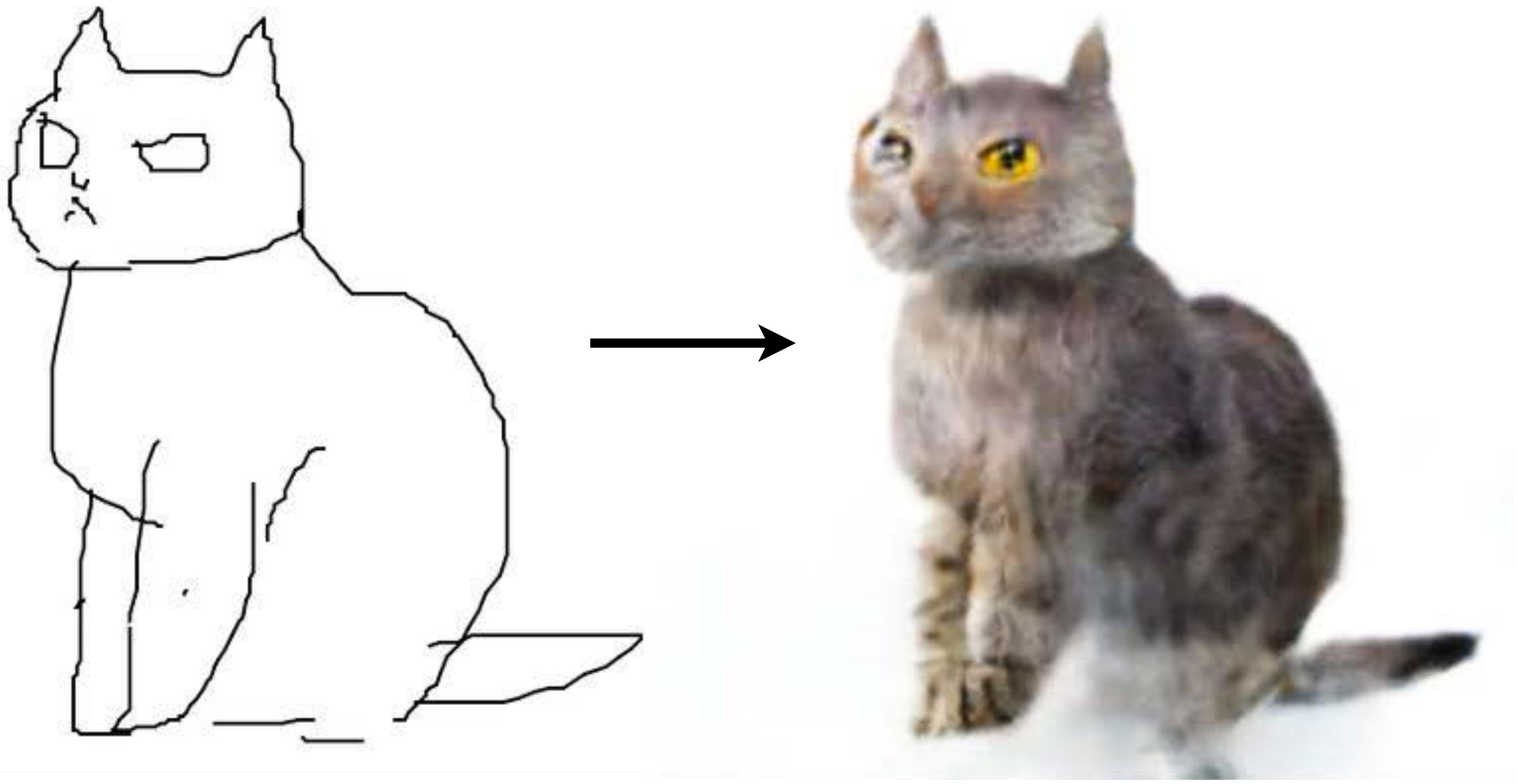
Lecture 16: Deep Generative Models III

Speaker: Phillip Isola



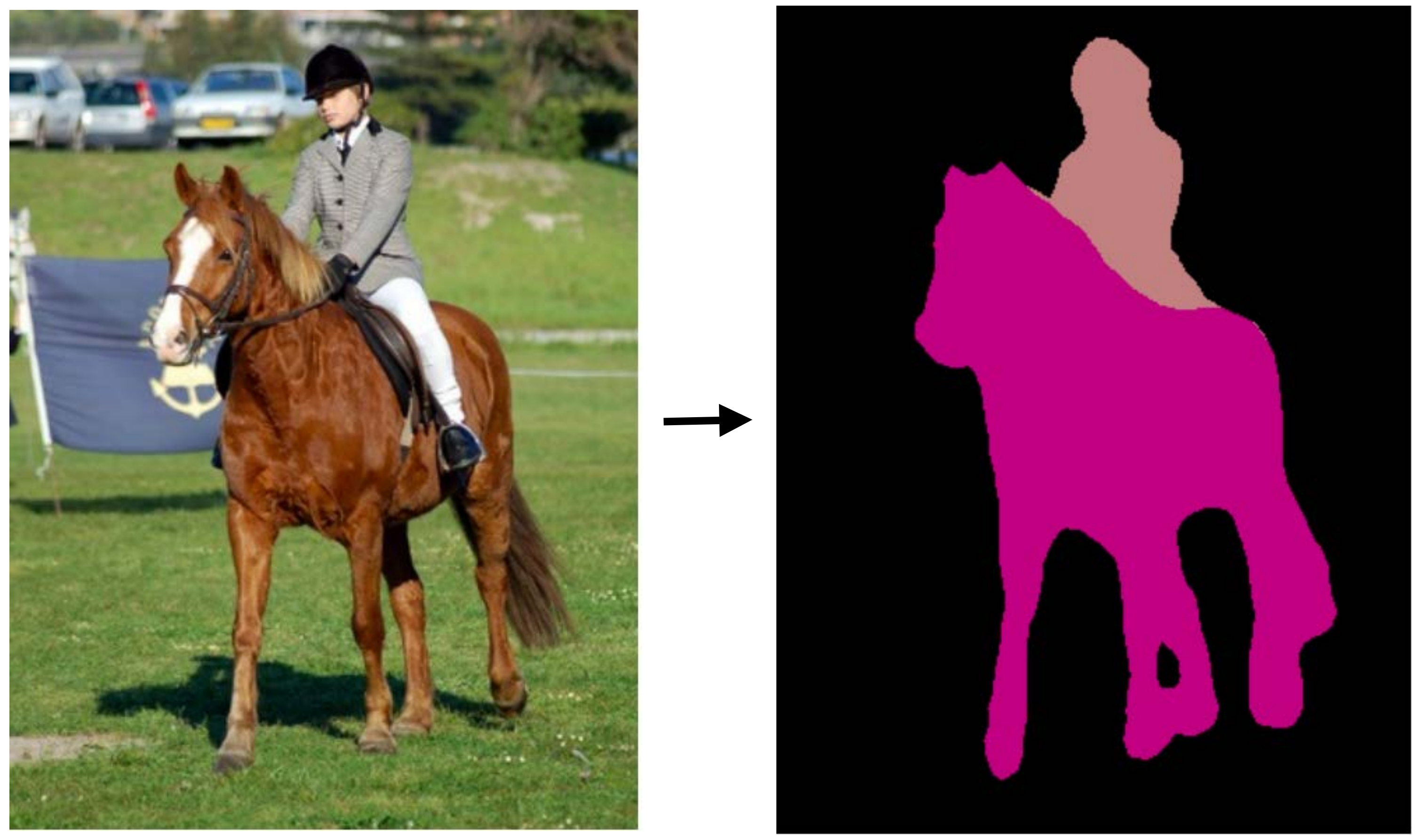
Created with edges2cats; Pix2Pix.

Deep generative models III

- Structured prediction
  - Image-to-image
  - Text-to-text
  - Image-to-text
  - Text-to-Image
- Unpaired translation

Data prediction problems ("structured prediction")

Semantic segmentation



[Long et al. 2015, …]

Automatic speech recognition

→ `computers that talk`

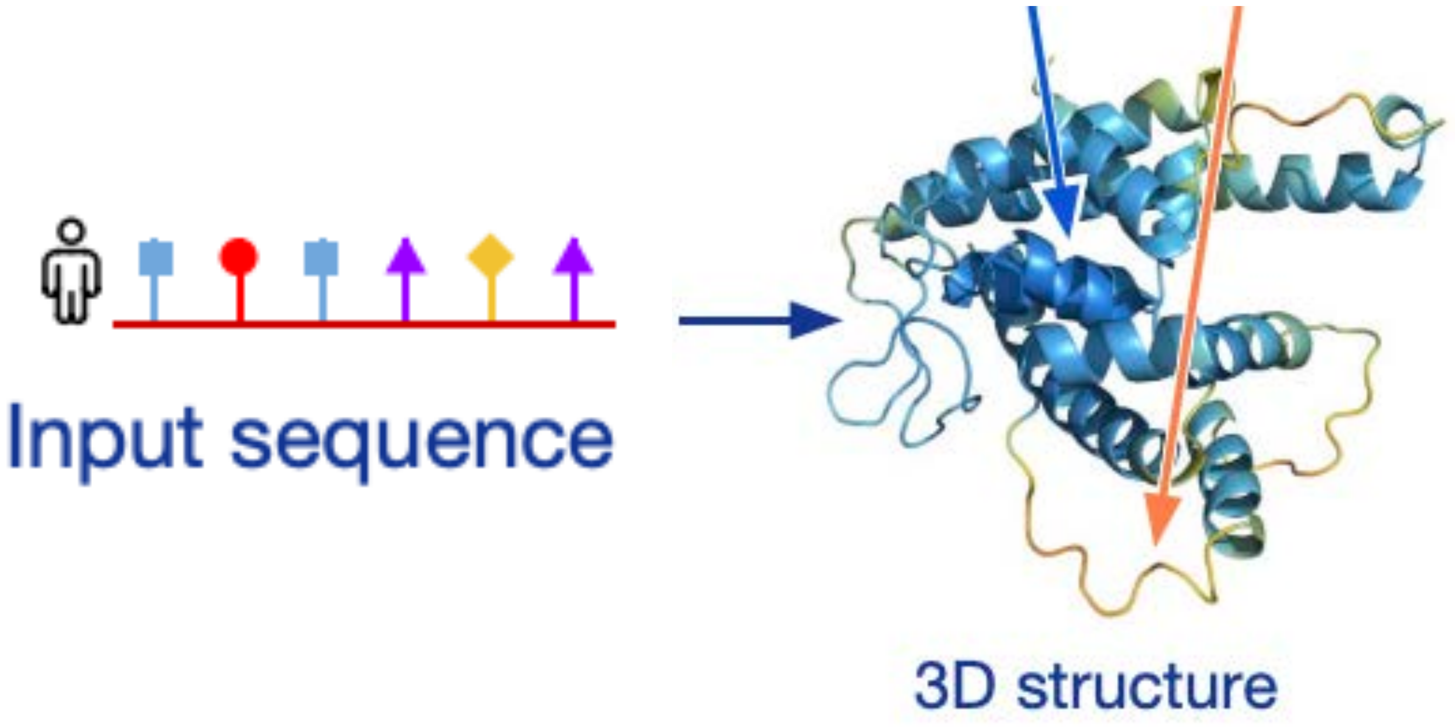
[Glass 2007, …]

Text-to-photo

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

[Reed et al. 2014, …]

Protein folding



[AlphaFold, Jumper et al. 2021, …]

© Long et al., Glass, Reed et al., AlphaFold, Jumper et al. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

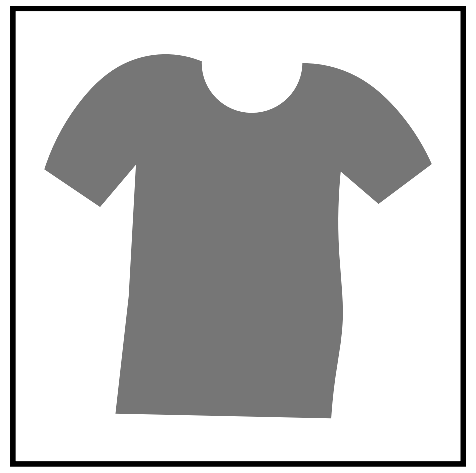
Structured prediction

X is high-dimensional

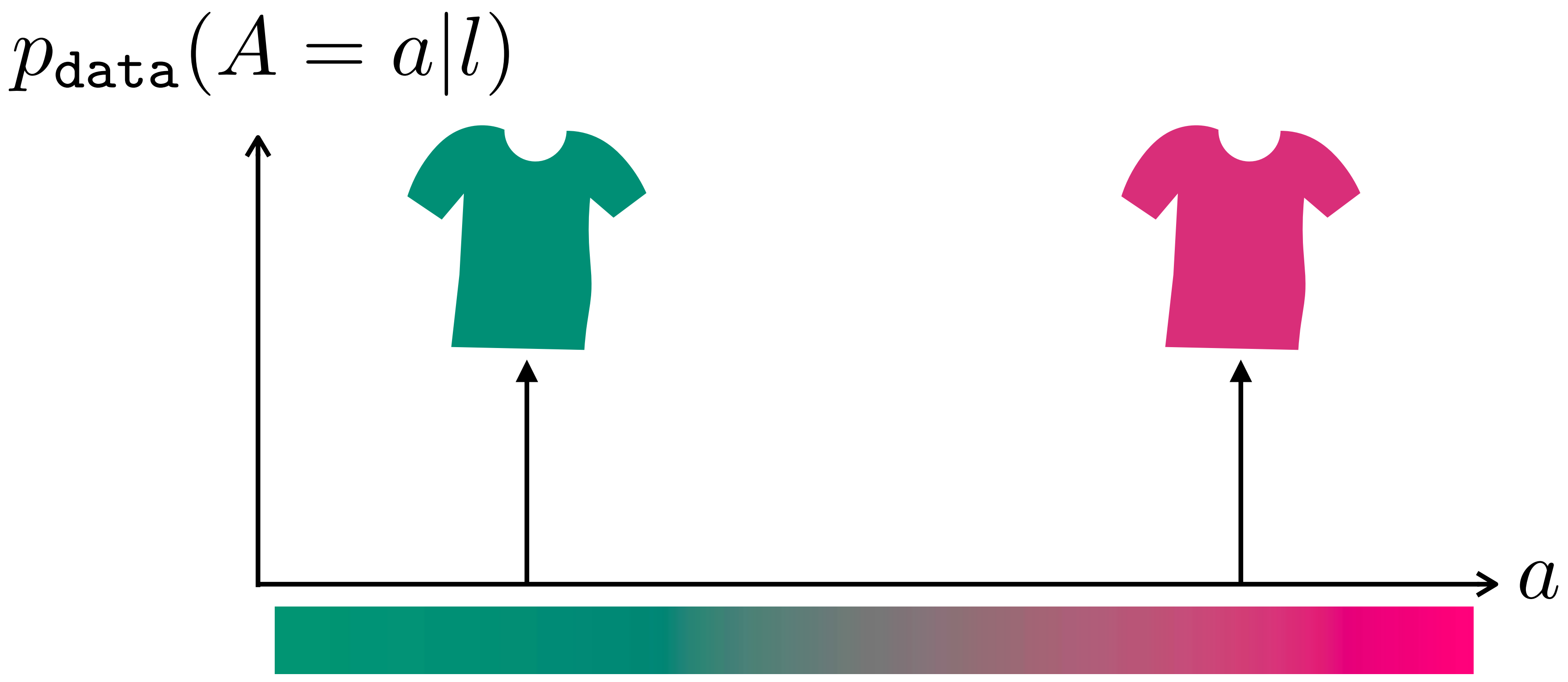
Model *joint* distribution of high-dimensional data $P(\mathbf{X}|\mathbf{Y}=\mathbf{y})$

In deep learning this is usually what we are interested in

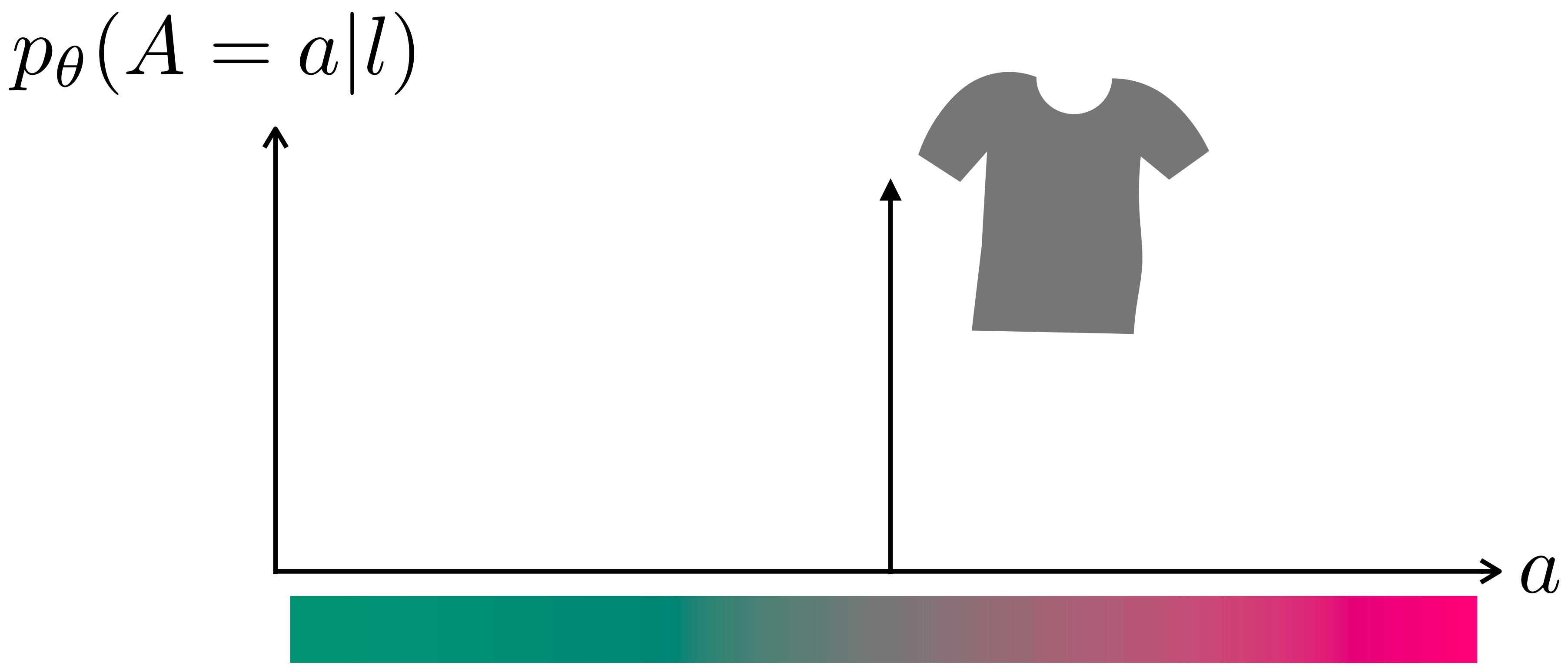
Unstructured: $\prod_i p(X_i|\mathbf{Y}=\mathbf{y})$



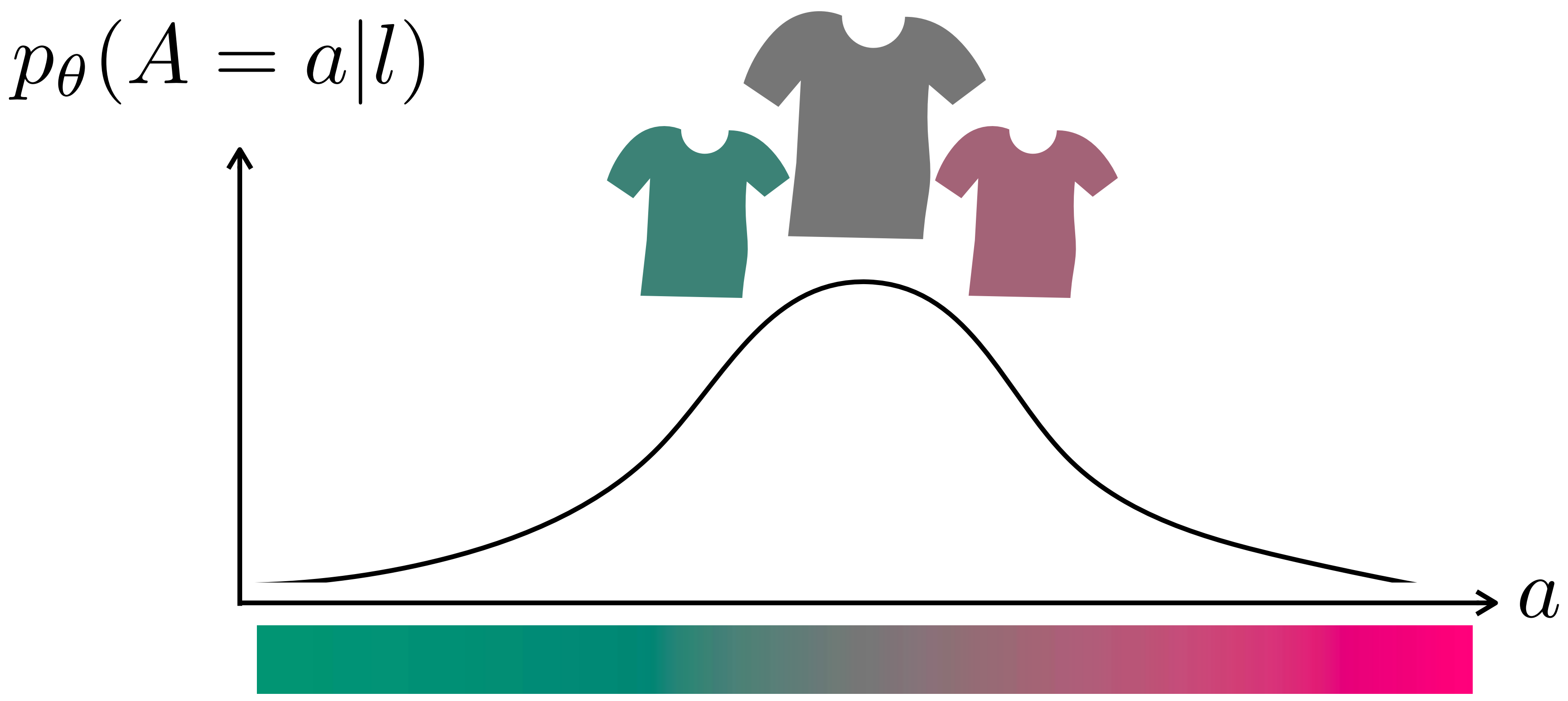
observation



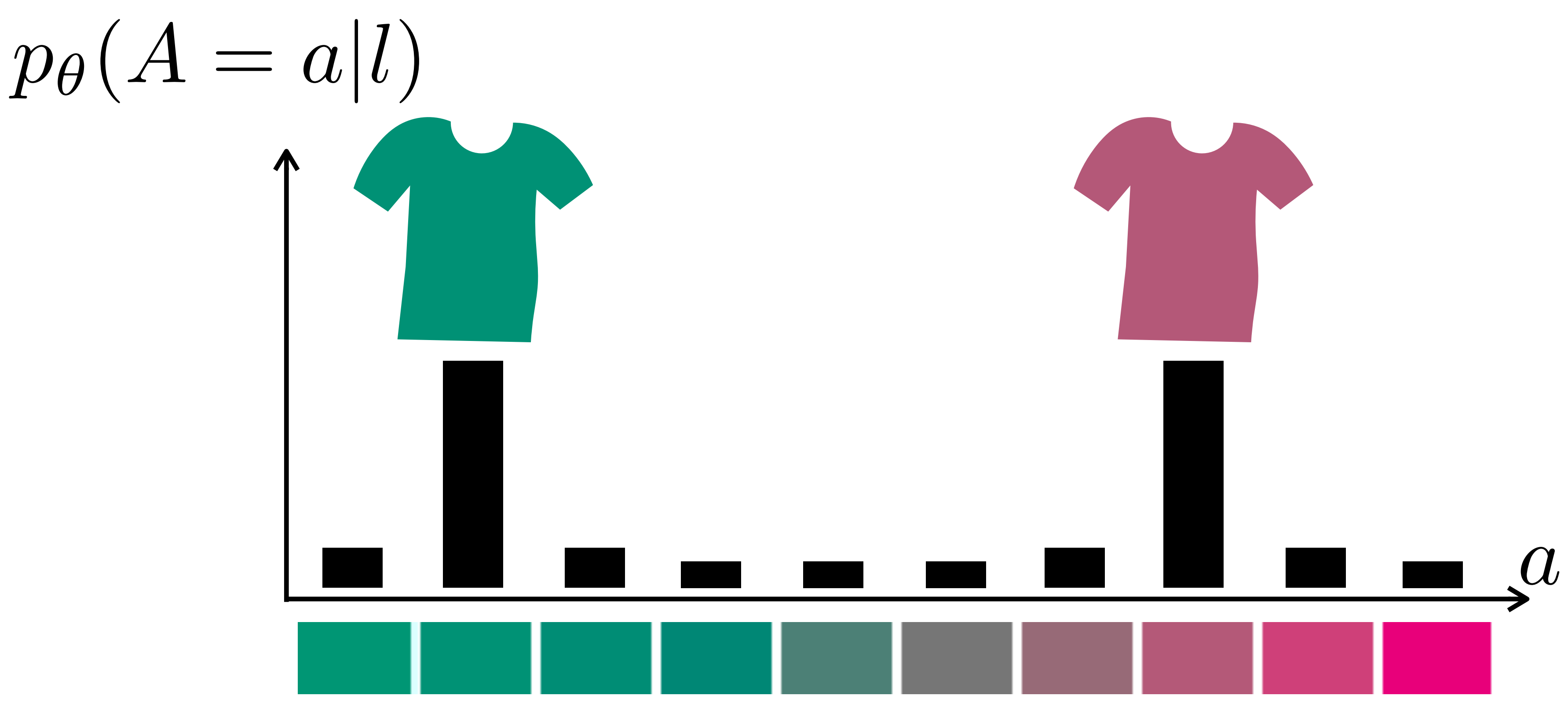
a) true data distribution



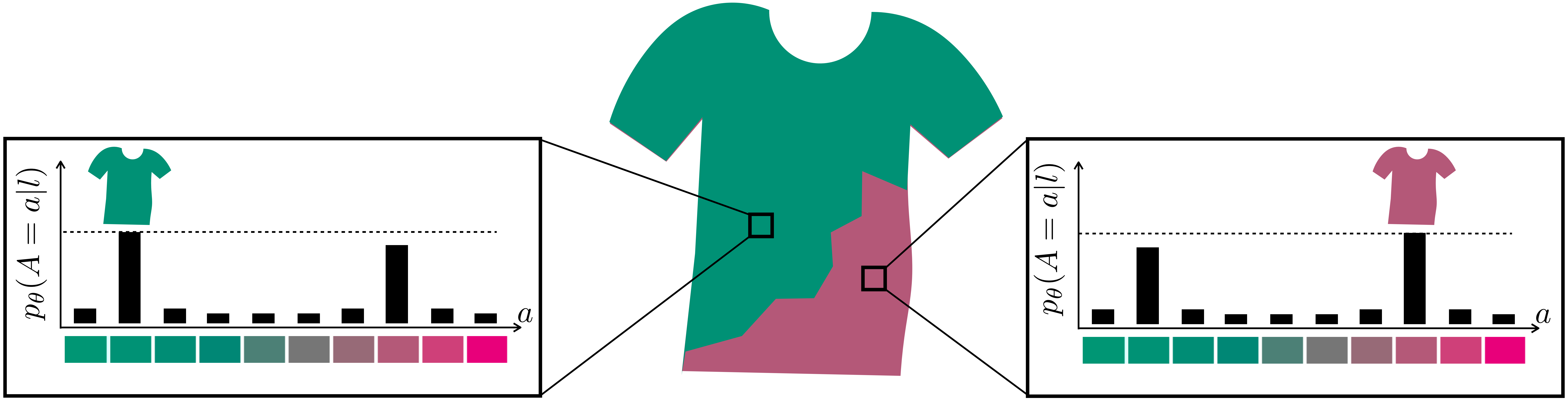
b) point prediction (aka regression)



d) Gaussian predictive distribution

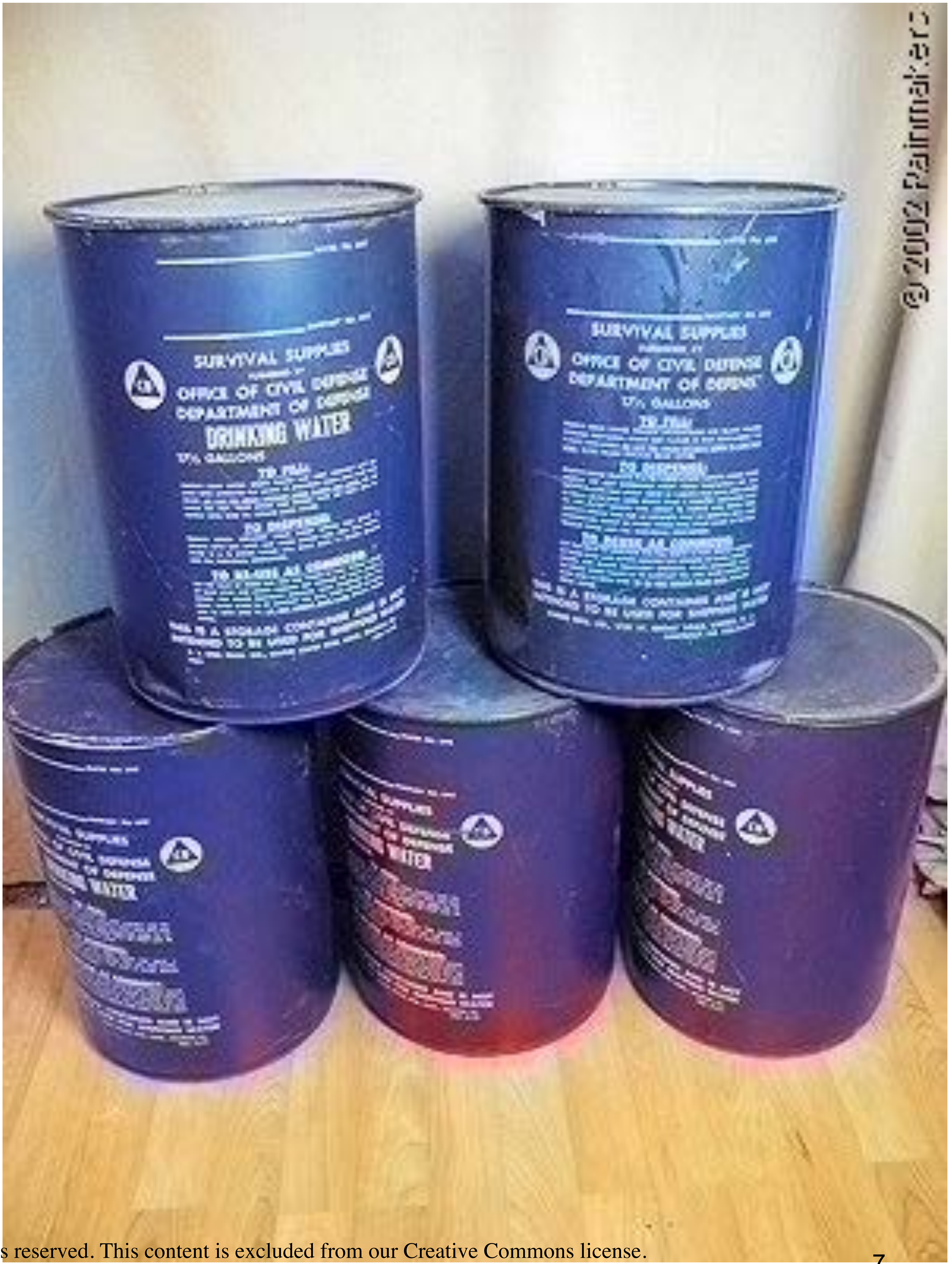


c) softmax distribution (aka classification)



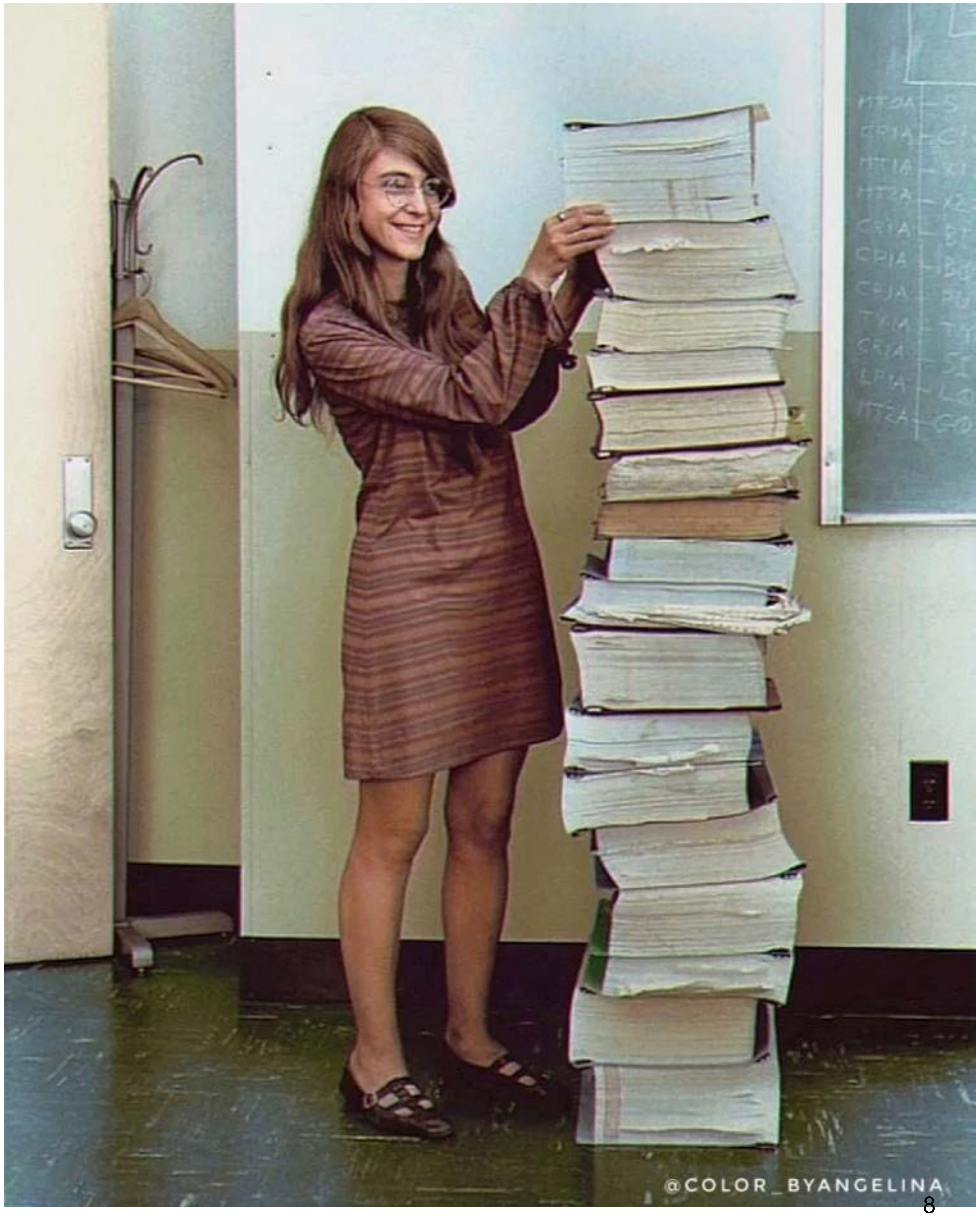
Generative models have two important properties for structured prediction:

1. They can model a multimodal distribution
2. They can model joint dependences between multidimensional predictions



Real or fake?

© painmakers. All rights reserved. This content is excluded from our Creative Commons license.
For more information, see https://ocw.mit.edu/help/faq-fair-use/



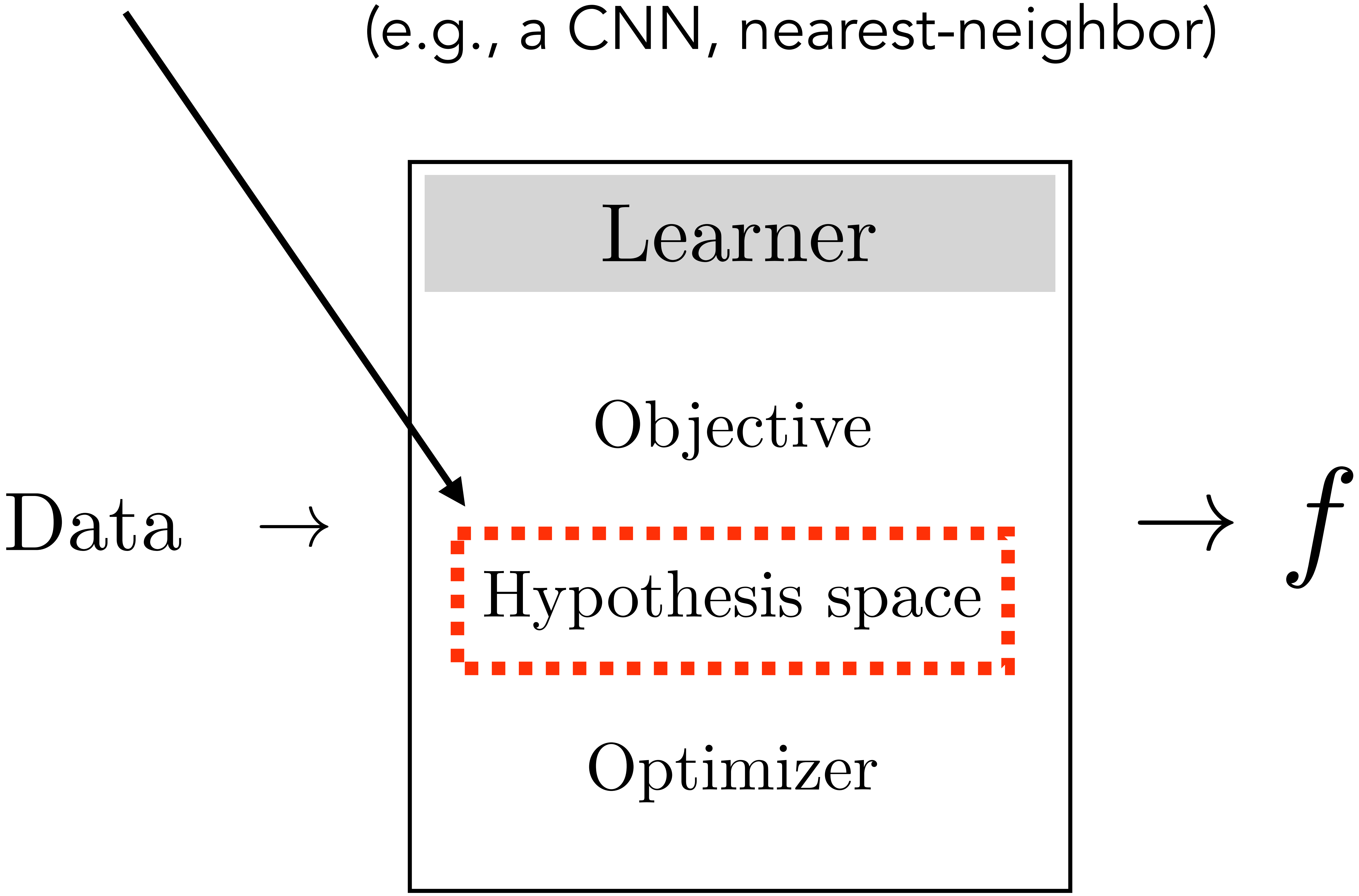
Real or fake?

© color_byangelina. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Deep learning in 2012

Use a **hypothesis space** that can model complex structure

(e.g., a CNN, nearest-neighbor)



Learner

Objective

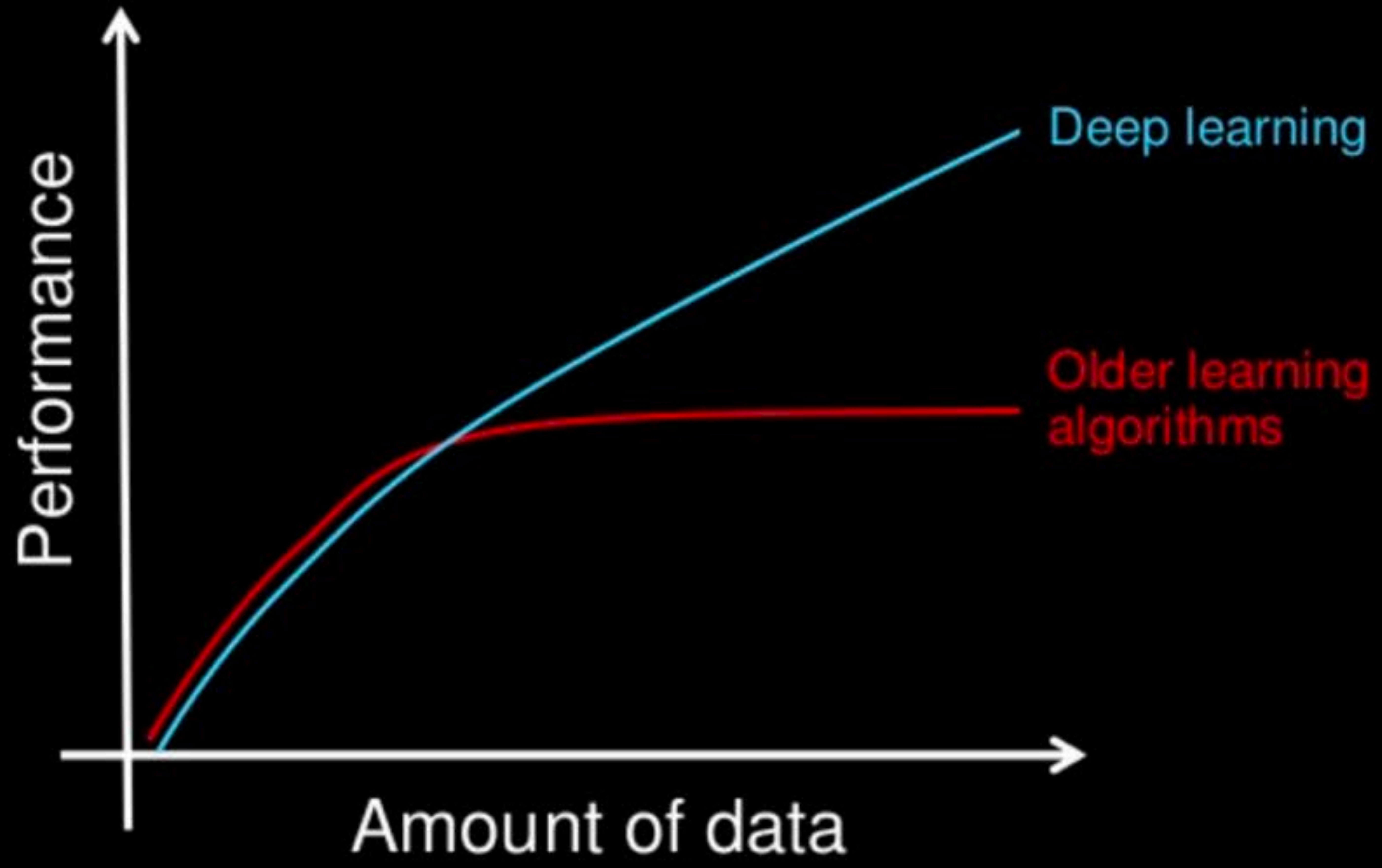
Data $\rightarrow$

Hypothesis space

$\rightarrow f$

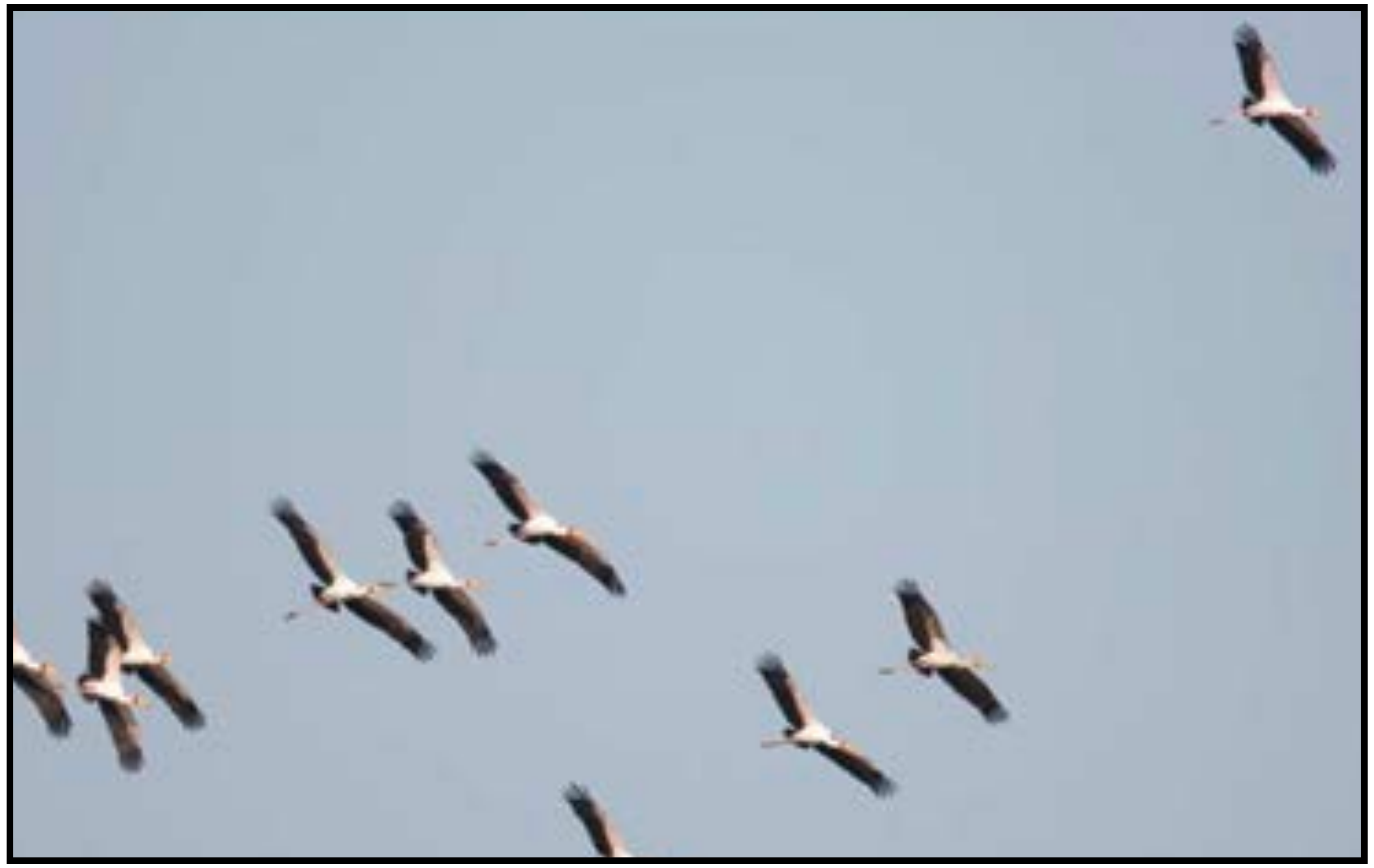
Optimizer

Why deep learning

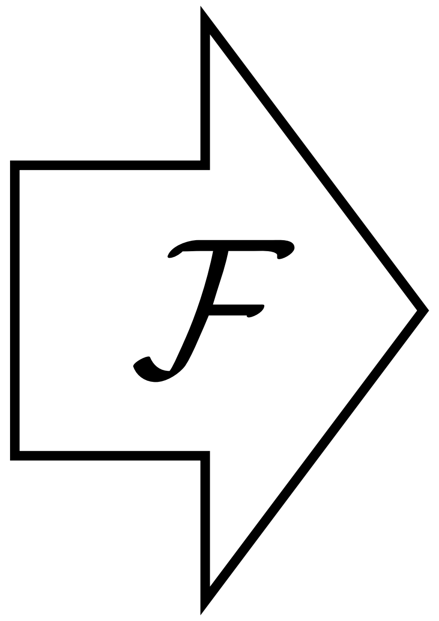


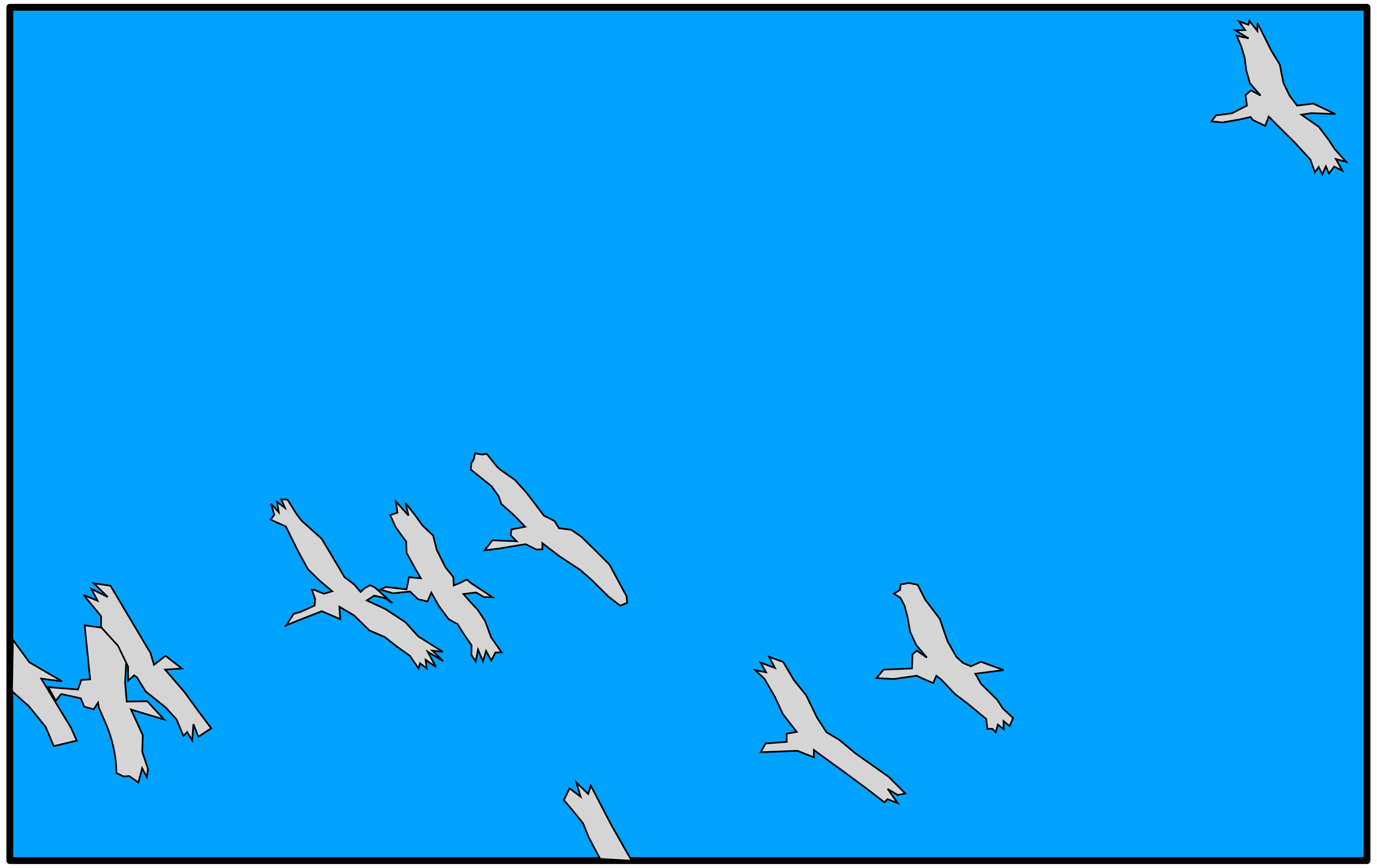
How do data science techniques scale with amount of data?

[Slide credit: Andrew Ng]



[Photo credit: Fredo Durand]





(Colors represent one-hot codes)

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

Hypothesis space

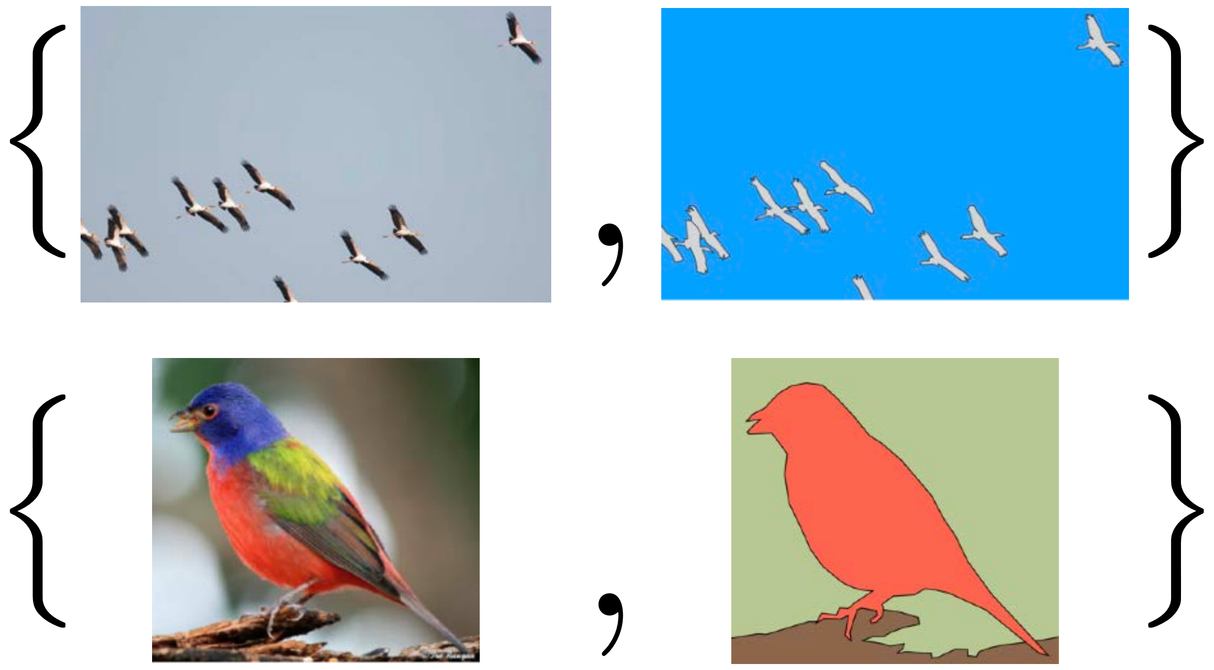
Objective function
(loss)

© Fred Durand. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Semantic Segmentation

Data

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



$\vdots$

$$\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$$
$$\mathbf{y} \in \mathbb{R}^{H \times W \times K}$$

$\rightarrow$

Learner

Objective

$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} H(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

Hypothesis space

Convolutional neural net

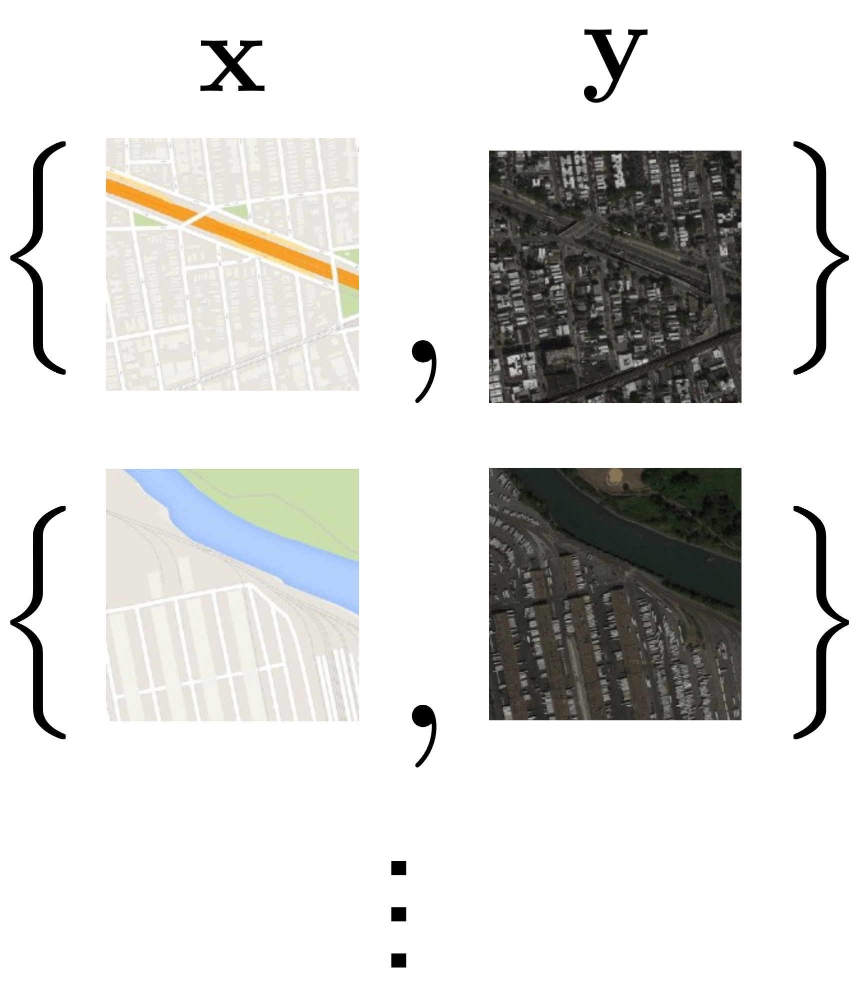
Optimizer

Stochastic gradient descent

$\rightarrow$ f

Sat2Map

Data



$$\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$$

$$\mathbf{y} \in \mathbb{R}^{H\times W\times 3}$$

$\rightarrow$

Learner

Objective

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} (\, f_\theta(\mathbf{x})_i - y_i)^2$$

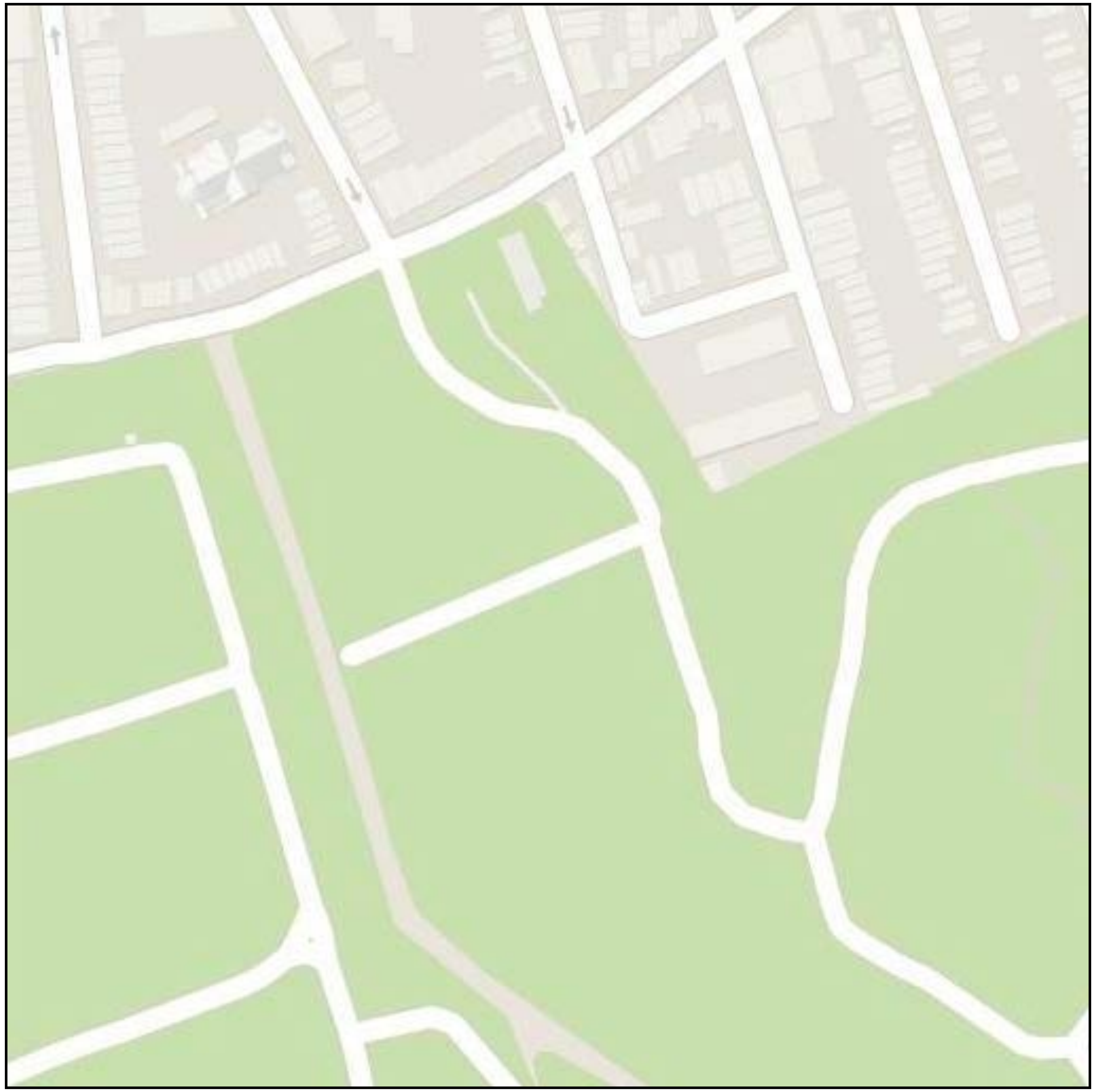
Hypothesis space

Convolutional neural net

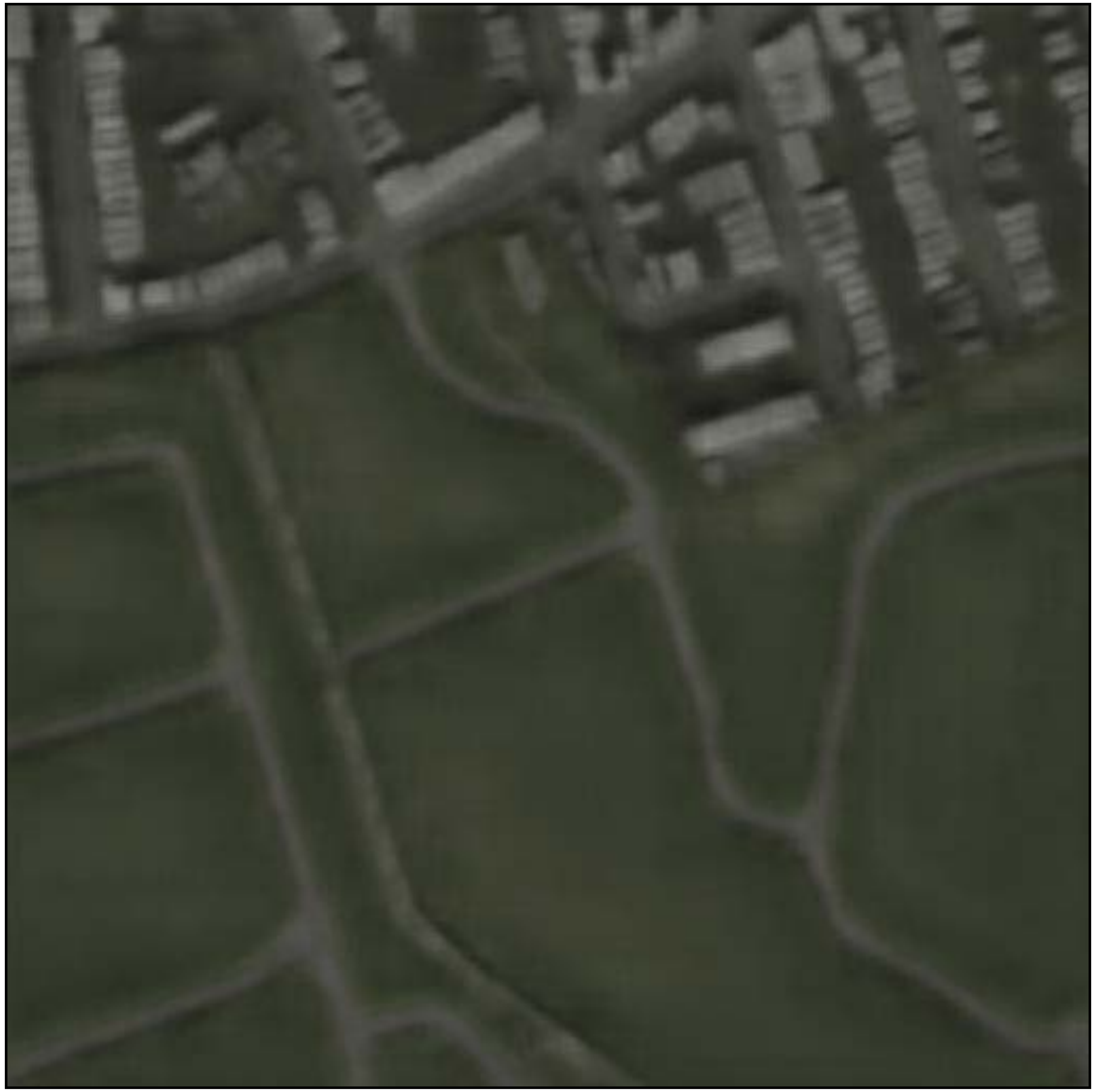
Optimizer

Stochastic gradient descent

$\rightarrow f$



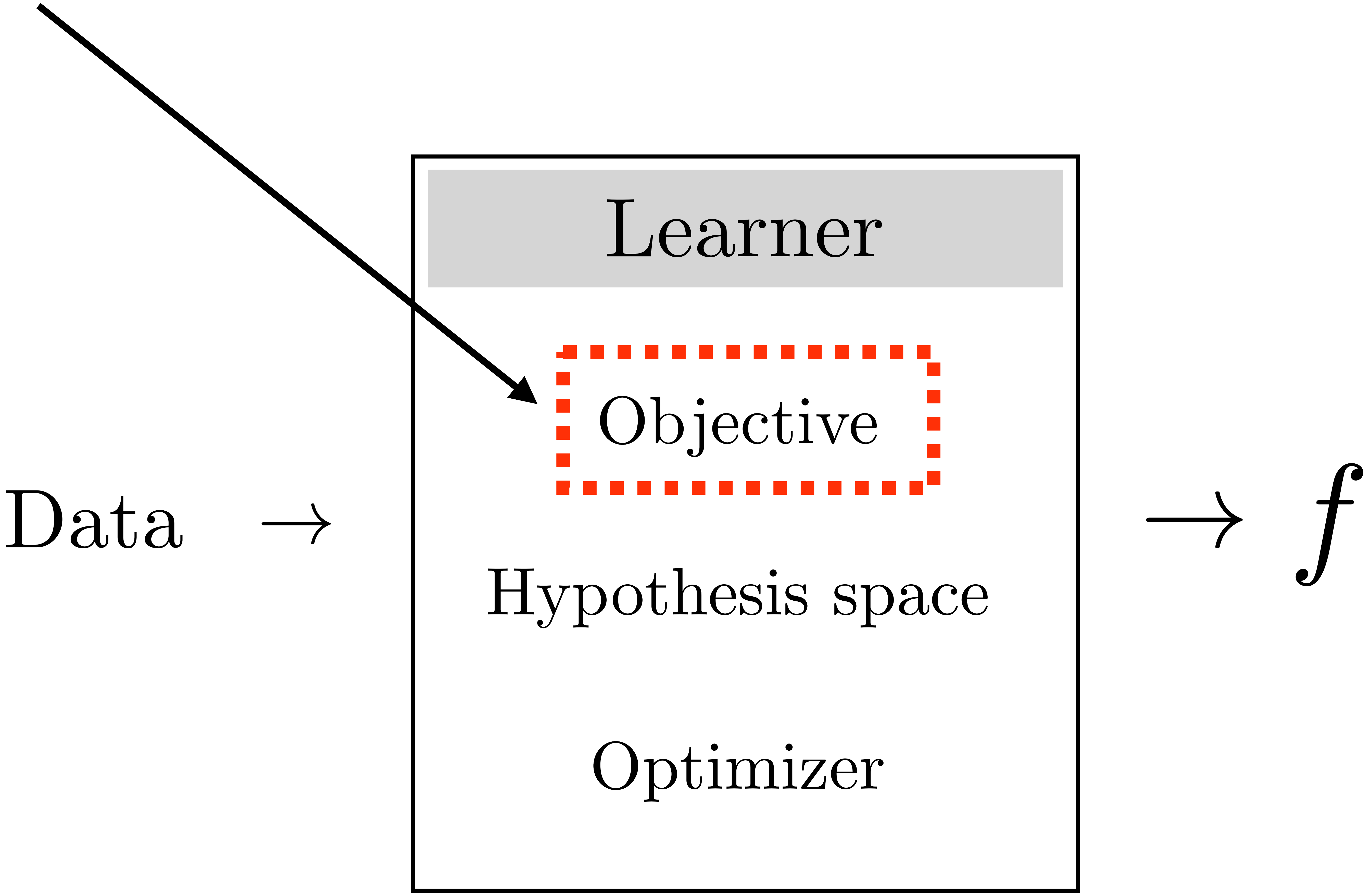
Input



Deep net output

Structured prediction

Use an **objective** that can model structure! (e.g., a graphical model, a GAN, etc)



Learner

Objective

Data →

Hypothesis space

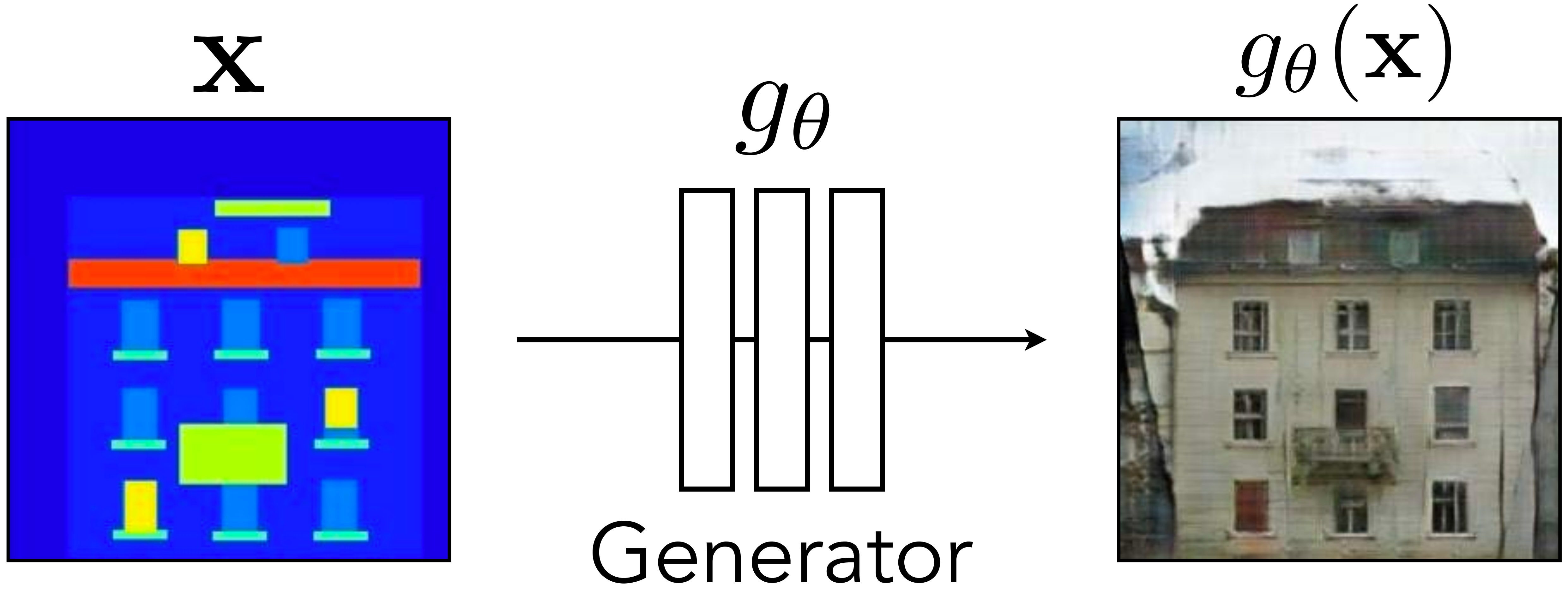
Optimizer

→ f

Image-to-image

- → with conditional GANs
- → with conditional VAEs

Conditional GAN



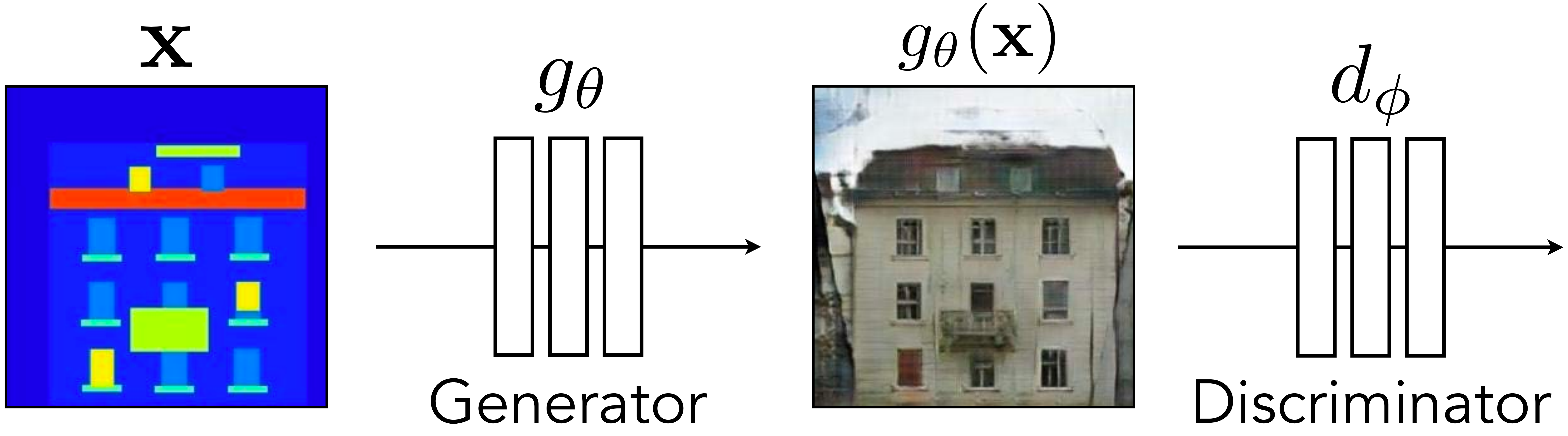
$\mathbf{x}$

g_θ

Generator

$g_\theta(\mathbf{x})$

For example: pix2pix [Isola et al. 2017]



$\mathbf{x}$

g_θ

Generator

$g_\theta(\mathbf{x})$

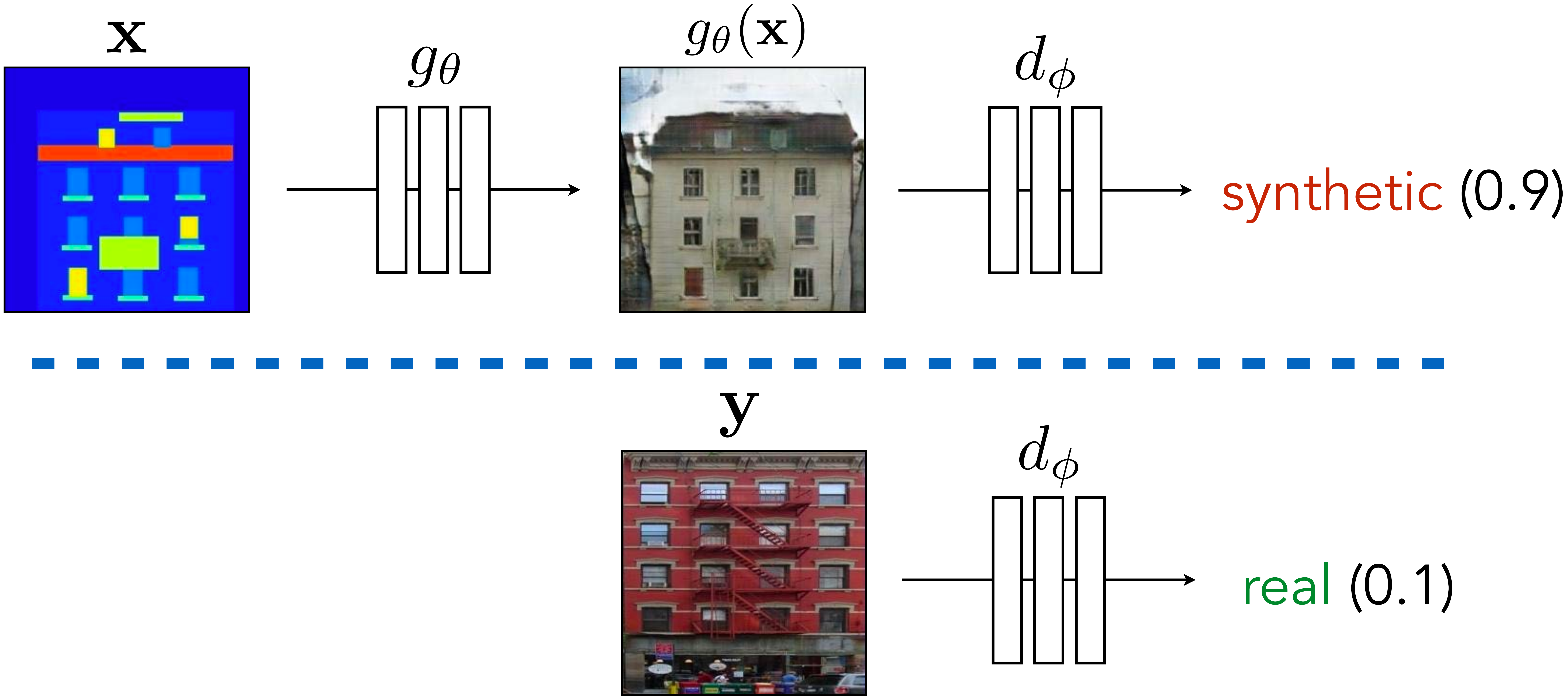
d_ϕ

Discriminator

real or synthetic?

g tries to synthesize fake images that fool d

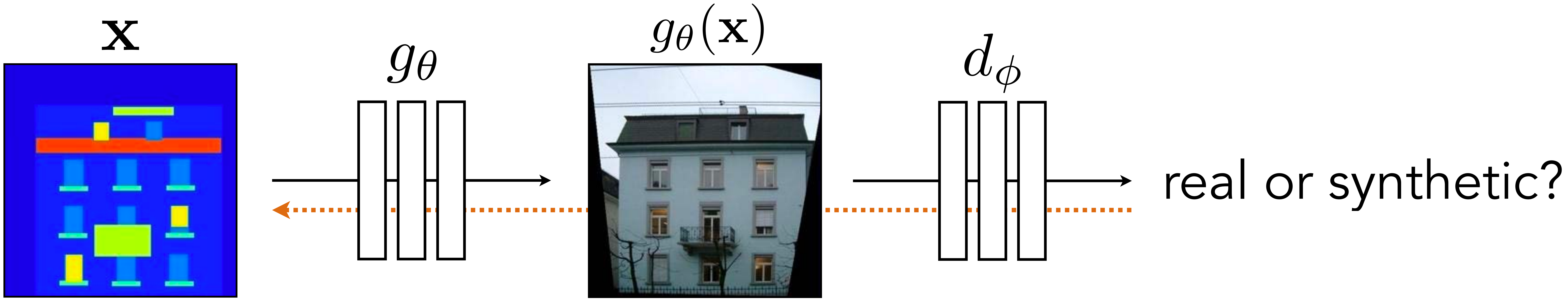
d tries to identify the fakes



synthetic (0.9)

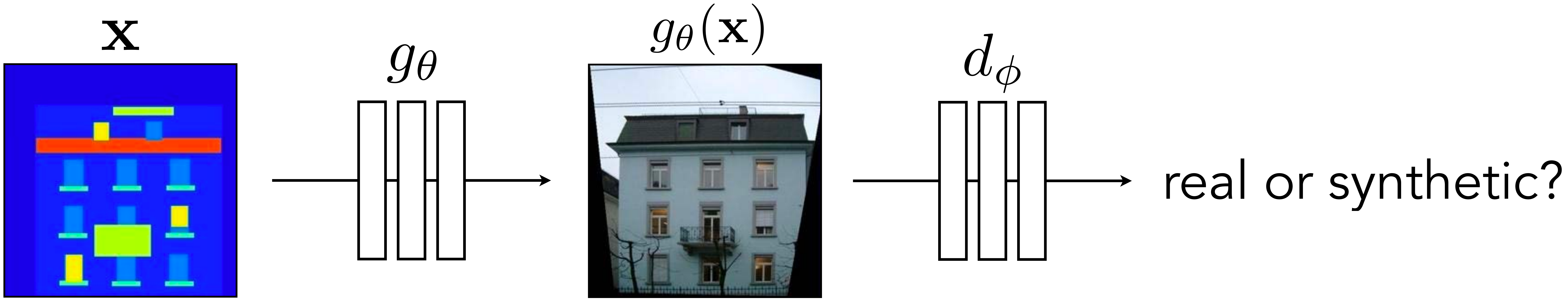
real (0.1)

$$d^*_\phi = \arg\max_\phi \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\boxed{\log d_\phi(g_\theta(\mathbf{x}))} + \boxed{\log(1 - d_\phi(\mathbf{y}))}\right]$$



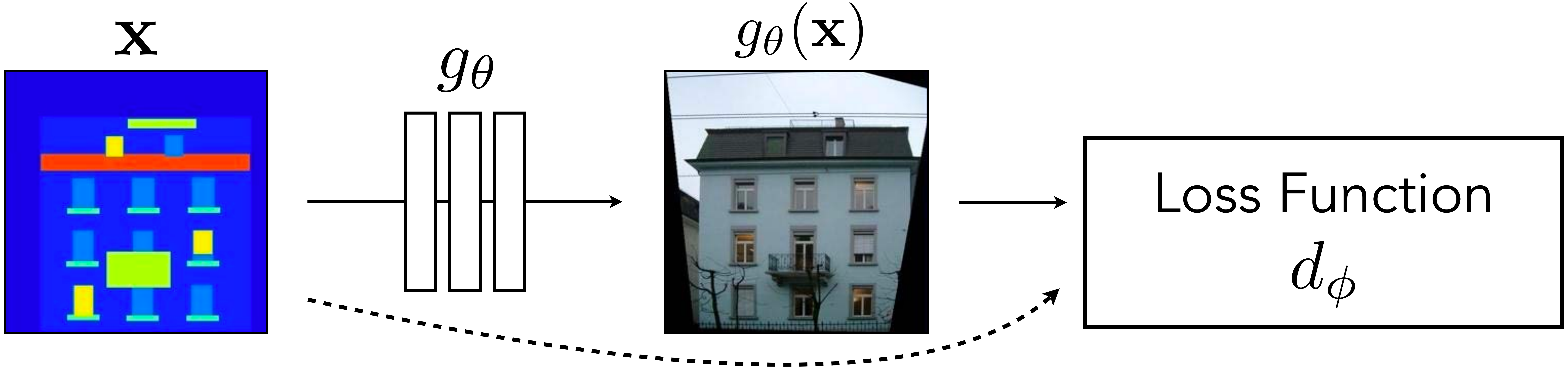
g tries to synthesize fake images that *fool* d:

$$g_\theta^* = \boxed{\arg\min_\theta} \mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(g_\theta(\mathbf{x}))]$$



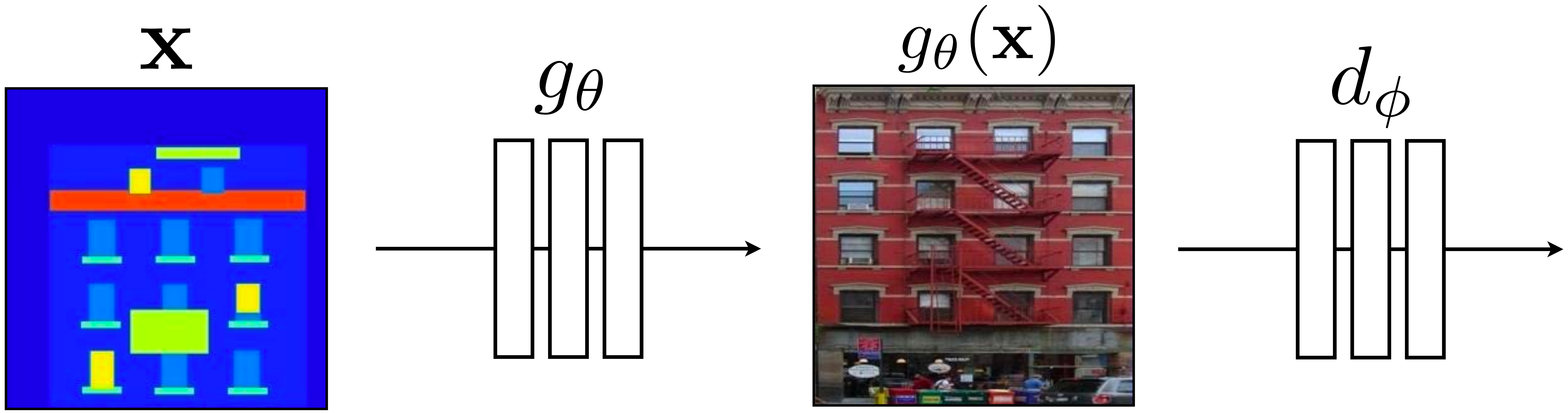
g tries to synthesize fake images that *fool* the *best* d:

$$\arg \boxed{\min_{\theta}} \boxed{\max_{\phi}} \mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(g_\theta(\mathbf{x})) + \log(1 - d_\phi(\mathbf{y}))]$$



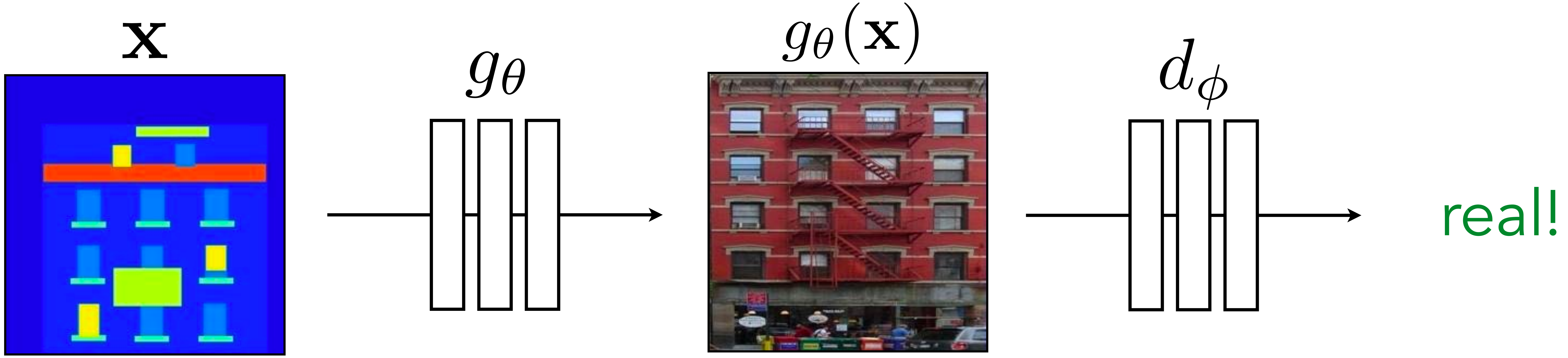
g's perspective: d is a loss function.

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

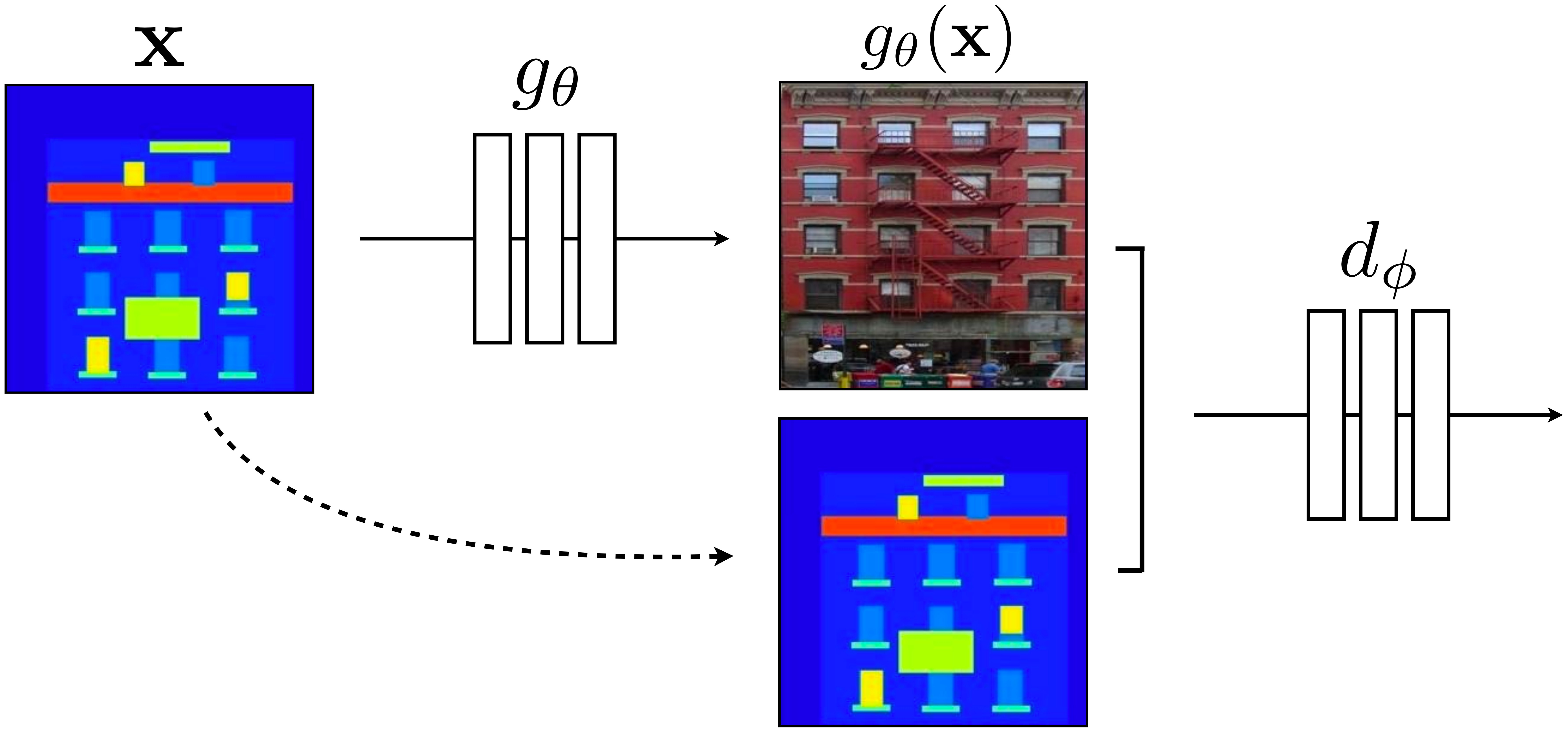


real or synthetic?

$$\arg\min_{\theta} \max_{\phi} \mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(g_\theta(\mathbf{x})) + \log(1 - d_\phi(\mathbf{y}))]$$

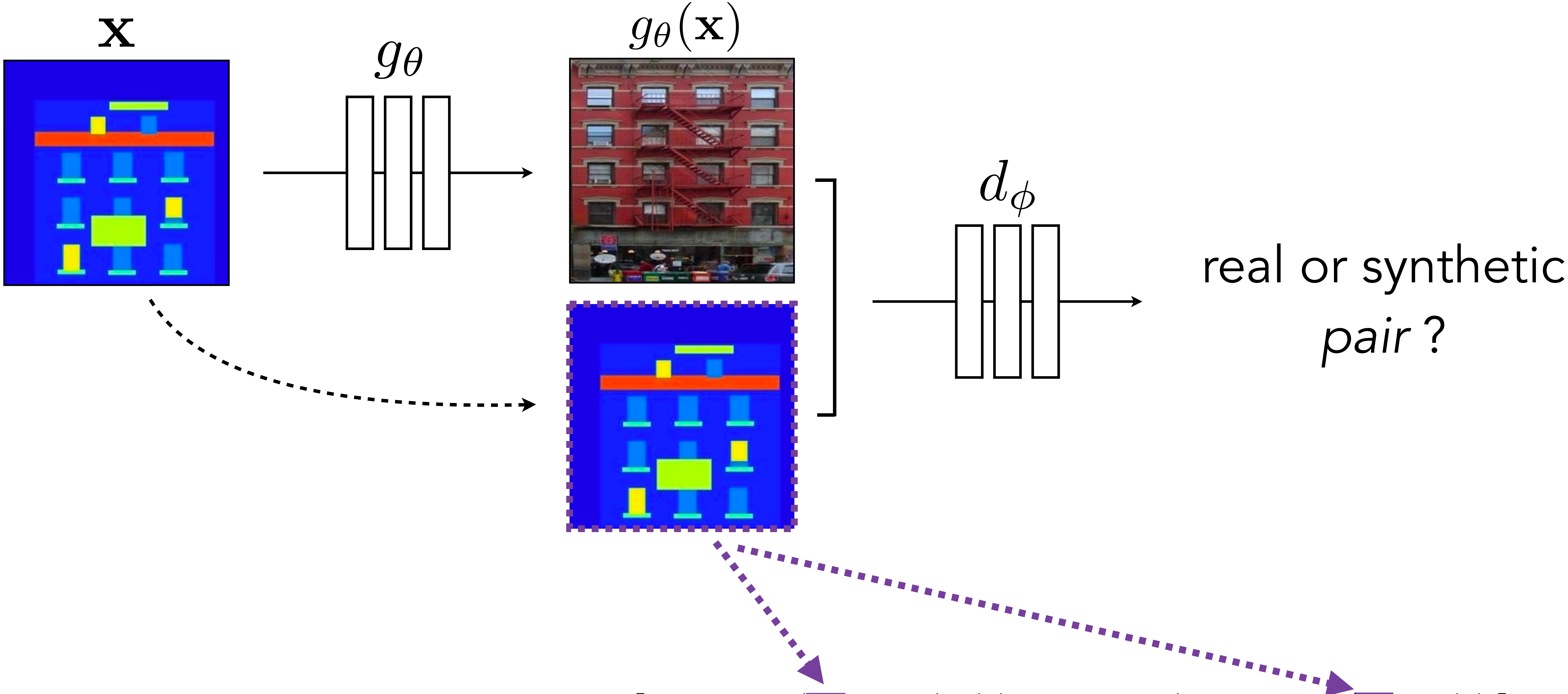


$$\arg\min_{\theta}\max_{\phi}\mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(g_\theta(\mathbf{x}))+\log(1-d_\phi(\mathbf{y}))]$$



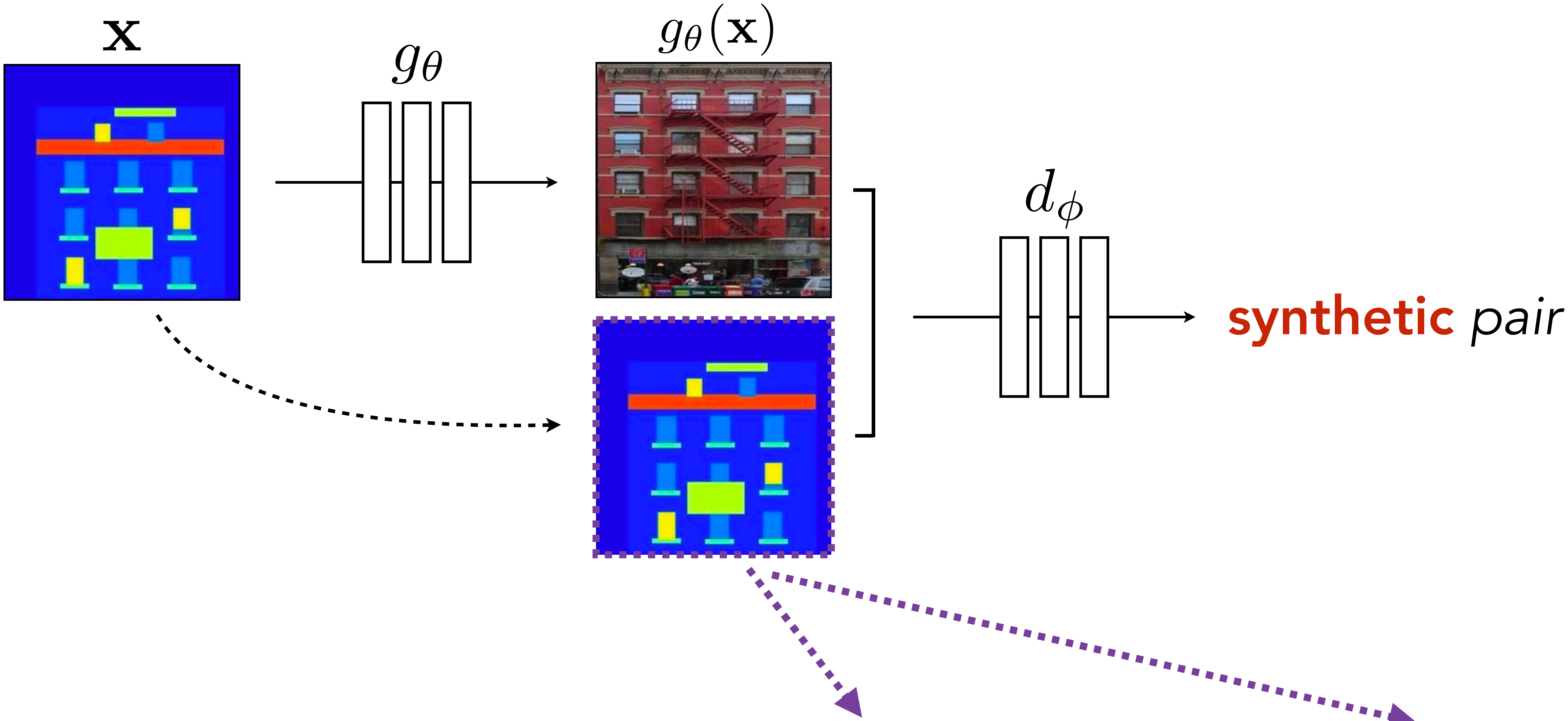
real or synthetic *pair* ?

$$\arg\min_\theta \max_\phi \mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(g_\theta(\mathbf{x})) + \log(1 - d_\phi(\mathbf{y}))]$$



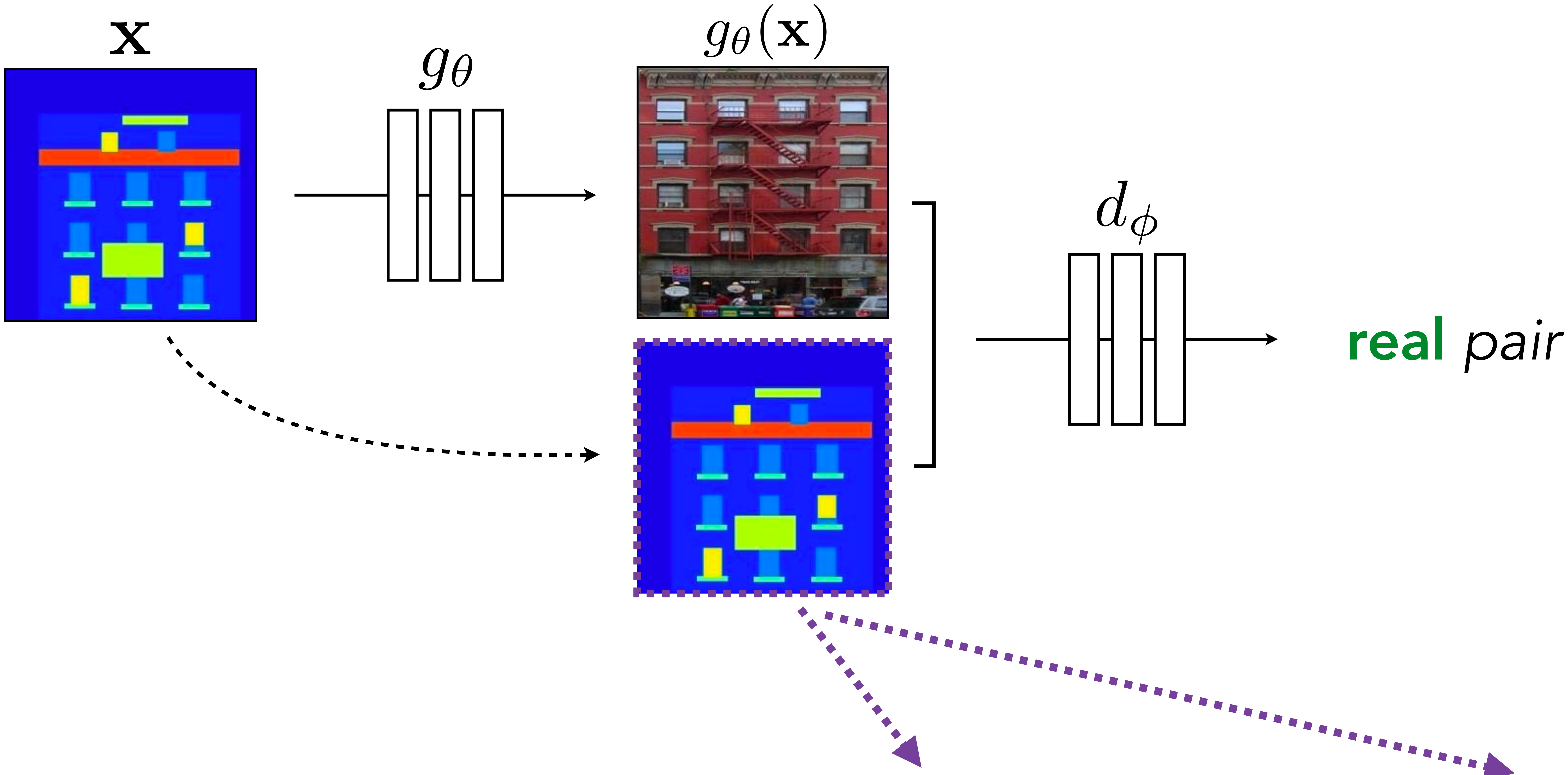
real or synthetic *pair*?

$$\arg\min_{\theta}\max_{\phi}\mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(\boxed{\mathbf{x}}, g_\theta(\mathbf{x})) + \log(1 - d_\phi(\boxed{\mathbf{x}}, \mathbf{y}))]$$

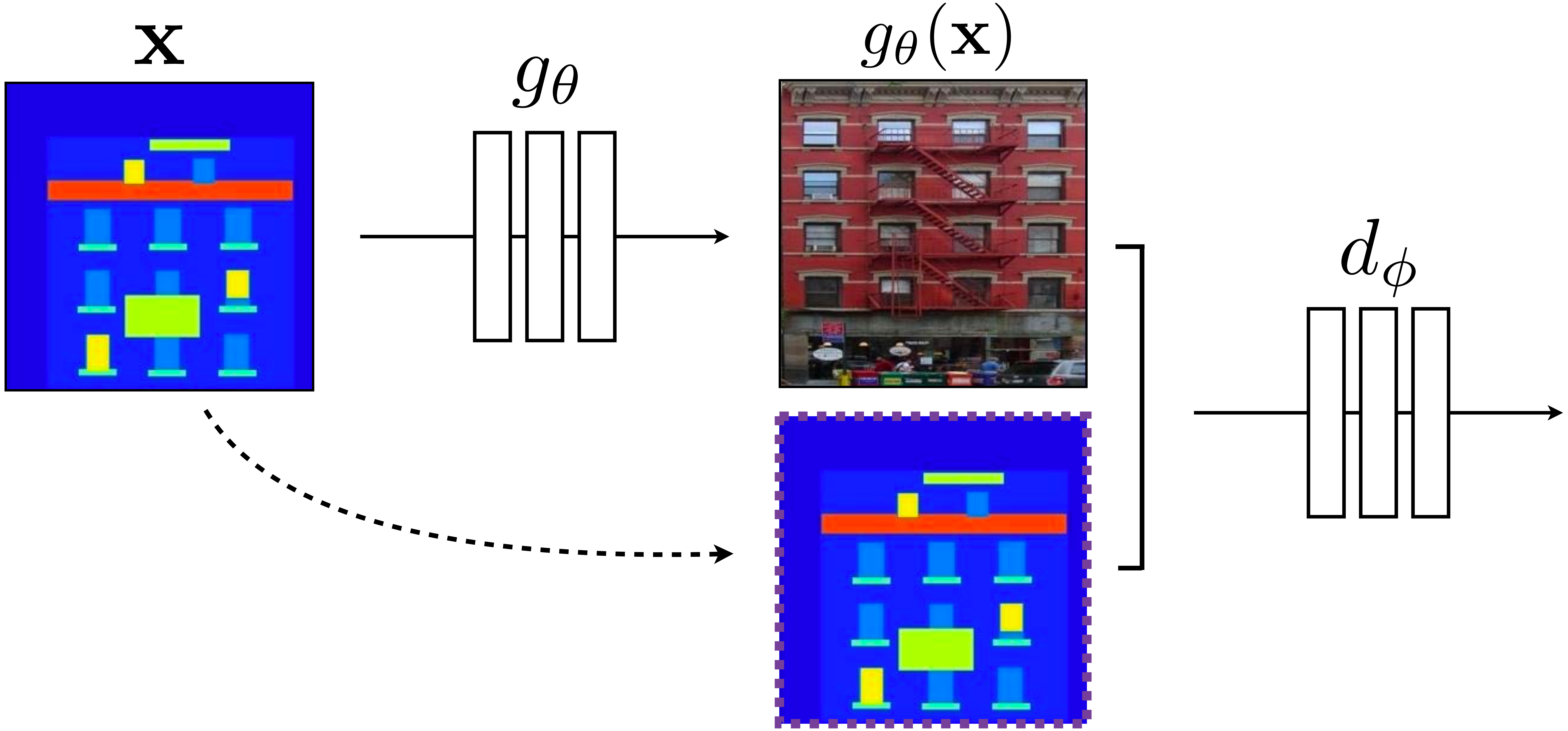


synthetic *pair*

$$\arg\min_{\theta}\max_{\phi}\mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(\boxed{\mathbf{x}}, g_\theta(\mathbf{x})) + \log(1 - d_\phi(\boxed{\mathbf{x}}, \mathbf{y}))]$$



$$\arg\min_{\theta}\max_{\phi}\mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(\boxed{\mathbf{x}}, g_\theta(\mathbf{x})) + \log(1 - d_\phi(\boxed{\mathbf{x}}, \mathbf{y}))]$$

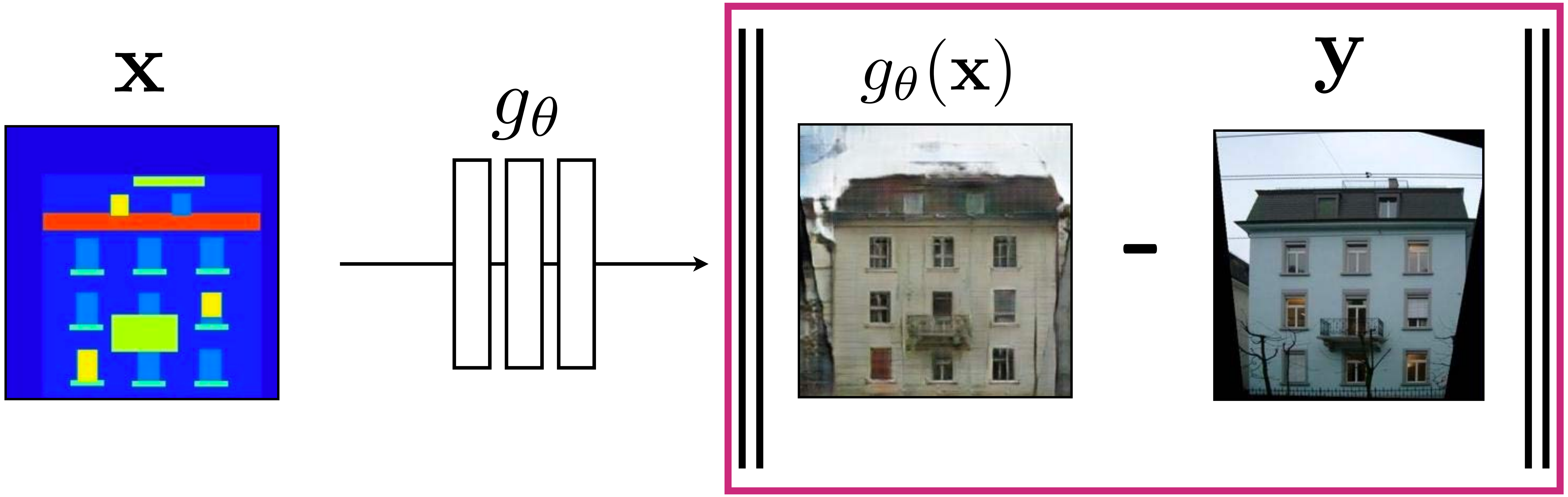


real or synthetic *pair* ?

$$\boxed{\arg\min_\theta \max_\phi \mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(\mathbf{x}, g_\theta(\mathbf{x})) + \log(1 - d_\phi(\mathbf{x},\mathbf{y}))]}$$

Training Details: Loss function

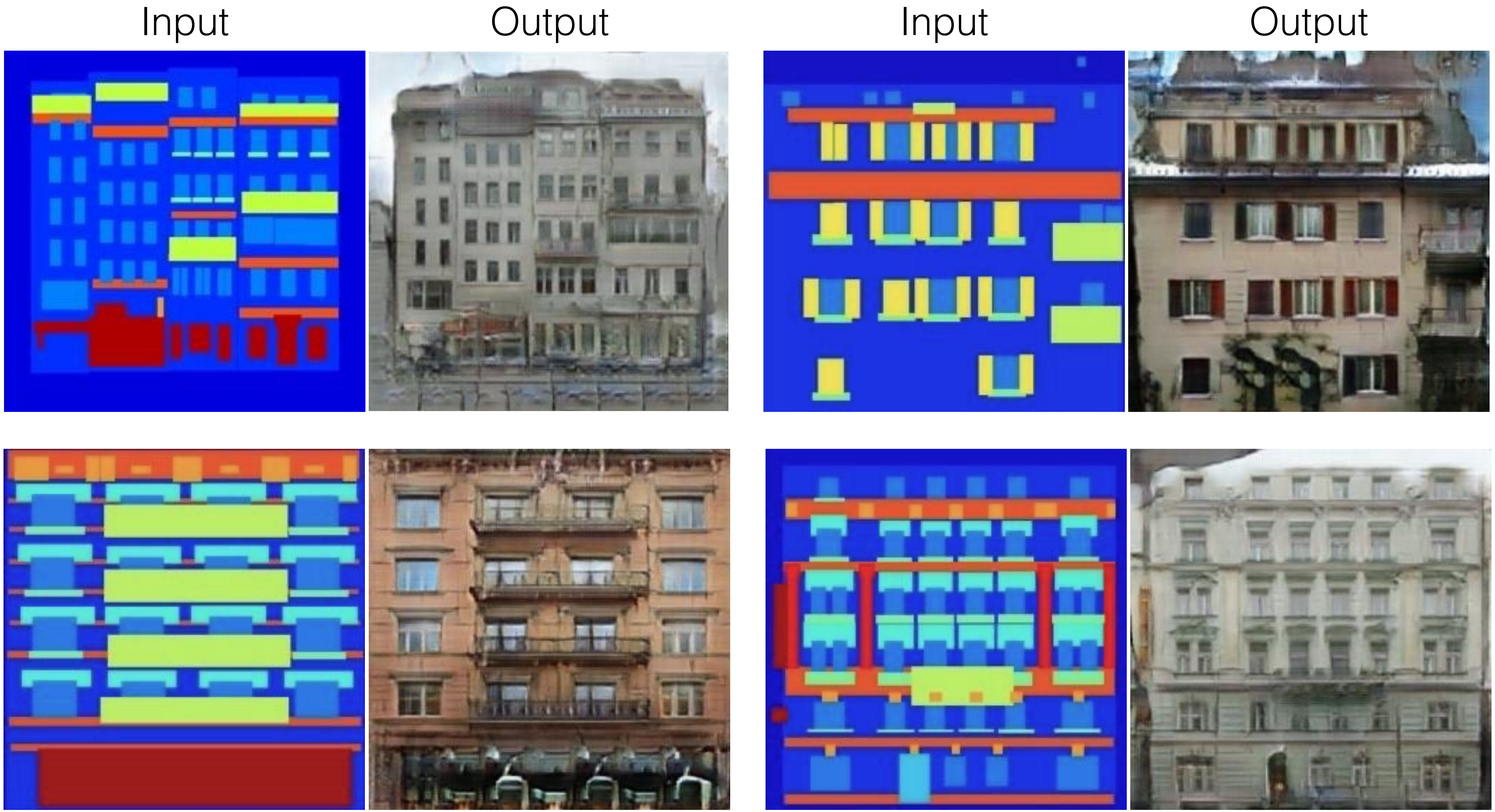
$$g_\theta^* = \arg\min_\theta \max_\phi \mathcal{L}_{\texttt{cGAN}}(\theta, \phi) + \boxed{\lambda \mathcal{L}_{\texttt{L1}}(\theta)}$$



Stable training + fast convergence

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

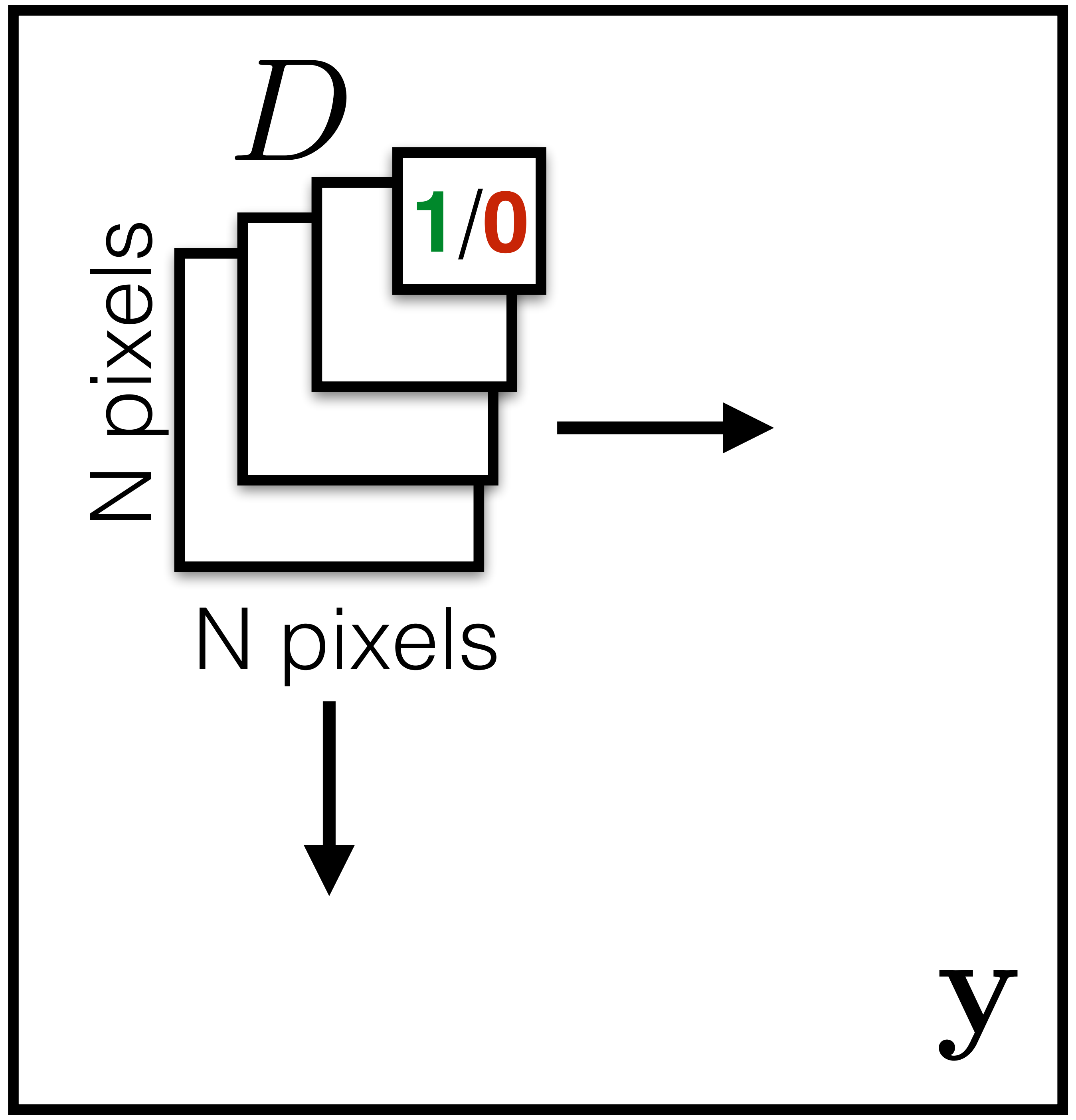
Labels → Facades

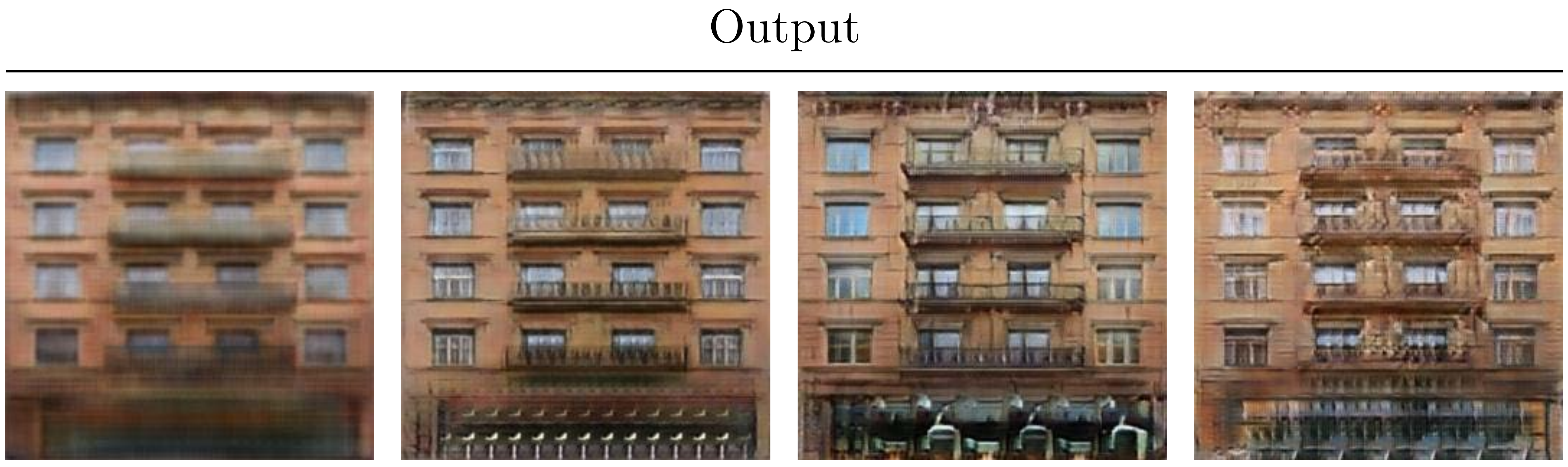


Input Output Input Output

Data from [Tylecek, 2013]

Patch Discriminator

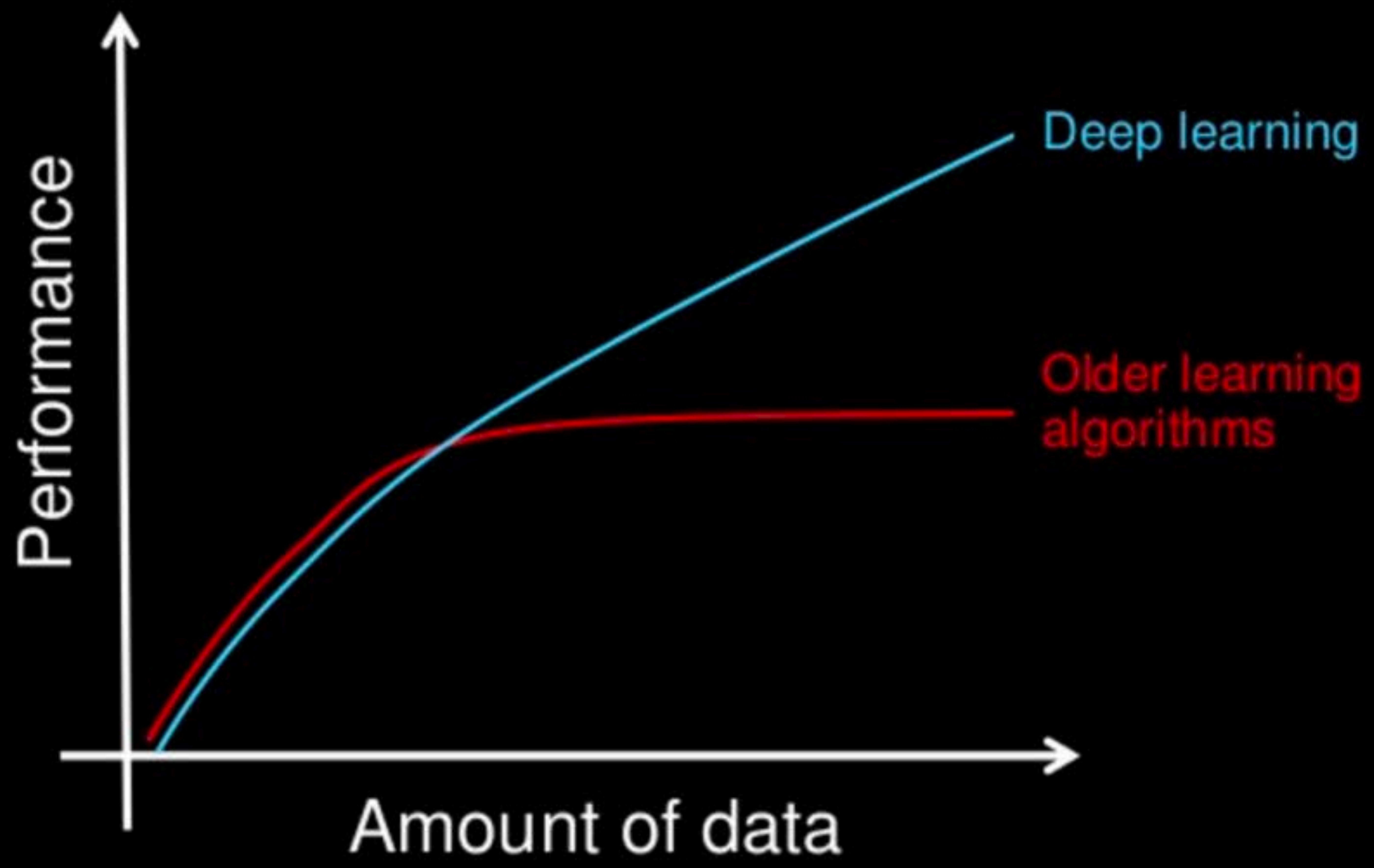




1×1 16×16 70×70 286×286

Discriminator patch size

Why deep learning

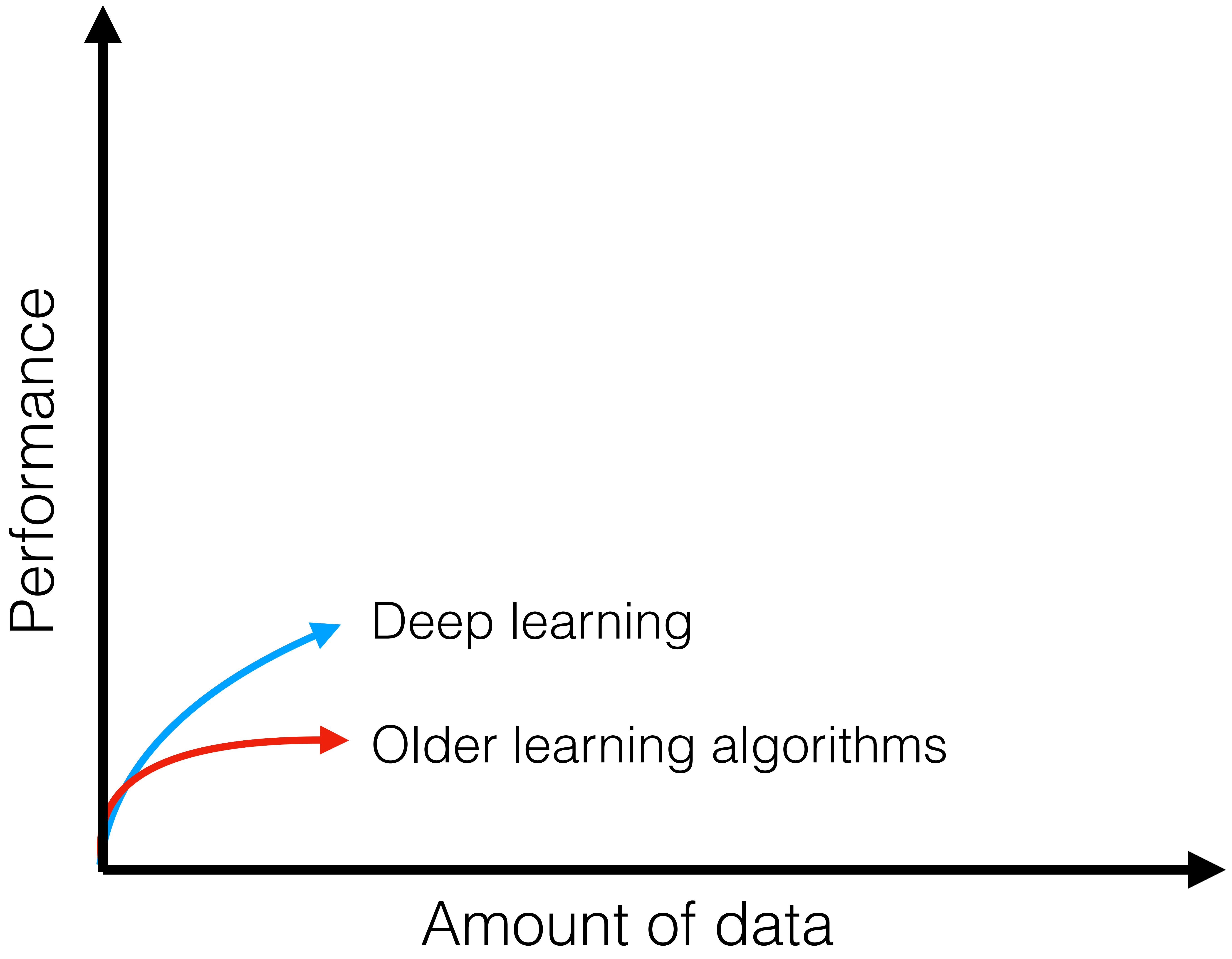


How do data science techniques scale with amount of data?

[Slide credit: Andrew Ng]

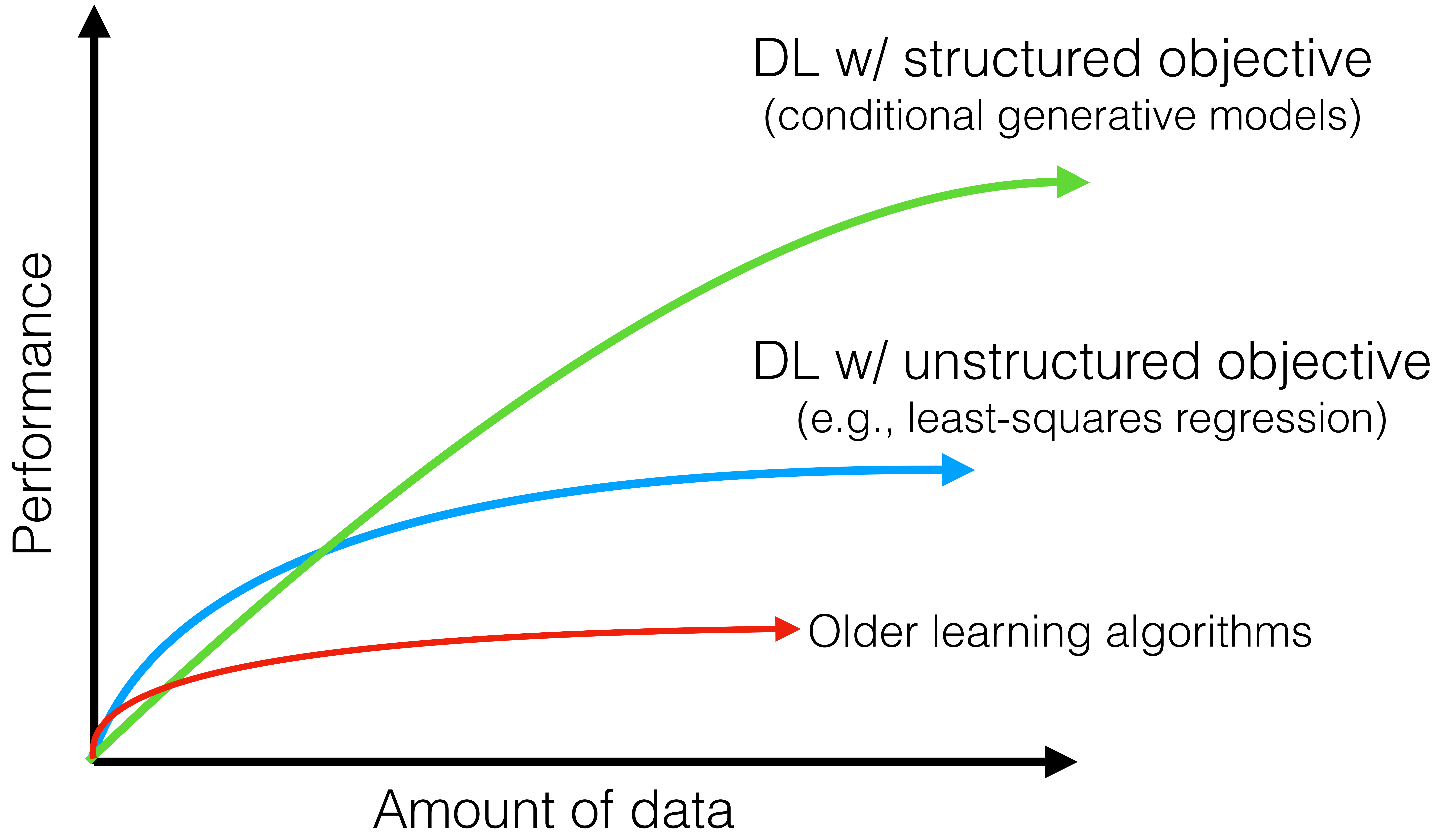
Why structured objectives

(cartoon)

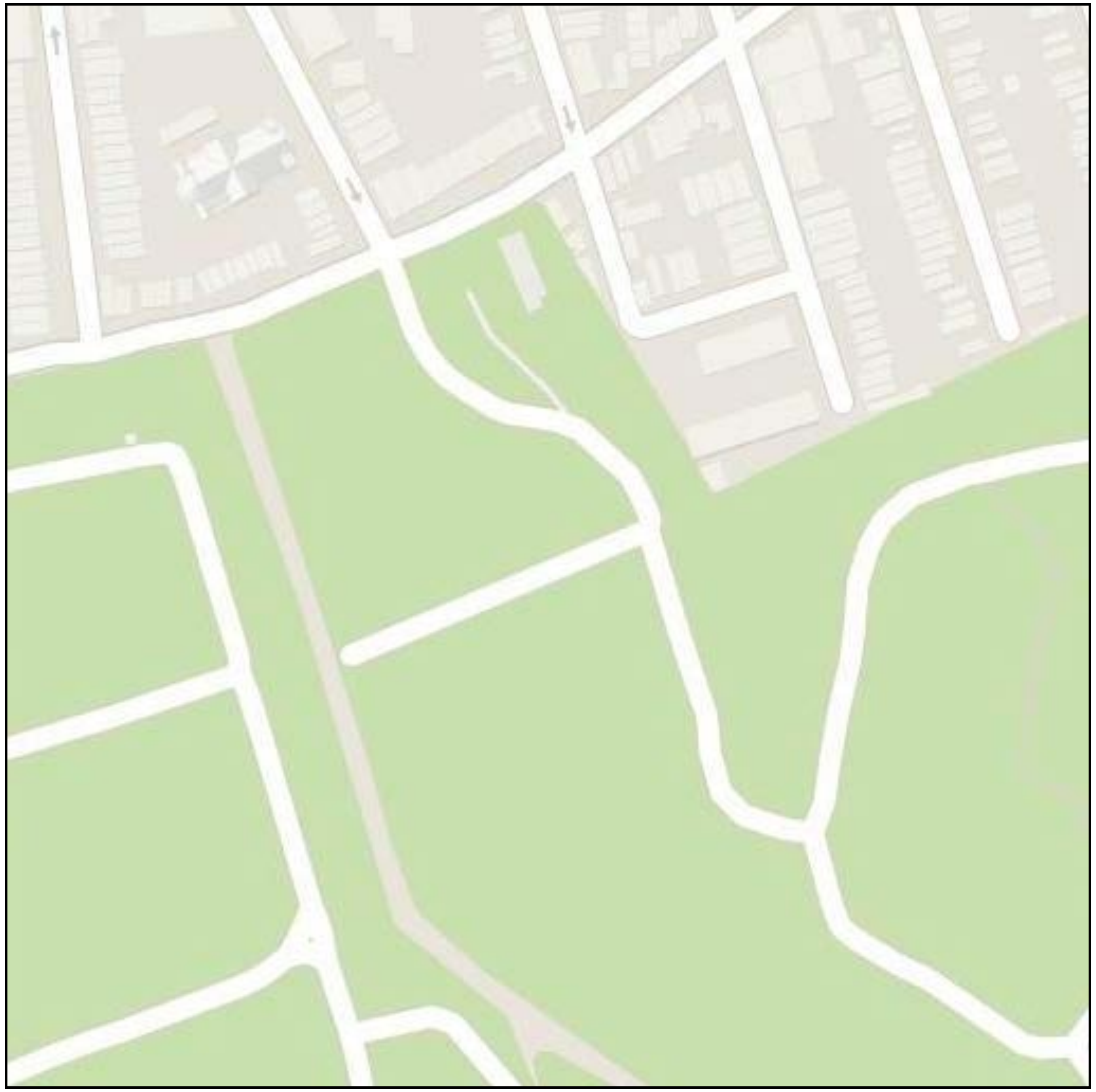


Why structured objectives

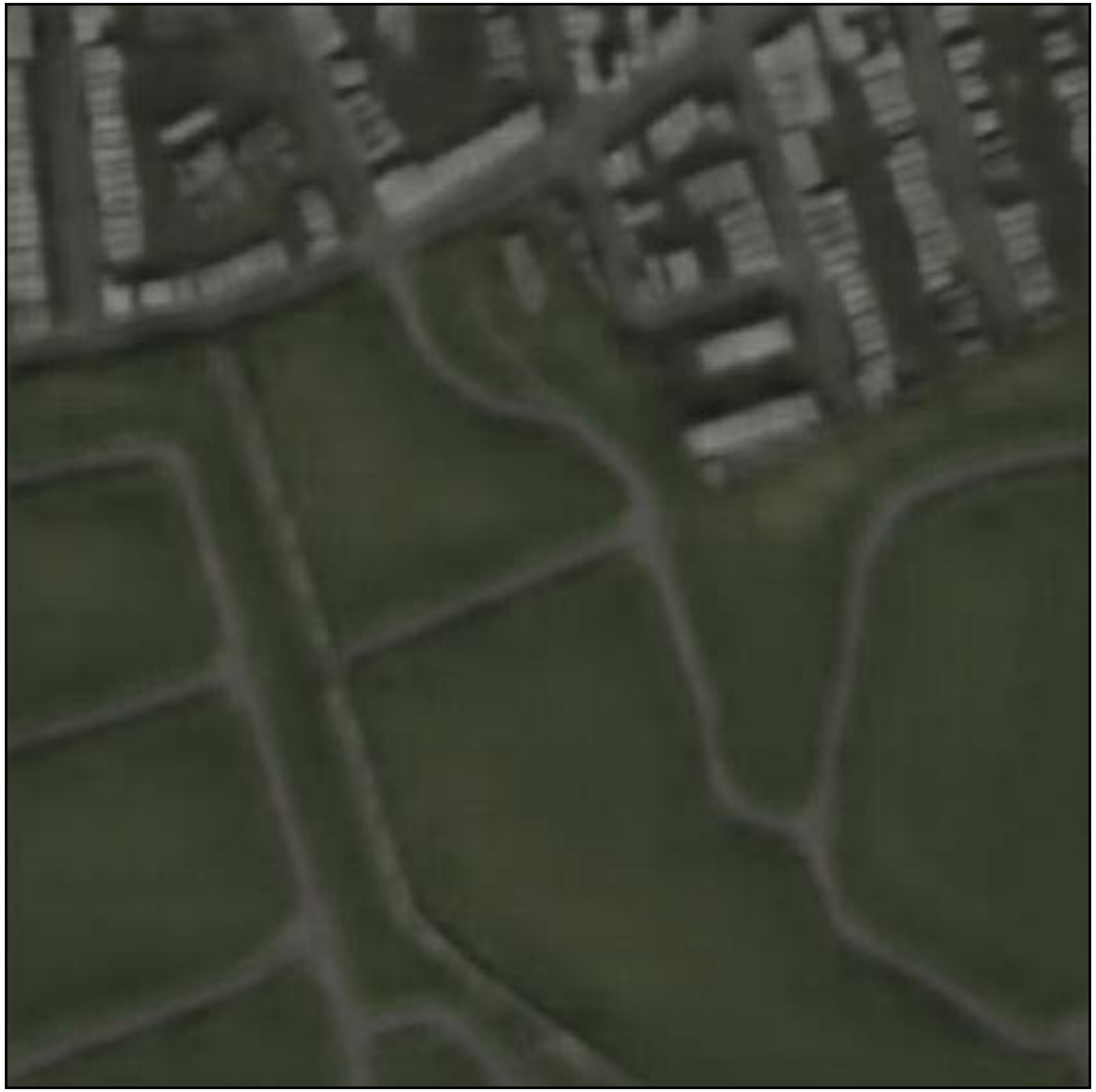
(cartoon)

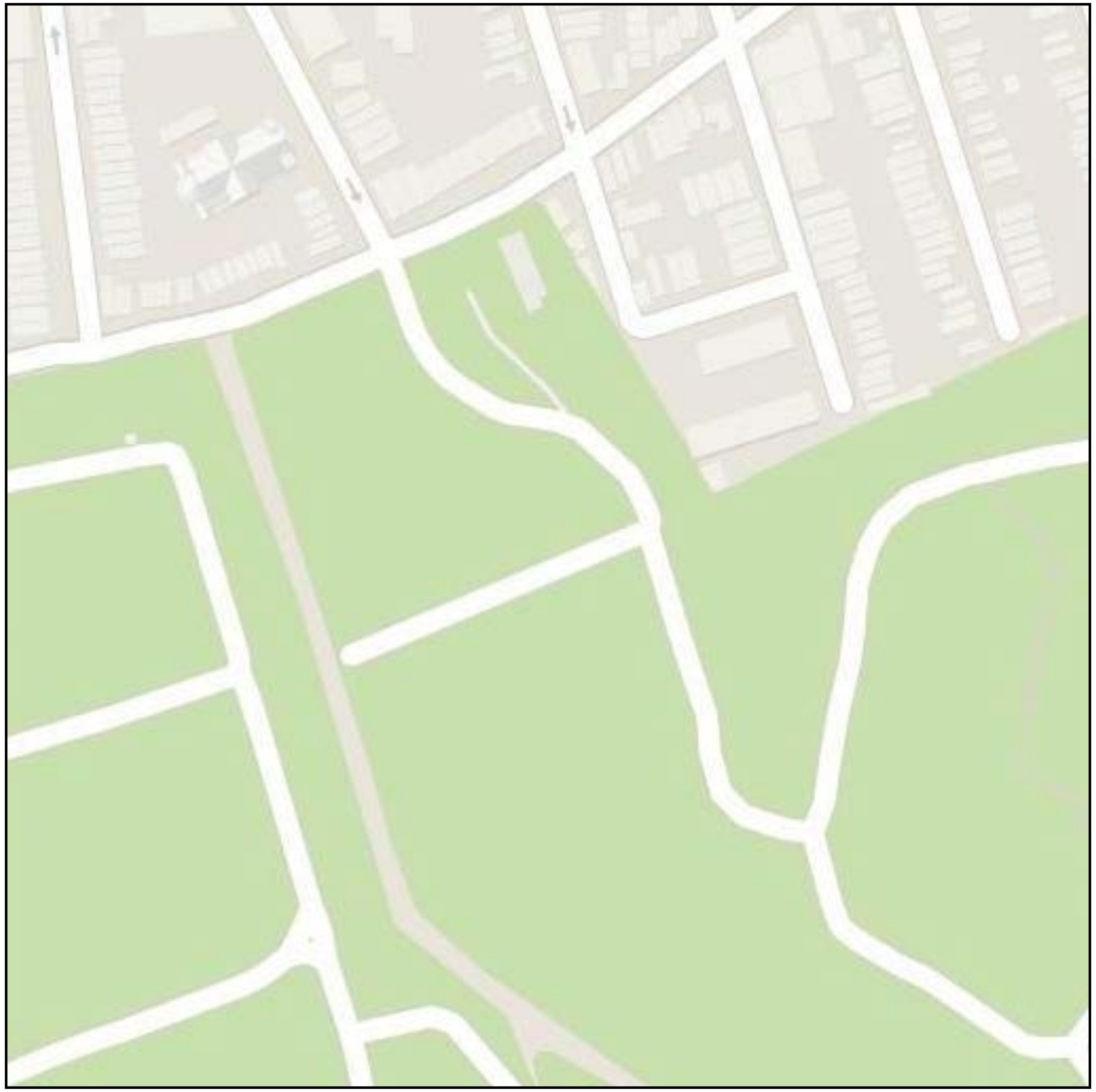


Input

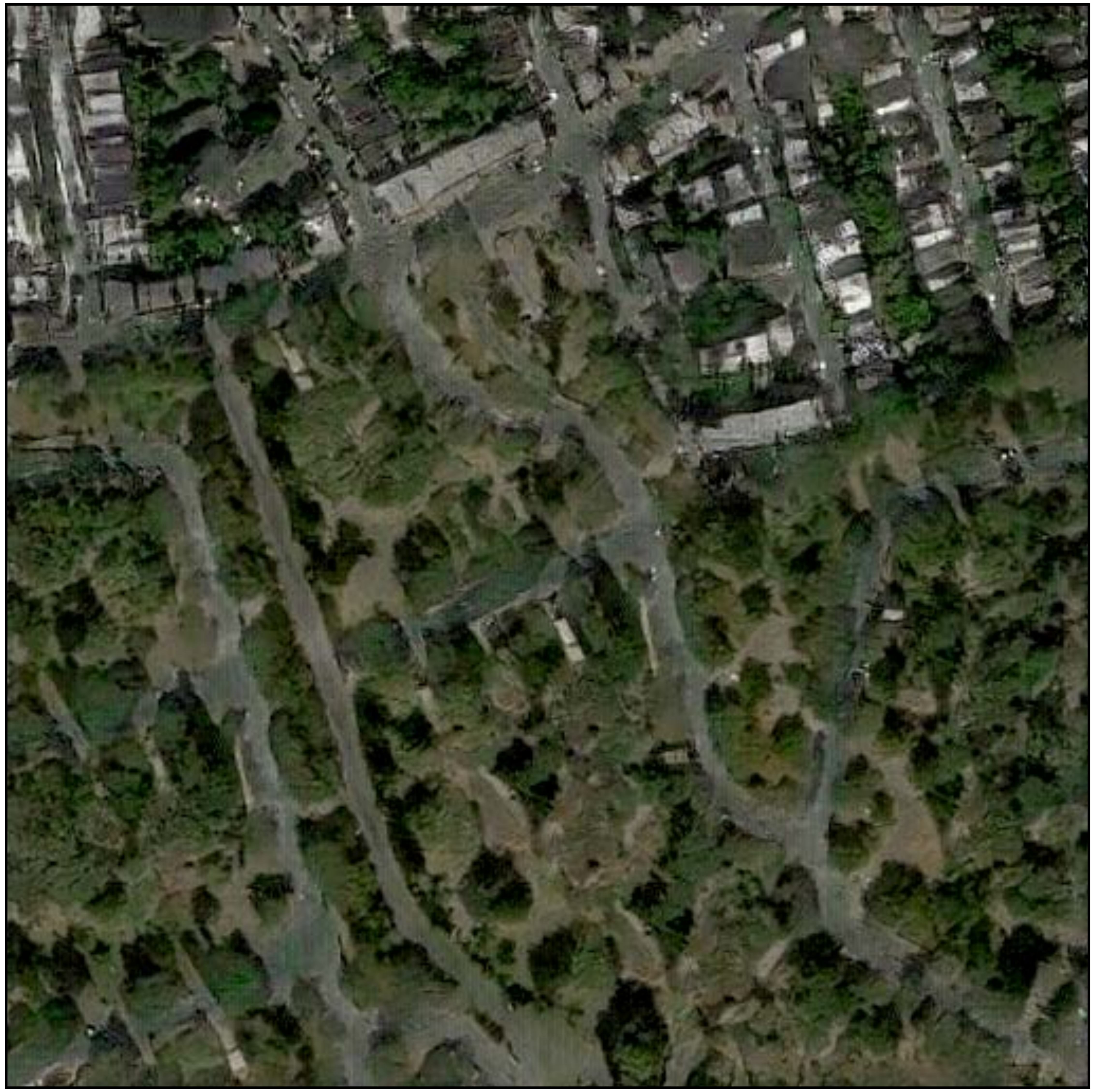


Unstructured prediction (L1)





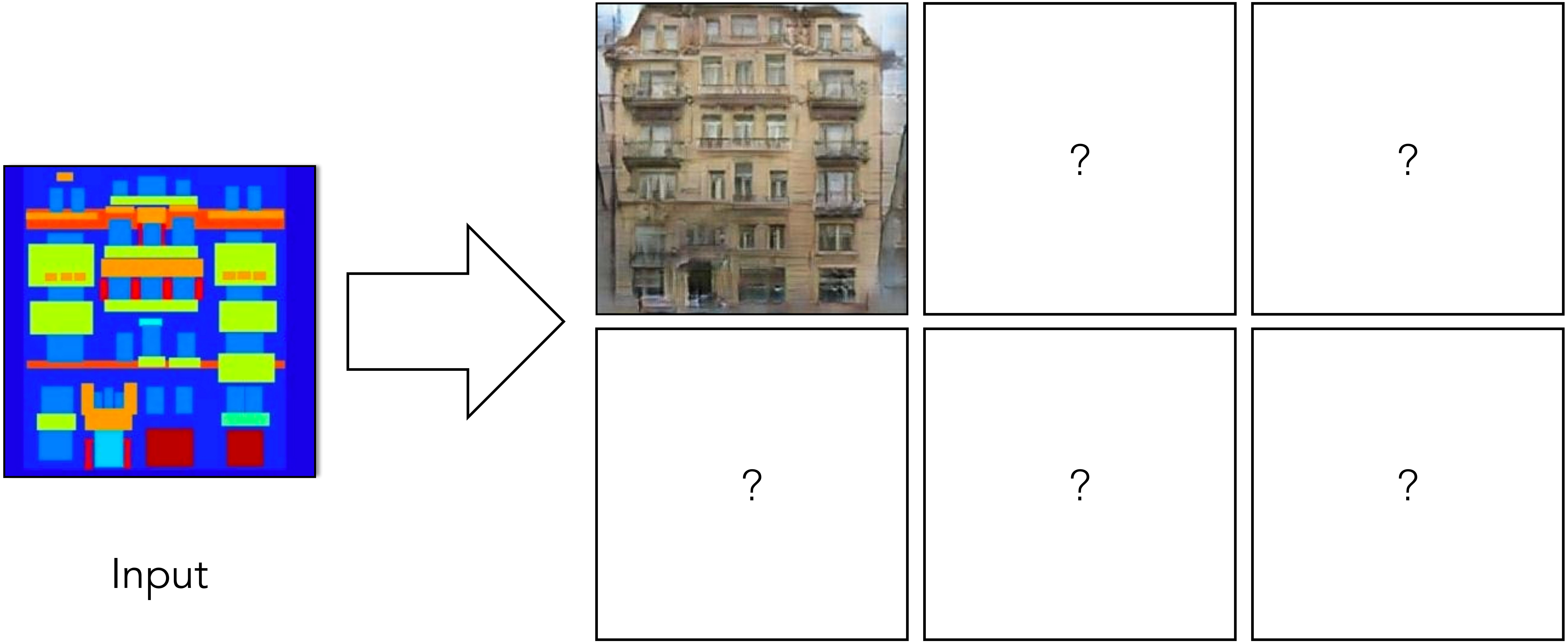
Input



Structured Prediction (cGAN)

Conditional VAE

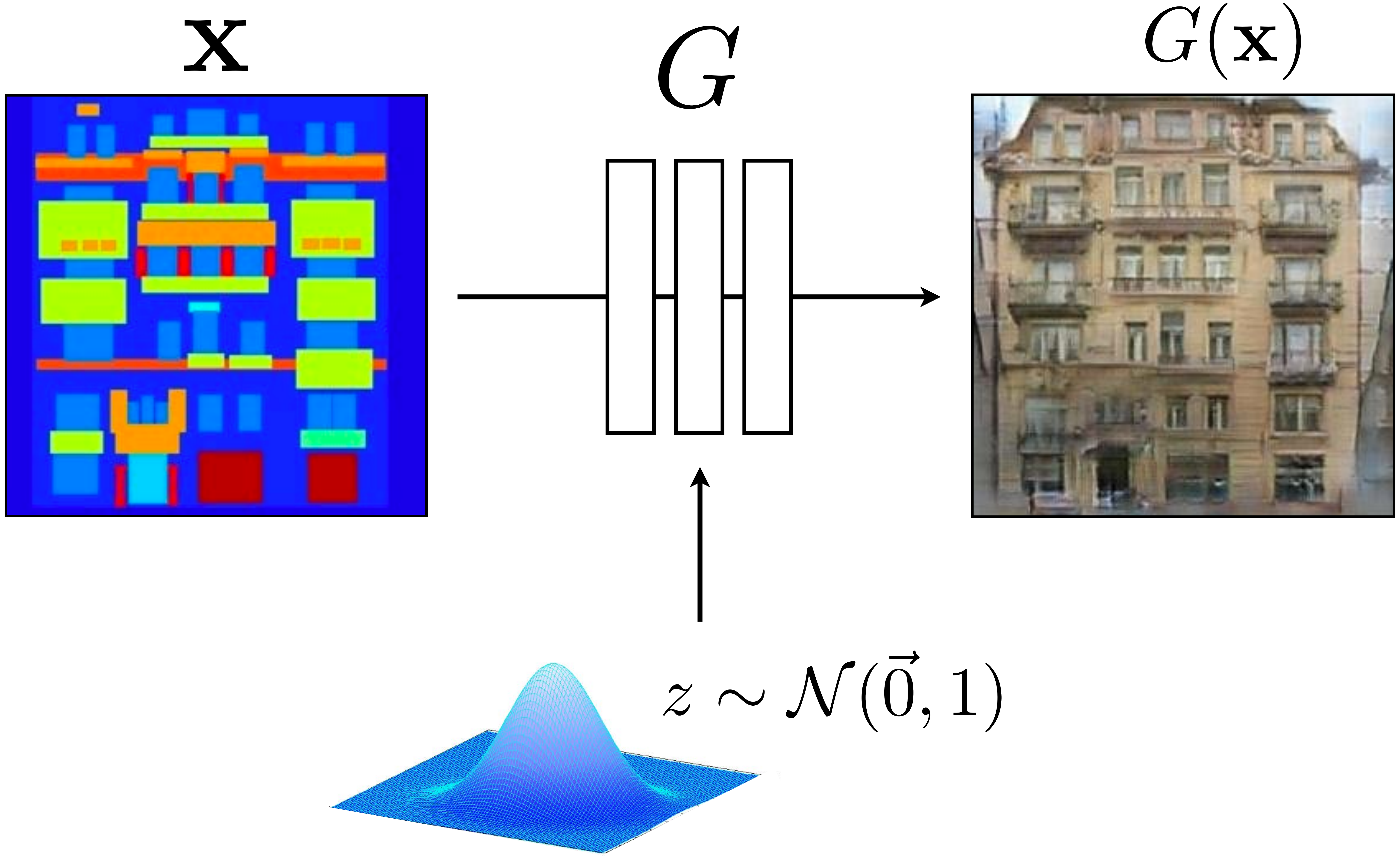
Modeling multiple possible outputs

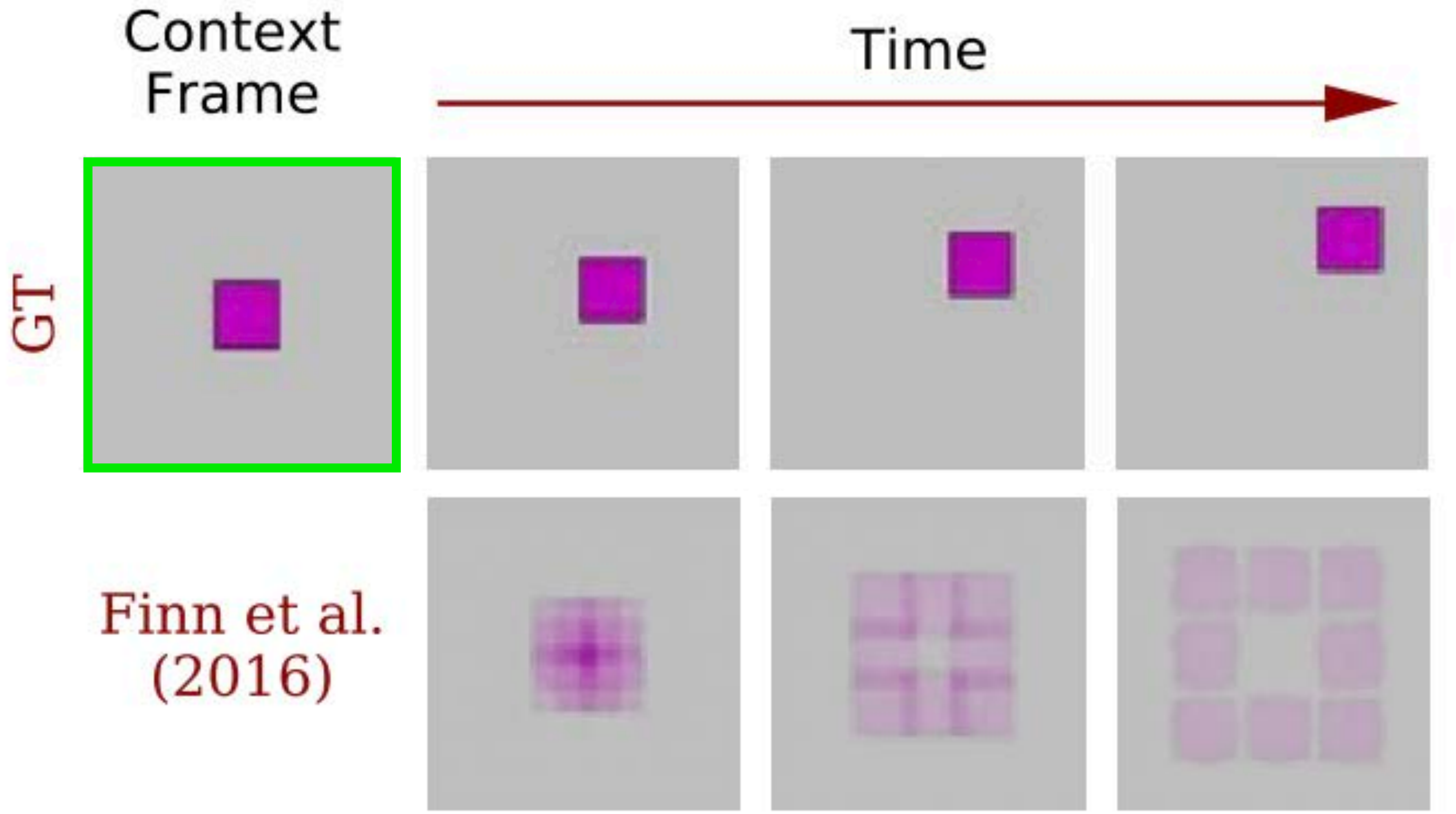


Input

Possible outputs

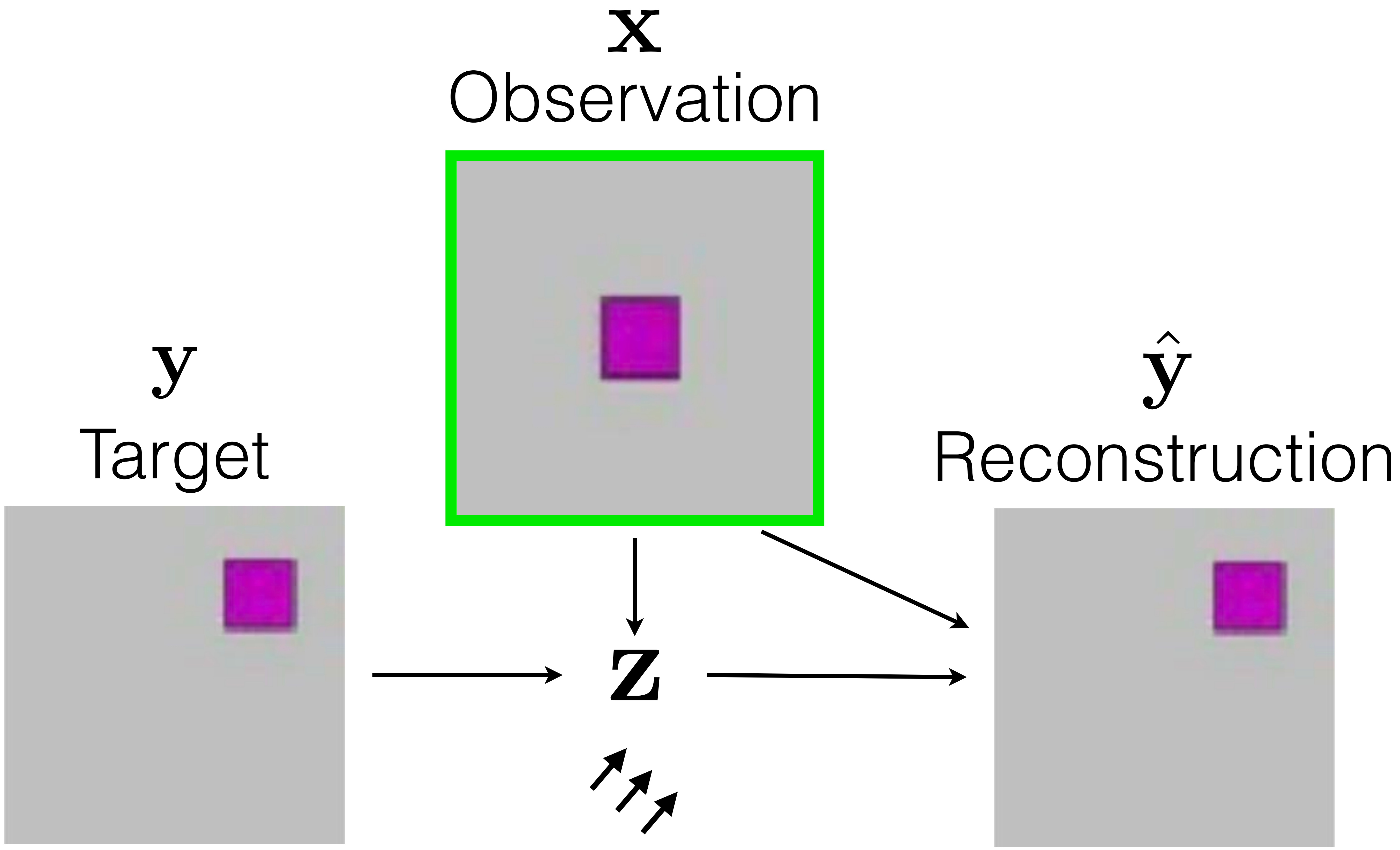
Modeling multiple possible outputs





Example from [Babaeizadeh et al., ICLR 2018]

see also [Walker et al., ECCV 2016], [Xue*, Wu*, et al., NIPS 2016]



$\mathbf{x}$
Observation

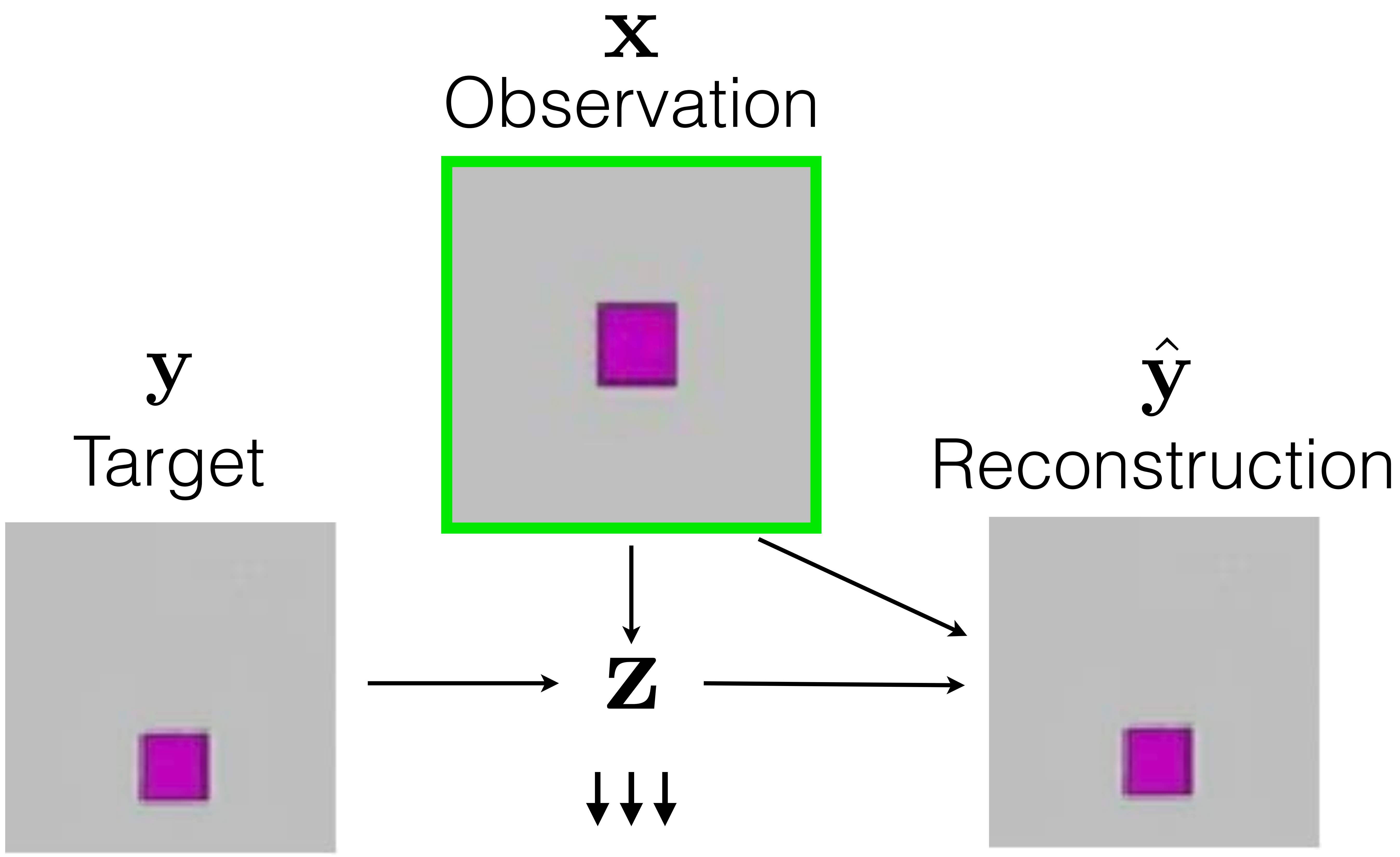
$\mathbf{y}$
Target

$\mathbf{z}$

$\hat{\mathbf{y}}$
Reconstruction

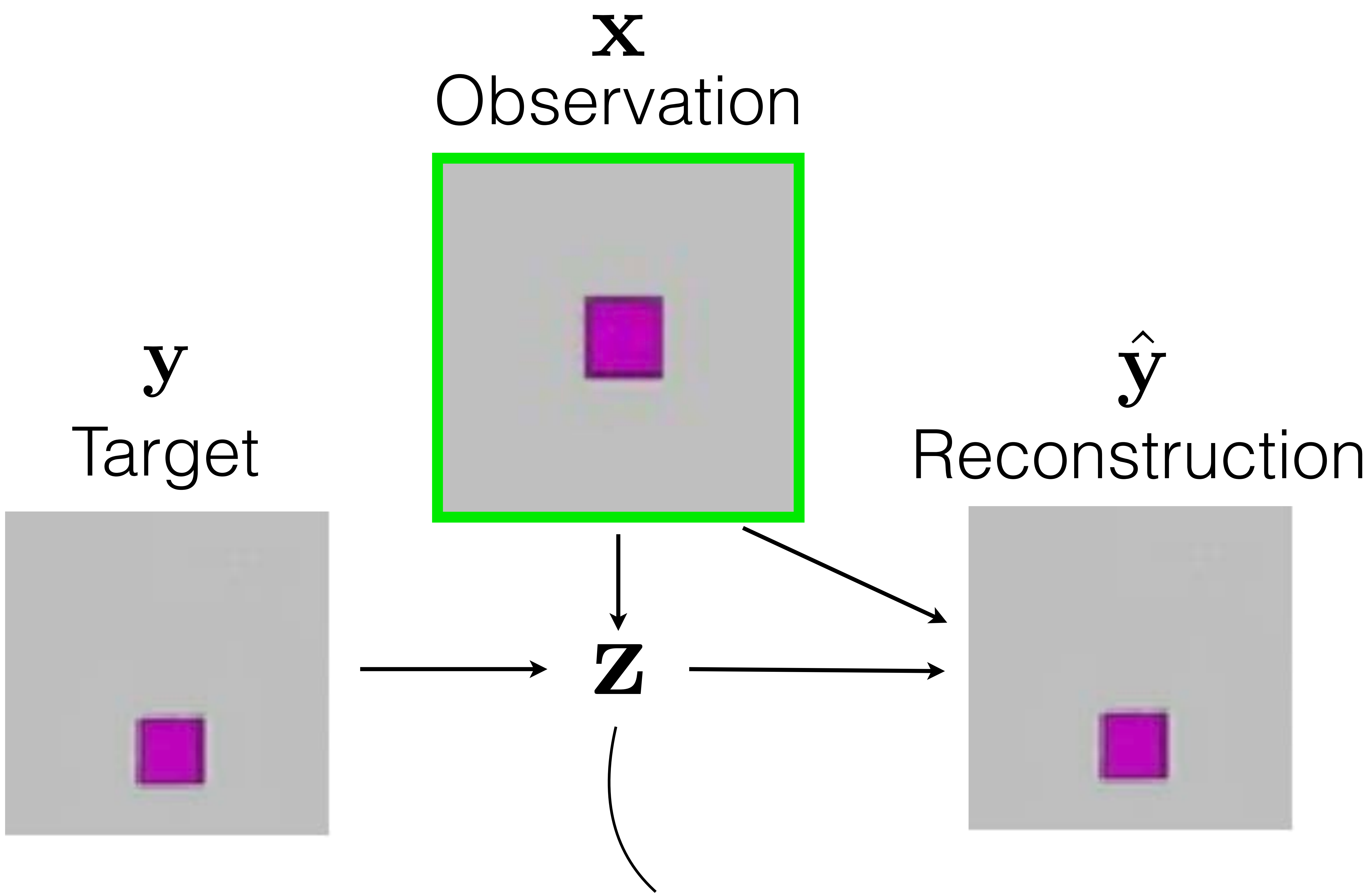
Example from [Babaeizadeh et al., ICLR 2018]

see also [Walker et al., ECCV 2016], [Xue*, Wu*, et al., NIPS 2016]



Example from [Babaeizadeh et al., ICLR 2018]

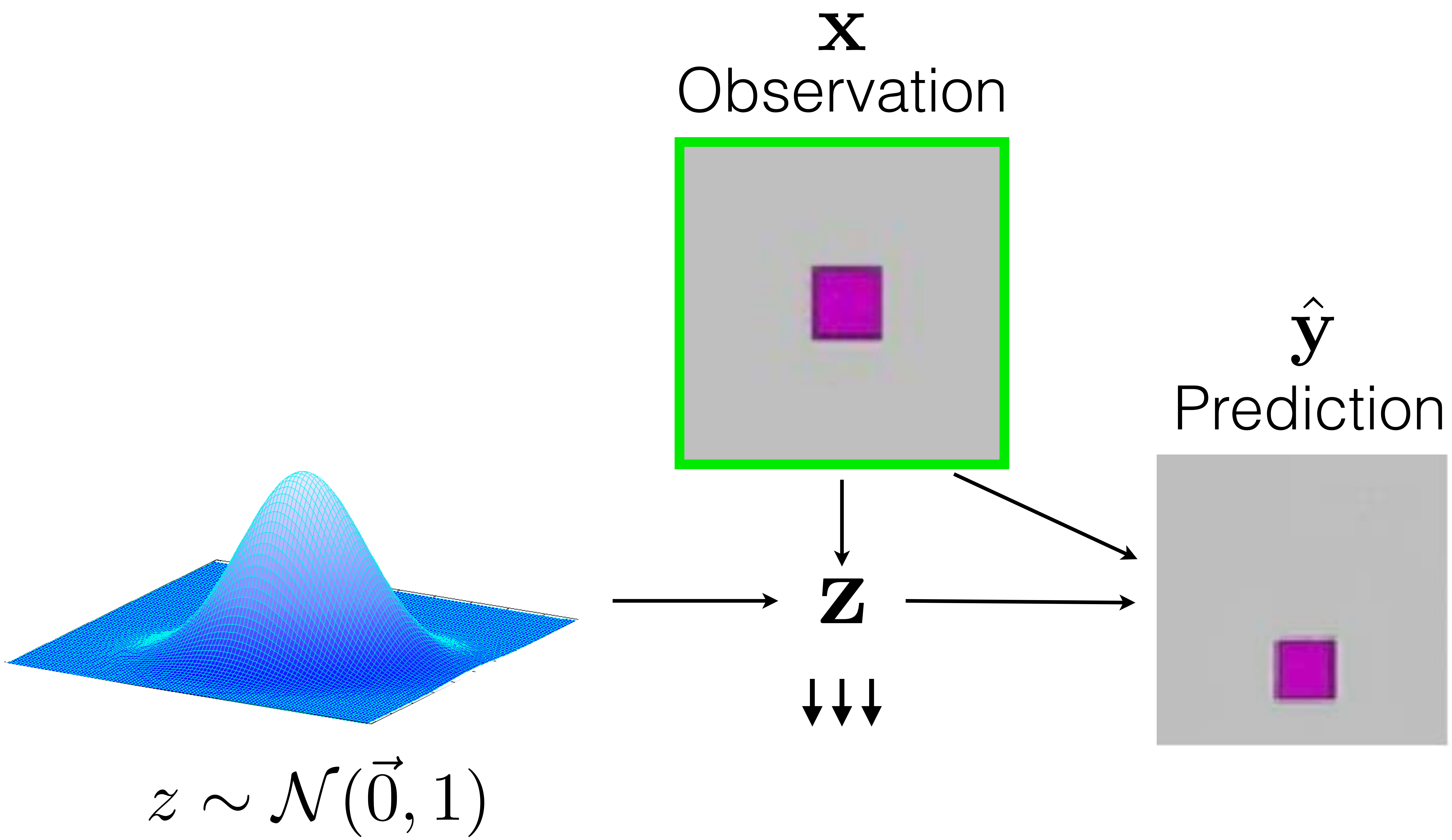
see also [Walker et al., ECCV 2016], [Xue*, Wu*, et al., NIPS 2016]



z learns to encode the missing information necessary to predict y from x, i.e. the direction in which the purple box moves

Example from [Babaeizadeh et al., ICLR 2018]

see also [Walker et al., ECCV 2016], [Xue*, Wu*, et al., NIPS 2016]



$\mathbf{x}$

Observation

$\hat{\mathbf{y}}$

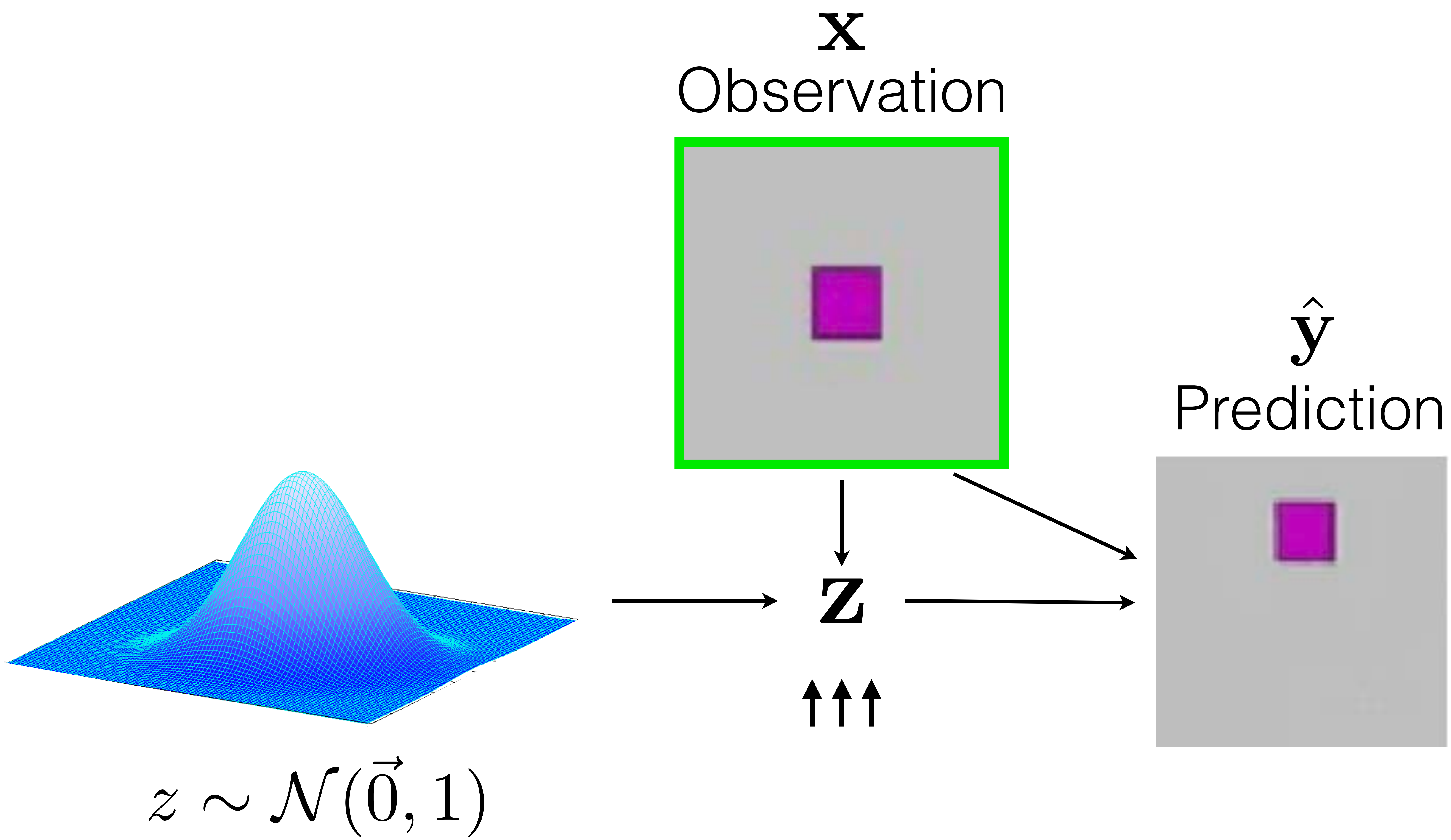
Prediction

$\mathbf{z}$

$z \sim \mathcal{N}(\vec{0}, 1)$

Example from [Babaeizadeh et al., ICLR 2018]

see also [Walker et al., ECCV 2016], [Xue*, Wu*, et al., NIPS 2016]



$\mathbf{x}$

Observation

$\hat{\mathbf{y}}$

Prediction

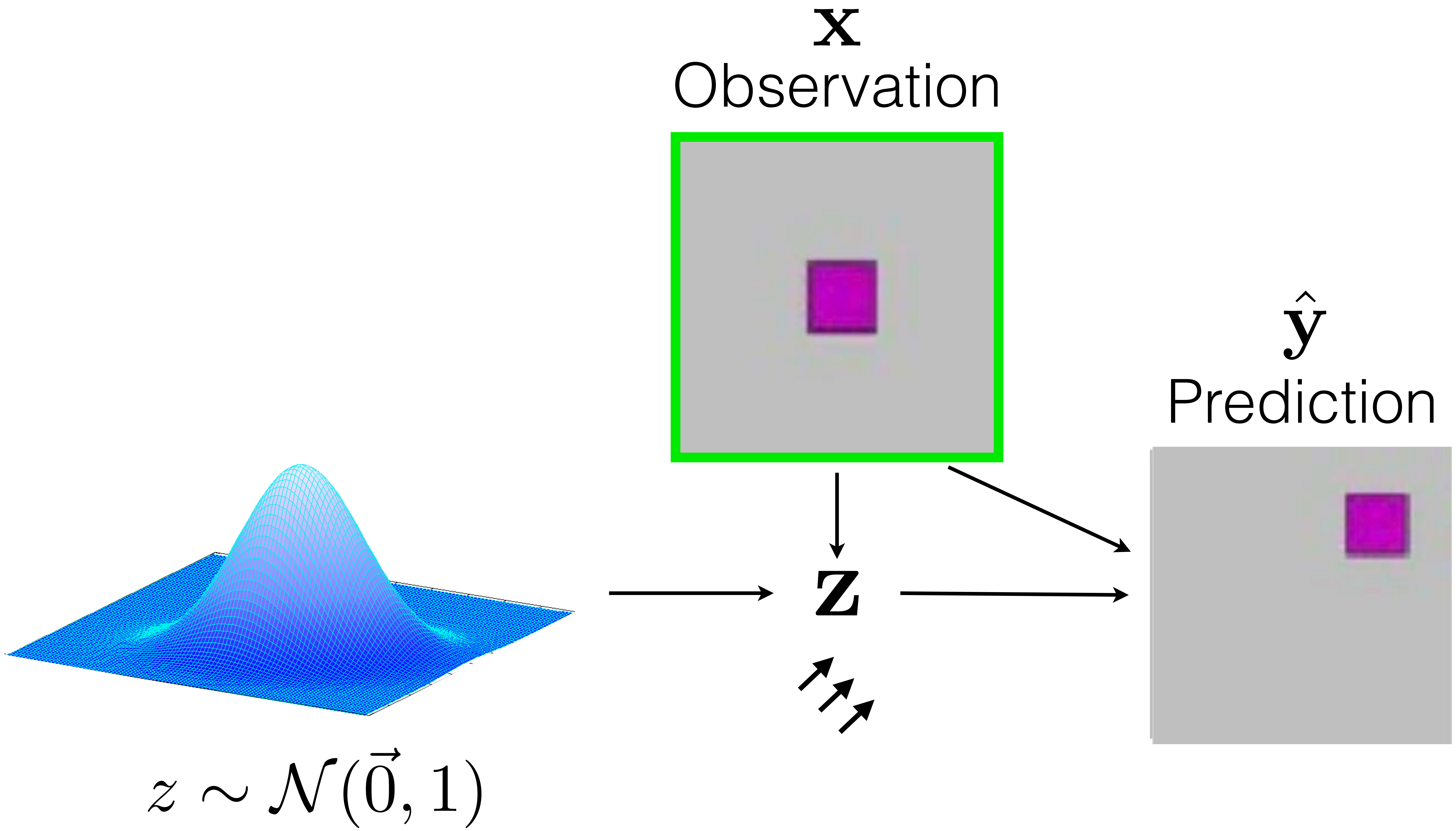
$\mathbf{z}$

↑↑↑

$z \sim \mathcal{N}(\vec{0}, 1)$

Example from [Babaeizadeh et al., ICLR 2018]

see also [Walker et al., ECCV 2016], [Xue*, Wu*, et al., NIPS 2016]



$\mathbf{x}$

Observation

$\hat{\mathbf{y}}$

Prediction

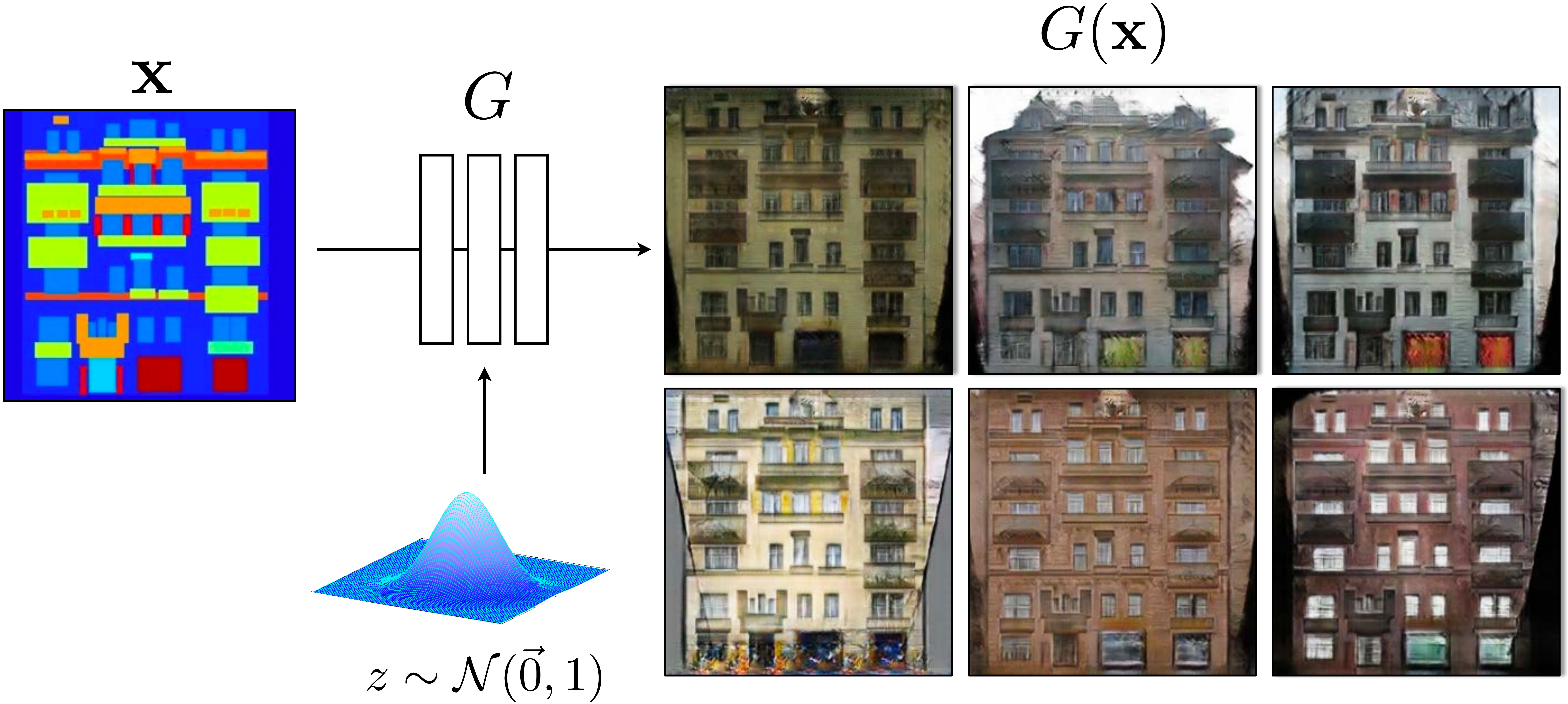
$\mathbf{z}$

$z \sim \mathcal{N}(\vec{0}, 1)$

Example from [Babaeizadeh et al., ICLR 2018]

see also [Walker et al., ECCV 2016], [Xue*, Wu*, et al., NIPS 2016]

Modeling multiple possible outputs



[BiCycleGAN, Zhu et al., NeurIPS 2017]

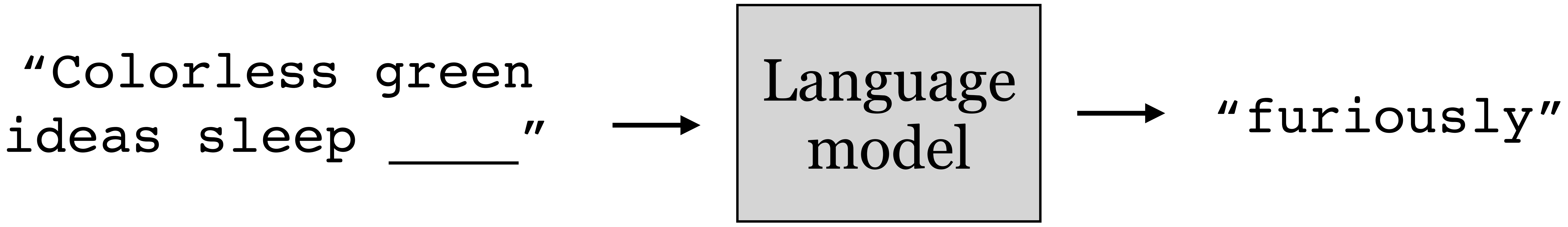
Concept: **You can control your data either via explicit inputs or latent variables**

Text-to-text

↳ with autoregressive models

Autoregressive models are conditional generative models

1. **Training**



“Colorless green ideas sleep ____”

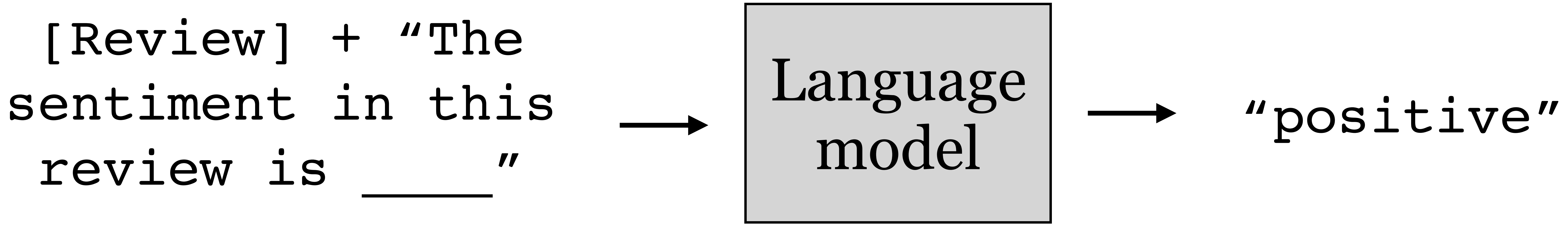
Language model

“furiously”

(Predict next characters)

Autoregressive models are conditional generative models

2. **Predicting**



`[Review] + “The sentiment in this review is ____”`

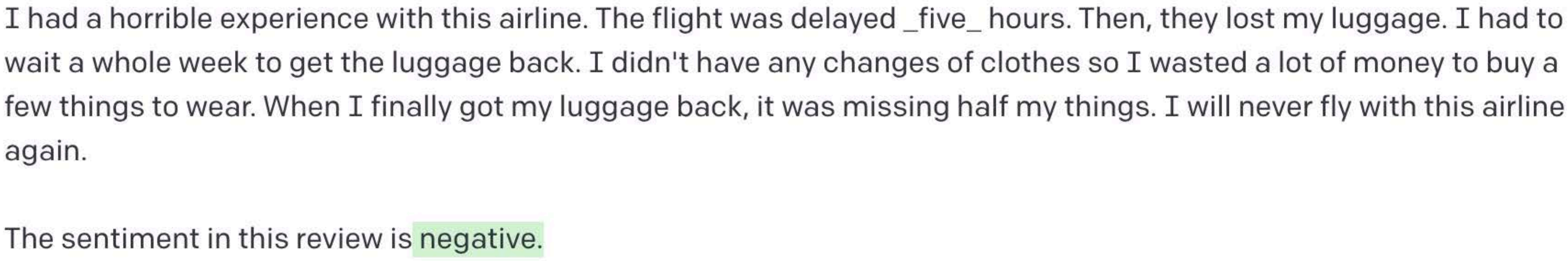
Language model

`“positive”`

(Predict next characters)

GPT-3

[Brown et al., 2020] https://arxiv.org/pdf/2005.14165.pdf

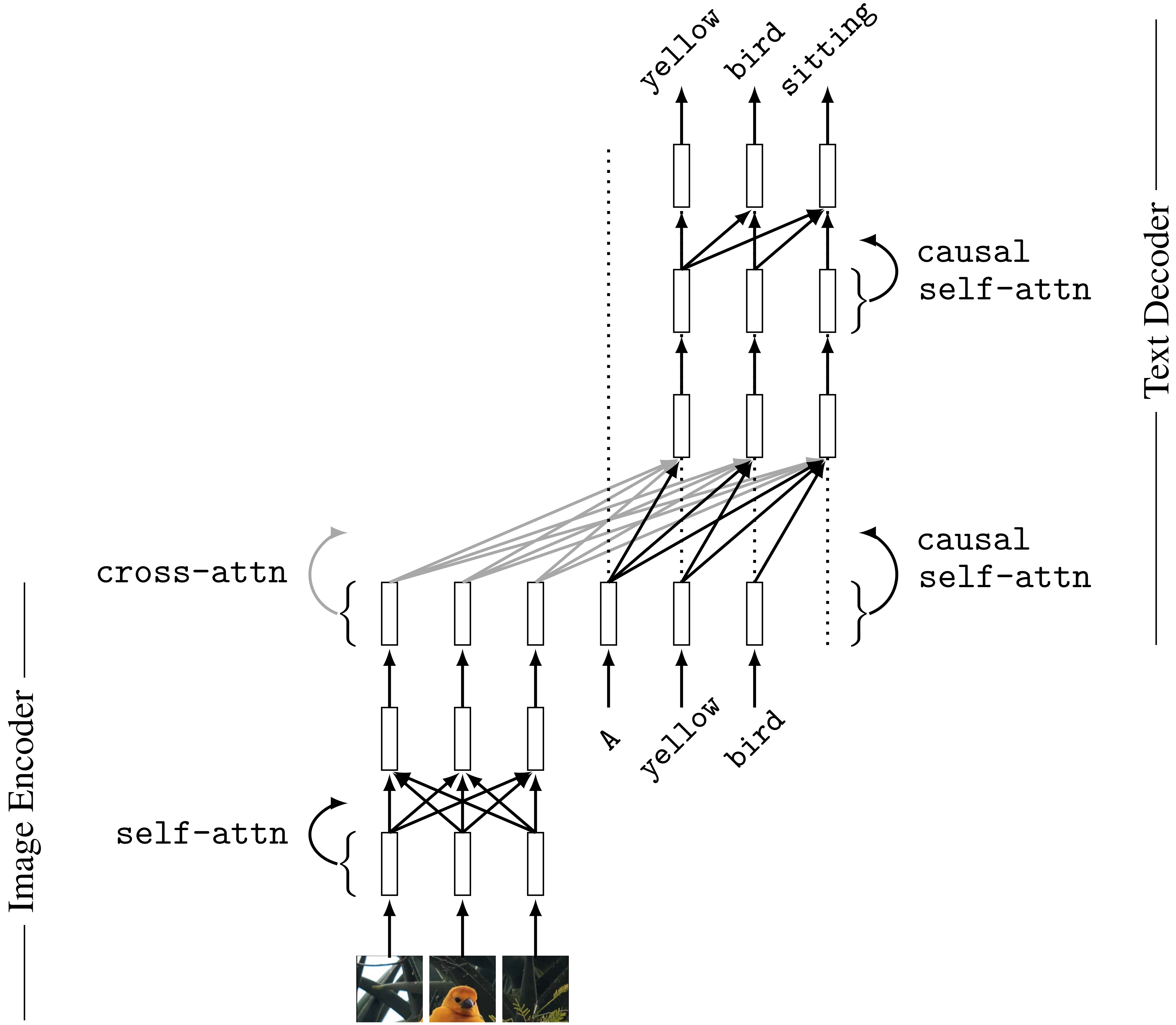


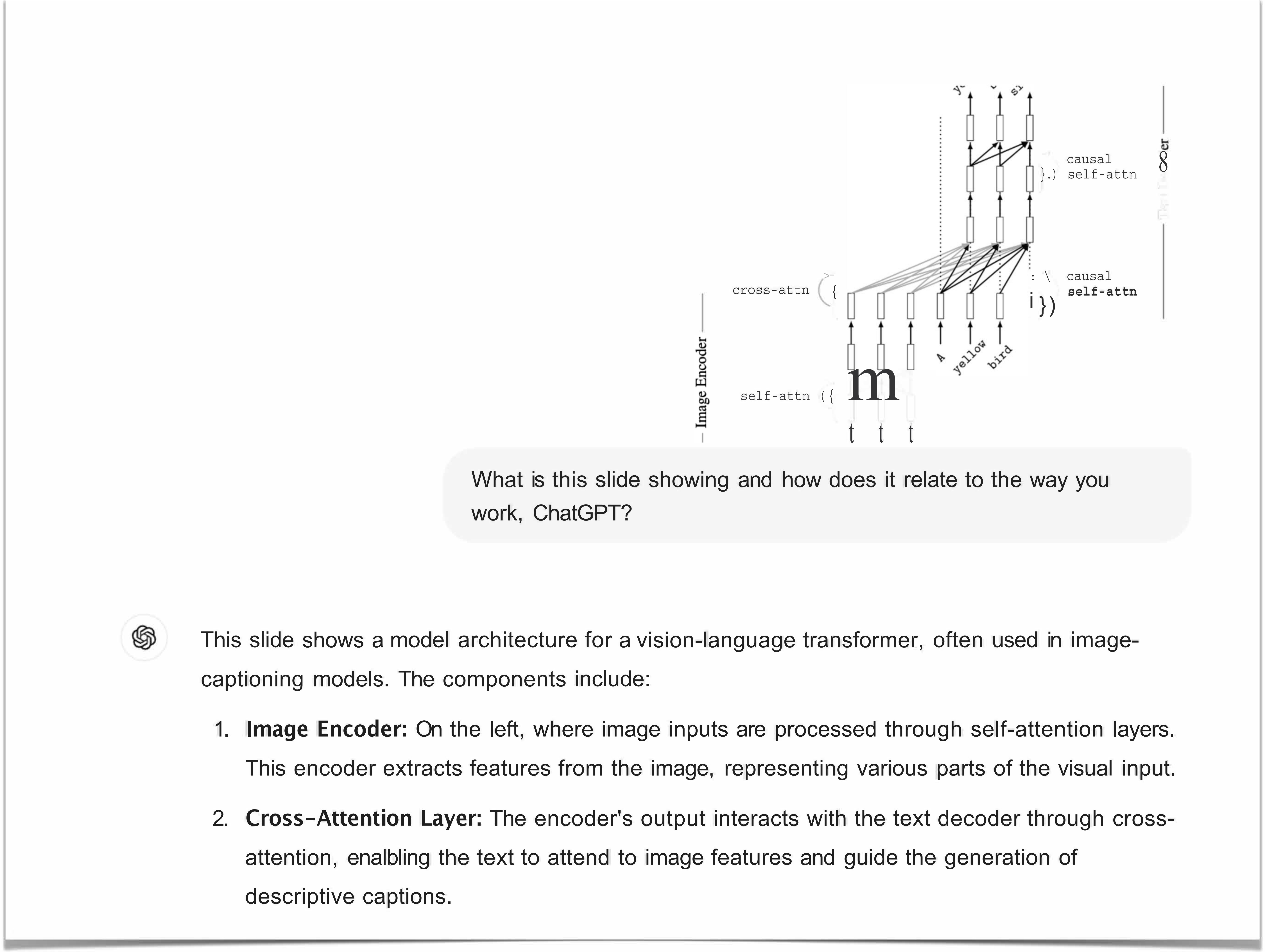
I had a horrible experience with this airline. The flight was delayed _five_ hours. Then, they lost my luggage. I had to wait a whole week to get the luggage back. I didn't have any changes of clothes so I wasted a lot of money to buy a few things to wear. When I finally got my luggage back, it was missing half my things. I will never fly with this airline again.

The sentiment in this review is negative.

Image-to-text

↳ with an autoregressive model


yellow
bird
sitting
causal
self-attn
Text Decoder
cross-attn
causal
self-attn
A
yellow
bird
self-attn
Image Encoder



What is this slide showing and how does it relate to the way you work, ChatGPT?

This slide shows a model architecture for a vision-language transformer, often used in image-captioning models. The components include:

1. **Image Encoder:** On the left, where image inputs are processed through self-attention layers. This encoder extracts features from the image, representing various parts of the visual input.
2. **Cross-Attention Layer:** The encoder's output interacts with the text decoder through cross-attention, enalbling the text to attend to image features and guide the generation of descriptive captions.

Created with ChatGPT.

Does ChatGPT know how its own brain works?

How This Relates to How I Work

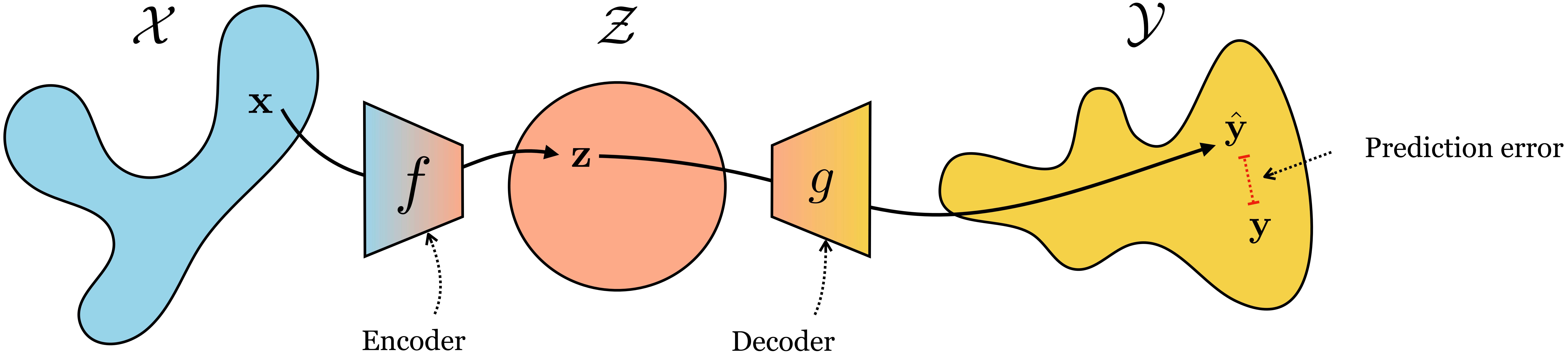
Similar to this architecture, my responses often involve "attention" to different parts of your input. Instead of visual data, I focus on your conversational context and keywords, determining the next part of the response based on previous parts of the conversation. In this way, I function similarly to a text decoder with attention mechanisms to retain consistency and relevance in dialogue. However, unlike this model's image encoding, I don't process visual data directly, but rely on text representations and information you share about images.

Created with ChatGPT.

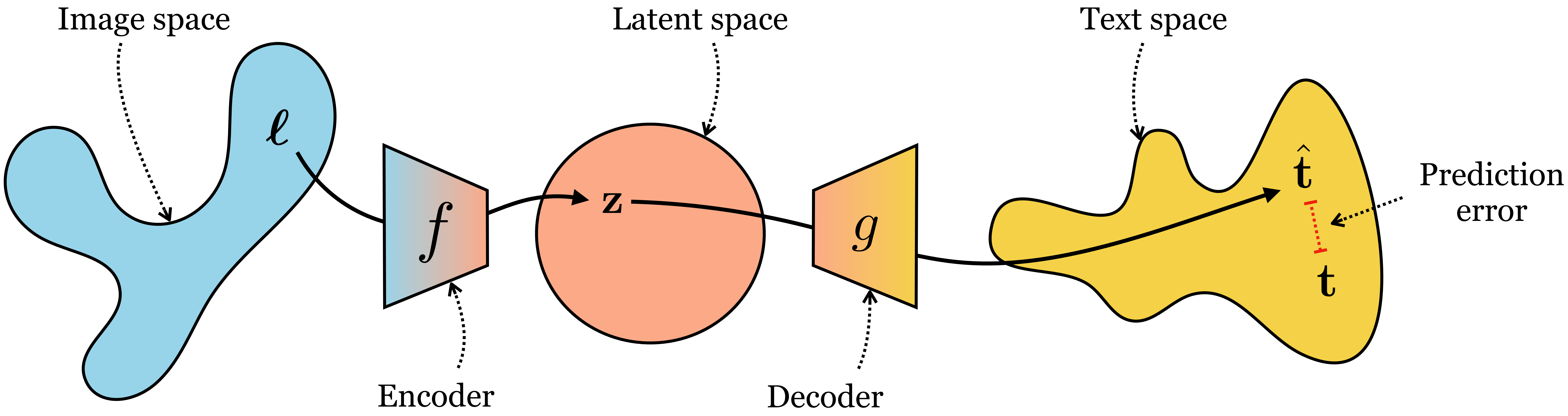
Text-to-image

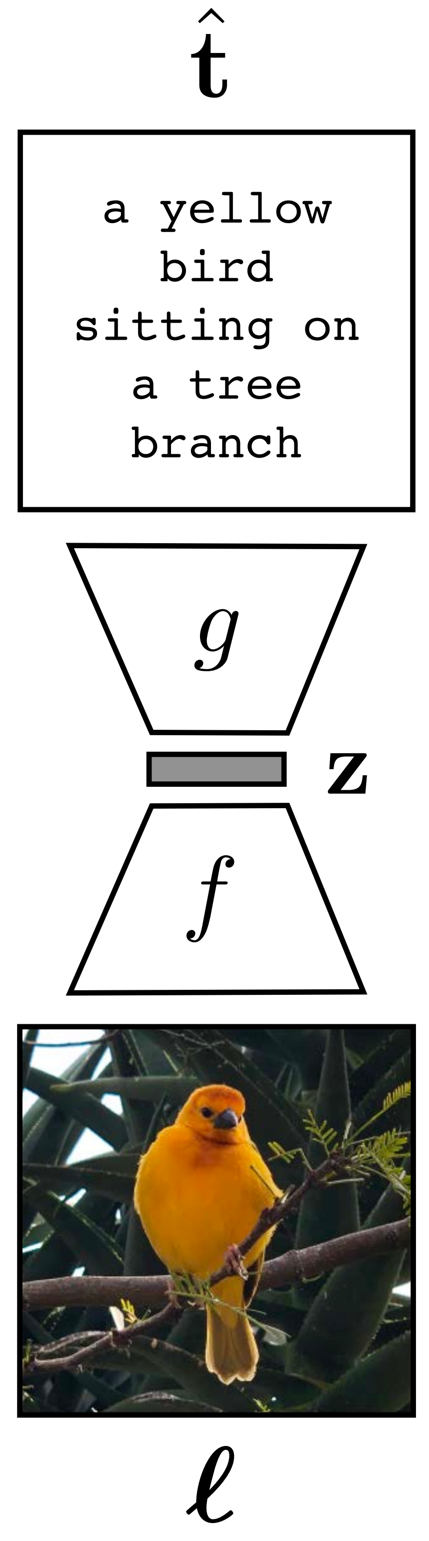
↳ with all models combined!

General strategy for domain translation

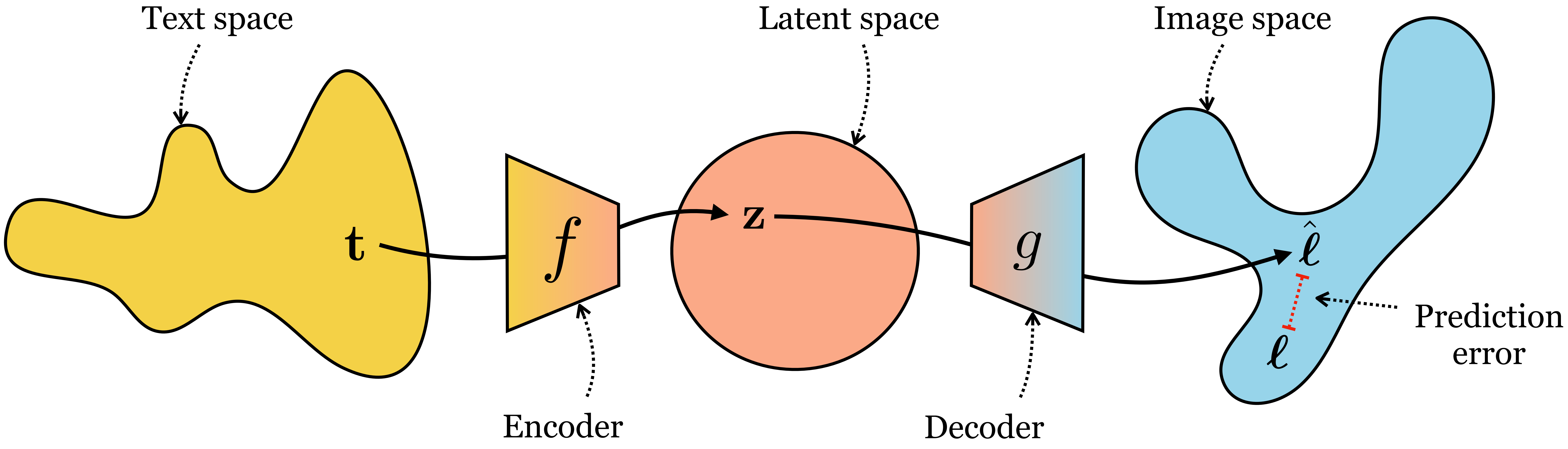


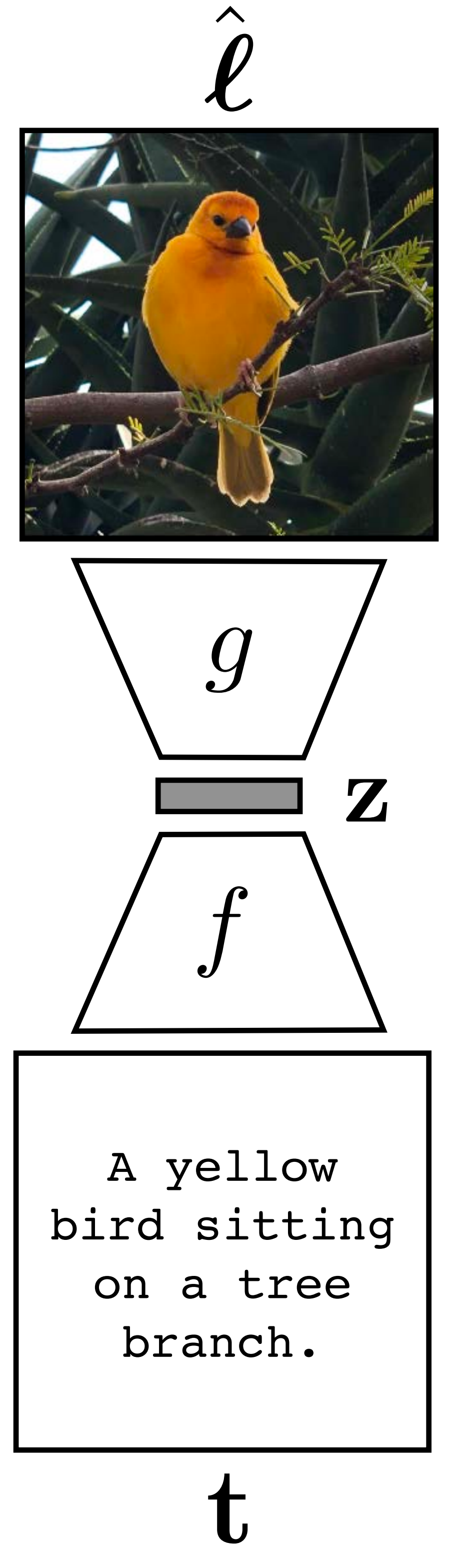
Text-to-image





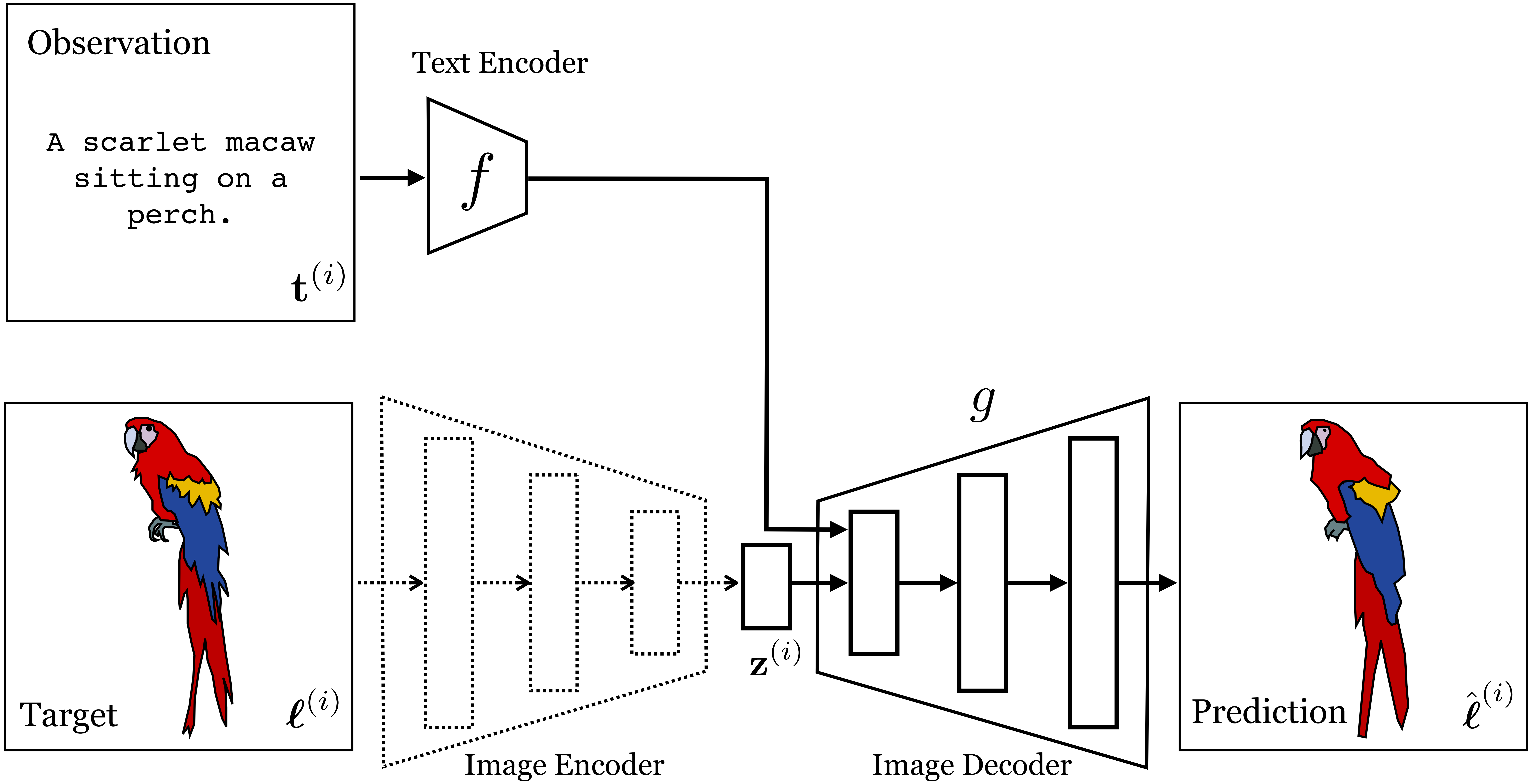
Text-to-image





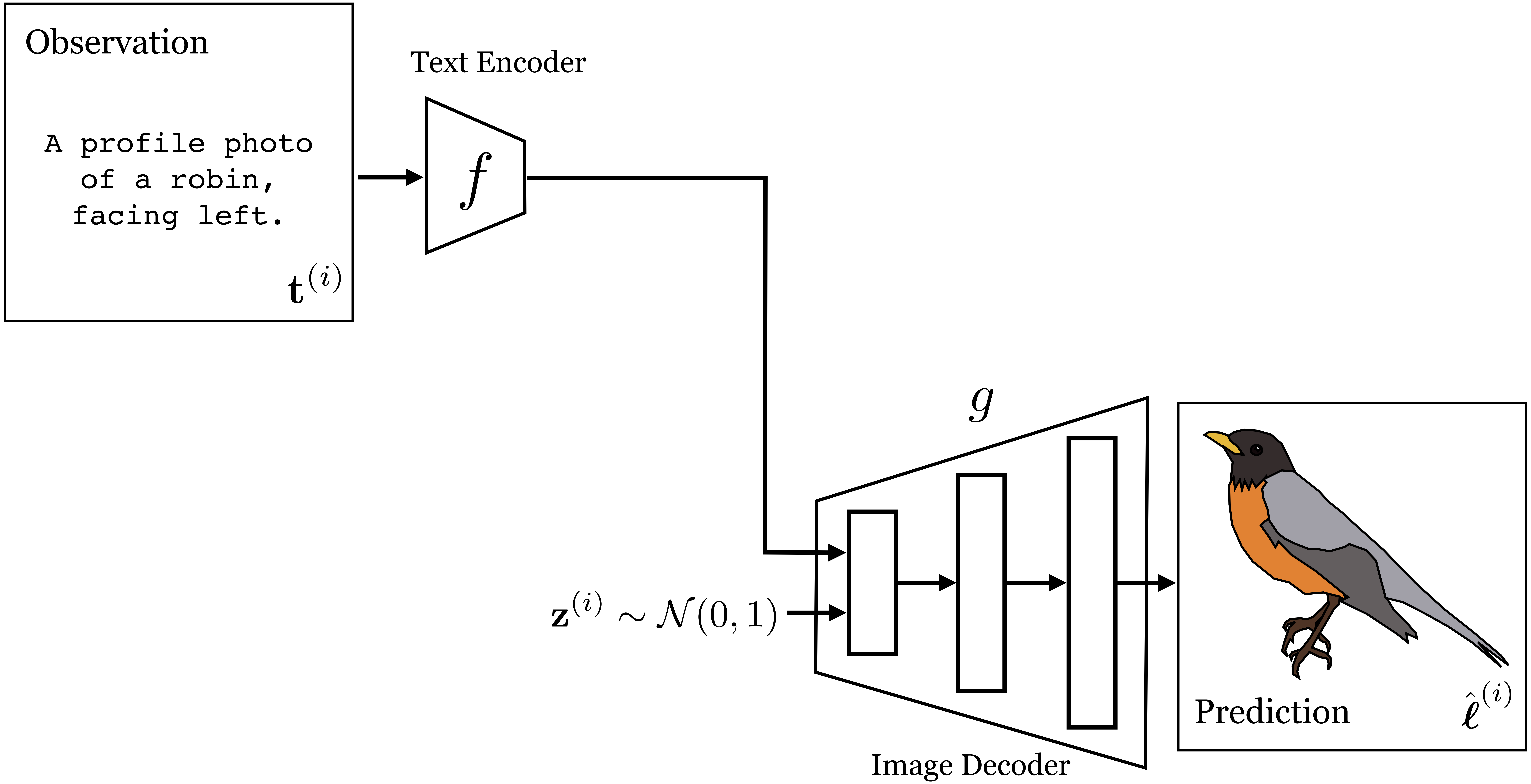
A yellow bird sitting on a tree branch.

Text-to-image VAE



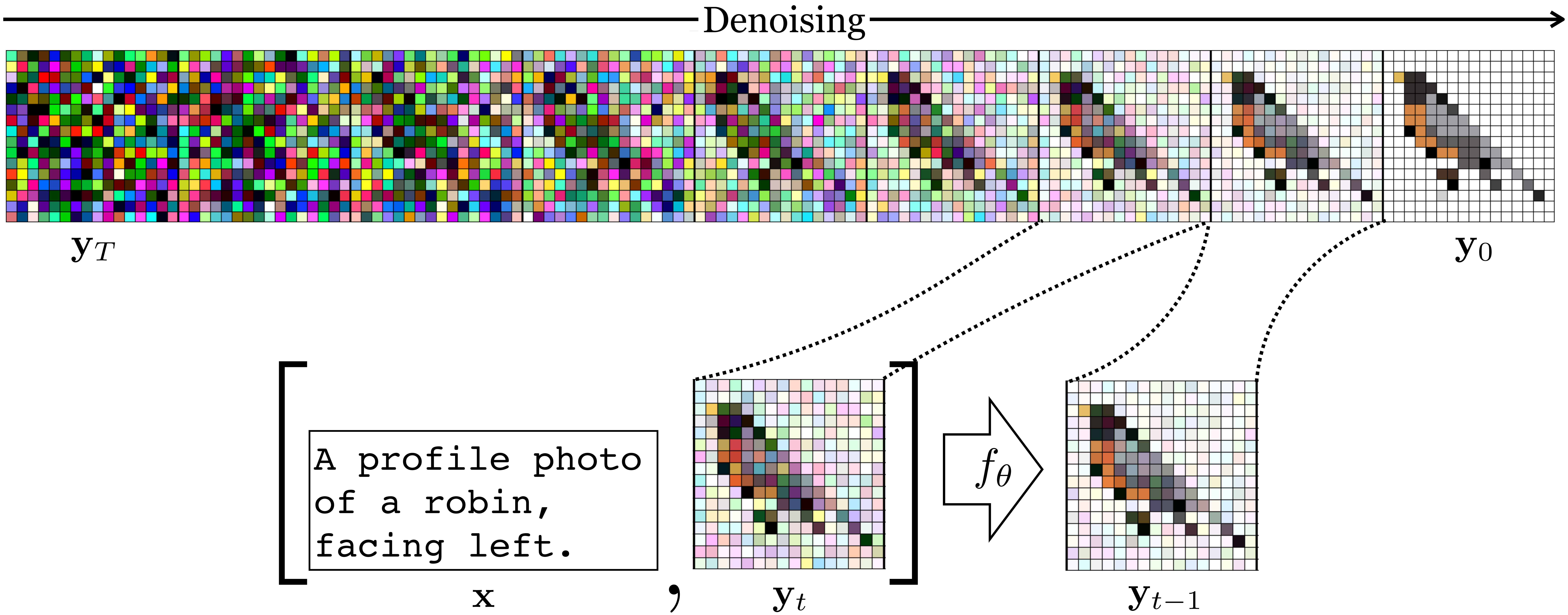
For example: DALL-E 1 [Ramesh et al. 2021]

Text-to-image VAE


Observation
A profile photo of a robin, facing left.
$\mathbf{t}^{(i)}$
Text Encoder
f
g
$\mathbf{z}^{(i)} \sim \mathcal{N}(0,1)$
Image Decoder
Prediction
$\hat{\boldsymbol{\ell}}^{(i)}$

For example: DALL-E 1 [Ramesh et al. 2021]

Text-to-image diffusion model

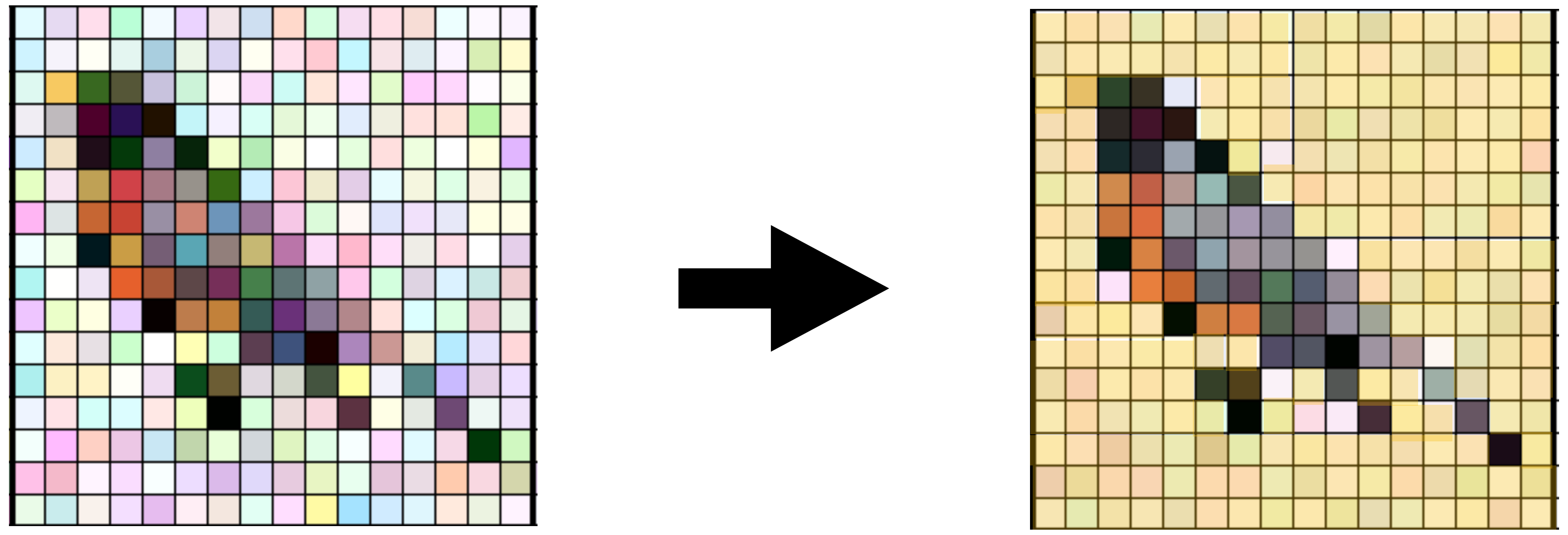


`A profile photo of a robin, facing left.`

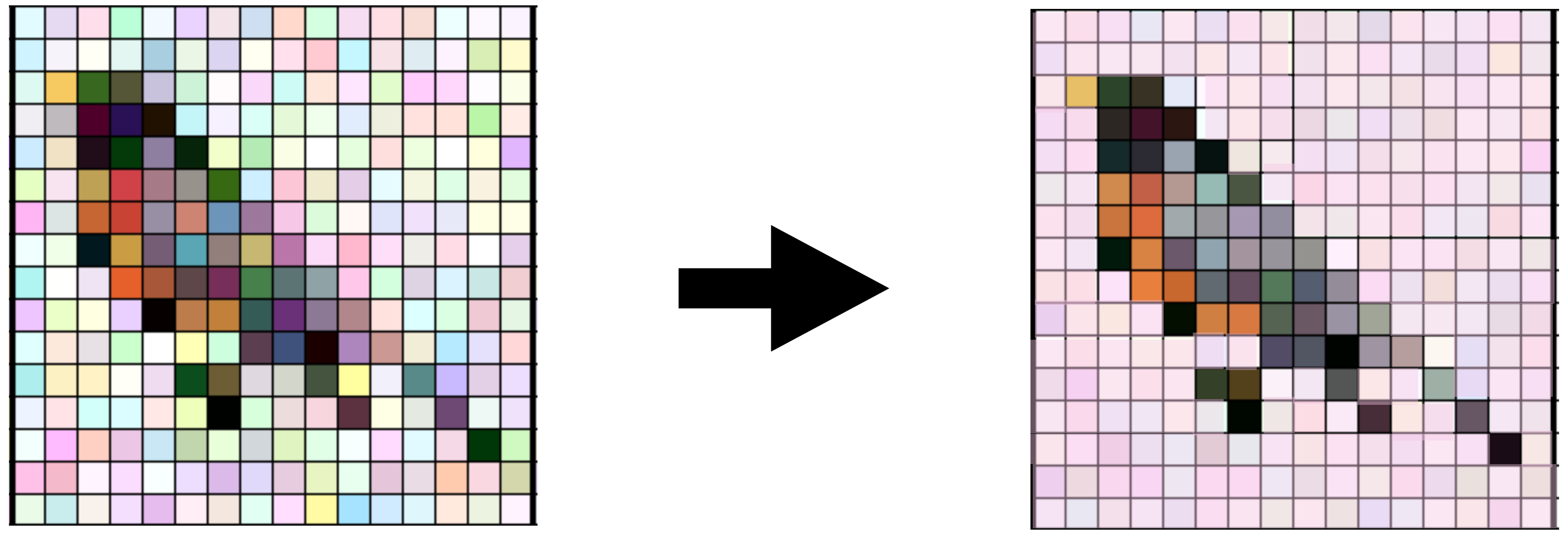
For example: DALL-E 2 [Ramesh et al. 2022], Stable Diffusion [Rombach*, Blattman* et al. 2022]

Text provides information as to what the noiseless image should look like

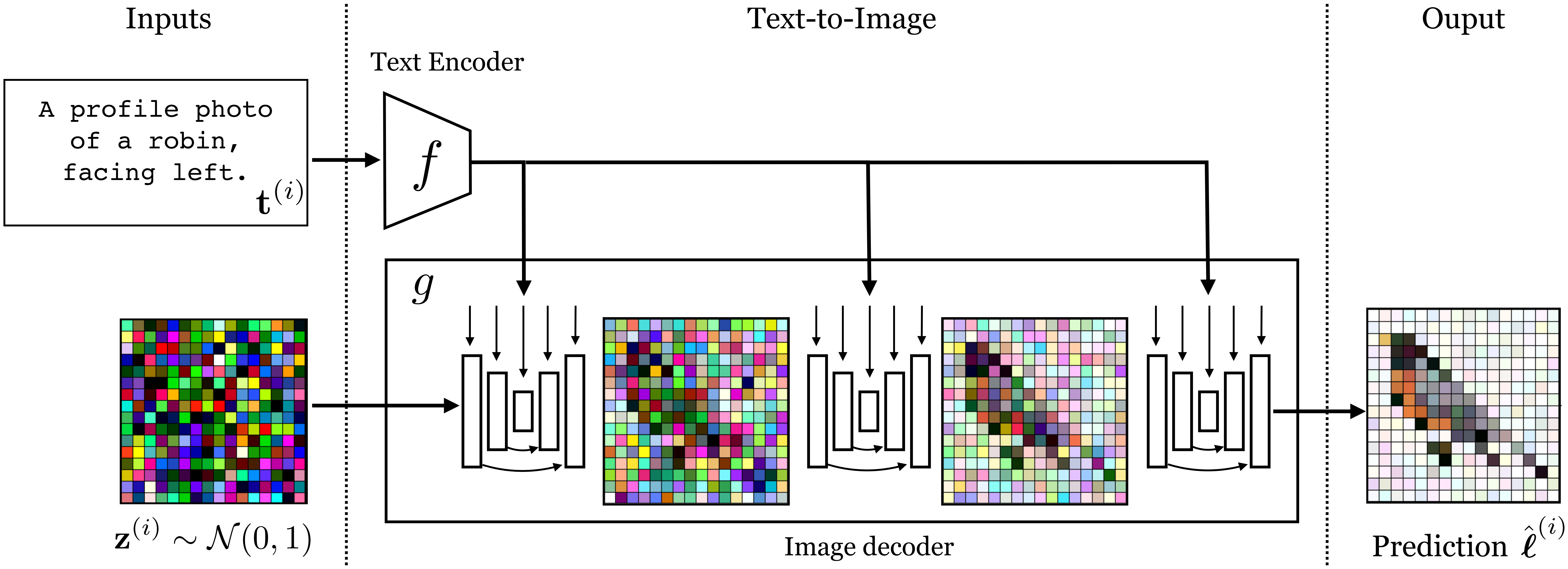
A robin on a
yellow background.



A robin on a
pink background.

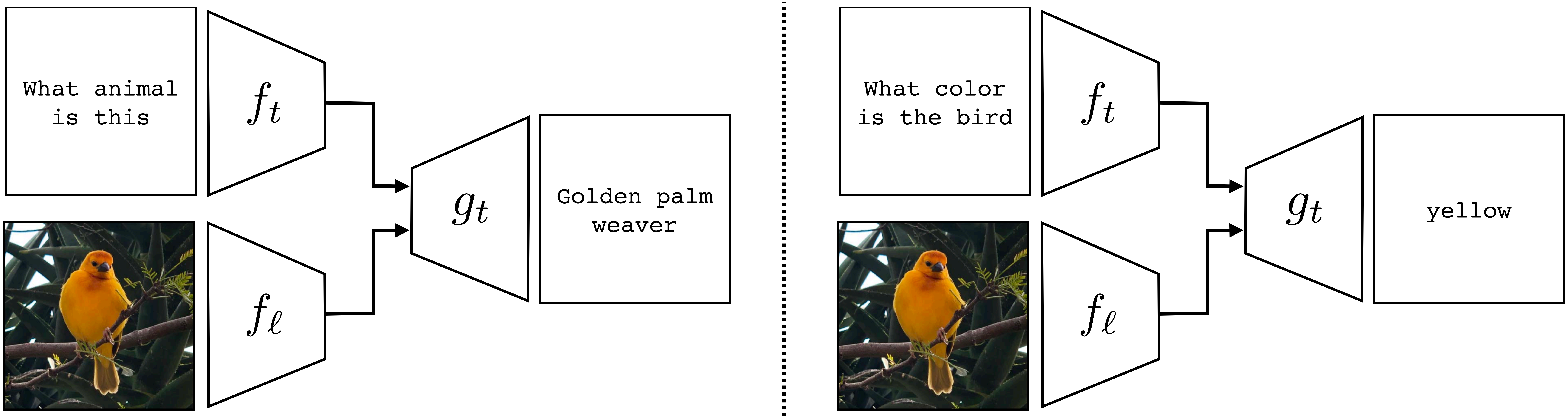


Text-to-image diffusion model


Inputs
Text-to-Image
Ouput
Text Encoder
A profile photo of a robin, facing left.
$\mathbf{t}^{(i)}$
f
g
$\mathbf{z}^{(i)} \sim \mathcal{N}(0,1)$
Image decoder
Prediction $\hat{\boldsymbol{\ell}}^{(i)}$

For example: DALL-E 2 [Ramesh et al. 2022], Stable Diffusion [Rombach*, Blattman* et al. 2022]

Visual Question Answering (VQA)



[Antol*, Agrawal* et al., ICCV 2015]

LLaVA

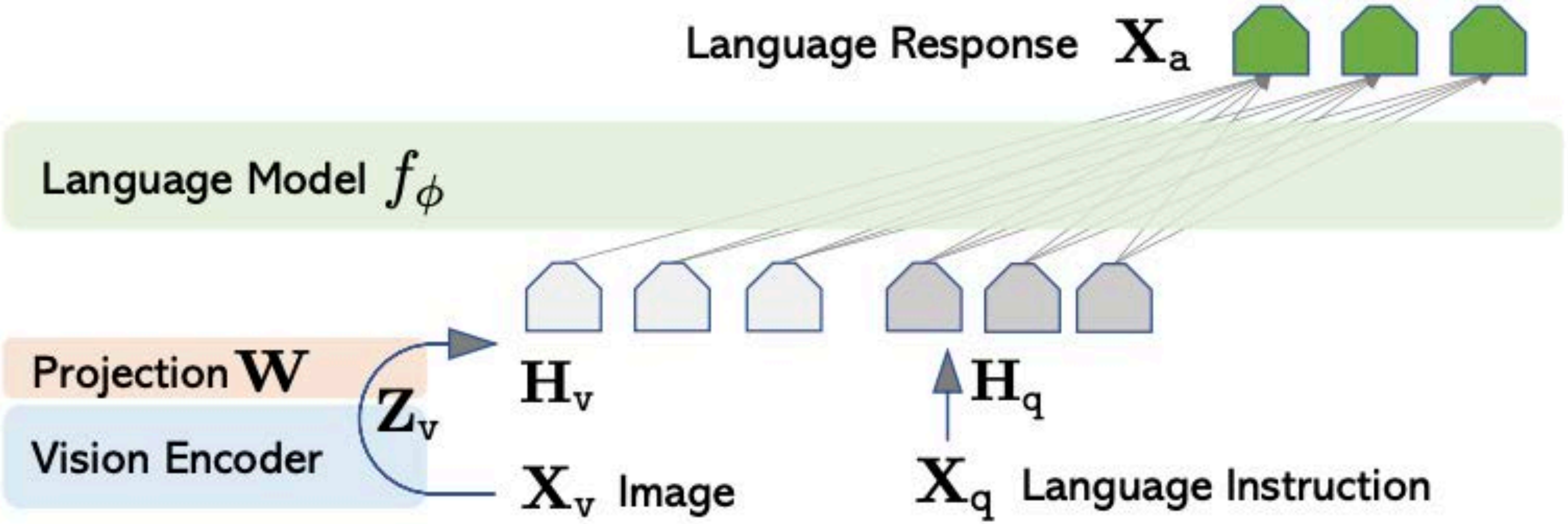
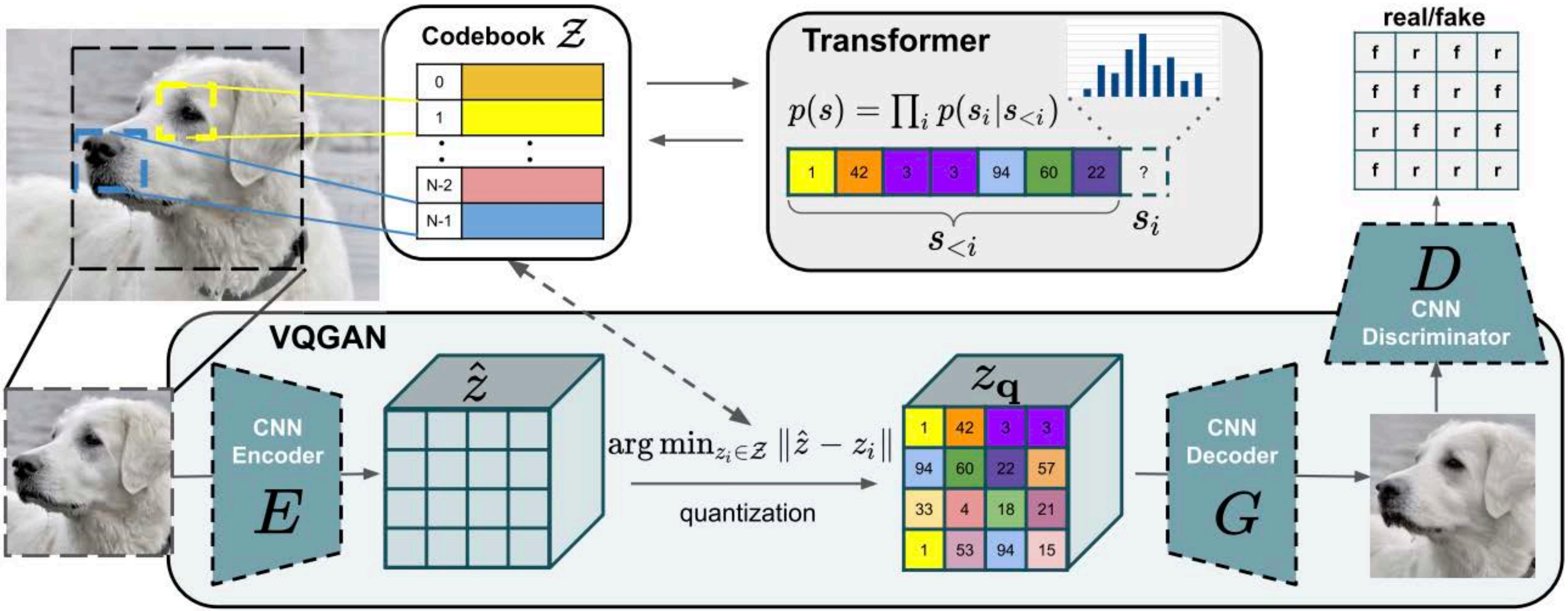


Figure 1: LLaVA network architecture.

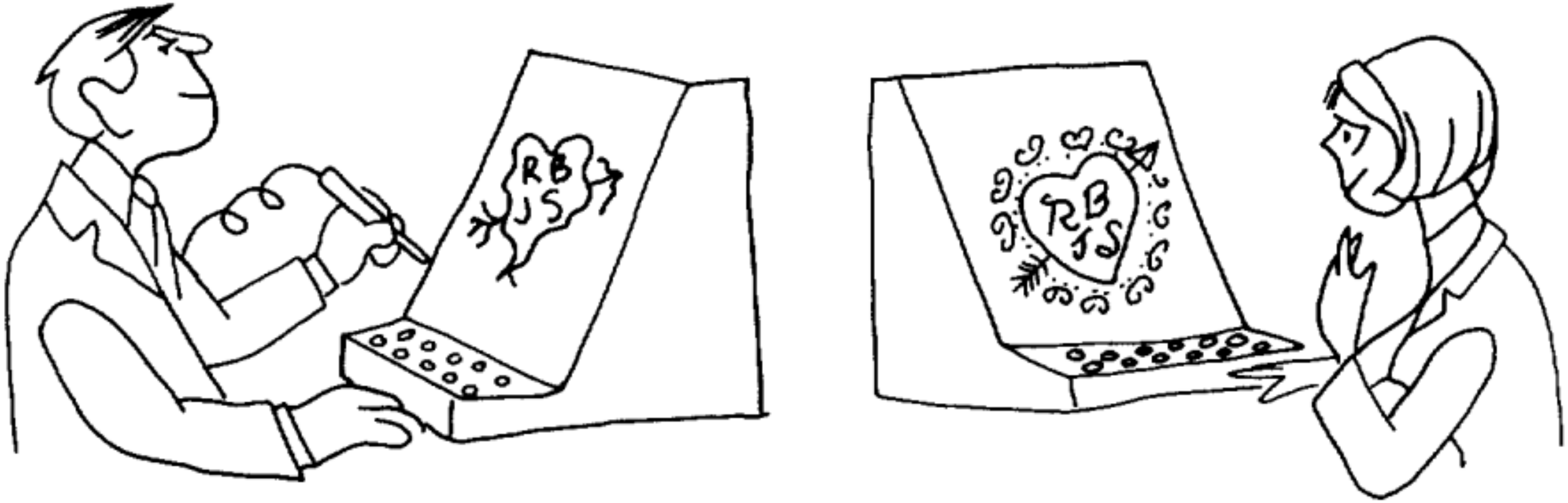
["Visual Instruction Tuning", Liu*, Li*, et al. 2023]

Mixing ideas from GANs, VAEs, and Autoregressive models



© Esser, et al. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

https://arxiv.org/pdf/2012.09841.pdf [“**Taming Transformers**”, Esser, Rombach, Ommer 2020]



Unpaired Translation

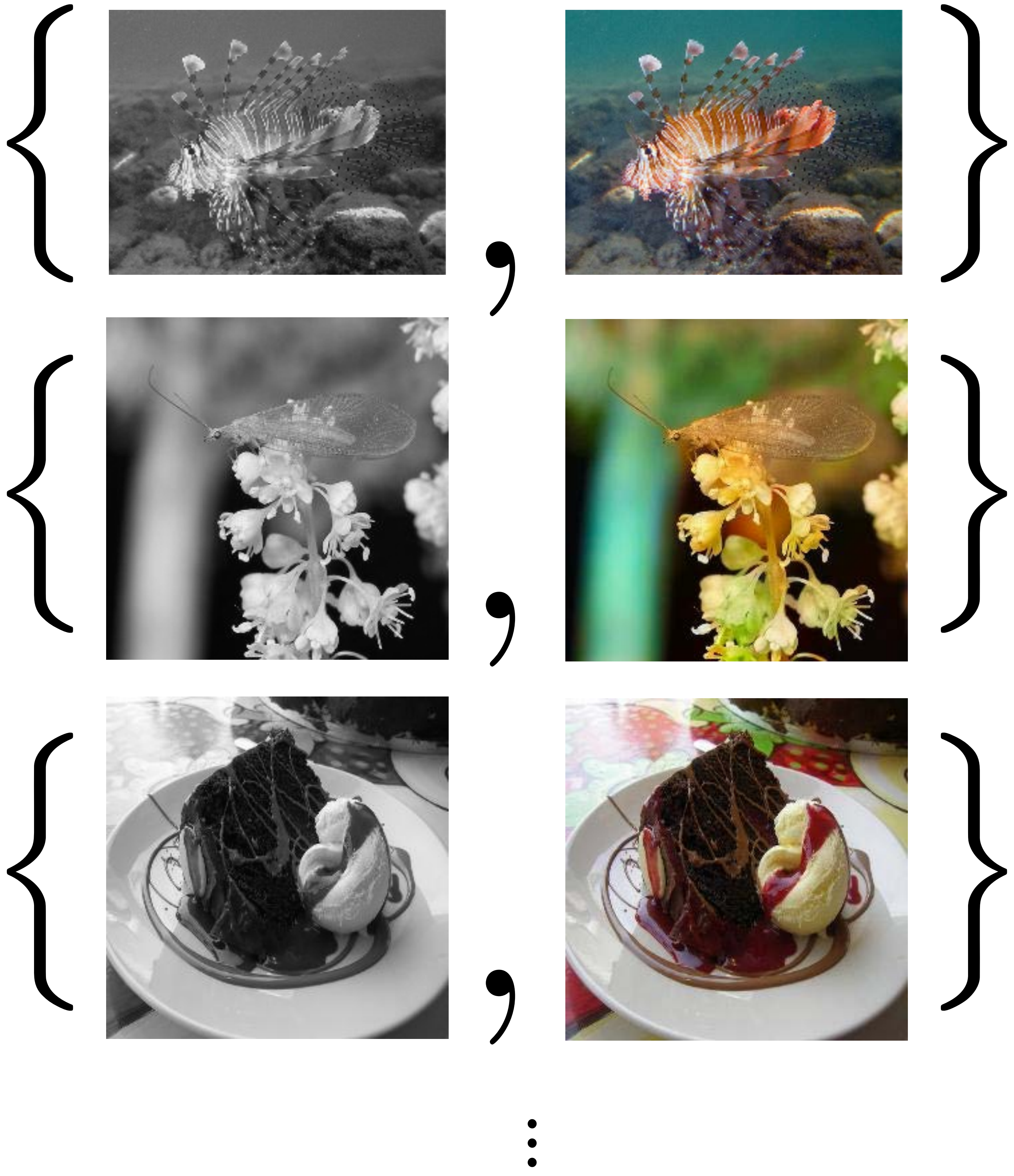
Cartoon © Science and Technology. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

[Includes slides from Jun-Yan Zhu, Taesung Park]

[Cartoon: The Computer as a Communication Device, Licklider & Taylor 1968]

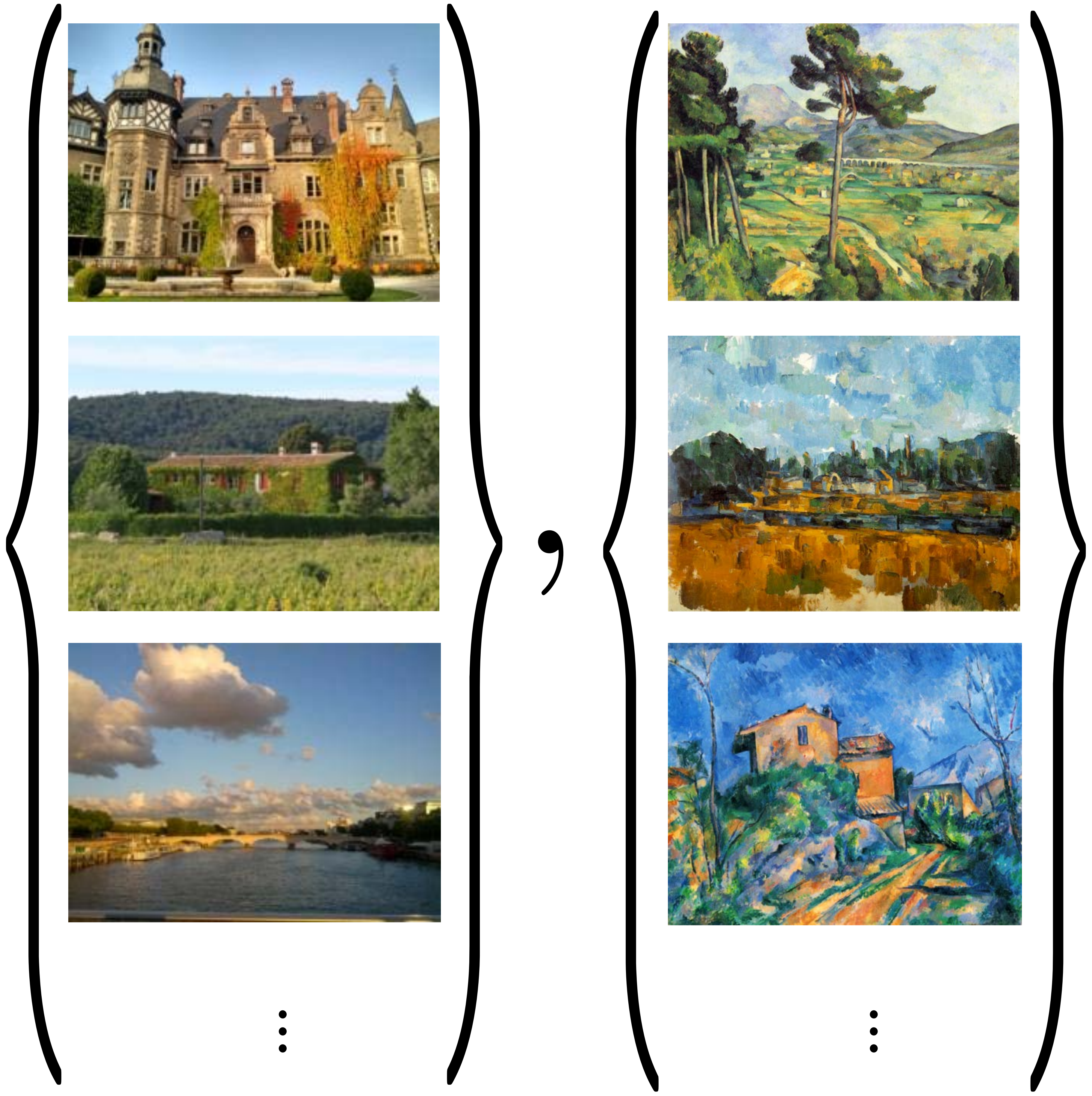
Paired

$\mathbf{x}^{(i)}$ $\mathbf{y}^{(i)}$

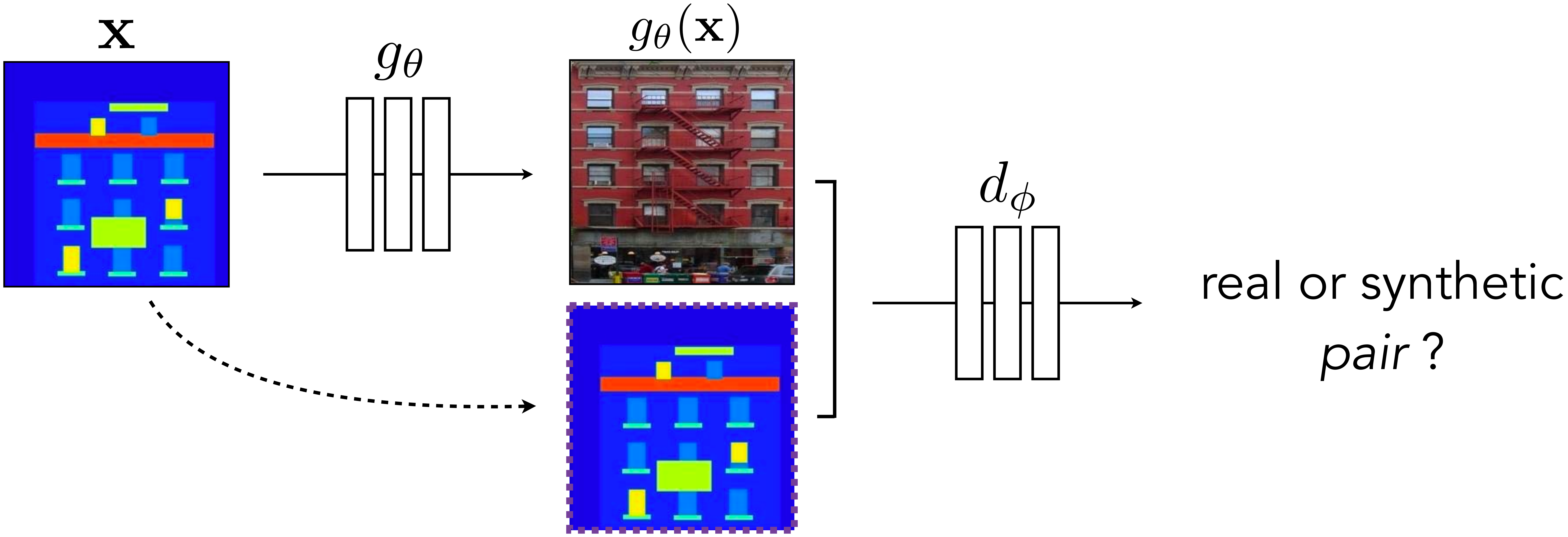


Unpaired

X Y



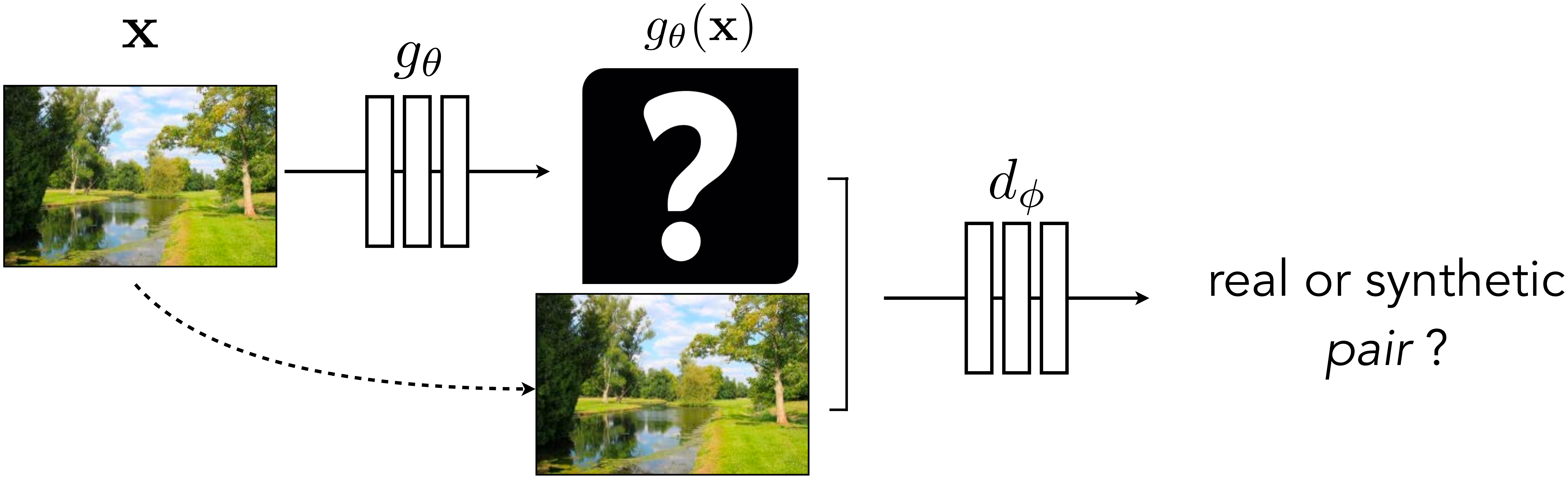
© Zhu, et al. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/



real or synthetic *pair* ?

$$\arg\min_\theta \max_\phi \mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(\mathbf{x}, g_\theta(\mathbf{x})) + \log(1 - d_\phi(\mathbf{x}, \mathbf{y}))]$$

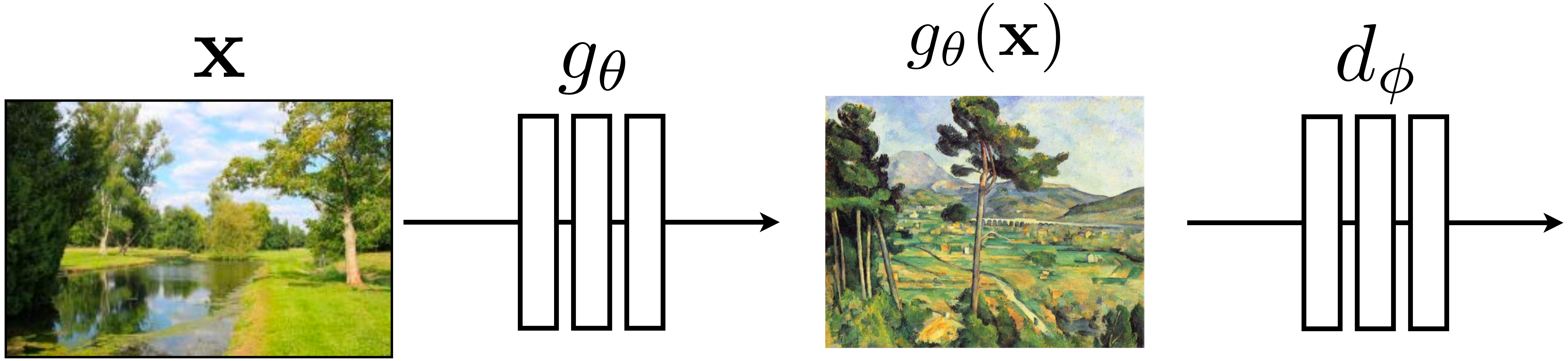
© Zhu, et al. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/



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

No input-output training pairs!

© Zhu, et al. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/



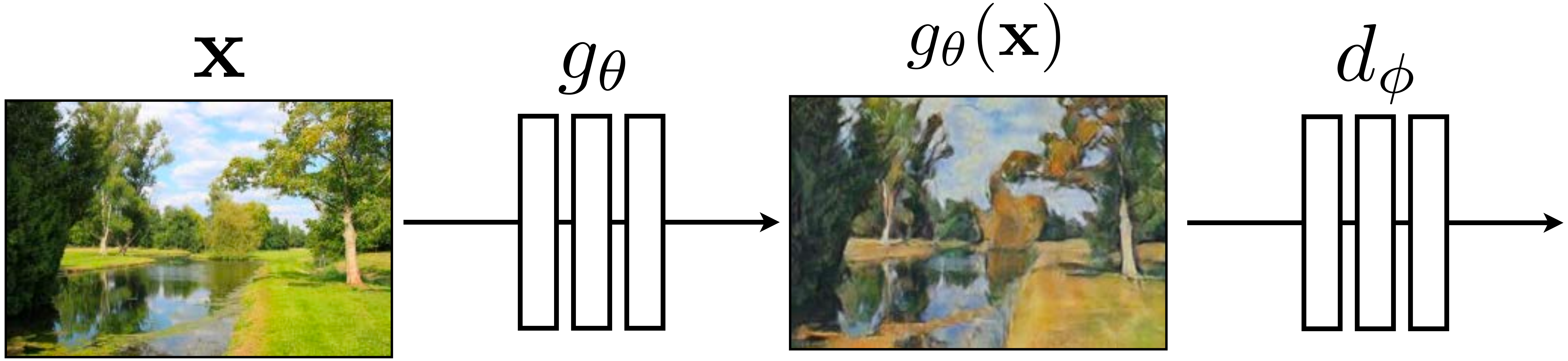
real painting
or synthetic?

$$\arg\min_{\theta}\max_{\phi}\mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(g_\theta(\mathbf{x})) + \log(1 - d_\phi(\mathbf{y}))]$$

Usually loss functions check if output matches a target *instance*

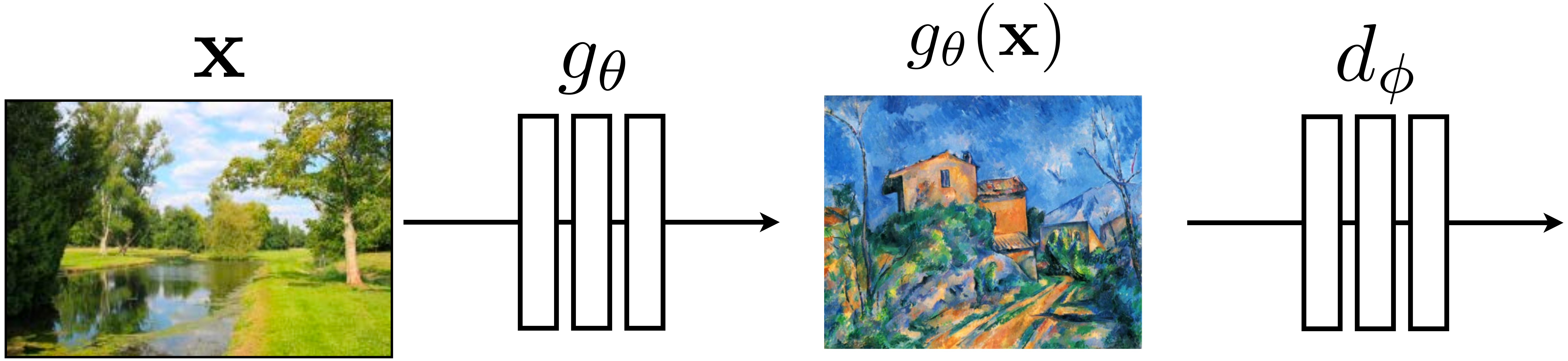
GAN loss checks if output is part of an admissible *set*

© Zhu, et al. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/



Real!

© Zhu, et al. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

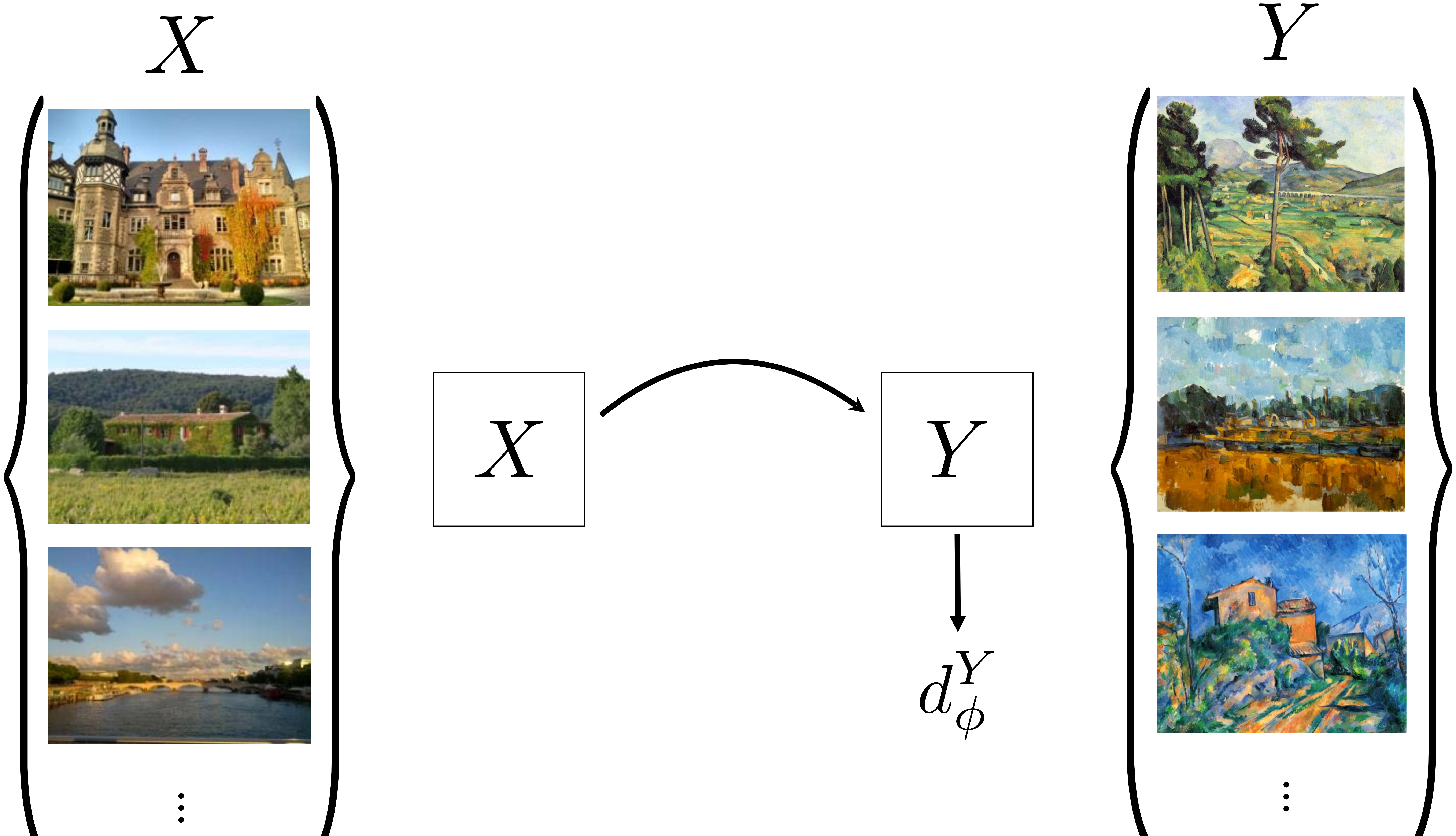


Real too!

Nothing to force output to correspond to input

© Zhu, et al. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

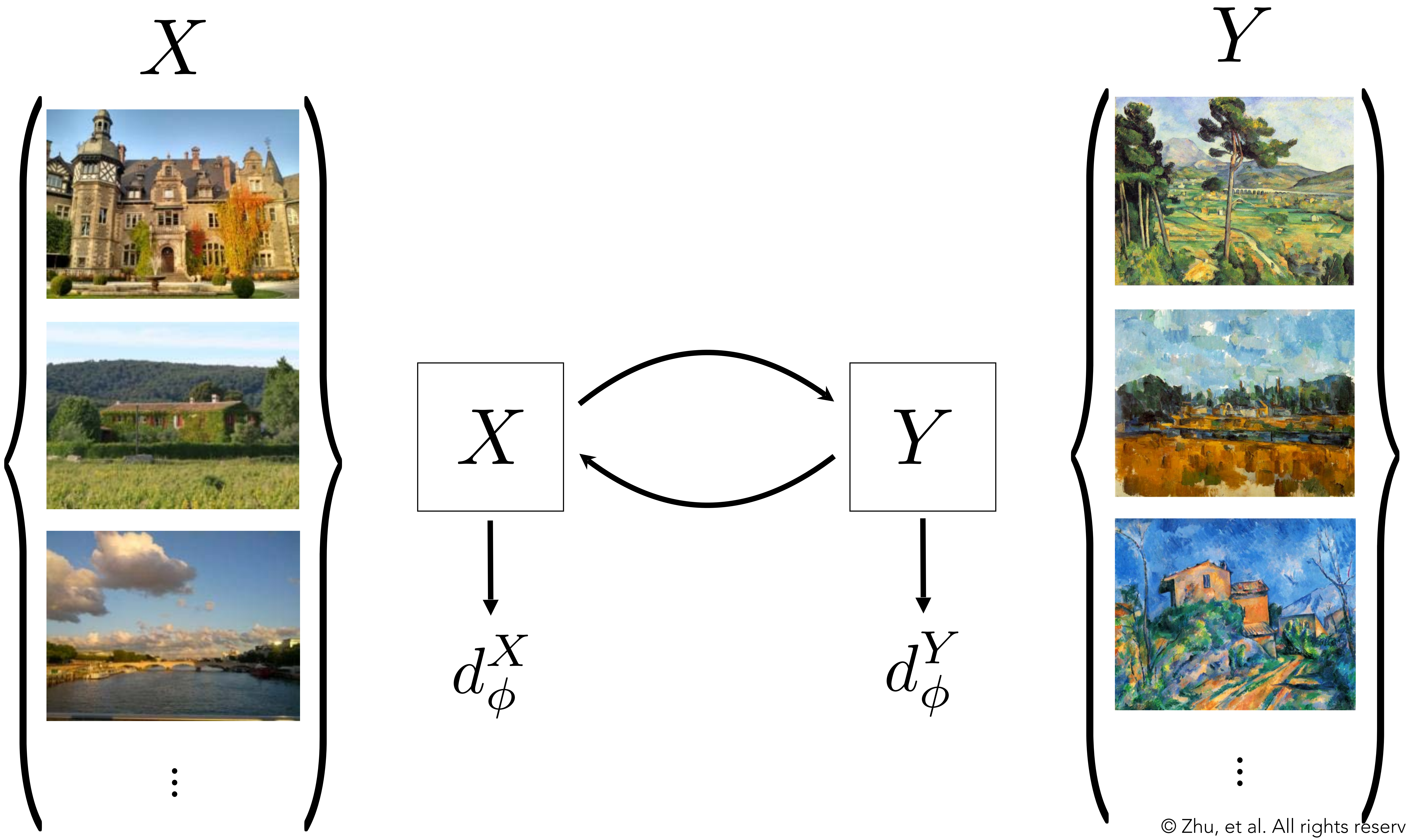
CycleGAN, or there and back aGAN



© Zhu, et al. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

[Zhu*, Park* et al. 2017], [Yi et al. 2017], [Kim et al. 2017]

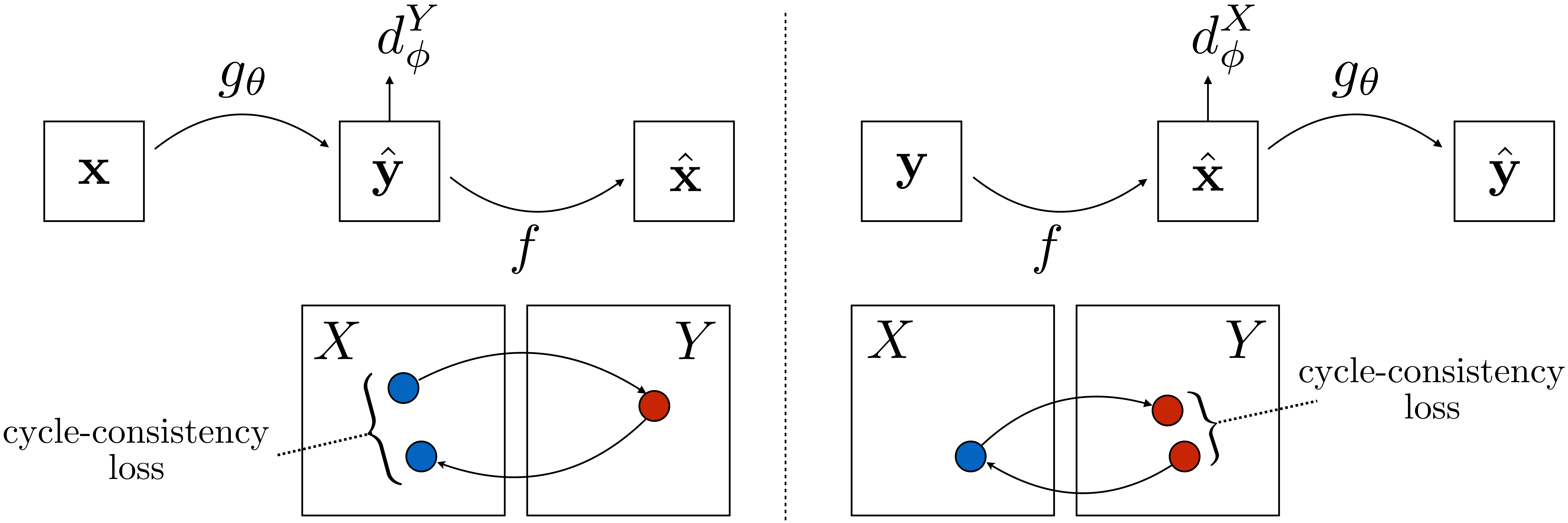
CycleGAN, or there and back aGAN



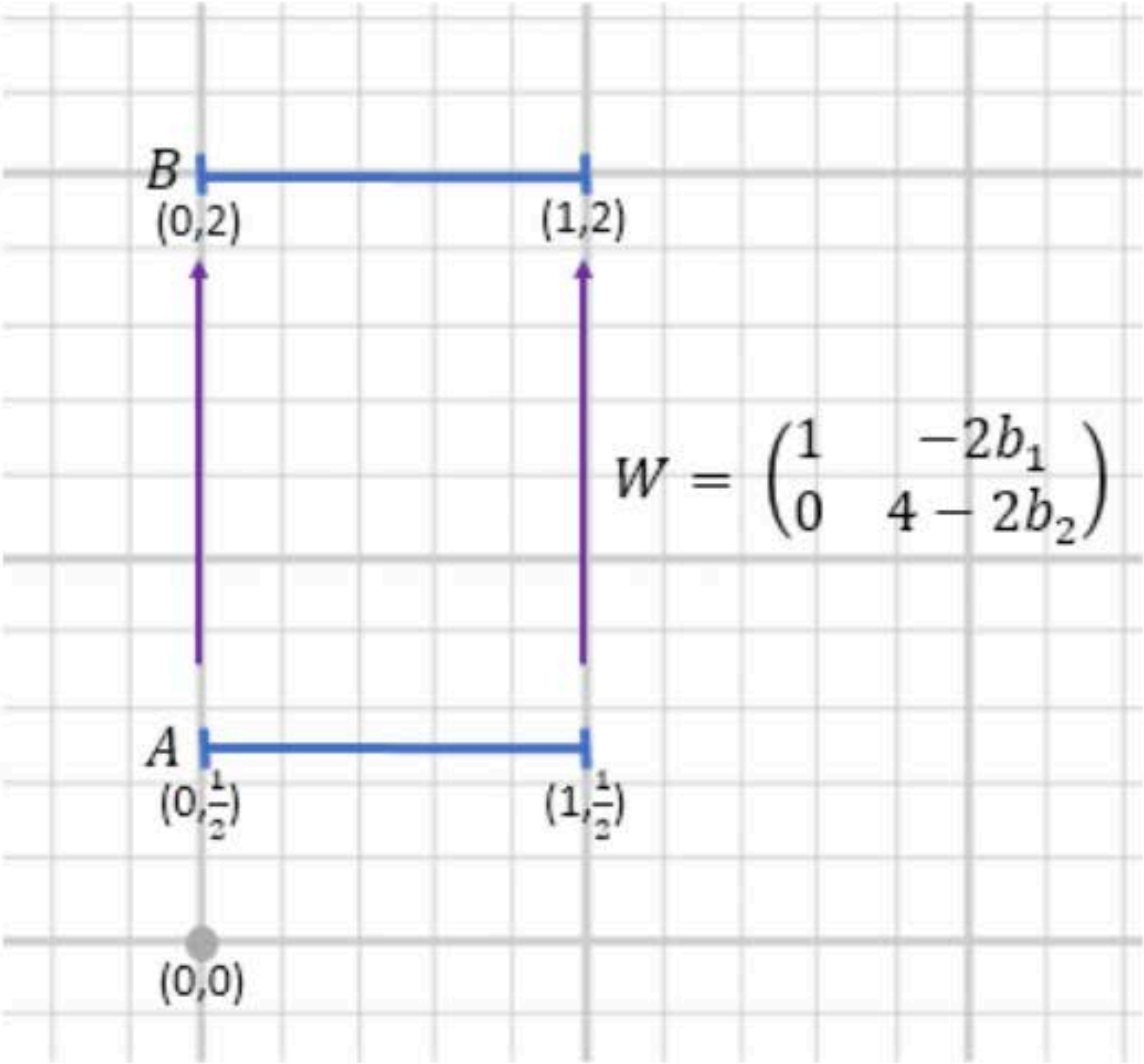
© Zhu, et al. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

[Zhu*, Park* et al. 2017], [Yi et al. 2017], [Kim et al. 2017]

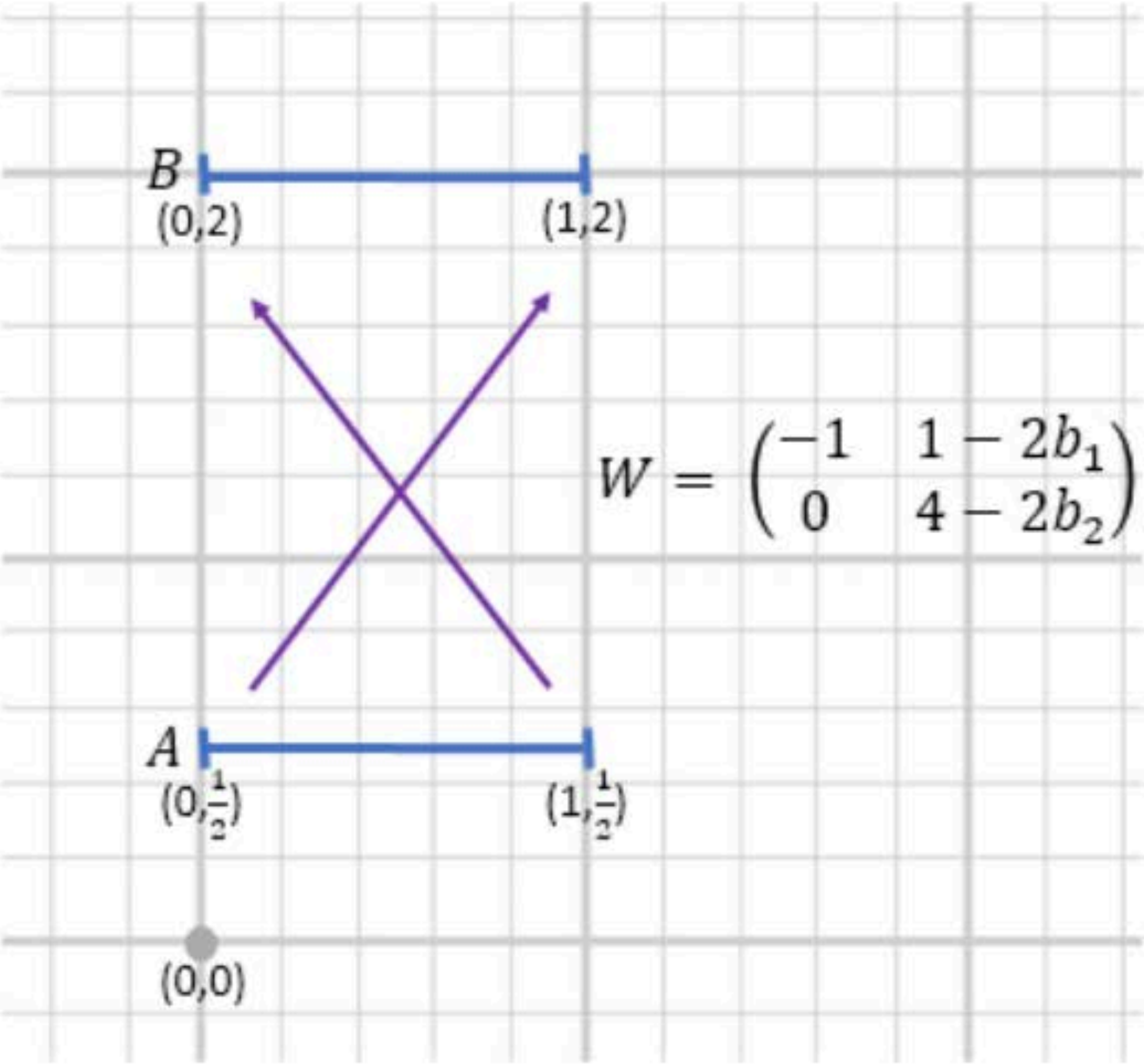
Cycle Consistency Loss



Unique solution?



$$W = \begin{pmatrix} 1 & -2b_1 \\ 0 & 4-2b_2 \end{pmatrix}$$

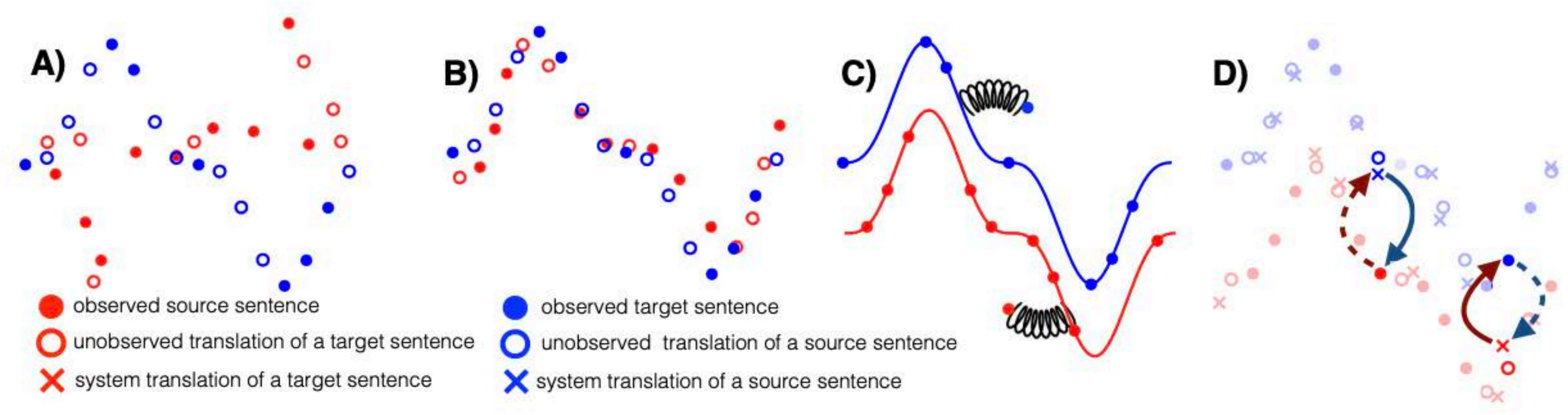


$$W = \begin{pmatrix} -1 & 1-2b_1 \\ 0 & 4-2b_2 \end{pmatrix}$$

© Galanti, Wolf, Benaim. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

[Galanti, Wolf, Benaim, ICLR 2018]

Unpaired translation in language


A)
B)
C)
D)
● observed source sentence
○ unobserved translation of a target sentence
× system translation of a target sentence
● observed target sentence
○ unobserved translation of a source sentence
× system translation of a source sentence

© Lample, et al. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

["Phrase-Based & Neural Unsupervised Machine Translation", Lample et al. 2018]

Concept: **paired data is not always necessary for grounding** (finding a unique mapping between two domains)

MIT OpenCourseWare
https://ocw.mit.edu

6.7960 Deep Learning
Fall 2024

For information about citing these materials or our Terms of Use, visit: https://ocw.mit.edu/terms