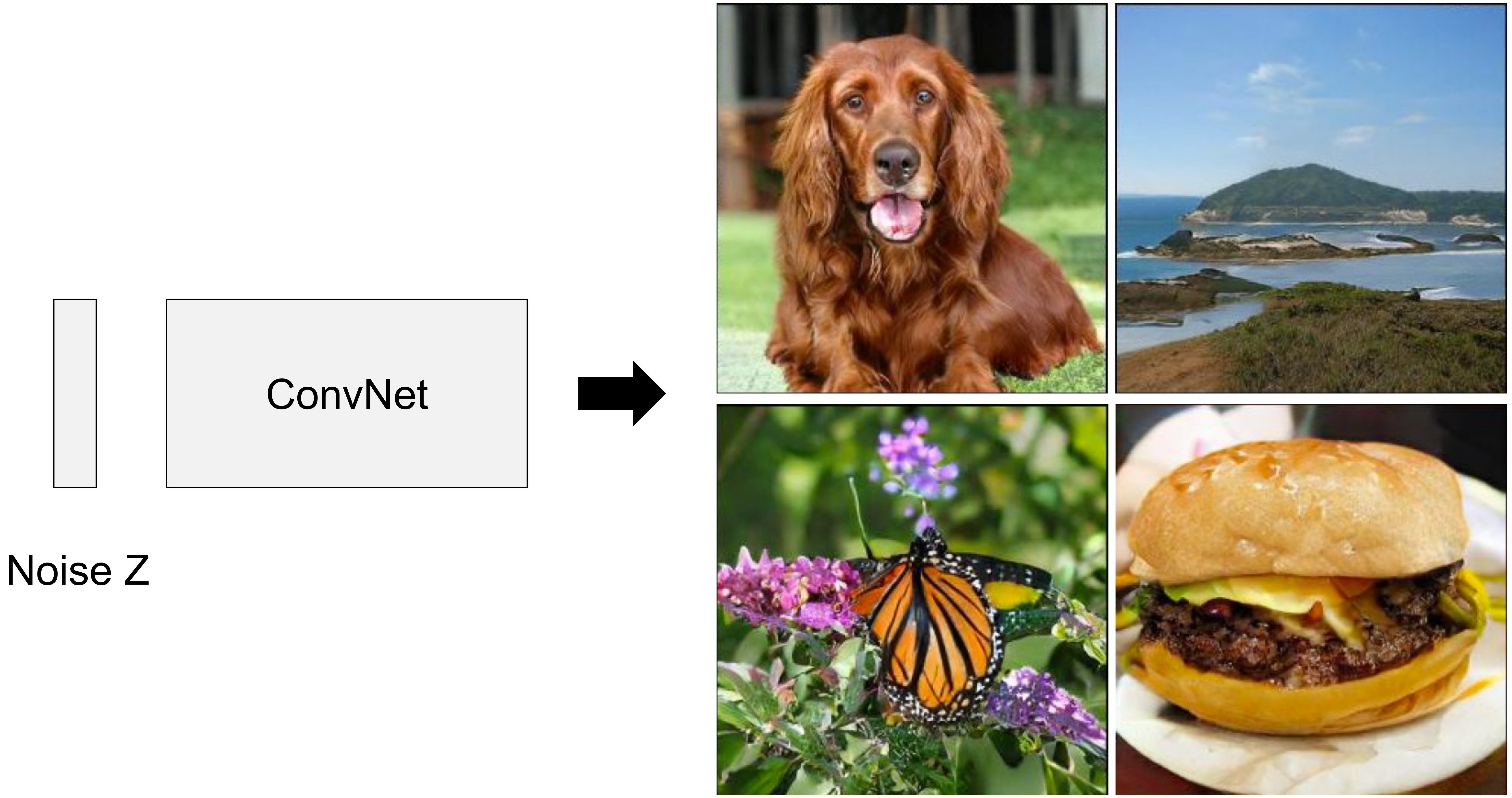
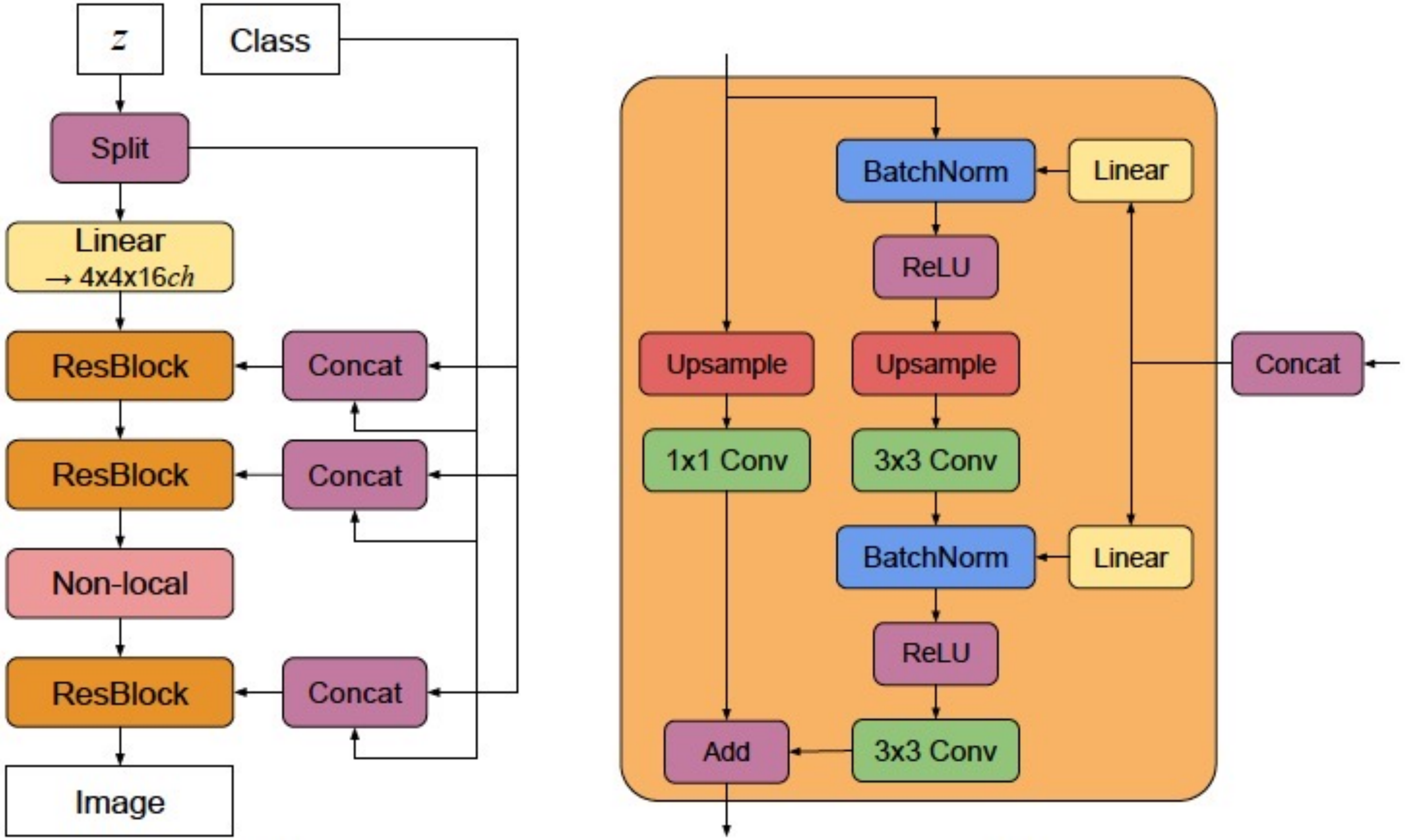
Conditional Generative Adversarial Networks

Xiaolong Wang

Last class



BigGAN: Class-Conditioned



(a) (b)

This Class

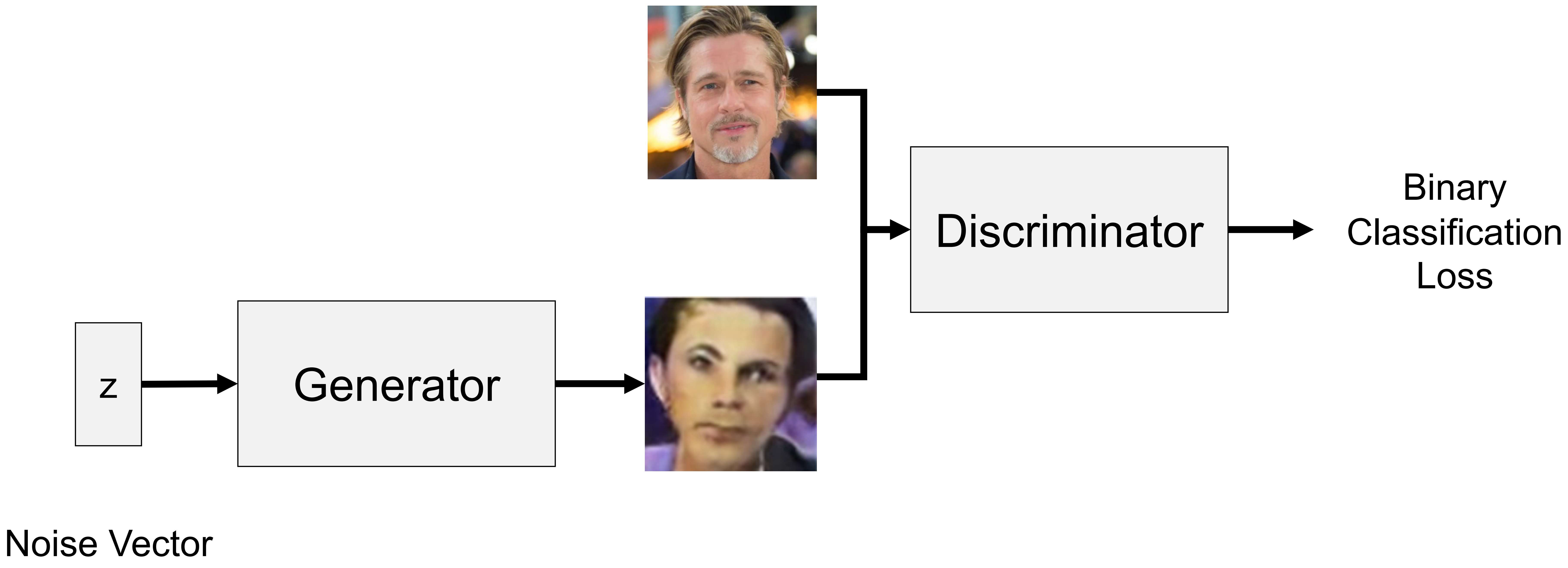
- Image-to-Image Translation: pix2pix

- Unpaired Image-to-Image Translation: CycleGAN

- Other Applications of Adversarial Learning

Image-to-Image Translation: pix2pix

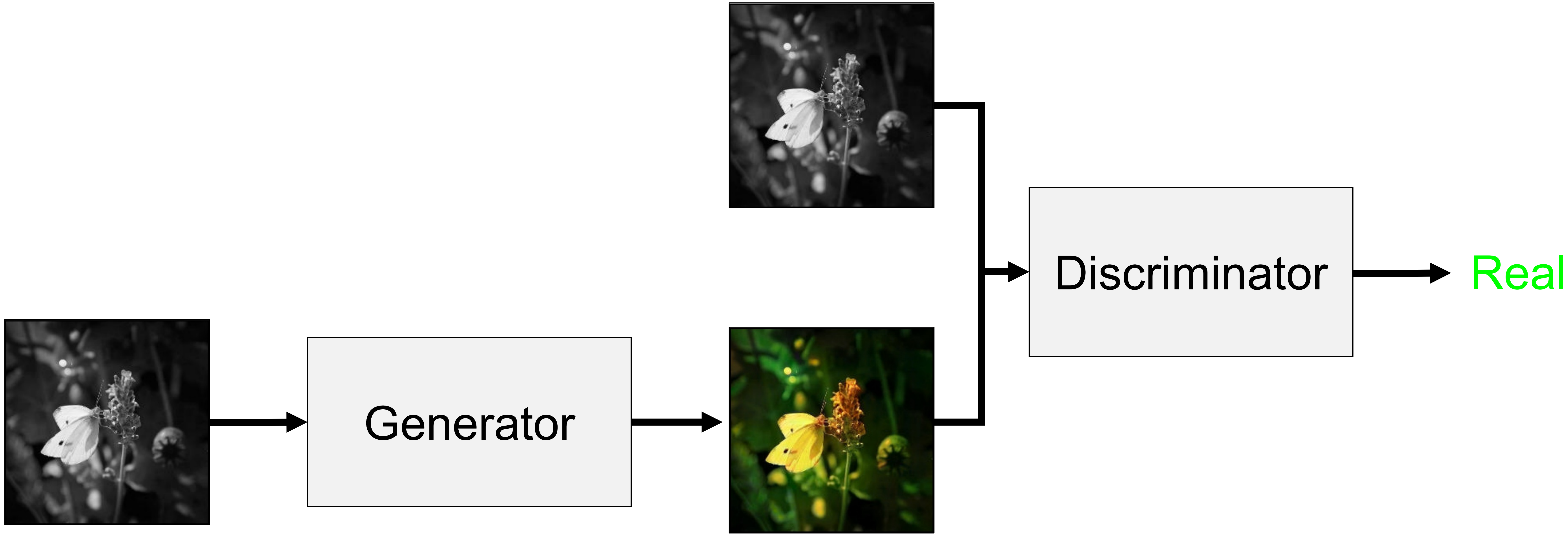
GANs



Goodfellow et al., 2014

Conditional GANs

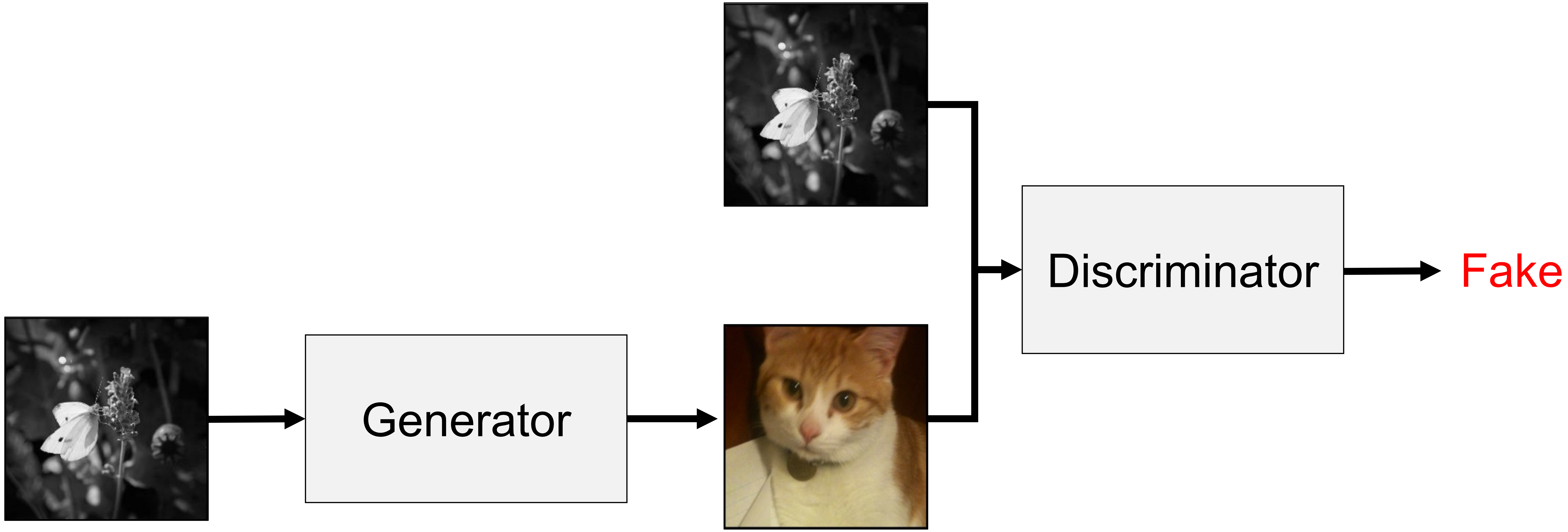
Conditional GANs



Generator takes an image as input, not noise.

Discriminator takes a pair of images as inputs, not just one image.

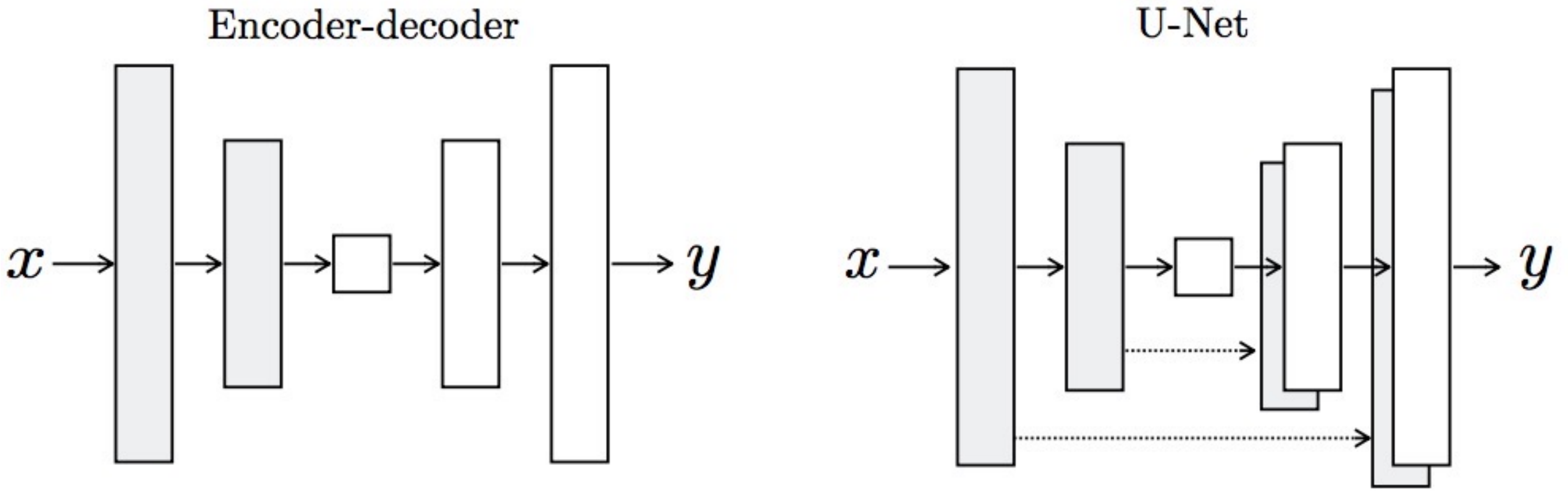
Conditional GANs



Generator takes an image as input, not noise.

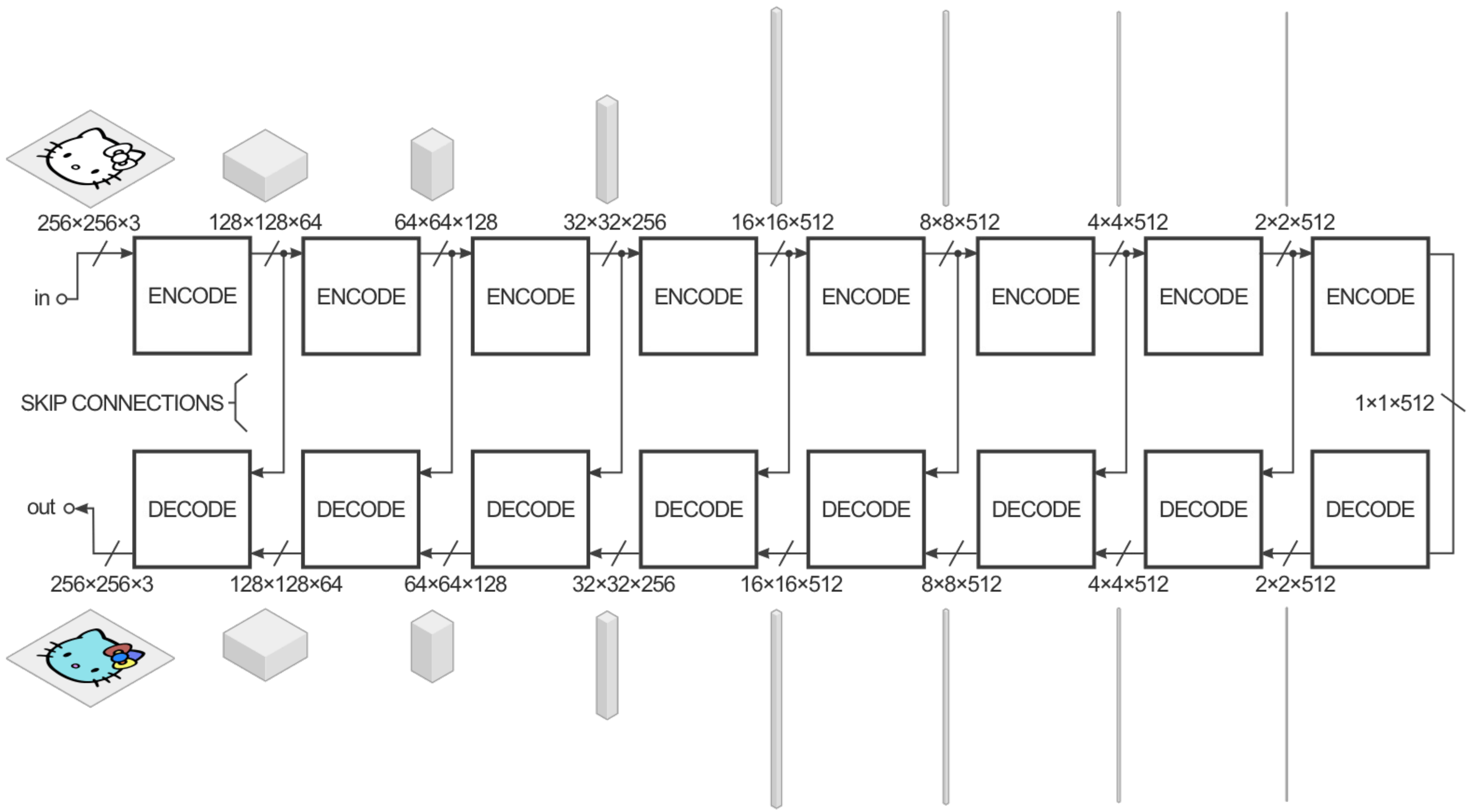
Discriminator takes a pair of images as inputs, not just one image.

Pix2Pix



Isola et al. Image-to-Image Translation with Conditional Adversarial Networks. 2017.

Image-to-image translation



Encode: convolution → BatchNorm → ReLU

Decode: transposed convolution → BatchNorm → ReLU

Image-to-image translation

Effect of adding skip connections to the generator

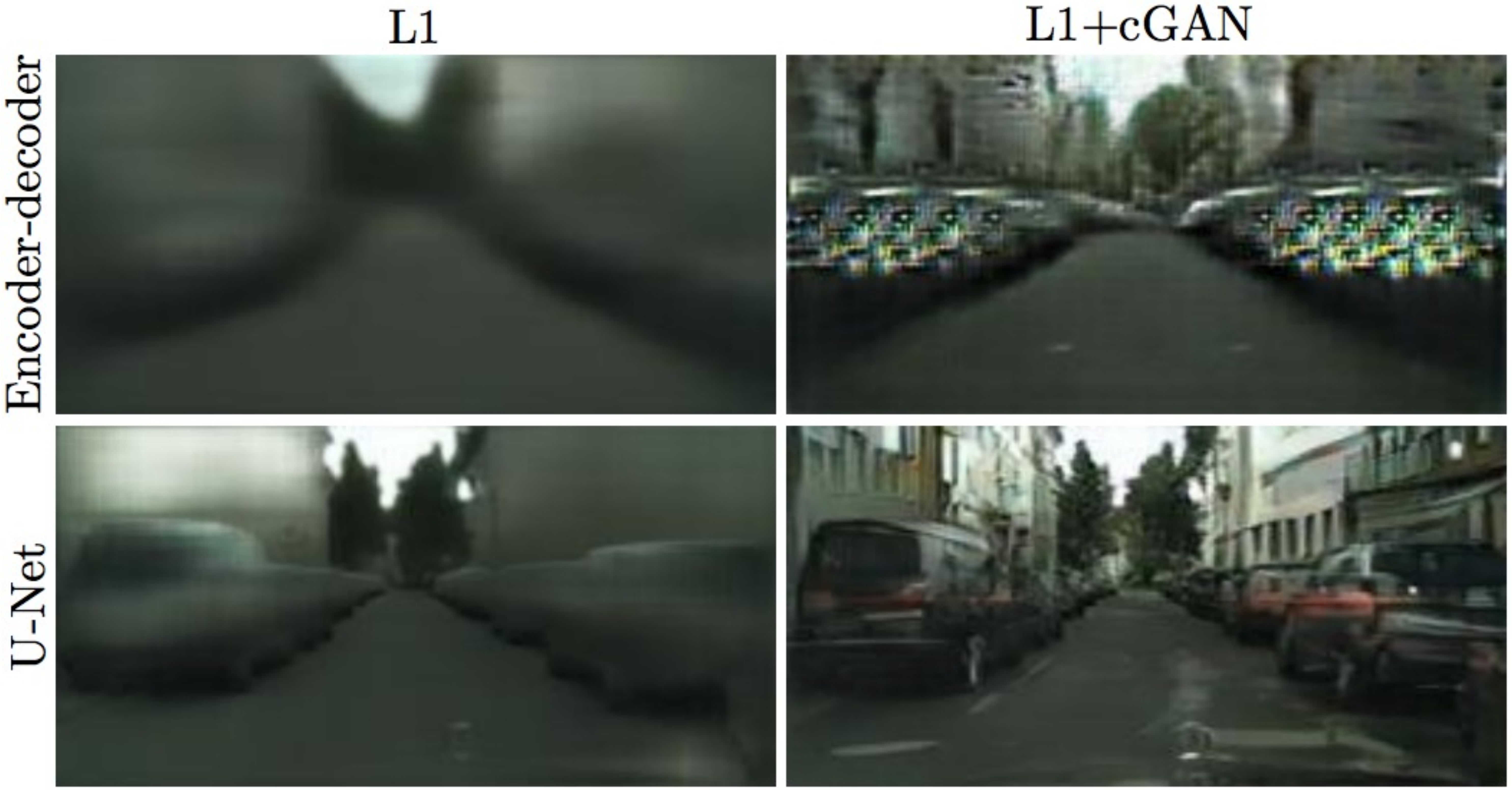


Image-to-image translation

- Generator loss: GAN loss plus L1 reconstruction penalty
- $G^* = \arg\min_G \max_D \mathcal{L}_{GAN}(G, D) + \lambda \sum_i \|y_i - G(x_i)\|_1$

Generated output $G(x_i)$ should be close to ground truth target y_i

Image-to-image translation

- Generator loss: GAN loss plus L1 reconstruction penalty

$$G^* = \arg\min_G \max_D \mathcal{L}_{GAN}(G, D) + \lambda \sum \|y_i - G(x_i)\|_1$$

Image-to-image translation: Results

- Day to night

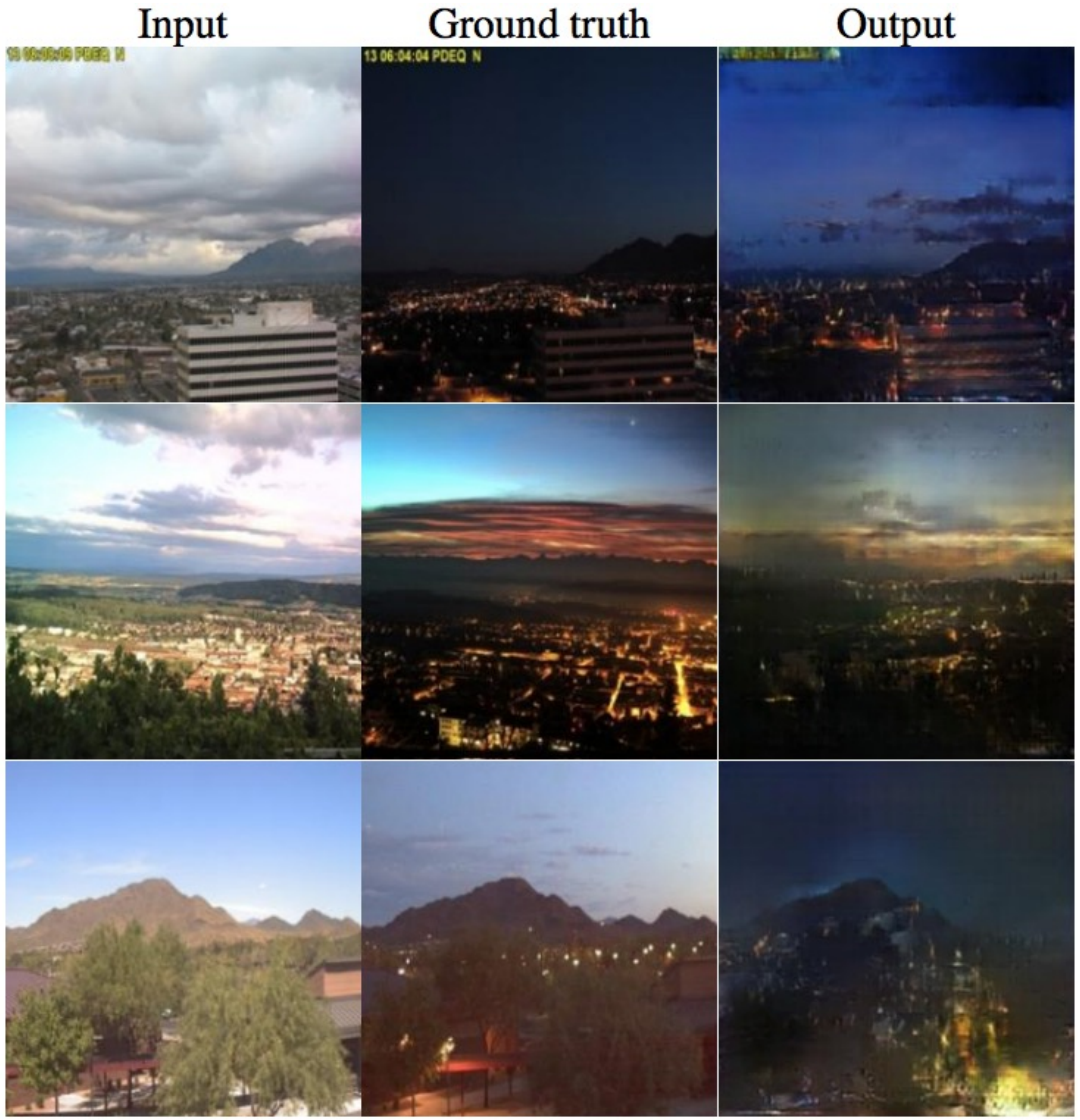
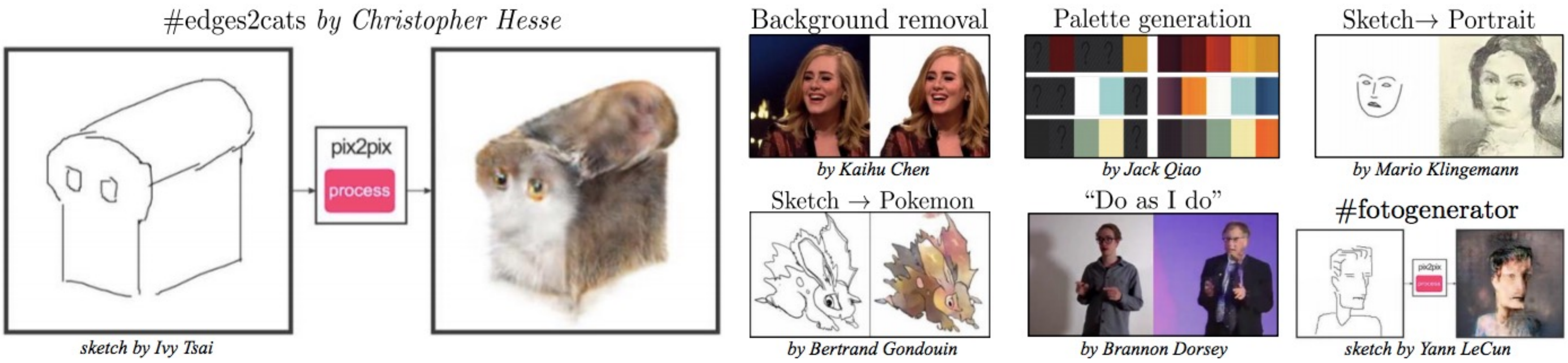


Image-to-image translation: Results

- Edges

Image-to-image translation: Results

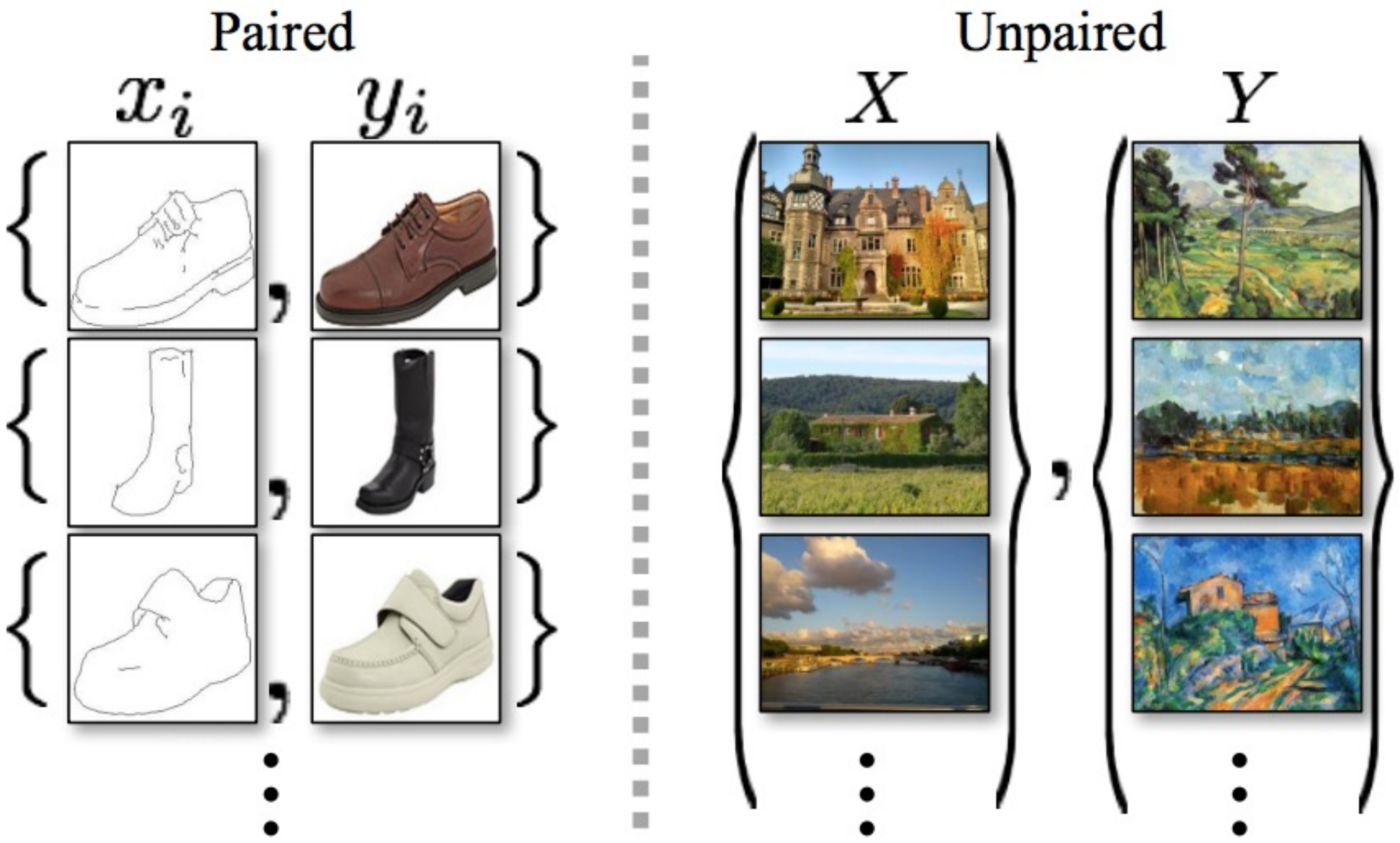
- pix2pix demo



Unpaired Image-to-Image Translation: CycleGAN

Unpaired image-to-image translation

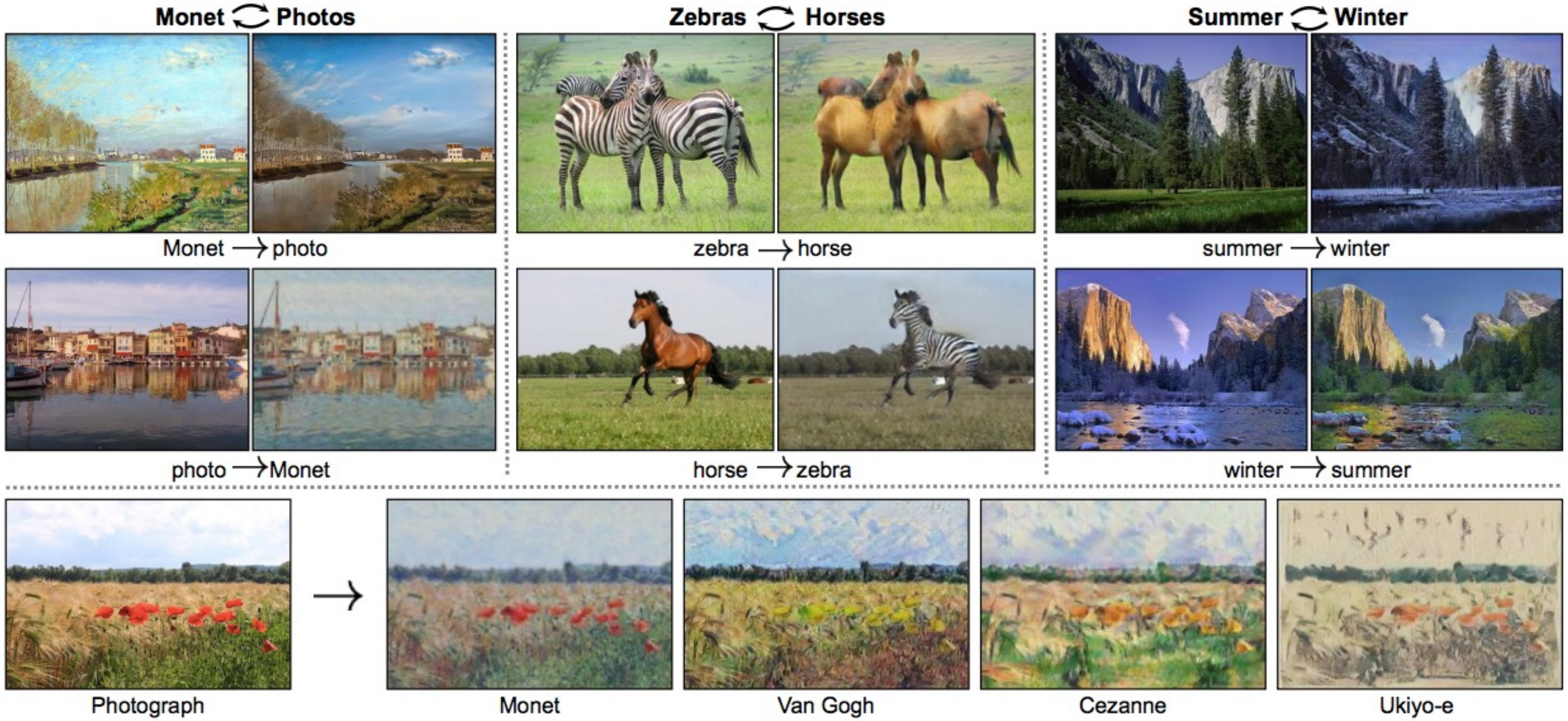
- Given two unordered image collections X and Y, learn to "translate" an image from one into the other and vice versa



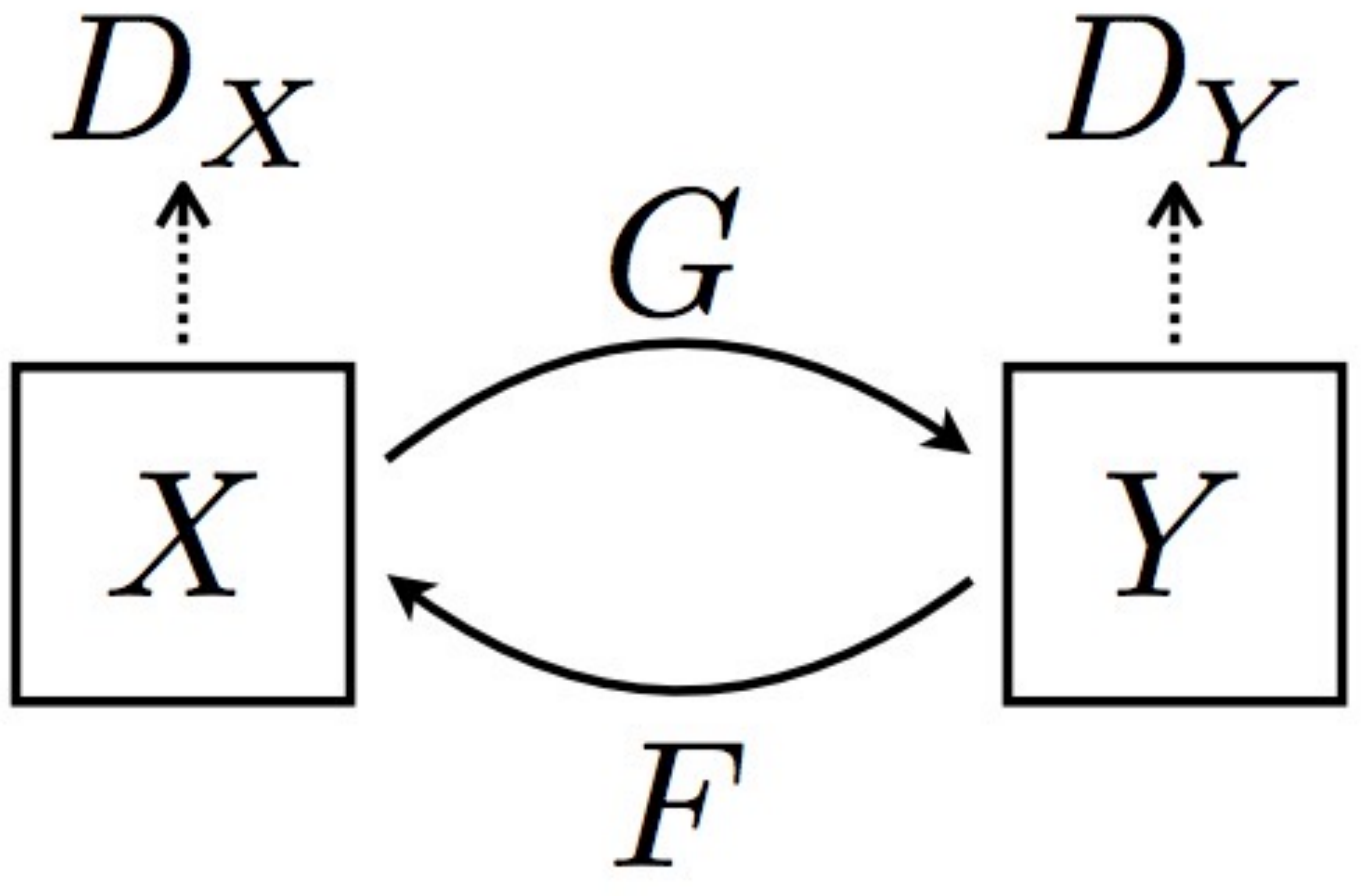
Zhu et al., 2017

Unpaired image-to-image translation

- Given two unordered image collections X and Y, learn to "translate" an image from one into the other and vice versa



CycleGAN



(a)

CycleGAN: Loss

- Requirements:
  - G translates from X to Y, F translates from Y to X
  - D_X recognizes images from X, D_Y from Y
  - We want $F(G(x)) \approx x$ and $G(F(y)) \approx y$
- CycleGAN discriminator loss: LSGAN

$$\mathcal{L}_{\text{GAN}}(D_Y) = \mathbb{E}_{y\sim p_{\text{data}}(y)}[(D_Y(y)-1)^2] + \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[D_Y(G(x))^2\right]$$

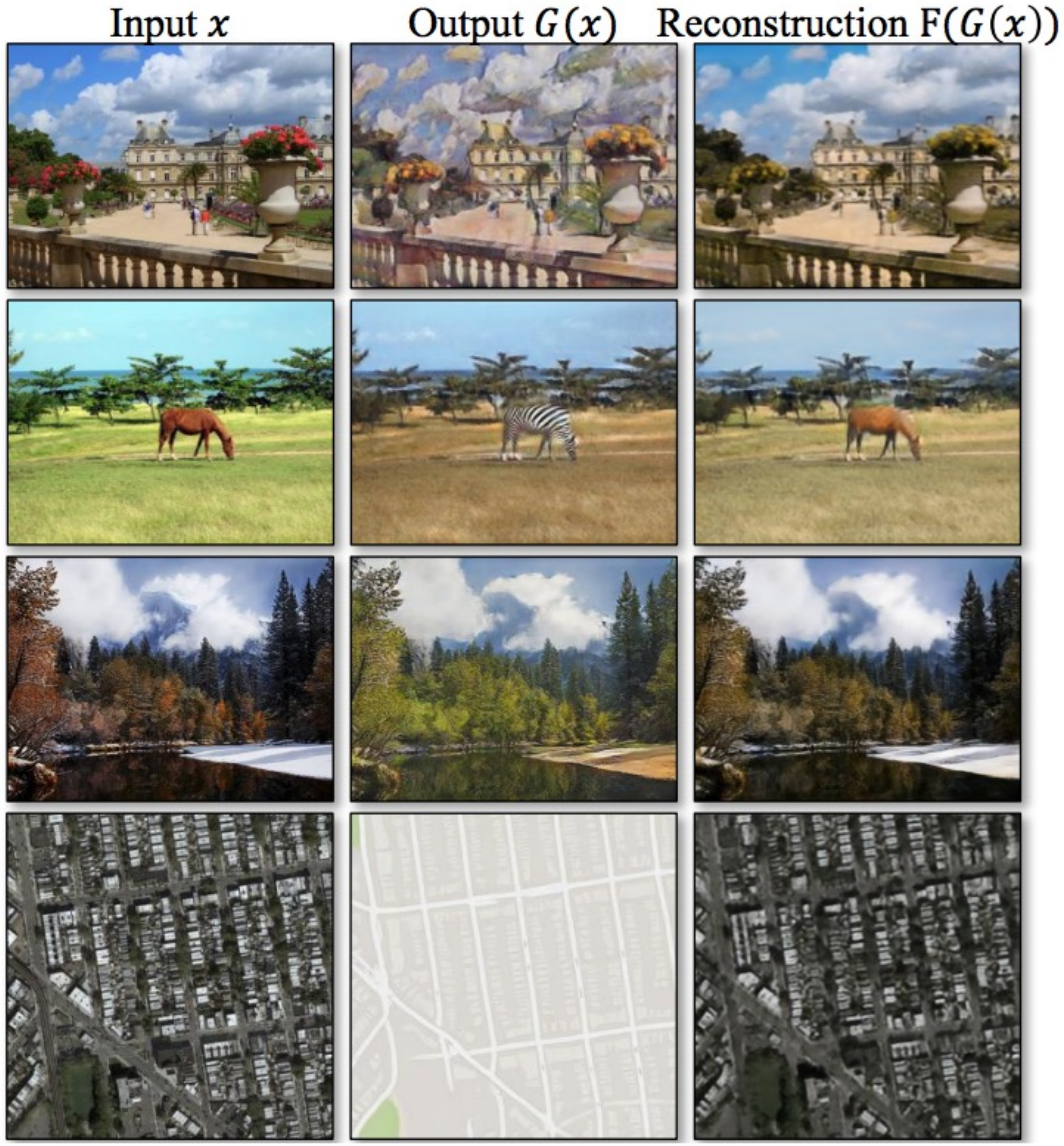
$$\mathcal{L}_{\text{GAN}}(D_X) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[(D_X(x)-1)^2] + \mathbb{E}_{y\sim p_{\text{data}}(y)}\left[D_X(F(y))^2\right]$$

- CycleGAN generator loss:

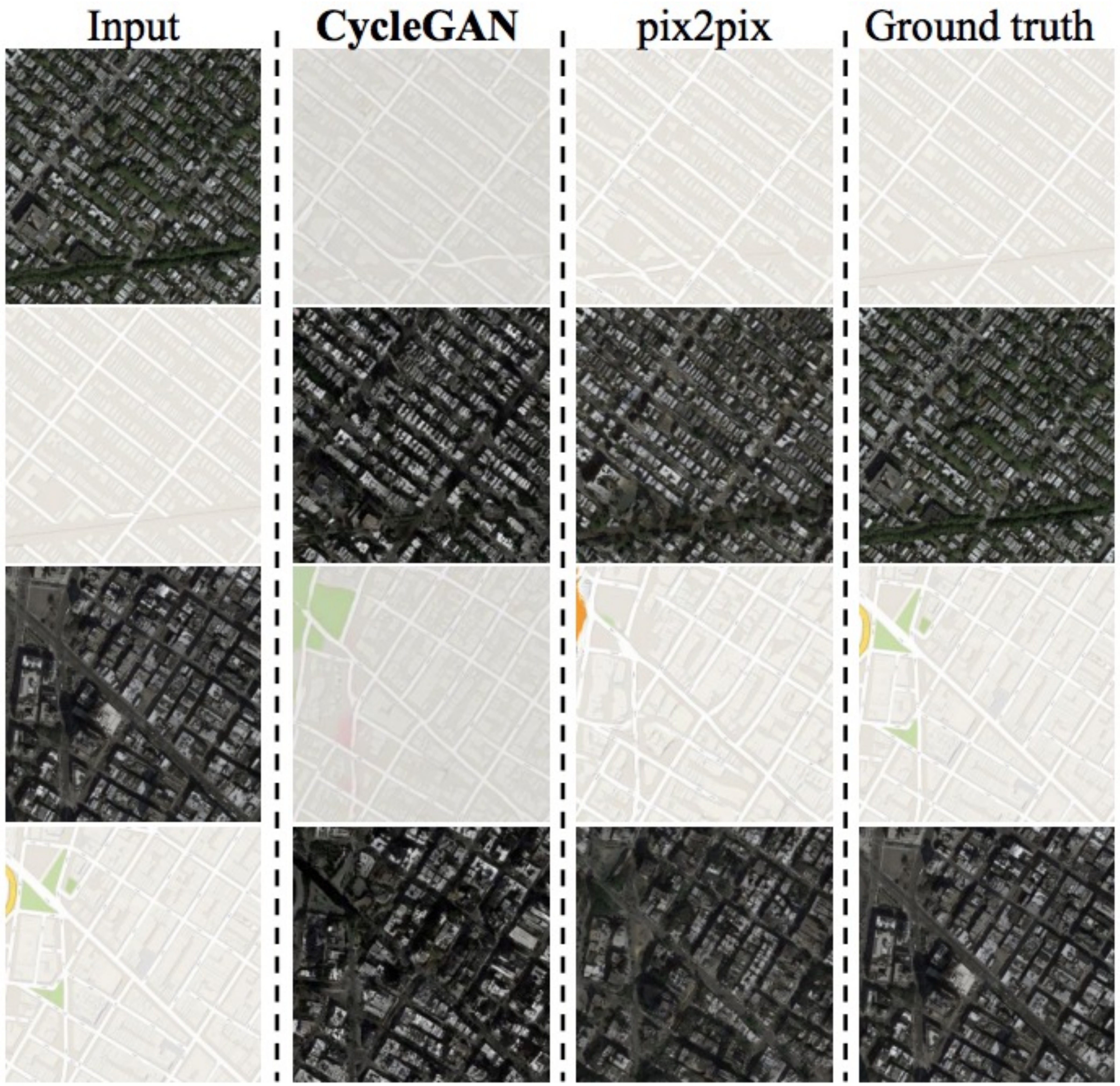
$$\mathcal{L}_{\text{cyc}}(G,F) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[D_Y(G(x)-1)^2] + \mathbb{E}_{y\sim p_{\text{data}}(y)}[D_X(F(y)-1)^2]$$

$$+\,\mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\left\|F(G(x))-x\right\|_1\right] + \mathbb{E}_{y\sim p_{\text{data}}(y)}\left[\left\|G(F(y))-y\right\|_1\right]$$

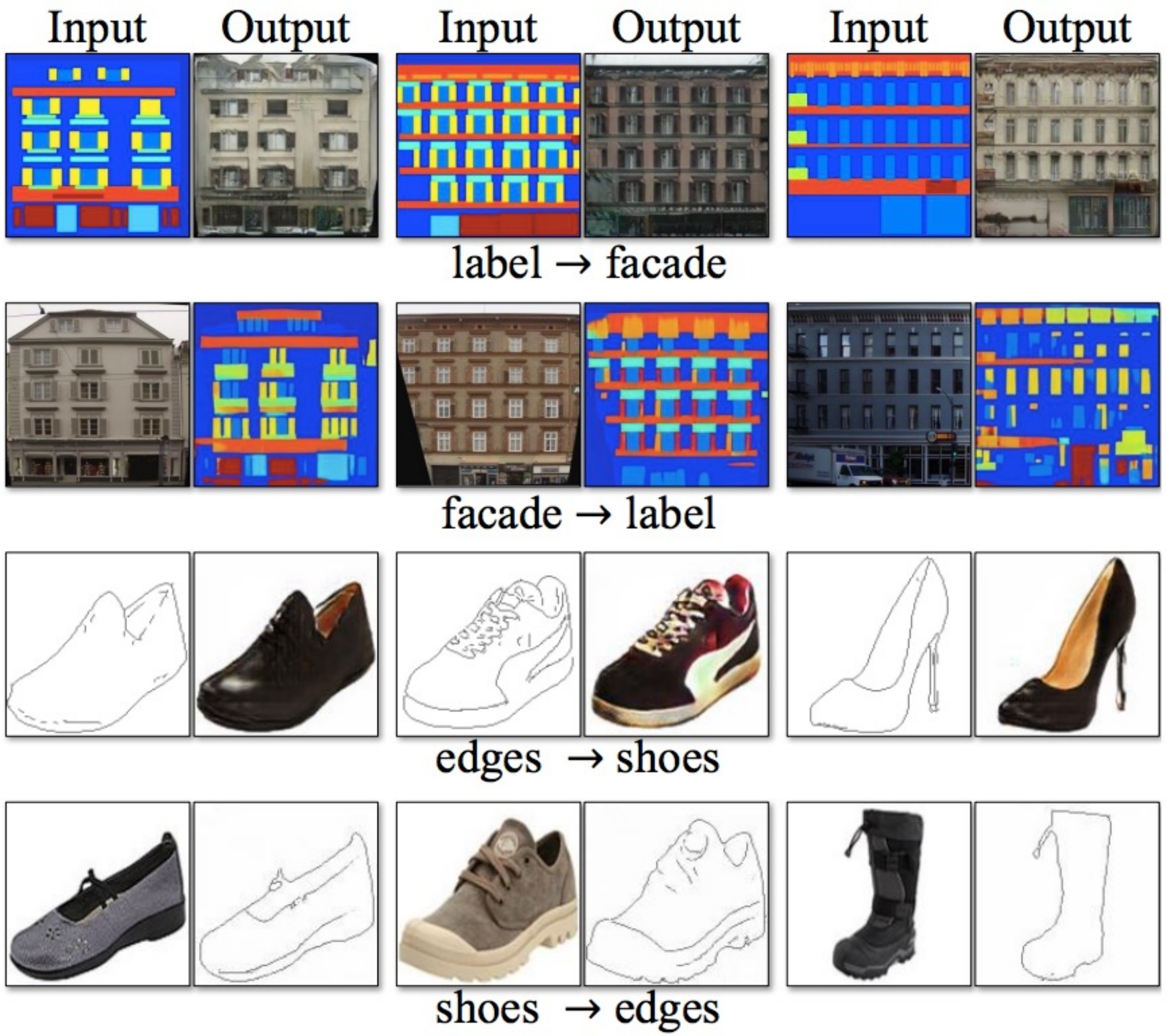
CycleGAN



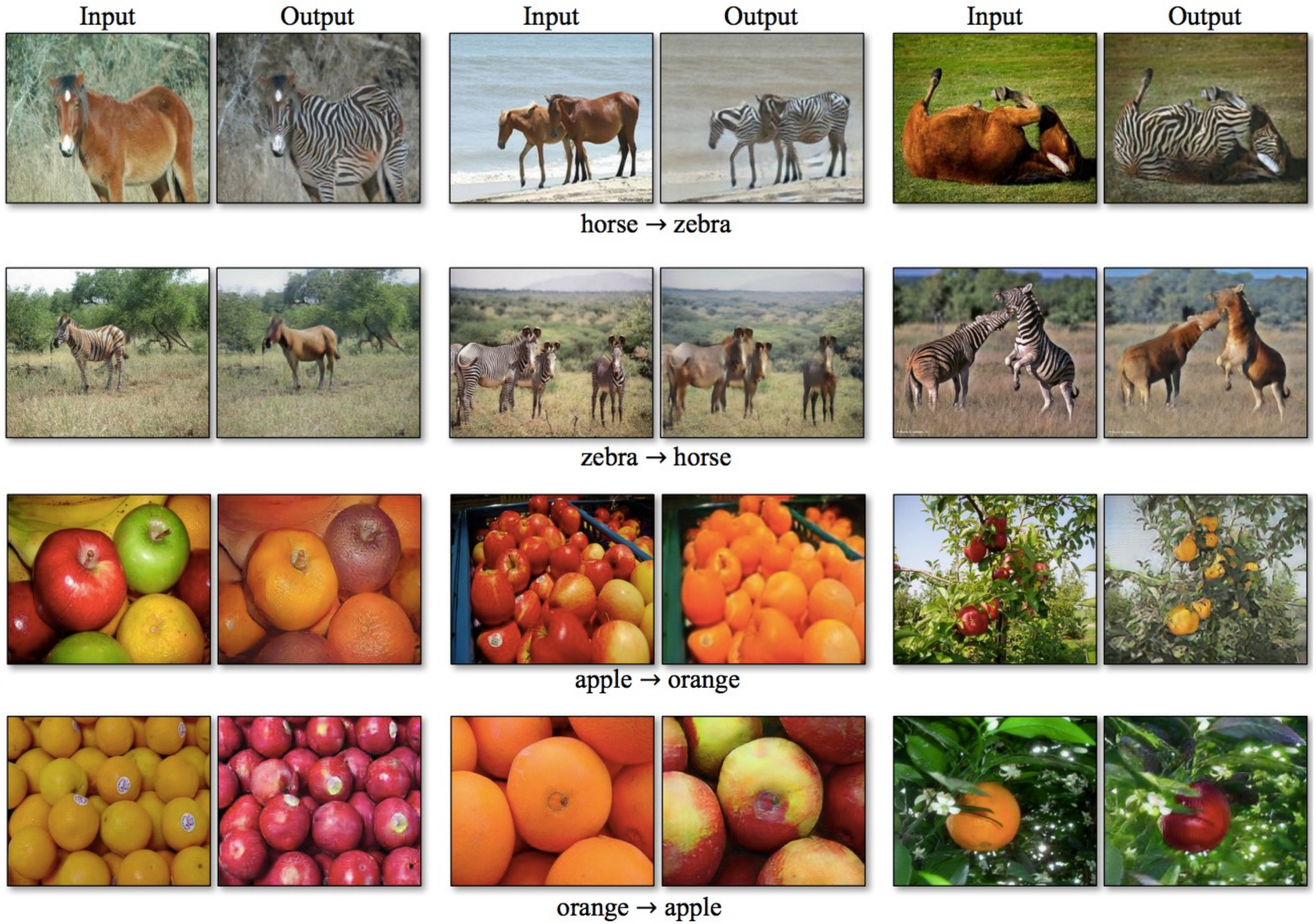
CycleGAN: Results



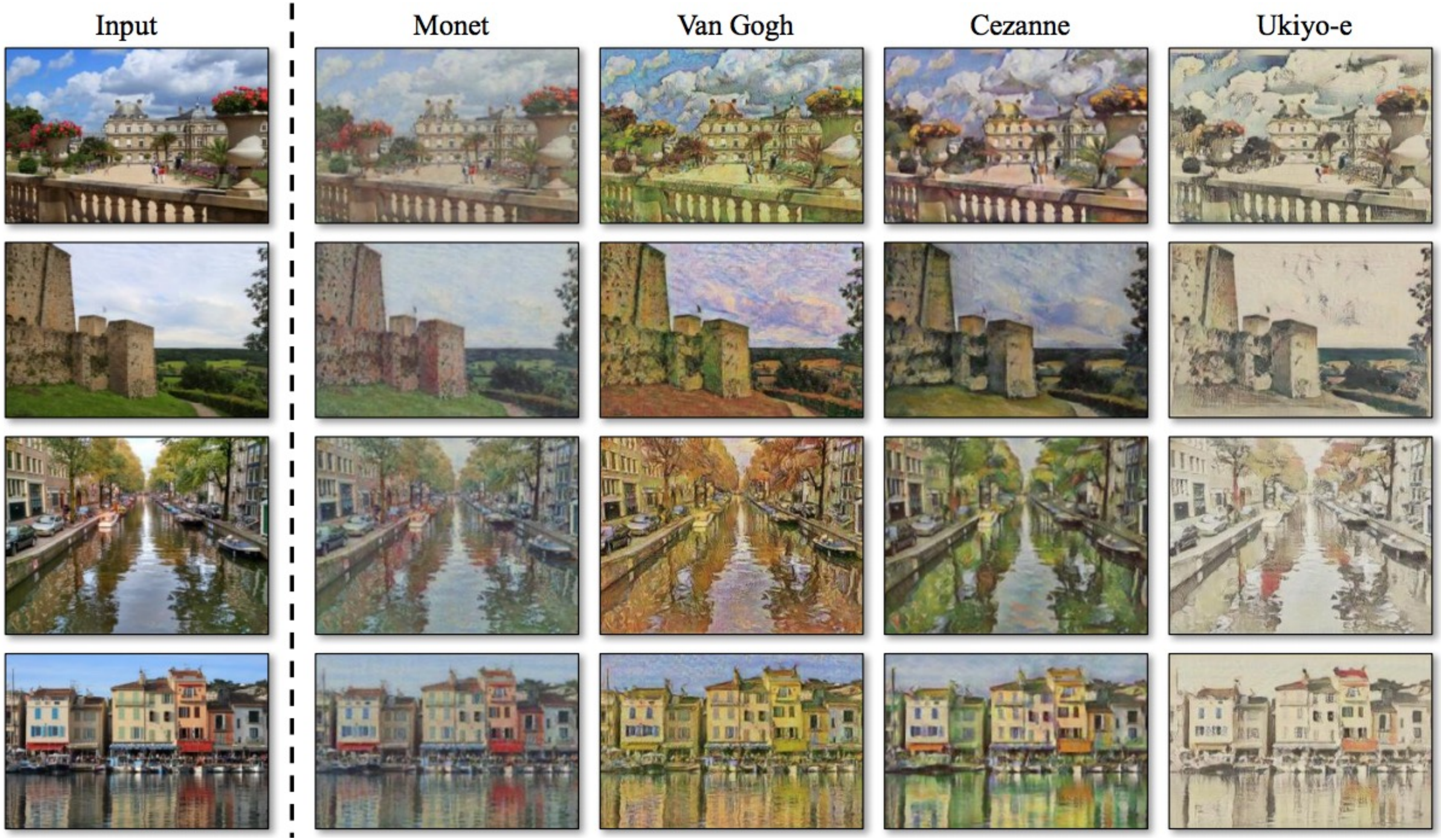
CycleGAN: Results



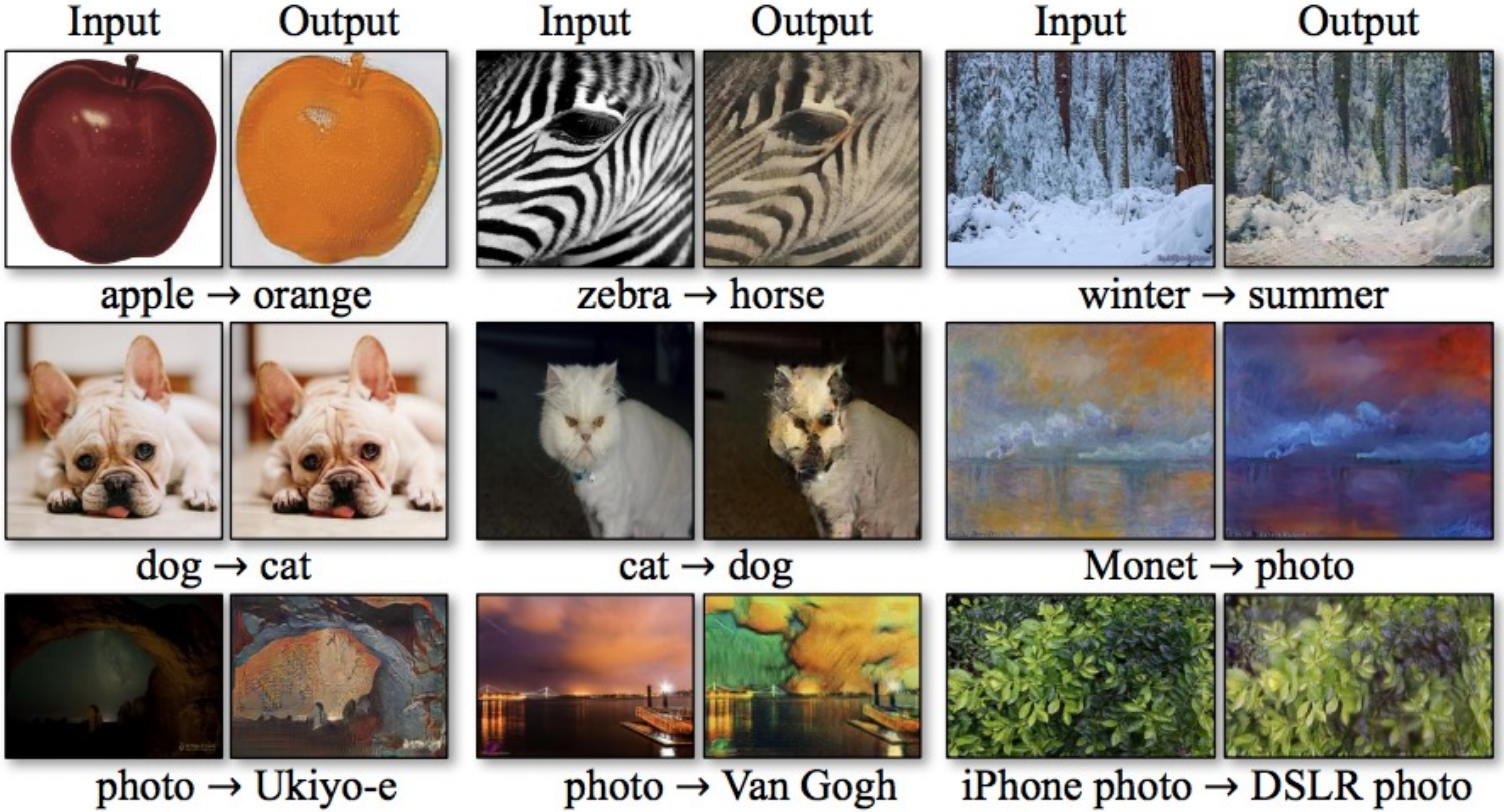
CycleGAN: Results



CycleGAN: Results



CycleGAN: Failure cases



CycleGAN: Failure cases

Input Output

horse → zebra

CycleGAN: Limitations

- Cannot handle shape changes (e.g., dog to cat)
- Can get confused on images outside of the training domains (e.g., horse with rider)
- Cannot close the gap with paired translation methods
- Does not account for the fact that one transformation direction may be more challenging than the other

High-resolution, high-quality pix2pix

Cascaded refinement network [5]

Our result

(a) Synthesized result

(b) Application: Change label types

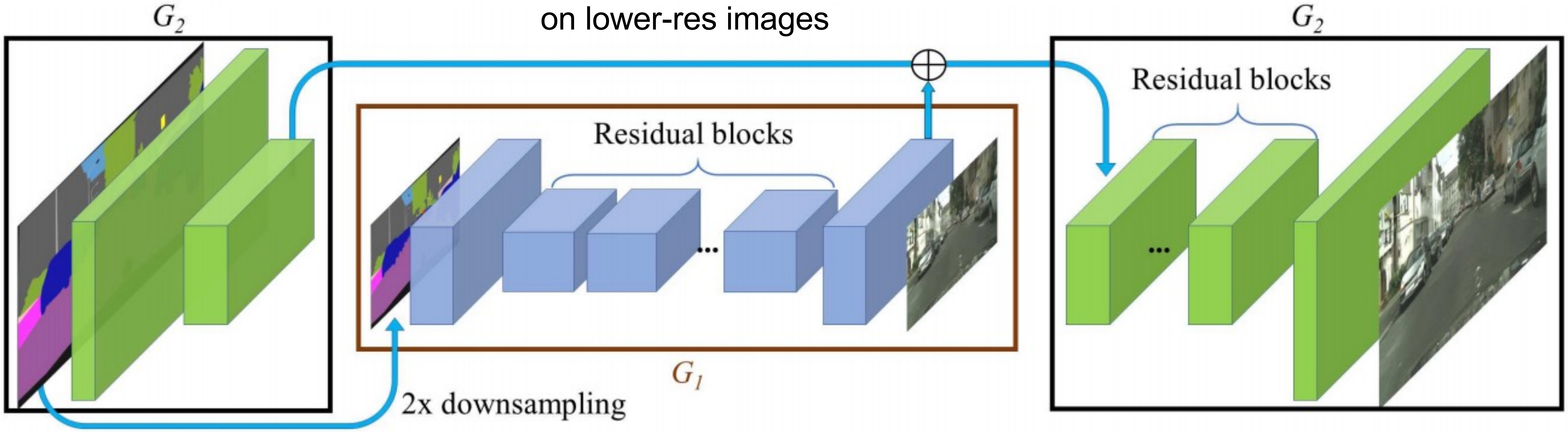
(c) Application: Edit object appearance

T.-C. Wang et al., [High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs](), CVPR 2018

High-resolution, high-quality pix2pix

- Two-scale generator architecture (up to 2048 x 1024 resolution)

First train *global generator* network (G1)
on lower-res images



Then append higher-res *enhancer network* (G2) blocks and train G1 and G2 jointly

T.-C. Wang et al., [High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs](), CVPR 2018

Human generation conditioned on pose

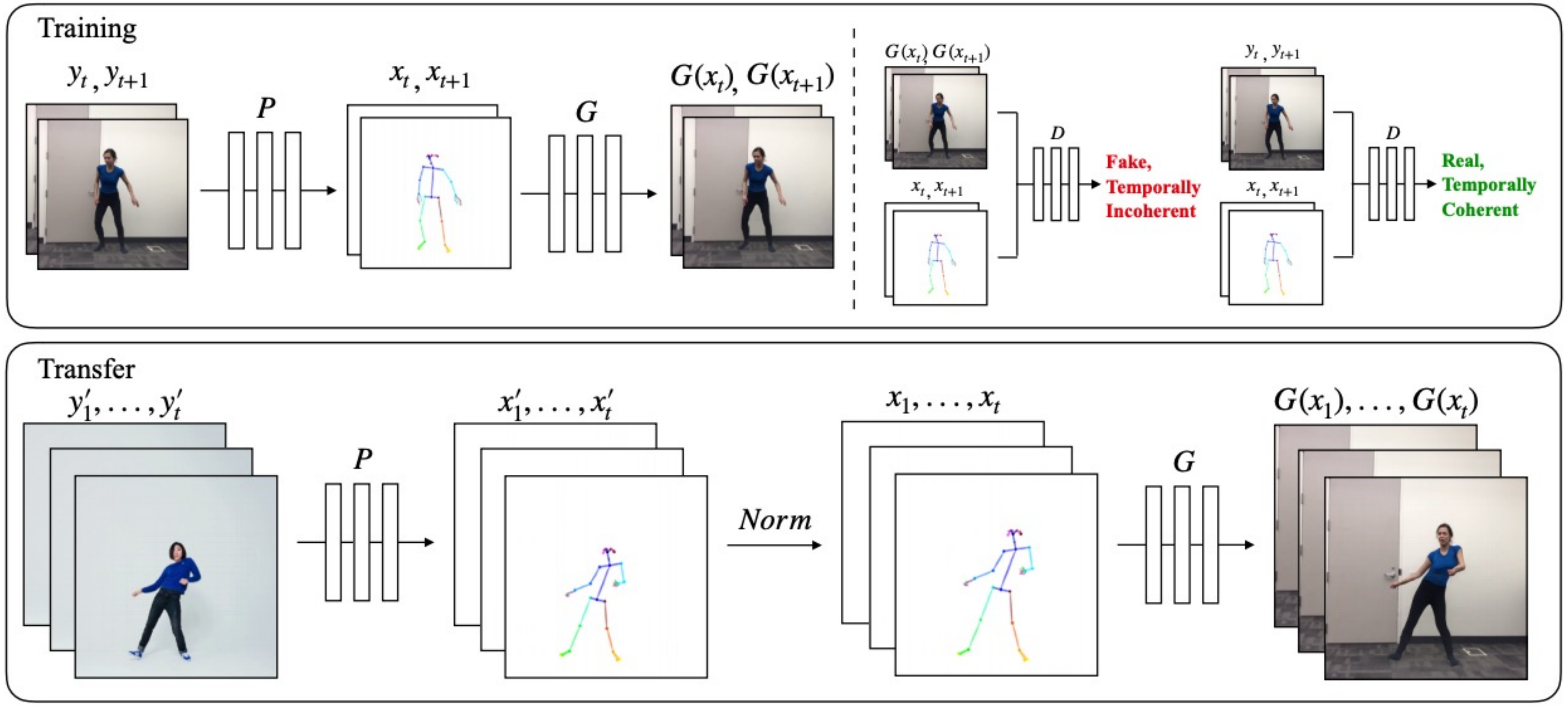
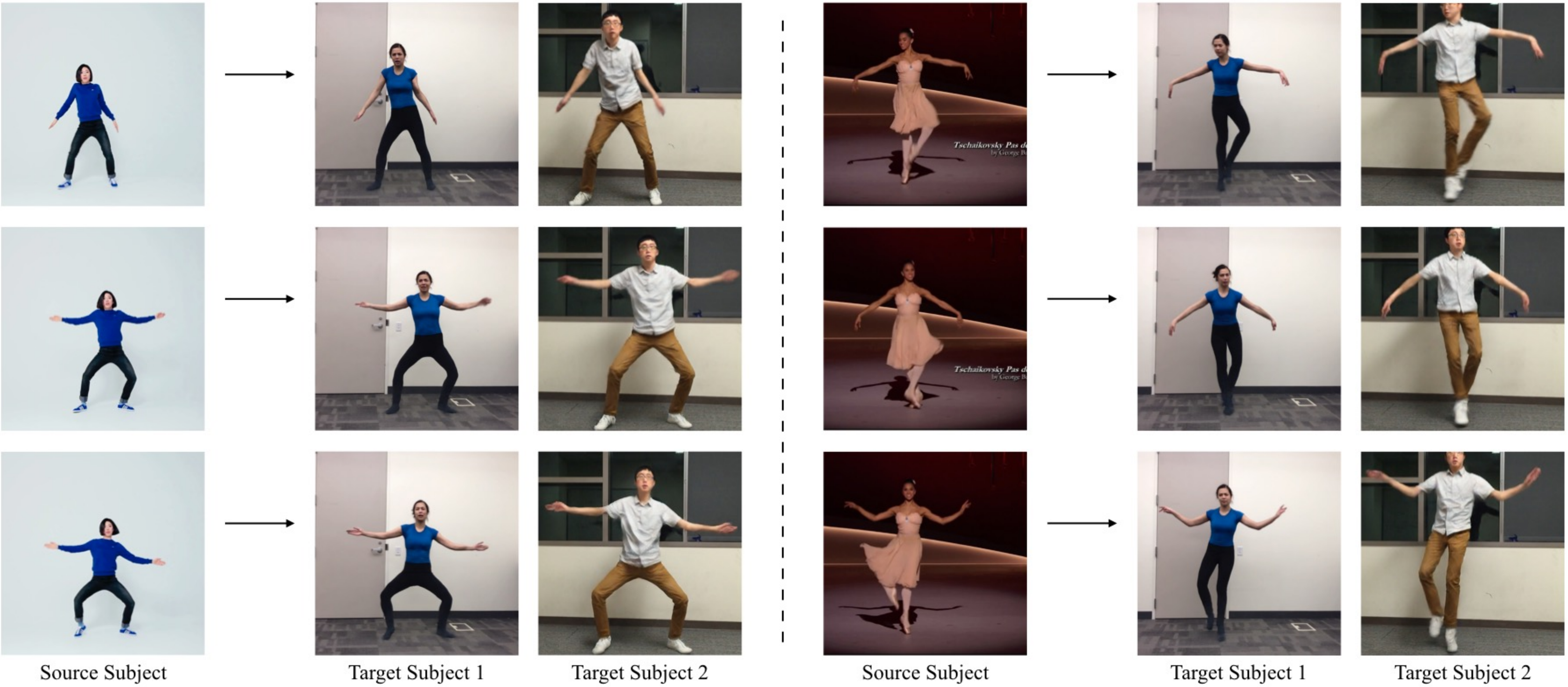


Figure 3: (Top) **Training**: Our model uses a pose detector P to create pose stick figures from video frames of the target subject. We learn the mapping G alongside an adversarial discriminator D which attempts to distinguish between the "real" correspondences $(x_t, x_{t+1}), (y_t, y_{t+1})$ and the "fake" sequence $(x_t, x_{t+1}), (G(x_t), G(x_{t+1}))$. (Bottom) **Transfer**: We use a pose detector P to obtain pose joints for the source person that are transformed by our normalization process $Norm$ into joints for the target person for which pose stick figures are created. Then we apply the trained mapping G.

C. Chan, S. Ginosar, T. Zhou, A. Efros. [Everybody Dance Now](). ICCV 2019



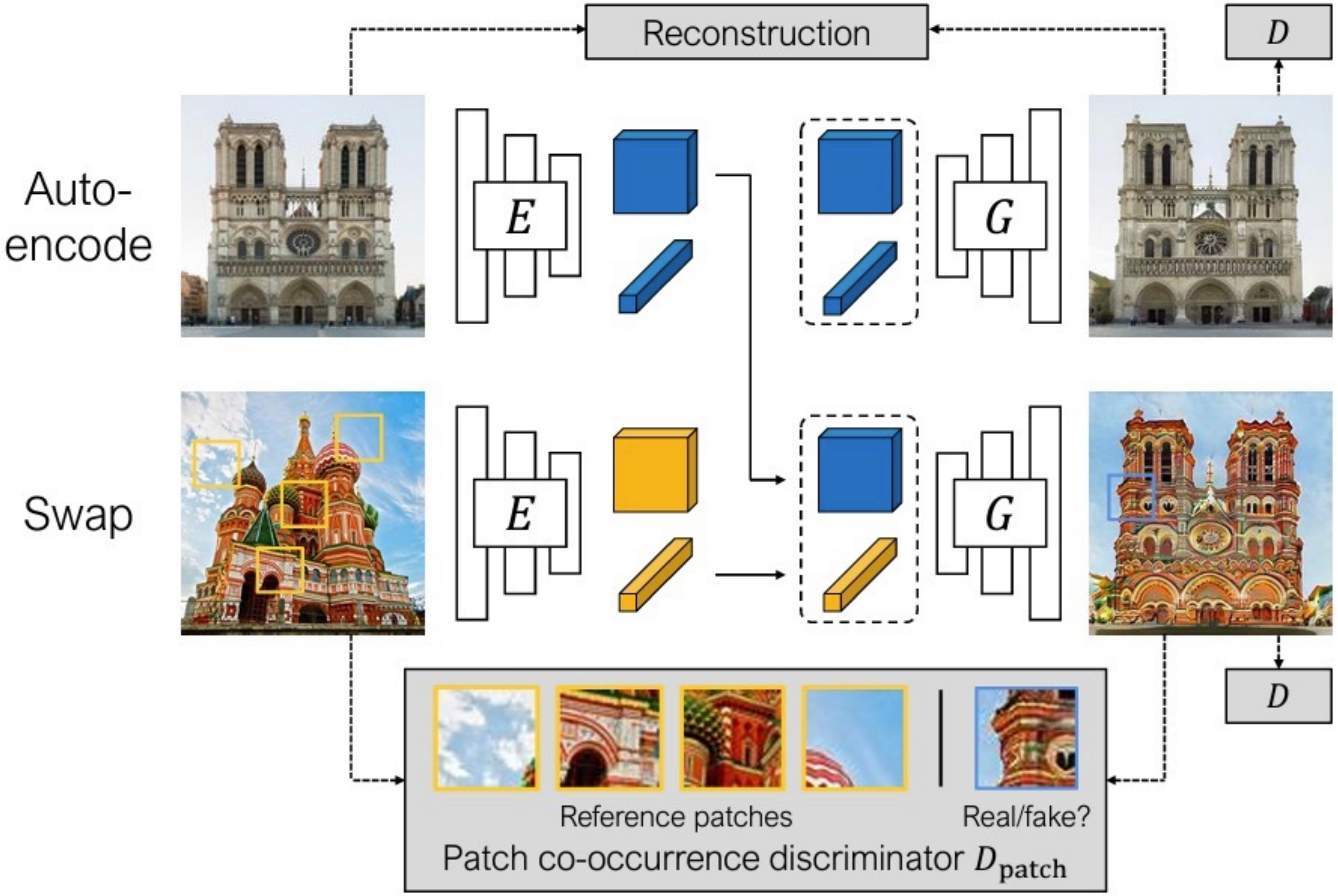
https://carolineec.github.io/everybody_dance_now/

C. Chan, S. Ginosar, T. Zhou, A. Efros. Everybody Dance Now. ICCV 2019


Source Video
Detected Pose
Source to Target 1: Result
Source to Target 2:

Other Applications of Adversarial Learning

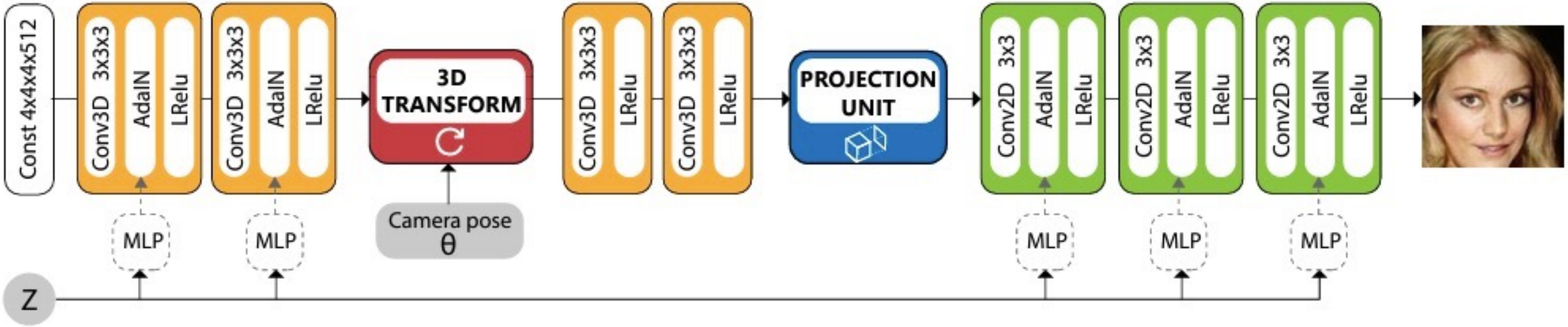
Swapping Autoencoder



Park et al. 2020

Swapping Autoencoder

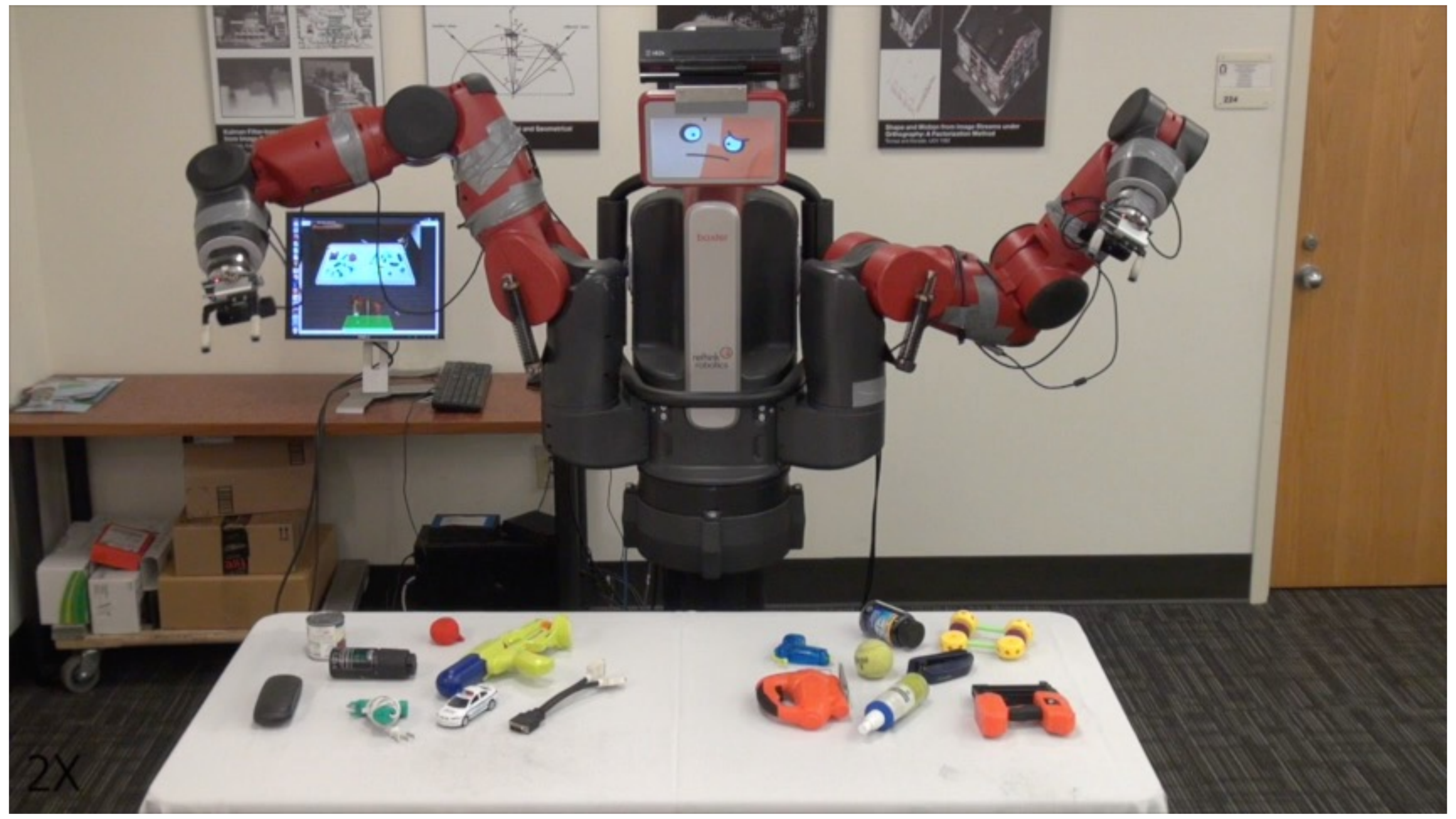
HoloGAN



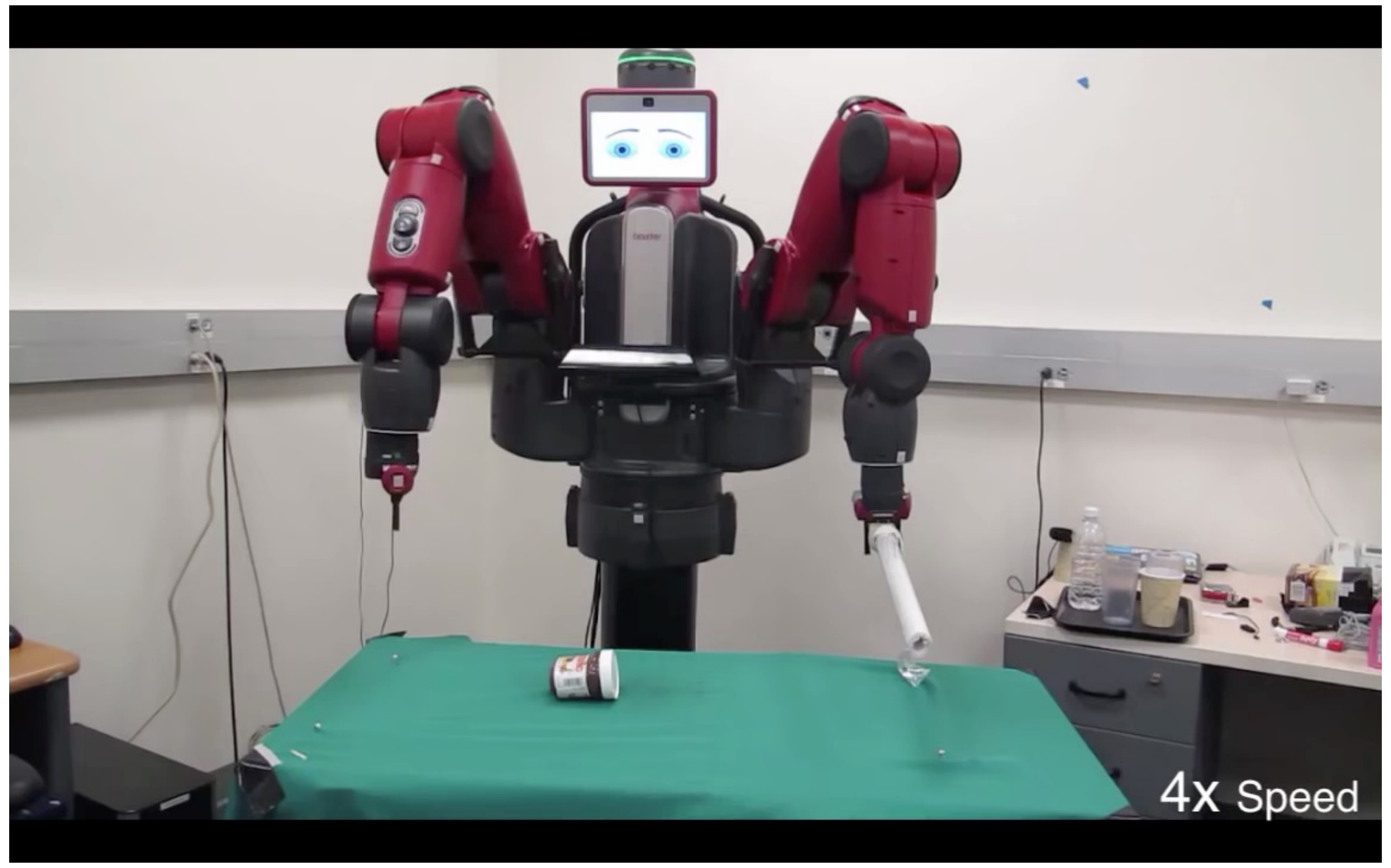
Const 4x4x4x512
Conv3D 3x3x3
AdaIN
LRelu
Conv3D 3x3x3
AdaIN
LRelu
3D TRANSFORM
Camera pose θ
Conv3D 3x3x3
LRelu
Conv3D 3x3x3
LRelu
PROJECTION UNIT
Conv2D 3x3
AdaIN
LRelu
Conv2D 3x3
AdaIN
LRelu
Conv2D 3x3
AdaIN
LRelu
MLP
MLP
MLP
MLP
MLP
Z

HoloGAN

Self-Supervised Robot Learning



Pinto et al. ICRA 2016

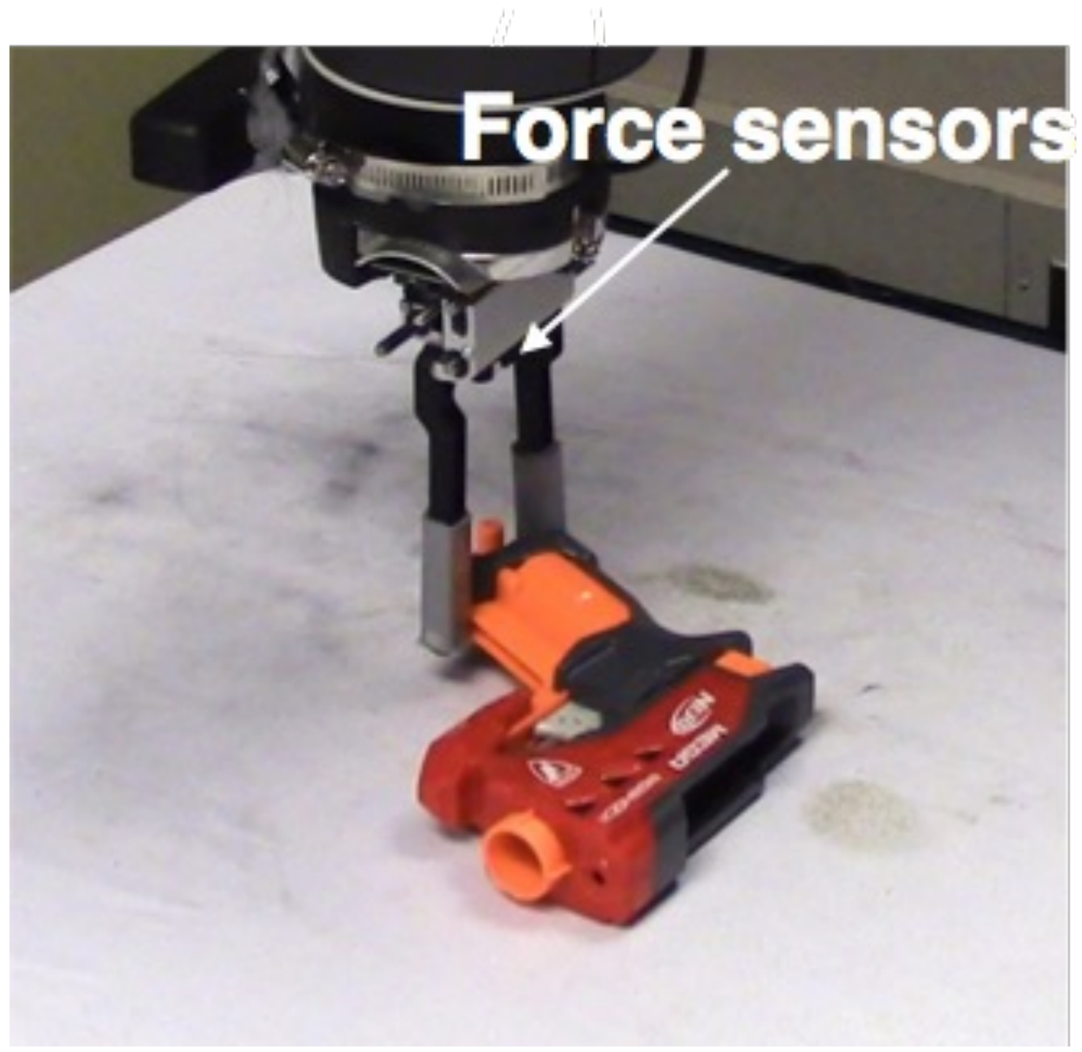


Agrawal et al. NIPS 2016

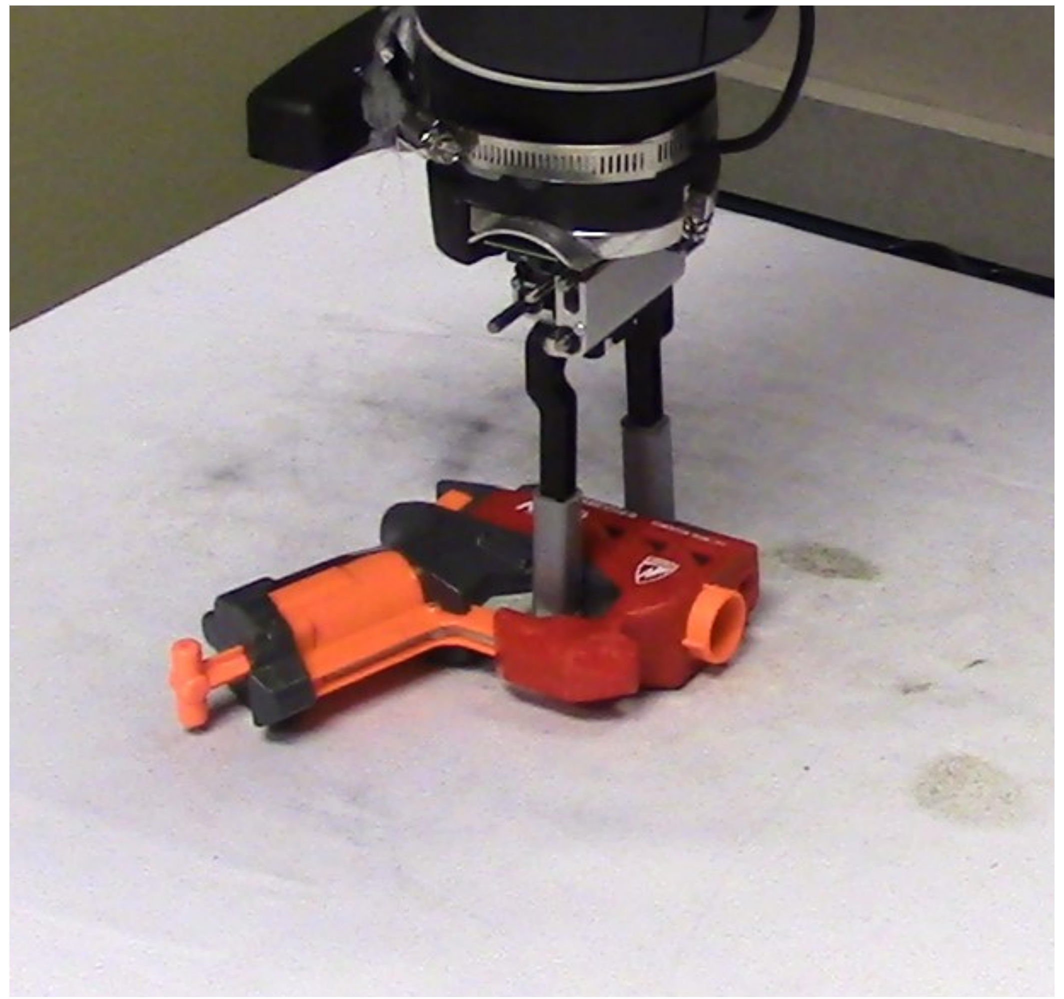
Levine et al. ISER 2016

Sensory supervision alone is weak

Hard to distinguish grasps:



vs



Pinto et al. Supervision via Competition: Robot Adversaries for Learning Tasks . ICRA 2017.

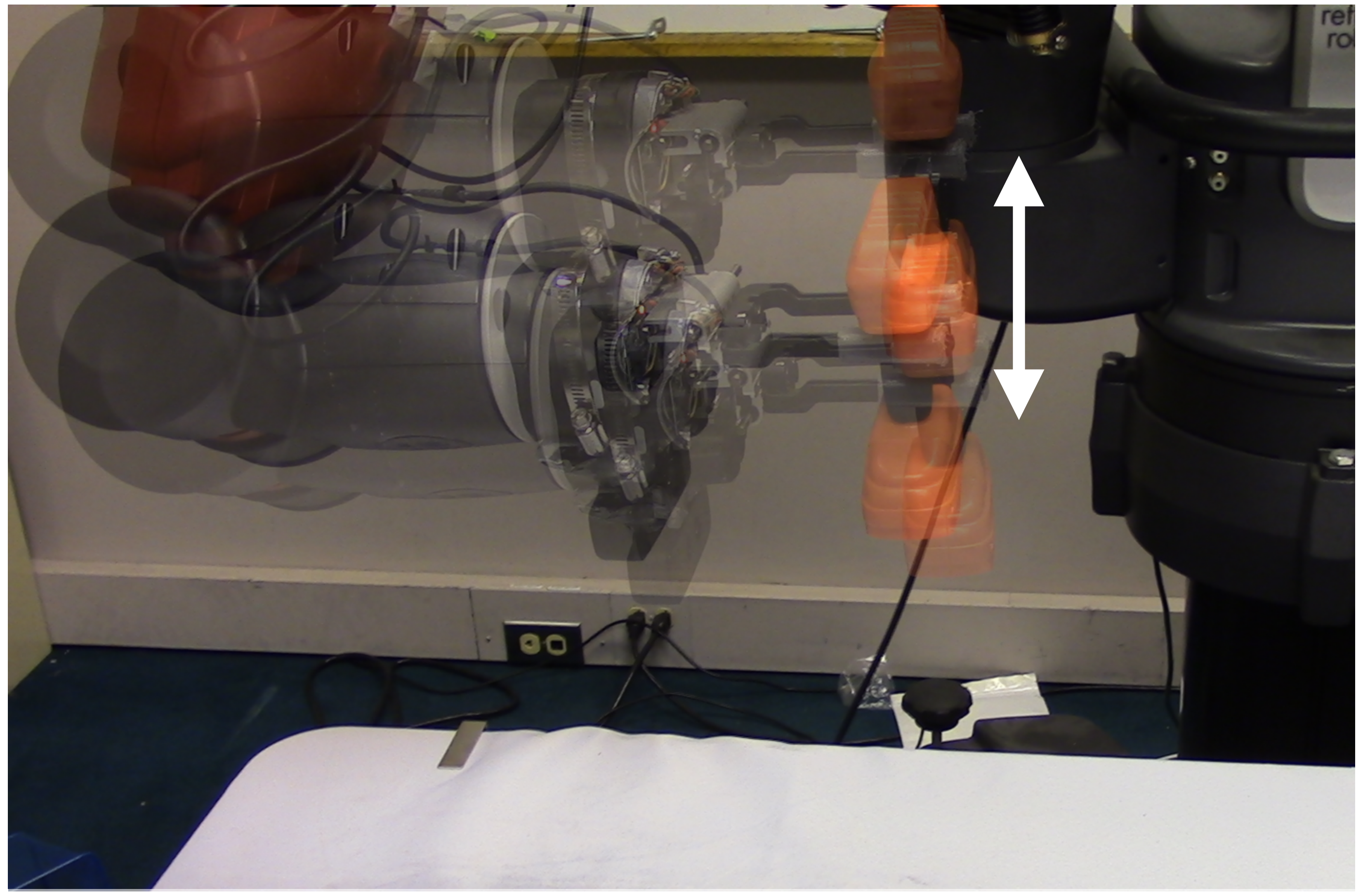
So what do humans do?

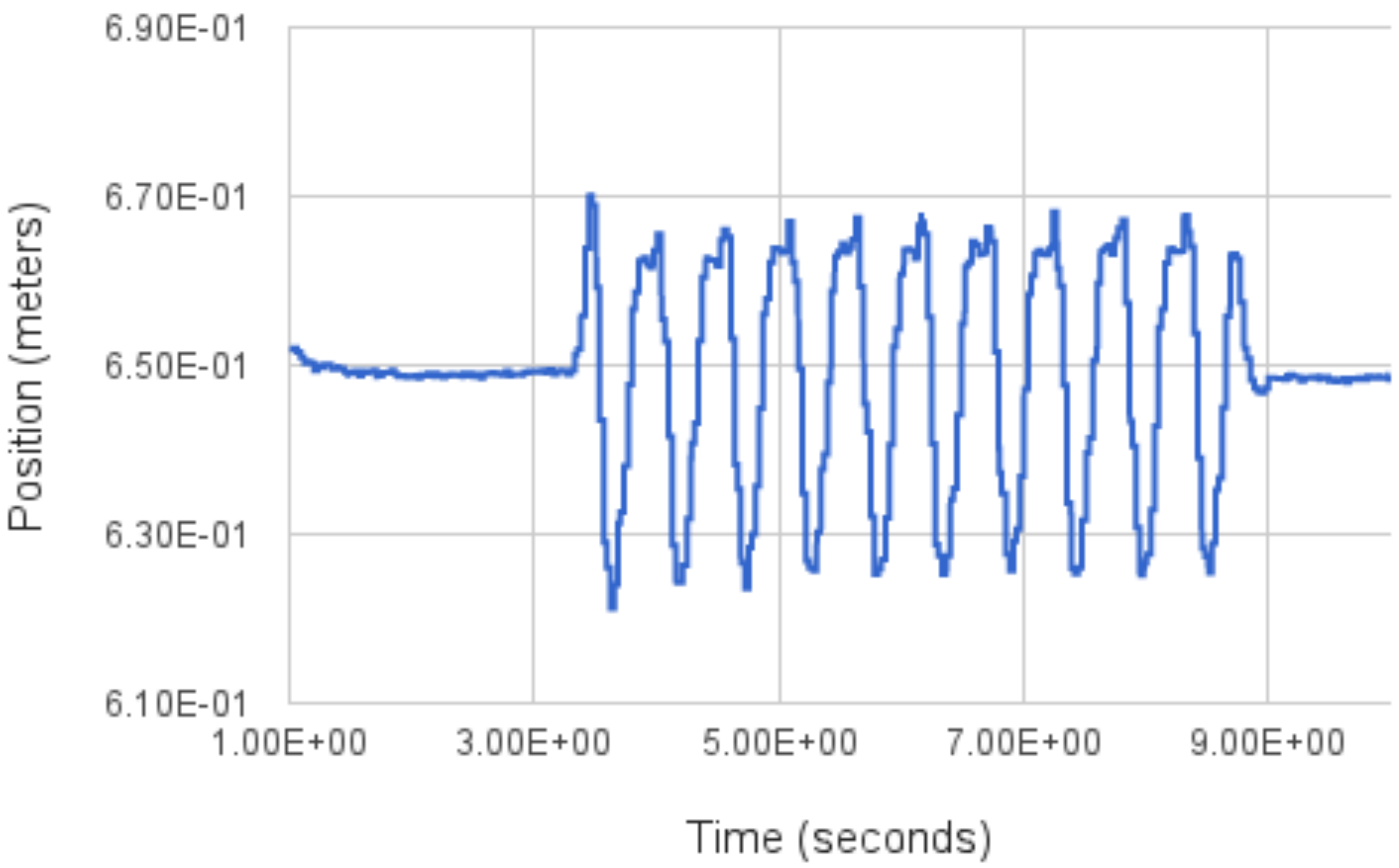
So what do humans do?

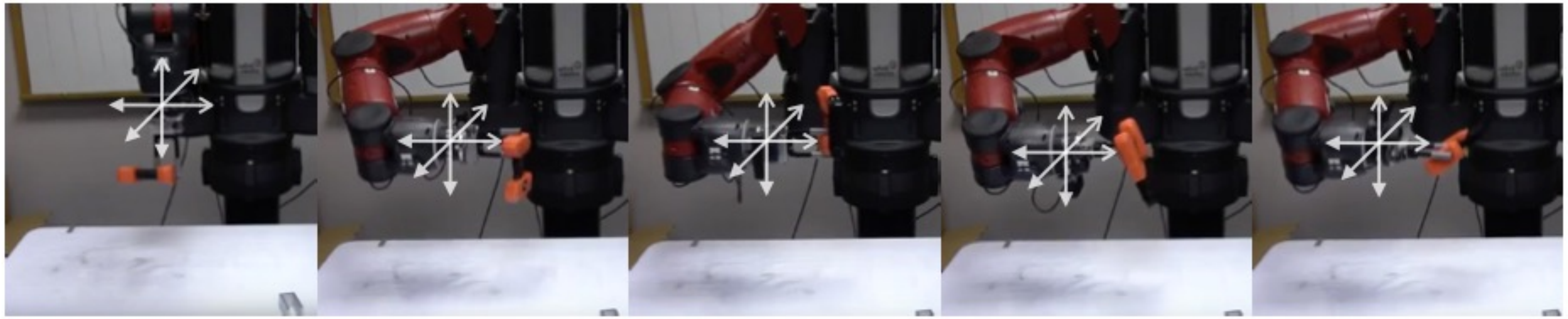
Key Idea

Extract **more** information with **Adversarial** agents.

An Adversary that Shakes

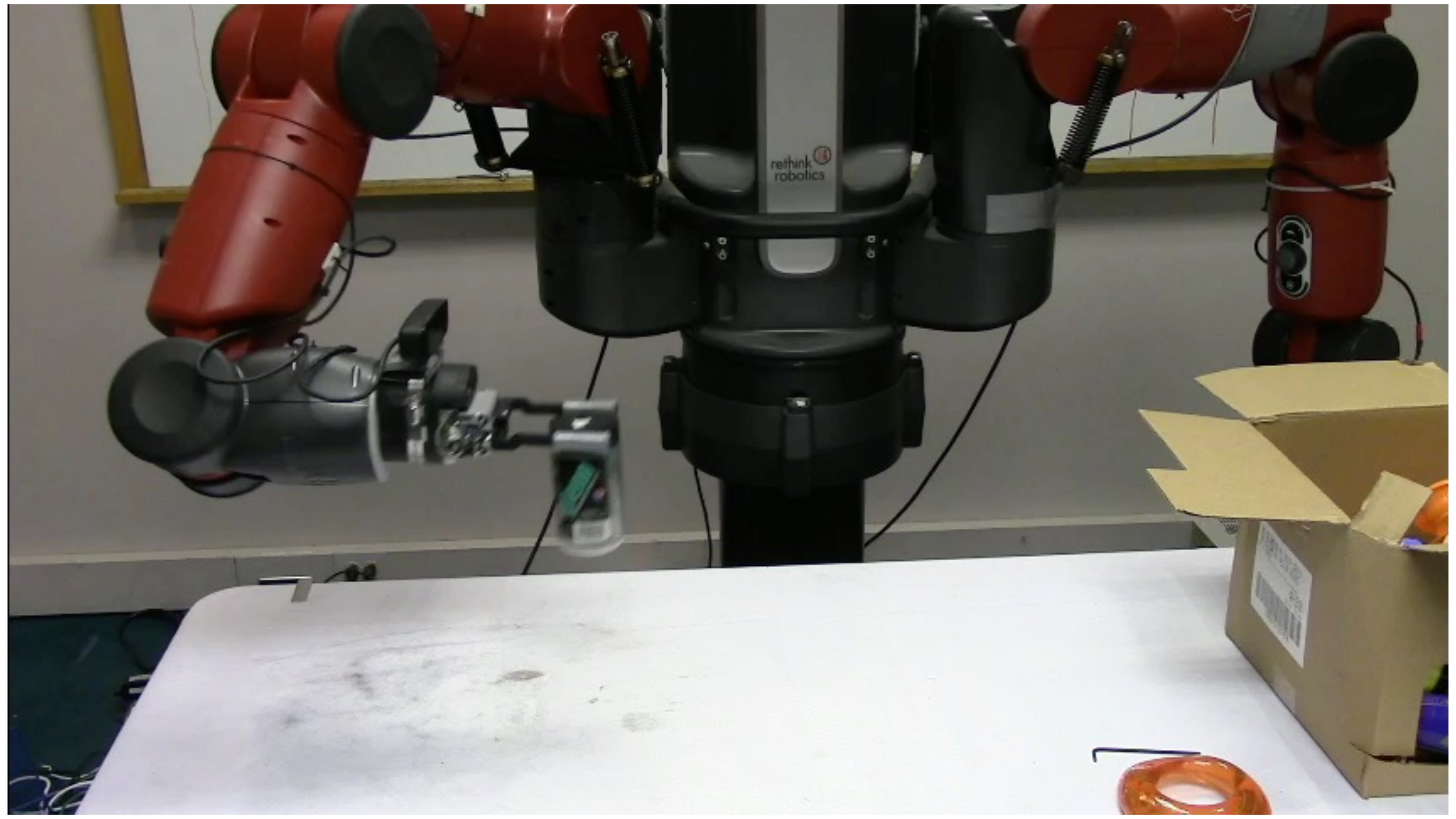




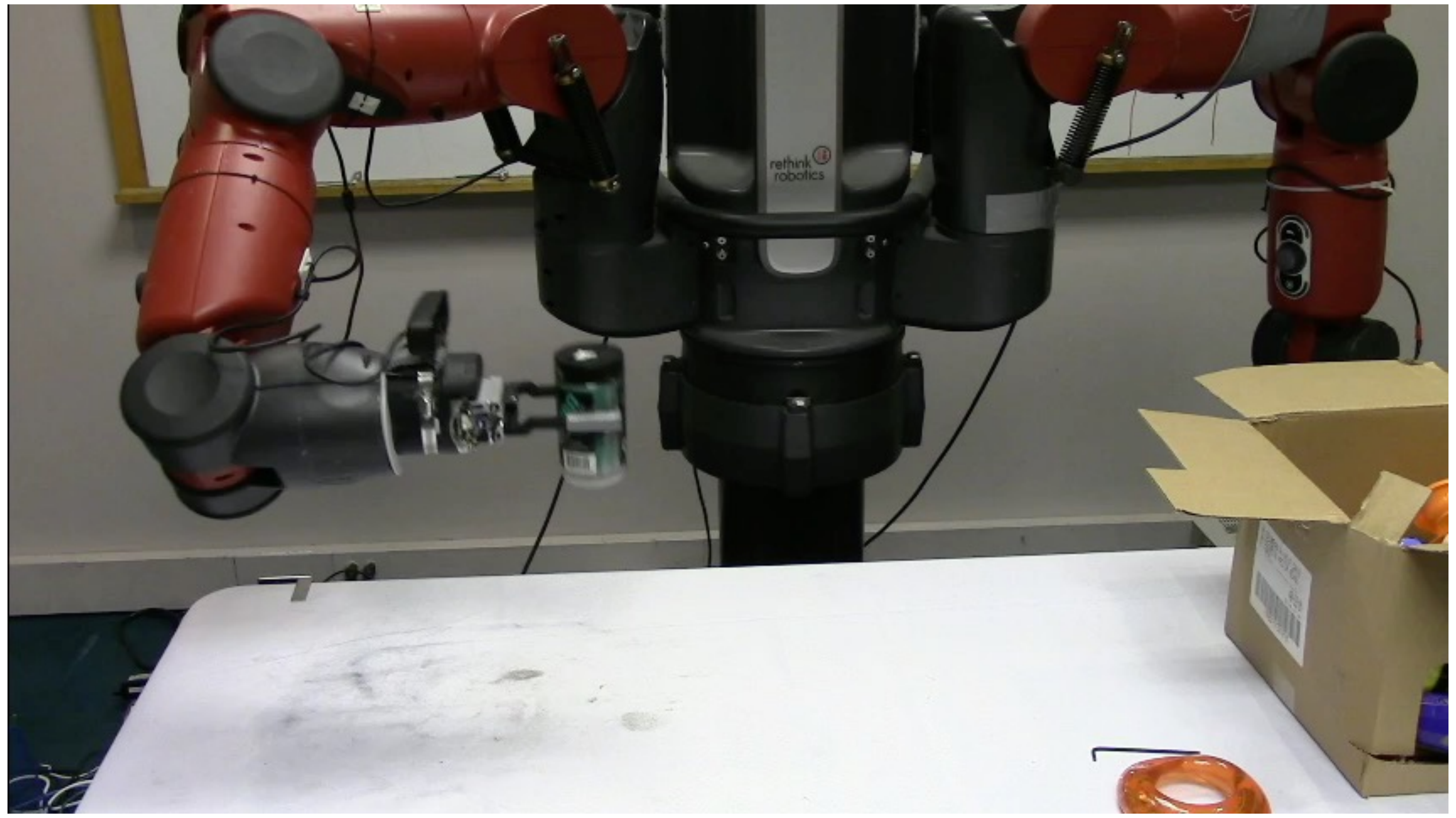


Destabilization of an unstable grasp by Shaking

Unstable Grasp

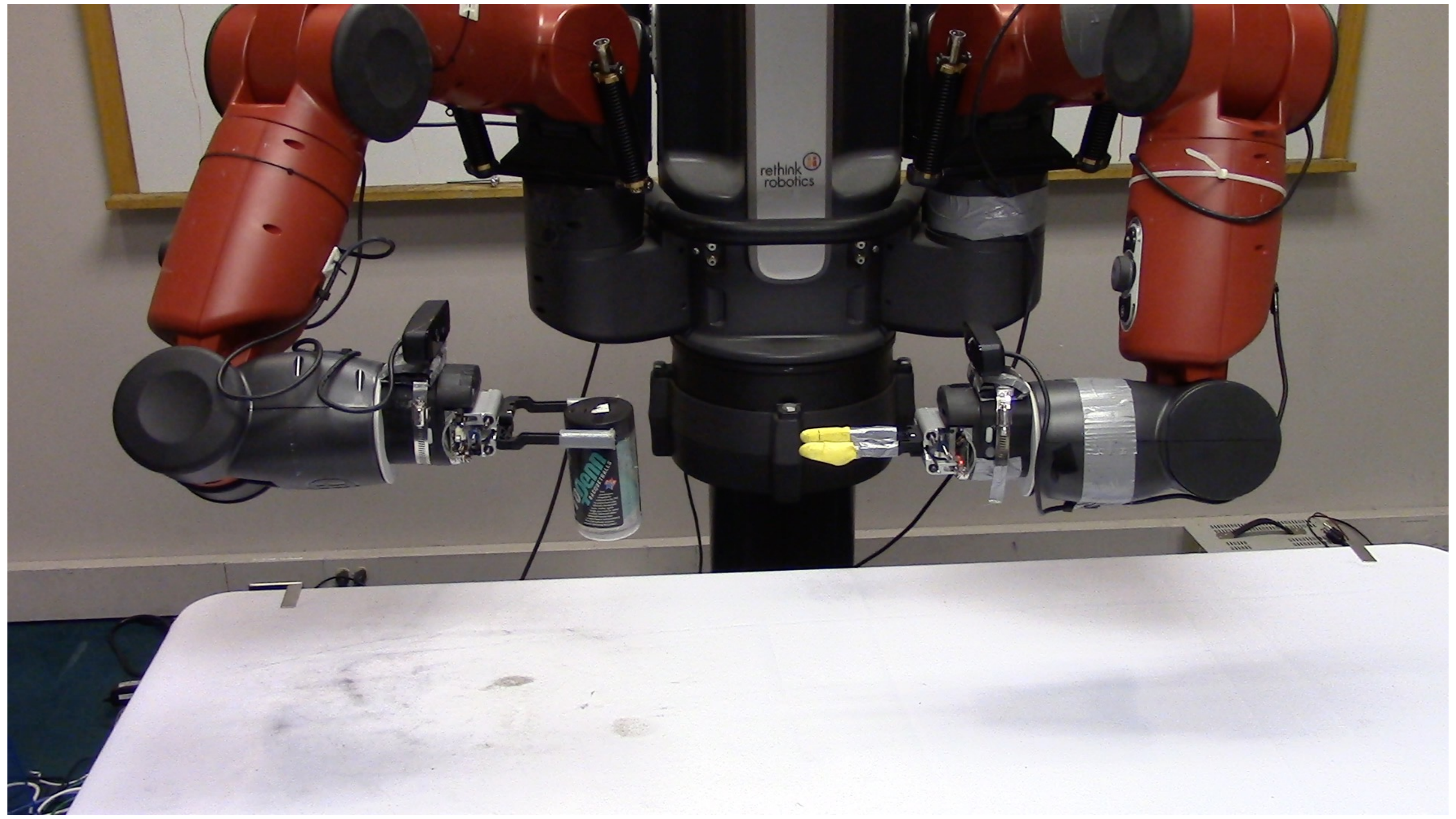


Stable Grasp

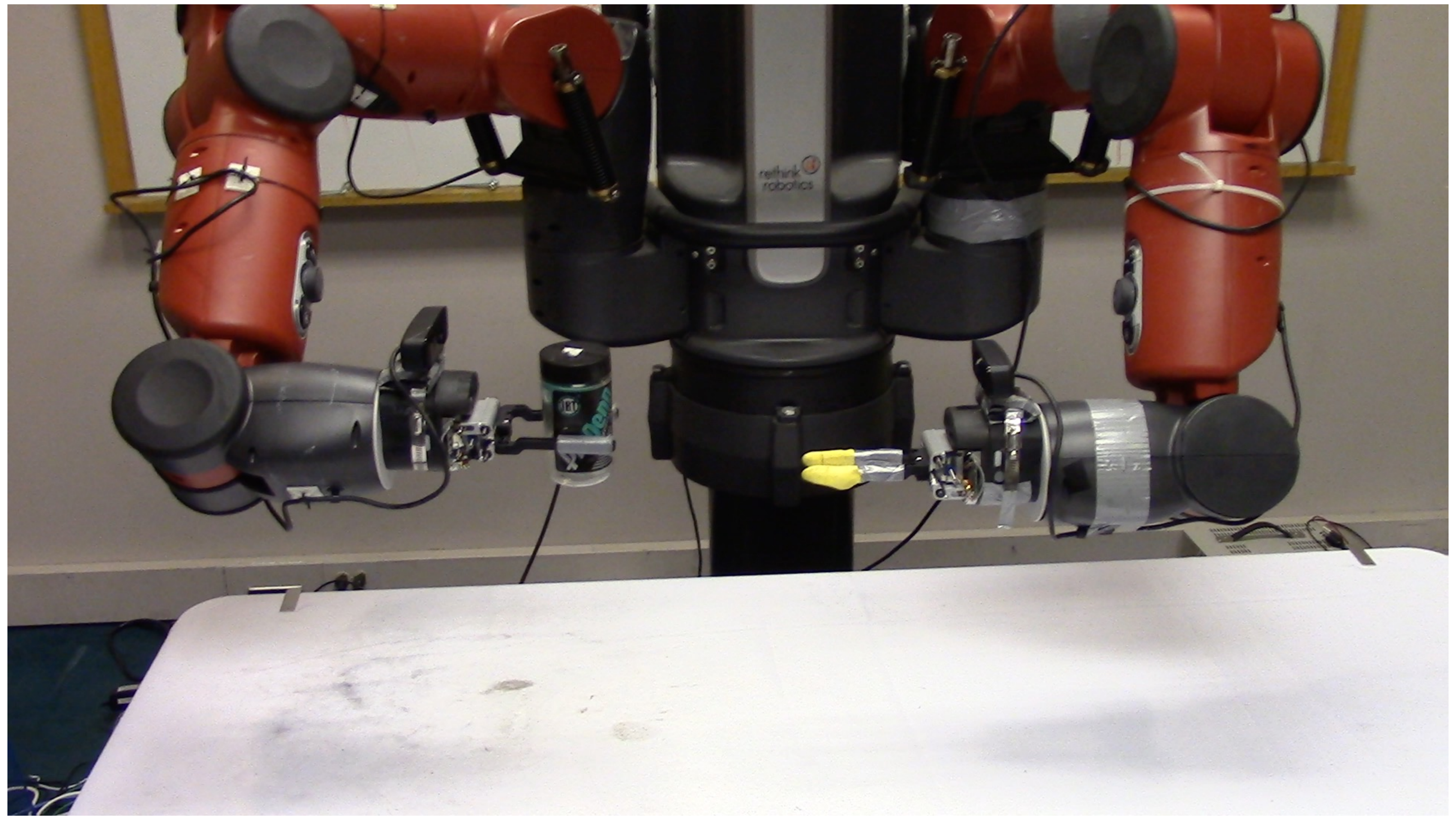


Destabilization of an unstable grasp by Snatching

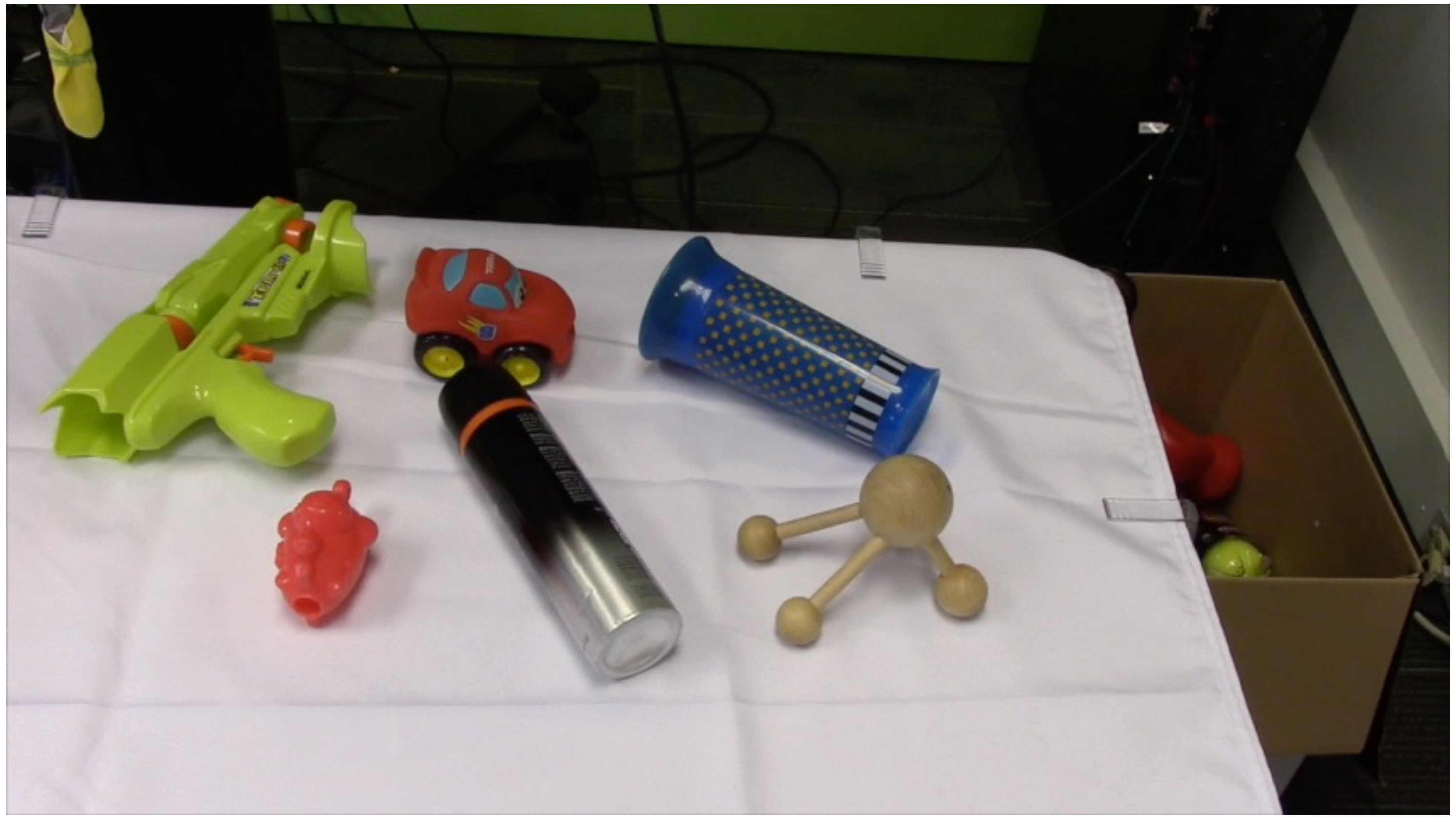
Unstable Grasp



Stable Grasp



Results



base	Shake	Snatch
68%	80%	82%

Summary

- Image-to-Image Translation: pix2pix
- Unpaired Image-to-Image Translation: CycleGAN
- Other Applications of Adversarial Learning