$\rightarrow$

Conditional GANs, Image-to-Image Translation

Jun-Yan Zhu

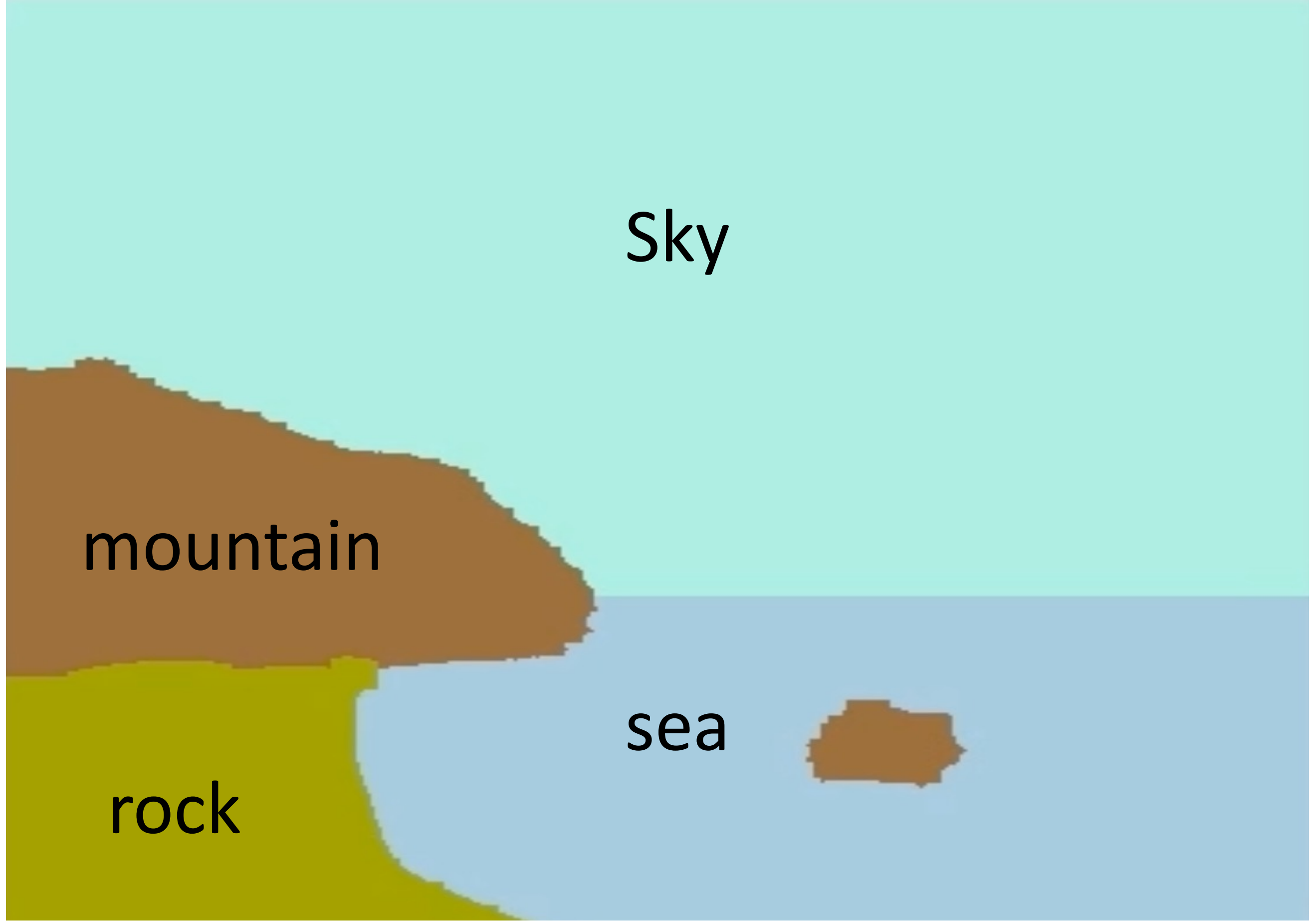
16-726, Spring 2022

Many slides from Phillip Isola, Ming-Yu Liu, Xun Huang, etc.

Logistics

- HW2 gather town party Mon 8-10 pm
- No class next week (due to Spring break)
- HW 1 Class Choice Award:
  – Vote by the end of Wed.
  – Winner will be announced on 03/14 (Mon)

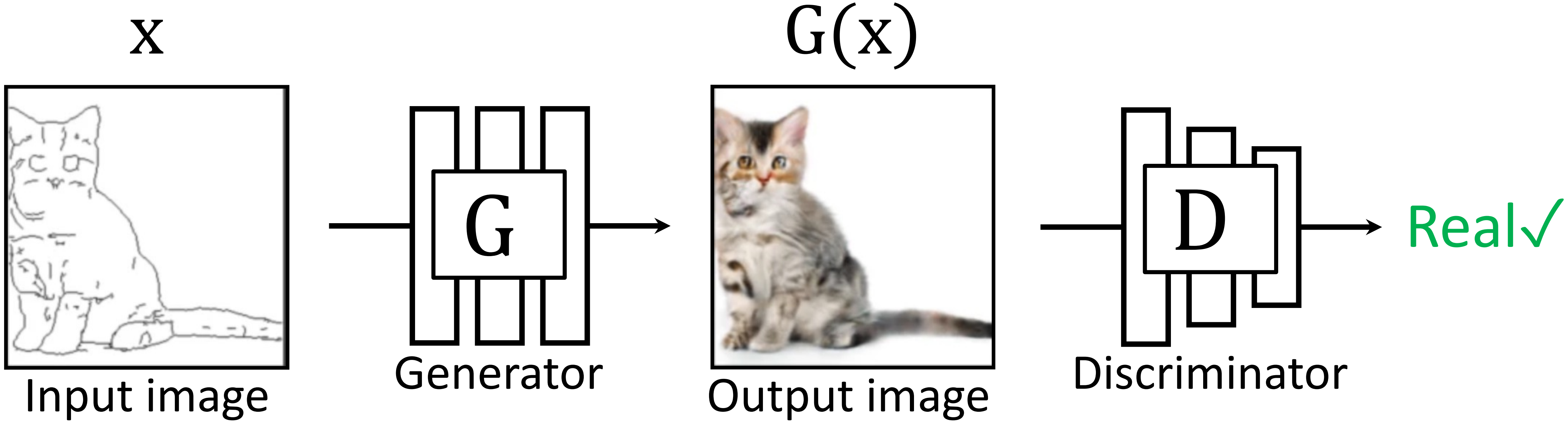
Problem Statement



Input

Output

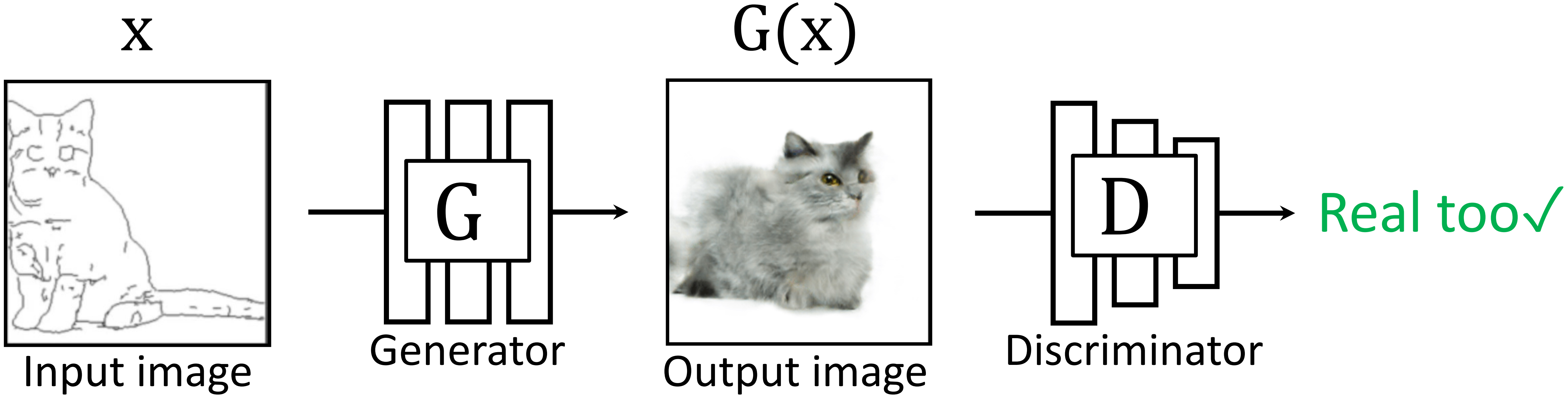
<u>Goal</u>: synthesize a realistic photograph given an input image



Learning objective

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

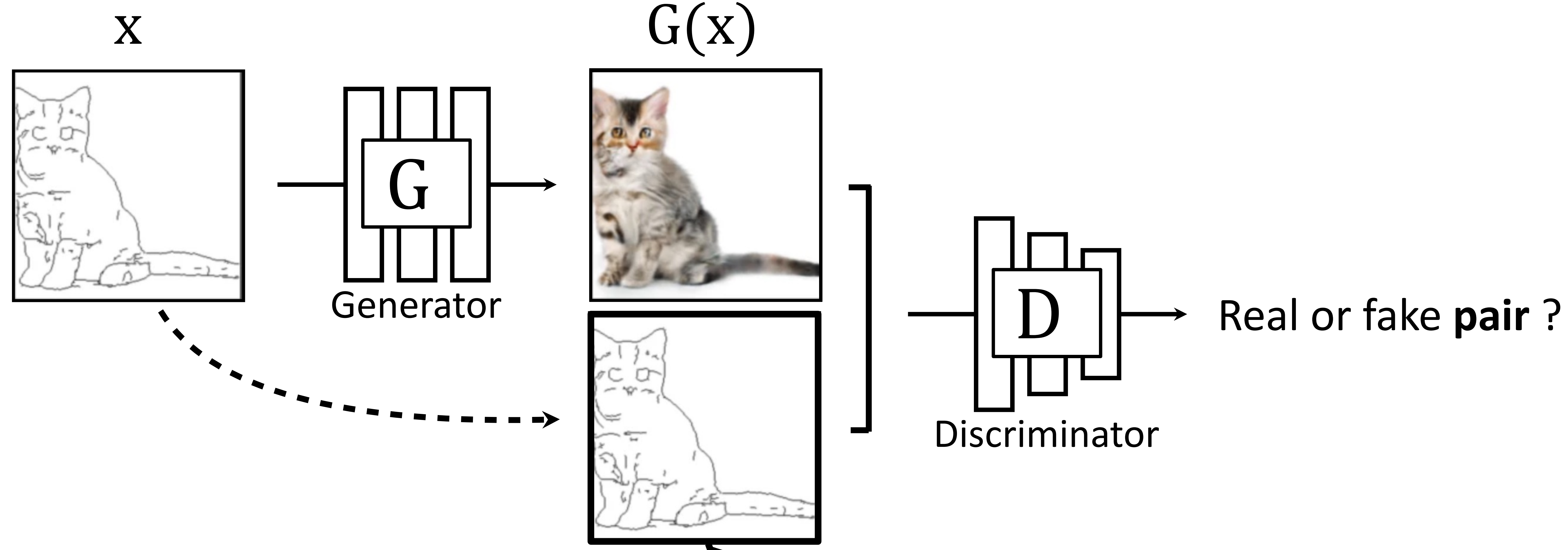
Pix2pix [Isola et al., 2016]



Learning objective

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

Pix2pix [Isola et al., 2016]

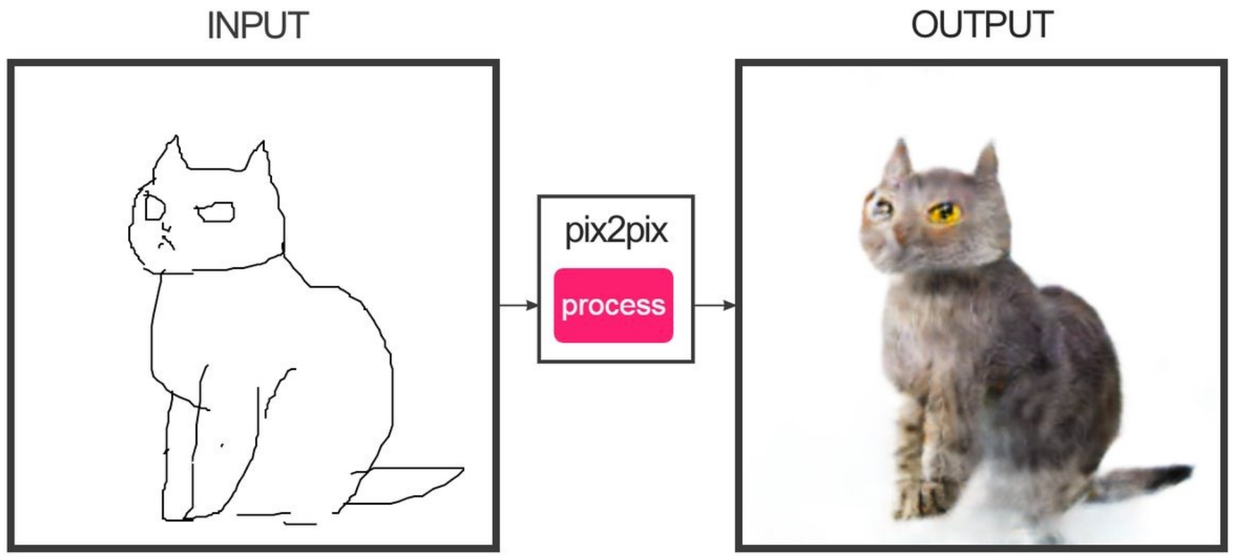


Learning objective

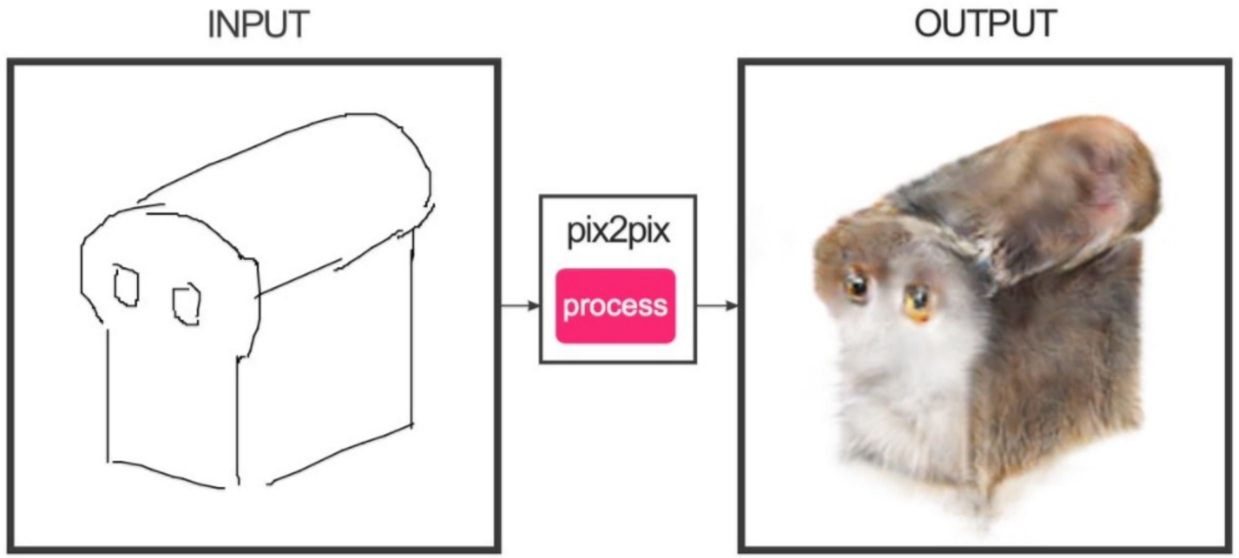
$$\min_G \max_D \mathbb{E}_x[\log(1-D(\boxed{x}, G(x))]+\mathbb{E}_{x,y}[\log D(\boxed{x}, y)]$$

Pix2pix [Isola et al., 2016]

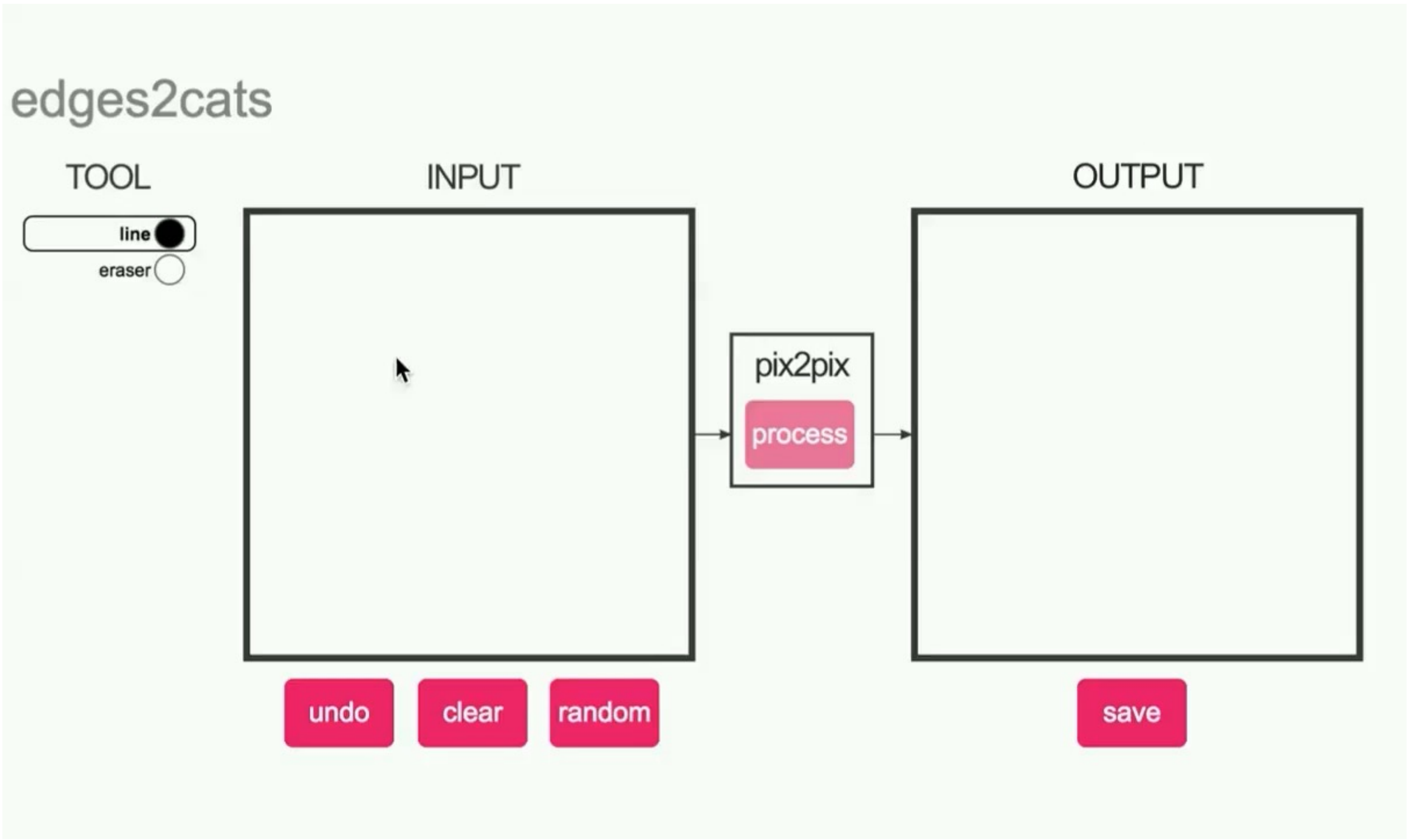
#edges2cats [Christopher Hesse]



@gods_tail



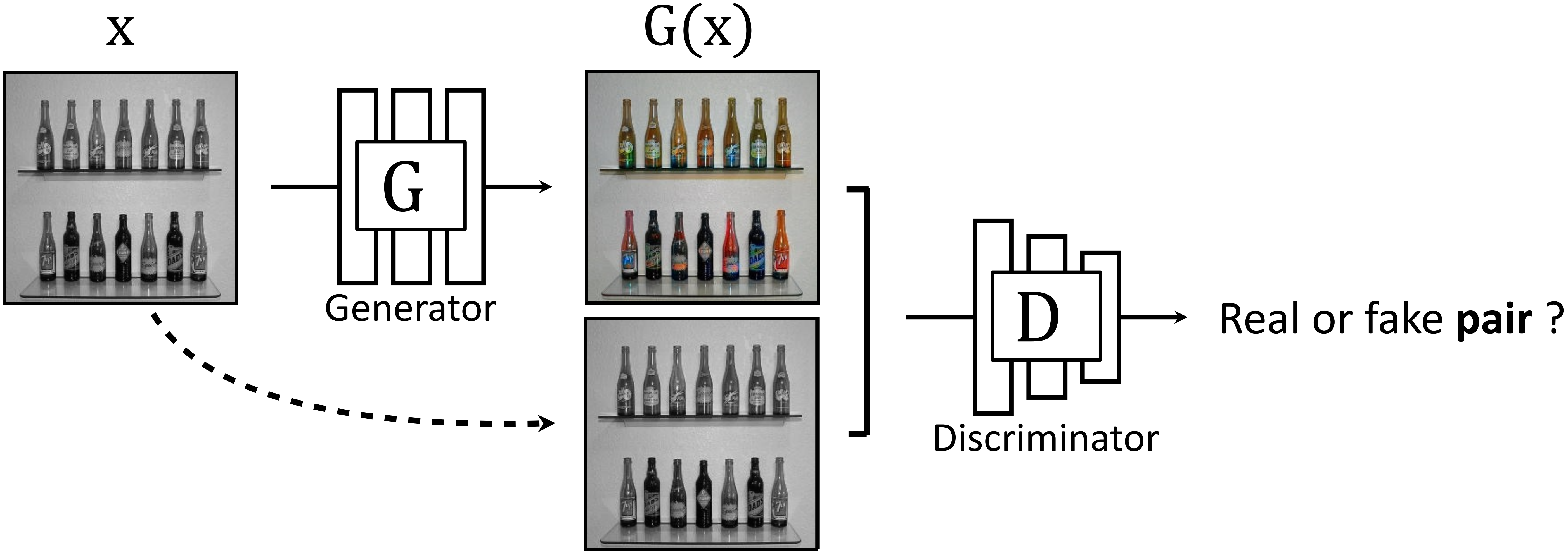
Ivy Tasi @ivymyt



@matthematician

Vitaly Vidmirov @vvid

https://affinelayer.com/pixsrv/



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

Input: **Sketch** → Output: **Photo**

Pix2pix [Isola et al., 2016]

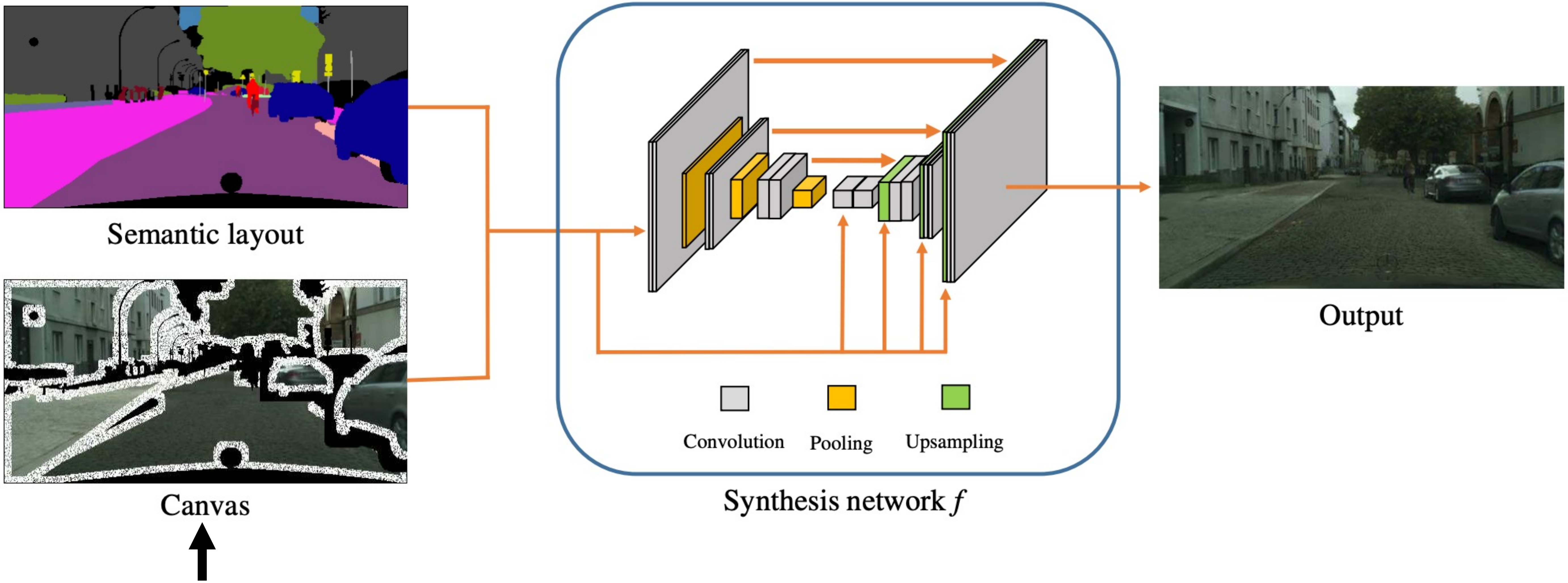
Automatic Colorization with pix2pix

Input Output Input Output Input Output

Data from [Russakovsky et al. 2015]

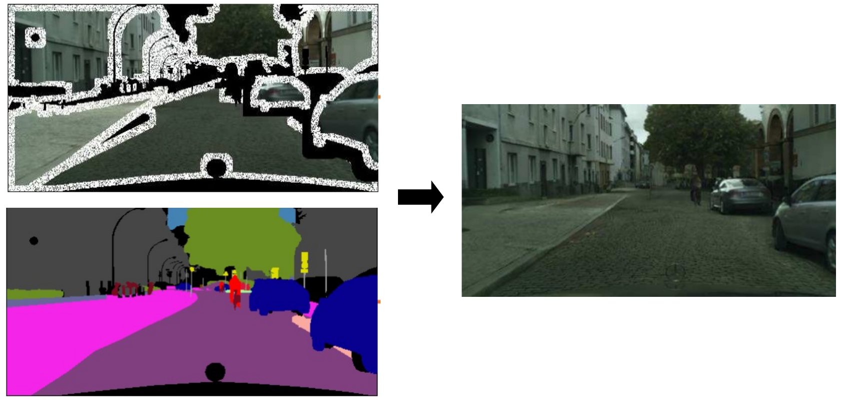
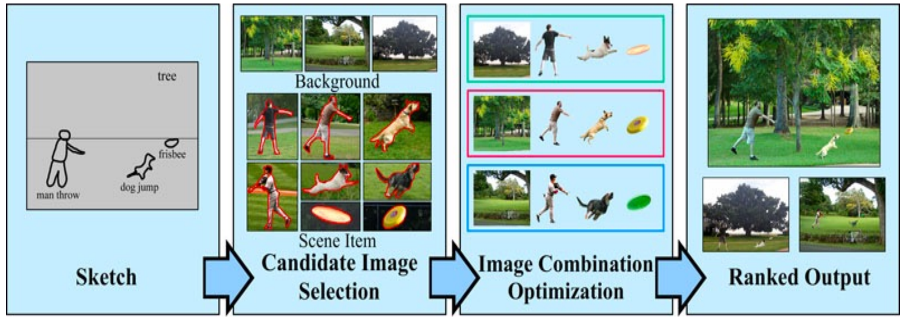
Learning vs. Exemplar-based

Hybrid Method



Output from exemplar-based method

Semi-Parametric Image Synthesis [Qi et al., 2018]

Learning-based	Hybrid method	Exemplar-based
		
[Isola et al], [Wang et al] [Park et al], SEAN [Zhu et al]	SIMS [Qi et al]	[Johnson et al], [Lalonde et al] [Tao et al], [Bansal et al]

	Learning-based	Hybrid method	Exemplar-based
Speed	🙂	🙁	🙁
Local realism	😐	🙂	🙂
Global realism	🙂	😐	🙁
Match Input	🙂	😐	😐

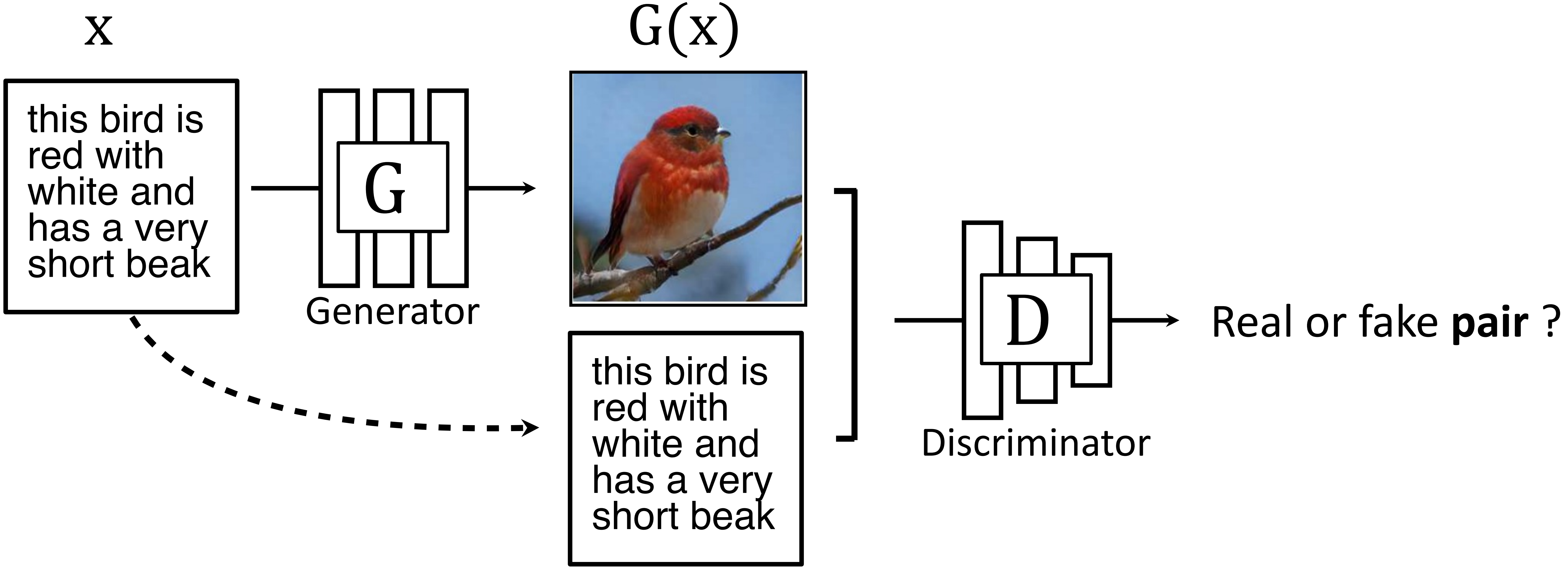
Discussion

Summary

- Intuitive user inputs.
- Realistic outputs.
- Used by visual artists.

Challenges

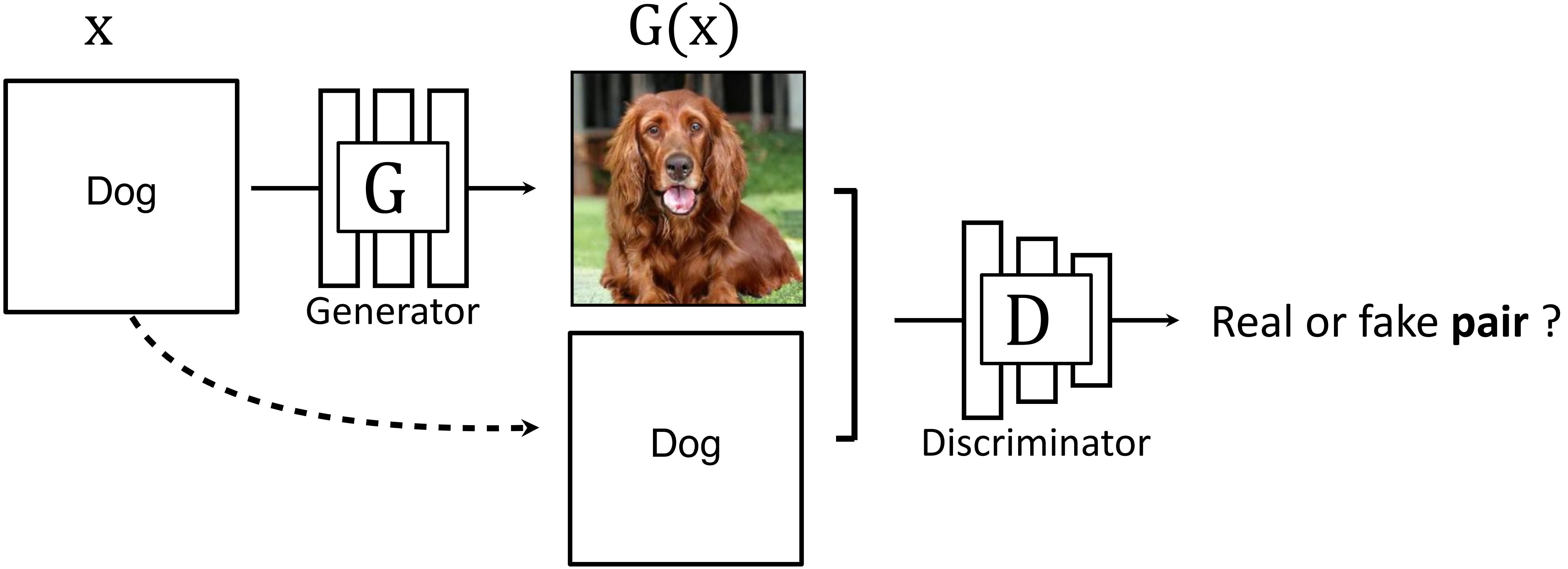
- Fine-grained controls (texture, 3D, and lighting).
- High-resolution output (4K).
- Model efficiency on mobile devices.
- Video Control.



Input: **Text** → Output: **Photo**

Text-to-Image Synthesis

StackGAN, StackGAN++ [Zhang et al. 2016 and 2017], AttnGAN [Xu et al., 2018]

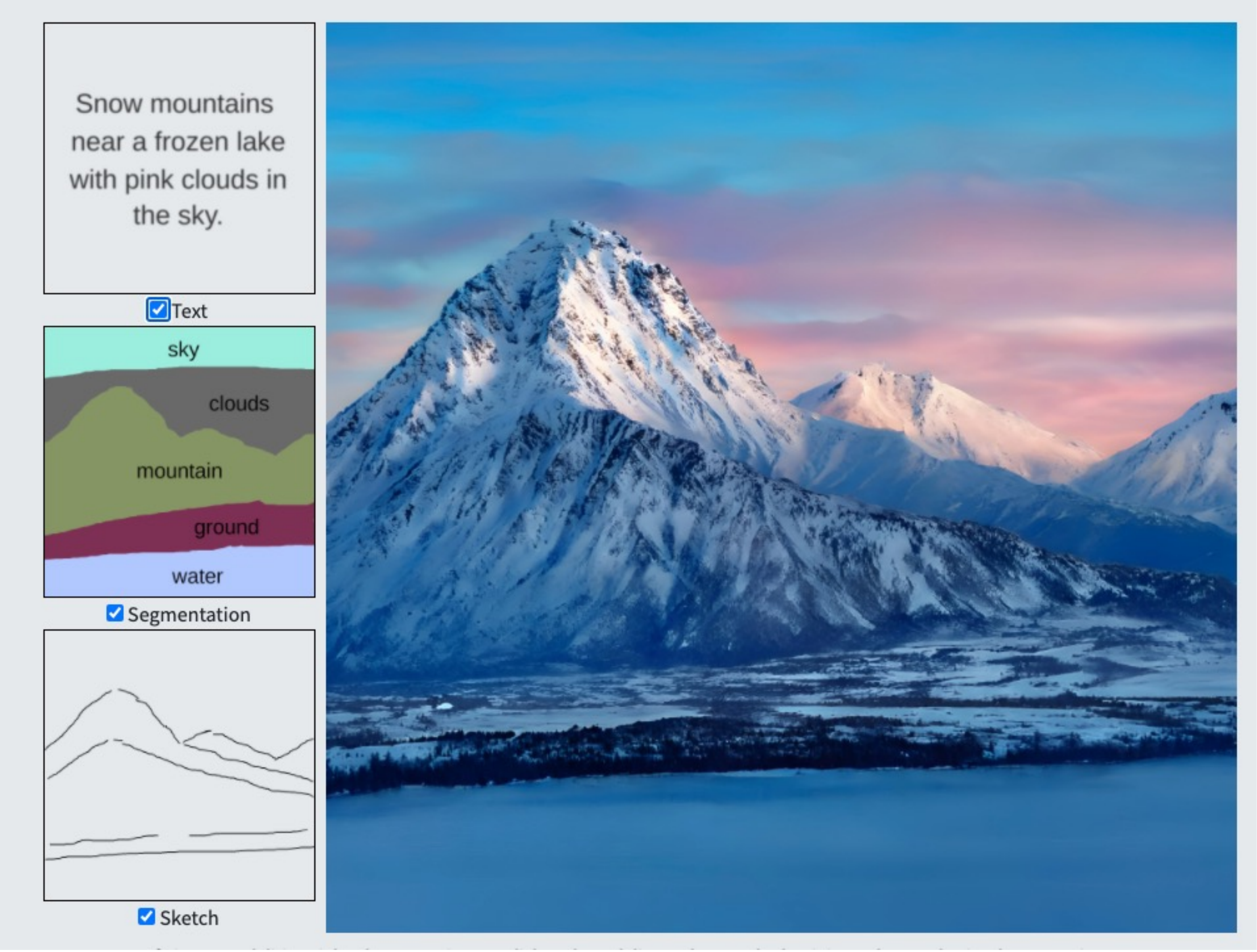


Input: **Class** → Output: **Photo**

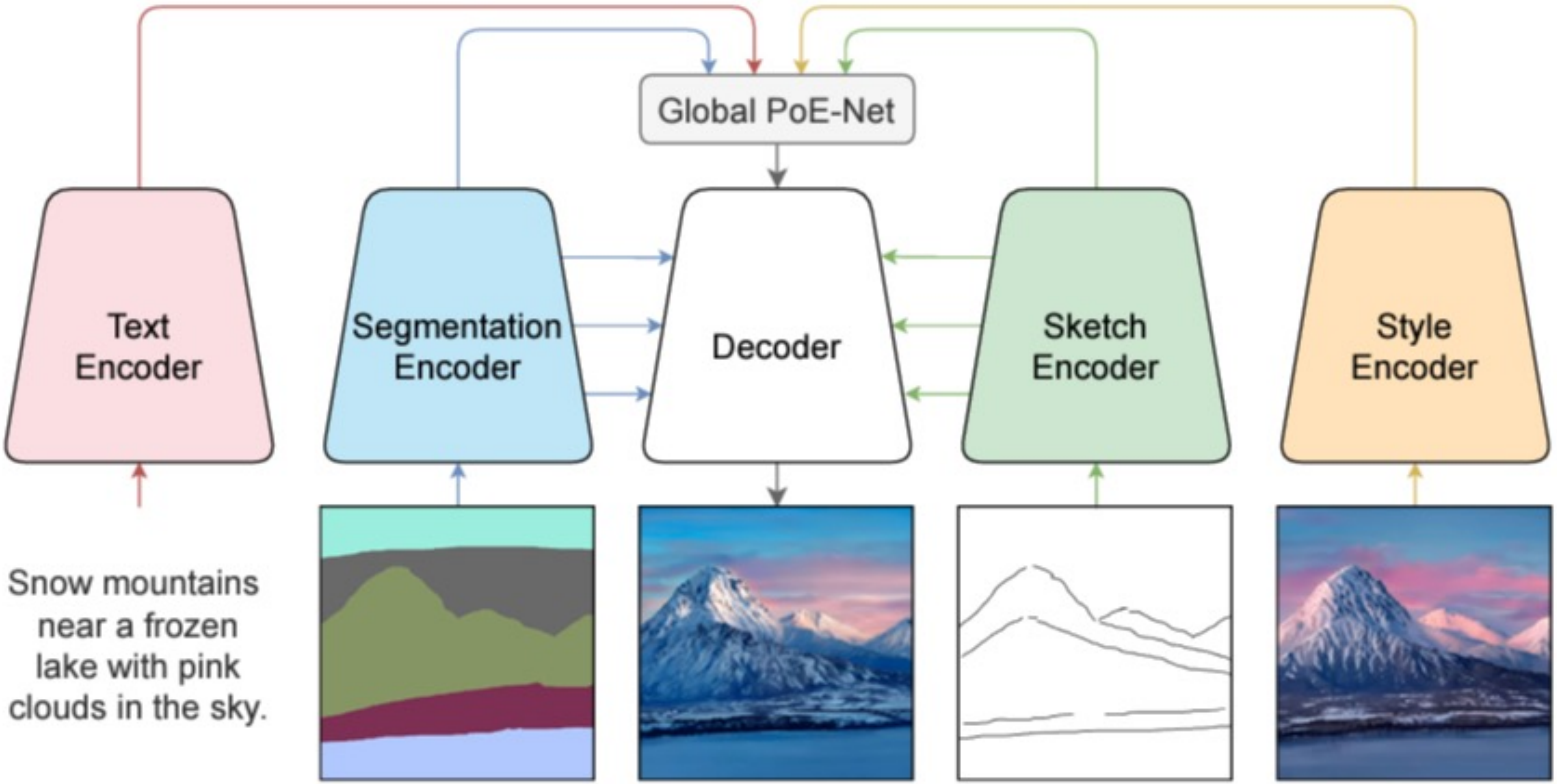
Class-conditional GANs

cGANs [Mirza and Osindero. 2014], SAGAN [Zhang et al., 2018], BigGAN [Brock et al., 2019]

StyleGAN-XL [Sauer et al., 2022]



Multimodal Conditional Image Synthesis with Product-of-Experts GANs [Huang et al., 2021]

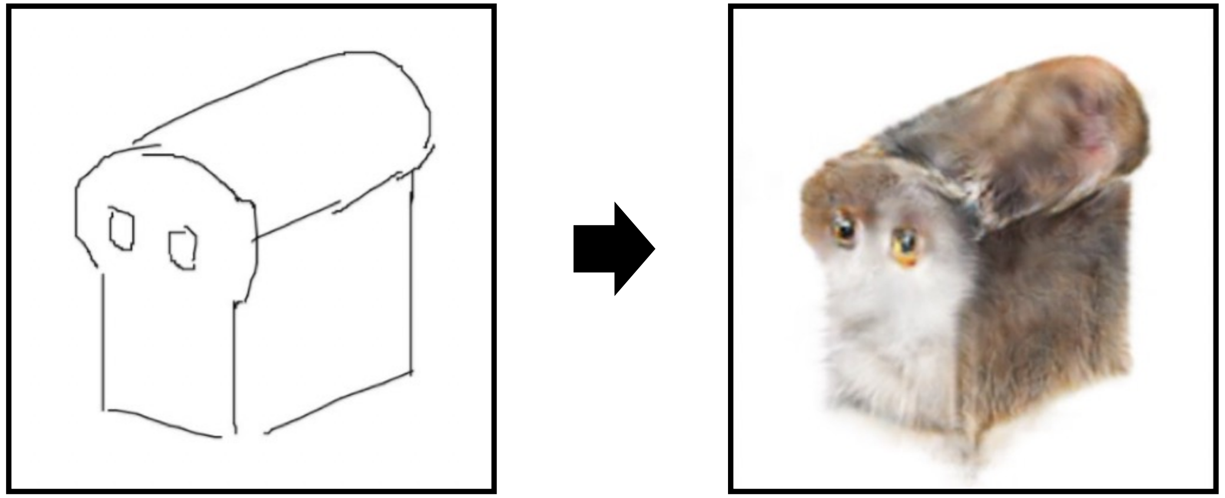


Multimodal Conditional Image Synthesis with Product-of-Experts GANs [Huang et al., 2021]

GauGAN2 Demo

http://gaugan.org/gaugan2/

Supervised Learning Approach



Edges2cats

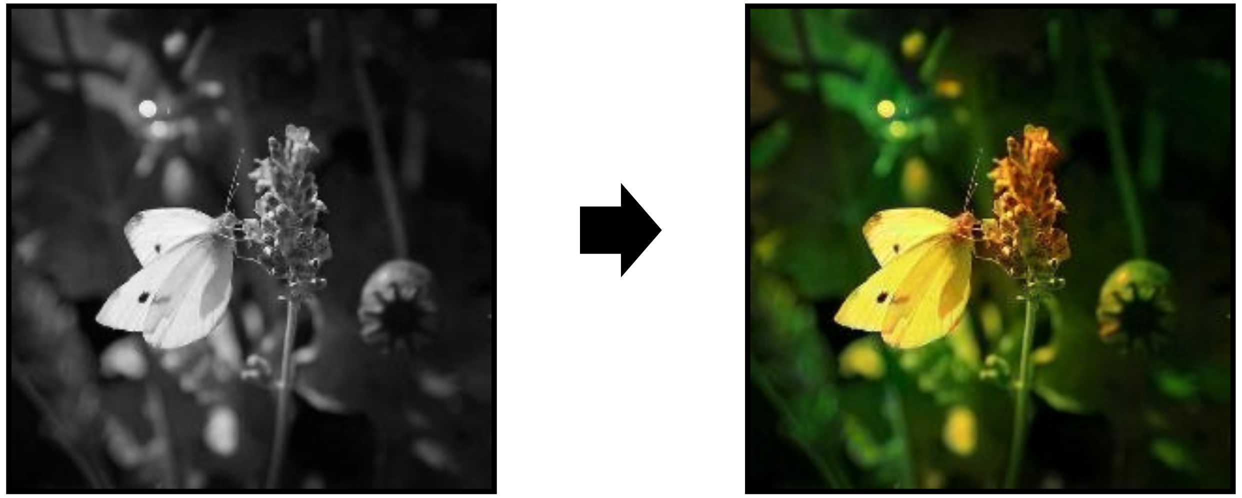
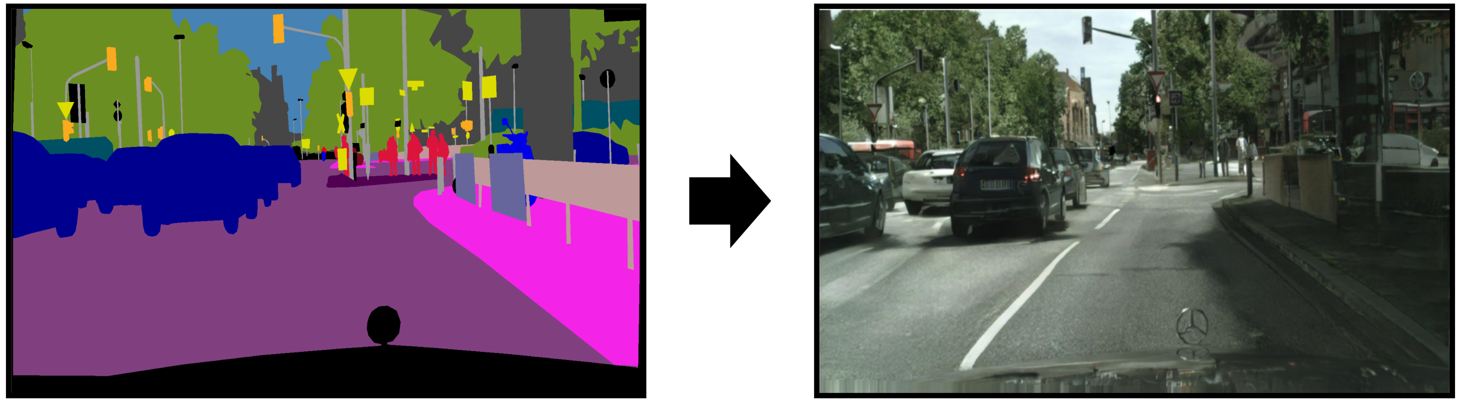
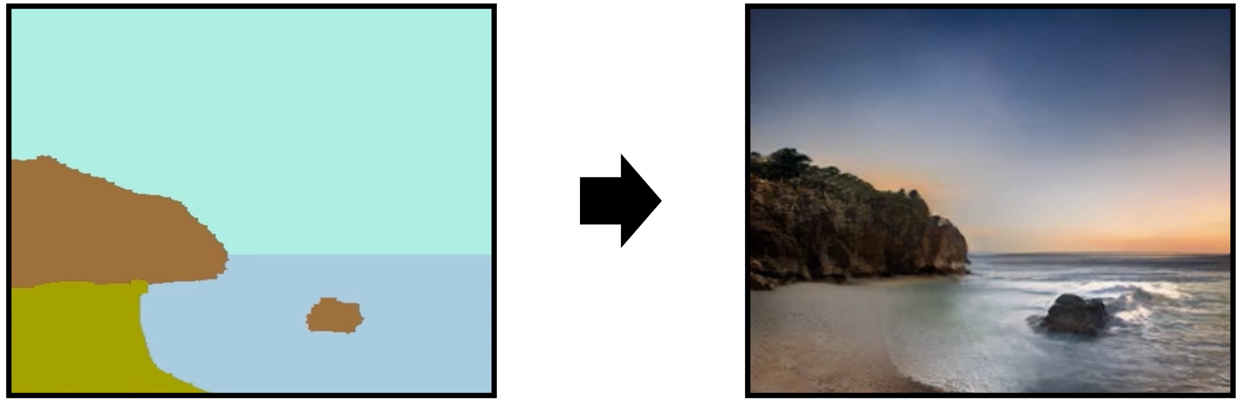


Image colorization

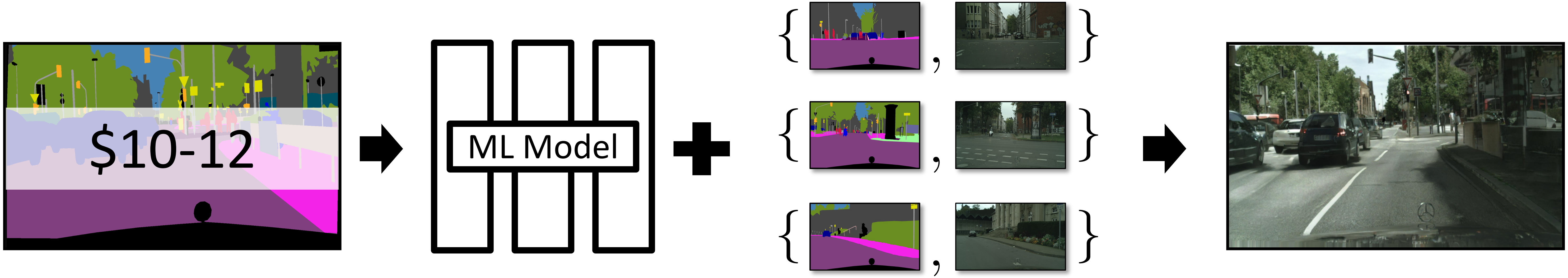


Street view images

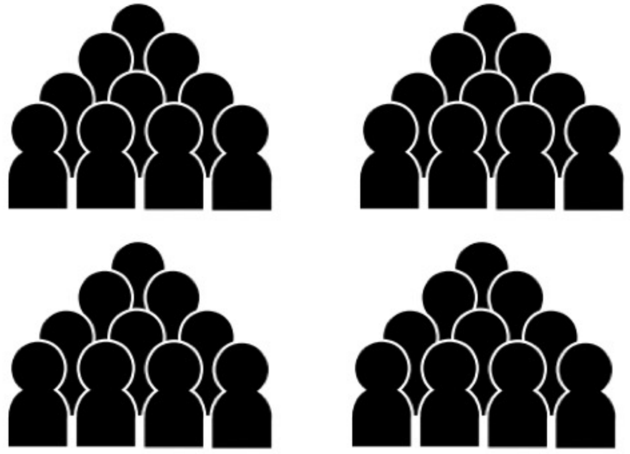


Natural outdoor images

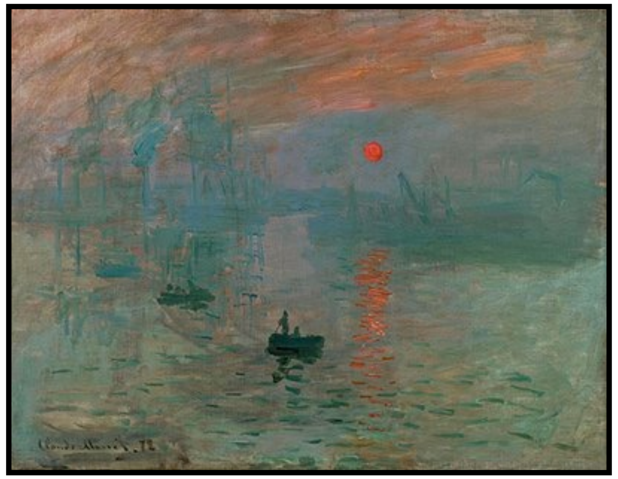
Supervised Learning Approach



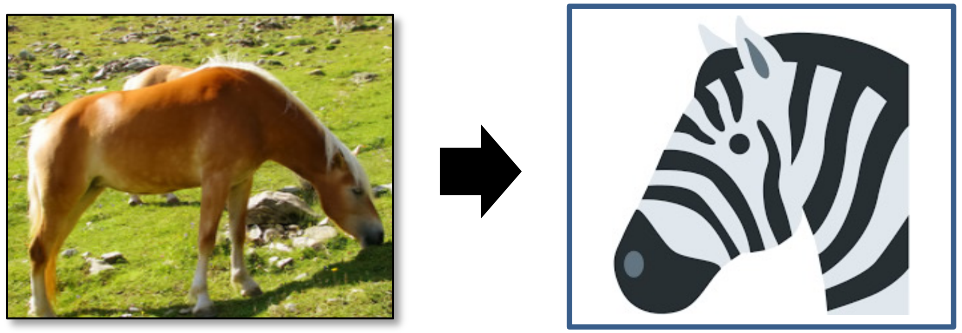
User Input | Learning algorithm | Labeled data | Visual Content



Expensive labor



Artistic authoring



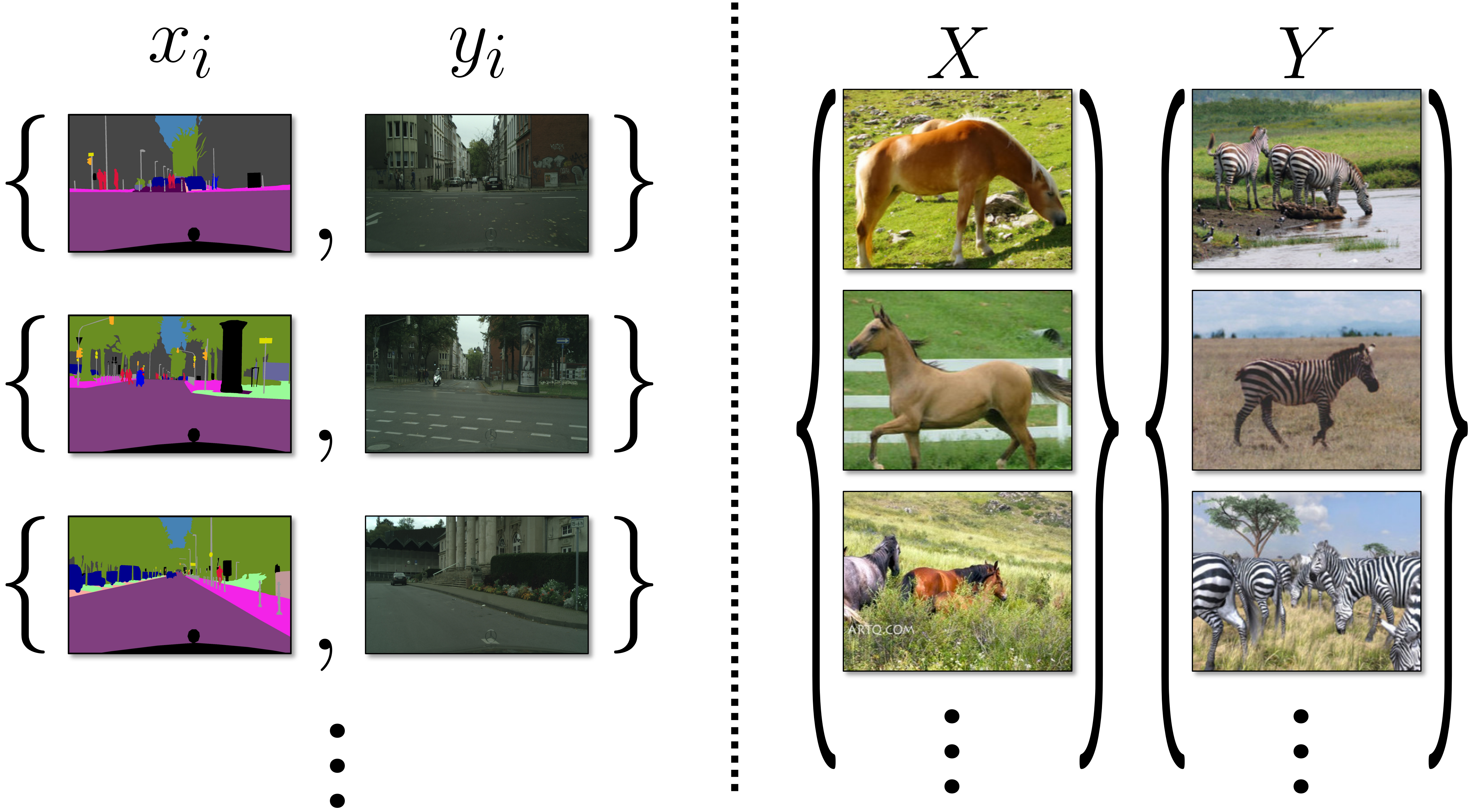
horse

zebra

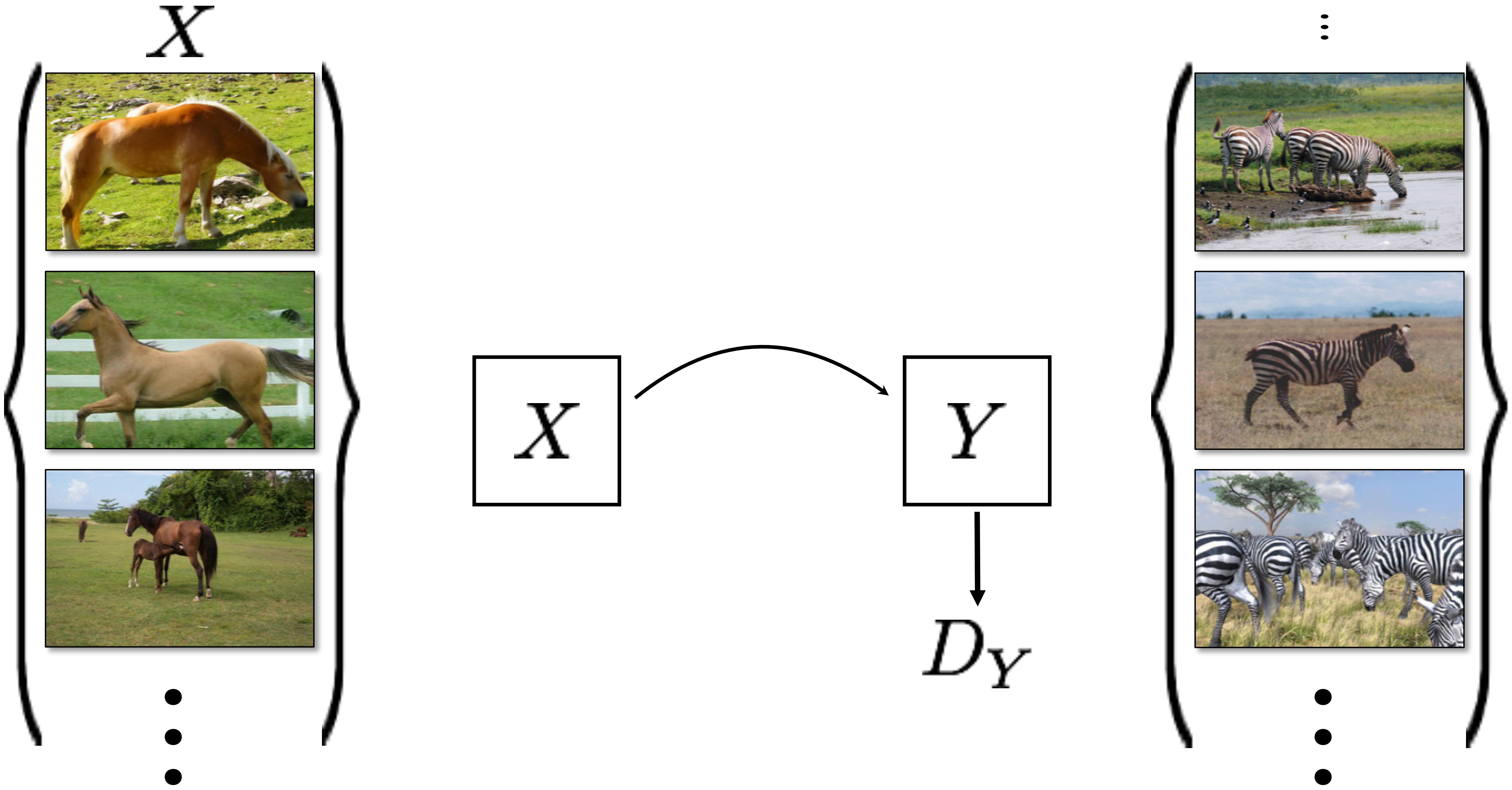
Infeasible

Supervised

Unsupervised



Unsupervised Learning of $p(y \mid x)$

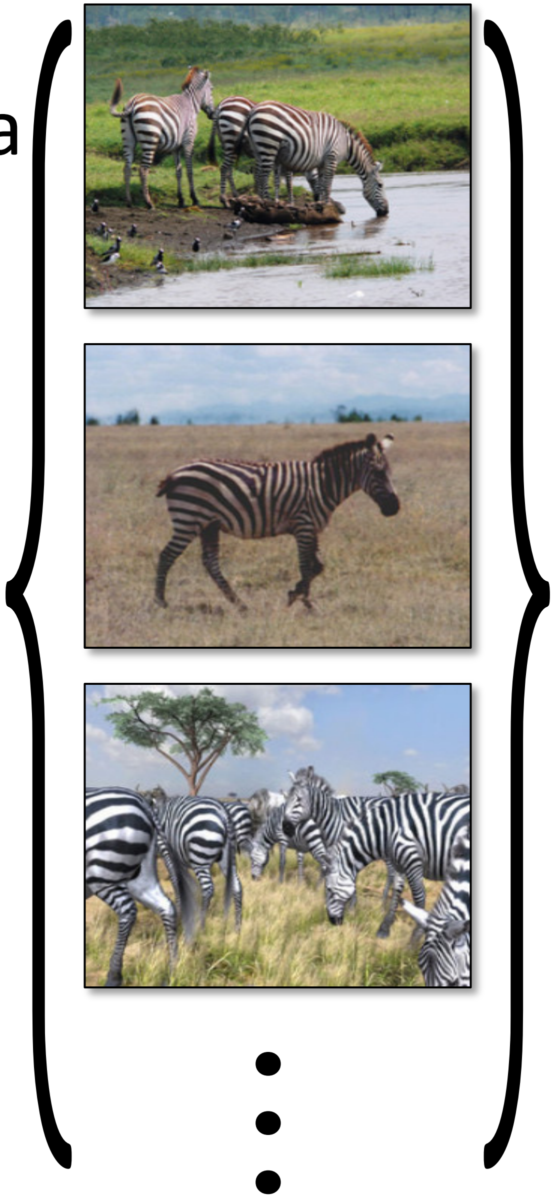


[Zhu*, Park*, Isola, and Efros, 2017]

Unsupervised Learning of $p(y \mid x)$

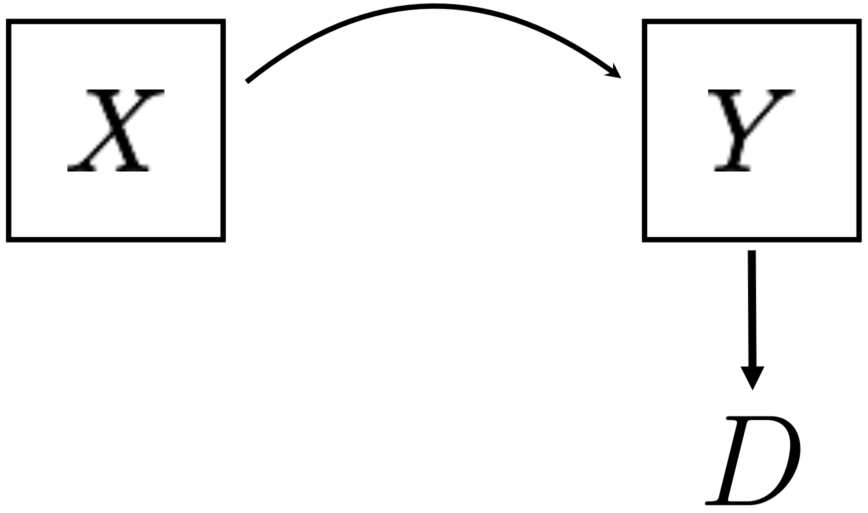
X

Y



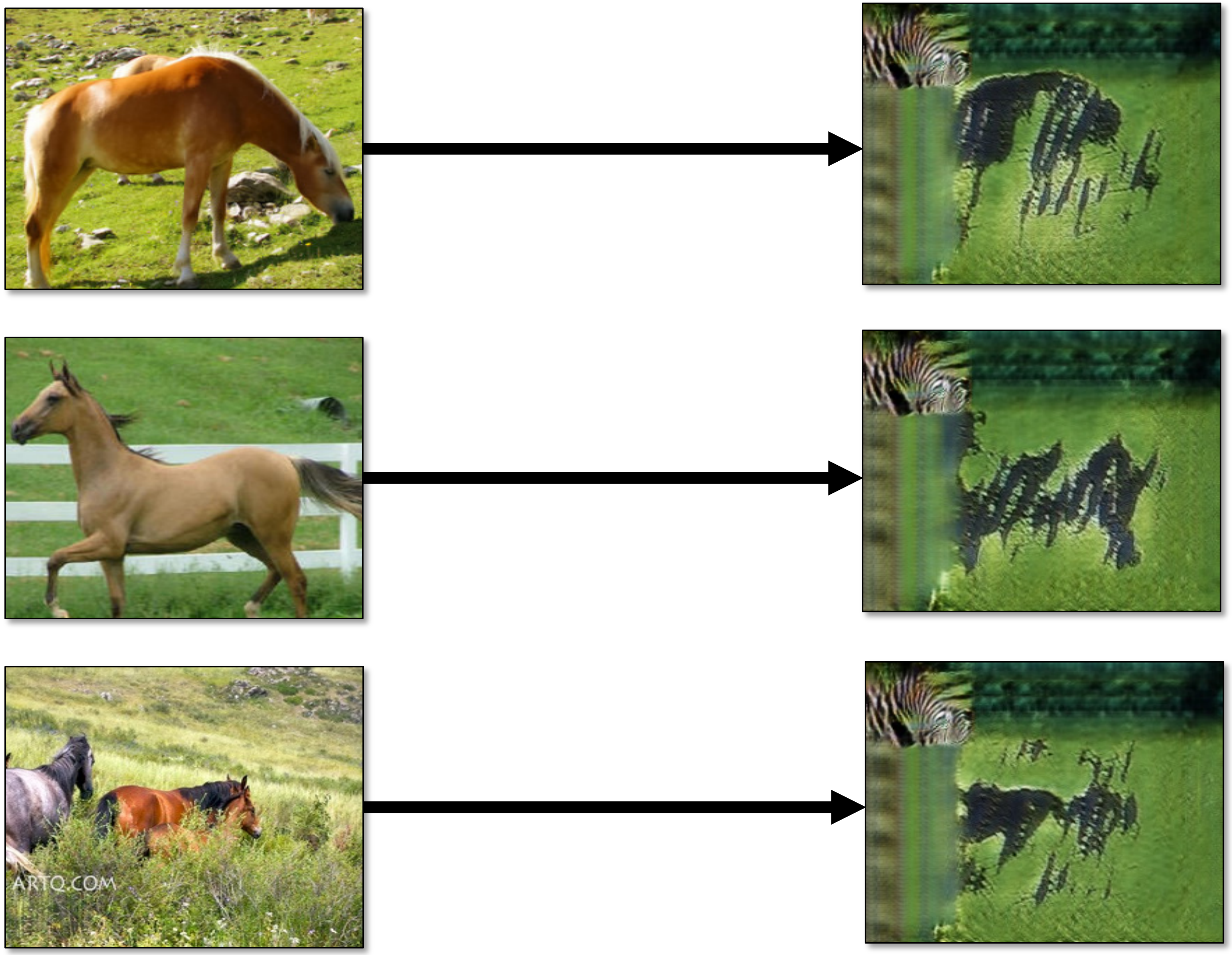
fake zebra　　　　real zebra

$$\mathbb{E}_x \log(1-D(\underline{G(x)}))+\mathbb{E}_y \log D(\underline{y})$$



Discriminator

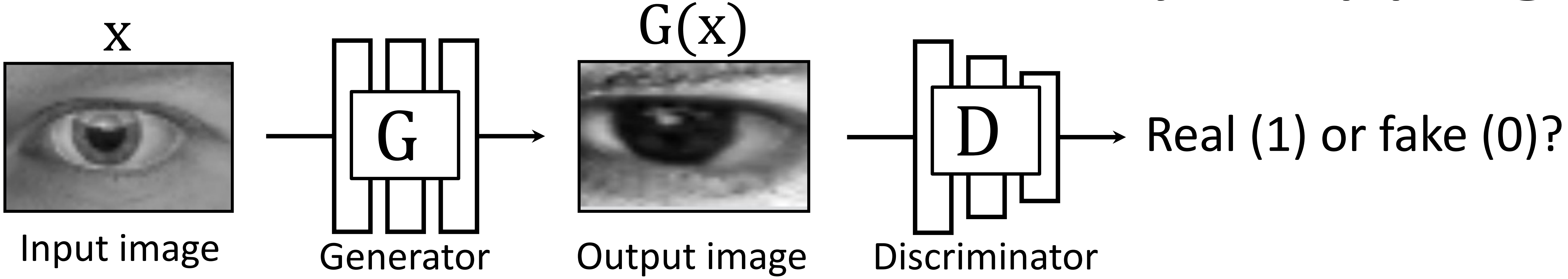
Unsupervised Learning of $p(y \mid x)$



- artifacts
- ignore inputs

[Goodfellow et al. 2014]

Additional Constraint: Identity Mapping

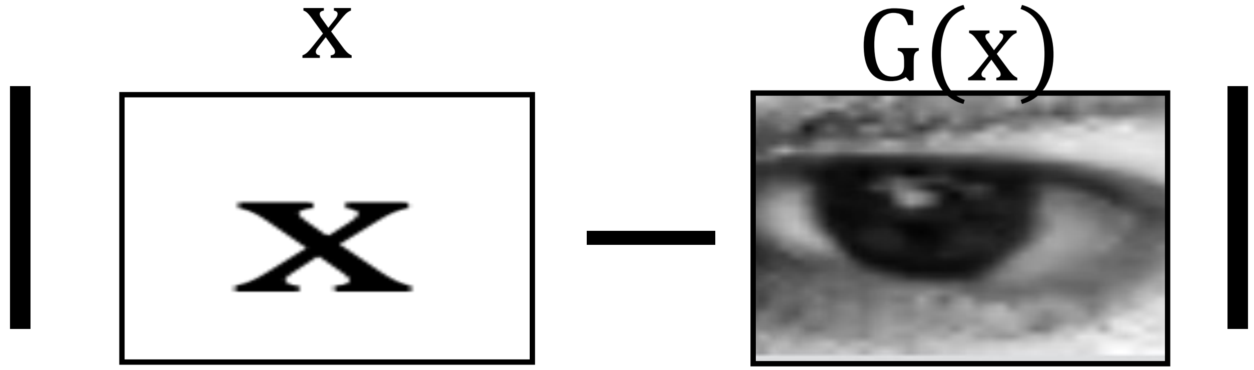


Adversarial loss

$$\mathbb{E}_x \log(1 - D_Y(G(x))) + \mathbb{E}_y \log D_Y(y)$$

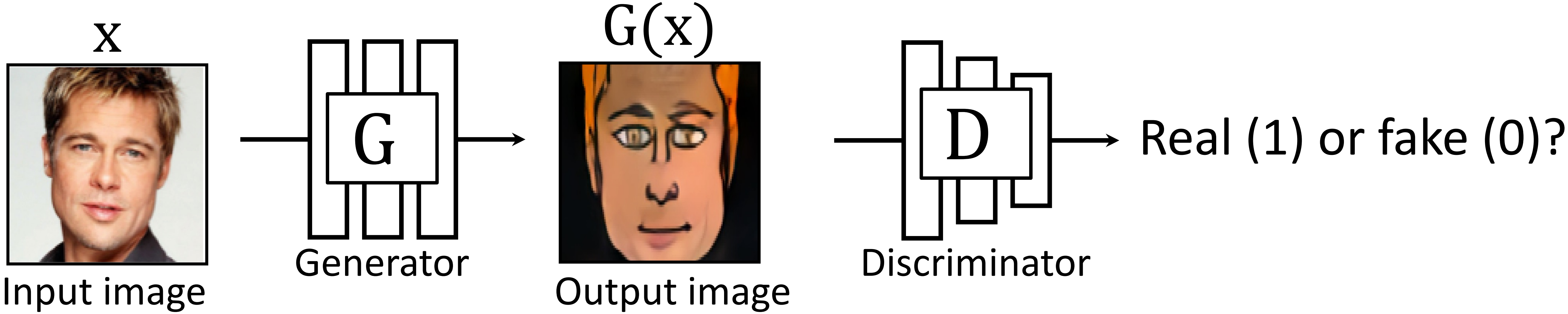
Self-Regularization loss

$$\mathbb{E}_x ||G(x) - x||_1$$



SimGAN [Shrivastava et al., 2017]

Additional Constraint: Feature Loss

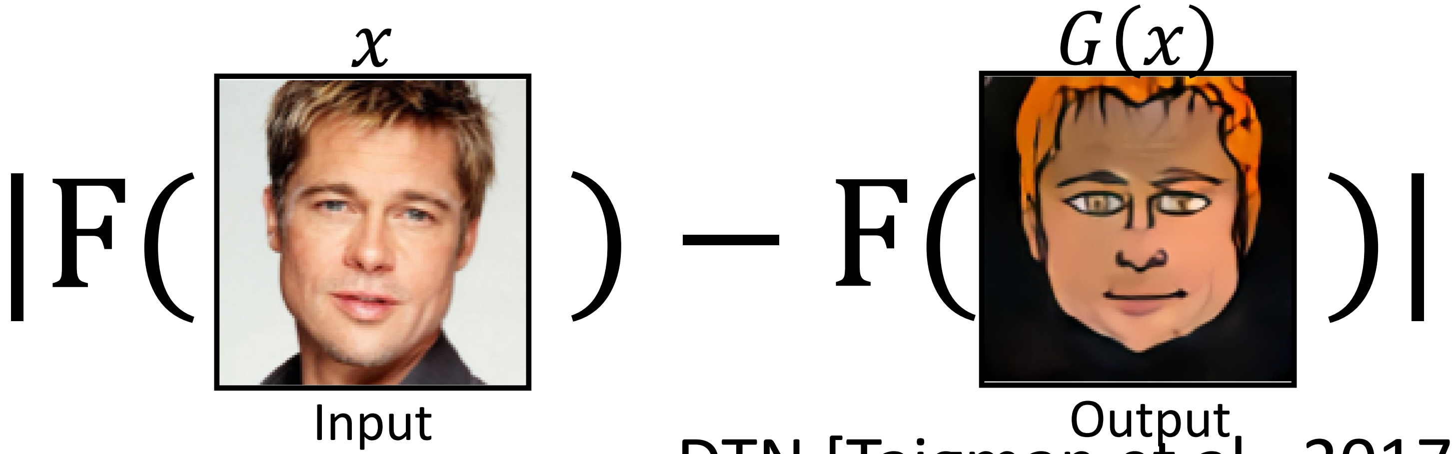


Adversarial loss

$$\mathbb{E}_x \log(1 - D_Y(G(x))) + \mathbb{E}_y \log D_Y(y)$$

Feature loss

$$\mathbb{E}_x ||F(G(x)) - F(x)||$$



Requires F to work across two domains

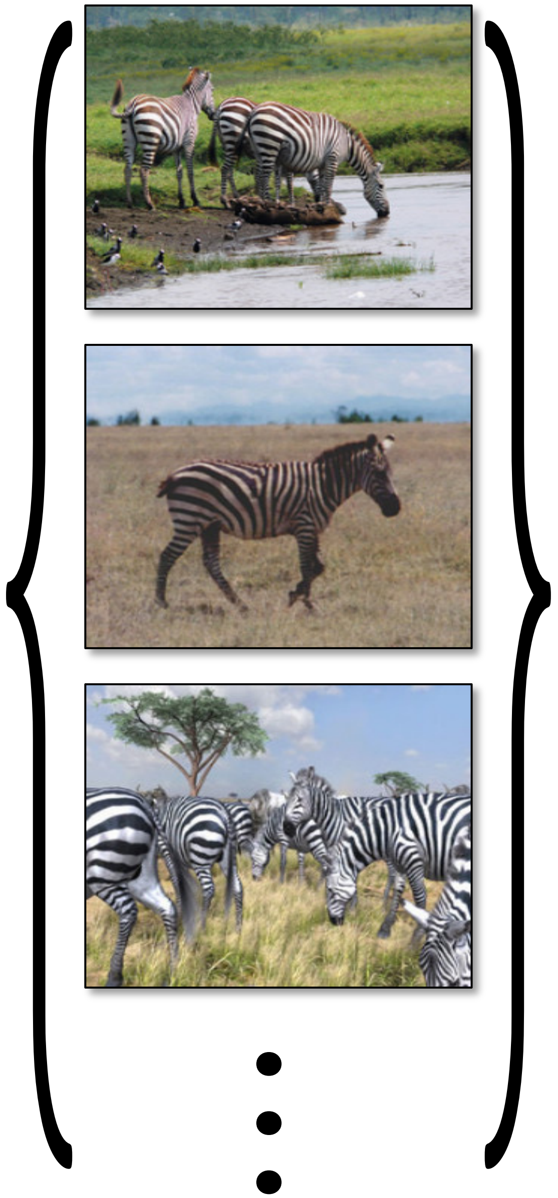
DTN [Taigman et al., 2017]

Additional Constraint: Cycle-Consistency

X

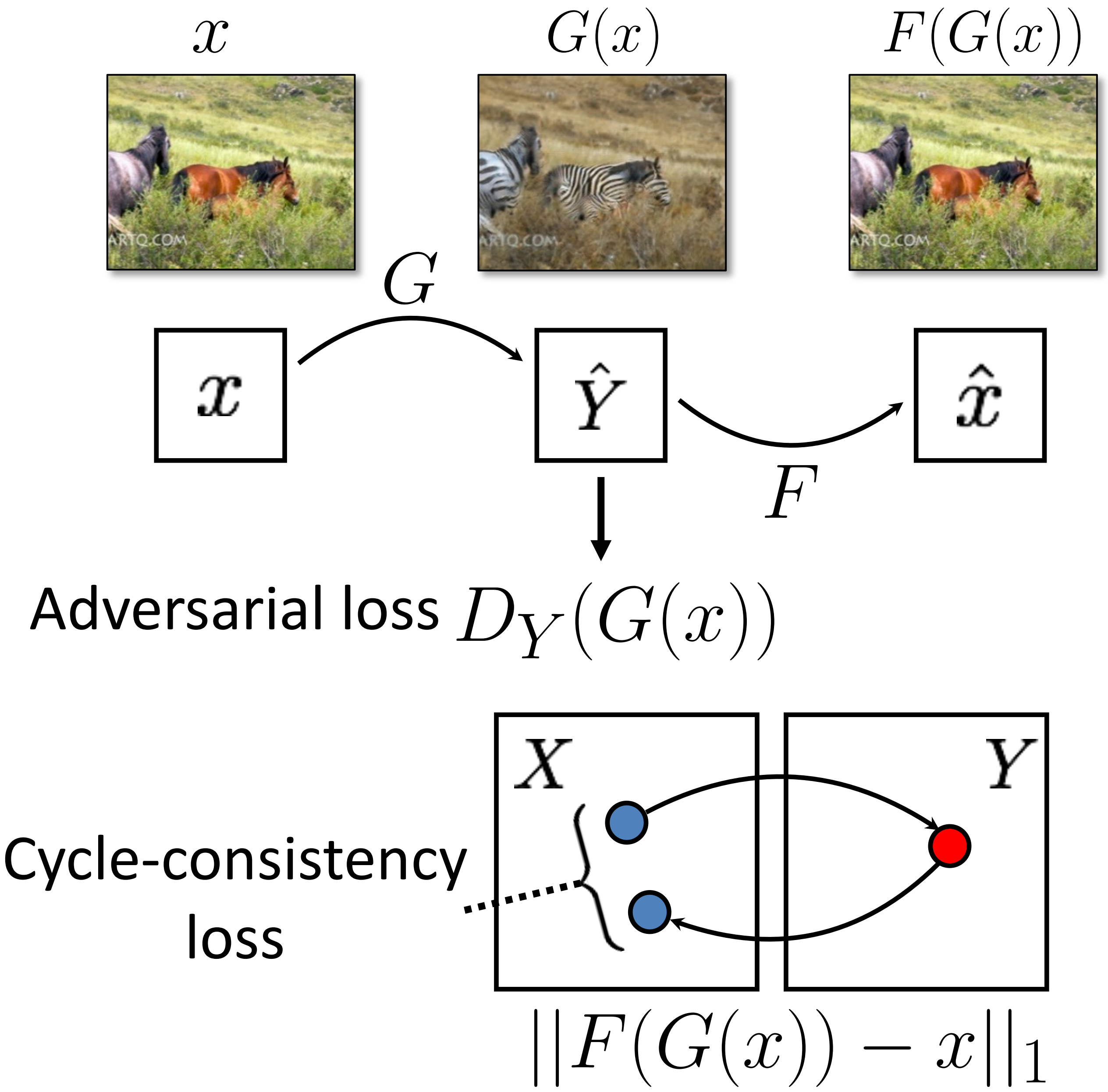
$$\begin{tikzcd} X \arrow[r, bend left] & Y \arrow[l, bend left] \arrow[d] \\ & D_Y \end{tikzcd}$$

Y



CycleGAN [Zhu*, Park* et al., ICCV 2017]

Cycle-Consistent Adversarial Networks



Adversarial loss

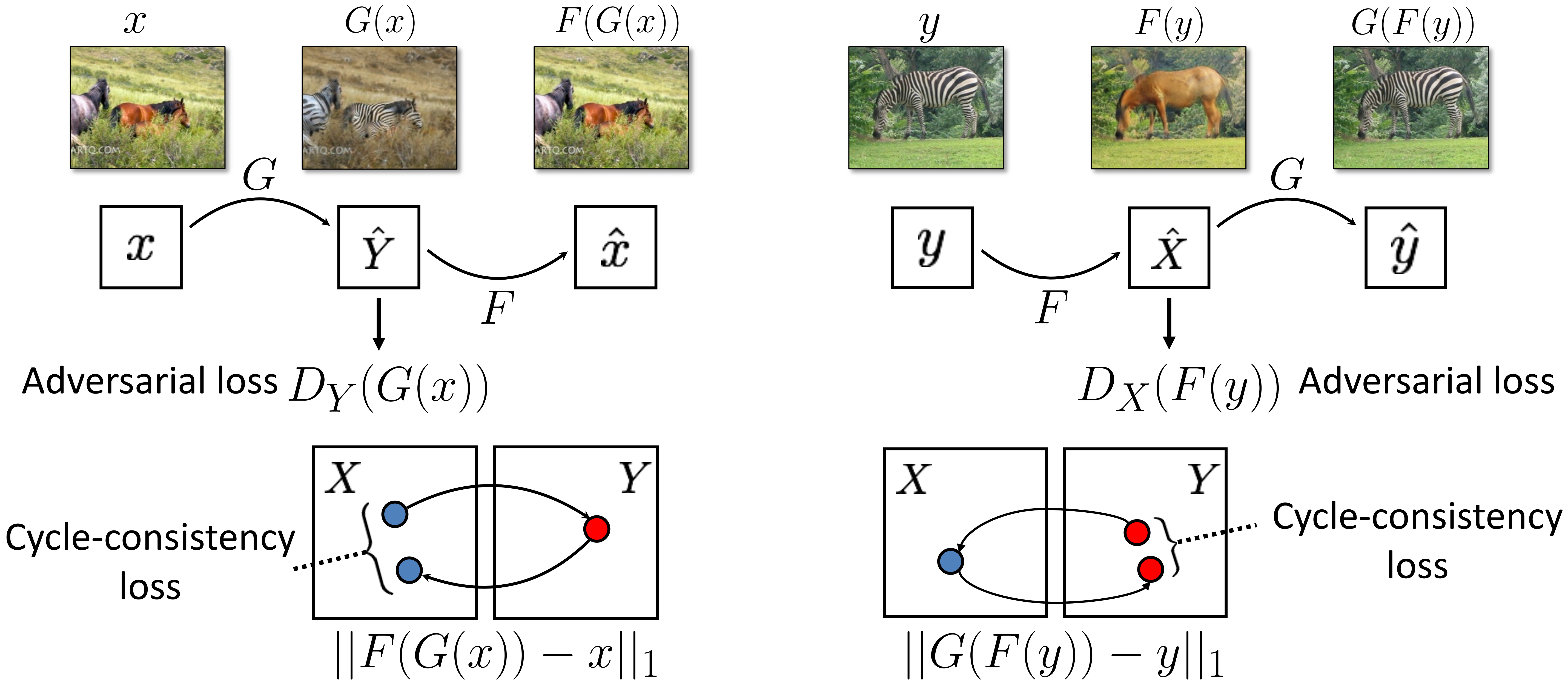
$$\mathbb{E}_x \log(1 - D_Y(G(x))) + \mathbb{E}_y \log D_Y(y)$$

Cycle-consistency loss

$$\mathbb{E}_x ||F(G(x)) - x||_1$$

CycleGAN [Zhu*, Park* et al., ICCV 2017]

Cycle-Consistent Adversarial Networks



CycleGAN [Zhu*, Park* et al., ICCV 2017]

Results

Horse → Zebra

Orange → Apple

Monet’s paintings → photographic style

Monet’s paintings → photographic style

Collection Style Transfer


Photograph ©Alexei Efros


Monet


Van Gogh


Cezanne


Ukiyo-e

Improving the Realism of CG Rendering

CG Game: Grand Theft Auto

Street view images in German cities

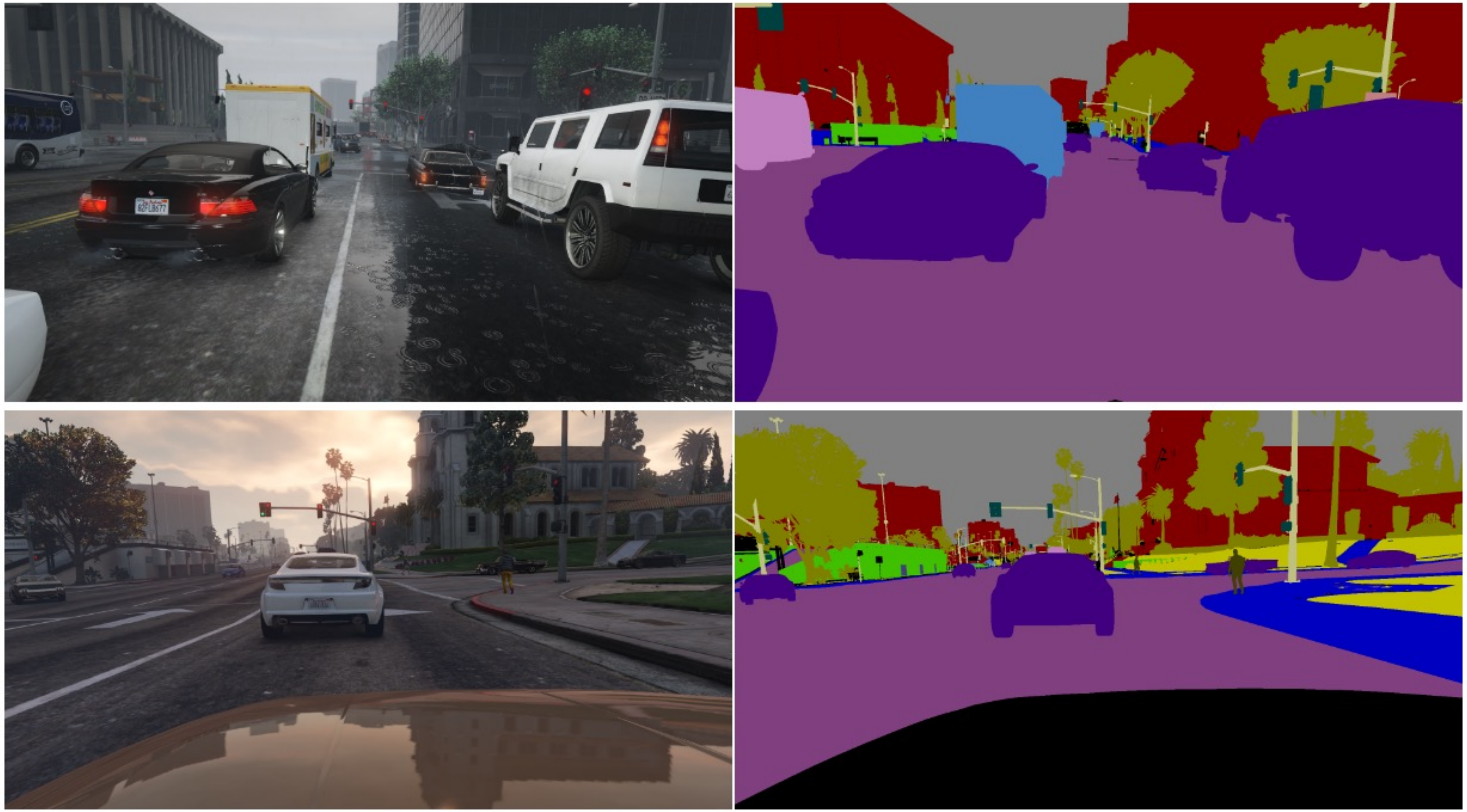
Data from [Richter et al., 2016], [Cordts et al, 2016]

Improving the Realism of CG Rendering

Output image with German street view style

Input CG image

Domain Adaptation with CycleGAN



CG images

Free segmentation labels

Data and labels from [Richter et al. 2016]

Domain Adaptation with CycleGAN

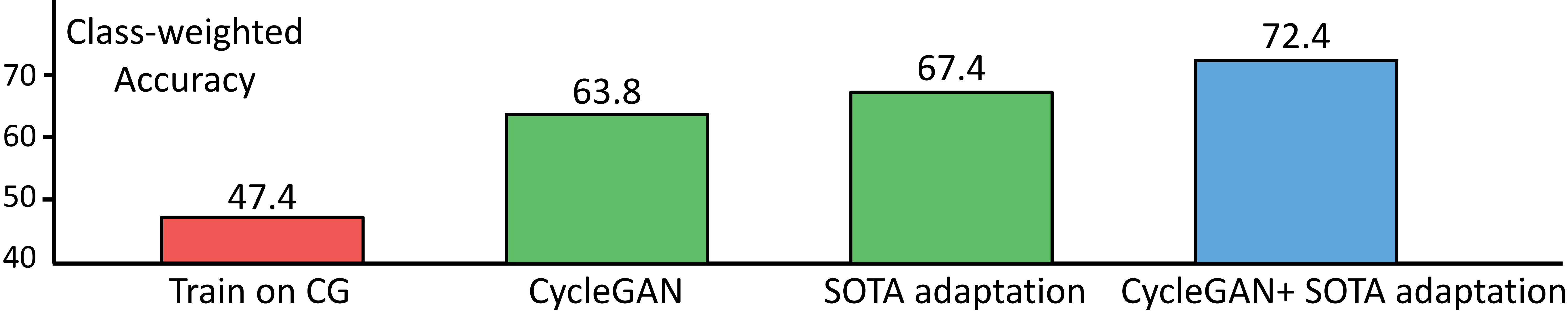
Train on CG data

Test on real images

Domain Adaptation with CycleGAN

Train on CycleGAN images

Test on real images



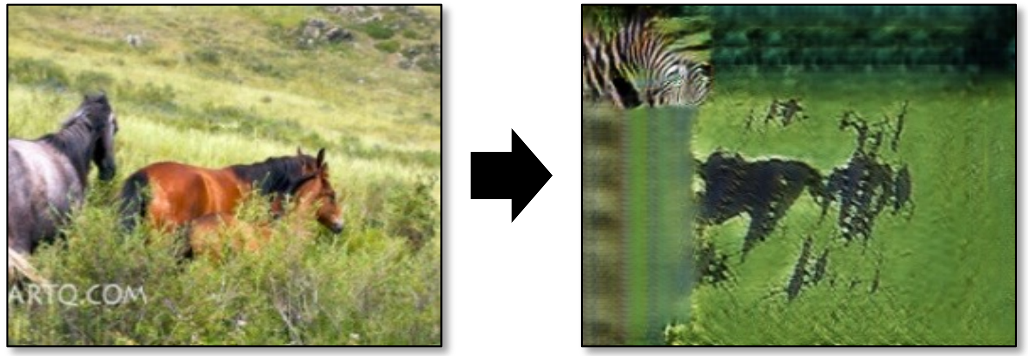
CycleGAN [Zhu*, Park* et al., ICCV 2017]
CycADA [Hoffman et al.,. ICML 2018]

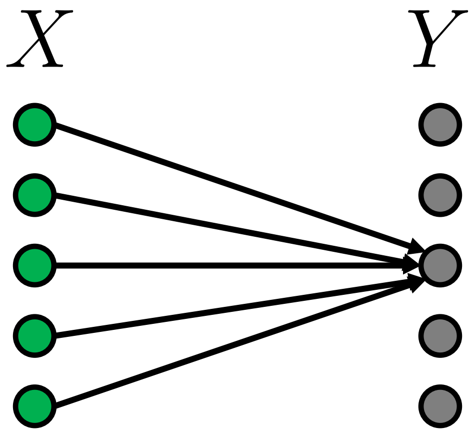
Why CycleGAN works

Why CycleGAN works

Adversarial loss

$$\mathbb{E}_x \log(1-D_Y(G(x)))+\mathbb{E}_y \log D_Y(y)$$

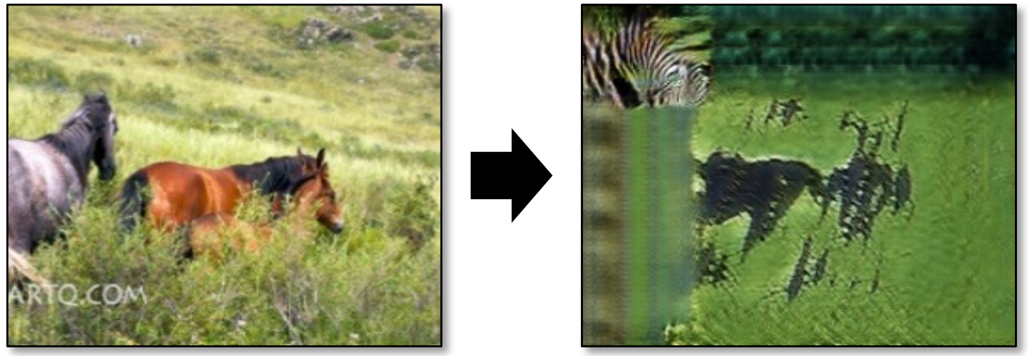




Why CycleGAN works

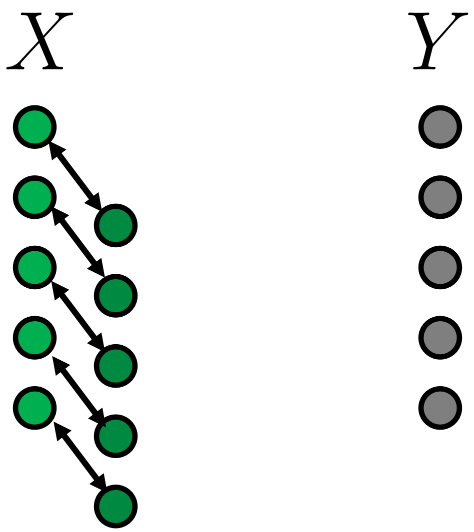
Adversarial loss

$$\mathbb{E}_x \log(1 - D_Y(G(x))) + \mathbb{E}_y \log D_Y(y)$$



Cycle-consistency loss

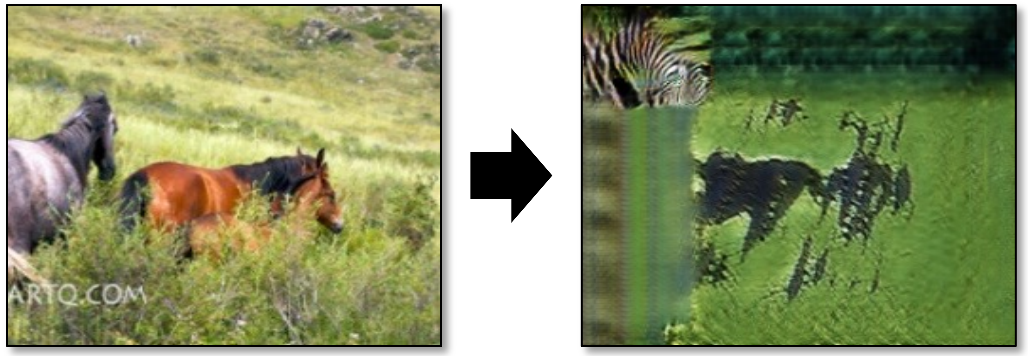
$$\mathbb{E}_x ||F(G(x)) - x||_1$$



Why CycleGAN works

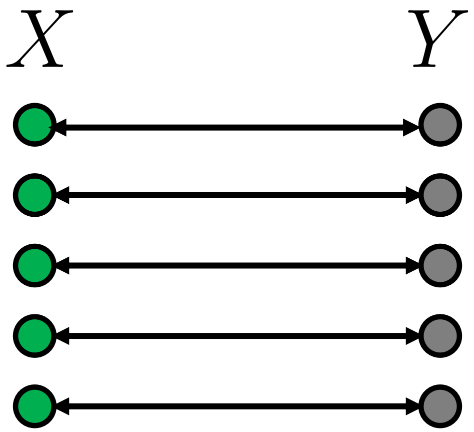
Adversarial loss

$$\mathbb{E}_x \log(1 - D_Y(G(x))) + \mathbb{E}_y \log D_Y(y)$$



Cycle-consistency loss

$$\mathbb{E}_x ||F(G(x)) - x||_1$$



Full objective

Why CycleGAN works

Adversarial loss

$$\mathbb{E}_x \log(1 - D_Y(G(x))) + \mathbb{E}_y \log D_Y(y)$$

Cycle-consistency loss

$$\mathbb{E}_x ||F(G(x)) - x||_1$$

x

$G(x)$

$F(G(x))$

Auto-encoder w/ domain prior

x — Encoder G → code $\hat{Y}$ — Decoder F → $\hat{x}$

Constraint: $\text{len}(x) = 30$ $G(x) \sim p_{data}(Y)$

[Hinton and Salakhutdinov. Science 2006]

Why CycleGAN works

Adversarial loss

$$\mathbb{E}_x \log(1 - D_Y(G(x))) + \mathbb{E}_y \log D_Y(y)$$

Under-constrained problem

Cycle-consistency loss

$$\mathbb{E}_x ||F(G(x)) - x||_1$$

Prior of G

<u>A strong regularizer</u>

Assumption: simple invertible function

Probabilistic Interpretation : Upper bound of conditional entropy $H(y|x)$

[Li et al. 2017]

Why CycleGAN works

Adversarial loss

$$\mathbb{E}_x \log(1 - D_Y(G(x))) + \mathbb{E}_y \log D_Y(y)$$

Cycle-consistency loss

$$\mathbb{E}_x||F(G(x)) - x||_1$$

flip the image

$P \circ G$

$F \circ P^{-1}$

flip the image again

<u>Invertible Perturbation</u>

Adversarial loss: images are horizontally symmetric

Cycle-consistency loss : $||F \circ P^{-1}(P \circ G(x)) - x||$

Applications of CycleGAN

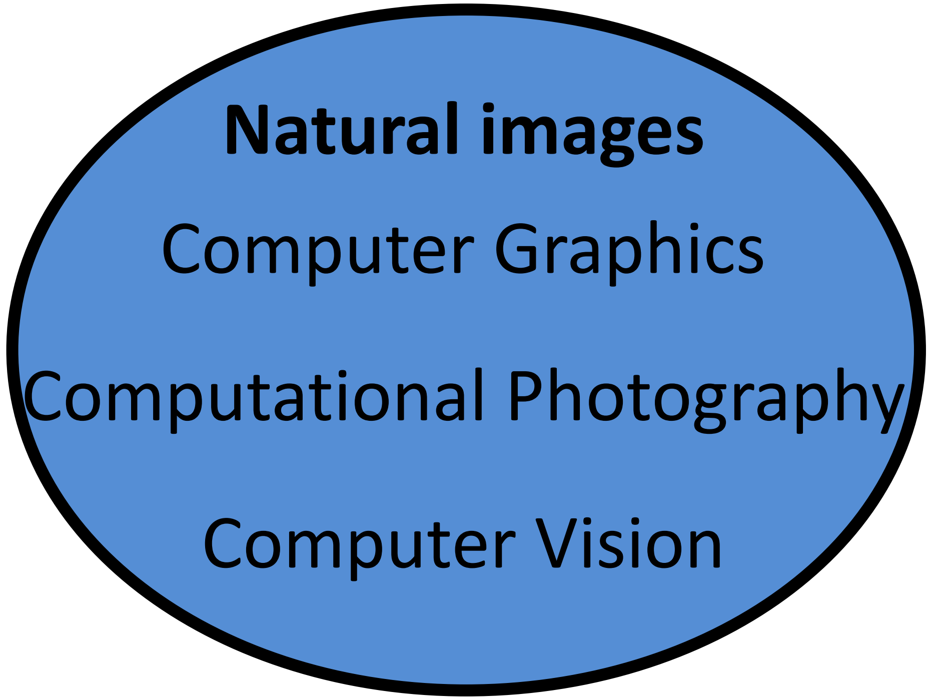

Natural images
Computer Graphics
Computational Photography
Computer Vision

Photo Enhancement

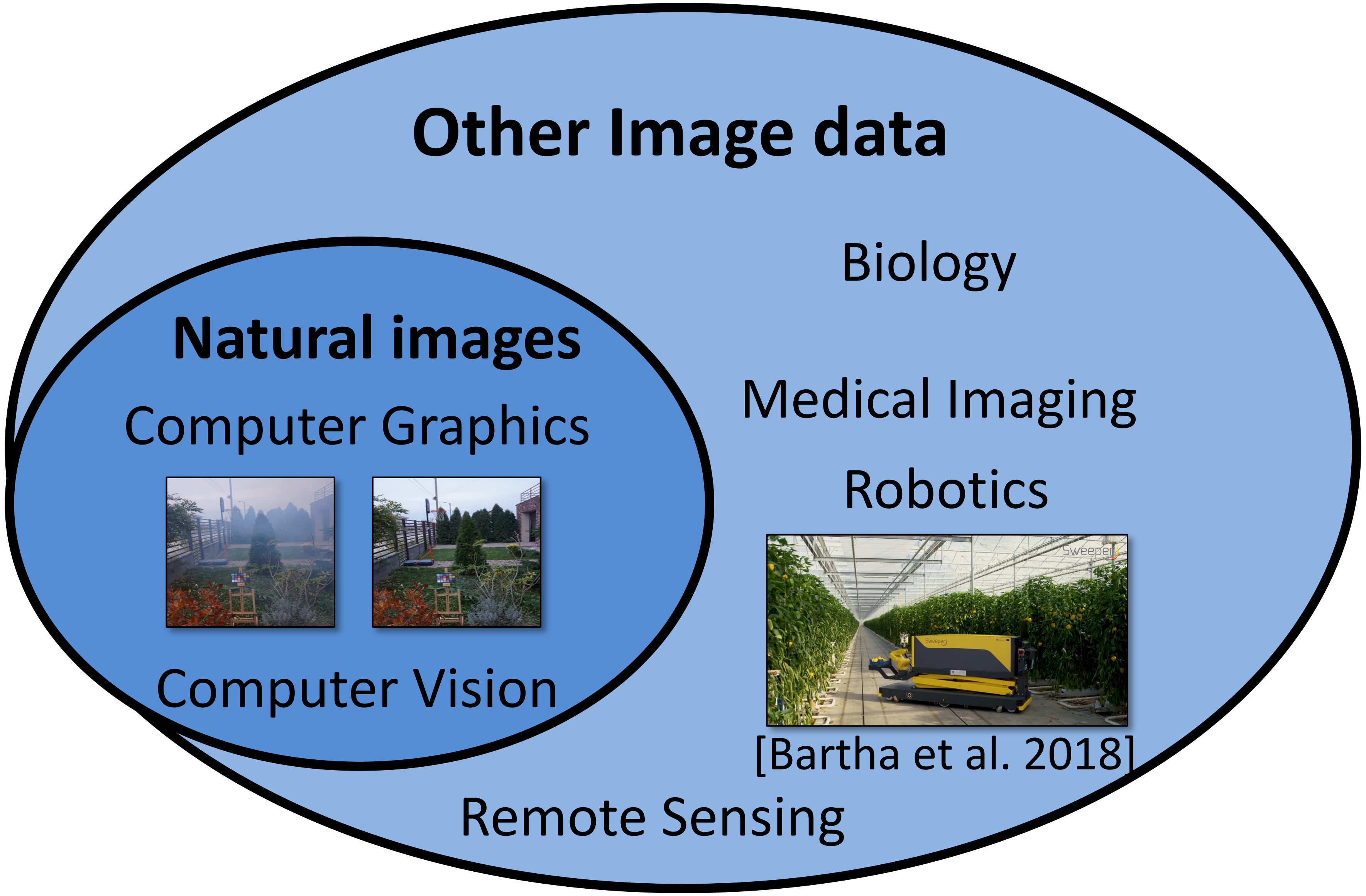
[Ignatov et al. CVPR 2018]

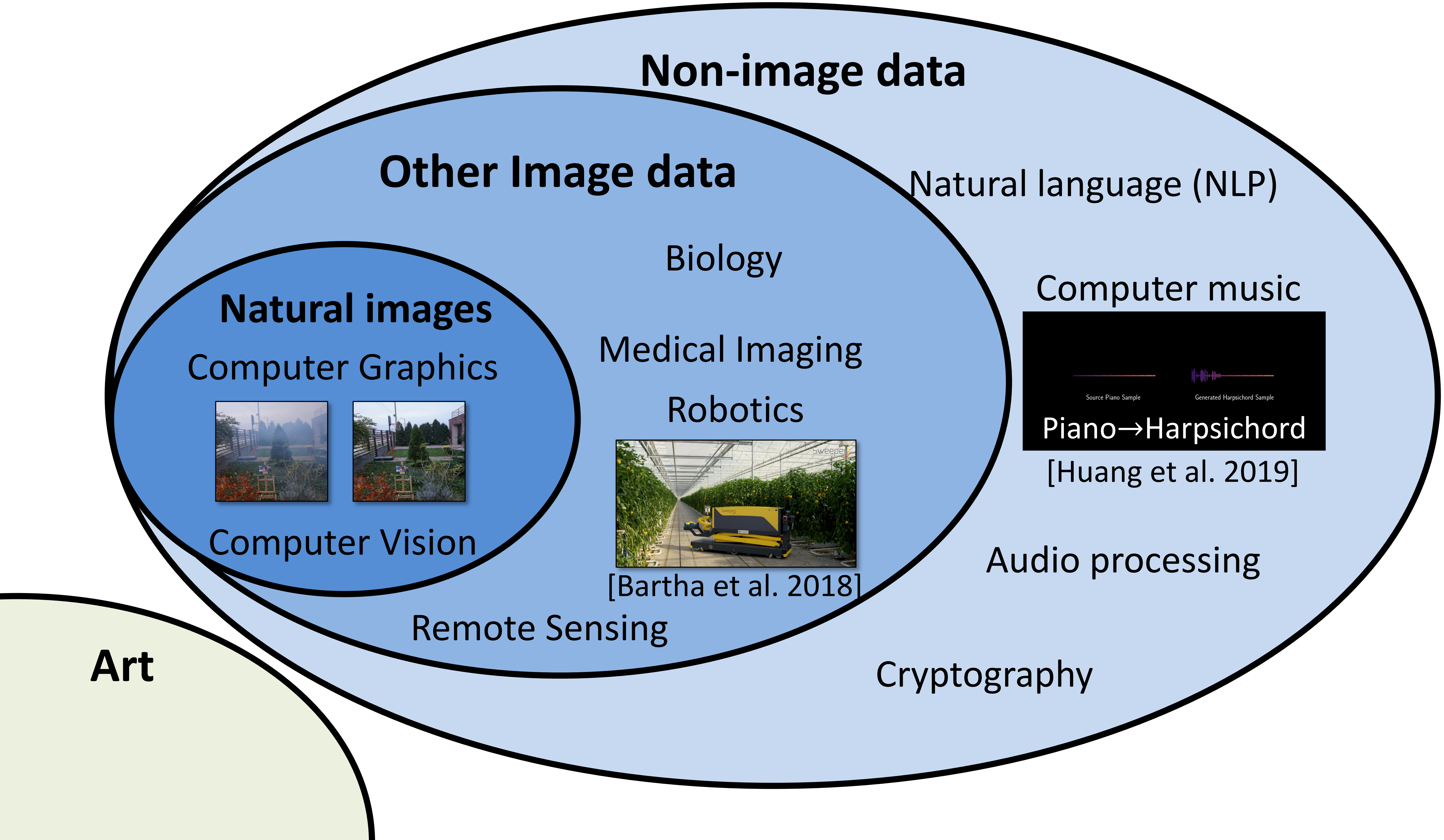
Image Dehazing

Foggy image

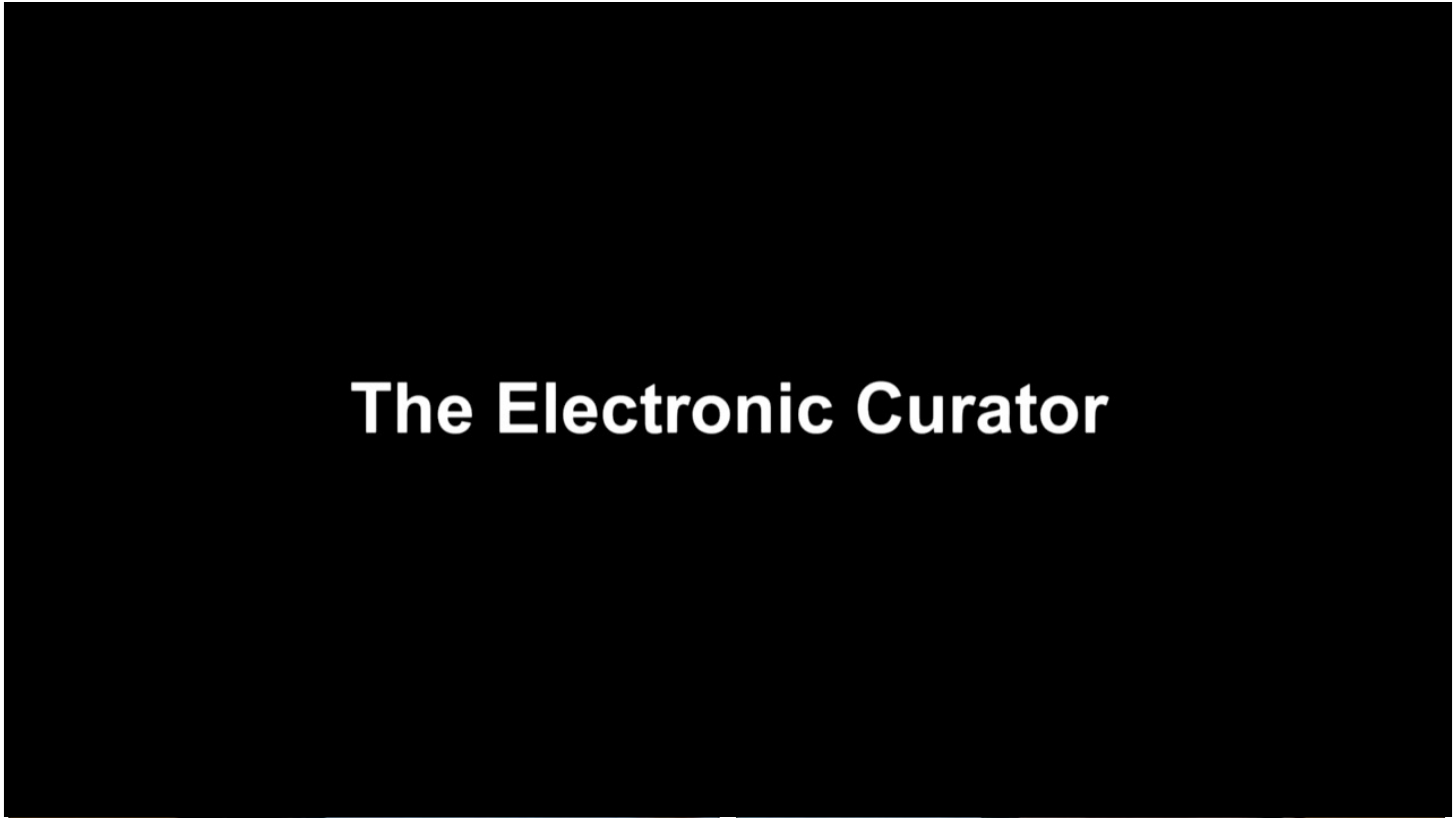
Clear image

[Engin et al. CVPRW 2018]

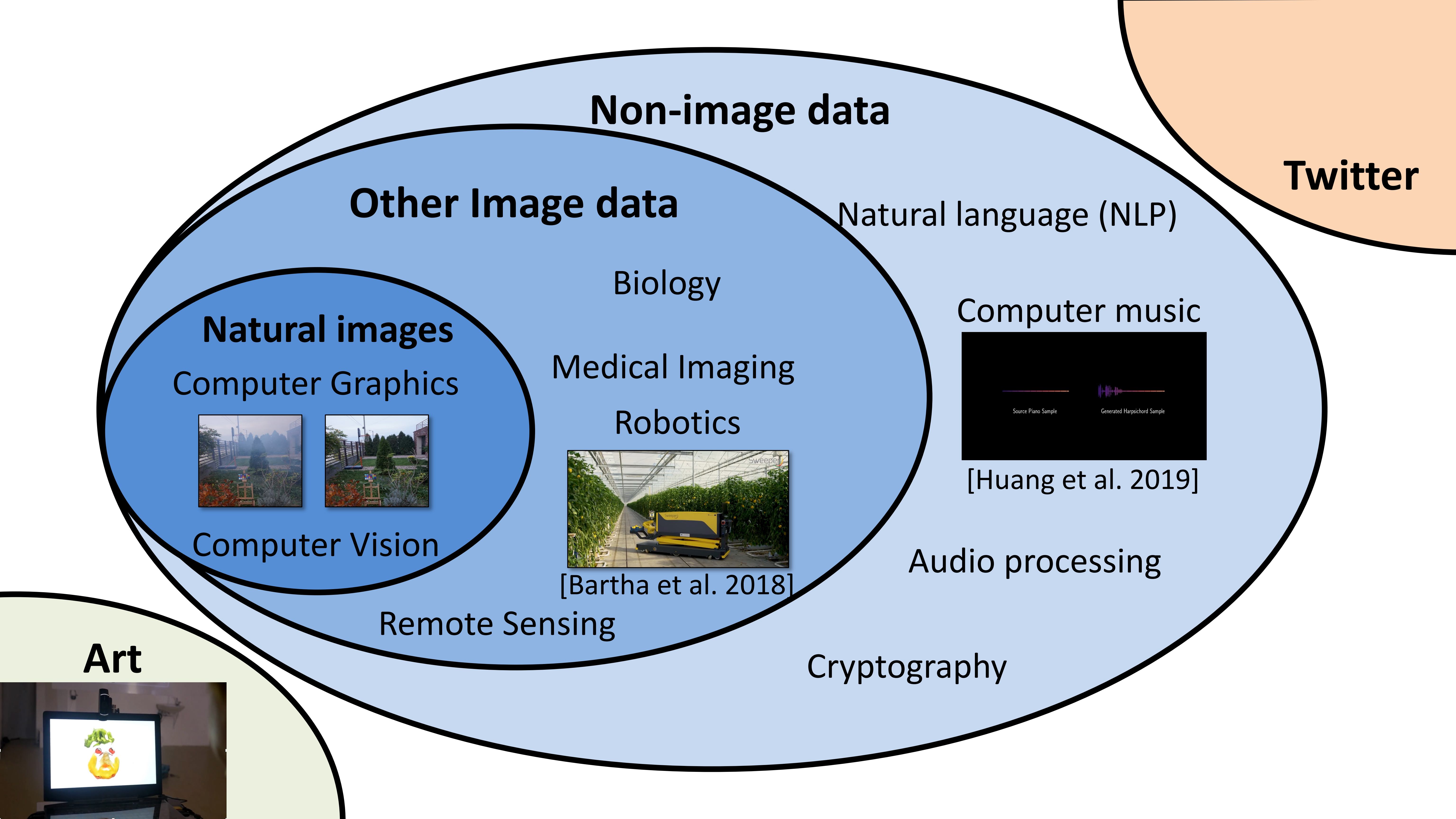

Other Image data
Natural images
Computer Graphics
Computer Vision
Biology
Medical Imaging
Robotics
[Bartha et al. 2018]
Remote Sensing


Non-image data
Other Image data
Natural language (NLP)
Biology
Natural images
Computer Graphics
Medical Imaging
Robotics
Computer music
Source Piano Sample
Generated Harpsichord Sample
Piano→Harpsichord
[Huang et al. 2019]
Computer Vision
[Bartha et al. 2018]
Audio processing
Remote Sensing
Art
Cryptography

Artistic Applications



“The Electronic Curator” © Eran Hadas and Eyal Gruss, israel


Non-image data
Twitter
Other Image data
Natural language (NLP)
Biology
Natural images
Computer music
Computer Graphics
Medical Imaging
Robotics
Source Piano Sample
Generated Harpsichord Sample
Sweeper
[Huang et al. 2019]
Computer Vision
Audio processing
[Bartha et al. 2018]
Remote Sensing
Art
Cryptography

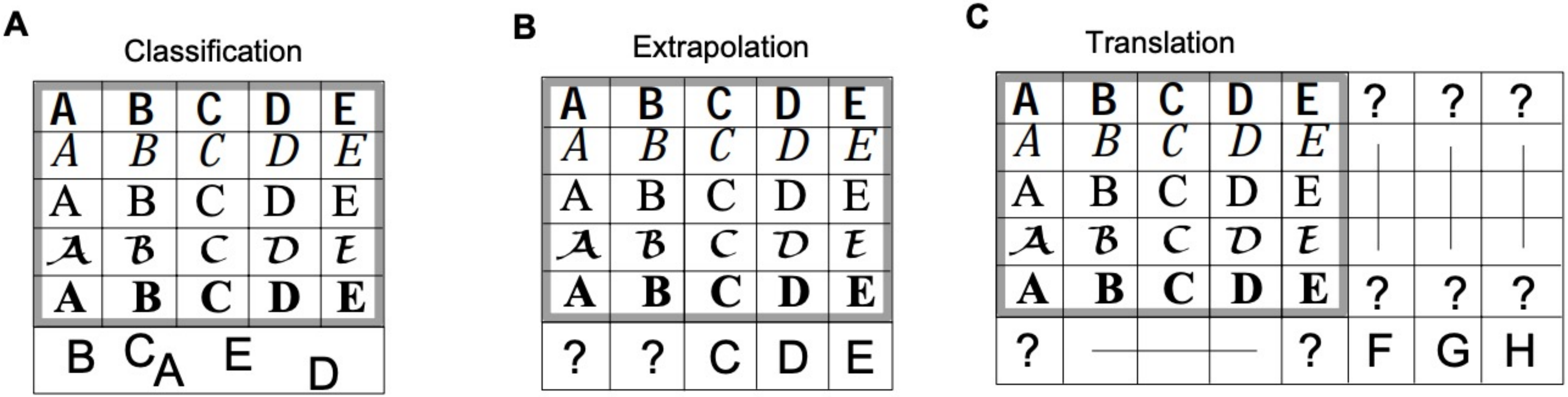
Latest from #CycleGAN

Input dog | Output cat | Input cat | Output dog

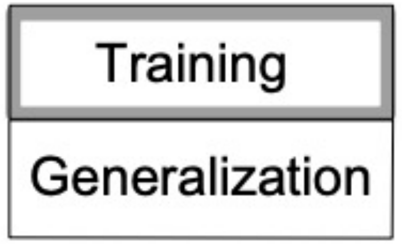
CycleGAN with modified architectures © itok_msi

Style and Content Disentanglement

Style and Content Separation



Domain Adaptation | Paired Image-to-Image Translation | Unpaired Image-to-Image Translation



Separating Style and Content
[Tenenbaum and Freeman 1996]

$$y_k^{sc} = \sum_{i=1}^{I} \sum_{j=1}^{J} w_{ijk} a_i^s b_j^c .$$

Style and Content

Adversarial loss

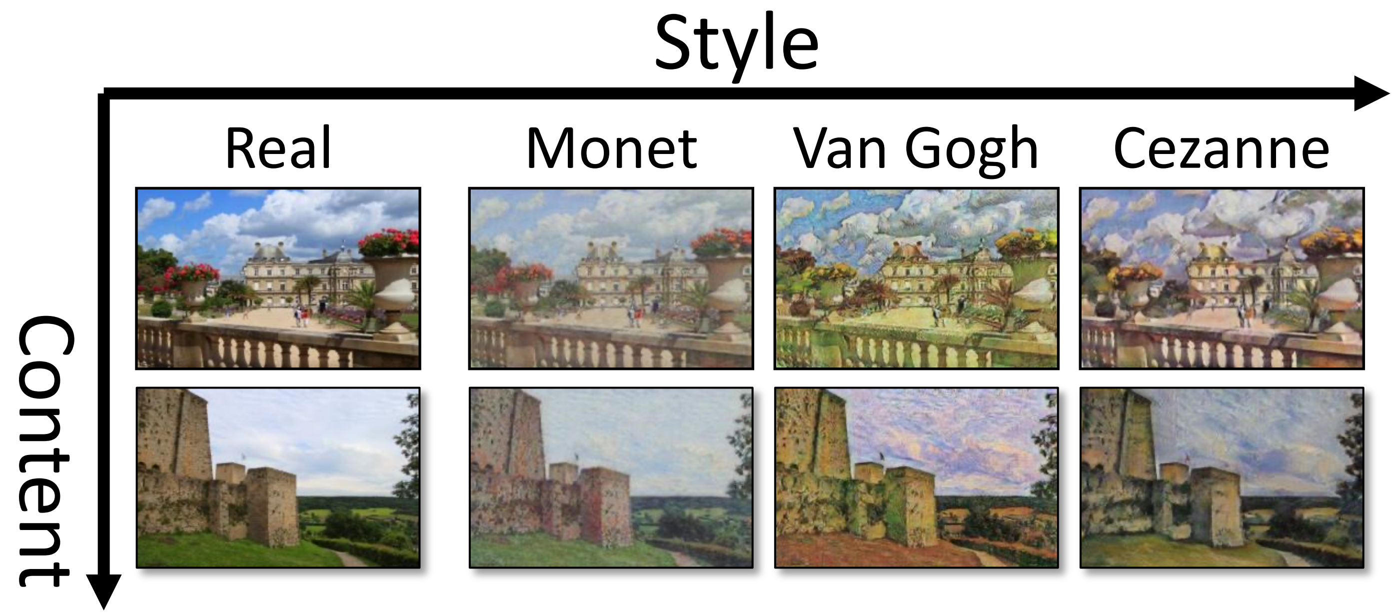
$$\mathbb{E}_x \log(1 - D_Y(G(x))) + \mathbb{E}_y \log D_Y(y)$$

$p(x) \rightarrow p(y)$ change **style**

Cycle-consistency loss

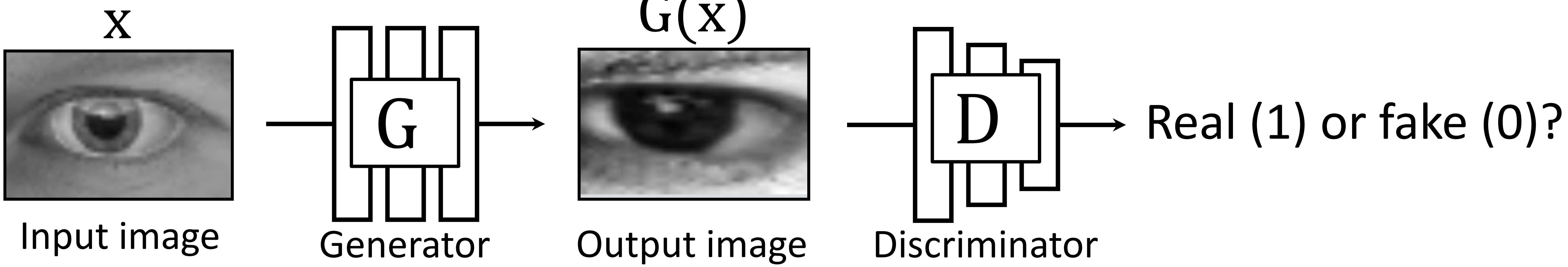
$$\mathbb{E}_x ||F(G(x)) - x||_1$$

Bidirectional: preserve **content**



Separating Style and Content
[Tenenbaum and Freeman 1996]

Style and Content

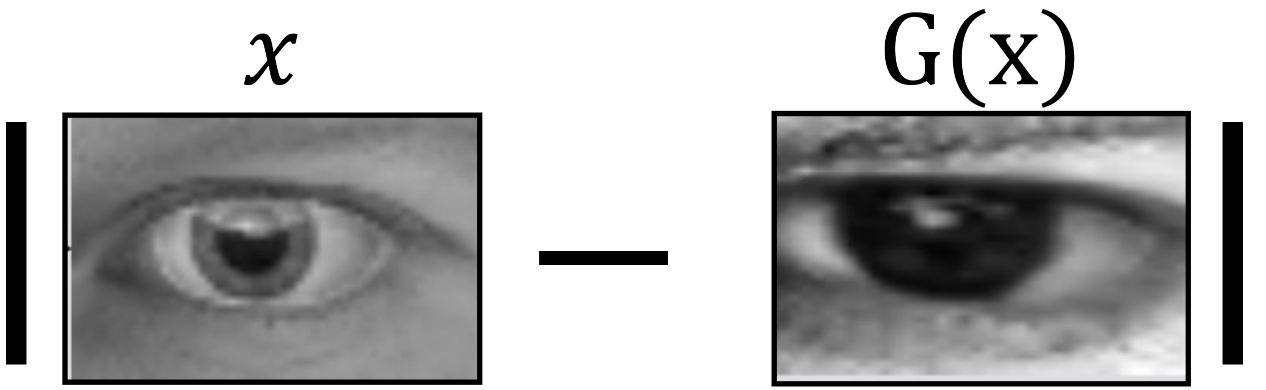


Adversarial loss (change style)

$$\mathbb{E}_x \log(1 - D_Y(G(x))) + \mathbb{E}_y \log D_Y(y)$$

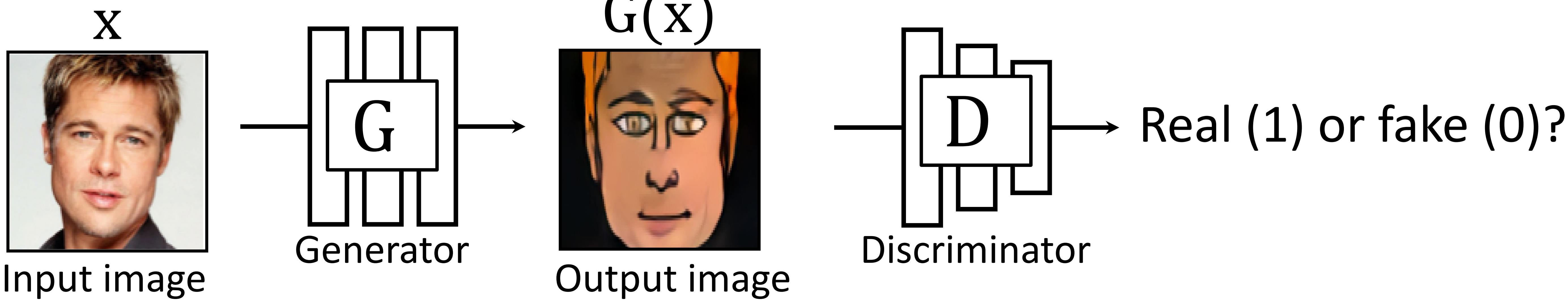
L1 loss (preserve content in pixel space)

$$\mathbb{E}_x ||G(x) - x||_1$$



SimGAN [Shrivastava et al., 2017]

Style and Content



Adversarial loss (change style)

$$\mathbb{E}_x \log(1 - D_Y(G(x))) + \mathbb{E}_y \log D_Y(y)$$

Feature loss (Preserve content in feature space)

$$\mathbb{E}_x ||F(G(x)) - F(x)||$$

x

$|F($  $) - F($  $)|$

$G(x)$

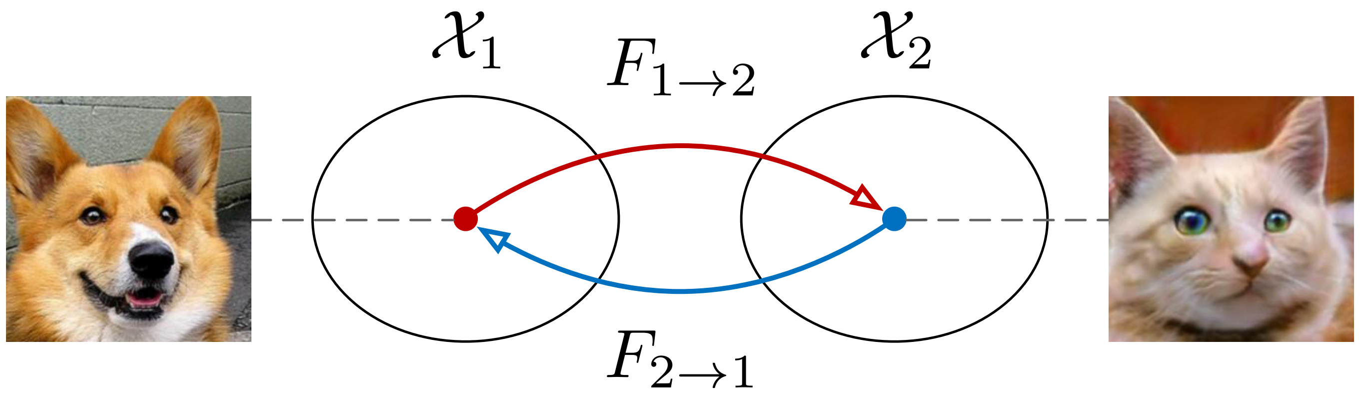
Input

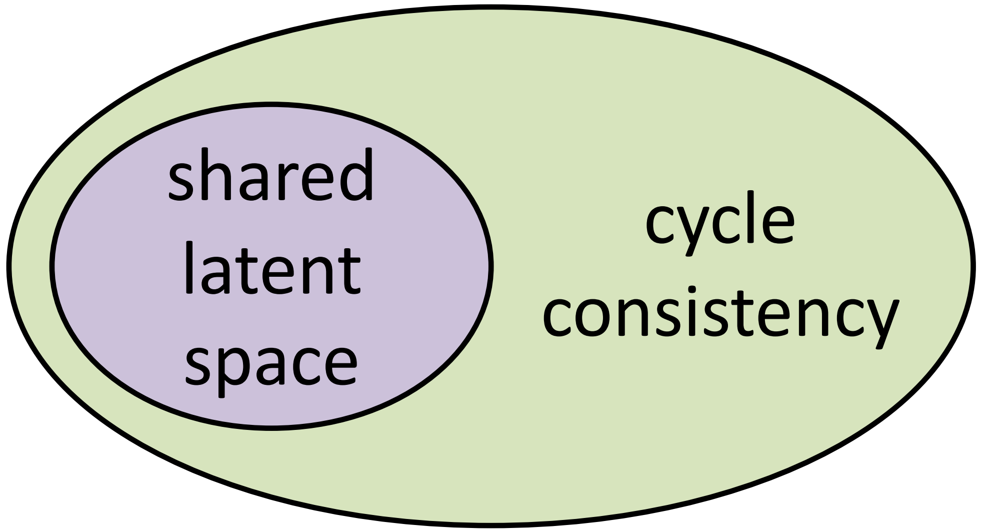
Output

DTN [Taigman et al., 2017]

CycleGAN and UNIT

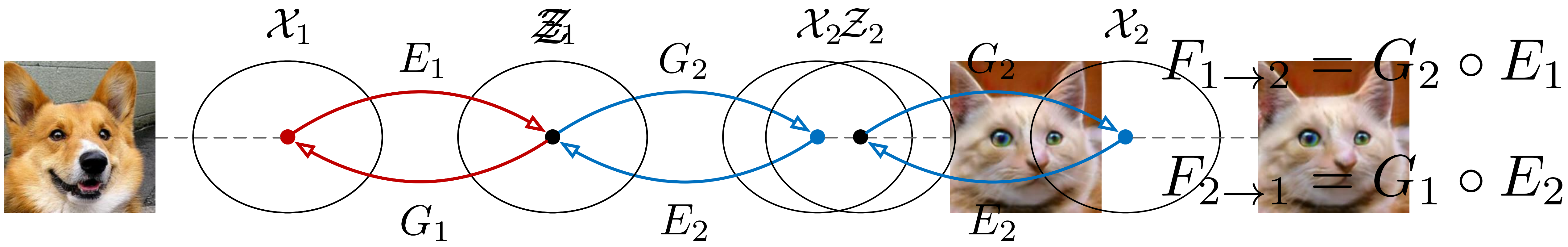
- CycleGAN (**cycle consistency**)





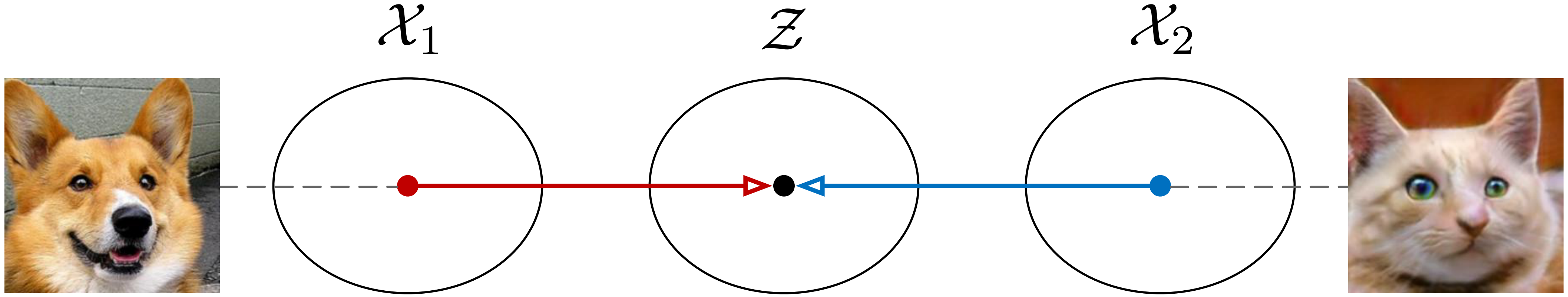
- UNIT (**shared latent space**) [Liu et al. 2017]

shared latent space $\Longrightarrow$ cycle consistency



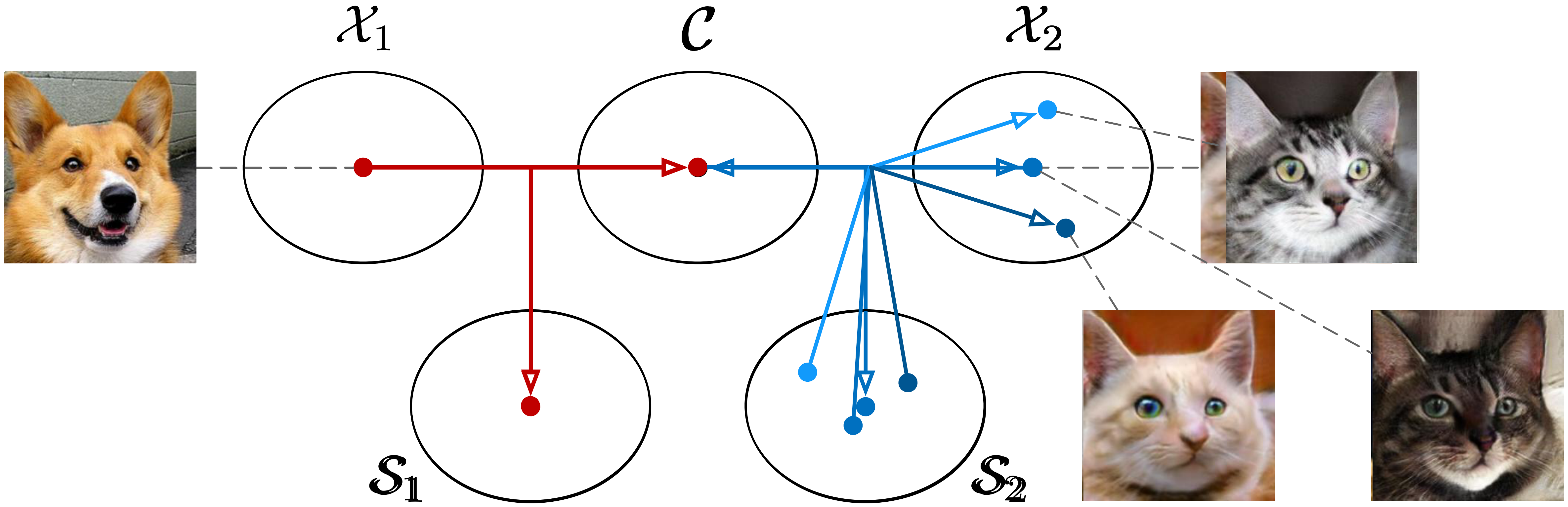
Disentangling the Latent Space

- UNIT
  - A single **shared, domain-invariant** latent space $\mathcal{Z}$

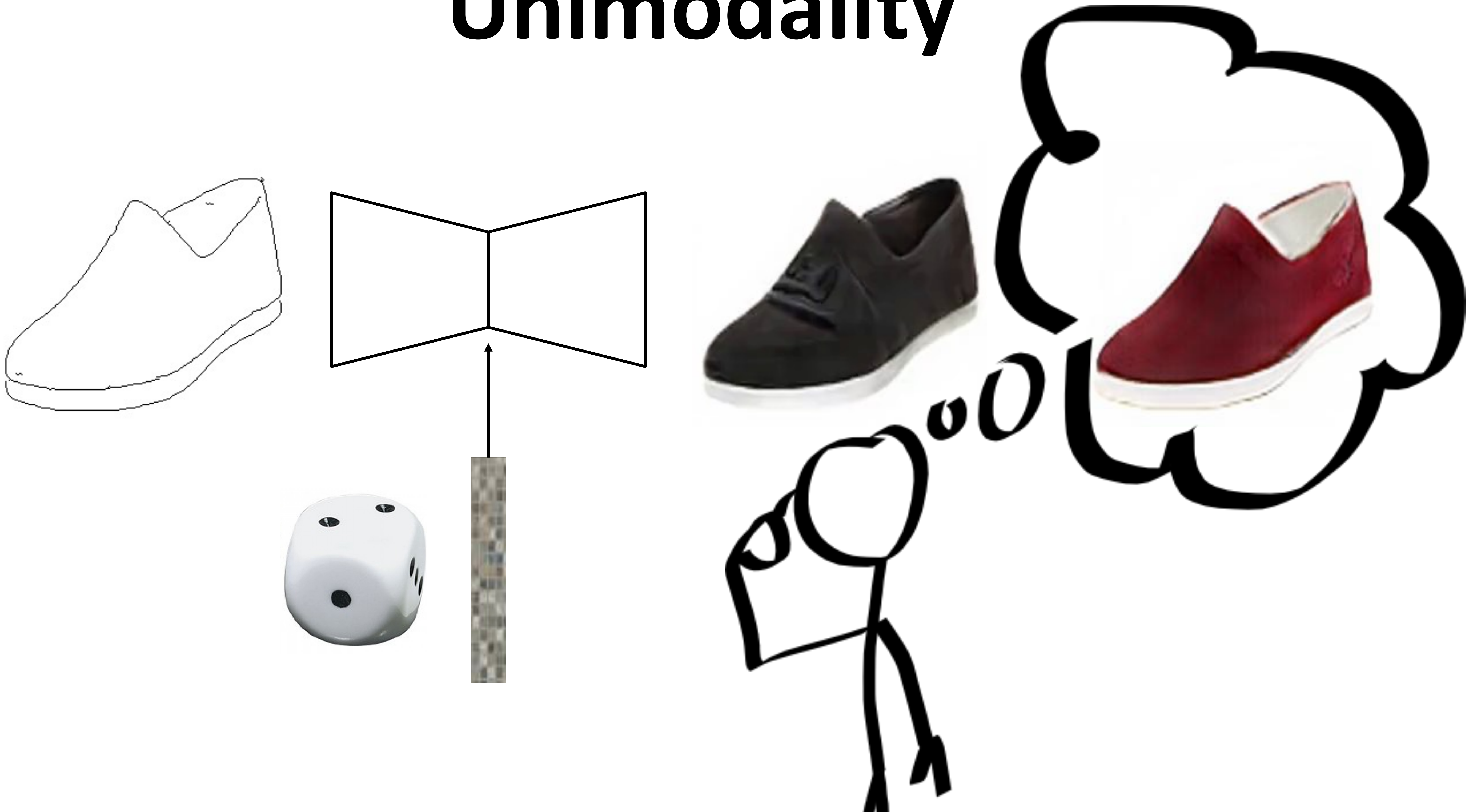


Disentangling the Latent Space

- Multimodal UNIT (MUNIT)
  - A **content** space $\mathcal{C}$ that is **shared, domain-invariant**
  - Two **style** spaces $\mathcal{S}_1, \mathcal{S}_2$ that are **unshared, domain-specific**



Unimodality

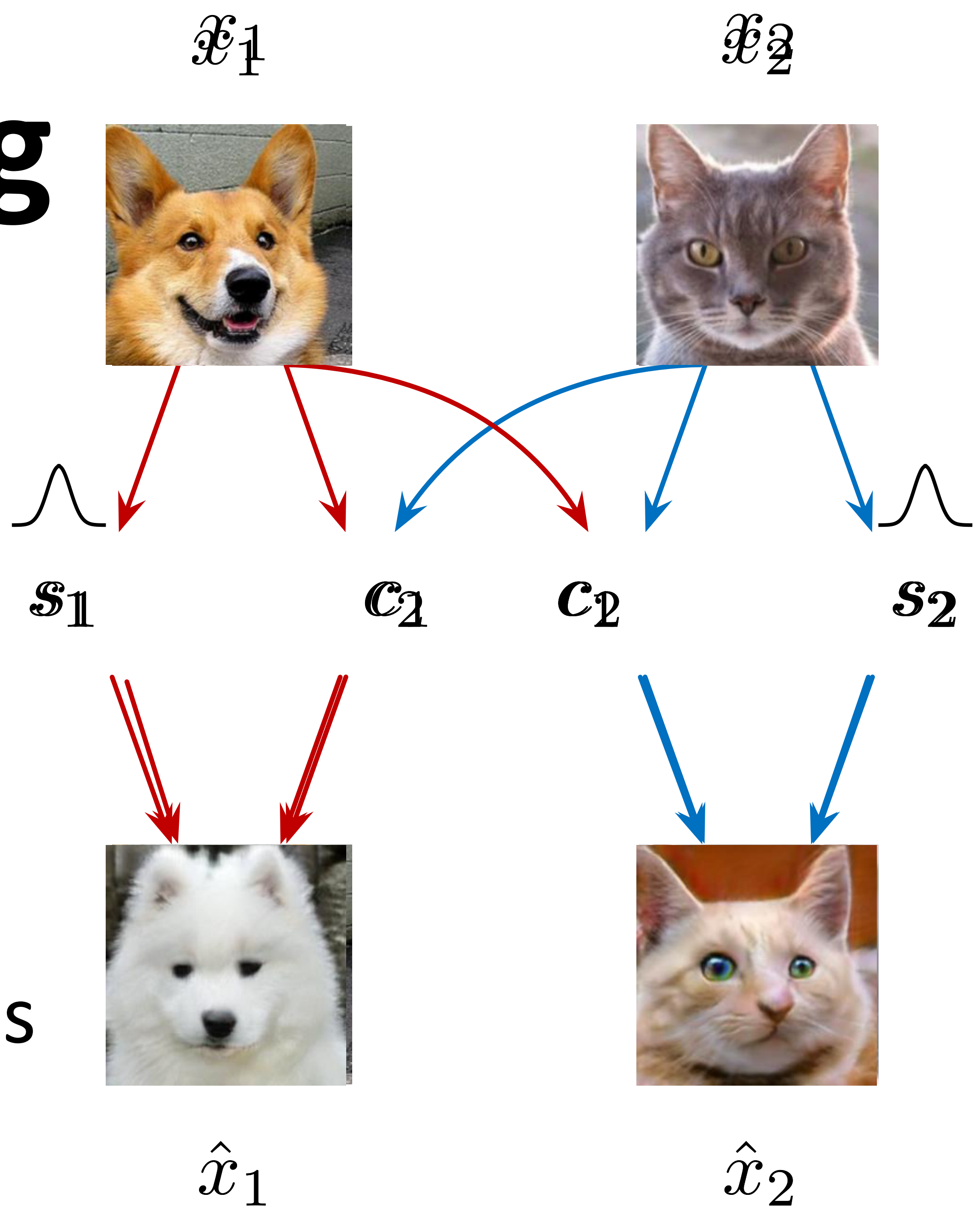


Towards Multimodality

...

Training

- Notations:
  - x: images
  - c: content
  - s: style
- Loss:
  - Bidirectional reconstruction loss
    - Image reconstruction loss
    - Latent reconstruction loss
  - GAN loss

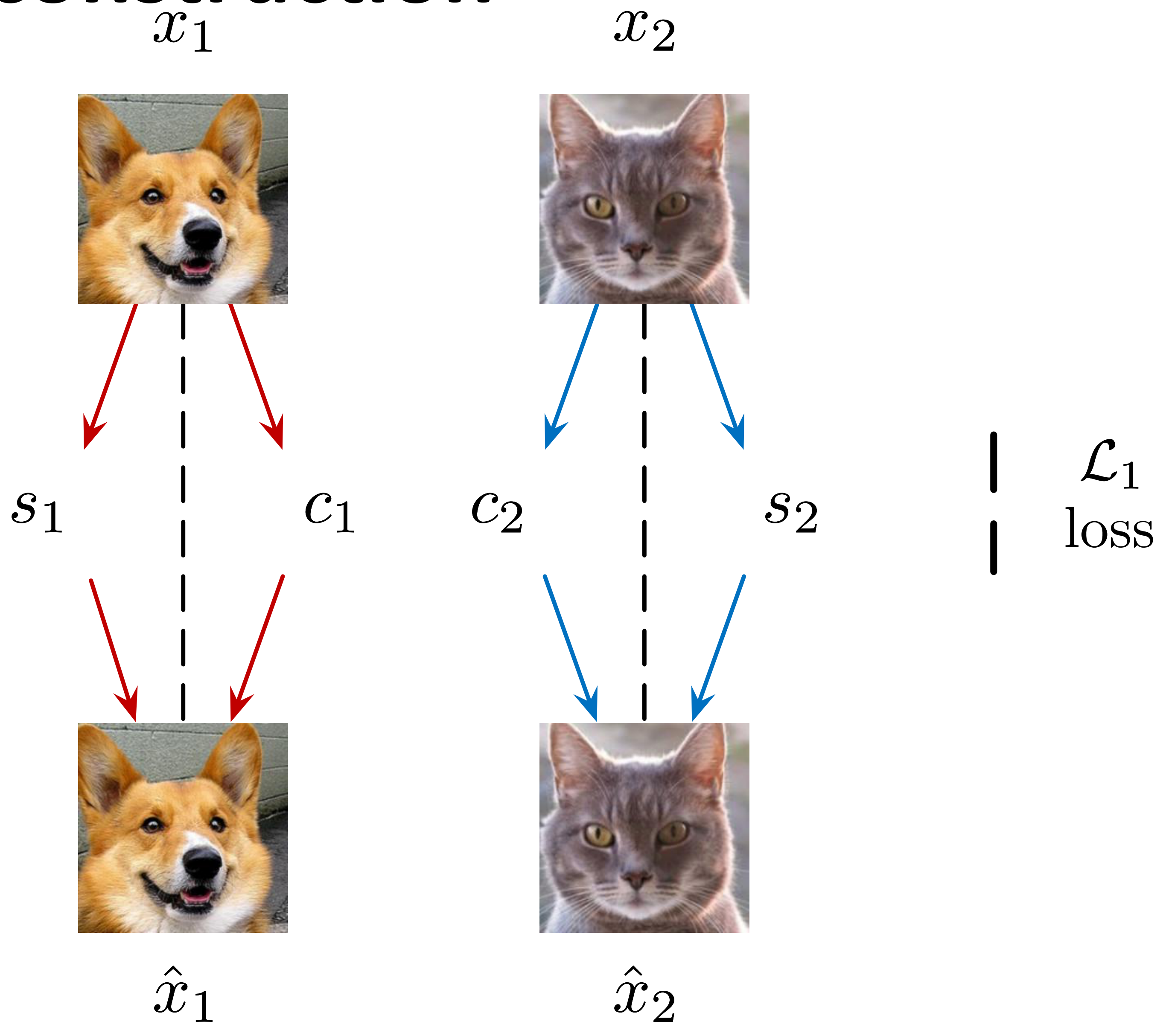


Within-domain reconstruction
Cross-domain translation

Bidirectional Reconstruction Loss: Image Reconstruction

Notations:

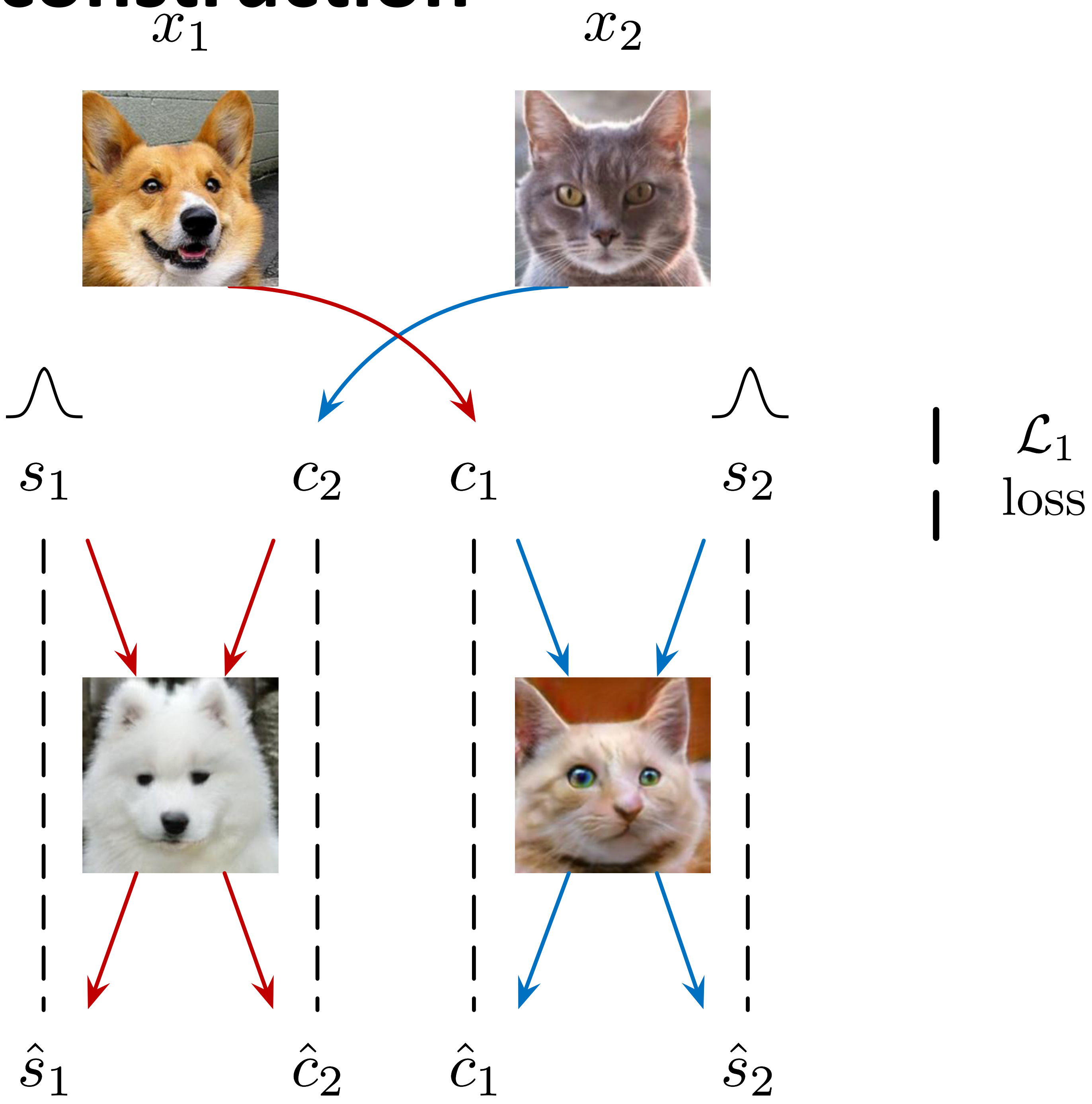
- x: images
- c: content
- s: style



Bidirectional Reconstruction Loss: Image Reconstruction

Notations:

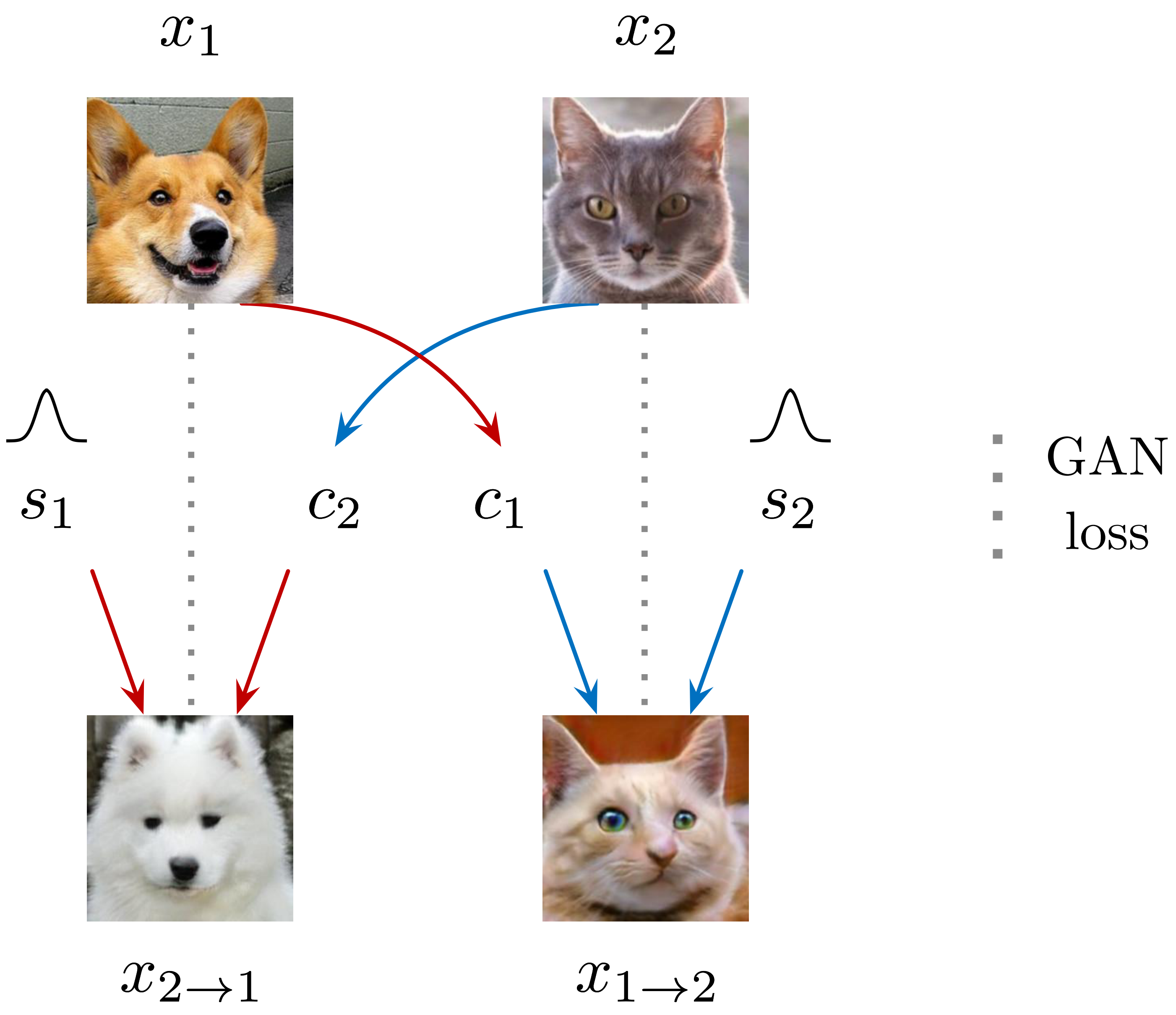
- x: images
- c: content
- s: style



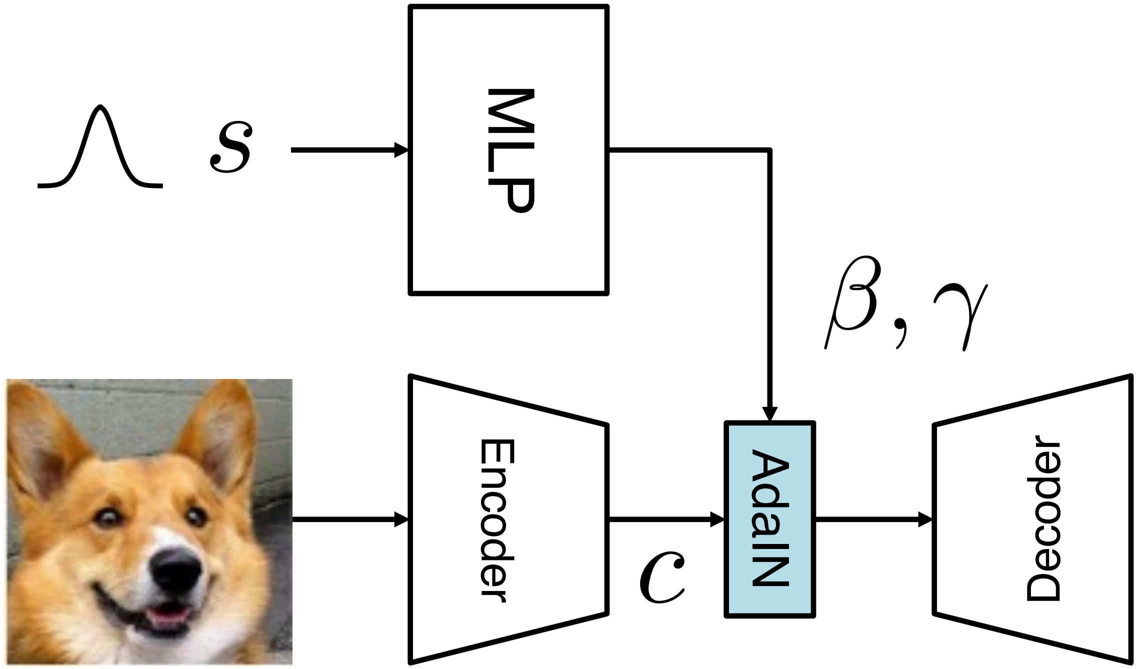
GAN Loss

Notations:

- x: images
- c: content
- s: style



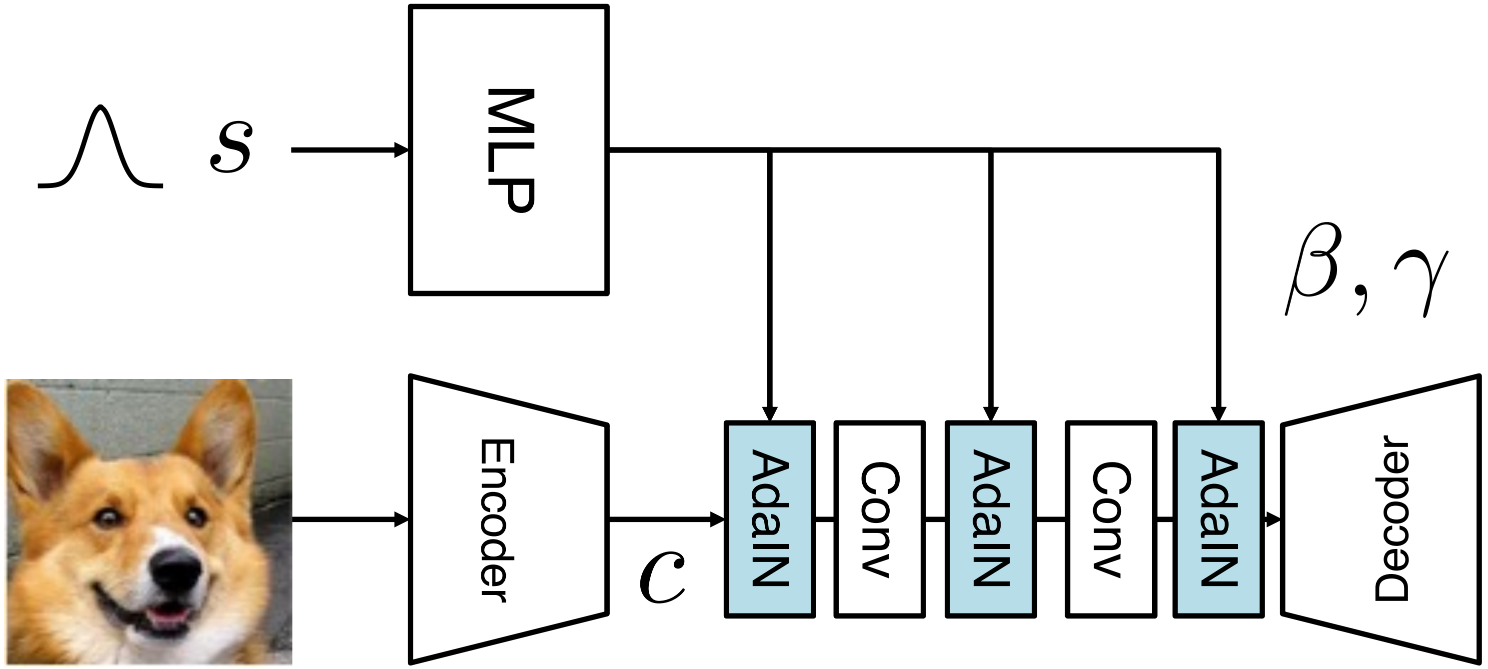
AdaIN in a Generative Network



$$\mathrm{AdaIN}(c, s) = \gamma\left(\frac{c - \mu(c)}{\sigma(c)}\right) + \beta$$

AdaIN in a generative network

AdaIN in a Generative Network



$$\mathrm{AdaIN}(c, s) = \gamma\left(\frac{c - \mu(c)}{\sigma(c)}\right) + \beta$$

AdaIN in a generative network

Sketches <-> Photo

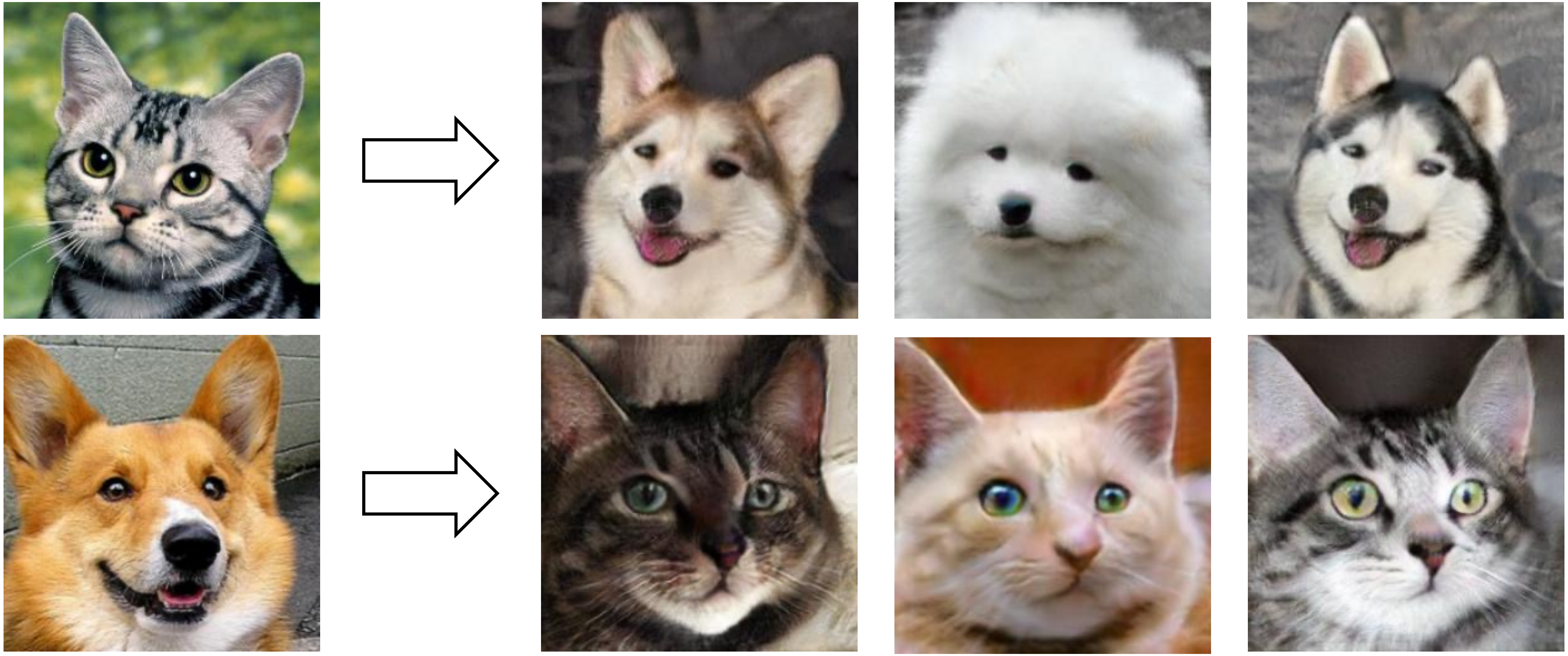
Input

Outputs

Cats ↔ Dogs

Input

Outputs

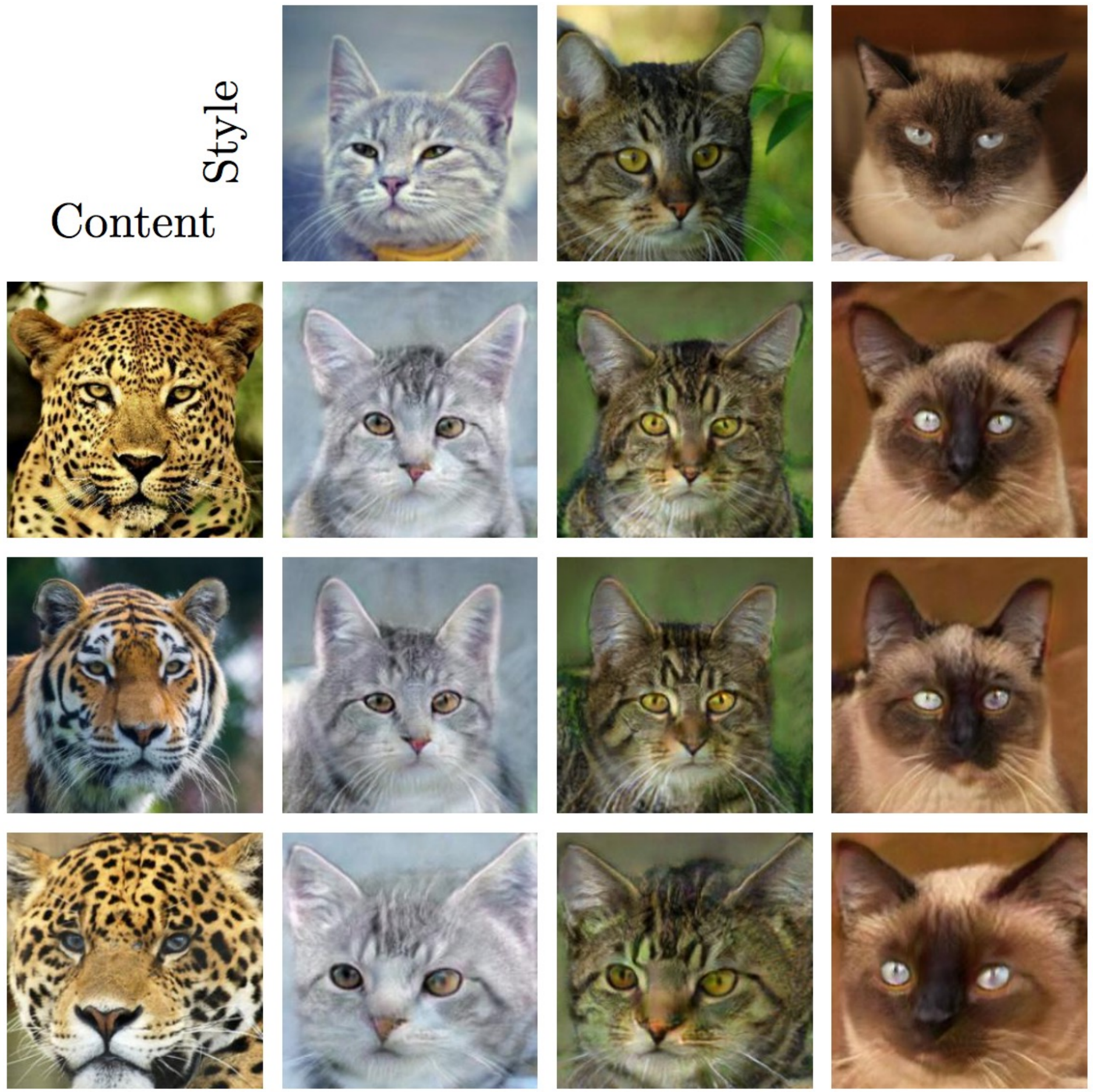


Synthetic ↔ Real

Input

Outputs

Example-guided Translation



Example-guided Translation

Thank You!

16-726, Spring 2022

https://learning-image-synthesis.github.io/sp22/