Seeing is NOT! Believing: Generating and Detecting Fakes

teddy bears mixing sparkling chemicals as mad scientists in a steampunk style –DALL-E 2

Computational Photography

Derek Hoiem, University of Illinois

Kinds of fakes

Danger Level

- Traditional CG

Yellow: Hard to make, easy to detect automatically

- Manipulated images
  - Photoshop
  - Image-based relighting, etc.

Orange: Easy to make for images, hard for video; harder to detect automatically

- Deep fakes

Red: Very easy to make for images or video; hard to detect automatically

CG vs. Real: Can you do it?

- http://area.autodesk.com/fakeorfoto/
- I can’t! (I got 4/12 this time)

Detecting Fakes -- Why It Matters: Trust and Information

- Top AI researchers race to detect 'deepfake' videos: 'We are outgunned'
- The Impact Of Fake Images
- Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes

"On April 14, 2020, the hashtag #TellTheTruthBelgium caught media attention. A video showed a nearly 5 min speech of Belgian premier Sophie Wilmès, depicting the COVID-19 pandemic as a consequence of environmental destruction. The environmental movement Extinction Rebellion used deepfake technology to alter a past address to the nation that Sophie Wilmès held previously (Extinction Rebellion, 2020; Galindo, 2020)."

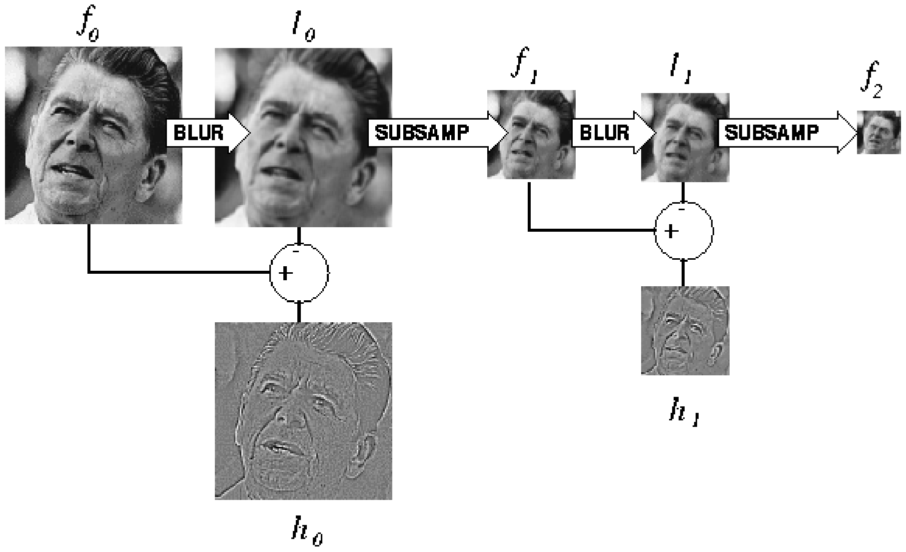
Automatically Detecting CG

- Sketch of approach
  - Intuition: natural images have predictable statistics (e.g., power law for frequency); CG images may have different statistics due to difficulty in creating detail
  - Decompose the image into wavelet coefficients and compute statistics of these coefficients

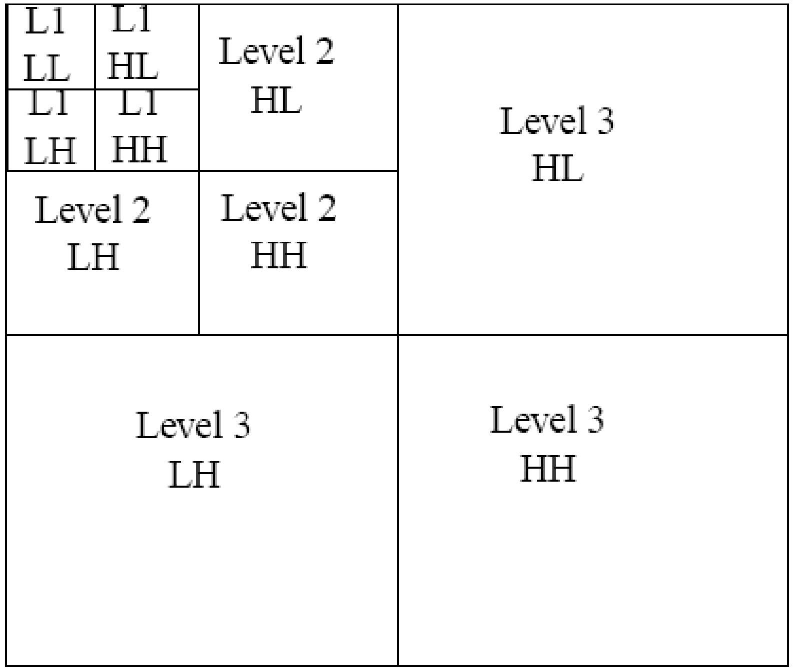
Lyu and Farid 2005: “How Realistic is Photorealistic?”

2D Wavelets

Kind of like the Laplacian pyramid, except broken down into horizontal, vertical, and diagonal frequency



Laplacian Pyramid



Wavelet Pyramid

2D Wavelet Transform

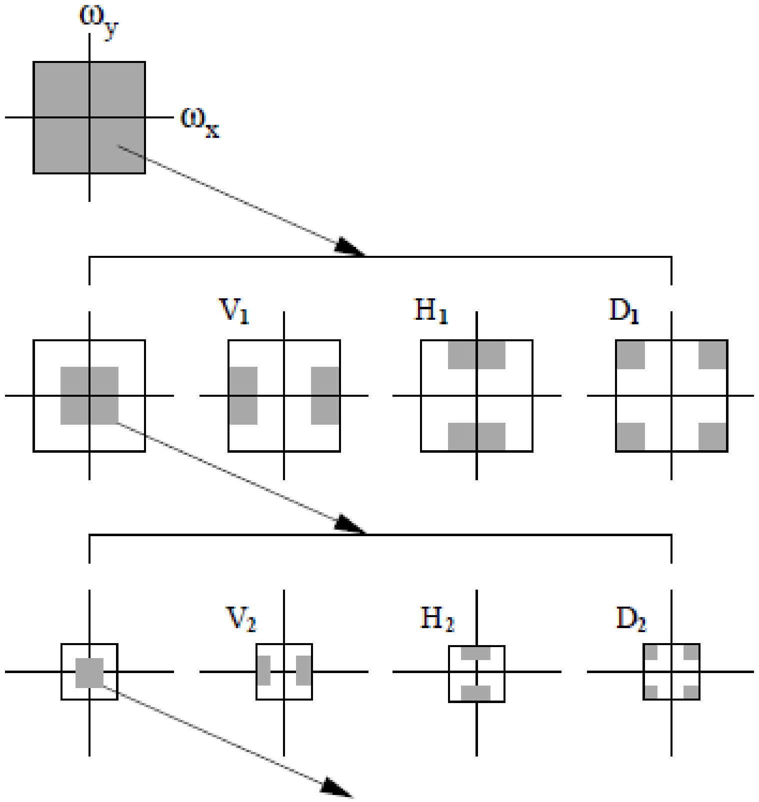
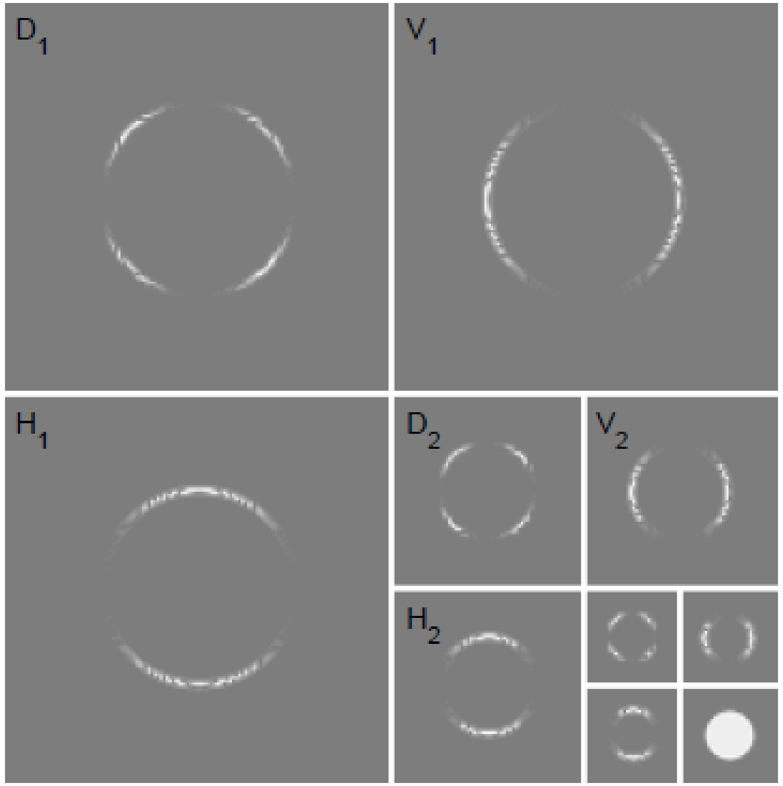


Illustration of procedure



Wavelet decomposition of disc image

Figure from Lyu and Farid 2005: “How Realistic is Photorealistic?”

Automatically Detecting CG

- Sketch of approach
  - Intuition: natural images have predictable statistics (e.g., power law for frequency); CG images may have different statistics due to difficulty in creating detail
  - Decompose the image into wavelet coefficients and compute statistics of these coefficients
  - Train a classifier to distinguish between CG and Real based on these features
    - Train RBF SVM with 32,000 real images and 4,800 fake images
    - Real images from http://www.freefoto.com
    - Fake images from http://www.raph.com and http://www.irtc.org/irtc/

Lyu and Farid 2005: “How Realistic is Photorealistic?”

Results

- 98.8% test accuracy on real images
- 66.8% test accuracy on fake images
- 10/14 on fakeorfoto.com

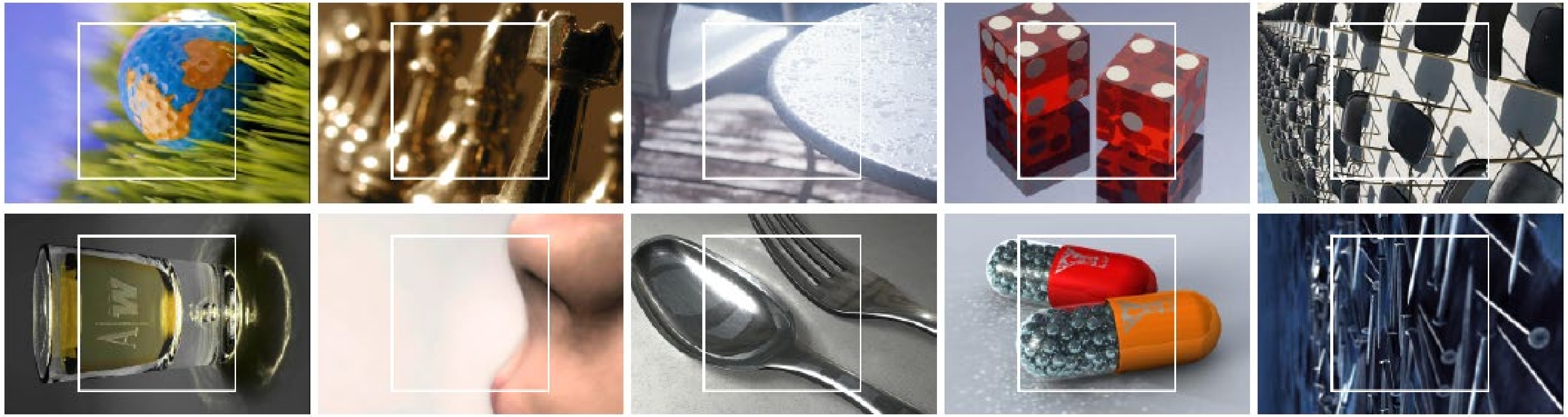
Lyu and Farid 2005: “How Realistic is Photorealistic?”

Results

- Fake-or-photo.com: Correct

Real Photos

CG



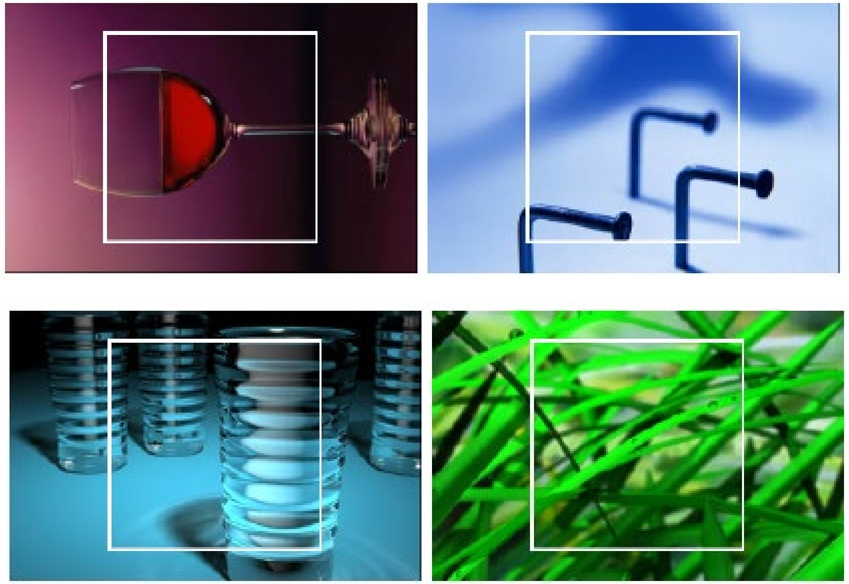
Lyu and Farid 2005: “How Realistic is Photorealistic?”

Results

- Fake-or-photo.com: Wrong

Real photos misclassified as CG

CG misclassified as real photos



Lyu and Farid 2005: “How Realistic is Photorealistic?”

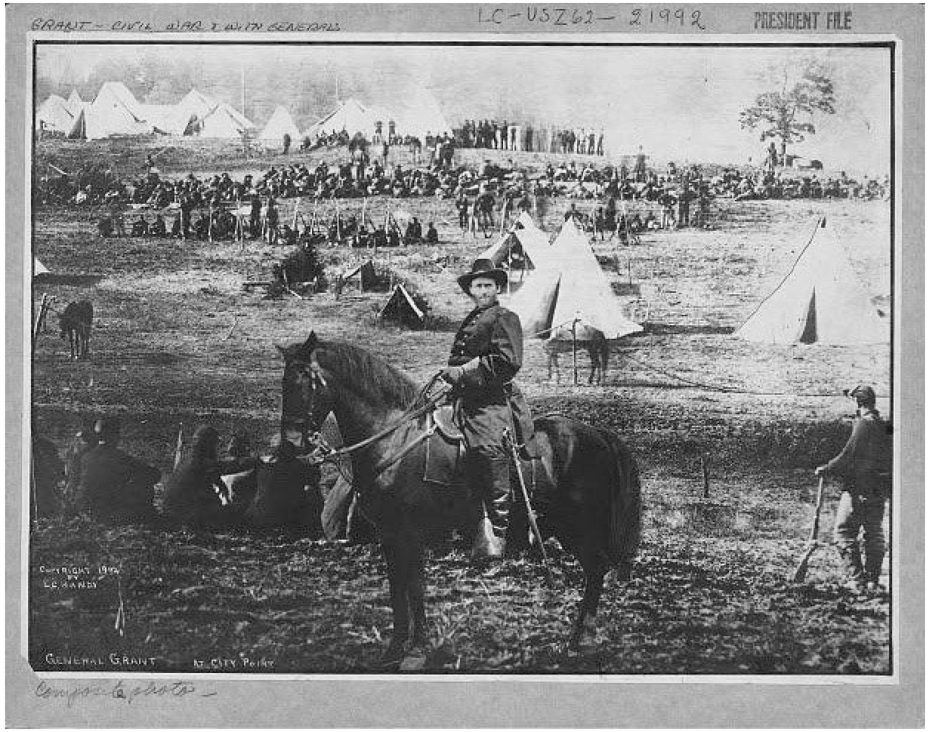
Photographic forgeries are an old problem

Examples collected by Hany Farid: http://www.fourandsix.com/photo-tampering-history/ (site no longer available)

Iconic Portrait of Lincoln (1860)

“While photographs may not lie, liars may photograph.”

Lewis Hine (1909)



General Grant in front of Troops (1864)

Mussolini in a Heroic Pose (1942)

1950: Doctored photo of Senator Tydings talking with Browder, the leader of the communist party, contributed to Tydings' electoral defeat

Gang of Four are removed (1976)

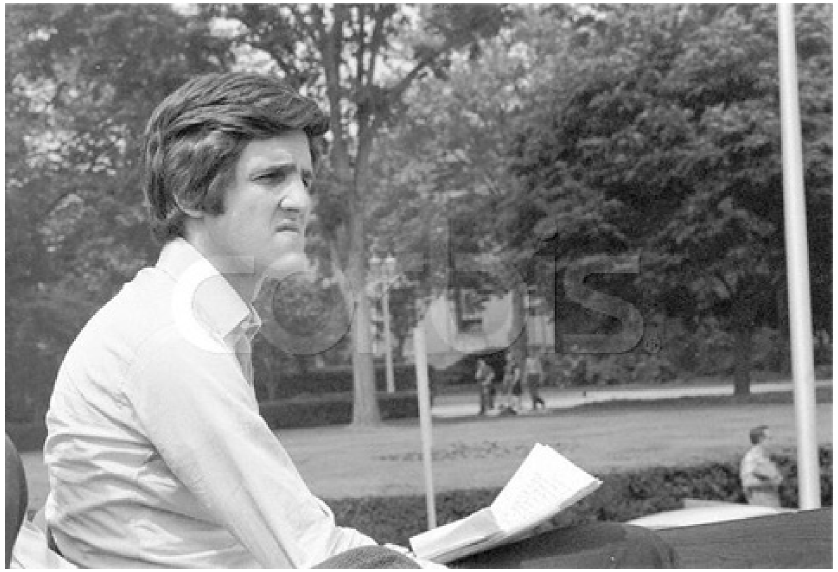
1989 composite of Oprah and Ann-Margret (without either’s permission)

Photo from terrorist attack in 1997 in Hatshepsut, Egypt

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)

Caption: “Actress and Anti-war activist Jane Fonda speaks to a crowd of Vietnam veterans, as activist and former Vietnam vet John Kerry listens and prepares to speak next concerning the war in Vietnam.” (AP Photo)



Kerry at Rally for Peace 1971

Fonda at rally in 1972

2005: USA Today SNAFU

2006: Photo by Adnan Hajj of strikes on Lebanon (original on right)
Later, all of Hajj's photos were removed from AP and a photo editor was fired.

2007 Retouching is “completely in line with industry standards”

The French Magazine Paris Match altered a photograph of French President Nicolas Sarkozy by removing some body fat. (2007)

Claimed Photo

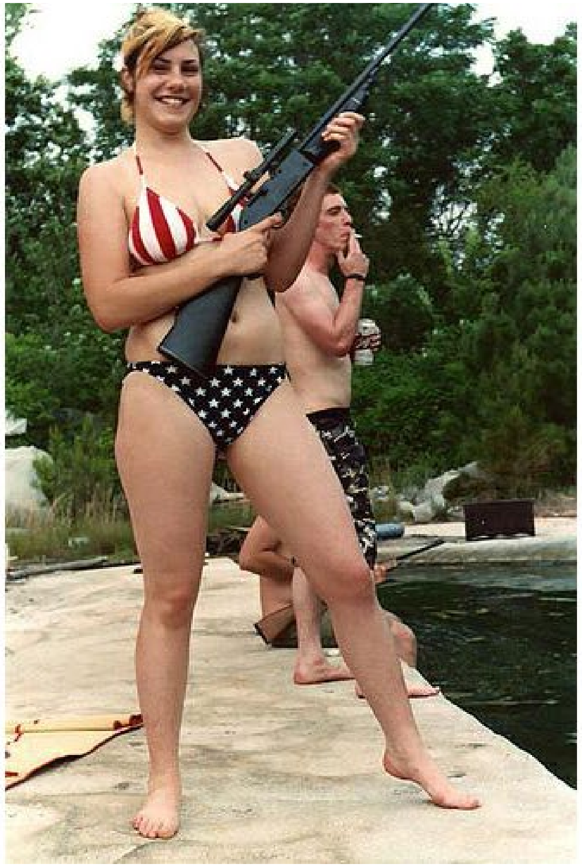
Poster

Overlay

2007: Zhou Zhenglong claimed to take 71 photos of the nearly extinct South China tiger

Similar scandal in 2011 from Terje Helleso who won Swedish Env. Prot. award

(2012) A Russian newspaper distributed by a pro-Kremlin group printed a photograph showing blogger/activist Aleksei Navalny standing beside Boris A. Berezovsky, an exiled financier being sought by Russian police.


2008

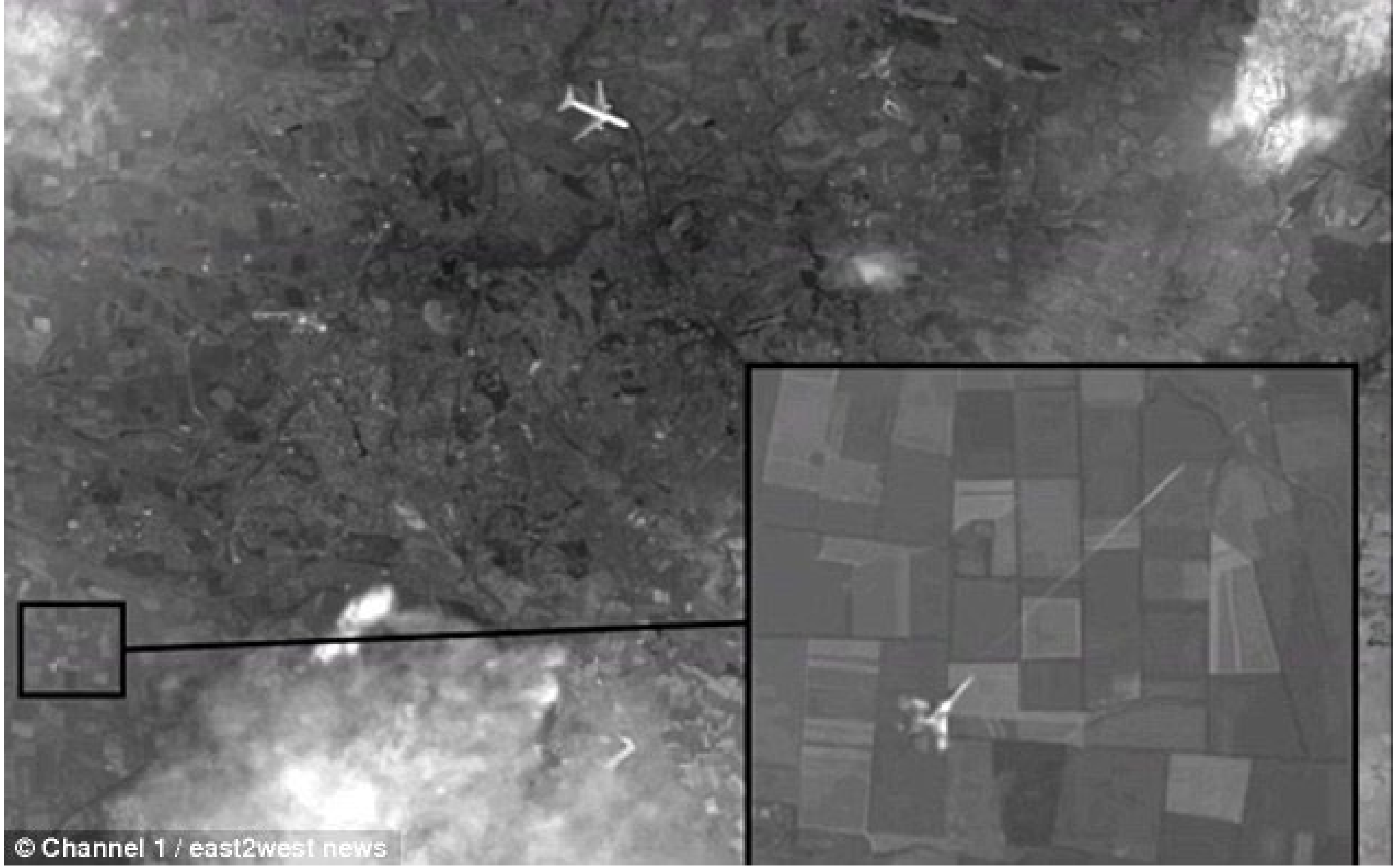
“Evidence” that Malaysian politician Jeffrey Wong Su En was knighted by the Queen (2010)

Cloning sand to remove shadow. Miguel Tovar – banned from AP, all his photos removed (2011)

Photo from Korean Central News Agency, determined to be composite (people don’t appear wet) – was attempt to get sympathy for North Korea to get more international aid

2013: fake floors, counter, appliances digitally added for listing in Luis Ortiz’s show “Million Dollar Listing New York”

A farmer from Hunan province, China was sentenced to 12 years in prison and fined 500,000 yuan after receiving 453,00 yuan (US$73,000) in blackmail payments out of an attempted 9.47 million yuan. He had mailed to more than 200 officials pornographic photos into which he had inserted them using photo editing software. He threatened to publicize the photos unless he was paid. One of the victims claimed that he made the demanded payment before he “sensed later on that the man inside the photo was actually not me.”



Nov 2014: Russian state media ran a story with “proof” that a Ukranian jet shot down the Malaysian airlines plane. Photo is composed of Google Earth imagery, Yandex maps, and a stock photo of a Boeing jet.

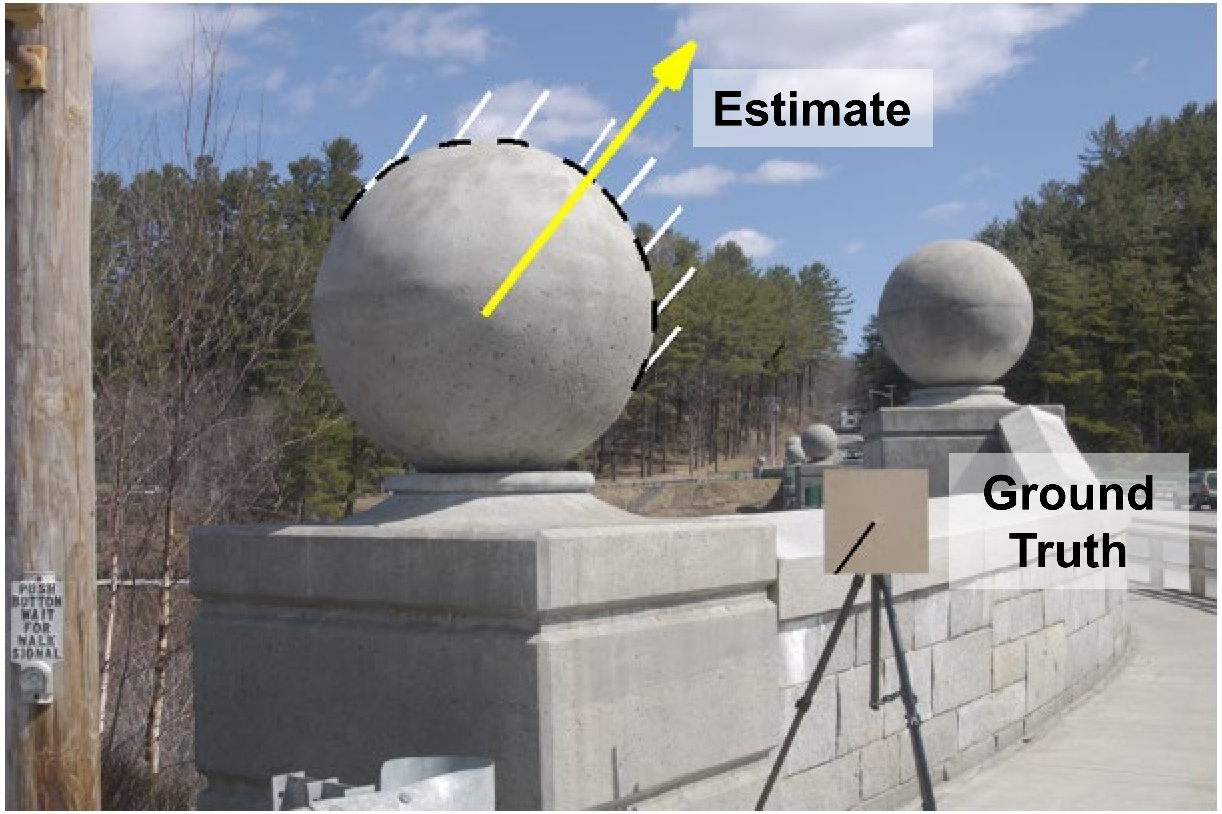
Detecting forgeries

- Work by Hany Farid and colleagues
- Method 1: 2D light from occluding contours

Estimating lighting direction

Method 1: 2D direction from occluding contour

- Provide at least 3 points on occluding contour (surface has 0 angle in Z direction)
- Estimate light direction from brightness



Estimating lighting direction

Estimating lighting direction

- Average error: 4.8 degrees

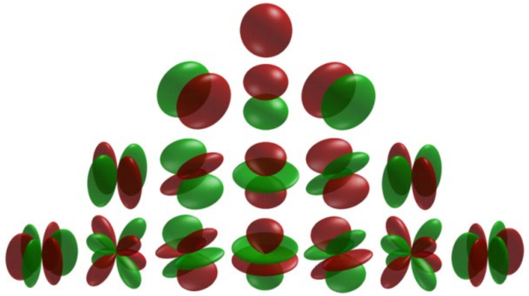
Method 2: Light from Eyes

Farid – “Seeing is not believing”, IEEE Spectrum 2009

Estimating Lighting from Eyes

Method 3: Complex light with spherical harmonics

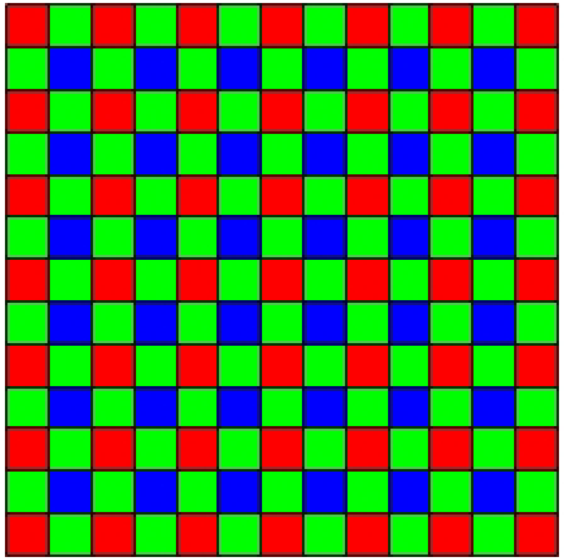
- Spherical harmonics parameterize complex lighting environment
- Same method as occluding contours, but need 9 points



Method 3: Complex light with spherical harmonics

Method 4: Demosaicking Prediction

- In demosaicking, RGB values are filled in based on surrounding measured values
- Filled in values will be correlated in a particular way for each camera
- Local tampering will destroy these correlations



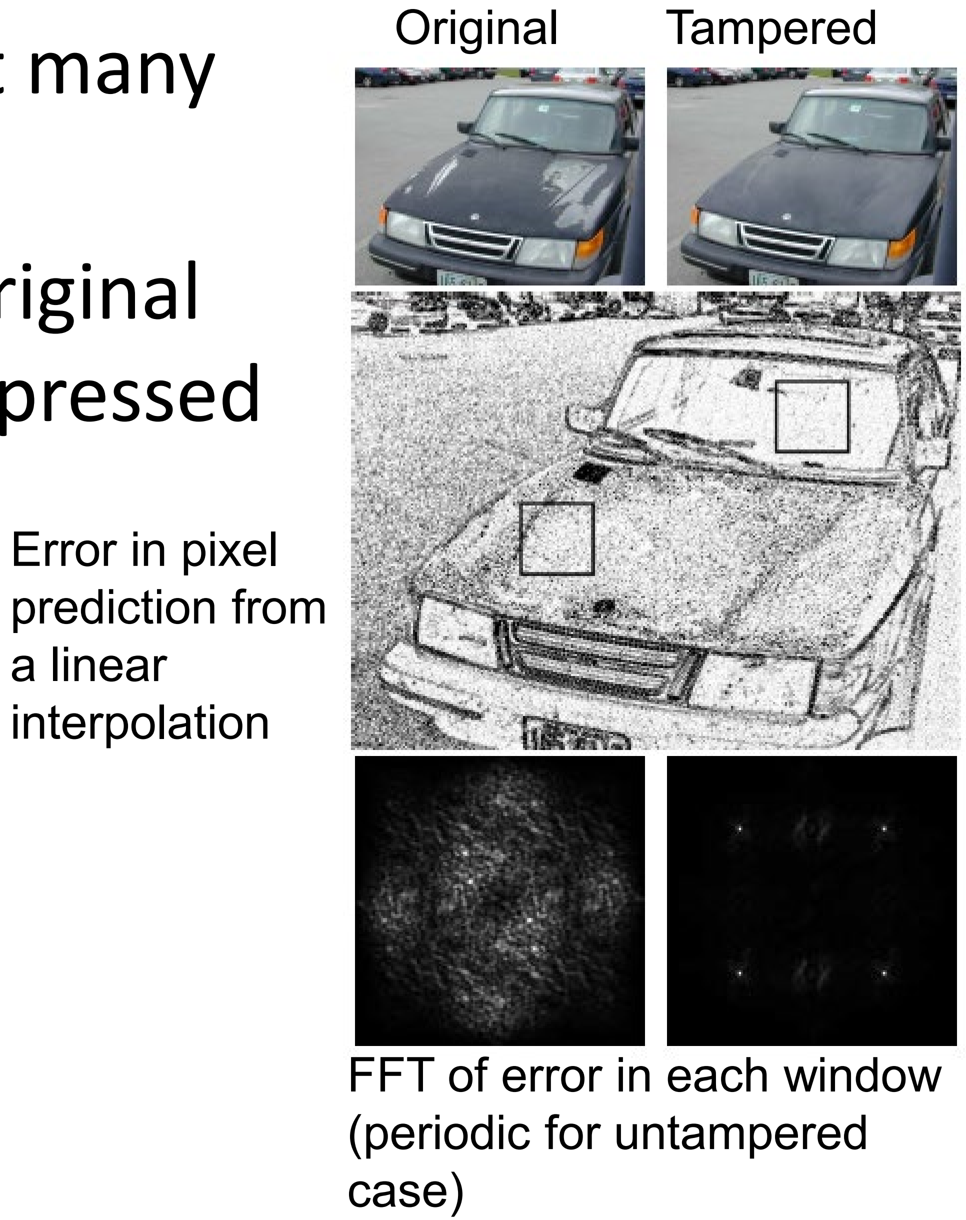
Bayer filter

© 2000 How Stuff Works

Farid: "Photo Fakery and Forensics" 2009

Demosaicking prediction

- Upside: can detect many kinds of forgery
- Downside: need original resolution, uncompressed image



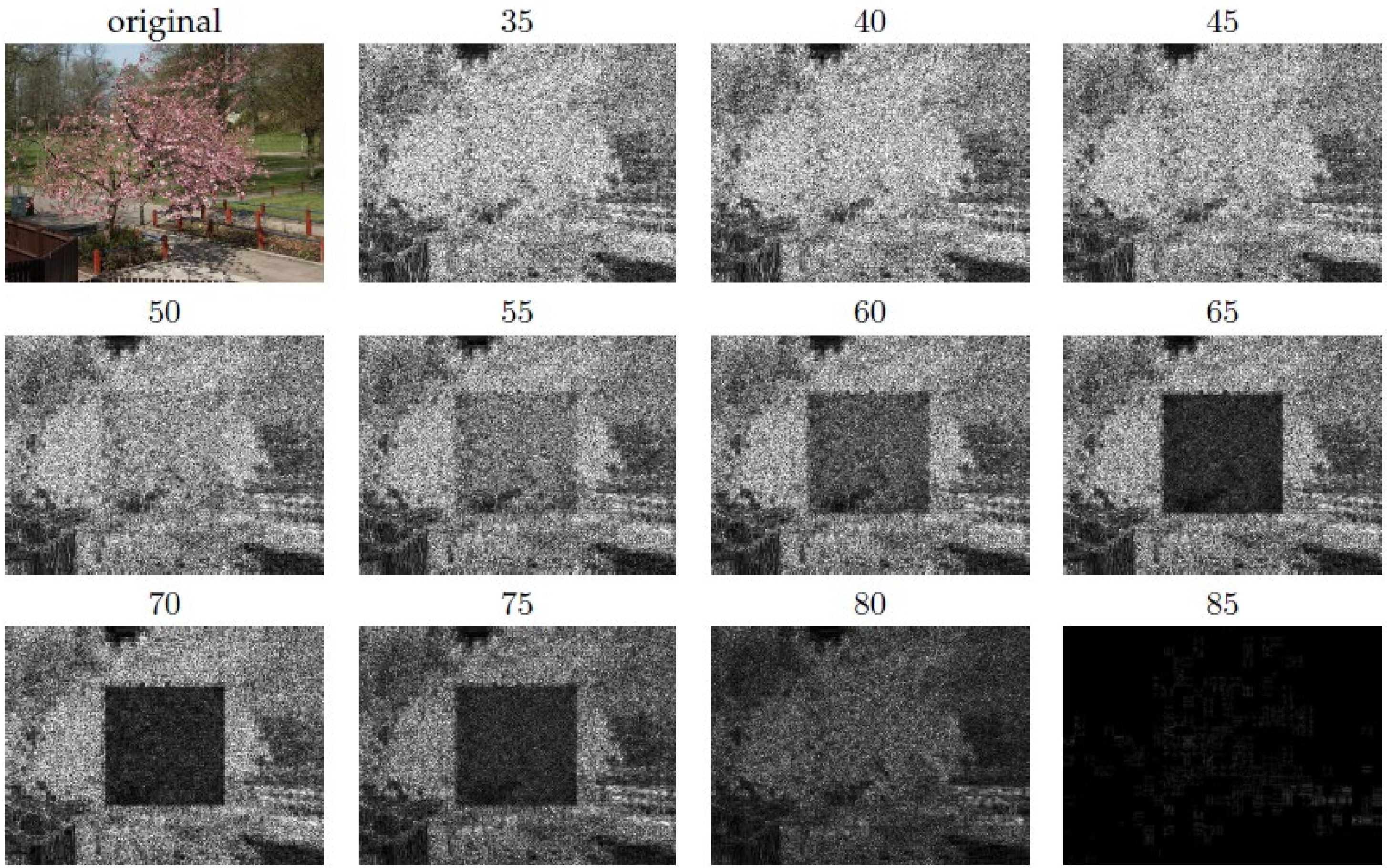
Method 5: JPEG Ghosts

- JPEG compresses 8x8 blocks by quantizing DCT coefficients to some level
  - E.g., coefficient value is 23, quantization = 7, quantized value = 3, error = 23-21=2
- Resaving a JPEG at the same quantization will not cause error, but resaving at a lower *or higher* quantization generally will
  - Value = 21; quantization = 13; error = 5
  - Value = 21; quantization = 4; error = 1

Farid: “Photo Fakery and Forensics” 2009

JPEG Ghosts

- Original is saved at 85 quality, center square is cut out and compressed at 65 quality; then image is resaved at given qualities



Pixel error for image saved at various JPEG qualities

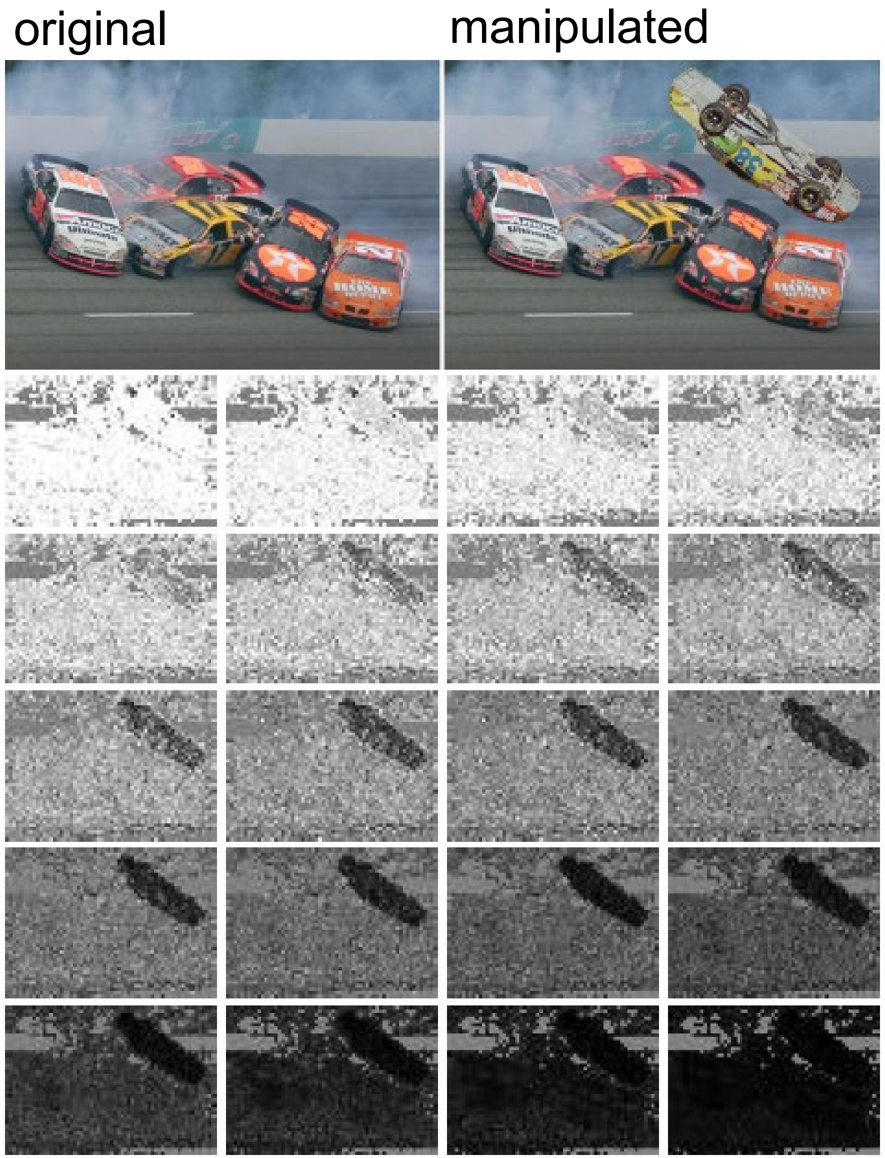
JPEG Ghosts

- If there is enough difference between the quality of the pasted region and the final saved quality, the pasted region can be detected with high accuracy

Table 2: JPEG ghost detection accuracy (%)

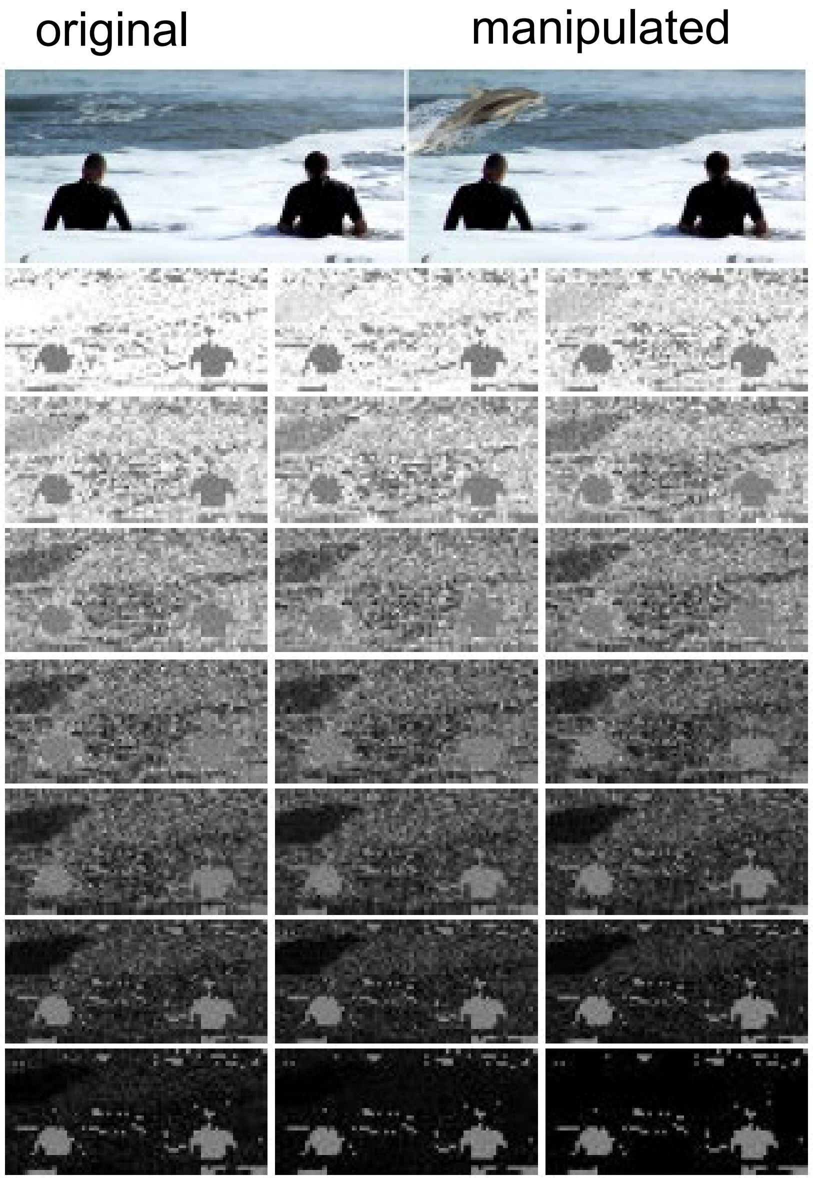
	$Q_1 - Q_0$					
size	0	5	10	15	20	25
200×200	99.2	14.8	52.6	88.1	93.8	99.9
150×150	99.2	14.1	48.5	83.9	91.9	99.8
100×100	99.1	12.6	44.1	79.5	91.1	99.8
50×50	99.3	5.4	27.9	58.8	77.8	97.7

JPEG Ghosts



Pixel error for manipulated image saved at various JPEG qualities

JPEG Ghosts



Pixel error for manipulated image saved at various JPEG qualities

Generating fake images with deep networks

- Pix2Pix
- CycleGAN
- StyleGAN
- Diffusion Networks

Slides adapted from Isola, Zhu

pix2pix: Image-to-Image Translation

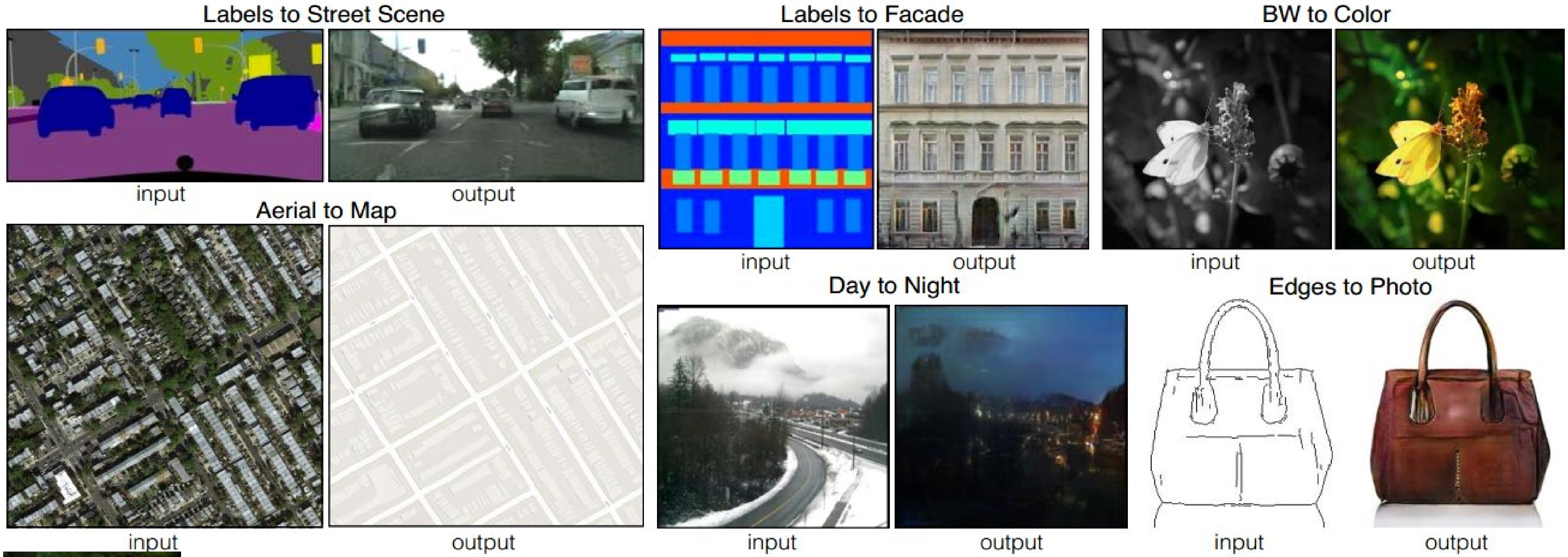
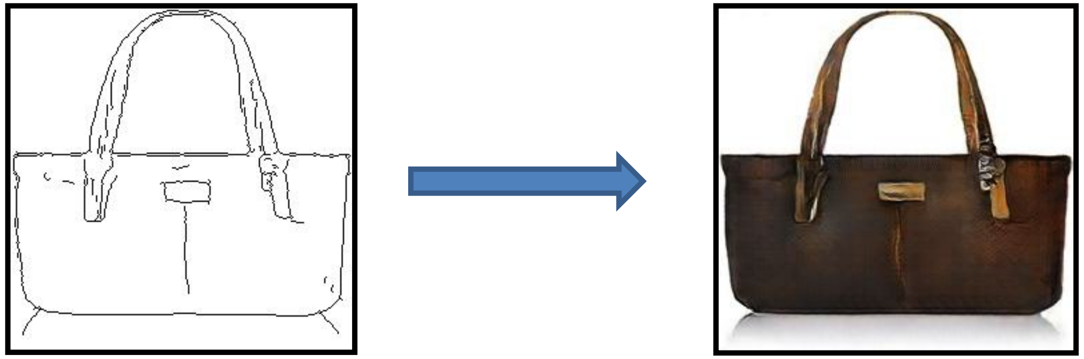


Image-to-image Translation with Conditional Adversarial Nets

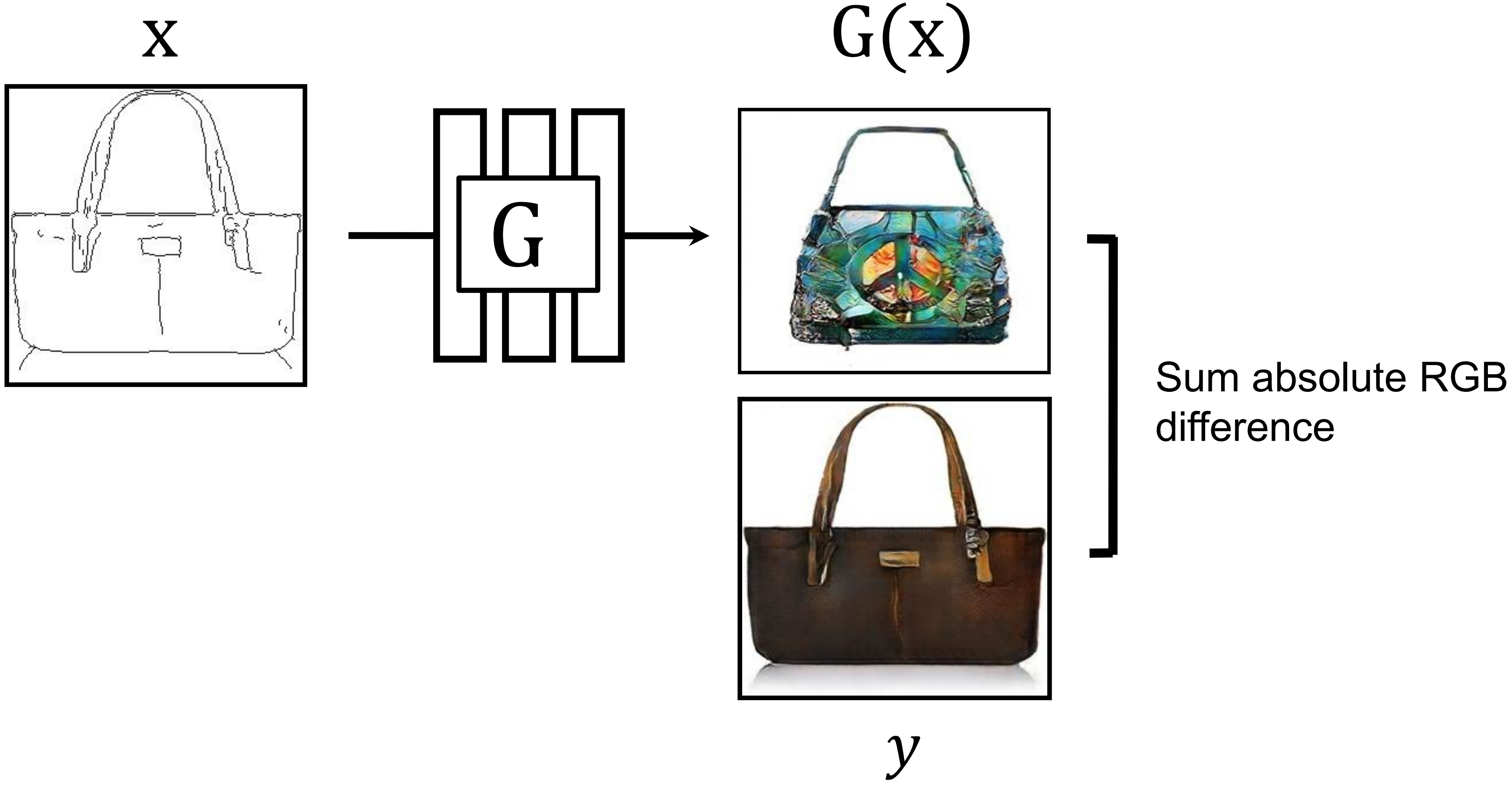
Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros. CVPR 2017

Image to image translation (pix2pix)

Train a conditional generator to translate from one image domain to another

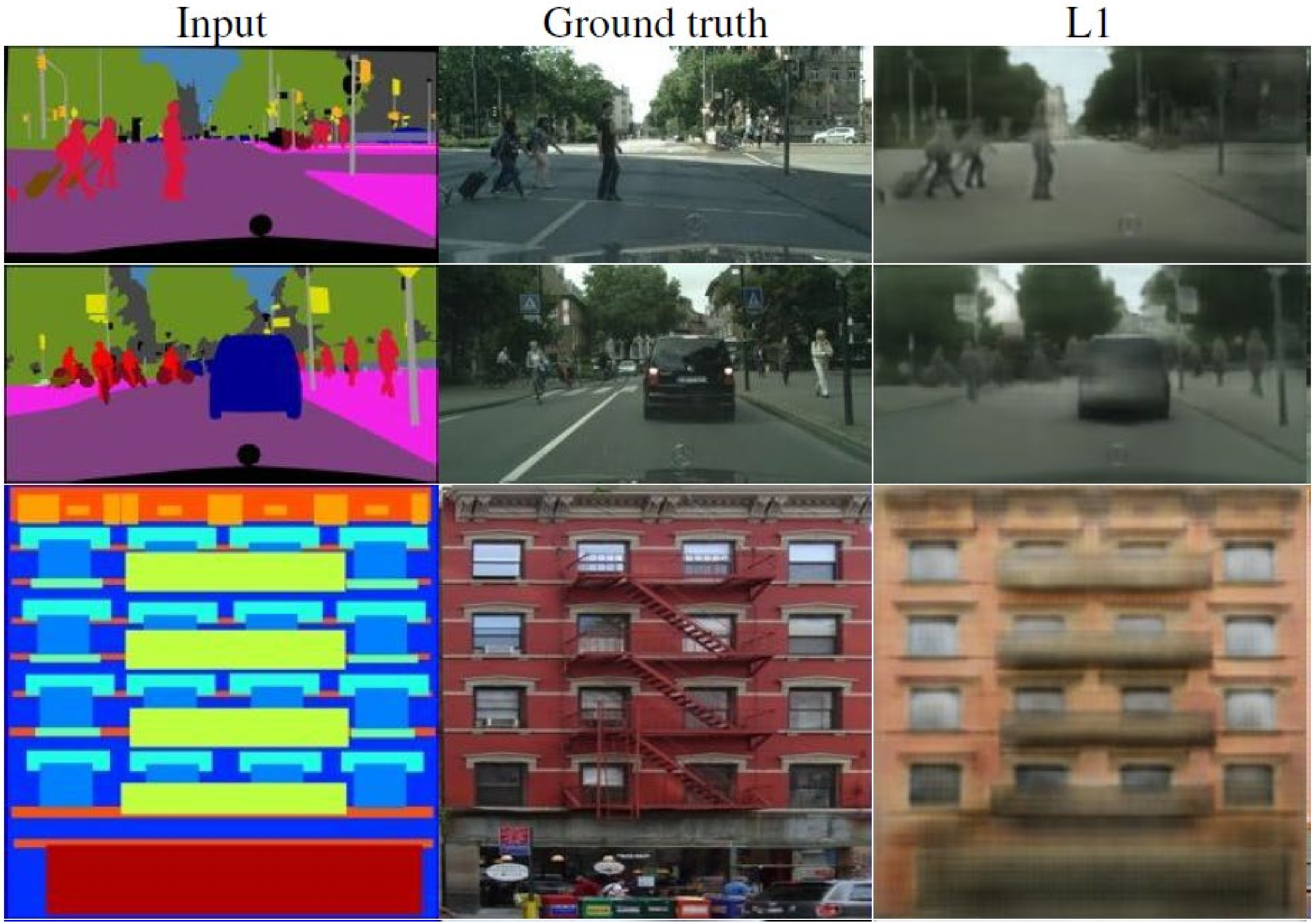


Objective 1: L1 Loss

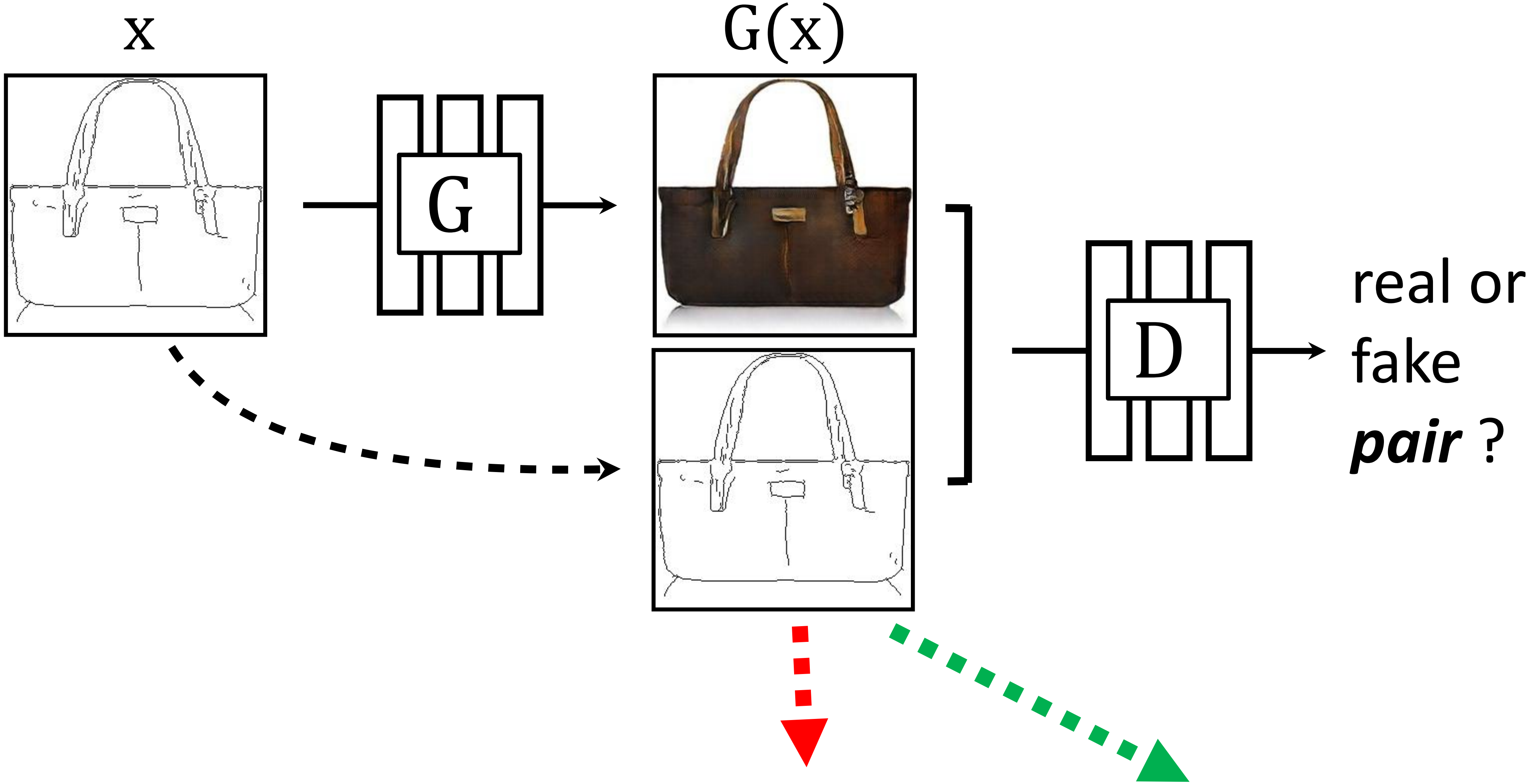


$$L_{\mathrm{L1}(G)} = \mathbb{E}_{x,y}\, ||y - G(x)||_1$$

L1 objective tends to produce slightly blurry results



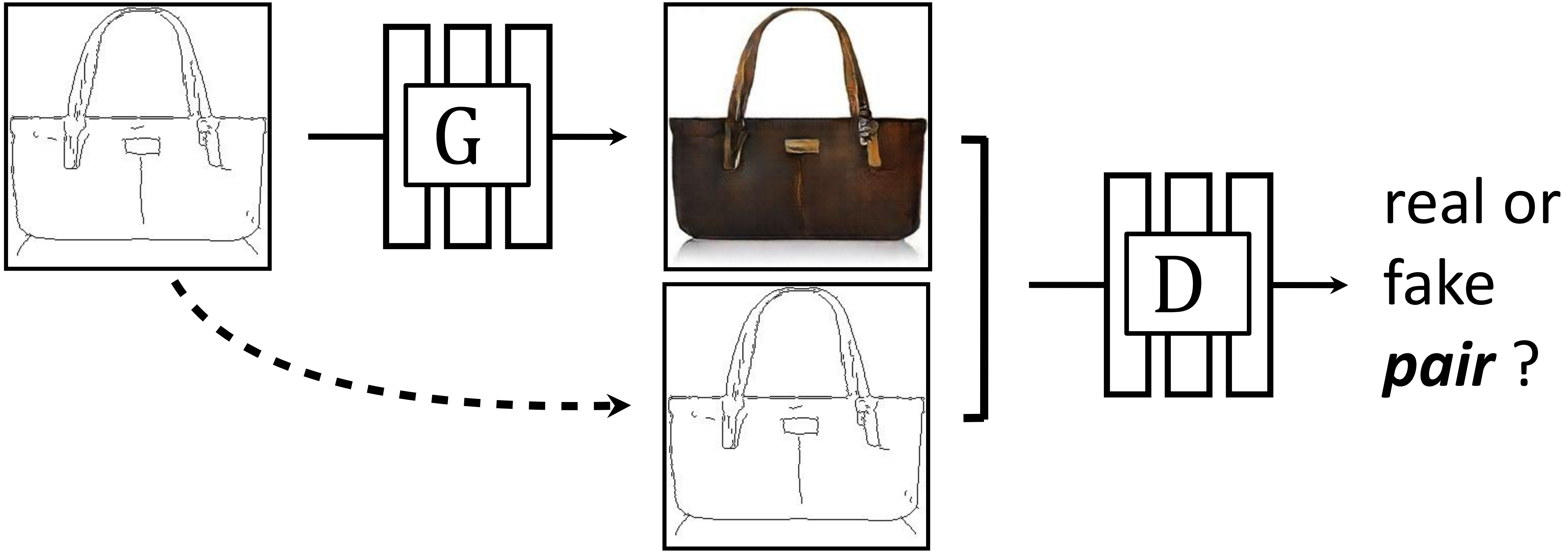
Objective 2: Paired Adversarial Loss



$$L_{GAN}(\mathrm{G}, \mathrm{D}) = \mathbb{E}_{x,y}[\log D\underbrace{(x, G(x))}_{\text{fake pair}} + \log(\underbrace{1 - D(x, y)}_{\text{real pair}})]$$

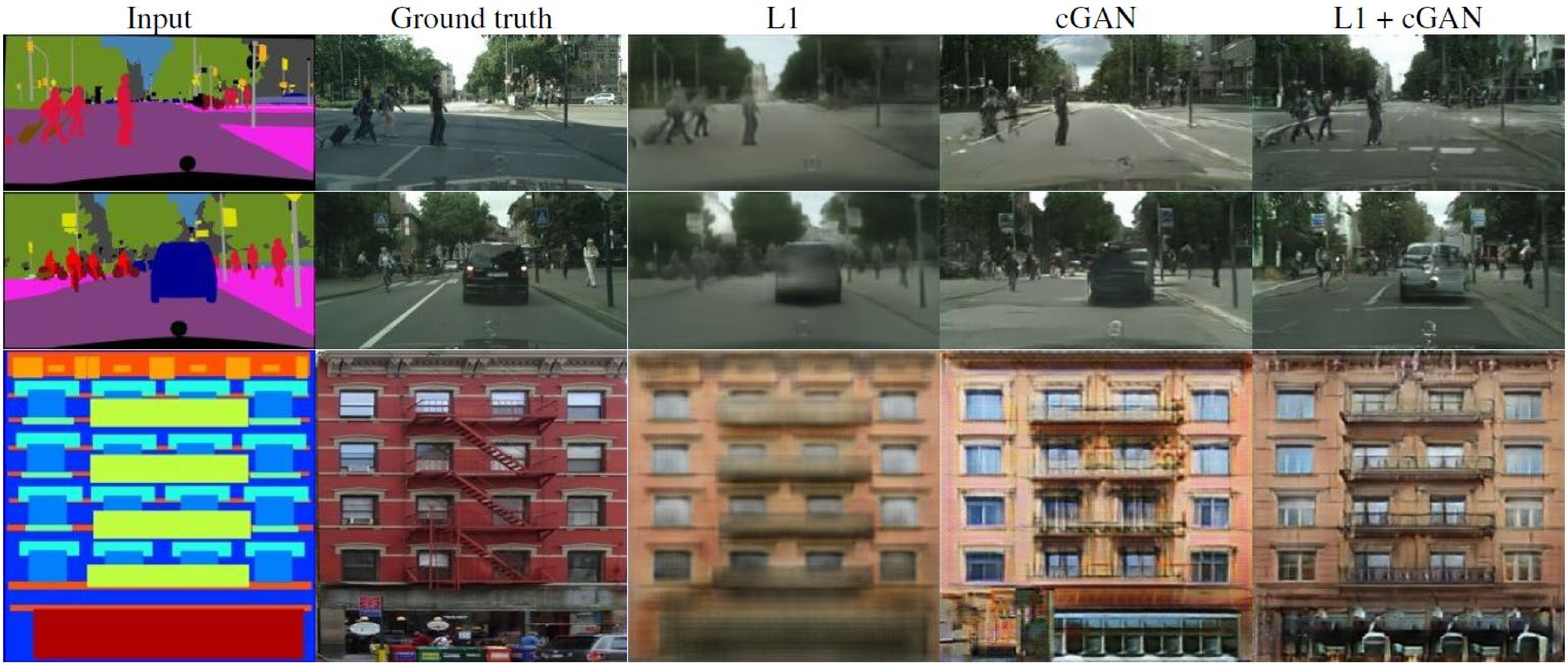
By itself, cGAN has some high texture artifacts

Combined Objective



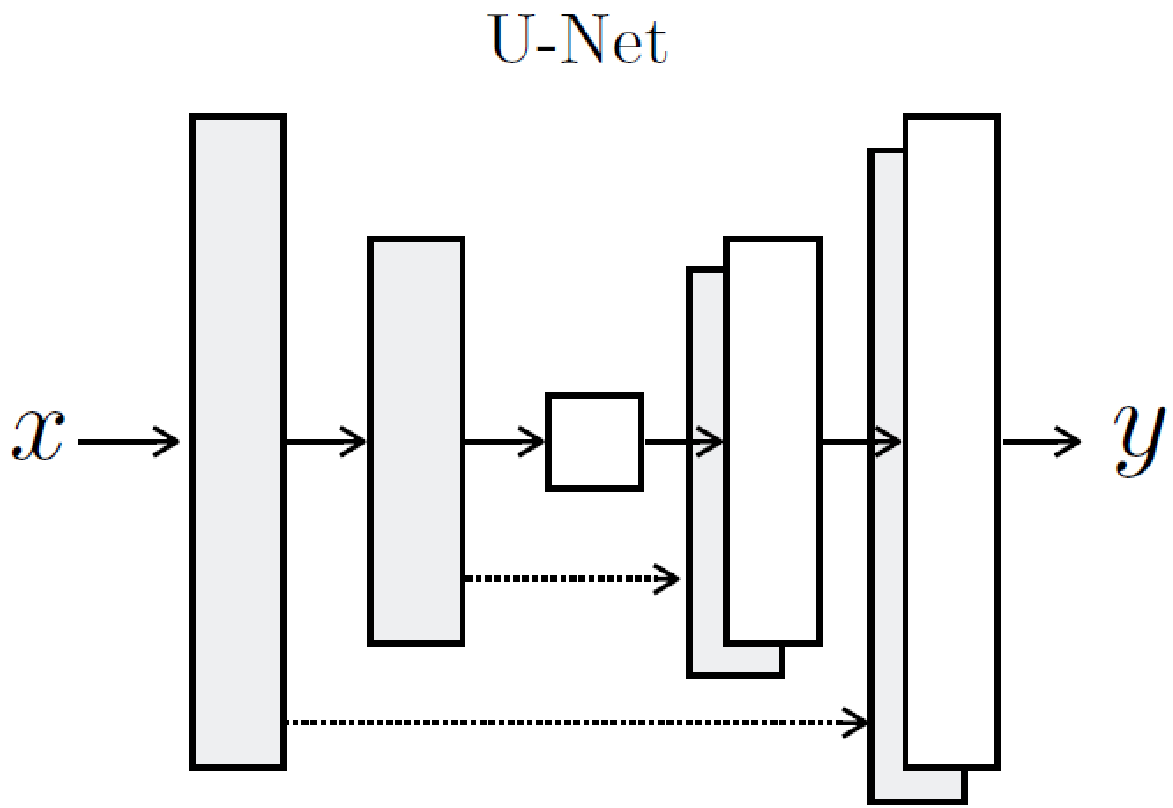
$$G^* = \min_G \max_D L_{GAN}(G, D) + \lambda L_1(G)$$

Combined objective works best



Design Choices

U-Net Encoder/Decoder helps preserve detail

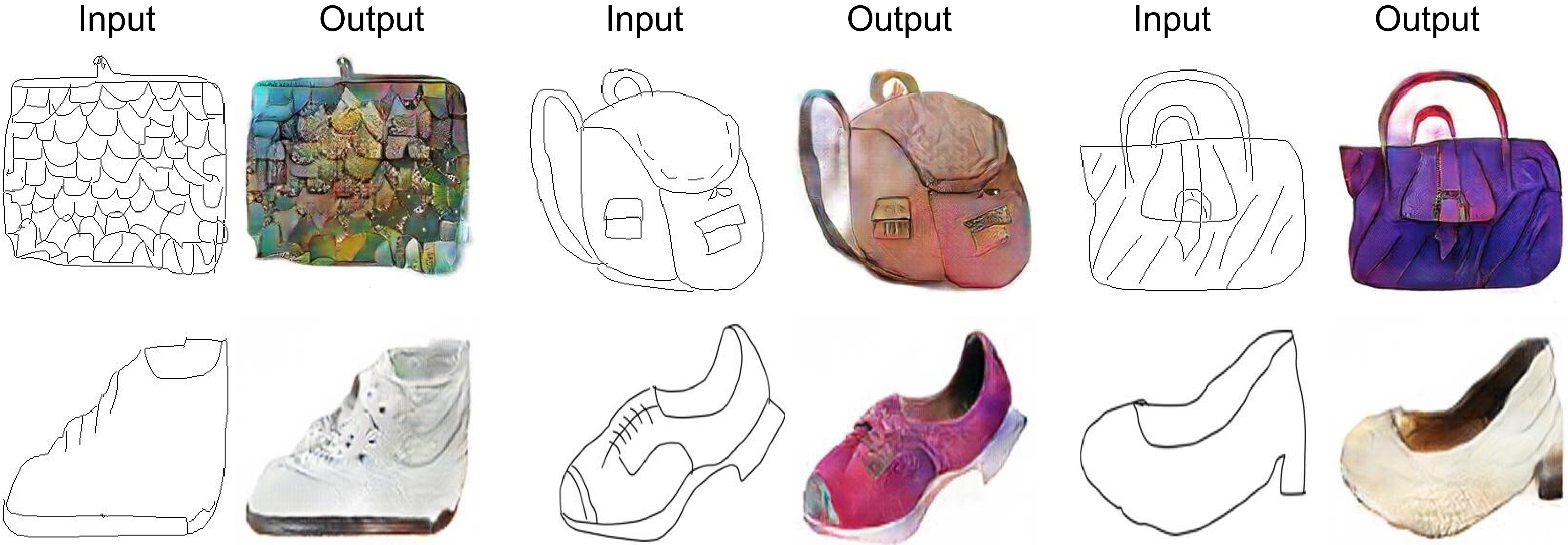


Design Choices

PatchGAN: Discriminator classifies NxN patches so that it focuses on details/texture that L1 loss doesn't capture

- NxN = 70x70 works well in experiments
- Average responses across patches

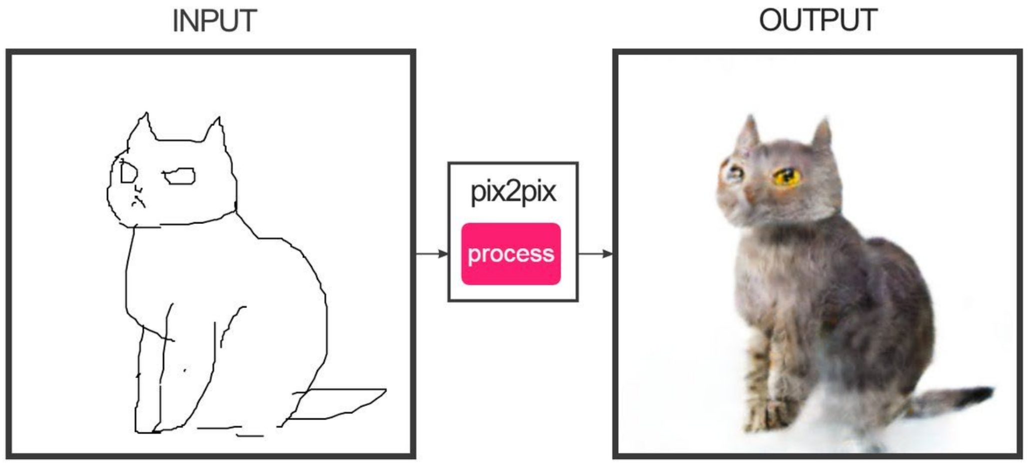
Sketches → Images



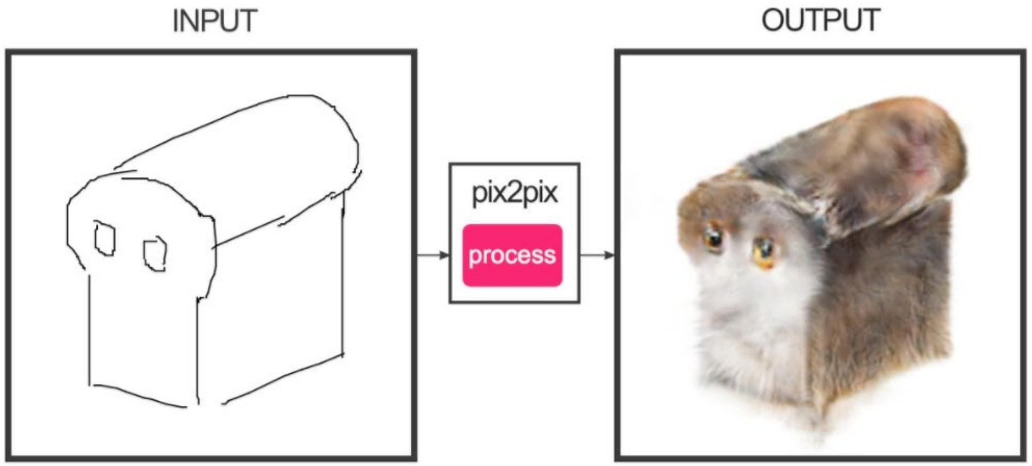
Trained on Edges → Images

Data from [Eitz, Hays, Alexa, 2012]

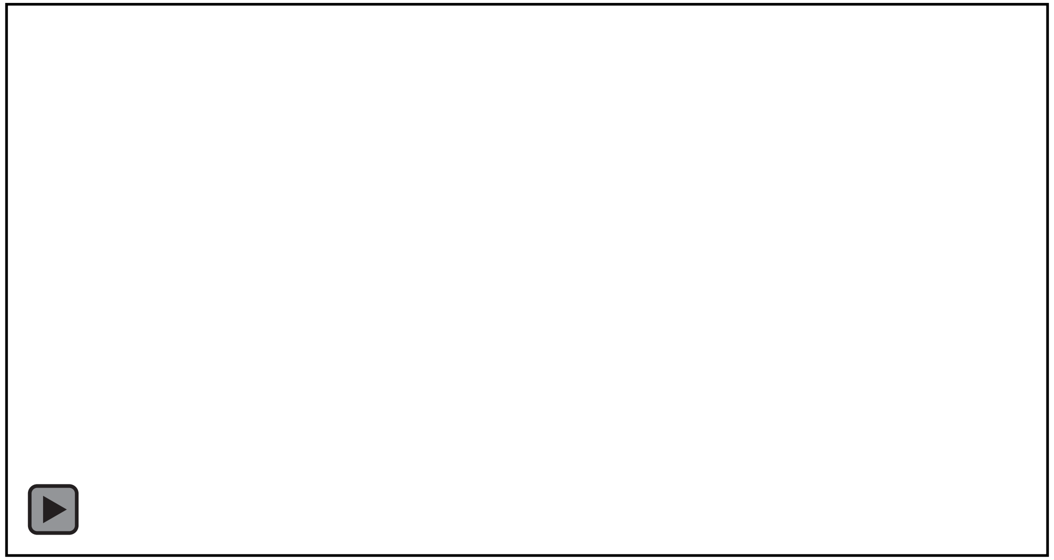
#edges2cats [Christopher Hesse]



@gods_tail



Ivy Tasi
@ivymyt

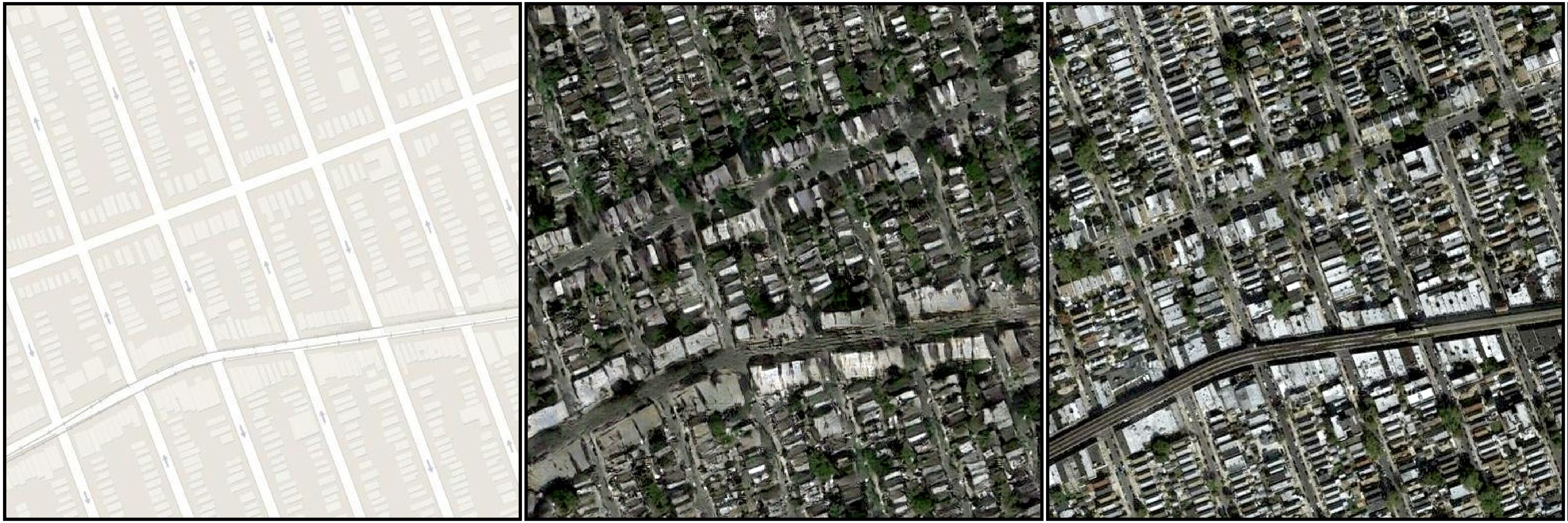


@matthematician

Vitaly Vidmirov @vvid

https://affinelayer.com/pixsrv/

Input Output Groundtruth



Data from [maps.google.com]

BW → Color

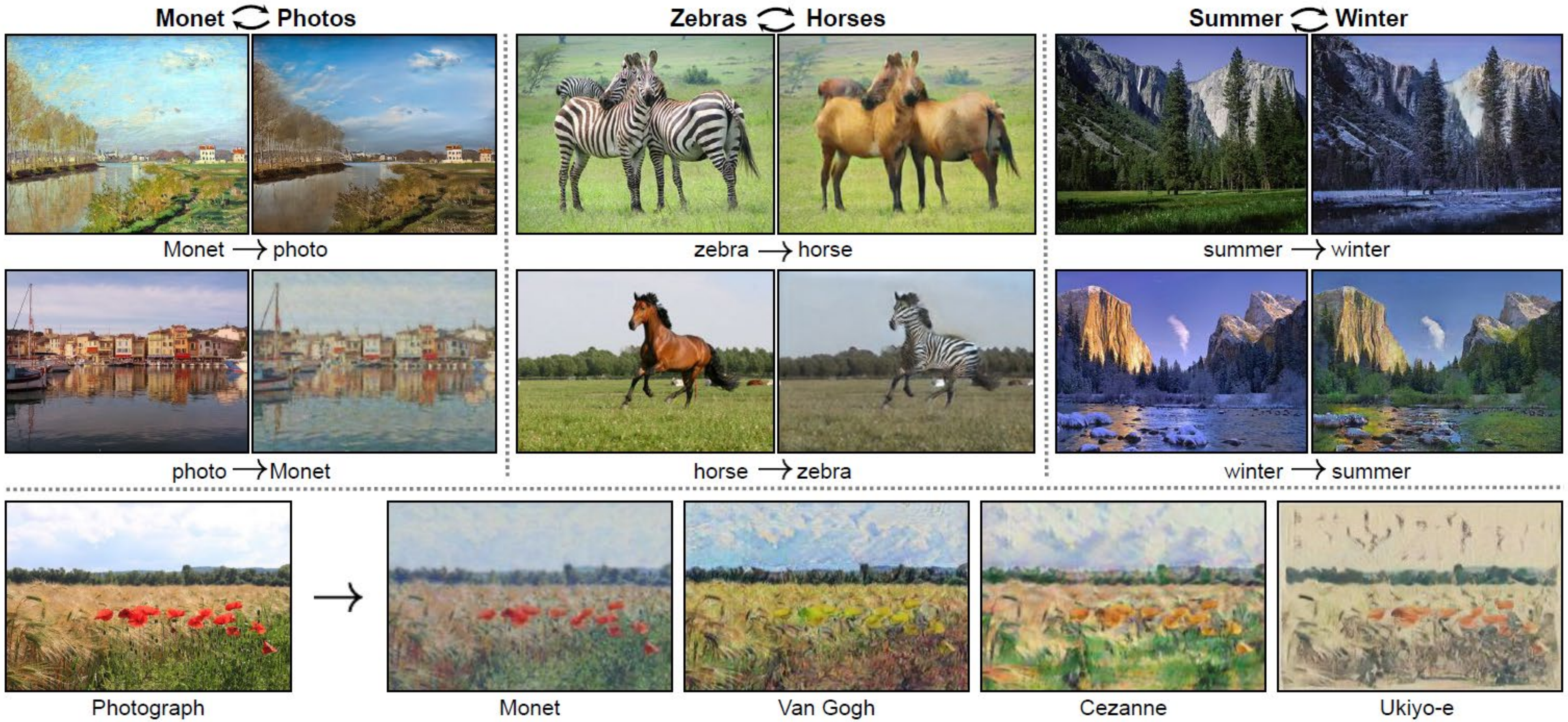
Input Output Input Output Input Output

Data from [Russakovsky et al. 2015]

Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

ICCV 2017

Jun-Yan Zhu* Taesung Park* Phillip Isola Alexei A. Efros
Berkeley AI Research (BAIR) laboratory, UC Berkeley



Cycle GAN

- Hard to get exact image domain translations for training, but easy to get unmatched sets of images
- Key idea: if you translate an image and then translate it back, you should get the original

English

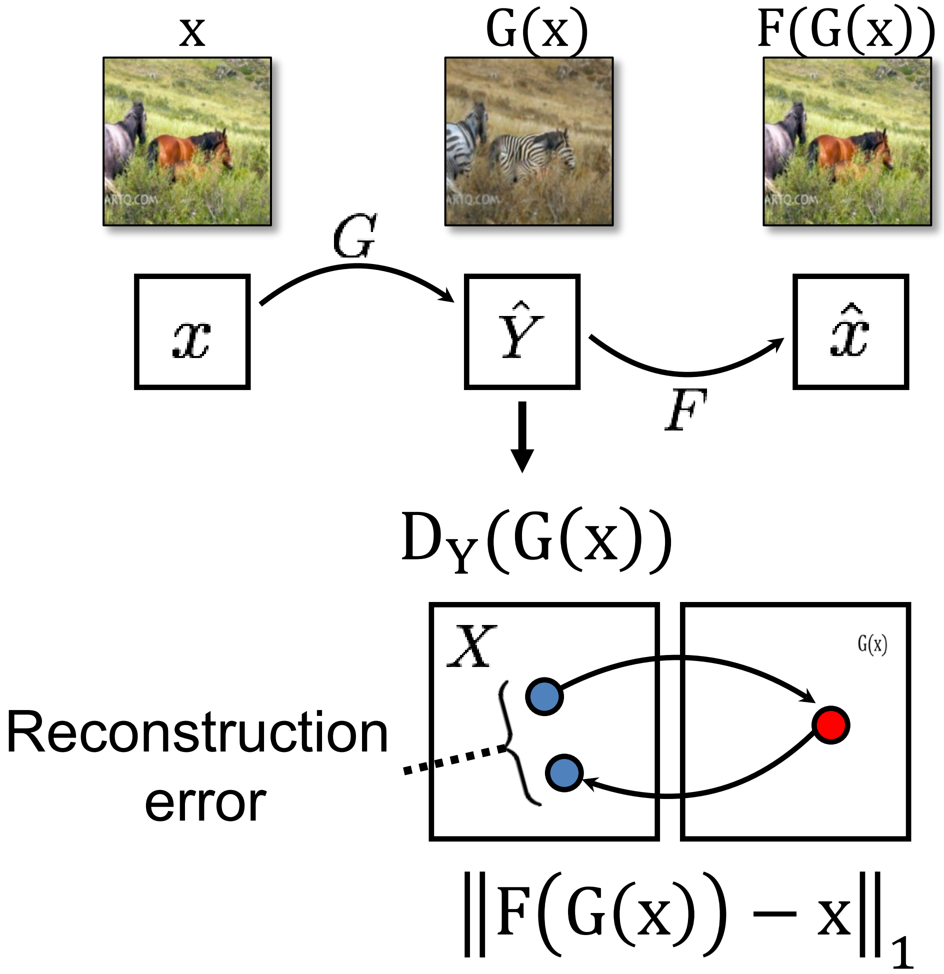
I gave my cat a warm bath

Arabic

أعطيت قطتي حمام دافئ

My cat was given a warm bath

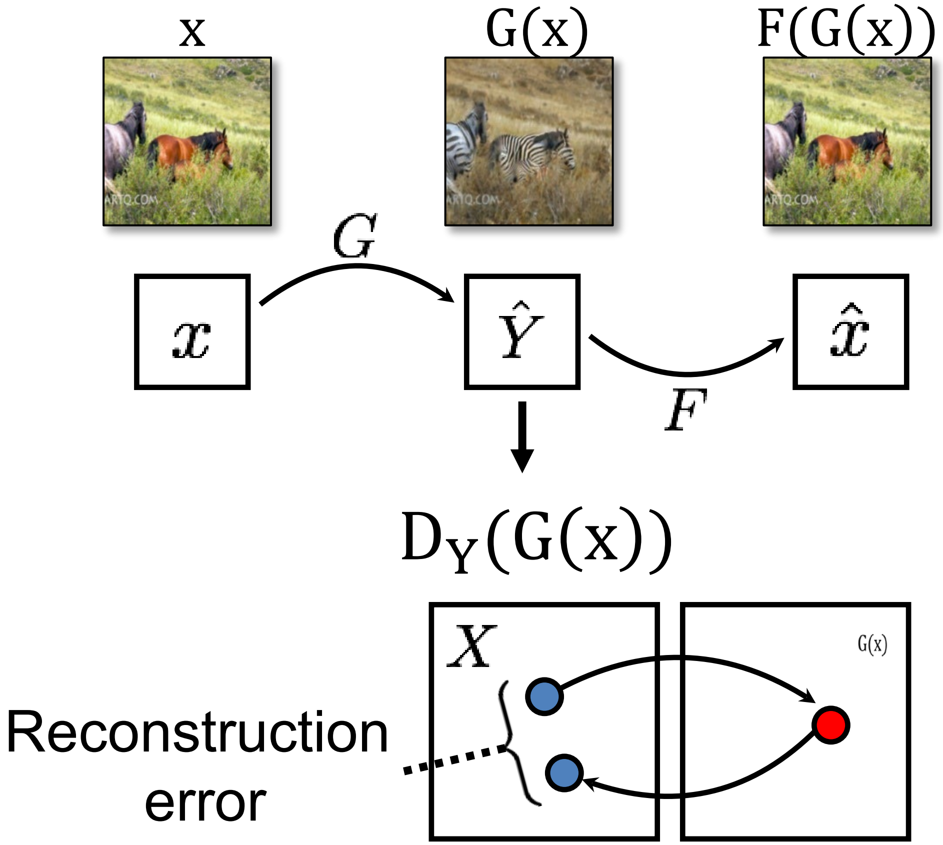
Cycle Consistency Loss



Slide credit:Jun-Yan Zhu

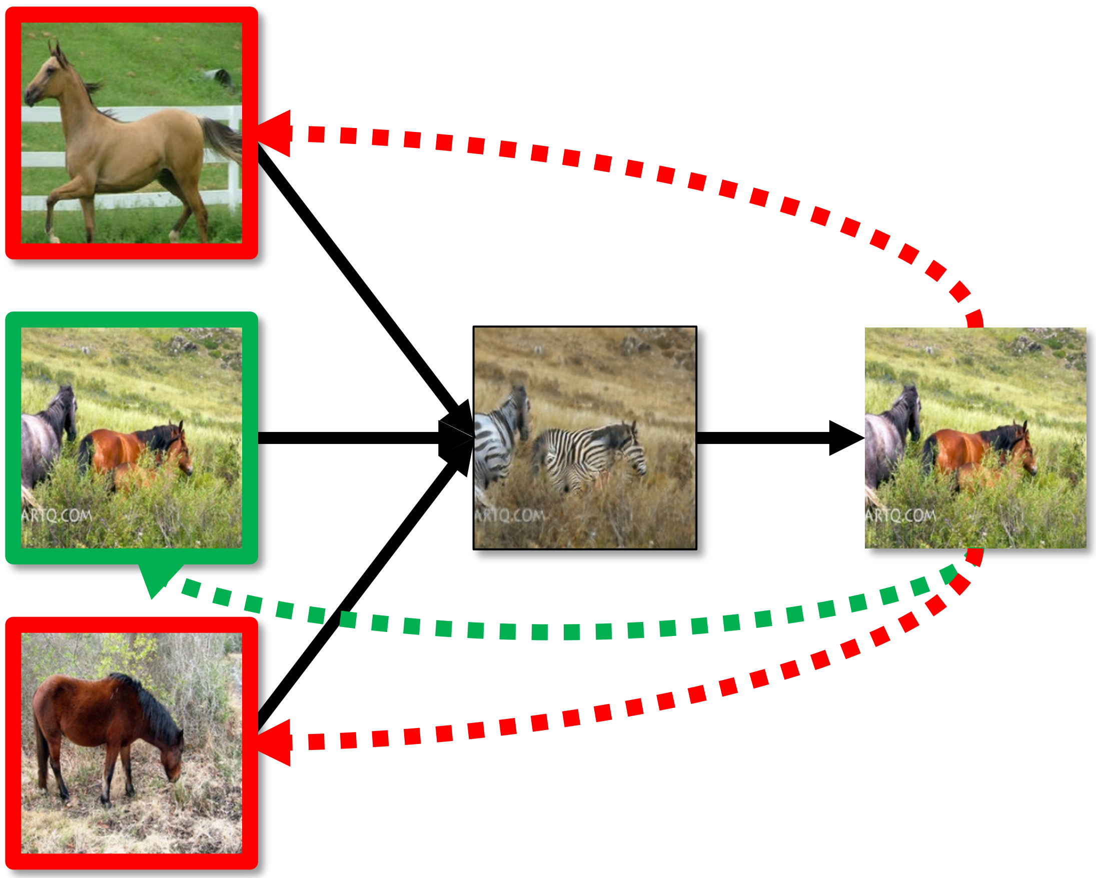
[Zhu*, Park*, Isola, and Efros, ICCV 2017]

Cycle Consistency Loss



$$\|F(G(x)) - x\|_1$$

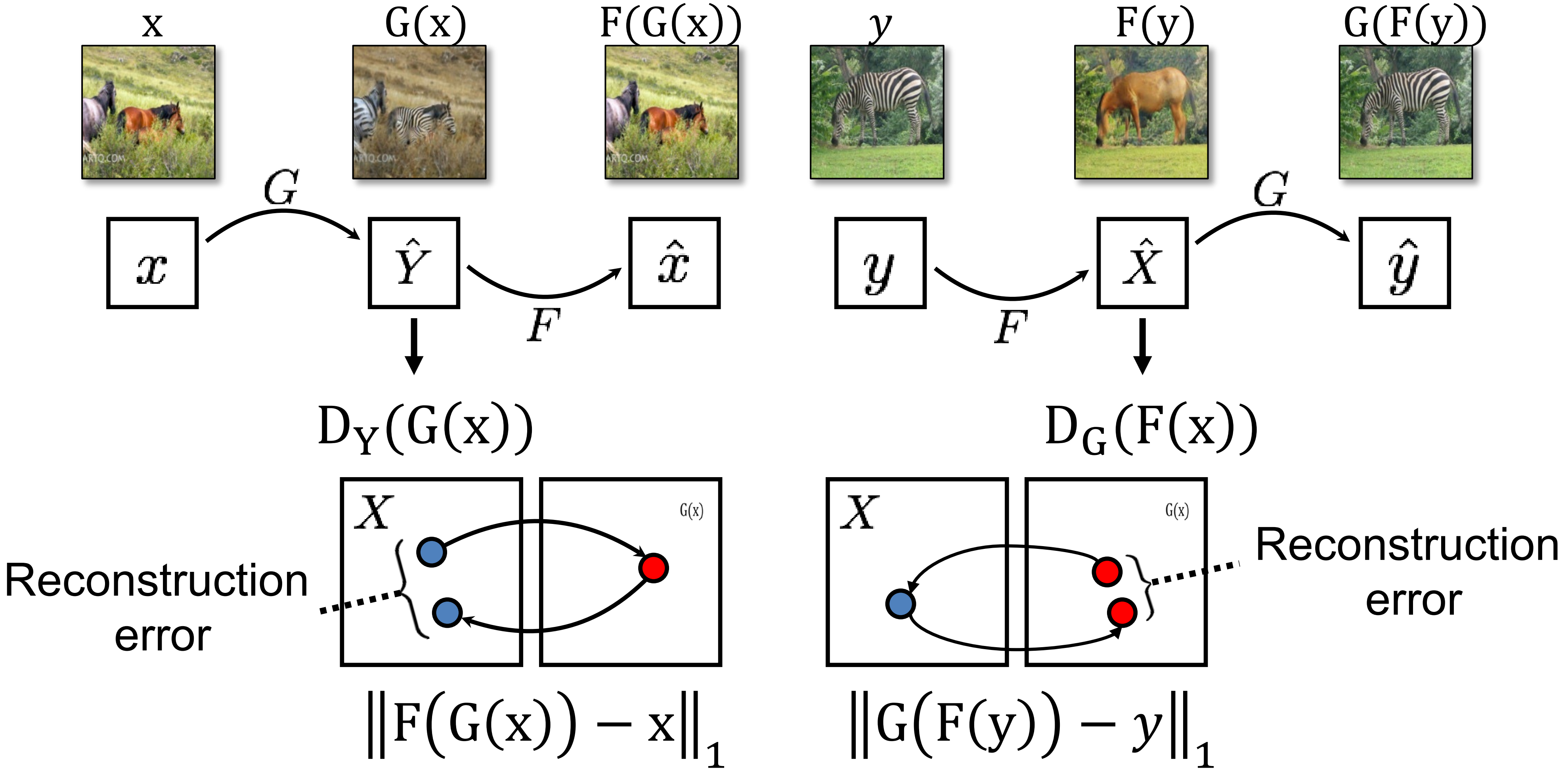
Small Large cycle loss



Slide credit:Jun-Yan Zhu

[Zhu*, Park*, Isola, and Efros, ICCV 2017]

Cycle Consistency Loss



Slide credit:Jun-Yan Zhu

[Zhu*, Park*, Isola, and Efros, ICCV 2017]

Cycle GAN: Full Objective

Produce images that look like each domain (according to discriminators) and complete a cycle

For L_{GAN} a squared loss is used instead of log loss

$$
\begin{aligned}
\mathcal{L}(G, F, D_X, D_Y) = & \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) \\
& + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) \\
& + \lambda \mathcal{L}_{\text{cyc}}(G, F),
\end{aligned}
$$

Collection Style Transfer

Photograph
@ Alexei Efros

Ukiyo-e   Cezanne

Van Gogh   Monet

Results slides credit:Jun-Yan Zhu


Input
Monet
Van Gogh
Cezanne
Ukiyo-e

Monet's paintings → photos

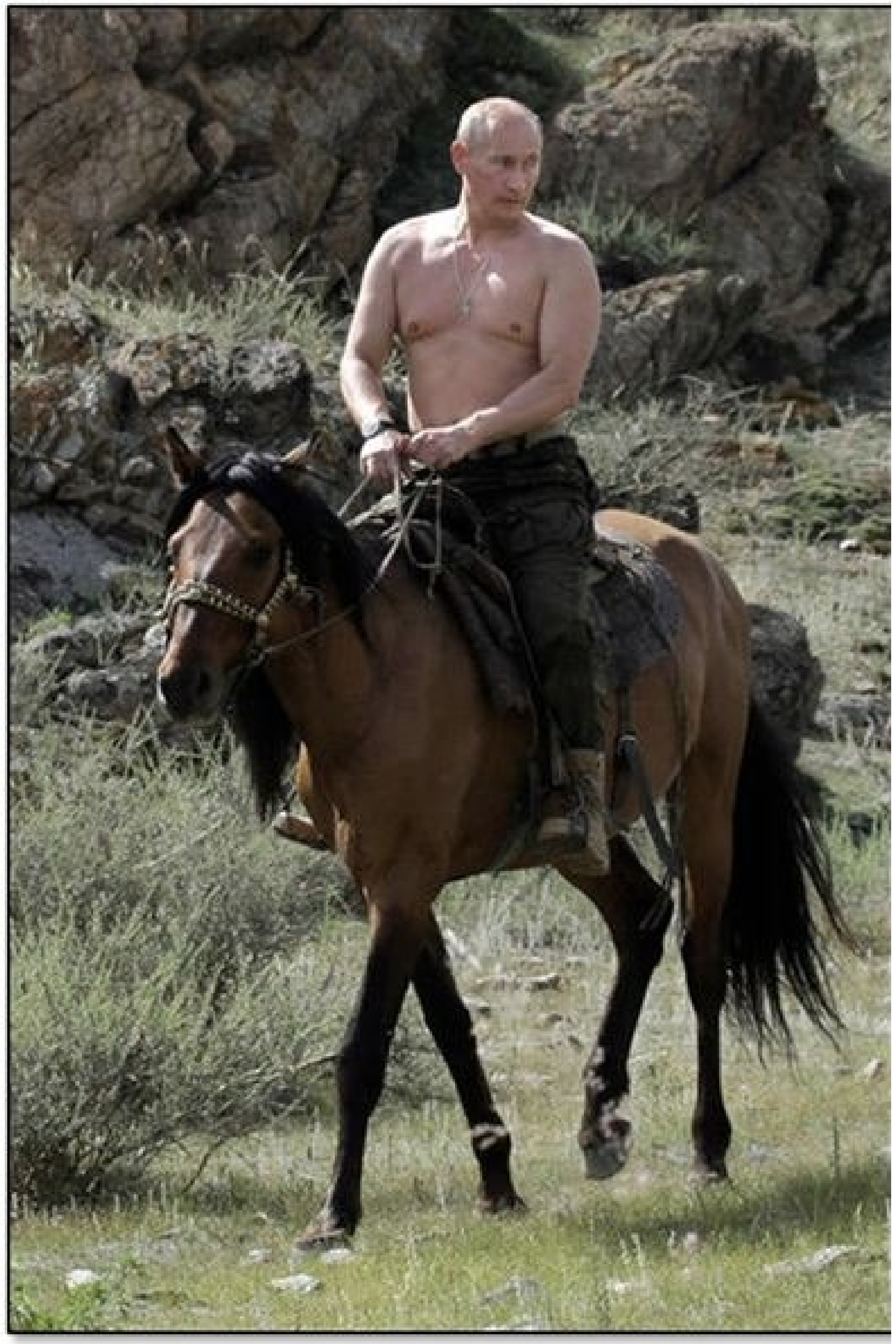
Monet’s paintings → photos





CycleGAN Horse -> Zebra

https://youtu.be/9reHvktowLY



Everybody Dance Now

ICCV 2019

Caroline Chan* Shiry Ginosar Tinghui Zhou† Alexei A. Efros

UC Berkeley

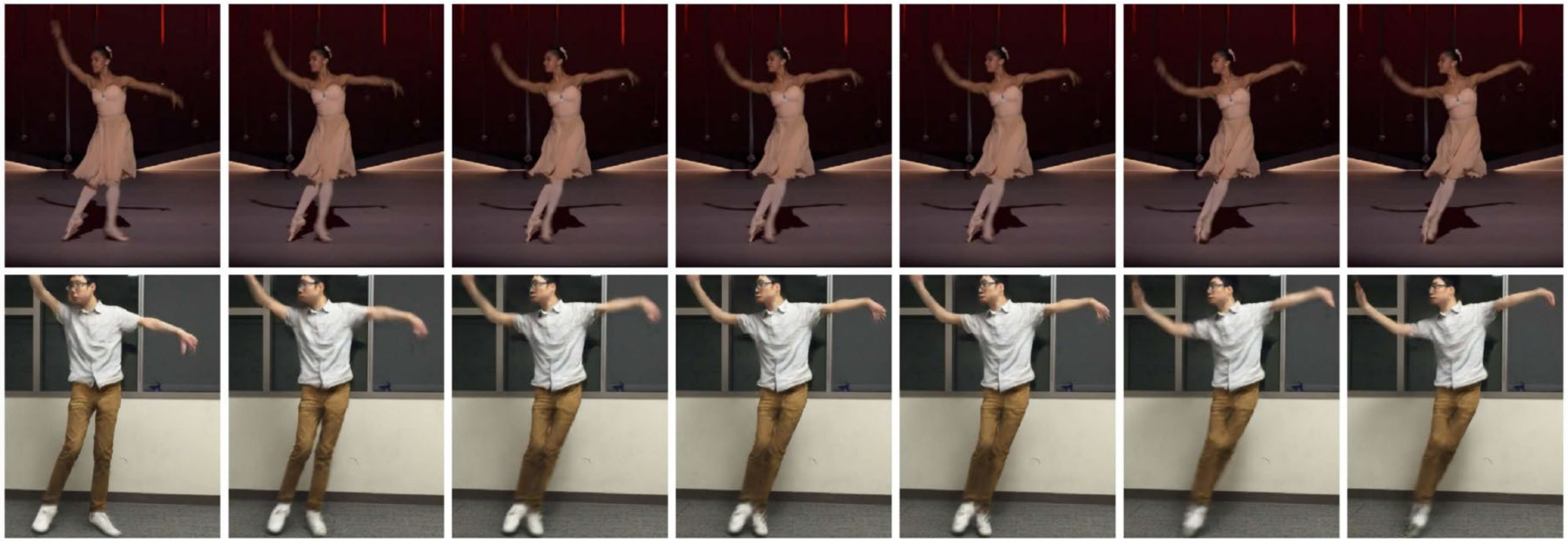
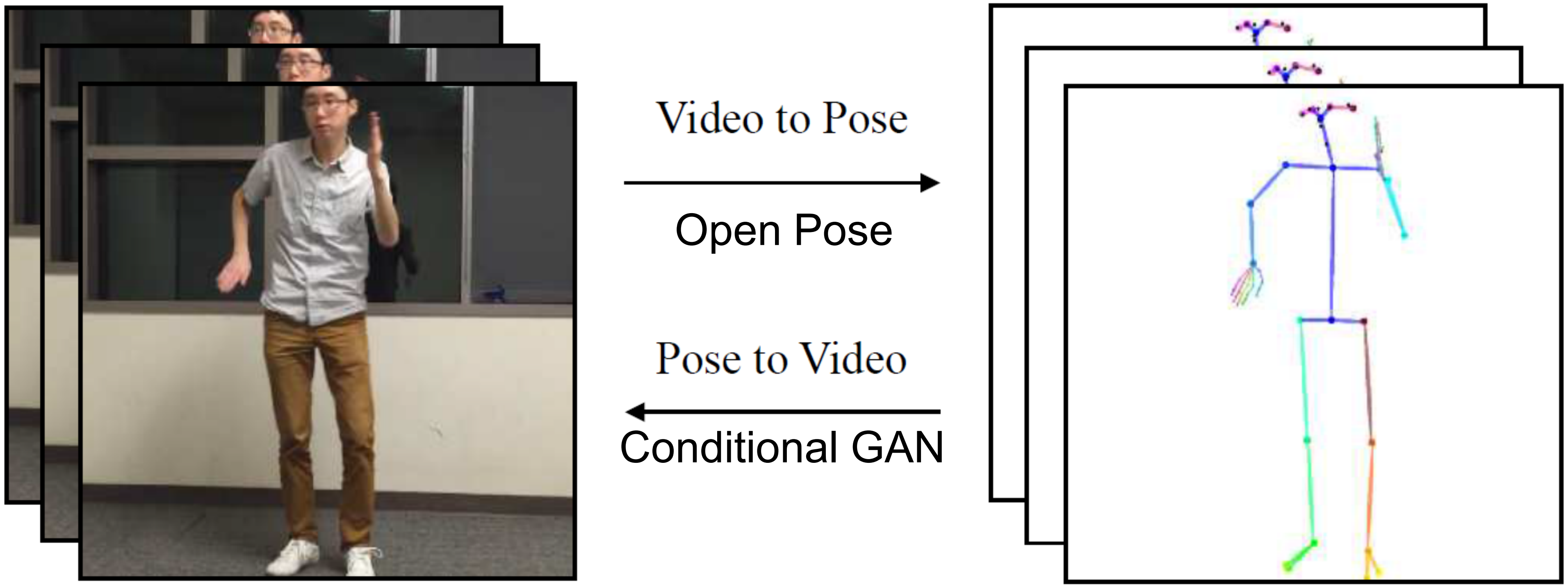
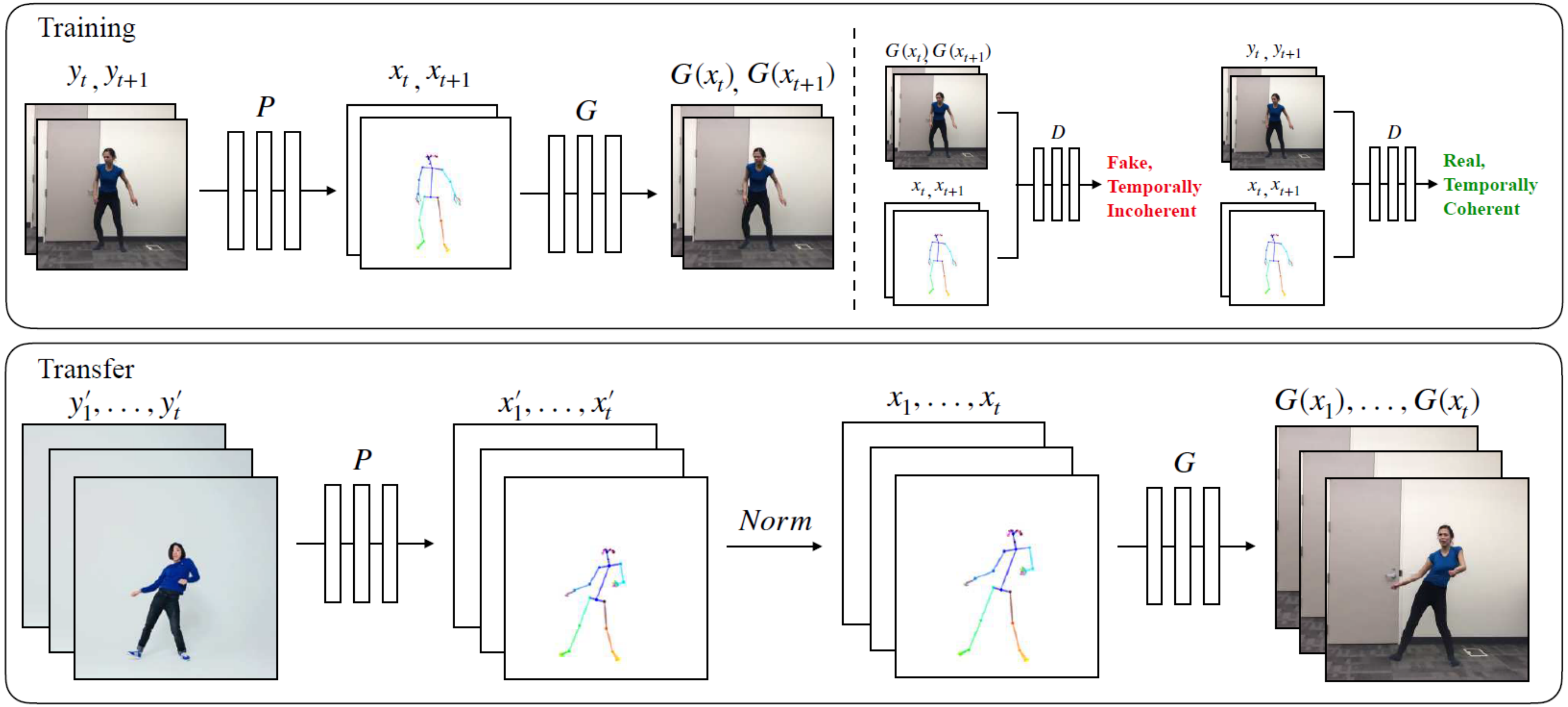


Figure 1: **"Do as I Do" motion transfer:** given a YouTube clip of a ballerina (top), and a video of a graduate student performing various motions, our method transfers the ballerina's performance onto the student (bottom). Video: `https://youtu.be/mSaIrz8lM1U`

Everybody Dance Now



Everybody Dance Now

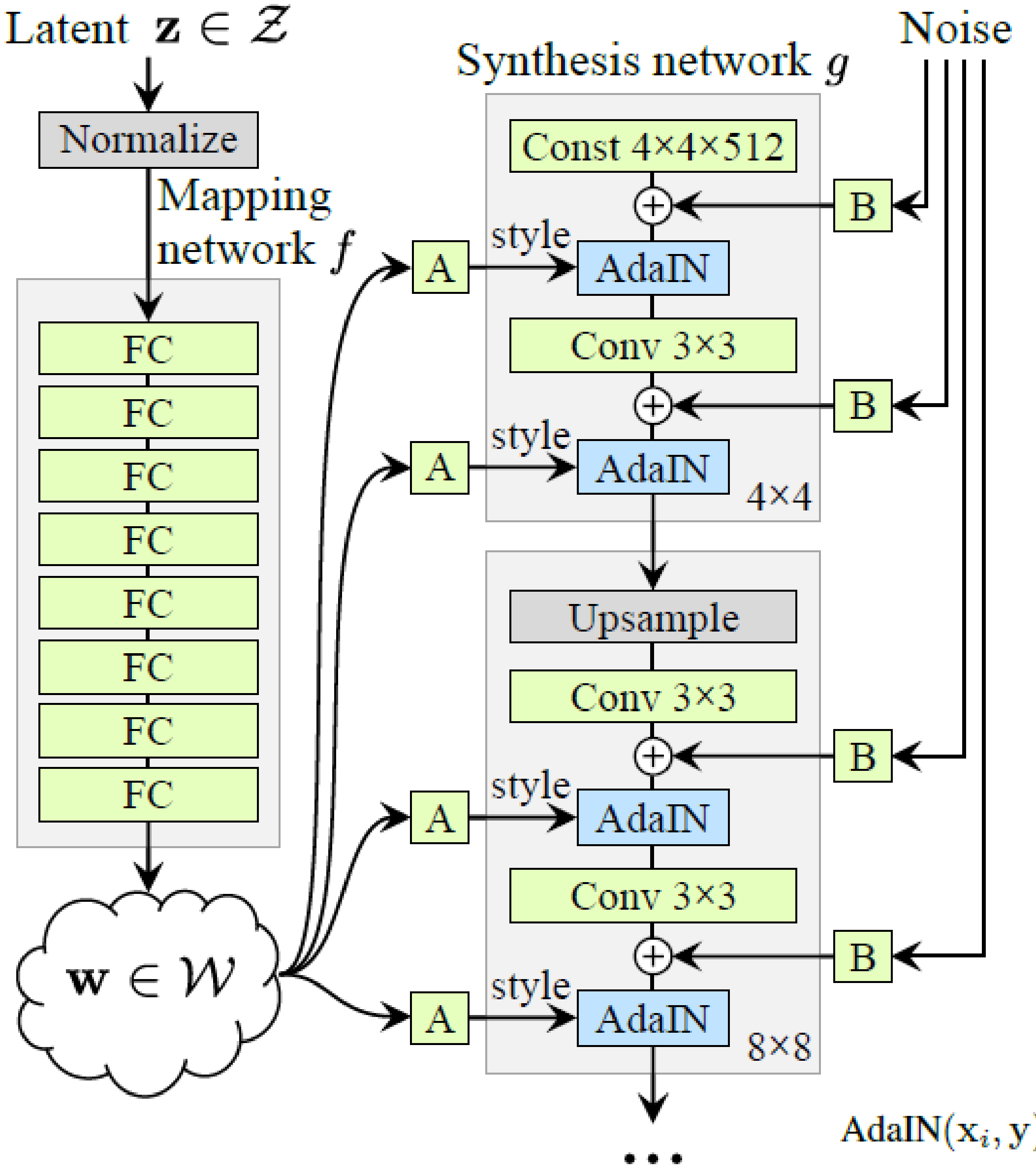


- Optimize a body GAN, face GAN, and temporal smoothness
- Discriminator conditions on pose and previous image and uses a perceptual distance for loss

Everybody Dance Now Video

https://www.youtube.com/watch?v=PCBTZh41Ris

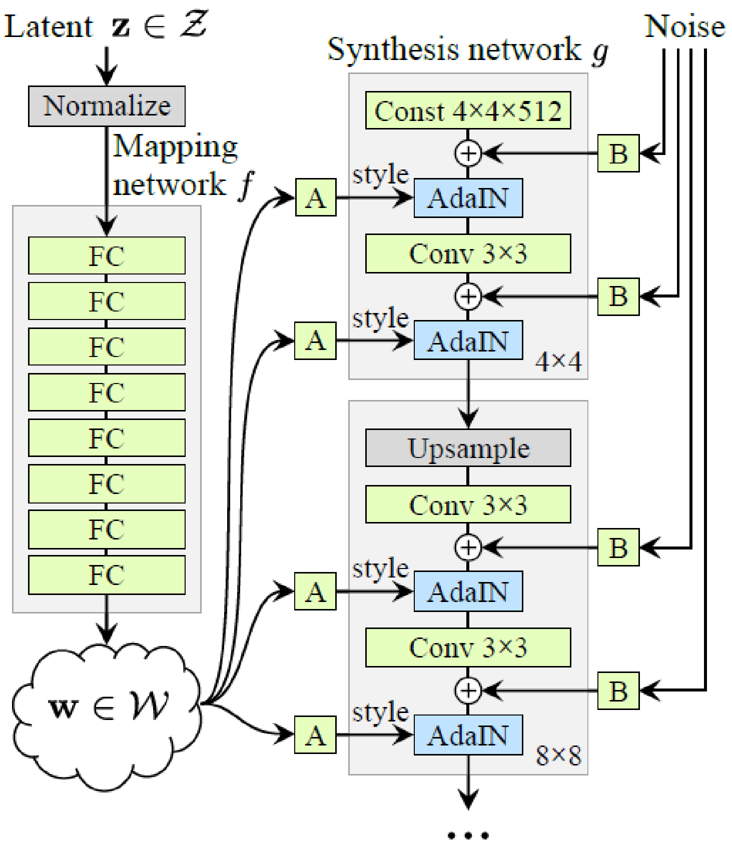
StyleGAN (Karras et al. CVPR 2019)



$$\mathrm{AdaIN}(\mathbf{x}_i, \mathbf{y}) = \mathbf{y}_{s,i} \frac{\mathbf{x}_i - \mu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)} + \mathbf{y}_{b,i}$$

Style Mixing

- Switch from one latent code to another at a random point in the synthesis network



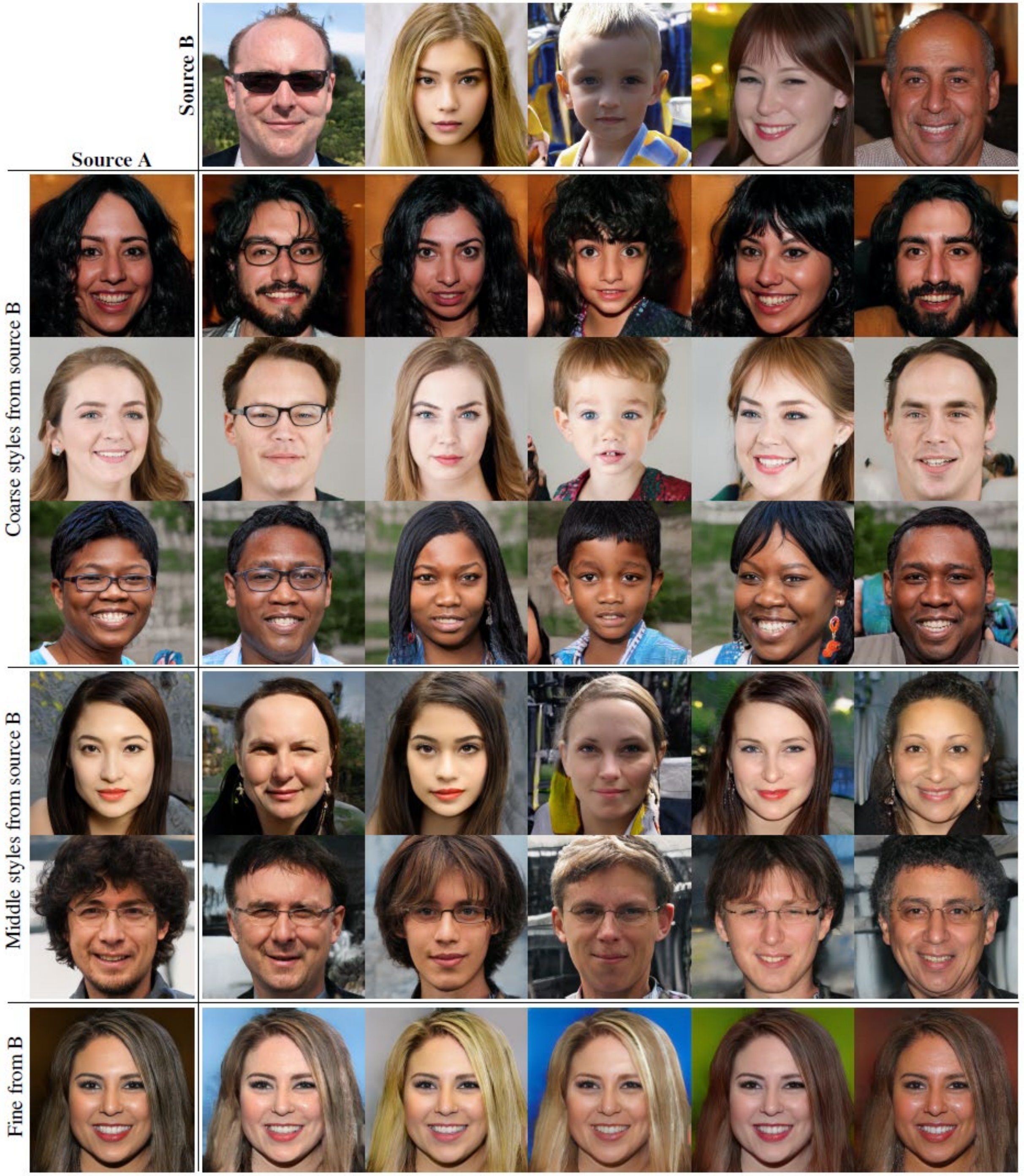


Figure 10. Uncurated set of images produced by our style-based generator (config F) with the LSUN BEDROOM dataset at 256^2. FID computed for 50K images was 2.65.

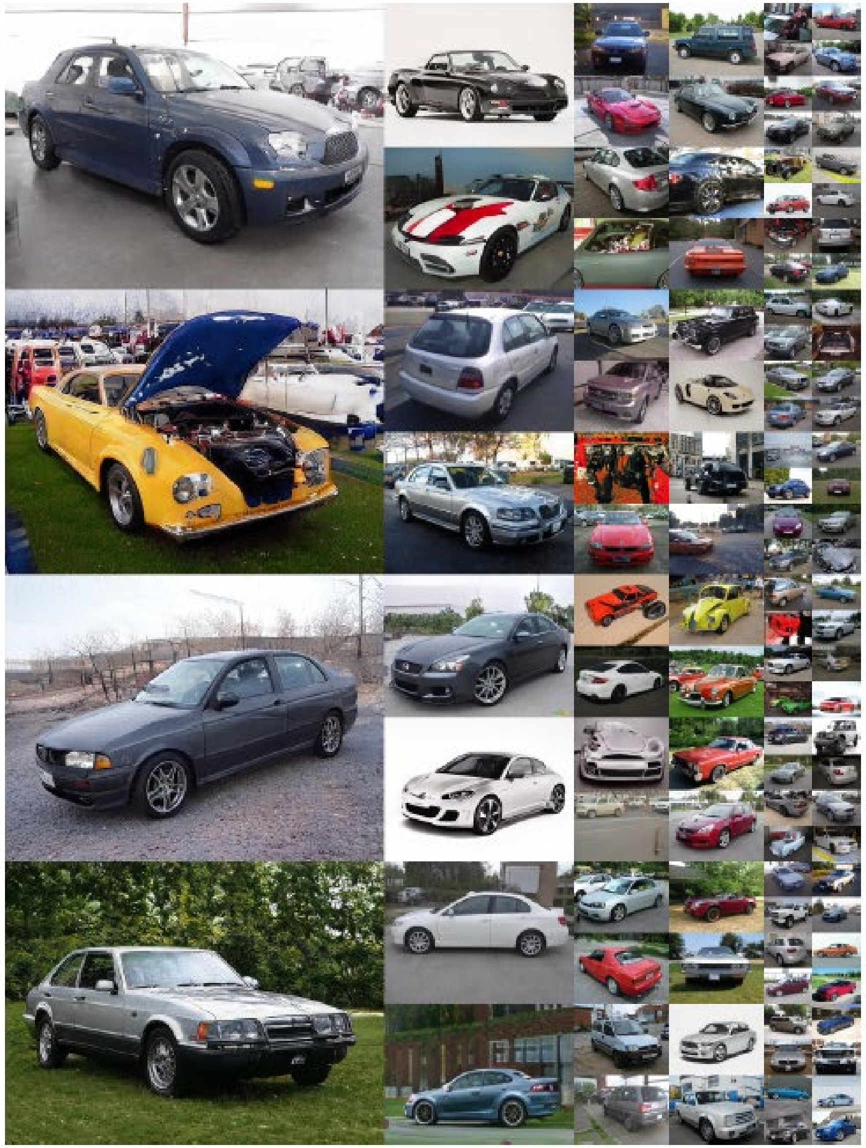


Figure 11. Uncurated set of images produced by our style-based generator (config F) with the LSUN CAR dataset at 512×384. FID computed for 50K images was 3.27.

Language Models and Diffusion Networks – Awesome short vids by Steve Seitz

- **Text to Image: Parti, Dall-E 2, Imagen**

https://www.youtube.com/watch?v=GYyP7Ova8KA&list=PLWfDJ5nla8UpwShx-lzLJqcp575fKpsSO&index=22

- **Text to Image: Part 2 -- Diffusion**

https://www.youtube.com/watch?v=lyodbLwb2lY&list=PLWfDJ5nla8UpwShx-lzLJqcp575fKpsSO&index=23

How to detect deep fakes?

- Google is creating DeepFake data for researchers: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html
- Mazaheri and Roy-Chowdhury (WACV 2022) report 99% accuracy in detection (when able to train on fake samples)
- Deep fake detection article: https://nerdist.com/article/deepfake-detector/

  https://youtu.be/RoGHVI-w9bE
- “Everybody dance now” provides a classifier to identify videos produced by their system
- Whose responsibility is it to detect fake images?

Summary

- Lots of fun and creative uses for generating images
- But digital forgeries are an increasingly major problem as it becomes easier to fake images
- A variety of automatic and semi-automatic methods are available for detection of well-done manual forgeries
  - Checking lighting consistency
  - Checking demosaicking consistency (for high quality images)
  - Checking JPEG compression level consistency (for low quality images)
- “Deep fakes” have recently become effective, and deep fake detection is a new challenge

Upcoming

- Next: How the Kinect Works
- After that: Neural Rendering Fields (NeRF)