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

Presenter: Xiao Fang

*Acknowledgement: Many of the following slides are adapted from slides by Junyan Zhu.

Image-to-Image Translation

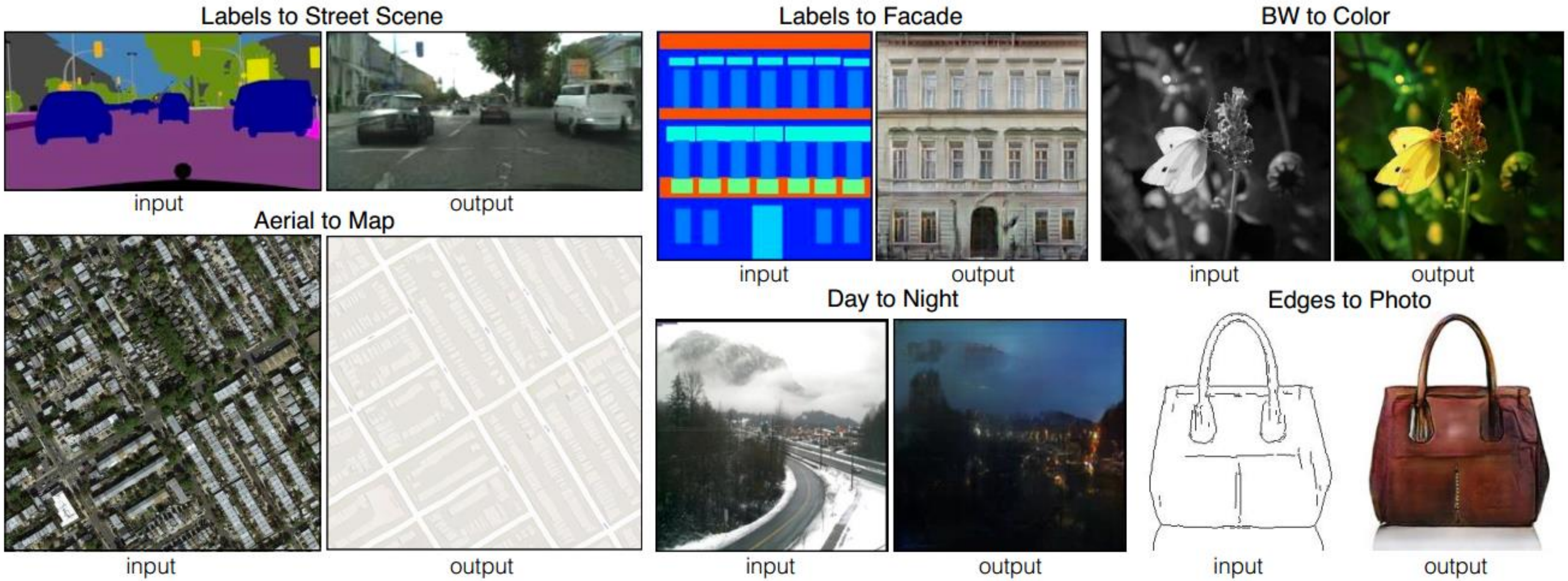
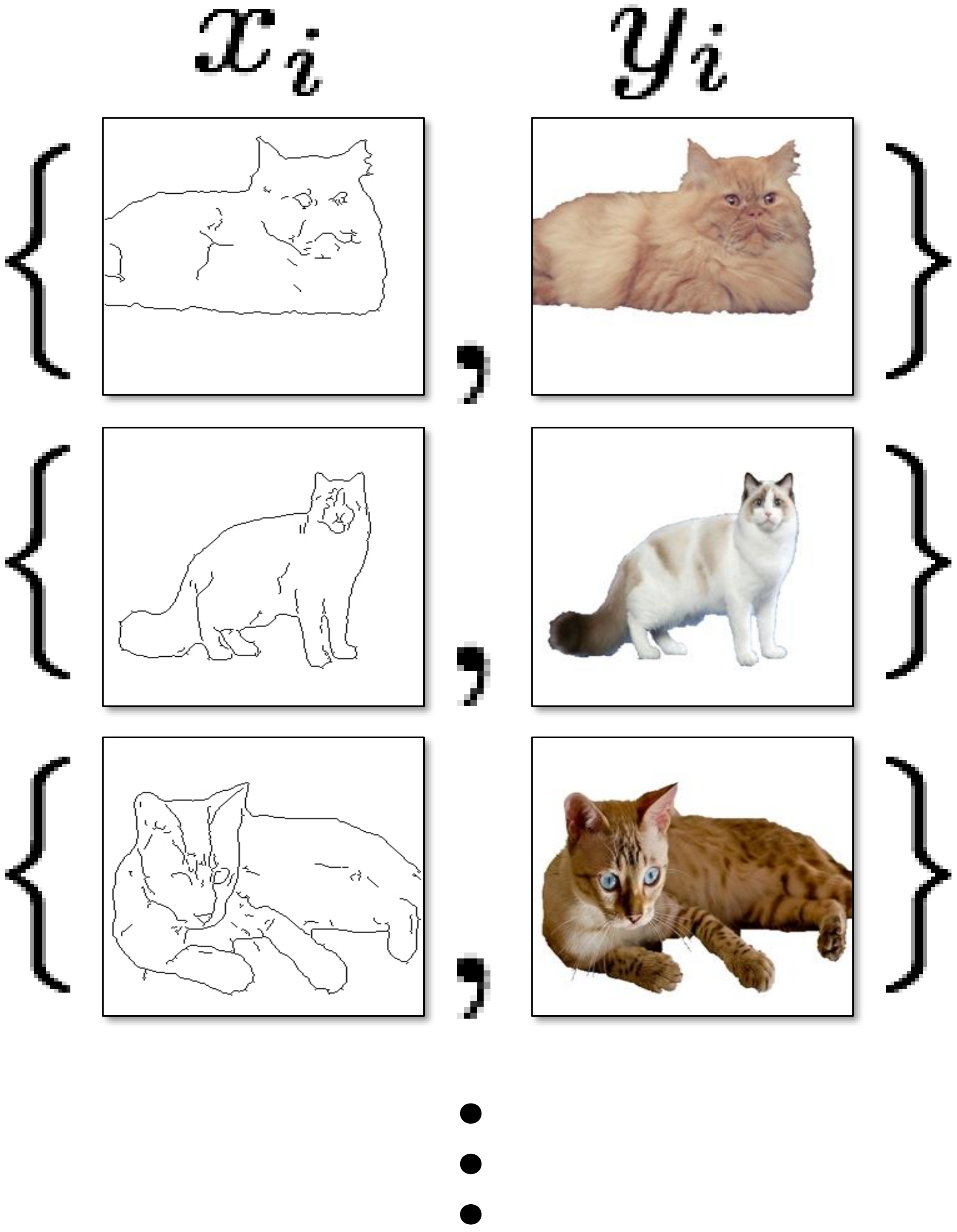
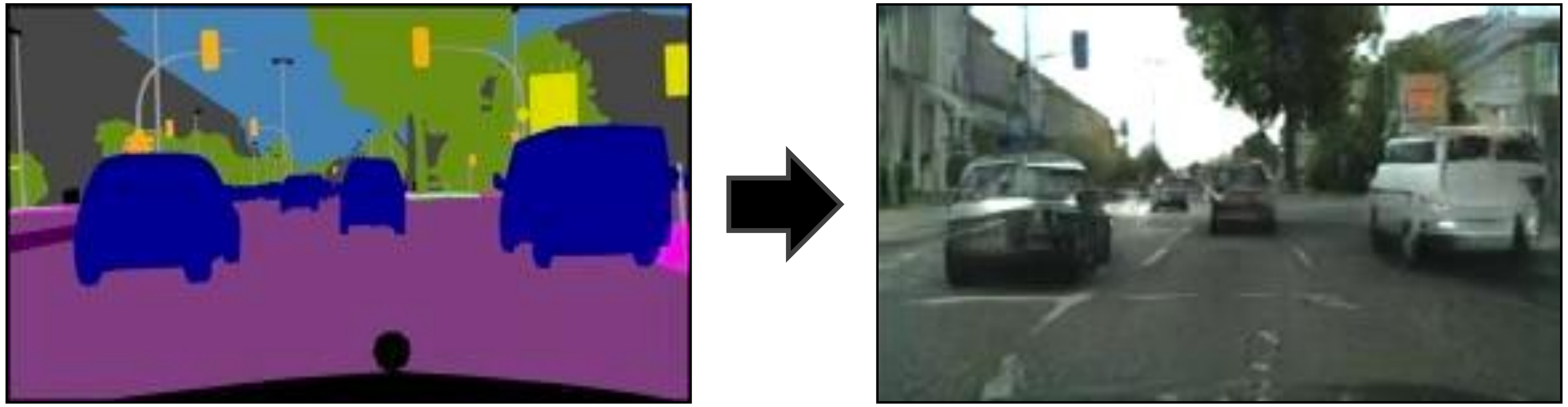


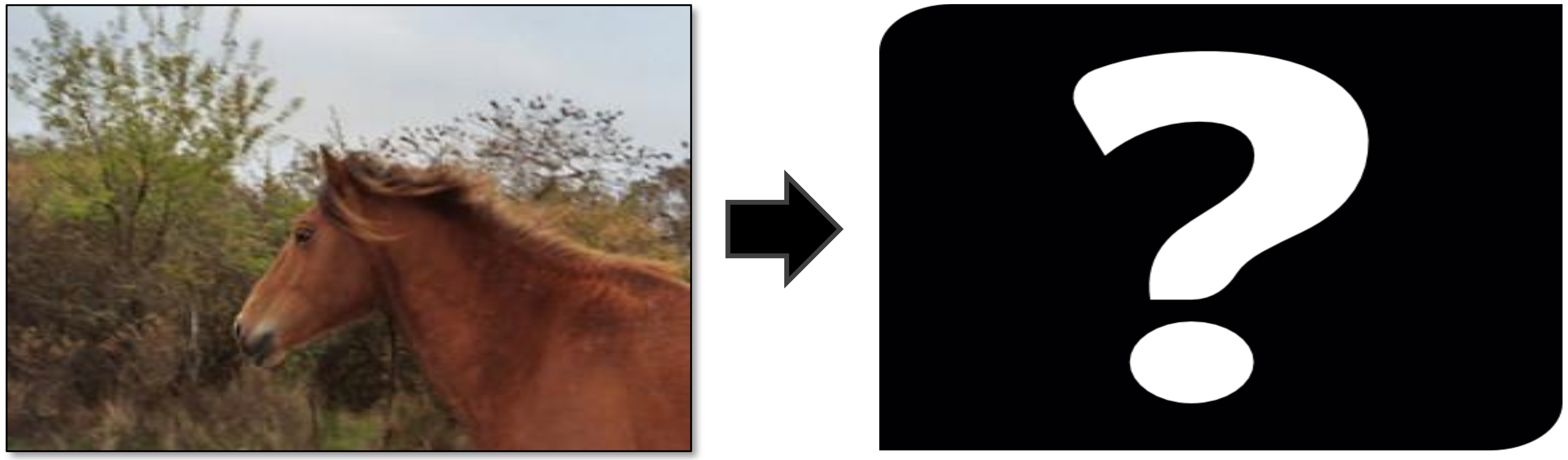
Image-to-Image Translation with Conditional Adversarial Networks, CVPR 2017

Paired



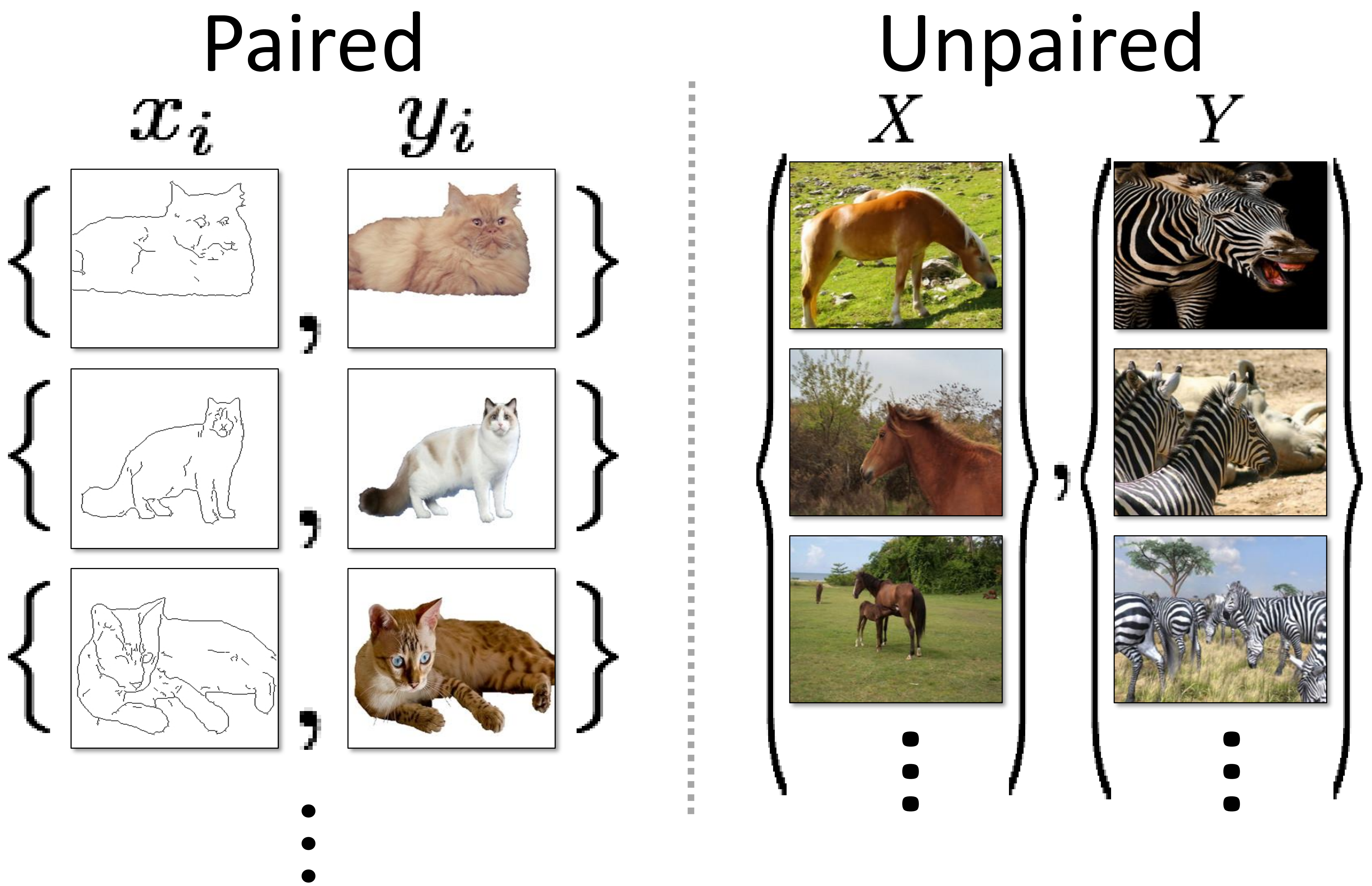


Label ↔ photo: per-pixel labeling

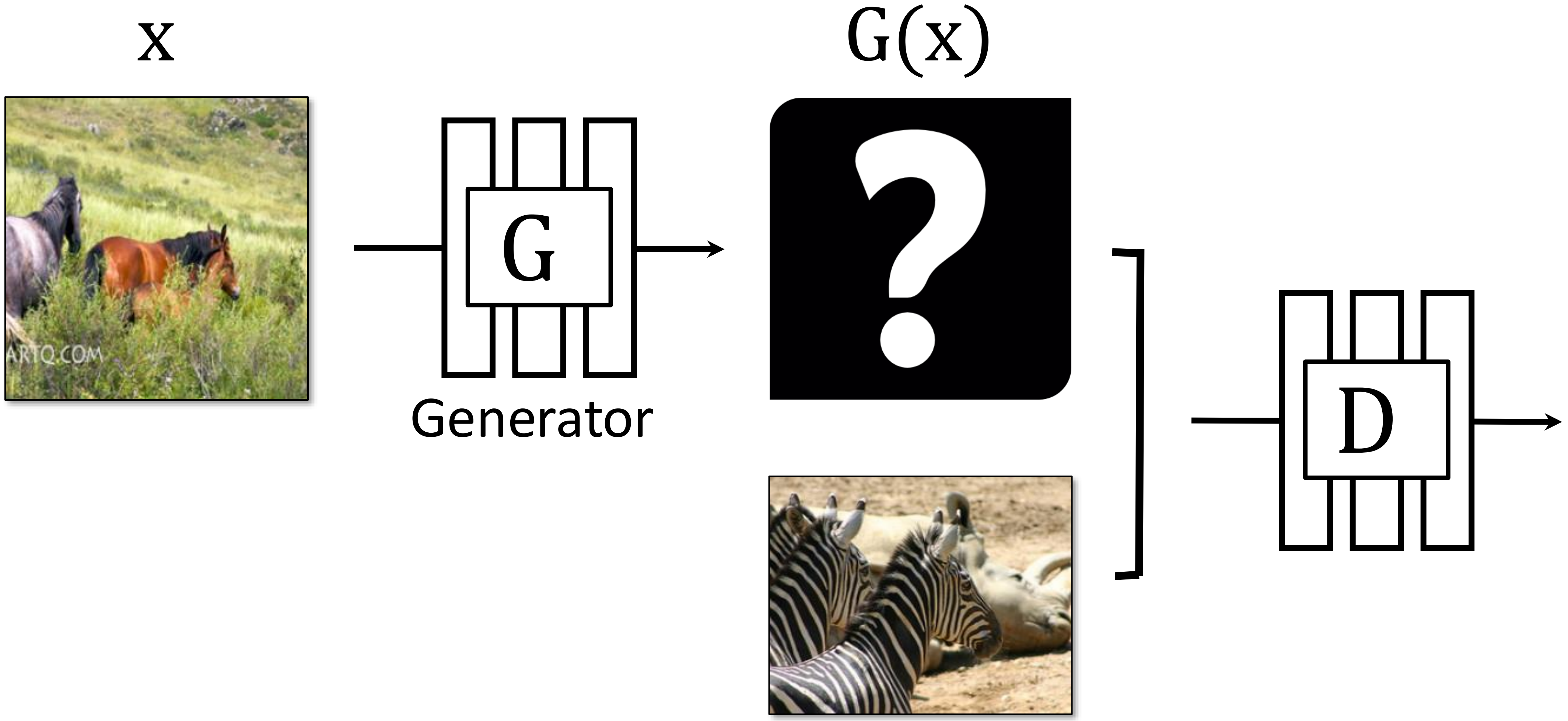


Horse ↔ zebra: how to get zebras?

- Expensive to collect pairs.
- Impossible in many scenarios.

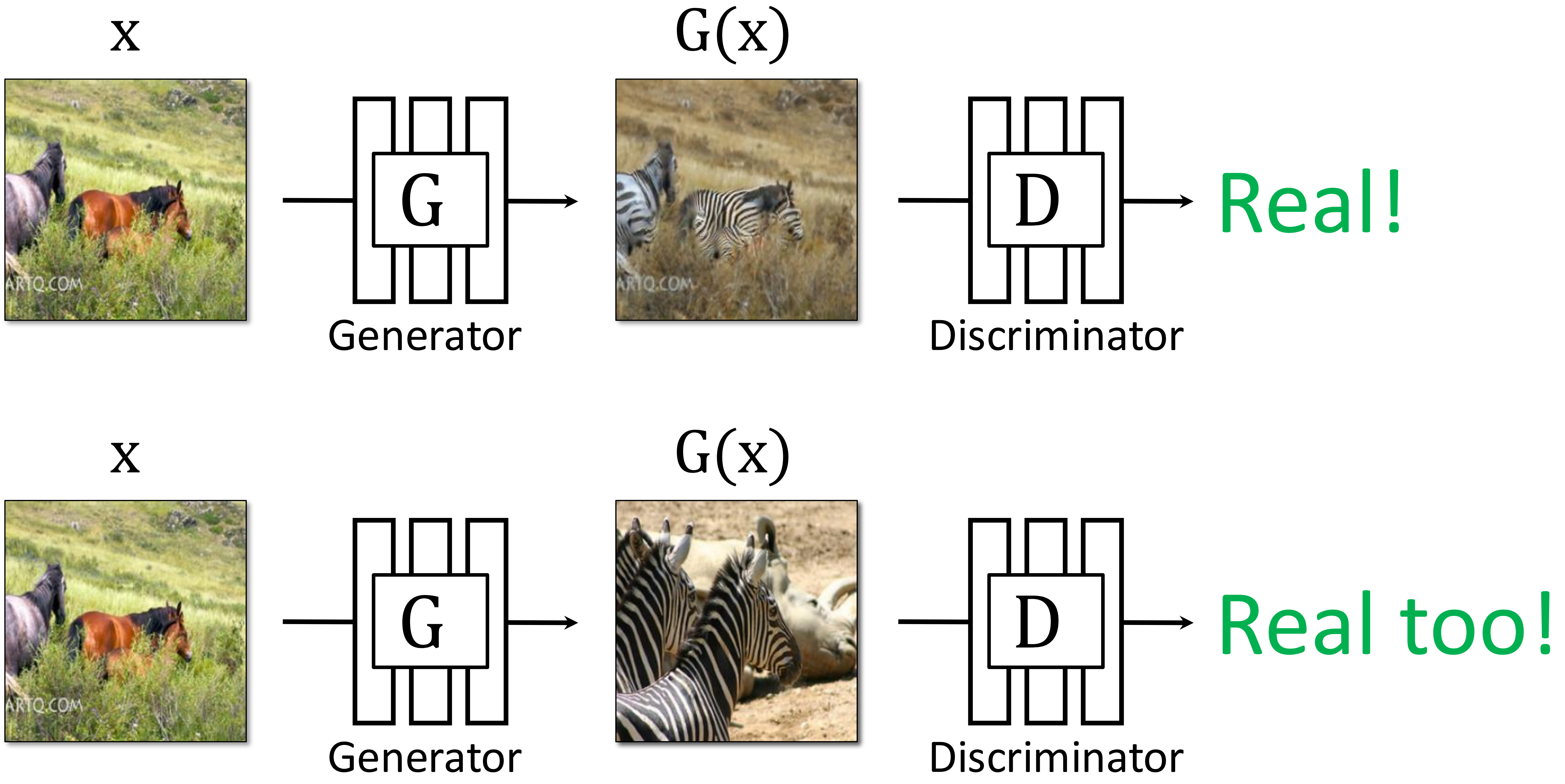

Paired
x_i
y_i
Unpaired
X
Y

Generative Adversarial Networks

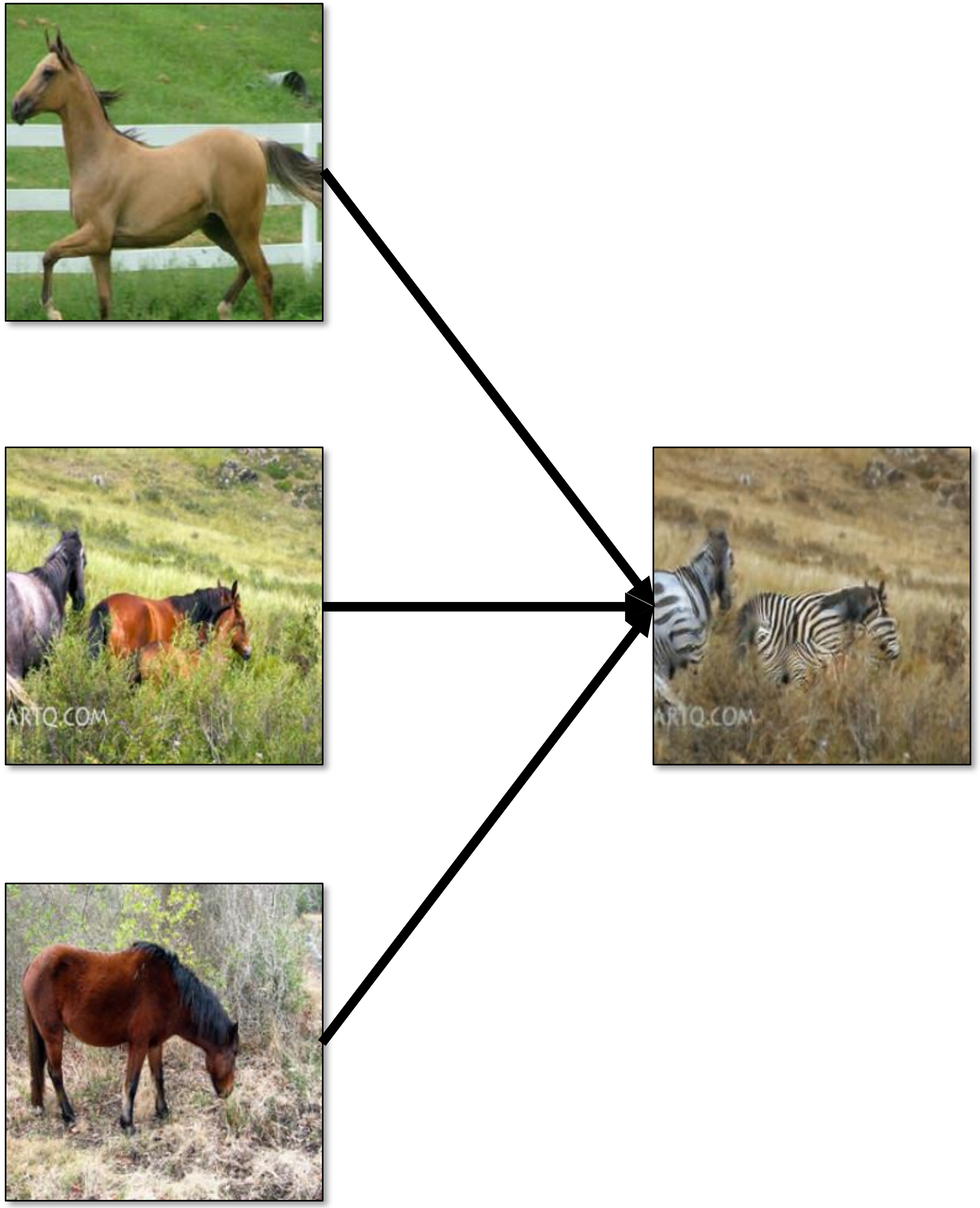


$$\mathcal{L}_{D_Y} = -\mathbb{E}_{y\sim p_{\text{data}}(y)}\left[\log D_Y(y)\right] - \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log\left(1 - D_Y\left(G_{X\to Y}(x)\right)\right)\right]$$

$$\mathcal{L}_{G_{X\to Y}} = -\mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log D_Y\left(G_{X\to Y}(x)\right)\right]$$

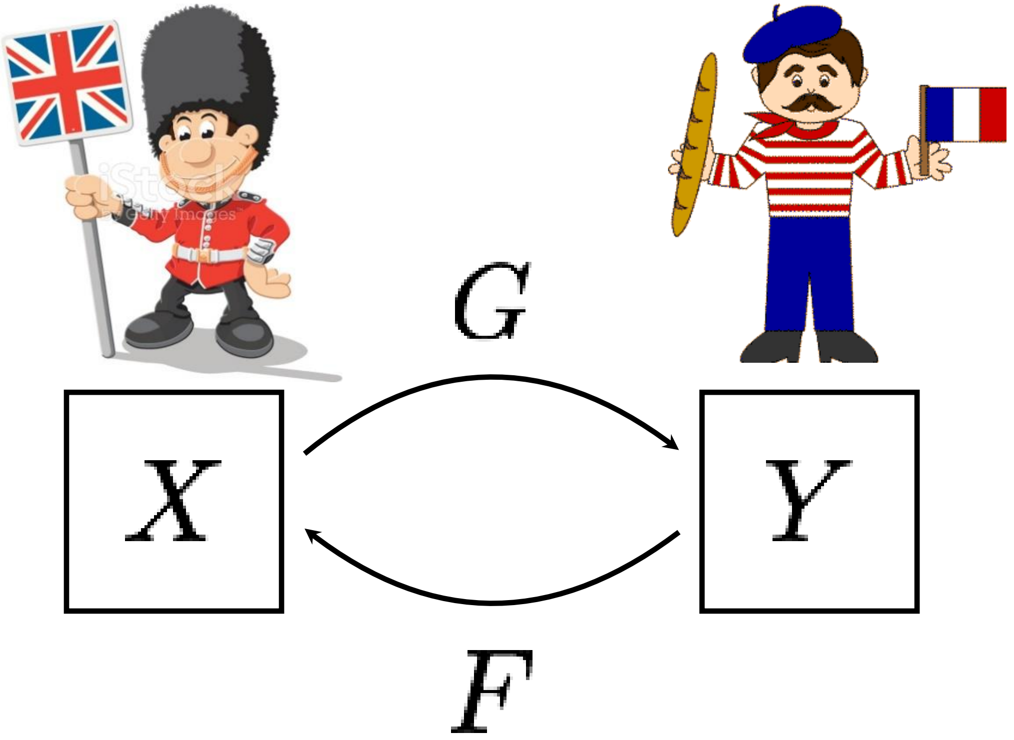


GANs **DO NOT** force output to correspond to input



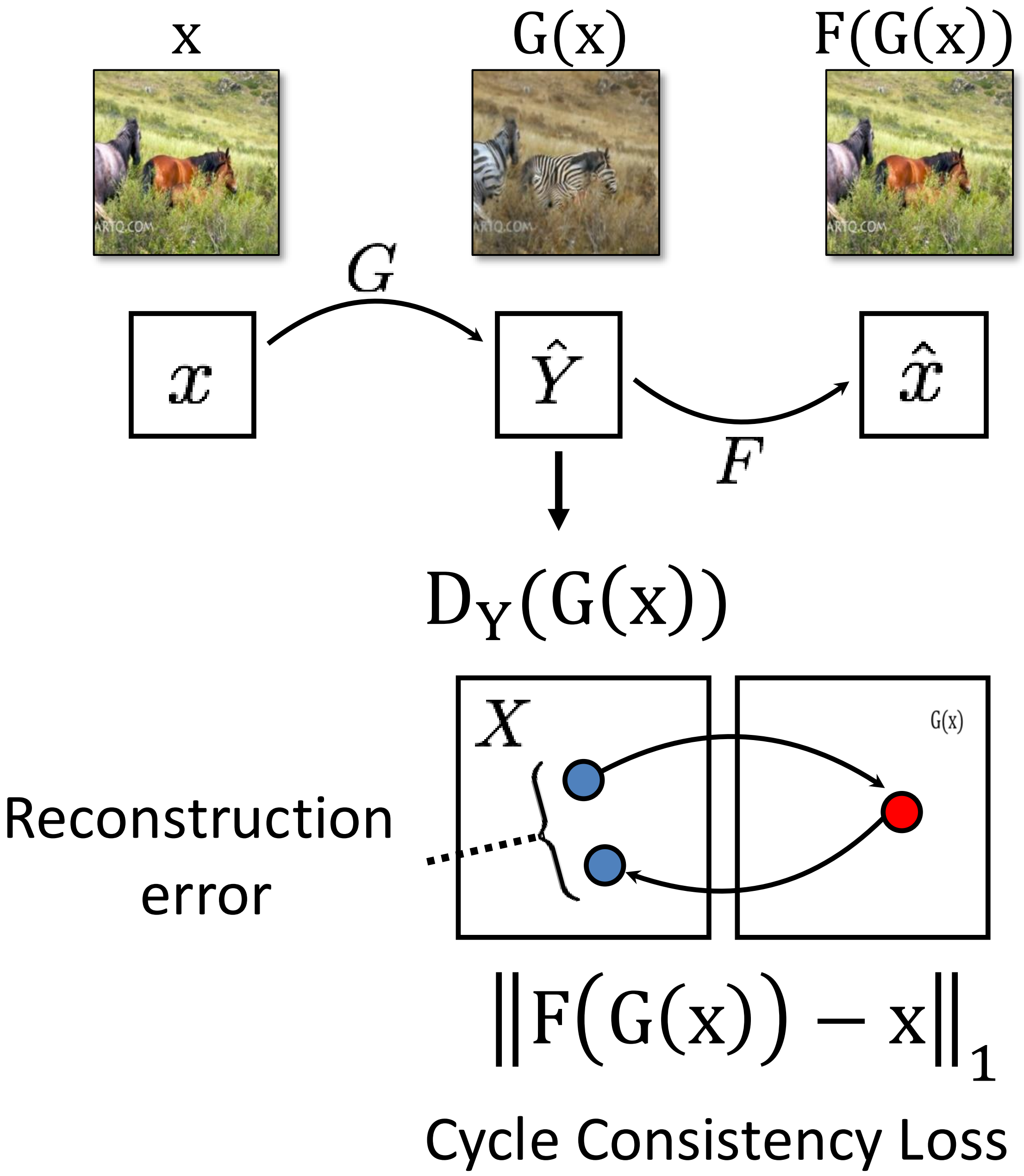
mode collapse!

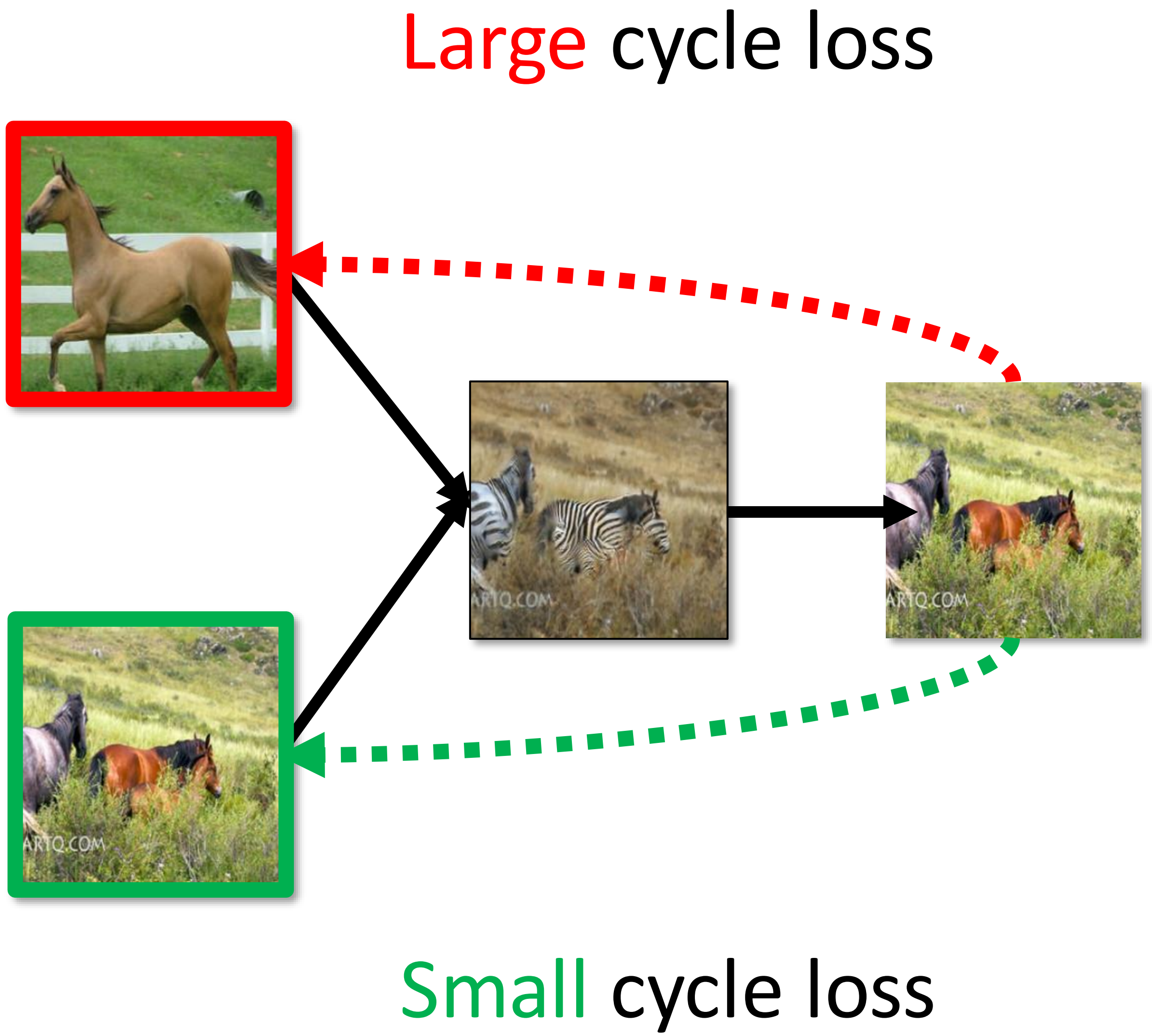
Cycle-Consistent Adversarial Networks



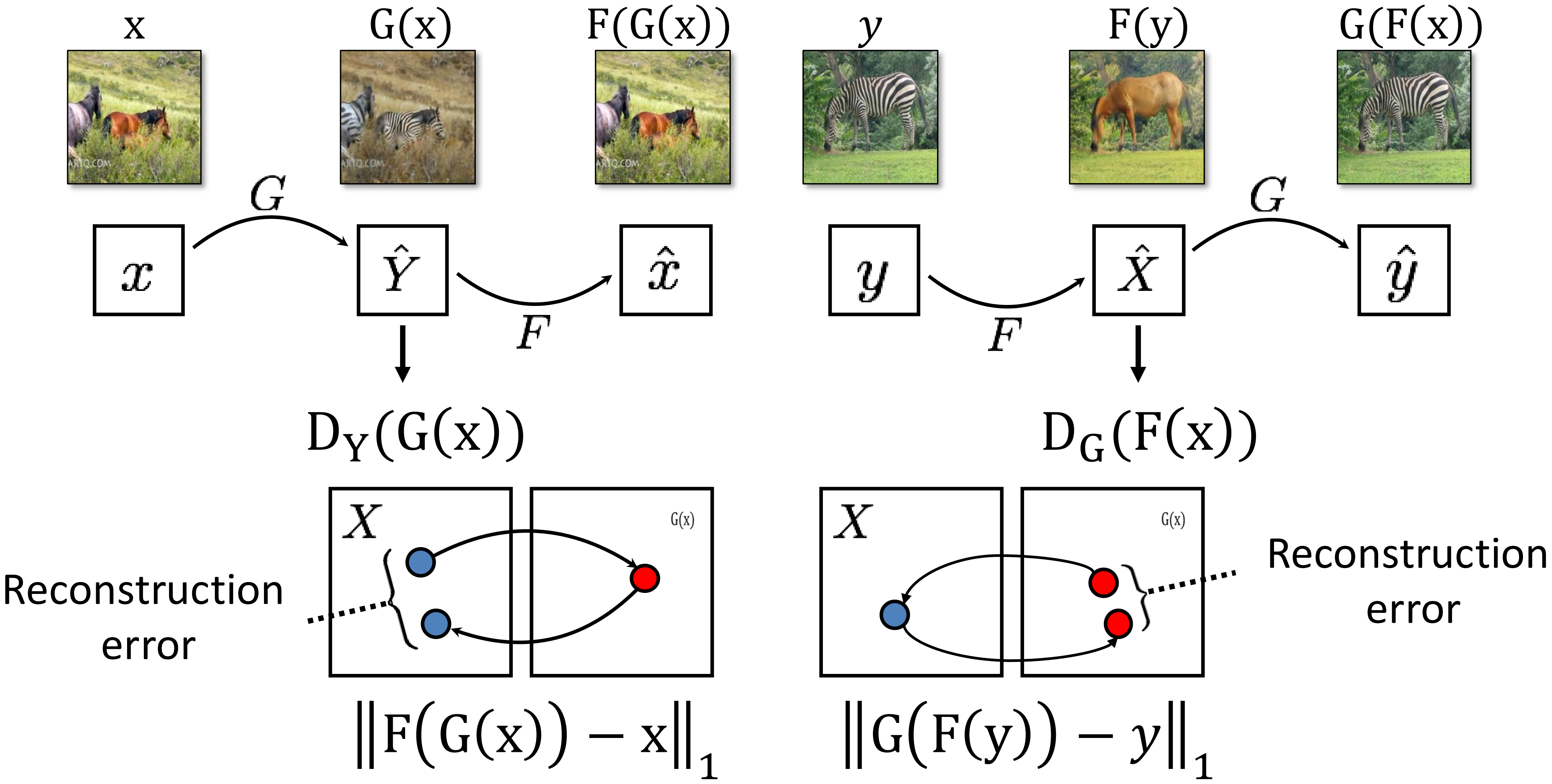
[Mark Twain, 1903]

Cycle-Consistent Adversarial Networks

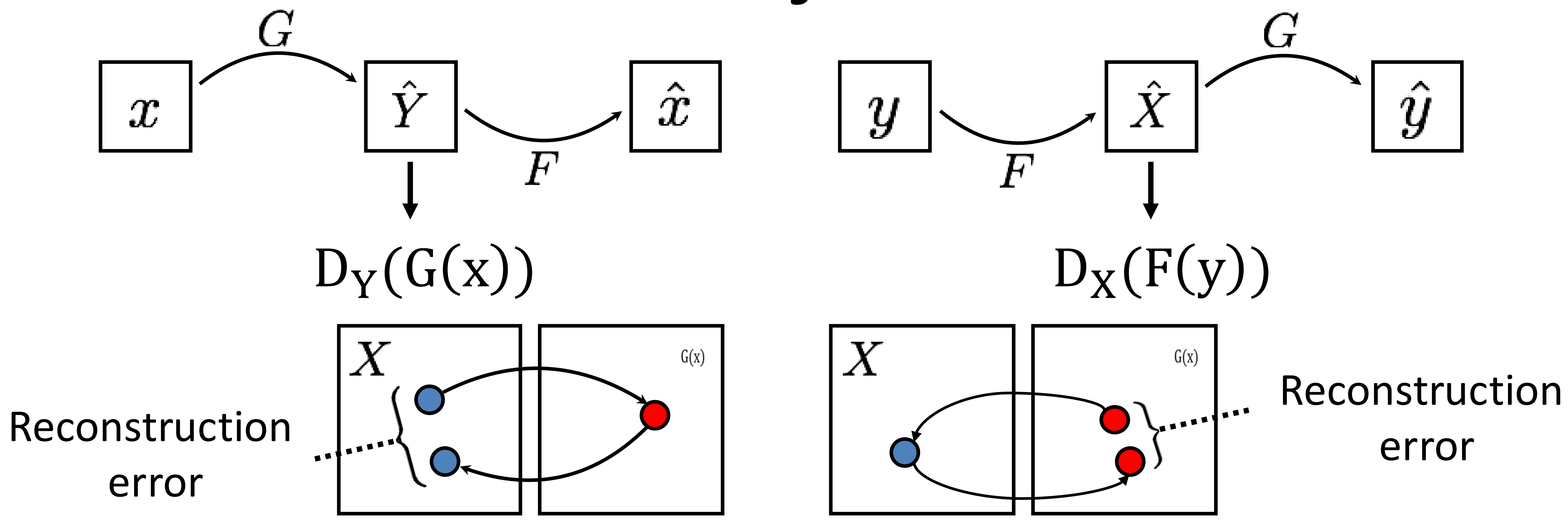




Cycle-Consistent Adversarial Networks



Full Objective



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

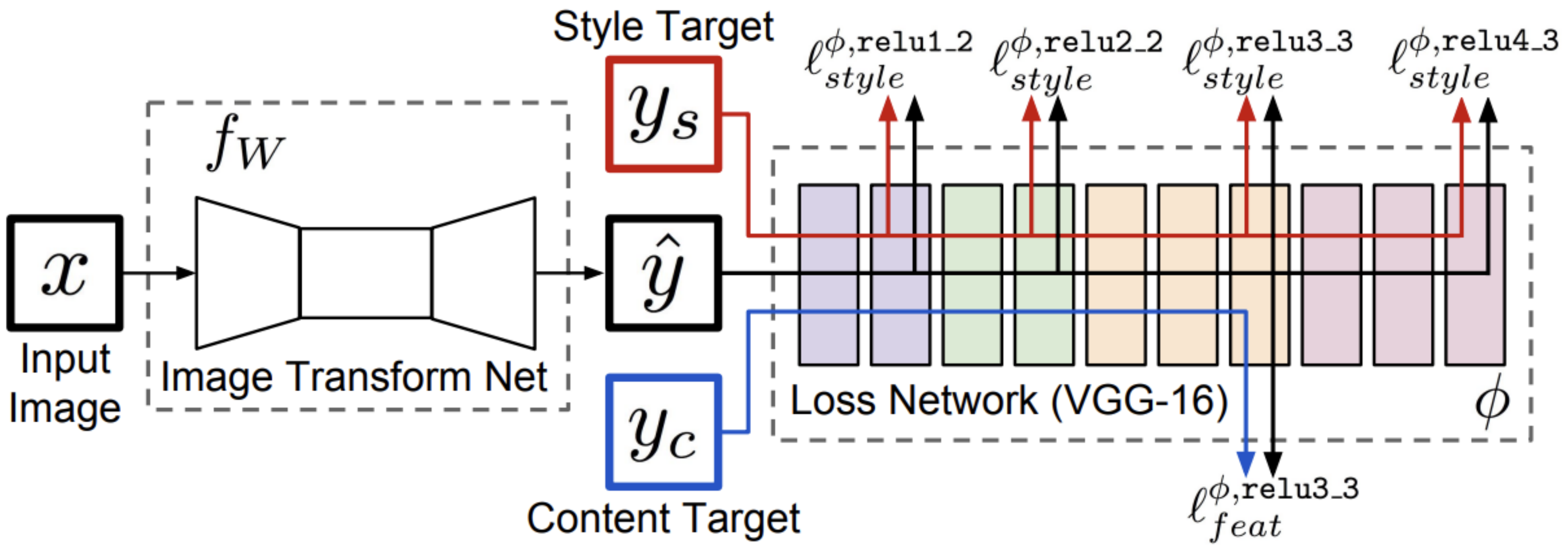
$$\mathcal{L}_D = \mathcal{L}_{D_X} + \mathcal{L}_{D_Y}$$

$$\mathcal{L}_{\text{cycle}} = \mathbb{E}_{x \sim p_{\text{data}}(x)}\left[\|F_{Y\to X}(G_{X\to Y}(x)) - x\|_1\right] + \mathbb{E}_{y \sim p_{\text{data}}(y)}\left[\|G_{X\to Y}(F_{Y\to X}(y)) - y\|_1\right]$$

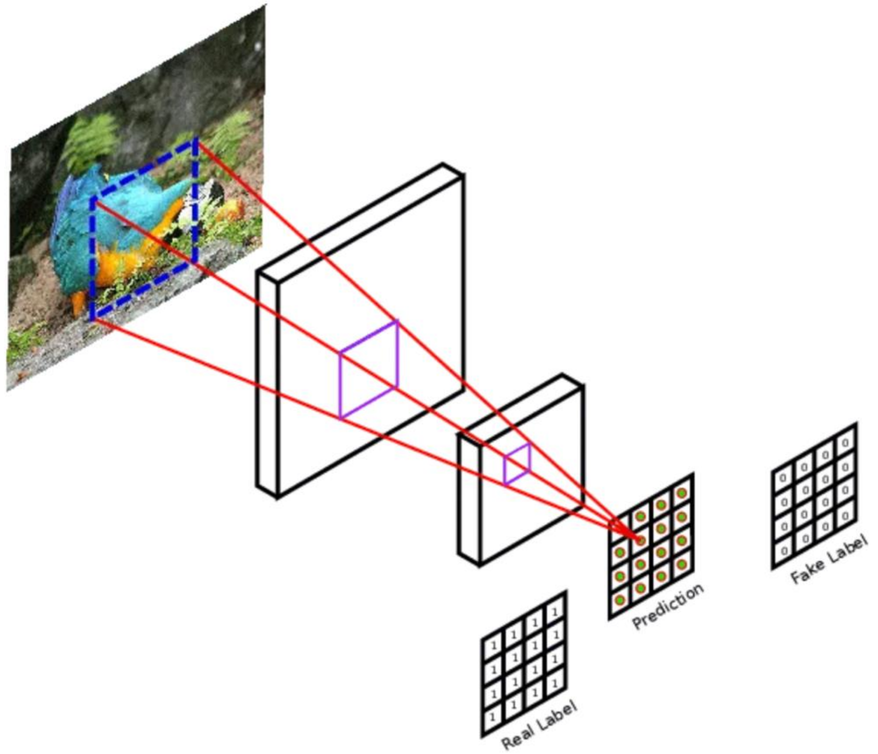
$$\mathcal{L}_G = \mathcal{L}_{G_{X\to Y}} + \mathcal{L}_{G_{Y\to X}} + \lambda \mathcal{L}_{\text{cycle}}$$

Implementation

Generator



Discriminator (PatchGAN)



Loss	Map → Photo % Turkers labeled *real*	Photo → Map % Turkers labeled *real*
CoGAN [30]	0.6% ± 0.5%	0.9% ± 0.5%
BiGAN/ALI [8, 6]	2.1% ± 1.0%	1.9% ± 0.9%
SimGAN [45]	0.7% ± 0.5%	2.6% ± 1.1%
Feature loss + GAN	1.2% ± 0.6%	0.3% ± 0.2%
CycleGAN (ours)	**26.8% ± 2.8%**	**23.2% ± 3.4%**

AMT ‘real vs fake’ test on maps ↔ aerial

Loss	Per-pixel acc.	Per-class acc.	Class IOU
CoGAN [30]	0.40	0.10	0.06
BiGAN/ALI [8, 6]	0.19	0.06	0.02
SimGAN [45]	0.20	0.10	0.04
Feature loss + GAN	0.06	0.04	0.01
CycleGAN (ours)	**0.52**	**0.17**	**0.11**

FCN scores on cityscapes labels→ photos

Loss	Per-pixel acc.	Per-class acc.	Class IOU
CoGAN [30]	0.45	0.11	0.08
BiGAN/ALI [8, 6]	0.41	0.13	0.07
SimGAN [45]	0.47	0.11	0.07
Feature loss + GAN	0.50	0.10	0.06
CycleGAN (ours)	**0.58**	**0.22**	**0.16**

Classification performance of cityscapes photo→labels

Applications

Painting -> Photos

Object Transfiguration

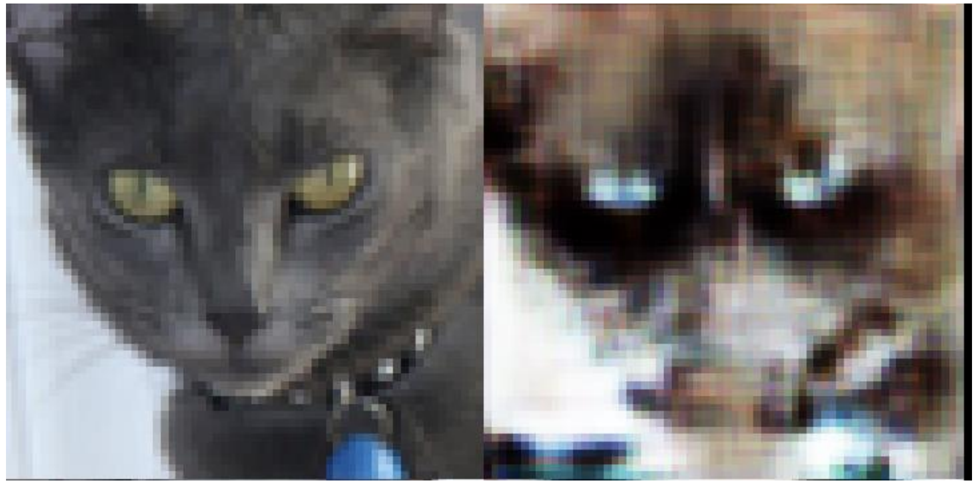
GTA -> Cityscapes

Season Transfer

Failure Cases

Out-of-Distribution

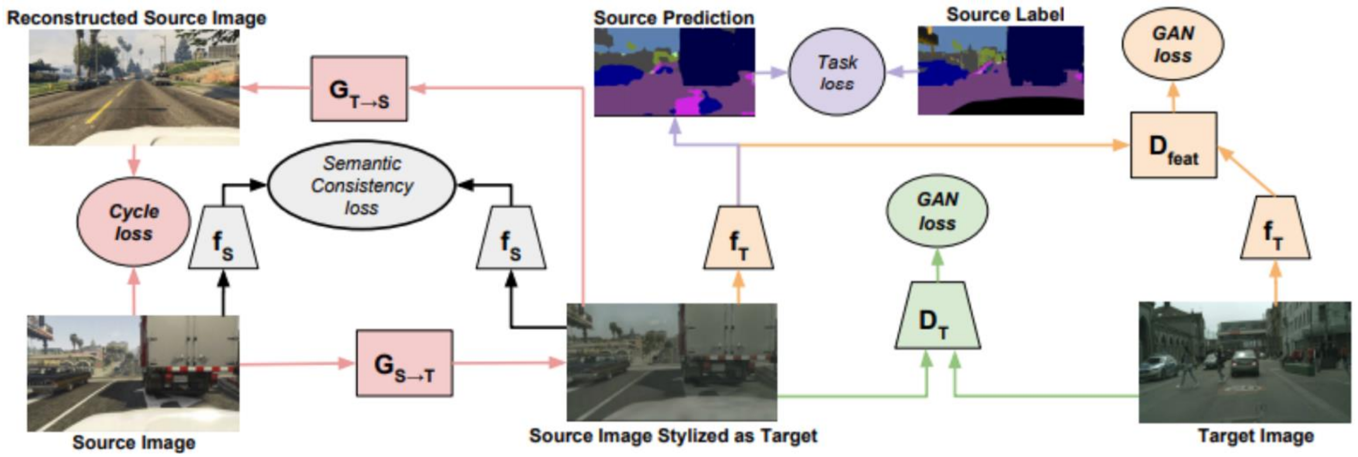
Fail to change geometry
(dog -> cat)



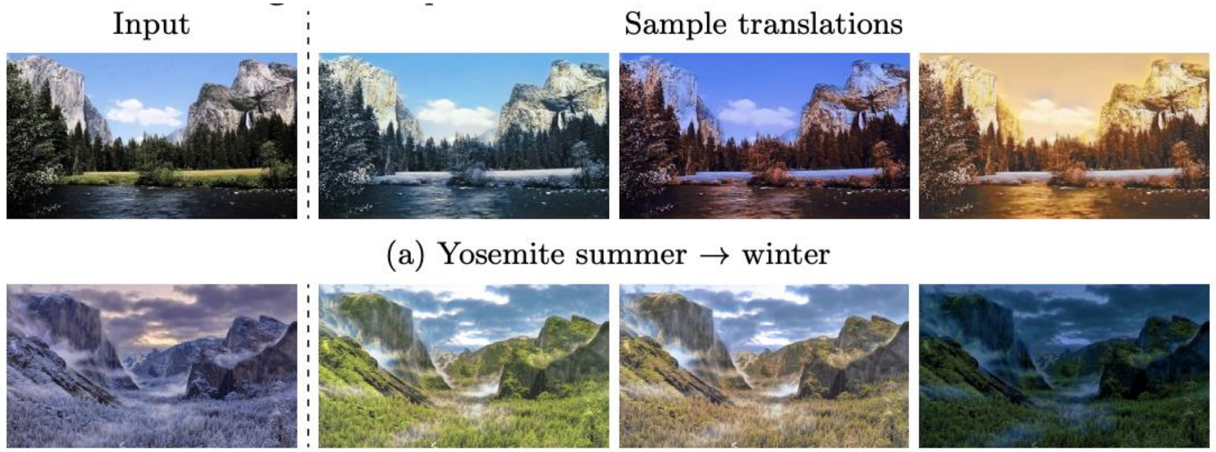
Fail to preserve perceptual details

Semantic ambiguity

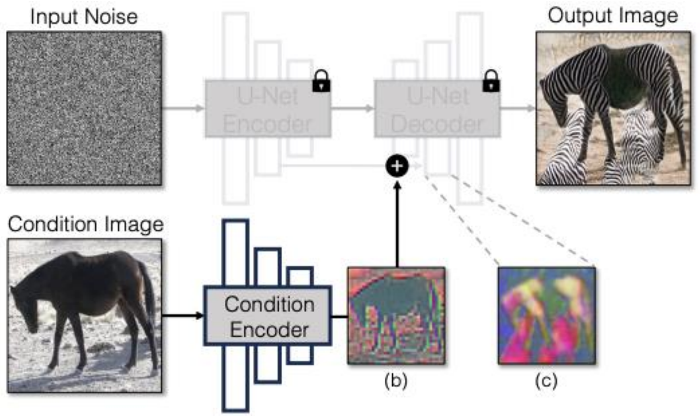
Beyond CycleGAN



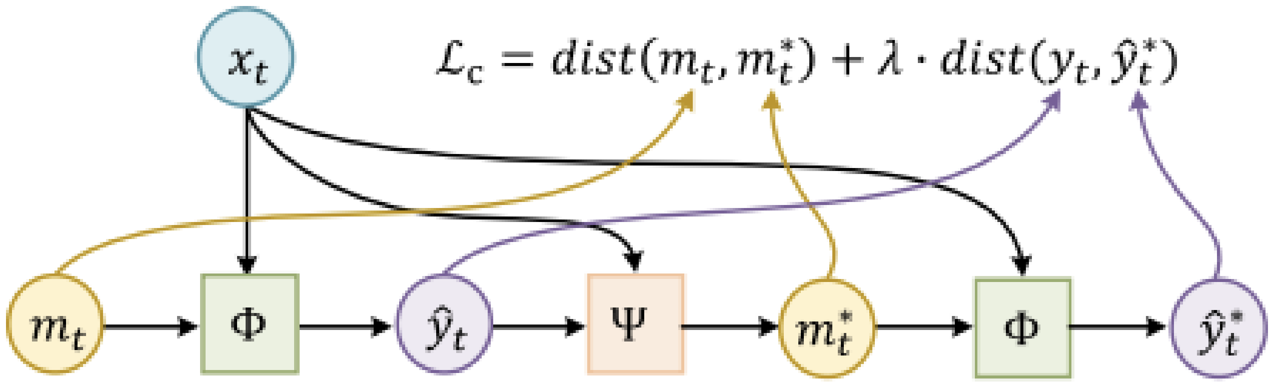
Domain Adaptation
(Hoffman et al., 2018)



Multimodal Translation
(Huang et al., 2018)



Translation using Stable Diffusion
(Gaurav et al., 2024)



Video Inpainting
(Wu et al., 2023)

Q & A

- How does this method compare with the method used when I ask ChatGPT to change an image to look like a Studio Ghibli image?

- This paper claims that paired training data is unavailable, but when training discriminator it still uses the GAN setting, which still requires paired training data as I understand. Could you clarify this confusing point for me?

Thank You!