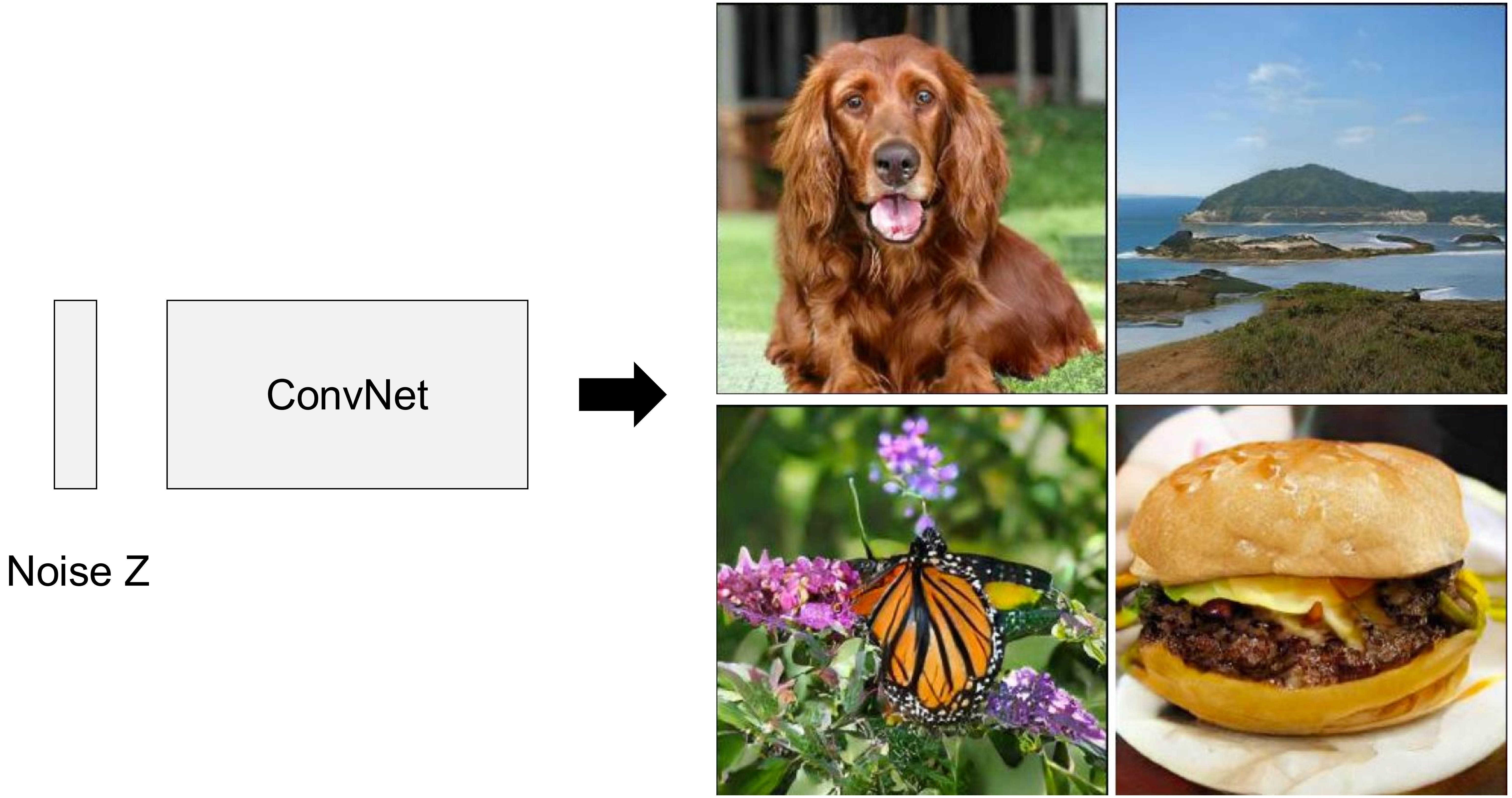
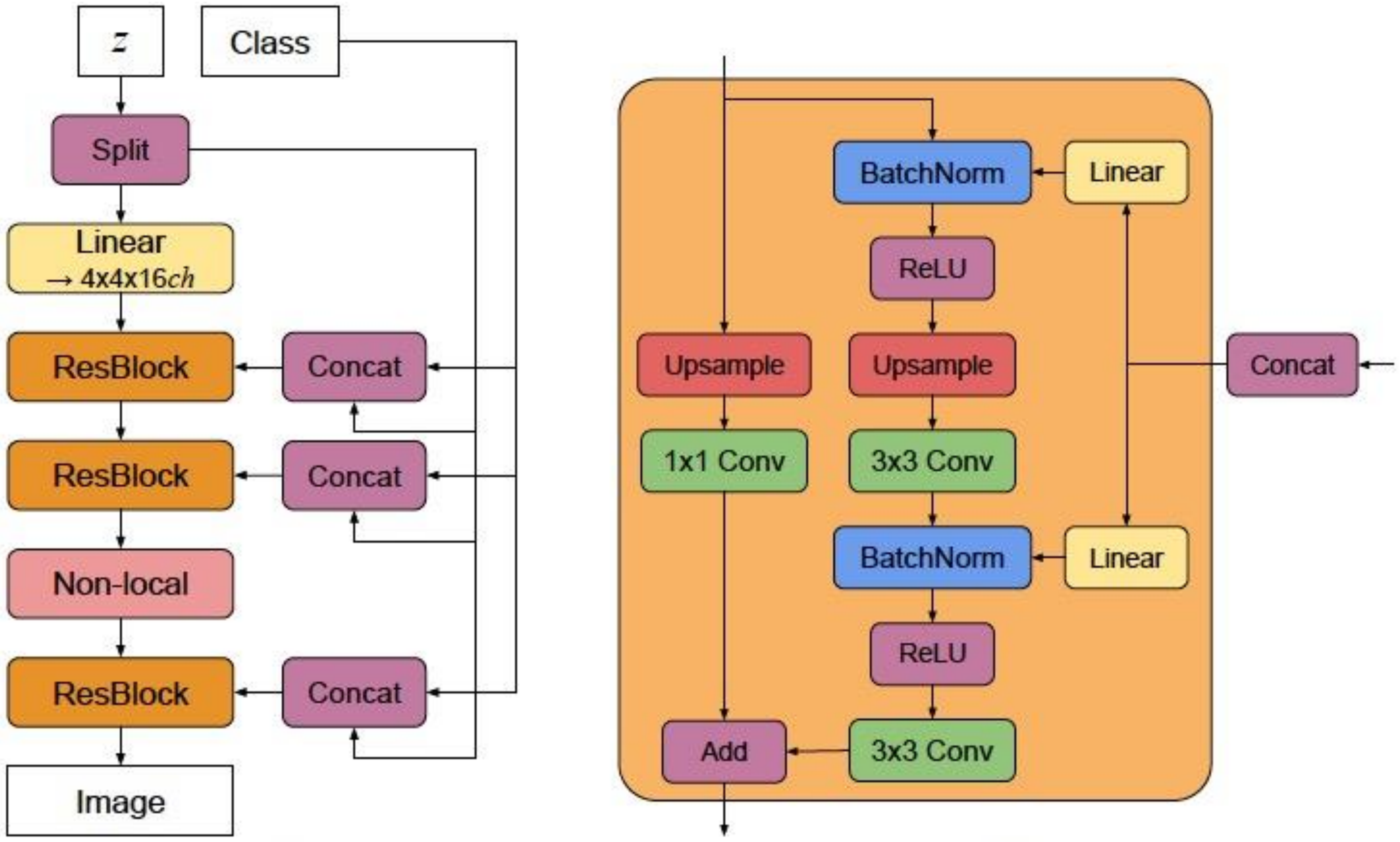
Conditional GAN and Variational Auto-Encoders

Xiaolong Wang

Last class



BigGAN: Class-Conditioned



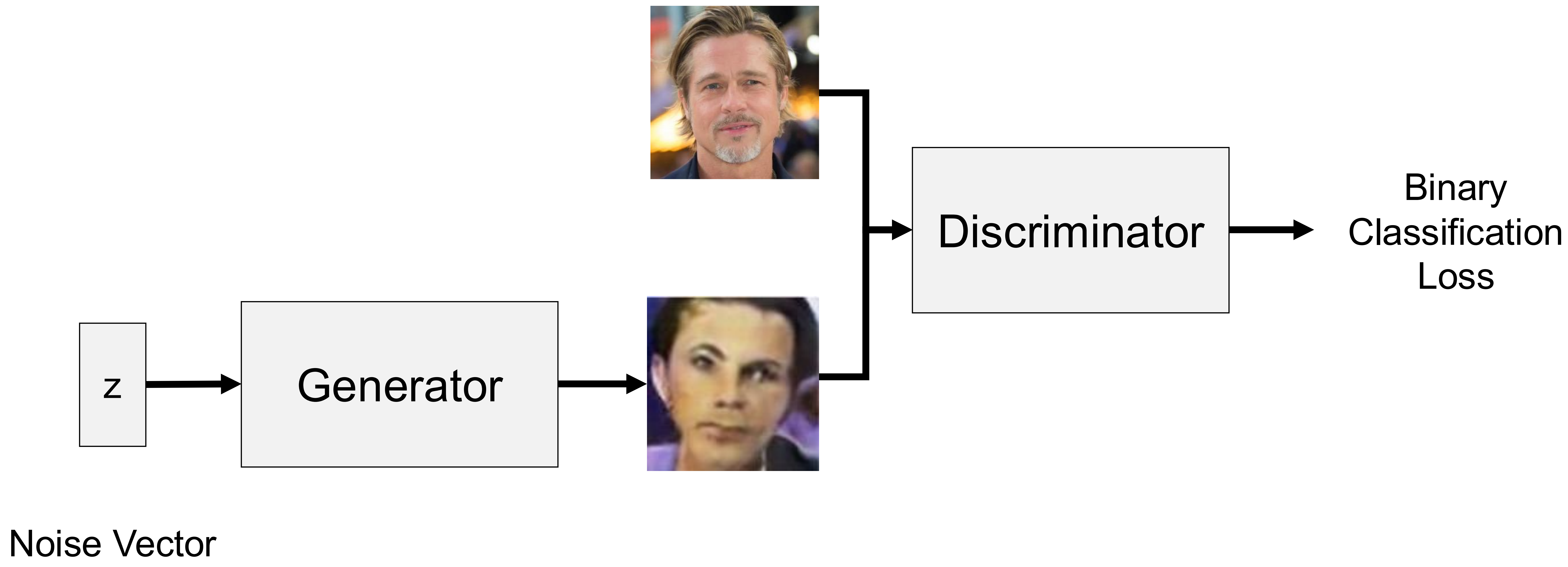
(a) (b)

This Class

- Image-to-Image Translation: pix2pix
- Unpaired Image-to-Image Translation: CycleGAN
- Variational Autoencoder (VAE)

Image-to-Image Translation: pix2pix

GANs



Goodfellow et al., 2014

Conditional GANs

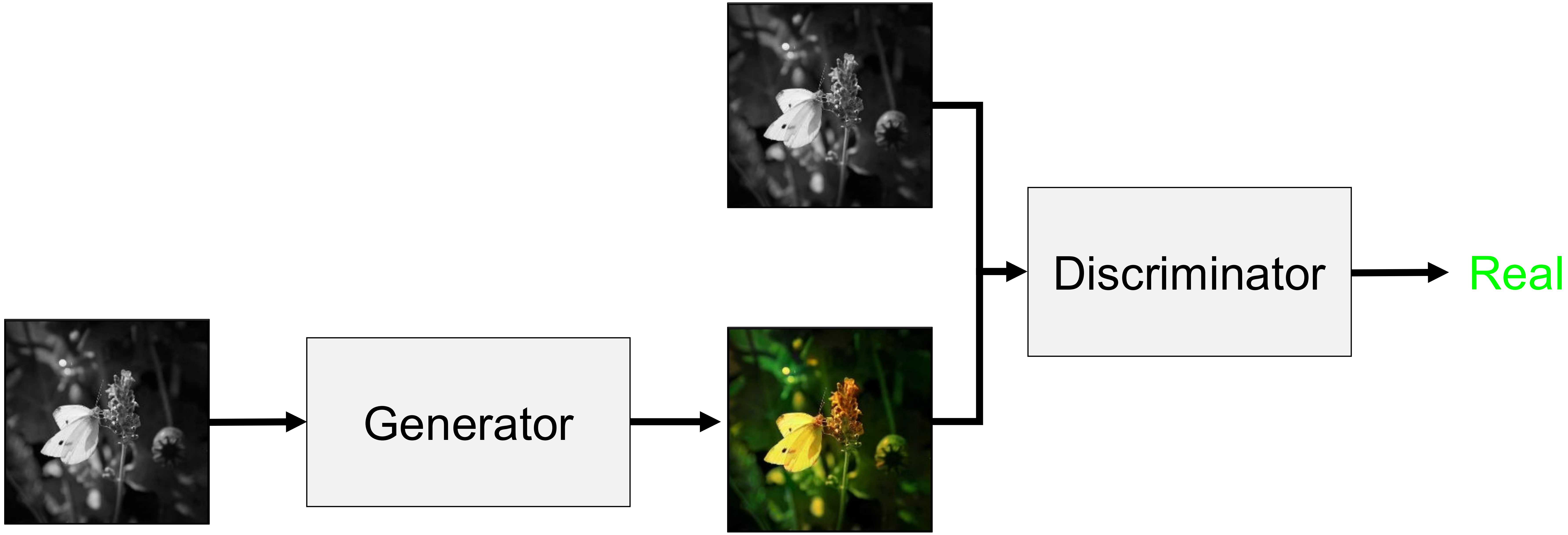
Edges to Photo

input output

BW to Color

input output

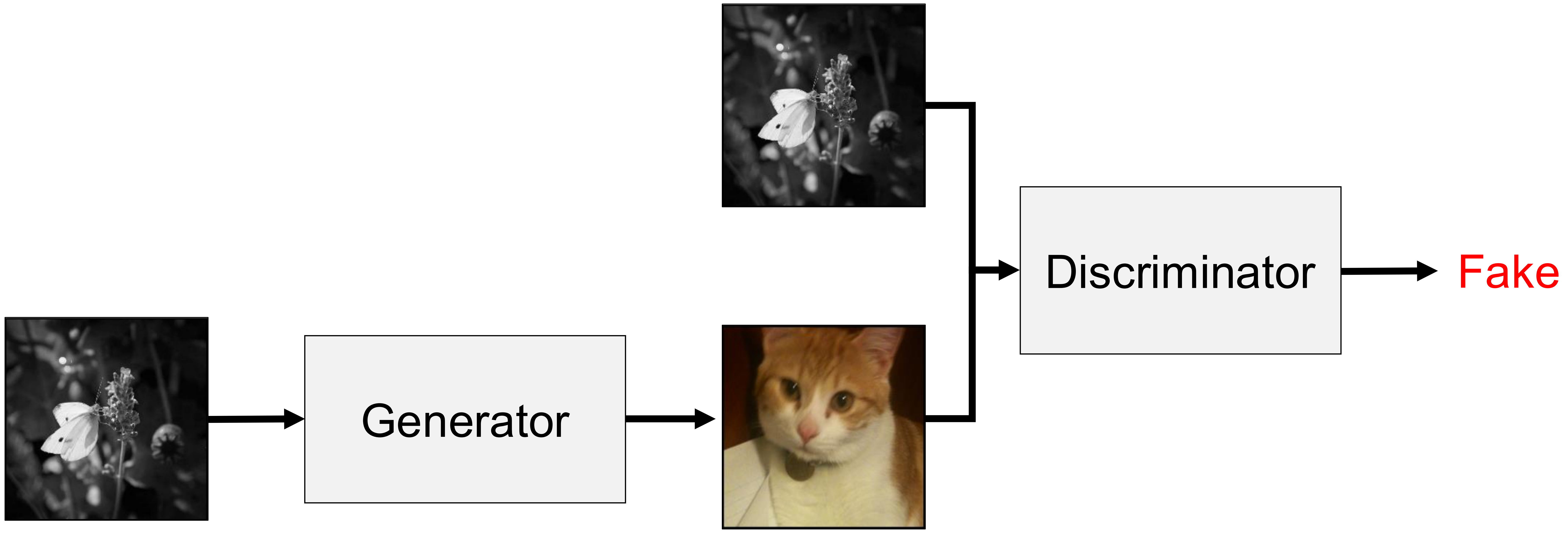
Conditional GANs



Generator takes an image as input, not noise.

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

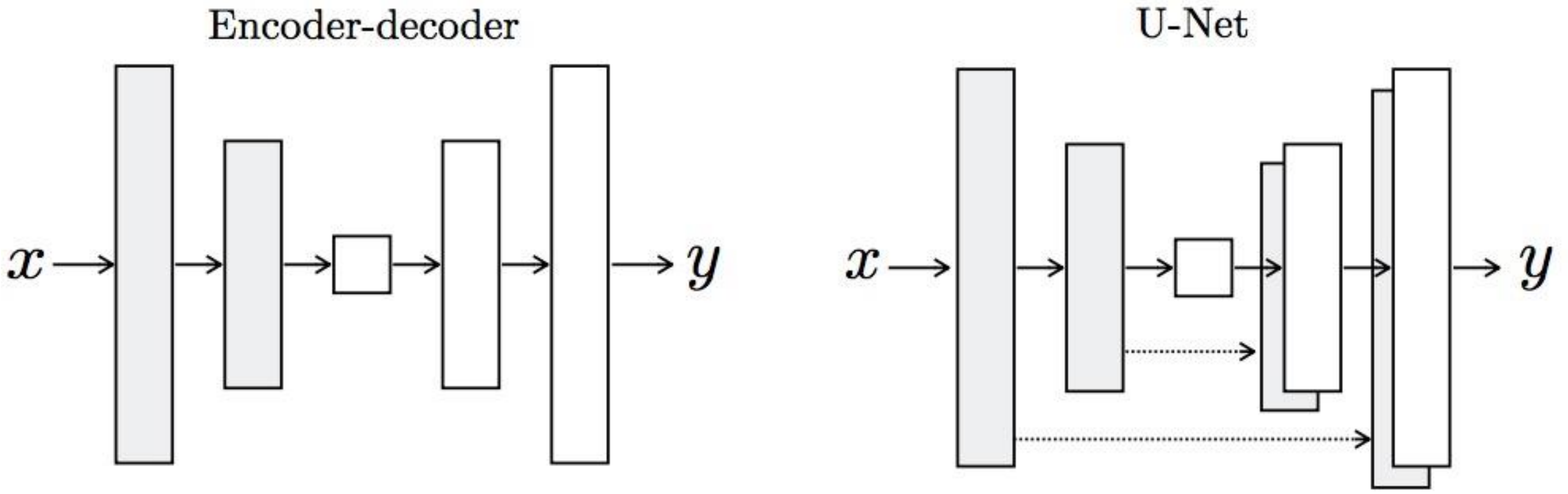
Conditional GANs



Generator takes an image as input, not noise.

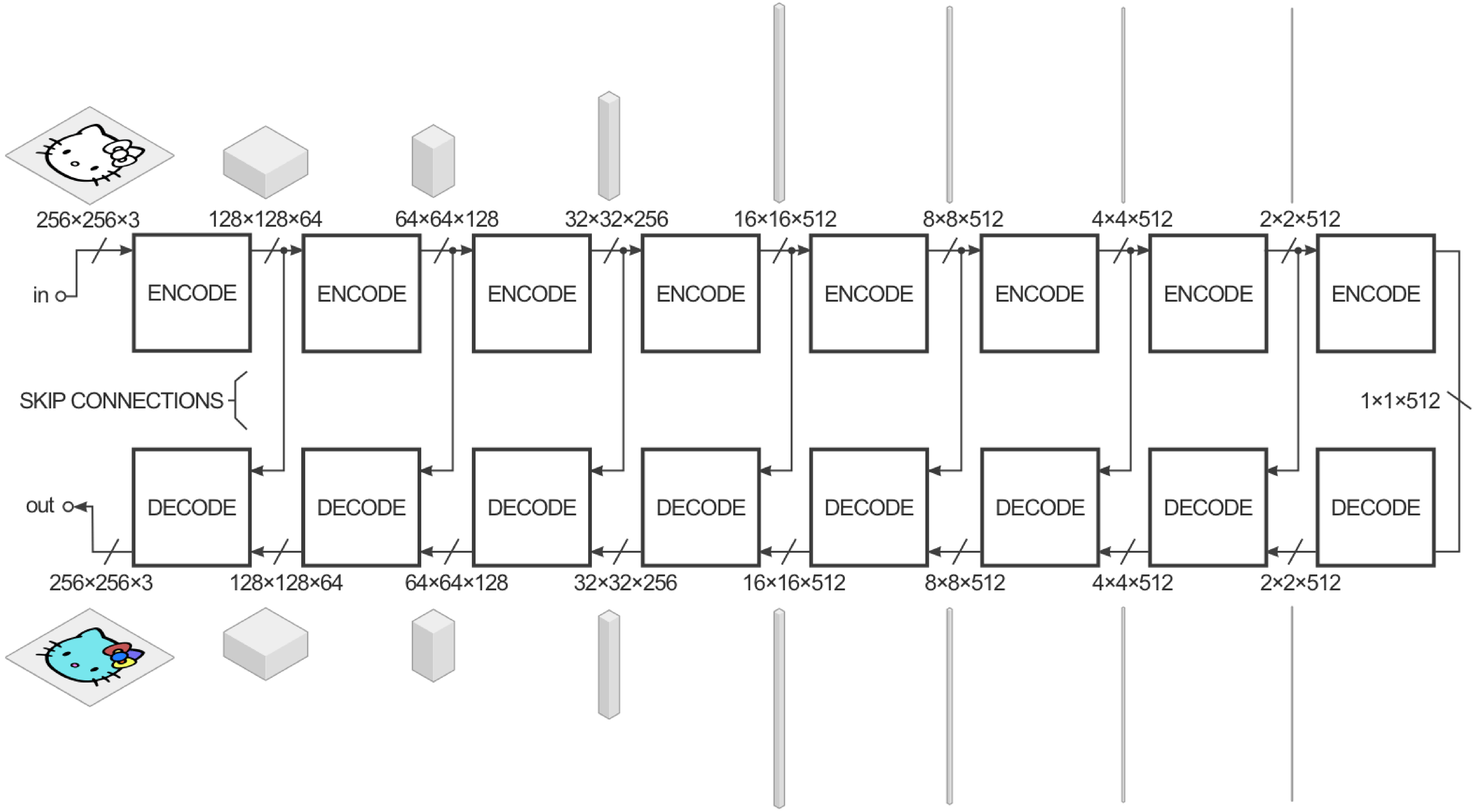
Discriminator takes a pair of images as inputs, not just one image.

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

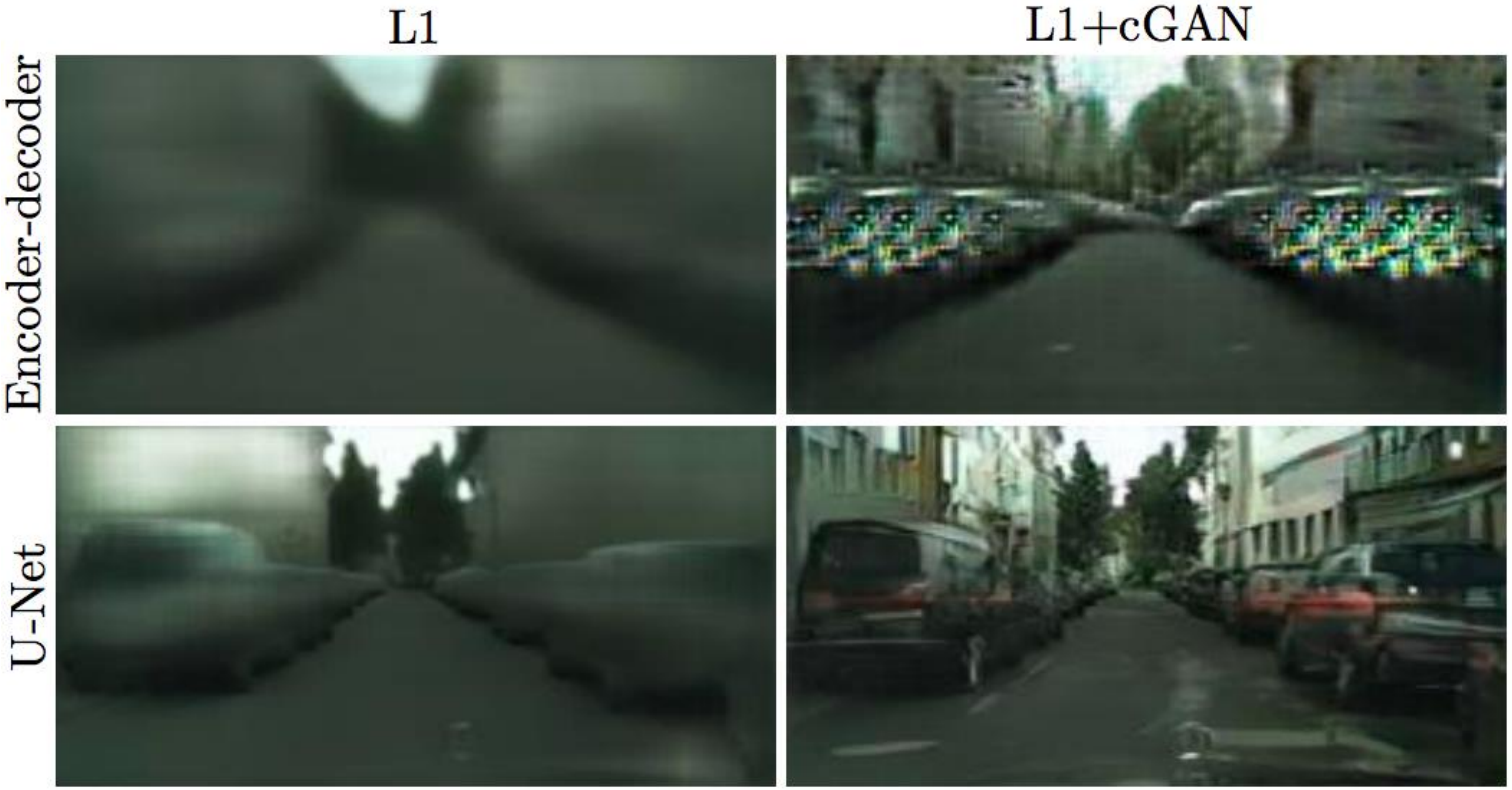


Image-to-image translation

- Generator loss: GAN loss plus L1 reconstruction penalty

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

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

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

Image-to-image translation: Results

- Day to night

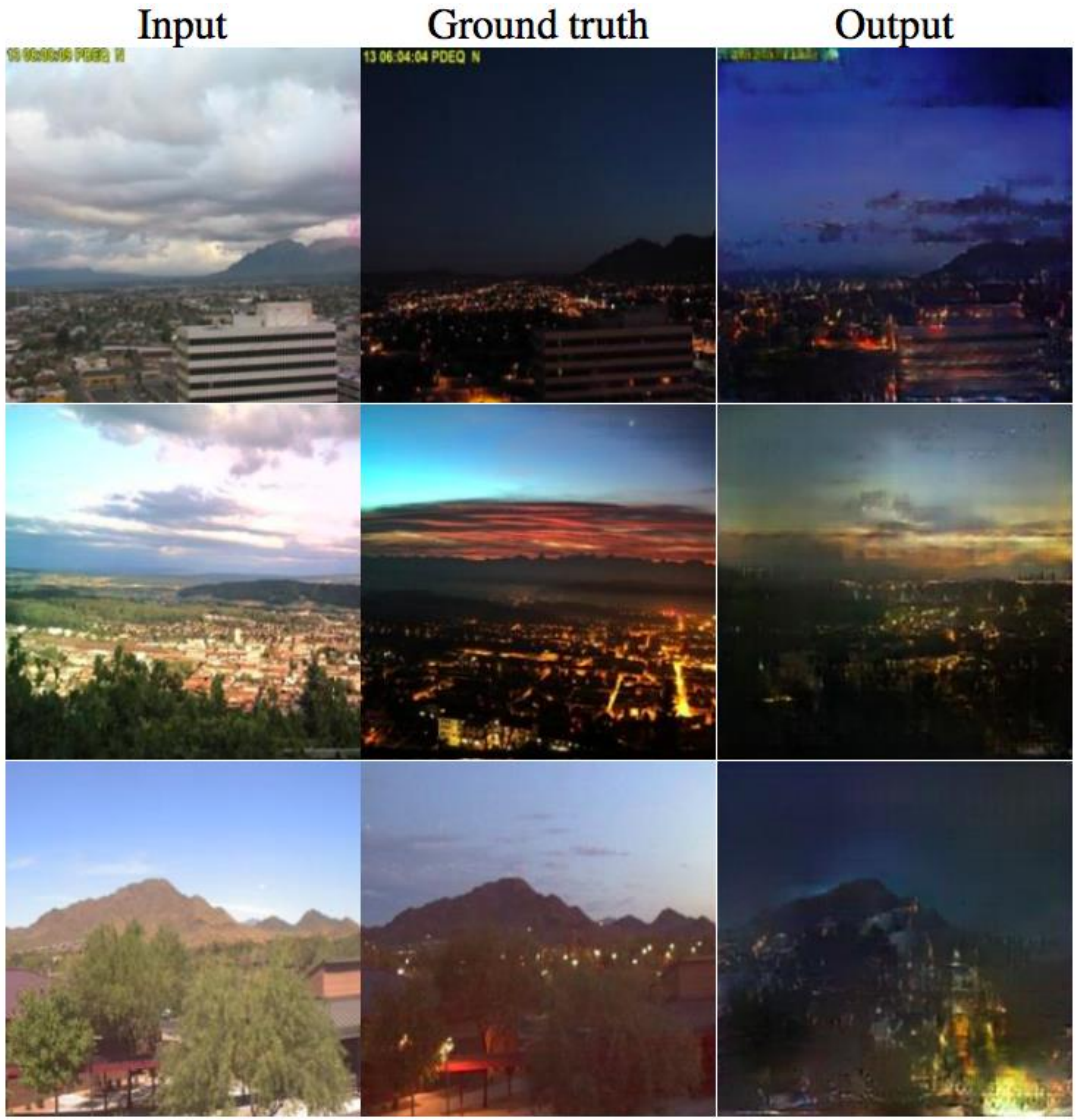
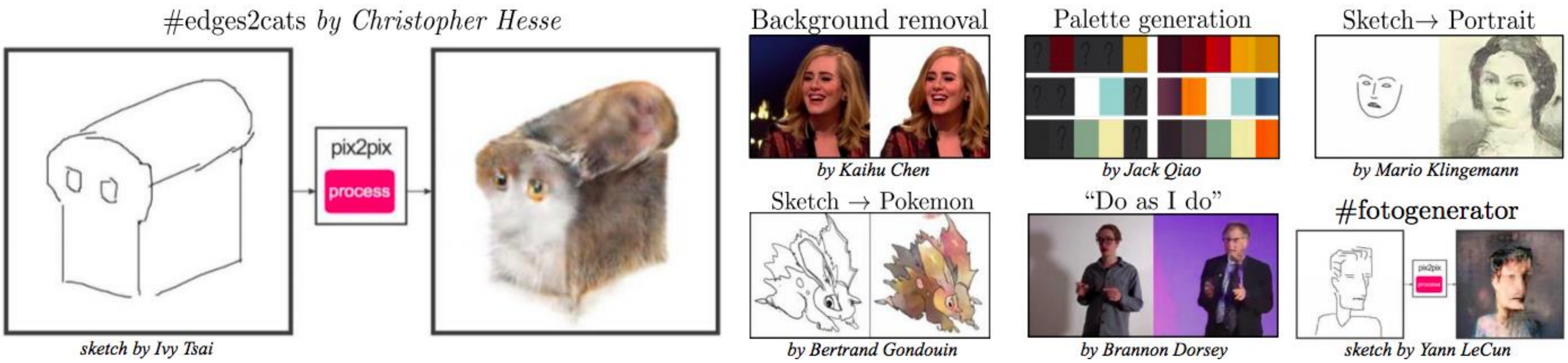


Image-to-image translation: Results

- Edges

Image-to-image translation: Results

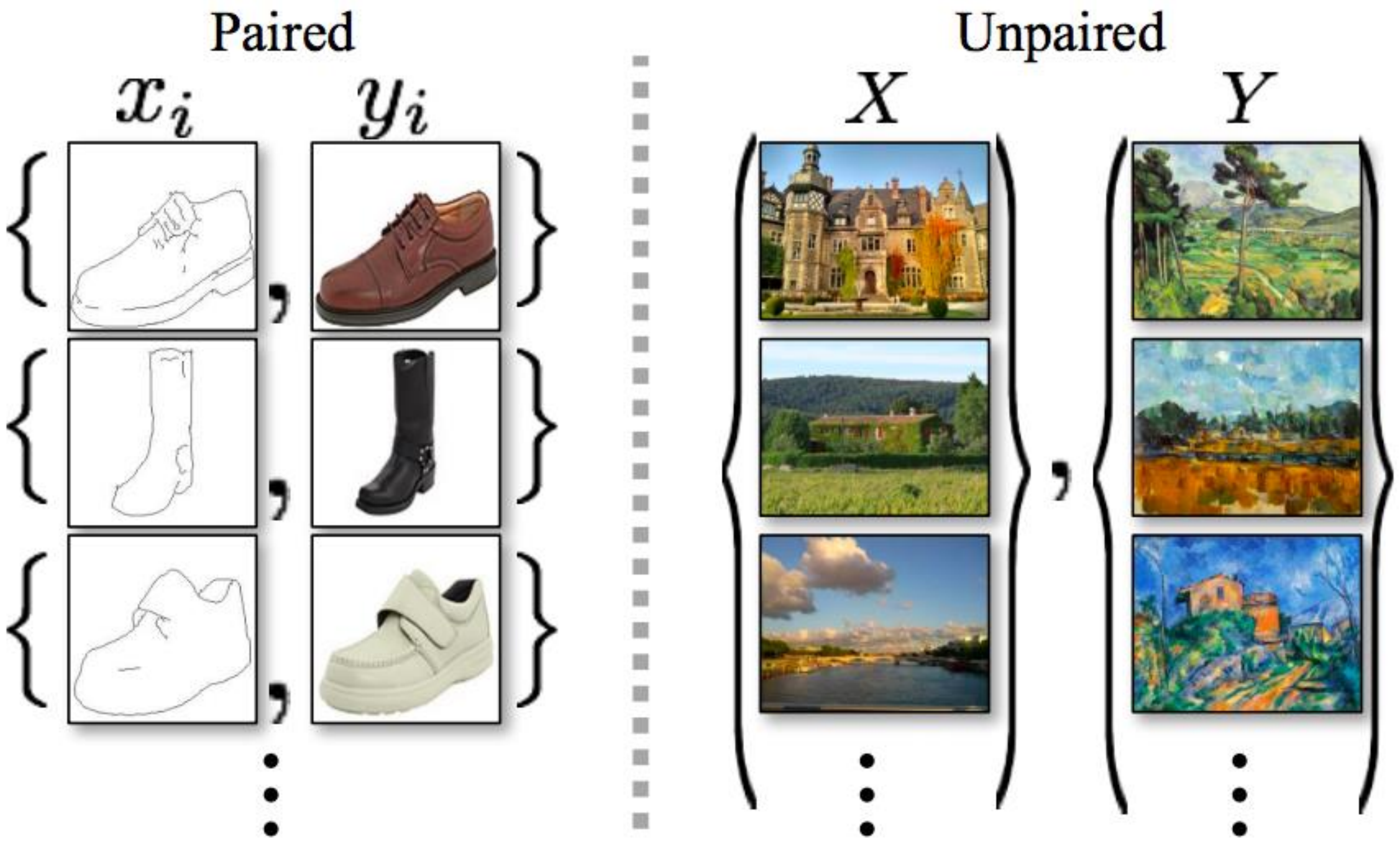
- [pix2pix demo]()



Unpaired Image-to-Image Translation: CycleGAN

Unpaired image-to-image translation

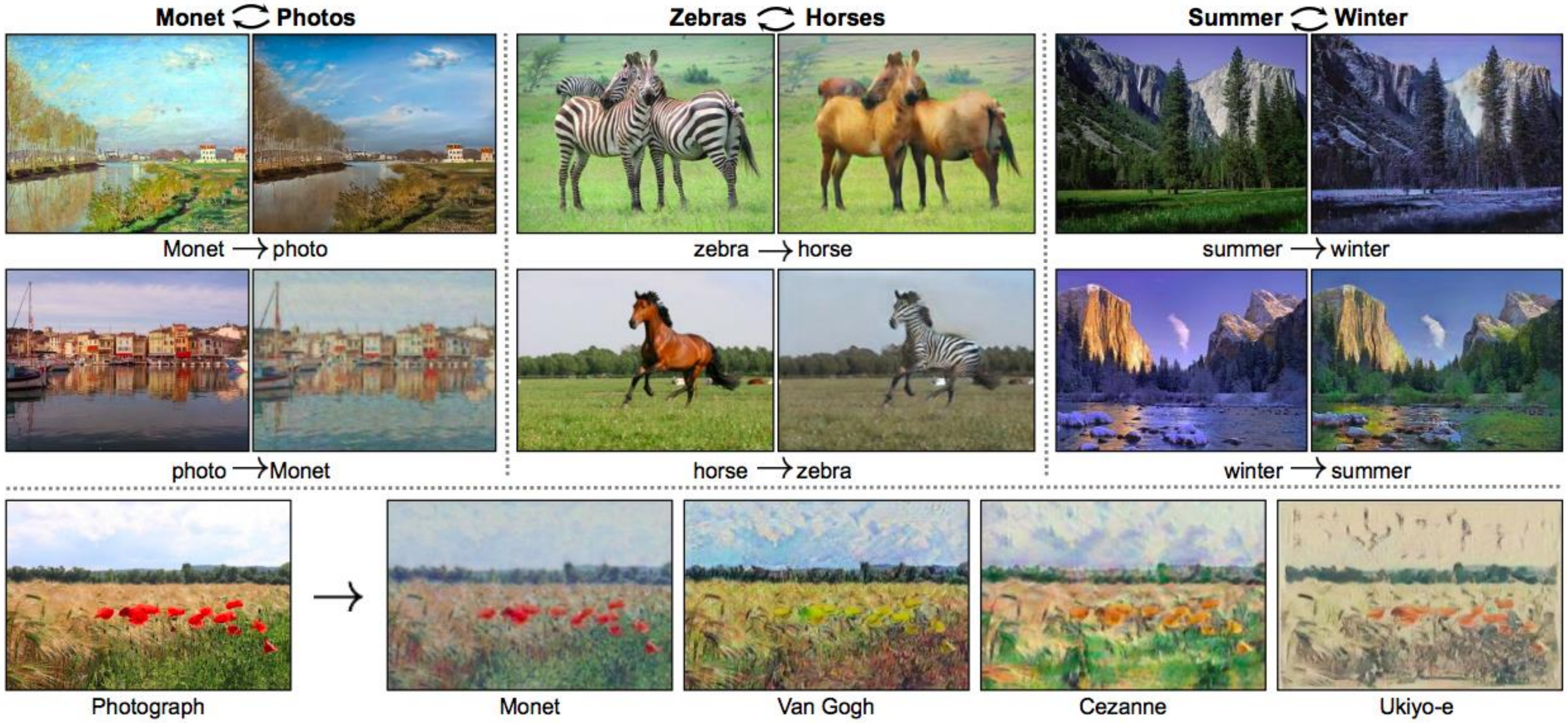
- Given two unordered image collections X and Y, learn to "translate" an image from one into the other and vice versa



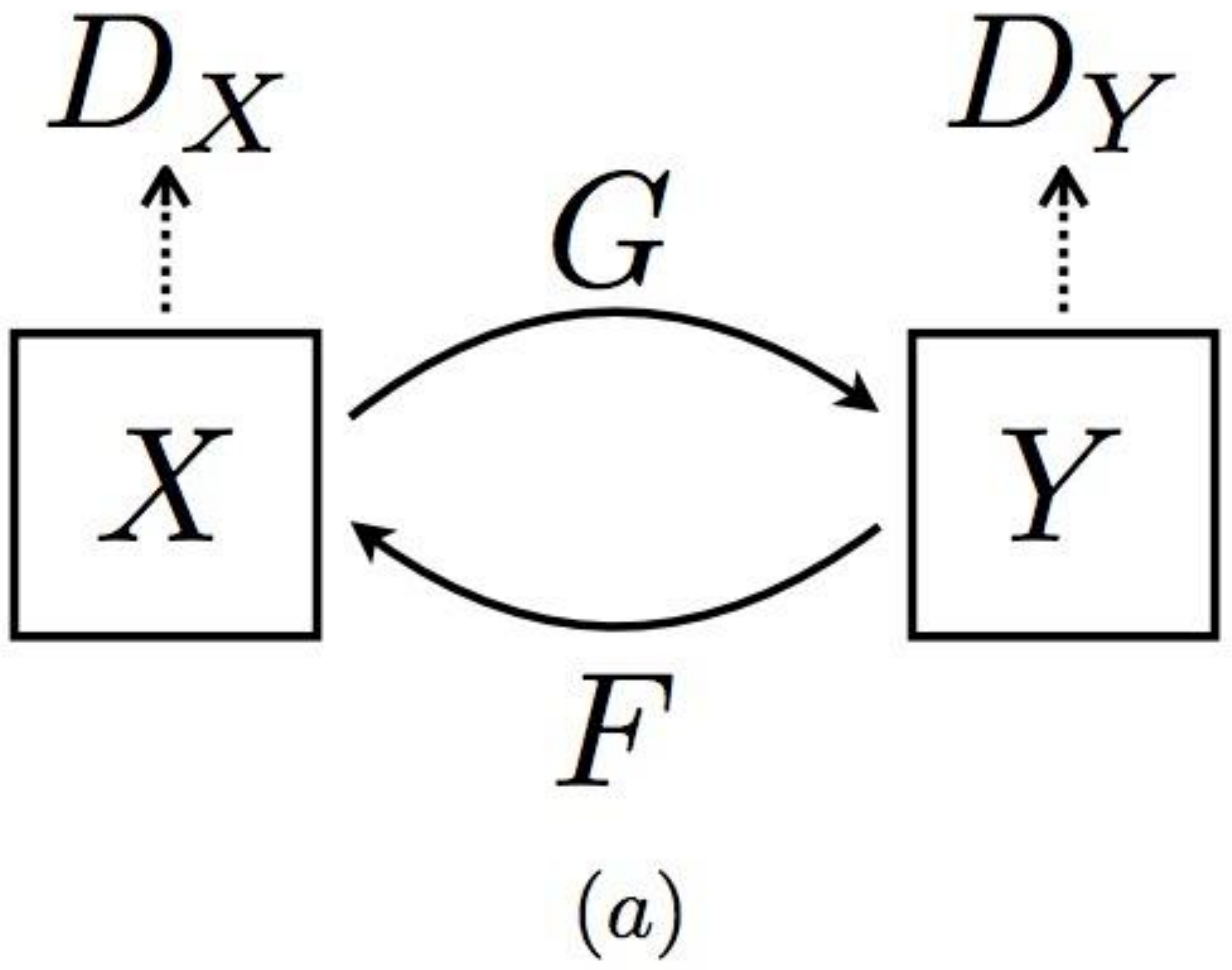
Zhu et al., 2017

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



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

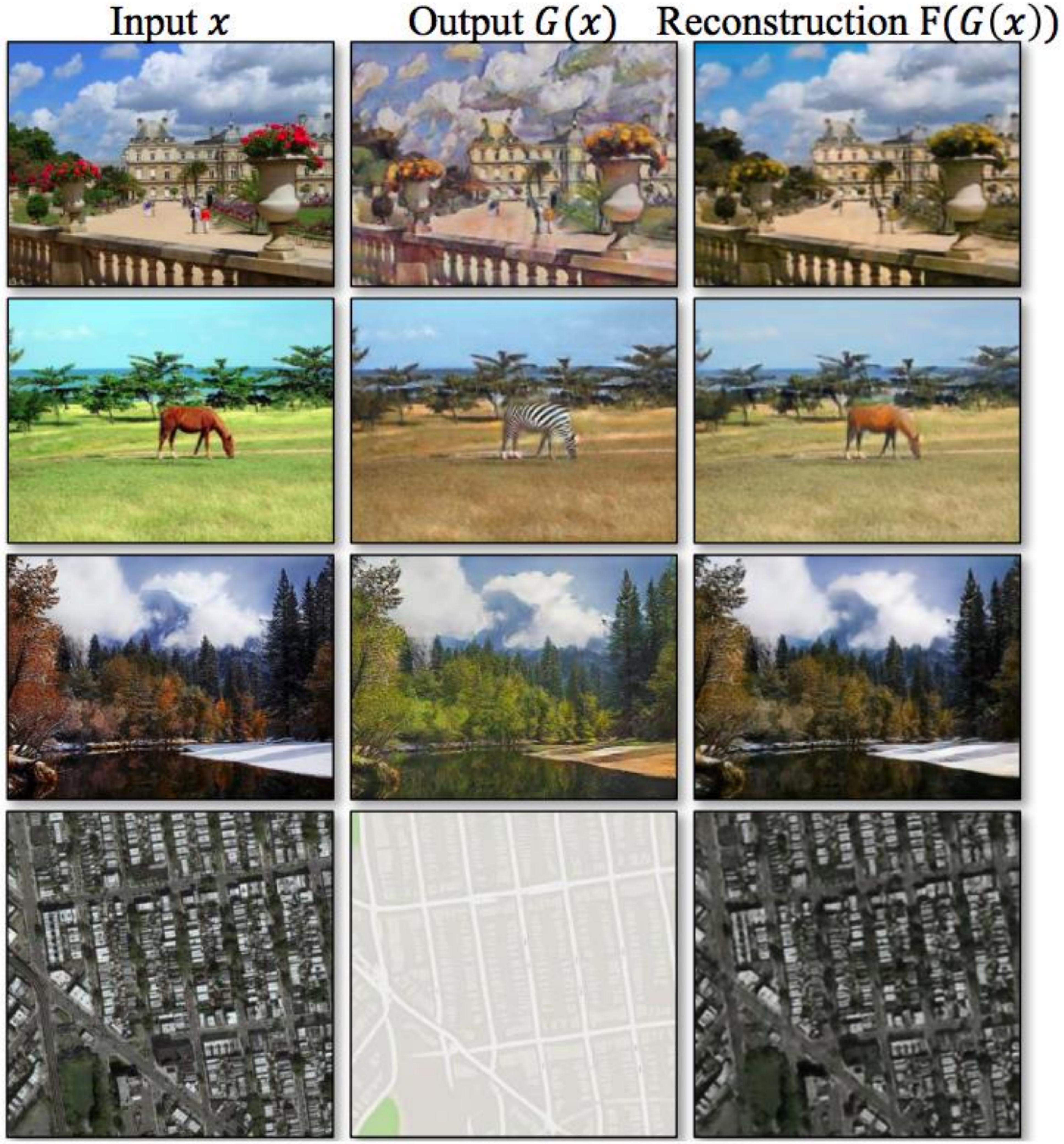
$$\mathcal{L}_{\mathrm{GAN}}(D_Y) = \mathbb{E}_{y\sim p_{\mathrm{data}}(y)}[(D_Y(y)-1)^2] + \mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\left[D_Y\big(G(x)\big)^2\right]$$

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

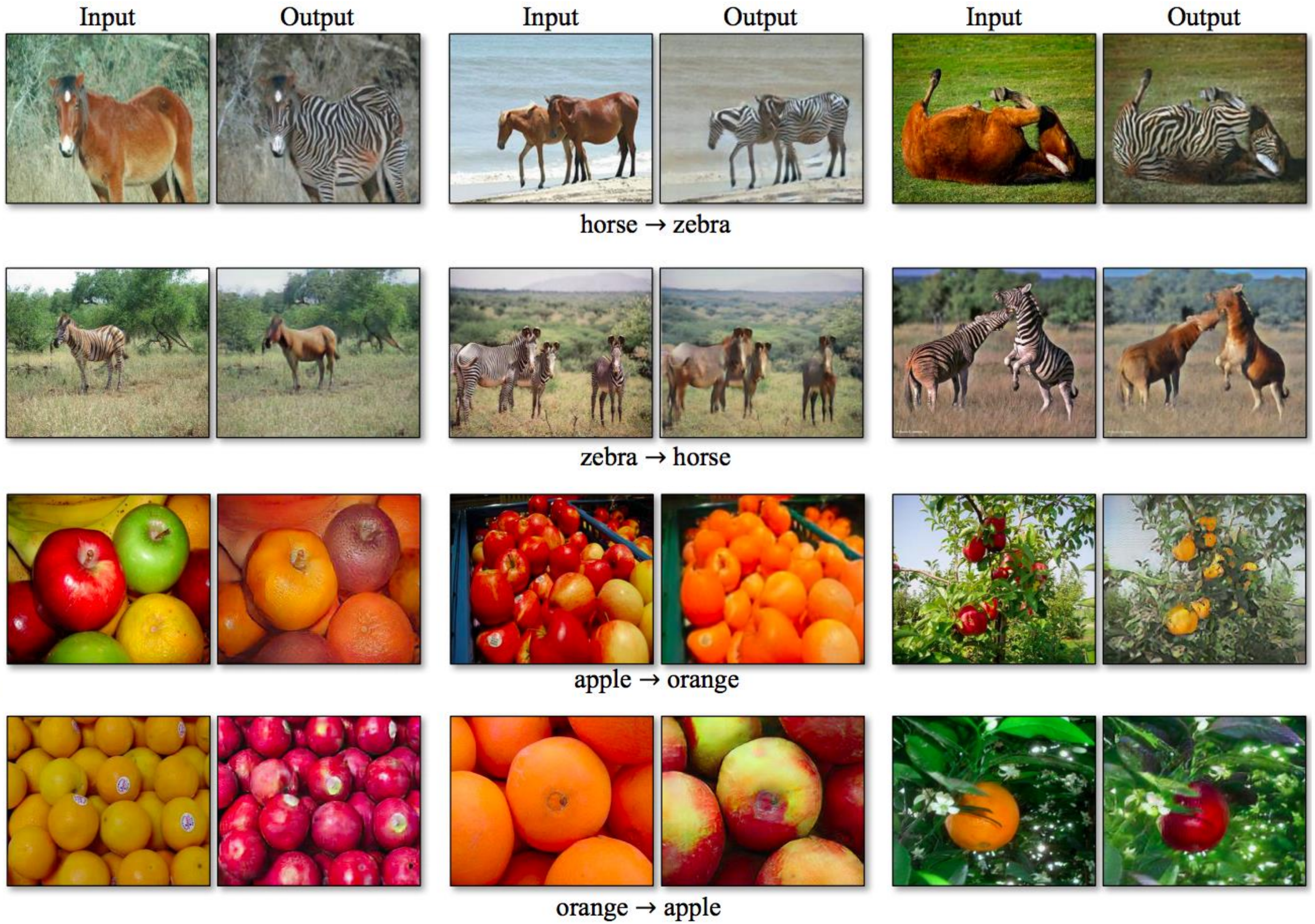
- CycleGAN generator loss:

$$\mathcal{L}_{\mathrm{cyc}}(G,F) = \mathbb{E}_{x\sim p_{\mathrm{data}}(x)}[(D_Y(G(x))-1)^2] + \mathbb{E}_{y\sim p_{\mathrm{data}}(y)}[(D_X(F(y))-1)^2]$$
$$+ \mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\left[\left\|F\big(G(x)\big)-x\right\|_1\right] + \mathbb{E}_{y\sim p_{\mathrm{data}}(y)}\left[\left\|G\big(F(y)\big)-y\right\|_1\right]$$

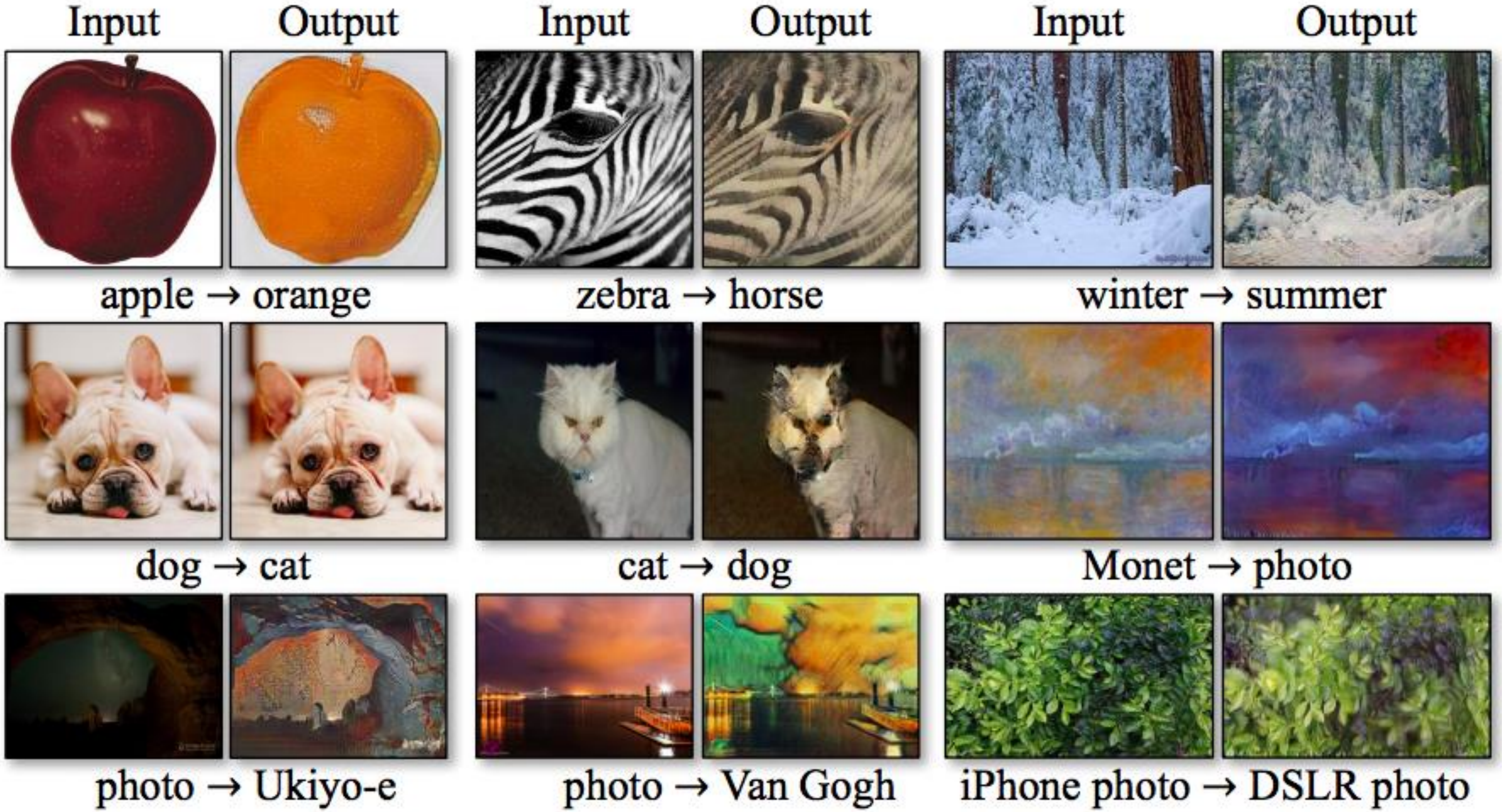
CycleGAN



CycleGAN: Results



CycleGAN: Failure cases



CycleGAN: Failure cases

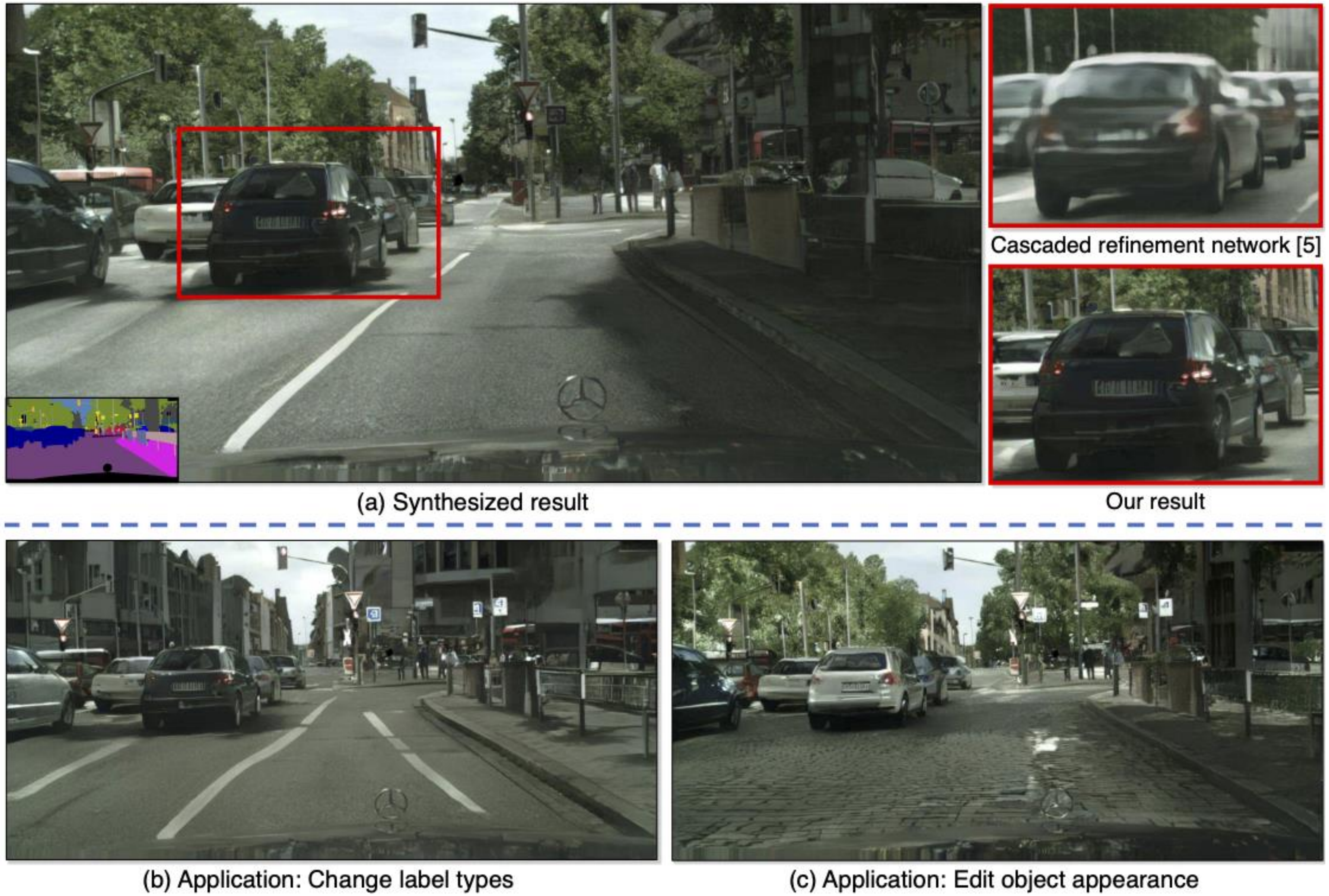
Input Output

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

High-resolution, high-quality pix2pix

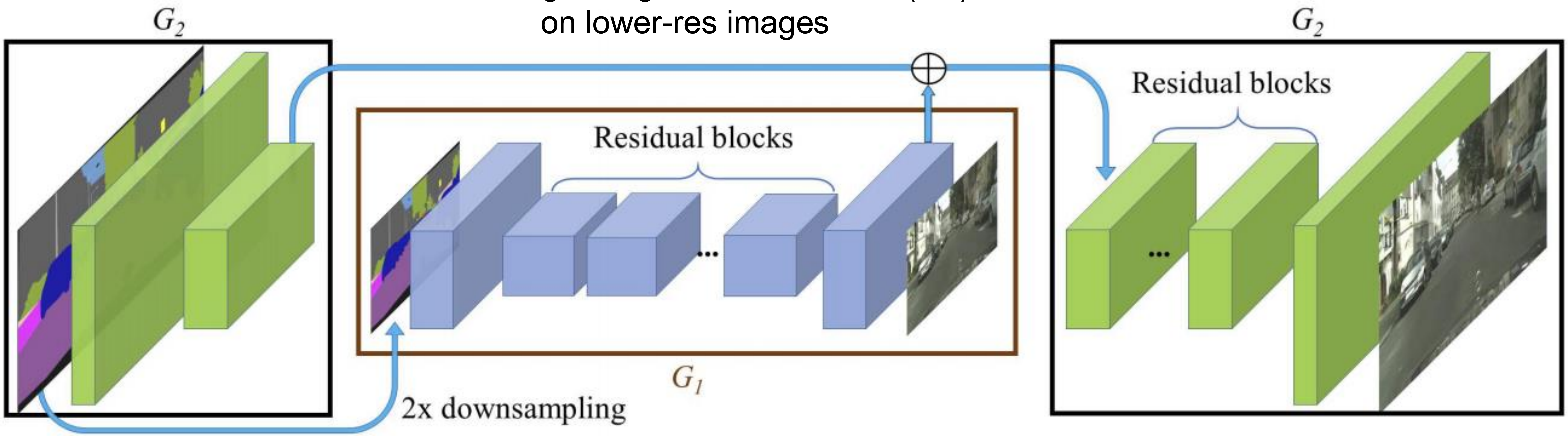


T.-C. Wang et al., [High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs](), CVPR 2018

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

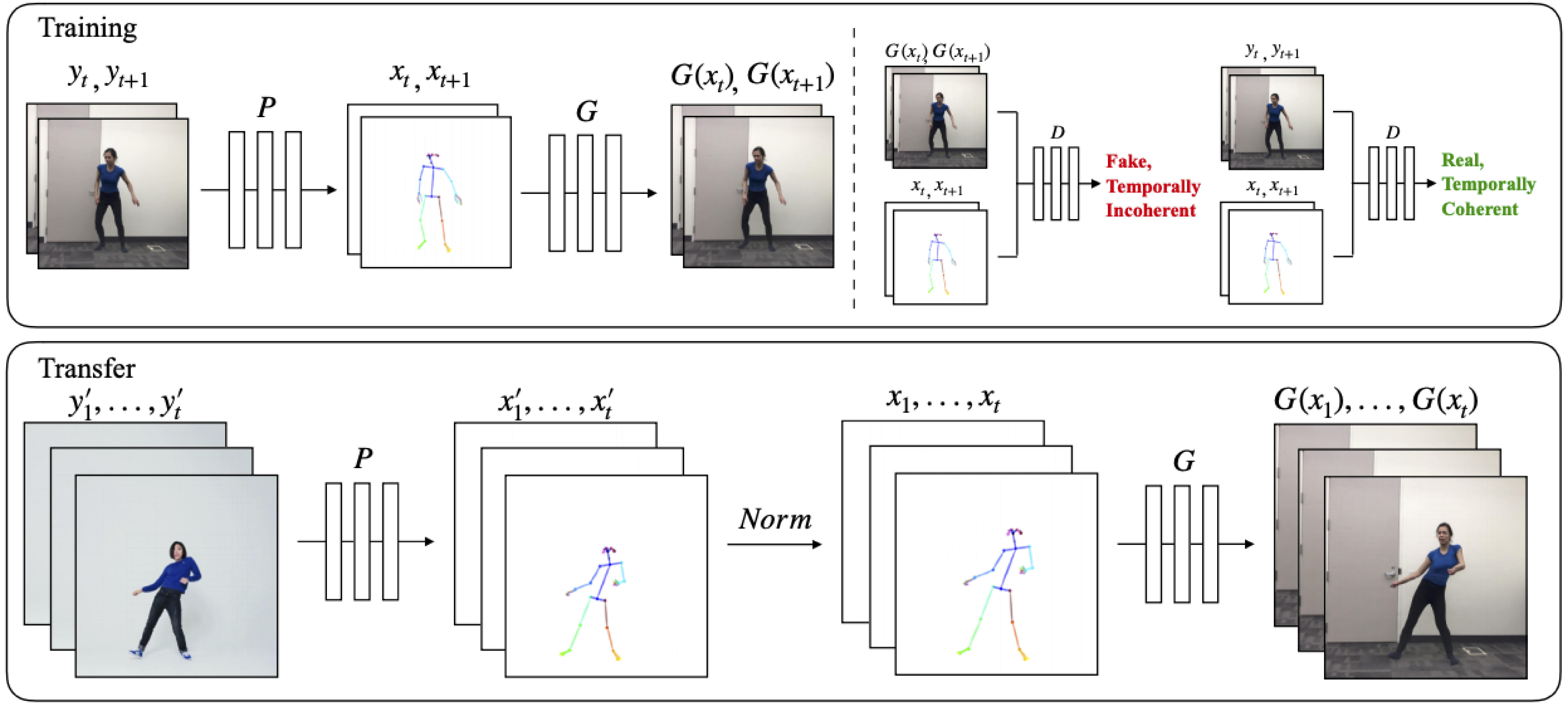
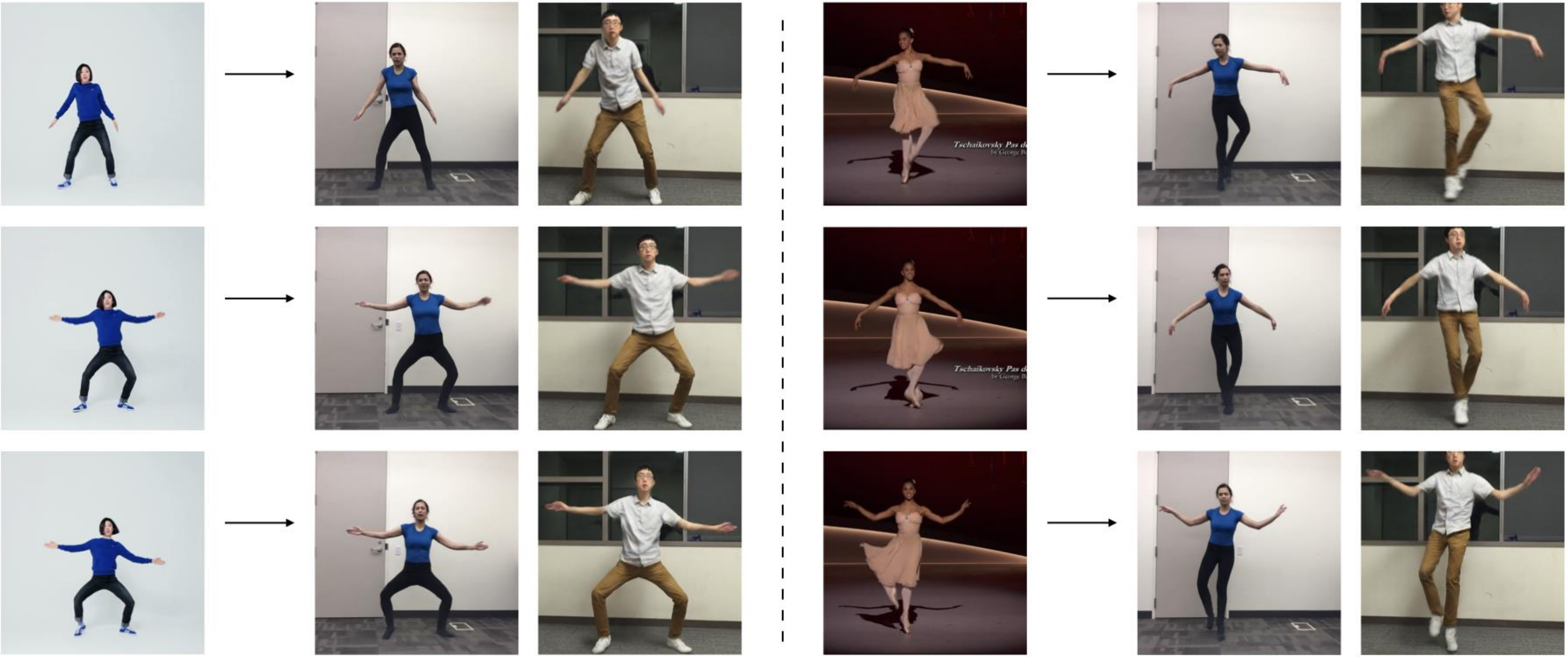


Figure 3: (Top) **Training**: Our model uses a pose detector P to create pose stick figures from video frames of the target subject. We learn the mapping G alongside an adversarial discriminator D which attempts to distinguish between the "real" correspondences $(x_t, x_{t+1}), (y_t, y_{t+1})$ and the "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



Source Subject Target Subject 1 Target Subject 2 Source Subject Target Subject 1 Target Subject 2

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

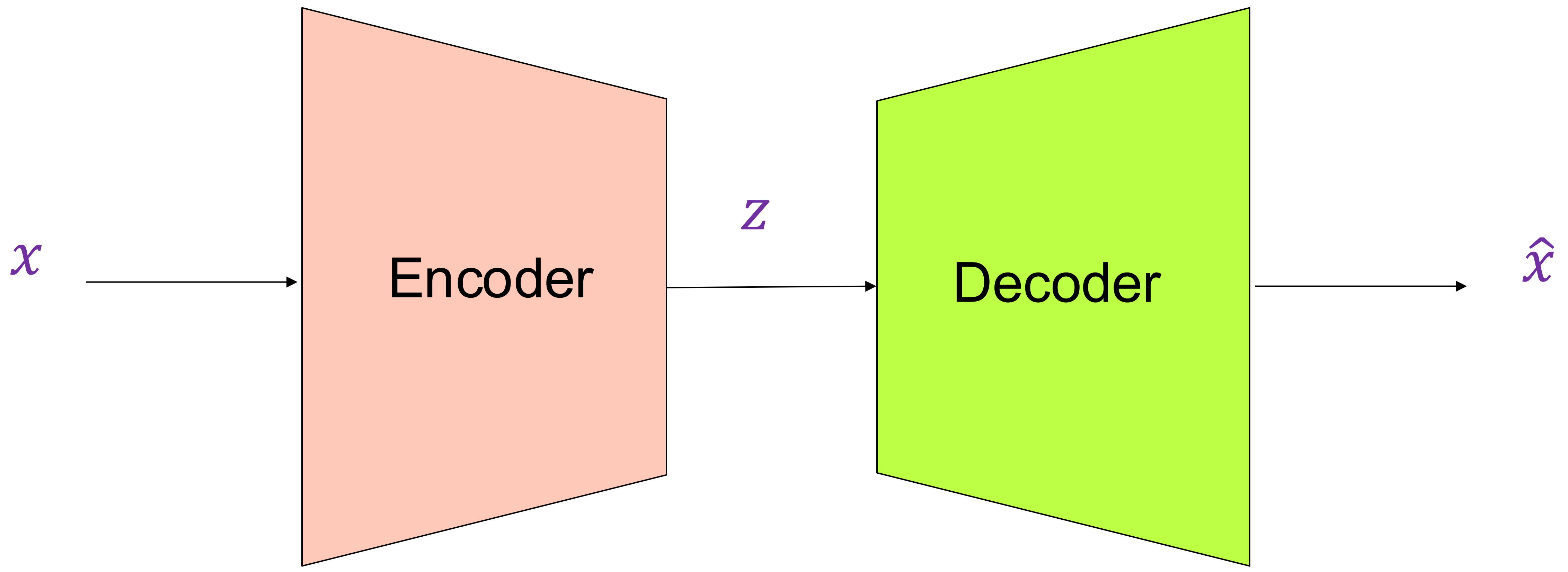
C. Chan, S. Ginosar, T. Zhou, A. Efros. Everybody Dance Now. ICCV 2019


Source Video
Detected Pose

Variational Autoencoder (VAE)

Variational autoencoders: Overview

- Probabilistic formulation based on *variational Bayes* framework
- At training time, jointly learn *encoder* and *decoder* by maximizing (a bound on) the data likelihood
- At test time, discard encoder and use decoder to sample from the learned distribution



D. Kingma and M. Welling, [Auto-Encoding Variational Bayes](), ICLR 2014

Variational autoencoders: Overview

- Probabilistic generative model of the data distribution:

Latent code Data point

$$z \rightarrow x$$

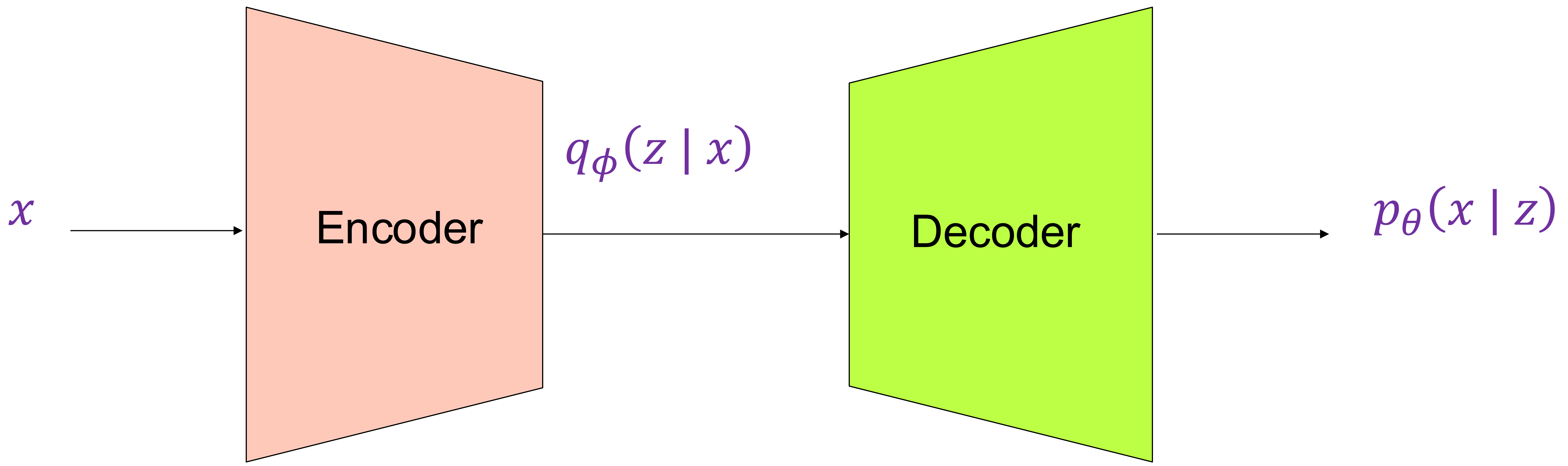
Sample z from *prior* $p(z)$ (usually Gaussian)

Sample x from *conditional* $p(x|z)$

Try to approximate the conditional with neural network

Variational autoencoders: Training

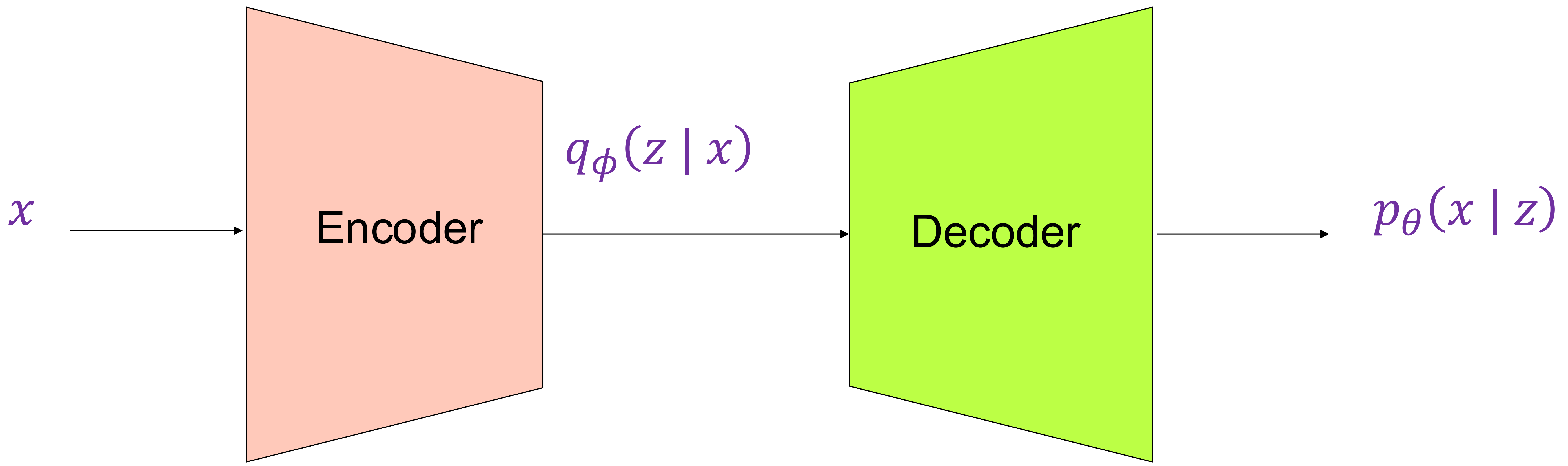
- **Encoder:** given inputs x, output $q_\phi(z \mid x)$
  - Specifically, output mean and (diagonal) covariance, or $\mu_{z|x}$ and $\Sigma_{z|x}$, so that $q_\phi(z \mid x) = N(\mu_{z|x}, \Sigma_{z|x})$
  - Approximate $q_\phi(z \mid x)$ to $N(0, I)$
- **Decoder:** given z, which is sampled from $q_\phi(z \mid x)$, then output $p_\theta(x \mid z)$



Variational autoencoders: Training

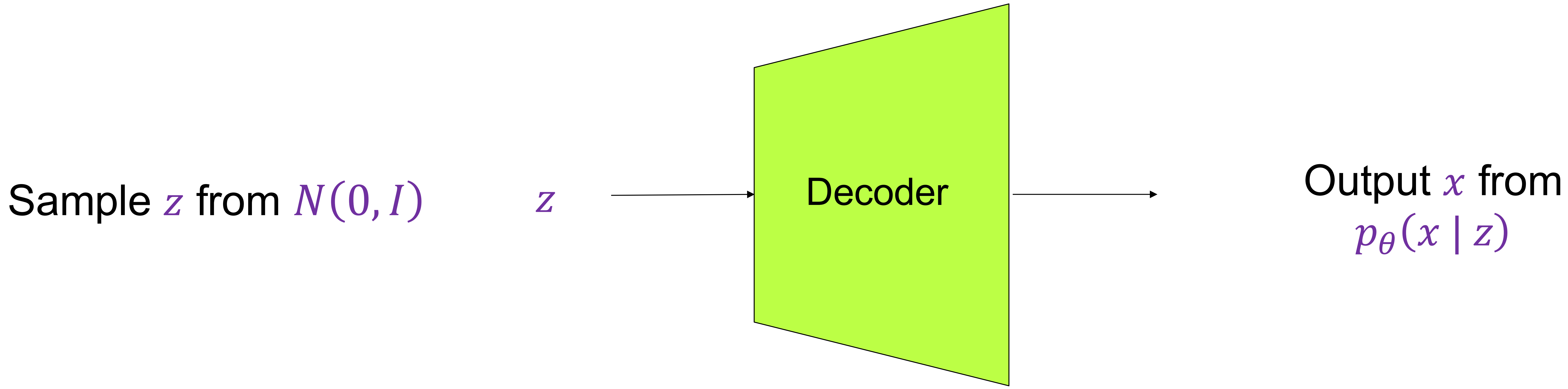
- Objective: maximize the *variational lower bound* on the data likelihood:

$$\log p_\theta(x) \;\geq\; \mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] \;-\; D_{KL}\left(q_\phi(z|x) \parallel N(0,I)\right)$$



Variational autoencoders: Testing

- At test time, discard encoder and use decoder to sample z from $N(0, I)$ and obtain output $p_\theta(x \mid z)$



Variational autoencoders: Generating data

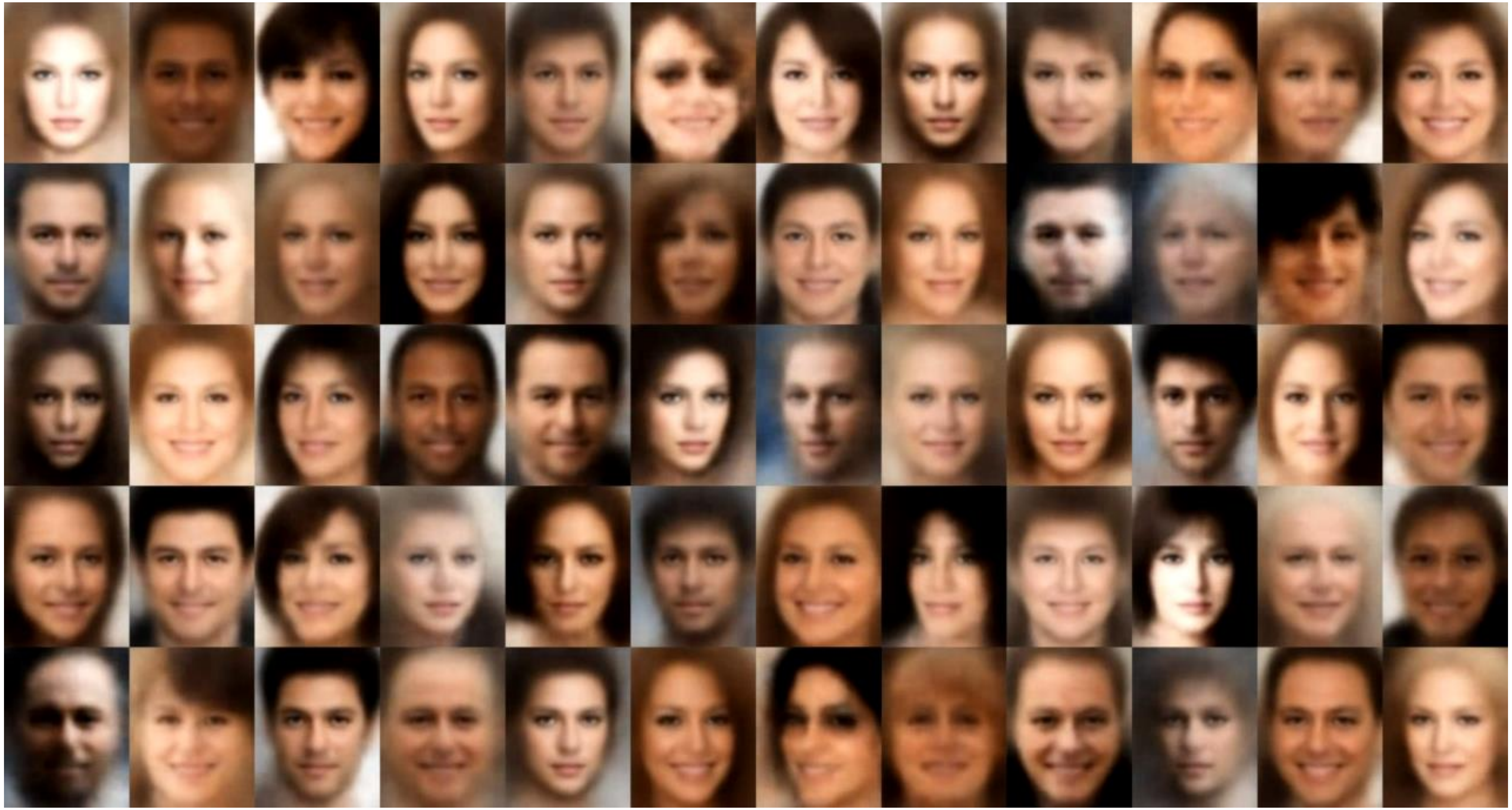
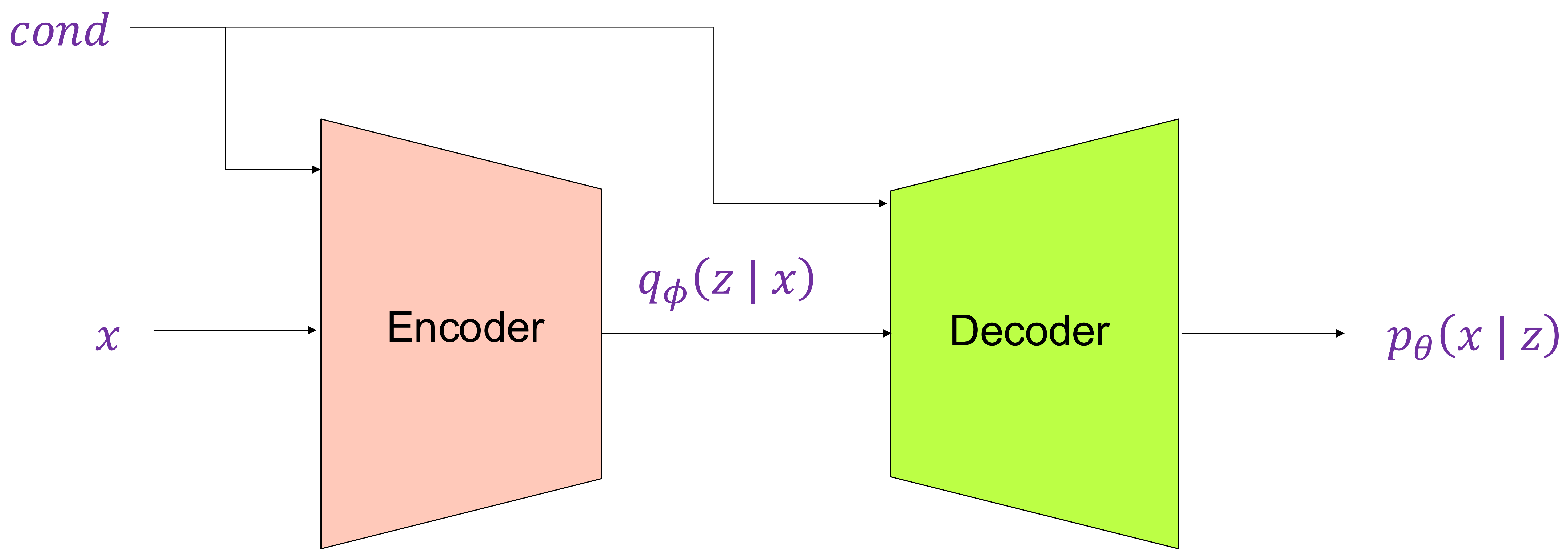
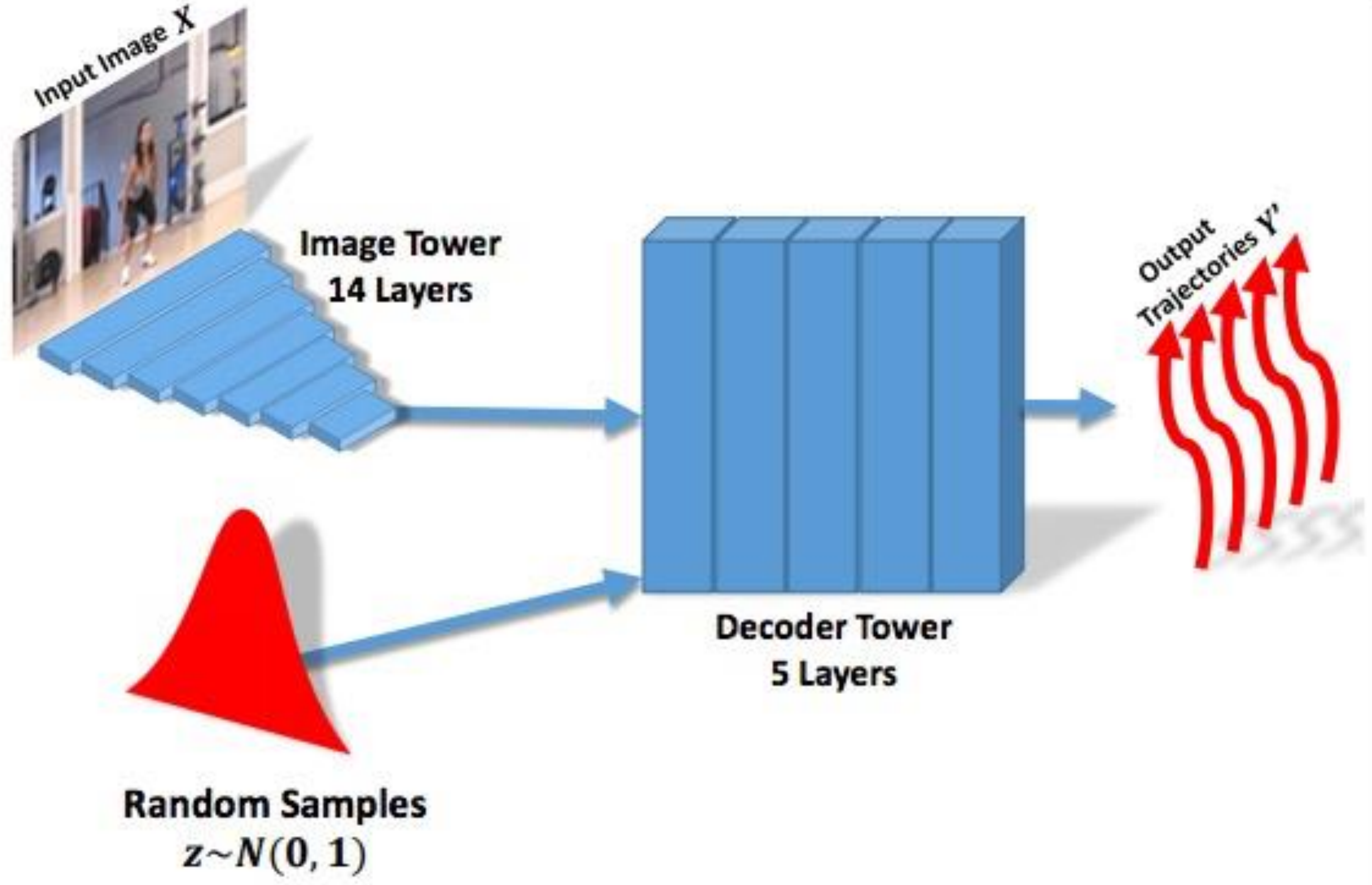


Image source

Conditional VAE



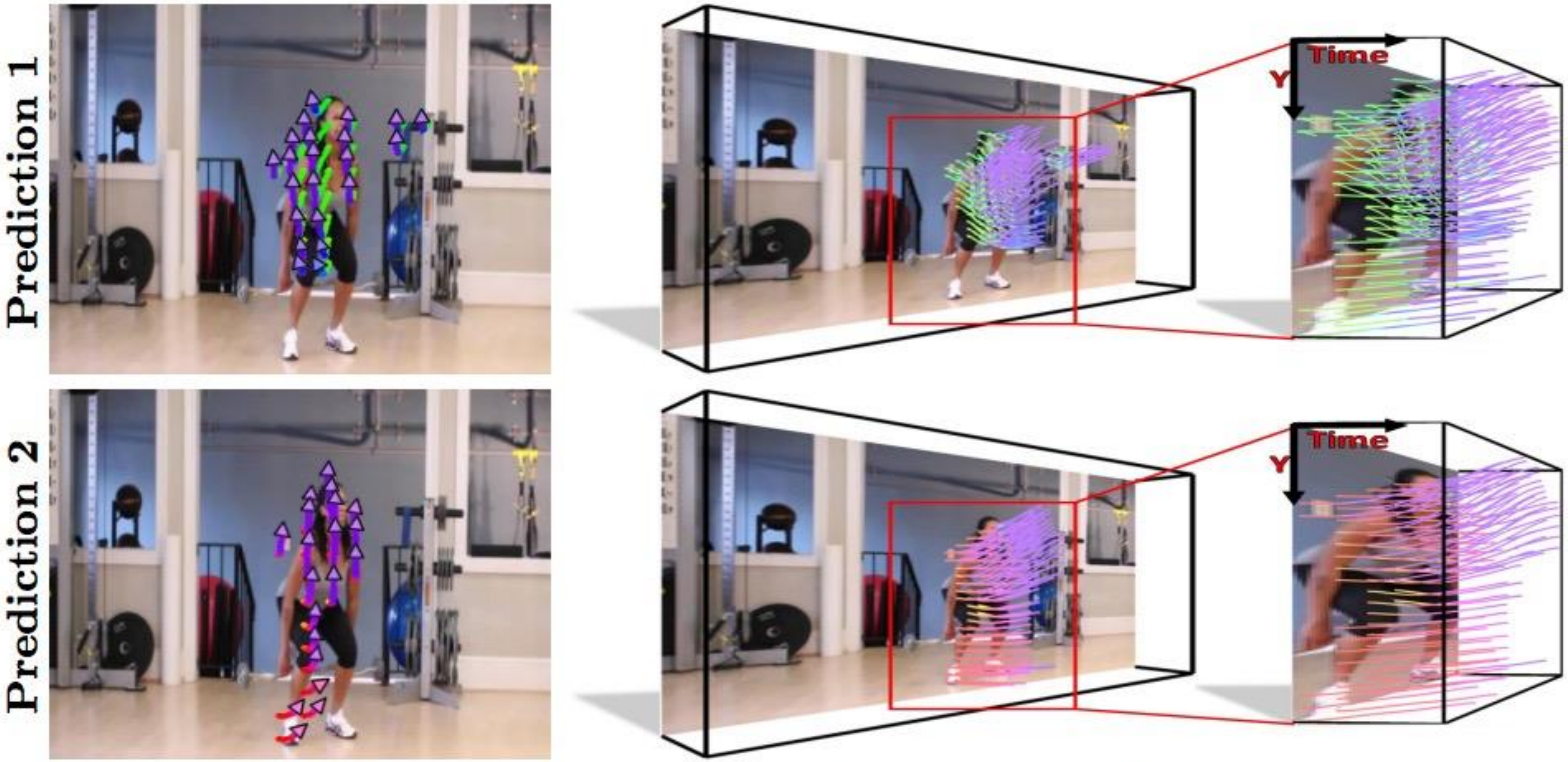
Conditional VAE for Video Prediction



(a) Testing Architecture

Walker et al. 2016.

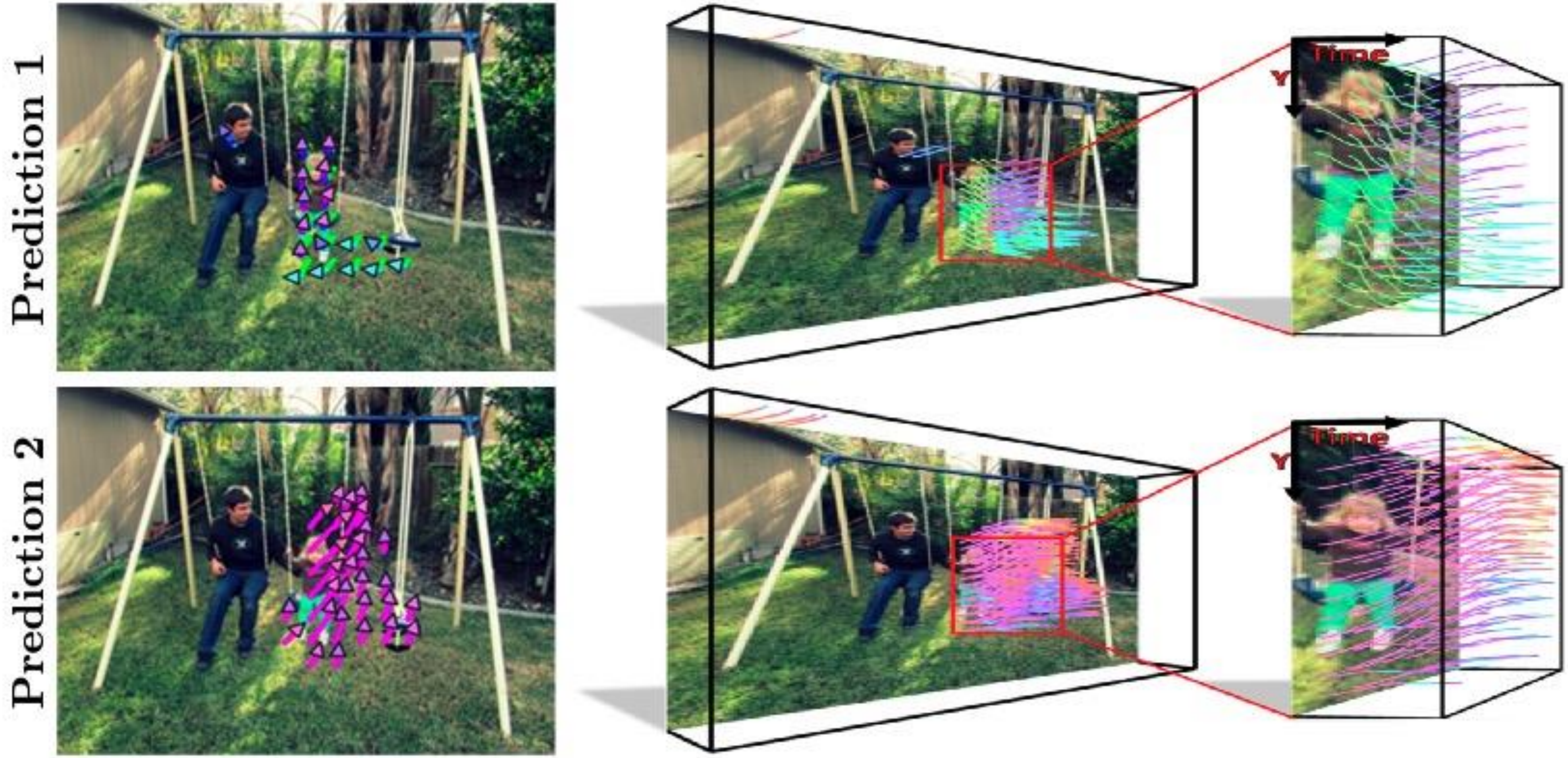
Conditional VAE for Video Prediction



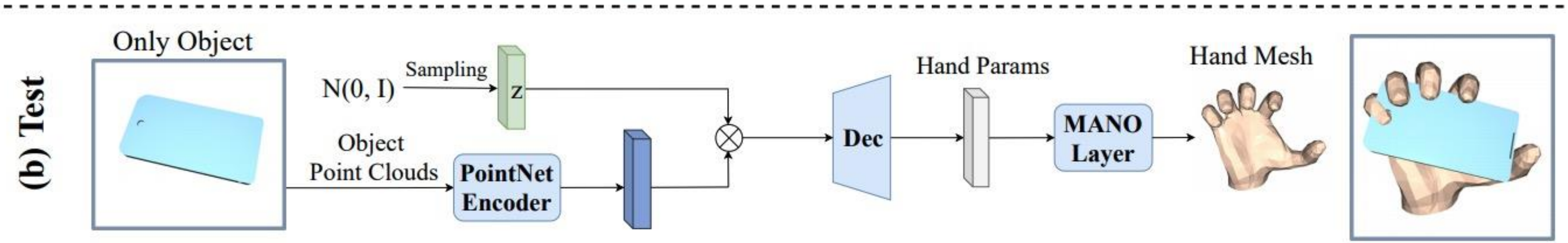
(a) Trajectories on Image (b) Trajectories in Space-Time

Walker et al. An Uncertain Future: Forecasting from Static Images Using Variational Autoencoders. 2016.

Conditional VAE for Video Prediction

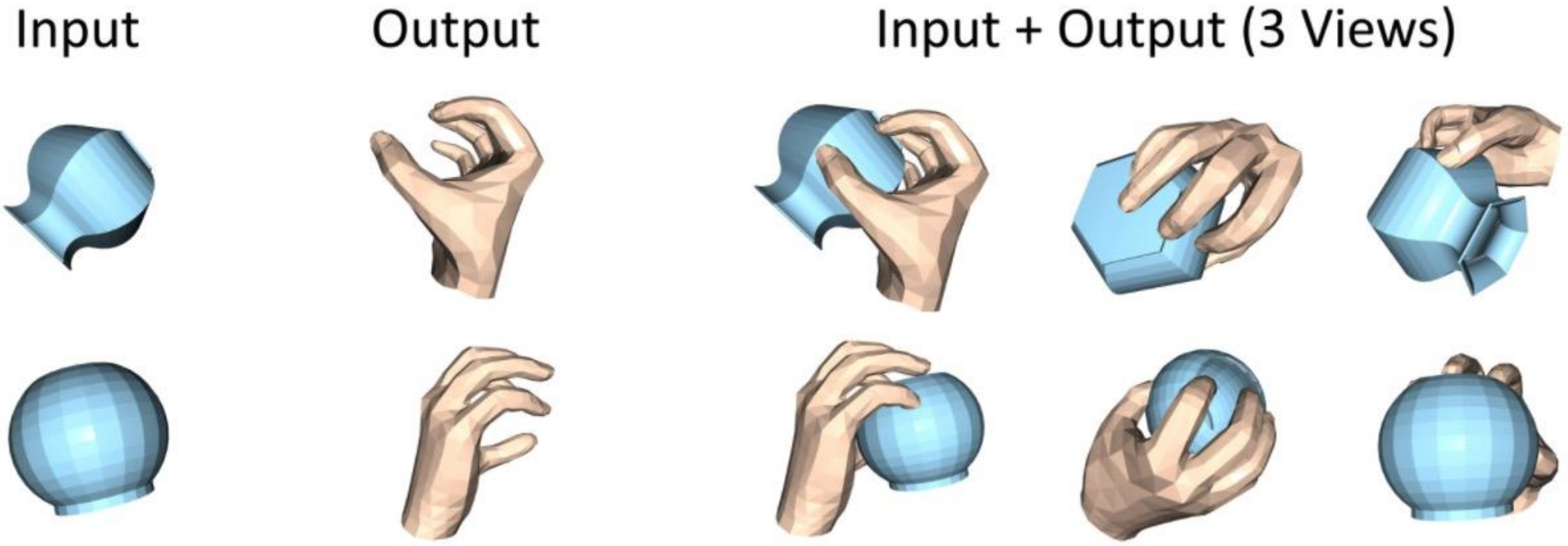


Conditional VAE for Grasp Generation



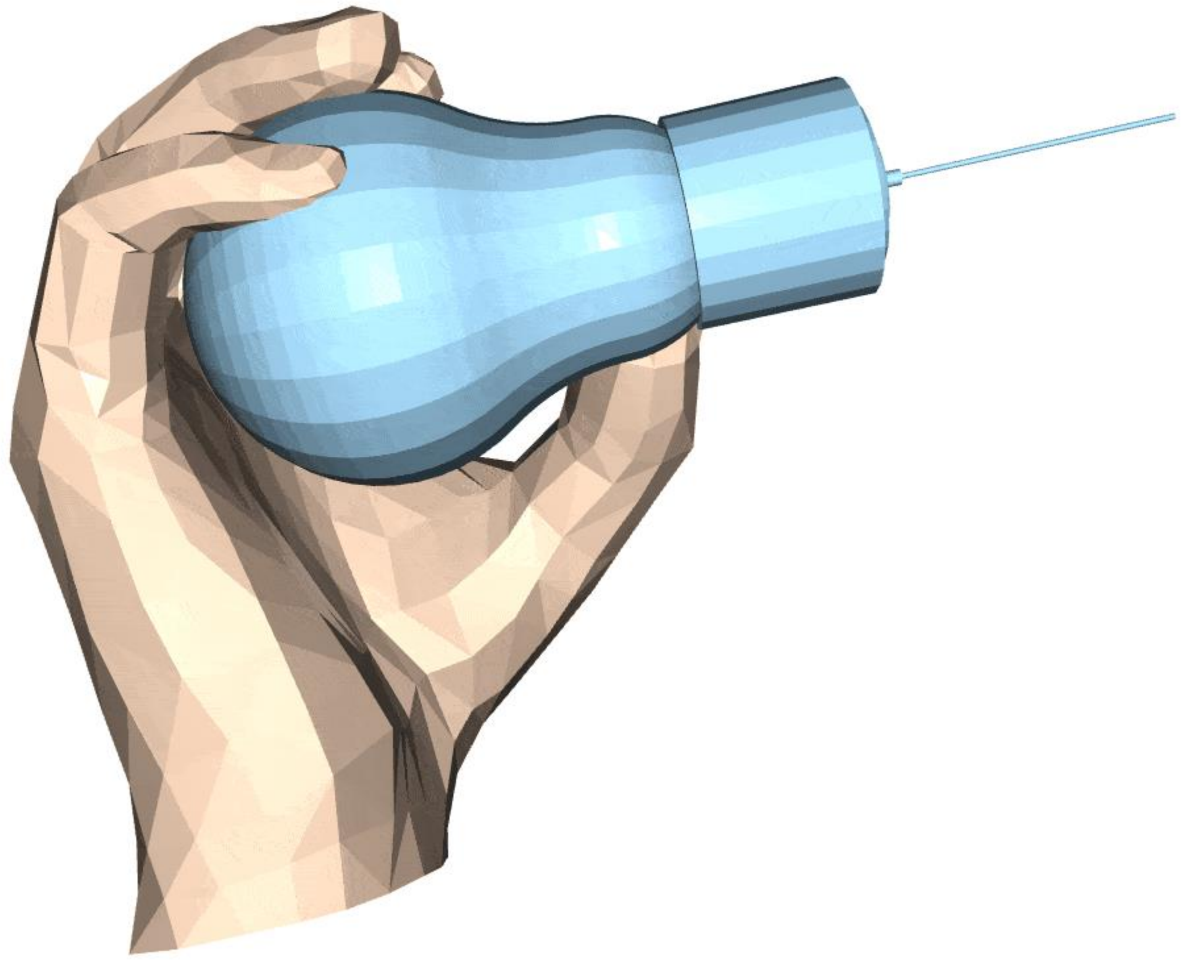
Jiang et al. 2021.

Conditional VAE for Grasp Generation



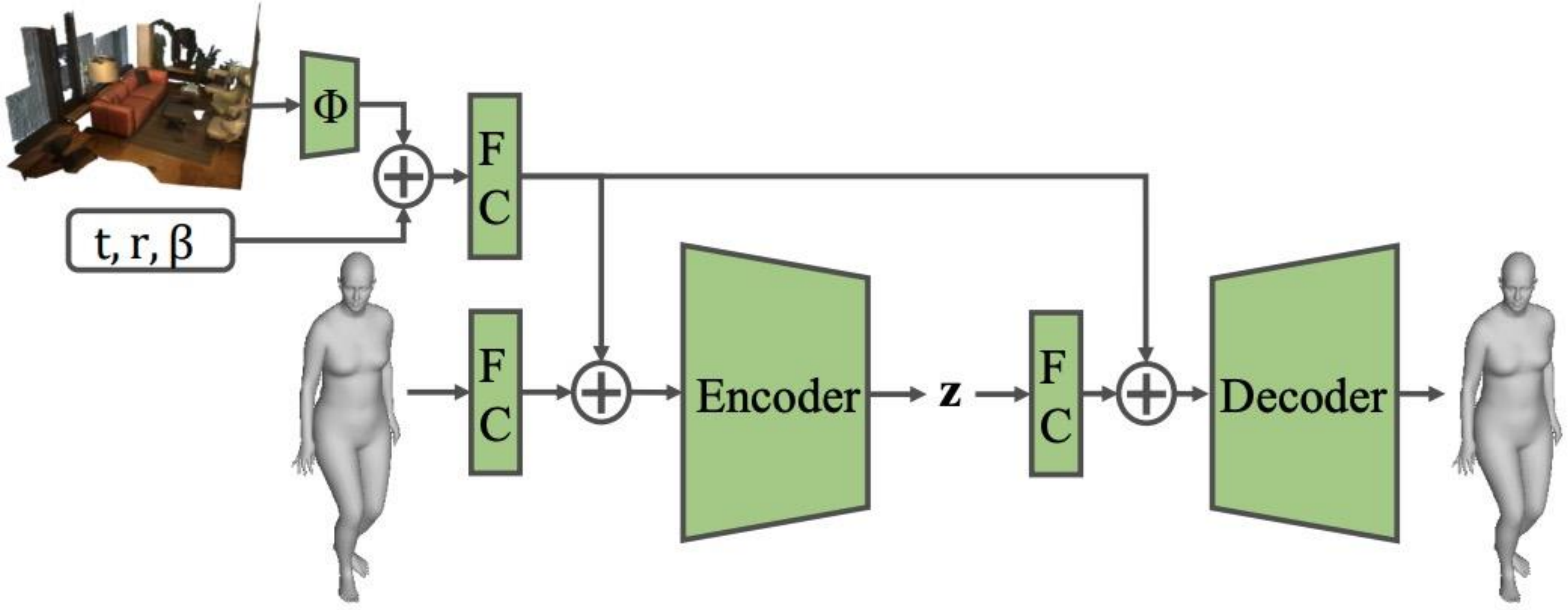
Jiang et al. 2021.

Conditional VAE for Grasp Generation



Jiang et al. 2021.

Conditional VAE for Human Motion Synthesis

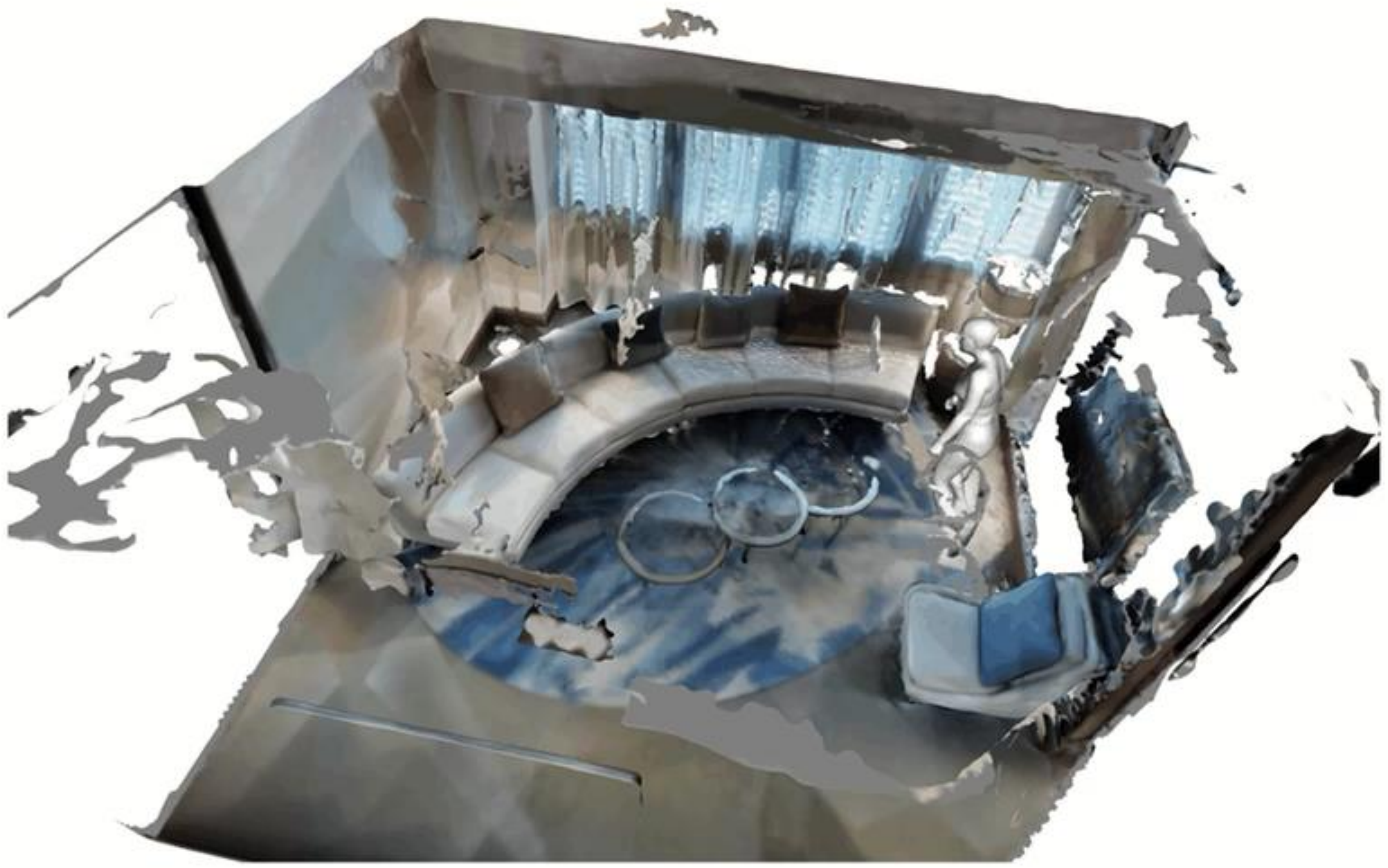


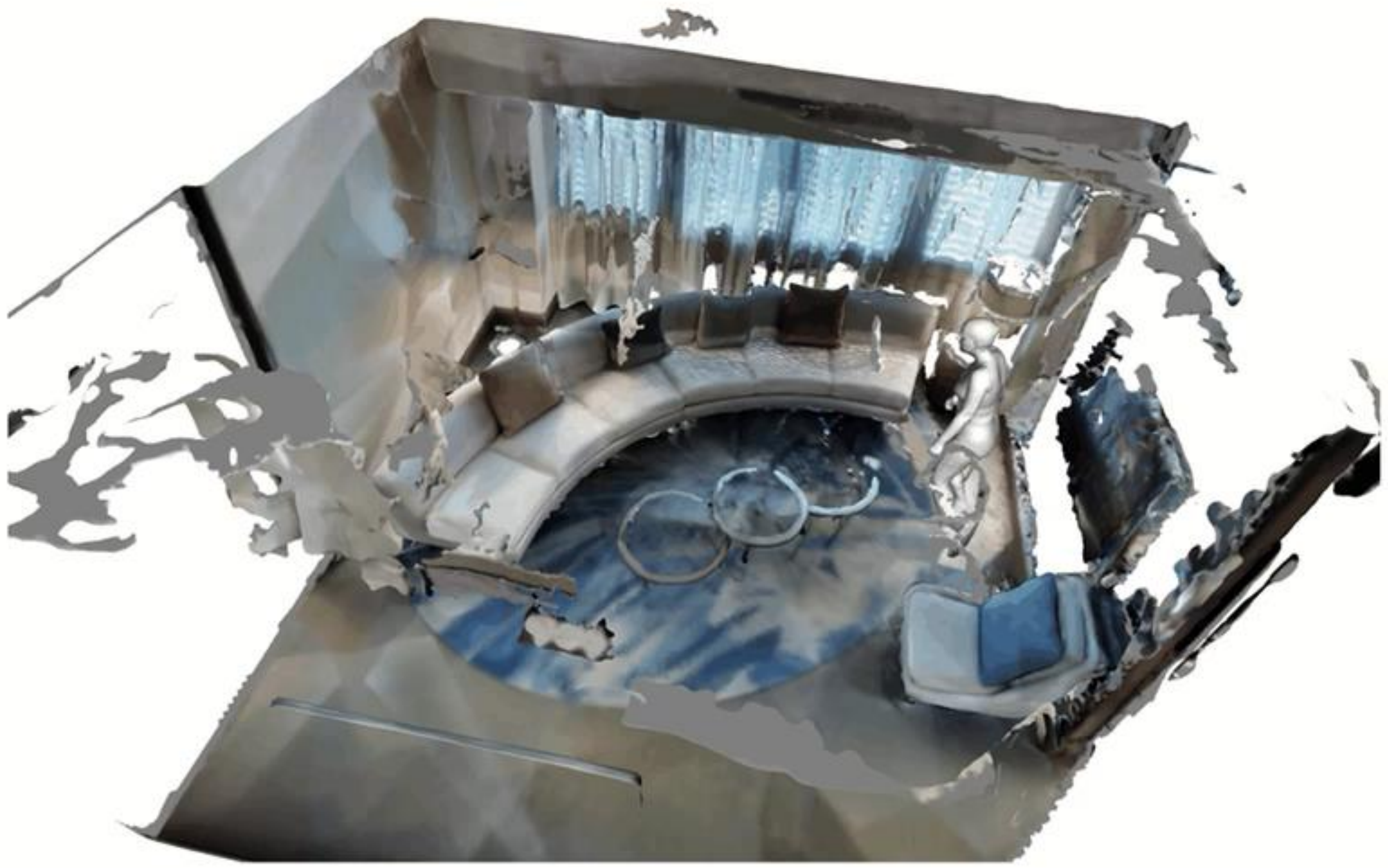
Wang et al. 2021.

Conditional VAE for Human Motion Synthesis

Wang et al. 2021.

Conditional VAE for Human Motion Synthesis





Wang et al. 2021.

Summary

- Image-to-Image Translation: pix2pix
- Unpaired Image-to-Image Translation: CycleGAN
- Variational Autoencoder (VAE)