

Lecture 24: Image forensics

Announcements

- Final project guidelines are on webpage.
- Sign up for a presentation time slot [here](#).
- PS9 (on NeRF) will be released by tomorrow.
- Shorter than usual (to give you time for the project).

Fake images in the news

THE WALL STREET JOURNAL.
English Edition | Print Edition | Video | Audio | Latest Headlines | More

Home World U.S. Politics Economy Business Tech Markets Opinion Books & Arts Real Estate Life & Work Style Sports Search

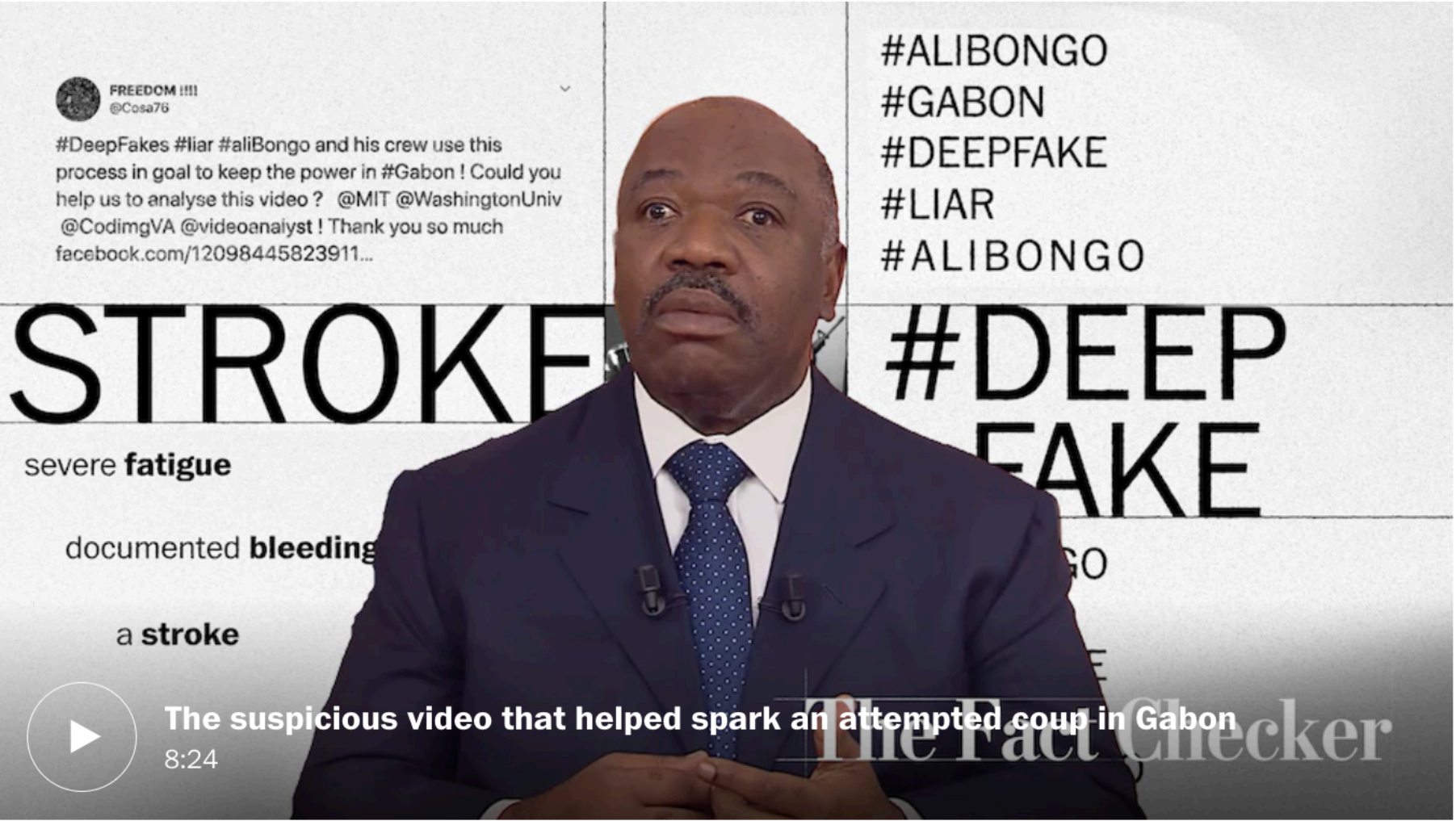


Paparazzi Photos Were the Scourge of Celebrities. Now, It's AI.
Researchers say advancements in artificial intelligence could be used to stoke misinformation about public figures. A recent image had even experts fooled.

The Washington Post
Democracy Dies in Darkness

How misinformation helped spark an attempted coup in Gabon

Analysis by [Sarah Cahlan](#)
Video reporter
February 13, 2020 at 3:00 a.m. EST



STROKE
severe **fatigue**
documented **bleeding**
a **stroke**

#ALIBONGO
#GABON
#DEEPPFAKE
#LIAR
#ALIBONGO
#DEEP FAKE

The suspicious video that helped spark an attempted coup in Gabon
8:24
The Fact Checker

Gabon's president was ill. He had not been seen in public for months. A week after his first video address, there was an attempted coup. (Video: Sarah Cahlan/The Washington Post)

Text-to-image models make it easy



“Catholic Pope Francis wearing Balenciaga puffy jacket in drill rap music video, throwing up gang signs with hands, taken using a Canon EOS R camera with a 50mm f/1.8 lens, f/2.2 aperture, shutter speed 1/200s, ISO 100 and natural light, Full Body, Hyper Realistic Photography, Cinematic, Cinema, Hyperdetail, UHD, Color Correction, hdr, color grading, hyper realistic CG animation --ar 4:5 --upbeta --q 2 --v 5.”

But image manipulation also has a long history



Abraham Lincoln?



John C. Calhoun

But image manipulation also has a long history



From Forrest Gump, 1994

Malicious image manipulation



Malicious image manipulation

Fonda Speaks To Vietnam Veterans At Anti-War Rally



Actress And Anti-War Activist Jane Fonda Speaks to a crowd of Vietnam Veterans as Activist and former Vietnam Vet John Kerry (LEFT) listens and prepares to speak next concerning the war in Vietnam (AP Photo)

Malicious image manipulation

Fonda Speaks To Vietnam Veterans At Anti-War Rally



Actress And Anti-War Activist Jane Fonda Speaks to a crowd of Vietnam Veterans as Activist and former Vietnam Vet John Kerry (LEFT) listens and prepares to speak next concerning the war in Vietnam (AP Photo)



Malicious image manipulation

Fonda Speaks To Vietnam Veterans At Anti-War Rally



Actress And Anti-War Activist Jane Fonda Speaks to a crowd of Vietnam Veterans as Activist and former Vietnam Vet John Kerry (LEFT) listens and prepares to speak next concerning the war in Vietnam (AP Photo)



his associates simply found photos of athletes on the Internet and either used those photos or used software such as PhotoShop to insert the applicants' faces onto the bodies of legitimate athletes. For example, as set forth in greater detail below, CW-1 explained to McGLASHAN that he would create a falsified athletic profile for McGLASHAN's son, something he told McGLASHAN he had "already done ... a million times," and which would involve him using "Photoshop and stuff" to deceive university admissions officers.

FBI affidavit on 2019 college admissions scandal

Malicious image manipulation



Malicious image manipulation



Detecting fake images



New image manipulation methods are emerging every day

ProGAN



2018

StyleGAN2



2019

2020

StyleGAN3



2021

DALL-E 2



2022

Midjourney



2023

2024

The challenge of fake image detection

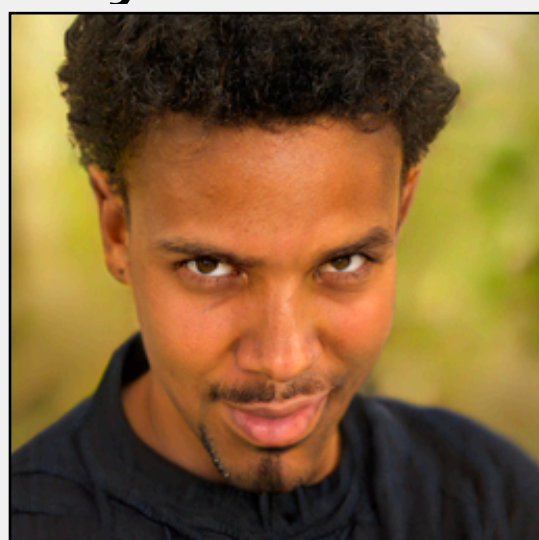
Training data

Images from known methods

ProGAN



StyleGAN2



Test data

Images from future methods

StyleGAN3



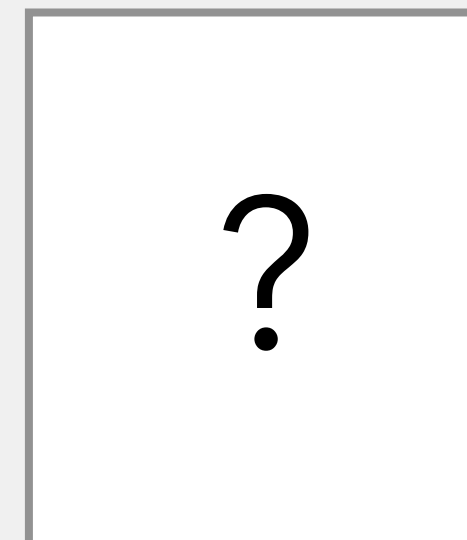
DALL-E 2



Midjourney



[github.com/.../
mycoolgenerator](https://github.com/.../mycoolgenerator)



2018

2019

2020

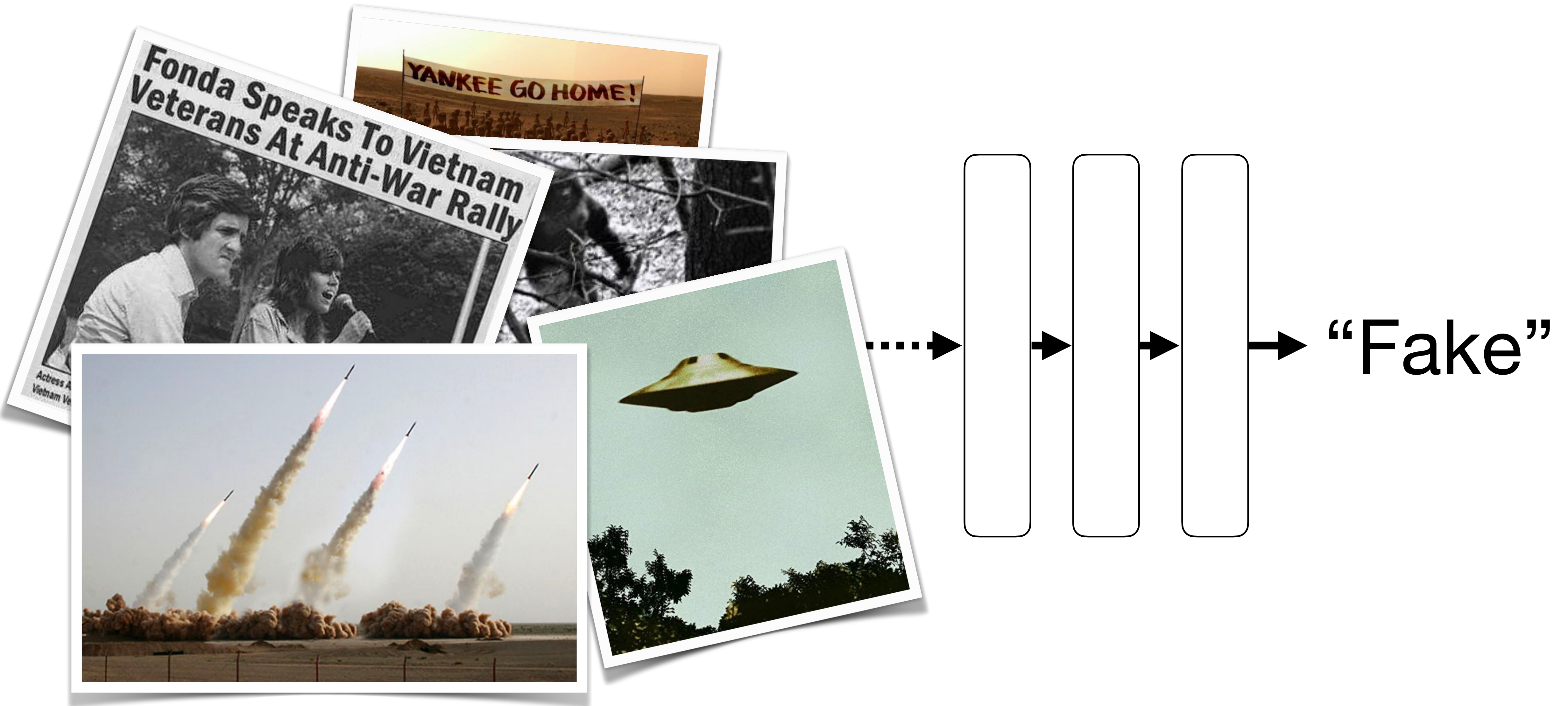
2021

2022

2023

2024

Hard to directly use supervised learning!



Strategy #1: physical models

Self-consistent lighting direction

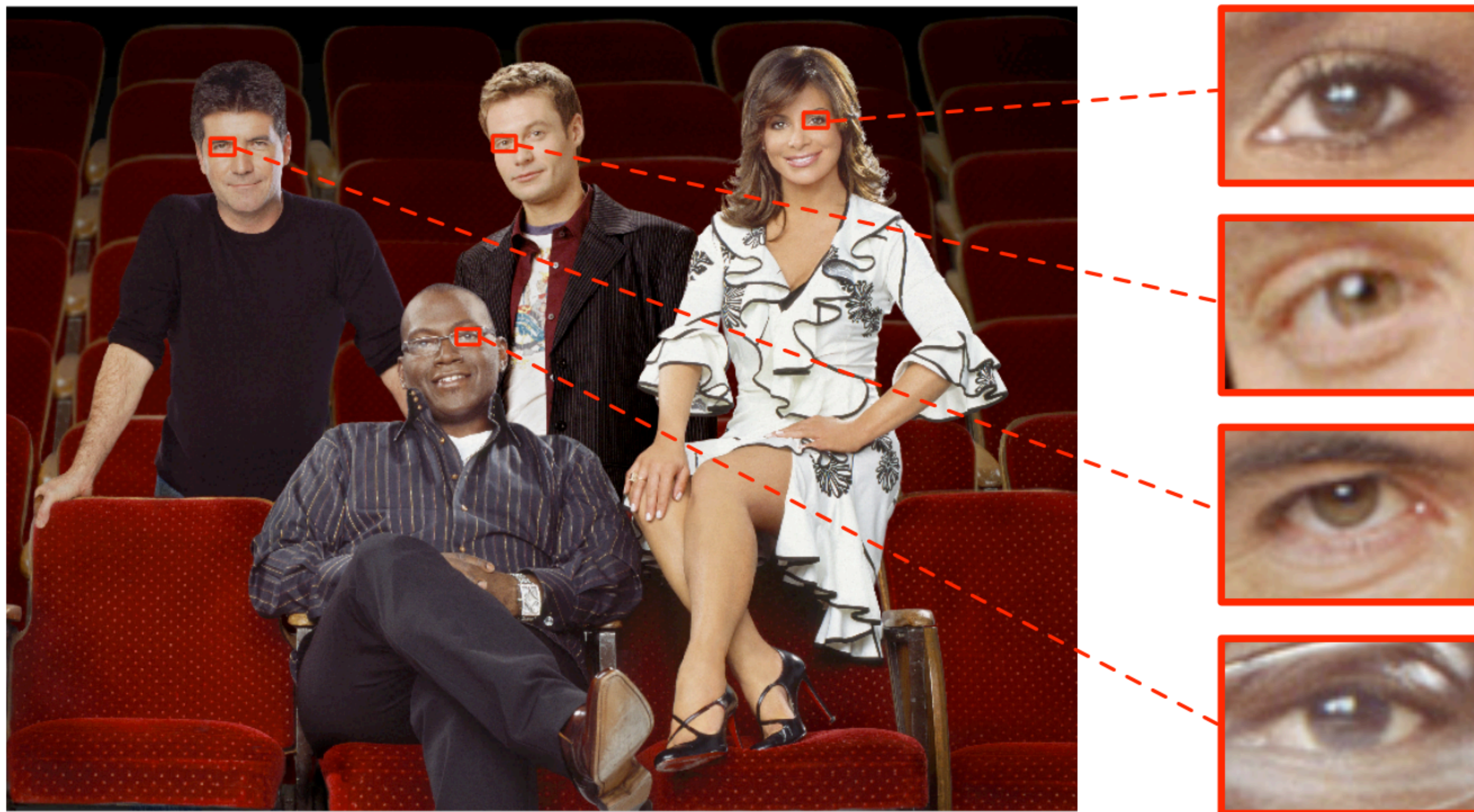
Fake photo



Real photo



Specular reflections



[Johnson and Farid, 2007]

Strategy #2: low-level imaging properties

JPEG artifacts

- Cameras vary in how they do JPEG compression.
- When you quantize a floating point numbers:
 - Some do **round()**, others do **floor()** or **ceil()**
- If a photo seems to have *both* kinds of quantization, it's probably a fake:
e.g., a composite from images taken by different cameras!

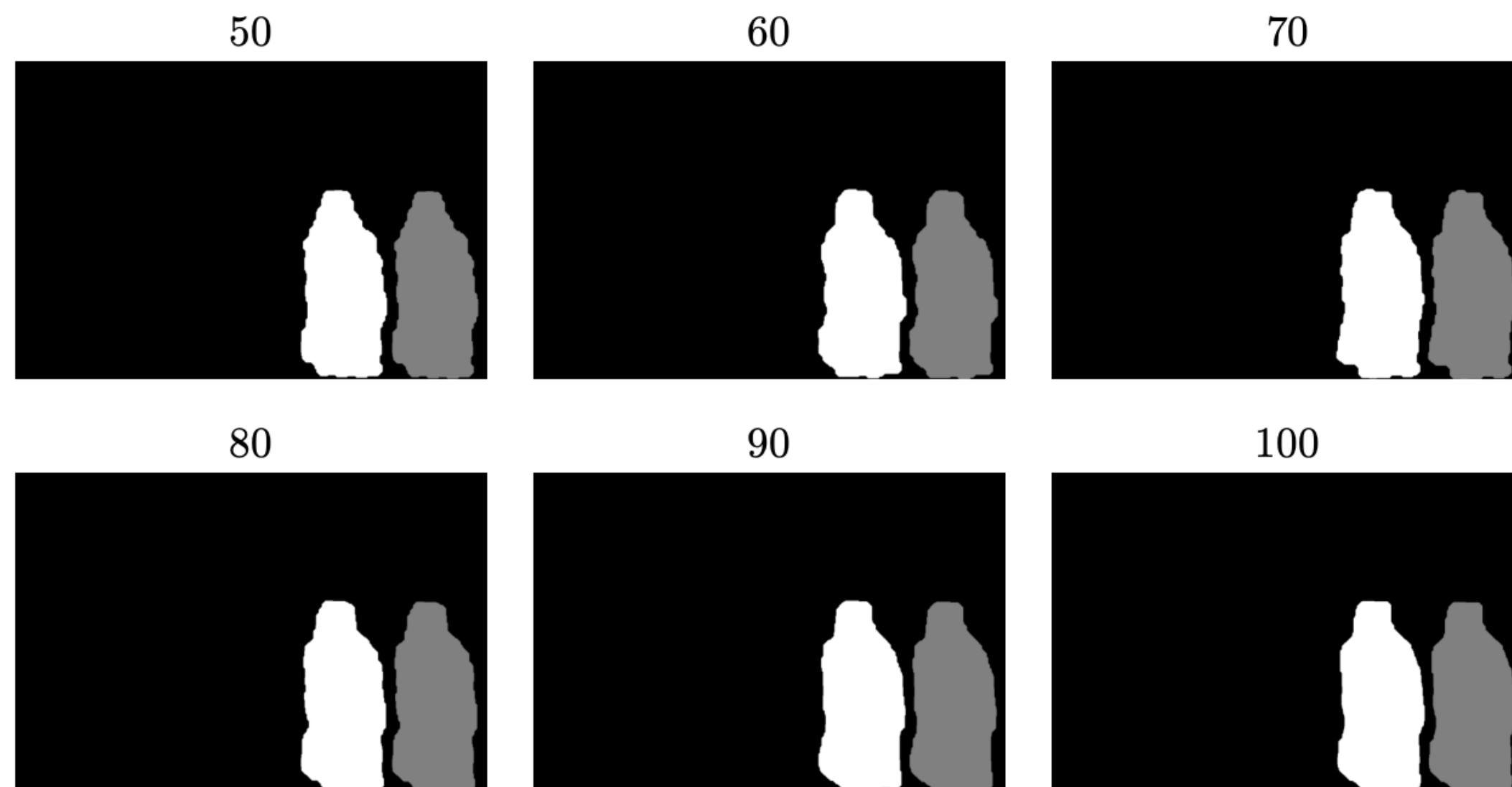


[Agarwal and Farid, “JPEG Dimples”, 2017]

Detecting duplicated image regions



- Traditional inpainting methods copy-and-paste image patches.
- Detect near-duplicated patches.
- But sensitive to postprocessing operations, like compression.



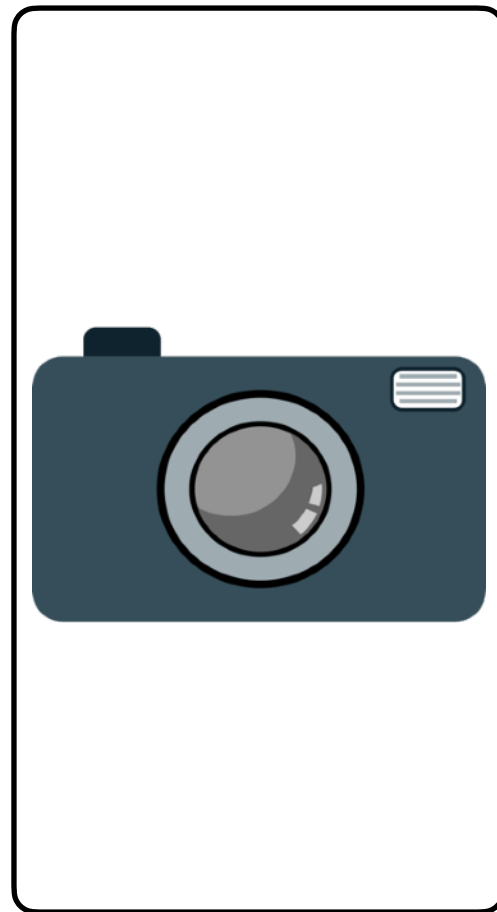
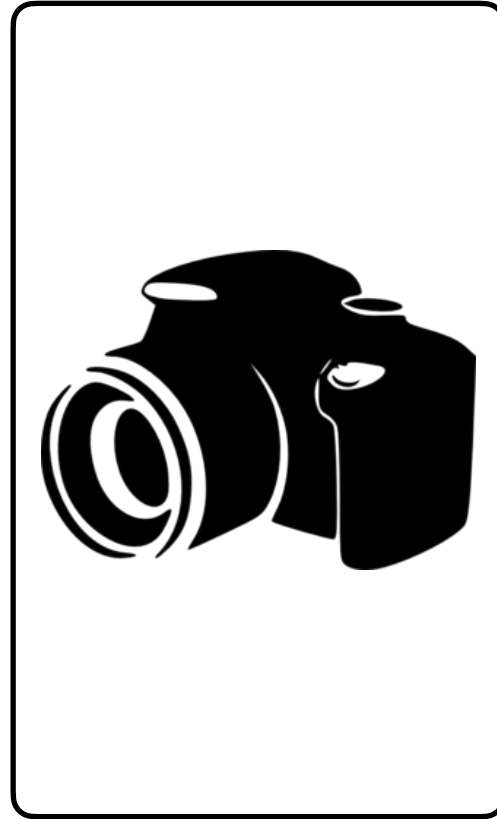
← amount of JPEG compression

[Popescu and Farid, 2004]

Strategy #3: learned anomaly detection

Instead of hand-crafting cues, can we learn to detect “anomalous” images, and flag suspicious images?







Inconsistent

Consistent

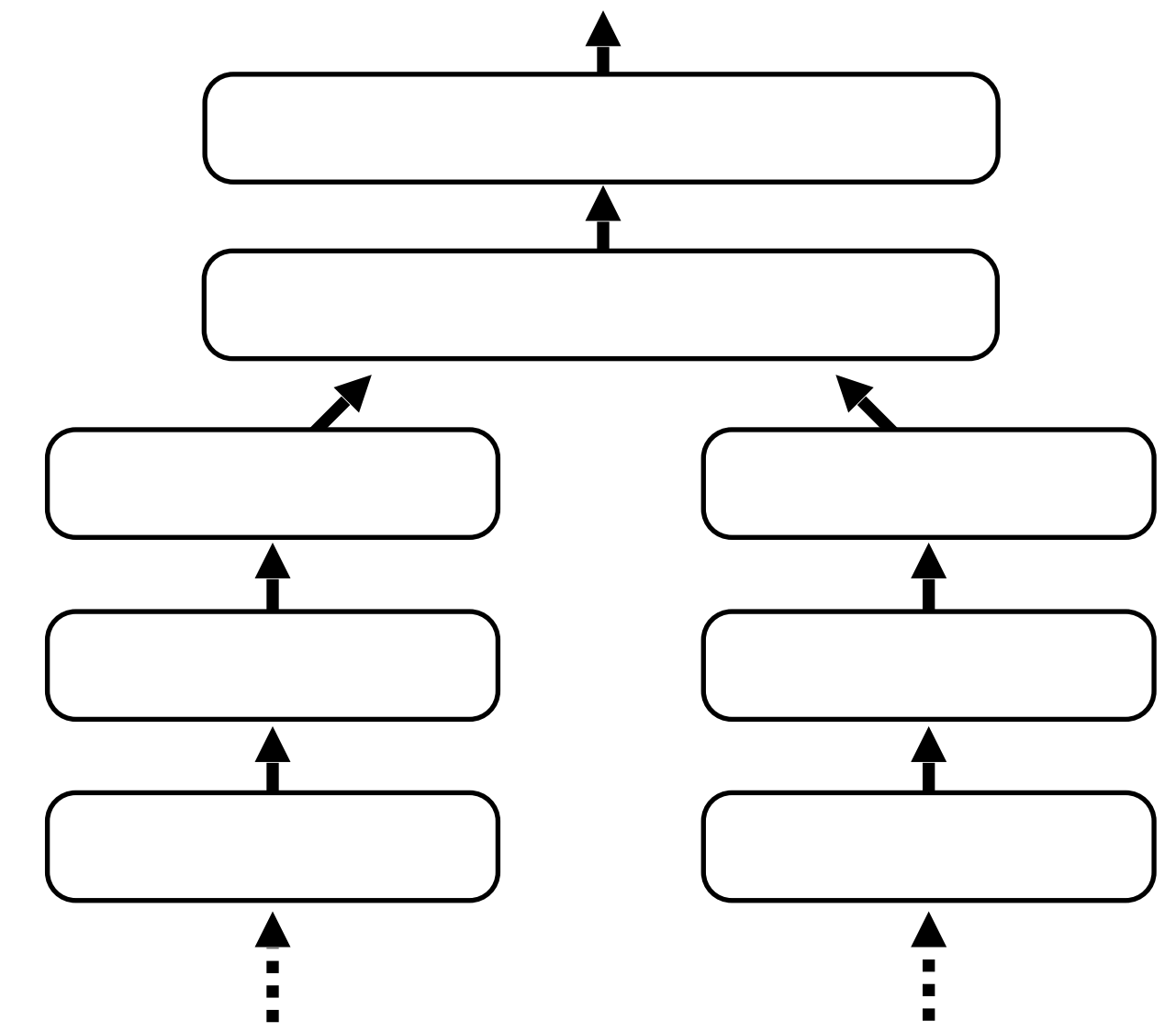
Predicting metadata consistency



CameraMake: Apple
CameraModel: iPhone 4s
ColorSpace: sRGB
ExifImageLength: 2448
ExifImageWidth: 3264
Flash: Flash did not fire
FocalLength: 107/2
WhiteBalance: Auto
ExposureTime: 1/2208
...

Same white balance?

~~Different~~



CameraMake: NIKON CORPORATION
CameraModel: NIKON D90
ColorSpace: sRGB
ExifImageLength: 2848
ExifImageWidth: 4288
Flash: Flash did not fire
FocalLength: 18/796
WhiteBalance: Auto
ExposureTime: 1/30
...

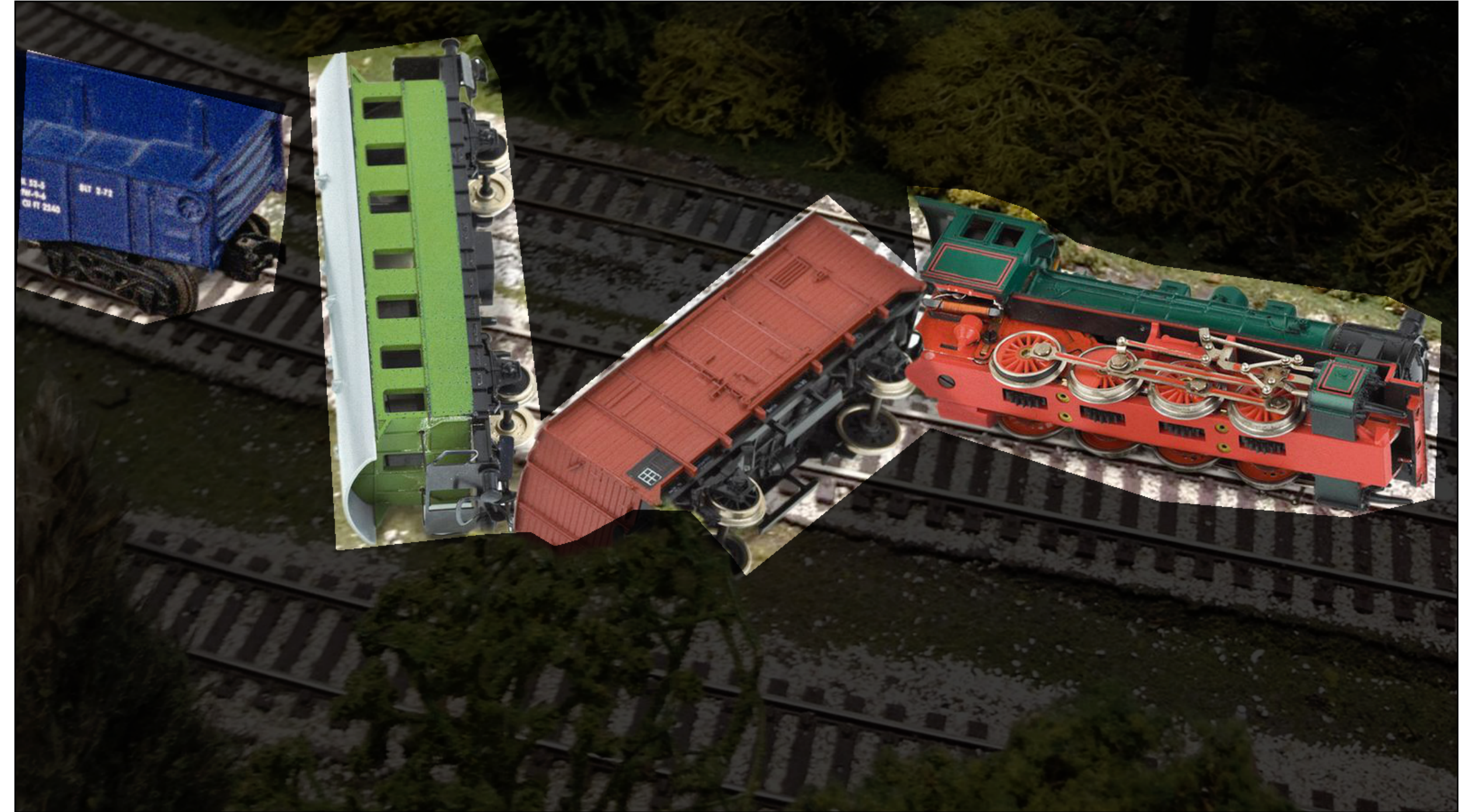


Input

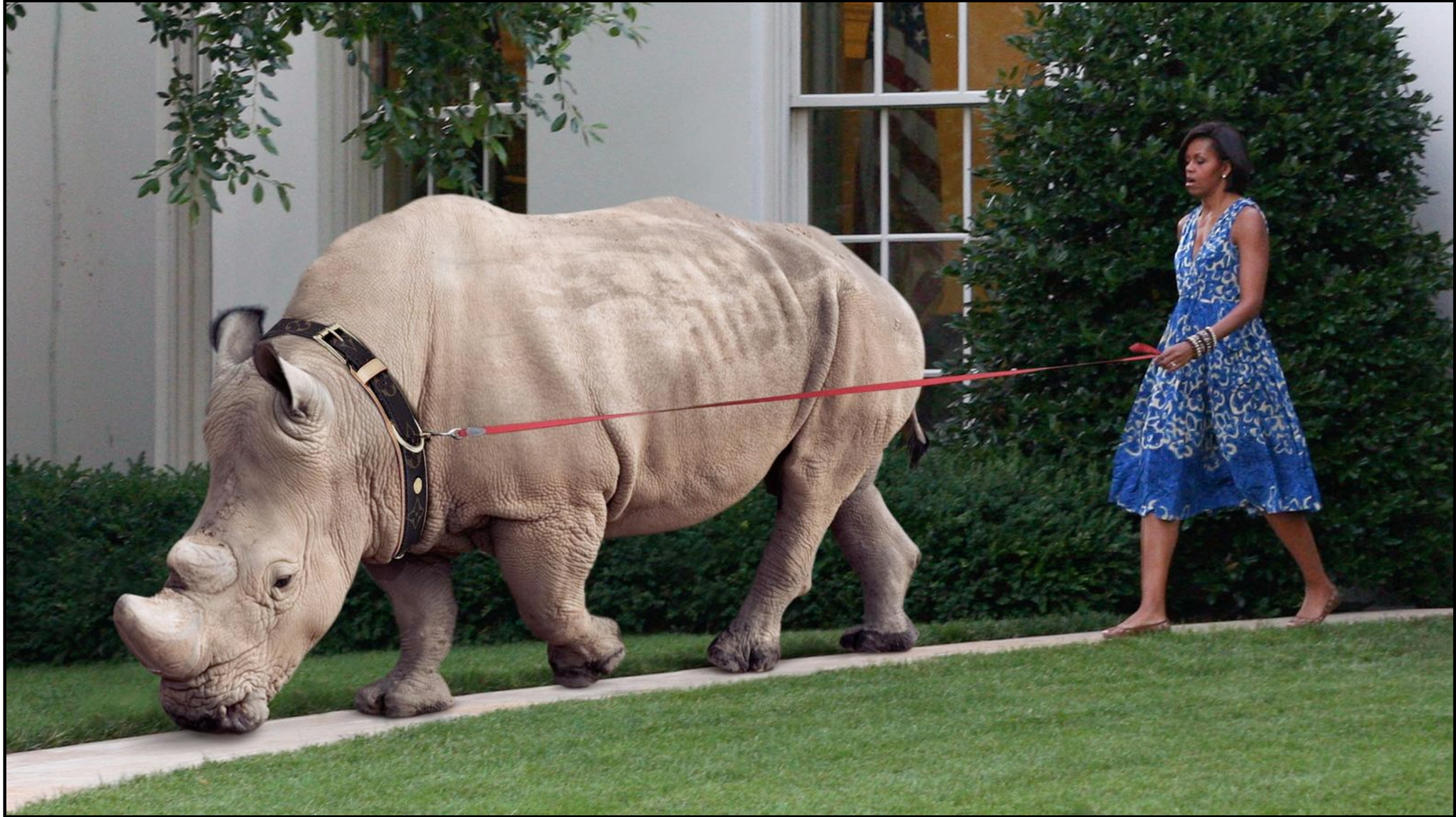
Photo source: TheOnion.com



Prediction

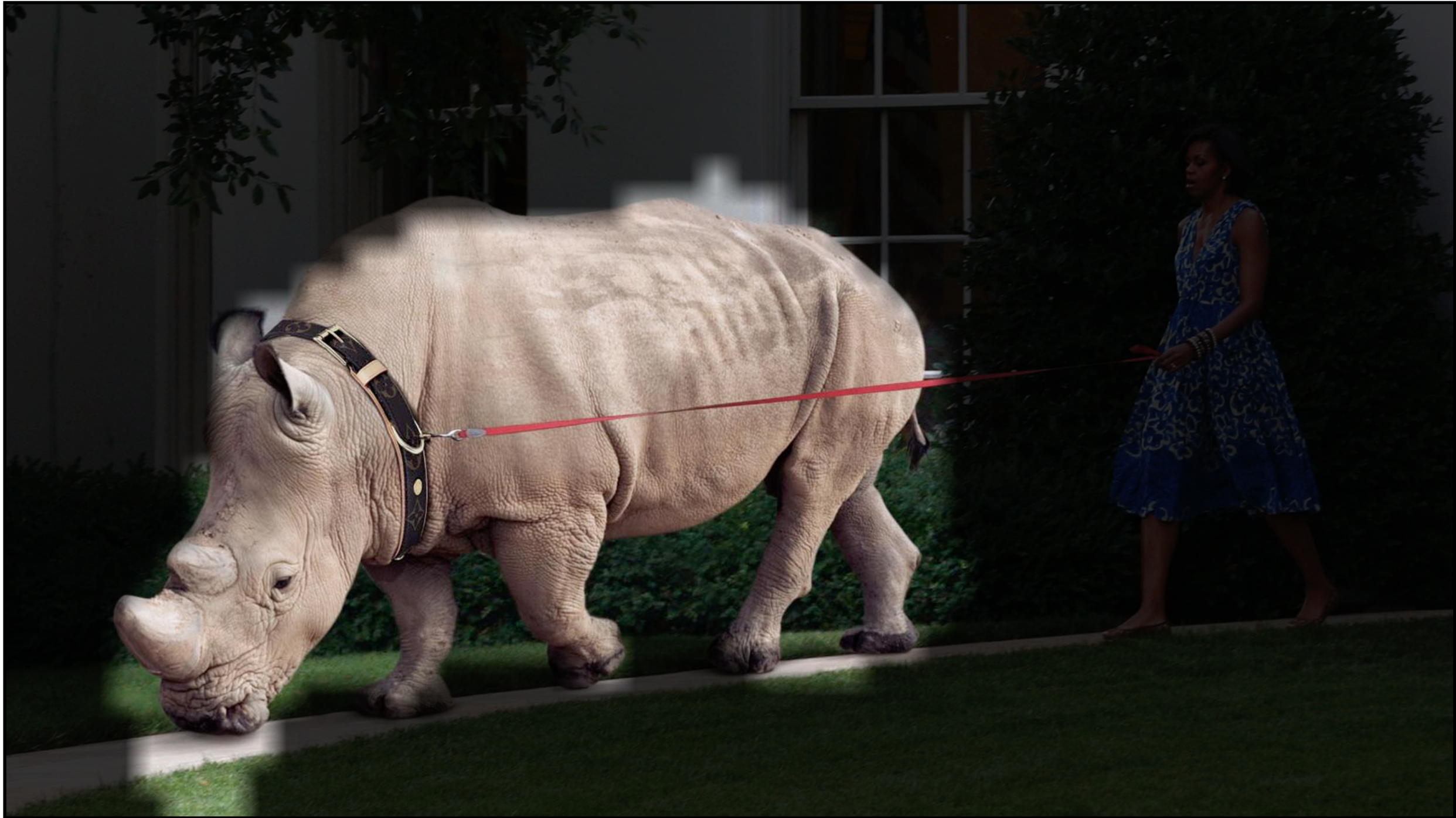


Ground truth

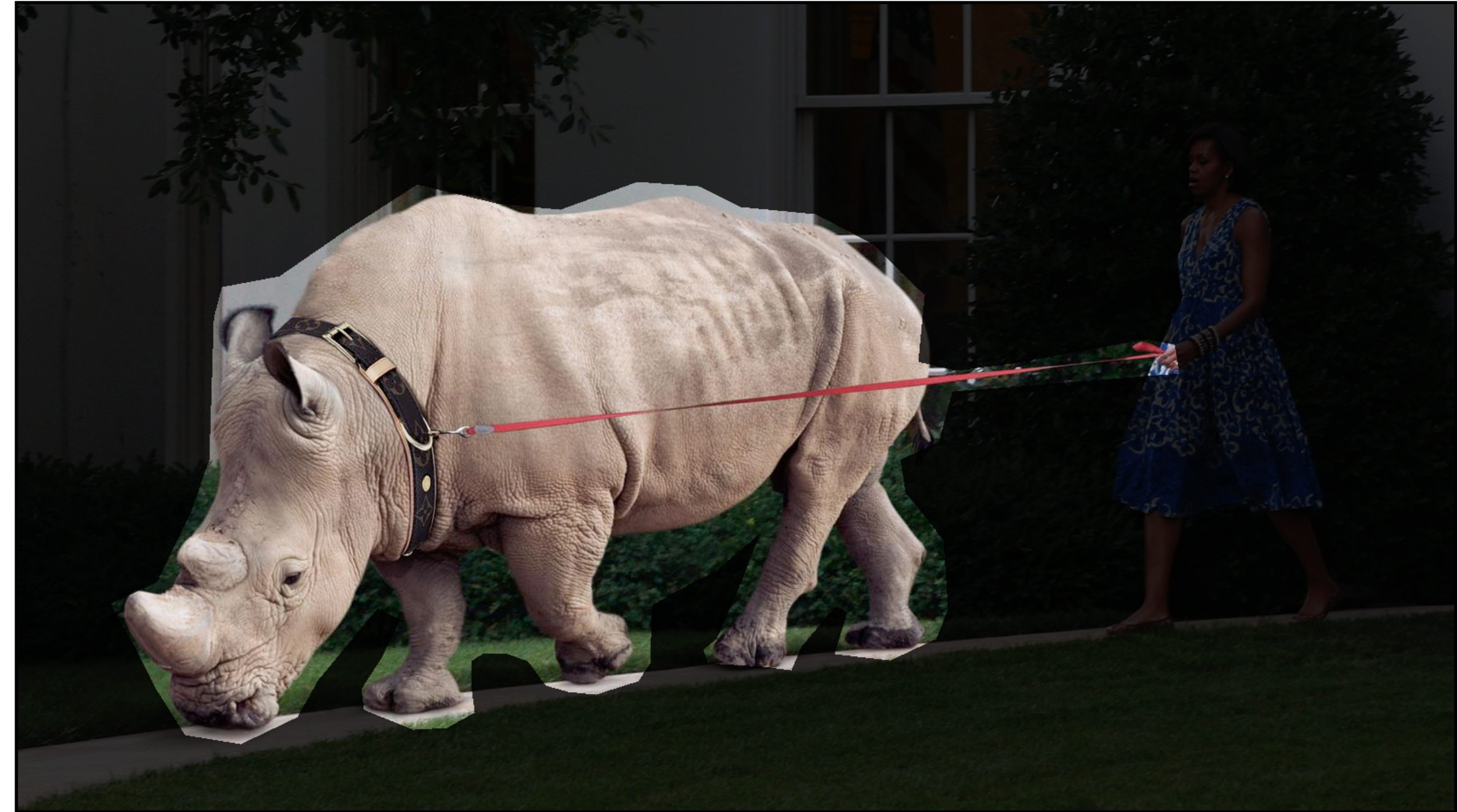


Input

Photo source: TheOnion.com



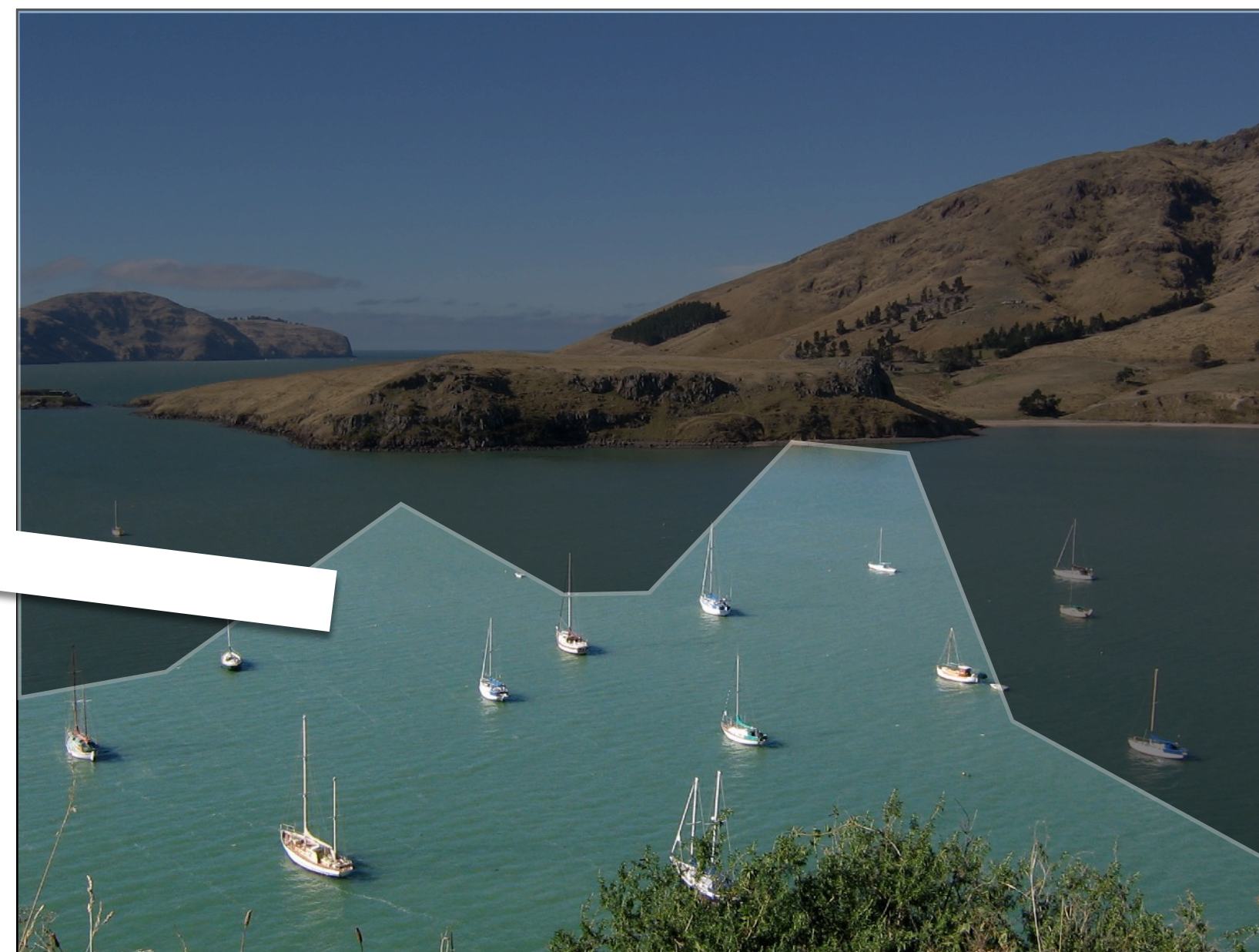
Prediction



Ground truth



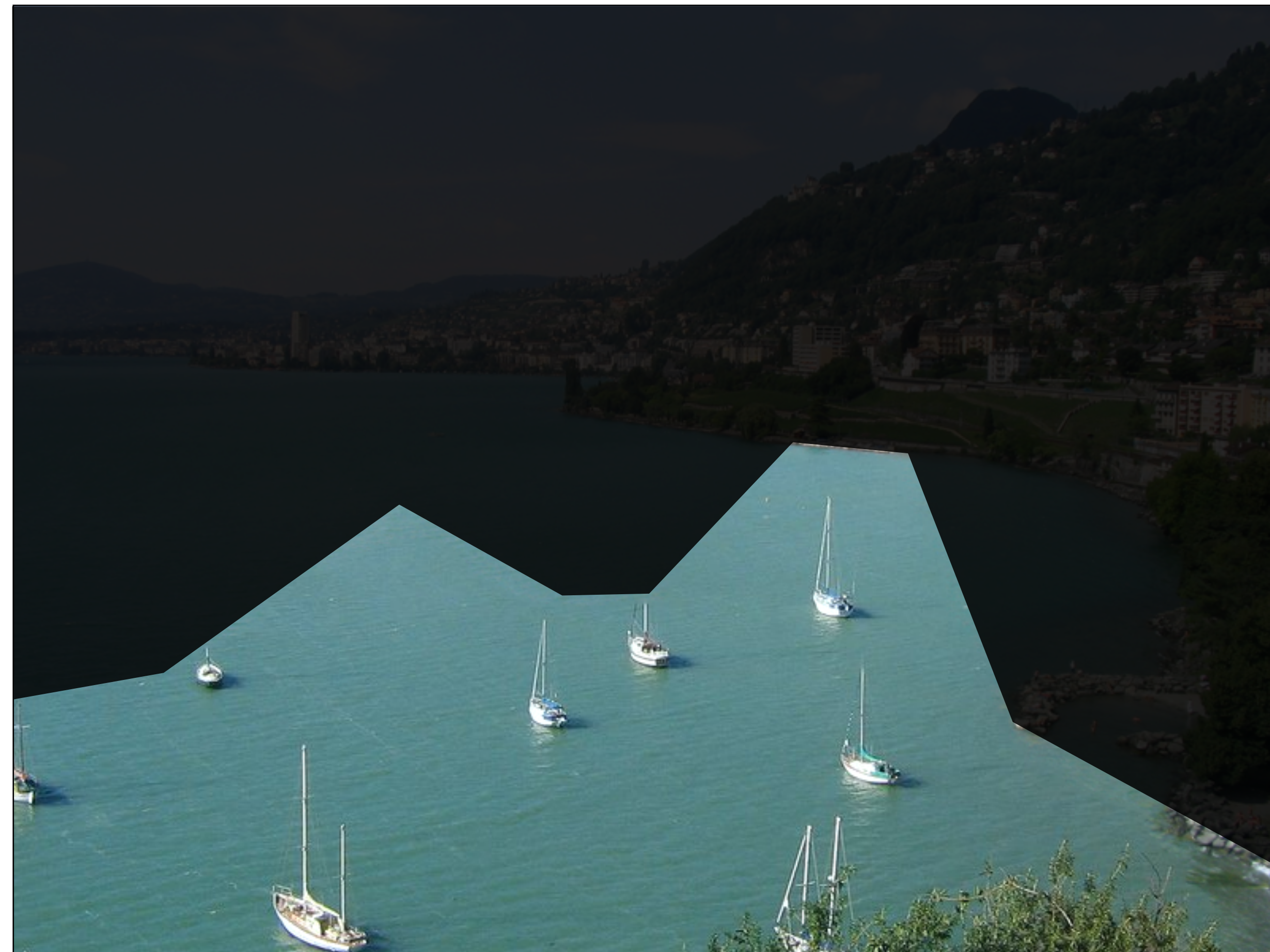
Input



(Hays & Efros 2009)



Prediction



Ground truth

Another approach: learning joint embeddings

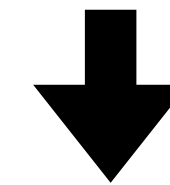


Make: NIKON
Model: NIKON D3200
Flash: Fired
Exposure Time: 1/500
Focal Length: 30.0mm
Exposure Program: Aperture Scene
Capture Type: Standard
...

Learning Joint Embeddings



Make: NIKON
Model: NIKON D3200
Flash: Fired
Exposure Time: 1/500
Focal Length: 30.0mm
Exposure Program: Aperture
Components Configuration: YCbCr
...



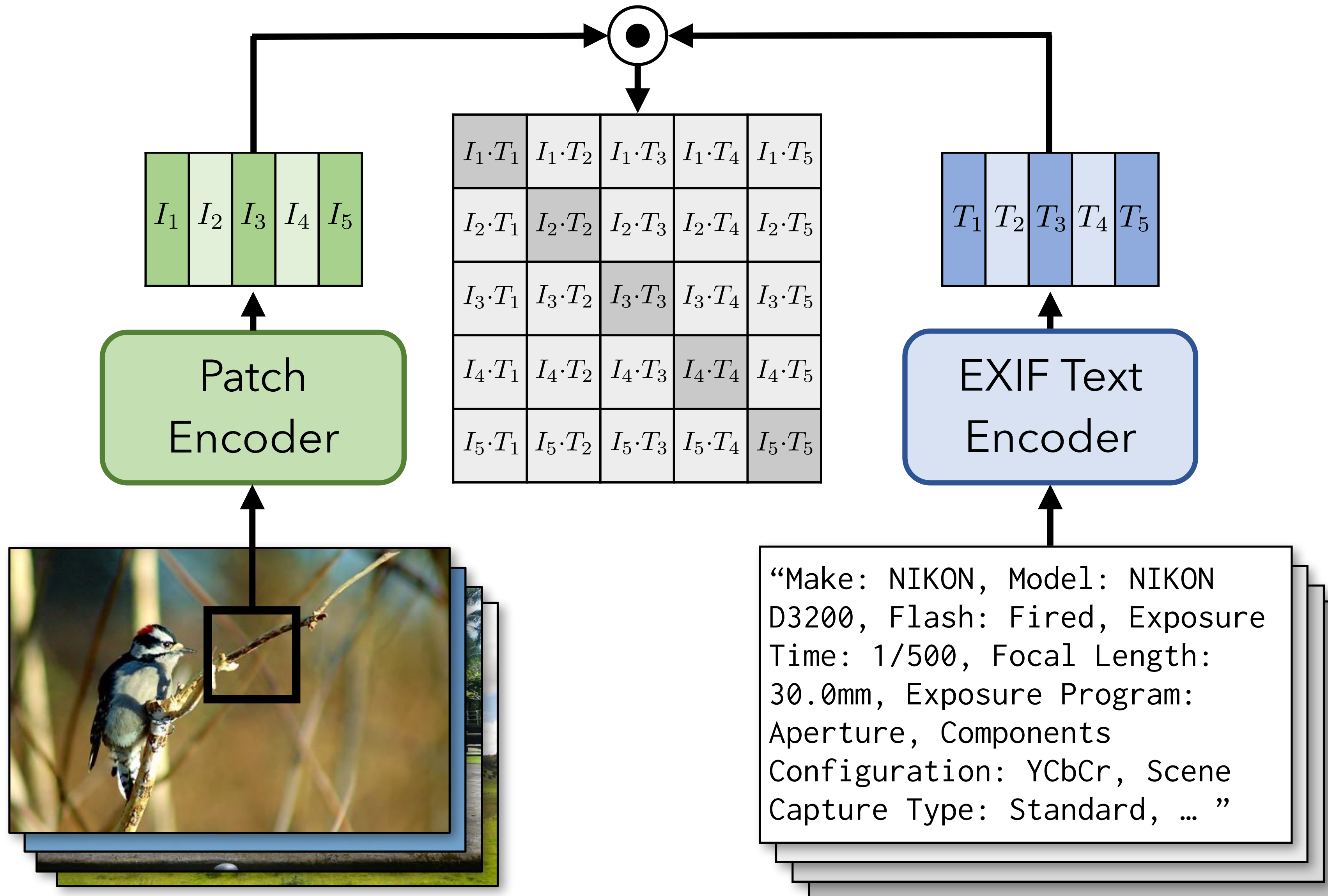
“Make: NIKON, Model: NIKON
D3200, Flash: Fired, Exposure
Time: 1/500, Focal Length:
30.0mm, Exposure Program:
Aperture, Components
Configuration: YCbCr, Scene
Capture Type: Standard, ...”

Learning Joint Embeddings



“Make: NIKON, Model: NIKON
D3200, Flash: Fired, Exposure
Time: 1/500, Focal Length:
30.0mm, Exposure Program:
Aperture, Components
Configuration: YCbCr, Scene
Capture Type: Standard, ...”

Learning Joint Embeddings



Linear classification evaluation

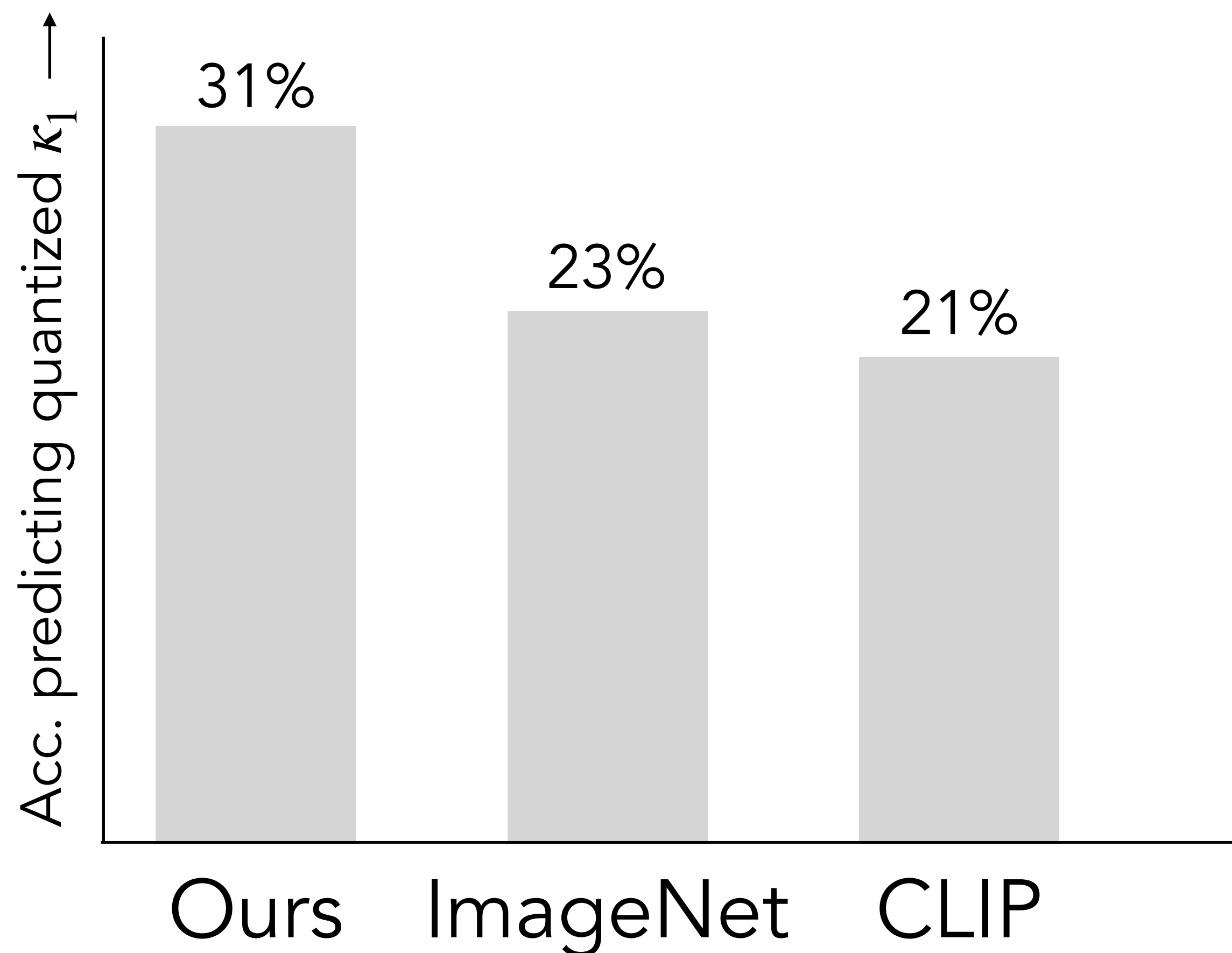


Radial distortion



Image manipulation

Linear classification evaluation



Radial distortion estimation
(Dresden dataset)

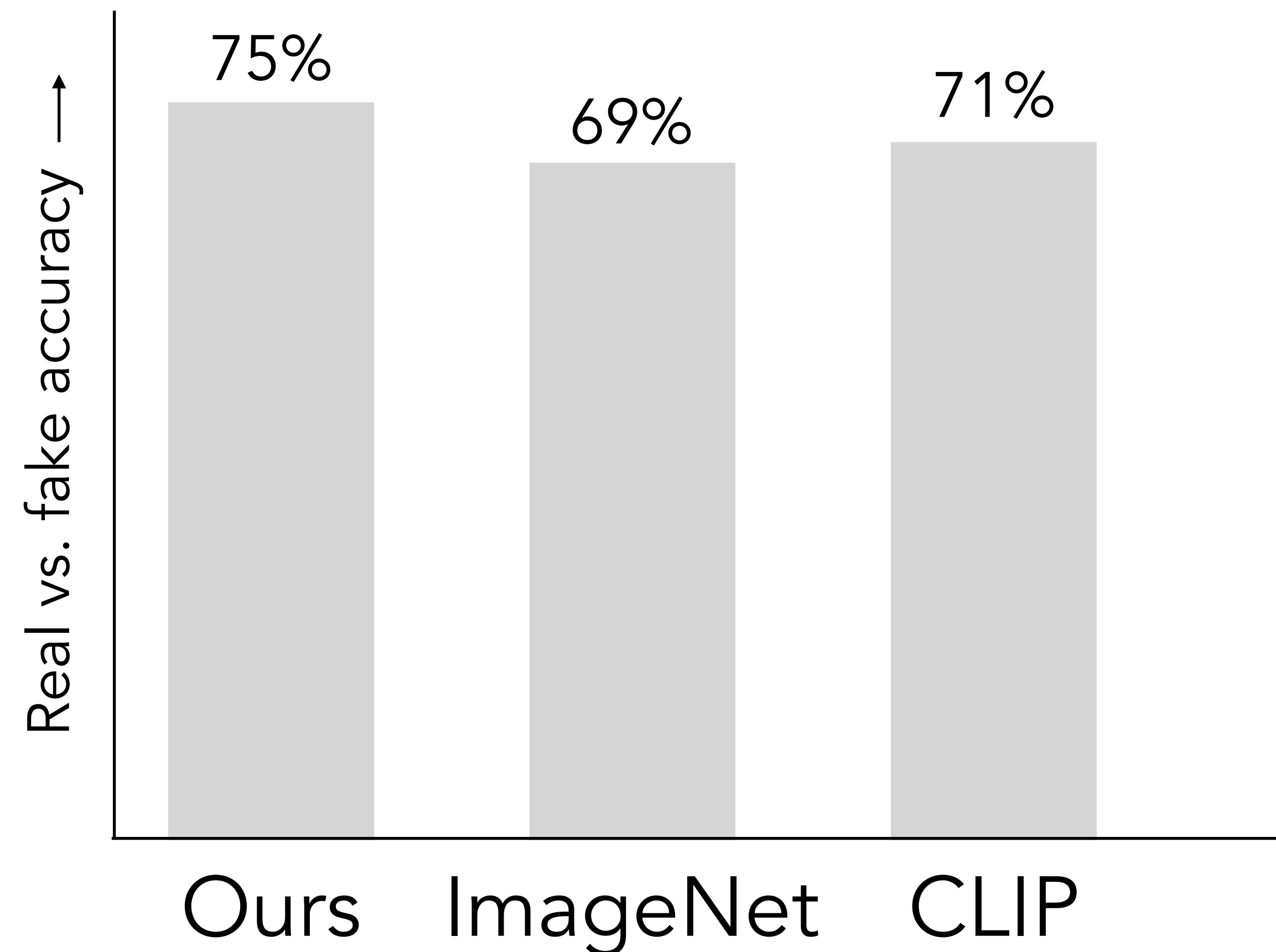
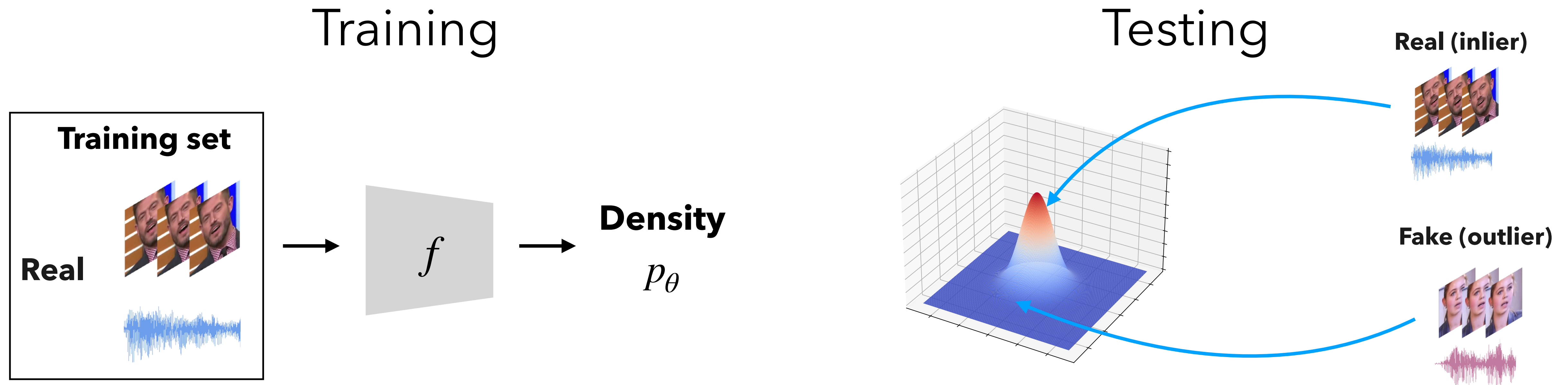


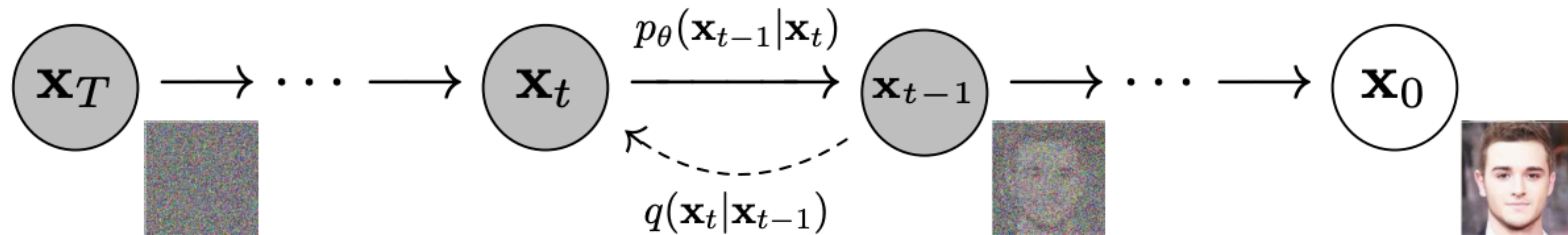
Image splice detection
(CASIA I dataset)

Video forensics as anomaly detection



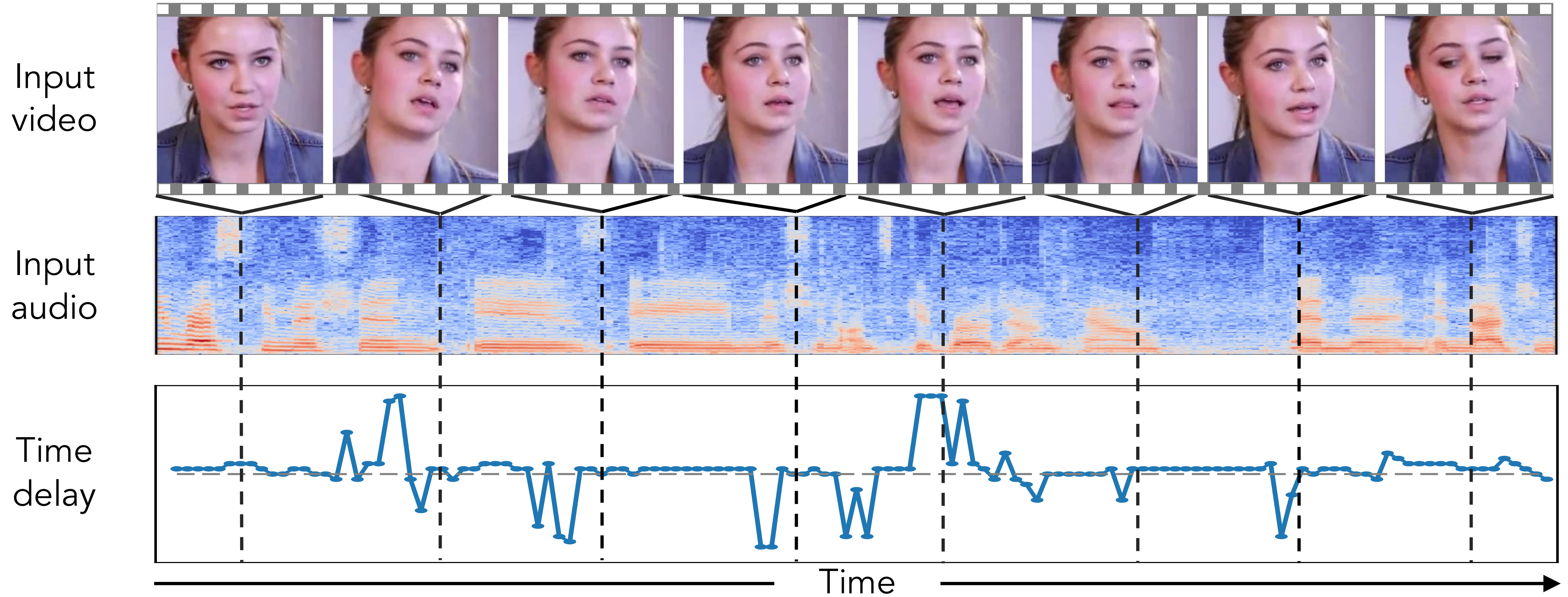
e.g., "deepfake" videos

Data representation



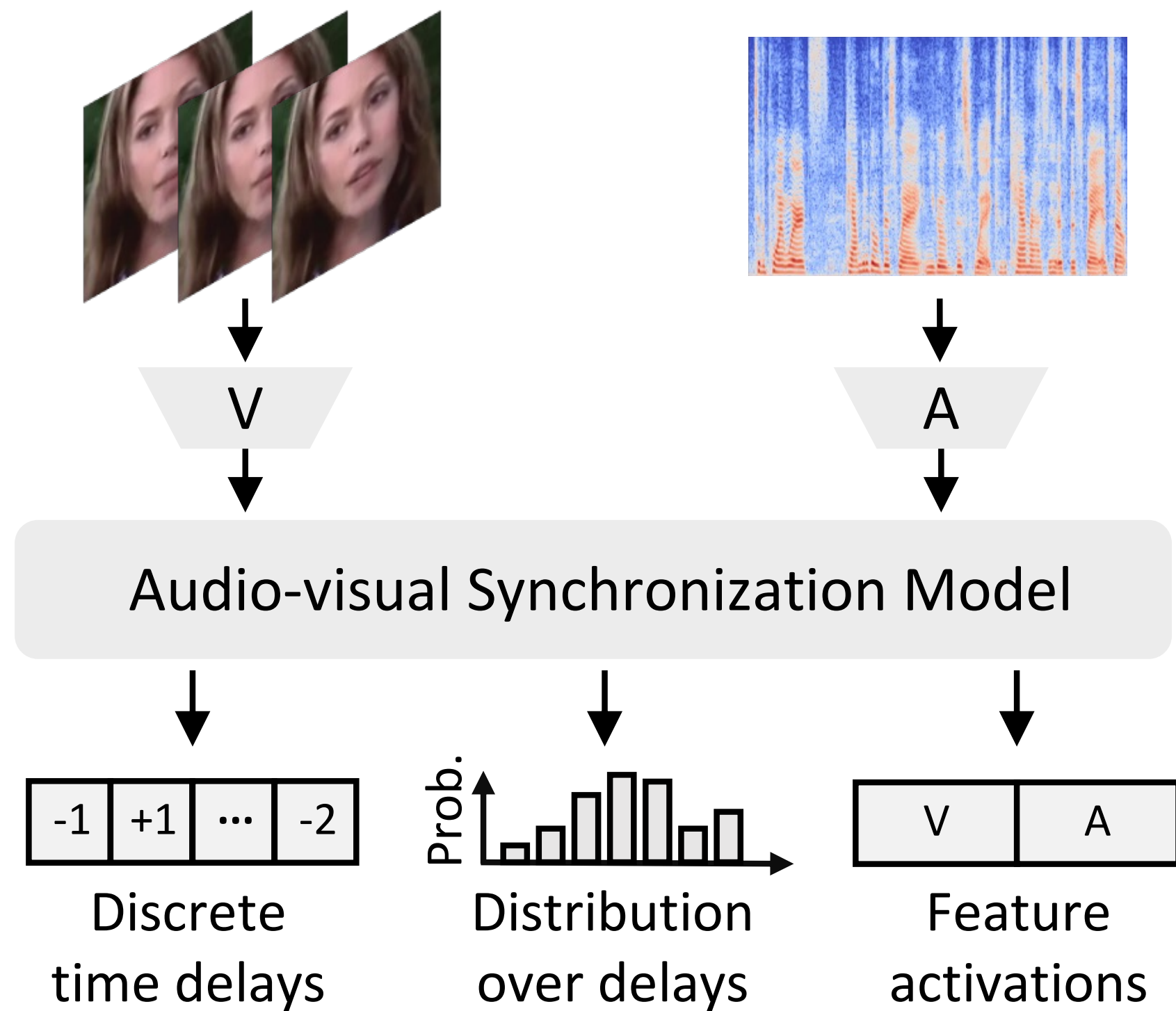
Raw pixels? Just as hard as generation!
Instead: self-supervised feature space.

Data representation



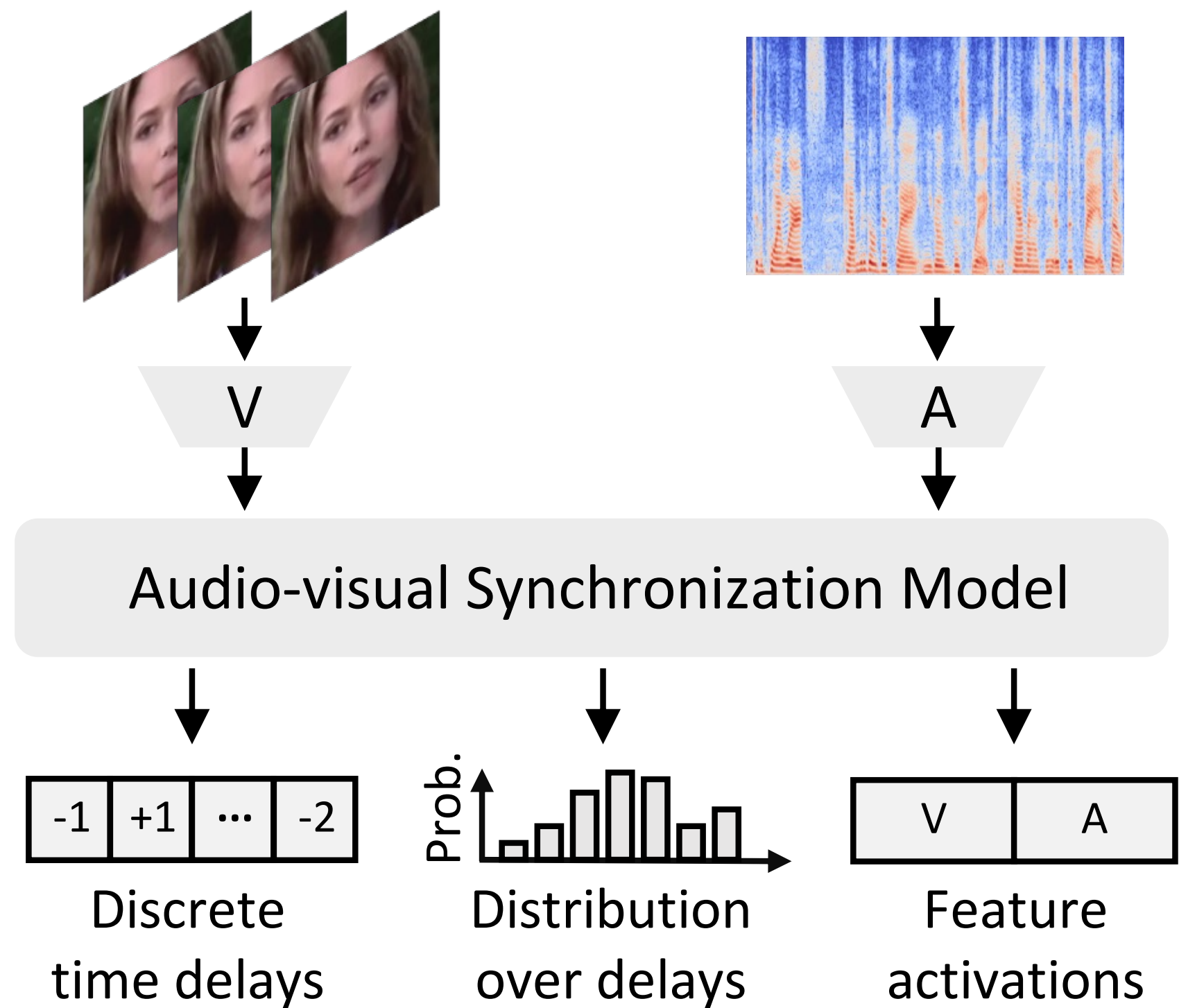
Learn the distribution for self-supervised features

Self-supervised feature learning

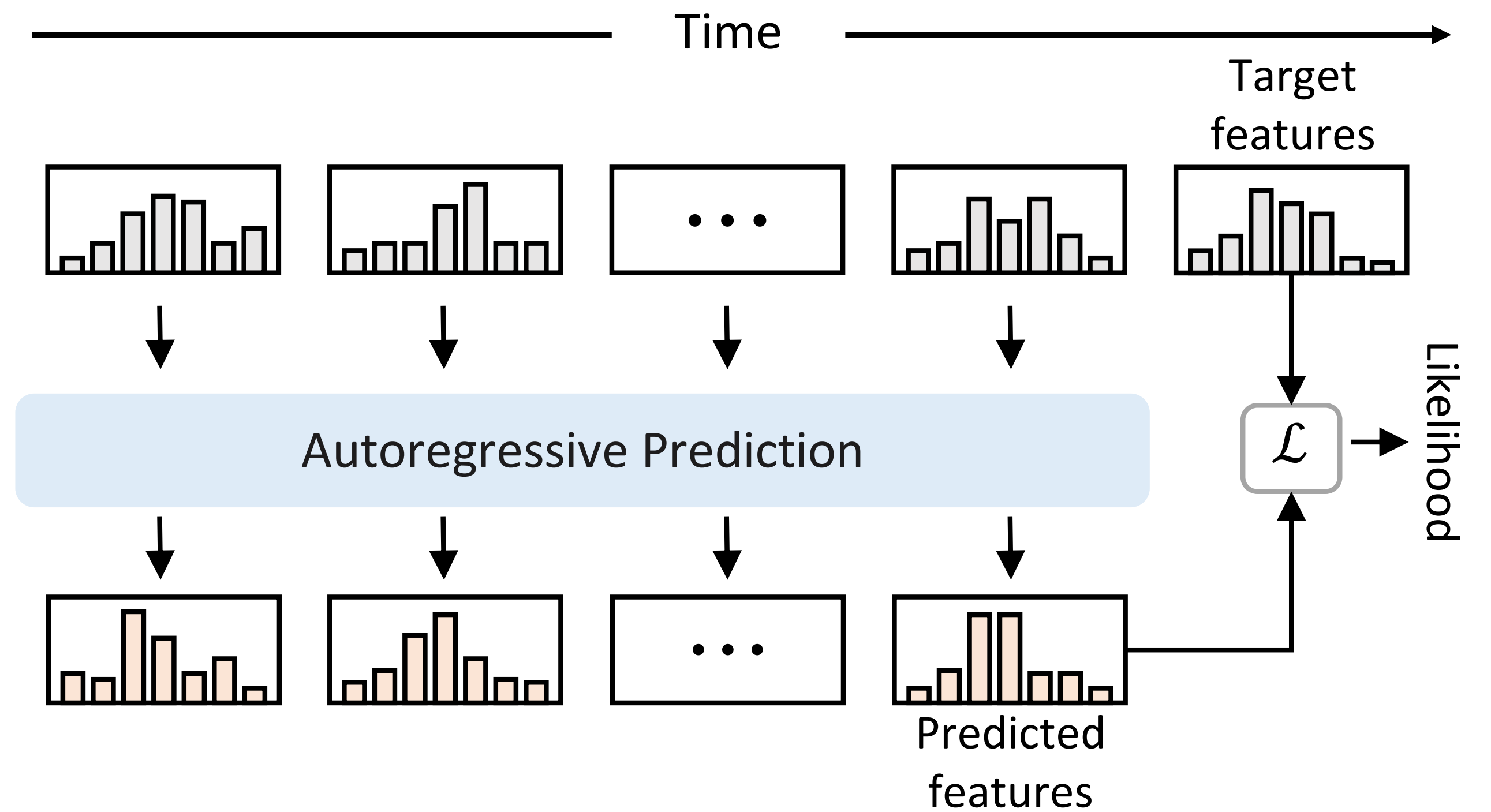


Learn the distribution for self-supervised features

Self-supervised feature learning



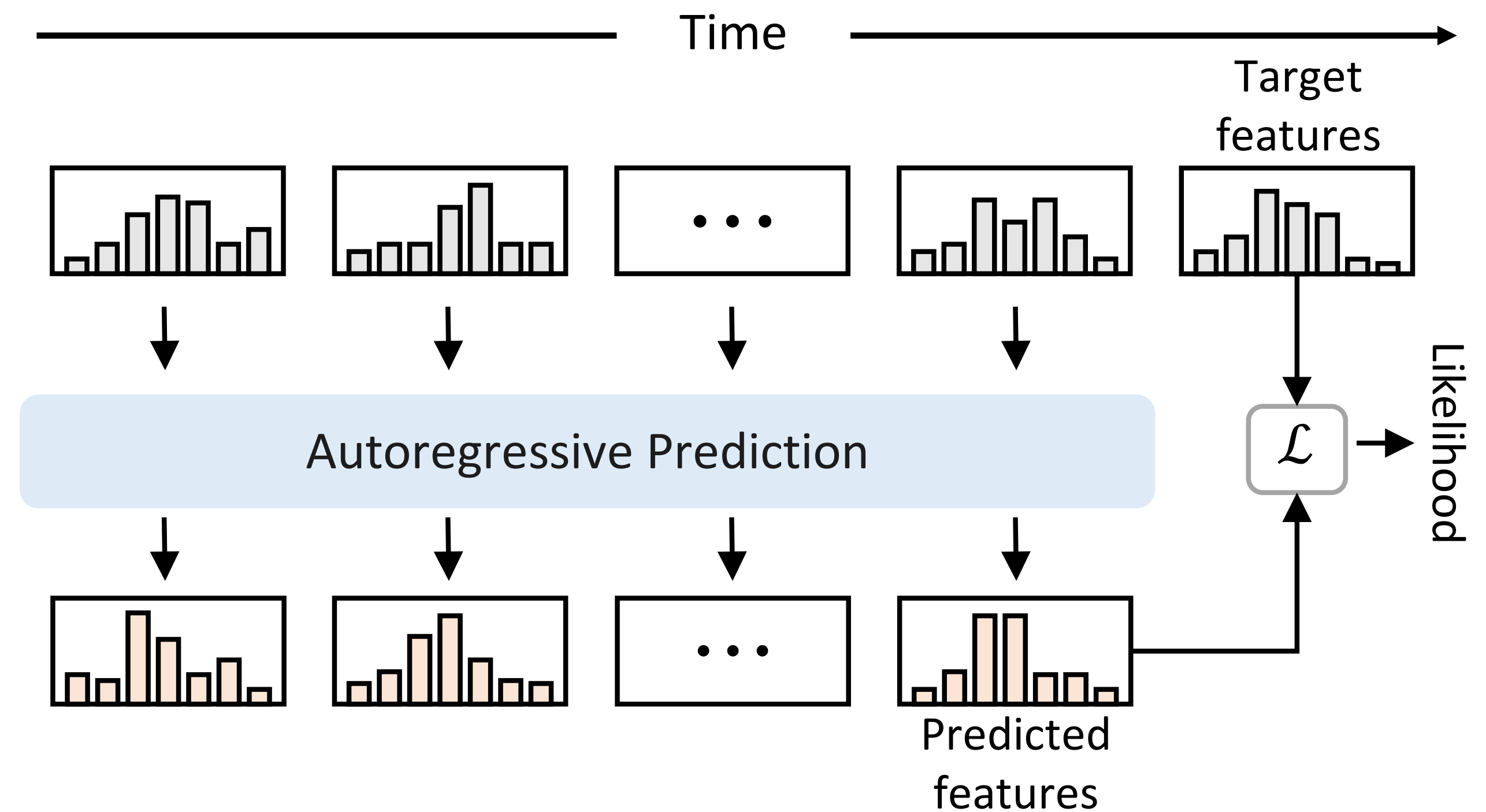
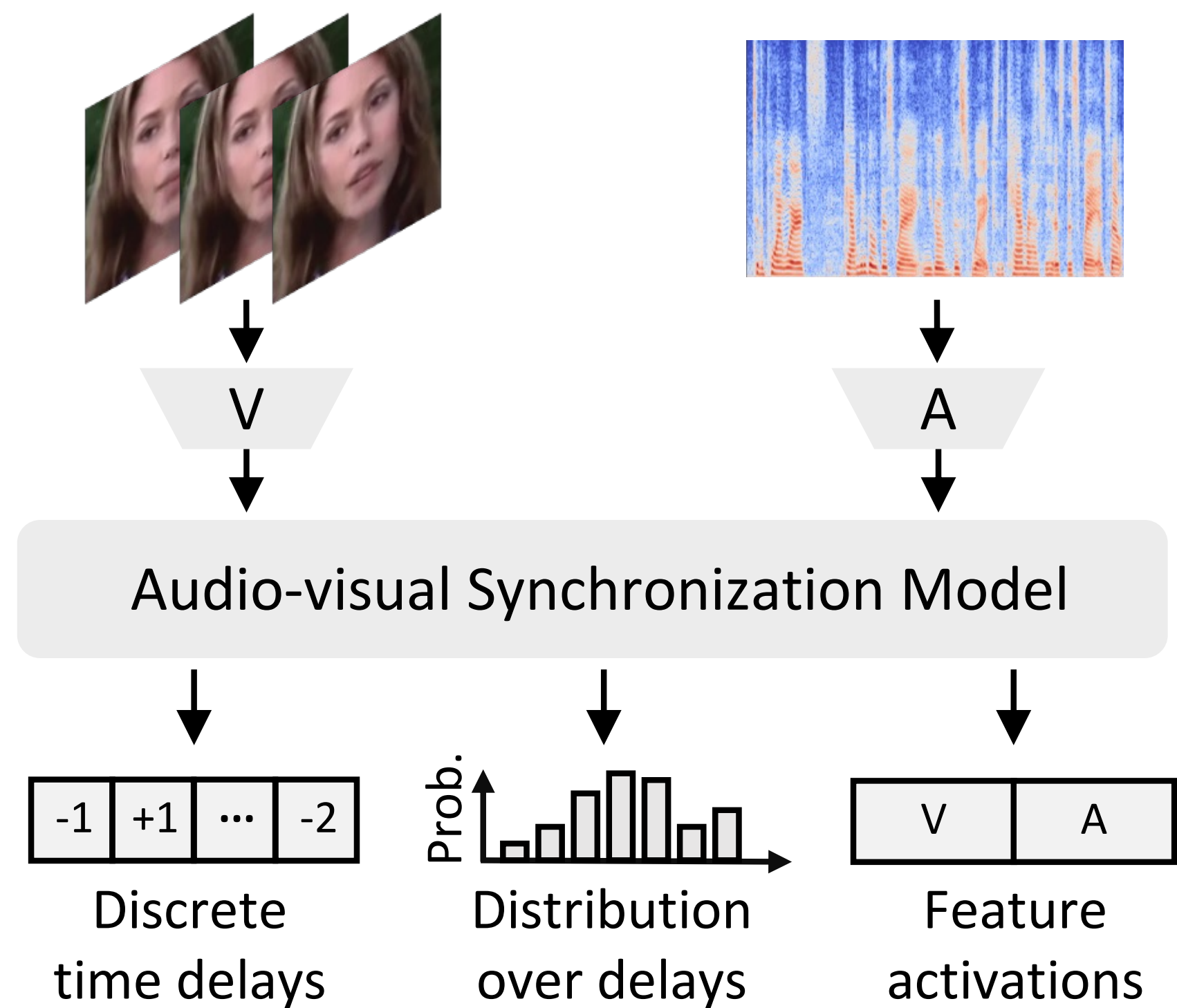
Audio-visual anomaly detection



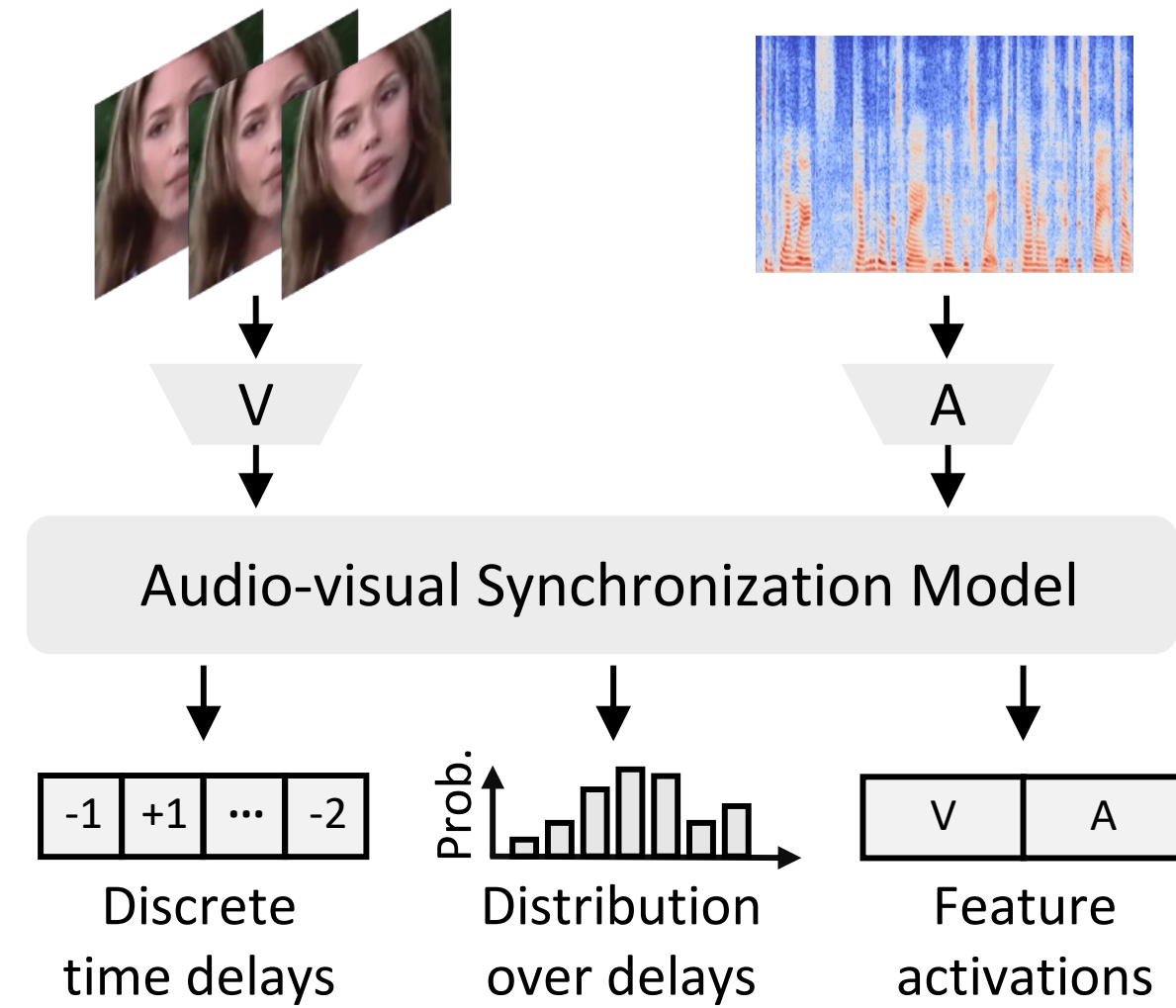
Learn the distribution for self-supervised features

Self-supervised feature learning

Audio-visual anomaly detection

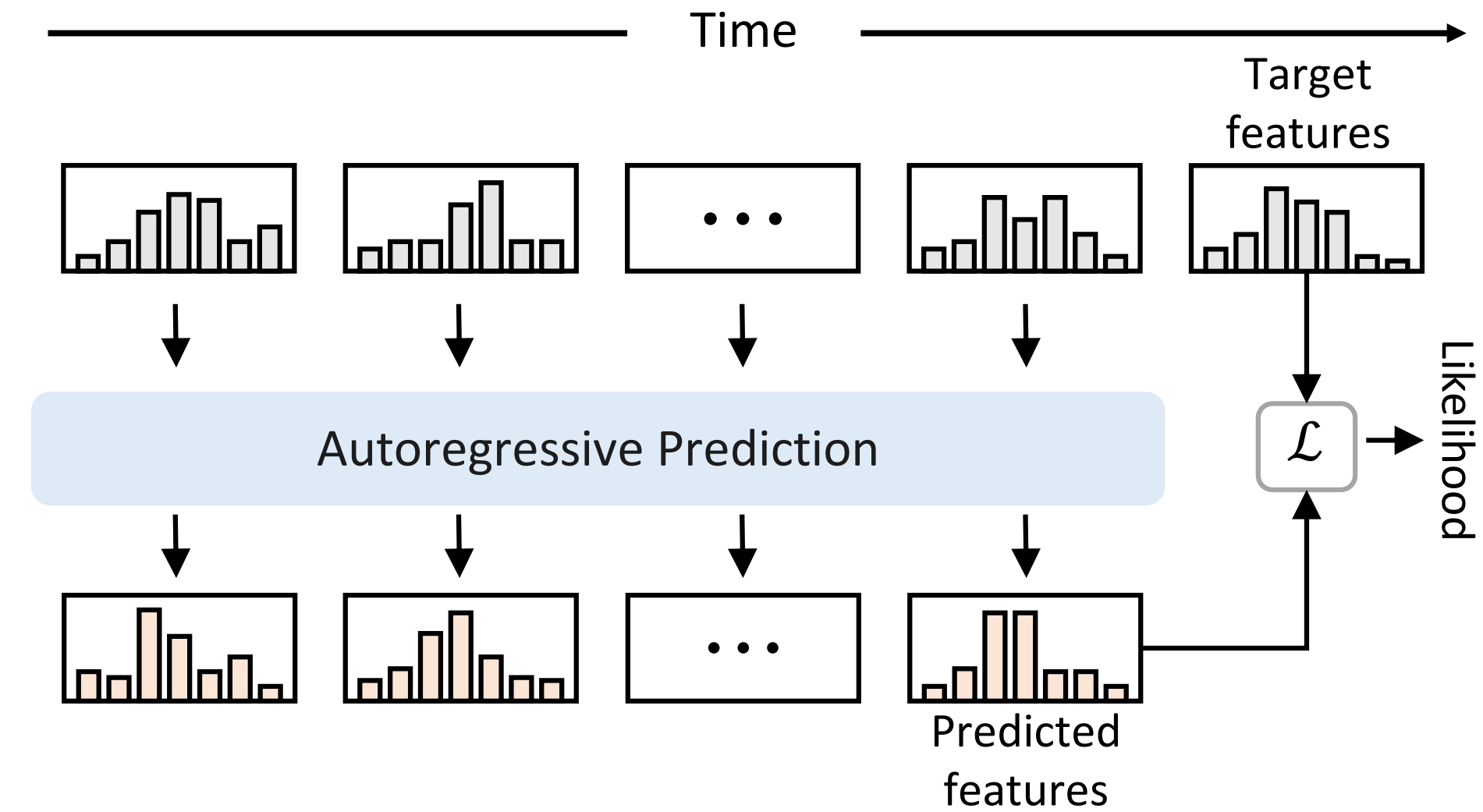


Learn the distribution for self-supervised features



Stage #1: Learning audio-visual synchronization feature sets:

$$S(i, j) = \frac{\exp(\phi(V_i, A_j))}{\sum_{k=i-\tau}^{i+\tau} \exp(\phi(V_i, A_k))}$$



Stage #2: Learning autoregressive model on self-supervised audio-visual feature sets:

$$p_{\theta}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) = \prod_{i=0}^{N-1} p_{\theta}(\mathbf{x}_{i+1} | \mathbf{x}_1, \dots, \mathbf{x}_i)$$

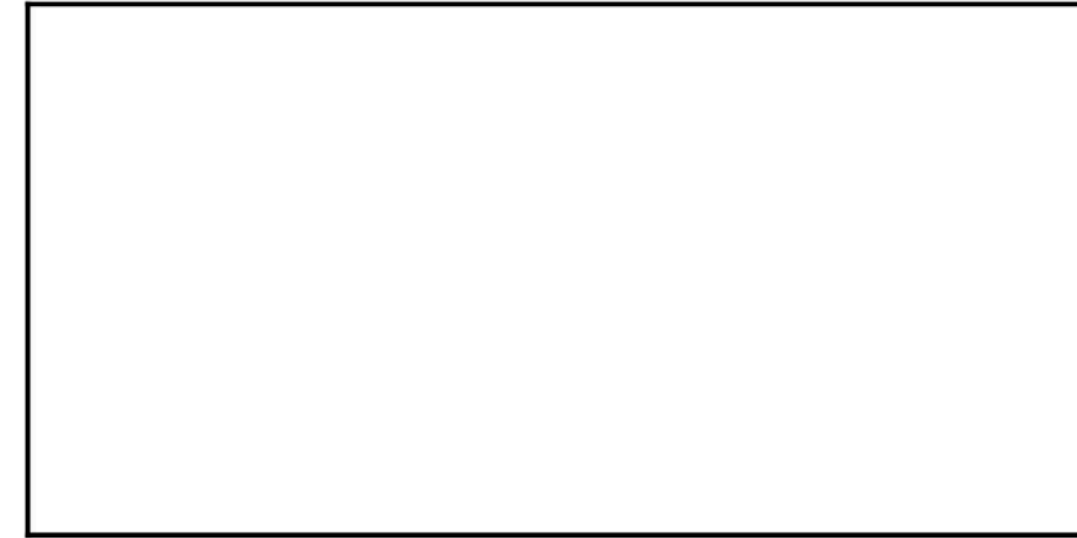
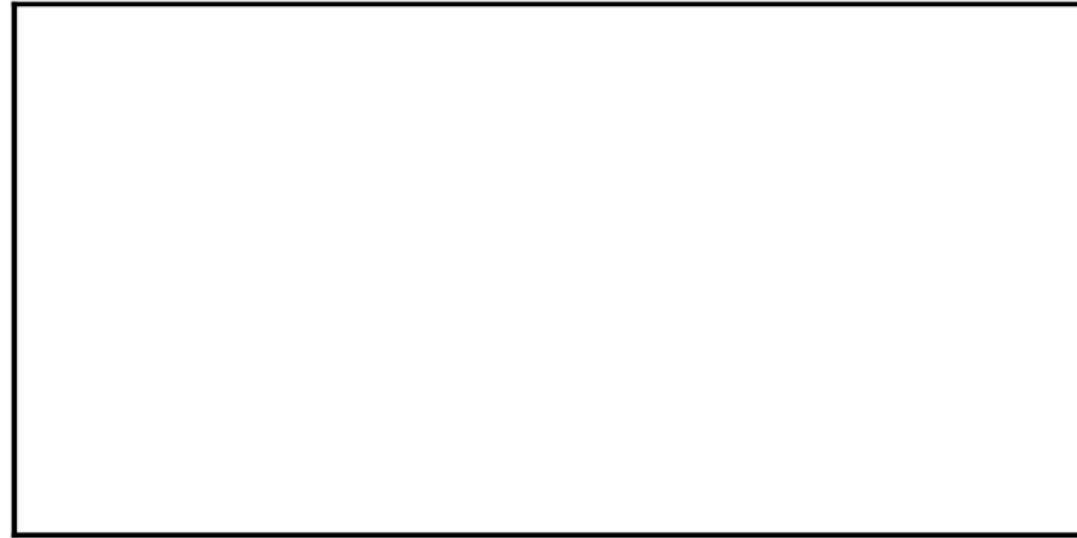
Real



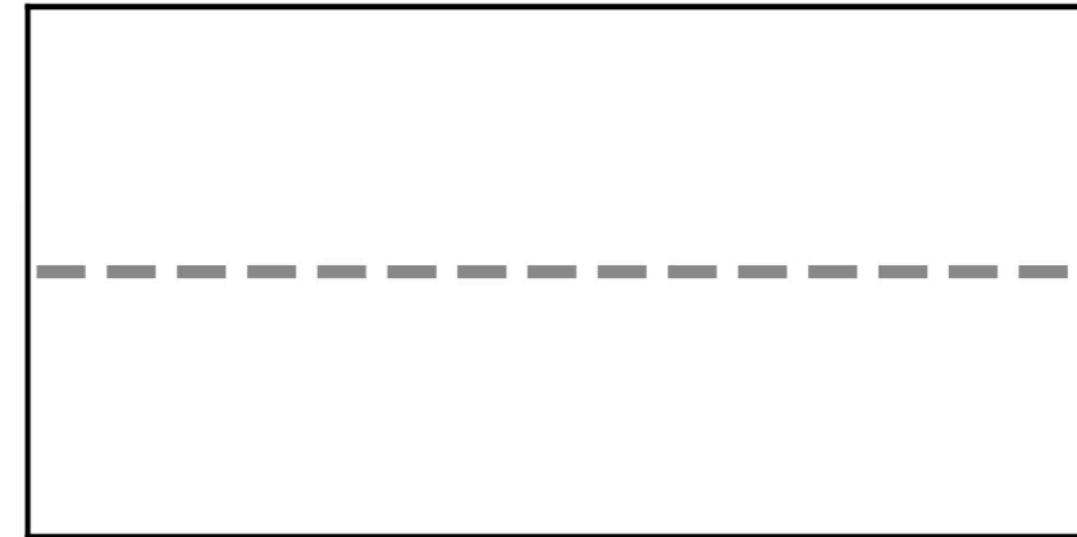
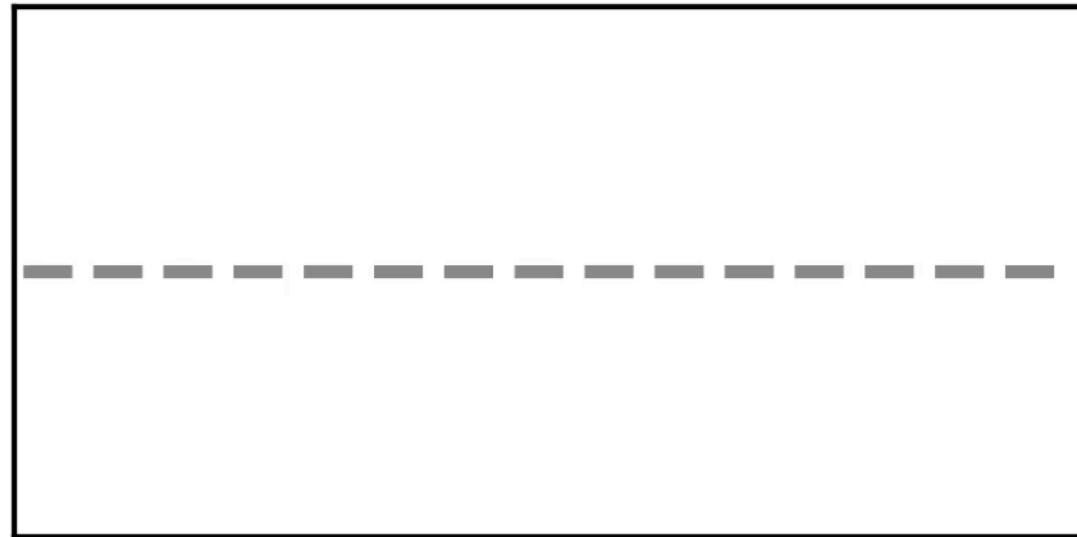
Fake



Sound



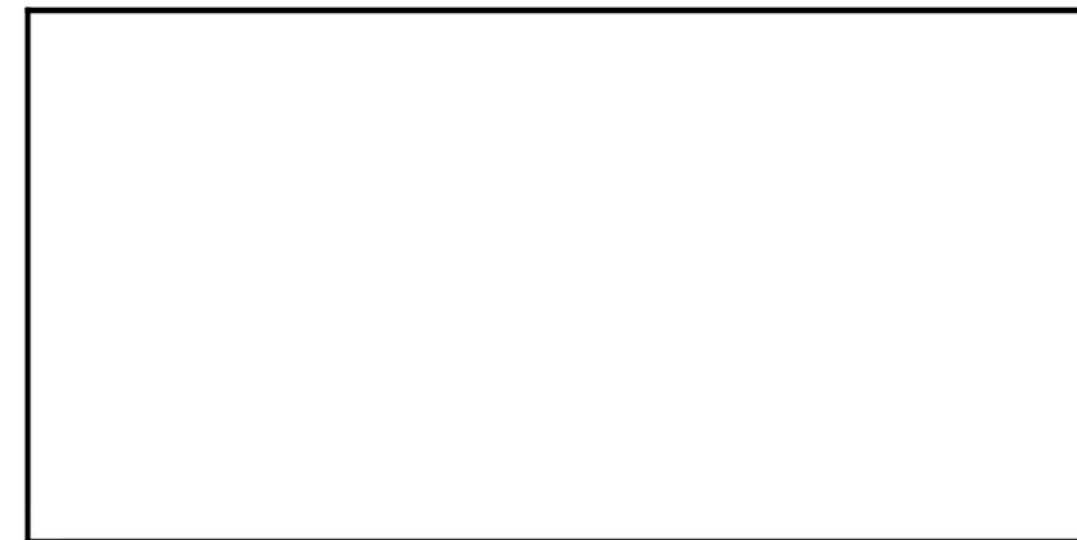
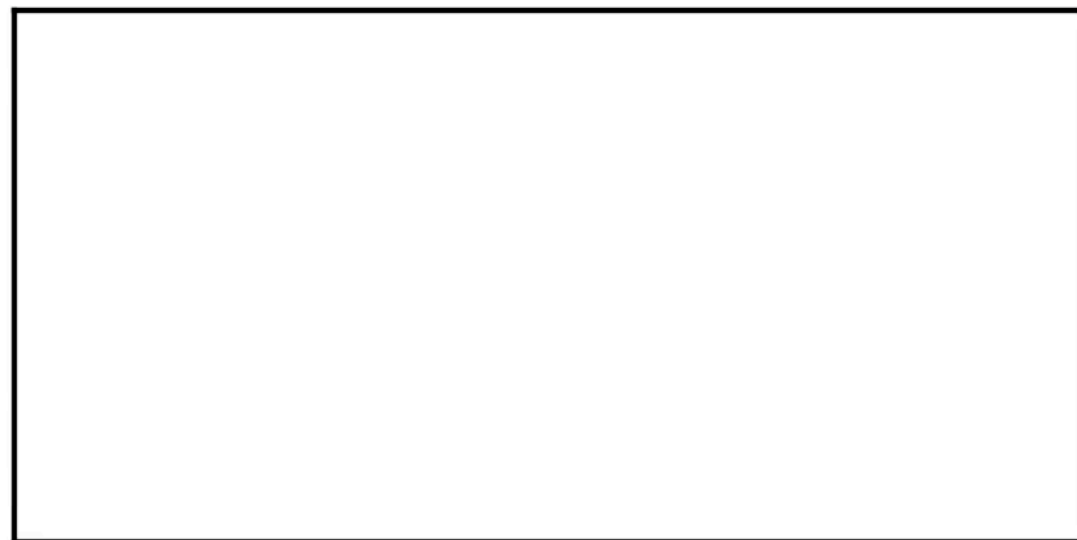
**Audio-visual
time delay**



+15

-15

**Anomaly
score**



460

230

0

0

Time

250

0

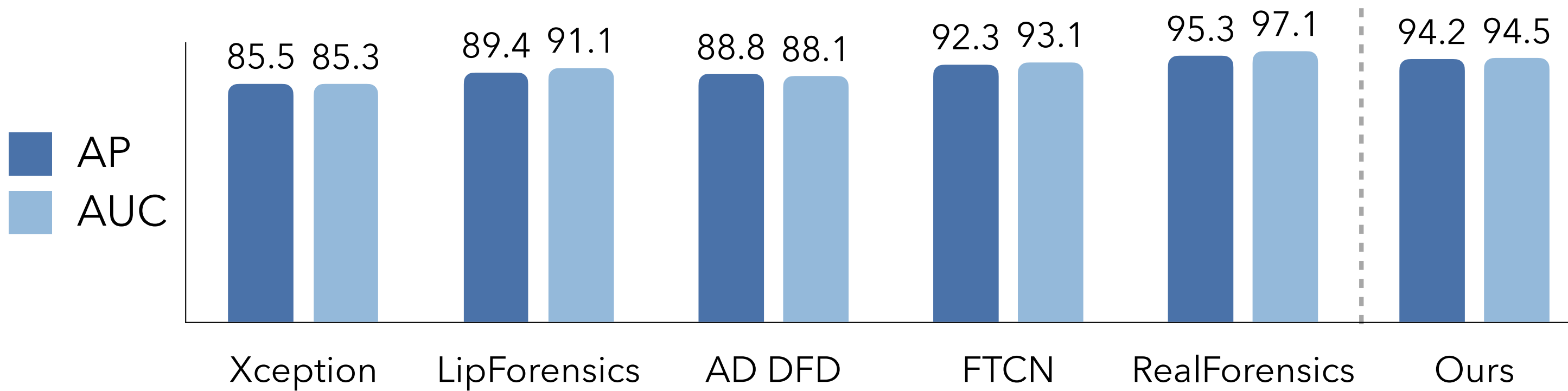
Time

250

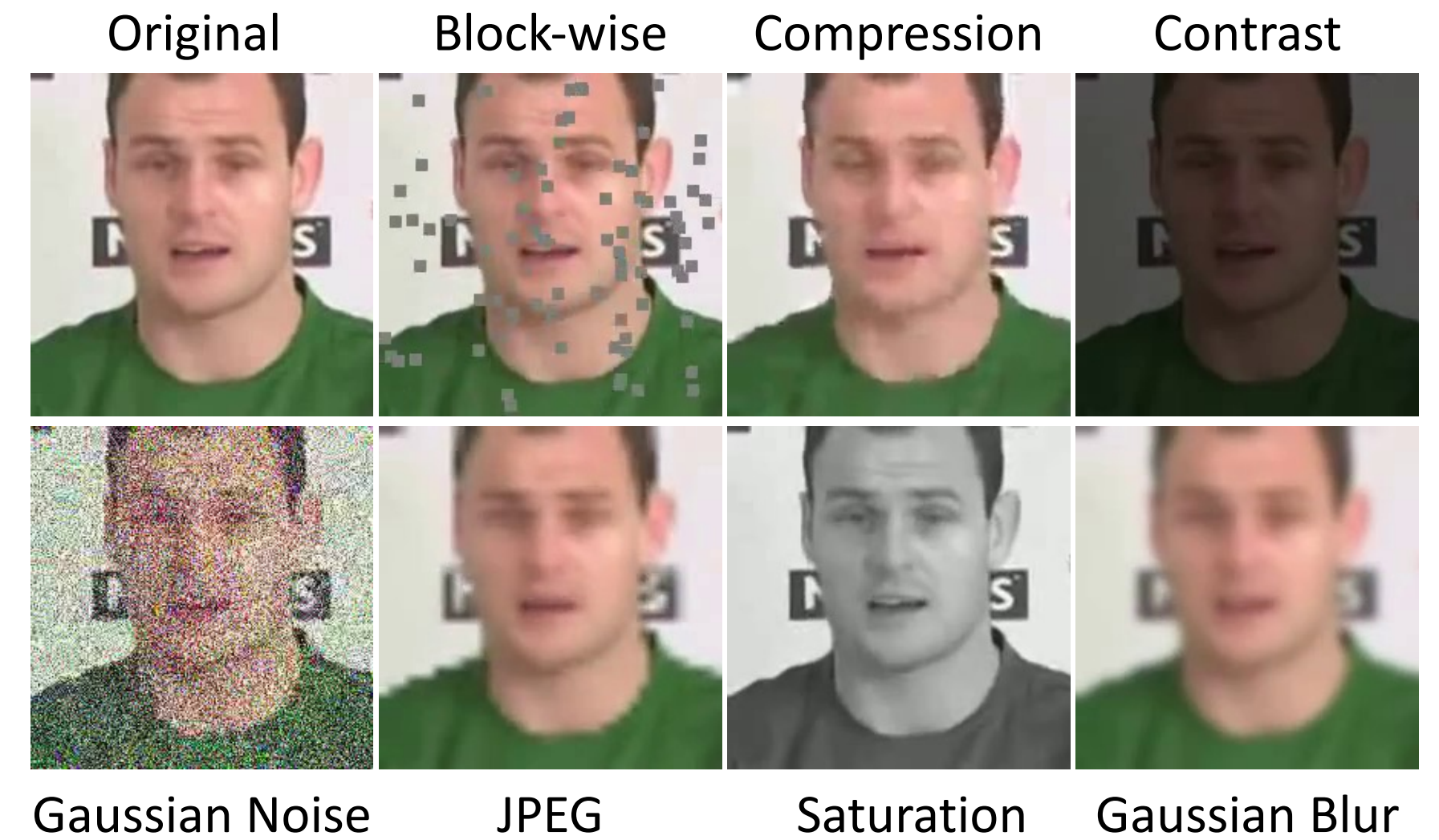
Results

FakeAVCeleb [Khalid et al., 2021]

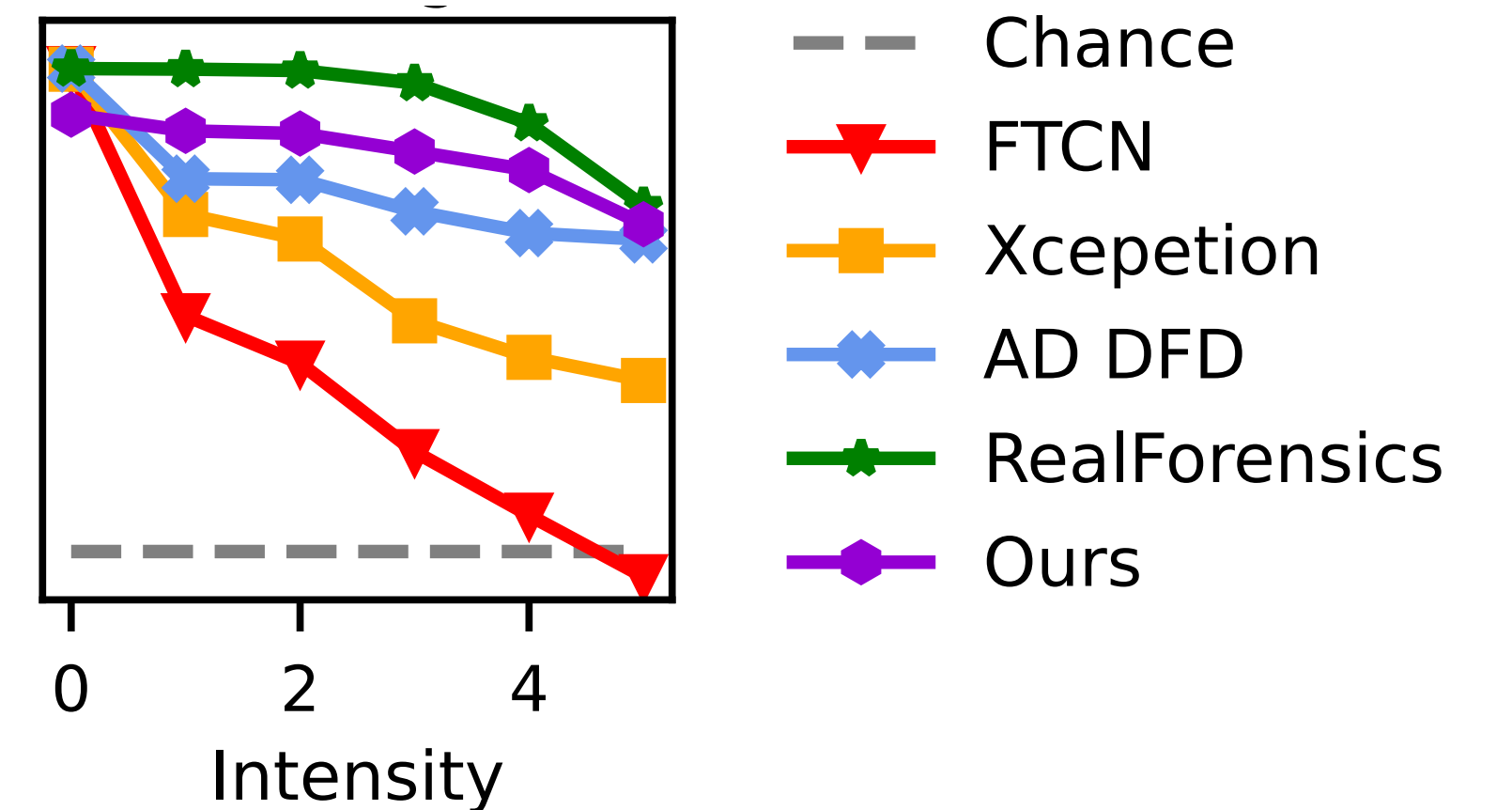
Supervised



Limitation: only works for out-of-sync lip motions (not face swaps)



Robustness to postprocessing

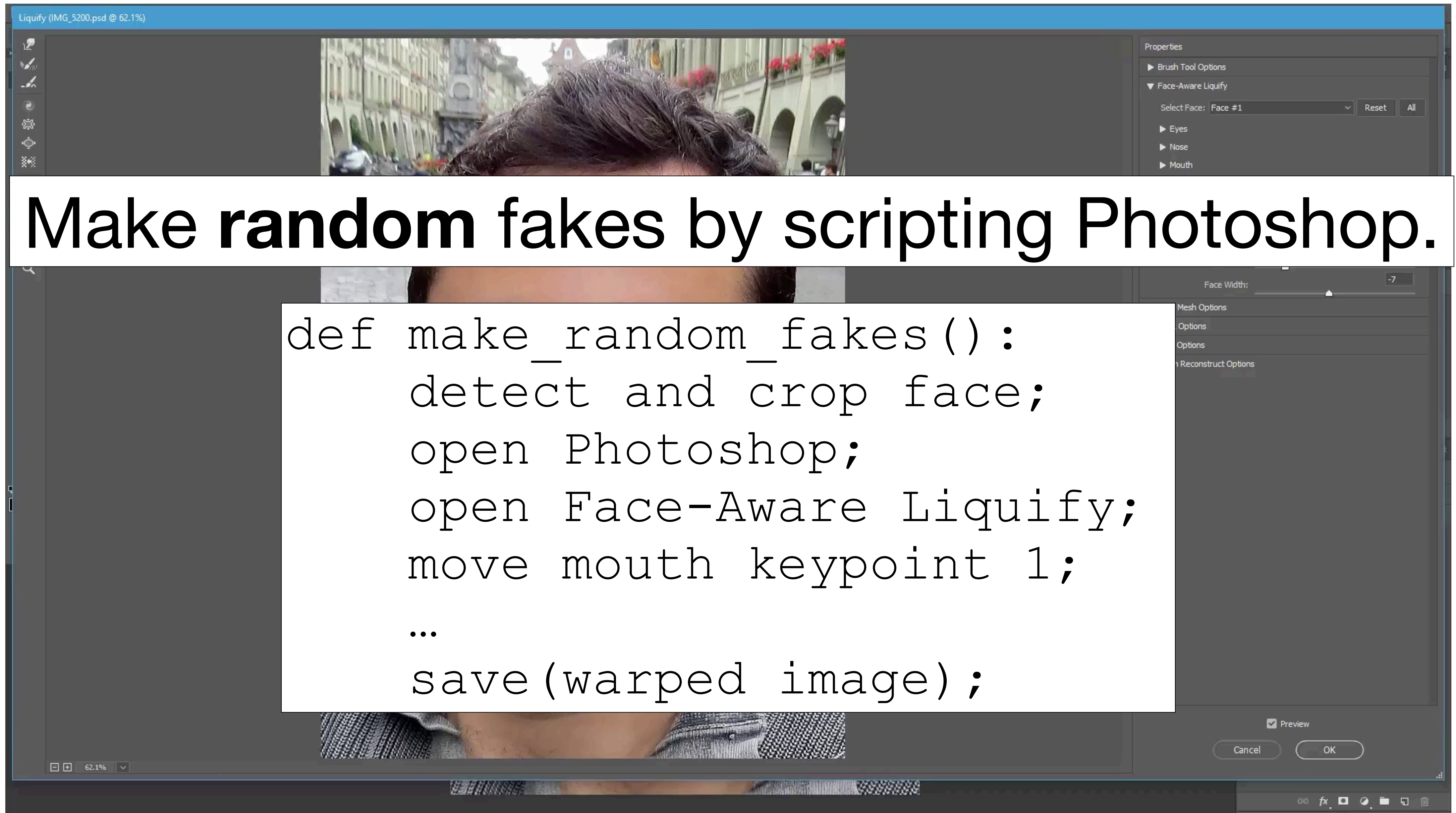


Strategy #4: supervised learning



[Wang et al., “Image Splice Detection via Learned Self-Consistency”, 2018]



The image shows a screenshot of the Adobe Photoshop interface, specifically the Face-Aware Liquify tool. The main window displays a photograph of a person's face. The right-hand side of the interface shows the 'Properties' panel with 'Face-Aware Liquify' selected. Below the main image, a white box contains a Python script. The script defines a function 'make_random_fakes()' which performs several steps: detecting and cropping the face, opening Photoshop, opening the Face-Aware Liquify tool, moving the mouth keypoint by 1 unit, and finally saving the warped image. The Photoshop interface also shows a 'Face Width' slider set to -7 and a 'Preview' checkbox checked. The bottom of the interface has 'Cancel' and 'OK' buttons.

Liquify (IMG_5200.psd @ 62.1%)

Make random fakes by scripting Photoshop.

```
def make_random_fakes():  
    detect and crop face;  
    open Photoshop;  
    open Face-Aware Liquify;  
    move mouth keypoint 1;  
    ...  
    save(warped image);
```

Face Width: -7

Mesh Options

Options

Options

Reconstruct Options

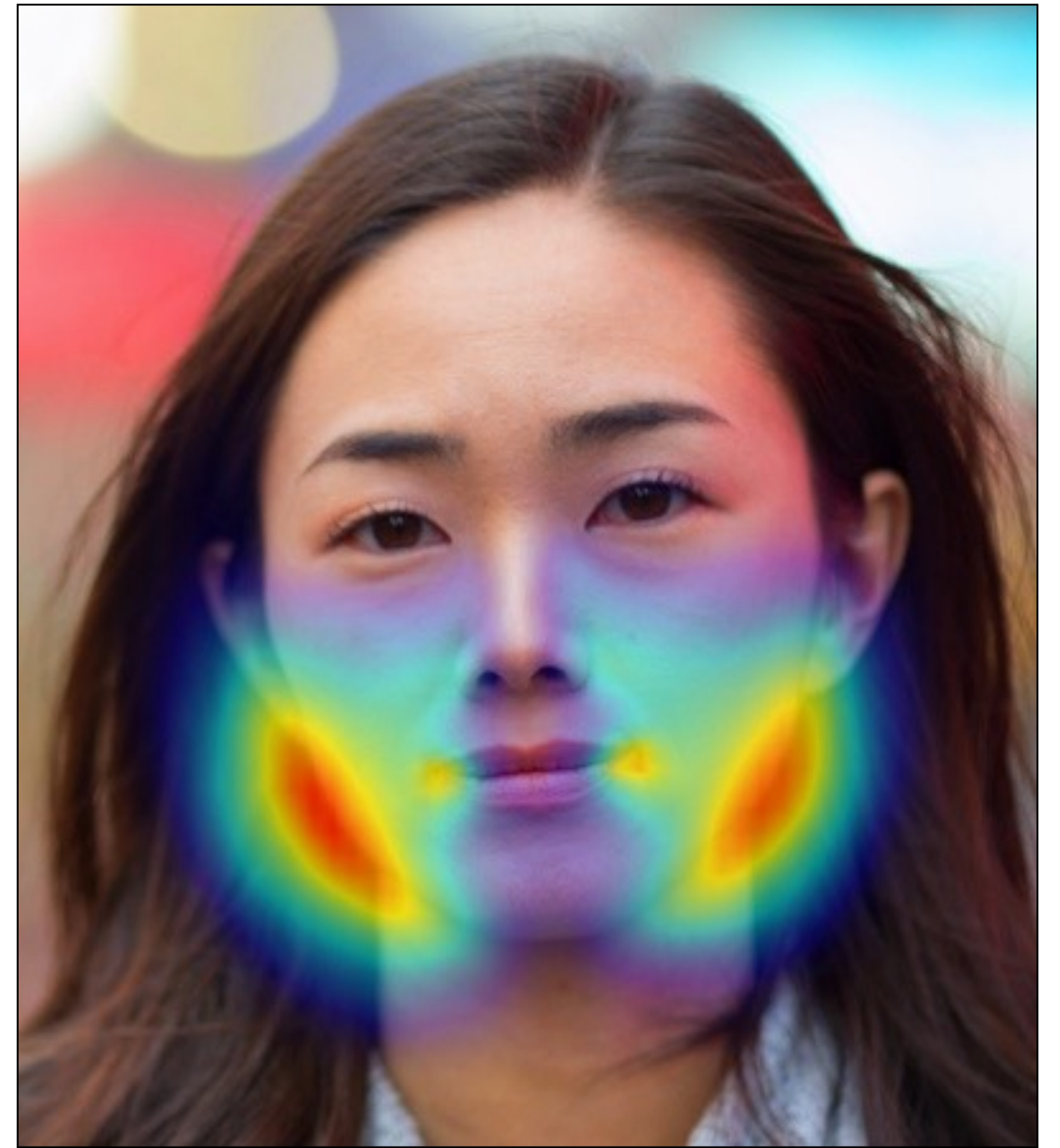
Preview

Cancel OK

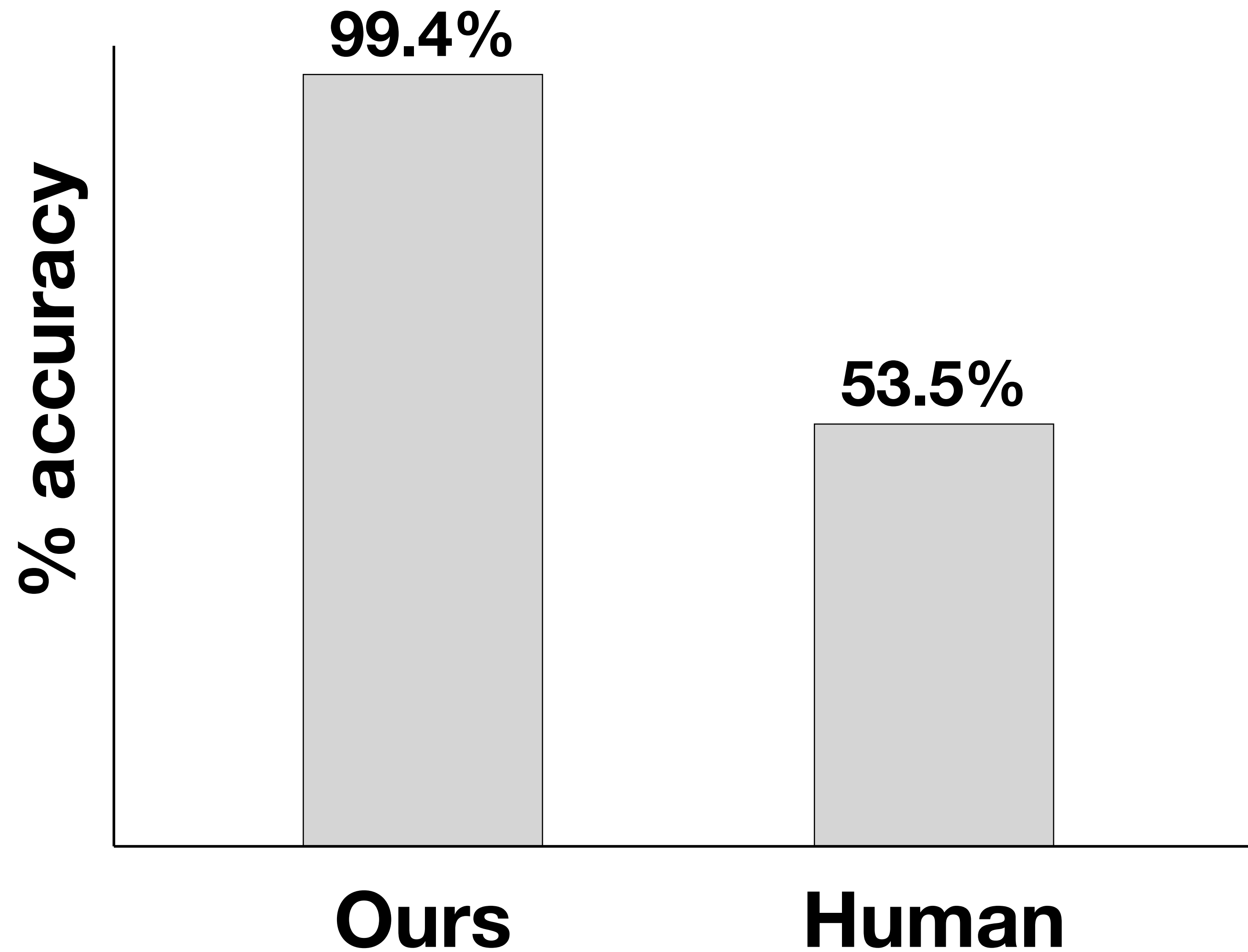
Photoshop Face-Aware Liquify tutorial. Source: https://youtu.be/5Qqv_C6iVvQ?t=86



Warp
detector



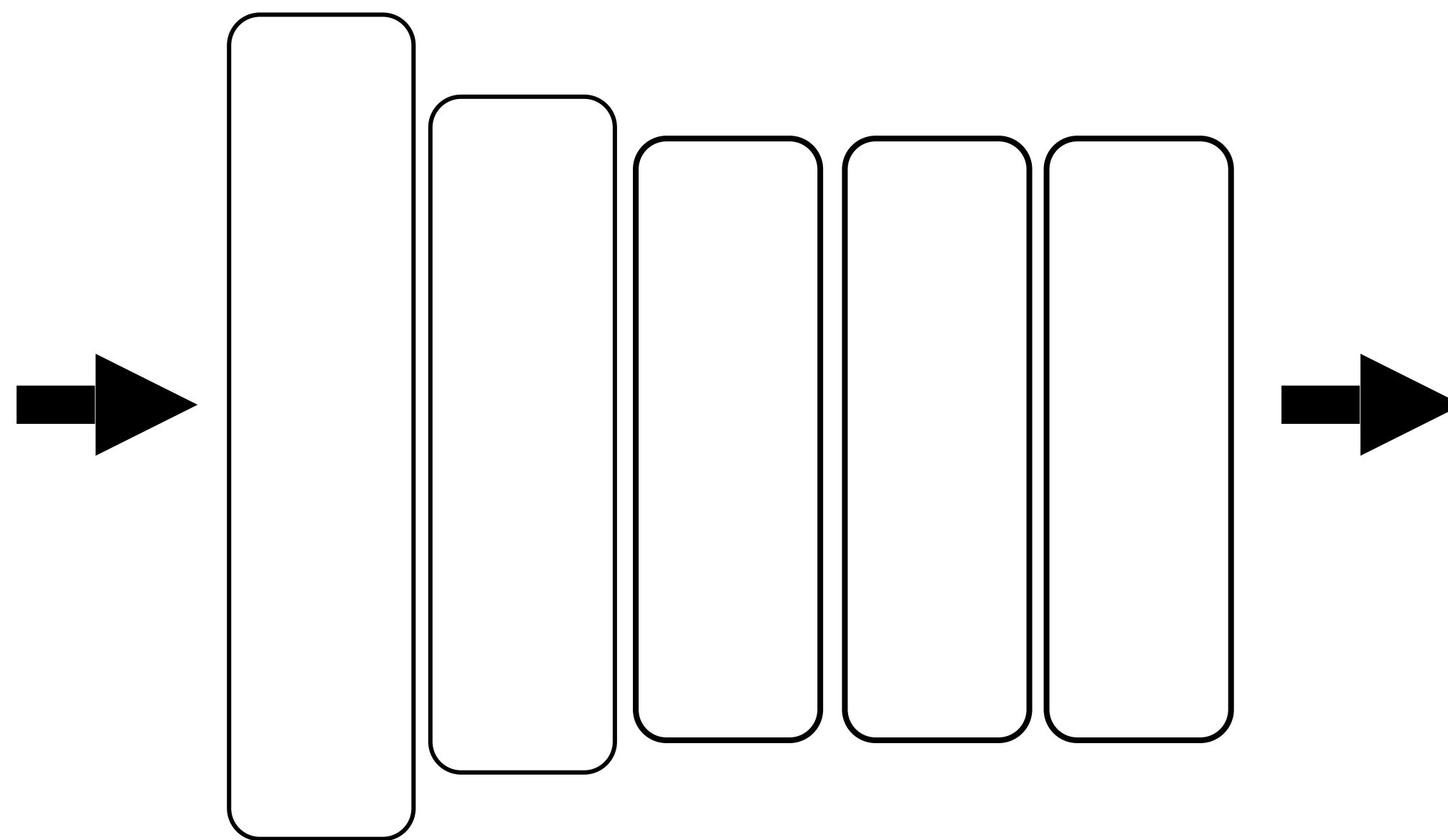
Real-or-fake classification



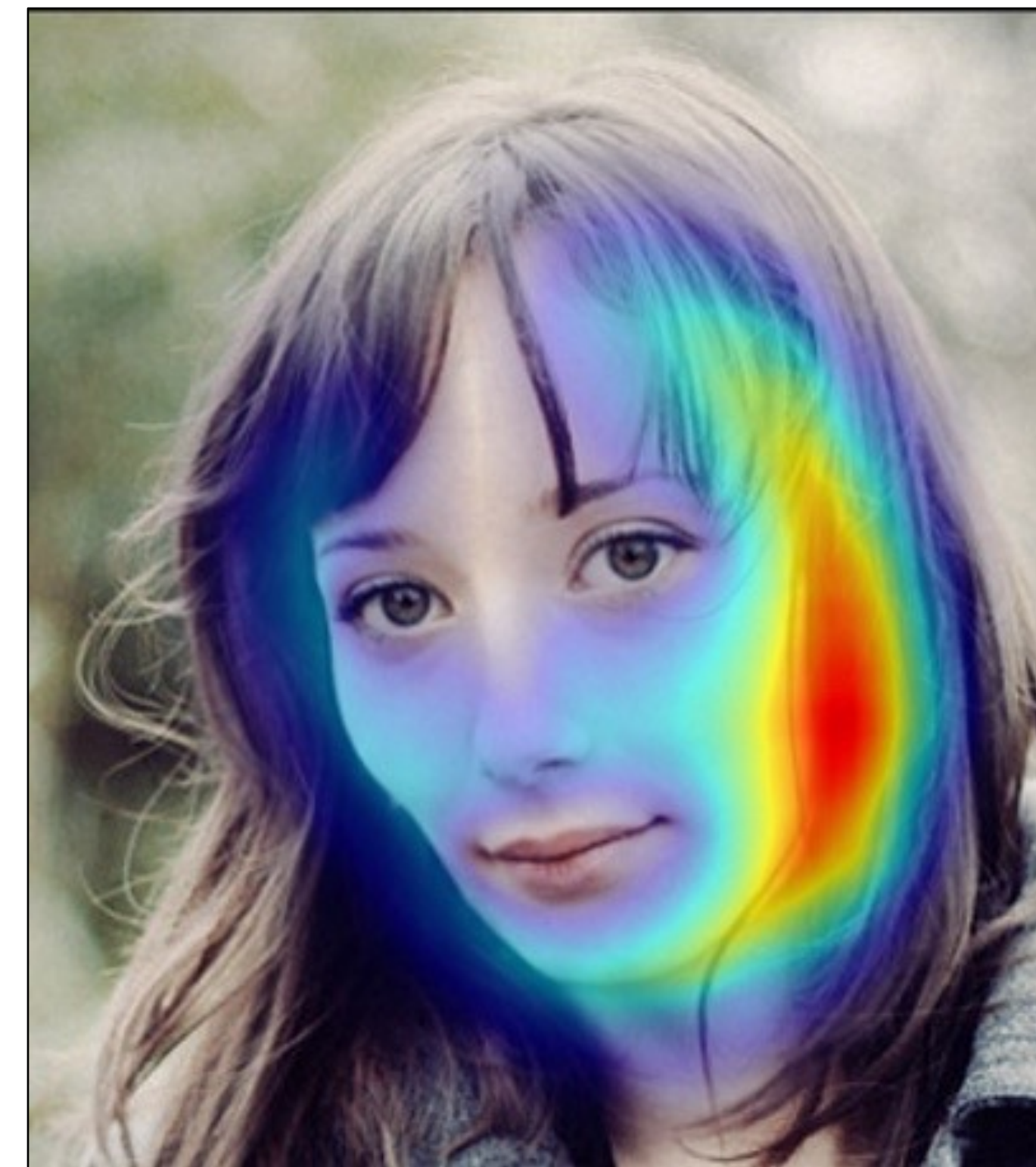
What moved where?



Manipulated Image



Dilated ResNet



Warp Prediction

What moved where?



Modified

Original

Optical Flow

Modified

Original

Optical Flow



Manipulated Photo



Flow Prediction



Suggested “Undo”



Original Photo



Manipulated vs. Original



Undo vs. Original



Manipulated Photo



Warp Prediction



Suggested “Undo”



Original Photo

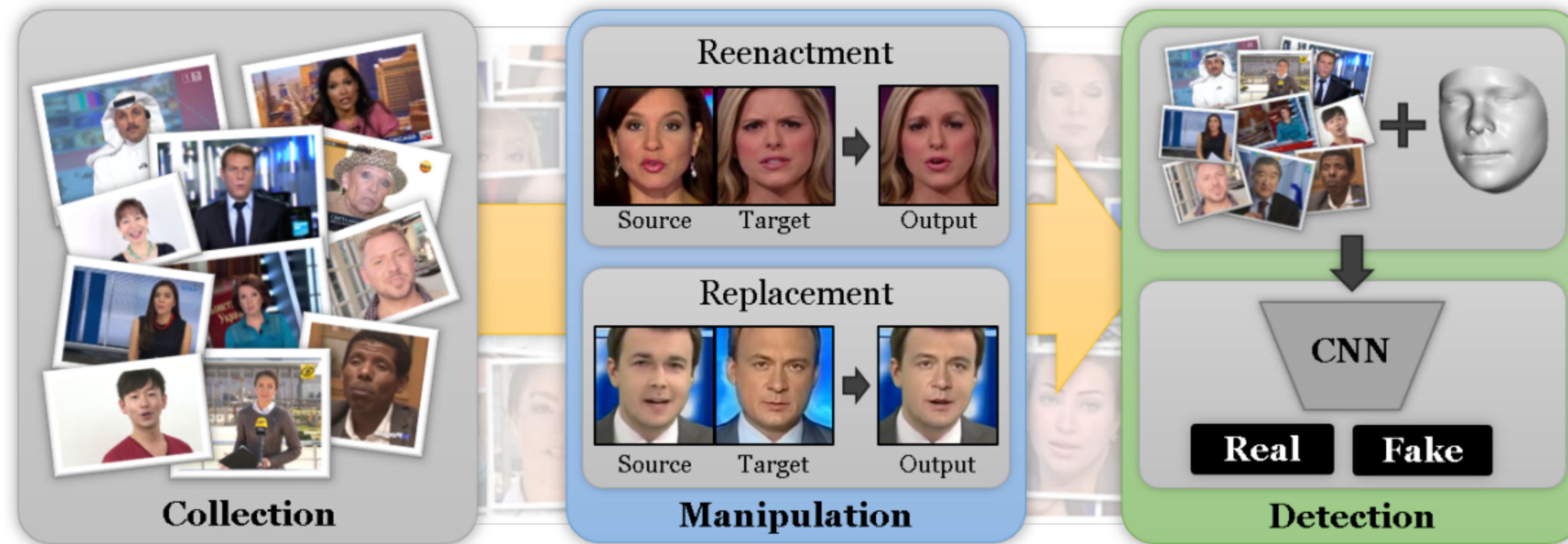


Suggested “Undo”



Manipulated Photo

Similar approaches for “deepfakes”



Create lots of deepfake videos, then learn to detect them.

[Rossler et al., “FaceForensics++”, 2019]

New challenges on the horizon

Celeb-DF: A New Dataset for DeepFake Forensics

Yuezun Li¹, Xin Yang¹, Pu Sun², Honggang Qi² and Siwei Lyu¹

¹University at Albany, State University of New York, USA

²University of Chinese Academy of Sciences, China

[Li et al., “Celeb-DF”, 2020]

The forensics generalization problem

New architectures & datasets



StyleGAN2 [Karras 2019]

New models



Cascaded refinement networks [Chen & Koltun 2017]

Lots of potential issues for “universal” detector:
dataset bias, domain adaptation, etc.

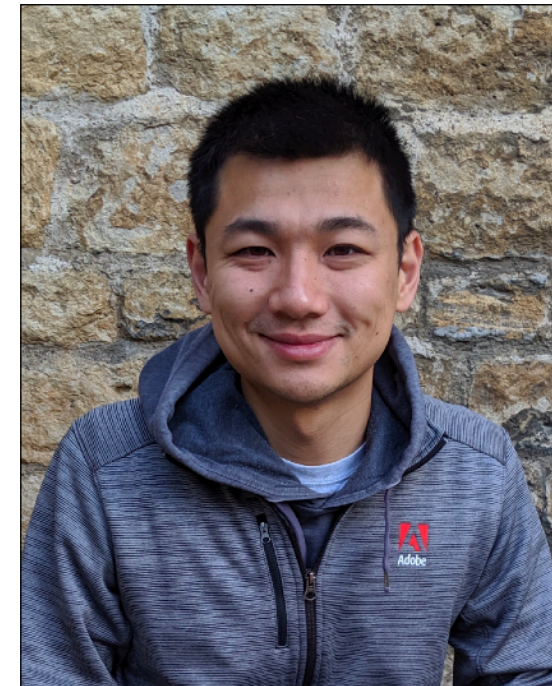
CNN-generated images are surprisingly easy to spot... for now



Sheng-Yu Wang



Oliver Wang



Richard Zhang

Andrew Owens



Alexei Efros

<https://peterwang512.github.io/CNNDetection>

Dataset of CNN-generated fakes

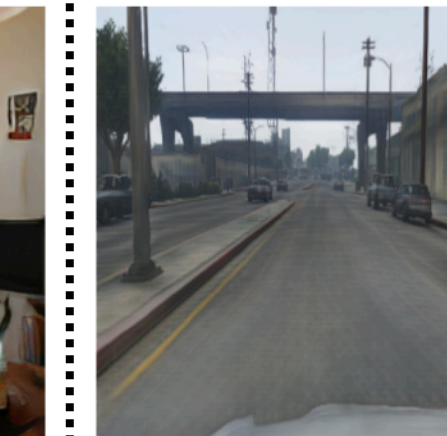
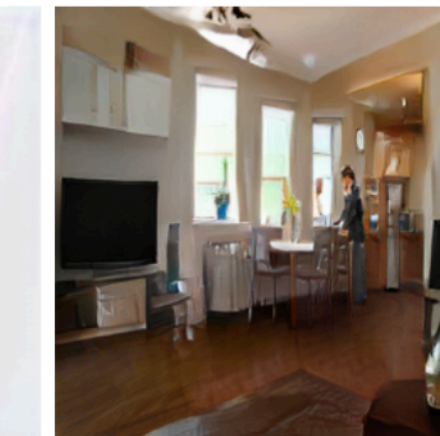
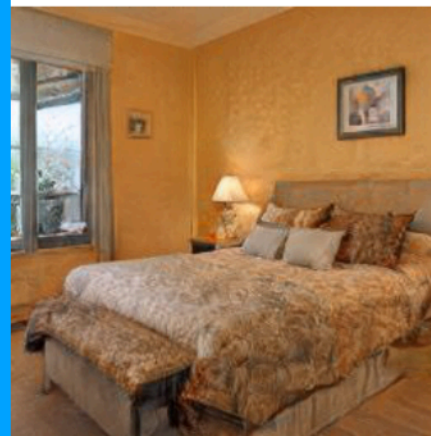
GANs

Perceptual loss

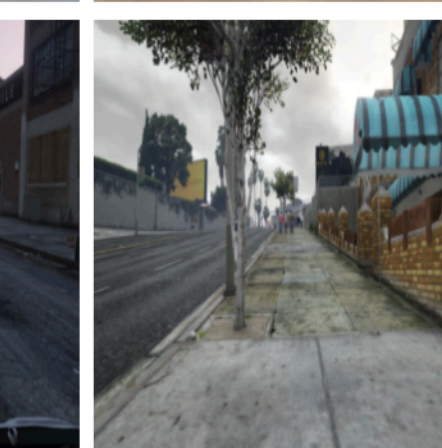
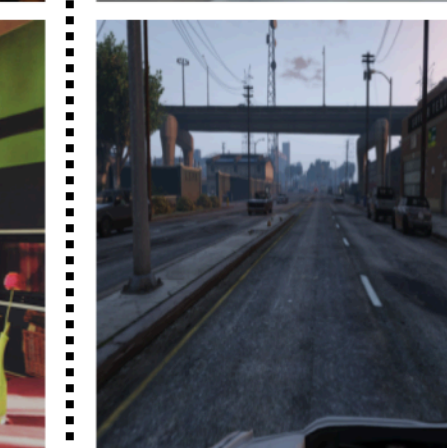
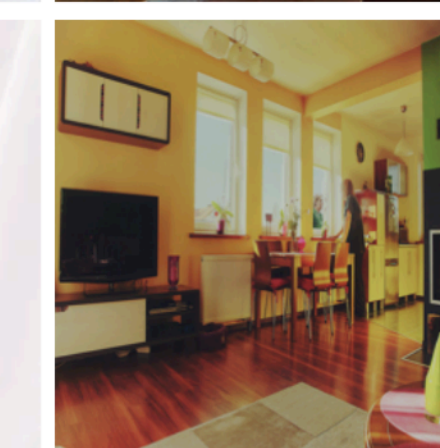
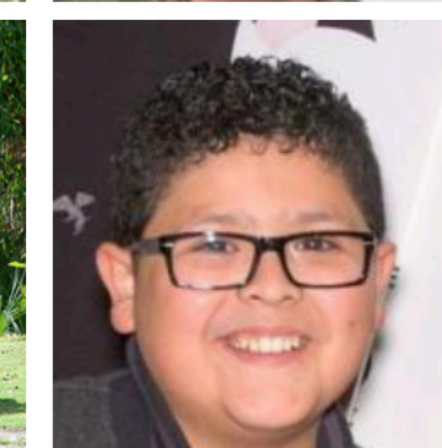
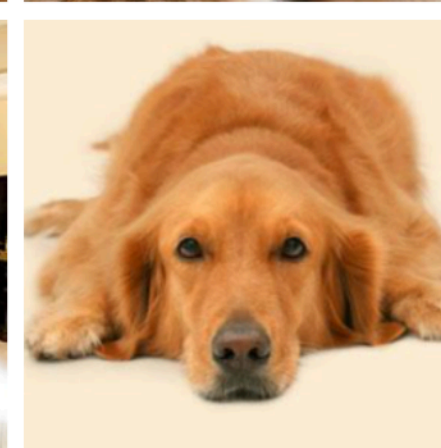
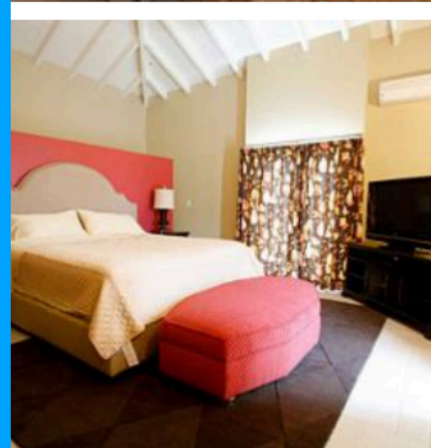
Low-level vision

Deep fakes

fake



real



ProGAN
(Karras 2018)

StyleGAN
(Karras 2018)

BigGAN
(Brock 2019)

CycleGAN
(Zhu 2017)

StarGAN
(Choi 2018)

GauGAN
(Park 2019)

Cascaded refinement
(Chen 2017)

IMLE
(Li 2019)

Seeing in the dark
(Chen 2018)

Super-resolution
(Dai 2019)

Faceswap
(Anonymous 2018)
(Rossler 2019)

Dataset of CNN-generated fakes

fake



GANs



StyleGAN
(Karras 2018)

BigGAN
(Brock 2019)

CycleGAN
(Zhu 2017)

StarGAN
(Choi 2018)

GauGAN
(Park 2019)

Perceptual loss



Cascaded refinement
(Chen 2017)

IMLE
(Li 2019)

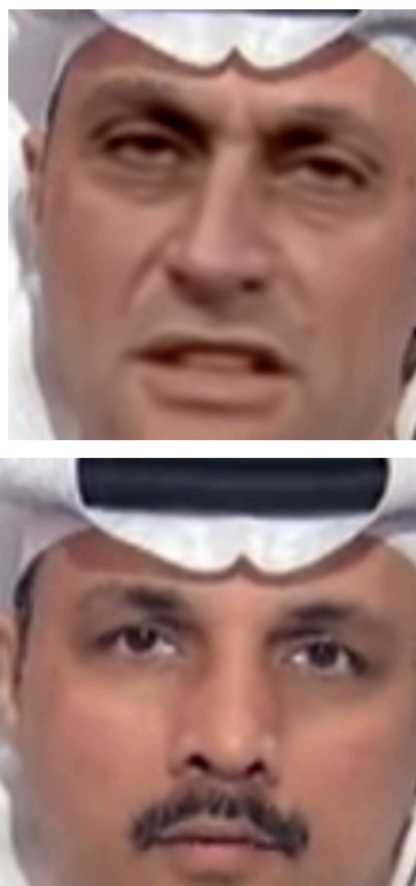
Low-level vision



Seeing in the dark
(Chen 2018)

Super-resolution
(Dai 2019)

Deep fakes



Faceswap
(Anonymous 2018)
(Rossler 2019)

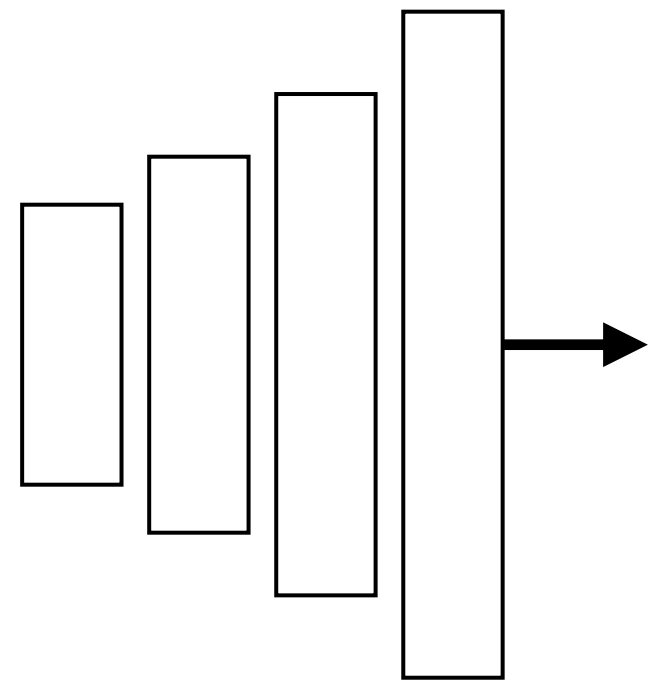
real



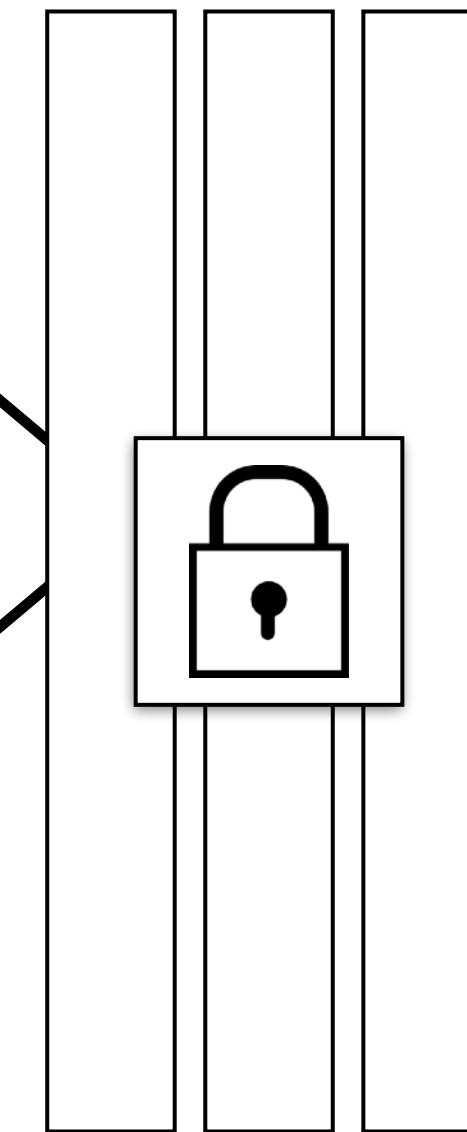
ProGAN
(Karras 2018)

How well do classifiers generalize?

ProGAN



Real images



Real vs. fake?

- Train with 720K images from 20 LSUN categories
- JPEG + Blurring data augmentation

How well do classifiers generalize?

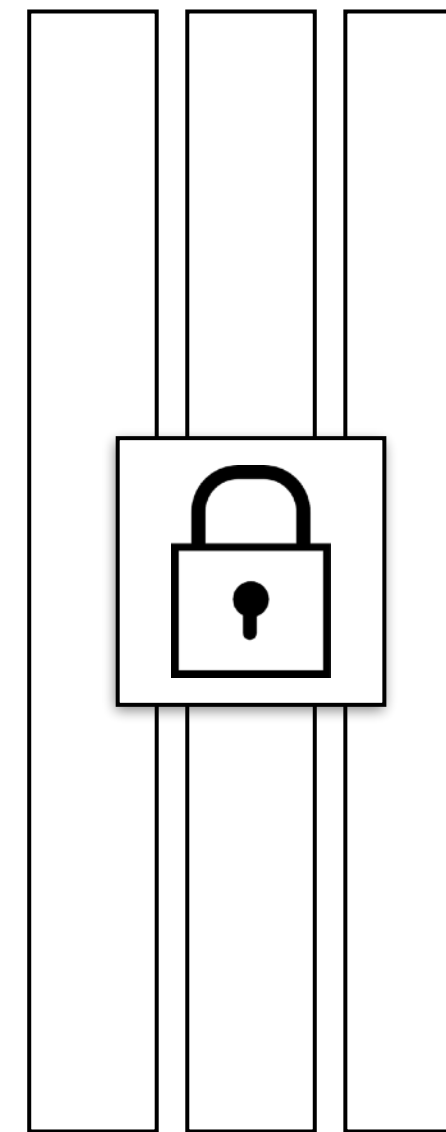
Synthesized
images from
another CNN



Real “target”
images



ProGAN detector



Real vs. fake?

How well do classifiers generalize?

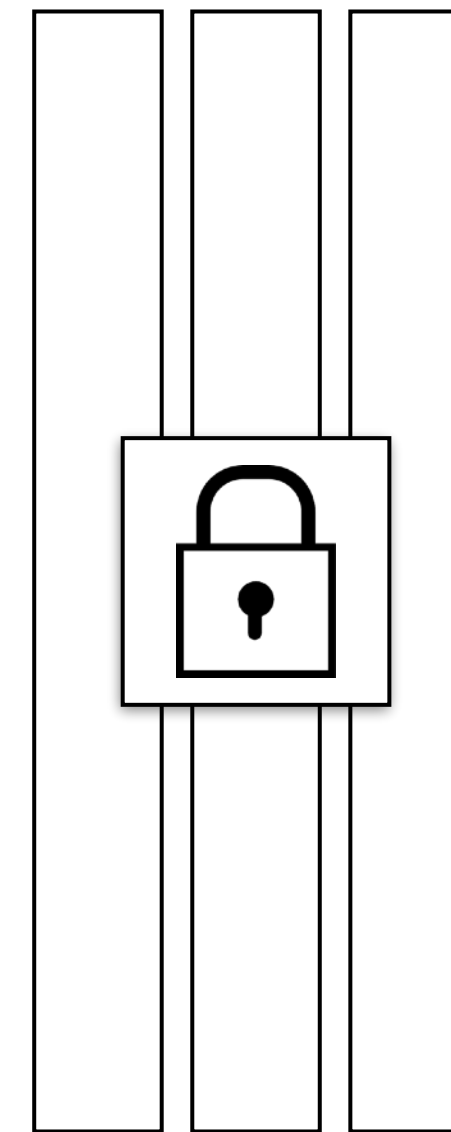
Images the CNN **actually** makes



Images the CNN **should** make



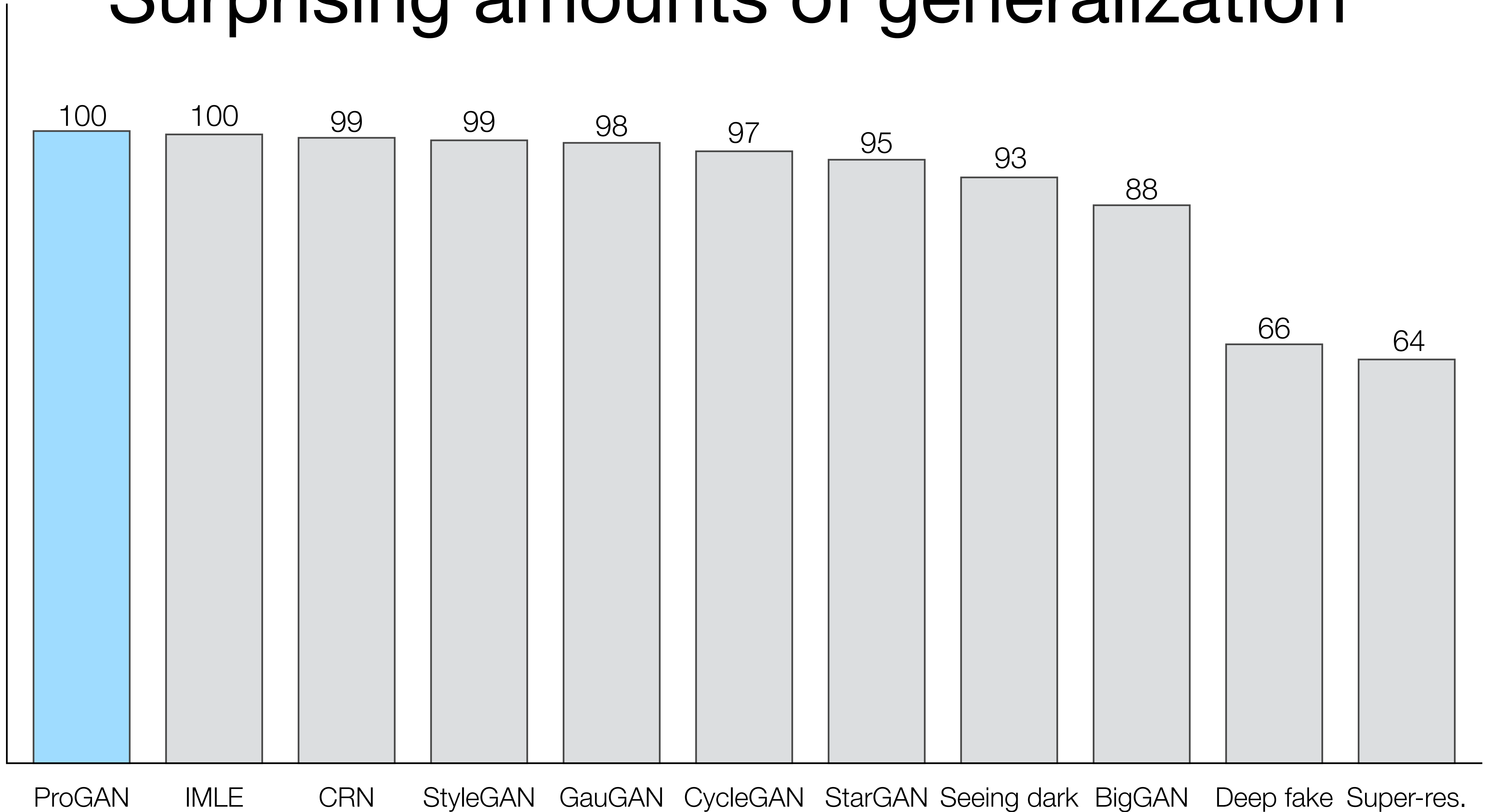
ProGAN detector



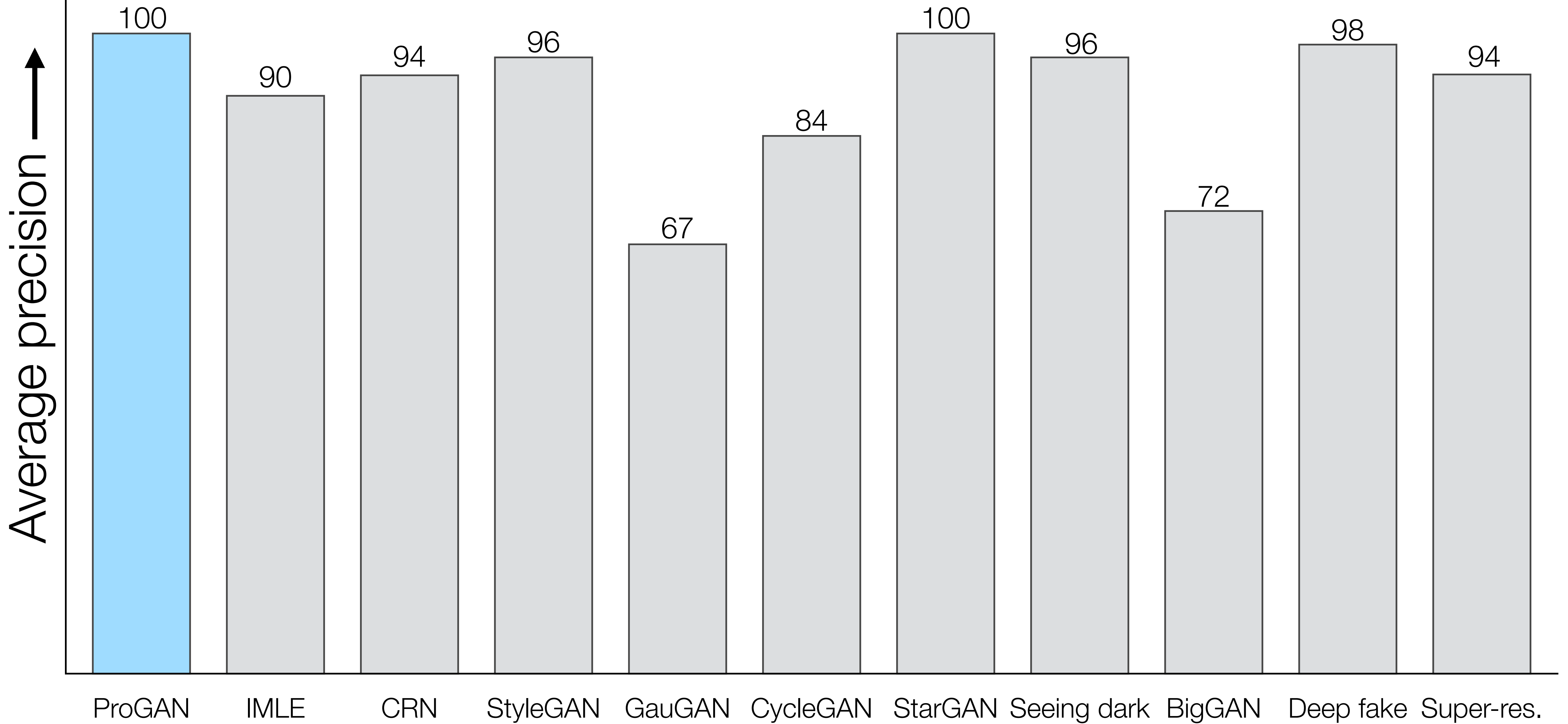
Real vs. fake?

Surprising amounts of generalization

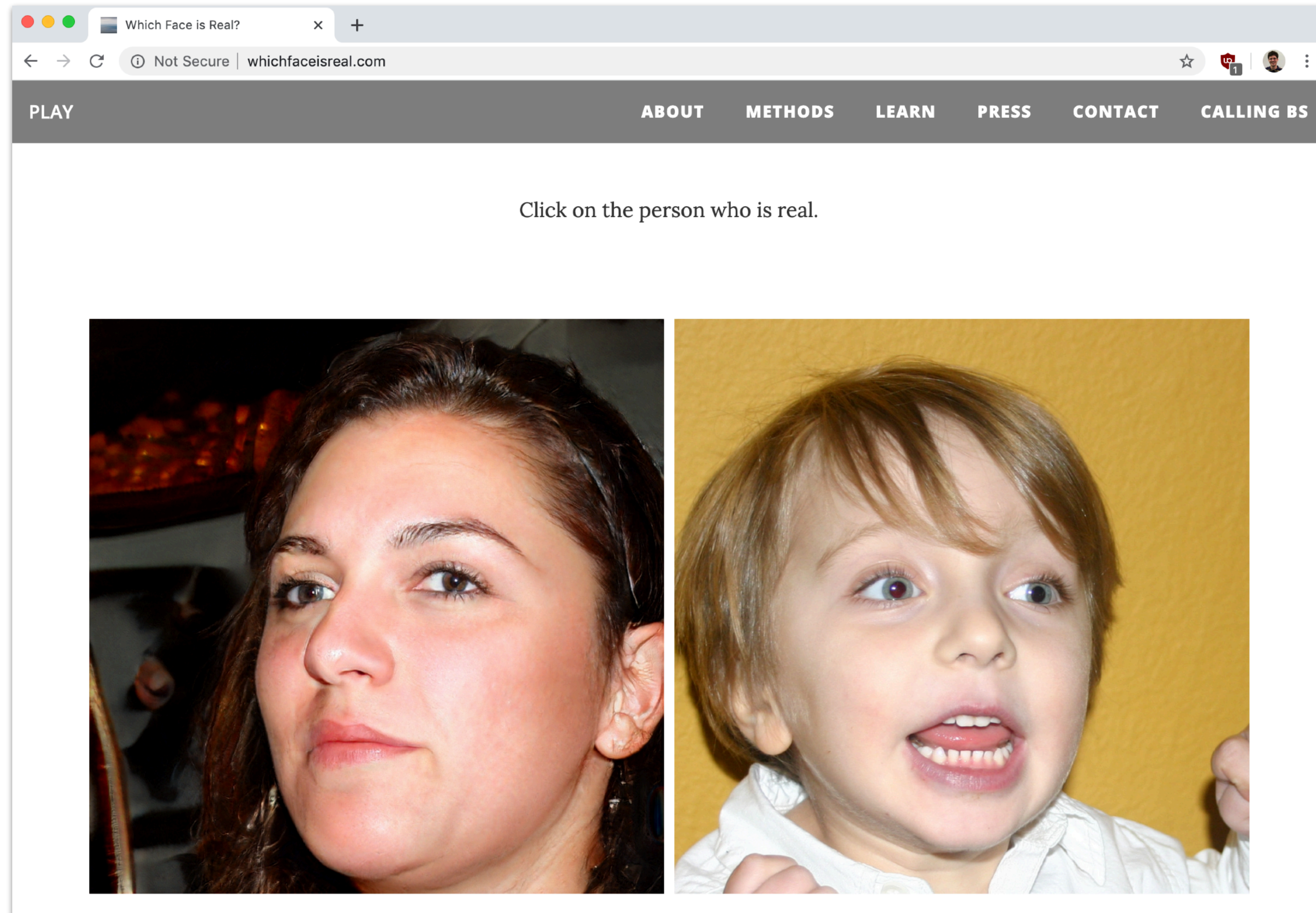
Average precision \uparrow



Generalization to other CNNs: no preprocessing



Generalization example



Detection accuracy: 93% AP

“Out-of-distribution” dataset:

- StyleGAN faces
- 1024x1024 JPEGs
- Use minimal preprocessing:
take 224x224 center crop

<http://whichfaceisreal.com> [West and Bergstrom 2019]

Generalization to StyleGAN3

NVlabs / **stylegan3-detector** Public

Notifications Fork 7 Star 109

Code Issues 1 Pull requests Actions Projects Security Insights

master 1 branch 0 tags

Go to file Code

Koki Nagano Add Unisi and kitware teams' results c84252f on Oct 26, 2021 6 commits

- figs Add Unisi and kitware teams' results 2 years ago
- slides Add Unisi and kitware teams' results 2 years ago
- README.md Add Unisi and kitware teams' results 2 years ago

README.md

StyleGAN3 Synthetic Image Detection

Overview

While new generator models, such as [StyleGAN3](#), enable new media synthesis capabilities, they may also present a new challenge for AI forensics algorithms for detection, attribution, and characterization of synthetic media.

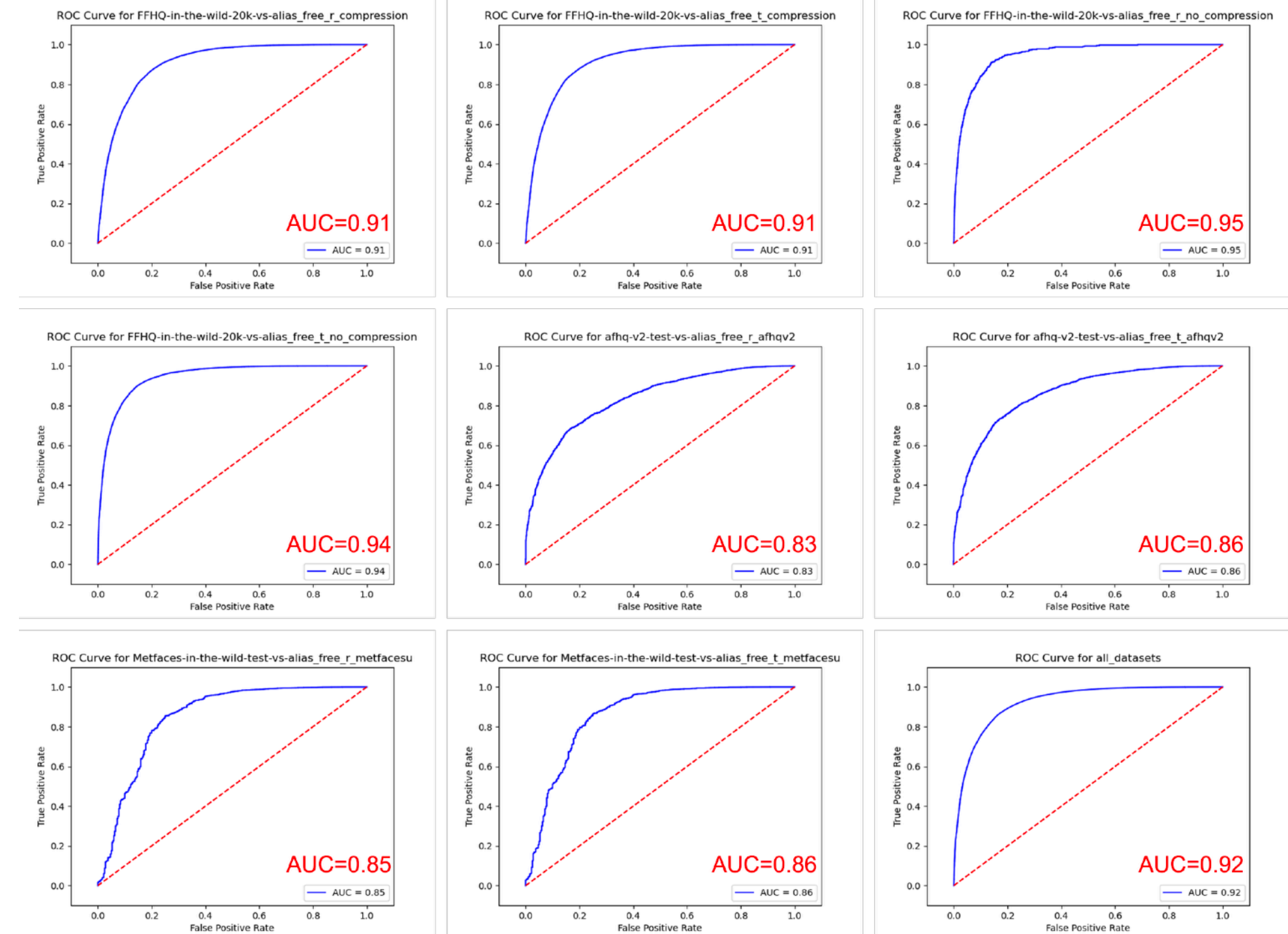
As part of DARPA's [Semantic Forensics](#) (SemaFor, for short) program, NVIDIA has been collaborating with digital forensics experts and researchers to help advance the capabilities to verify the authenticity and provenance of

About
No description, website, or topics provided.

Readme
109 stars
9 watching
7 forks
Report repository

Releases
No releases published

Packages
No packages published

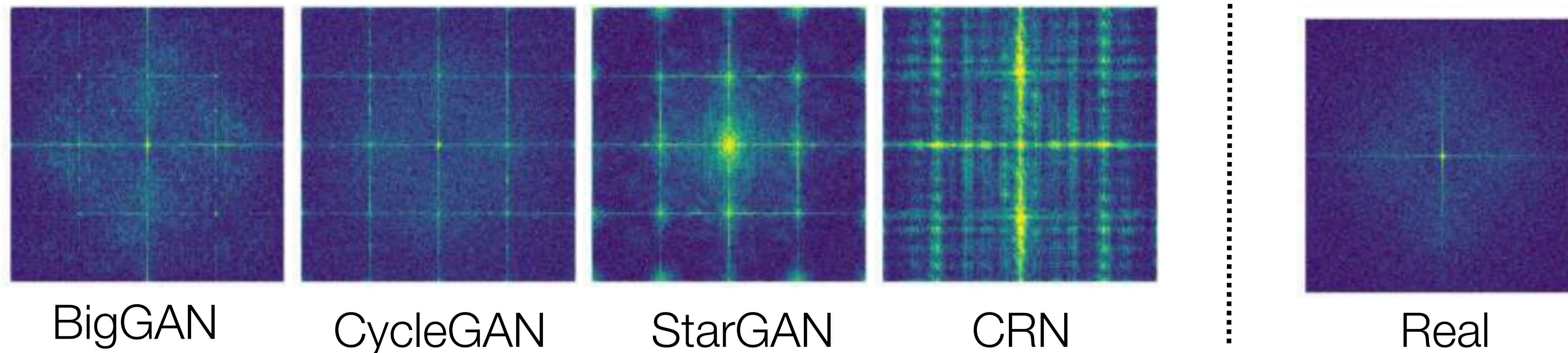


A model trained on a model from 2019 (ProGAN) generalizes to a (similar) model in 2021 (StyleGAN3)

Implications

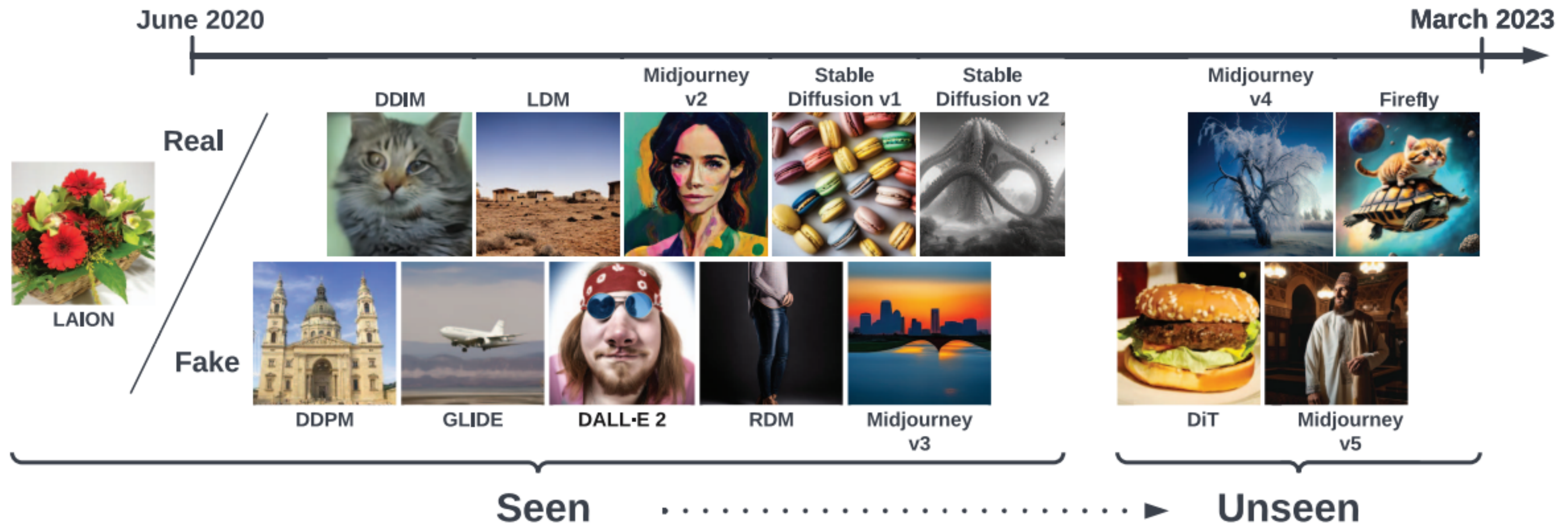
- Suggests CNN-generated images have common artifacts
- These artifacts can be detected with a simple classifier!
- But what *are* these artifacts?

Average Fourier magnitude (after high pass filtering)



Example from literature: checkerboard/aliasing artifacts [Xu Zhang et al. 2019]

Need online “open world” detection



Source: [Epstein et al., “Online Detection of AI-Generated Images”, 2023],

See also [Girish et al., “Towards discovery and attribution of open-world gan generated images”, 2021]

What's real and what's fake?



["The suspicious video that helped spark an attempted coup in Gabon" Washington Post. 2020]

<https://www.youtube.com/watch?v=F5vzKs4z1dc>

Challenges on the horizon

- Lots of ways to make fake images.
- If we know what methods were used, there's a good chance we can succeed.
- But it's hard to capture all of them!
- False positives are still a huge problem.
- So are postprocessing operations, like cropping and compression.
- Need methods that can handle unseen models.
- Alternative approaches: watermarking, signatures, etc.

Open-ended discussion

- How susceptible are people to fake images?
- Is there any hope of detecting “most” fake images?
- Under what situations might it be important and/or feasible?
- How do we deal with false positives?