Paired image-to-image translation

Phillip Isola
OpenAI/MIT
6/22/18

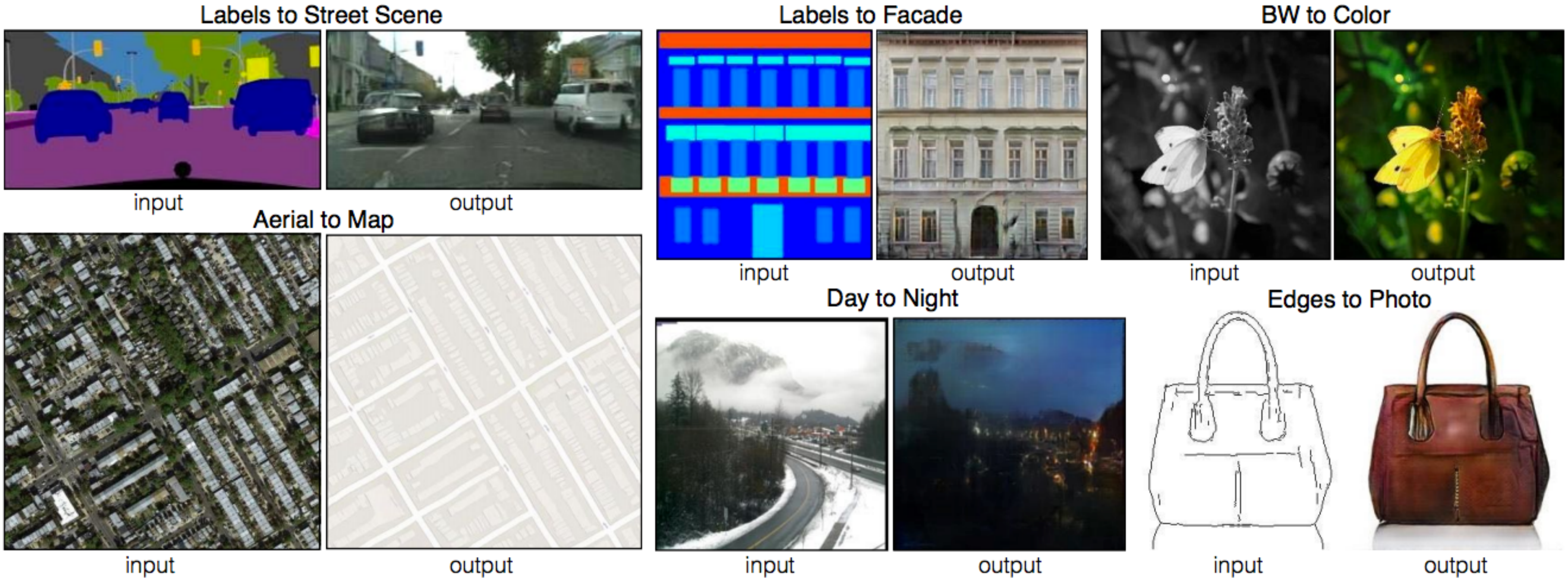
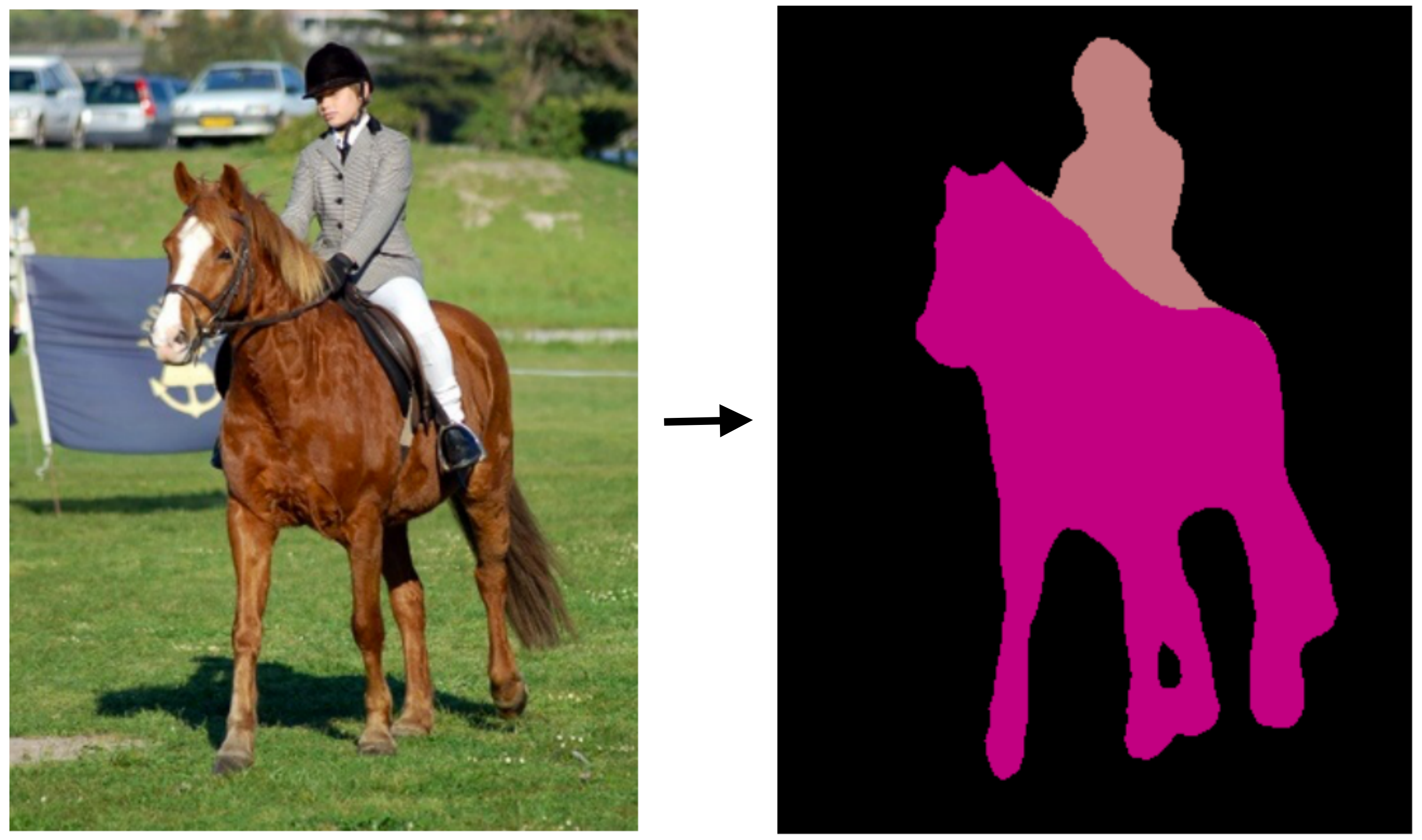


Image-to-Image Translation

Image-to-Image Translation

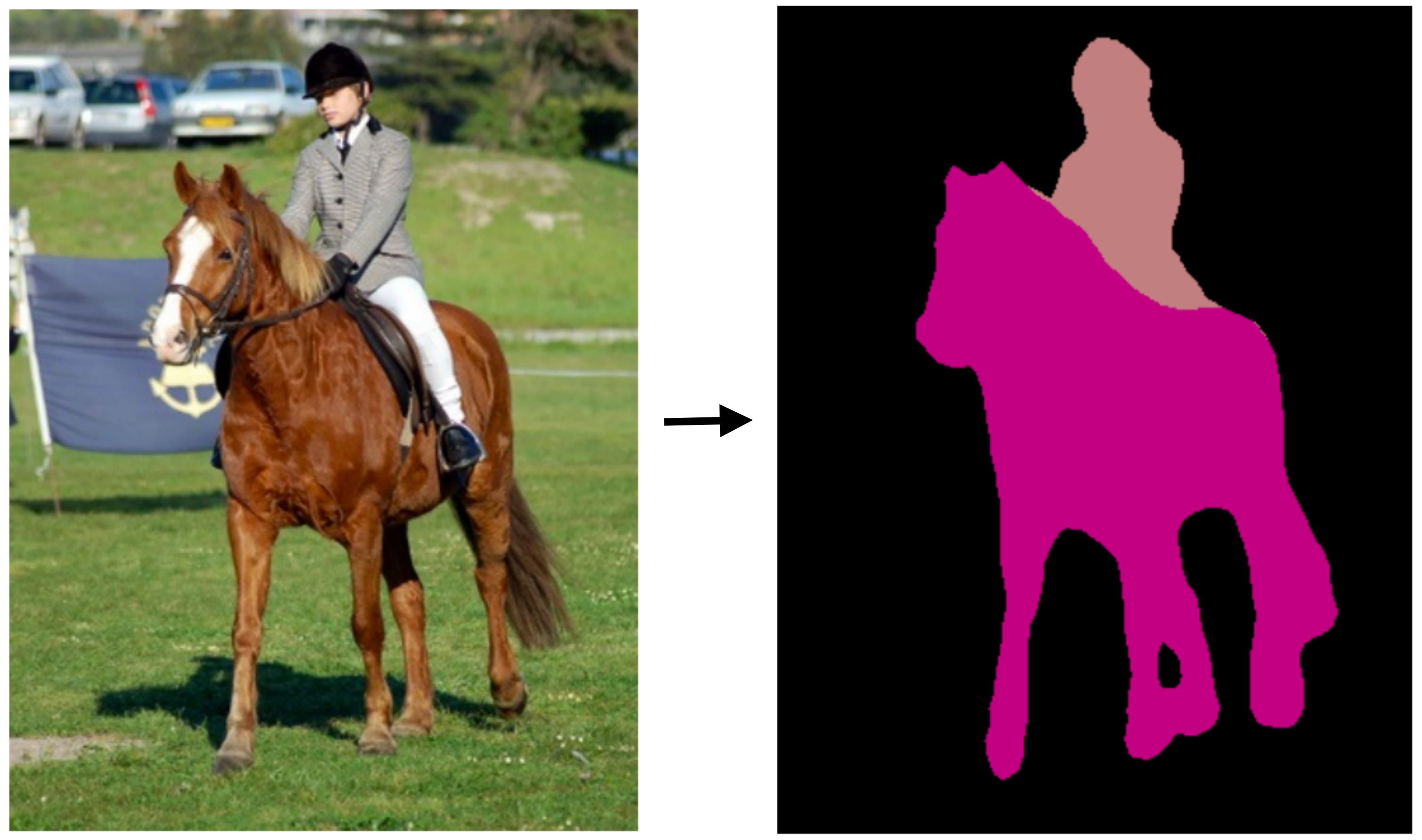
Object labeling



[Long et al. 2015]

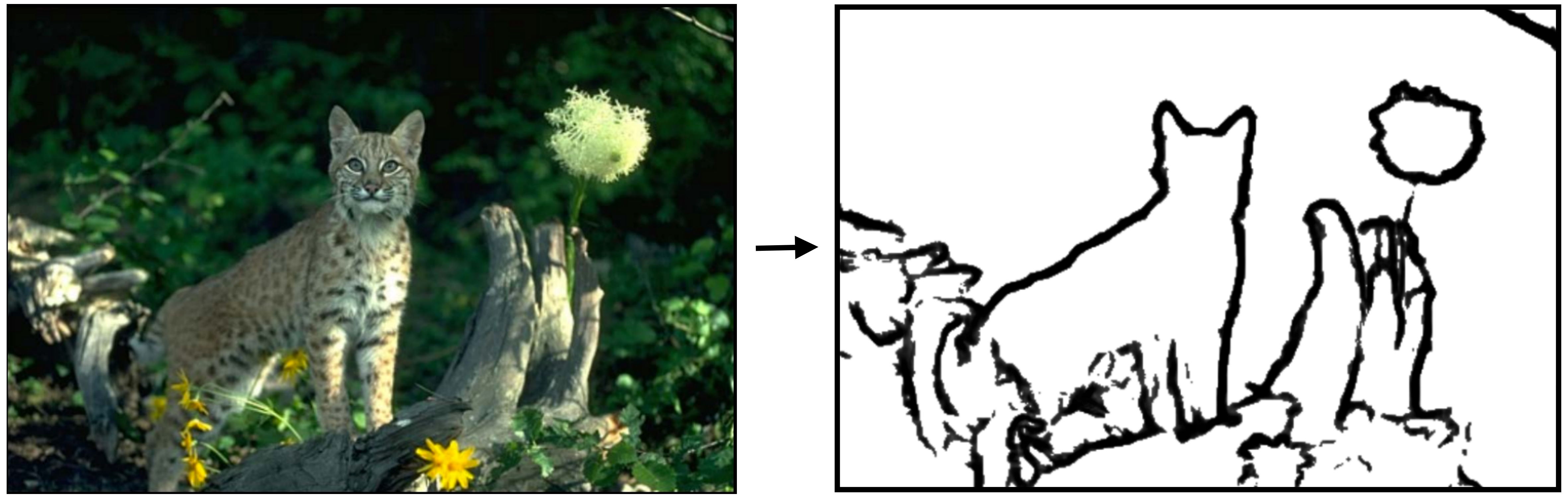
Image-to-Image Translation

Object labeling



[Long et al. 2015]

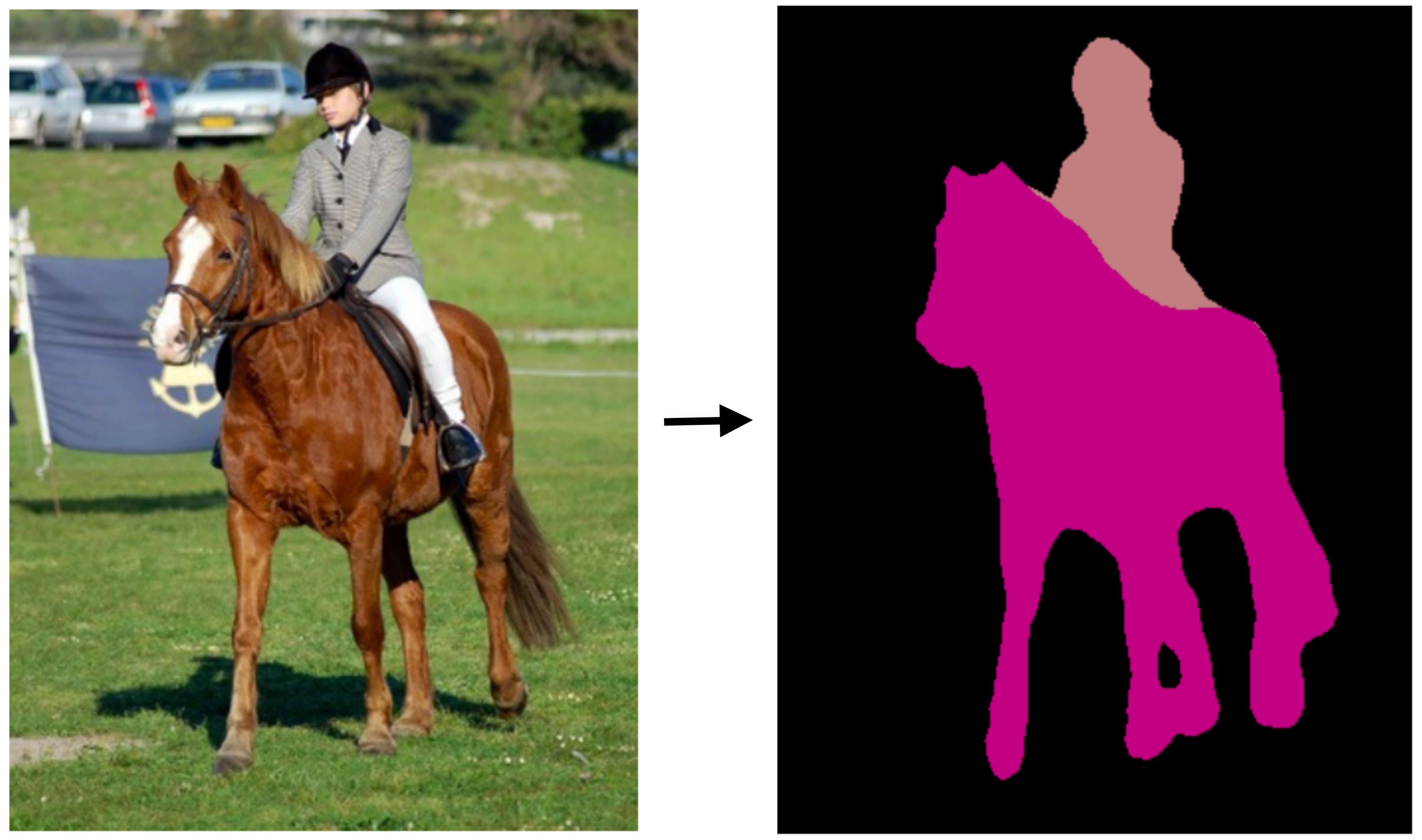
Edge Detection



[Xie et al. 2015]

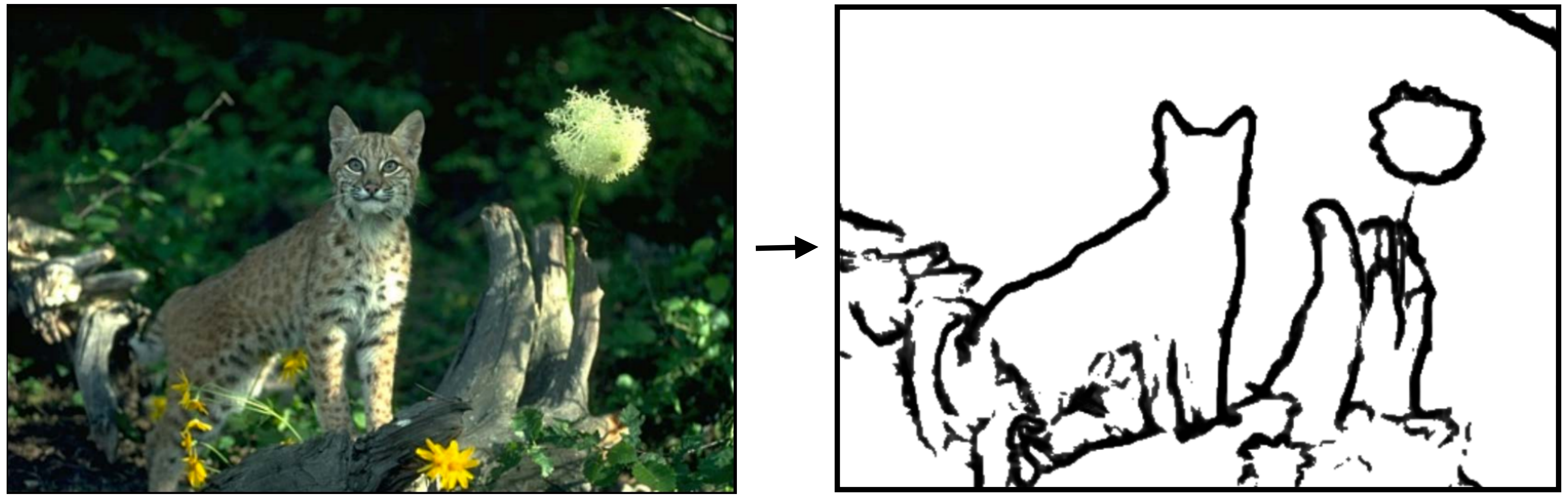
Image-to-Image Translation

Object labeling



[Long et al. 2015]

Edge Detection



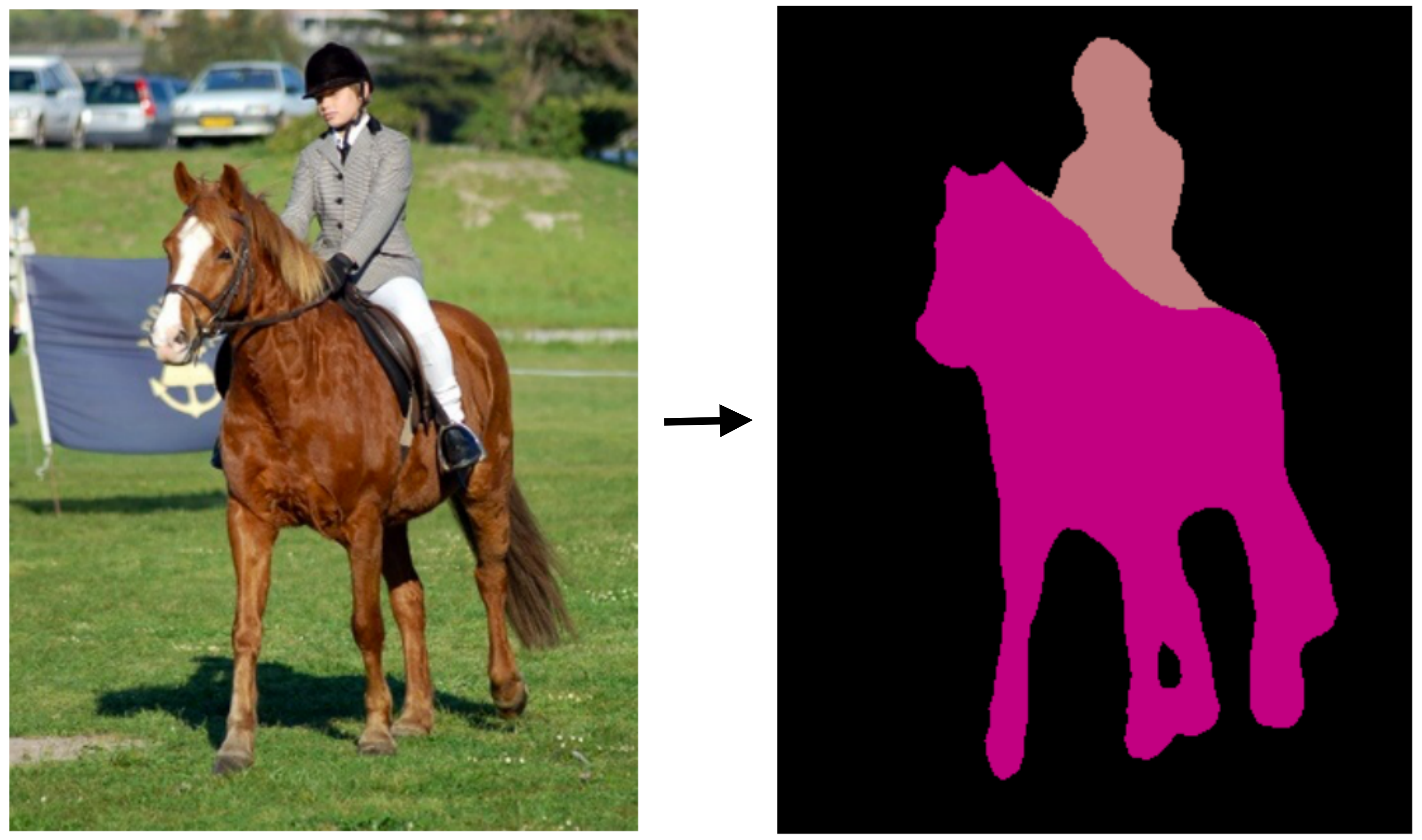
[Xie et al. 2015]

Season change

[Laffont et al. 2014]

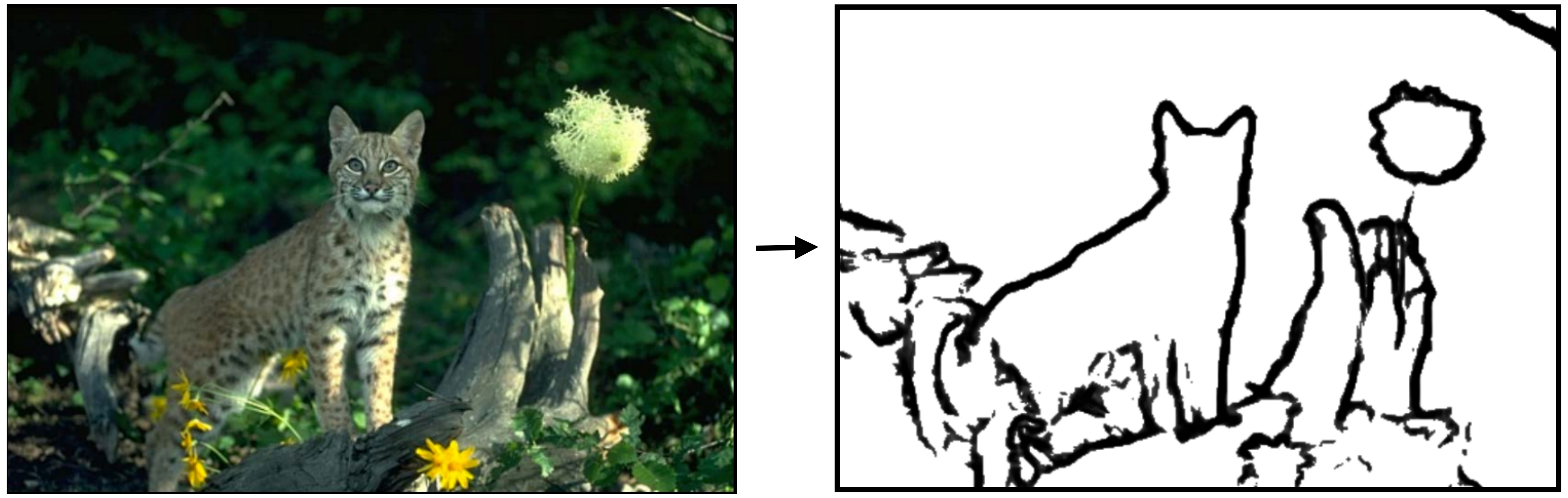
Image-to-Image Translation

Object labeling



[Long et al. 2015]

Edge Detection

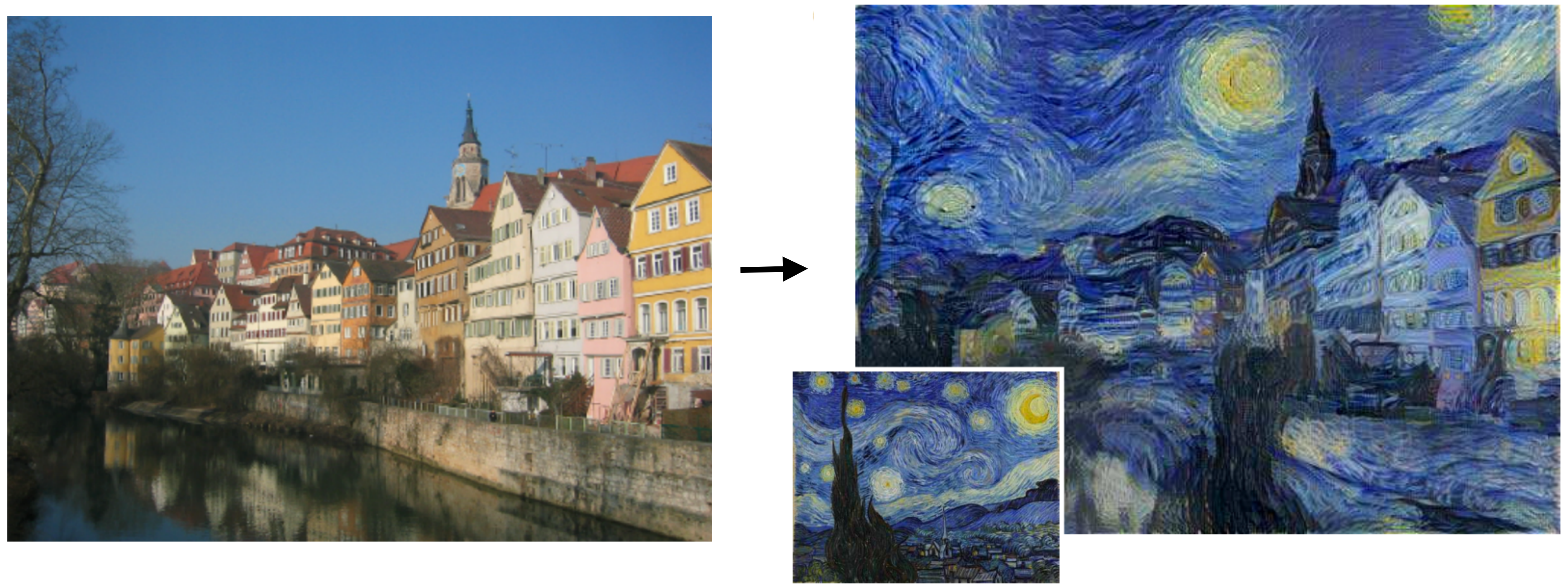


[Xie et al. 2015]

Season change

[Laffont et al. 2014]

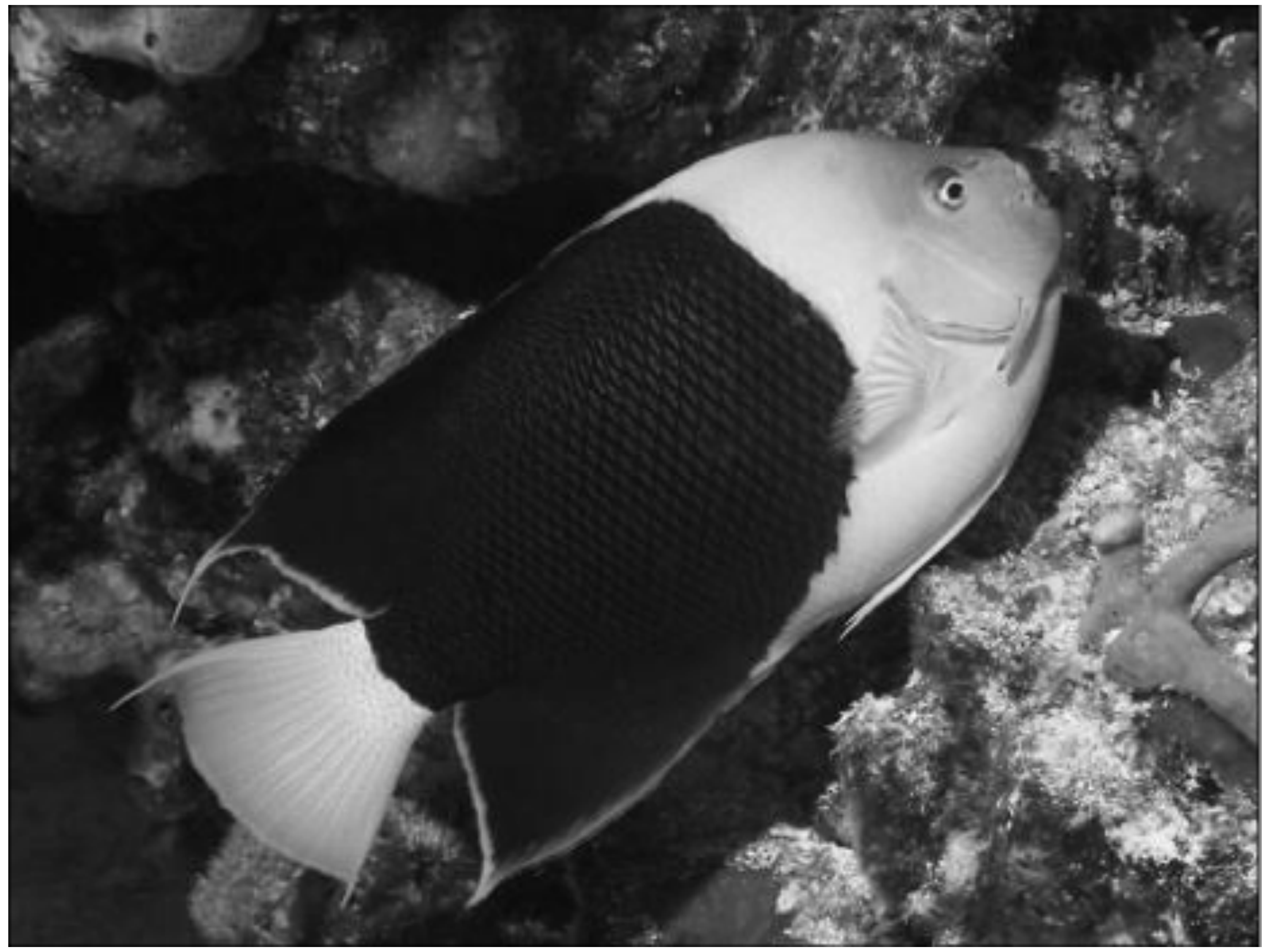
Artistic style transfer



[Gatys et al. 2016]

Paired Image-to-Image Translation

Input $\mathbf{x}$



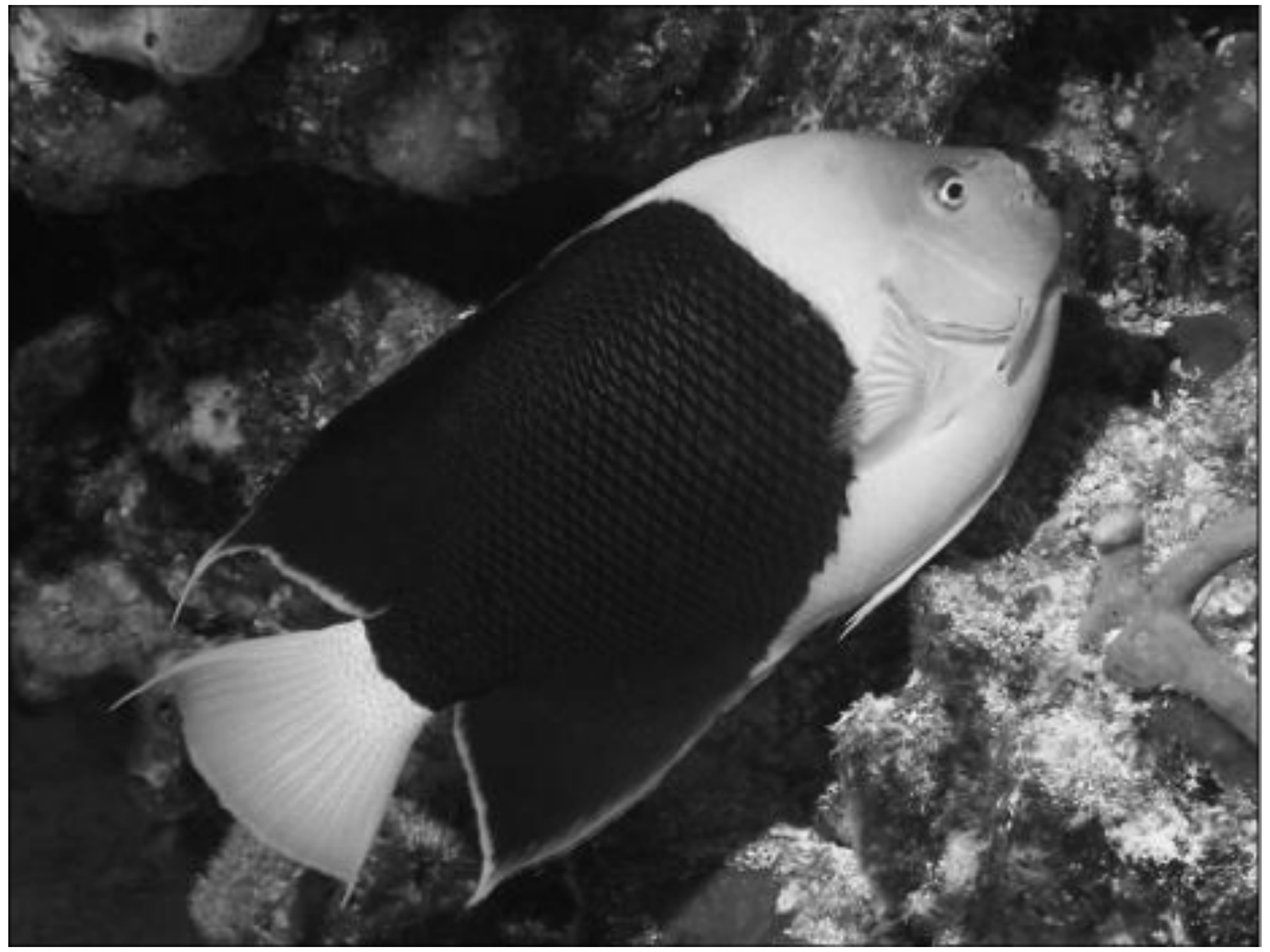
$\mathcal{F}$

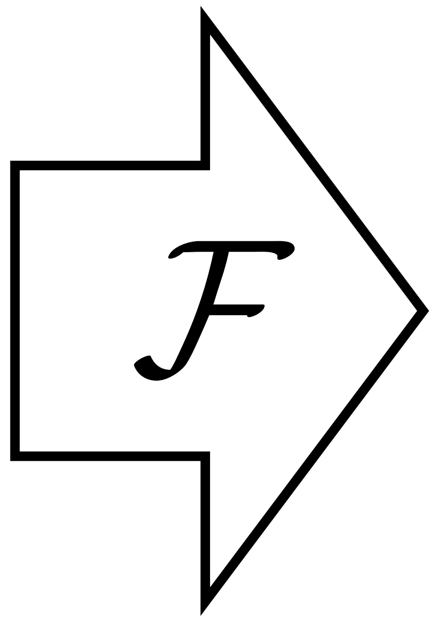
Output $\mathbf{y}$

[Zhang et al., ECCV 2016]

Paired Image-to-Image Translation

Input $\mathbf{x}$





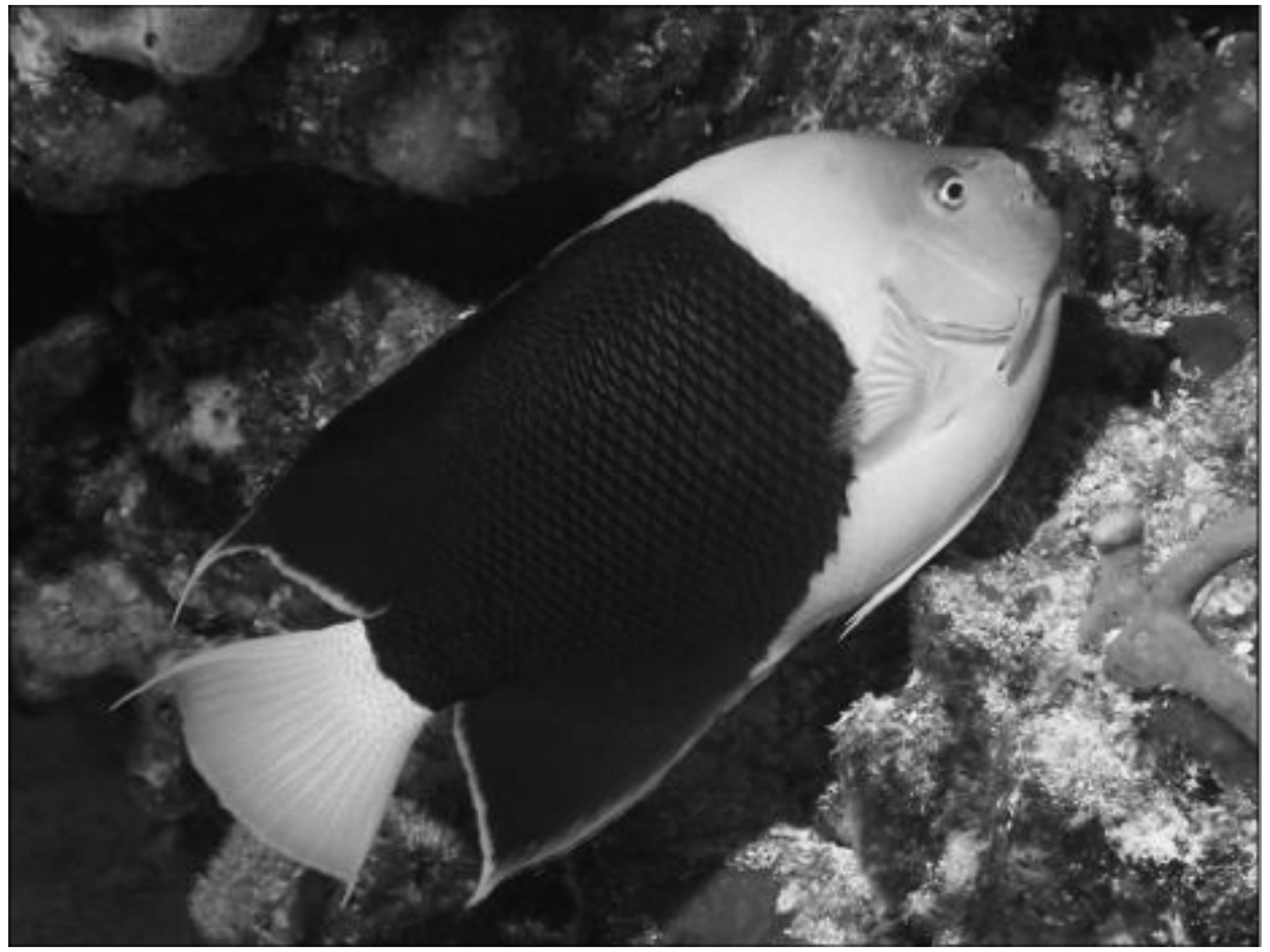
Output $\mathbf{y}$

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

[Zhang et al., ECCV 2016]

Paired Image-to-Image Translation

Input $\mathbf{x}$



$\mathcal{F}$

Output $\mathbf{y}$

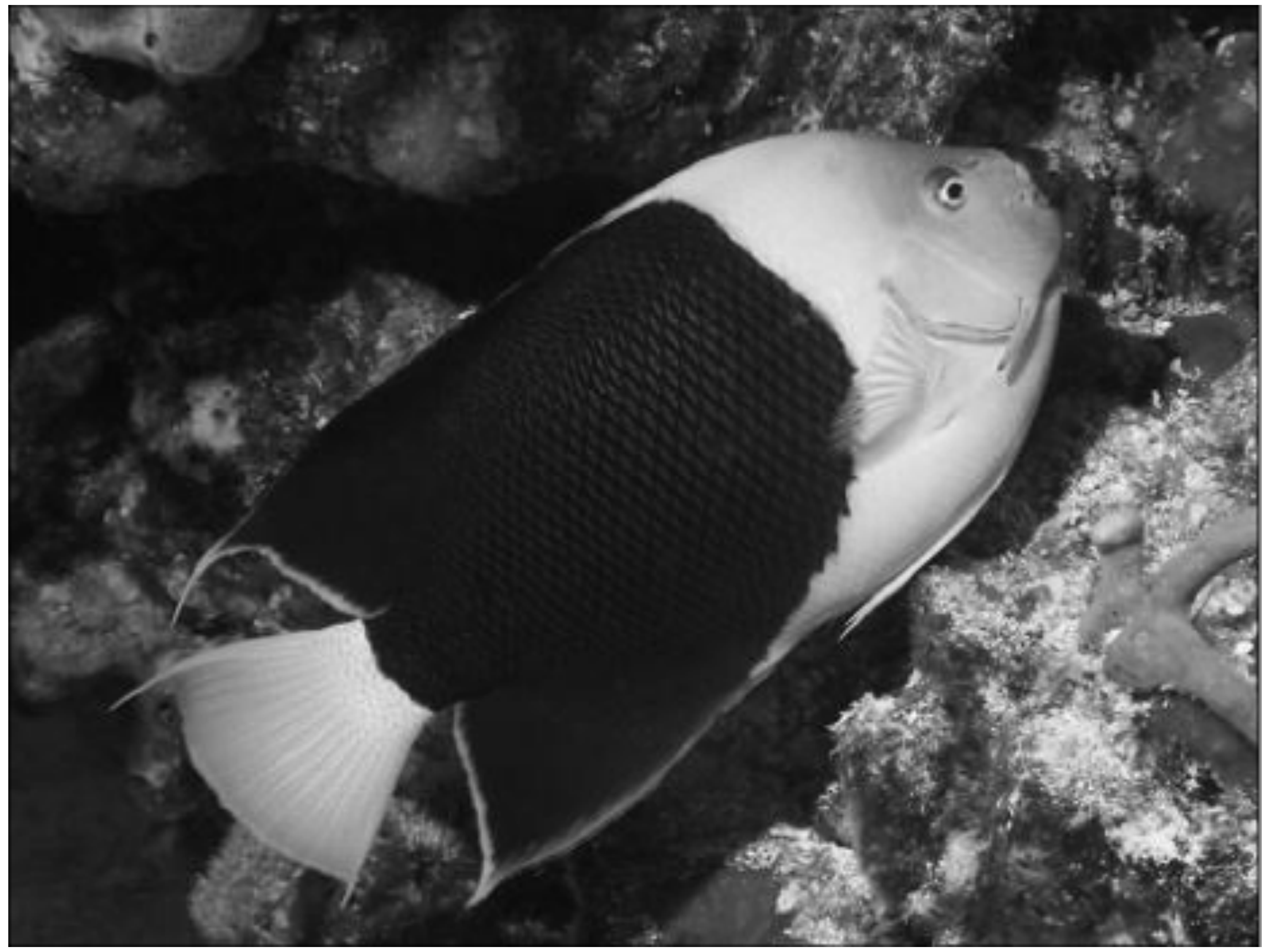
$$\arg\min_{\mathcal{F}} \mathbb{E}_{\mathbf{x},\mathbf{y}}[L(\mathcal{F}(\mathbf{x}), \mathbf{y})]$$

Neural Network

[Zhang et al., ECCV 2016]

Paired Image-to-Image Translation

Input $\mathbf{x}$



$\mathcal{F}$

Output $\mathbf{y}$

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

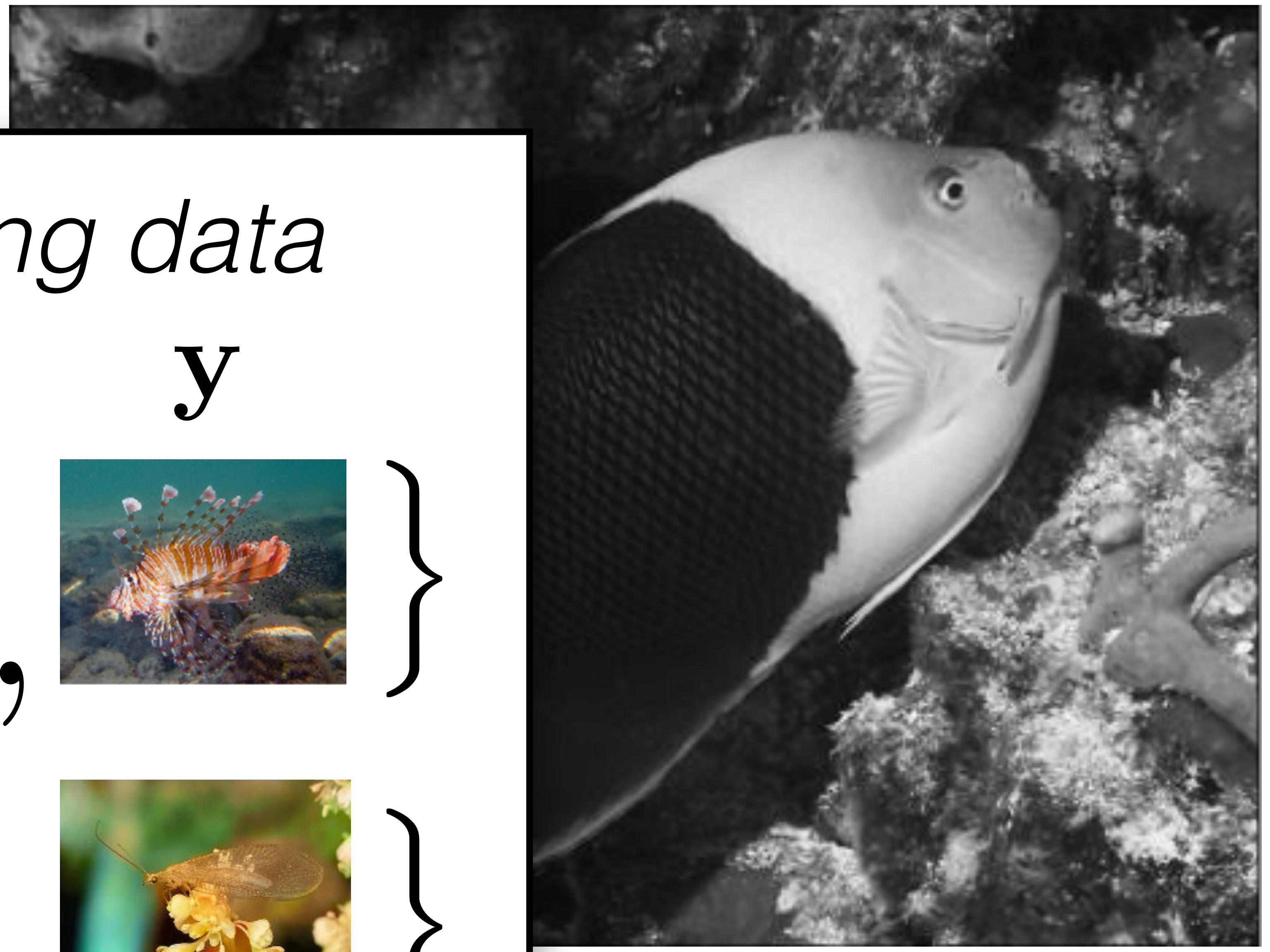
Objective function
(loss)

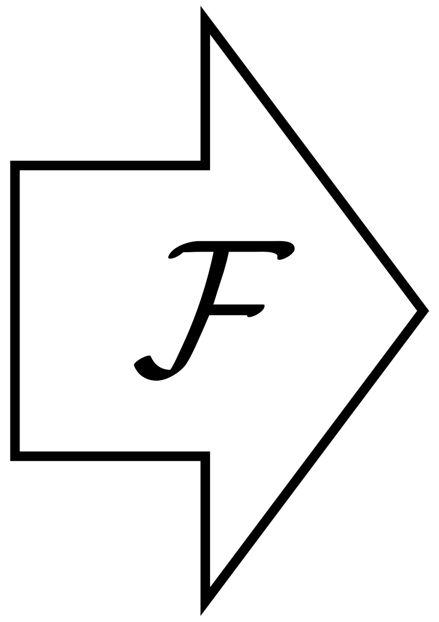
Neural Network

[Zhang et al., ECCV 2016]

Paired Image-to-Image Translation

Input $\mathbf{x}$





Output $\mathbf{y}$

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

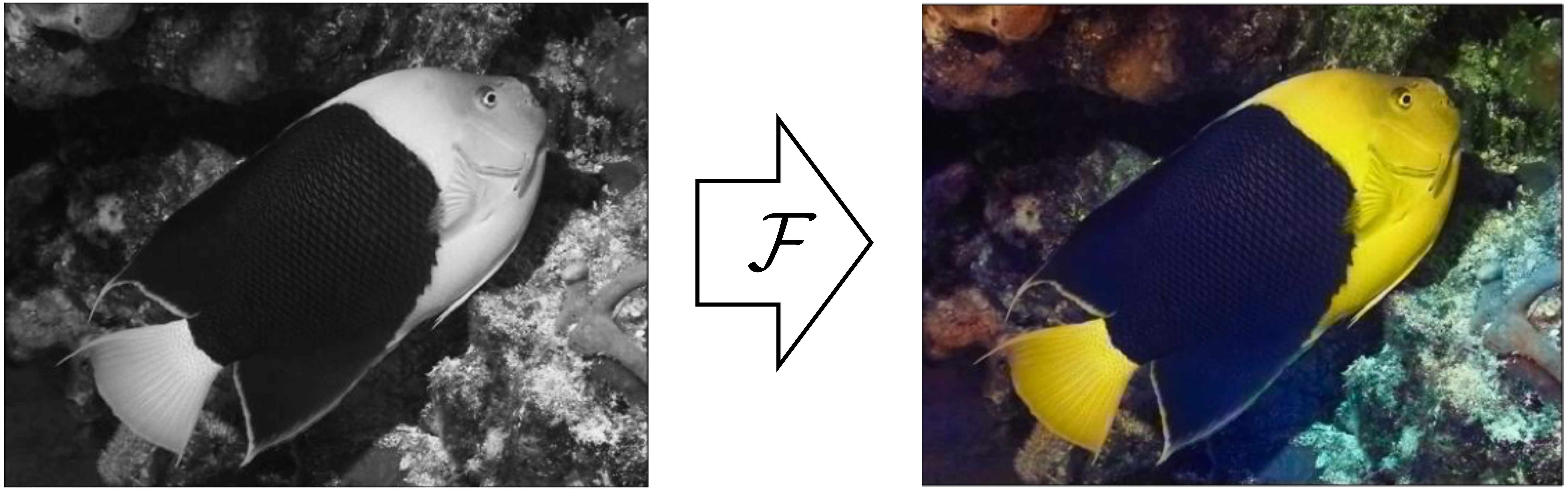
Objective function (loss)

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

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

Designing loss functions

Input

Output

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

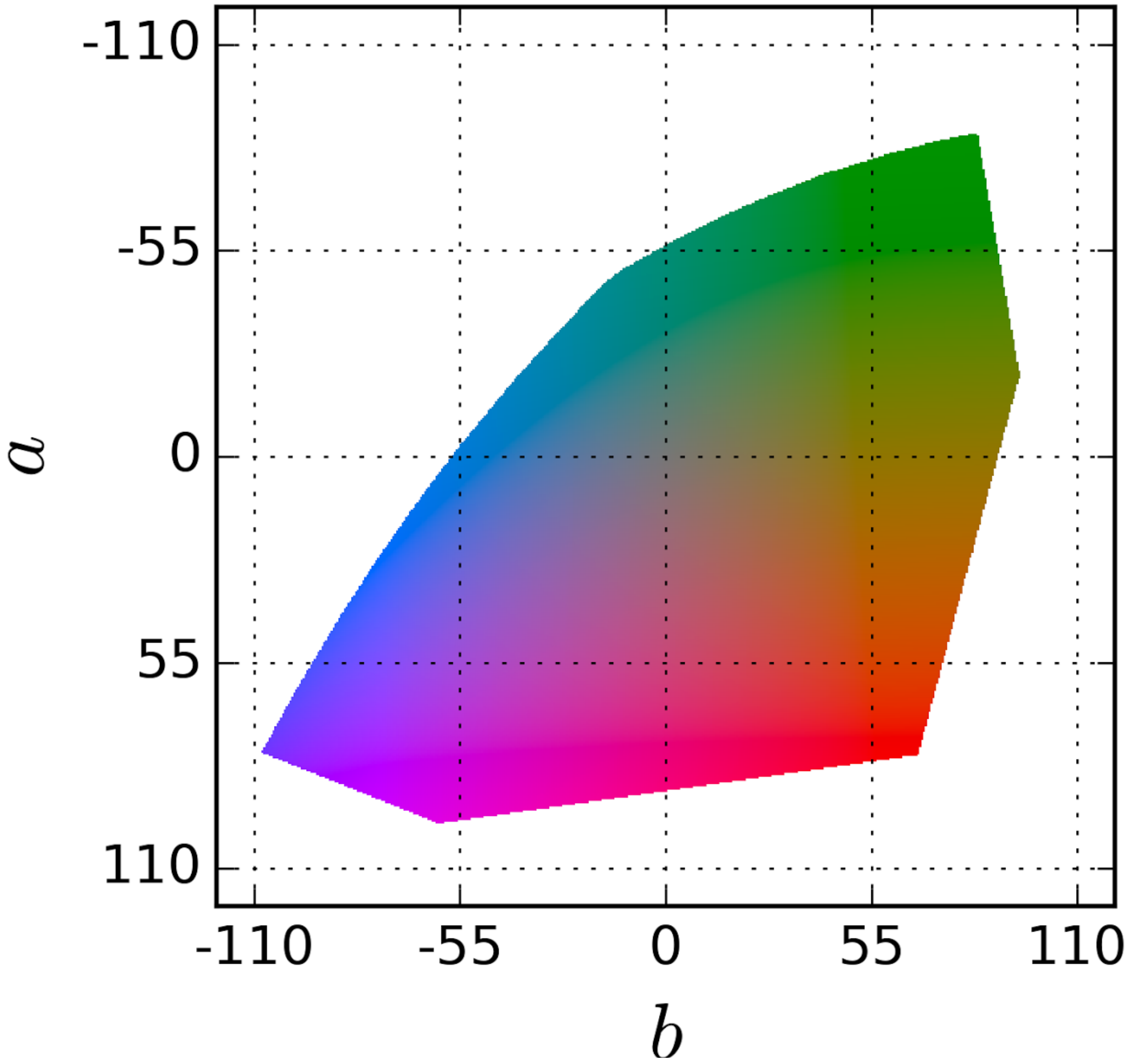
Designing loss functions

Input

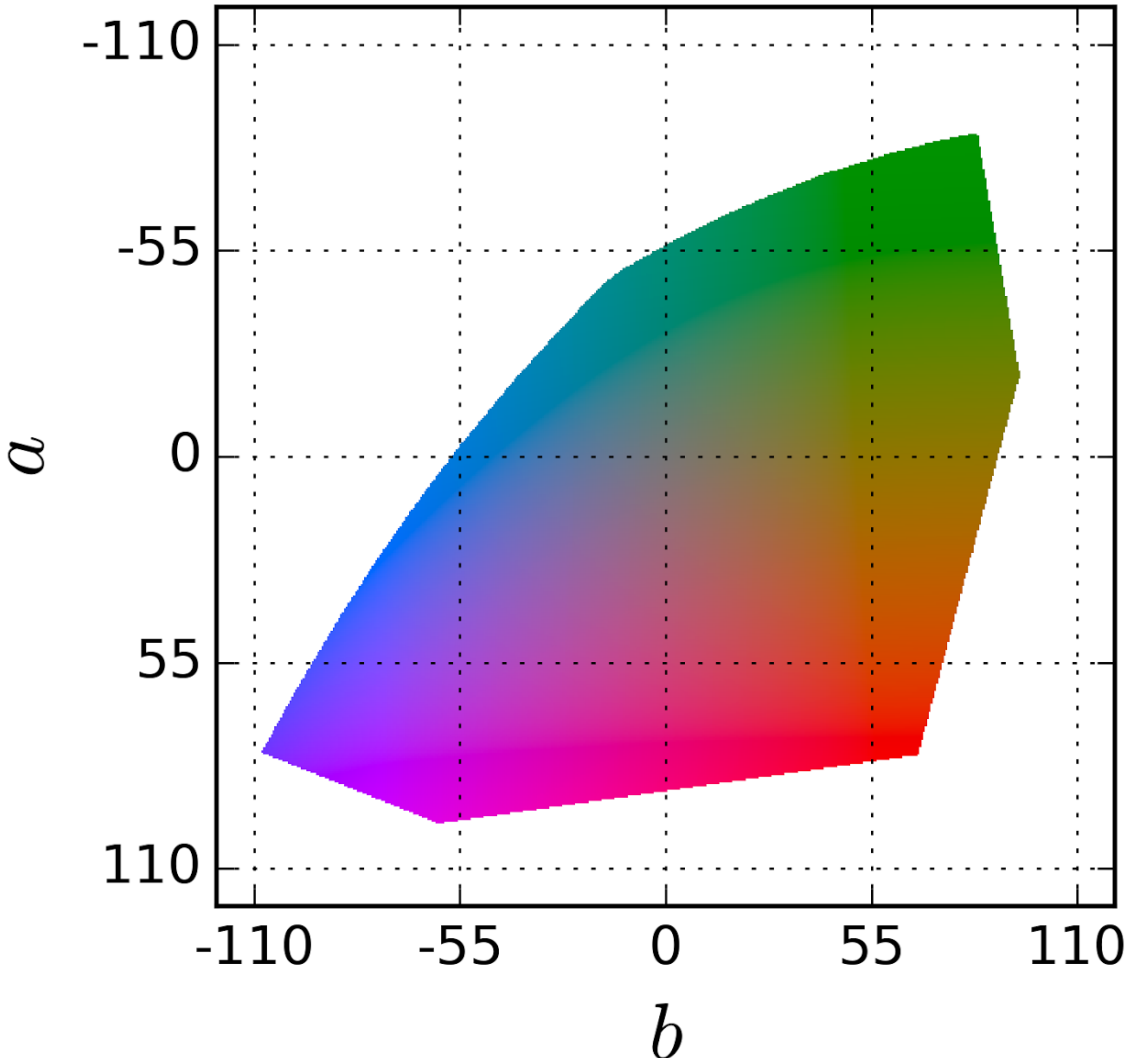
Output

Ground truth

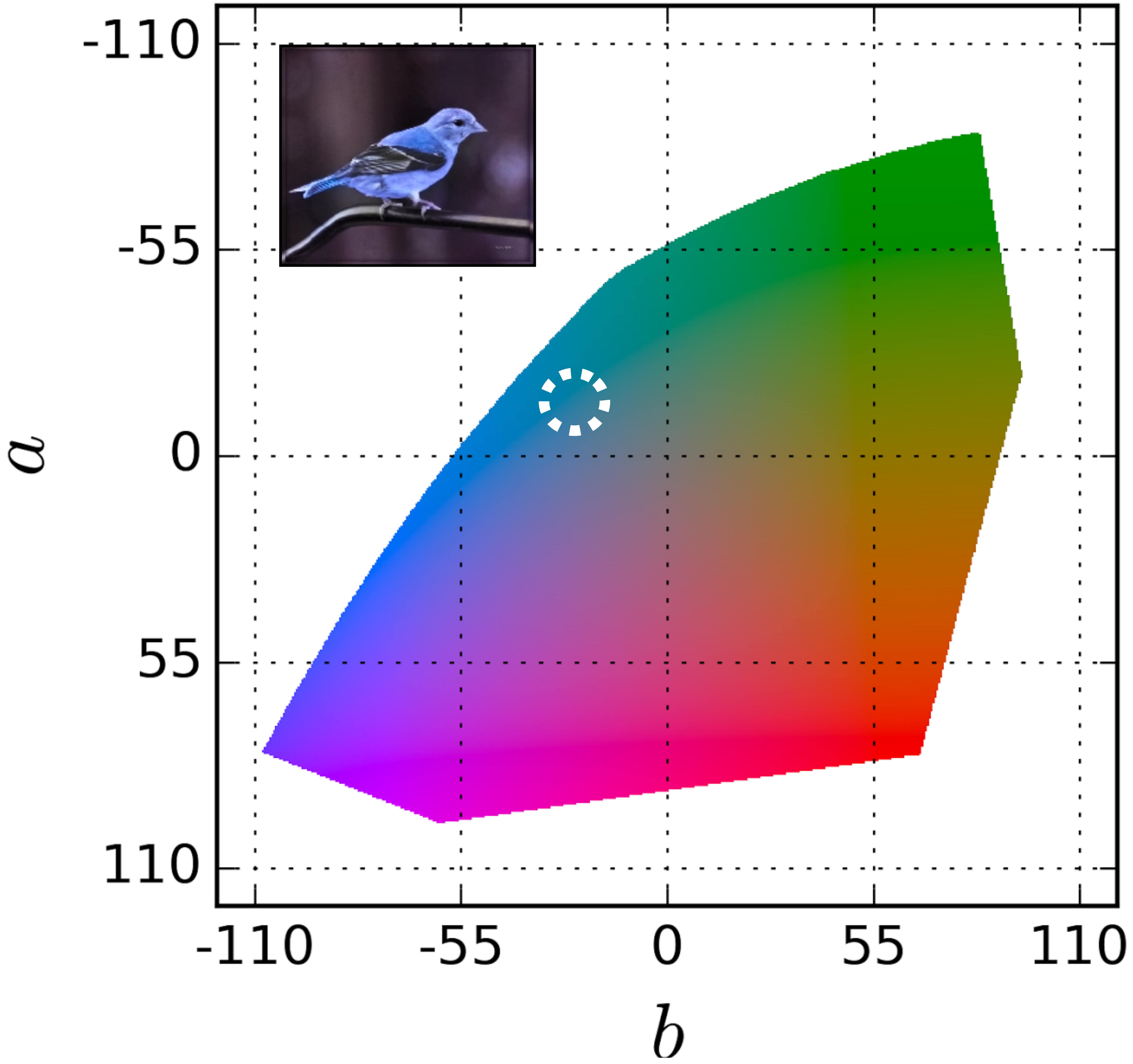
$$L_2(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2}\sum_{h,w} \|\mathbf{Y}_{h,w} - \widehat{\mathbf{Y}}_{h,w}\|_2^2$$



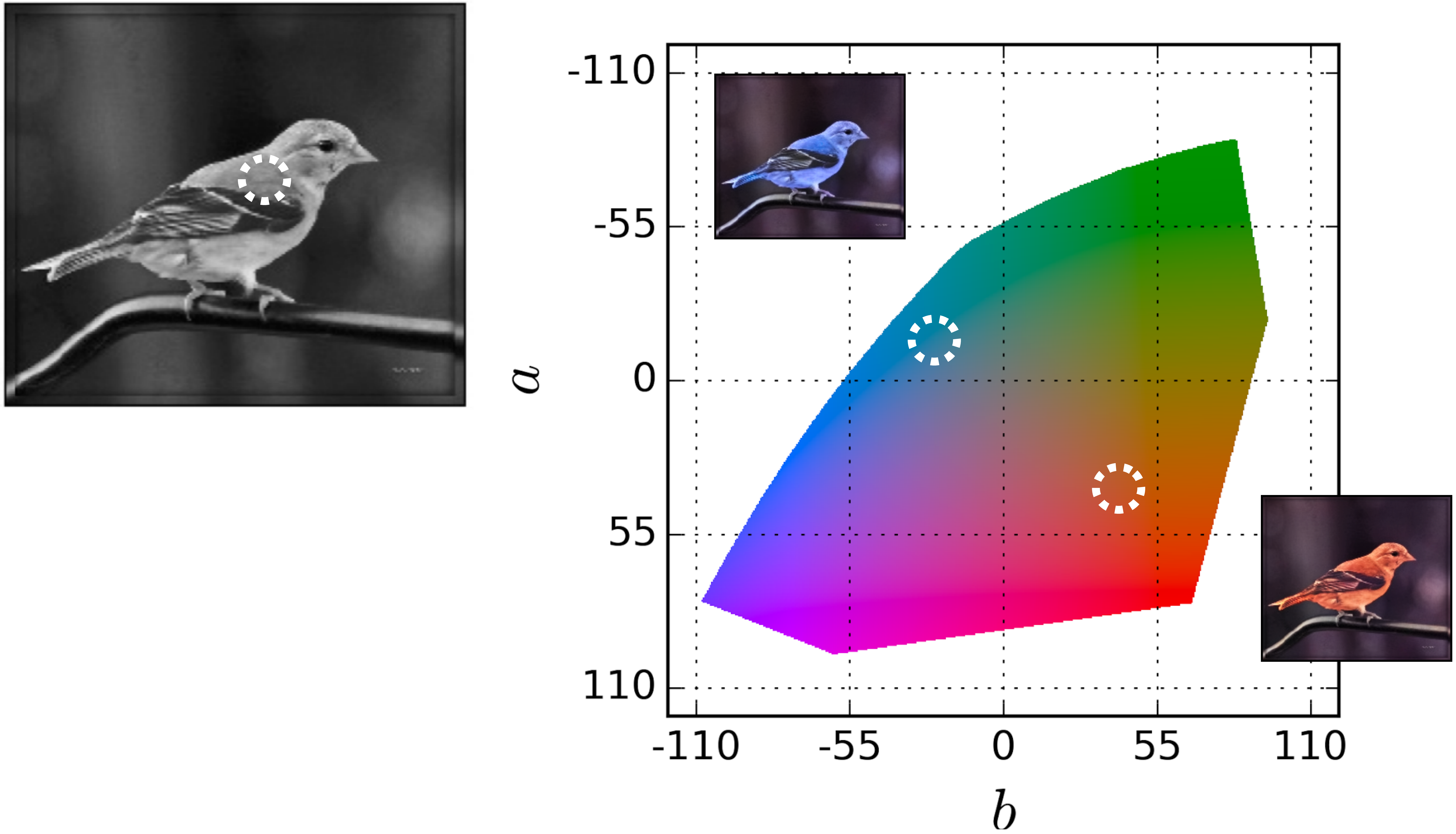
$$L_2(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \widehat{\mathbf{Y}}_{h,w}\|_2^2$$



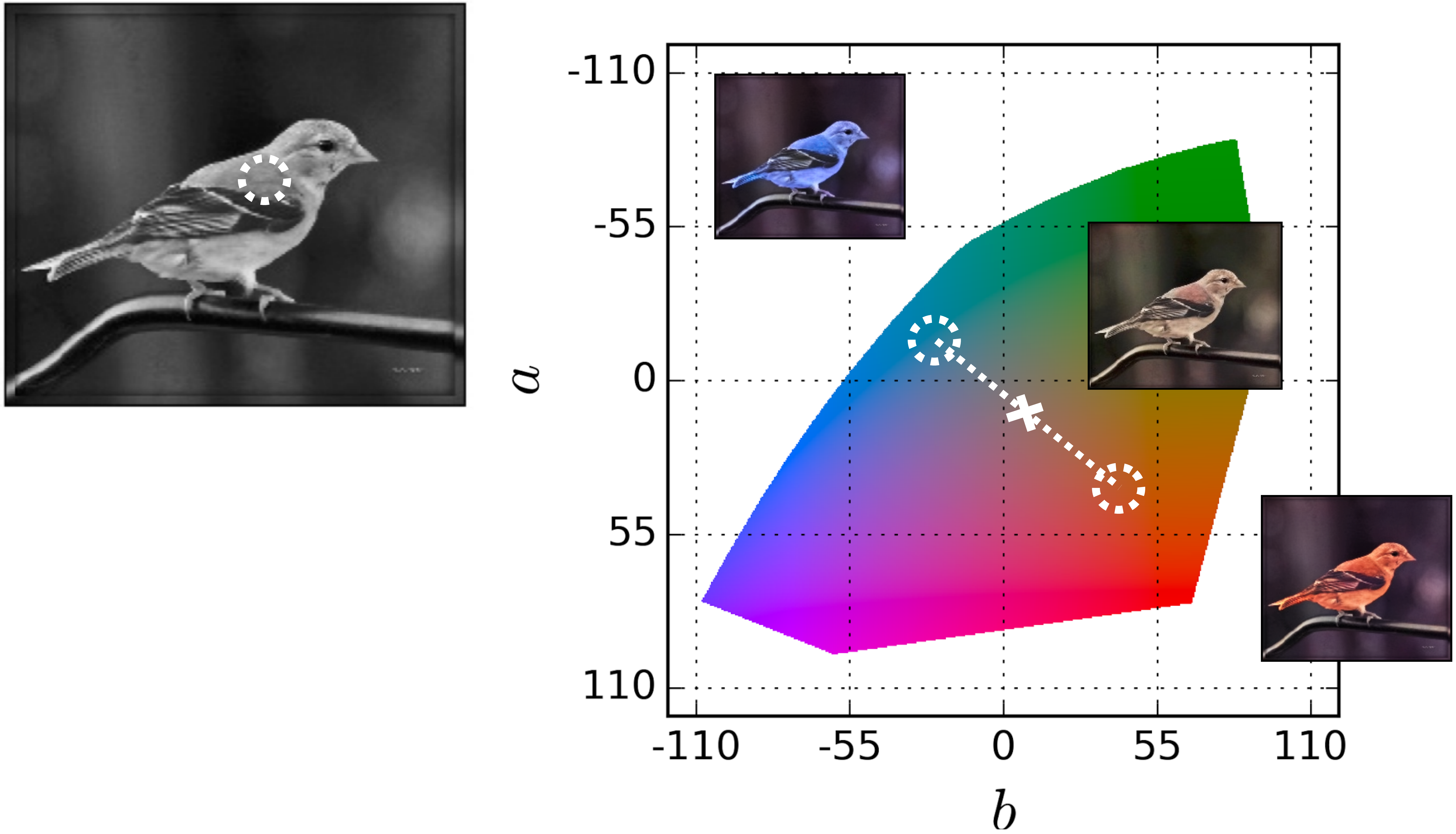
$$\mathrm{L}_2(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \widehat{\mathbf{Y}}_{h,w}\|_2^2$$



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



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



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

Designing loss functions

Input

Color distribution cross-entropy loss with colorfulness enhancing term.

Designing loss functions

Input

Zhang et al. 2016

Color distribution cross-entropy loss with colorfulness enhancing term.

Designing loss functions

Input

Zhang et al. 2016

Ground truth

Color distribution cross-entropy loss with colorfulness enhancing term.


DADS
DADS


DADS
DADS

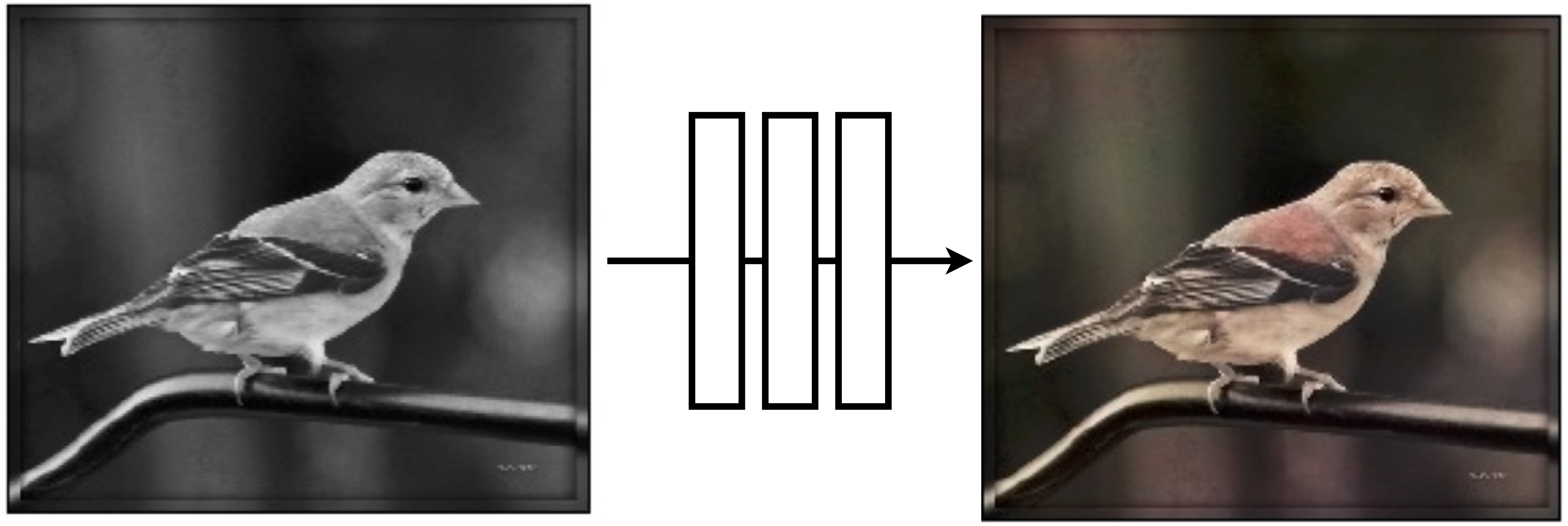
Designing loss functions

Designing loss functions

Be careful what you wish for!

Designing loss functions

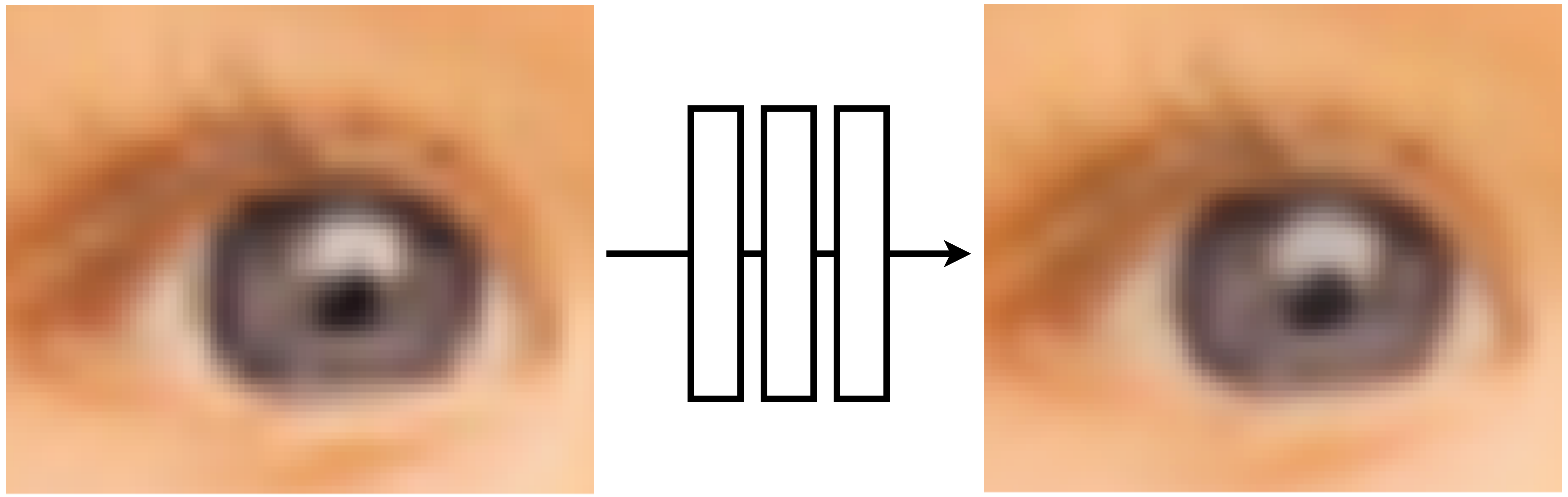
Image colorization



[Zhang, Isola, Efros, ECCV 2016]

L2 regression

Super-resolution

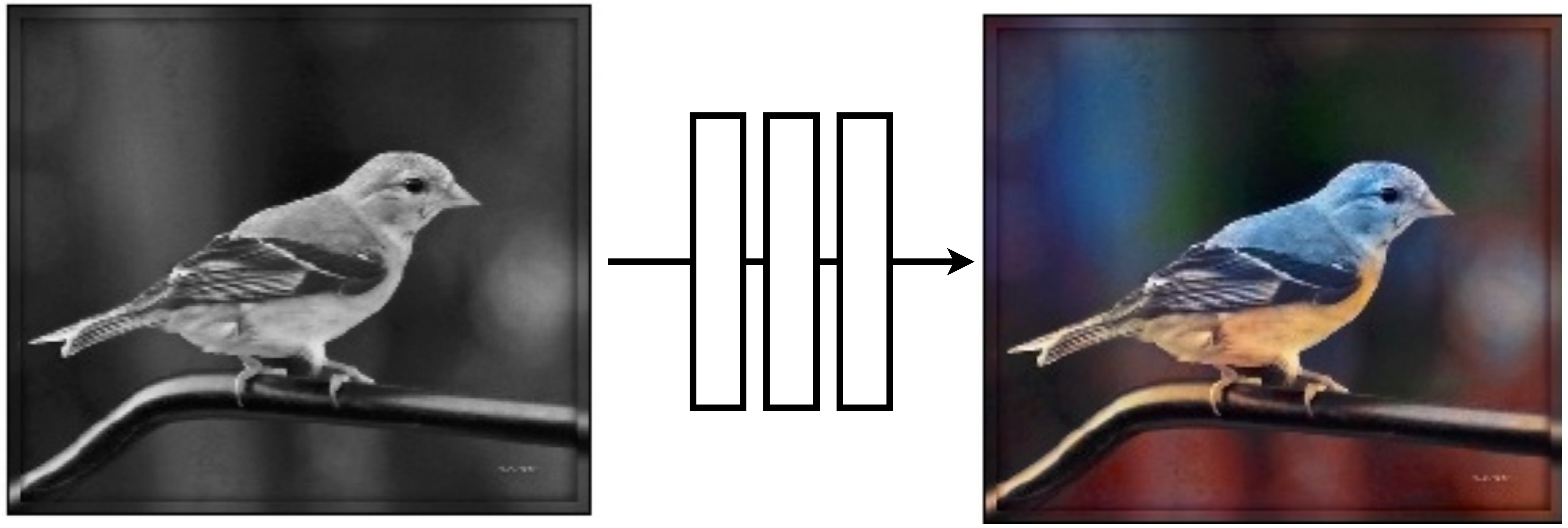


[Johnson, Alahi, Li, ECCV 2016]

L2 regression

Designing loss functions

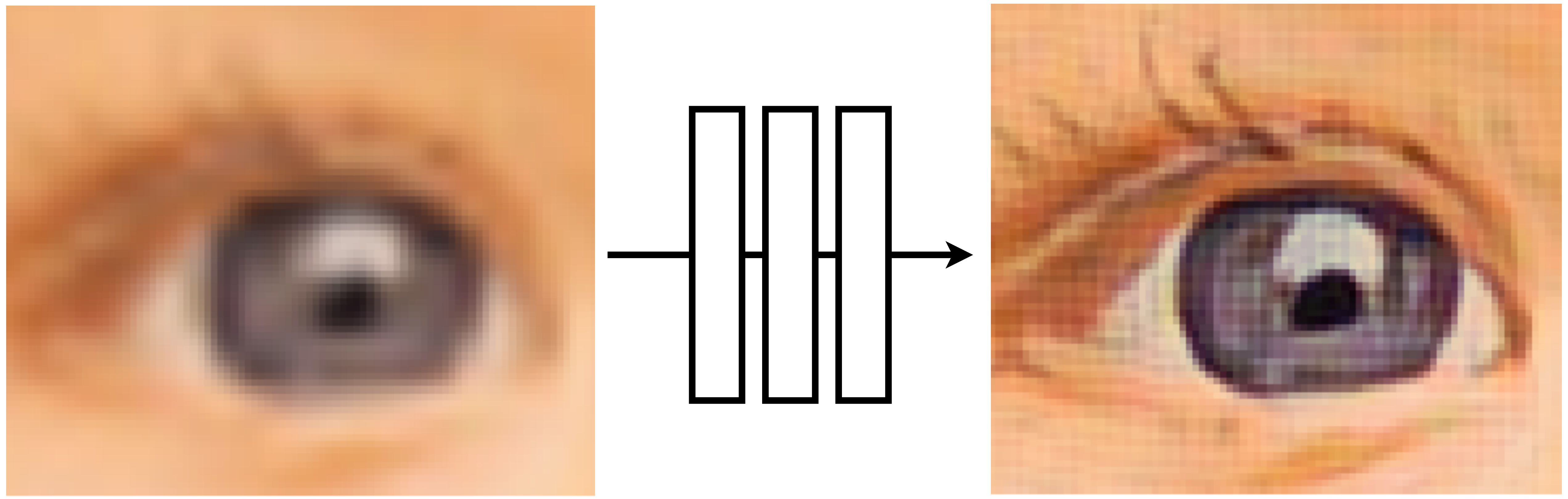
Image colorization



[Zhang, Isola, Efros, ECCV 2016]

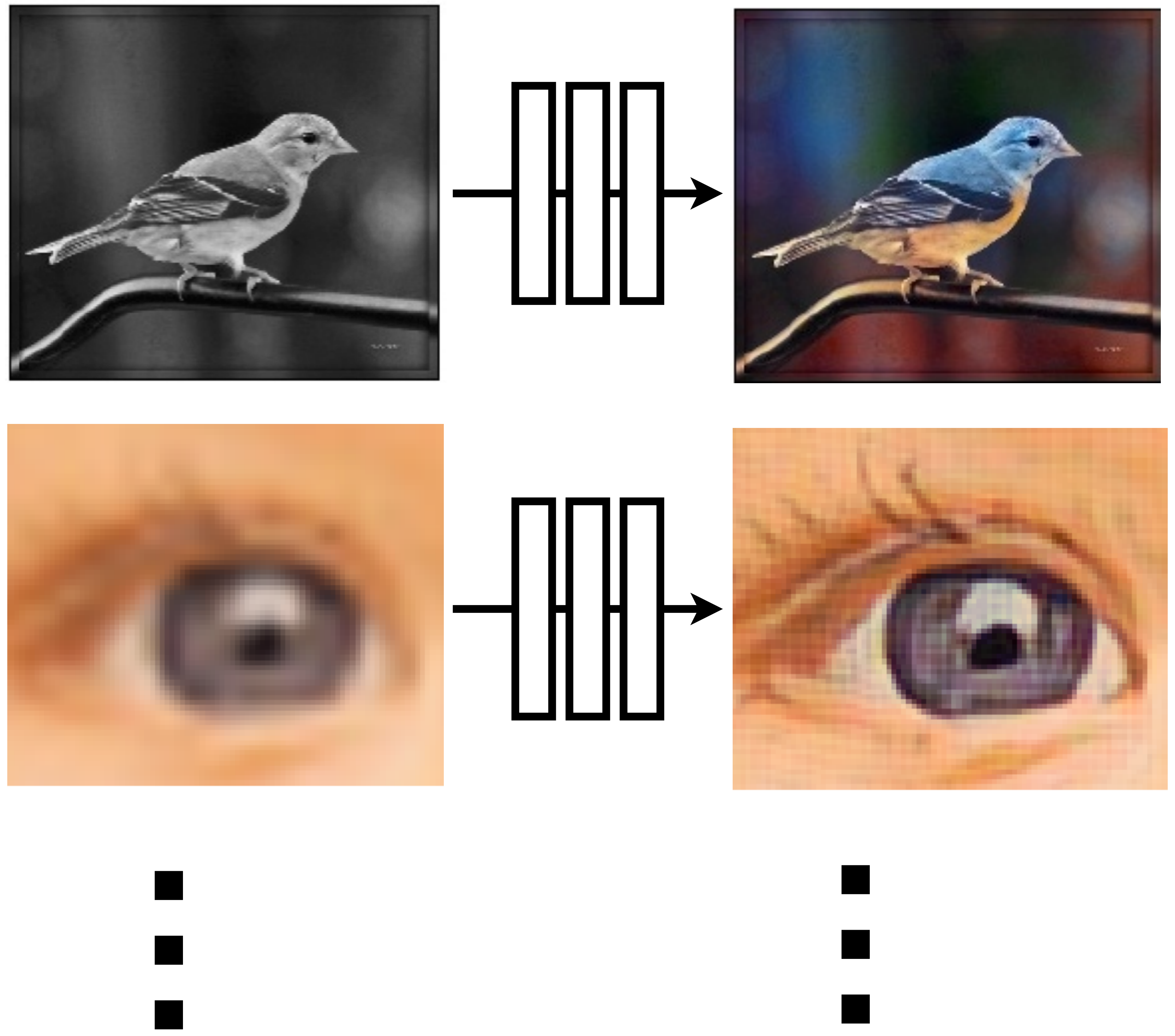
Cross entropy objective, with colorfulness term

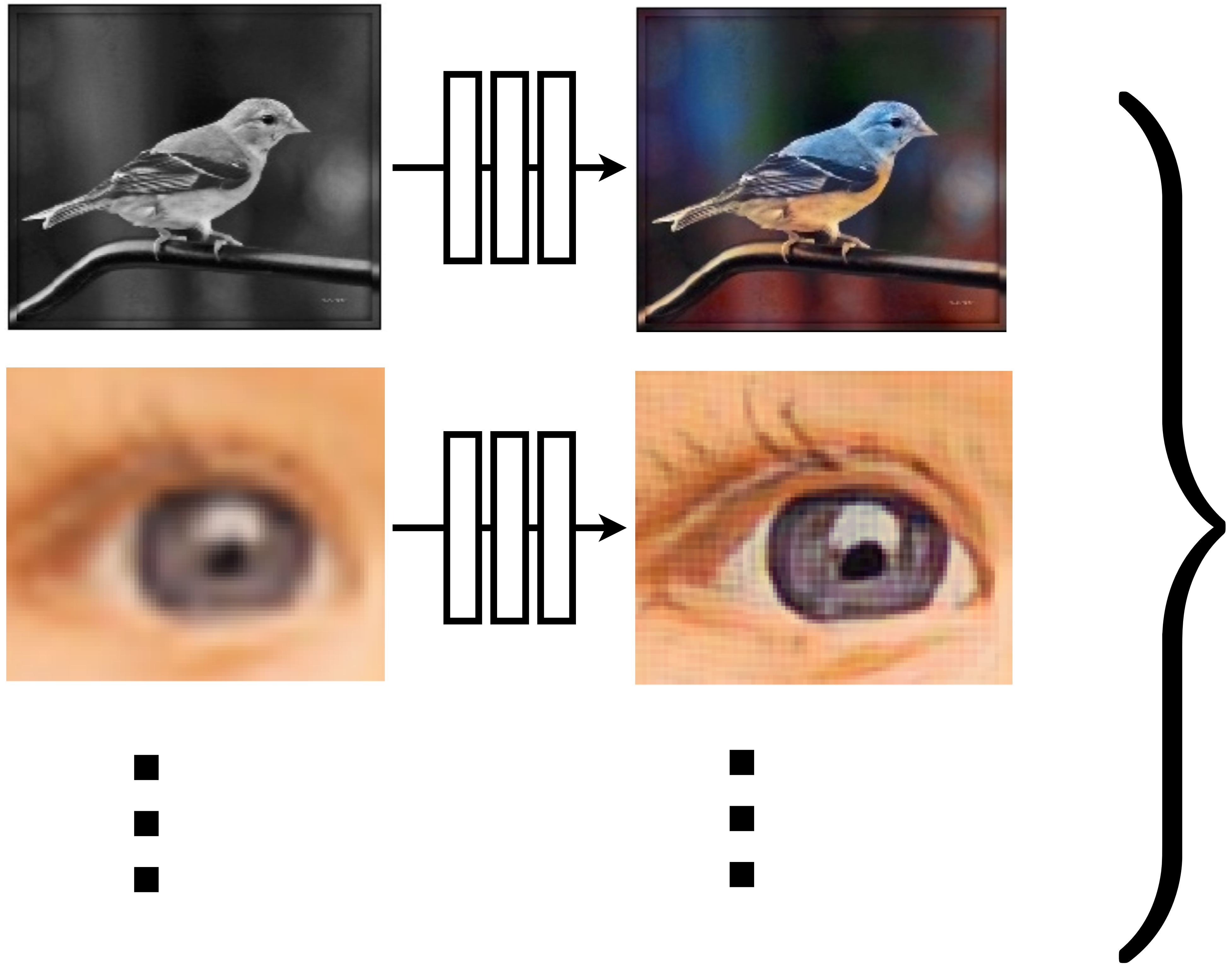
Super-resolution



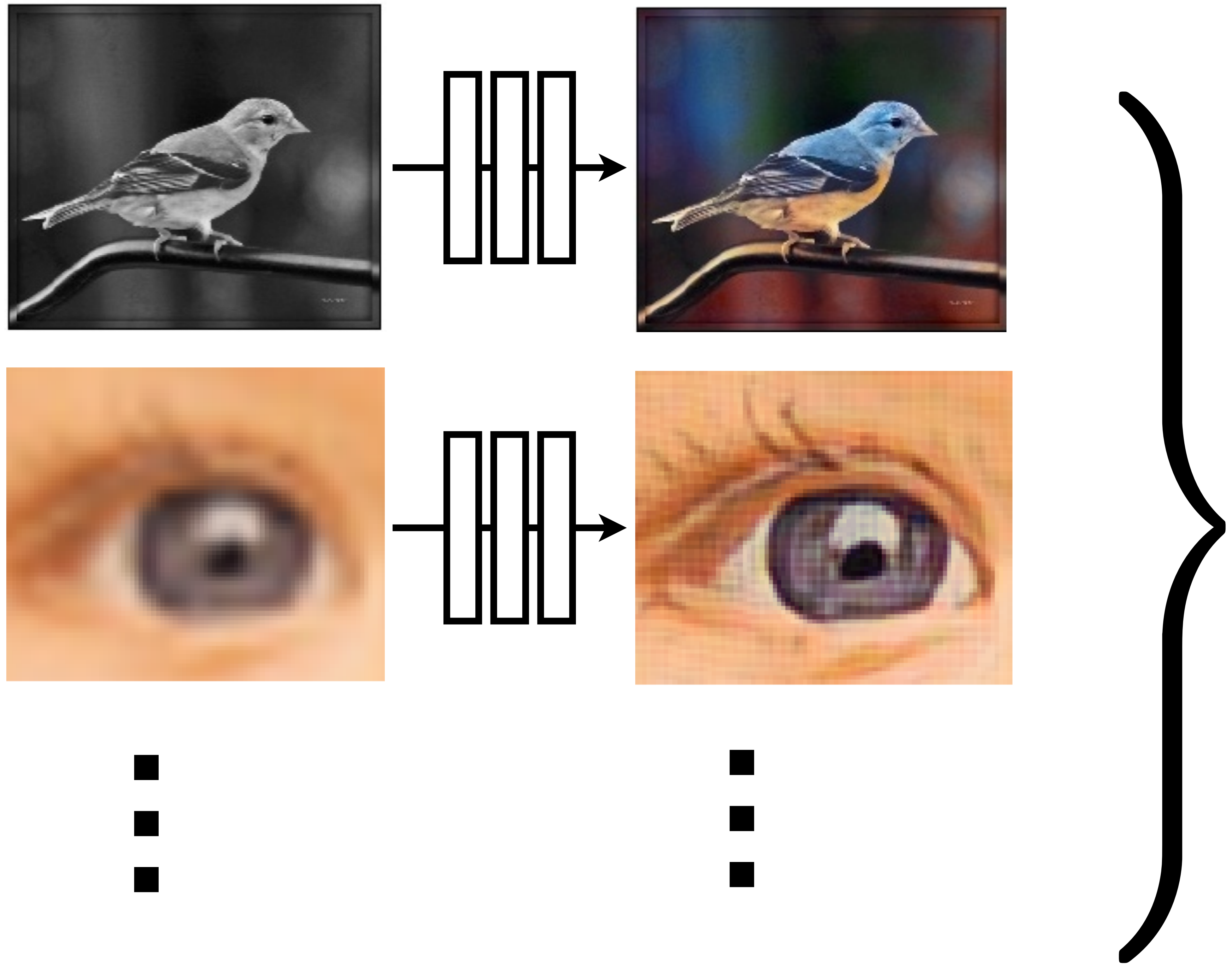
[Johnson, Alahi, Li, ECCV 2016]

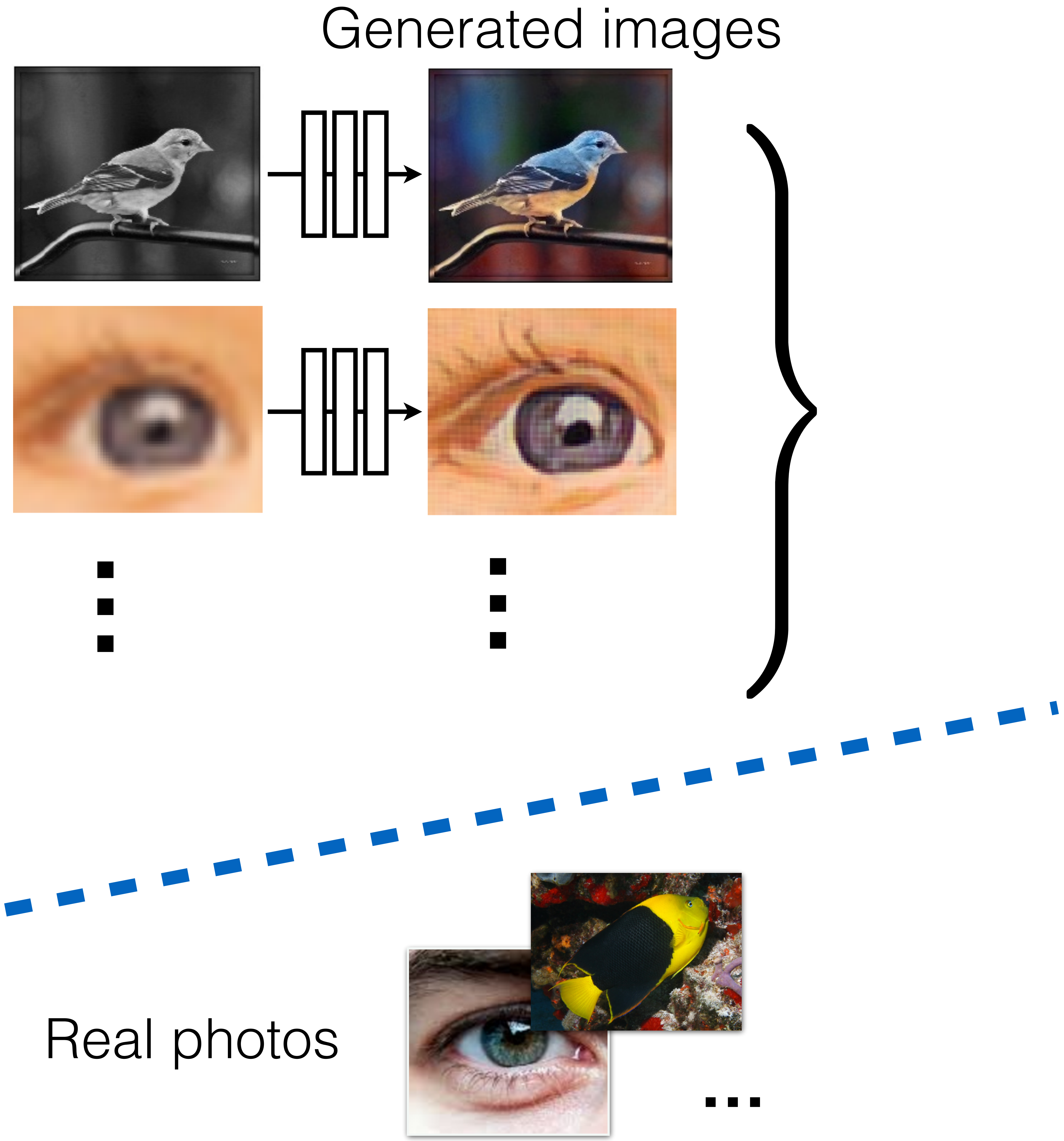
Deep feature covariance matching objective



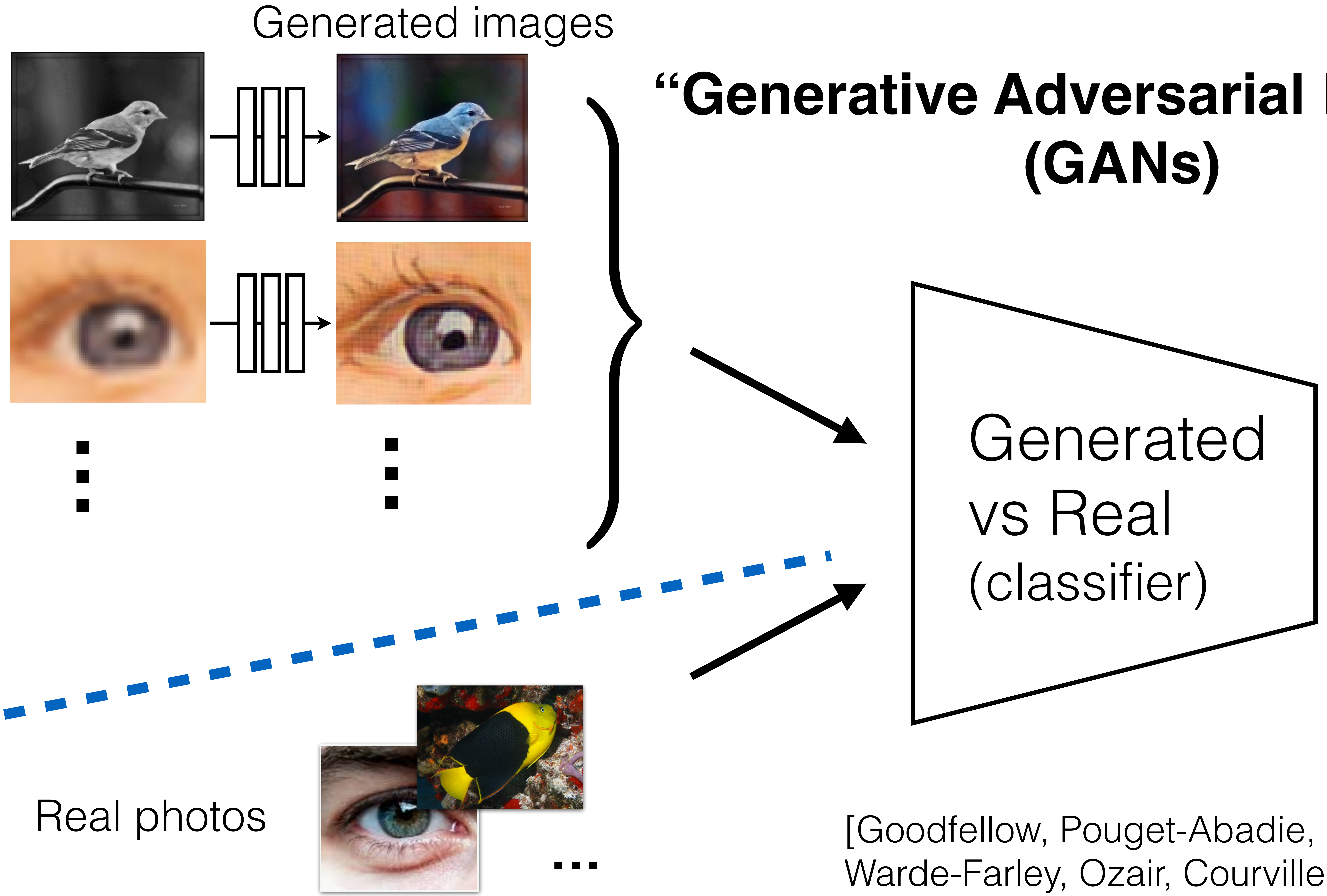


Universal loss?




Generated images
Real photos
...

“Generative Adversarial Network” (GANs)



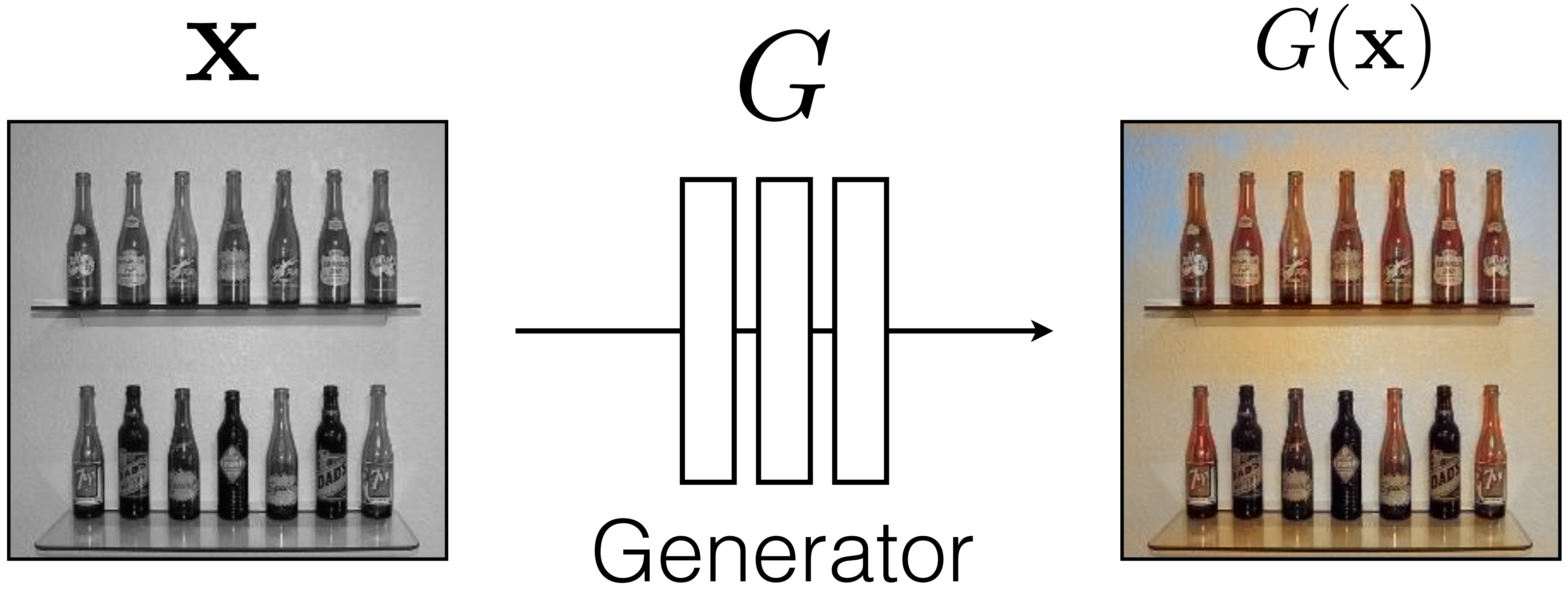
[Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, Bengio 2014]

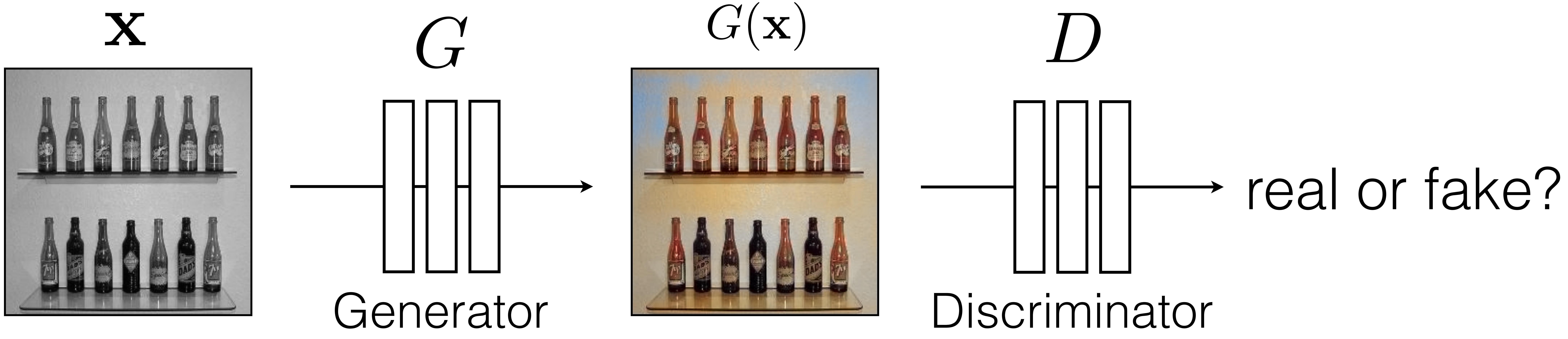
Conditional GANs

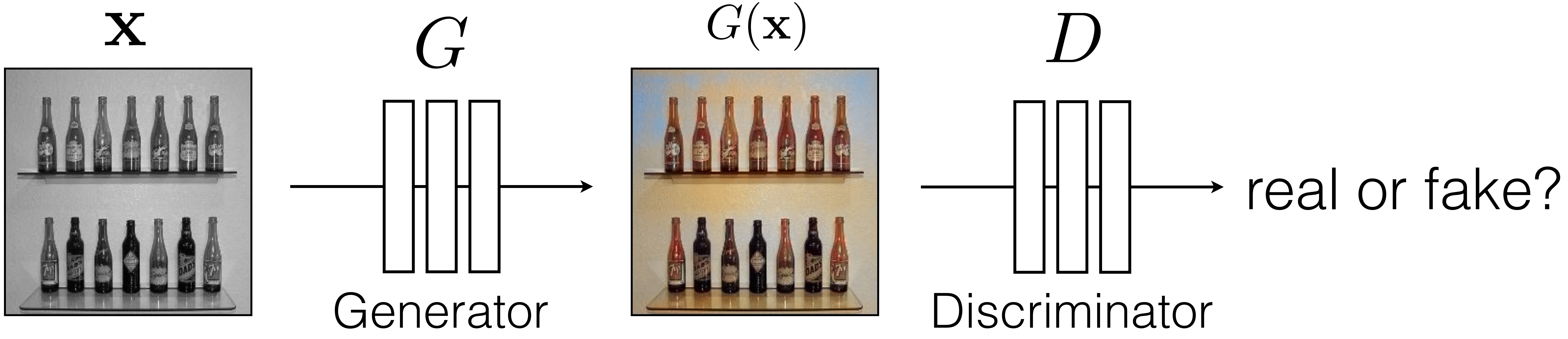
[Mirza et al. 2014] [Reed et al. 2016]

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

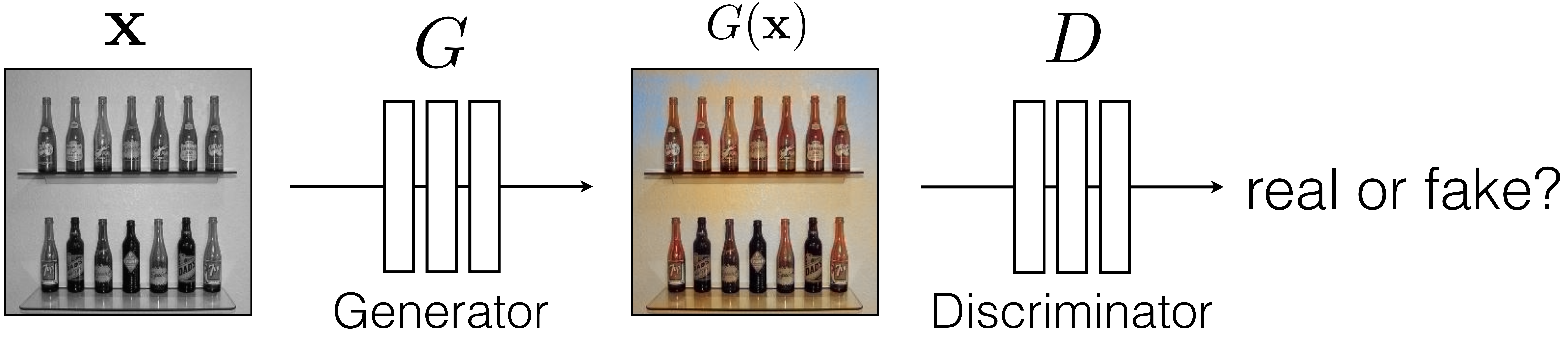
[…]


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


$\mathbf{x}$
G
Generator
$G(\mathbf{x})$
D
Discriminator
real or fake?

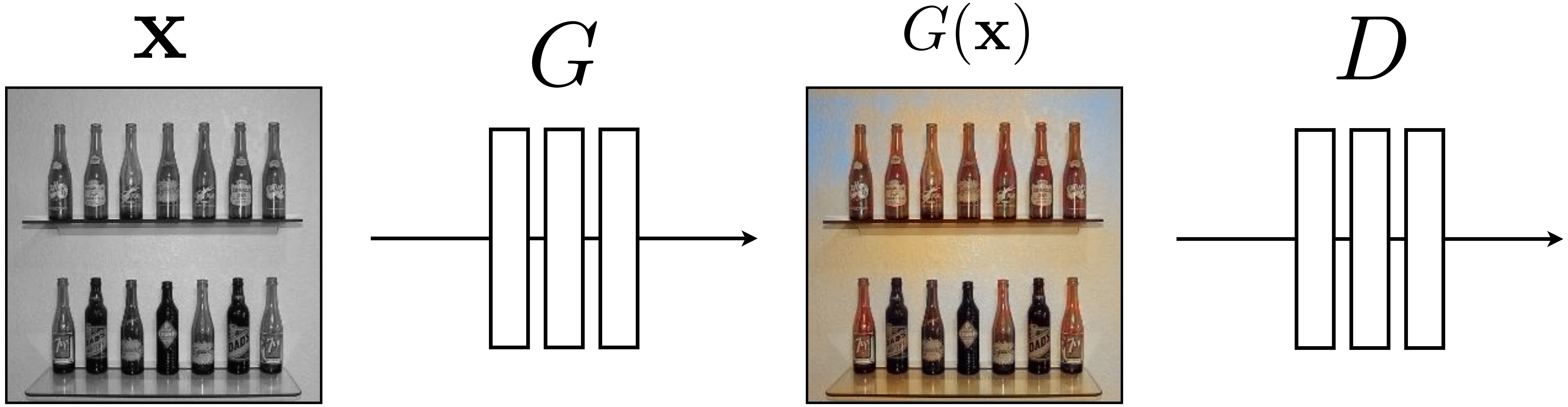


G tries to synthesize fake images that fool **D**



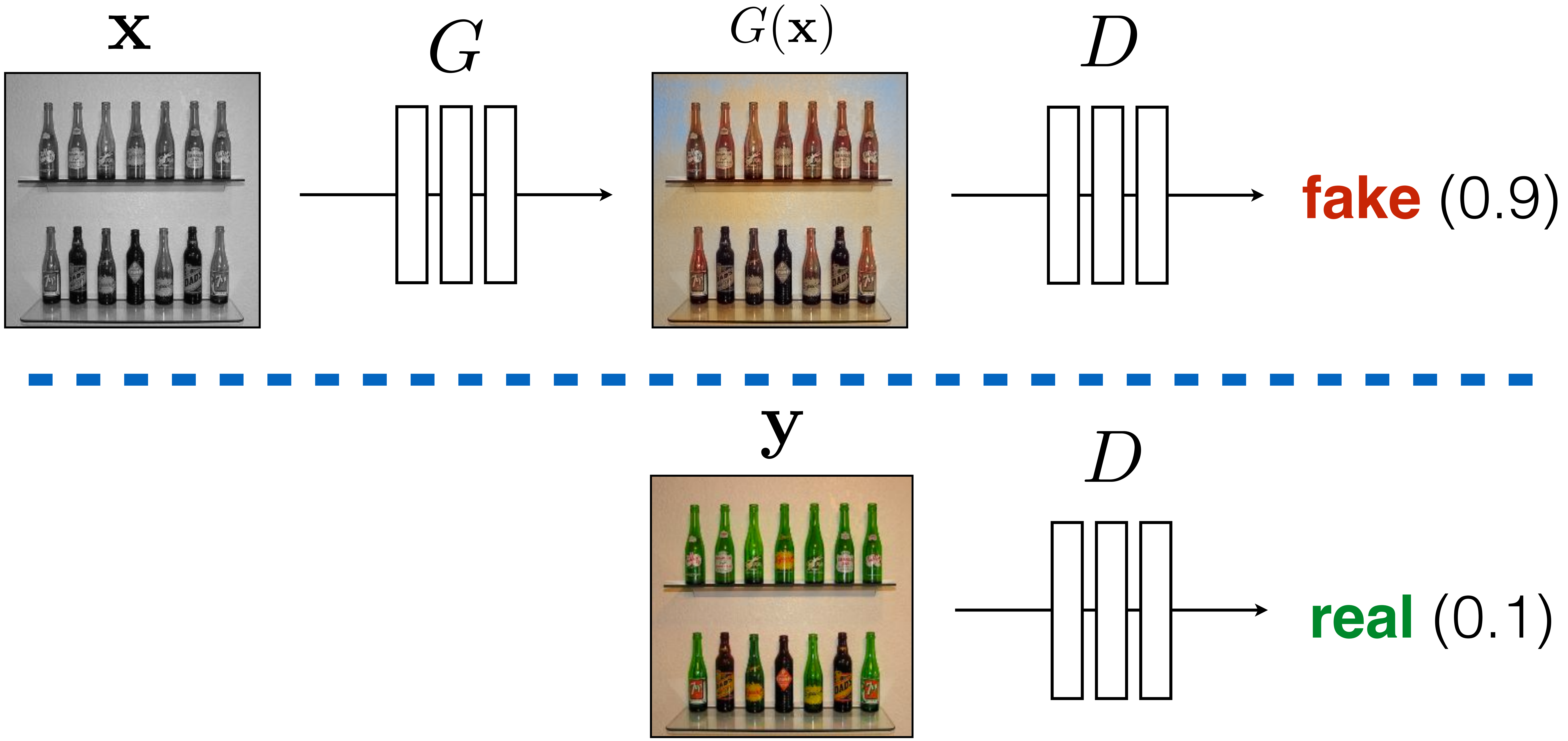
G tries to synthesize fake images that fool **D**

D tries to identify the fakes

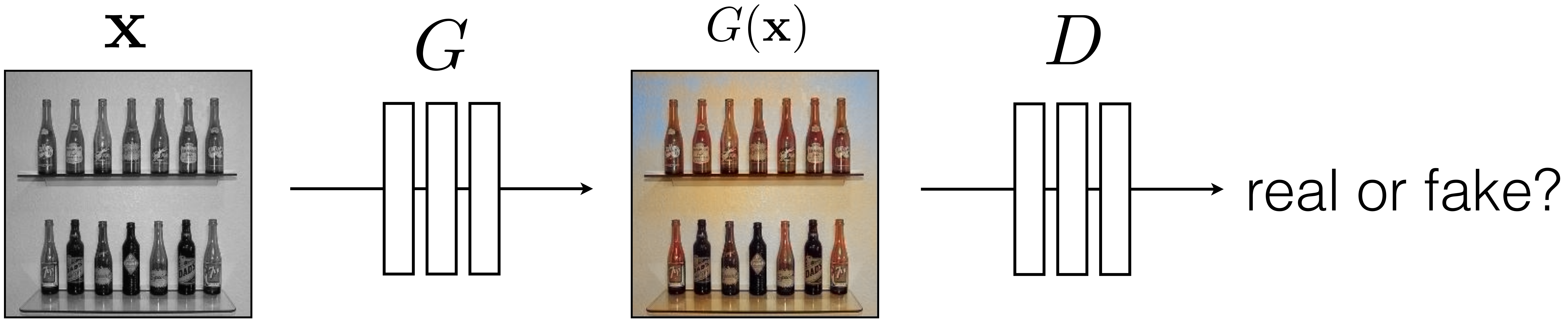


$$\arg\max_{D} \; \mathbb{E}_{\mathbf{x},\mathbf{y}}[\qquad\qquad\qquad\qquad\qquad\qquad]$$

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

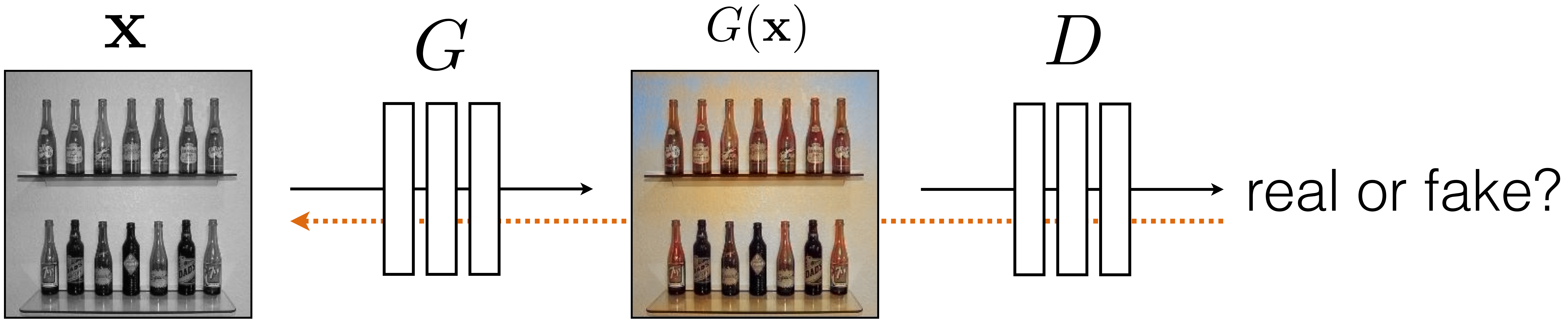


$$\arg\max_{D}\ \mathbb{E}_{\mathbf{x},\mathbf{y}}[\ \boxed{\log D(G(\mathbf{x}))} + \boxed{\log(1 - D(\mathbf{y}))}\ ]$$



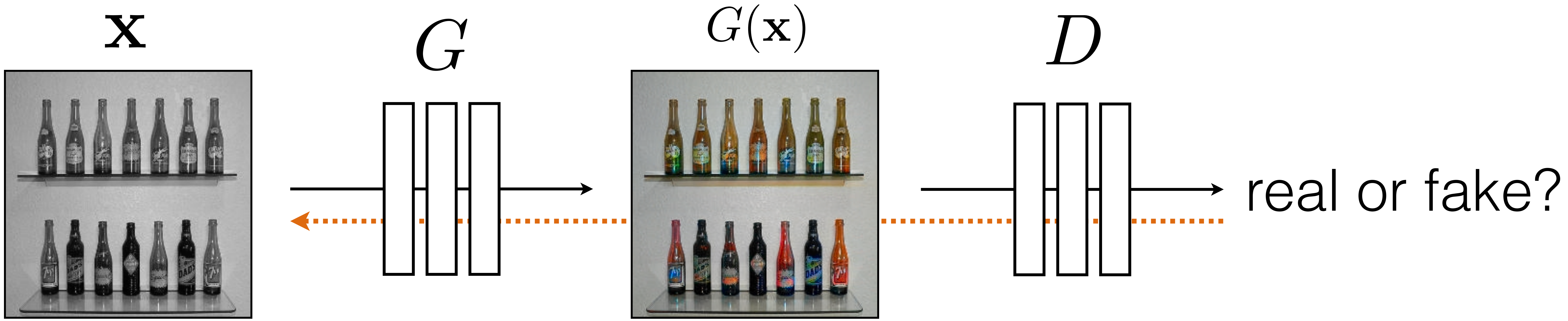
G tries to synthesize fake images that ***fool*** **D**:

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



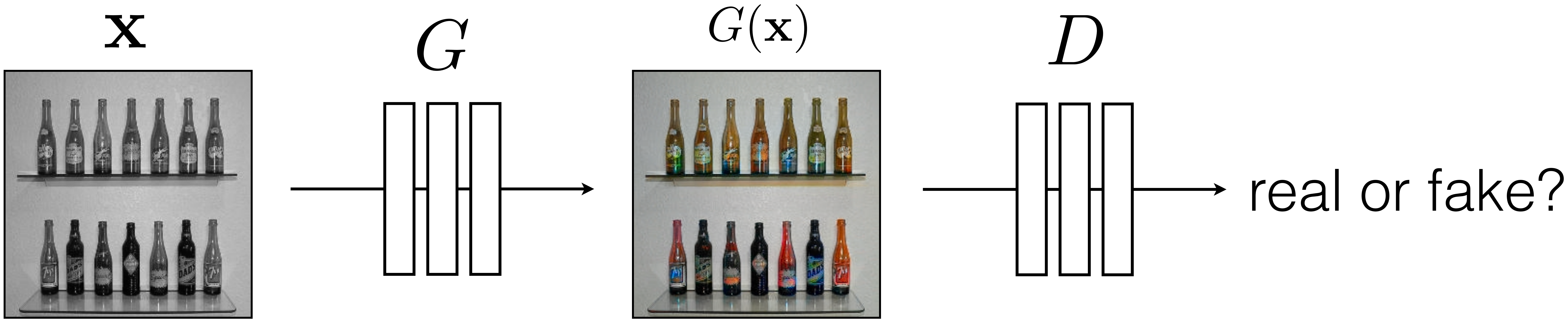
G tries to synthesize fake images that ***fool*** **D**:

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



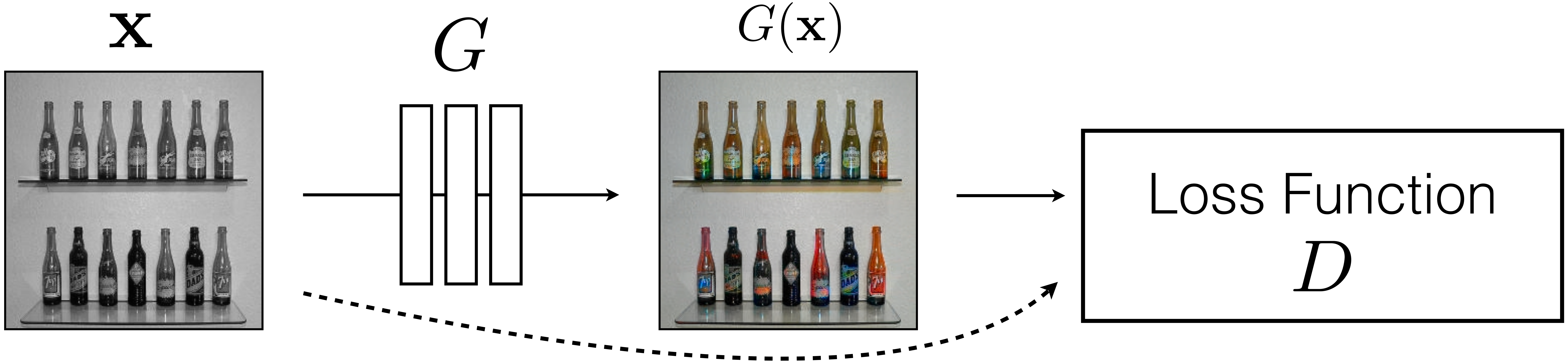
G tries to synthesize fake images that ***fool*** **D**:

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



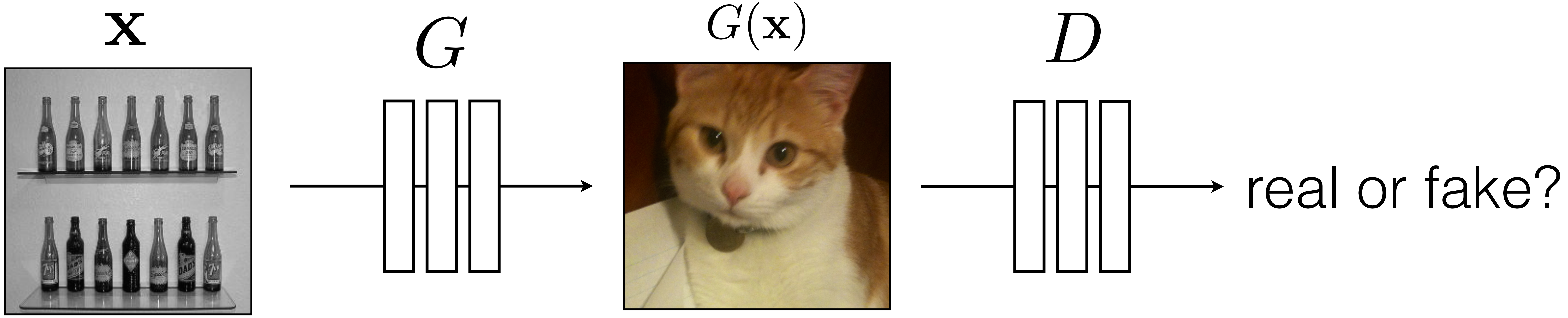
G tries to synthesize fake images that ***fool*** the ***best*** **D**:

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

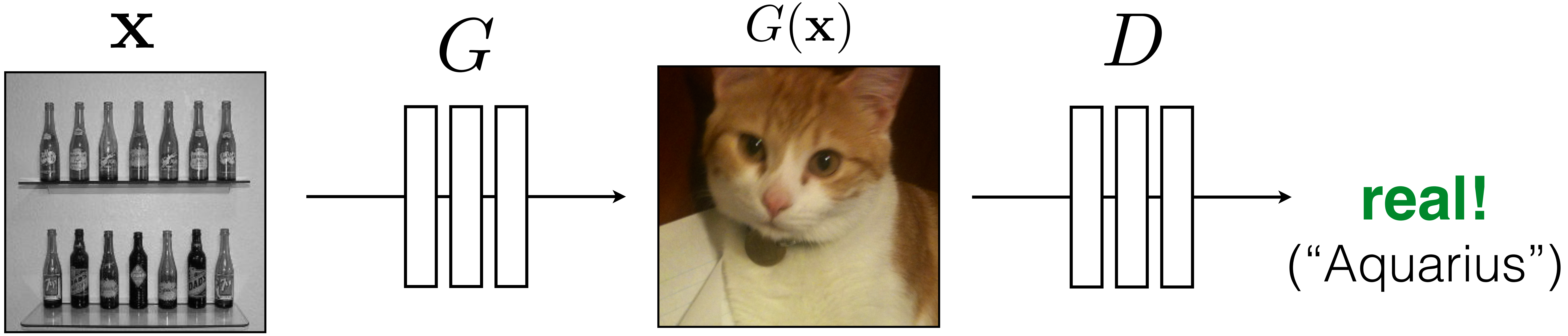


G's perspective: **D** is a loss function.

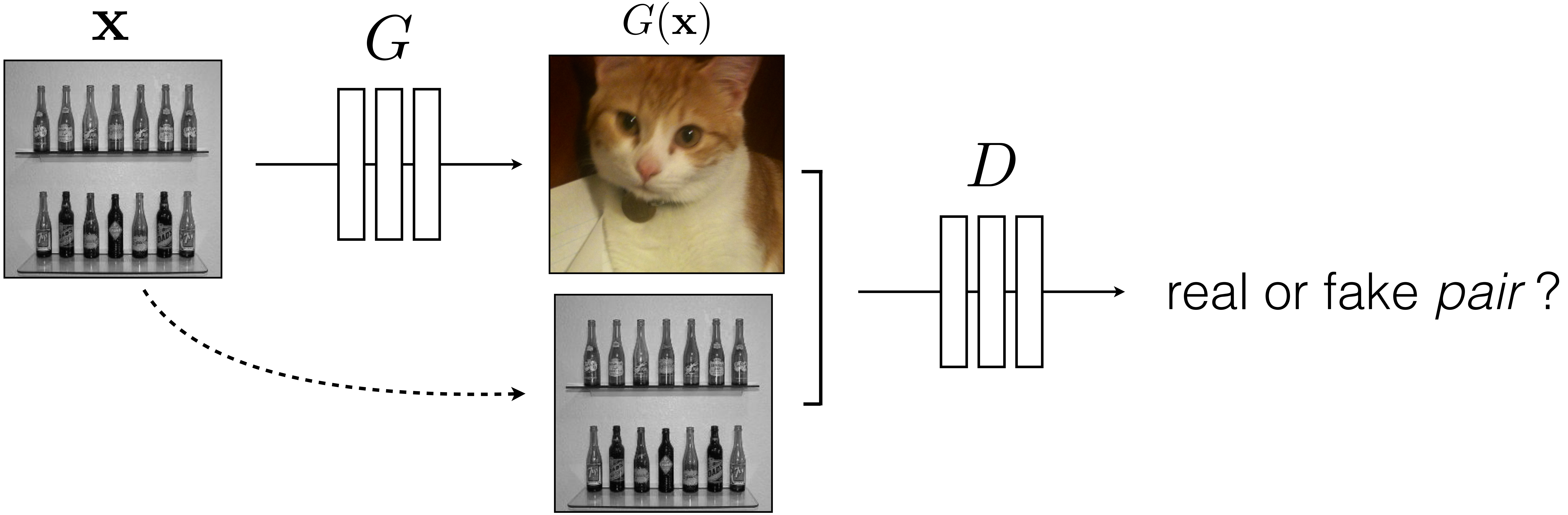
Rather than being hand-designed, it is *learned*.



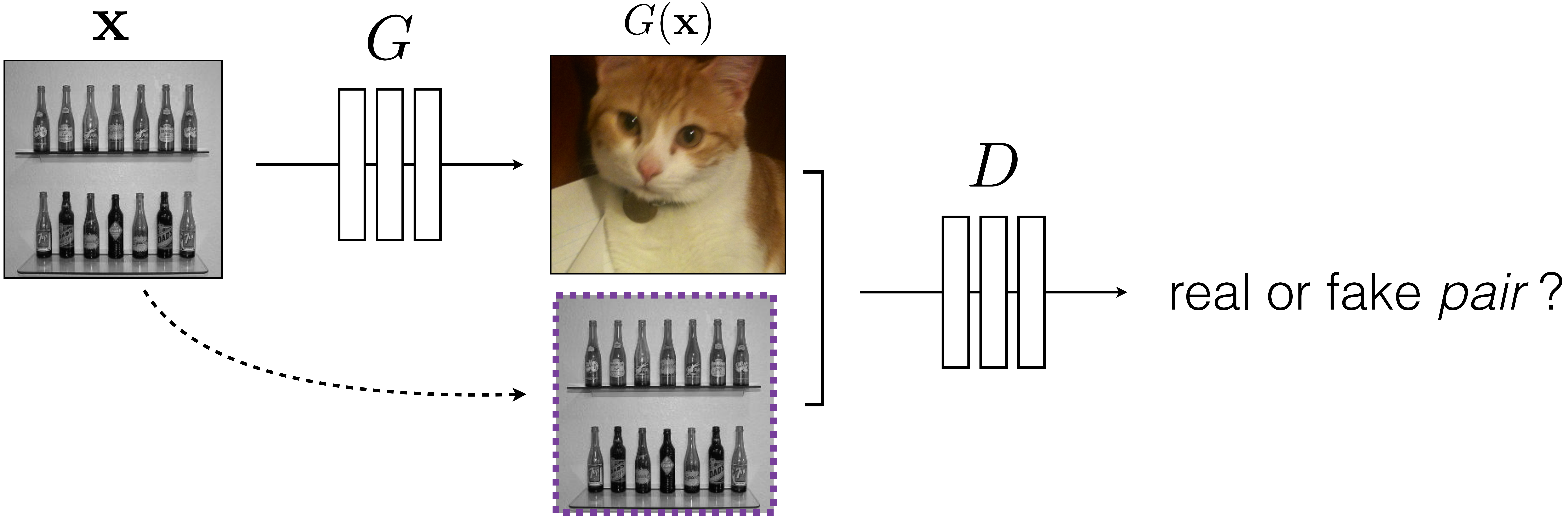
$$\arg\min_{G}\max_{D}\ \mathbb{E}_{\mathbf{x},\mathbf{y}}[\ \log D(G(\mathbf{x})) \quad + \quad \log(1 - D(\mathbf{y}))\ \ ]$$



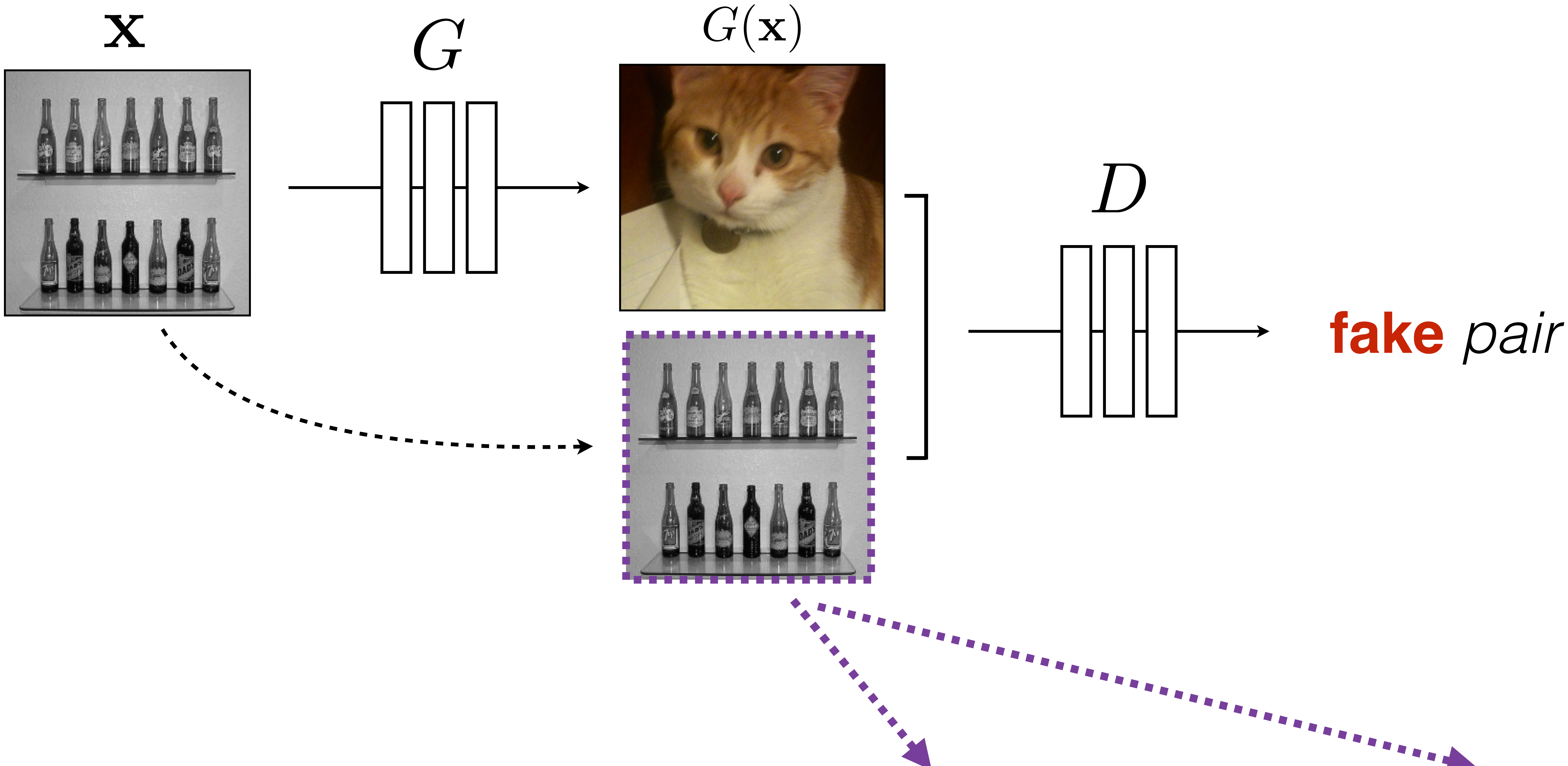
$$\arg\min_{G}\max_{D}\ \mathbb{E}_{\mathbf{x},\mathbf{y}}[\ \log D(G(\mathbf{x})) \quad + \quad \log(1 - D(\mathbf{y}))\ \ ]$$



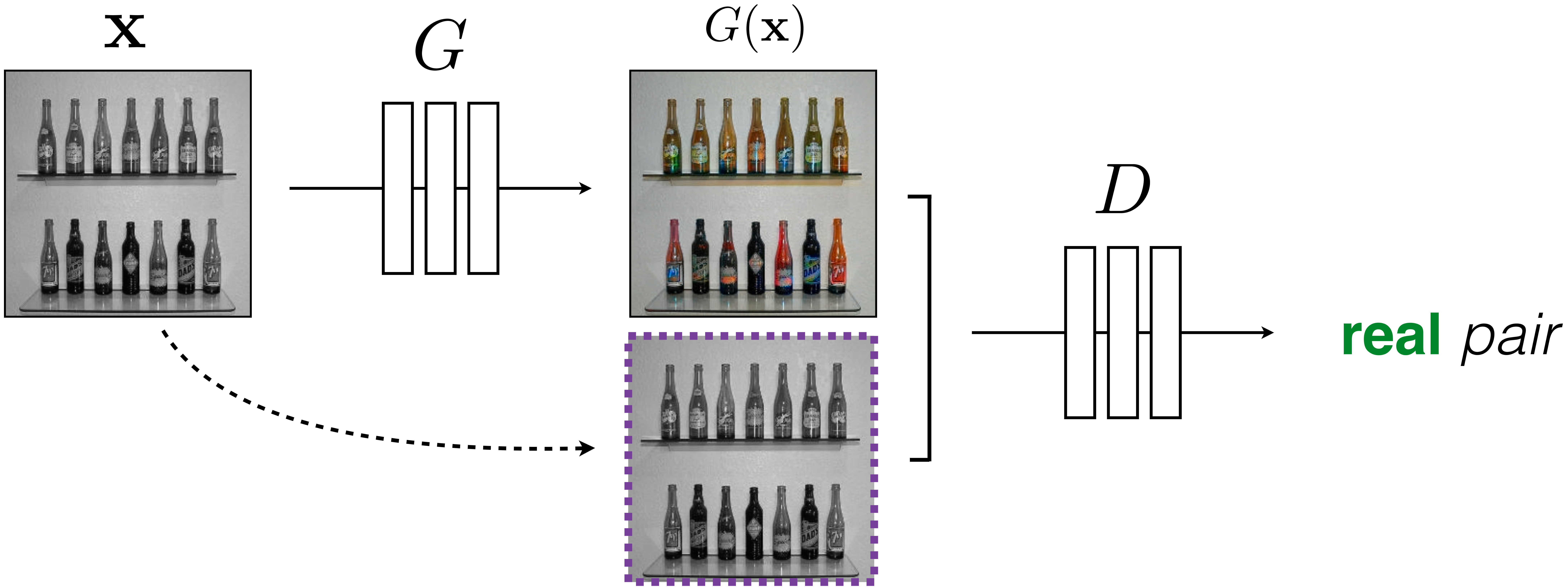
$$\arg\min_{G}\max_{D}\ \mathbb{E}_{\mathbf{x},\mathbf{y}}[\ \log D(G(\mathbf{x})) \quad + \quad \log(1 - D(\mathbf{y}))\ ]$$



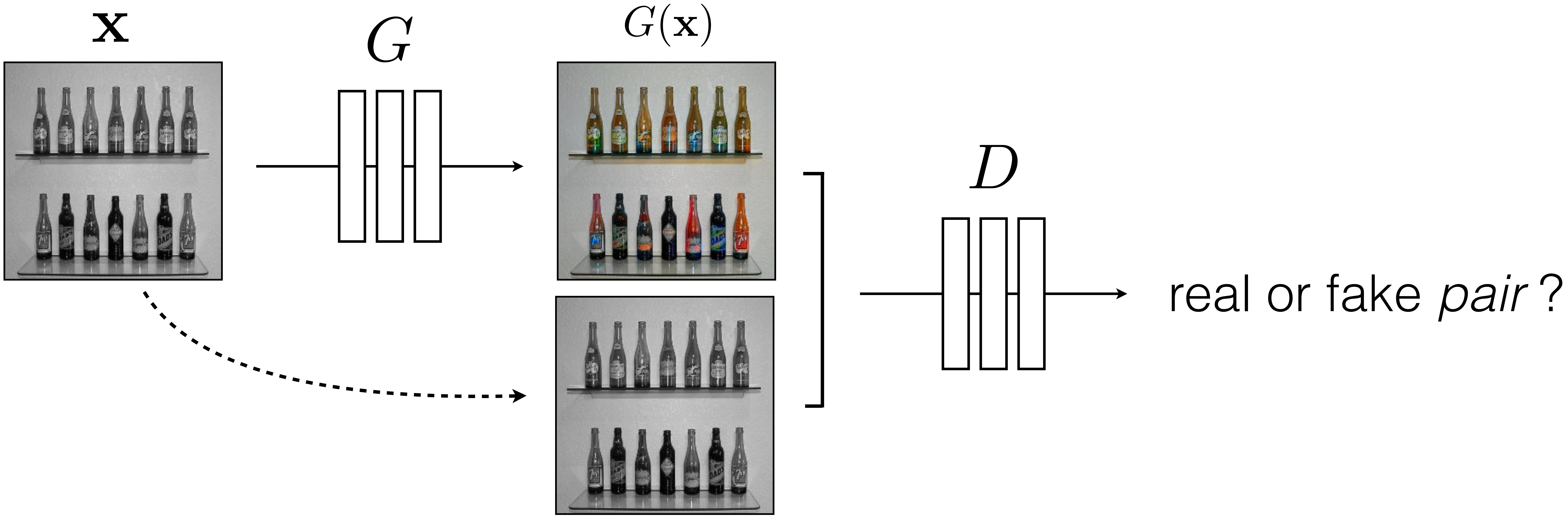
$$\arg\min_{G}\max_{D}\;\mathbb{E}_{\mathbf{x},\mathbf{y}}\big[\;\log D(\boxed{\mathbf{x}}, G(\mathbf{x})) + \log(1 - D(\boxed{\mathbf{x}}, \mathbf{y}))\;\big]$$



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



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



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

Training Details: Loss function

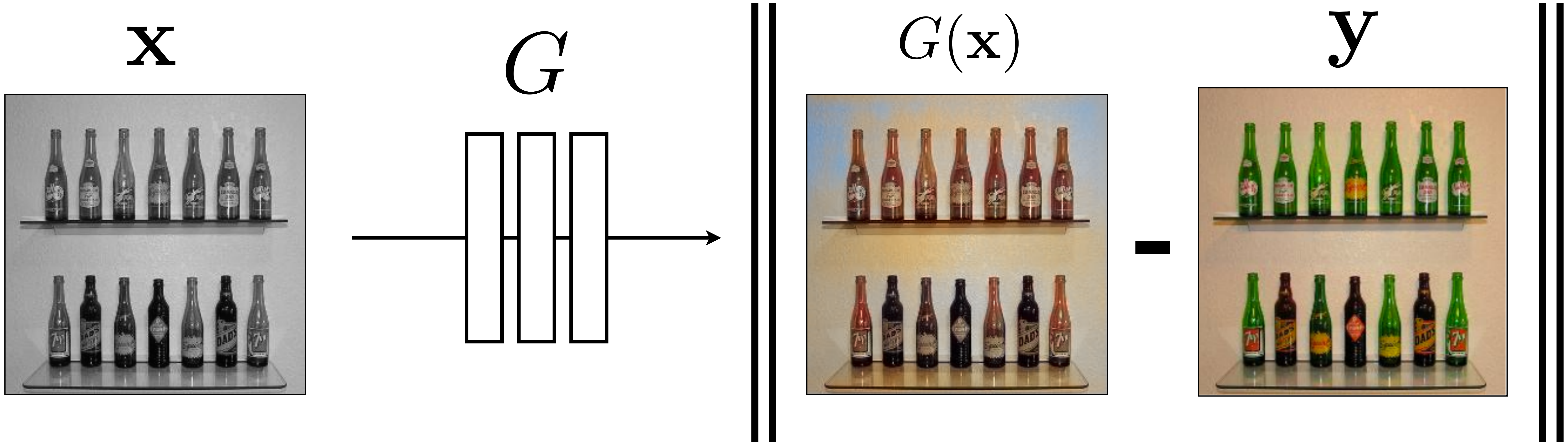
Conditional GAN

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

Training Details: Loss function

Conditional GAN

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

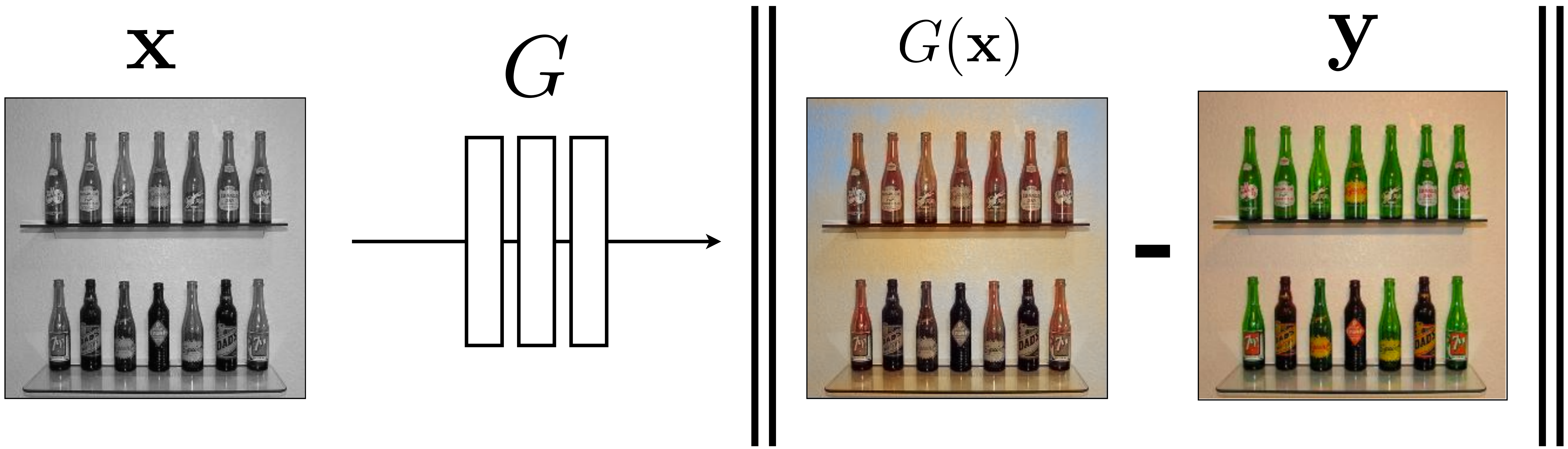


[c.f. Pathak et al. CVPR 2016]

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]

BW → Color

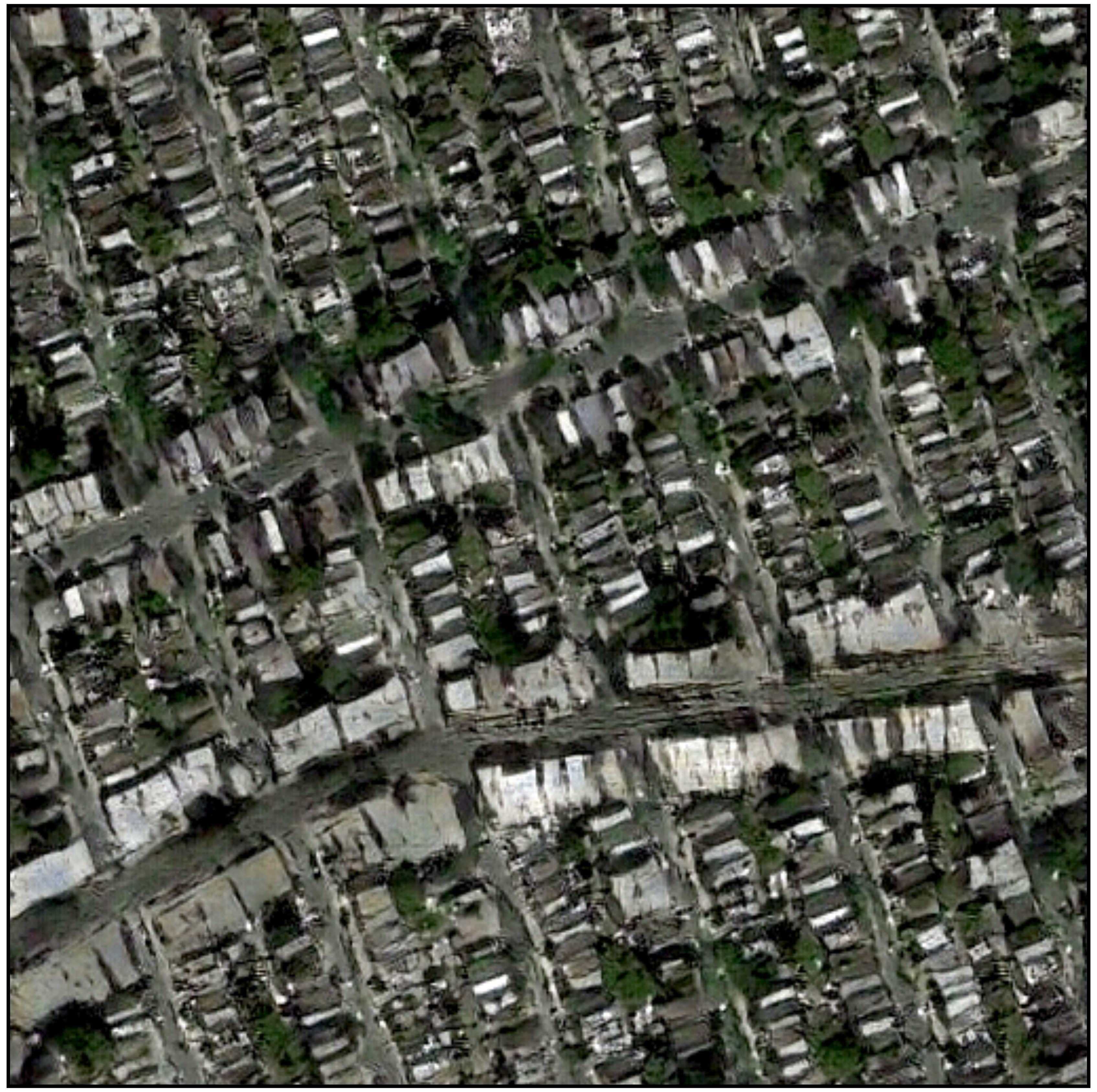
Data from [Russakovsky et al. 2015]

Input

Data from
[maps.google.com]

Input

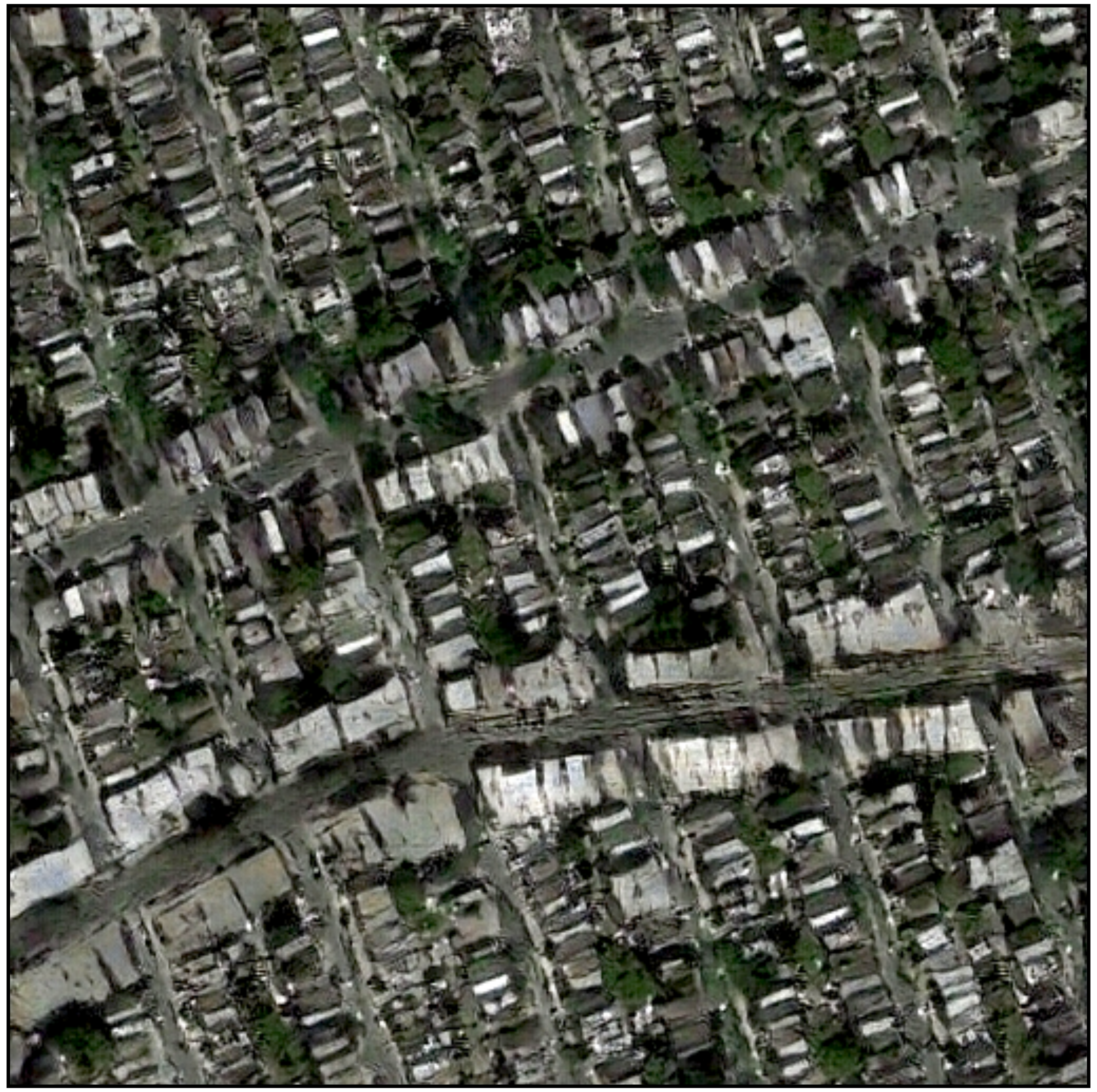
Output



Data from
[maps.google.com]

Input

Output



Groundtruth

Data from [maps.google.com]

Input

Output

Groundtruth

Data from [maps.google.com]

Input

Output

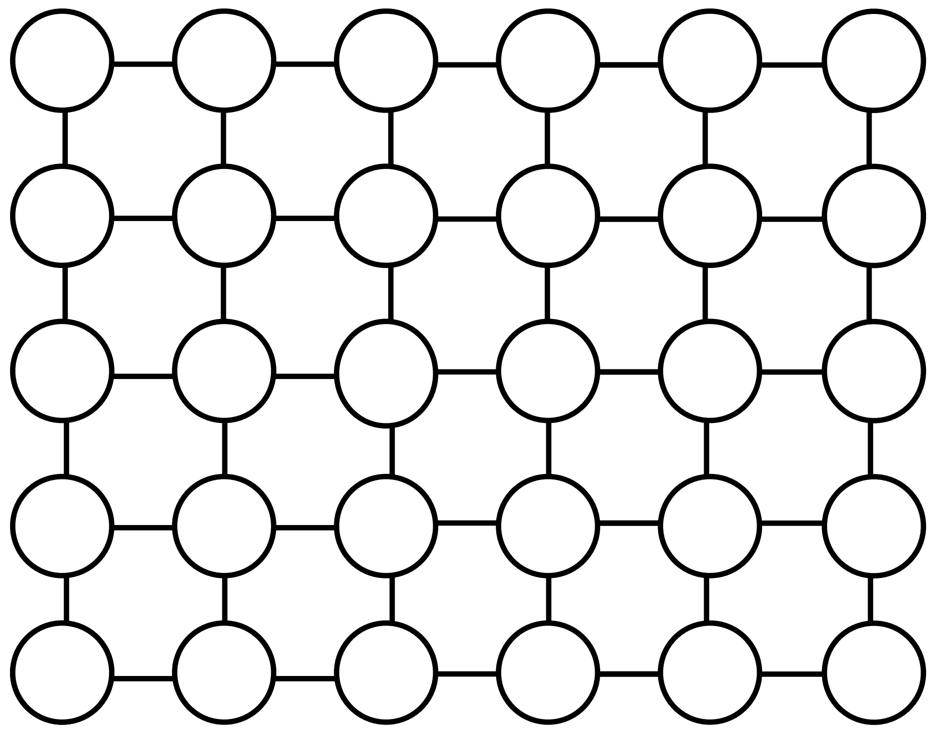
Groundtruth

Data from [maps.google.com]

Challenges in image-to-image translation

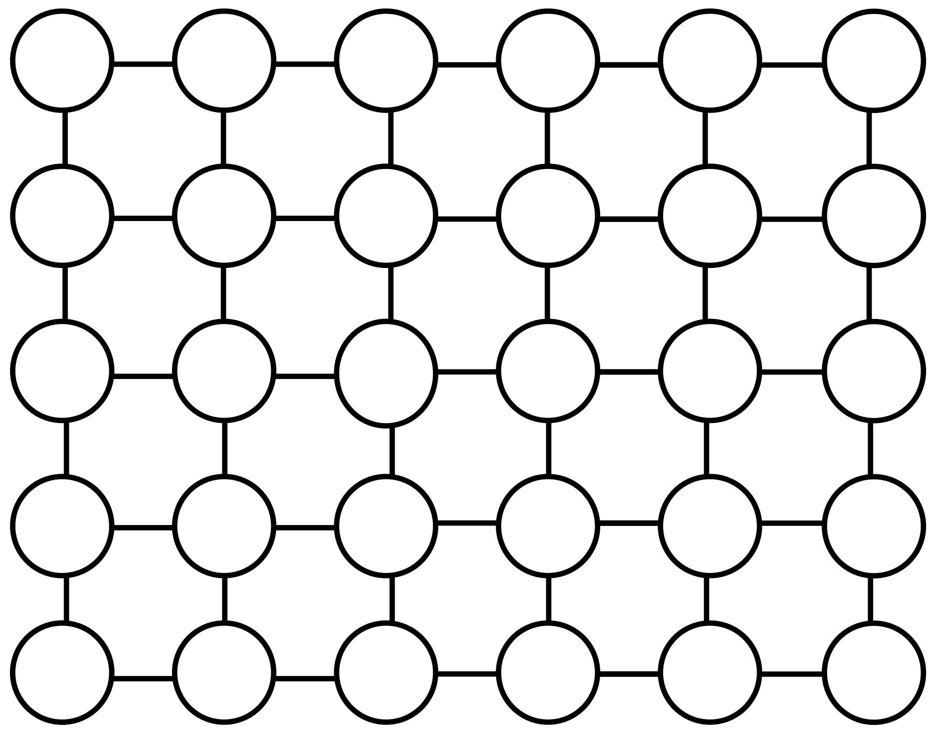
Challenges in image-to-image translation

1. Output is high-dimensional, structured object

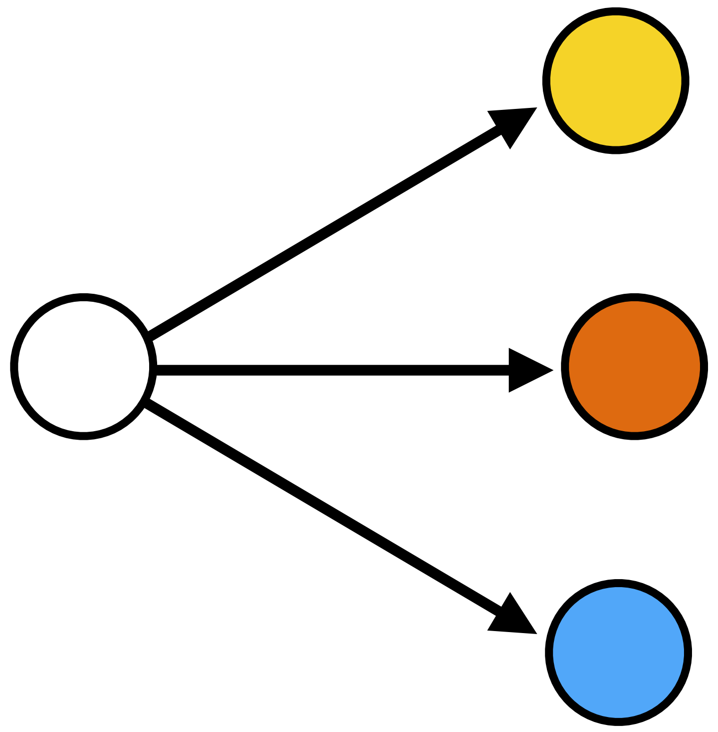


Challenges in image-to-image translation

1. Output is high-dimensional, structured object



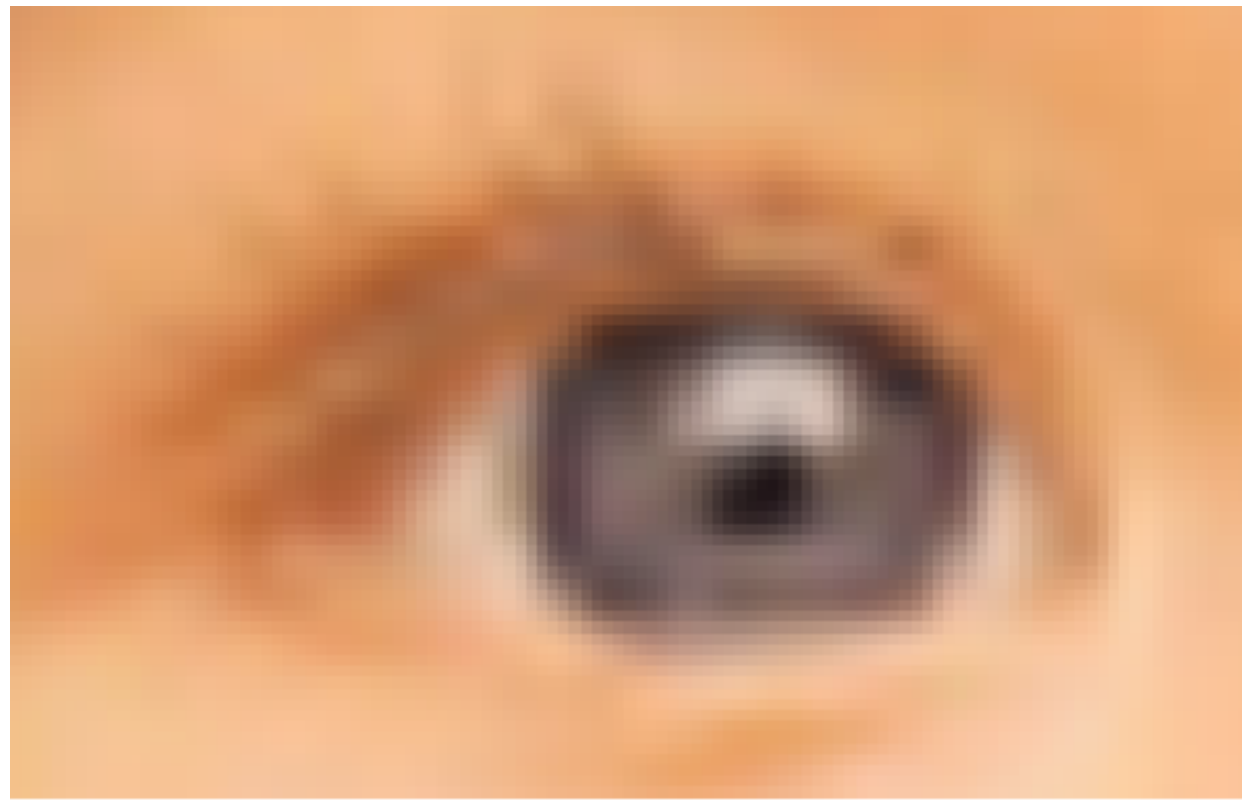
2. Uncertainty in mapping; many plausible outputs



Structured Prediction

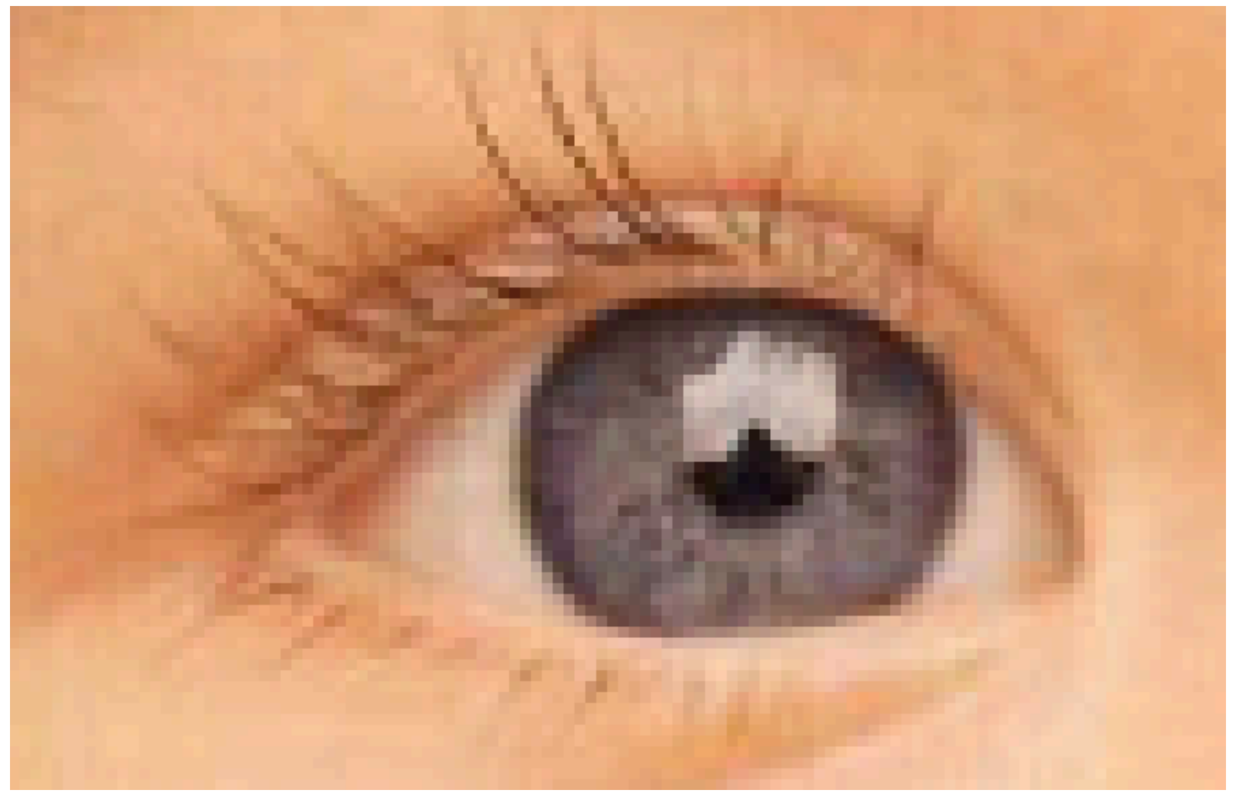
Input

$\mathbf{x}$



Target

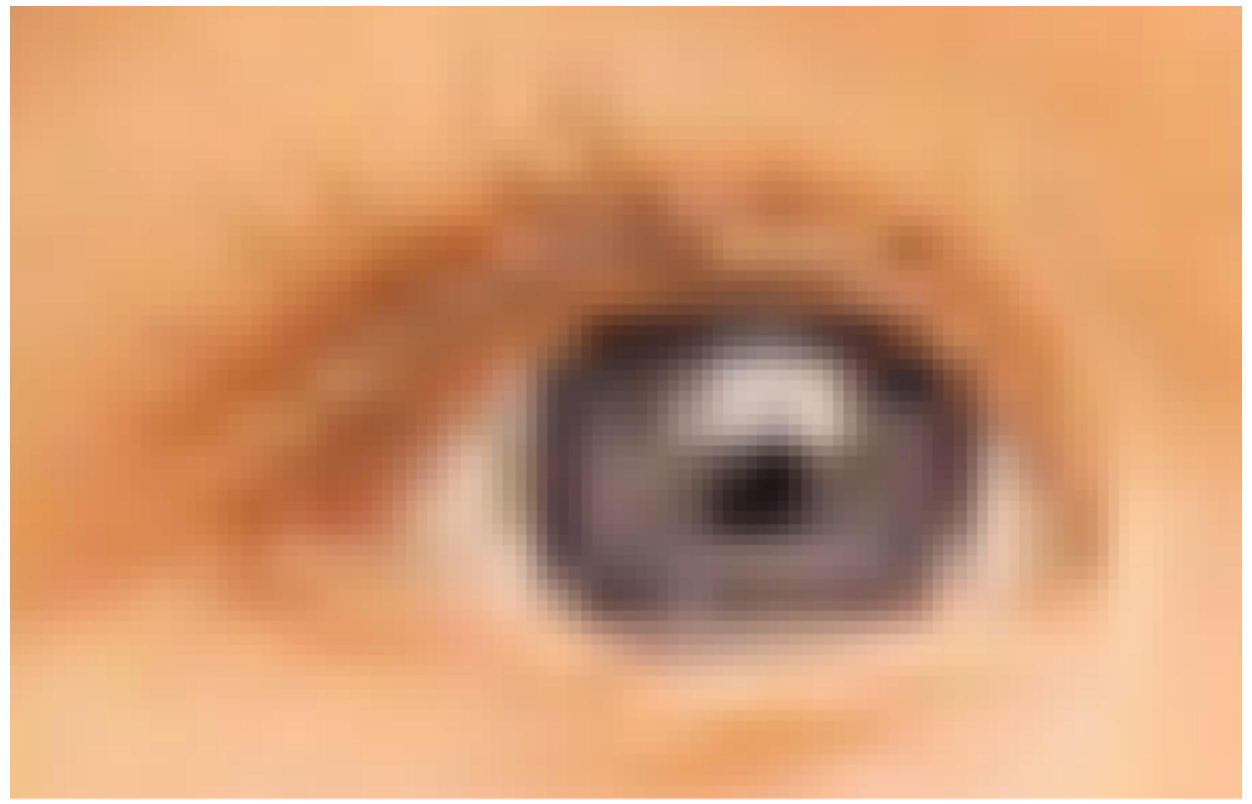
$\mathbf{y}$



Structured Prediction

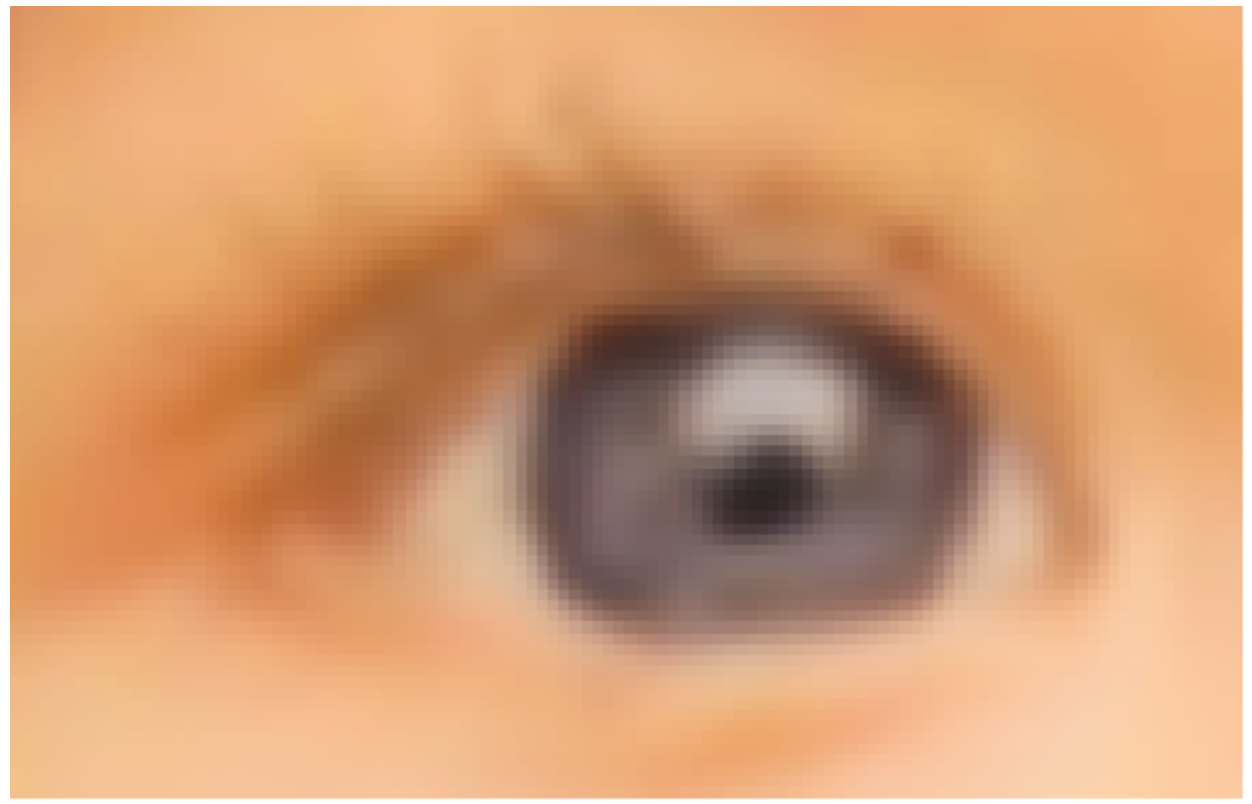
Input

$\mathbf{x}$



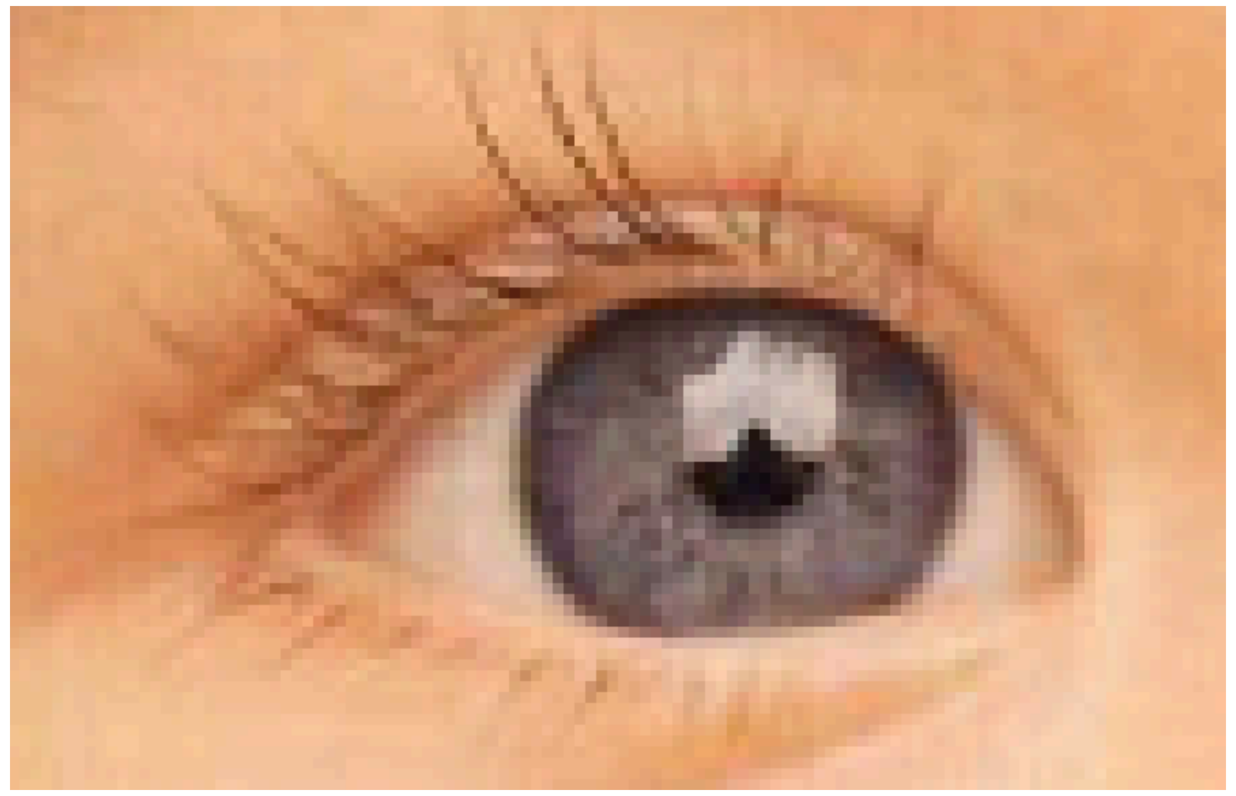
Output

$\hat{\mathbf{y}}$



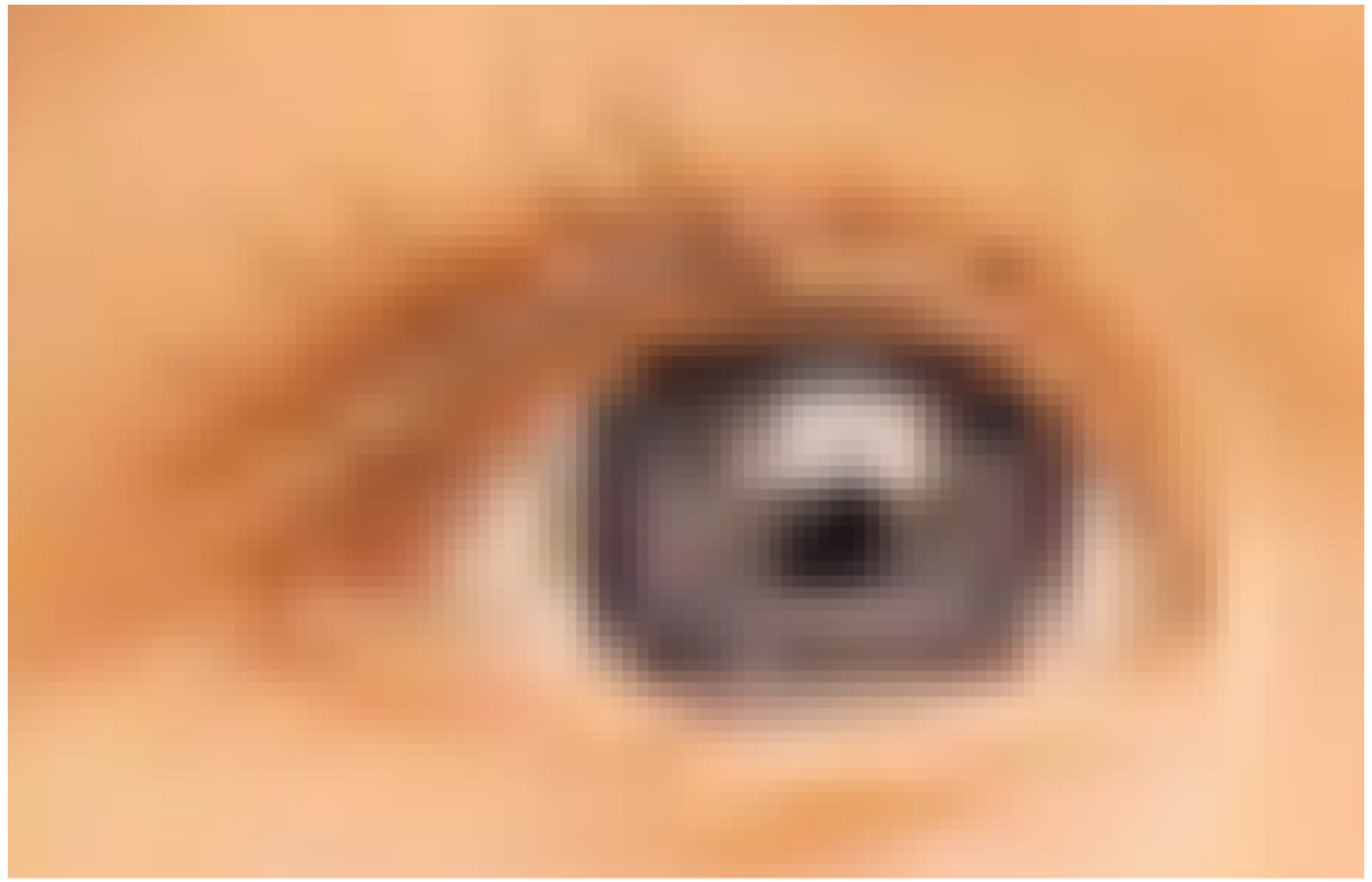
Target

$\mathbf{y}$

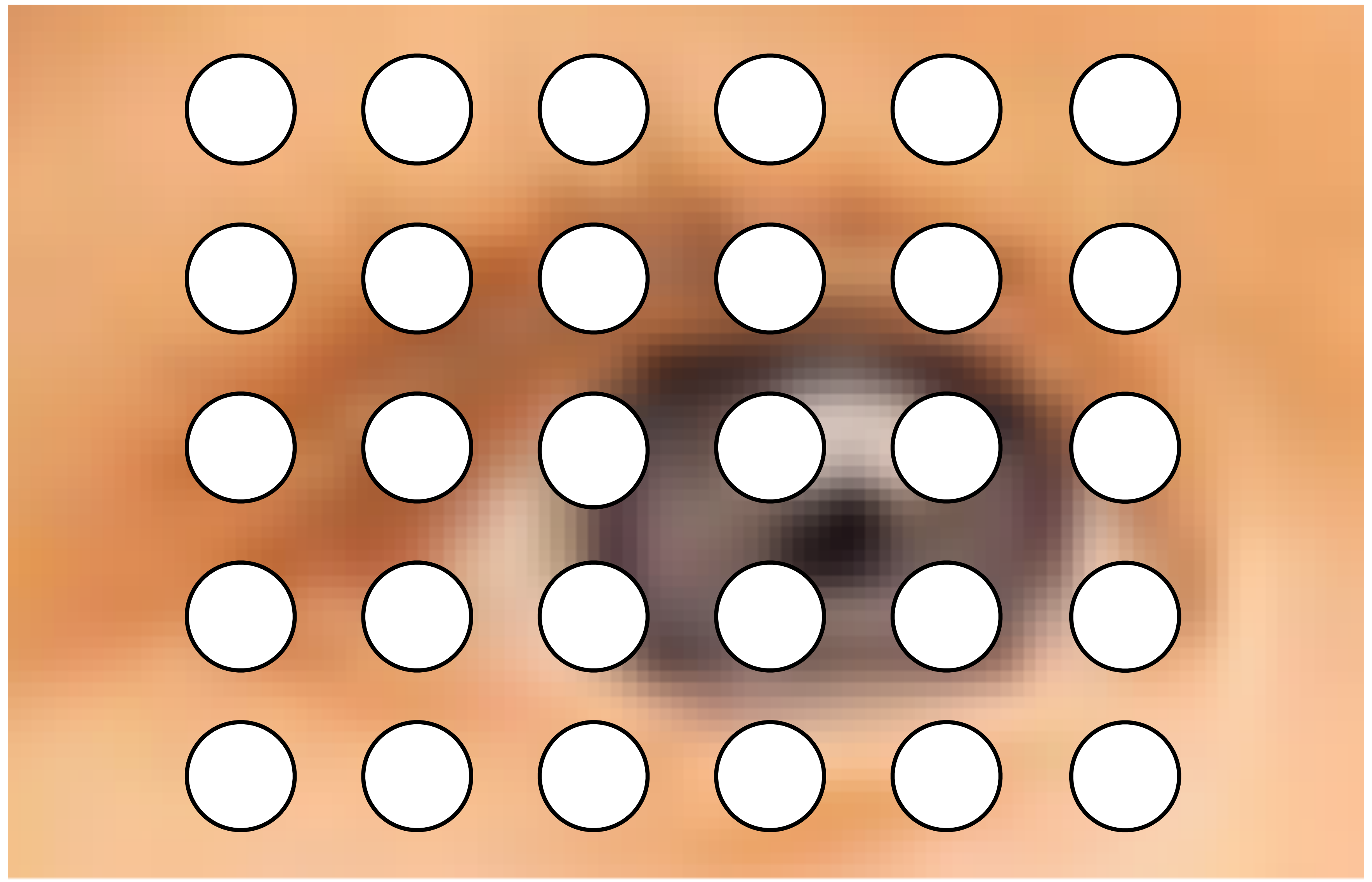


$$L(\hat{\mathbf{y}}, \mathbf{y}) = \|\hat{\mathbf{y}} - \mathbf{y}\|_2$$

Structured Prediction

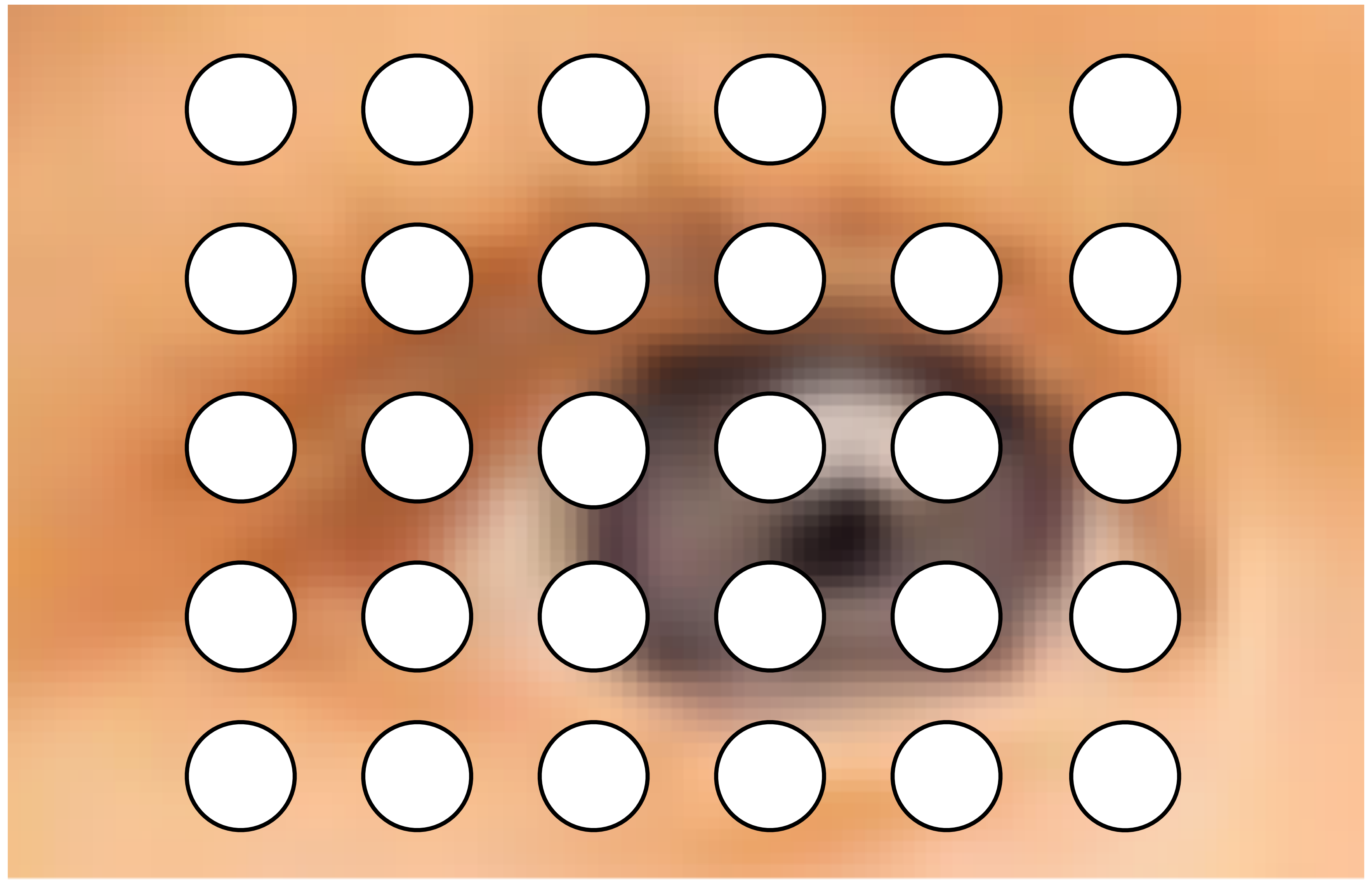


Structured Prediction



Each pixel treated as independent

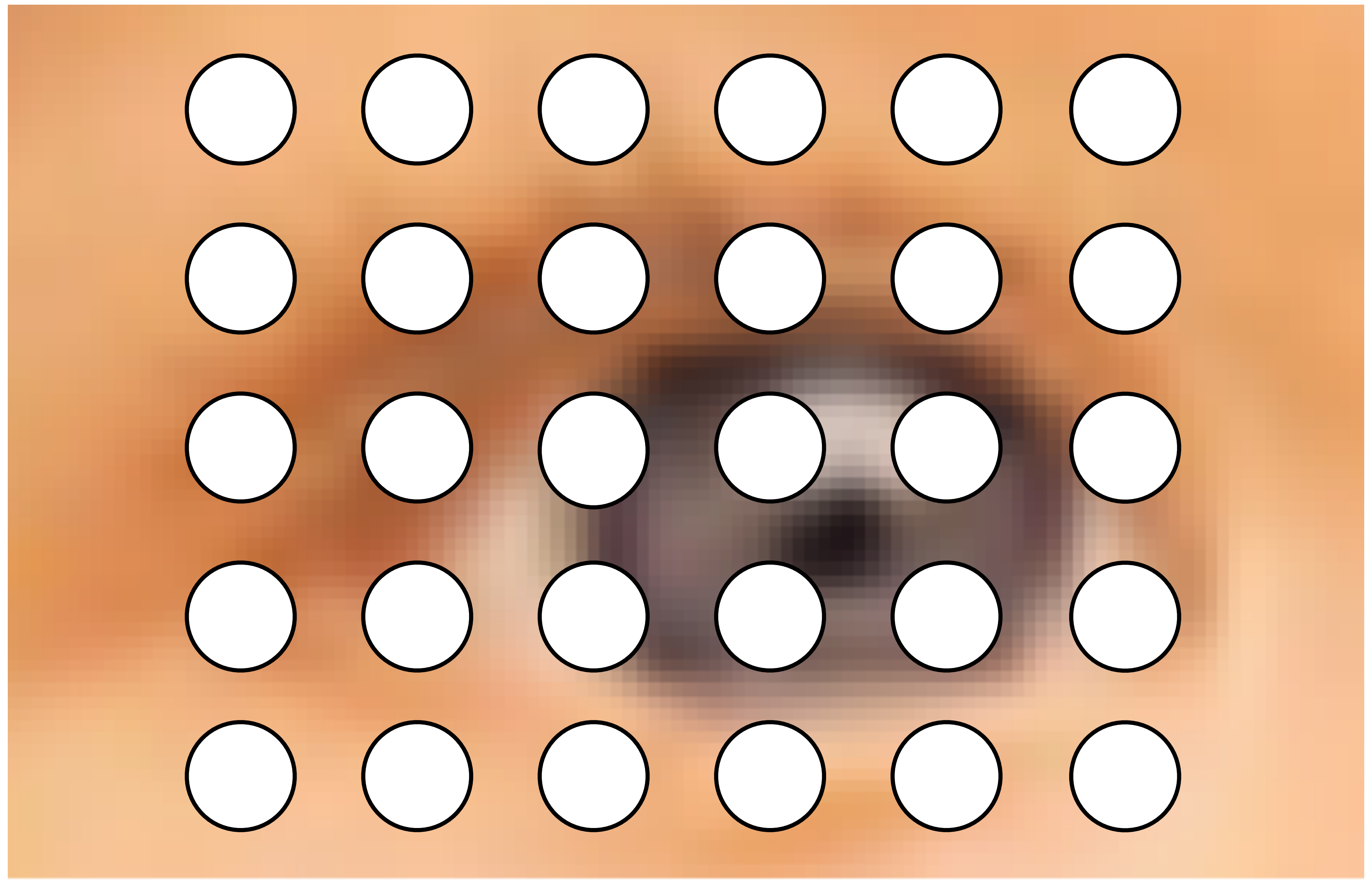
Structured Prediction



Each pixel treated as independent

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

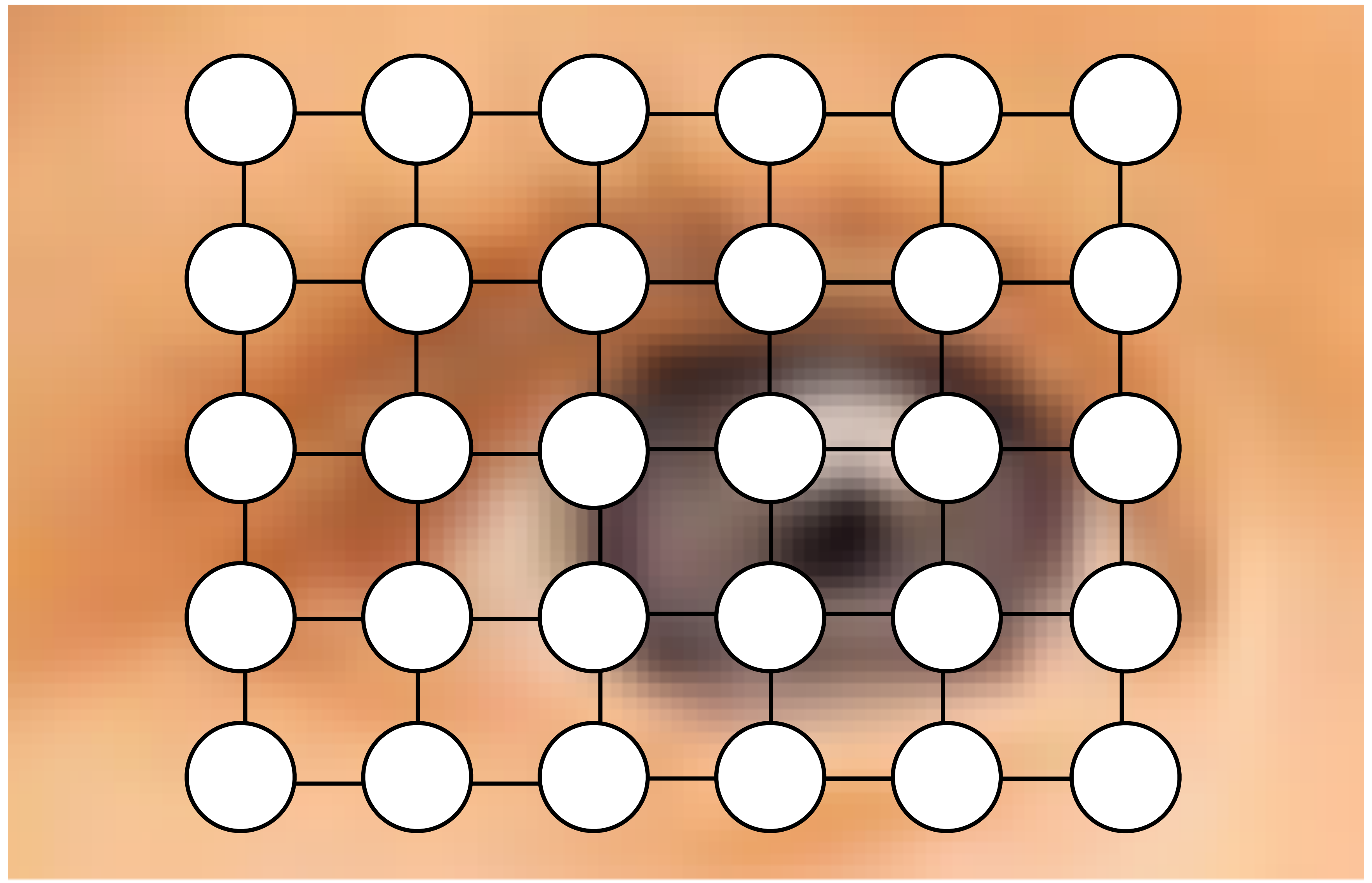
Structured Prediction



Each pixel treated as independent

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

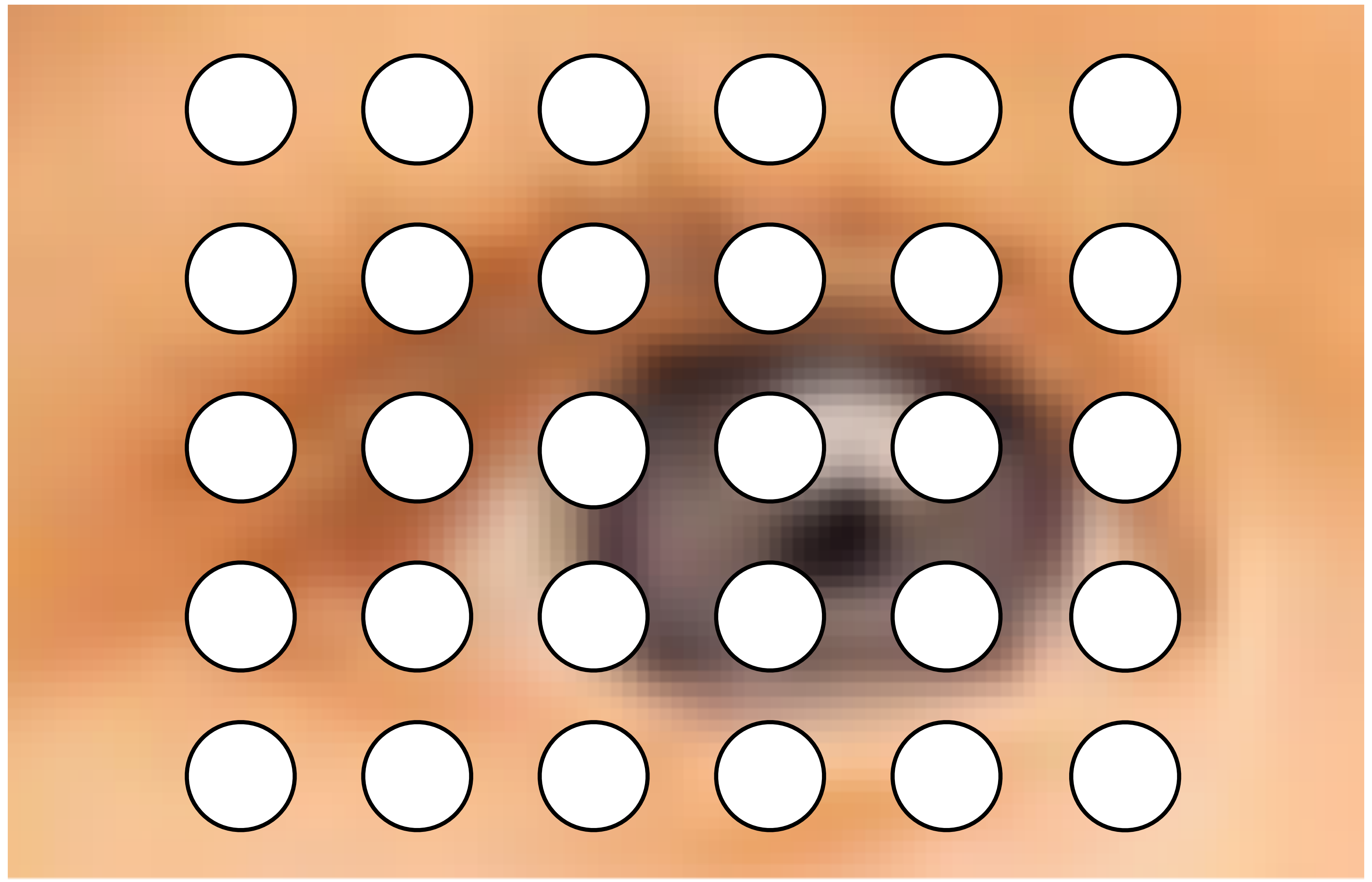
CRF

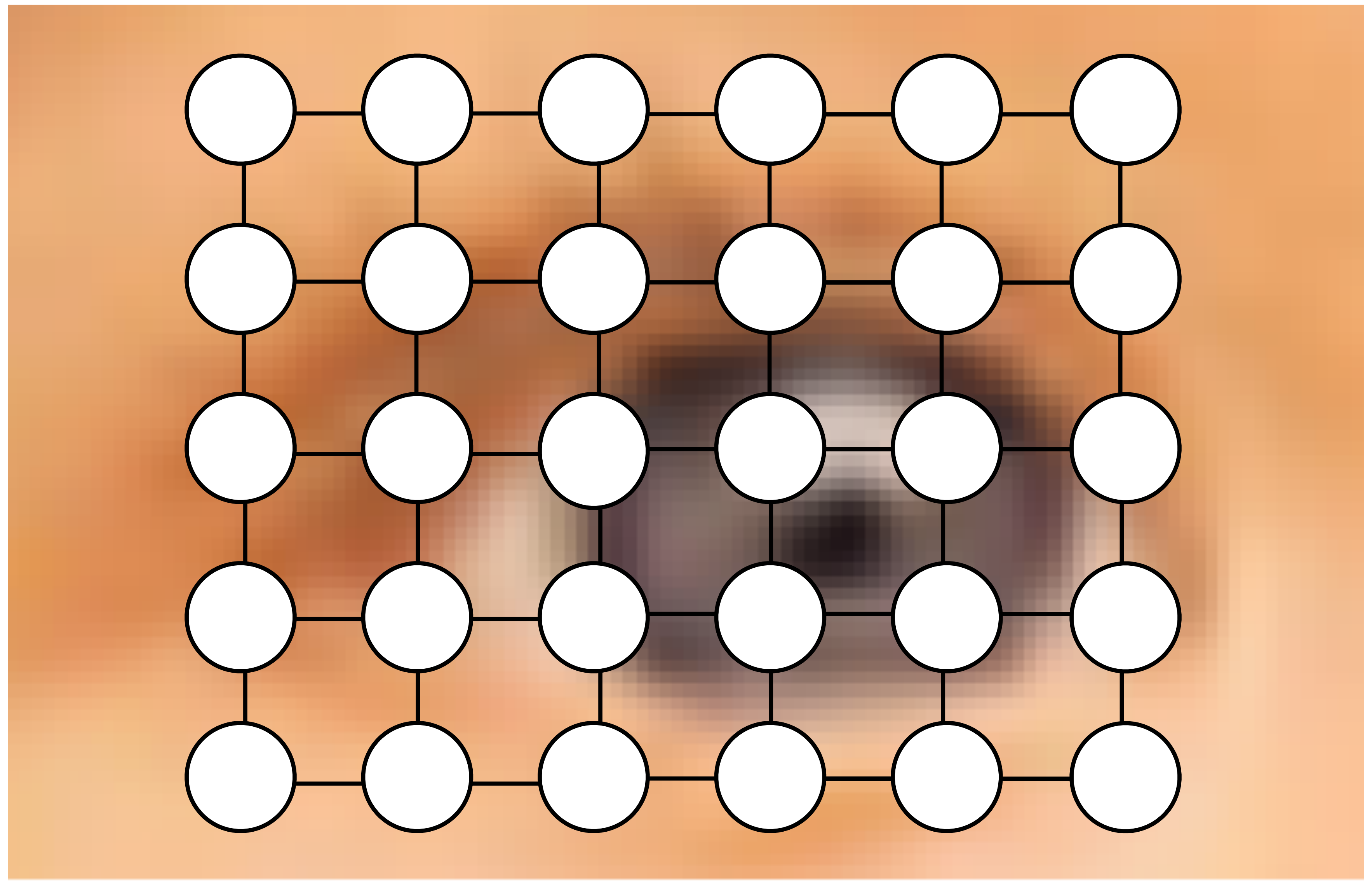


Models at pairwise configuration of pixels

Structured Prediction

CRF





Each pixel treated as independent

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

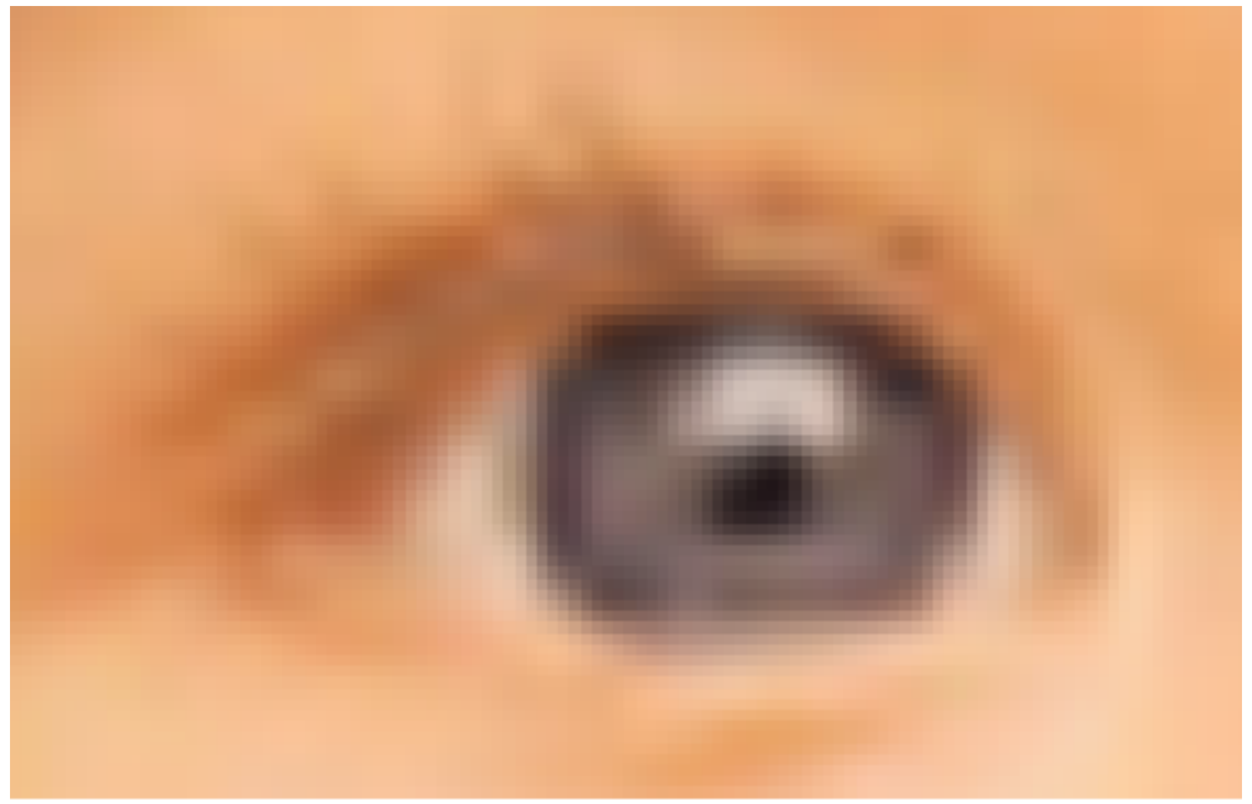
Models at pairwise configuration of pixels

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

"Perceptual Loss"

Input

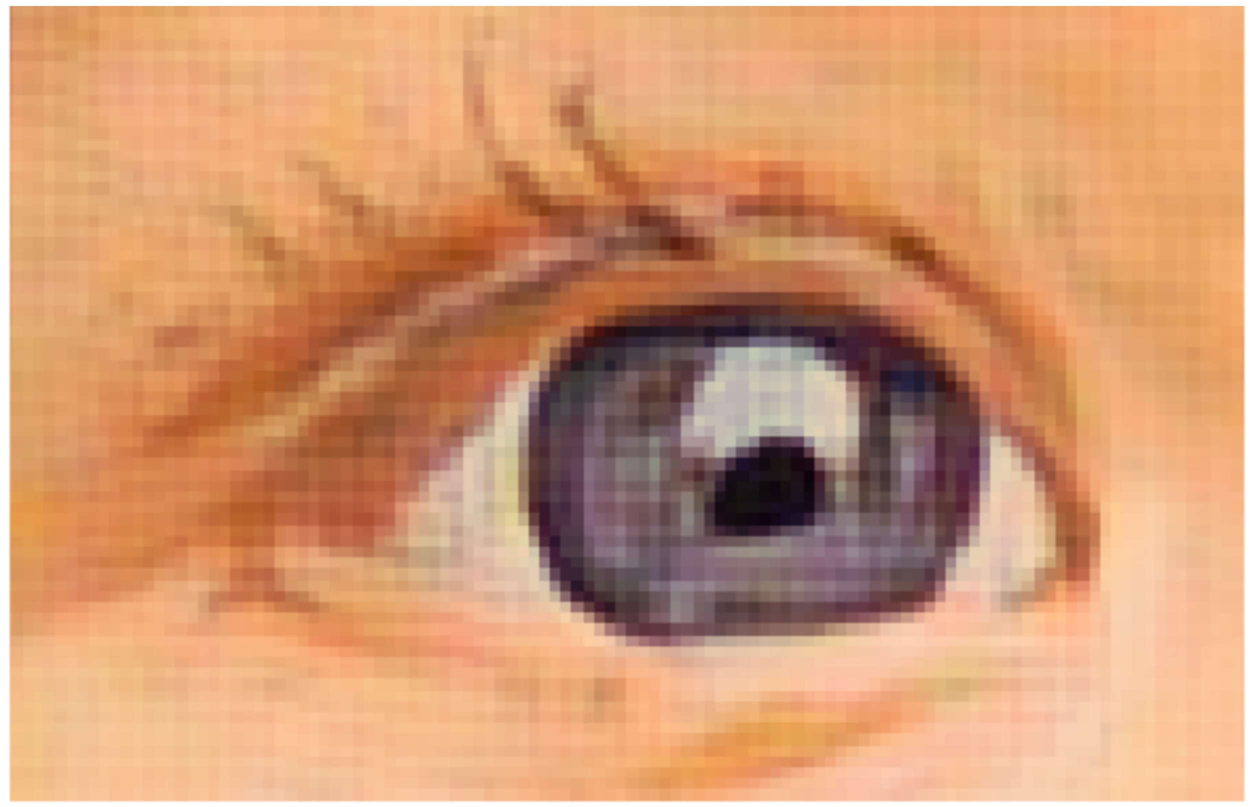
$\mathbf{x}$



Output

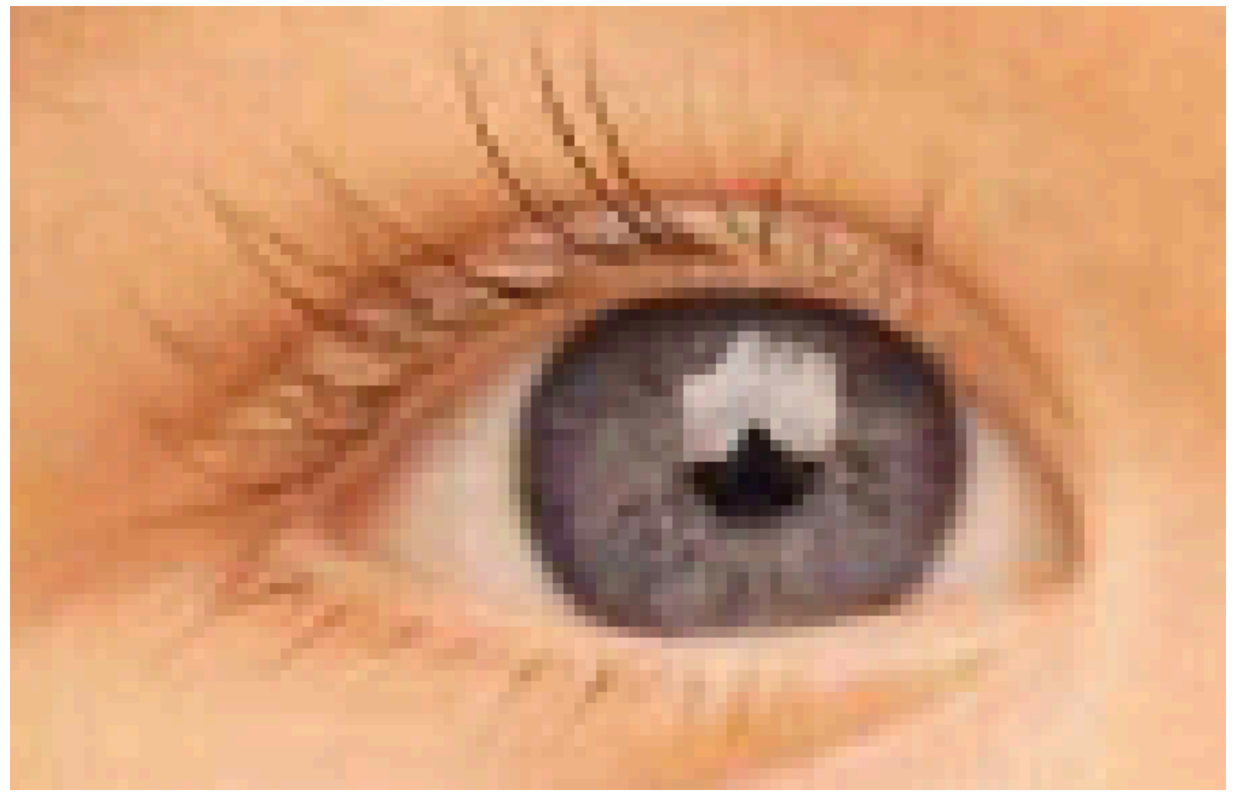
[Johnson, Alahi, Li 2016]

$\hat{\mathbf{y}}$



Target

$\mathbf{y}$



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

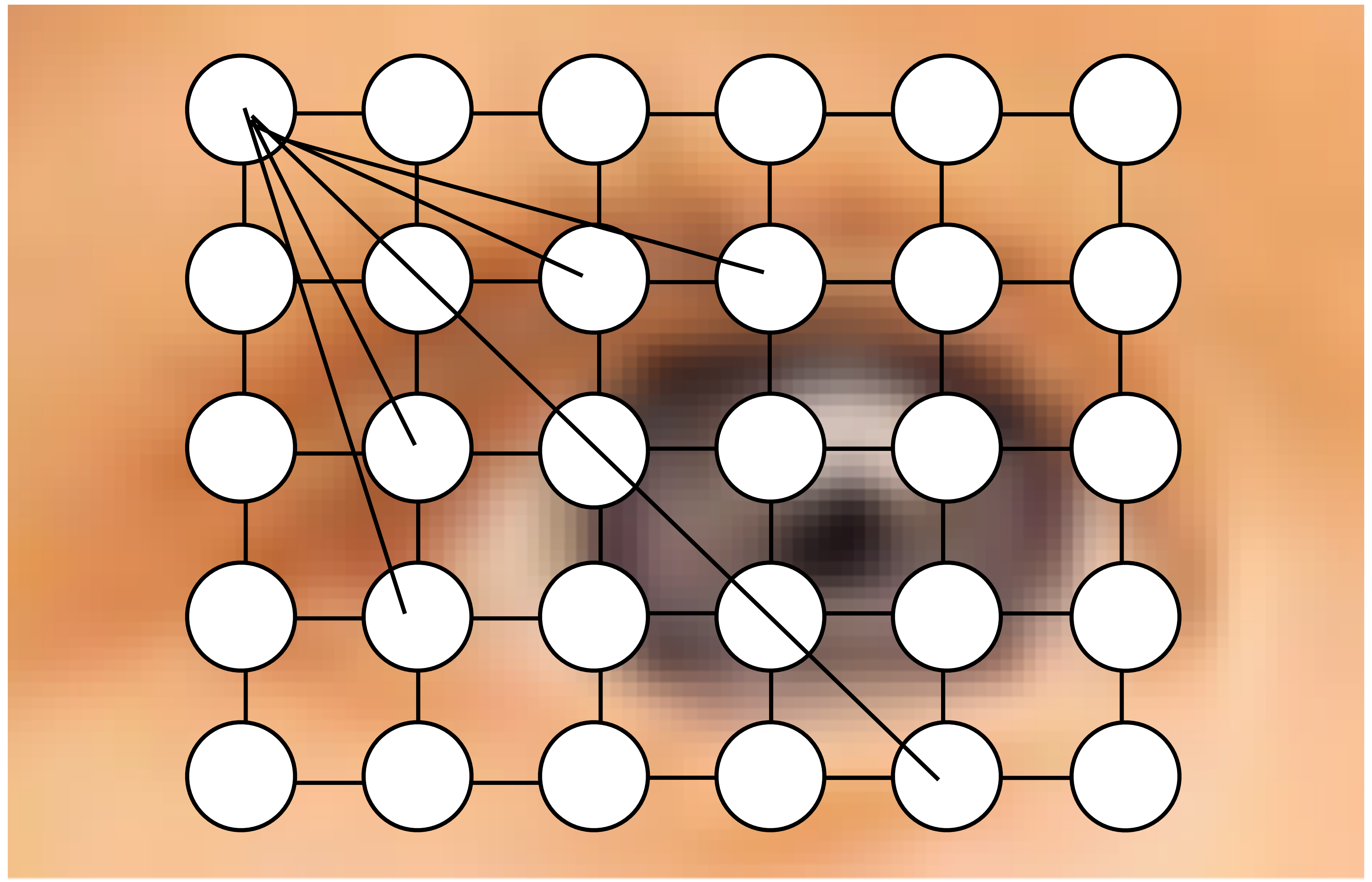
[Johnson, Alahi, Li, ECCV 2016]

[Chen & Koltun ICCV 2017]

[Zhang et al. CVPR 2018]

[Mostajabi, Maire, Shakhnarovich, arXiv 2018]

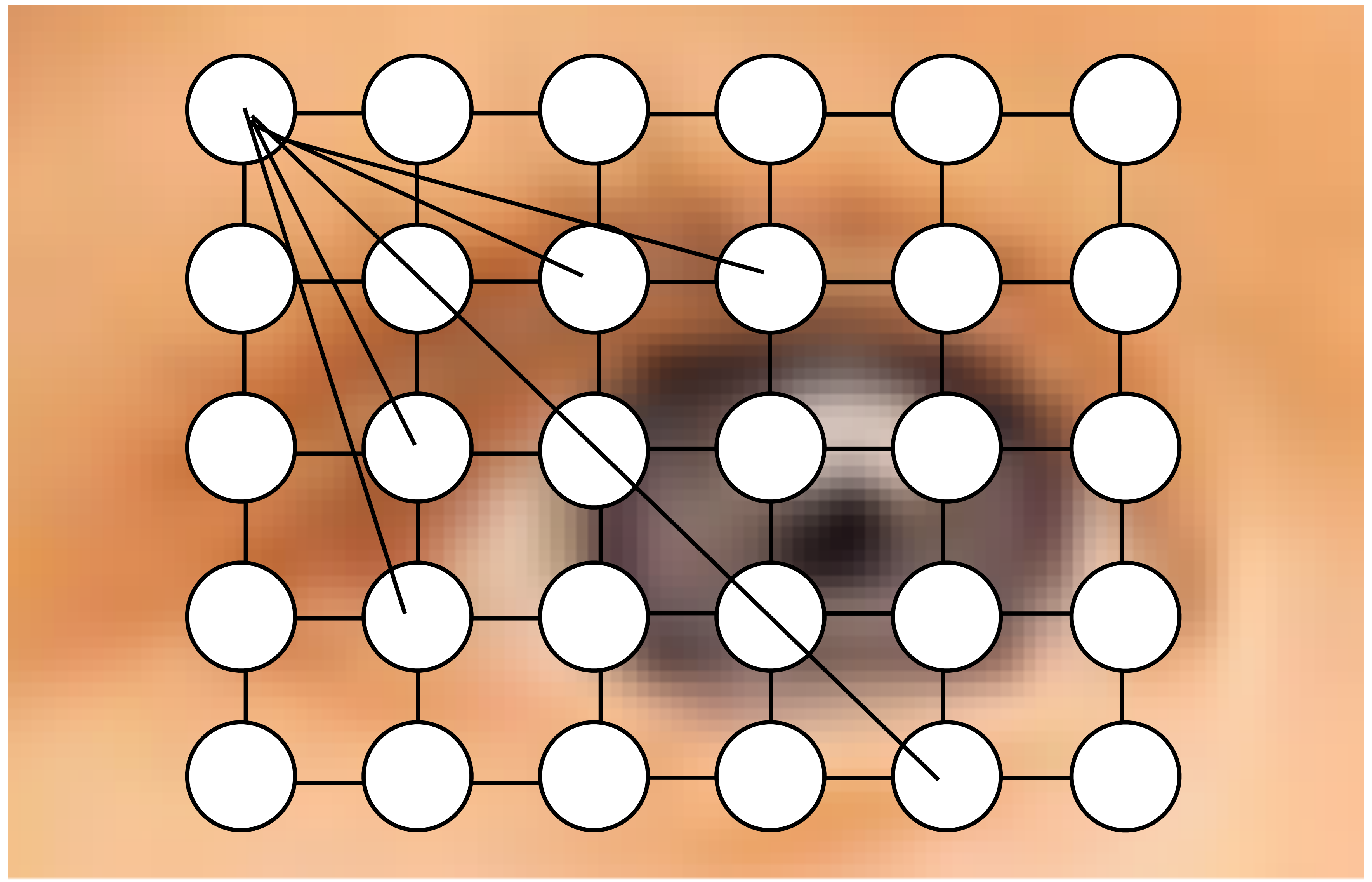
Structured Prediction



Model *joint* configuration of all pixels

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

Structured Prediction

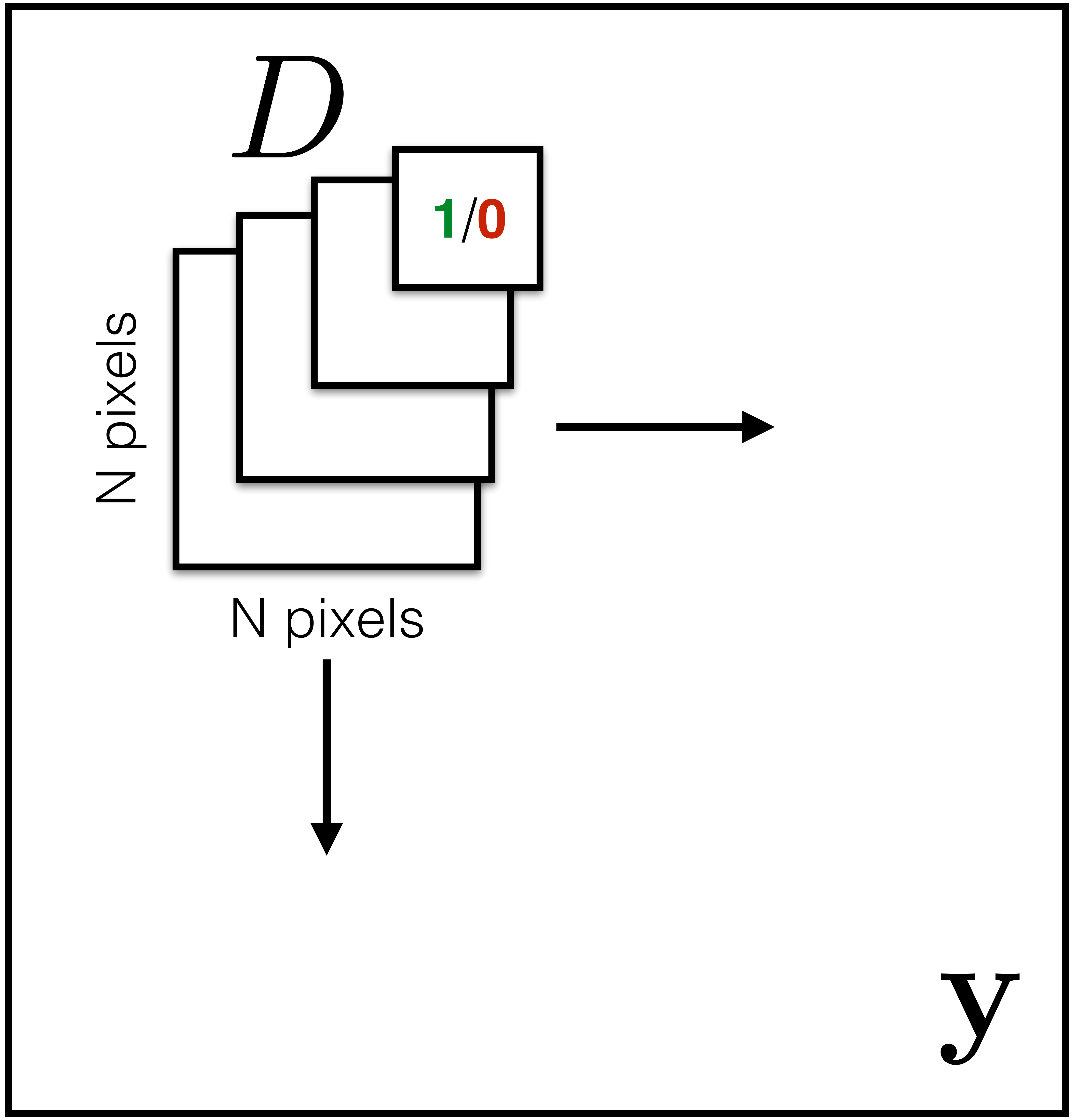


Model *joint* configuration
of all pixels

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

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

Patch Discriminator

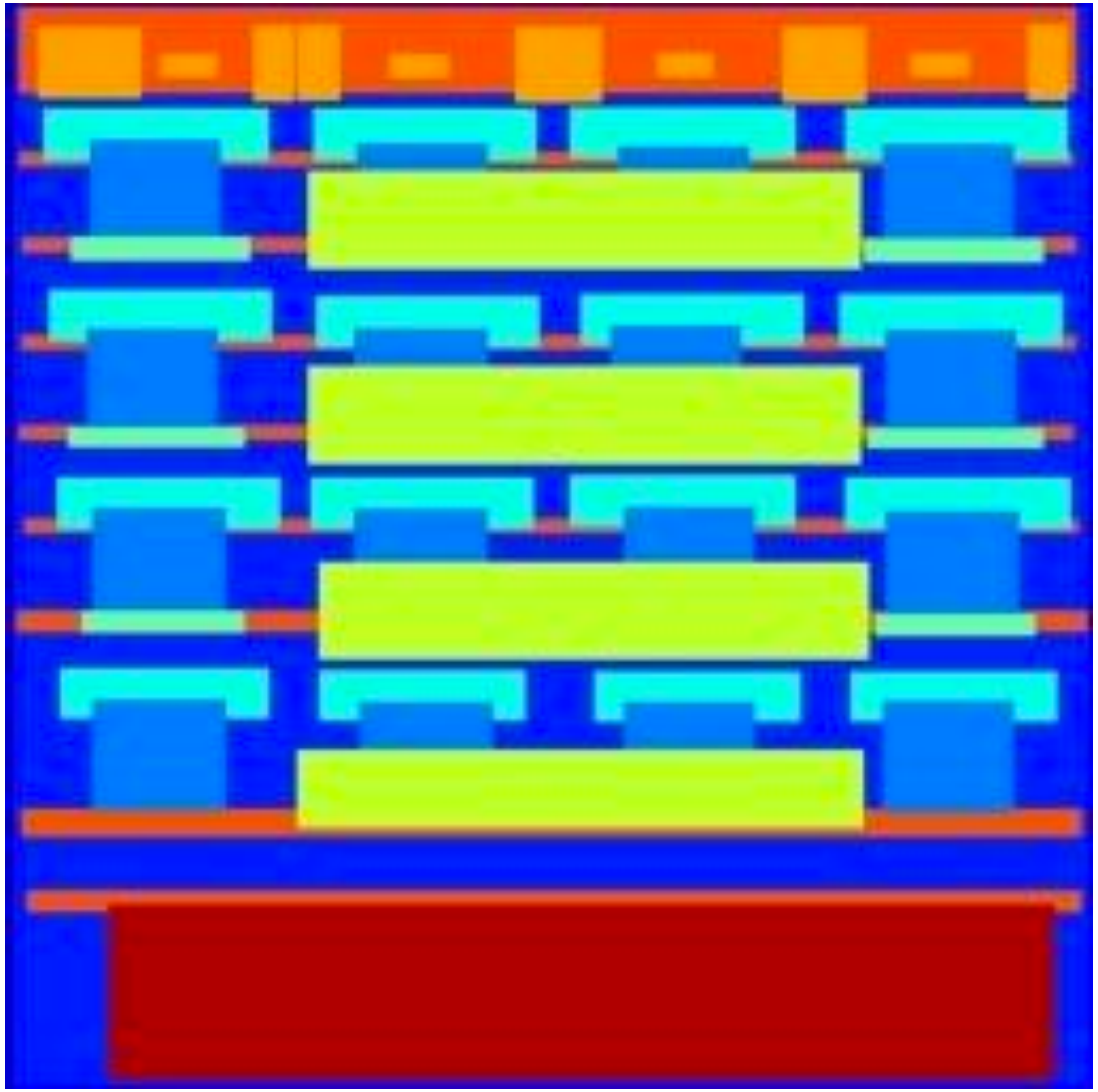


Rather than penalizing if output *image* looks fake, penalize if each overlapping *patch* in output looks fake

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

Labels → Facades

Input

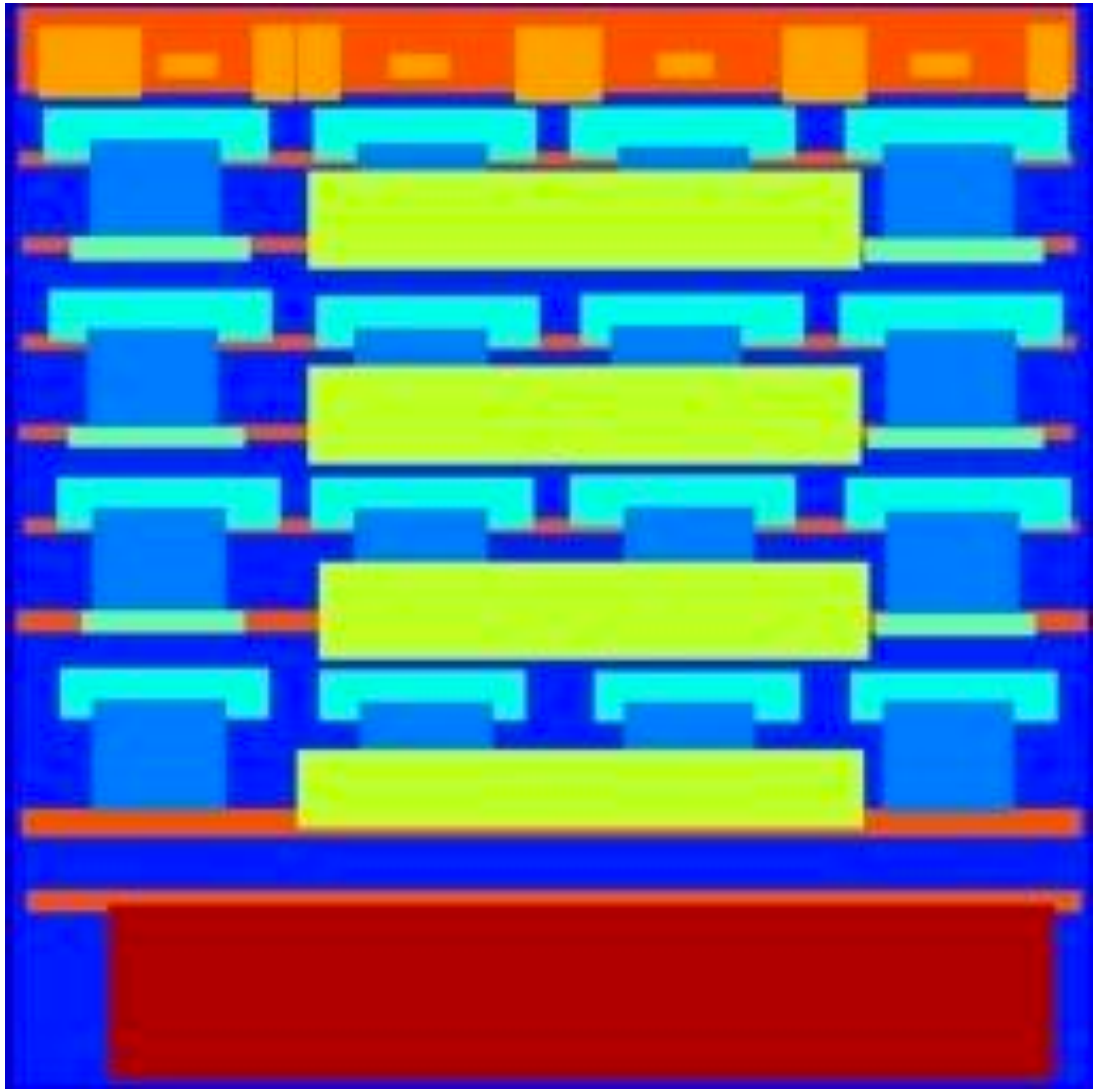


1x1 Discriminator

Data from [Tylecek, 2013]

Labels → Facades

Input

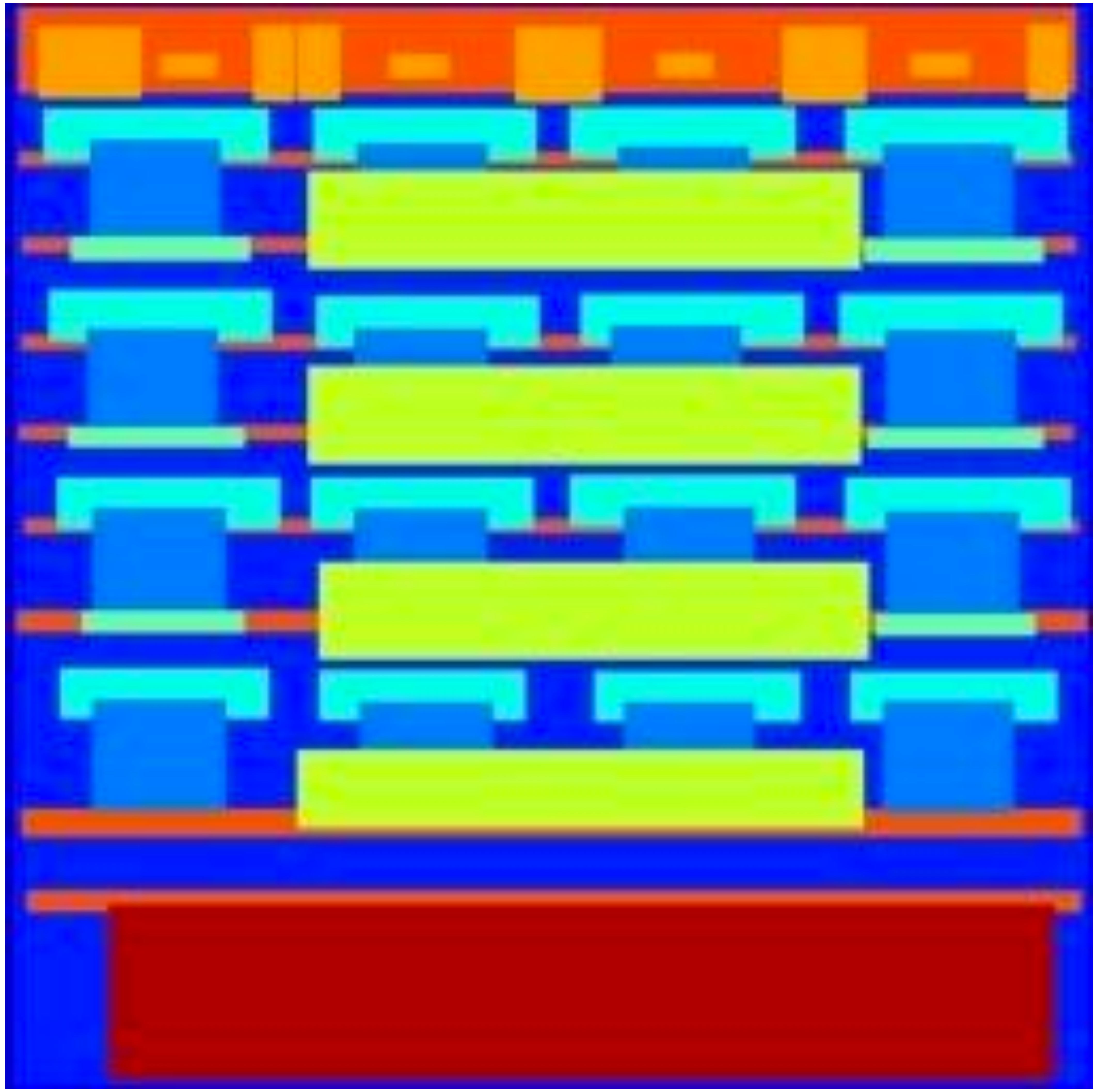


16x16 Discriminator

Data from [Tylecek, 2013]

Labels → Facades

Input

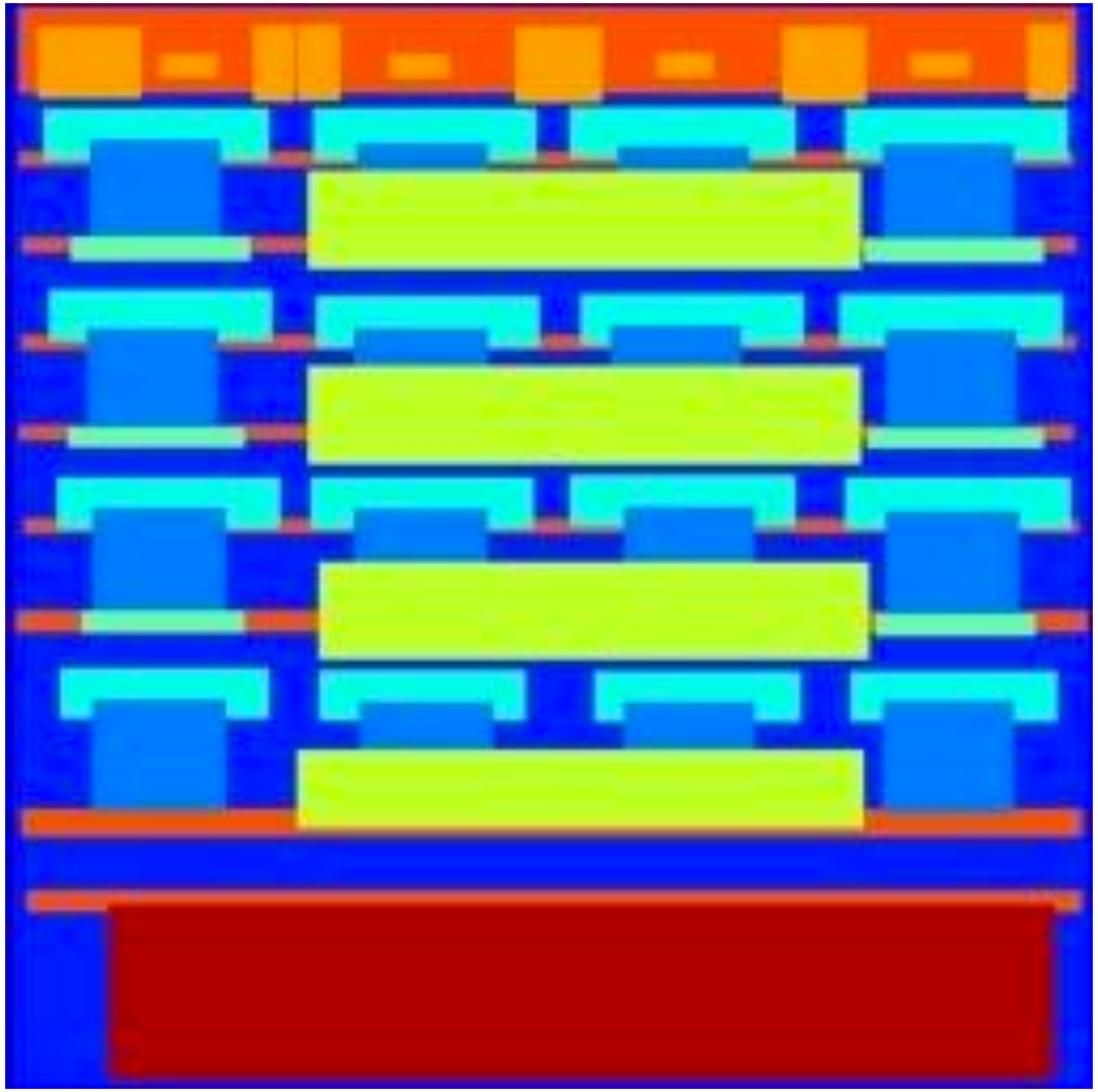


70x70 Discriminator

Data from [Tylecek, 2013]

Labels → Facades

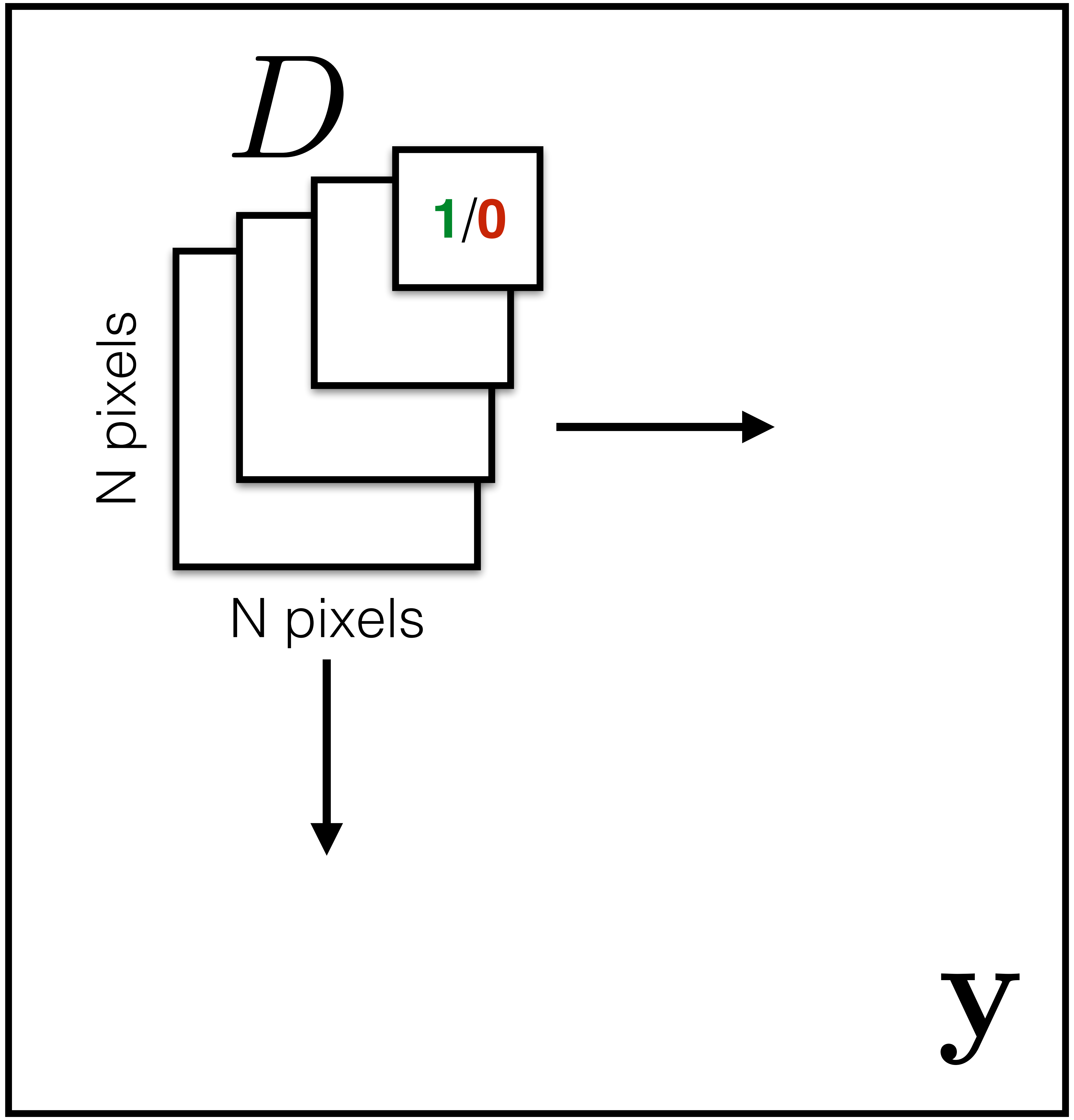
Input



Full image Discriminator

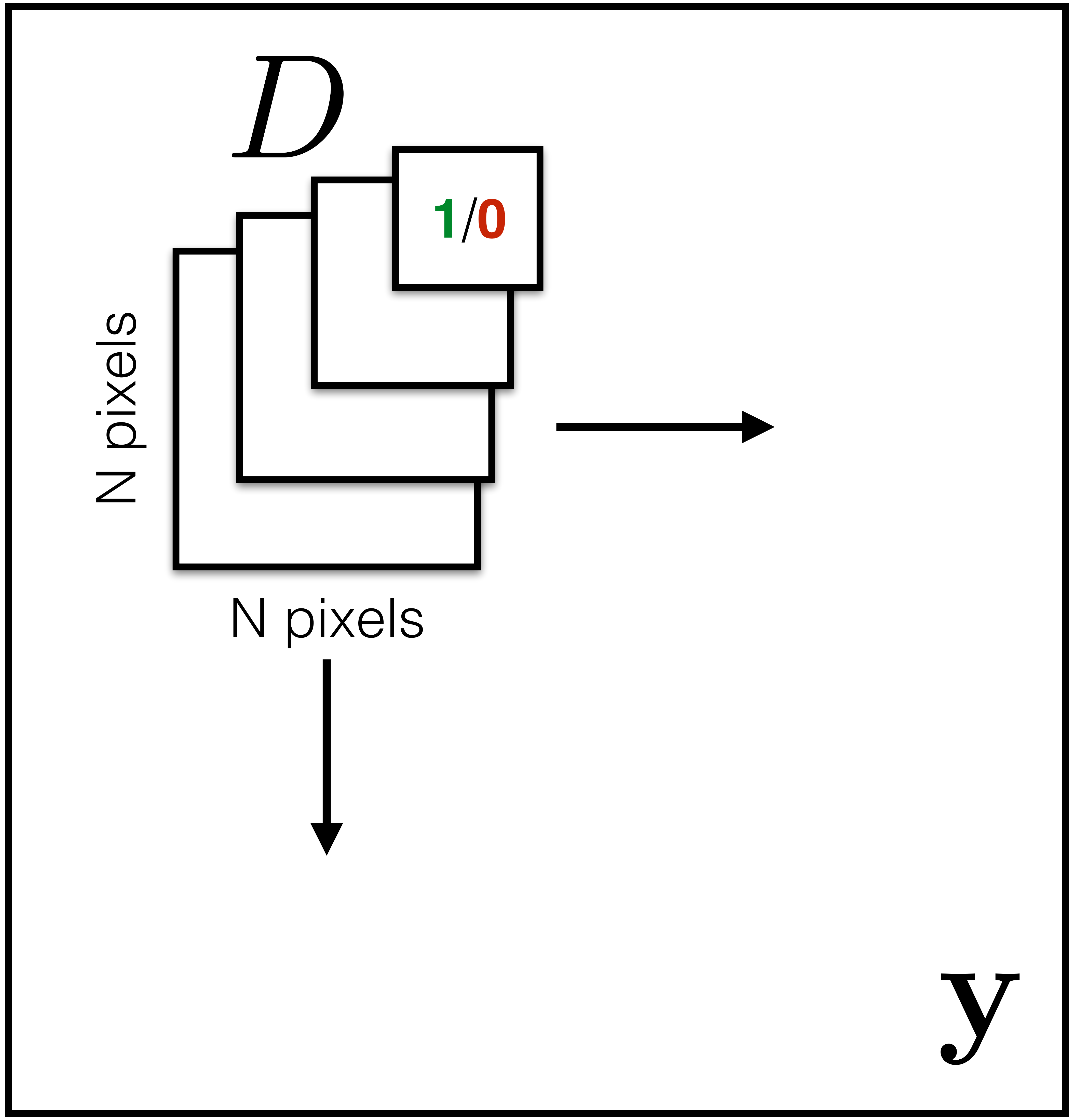
Data from [Tylecek, 2013]

Patch Discriminator



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

Patch Discriminator

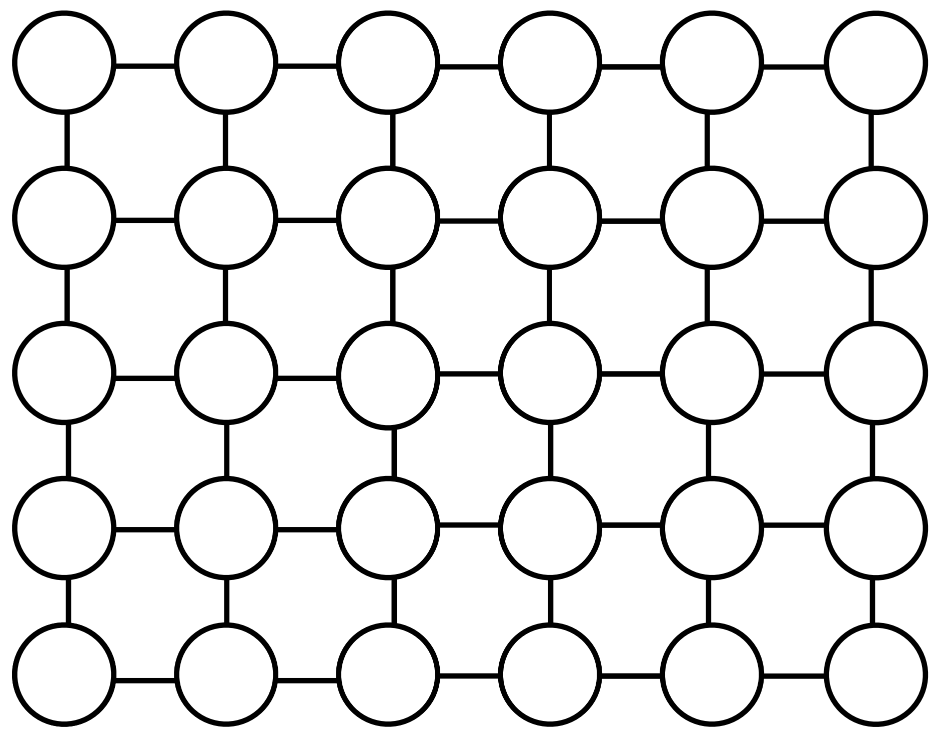


Rather than penalizing if output *image* looks fake, penalize if each overlapping *patch* in output looks fake

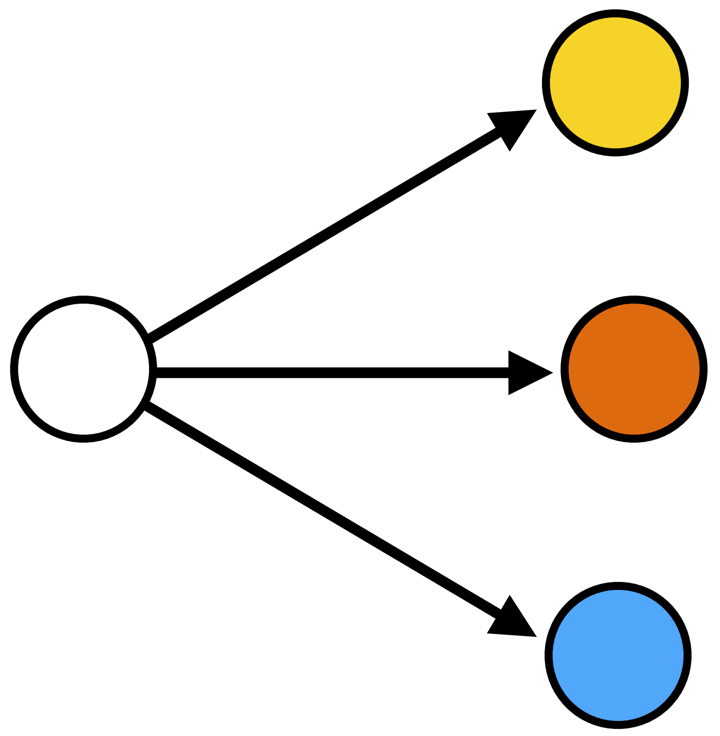
- Faster, fewer parameters
- More supervised observations
- Applies to arbitrarily large images

Challenges in image-to-image translation

1. Output is high-dimensional, structured object

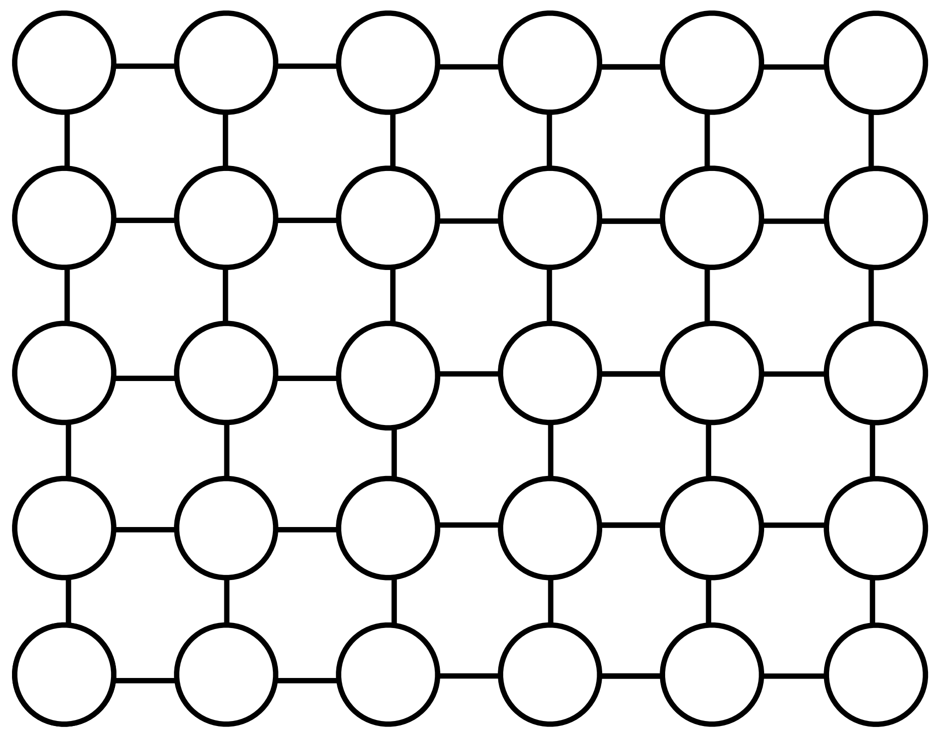


2. Uncertainty in mapping; many plausible outputs



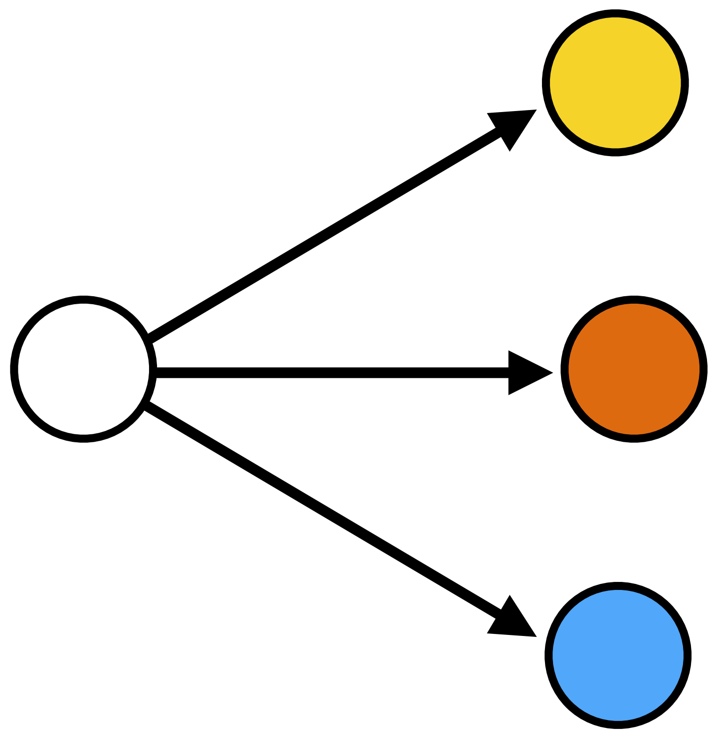
Challenges in image-to-image translation

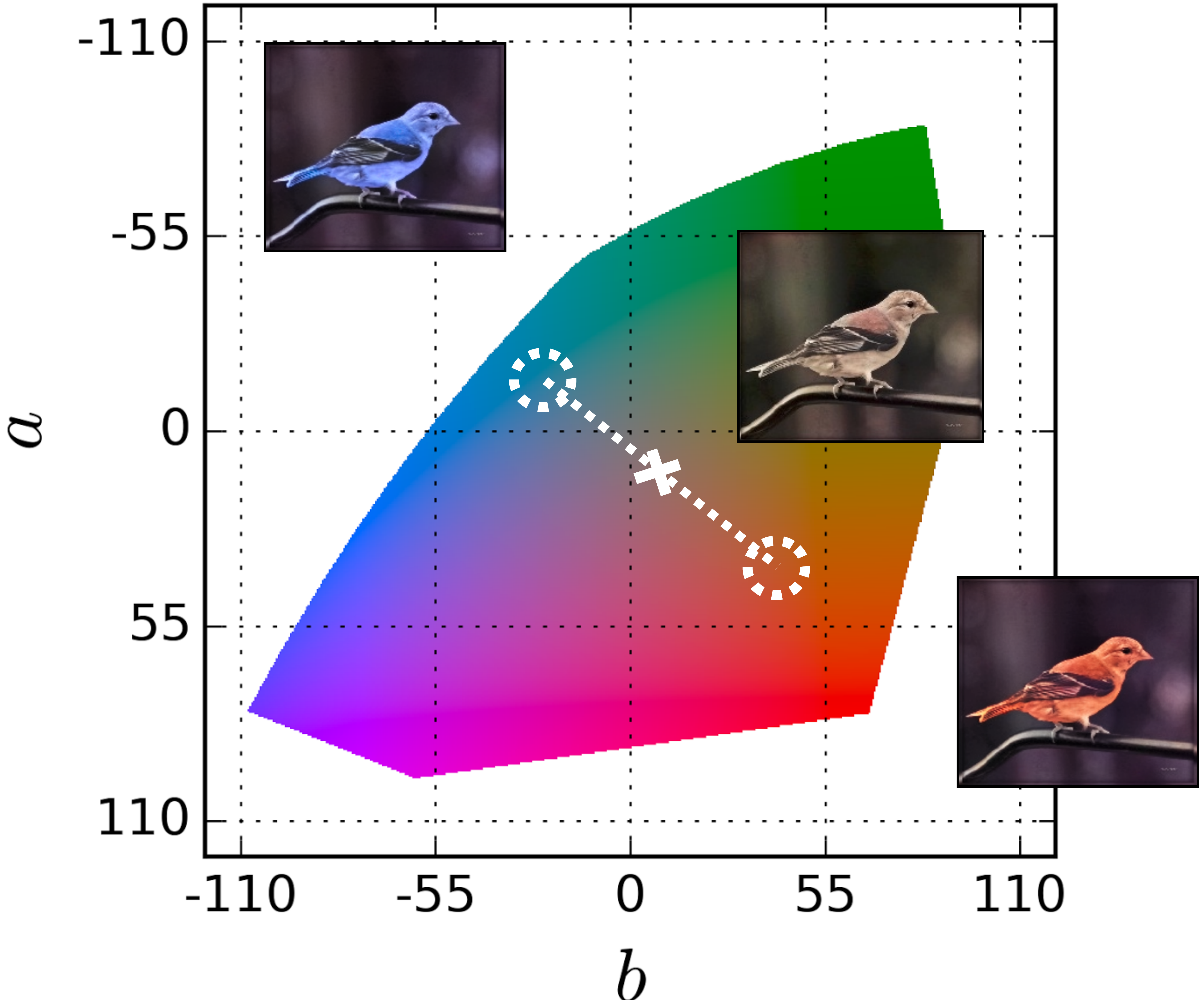
1. Output is high-dimensional, structured object



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

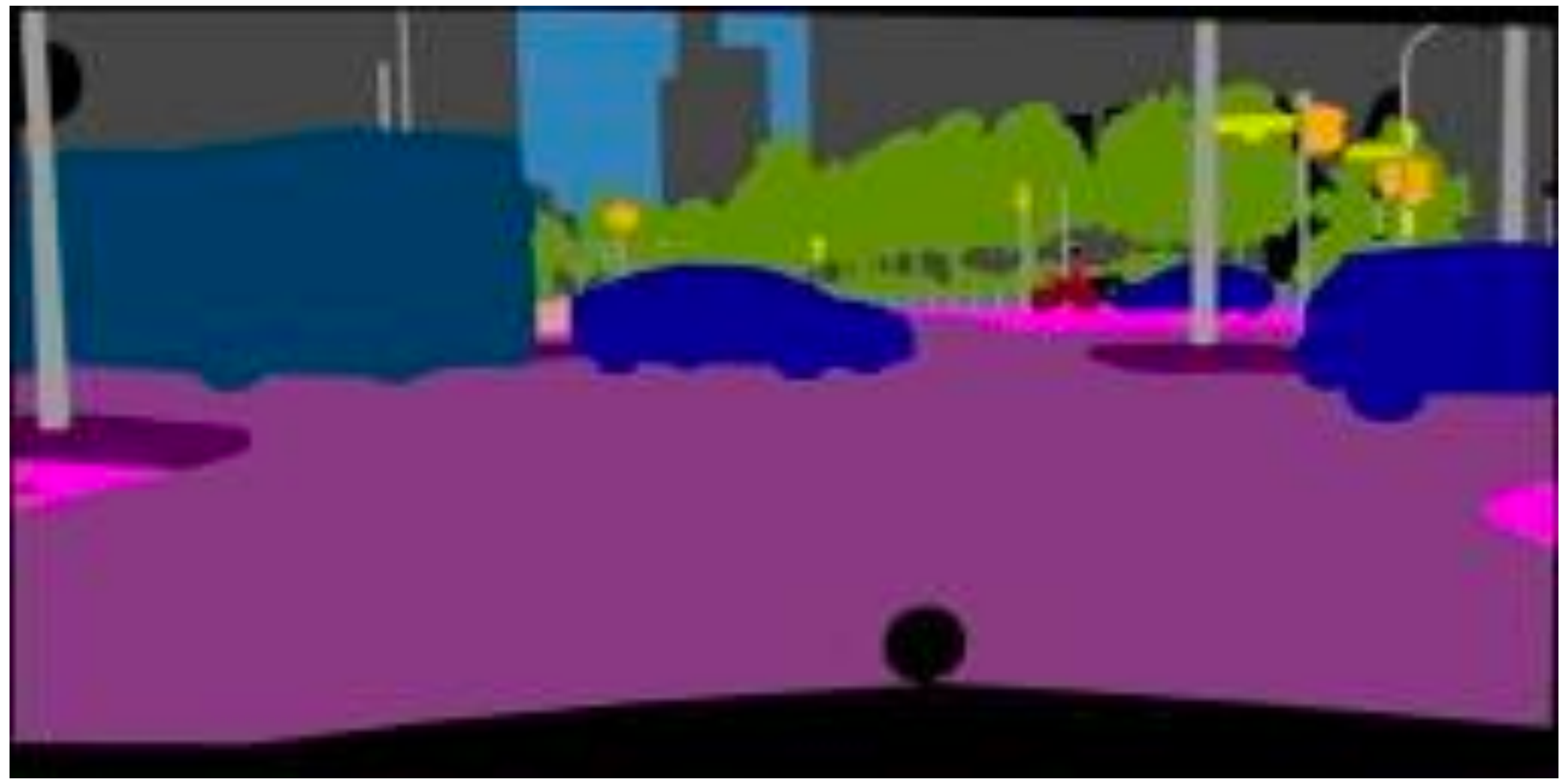
2. Uncertainty in mapping; many plausible outputs



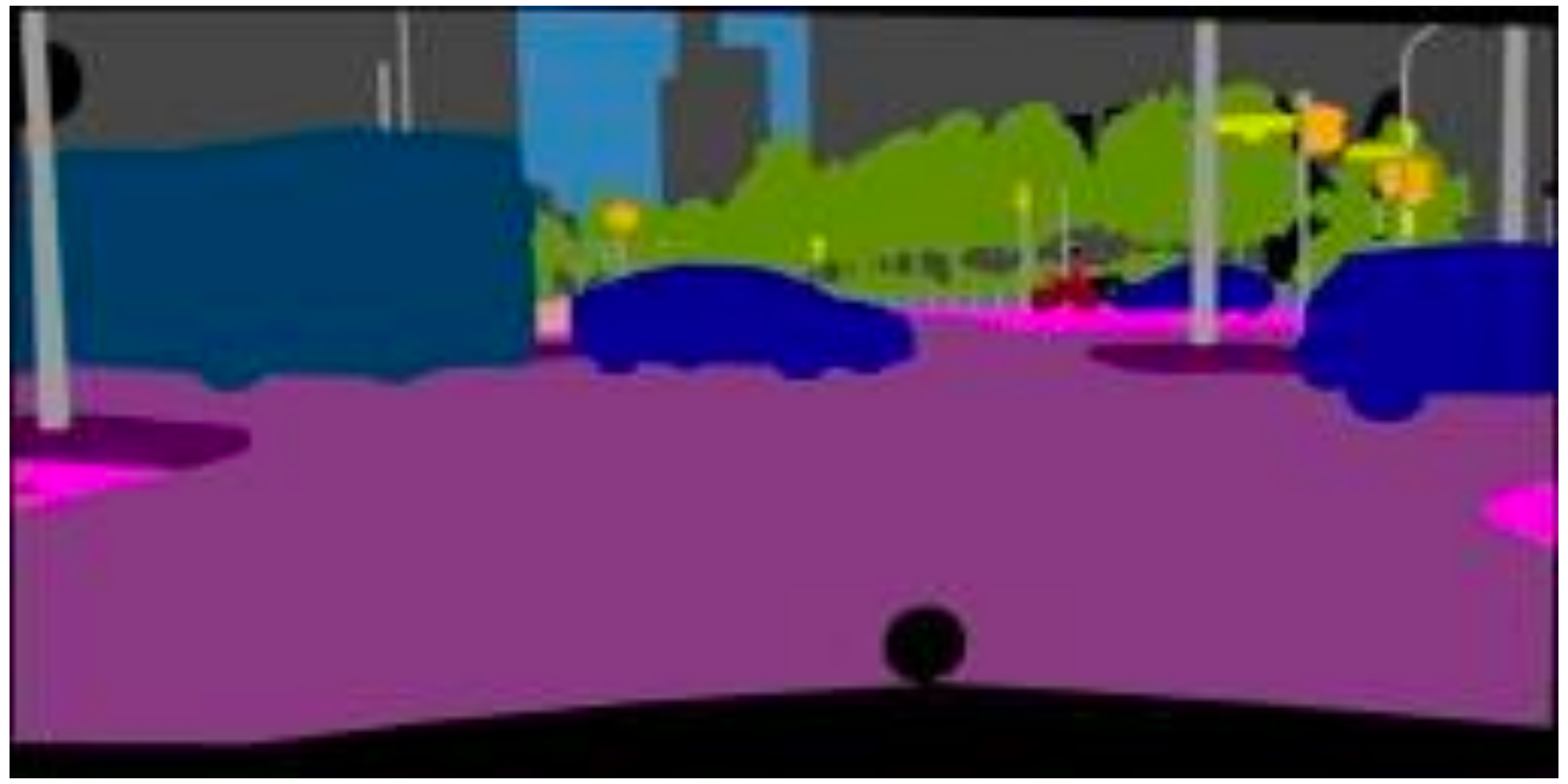


$$\mathrm{L}_2(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \| \mathbf{Y}_{h,w} - \widehat{\mathbf{Y}}_{h,w} \|_2^2$$

Input

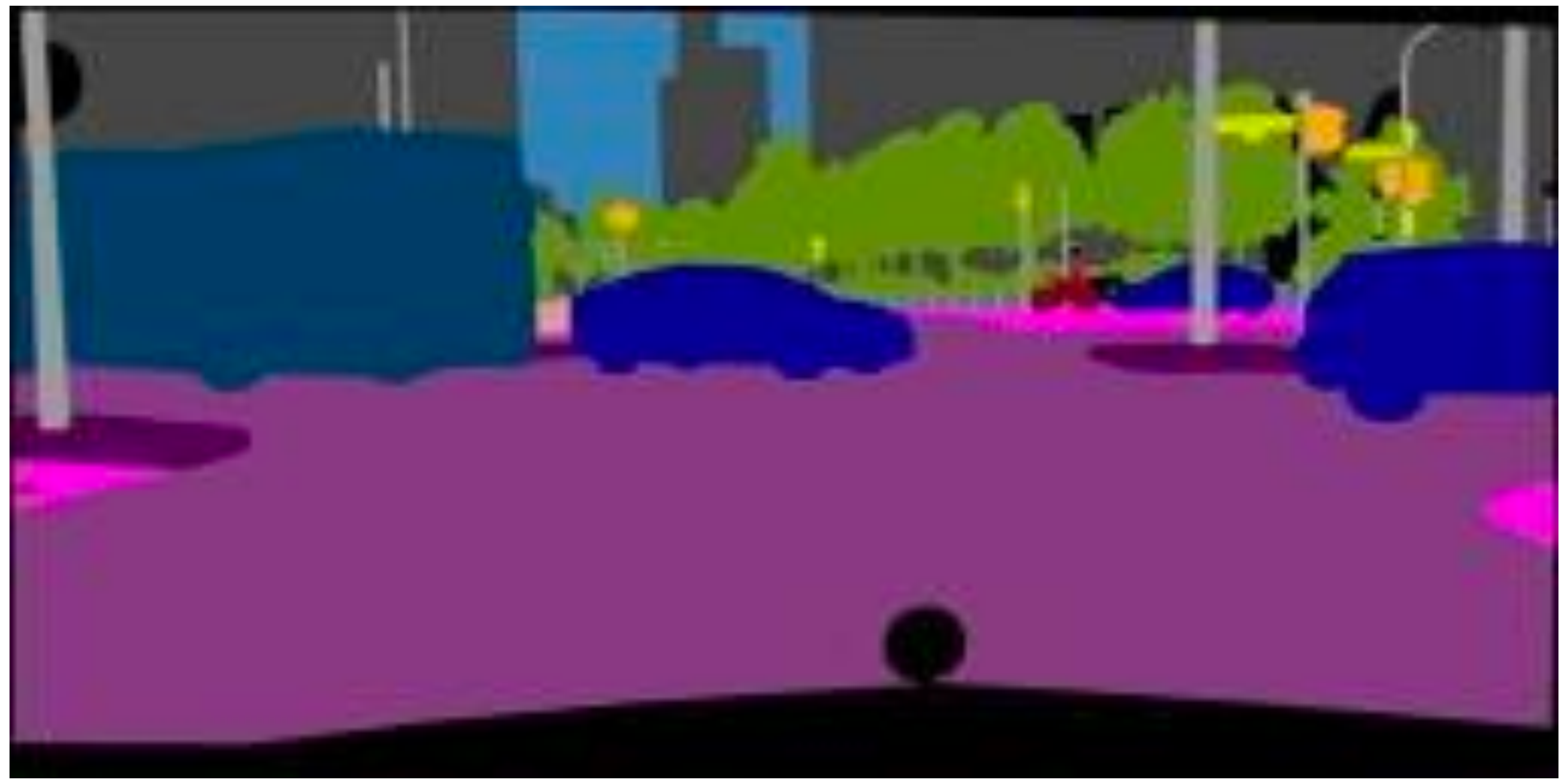


Input



L1

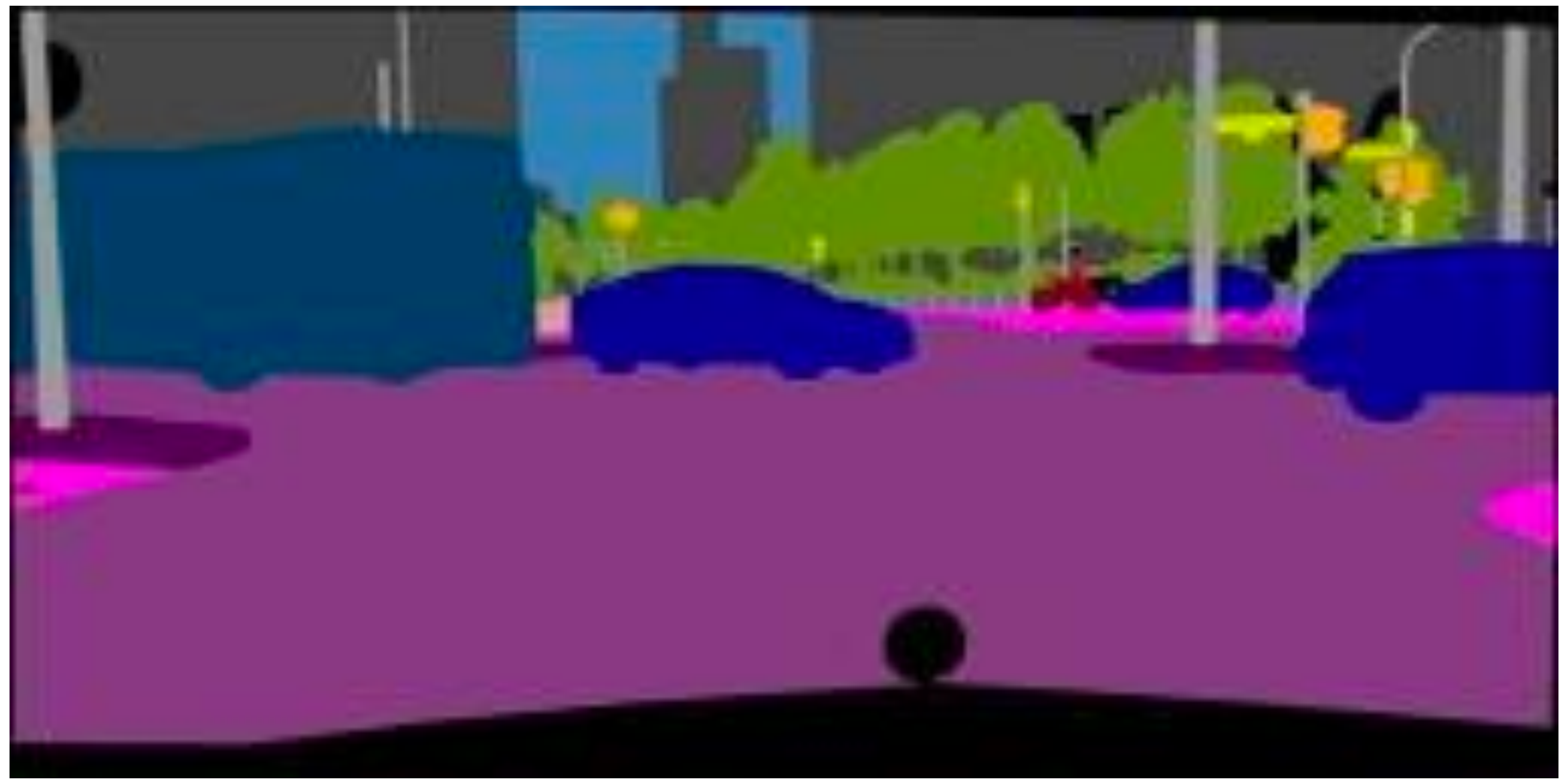
Input



L1

1x1 Discriminator

Input



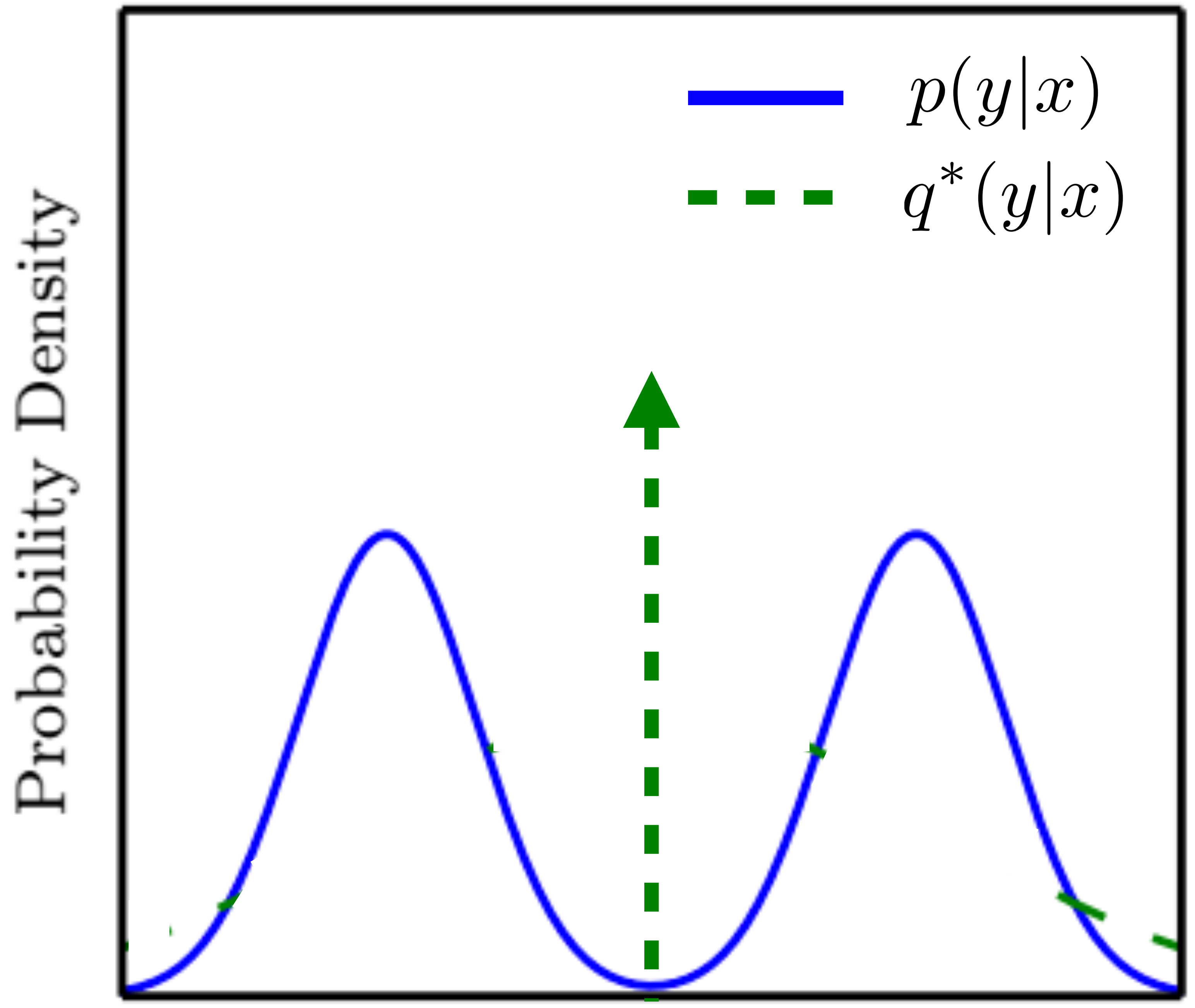
L1

1x1 Discriminator

“Unstructured” discriminator makes images colorful!

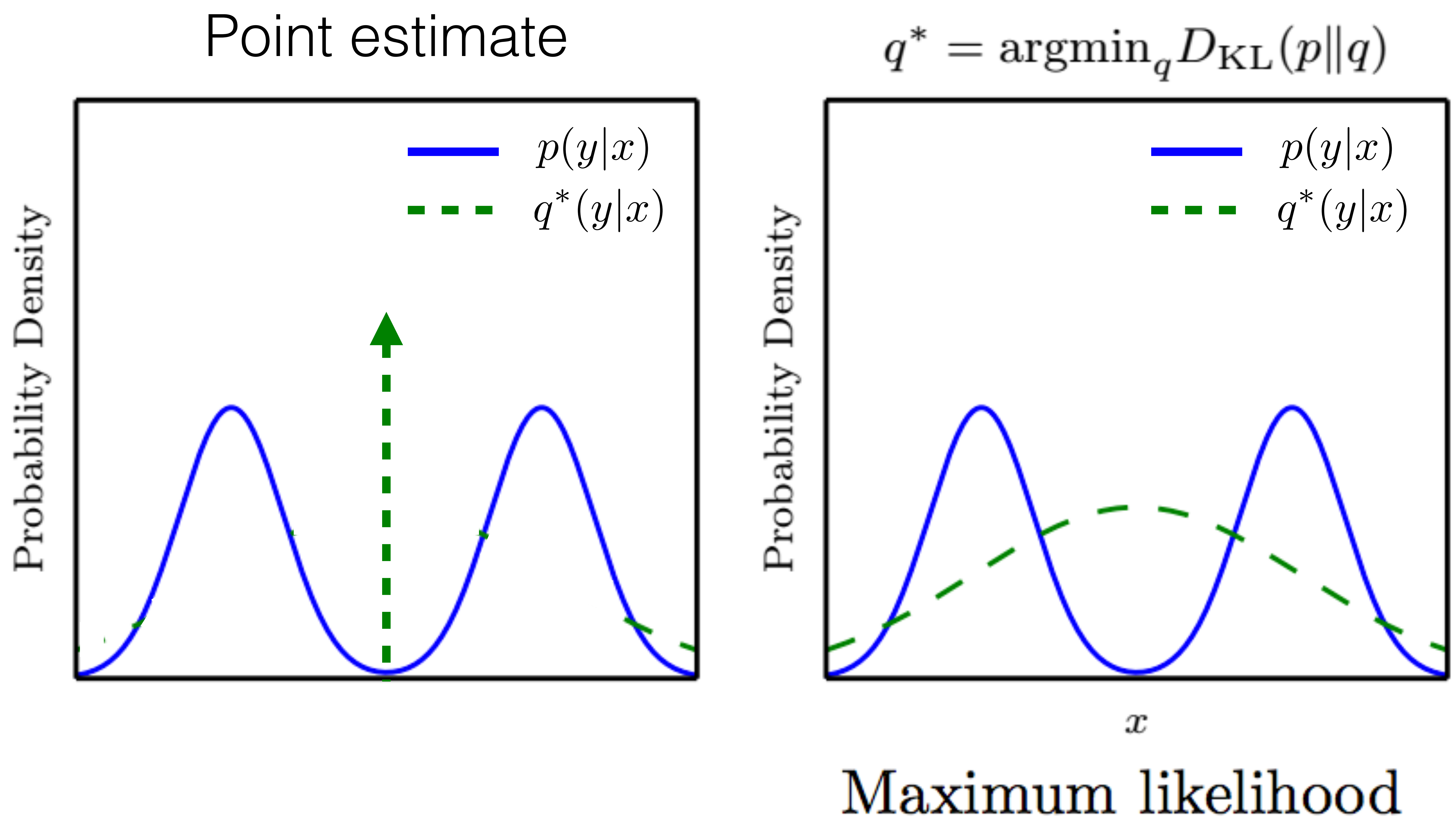
Mode seeking property

Point estimate



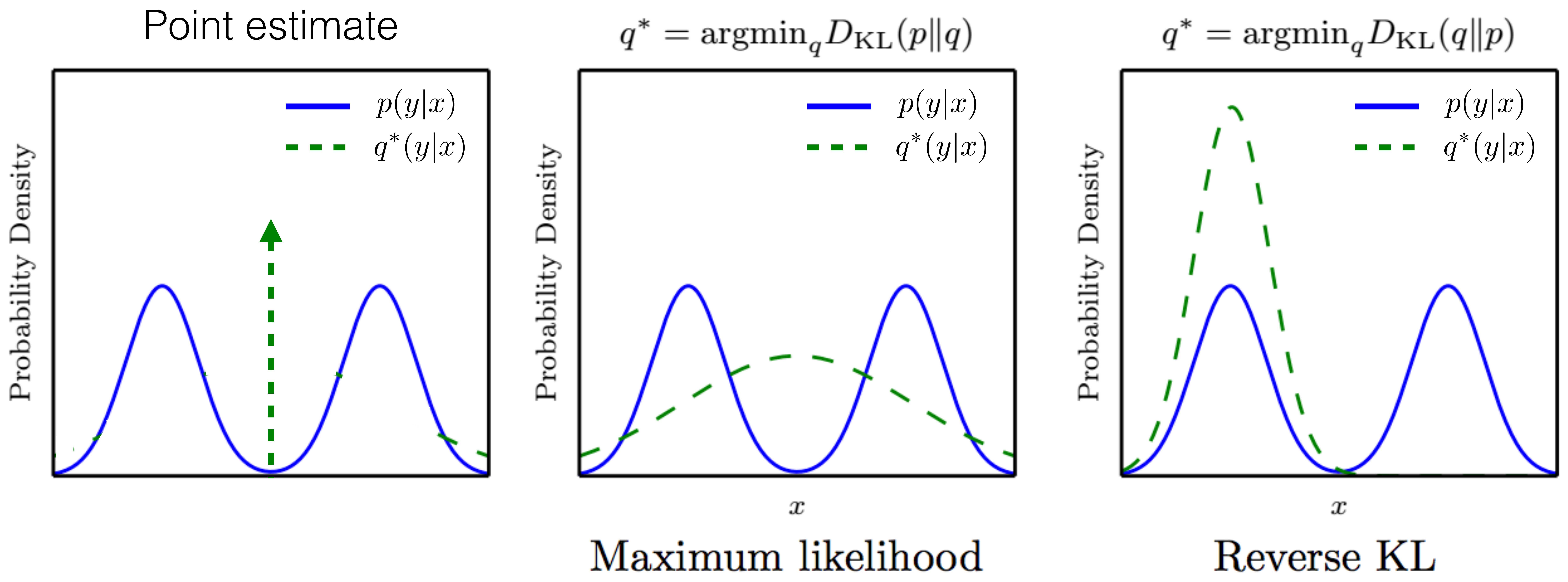
adapted from [Goodfellow, 2016]

Mode seeking property



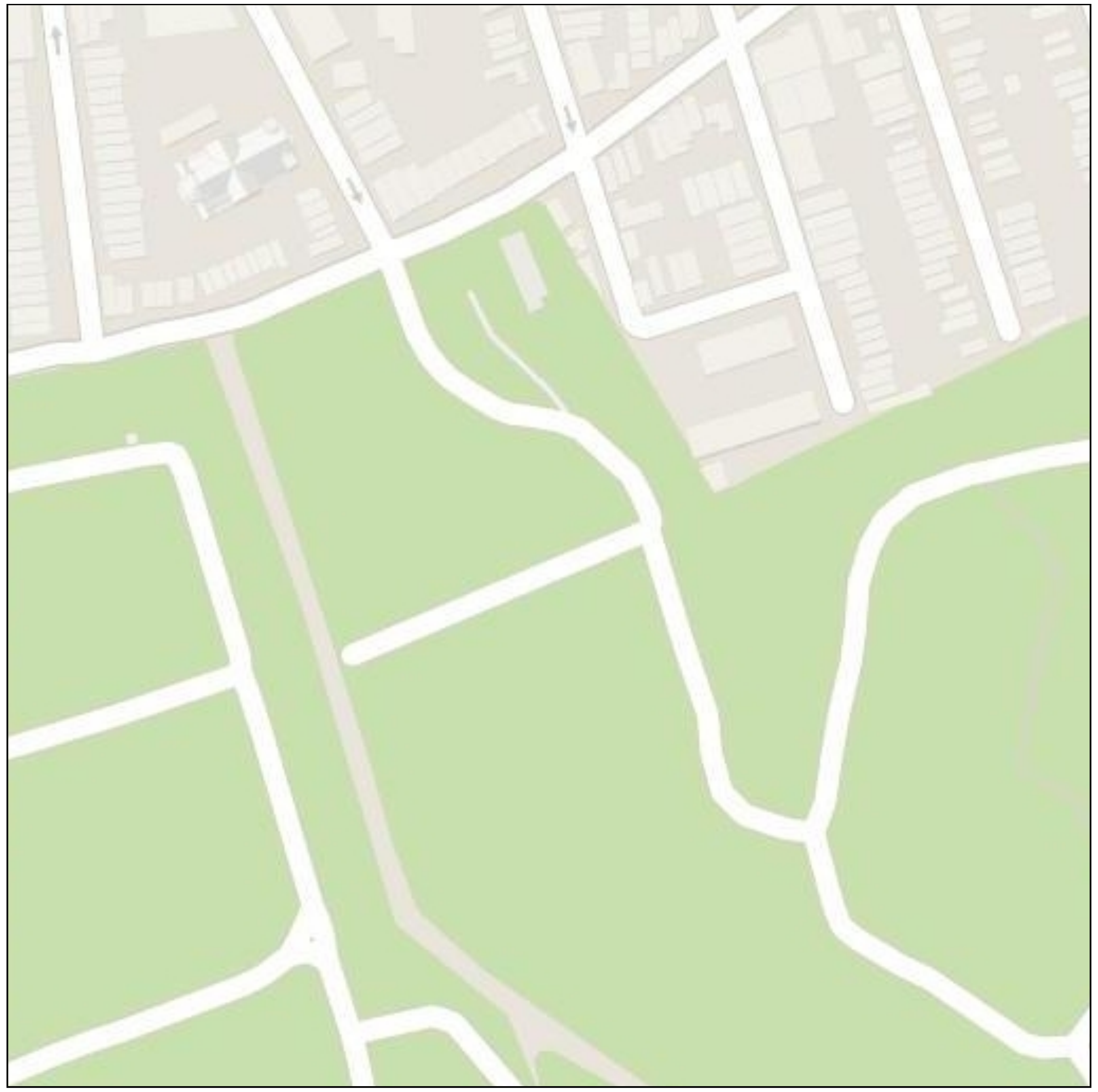
adapted from [Goodfellow, 2016]

Mode seeking property

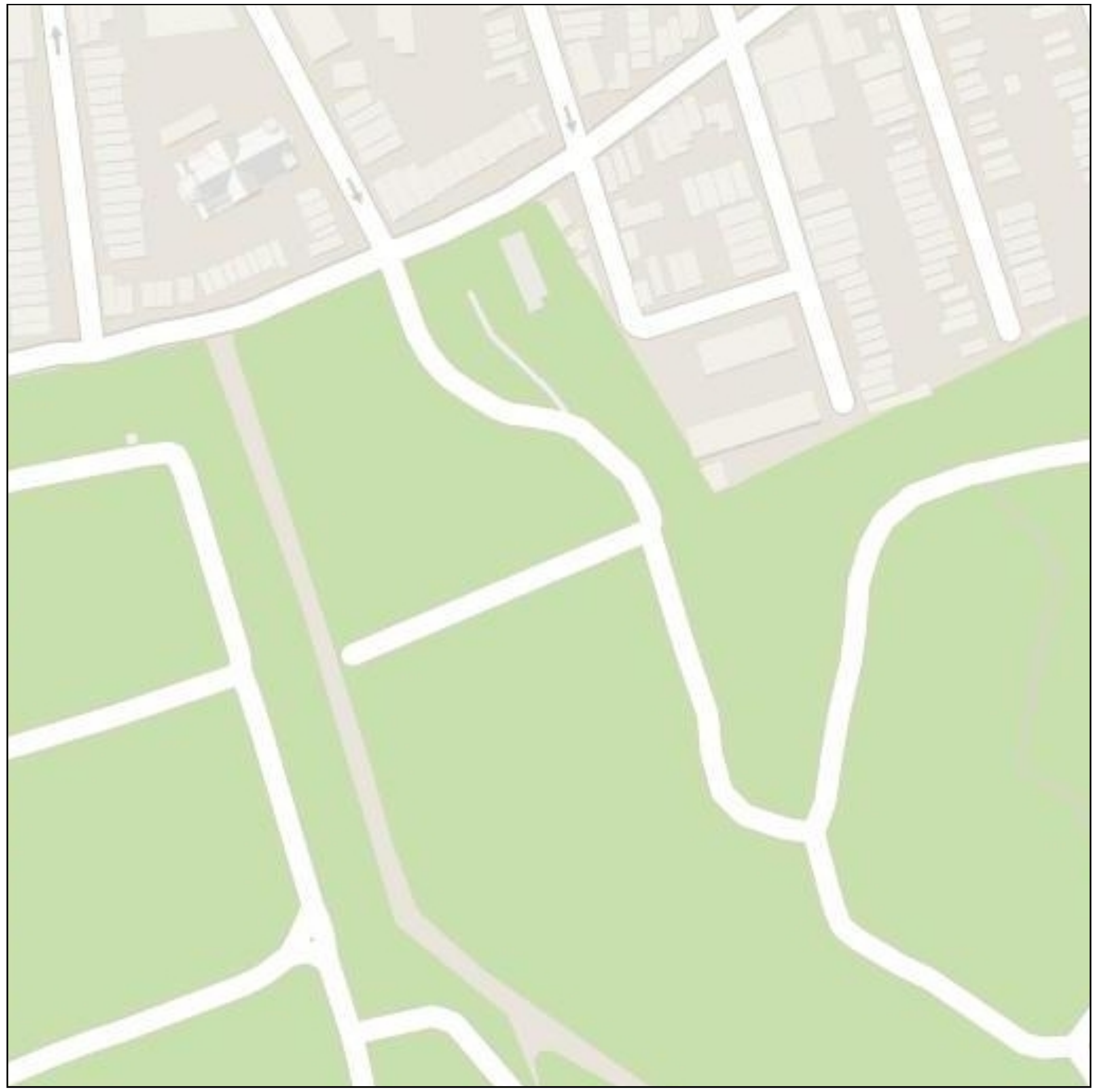


adapted from [Goodfellow, 2016]

Input

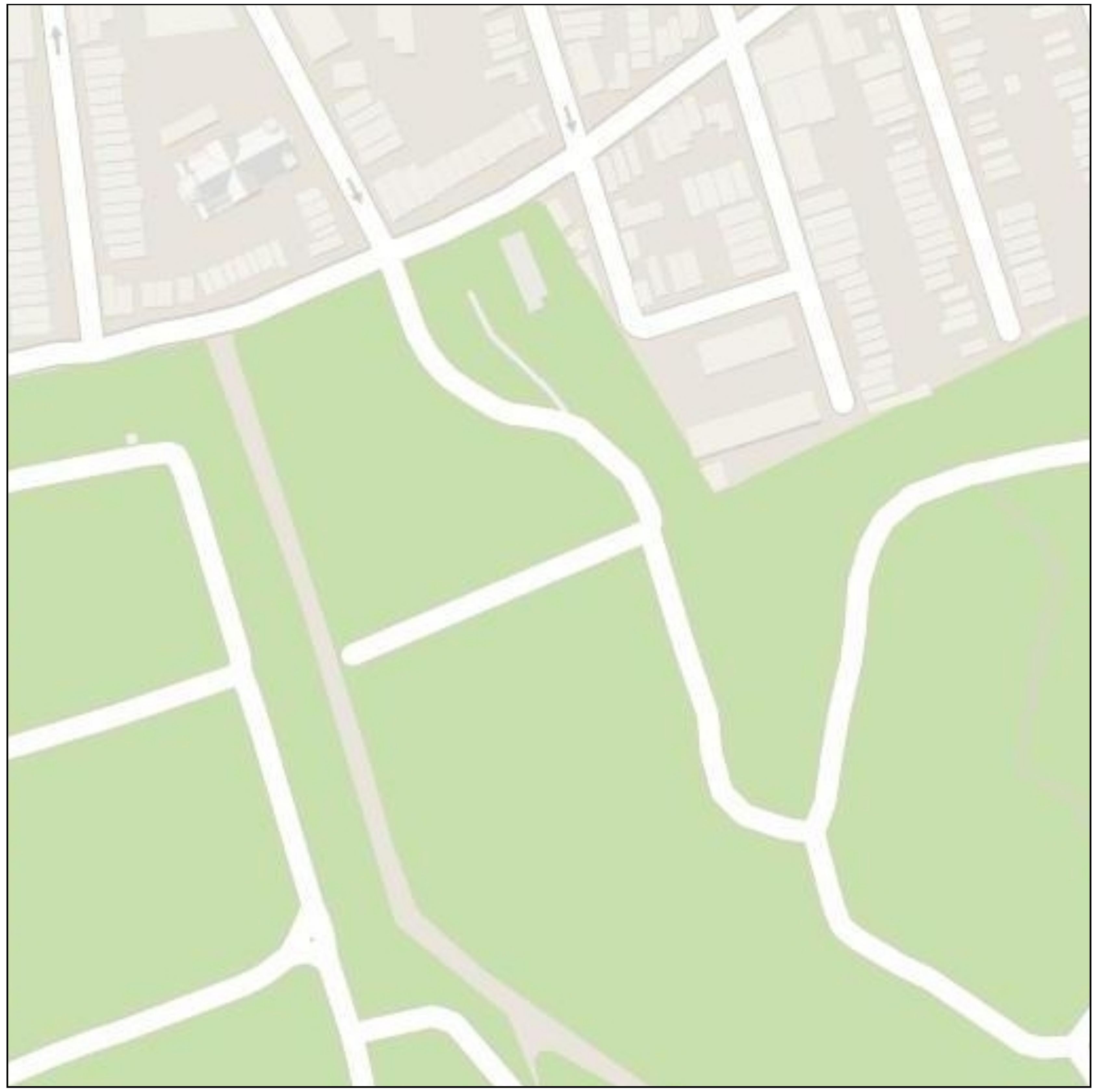


Input



Output

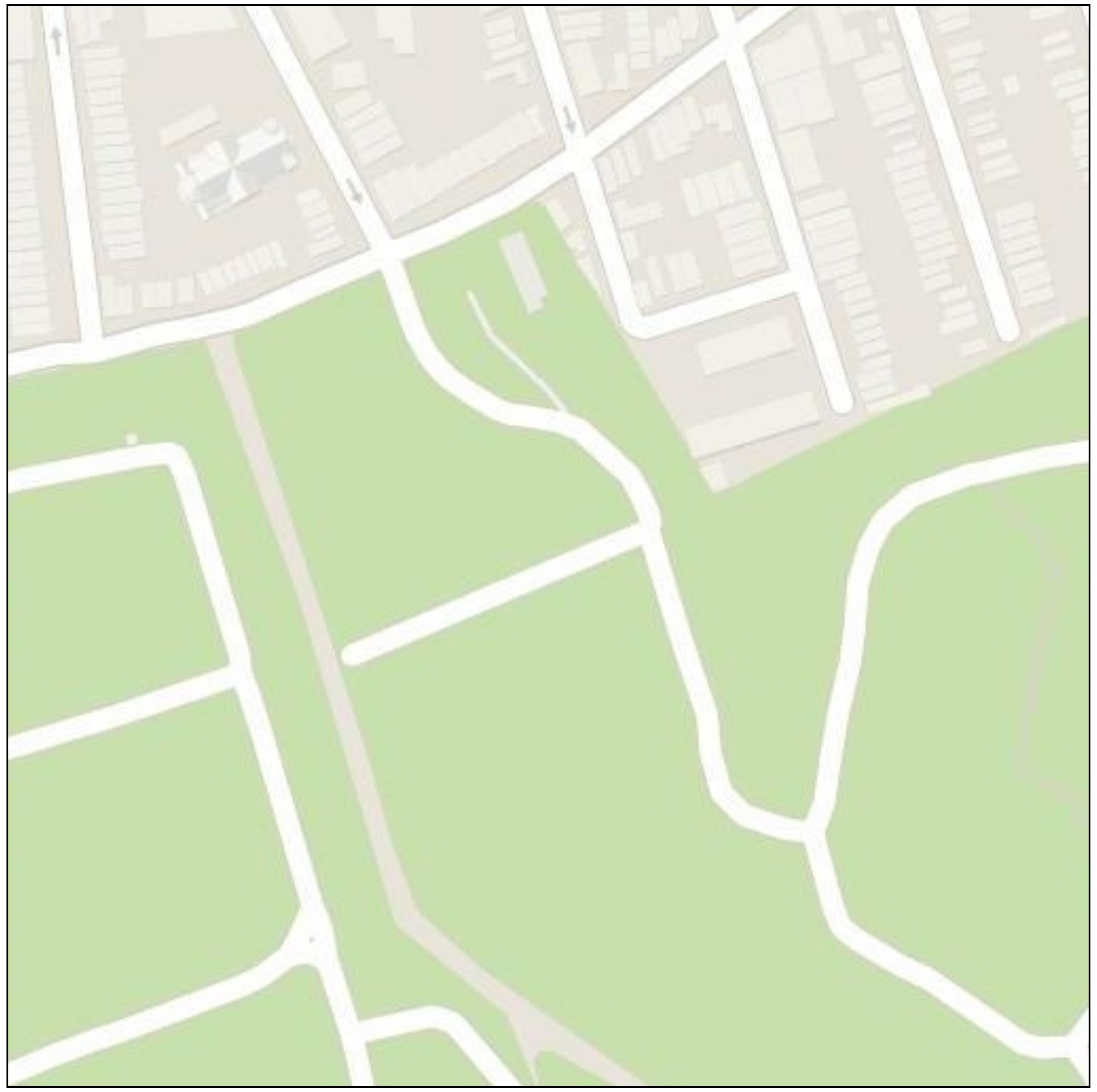
Input



Output

Groundtruth

Input

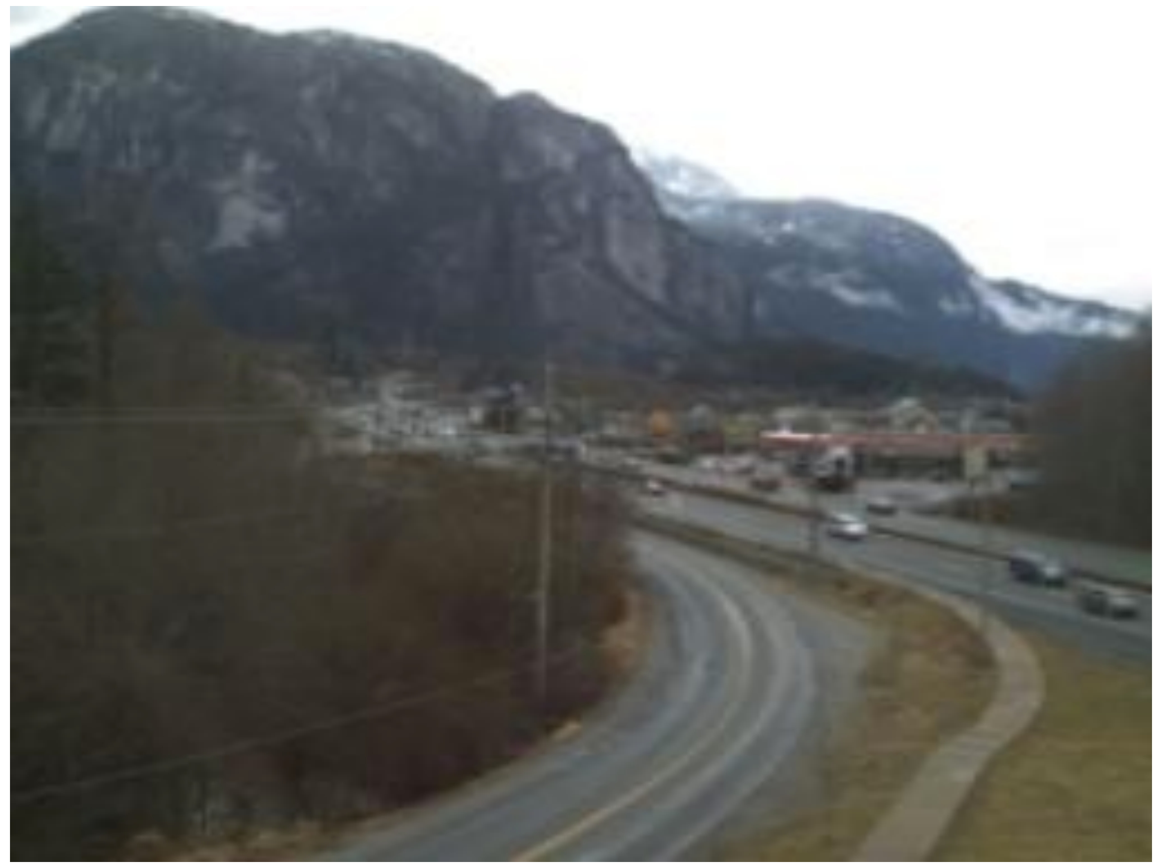


L1 Output

Groundtruth

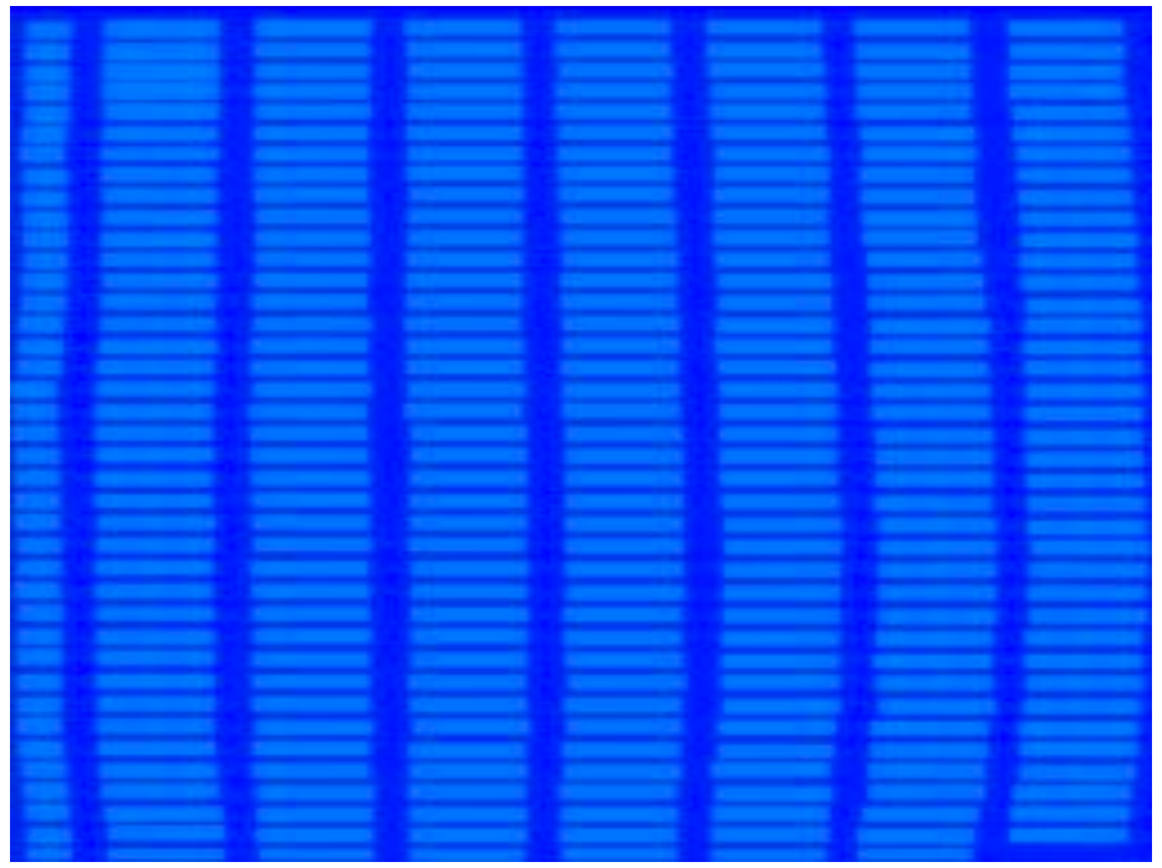
Hallucinations

Input

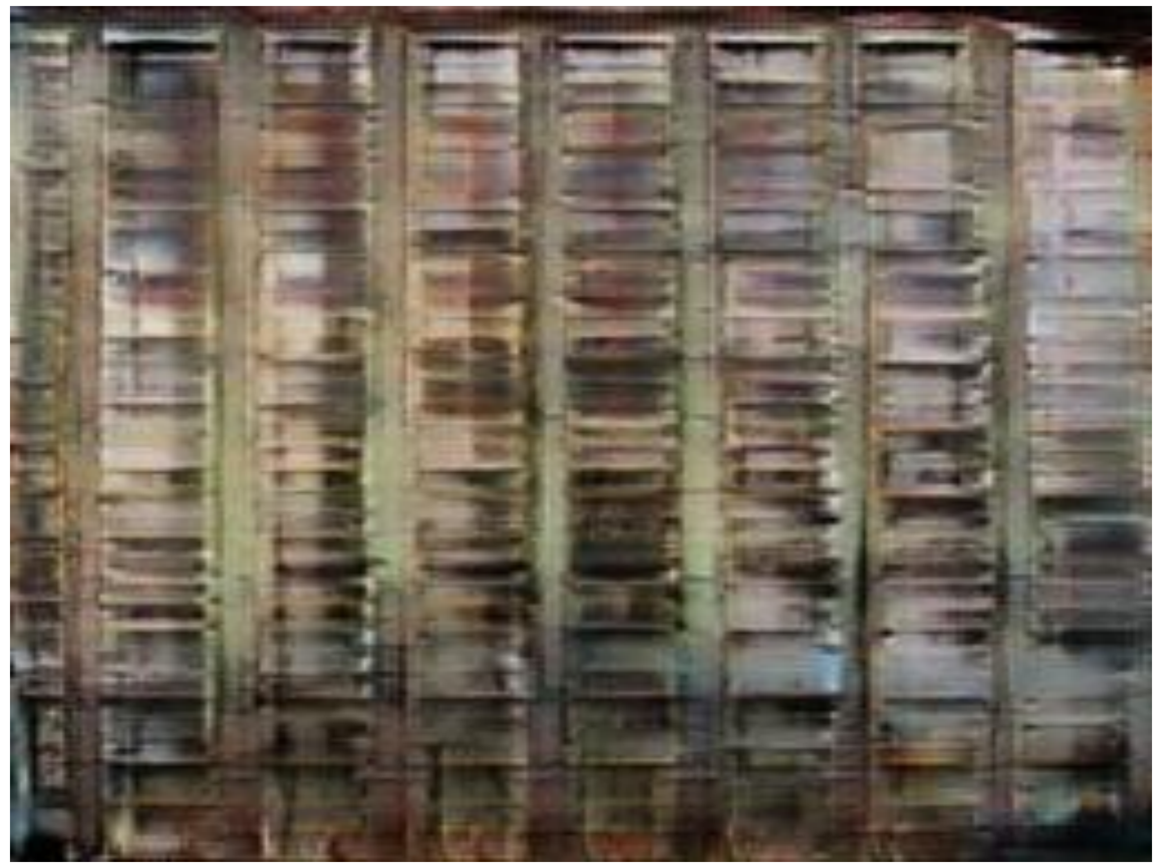


Output

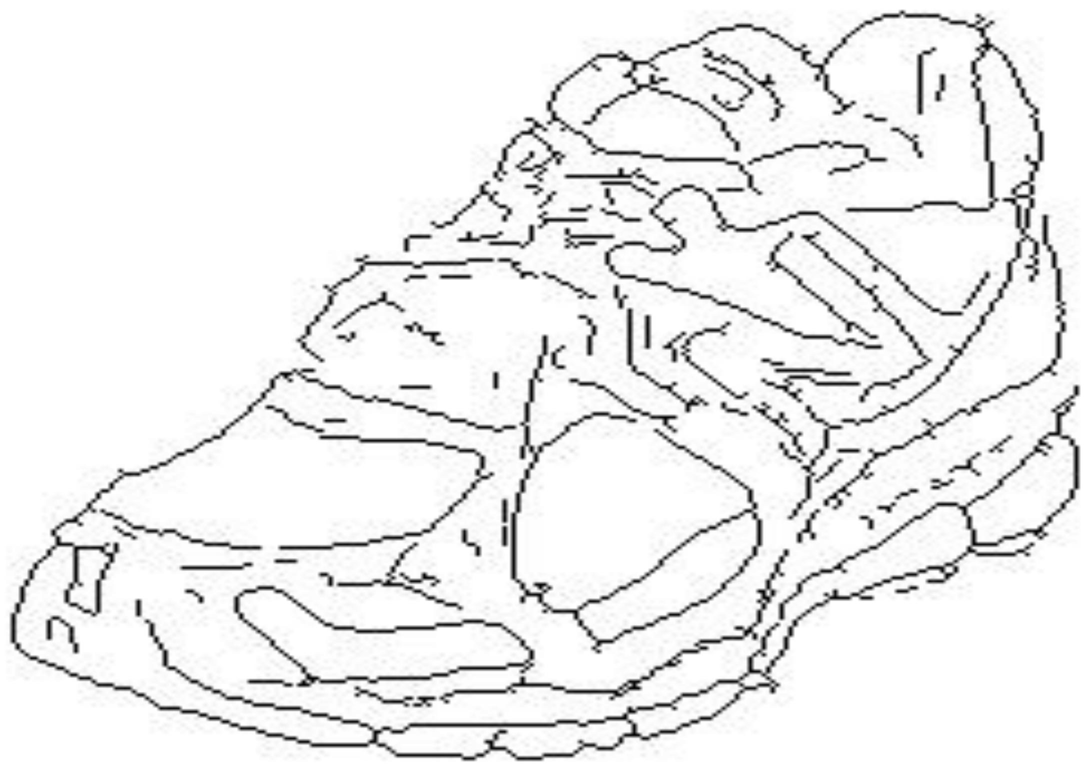
Input



Output

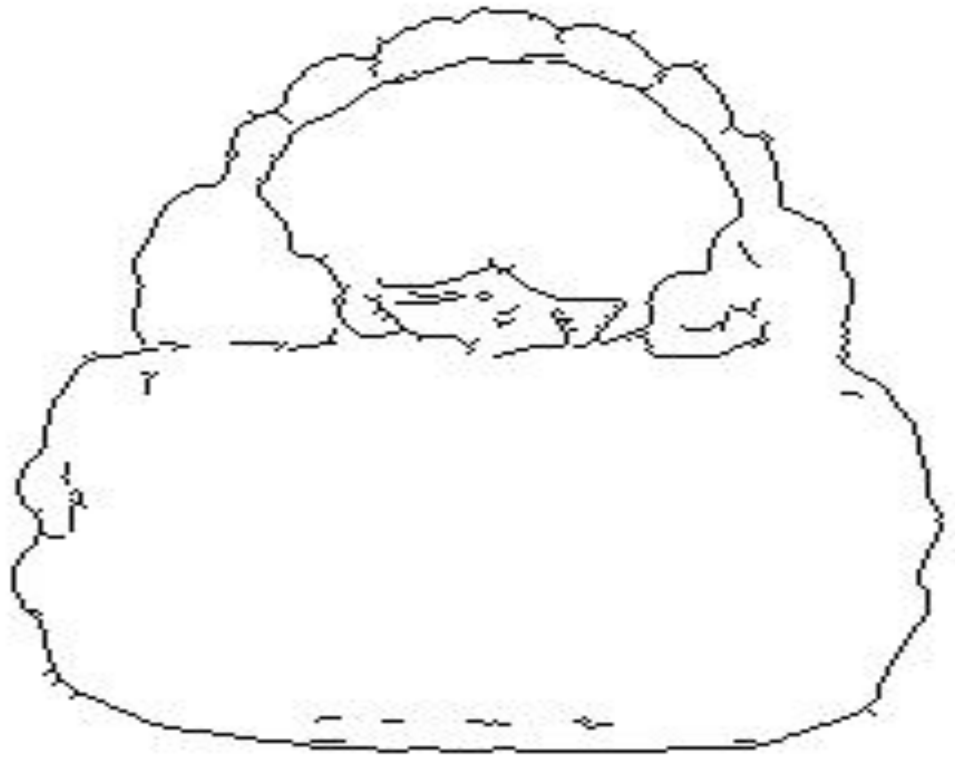


Input



Output

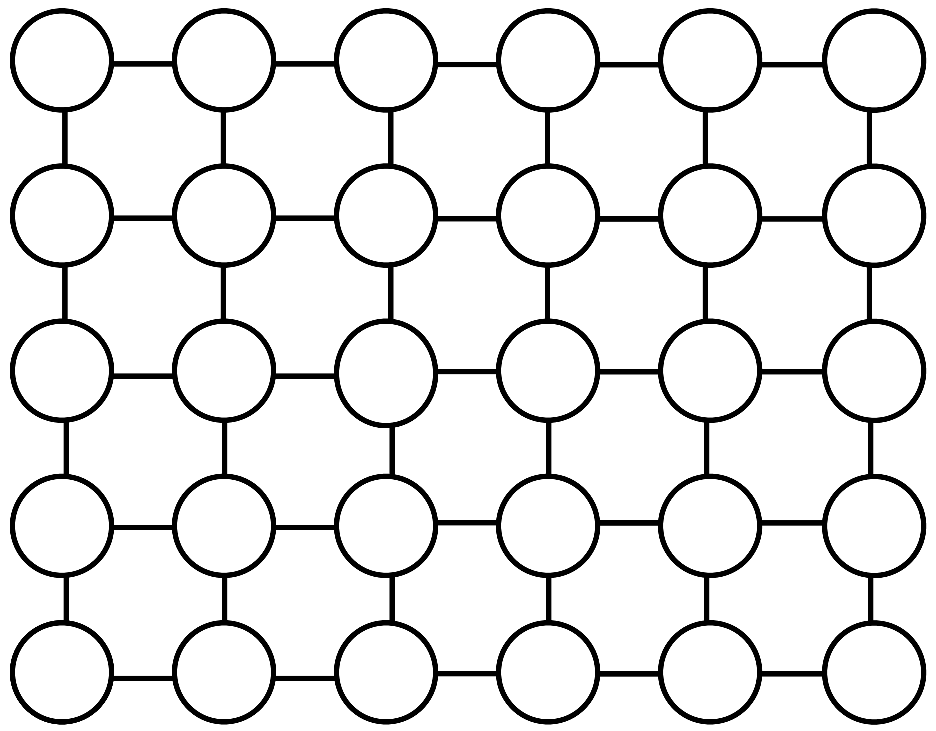
Input



Output

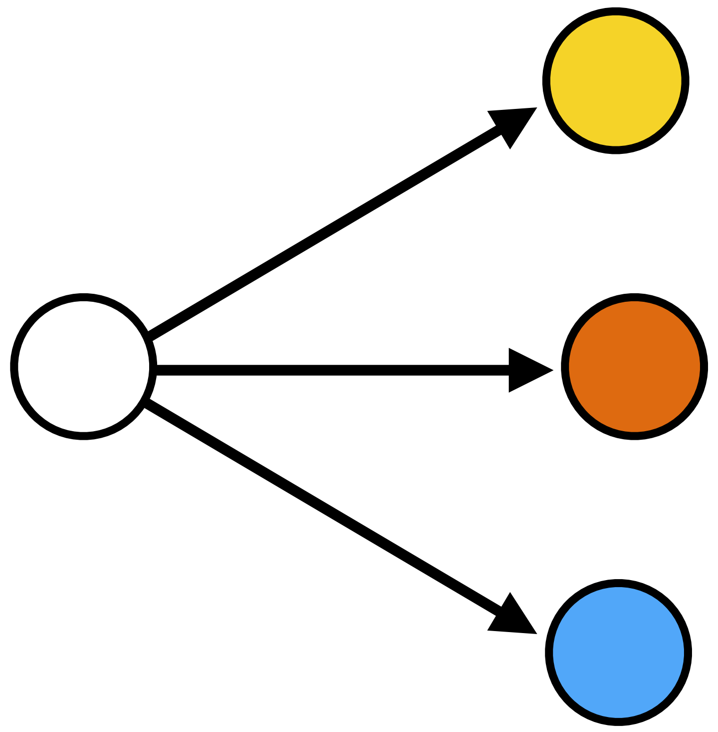
Challenges in image-to-image translation

1. Output is high-dimensional, structured object



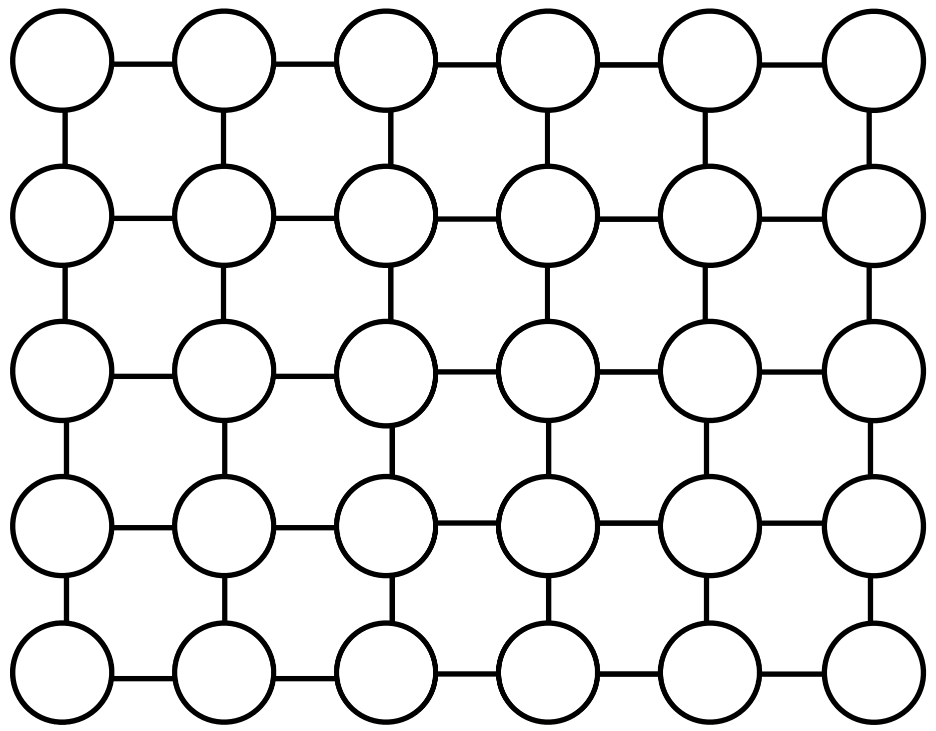
—> Use a deep net, D, to analyze output!

2. Uncertainty in mapping; many plausible outputs



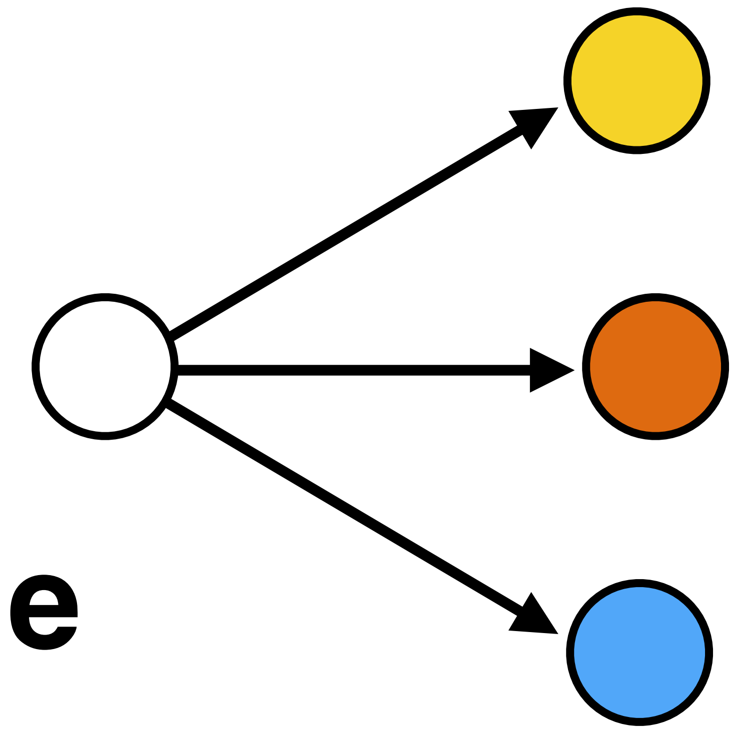
Challenges in image-to-image translation

1. Output is high-dimensional, structured object



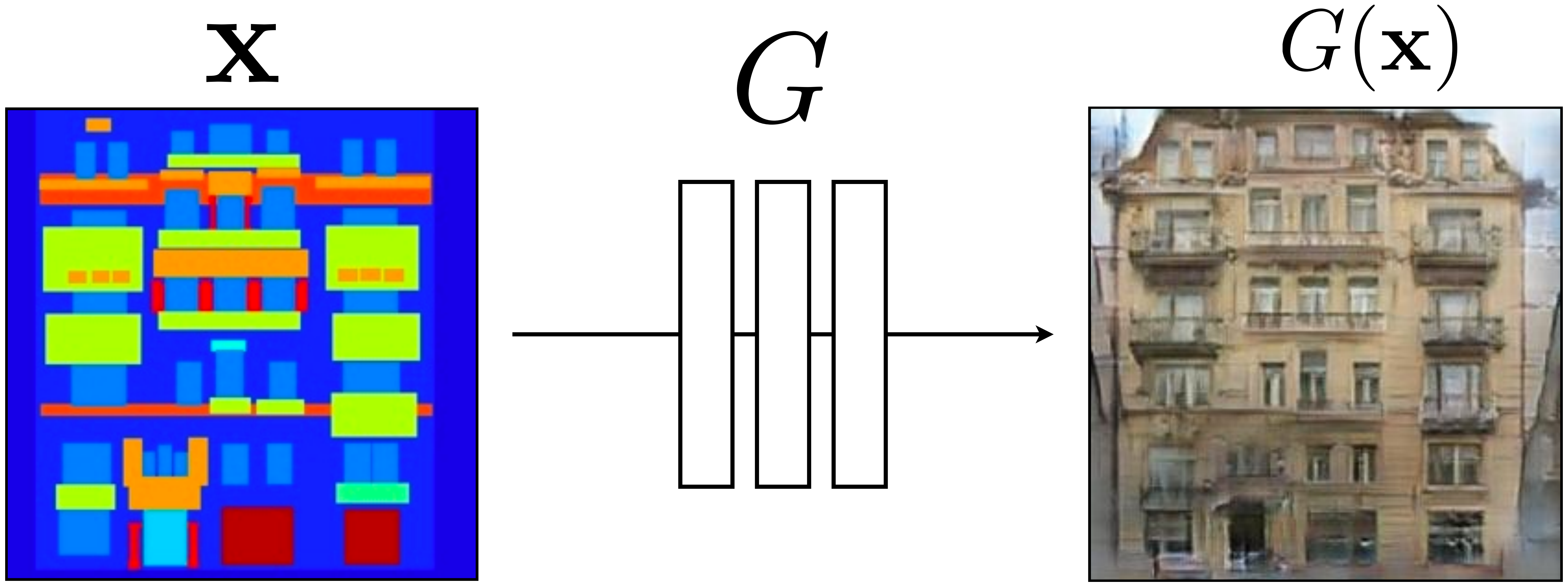
—> Use a deep net, D, to analyze output!

2. Uncertainty in mapping; many plausible outputs

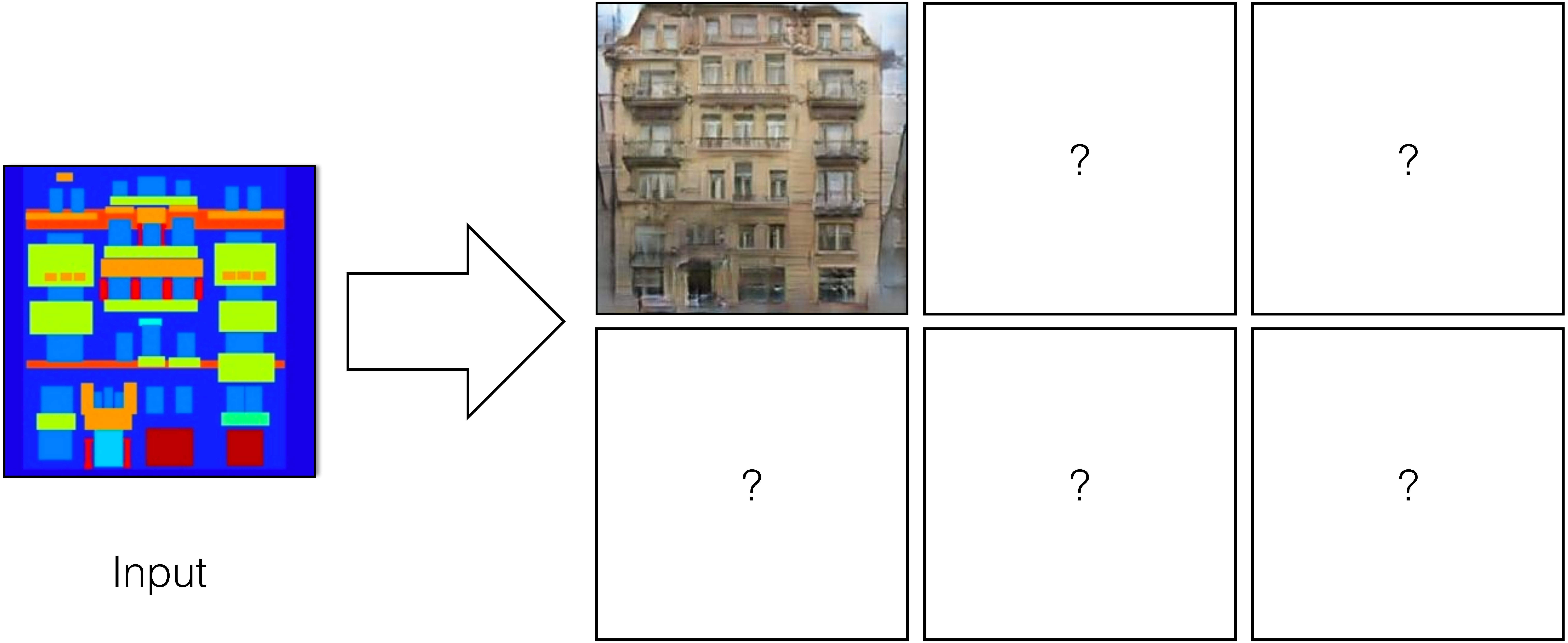


—> D only cares about "plausibility", doesn't hedge

Modeling multiple possible outputs



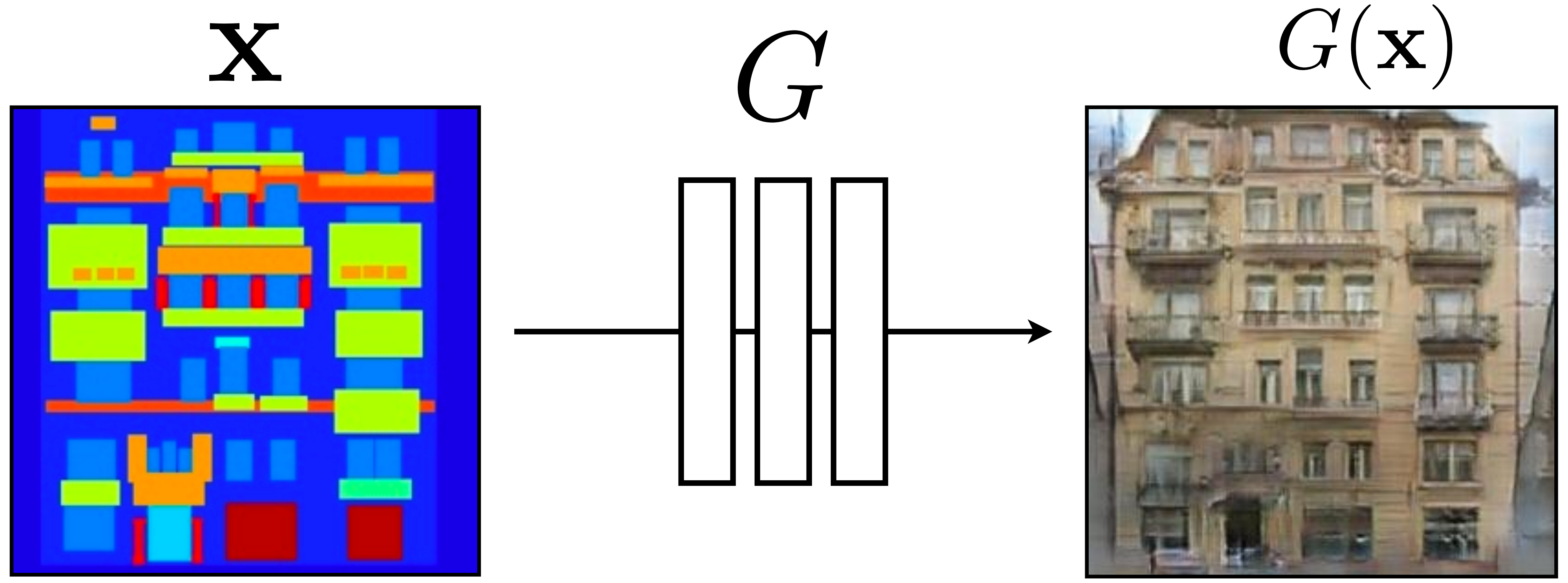
Modeling multiple possible outputs



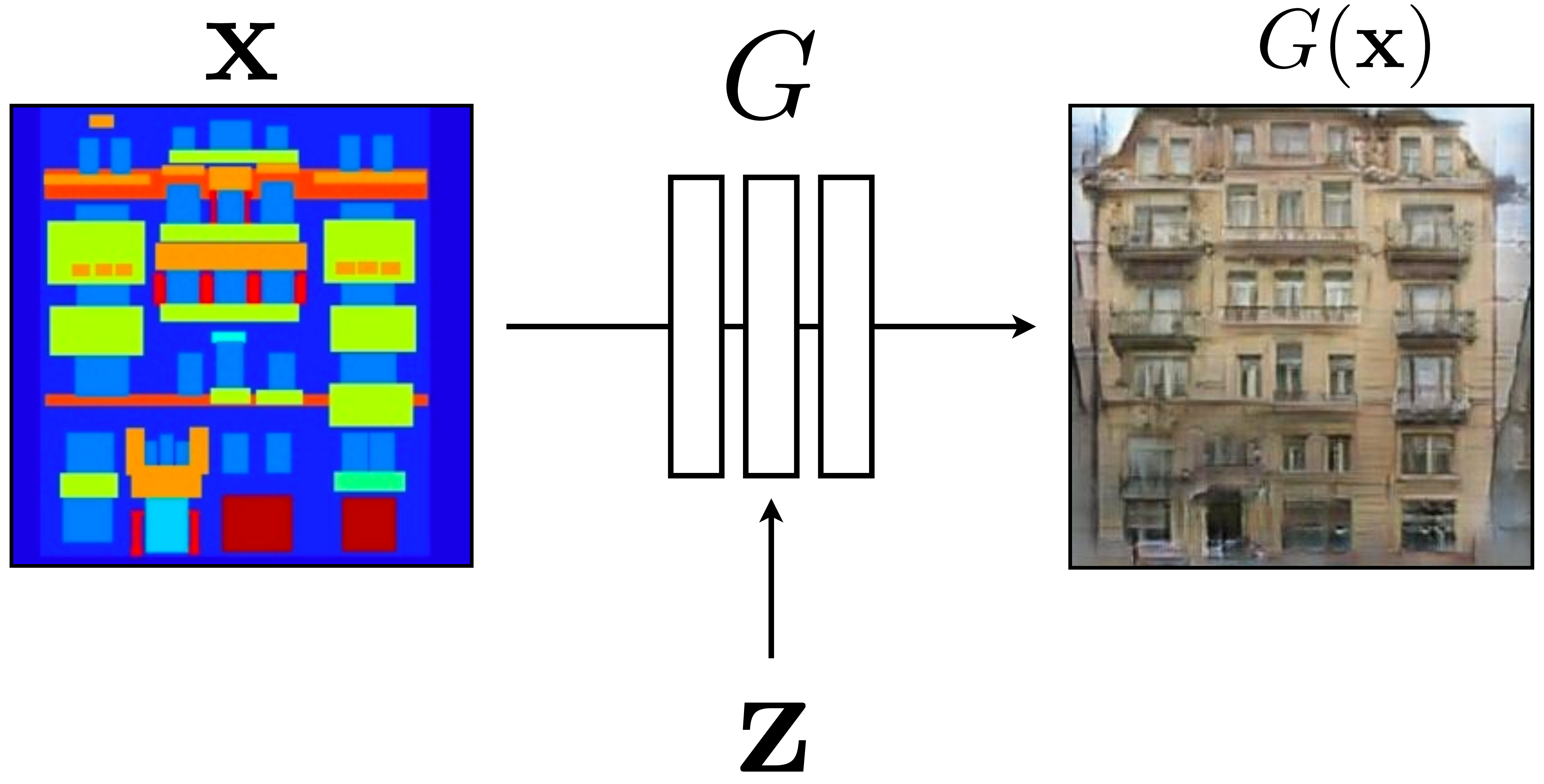
Input

Possible outputs

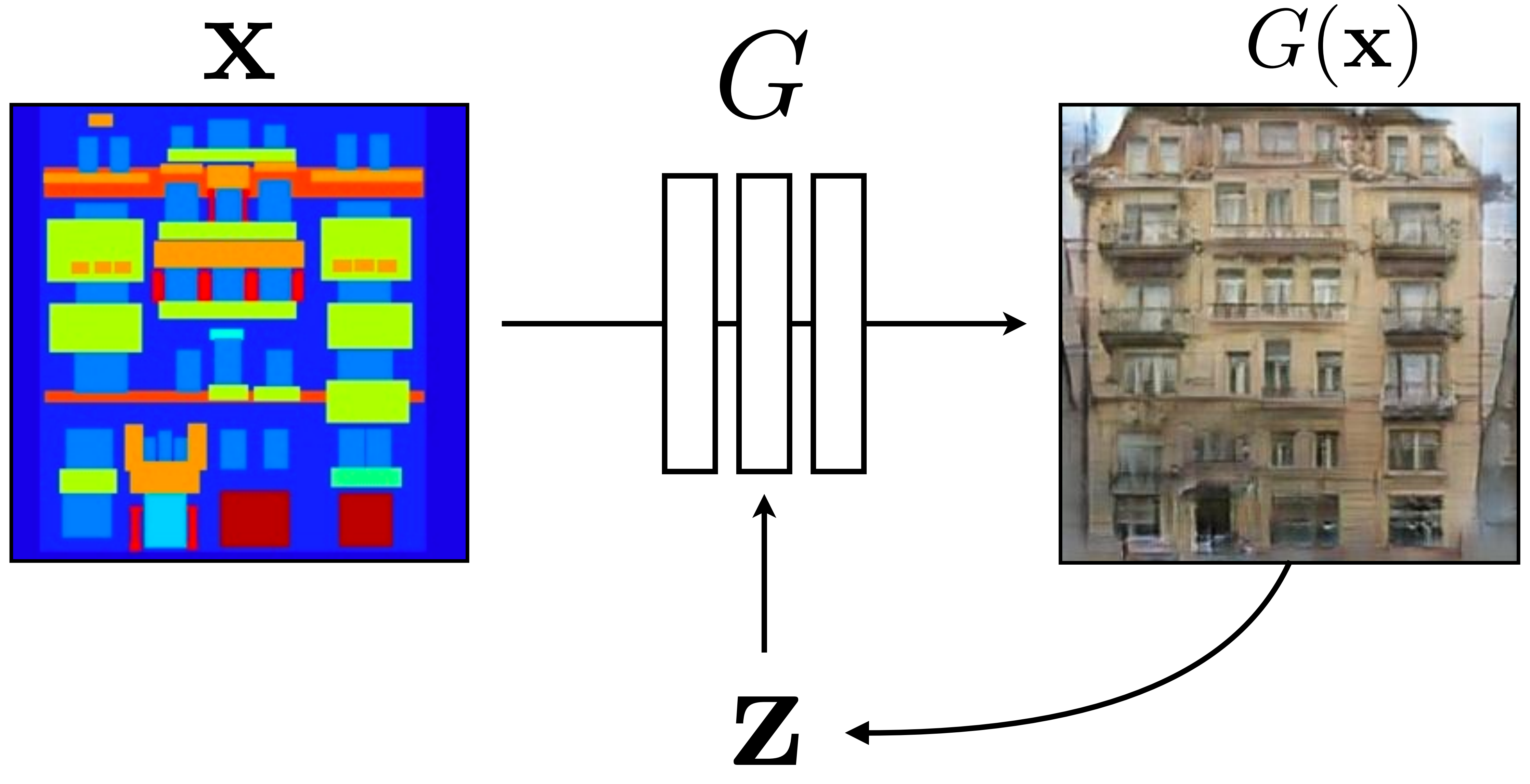
BiCycleGAN [Zhu et al., NIPS 2017]
(c.f. InfoGAN [Chen et al. 2016])



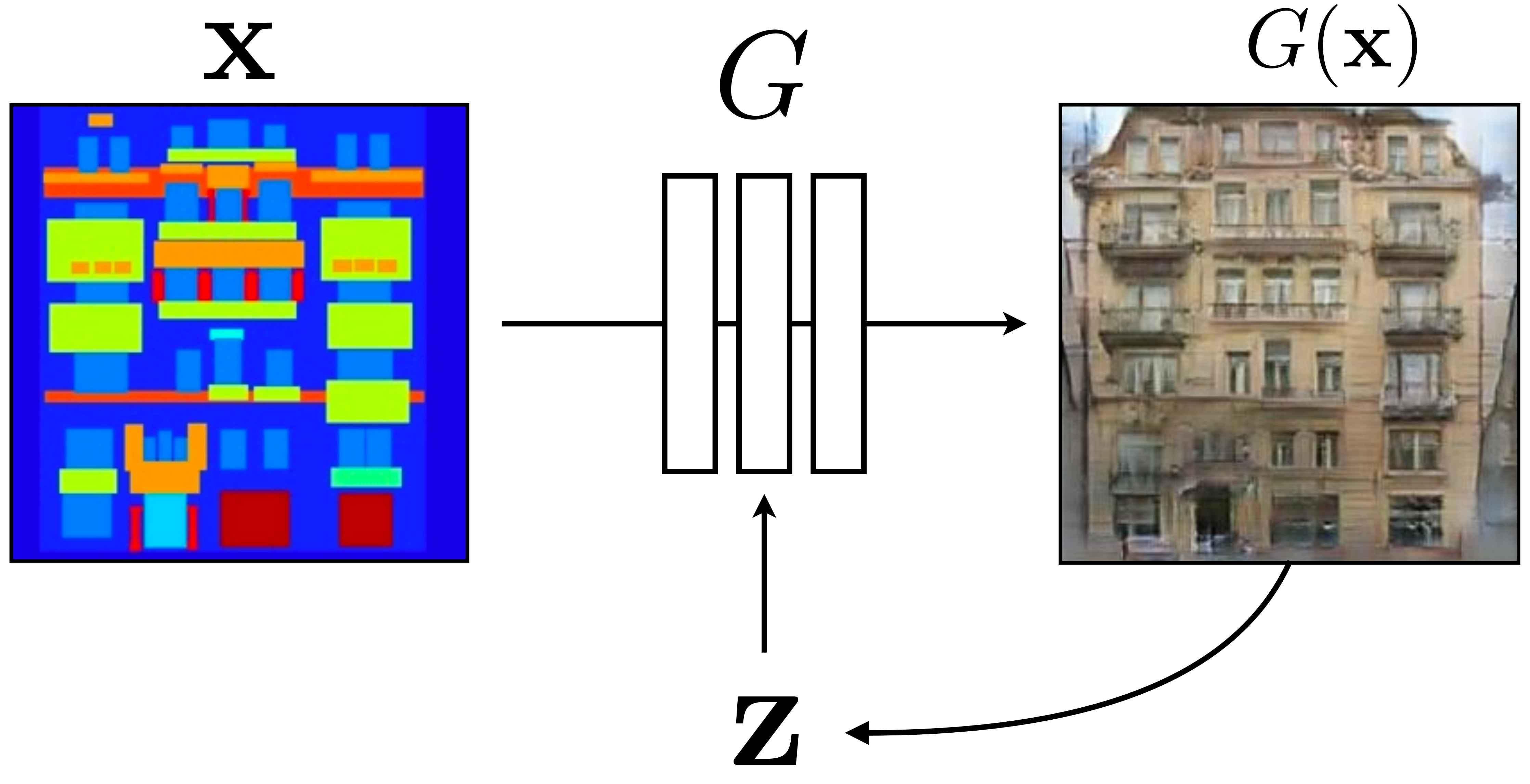
BiCycleGAN [Zhu et al., NIPS 2017]
(c.f. InfoGAN [Chen et al. 2016])



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

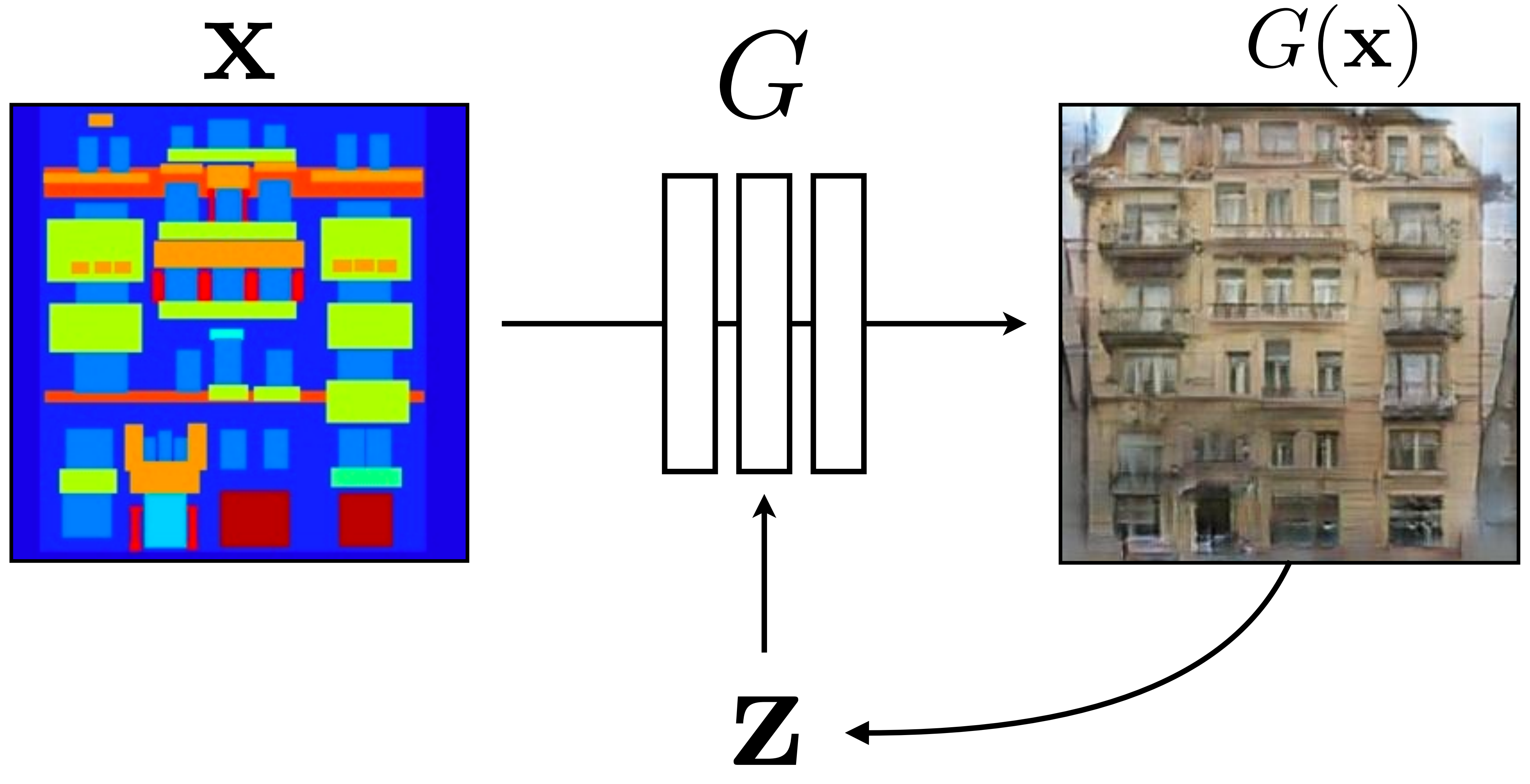


BiCycleGAN [Zhu et al., NIPS 2017]
(c.f. InfoGAN [Chen et al. 2016])

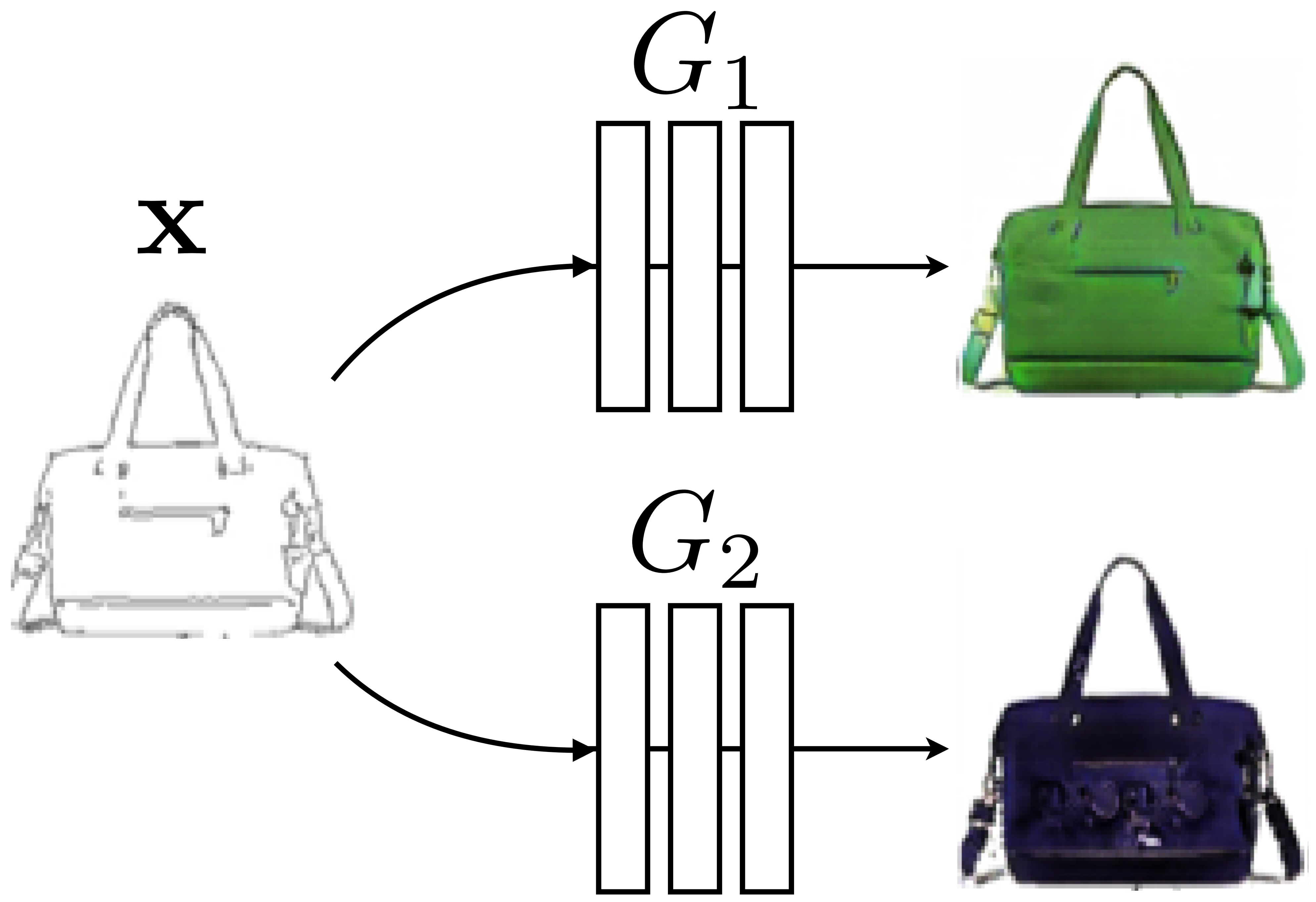


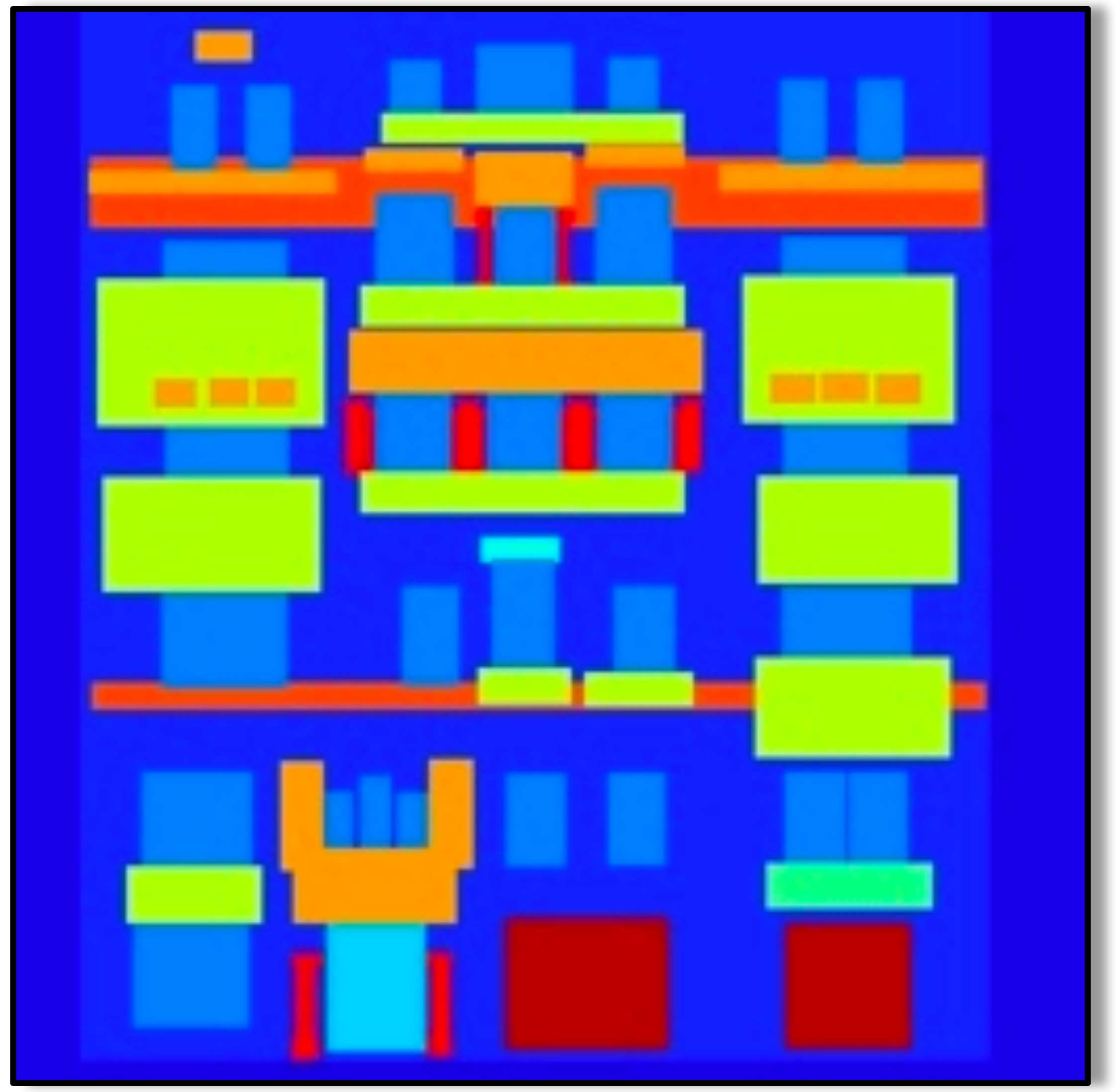
MAD-GAN [Ghosh et al., CVPR 2018]

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



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



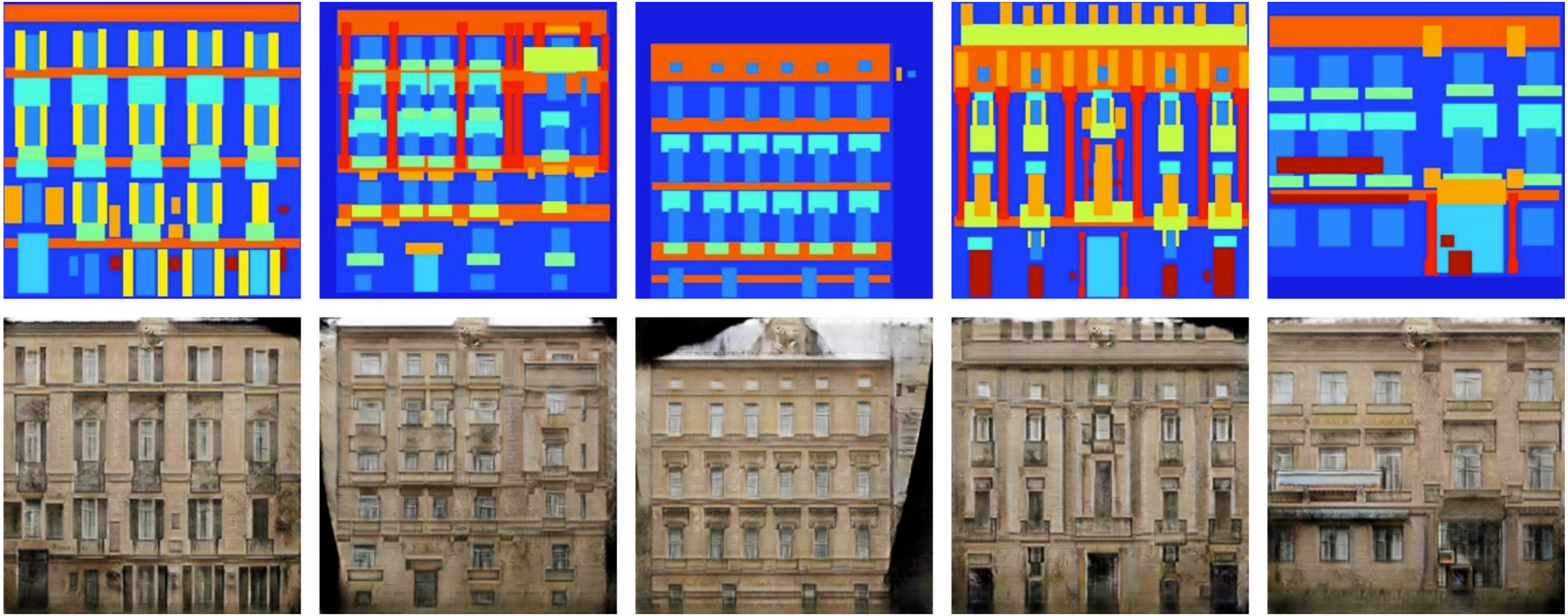


Labels

Randomly generated facades

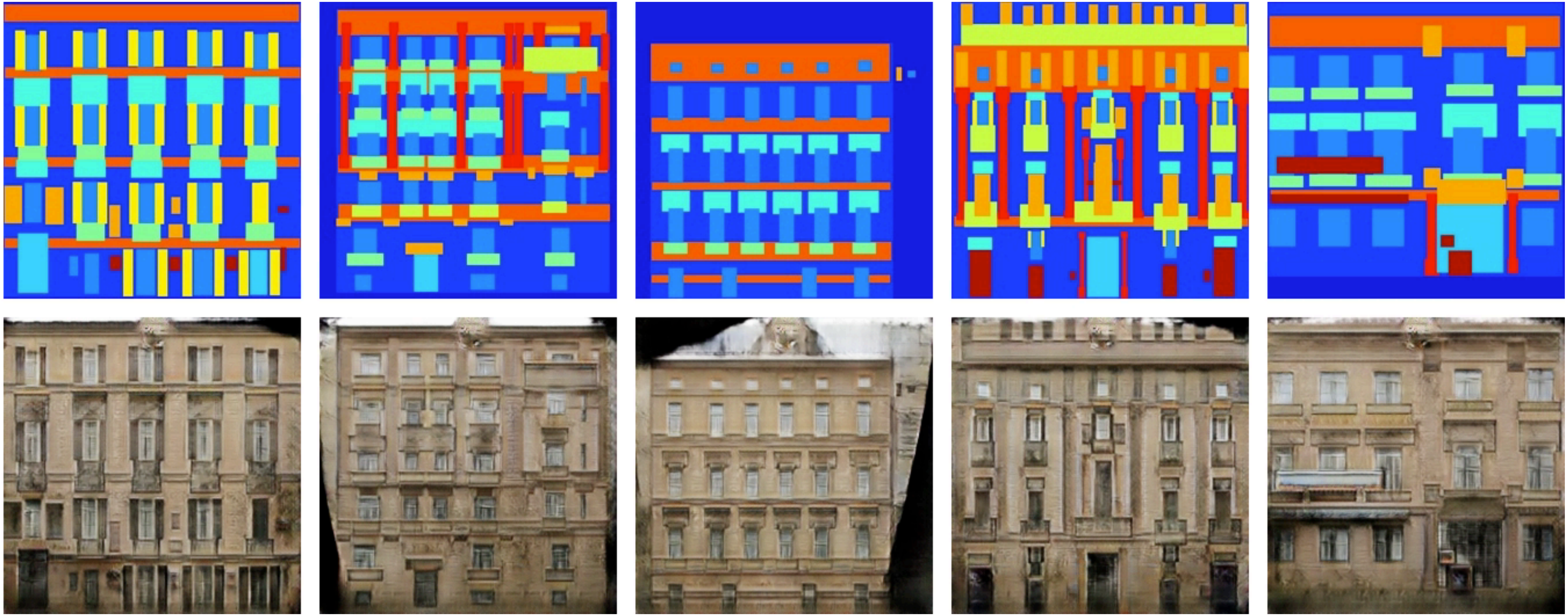
[BiCycleGAN, Zhu et al., NIPS 2017]

Latent space exploration



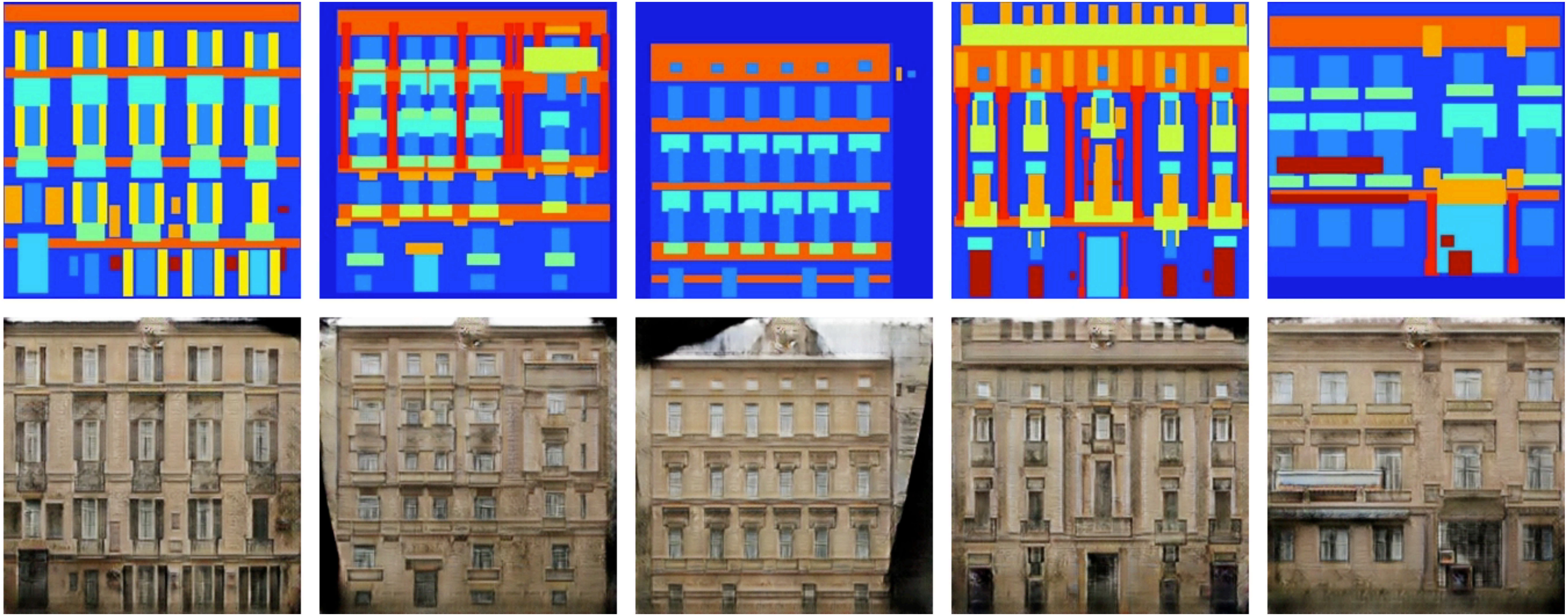
[BiCycleGAN, Zhu et al., NIPS 2017]

Latent space exploration



[BiCycleGAN, Zhu et al., NIPS 2017]

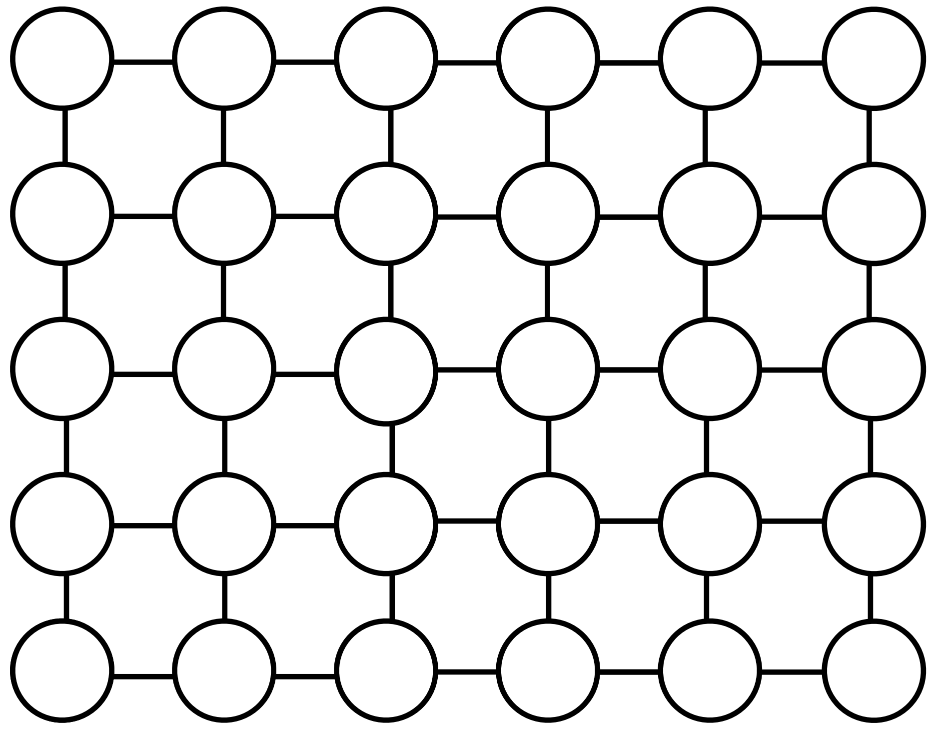
Latent space exploration



[BiCycleGAN, Zhu et al., NIPS 2017]

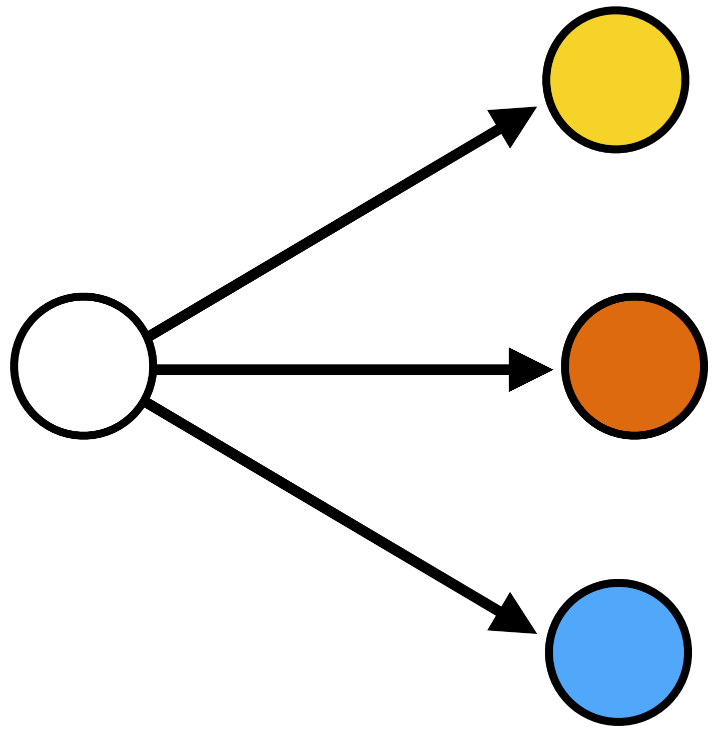
Challenges in image-to-image translation

1. Output is high-dimensional, structured object



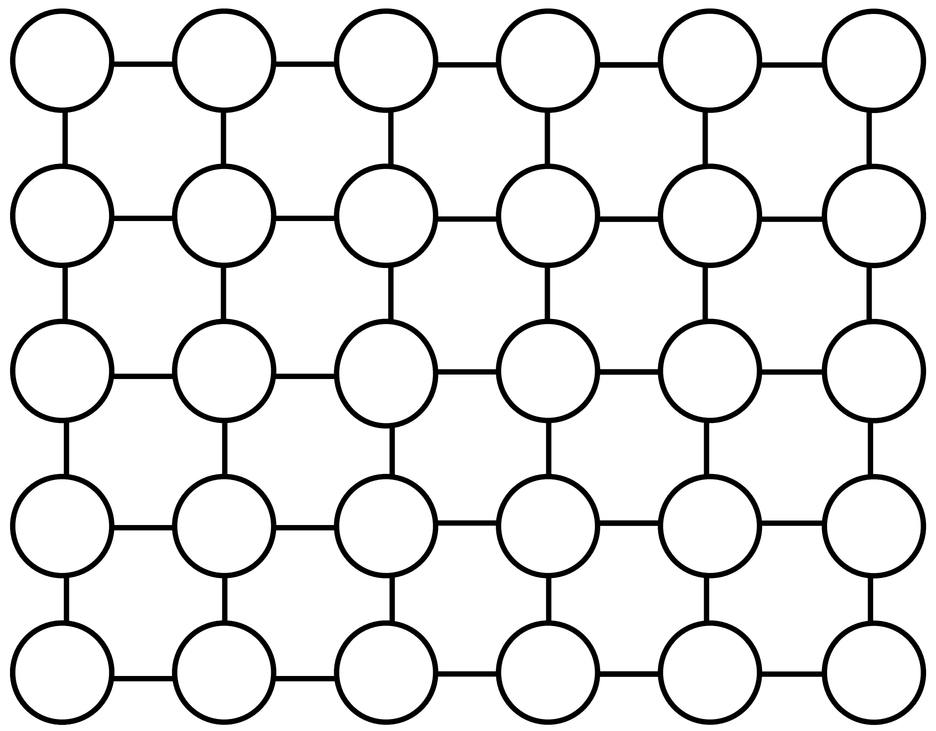
—> Use a deep net, D, to analyze output!

2. Uncertainty in mapping; many plausible outputs



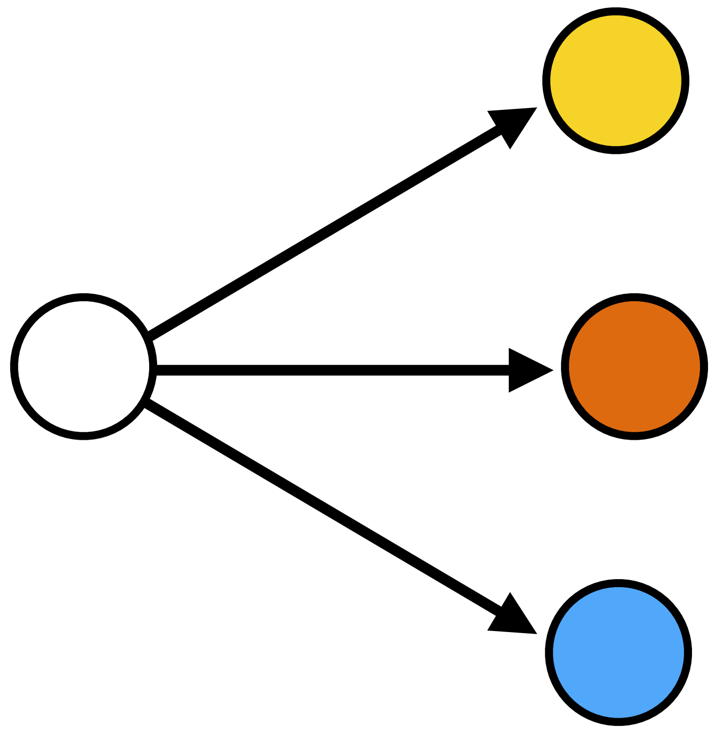
Challenges in image-to-image translation

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

Questions?