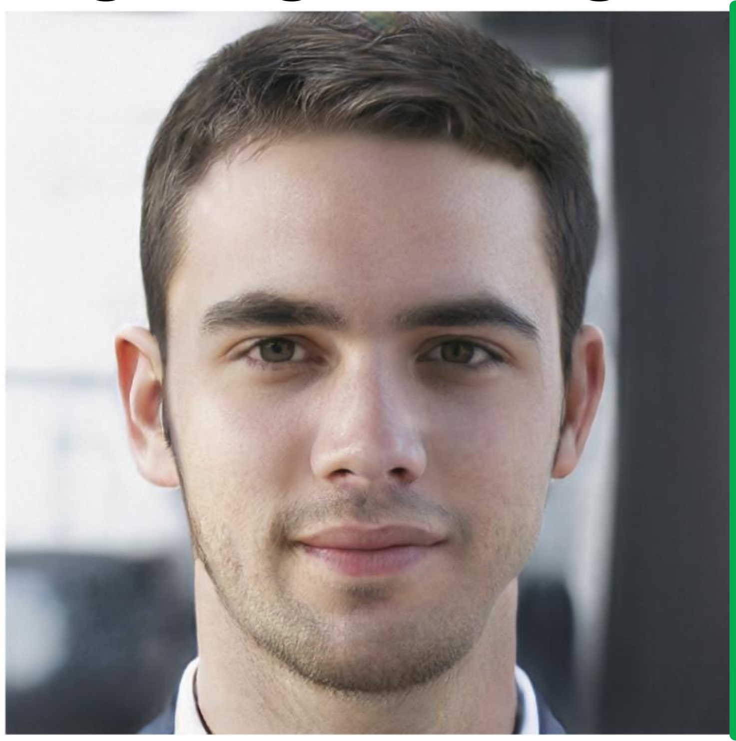
CS5670: Computer Vision

Guest Lecture - Jin Sun

Synthesizing images with generative adversarial networks (GANs)



Which face is real?

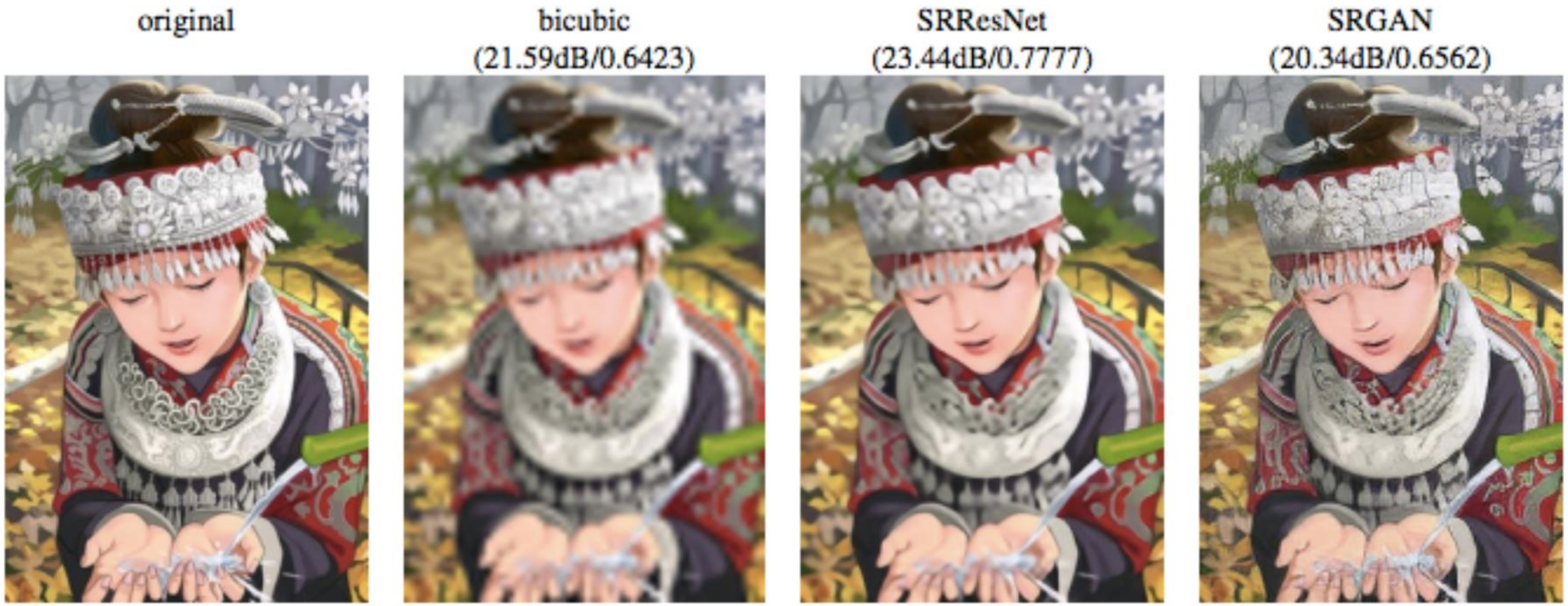
Slides from Philipp Isola

Announcements

- Project 5 due Friday, 5/10 at 11:59pm
- **Course evals** (you will receive a few bonus points!)
  - https://apps.engineering.cornell.edu/CourseEval
- Final exam in class next Monday, 5/6
  - Please arrange yourselves with at least one space between you and the closest person in the same row when you arrive

Motivation: Synthesizing images

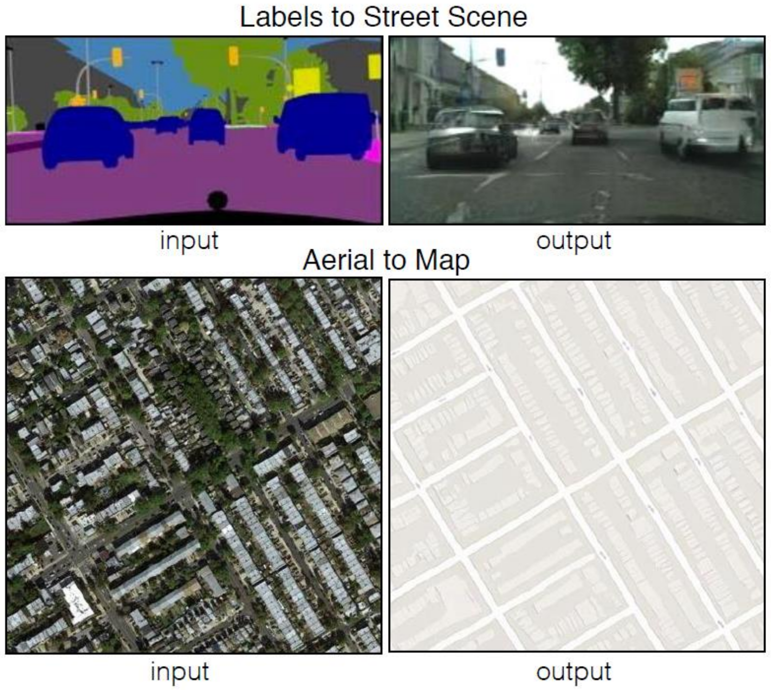
Single Image Super-Resolution



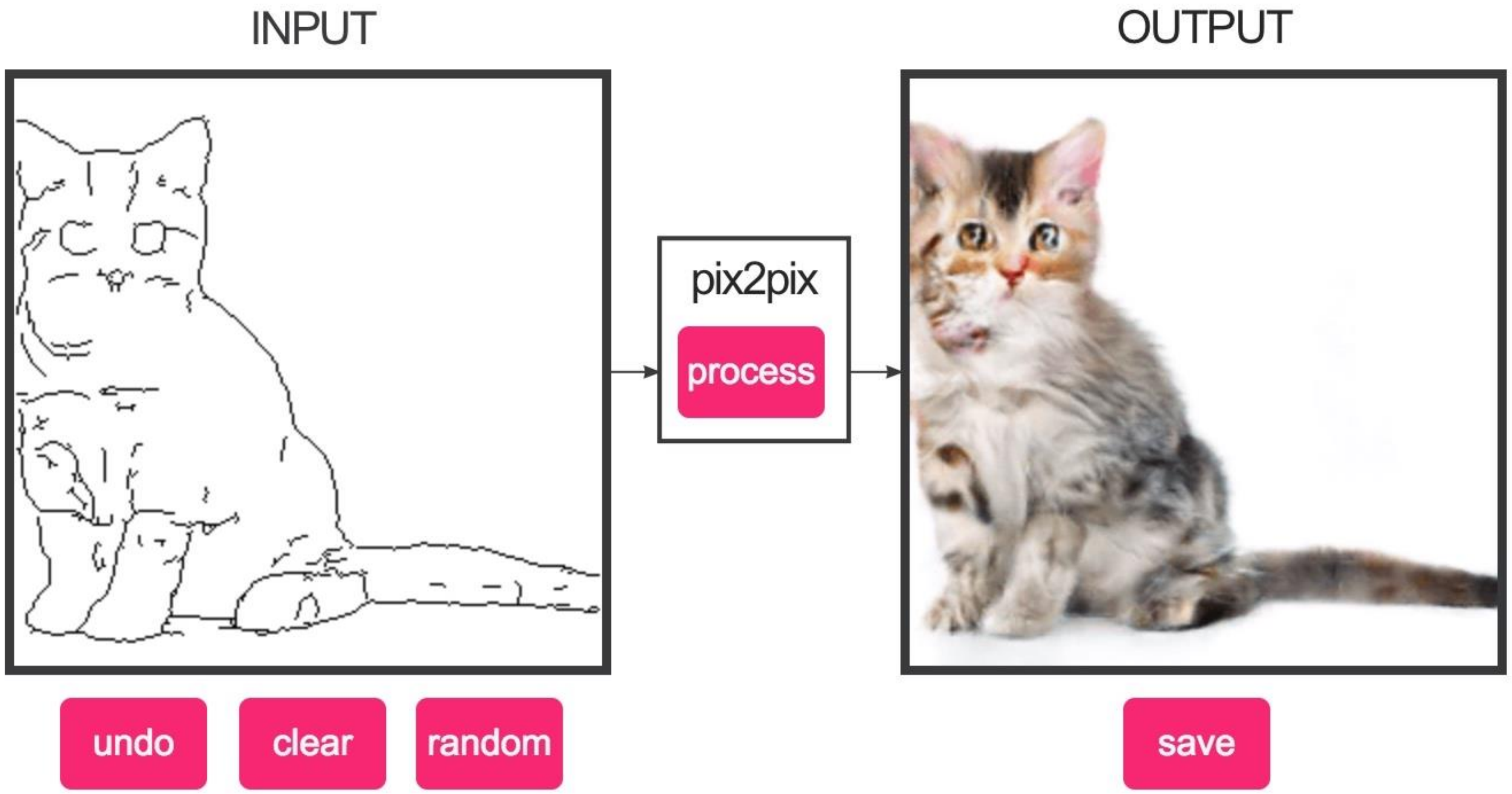
(Ledig et al 2016)

Motivation: Synthesizing images

Image to Image Translation



Demo



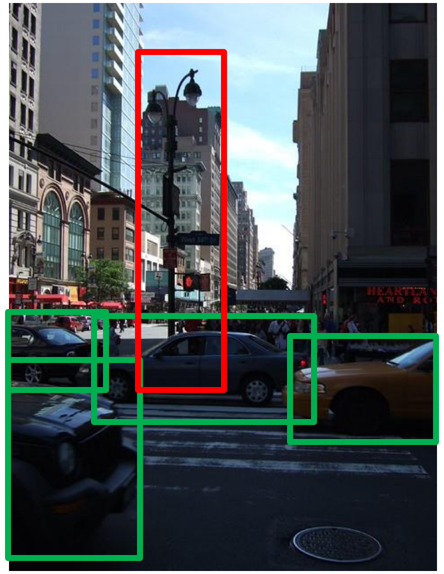
https://affinelayer.com/pixsrv/

Why Synthesizing Images

Computer vision is all about understanding the world from images

Understand →

← Synthesize



a busy downtown area with a lot of traffic and buildings.
this is a picture of a busy downtown cross walk with several cars in the flow of traffic.
cars passing on a street in the city.
...

"What I cannot create, I do not understand."

—Richard Feynman

Image classification

Learner

"Fish"

image X

label Y

Image classification

Learner

"Fish"

image X

label Y

Image classification

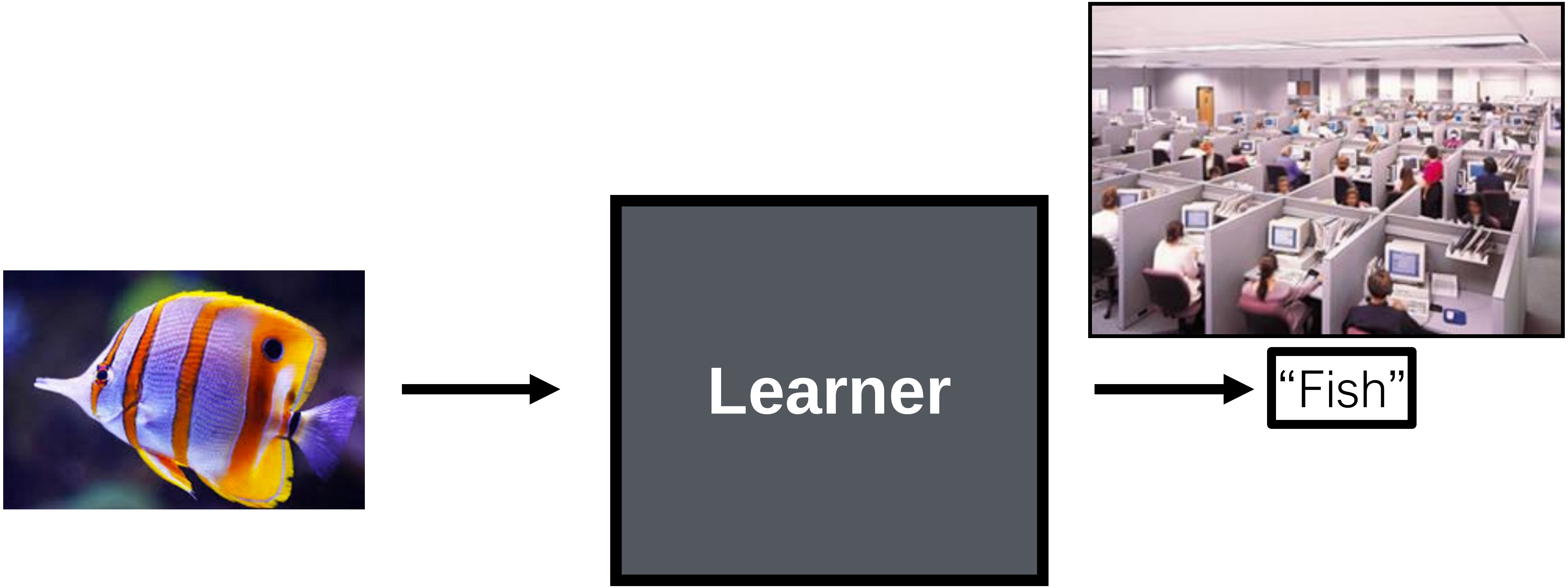


image X

label Y

Image classification

⋮

image X

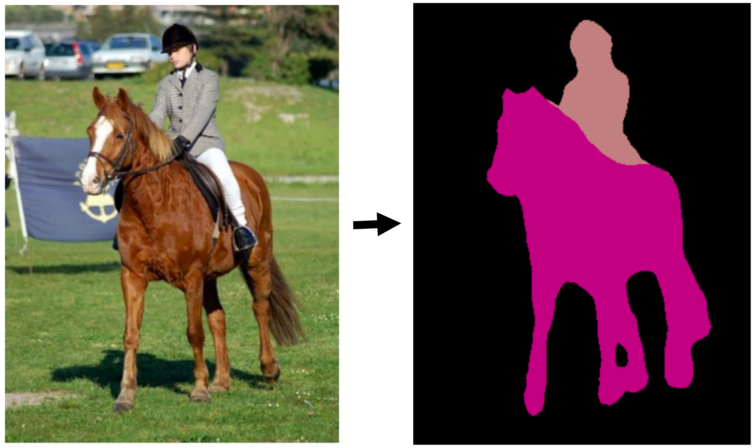
Learner

"Fish"

label Y

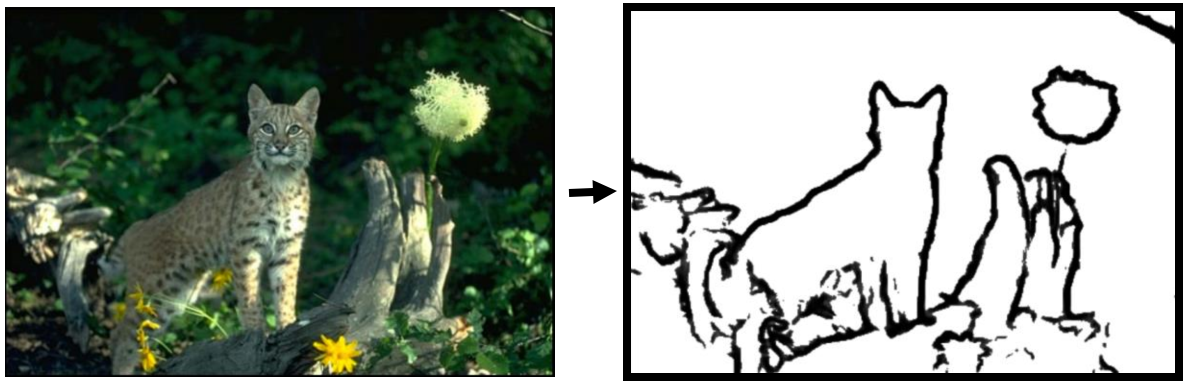
Image prediction ("structured prediction")

Object labeling



[Long et al. 2015, …]

Edge Detection



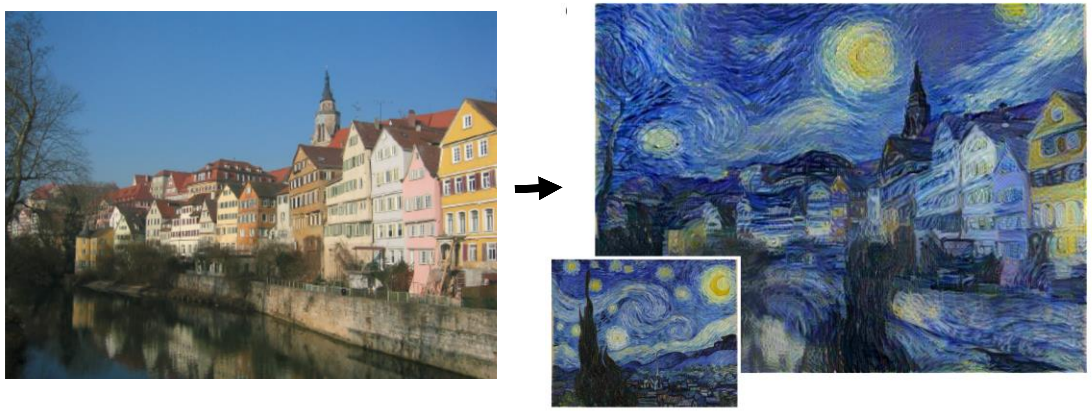
[Xie et al. 2015, …]

Text-to-photo

"this small bird has a pink breast and crown…" →

[Reed et al. 2014, …]

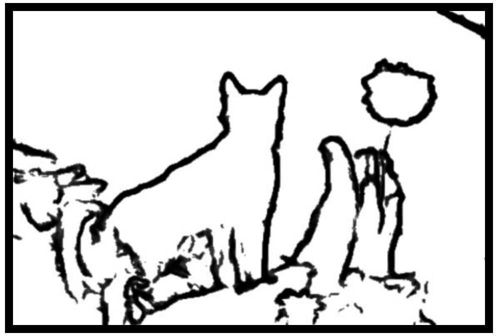
Style transfer



[Gatys et al. 2016, …]

Challenges

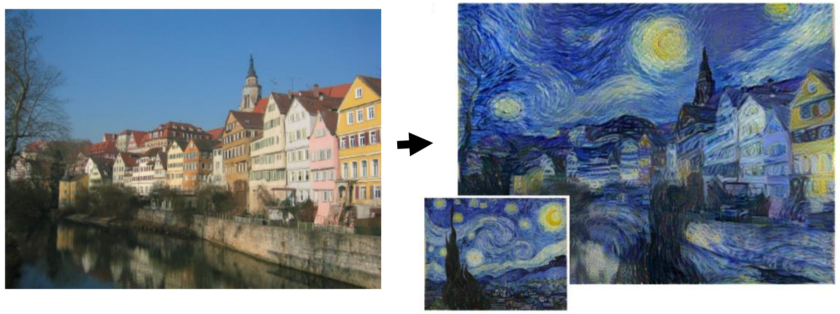
1. Output is high-dimensional and structured



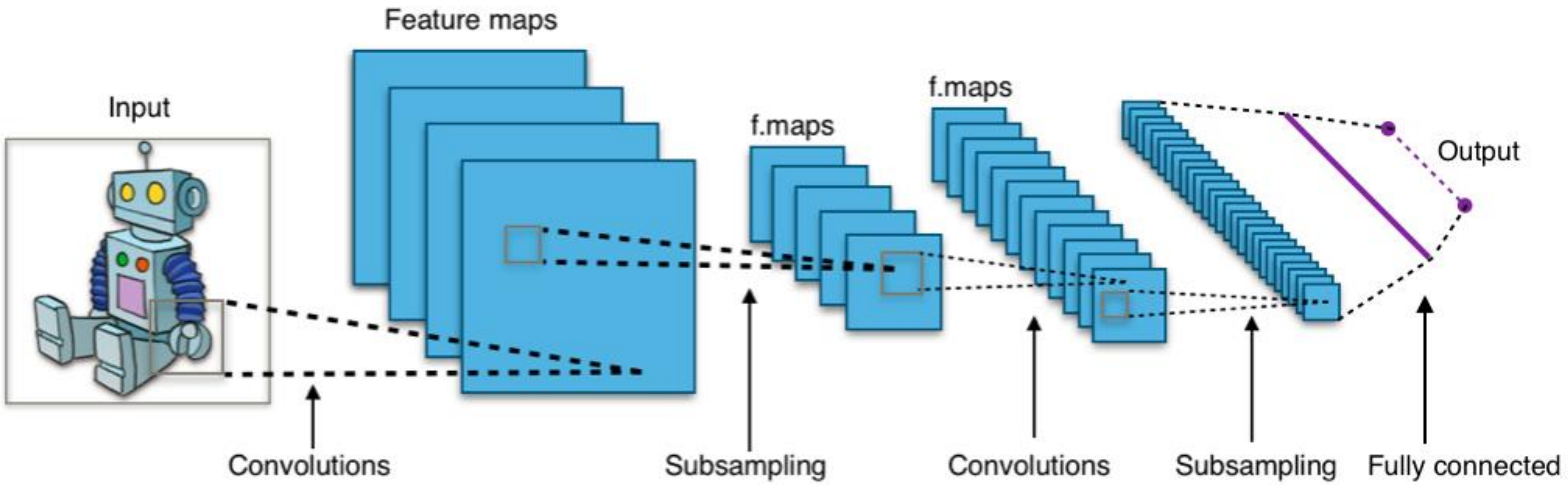
2. Uncertainty in mapping; many plausible outputs

"this small bird has a pink breast and crown…"

3. Lack of supervised training data



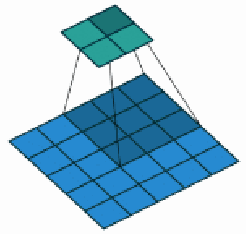
A standard CNN for classification



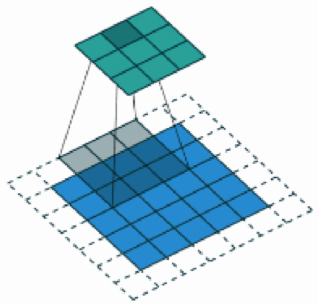
Need new network modules to generate same-sized output!

Deconvolution

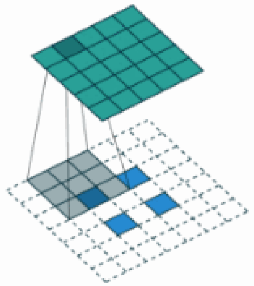
Compare to convolution with different params



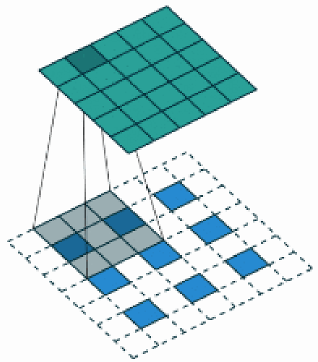
No padding, strides



Padding, strides



No padding, strides, transposed



Padding, strides, transposed

Deconvolution

Also known as: transpose conv, upconv, ...

Input

0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	3	0	3	0	0
0	0	0	0	0	0	0
0	0	1	0	1	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

Kernel

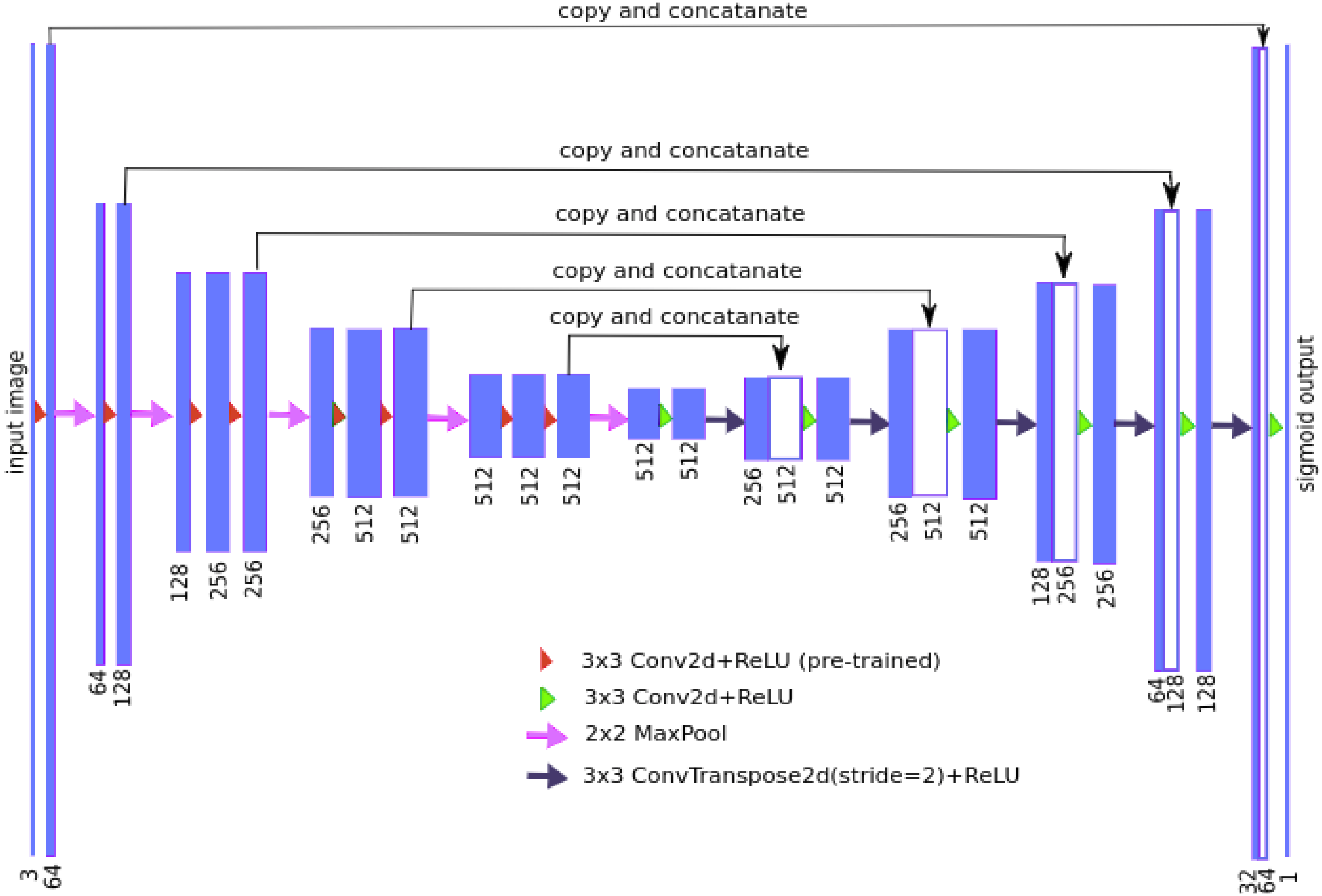
1	2	3
0	1	0
2	1	2

Output

3	6	12	6	9
0	3	0	3	0
7	5	16	5	9
0	1	0	1	0
2	1	4	1	2

Unet

A popular network structure to generate same-sized output



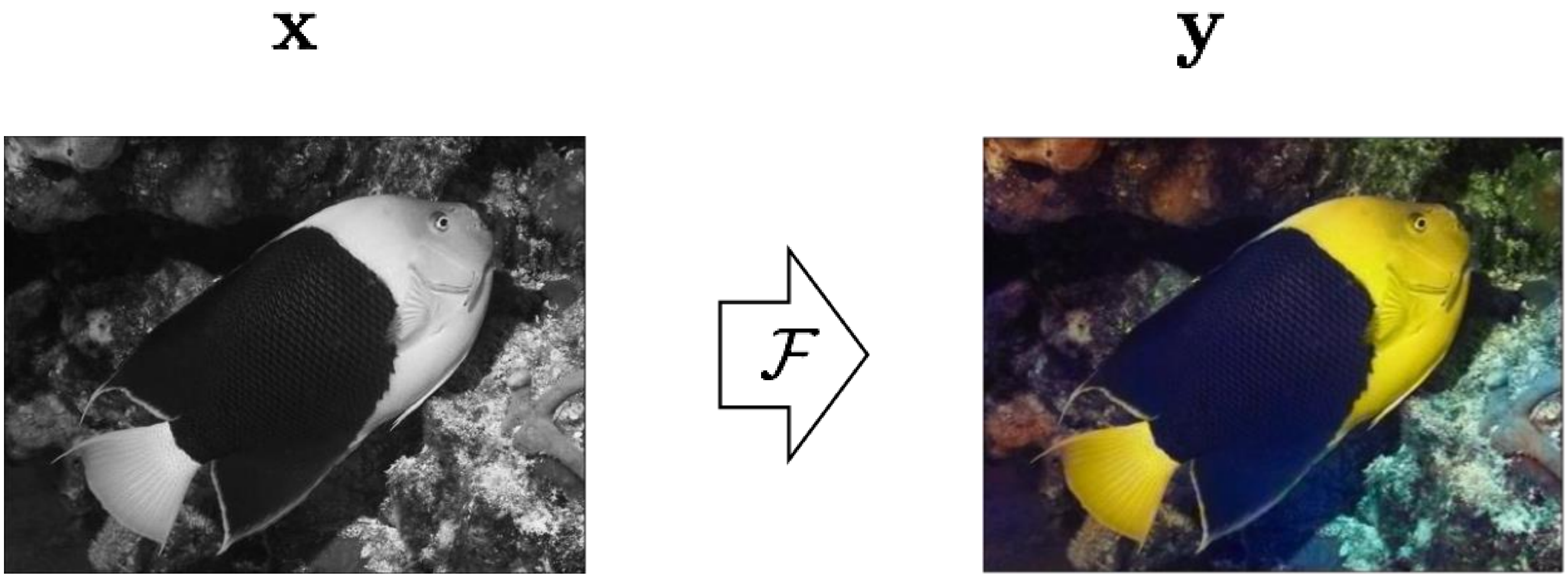
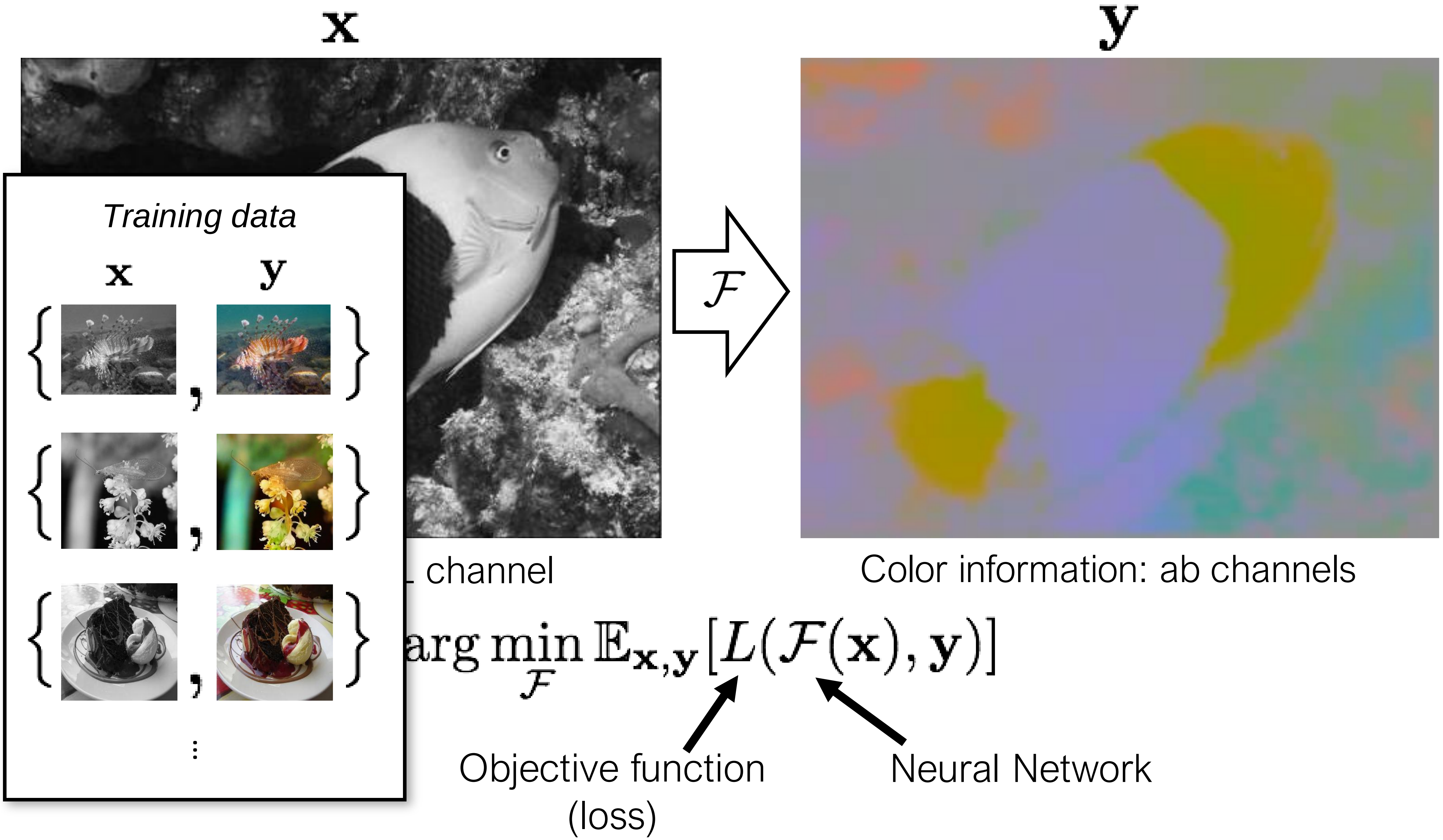


Image Colorization

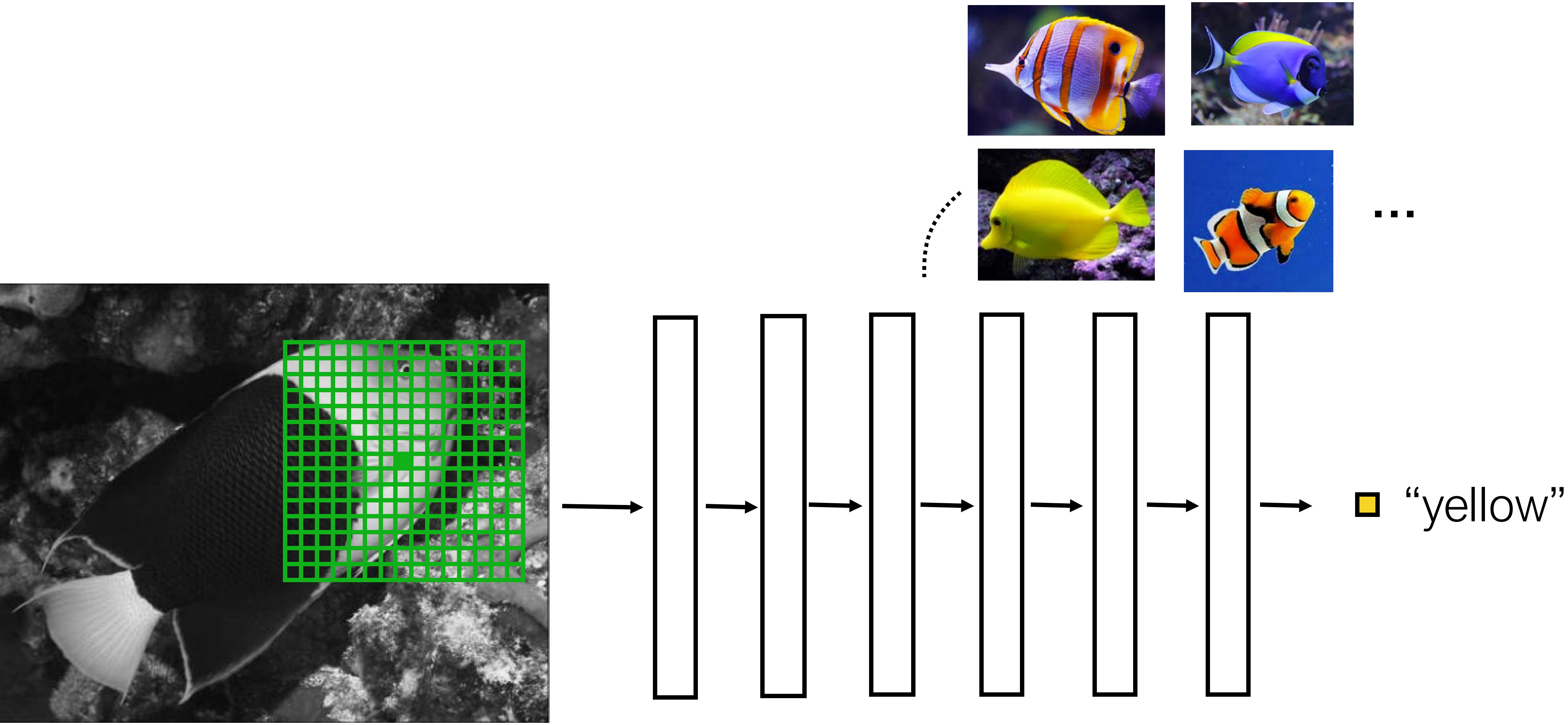
$\mathbf{x}$ $\mathbf{y}$

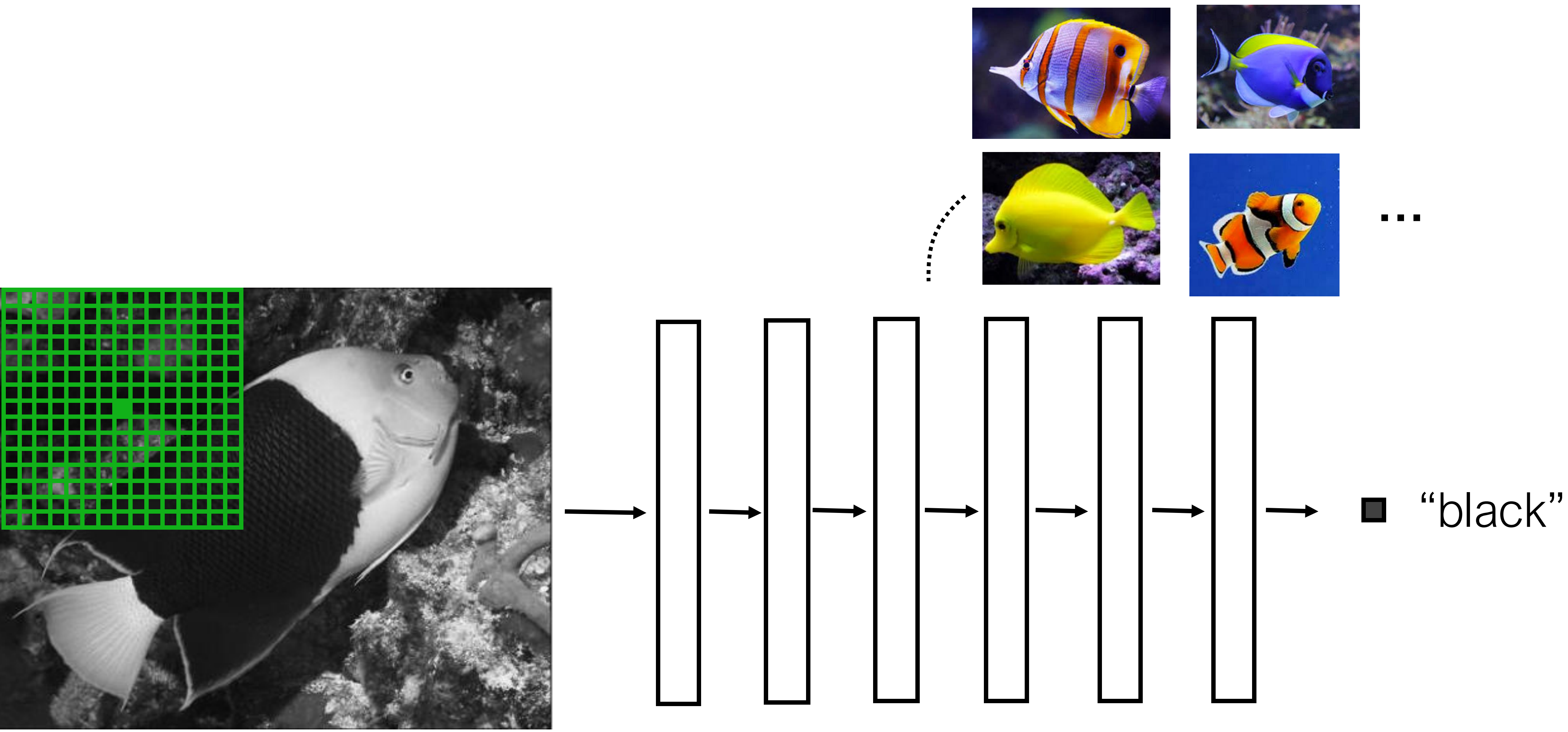
$$\arg\min_{\mathcal{F}} \mathbb{E}_{\mathbf{x},\mathbf{y}}[L(\mathcal{F}(\mathbf{x}), \mathbf{y})]$$

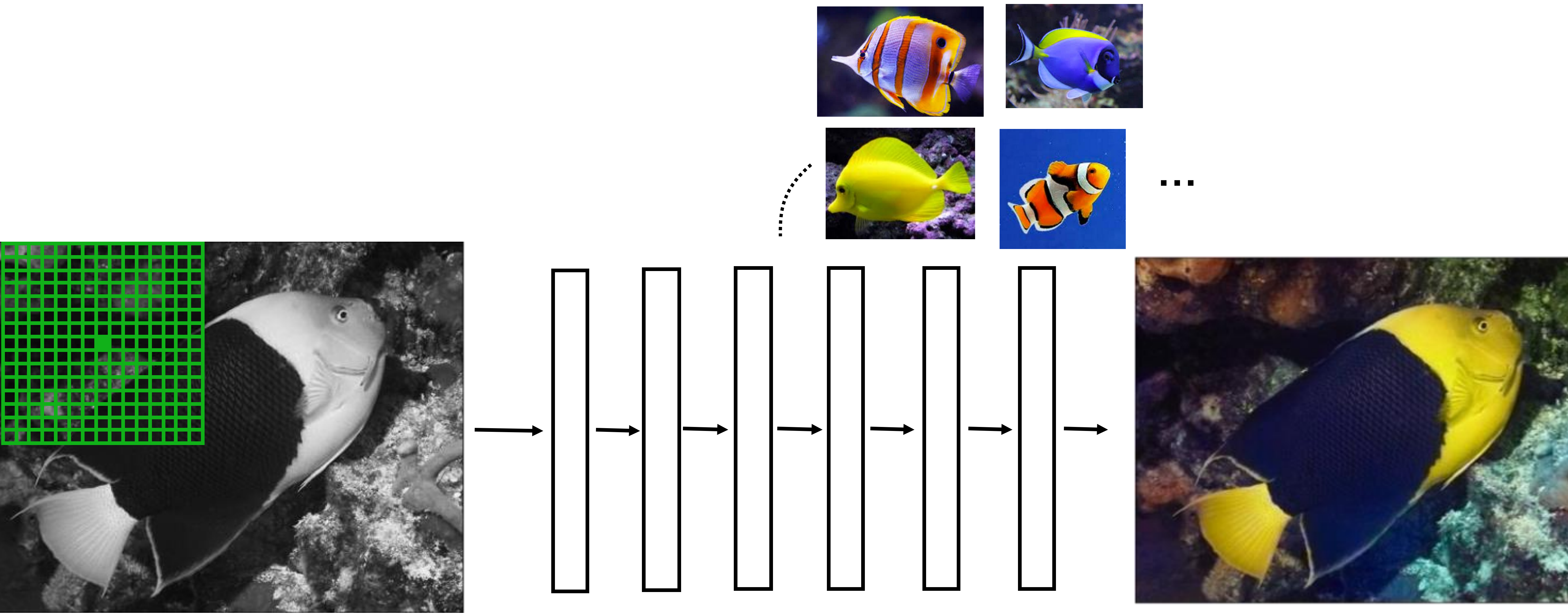
"**What** should I do" "**How** should I do it?"



x
y
Training data
x
y
channel
Color information: ab channels
$\arg\min_{\mathcal{F}} \mathbb{E}_{\mathbf{x},\mathbf{y}}[L(\mathcal{F}(\mathbf{x}),\mathbf{y})]$
Objective function (loss)
Neural Network


...
"yellow"


“black”



Basic loss functions

Prediction: $\hat{\mathbf{y}} = \mathcal{F}(\mathbf{x})$

Truth: $\mathbf{y}$

Classification (cross-entropy):

$$L(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_i \hat{\mathbf{y}}_i \log \mathbf{y}_i$$

How many extra bits it takes to correct the predictions

Least-squares regression:

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

How far off we are in Euclidean distance

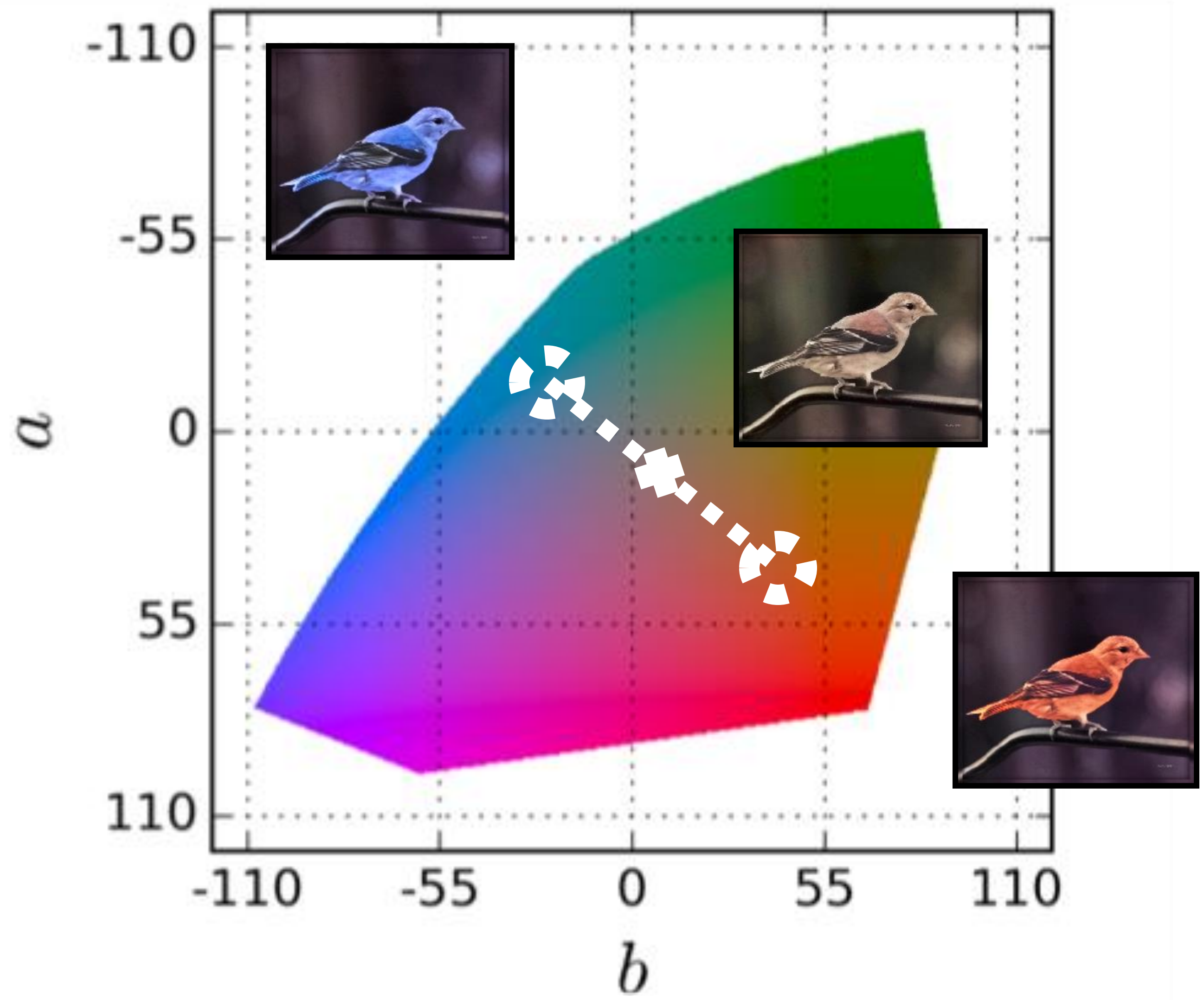
Designing loss functions

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

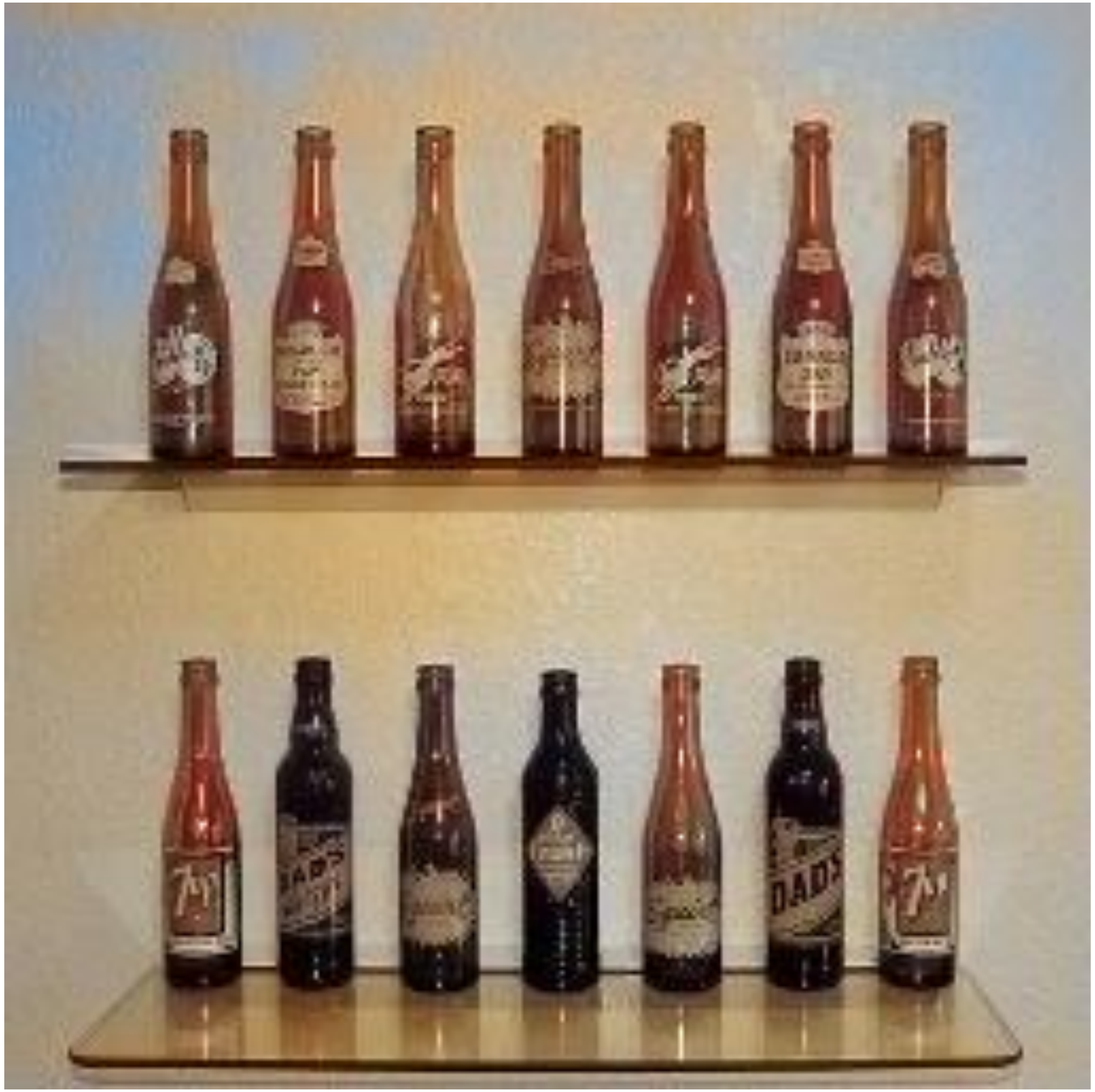
Input

Zhang et al. 2016

Ground truth

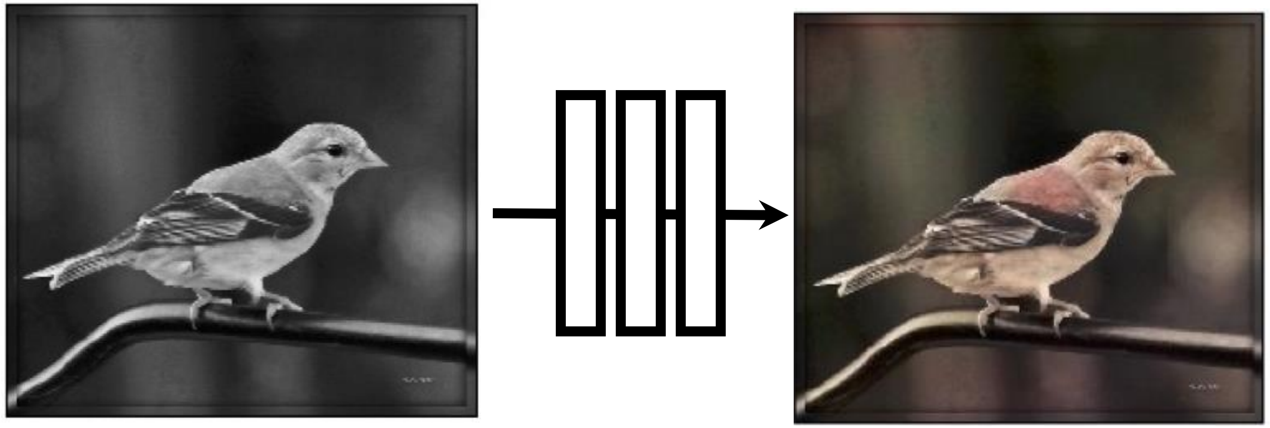
Color distribution cross-entropy loss with colorfulness enhancing term.

[Zhang, Isola, Efros, ECCV 2016]


DADS

Designing loss functions

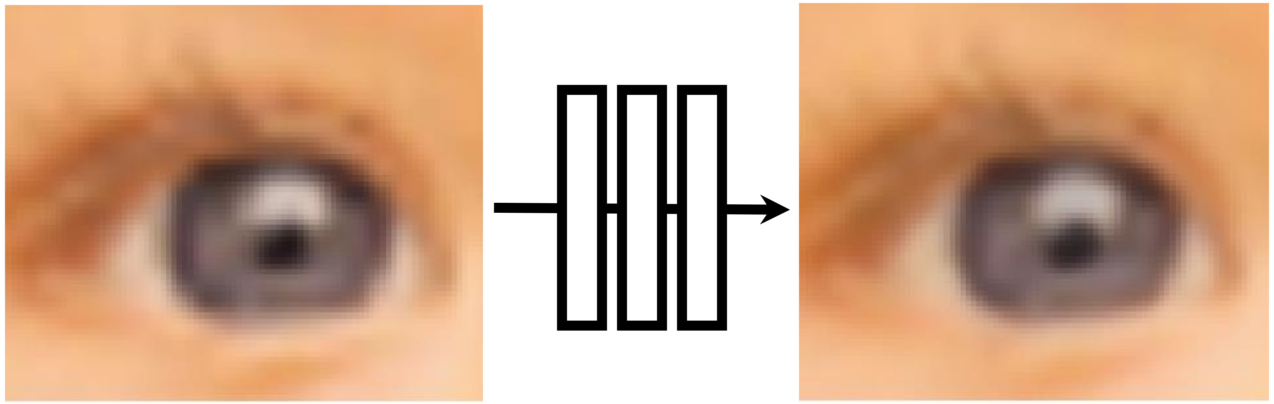
Image colorization



[Zhang, Isola, Efros, ECCV 2016]

L2 regression

Super-resolution

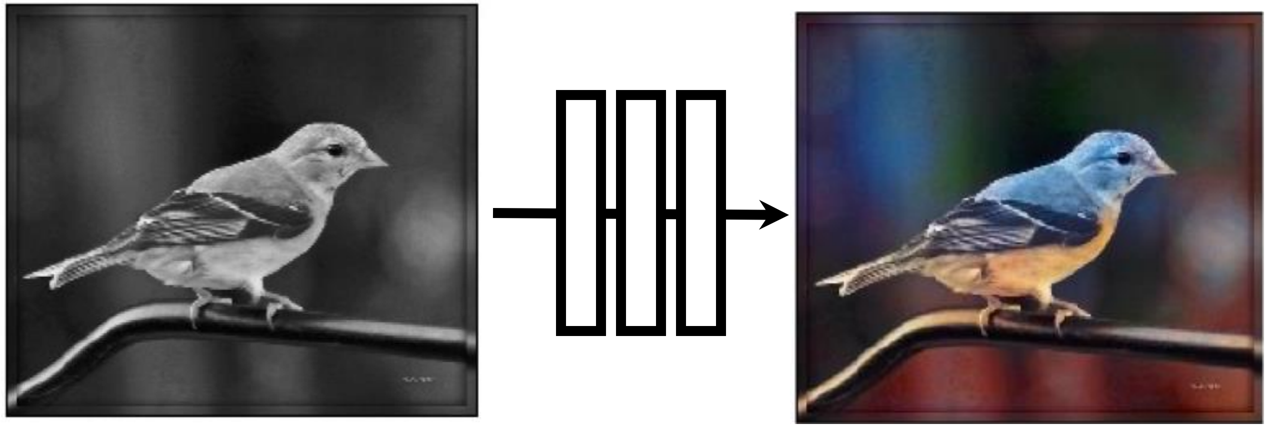


[Johnson, Alahi, Li, ECCV 2016]

L2 regression

Designing loss functions

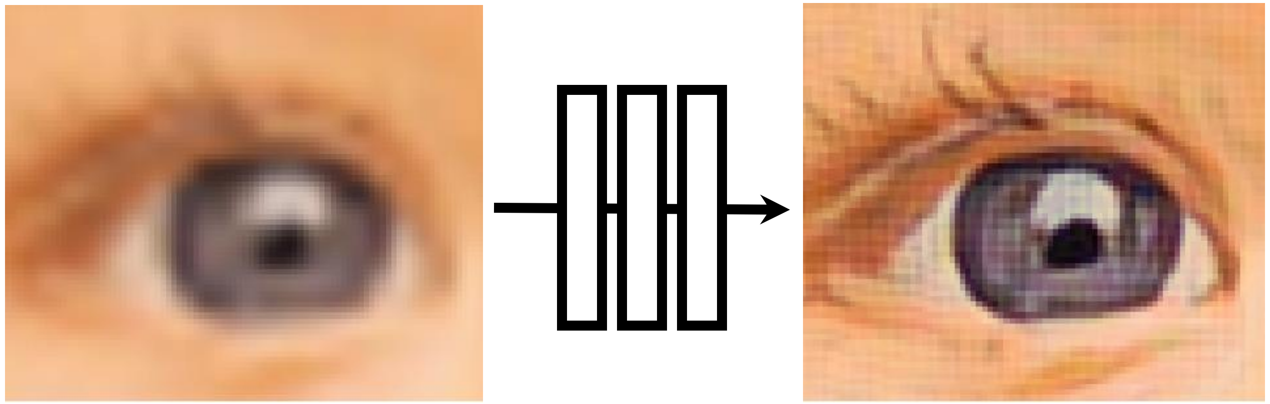
Image colorization



[Zhang, Isola, Efros, ECCV 2016]

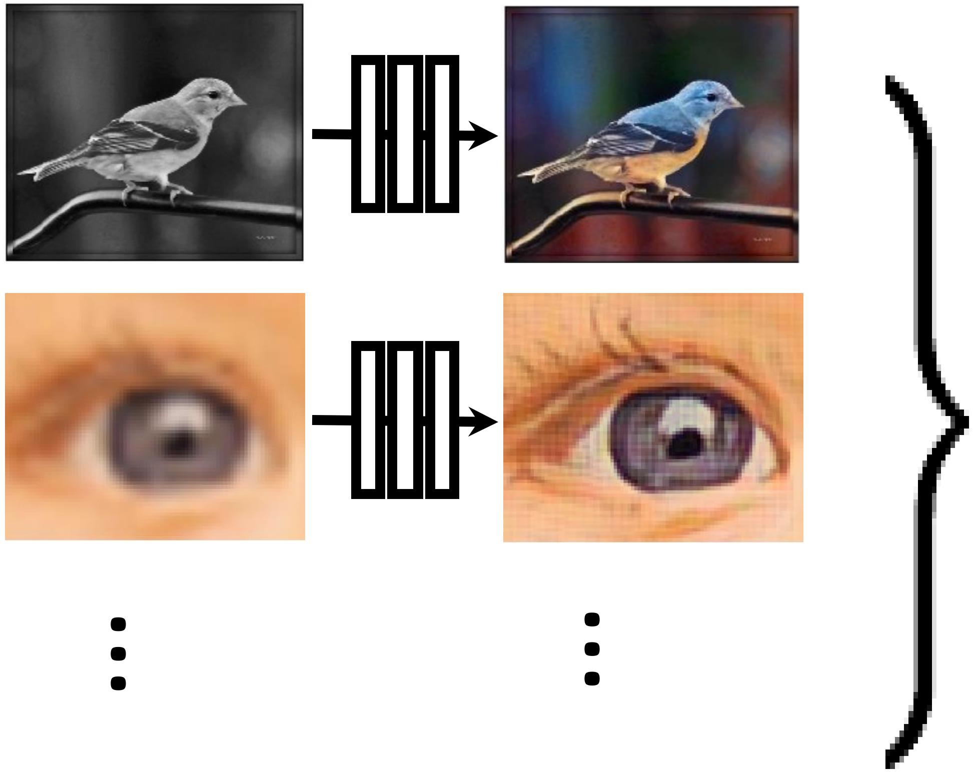
Cross entropy objective, with colorfulness term

Super-resolution



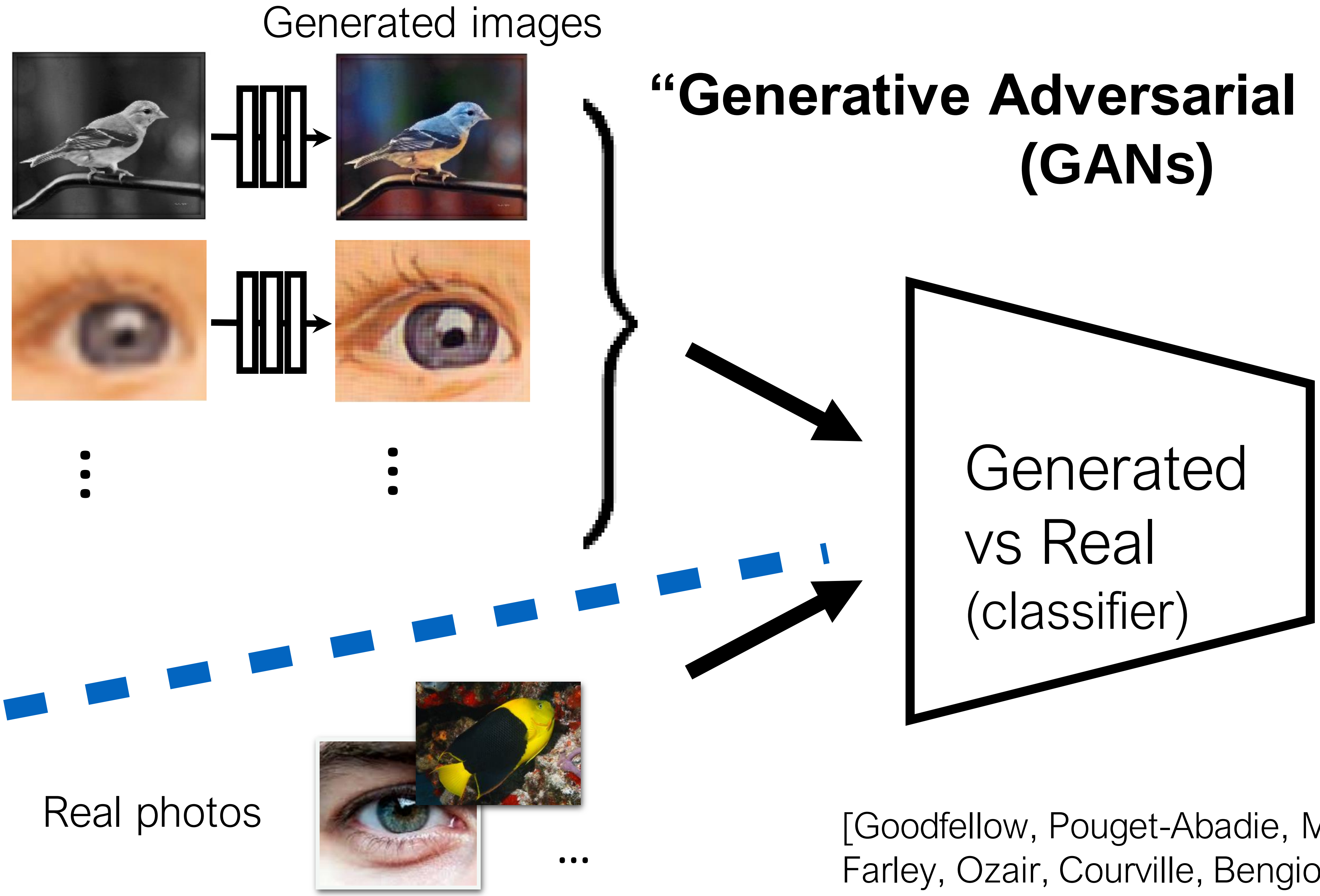
[Johnson, Alahi, Li, ECCV 2016]

Deep feature covariance matching objective



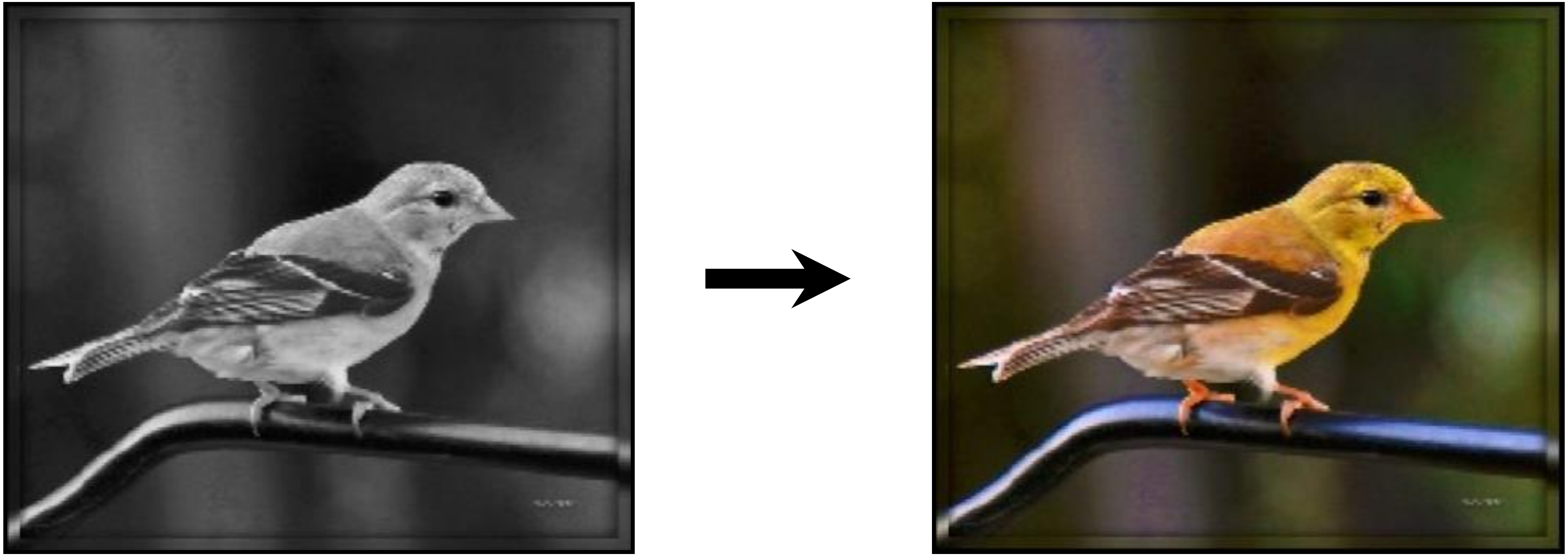
Universal loss?

"Generative Adversarial Network" (GANs)

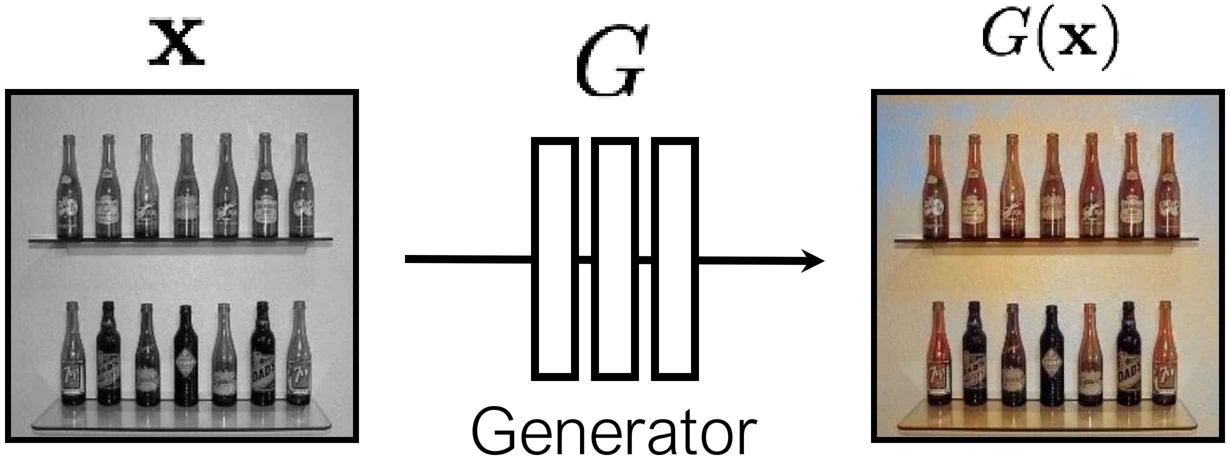


[Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, Bengio 2014]

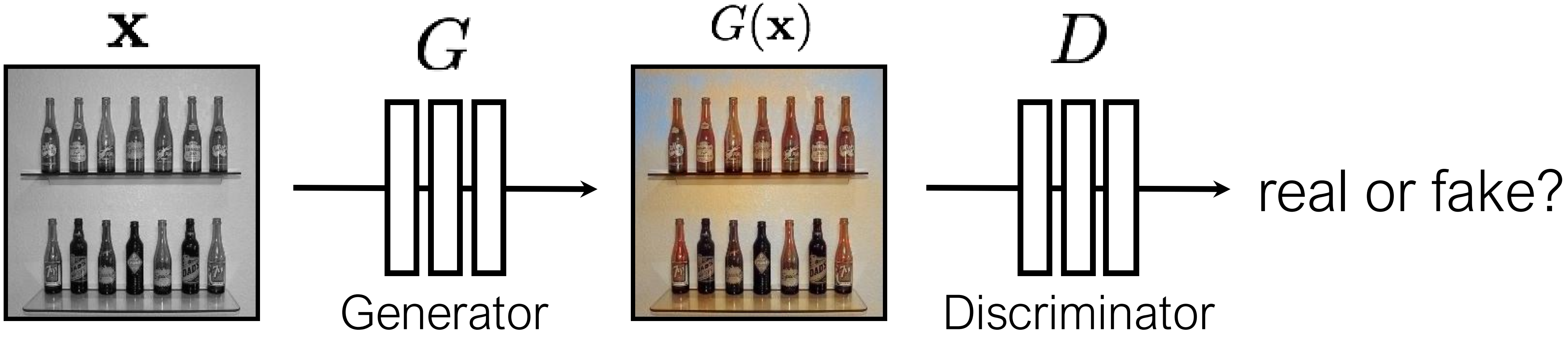
Conditional GANs



[Goodfellow et al., 2014]
[Isola et al., 2017]



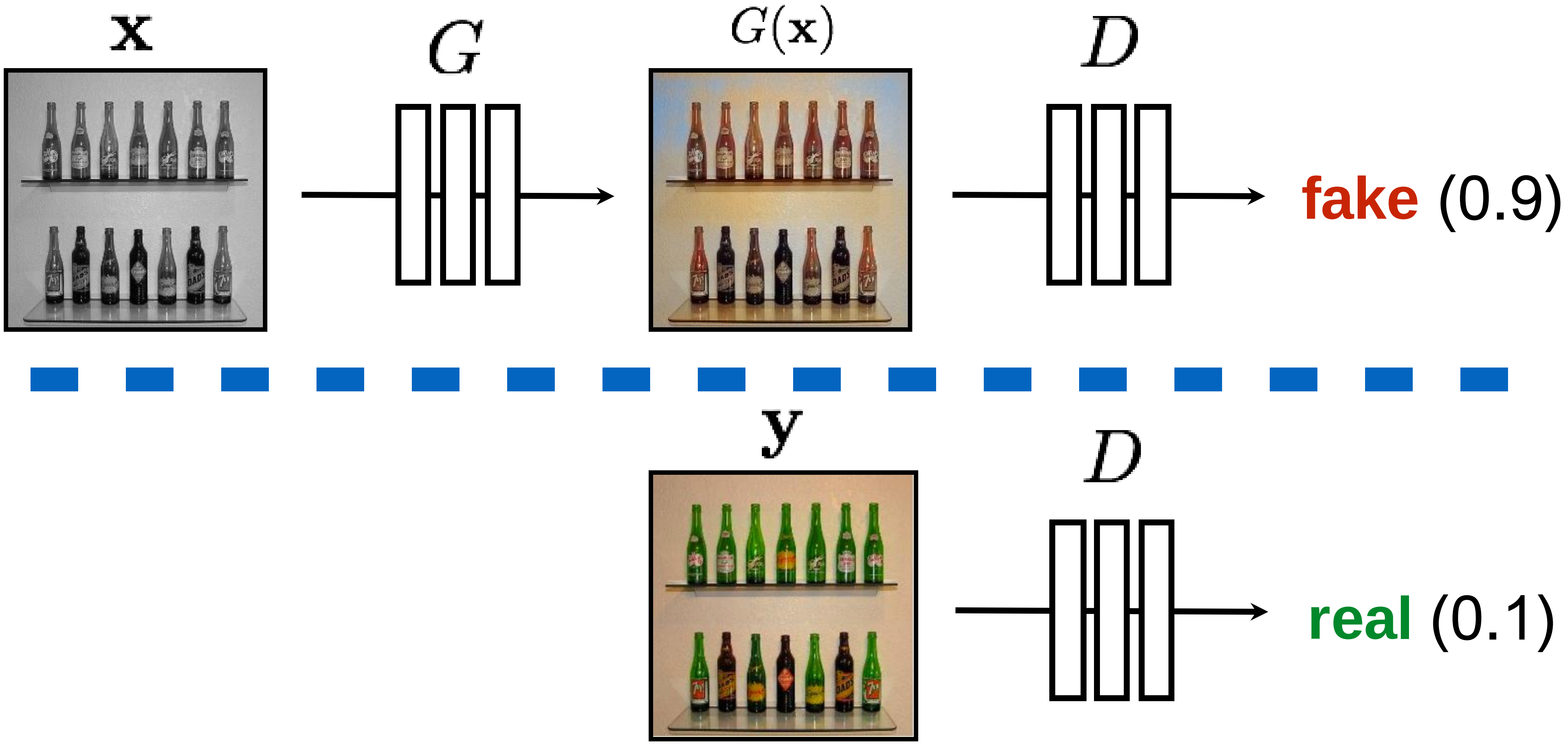
[Goodfellow et al., 2014]



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

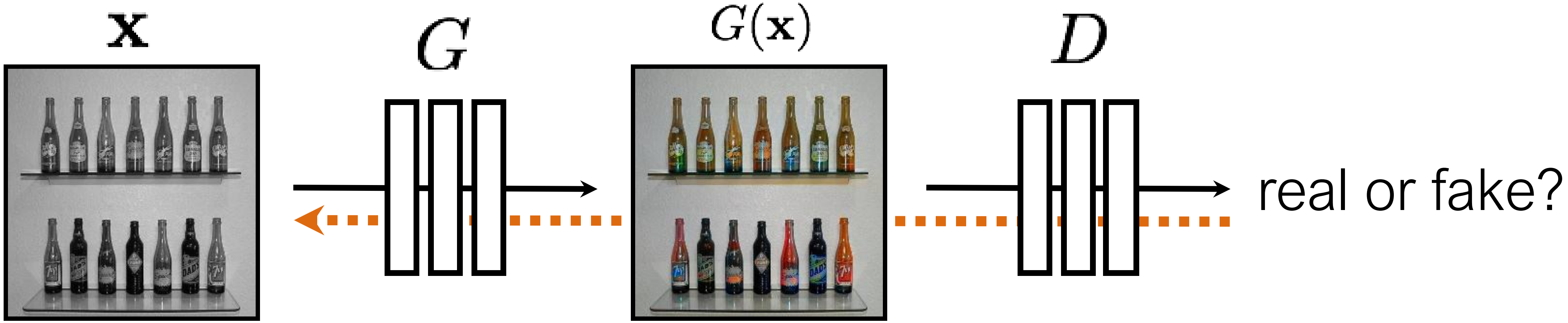
D tries to identify the fakes

[Goodfellow et al., 2014]



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

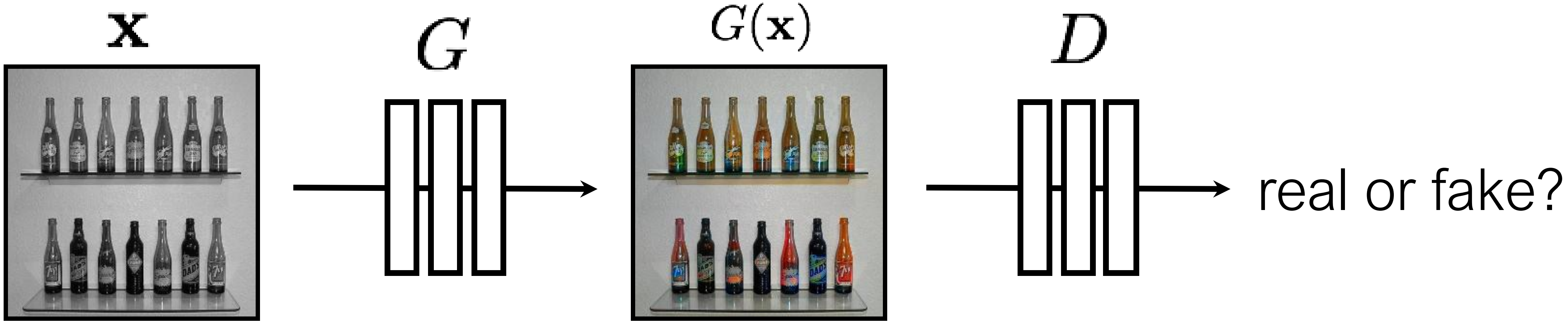
[Goodfellow et al., 2014]



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

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

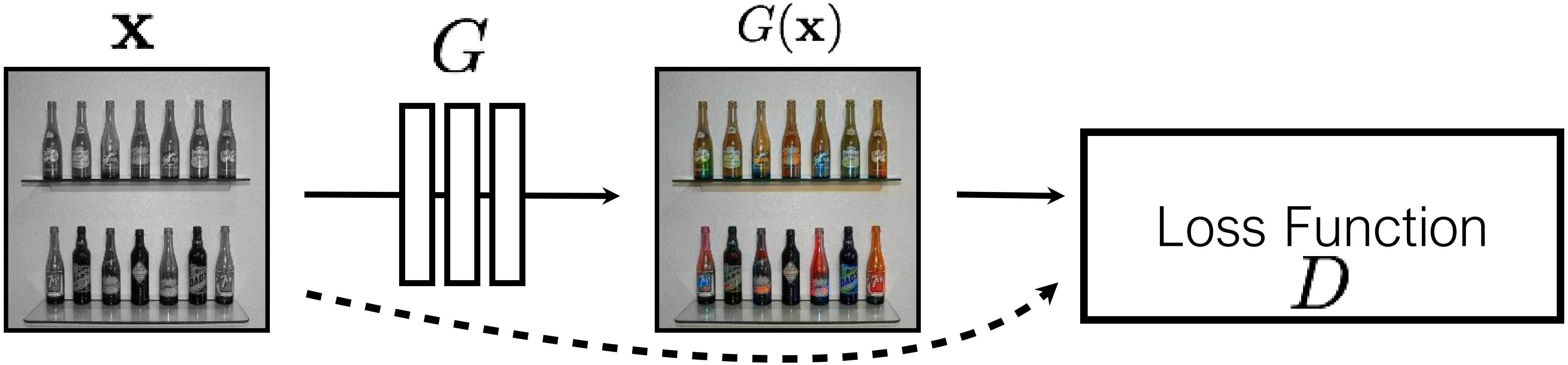
[Goodfellow et al., 2014]



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

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

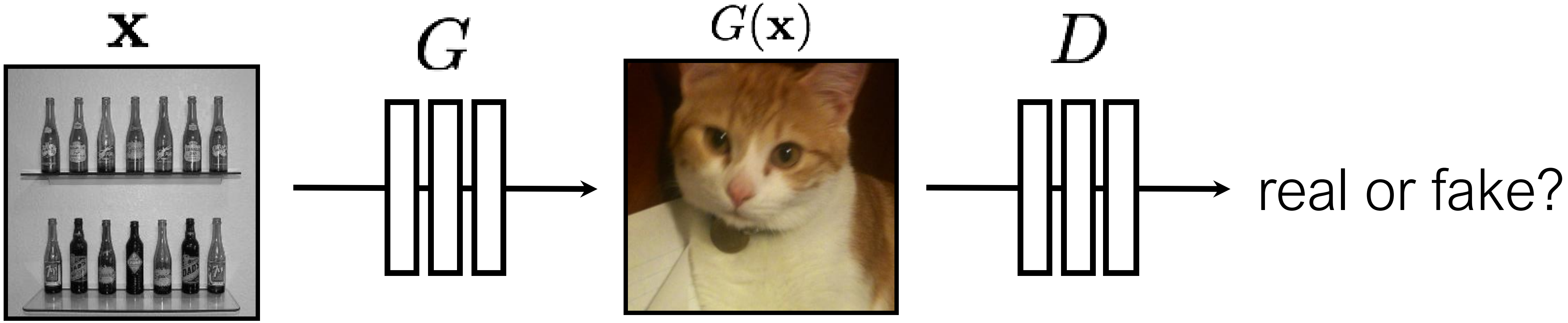
[Goodfellow et al., 2014]



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

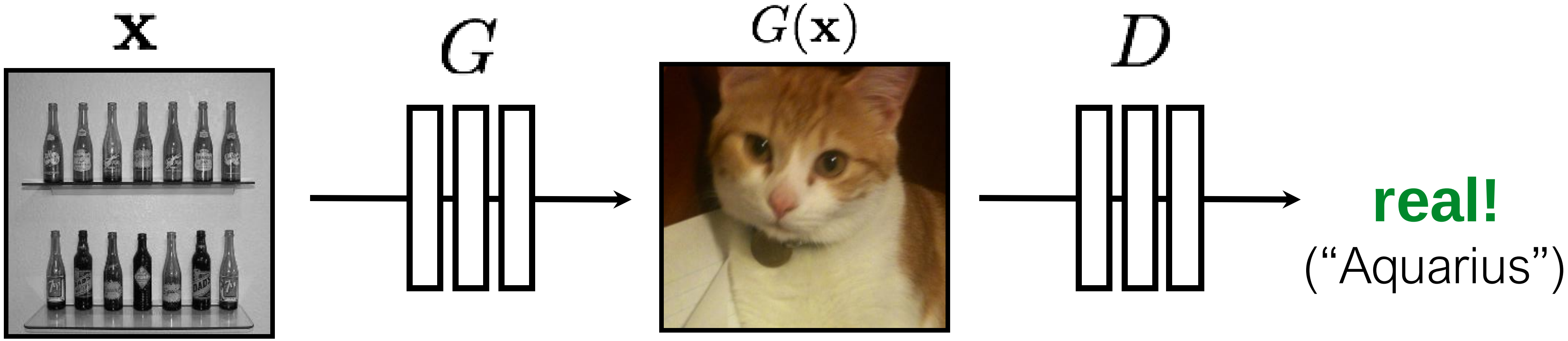
Rather than being hand-designed, it is *learned*.

[Goodfellow et al., 2014]
[Isola et al., 2017]



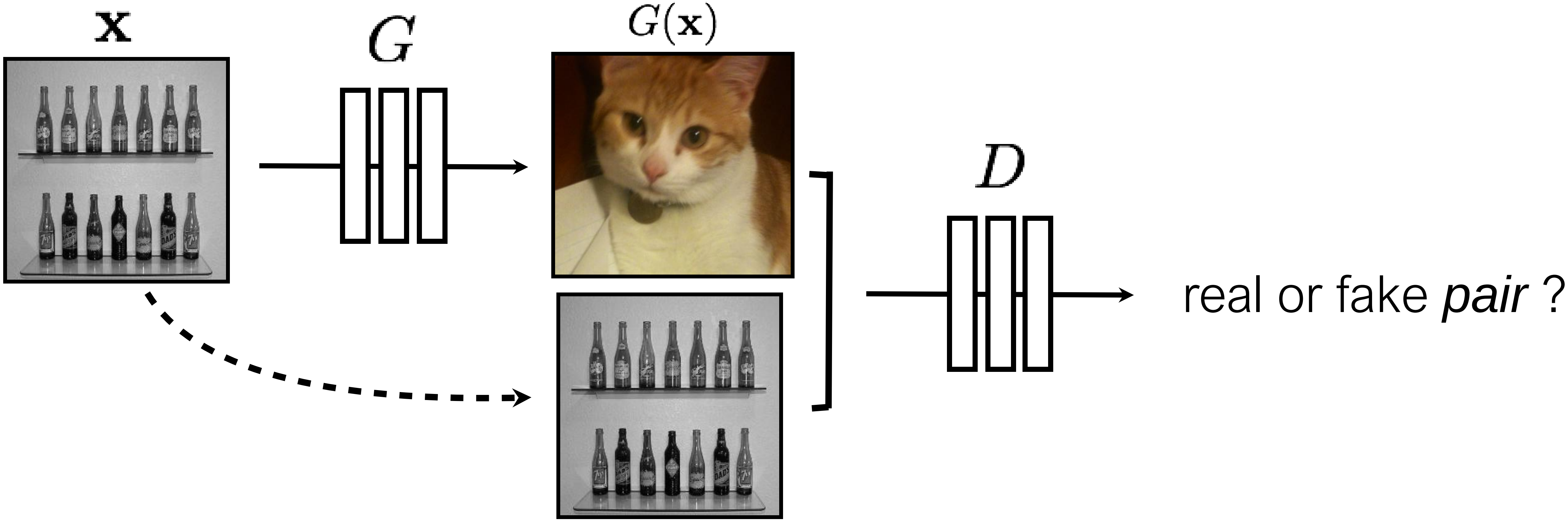
$$\arg\min_G \max_D \; \mathbb{E}_{\mathbf{x},\mathbf{y}}[\; \log D(G(\mathbf{x})) \;+\; \log(1 - D(\mathbf{y})) \;]$$

[Goodfellow et al., 2014]



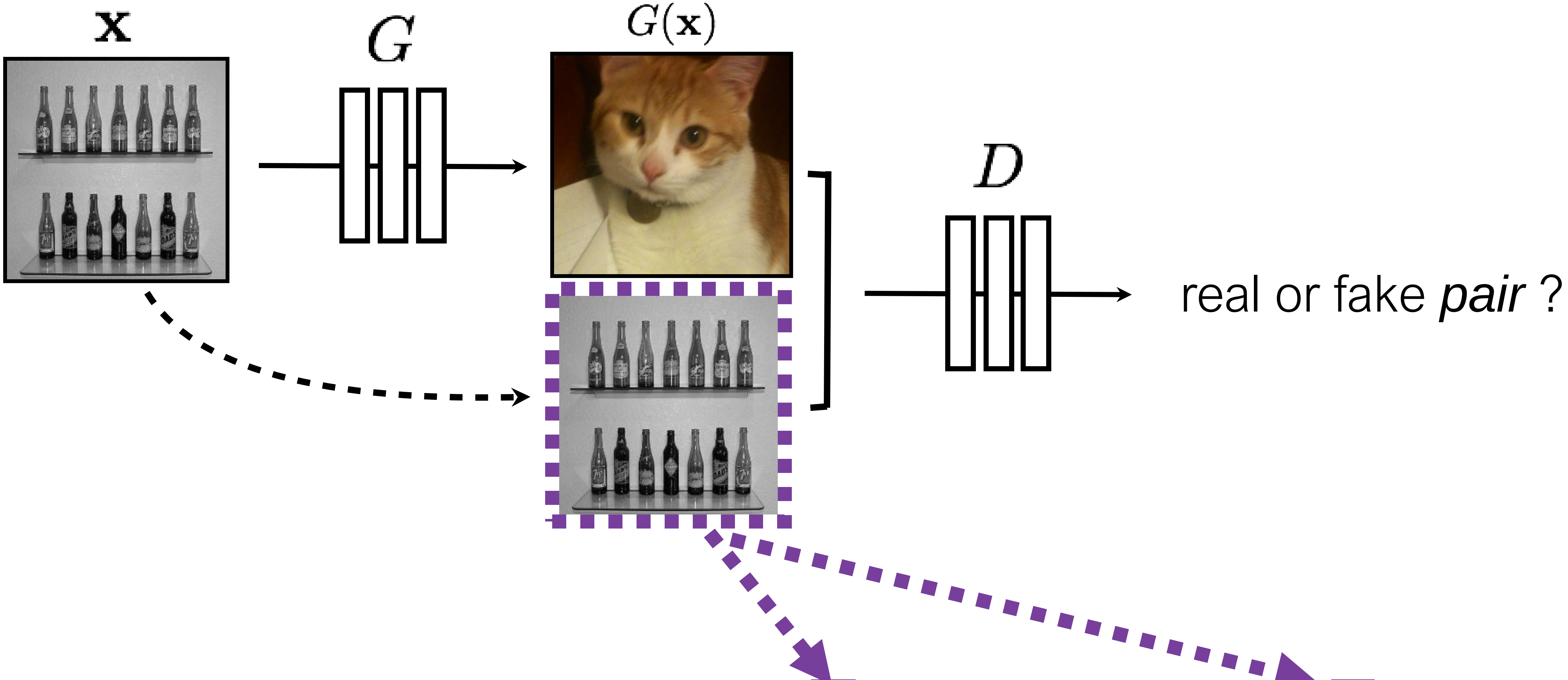
$$\arg\min_{G}\max_{D}\ \mathbb{E}_{\mathbf{x},\mathbf{y}}[\ \log D(G(\mathbf{x})) \quad + \quad \log(1 - D(\mathbf{y}))\ \ ]$$

[Goodfellow et al., 2014]



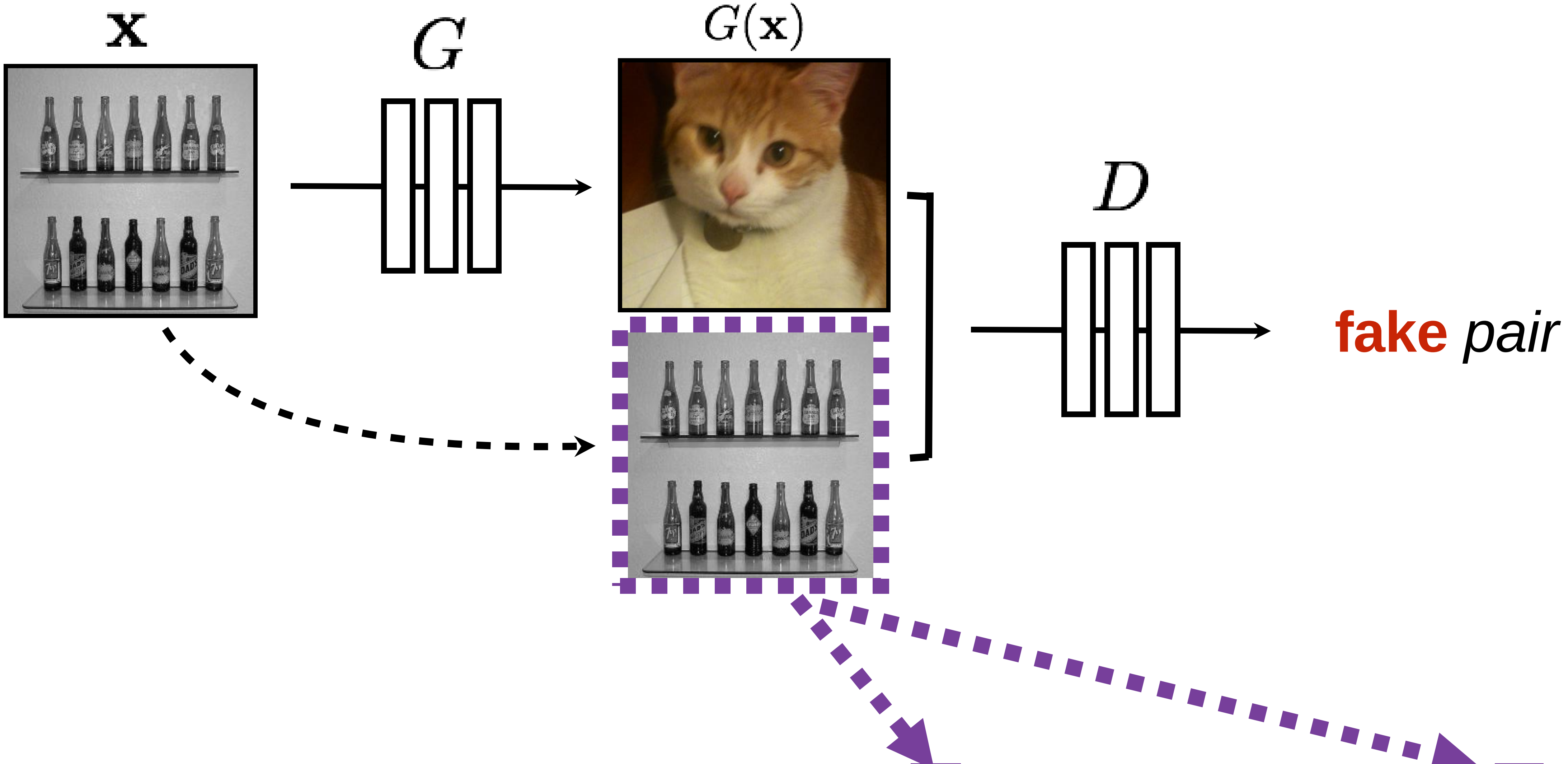
$$\arg\min_{G}\max_{D}\;\mathbb{E}_{\mathbf{x},\mathbf{y}}[\;\log D(G(\mathbf{x}))\quad+\quad\log(1-D(\mathbf{y}))\;]$$

[Goodfellow et al., 2014]
[Isola et al., 2017]



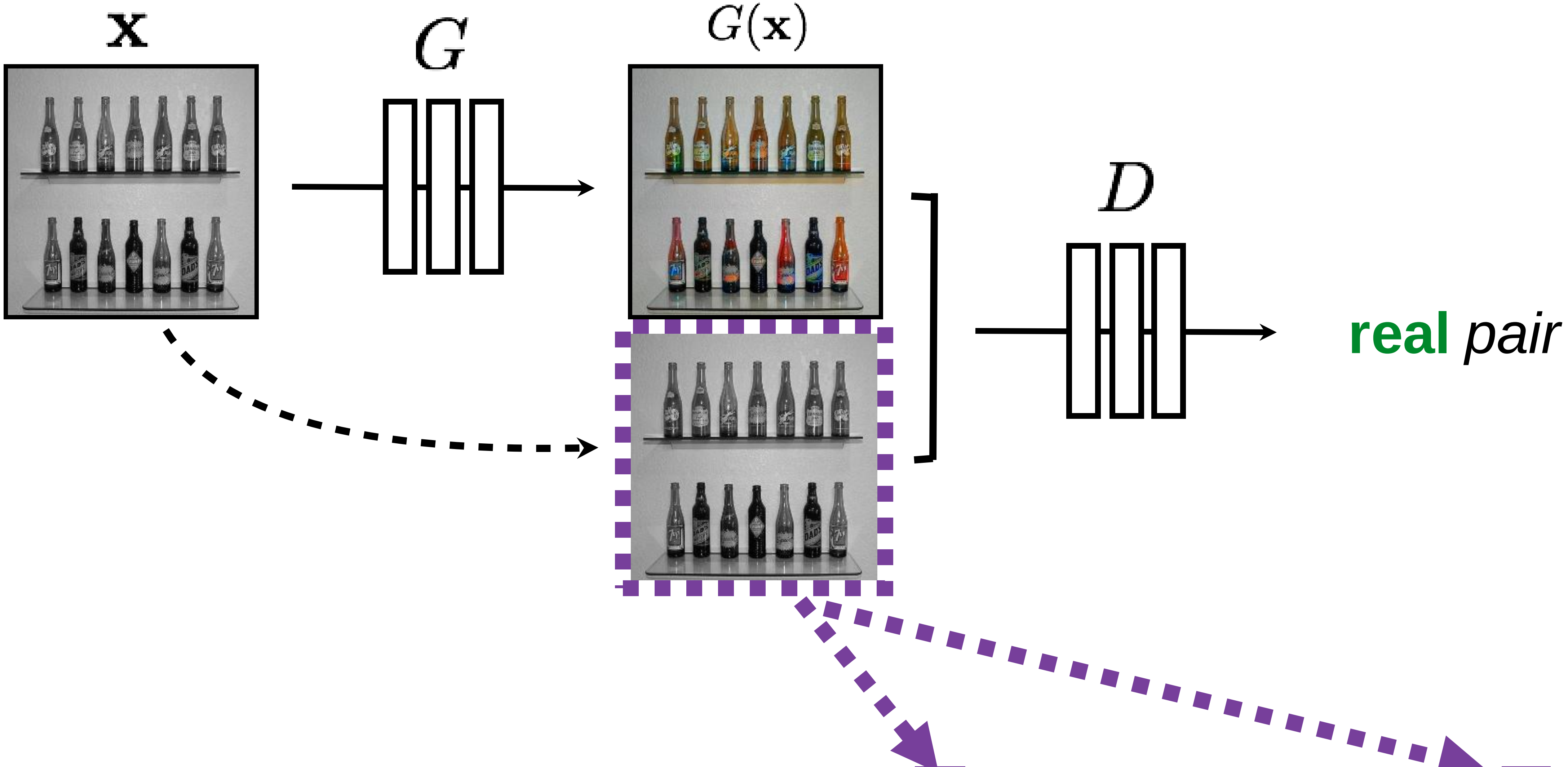
$$\arg\min_{G}\max_{D}\ \mathbb{E}_{\mathbf{x},\mathbf{y}}[\ \log D(\boxed{\mathbf{x}}, G(\mathbf{x})) + \log(1 - D(\boxed{\mathbf{x}}, \mathbf{y}))\ ]$$

[Goodfellow et al., 2014]
[Isola et al., 2017]



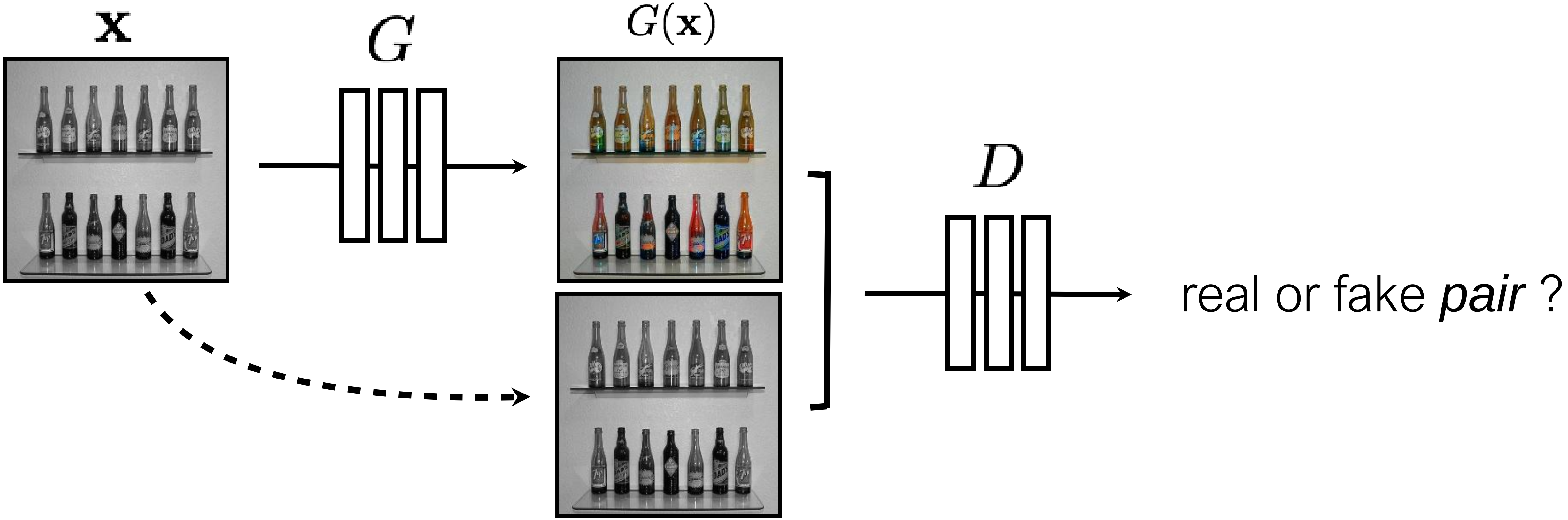
$$\arg\min_{G}\max_{D}\ \mathbb{E}_{\mathbf{x},\mathbf{y}}[\ \log D(\boxed{\mathbf{x}}, G(\mathbf{x})) + \log(1 - D(\boxed{\mathbf{x}}, \mathbf{y}))\ ]$$

[Goodfellow et al., 2014]
[Isola et al., 2017]



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

[Goodfellow et al., 2014]
[Isola et al., 2017]



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

[Goodfellow et al., 2014]
[Isola et al., 2017]

BW → Color

Input Output Input Output Input Output

Data from [Russakovsky et al. 2015]

Input

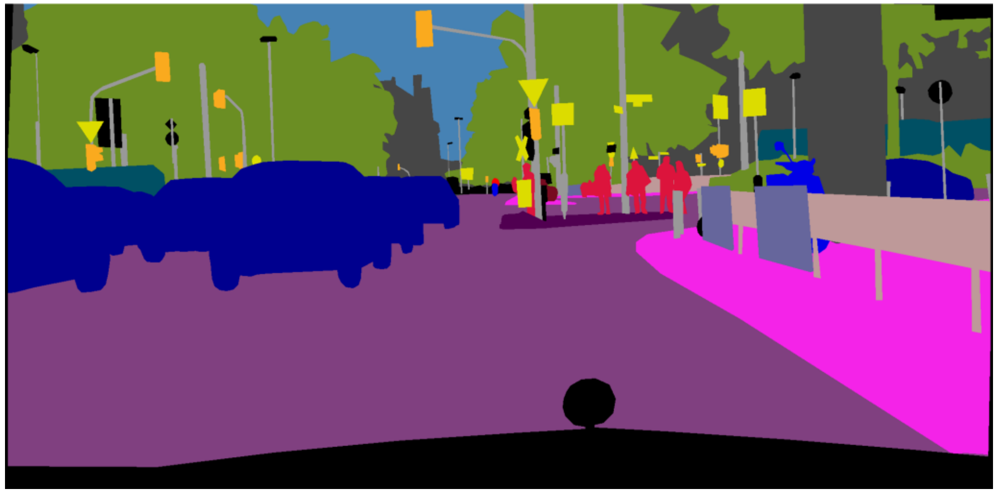
Output

Groundtruth

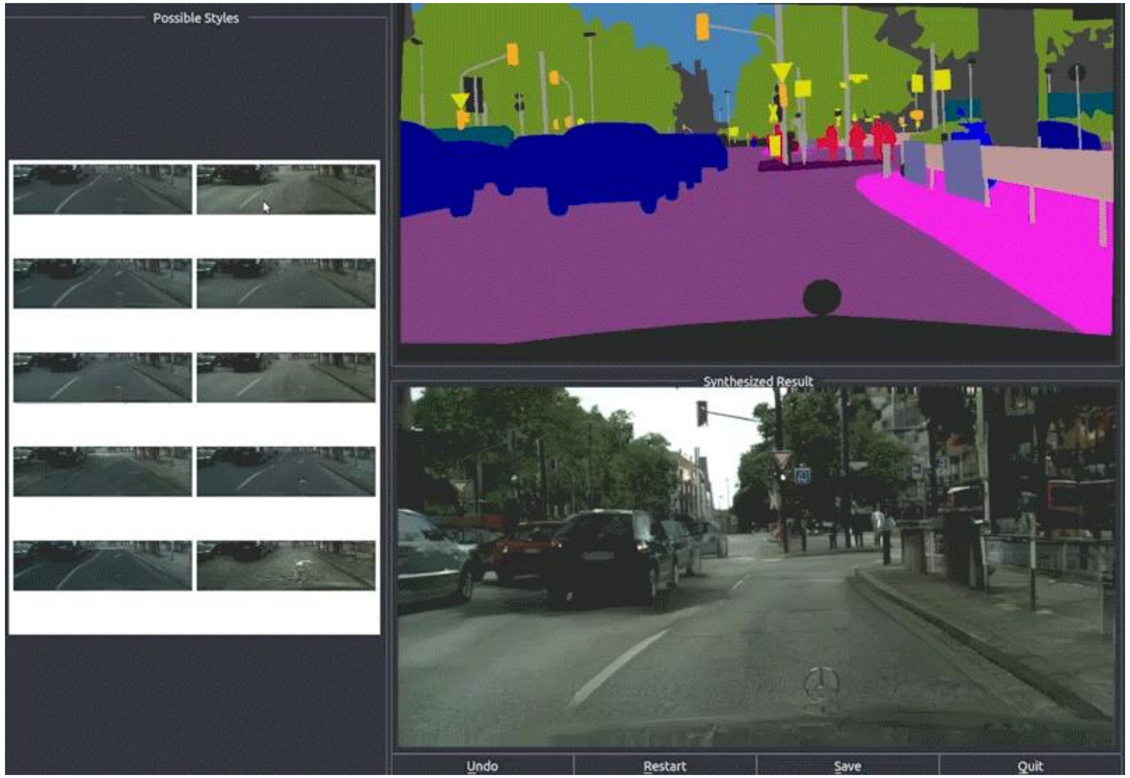
Data from
[maps.google.com]

Labels → Street Views

Input labels



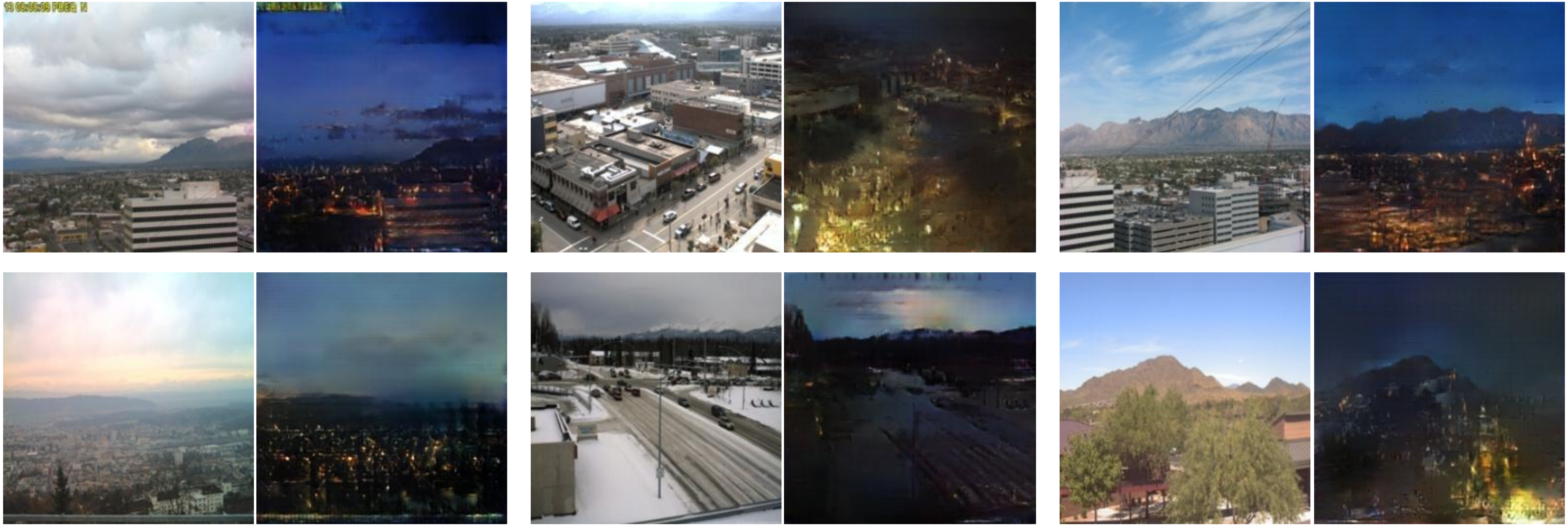
Synthesized image



Data from [Wang et al, 2018]

Day → Night

Input | Output | Input | Output | Input | Output

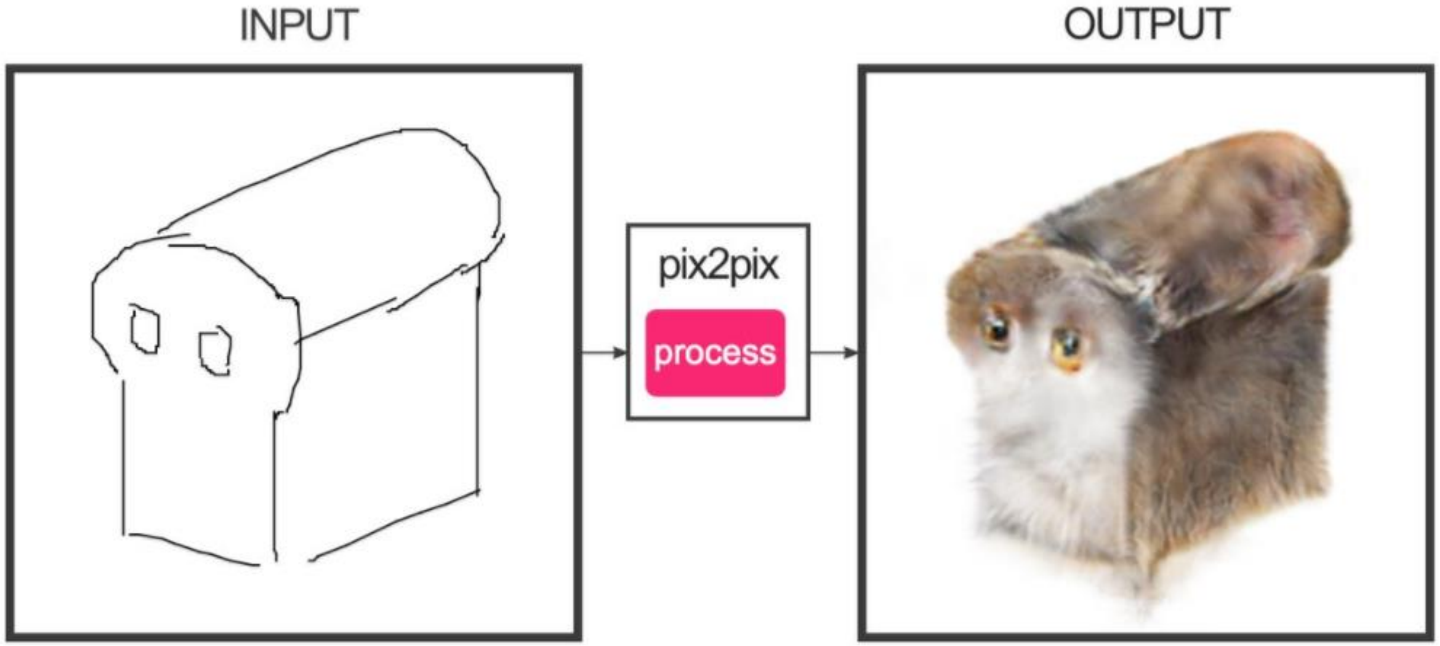


Data from [Laffont et al., 2014]

Edges → Images

Input Output Input Output Input Output

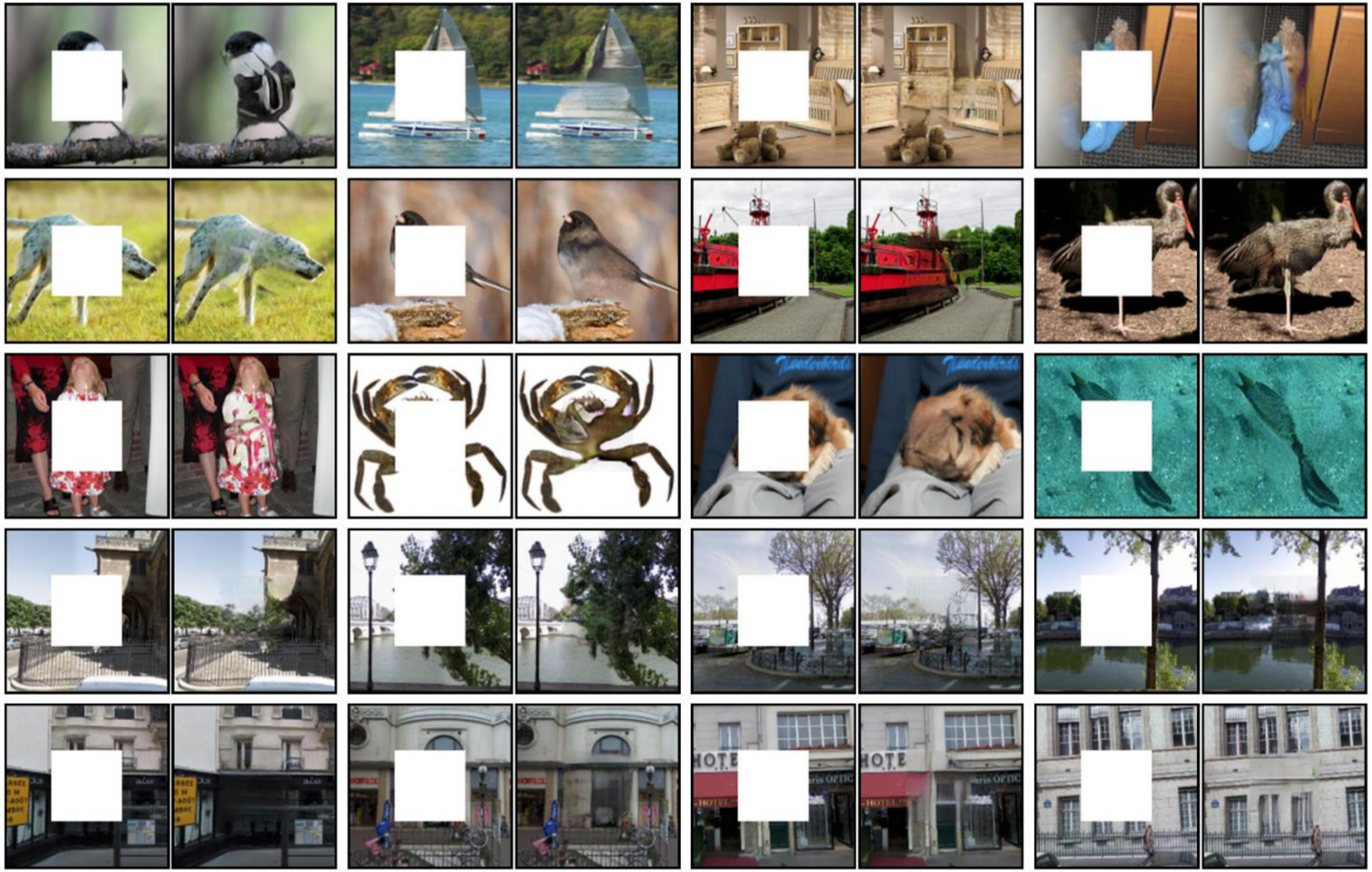
Edges from [Xie & Tu, 2015]



Ivy Tasi @ivymyt

Vitaly Vidmirov @vvid

Image Inpainting

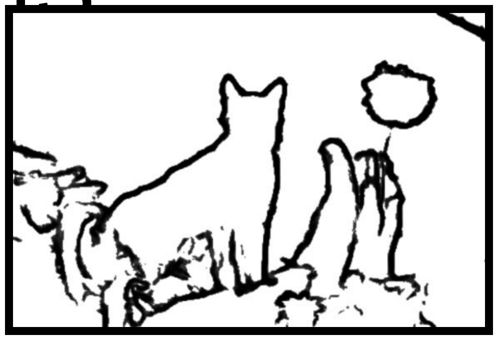


Data from [Pathak et al., 2016]

Pose-guided Generation

Data from [Ma et al., 2018]

Challenges —> Solutions



1. Output is high-dimensional, structured object

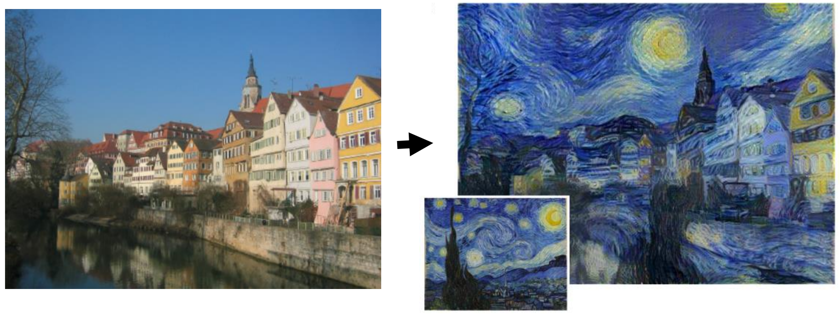
   —> Use a deep net, D, to analyze output!

“this small bird has a pink breast and crown…”

2. Uncertainty in mapping; many plausible outputs

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

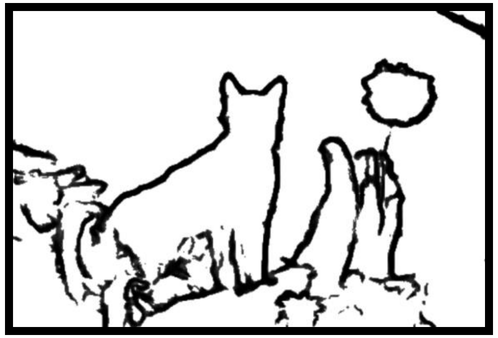
3. Lack of supervised training data



Challenges —> Solutions

1. Output is high-dimensional, structured object

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

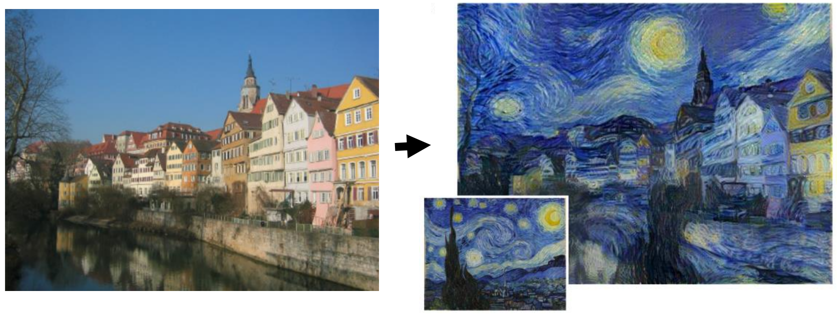


2. Uncertainty in mapping; many plausible outputs

“this small bird has a pink breast and crown…”

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

3. **Lack of supervised training data**



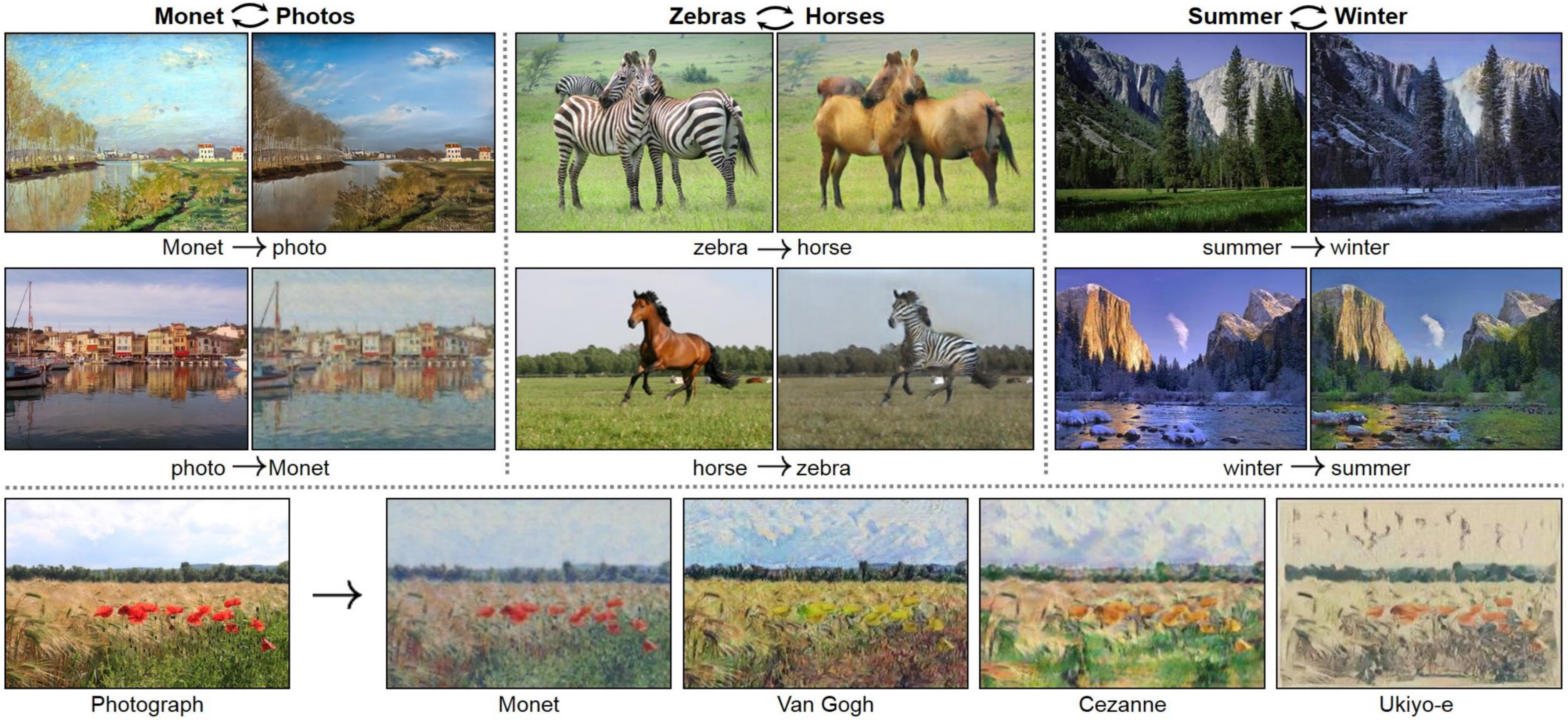
Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

Jun-Yan Zhu* **Taesung Park*** **Phillip Isola** **Alexei A. Efros**

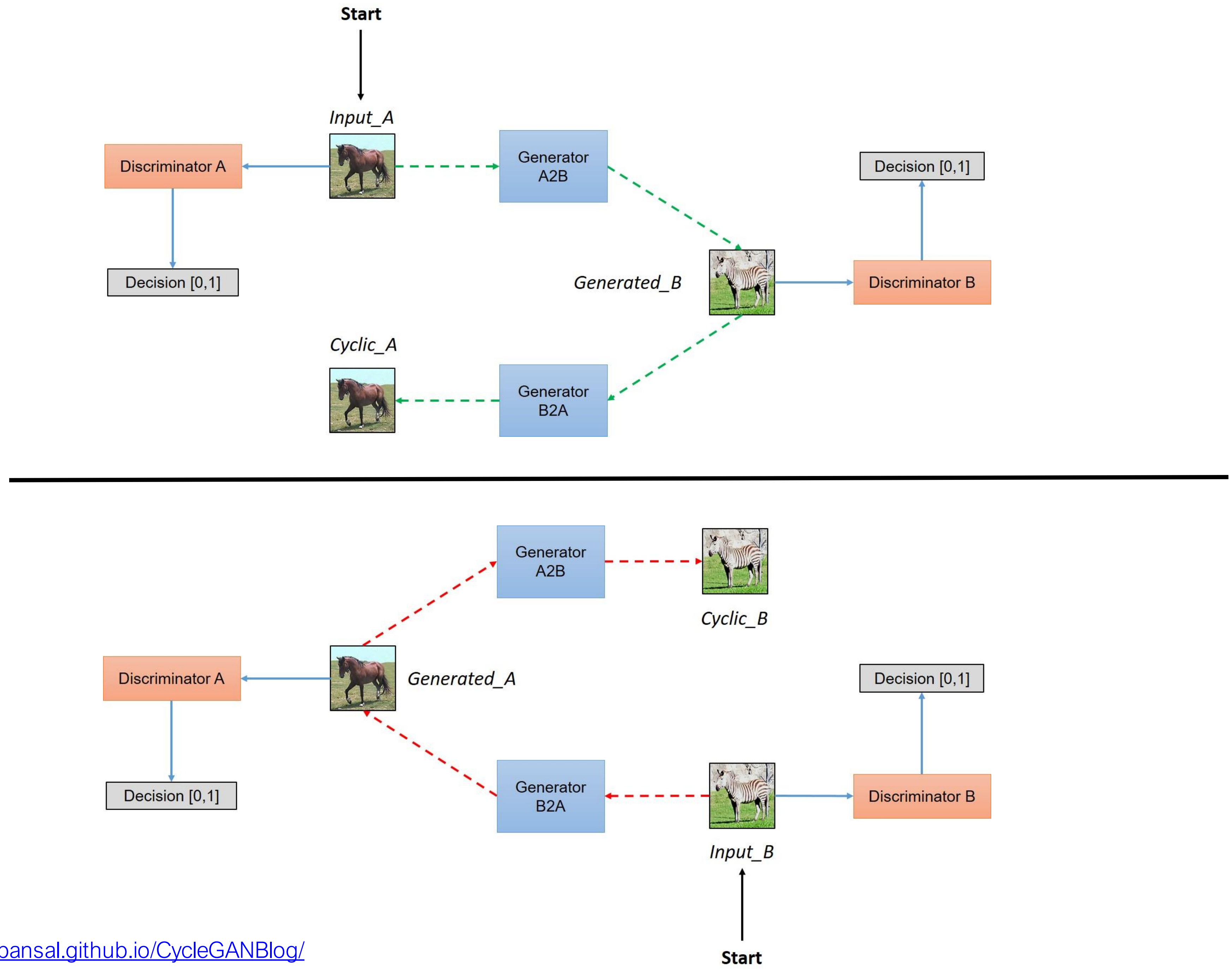
UC Berkeley

In ICCV 2017

[Paper] [Code (Torch)] [Code (PyTorch)]



https://junyanz.github.io/CycleGAN/



https://hardikbansal.github.io/CycleGANBlog/

StyleGAN

https://github.com/NVlabs/stylegan

Questions?