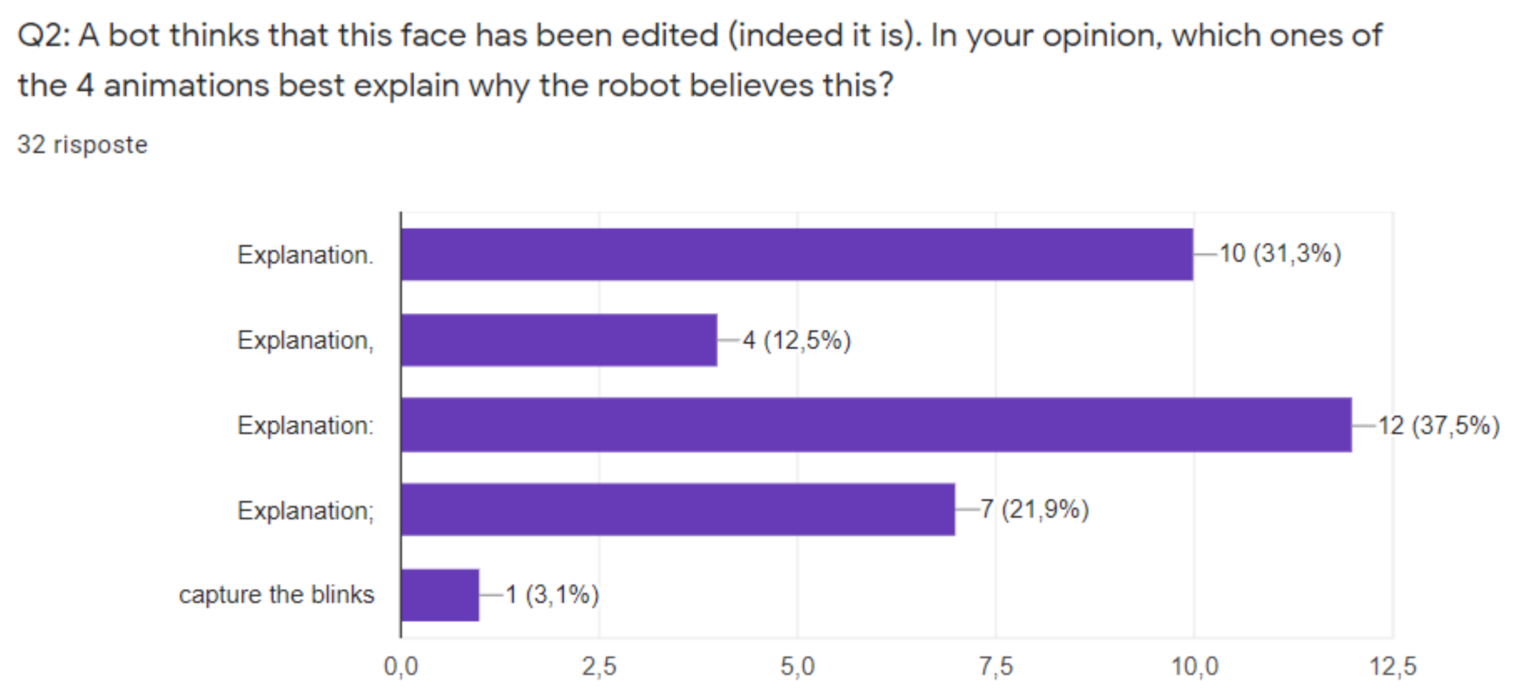



(a) GradCAM (b) LTPA lv. 2 (c) SHAP (d) Bonettini

# Overview

- Introduction
- Approach
- Experiments
- Results




| | $I$ [0, 1] | $V$ [0, 1] | $\tau$ [-1, 1] | $\rho$ [-1, 1] | $\mu$ [0, 1] |
|---|---|---|---|---|---|
| Bonettini | 0.4821 | 0.0951 | 0.7390 | 0.1262 | 0.5286 |
| GradCAM | 0.0689 | 0.0135 | **0.8756** | **0.7489** | **0.8666** |
| LTPA | 0.0616 | 0.0108 | 0.7991 | 0.3333 | 0.6386 |
| SHAP | 0.0563 | 0.0302 | 0.4496 | 0.2326 | 0.7348 |

# Introduction

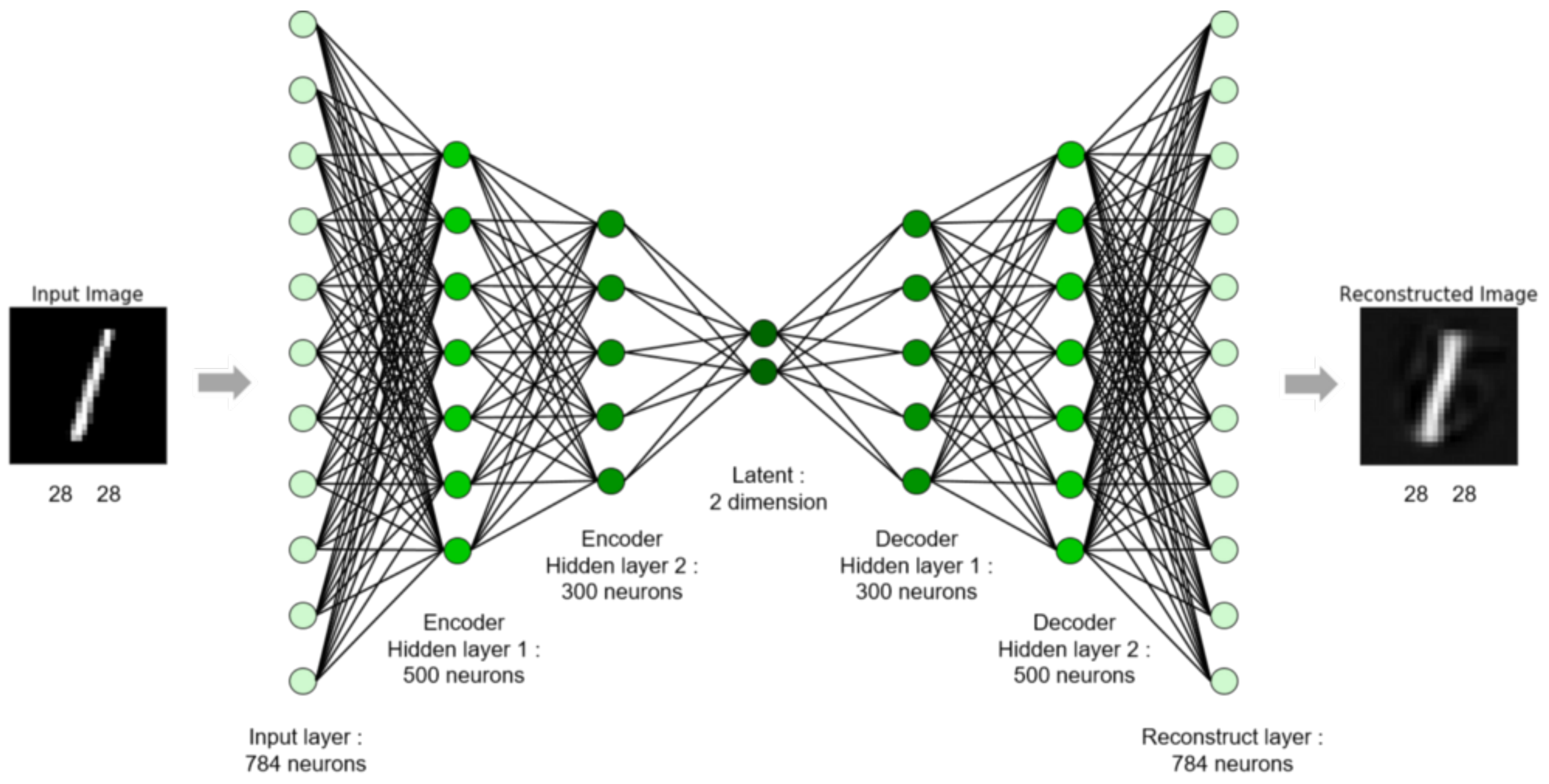
# Deepfakes

- Replacing faces in videos


[Deepfake Detection Challenge, 2019]

# Deepfakes

- Replacing faces in videos
- Deep learning technique



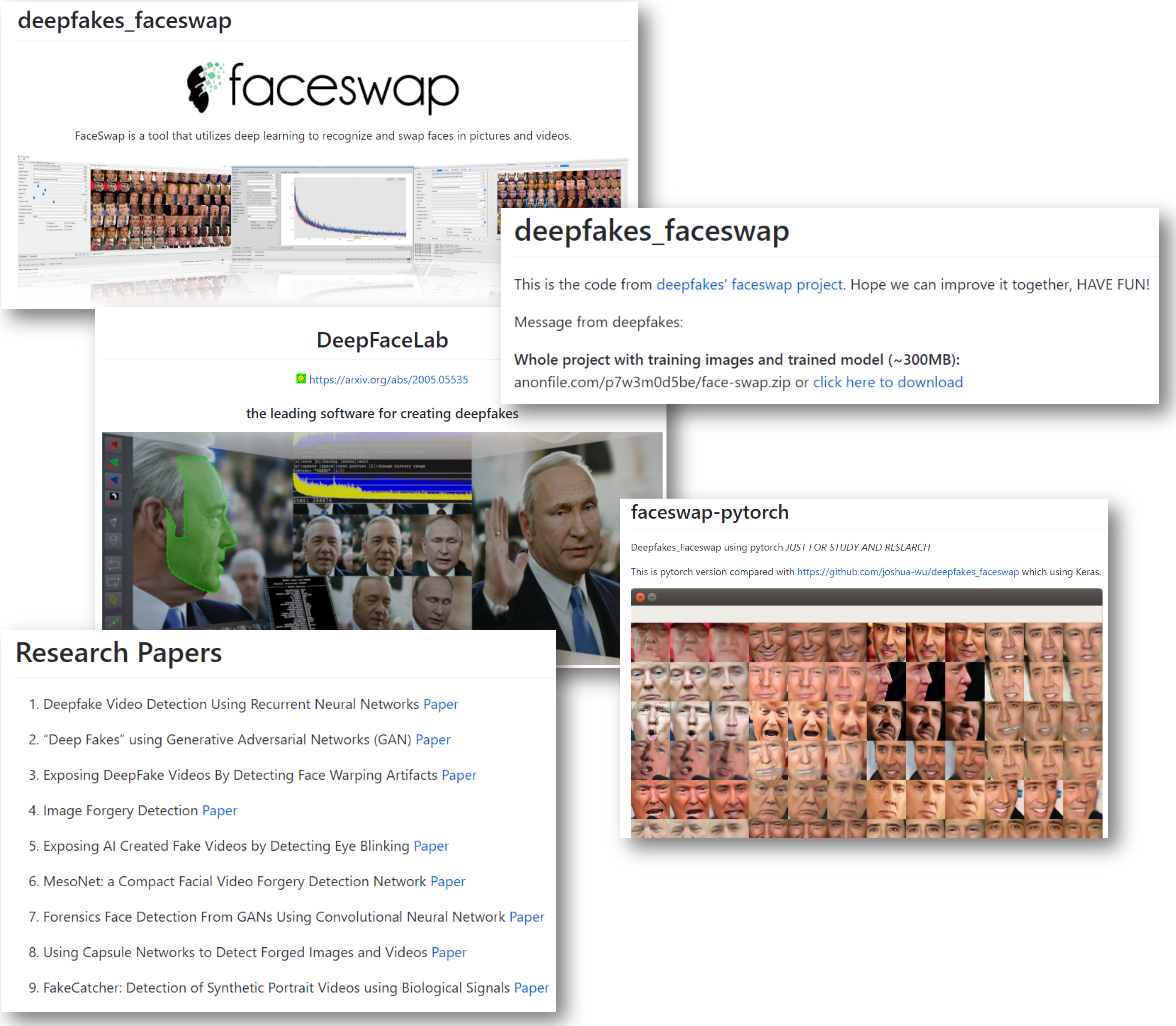

https://mc.ai/auto-encoder-in-biology/

# Deepfakes

- Replacing faces in videos
- Deep learning technique
- Initially to generate adult contents

# Deepfakes

- Replacing faces in videos
- Deep learning technique
- Initially to generate adult contents
- No official implementation



# Deepfakes

Why is it important to detect them?

# Deepfakes

Why is it important to detect them?

- disinformation


[https://leidenlawblog.nl/articles/wash-your-hands-often-and-your-newsfeed-even-more-disinformation-in-covid-19-times]

# Deepfakes

Why is it important to detect them?

- disinformation
- online abuse


[https://removify.com.au/blog/online-abuse-what-it-is-and-how-to-deal-with-it/]

# Deepfakes

Why is it important to detect them?

- disinformation
- online abuse
- financial fraud

[https://economictimes.indiatimes.com/wealth/plan/beware-of-financial-fraud-on-social-media/articleshow/51468572.cms?from=mdr]

# Deepfakes

Why is it important to detect them?

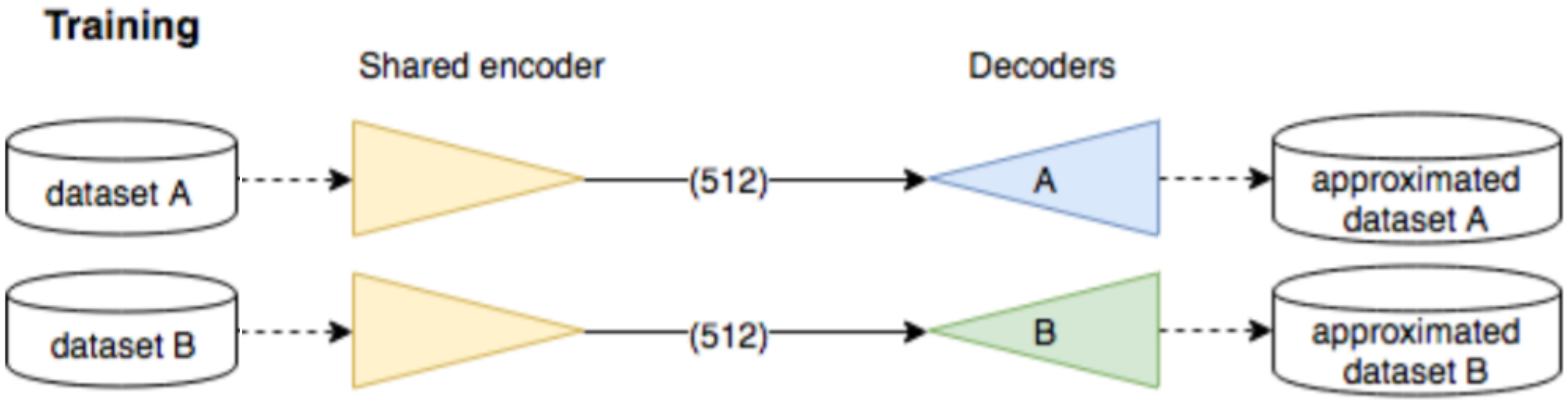
- disinformation
- online abuse
- financial fraud
- law enforcement

[http://www.forensicsciencesimplified.org/av/how.html]

# Deepfakes

Let's build our deepfake!

# Deepfakes

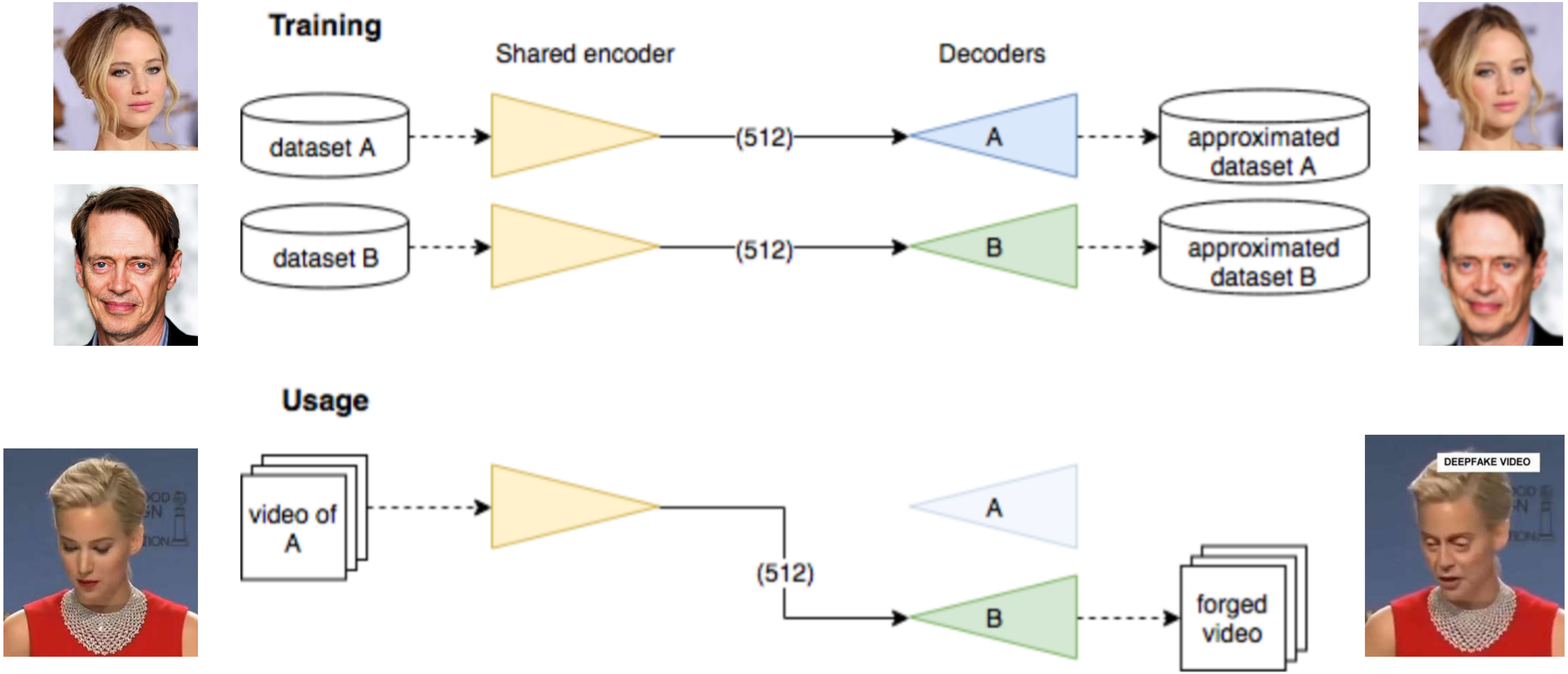



Jennifer Lawrence

Steve Buscemi

[Afchar et al., “MesoNet: a Compact Facial Video Forgery Detection Network”, 2018]

# Deepfakes




Jennifer Lawrence

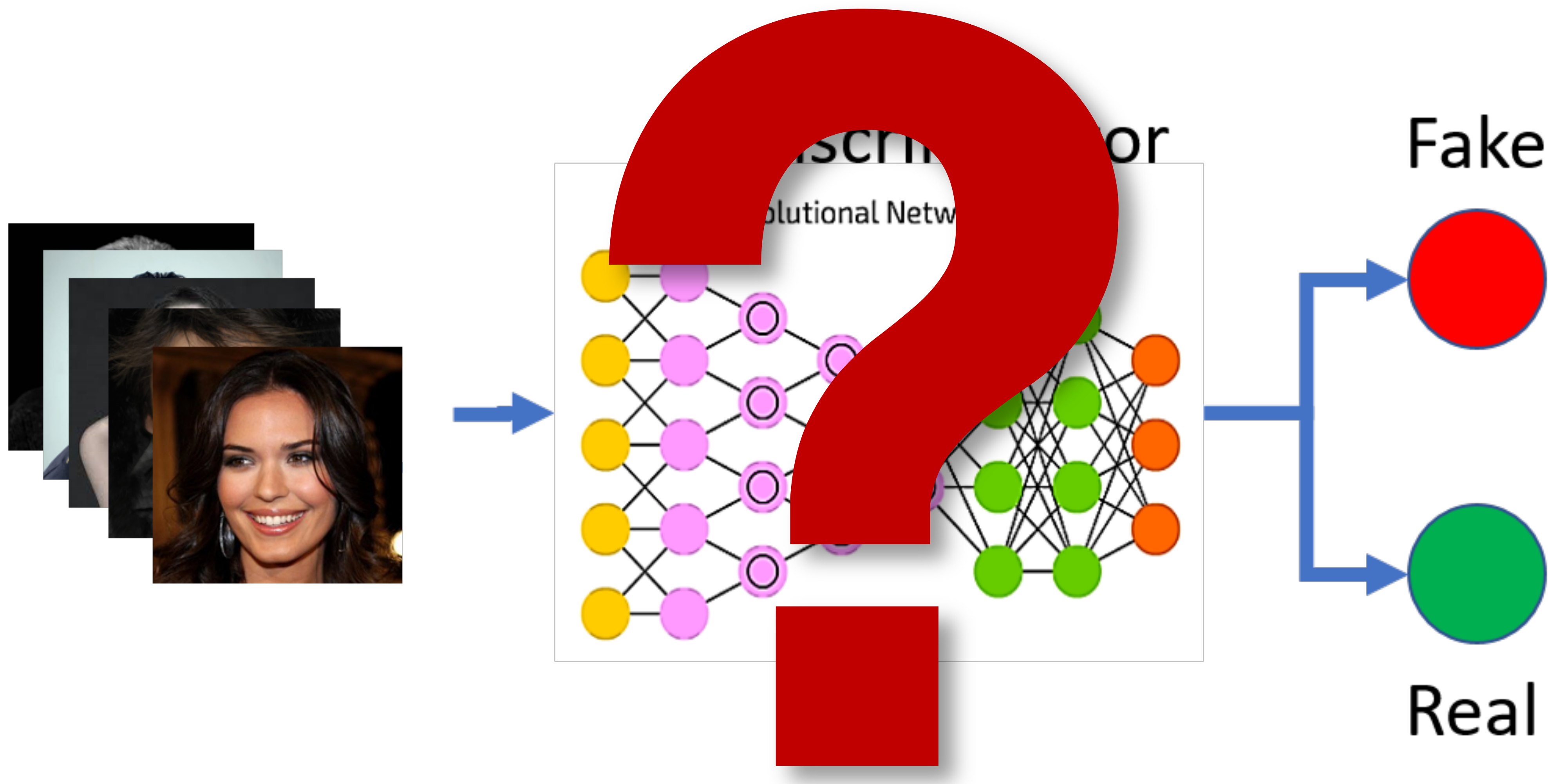
Steve Buscemi

Jennifer Buscemi (?)

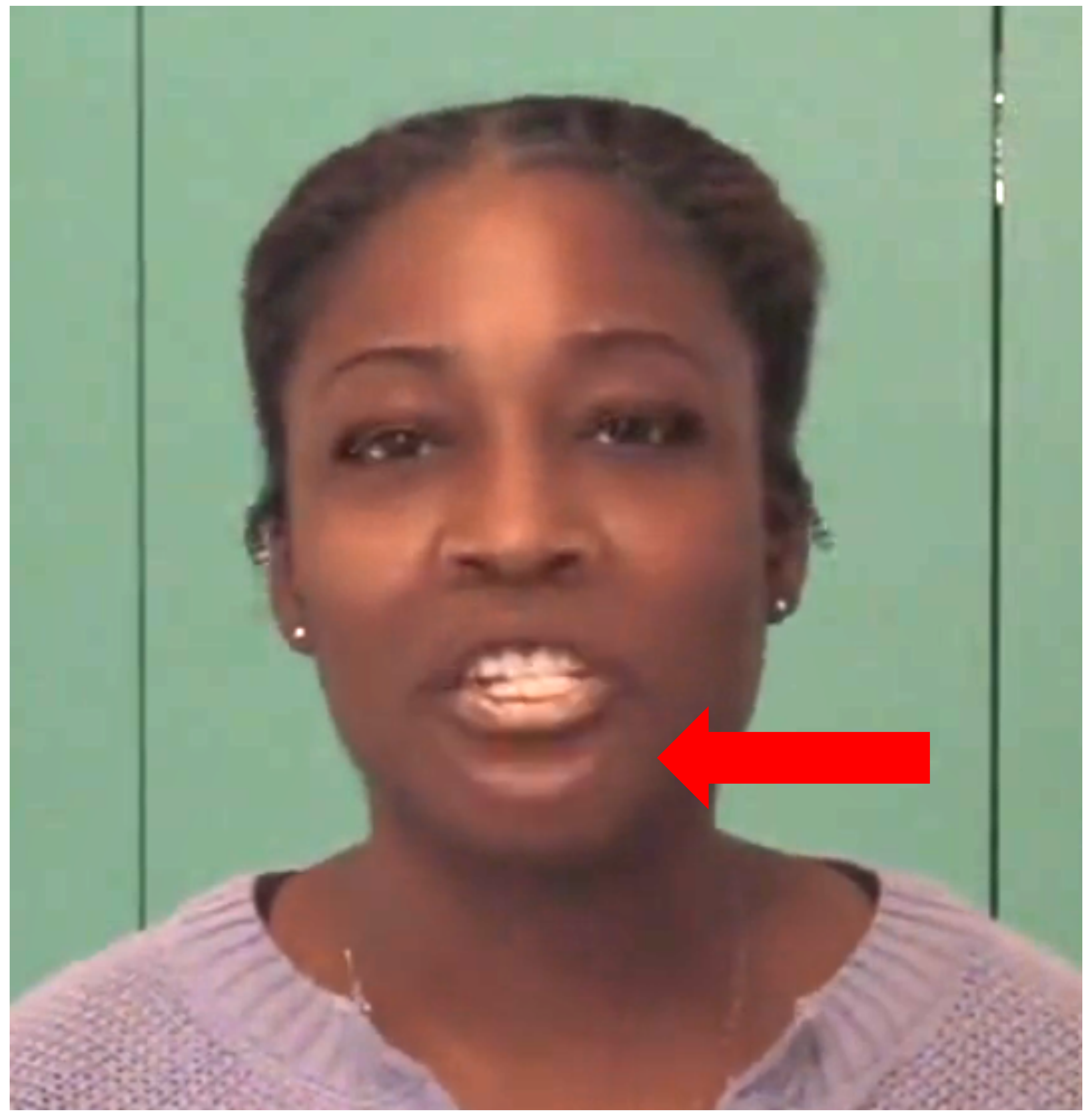
[Afchar et al., "MesoNet: a Compact Facial Video Forgery Detection Network", 2018]

# Explainability problem

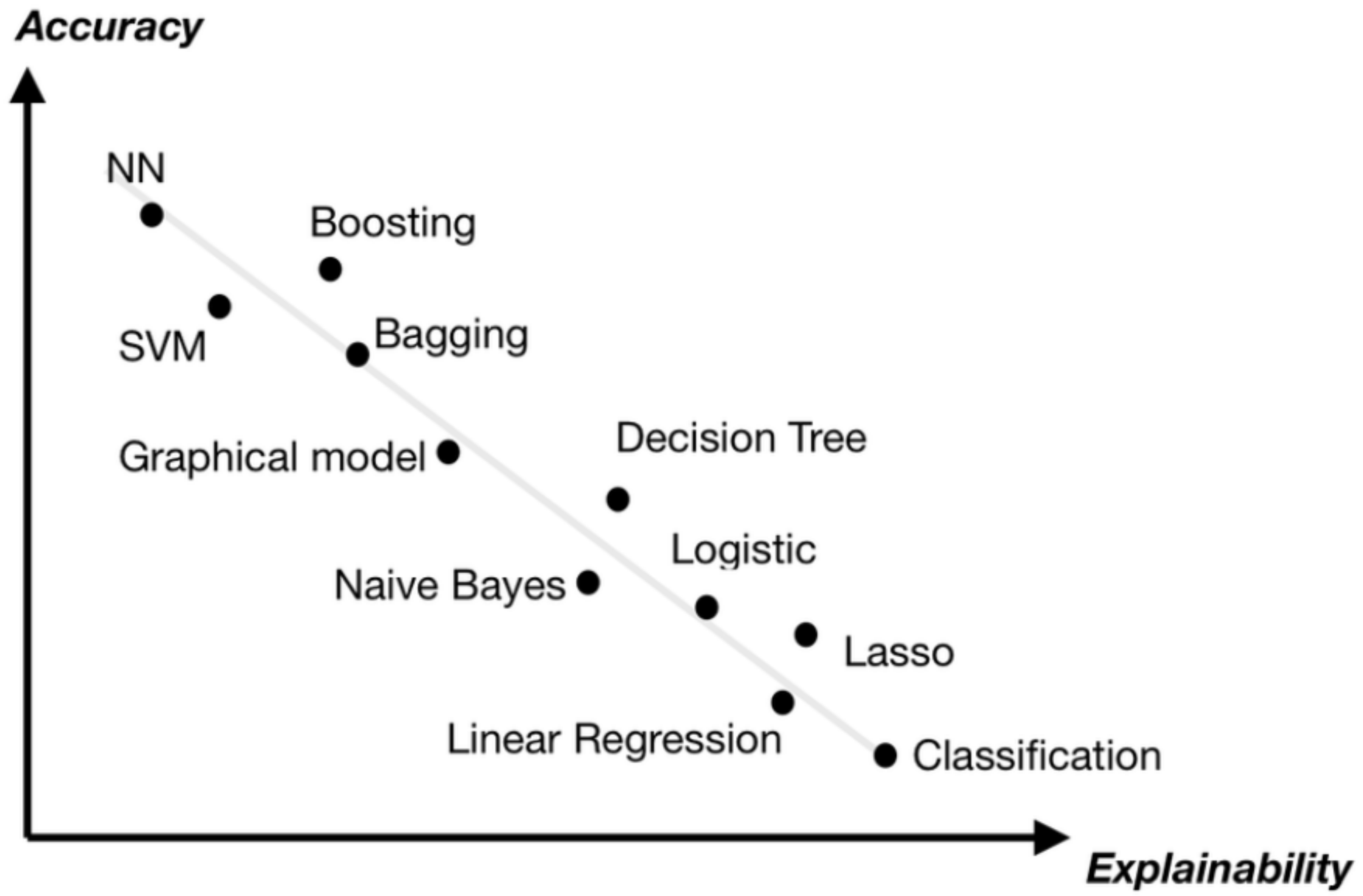
- Also detectors use deep learning
- Should we trust them?



# Explainability problem



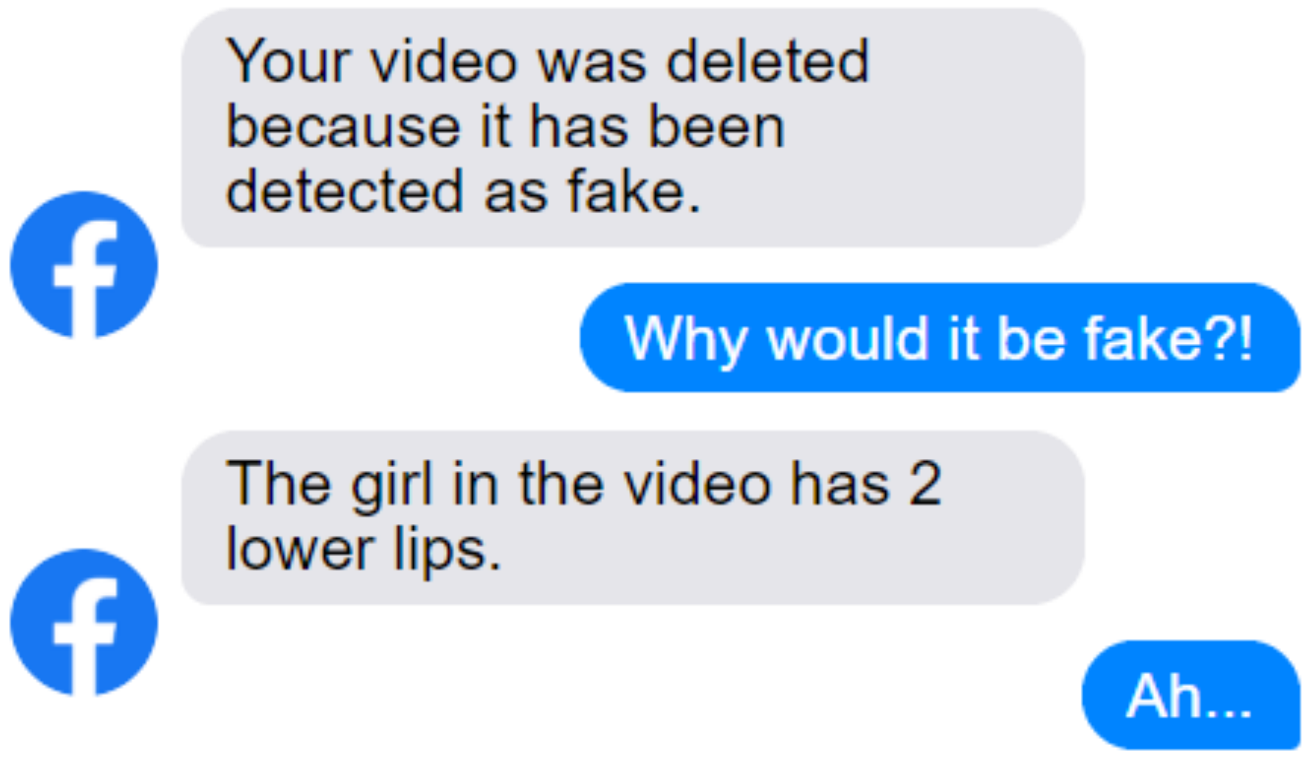
- «correct prediction for the correct reason»
- Complexity-interpretability trade off




[Alexandre Duval. Explainable articial intelligence (xai), 04 2019.]

# Explainability problem

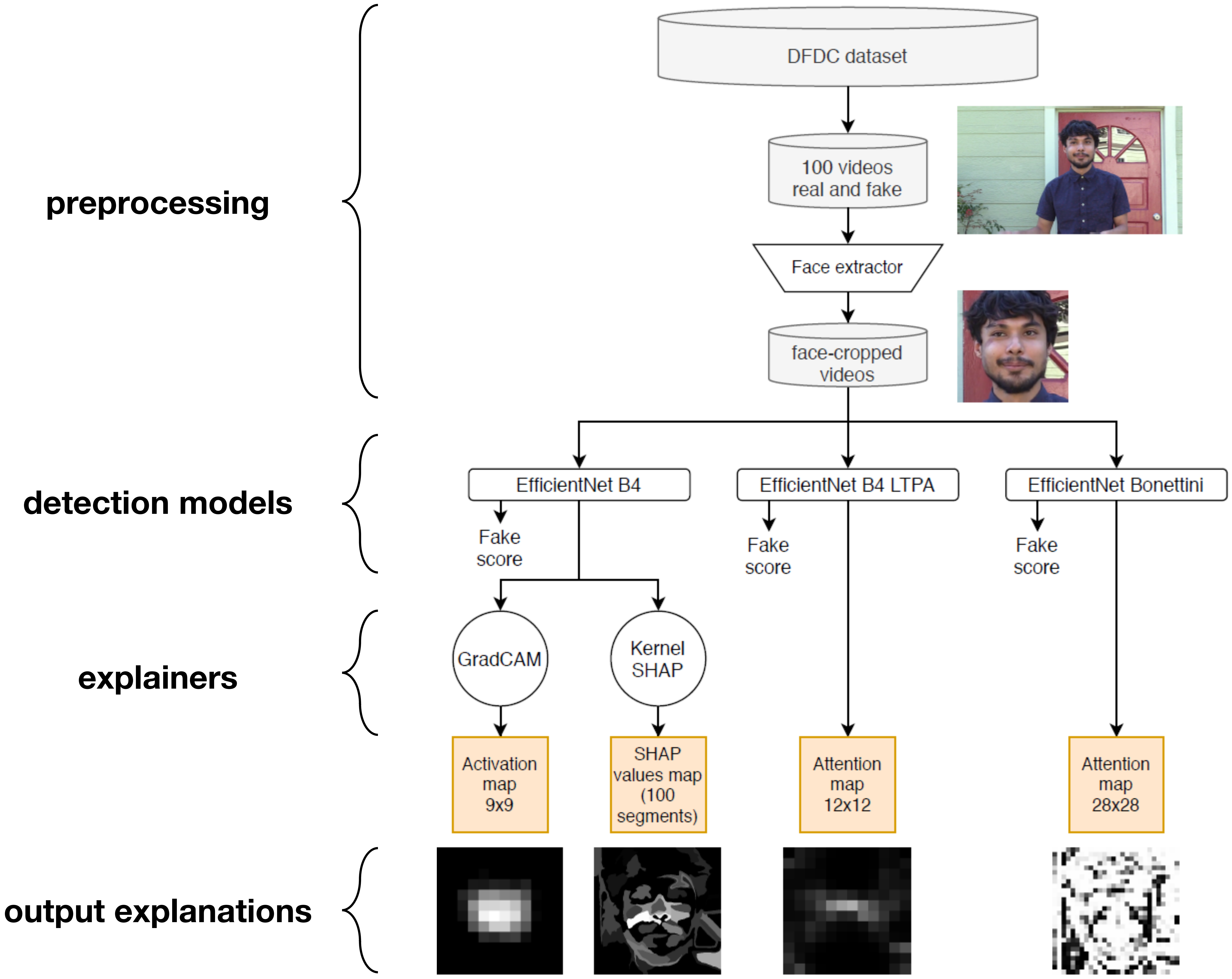
- Why do we need it:
  - law enforcement
  - journalists
  - dispute resolution in social media

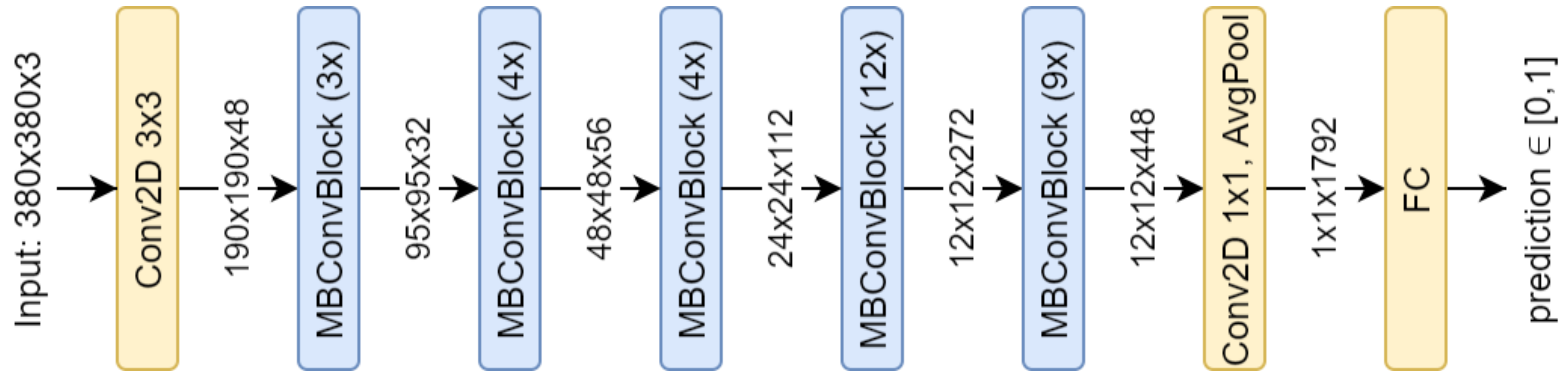

[https://www.shutterstock.com/it/image-illustration/gavel-isolated-on-white-background-3d-1437862448]
[https://www.aa.com.tr/en/asia-pacific/sri-lankan-journalists-fear-new-gov-t-silencing-dissent/1670591]

# Approach

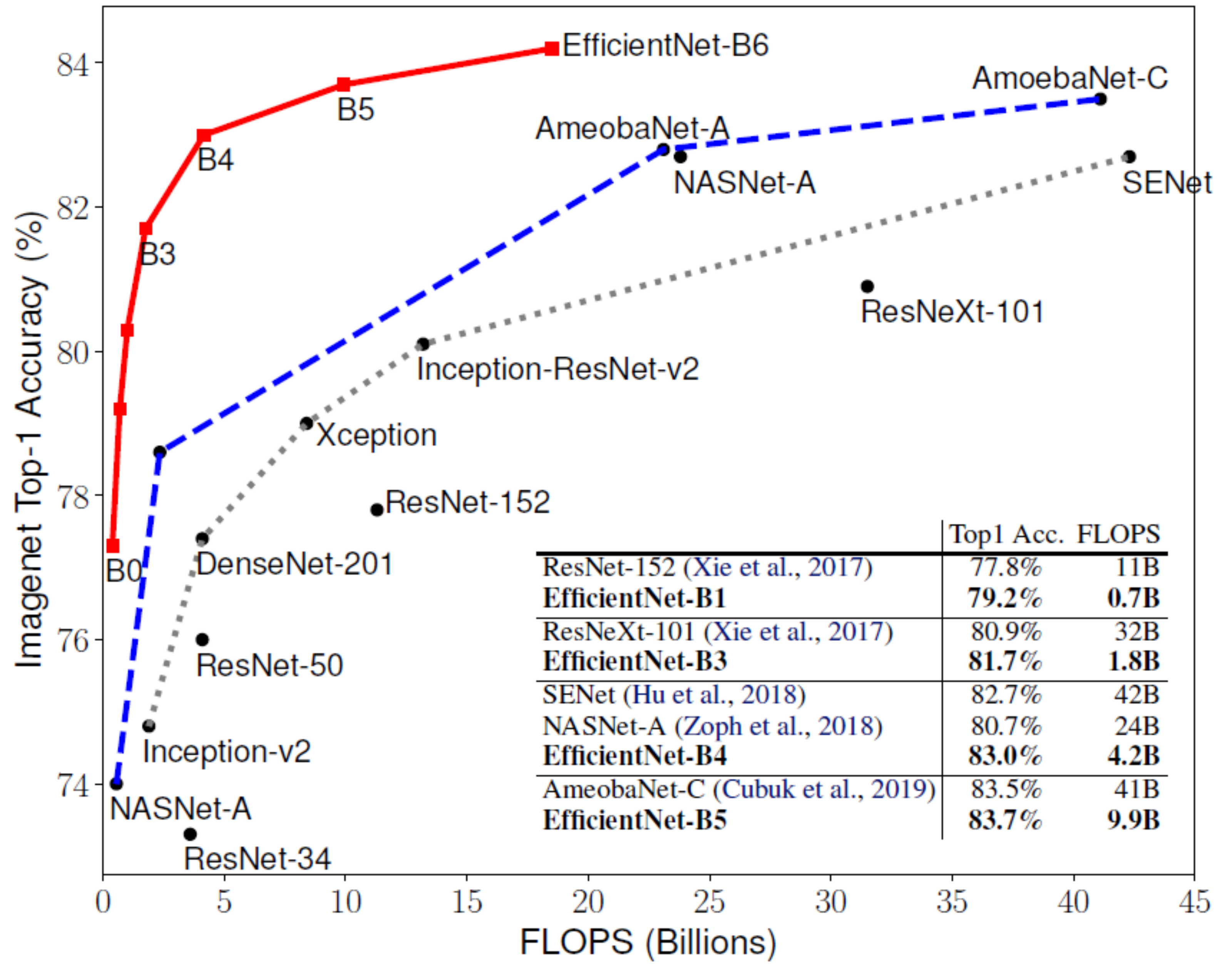
# Approach: overview



# Approach: detection model




- EfficientNet as a backbone CNN
- Powerful and lightweight
- Winner's solution in Deepfake Detection Challenge

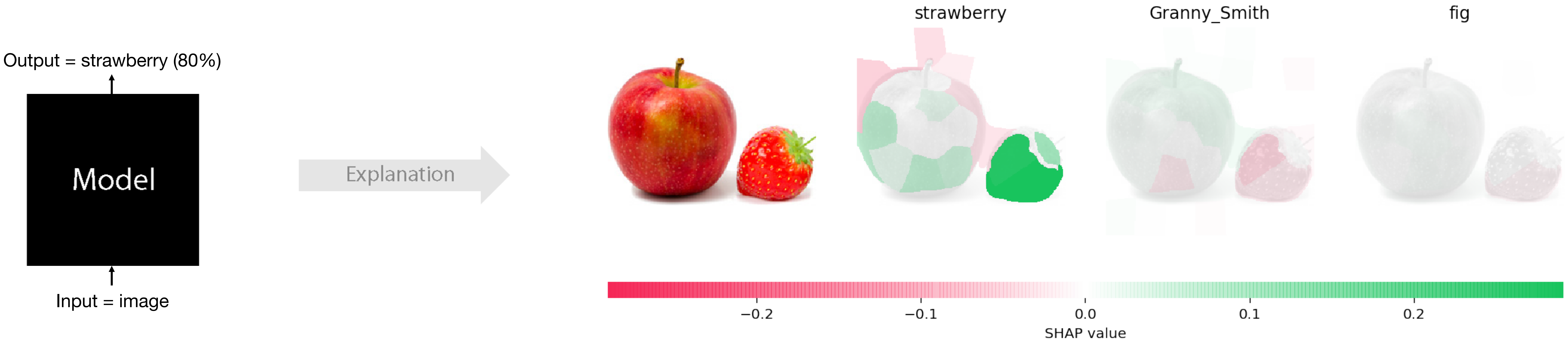



[Ecientnet: Rethinking model scaling for convolutional neural networks, Tan and Le, 2019]

# Approach: explainers

SHAP

- Black-box (SHAP)




[A unified approach to interpreting model predictions, Lundberg and Lee, 2017]

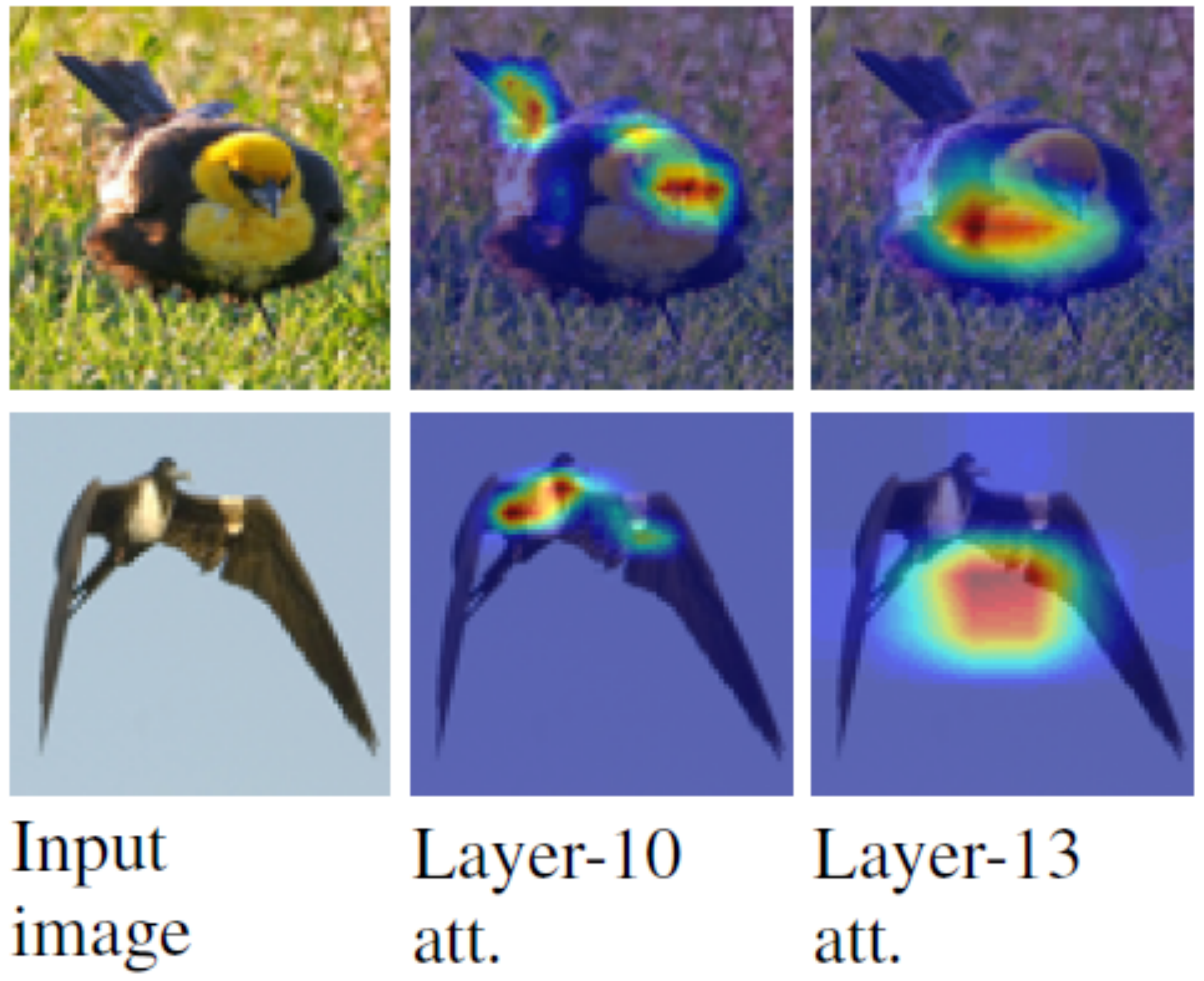
# Approach: explainers

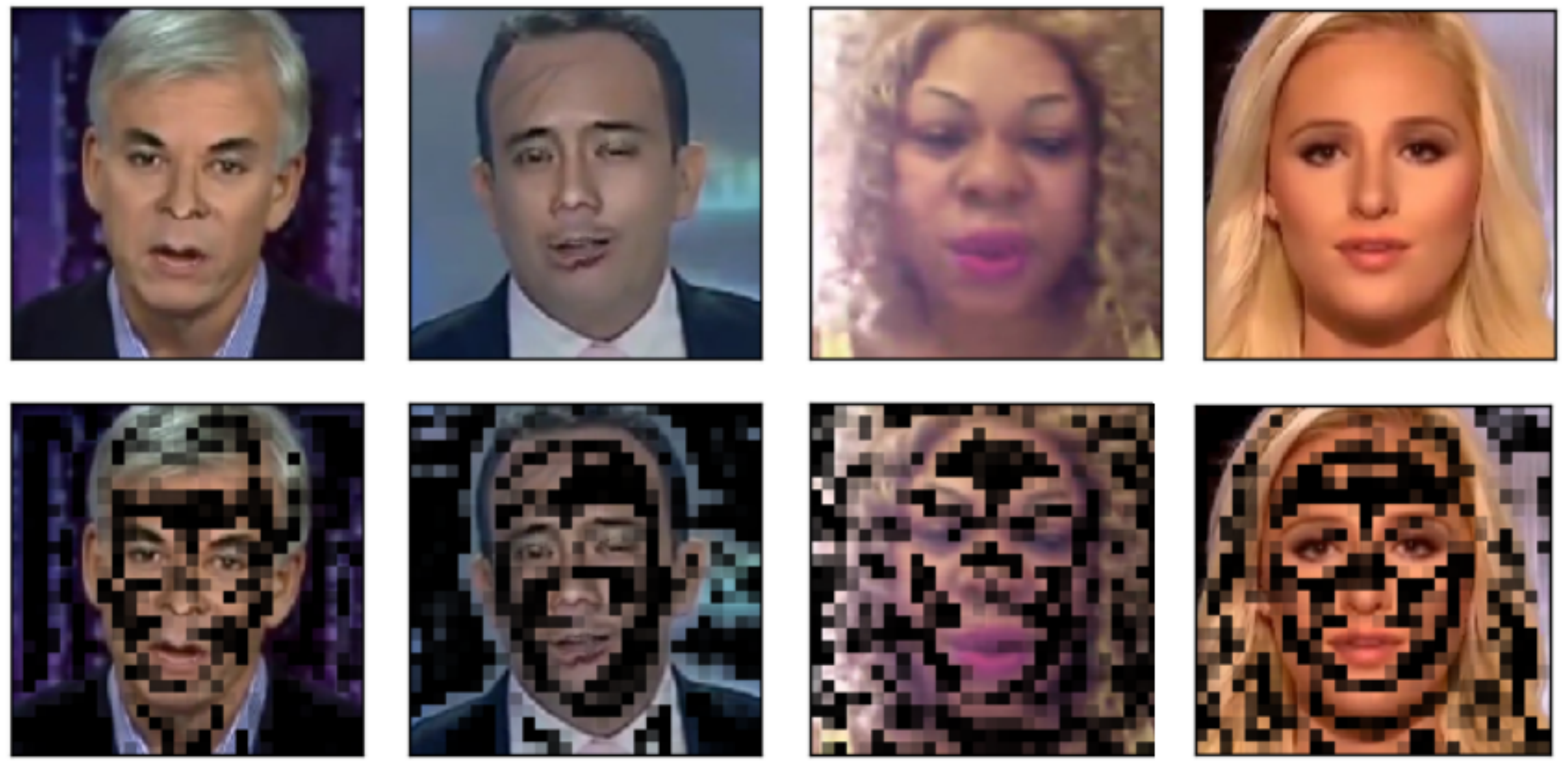
- Black-box (SHAP)
- White-box (GradCAM)


[Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, Ramprasaath et al., 2019]

# Approach: explainers

- Black-box (SHAP)
- White-box (GradCAM)
- Self-attention (LTPA, Bonettini)

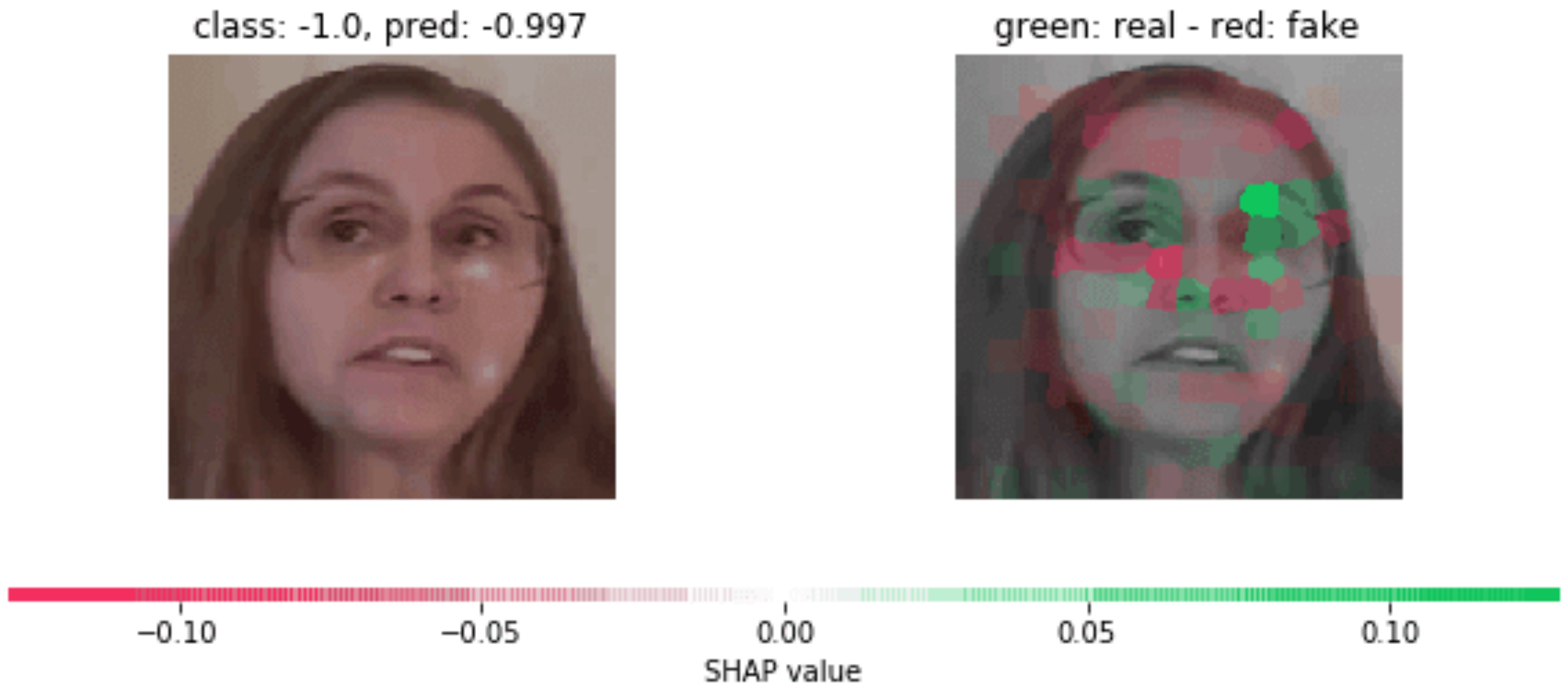


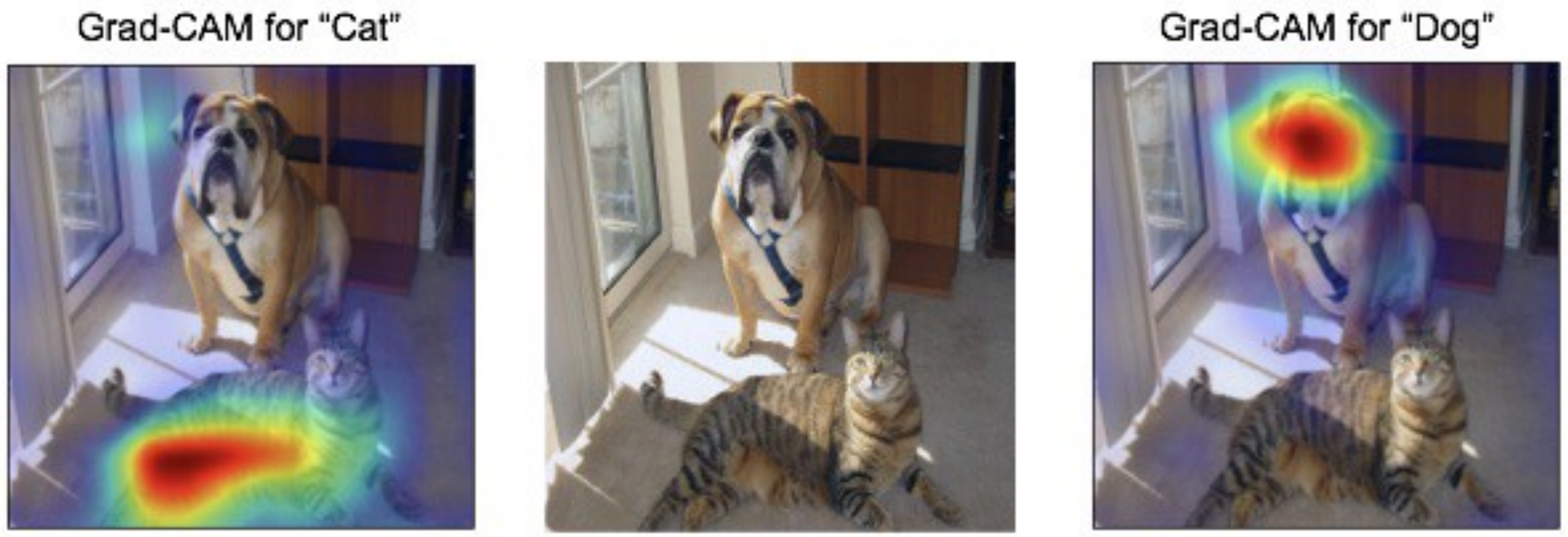
[Learn to pay attention, Jetley et al., 2018]
[Video face manipulation detection through ensemble of cnns, Bonettini et al., 2020]

# Approach: SHAP

- Model-agnostic
- Kernel SHAP for image classification
    - Segmentation
    - SHAP values assignment
- Extension: 3D segmentation for video classification

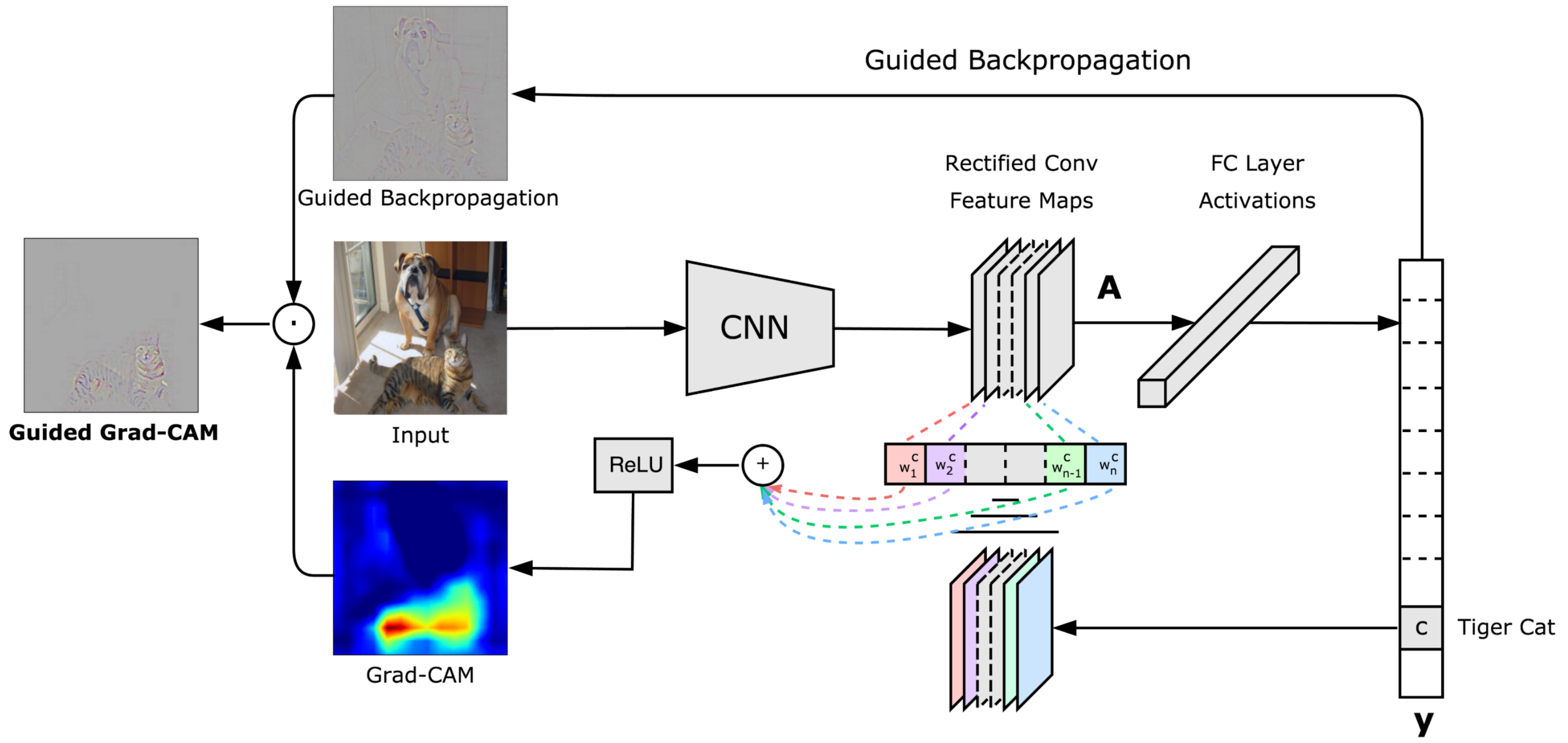


# Approach: GradCAM



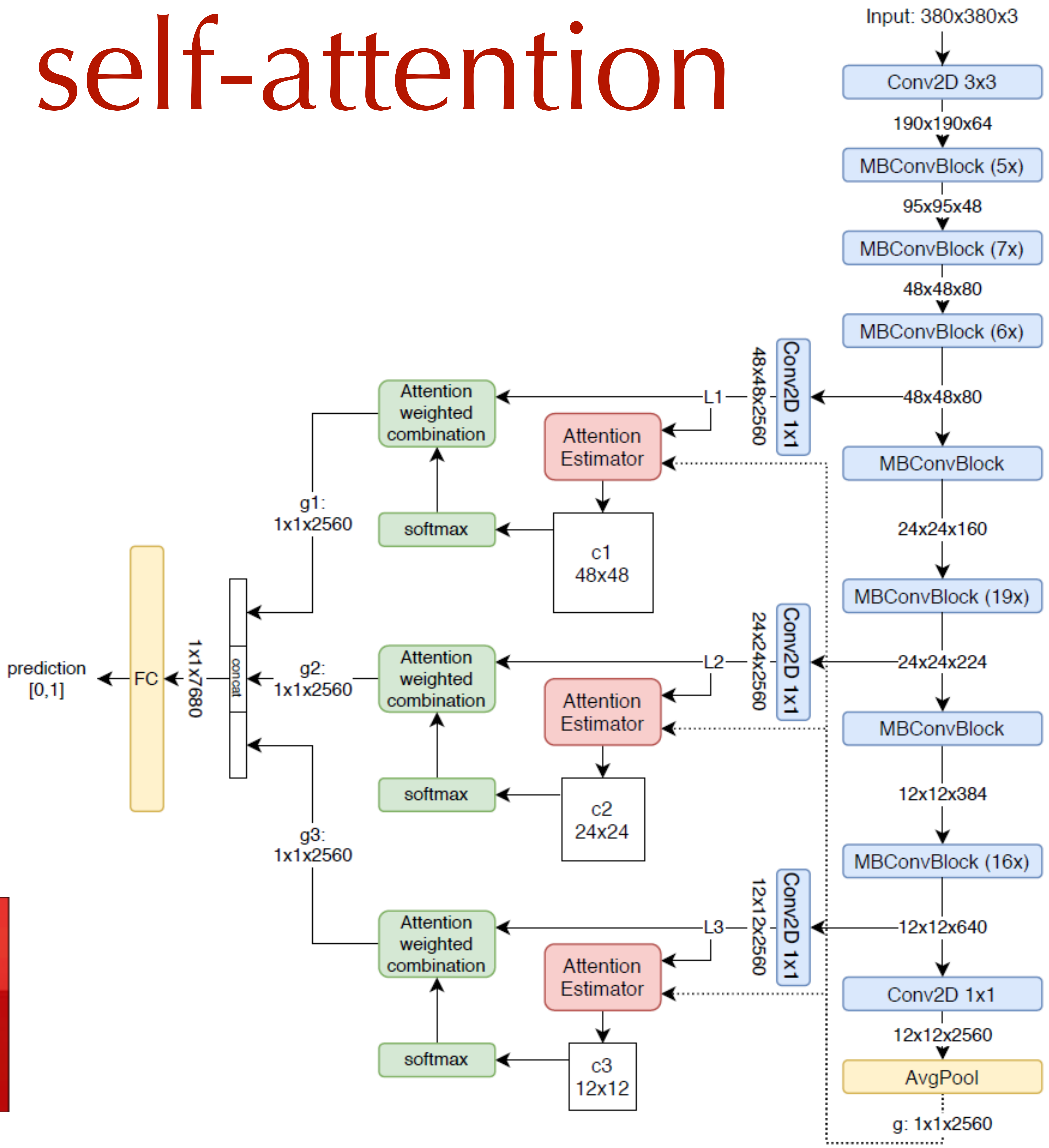

- Class Activation Mapping
- Neural network gradients
- Binary classification extension

Grad-CAM for 'fake'    Grad-CAM for 'real'



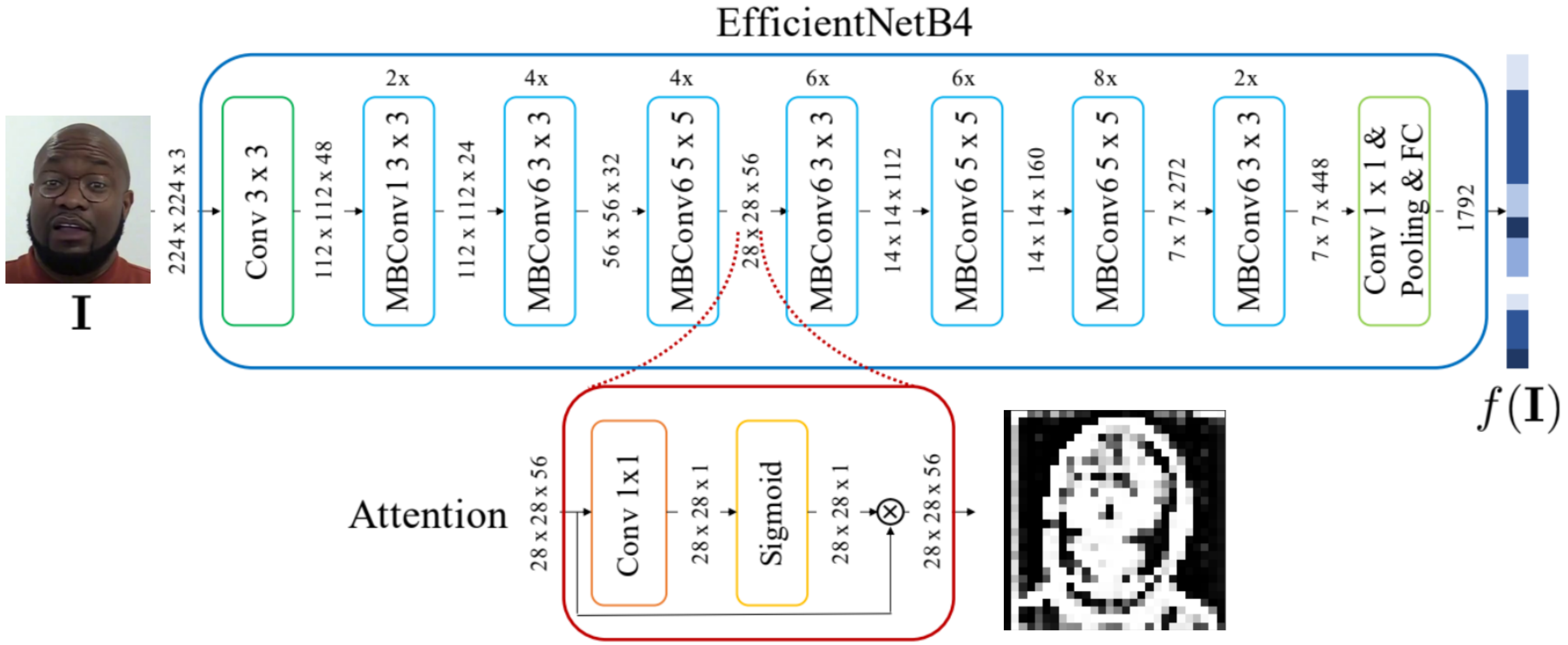
# Approach: self-attention

- Learn To Pay Attention (LTPA)
  - 3 attention maps

# Approach: self-attention

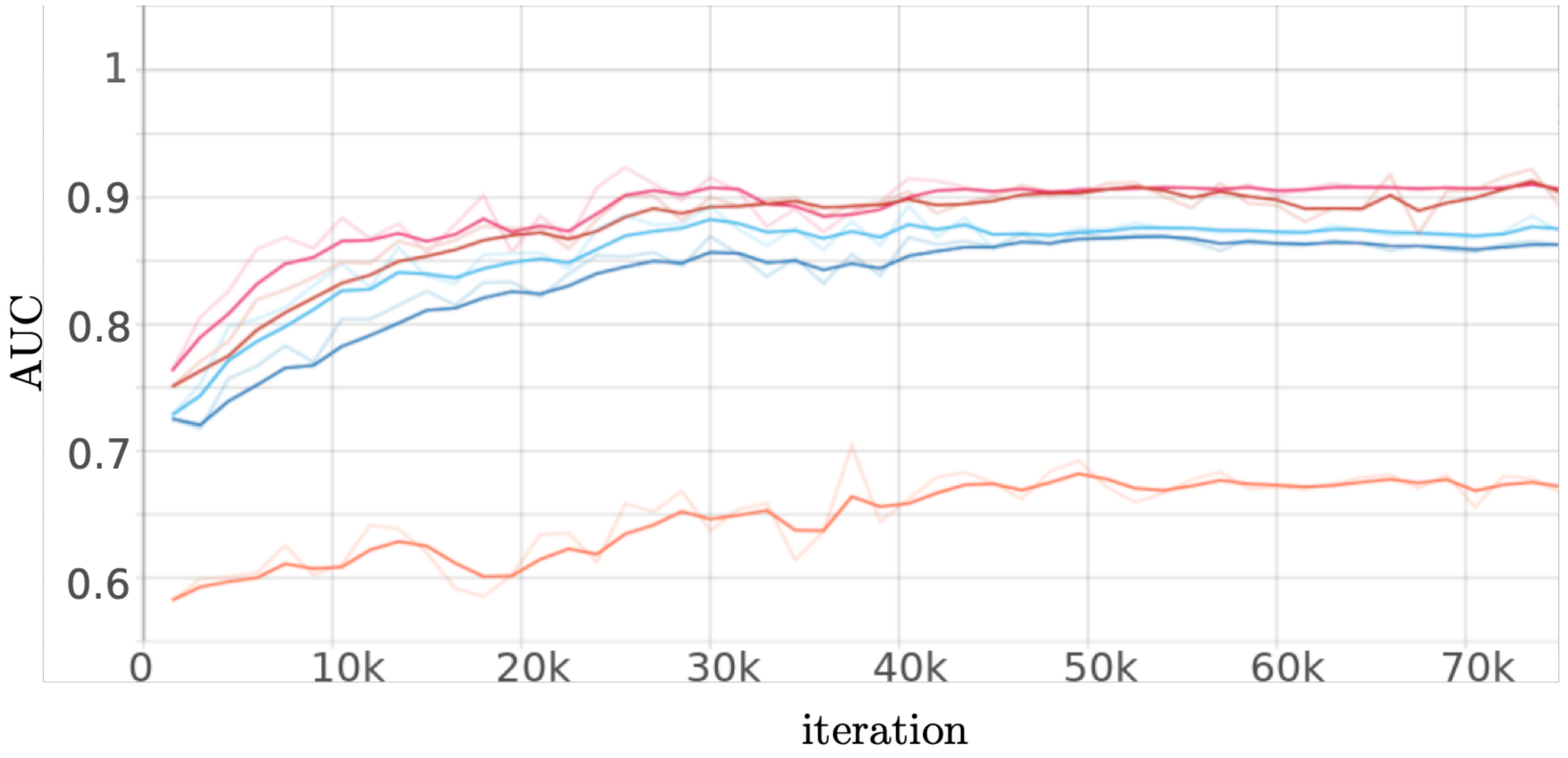
- Learn To Pay Attention (LTPA)
  - 3 attention maps
- Bonettini
  - Single attention map



# Experiments

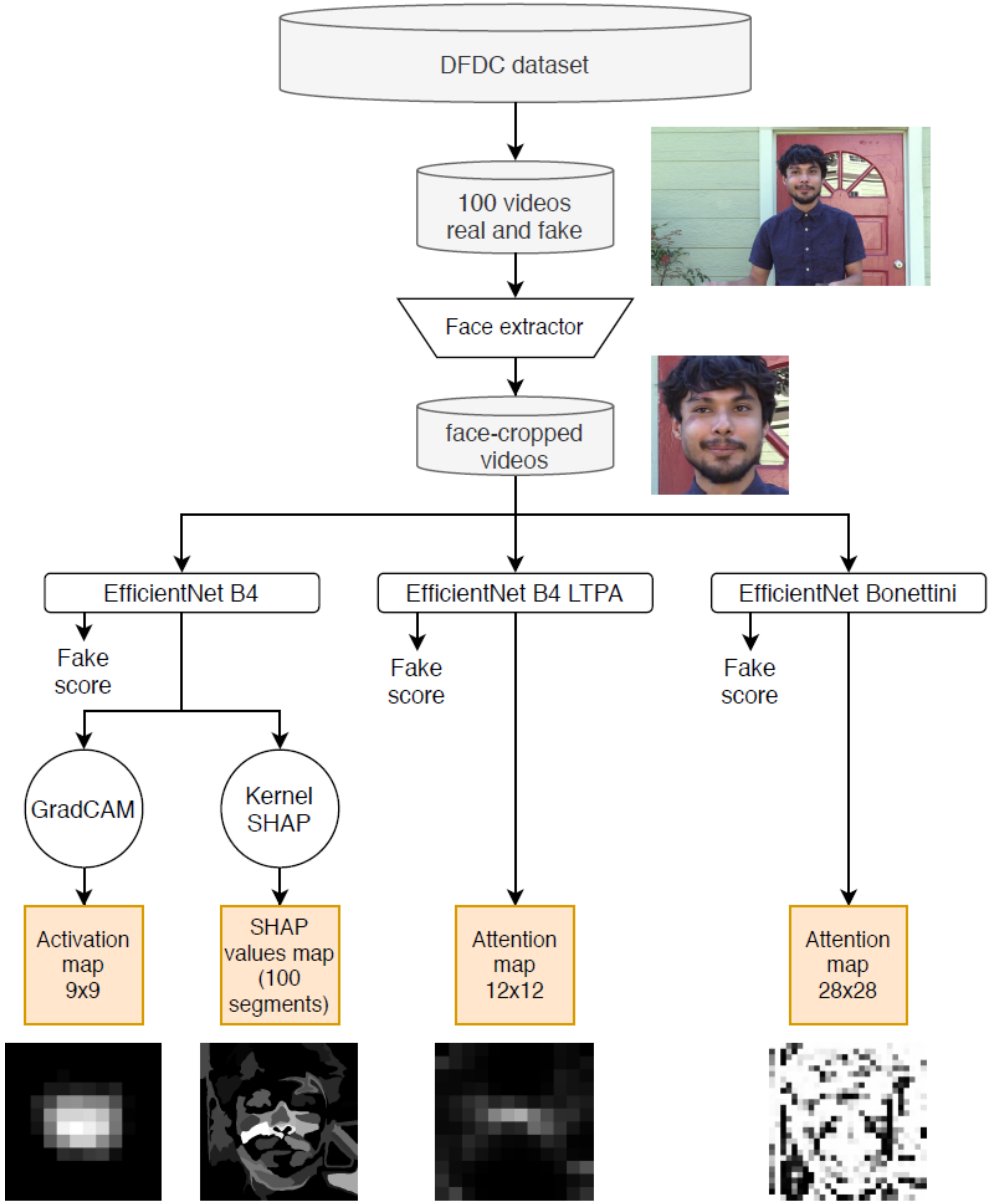
# Experiments: setup

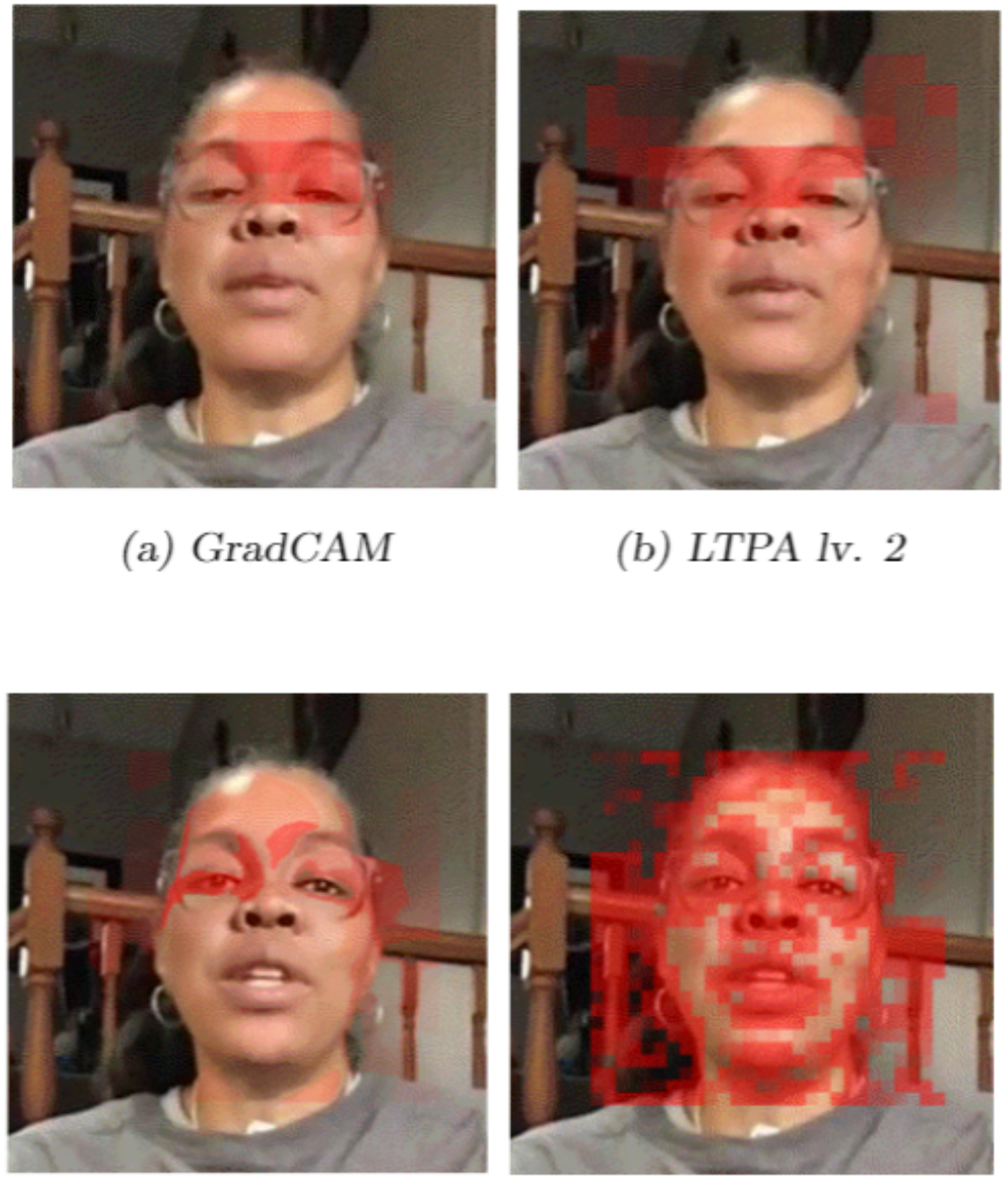
- Dataset (DFDC)
- Training




| Model | Accuracy (balanced) |
|---|---|
| EfficientNet B4, 224×224 | 0.888 |
| **EfficientNet B4, 380×380** | **0.931** |
| EfficientNet B7, 224×224 | 0.906 |
| **EfficientNet B7, 380×380** | **0.926** |
| EfficientNet B4, LTPA, 224×224 | 0.879 |
| **EfficientNet B4, LTPA, 380×380** | **0.929** |
| EfficientNet B7, LTPA, 224×224 | 0.893 |
| EfficientNet B7, LTPA, 380×380 | 0.904 |

# Experiments: explanations



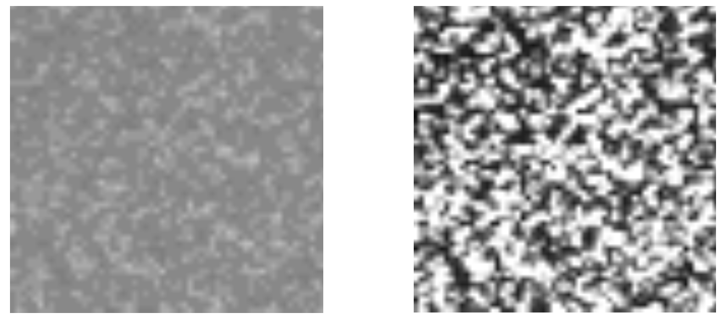



(a) GradCAM (b) LTPA lv. 2

(c) SHAP (d) Bonettini

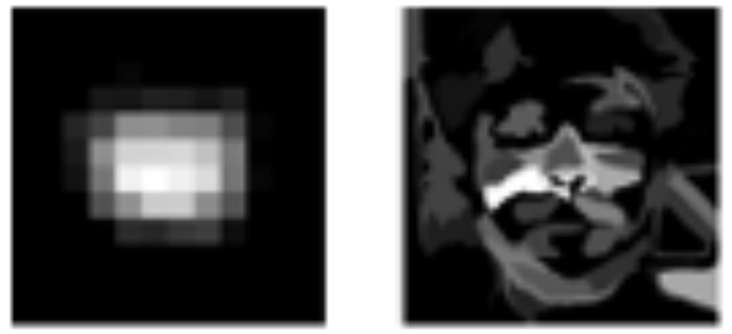
# Experiments: evaluation

- Metrics

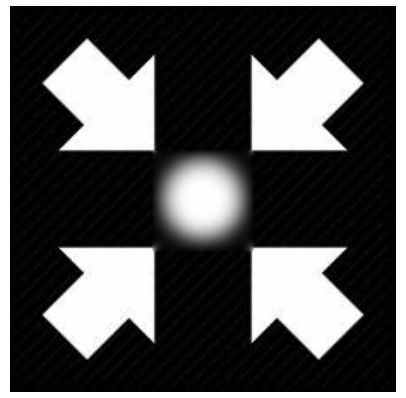


  - Variance $$V = avg_{f \in frames}(var(f))$$

  - Inter-frame consistency $$\tau = avg_{f \in frames}(PCC_{f,f+1})$$



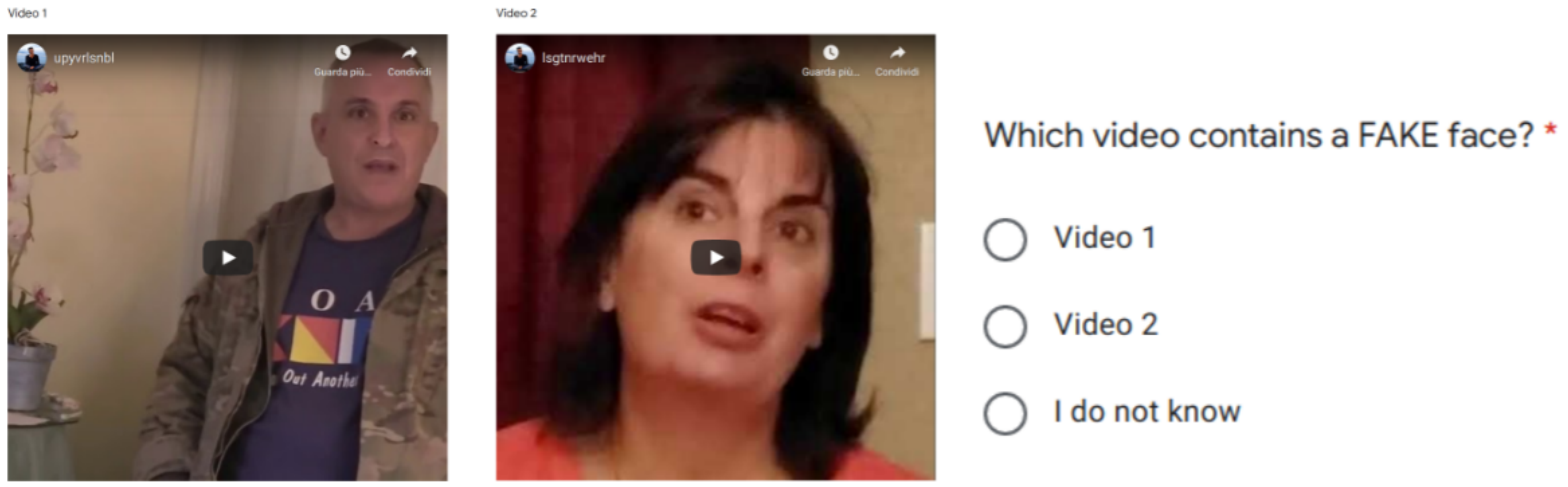
  - Intra-frame consistency $$\rho = avg_{f \in frames}\left(\frac{avg_{s \in S}(a_{0.1l \cdot s}(f))}{a_{0,0}(f)}\right)$$

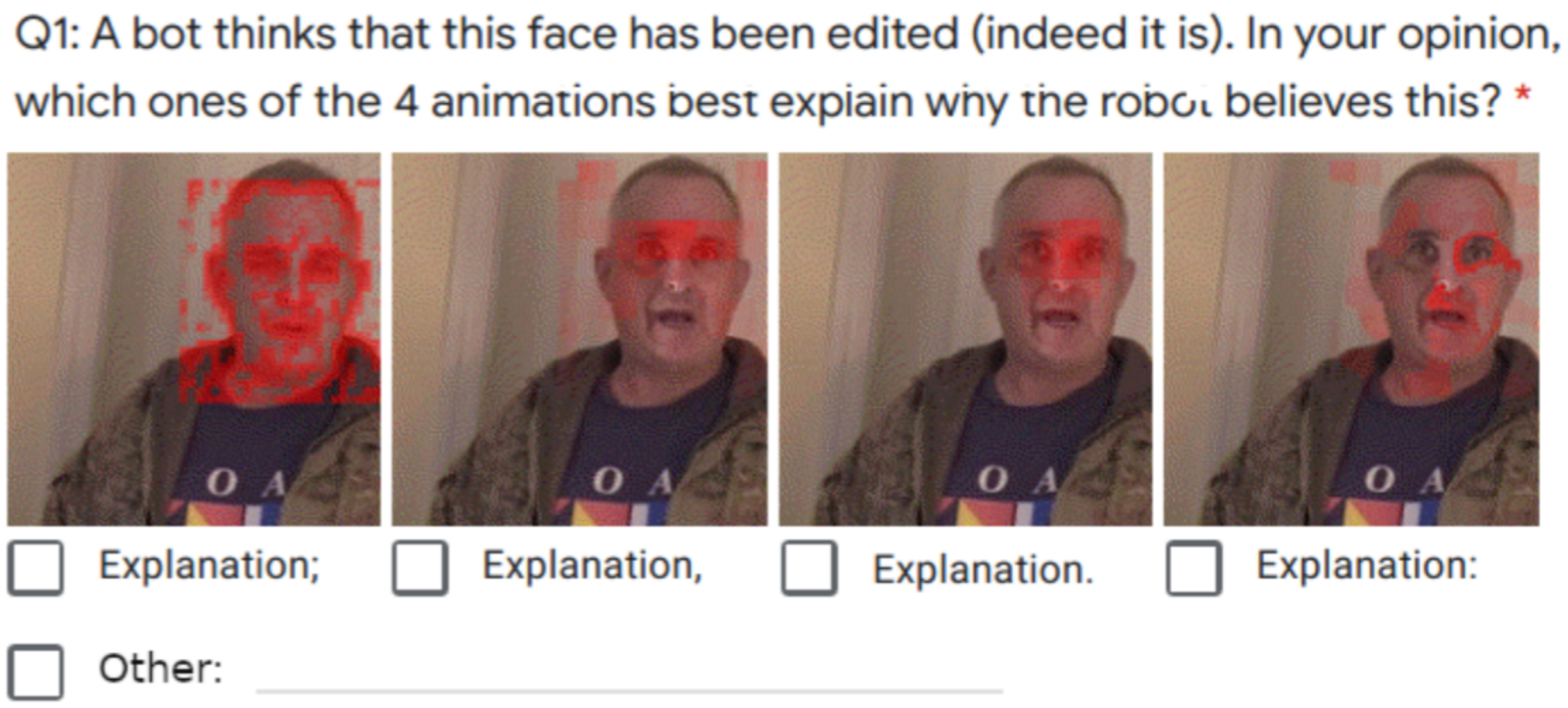


  - Centredness $$\mu = avg_{f \in frames}\left(\frac{I(inner\ 50\%)}{I(full\ frame)}\right)$$

# Experiments: evaluation

- User study
  - 20 real and 20 fake videos
  - 20 sections
  - 2 questions per section



# Results

# Results: metrics

- Average over 58 fake videos
- GradCAM performs best

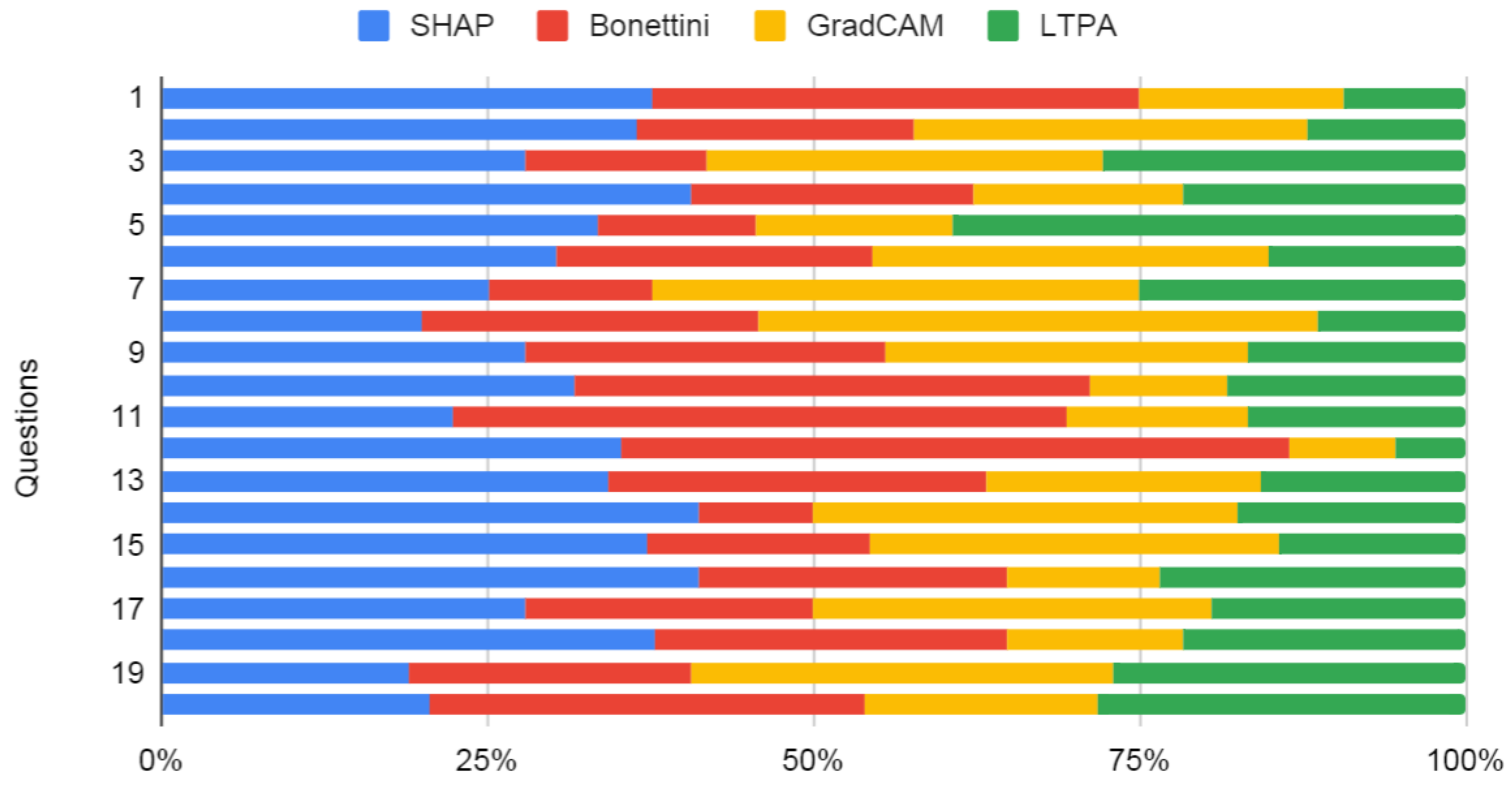
| | $V$ [0, 1] | $\tau$ [-1, 1] | $\rho$ [-1, 1] | $\mu$ [0, 1] |
|---|---|---|---|---|
| Bonettini | 0.0951 | 0.7390 | 0.1262 | 0.5286 |
| GradCAM | 0.0135 | **0.8756** | **0.7489** | **0.8666** |
| LTPA | 0.0108 | 0.7991 | 0.3333 | 0.6386 |
| SHAP | 0.0302 | 0.4496 | 0.2326 | 0.7348 |

# Results: user study

- Number of answers: 67
- Accuracy: 85%


- Preferred explainer: SHAP
- Statistical “sign test” for validation
- Preference dependent on video

| | | | |
|---|---|---|---|
| **Number of answers** | 67 | | |
| **Screen used** | Large | 43% | |
| | Small | 57% | |
| **Are you familiar with deepfakes?** | Yes | 37% | |
| | Heard of it | 33% | |
| | No | 30% | |
| **Correct video identification** | 85% | | |
| **Explainer choices** | GradCAM | 165 | 23% |
| | **SHAP** | **221** | **31%** |
| | LTPA | 137 | 19% |
| | Bonettini | 185 | 26% |

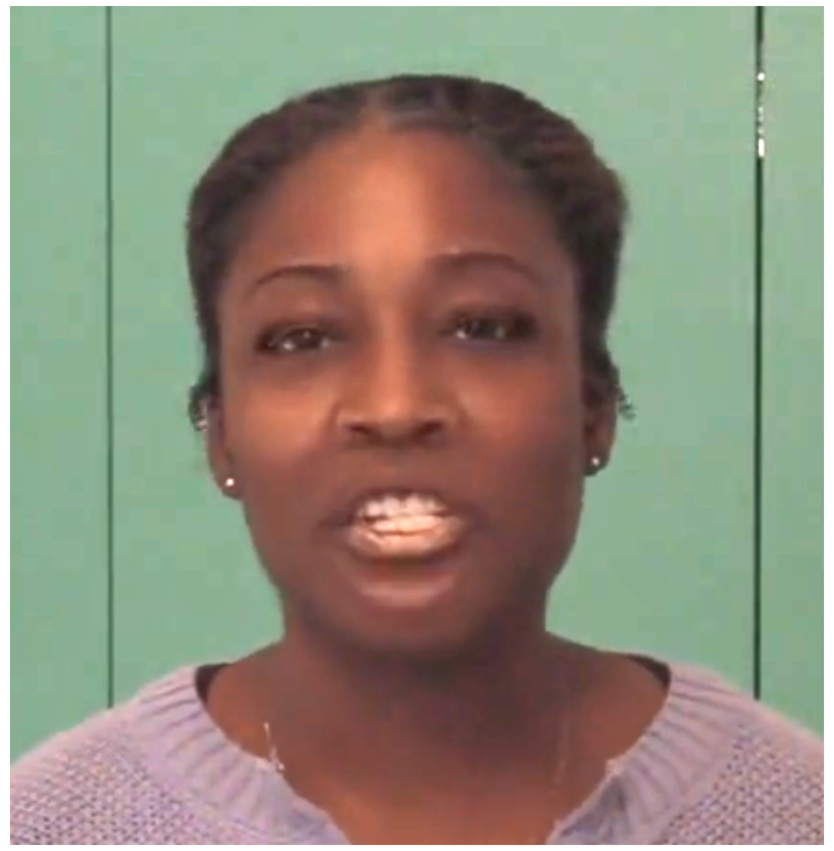


# Conclusion

# Conclusion

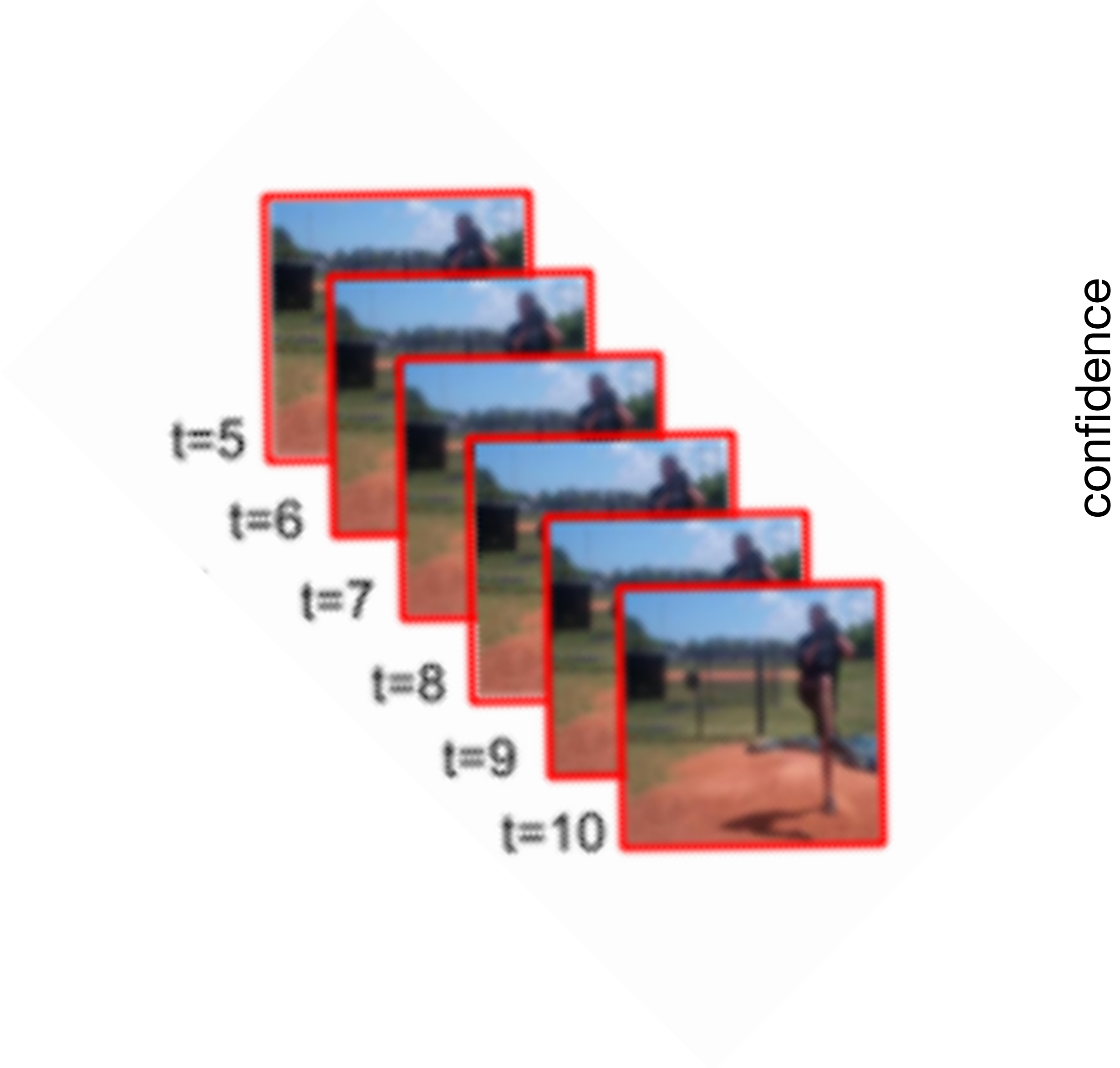
- We implemented and extended 4 explanation techniques
- We defined intrinsic and extrinsic metrics
- We empirically compared the explainers based on them
- We performed a user survey
- Human perception is not always aligned with objective metrics

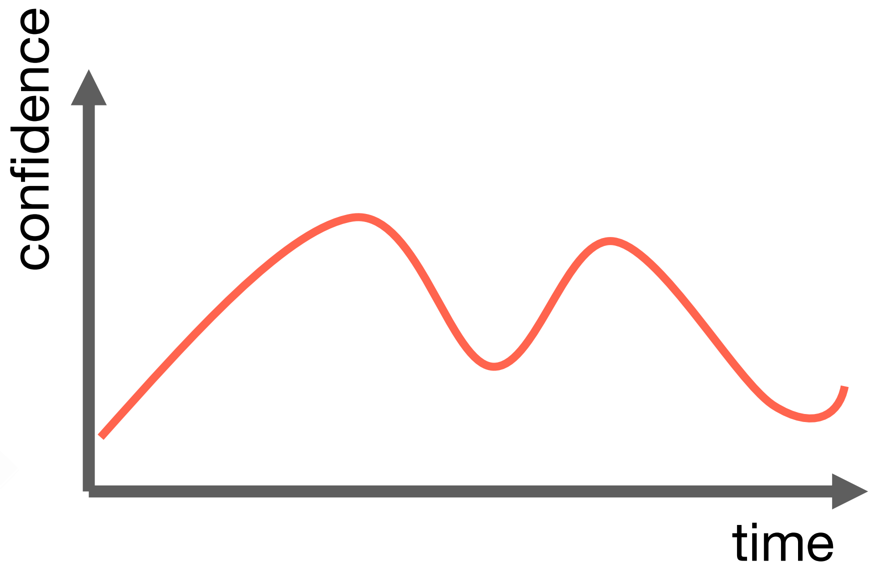
# Future work



« LIPS ARE FAKE! »

- Captioning explanation maps
- Weighting explanations on classifier's confidence



# Thank you