CMP717
Image Processing

Image to Image Translation

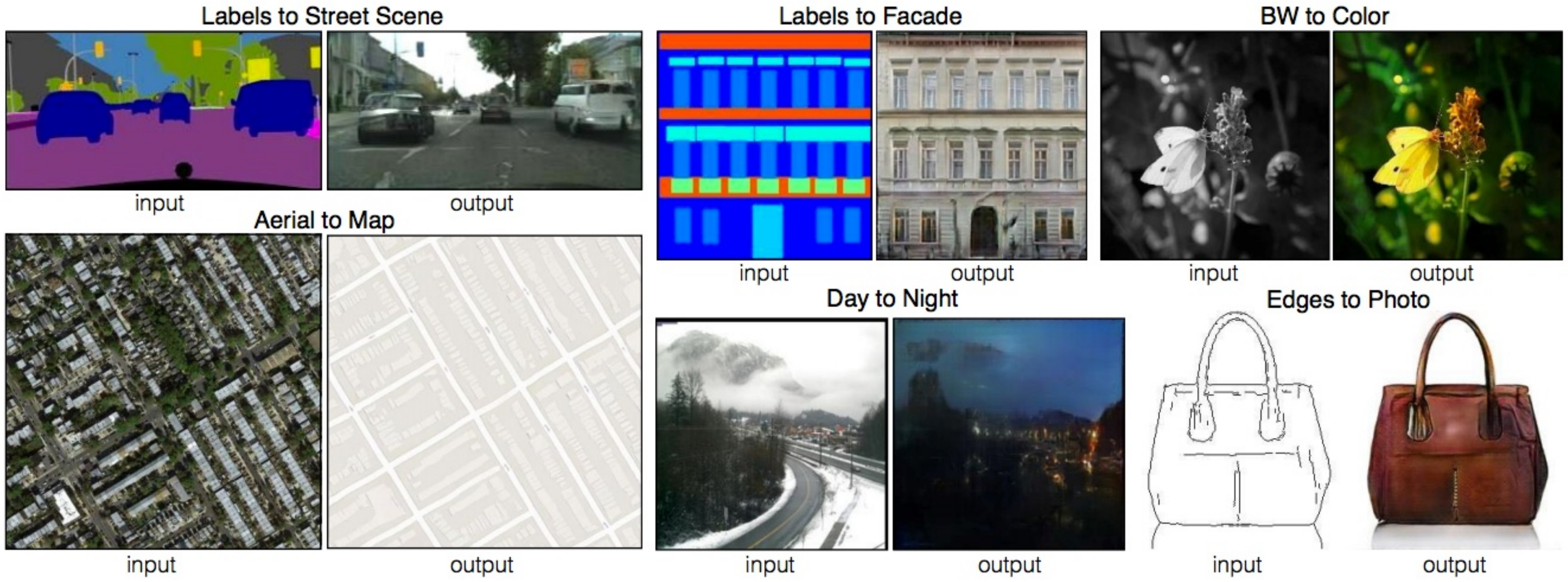
Erkut Erdem
Hacettepe University
Computer Vision Lab (HUCVL)

Outline

- Paired image-to-image translation
- Unpaired image-to-image translation

Outline

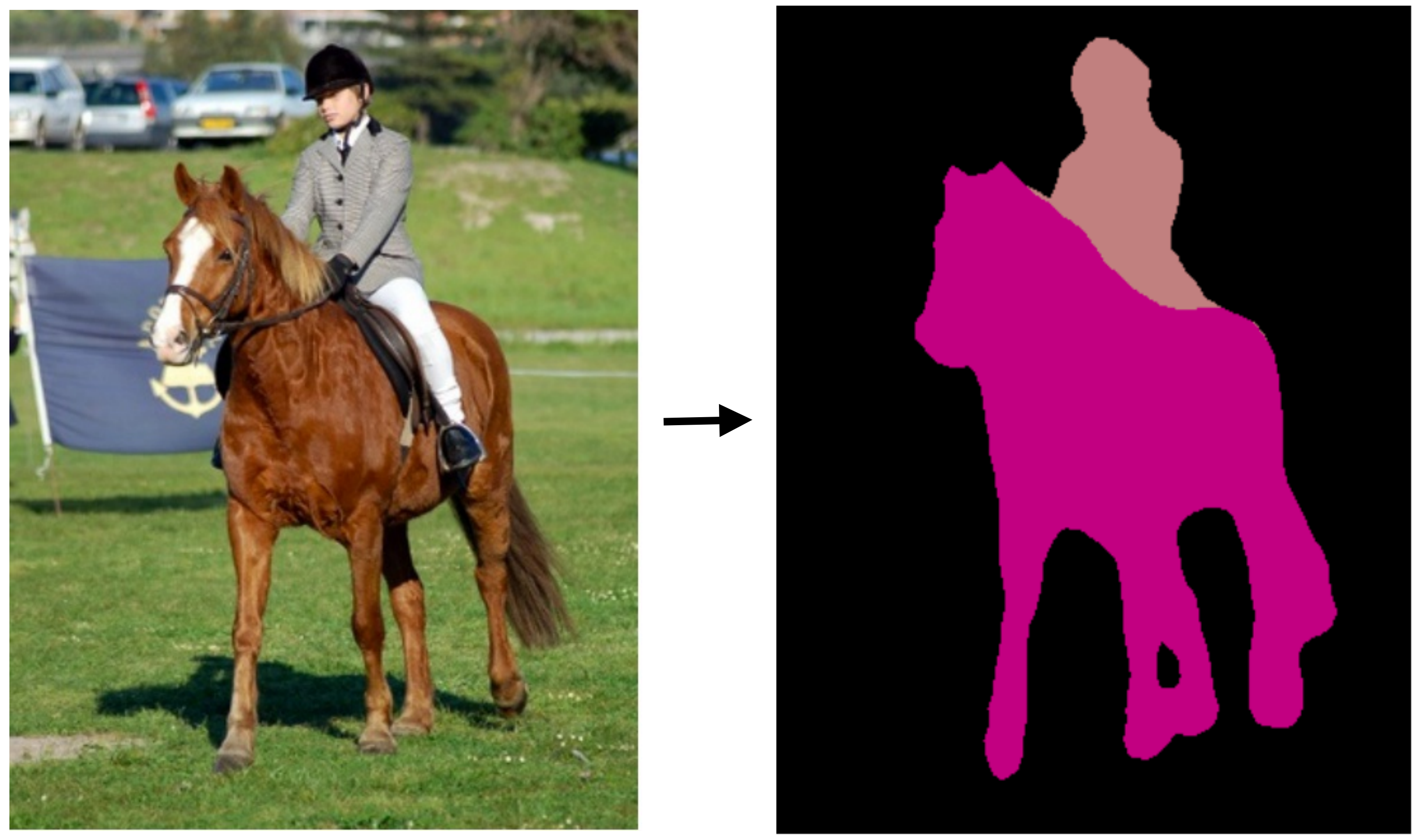
- Paired image-to-image translation
- Unpaired image-to-image translation



Acknowledgement: The slides adapted from CVPR 2018 Tutorial on GANs by Philip Isola on paired image to image translation.

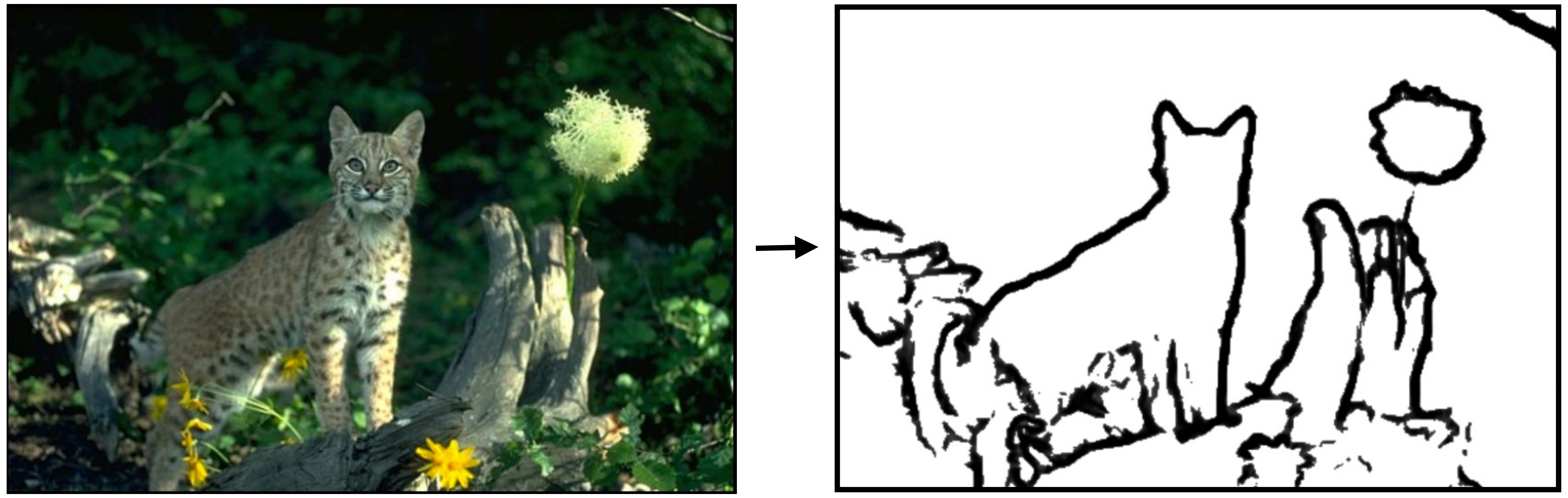
Image-to-Image Translation

Object labeling



[Long et al. 2015]

Edge Detection

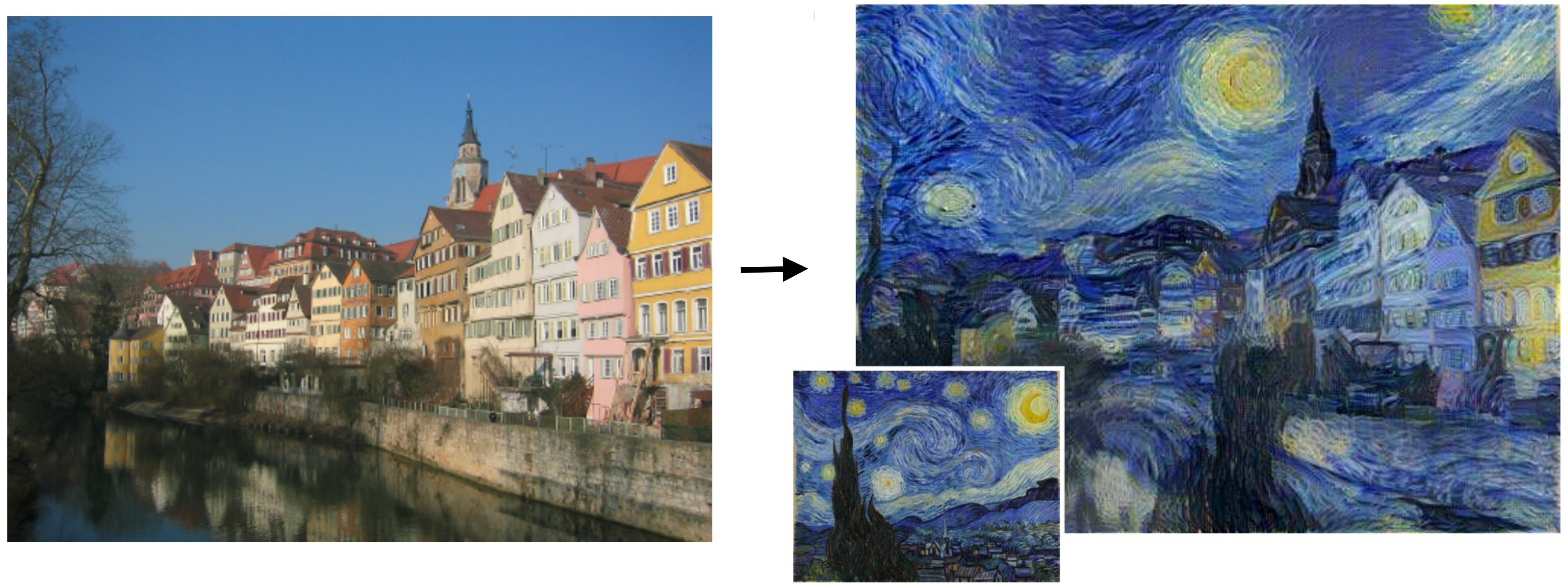


[Xie et al. 2015]

Season change

[Laffont et al. 2014]

Artistic style transfer



[Gatys et al. 2016]

Paired Image-to-Image Translation

Training data

$\mathbf{x}$ $\mathbf{y}$

⋮

Input $\mathbf{x}$

$\mathcal{F}$

Output $\mathbf{y}$

$$\arg\min_{\mathcal{F}} \mathbb{E}_{\mathbf{x},\mathbf{y}}[L(\mathcal{F}(\mathbf{x}), \mathbf{y})]$$

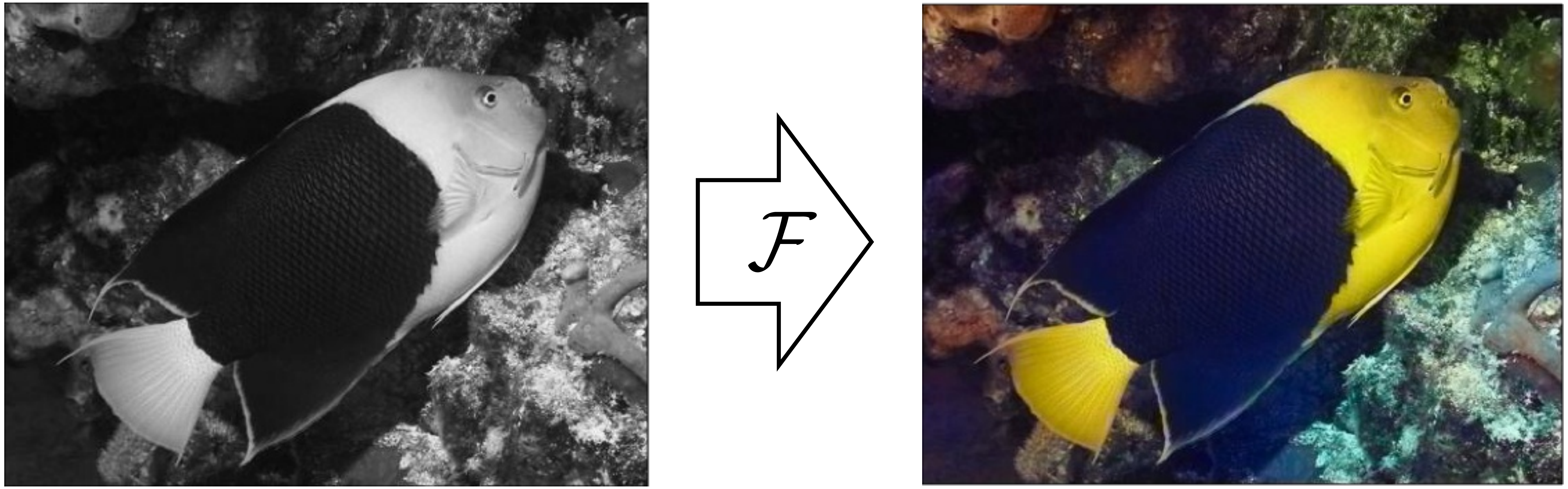
Objective function (loss)

Neural Network

[Zhang et al., ECCV 2016]

Paired Image-to-Image Translation

Input $\mathbf{x}$ Output $\mathbf{y}$



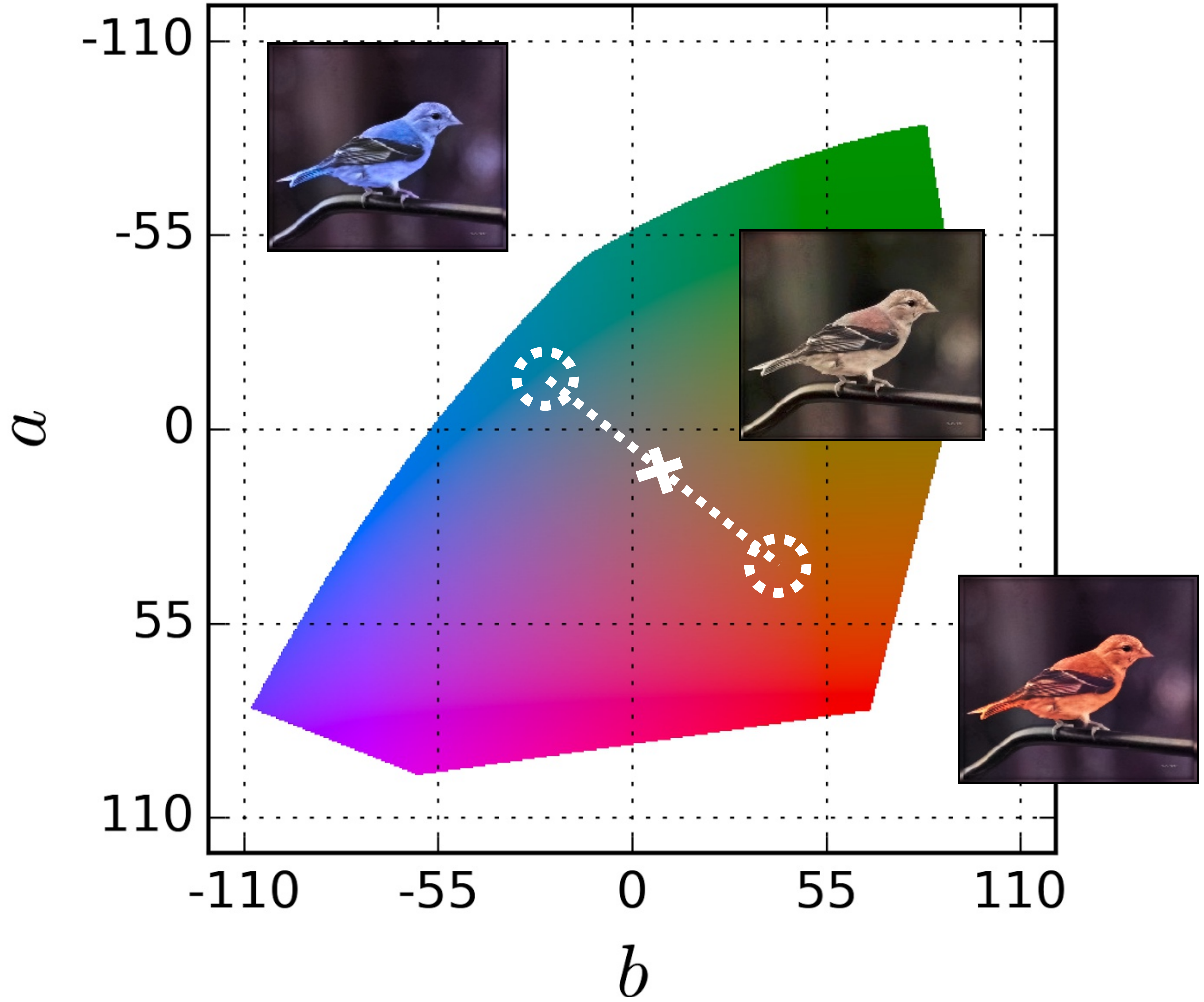
$$\arg\min_{\mathcal{F}} \mathbb{E}_{\mathbf{x},\mathbf{y}}[L(\mathcal{F}(\mathbf{x}),\mathbf{y})]$$

“**What** should I do” “**How** should I do it?”

Designing loss functions

Input Output Ground truth

$$L_2(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \widehat{\mathbf{Y}}_{h,w}\|_2^2$$



$$\mathrm{L}_2(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \widehat{\mathbf{Y}}_{h,w}\|_2^2$$

Designing loss functions

Input

Zhang et al. 2016

Ground truth

Color distribution cross-entropy loss with colorfulness enhancing term.

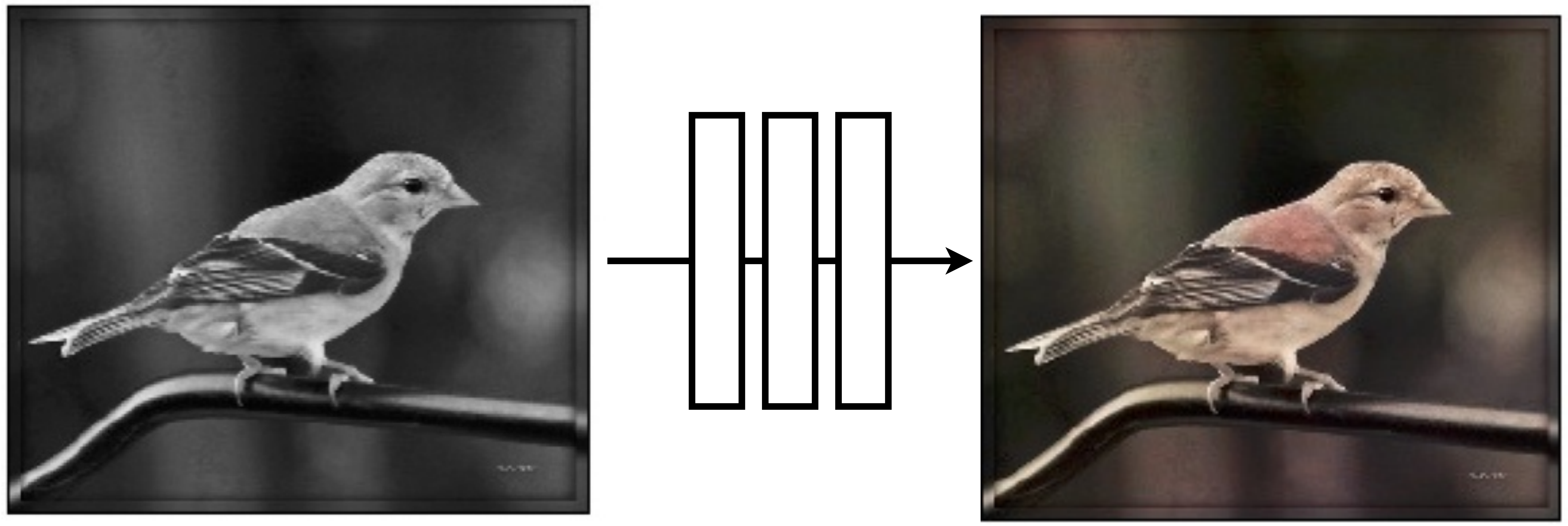

DADS
DADS

Designing loss functions

Be careful what you wish for!

Designing loss functions

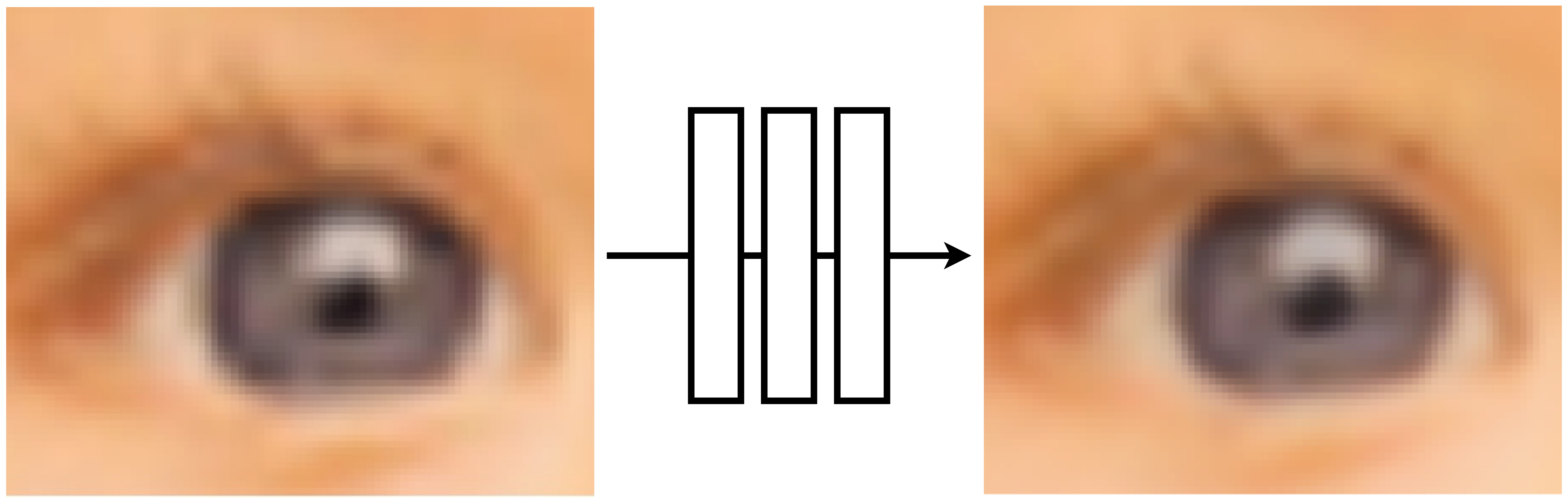
Image colorization



[Zhang, Isola, Efros, ECCV 2016]

L2 regression

Super-resolution

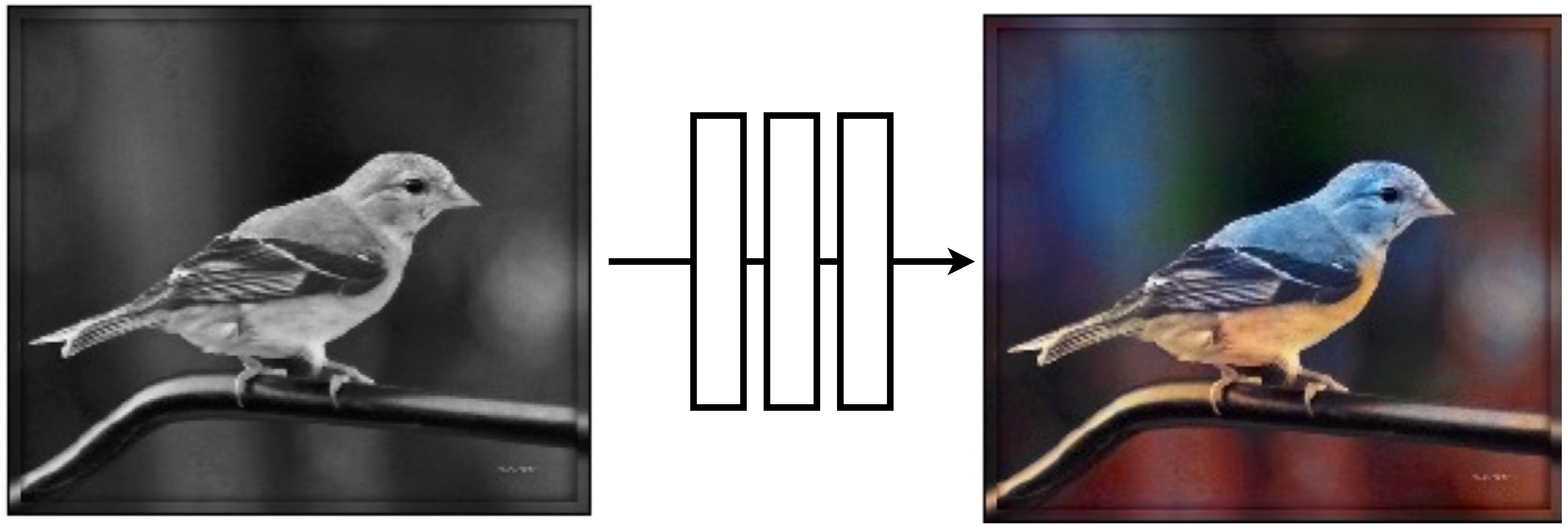


[Johnson, Alahi, Li, ECCV 2016]

L2 regression

Designing loss functions

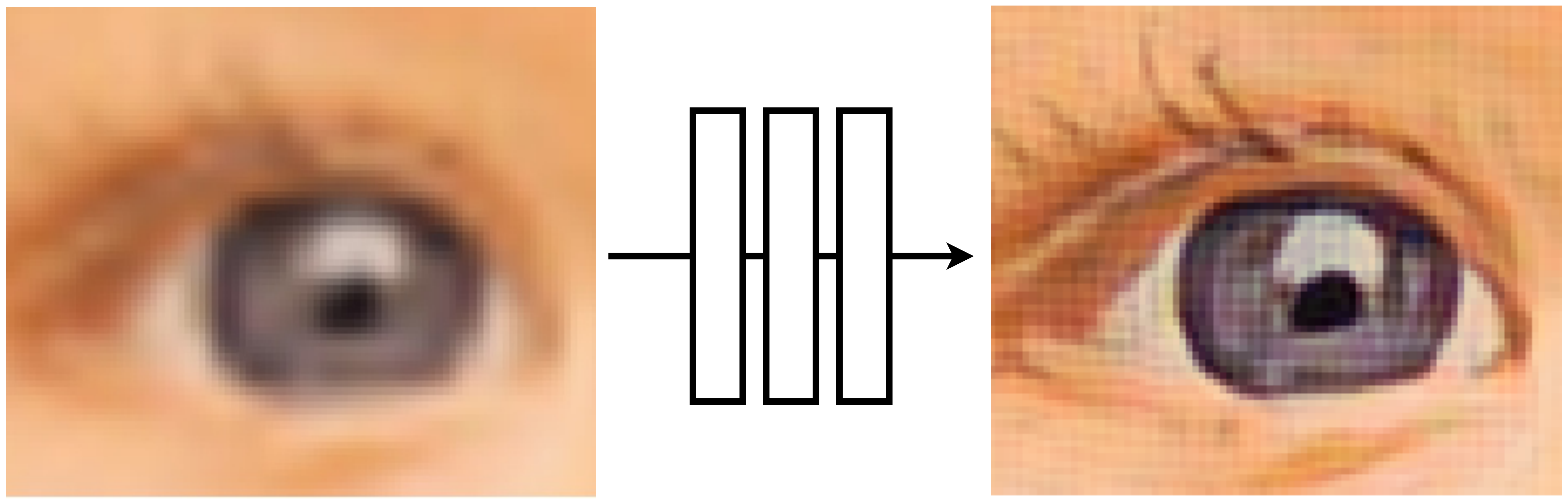
Image colorization



[Zhang, Isola, Efros, ECCV 2016]

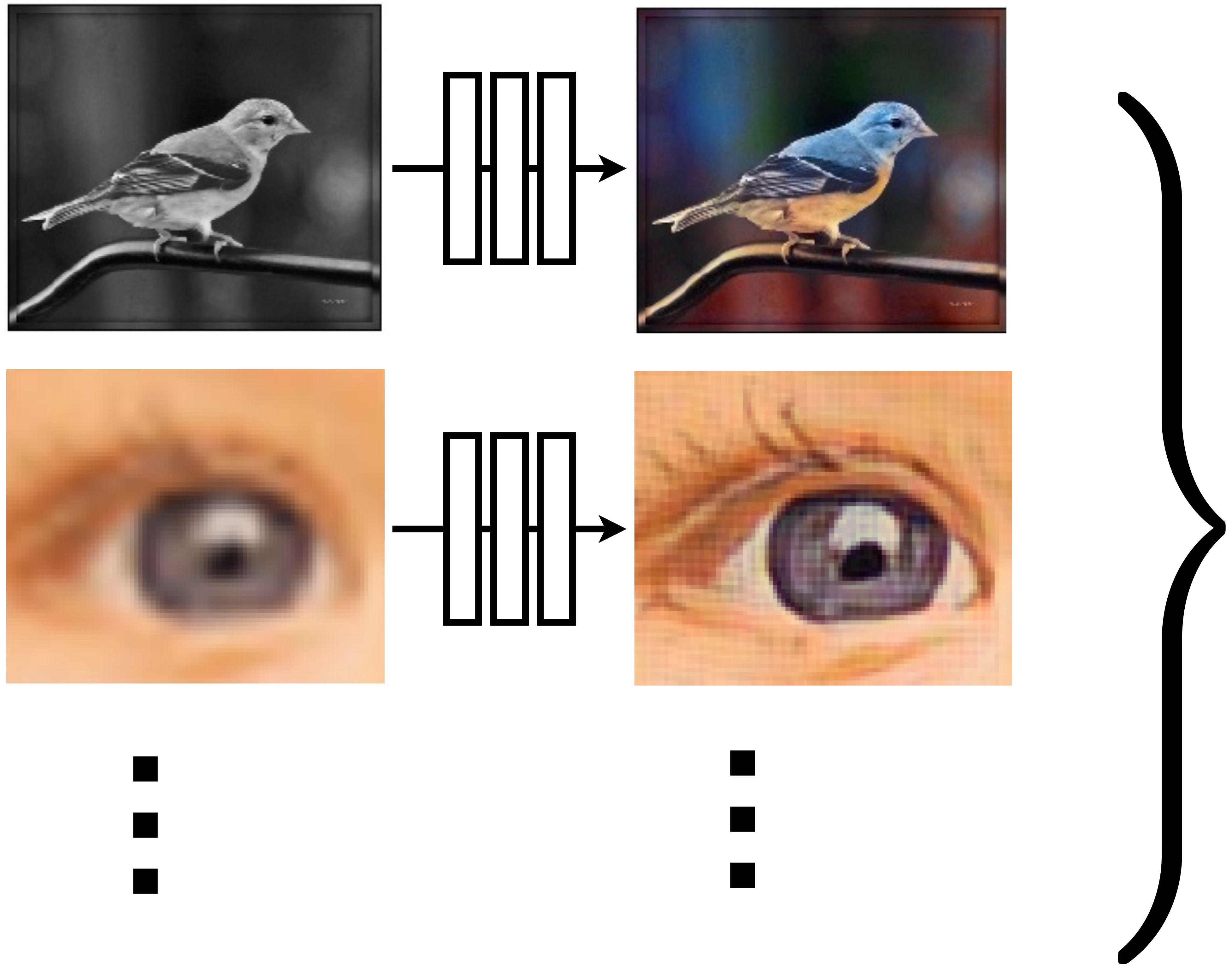
Cross entropy objective, with colorfulness term

Super-resolution



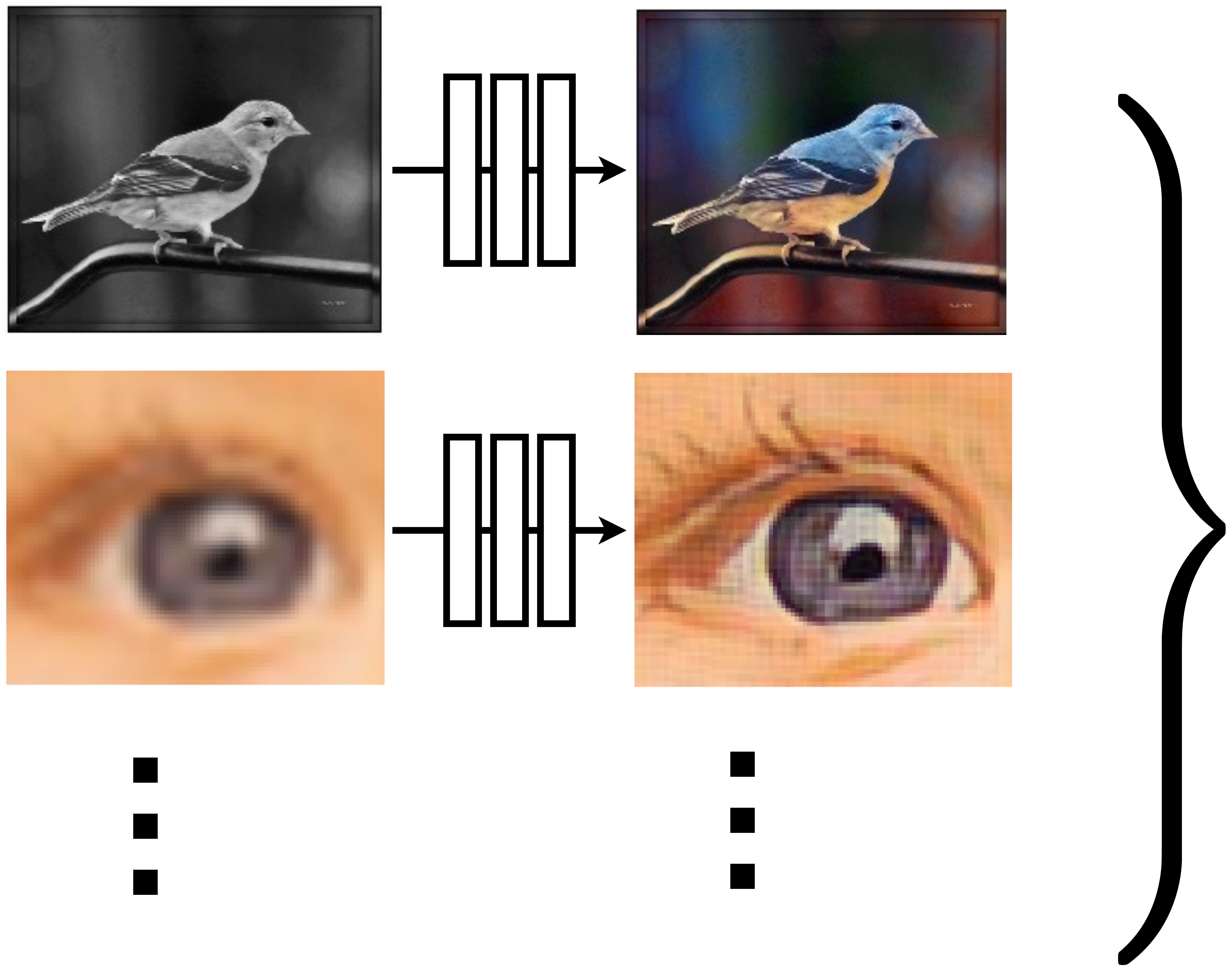
[Johnson, Alahi, Li, ECCV 2016]

Deep feature covariance matching objective

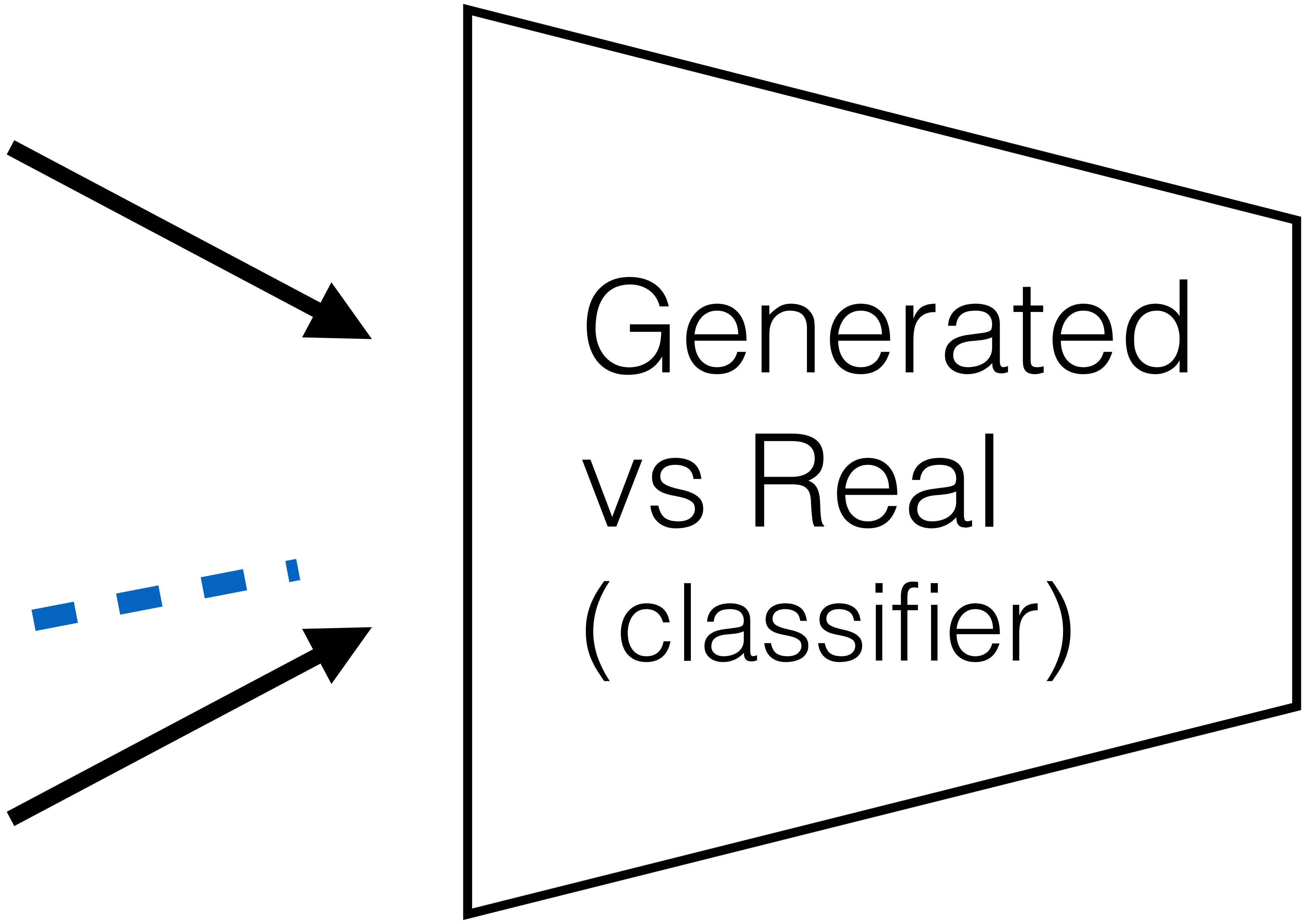


Universal loss?

Generated images



“Generative Adversarial Network” (GANs)

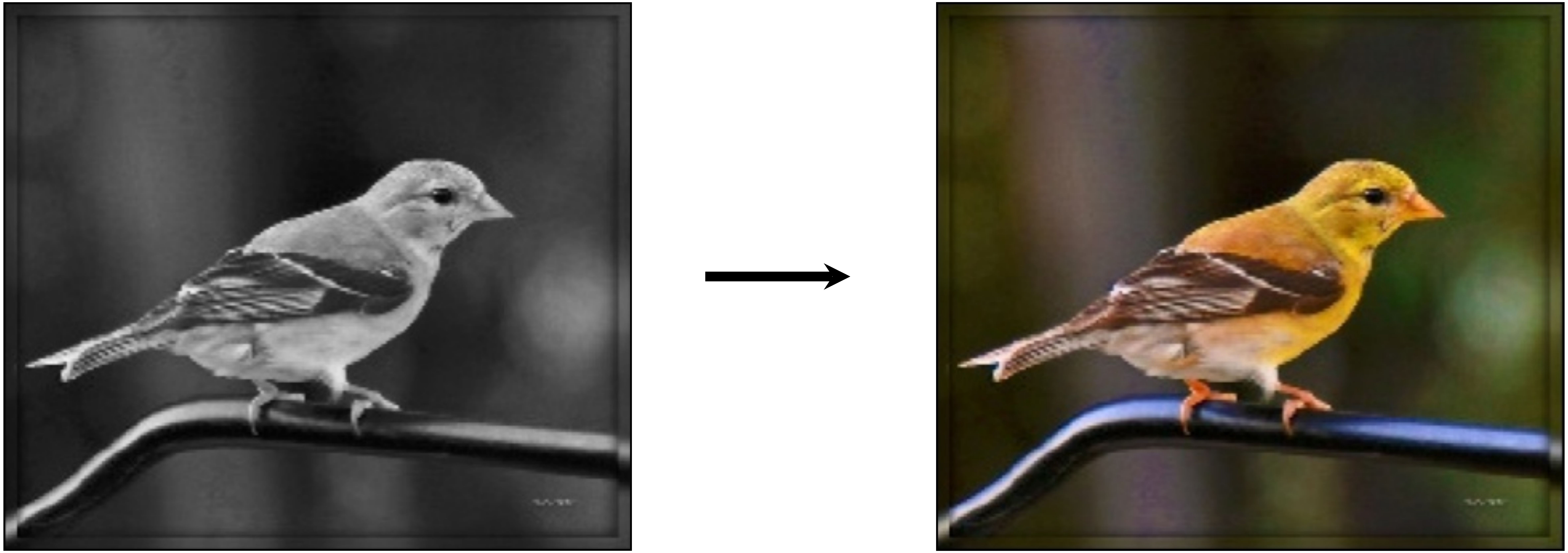


Real photos

...

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

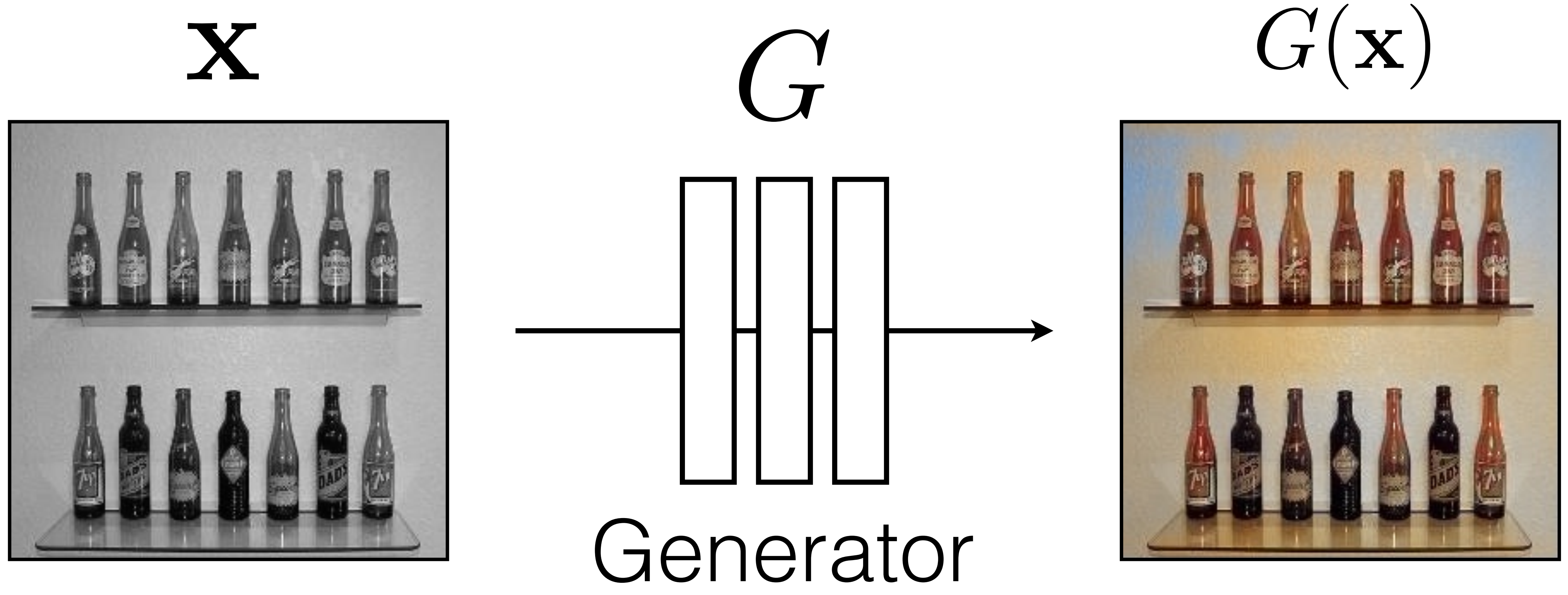
Conditional GANs

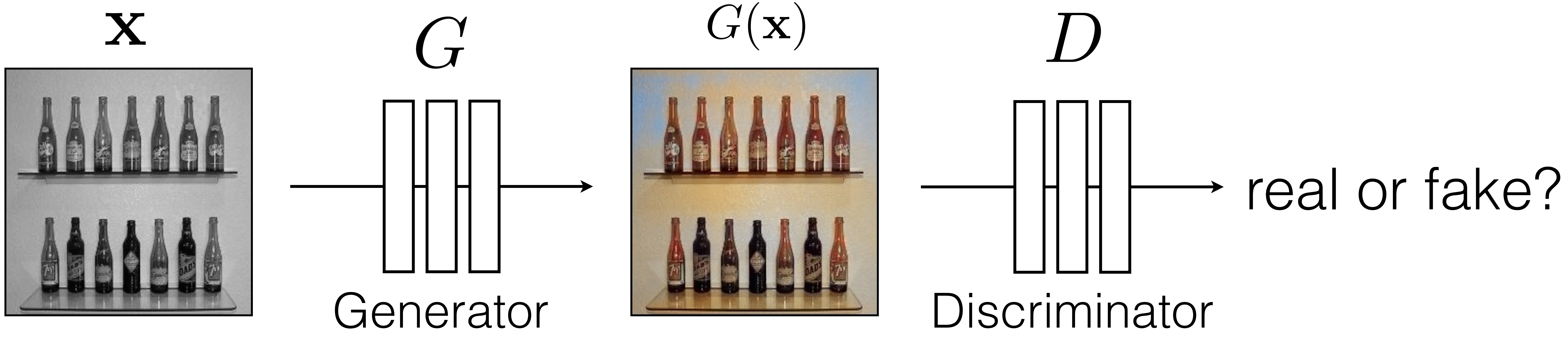


[Mirza et al. 2014] [Reed et al. 2016]

[Ledig et al. 2017] [Isola et al. 2017]

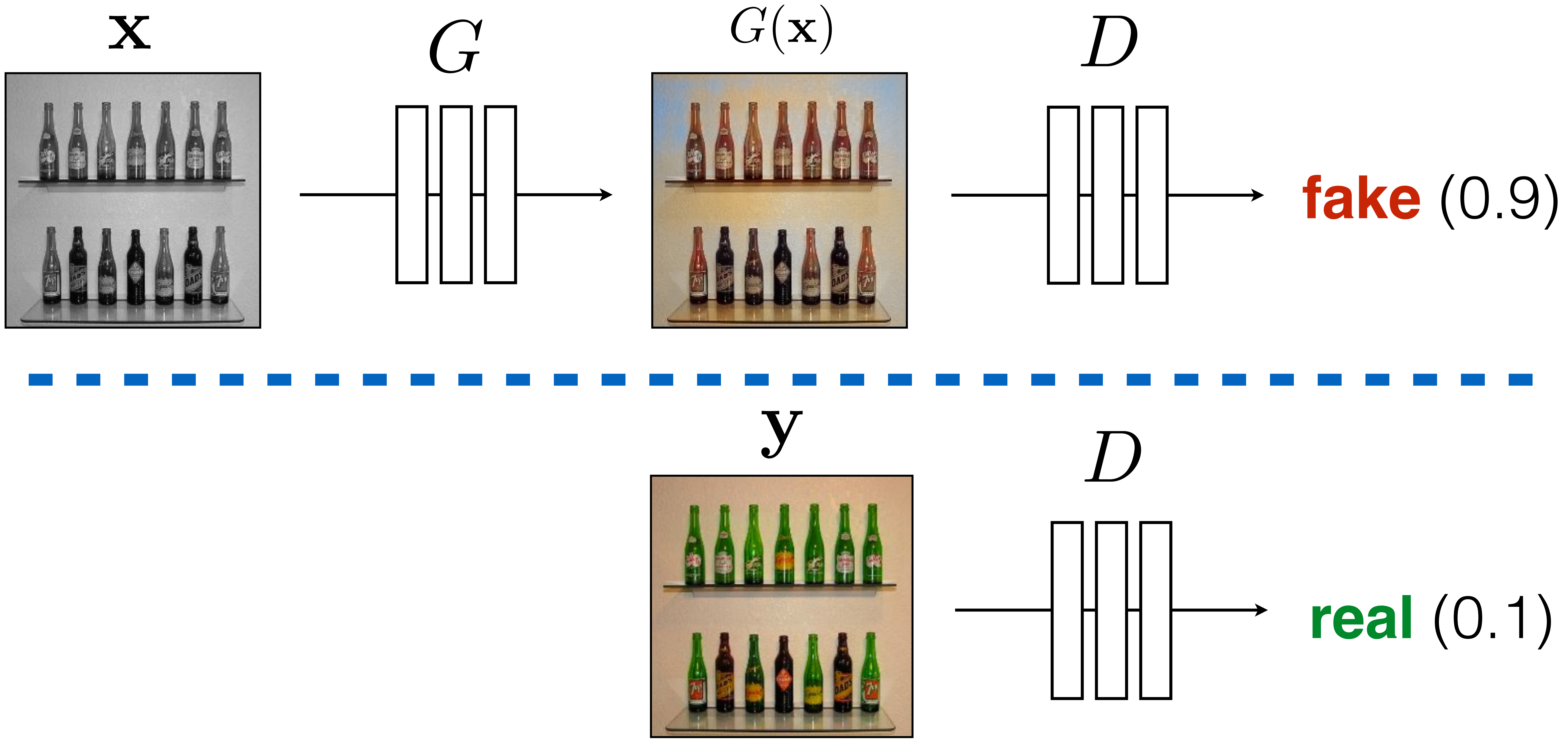
[…]


$\mathbf{x}$
G
Generator
$G(\mathbf{x})$

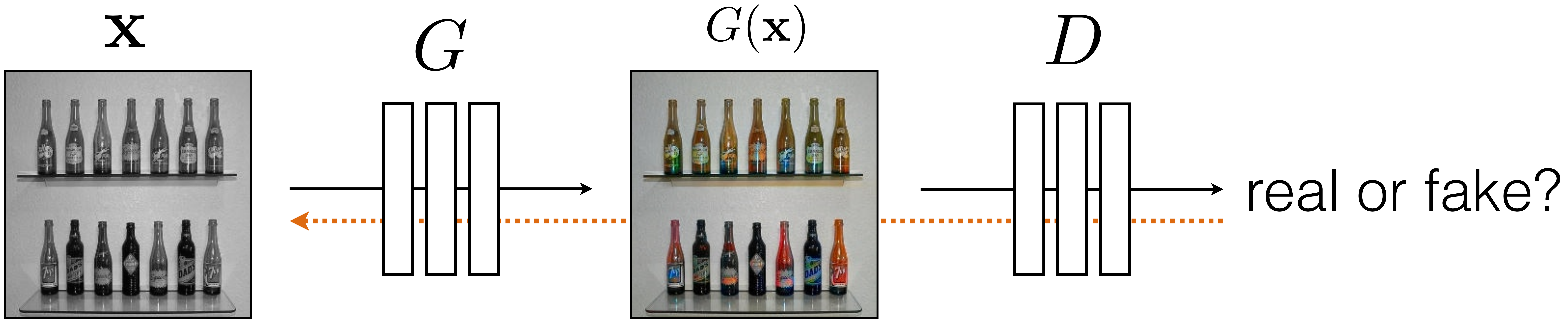


G tries to synthesize fake images that fool **D**

D tries to identify the fakes

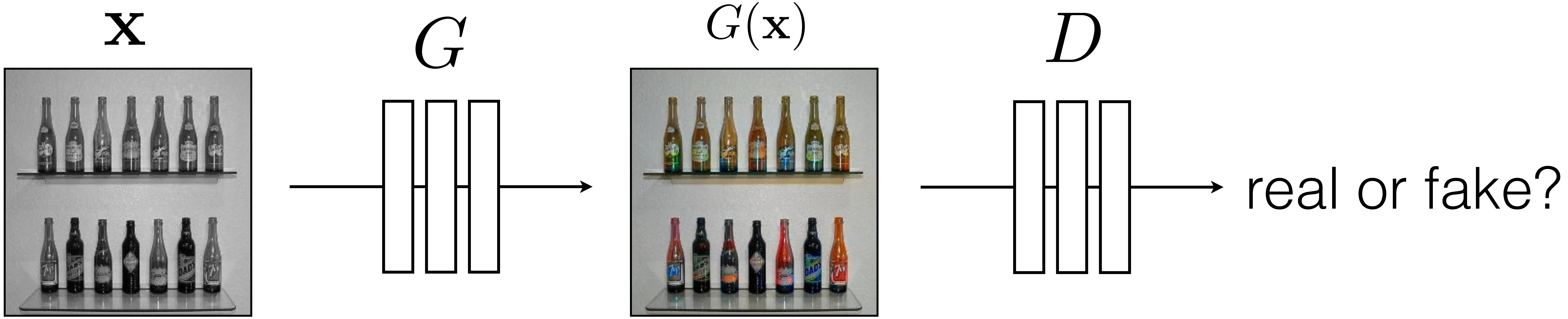


$$\arg\max_{D} \ \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\ \boxed{\log D(G(\mathbf{x}))} \ + \ \boxed{\log(1 - D(\mathbf{y}))} \ \right]$$



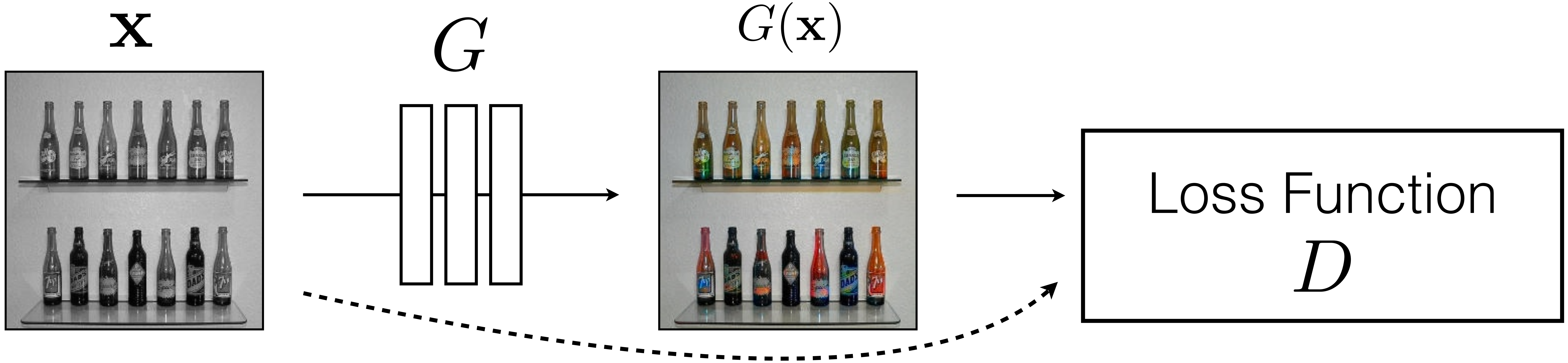
G tries to synthesize fake images that ***fool*** **D**:

$$\arg\boxed{\min_{G}}\ \mathbb{E}_{\mathbf{x},\mathbf{y}}[\ \log D(G(\mathbf{x})) \quad + \quad \log(1 - D(\mathbf{y}))\ \ ]$$



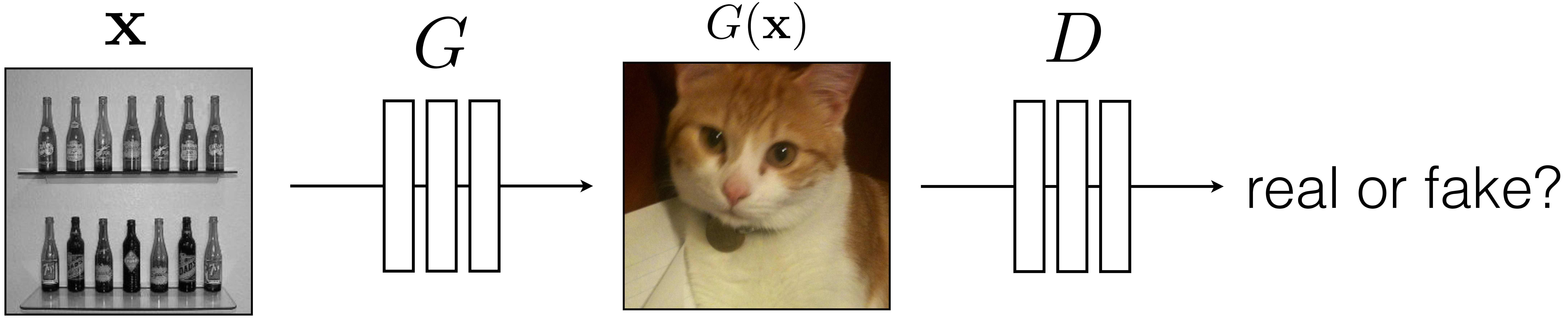
G tries to synthesize fake images that ***fool*** the ***best*** **D**:

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

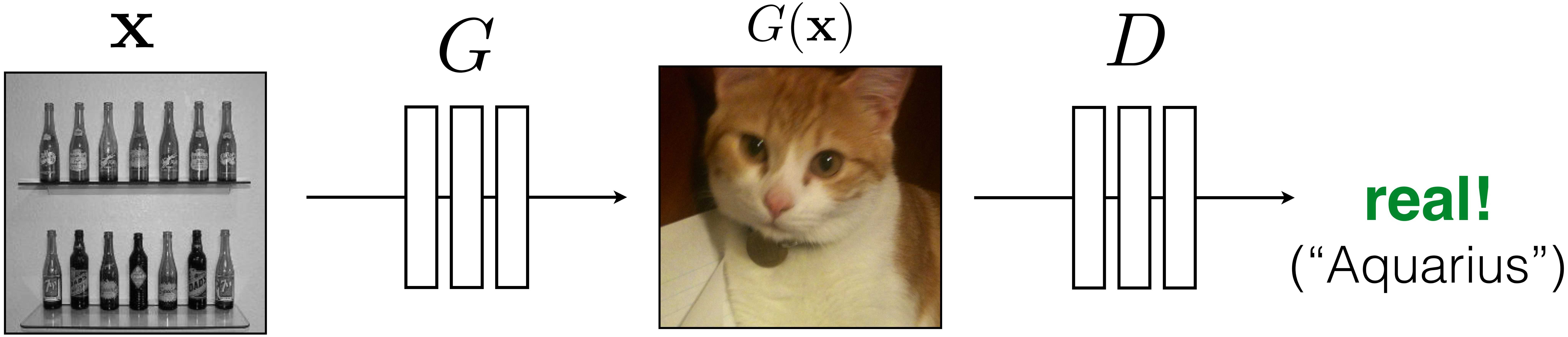


G's perspective: **D** is a loss function.

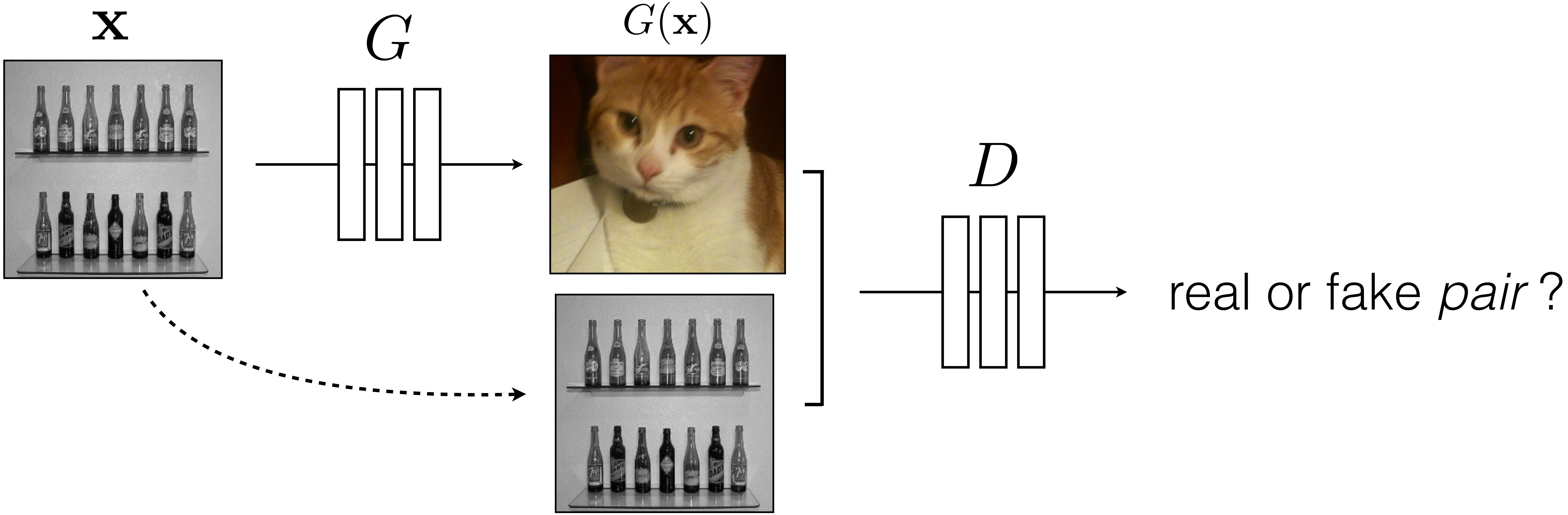
Rather than being hand-designed, it is *learned*.



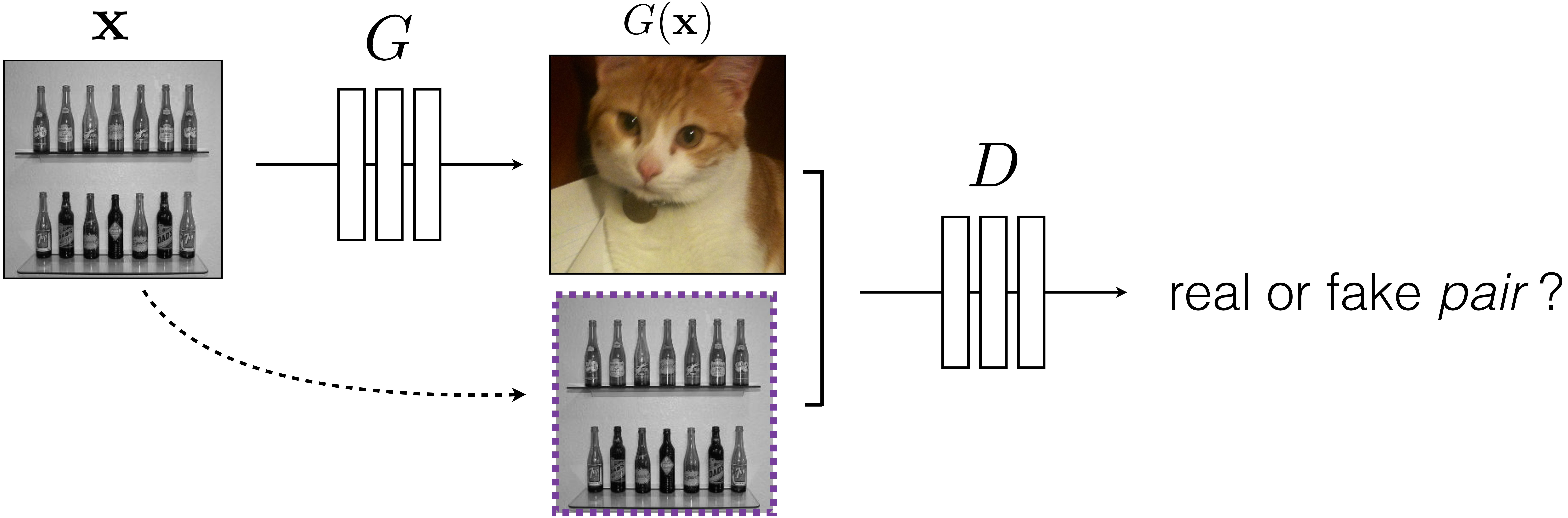
$$\arg\min_{G} \max_{D} \ \mathbb{E}_{\mathbf{x},\mathbf{y}}[\ \log D(G(\mathbf{x})) \quad + \quad \log(1 - D(\mathbf{y})) \ ]$$



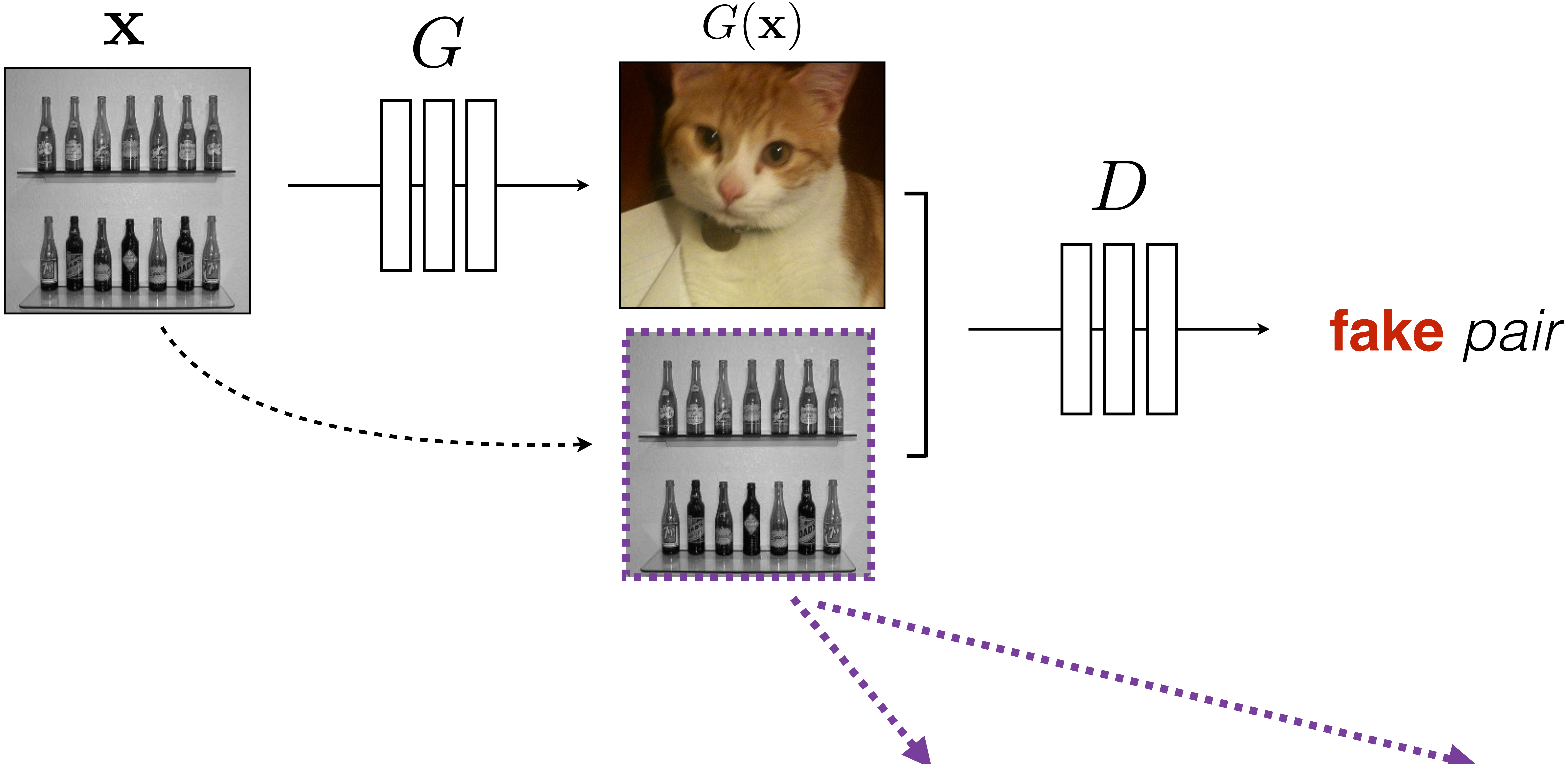
$$\arg\min_G \max_D \; \mathbb{E}_{\mathbf{x},\mathbf{y}}[\; \log D(G(\mathbf{x})) \quad + \quad \log(1 - D(\mathbf{y})) \;]$$



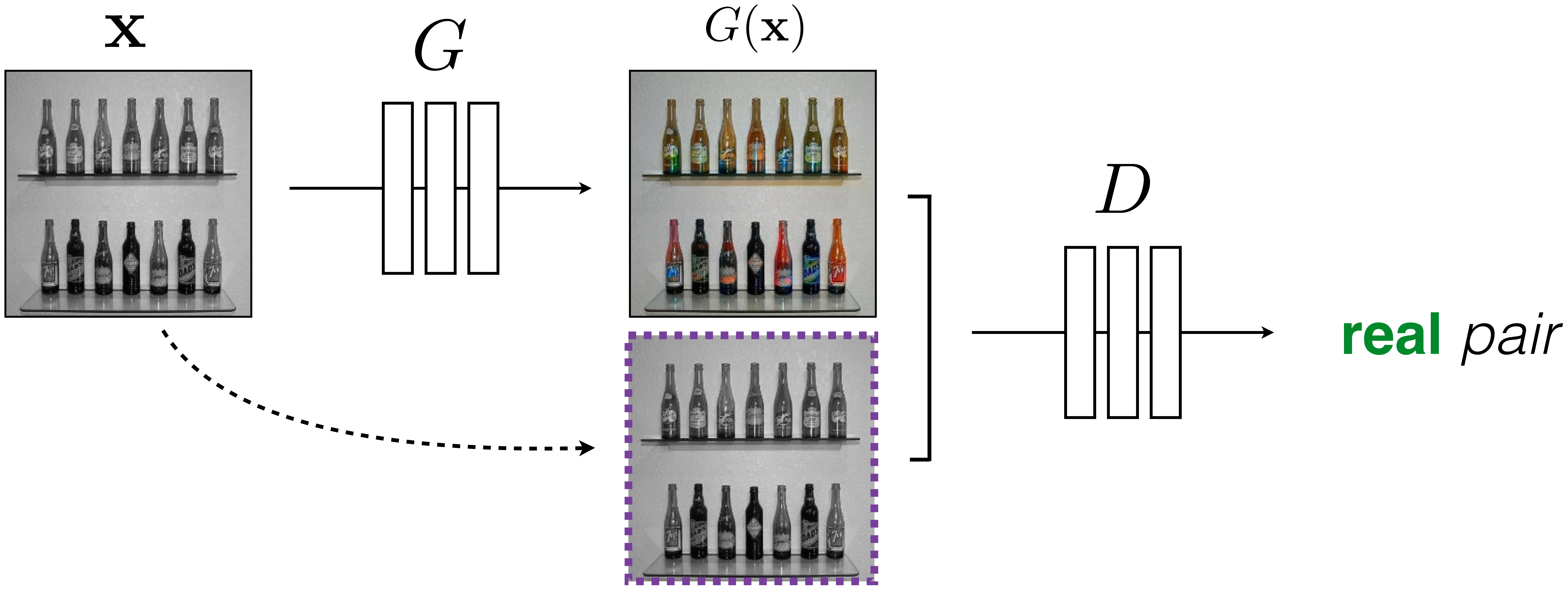
$$\arg\min_{G}\max_{D}\ \mathbb{E}_{\mathbf{x},\mathbf{y}}[\ \log D(G(\mathbf{x})) \quad + \quad \log(1 - D(\mathbf{y}))\ \ ]$$



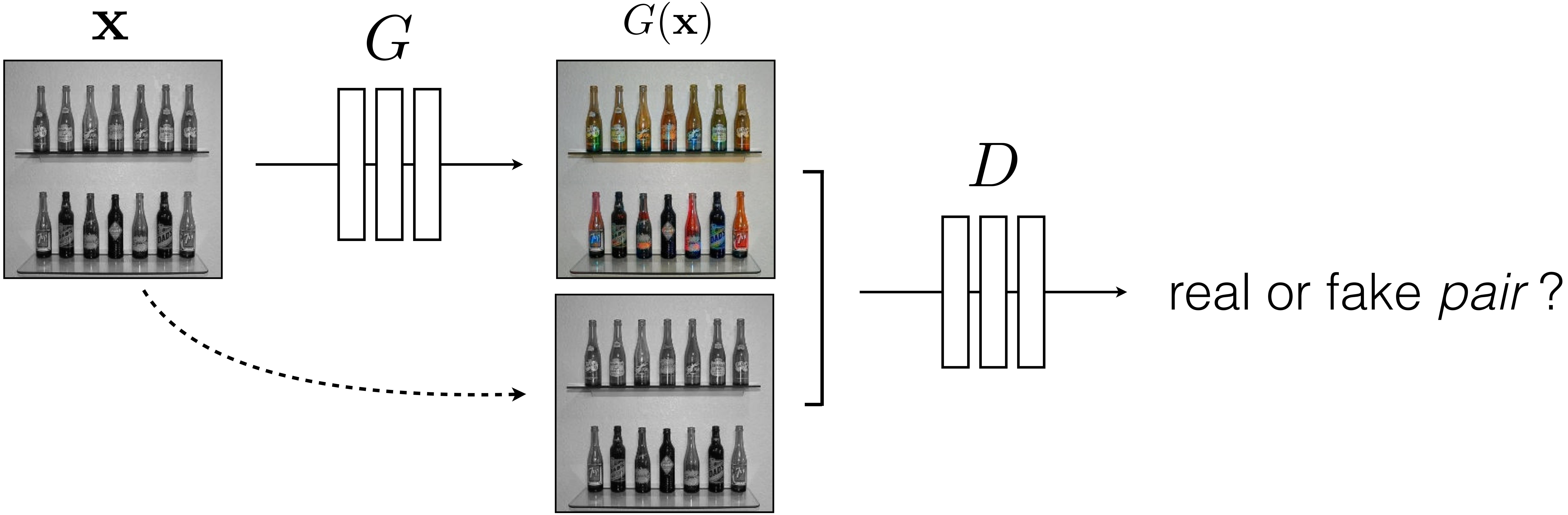
$$\arg\min_{G}\max_{D}\ \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\ \log D(\boxed{\mathbf{x}}, G(\mathbf{x})) + \log(1 - D(\boxed{\mathbf{x}}, \mathbf{y}))\ \right]$$



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



$$\arg\min_{G}\max_{D}\ \mathbb{E}_{\mathbf{x},\mathbf{y}}[\ \log D(\boxed{\mathbf{x}}, G(\mathbf{x})) + \log(1 - D(\boxed{\mathbf{x}}, \mathbf{y}))\ ]$$



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

Training Details: Loss function

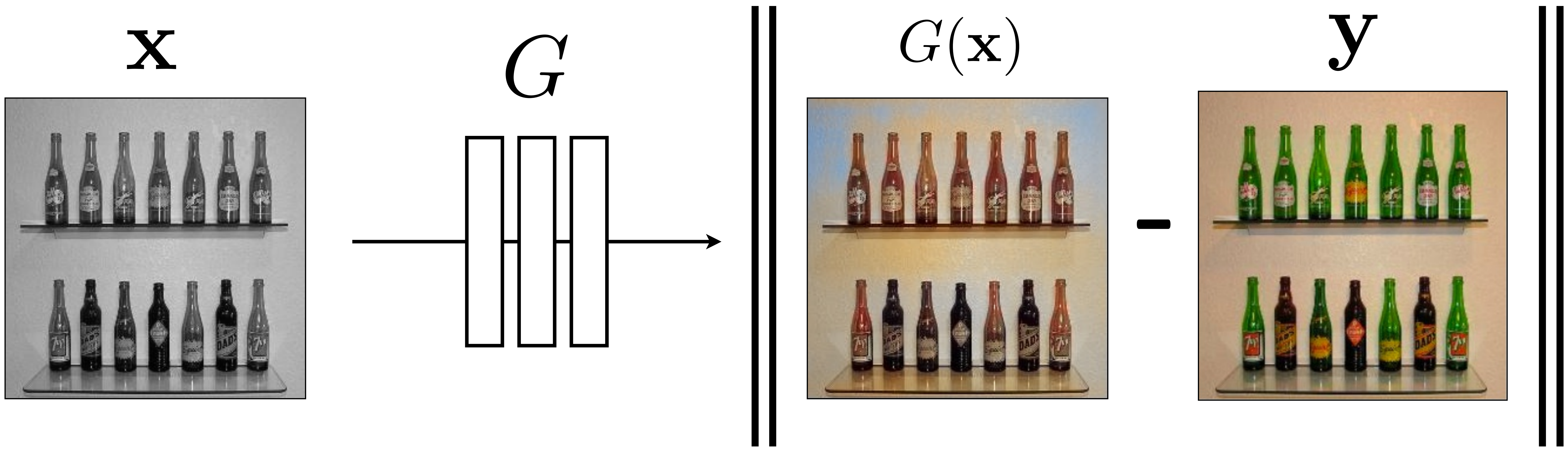
Conditional GAN

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda\mathcal{L}_{L1}(G).$$

Training Details: Loss function

Conditional GAN

$$G^* = \arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G).$$



Stable training + fast convergence

[c.f. Pathak et al. CVPR 2016]

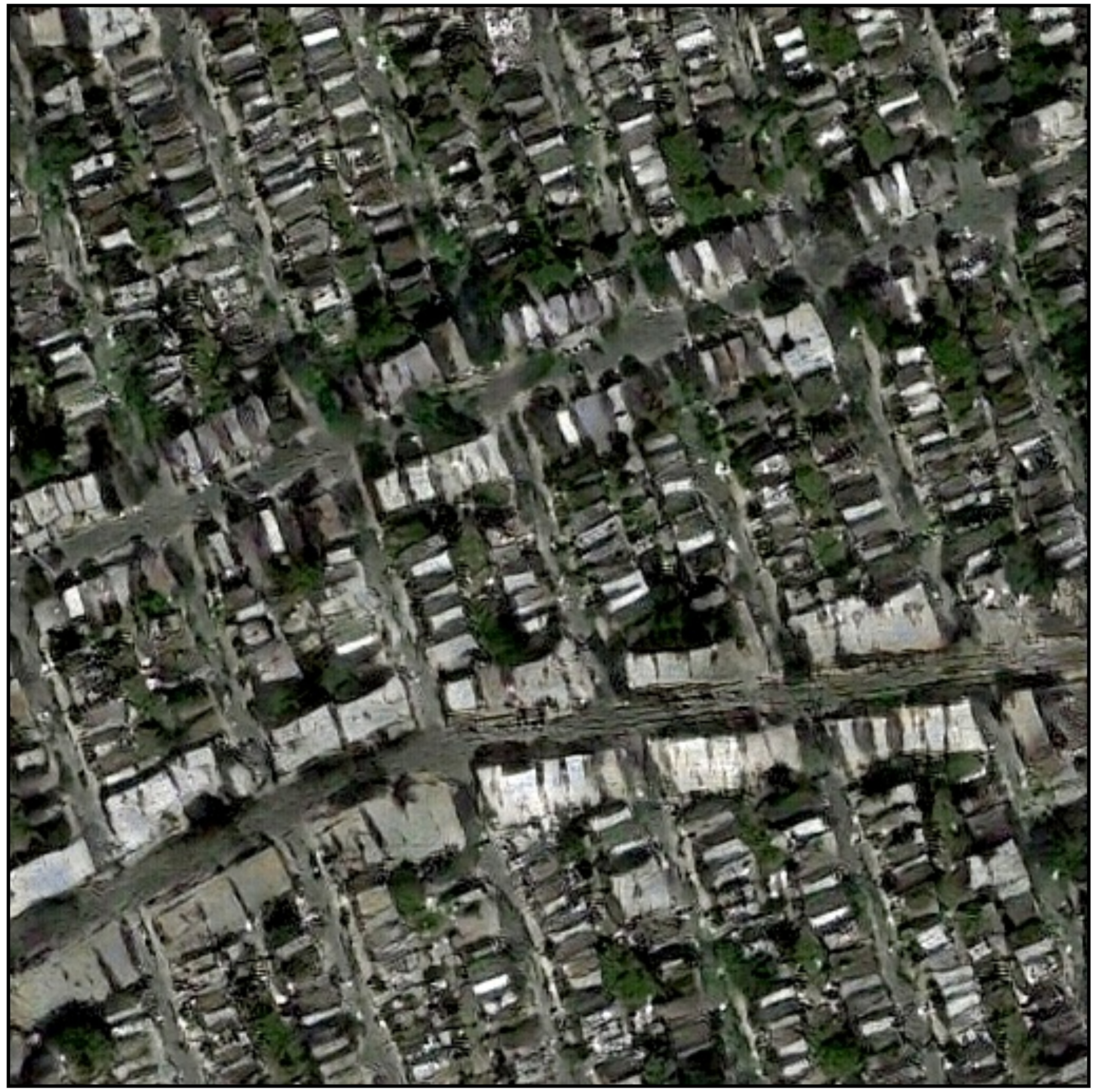
BW → Color

Input Output Input Output Input Output

Data from [Russakovsky et al. 2015]

Input

Output



Groundtruth

Data from
[maps.google.com]

Input

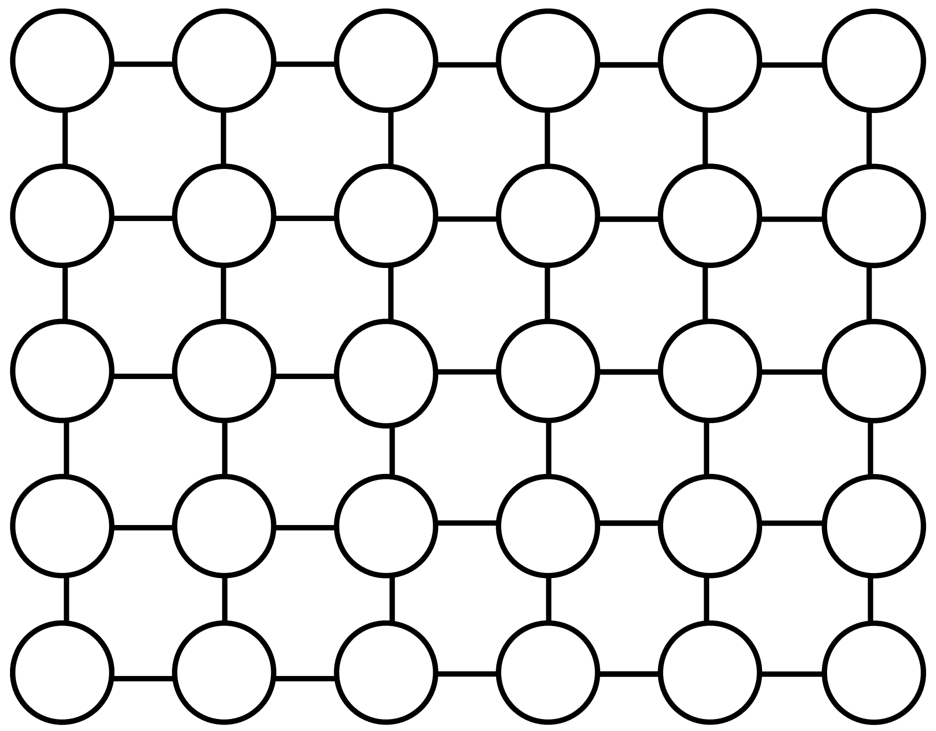
Output

Groundtruth

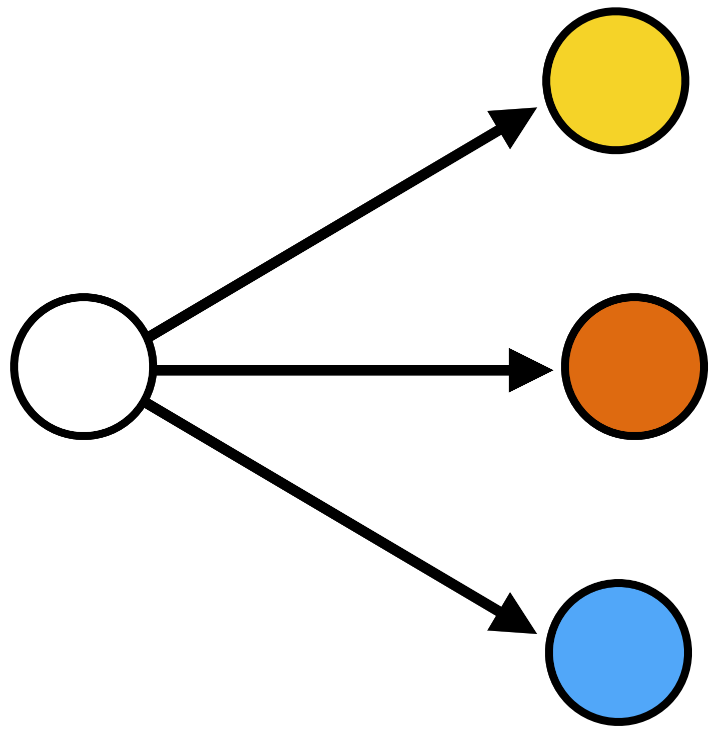
Data from [maps.google.com]

Challenges in image-to-image translation

1. Output is high-dimensional, structured object



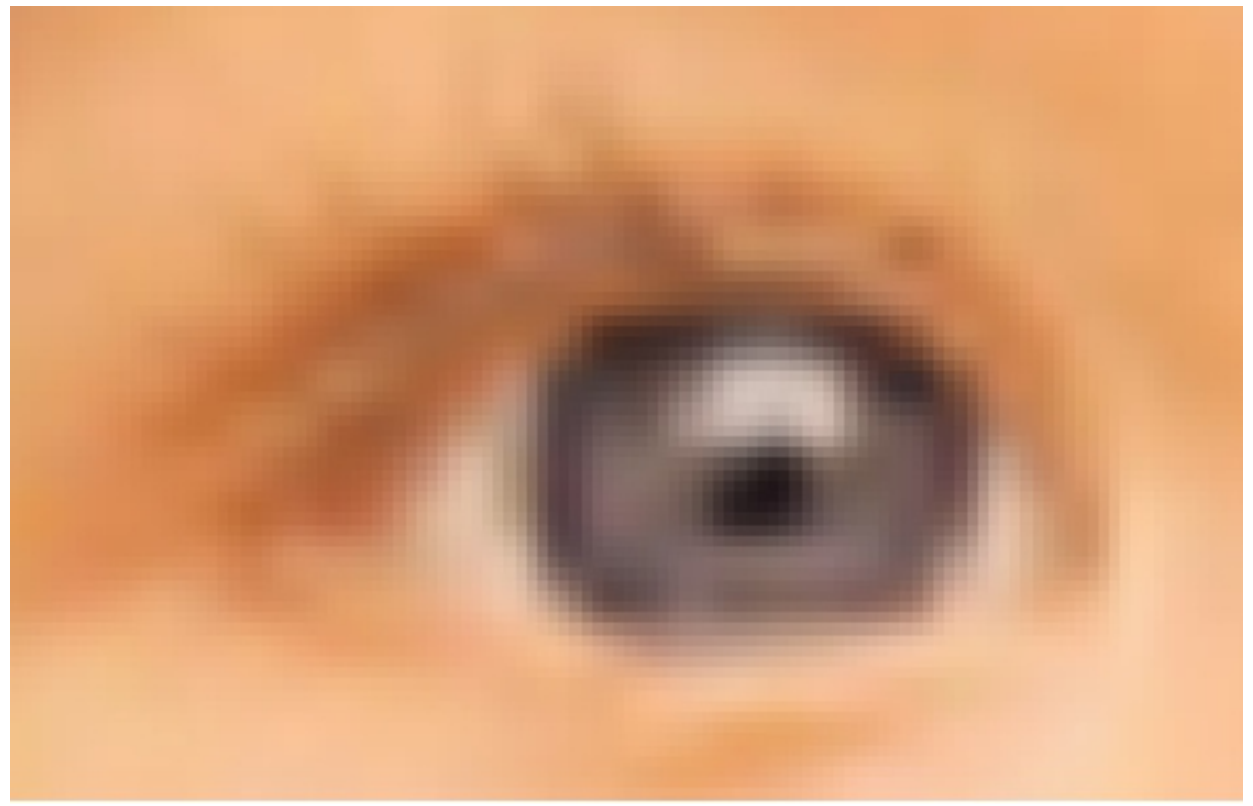
2. Uncertainty in mapping; many plausible outputs



Structured Prediction

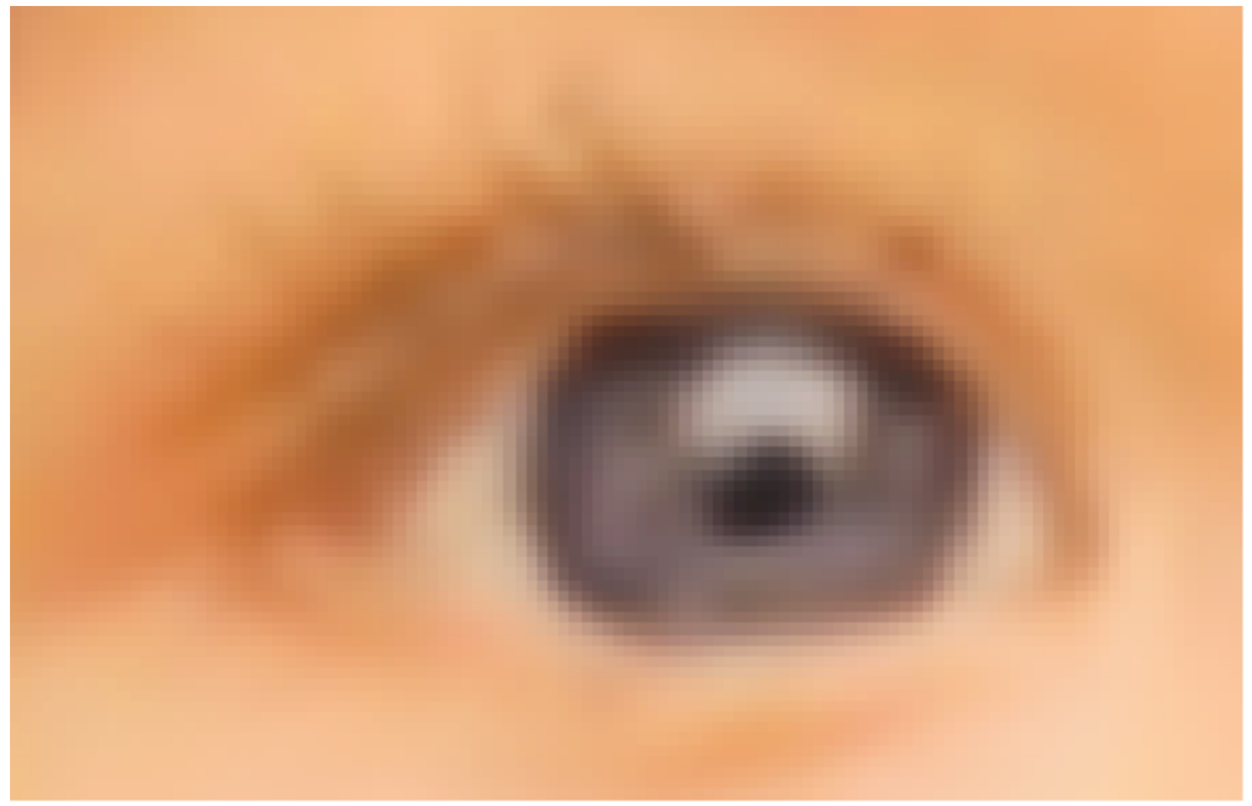
Input

$\mathbf{x}$



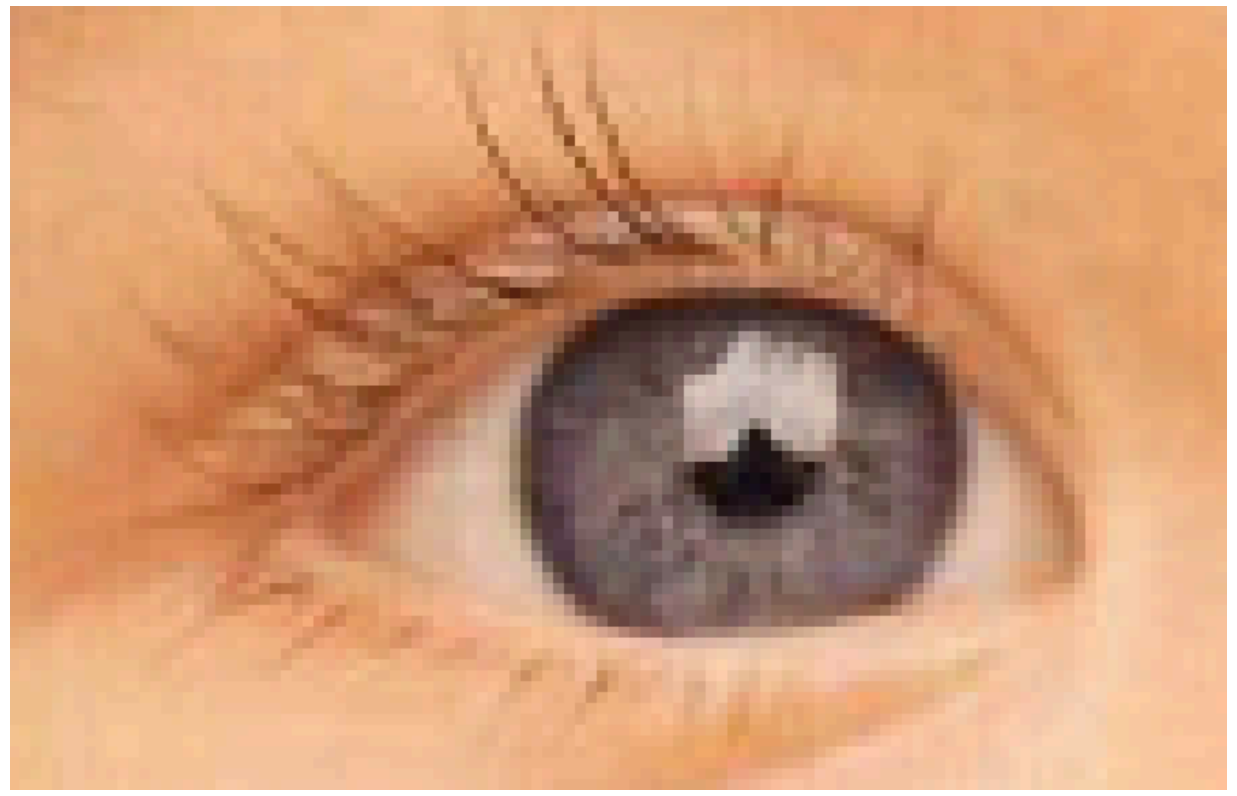
Output

$\hat{\mathbf{y}}$



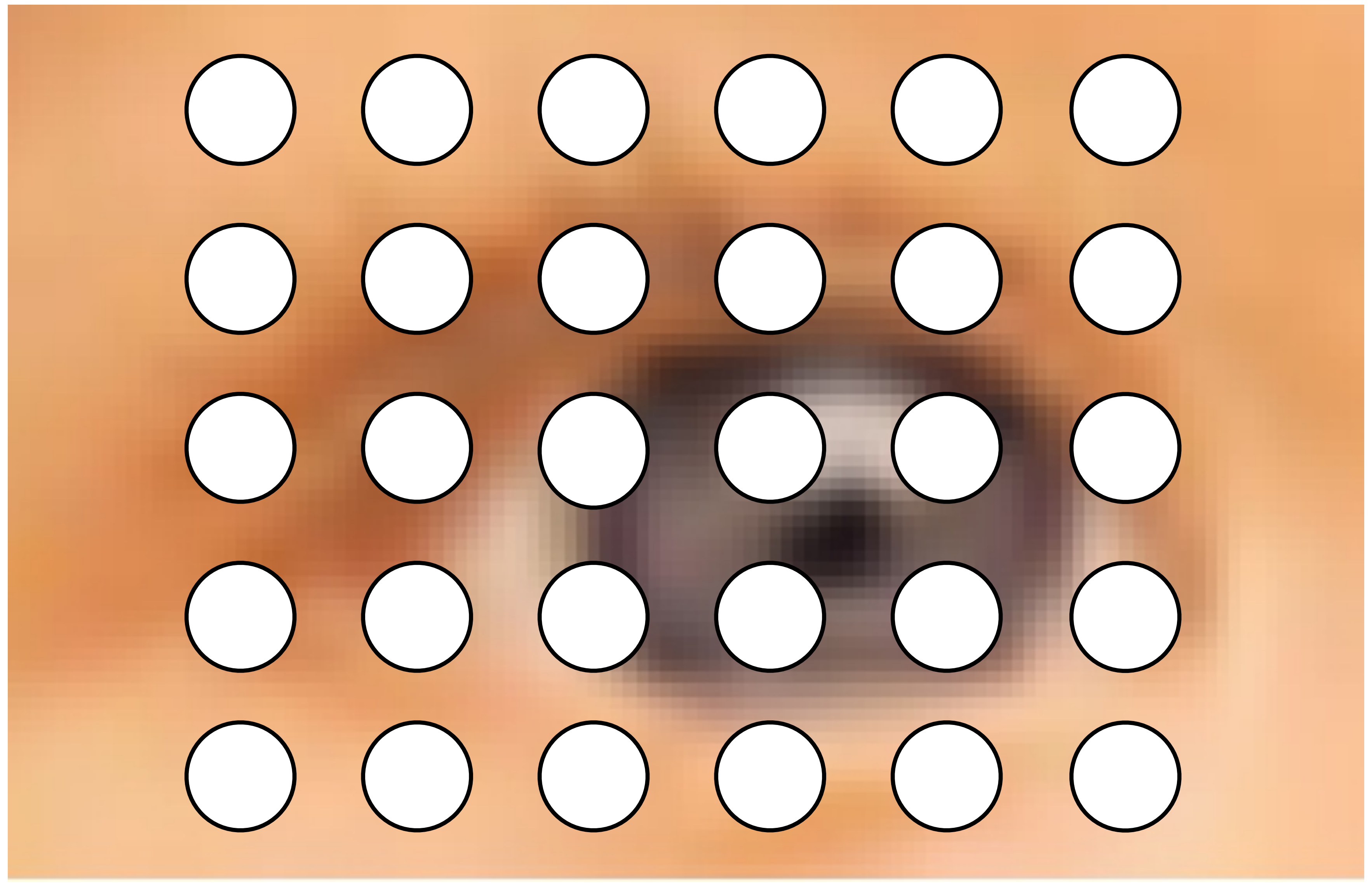
Target

$\mathbf{y}$



$$L(\hat{\mathbf{y}}, \mathbf{y}) = \|\hat{\mathbf{y}} - \mathbf{y}\|_2$$

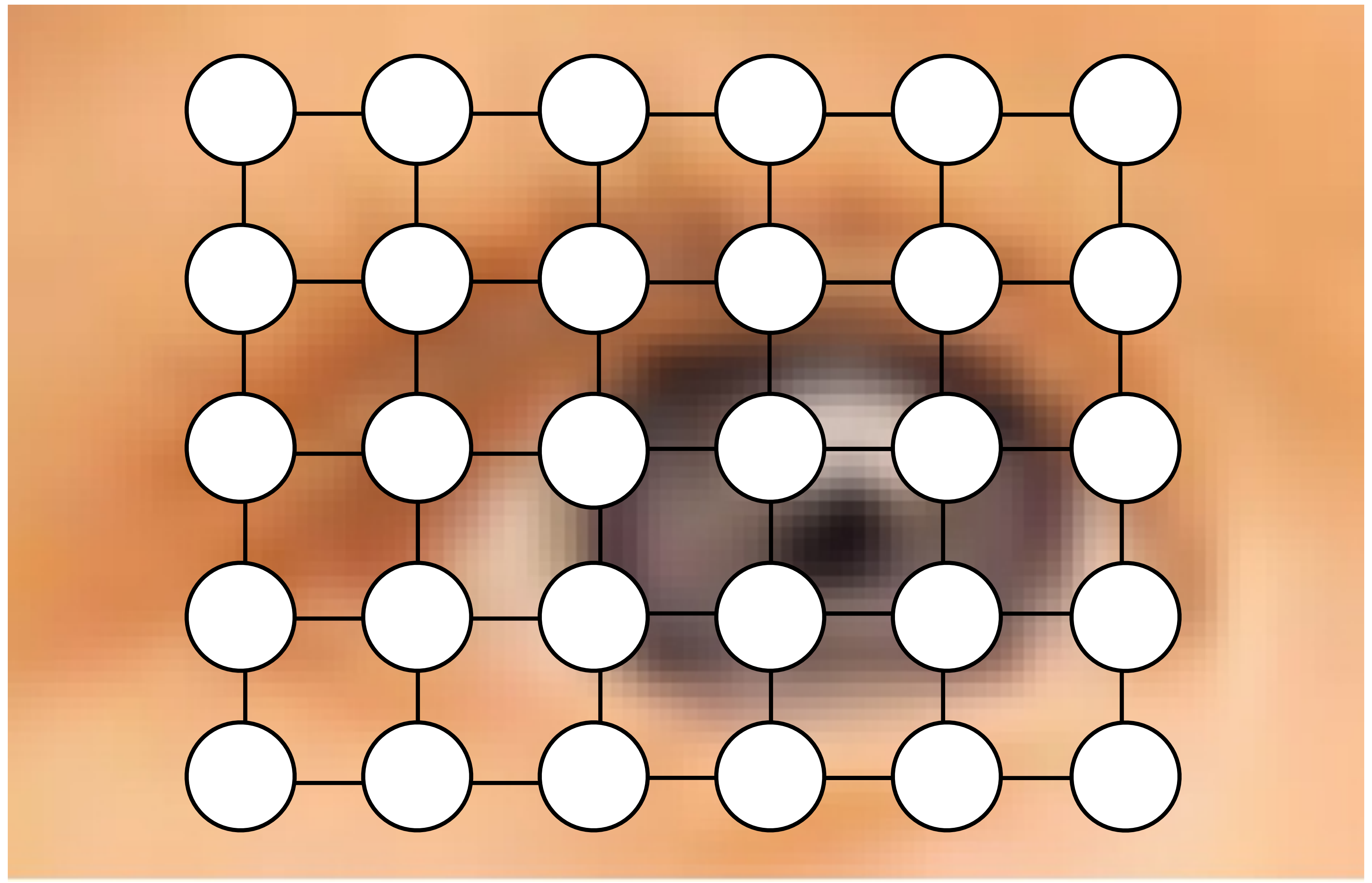
Structured Prediction



Each pixel treated as independent

$$\prod_i p(y_i|\mathbf{x})$$

CRF



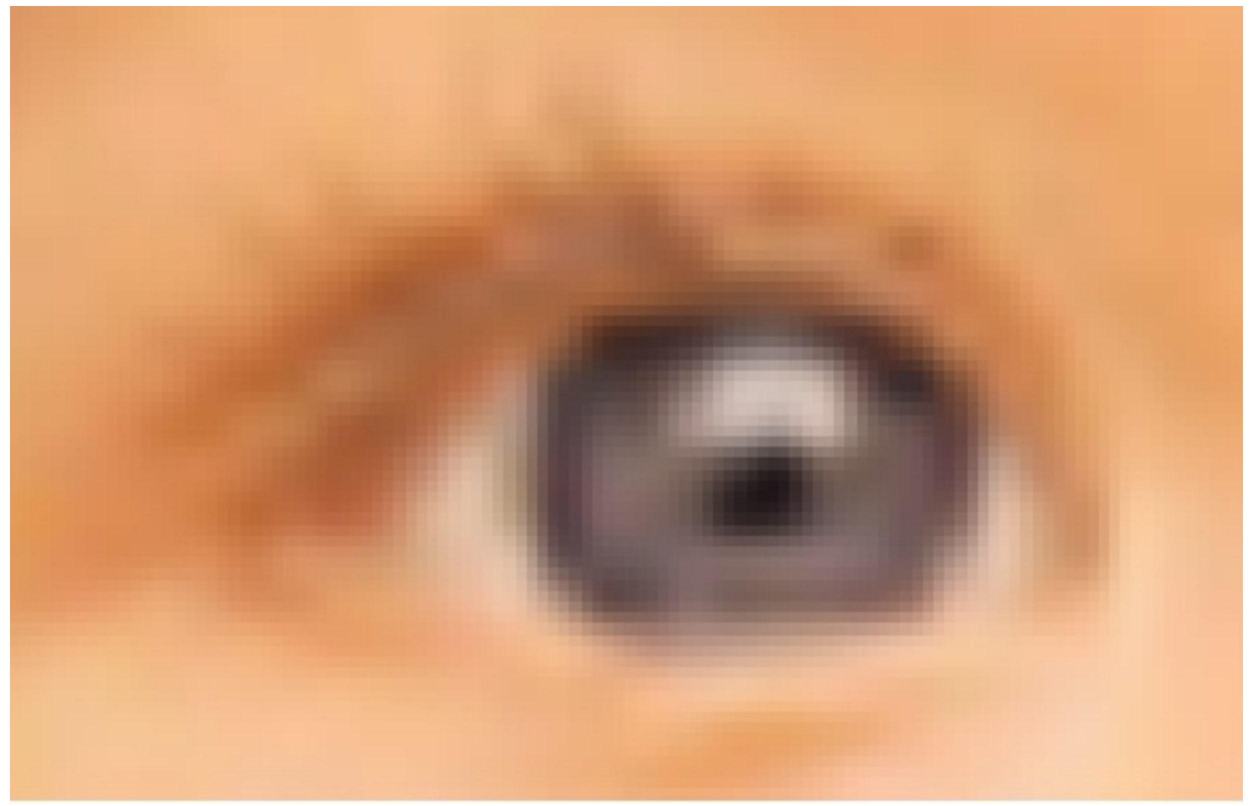
Models at pairwise configuration of pixels

$$\frac{1}{Z}\prod_{i,j} p(y_i, y_j|\mathbf{x})$$

"Perceptual Loss"

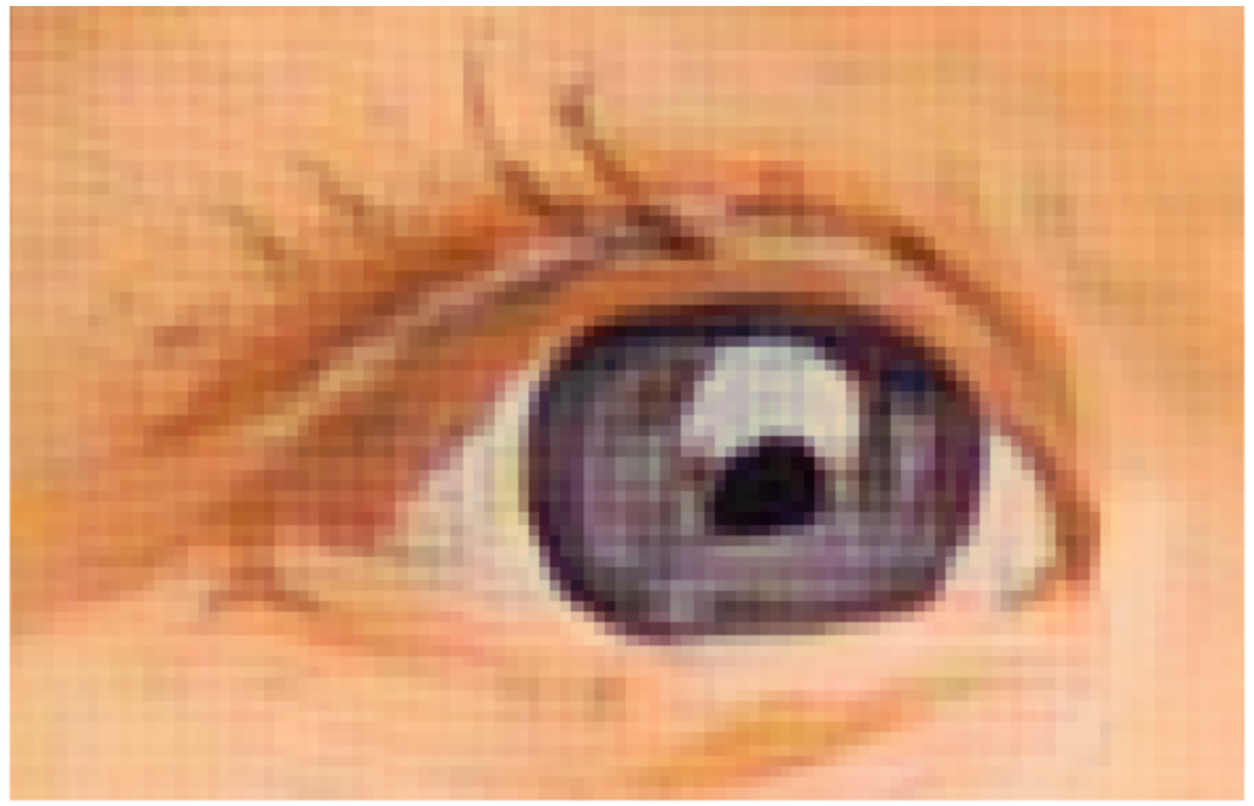
Input

$\mathbf{x}$



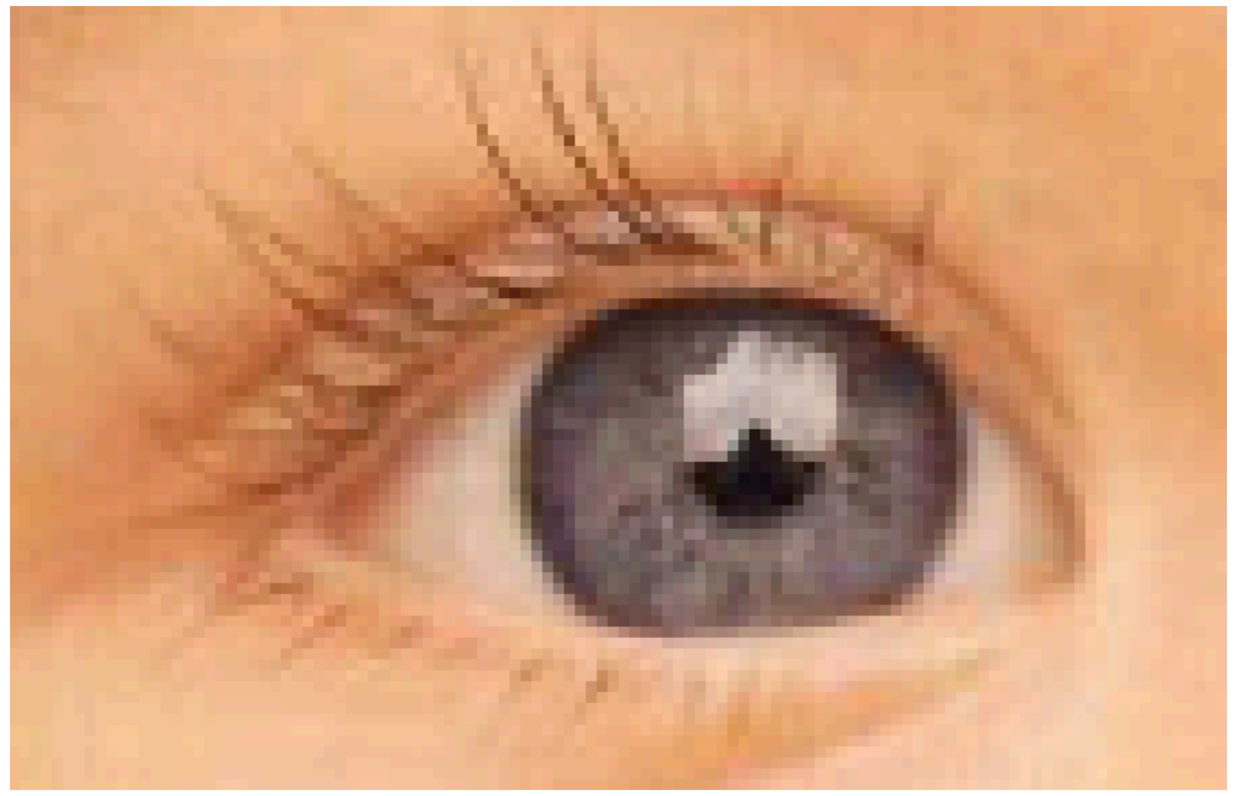
Output
[Johnson, Alahi, Li 2016]

$\hat{\mathbf{y}}$



Target

$\mathbf{y}$



$$L(\hat{\mathbf{y}}, \mathbf{y}) = \|\phi(\hat{\mathbf{y}}) - \phi(\mathbf{y})\|_2$$

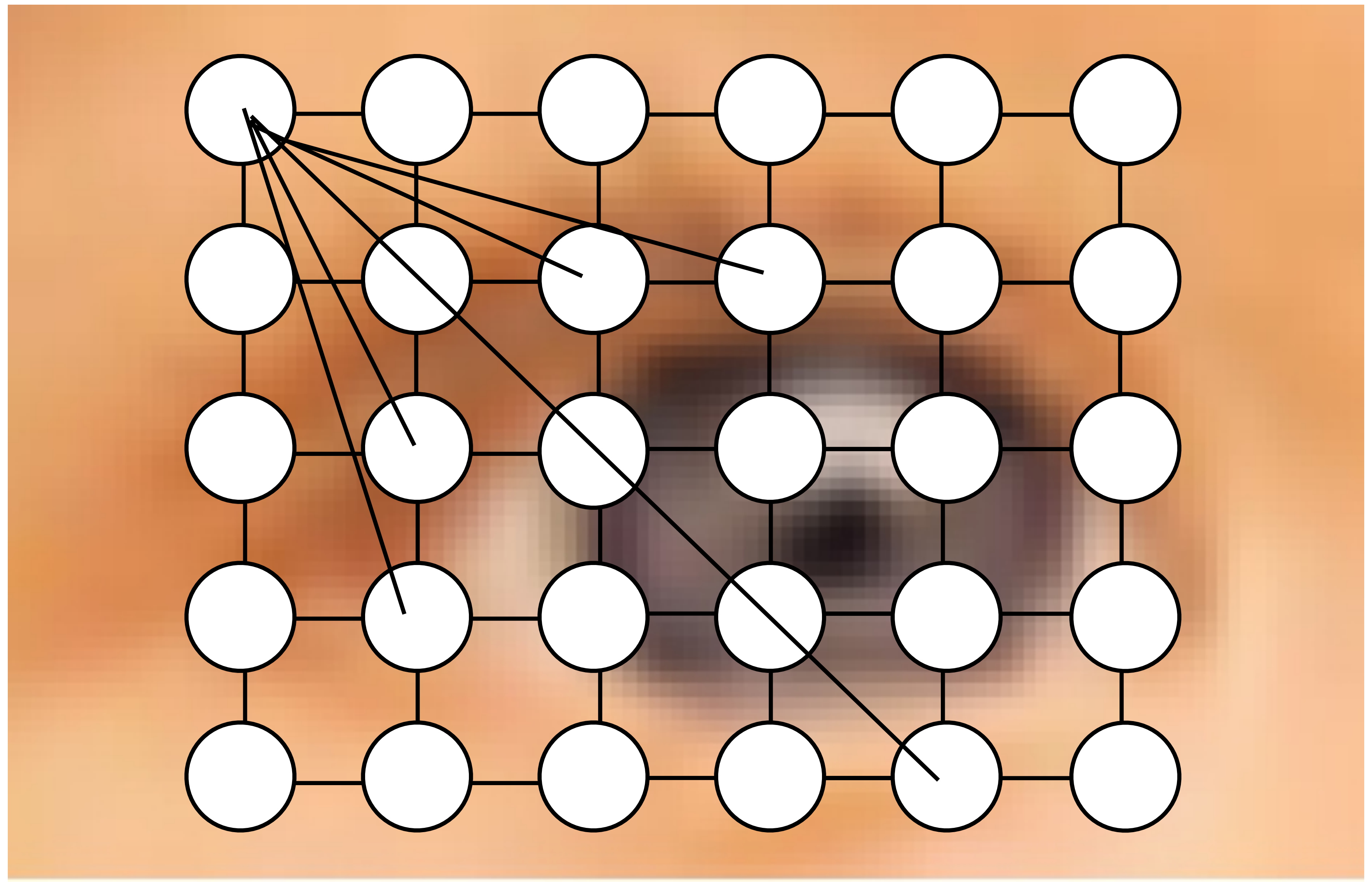
[Johnson, Alahi, Li, ECCV 2016]

[Chen & Koltun ICCV 2017]

[Zhang et al. CVPR 2018]

[Mostajabi, Maire, Shakhnarovich, arXiv 2018]

Structured Prediction

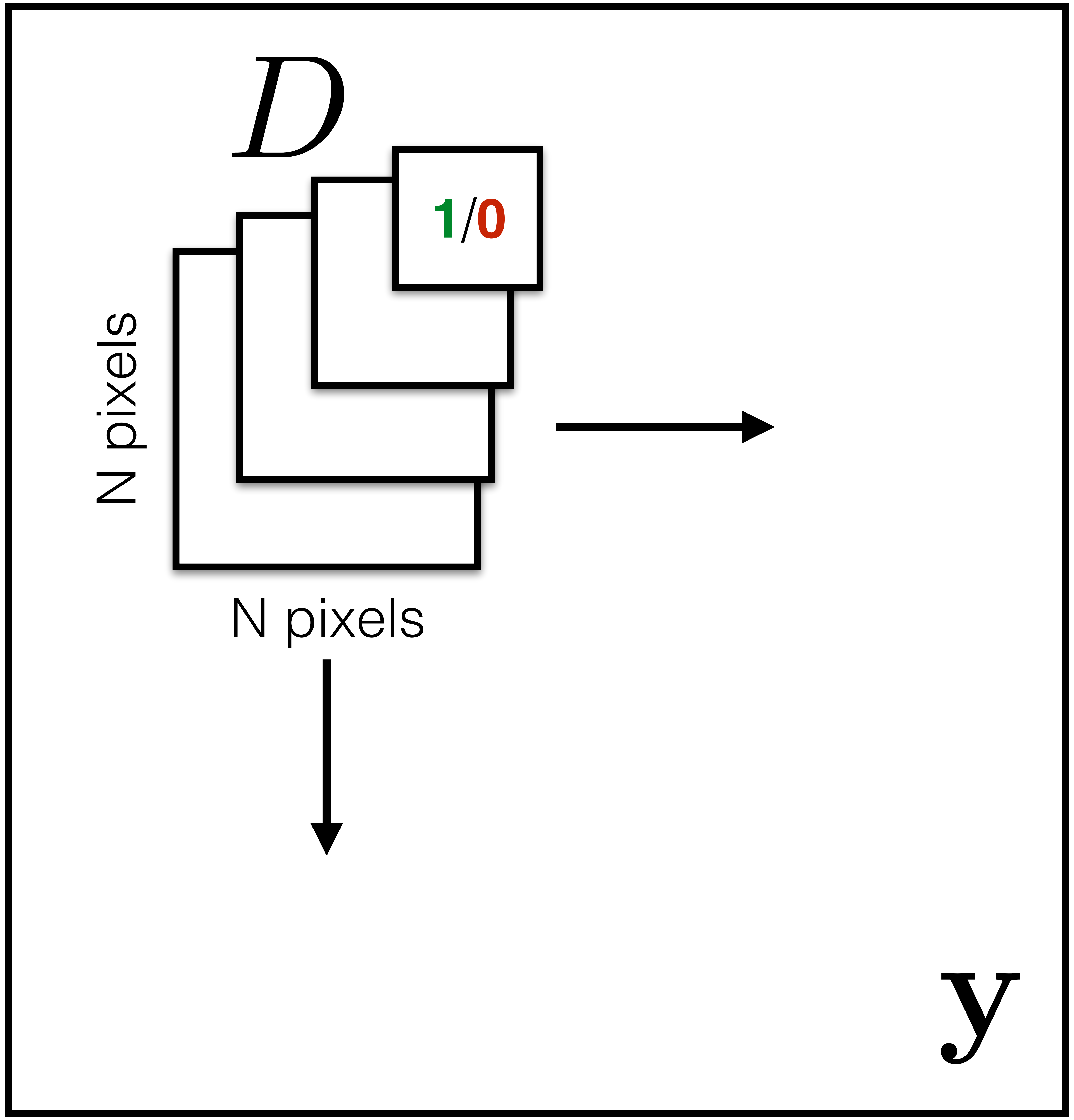


Model *joint* configuration
of all pixels

$$p(\mathbf{y}|\mathbf{x})$$

A GAN, with sufficient capacity, samples from the full joint distribution (at equilibrium)

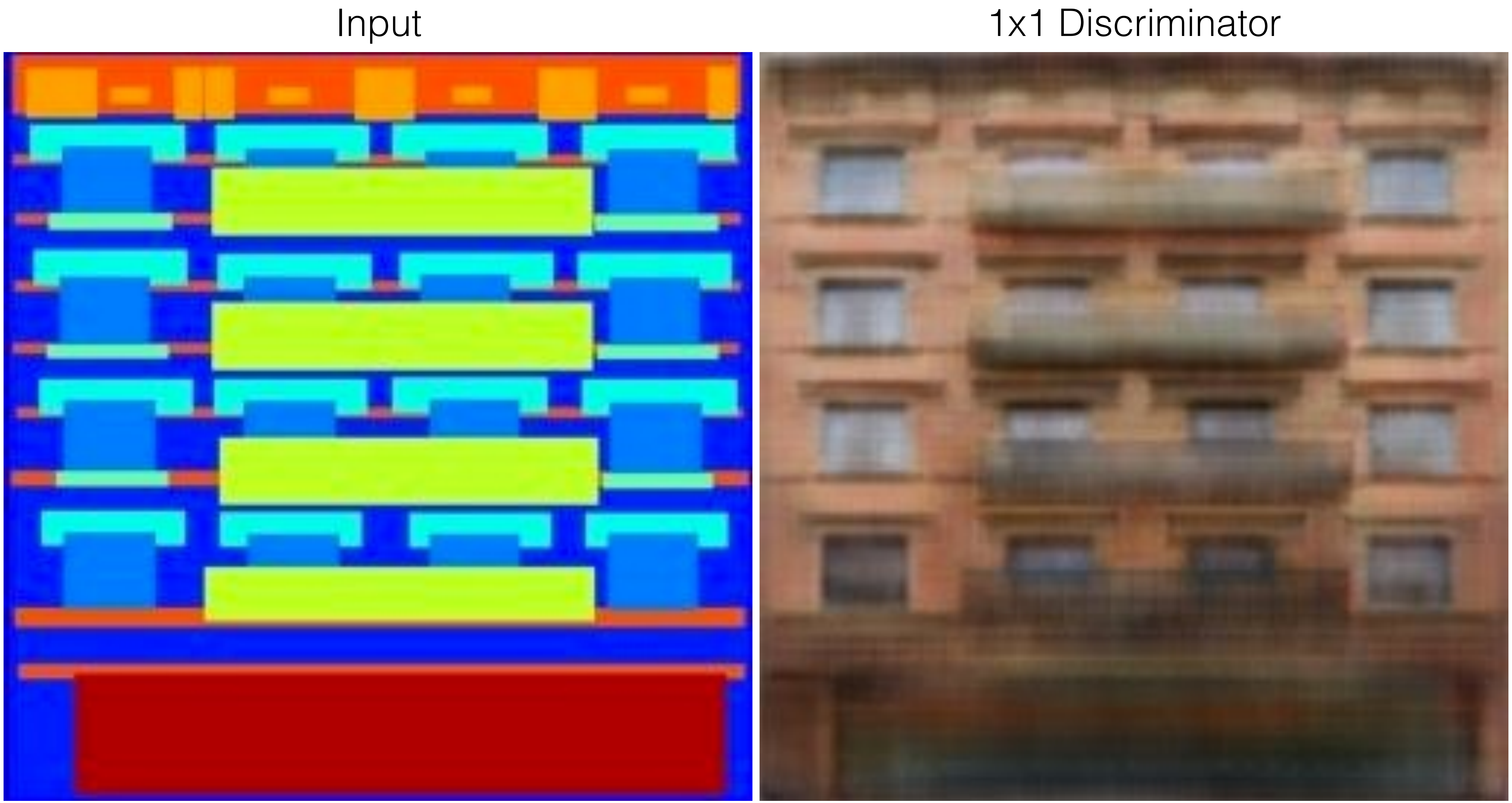
Patch Discriminator



Rather than penalizing if output *image* looks fake, penalize if each overlapping *patch* in output looks fake

[Li & Wand 2016]
[Shrivastava et al. 2017]
[Isola et al. 2017]

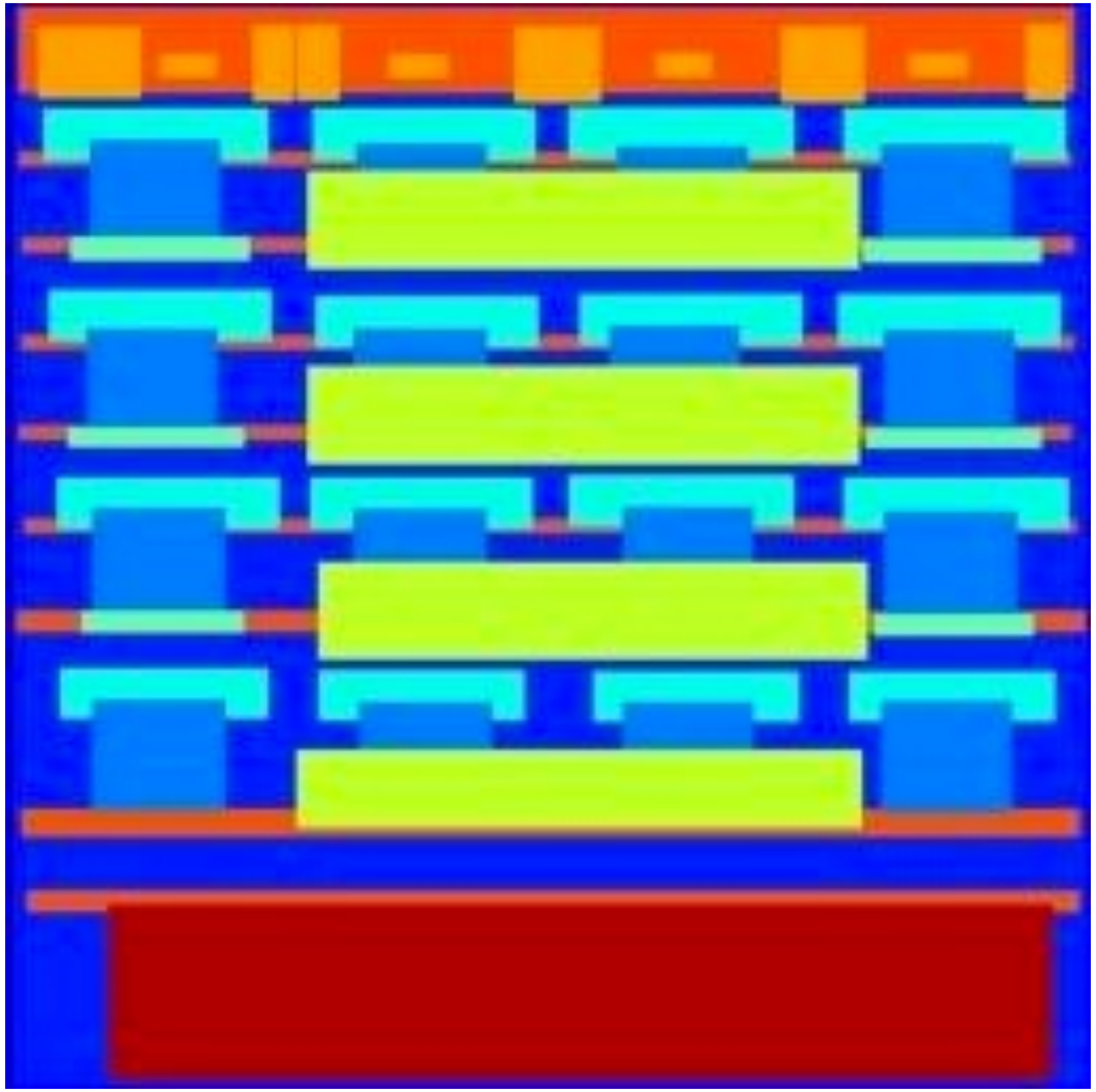
Labels → Facades



Data from [Tylecek, 2013]

Labels → Facades

Input

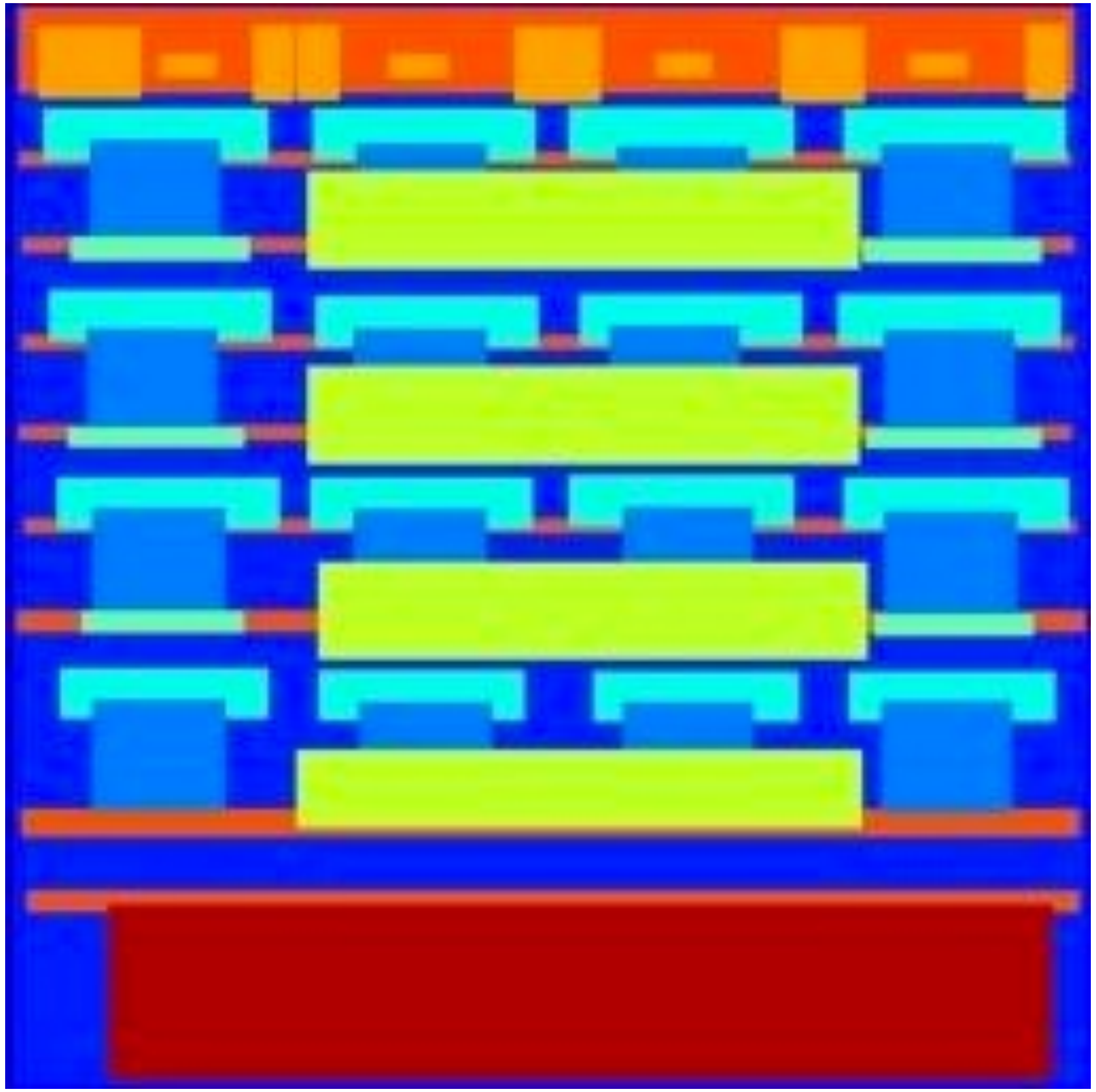


16x16 Discriminator

Data from [Tylecek, 2013]

Labels → Facades

Input

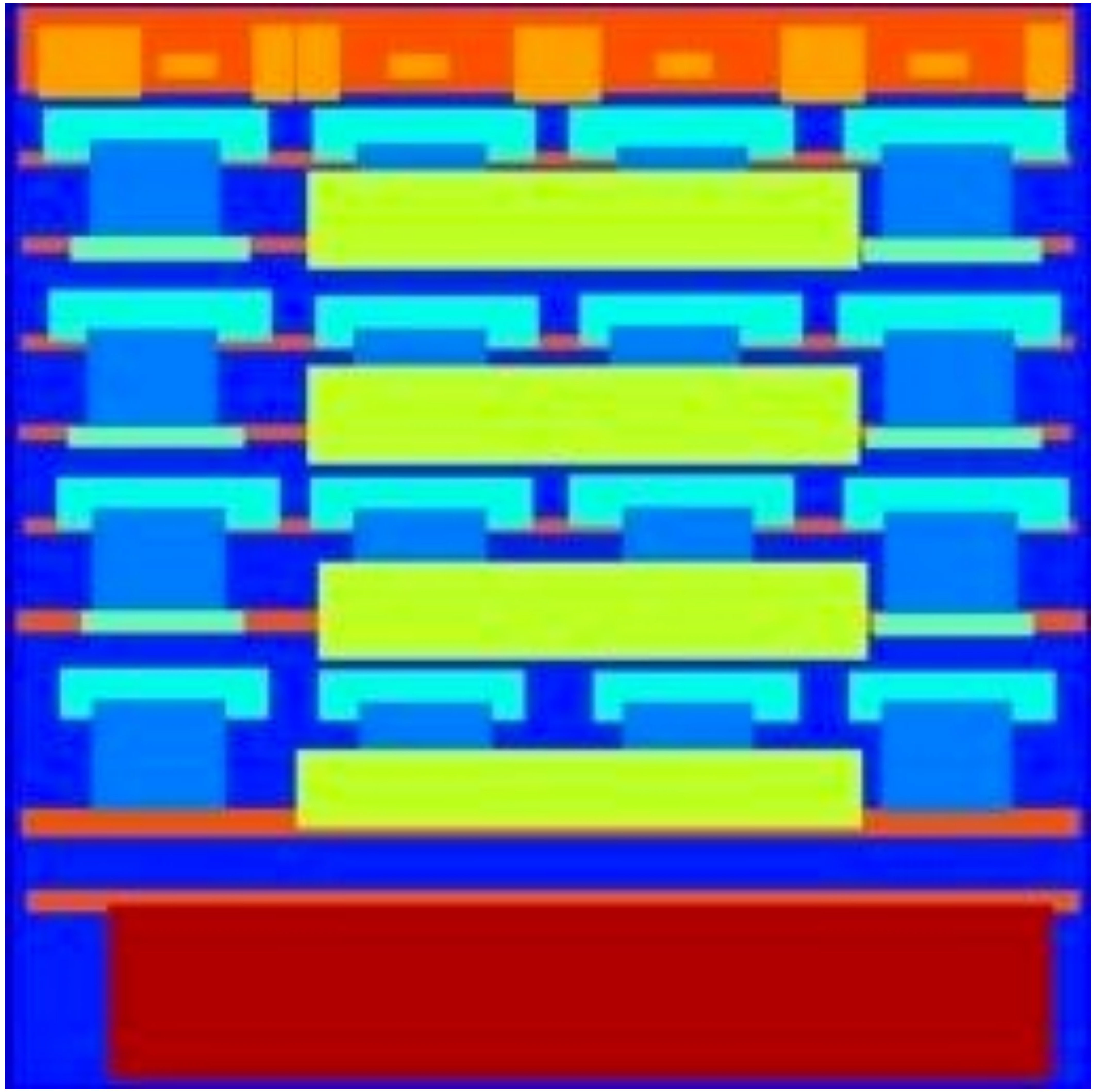


70x70 Discriminator

Data from [Tylecek, 2013]

Labels → Facades

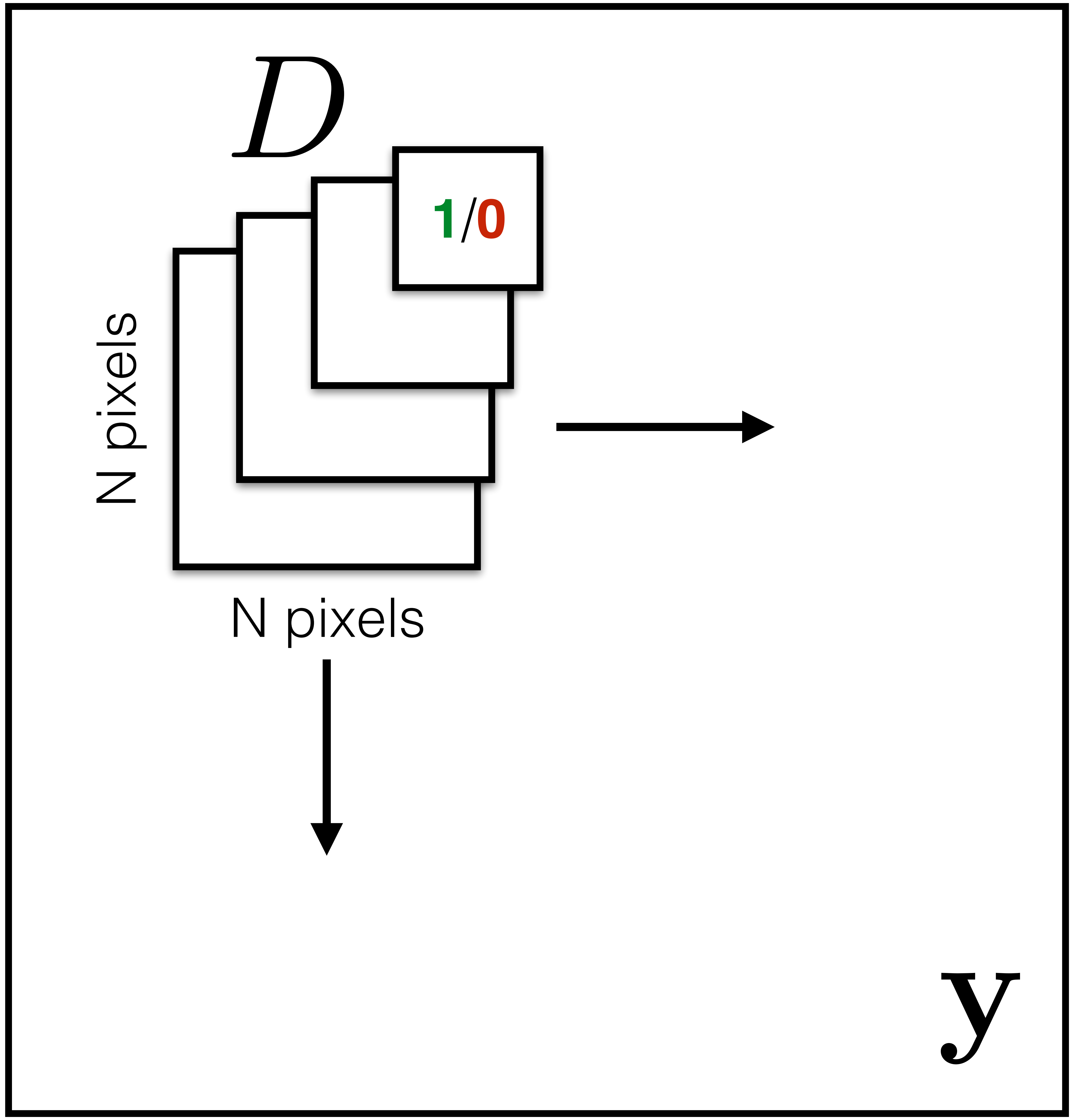
Input



Full image Discriminator

Data from [Tylecek, 2013]

Patch Discriminator

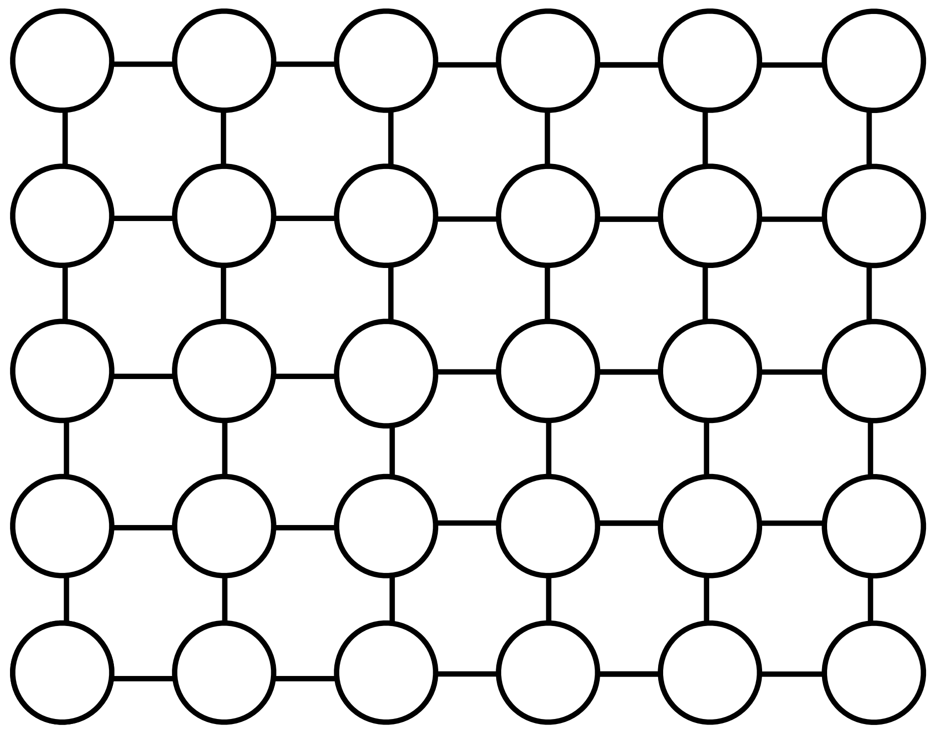


Rather than penalizing if output *image* looks fake, penalize if each overlapping *patch* in output looks fake

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

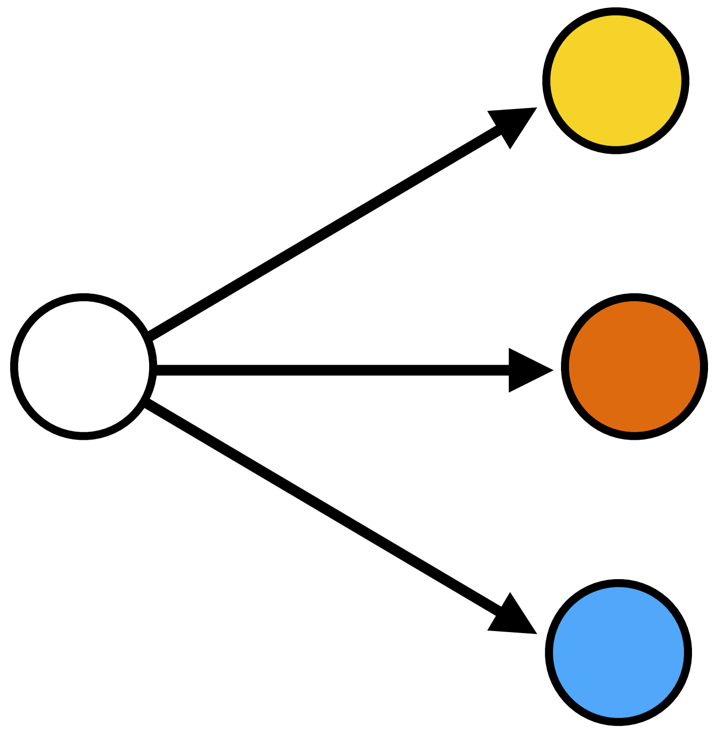
Challenges in image-to-image translation

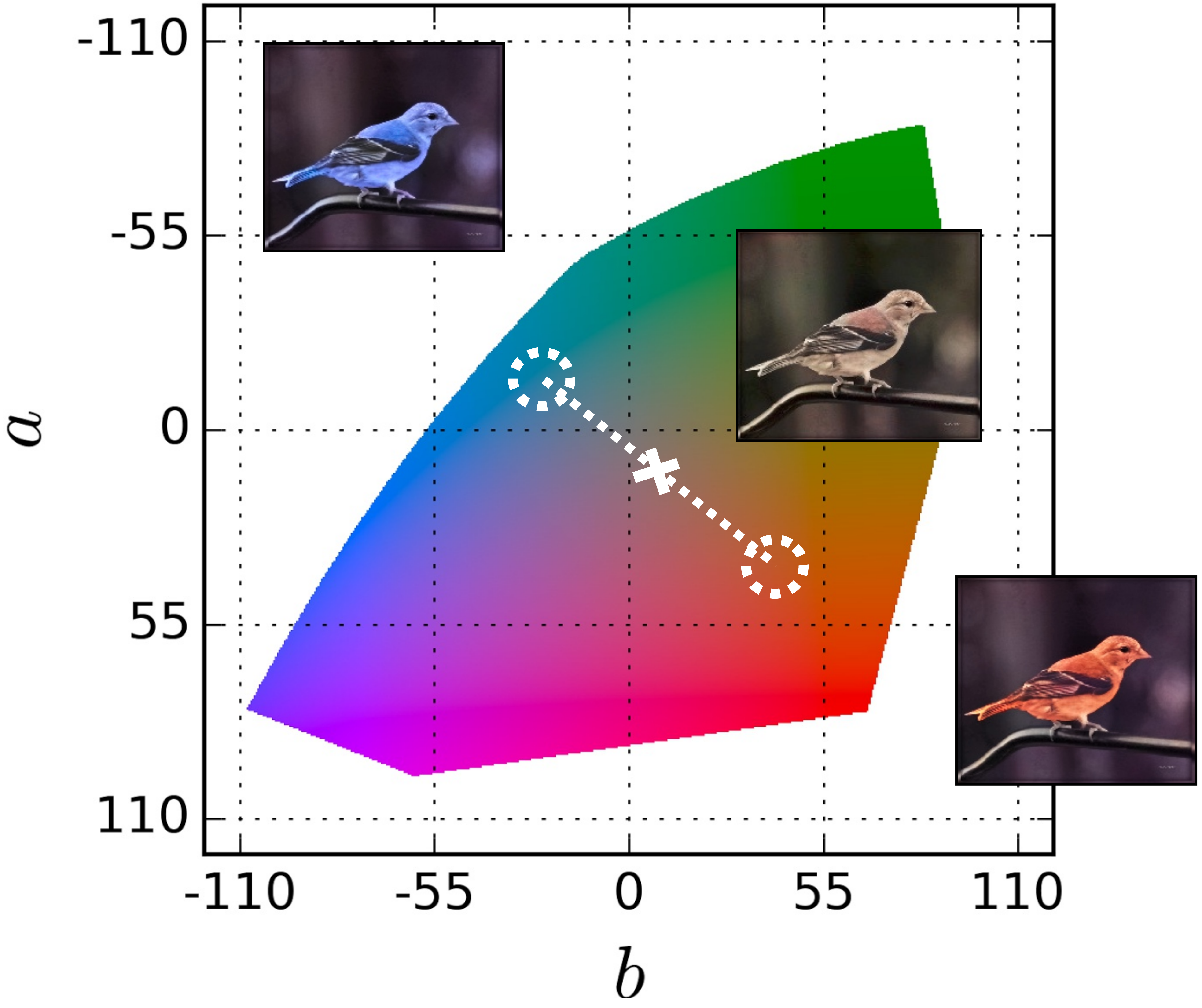
1. Output is high-dimensional, structured object



—> Use a deep net, D, to analyze output!

2. Uncertainty in mapping; many plausible outputs



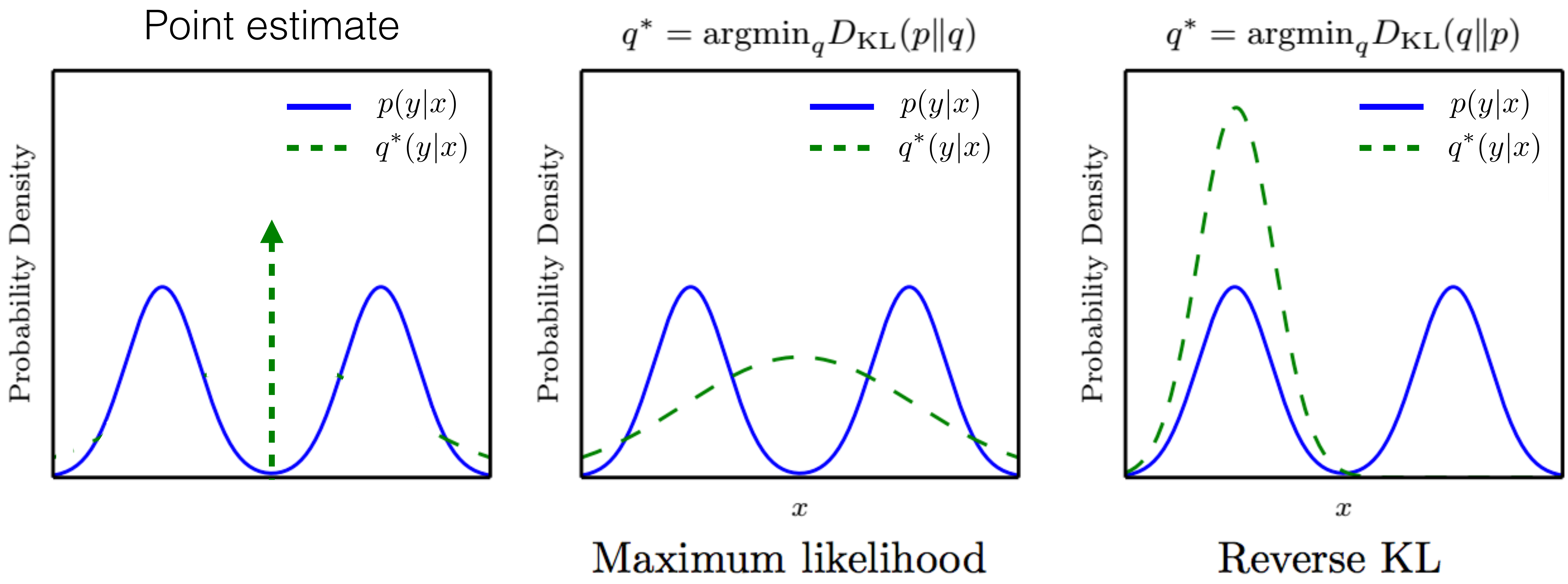


$$L_2(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \widehat{\mathbf{Y}}_{h,w}\|_2^2$$

Input L1 1x1 Discriminator

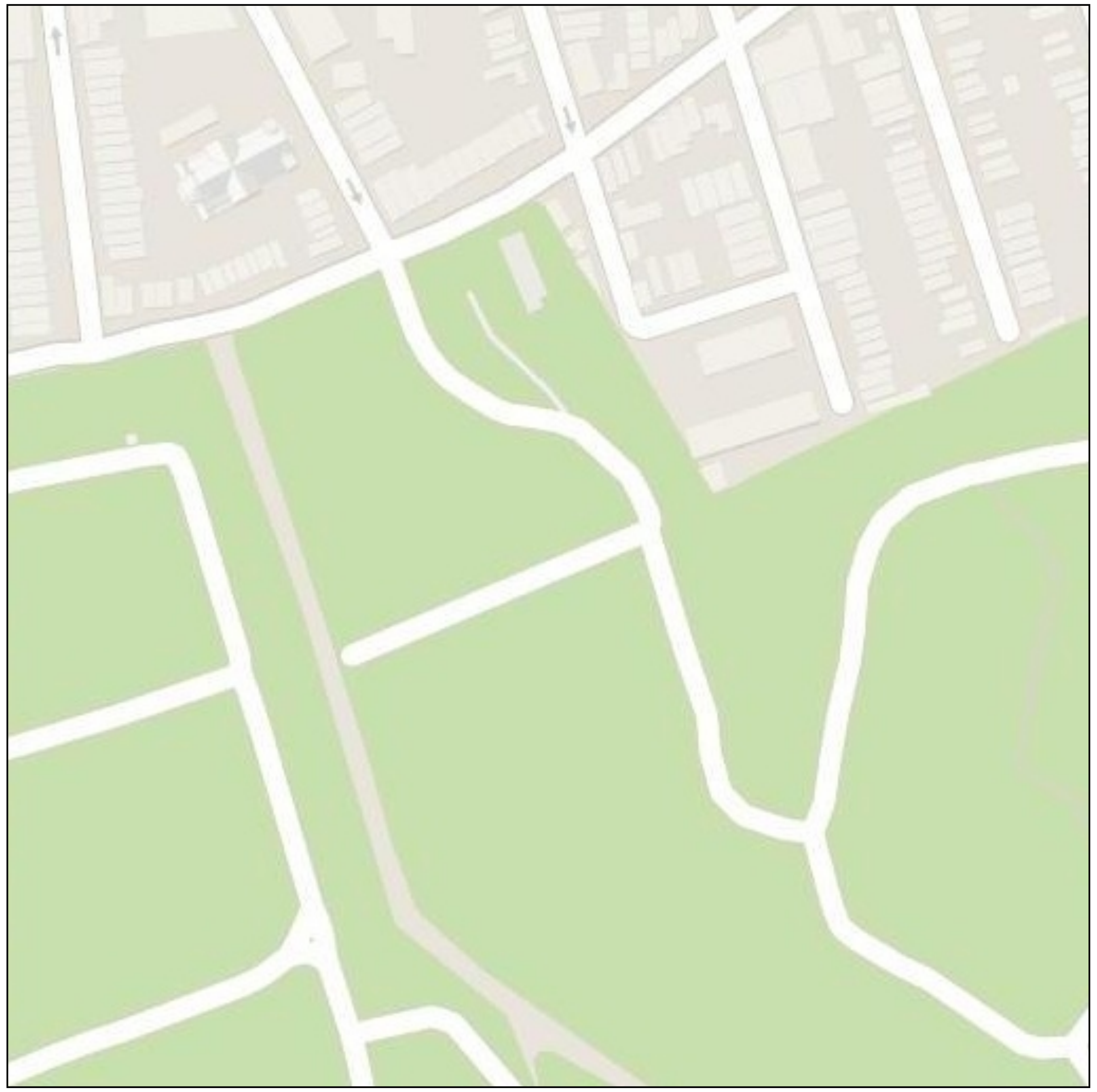
“Unstructured” discriminator makes images colorful!

Mode seeking property



adapted from [Goodfellow, 2016]

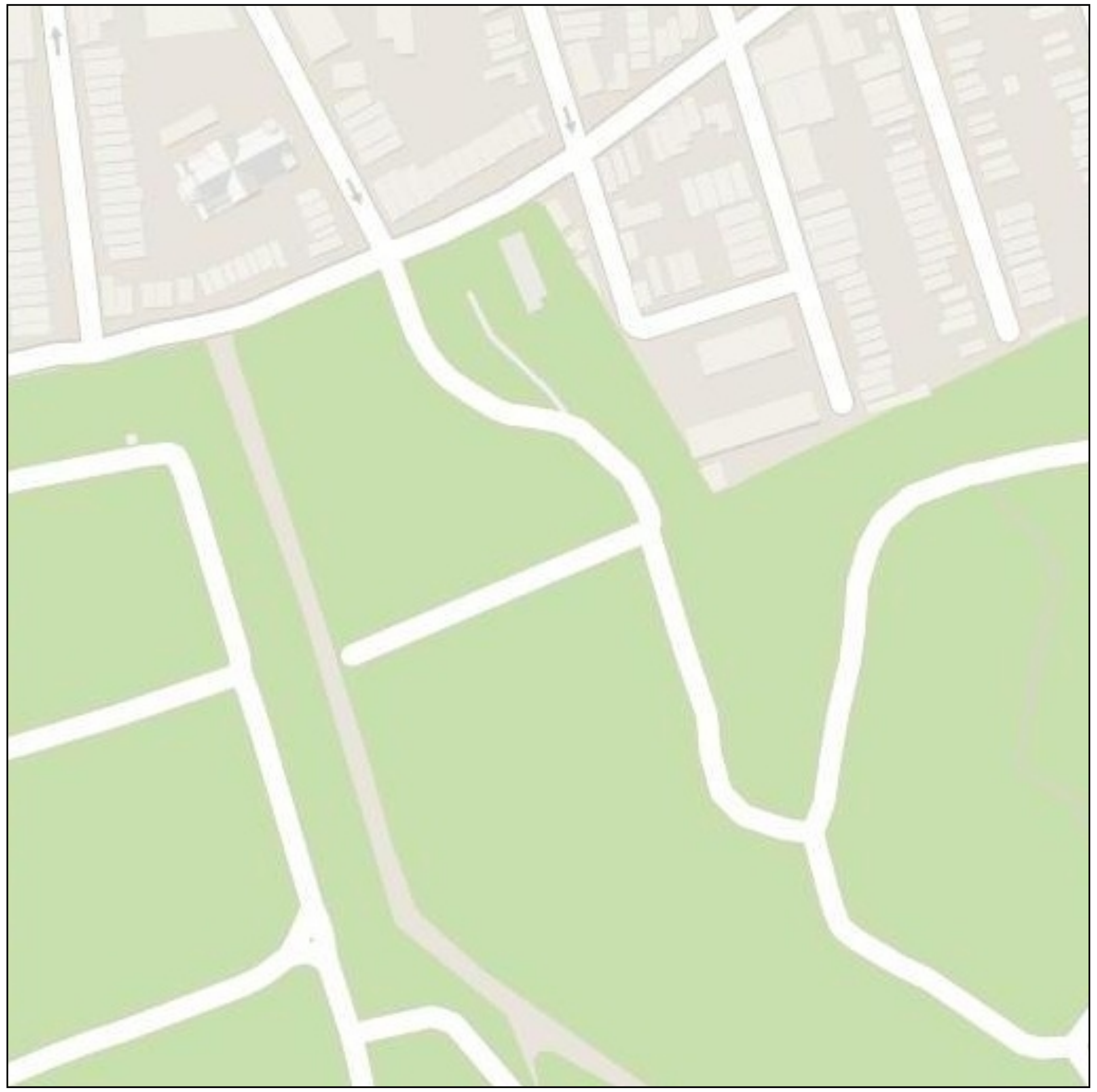
Input



Output

Groundtruth

Input

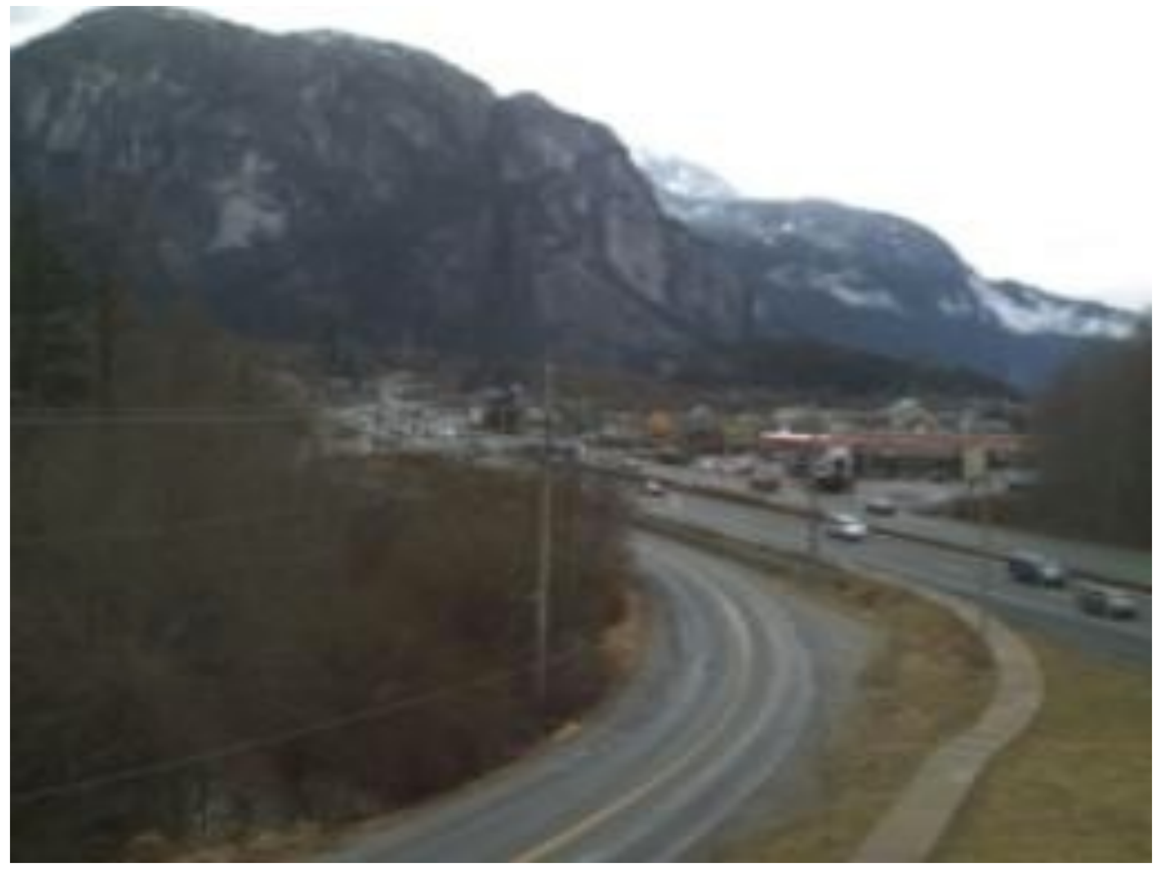


L1 Output

Groundtruth

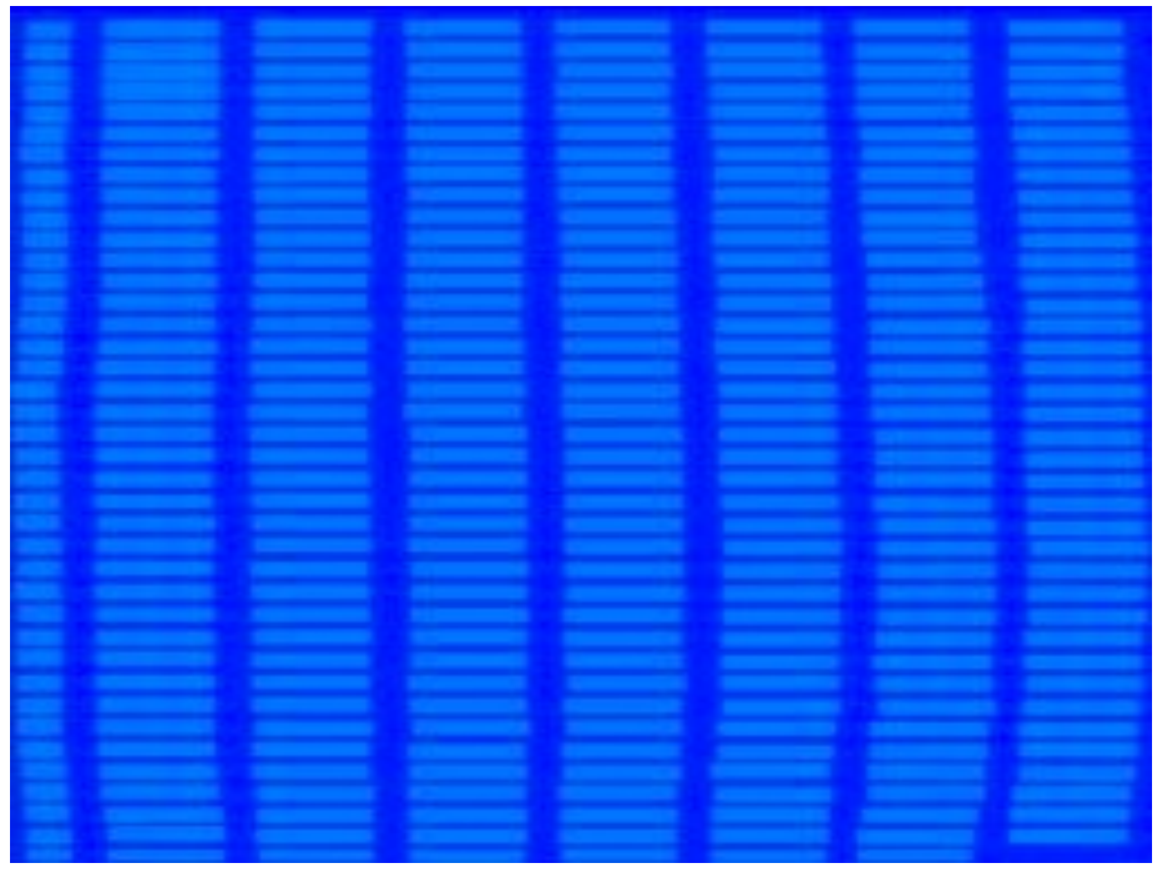
Hallucinations

Input

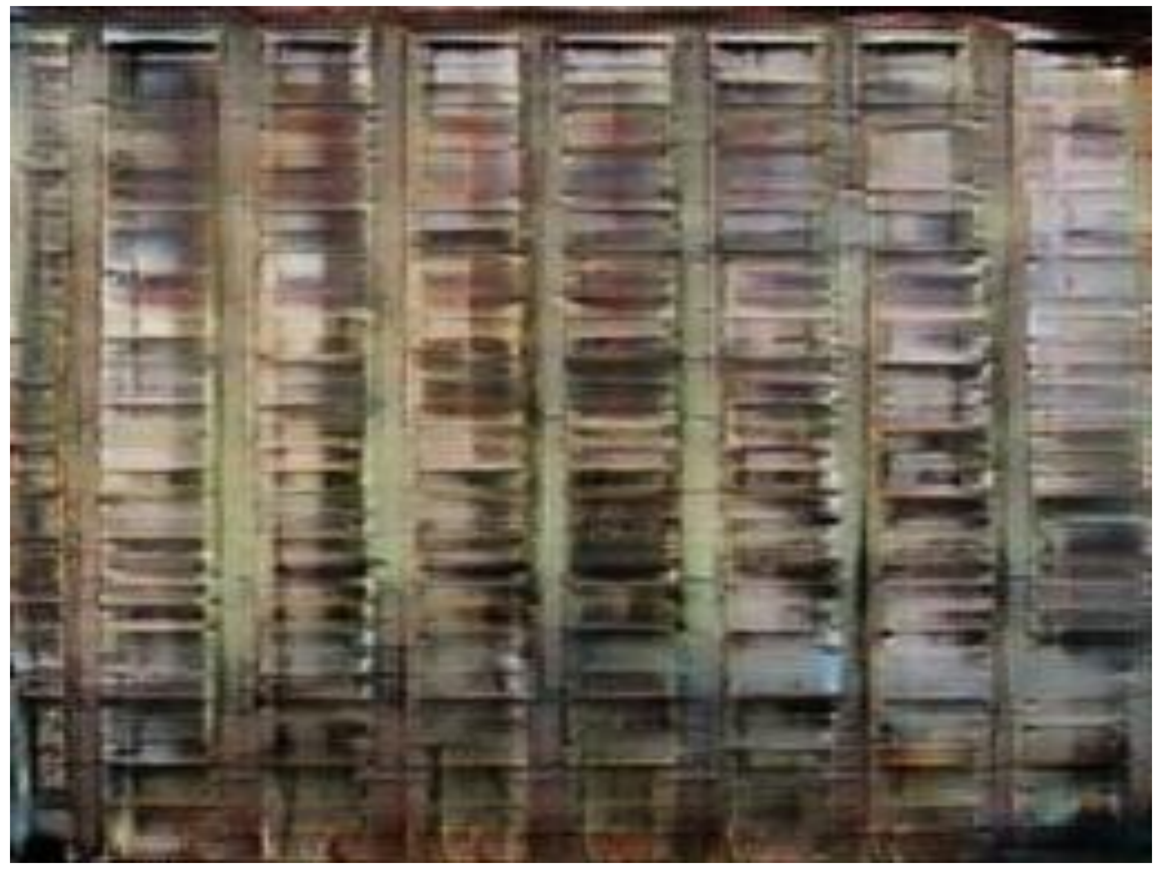


Output

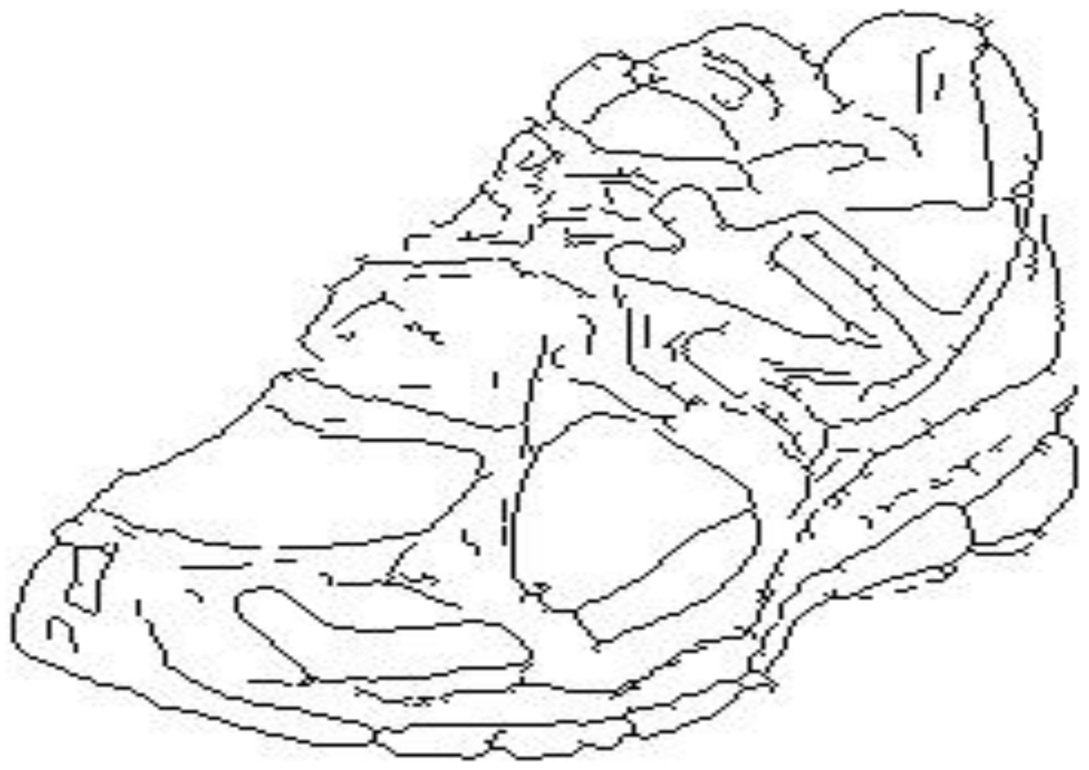
Input



Output

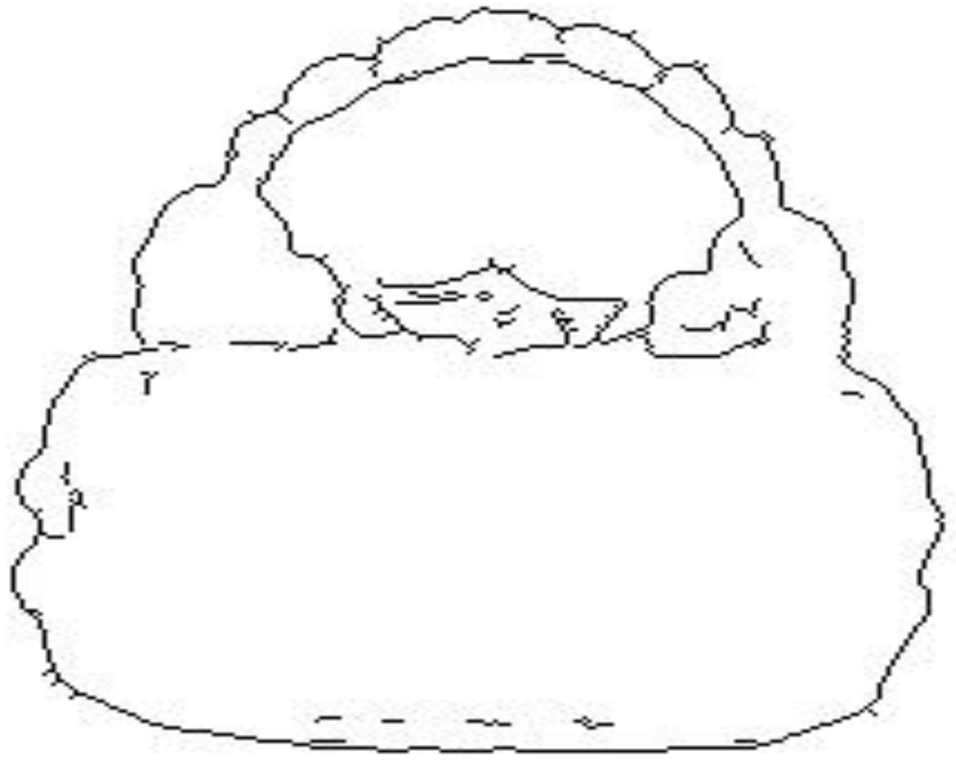


Input



Output

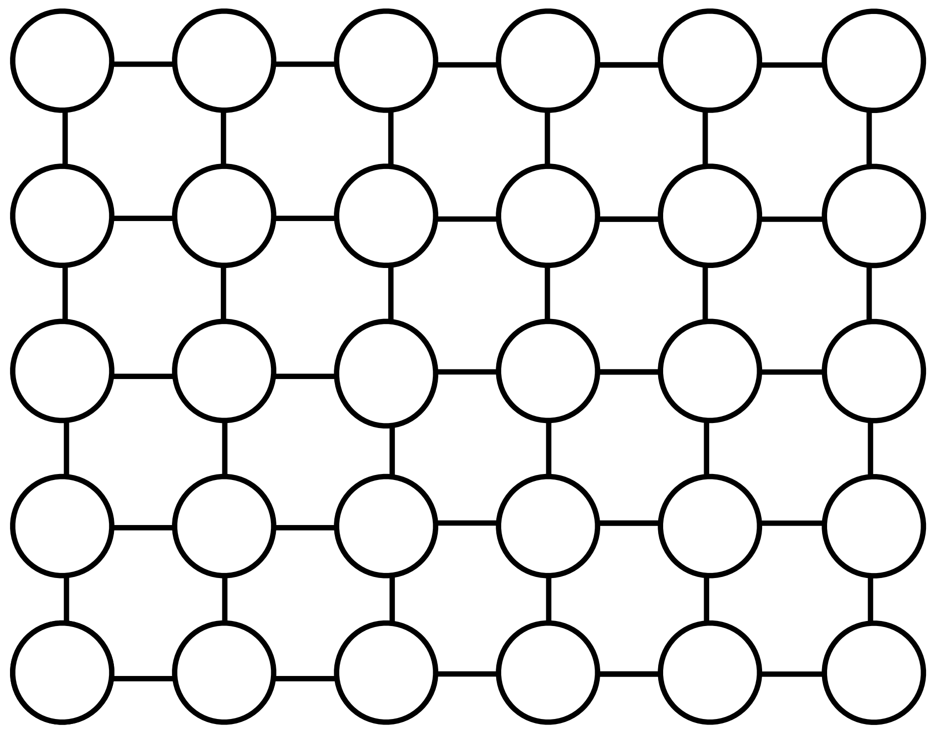
Input



Output

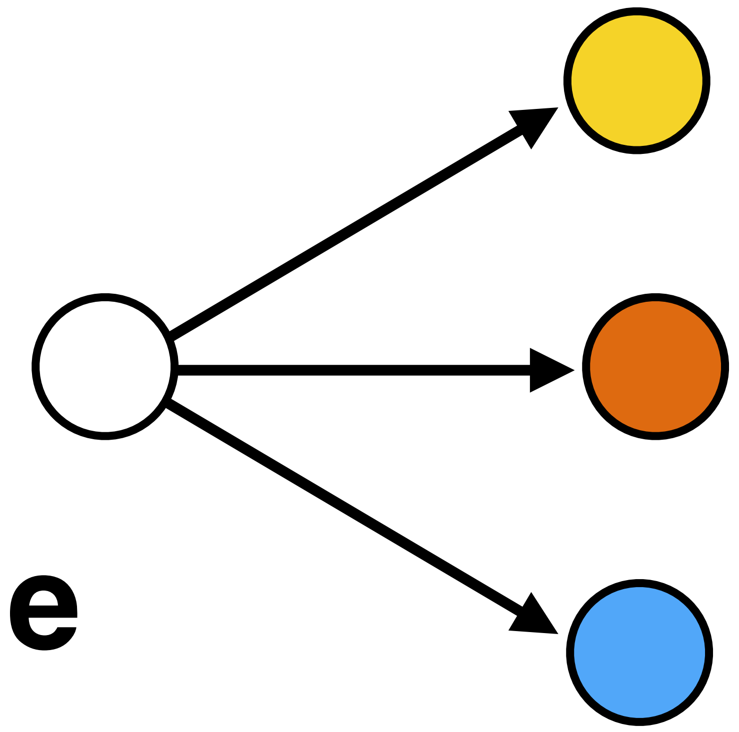
Challenges in image-to-image translation

1. Output is high-dimensional, structured object



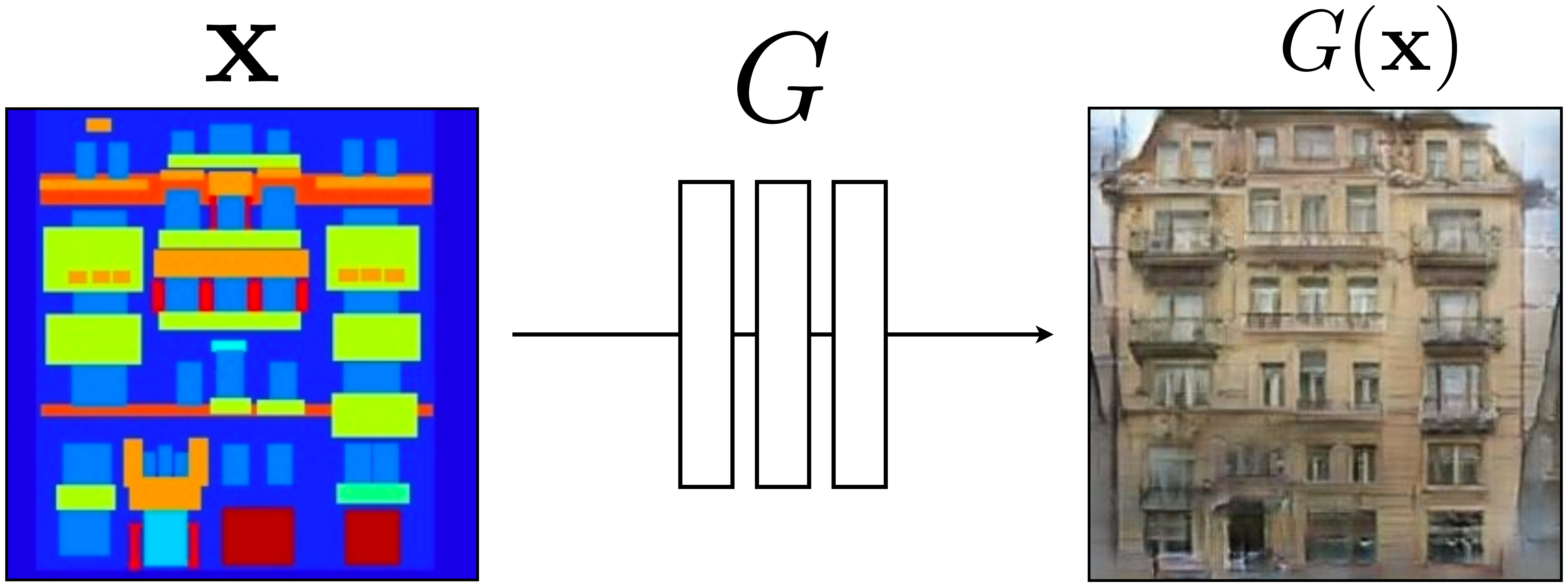
—> Use a deep net, D, to analyze output!

2. Uncertainty in mapping; many plausible outputs

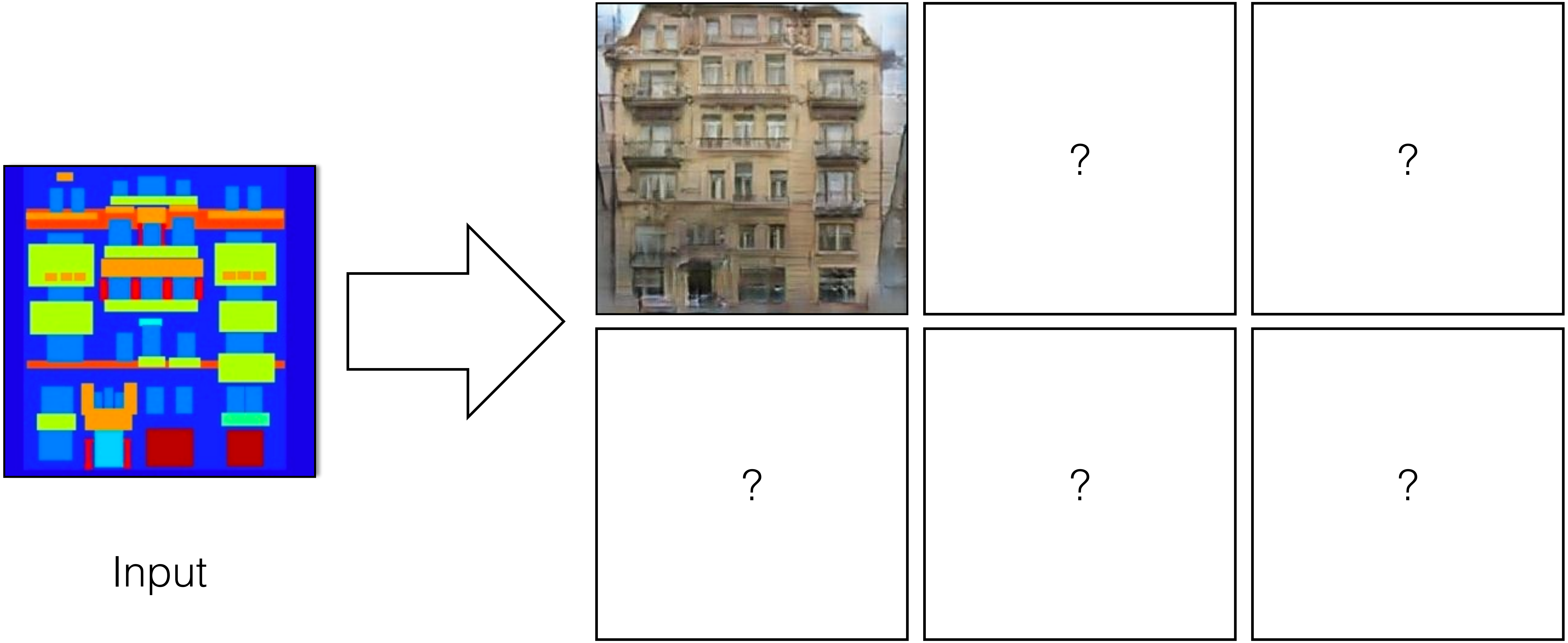


—> D only cares about “plausibility”, doesn’t hedge

Modeling multiple possible outputs



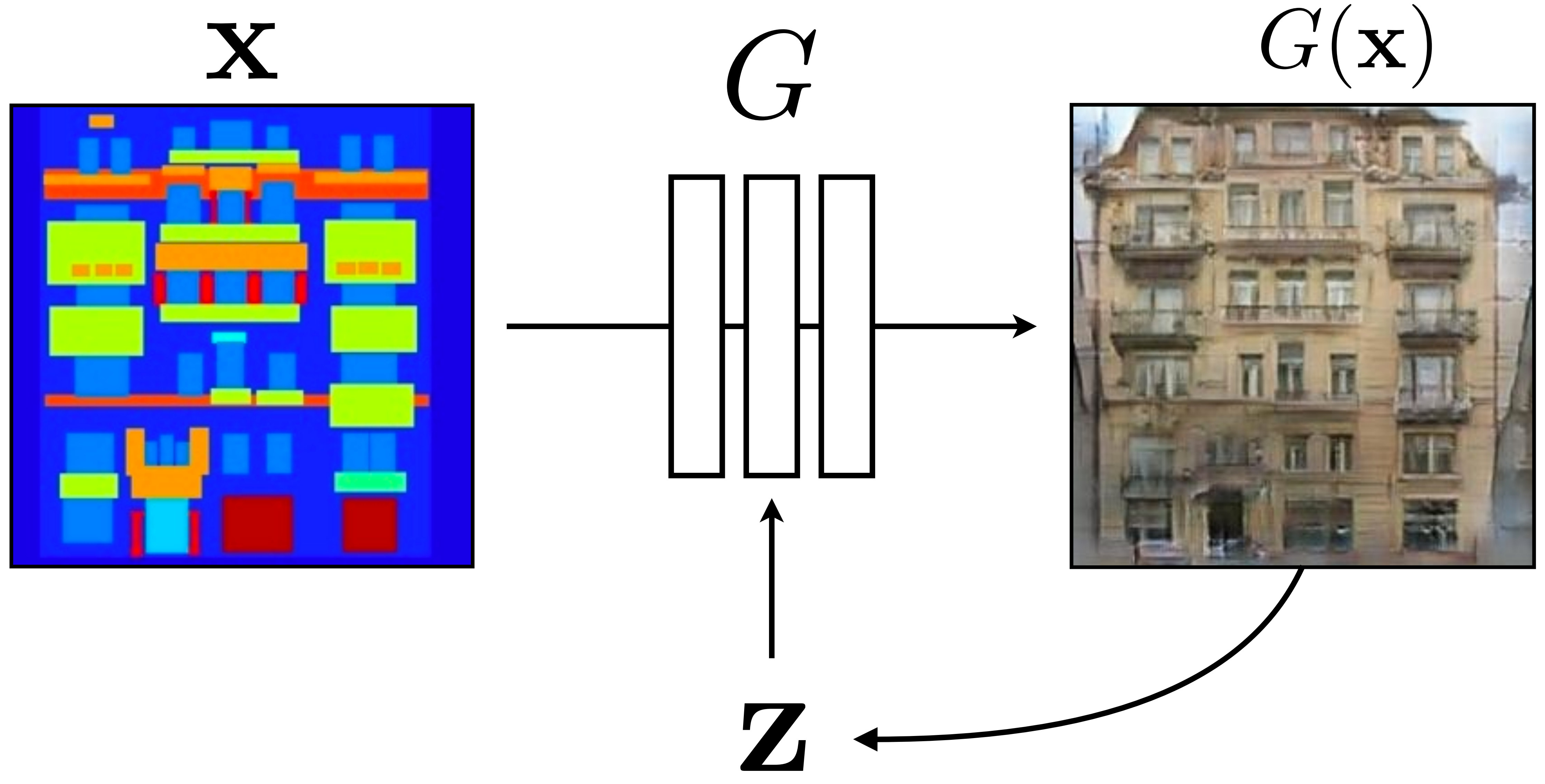
Modeling multiple possible outputs



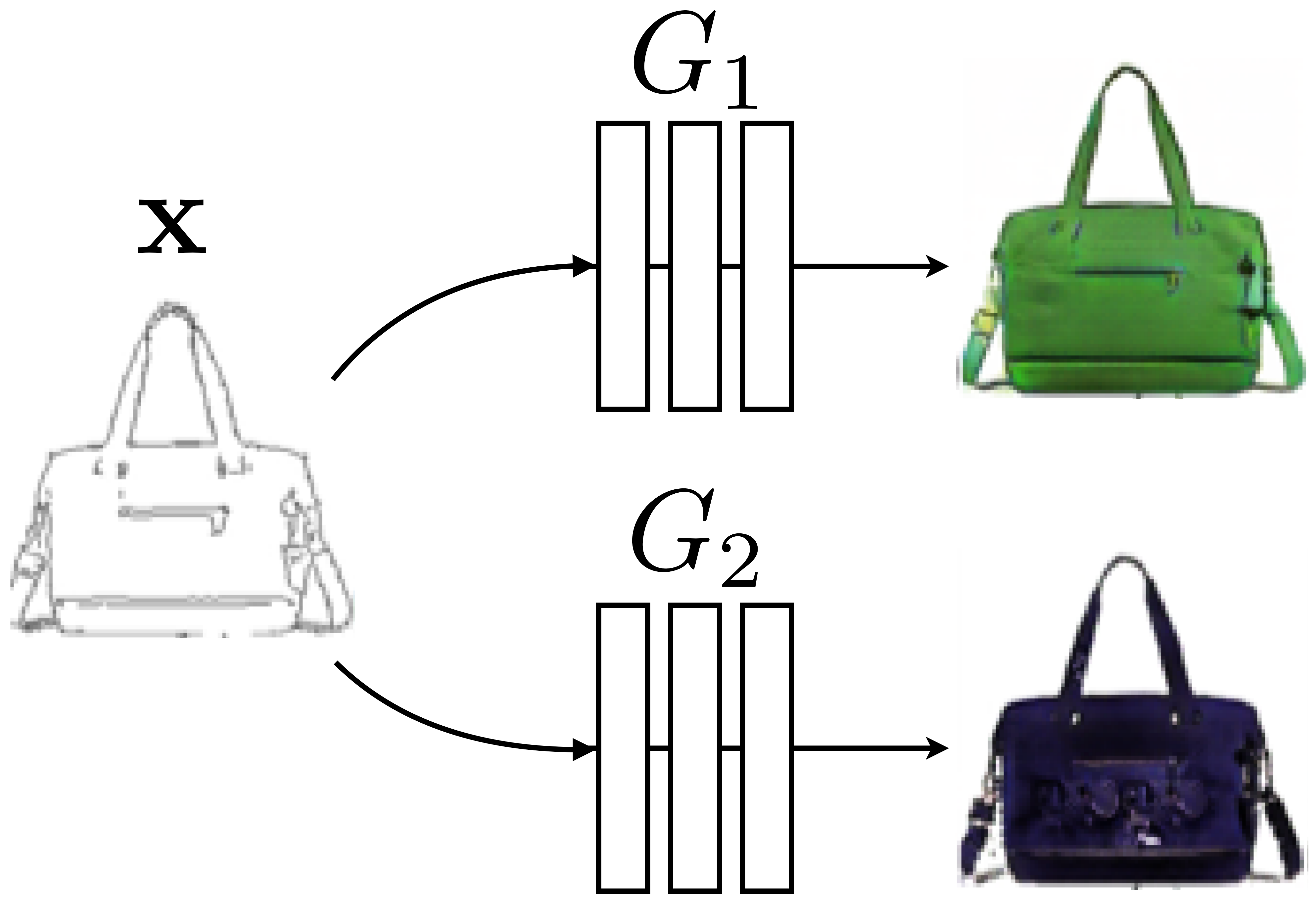
Input

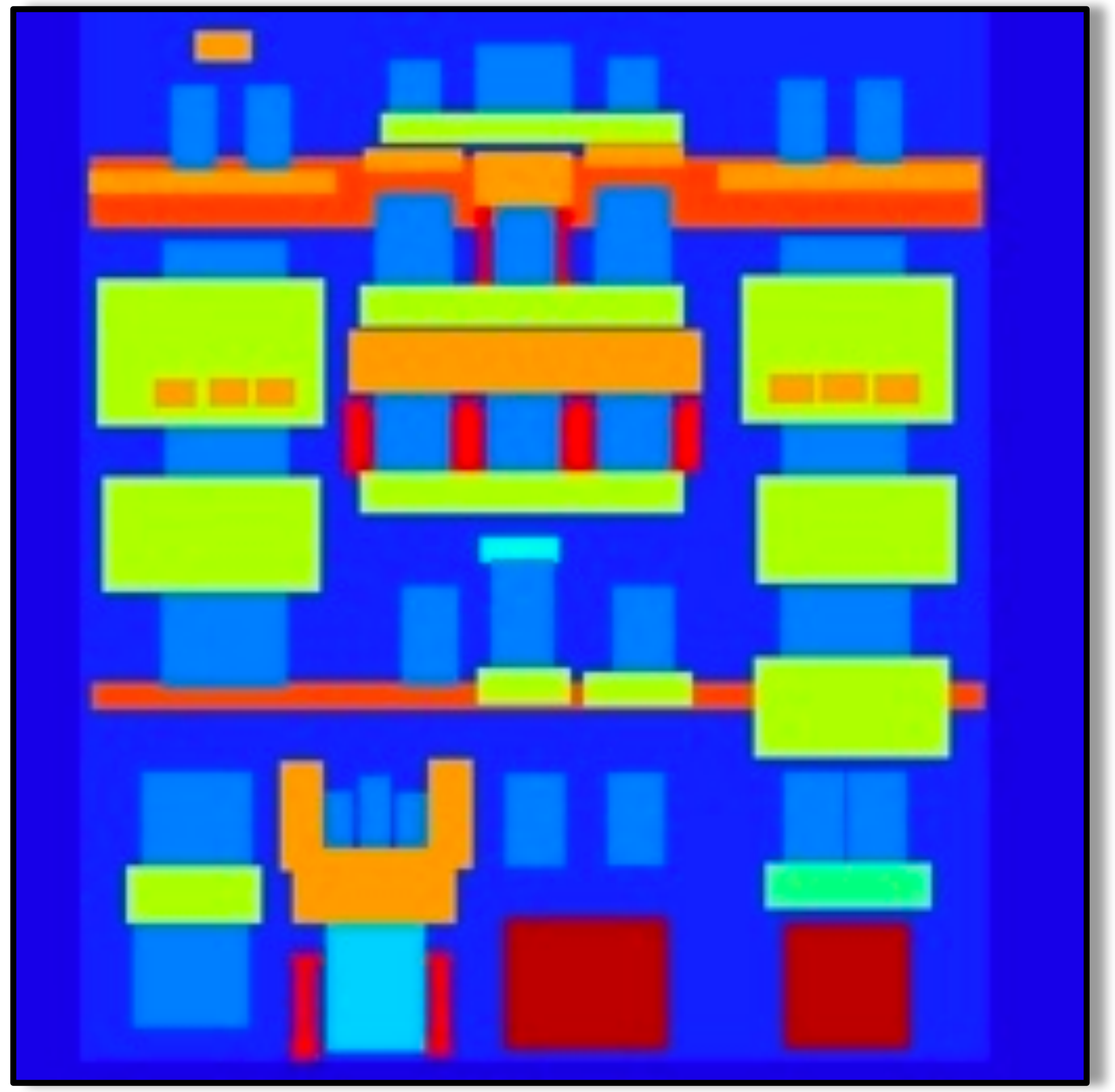
Possible outputs

BiCycleGAN [Zhu et al., NIPS 2017]
(c.f. InfoGAN [Chen et al. 2016])



MAD-GAN [Ghosh et al., CVPR 2018]



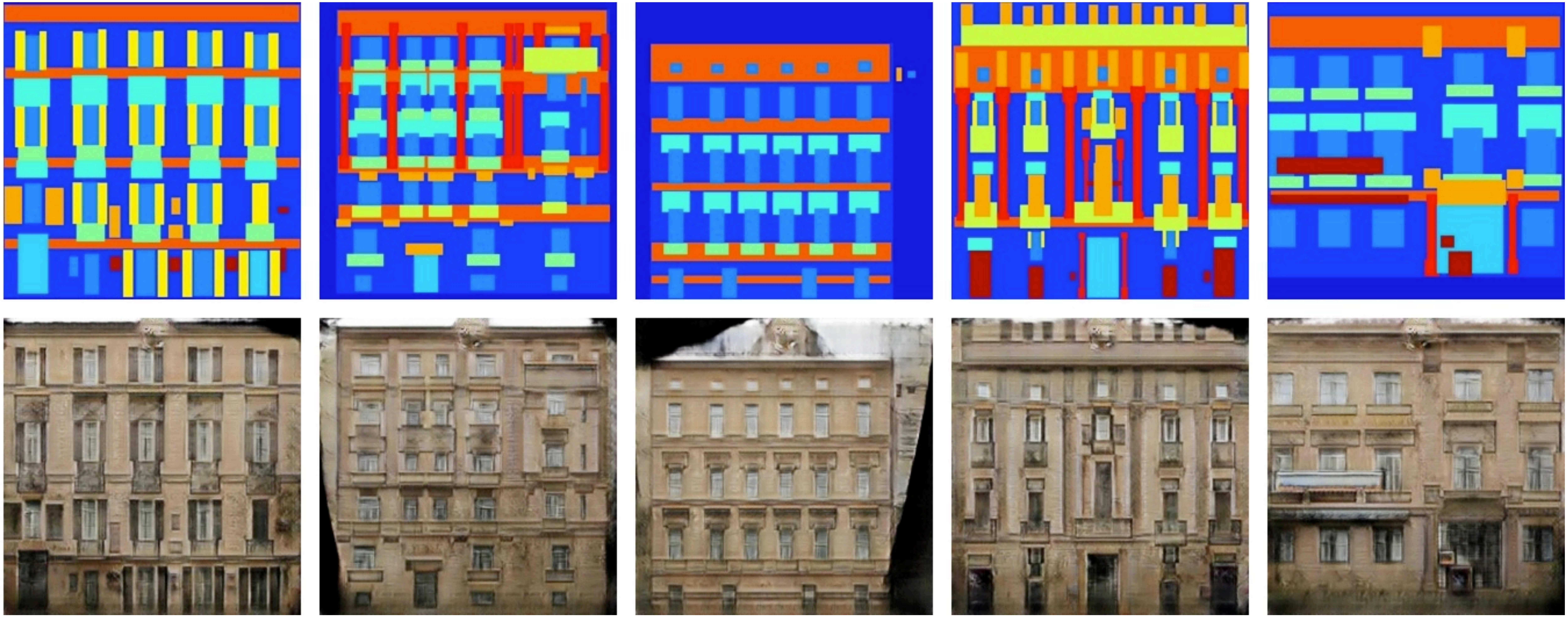


Labels

Randomly generated facades

[BiCycleGAN, Zhu et al., NIPS 2017]

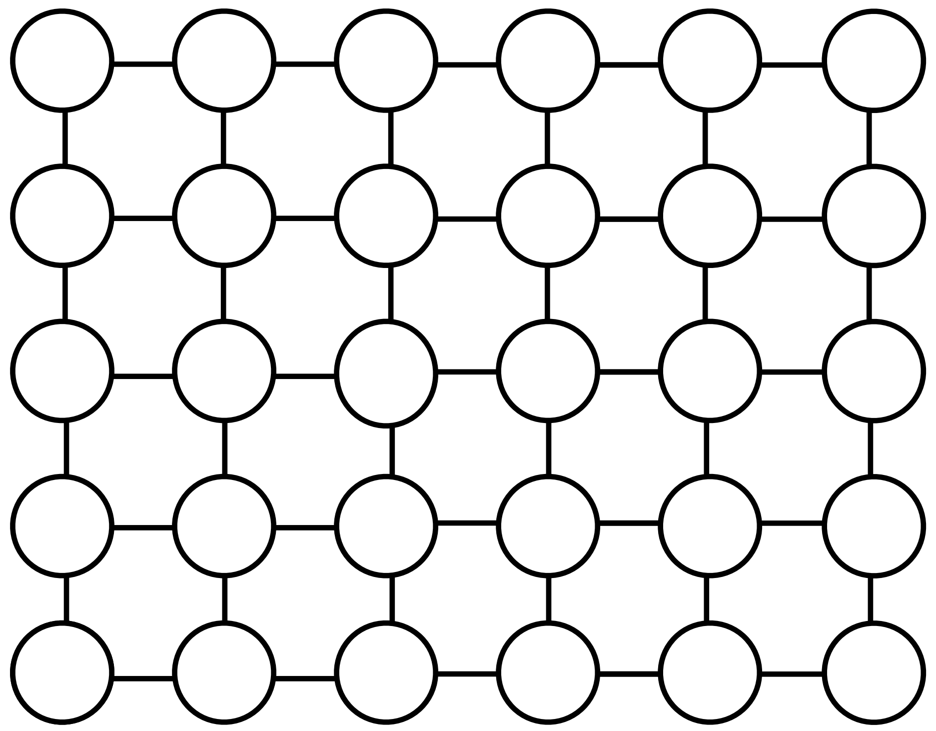
Latent space exploration



[BiCycleGAN, Zhu et al., NIPS 2017]

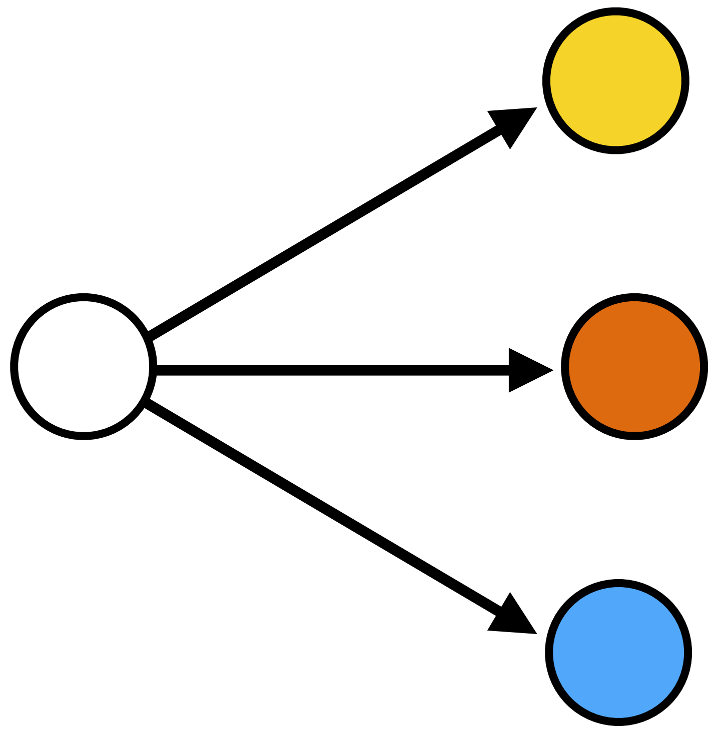
Challenges in image-to-image translation

1. Output is high-dimensional, structured object



   —> Use a deep net, D, to analyze output!

2. Uncertainty in mapping; many plausible outputs



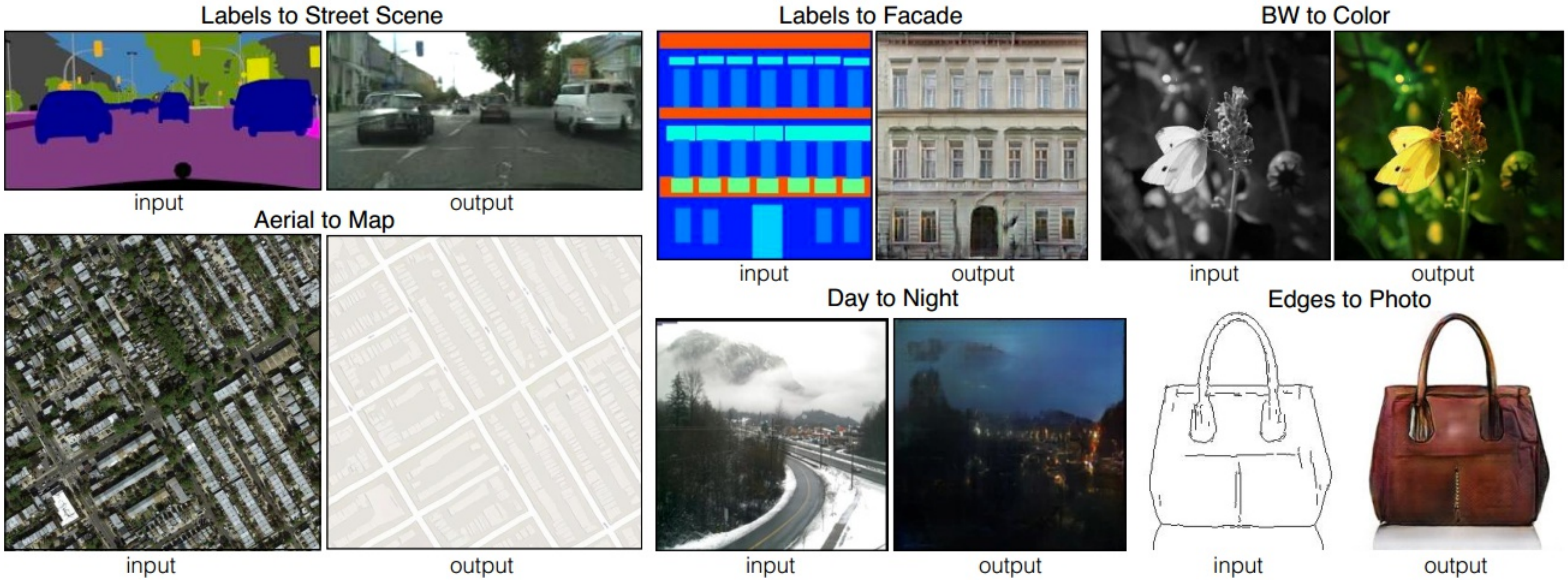
   —> Can model the *distribution* of possibilities

Outline

- Paired image-to-image translation
- Unpaired image-to-image translation

Acknowledgement: The slides adapted from the ones prepared by Jun-Yan Zhu and Taesung Park

Image-to-Image Translation with **pix2pix**



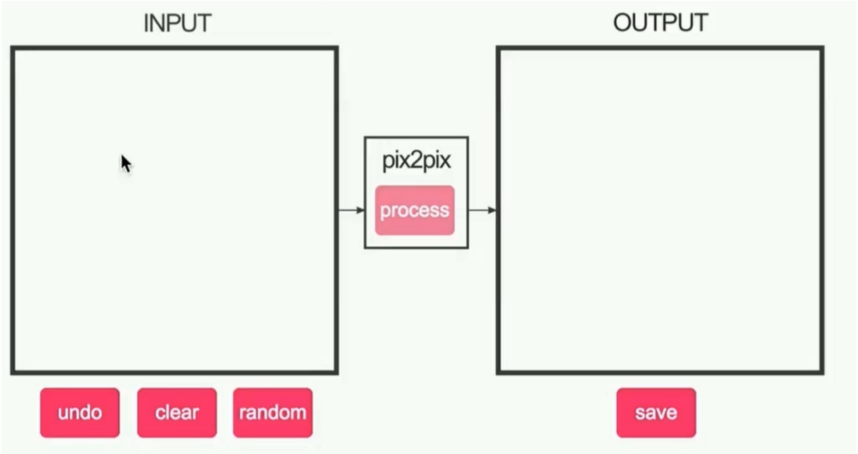
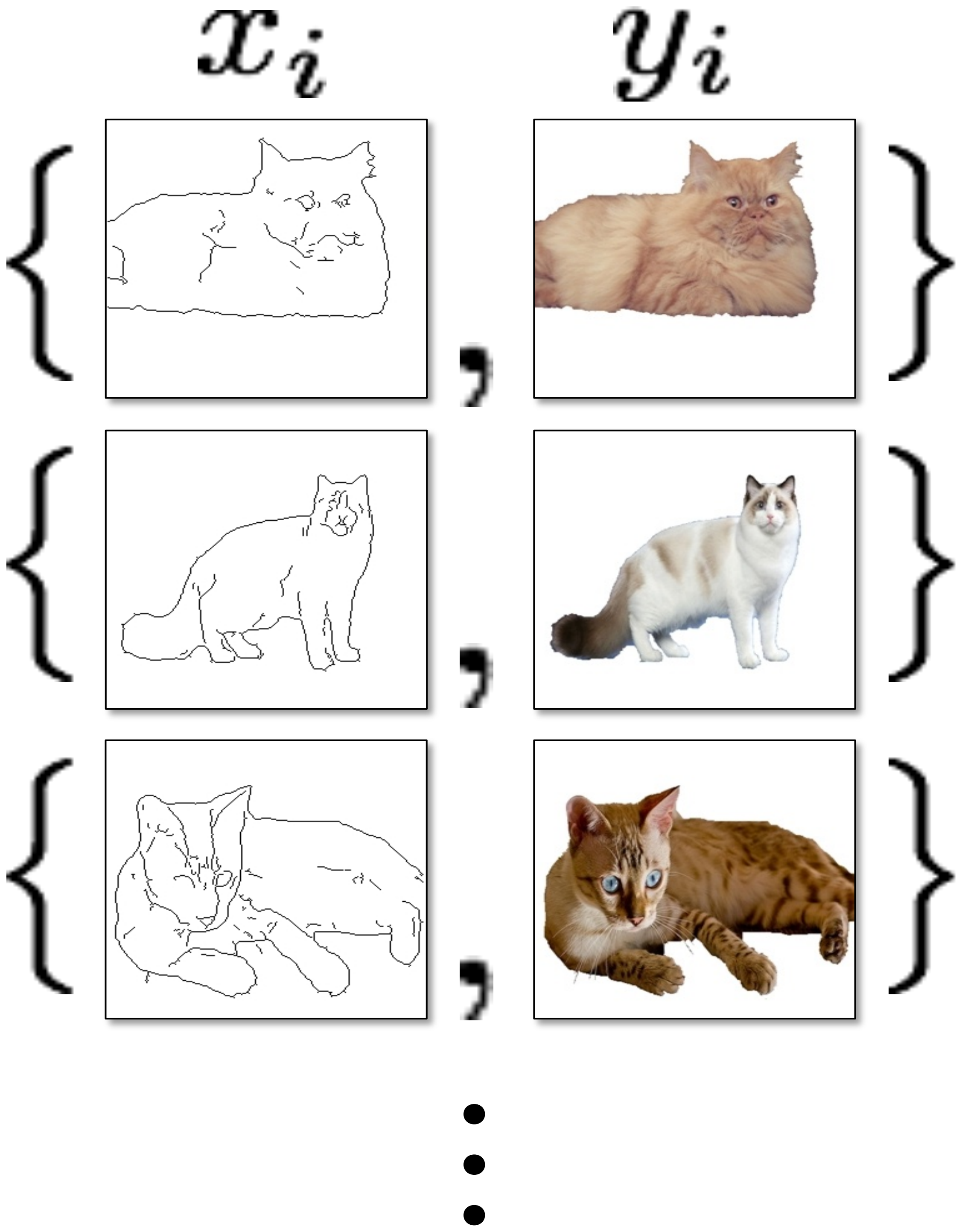


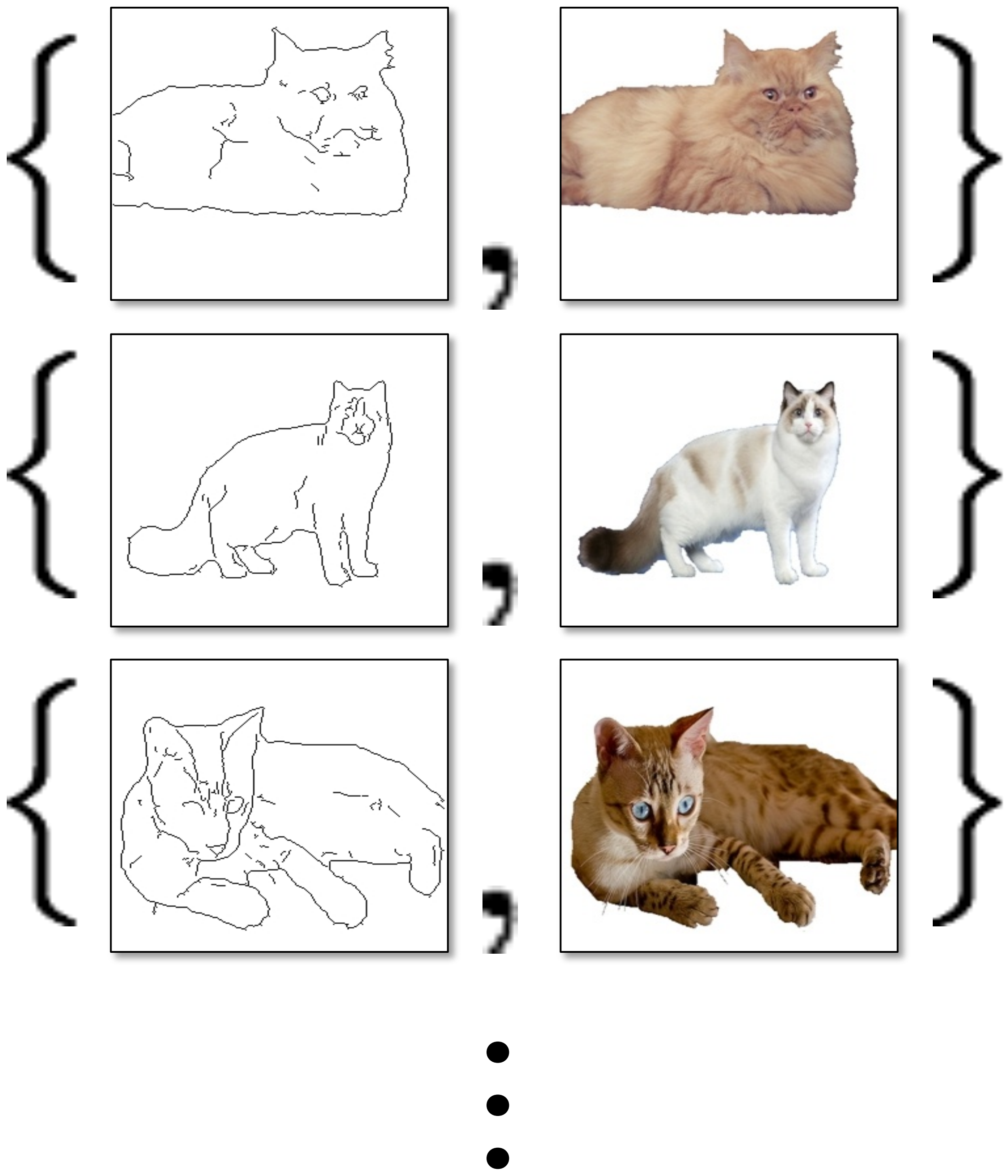
Image-to-image Translation with Conditional Adversarial Nets
Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros. CVPR 2017

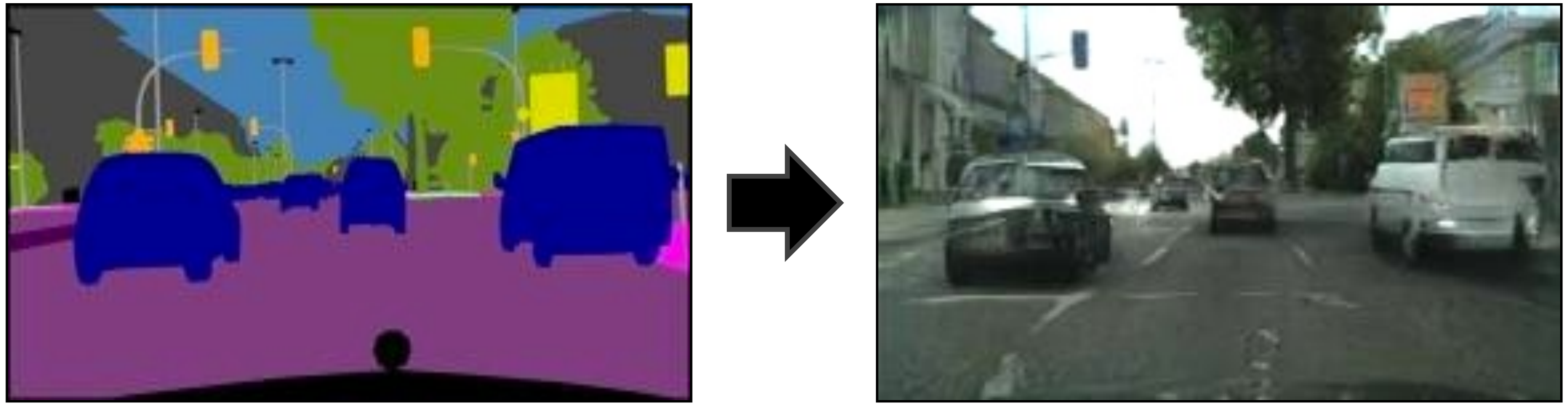
Paired



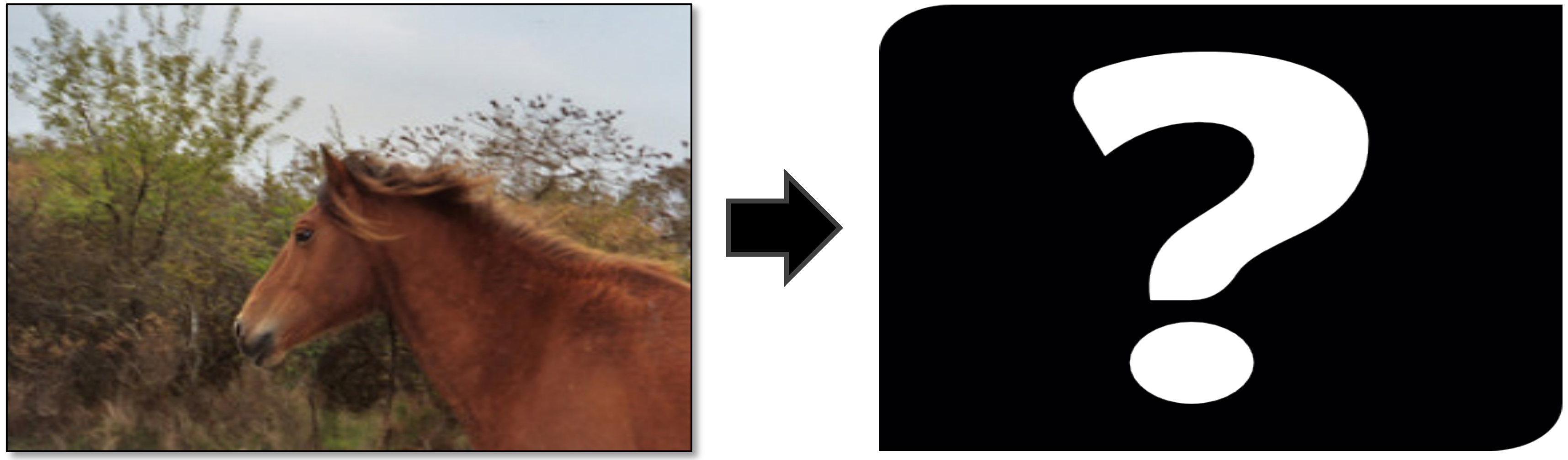
Paired

x_i y_i



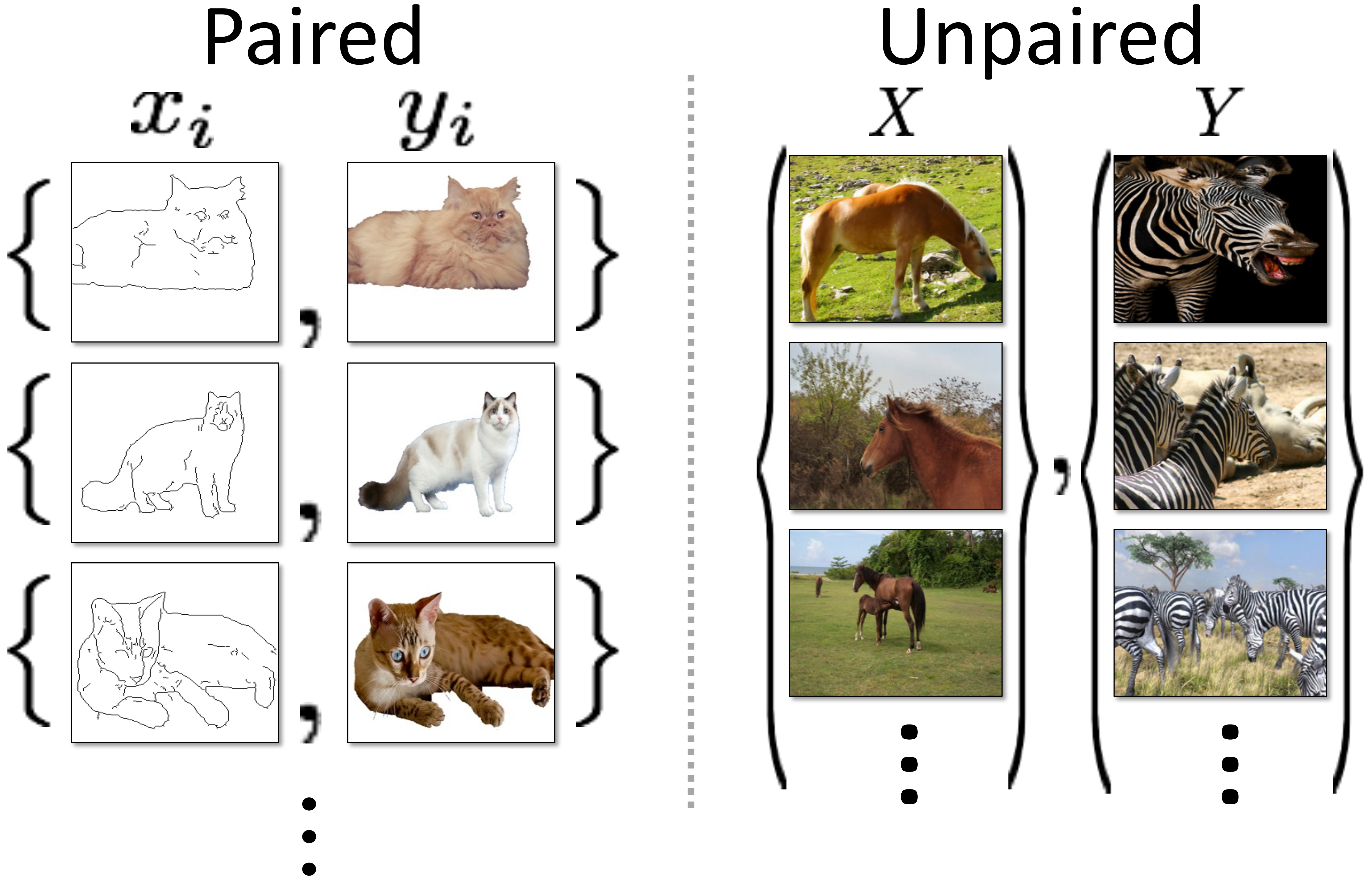


Label ↔ photo: per-pixel labeling

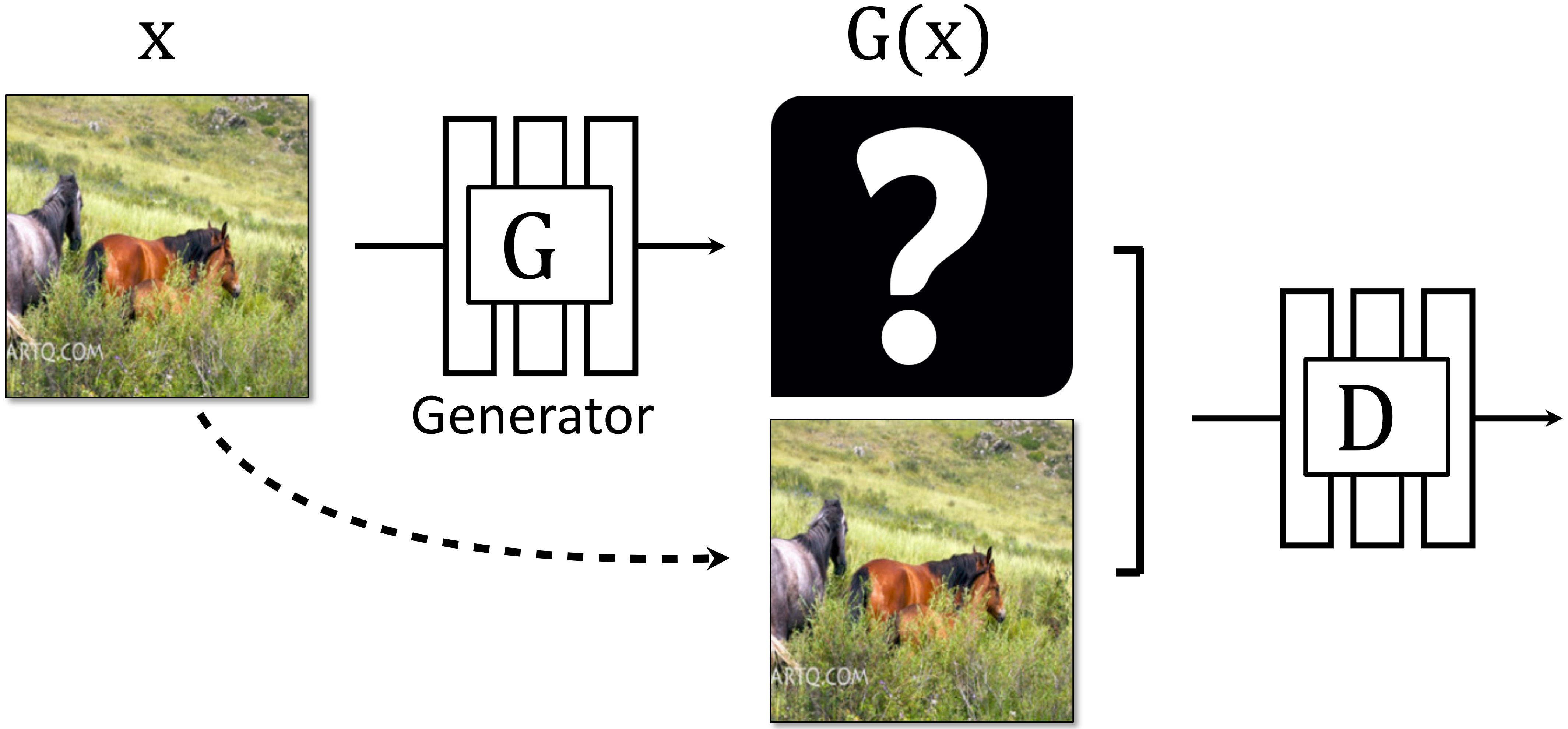


Horse ↔ zebra: how to get zebras?

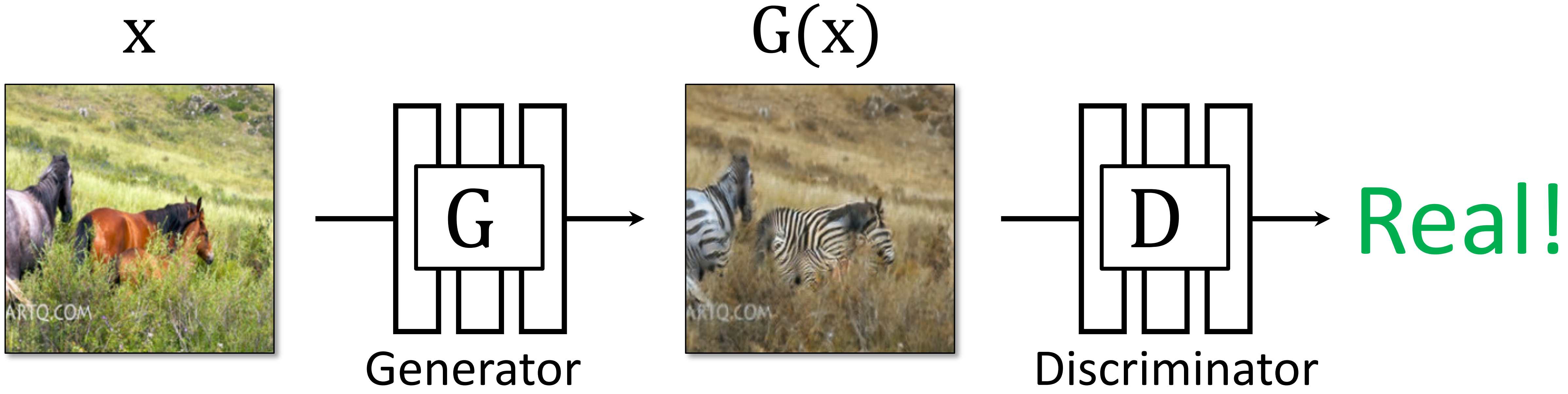
- Expensive to collect pairs.
- Impossible in many scenarios.



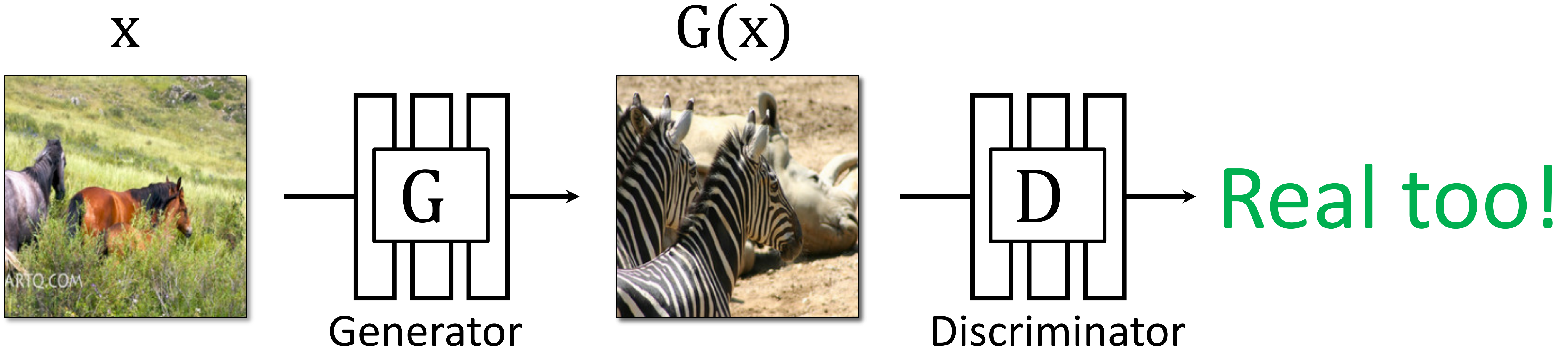
Paired
Unpaired
x_i
y_i
X
Y



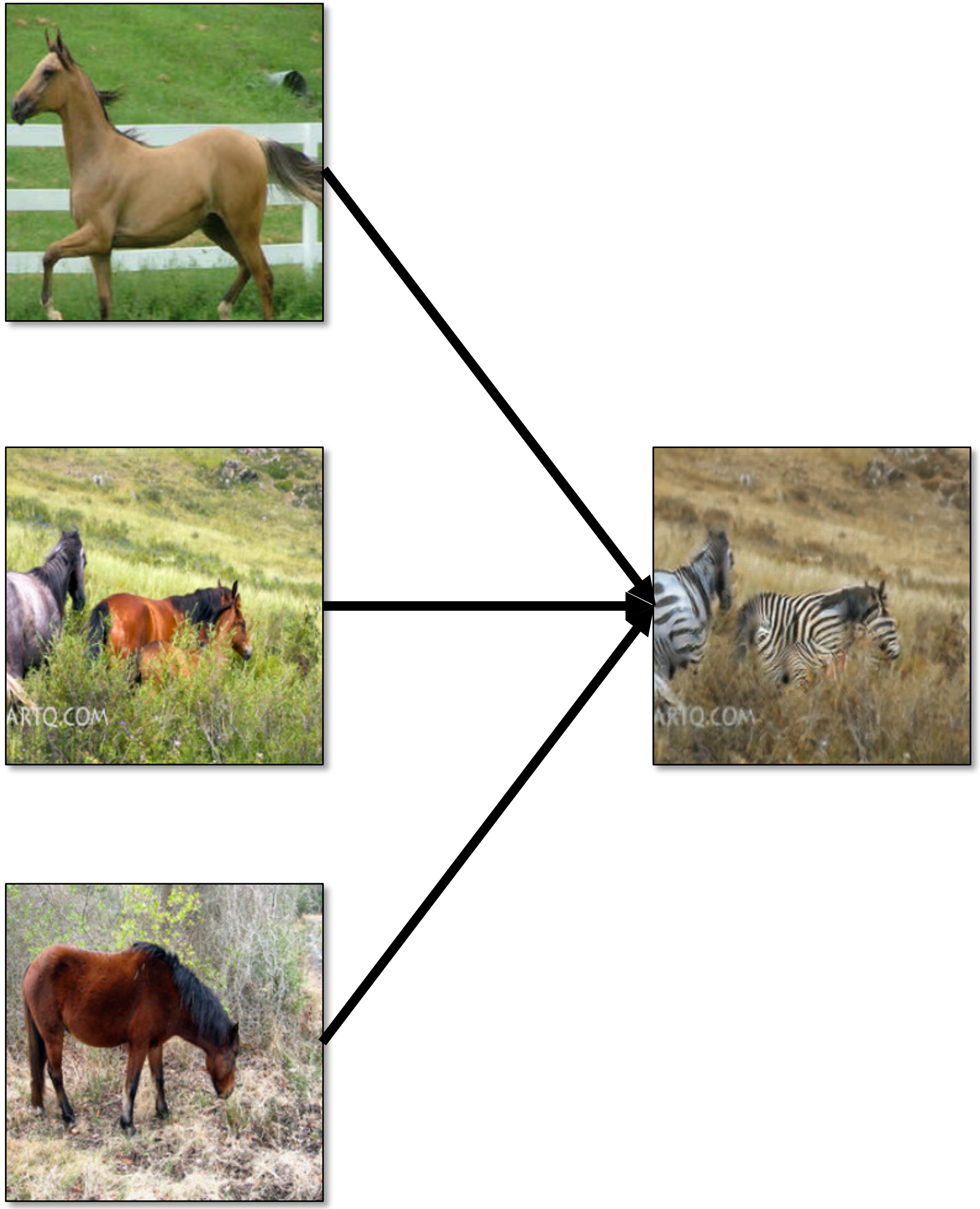
No input-output pairs!



x
G(x)
G
Generator
D
Discriminator
Real!

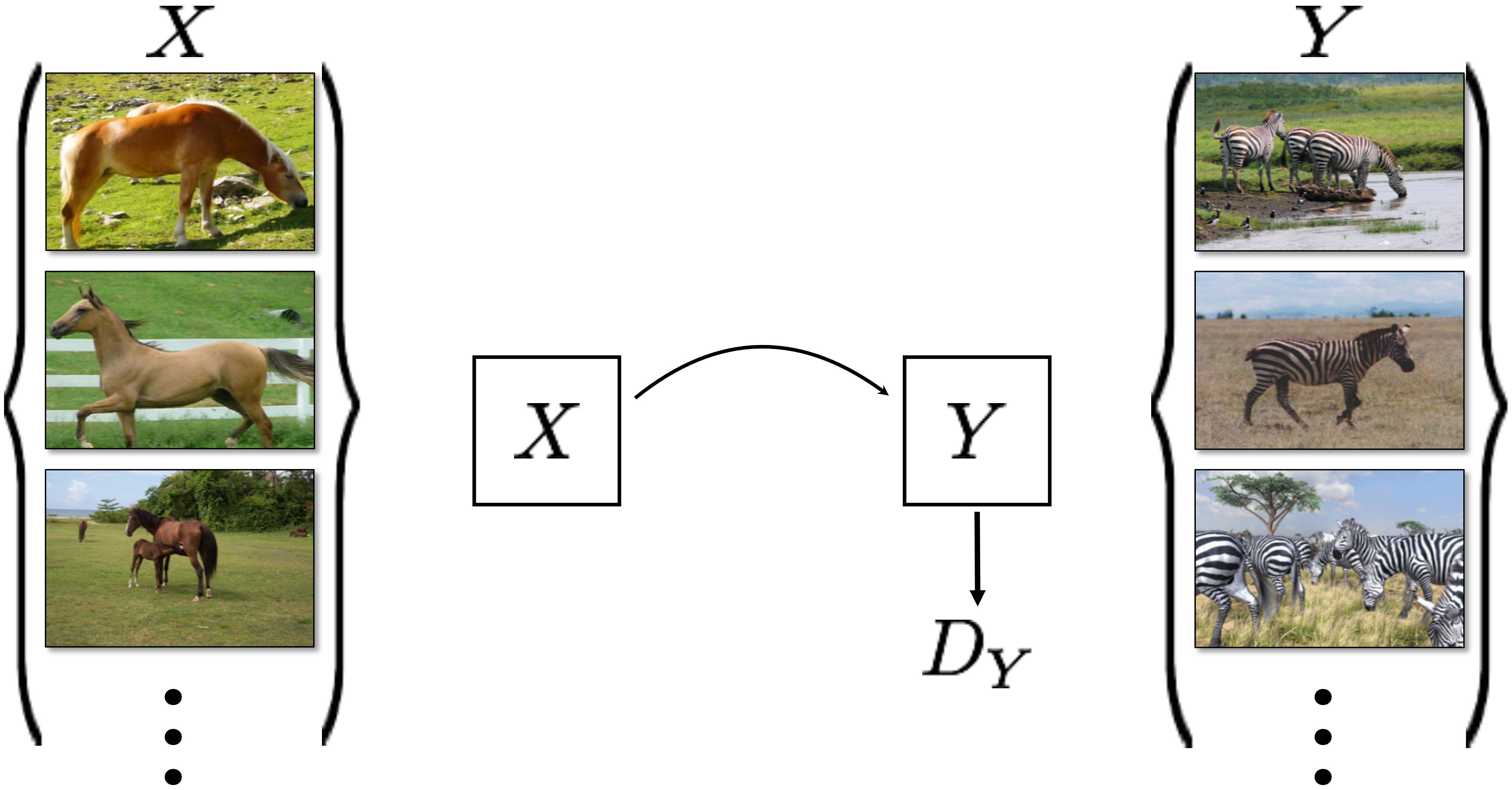


GANs do **not** force output to correspond to input



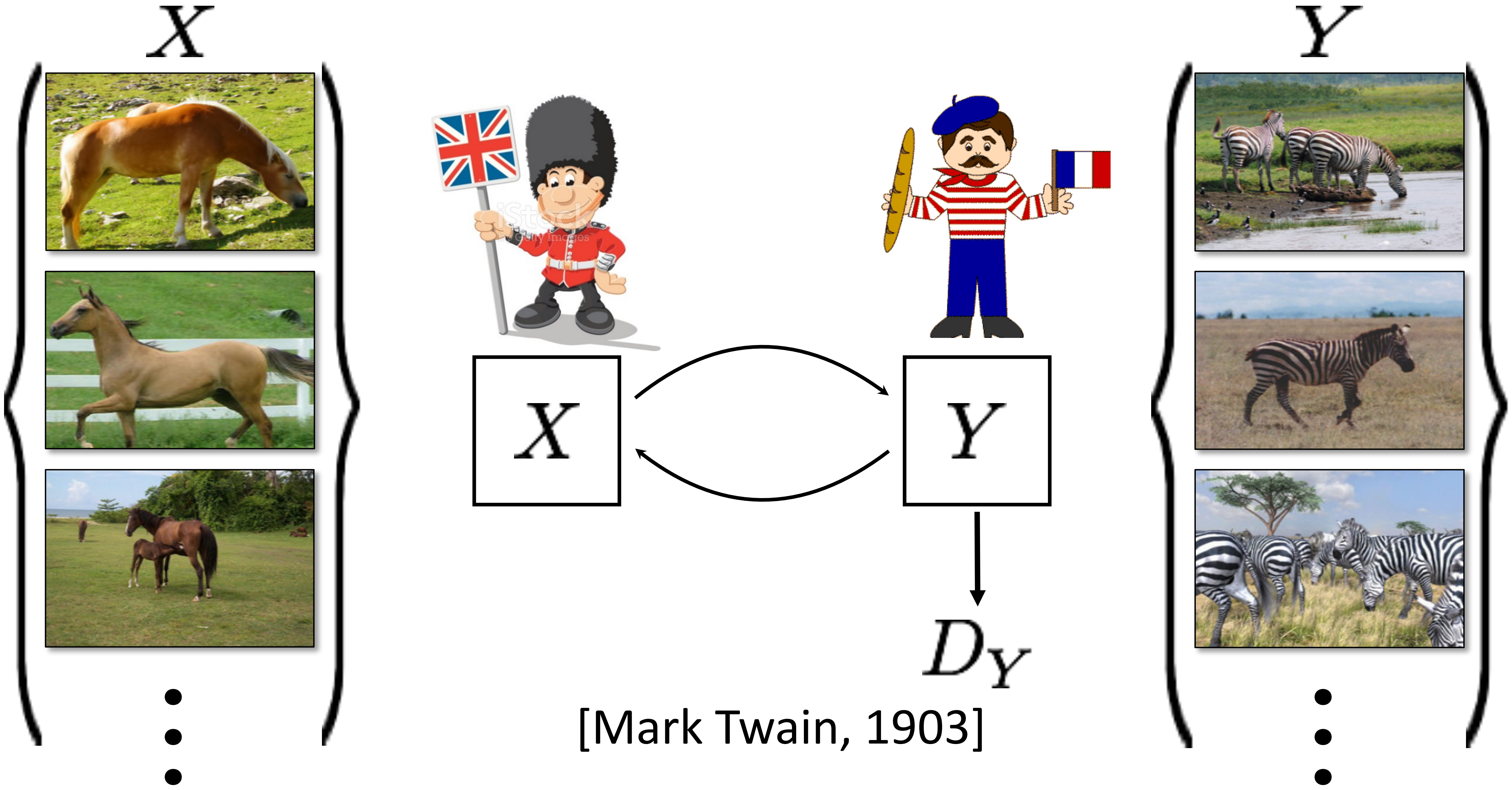
mode collapse!

Cycle-Consistent Adversarial Networks



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

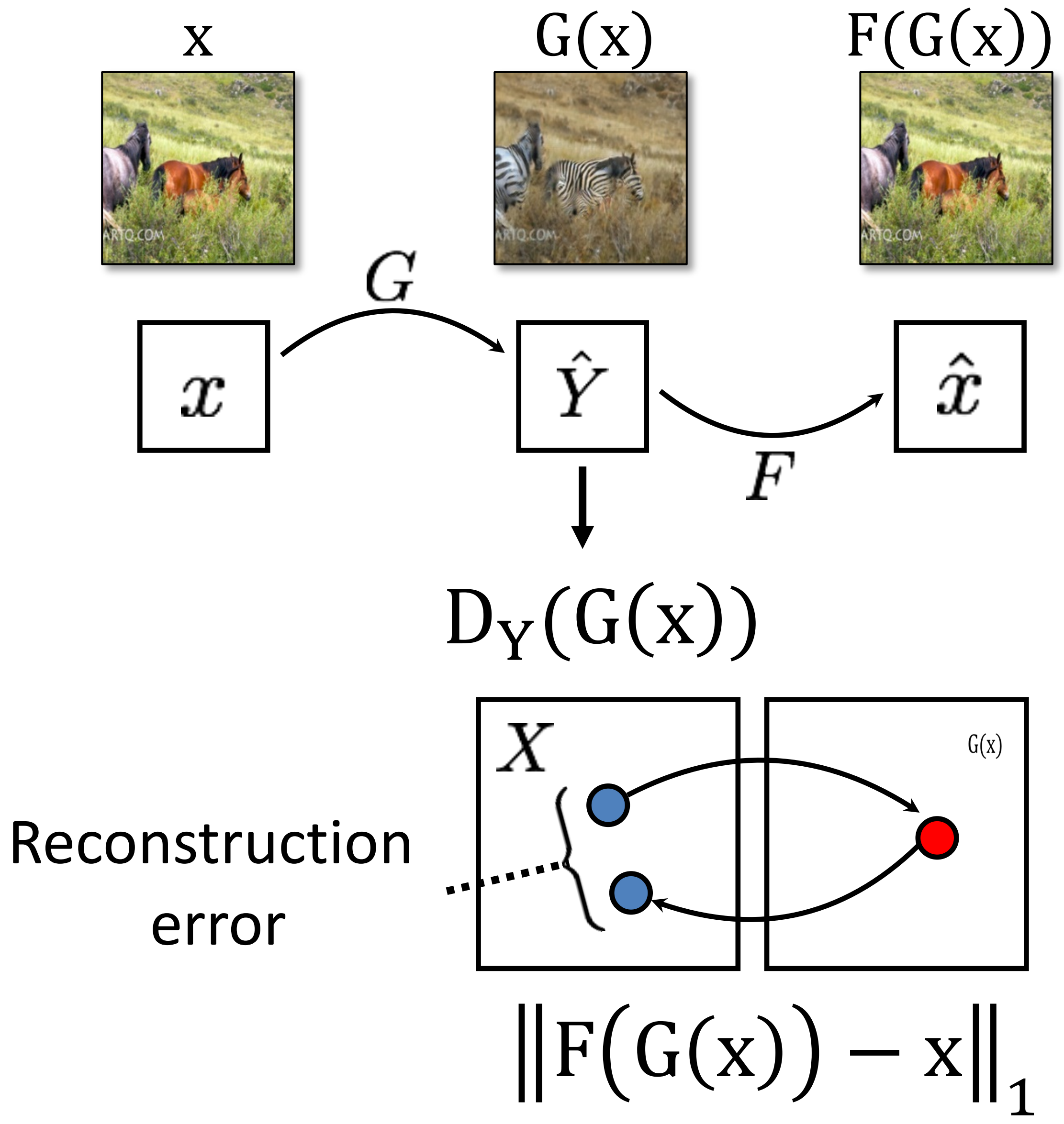
Cycle-Consistent Adversarial Networks



[Mark Twain, 1903]

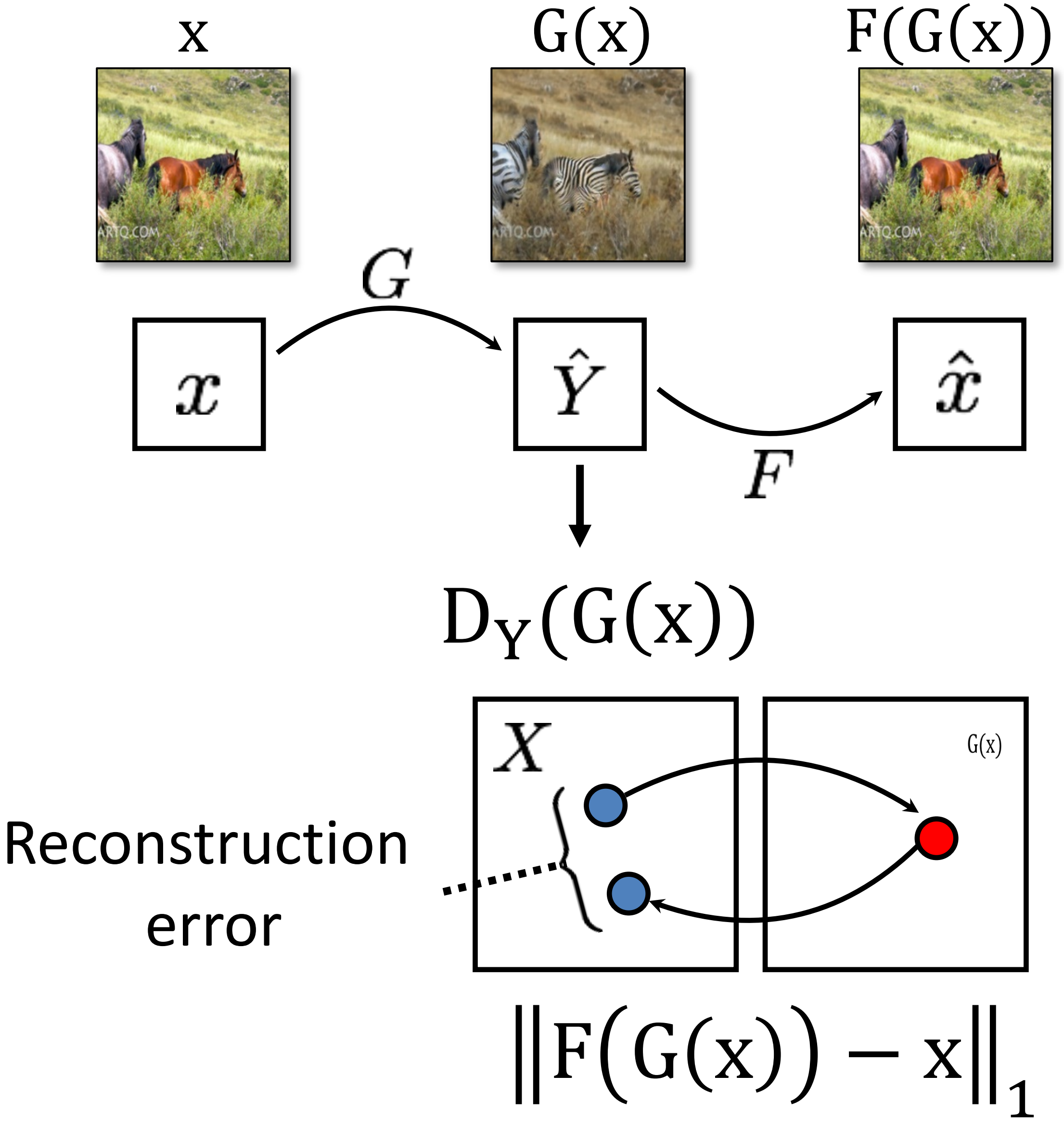
[Zhu*, Park*, Isola, and Efros, ICCV 2017]

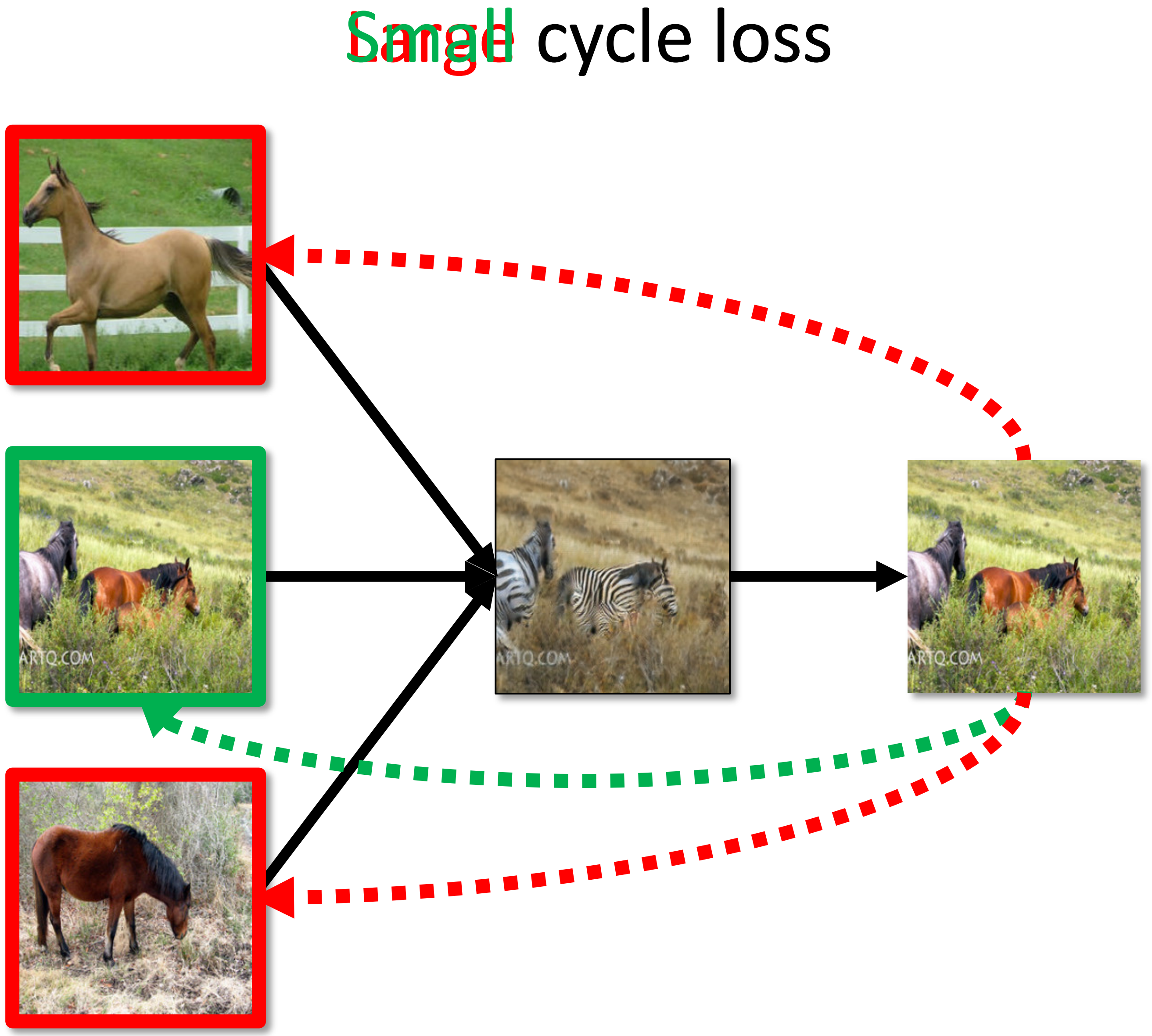
Cycle-Consistent Adversarial Networks



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

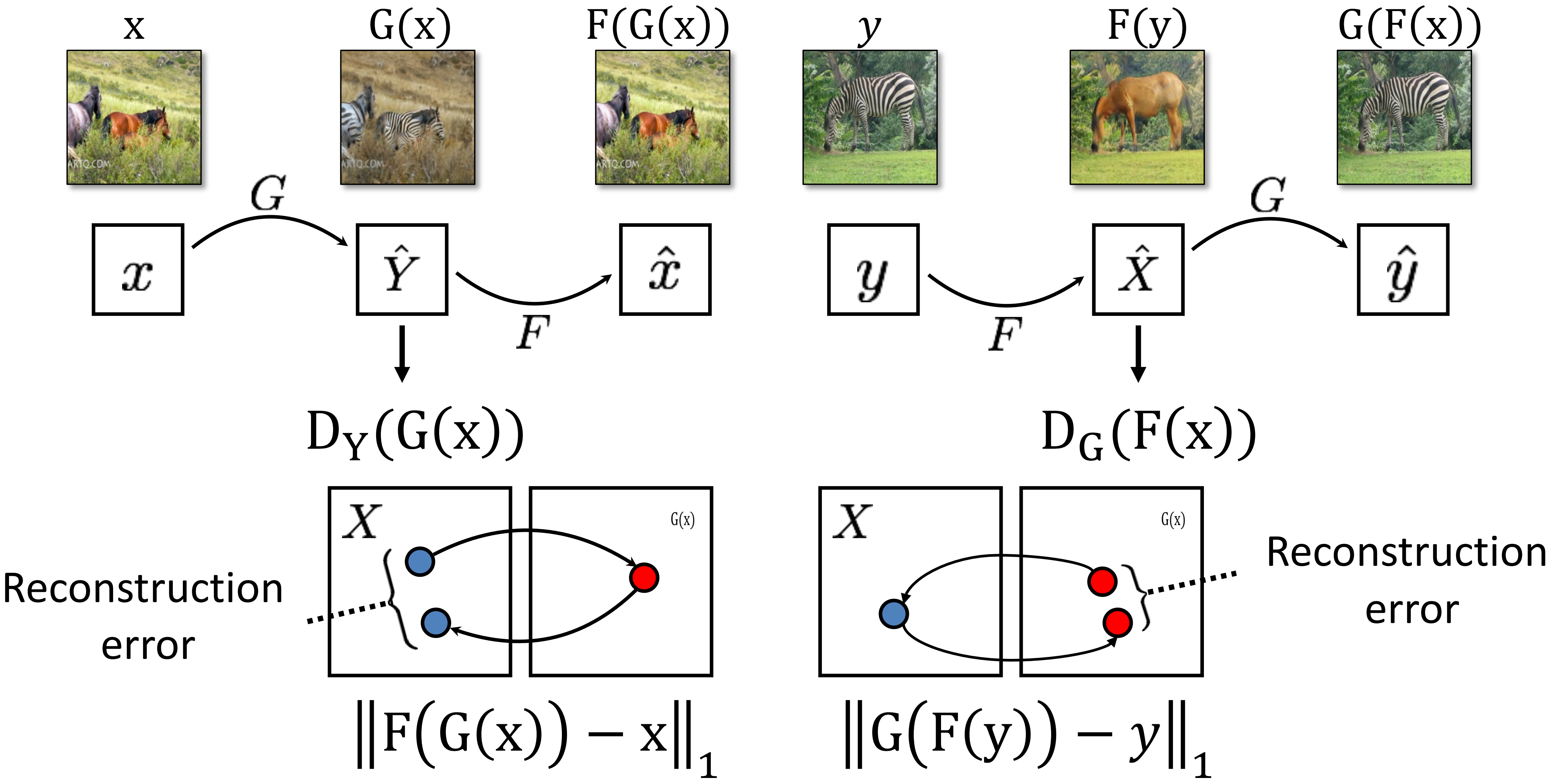
Cycle Consistency Loss





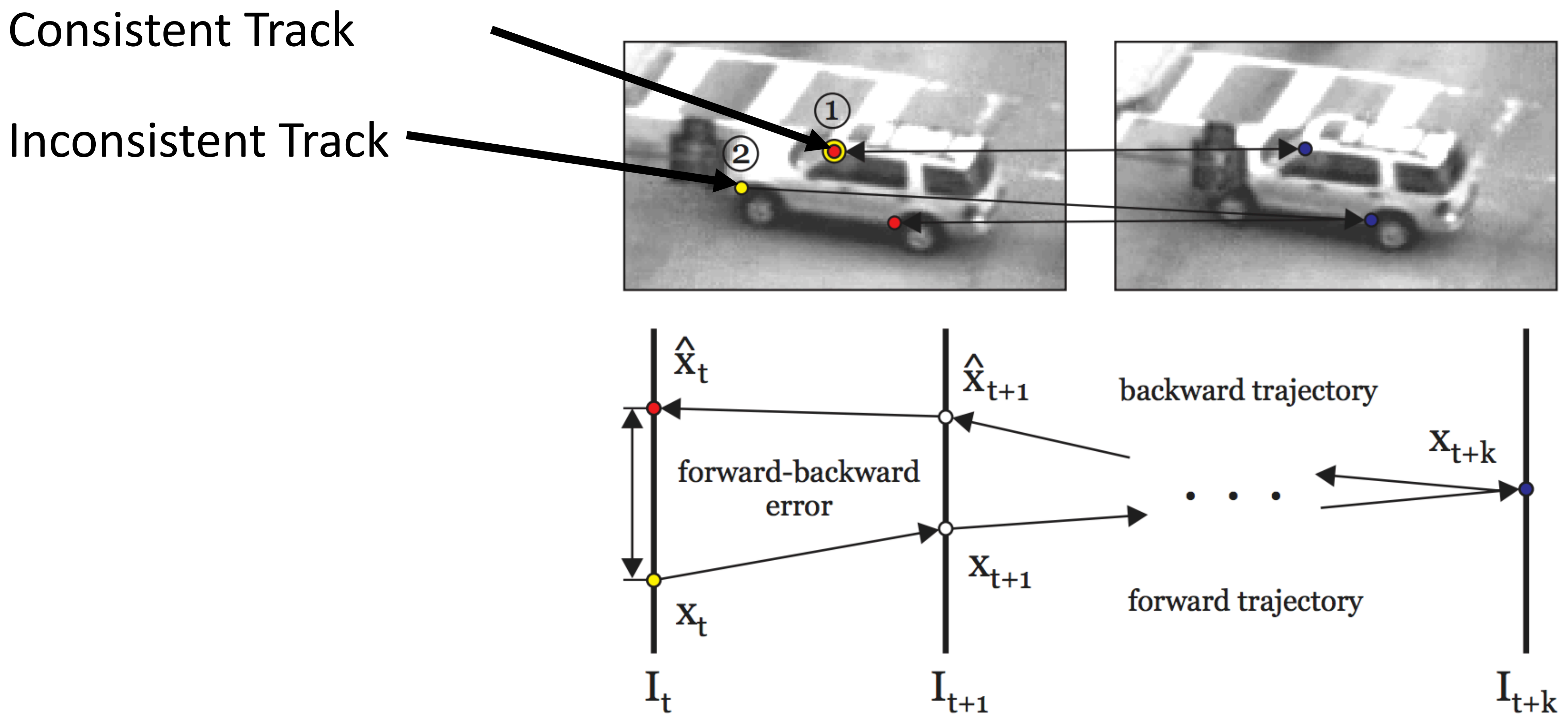
[Zhu*, Park*, Isola, and Efros, ICCV 2017]

Cycle Consistency Loss



See similar formulations [Yi et al. 2017], [Kim et al. 2017] [Zhu*, Park*, Isola, and Efros, ICCV 2017]

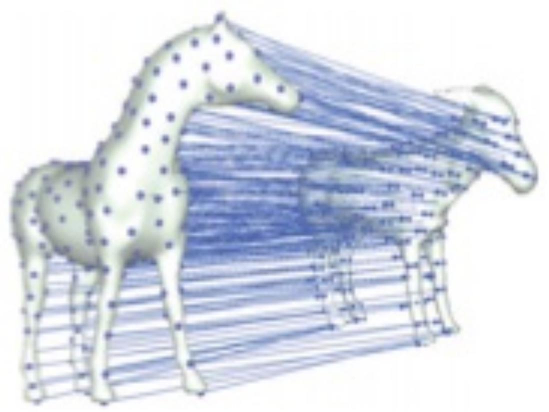
Cycle Consistency in Vision



Forward-Backward Error: Automatic Detection of Tracking Failures. ICPR 10'
Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas.
Also see [Sundaram, Brox, Keutzer, ECCV 10']

Cycle Consistency in Vision

Shape Matching

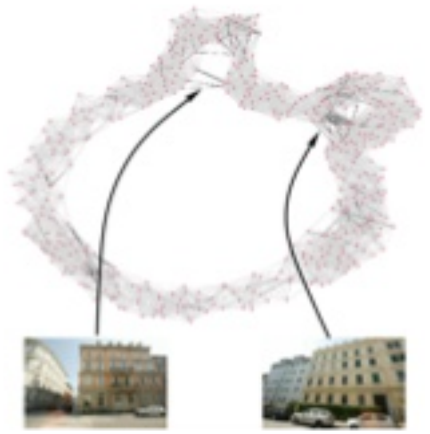


Huang *et al*, SGP'13

Co-segmentation

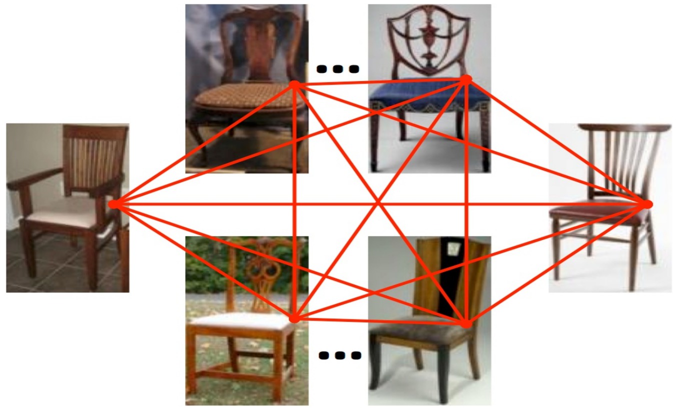
Wang *et al*, ICCV'13

SfM

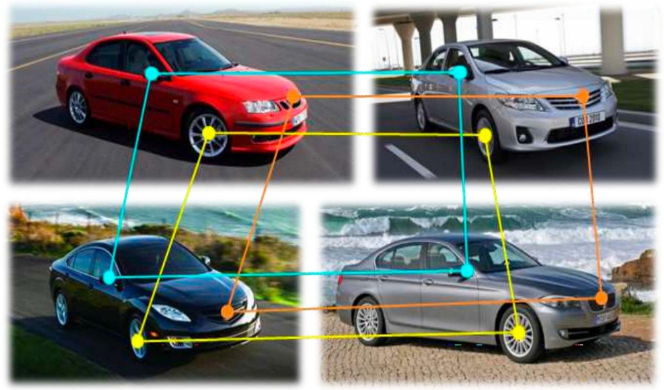


Zach *et al*, CVPR'10

Collection Correspondence



Zhou *et al*, CVPR'15



Zhou *et al*, ICCV'15

slides credit @Tinghui Zhou

Results

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 photo→labels


Collection Style Transfer

Photograph
@ Alexei Efros

Monet

Van Gogh

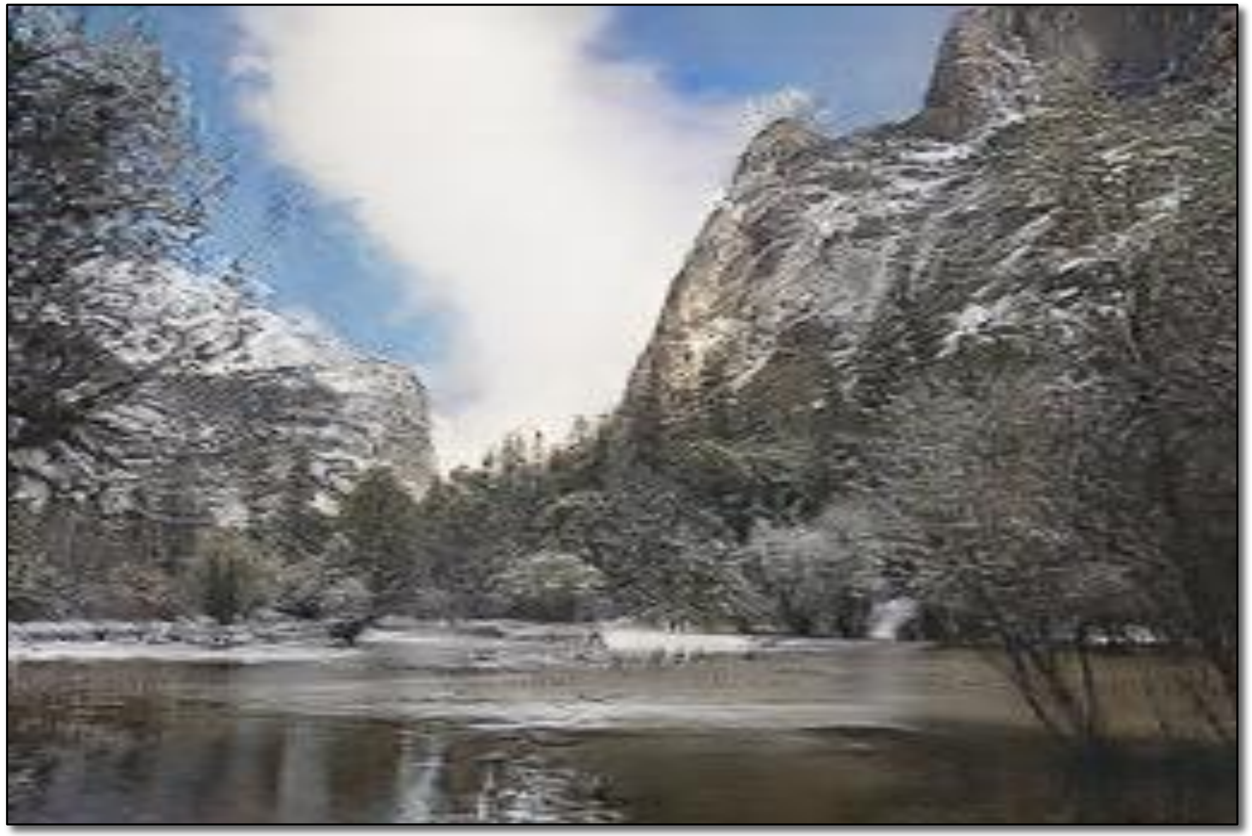
Cezanne

Ukiyo-e


Input
Monet
Van Gogh
Cezanne
Ukiyo-e

Monet's paintings → photos

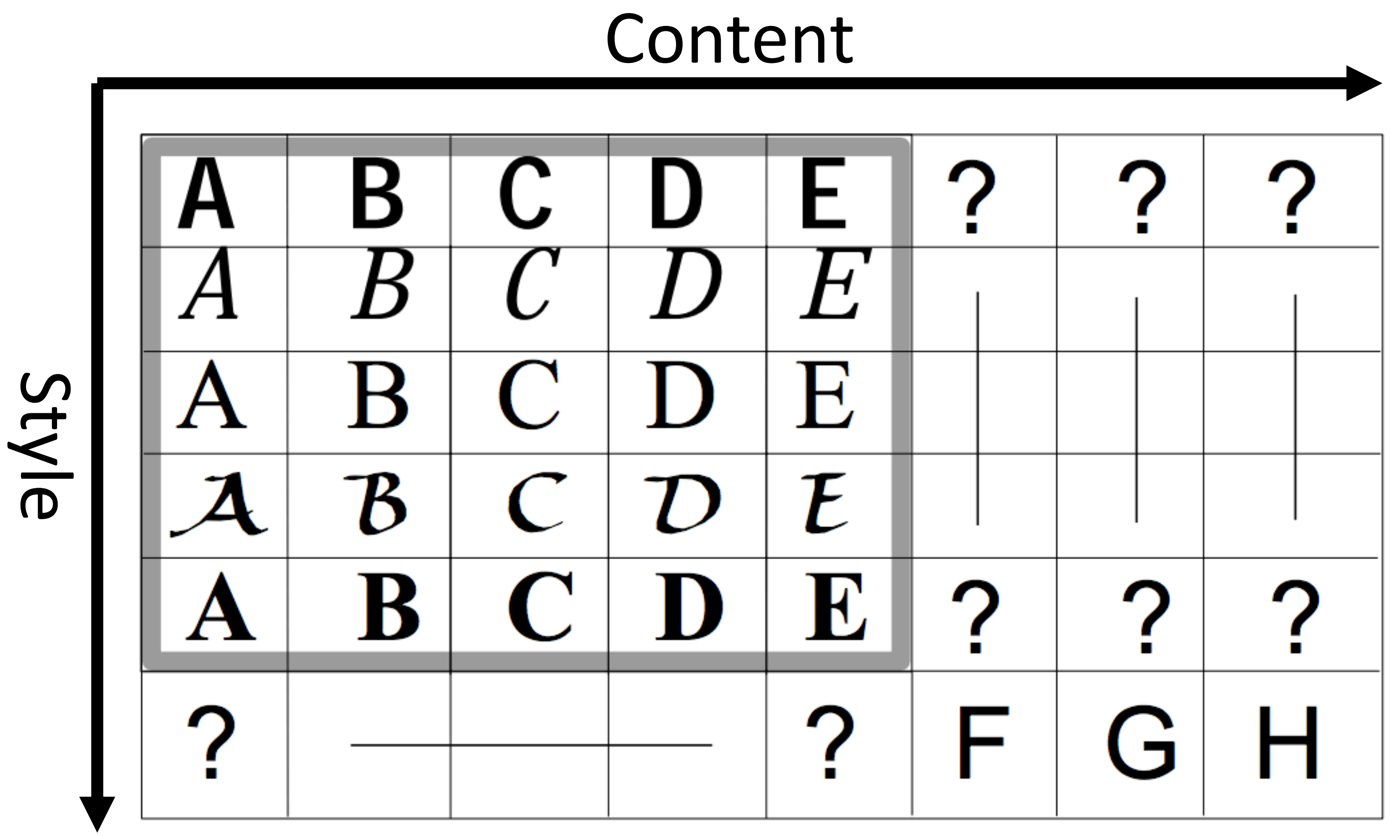
Monet’s paintings → photos

Why CycleGAN works

Style and Content Separation

Paired Separation



Separating Style and Content with
Bilinear Models
[Tenenbaum and Freeman 2000']

Unpaired Separation

Adversarial Loss: change the style

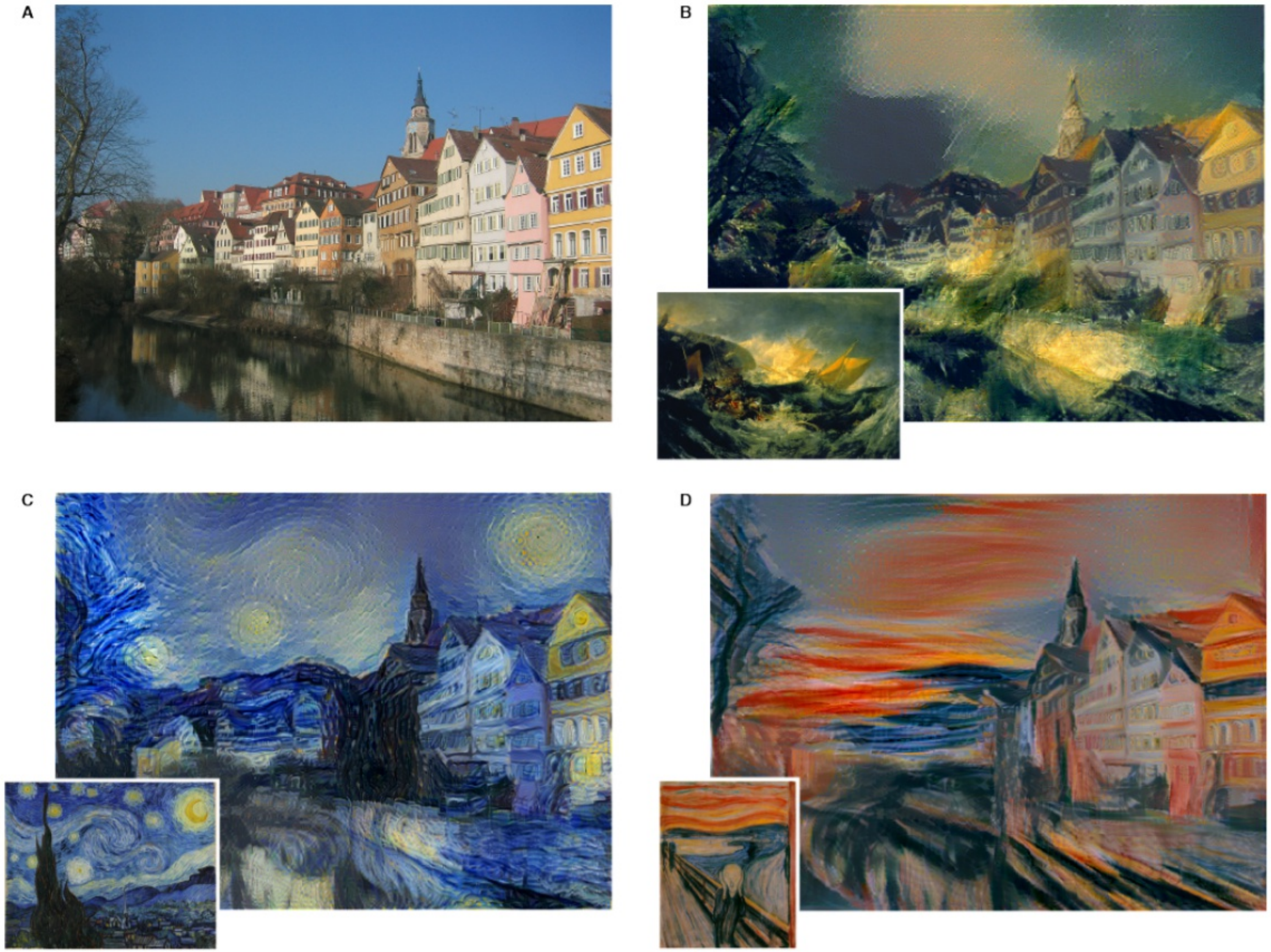
$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) =& \mathbb{E}_{y \sim p_{\text{data}}(y)}[\log D_Y(y)] \\ &+ \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x))]\end{aligned}$$

Cycle Consistency Loss: preserve the content

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G, F) =& \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] \\ &+ \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1].\end{aligned}$$

Two empirical assumptions:
- content is easy to keep.
- style is easy to change.

Neural Style Transfer [Gatys et al. 2015]



Style and Content:

- Content: feature difference
- Style: Gram Matrix difference
- Both losses are hard-coded.

PRISMA


Input
Style Image I
Style image II
Entire collection
CycleGAN
Photo → Van Gogh
Input
Style image I
Style image II
Entire collection
CycleGAN
horse → zebra

Applications

CG2Real: GTA5 → real streetview

GTA5 CG Input

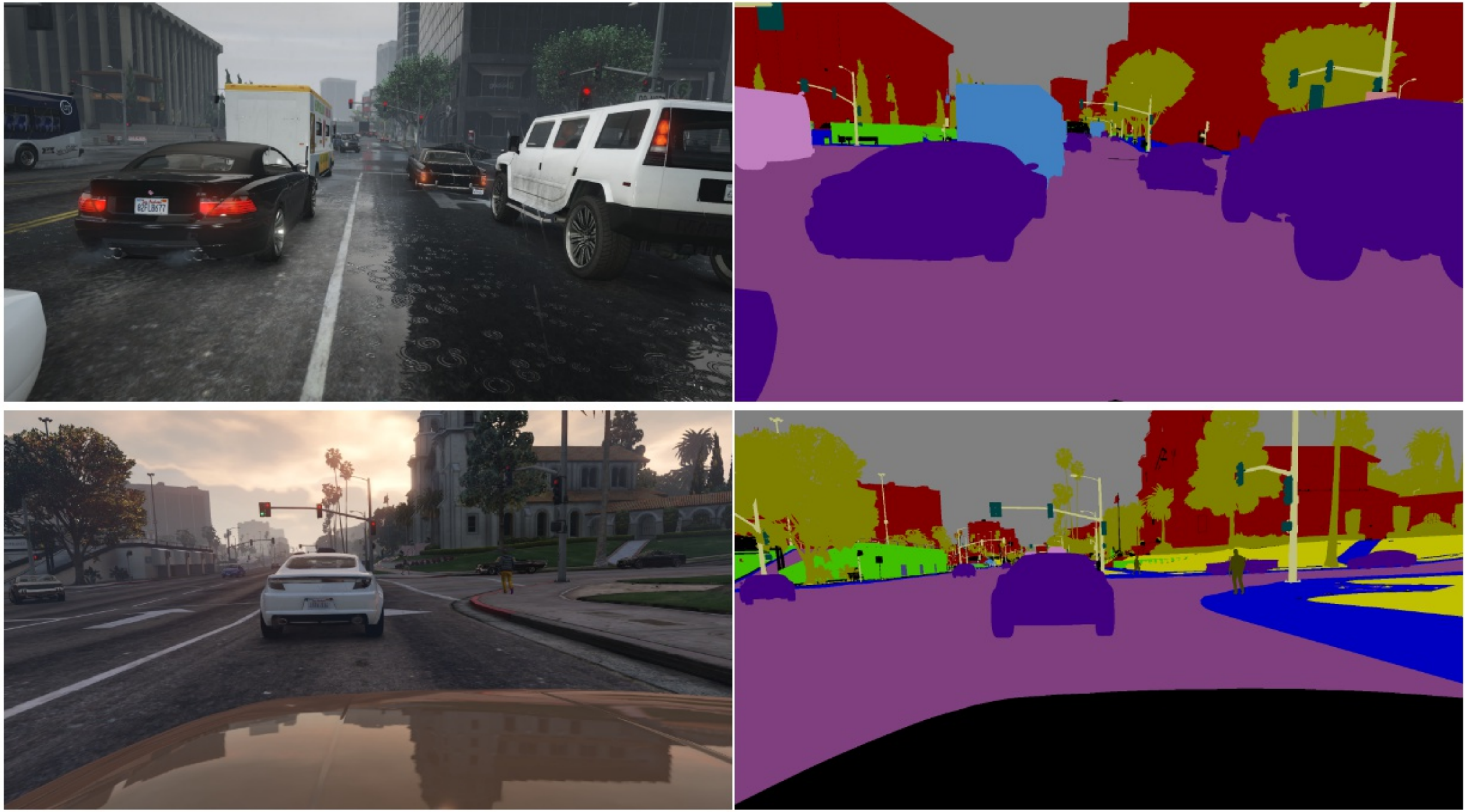
Output

Inspired by [Johnson et al. 2011]

Real2CG: real streetview → GTA

Cityscape Input　　　　Output

Synthetic Data as Supervision



GTA5 images | Segmentation labels

[Richter*, Vineet* et al. 2016] [Krähenbühl et al. 2018]

Domain Adaptation with CycleGAN

Train on GTA5 data

Test on real images

	meanIOU	Per-pixel accuracy
Oracle (Train and test on Real)	60.3	93.1
Train on CG, test on Real	17.9	54.0

See Judy Hoffman's talk at 14:30 "Adversarial Domain Adaptation"

Domain Adaptation with CycleGAN

GTA5 data + Domain adaptation

Test on real images

	meanIOU	Per-pixel accuracy
Oracle (Train and test on Real)	60.3	93.1
Train on CG, test on Real	17.9	54.0
FCN in the wild [Previous STOA]	27.1	-

See Judy Hoffman's talk at 14:30 "Adversarial Domain Adaptation"

Domain Adaptation with CycleGAN

Train on CycleGAN data

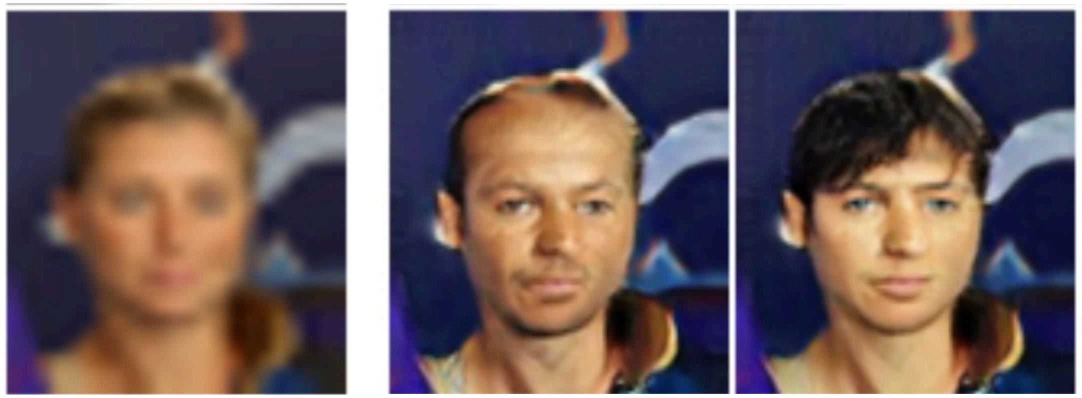
Test on real images

	meanIOU	Per-pixel accuracy
Oracle (Train and test on Real)	60.3	93.1
Train on CG, test on Real	17.9	54.0
FCN in the wild [Previous STOA]	27.1	-
Train on CycleGAN, test on Real	**34.8**	**82.8**

See Judy Hoffman's talk at 14:30 "Adversarial Domain Adaptation"

Applications and Extentions

Attribute Editing [Lu et al.]



Low-res Bald Bangs

arXiv:1705.09966

Object Editing [Liang et al.]

Mask Input Output

arXiv:1708.00315

Front/Character Transfer [Ignatov et al.]

Input output

arXiv: 1801.08624

Data generation [Wang et al.]

samples by CycleWGAN

arXiv:1707.03124

Image Dehazing

Cycle-Dehaze: Enhanced CycleGAN for Single Image Dehazing. CVPRW 2018
Deniz Engin∗ Anıl Genc∗, Hazım Kemal Ekenel

Manipulating Natural Scenes*

Manipulating Attributes of Natural Scenes via Hallucination

LEVENT KARACAN, Hacettepe University and Iskenderun Technical University, Turkey
ZEYNEP AKATA, University of Tübingen, Germany
AYKUT ERDEM and ERKUT ERDEM, Hacettepe University, Turkey

Fig. 1. Given a natural image, our approach can hallucinate different versions of the same scene in a wide range of conditions, e.g., *night*, *sunset*, *winter*, *spring*, *rain*, *fog*, or even a combination of those. First, we utilize a generator network to imagine the scene with respect to its semantic layout and the desired set of attributes. Then, we directly transfer the scene characteristics from the hallucinated output to the input image, without the need for a reference style image.

In this study, we explore building a two-stage framework for enabling users to directly manipulate high-level attributes of a natural scene. The key to our approach is a deep generative network that can hallucinate images of a scene as if they were taken in a different season (e.g., during winter), weather condition (e.g., on a cloudy day), or at a different time of the day (e.g., at sunset). Once the scene is hallucinated with the given attributes, the corresponding look is then transferred to the input image while preserving the semantic details intact, giving a photo-realistic manipulation result. As the proposed framework hallucinates what the scene will look like, it does not require any reference style image as commonly utilized in most of the appearance or style transfer approaches. Moreover, it allows to simultaneously manipulate a given scene according to a diverse set of transient attributes within a single model, eliminating the need of training multiple networks per each translation task. Our comprehensive set of qualitative and quantitative results demonstrates the effectiveness of our approach against the competing methods.

CCS Concepts: • **Computing methodologies → Neural networks**; **Image manipulation**; **Image representations**;

Additional Key Words and Phrases: Image generation, style transfer, generative models, visual attributes

ACM Reference format:
Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. 2019. Manipulating Attributes of Natural Scenes via Hallucination. *ACM Trans. Graph.* 39, 1, Article 7 (November 2019), 17 pages.
https://doi.org/10.1145/3368312

1 INTRODUCTION

"The trees, being partly covered with snow, were outlined indistinctly against the grayish background formed by a cloudy sky, barely whitened by the moon."

—Honore de Balzac (Sarrasine, 1831)

This work was supported in part by TUBA GEBIP fellowship awarded to E. Erdem. We would like to thank NVIDIA Corporation for the donation of GPUs used in this research. This work has been partially funded by the DFG-EXC-Nummer 2064/1-Projektnummer 390727645.
Authors' addresses: L. Karacan, Hacettepe University, Ankara, Turkey and Iskenderun Technical University, Hatay, Turkey; email: karacan@cs.hacettepe.edu.tr; Z. Akata, University of Tübingen, Tubingen, Germany; email: zeynep.akata@uni-tuebingen.de; A. Erdem and E. Erdem, Hacettepe University, Ankara, Turkey; emails: {aykut, erkut}@cs.hacettepe.edu.tr.

0730-0301/2019/11-ART7 $15.00
https://doi.org/10.1145/3368312

* Joint work with Levent Karacan, Zeynep Akata and Aykut Erdem

What does this scene look like on a cloudy day?

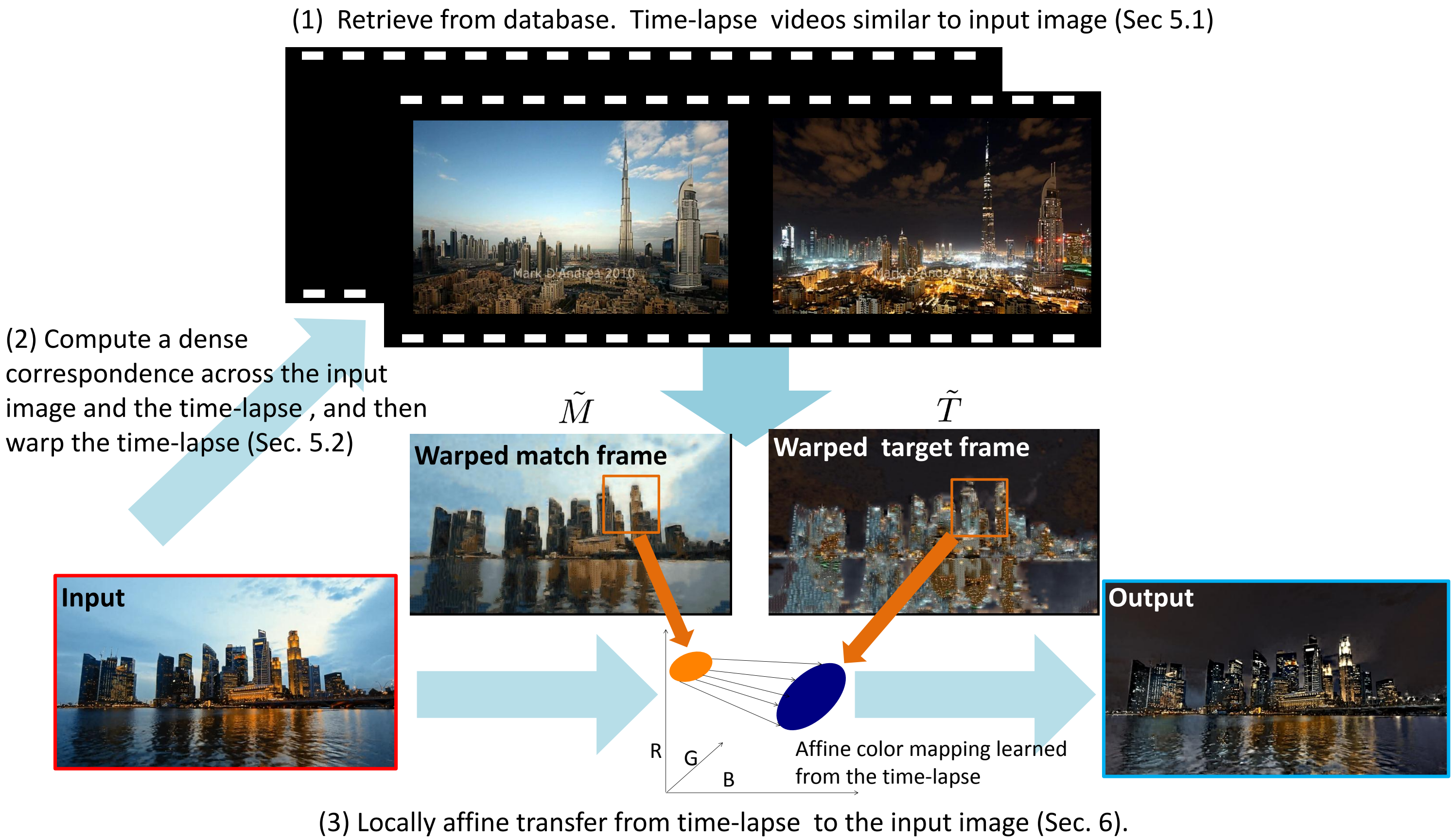

... like this.

Solving this problem

- requires to learn the relation between transient scene attributes and scene elements
- demands for a dataset suitable for this task

Related Work – Attribute Manipulation

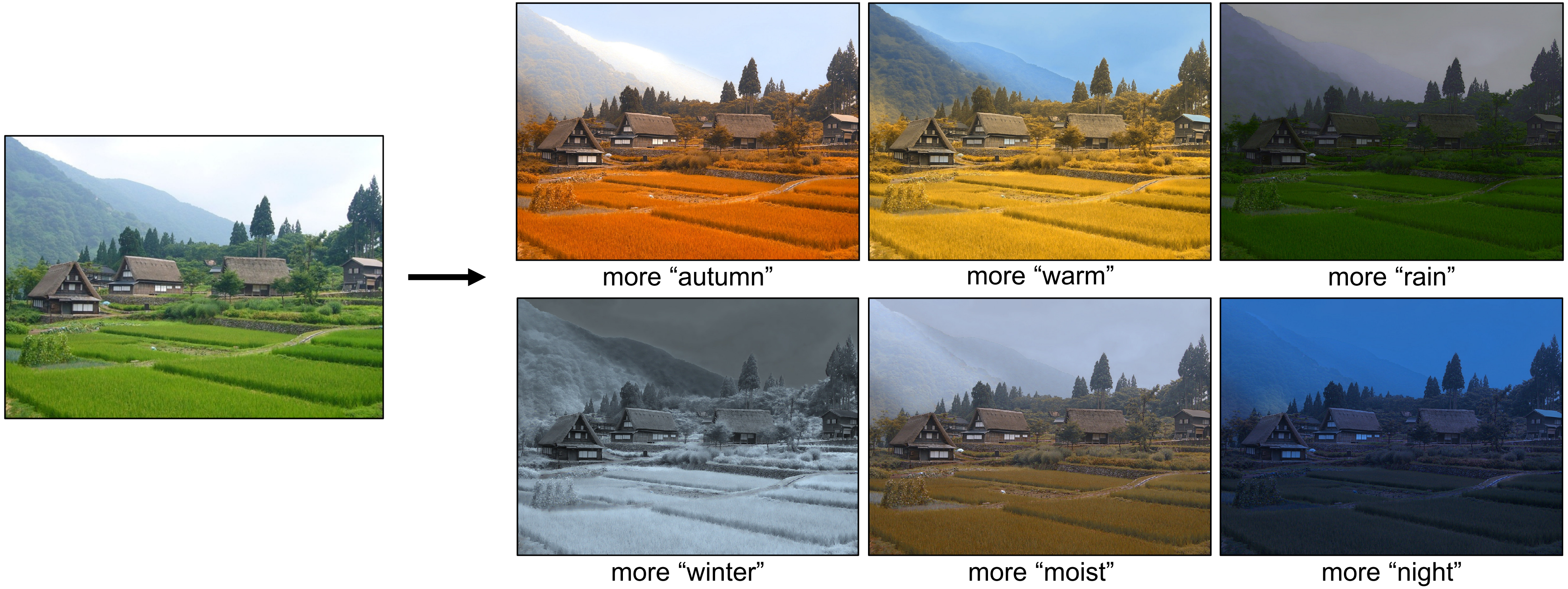
- Different times of a day [Shih et al., 2013]



- An examplar-based local appearance transfer approach

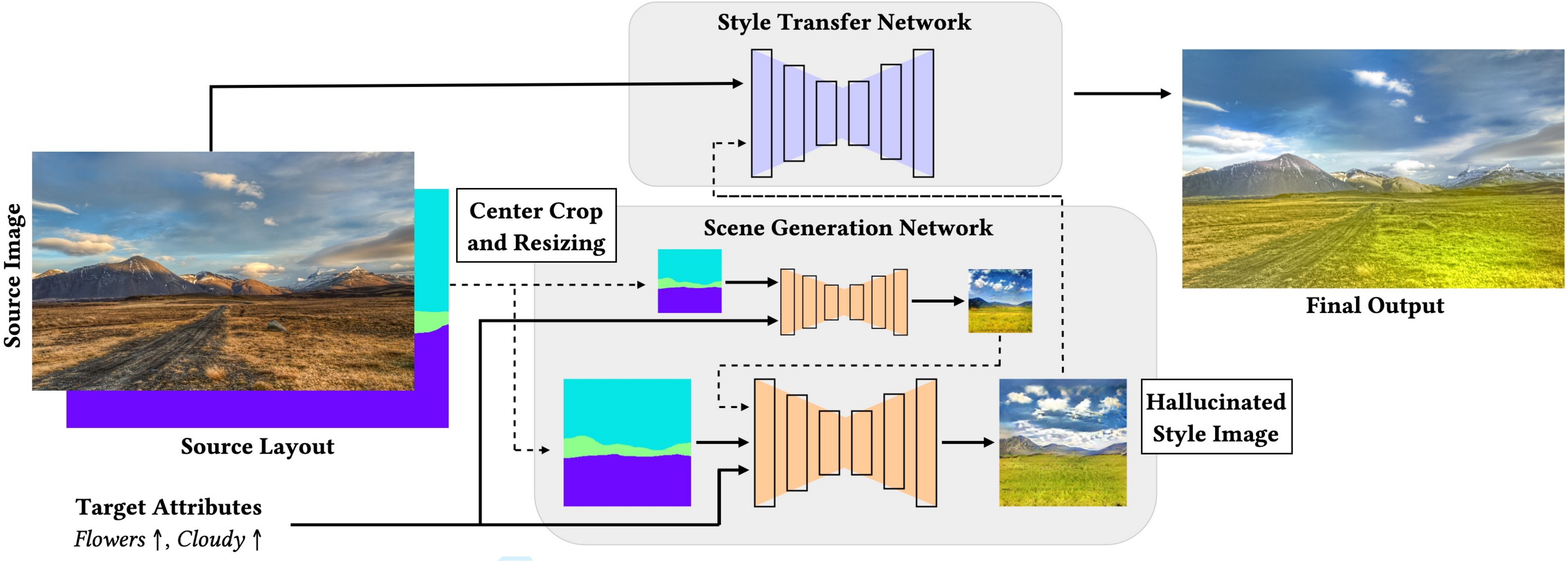
Related Work– Attribute Manipulation

- Editing scene attributes, [Laffont et al., 2014]

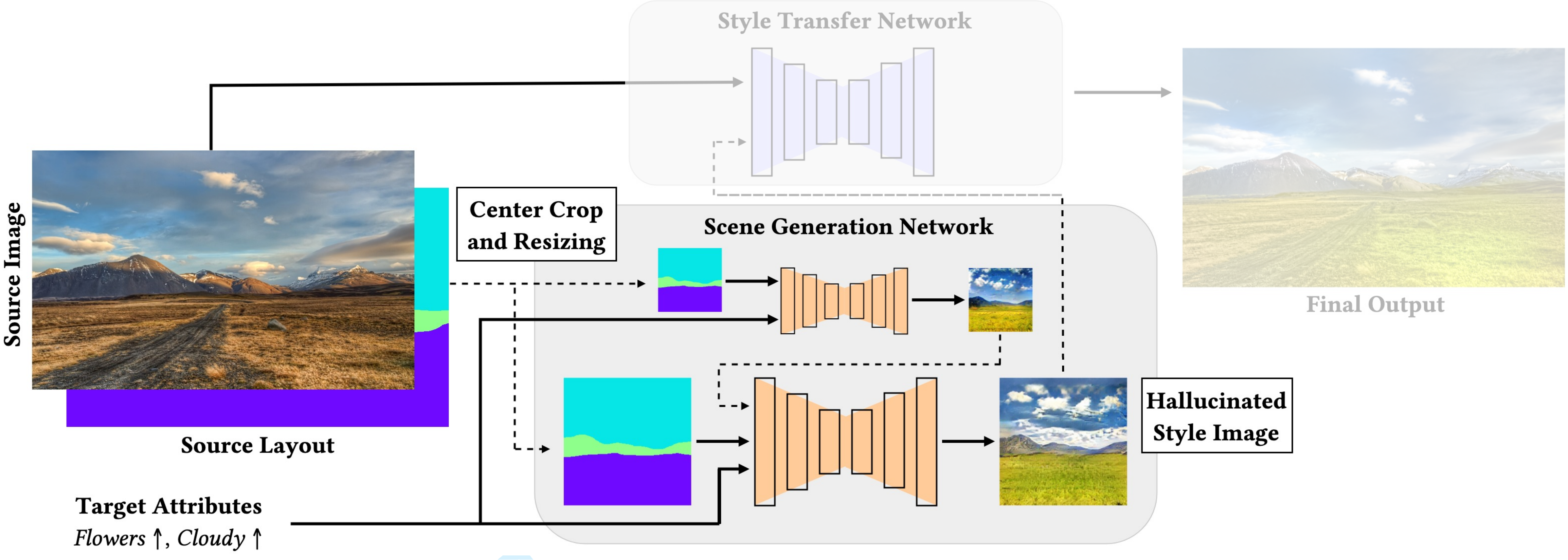


- An examplar-based local appearance transfer approach

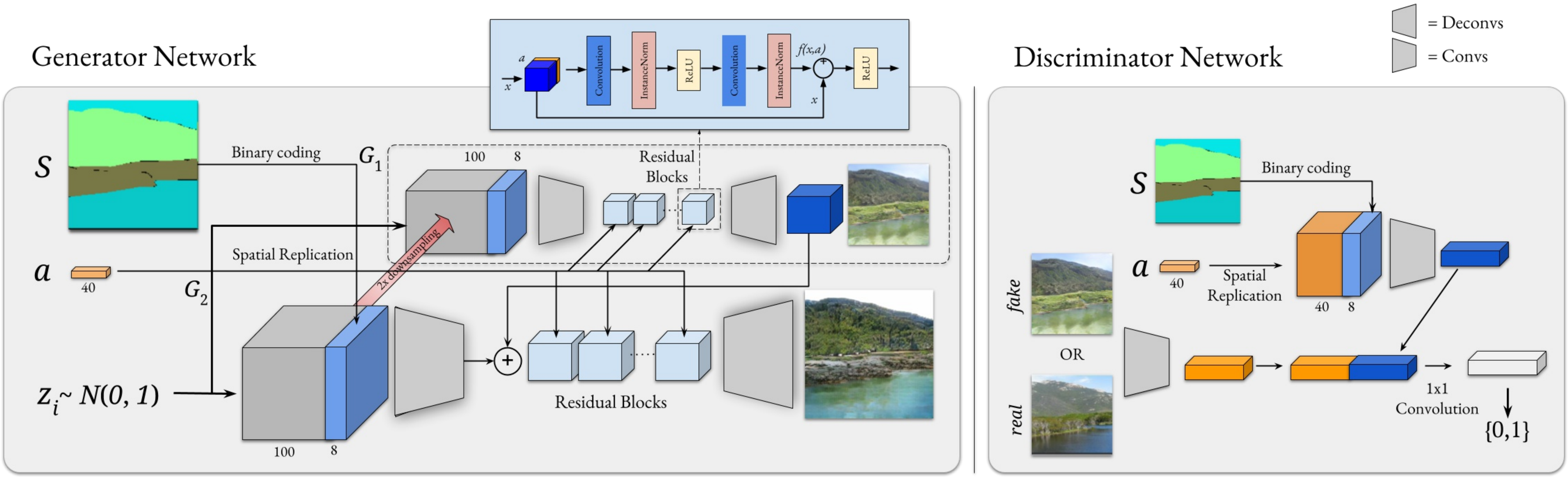
Proposed Framework



Proposed Framework

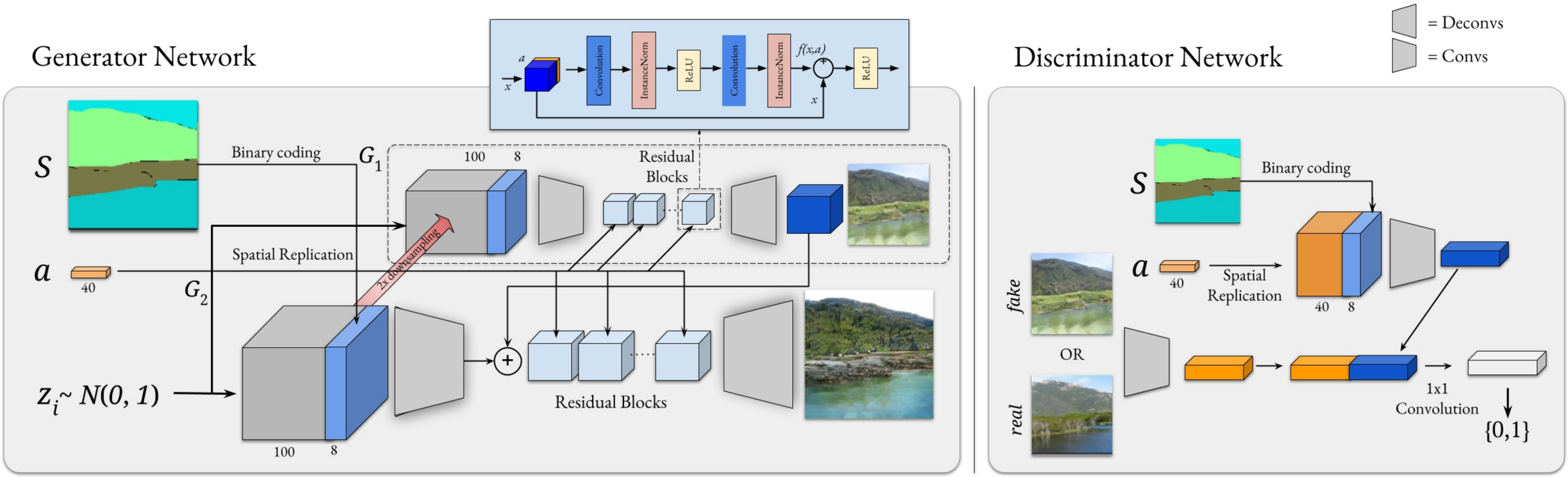


Scene Generation Network (SGN)



- A multiscale strategy similar to that in Pix2pixHD [Wang et al. 2018]
- Generator network: a coarse-scale and a fine-scale generator subnets
- Discriminator: 3 discriminator subnets that operate at 3 different image scales

Scene Generation Network (SGN)



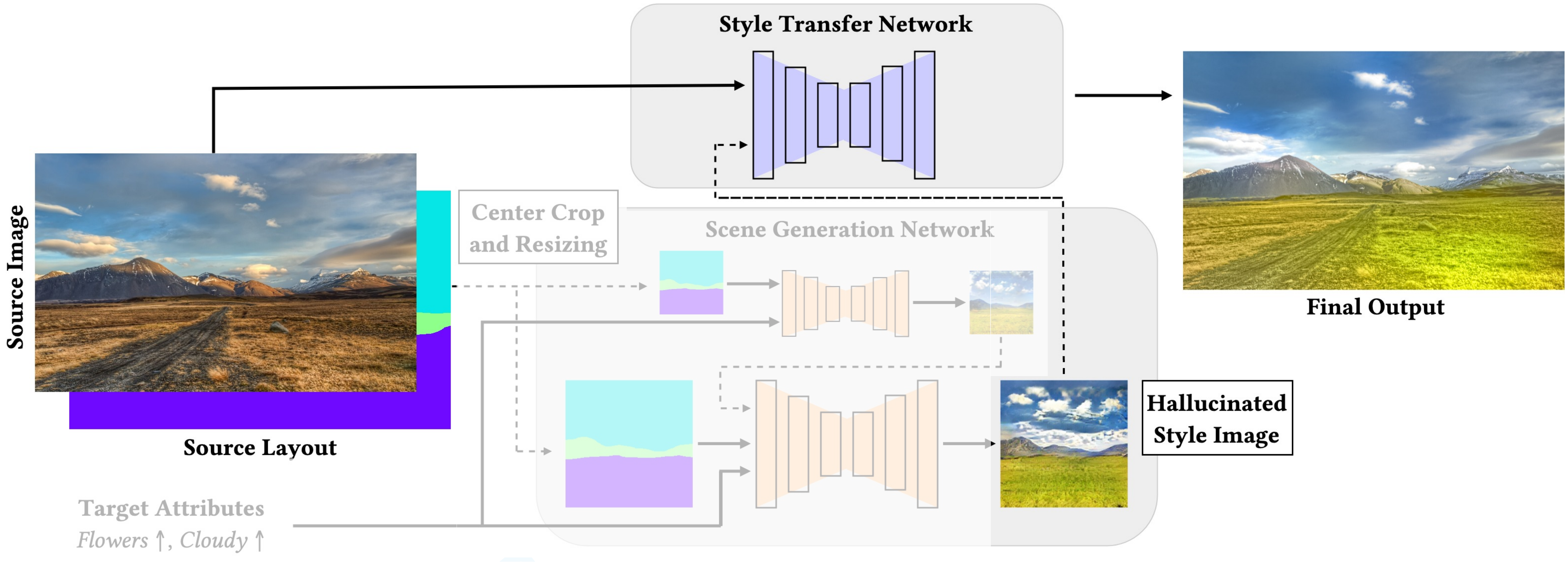
- Our multi-scale generator network consists of a coarse-scale generator and a fine-scale generator.
- Our multi-scale discriminator includes 3 different discriminators with similar network structures that operate at 3 different image scales.

Improved Training of SGNs

- Relative Negative Mining (RNM)
  - A "real pair" (real image paired with right conditions) should score higher than a "fake pair" (either image is fake or context information mismatches)
  - During training SGN, sample mismatching layouts as well.

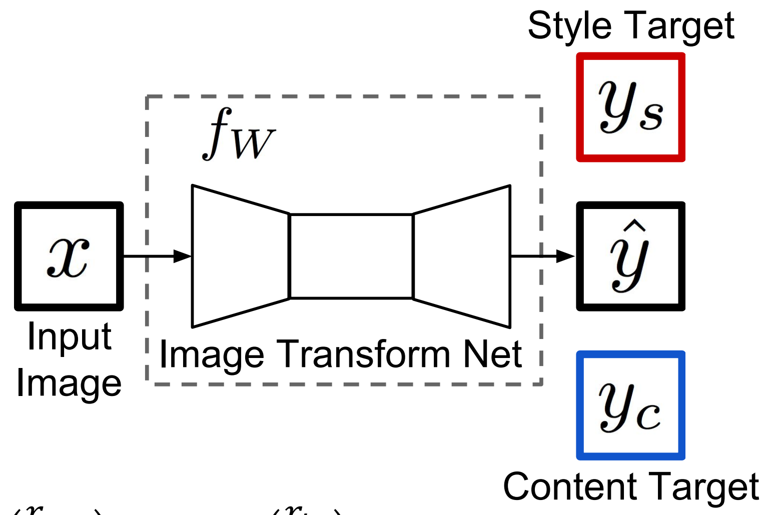
- Layout-Invariant Perceptual Loss
  - $\mathbb{E}_P = \mathbb{E}_{z \sim p_z(z); x,s,a \sim p_{data}(x,s,a)} \left[\left\| f_P(x) - f_P\big(G(z,s,a)\big) \right\|_2^2 \right]$
  - f is the CNN encoder for the scene parser network [Zhou et al., 2018]

Proposed Framework



Style Transfer Network

- DPST [Luan et al., 2017]
  - semantic segmentation to avoid content mismatch (transfer statistics within each category)
  - locally affine model as a photorealism regularization



$$\begin{pmatrix} r_{out} \\ g_{out} \\ b_{out} \end{pmatrix} = A_{3\times3} \begin{pmatrix} r_{in} \\ g_{in} \\ b_{in} \end{pmatrix} + B_{3\times1}$$

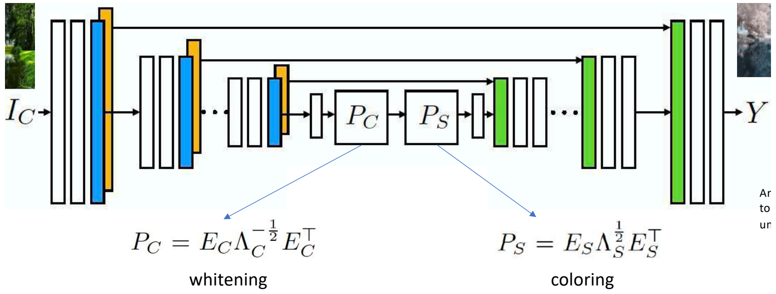
style

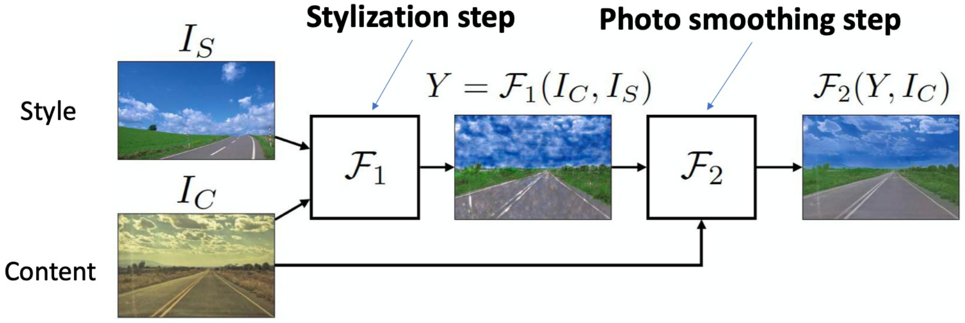
content

stylized content

Style Transfer Network

- FPST [Li et al. 2018]
  - models photo style transfer as a close-form function mapping
  - covariance matrix of deep features encodes the style information





style

content

stylized content

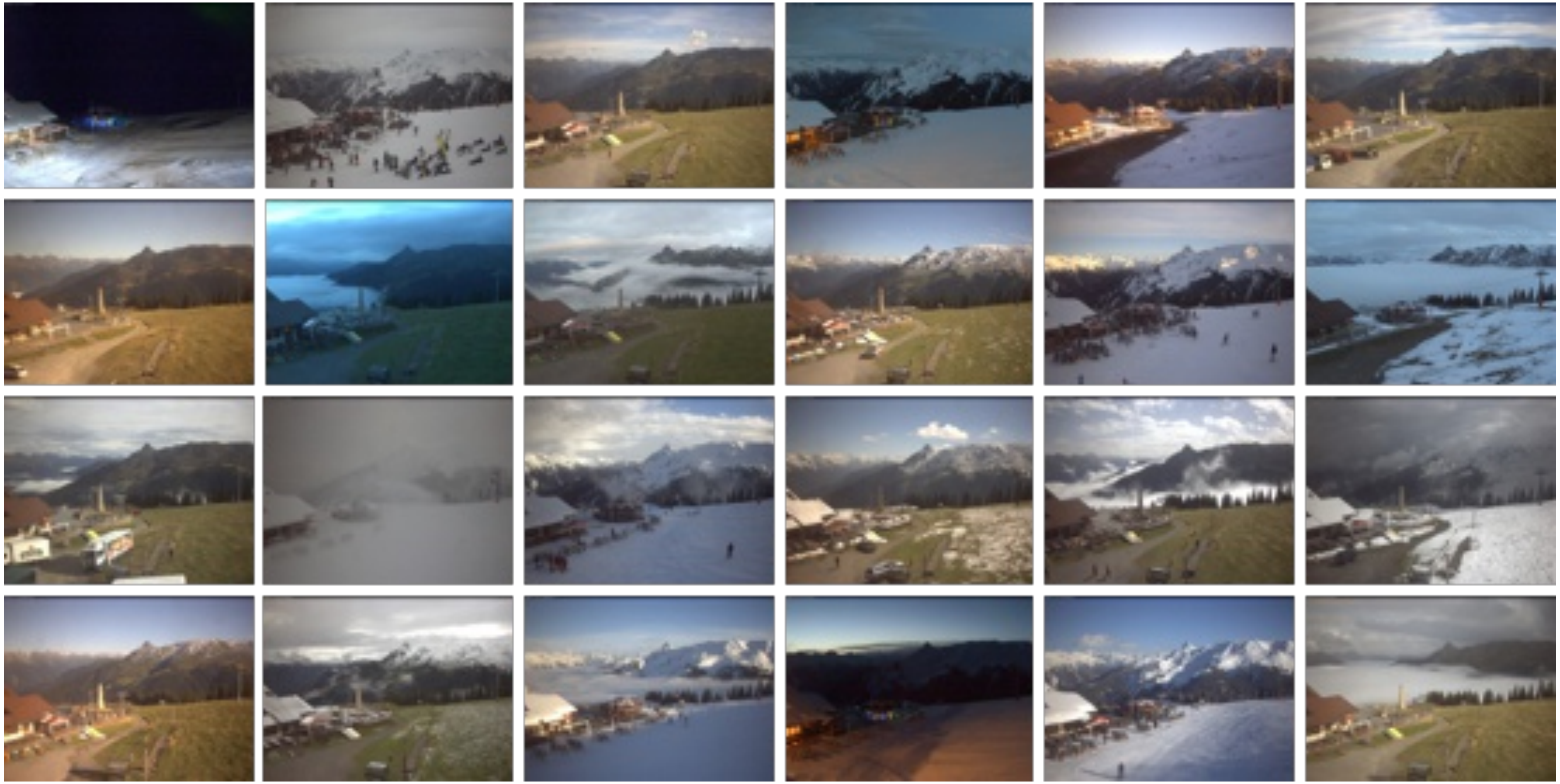
Training Data

- A collection of images from ADE20K [Zhou et al., 2017] and Transient Attributes [Laffont et al., 2014]

- 9,201 images corresponding to outdoor scenes from ADE20K dataset
  - Semantic layouts and predicted scene attributes
- 8,571 images from Transient Attributes dataset
  - Scene attributes and predicted semantic layouts

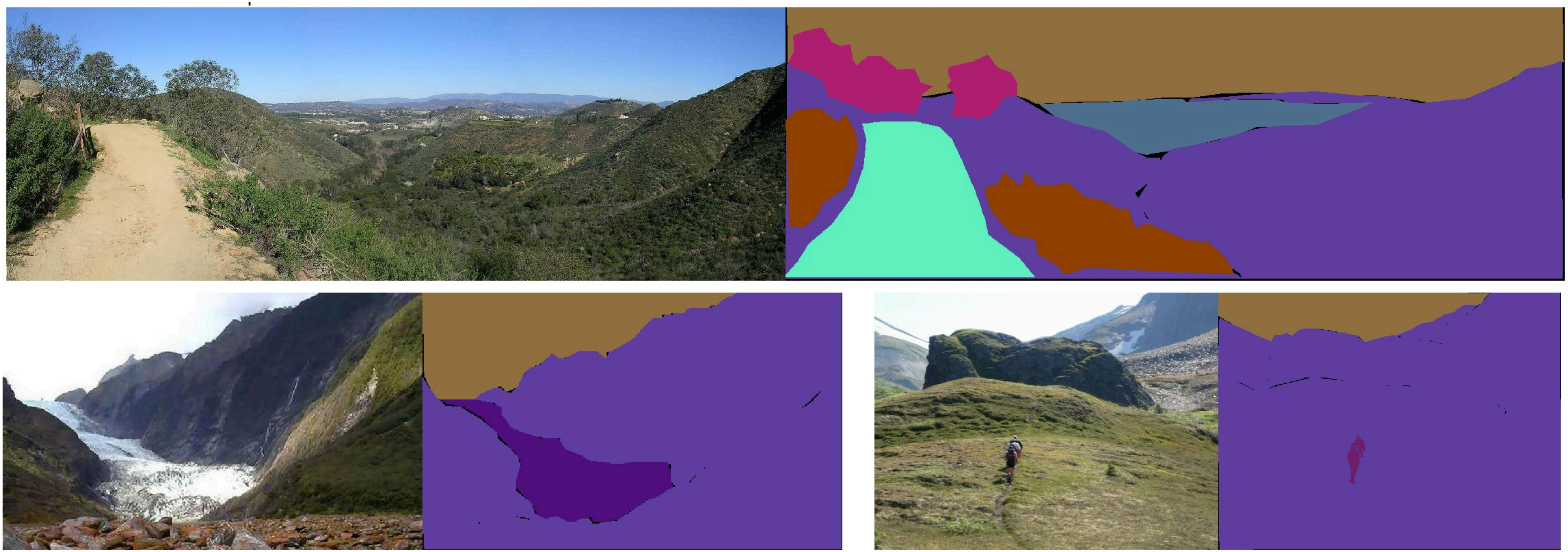
- In total 17,772 outdoor images with 150 semantic categories and 40 transient attributes
  - 1,338 images are used for testing

Transient Attributes Dataset [Laffont et al., 2014]



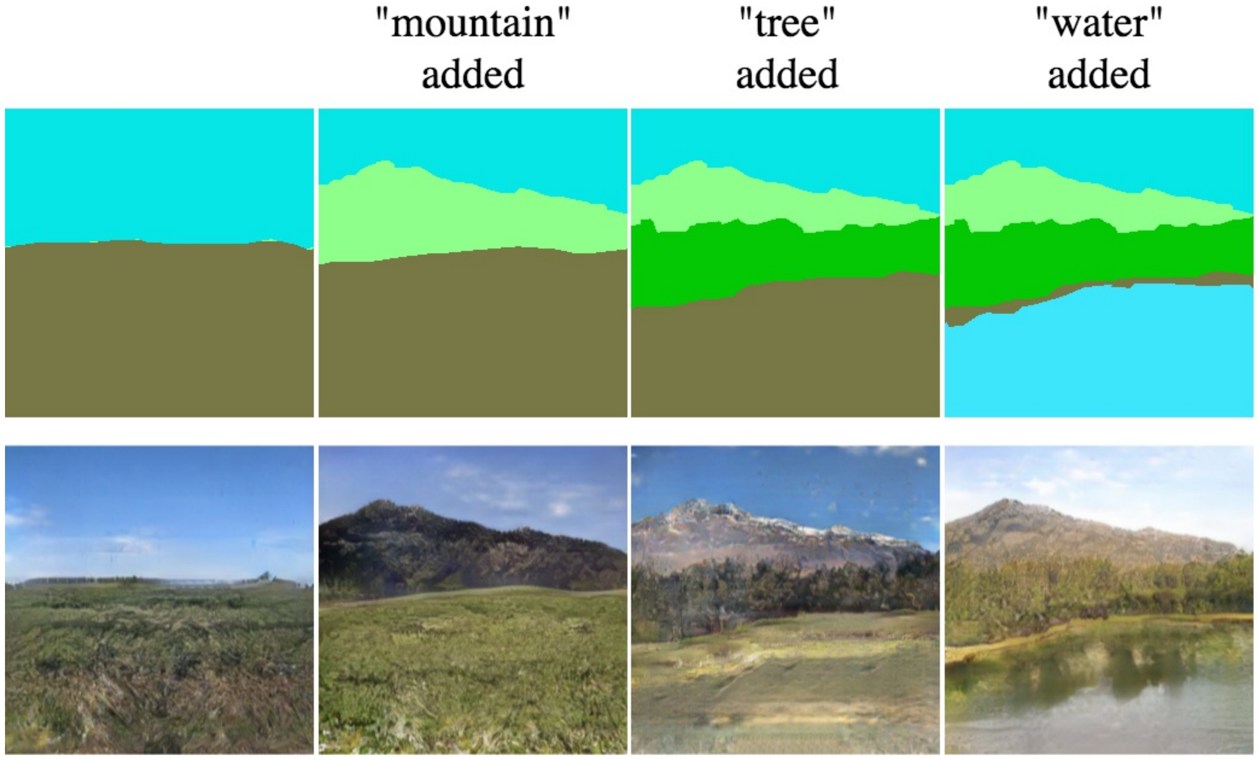
- 101 webcams 8571 outdoor scenes
- 40 transient attributes for each image

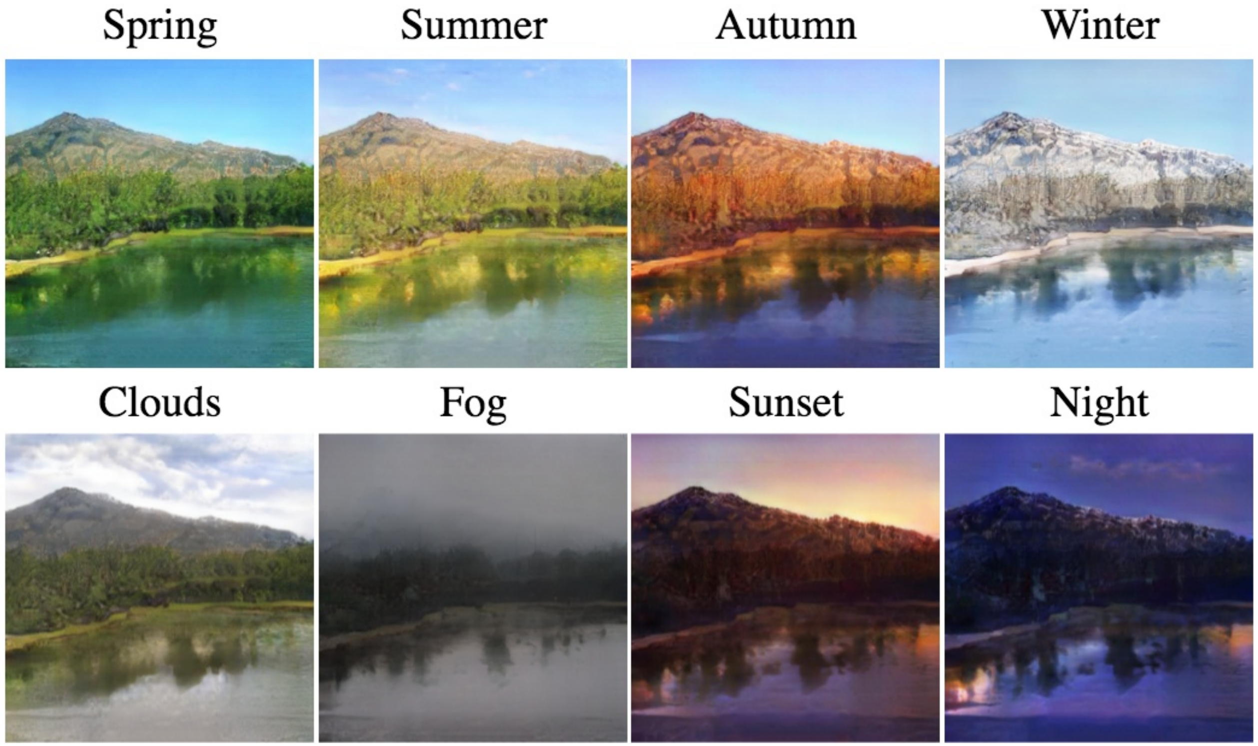
ADE20K Dataset [Zhou et al., 2017]



- 9,201 outdoor images
- 150 semantic categories

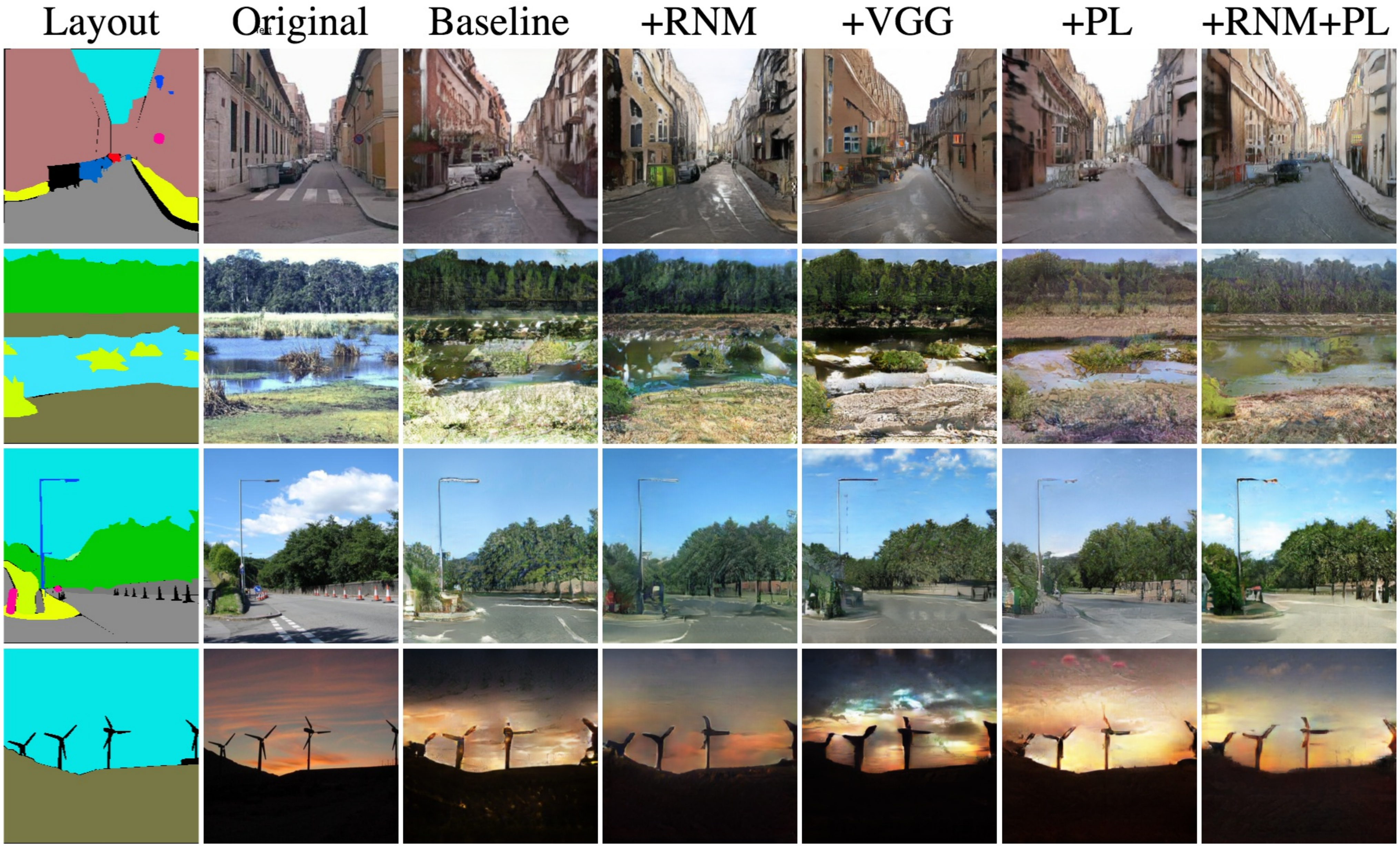
Sample generations



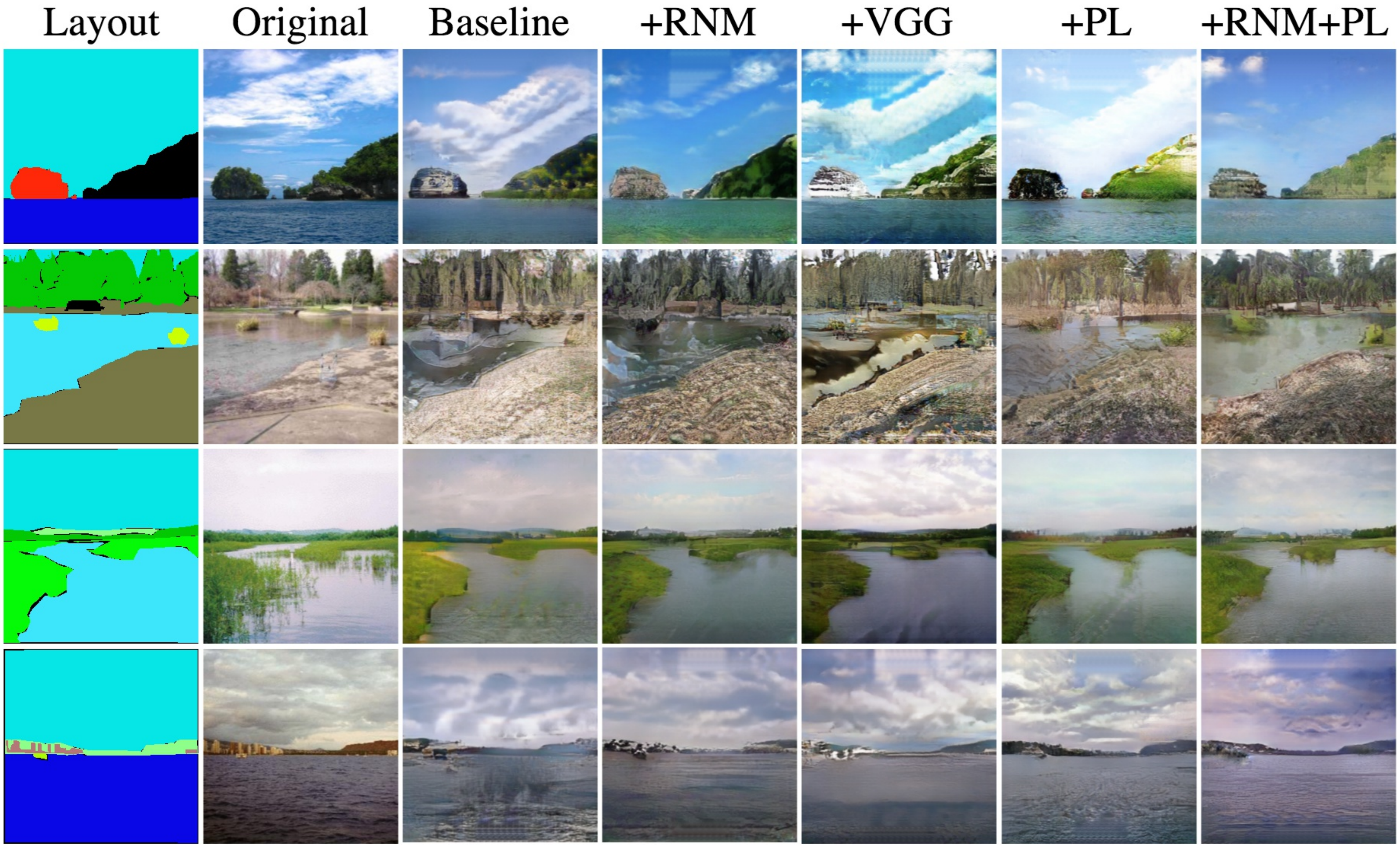


Comparison against pix2pix and pix2pixHD

Ablation Study



Ablation Study



Quantitative Analysis of SGN

Model	IS	FID	Att. MSE	Seg. Acc.
SGN	3.91	43.77	0.016	67.70
+RNM	3.89	41.84	0.016	70.11
+VGG	3.80	41.87	0.016	67.42
+PL	4.15	36.42	**0.015**	70.44
+RNM+PL	**4.19**	**35.02**	**0.015**	**71.80**
Original	5.77	0.00	0.010	75.64

	Model	IS	FID	Seg. Acc.
Coarse	Pix2pix	3.26	76.40	61.93
	Pix2pixHD	4.20	47.86	**75.57**
	Ours	**4.19**	**35.02**	71.80
	Original	5.77	0.00	75.64
Fine	Pix2pixHD	4.87	50.85	**76.17**
	Ours	**5.05**	**36.34**	74.60
	Original	7.37	0.00	77.14

- IS and FID to measure photorealism
- Attribute and segmentation predictions to measure consistency with the given contextual cues
- A user study containing 200 test questions was performed
- 66% of the users picked our results as more realistic.

Semantic layout


night
prediction


sunset
prediction


snow
prediction

prediction

Spring and clouds

prediction

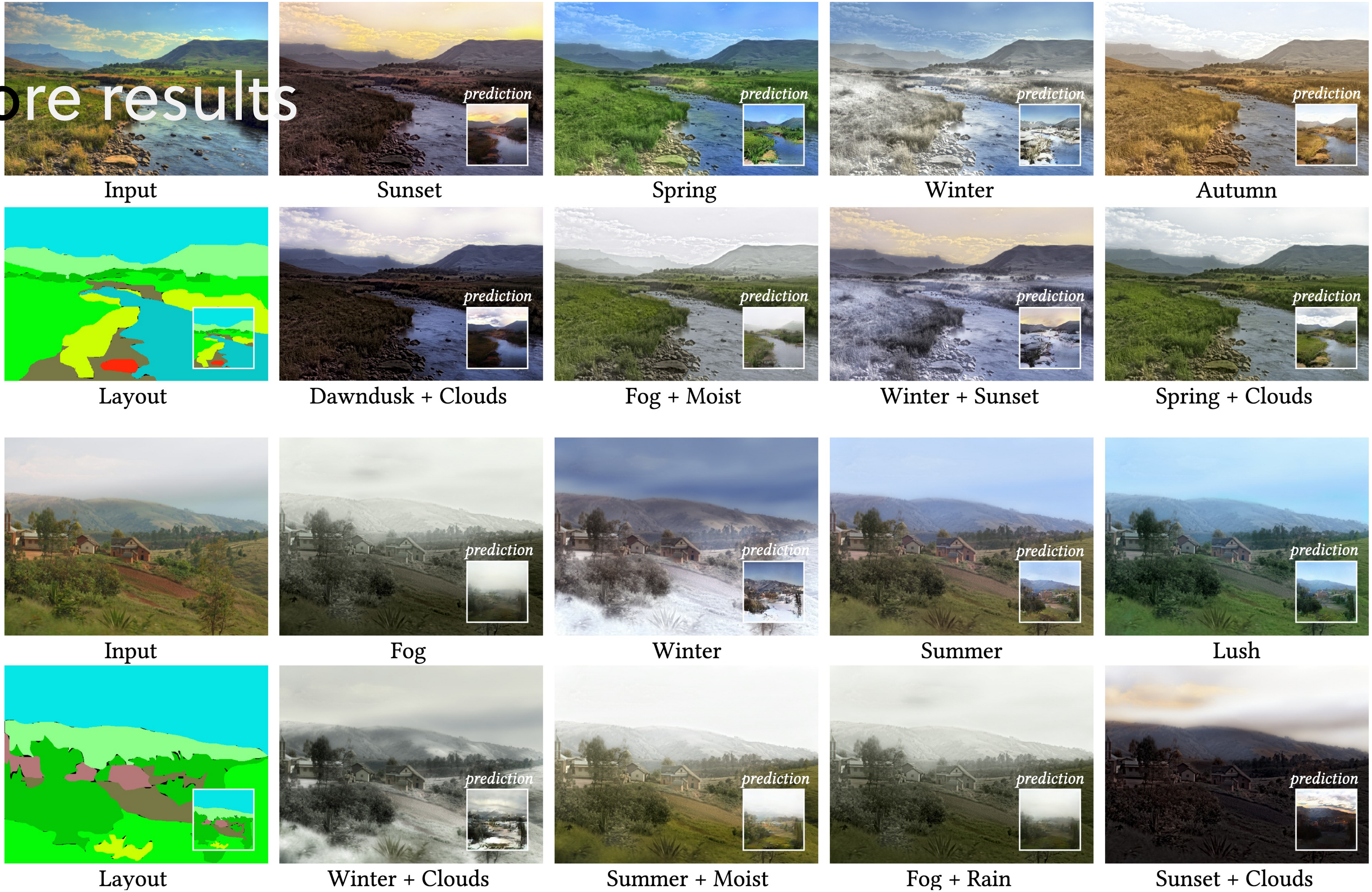
Moist, rain and fog

prediction

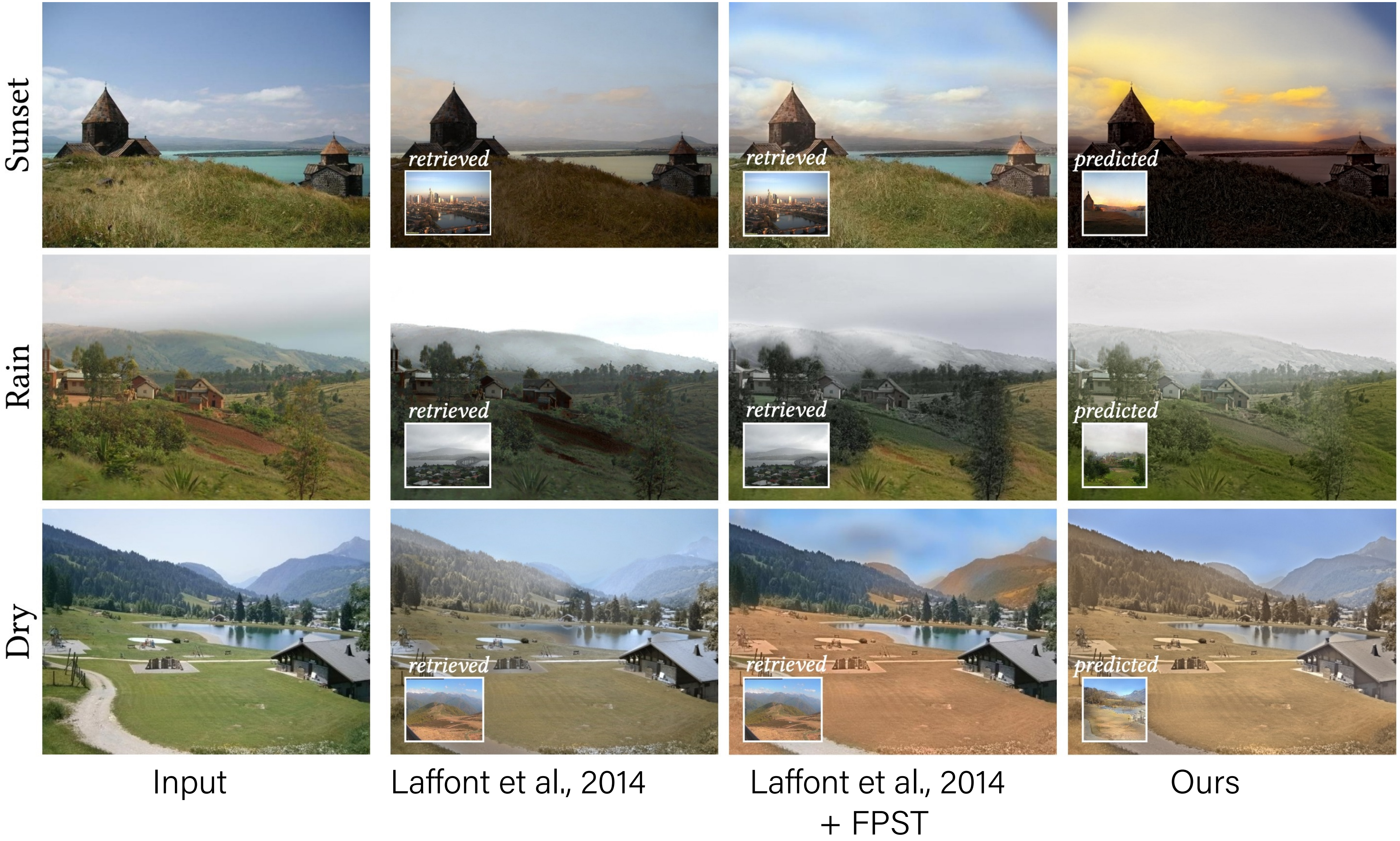
flowers

prediction

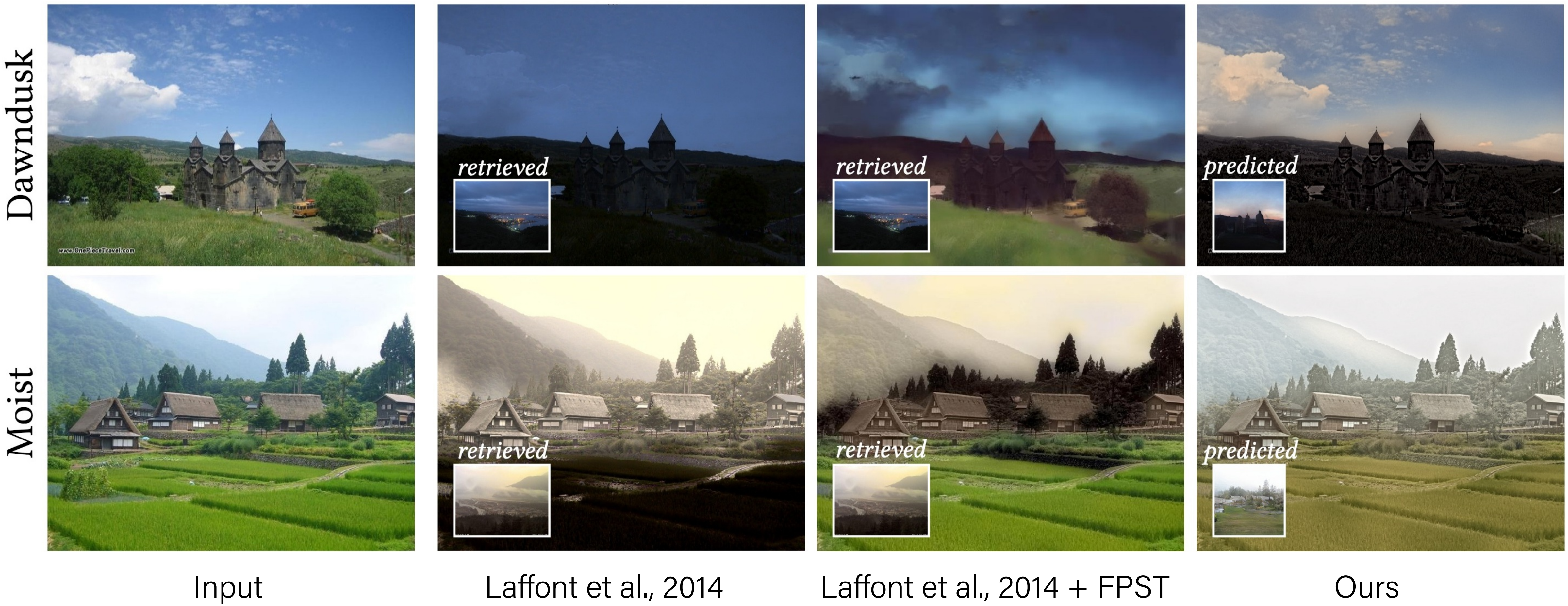
More results



Comparison against the state-of-the-art

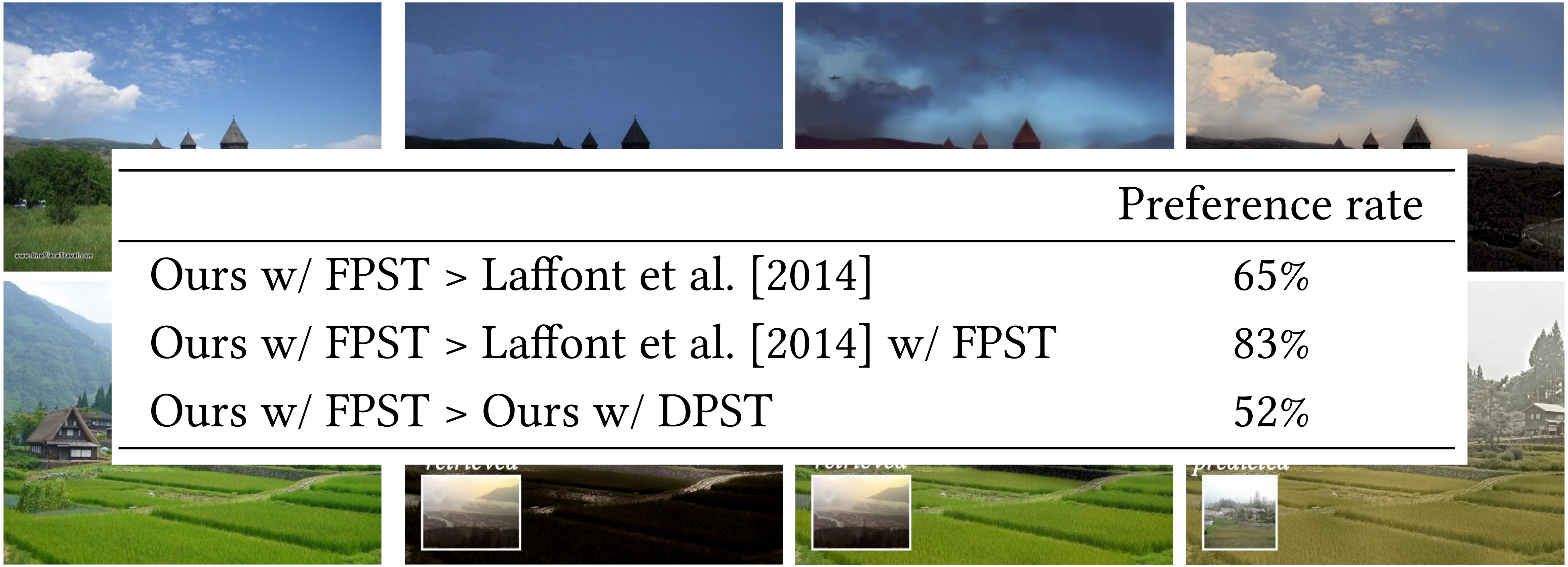


Comparison against the state-of-the-art



Our results are favored 65% of the time by the users on 60 different test questions.

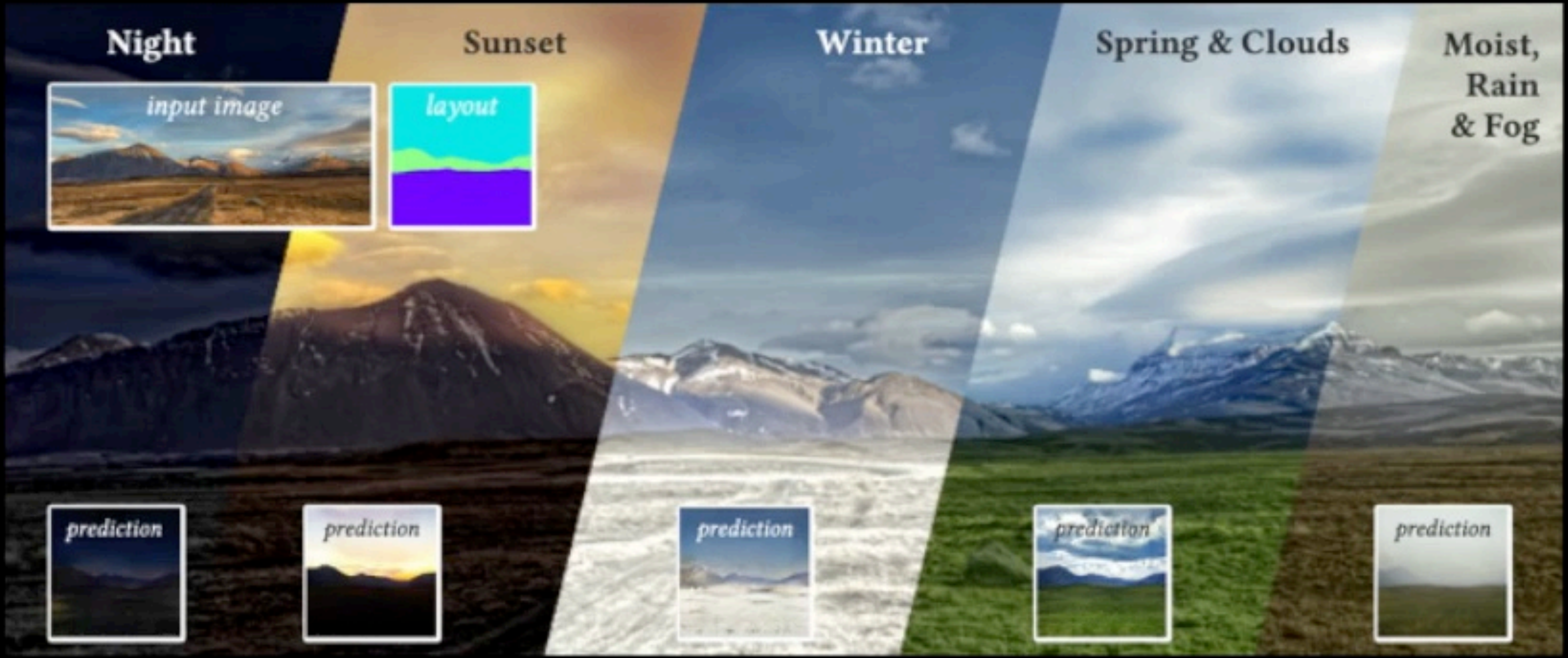
Comparison against the state-of-the-art



	Preference rate
Ours w/ FPST > Laffont et al. [2014]	65%
Ours w/ FPST > Laffont et al. [2014] w/ FPST	83%
Ours w/ FPST > Ours w/ DPST	52%

Our results are favored 65% of the time by the users on 60 different test questions.

Demo



Manipulating Attributes of Natural Scenes via Hallucination

Levent Karacan, Zeynep Akata, Aykut Erdem, Erkut Erdem

ACM Transactions on Graphics

Manipulating MR Images*

IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 38, NO. 10, OCTOBER 2019 2375

Image Synthesis in Multi-Contrast MRI With Conditional Generative Adversarial Networks

Salman UH. Dar, *Student Member, IEEE*, Mahmut Yurt, Levent Karacan, Aykut Erdem, Erkut Erdem, and Tolga Çukur, *Senior Member, IEEE*

***Abstract*—Acquiring images of the same anatomy with multiple different contrasts increases the diversity of diagnostic information available in an MR exam. Yet, the scan time limitations may prohibit the acquisition of certain contrasts, and some contrasts may be corrupted by noise and artifacts. In such cases, the ability to synthesize unacquired or corrupted contrasts can improve diagnostic utility. For multi-contrast synthesis, the current methods learn a nonlinear intensity transformation between the source and target images, either via nonlinear regression or deterministic neural networks. These methods can, in turn, suffer from the loss of structural details in synthesized images. Here, in this paper, we propose a new approach for multi-contrast MRI synthesis based on conditional generative adversarial networks. The proposed approach preserves intermediate-to-high frequency details via an adversarial loss, and it offers enhanced synthesis performance via pixel-wise and perceptual losses for registered multi-contrast images and a cycle-consistency loss for unregistered images. Information from neighboring cross-sections are utilized to further improve synthesis quality. Demonstrations on T_1- and T_2- weighted images from healthy subjects and patients clearly indicate the superior performance of the proposed approach compared to the previous state-of-the-art methods. Our synthesis approach can help improve the quality and versatility of the multi-contrast MRI exams without the need for prolonged or repeated examinations.**

***Index Terms*—Generative adversarial network, image synthesis, multi-contrast MRI, pixel-wise loss, cycle-consistency loss.**

I. Introduction

MAGNETIC resonance imaging (MRI) is pervasively used in clinical applications due to the diversity of contrasts it can capture in soft tissues. Tailored MRI pulse sequences enable the generation of distinct contrasts while imaging the same anatomy. For instance, T_1-weighted brain images clearly delineate gray and white matter tissues, whereas T_2-weighted images delineate fluid from cortical tissue. In turn, multi-contrast images acquired in the same subject increase the diagnostic information available in clinical and research studies. However, it may not be possible to collect a full array of contrasts given considerations related to the cost of prolonged exams and uncooperative patients, particularly in pediatric and elderly populations [1]. In such cases, acquisition of contrasts with relatively shorter scan times might be preferred. Even then a subset of the acquired contrasts can be corrupted by excessive noise or artifacts that prohibit subsequent diagnostic use [2]. Moreover, cohort studies often show significant heterogeneity in terms of imaging protocol and the specific contrasts that they acquire [3]. Thus, the ability to synthesize missing or corrupted contrasts from other successfully acquired contrasts has potential value for enhancing multi-contrast MRI by increasing availability of diagnostically-relevant images, and improving analysis tasks such as registration and segmentation [4].

Cross-domain synthesis of medical images has recently been gaining popularity in medical imaging. Given a subject's image x in X (source domain), the aim is to accurately estimate the respective image of the same subject y in Y (target domain). Two main synthesis approaches are registration-based [5]–[7] and intensity-transformation-based methods [8]–[24]. Registration-based methods start by generating an atlas based on a co-registered set of images, x_1 and y_1, respectively acquired in X and Y [5]. These methods further make the assumption that within-domain images from separate subjects are related to each other through a geometric warp. For synthesizing y_2 from x_2, the warp that transforms x_1 to x_2 is estimated, and this warp is then applied

Manuscript received January 7, 2019; revised February 19, 2019; accepted February 22, 2019. Date of publication February 26, 2019; date of current version October 1, 2019. The work of T. Çukur was supported by a European Molecular Biology Organization Installation Grant (IG 3028), by a TUBITAK 1001 Grant (118E256), by a BAGEP fellowship awarded, by a TUBA GEBIP fellowship and Nvidia Corporation under GPU grant. The work of E. Erdem was supported by a separate TUBA GEBIP fellowship. *(Corresponding author: Tolga Çukur.)*

S. U. Dar and M. Yurt are with the Department of Electrical and Electronics Engineering, Bilkent University, TR-06800 Ankara, Turkey, and also with the National Magnetic Resonance Research Center, Bilkent University, TR-06800 Ankara, Turkey.

L. Karacan, A. Erdem, and E. Erdem are with the Department of Computer Engineering, Hacettepe University, TR-06800 Ankara, Turkey.

T. Çukur is with the Department of Electrical and Electronics Engineering, Bilkent University, TR-06800 Ankara, Turkey, also with the National Magnetic Resonance Research Center, Bilkent University, TR-06800 Ankara, Turkey, and also with the Neuroscience Program, Sabuncu Brain Research Center, Bilkent University, TR-06800 Ankara, Turkey (e-mail: cukur@ee.bilkent.edu.tr).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

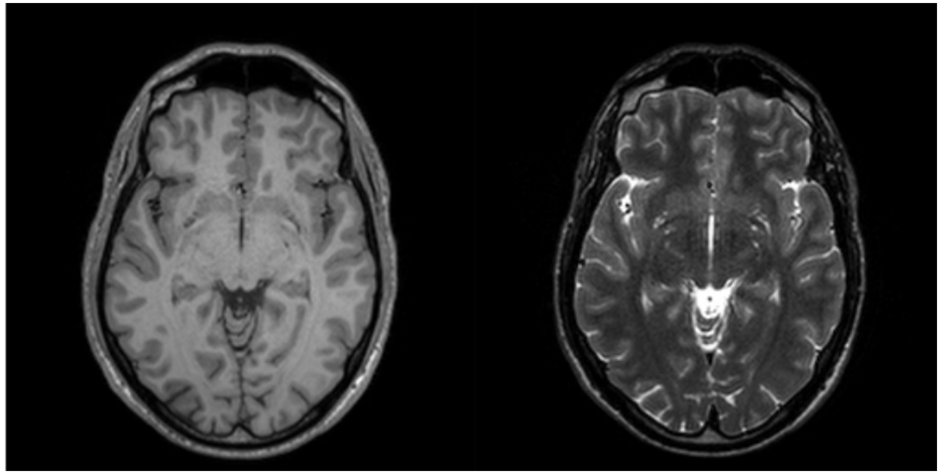
Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2901750

* Joint works with Salman UH. Dar, Mahmut Yurt, Levent Karacan, Tolga Cukur and Aykut Erdem

Motivation

- Acquiring multi-contrast MR images of a patient increases the diversity of diagnostic information for the radiologists.



- Cost of prolonged exams or uncooperative patients might prohibit the acquisition of full array of contrasts.

- Can we automatically synthesize unacquired or corrupted contrasts from successfully acquired contrast(s) to help diagnosis?

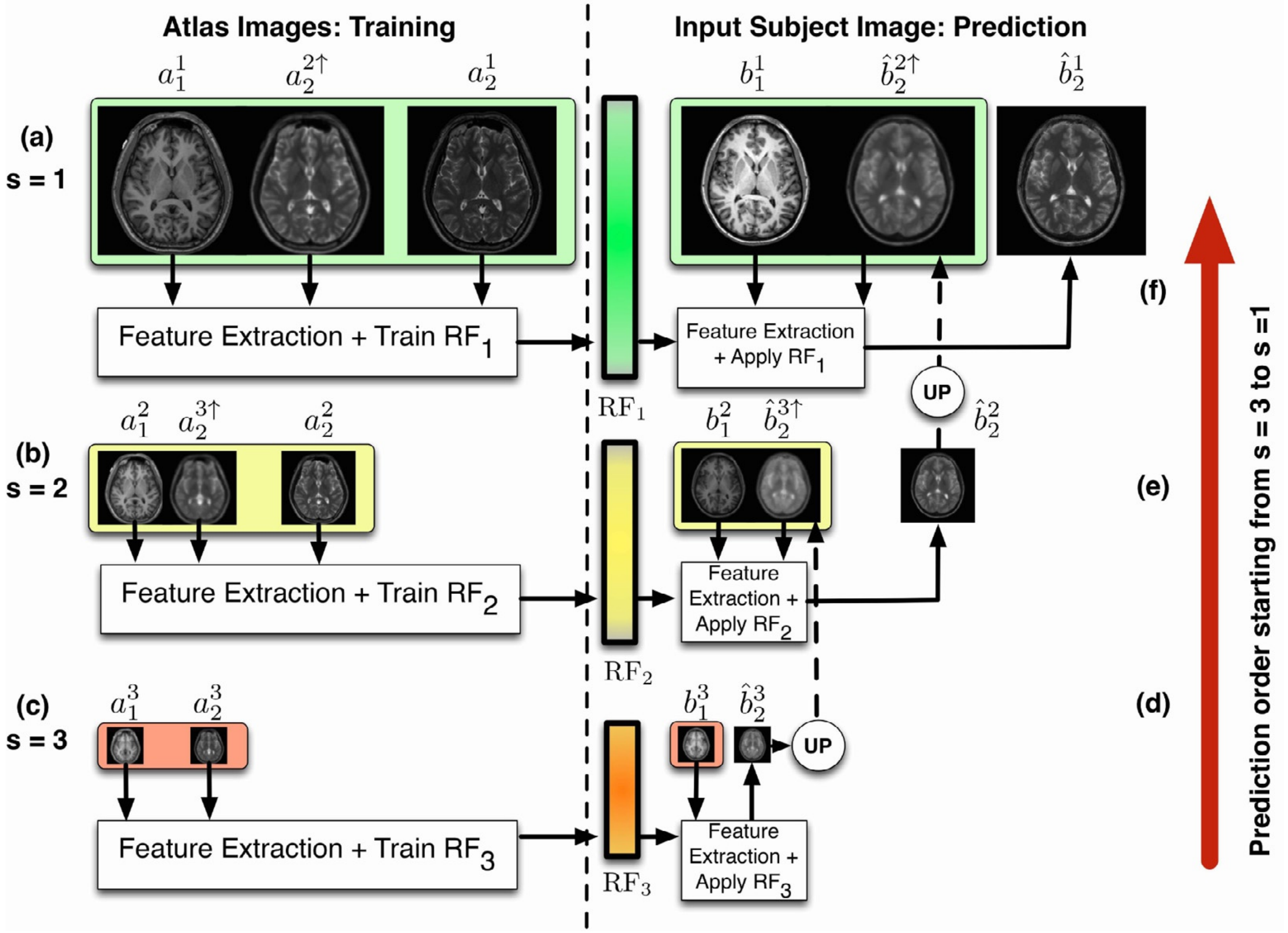
Our approaches

- We cast MRI synthesis as an image-to-image translation problem

- We propose two different MRI synthesis models
  - pGAN (Dar et al., 2019) – a variant of pix2pix model (single source – single output)
  - cGAN (Dar et al., 2019) – a variant of CycleGAN model (single source – single output)

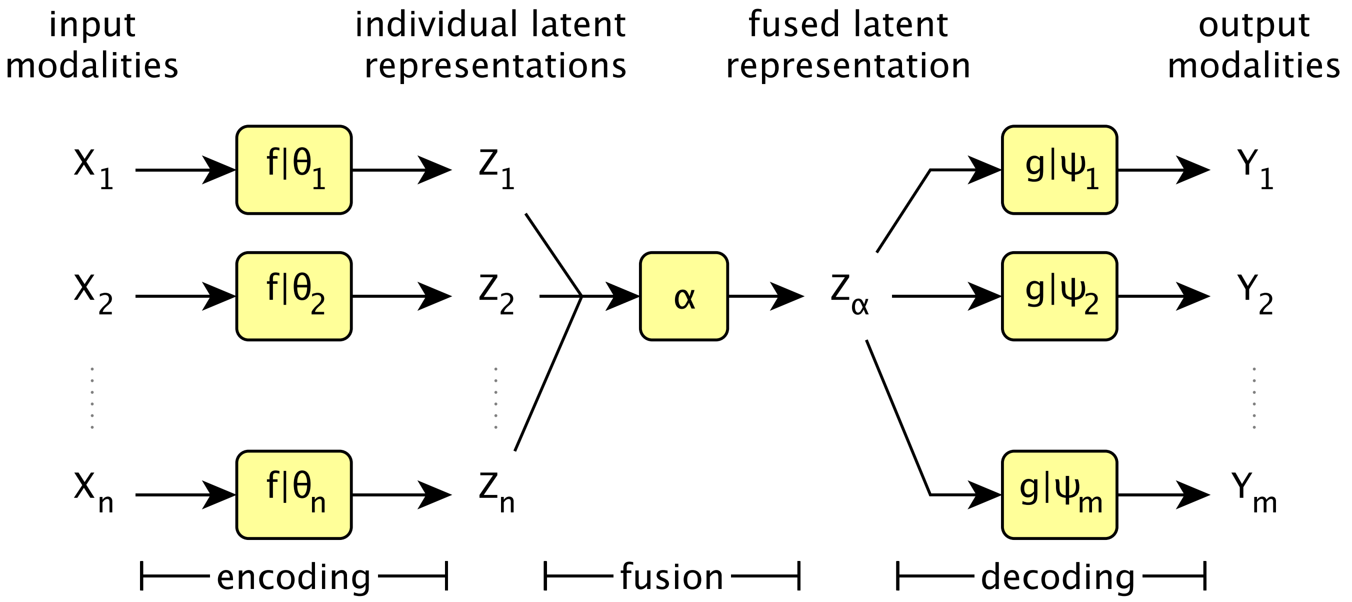
Related work

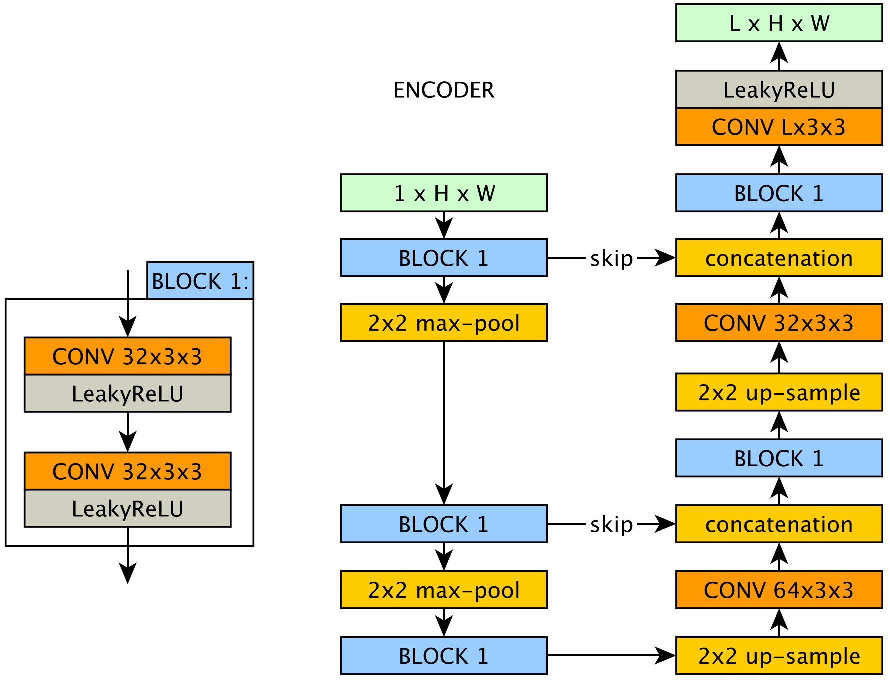
- REPLICA (Jog et al., Medical Image Analysis 2017)
  - a supervised random forest image synthesis approach
  - learns a nonlinear regression function to predict target contrast from a source contrast
  - Considers a multi-scale processing strategy



Related work

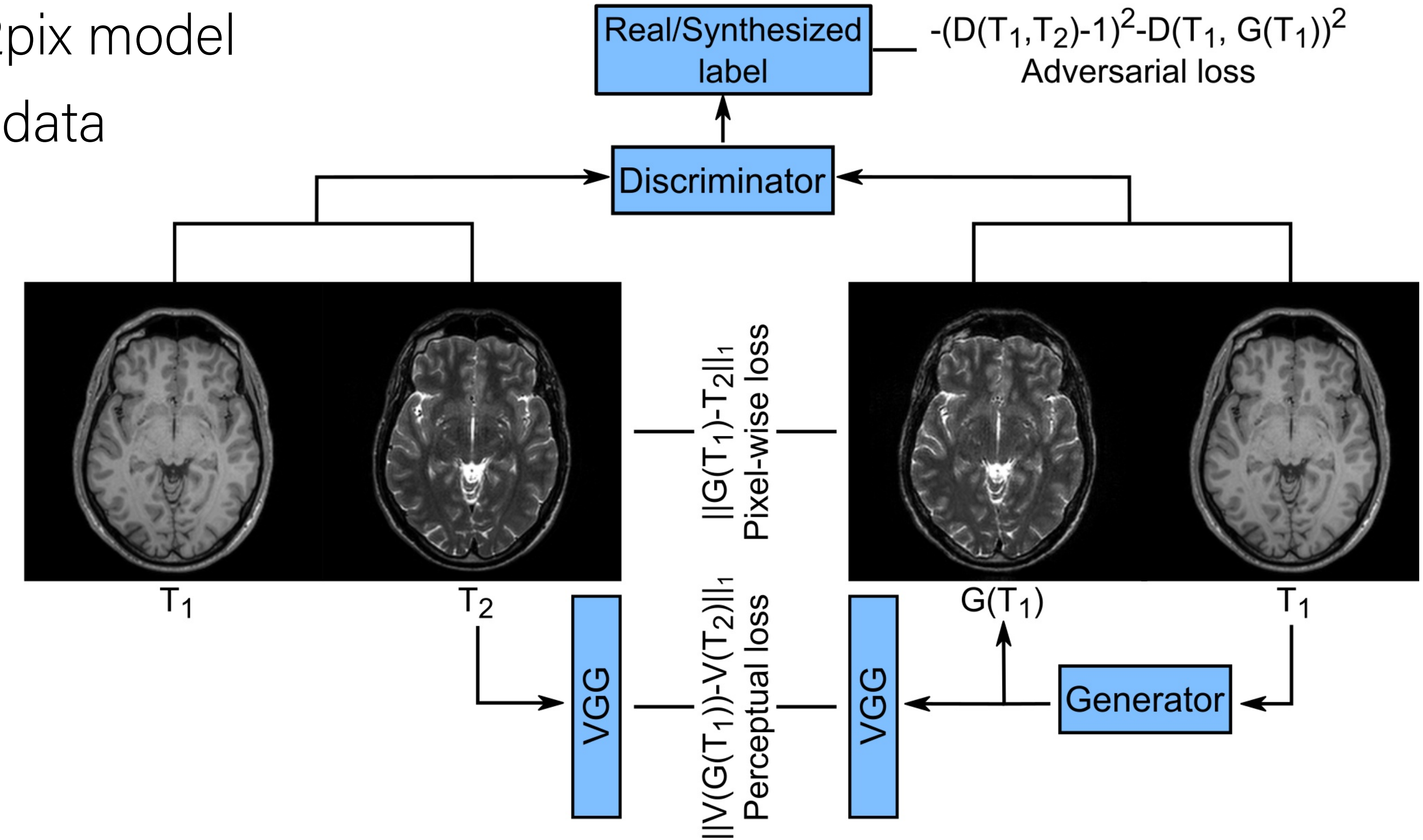
- Multimodal (Chartsias et al., IEEE Trans. Medical Imaging 2018)
  - a multi-input, multi-output fully convolutional neural network model
  - learns to embed all input modalities into a common latent space, which is used for MRI synthesis





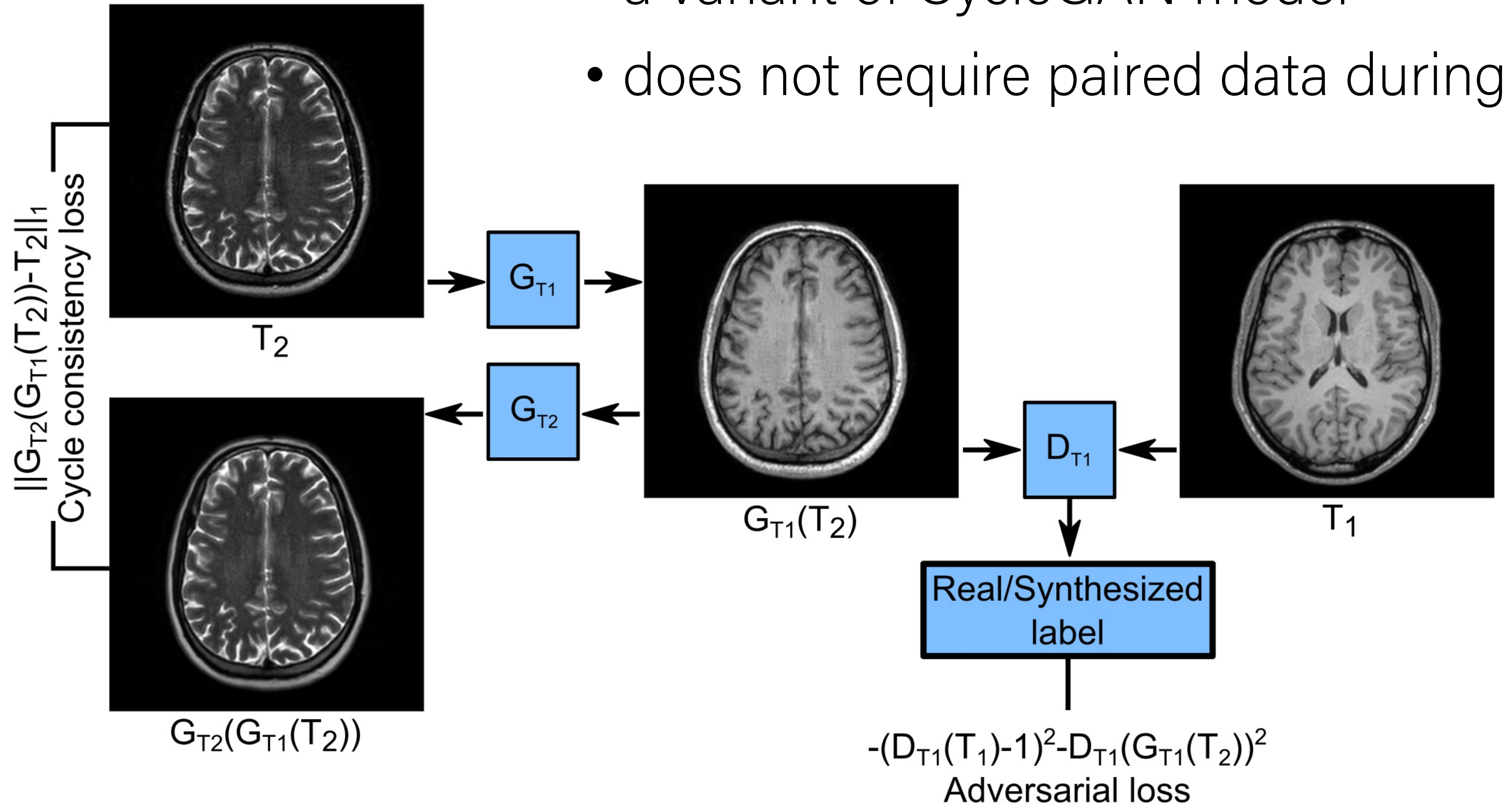
Multi-Contrast MRI systhesis with pGAN

- a variant of pix2pix model
- requires paired data during training



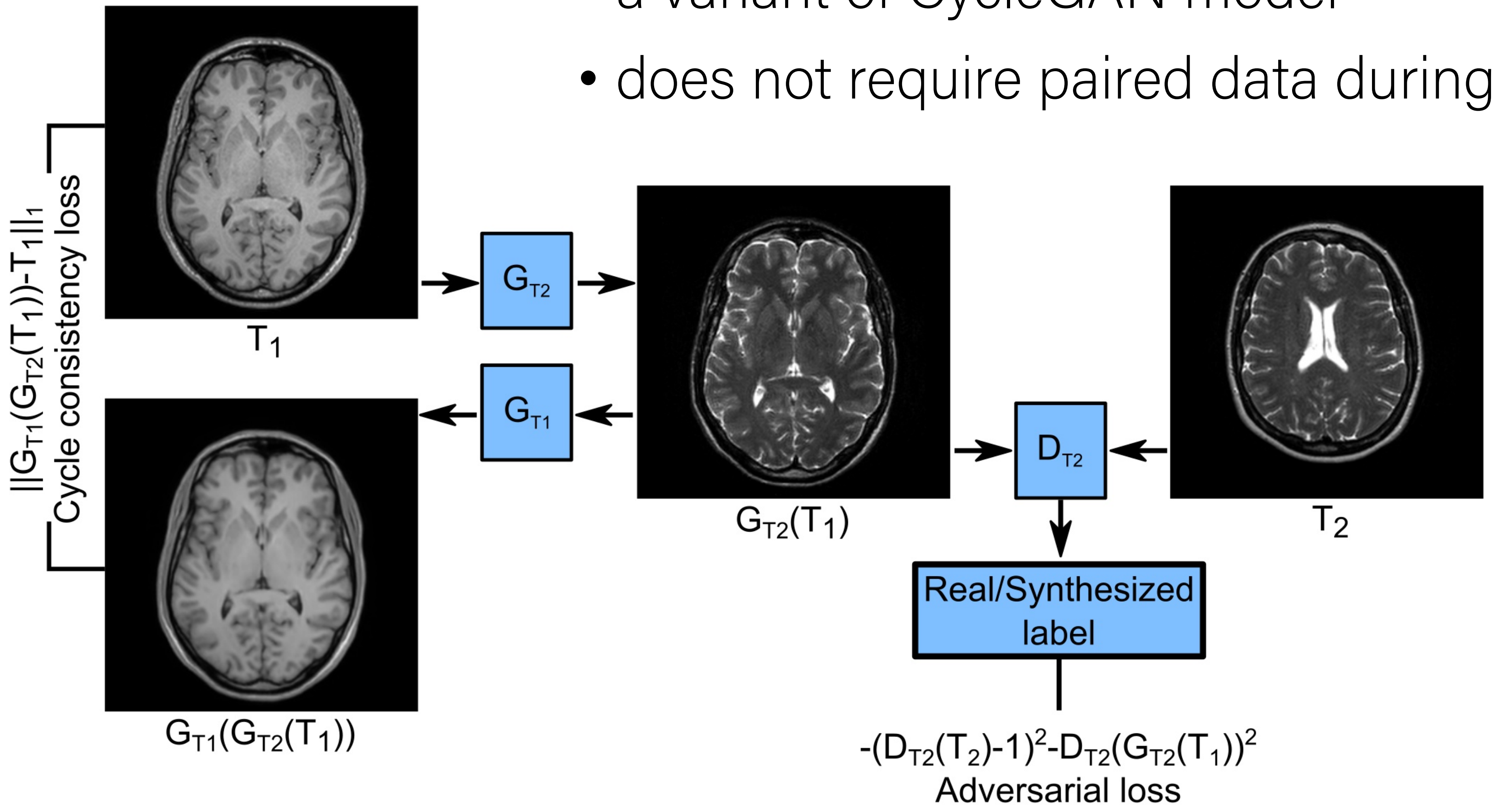
Multi-Contrast MRI systhesis with cGAN

- a variant of CycleGAN model
- does not require paired data during training

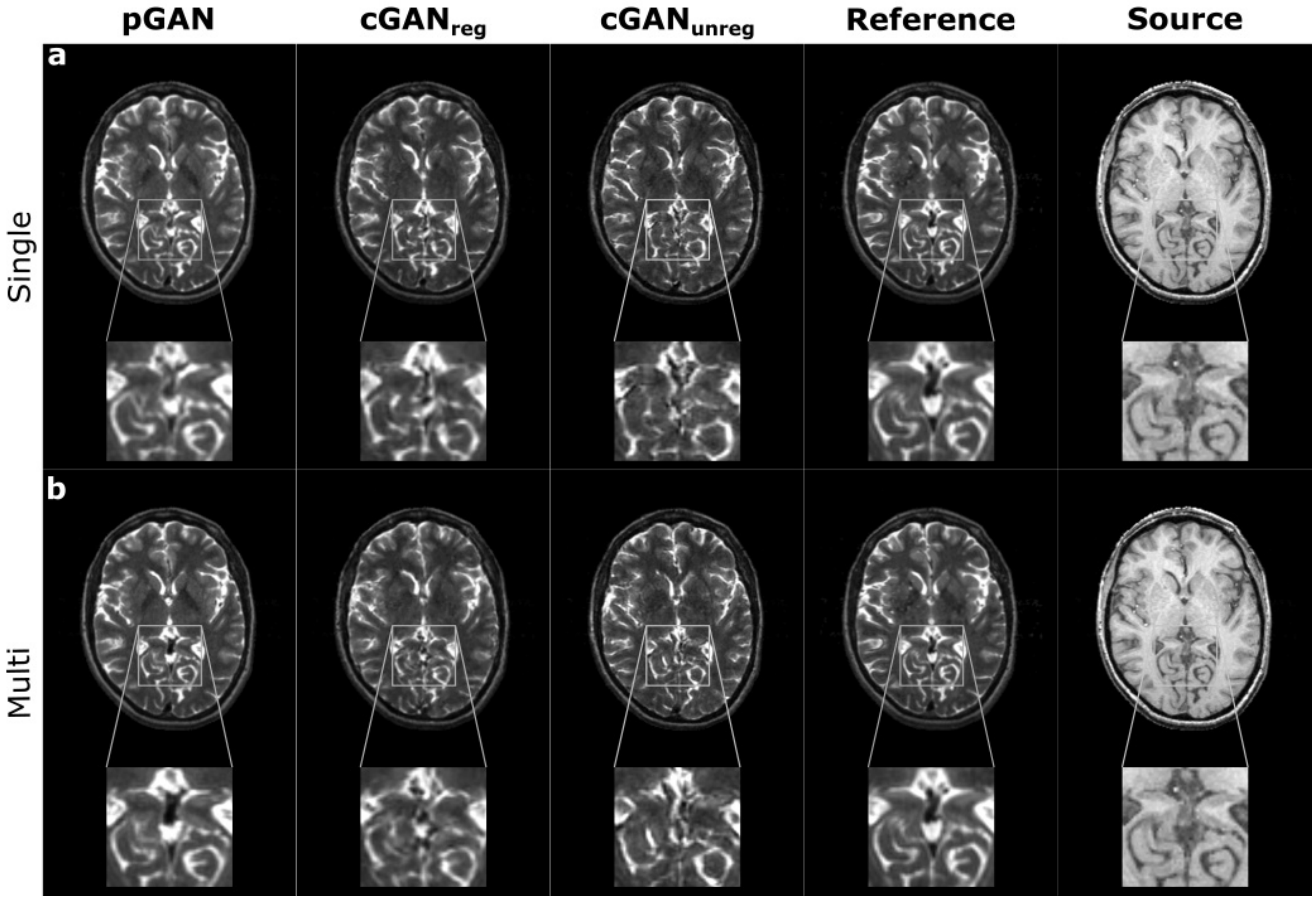


Multi-Contrast MRI systhesis with cGAN

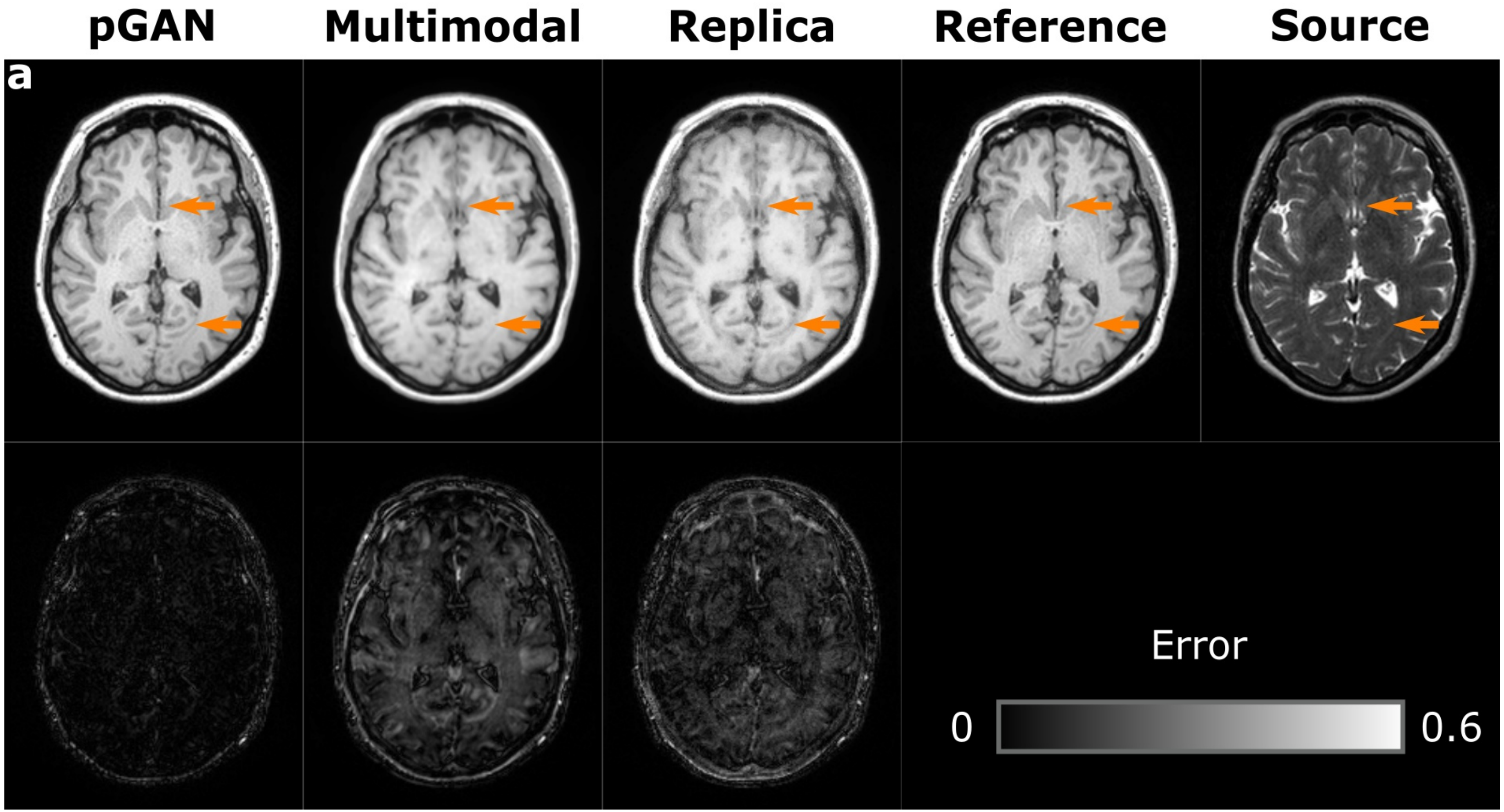
- a variant of CycleGAN model
- does not require paired data during training



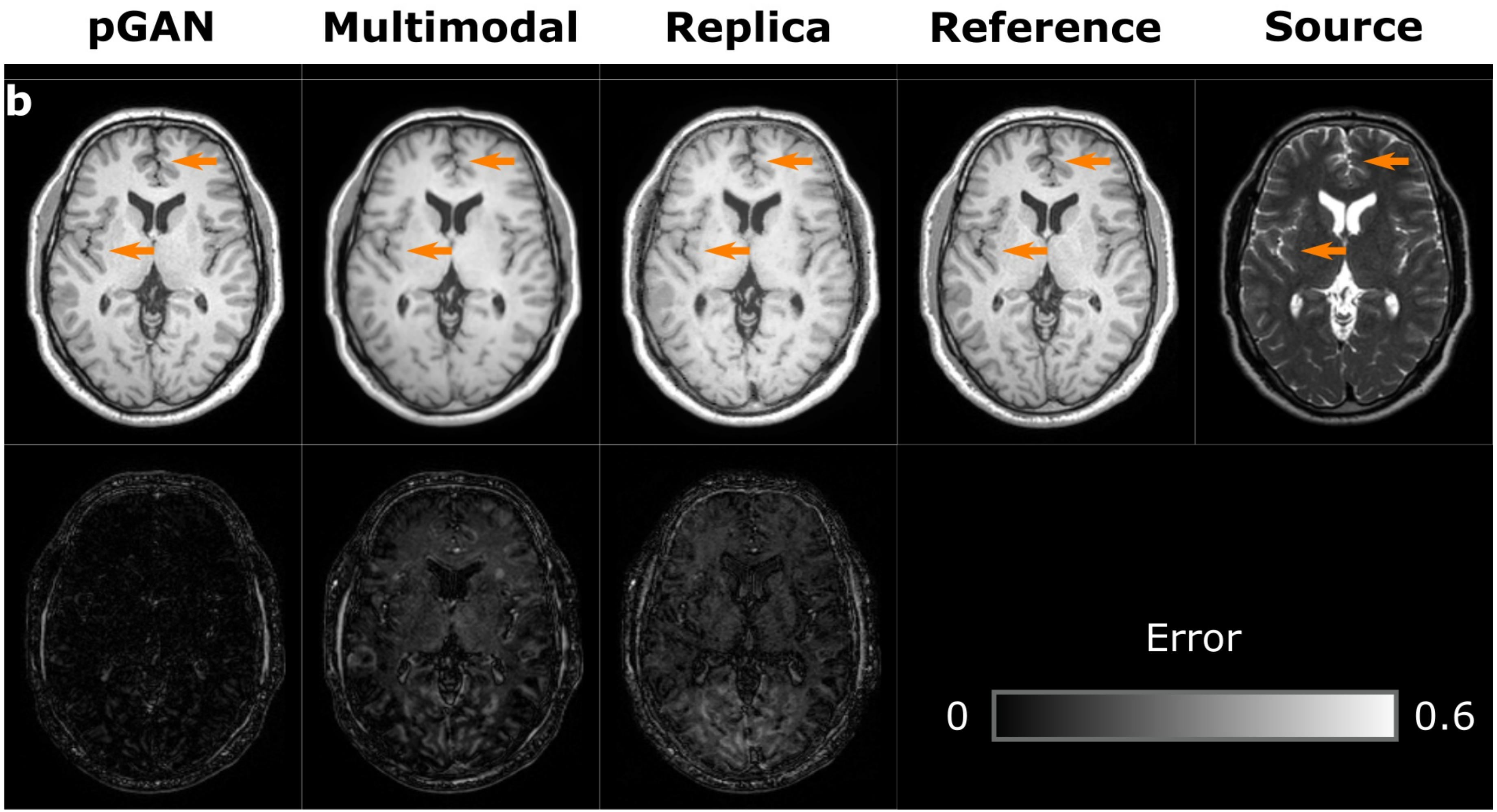
Qualitative Results



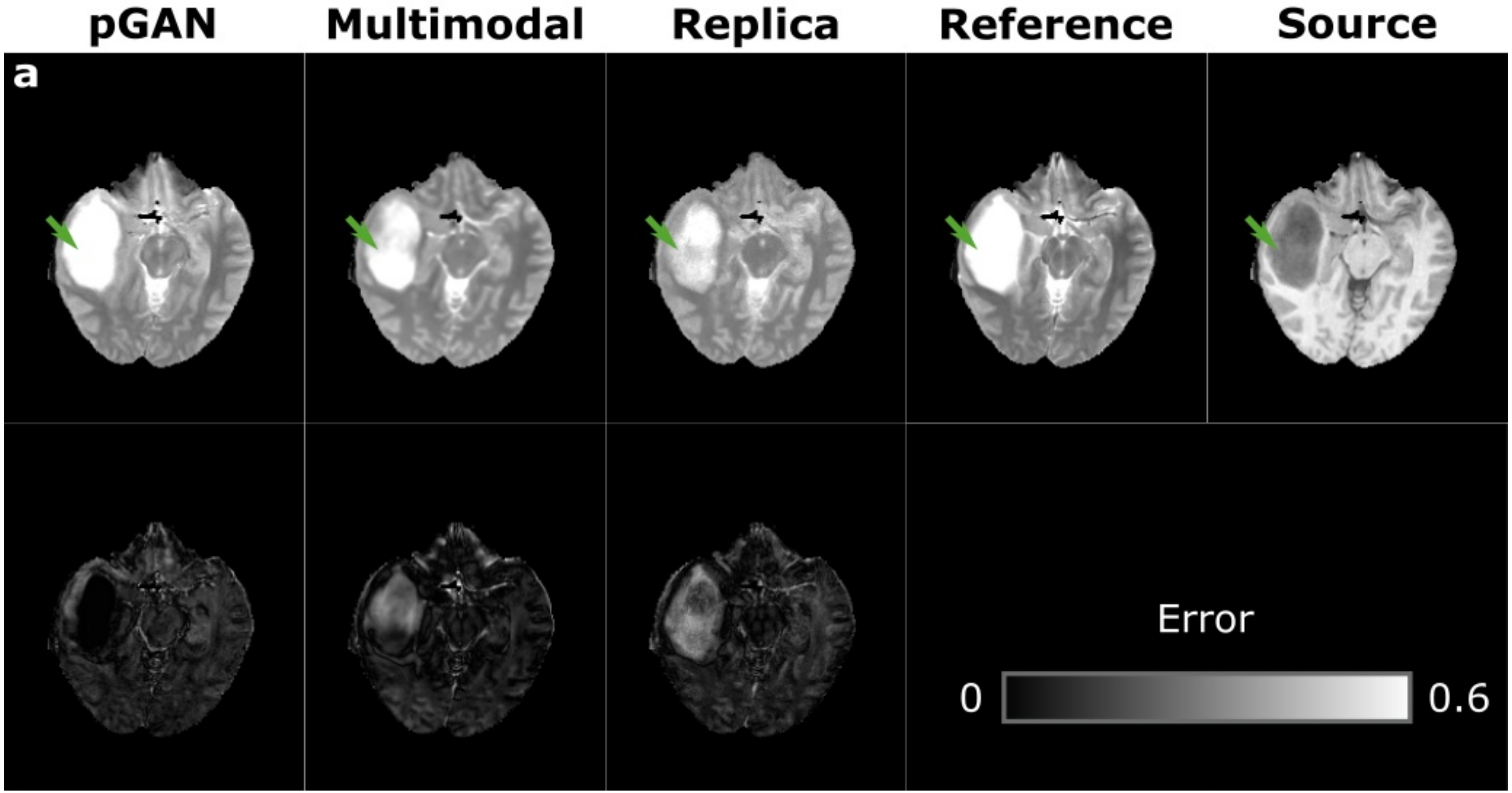
Comparison against the state-of-the-art



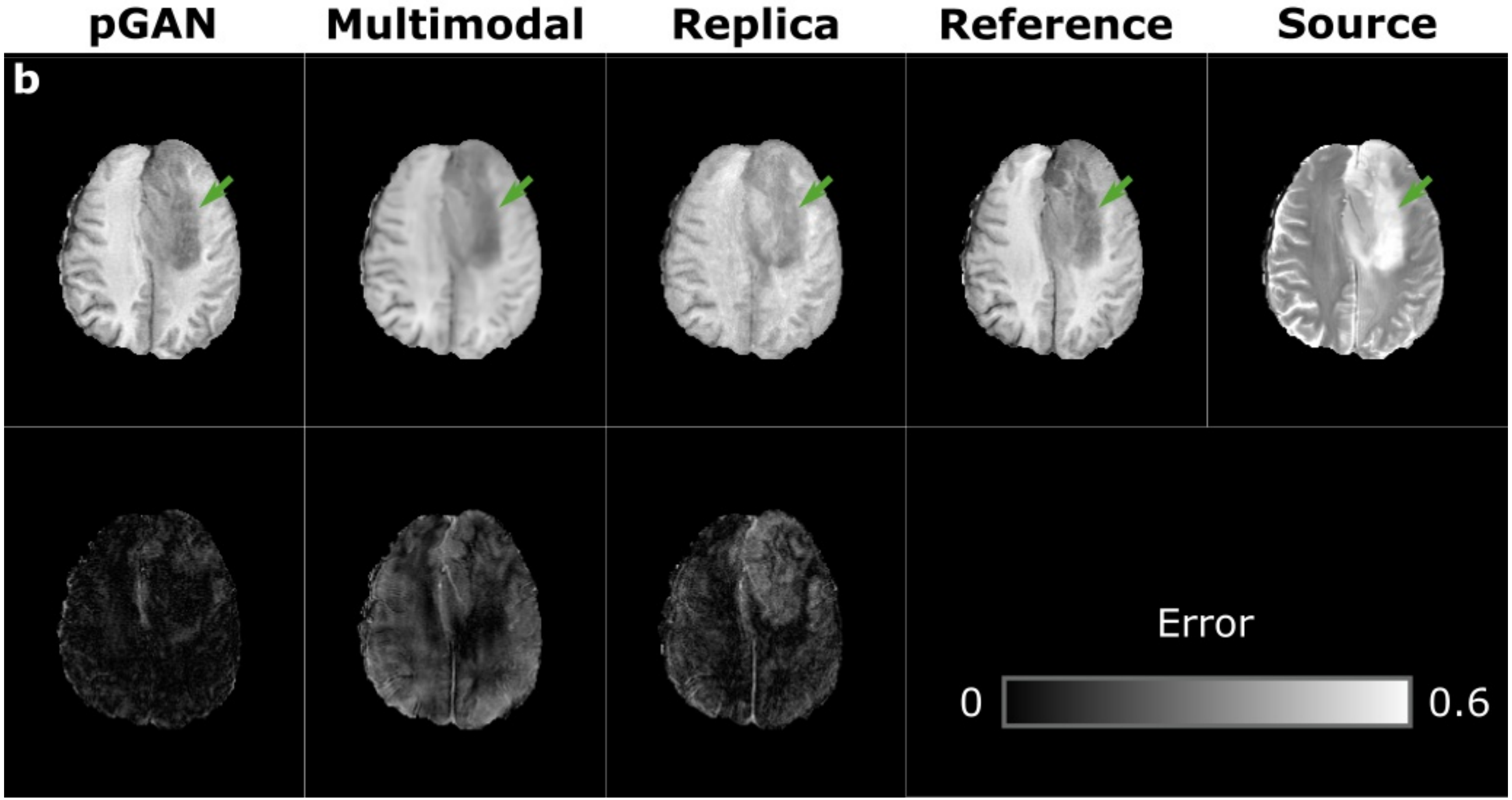
Comparison against the state-of-the-art



Comparison against the state-of-the-art



Comparison against the state-of-the-art



Comparison against the state-of-the-art

QUALITY OF SYNTHESIS IN THE MIDAS DATASET

	pGAN		**Replica**		**Multimodal**	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
$T_1 \rightarrow T_{2\#}$	**0.926 ±0.014**	**29.34 ±0.592**	0.877 ±0.027	26.18 ±0.638	0.924 ±0.012	28.33 ±0.501
$T_{1\#} \rightarrow T_2$	0.883 ±0.027	**27.49 ±0.643**	0.838 ±0.039	25.27 ±0.468	**0.889 ±0.020**	26.73 ±0.461
$T_2 \rightarrow T_{1\#}$	**0.920 ±0.016**	**28.16 ±1.303**	0.840 ±0.028	20.00 ±1.207	0.886 ±0.022	22.13 ±1.325
$T_{2\#} \rightarrow T_1$	**0.887 ±0.023**	**27.42 ±1.127**	0.827 ±0.031	20.29 ±1.066	0.872 ±0.020	23.08 ±1.280

Boldface marks the model with the highest performance.

Comparison against the state-of-the-art

QUALITY OF SYNTHESIS IN THE IXI DATASET

	pGAN		**Replica**		**Multimodal**	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
$T_1 \rightarrow T_{2\#}$	**0.948 ±0.014**	**29.77 ±1.568**	0.912 ±0.028	25.40 ±2.084	0.936 ±0.015	27.72 ±0.910
$T_{1\#} \rightarrow T_2$	**0.917 ±0.012**	**27.89 ±0.887**	0.863 ±0.023	24.08 ±1.427	0.898 ±0.014	26.11 ±0.769
$T_2 \rightarrow T_{1\#}$	**0.926 ±0.013**	**27.27 ±0.960**	0.865 ±0.013	20.46 ±0.921	0.895 ±0.015	22.61 ±1.105
$T_{2\#} \rightarrow T_1$	**0.953 ±0.012**	**29.55 ±1.423**	0.887 ±0.033	21.82 ±1.600	0.936 ±0.017	25.91 ±1.689

Boldface marks the model with the highest performance.

Comparison against the state-of-the-art

QUALITY OF SYNTHESIS IN THE BRATS DATASET

	pGAN		**Replica**		**Multimodal**	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
$T_1 \rightarrow T_2$	**0.946 ±0.009**	**27.19 ±1.456**	0.924 ±0.014	24.64 ±1.615	0.939 ±0.011	25.09 ±1.013
$T_2 \rightarrow T_1$	**0.940 ±0.009**	**25.80 ±1.867**	0.917 ±0.007	24.49 ±1.230	0.935 ±0.010	23.78 ±2.080

Boldface marks the model with the highest performance.