Self-supervision, Meta-supervision, Curiosity: Making Computers Study Harder

Alexei (Alyosha) Efros

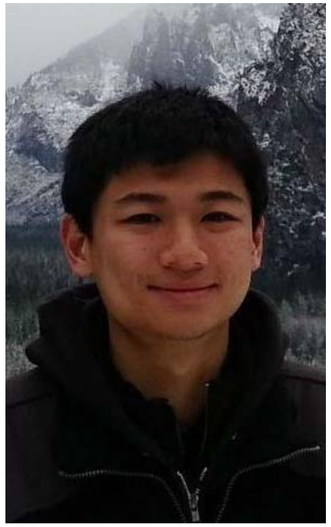
UC Berkeley

BAIR

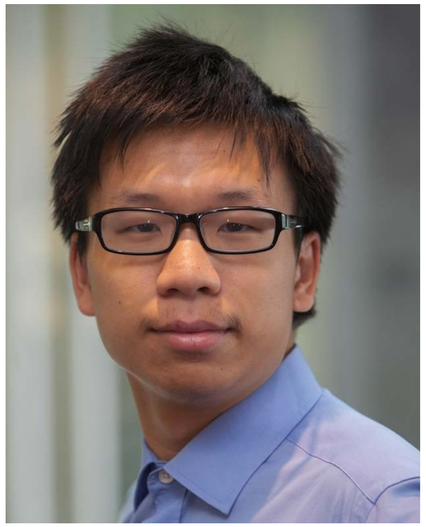
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

the brains behind the research

Dr. Phillip Isola
→ MIT



Richard Zhang

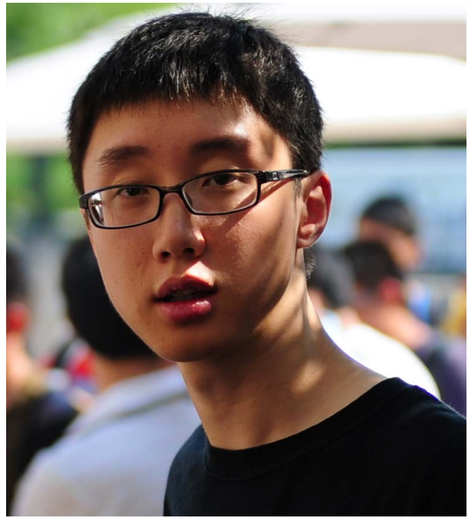


Jun-Yan Zhu
→ MIT

Taesung Park

Deepak Pathak

Pulkit Agarawal



Tinghui Zhou

Carl Doersch

Direct Supervised Learning

Convolutional Neural Network

“Collie”

image *X*

label *Y*

Direct Supervised Learning

Convolutional Neural Network

“Collie”

image *X*

label *Y*

Gatys et al, 2017

Are we fooling ourselves?

- E.g. action recognition
  - Very hard to improve on single frame classifiers
  - Consider “opening fridge” action:







Dataset bias is a problem, but so is our complacency

example by David Fouhey

direct semantic supervision considered harmful

- Dataset Bias will not go away
  - Data is finite
  - Machine learning people don’t care

- Need to make **better use** of the data we have
  - Direct supervision == memorization
  - Need to make the computer **study harder**

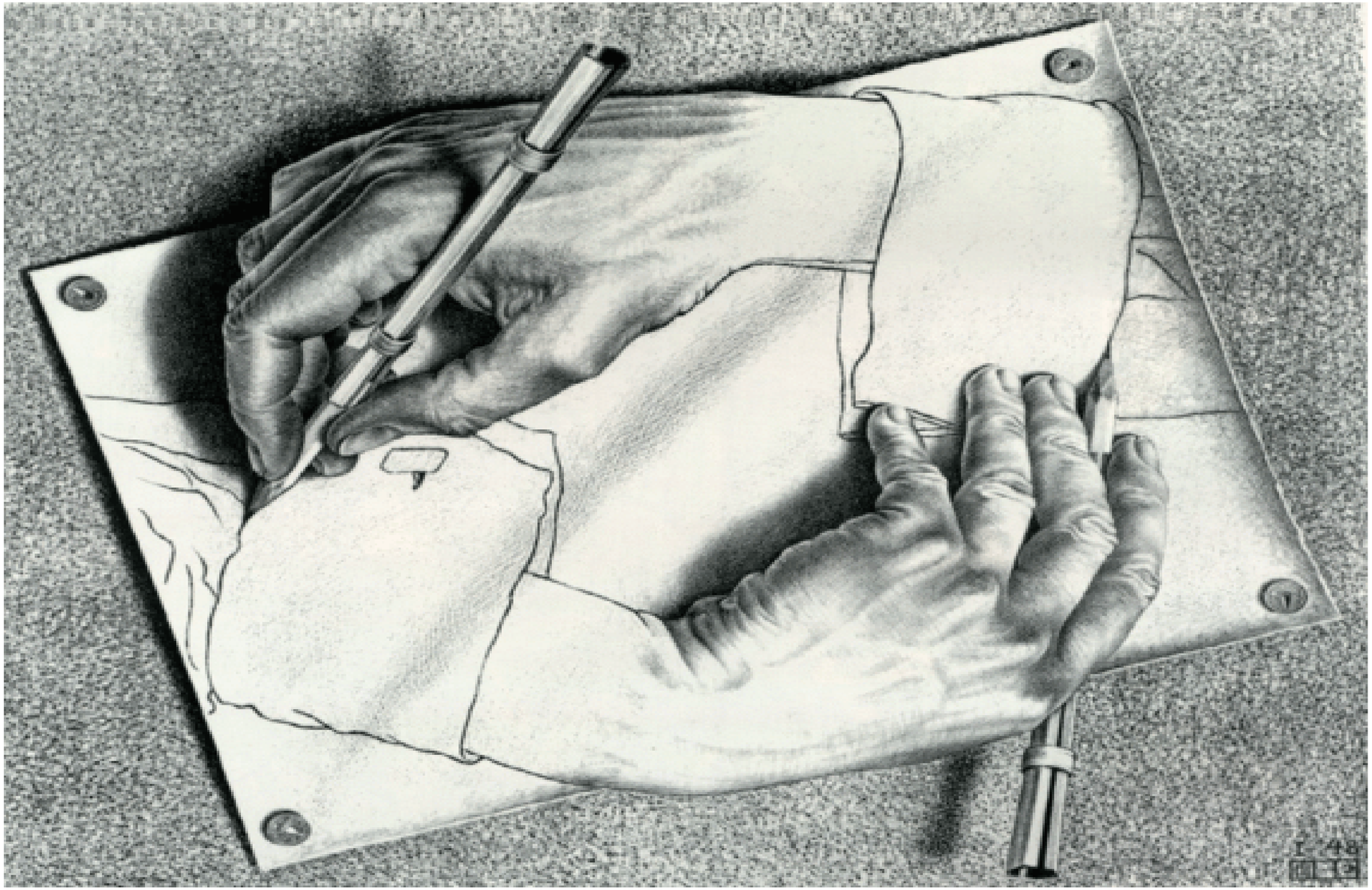
Making computer study harder

- Self-supervision
  – Data as its own supervision

- Meta-supervision
  – Supervise **constraints** on the data

- Curiosity / Intrinsic Motivation
  – Incentivize general learning

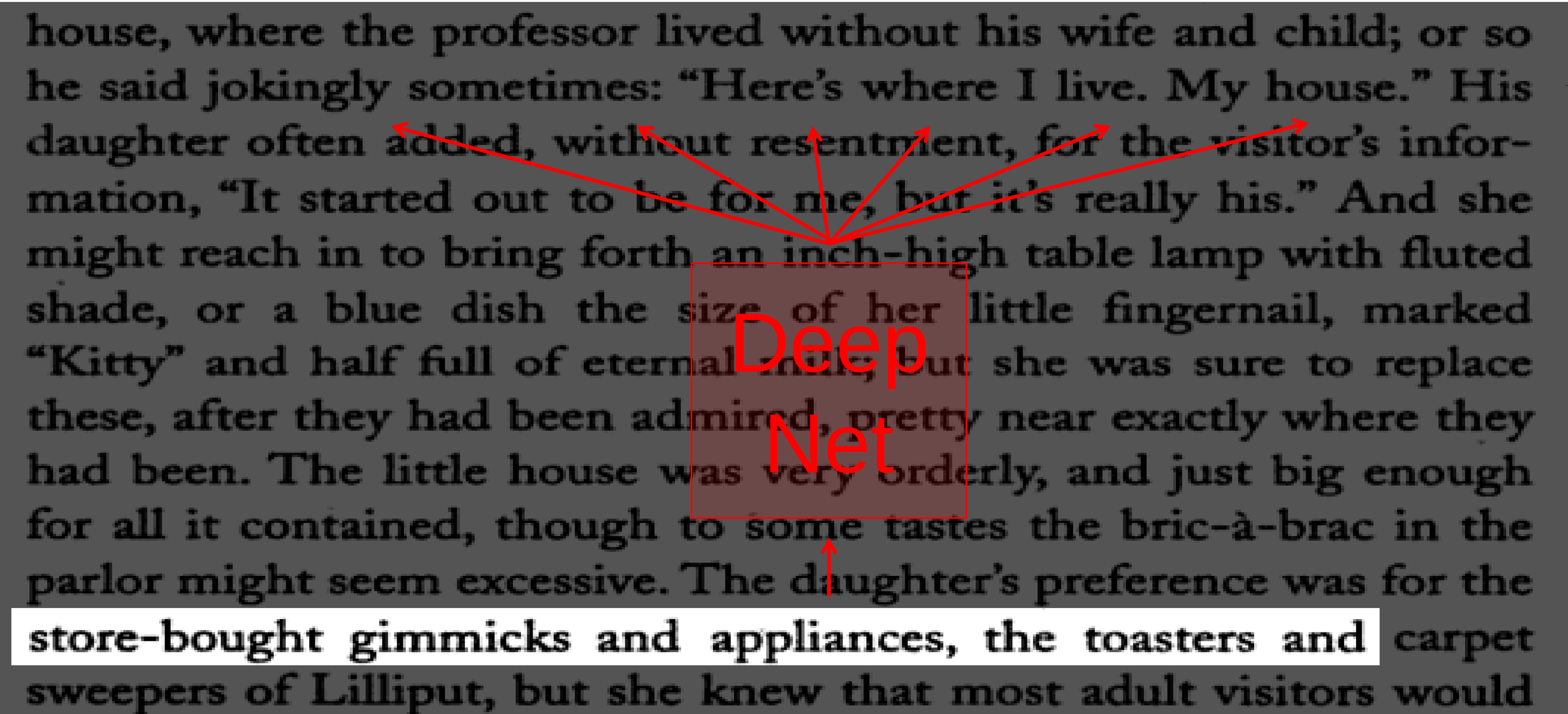
Self-Supervision: data as its own supervision



Drawing Hands, M.C. Escher, 1948

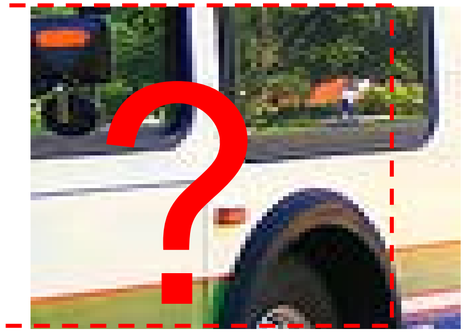
Context as Supervision

[Collobert & Weston 2008; Mikolov et al. 2013]



Context Prediction for Images

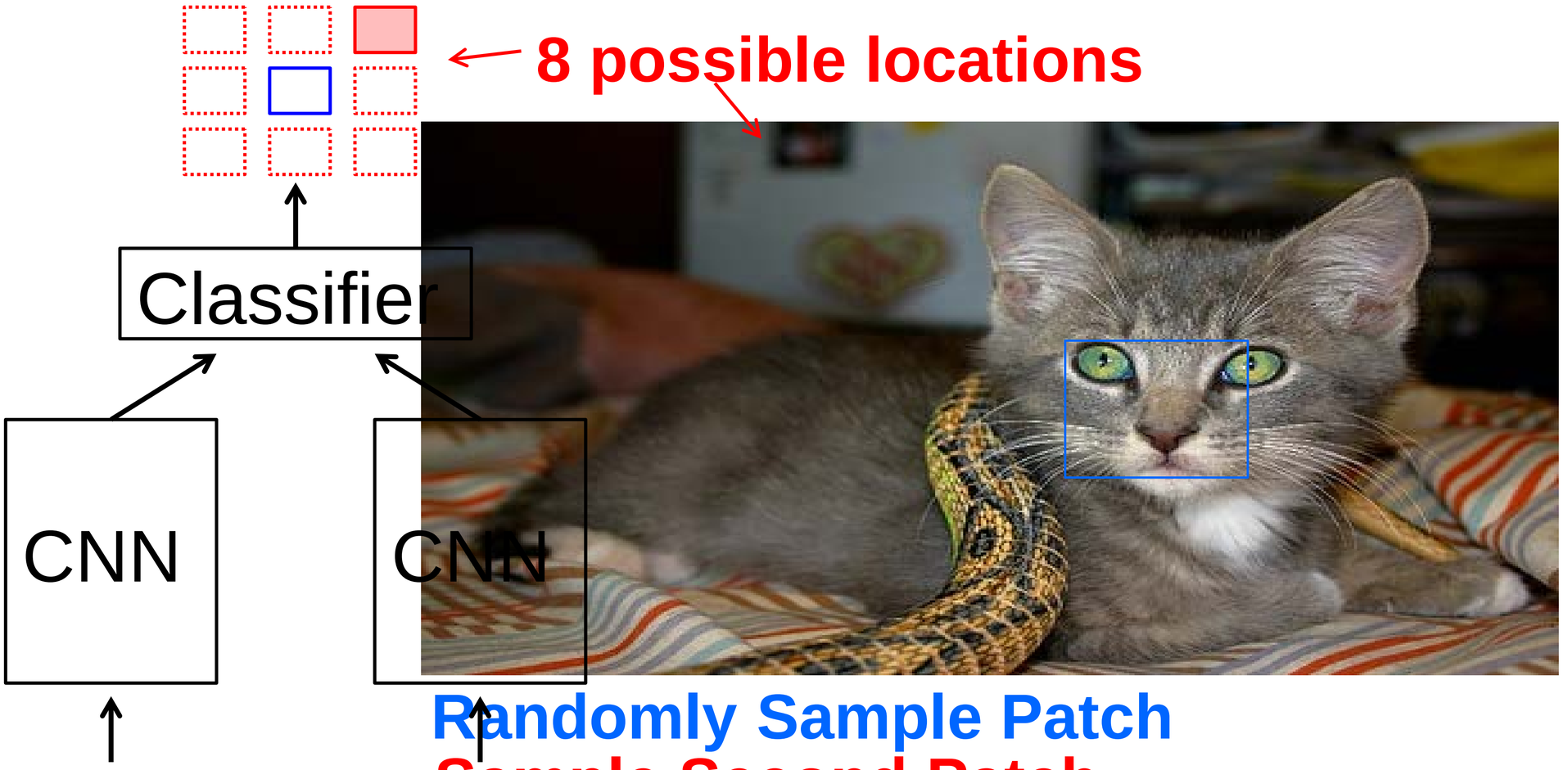
A

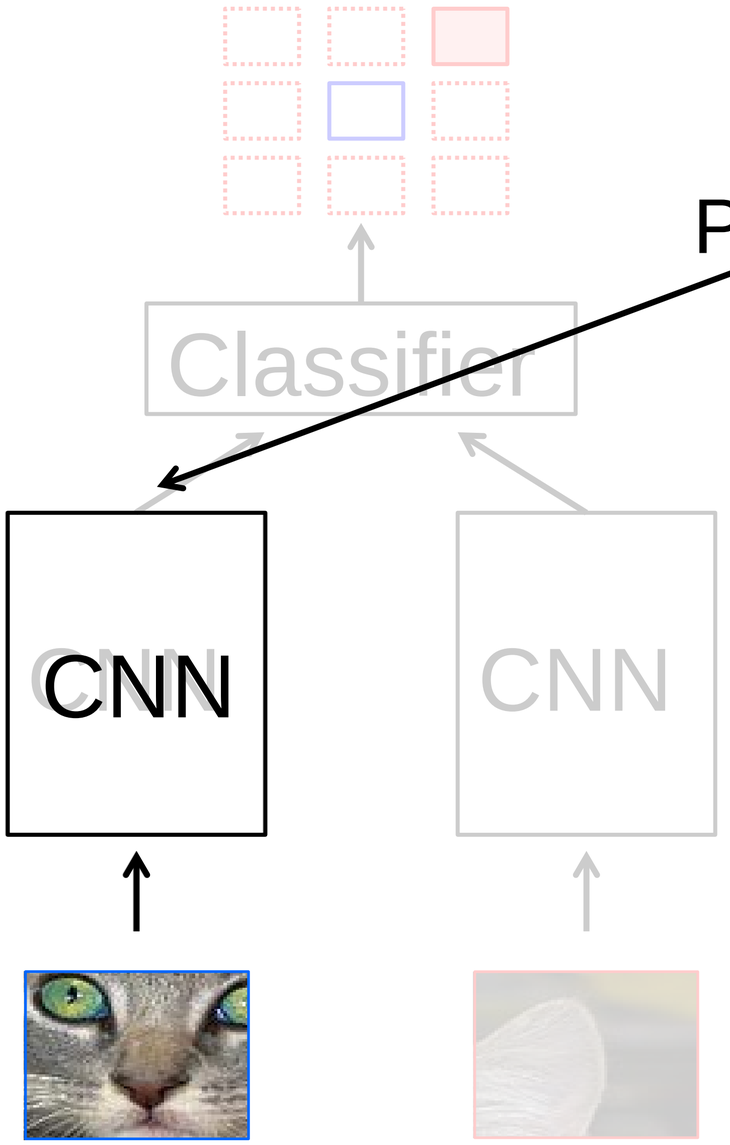


B

Semantics from a non-semantic task

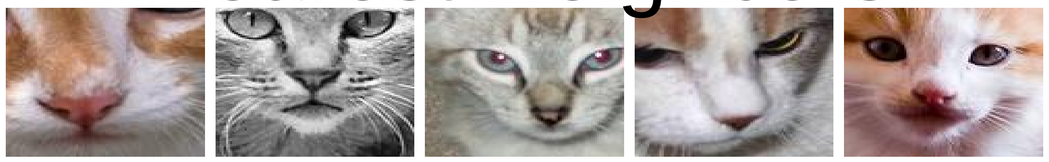
Relative Position Task





Input

Nearest Neighbors



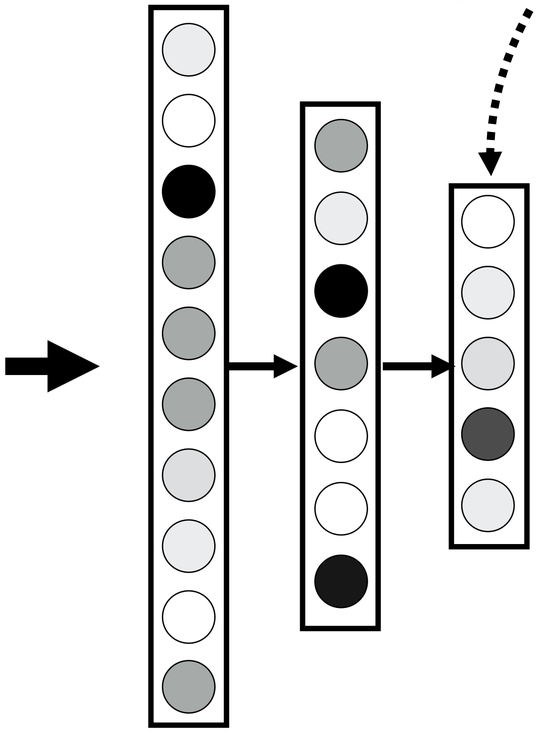
Correspondence ***across*** instances!

Representation Learning

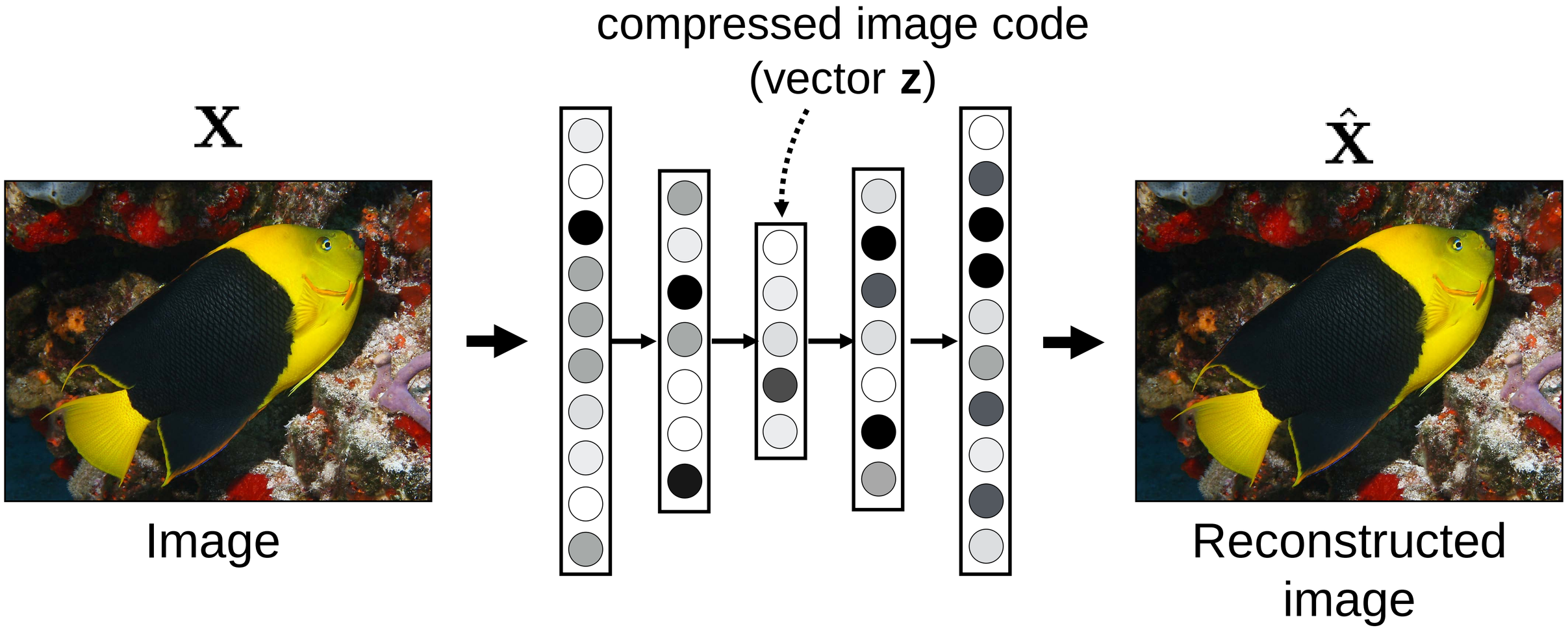
compressed image code
(vector **z**)

x

Image



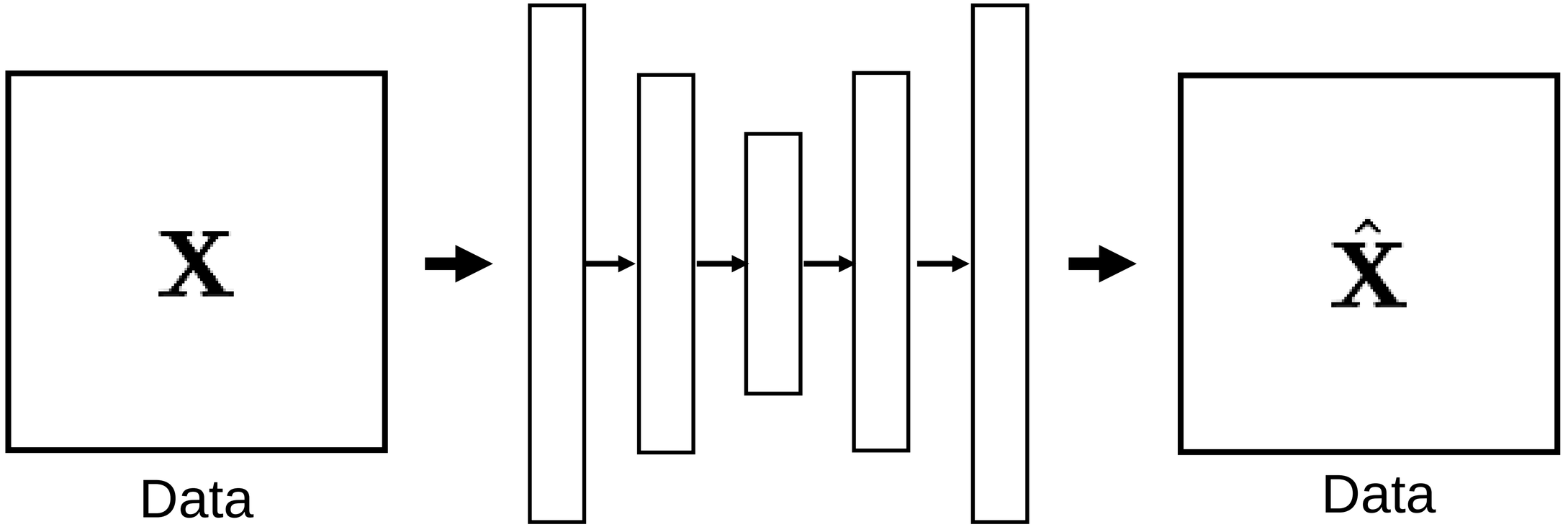
Representation Learning



"Autoencoder"

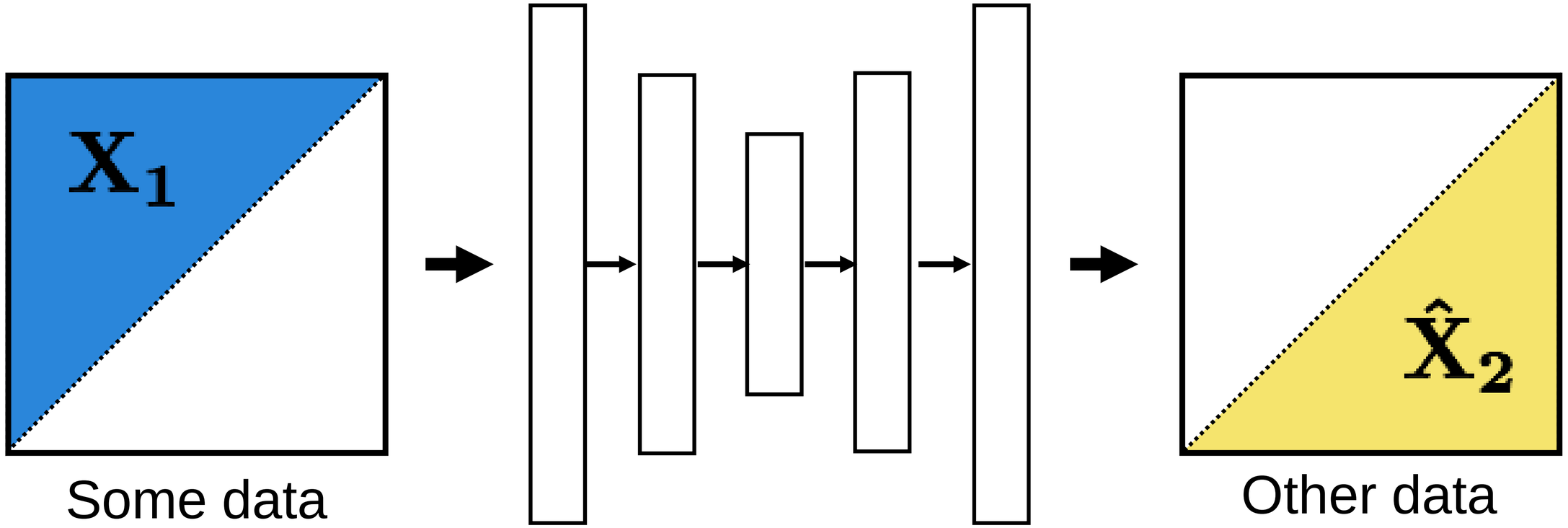
[e.g., Hinton & Salakhutdinov, Science 2006]

Data compression

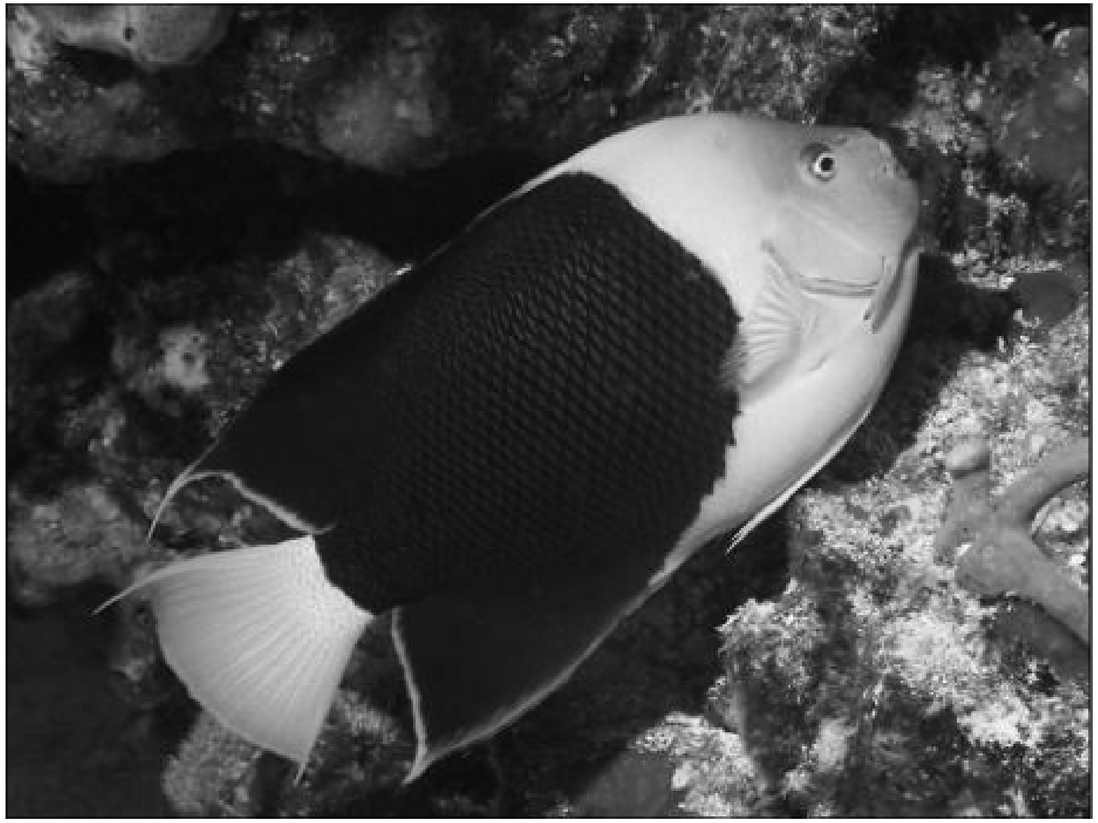


[Hinton & Salakhutdinov, Science 2009]

Data *prediction*



see also [Vincent et al., 2008]



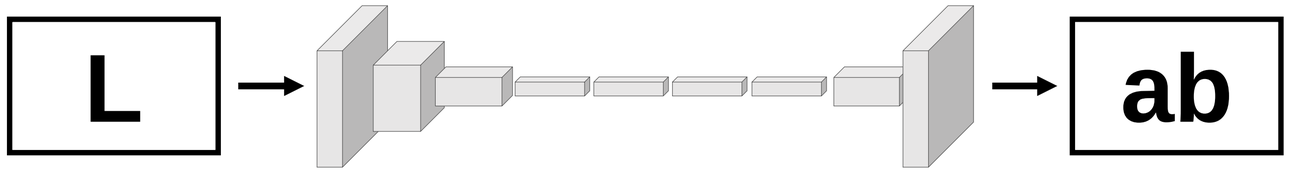
$\mathcal{F}$

Grayscale image: L channel

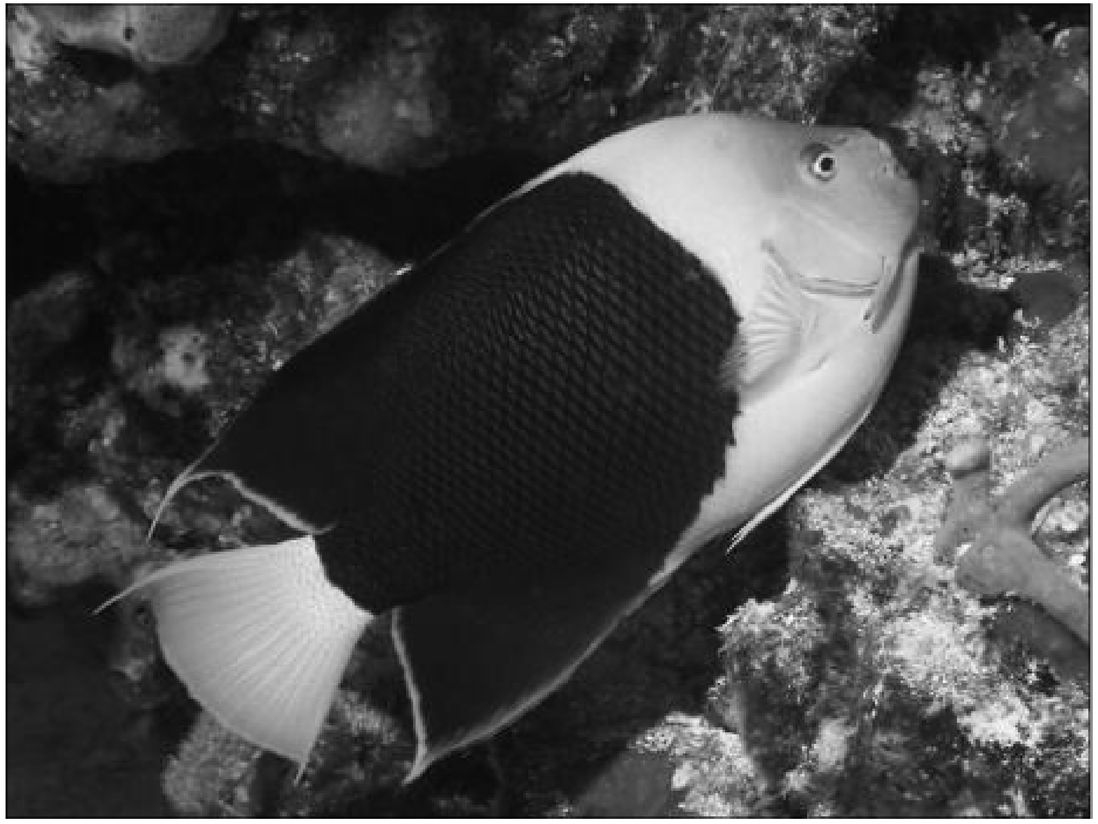
$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

Color information: ab channels

$\hat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$



[Zhang, Isola, Efros, ECCV 2016]



$\mathcal{F}$

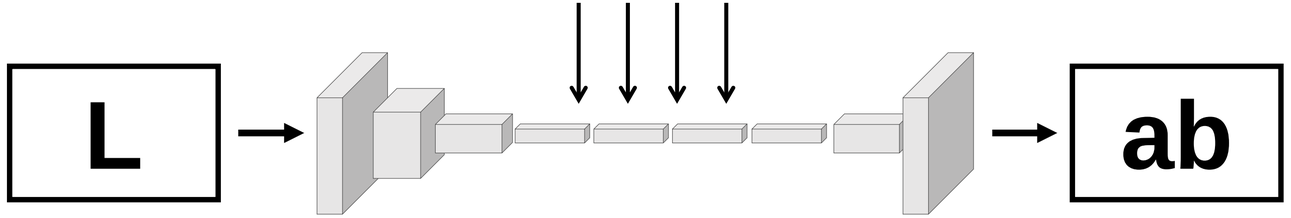
Grayscale image: L cha

$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

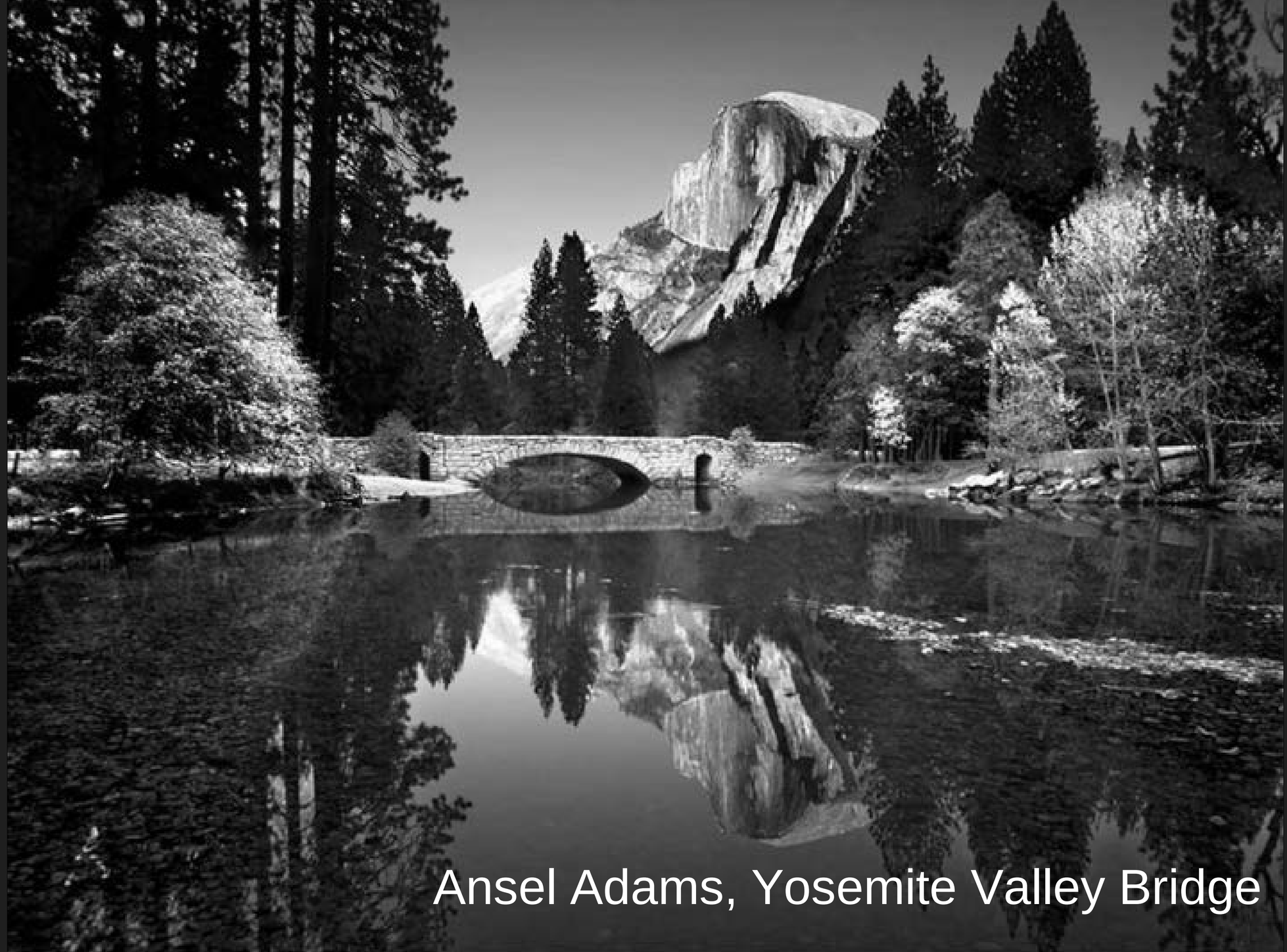
Semantics? Higher-level abstraction?

information: ab channels

$\hat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$

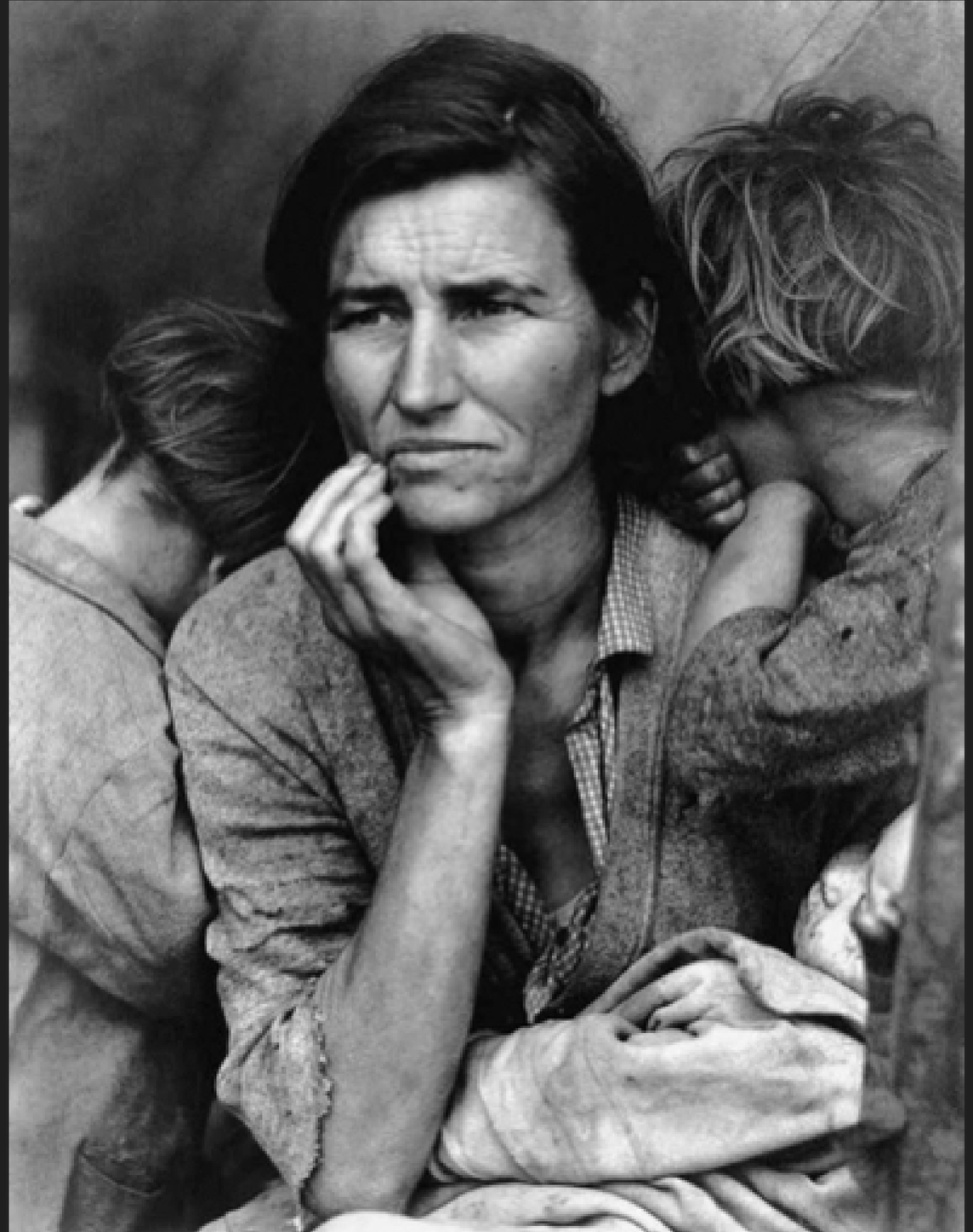


[Zhang, Isola, Efros, ECCV 2016]

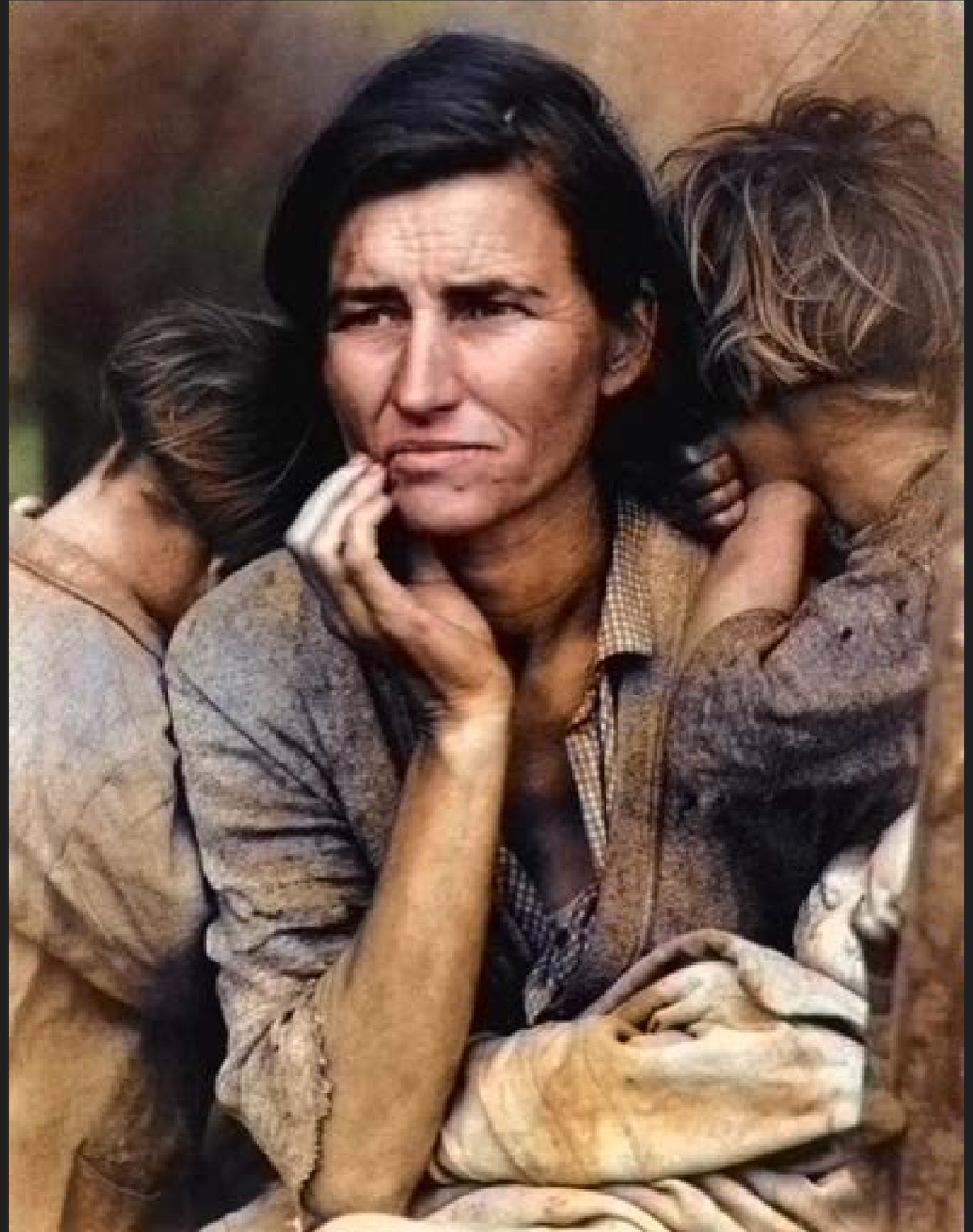


Ansel Adams, Yosemite Valley Bridge


Our result



Migrant Mother
Dorothea Lange
1936

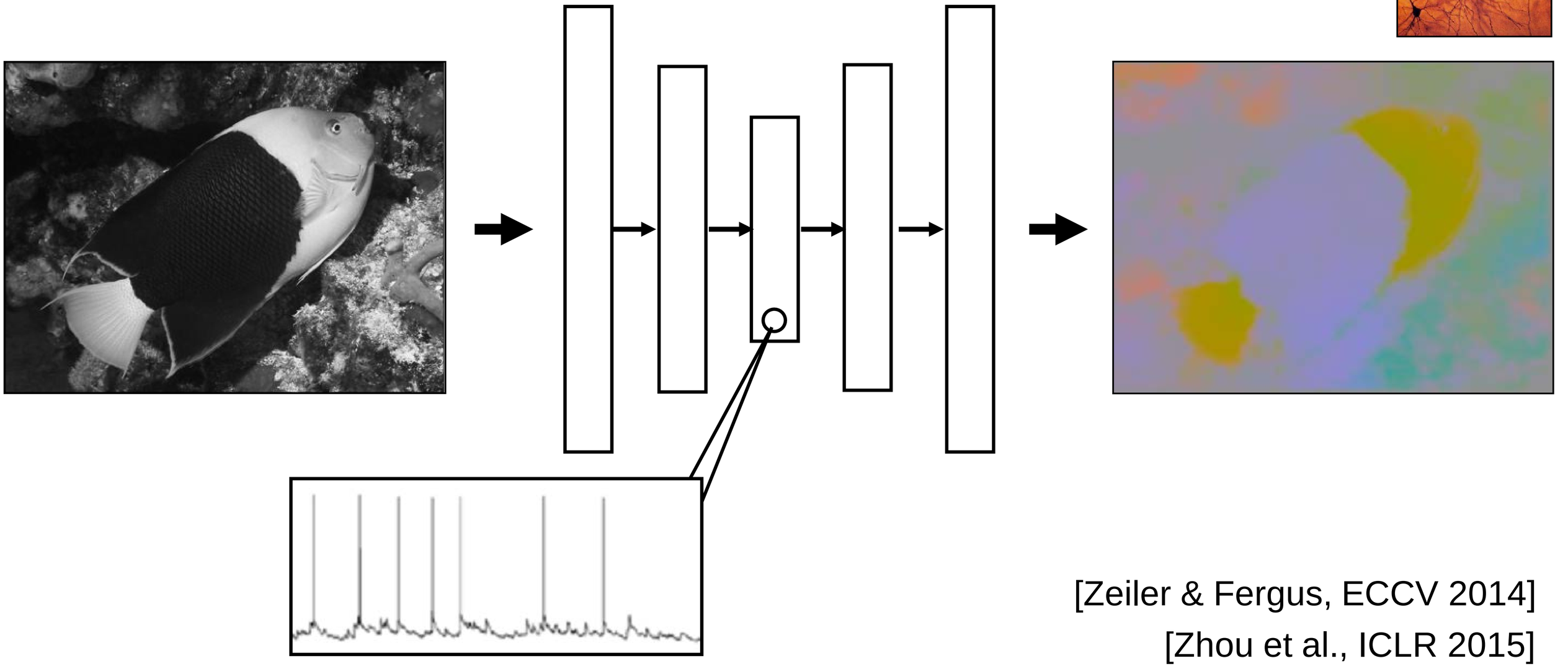


Our result

Instructive failure

Instructive failure

Deep Net “Electrophysiology”



[Zeiler & Fergus, ECCV 2014]

[Zhou et al., ICLR 2015]

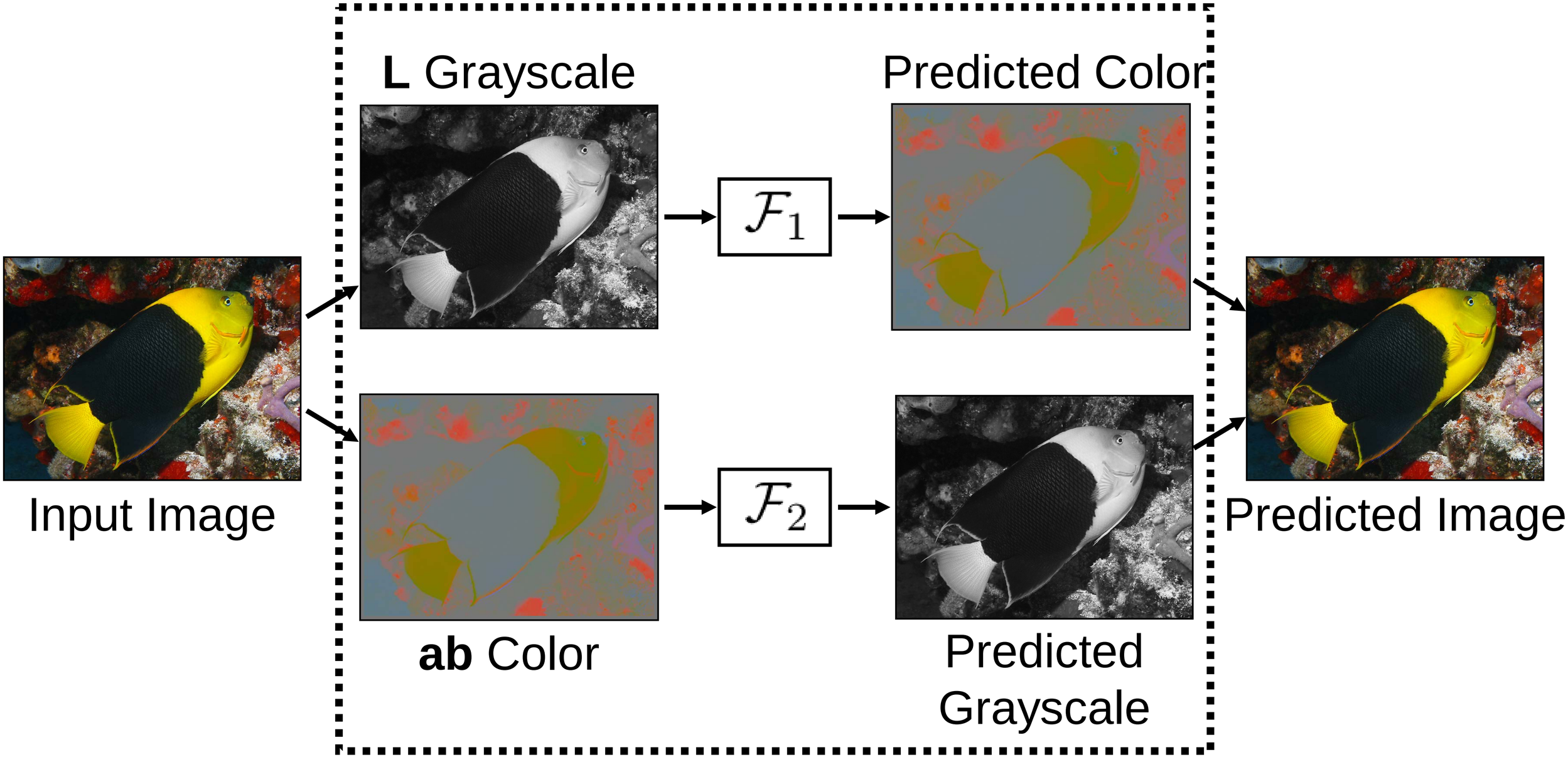
Stimuli that drive selected neurons (conv5 layer)

faces

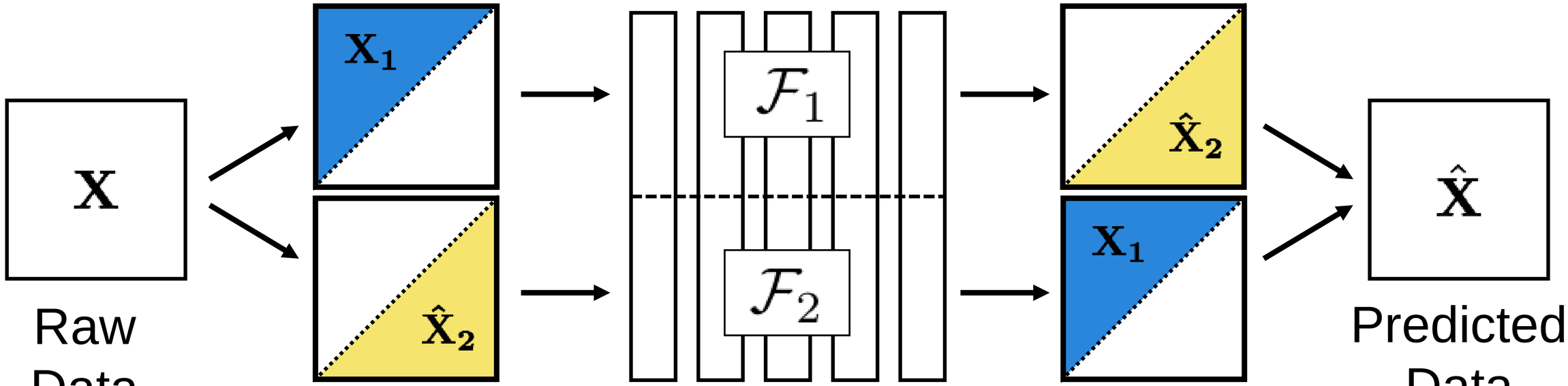
dog faces

flowers

“Autoencoder”

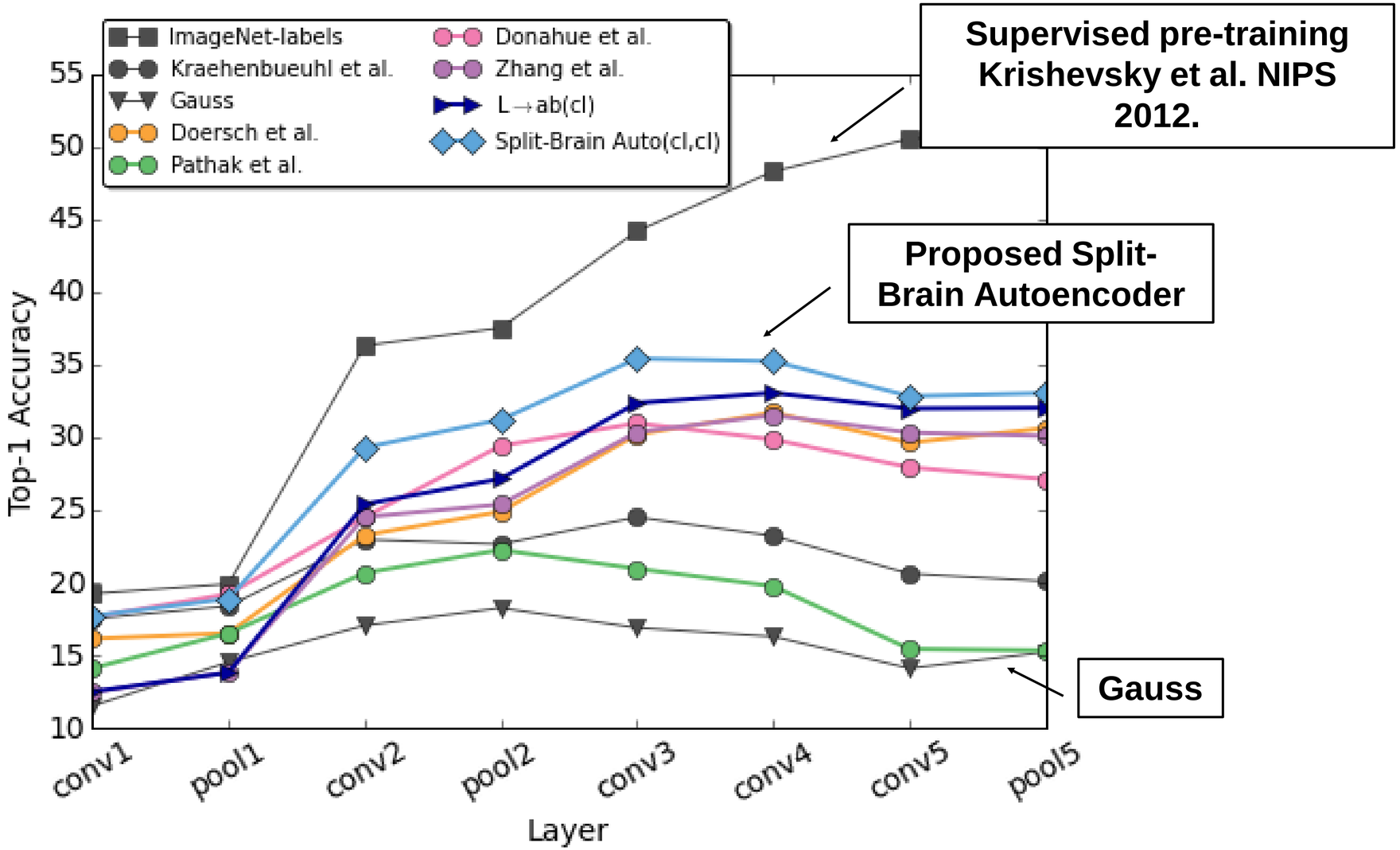


“Split-Brain Autoencoder” (CVPR’17)



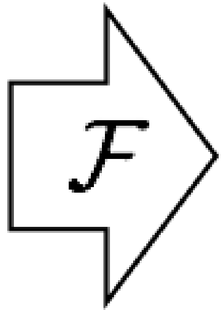
[Zhang, Isola, Efros, CVPR 2017]

ImageNet Classification Performance per Layer



How to evaluate predictions?

Input



Output

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

Objective function (loss)

Neural Network

How to evaluate predictions?

Input

$\mathcal{F}$

Output

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

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

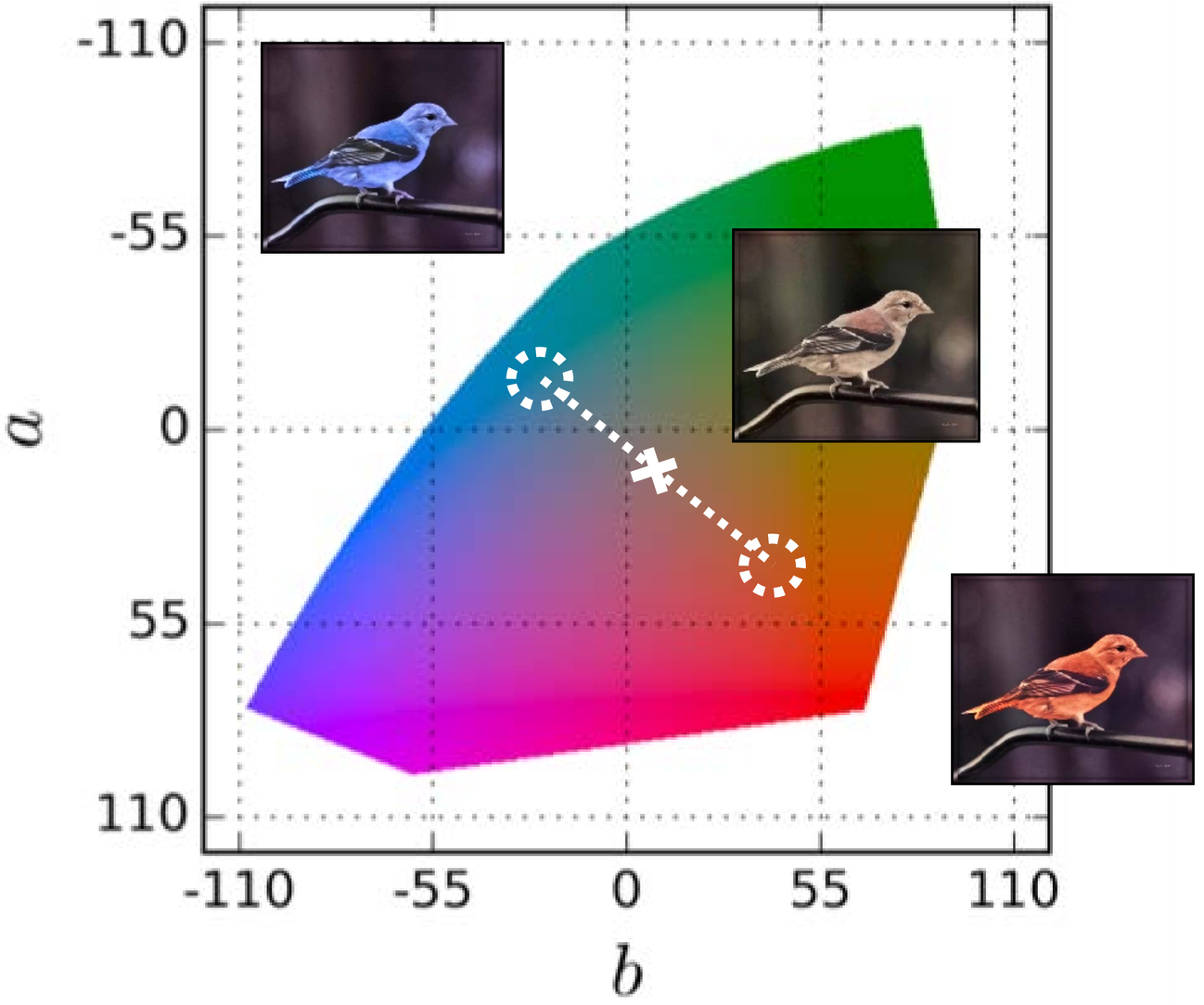
Designing objective functions

Input

Output

Ground truth

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



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

Designing objective functions

Input

Zhang et al. 2016

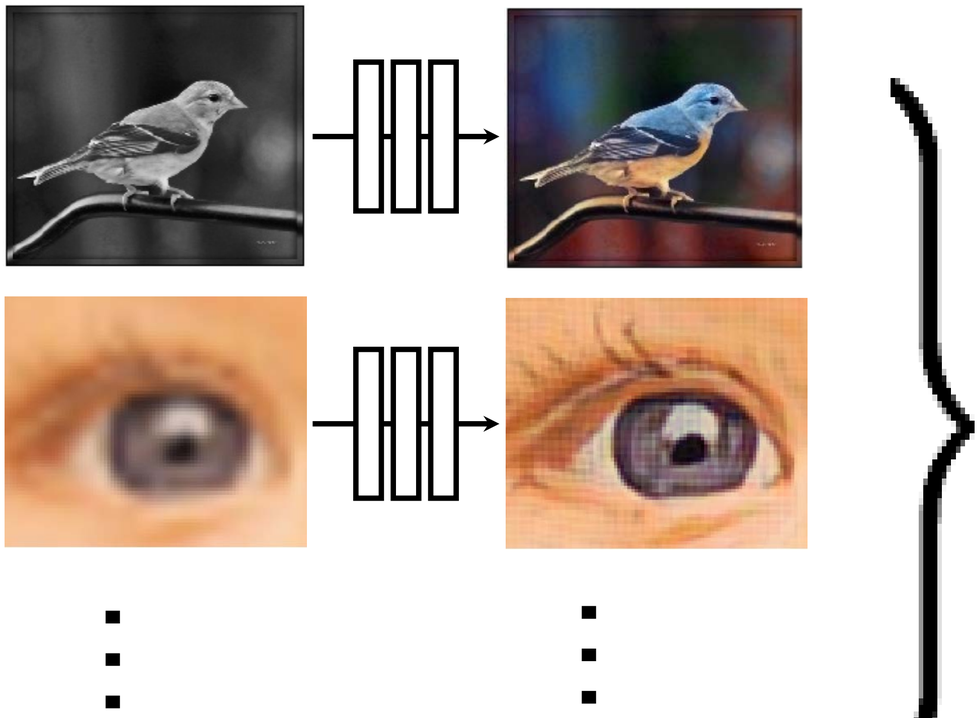
Ground truth

Color distribution cross-entropy loss with colorfulness enhancing term.

[Zhang, Isola, Efros, ECCV 2016]

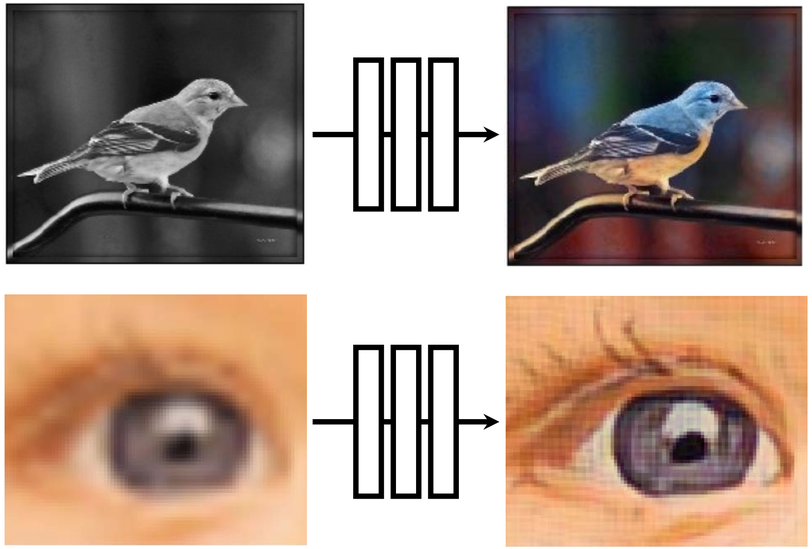
Designing objective functions

Be careful what you wish for!



Universal loss?

Generated images



Real photos

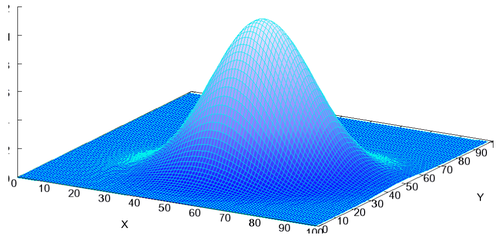
…

“Generative Adversarial Network” (GANs)

Generated vs Real (classifier)

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

GANs



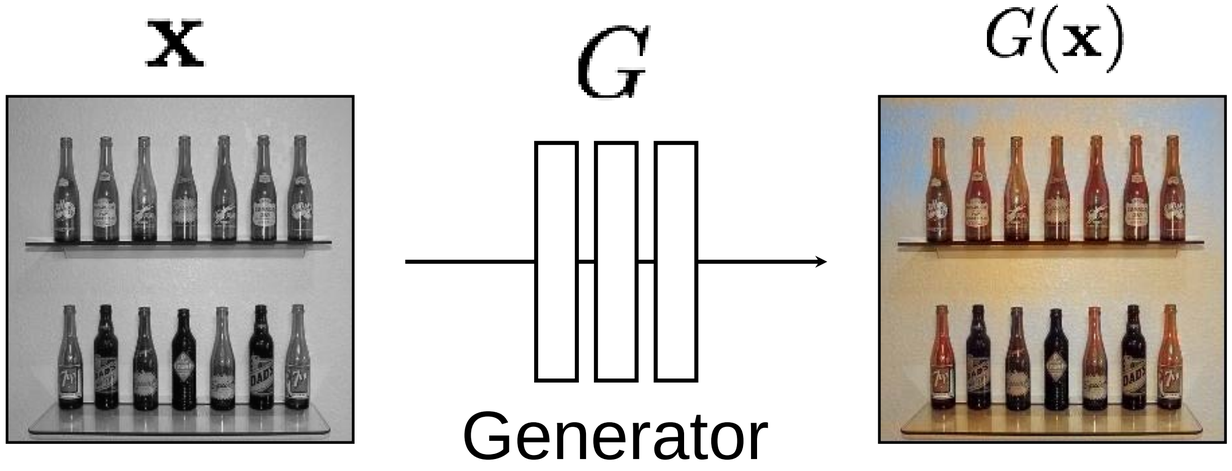
Z

[Goodfellow et al., 2014]

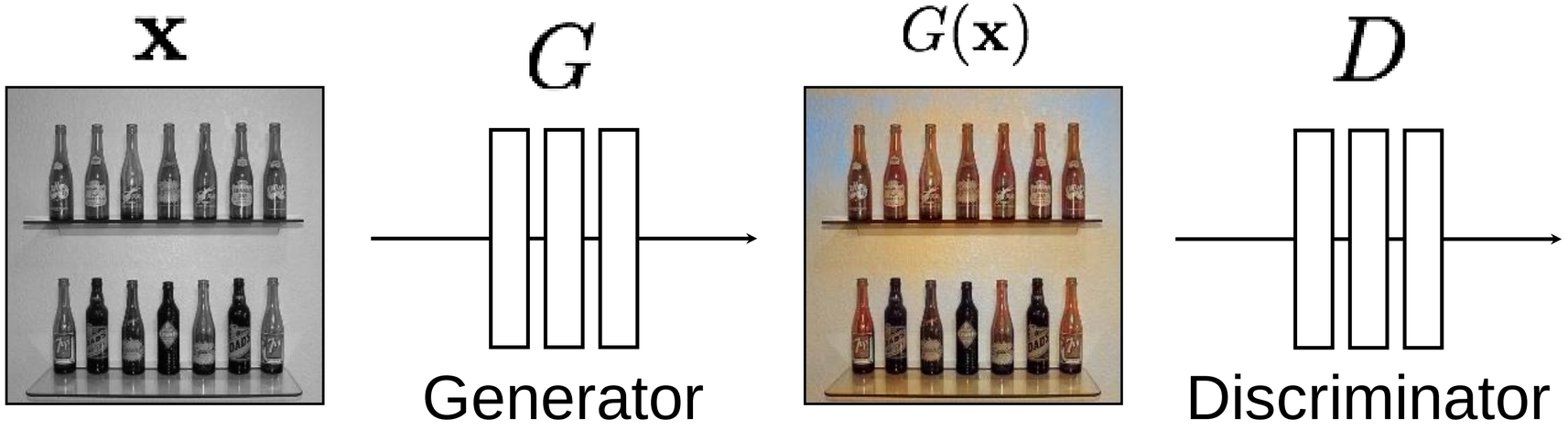
Conditional GANs

[Goodfellow et al., 2014]

[Isola et al., 2017]



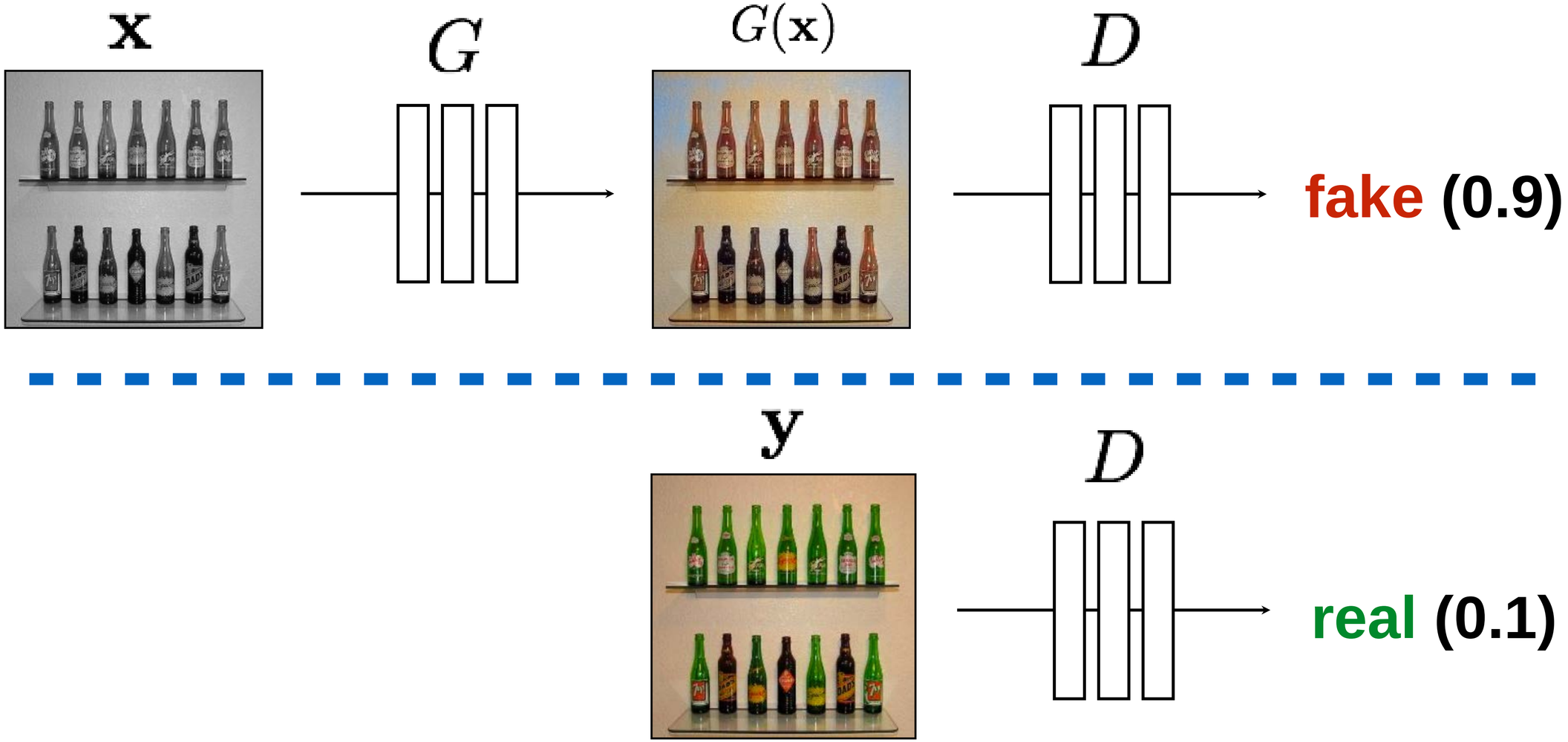
[Goodfellow et al., 2014]



G tries to synthesize fake images that fool D

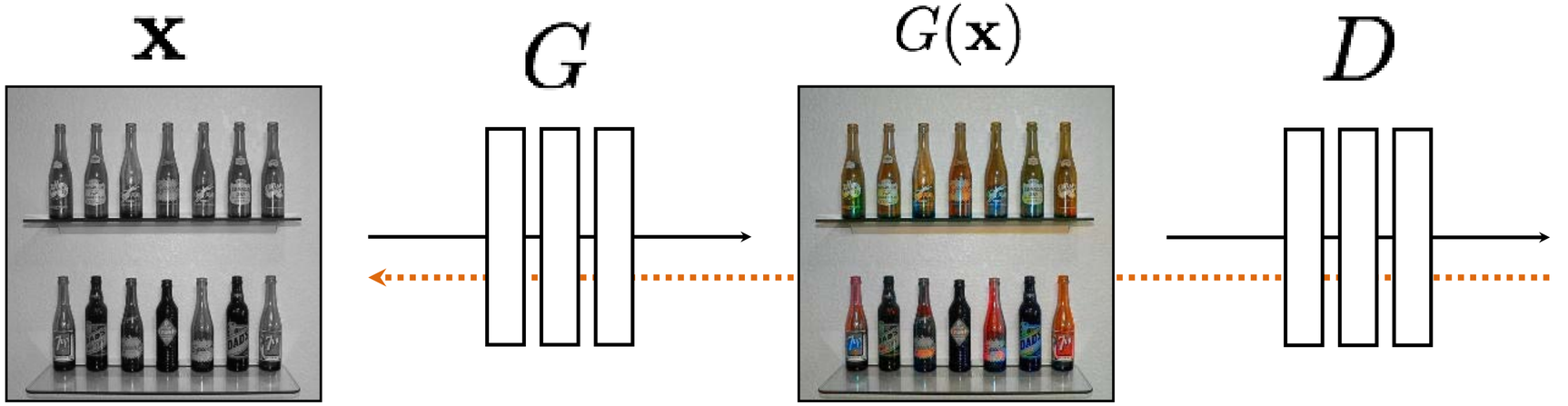
D tries to identify the fakes

[Goodfellow et al., 2014]



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

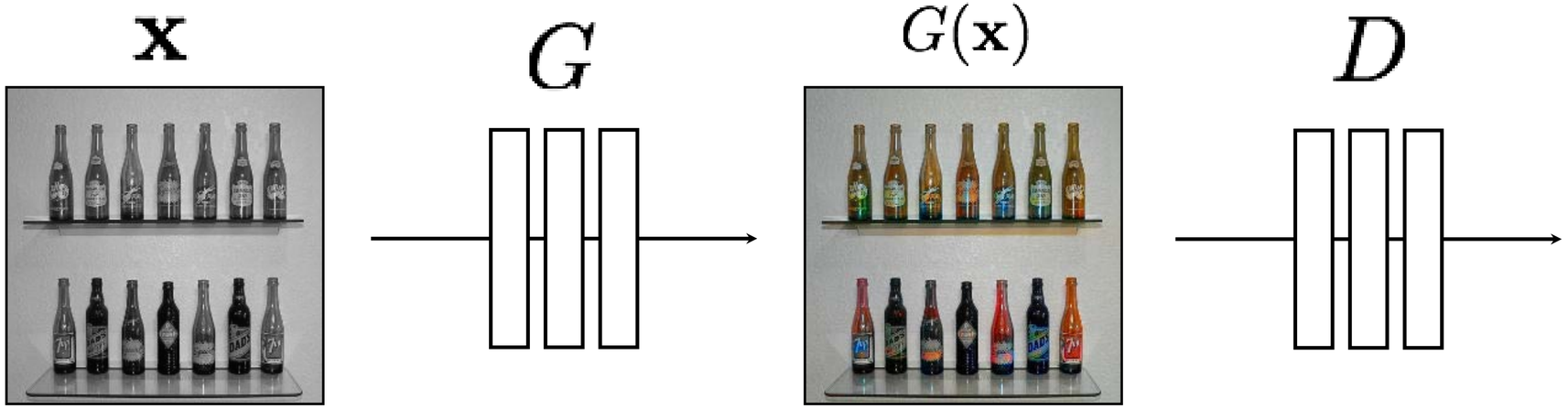
[Goodfellow et al., 2014]



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

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

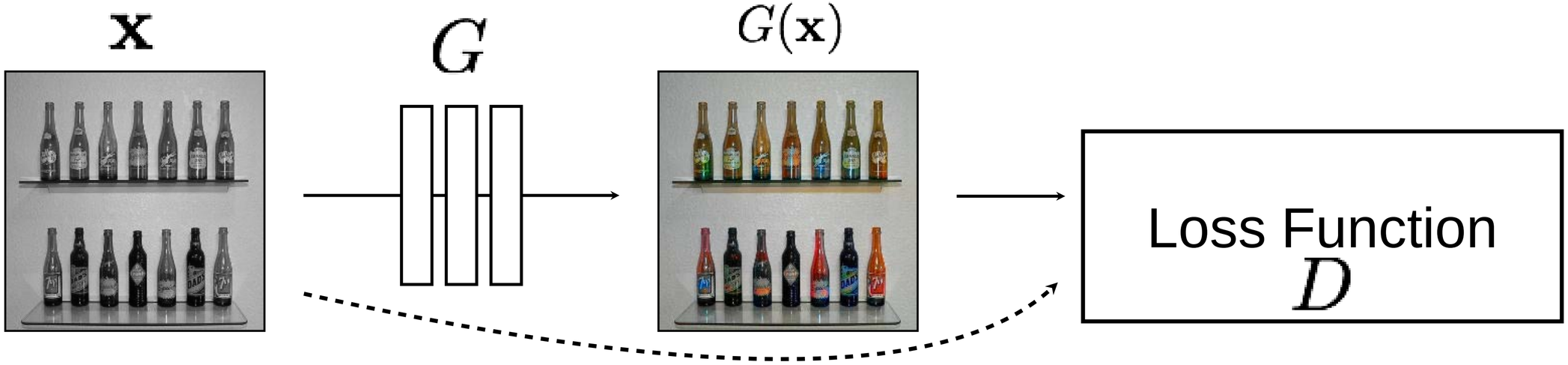
[Goodfellow et al., 2014]



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

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

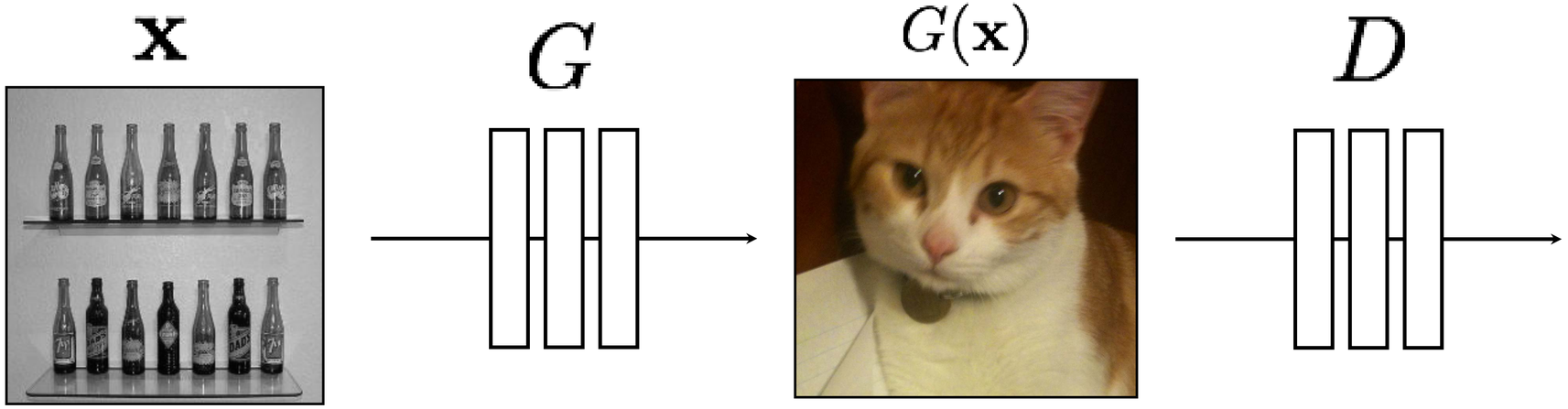
[Goodfellow et al., 2014]



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

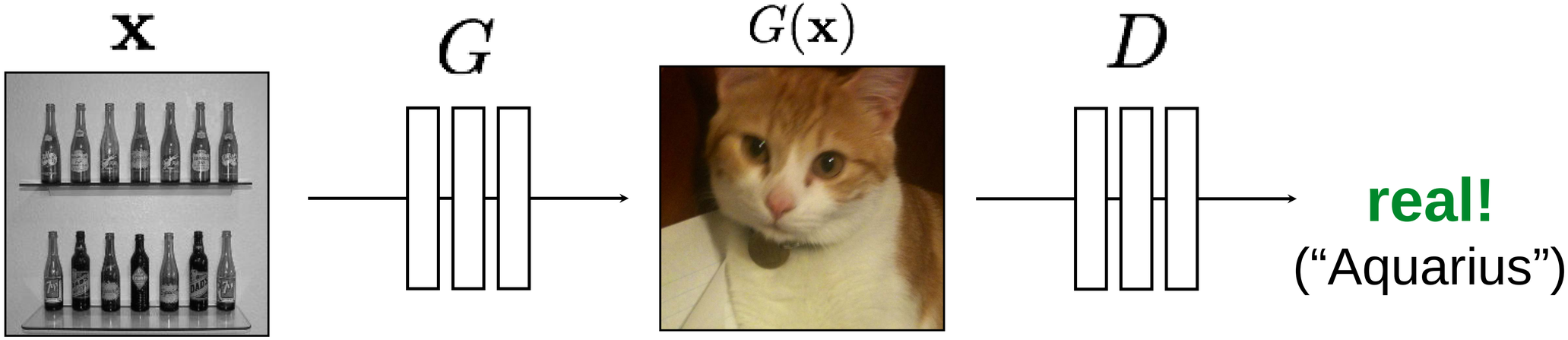
Rather than being hand-designed, it is *learned*.

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



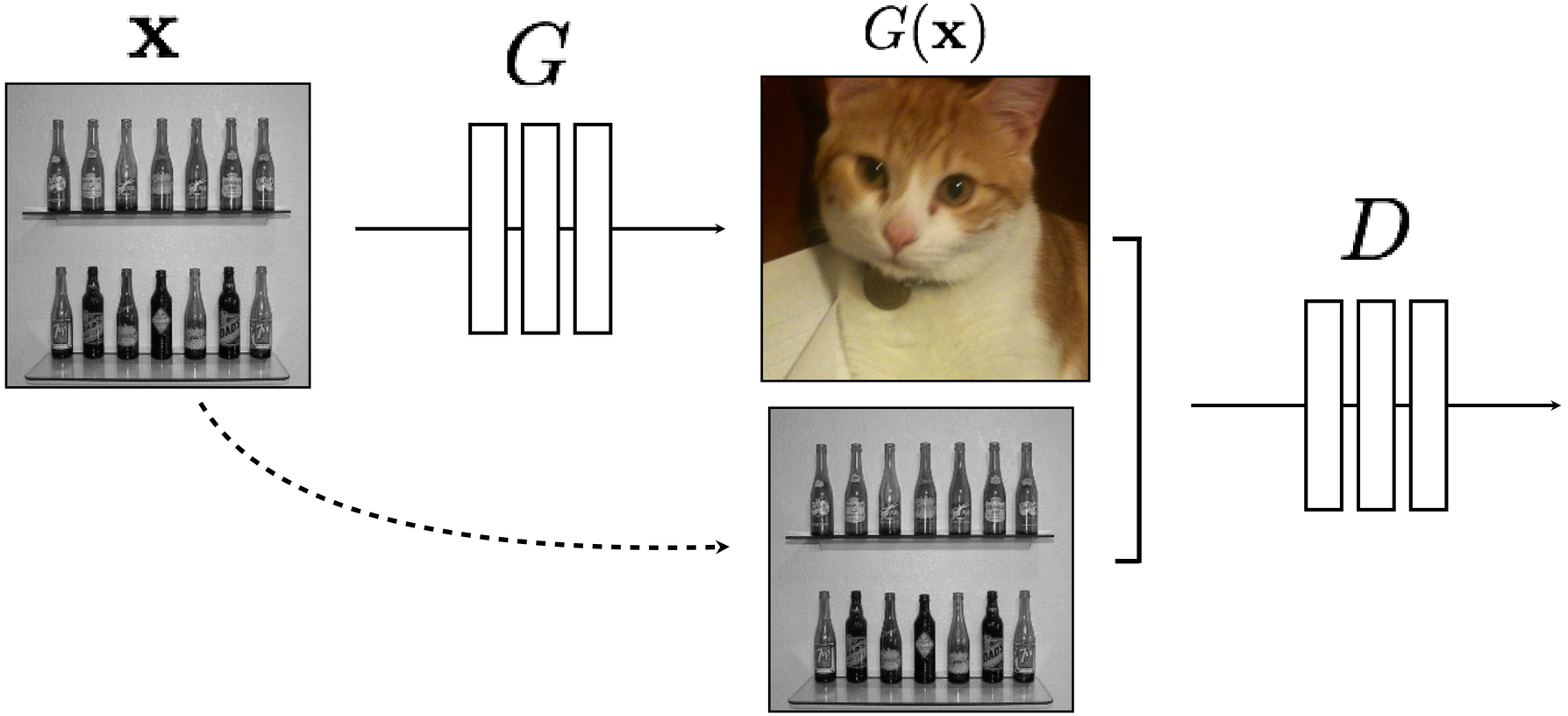
$$\arg\min_G \max_D \; \mathbb{E}_{\mathbf{x},\mathbf{y}}[\; \log D(G(\mathbf{x})) \quad + \quad \log(1 - D(\mathbf{y})) \;\;]$$

[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]



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

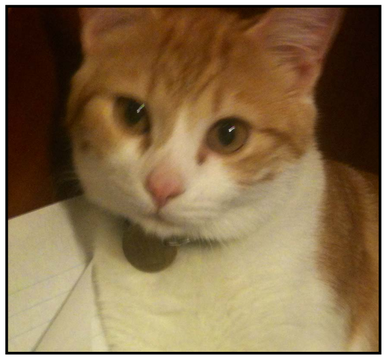
[Goodfellow et al., 2014]

[Isola et al., 2017]

$\mathbf{x}$

G

$G(\mathbf{x})$



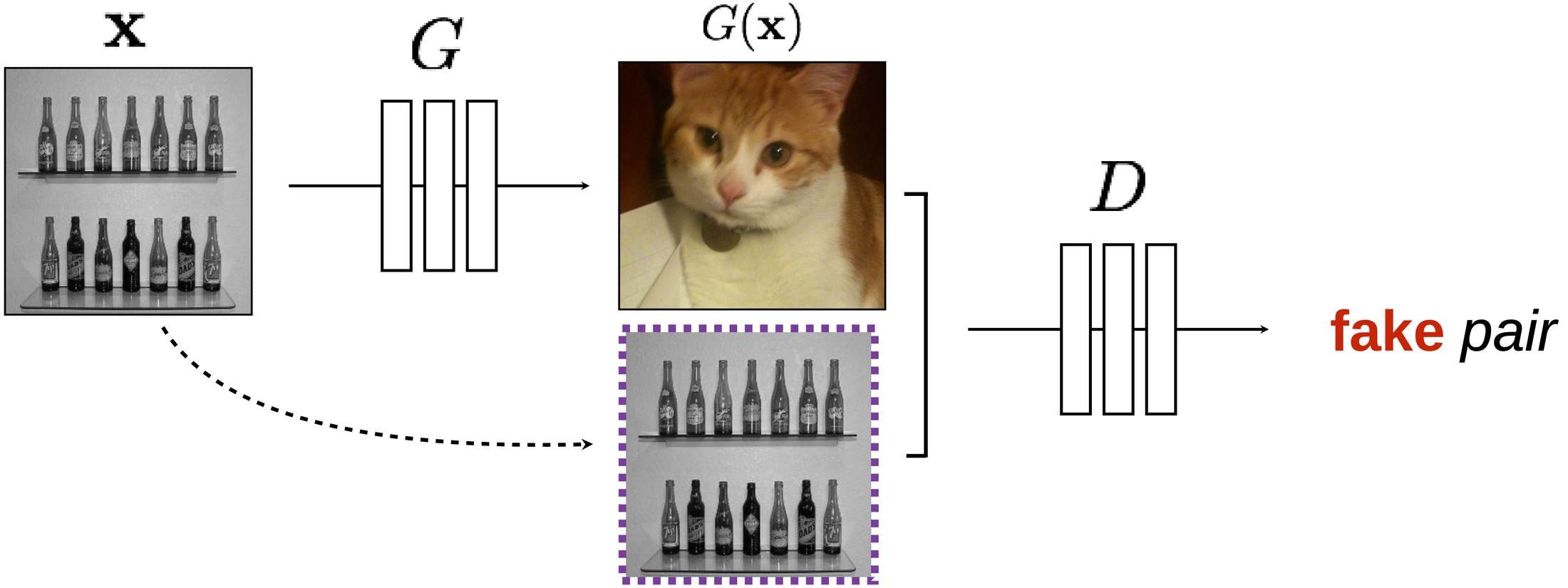
D

real or fake *pair* ?

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

[Goodfellow et al., 2014]

[Isola et al., 2017]



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

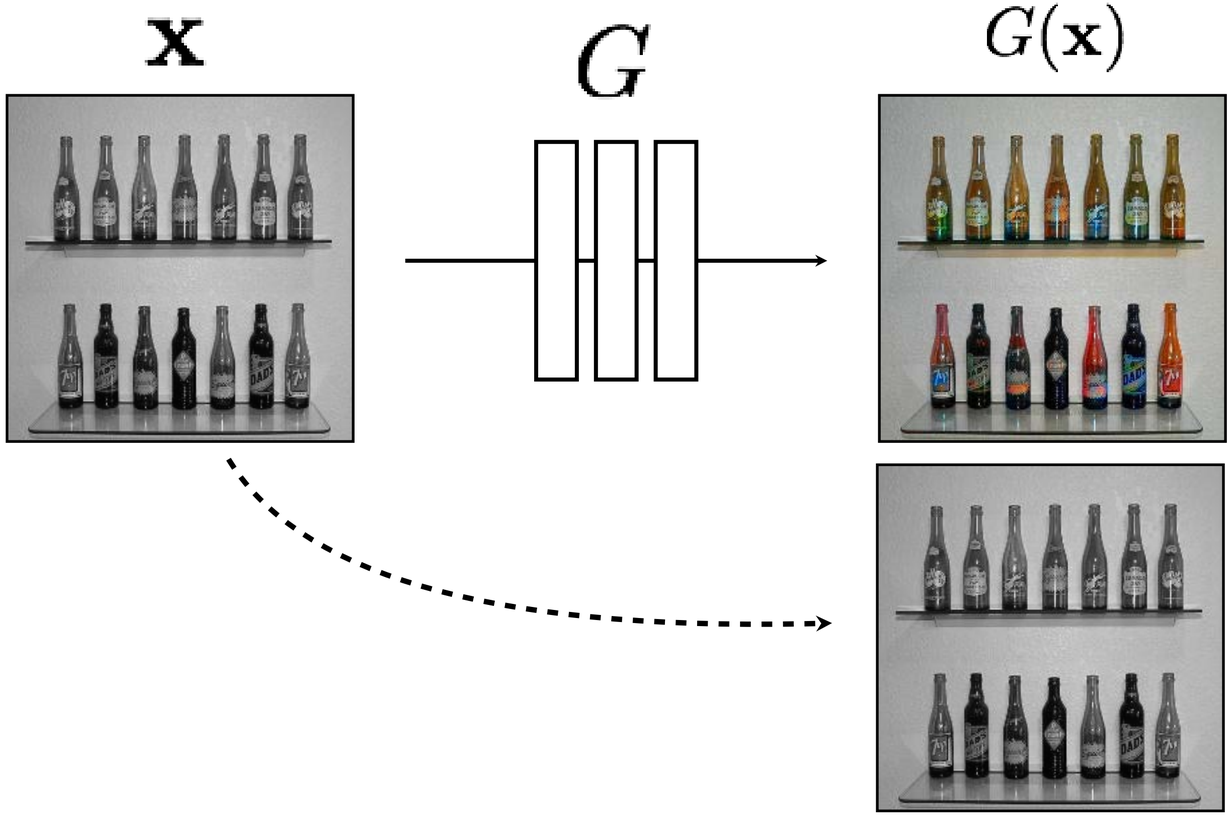
[Goodfellow et al., 2014]

[Isola et al., 2017]

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

[Goodfellow et al., 2014]

[Isola et al., 2017]



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

[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]

Input

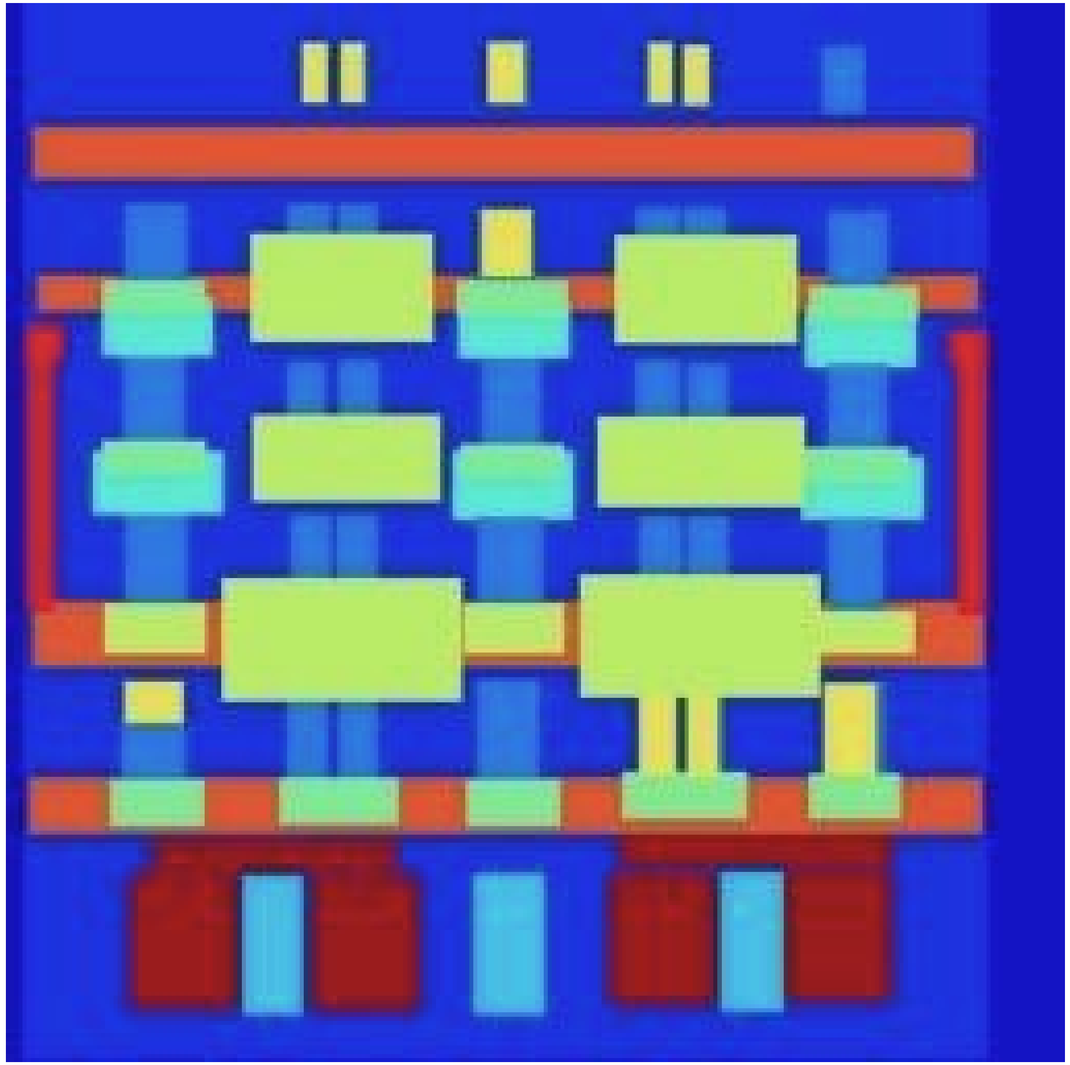
Output

Groundtruth

Data from [maps.google.com]

Labels → Facades

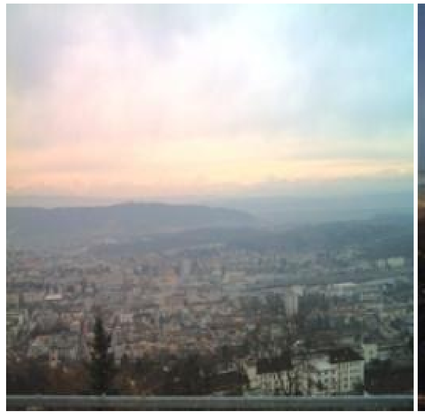
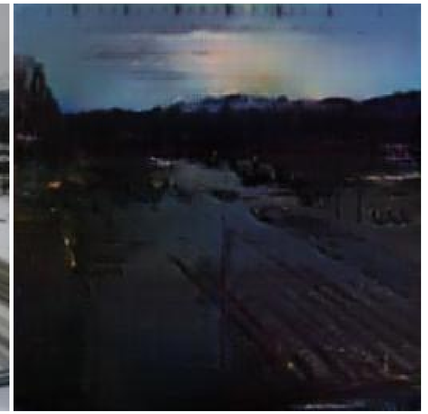
Input



Output

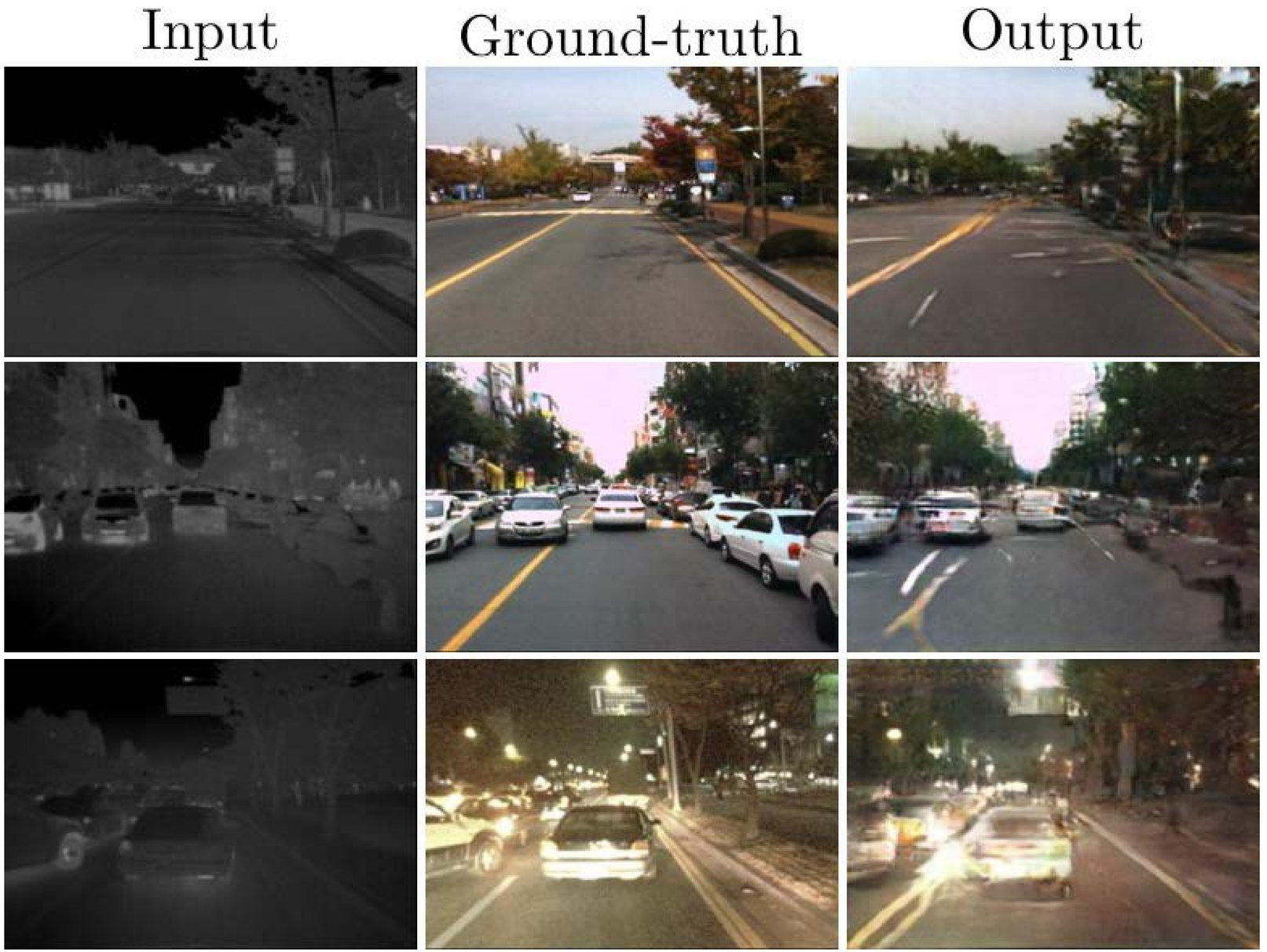
Data from [Tylecek, 2013]

Day → Night

Input	Output	Input	Output	Input	Output
					
					

Data from [Laffont et al., 2014]

Thermal → RGB



Edges → Images

Input Output Input Output Input Output

Edges from [Xie & Tu, 2015]

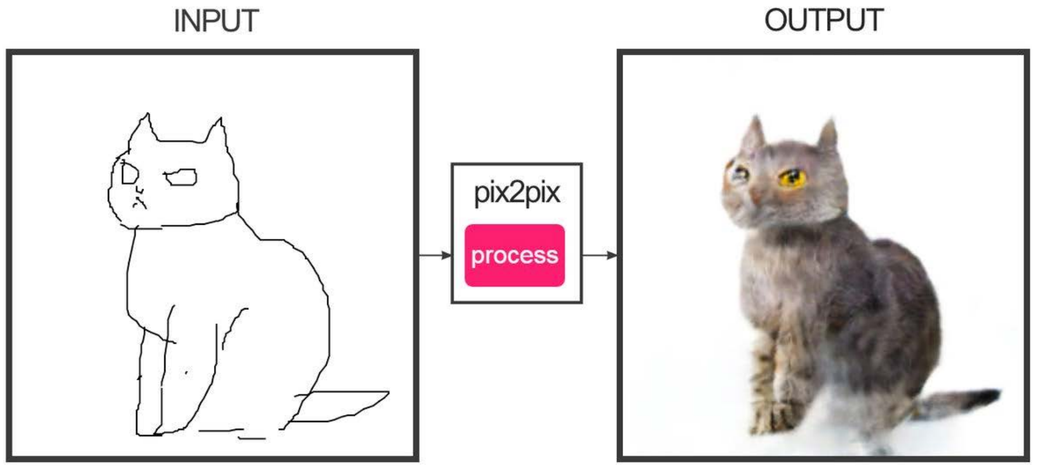
Sketches → Images

Input | Output | Input | Output | Input | Output

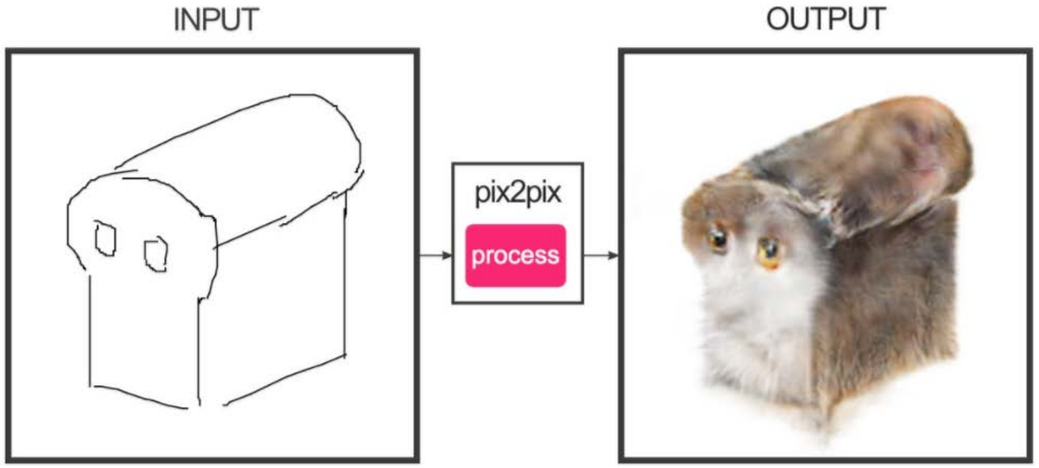
Trained on Edges → Images

Data from [Eitz, Hays, Alexa, 2012]

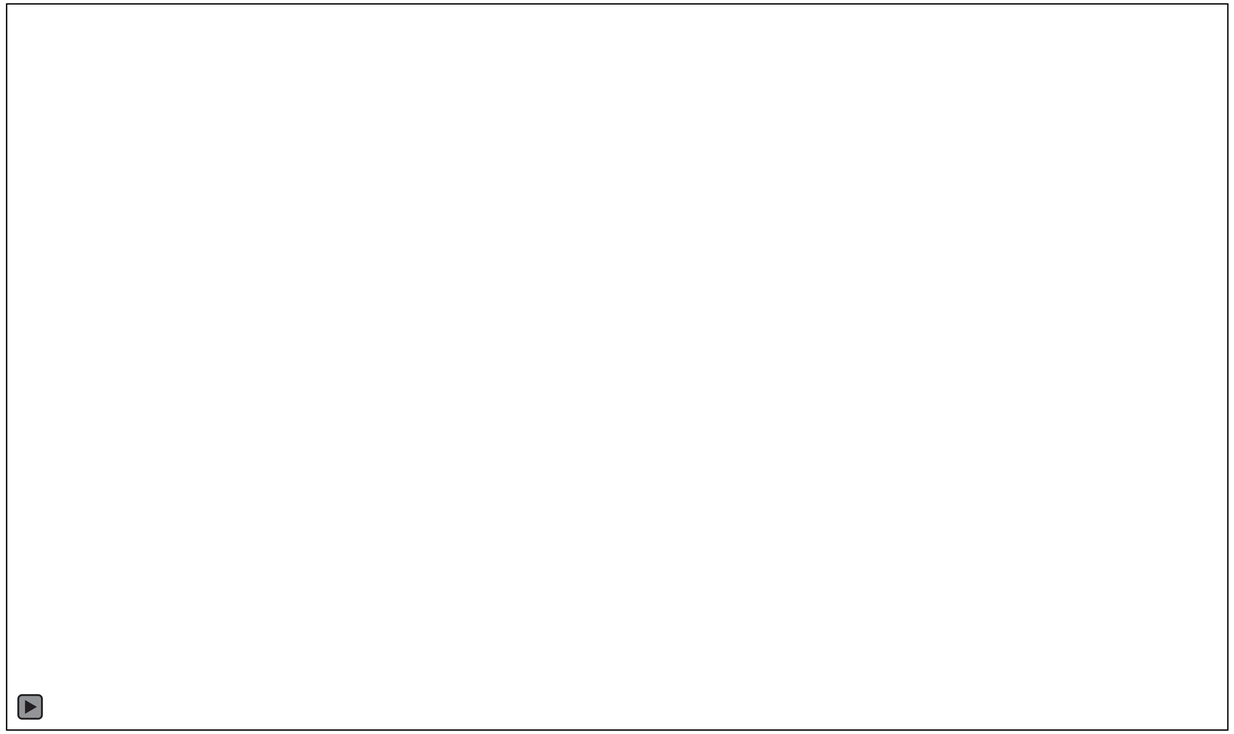
#edges2cats [Christopher Hesse]



@gods_tail



Ivy Tasi @ivymyt



@matthematician

Meta-supervision

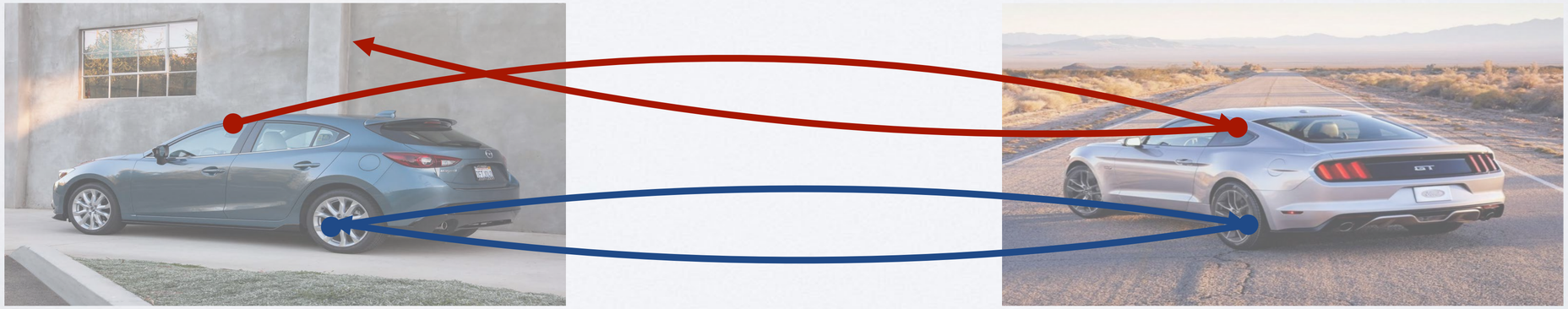
Don’t tell me what to do, tell me how to be well-behaved

Cycle-consistency

- Composite flows along a cycle should be zero

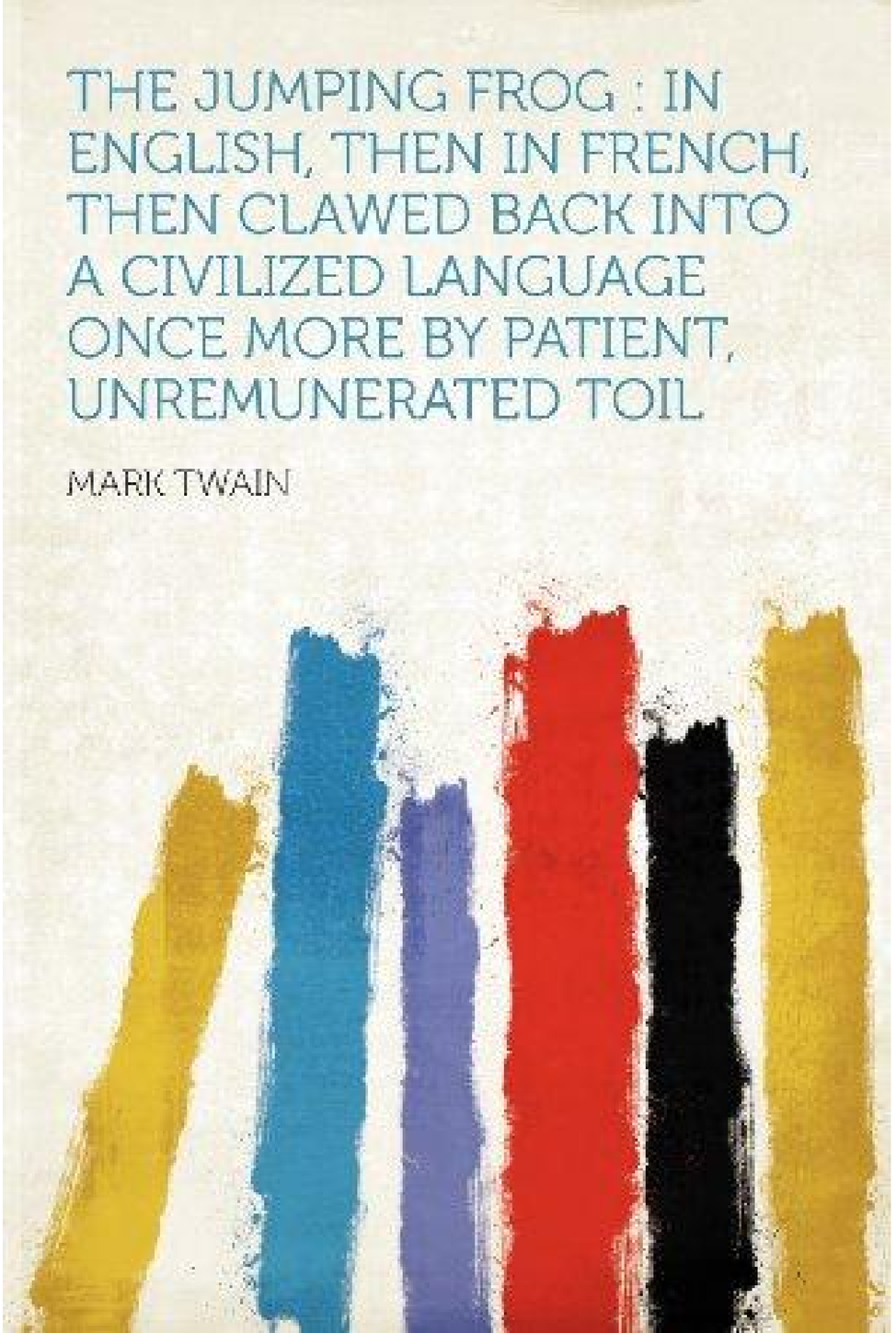
Cycle-consistency

- Composite flows along a cycle should be zero
- 2-cycle consistency: $F_{i,j} \circ F_{j,i} = 0$



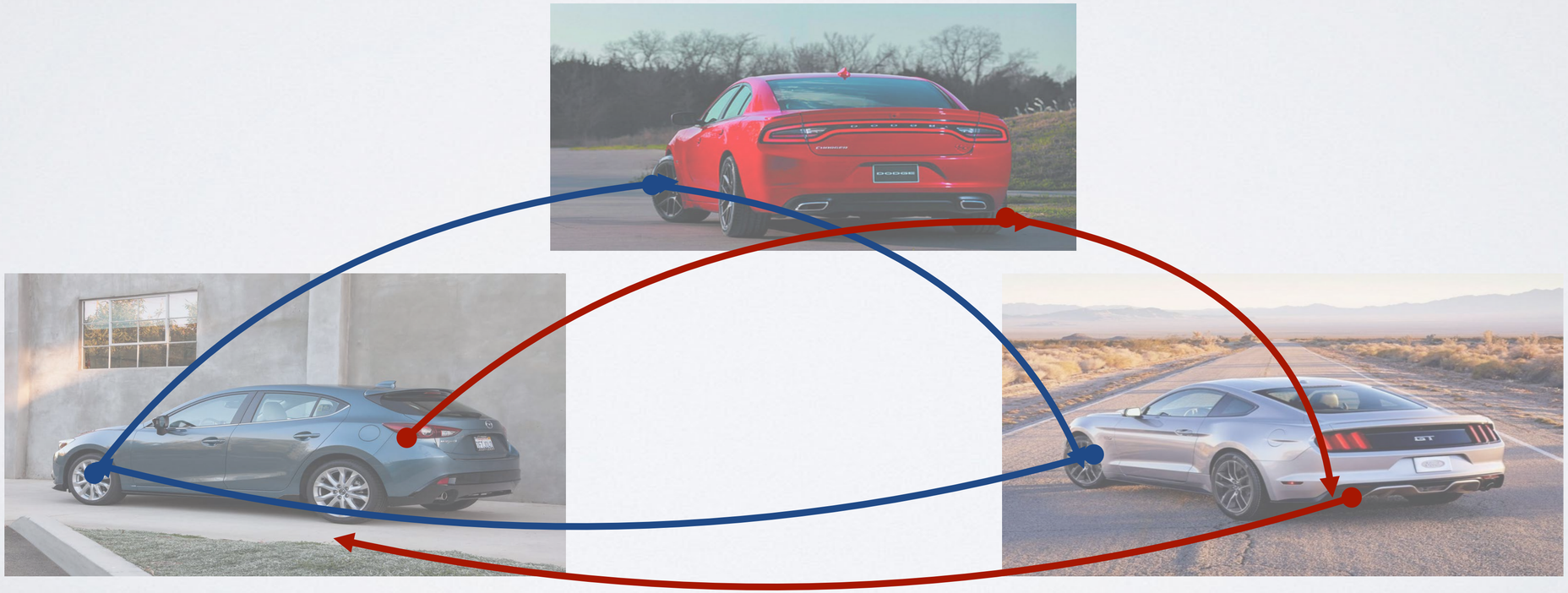
[Twain,1903]

- ***Original***: *‘Well,’ he says, ‘I don’t see no p’ints about that frog that’s any better’n any other frog.’*
- ***Back Translation***: *“Eh bien! I no saw not that that frog had nothing of better than each frog.”*



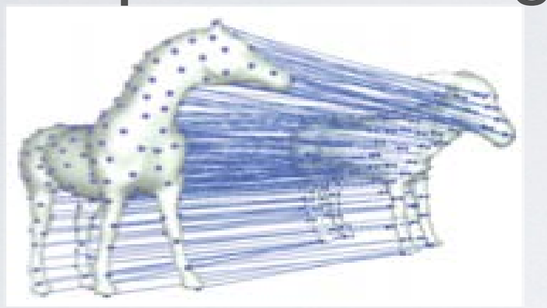
Cycle-consistency

- Composite flows along a cycle should be zero
- 2-cycle consistency: $F_{i,j} \circ F_{j,i} = 0$
- 3-cycle consistency: $F_{i,k} \circ F_{k,j} \circ F_{j,i} = 0$



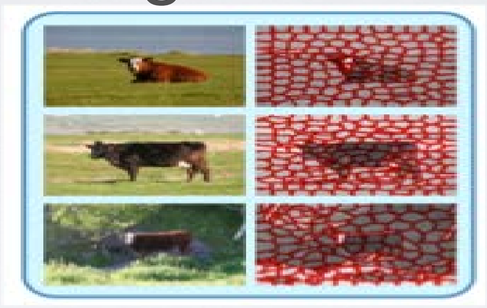
Cycle Consistency in Vision

Shape Matching



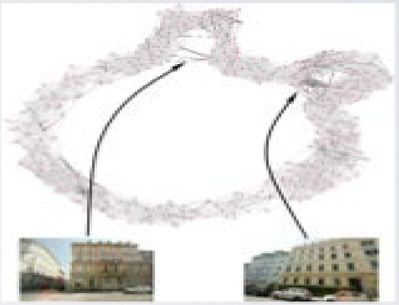
Huang *et al,* SGP’13

Co-segmentation



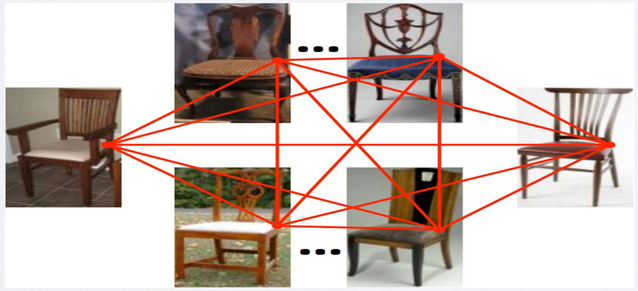
Wang *et al,* ICCV’13

SfM

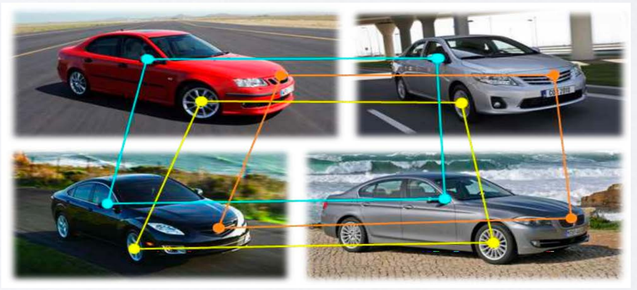


Zach *et al*, CVPR’10

Collection Correspondence



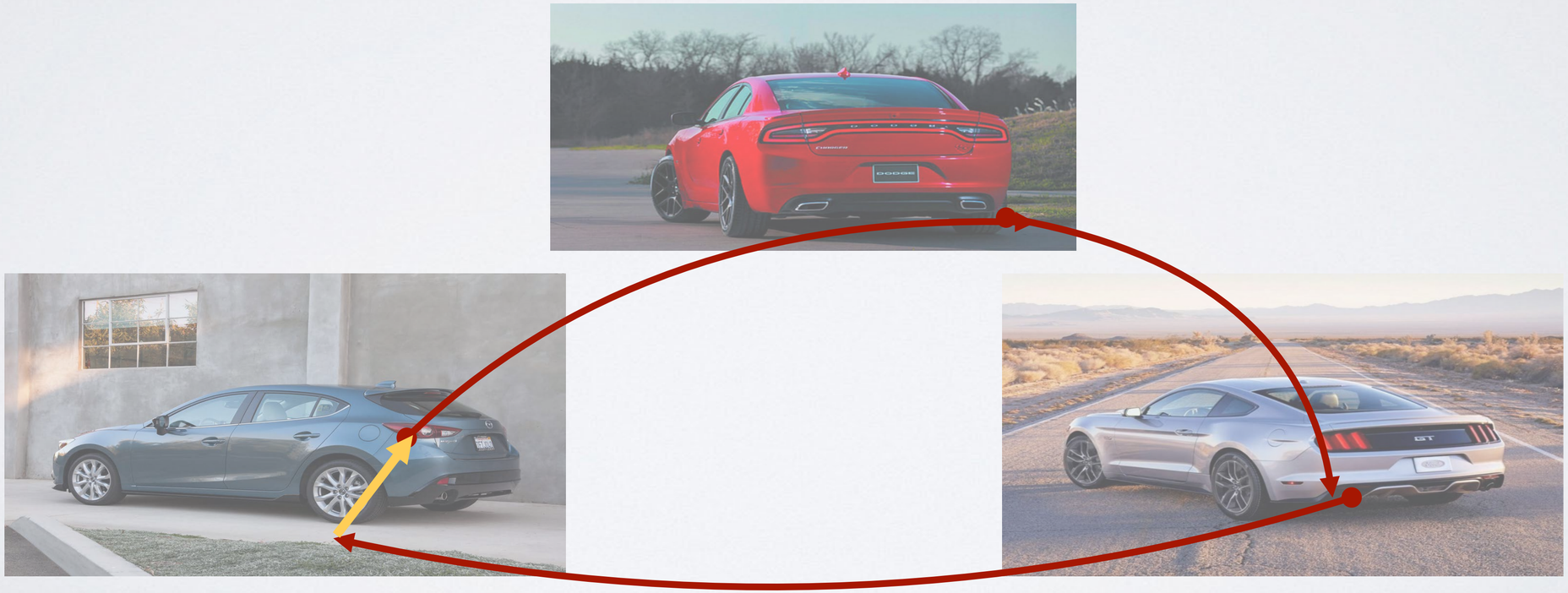
Zhou *et al,* CVPR’15



Zhou *et al,* ICCV’15

Cycle-consistency as Supervision

- Composite flows along a cycle should be zero
- 2-cycle consistency: $F_{i,j} \circ F_{j,i} = 0$
- 3-cycle consistency: $F_{i,k} \circ F_{k,j} \circ F_{j,i} = 0$



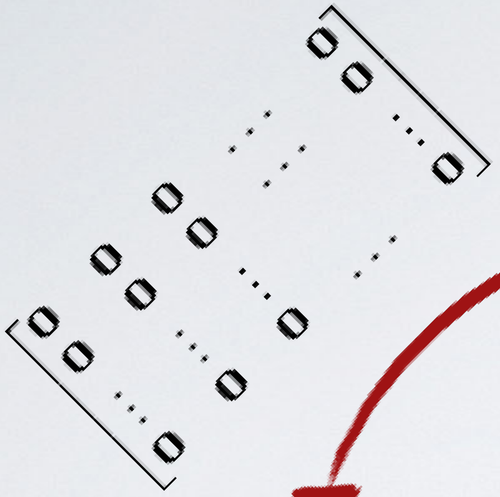
Cycle-consistency as Supervision

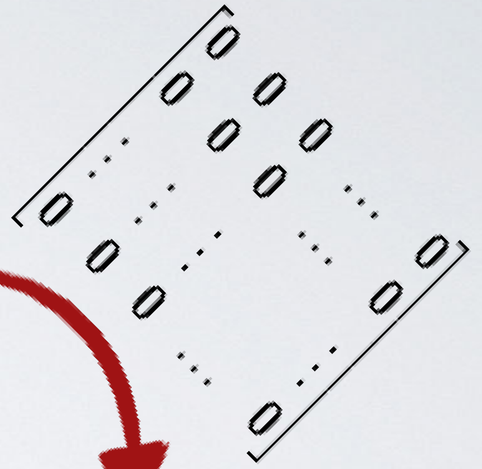
- Composite flows along a cycle should be zero
- 2-cycle consistency: $F_{i,j} \circ F_{j,i} = 0$
- 3-cycle consistency: $F_{i,k} \circ F_{k,j} \circ F_{j,i} = 0$

CNN

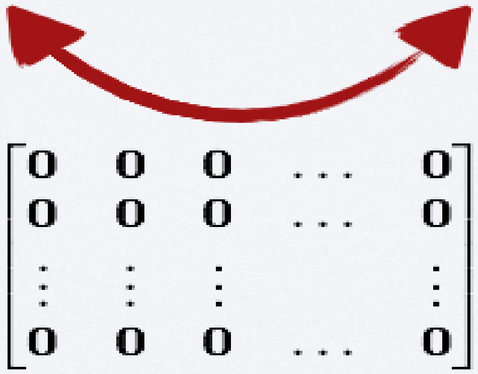
Amount of inconsistency

Could be consistent but wrong…

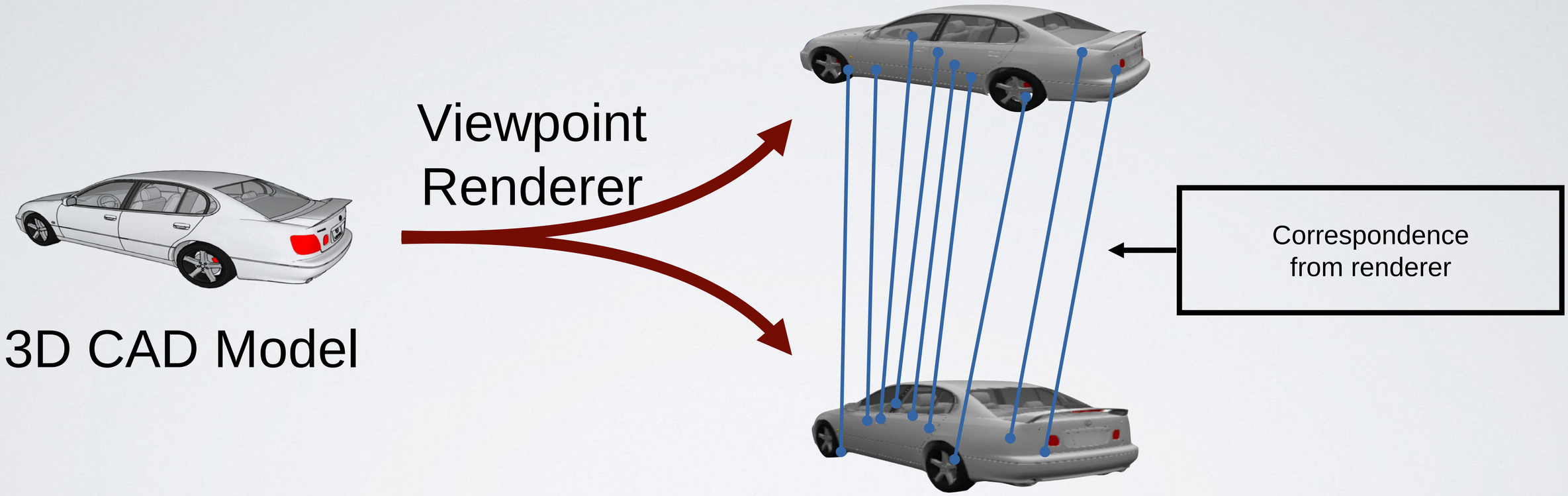




Need an anchor edge!



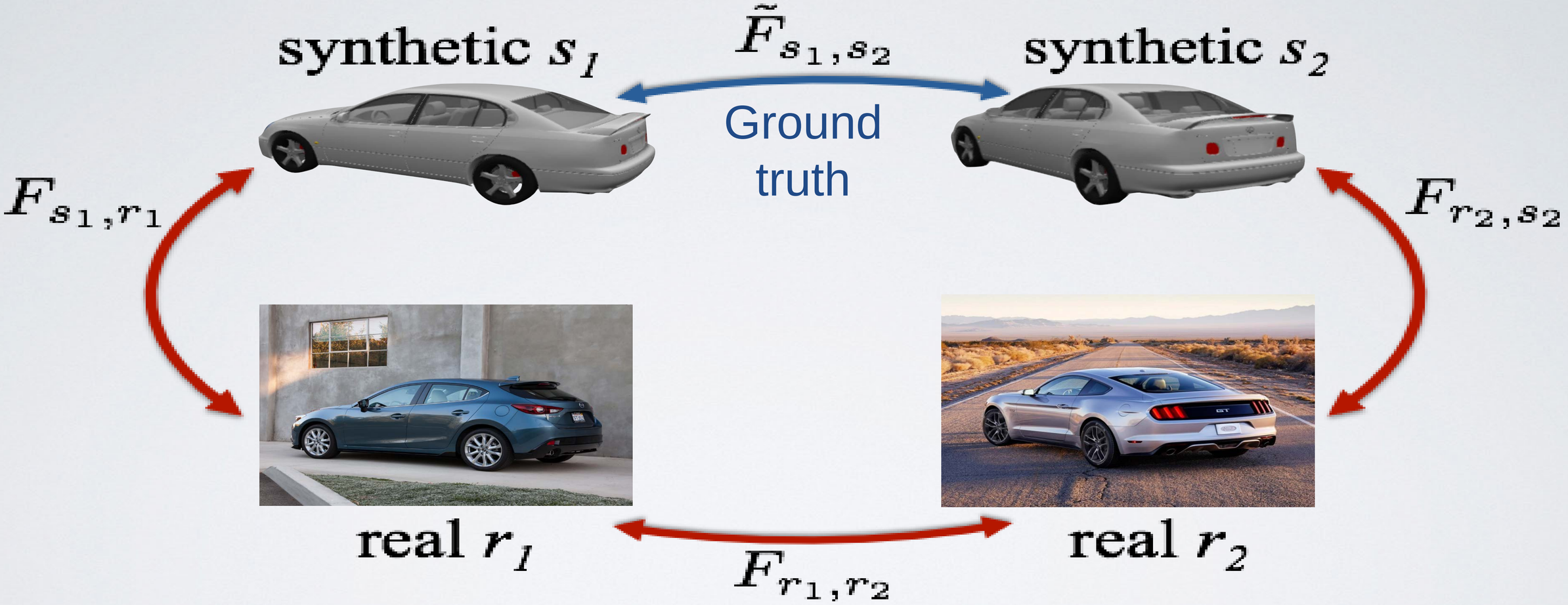
Synthetic Correspondence as the Anchor



Learning Dense Correspondence via 3D-guided Cycle Consistency

Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qixing Huang, Alexei A. Efros , CVPR'16

3D-guided Cycle Consistency



$$\tilde{F}_{s_1,s_2} = F_{s_1,r_1} \circ F_{r_1,r_2} \circ F_{r_2,s_2}$$

Learning Dense Correspondence via 3D-guided Cycle Consistency

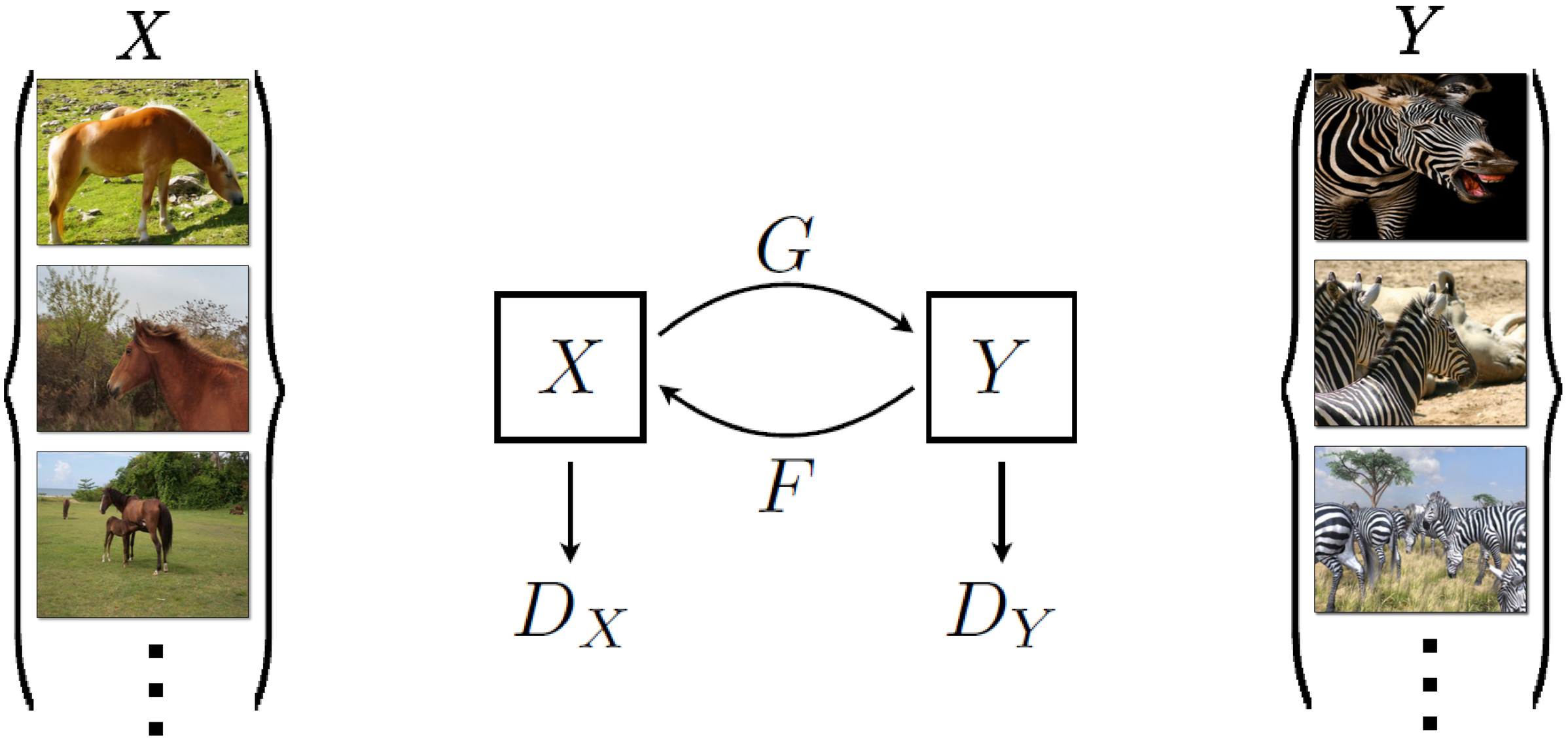
Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qixing Huang, Alexei A. Efros , CVPR'16

OUR RESULT

Learning Dense Correspondence via 3D-guided Cycle Consistency

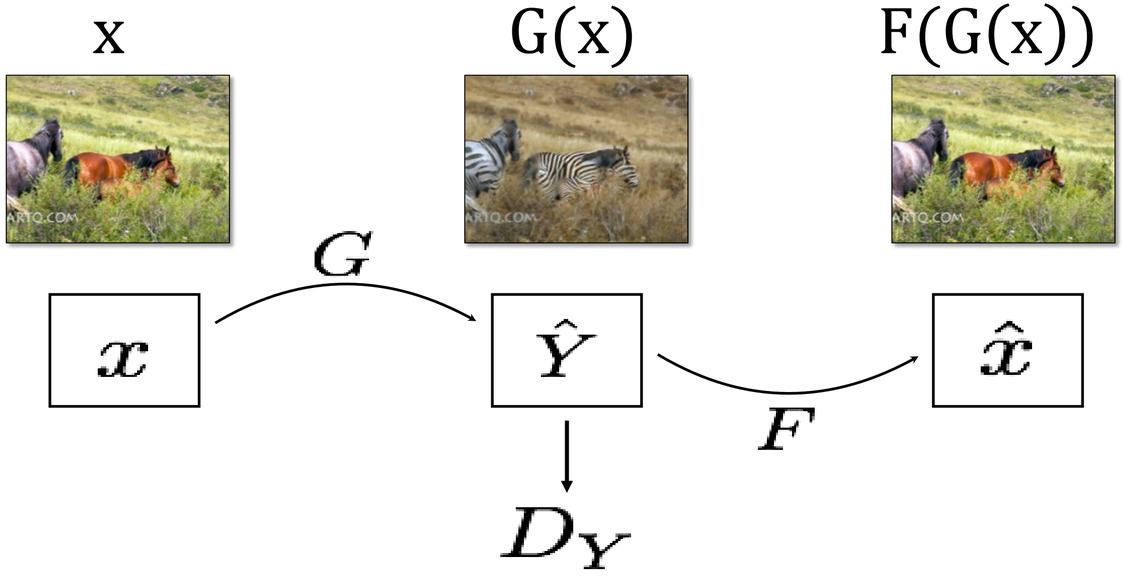
Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qixing Huang, Alexei A. Efros , CVPR'16

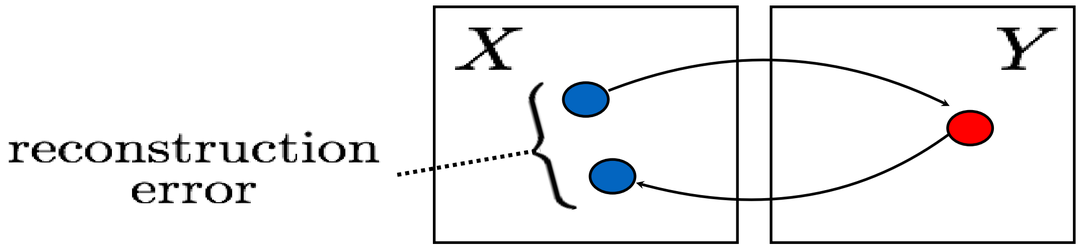
CycleGAN, or “there and back aGAN”



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

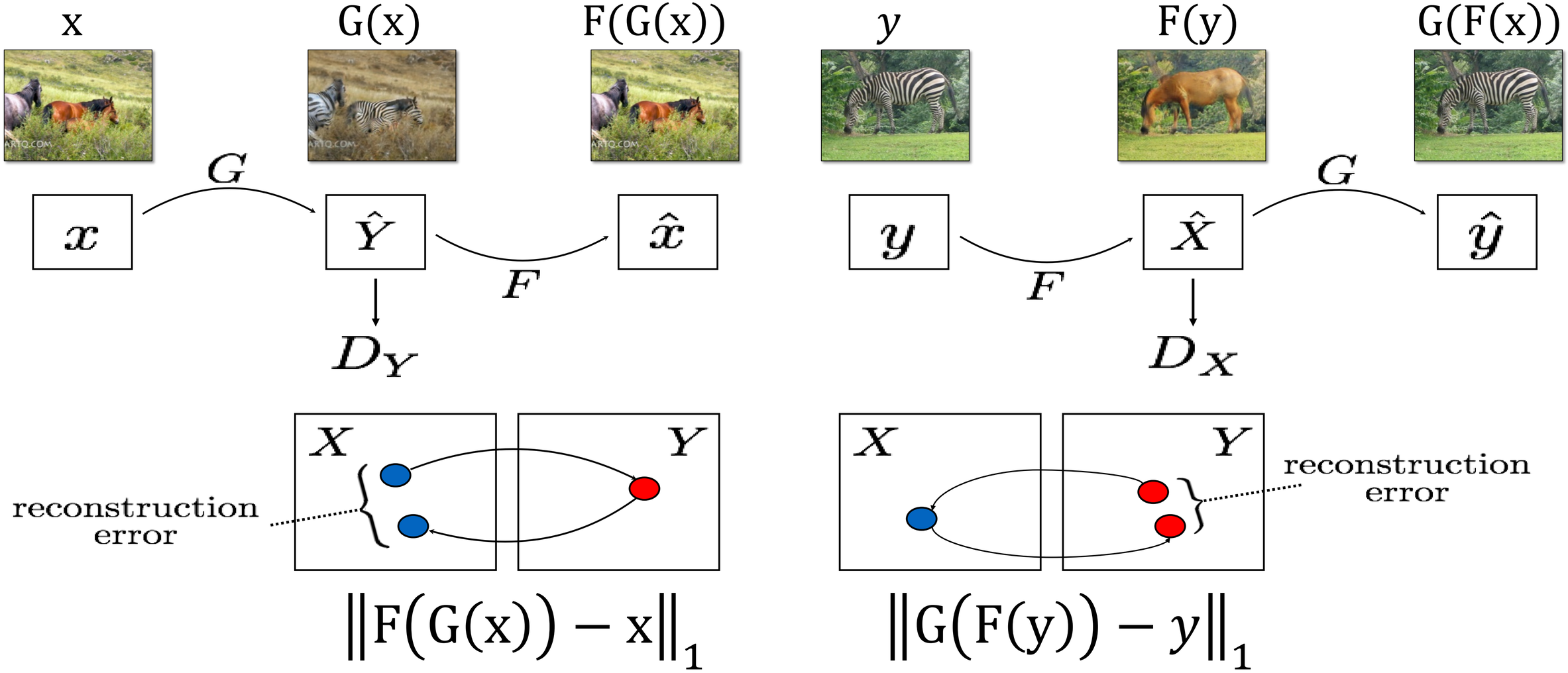
Cycle-Consistency Loss





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

Cycle-Consistency Loss



Video

Failure

Collection Style Transfer

Photograph
@ Alexei Efros

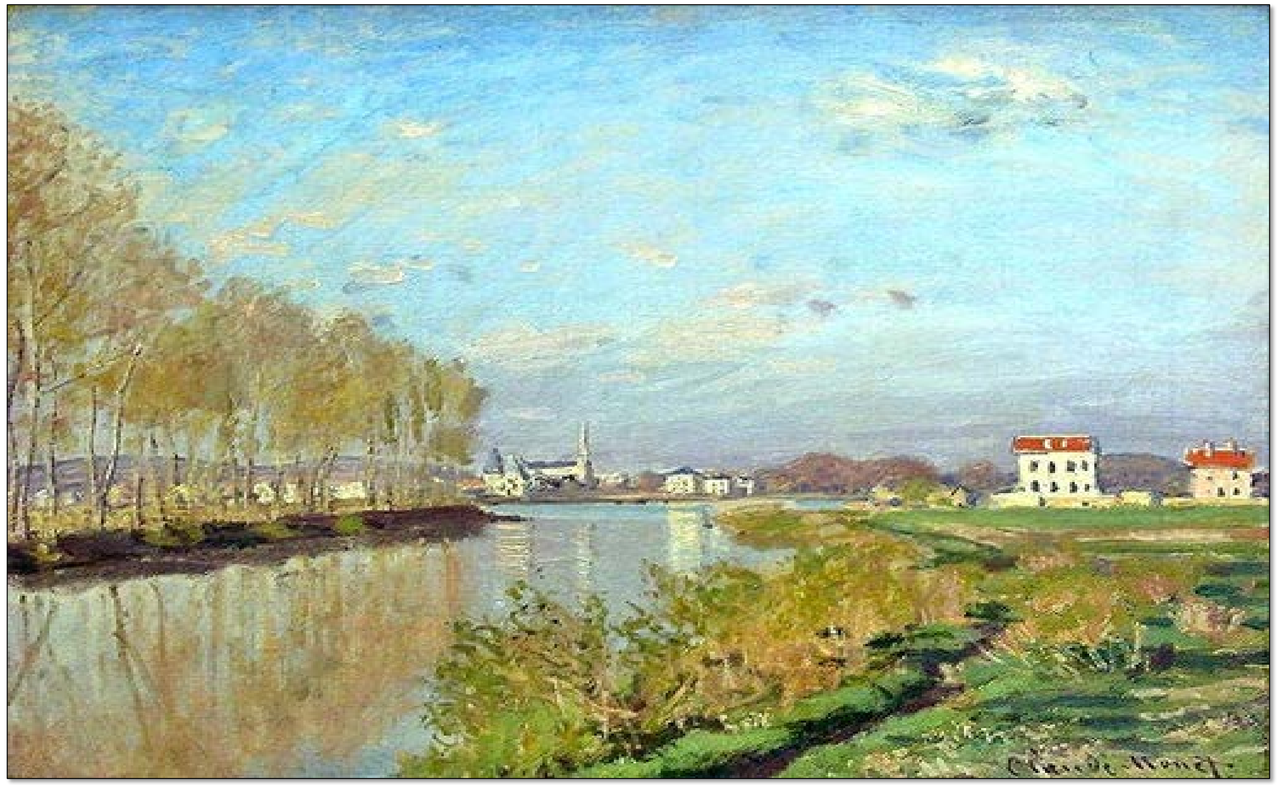
Monet

Van Gogh

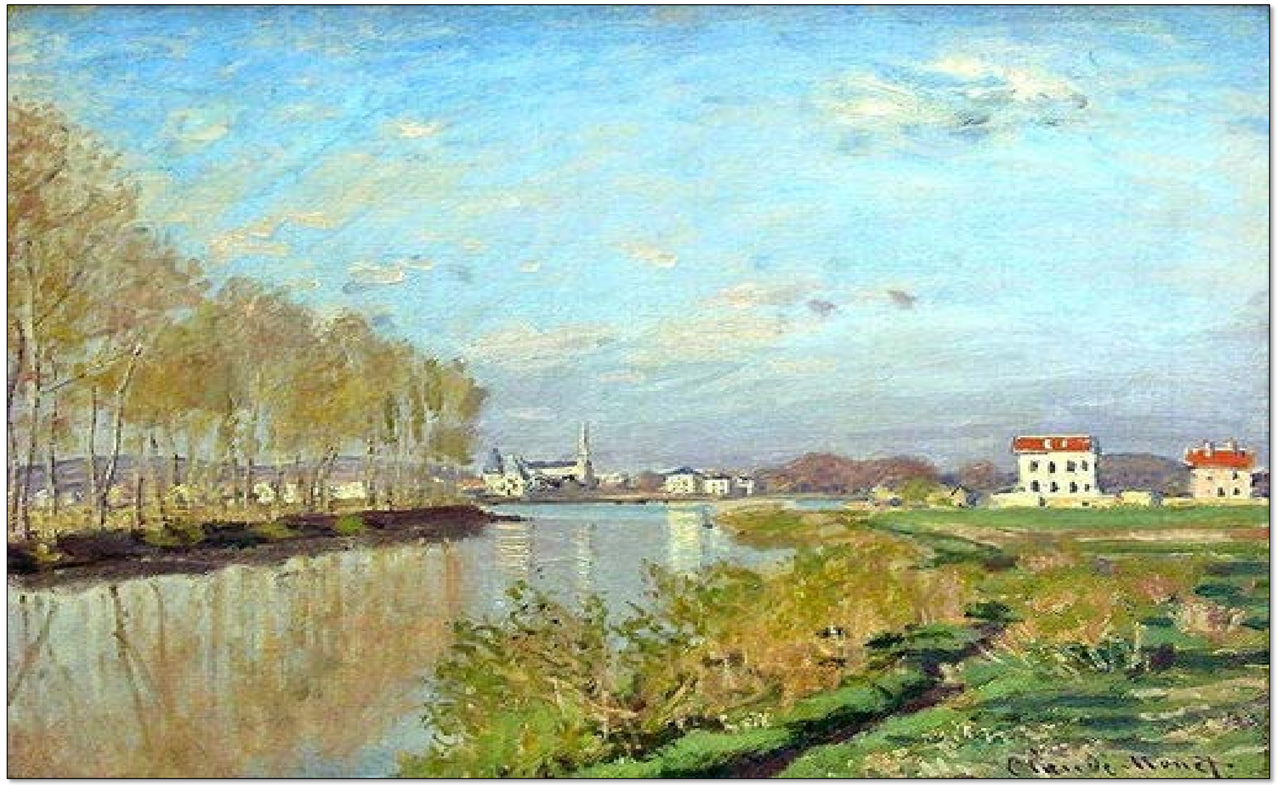
Cezanne

Ukiyo-e

Monet → photo



Monet → photo



Video Game to Real

Grand Theft Auto

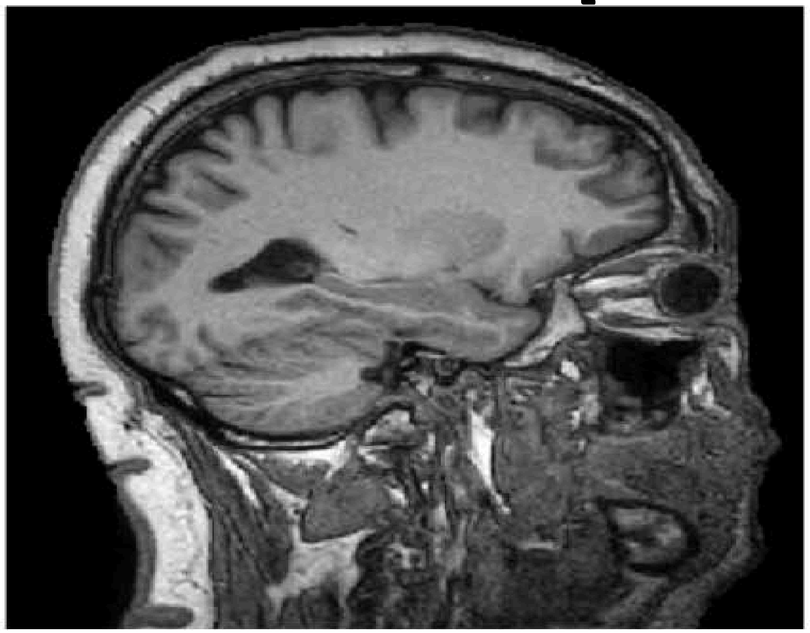
→

→

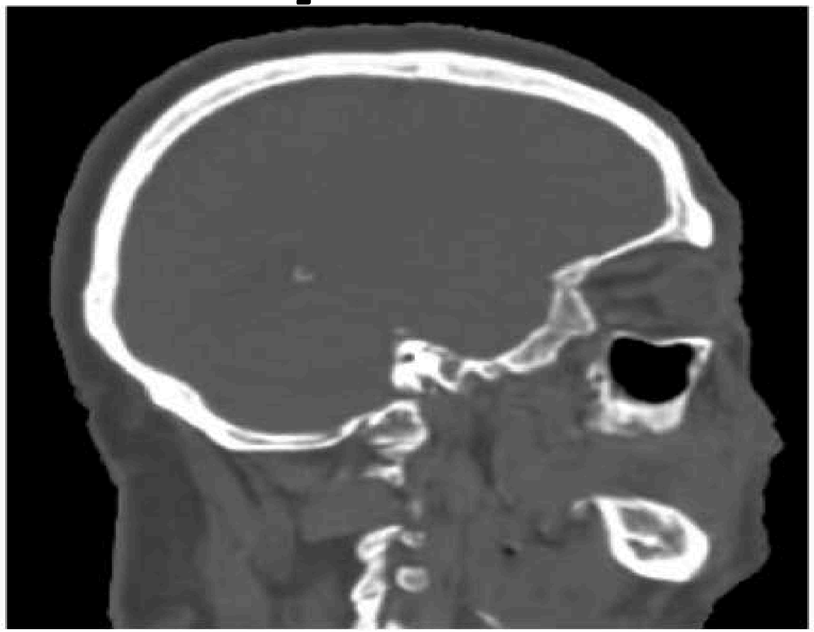
Real to Video Game

Applications of CycleGAN

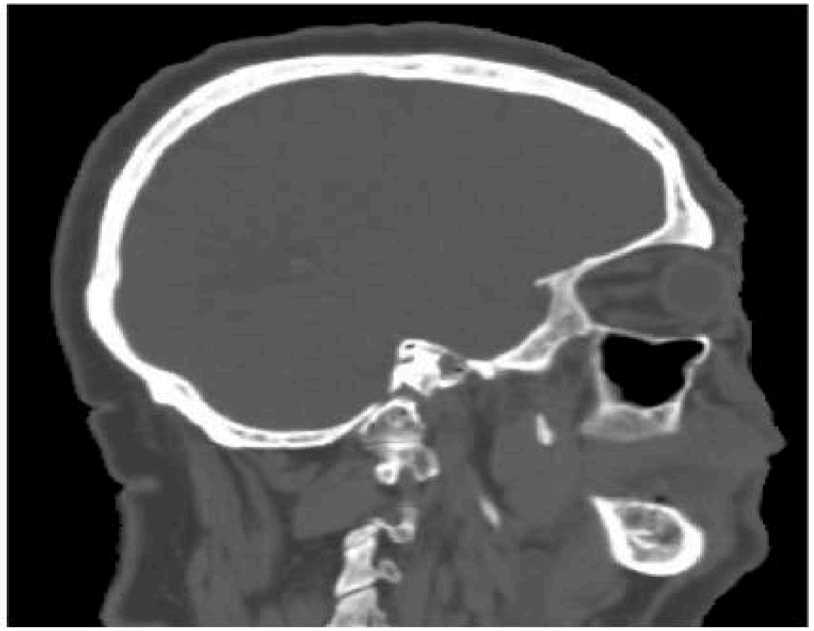
MR → CT [Wolterink et al] arxiv: 1708.01155



Input MR



Generated CT



Ground truth CT

- **MRI reconstruction** [Quan et al.] arxiv:1709.00753
- **Cardiac MR images from CT** [Chartsias et al. 2017]

Thank You!