Ideal models (Dream)

Pros: good sample, fast sample, Exact/fast likelihoods
good coverage, easy to training, learn low-dimensional latent representation.
Cons: no cons

Autoregressive models

Pros: Exact likelihoods, good samples, good coverage
Cons: Slow to evaluate or sample

VAEs

Pros: Cheap to sample, good coverage
Cons: Blurry samples (in practice)

GANs

Pros: Cheap to sample, fast to train
Cons: No likelihoods (density), hard to train

Diffusion models

Pros: Easy to train, good samples
Cons: slow to sample; slow/hard to compute likelihoods

[adapted from Phillip Isola and David Duvenaud]

No Free Lunch

Stable Training <-> Slow Inference

Hybrid Models

VQ-VAE2: VAE + autoregressive
VQ-GAN: VAE, GANs, Perceptual loss + autoregressive
Latent Diffusion: VAE, GANs, perceptual loss + diffusion models

Base model (feature): autoregressive, diffusion
Upsampler (feature->image): VAE, GANs, Perceptual loss

Model Distillation

Teacher: Multi-step Models (e.g., Latent Diffusion Models)
Student: Single-Step Model (e.g., Conditional GANs)
Distillation Loss: GAN Loss, Perceptual Loss, …

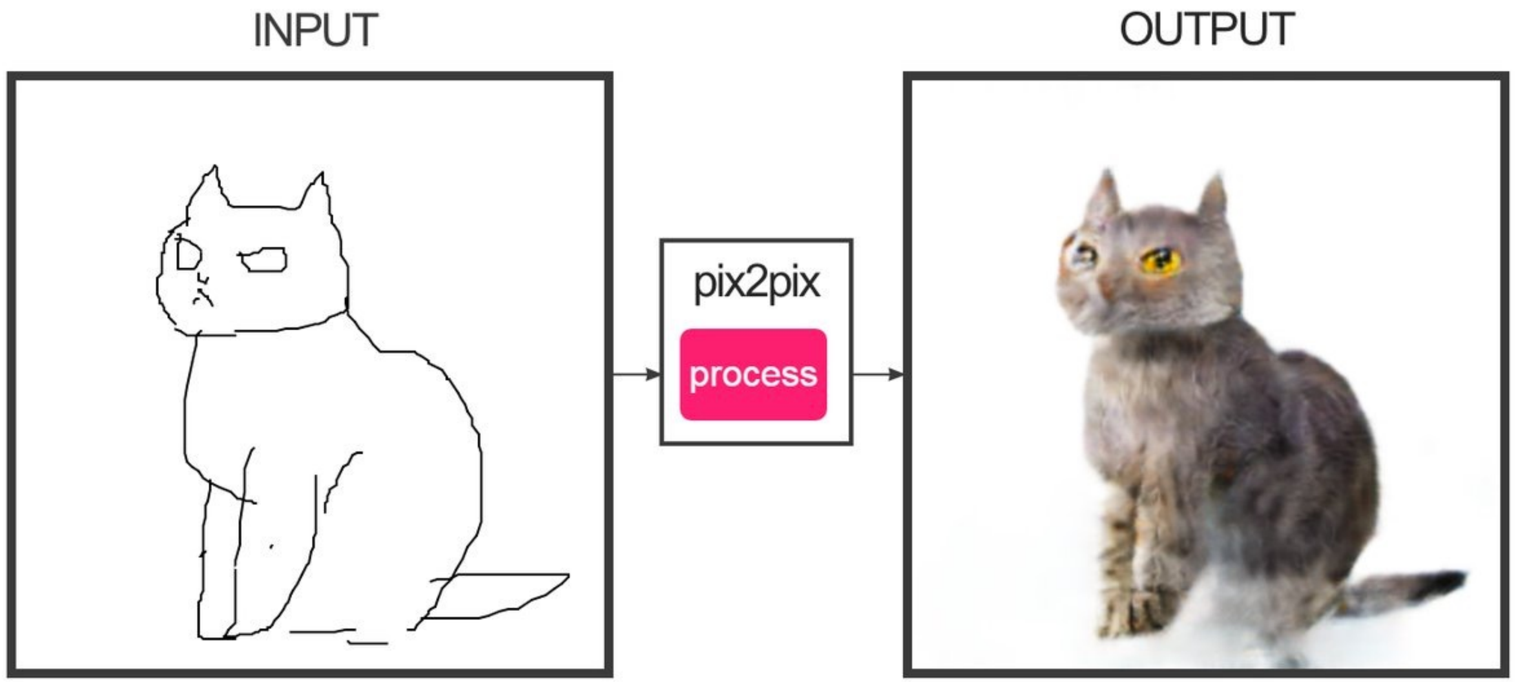
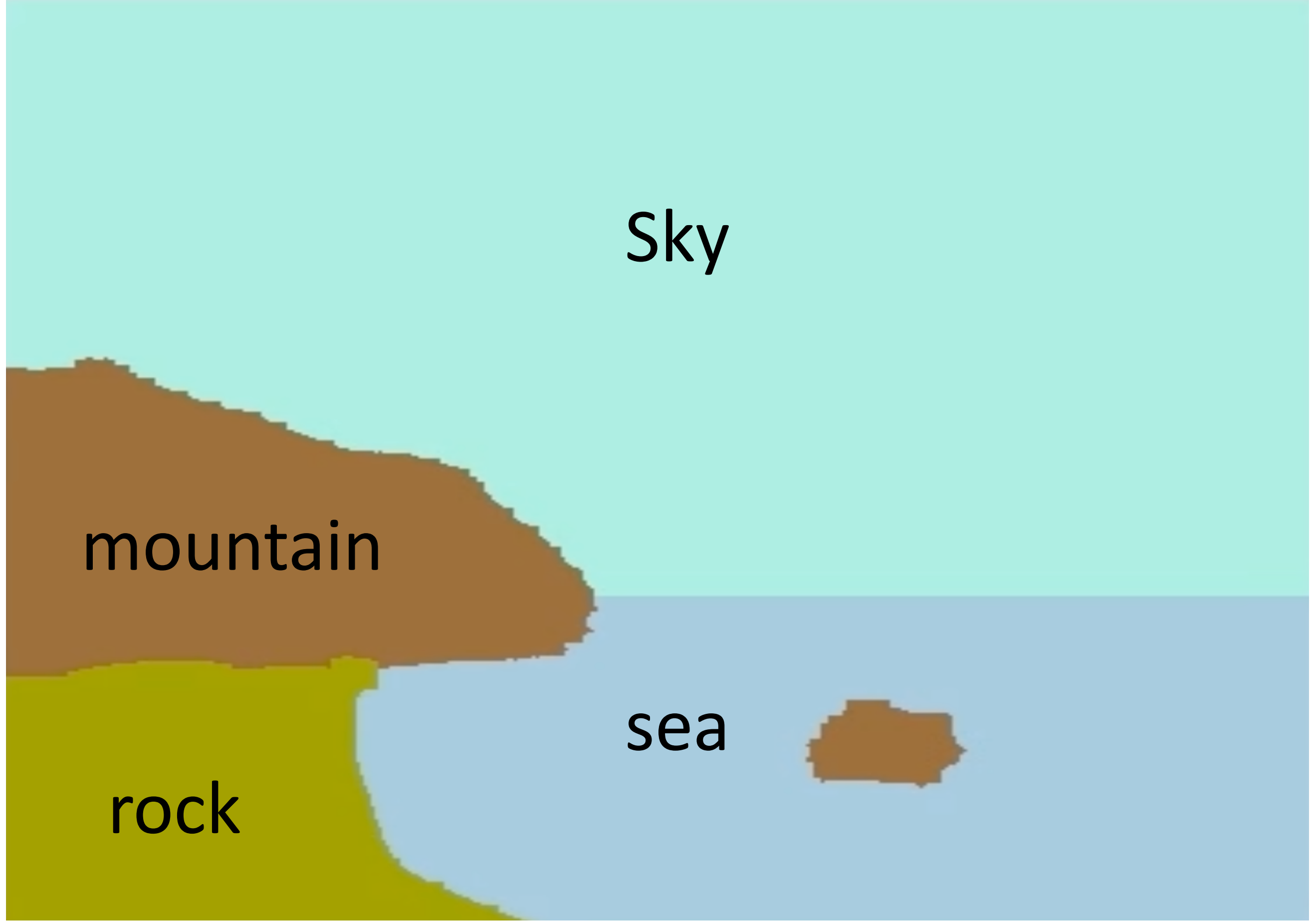


Image-to-Image Translation
Conditional Generative Models

Jun-Yan Zhu

16-726, Spring 2025

Problem Statement

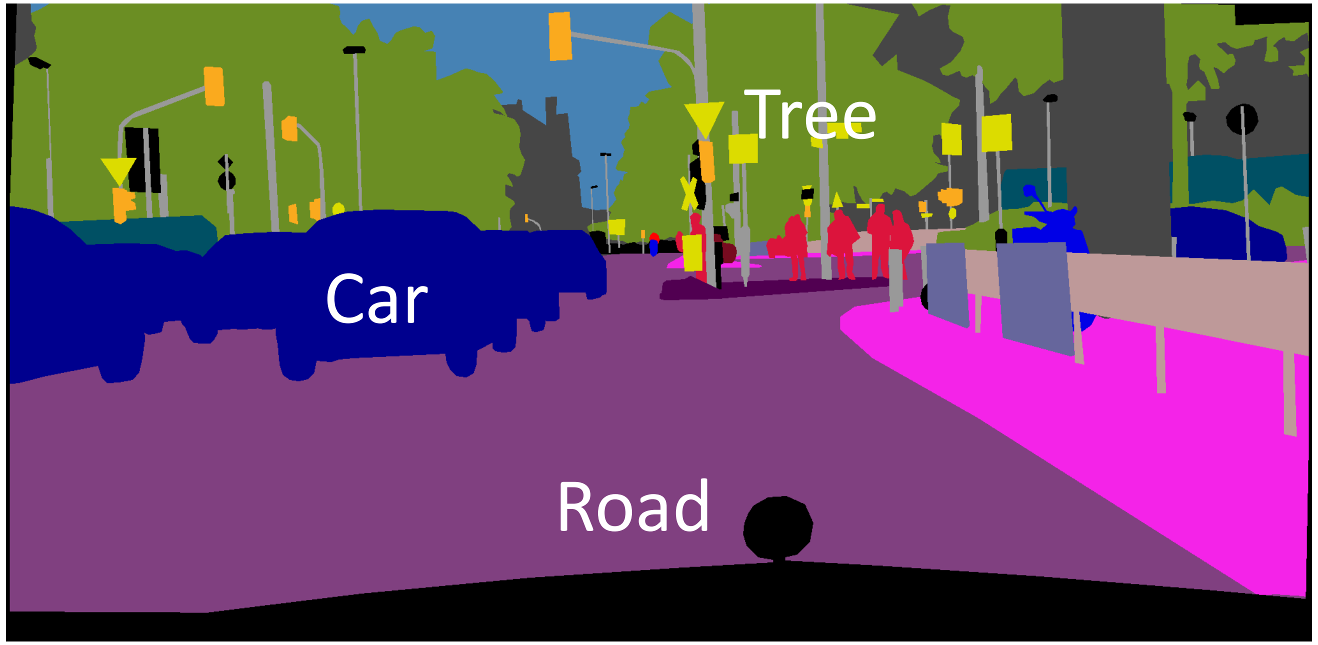


Input

Output

Goal: synthesize a photograph given an input image

Problem Statement



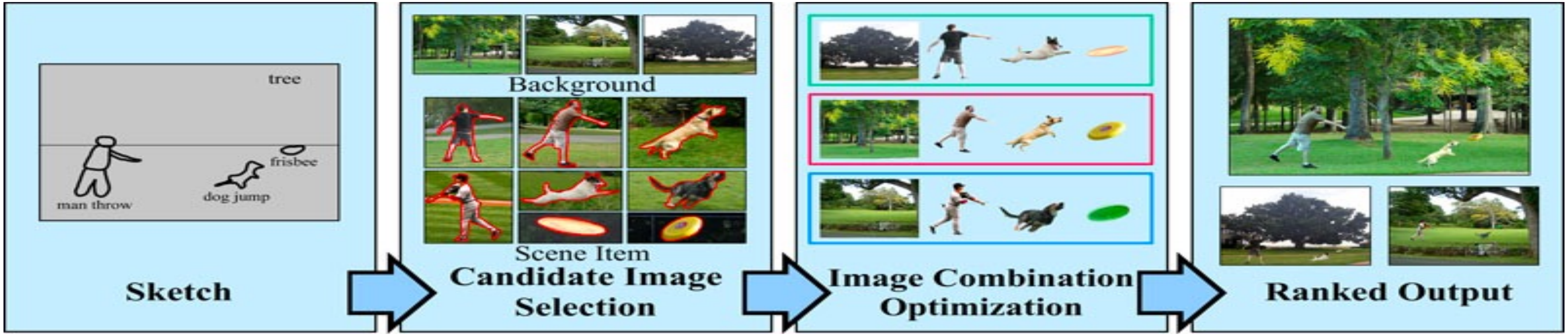
Input

Output

Goal: synthesize a photograph given an input image

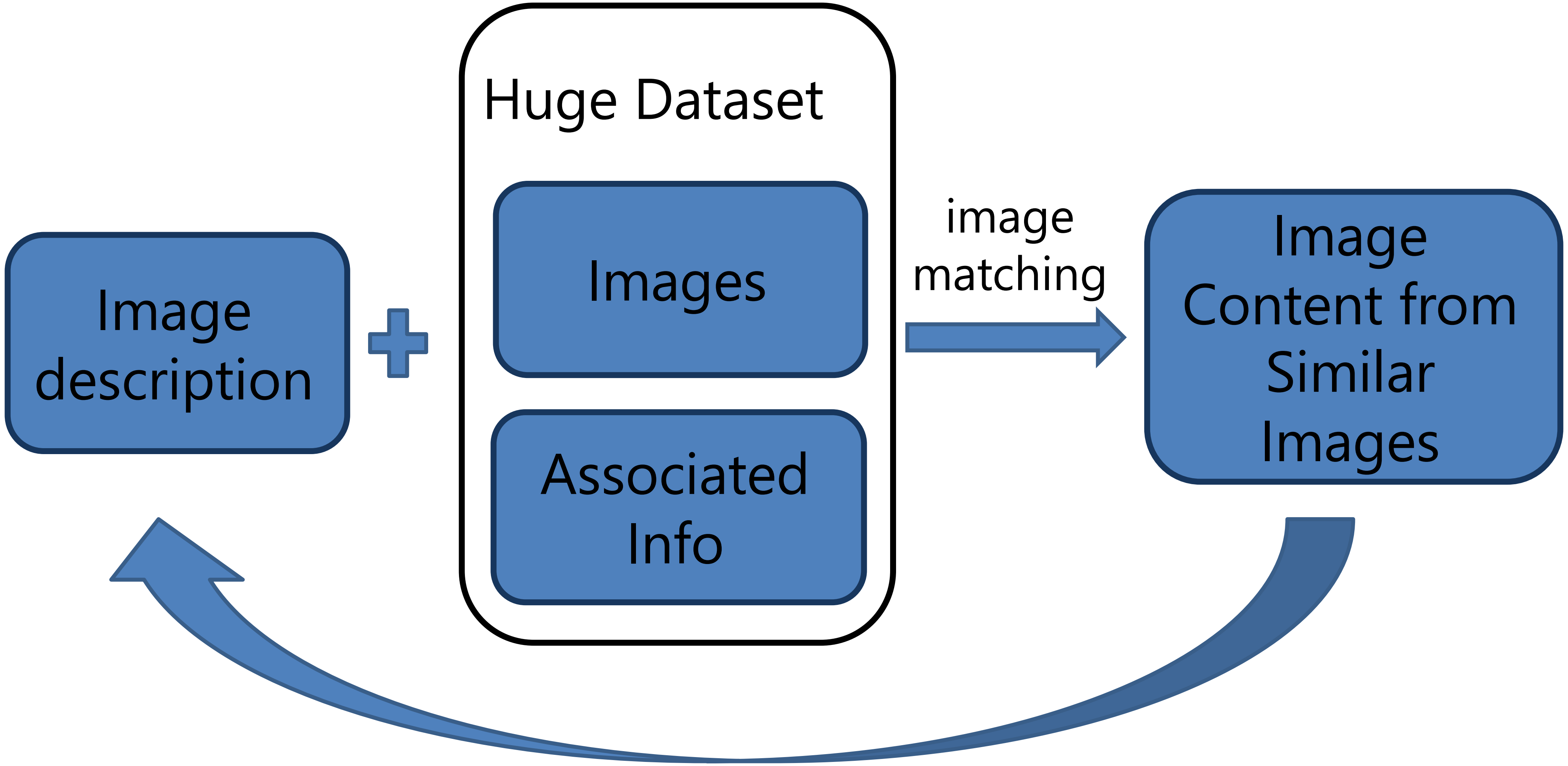
Early work (Example-based)

Semantic Photo Synthesis [Johnson et al., Eurographics 2006]



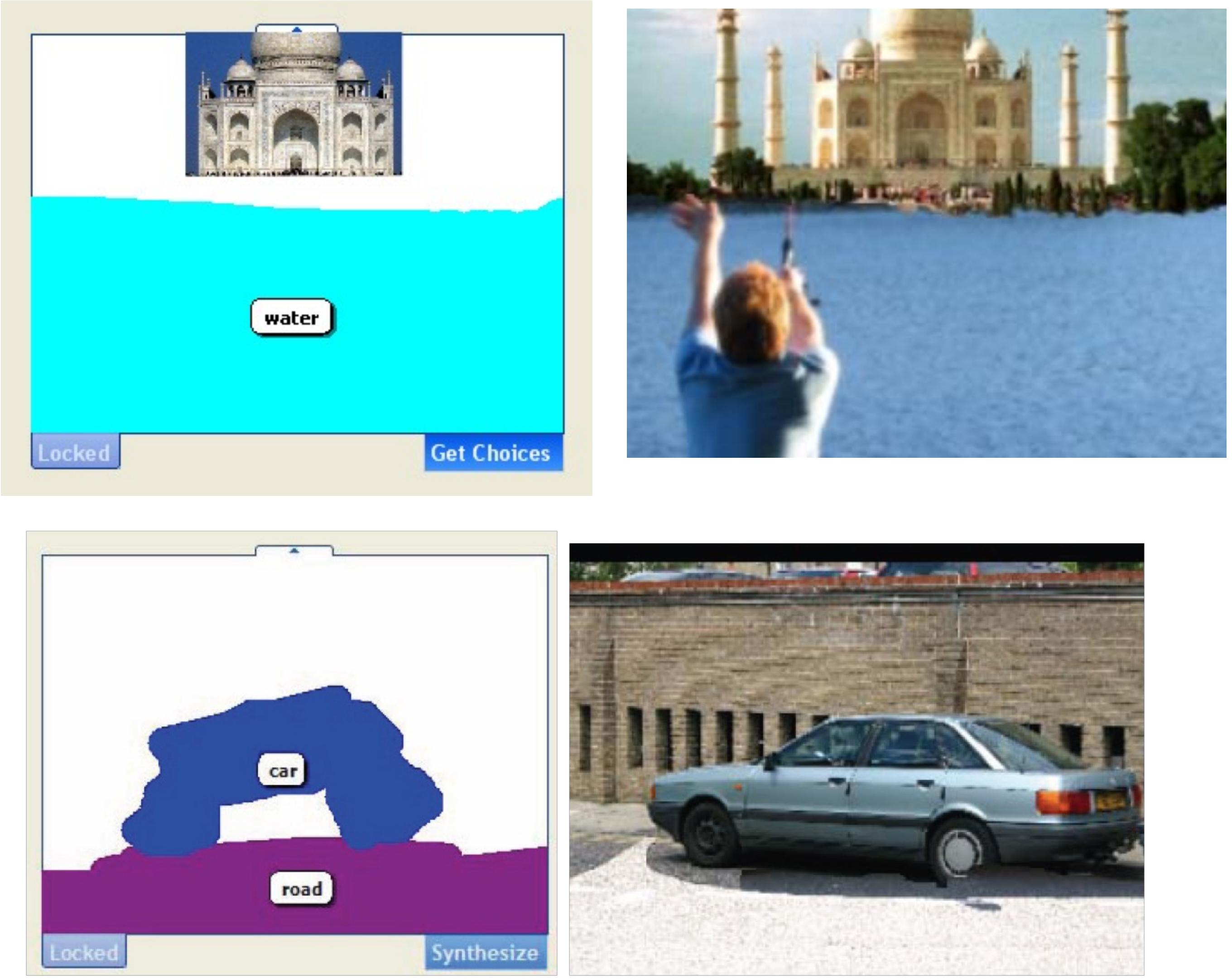
Sketch2Photo [Tao et al., SIGGRAPH Asia 2009]

Semantic Photo Synthesis



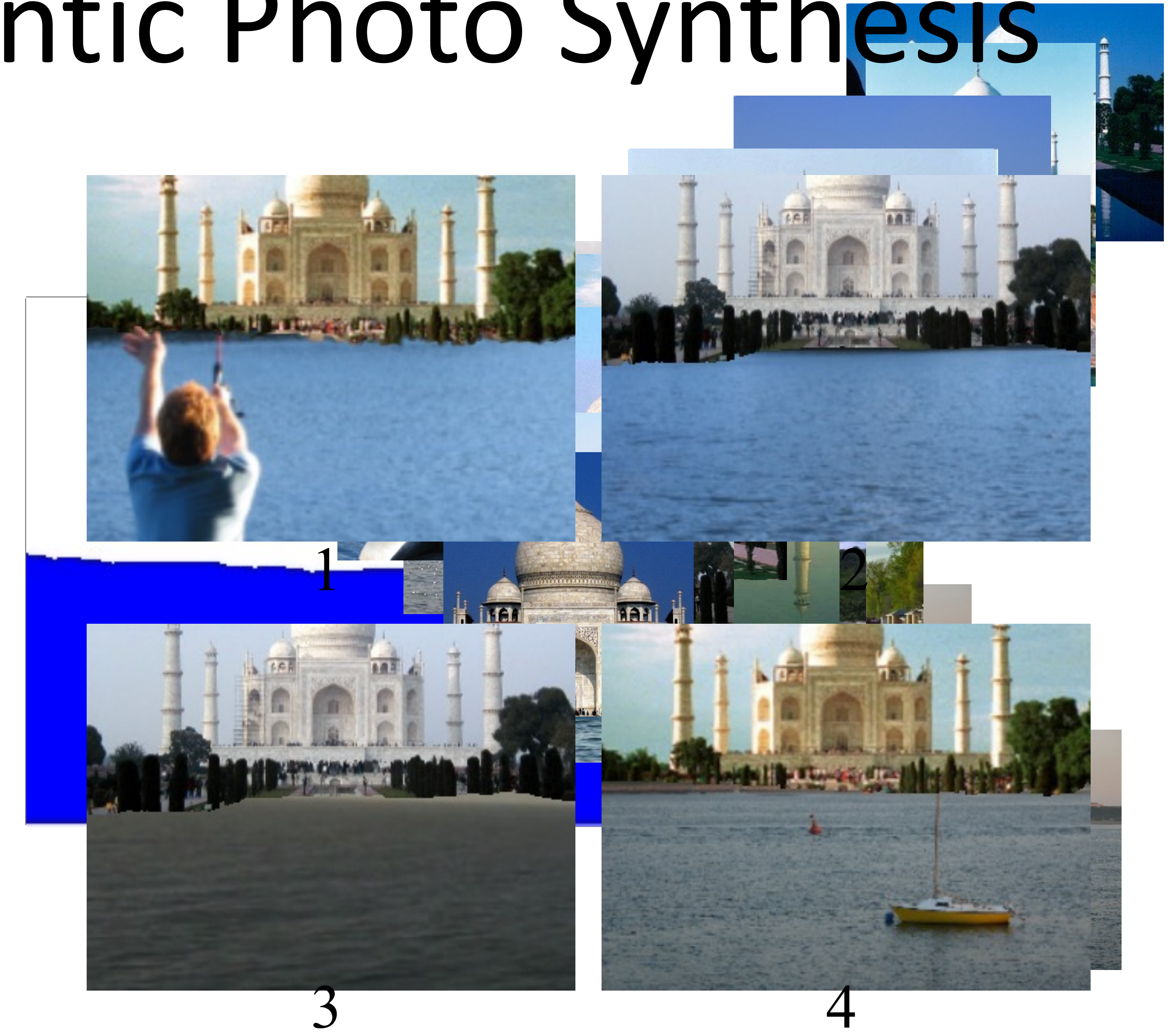
M. Johnson, G. Brostow, J. Shotton, O. A. c, and R. Cipolla, “Semantic Photo Synthesis,” Eurographics 2006

Semantic Photo Synthesis [EG'06]

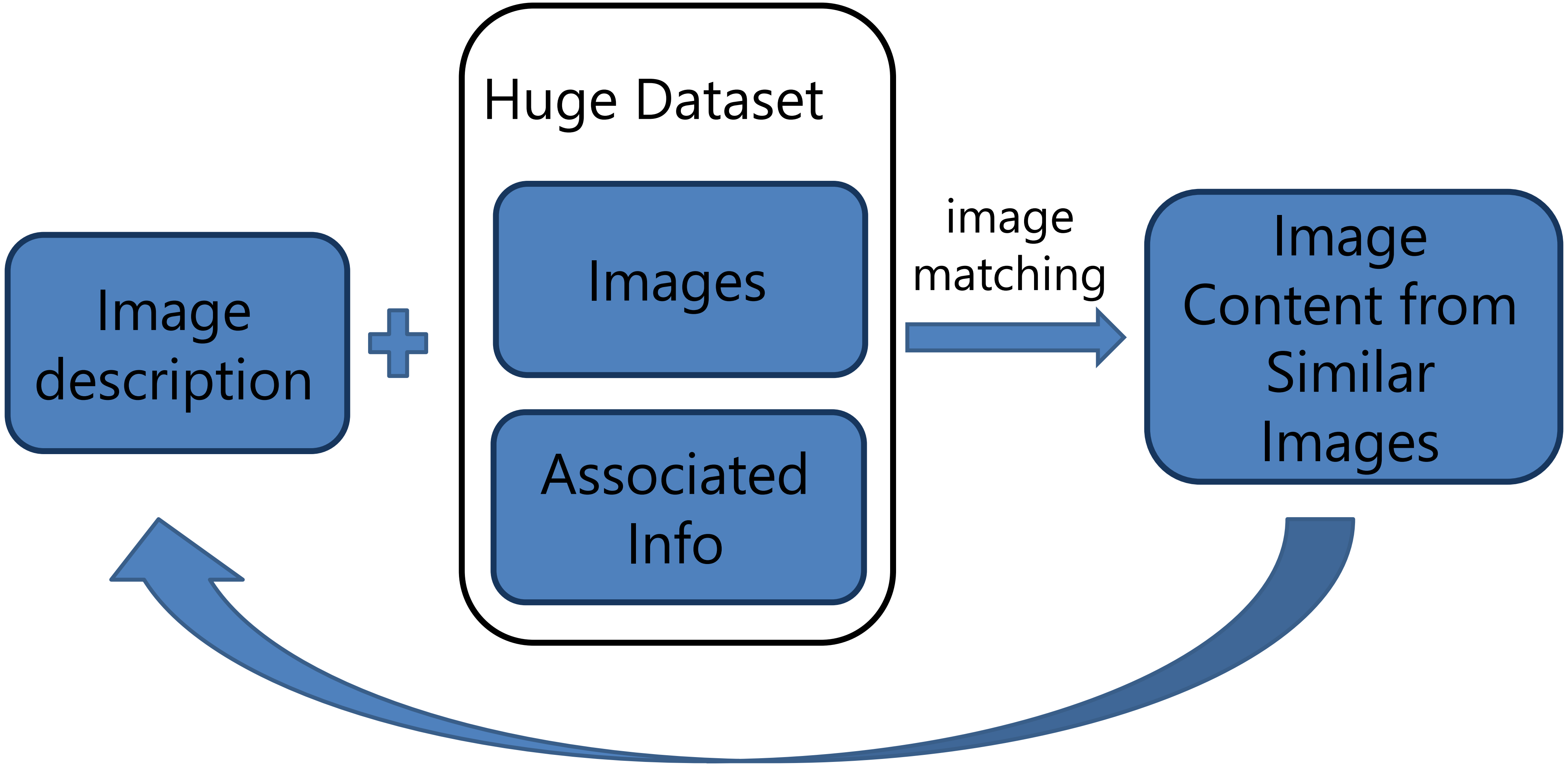


M. Johnson, G. Brostow, J. Shotton, O. A. c, and R. Cipolla, "Semantic Photo Synthesis," Eurographics 2006

Semantic Photo Synthesis



Semantic Photo Synthesis

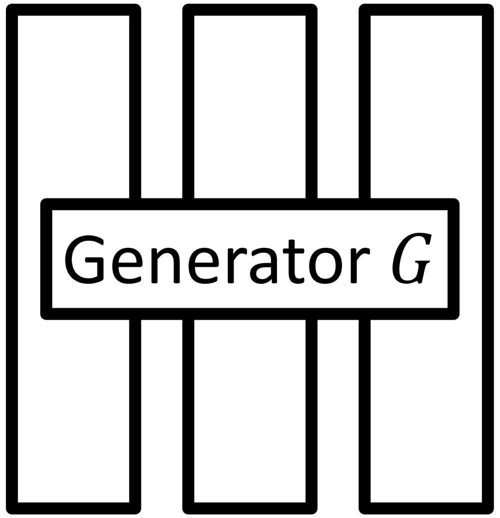


M. Johnson, G. Brostow, J. Shotton, O. A. c, and R. Cipolla, "Semantic Photo Synthesis," Computer Graphics Forum Journal (Eurographics 2006), vol. 25, no. 3, 2006.

Learning-based methods

Loss functions for Image Synthesis

Input x



Learnable rendering

Output Image $\mathrm{G}(x)$

What is a good objective $\mathcal{L}$?

- What is a good loss?
- How to calculate it efficiently?
- How to collect data (x, y)?

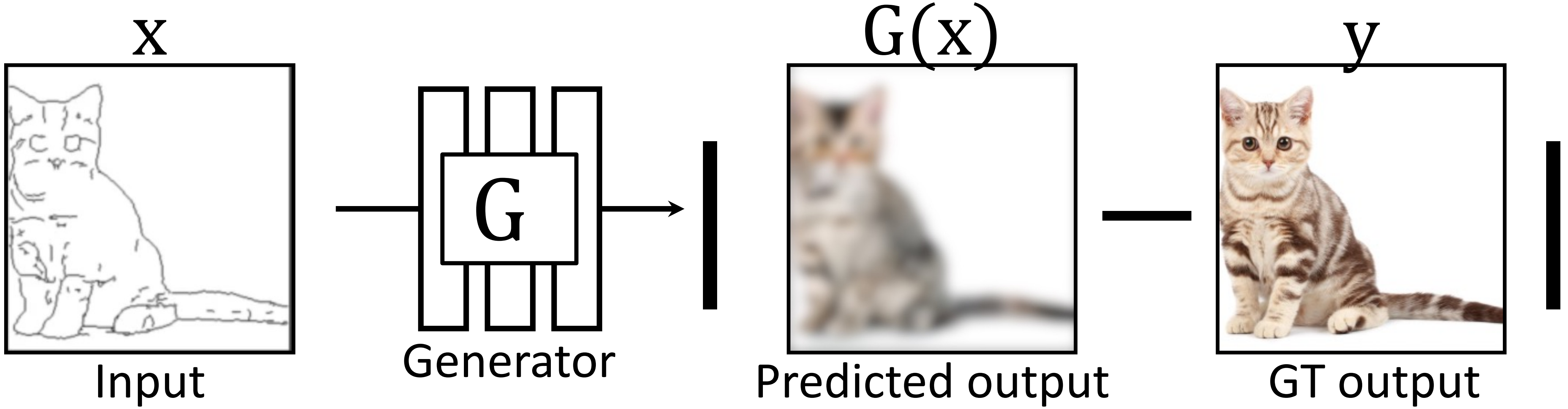
Problem Statement

Loss function

$$\arg\min_G \mathcal{L}(G(x), y)$$

Generator Input Info Output image

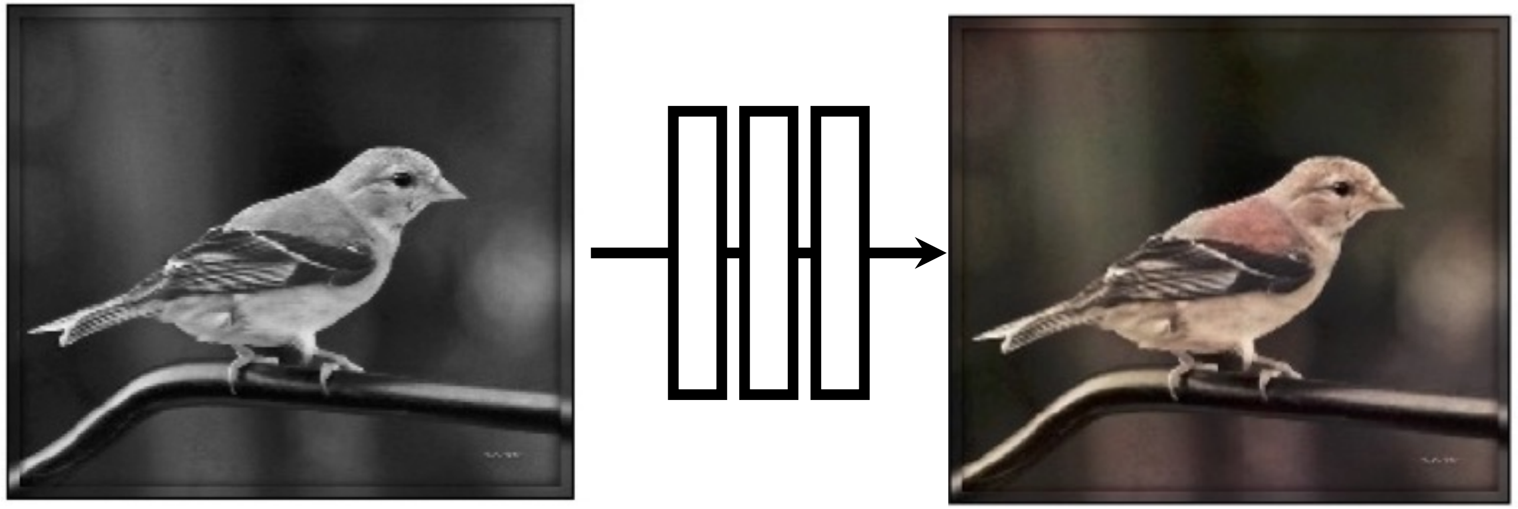
Designing Loss Functions



L2 regression $\arg\min_{G} \mathbb{E}[\,||G(x) - y||\,]$

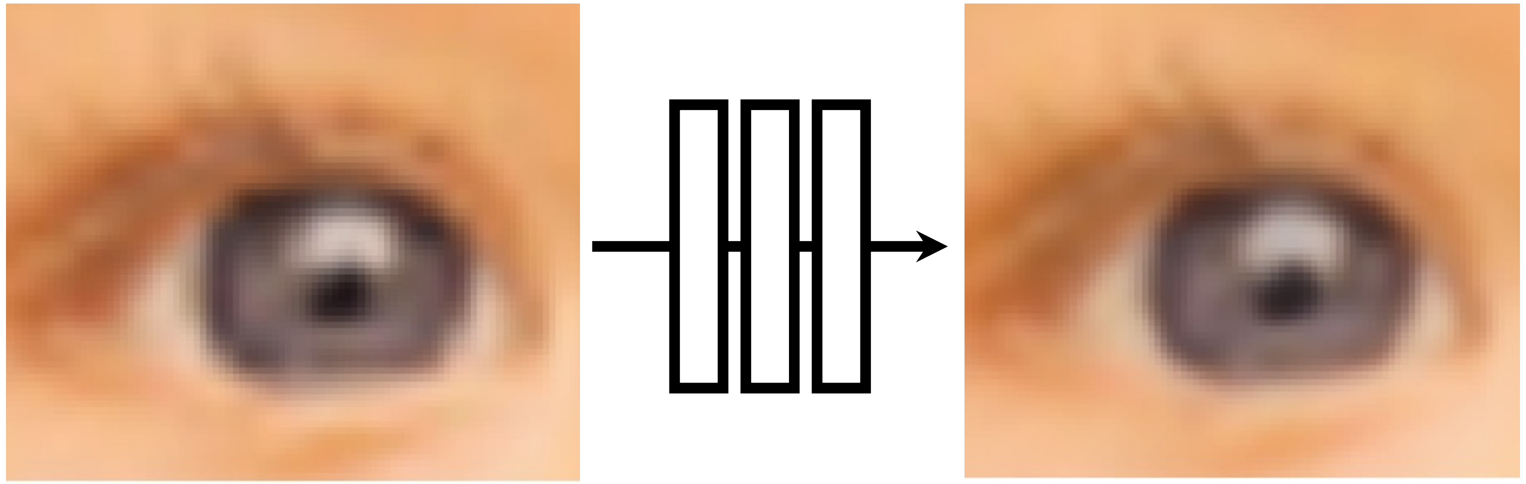
Designing Loss Functions

Image colorization



L2 regression

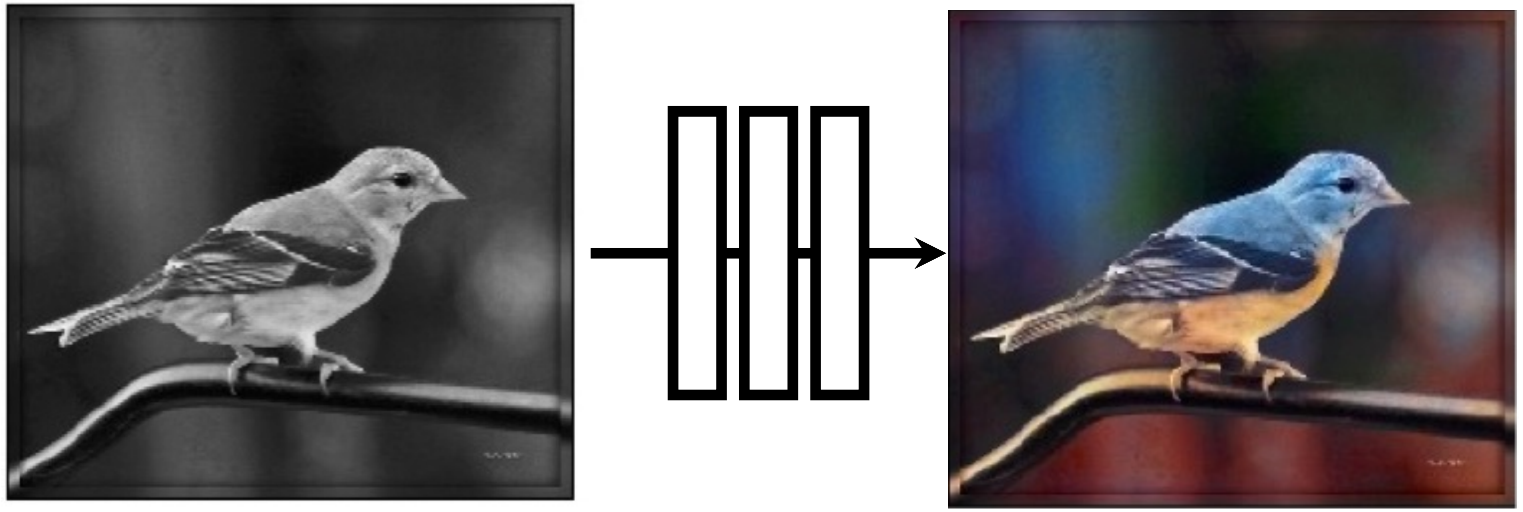
Super-resolution



L2 regression

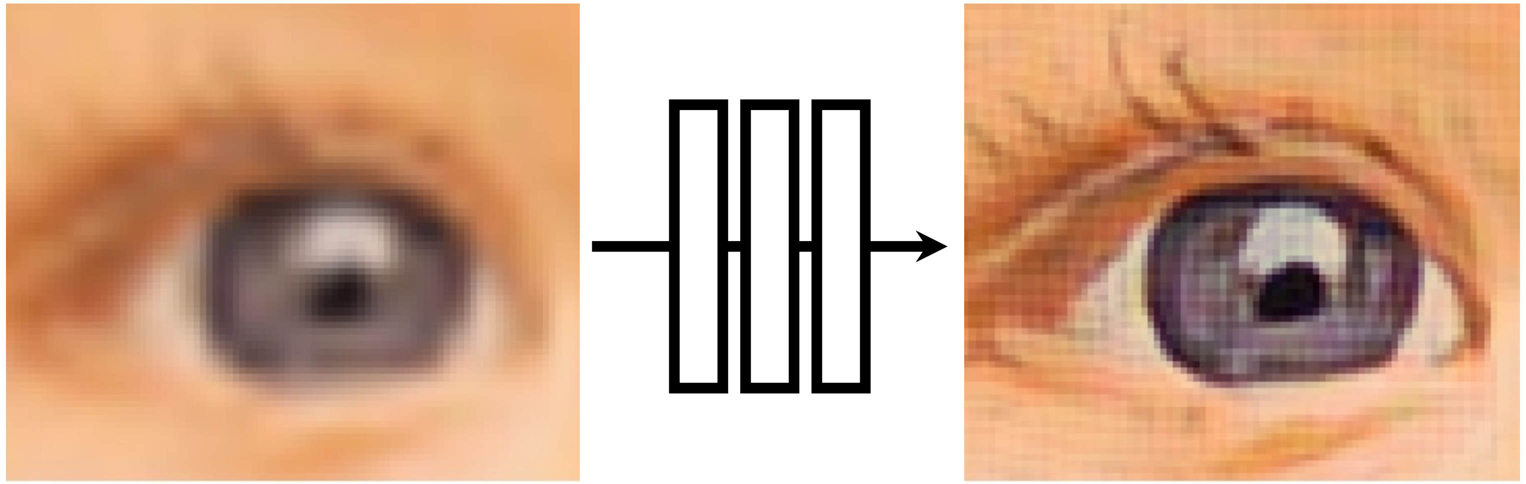
Designing Loss Functions

Image colorization



[Zhang et al. 2016]

Super-resolution



[Gatys et al., 2016], [Johnson et al. 2016], [Dosovitskiy and Brox. 2016]

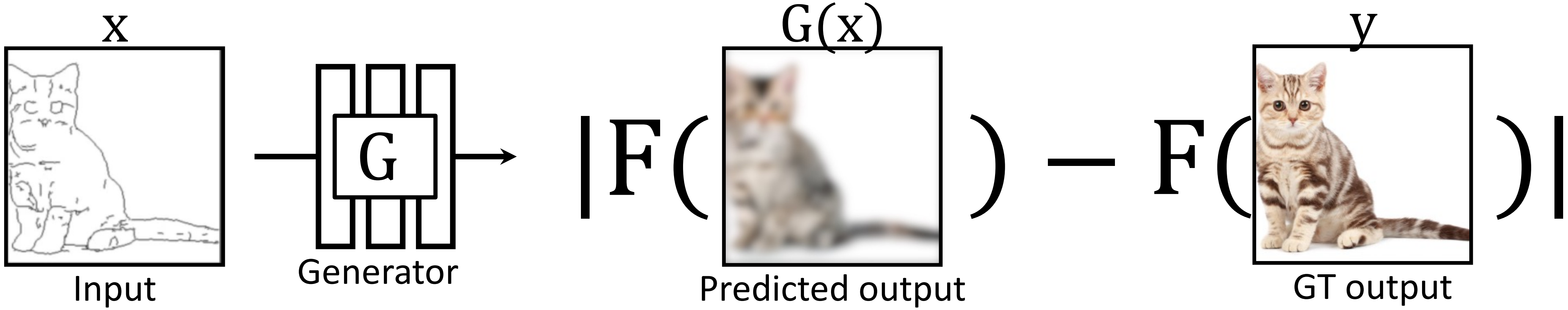
Classification Loss:

Cross entropy objective, with colorfulness term

Feature/Perceptual loss

Deep feature matching objective

CNNs as a Perceptual Metric



F is a deep network (e.g., ImageNet classifier)

Perceptual Loss

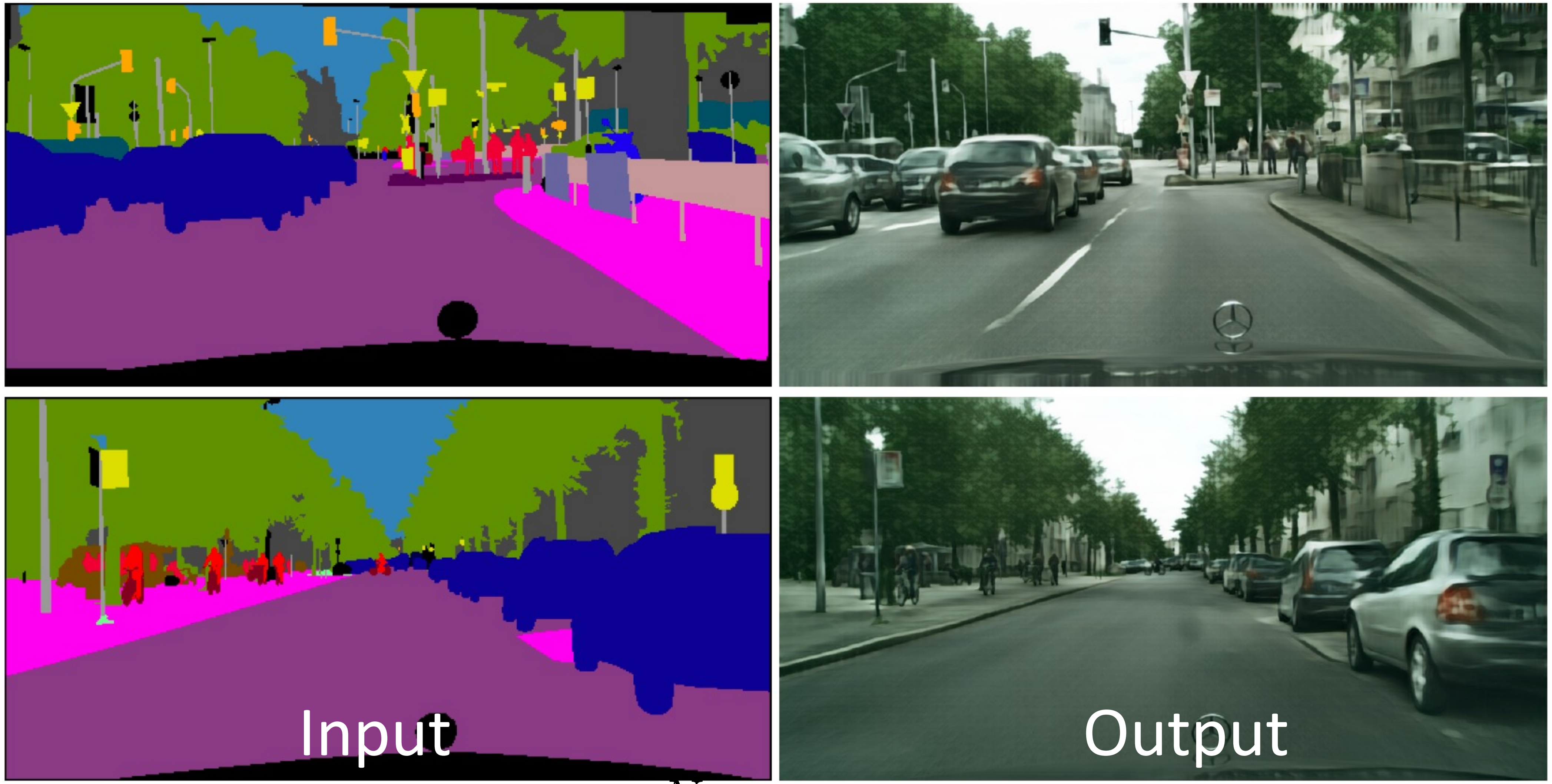
weight

(i)-th layer

$$\arg\min_{G} \mathbb{E}_{(x,y)} \sum_{i=1}^{N} \lambda_i \frac{1}{M_i} ||F^{(i)}(G(x)) - F^{(i)}(y))||_2^2$$

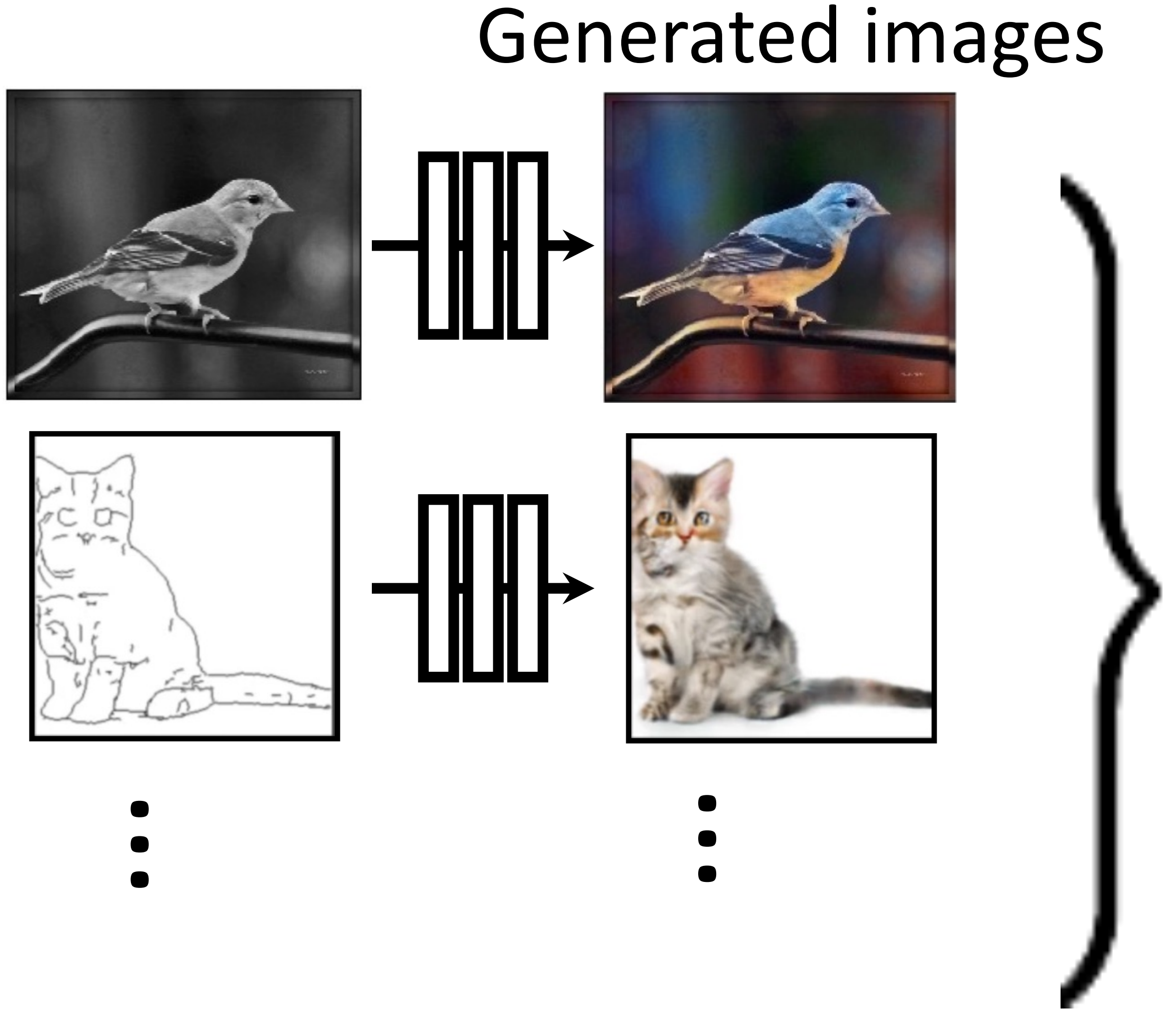
The number of elements in the (i)-th layer

Learning with Perceptual Loss



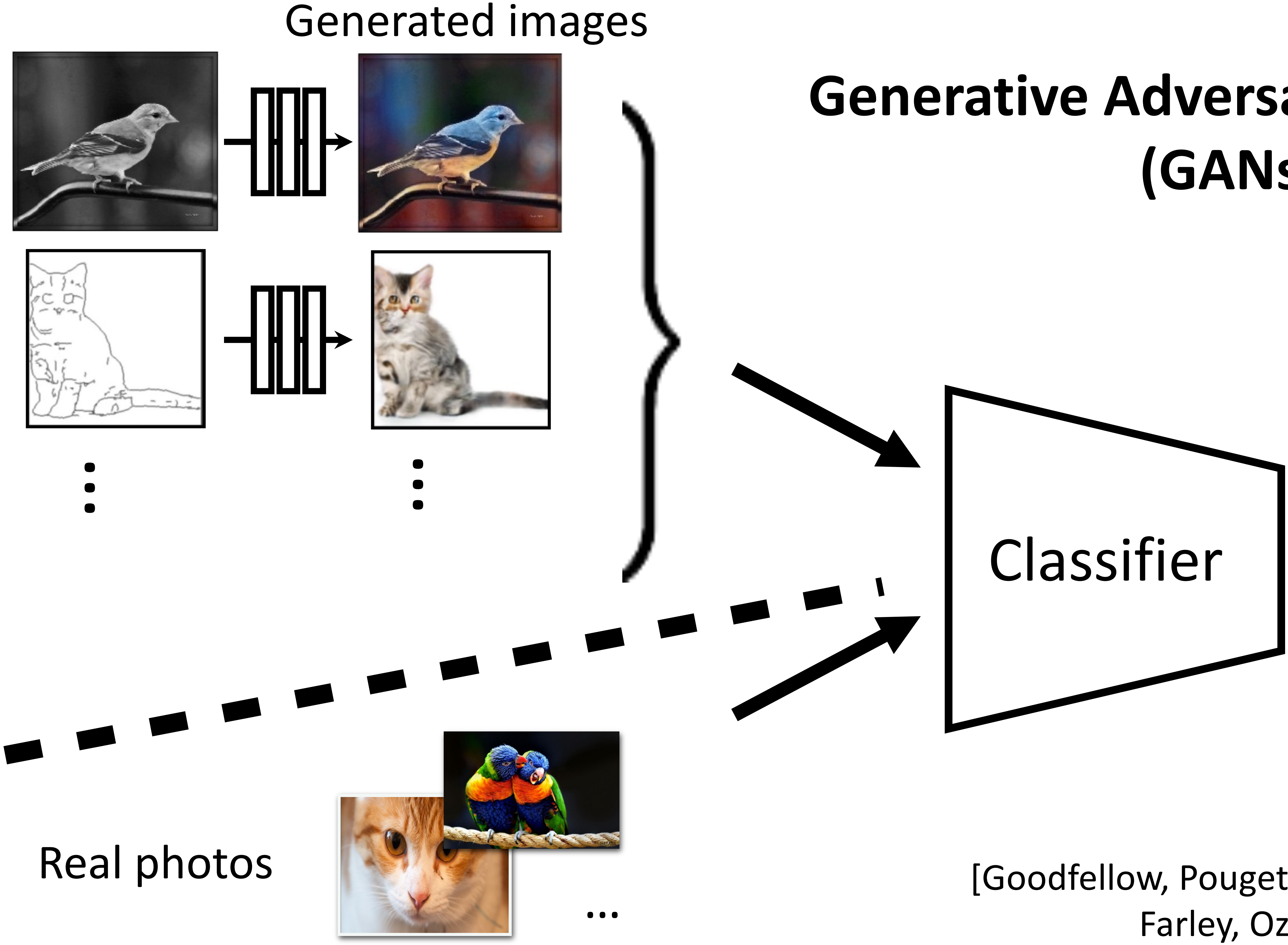
Training objective: $\arg\min_{G} \mathbb{E}_{(x,y)} \sum_{i=1}^{N} \lambda_i \frac{1}{M_i} ||F^{(i)}(G(x)) - F^{(i)}(y))||_2^2$

CRN [Chen and Koltun, 2017]

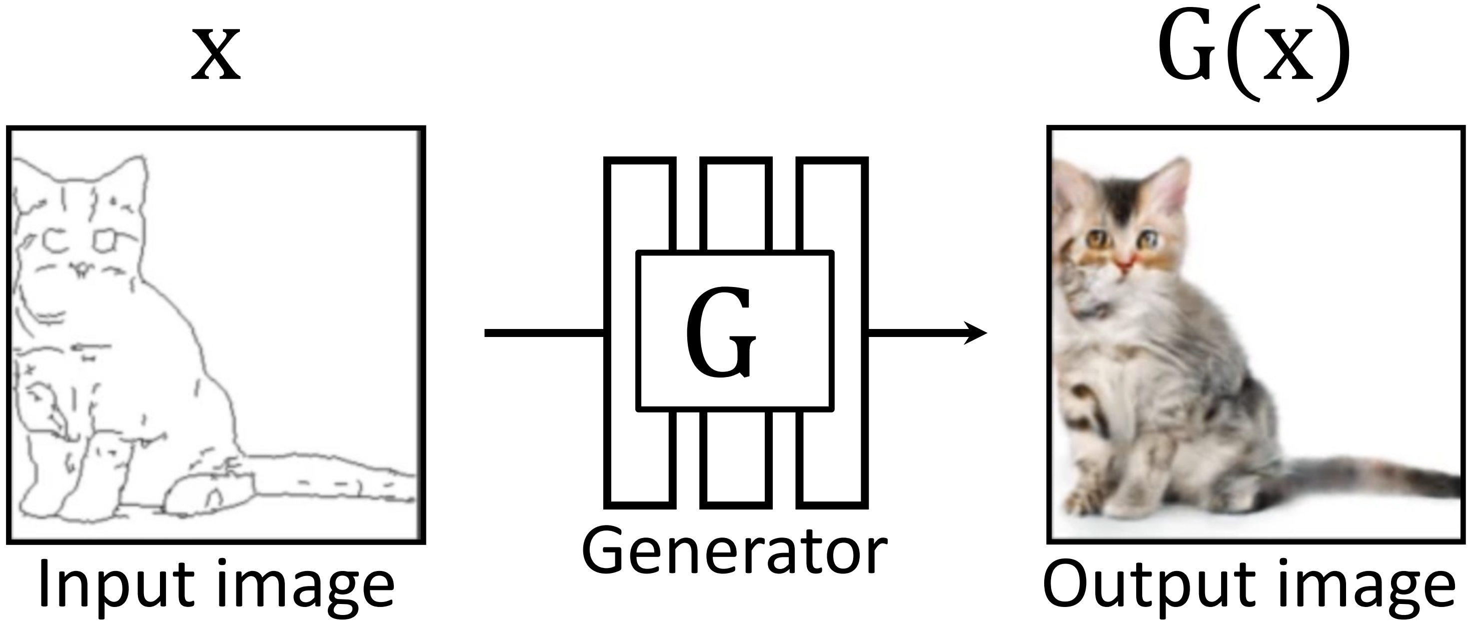


Universal loss?

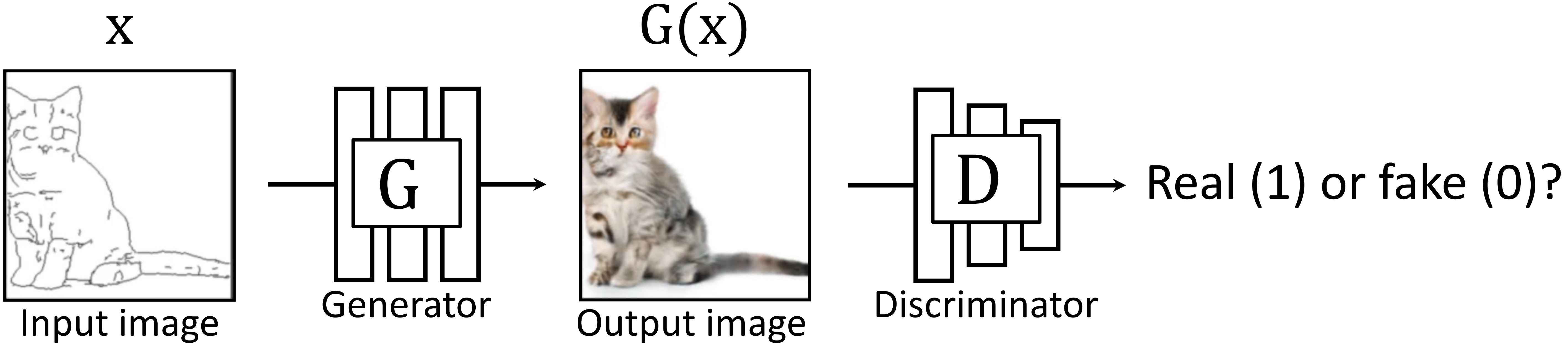
Generative Adversarial Network (GANs)



[Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, Bengio 2014]



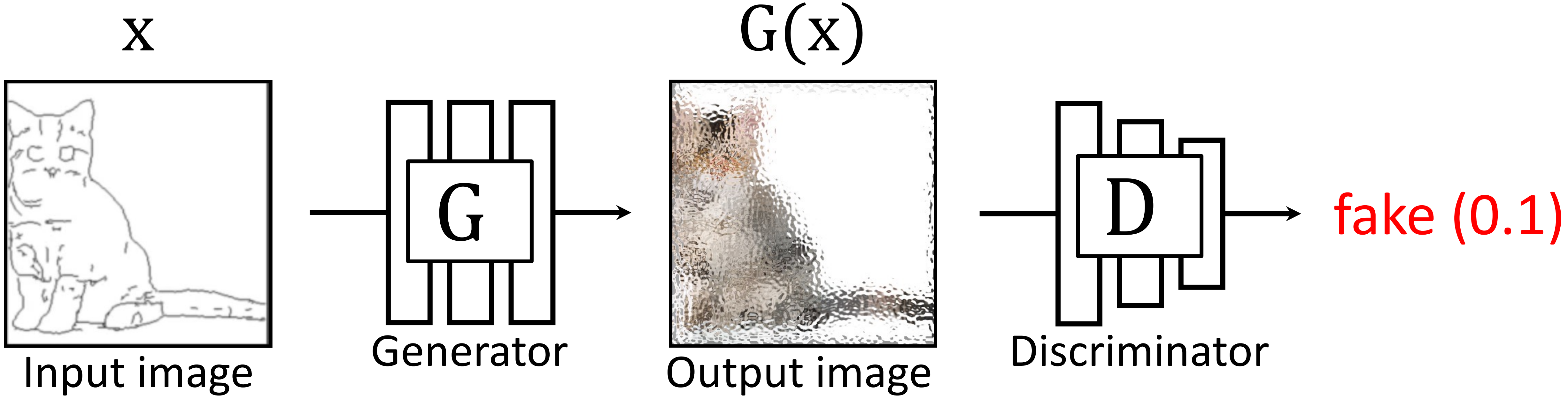
[Goodfellow et al. 2014]



A two-player game:

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

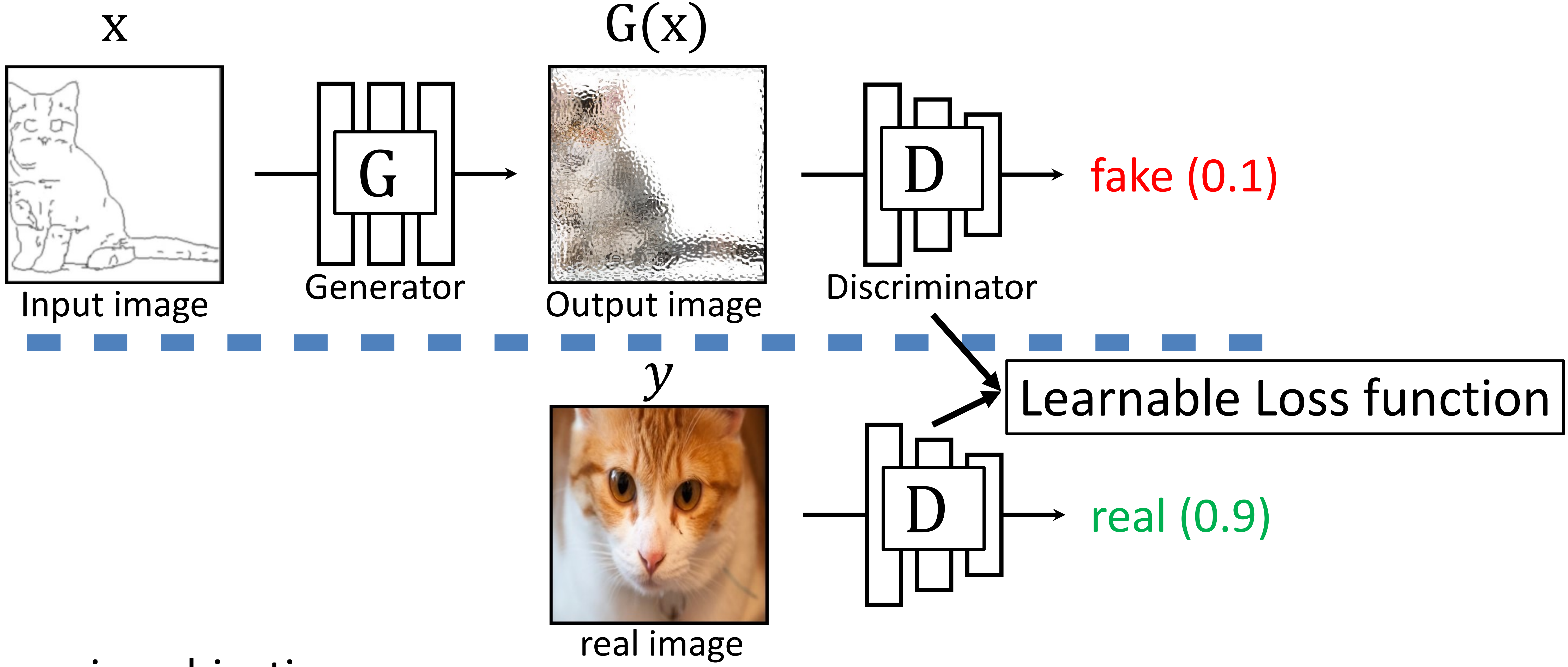
[Goodfellow et al. 2014]



Learning objective

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

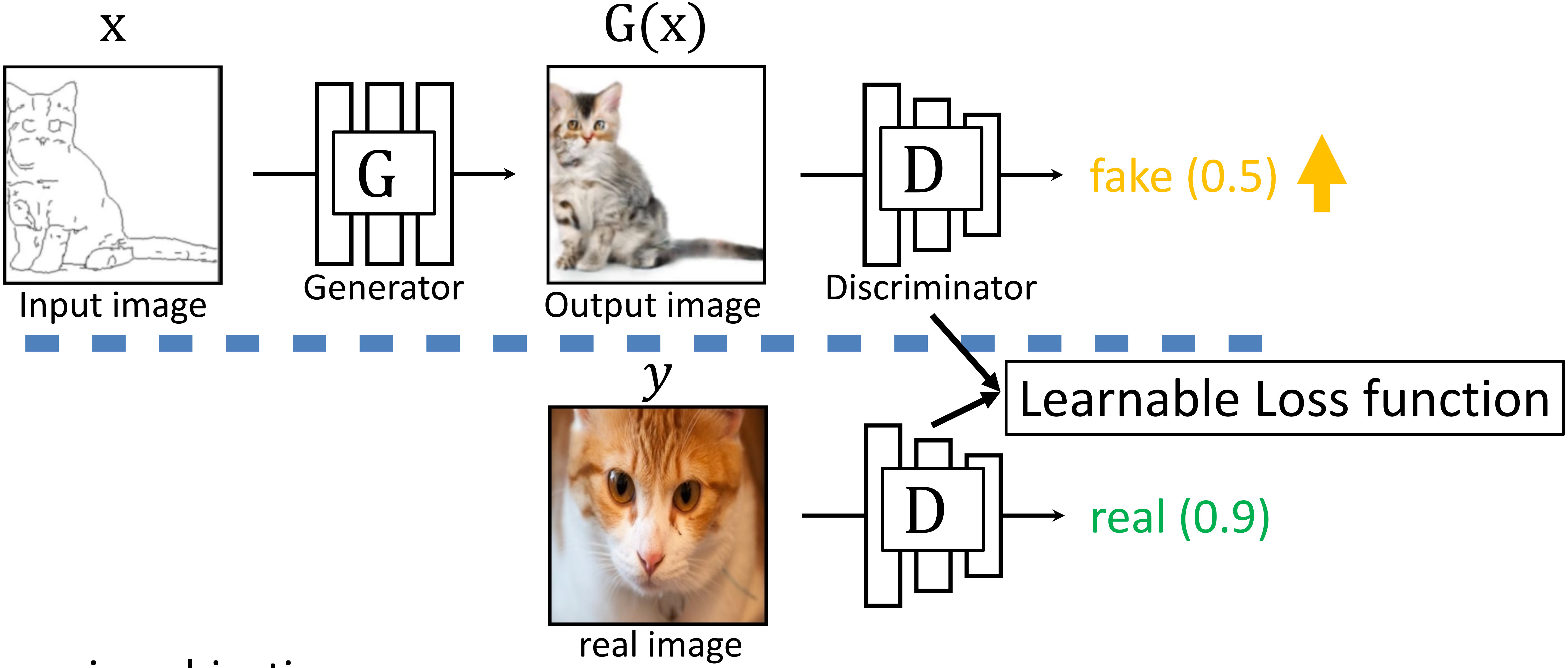
[Goodfellow et al. 2014]



Learning objective

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

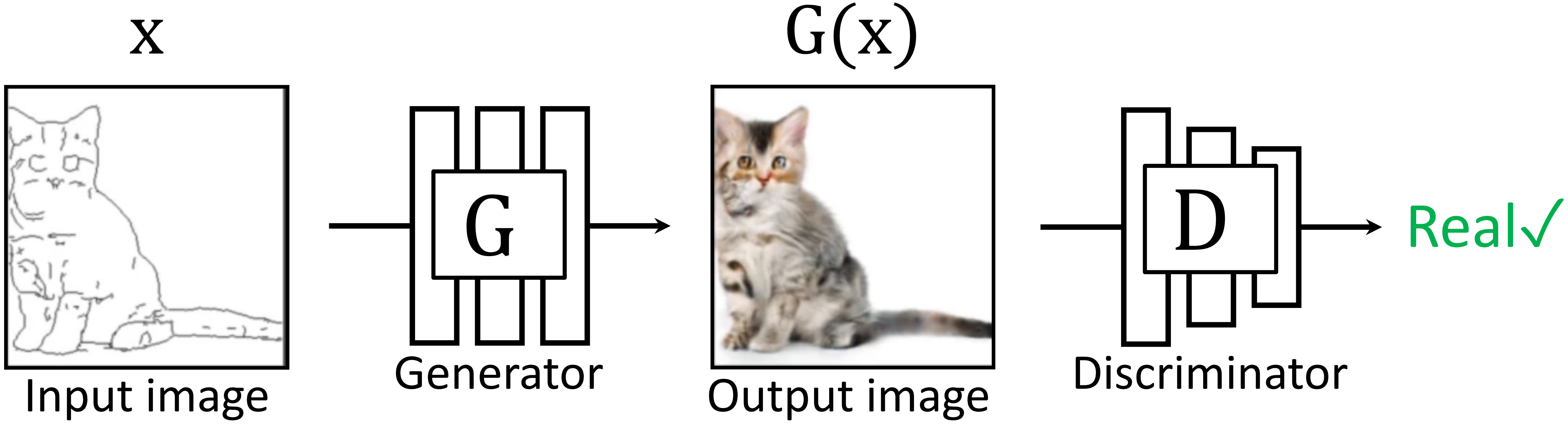
[Goodfellow et al. 2014]



Learning objective

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

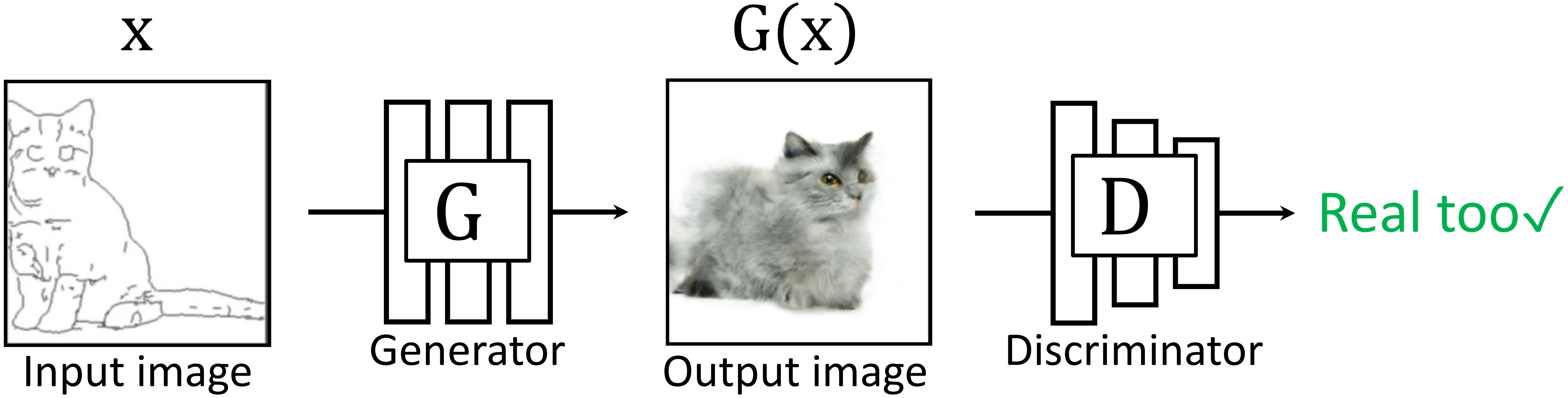
[Goodfellow et al. 2014]



Learning objective

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

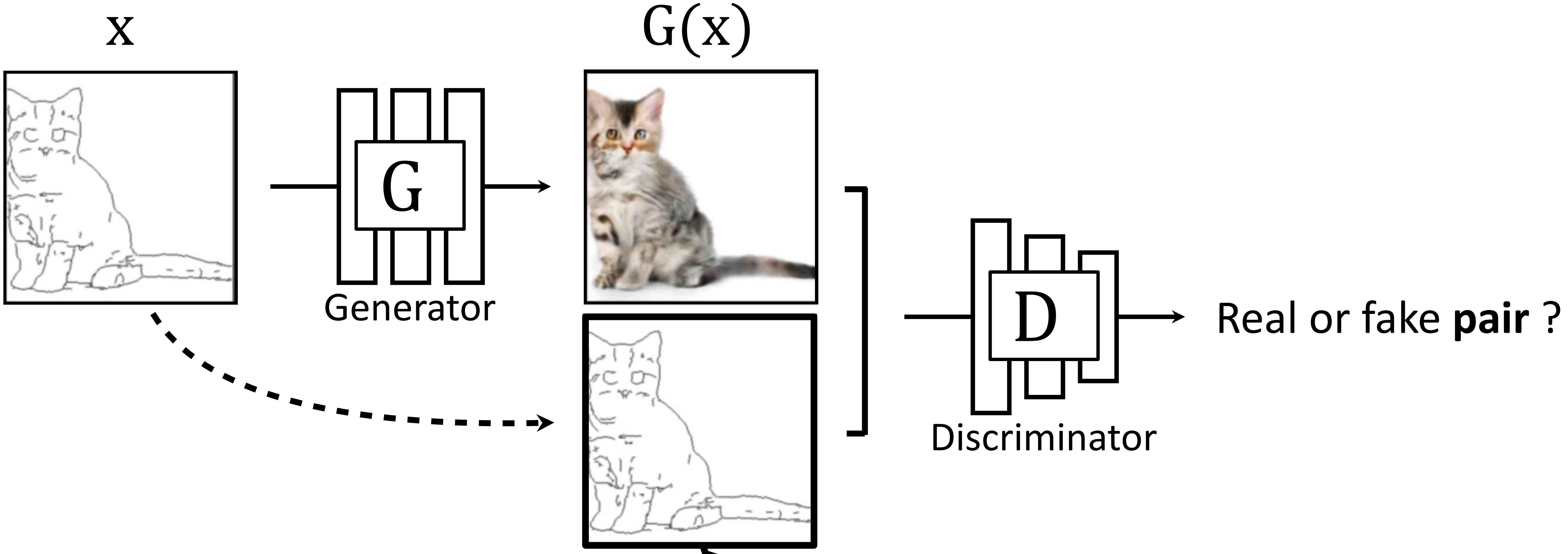
Pix2pix [Isola et al., 2016]



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]

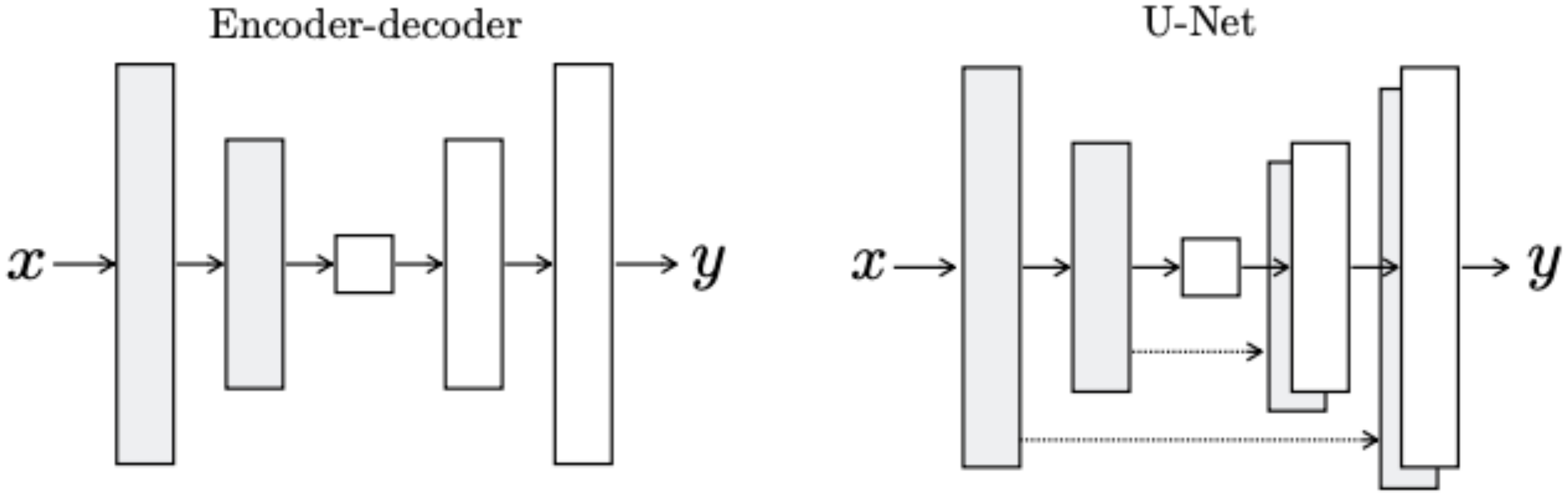


Learning objective

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

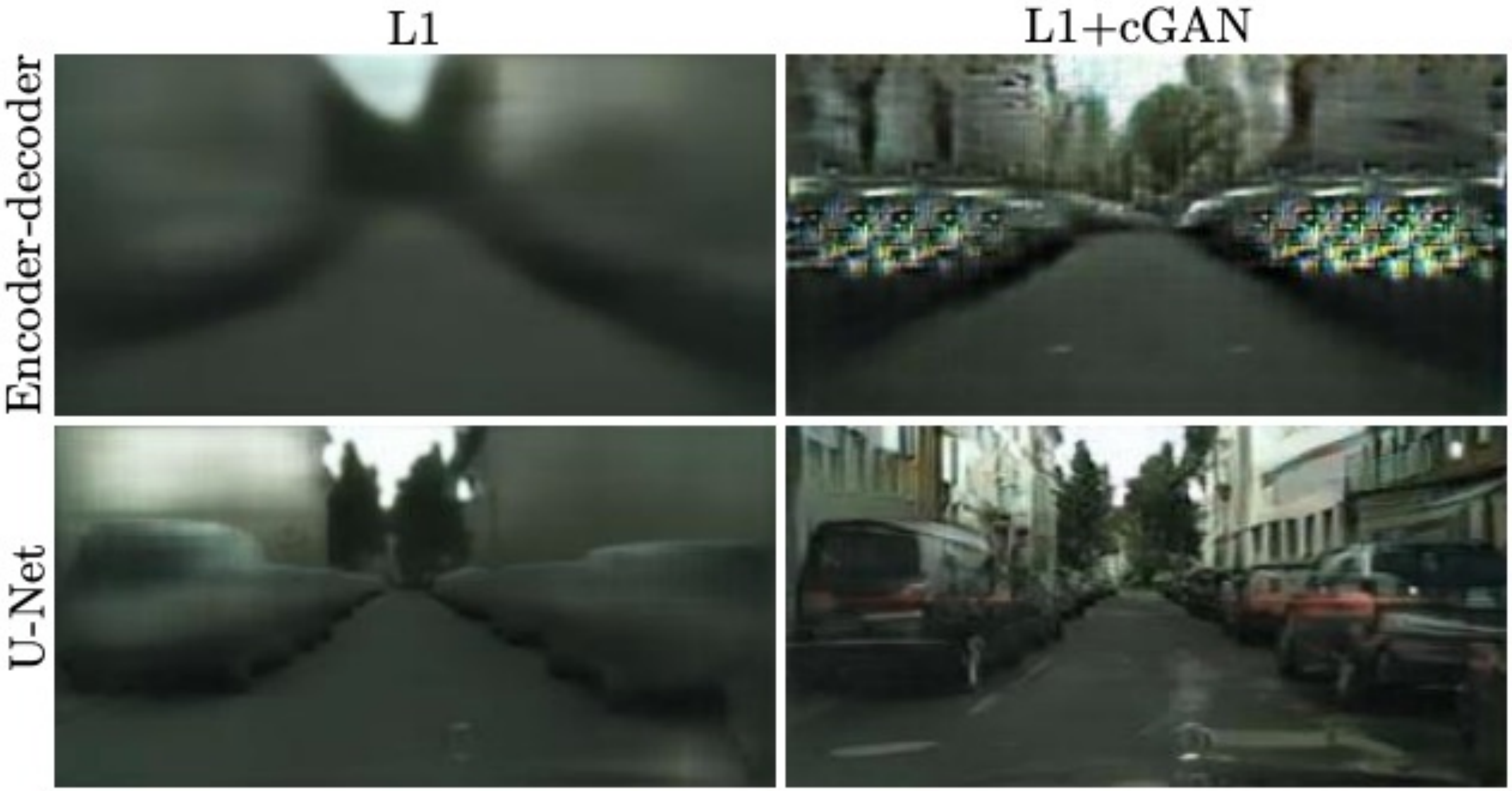
Pix2pix [Isola et al., 2016]

pix2pix Generator (U-Net)



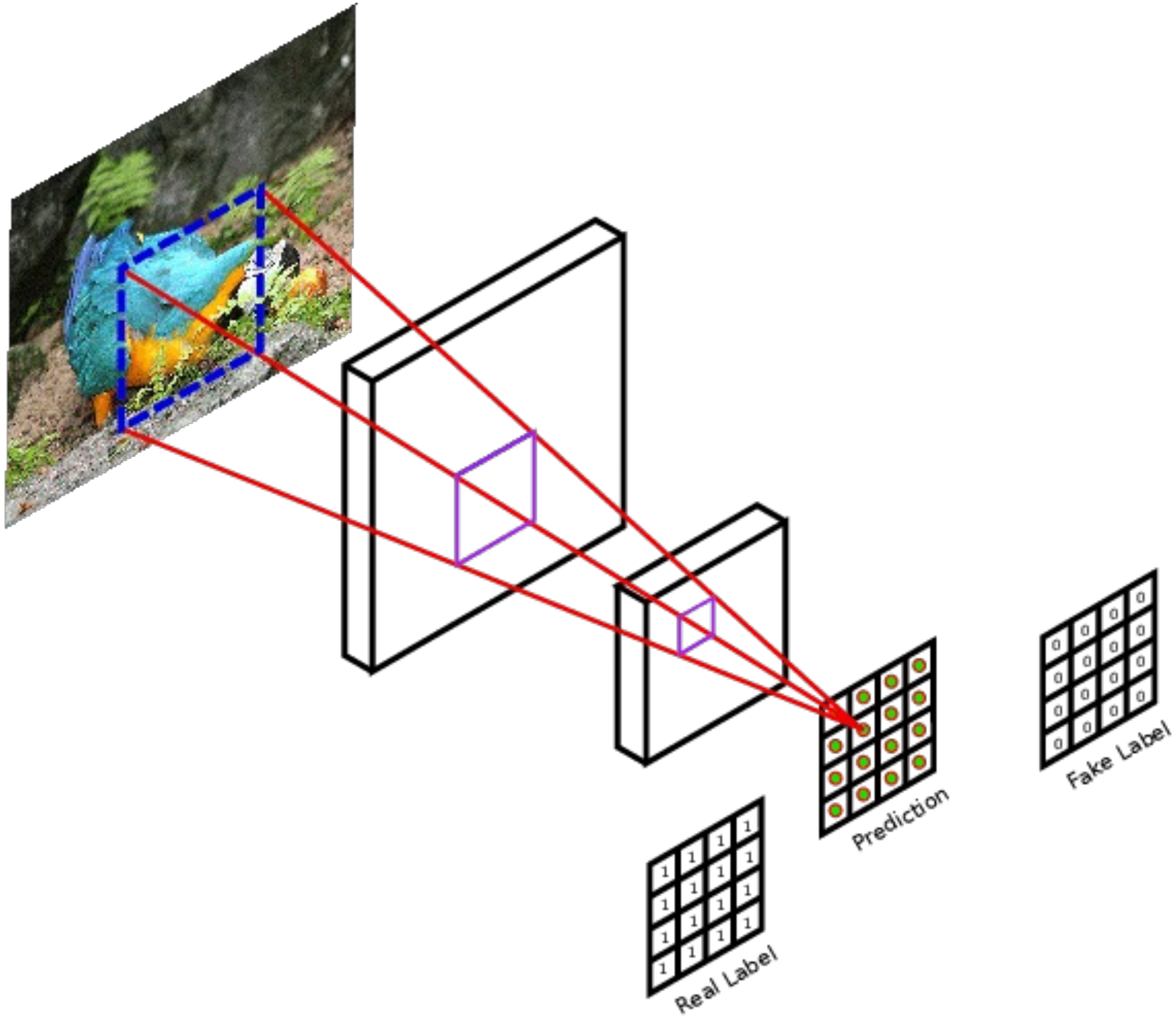
U-Net [Ronneberger et al.]: popular CNN backbone for biomedical image segmentation
U-Net: preserve high-frequency information (e.g., edge) of the input image.
Encoder-decoder: lose high-frequency details due to the information bottleneck

pix2pix Generator (U-Net)



Generator design is critical for image quality.
cGAN (conditional GANs) loss: capture realism. L1 loss stabilizes training (faster convergence)

pix2pix Discriminator (PatchGAN)

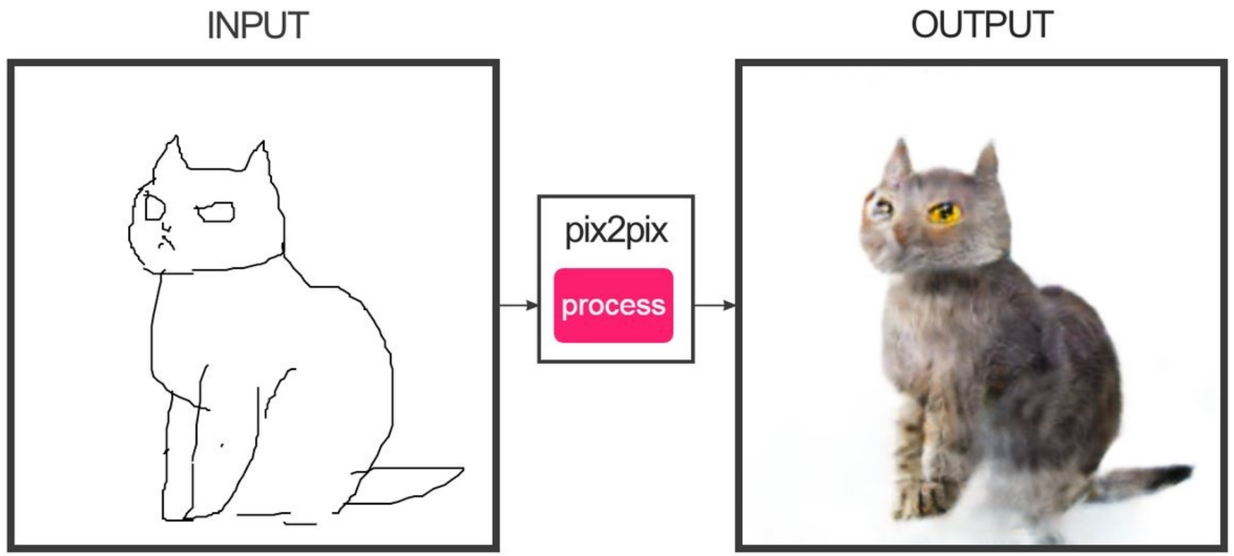


- Rather than penalizing if output *image* looks fake, penalize if each overlapping *patches* looks fake
- Focus on <u>local</u> visual cues (color, textures).
- <u>Global</u> structure: the input image has already encoded global structure. L1 loss can help as well.

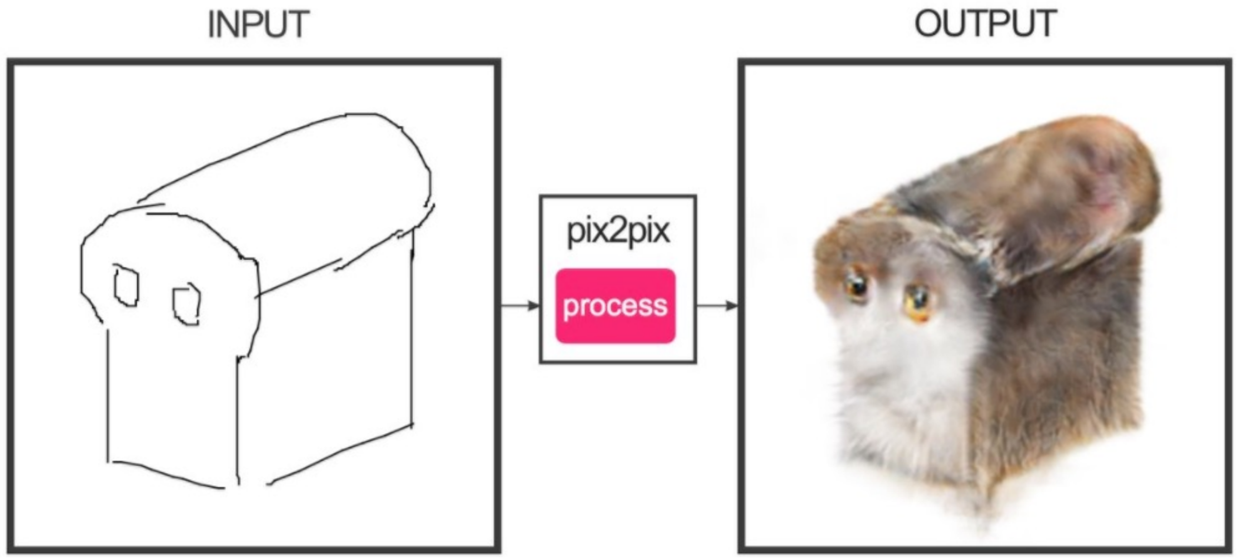
Advantages

- Faster, fewer parameters
- More supervised observations
- Applies to arbitrarily large images

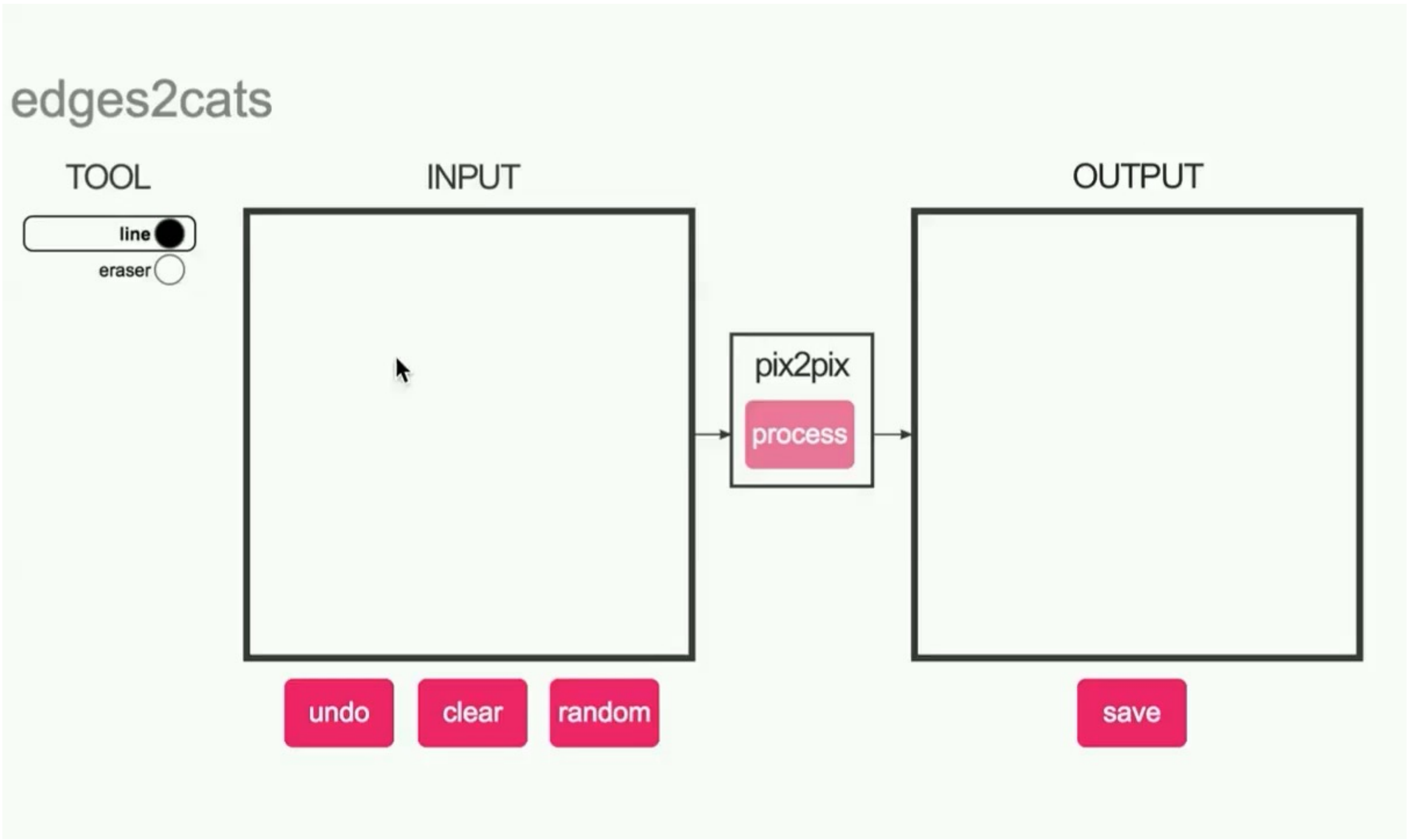
#edges2cats [Christopher Hesse]



@gods_tail



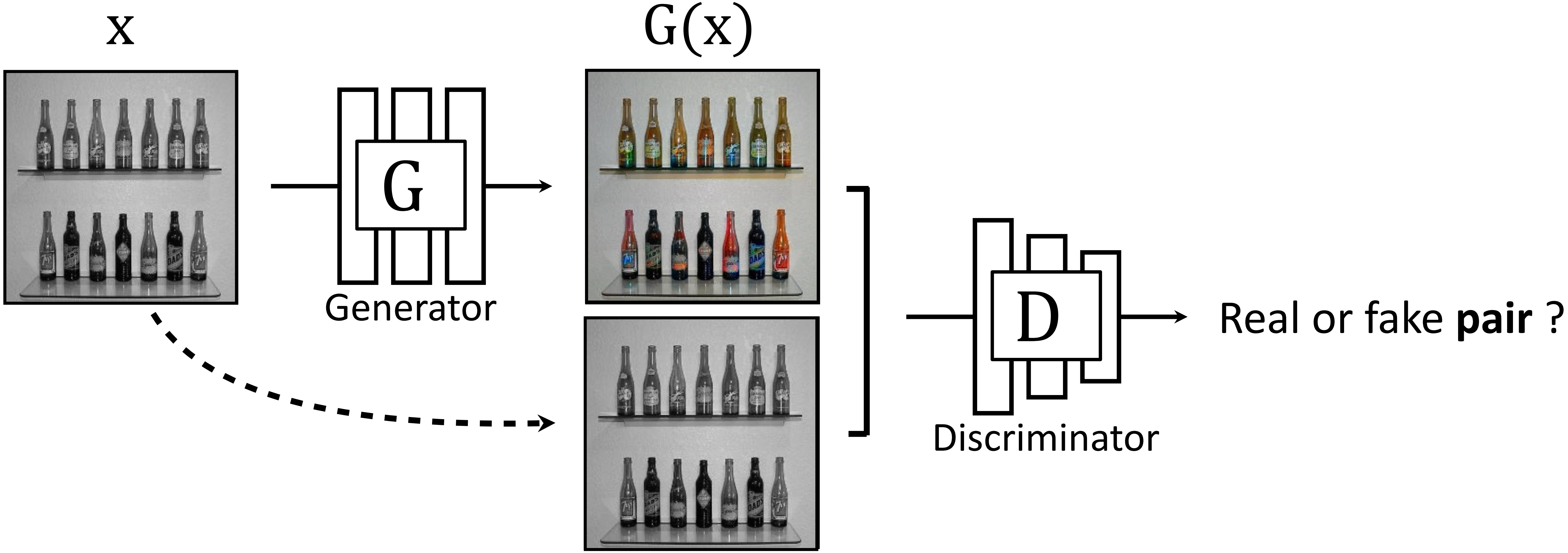
Ivy Tasi @ivymyt



@matthematician

Vitaly Vidmirov @vvid

https://affinelayer.com/pixsrv/



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

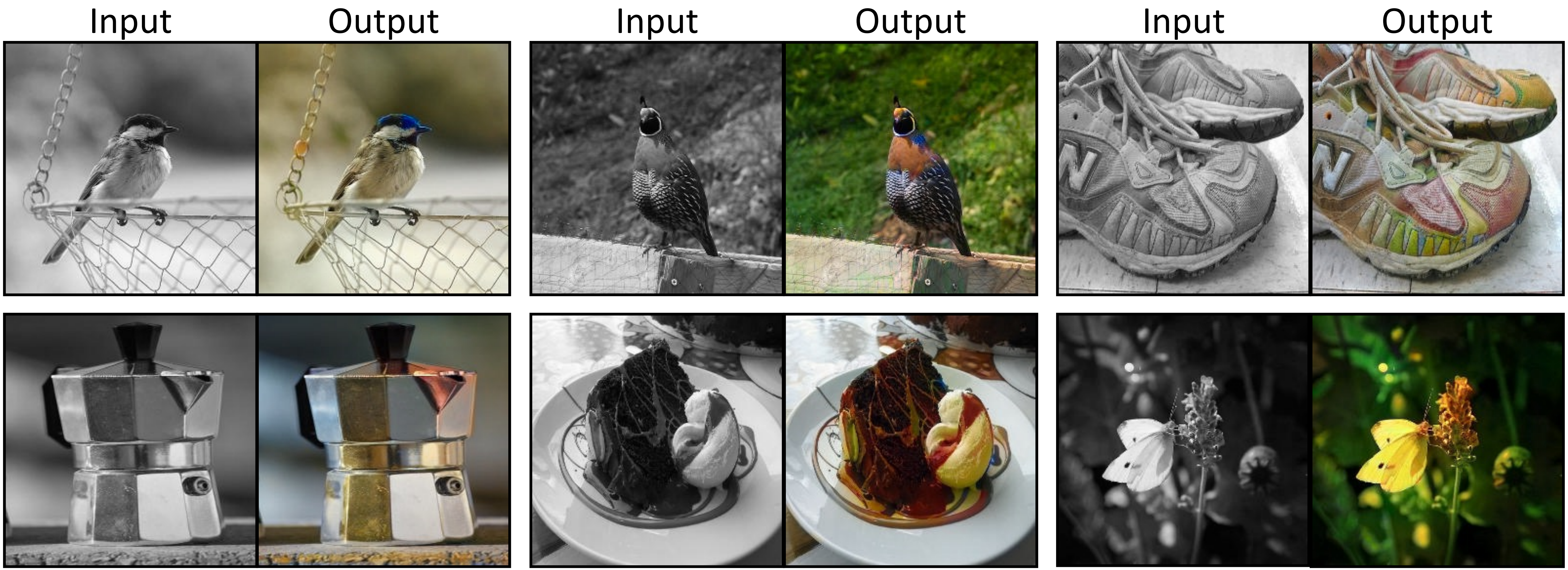
Input: **Sketch** → Output: **Photo**

Automatic Colorization with pix2pix

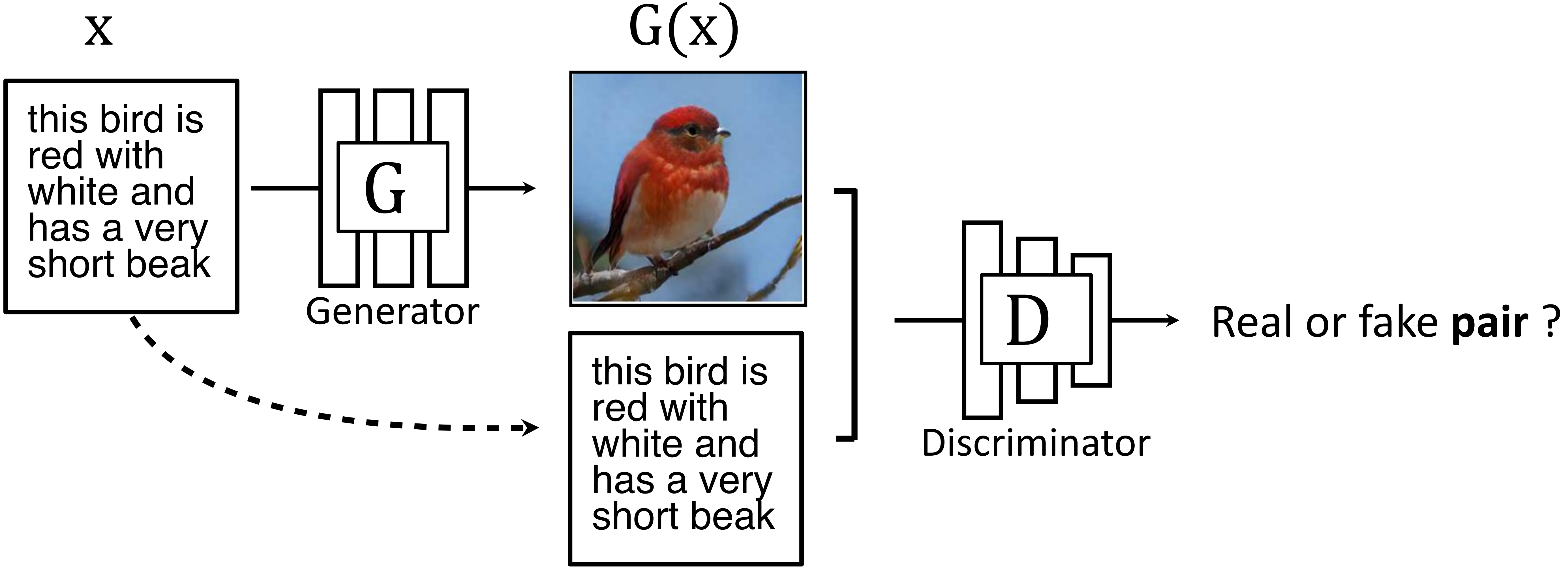
Input Output Input Output Input Output

Data from [Russakovsky et al. 2015]

Automatic Colorization with pix2pix



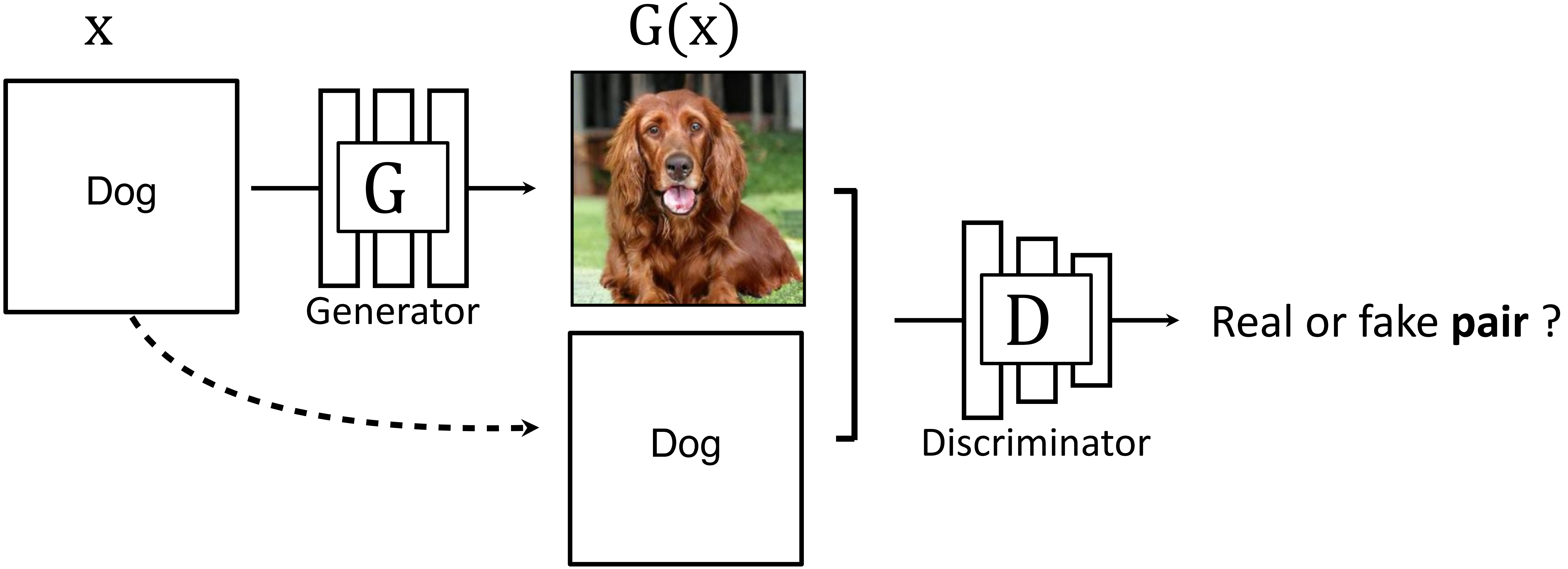
Data from [Russakovsky et al. 2015]



Input: **Text** $\rightarrow$ Output: **Photo**

Text-to-Image Synthesis

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



Input: **Class** → Output: **Photo**

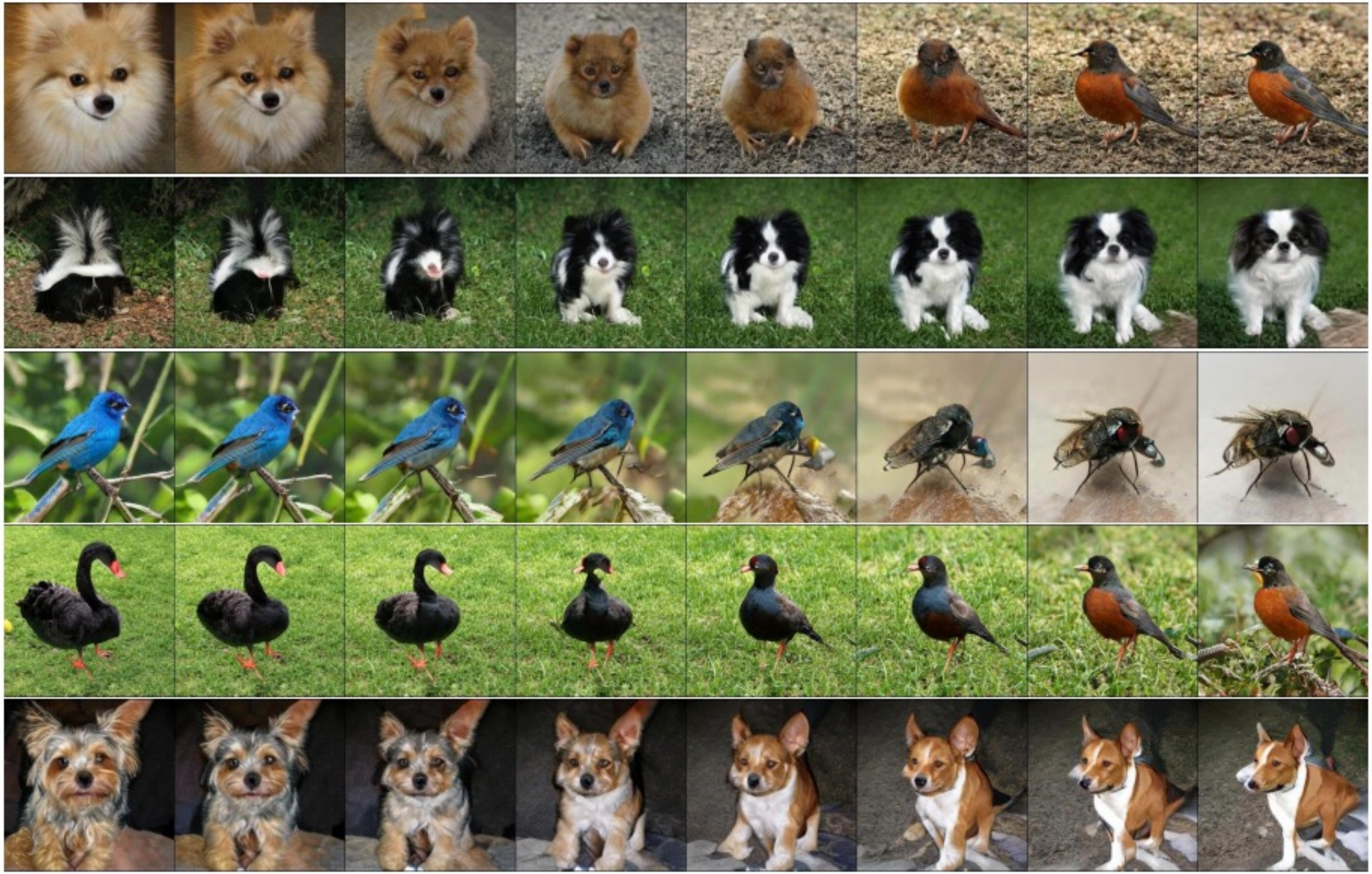
<u>Class-conditional GANs</u>

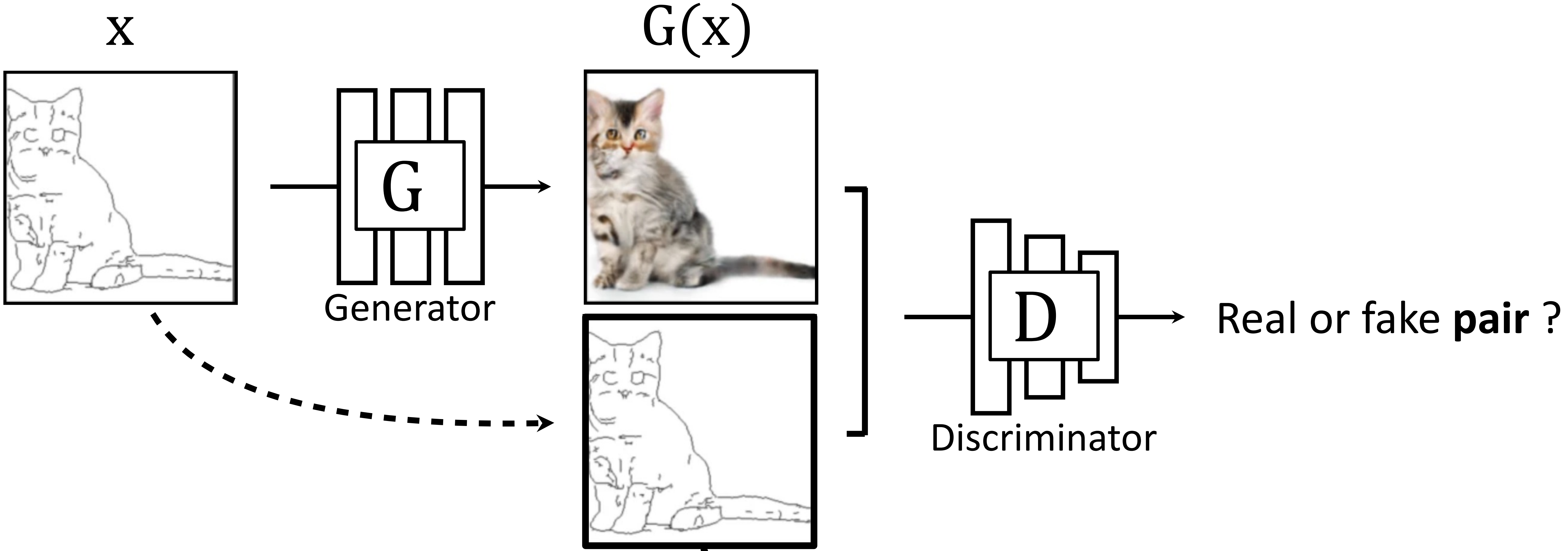
cGANs [Mirza and Osindero. 2014], SAGAN [Zhang et al., 2018], BigGAN [Brock et al., 2019]

StyleGAN-XL [Sauer et al., 2022]

BigGAN

(a) 128×128 (b) 256×256 (c) 512×512 (d)





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]

Limitations

- One-to-one mapping.
- Low-resolution output.
- Requires paired training data

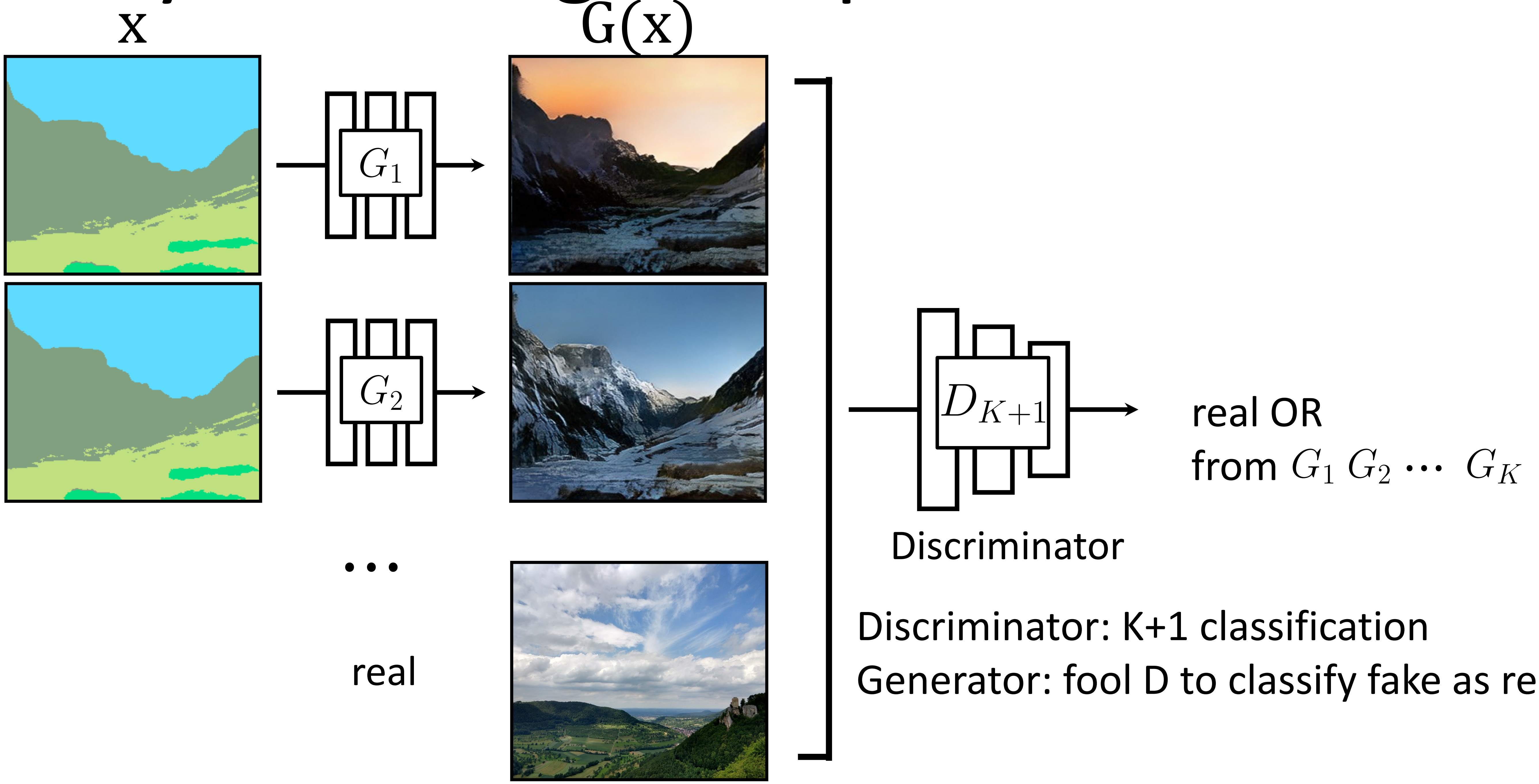
Improving Conditional GANs

- **Multimodal synthesis.**
- High-resolution synthesis.
- Model training without pairs (next lecture)

Group Discussion

- https://docs.google.com/forms/d/e/1FAIpQLSdfSXMRLddytfNaDOFAORBaOTxLPQTLWTELxpyAr8NJrFcZhA/viewform?usp=sharing

Synthesizing Multiple Results



Discriminator: K+1 classification

Generator: fool D to classify fake as real

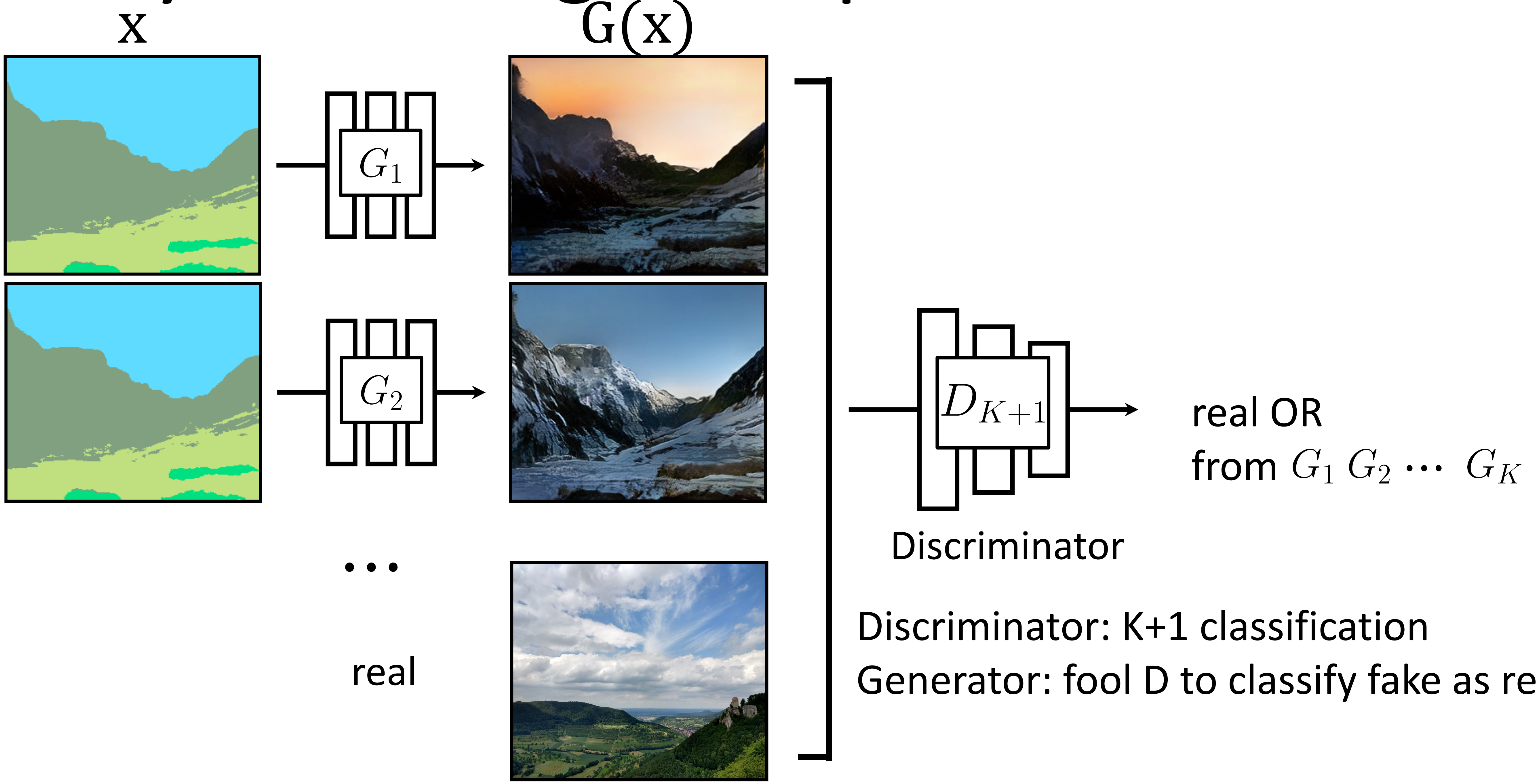
Multi-agent Diverse GANs [Ghosh et al., CVPR 2018]

Synthesizing Multiple Results

Night input Day output 1 Day output 2 Day output 3

Multi-agent Diverse GANs [Ghosh et al., CVPR 2018]

Synthesizing Multiple Results



Discriminator: K+1 classification

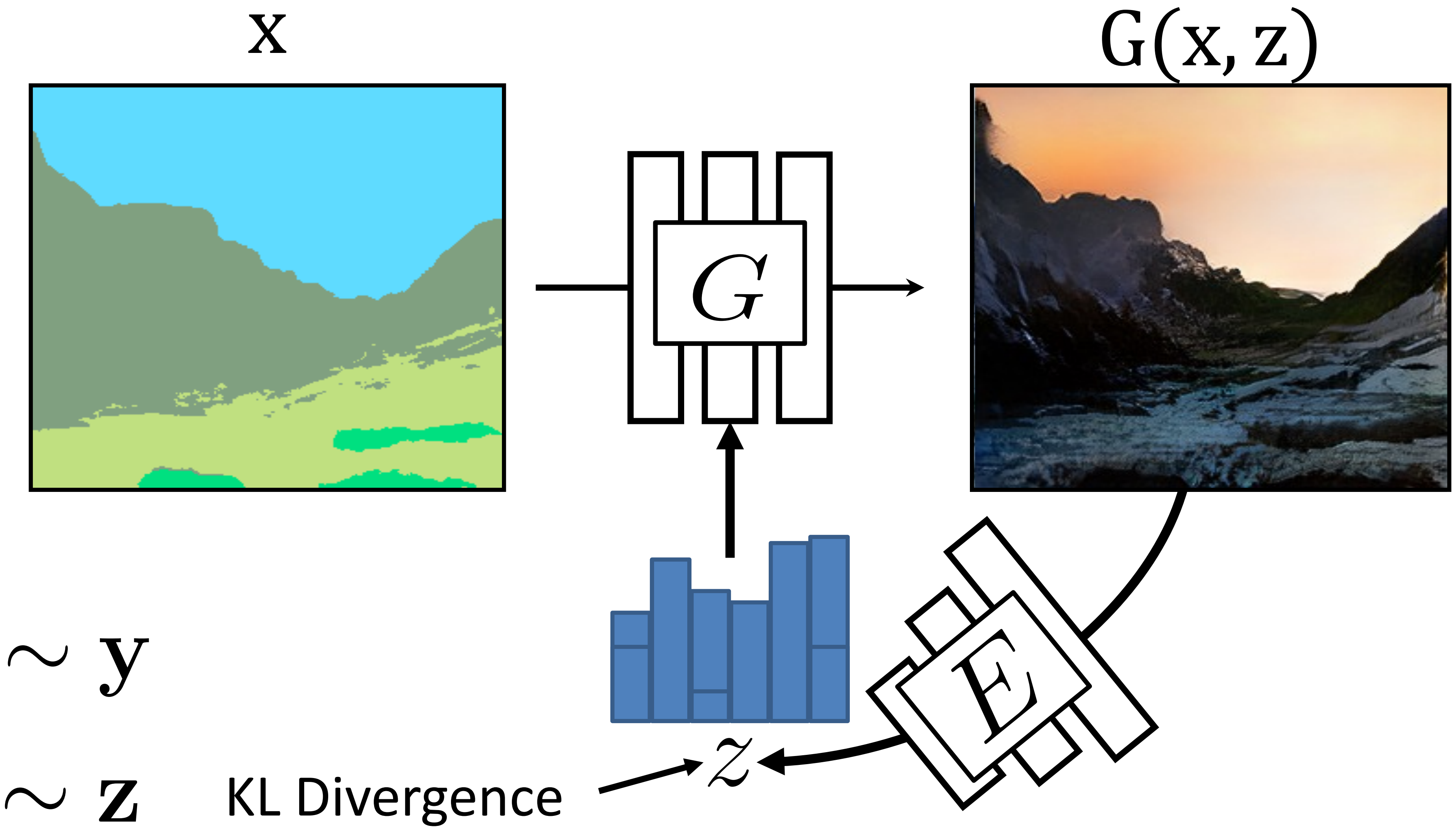
Generator: fool D to classify fake as real

Multi-agent Diverse GANs [Ghosh et al., CVPR 2018]

Synthesizing Multiple Results

Multi-agent Diverse GANs [Ghosh et al., CVPR 2018]

Synthesizing Multiple Results



$$G(\mathbf{x}, E(\mathbf{y})) \sim \mathbf{y}$$

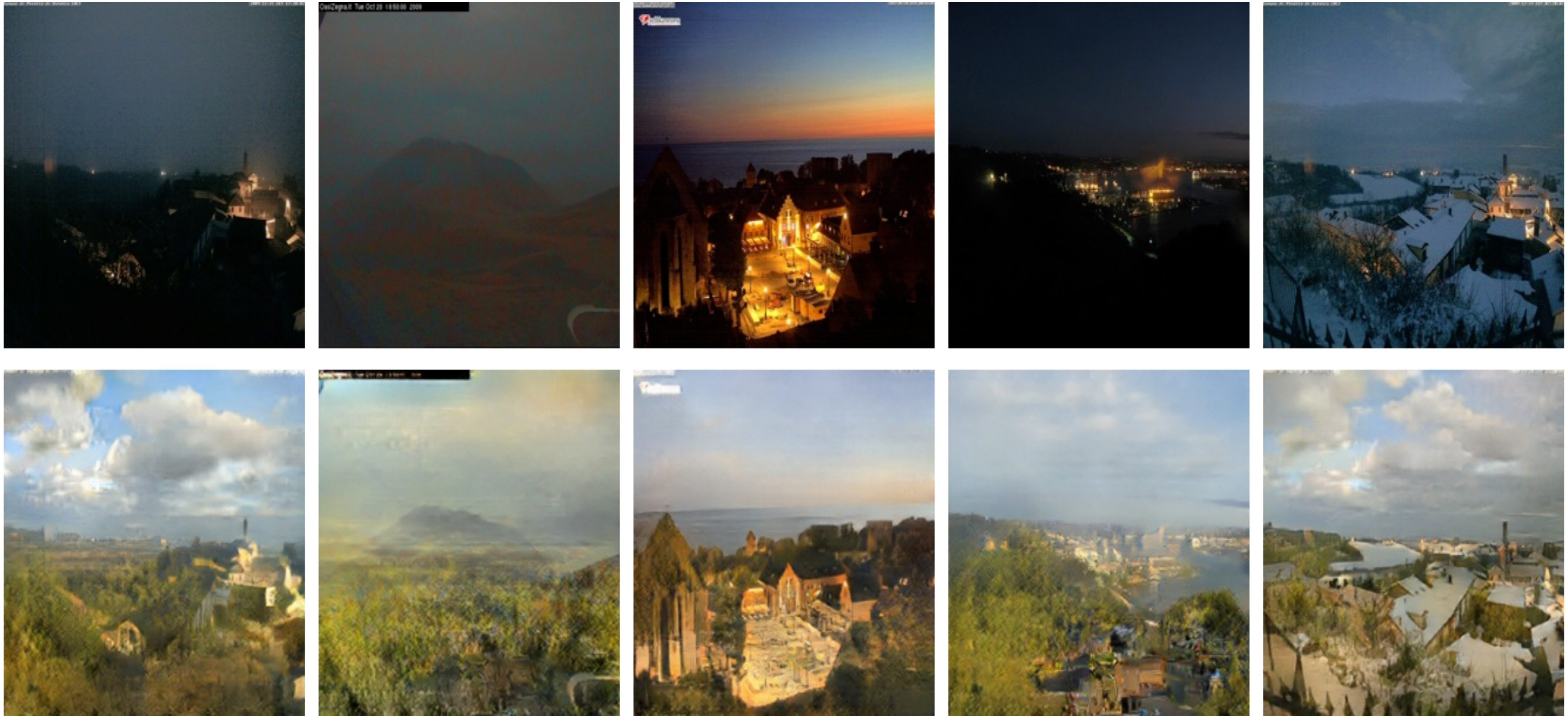
$$E(G(\mathbf{x}, \mathbf{z})) \sim \mathbf{z}$$

VAE-GAN [Larsen et al., 2016], BicycleGAN [Zhu et al., 2017]

Synthesizing Multiple Results

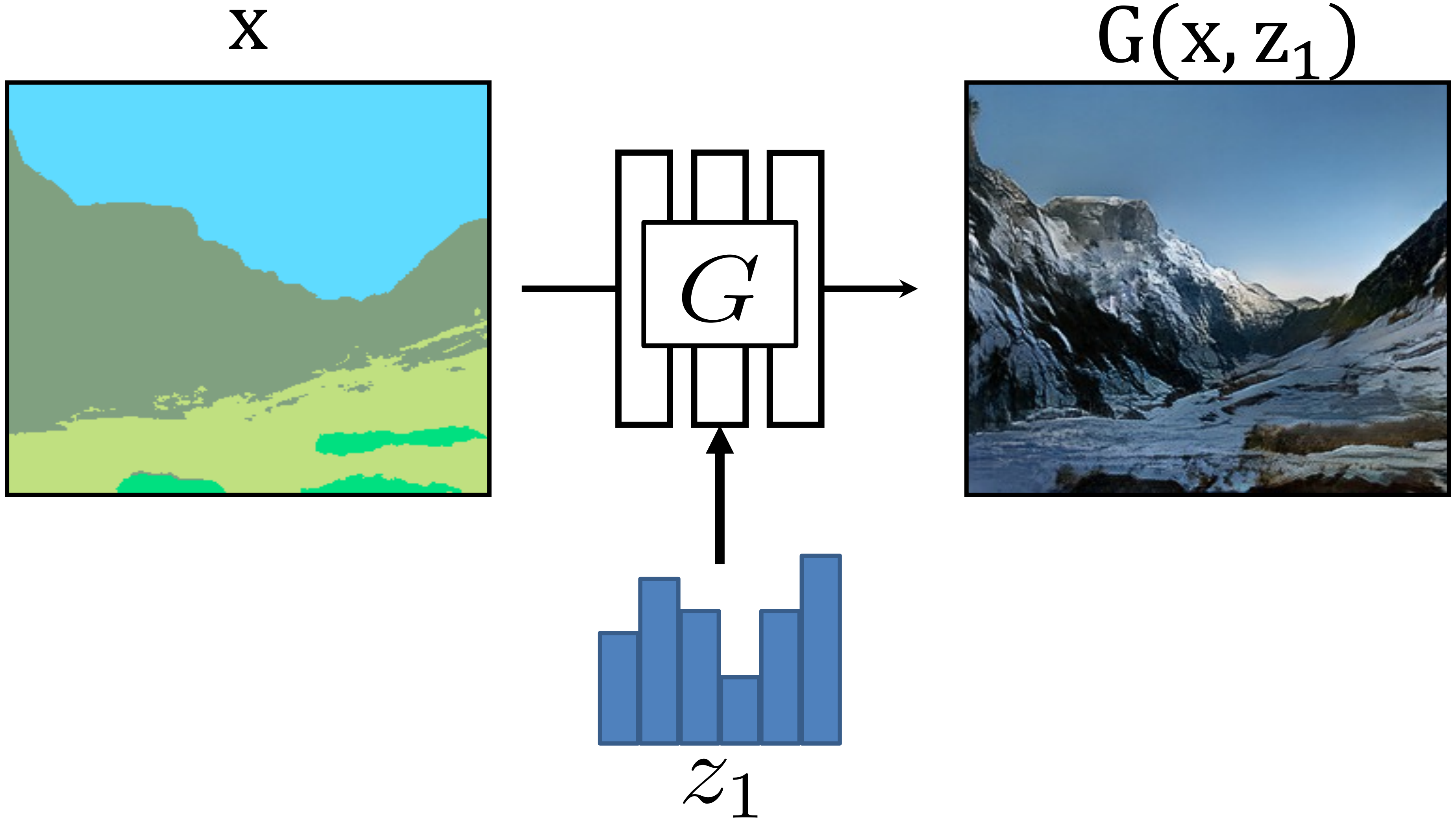
BicycleGAN [Zhu et al., 2017]

Synthesizing Multiple Results



BicycleGAN [Zhu et al., 2017]

Synthesizing Multiple Results

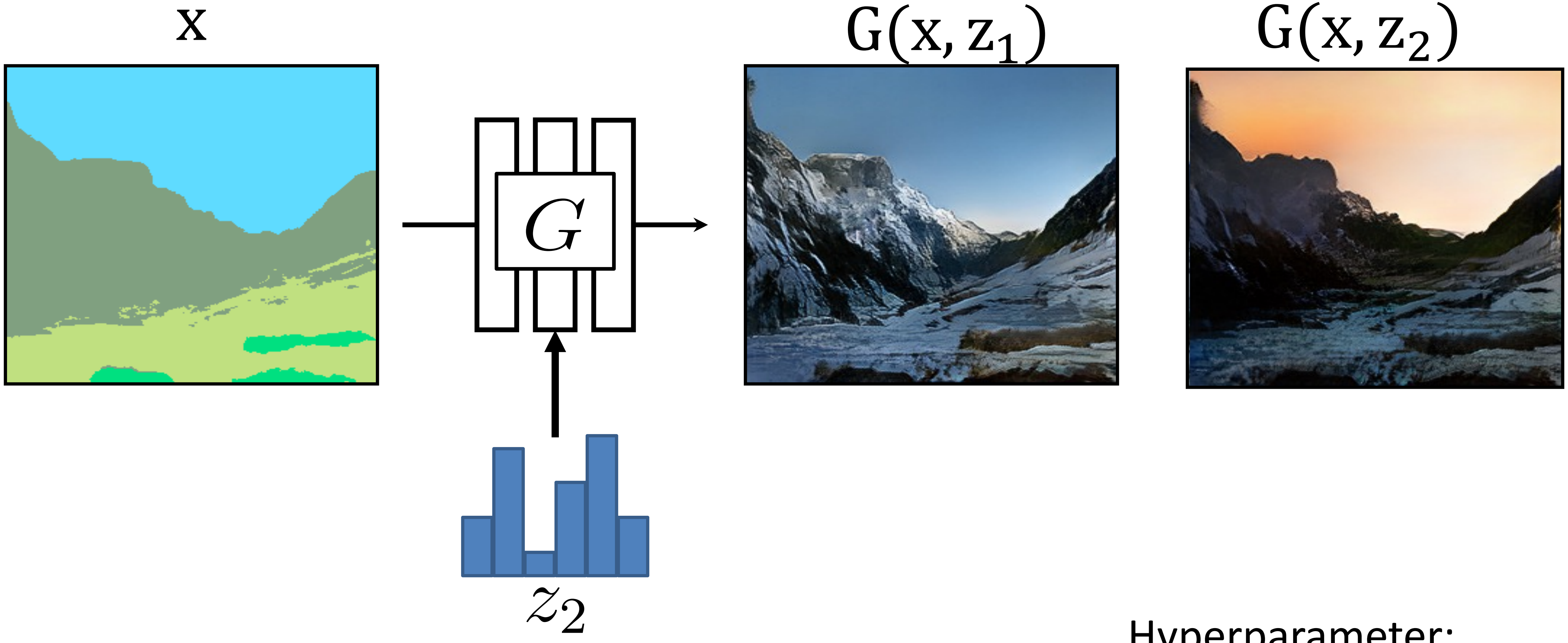


Hyperparameter:
Degree of diversity

$$\max_G \ \mathcal{L}_{\boldsymbol{z}}(G) = \mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\left[\min\left(\frac{\|G(\boldsymbol{x}, \boldsymbol{z}_1) - G(\boldsymbol{x}, \boldsymbol{z}_2)\|}{\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|}, \tau\right)\right],$$

Diversity-Sensitive GAN [Yang et al., 2019]

Synthesizing Multiple Results



$$\max_G \mathcal{L}_{\boldsymbol{z}}(G) = \mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\min\left(\frac{\| \quad - \quad \|}{\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|}, \tau\right)\right]$$

Hyperparameter:
Degree of diversity

Diversity-Sensitive GAN [Yang et al., 2019]

Synthesizing Multiple Results

Hyperparameter:
Degree of diversity

$$\max_{G} \mathcal{L}_{\boldsymbol{z}}(G) = \mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\min\left(\frac{\|G(\boldsymbol{x},\boldsymbol{z}_1) - G(\boldsymbol{x},\boldsymbol{z}_2)\|}{\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|}, \tau\right)\right]$$

Diversity-Sensitive GAN [Yang et al., 2019]

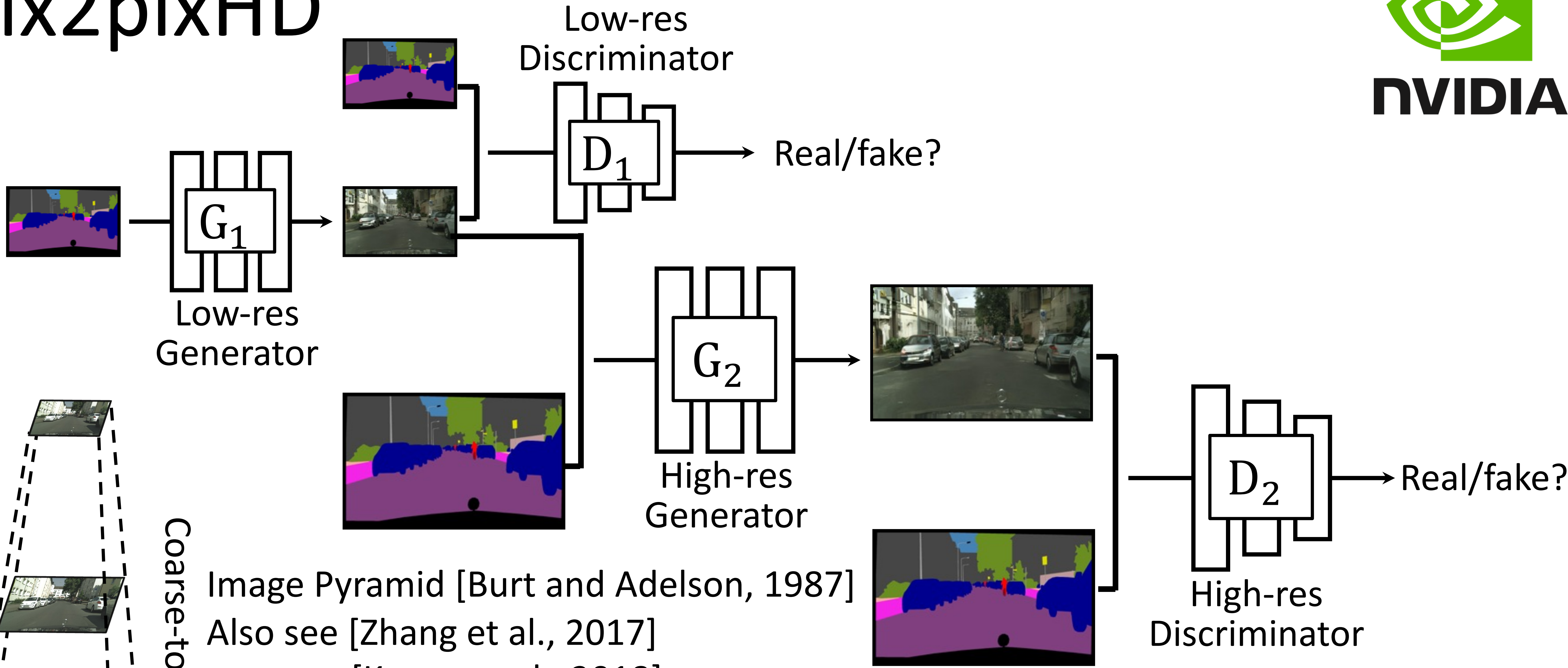
Improving Conditional GANs

- Multimodal synthesis.
- **High-resolution synthesis.**
- Model training without pairs (next lecture)

The Curse of Dimensionality

Pix2pix output

pix2pixHD



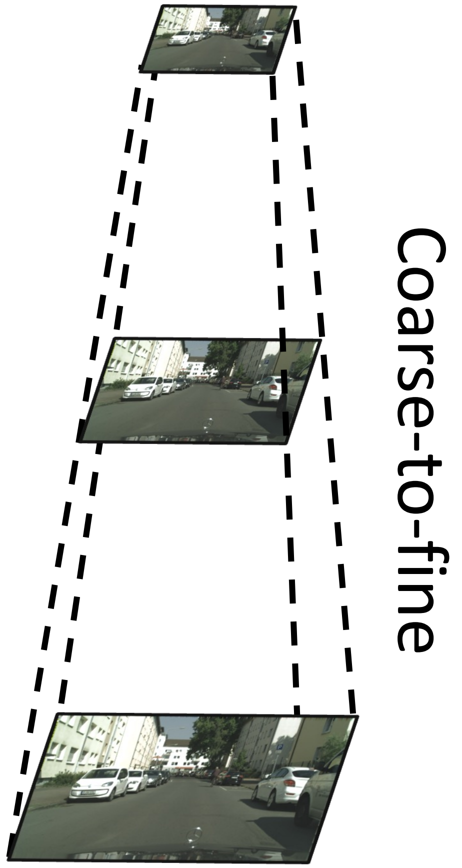
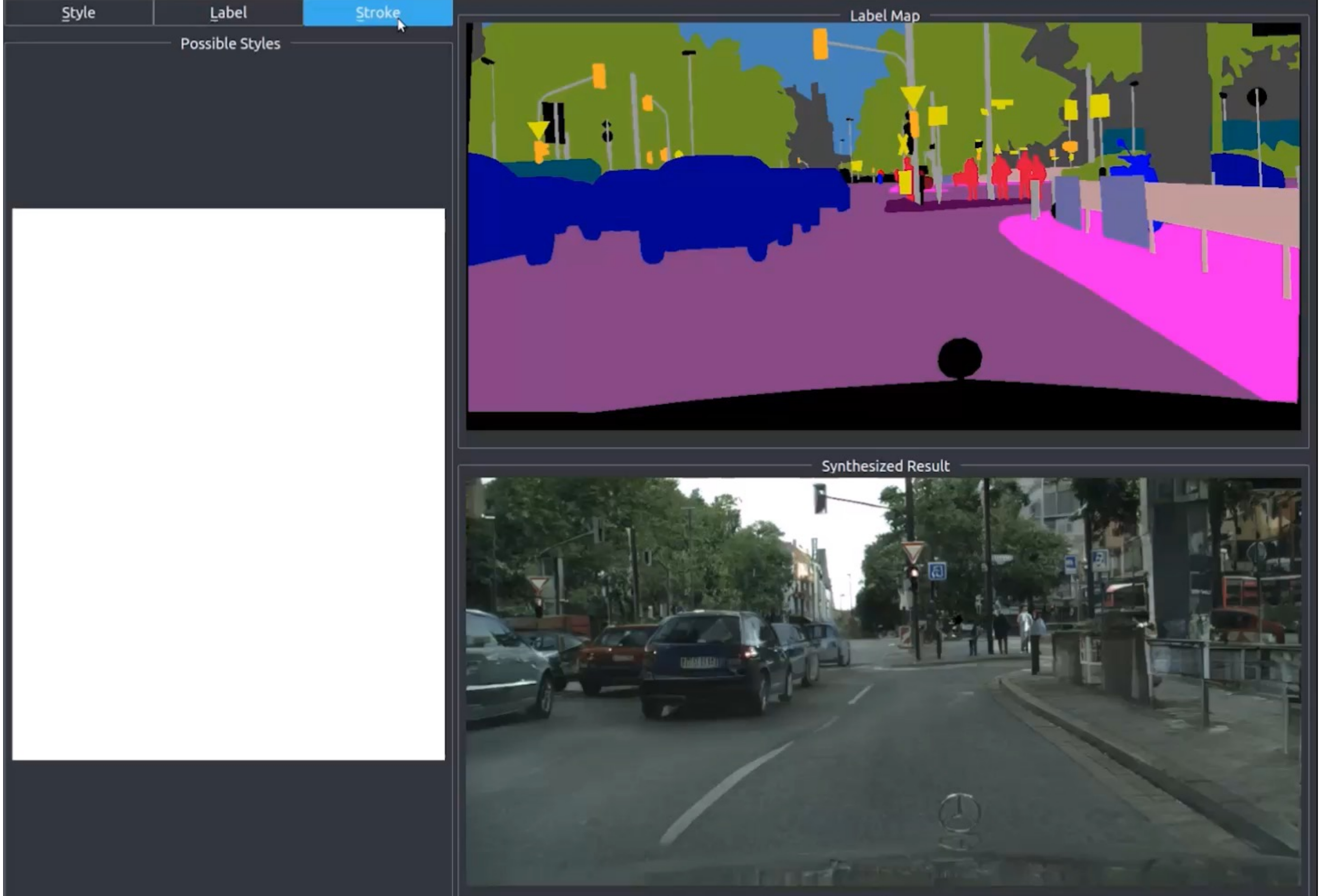


Image Pyramid [Burt and Adelson, 1987]
Also see [Zhang et al., 2017]
[Karras et al., 2018]

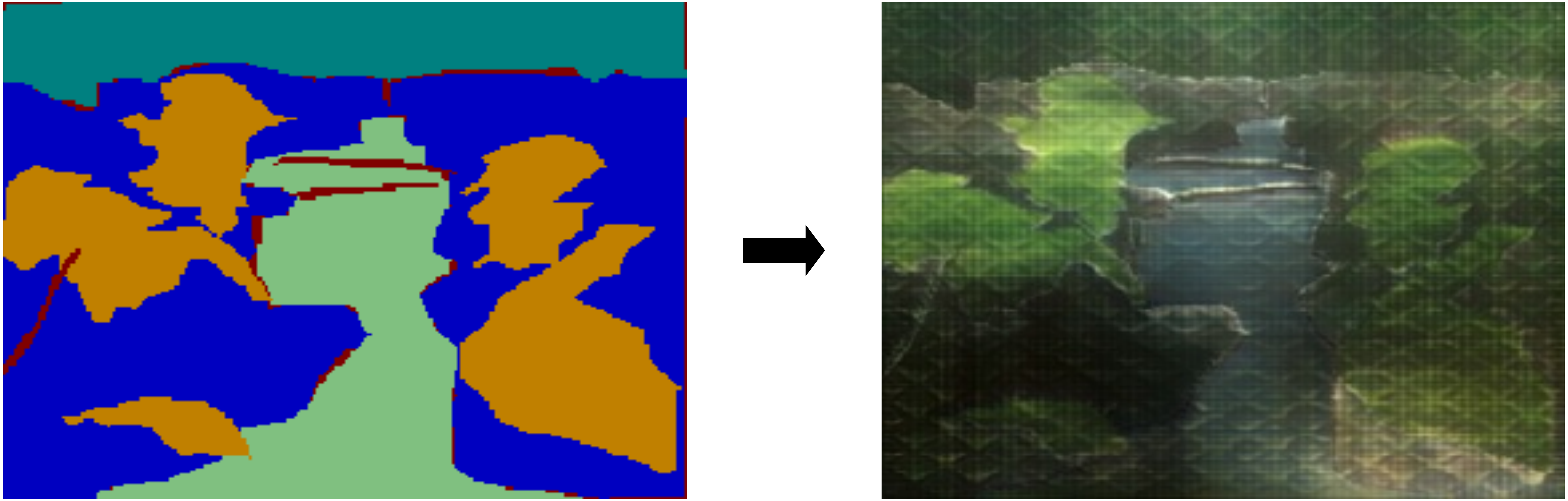
Objective: Multi-scale GANs loss + Perceptual Loss
+ Feature Matching Loss (with Discriminator's features)

pix2pixHD [Wang et al., 2018]

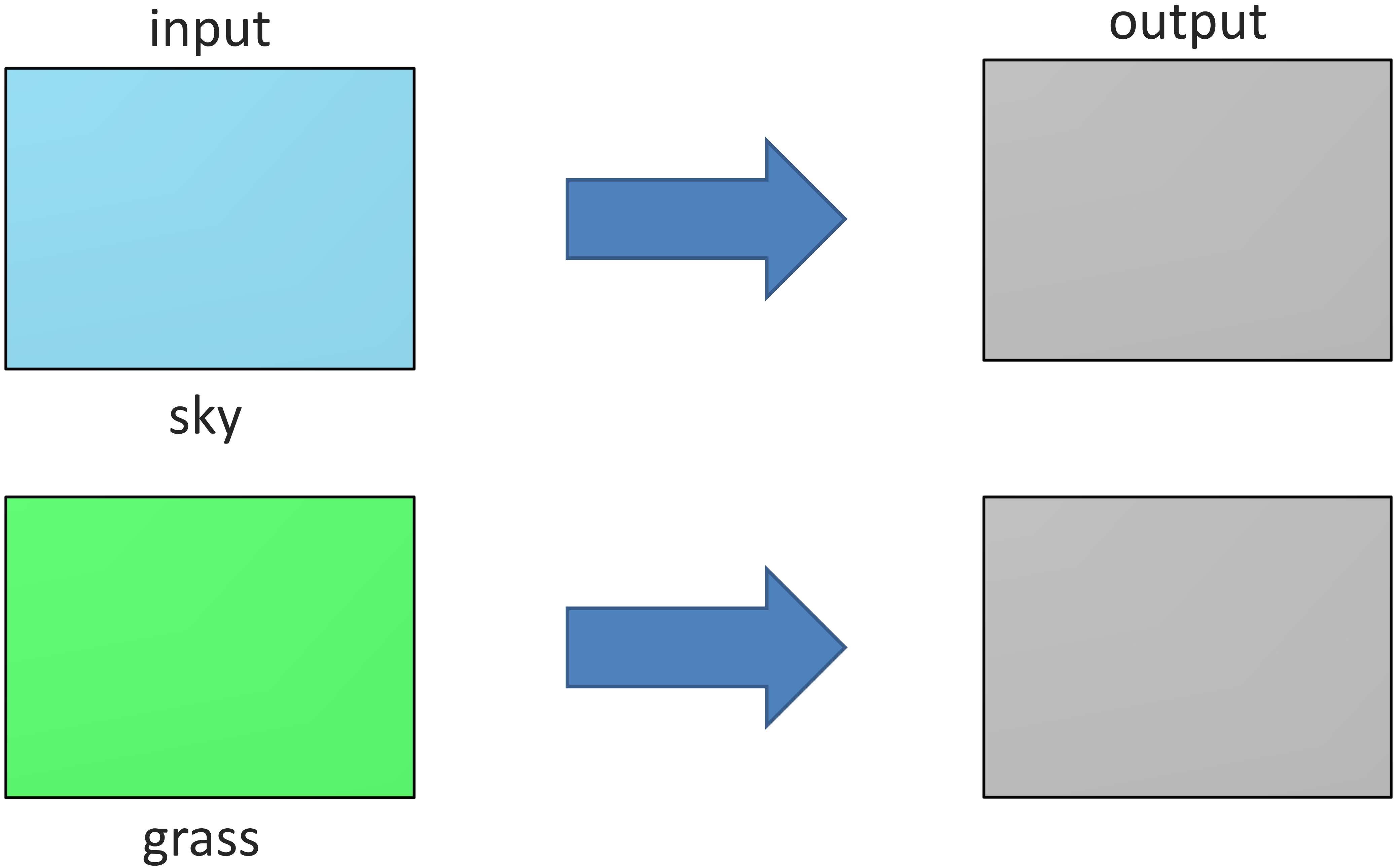
pix2pixHD: 2048×1024


Style
Label
Stroke
Possible Styles
Label Map
Synthesized Result

Conditional Image Synthesis in the Wild

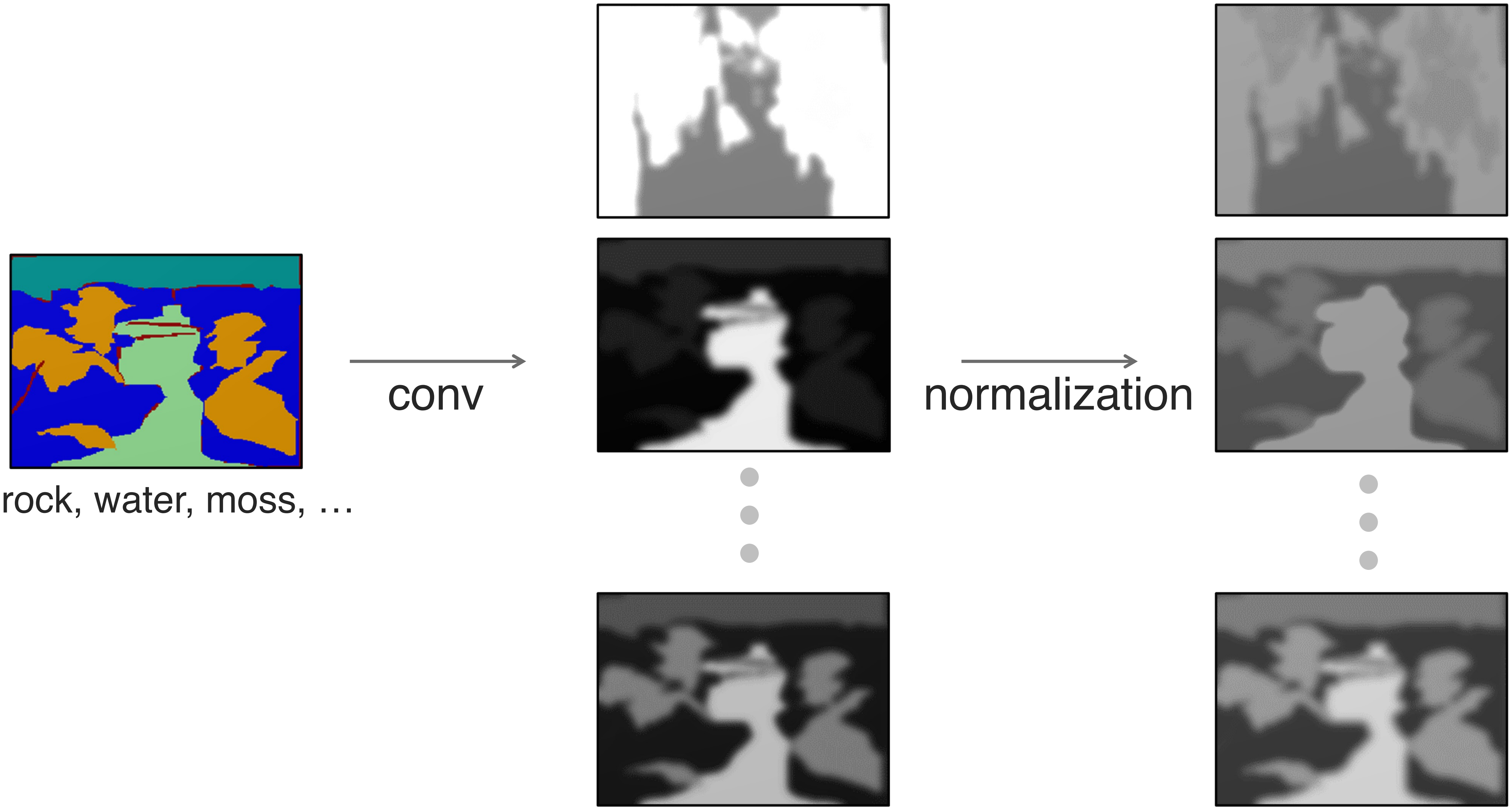


pix2pixHD [Wang et al., 2018]

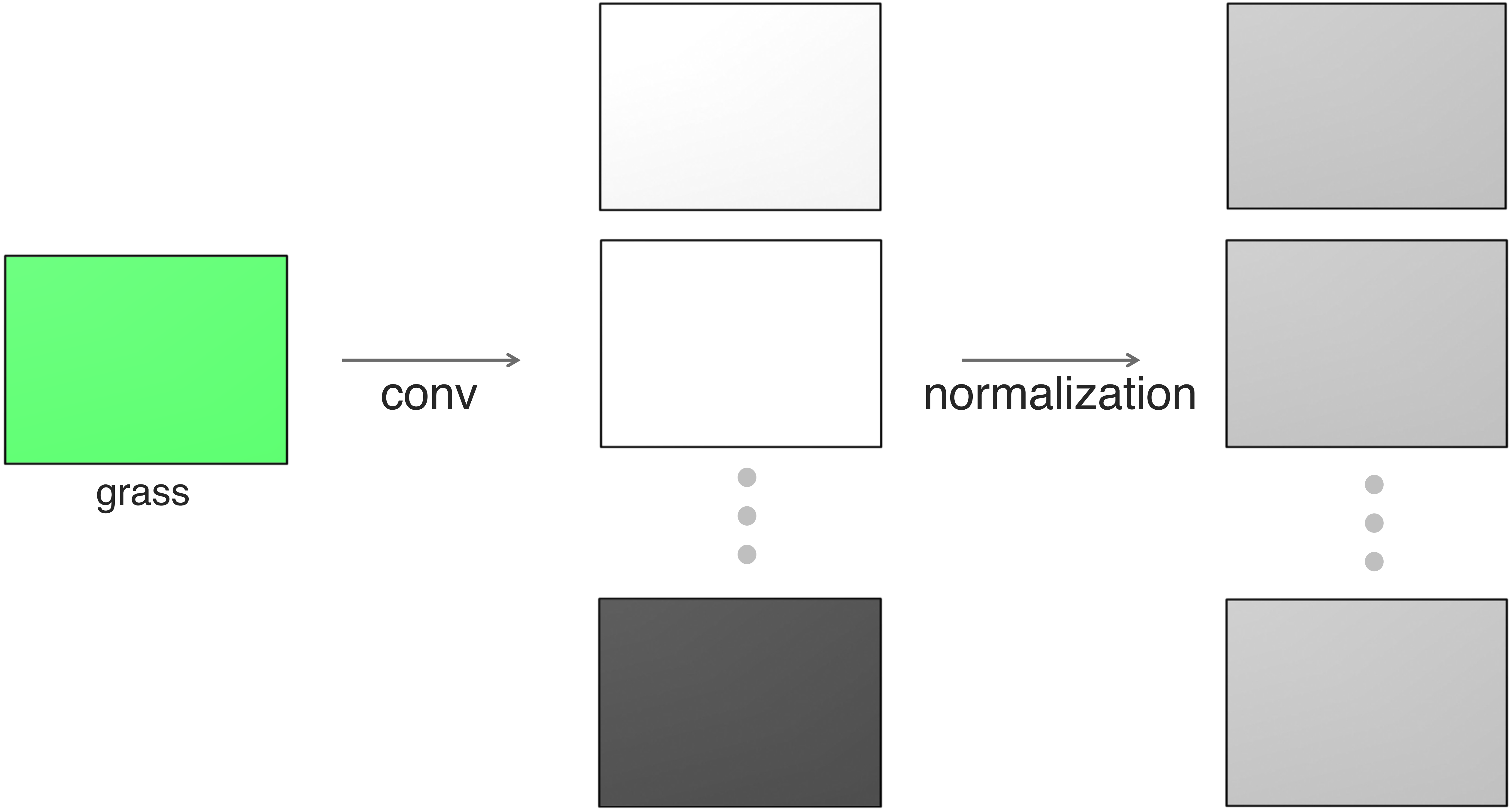


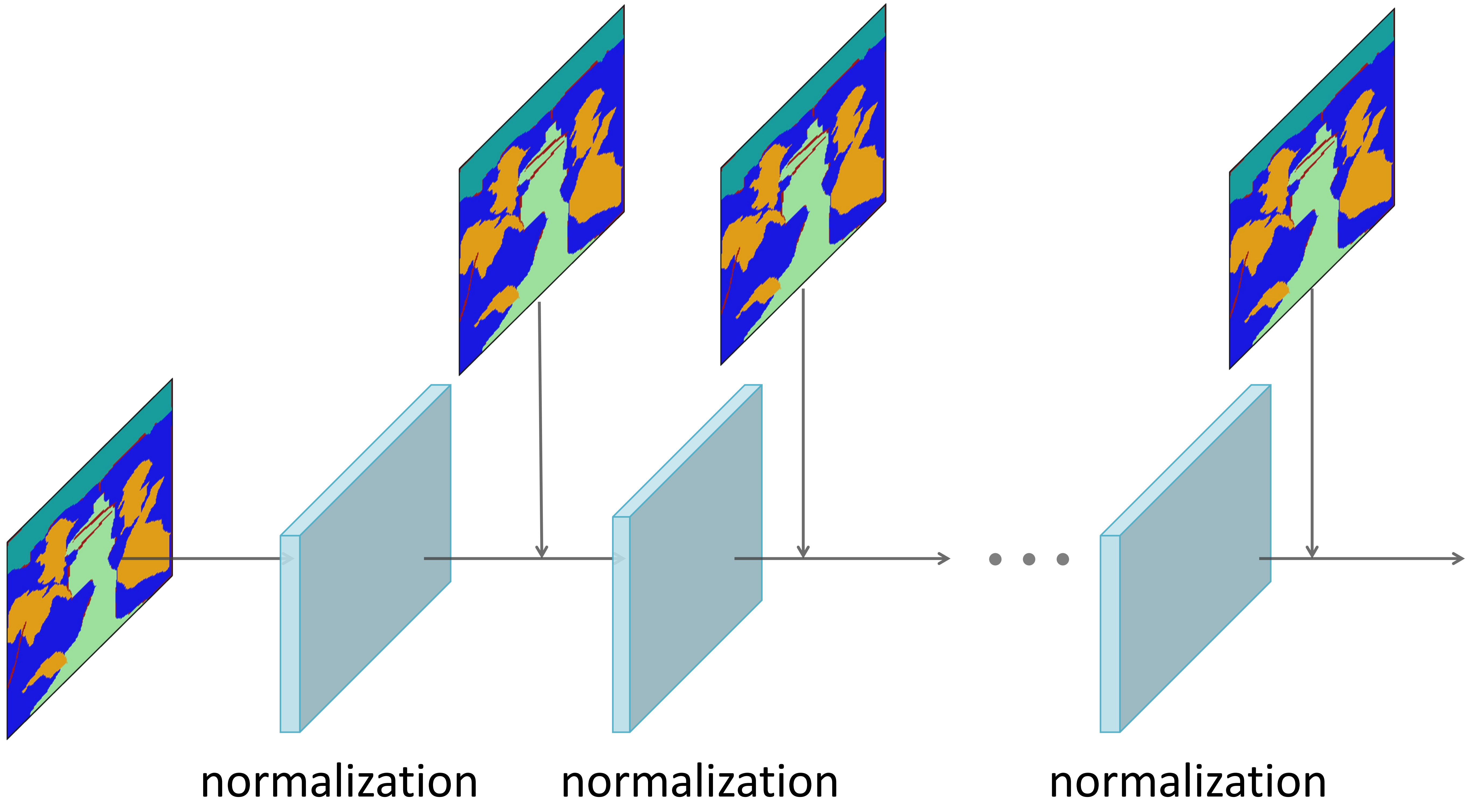
pix2pixHD [Wang et al., 2018]

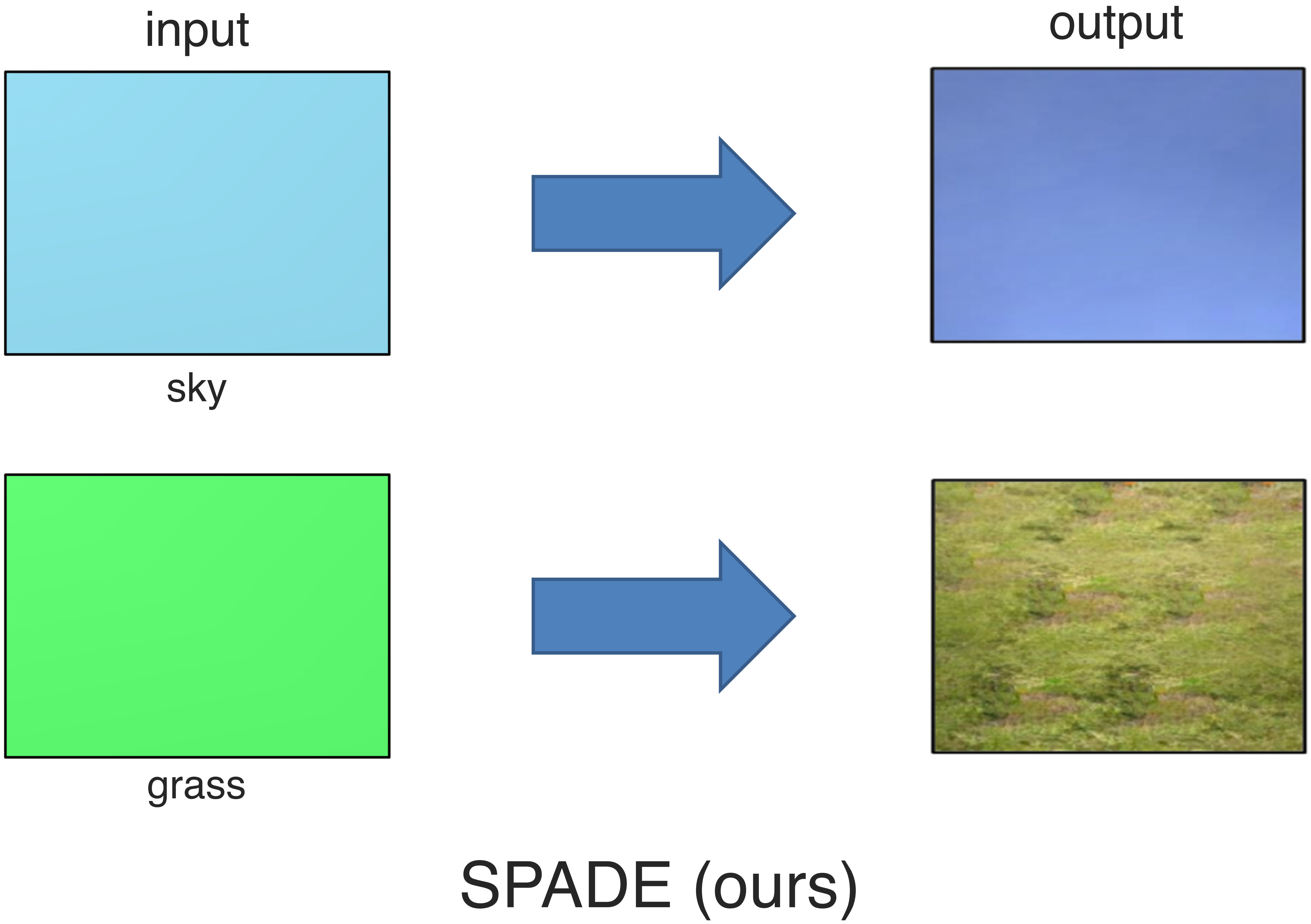
Problem with standard networks



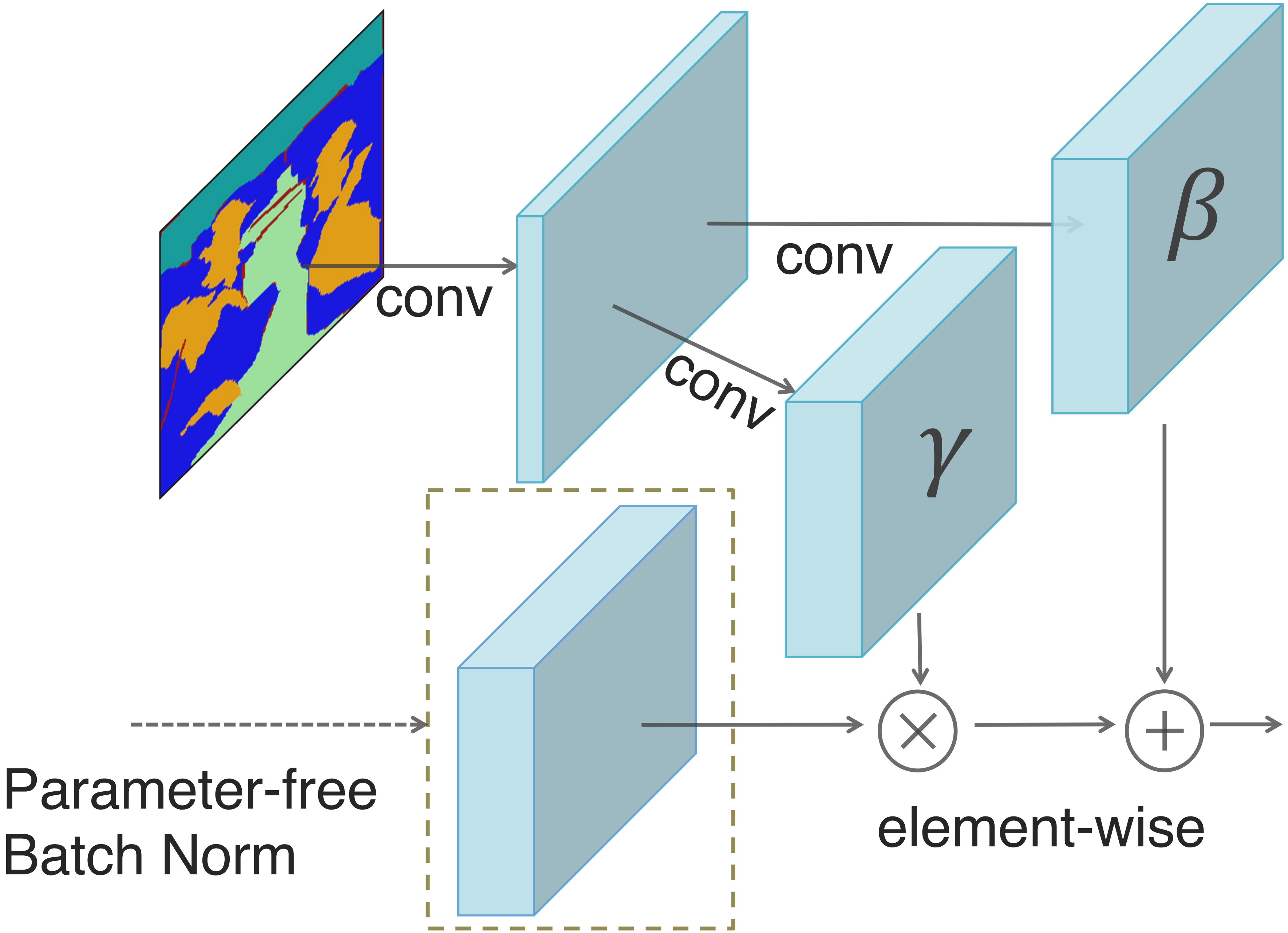
Problem with standard networks




normalization
normalization
normalization


input
output
sky
grass
SPADE (ours)

SPADE (SPAtially ADaptive DEnormalization)



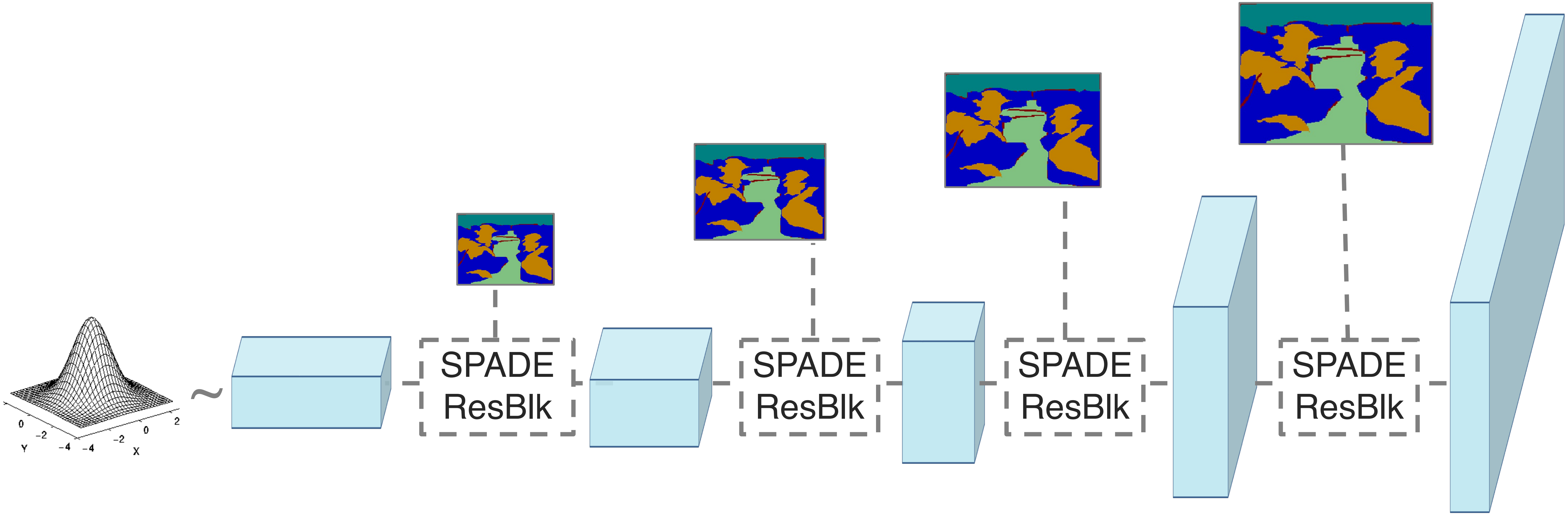
SPADE (SPAtially ADaptive DEnormalization)

Batch Norm (Ioffe et al. 2015)

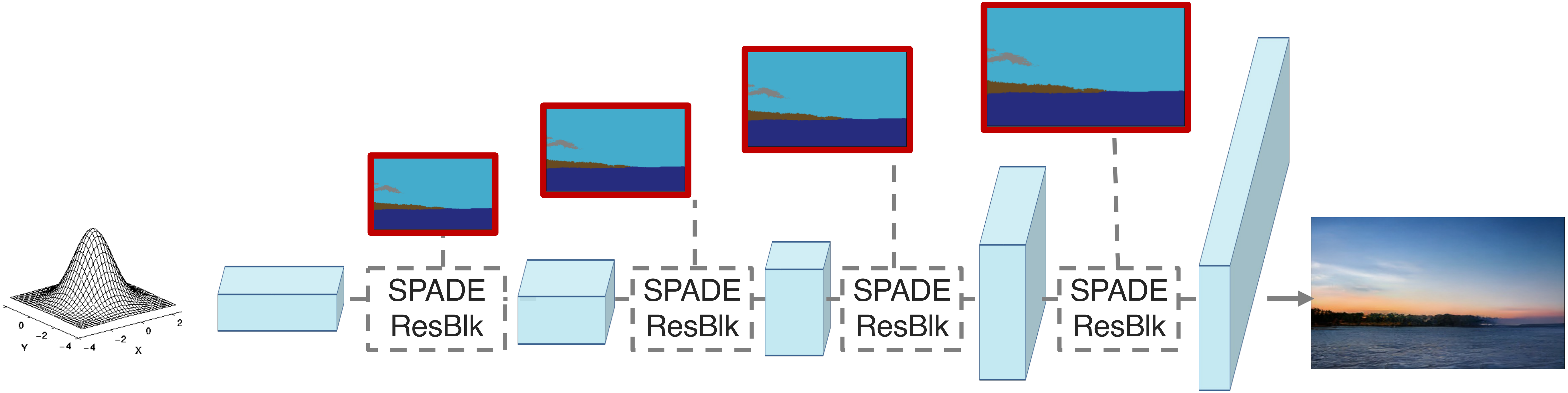
$$y = \underbrace{\frac{x - \mu}{\sigma}}_{\text{normalization}} \cdot \underbrace{\gamma + \beta}_{\text{affine transform}}$$

See other adaptive/conditional normalization: conditional BN (Dumoulin et al.), AdaIN (Huang and Belongie), SFT (Wang et al.)

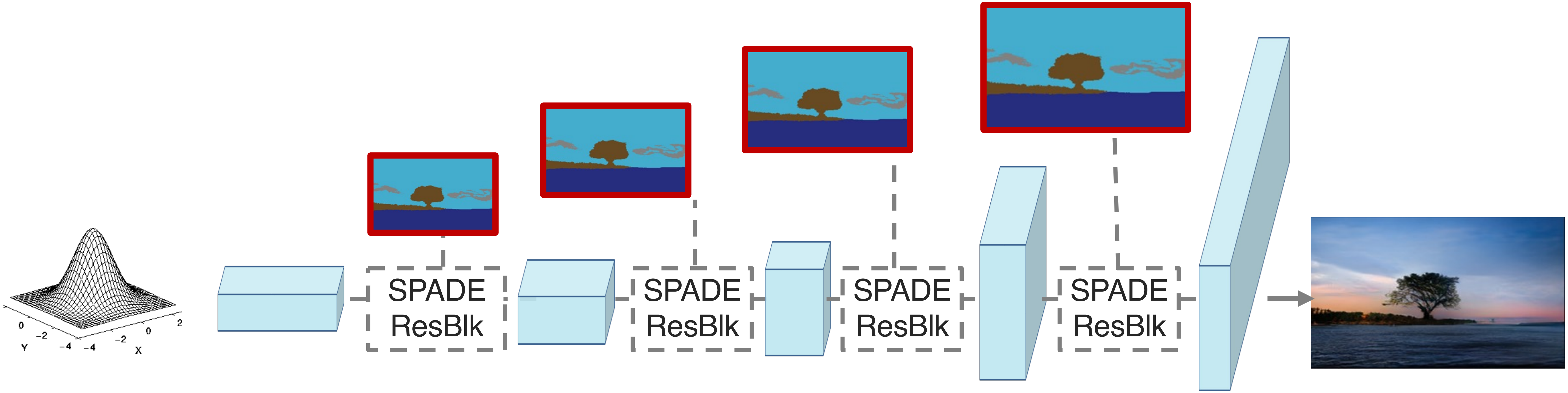
Generator



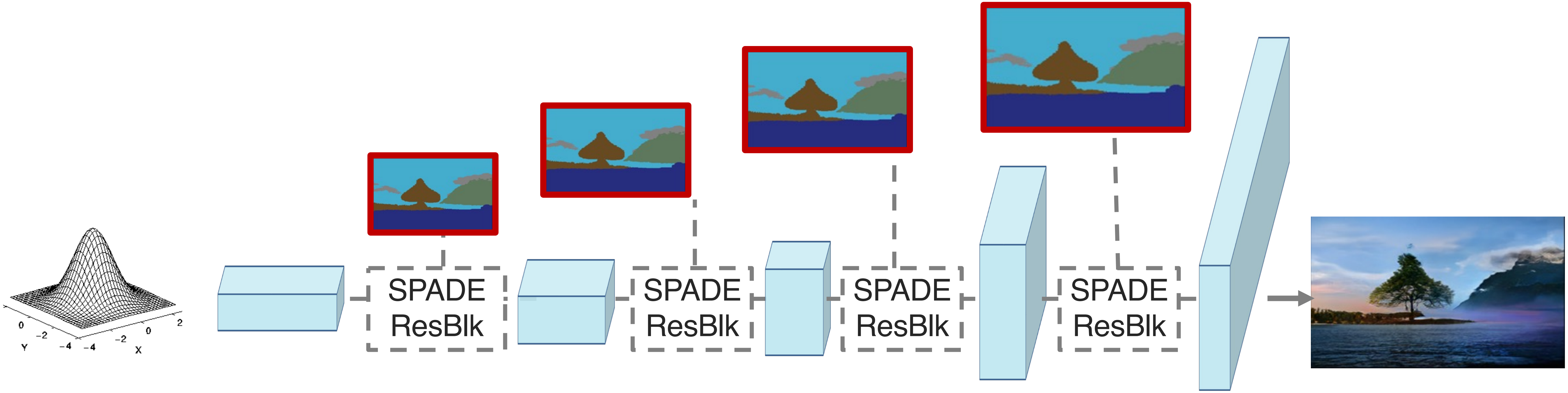
Semantic Control



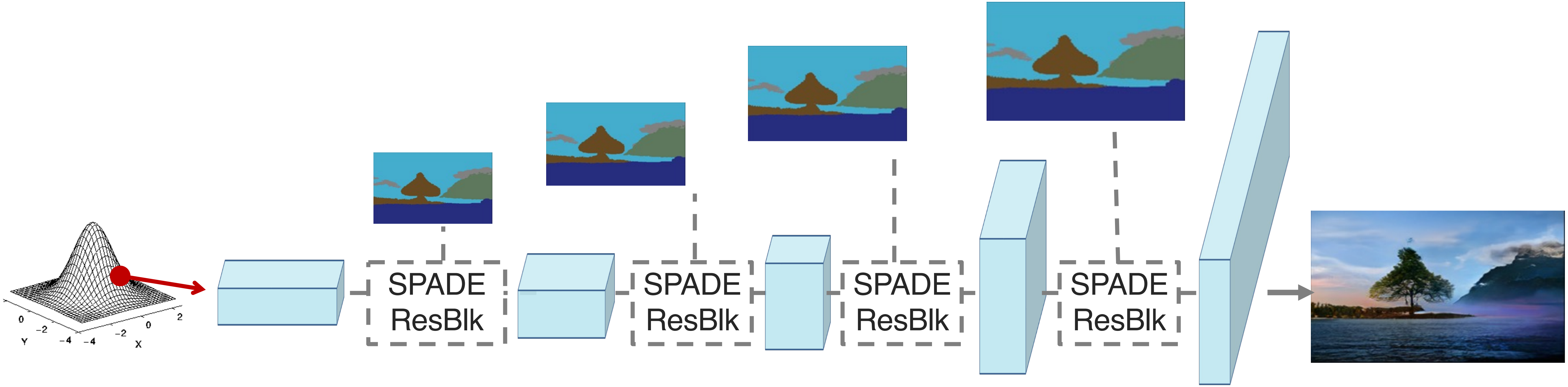
Semantic Control



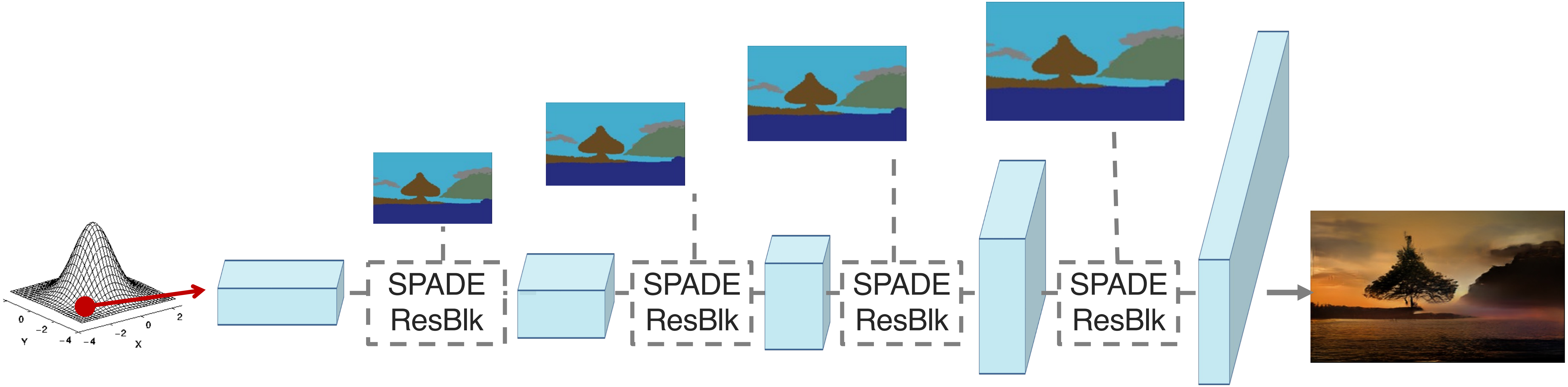
Semantic Control



Style Control

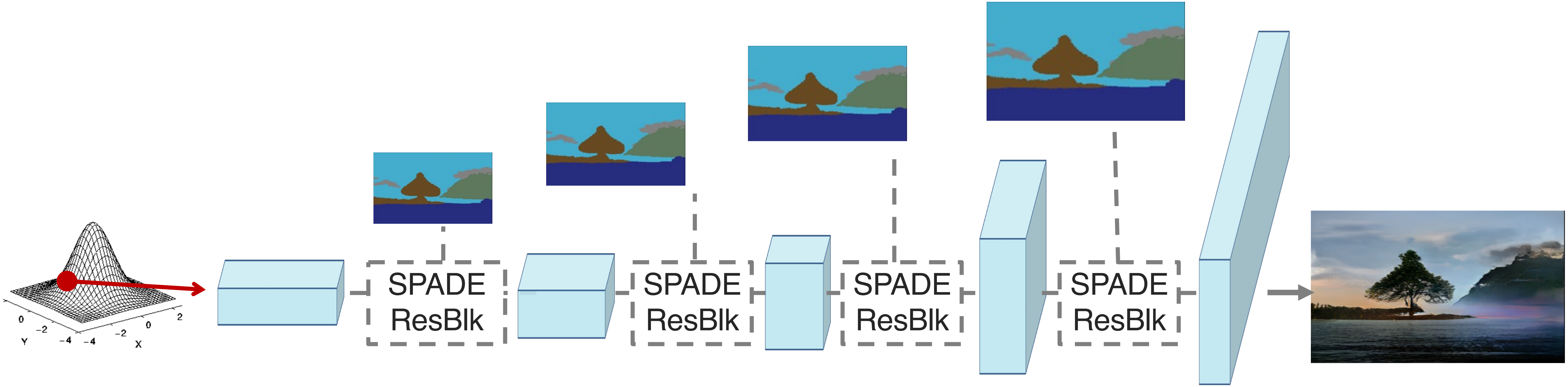


Style Control

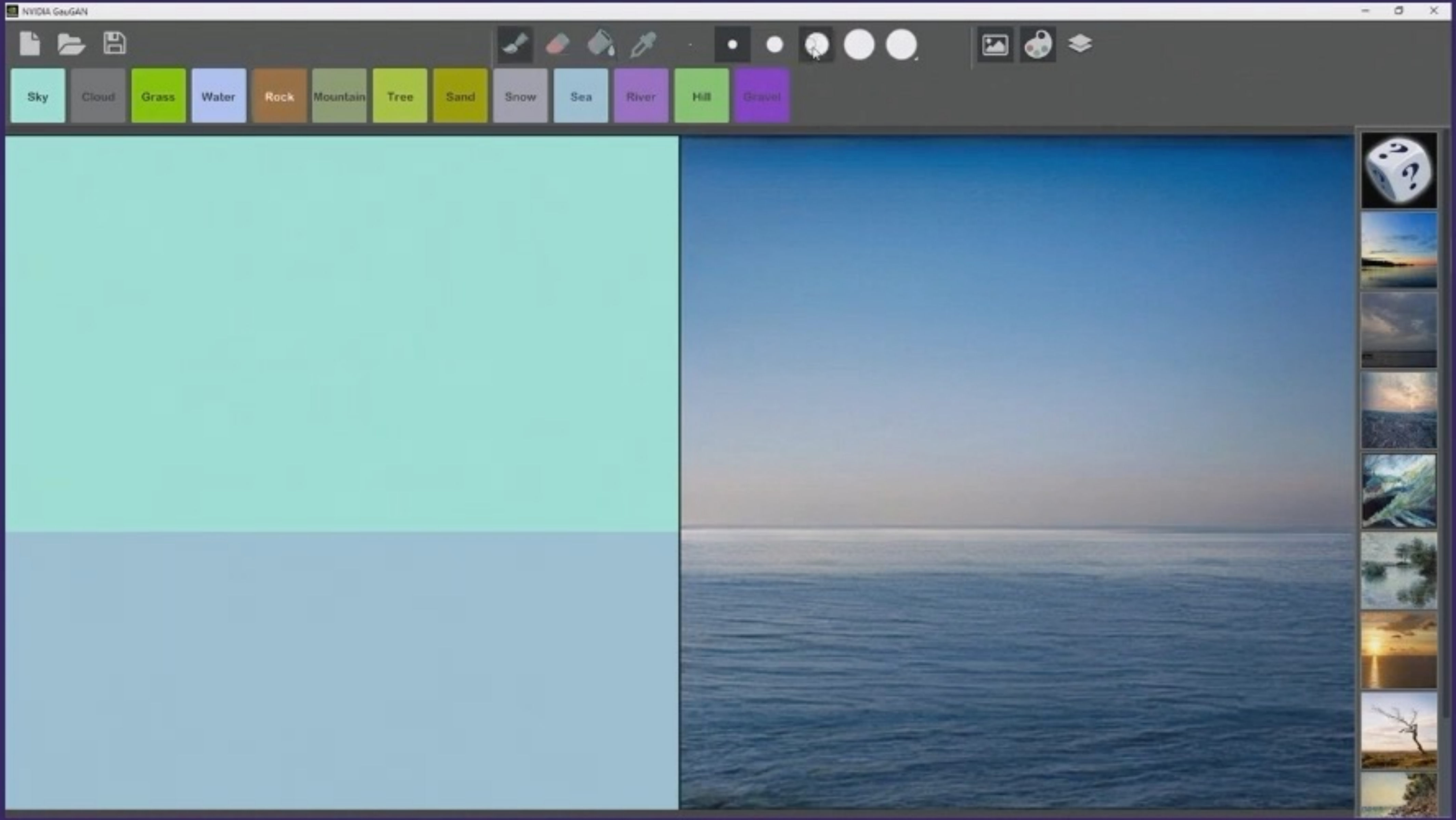


Style Manipulation

Style Control



Style Manipulation

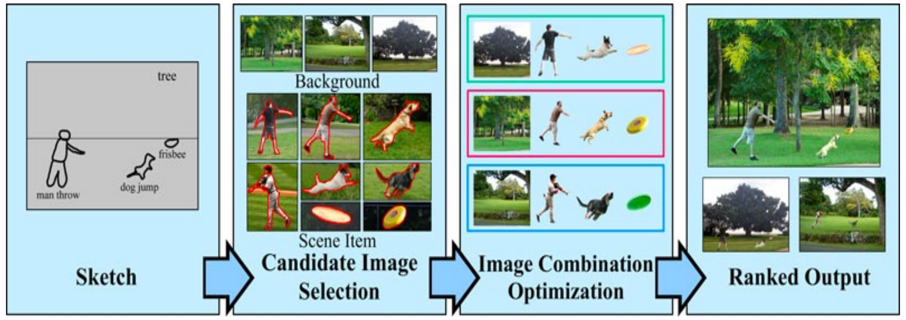


thrive
SIGGRAPH2019
LOS ANGELES 28 JULY - 1 AUGUST

SIGGRAPH 2019 Real-time Live! “Best of Show Award" and "Audience Choice Award"

By Darek Zabrocki, Concept Designer and Illustrator

Learning vs. Exemplar-based

	Learning-based	Exemplar-based
		
	[Isola et al], [Wang et al] [Park et al], SEAN [Zhu et al]	[Johnson et al], [Lalonde et al] [Tao et al], [Bansal et al]
Speed		
Local realism		
Global realism		
Match Input		

Thank You!

16-726, Spring 2025