Visual Forensics and Societal Impacts

Jun-Yan Zhu

16-726 Learning-based Image Synthesis, Spring 2023

Many slides were adopted from Richard Zhang, Sheng-Yu Wang, Frédo Durand, Alyosha Efros, etc.

MGM Lion (https://www.snopes.com/fact-check/leo-the-lion-mgm-logo/)

Topics

- Fake Images and Forensics
- Copyrights/Law
- Biases
- …

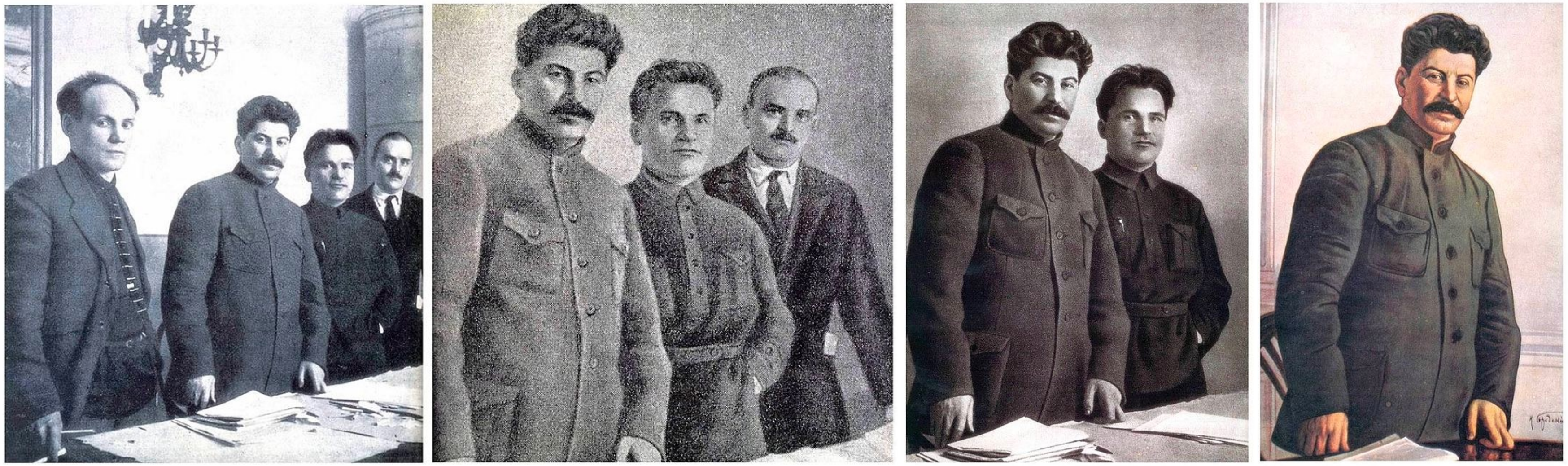
Visual Forensics

+ deep/shallow fake

+ misinformation

Photo manipulation old as photography

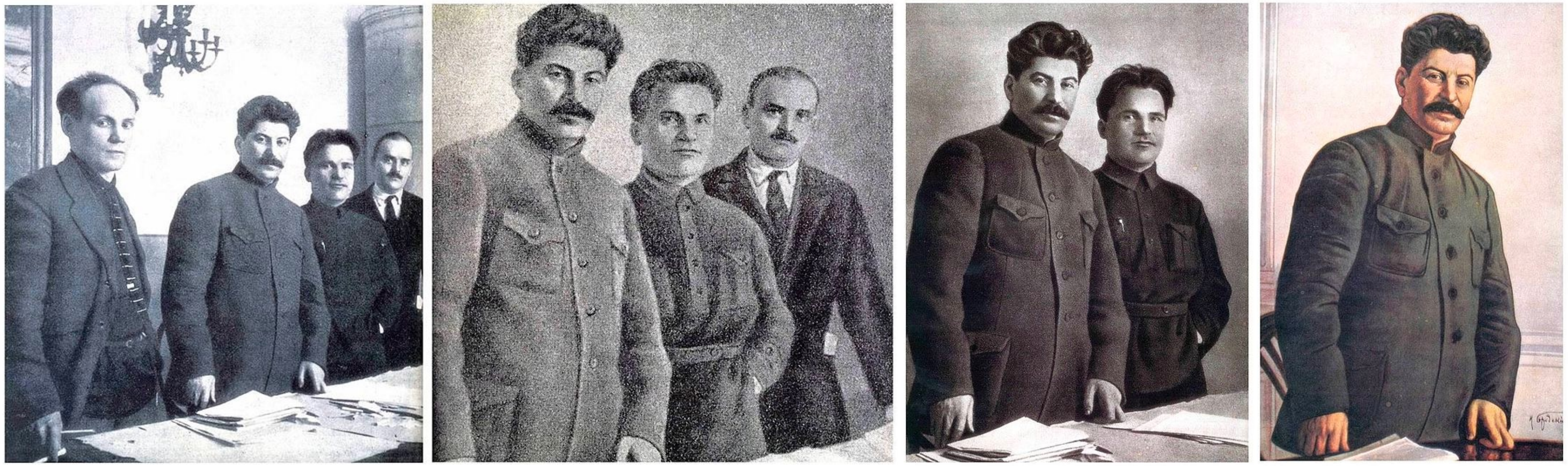
Joseph Stalin



Slides credit: Richard Zhang

AI: democratizing image **editing**

Joseph Stalin

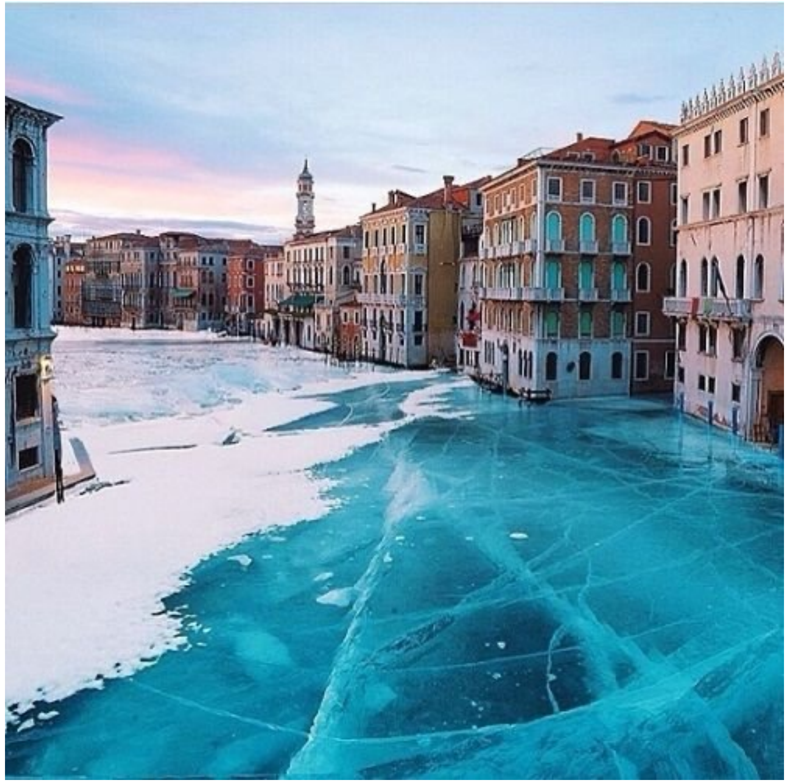


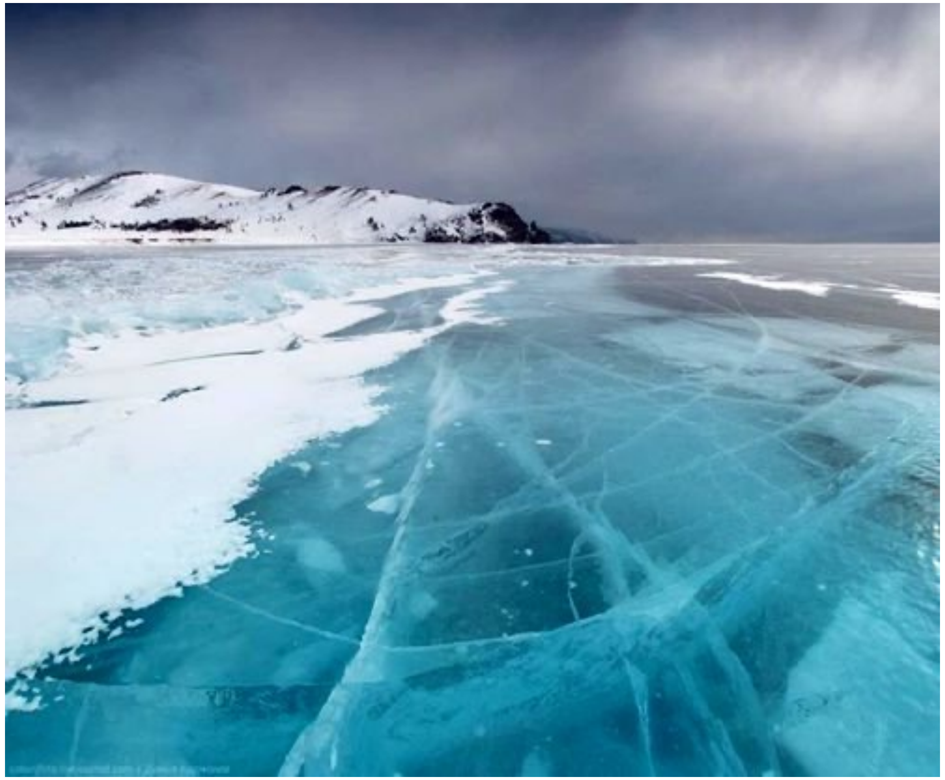
“DeepFakes”

Slides credit: Richard Zhang

More fake photos/videos

- https://www.quora.com/What-are-some-of-the-most-widely-circulated-fake-pictures

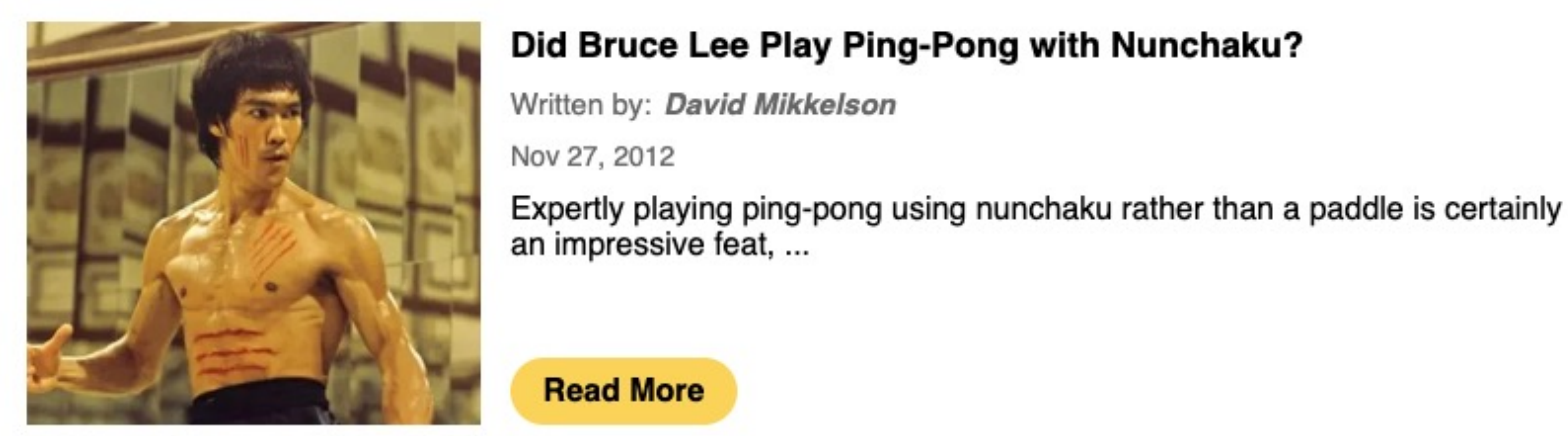




Slides credit: Frédo Durand

More fake photos/videos

- https://www.snopes.com/fact-check/category/photos/



Did Bruce Lee Play Ping-Pong with Nunchaku?

Written by: ***David Mikkelson***

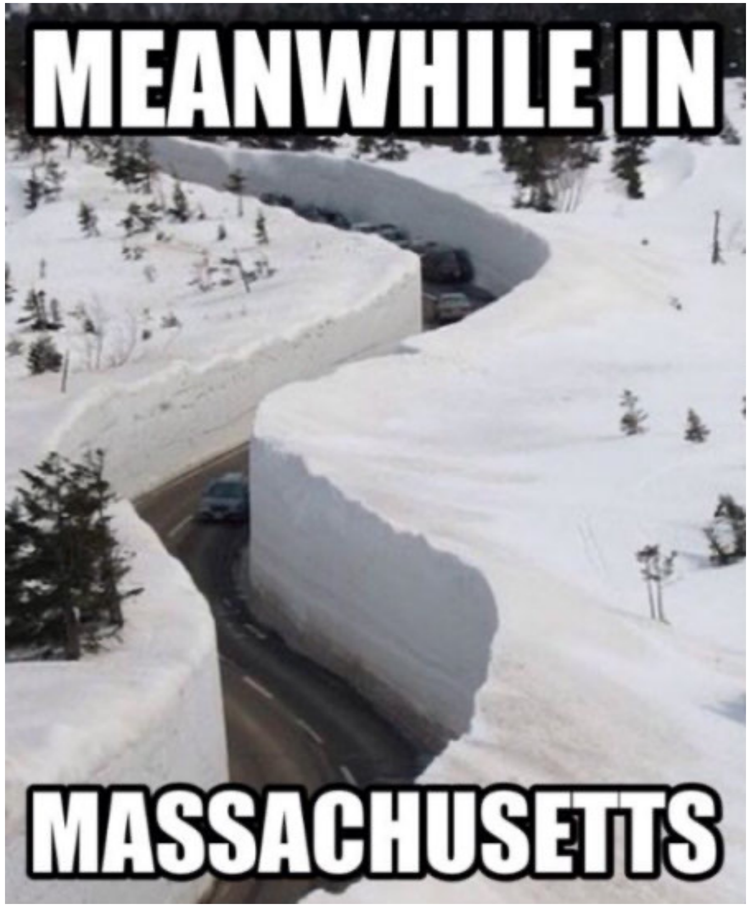
Nov 27, 2012

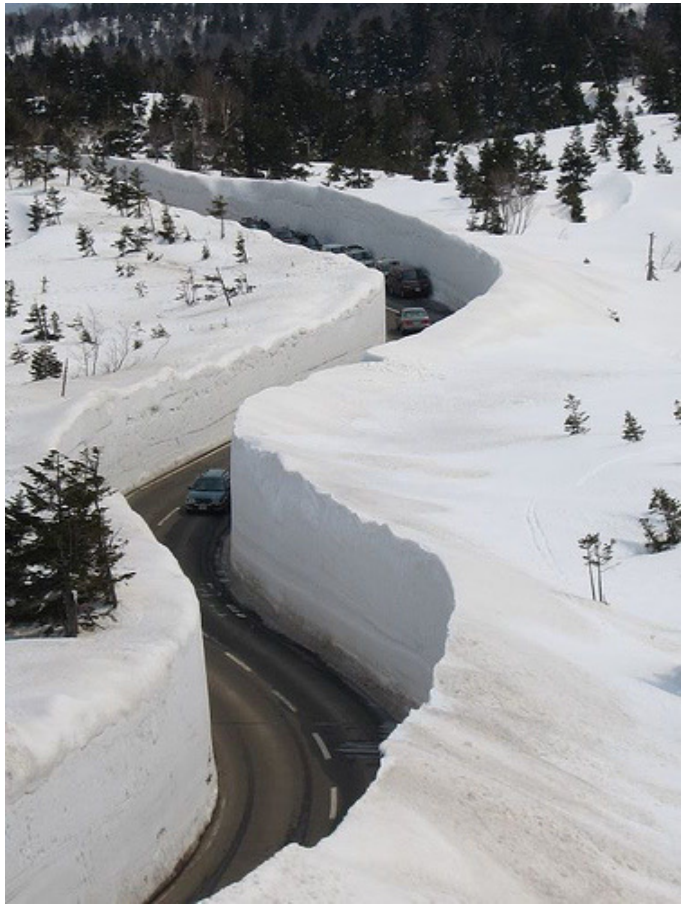
Expertly playing ping-pong using nunchaku rather than a paddle is certainly an impressive feat, ...

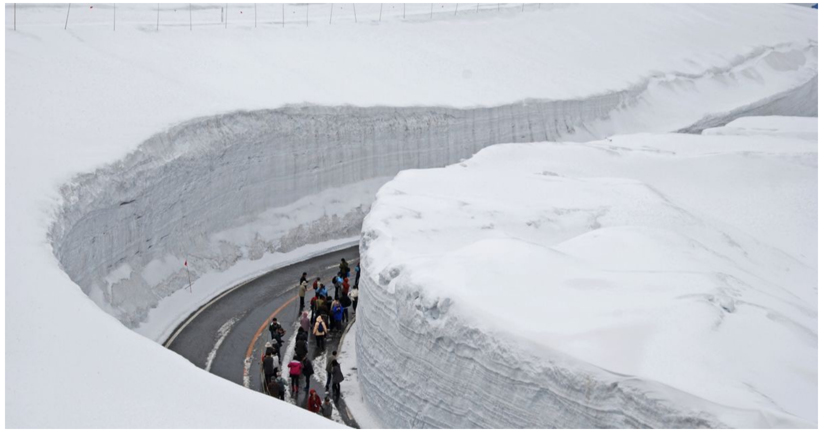
Bruce Lee plays Ping Pong?

https://www.snopes.com/fact-check/bruce-lee-ping-pong/

Miscaptioned photos

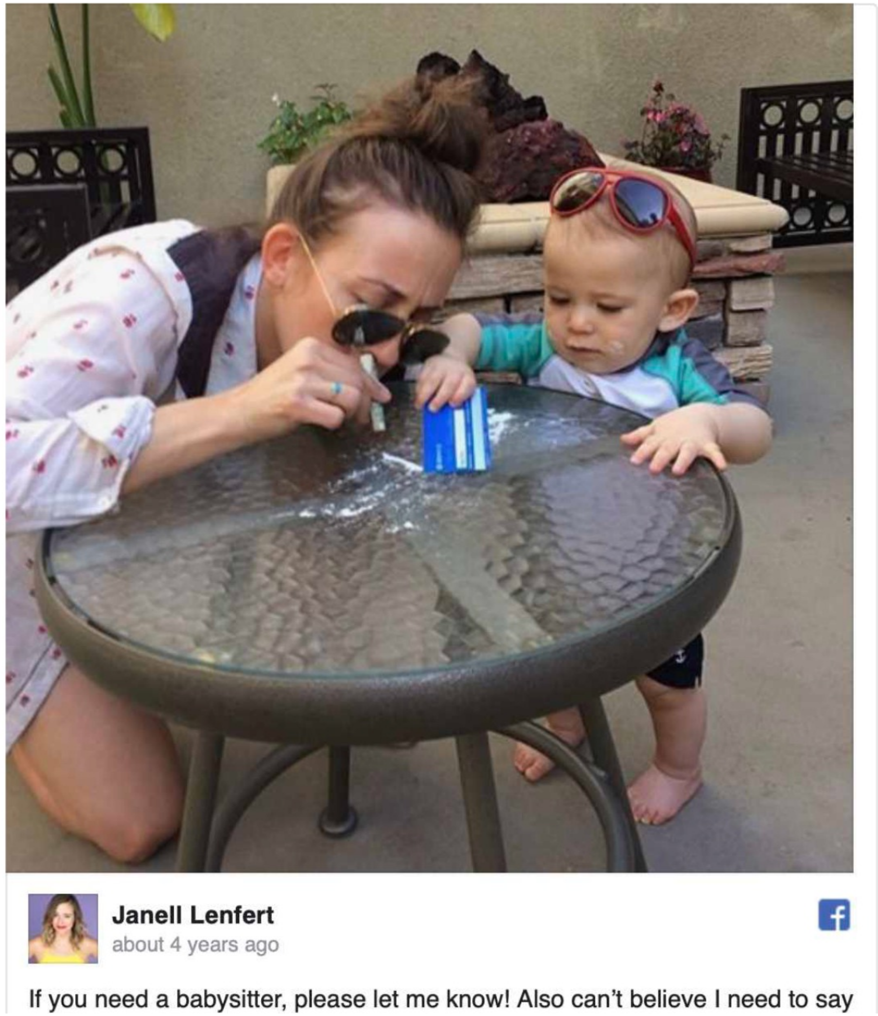






https://www.snopes.com/fact-check/snow-walls/

Context matters

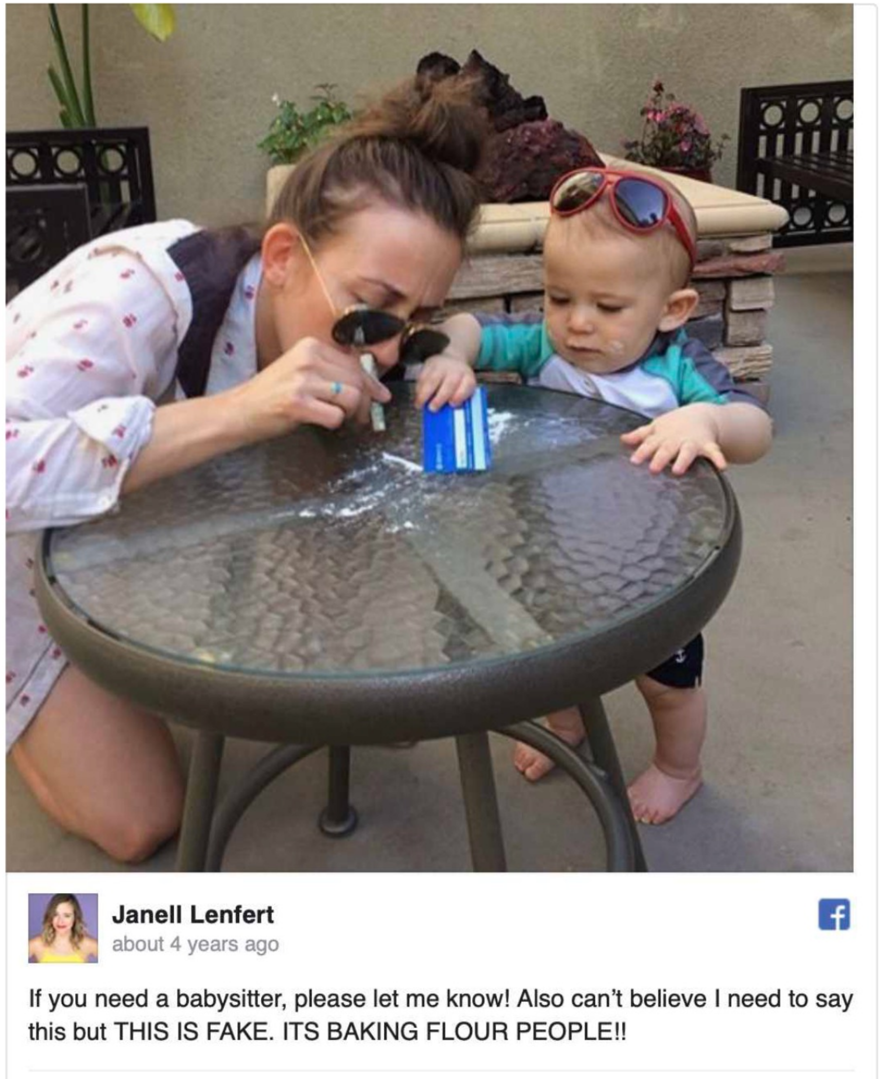


Janell Lenfert
about 4 years ago

If you need a babysitter, please let me know! Also can’t believe I need to say

Slides credit: Frédo Durand

Context matters

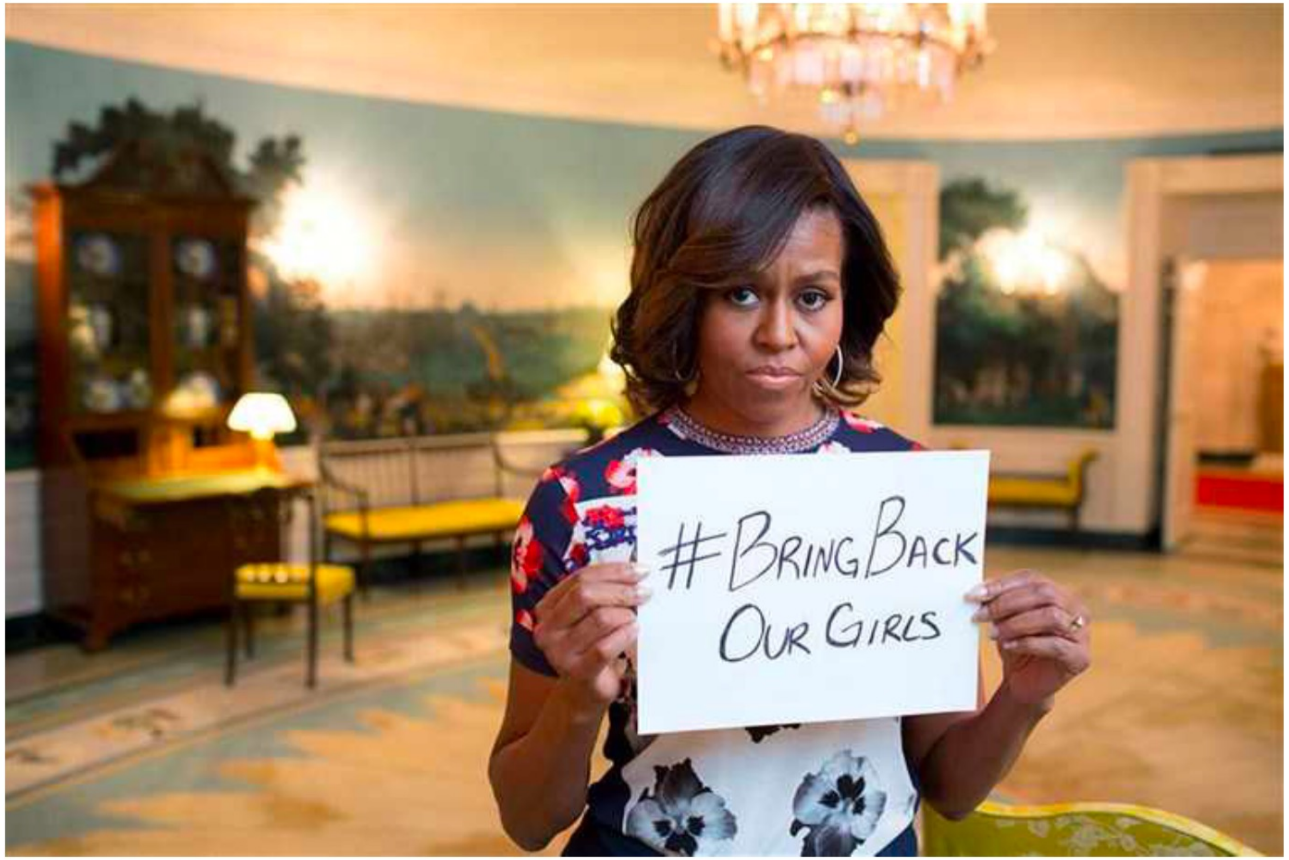


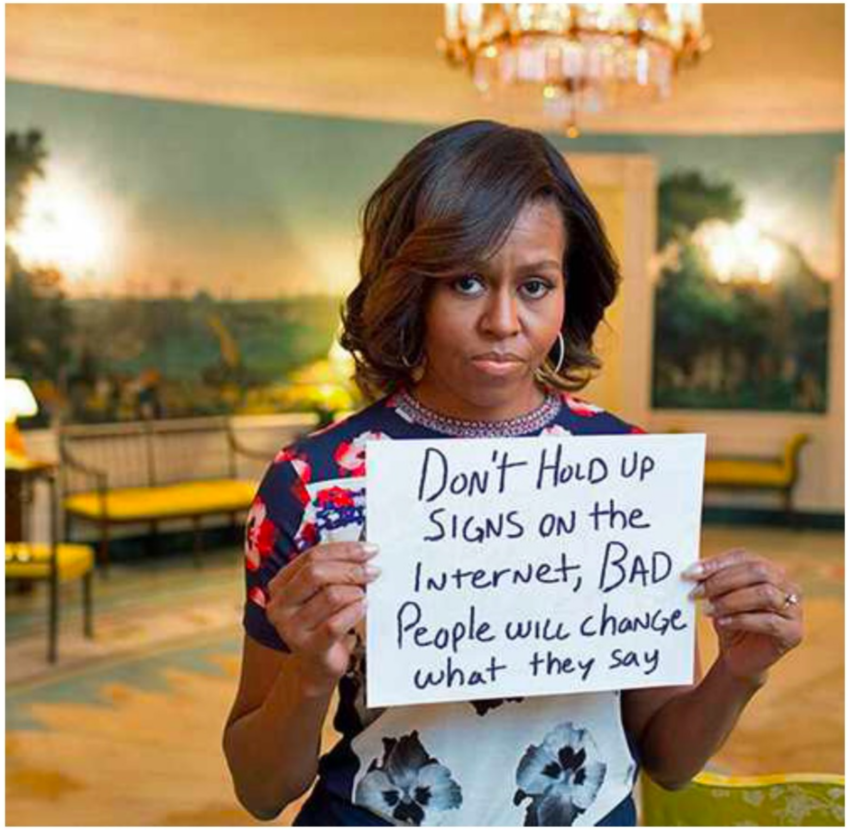
Janell Lenfert
about 4 years ago

If you need a babysitter, please let me know! Also can't believe I need to say this but THIS IS FAKE. ITS BAKING FLOUR PEOPLE!!

Slides credit: Frédo Durand

Don’t hold up signs





Slides credit: Frédo Durand

News and Fake news

- Photo editing / deep fake
- Photo retouching
- Fake caption (time/place/people)
- Selective choice of photos to take, publish
- Choice of topics to cover and emphasize

- Why do lawyers/scholars fake less often?

Slides credit: Frédo Durand

Detect Shallow Fake

Early works in Visual Forensics

- Detect alterations
  - e.g., inconsistent lighting, inconsistent noise, cloning boundary, etc.
- Analyze patterns:
  - physical, geometric, optical, sensor, and image file properties

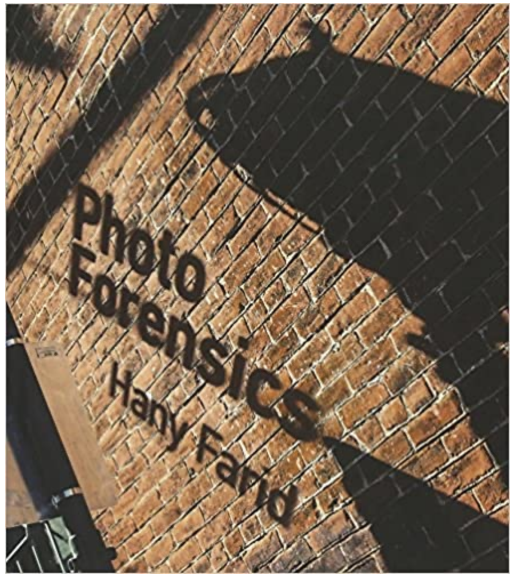
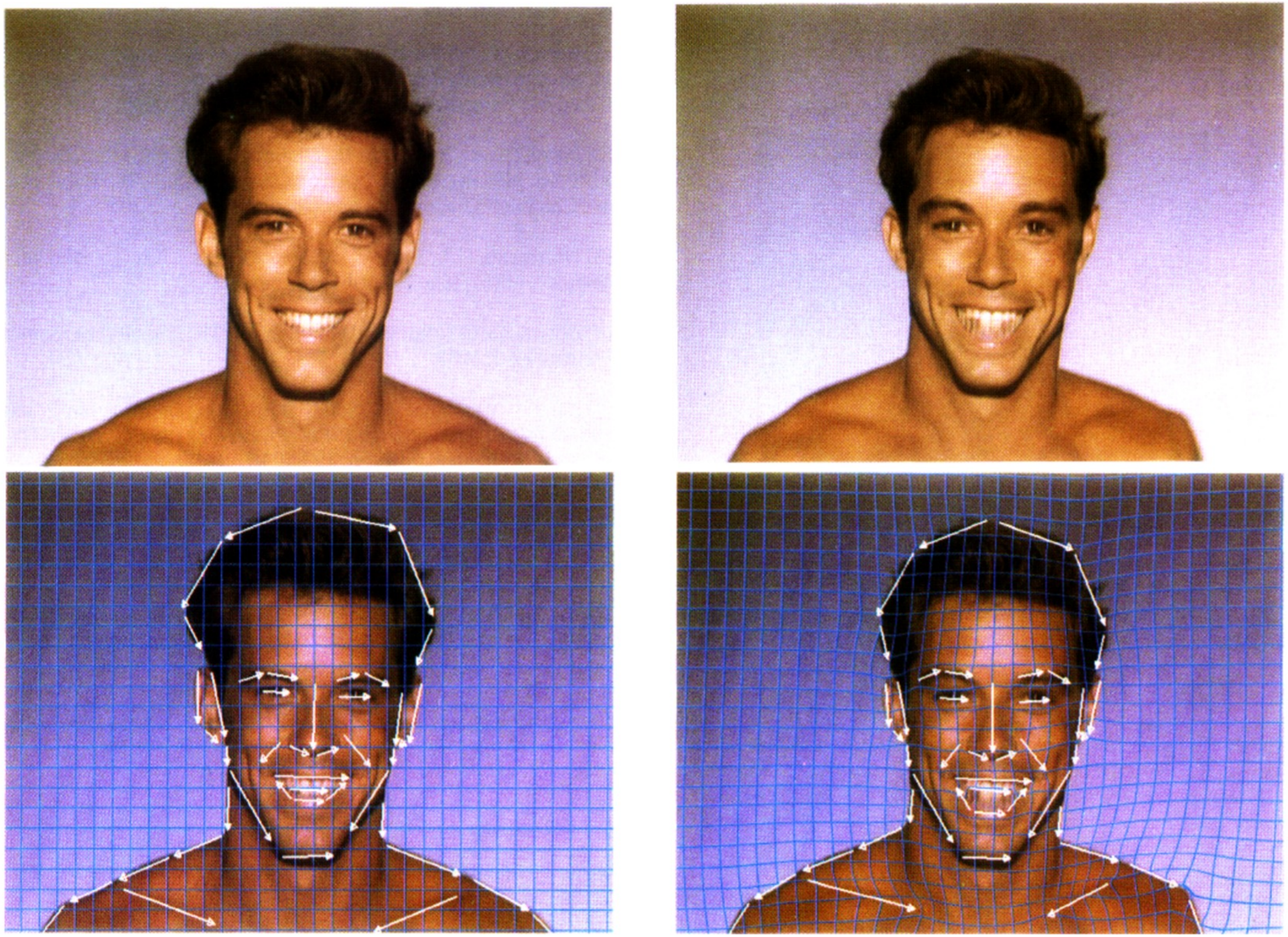


Photo Forensics (The MIT Press)
Hany Farid (https://farid.berkeley.edu/)

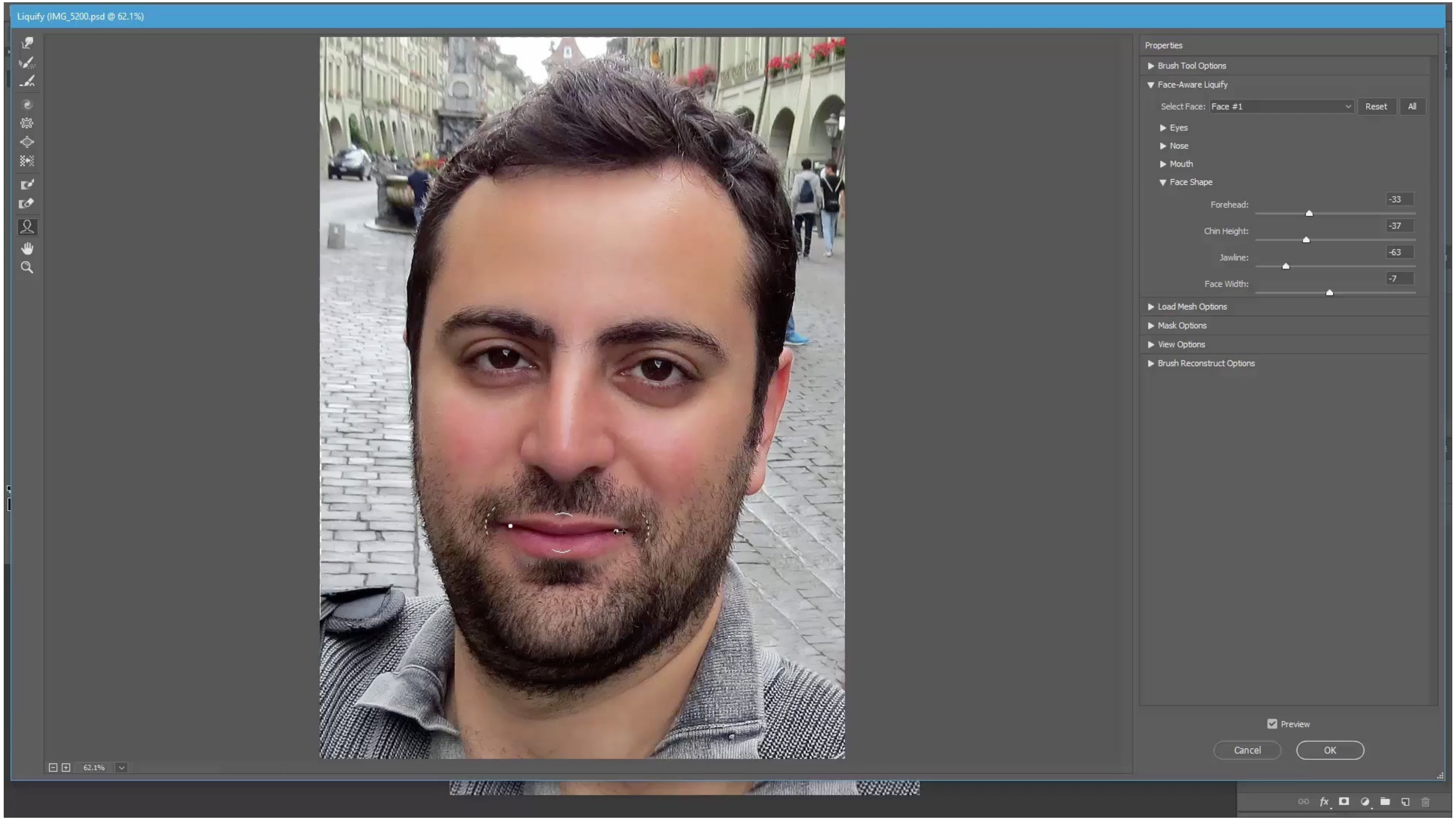
“Shallow” fakes



Feature-Based Image Metamorphosis (Beier et al. 1992)

Slides credit: Richard Zhang

Photoshop Face-Aware Liquify



Select Face: Face #1

Forehead: -33

Chin Height: -37

Jawline: -63

Face Width: -7

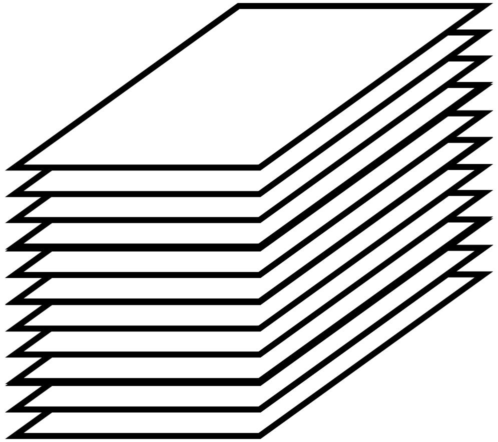
☑ Preview

Slides credit: Richard Zhang

https://www.youtube.com/watch?v=5Qqv_C6iVvQ

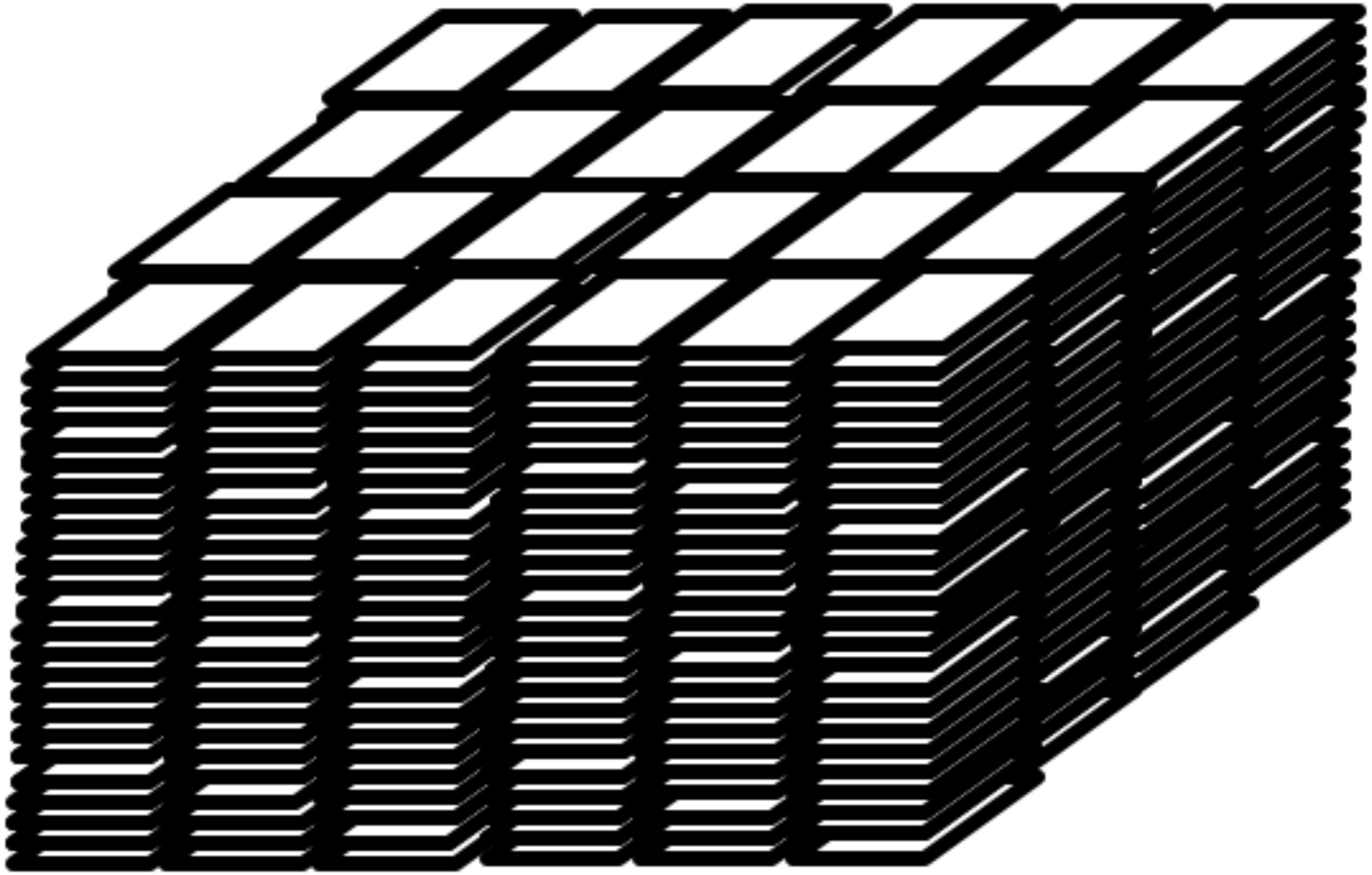
Supervised Learning?

<< Billions



Labeled manipulated images

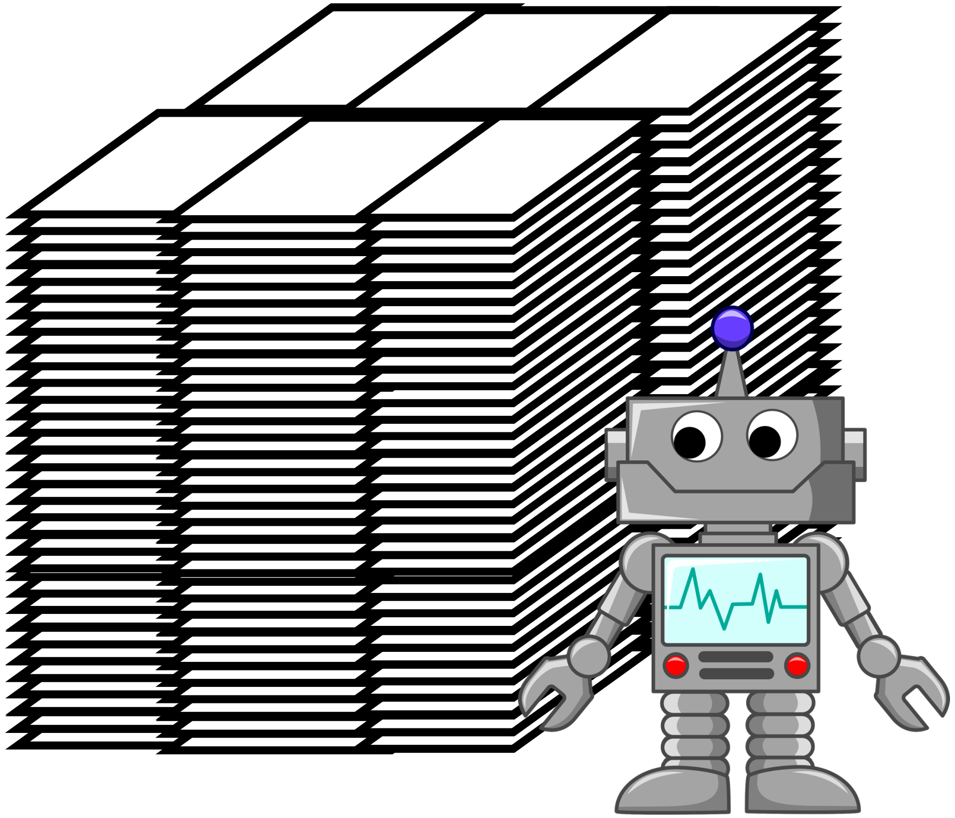
Billions!



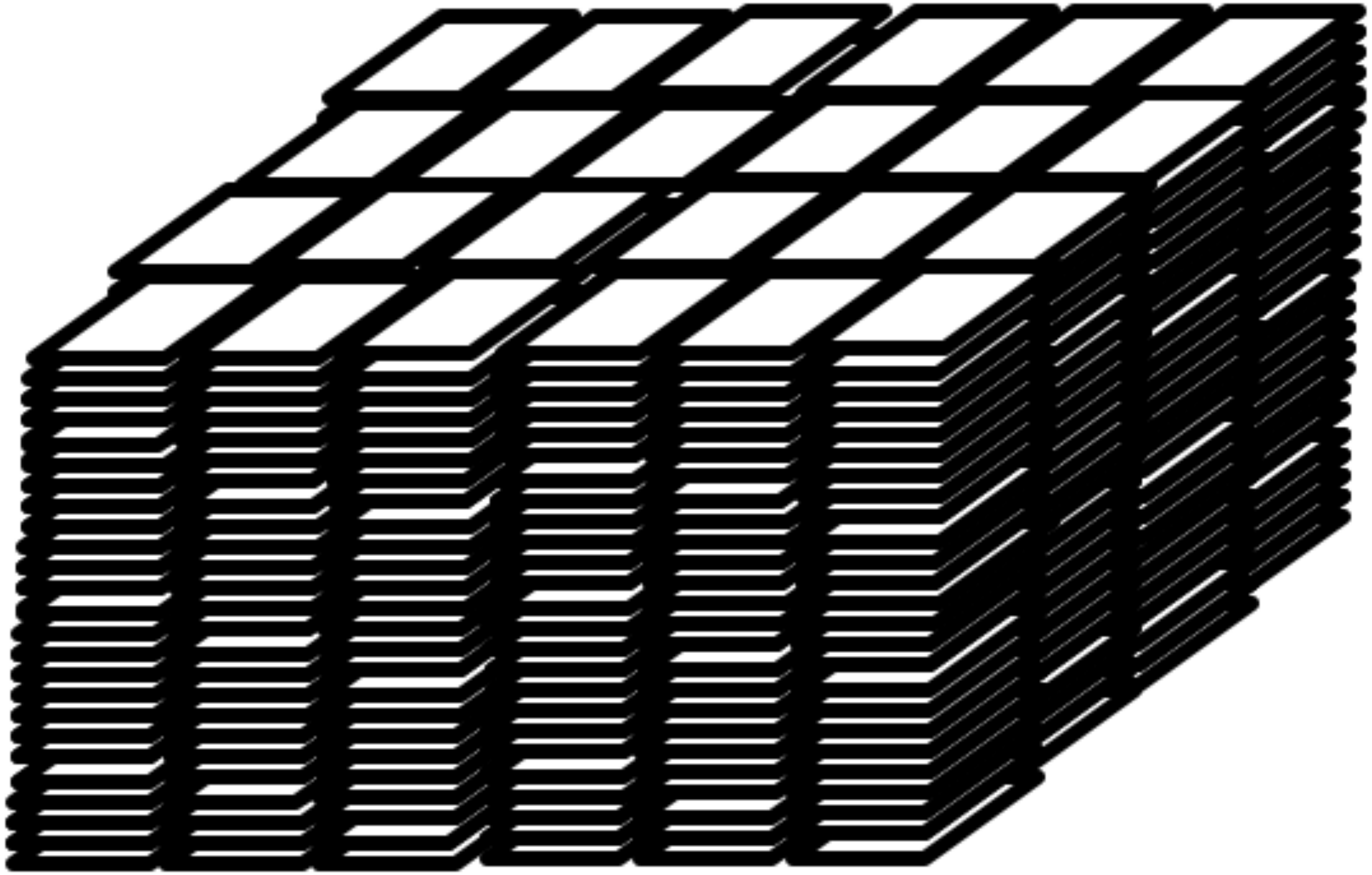
Natural Images

Slides credit: Richard Zhang

Supervised Learning with Self-Generated Data



Self-generated data

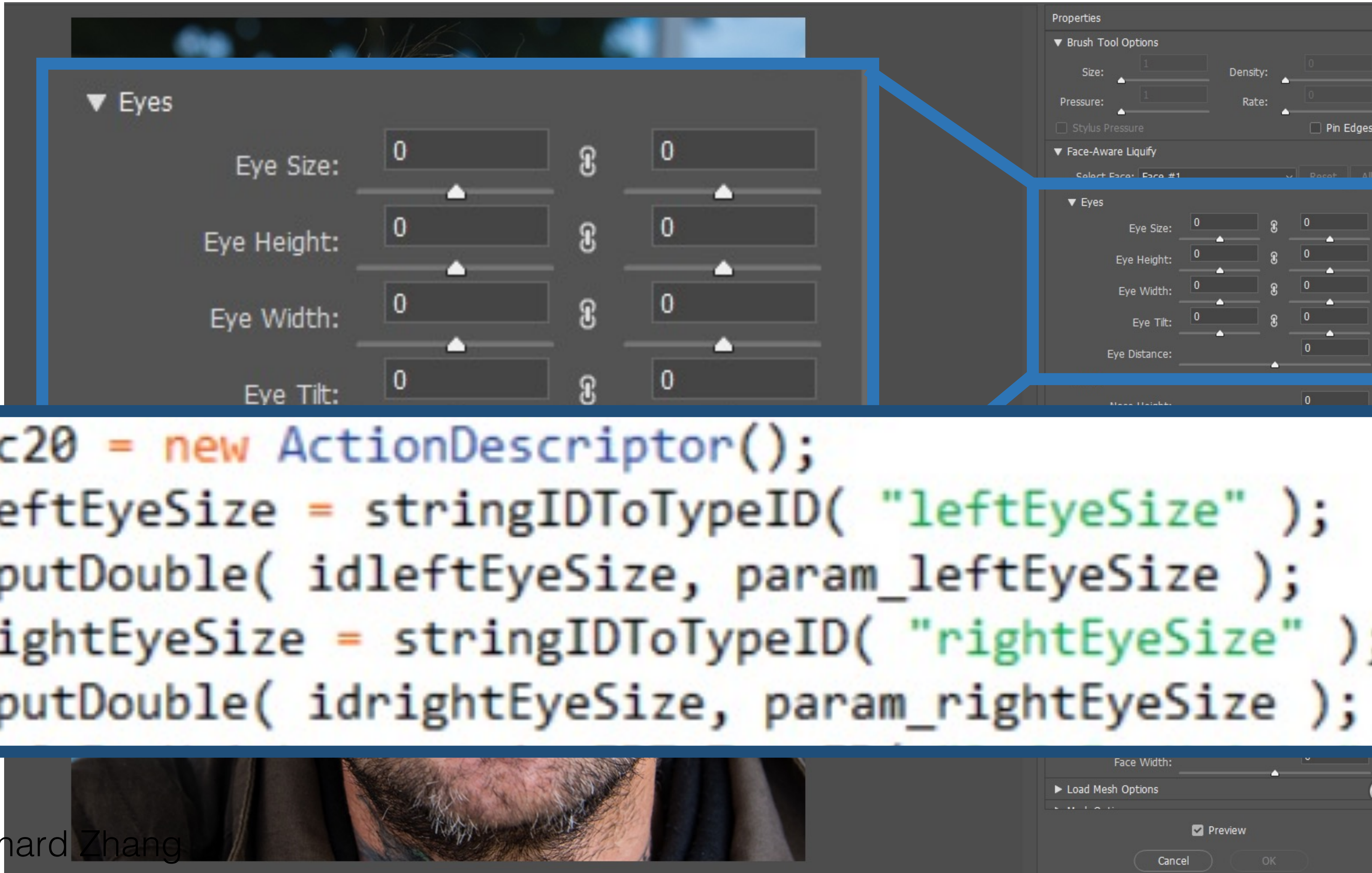


Billions!

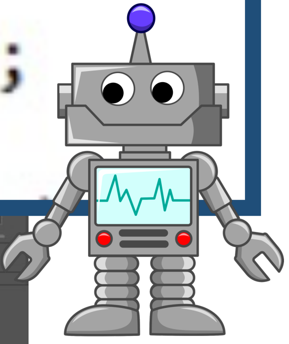
Natural Images

Slides credit: Richard Zhang

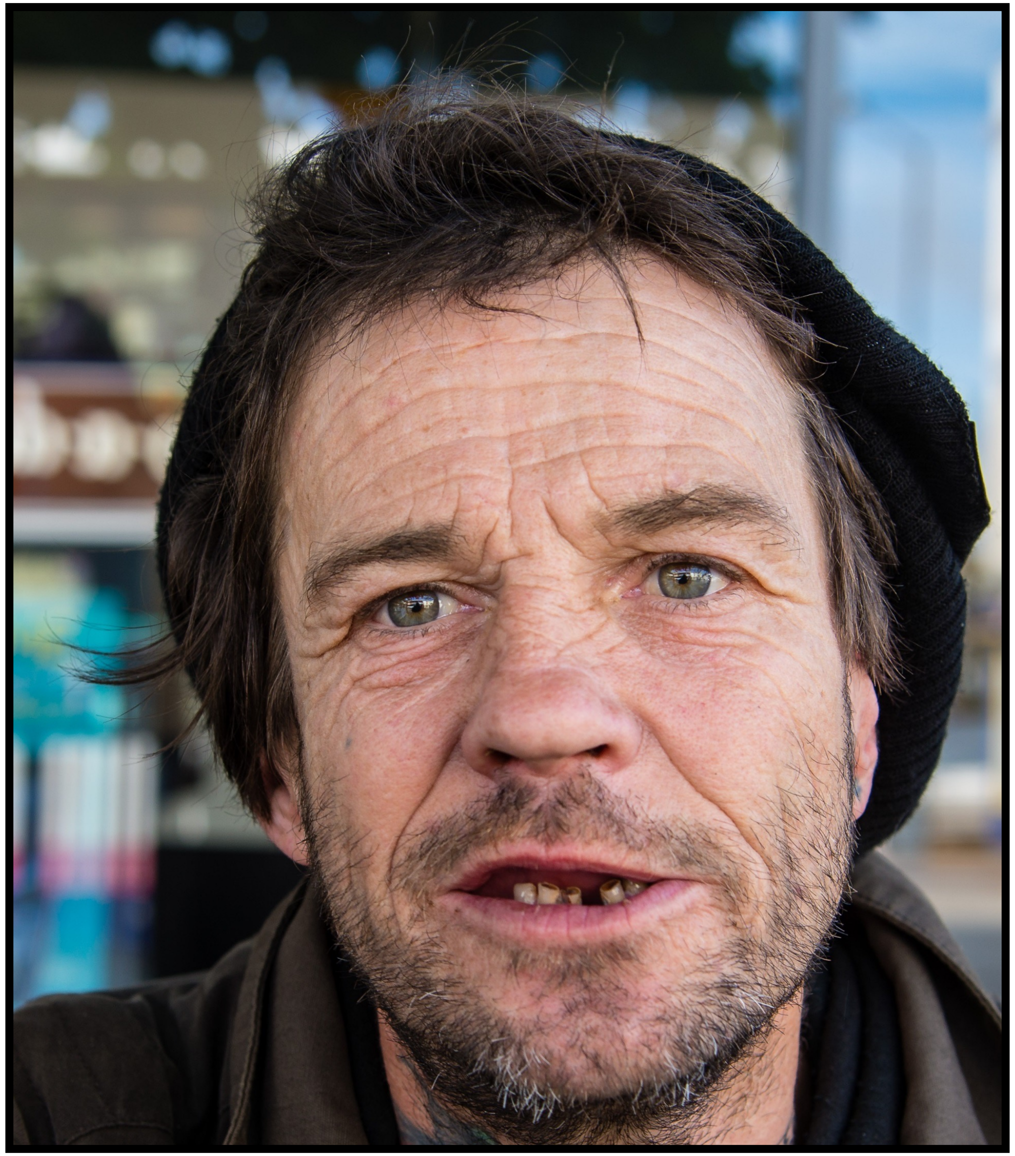
Scripting Photoshop



```
var desc20 = new ActionDescriptor();
var idleftEyeSize = stringIDToTypeID( "leftEyeSize" );
desc20.putDouble( idleftEyeSize, param_leftEyeSize );
var idrightEyeSize = stringIDToTypeID( "rightEyeSize" );
desc20.putDouble( idrightEyeSize, param_rightEyeSize );
```

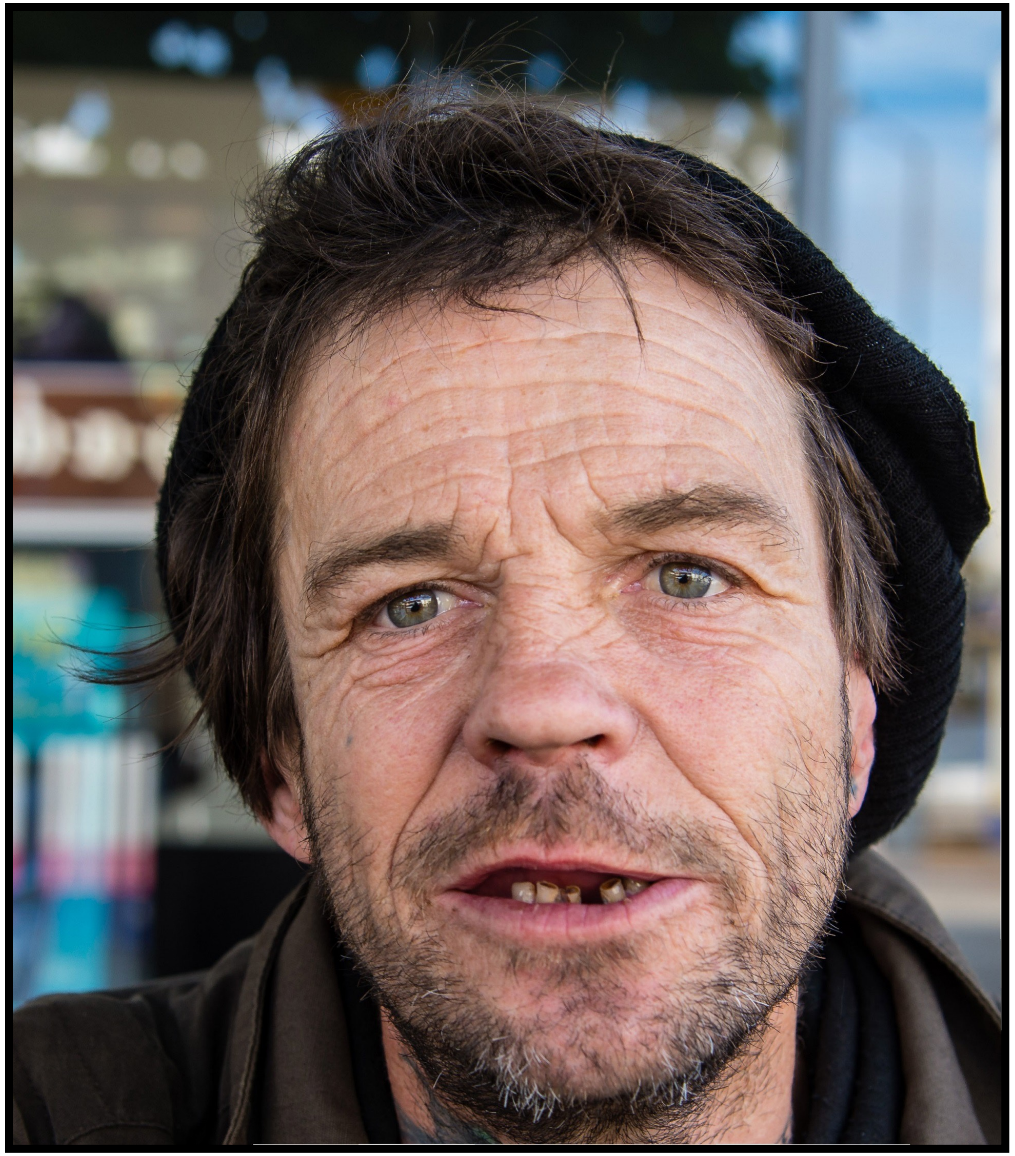


Slides credit: Richard Zhang

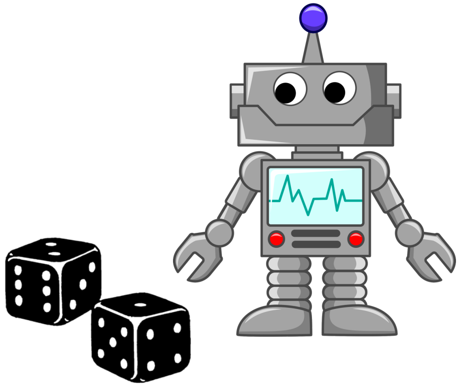


Original

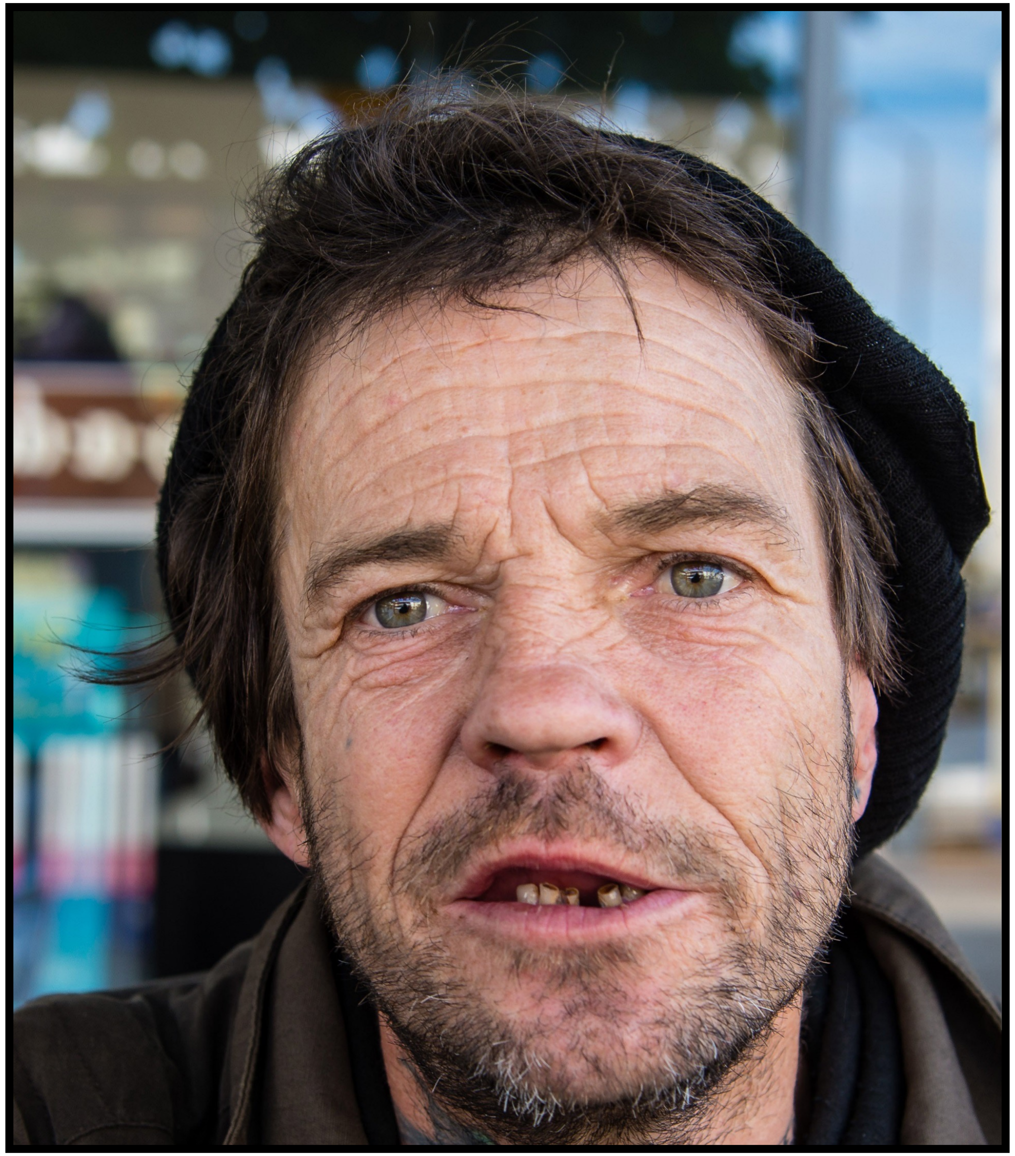
Slides credit: Richard Zhang



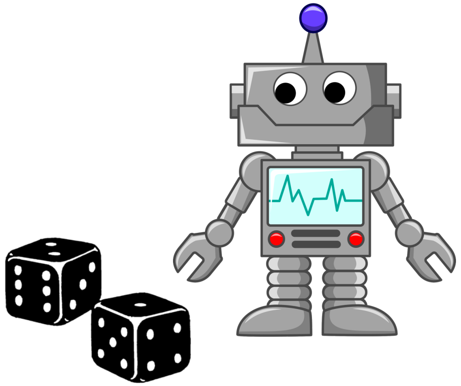
#1 modification



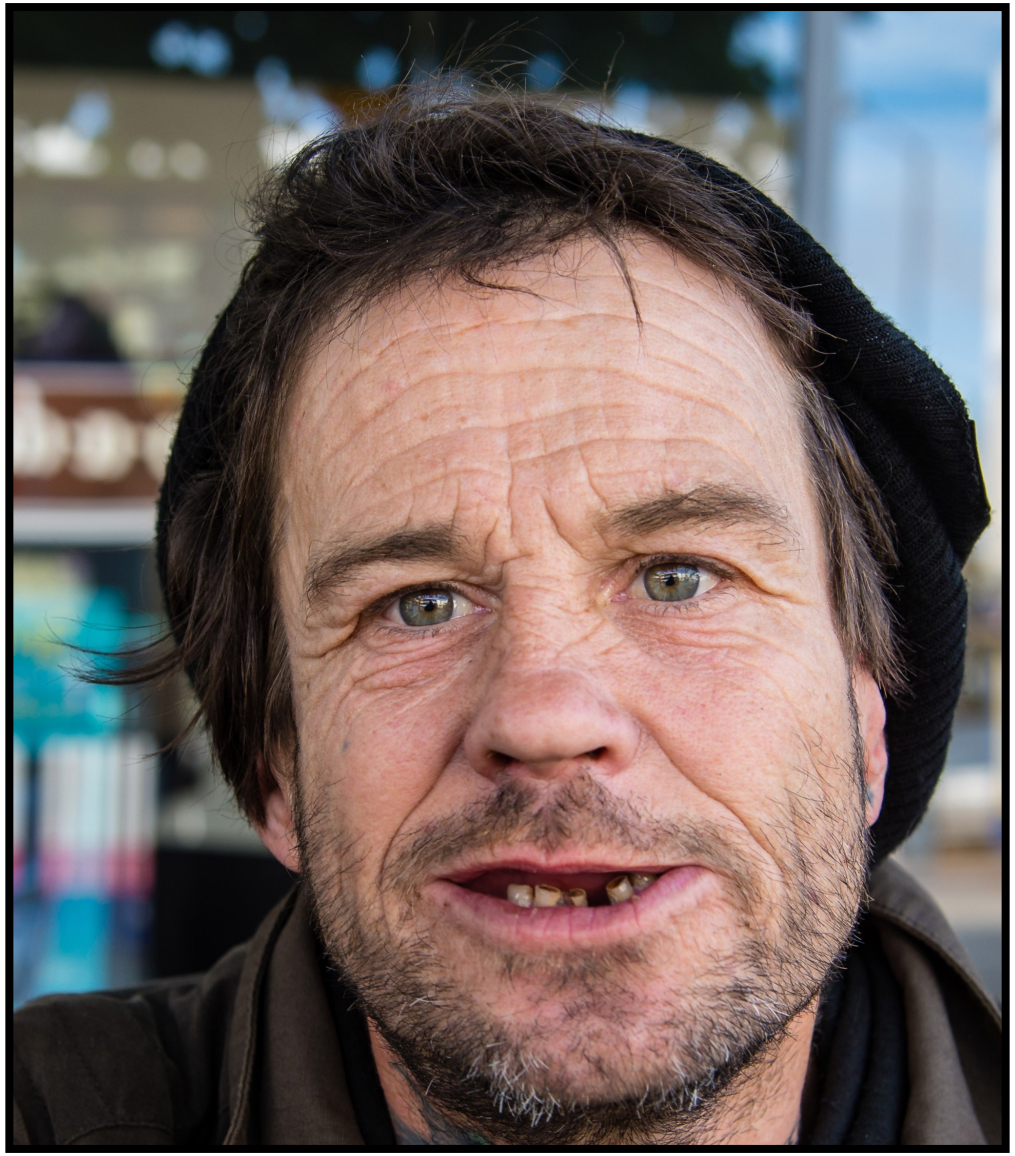
Slides credit: Richard Zhang



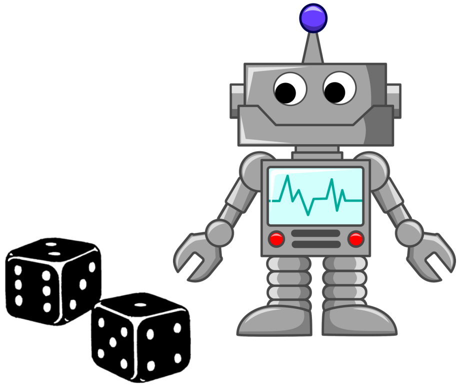
#2 modification



Slides credit: Richard Zhang

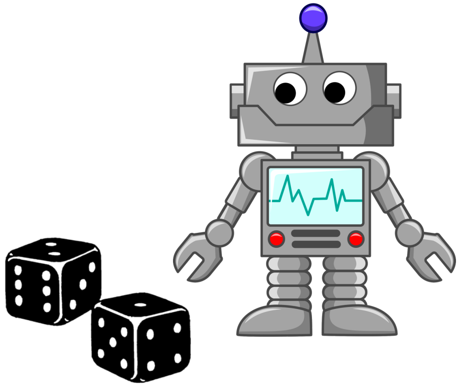


#3 modification



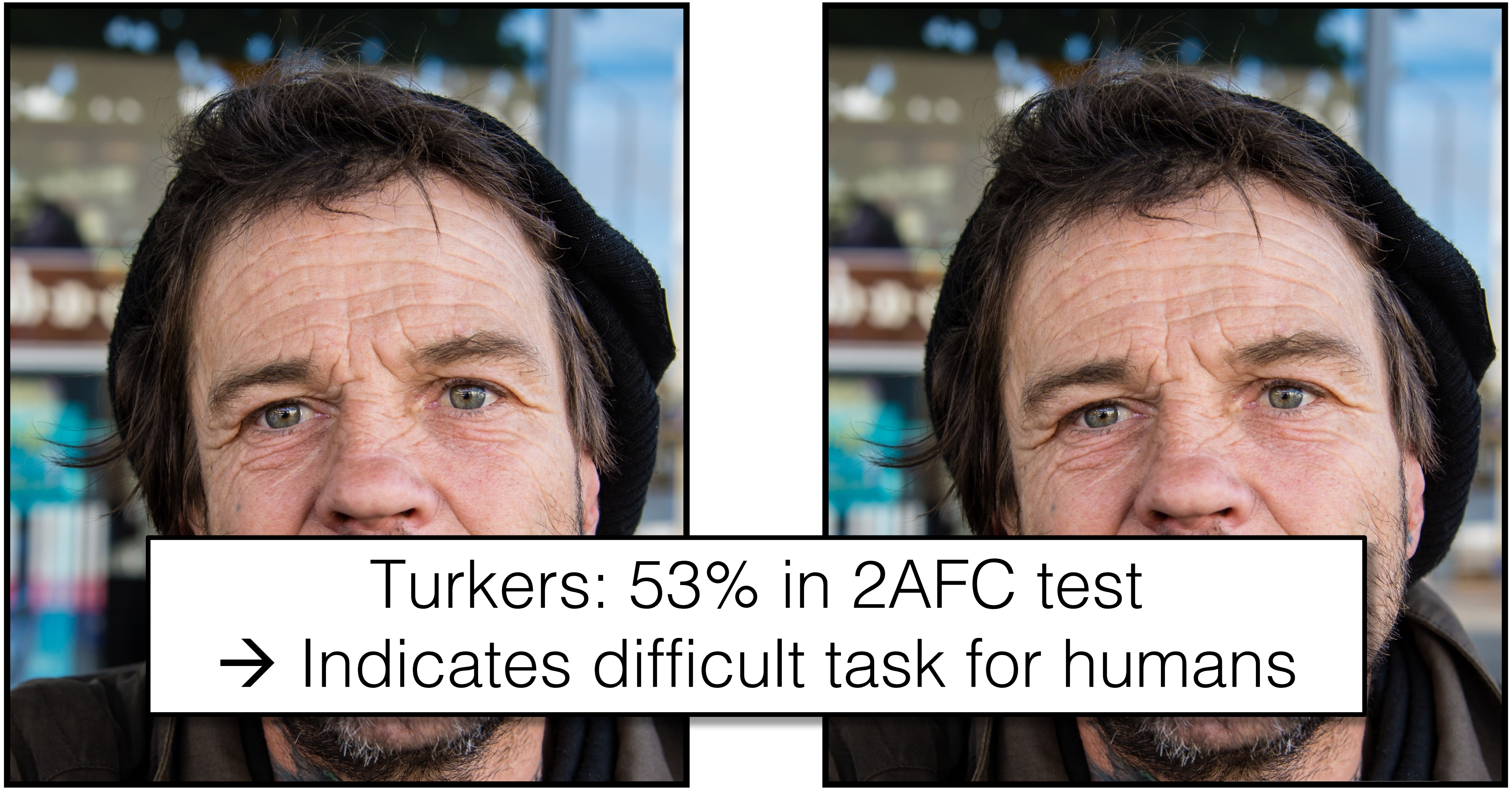
Slides credit: Richard Zhang

#4 modification



Slides credit: Richard Zhang

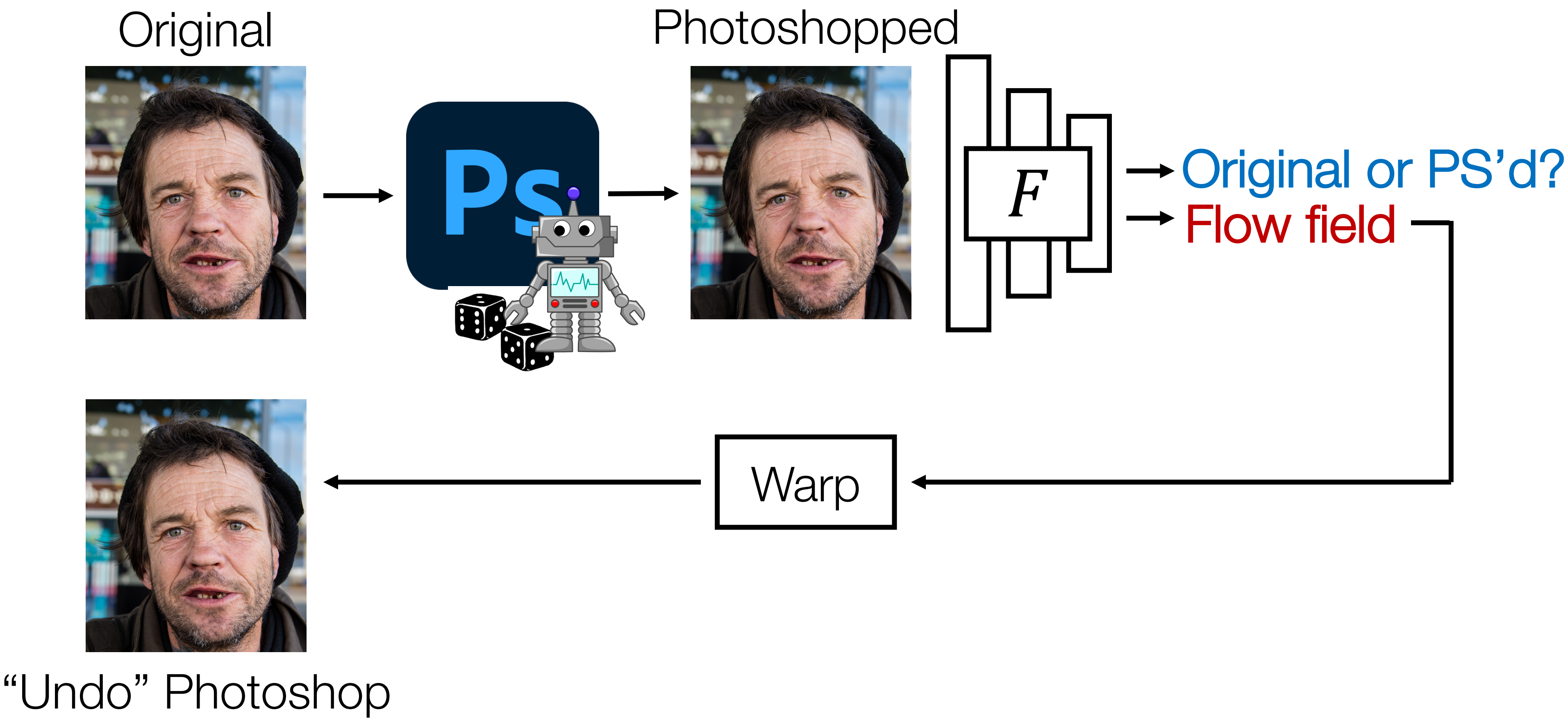
User study: Which image is modified?



Turkers: 53% in 2AFC test
→ Indicates difficult task for humans

Slides credit: Richard Zhang

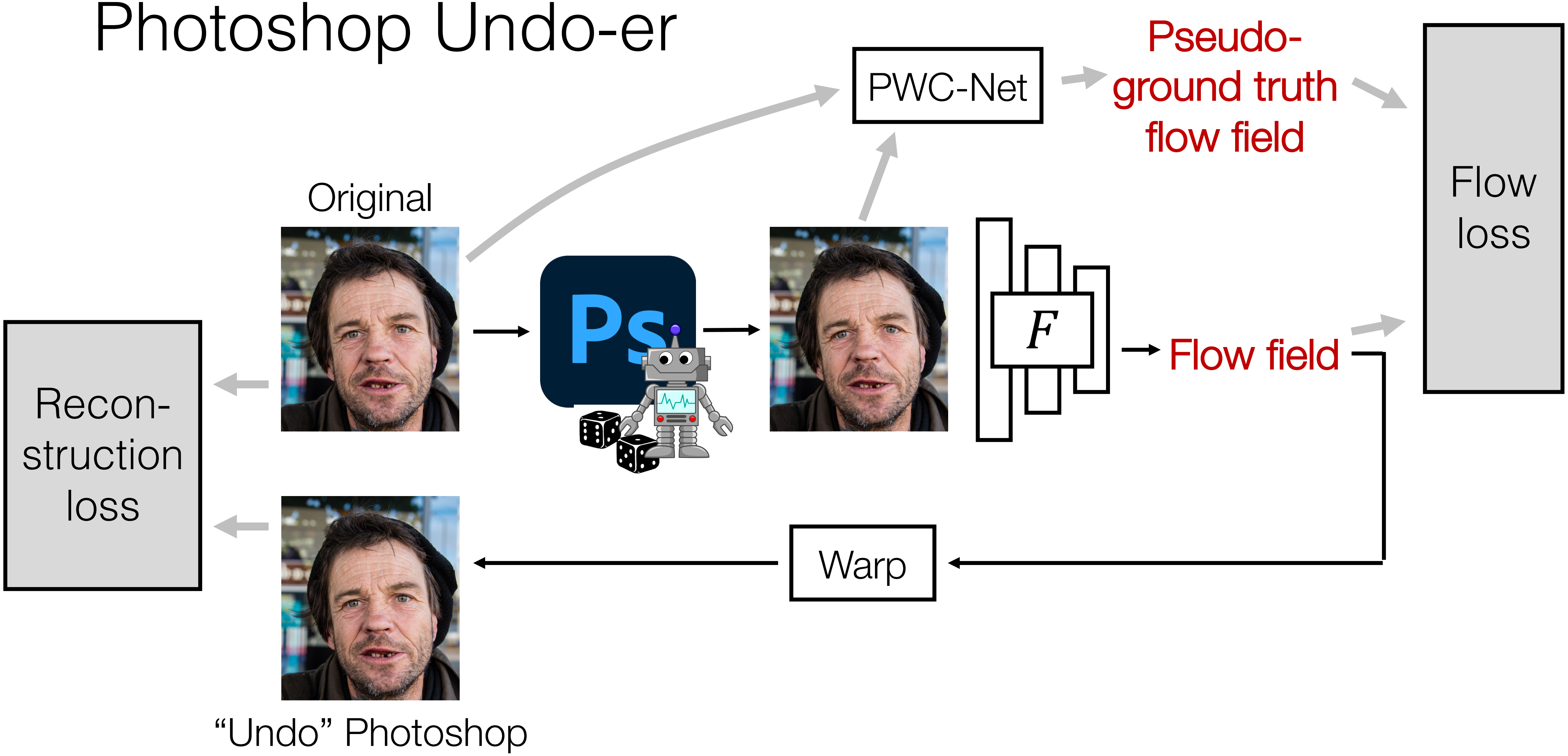
Photoshop Detector



Slides credit: Richard Zhang

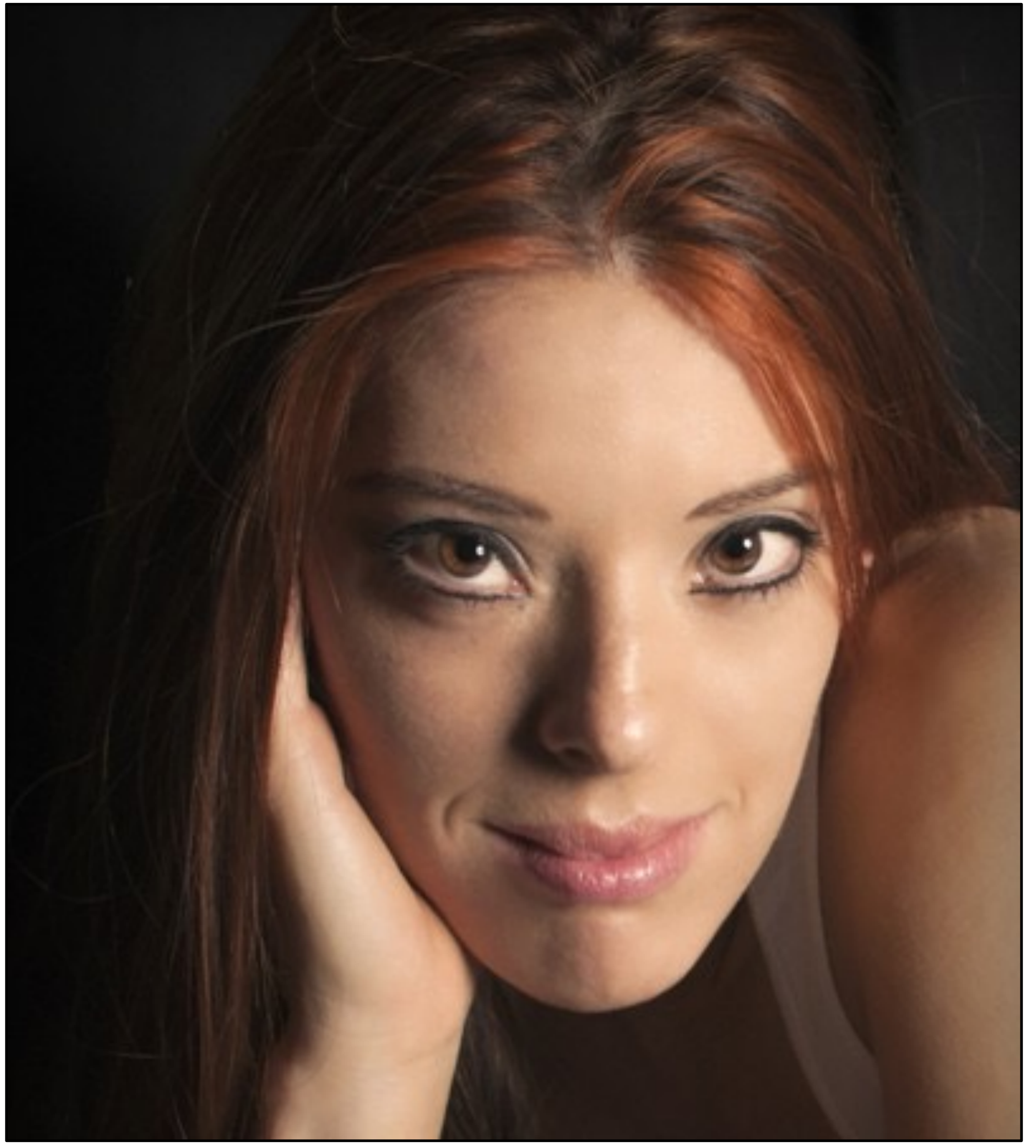
Detecting Photoshopped Faces by Scripting Photoshop [Wang et al., ICCV 2019]

Photoshop Undo-er



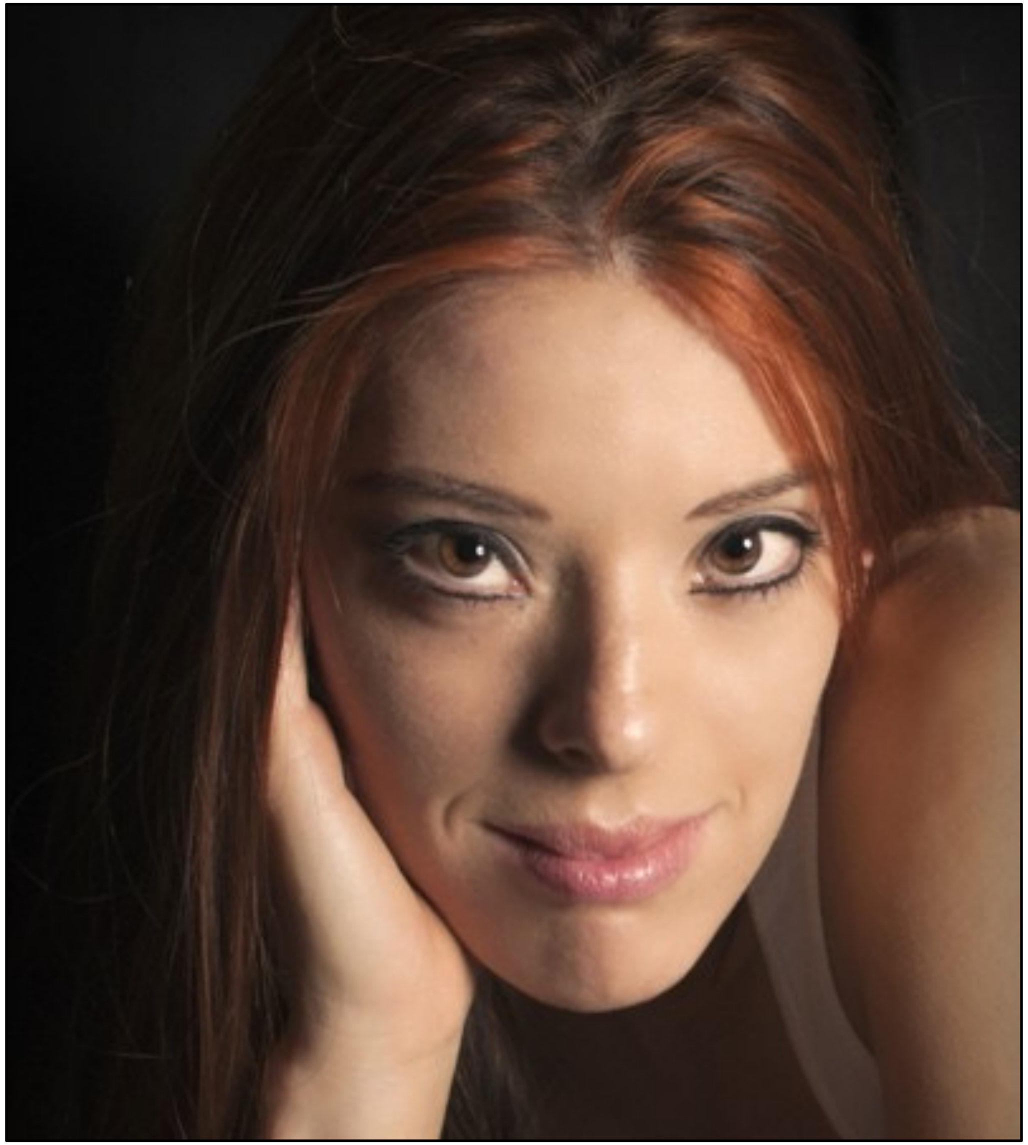
Slides credit: Richard Zhang

Detecting Photoshopped Faces by Scripting Photoshop [Wang et al., ICCV 2019]



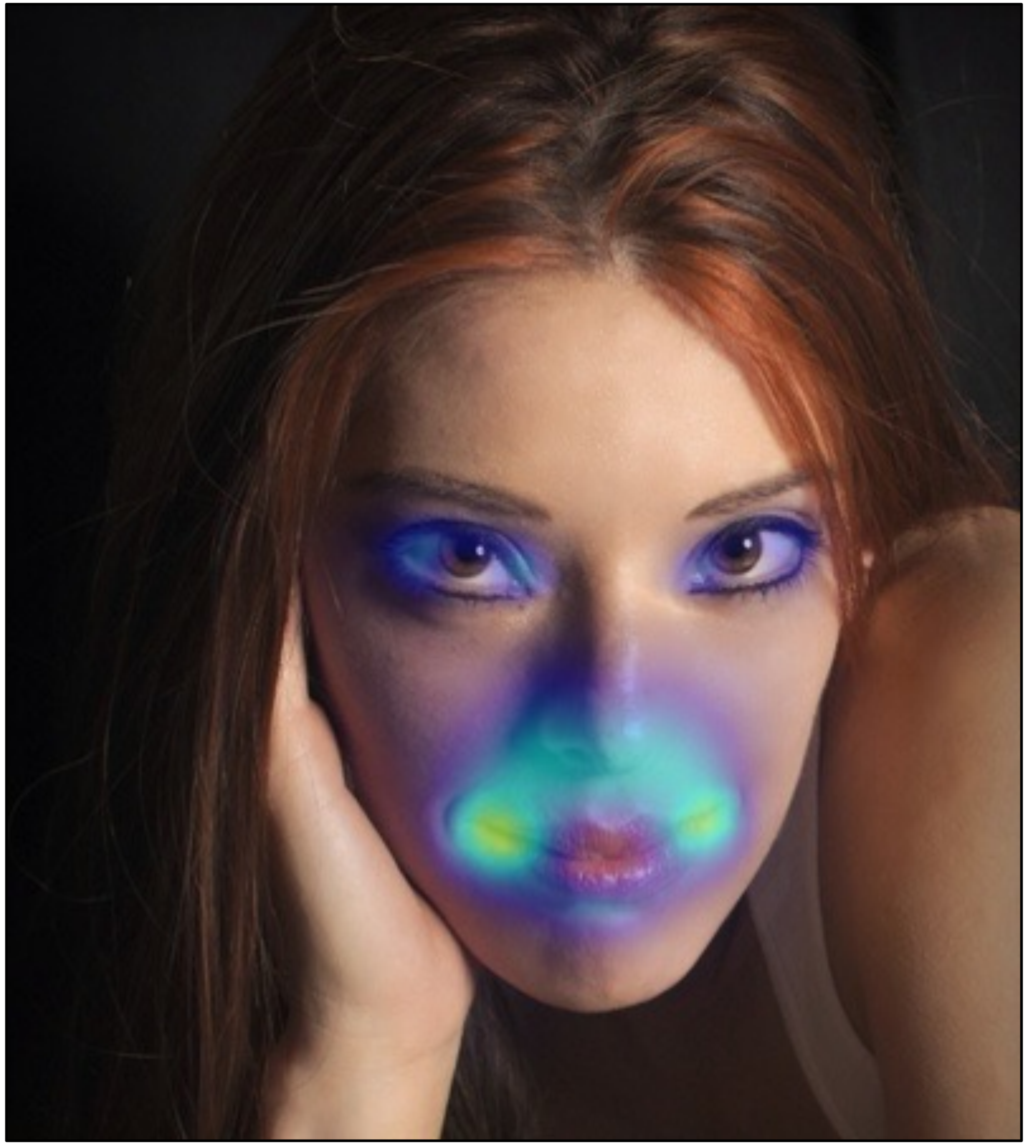
Manipulated?

Slides credit: Richard Zhang



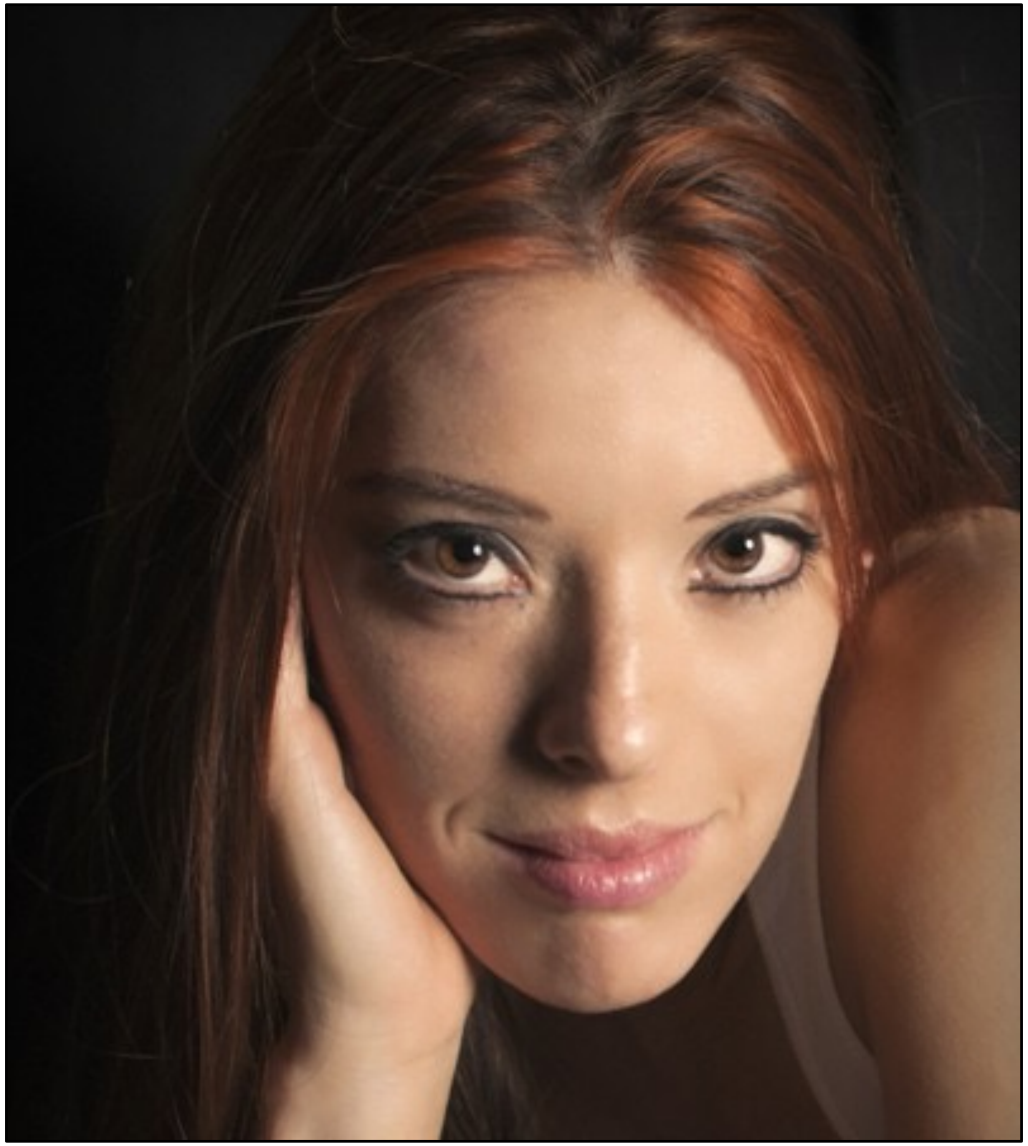
Manipulated? **Yes**

Slides credit: Richard Zhang



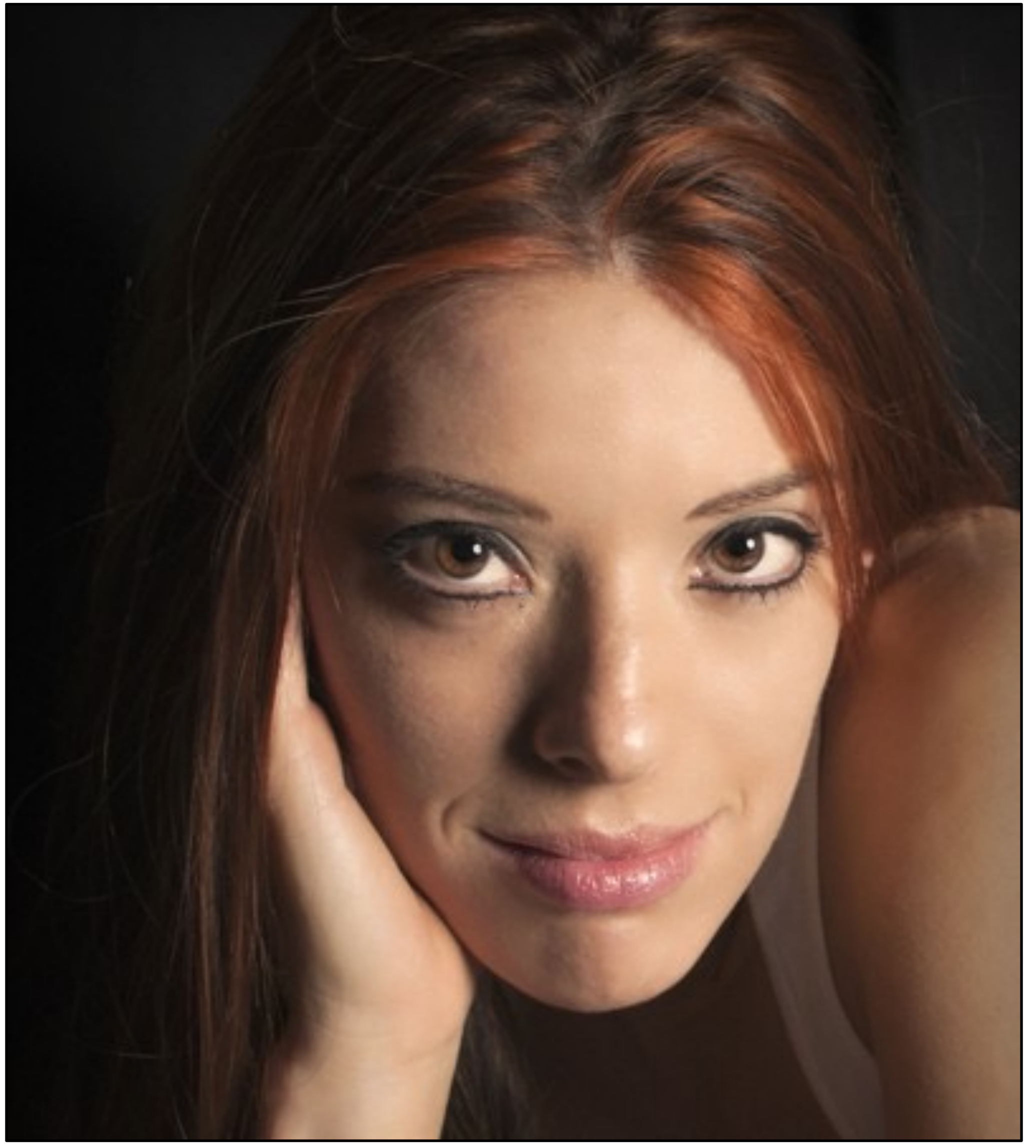
Flow prediction

Slides credit: Richard Zhang



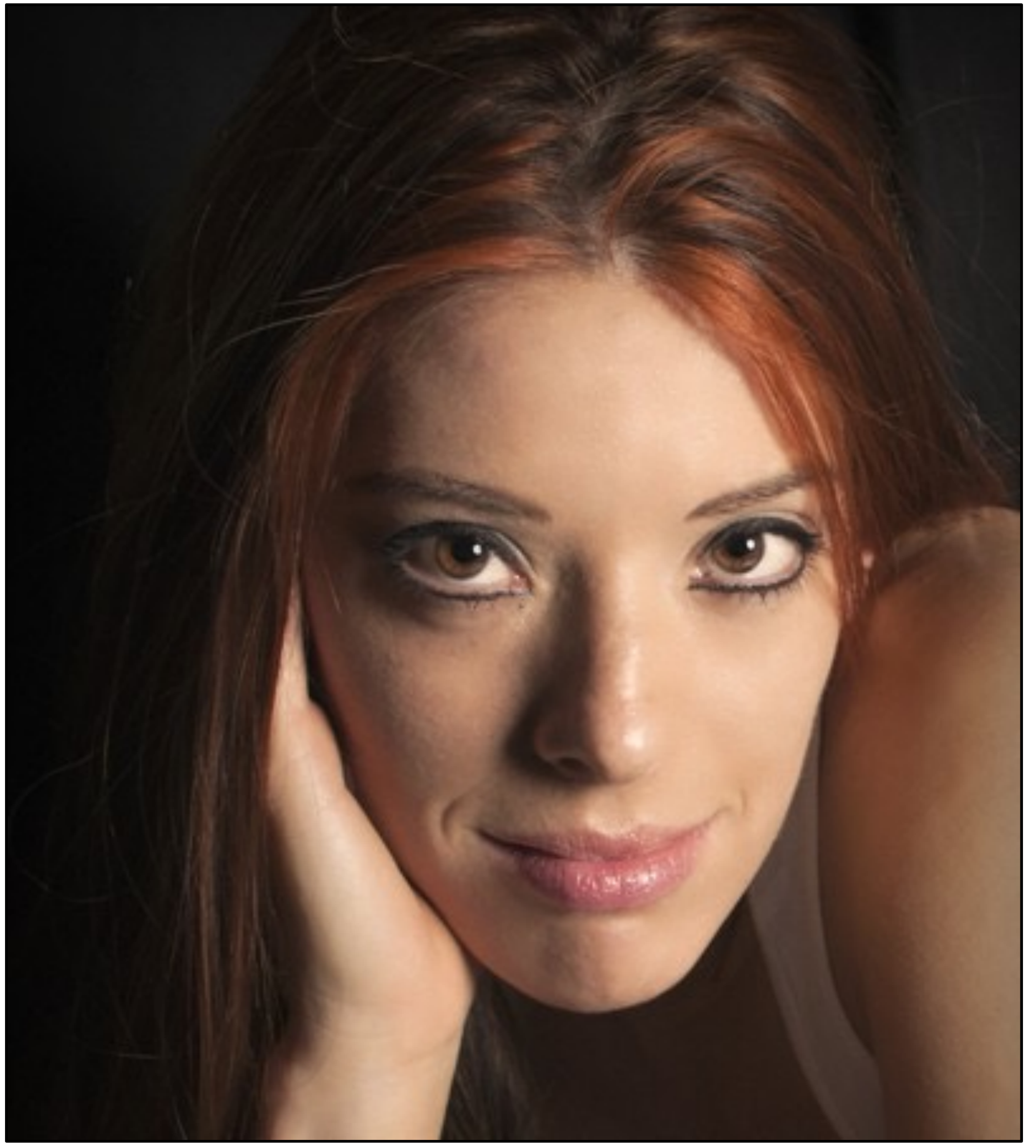
Suggested “undo”

Slides credit: Richard Zhang



Original

Slides credit: Richard Zhang

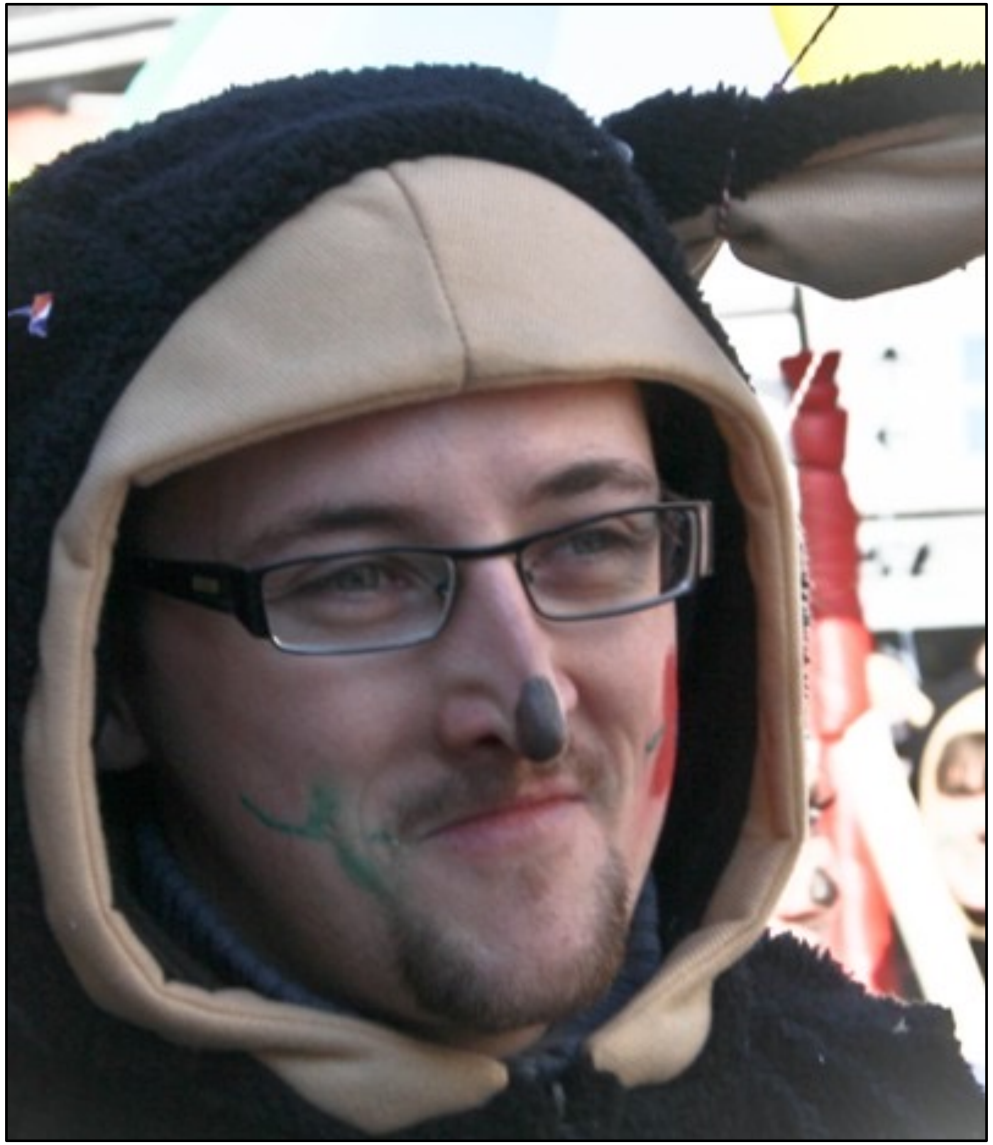


Manipulated vs. Original

Slides credit: Richard Zhang

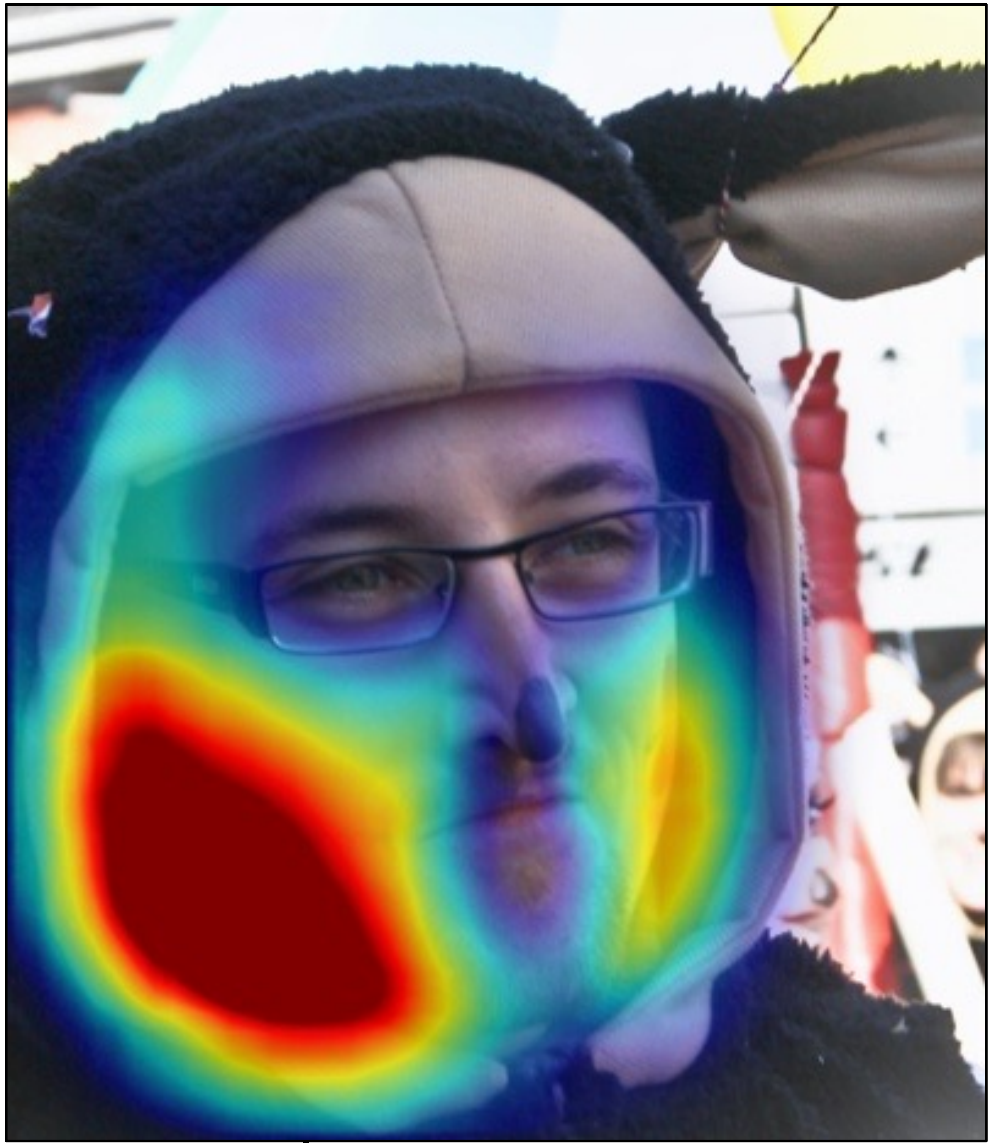
Undo vs. Original

Slides credit: Richard Zhang



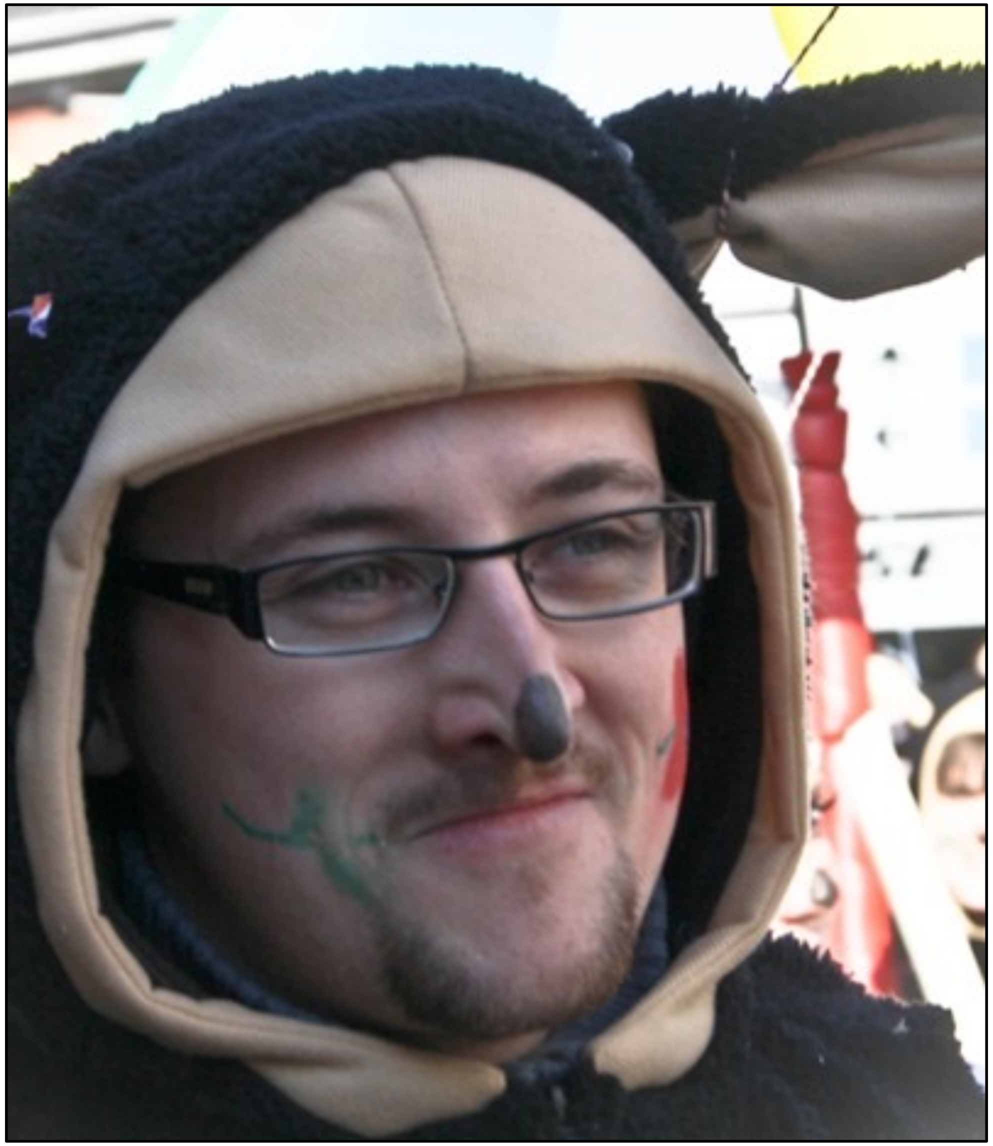
Manipulated

Slides credit: Richard Zhang



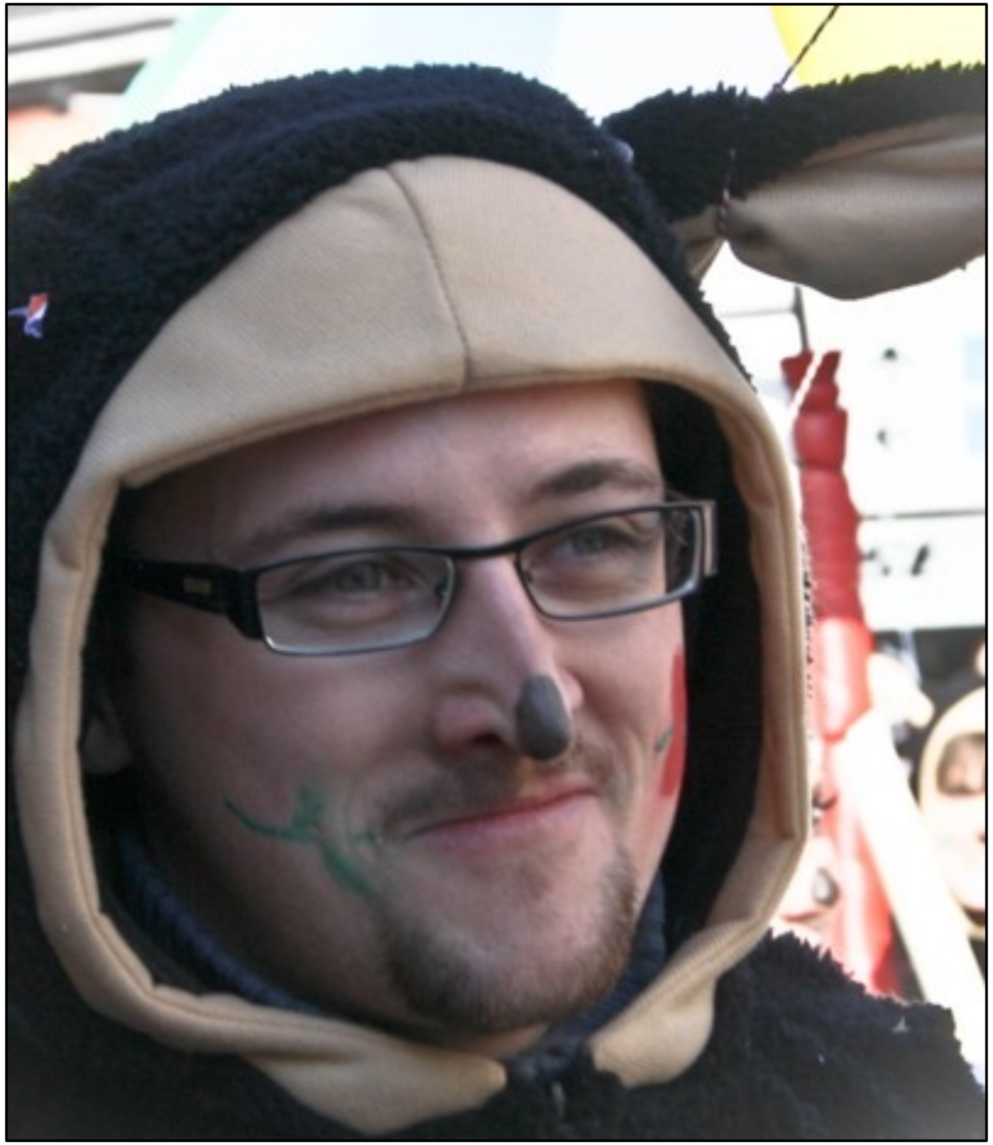
Flow prediction

Slides credit: Richard Zhang



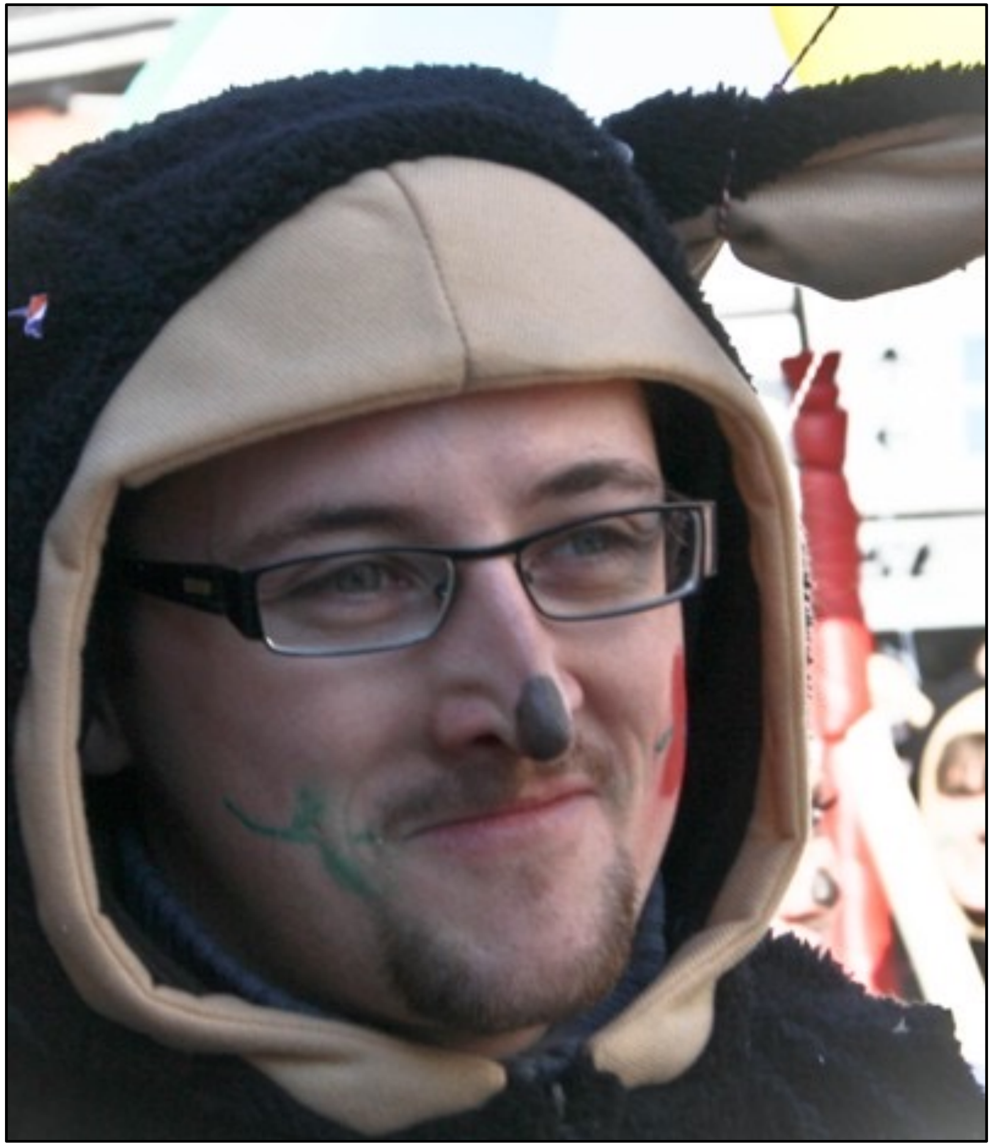
Suggested “undo”

Slides credit: Richard Zhang



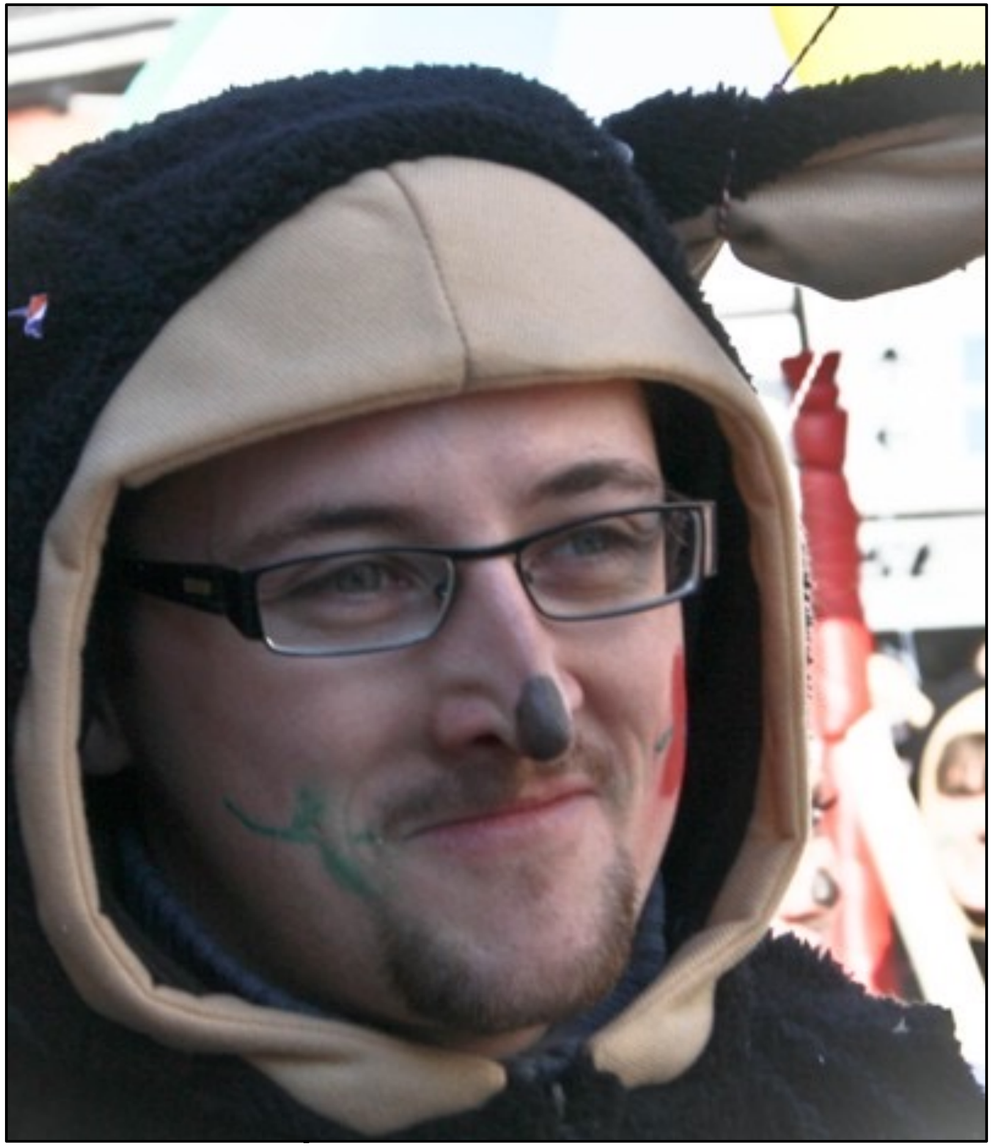
Original

Slides credit: Richard Zhang



Manipulated vs. Original

Slides credit: Richard Zhang



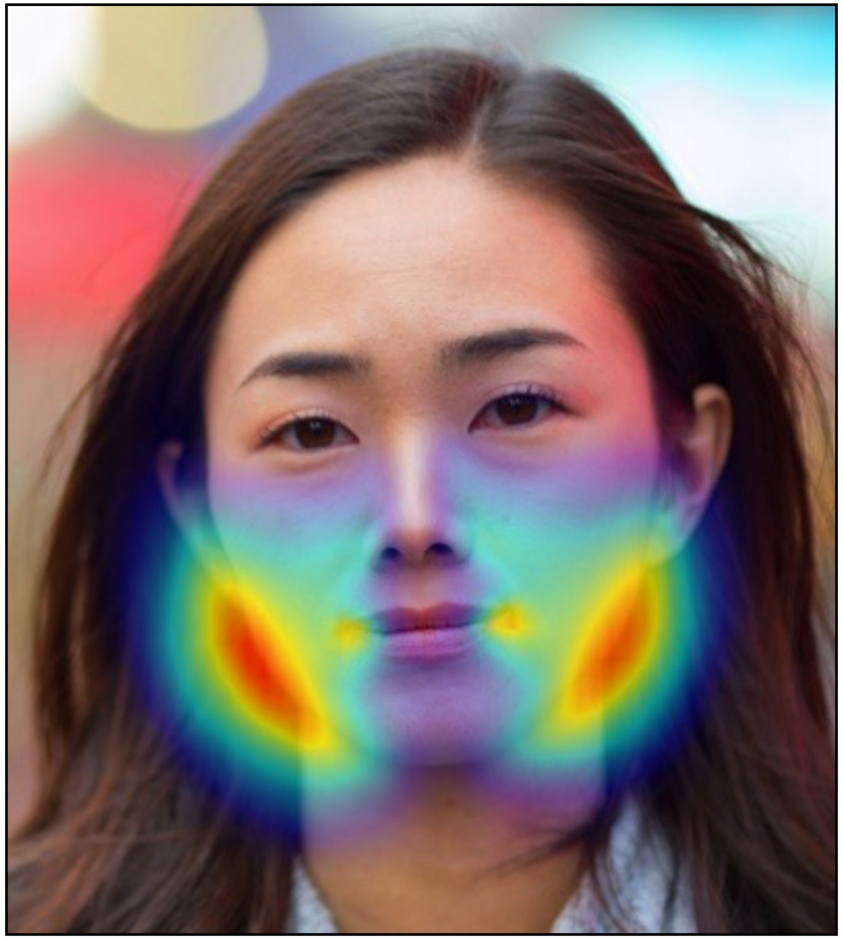
Undo vs. Original

Artist-generated example

Manipulated

Slides credit: Richard Zhang

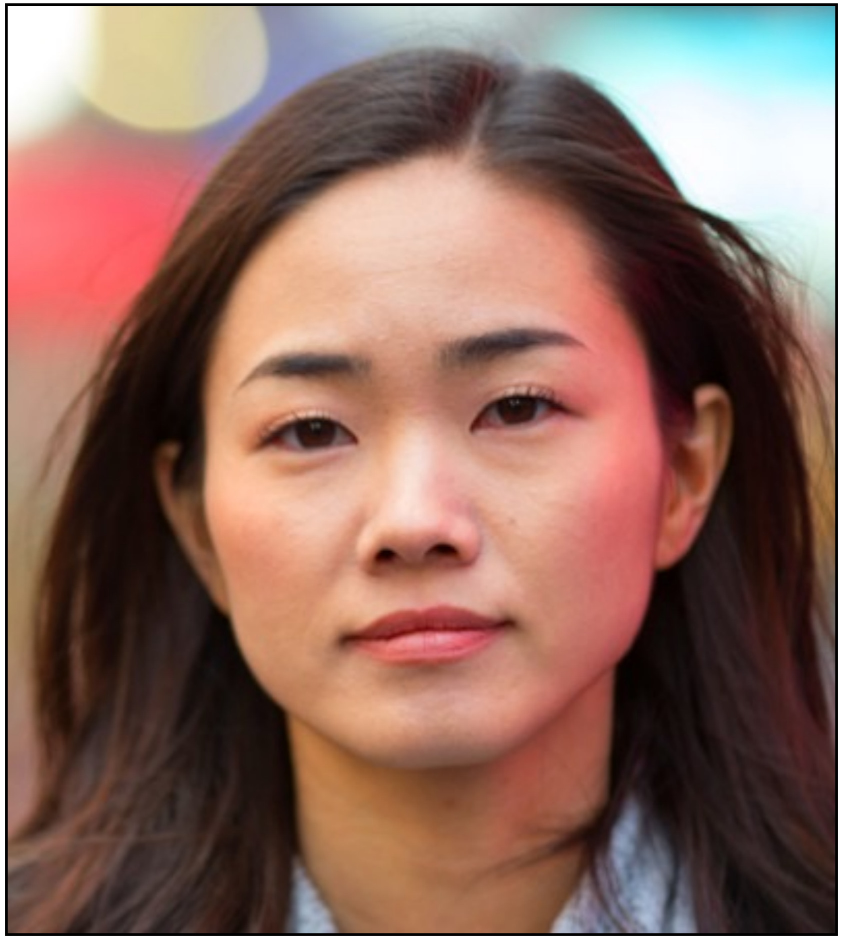
Artist-generated example



Flow prediction

Slides credit: Richard Zhang

Artist-generated example



Suggested “undo”

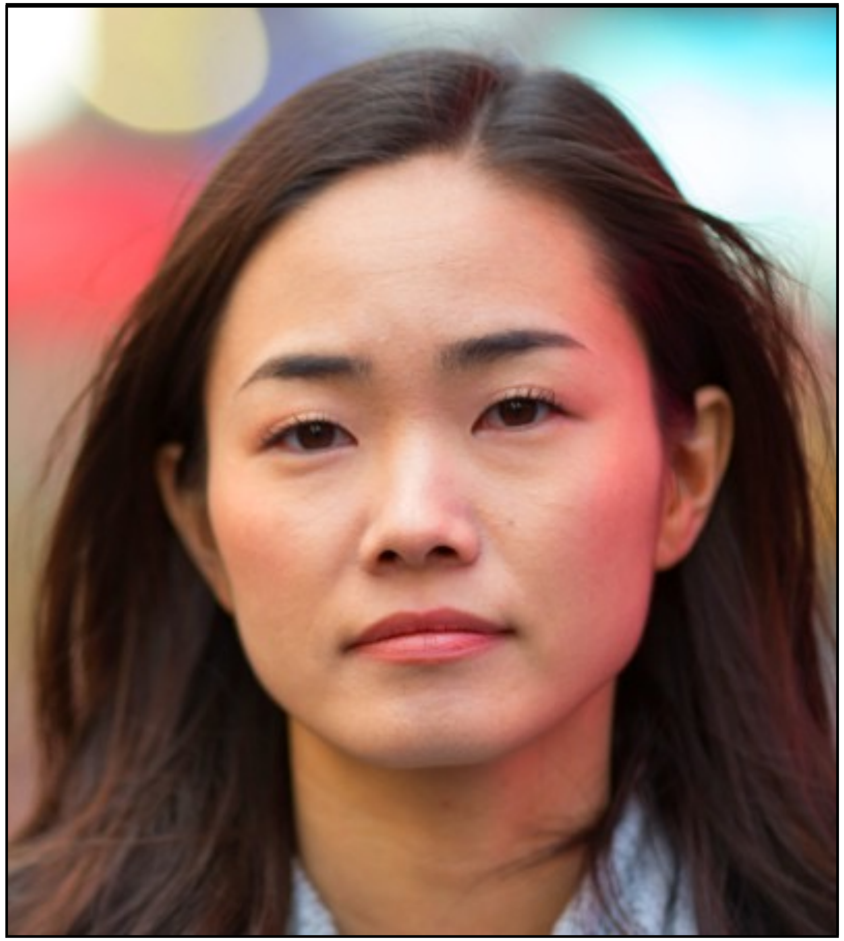
Slides credit: Richard Zhang

Artist-generated example

Original photo

Slides credit: Richard Zhang

Artist-generated example



Manipulated vs. Original

Slides credit: Richard Zhang

Artist-generated example

Undo vs. Original

Slides credit: Richard Zhang

Senses of generalization

- Post-processing
- Heldout artist data
- Different warp domain
- Different image domains

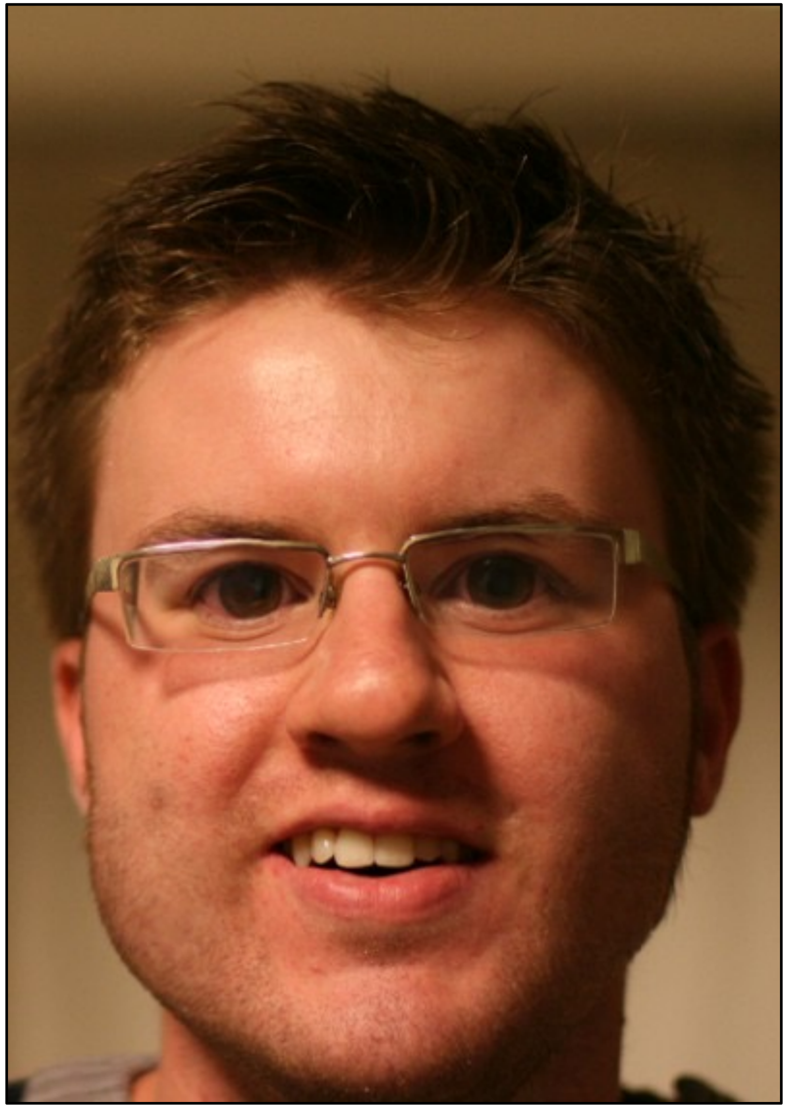
Slides credit: Richard Zhang

Snapchat warps

Original Photo

Slides credit: Richard Zhang

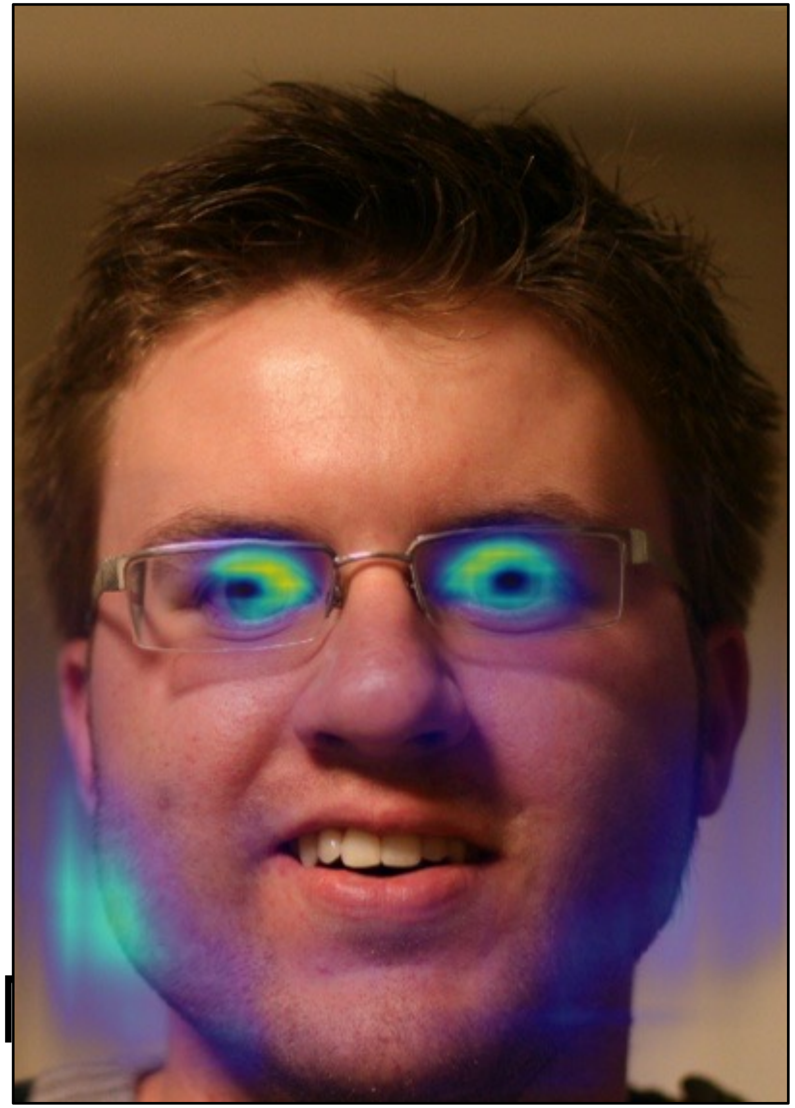
Snapchat warps



Manipulated Photo

Slides credit: Richard Zhang

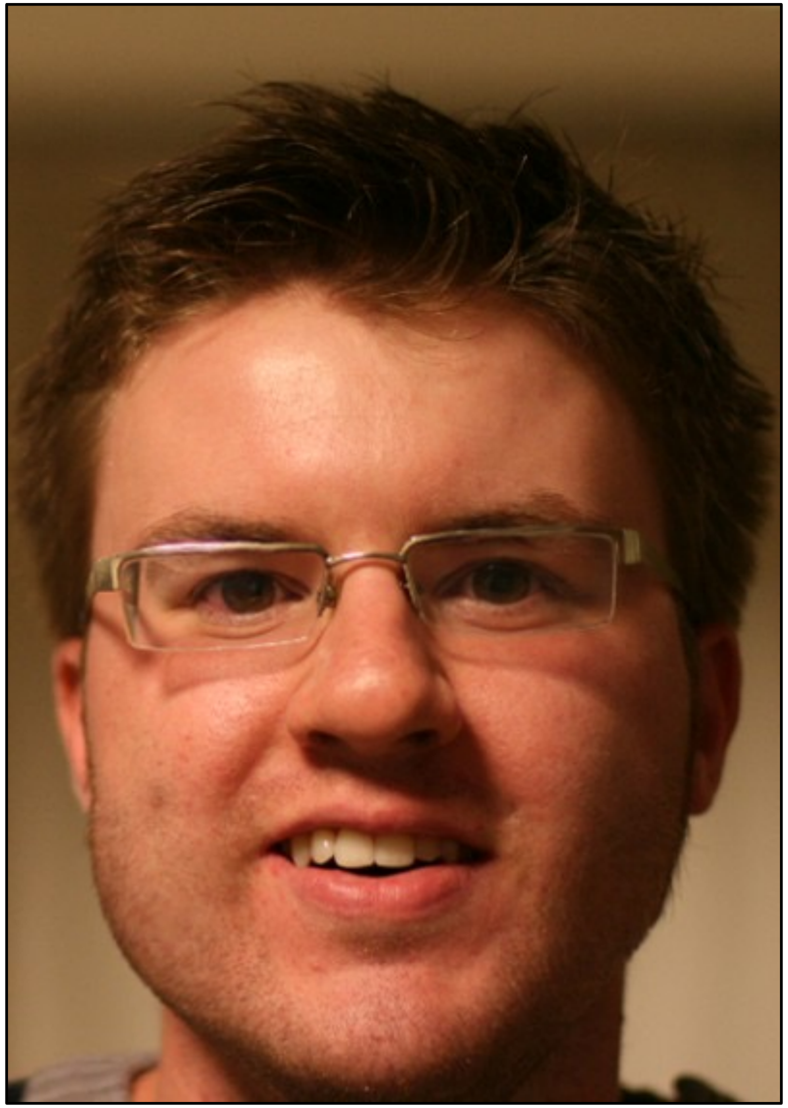
Snapchat warps



Flow Prediction

Slides credit: Richard Zhang

Snapchat warps



Suggested “Undo”

Slides credit: Richard Zhang

Snapchat warps

Some generalization across warp methods

Original Photo

Slides credit: Richard Zhang

Different image domain

Slides credit: Richard Zhang

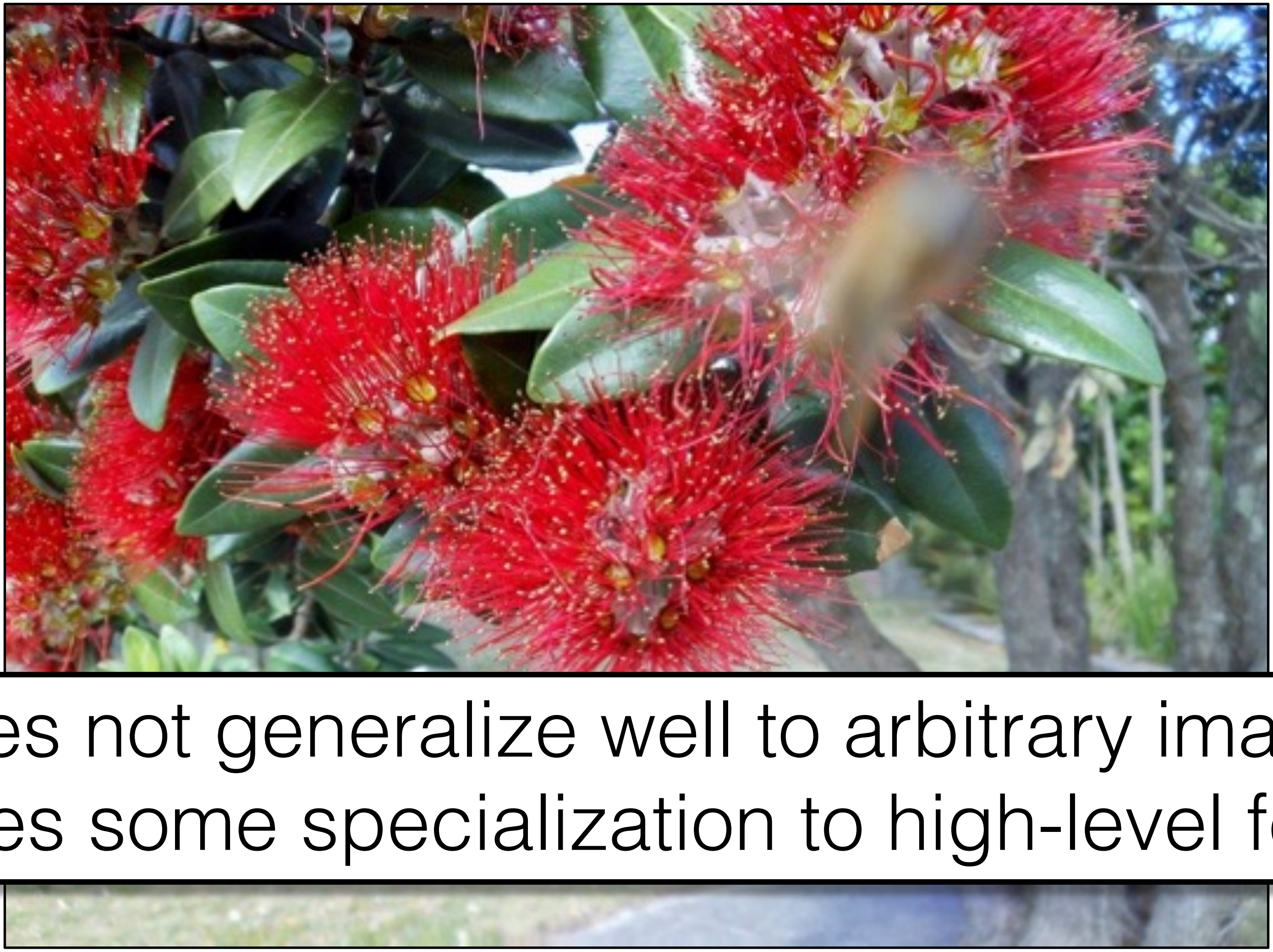
Different image domain

Slides credit: Richard Zhang

Different image domain

Slides credit: Richard Zhang

Predicted warp (not successful)



Does not generalize well to arbitrary image;
Indicates some specialization to high-level features

Slides credit: Richard Zhang

Discussion

- Given a relatively static tool, directly specialize
- Representation learns a combination of low and high-level cues
- Data augmentation helps generalization

Slides credit: Richard Zhang

Detect Deep Fake

Making fake images is getting easier

“Deepfakes”

GANs

Can we create a “universal” detector?

Slides credit: Richard Zhang

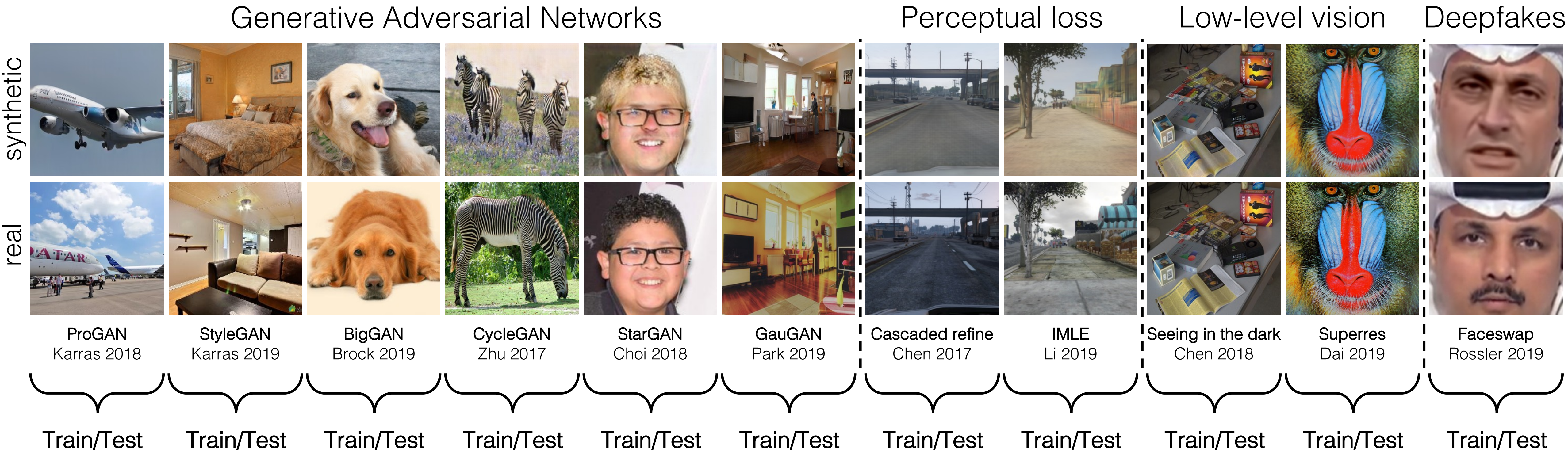
DeepFakes (https://github.com/deepfakes/faceswap)

Face2Face (Thies et al. 2016)

Lip-syncing Obama (Suwajanakorn et al. 2017)

Slides credit: Richard Zhang

Dataset of CNN-generated fakes



Generative Adversarial Networks

Perceptual loss

Low-level vision

Deepfakes

synthetic

real

ProGAN
Karras 2018

StyleGAN
Karras 2019

BigGAN
Brock 2019

CycleGAN
Zhu 2017

StarGAN
Choi 2018

GauGAN
Park 2019

Cascaded refine
Chen 2017

IMLE
Li 2019

Seeing in the dark
Chen 2018

Superres
Dai 2019

Faceswap
Rossler 2019

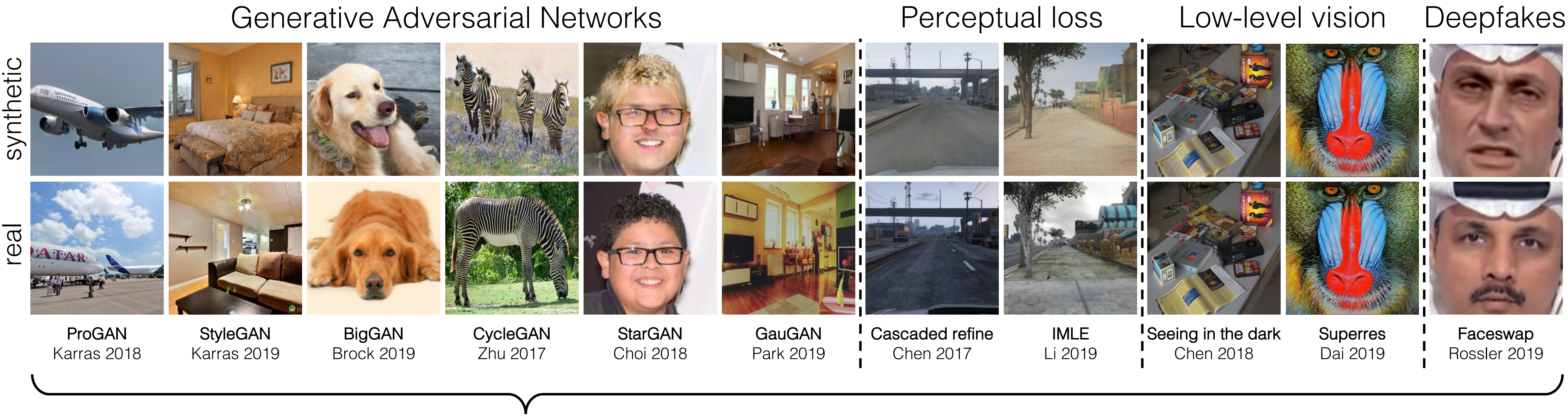
Train/Test Train/Test Train/Test Train/Test Train/Test Train/Test Train/Test Train/Test Train/Test Train/Test Train/Test

Test: ??? et al. 20XX

Can we create a “universal” detector?

Slides credit: Richard Zhang

Dataset of CNN-generated fakes

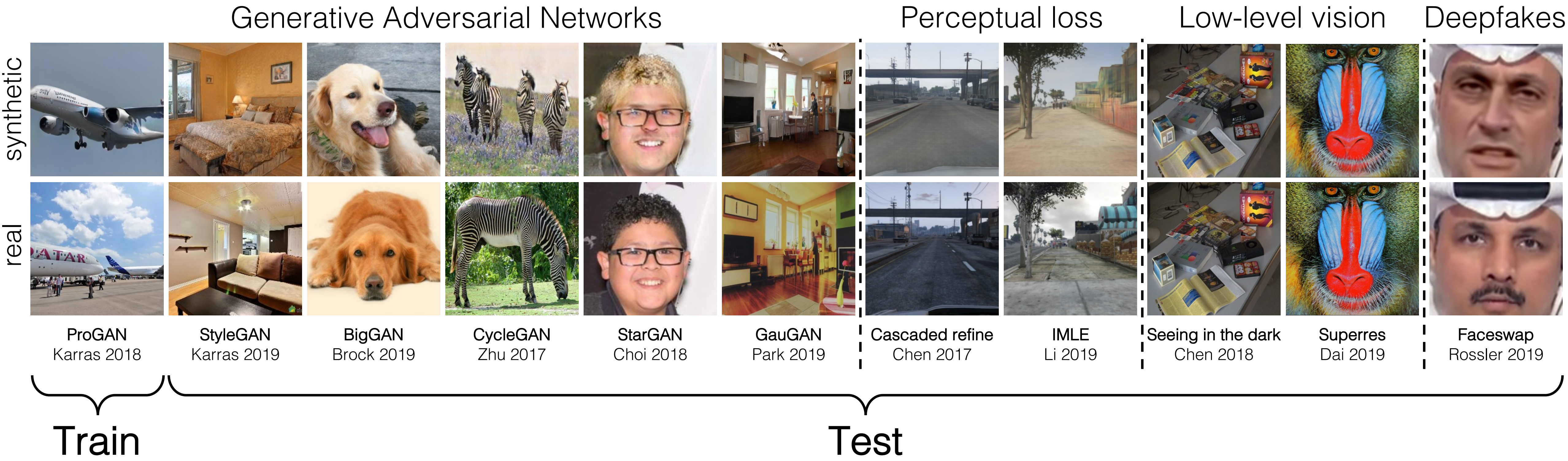


Train ⟶ **Test:** ??? et al. 20XX

Can we create a “universal” detector?

Slides credit: Richard Zhang

Dataset of CNN-generated fakes

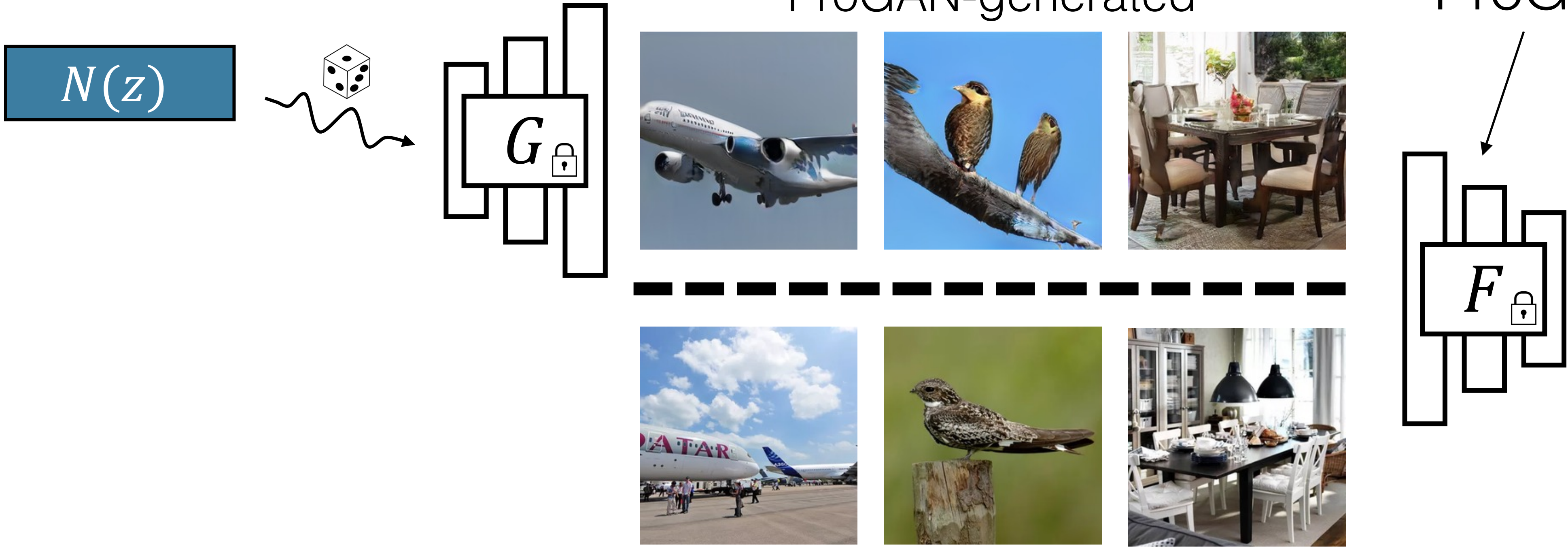


Many **differences** (architecture, dataset, objective)

Training on ProGAN

ProGAN-generated

ProGAN detector



Real vs. fake

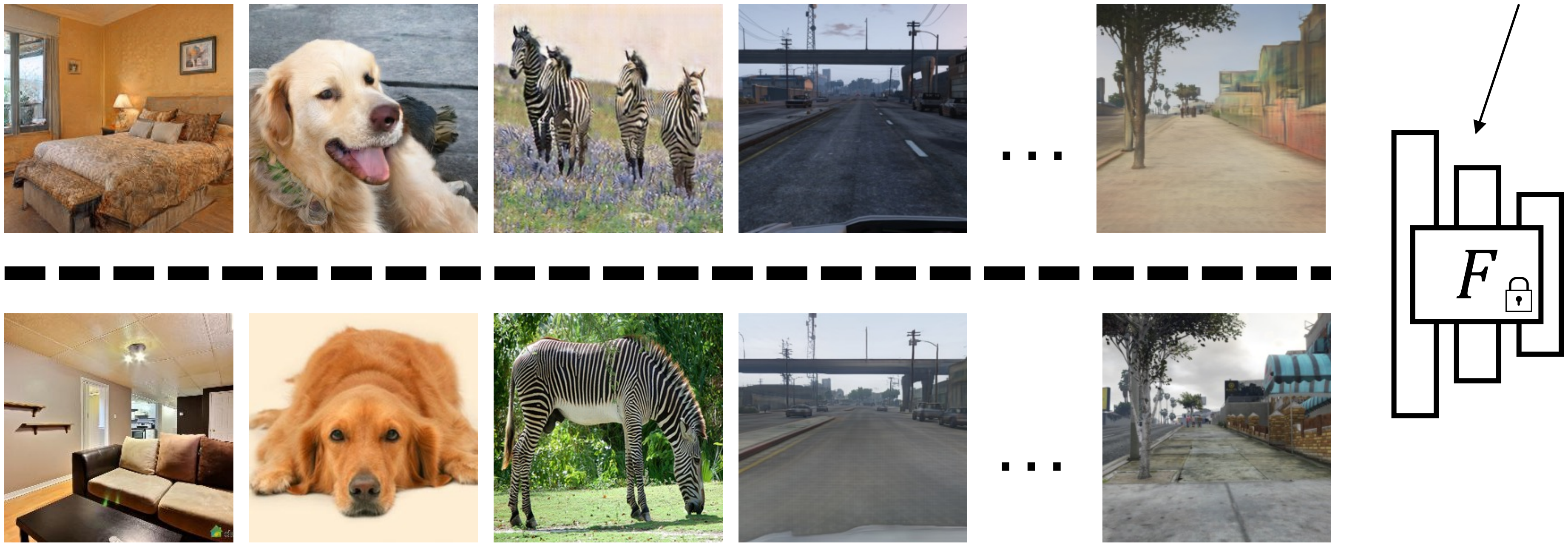
720K real images, 20 categories from LSUN

CNN-generated images are surprisingly easy to spot... for now [Wang et al., CVPR 2020]

Testing across architectures

Synthesized images from **other** CNNs

ProGAN detector

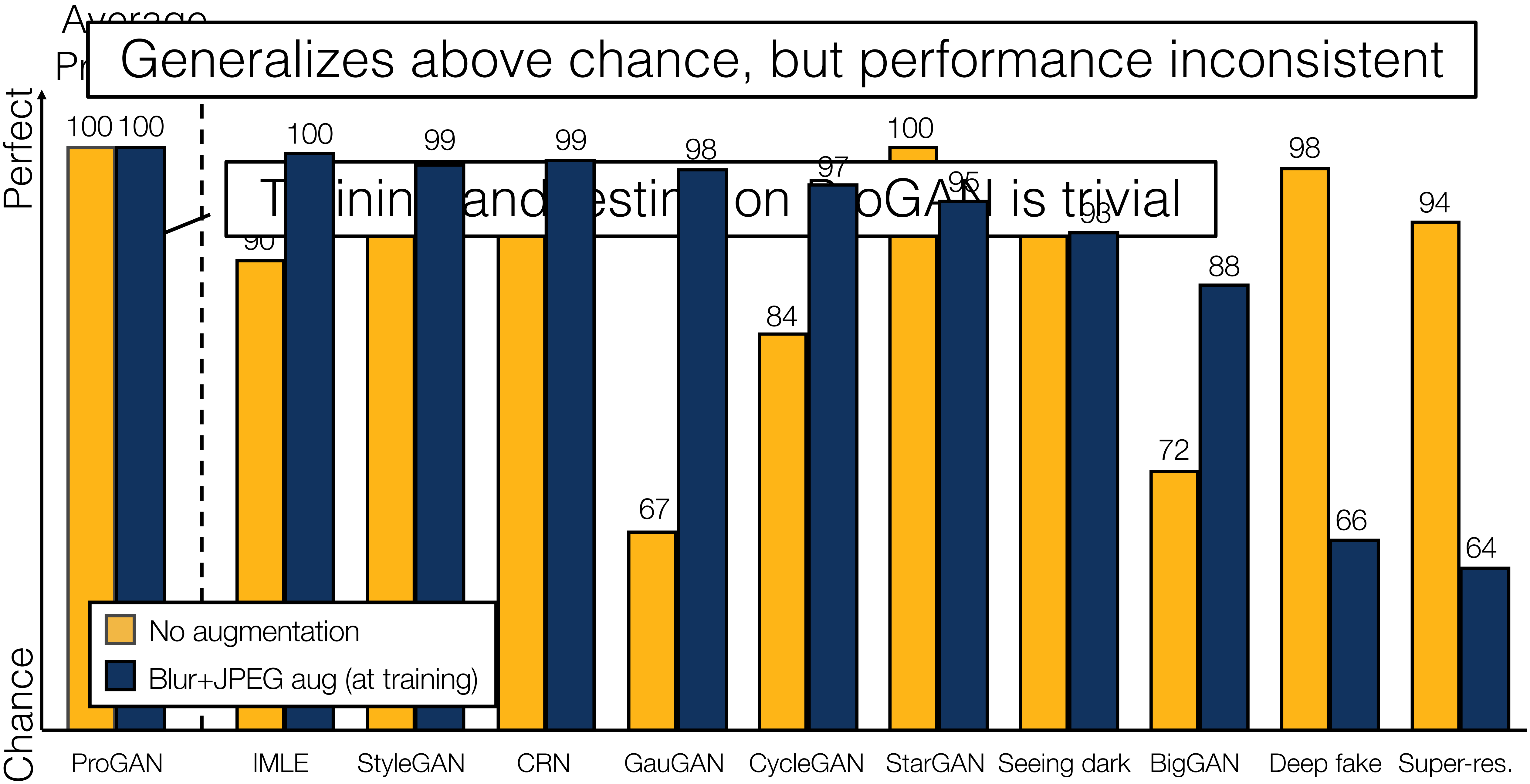


Real vs. fake

Real images

CNN-generated images are surprisingly easy to spot... for now [Wang et al., CVPR 2020]

Generalizes above chance, but performance inconsistent

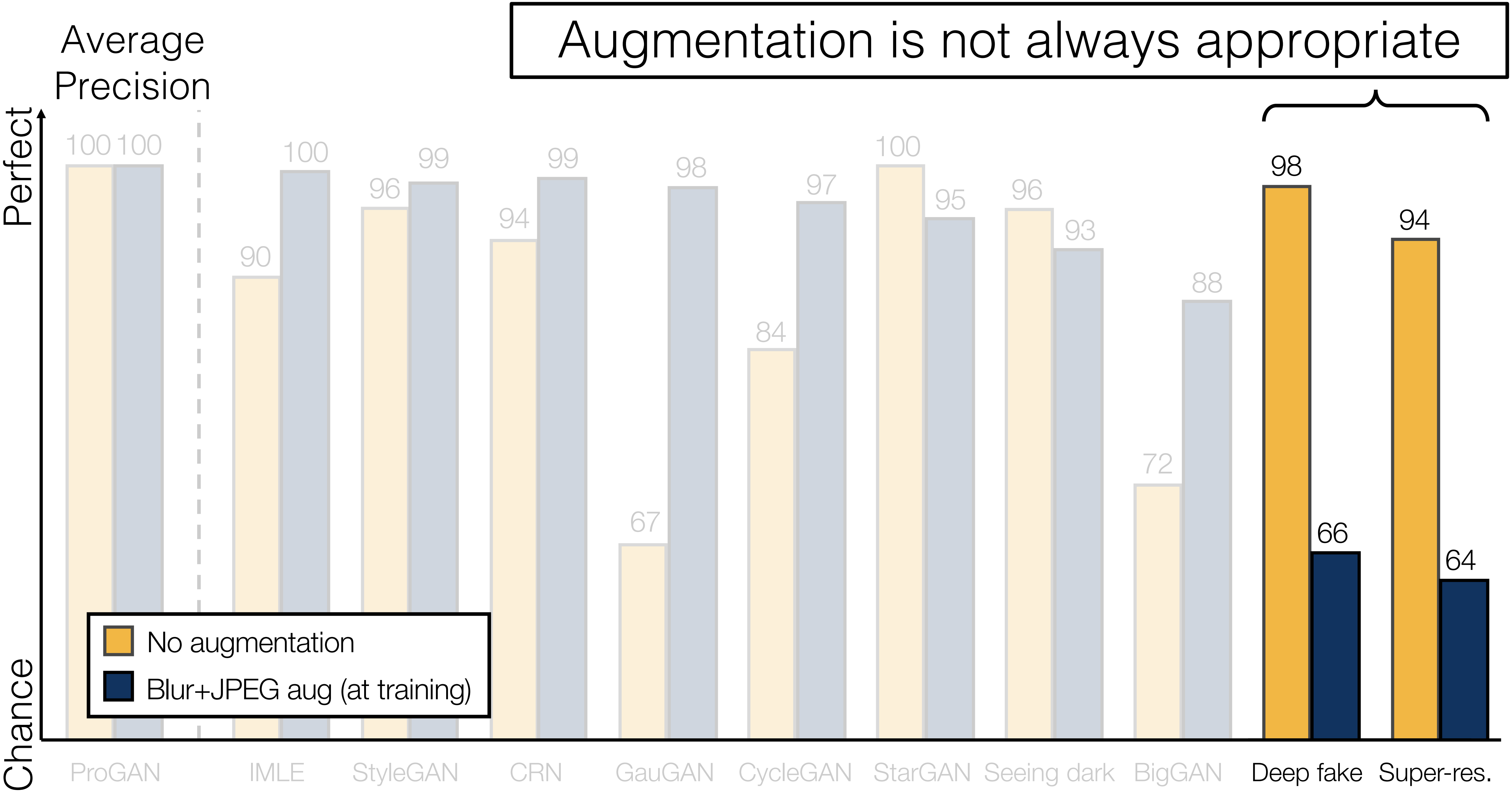


Training and testing on ProGAN is trivial

Slides credit: Richard Zhang

[Wang et al., CVPR 2020]

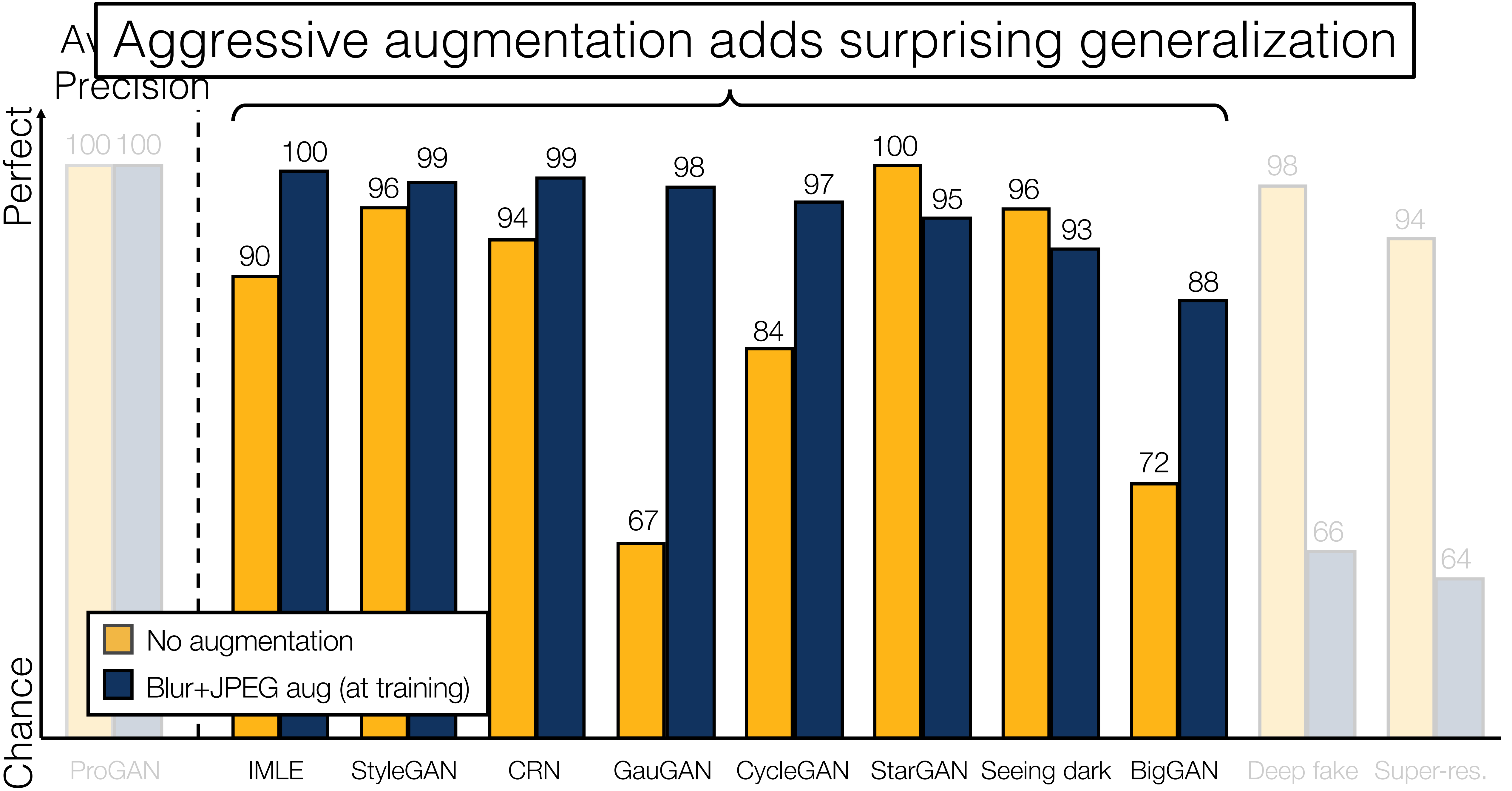
Augmentation is not always appropriate



Slides credit: Richard Zhang

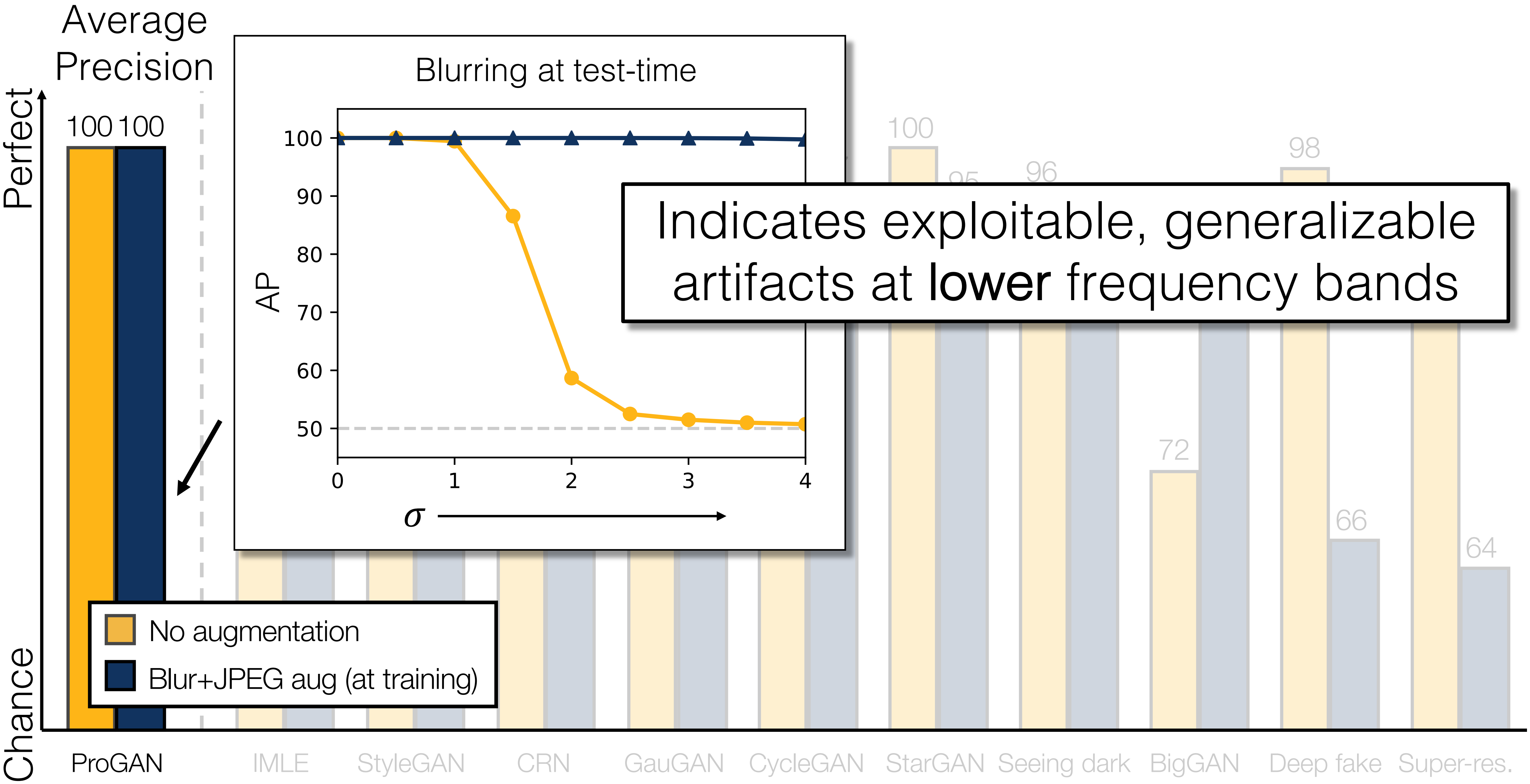
[Wang et al., CVPR 2020]

Aggressive augmentation adds surprising generalization



Slides credit: Richard Zhang

[Wang et al., CVPR 2020]



Indicates exploitable, generalizable artifacts at **lower** frequency bands

Slides credit: Richard Zhang

[Wang et al., CVPR 2020]

Discussion

- Suggests CNN-generated images have common artifacts
- Artifacts can be detected by a simple classifier!
  - StyleGAN2 (released after the paper): 100% AP on FFHQ
  - Maybe generalizes beyond CNNs [Chai et al. ECCV 2020]
  - Note: AP is computed on a collection of images;
    a real/fake decision on a per-image basis is more difficult
- Situation may not persist
  - GANs train with a discriminator
  - Future architecture changes (does not generalize well to Diffusion models)

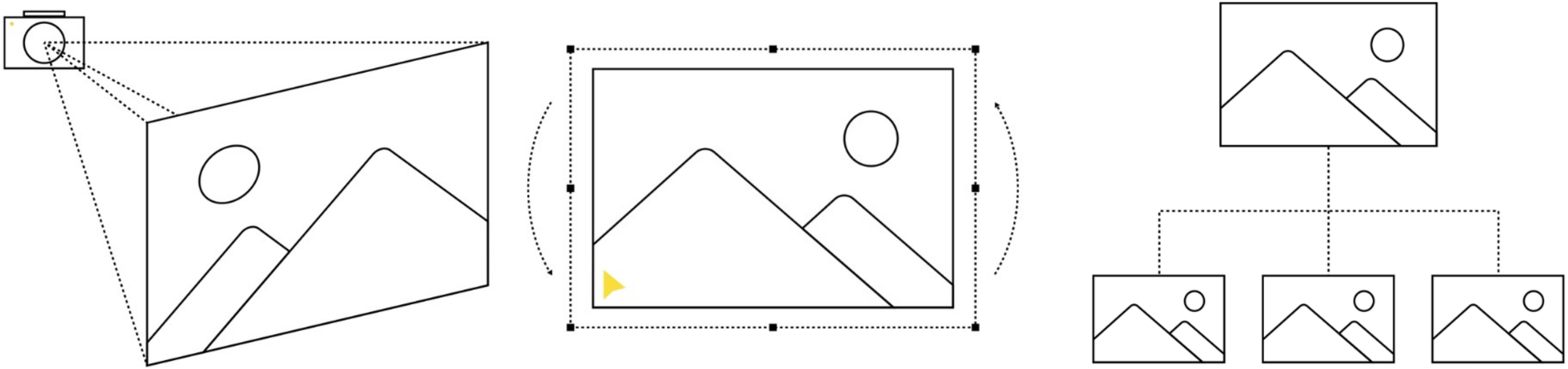
Slides credit: Richard Zhang

Discussion

- Suggests a multi-prong approach
  - For rapidly evolving tools, continuously training and generalize
  - For relatively static tools, specialize
- Synthesis and manipulations for creative uses
- Detection is only a piece of the puzzle
  - e.g., Content Authenticity Initiative: https://contentauthenticity.org/, collaboration between Adobe, New York Times, and Twitter

Slides credit: Richard Zhang

Content Authenticity (prove what is real)



Capture

Editing

Publishing and sharing

https://contentauthenticity.org/

Copyrights

+ Disclaimer: I am not a lawyer

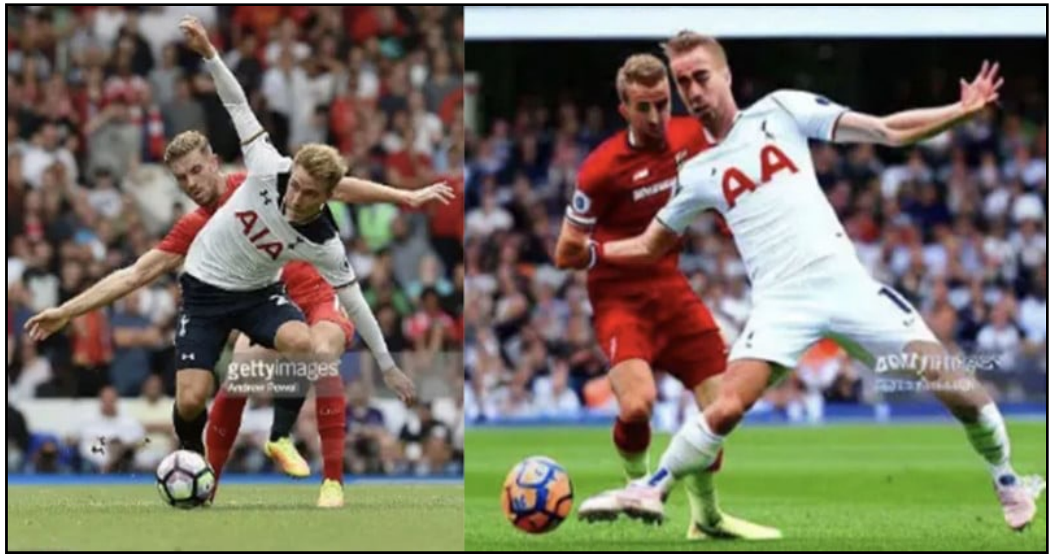
+ Human creator’s rights

+ Diverse opinions

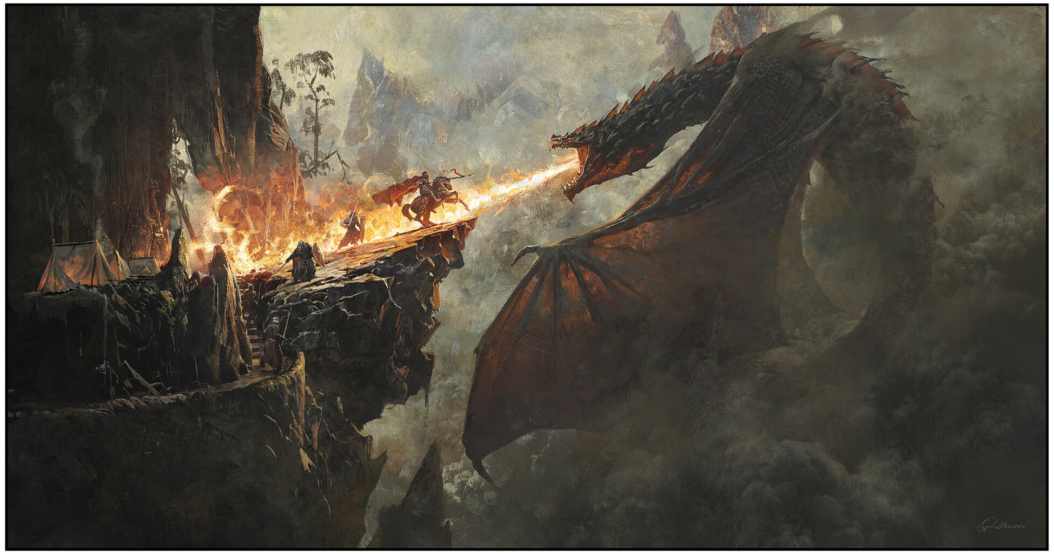
+ Evolving landscape

Copyrighted content?

- Copyrighted images
- Company IPs / logos
- Artist styles of living artists

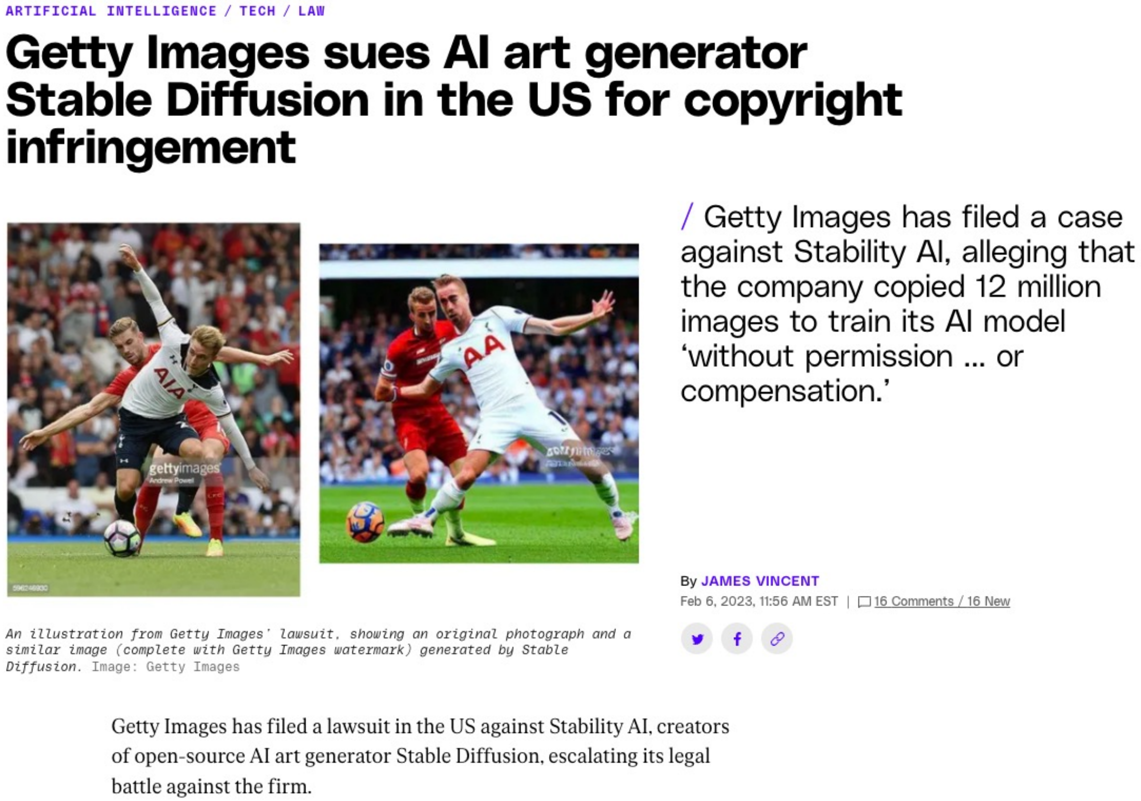


Getty Images



Greg Rutkowski

Ongoing Legal Battles



ARTIFICIAL INTELLIGENCE / TECH / LAW

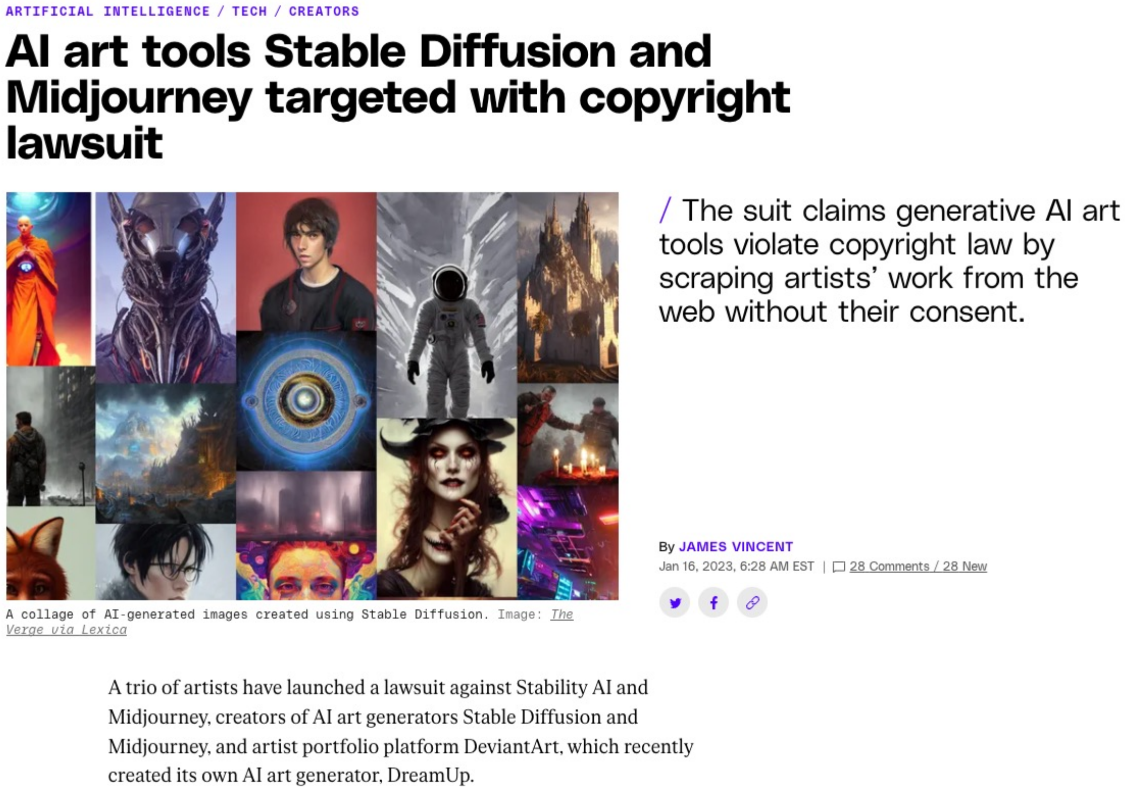
Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement

/ Getty Images has filed a case against Stability AI, alleging that the company copied 12 million images to train its AI model 'without permission ... or compensation.'

By JAMES VINCENT
Feb 6, 2023, 11:56 AM EST | 16 Comments / 16 New

An illustration from Getty Images' lawsuit, showing an original photograph and a similar image (complete with Getty Images watermark) generated by Stable Diffusion. Image: Getty Images

Getty Images has filed a lawsuit in the US against Stability AI, creators of open-source AI art generator Stable Diffusion, escalating its legal battle against the firm.



ARTIFICIAL INTELLIGENCE / TECH / CREATORS

AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit

/ The suit claims generative AI art tools violate copyright law by scraping artists' work from the web without their consent.

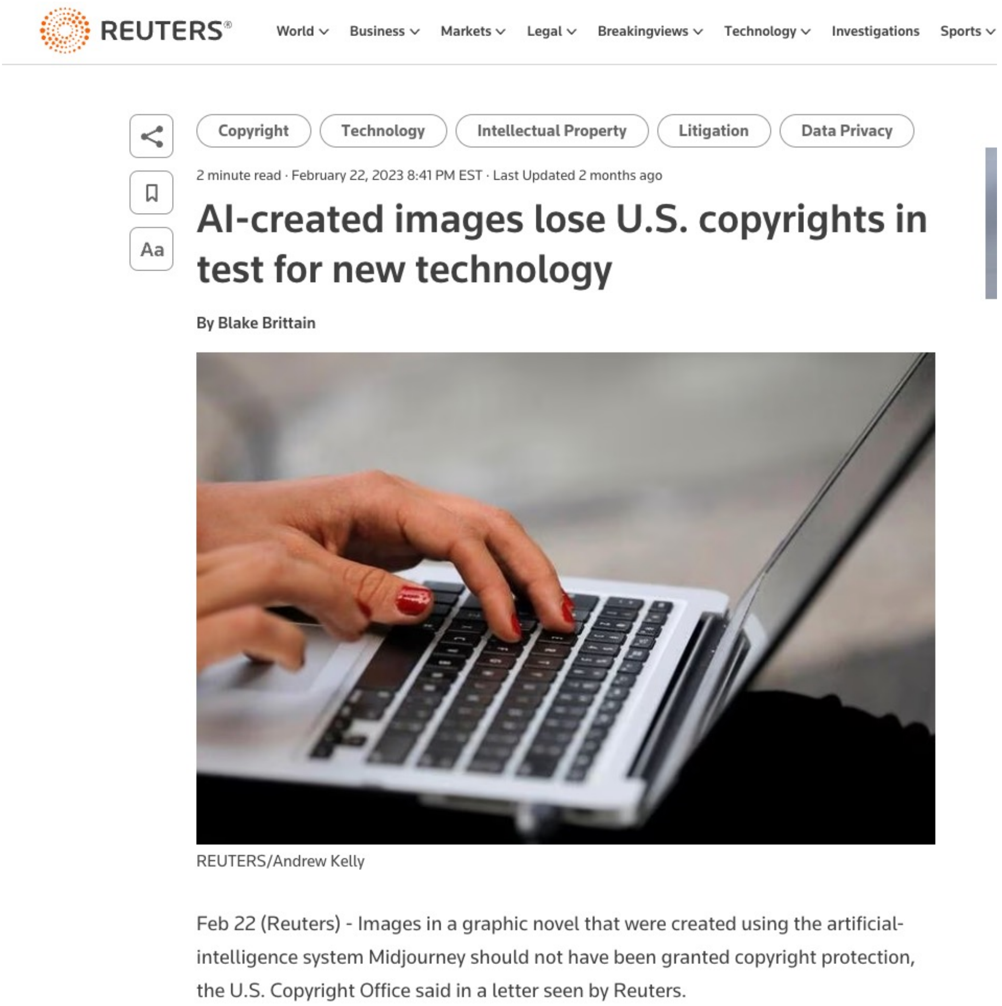
By JAMES VINCENT
Jan 16, 2023, 6:28 AM EST | 28 Comments / 28 New

A collage of AI-generated images created using Stable Diffusion. Image: *The Verge via Lexica*

A trio of artists have launched a lawsuit against Stability AI and Midjourney, creators of AI art generators Stable Diffusion and Midjourney, and artist portfolio platform DeviantArt, which recently created its own AI art generator, DreamUp.

Source: The Verge

Ongoing Legal Battles



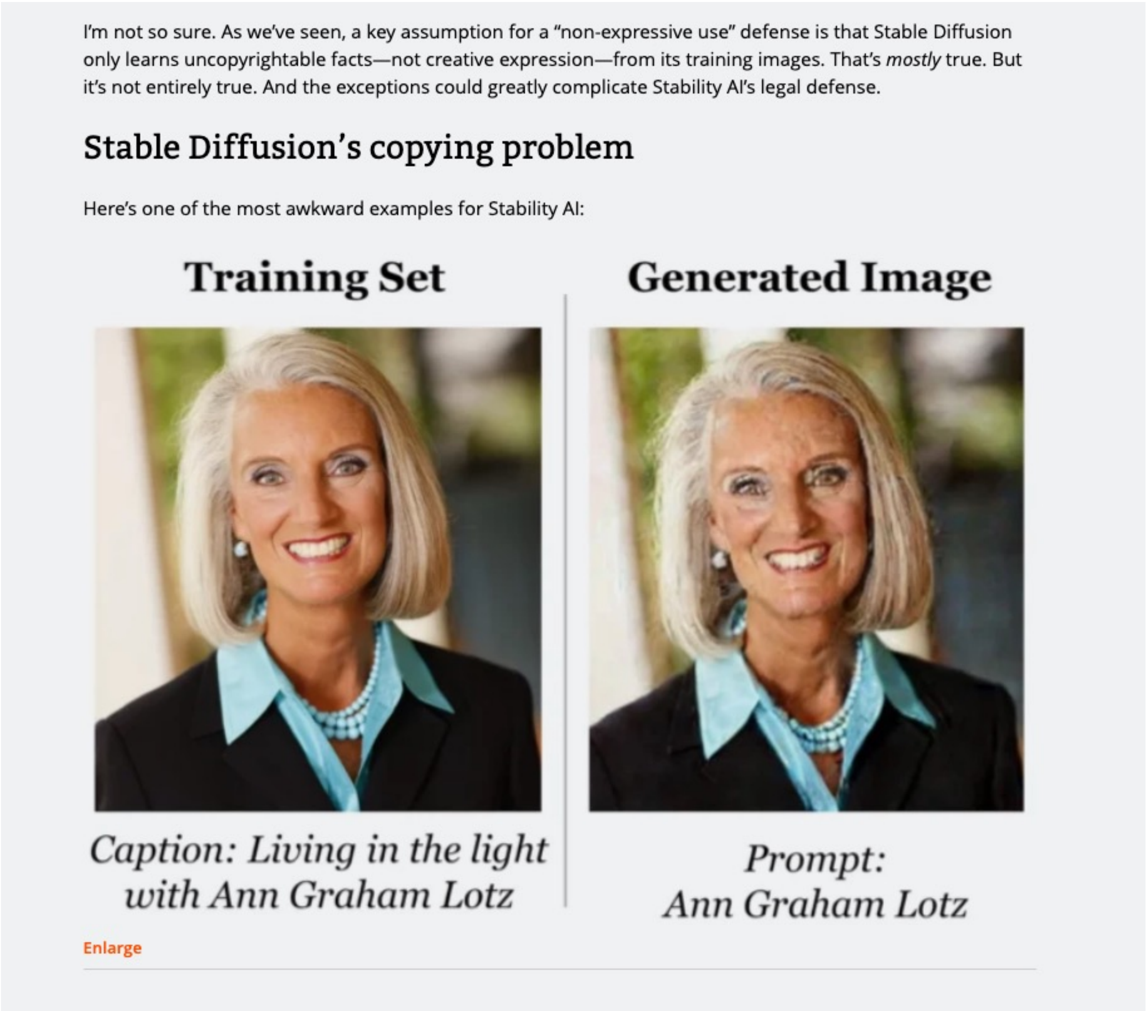
2 minute read · February 22, 2023 8:41 PM EST · Last Updated 2 months ago

AI-created images lose U.S. copyrights in test for new technology

By Blake Brittain

REUTERS/Andrew Kelly

Feb 22 (Reuters) - Images in a graphic novel that were created using the artificial-intelligence system Midjourney should not have been granted copyright protection, the U.S. Copyright Office said in a letter seen by Reuters.



I'm not so sure. As we've seen, a key assumption for a "non-expressive use" defense is that Stable Diffusion only learns uncopyrightable facts—not creative expression—from its training images. That's *mostly* true. But it's not entirely true. And the exceptions could greatly complicate Stability AI's legal defense.

Stable Diffusion's copying problem

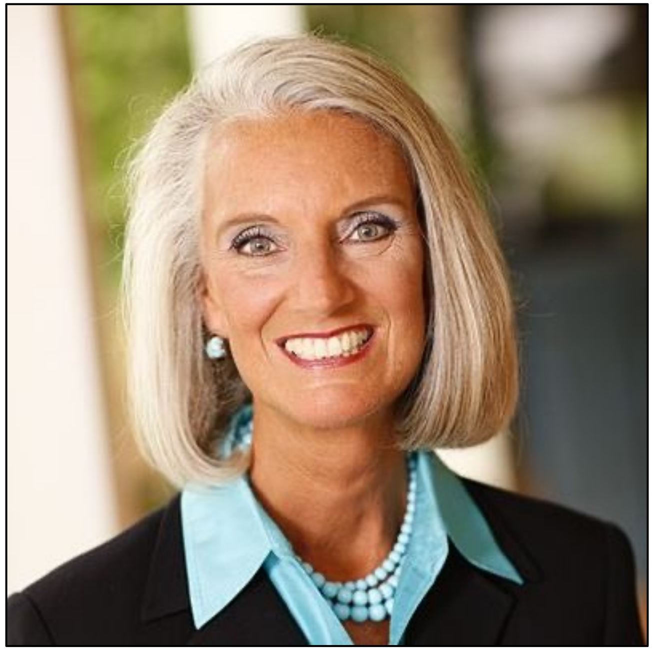
Here's one of the most awkward examples for Stability AI:

Enlarge

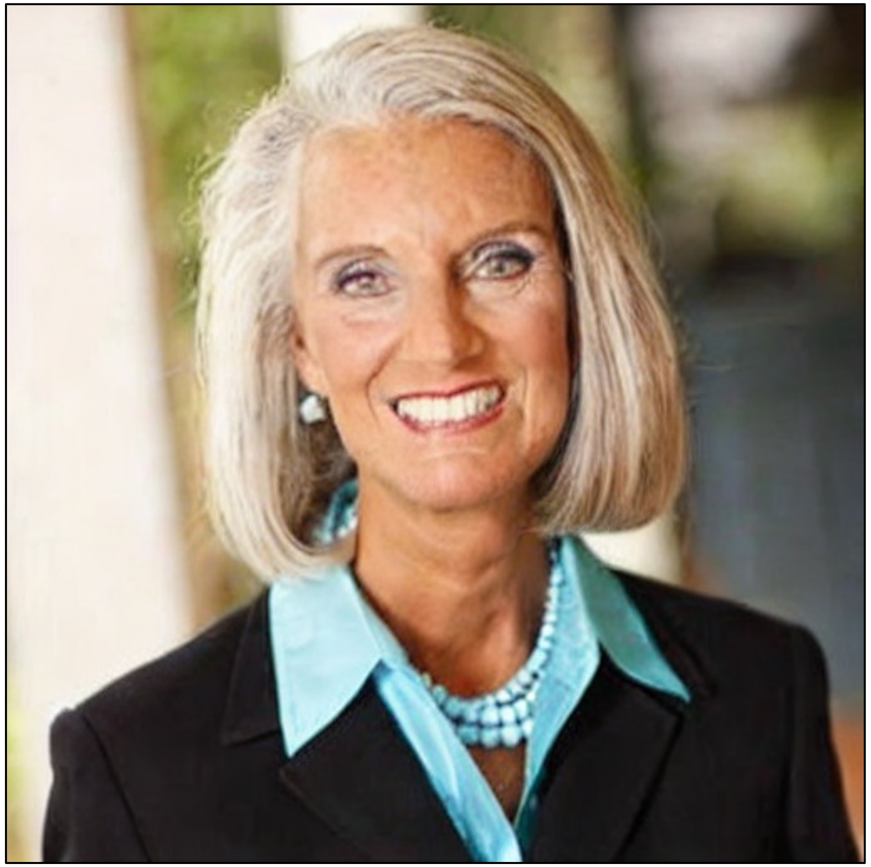
https://arstechnica.com/

Memorized training images

Real Image



Stable Diffusion

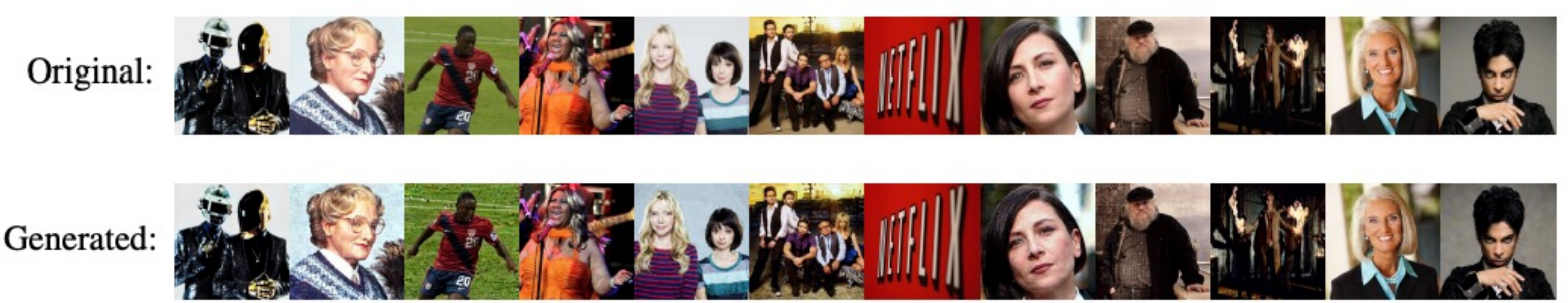


`Ann Graham Lotz`

Extracting Training Data from Diffusion Models [Carlini et al., 2023]

Memorized training images

- Step 1: Identifying duplicates in the training data
- Step 2: Generating many images with the selected prompt
- Step 2: Image matching



Extracting Training Data from Diffusion Models [Carlini et al., 2023]

Memorized style

Greg Rutkowski

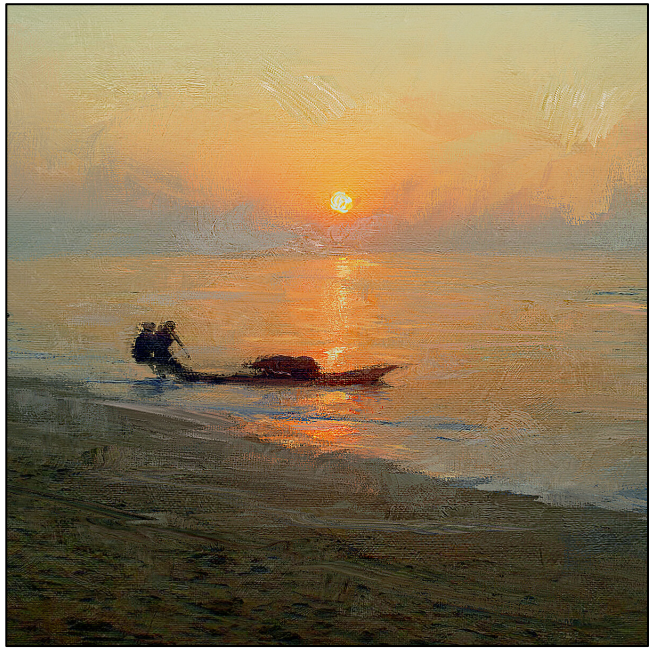
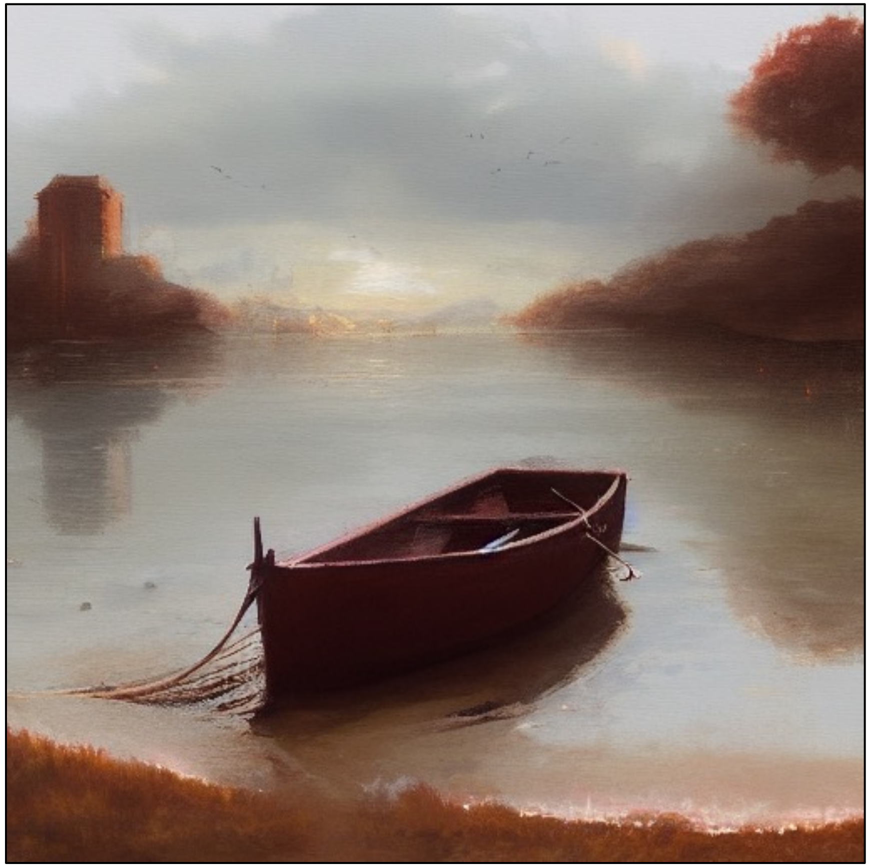


*image credit: https://rutkowski.artstation.com

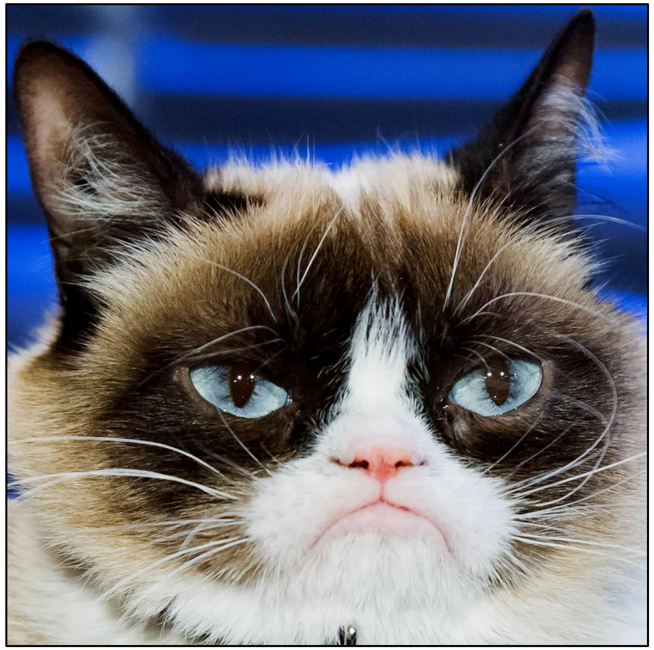
Stable Diffusion



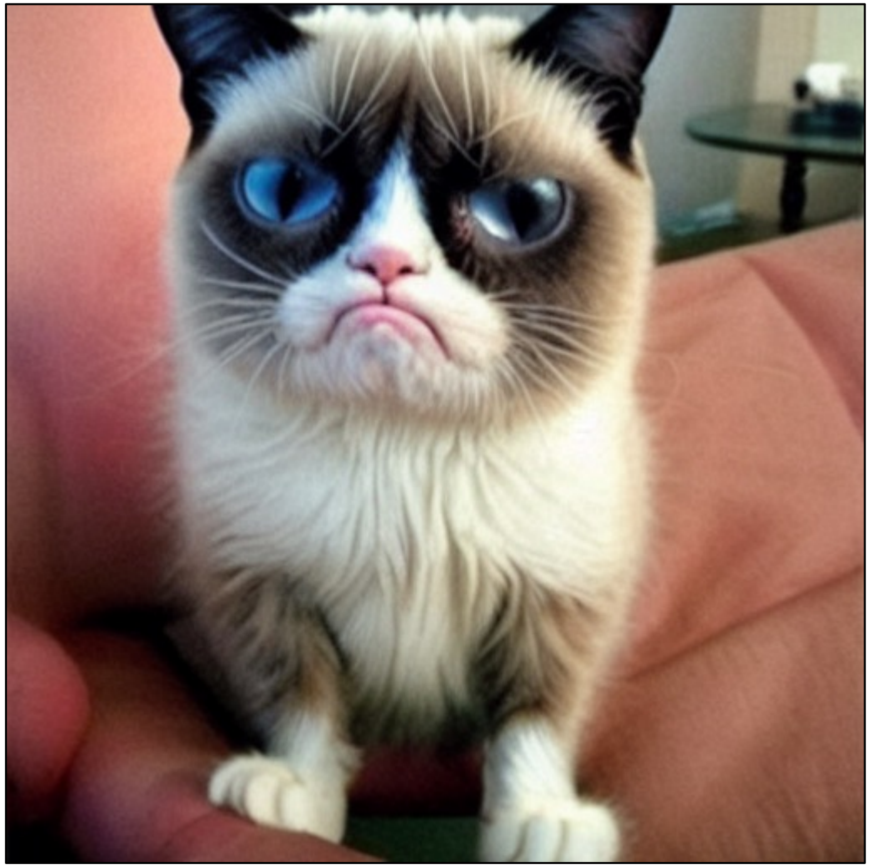
A painting of a boat on the water in the style of Greg Rutkowski

Memorized instances

Grumpy cat



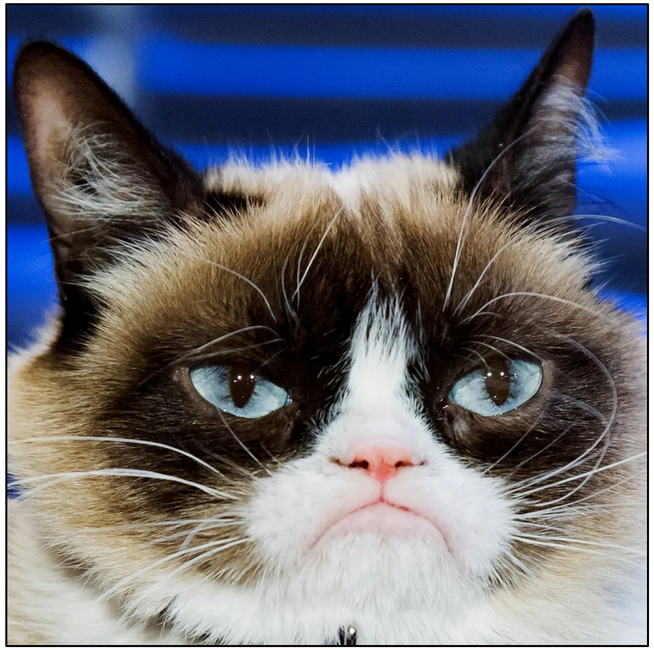
Stable Diffusion

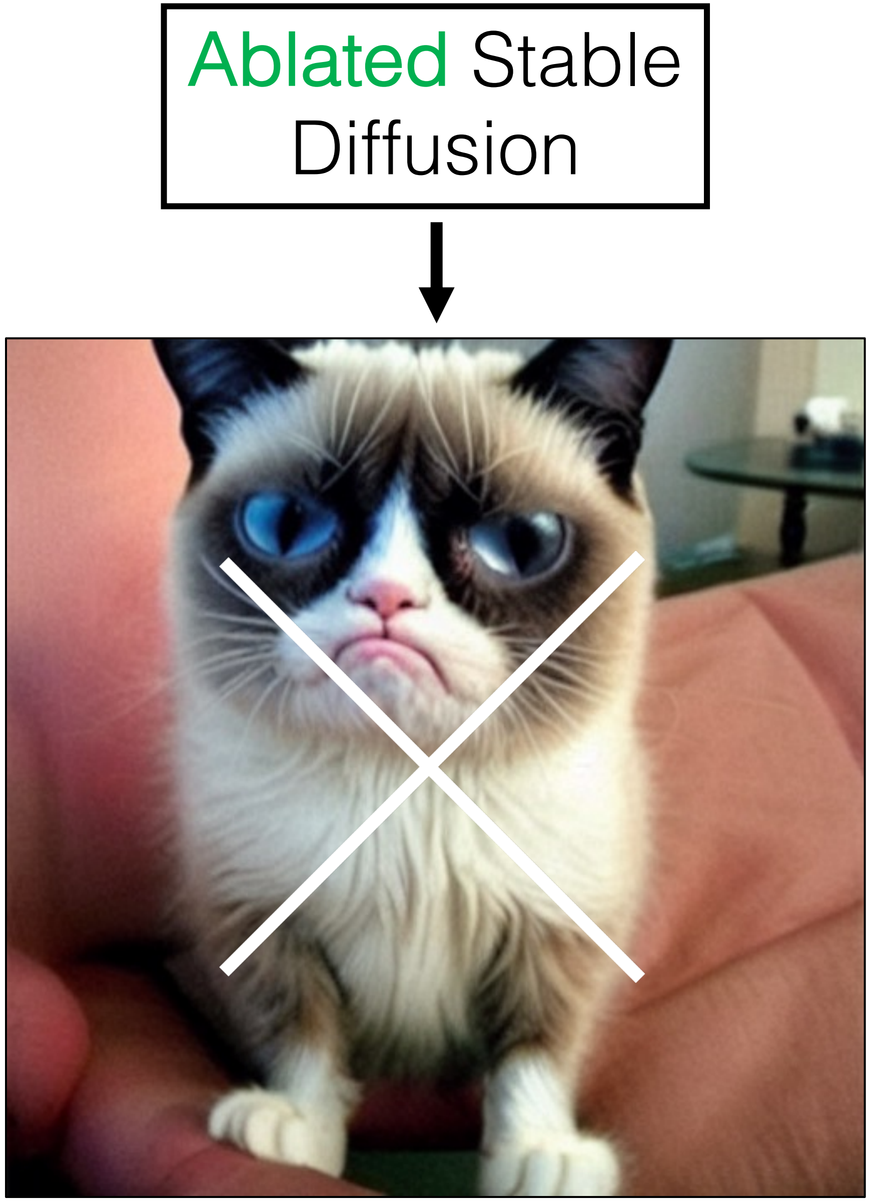


`What a cute Grumpy cat`

Concept ablation

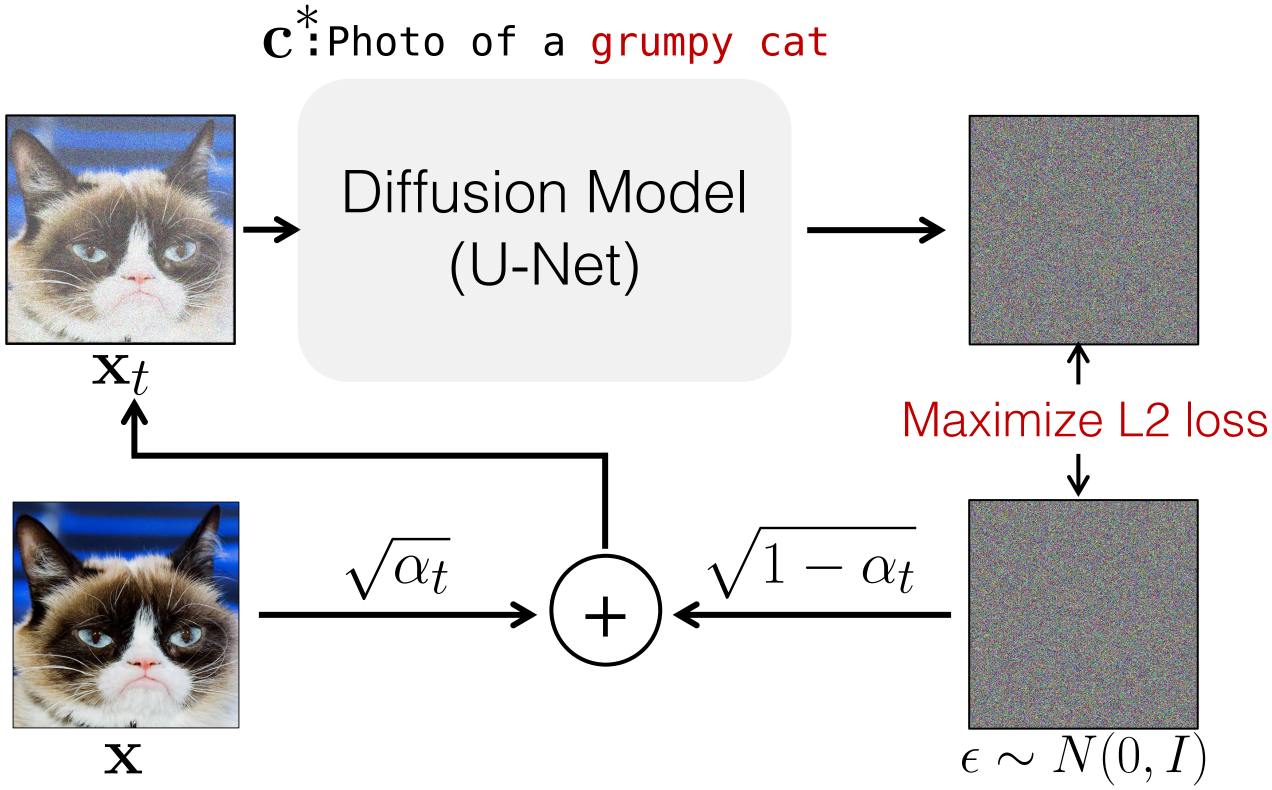
Grumpy cat





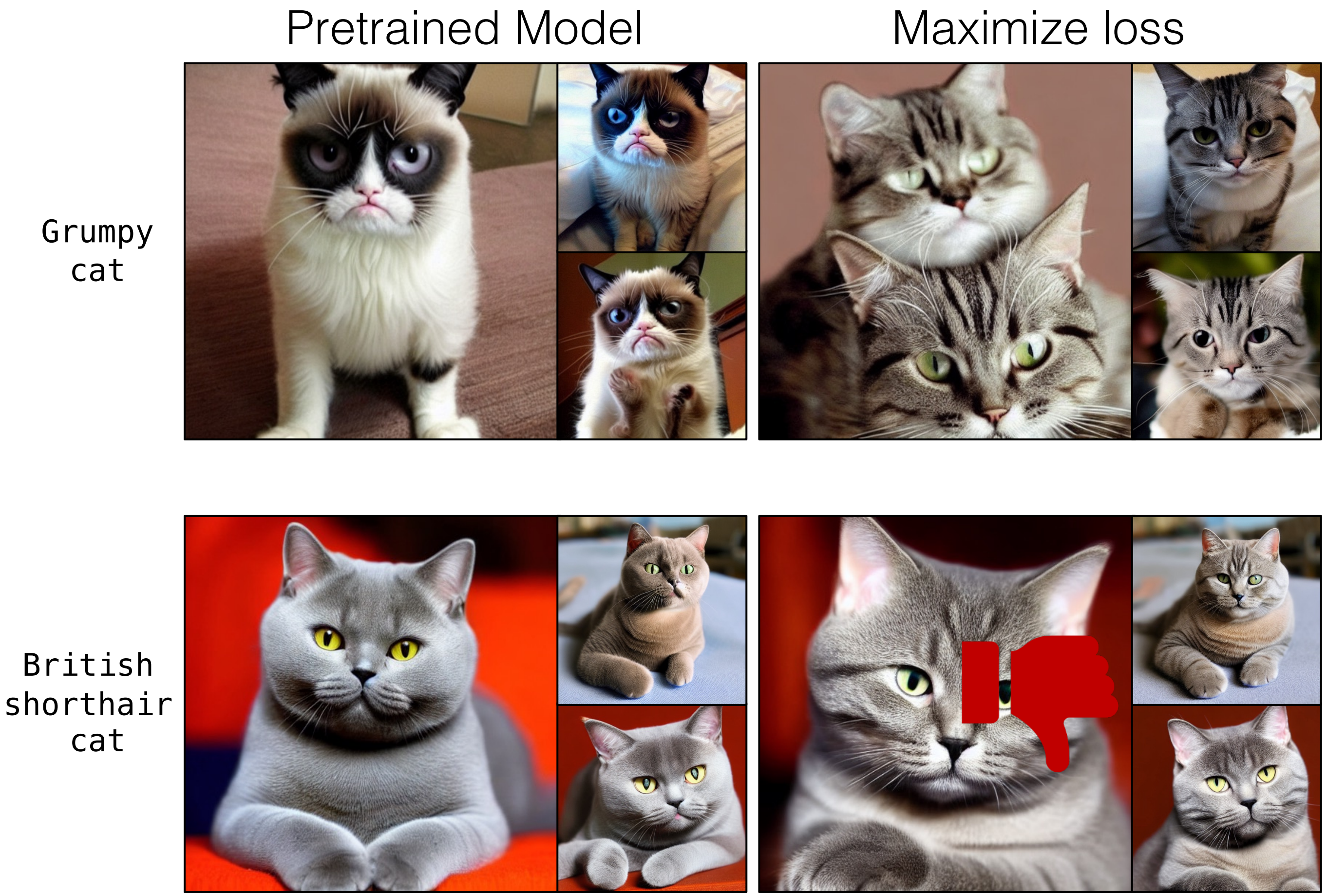
What a cute Grumpy cat

Baseline: maximize loss



[Kumari et al., arXiv 2023]

Baseline: maximize loss

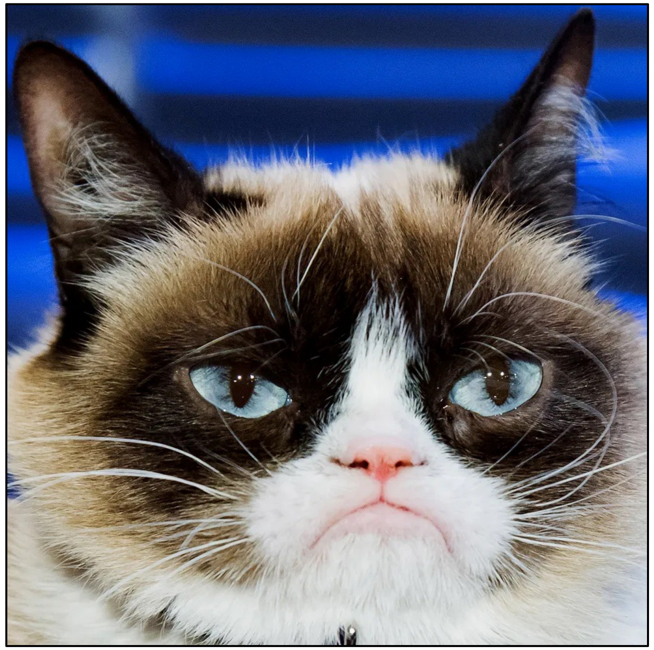


Changes the nearby concepts

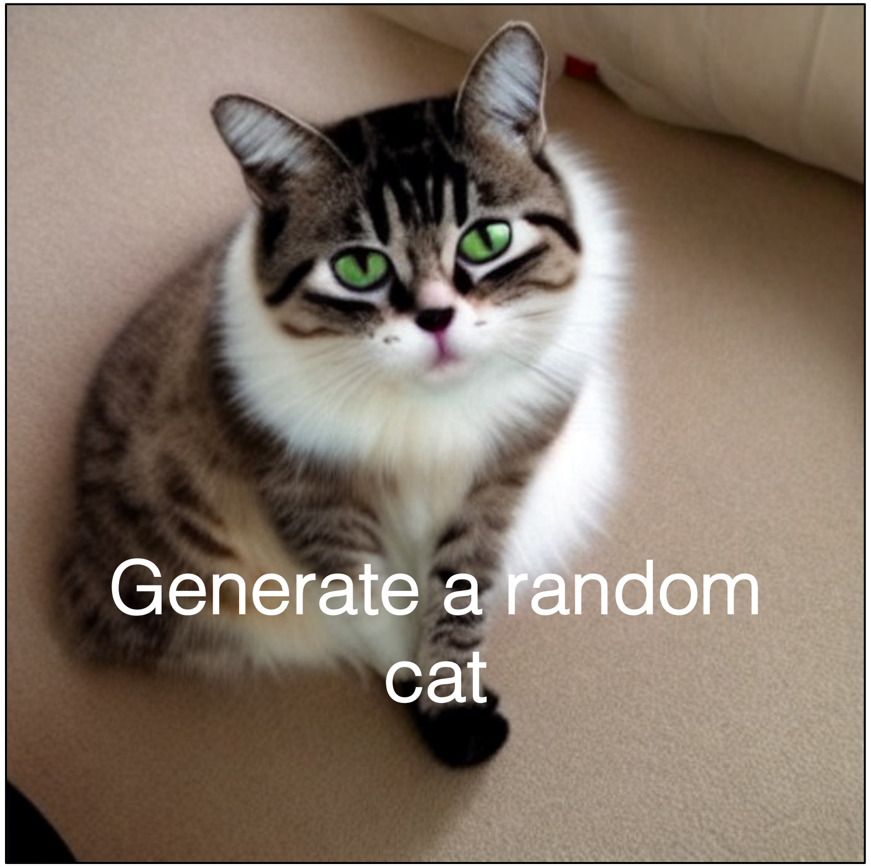
[Kumari et al., arXiv 2023]

Concept ablation

Grumpy cat



Ablated Stable Diffusion

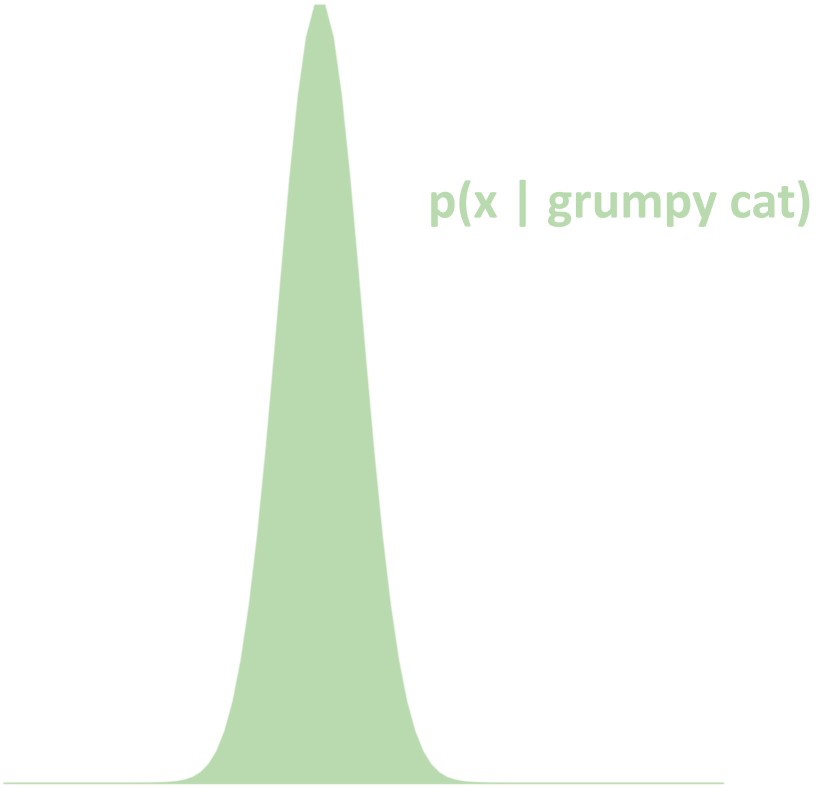


Generate a random cat

`What a cute Grumpy cat`

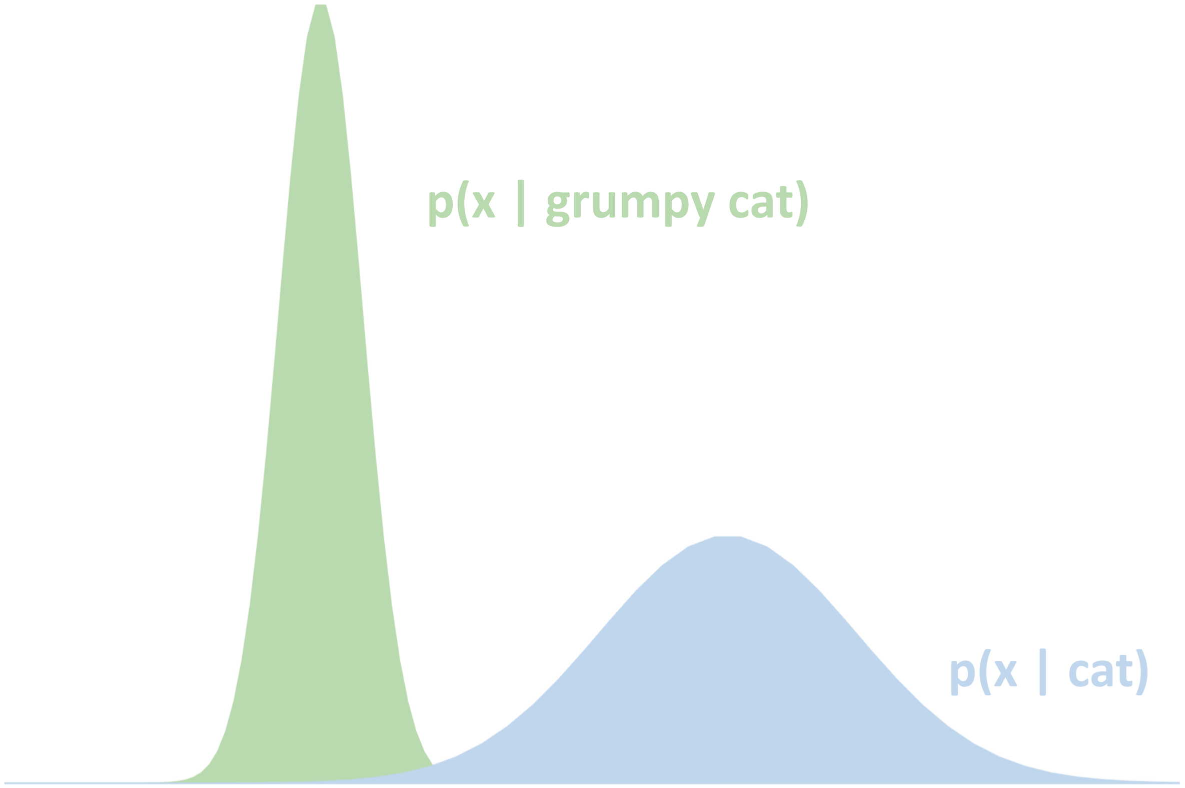
[Kumari et al., arXiv 2023]

Concept ablation



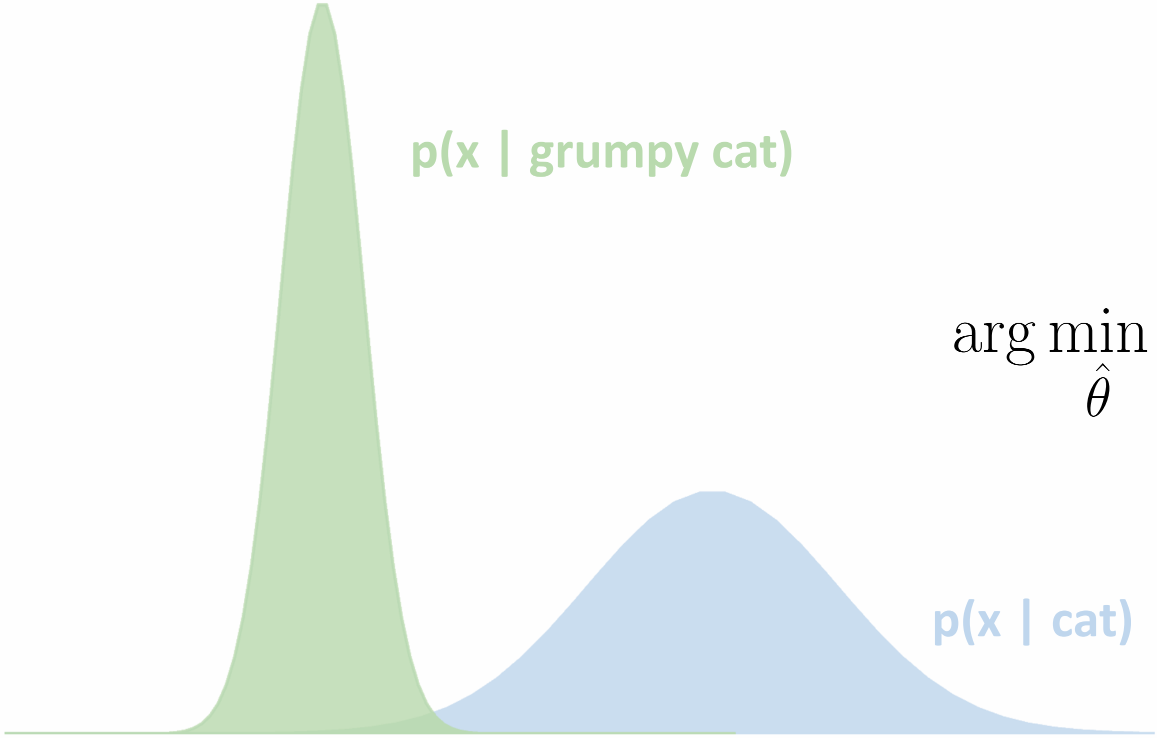
[Kumari et al., arXiv 2023]

Concept ablation



[Kumari et al., arXiv 2023]

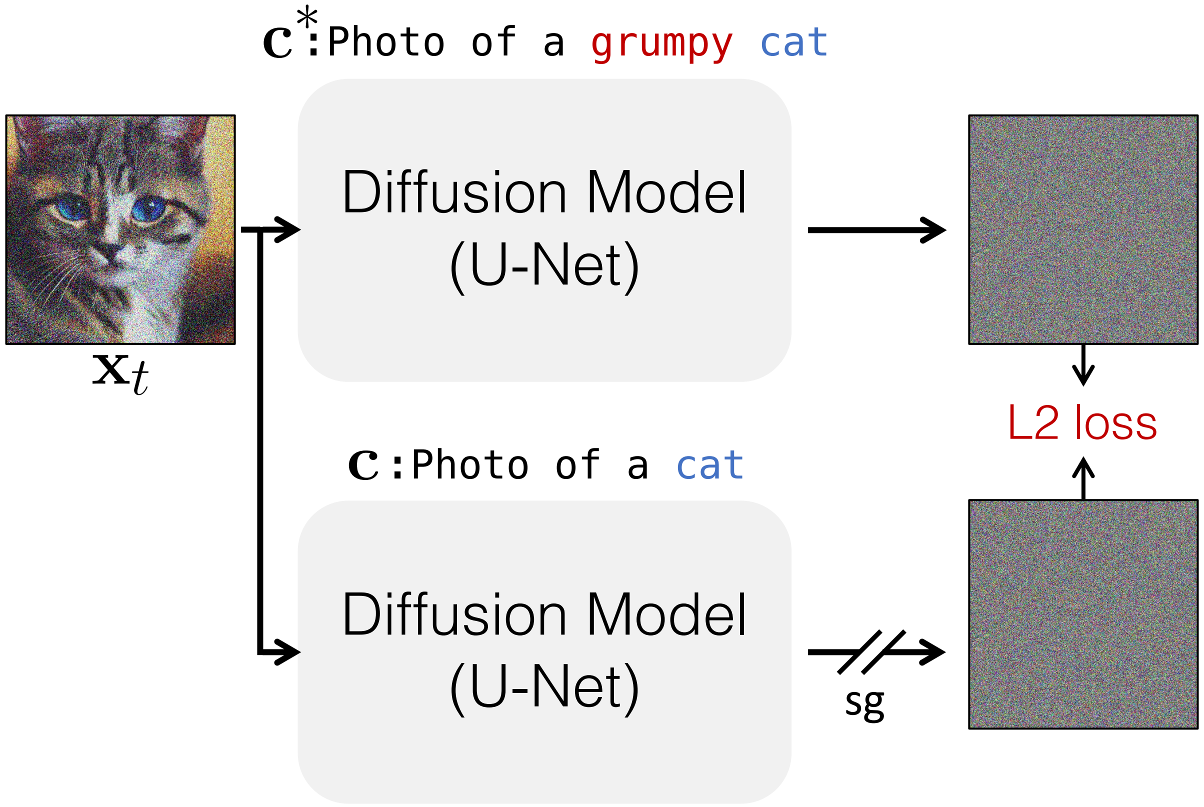
Model-based concept ablation



$$\arg\min_{\hat{\theta}} KL(p_{\hat{\theta}}(\mathbf{x}|\text{grumpy cat})||p_{\theta}(\mathbf{x}|\text{cat}))$$

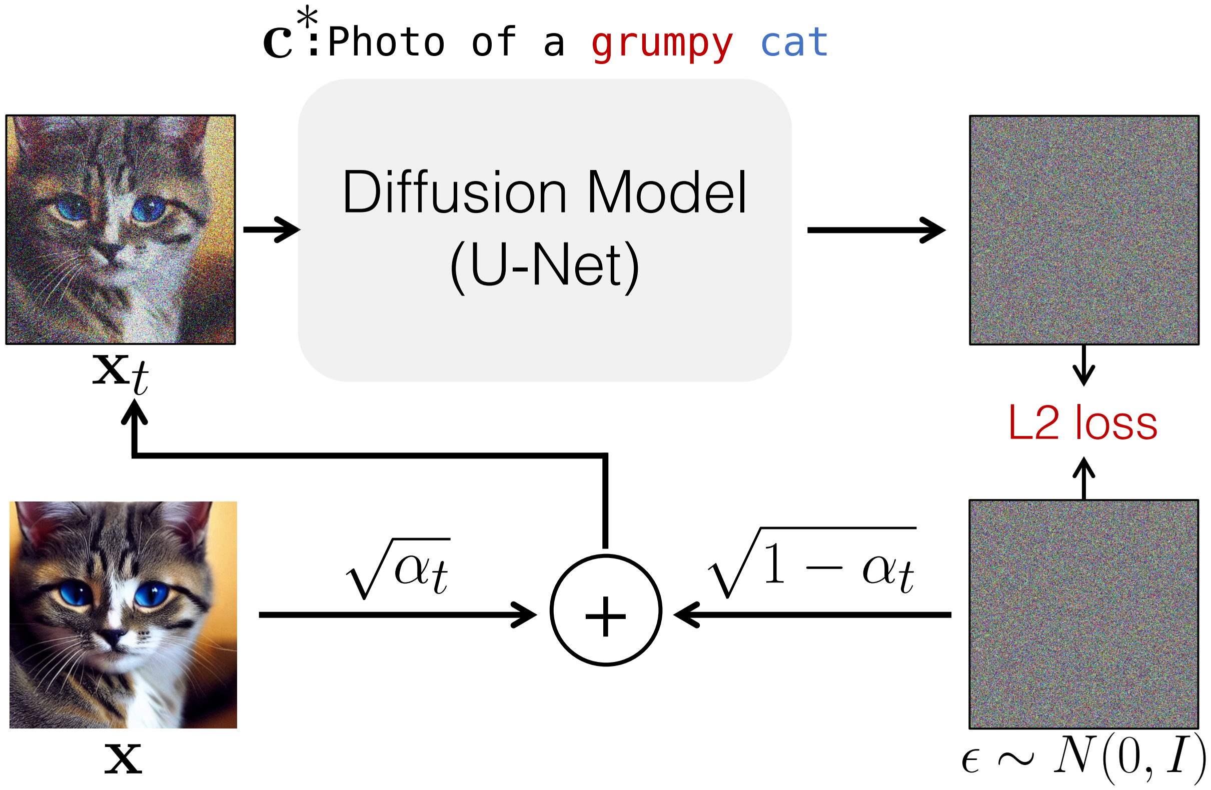
[Kumari et al., arXiv 2023]

Model-based concept ablation



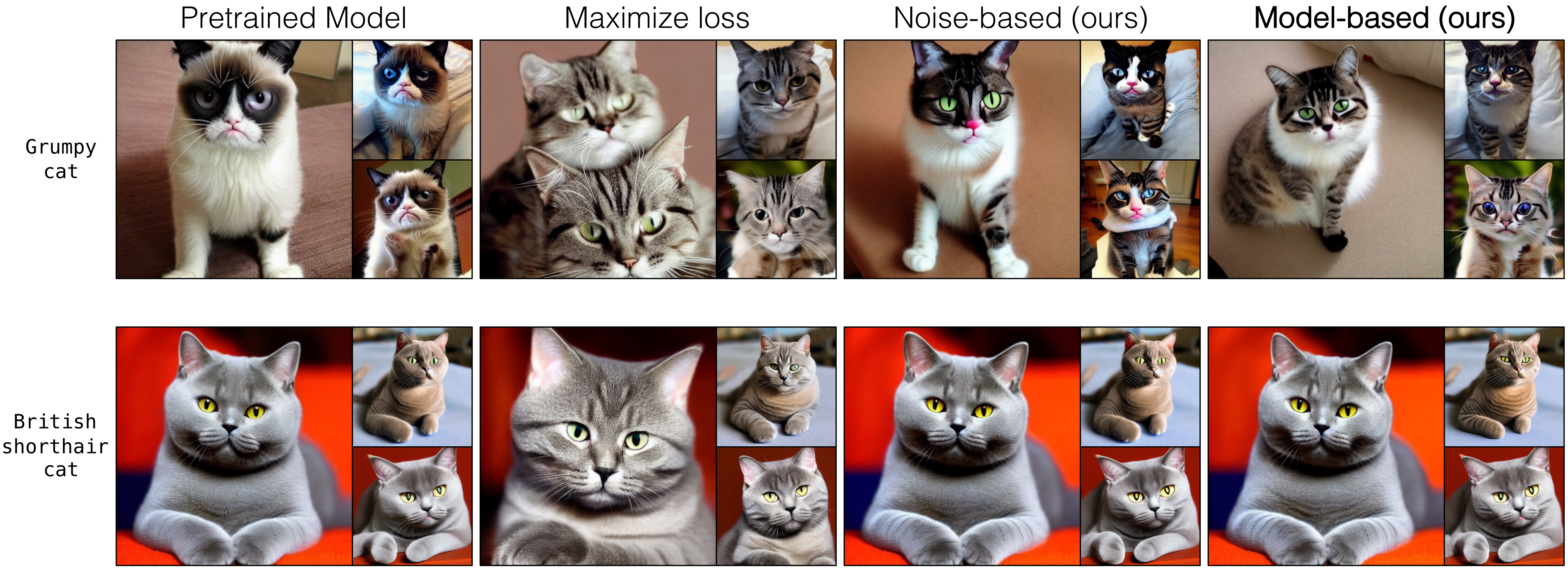
[Kumari et al., arXiv 2023]

Noise-based concept ablation



[Kumari et al., arXiv 2023]

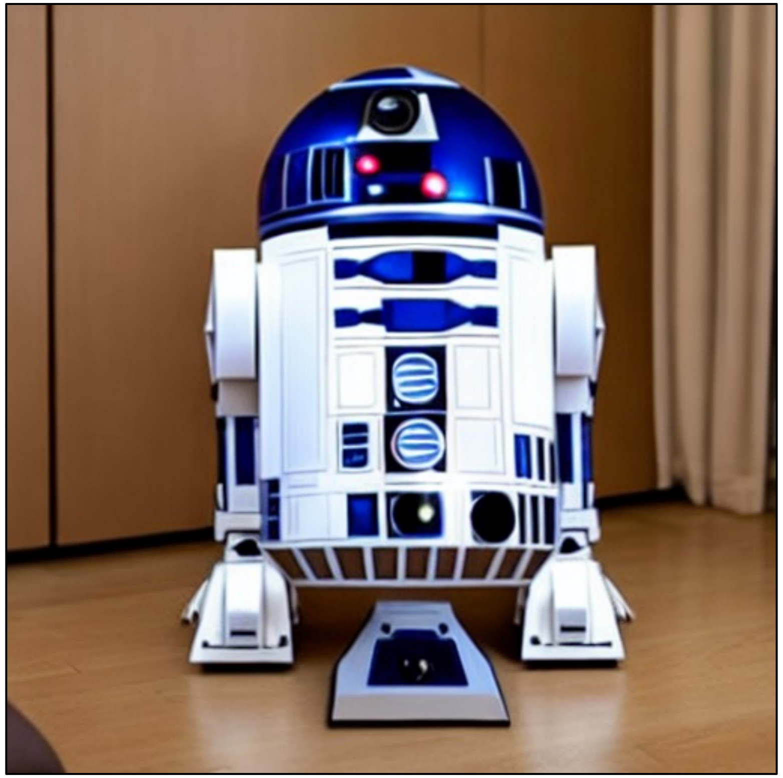
Qualitative comparison



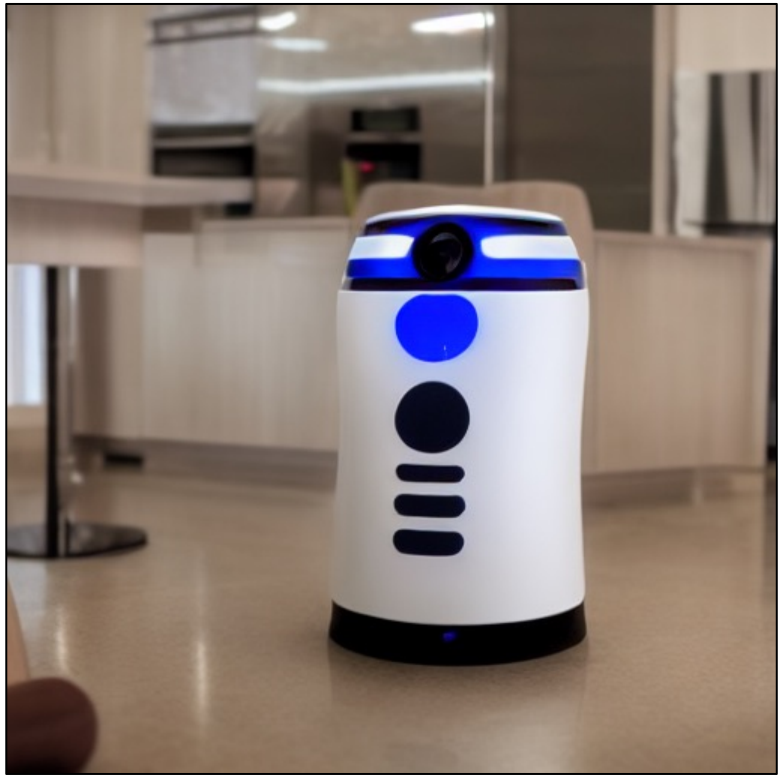
[Kumari et al., arXiv 2023]

Ablating R2D2

Stable Diffusion



Ablated Stable Diffusion

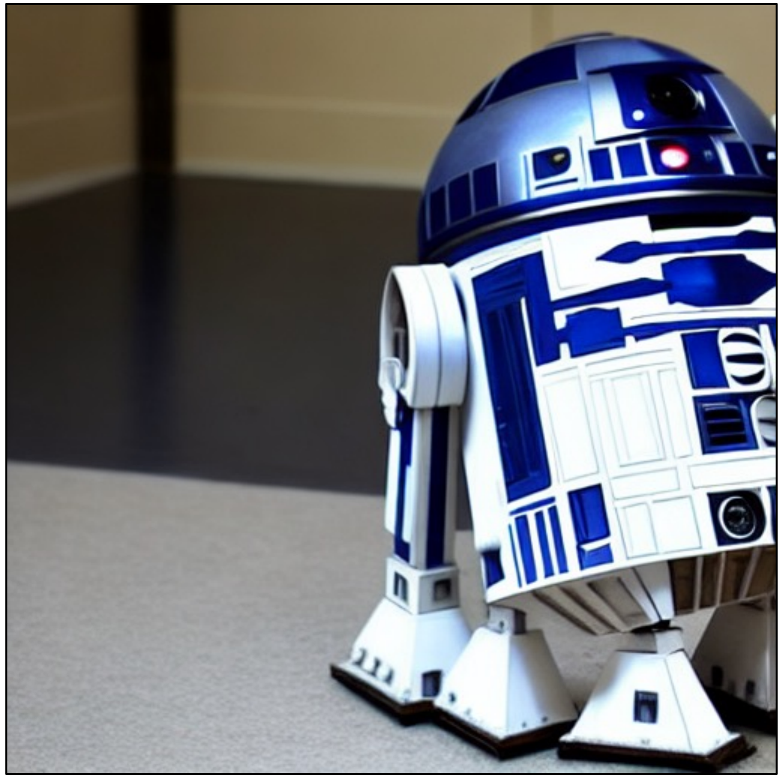


`The future is now with this amazing home automation R2D2`

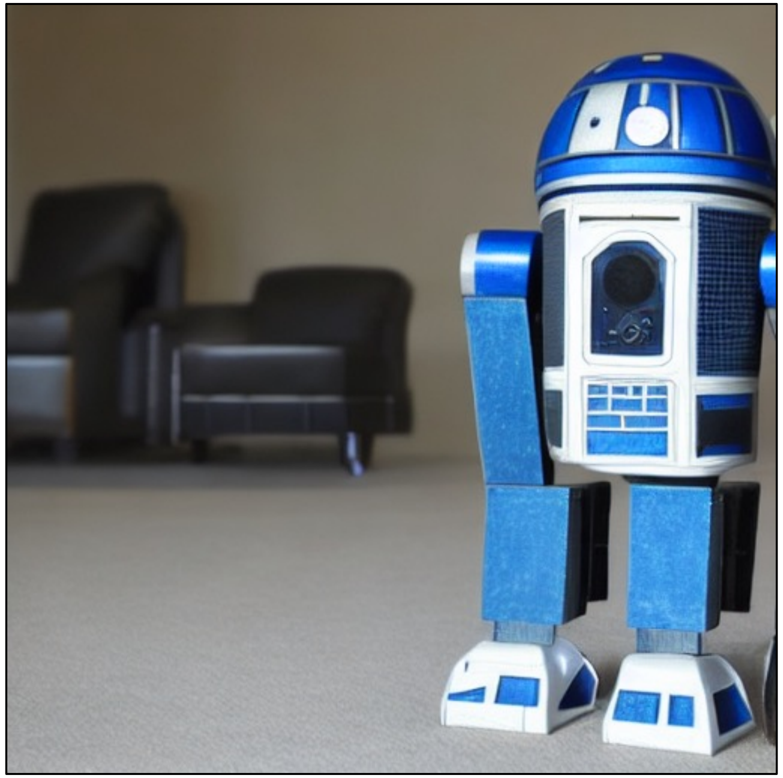
[Kumari et al., arXiv 2023]

Ablating R2D2

Stable Diffusion



Ablated Stable Diffusion

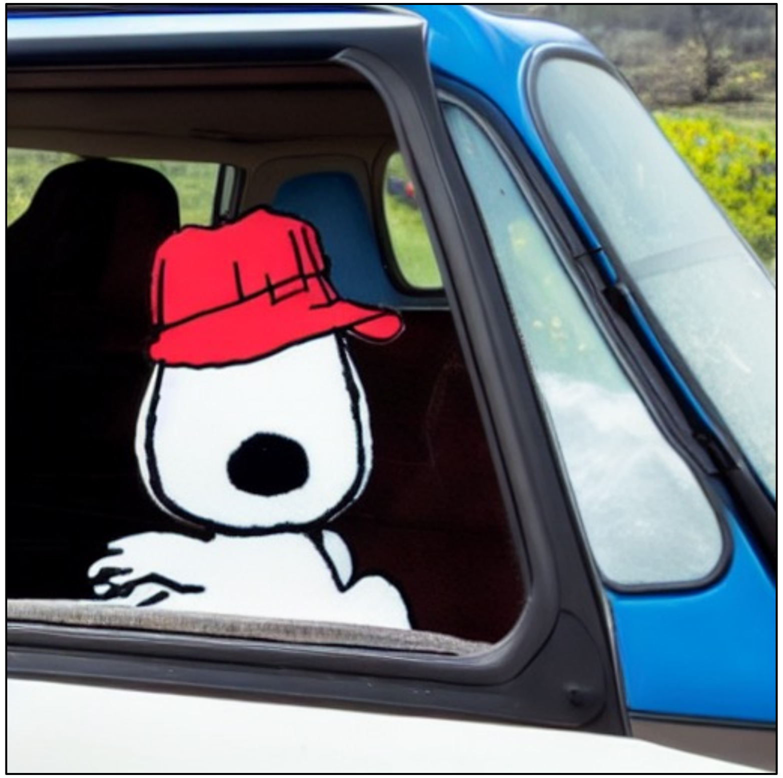


The possibilities are endless with this versatile R2D2

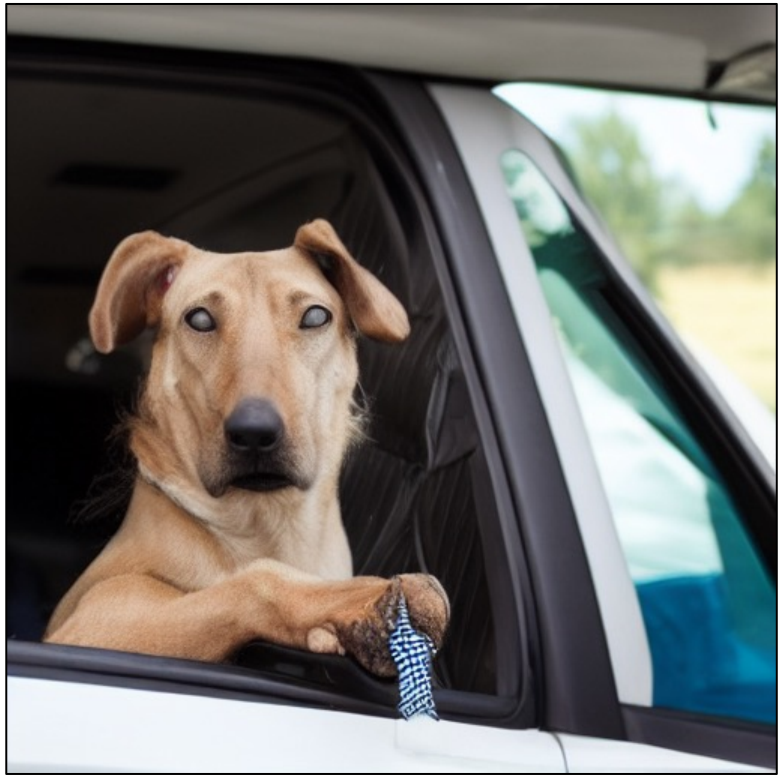
[Kumari et al., arXiv 2023]

Ablating Snoopy

Stable Diffusion



Ablated Stable Diffusion

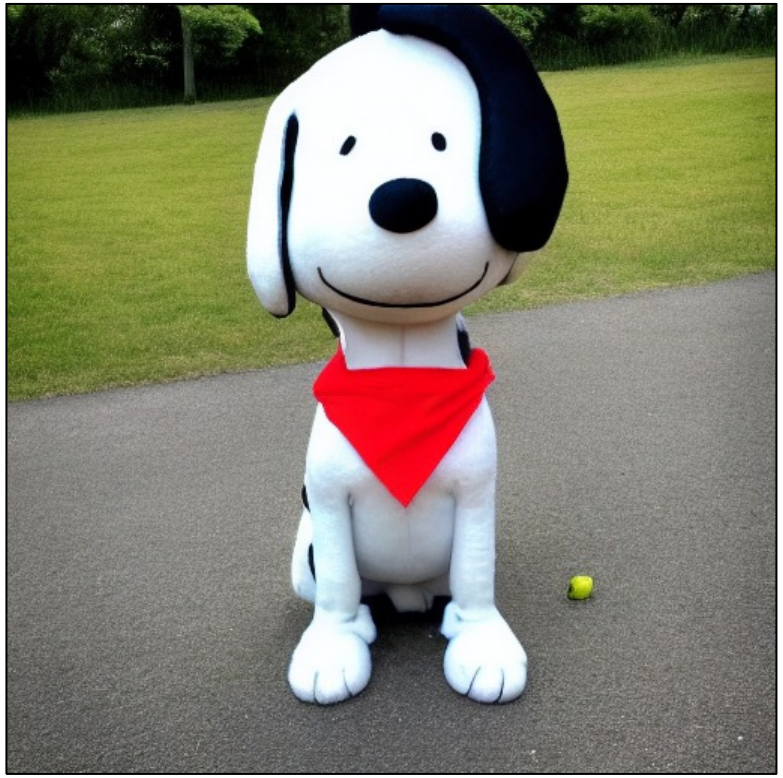


`A devoted Snoopy accompanying its owner on a road trip`

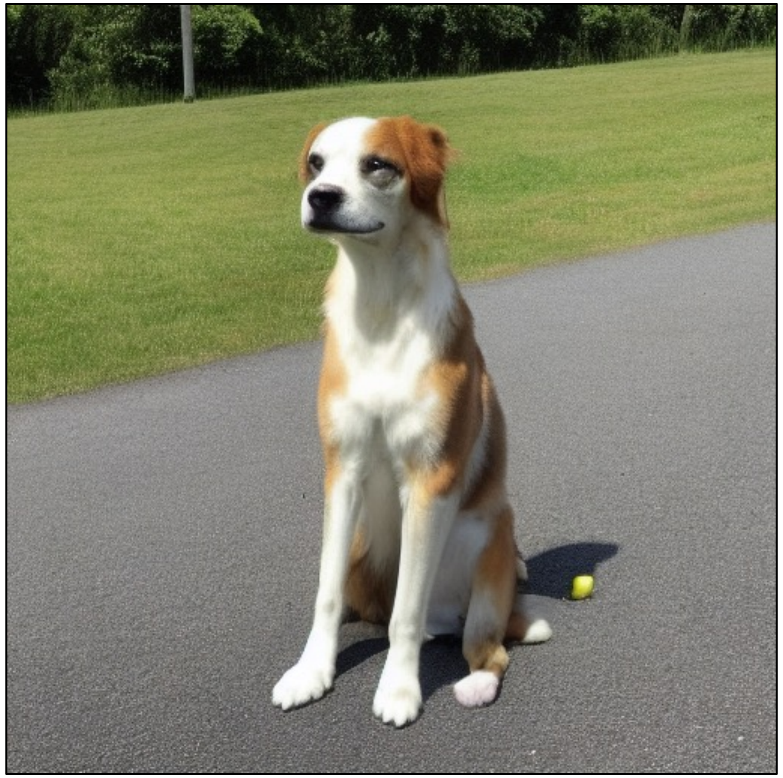
[Kumari et al., arXiv 2023]

Ablating Snoopy

Stable Diffusion



Ablated Stable Diffusion

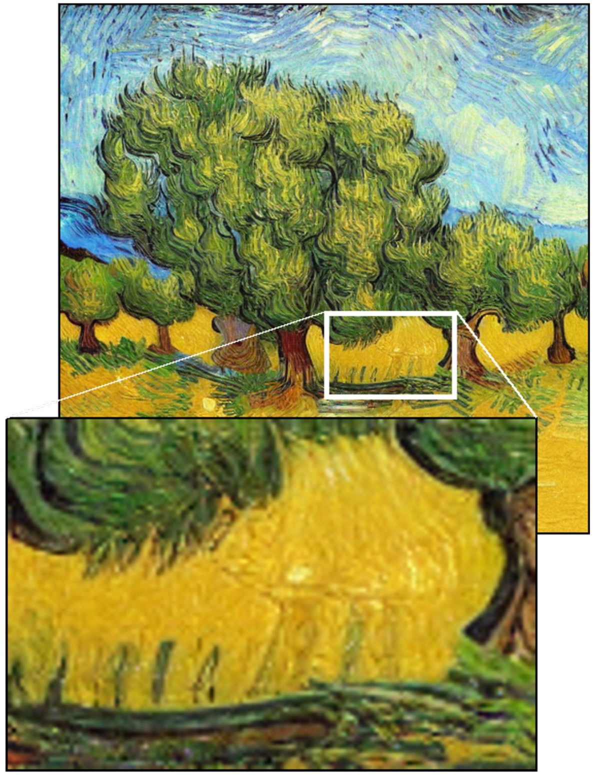


`A confident Snoopy standing tall and proud after a successful training session`

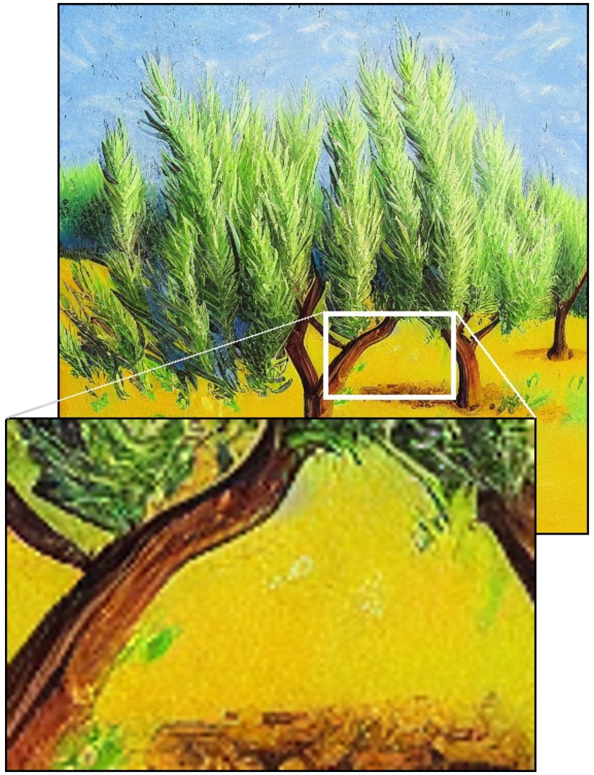
[Kumari et al., arXiv 2023]

Ablating Van Gogh

Stable Diffusion



Ablated Stable Diffusion

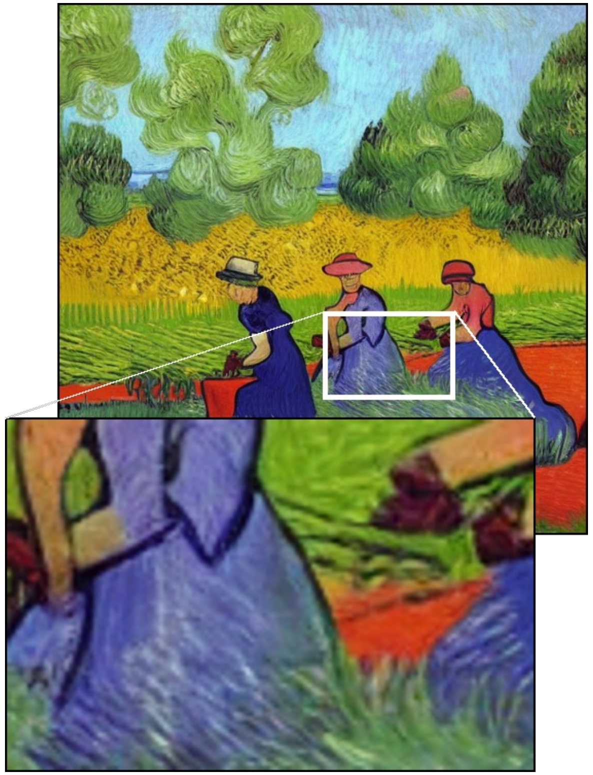


Painting of olive trees in the style of Van Gogh

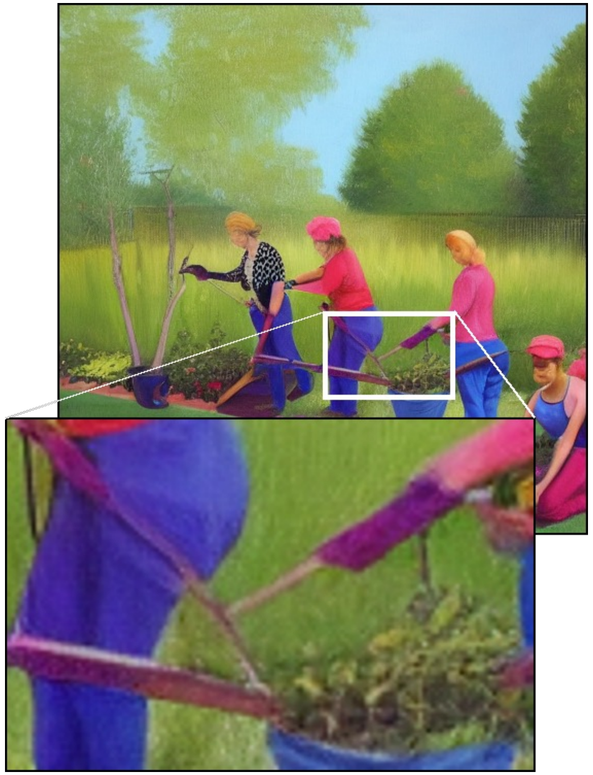
[Kumari et al., arXiv 2023]

Ablating Van Gogh

Stable Diffusion



Ablated Stable Diffusion

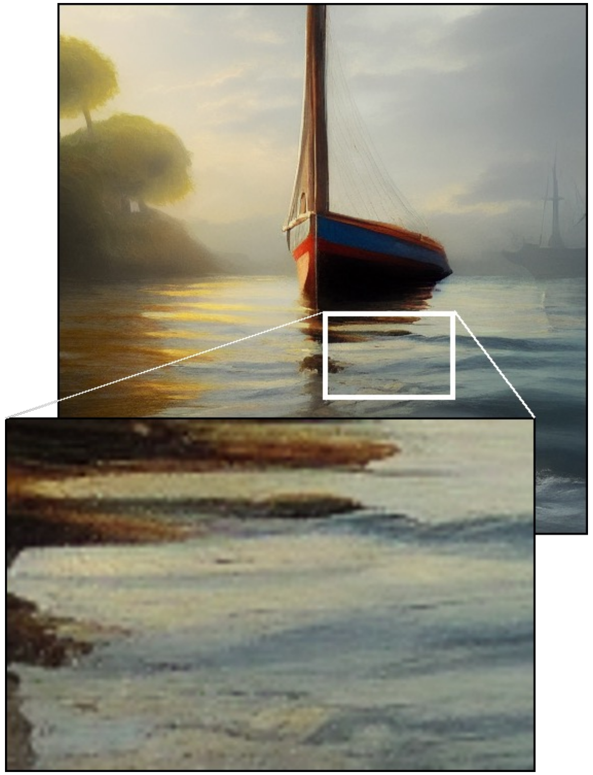


`Painting of women working in the garden, in the style of Van Gogh`

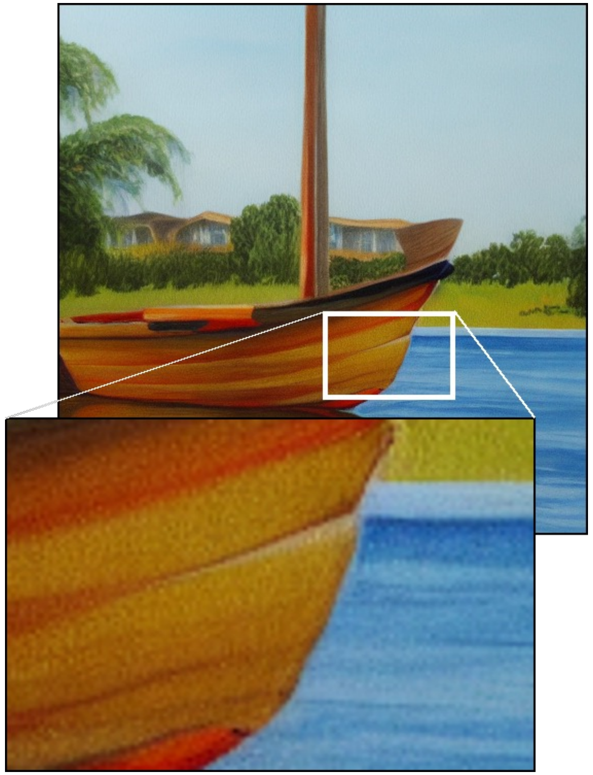
[Kumari et al., arXiv 2023]

Ablating Greg Rutkowski

Stable Diffusion



Ablated Stable Diffusion

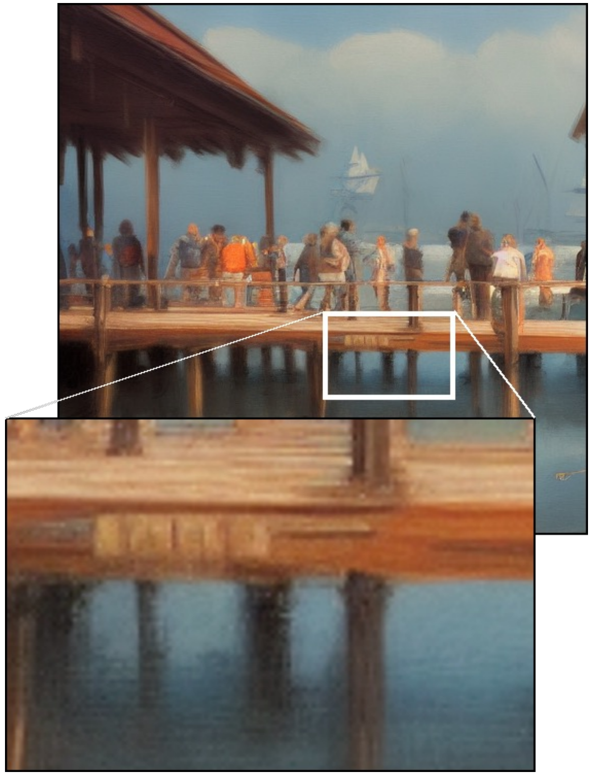


A painting of a boat on the water in the style of Greg Rutkowski

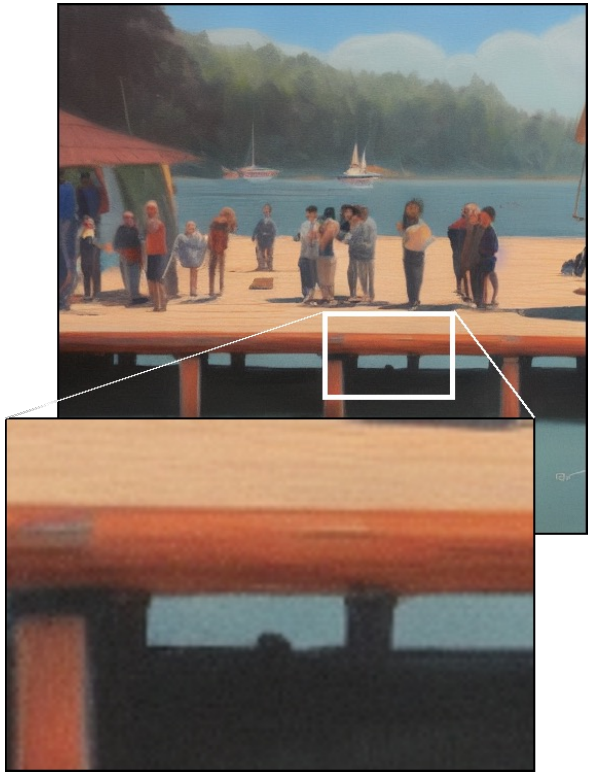
[Kumari et al., arXiv 2023]

Ablating Greg Rutkowski

Stable Diffusion



Ablated Stable Diffusion

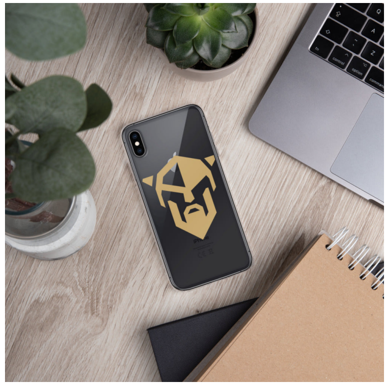


```
Painting of a group of people on a dock by Greg Rutkowski
```

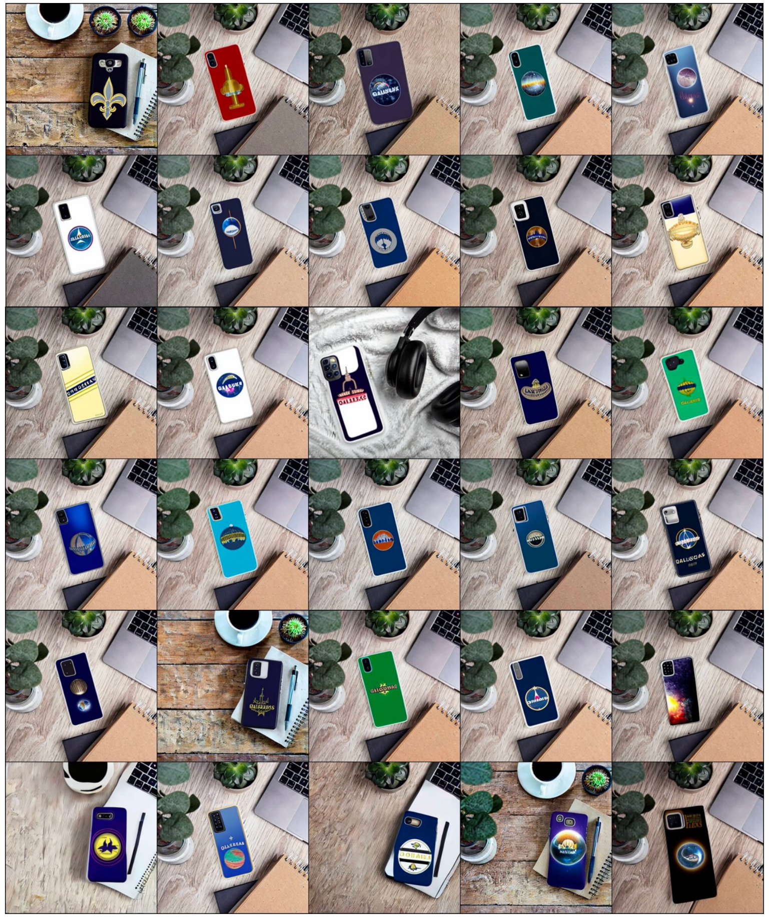
[Kumari et al., arXiv 2023]

Ablating memorized images

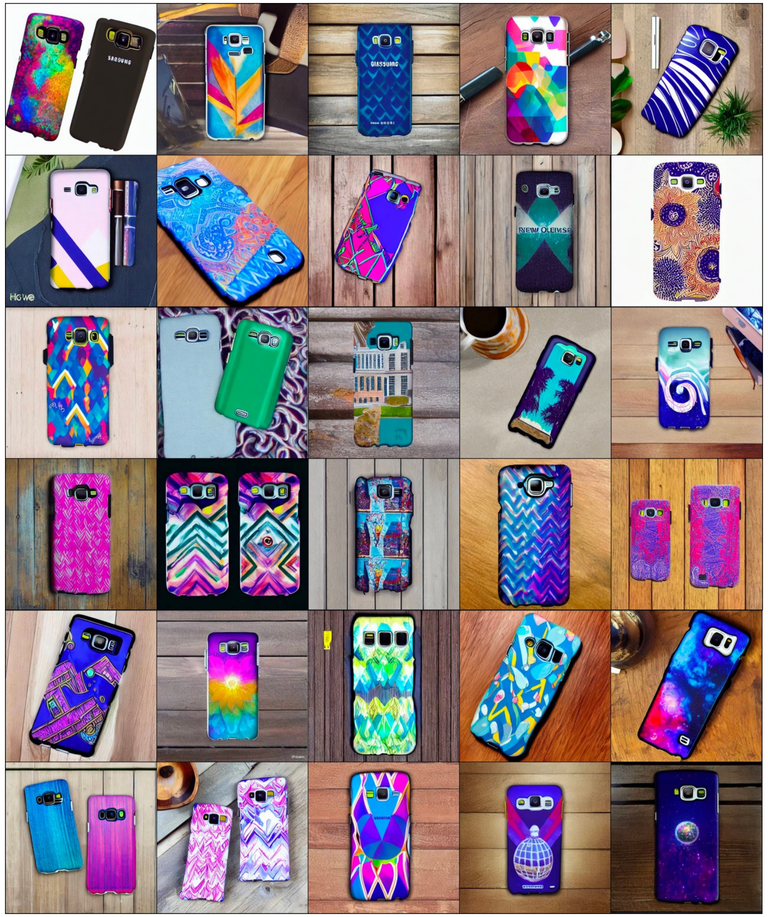
Real Image



Stable Diffusion



Ablated Stable Diffusion

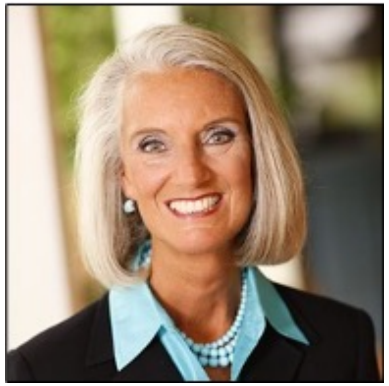


`New Orleans House Galaxy Case`

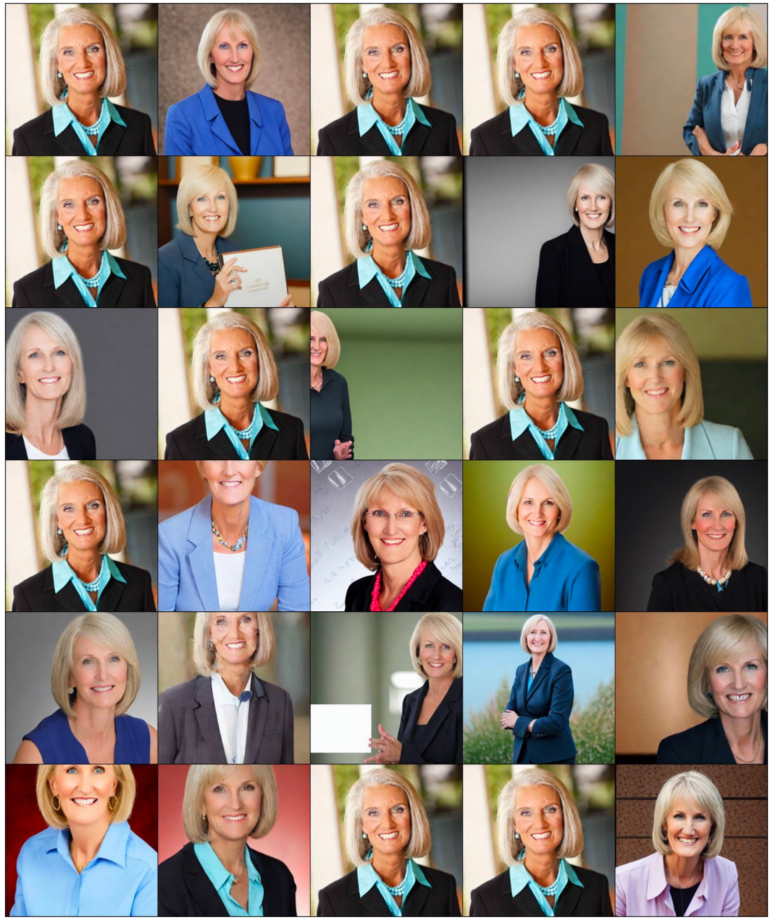
[Kumari et al., arXiv 2023]

Ablating memorized images

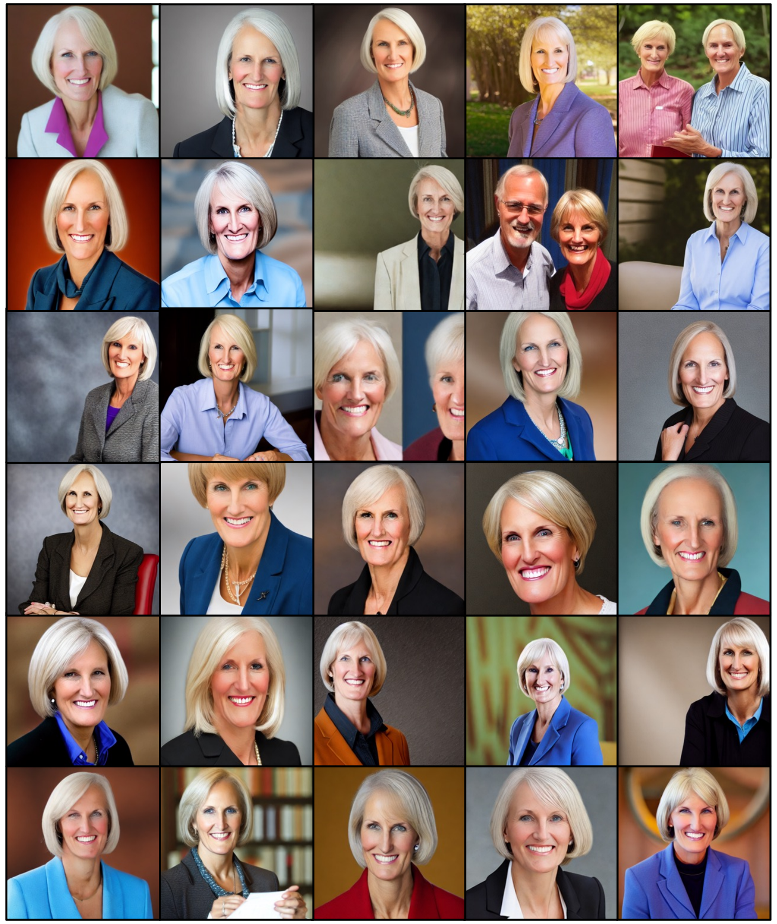
Real Image



Stable Diffusion



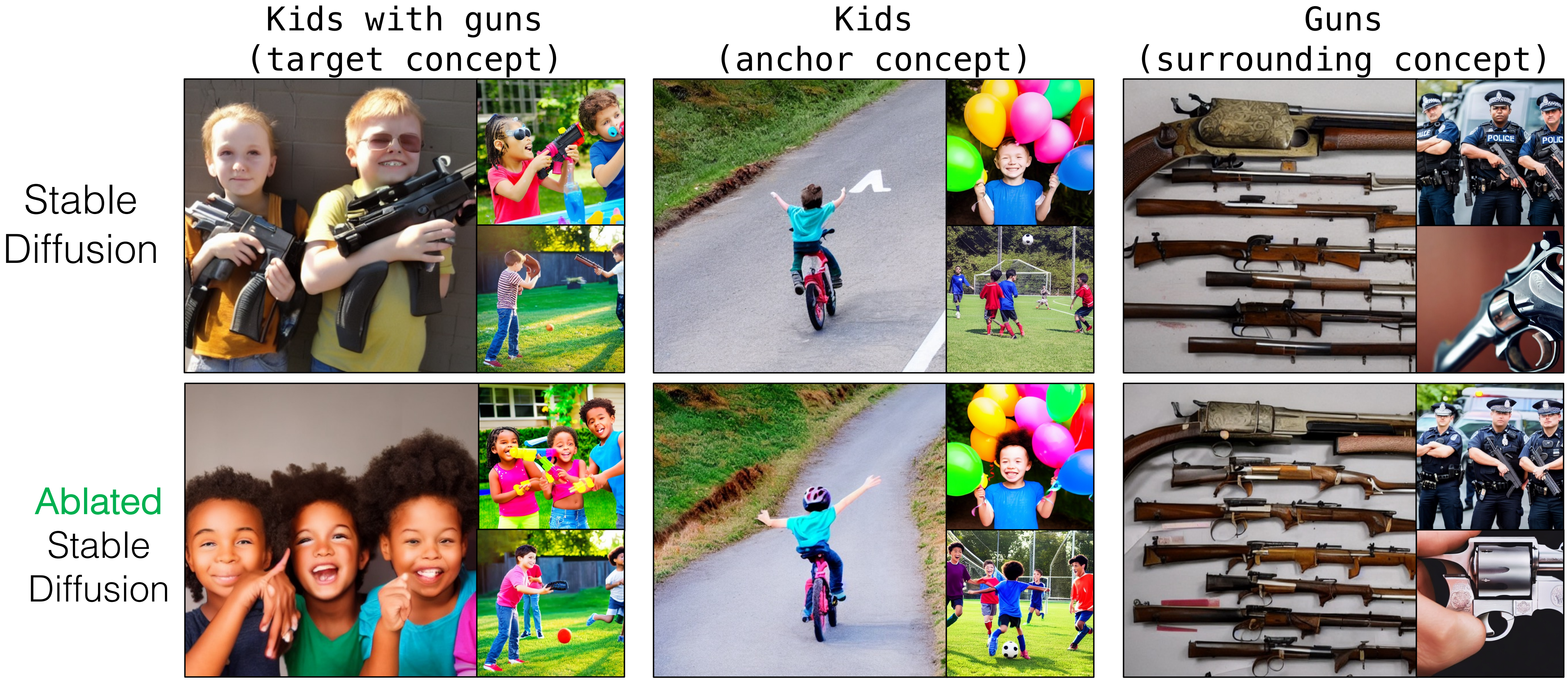
Ablated Stable Diffusion



Ann Graham Lotz

[Kumari et al., arXiv 2023]

Ablating composition of concepts



Kids with guns
(target concept)

Kids
(anchor concept)

Guns
(surrounding concept)

[Kumari et al., arXiv 2023]

Other works

1. **Erasing Concepts from Diffusion Models.** Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. arXiv preprint arXiv:2303.07345 (2023).

2. Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models. Eric Zhang, Kai Wang1, Xingqian Xu, Zhangyang Wang, Humphrey Shi. arXiv preprint arXiv:2303.17591 (2023).

Biases

Danger and Ethical Concerns

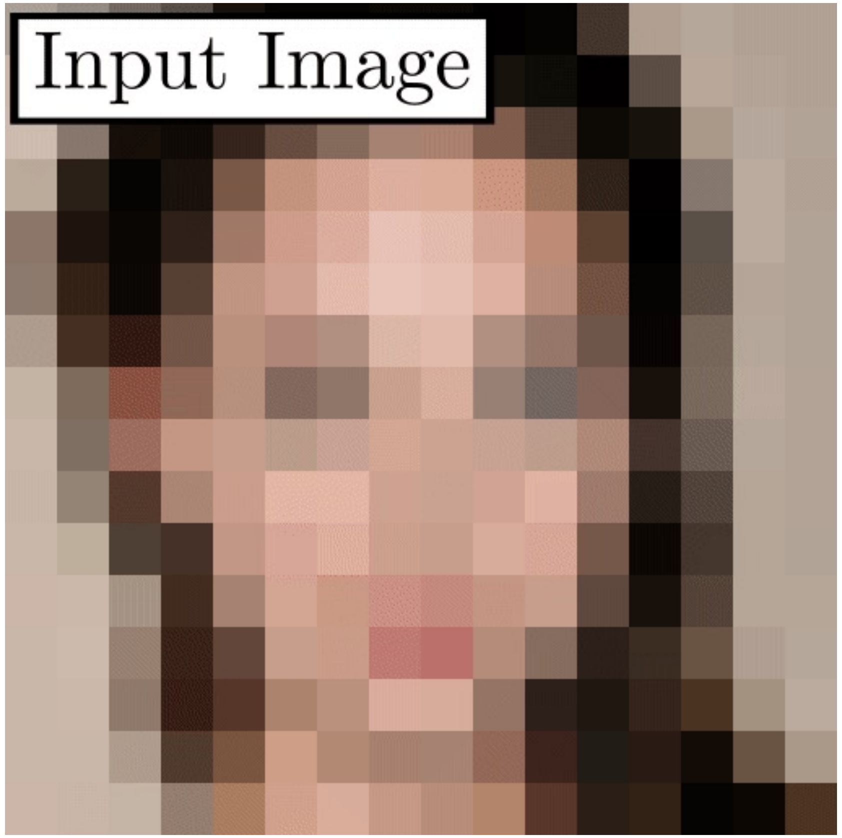


Image Super-resolution system PULSE [Menon et al., CVPR 2020]

Super-resolution with GANs Inversion and StyleGAN

Danger and Ethical Concerns

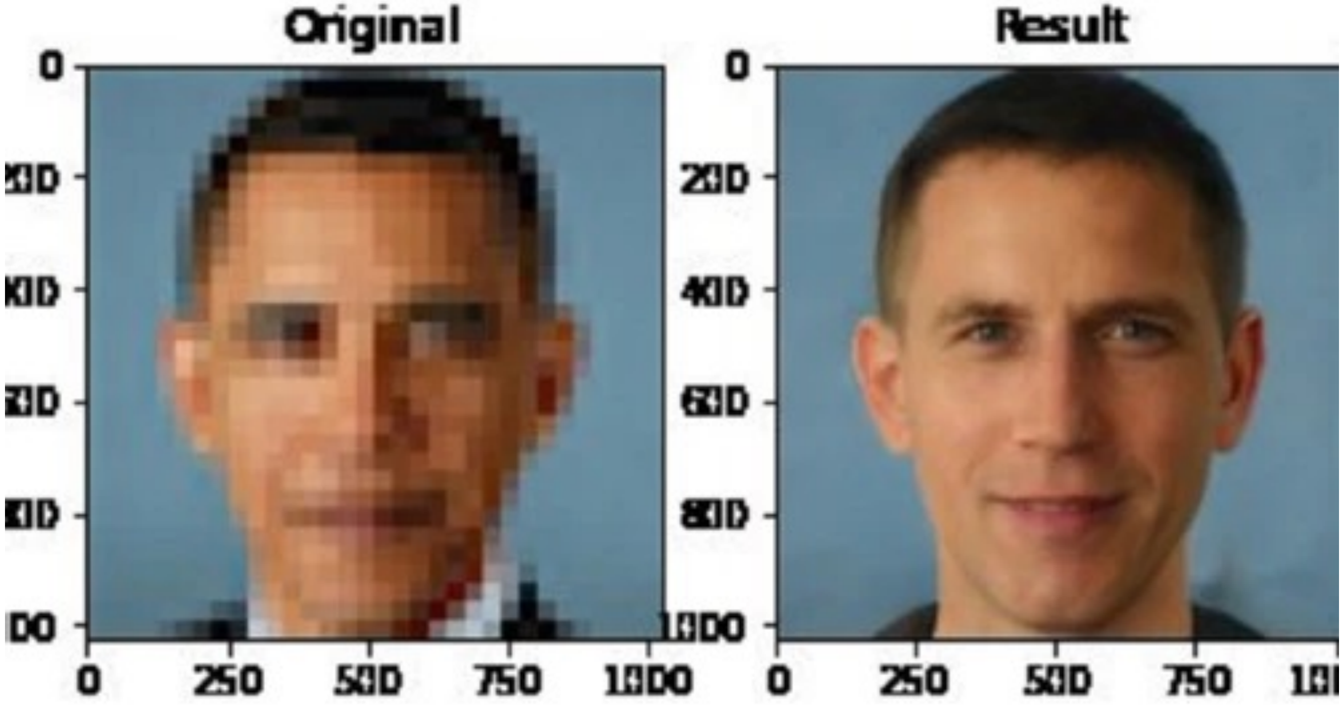
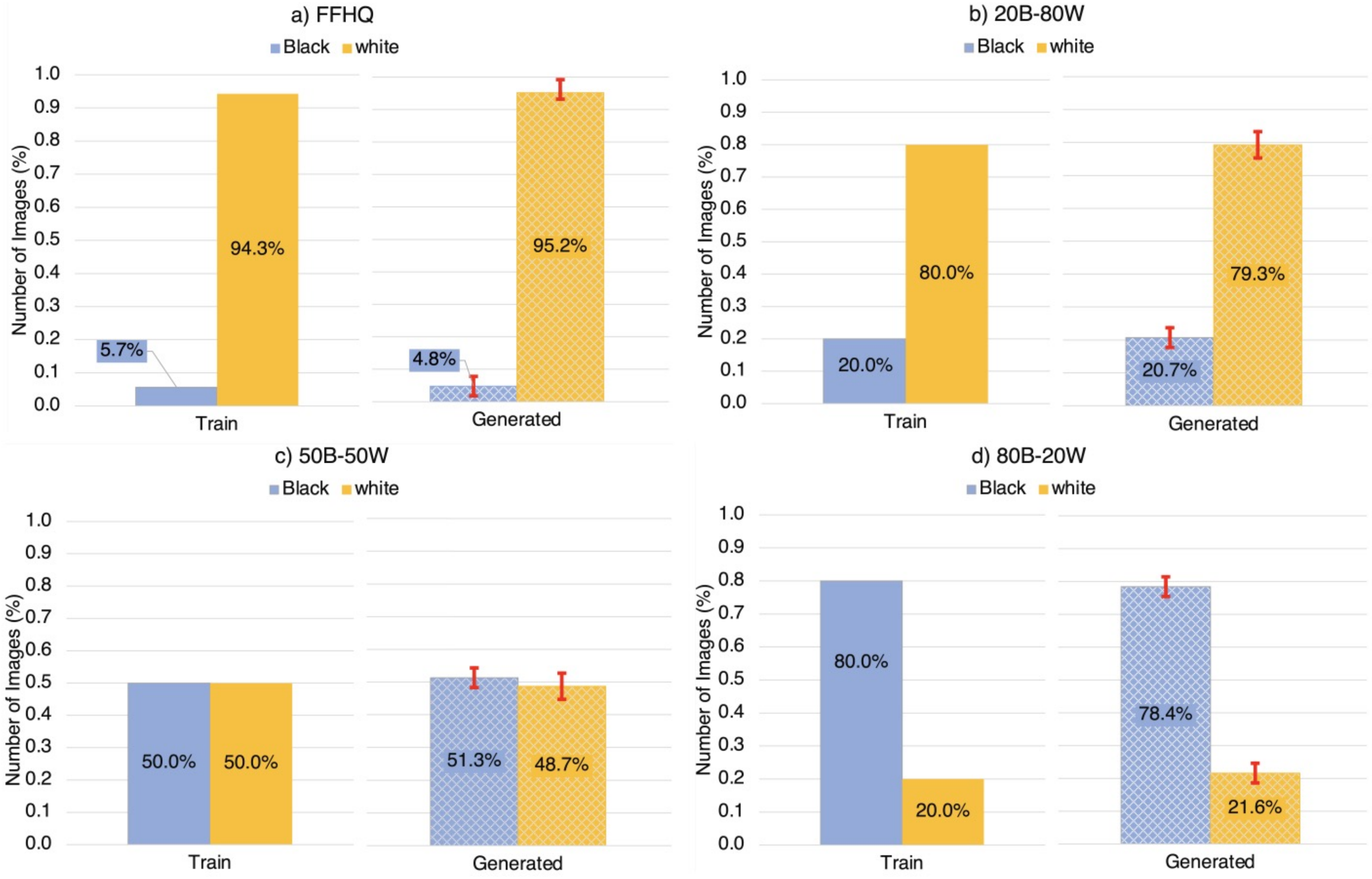


Image Super-resolution system PULSE [Menon et al., CVPR 2020]

GAN models



Studying Bias in GANs through the Lens of Race [Maluleke et al., 2022]

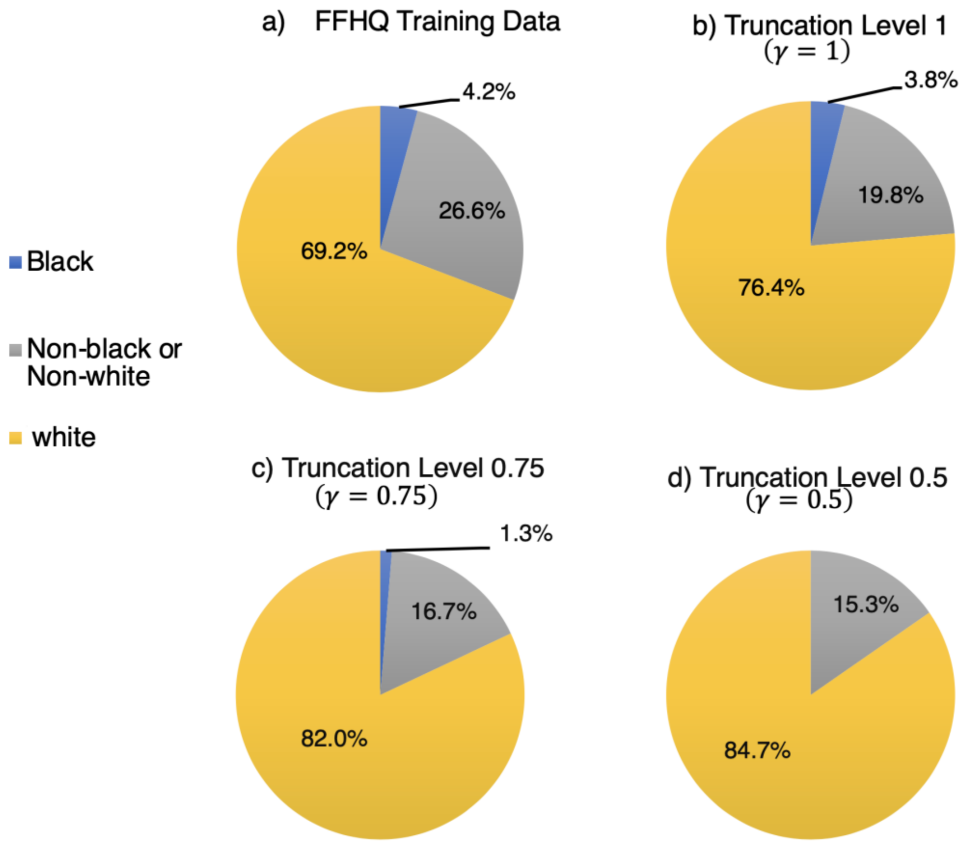
GAN models

Truncation Trick reduces diversity

$$w' = \gamma \, w + (1 - \gamma) \, \bar{w}$$

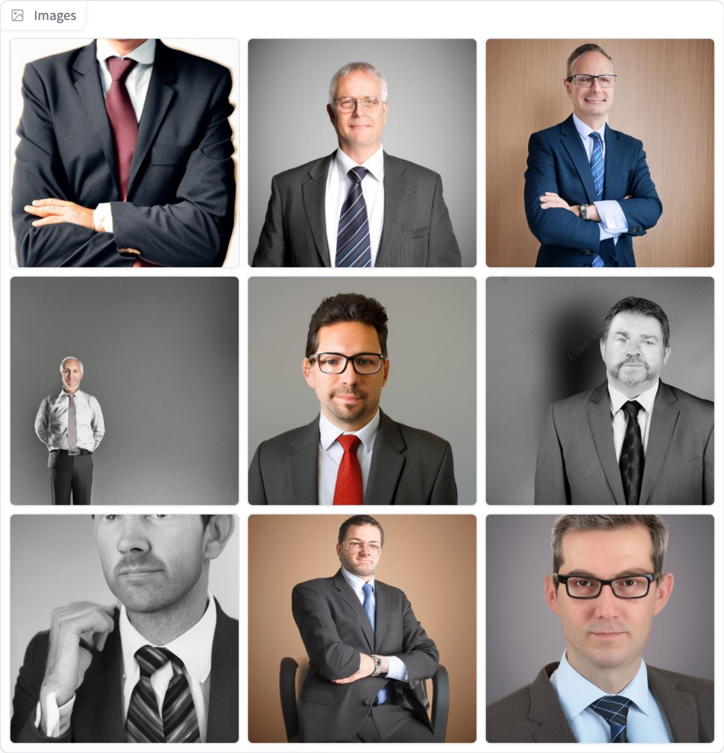
↑ Sampled code

↑ Average code

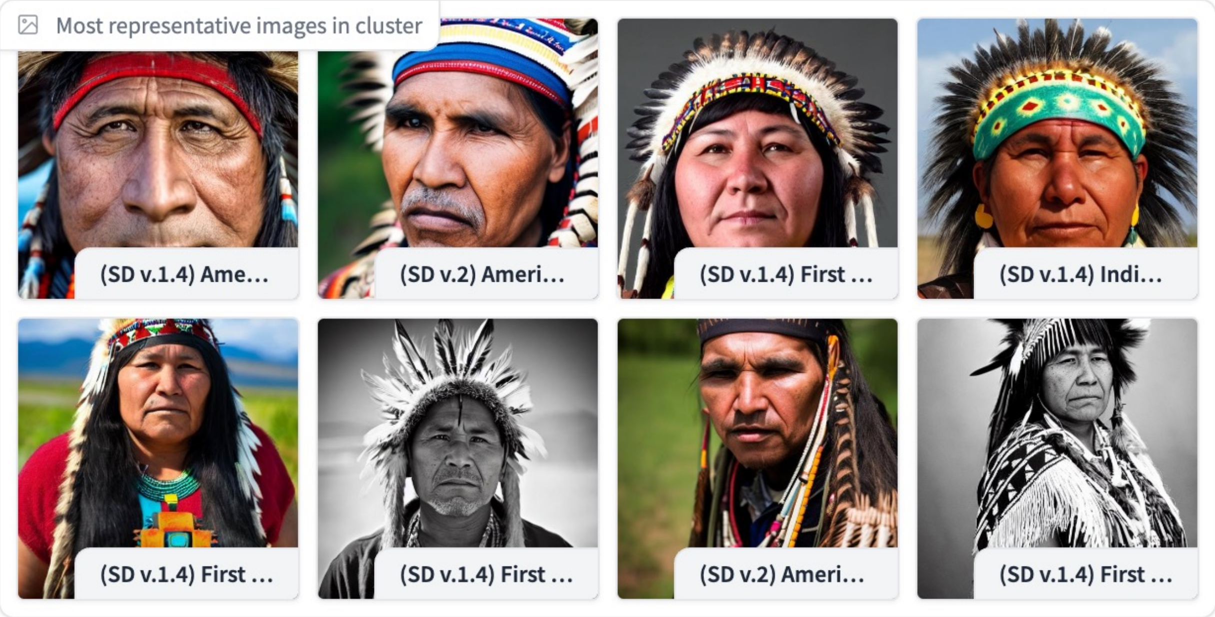


Studying Bias in GANs through the Lens of Race [Maluleke et al., 2022]

Text-to-Image Models



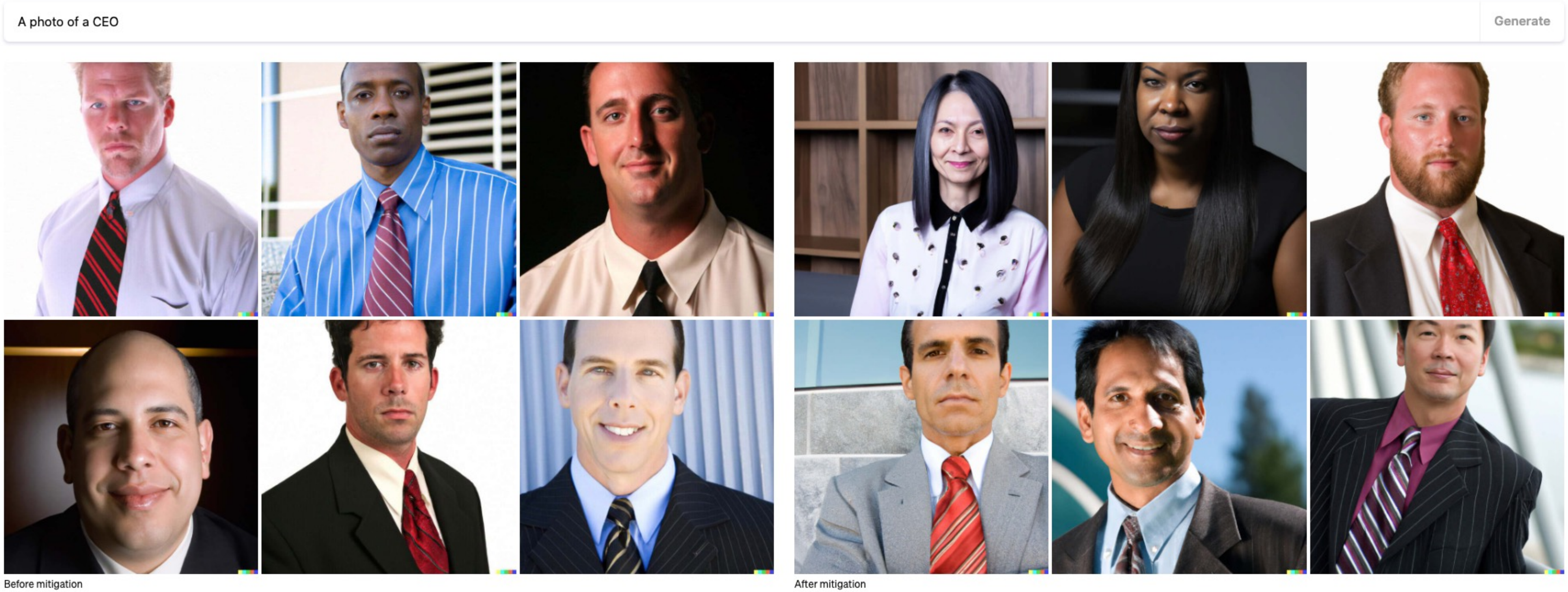
Managers



Native Americans

https://www.technologyreview.com/2023/03/22/1070167/these-news-tool-let-you-see-for-yourself-how-biased-ai-image-models-are/

Text-to-Image Models (quick fixes?)



A photo of a CEO

Before mitigation

After mitigation

https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2

Quick fixes or long-term solutions?