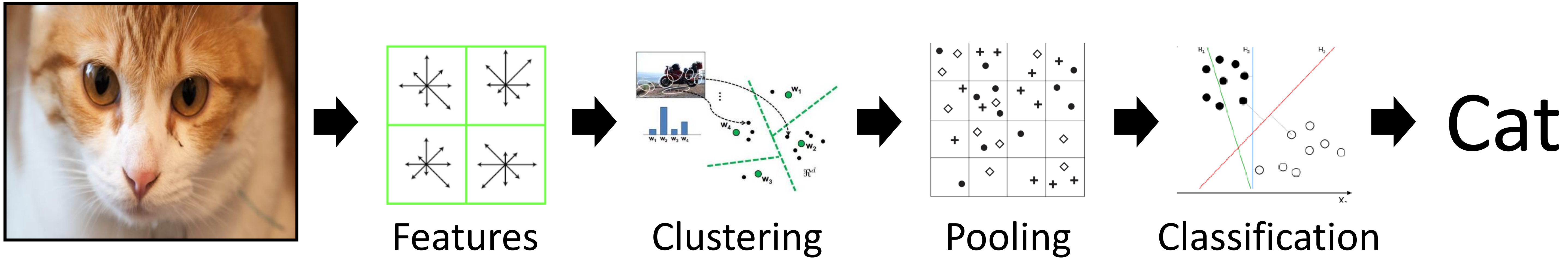
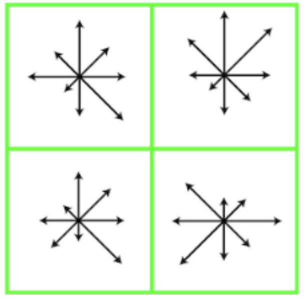
Learning to Generate Images

Jun-Yan Zhu
Ph.D. at UC Berkeley
Postdoc at MIT CSAIL

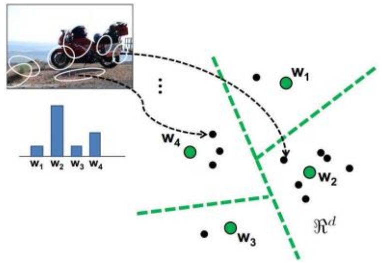
Computer Vision before 2012



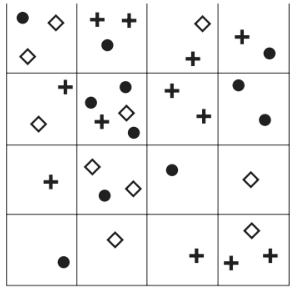
Computer Vision Now



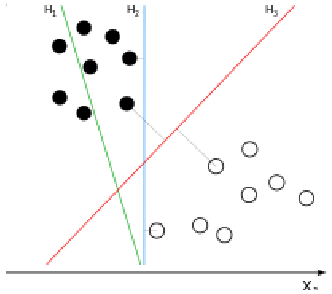
Features



Clustering

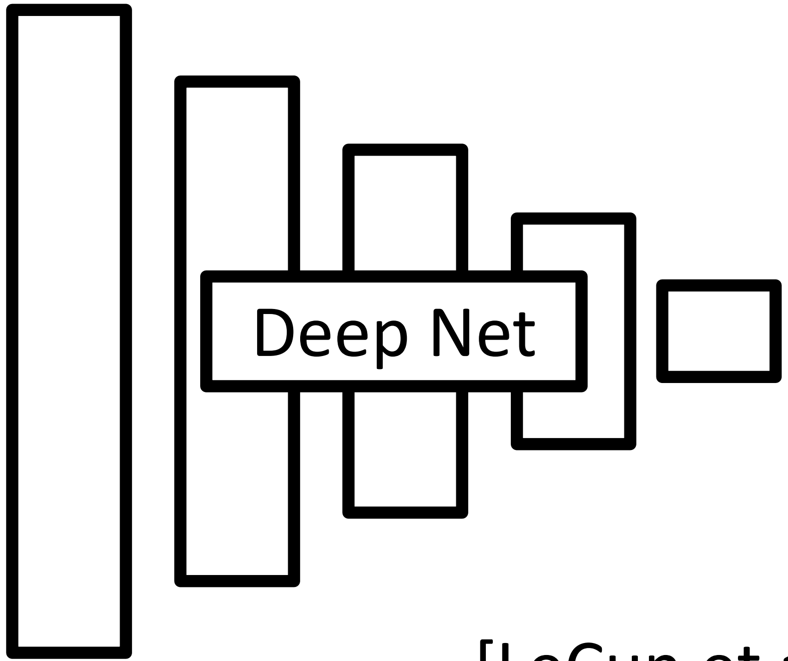


Pooling



Classification

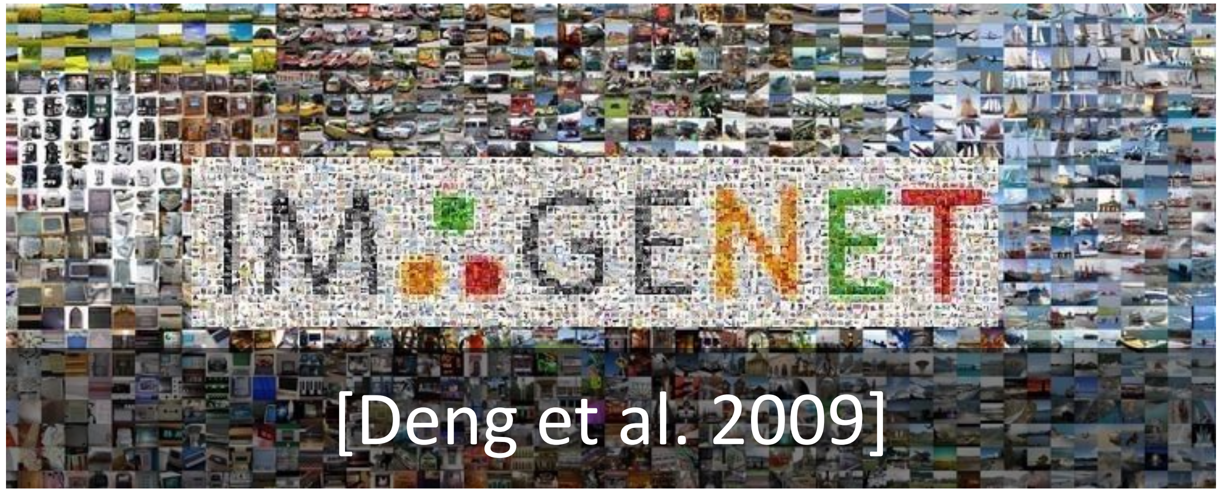
Cat

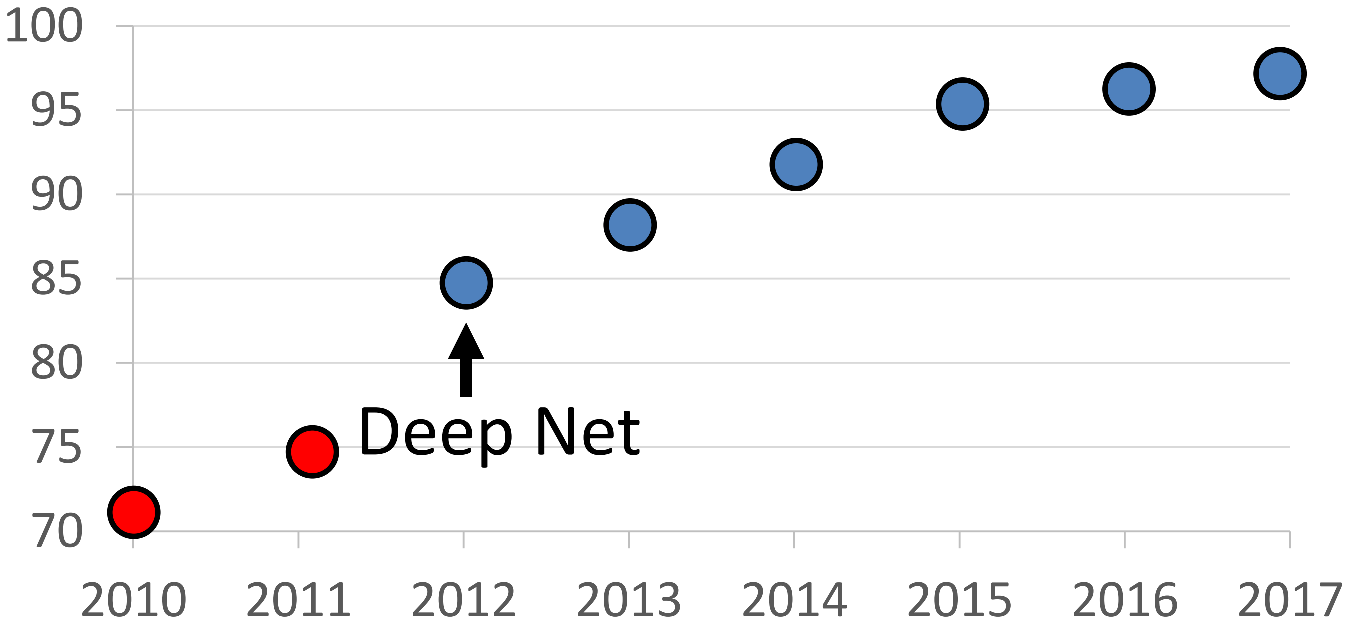


Cat

[LeCun et al, 1998], [Krizhevsky et al, 2012]

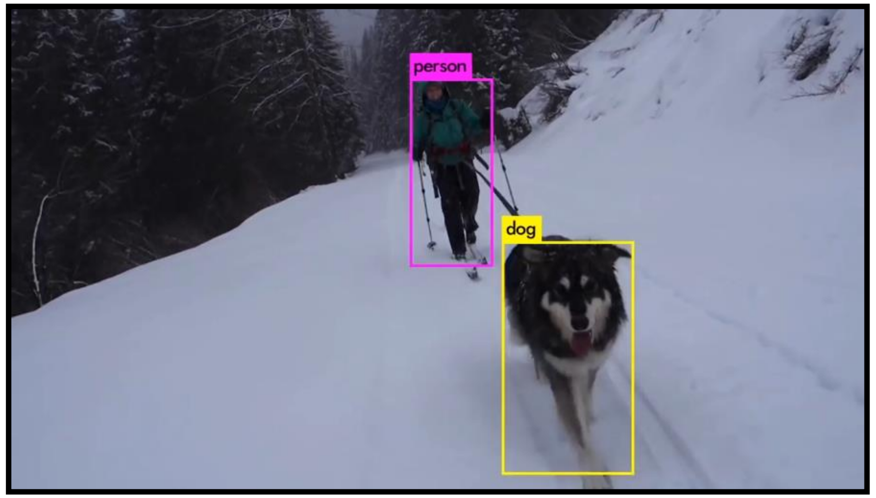
Deep Learning for Computer Vision





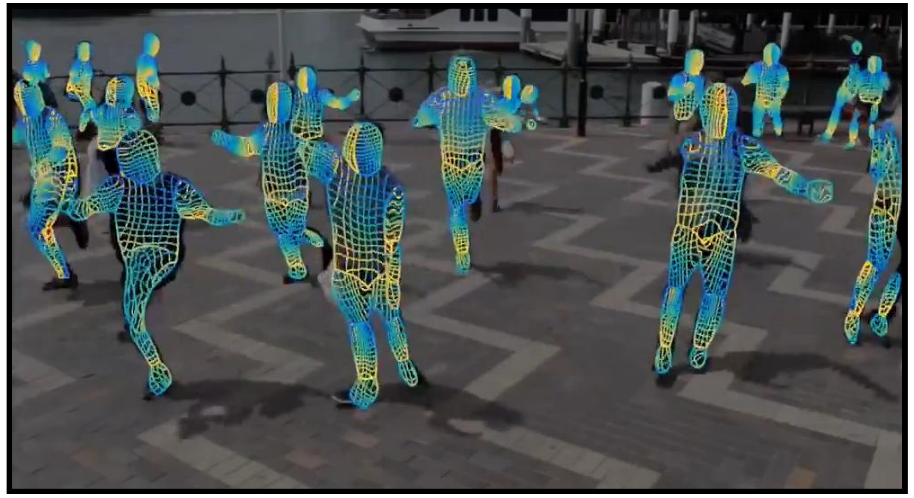
Top 5 **accuracy** on ImageNet benchmark

Object detection



[Redmon et al., 2018]

Human understanding



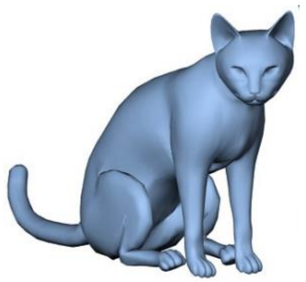
[Güler et al., 2018]

Autonomous driving

[Zhao et al., 2017]

Can Deep Learning **Help** Graphics?

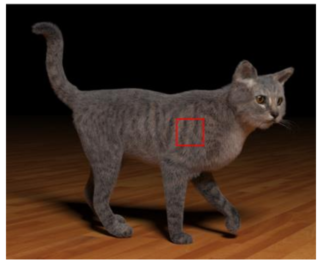
Cat



Modeling

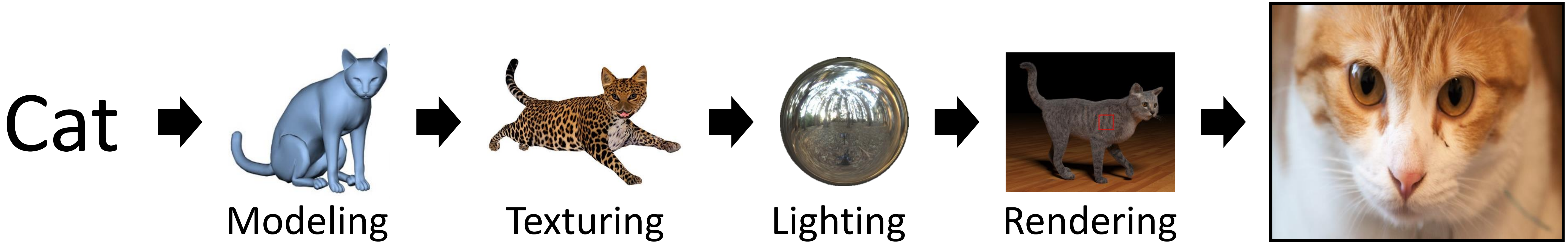
Texturing

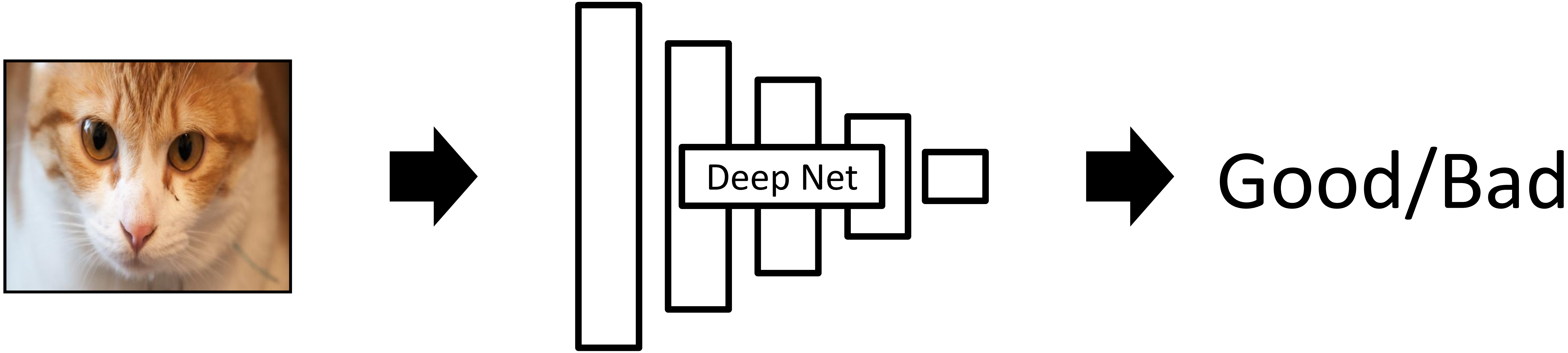
Lighting



Rendering

Can Deep Learning **Help** Graphics?



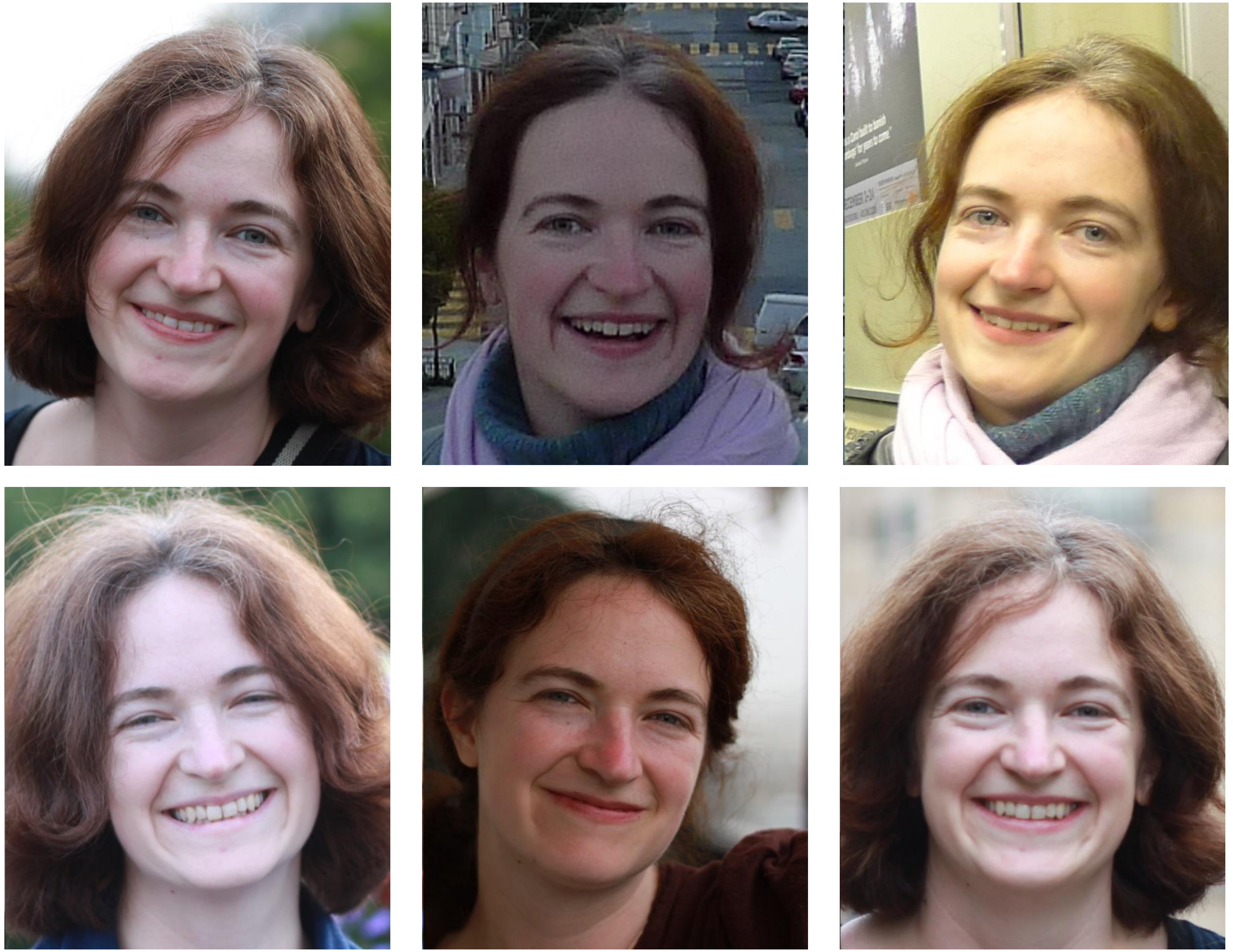


Selecting the most attractive expressions

Photos

⋮

101 Photos



[Zhu et al. SIGGRAPH Asia 2014]

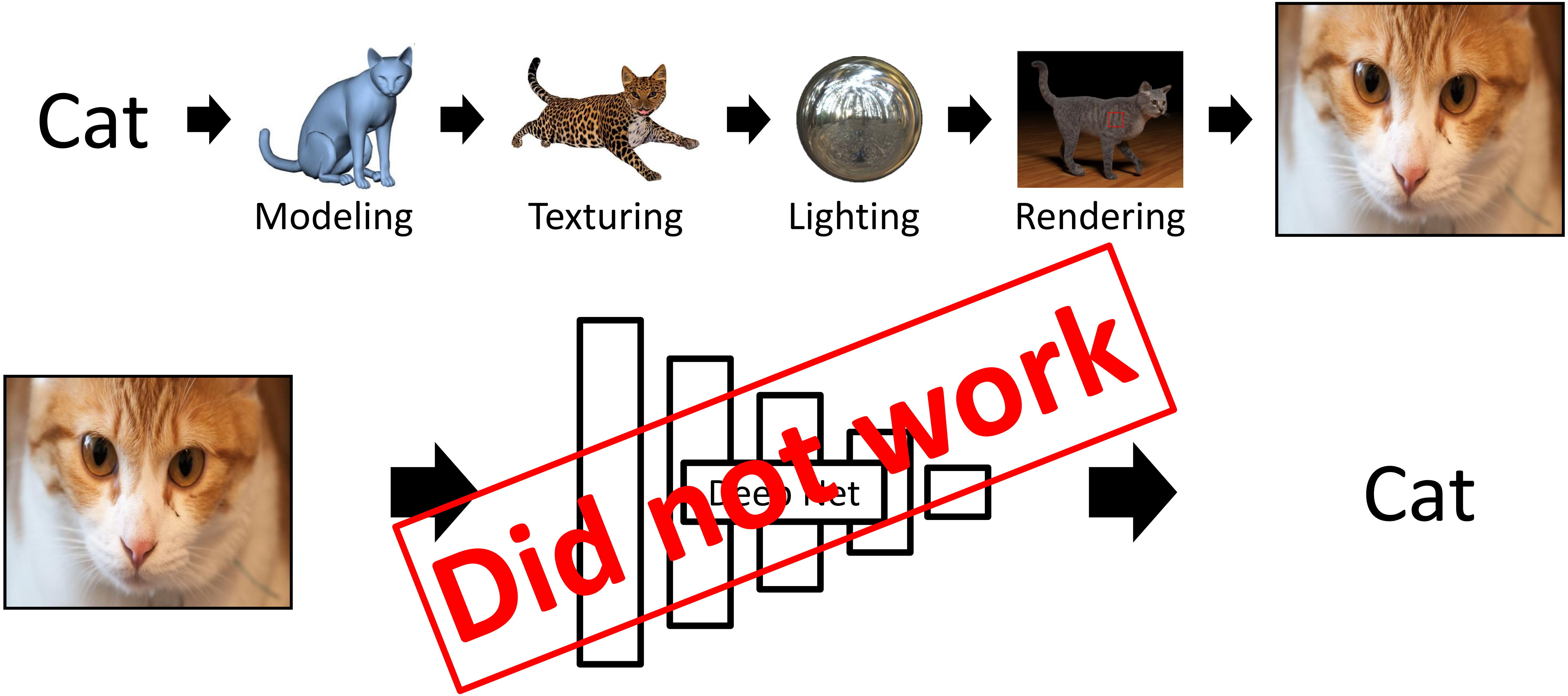
Selecting the most realistic composites

Most realistic composites

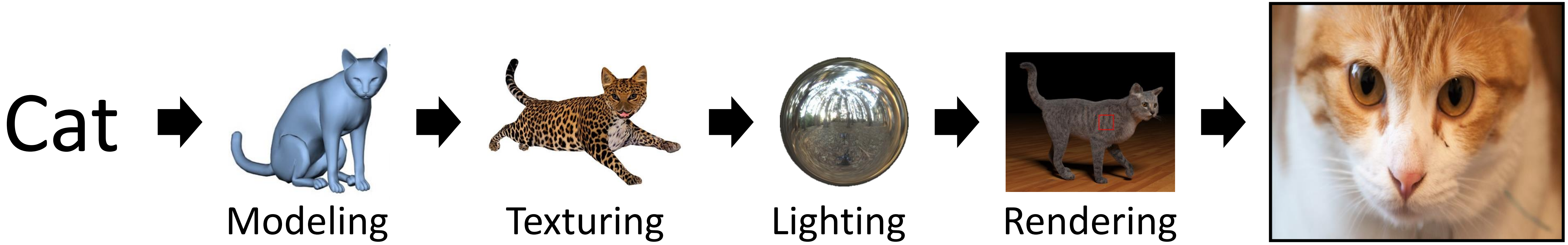
Least realistic composites

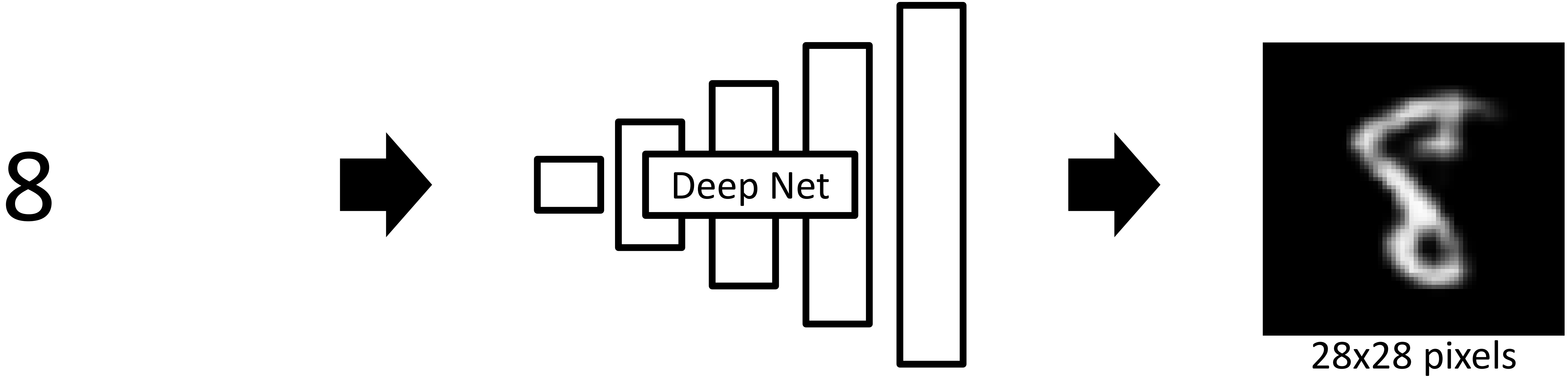
[Zhu et al. ICCV 2015]

Can Deep Learning **Help** Graphics?



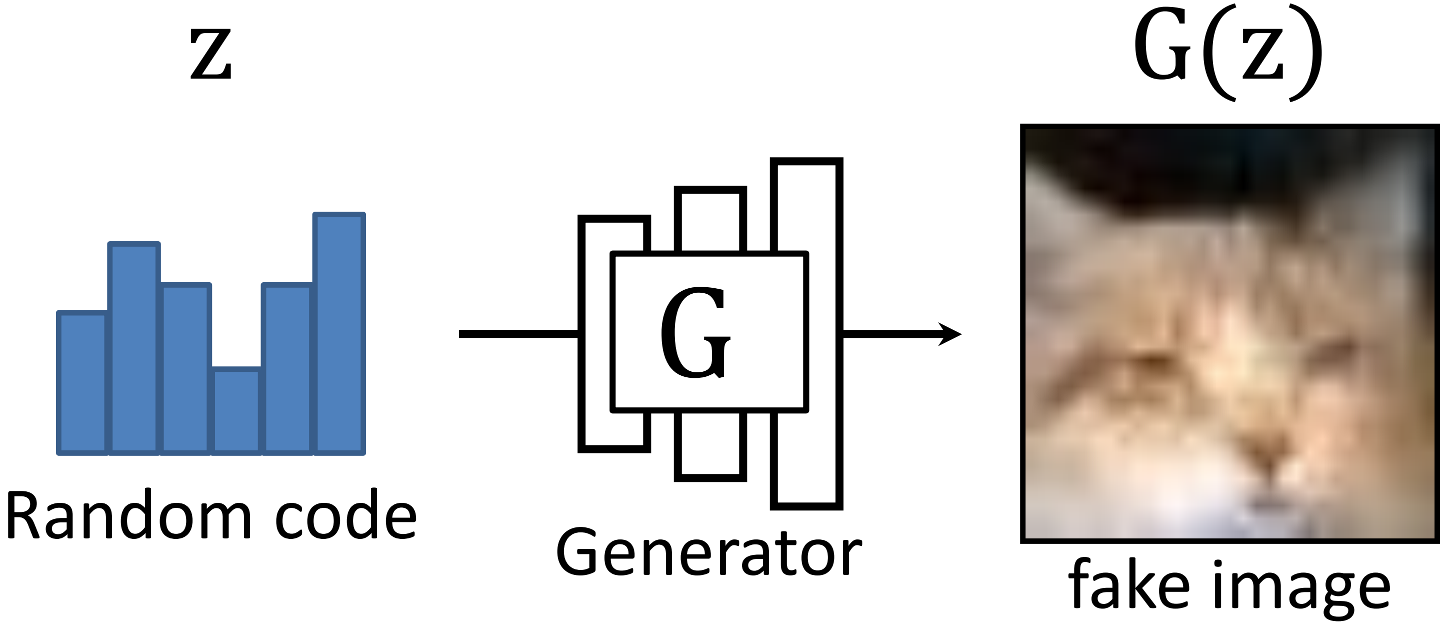
Generating images is hard!



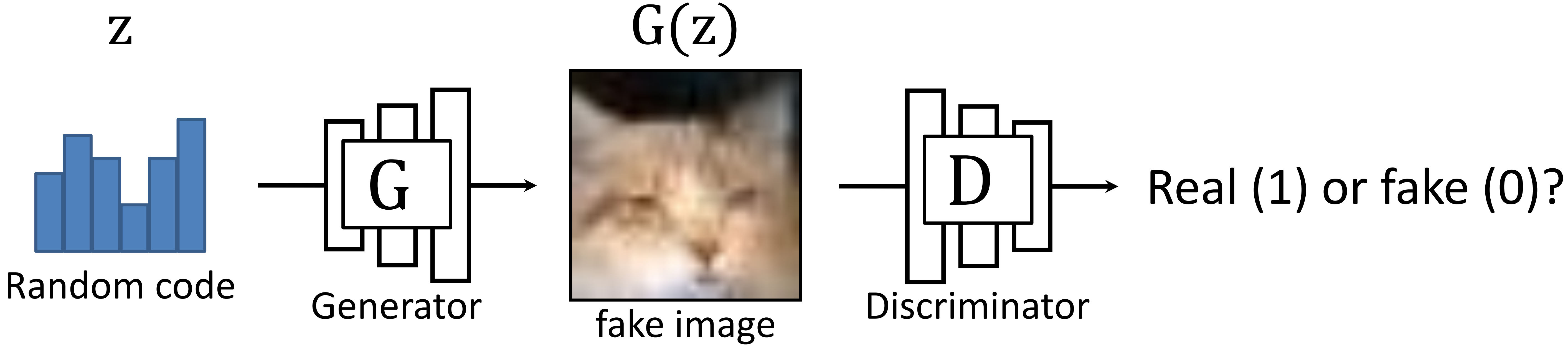


Generative Adversarial Networks (**GANs**)

[Goodfellow et al. 2014]



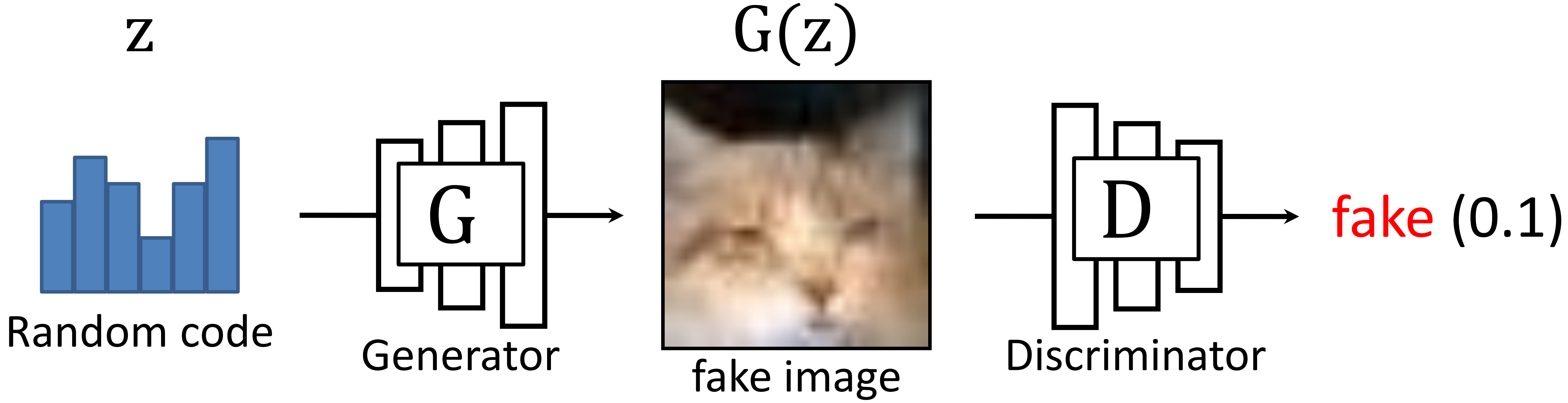
[Goodfellow et al. 2014]



A two-player game:

- G tries to generate fake images that can fool D.
- D tries to detect fake images.

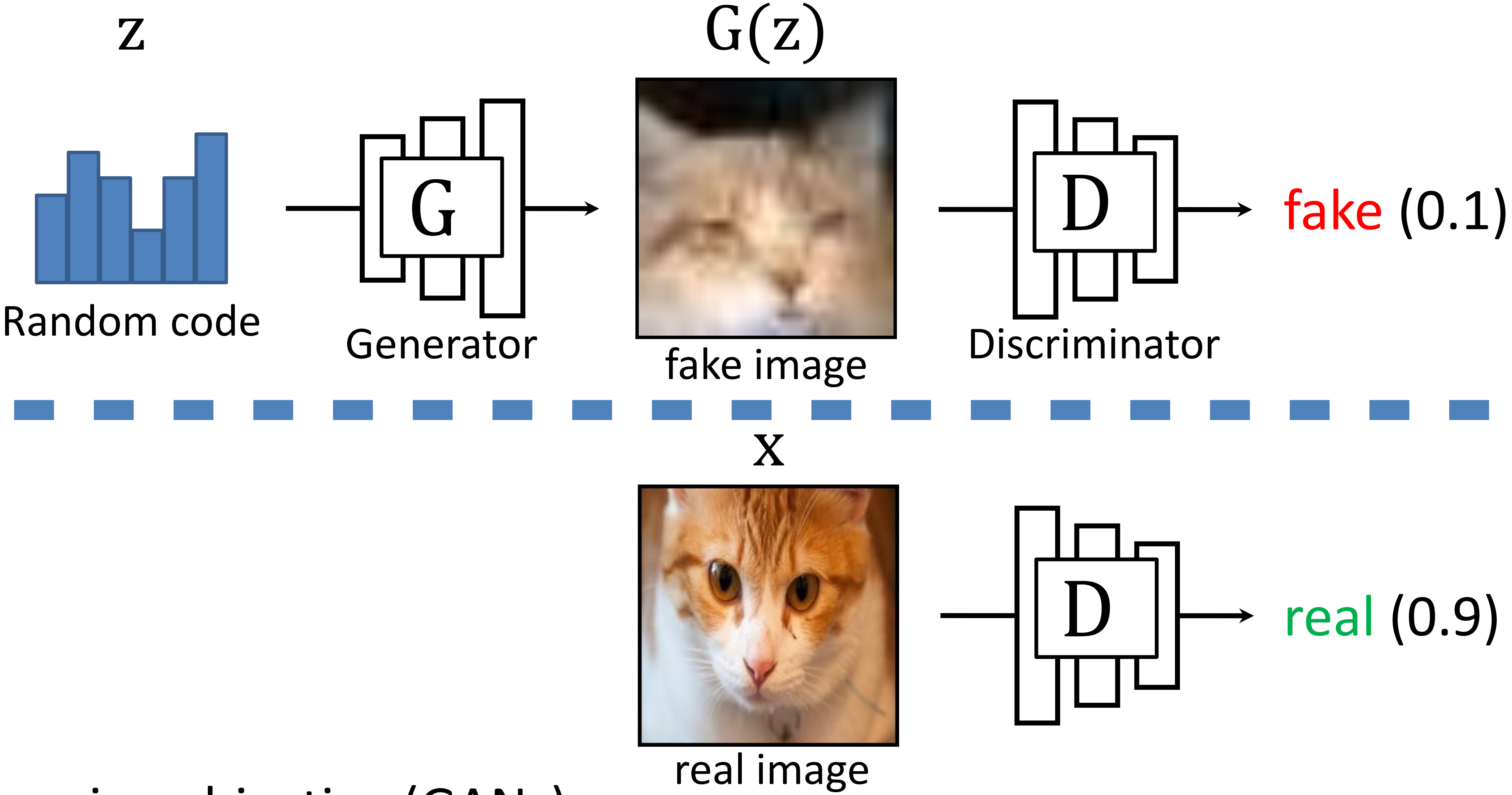
[Goodfellow et al. 2014]



Learning objective (GANs)

$$\min_{G} \max_{D} \mathbb{E} \left[\boxed{\log D(G(z))} \right.$$

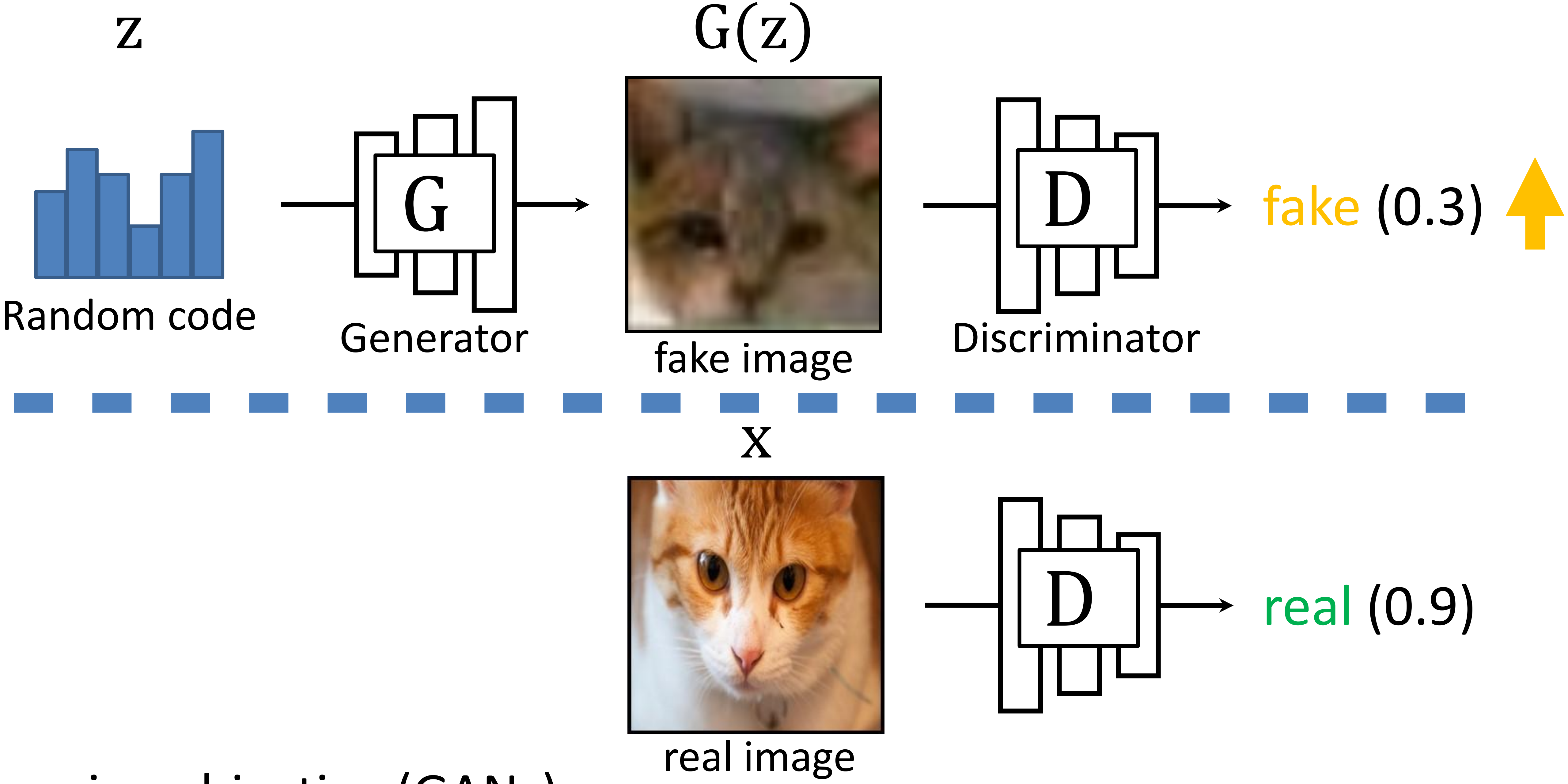
[Goodfellow et al. 2014]



Learning objective (GANs)

$$\min_{G} \max_{D} \mathbb{E} \left[\boxed{\log D(G(z))} + \boxed{\log\left(1 - D(x)\right)} \right]$$

[Goodfellow et al. 2014]



Learning objective (GANs)

$$\min_{G} \max_{D} \mathbb{E} \left[\boxed{\log D(G(z))} + \log\left(1 - D(x)\right) \right]$$

[Goodfellow et al. 2014]

Limitations of GANs

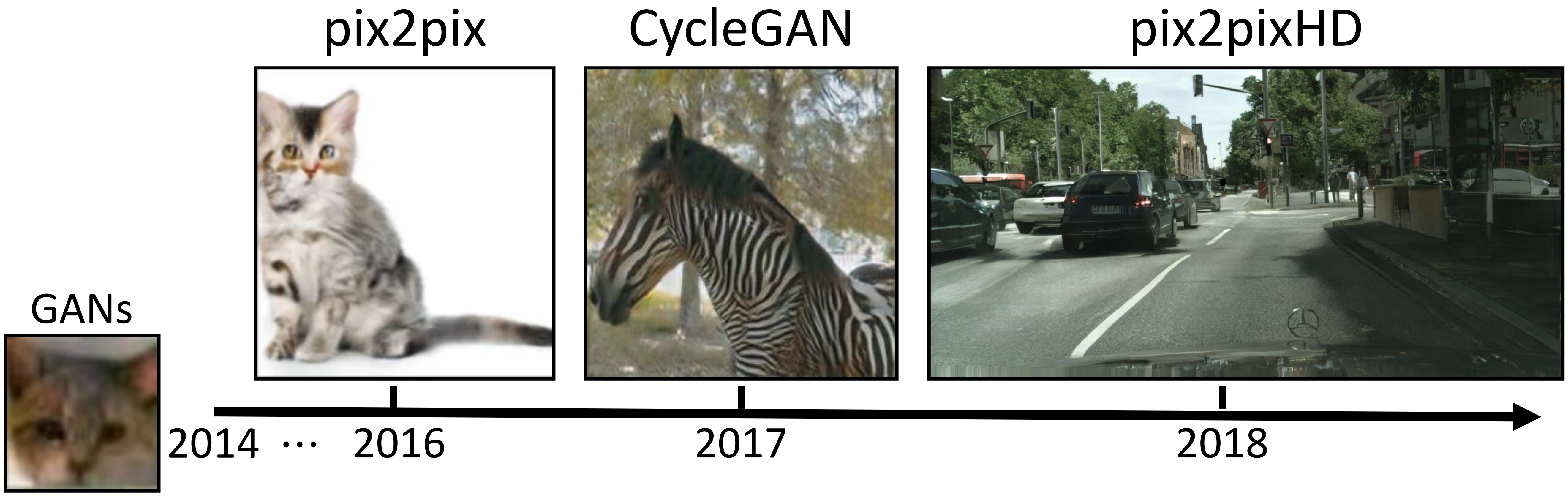
- No user control.

- Low resolution and quality.

Contributions

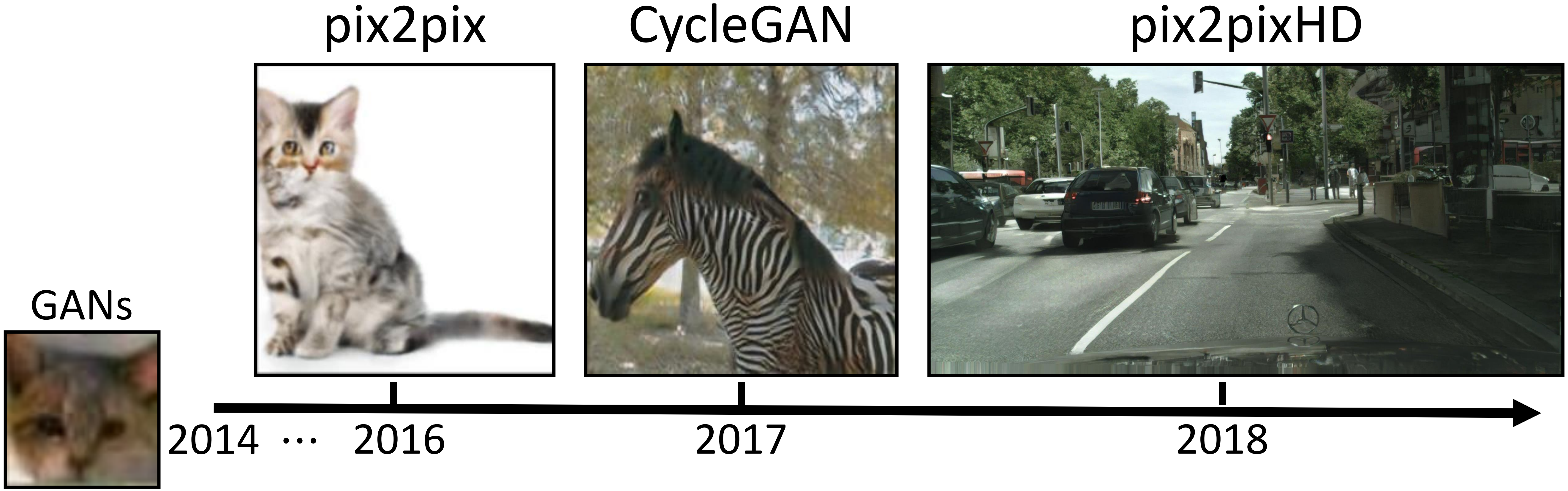
Co-authors:
Phillip Isola, Taesung Park, Ting-Chun Wang
Richard Zhang, Tinghui Zhou, Ming-Yu Liu, Andrew Tao
Jan Kautz, Bryan Catanzaro, Alexei A. Efros

Goals: Improve Control, Quality, and Resolution

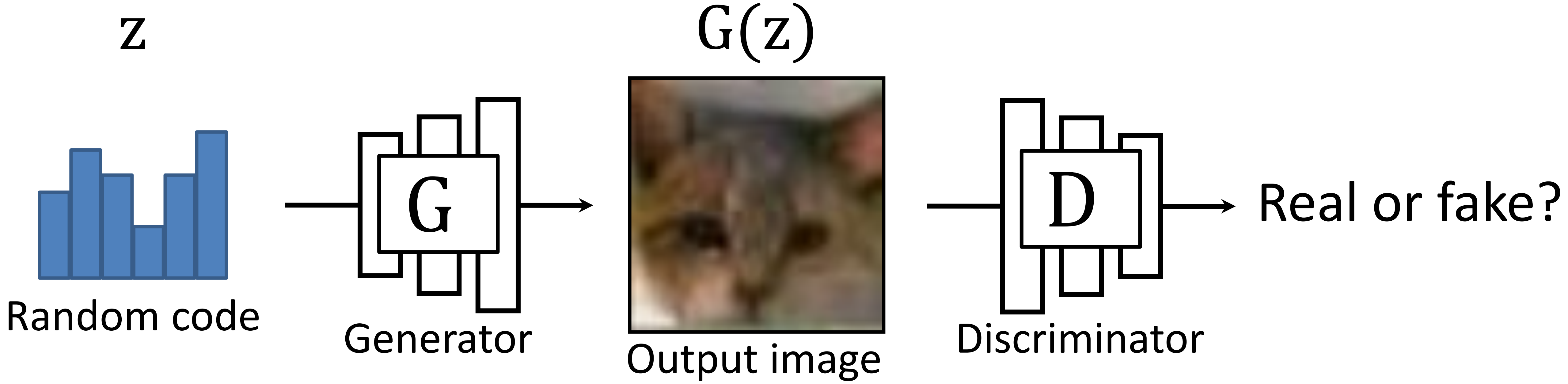


- Conditional on user inputs.
- Learning without pairs.
- High quality and resolution.

Goals: Improve Control, Quality, and Resolution



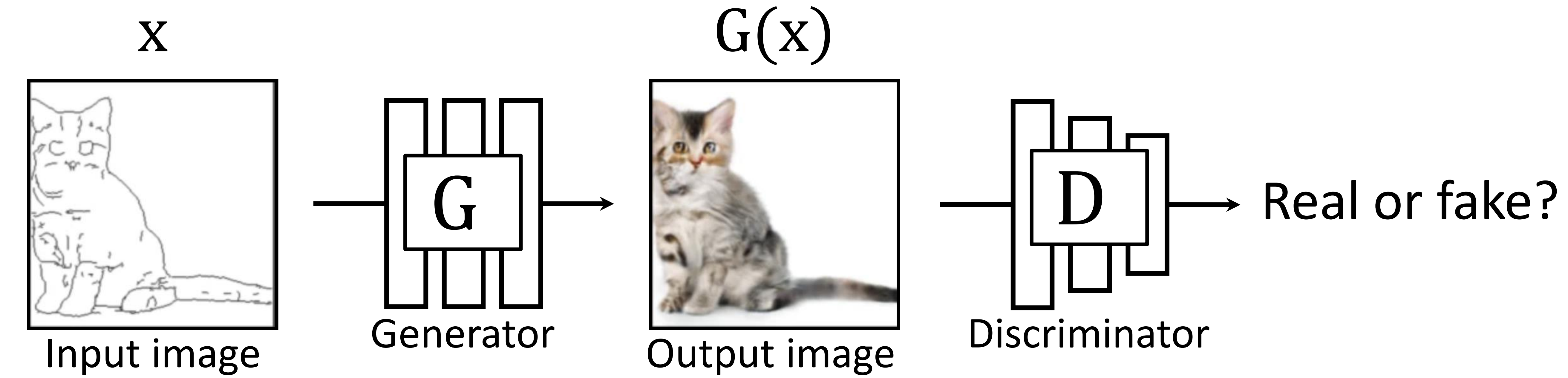
- Conditional on user inputs.
- Learning without pairs.
- High quality and resolution.



Learning objective (GANs)

$$\min_{G} \max_{D} \mathbb{E} \left[\log D(G(z)) + \log\left(1 - D(x)\right)\right]$$

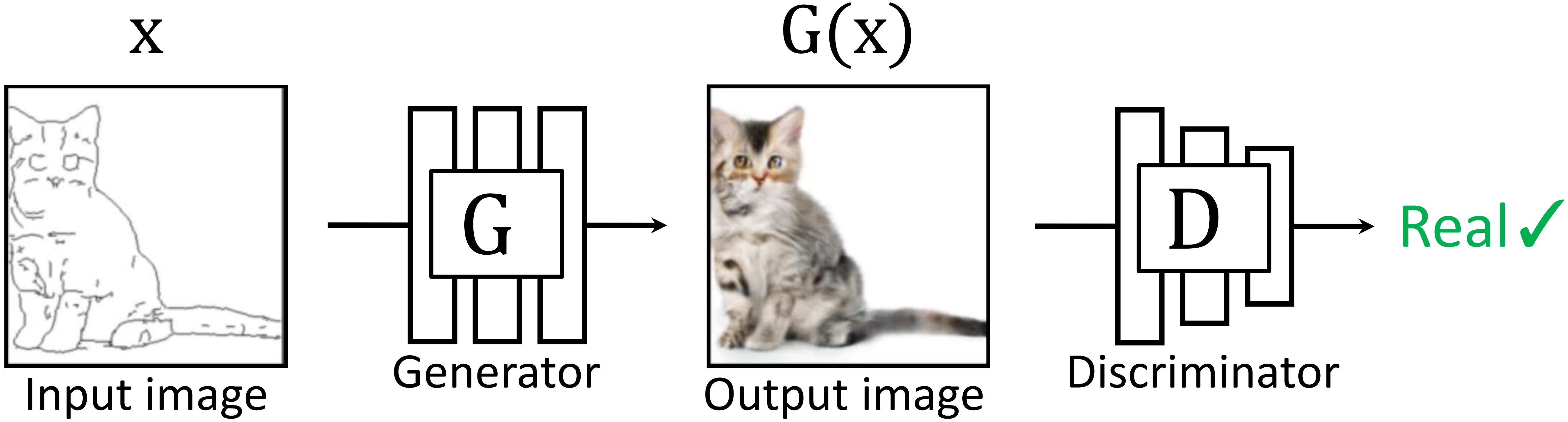
[Goodfellow et al. 2014]



Learning objective (pix2pix)

$$\min_{G} \max_{D} \mathbb{E} \left[\log D(G(x)) + \log\left(1 - D(y)\right)\right]$$

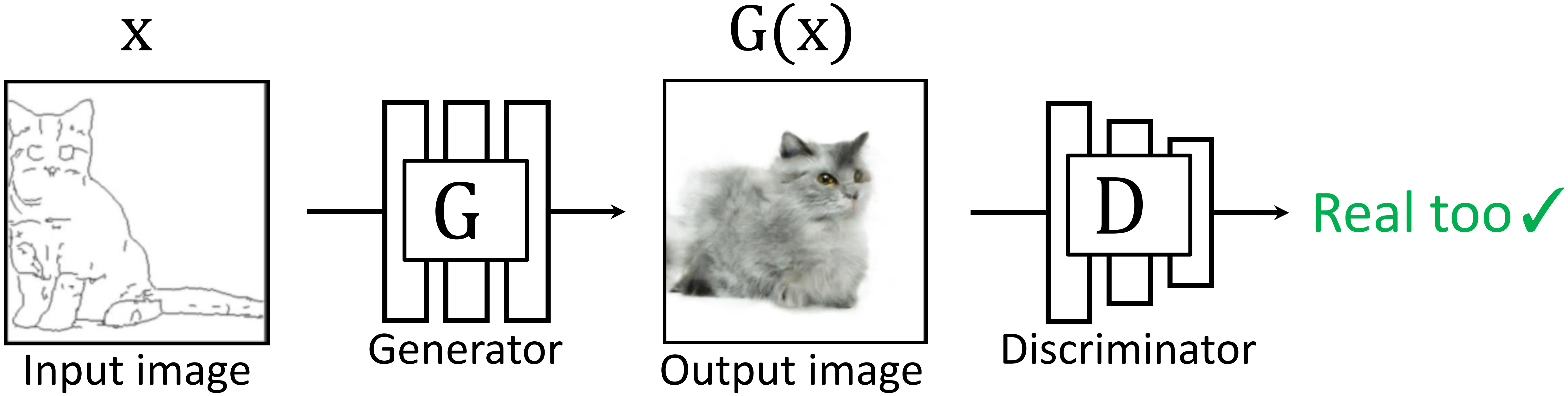
[Isola, Zhu, Zhou, Efros, 2016]



Learning objective (pix2pix)

$$\min_G \max_D \mathbb{E}\left[\log D(G(x)) + \log\left(1 - D(y)\right)\right]$$

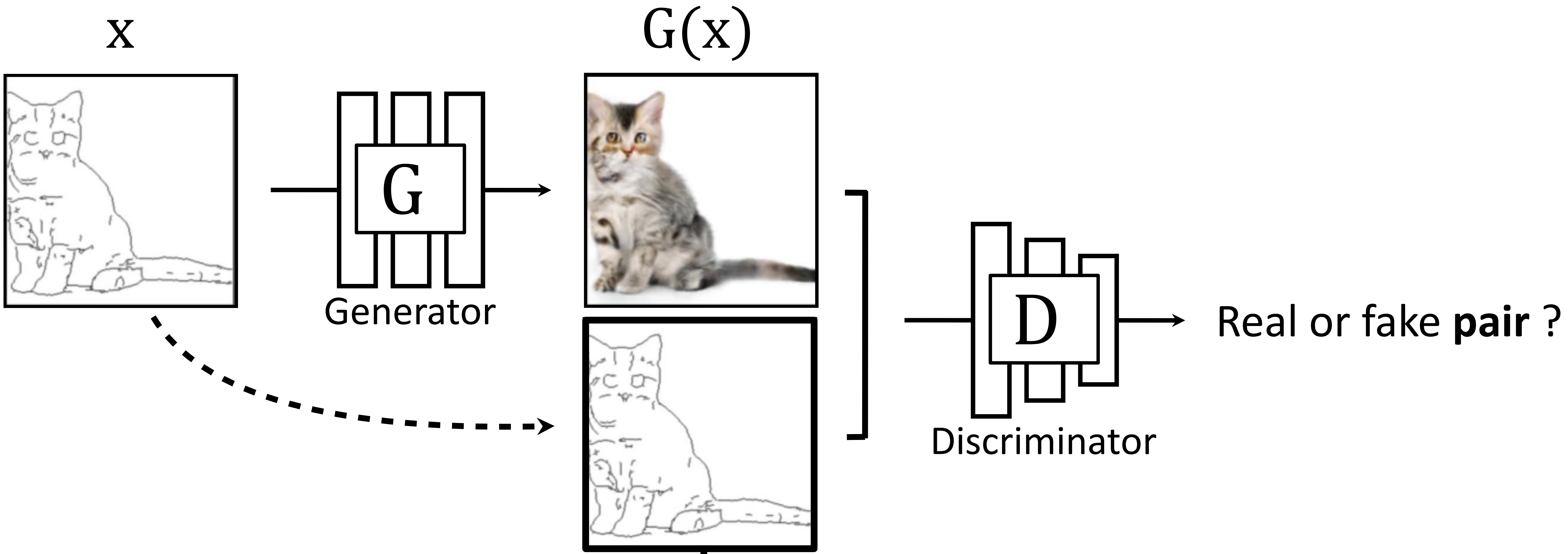
[Isola, Zhu, Zhou, Efros, 2016]



Learning objective (pix2pix)

$$\min_G \max_D \mathbb{E}\left[\log D(G(x)) + \log\left(1 - D(y)\right)\right]$$

[Isola, Zhu, Zhou, Efros, 2016]

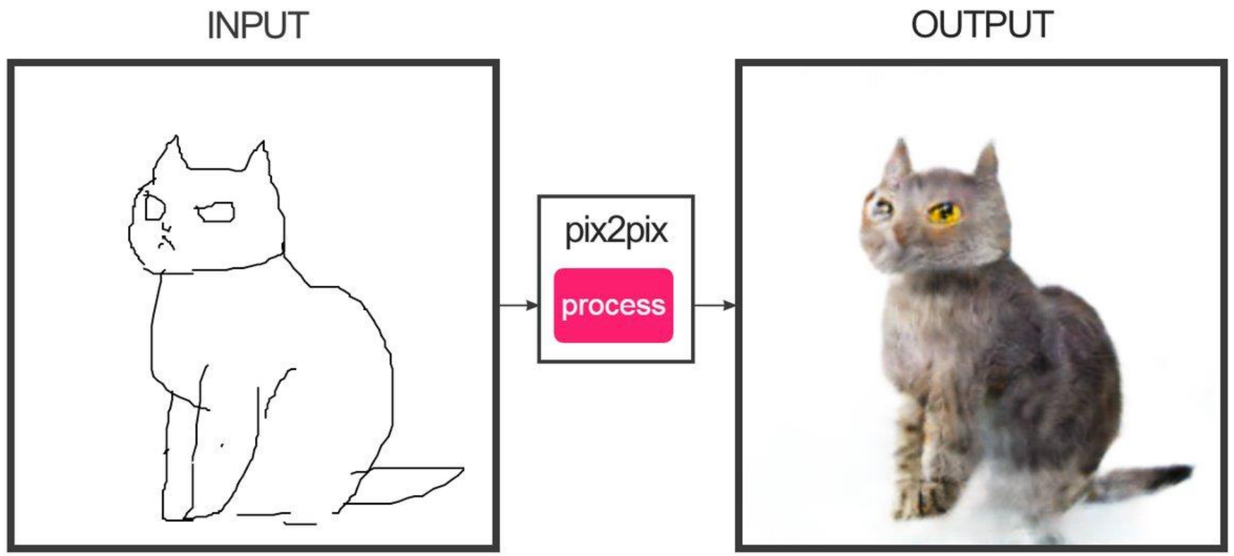


Learning objective (pix2pix)

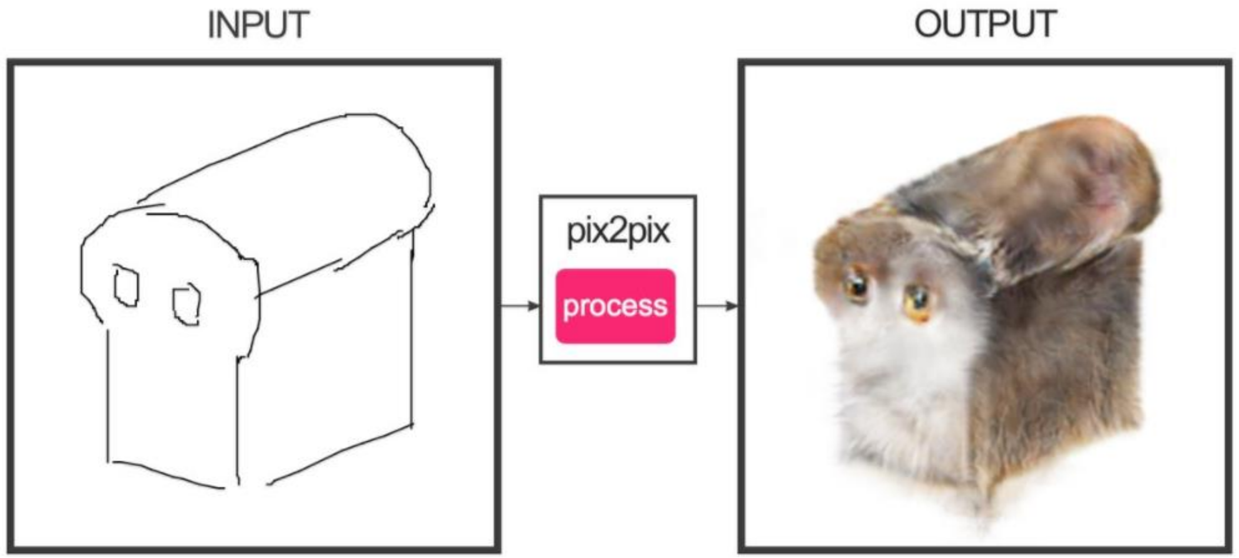
$$\min_G \max_D \mathbb{E}\left[\log D(\boxed{x}, G(x)) + \log\left(1 - D(\boxed{x}, y)\right)\right]$$

[Isola, Zhu, Zhou, Efros, 2016]

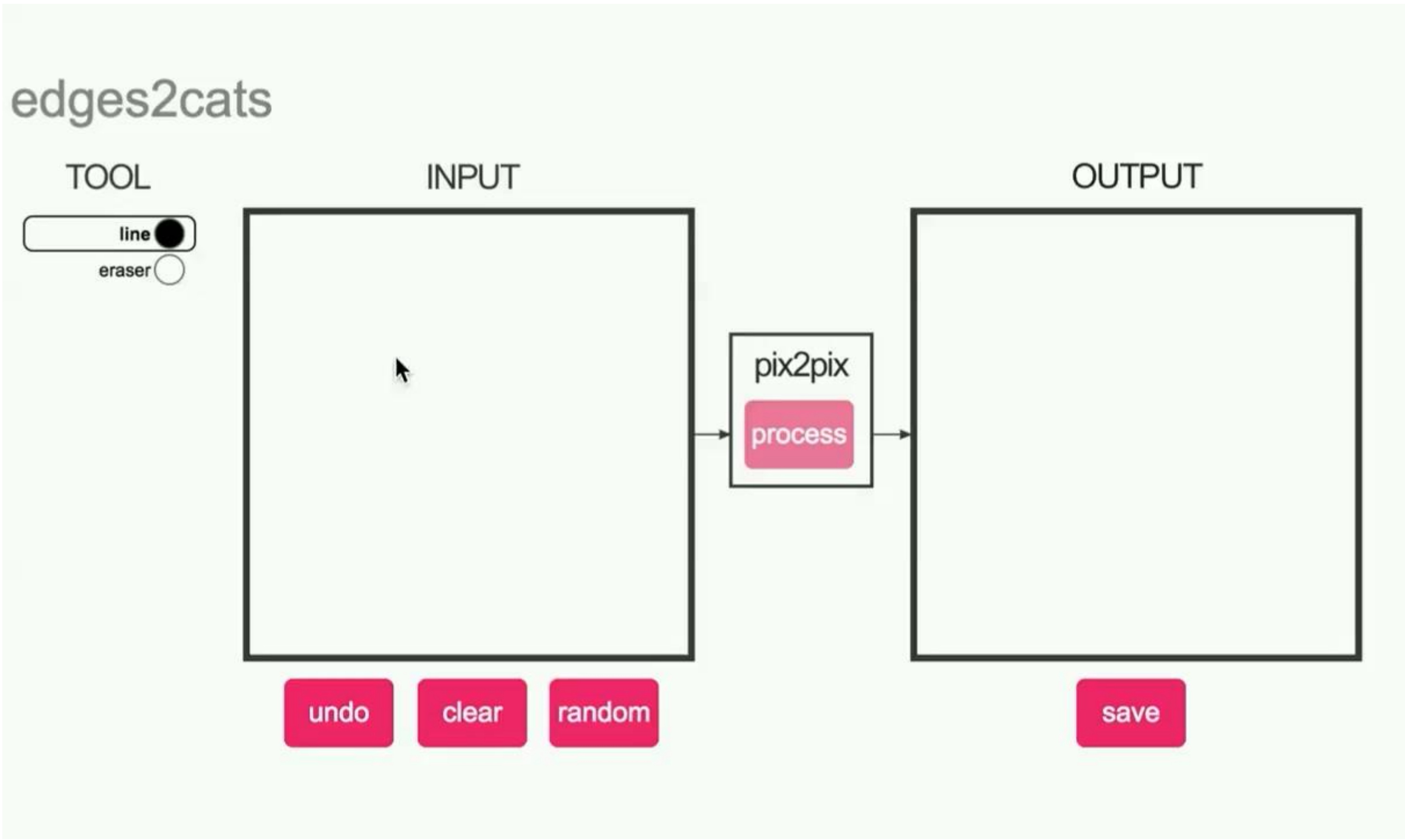
#edges2cats [Christopher Hesse]



@gods_tail



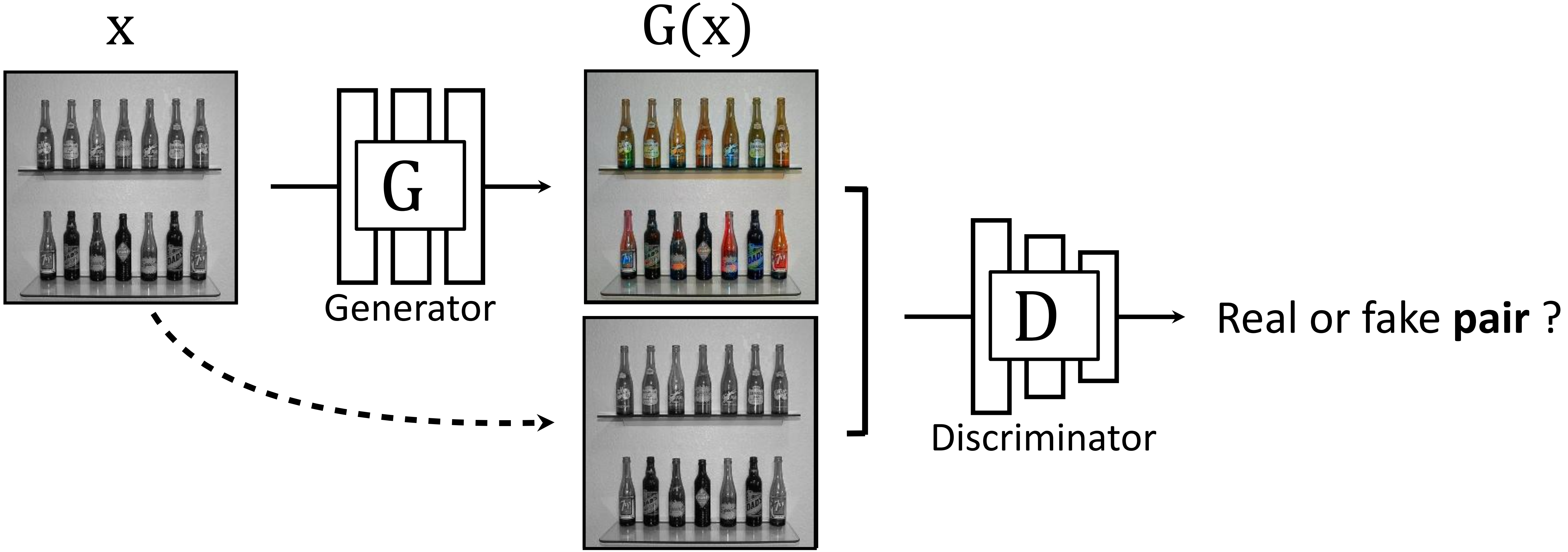
Ivy Tasi @ivymyt



@matthematician

Vitaly Vidmirov @vvid

https://affinelayer.com/pixsrv/



Input: **Sketch** → Output: **Photo**

Input: **Grayscale** → Output: **Color**

[Isola, Zhu, Zhou, Efros, 2016]

Automatic Colorization with pix2pix

Input Output Input Output Input Output

Data from [Russakovsky et al. 2015]

Interactive Colorization

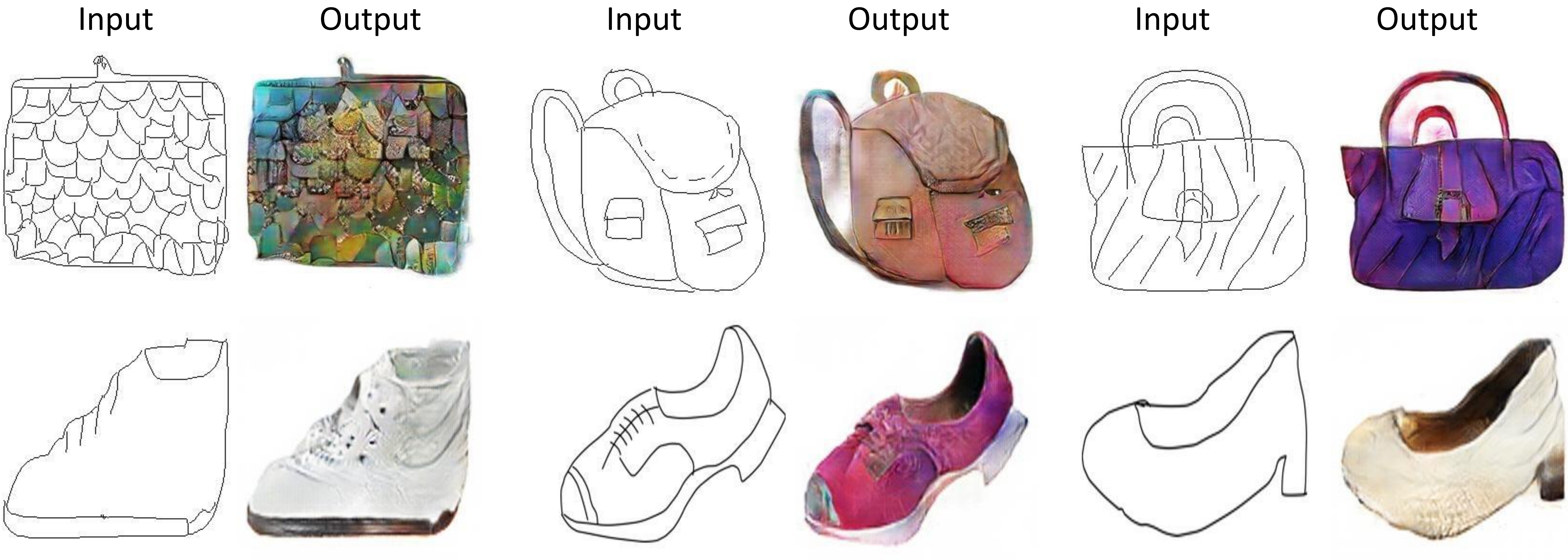
[Zhang*, Zhu*, Isola, Geng, Lin, Yu, Efros, 2017]

Edges → Images

Input Output Input Output Input Output

Edges from [Xie & Tu, 2015]

Sketches → Images



Trained on Edges → Images

Data from [Eitz, Hays, Alexa, 2012]

Input

Output

Groundtruth

Data from
[maps.google.com]

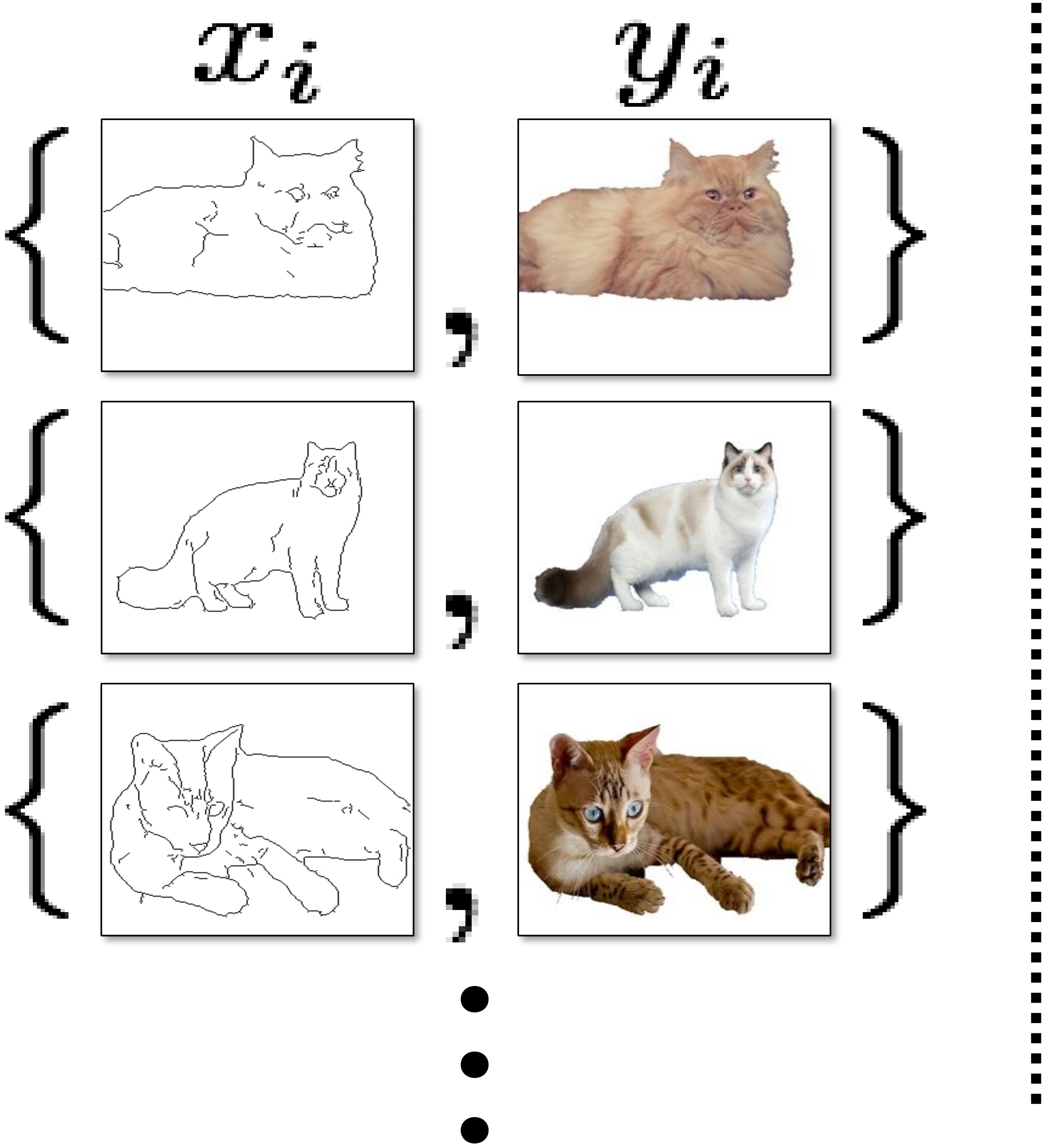
Input

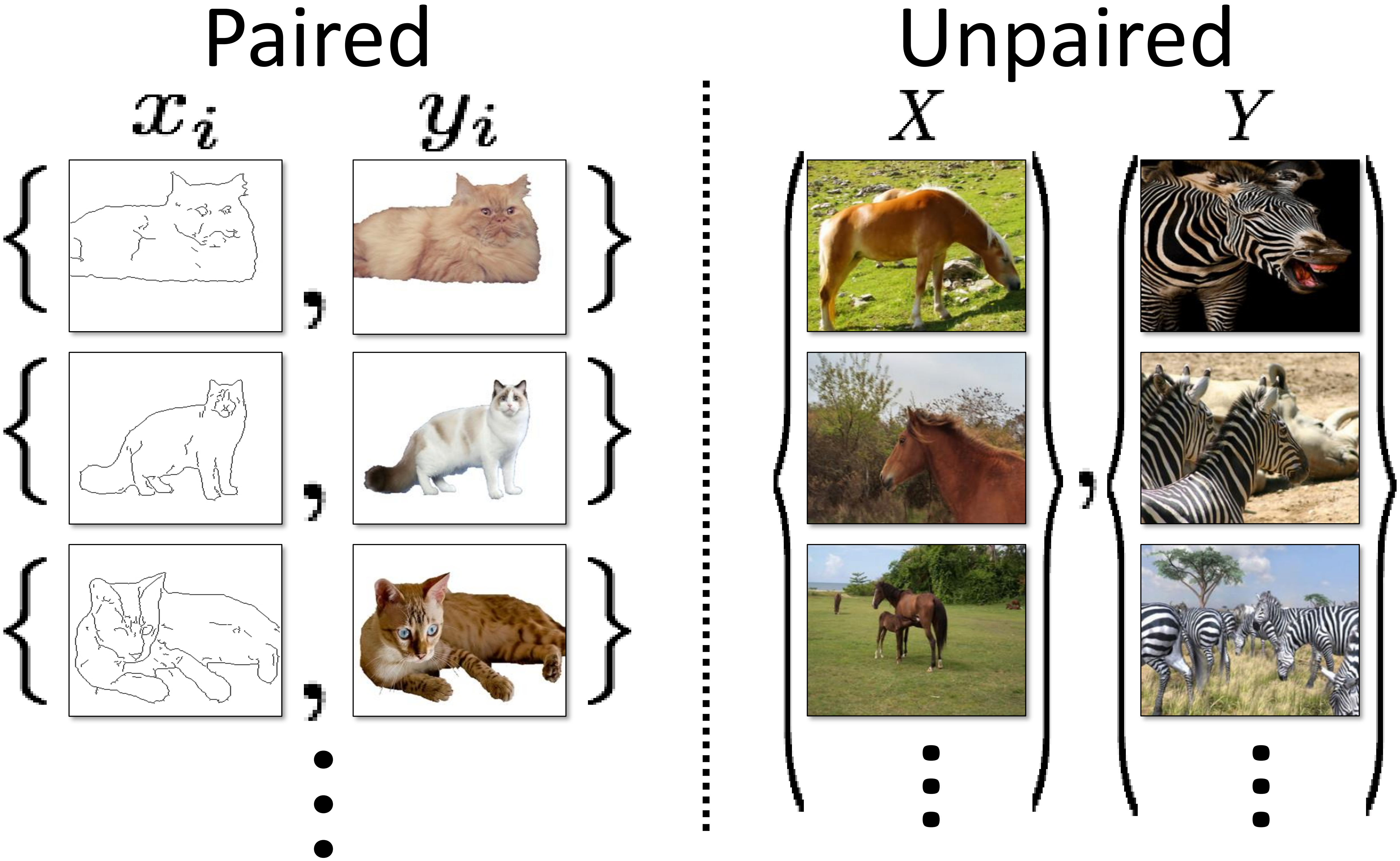
Output

Groundtruth

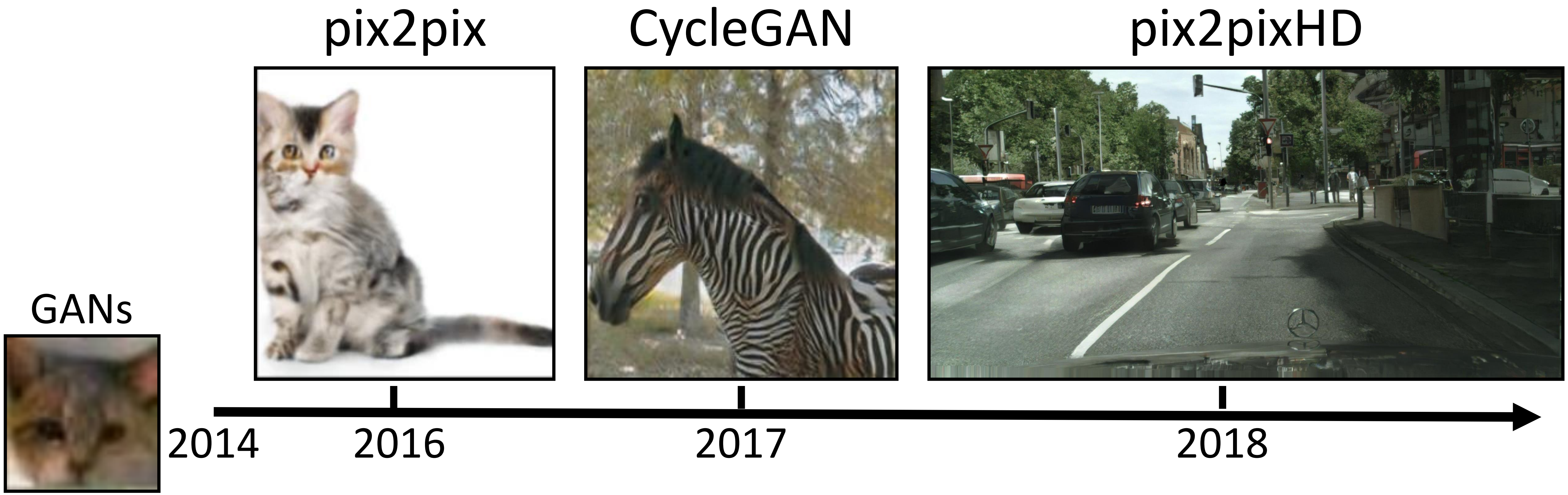
Data from [maps.google.com]

Paired




Paired
x_i
y_i
Unpaired
X
Y

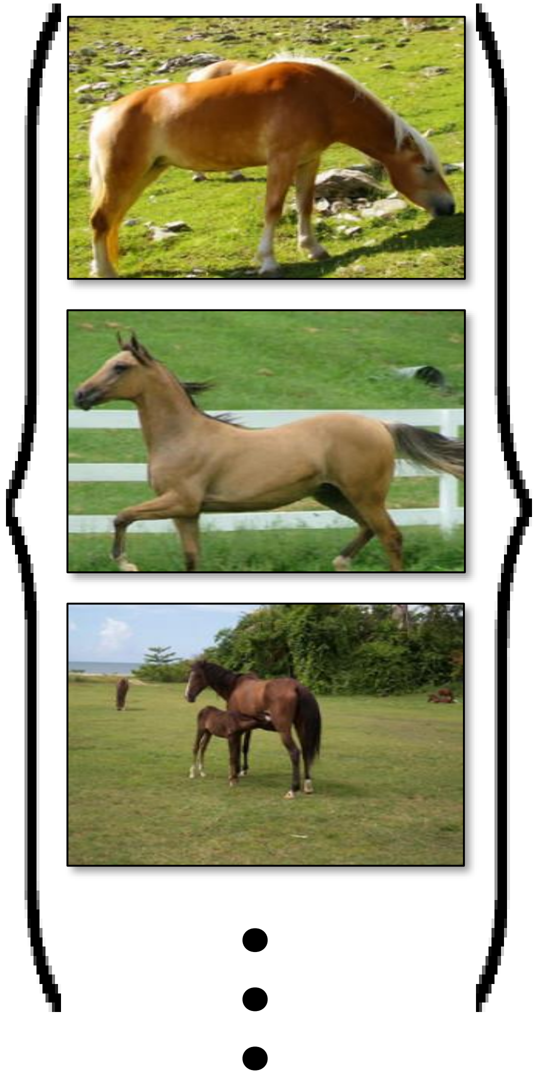
Goals: Improve Control, Quality, and Resolution

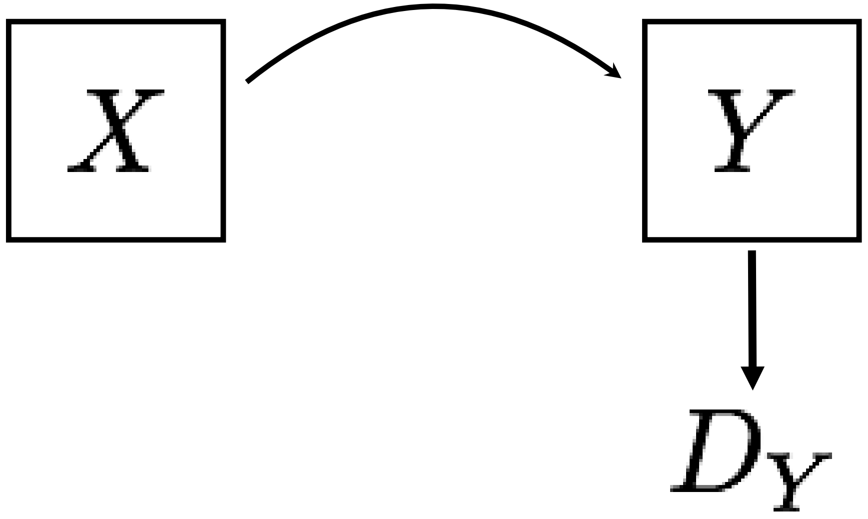


- Conditional on user inputs.
- Learning without pairs.
- High quality and resolution.

Cycle-Consistent Adversarial Networks

X

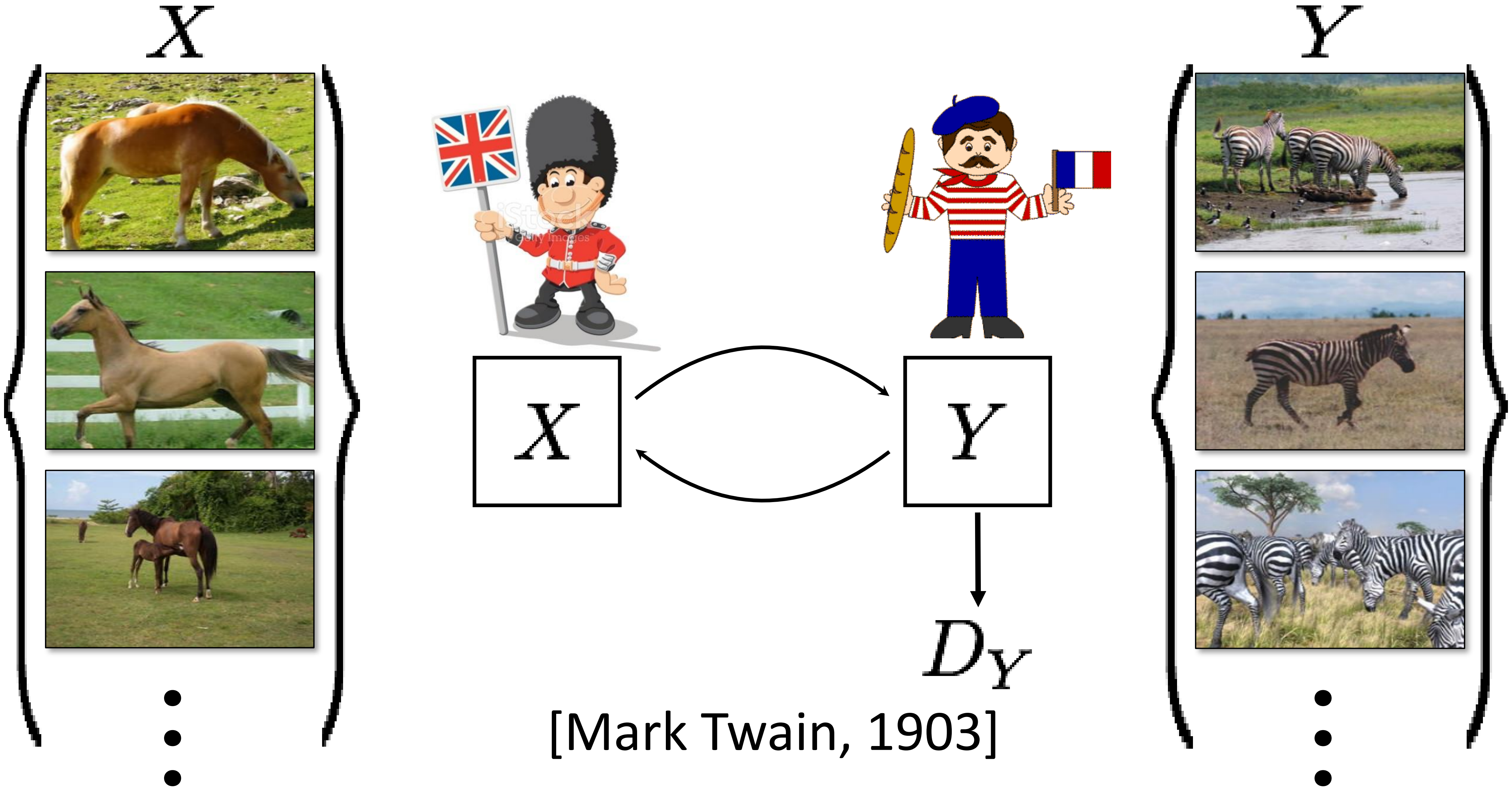




Y

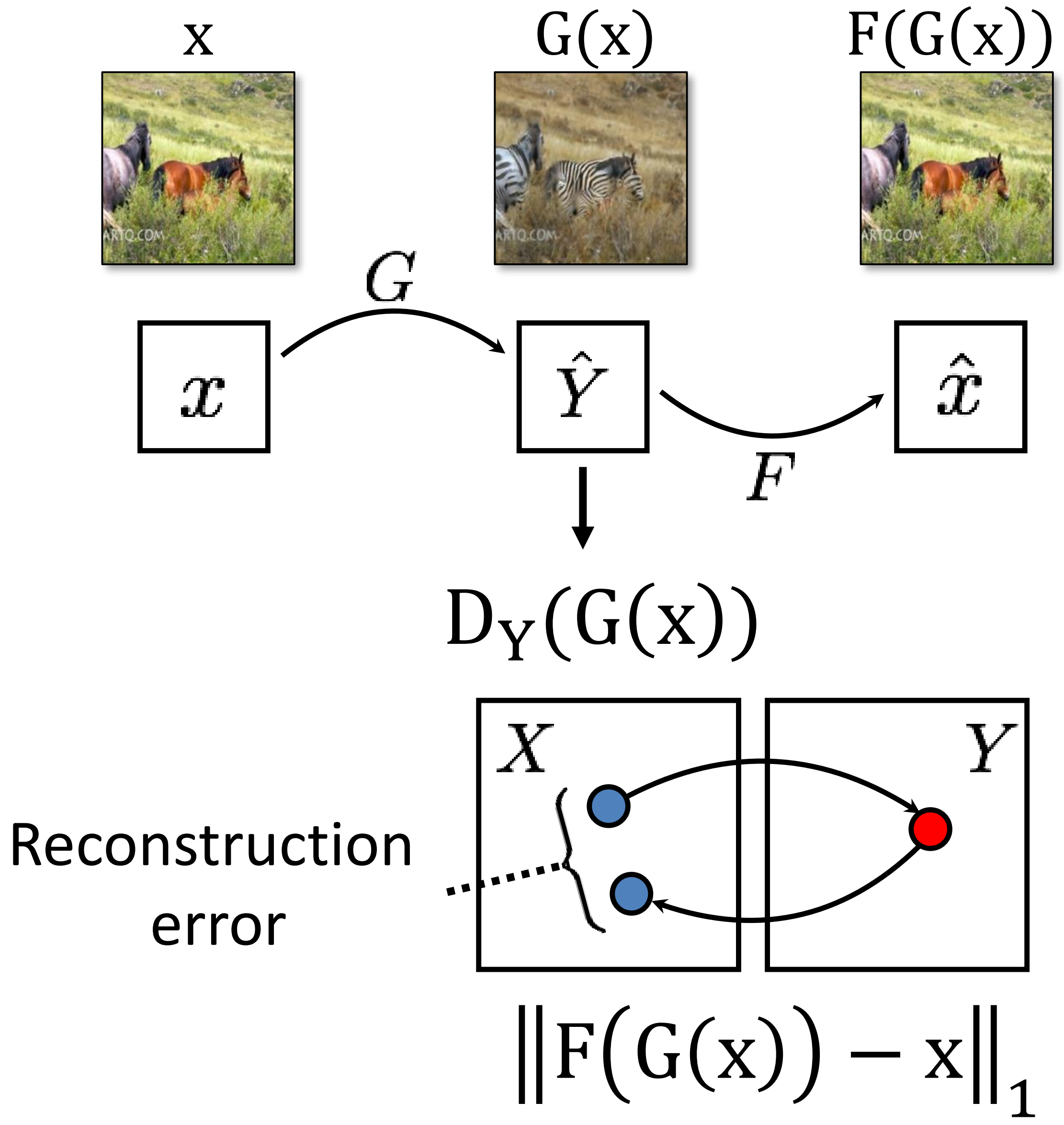
[Zhu*, Park*, Isola, and Efros, 2017]

Cycle-Consistent Adversarial Networks



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

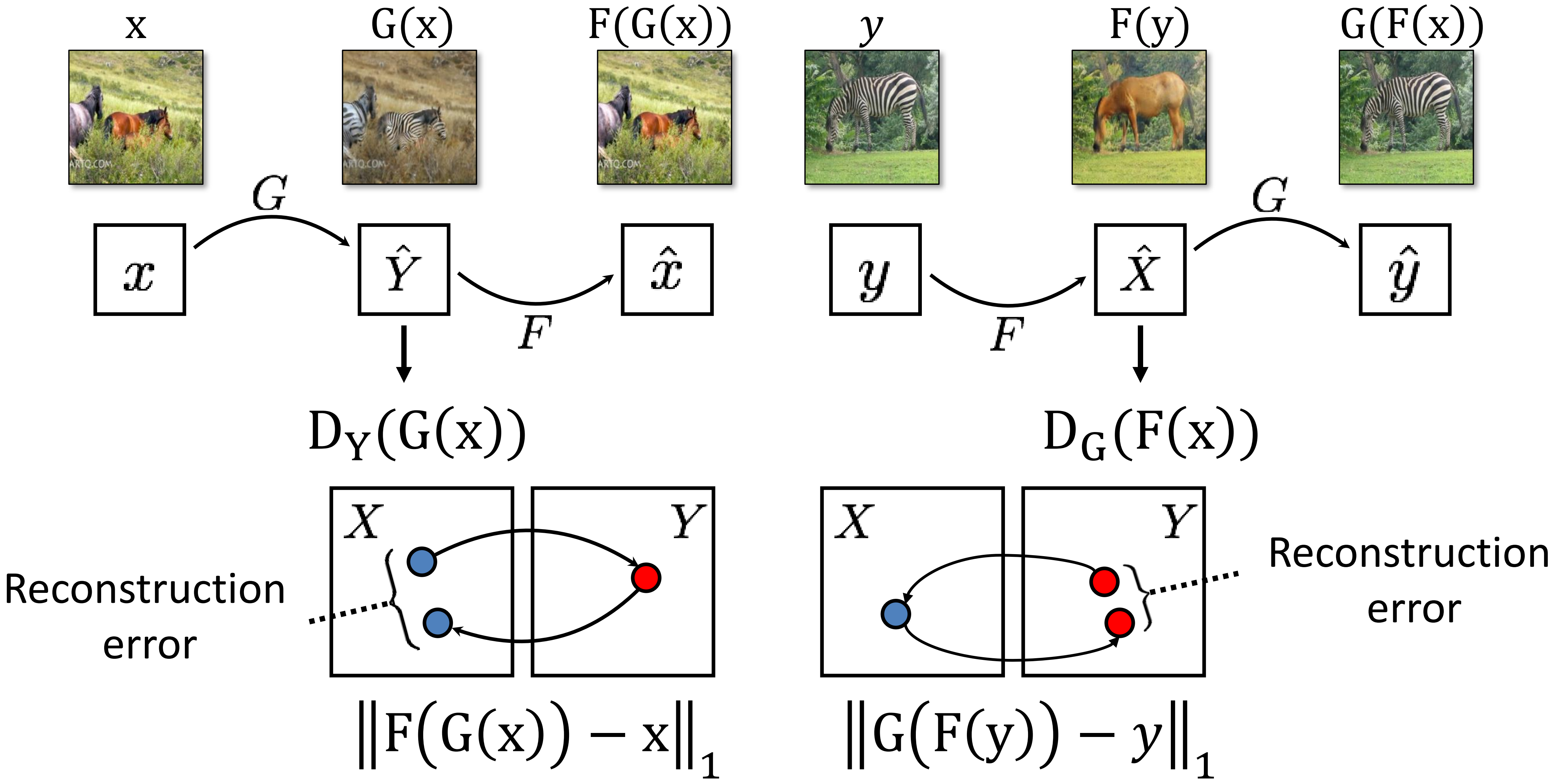
Cycle Consistency Loss



See also [Yi et al., 2017], [Kim et al, 2017]

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

Cycle Consistency Loss



See also [Yi et al., 2017], [Kim et al, 2017]

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

Horse → Zebra

Orange → Apple

Collection Style Transfer

Photograph ©Alexei Efros

Monet

Van Gogh

Cezanne

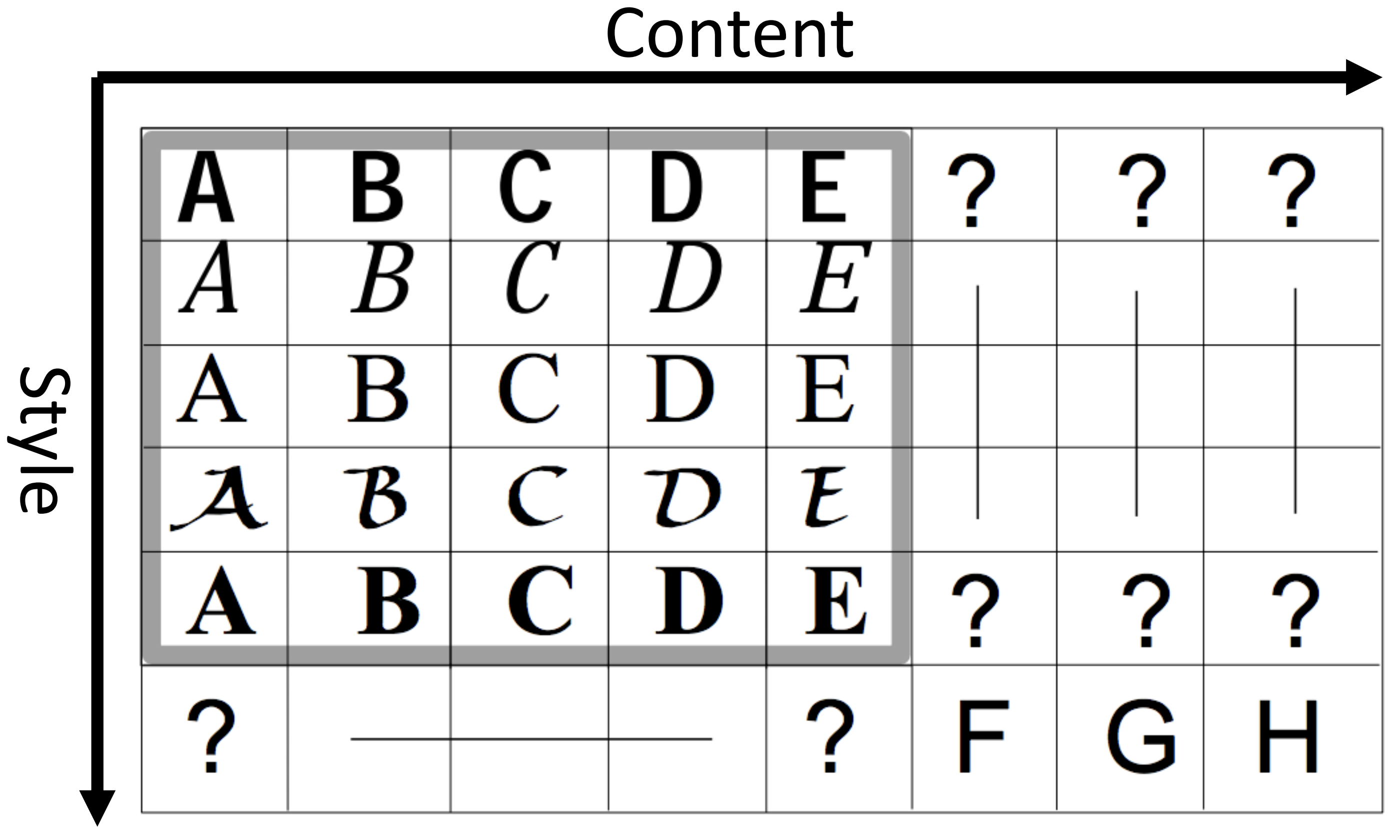
Ukiyo-e

Monet's paintings → photographic style

Why CycleGAN works

Style and Content Separation

Paired Separation



Separating Style and Content with Bilinear Models
[Tenenbaum and Freeman 2000']

Unpaired Separation

Adversarial Loss: change the style

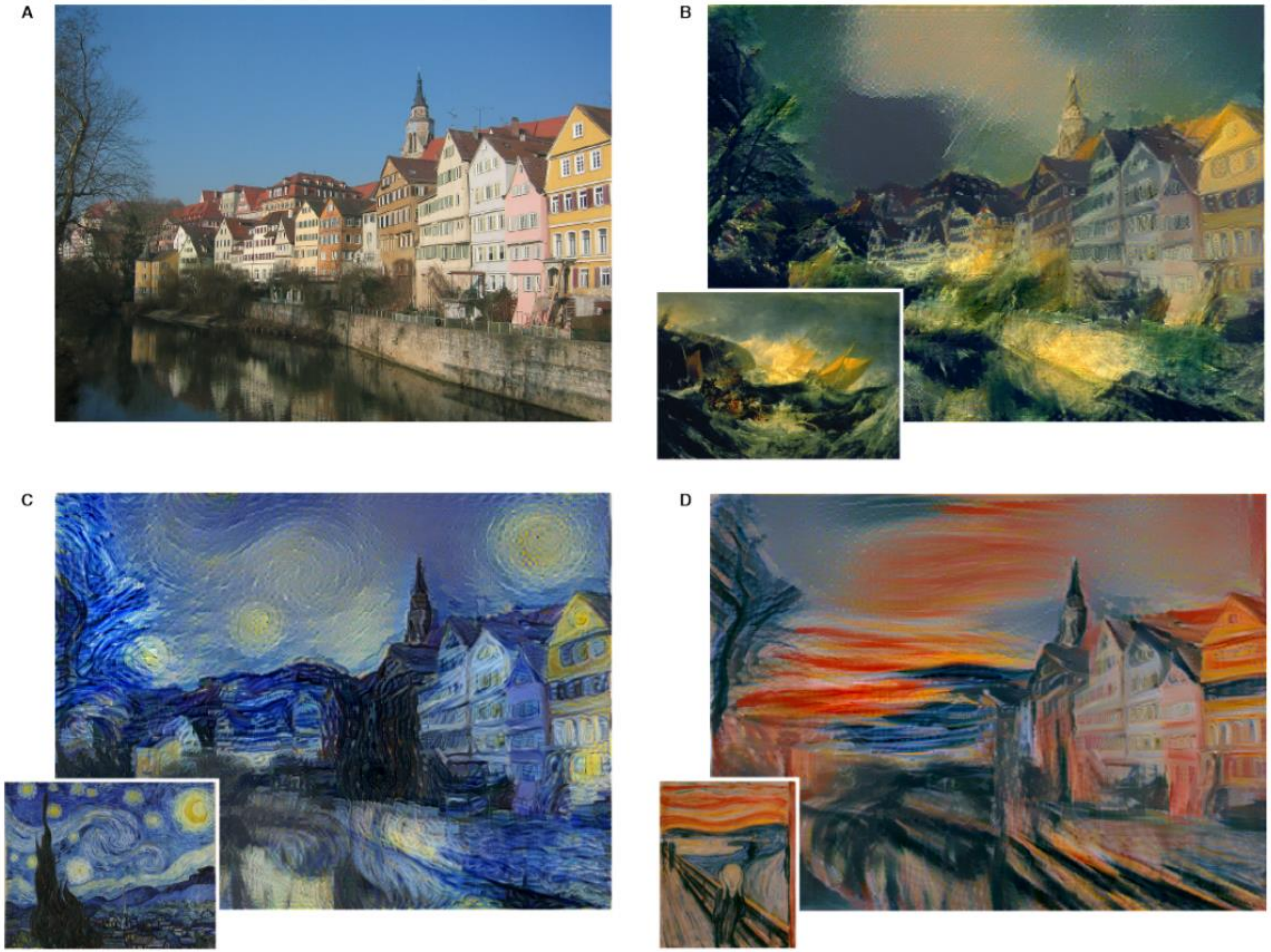
$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = &\mathbb{E}_{y\sim p_{\text{data}}(y)}[\log D_Y(y)] \\ &+\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x))]\end{aligned}$$

Cycle Consistency Loss: preserve the content

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G, F) = &\mathbb{E}_{x\sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] \\ &+\mathbb{E}_{y\sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1].\end{aligned}$$

Two empirical assumptions:
- content is easy to keep.
- style is easy to change.

Neural Style Transfer [Gatys et al. 2015]



Style and Content:

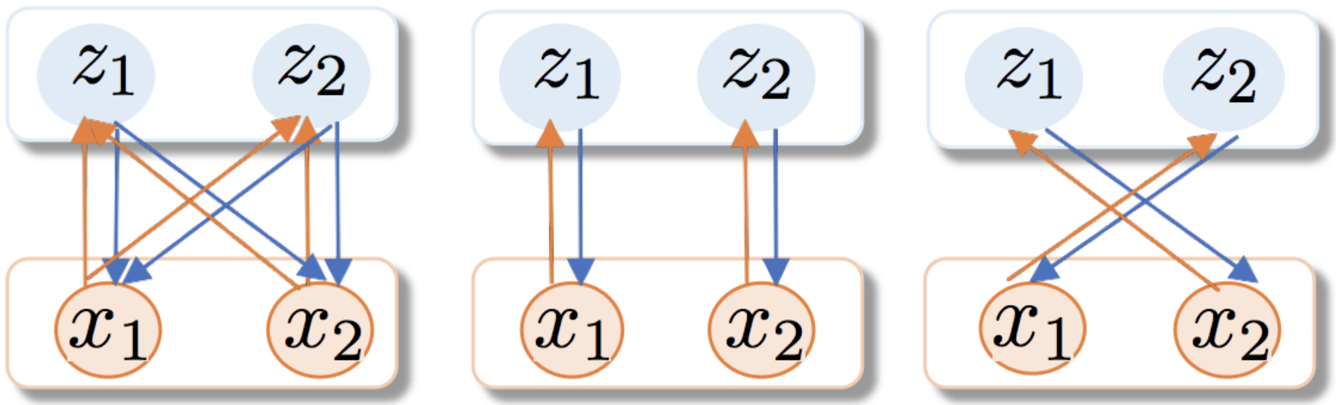
- Content: feature difference
- Style: Gram Matrix difference
- Both losses are hard-coded.

PRISMA

Photo → Van Gogh

horse → zebra

Cycle Loss **upper bounds** Conditional Entropy



	z_1	z_2
x_1	$\delta/2$	$(1-\delta)/2$
x_2	$(1-\delta)/2$	$\delta/2$

	z_1	z_2
x_1	1/2	0
x_2	0	1/2

	z_1	z_2
x_1	0	1/2
x_2	1/2	0

High Conditional Entropy

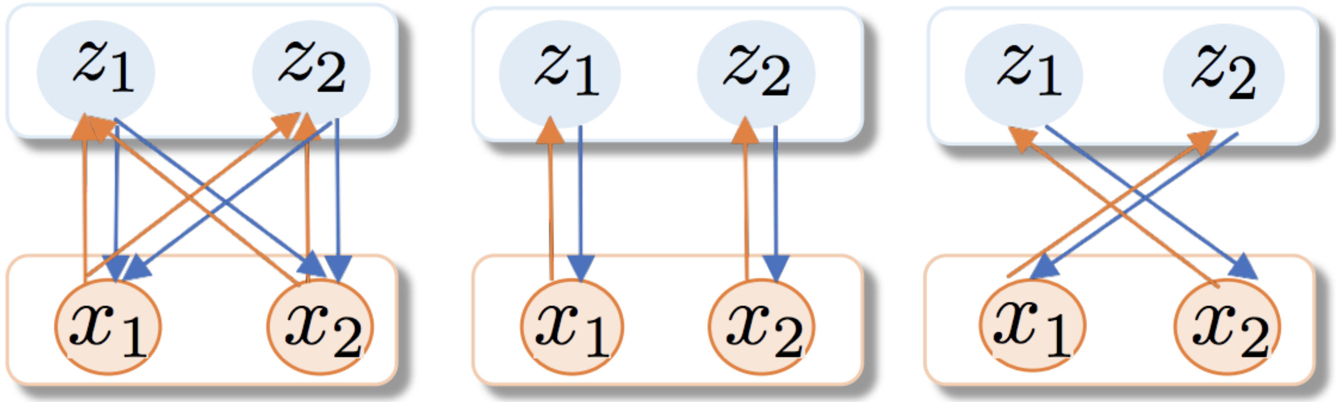
Low Conditional Entropy

Conditional Entropy

$$H^\pi(\boldsymbol{x}|\boldsymbol{z}) \triangleq -\mathbb{E}_{\pi(\boldsymbol{x},\boldsymbol{z})}[\log \pi(\boldsymbol{x}|\boldsymbol{z})]$$

"ALICE: Towards Understanding Adversarial Learning for Joint Distribution Matching" [Li et al. NIPS 2017]. Also see [Tiao et al. 2018] "CycleGAN as Approximate Bayesian Inference"

Cycle Loss **upper bounds** Conditional Entropy



	z_1	z_2
x_1	$\delta/2$	$(1-\delta)/2$
x_2	$(1-\delta)/2$	$\delta/2$

	z_1	z_2
x_1	1/2	0
x_2	0	1/2

	z_1	z_2
x_1	0	1/2
x_2	1/2	0

Conditional Entropy

$$H^{\pi}(\boldsymbol{x}|\boldsymbol{z}) \triangleq -\mathbb{E}_{\pi(\boldsymbol{x},\boldsymbol{z})}[\log \pi(\boldsymbol{x}|\boldsymbol{z})]$$

Lemma 3 *For joint distributions $p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$ or $q_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{z})$, we have*

$$\boxed{H^{q_{\boldsymbol{\phi}}}(\boldsymbol{x}|\boldsymbol{z})} \triangleq -\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{z})}[\log q_{\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{z})] = -\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{z})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})] - \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z})}[\mathrm{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{z})\|p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}))]$$
$$\boxed{\leq} -\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{z})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})] \triangleq \boxed{\mathcal{L}_{\text{Cycle}}(\boldsymbol{\theta}, \boldsymbol{\phi}).} \quad (6)$$

“ALICE: Towards Understanding Adversarial Learning for Joint Distribution Matching” [Li et al. NIPS 2017]. Also see [Tiao et al. 2018] “CycleGAN as Approximate Bayesian Inference”

Customizing Gaming Experience

Grand Theft Auto v (GTA5)

Street view images in German cities

Data from [Richter et al., 2016], [Cordts et al, 2016]

Customizing Gaming Experience

Output image with German street view style

Input GTA5 CG

Domain Adaptation with CycleGAN

Train on GTA5 data

Test on real images

	meanIOU	Per-pixel accuracy
Oracle (Train and test on Real)	60.3	93.1
Train on CG, test on Real	17.9	54.0

See Judy Hoffman's talk at 14:30 "Adversarial Domain Adaptation"

Domain Adaptation with CycleGAN

GTA5 data + Domain adaptation

Test on real images

	meanIOU	Per-pixel accuracy
Oracle (Train and test on Real)	60.3	93.1
Train on CG, test on Real	17.9	54.0
FCN in the wild [Previous STOA]	27.1	-

See Judy Hoffman's talk at 14:30 "Adversarial Domain Adaptation"

Domain Adaptation with CycleGAN

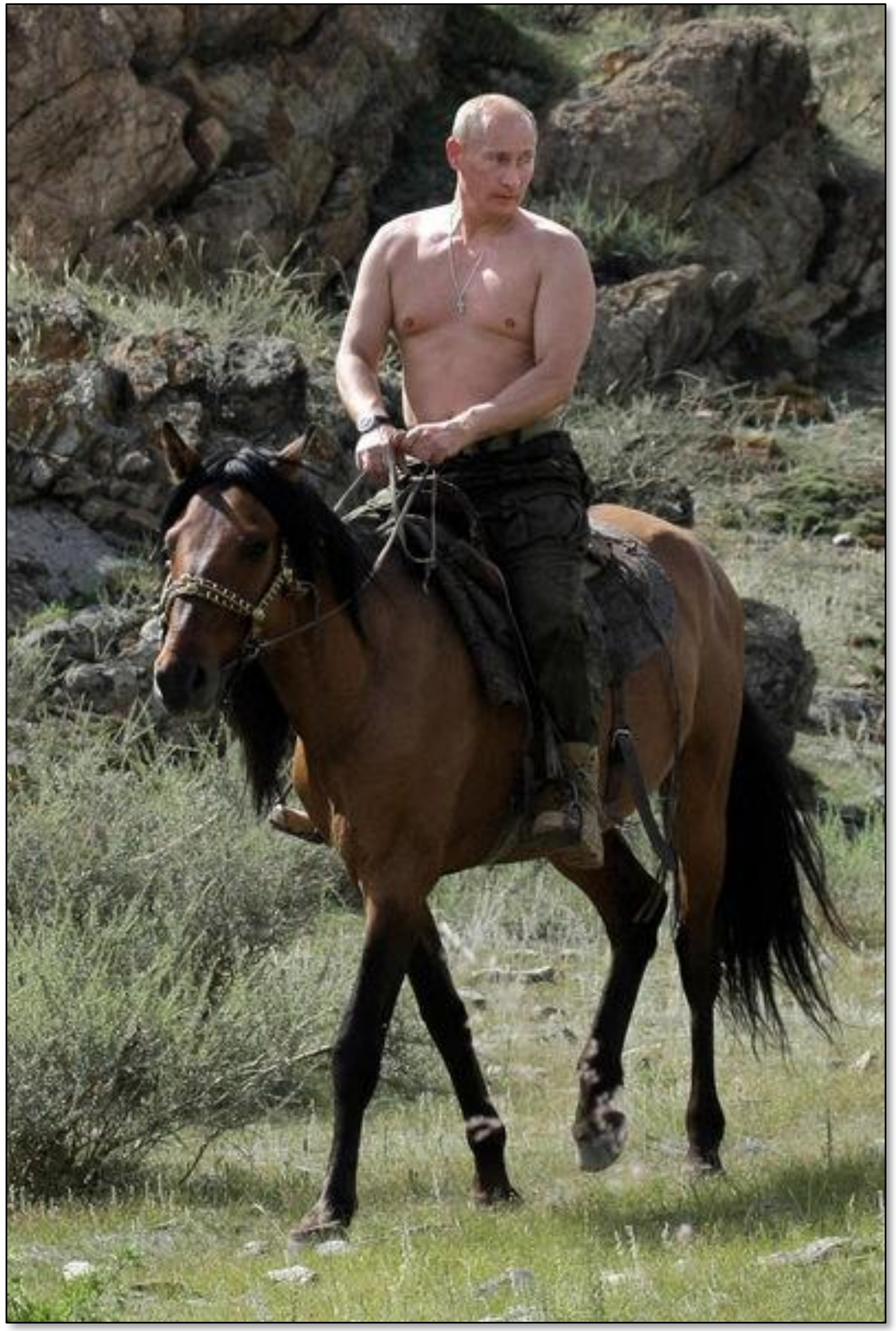
Train on CycleGAN data

Test on real images

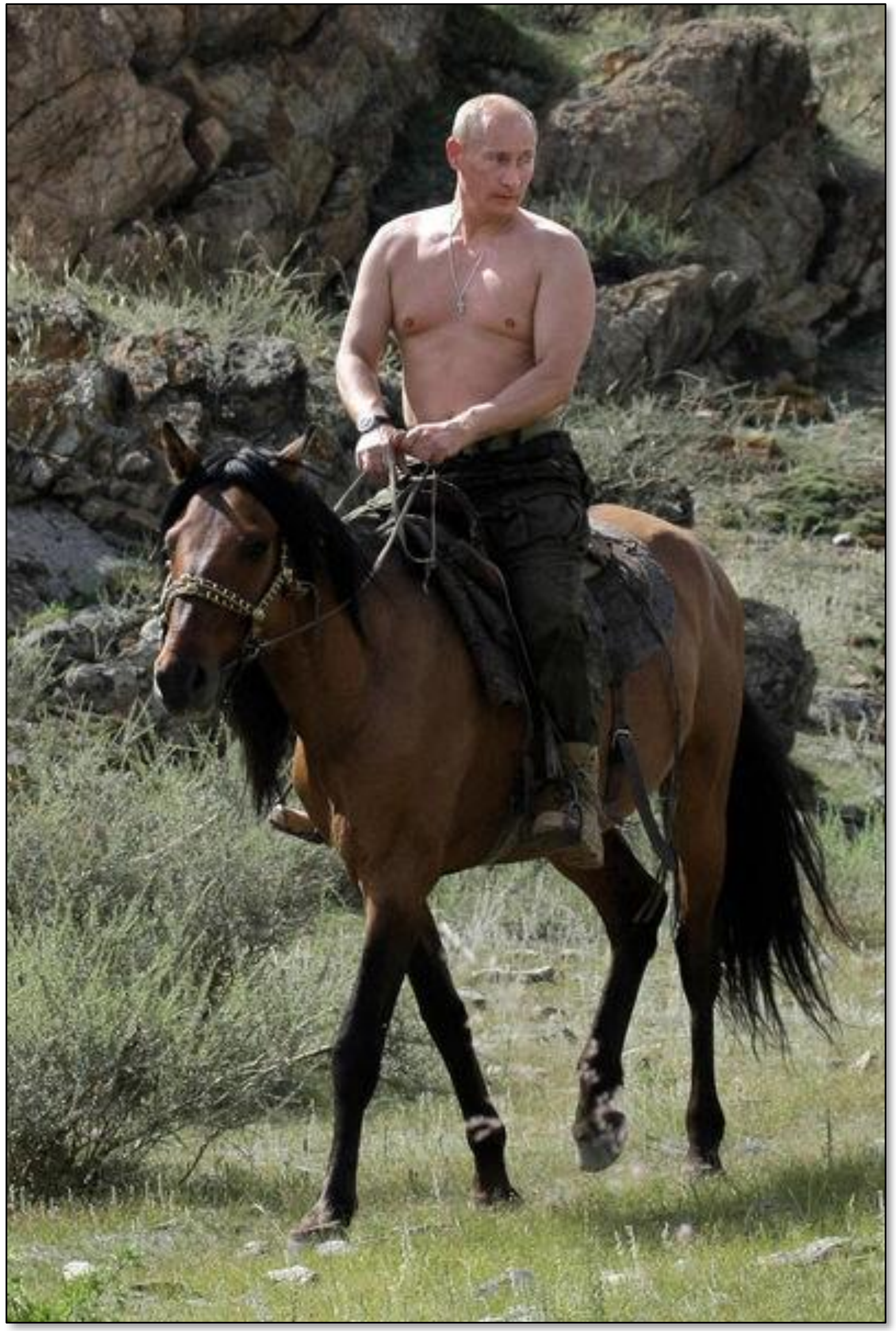
	meanIOU	Per-pixel accuracy
Oracle (Train and test on Real)	60.3	93.1
Train on CG, test on Real	17.9	54.0
FCN in the wild [Previous STOA]	27.1	-
Train on CycleGAN, test on Real	**34.8**	**82.8**

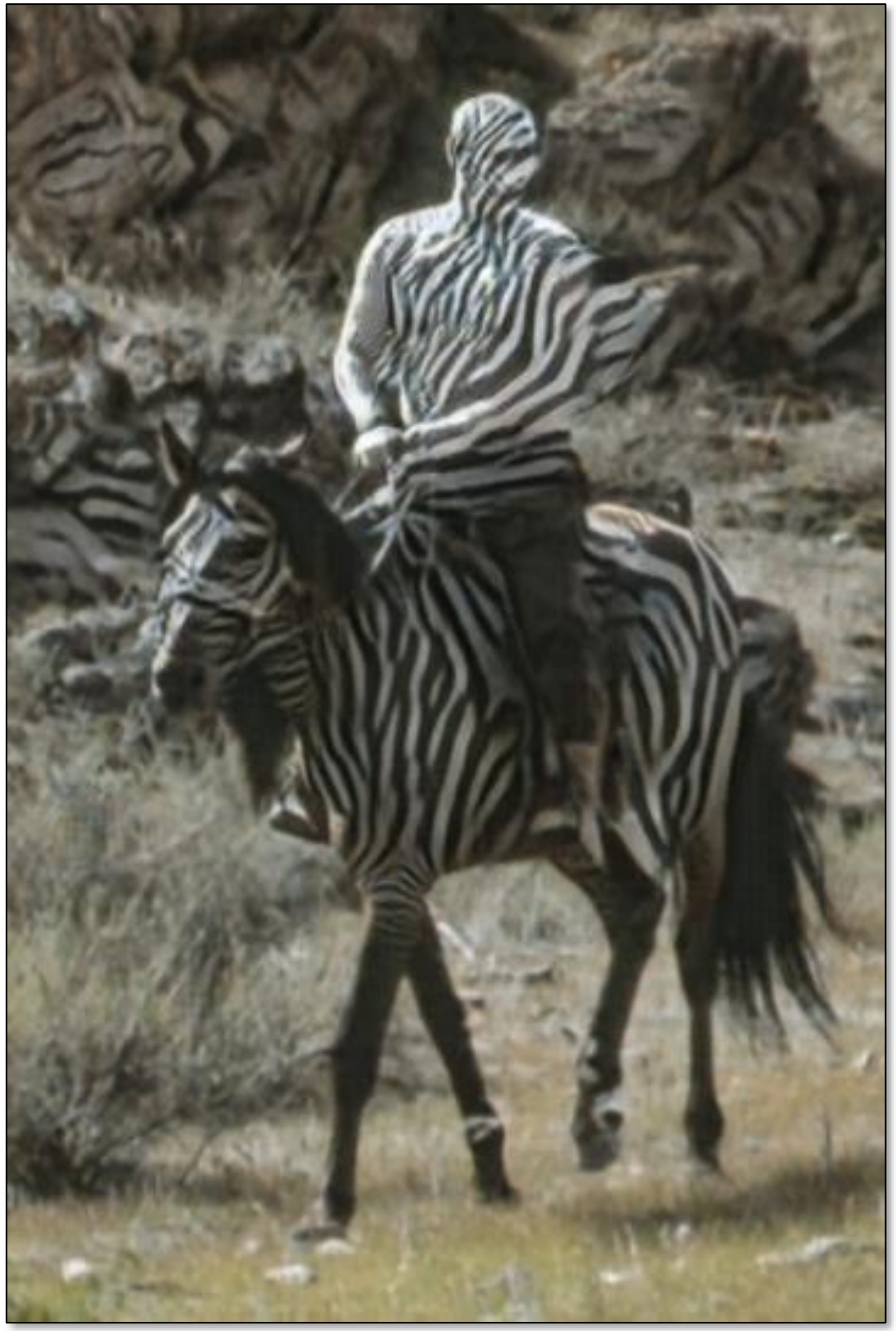
See Judy Hoffman's talk at 14:30 "Adversarial Domain Adaptation"

Failure case



Failure case





Open Source CycleGAN and pix2pix

pytorch-CycleGAN-and-pix2pix

Image-to-image translation in PyTorch (e.g., horse2zebra, edges2cats, and more)

Python ★ 4.9k 1.1k

CycleGAN

Software that can generate photos from paintings, turn horses into zebras, perform style transfer, and more.

Lua ★ 6.9k 1k

- Among the most **popular** GitHub research projects since 2017.
- Among the most **cited** papers in Graphics/CV/ML since 2017.

CycleGAN in Classes

Stanford University

Berkeley UNIVERSITY OF CALIFORNIA

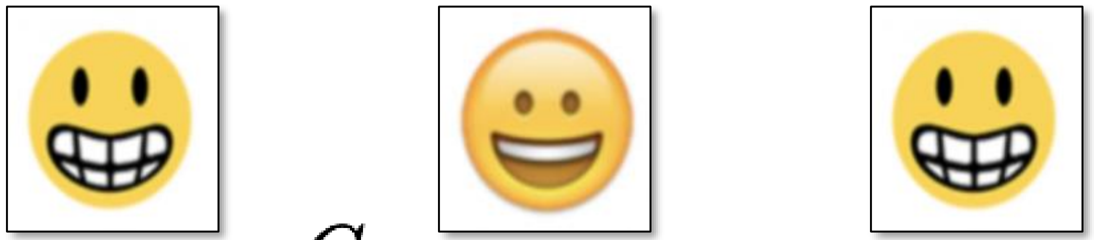
UNIVERSITY OF TORONTO

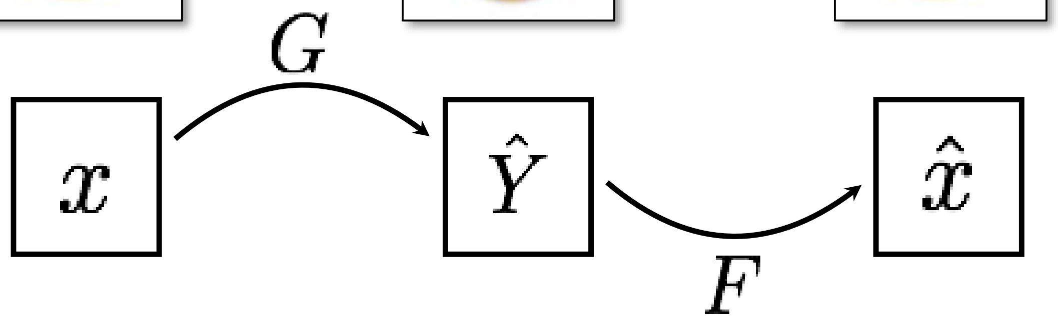
UDACITY

Faster AI

CycleGAN results by students

MS emoji Apple emoji MS emoji





© Roger Grosse, UoT

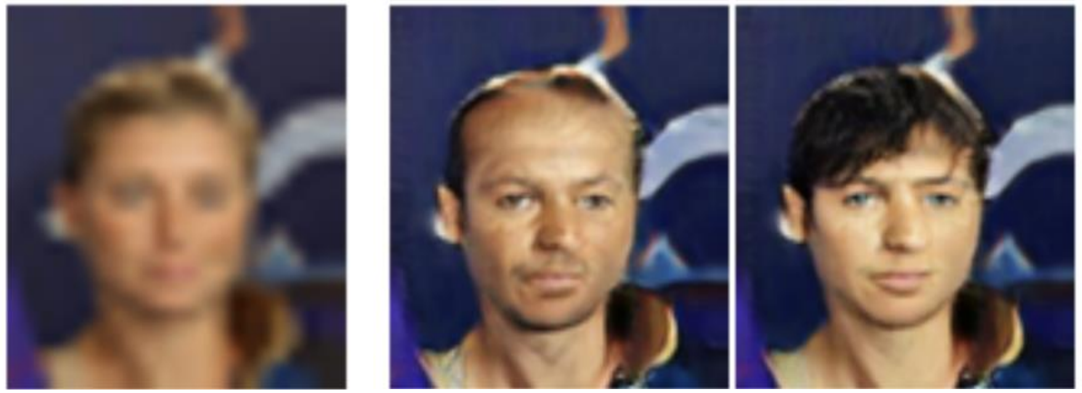
Input photo

Stained glass art

© Alena Harley, FastAI

Applications and Extentions

Attribute Editing [Lu et al.]



Low-res Bald Bangs

arXiv:1705.09966

Object Editing [Liang et al.]

Mask Input Output

arXiv:1708.00315

Front/Character Transfer [Ignatov et al.]

Input output

arXiv: 1801.08624

Data generation [Wang et al.]

samples by CycleWGAN

arXiv:1707.03124

Photo Enhancement

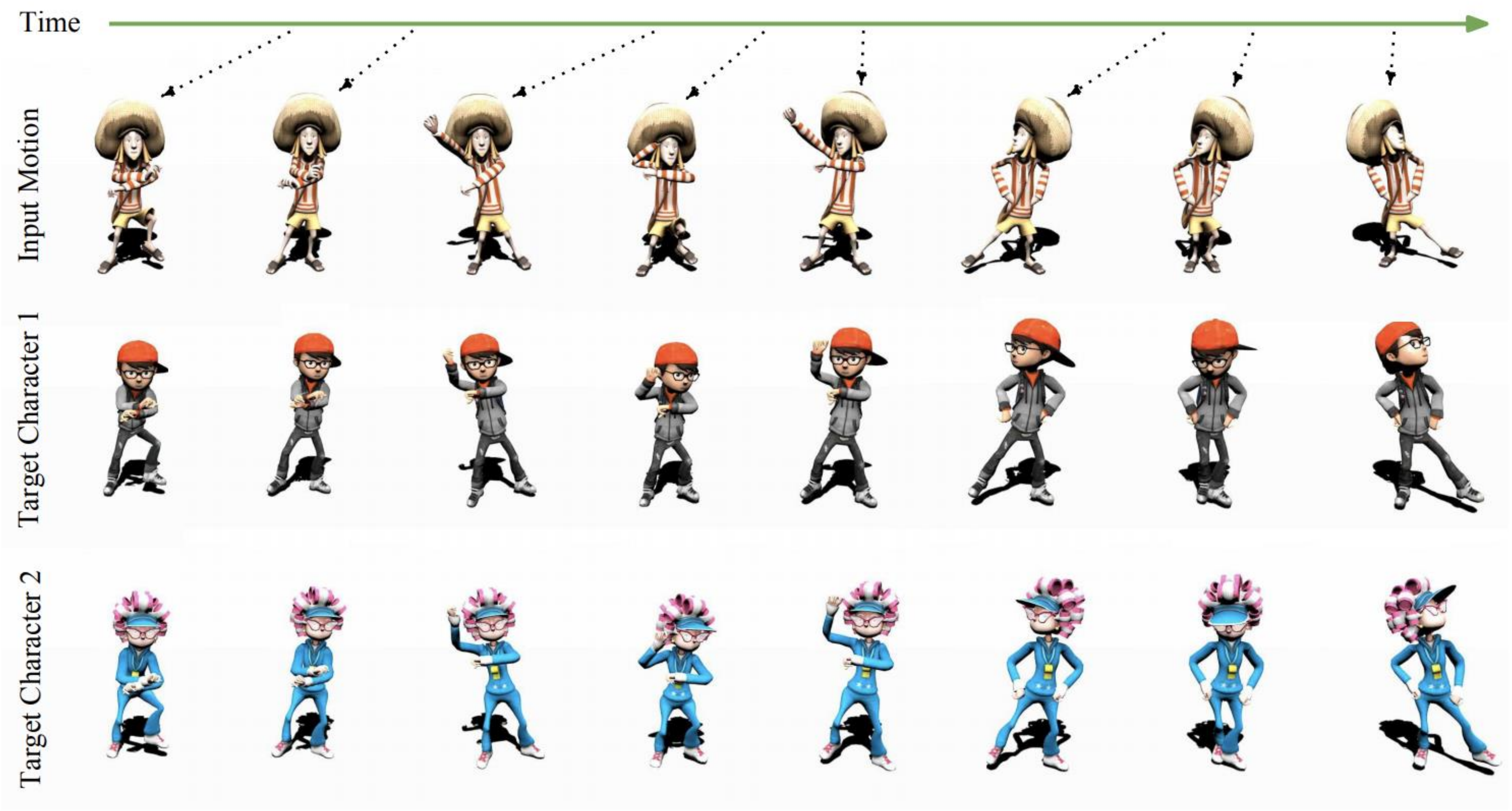
WESPE: Weakly Supervised Photo Enhancer for Digital Cameras. arxiv 1709.01118

Andrey Ignatov, Nikolay Kobyshev, Kenneth Vanhoey, Radu Timofte, Luc Van Gool

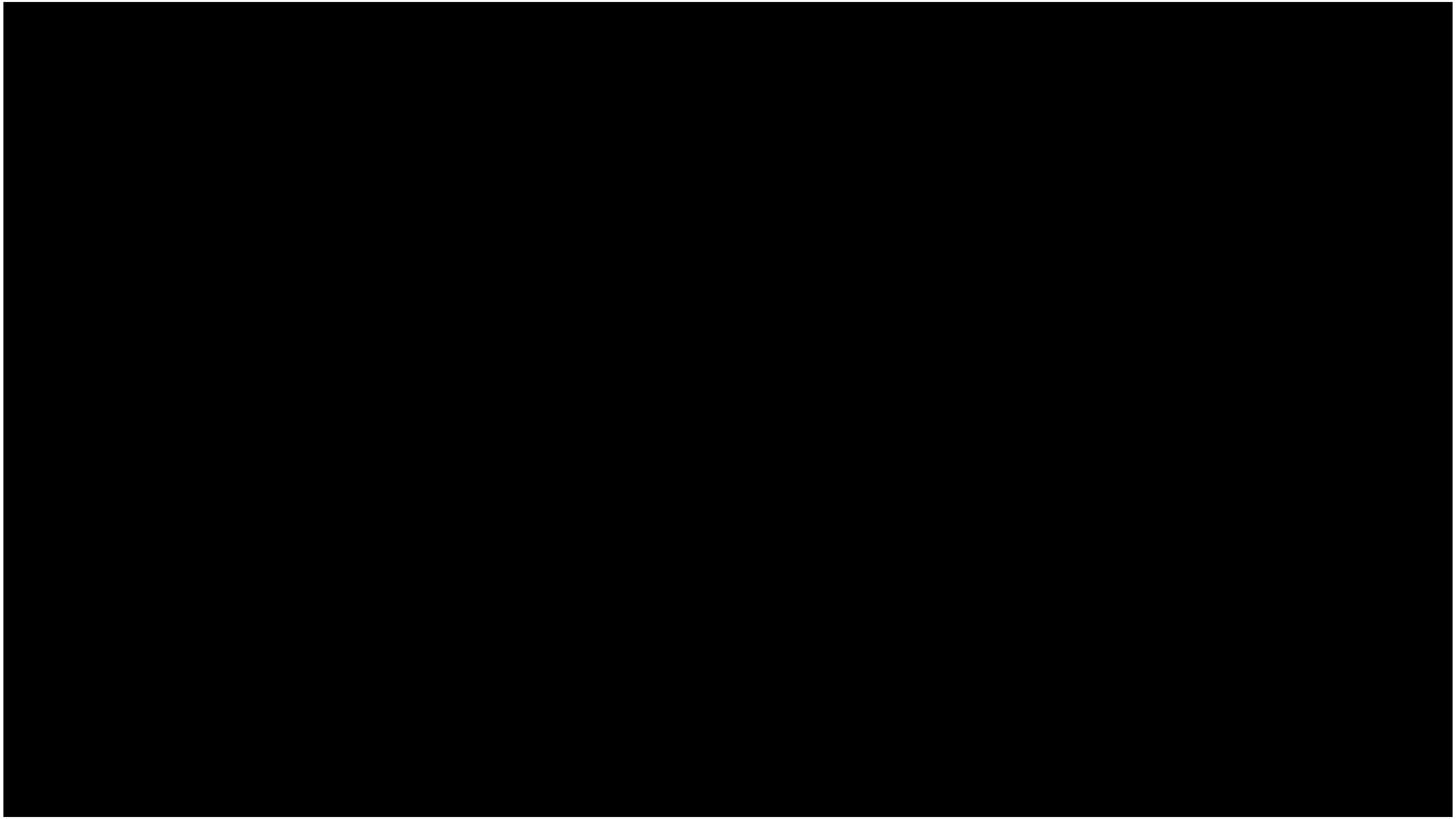
Image Dehazing

Cycle-Dehaze: Enhanced CycleGAN for Single Image Dehazing. CVPRW 2018
Deniz Engin∗ Anıl Genc∗, Hazım Kemal Ekenel

Unsupervised Motion Retargeting



Neural Kinematic Networks for Unsupervised Motion Retargetting. CVPR 2018 (oral)
Ruben Villegas, Jimei Yang, Duygu Ceylan, Honglak Lee



Neural Kinematic Networks for Unsupervised Motion Retargetting. CVPR 2018 (oral)
Ruben Villegas, Jimei Yang, Duygu Ceylan, Honglak Lee

Applications Beyond Computer Vision

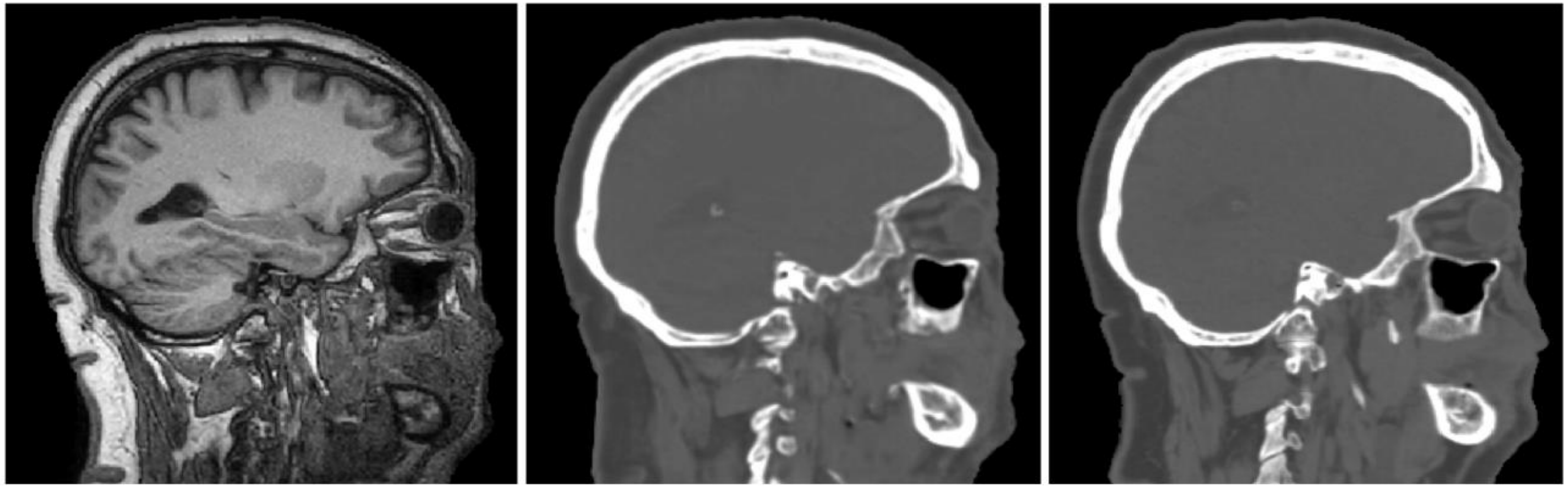
- Medical Imaging and Biology [Wolterink et al., 2017]
- Voice conversion [Fang et al., 2018, Kaneko et al., 2017]
- Cryptography [CipherGAN: Gomez et al., ICLR 2018]
- Robotics
- NLP: Unsupervised machine translation.
- NLP: Text style transfer.

…

Deep MR to CT Synthesis using Unpaired Data

Jelmer M. Wolterink[1]✉, Anna M. Dinkla[2], Mark H.F. Savenije[2], Peter R. Seevinck[1], Cornelis A.T. van den Berg[2], Ivana Išgum[1]

[1] Image Sciences Institute, University Medical Center Utrecht, The Netherlands
j.m.wolterink@umcutrecht.nl
[2] Department of Radiotherapy, University Medical Center Utrecht, The Netherlands



Input MR | Generated CT | Ground truth CT

Latest from #CycleGAN

Input dog

Output cat

Input cat

Output dog

CycleGAN for Customized Gaming

Battle royale games

Fortnite Input

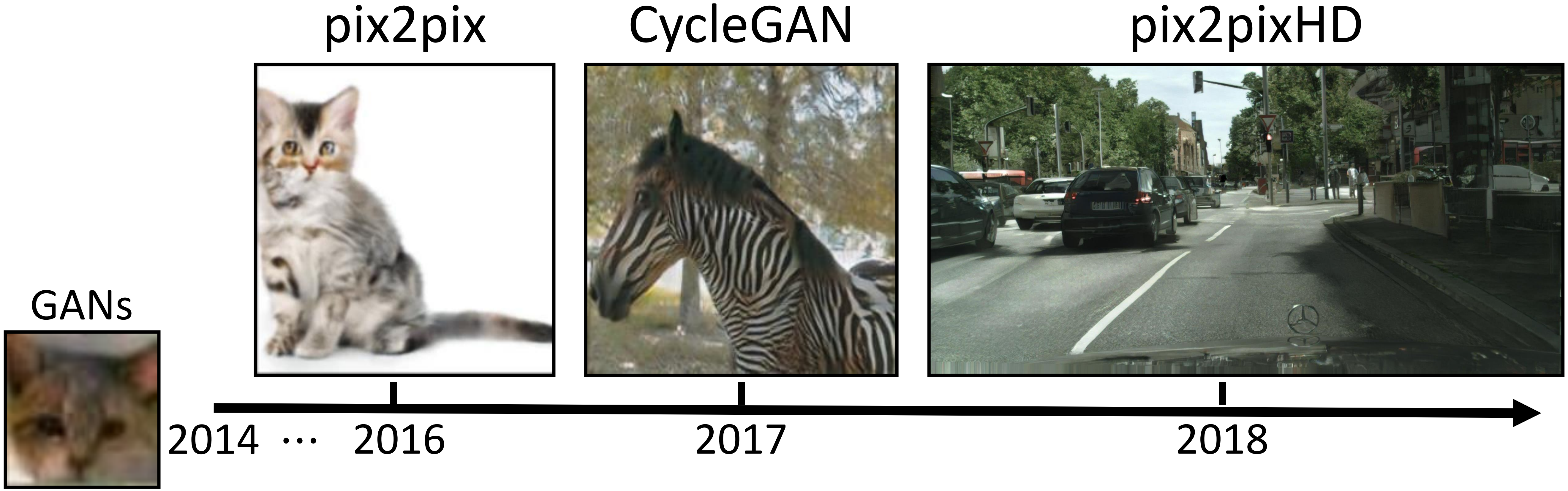
+

PUBG Style

=

Final result

Goals: Improve Control, Quality, and Resolution

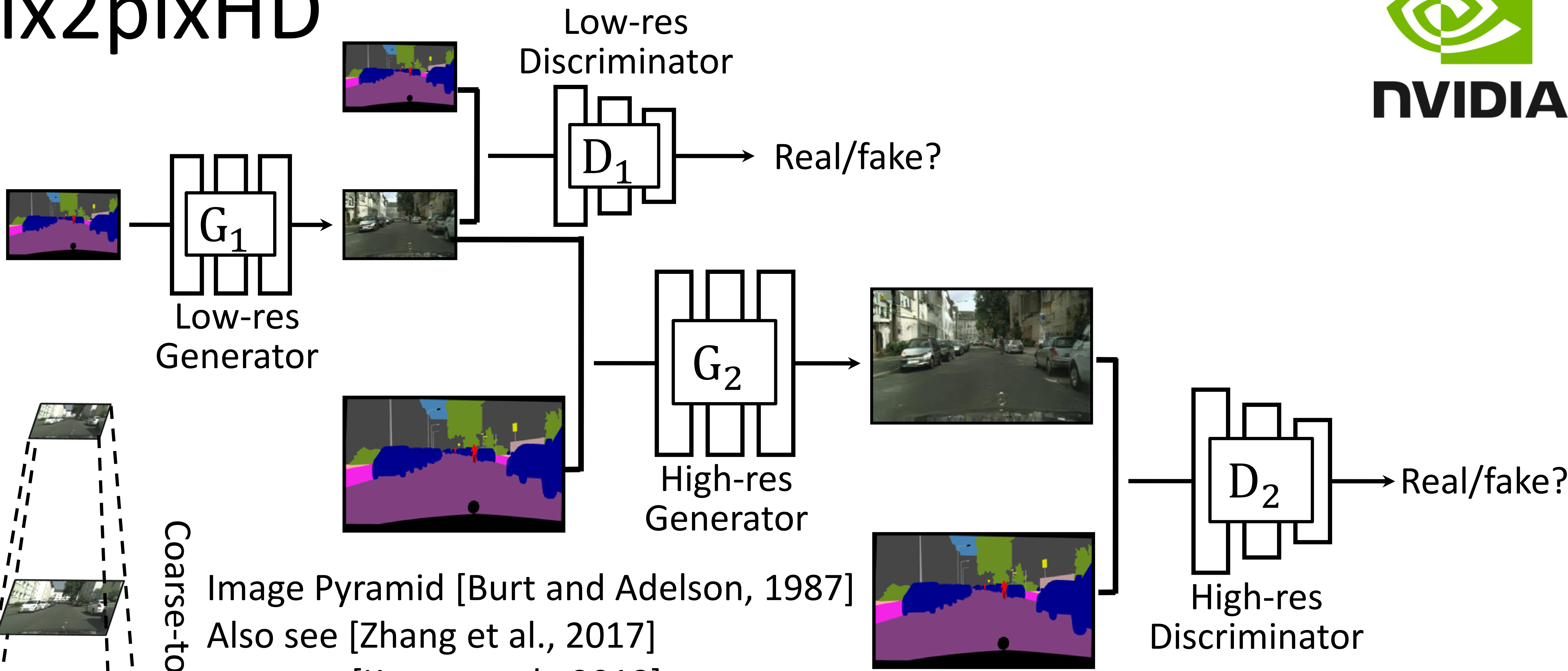


- Conditional on user inputs.
- Learning without pairs.
- High quality and resolution.

The Curse of Dimensionality

Pix2pix output

pix2pixHD



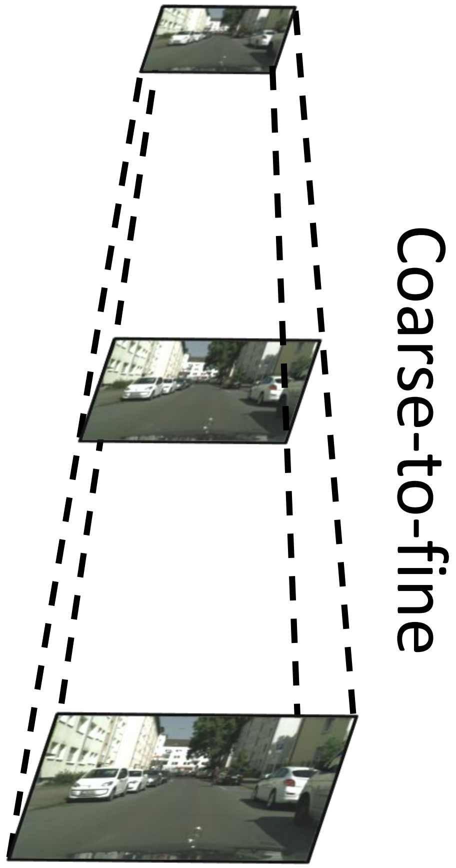


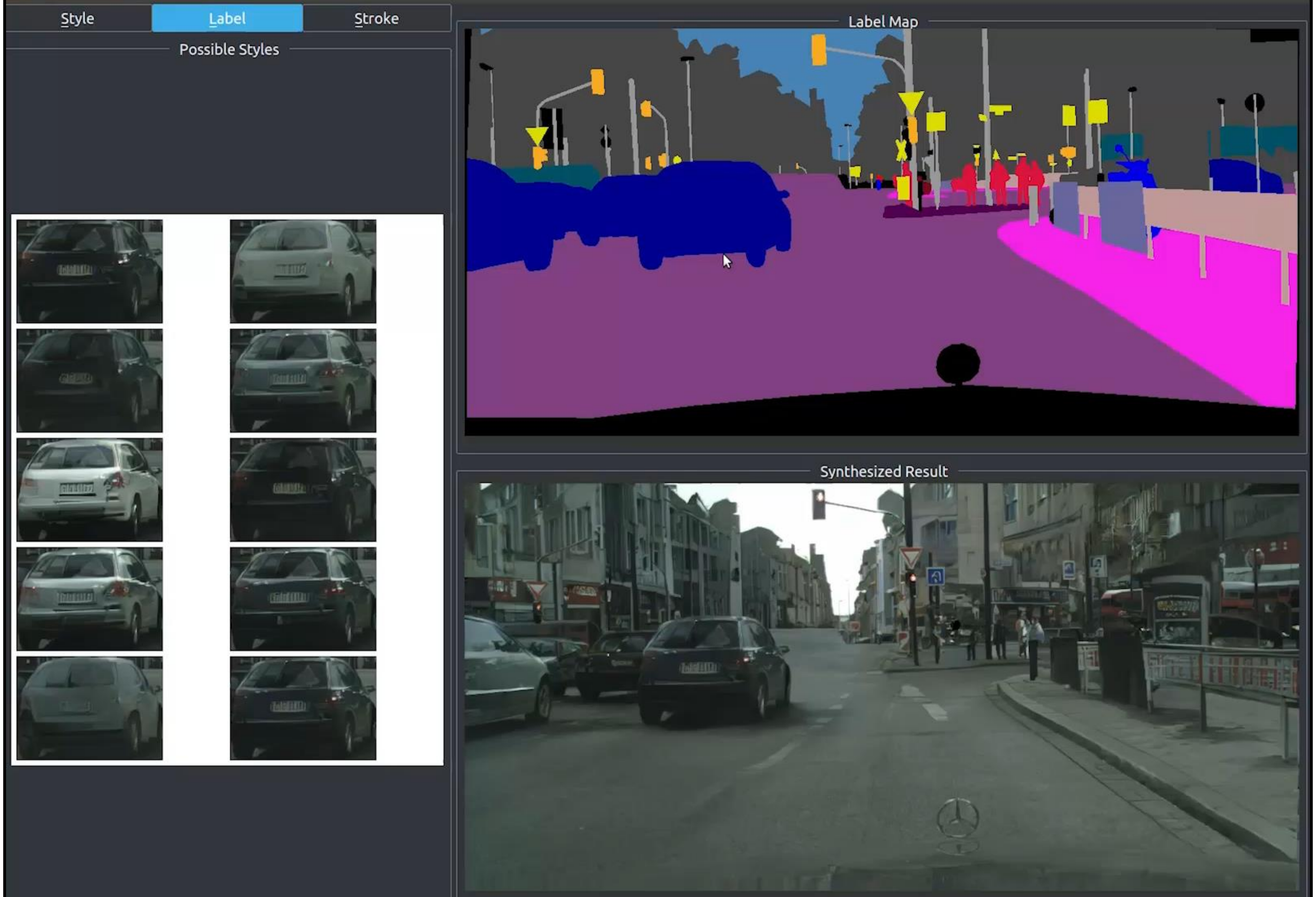
Image Pyramid [Burt and Adelson, 1987]

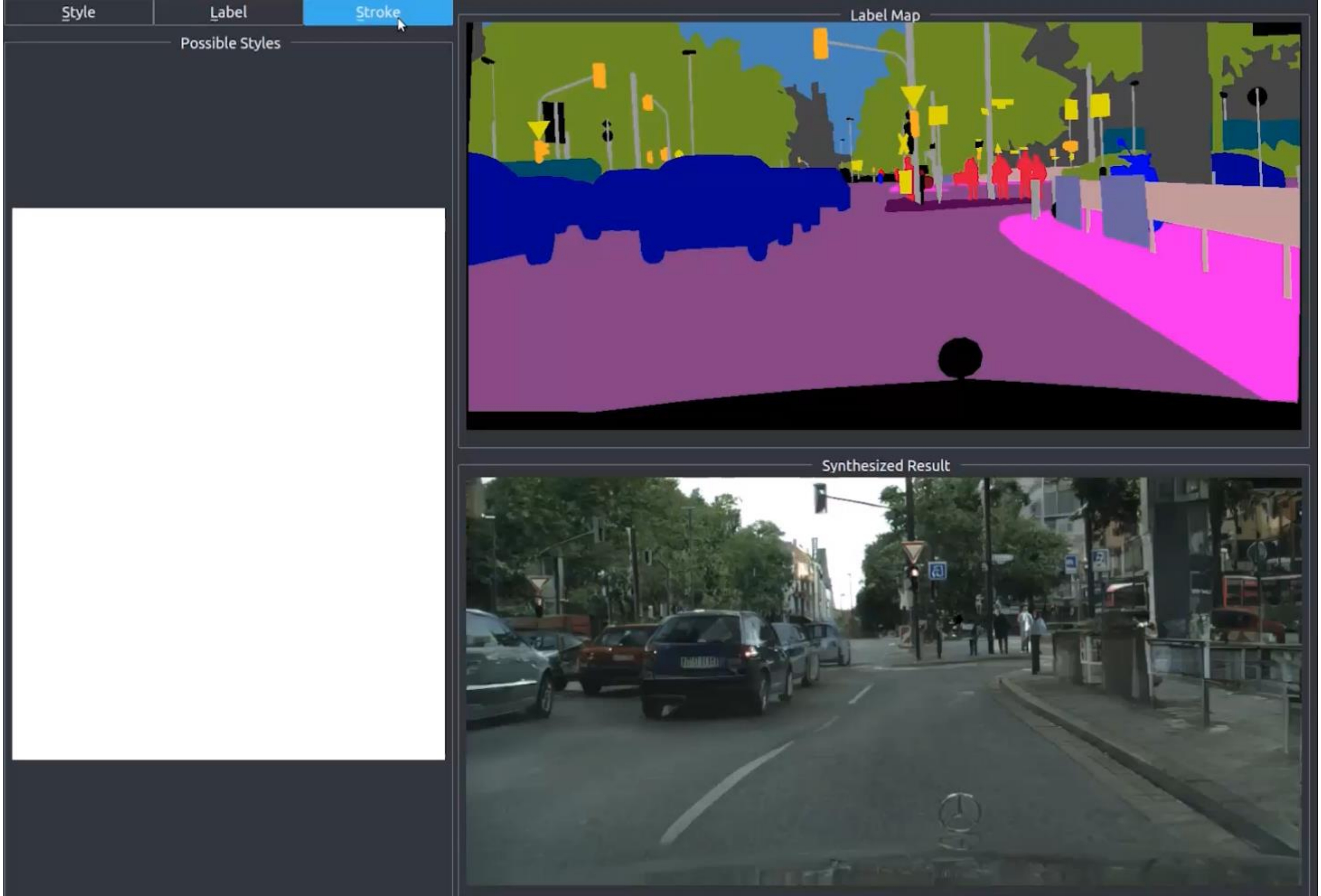
Also see [Zhang et al., 2017]

[Karras et al., 2018]

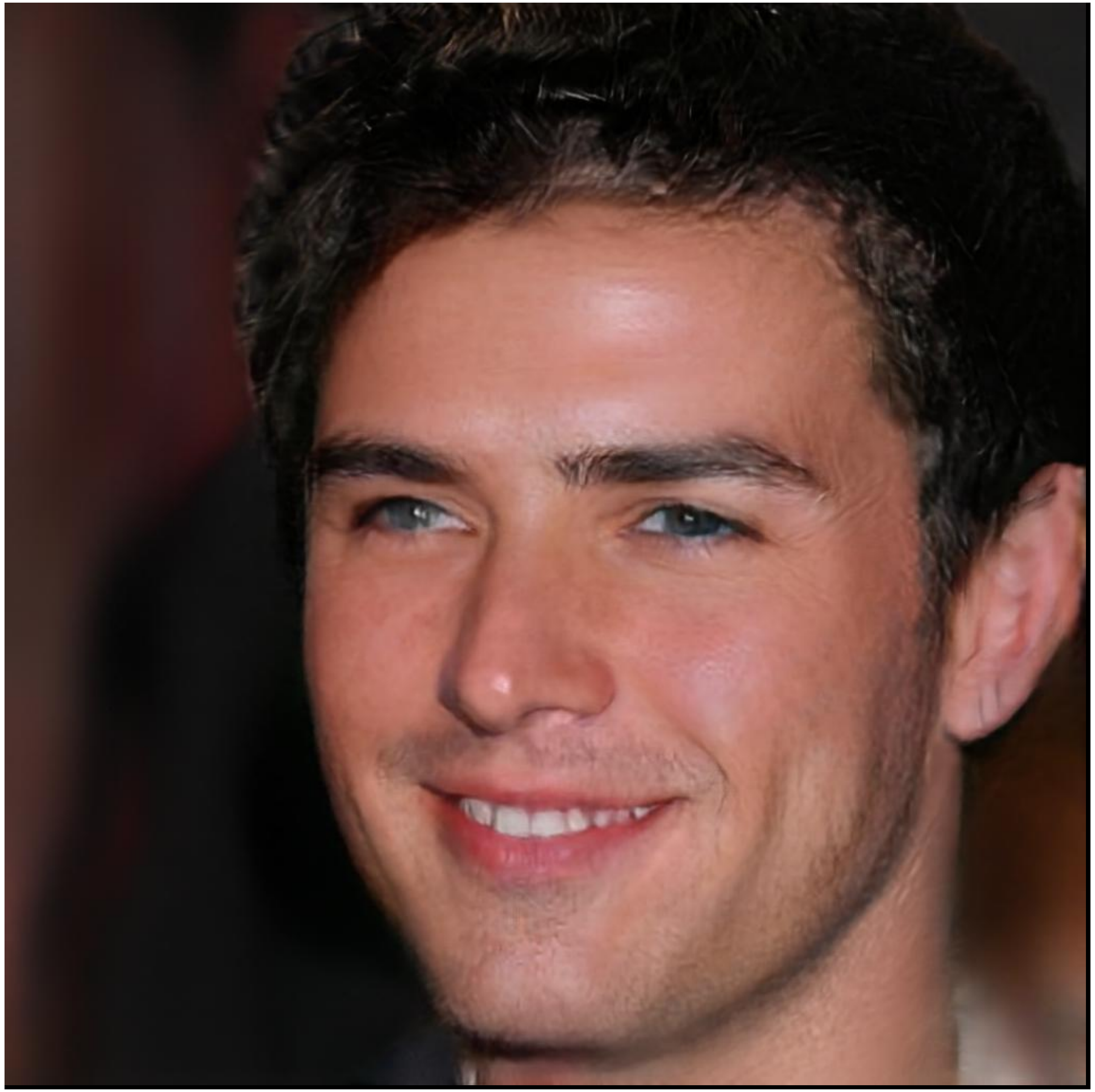
[Wang, Liu, Zhu, Tao, Kautz. Catanzaro, 2018]

pix2pixHD: 2048×1024


Style
Label
Stroke
Possible Styles
Label Map
Synthesized Result

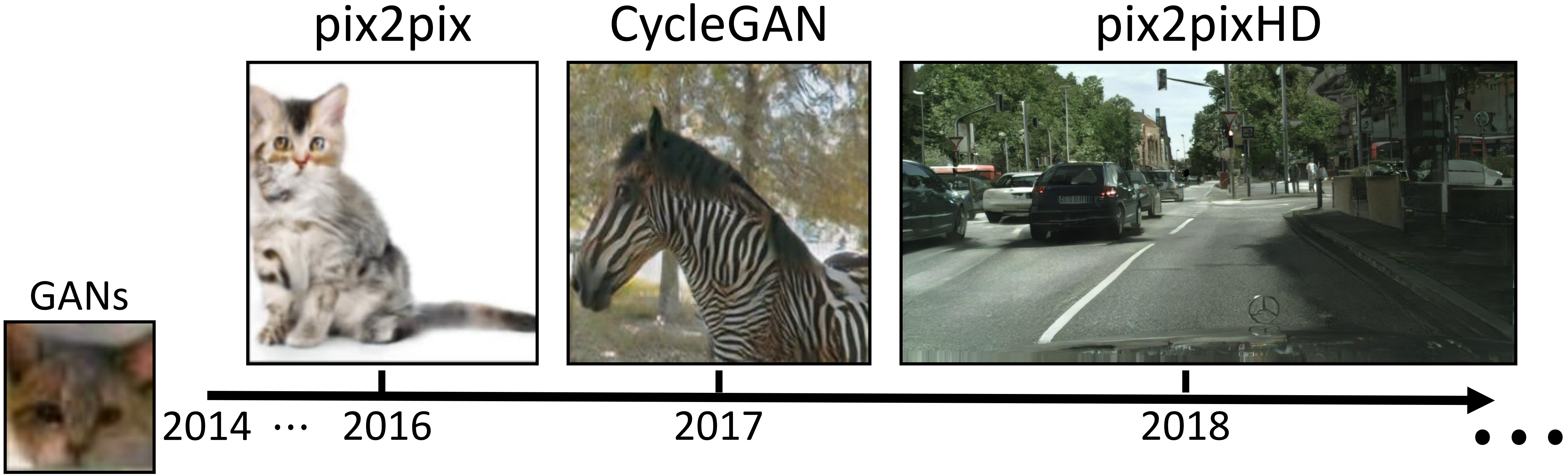

Style
Label
Stroke
Possible Styles
Label Map
Synthesized Result

pix2pixHD for sketch→ face

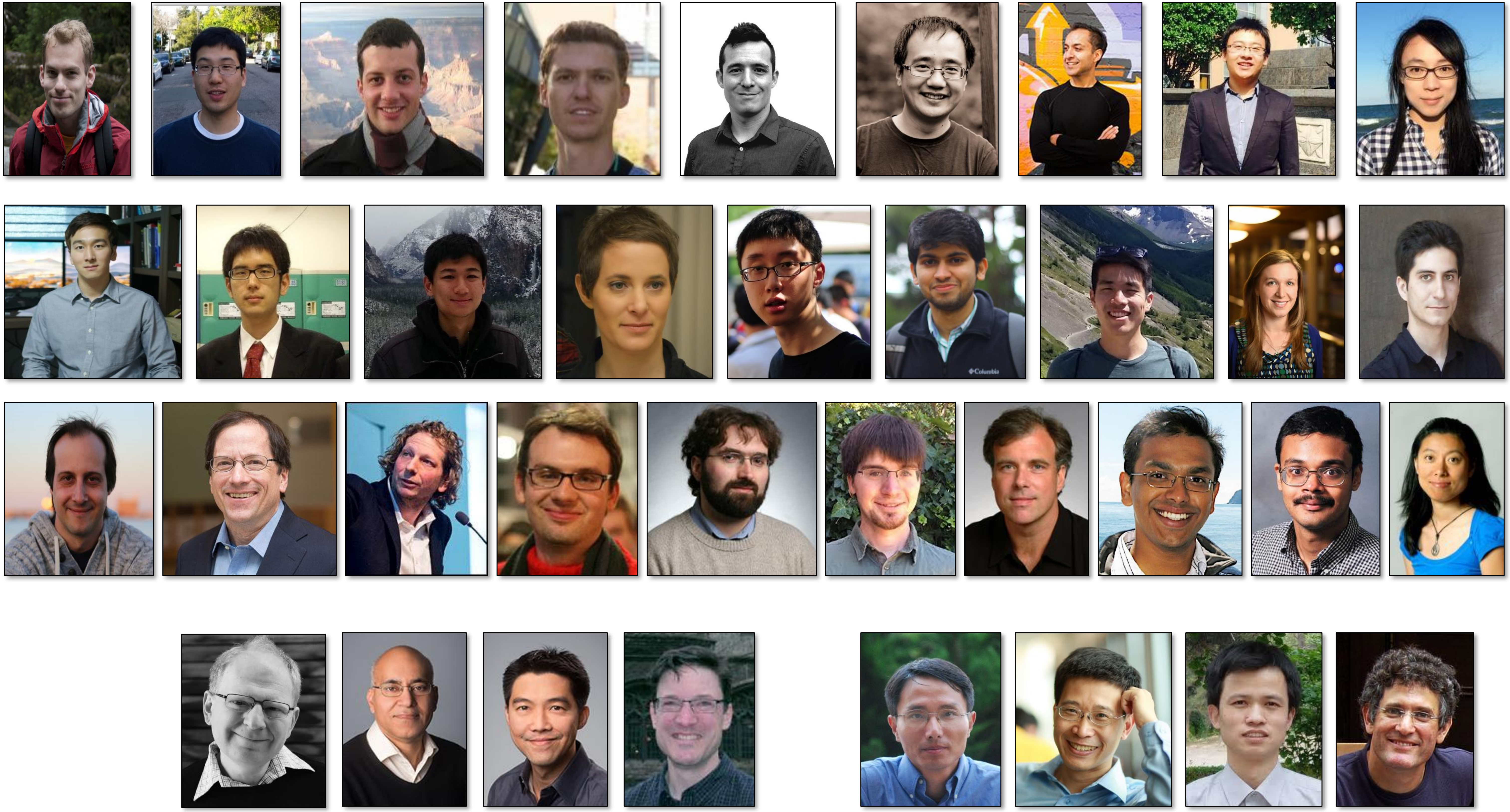


Improve Control, Quality and Resolution

Still a long way to go



- **Learning** to generate images from trillions of photos.
- Help more people tell their own visual **stories**.



Thank You!