Generative Models; Vision & Language

Gül Varol

IMAGINE team, École des Ponts ParisTech

gul.varol@enpc.fr

http://imagine.enpc.fr/~varolg/

@RecVis, 21.11.2023

With many slides from: L. Lazebnik, F.-F. Li, J. Johnson, S. Yeung, F. Fleuret and others

Advanced topics in vision

Previously

1) **[J. Ponce]** Camera geometry, image processing

2) **[G. Varol]** Instance-level recognition

3) **[A. Joulin]** Supervised learning; Introduction to deep learning

4) **[G. Varol]** Neural networks for visual recognition

5) **[G. Varol]** Object detection, Segmentation, Human pose

6) **[J. Sivic]** Efficient visual search

Today

7) **[G. Varol]** Generative models; Vision & language

Next weeks

8) **[I. Laptev]** Weakly-, self-supervised learning; Robotics (Nov 28)

9) **[C. Schmid]** Videos (Dec 5)

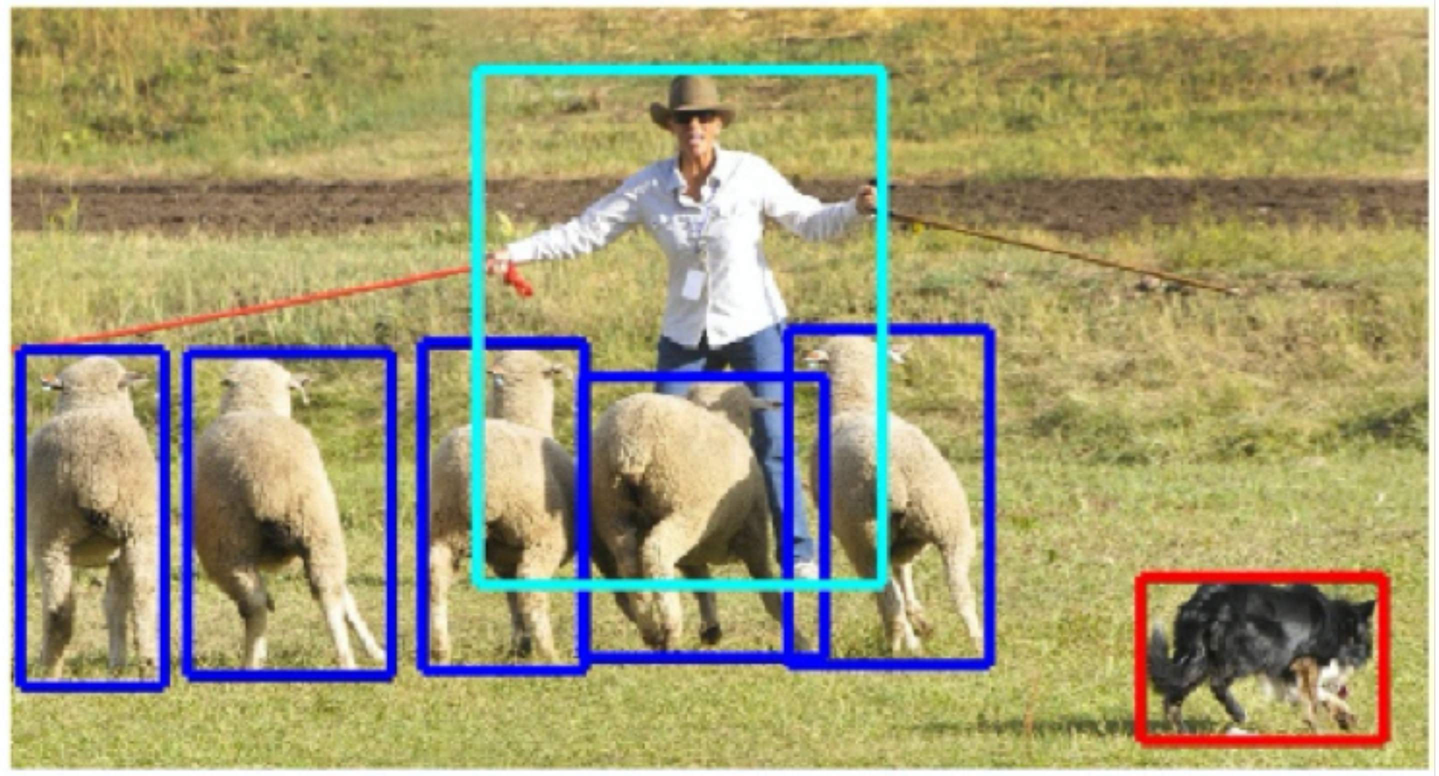
10) **[M. Aubry]** 3D (Dec 12)

Recap: Visual recognition so far

Image Classification

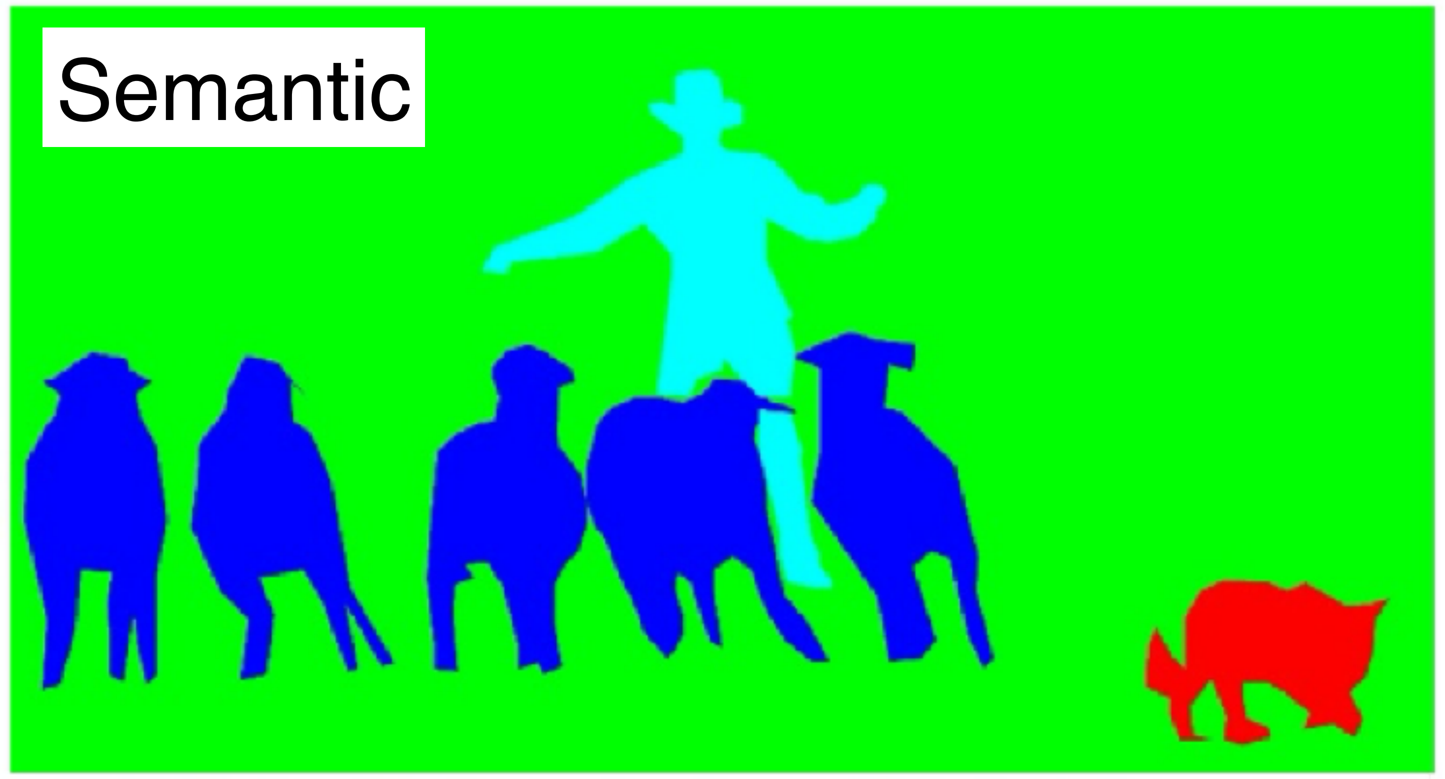
‣ Class labels

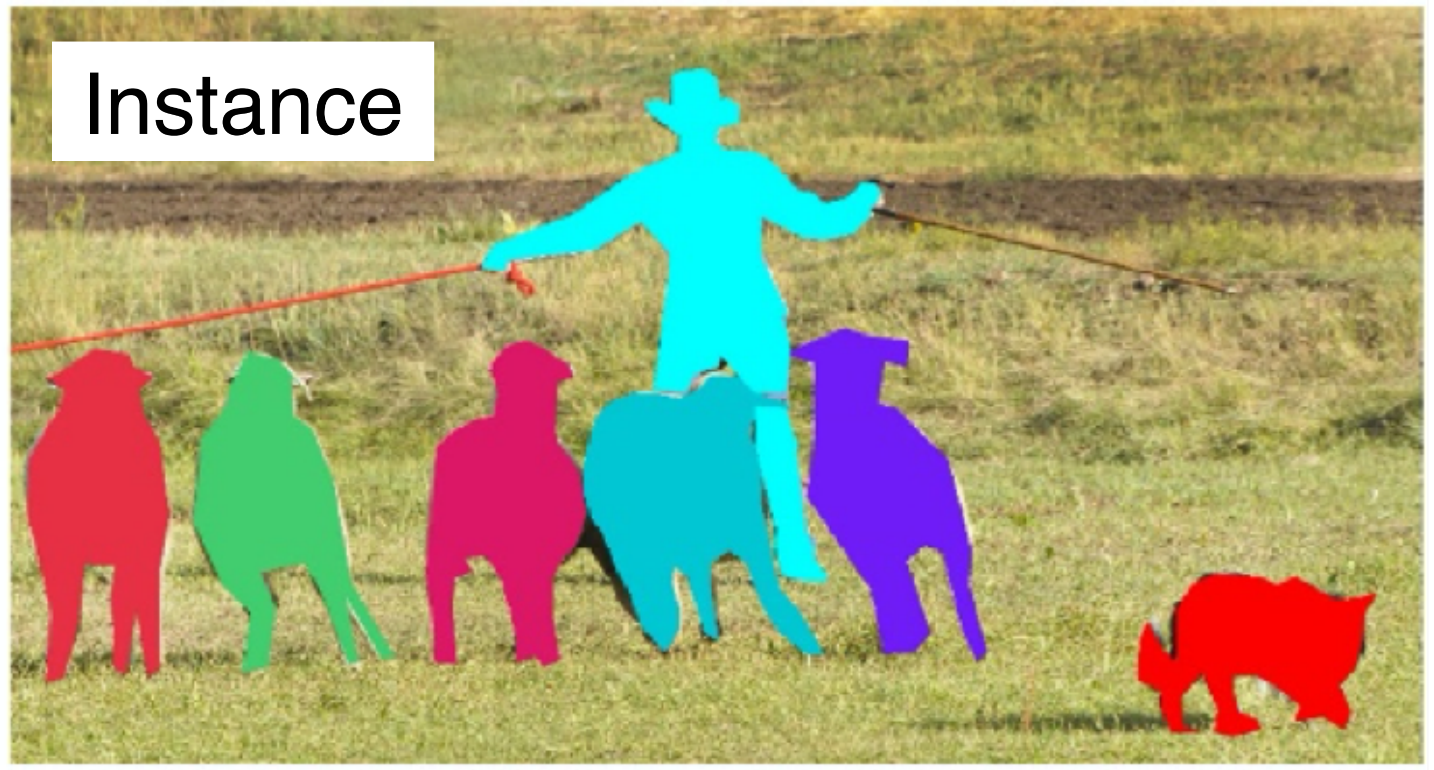
Object Detection



‣ Bounding box

Segmentation

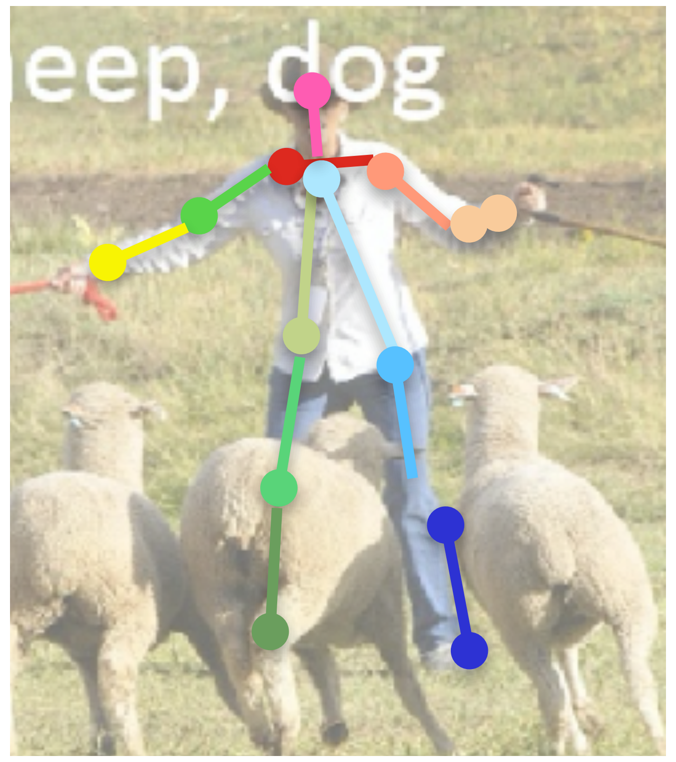




Panoptic,
Promptable …

‣ Pixel-wise labels

Human Pose



‣ Structured output

- **Objects in images**
- **Symbolic object categories**

Lin *et al.* “Microsoft COCO: Common Objects in Context”

Today

Analysis

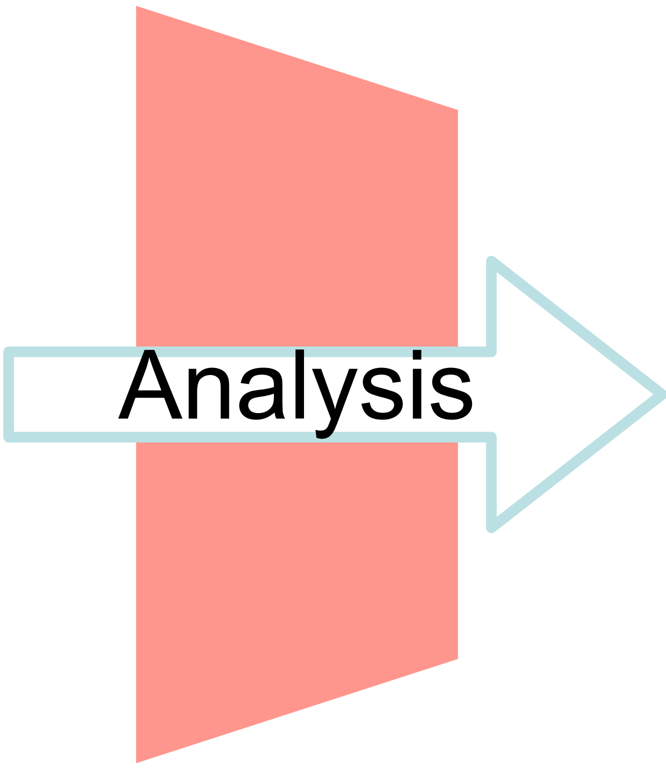
- Class labels
- Bounding box
- Pixel-wise labels
- Structured output

<u>Part 1: Generative models</u>

- Image

Synthesis

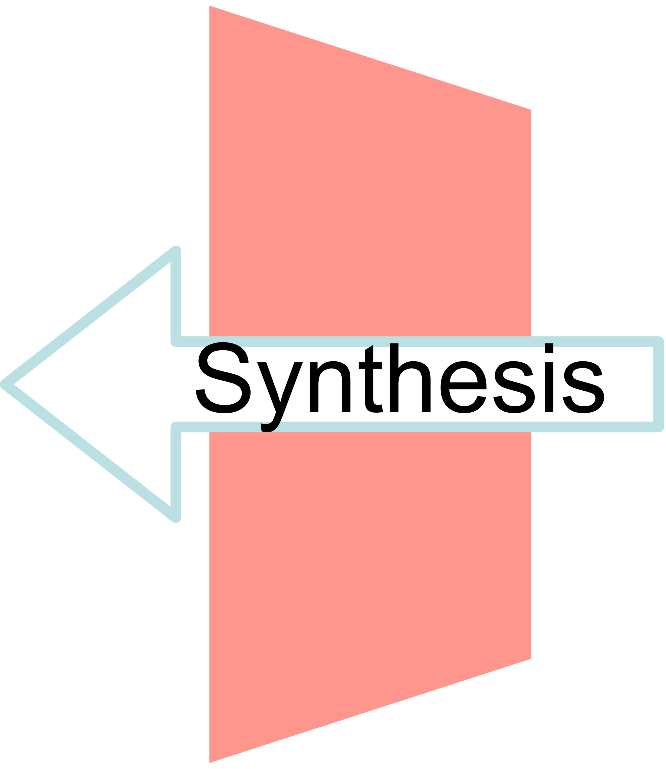
Today



- Class labels
- Bounding box
- Pixel-wise labels
- Structured output

- **Symbolic (object) categories**

Part 1: Generative models



- Image

Part 2: Vision & Language

- **Free-form text (language)**

Agenda

1. Generative neural networks

- VAE: Variational autoencoders
- GAN: Generative adversarial networks
- Diffusion models

2. Vision & language

- Text-to-image retrieval
- Text-to-image generation
- Image captioning

Agenda

1. Generative neural networks

- VAE: Variational autoencoders
- GAN: Generative adversarial networks
- Diffusion models

2. Vision & language

- Text-to-image retrieval
- Text-to-image generation
- Image captioning

Generative models

- Disclaimer: “Generative” is an overloaded term.
  - In this lecture’s context, we are concerned with generating (synthesizing) images with neural networks.

- A lot of buzz around a new term “Generative AI” (same thing)
  = models capable of generating media,
  typically text or images
  based on input text/prompt.

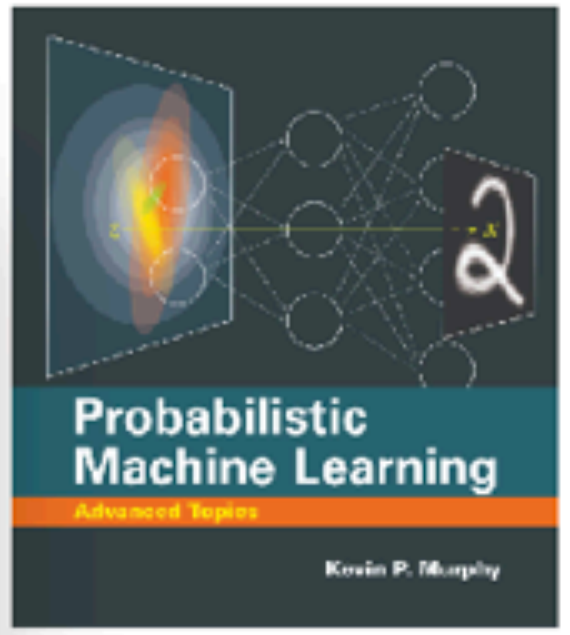
Further reading:

https://probml.github.io/pml-book/book2.html

probml.github.io/pml-book/book2.html

Probabilistic Machine Learning: Advanced Topics

by Kevin Patrick Murphy.
MIT Press, 2023.



Key links

- Short table of contents
- Long table of contents
- Preface
- Draft pdf of the main book, 2023-08-15. CC-BY-NC-ND license. (Please cite the official reference below.)
- Supplementary material
- Issue tracker.
- Code to reproduce most of the figures
- Acknowledgements
- Endorsements

If you use this book, please be sure to cite

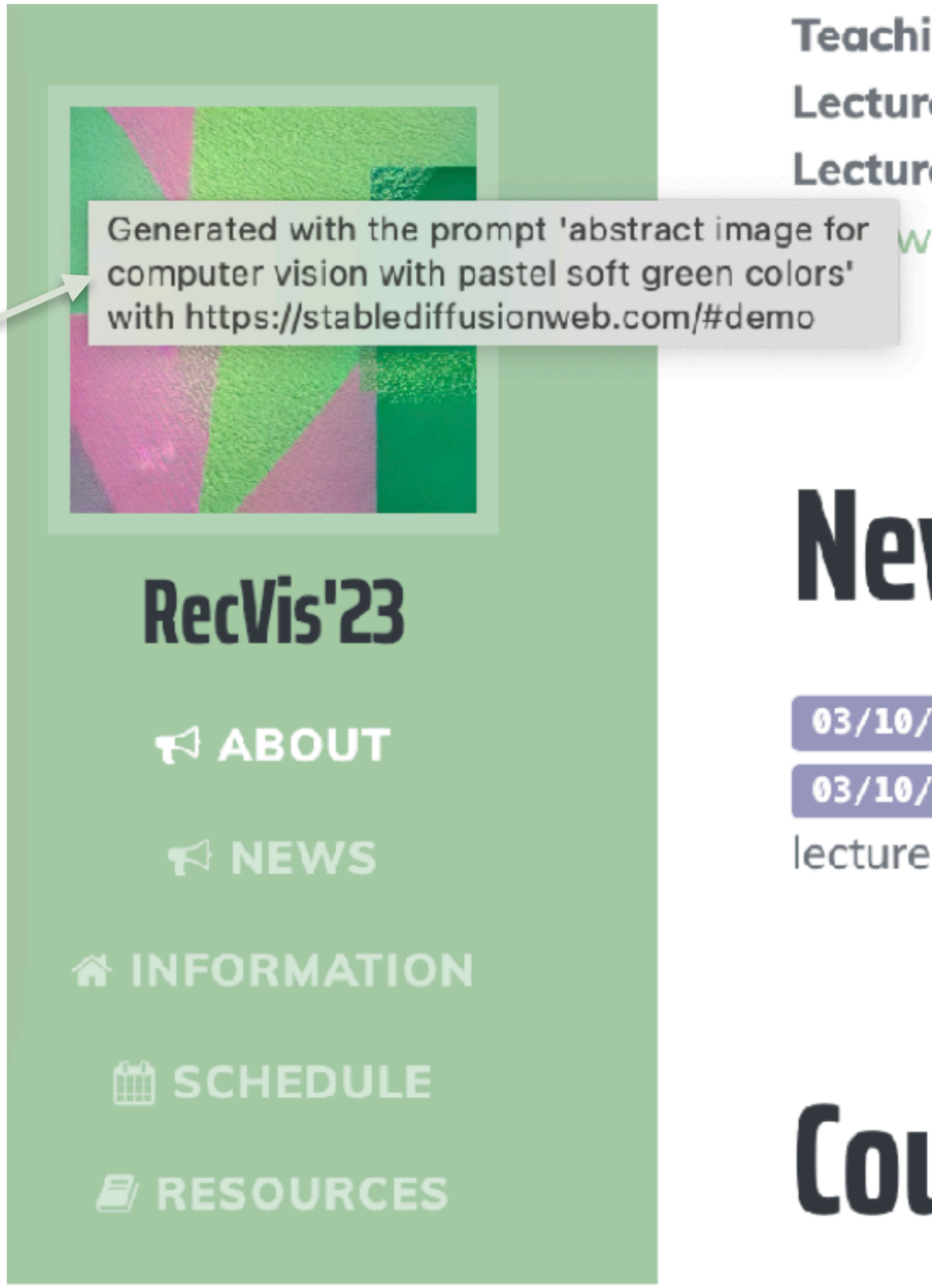
```
@book{pml2Book,
 author = "Kevin P. Murphy",
 title = "Probabilistic Machine Learning: Advanced Topics",
 publisher = "MIT Press",
 year = 2023,
 url = "http://probml.github.io/book2"
}
```

Downloads since 2022-02-28. downloads 158k

IV Generation 761
20 Generative models: an overview 763
21 Variational autoencoders 779
22 Autoregressive models 815
23 Normalizing flows 823
24 Energy-based models 843
25 Diffusion models 861
26 Generative adversarial networks 887

Why Generative Models?

- Creativity/arts, super-resolution,...
- Can create synthetic data for training
- Can provide useful feature representations
- Data compression
- ...



Generative (image synthesis) tasks

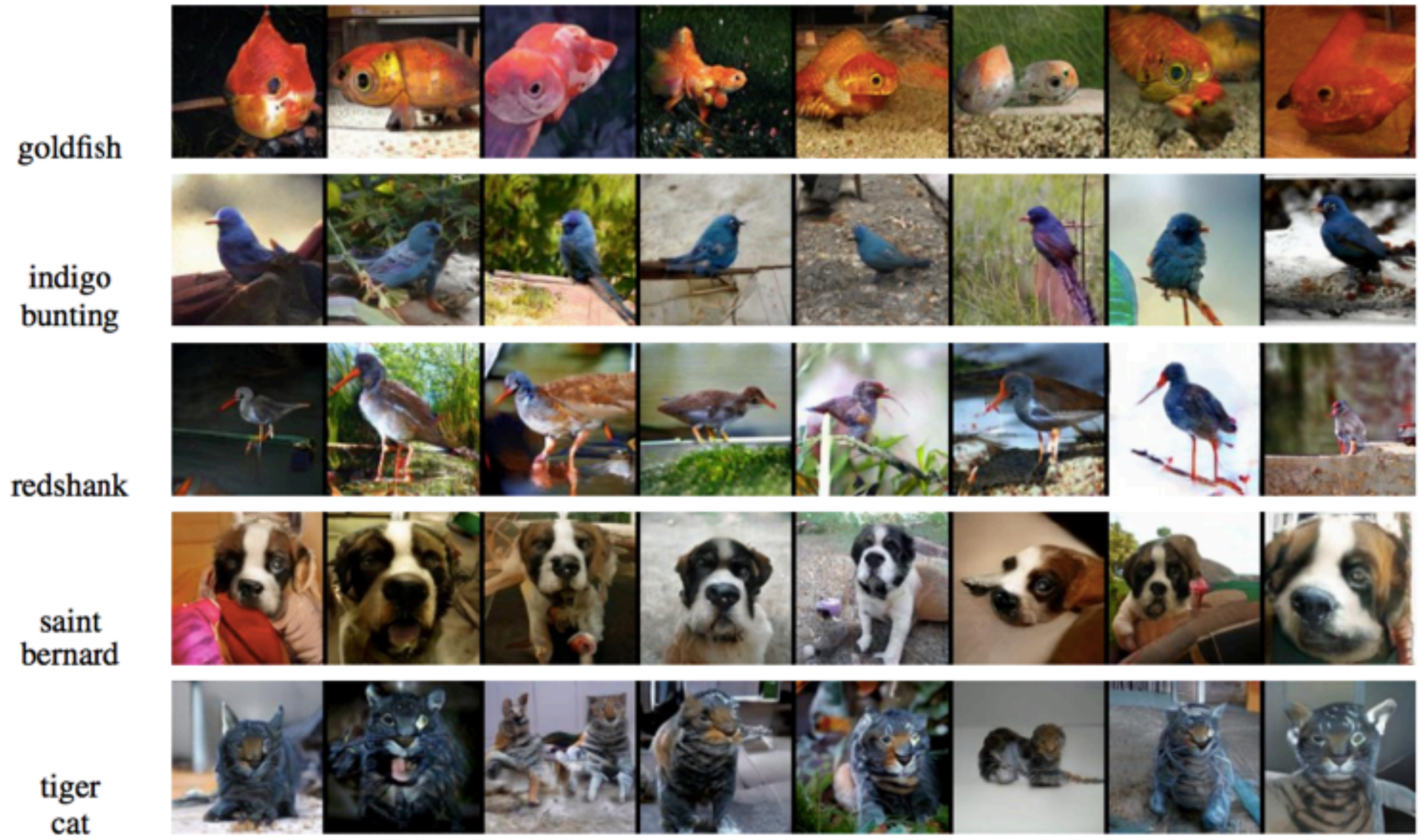
- Unconditional
- Conditioned on class label
- Conditioned on image
- Conditioned on text
- …

Generative tasks

- **Unconditional** generation: learn to sample from the distribution represented by the training set
  - *Unsupervised learning* task

Generative tasks

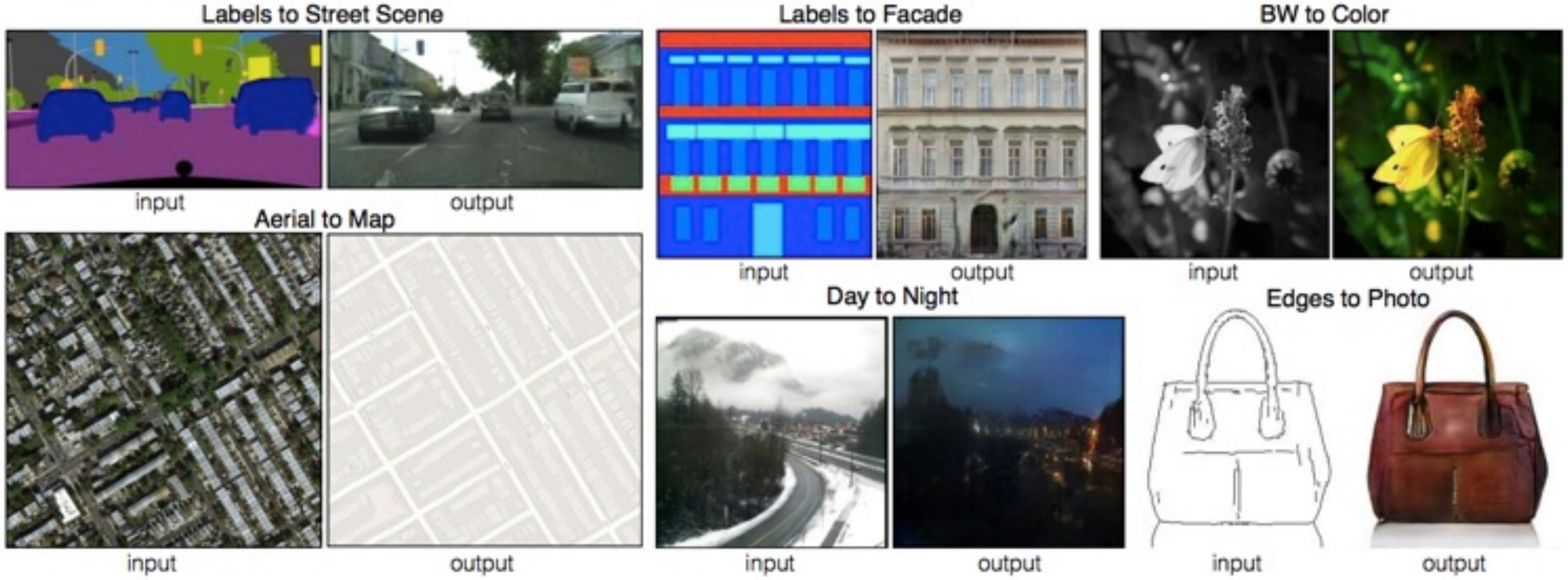
- Generation conditioned on **class label**



[Figure source](#)

Generative tasks

- Generation conditioned on **image**
  or *image-to-image translation*



P. Isola, J.-Y. Zhu, T. Zhou, A. Efros, [Image-to-Image Translation with Conditional Adversarial Networks](), CVPR 2017

Generative tasks

- Generation conditioned on **text**

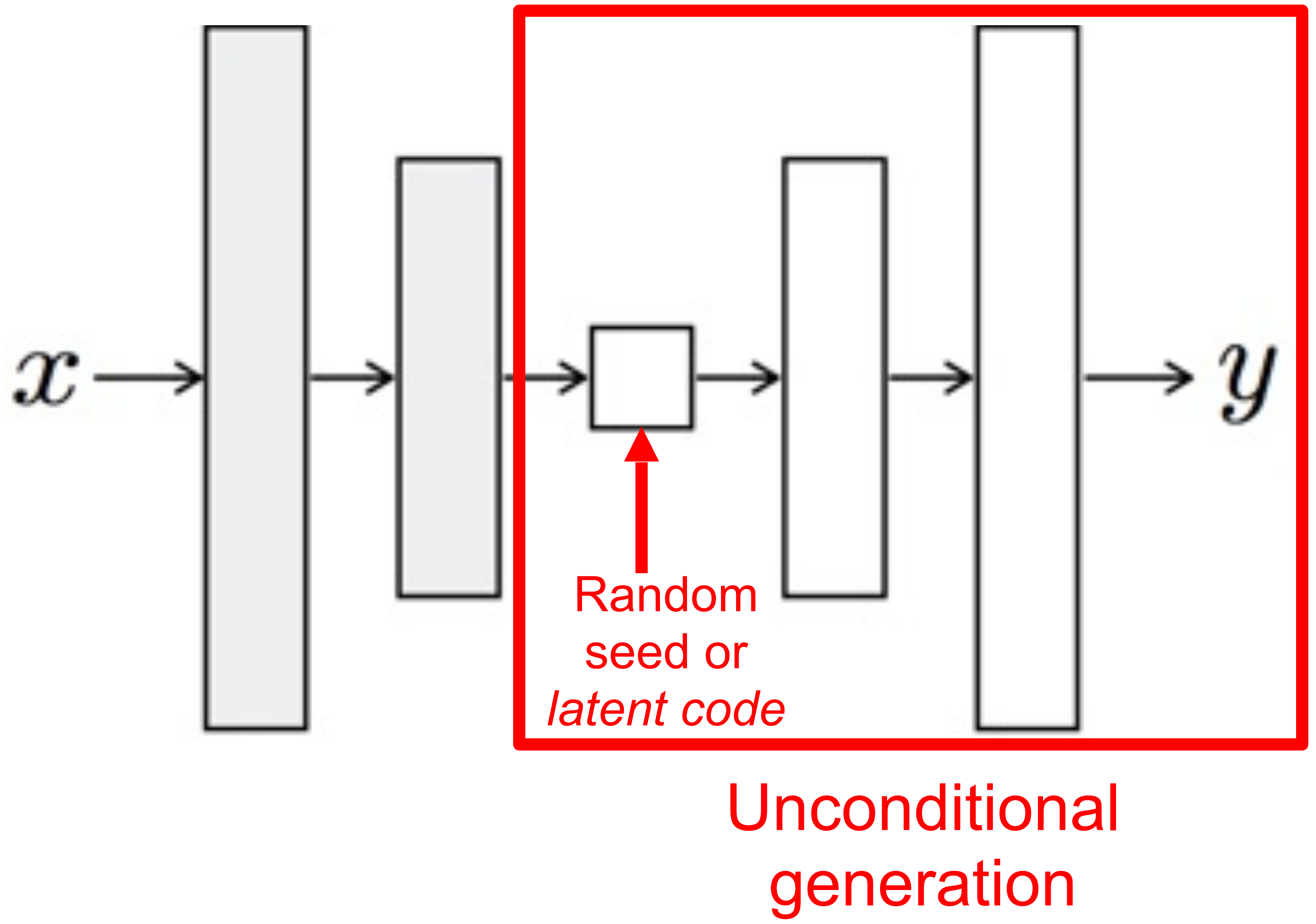
Vibrant portrait painting of Salvador Dali with a robotic half face

A close up of a handpalm with leaves growing from it

A. Ramesh et al. [Hierarchical text-conditional image generation with CLIP latents](). 2022

Designing a network for generative tasks

1. We need an architecture that can generate an image
   - Recall upsampling architectures for dense prediction



Designing a network for generative tasks

1. We need an architecture that can generate an image
   - Recall upsampling architectures for dense prediction

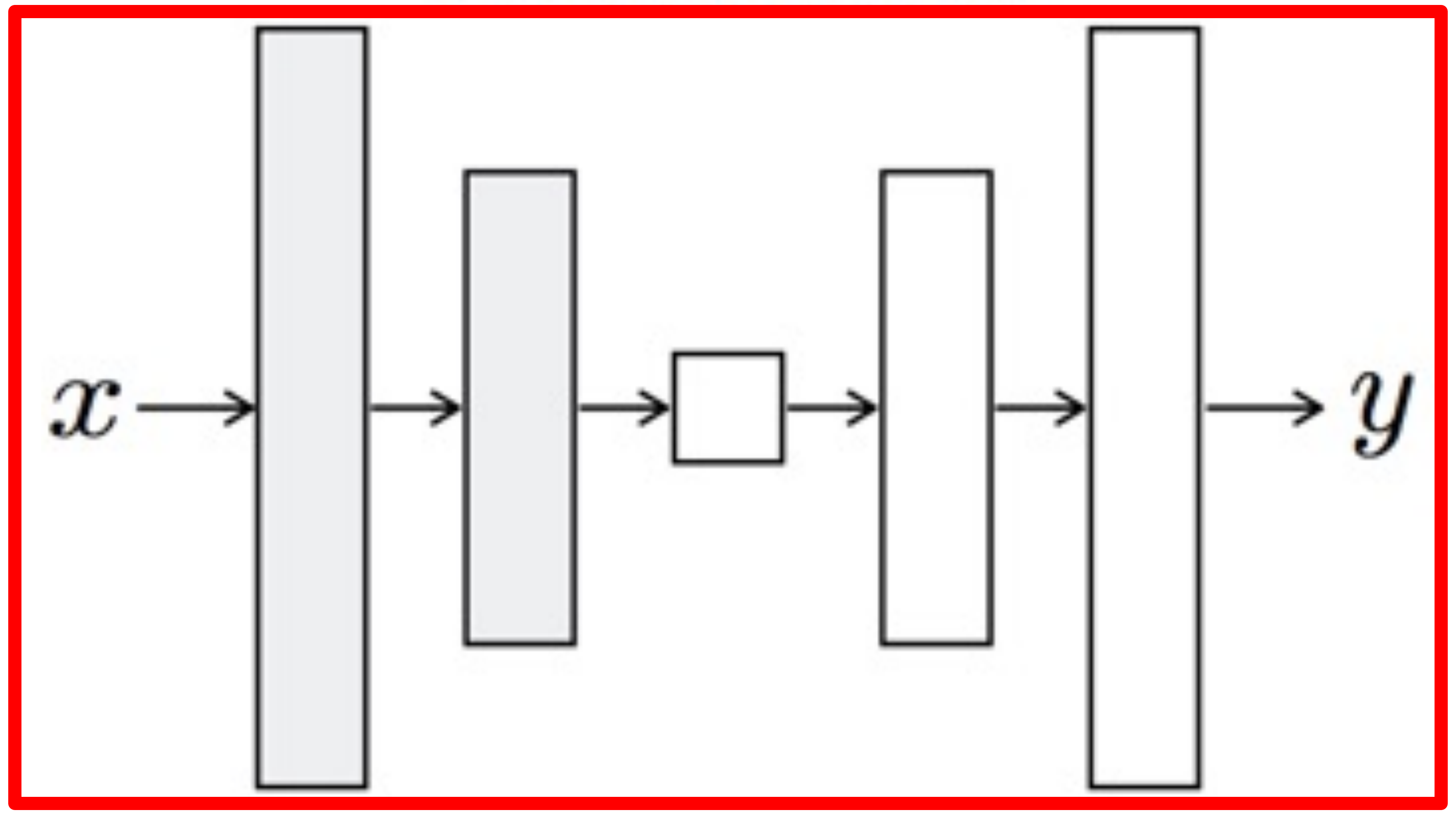


Image-to-image translation

Designing a network for generative tasks

1. We need an architecture that can generate an image
   - Recall upsampling architectures for dense prediction

2. We need to design the right loss function and training framework

Learning to sample

- Given training data, generate new samples from same distribution

Training data $x \sim p_{\text{data}}$

Generated samples $x \sim p_{\text{model}}$

We want to learn p_{model} that matches p_{data}

Adapted from [Stanford CS231n]

Agenda

1. Generative neural networks

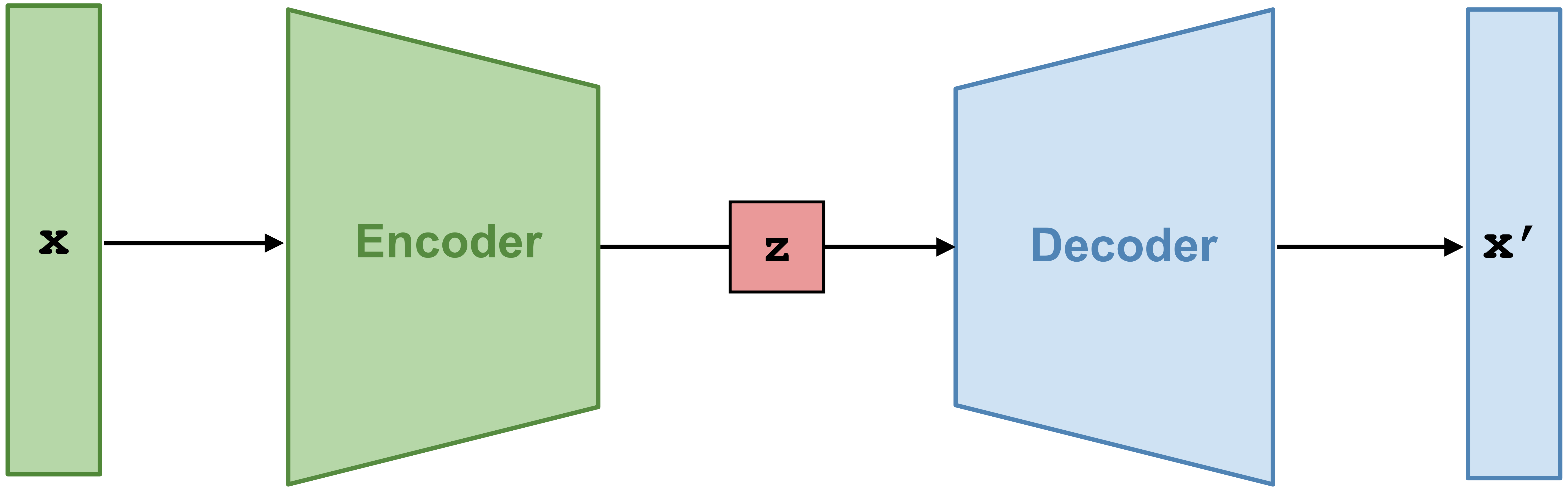
- VAE: Variational autoencoders
- GAN: Generative adversarial networks
- Diffusion models

2. Vision & language

- Text-to-image retrieval
- Text-to-image generation
- Image captioning

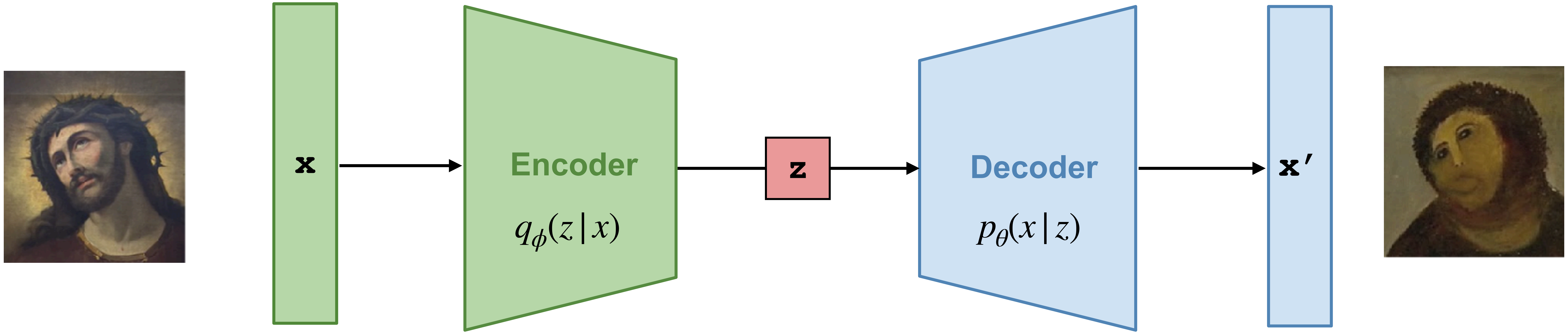
What is an autoencoder?

- Encoder + Decoder with a bottleneck $\mathtt{z}$
- Reconstruction loss ($\mathtt{x}'$, $\mathtt{x}$)



Variational Autoencoders (VAEs)

- Autoencoder with structured bottleneck
- At training time, jointly learn *encoder* and *decoder* by maximizing a variational bound on the data likelihood, with 2 loss terms: (1) KL divergence and (2) Reconstruction



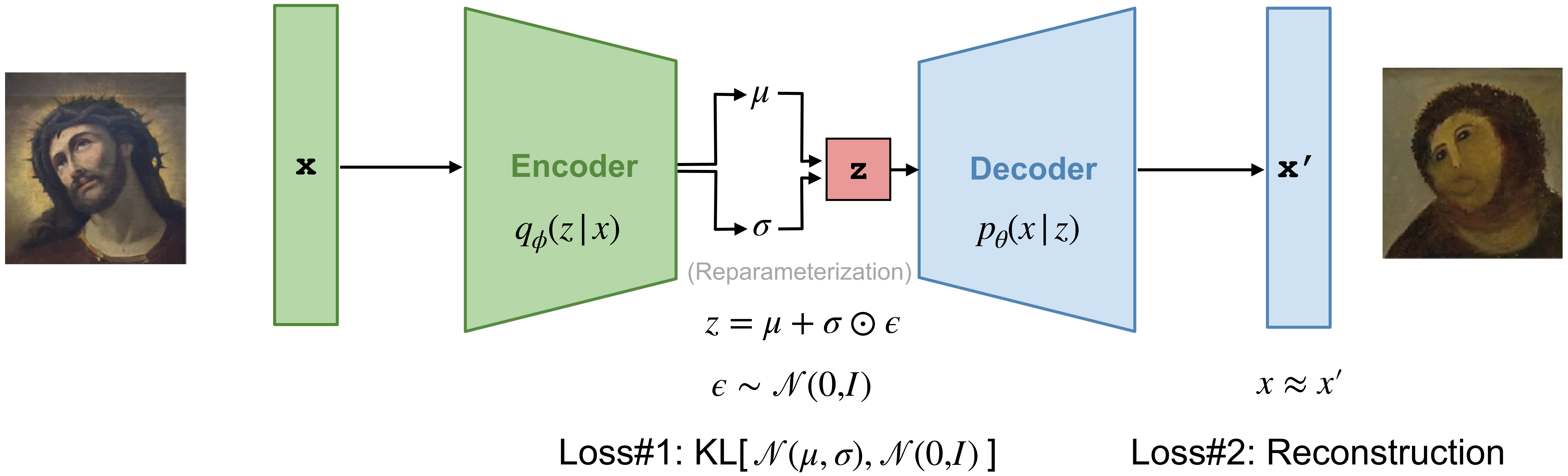
<u>Learns to:</u>
compress samples to a low-dimensional representation

<u>Learns to:</u>
generate samples

D. Kingma and M. Welling, [Auto-Encoding Variational Bayes](), ICLR 2014

Variational Autoencoders (VAEs)

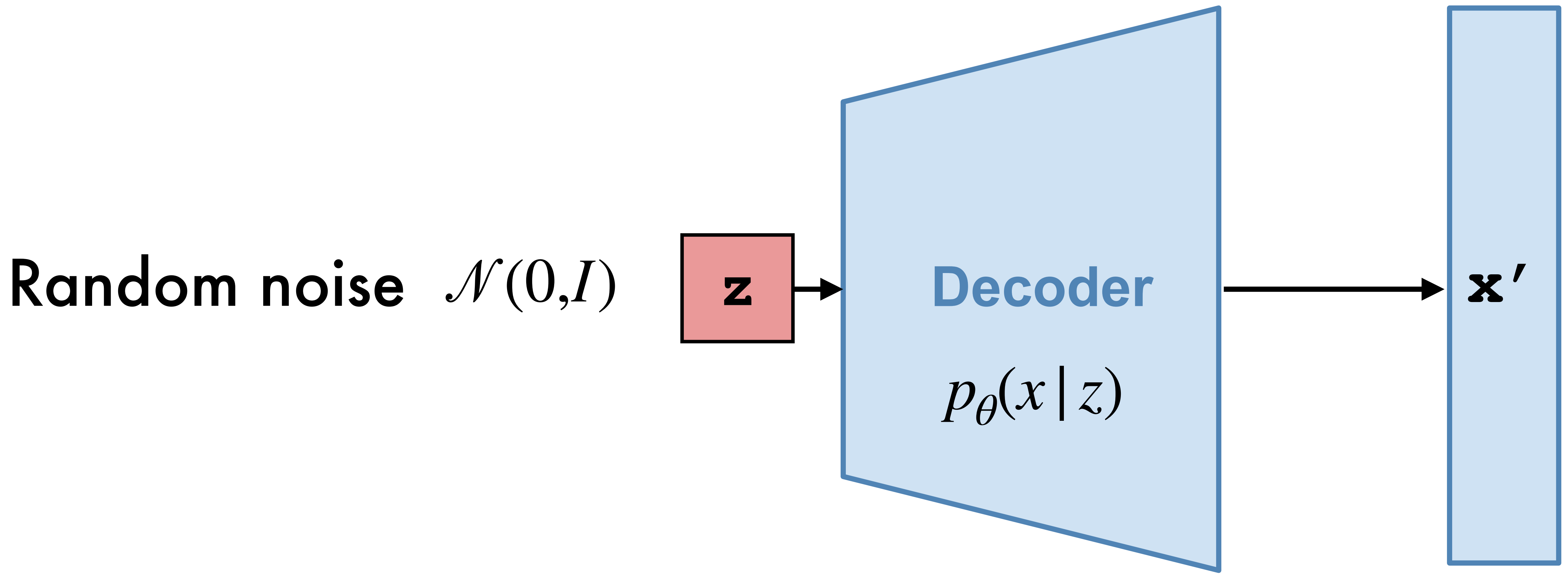
- Autoencoder with structured bottleneck
- At training time, jointly learn *encoder* and *decoder* by maximizing a variational bound on the data likelihood, with 2 loss terms: (1) KL divergence and (2) Reconstruction



D. Kingma and M. Welling, Auto-Encoding Variational Bayes, ICLR 2014

Variational Autoencoders (VAEs)

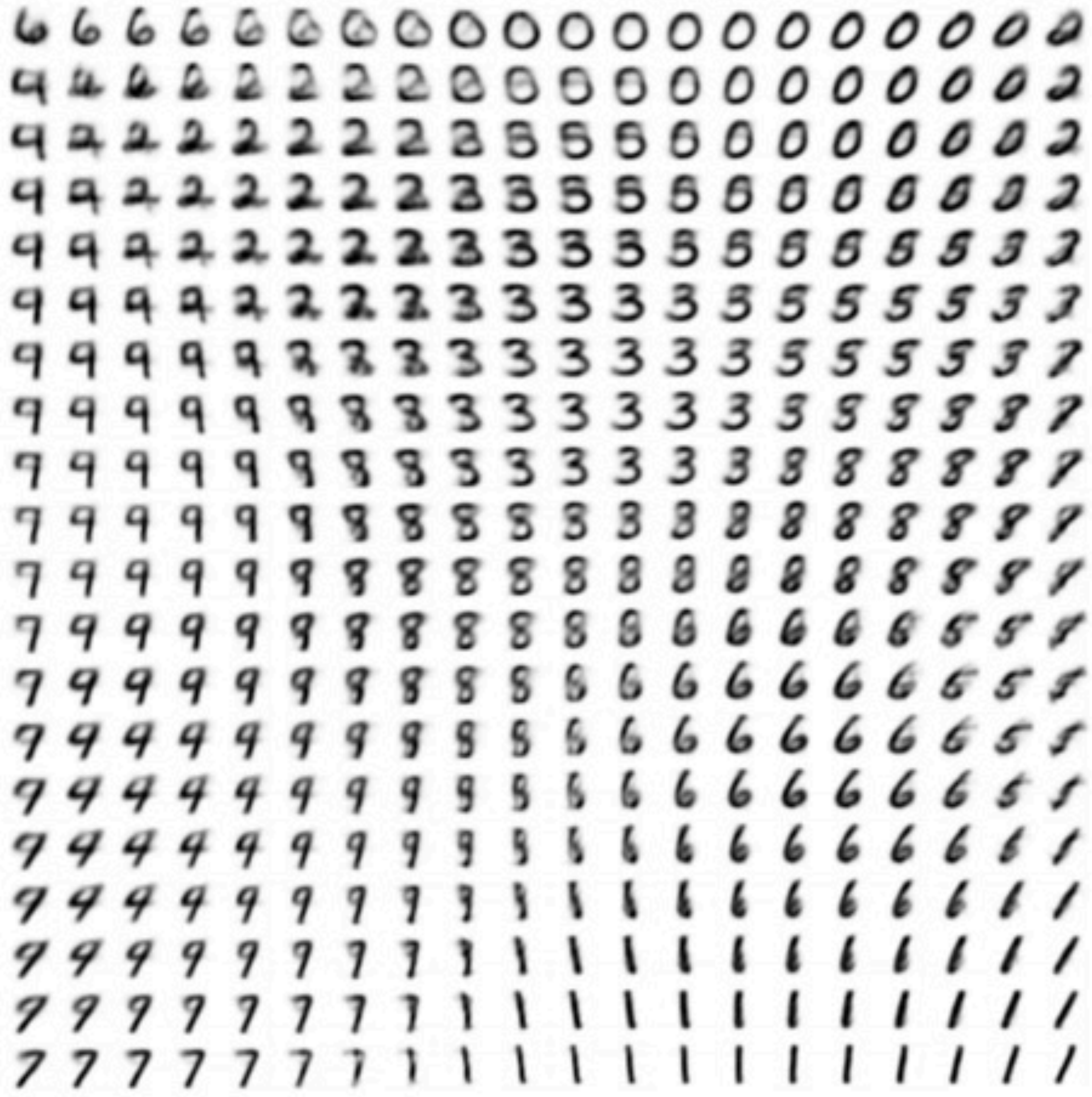
- Autoencoder with structured bottleneck
- At training time, jointly learn *encoder* and *decoder* by maximizing a variational bound on the data likelihood, with 2 loss terms: (1) KL divergence and (2) Reconstruction
- **At test time**, discard encoder and use *decoder* to sample from the learned distribution



D. Kingma and M. Welling, [Auto-Encoding Variational Bayes](), ICLR 2014

Original VAE results

- Learned 2D "manifolds":



D. Kingma and M. Welling, [Auto-Encoding Variational Bayes](), ICLR 2014

Original VAE results

Image source

VAE pros and cons

- Pros:
  - Principled mathematical formalism for generative models
  - Allows inference of code given image, can be useful for controlling the latent space

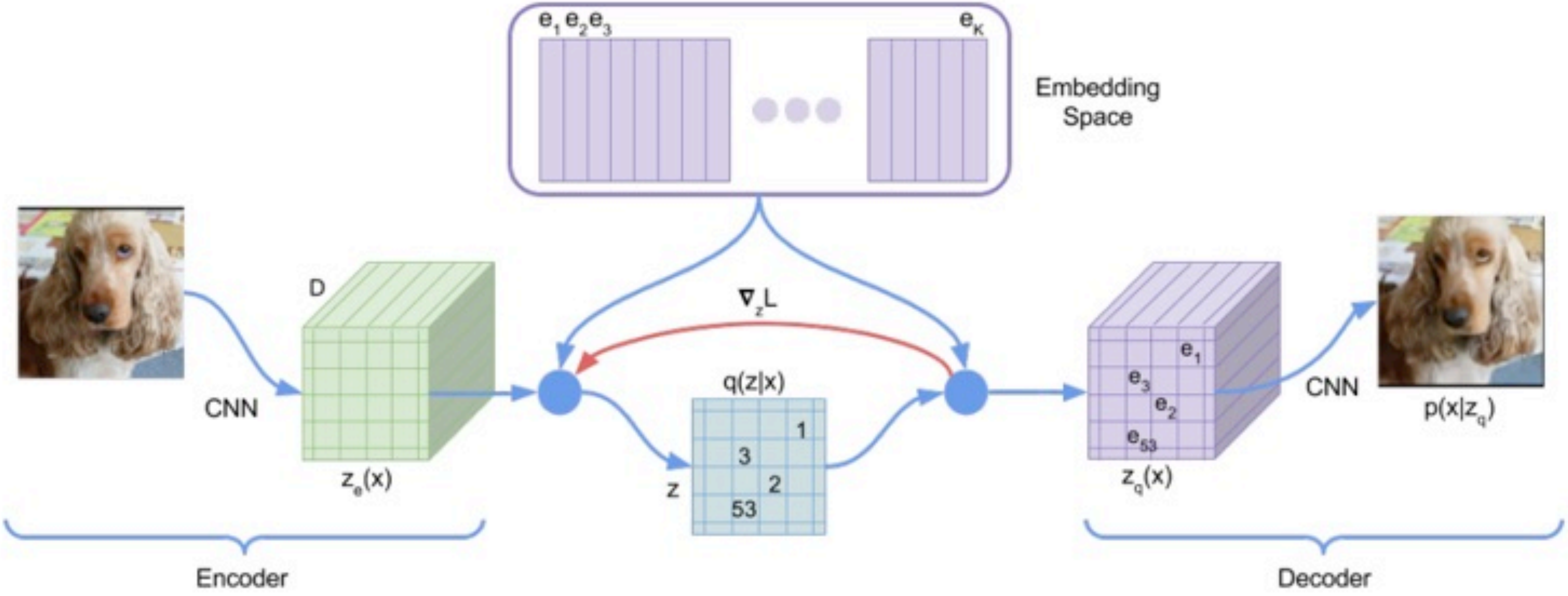
- Cons:
  - Samples blurrier and lower quality compared to GANs

- Active areas of research:
  - More powerful and flexible approximations for relevant probability distributions
  - Combining VAEs and GANs
  - Incorporating structure in latent variables, e.g., hierarchical or categorical distributions

Adapted from [J. Johnson]

Vector Quantised Variational AutoEncoder (VQ-VAE)

- "We show that a discrete latent model (VQ-VAE) performs as well as its continuous model counterparts in log-likelihood."



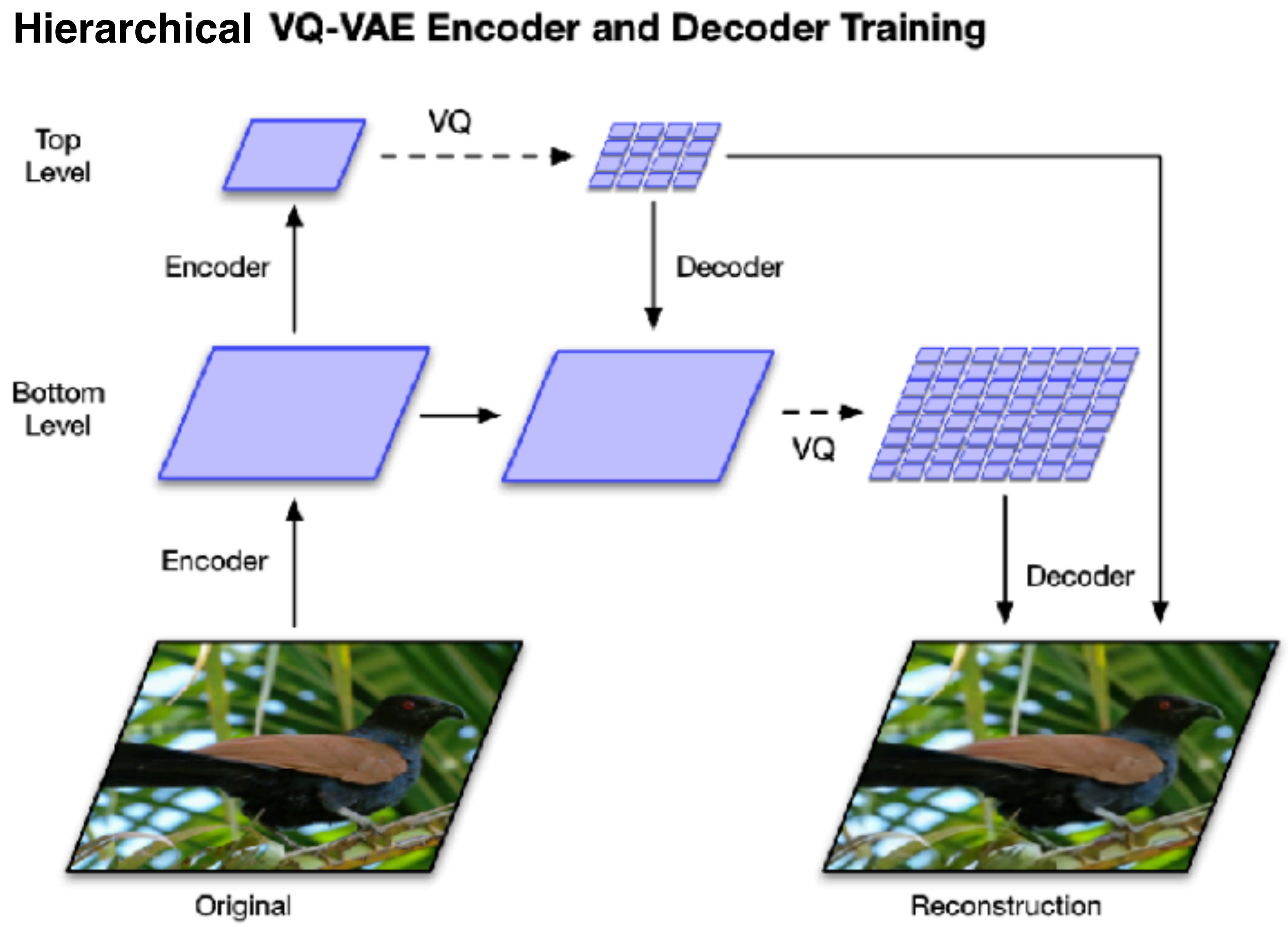
A. van den Oord, O. Vinyals, K. Kavukcuoglu, [Neural Discrete Representation Learning](), NeurIPS 2017

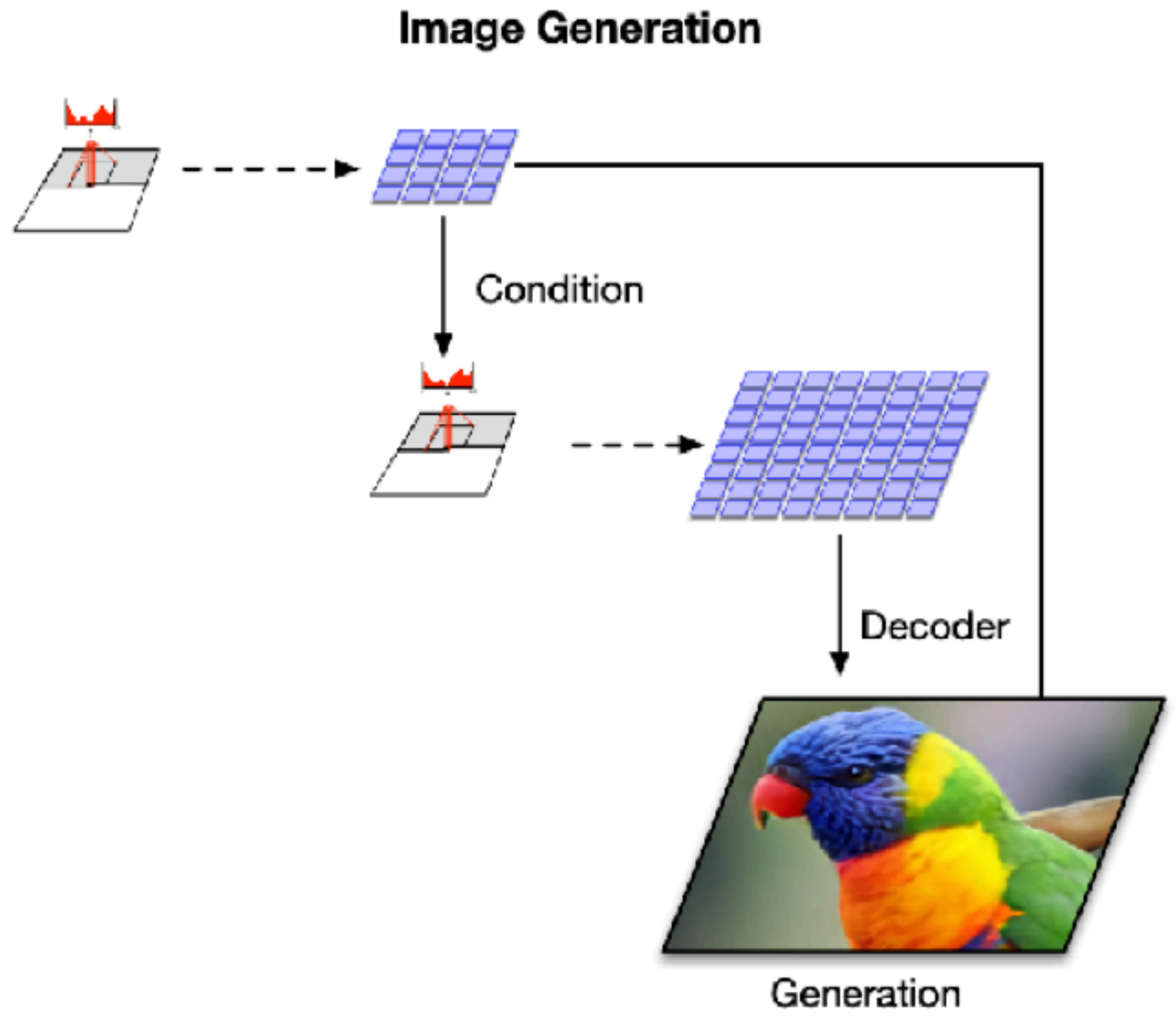
Generating better samples: VQ-VAE-2

- Combining VAE and autoregressive models:

Train a VAE-like model to generate multiscale grids of latent codes

Use a multiscale autoregressive model (PixelCNN) to sample in latent code space





A. Razavi, A. van den Oord, O. Vinyals, [Generating Diverse High-Fidelity Images with VQ-VAE-2](), NeurIPS 2019

Generating better samples: VQ-VAE-2

- 256 x 256 class-conditional samples, trained on ImageNet:

A. Razavi, A. van den Oord, O. Vinyals, [Generating Diverse High-Fidelity Images with VQ-VAE-2](), NeurIPS 2019

Generating better samples: VQ-VAE-2

- 1024 x 1024 generated faces, trained on FFHQ:

A. Razavi, A. van den Oord, O. Vinyals, [Generating Diverse High-Fidelity Images with VQ-VAE-2](), NeurIPS 2019

Combining VAEs and Transformers: DALL-E

- Train an encoder similar to **VQ-VAE** to compress images to 32x32 grids of discrete tokens (each assuming 8192 values)
- Concatenate with text strings, learn a joint sequential transformer model that can be used to **generate image based on text** prompt

We will come back to text-conditioning.

(a) a tapir made of accordion. a tapir with the texture of an accordion.

(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

(c) a neon sign that reads "backprop". a neon sign that reads "backprop". backprop neon sign

A. Ramesh et al., [Zero-Shot Text-to-Image Generation](), ICML 2021, https://openai.com/blog/dall-e/

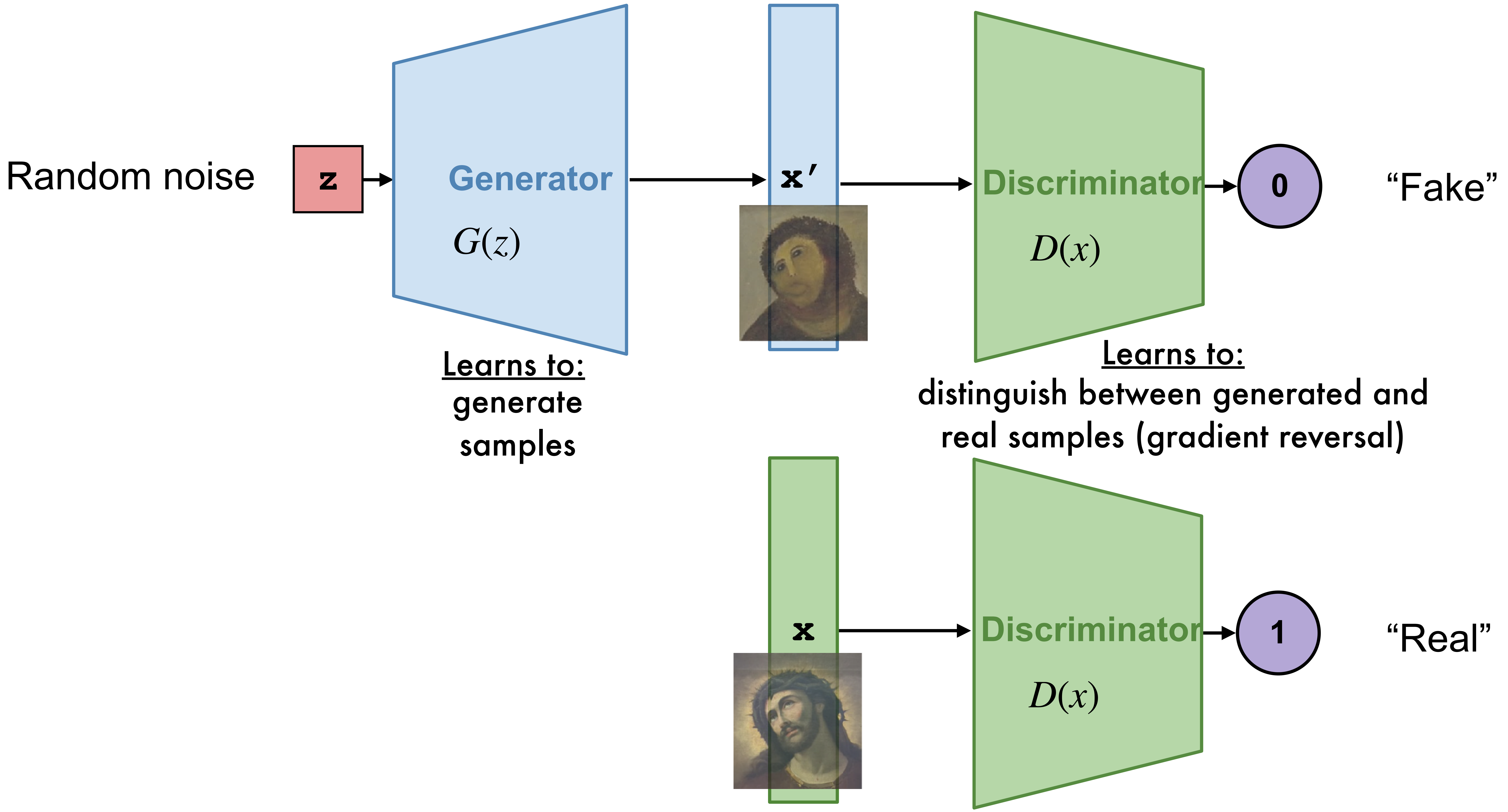
Agenda

1. **Generative neural networks**
    - VAE: Variational autoencoders
    - GAN: Generative adversarial networks
    - Diffusion models

2. **Vision & language**
    - Text-to-image retrieval
    - Text-to-image generation
    - Image captioning

Generative Adversarial Networks (GANs)

- Train two networks with opposing objectives:

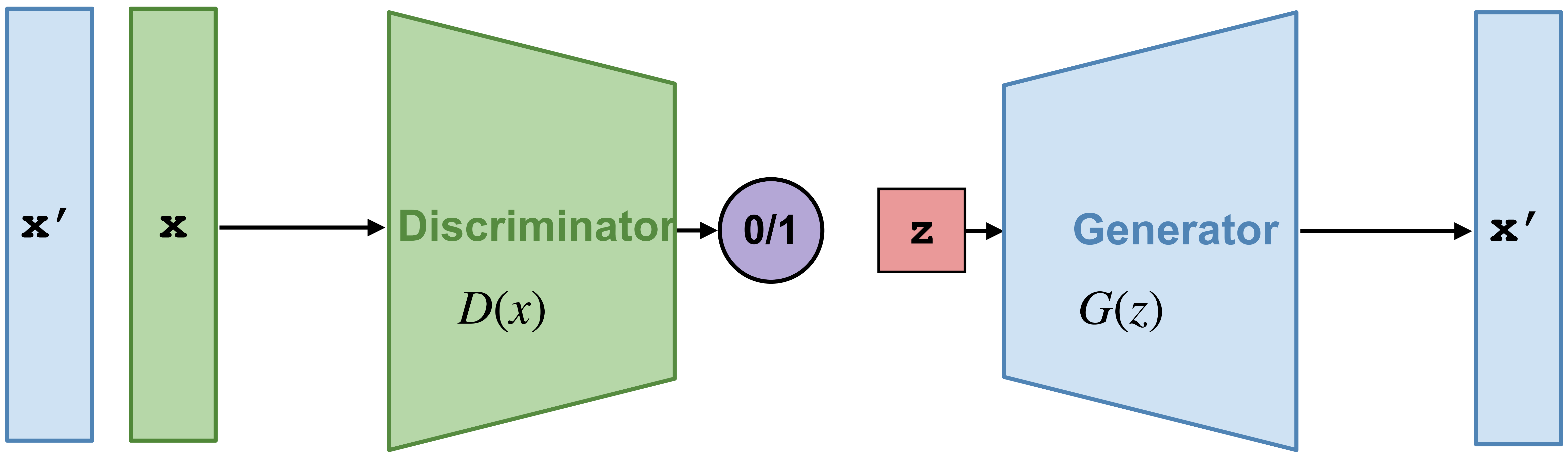


I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, [Generative adversarial nets](), NeurIPS 2014

Generative Adversarial Networks (GANs)

- Train two networks with opposing objectives:

GAN

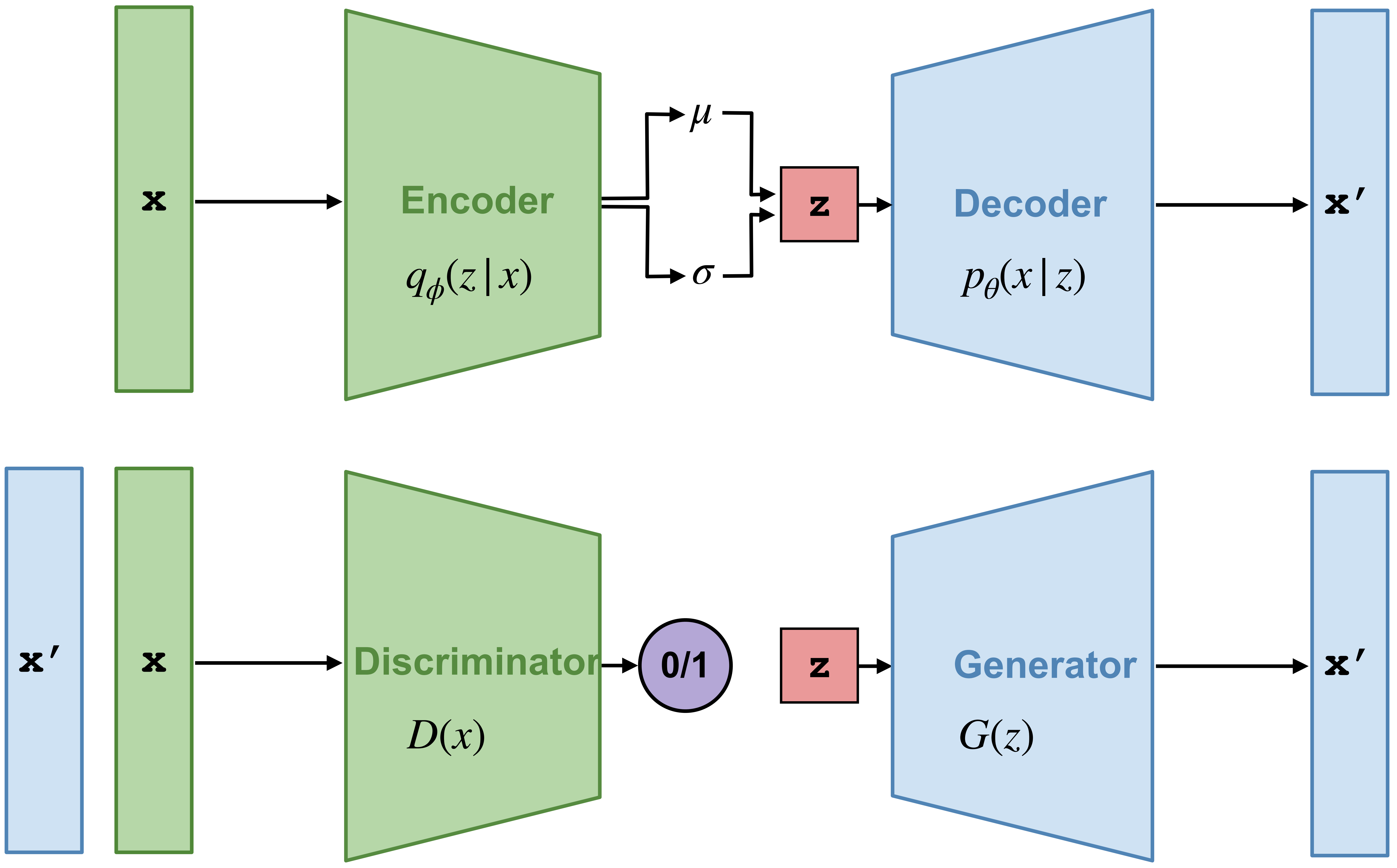


I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, [Generative adversarial nets](), NeurIPS 2014

Generative Adversarial Networks (GANs)

VAE

GAN
implicit density,
i.e., no well-defined
density p(x)



Adapted from: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

GAN objective

- Discriminator $D(x)$ outputs the probability that the sample x is *real.*
- We want $D(x)$ to be close to 1 for real data and close to 0 for fake.
- Expected conditional log likelihood for

real and generated data:

$$= \mathbb{E}_{x\sim p_{\text{data}}} \log D(x) + \mathbb{E}_{z\sim p} \log\left(1 - D\left(G(z)\right)\right)$$

We seed the generator with noise z
drawn from a simple distribution p
(Gaussian or uniform)

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, [Generative adversarial nets](), NeurIPS 2014

GAN objective

$$V(G, D) = \mathbb{E}_{x \sim p_{\text{data}}} \log D(x) + \mathbb{E}_{z \sim p} \log(1 - D(G(z)))$$

- The discriminator wants to correctly distinguish real and fake samples:

$$D^* = \arg\max_D V(G, D)$$

- The generator wants to fool the discriminator:

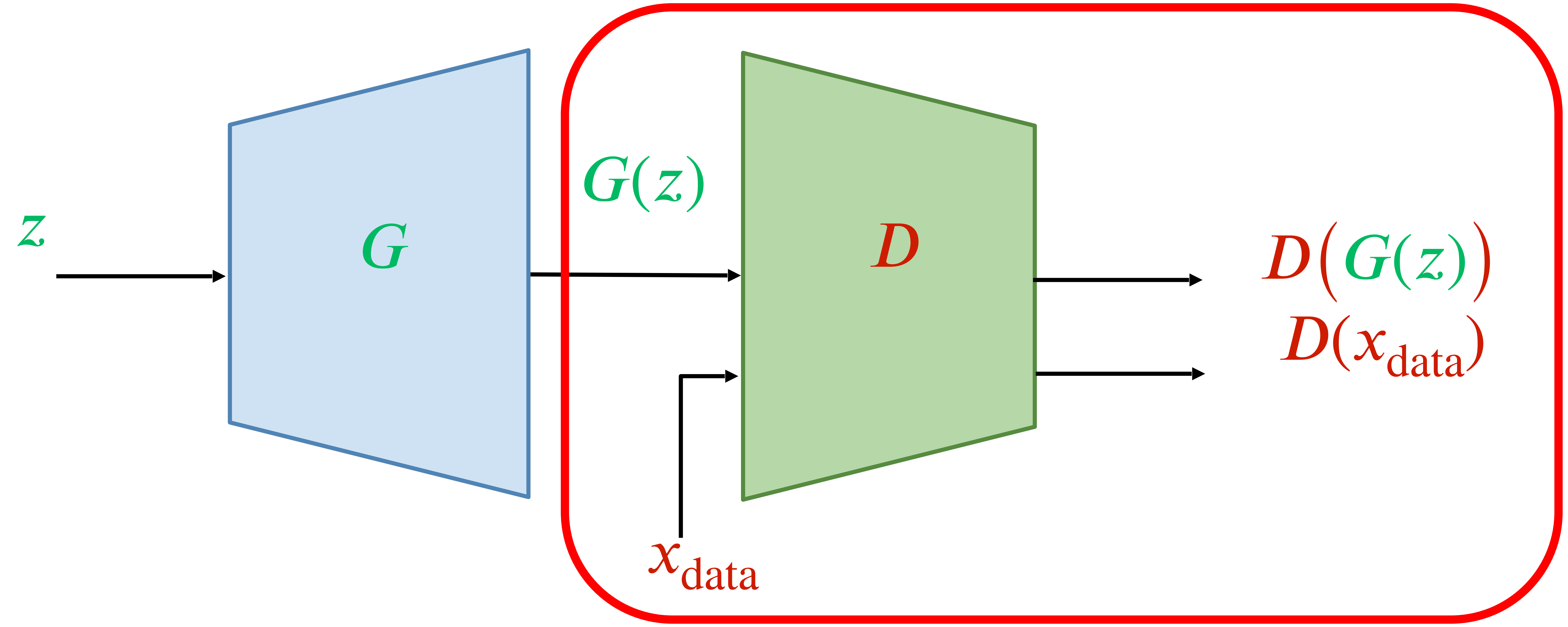
$$G^* = \arg\min_G V(G, D)$$

- Train the generator and discriminator jointly in a *minimax game*

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, [Generative adversarial nets](), NeurIPS 2014

GAN: Schematic picture

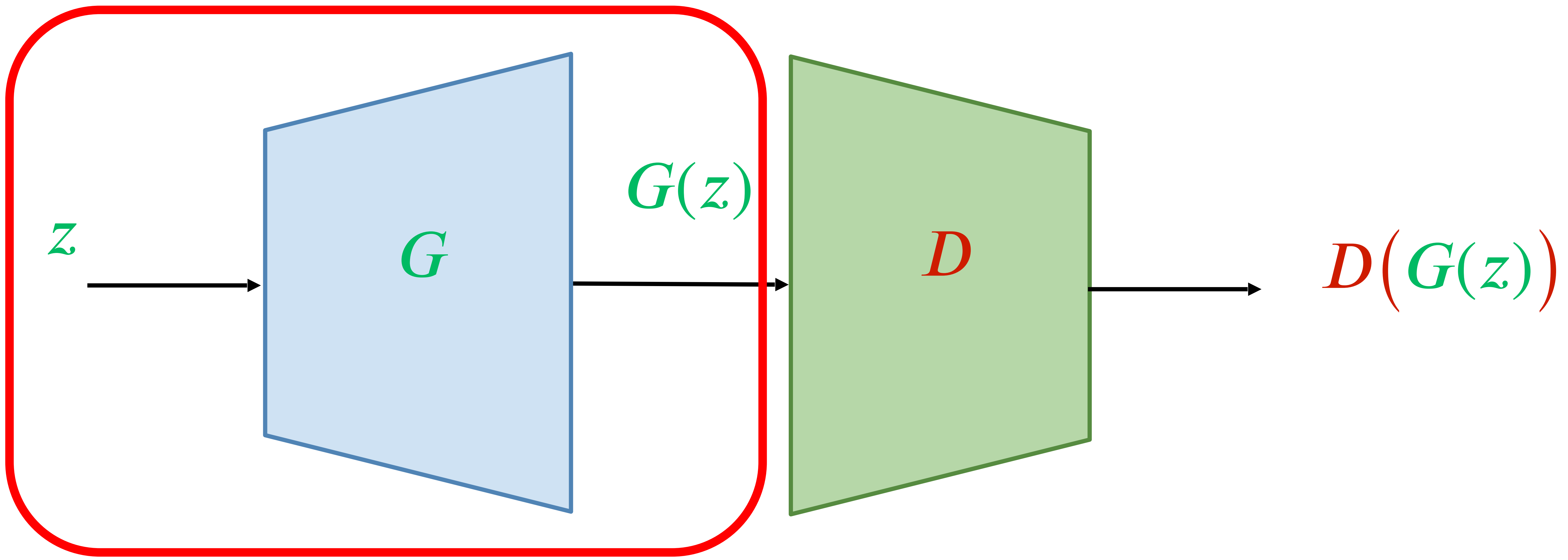
- Update discriminator: push $D(x_{\text{data}})$ close to 1 and $D\big(G(z)\big)$ close to 0
  - The generator is a "black box" to the discriminator



I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, NeurIPS 2014

GAN: Schematic picture

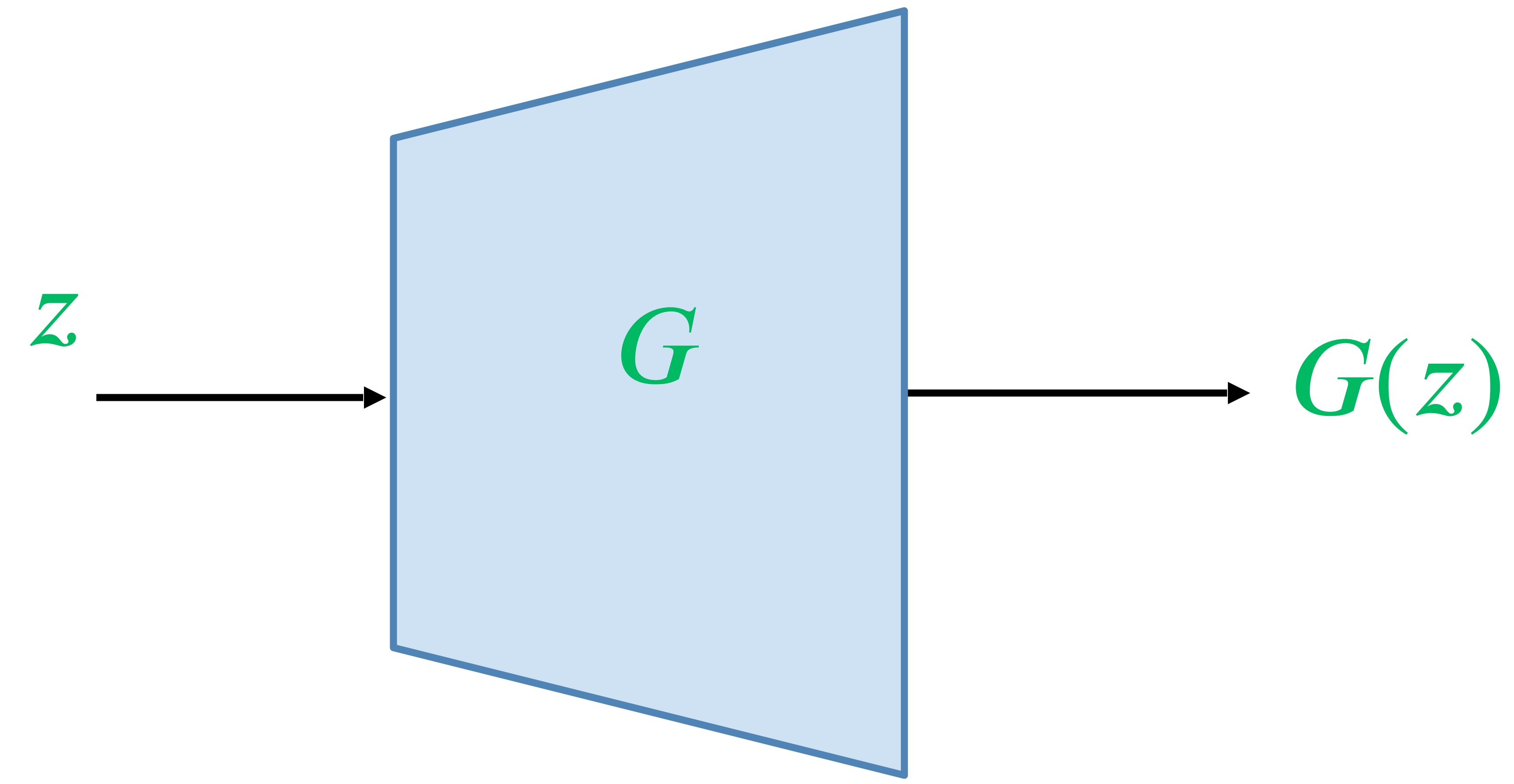
- Update generator: increase $D\big(G(z)\big)$
  - Requires back-propagating through the composed generator-discriminator network (i.e., the discriminator cannot be a black box)
  - The generator is exposed to real data only via the output of the discriminator (and its gradients)



I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, [Generative adversarial nets](), NeurIPS 2014

GAN:

- Test time – the discriminator is discarded

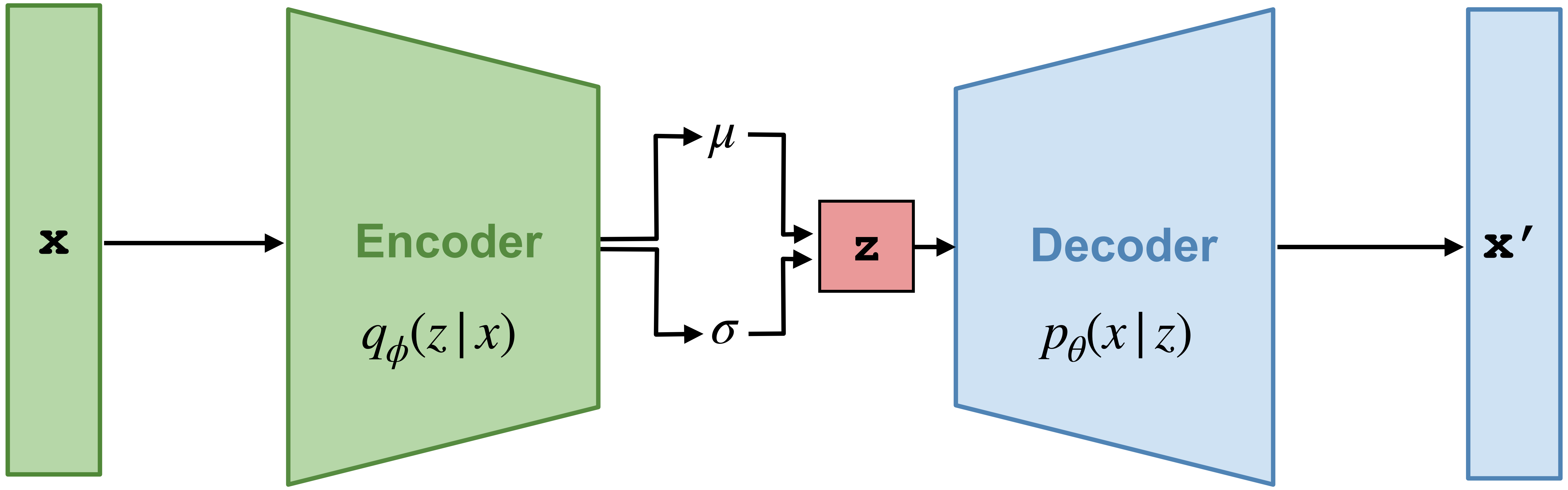


I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, [Generative adversarial nets](), NeurIPS 2014

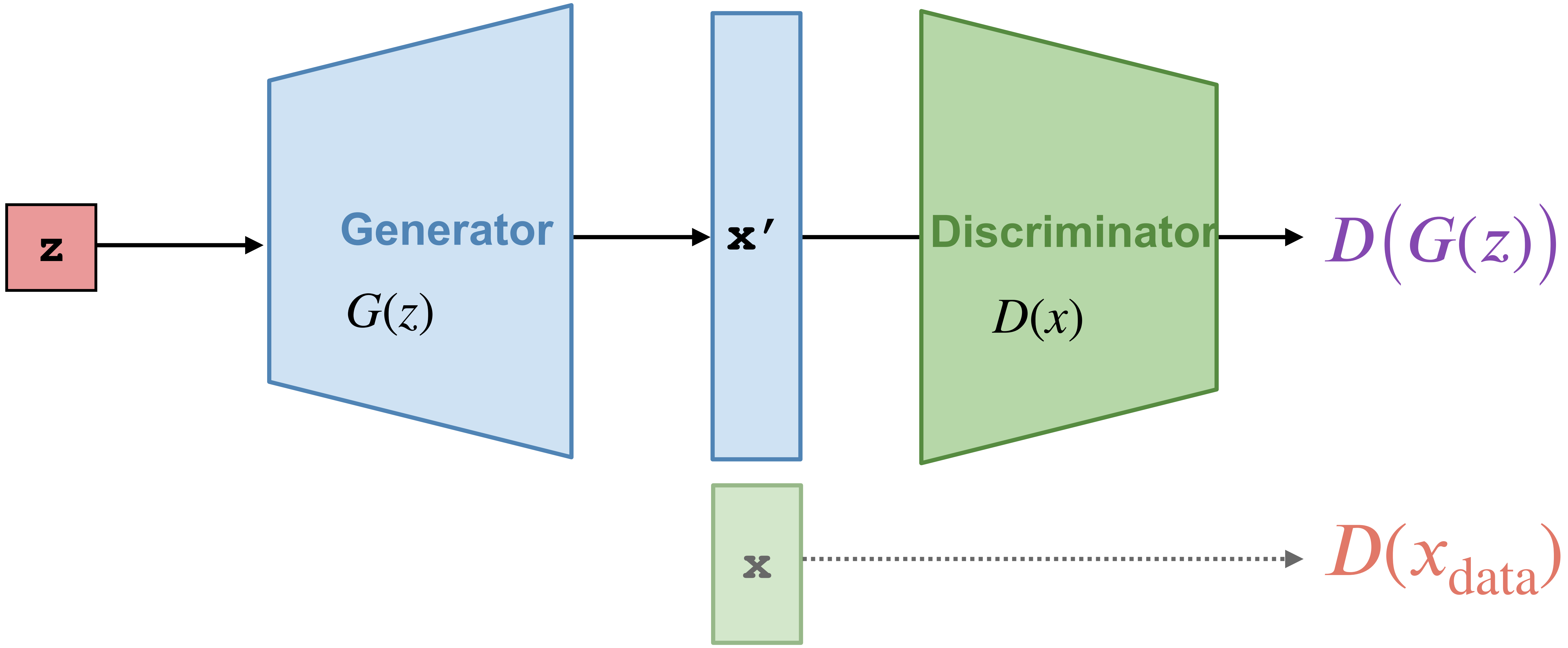
VAEs vs. GANs

VAEs offer more control over the latent space, explicit encoding

VAE training



GAN training
implicit density,
i.e., no well-defined
density p(x)

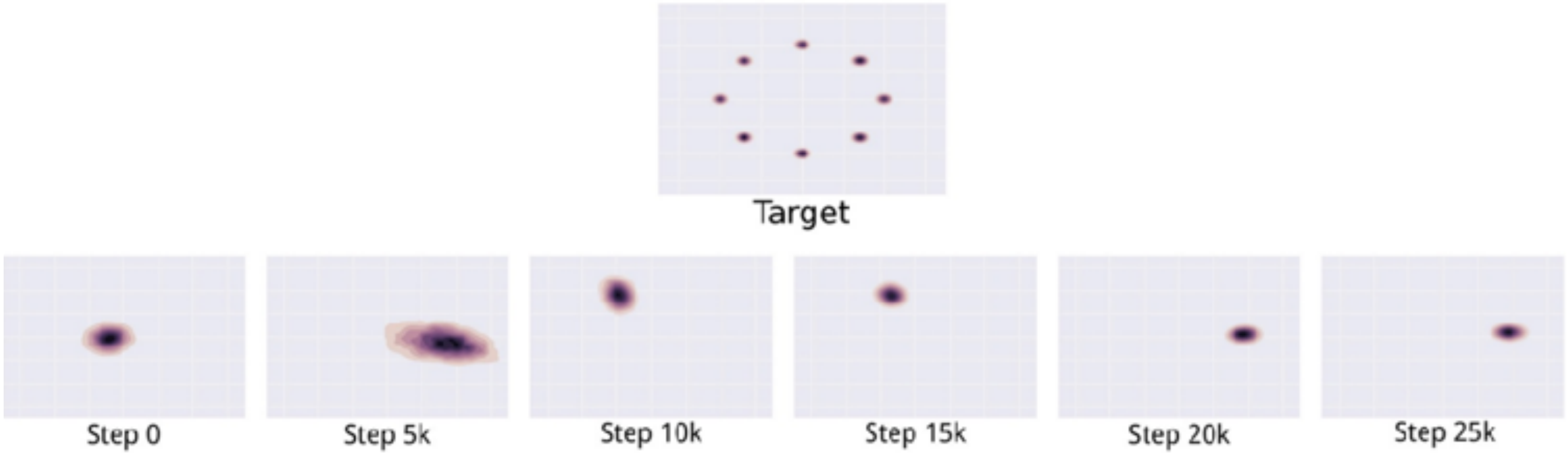


Problems with GAN training

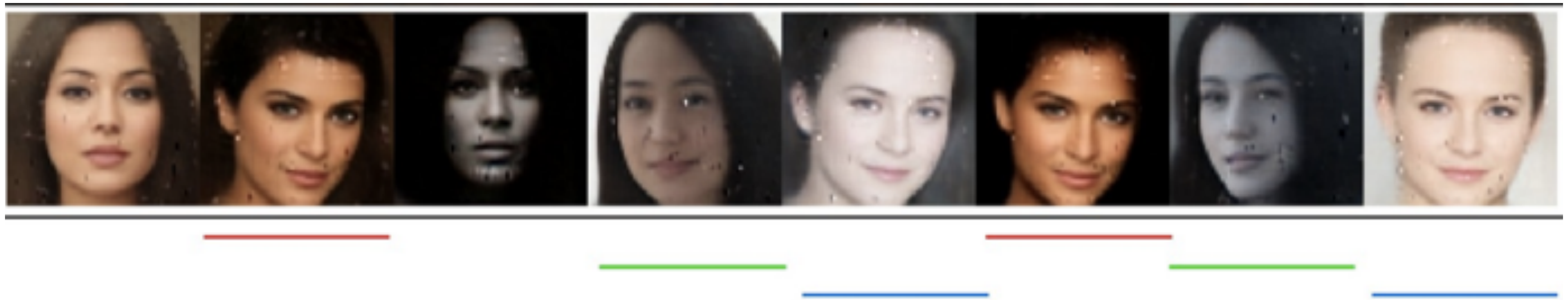
- Stability
  - Parameters can oscillate or diverge, generator loss does not correlate with sample quality
  - Behavior very sensitive to hyperparameter selection

Problems with GAN training

- Mode collapse
  - Generator ends up modeling only a small subset of the training data



Source



Source

Original GAN results

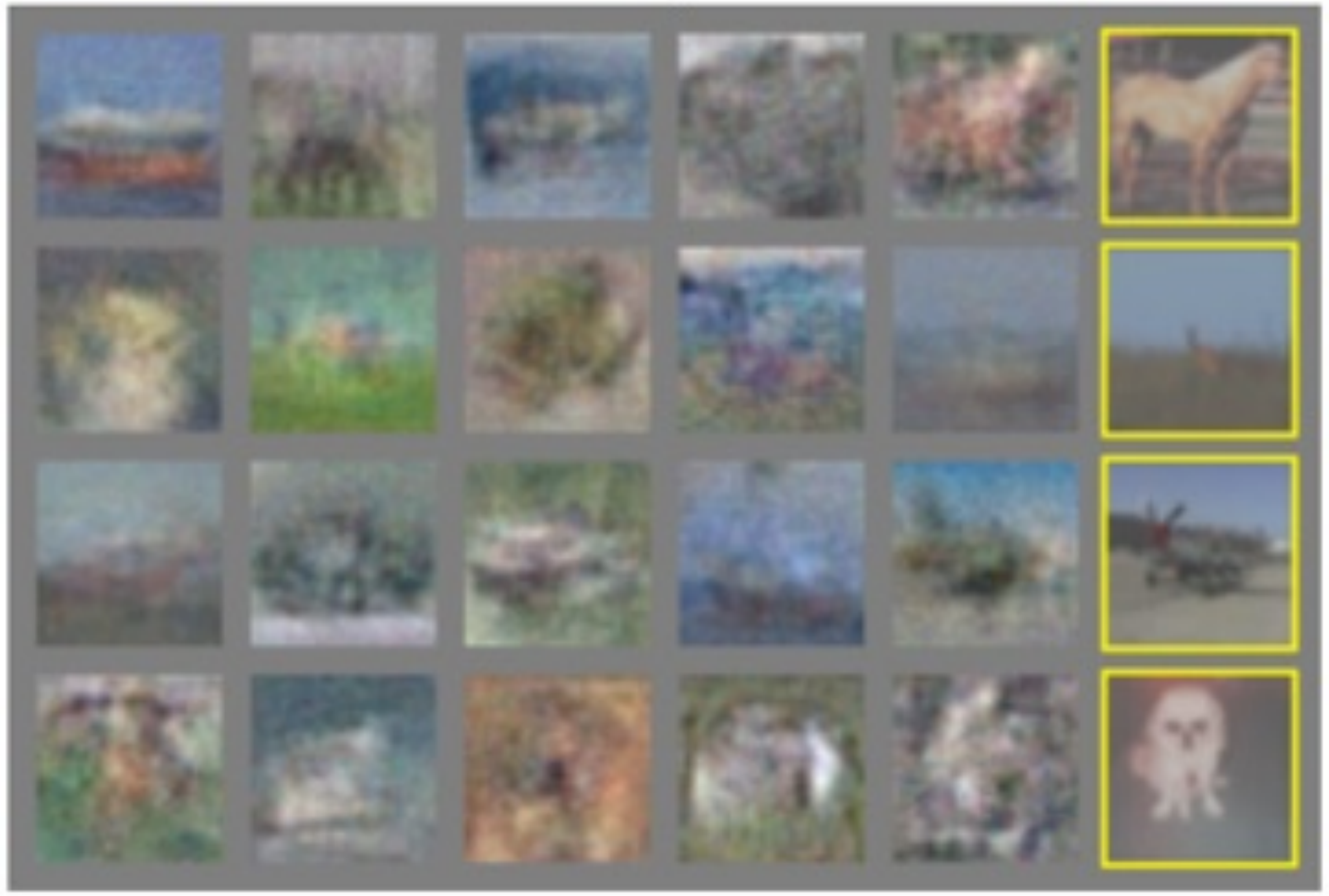
MNIST digits

Toronto Face Dataset

Nearest real image for sample to the left

Original GAN results

CIFAR-10 (FC networks)

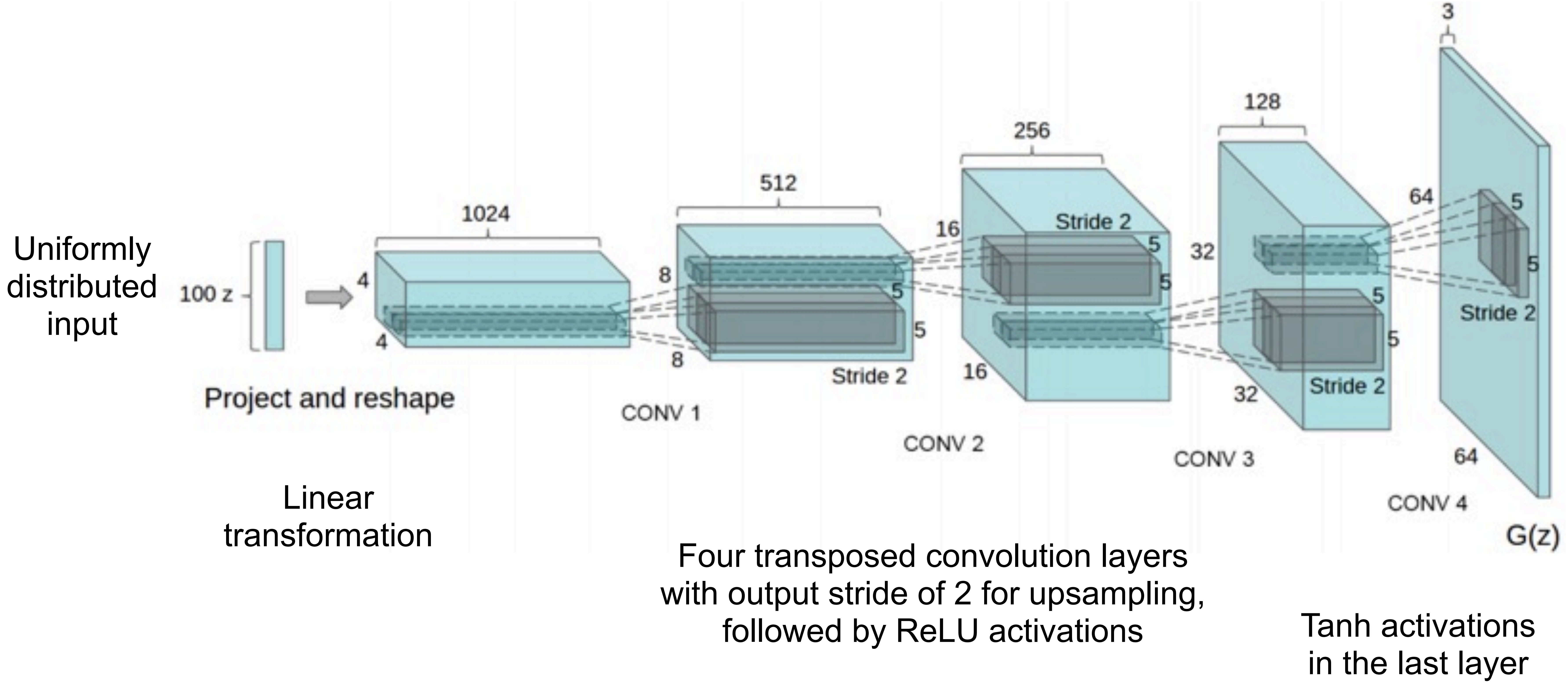


CIFAR-10 (conv networks)

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, NeurIPS 2014

DCGAN (Deep Convolutional GAN)

- Early, influential convolutional architecture for generator



A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, ICLR 2016

DCGAN

- Early, influential convolutional architecture for generator
- Discriminator architecture (empirically determined to give best training stability):
  - Don't use pooling, only strided convolutions
  - Use Leaky ReLU activations (sparse gradients cause problems for training)
  - Use only one FC layer before the softmax output
  - Use batch normalization after most layers (in the generator also)

A. Radford, L. Metz, S. Chintala, [Unsupervised representation learning with deep convolutional generative adversarial networks](), ICLR 2016

DCGAN results

Generated bedrooms after one epoch

A. Radford, L. Metz, S. Chintala, [Unsupervised representation learning with deep convolutional generative adversarial networks](), ICLR 2016

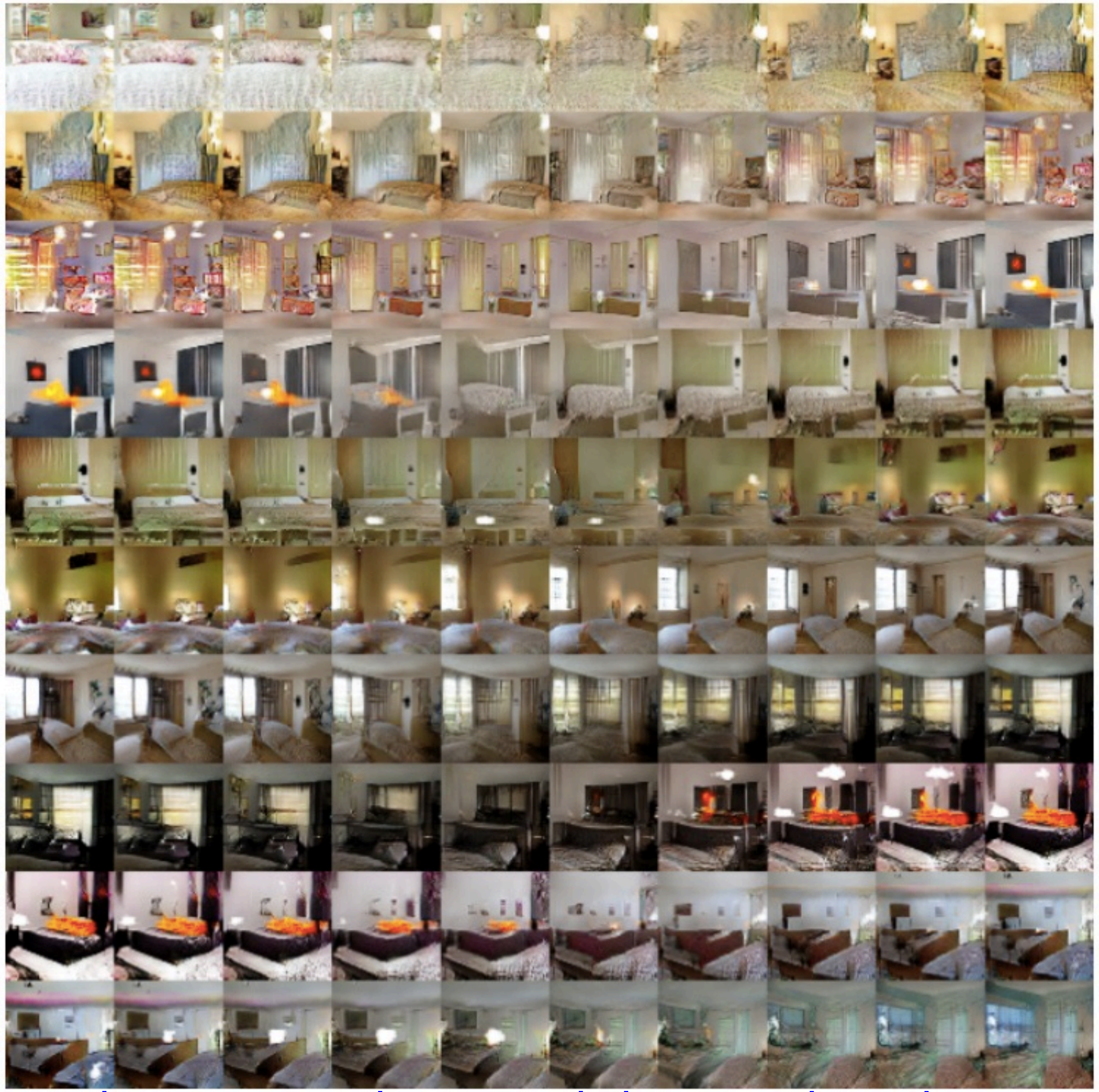
DCGAN results

Generated bedrooms after five epochs

A. Radford, L. Metz, S. Chintala, [Unsupervised representation learning with deep convolutional generative adversarial networks](), ICLR 2016

DCGAN results

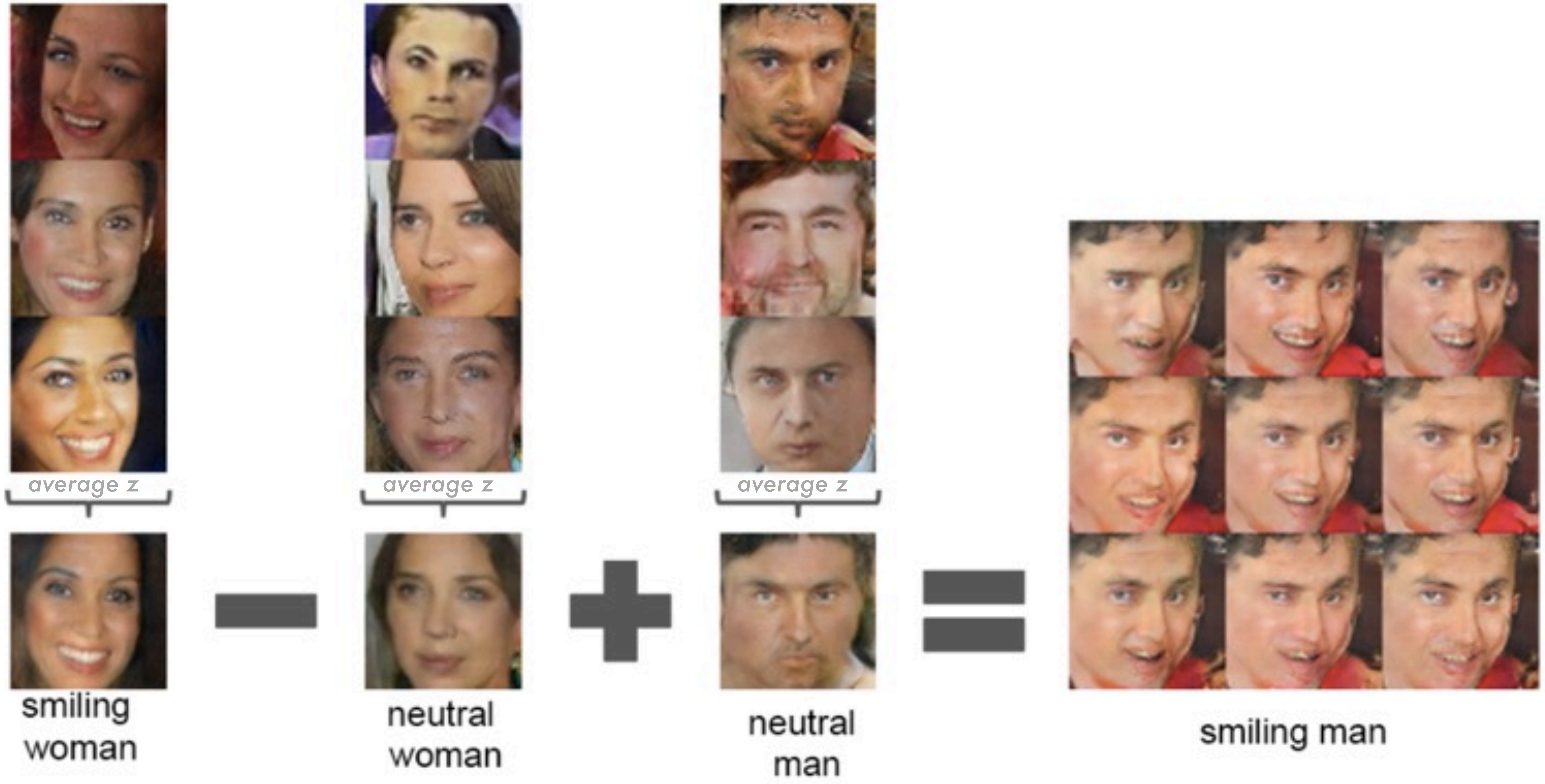
Interpolation between different points in the z space



A. Radford, L. Metz, S. Chintala, [Unsupervised representation learning with deep convolutional generative adversarial networks](), ICLR 2016

DCGAN results

- Vector arithmetic in the z space

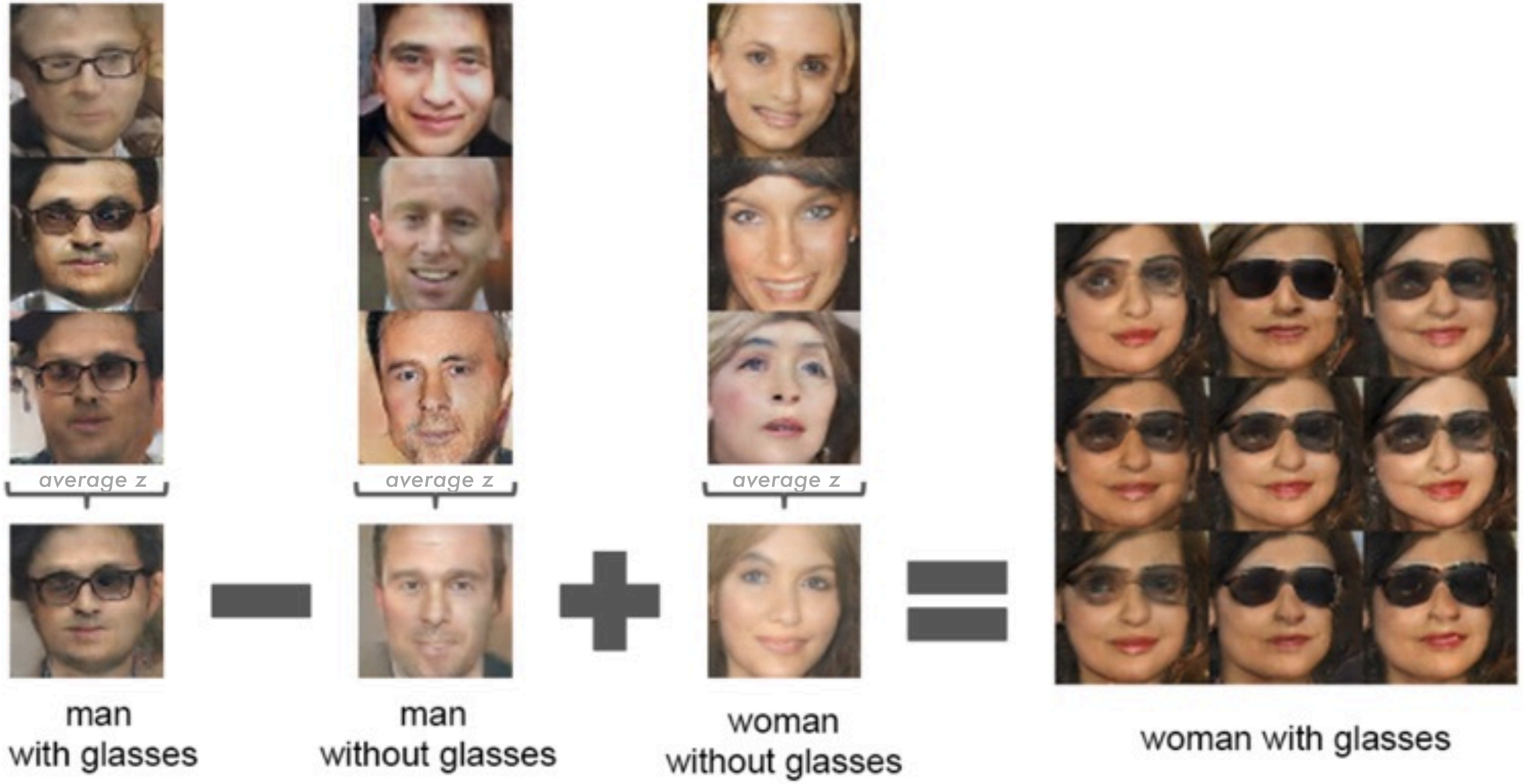


"Experiments working on only single samples per concept were unstable, but averaging the Z vector for three examplars showed consistent and stable generations that semantically obeyed the arithmetic."

A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, ICLR 2016

DCGAN results

- Vector arithmetic in the z space

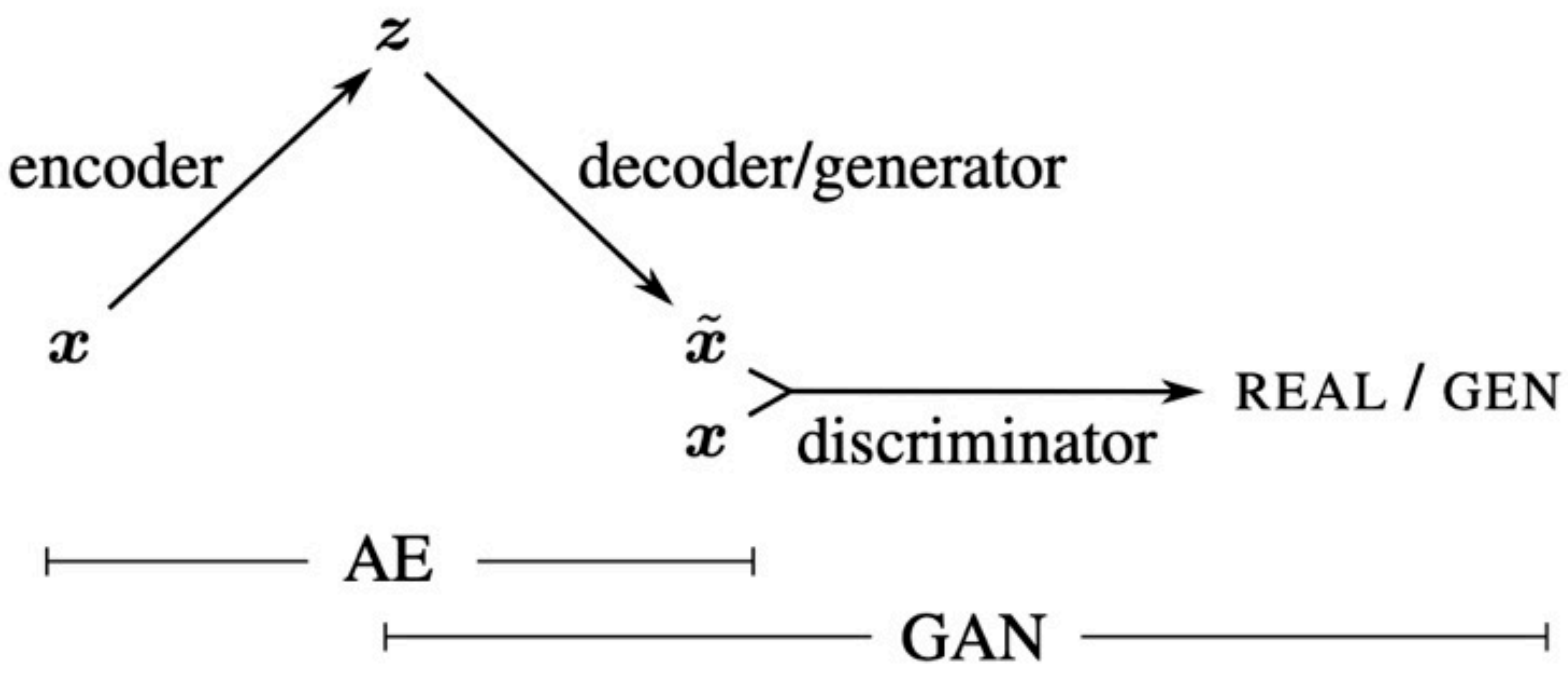


"Experiments working on only single samples per concept were unstable, but averaging the Z vector for three examplars showed consistent and stable generations that semantically obeyed the arithmetic."

A. Radford, L. Metz, S. Chintala, [Unsupervised representation learning with deep convolutional generative adversarial networks](), ICLR 2016

Hybrid approaches: e.g., Combining VAEs and GANs

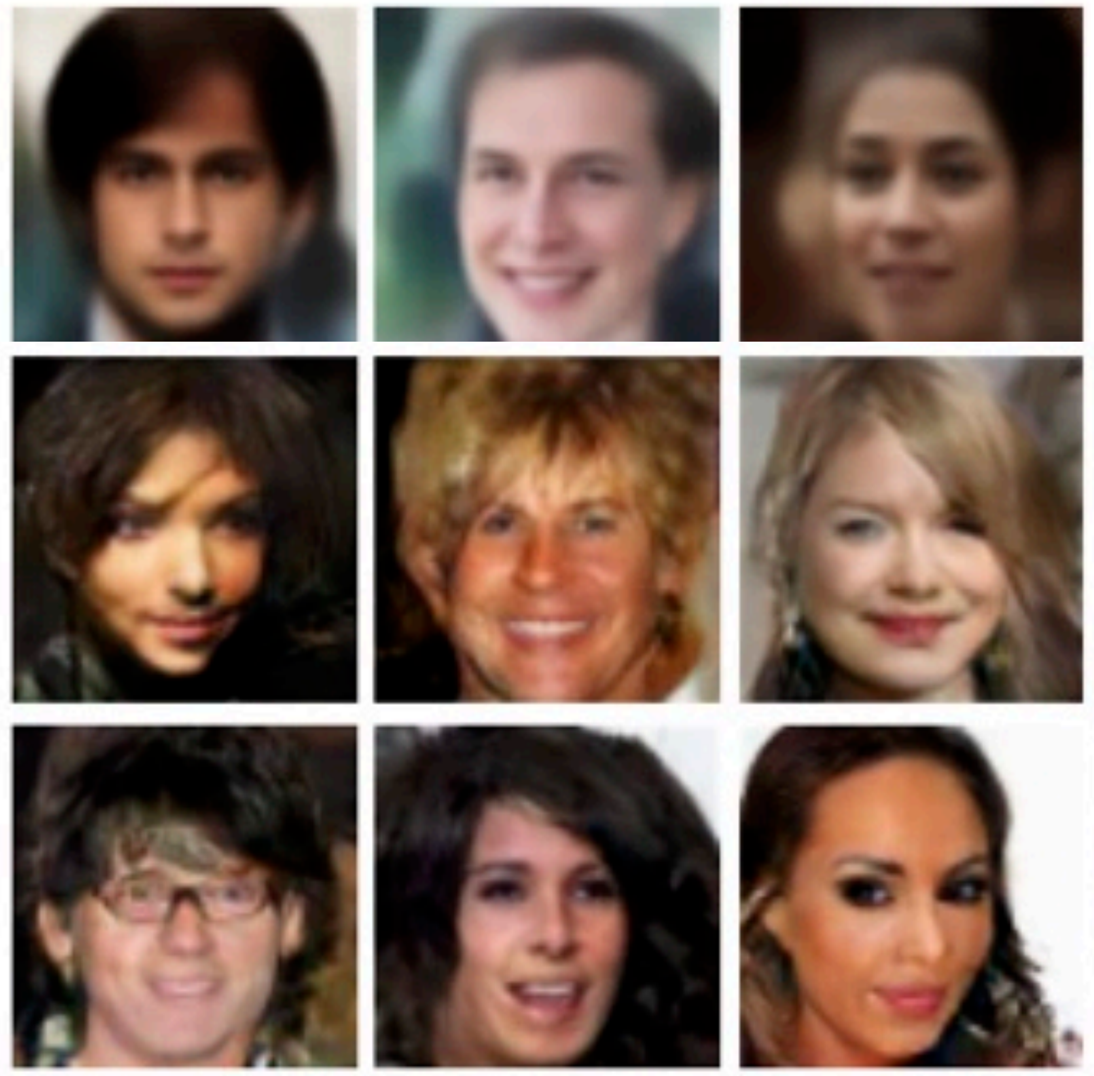
- Define decoder probability model $p_\theta(x \mid z)$ not in terms of reconstruction errors in pixel space, but in terms of errors in discriminator feature space



VAE

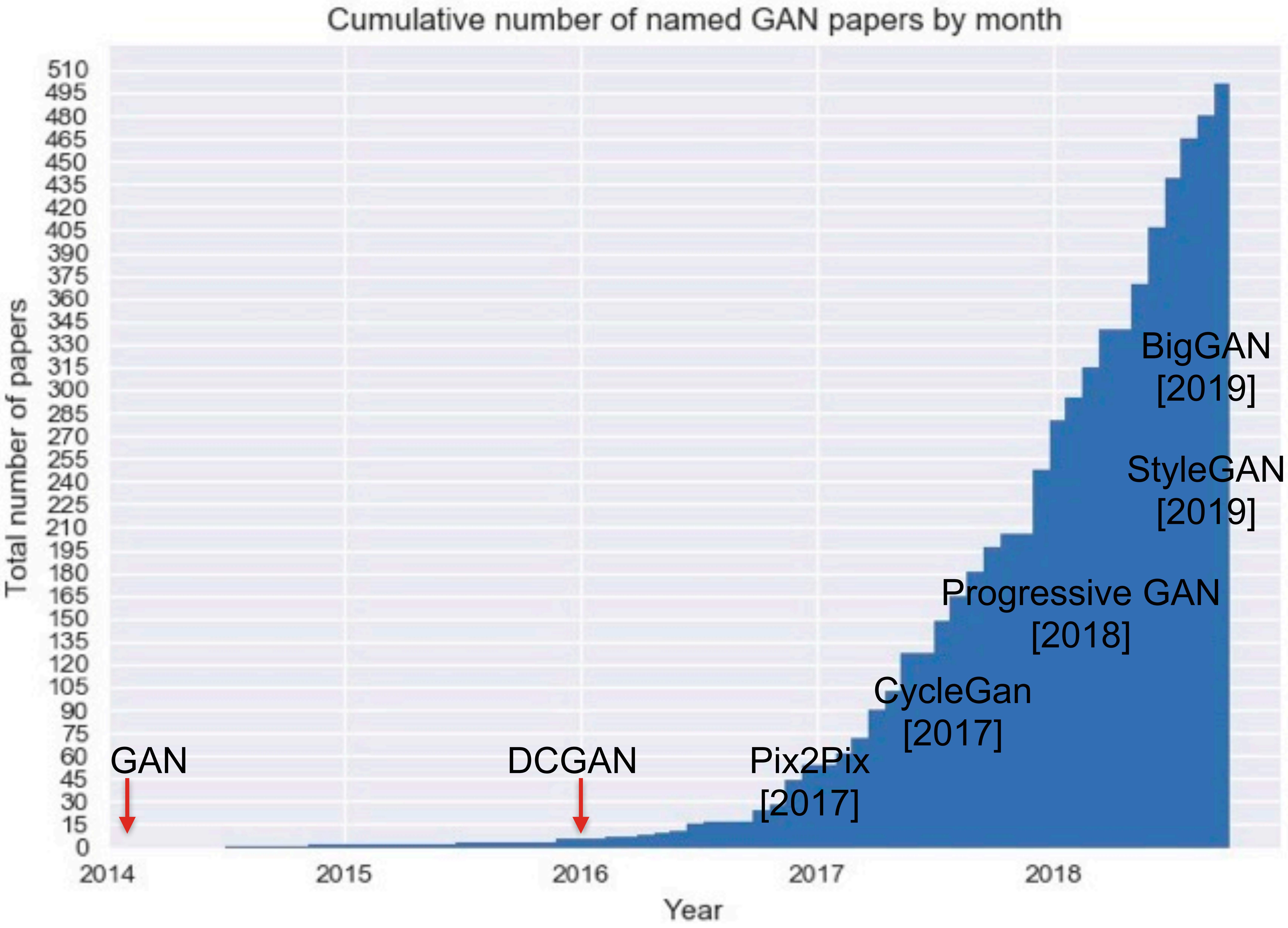
VAE/GAN

GAN



A. Larsen, S. Sonderby, H. Larochelle, O. Winther, Autoencoding beyond pixels using a learned similarity metric, ICML 2016

Fast-forwarding a little…



Source

Progress in GANs

- **Progressive GAN, StyleGAN, StyleGan2** (higher quality)

T. Karras, T. Aila, S. Laine, J. Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. ICLR 2018

T. Karras, S. Laine, T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. CVPR 2019

T. Karras et al. Analyzing and Improving the Image Quality of StyleGAN. CVPR 2020

- **GAN Dissection** (interpretability)

D. Bau et al. GAN Dissection: Visualizing and understanding generative adversarial networks. ICLR 2019

- **BigGan** (class-conditioned)

A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, ICLR 2019

- **Pix2Pix, CycleGan** (image-conditioned)

P. Isola, J.-Y. Zhu, T. Zhou, A. Efros, Image-to-Image Translation with Conditional Adversarial Networks, CVPR 2017

J.-Y. Zhu, T. Park, P. Isola, A. Efros, Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, ICCV 2017

…

Progress in GANs: Faces

Ian Goodfellow
@goodfellow_ian

4.5 years of GAN progress on face generation.
arxiv.org/abs/1406.2661 arxiv.org/abs/1511.06434
arxiv.org/abs/1606.07536 arxiv.org/abs/1710.10196
arxiv.org/abs/1812.04948

6:40 PM · Jan 14, 2019

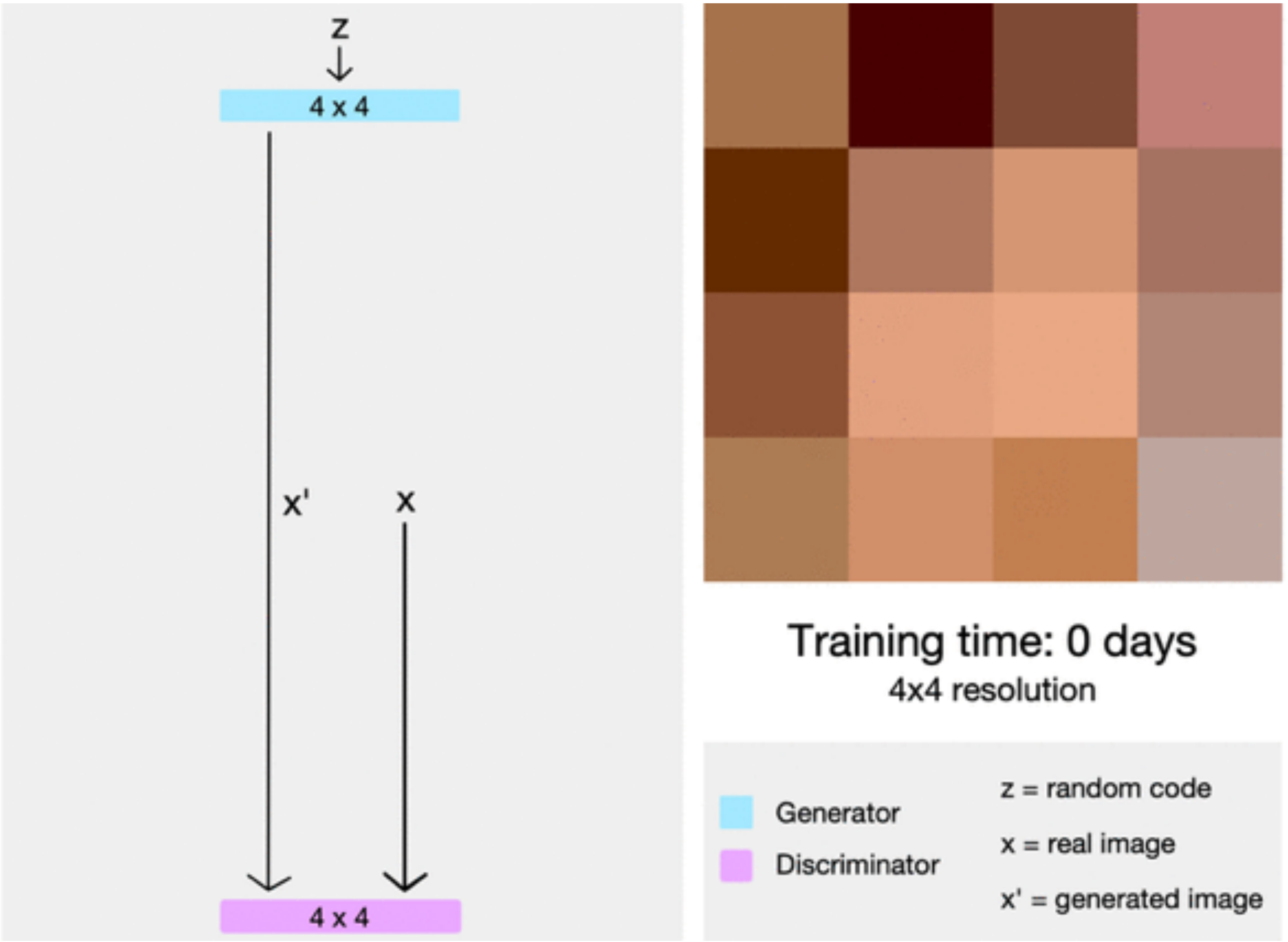
Progressive GANs

Realistic face images up to 1024 x 1024 resolution

T. Karras, T. Aila, S. Laine, J. Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. ICLR 2018

Progressive GANs

- Key idea: train lower-resolution models, gradually add layers corresponding to higher-resolution outputs



[Source]

T. Karras, T. Aila, S. Laine, J. Lehtinen. [Progressive Growing of GANs for Improved Quality, Stability, and Variation]. ICLR 2018

Progressive GANs: Results

256 x 256 results for LSUN categories

"A separate network was trained for each category using identical parameters."

StyleGAN: Results

Built on top of Progressive GAN

T. Karras, S. Laine, T. Aila. [A Style-Based Generator Architecture for Generative Adversarial Networks](). CVPR 2019

GAN Dissection

D. Bau et al. [GAN Dissection: Visualizing and understanding generative adversarial networks](). ICLR 2019

GAN Dissection

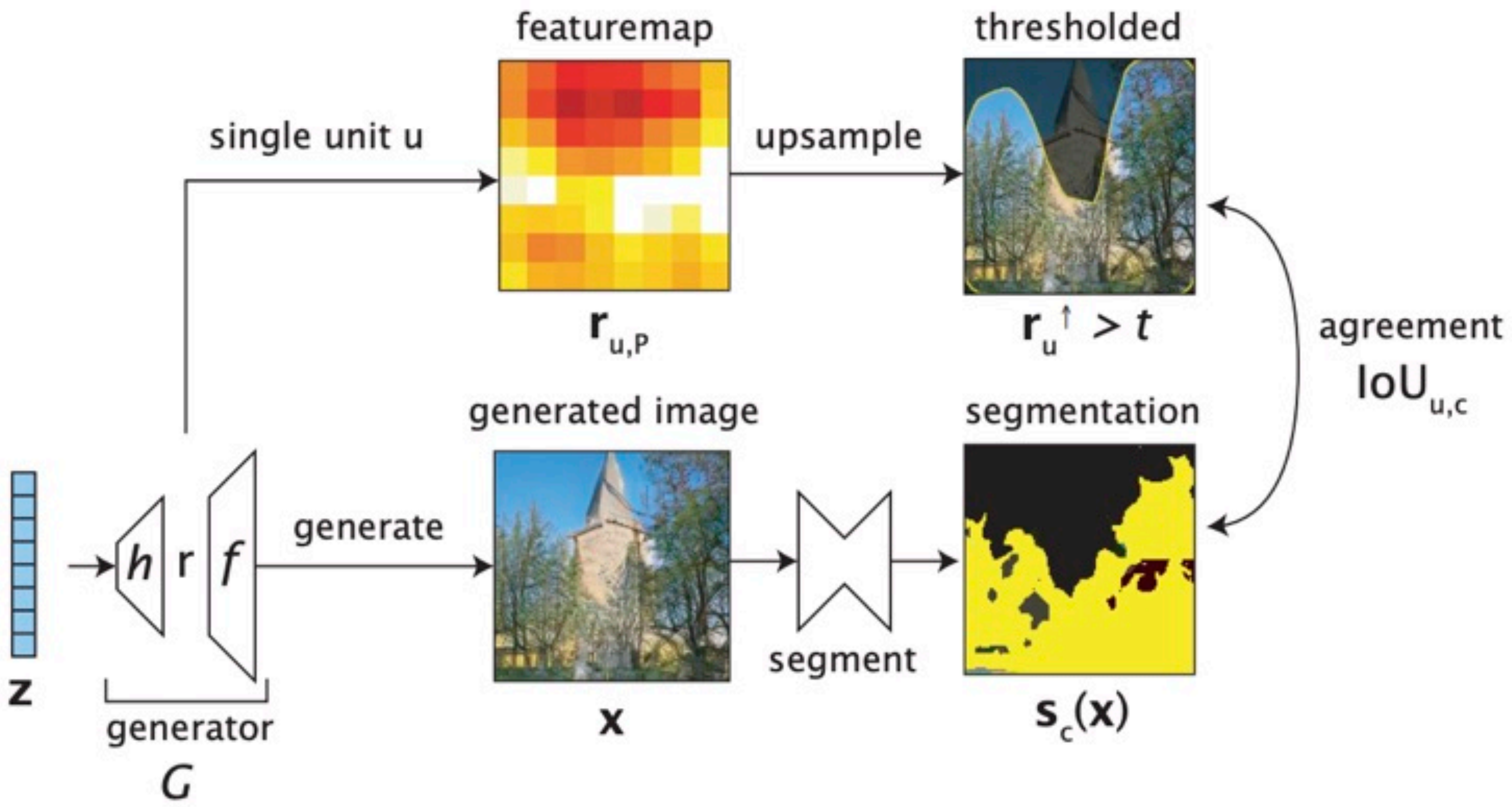
GAN dissection allows us to ask:

1. Does the network learn internal neurons that match meaningful concepts?
2. Do these sets of neurons merely correlate with objects, or does the GAN use those neurons to reason about objects?
3. Can causal neurons be manipulated to improve the output of a GAN?

D. Bau et al. [GAN Dissection: Visualizing and understanding generative adversarial networks](). ICLR 2019

GAN Dissection

- Dissection: measure agreement between a unit and a concept



D. Bau et al. [GAN Dissection: Visualizing and understanding generative adversarial networks](). ICLR 2019

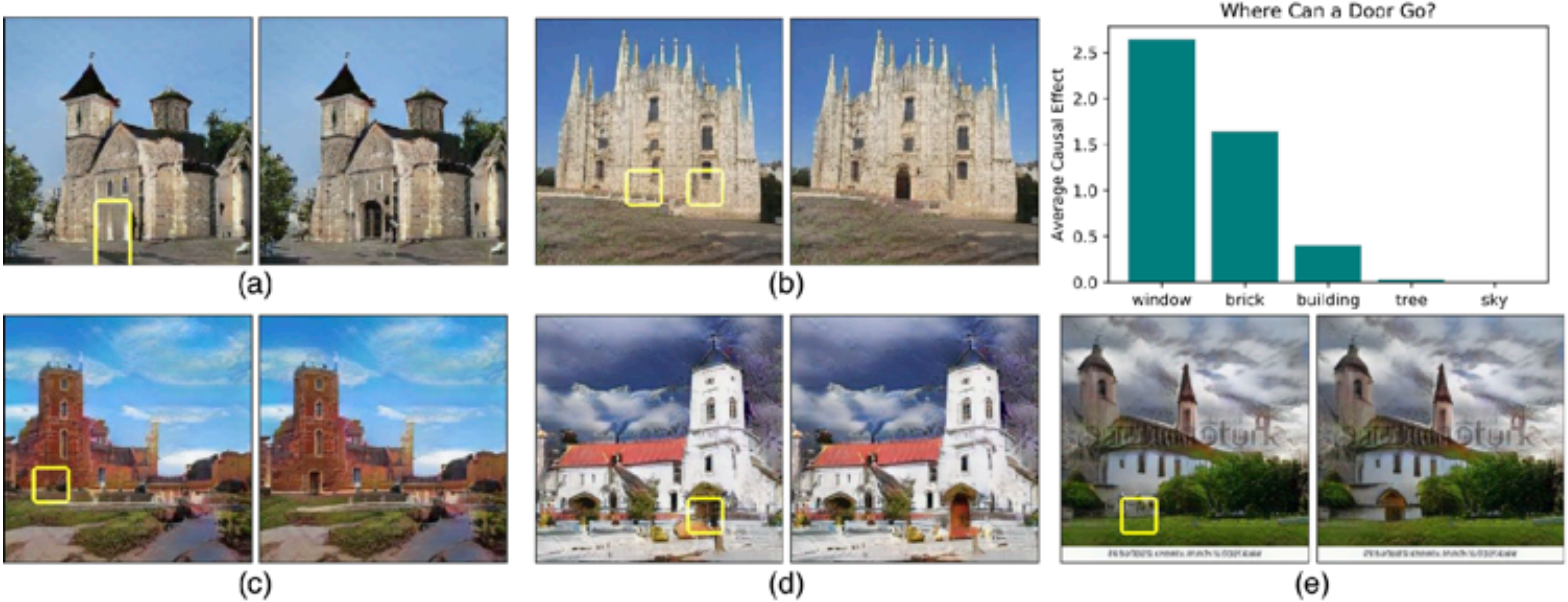
GAN Dissection

- Intervention: measure the causal effect of a set of units and a concept

D. Bau et al. [GAN Dissection: Visualizing and understanding generative adversarial networks](). ICLR 2019

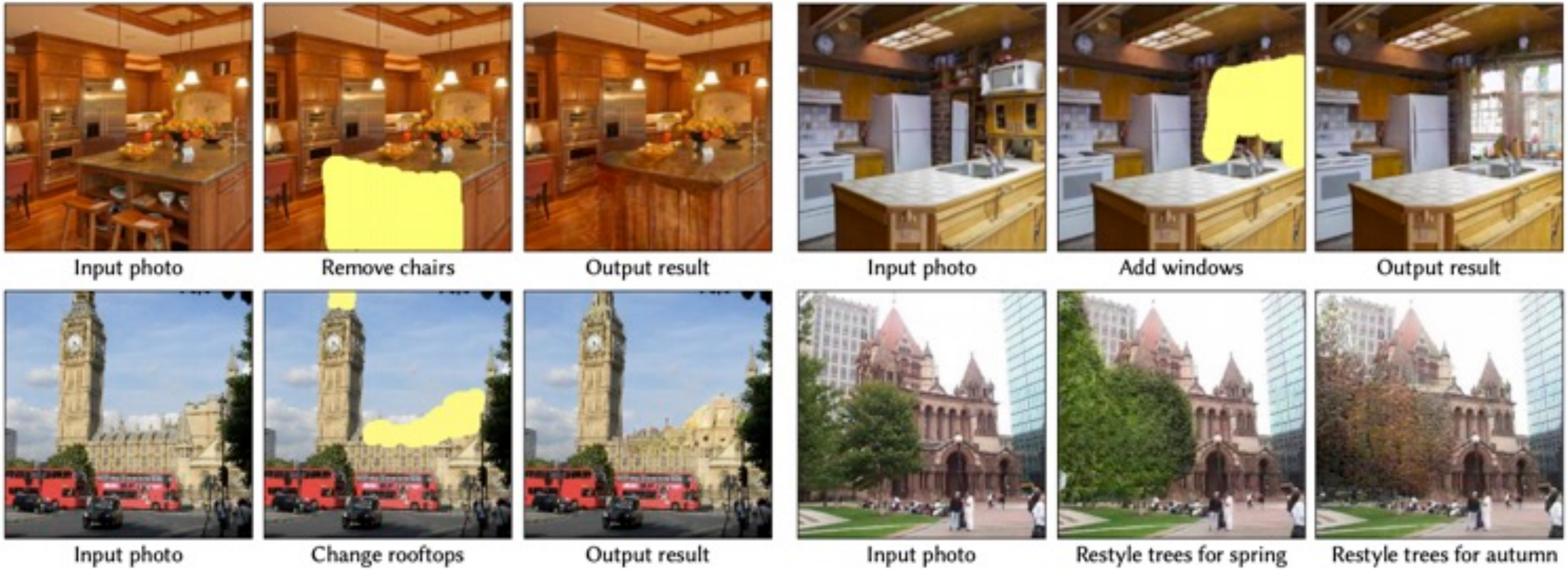
GAN Dissection

*“The network also understands when it can and cannot compose objects. For example, turning on neurons for a door in the proper location of a building will **add a door**. But doing the same in the sky or on a tree will typically have no effect. This structure can be quantified.”*



D. Bau et al. [GAN Dissection: Visualizing and understanding generative adversarial networks](). ICLR 2019

GANPaint demo



https://ganpaint.io/

(Demo currently not available)

D. Bau et al. [GAN Dissection: Visualizing and understanding generative adversarial networks](). ICLR 2019

Parenthesis:
(Conditional generation

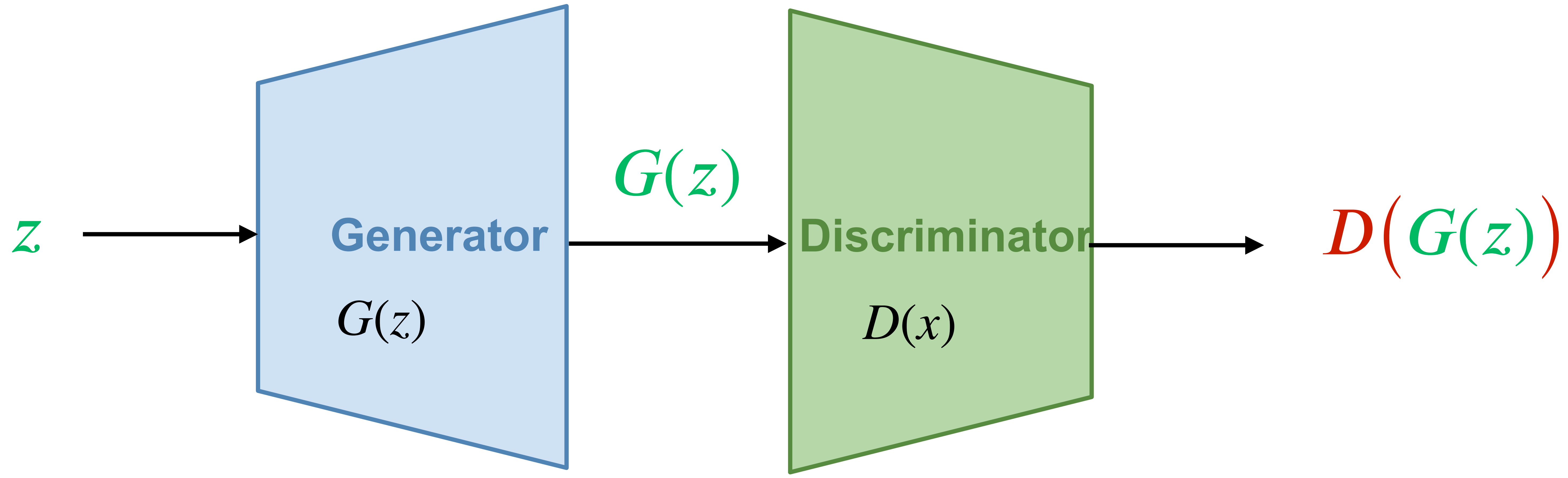
Conditional generation

- One may want to *control* the generation, e.g., instead of a random image sample, generating for a given:
    - category label (class conditioning),
    - natural language description (text conditioning),
    - …

- Simply add the condition as input.

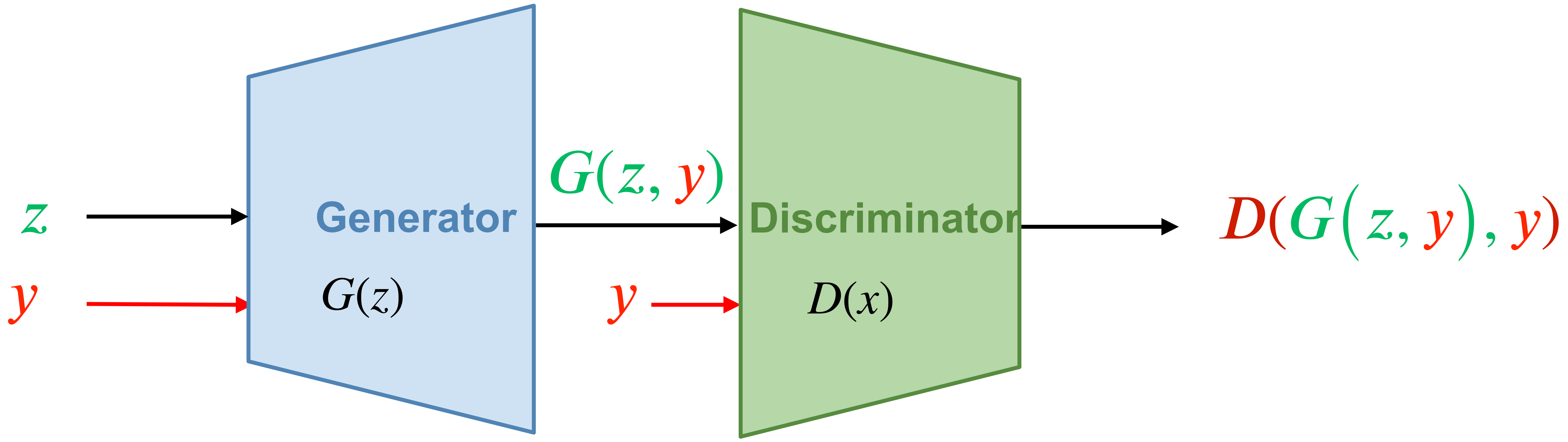
Conditional GANs:

- To condition the generation of samples on discrete side information (e.g., label) y, we need to add y as an input to **both generator and discriminator**



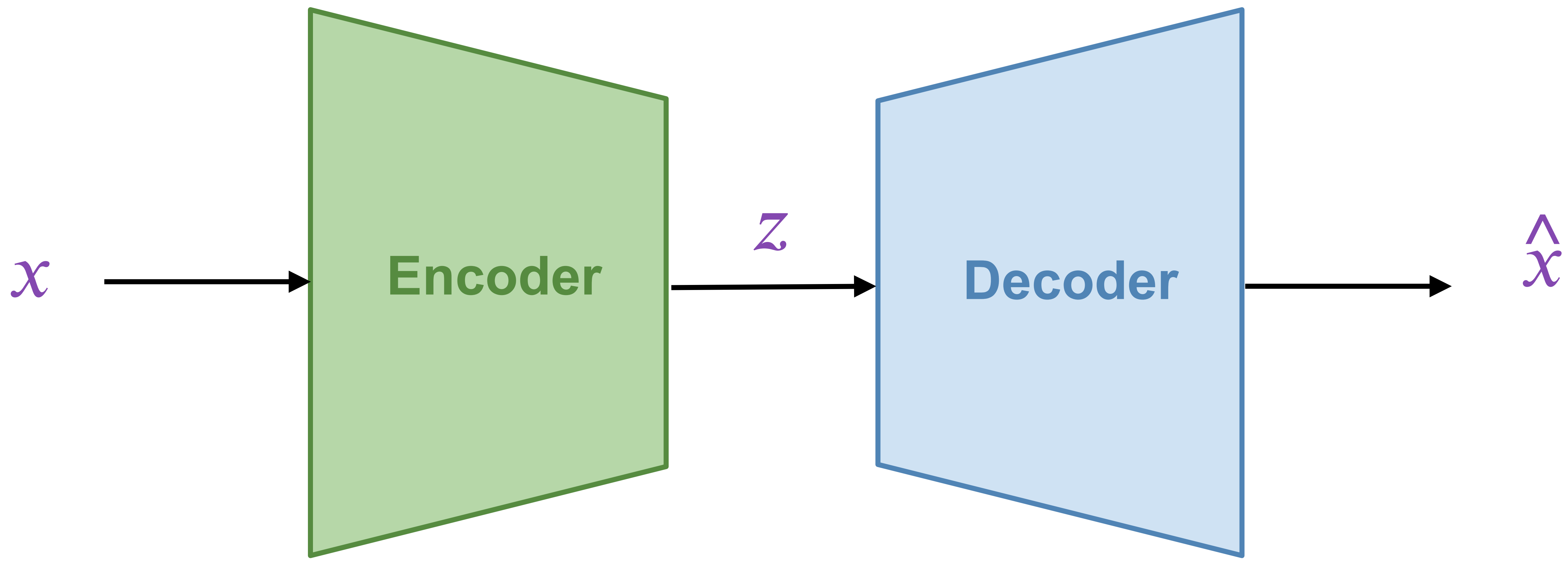
Conditional GANs:

- To condition the generation of samples on discrete side information (e.g., label) y, we need to add y as an input to **both generator and discriminator**



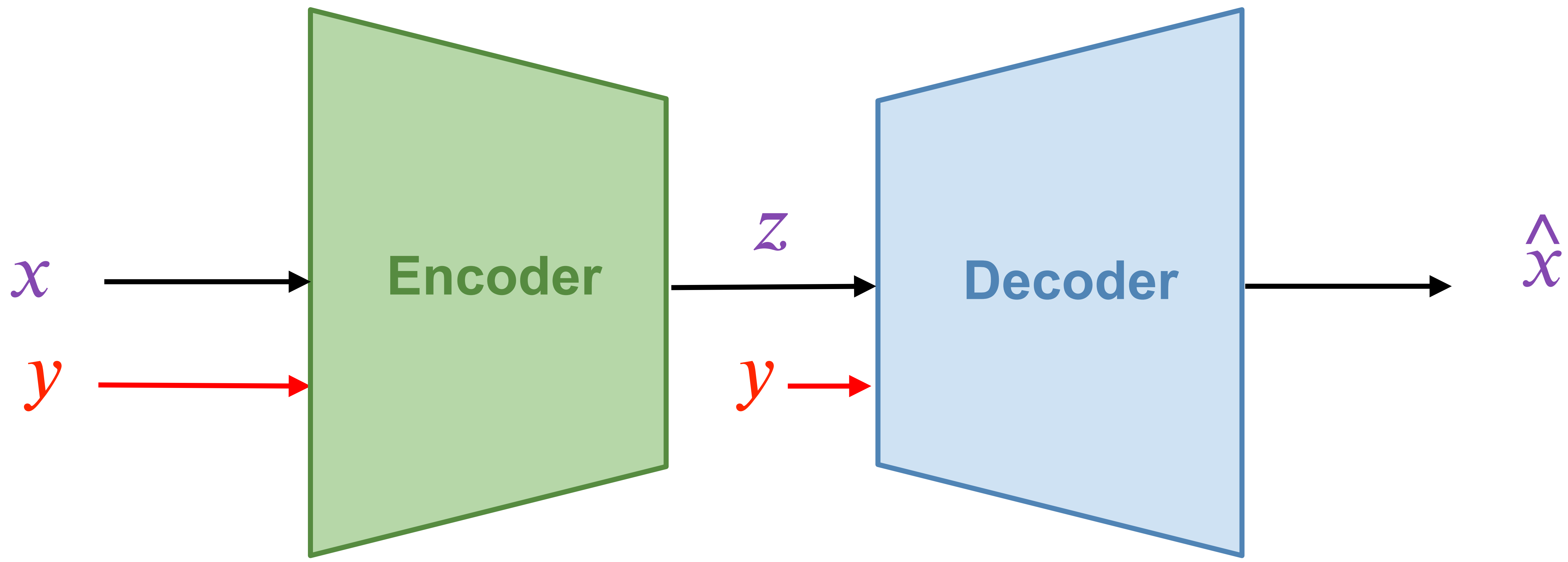
Conditional VAEs:

- To condition the generation of samples on discrete side information (e.g., label) y, we need to add y as an input to **both encoder and decoder**



Conditional VAEs:

- To condition the generation of samples on discrete side information (e.g., label) y, we need to add y as an input to **both encoder and decoder**



Parenthesis Closed: Conditional generation)

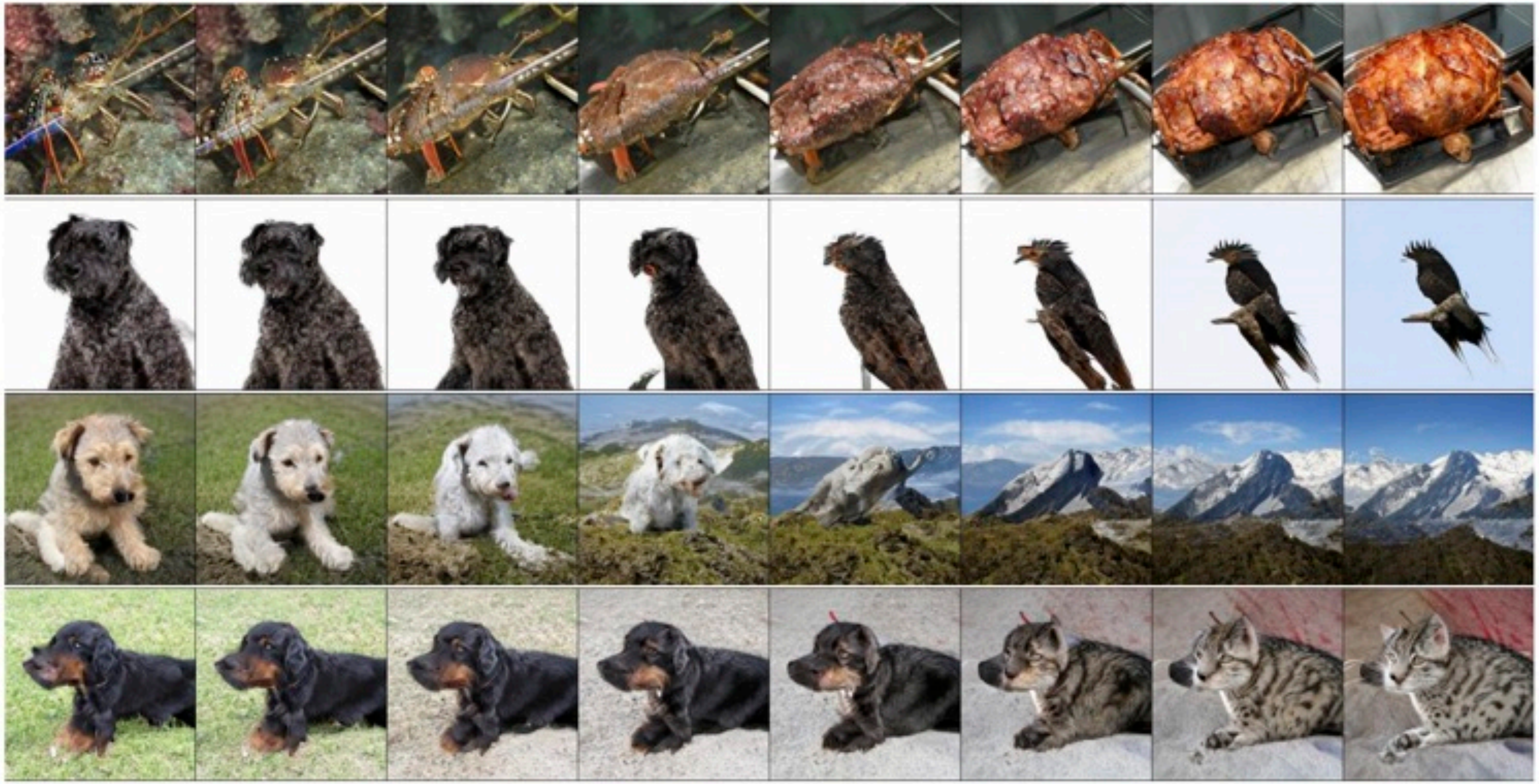
BigGAN

- Class-conditional generation of ImageNet images up to 512 x 512 resolution

A. Brock, J. Donahue, K. Simonyan, [Large scale GAN training for high fidelity natural image synthesis](), ICLR 2019

BigGAN: Results

- Interpolation between class c with noise z held constant:



BigGAN: Results

- Interpolation between c, z pairs:

BigGAN: Results

- Difficult classes:

- Human bodies are still difficult today in 2023

Progress in GANs

- **Progressive GAN, StyleGAN, StyleGan2** **(higher quality)**

T. Karras, T. Aila, S. Laine, J. Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. ICLR 2018

T. Karras, S. Laine, T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. CVPR 2019

T. Karras et al. Analyzing and Improving the Image Quality of StyleGAN. CVPR 2020

- **GAN Dissection** **(interpretability)**

D. Bau et al. GAN Dissection: Visualizing and understanding generative adversarial networks. ICLR 2019

- **BigGan** **(class-conditioned)**

A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, ICLR 2019

- **Pix2Pix, CycleGan** **(image-conditioned)**

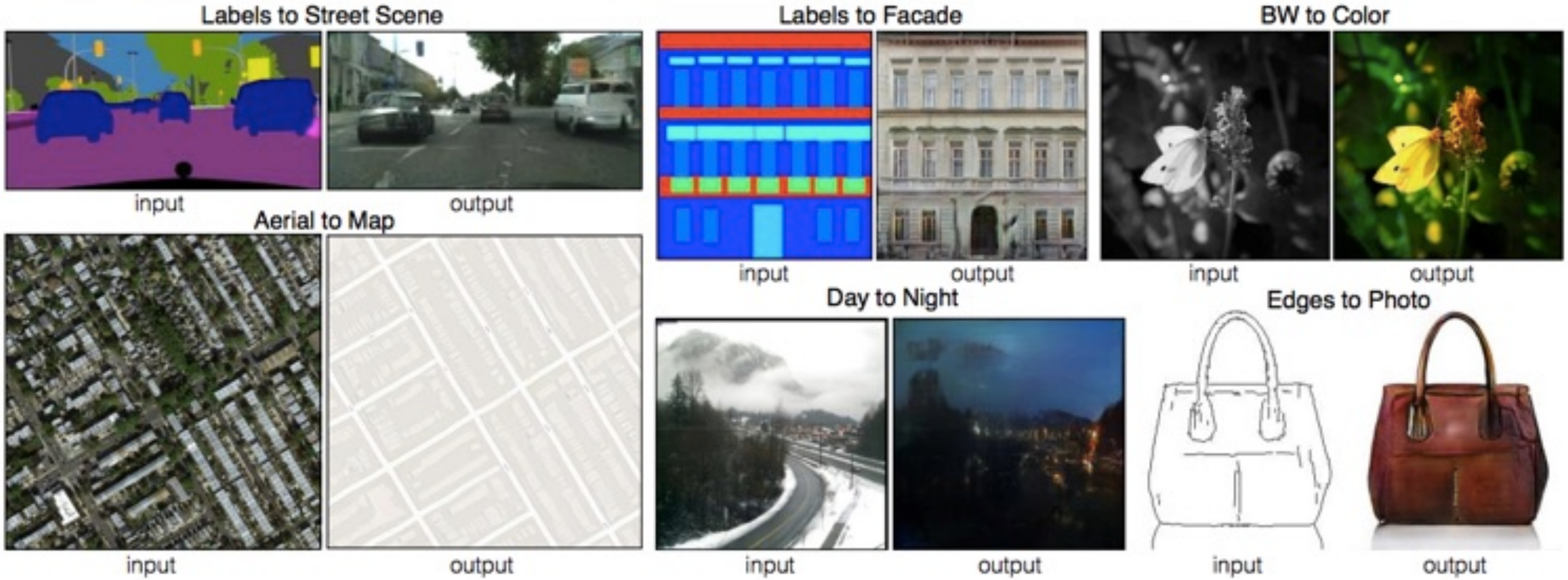
P. Isola, J.-Y. Zhu, T. Zhou, A. Efros, Image-to-Image Translation with Conditional Adversarial Networks, CVPR 2017

J.-Y. Zhu, T. Park, P. Isola, A. Efros, Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, ICCV 2017

…

Paired image-to-image translation

- Deterministic



P. Isola, J.-Y. Zhu, T. Zhou, A. Efros, [Image-to-Image Translation with Conditional Adversarial Networks](), CVPR 2017

Image-to-image translation

- Produce modified image y conditioned on input image x (note change of notation)
- Generator receives x as input
- Discriminator receives an $x, \ y$ pair and has to decide whether it is real or fake

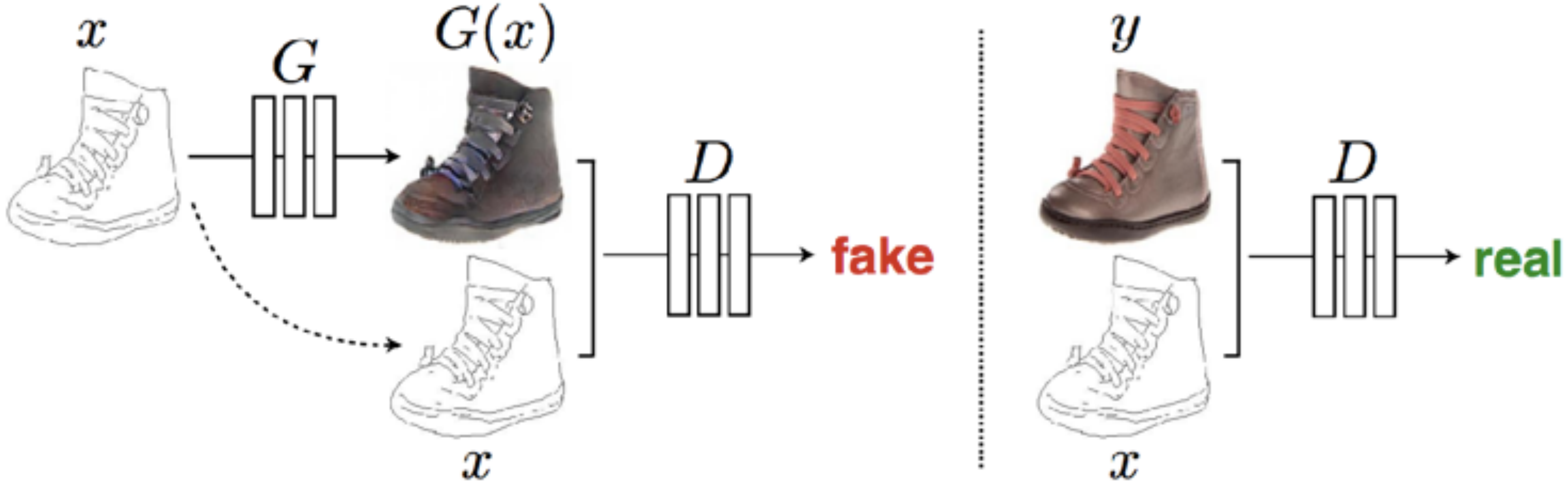
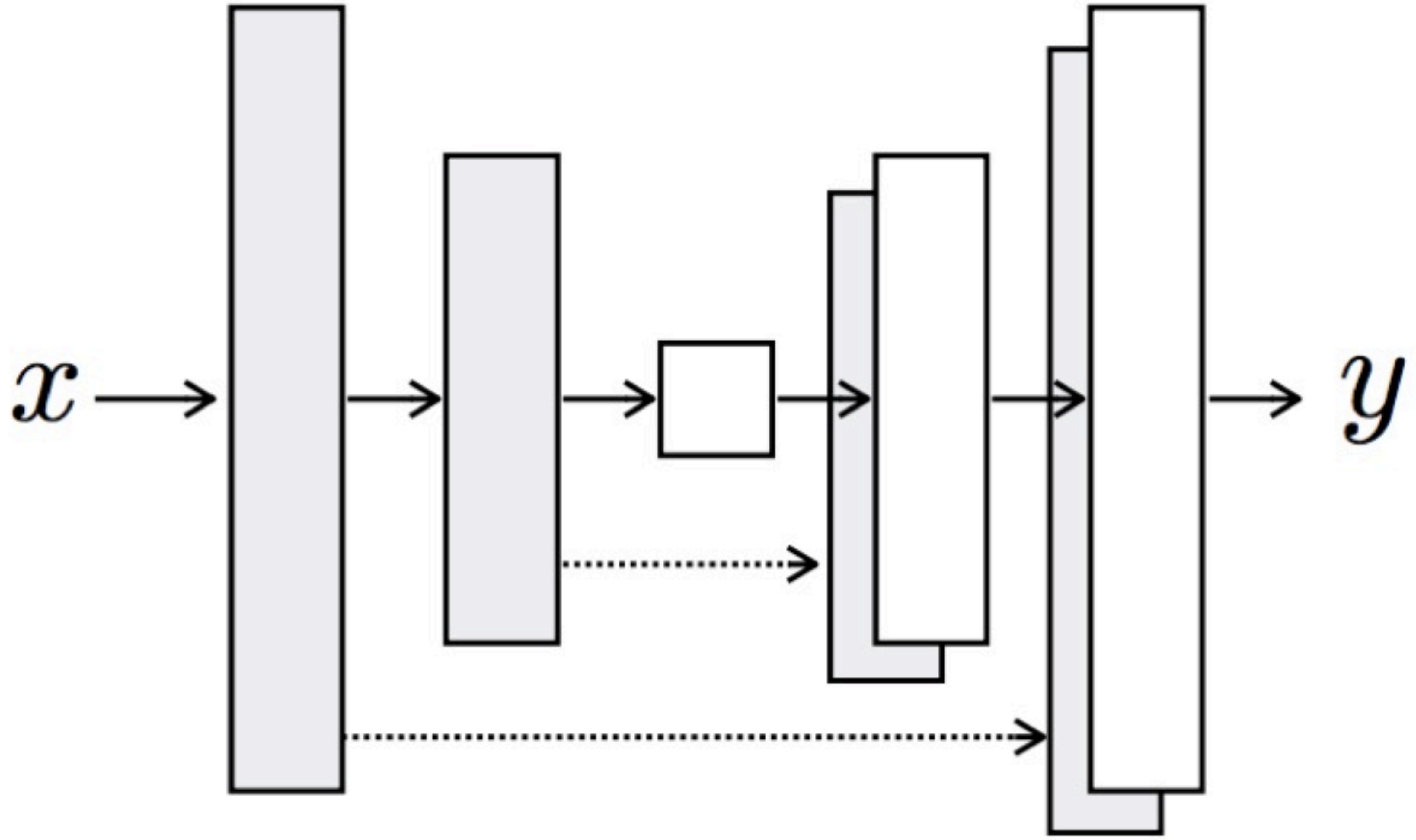


Image-to-image translation

- Generator architecture: U-Net



- Note: no z used as input, transformation is basically deterministic

Image-to-image translation

- Generator architecture: U-Net

Effect of adding skip connections to the generator

without
skip connections

with
skip connections

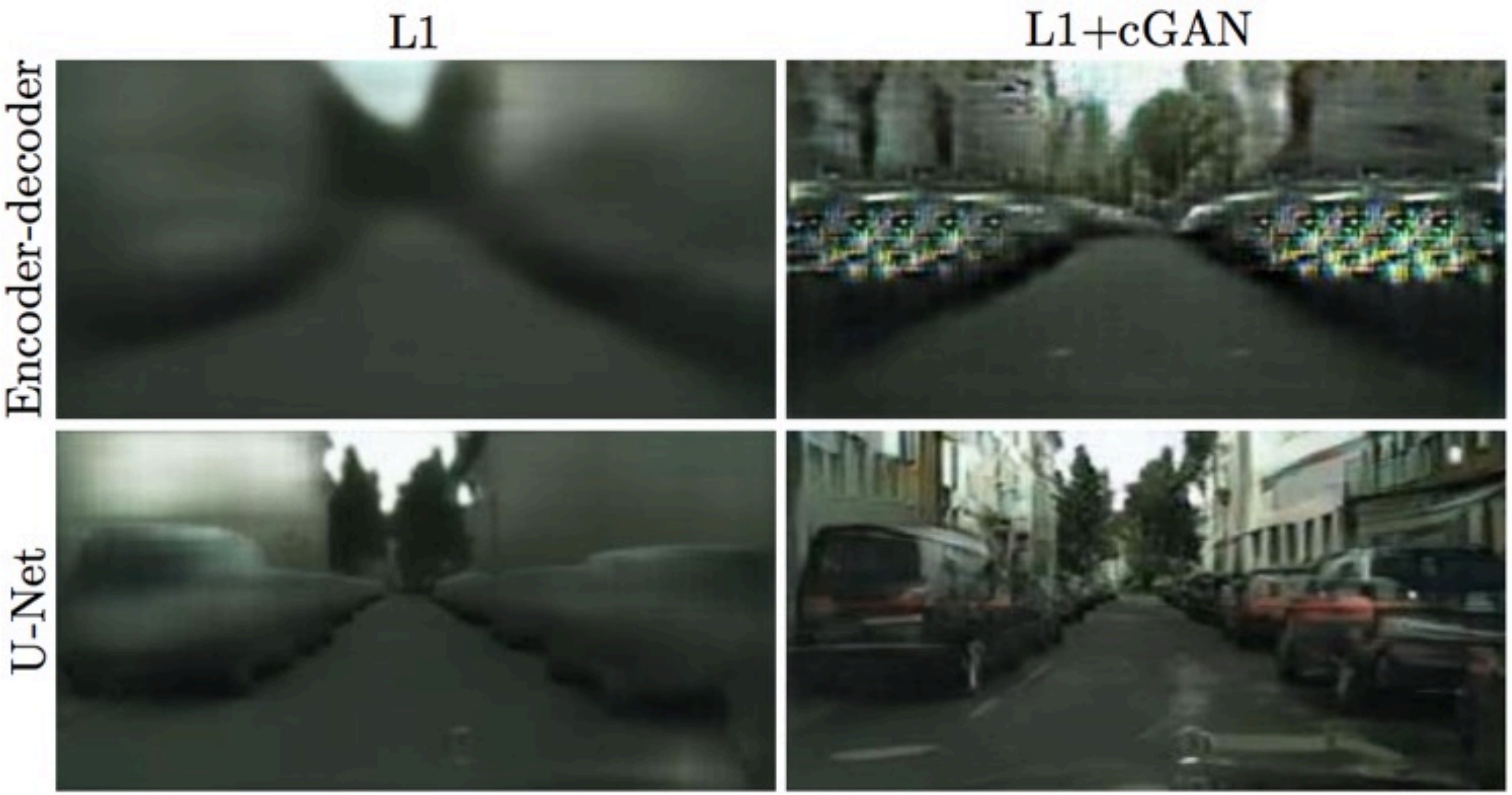


Image-to-image translation

- Generator loss: GAN loss plus L1 reconstruction penalty

$$G^* = \text{argmin}_G \max_D \mathscr{L}_{GAN}(G, D) + \lambda \sum_i \|y_i - G(x_i)\|_1$$

Image-to-image translation: Results

- Translating between maps and aerial photos

Image-to-image translation: Results

- Semantic labels to scenes

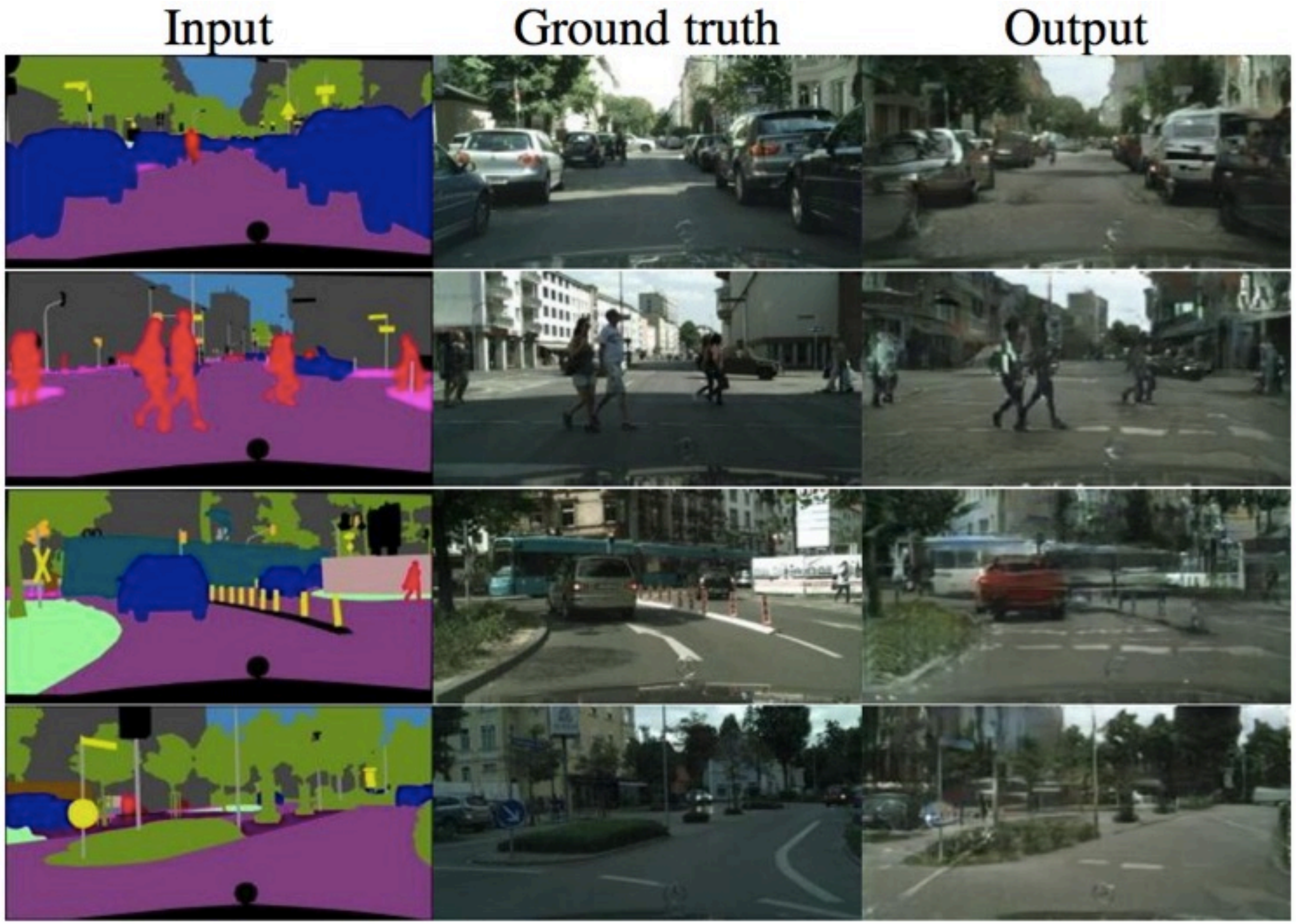


Image-to-image translation: Results

- Scenes to semantic labels

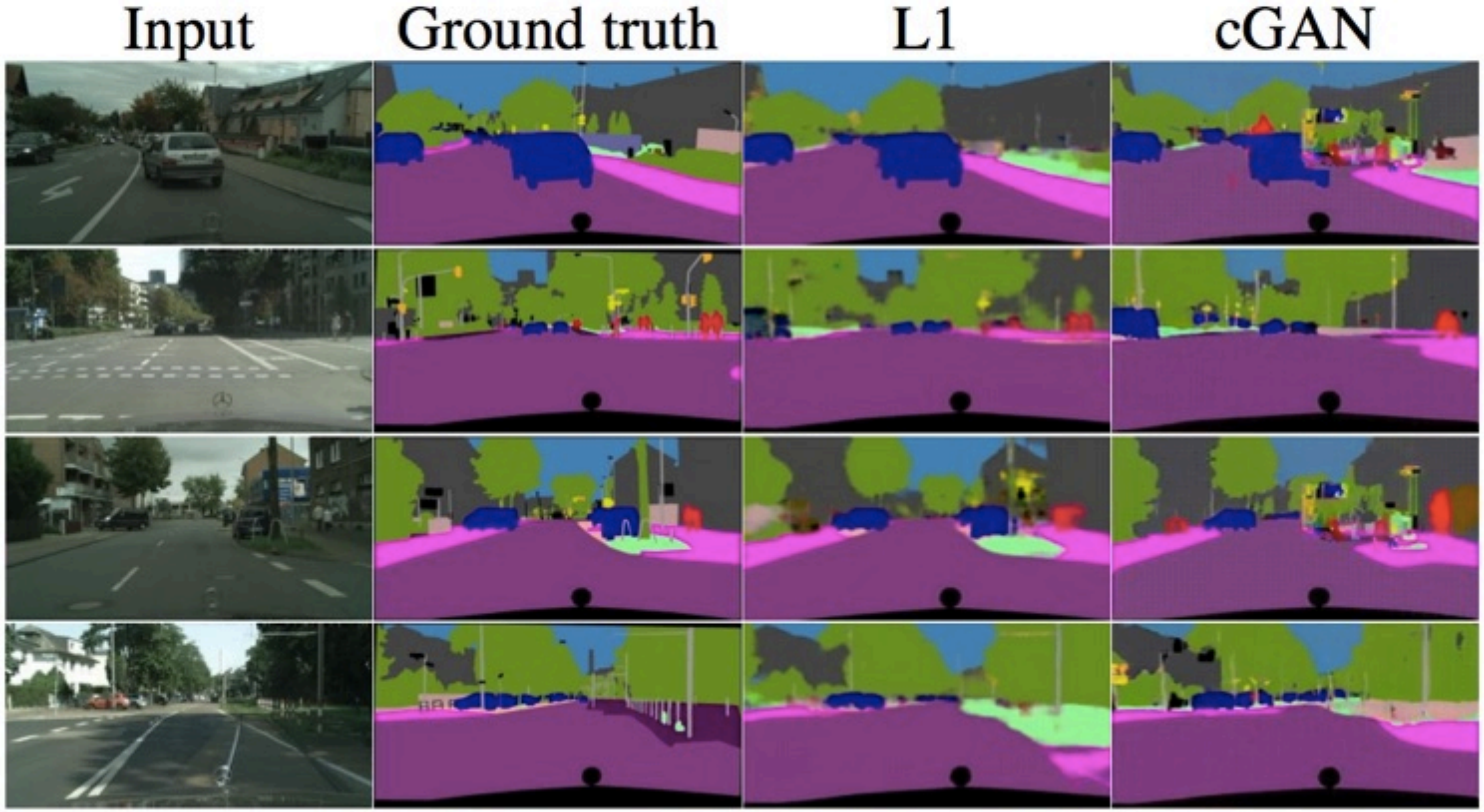


Image-to-image translation: Results

- Semantic labels to facades

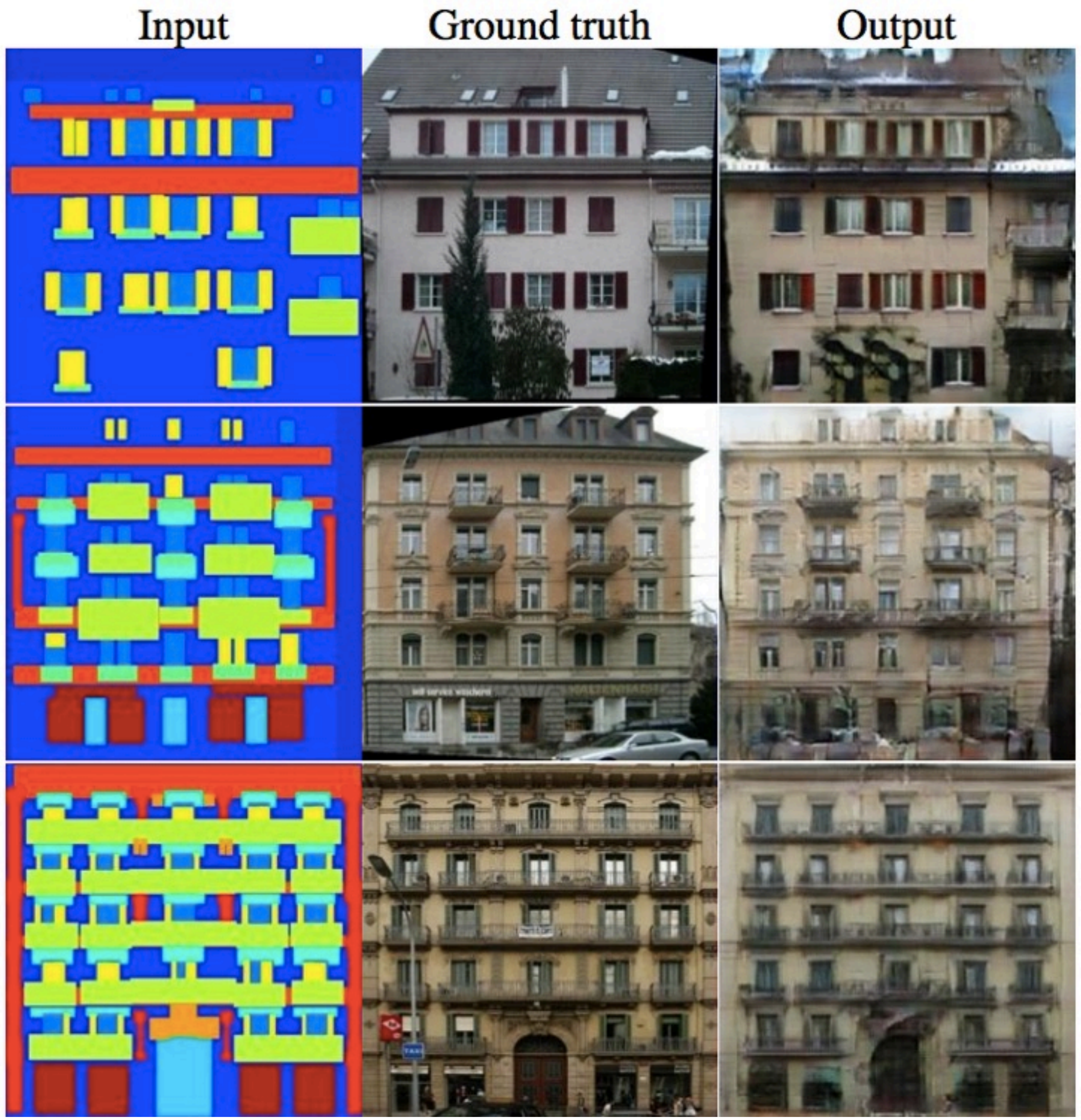


Image-to-image translation: Results

- Day to night

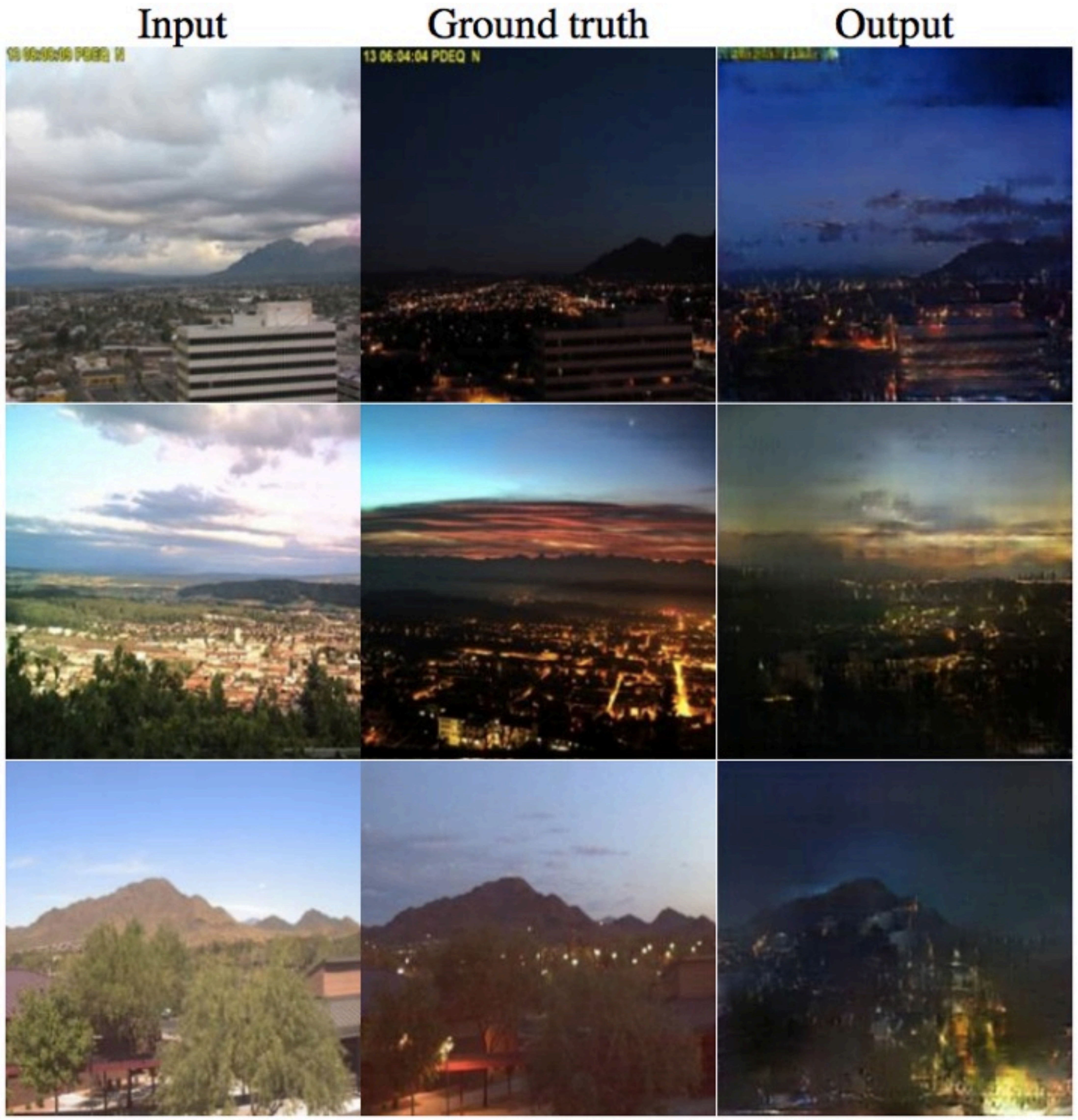


Image-to-image translation: Results

- Edges to photos

Image-to-image translation: Results

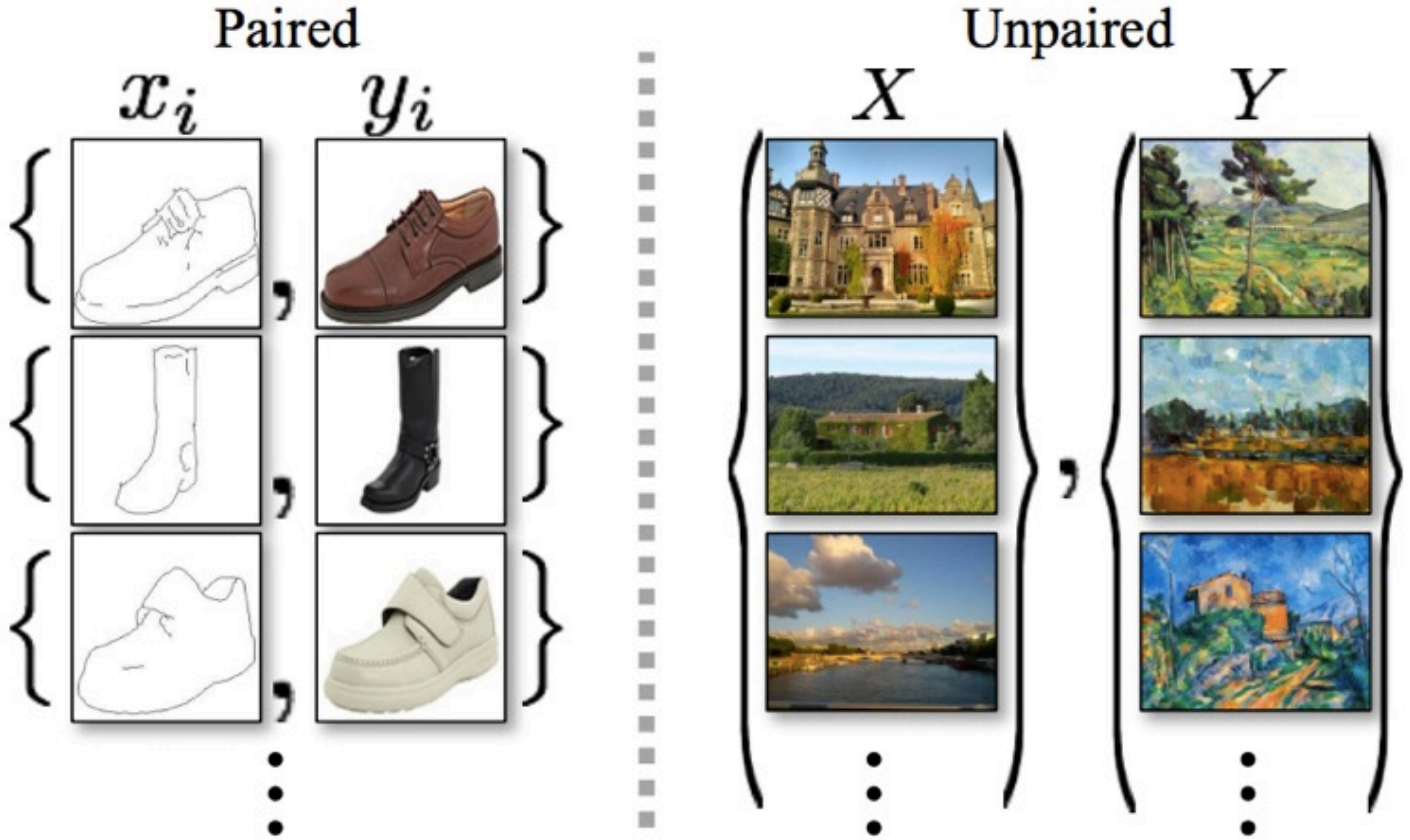
- [pix2pix demo]

Pix2pix: Limitations

- Visual quality could be improved
- Requires $x, \ y$ pairs for training
- Does not model conditional distribution $P(y \mid x)$, returns a single mode instead

Unpaired image-to-image translation

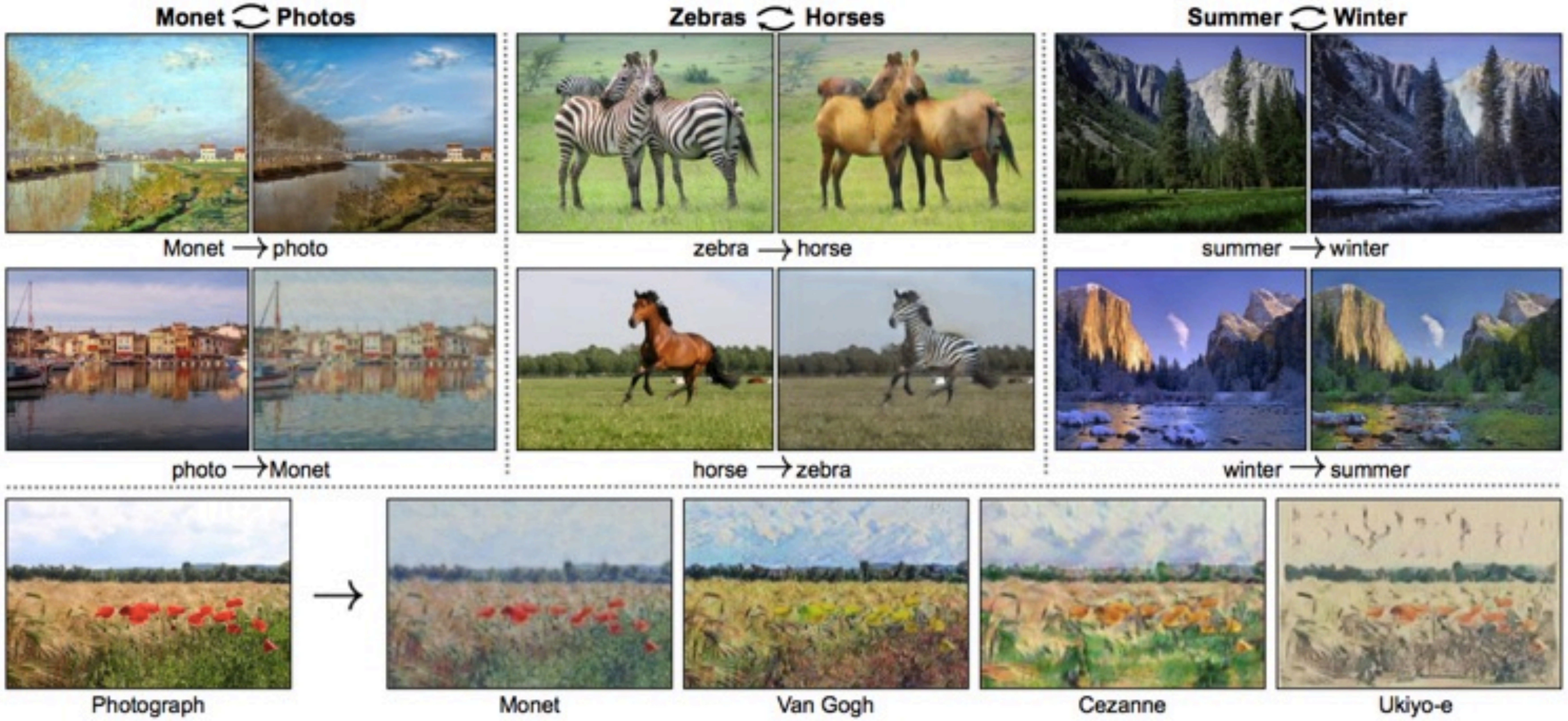
- Given two unordered image collections X and Y, learn to "translate" an image from one into the other and vice versa



J.-Y. Zhu, T. Park, P. Isola, A. Efros, Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, ICCV 2017

Unpaired image-to-image translation

- Given two unordered image collections X and Y, learn to "translate" an image from one into the other and vice versa

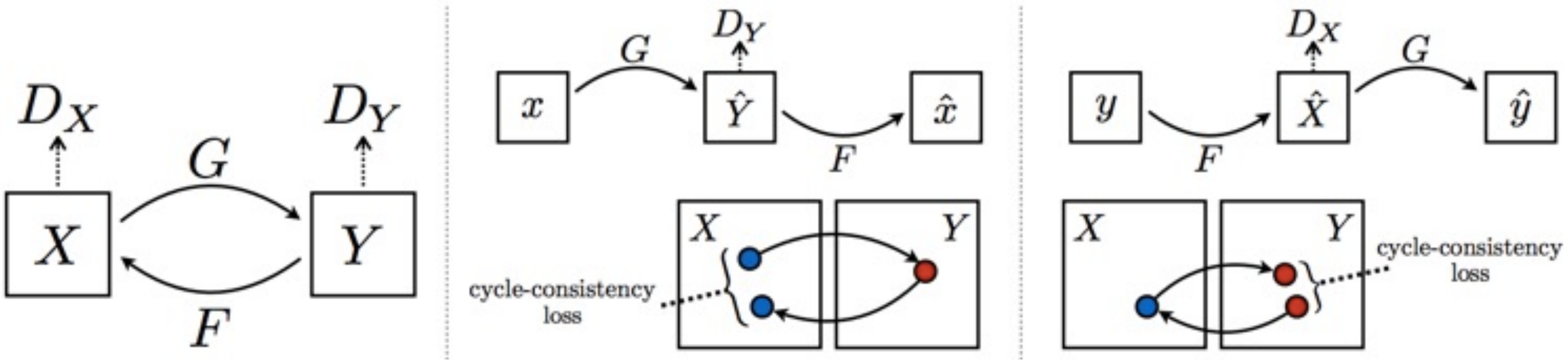


J.-Y. Zhu, T. Park, P. Isola, A. Efros, [Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks](), ICCV 2017

CycleGAN

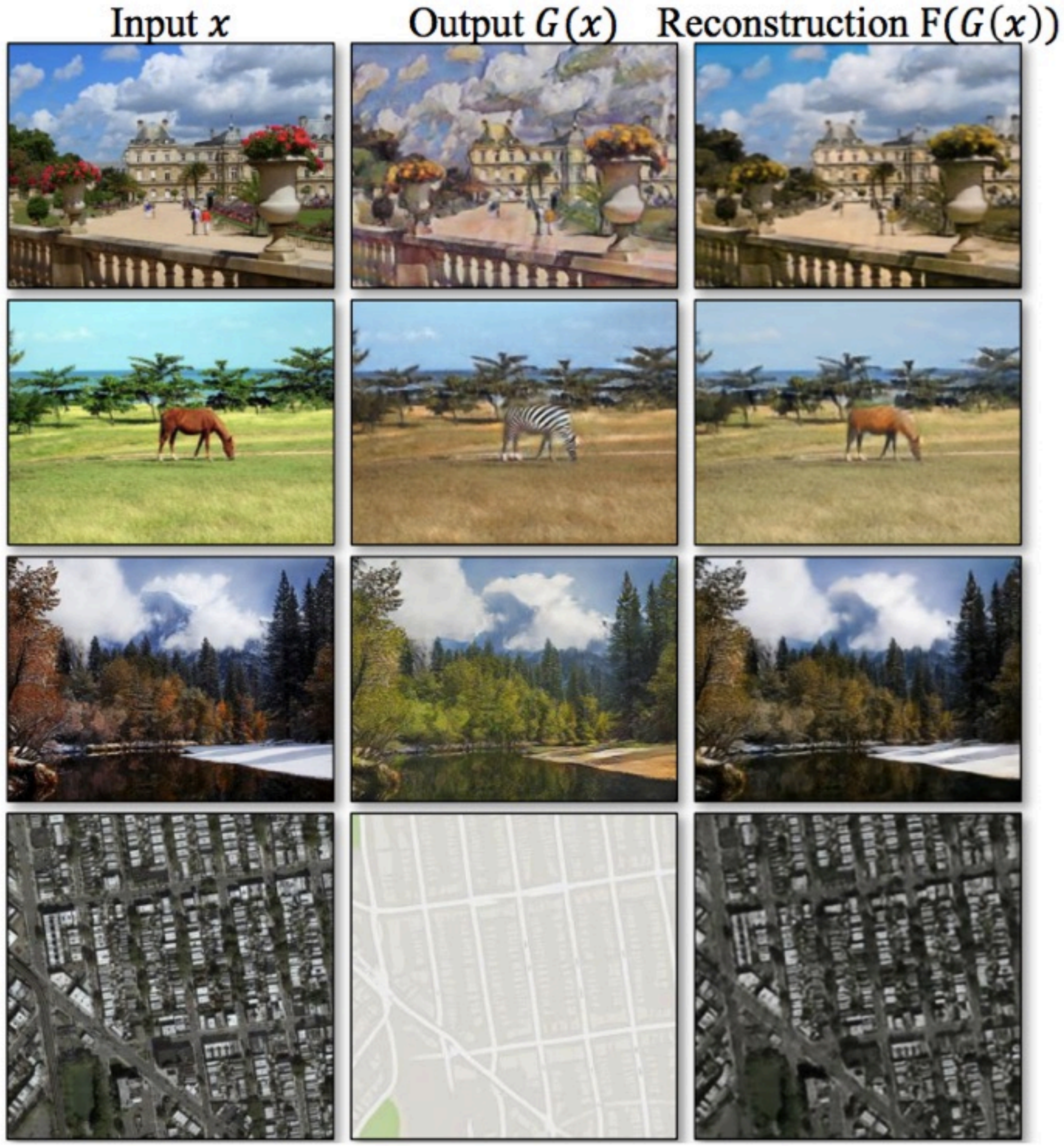
- Given: domains X and Y

- Train two generators F and G and two discriminators D_X and D_Y
  - G translates from X to Y, F translates from Y to X
  - D_X recognizes images from X, D_Y from Y
  - *Cycle consistency*: we want $F(G(x)) \approx x$ and $G(F(y)) \approx y$



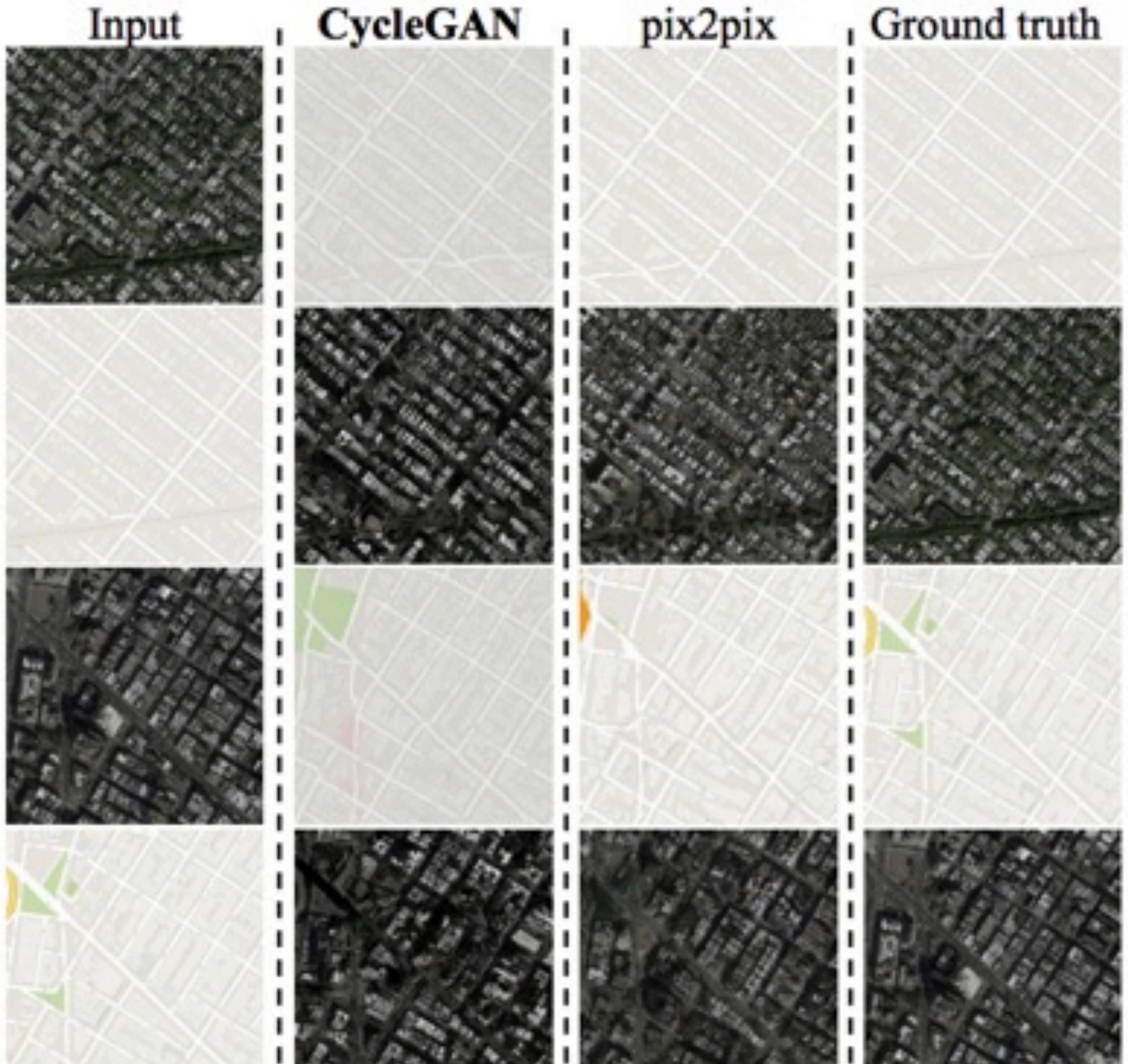
CycleGAN

- Illustration of cycle consistency:



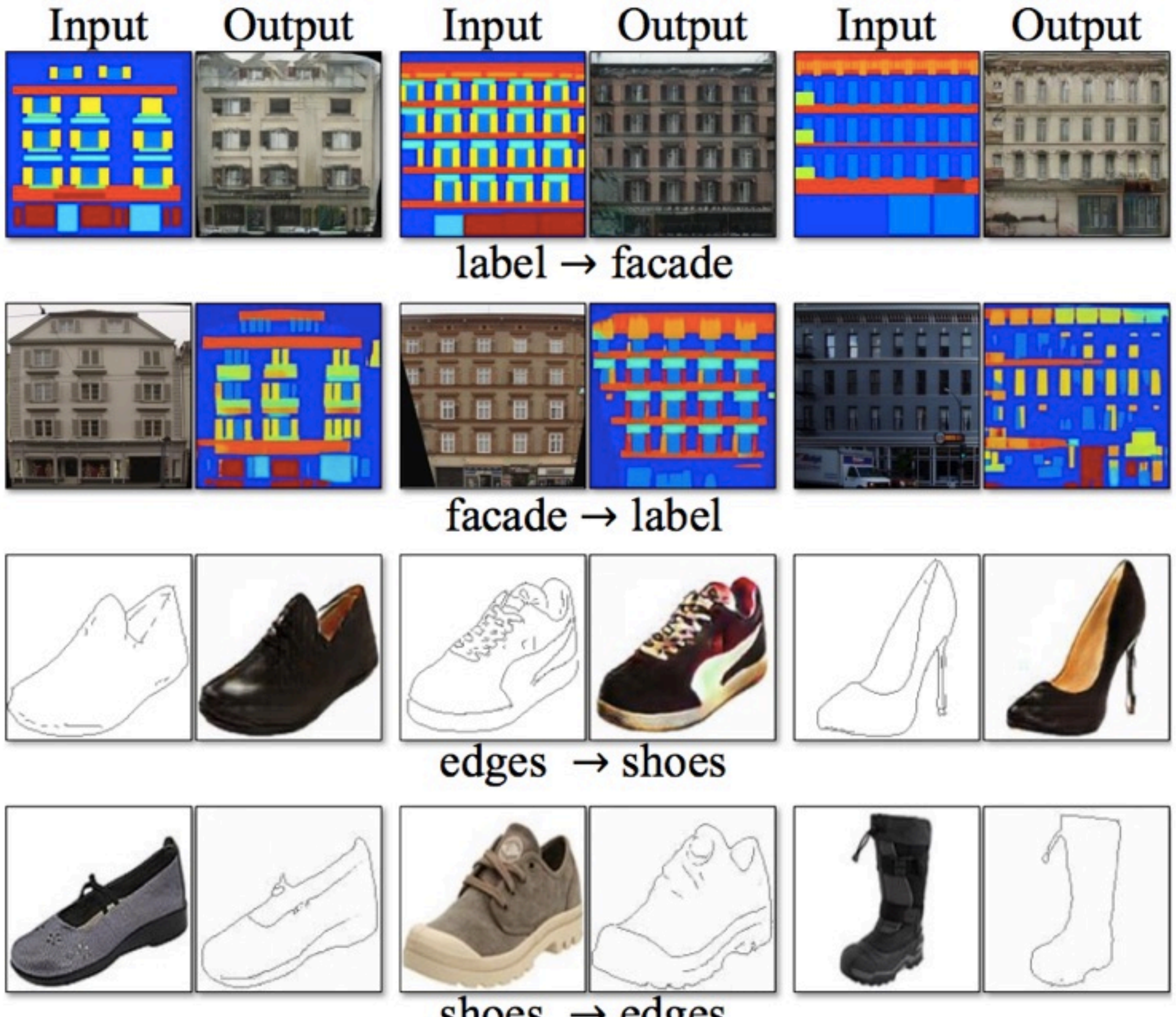
CycleGAN: Results

- Translation between maps and aerial photos



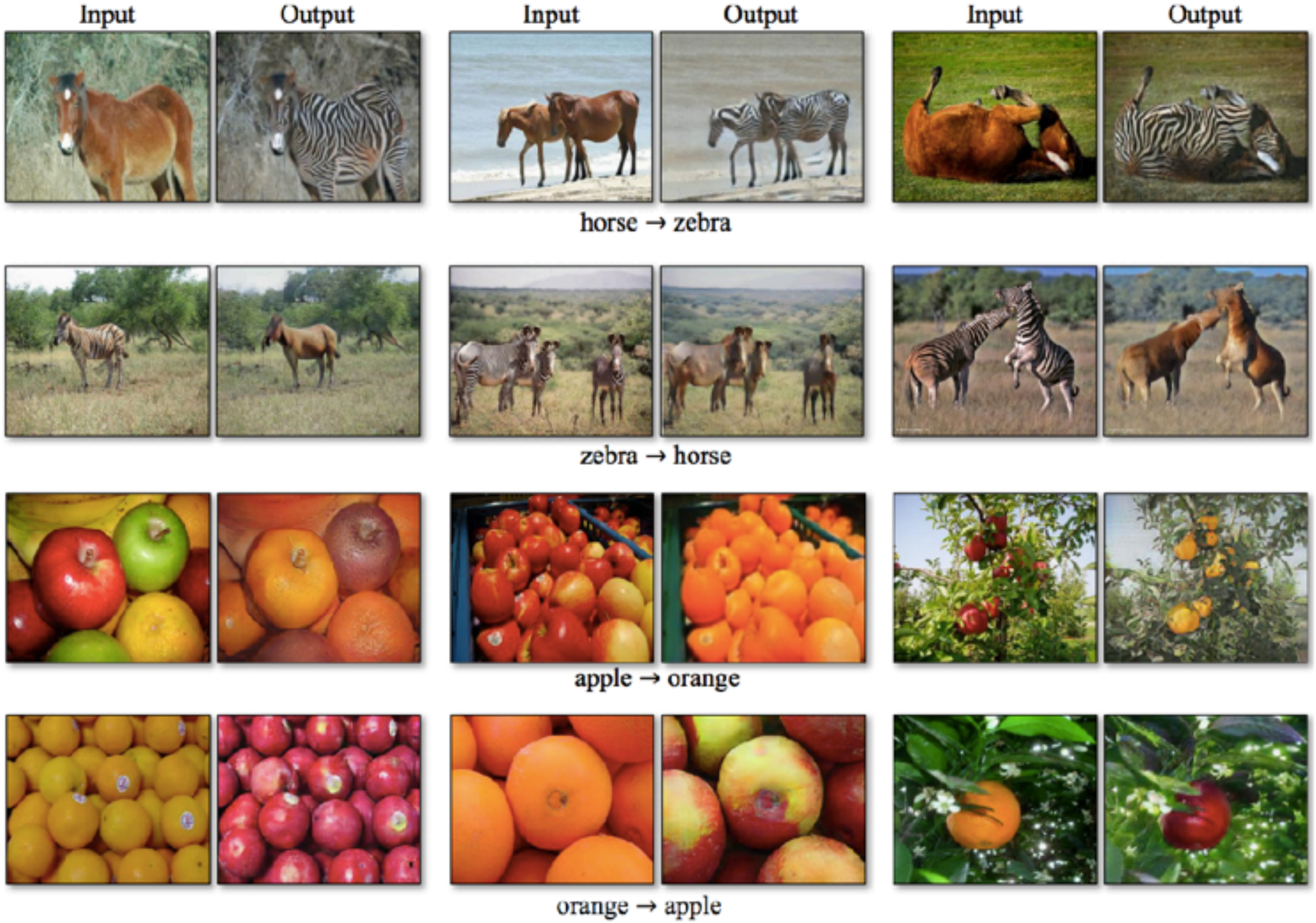
CycleGAN: Results

- Other pix2pix tasks



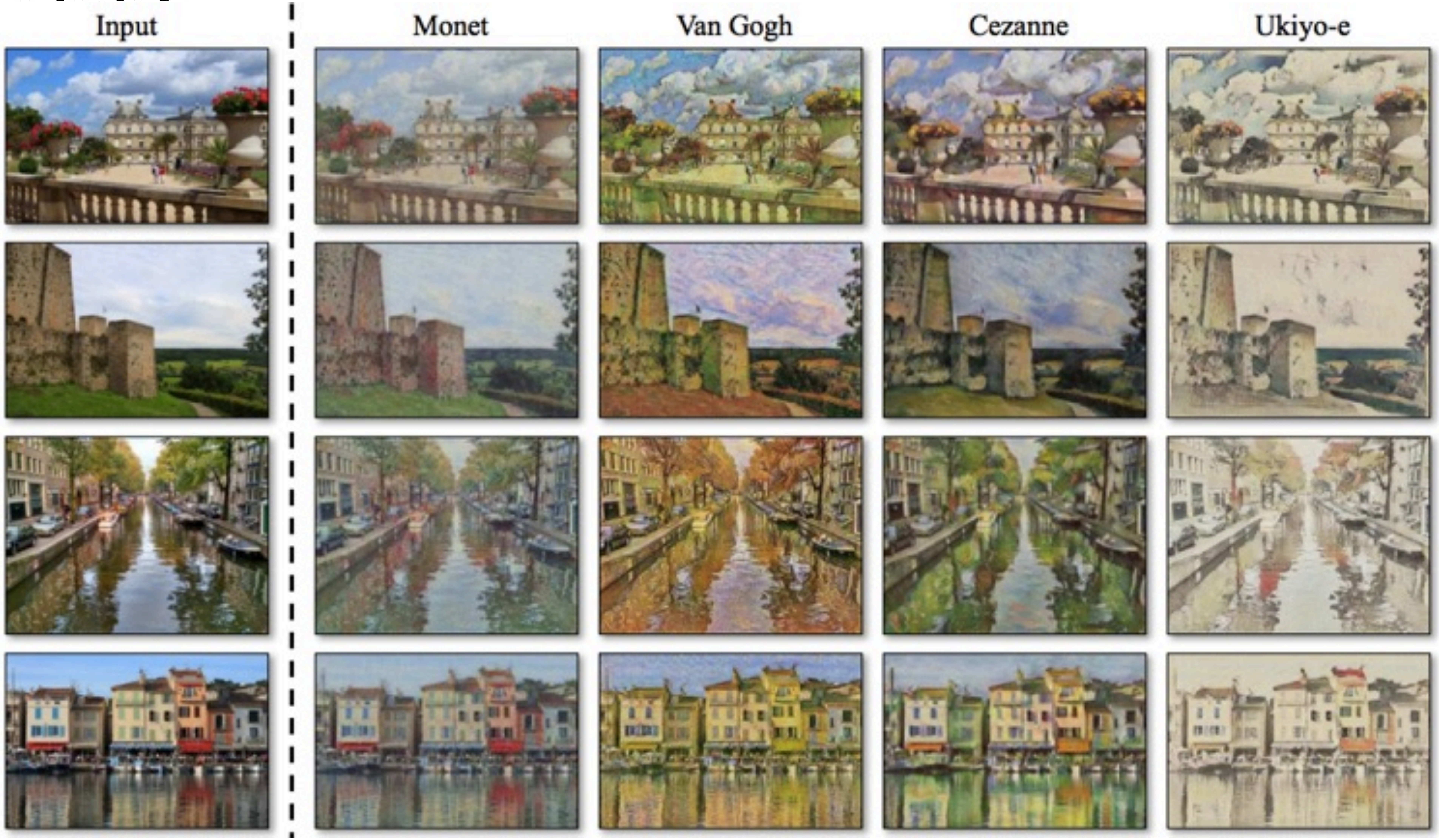
CycleGAN: Results

- Tasks for which paired data is unavailable

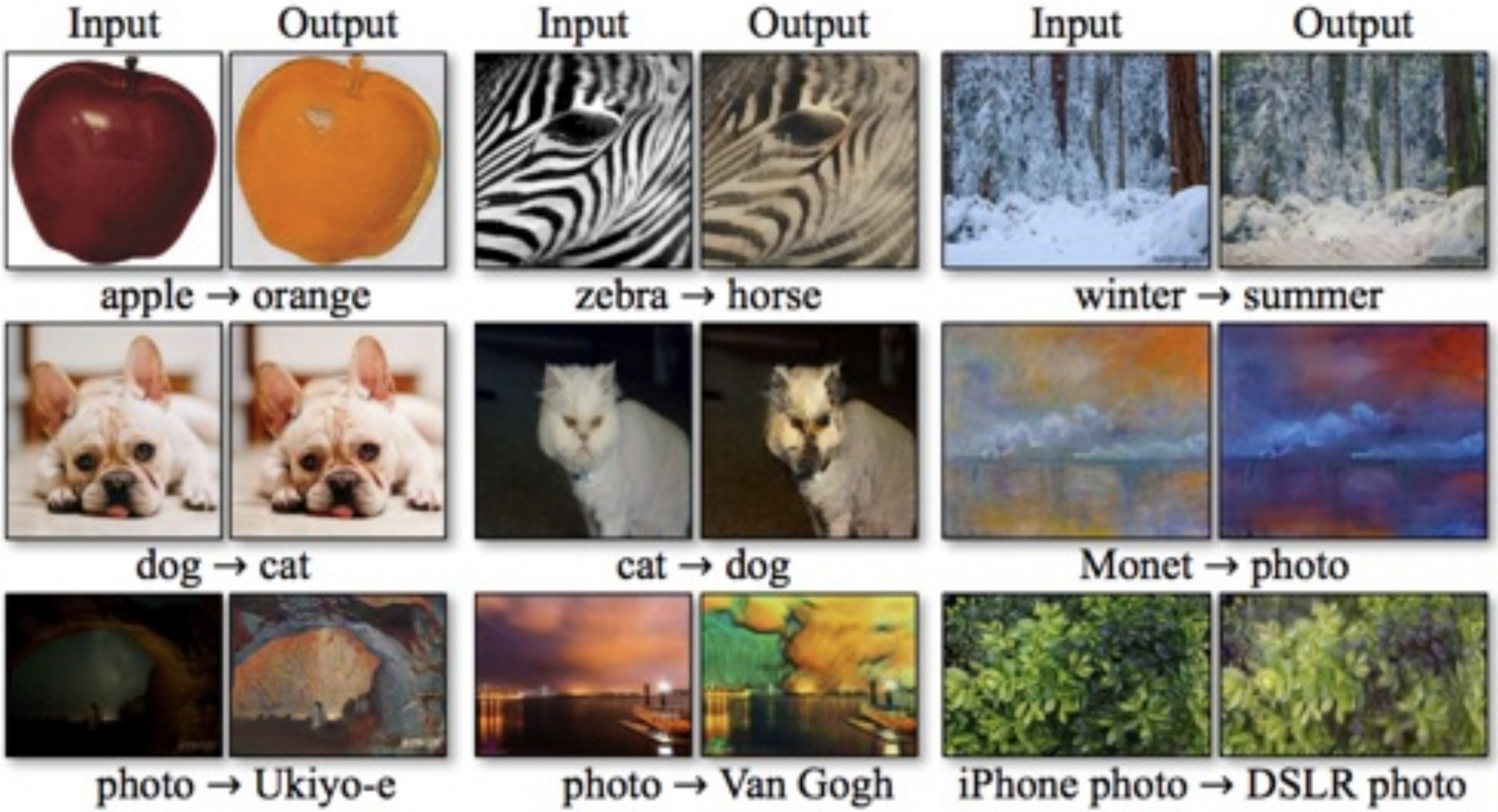


CycleGAN: Results

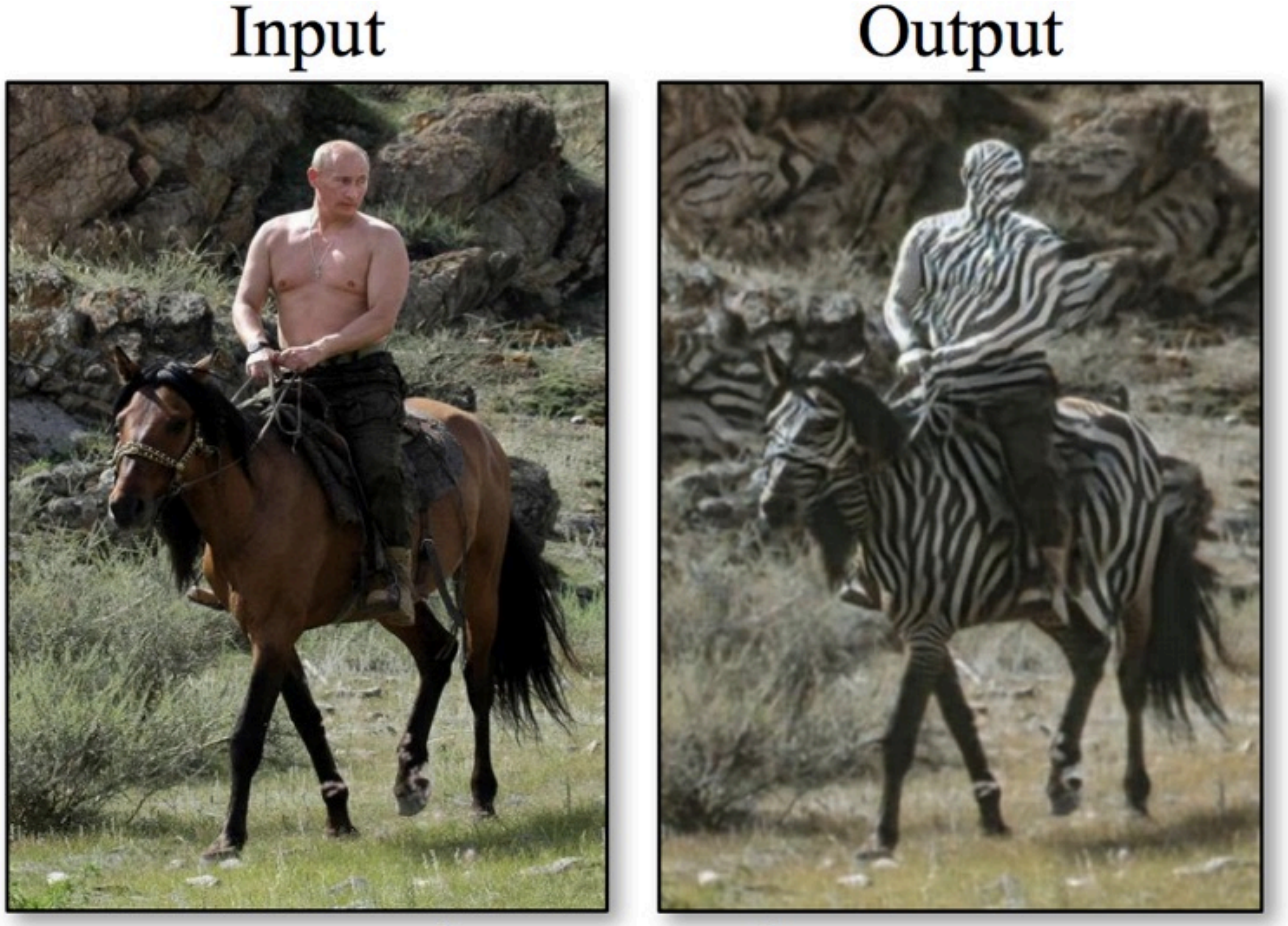
- Style transfer



CycleGAN: Failure cases



CycleGAN: Failure cases

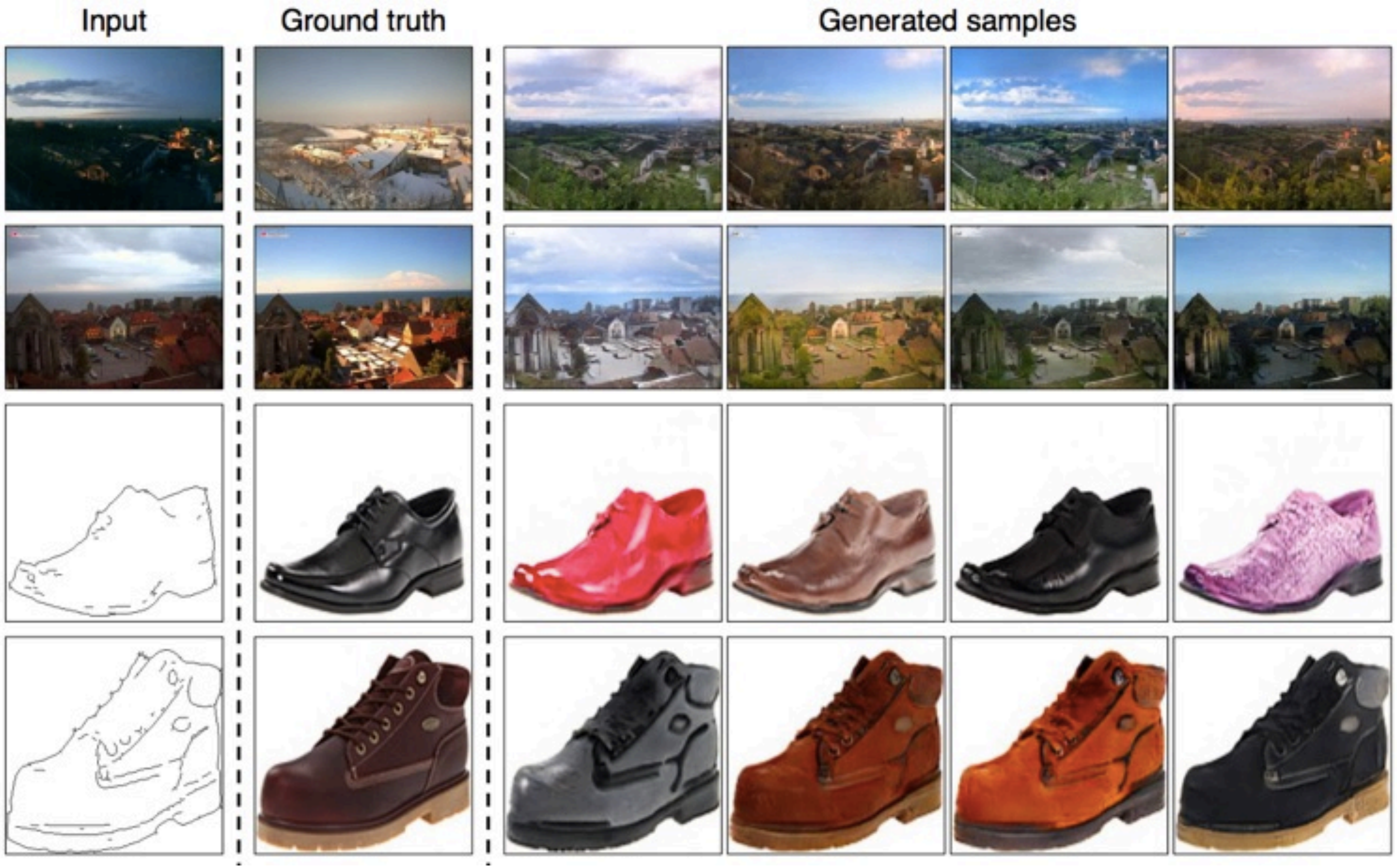


horse → zebra

CycleGAN: Limitations

- Cannot handle shape changes (e.g., dog to cat)
- Can get confused on images outside of the training domains (e.g., horse with rider)
- Cannot close the gap with paired translation methods
- Does not account for the fact that one transformation direction may be more challenging than the other

Multimodal image-to-image translation



J.Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, E. Shechtman, Toward Multimodal Image-to-Image Translation, NIPS 2017

Human generation conditioned on pose

https://www.youtube.com/watch?v=PCBTZh41Ris

C. Chan, S. Ginosar, T. Zhou, A. Efros. Everybody Dance Now. ICCV 2019 https://carolineec.github.io/everybody_dance_now/

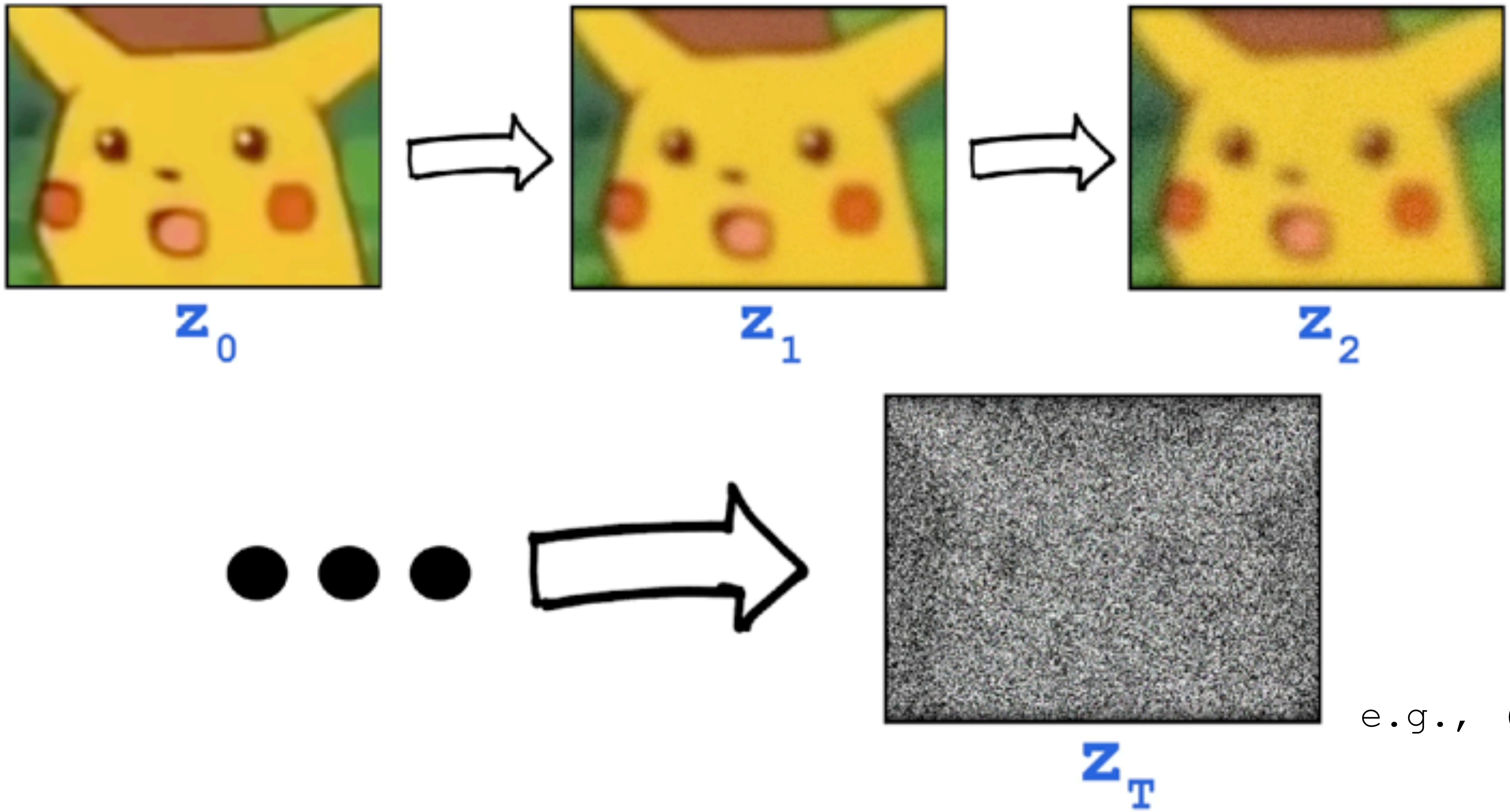
Agenda

1. **Generative neural networks**
    - VAE: Variational autoencoders
    - GAN: Generative adversarial networks
    - Diffusion models

2. **Vision & language**
    - Text-to-image retrieval
    - Text-to-image generation
    - Image captioning

Diffusion models: Forward (diffusion) process

"easy to convert structured data into noise"*



e.g., (T ~ 1000)

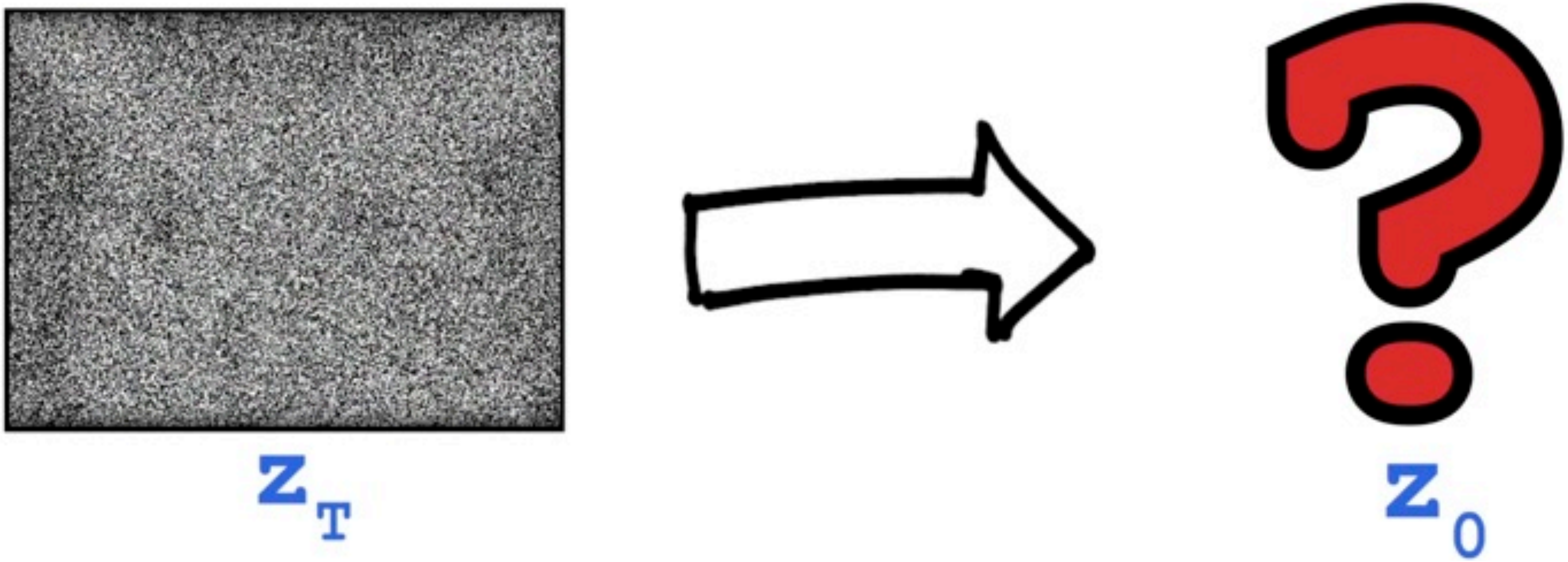
After enough steps, we have $z_T \sim N(0,1)$

*[Murphy 2023]

Edan Meyer: [Stable Diffusion - What, Why, How?]()

Diffusion models

"hard to convert noise into structured data"*

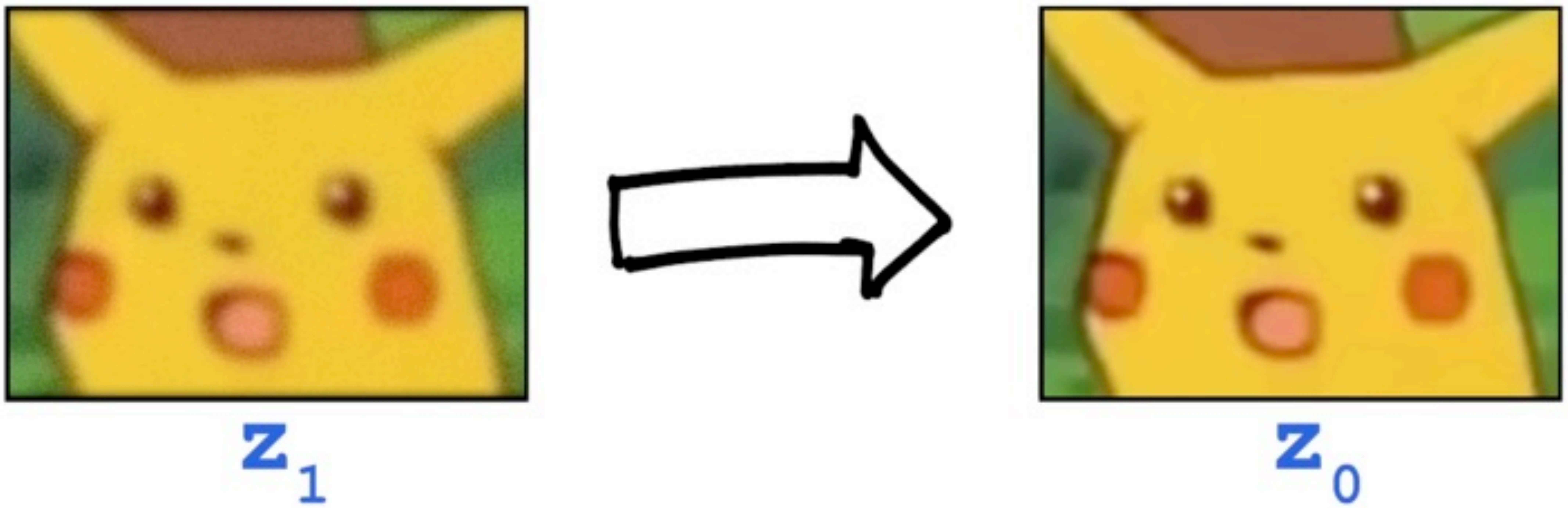


*[Murphy 2023]

Edan Meyer: [Stable Diffusion - What, Why, How?]

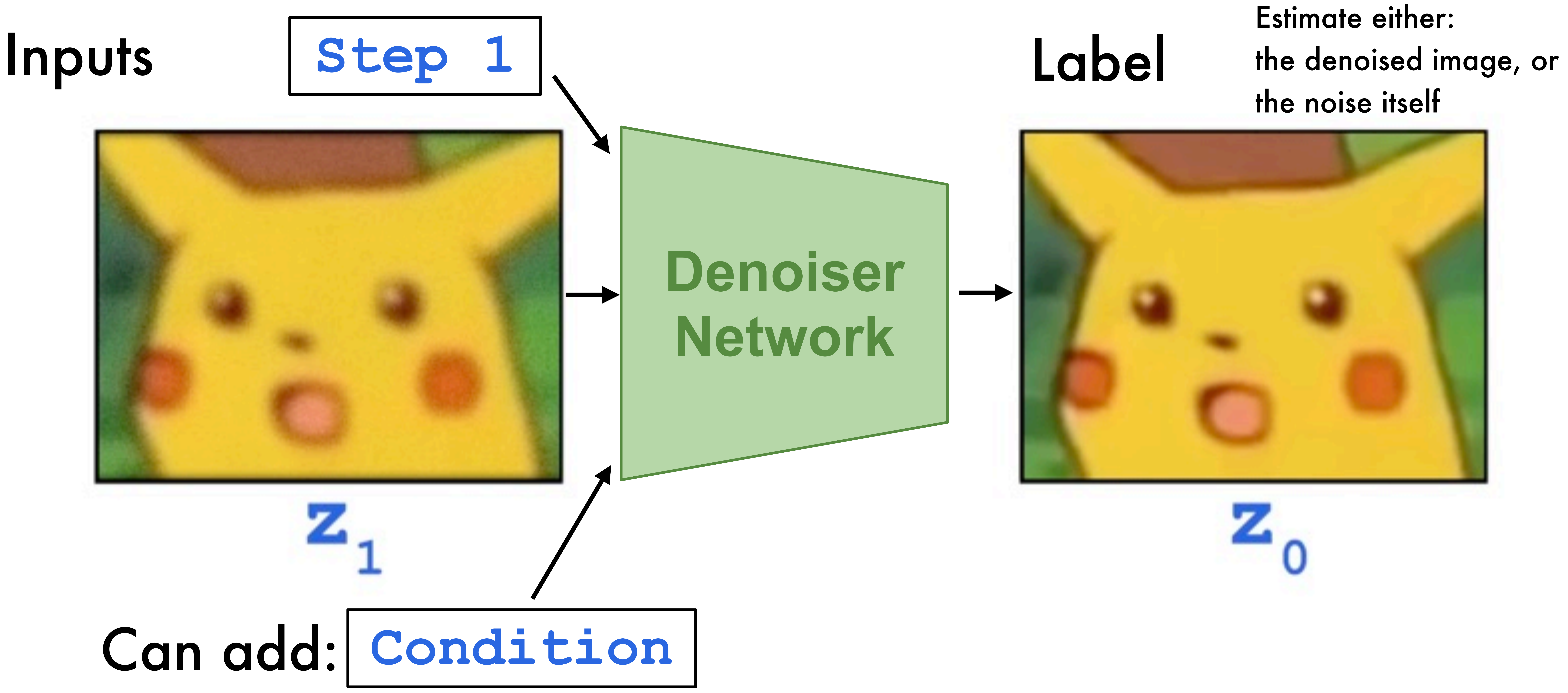
Diffusion models

Reverse (denoising) process

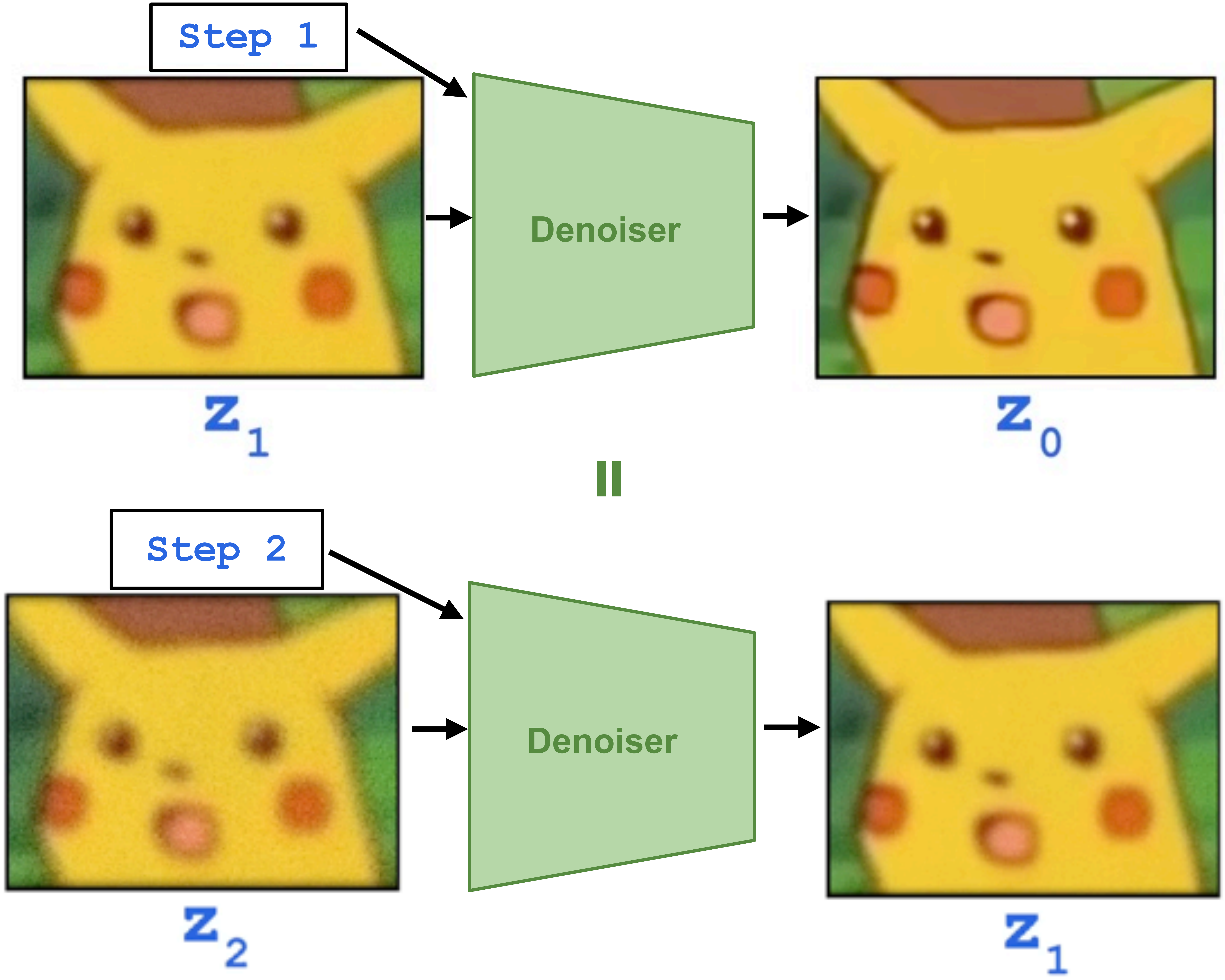


Edan Meyer: [Stable Diffusion - What, Why, How?]

Diffusion models: Learning to denoise/reverse



Diffusion models: Learning to denoise/reverse



Diffusion models: Test time

Diffusion models

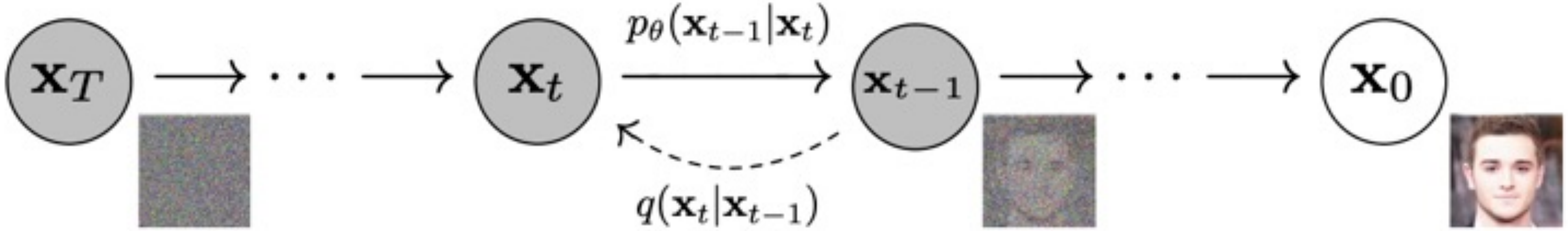


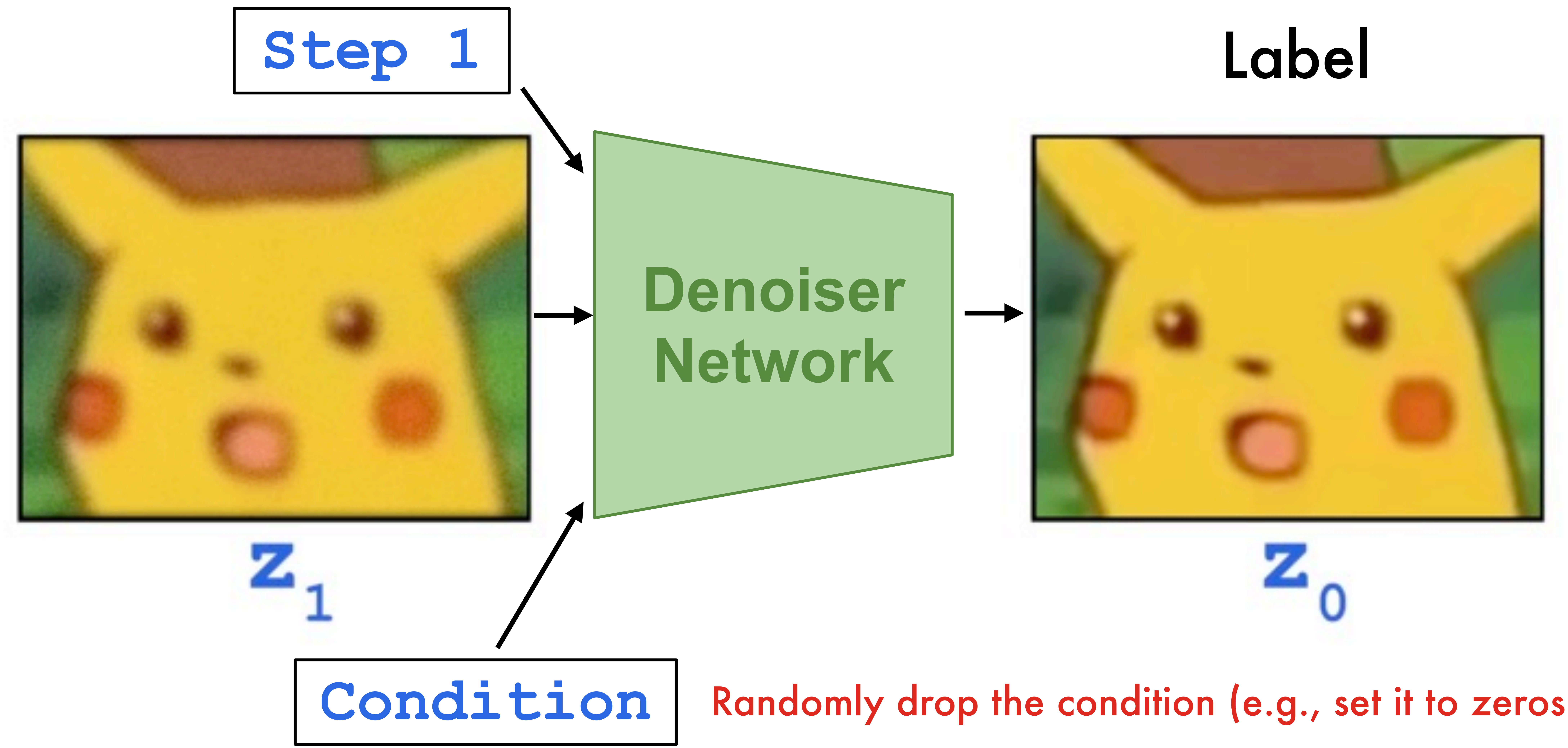
Figure 2: The directed graphical model considered in this work.

Unconditional CIFAR10 sample generation

"Noise schedule"?: linear, cosine etc

J. Ho et al. Denoising diffusion probabilistic models. NeurIPS 2020
Blog introduction: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Diffusion models: Conditioning



Randomly drop the condition (e.g., set it to zeros)!

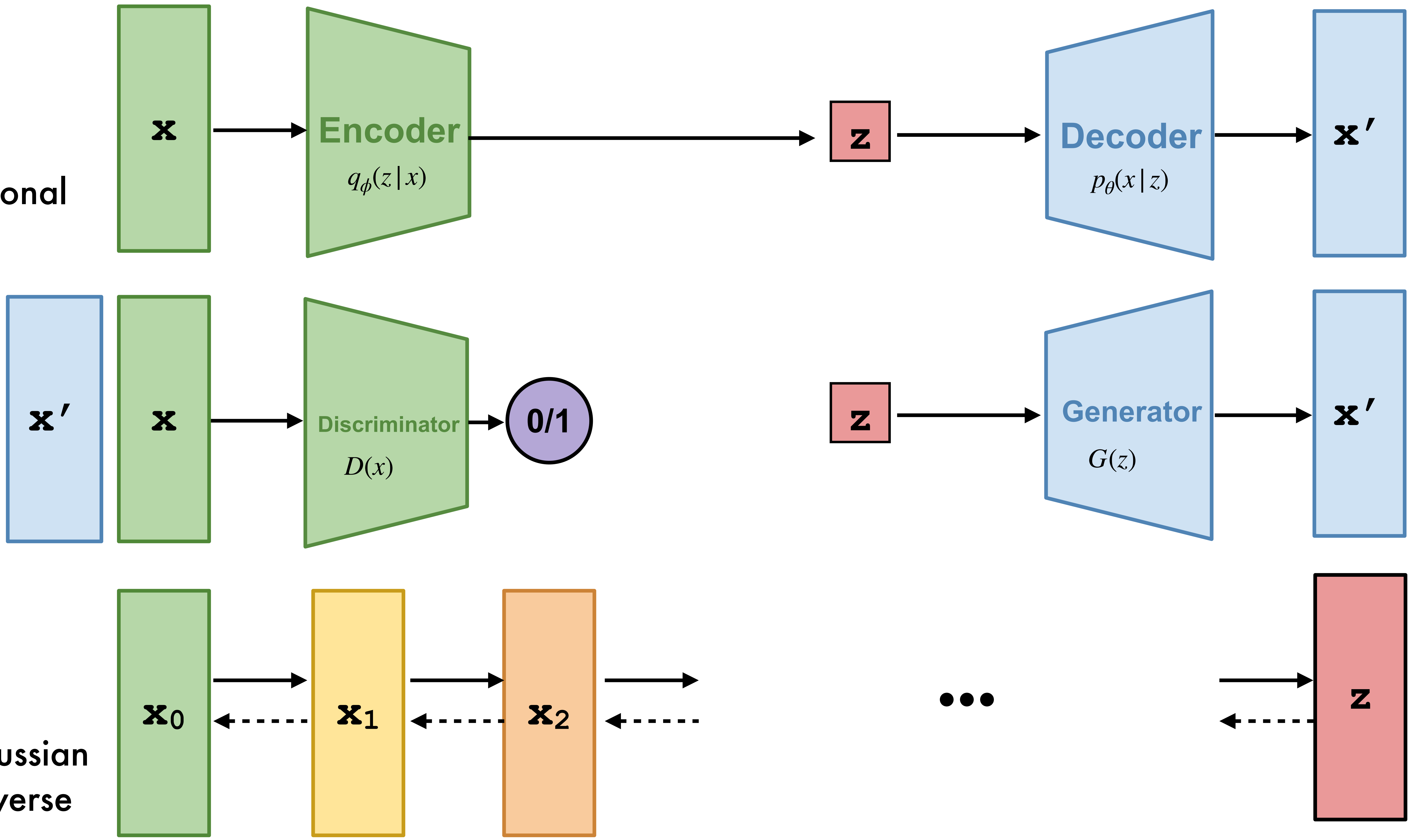
J. Ho and T. Salimans. **Classifier-free diffusion guidance**. In NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications, 2021.

Diffusion models vs GANs / VAEs

VAE
Maximize variational lower bound

GAN
Adversarial training

Diffusion
Gradually add Gaussian noise and then reverse



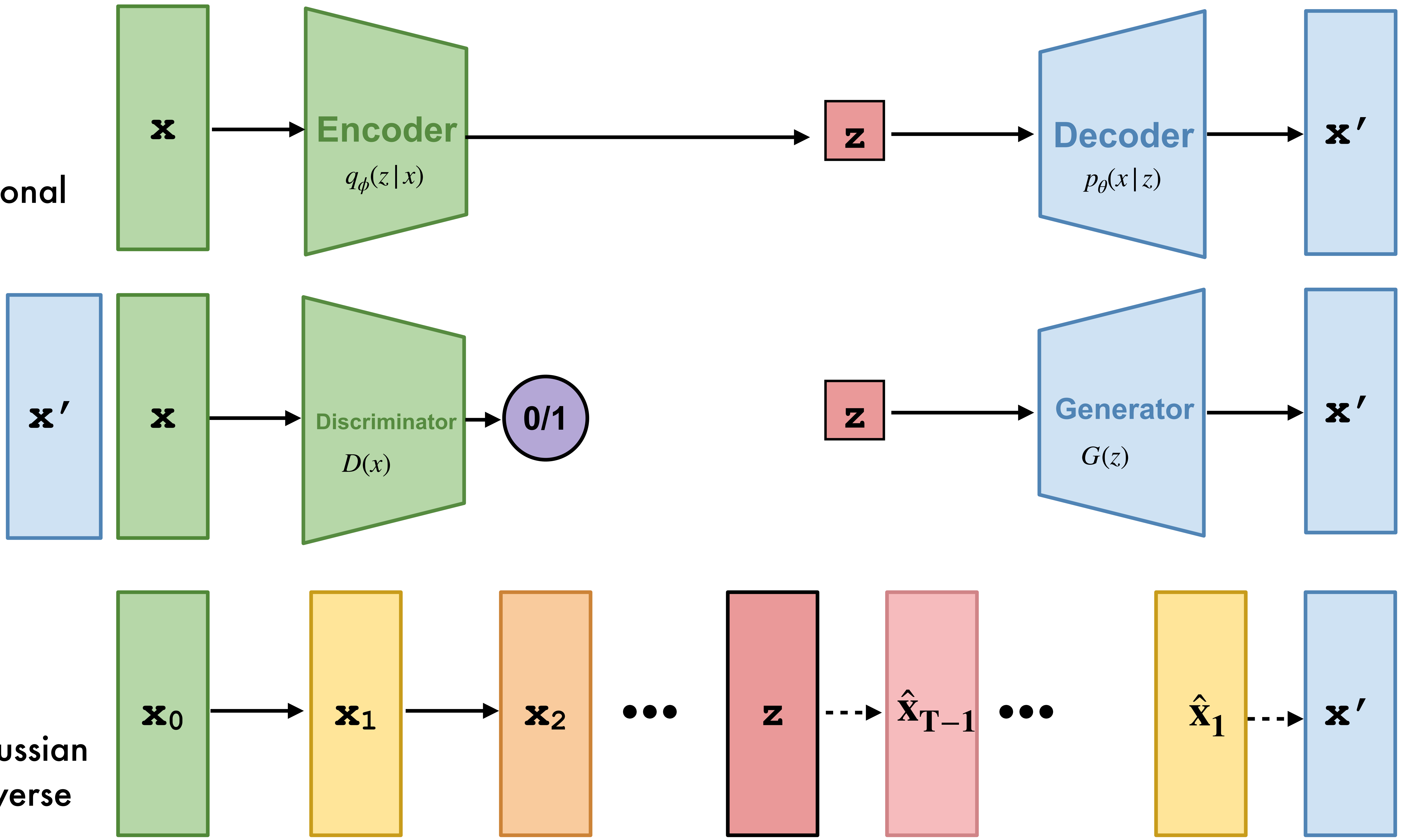
Source: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Diffusion models vs GANs / VAEs

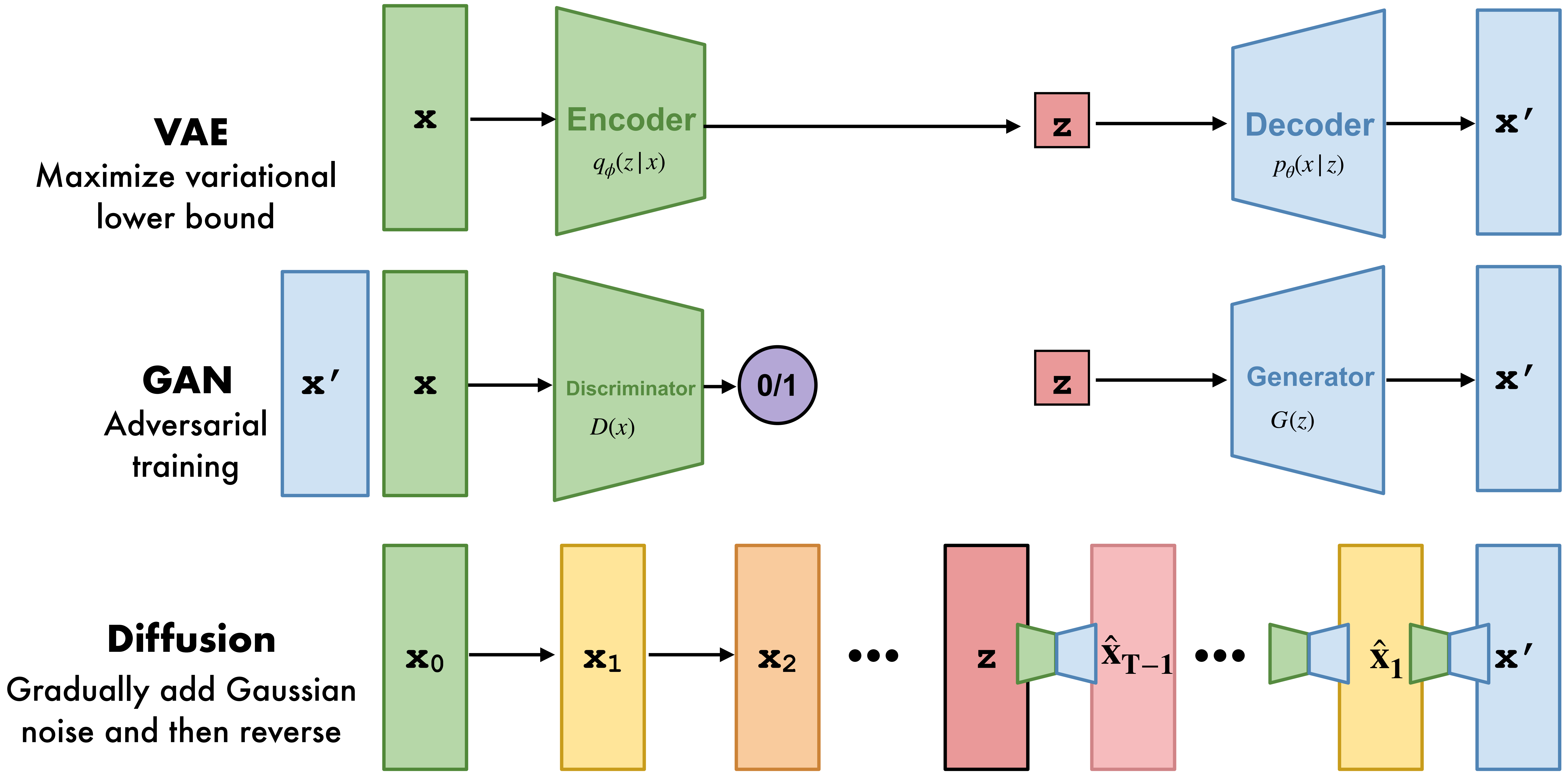
VAE
Maximize variational lower bound

GAN
Adversarial training

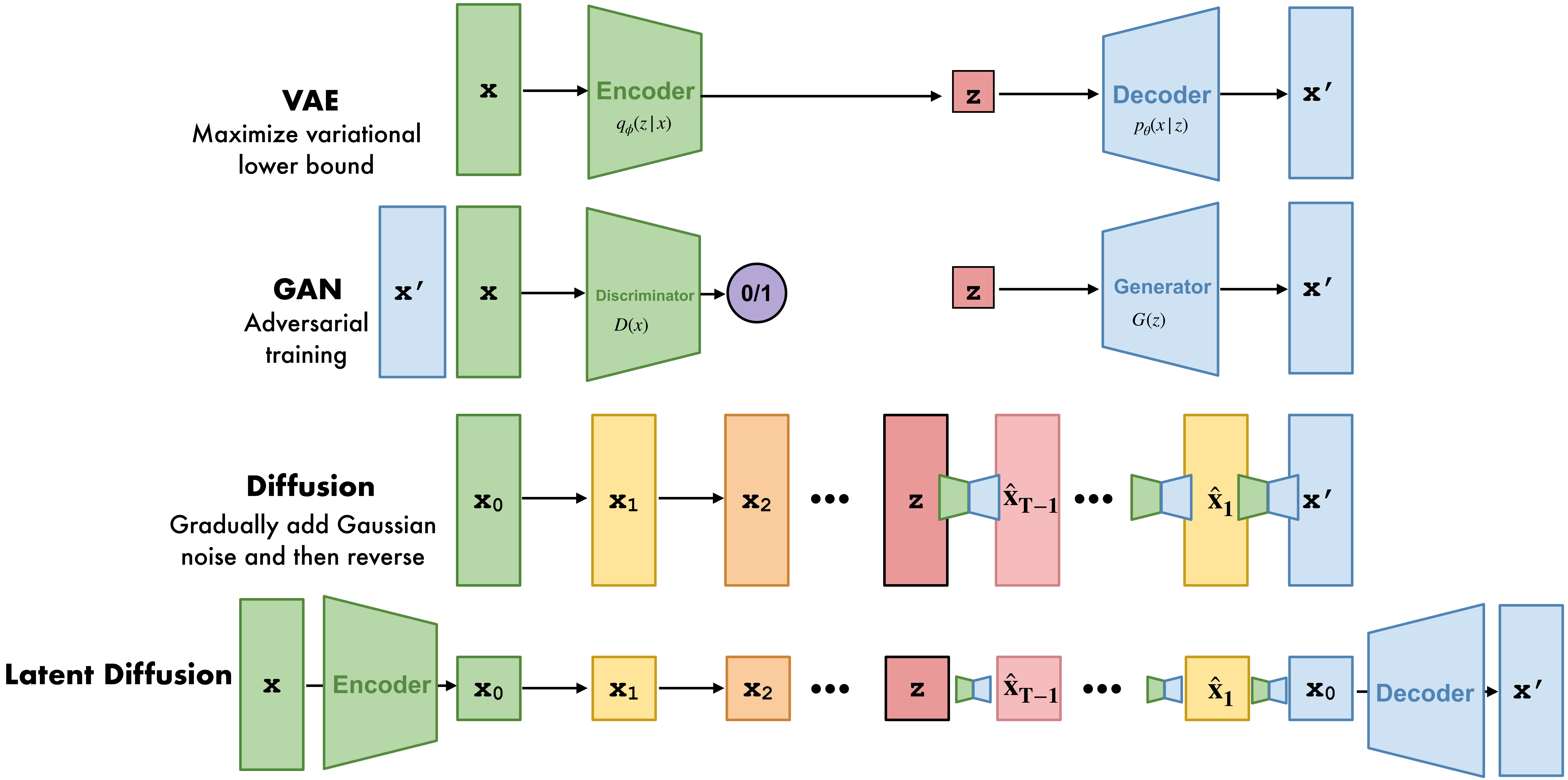
Diffusion
Gradually add Gaussian noise and then reverse



Diffusion models vs GANs / VAEs



Diffusion models: Latent diffusion



Trends (!)

23 May 2022

Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

Chitwan Saharia*, William Chan*, Saurabh Saxena†, Lala Li†, Jay Whang†, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho†, David J Fleet†, Mohammad Norouzi*

{sahariac,williamchan,mnorouzi}@google.com
{srbs,lala,jwhang,jonathanho,davidfleet}@google.com

Google Research, Brain Team
Toronto, Ontario, Canada

Imagen (Google)

13 Apr 2022

Hierarchical Text-Conditional Image Generation with CLIP Latents

Aditya Ramesh* OpenAI aramesh@openai.com
Prafulla Dhariwal* OpenAI prafulla@openai.com
Alex Nichol* OpenAI alex@openai.com
Casey Chu* OpenAI casey@openai.com
Mark Chen OpenAI mark@openai.com

Dall-E 2 (OpenAI)

Paid access with API, Sep. 2022

Dall-E 3 (OpenAI)

Paid access with ChatGPT+, Oct. 2023

20 Dec 2021

High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach[1]* Andreas Blattmann[1]* Dominik Lorenz[1] Patrick Esser[R] Björn Ommer[1]

[1]Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany [R]Runway ML

https://github.com/CompVis/latent-diffusion

Abstract

By decomposing the image formation process into a sequential application of denoising autoencoders, diffusion models (DMs) achieve state-of-the-art synthesis results on image data and beyond. Additionally, their formulati... lows for a guiding mechanism to control the ima... eration process without retraining. However, sinc... models typically operate directly in pixel space, op... tion of powerful DMs often consumes hundreds of ...

Latent Diffusion => Stable Diffusion (StabilityAI)

Open sourced, Aug. 2022

Diffusion models

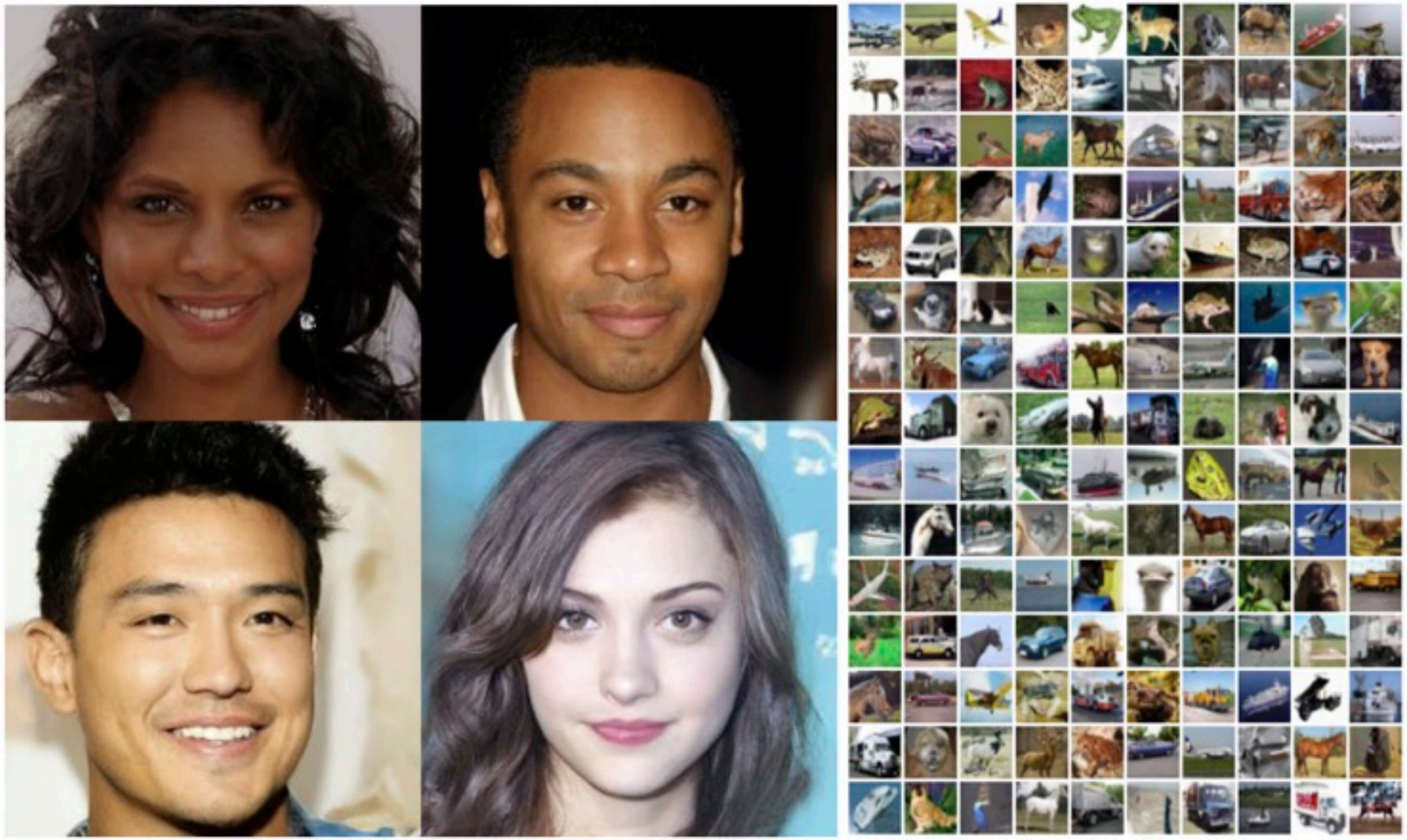


Figure 1: Generated samples on CelebA-HQ 256 × 256 (left) and unconditional CIFAR10 (right)

J. Ho et al. [Denoising diffusion probabilistic models](). NeurIPS 2020

Diffusion models

- "We can sample with as few as 25 forward passes while maintaining FIDs comparable to BigGAN"

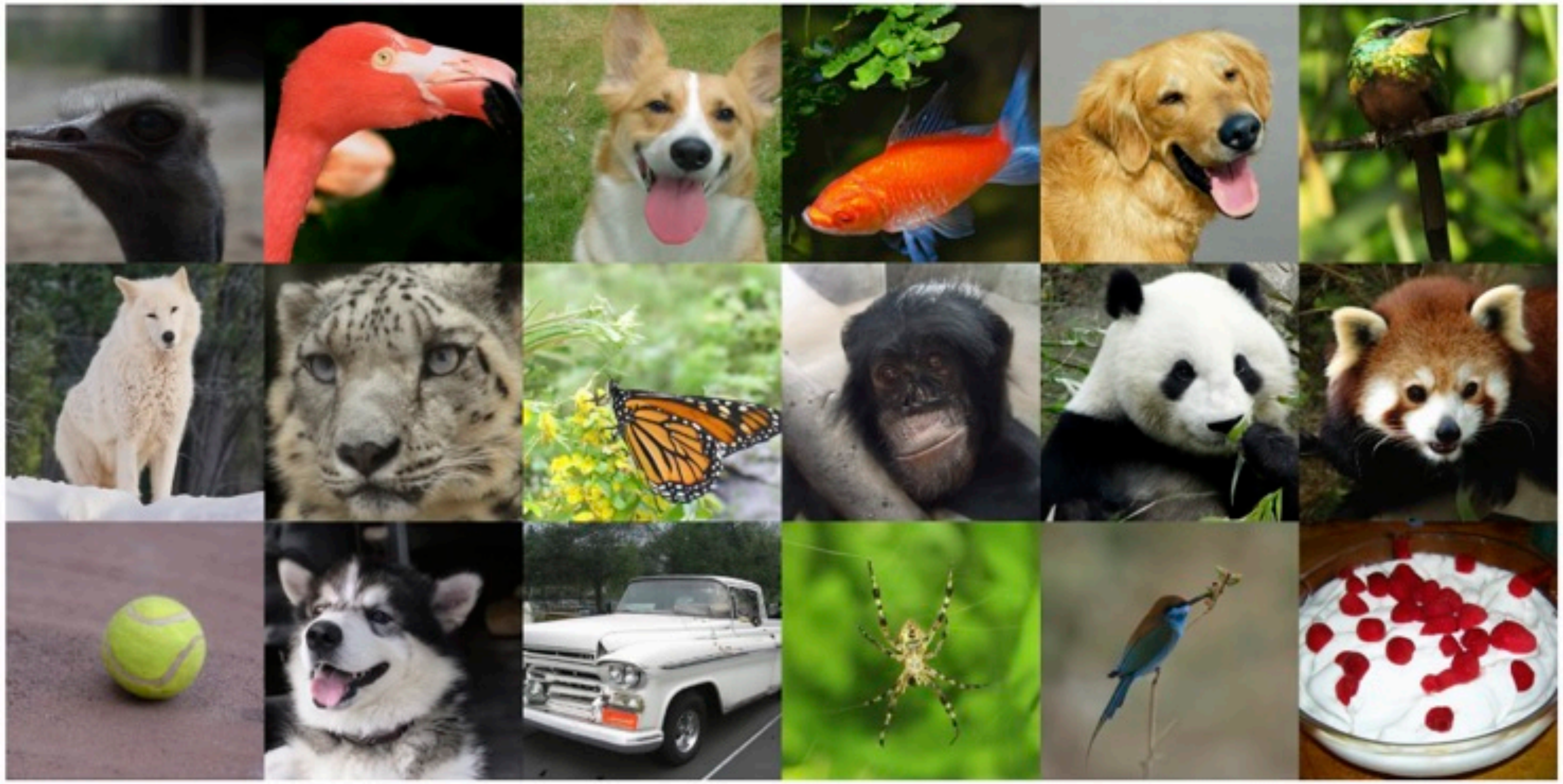
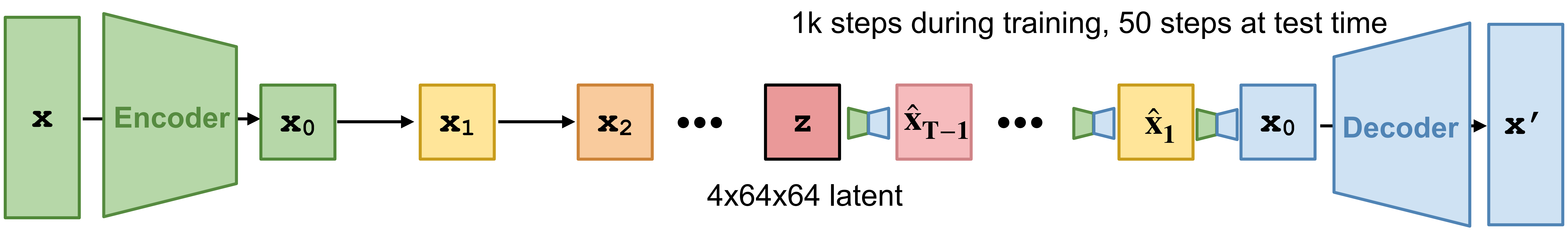


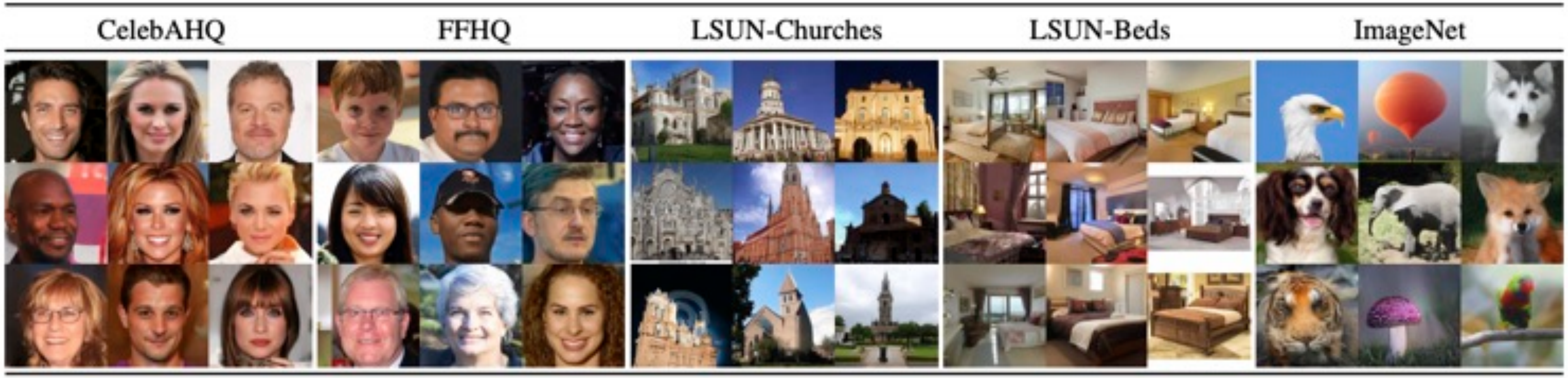
Figure 1: Selected samples from our best ImageNet 512×512 model (FID 3.85)

P. Dhariwal and A. Nichol. Diffusion Models Beat GANs on Image Synthesis. NeurIPS 2021

Latent diffusion models (aka Stable Diffusion)

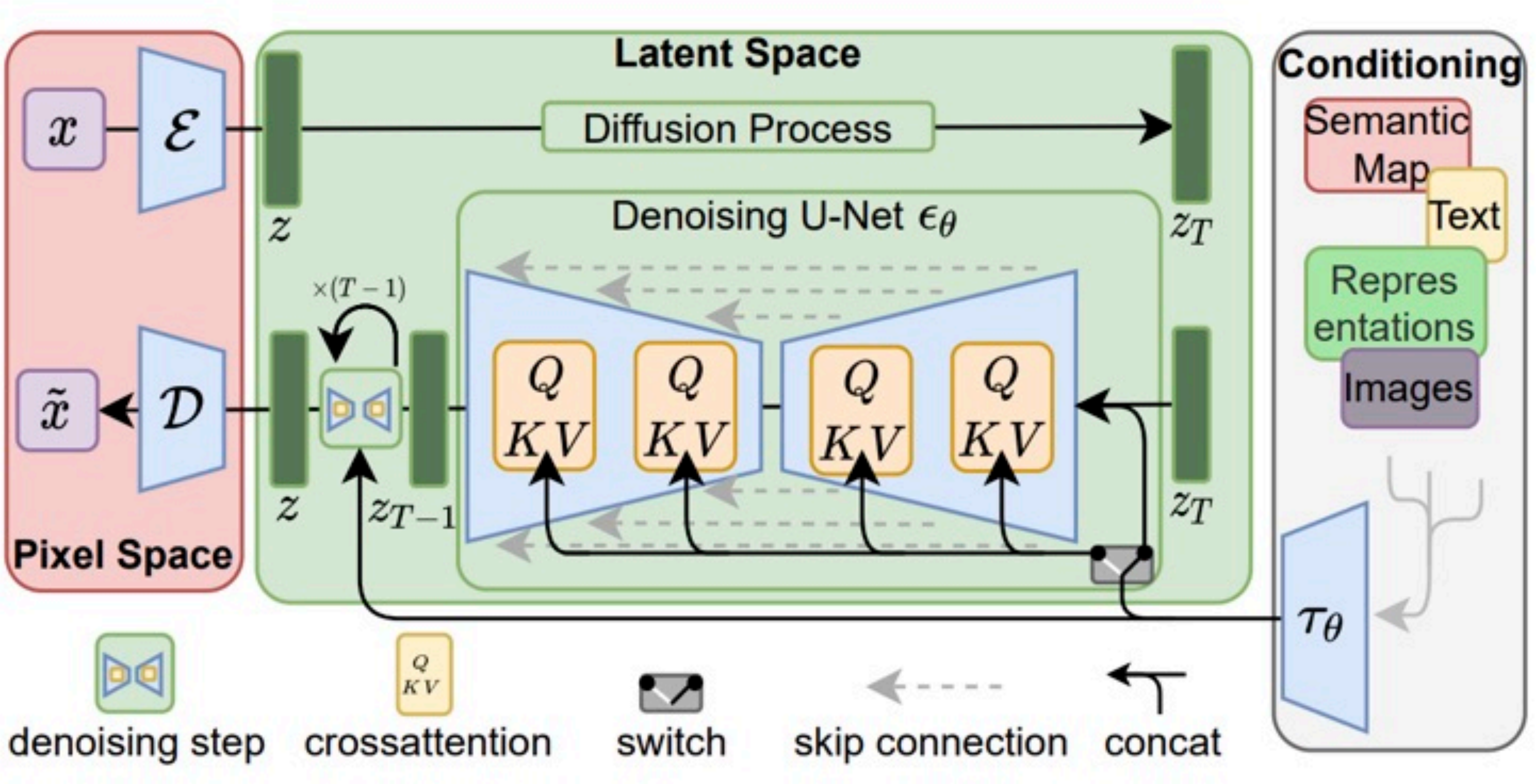
- Trained on a 2B subset of LAION5B dataset (crawl of the internet)
- Unconditional image synthesis, inpainting, stochastic super-resolution. General-purpose conditioning: class-conditional, text-to-image, layout-to-image…





R. Rombach et al. [High-Resolution Image Synthesis with Latent Diffusion Models](). CVPR 2022

Latent diffusion models (aka Stable Diffusion)



R. Rombach et al. [High-Resolution Image Synthesis with Latent Diffusion Models](). CVPR 2022

Further reading

https://arxiv.org/pdf/2208.11970.pdf

Understanding Diffusion Models: A Unified Perspective

Calvin Luo
Google Research, Brain Team
calvinluo@google.com

August 26, 2022

Contents

Introduction: Generative Models 1

Background: ELBO, VAE, and Hierarchical VAE 2

- **Evidence Lower Bound** 2
- **Variational Autoencoders** 4
- **Hierarchical Variational Autoencoders** 5

Variational Diffusion Models 6

- **Learning Diffusion Noise Parameters** 14
- **Three Equivalent Interpretations** 15

Score-based Generative Models 17

Guidance 20

- **Classifier Guidance** 21
- **Classifier-Free Guidance** 21

Other Generative Models

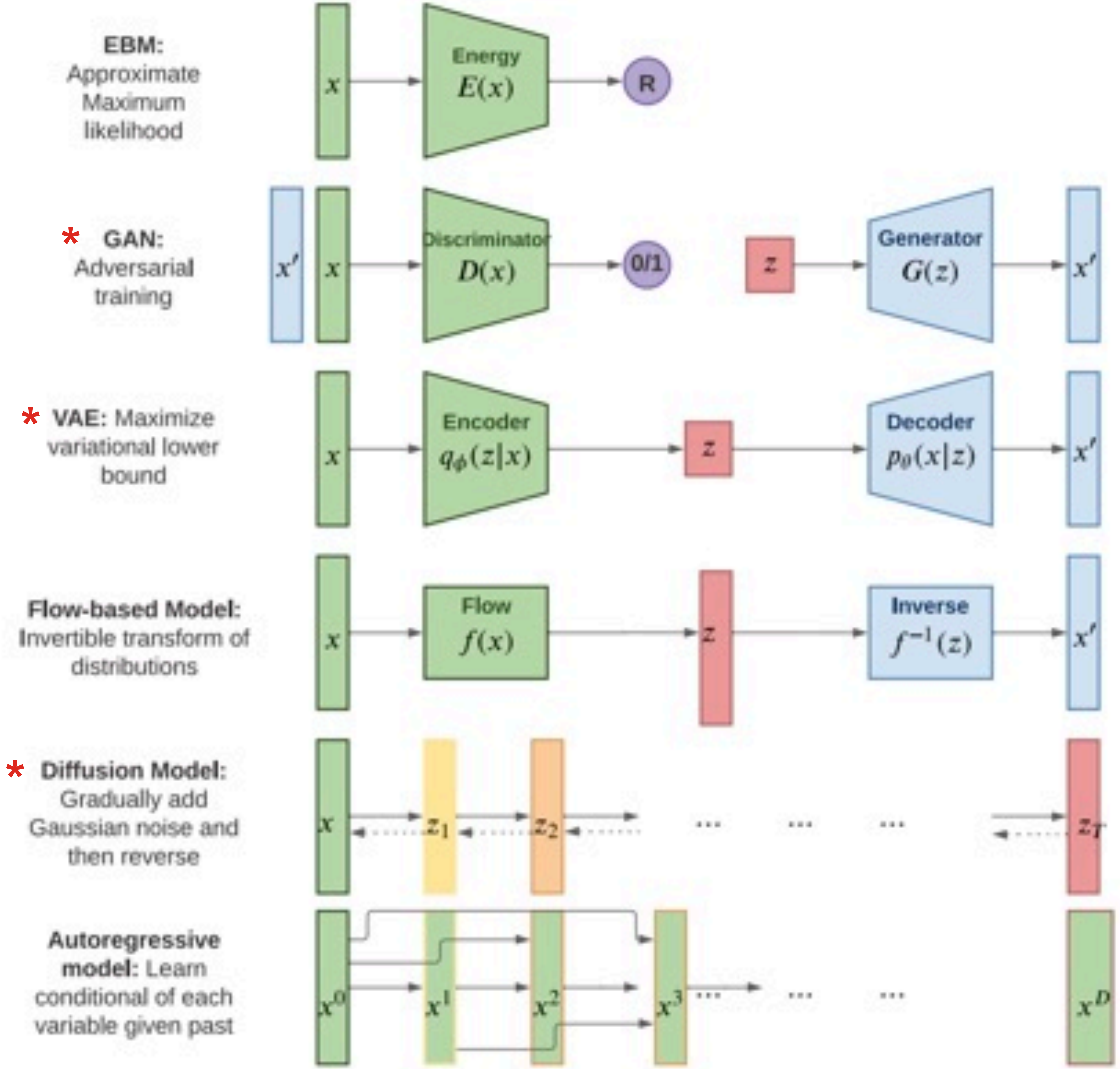


Fig. from Murphy 2023, adapted from

https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Agenda

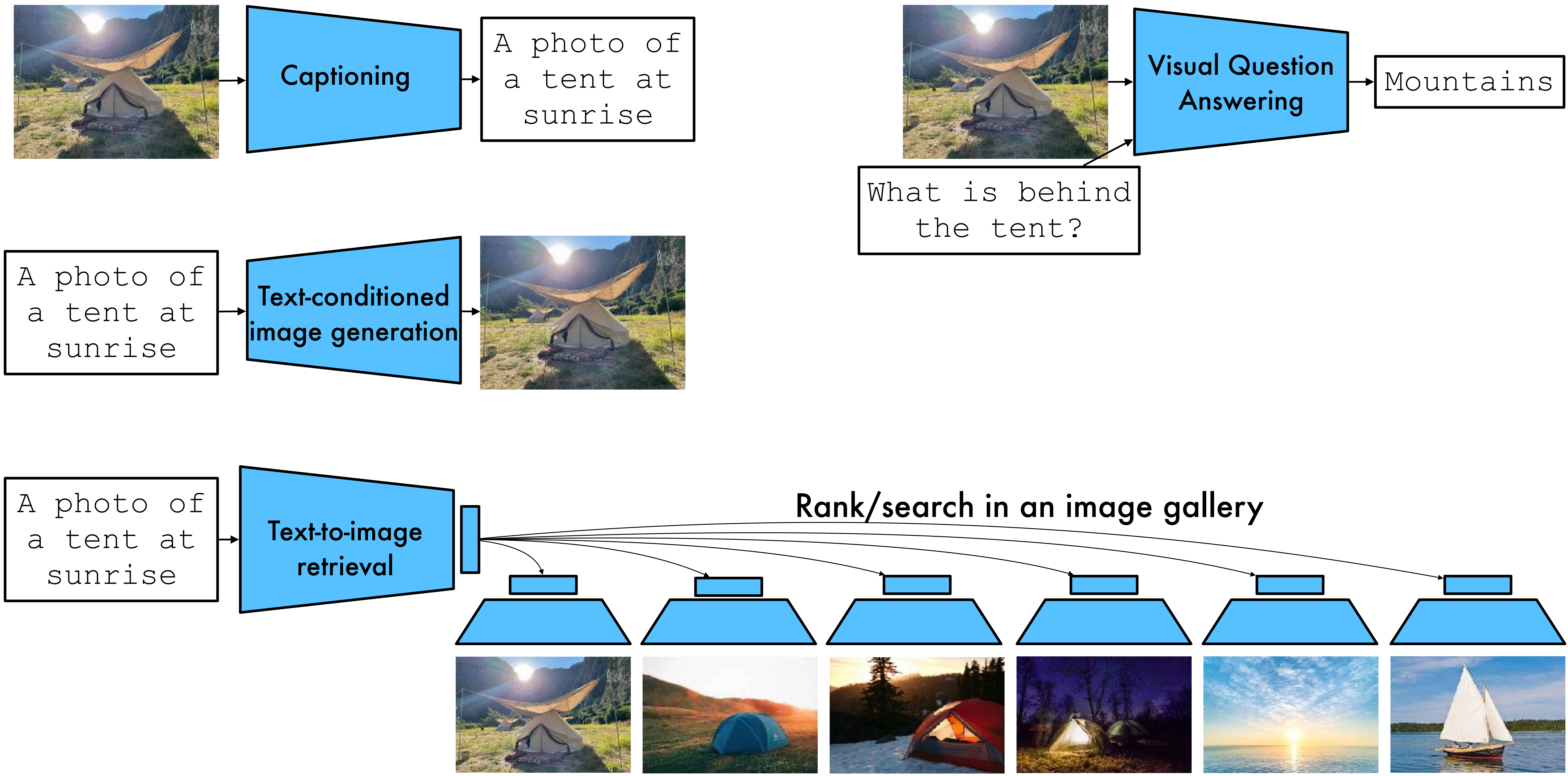
1. Generative neural networks

- VAE: Variational autoencoders
- GAN: Generative adversarial networks
- Diffusion models

2. Vision & language

- Text-to-image retrieval
- Text-to-image generation
- Image captioning

Vision & Language: Tasks

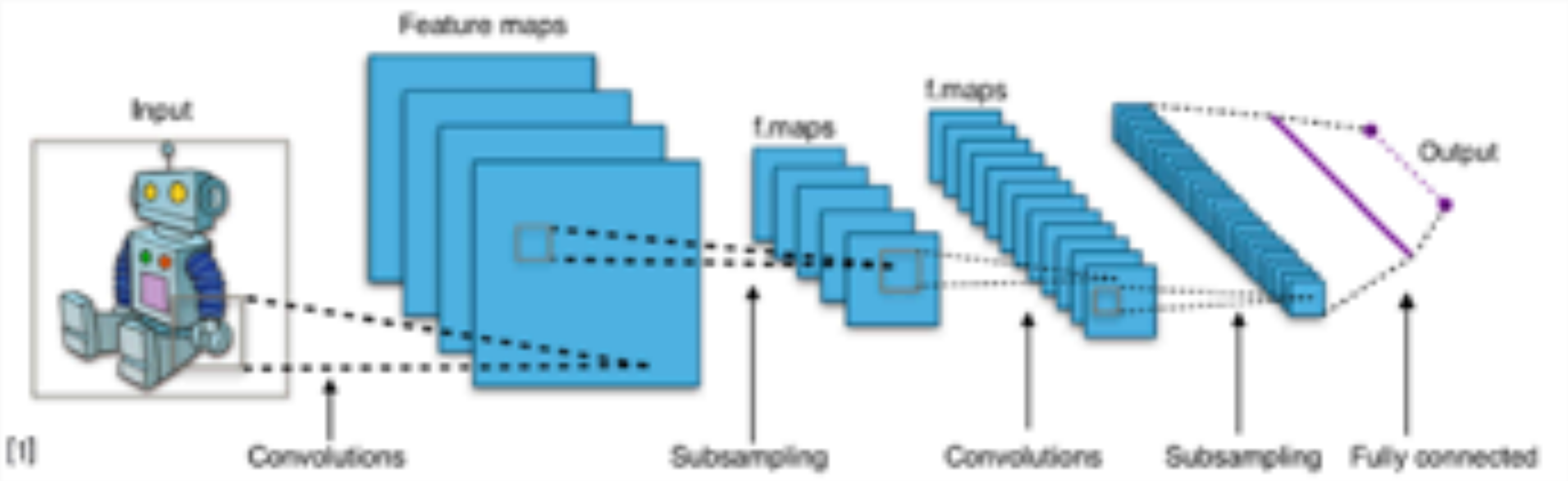


Vision & Language: Tasks

Before:
One architecture per field

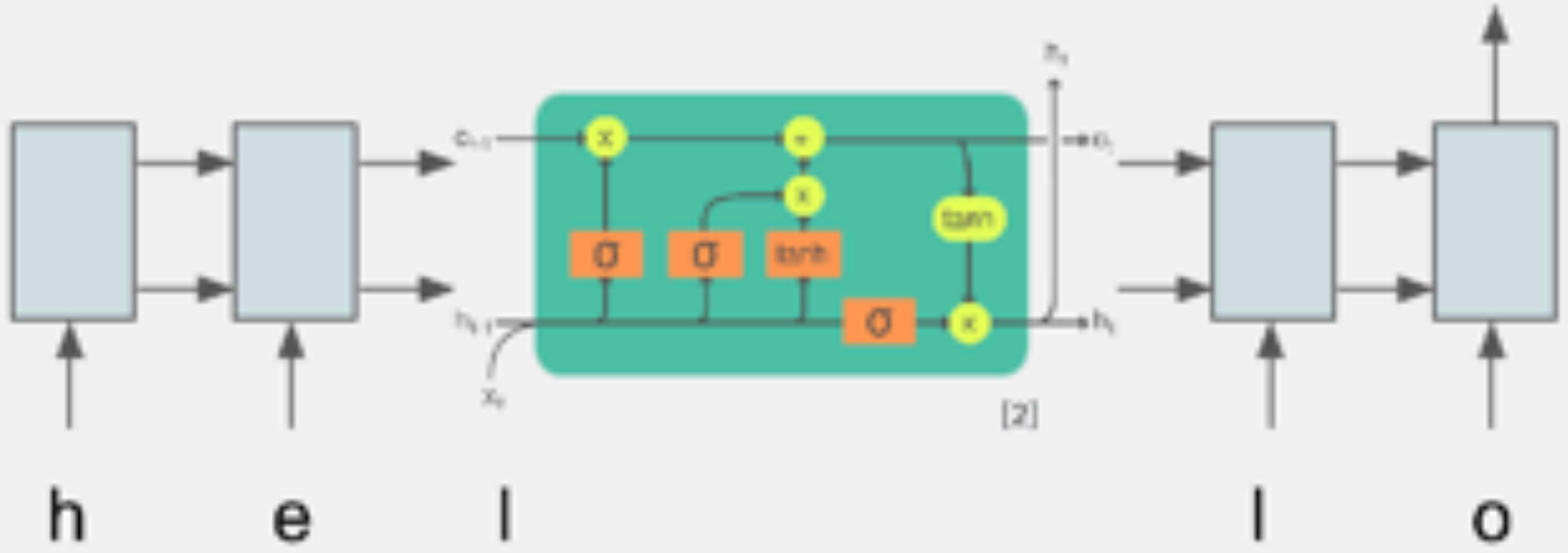
Computer Vision

Convolutional NNs (+ResNets)



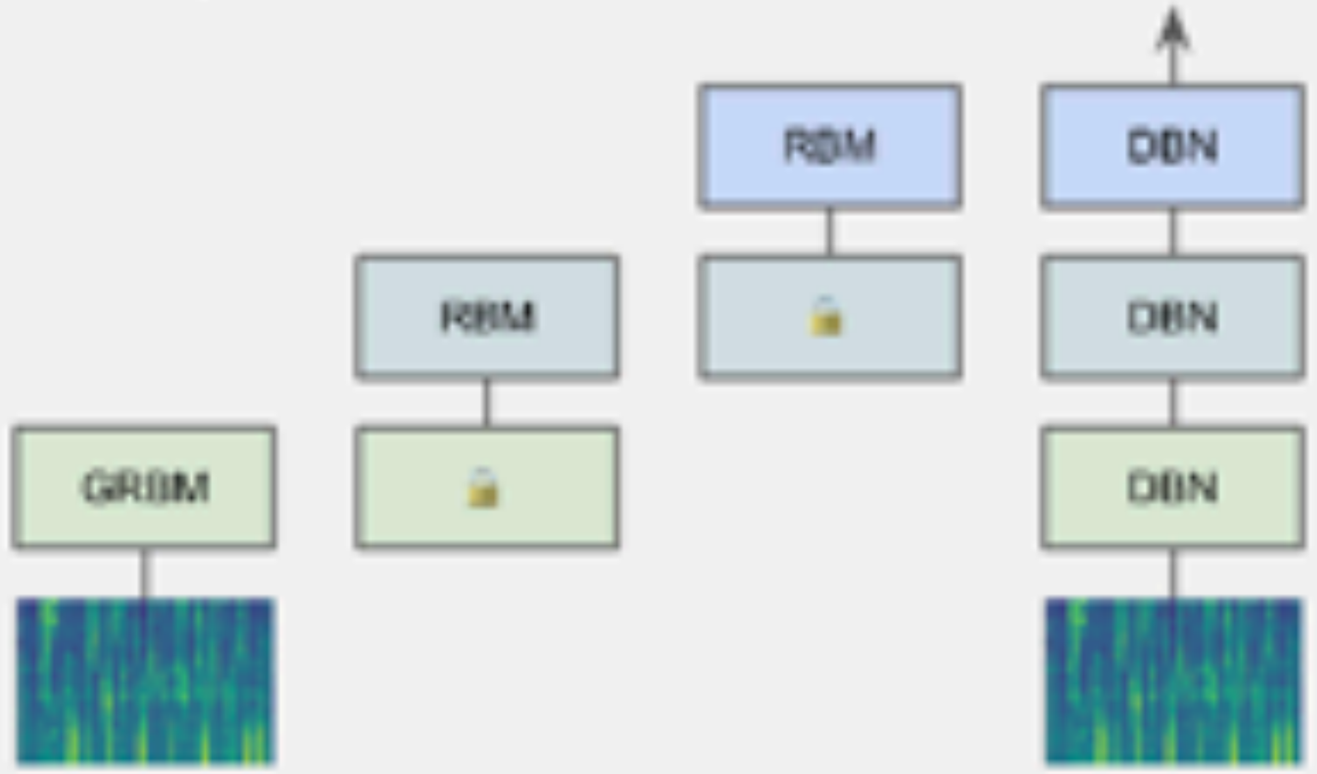
Natural Lang. Proc.

Recurrent NNs (+LSTMs)



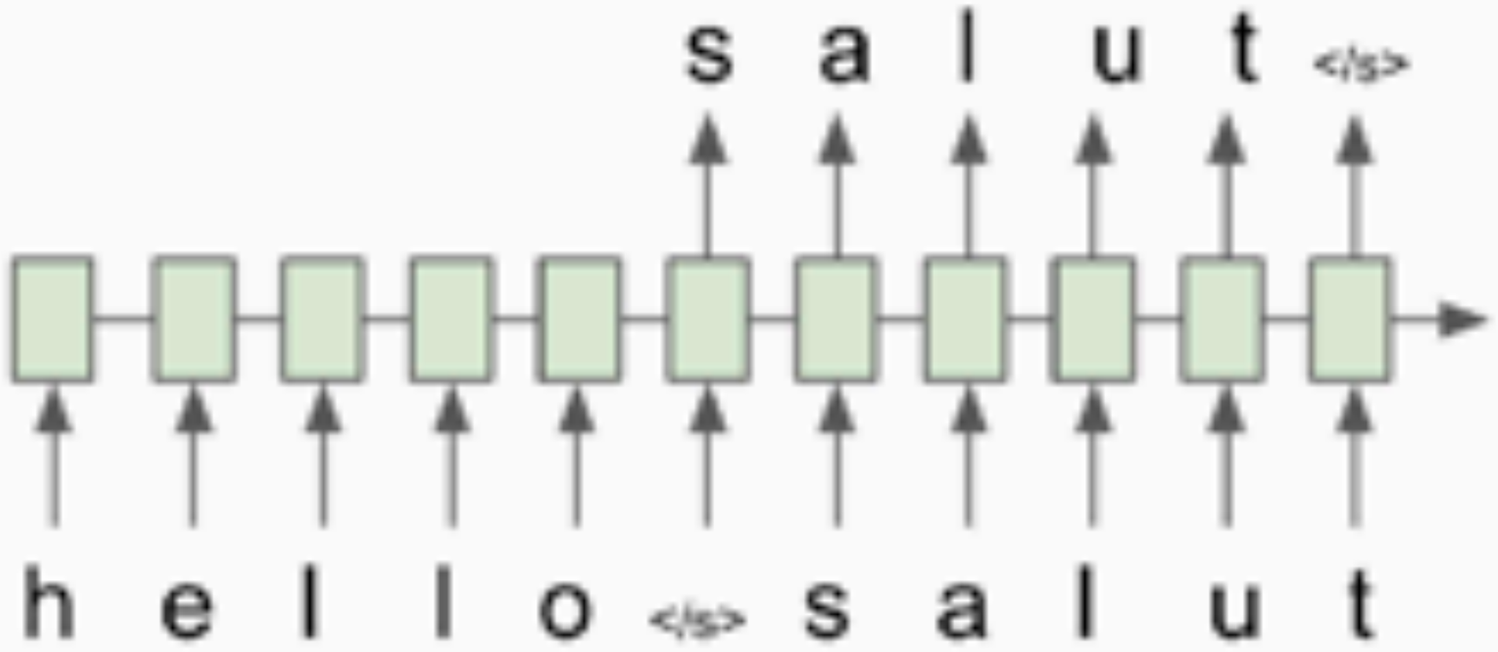
Speech

Deep Belief Nets (+non-DL)



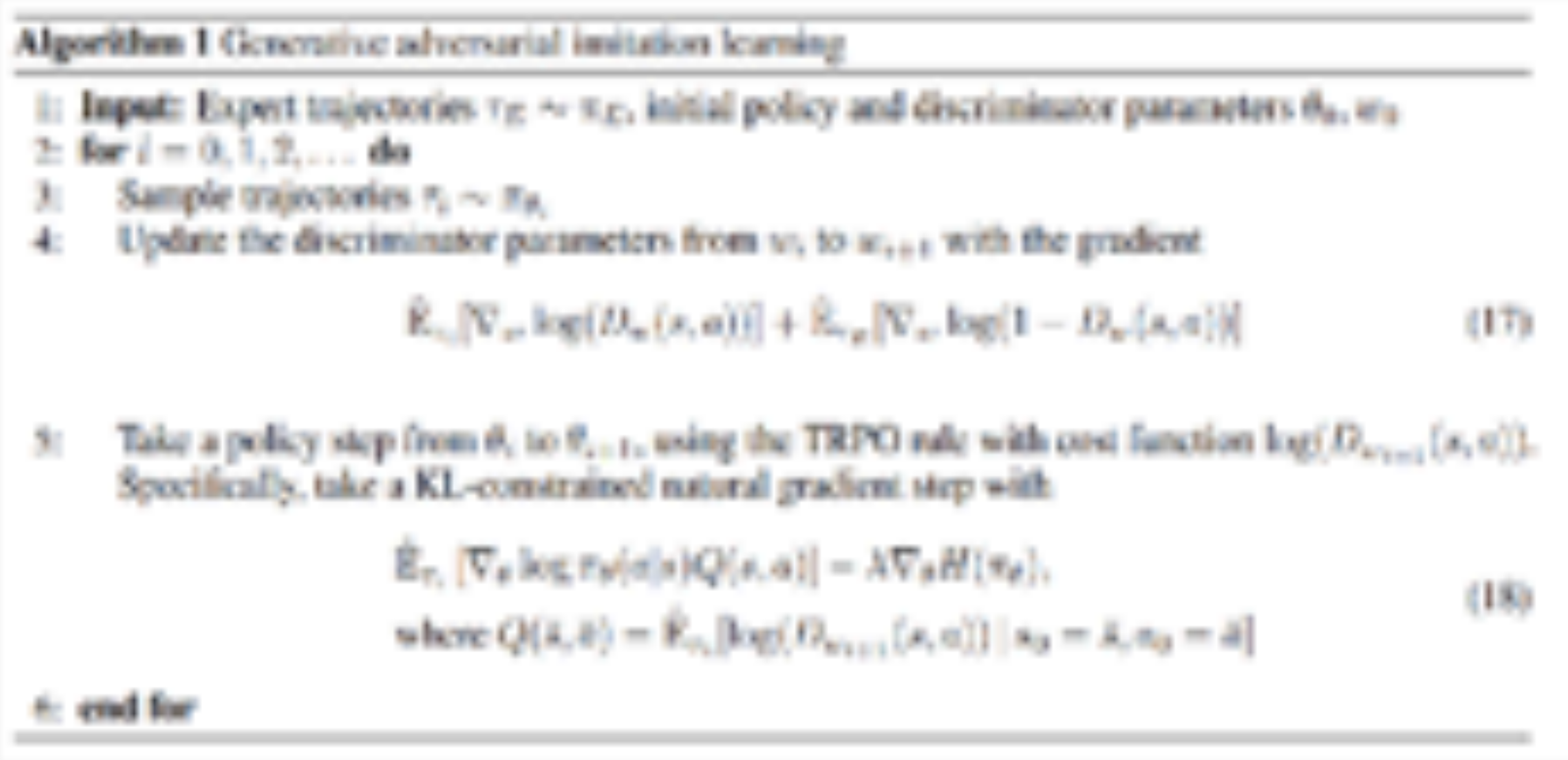
Translation

Seq2Seq



RL

BC/GAIL


Algorithm 1 Generative adversarial imitation learning

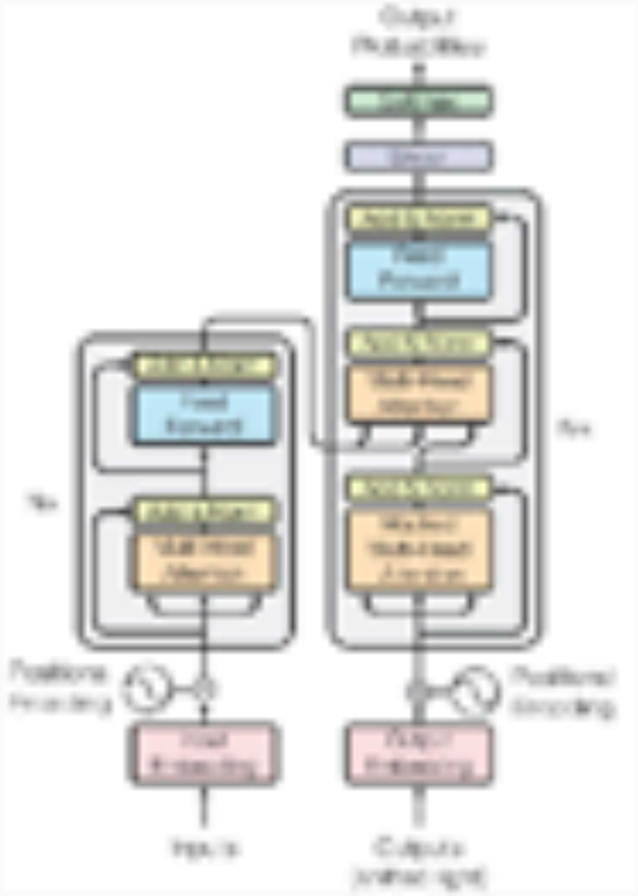
1: **Input:** Expert trajectories $\tau_E \sim \pi_E$, initial policy and discriminator parameters θ_0, w_0
2: **for** $i = 0, 1, 2, \ldots$ **do**
3: Sample trajectories $\tau_i \sim \pi_{\theta_i}$
4: Update the discriminator parameters from w_i to w_{i+1} with the gradient
5: Take a policy step from θ_i to θ_{i+1}, using the TRPO rule with cost function $\log(D_{w_{i+1}}(s, a))$. Specifically, take a KL-constrained natural gradient step with
6: **end for**

[1] CNN image CC-BY-SA by Aphex34 for Wikipedia https://commons.wikimedia.org/wiki/File:Typical_cnn.png
[2] RNN image CC-BY-SA by GChe for Wikipedia https://commons.wikimedia.org/wiki/File:The_LSTM_Cell.svg

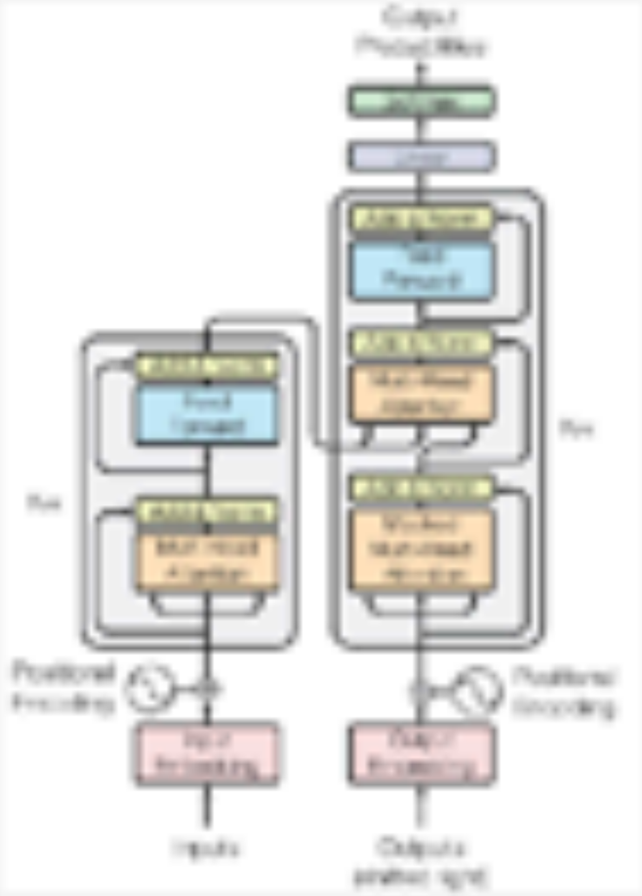
Credit: Lucas Beyer

After:
Unified architecture for all input types
(Transformers)

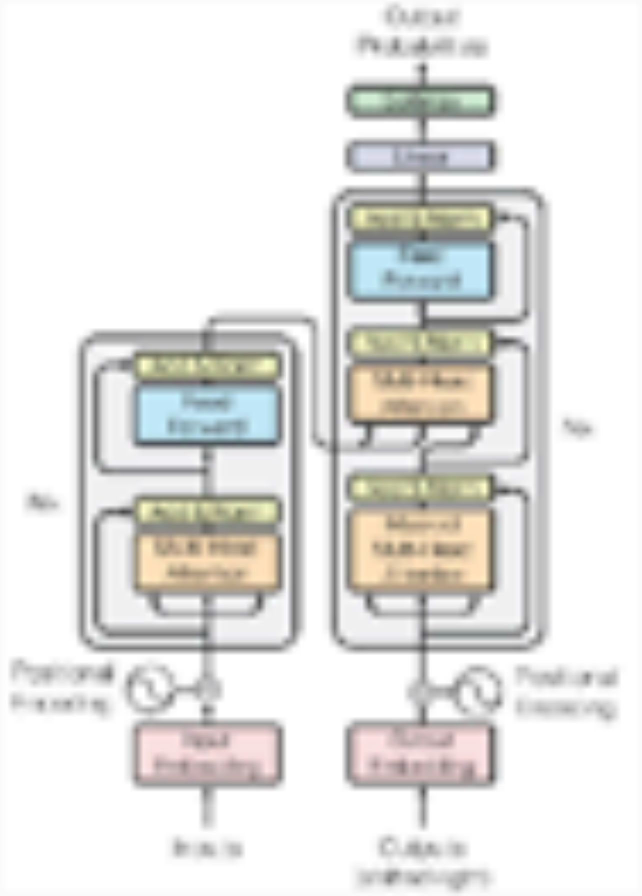
Computer Vision



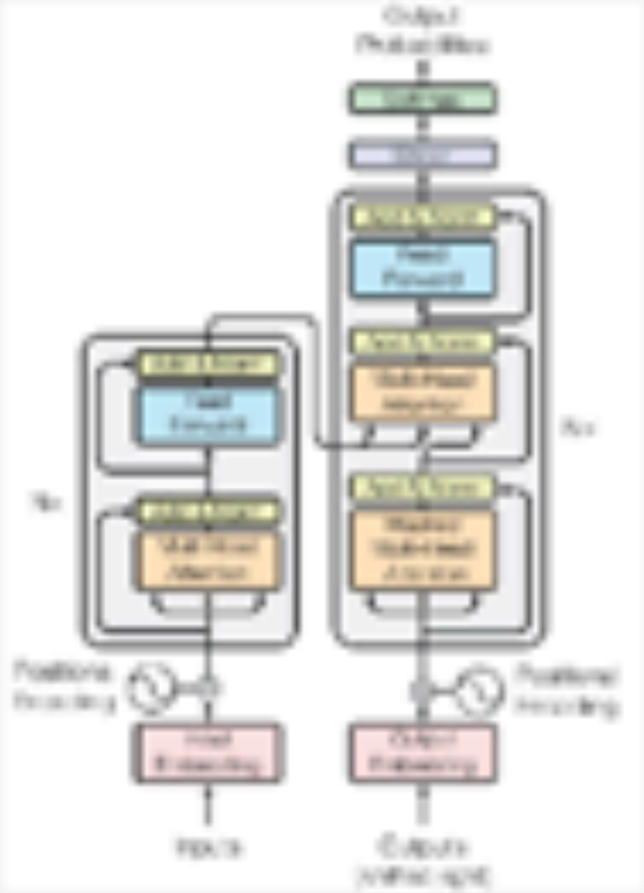
Natural Lang. Proc.



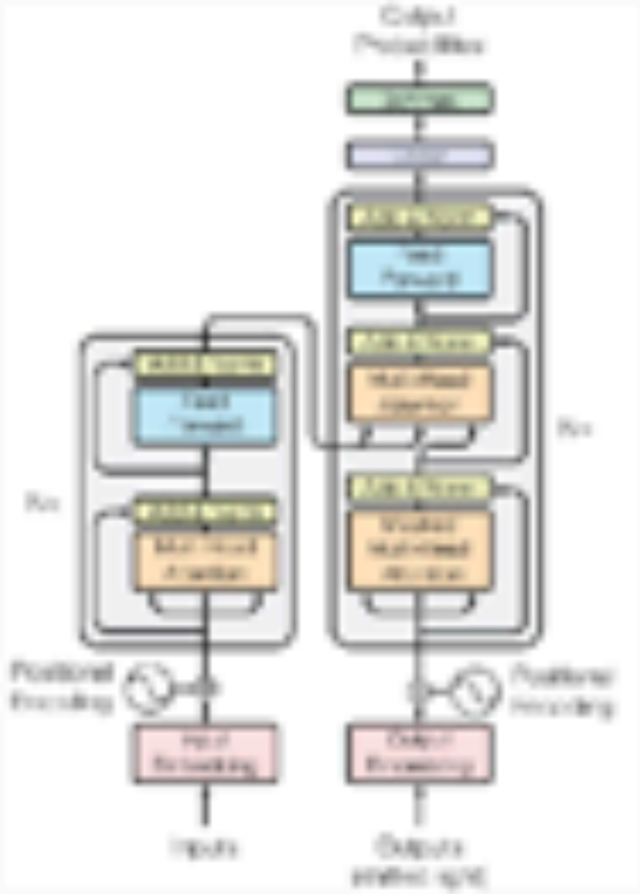
Reinf. Learning



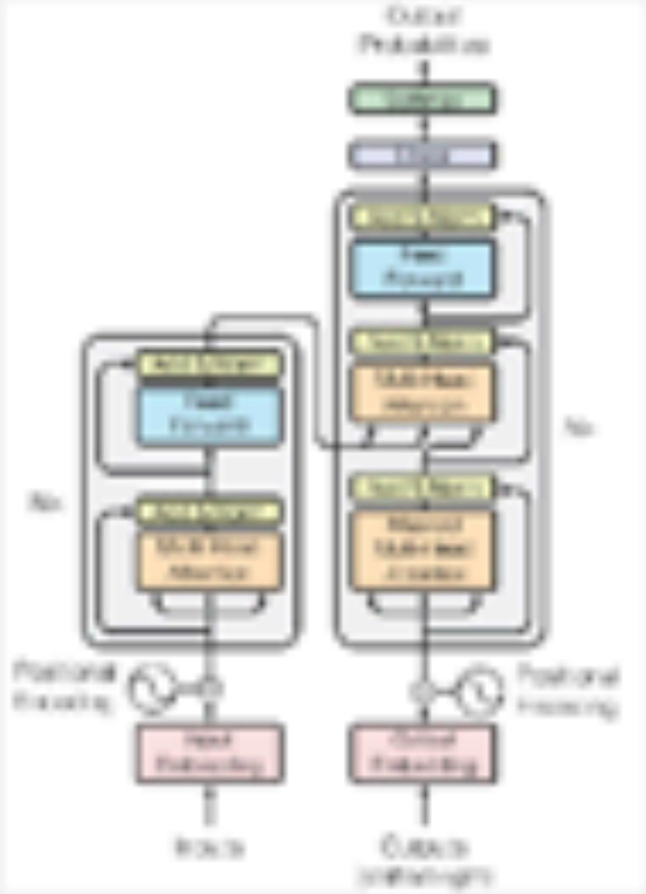
Speech



Translation



Graphs/Science



Transformer image source: "Attention Is All You Need" paper

Credit: Lucas Beyer

Agenda

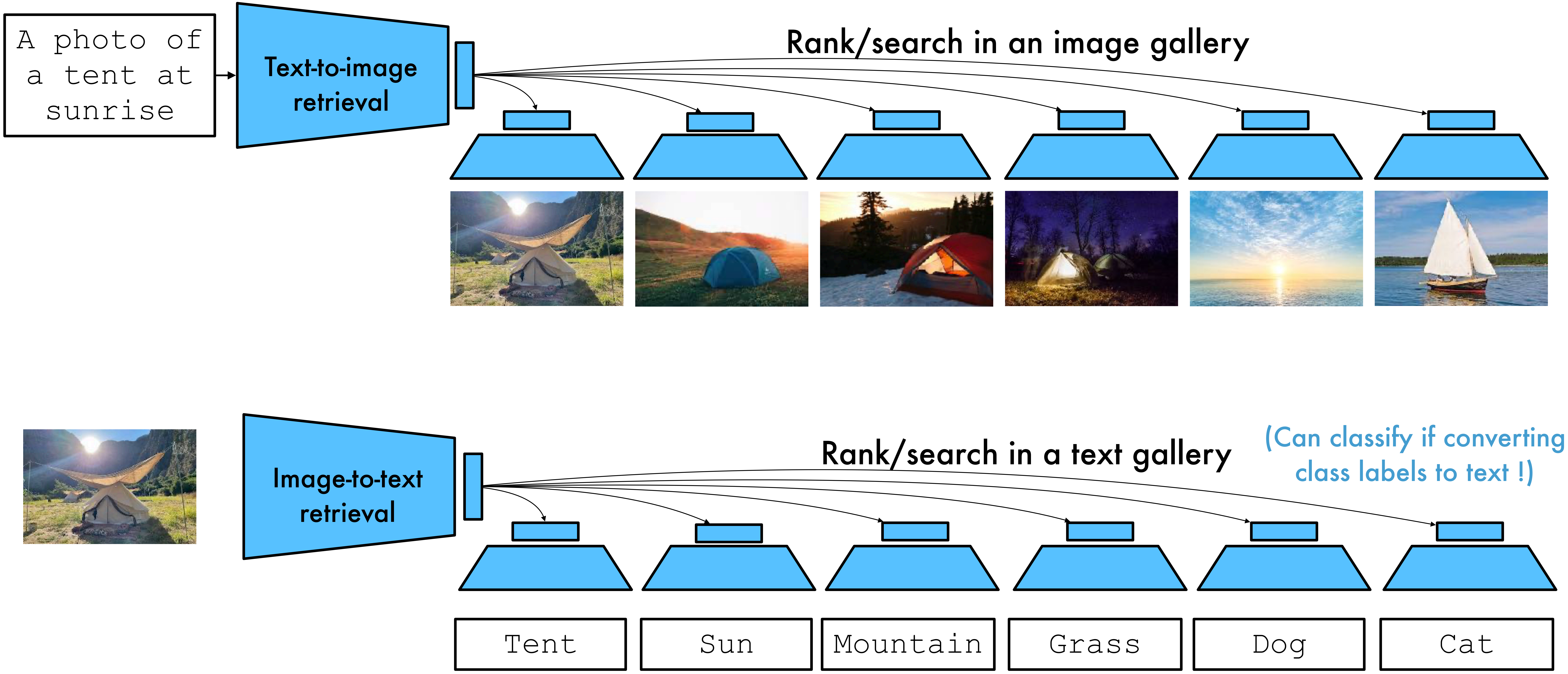
1. Generative neural networks

- VAE: Variational autoencoders
- GAN: Generative adversarial networks
- Diffusion models

2. Vision & language

- Text-to-image retrieval
- Text-to-image generation
- Image captioning

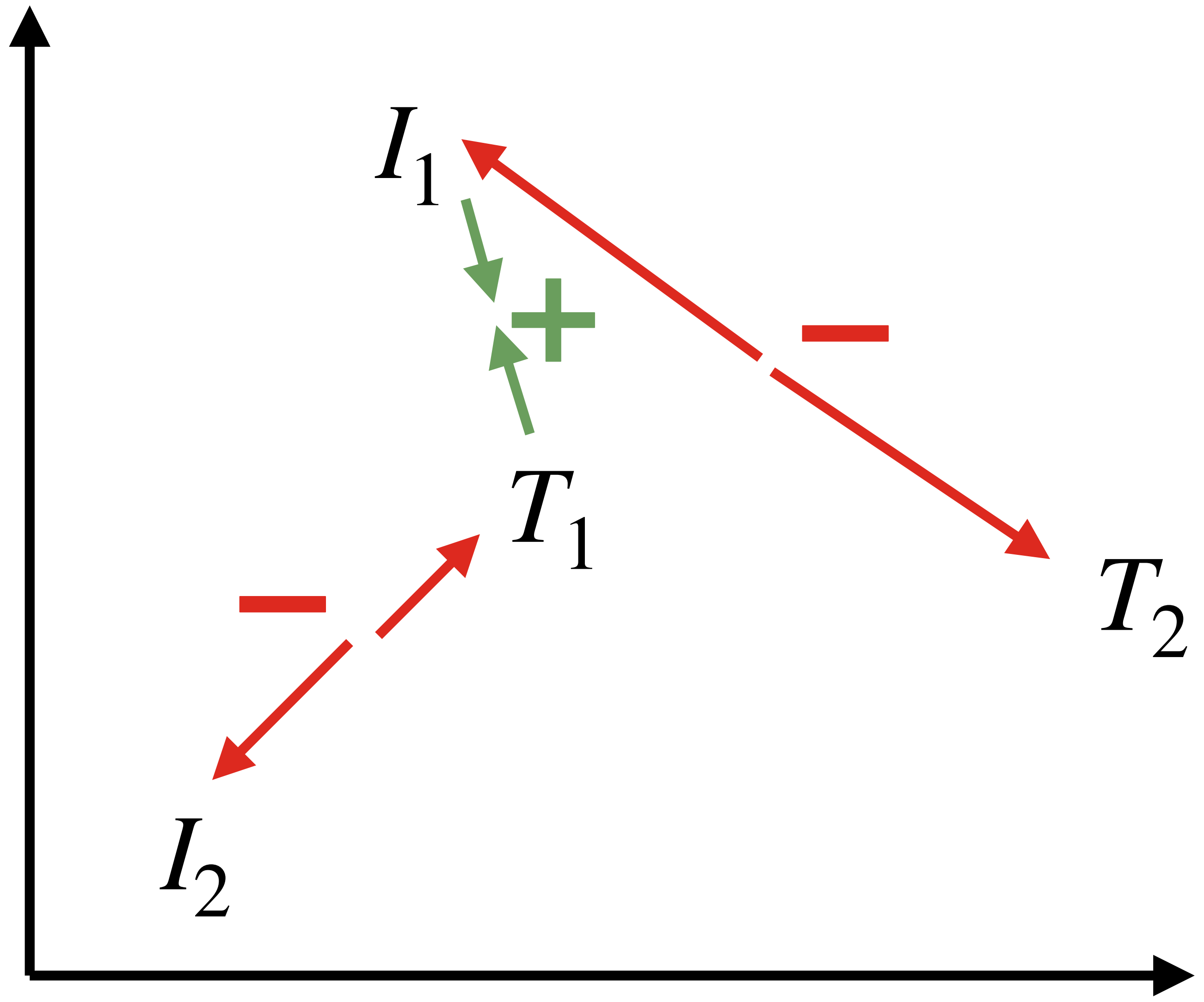
Text-image retrieval



Text-image retrieval: training

Training data: Text-image pairs (T_i, I_i)

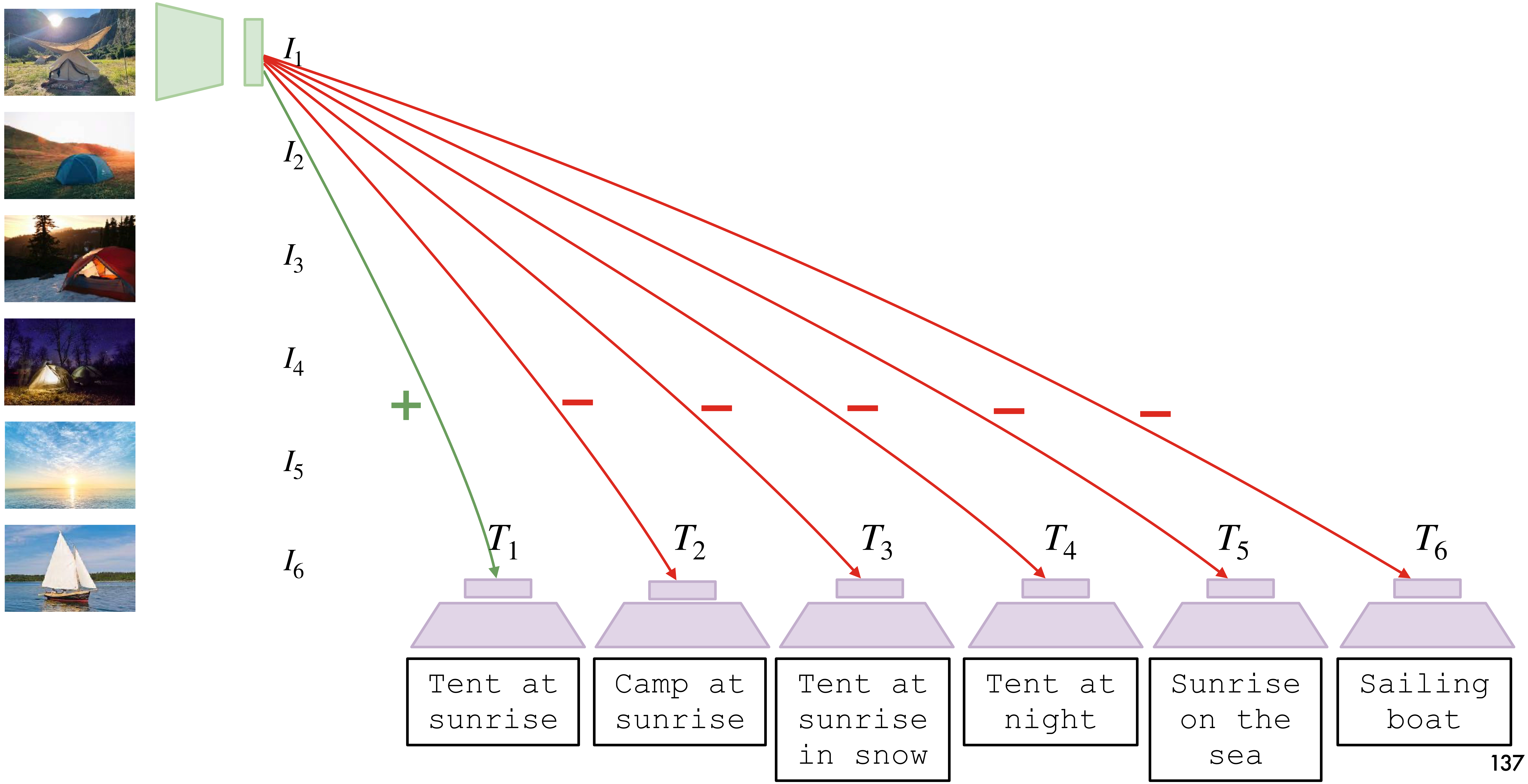
Goal: Learn a joint embedding space



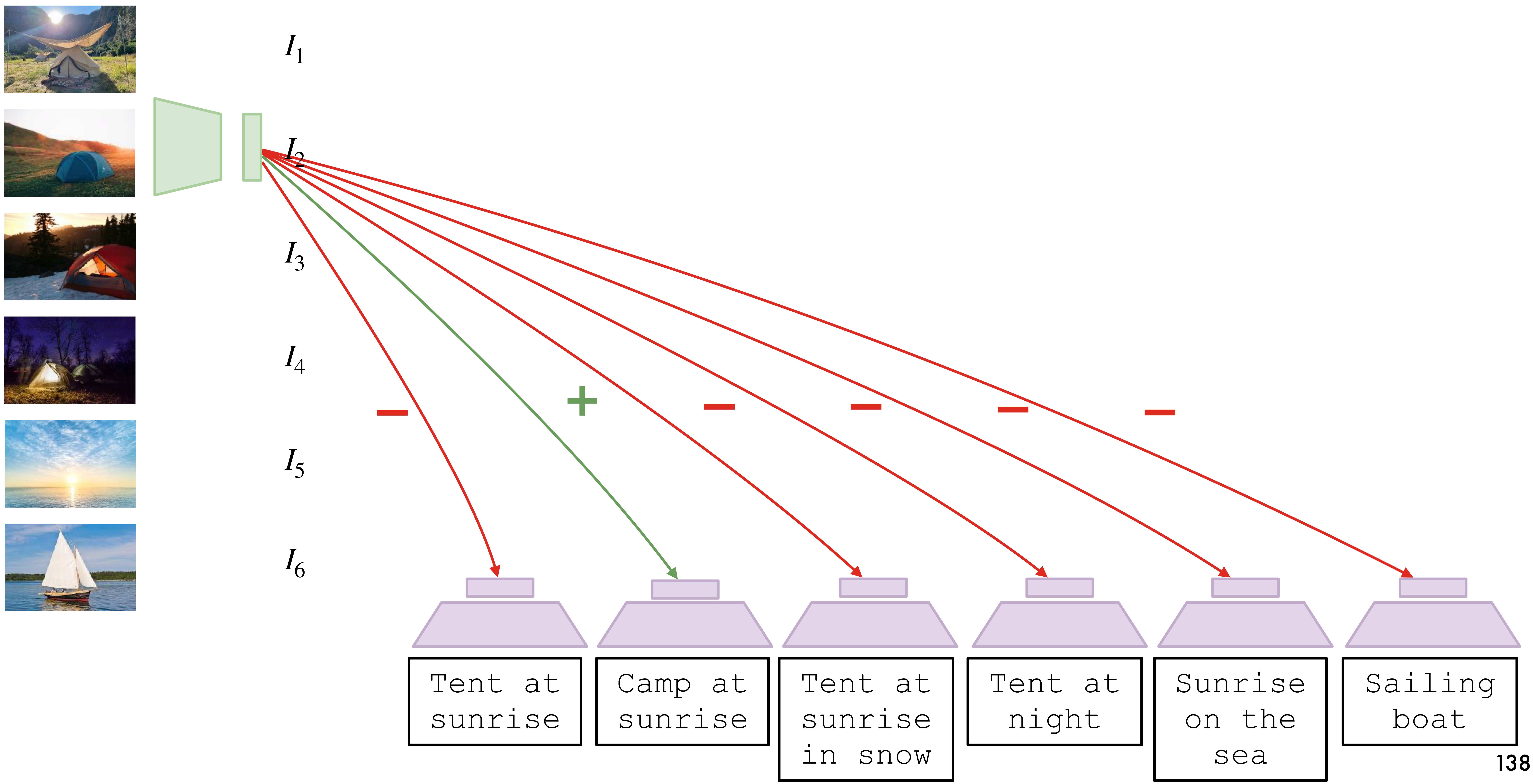
I_1	I_2	I_3	I_4	I_5	I_6

T_1	T_2	T_3	T_4	T_5	T_6
Tent at sunrise	Camp at sunrise	Tent at sunrise in snow	Tent at night	Sunrise on the sea	Sailing boat

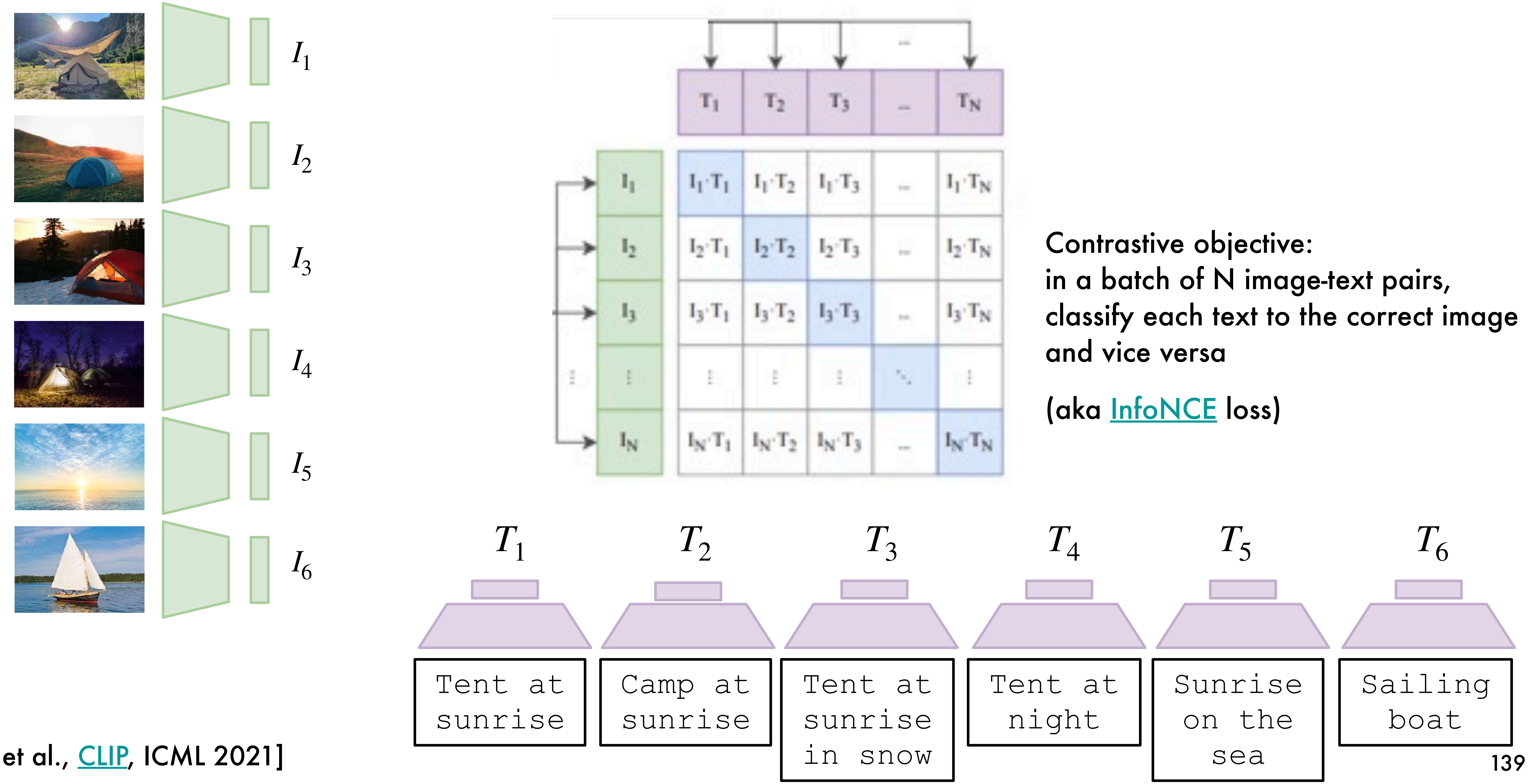
Text-image retrieval: training



Text-image retrieval: training



Text-image retrieval: training

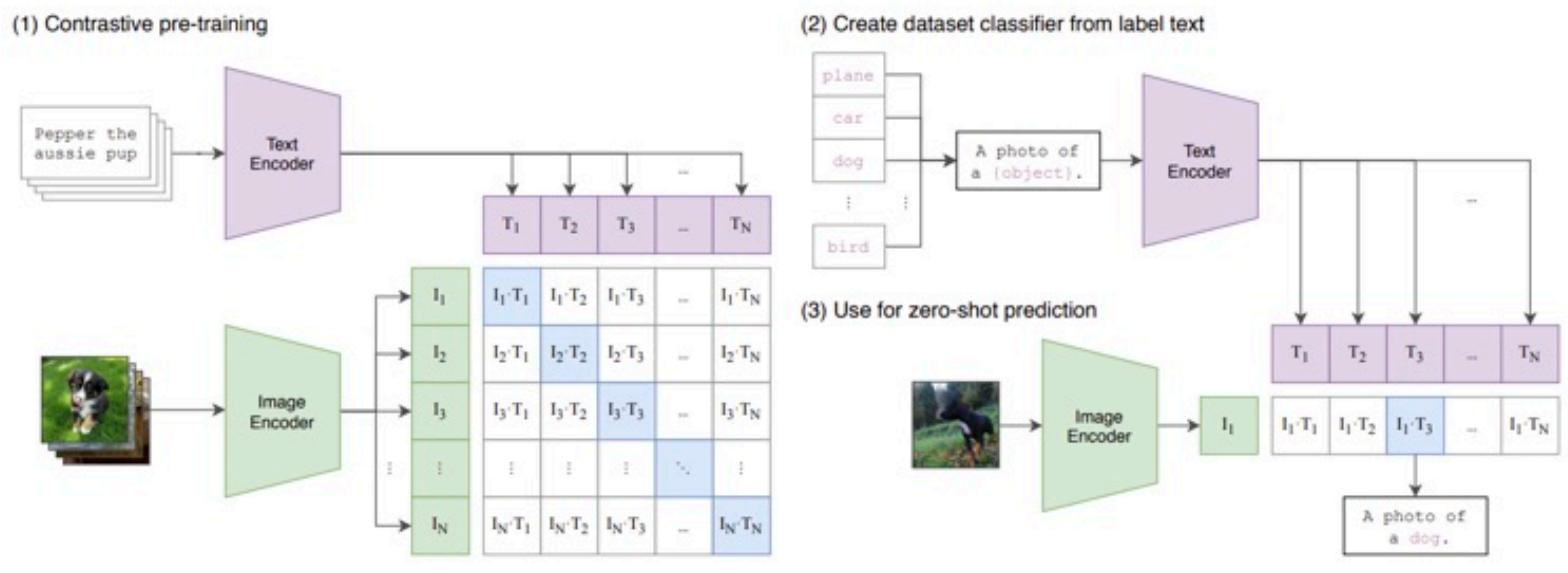


Contrastive objective:
in a batch of N image-text pairs, classify each text to the correct image and vice versa

(aka InfoNCE loss)

[Radford et al., CLIP, ICML 2021]

Contrastive Language-Image Pretraining (CLIP)



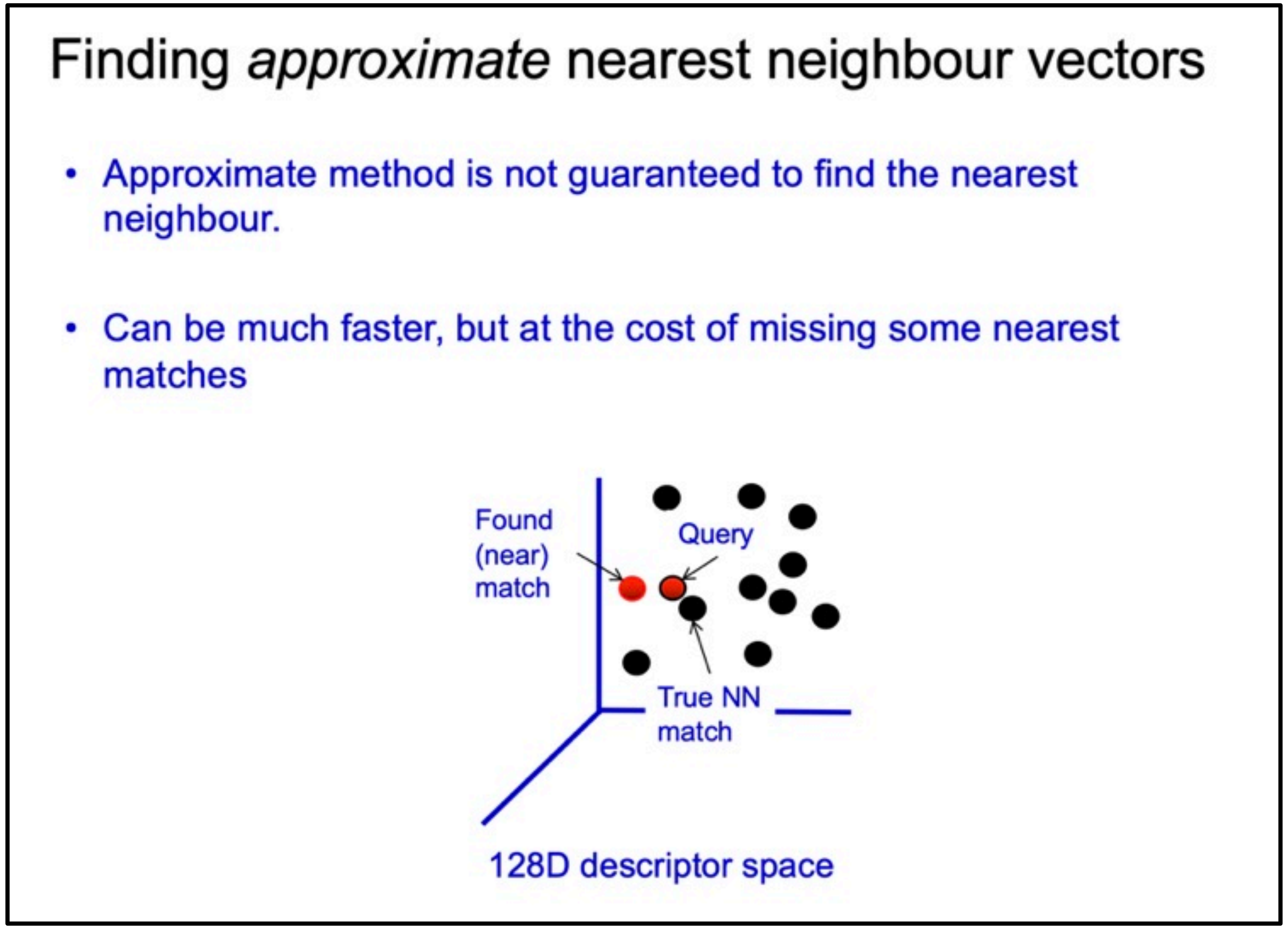
Radford et al., [Learning Transferable Visual Models From Natural Language Supervision](), ICML 2021 https://openai.com/blog/clip/

CLIP: Details

- Image encoders
  - ResNet-50 with self-attention layer on top of global average pooling
  - Vision transformer (ViT)
- Text encoder: GPT-style transformer with 63M parameters
- Dataset: 400M image-text pairs from the Web

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021 https://openai.com/blog/clip/

Remember last week: Efficient search



- Approximate nearest neighbor search if the gallery size is millions.

FAISS Library: https://github.com/facebookresearch/faiss

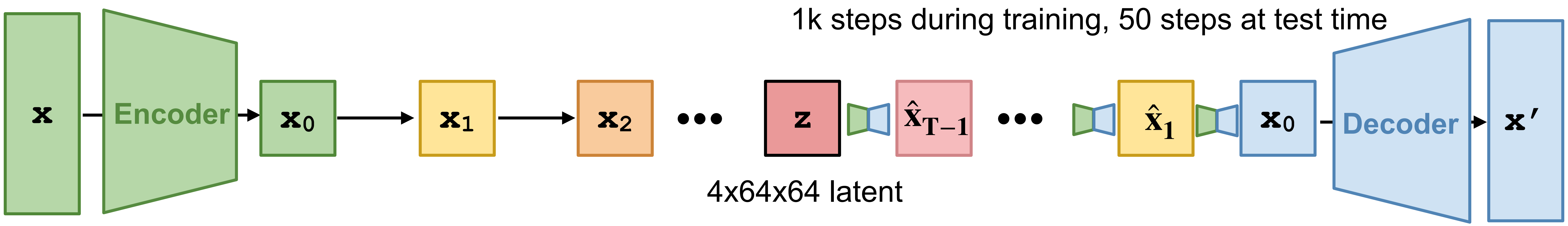
Agenda

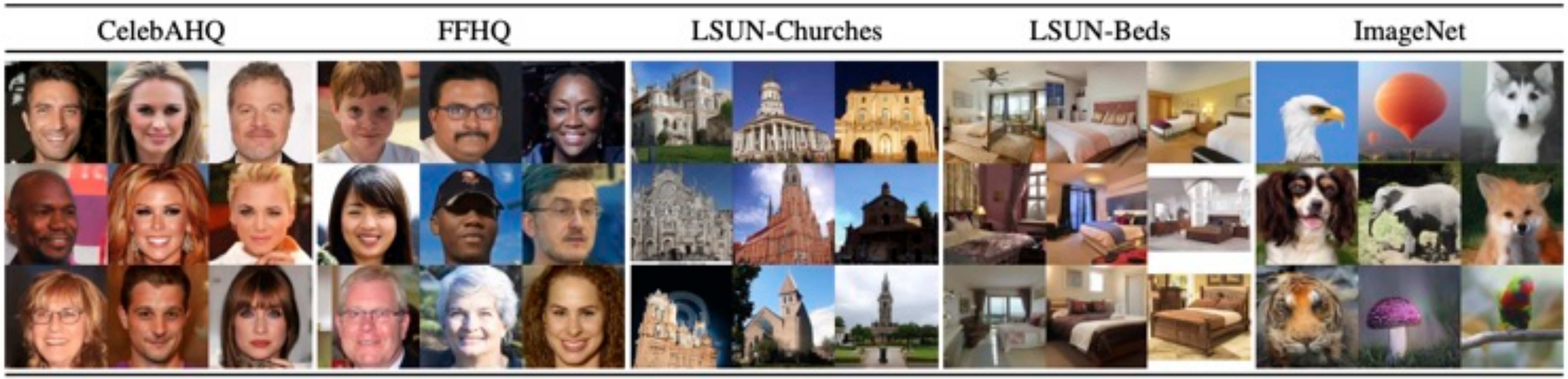
1. **Generative neural networks**
   - VAE: Variational autoencoders
   - GAN: Generative adversarial networks
   - Diffusion models

2. **Vision & language**
   - Text-to-image retrieval
   - Text-to-image generation
   - Image captioning

Latent diffusion models (aka Stable Diffusion)

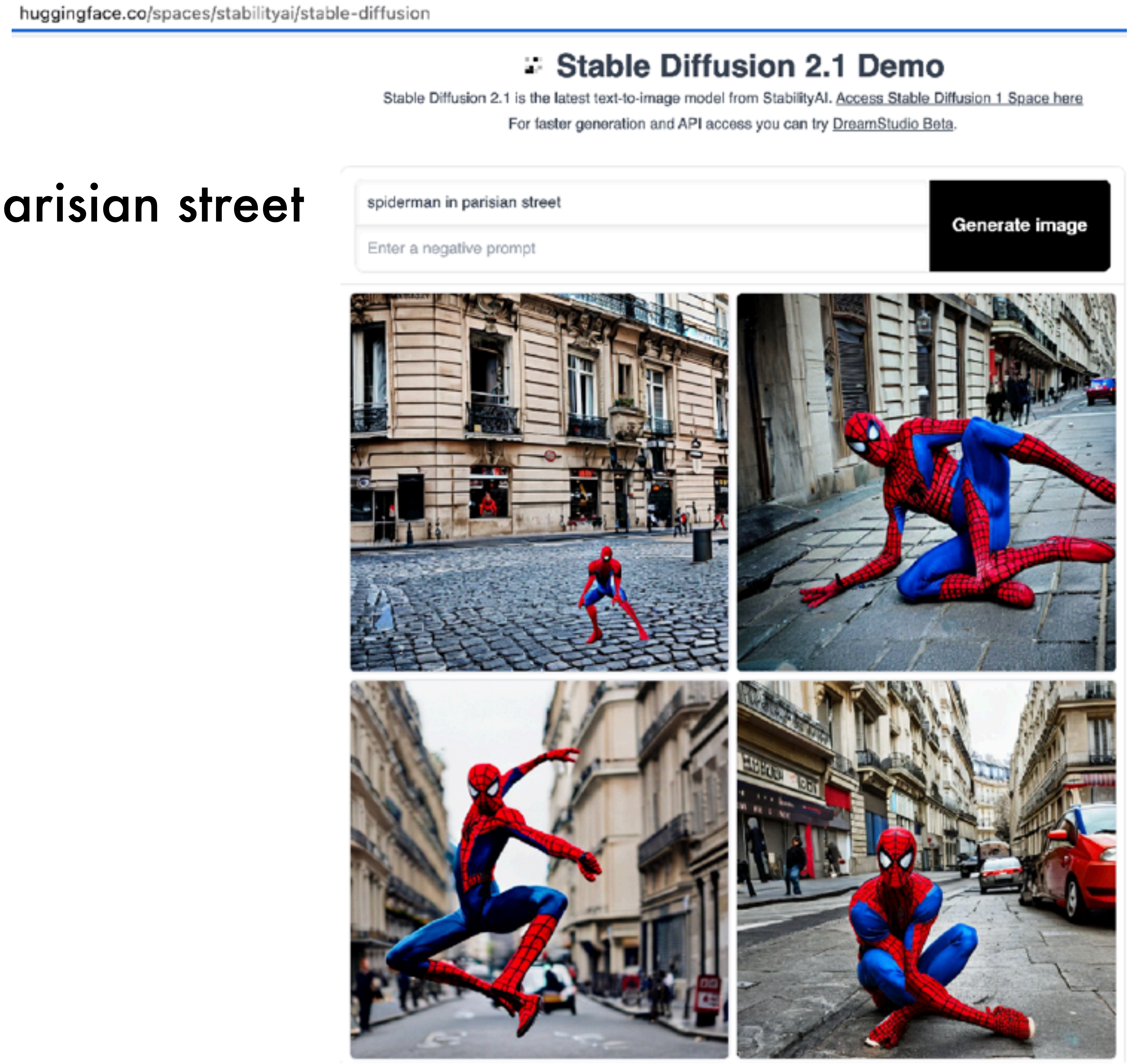
- Demo: https://huggingface.co/spaces/stabilityai/stable-diffusion





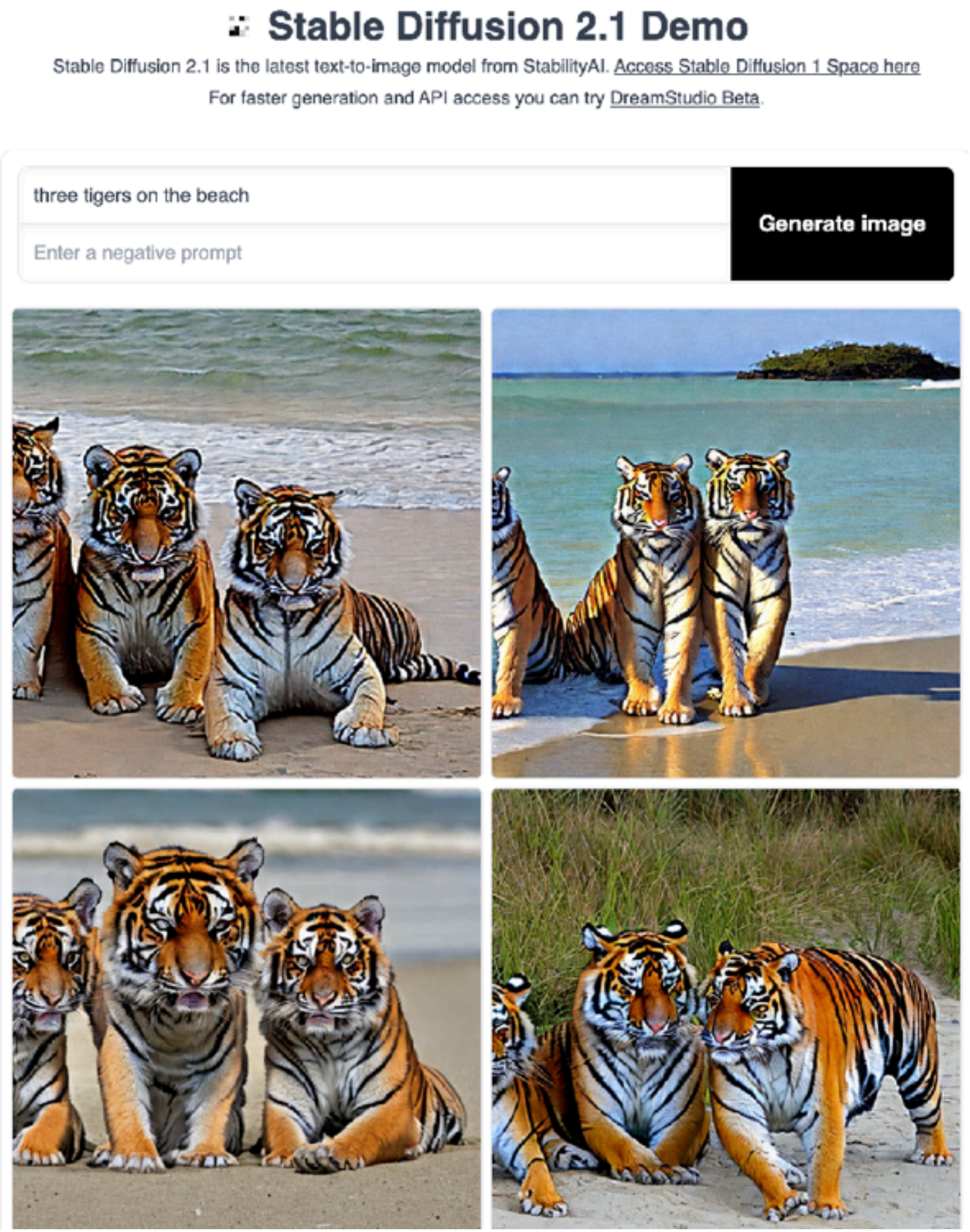
R. Rombach et al. [High-Resolution Image Synthesis with Latent Diffusion Models](). CVPR 2022

spiderman in parisian street



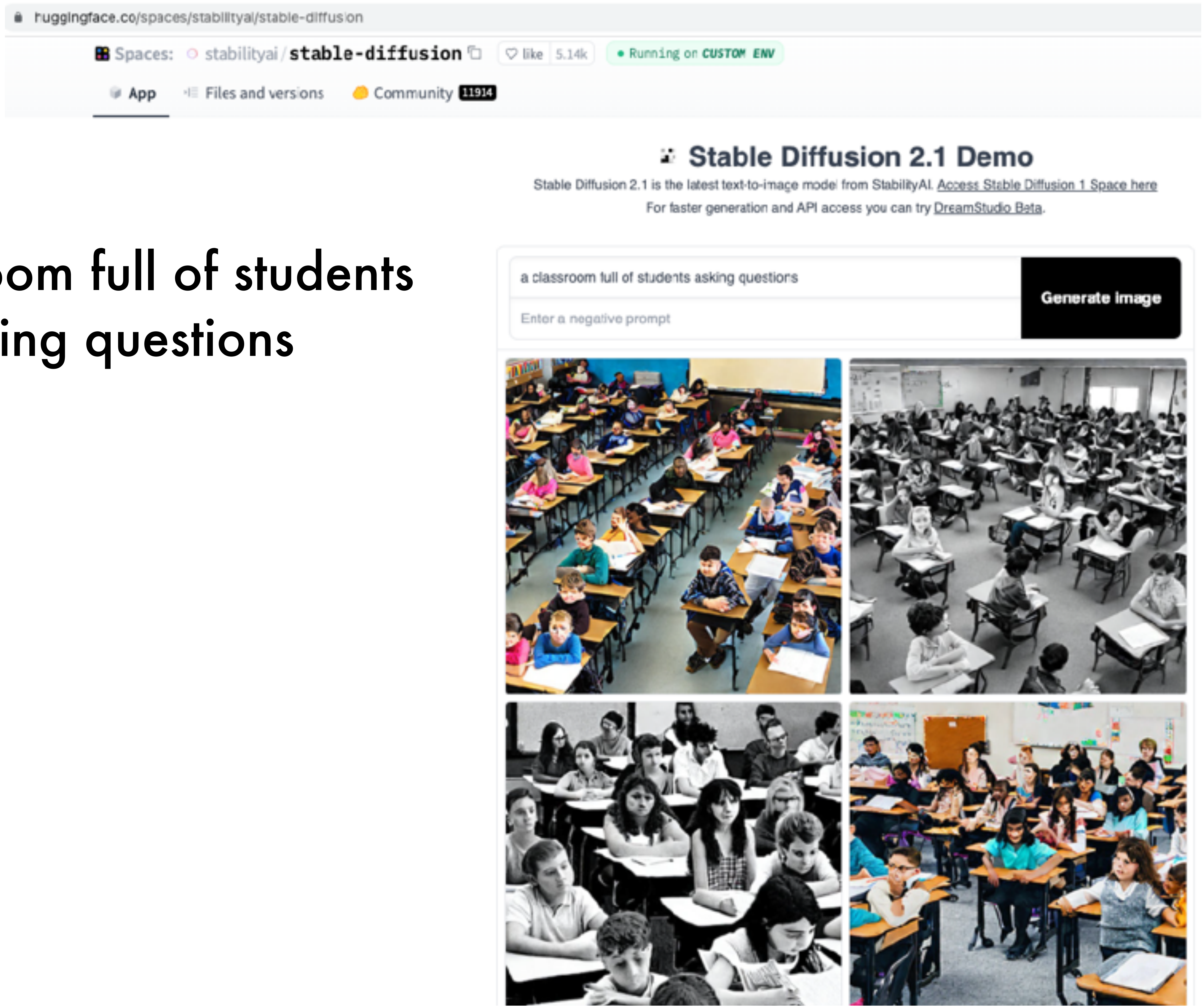
R. Rombach et al. [High-Resolution Image Synthesis with Latent Diffusion Models](). CVPR 2022

three tigers on the beach



R. Rombach et al. [High-Resolution Image Synthesis with Latent Diffusion Models](). CVPR 2022

a classroom full of students asking questions



R. Rombach et al. [High-Resolution Image Synthesis with Latent Diffusion Models](). CVPR 2022

DALL-E-2


vibrant portrait painting of Salvador Dalí with a robotic half face


a shiba inu wearing a beret and black turtleneck


a close up of a handpalm with leaves growing from it


an espresso machine that makes coffee from human souls, artstation


panda mad scientist mixing sparkling chemicals, artstation


a corgi's head depicted as an explosion of a nebula

A. Ramesh et al. [Hierarchical text-conditional image generation with CLIP latents](). 2022

DALL-E-2

“A closeup of a handpalm with leaves growing from it.”

Figure 19: Random samples from unCLIP for prompt “A close up of a handpalm with leaves growing from it.”

DALL-E-2

"Vibrant portrait painting of Salvador Dali with a robotic half face"

Figure 18: Random samples from unCLIP for prompt "Vibrant portrait painting of Salvador Dali with a robotic half face"

Imagen

“We discover that **large frozen language models** trained only on text data are surprisingly very effective text encoders for text-to-image generation, and that **scaling the size of frozen text encoder** improves sample quality significantly more than scaling the size of image diffusion model”

Sprouts in the shape of text ‘Imagen’ coming out of a fairytale book.

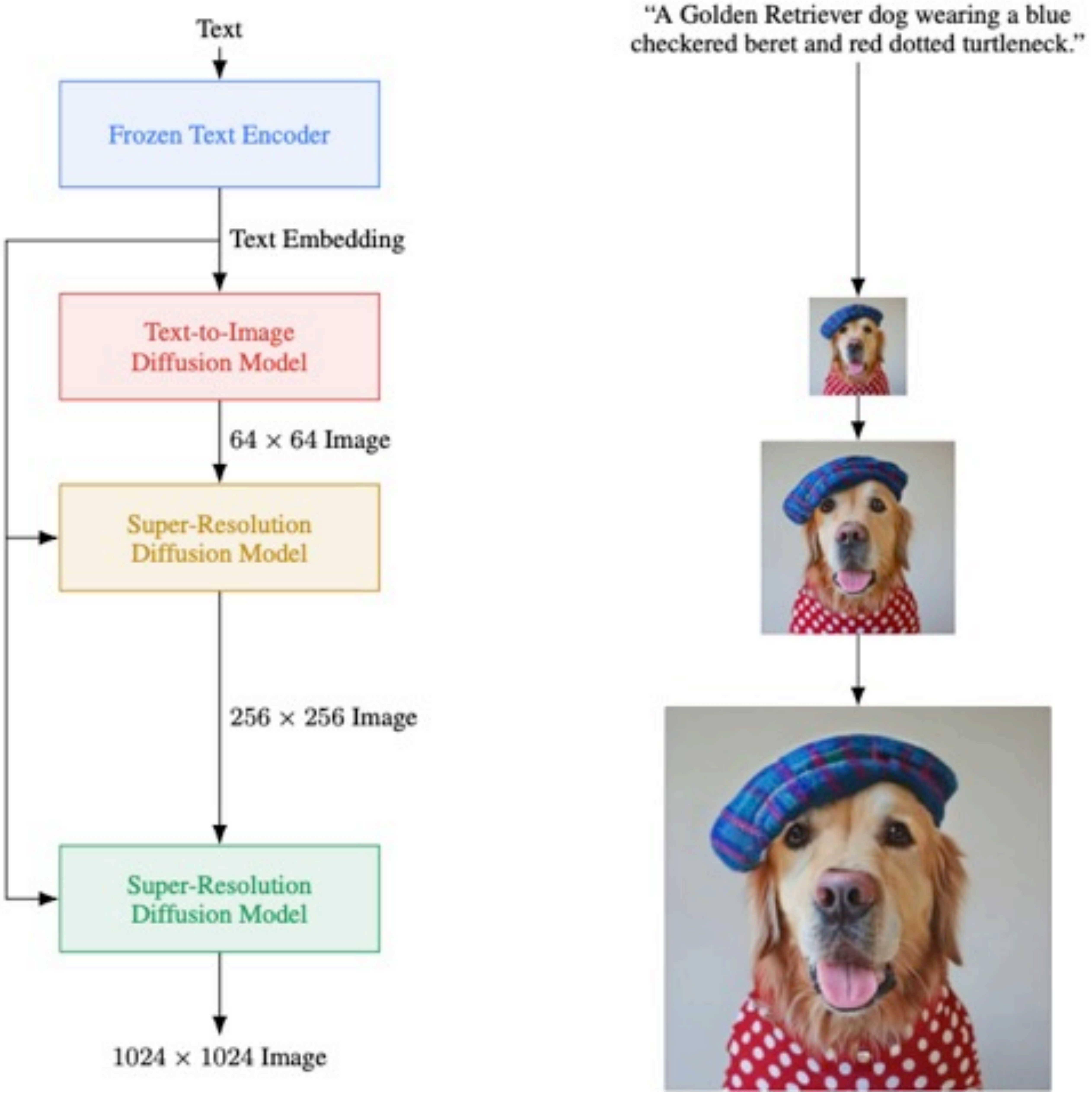
A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

C. Sharia et al. [Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding](). NeurIPS 2022

Imagen

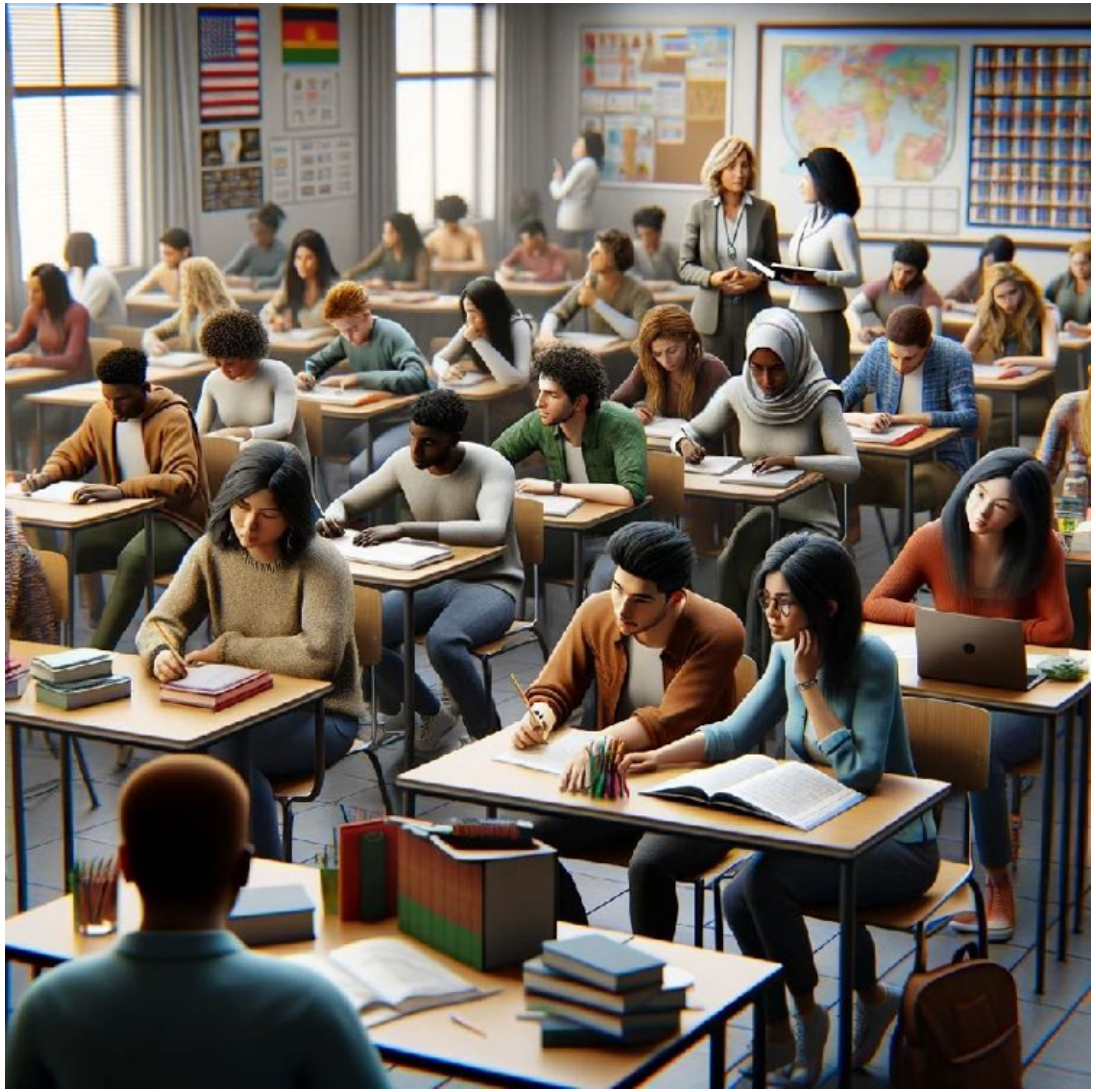
Cascade of conditional diffusion models



C. Sharia et al. [Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding](). NeurIPS 2022

DALL-E-3

"Classroom full of students"



Betker et al. [Improving Image Generation with Better Captions](). 2023

DALL-E-3

"*People dancing*"

"For DALL-E 3, we trained our own diffusion decoder on top of the latent space learned by the VAE trained by Rombach et al. (2022). We found that using a diffusion decoder here provided marked improvements to fine image details, for example text or human faces."

Betker et al. [Improving Image Generation with Better Captions](). 2023

DALL-E-3

"Photo of
Paris"

Car or boat?

Street or river?

Betker et al. [Improving Image Generation with Better Captions](). 2023

…

Stable Diffusion

Latent Diffusion

Dec 2021 / Aug 2022 (CVPR'22)

Imagen

Cascaded Diffusion

May 2022 (NeurIPS'22)

…

DALL-E

VQ-VAE + Transformers

Feb 2021 (ICML'21)

…

DALL-E-2

Diffusion + CLIP latents

Apr 2022

…

DALL-E-3

Latent Diffusion + Better captions + Other undocumented improvements

Oct 2023

[DALL-E] A. Ramesh et al., Zero-Shot Text-to-Image Generation, ICML 2021

[StableDiffusion] R. Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022

[Imagen] C. Sharia et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. NeurIPS 2022

[DALL-E-2] A. Ramesh et al. Hierarchical text-conditional image generation with CLIP latents. 2022

[DALL-E-3] Betker et al. Improving Image Generation with Better Captions. 2023

…

Agenda

1. Generative neural networks

- VAE: Variational autoencoders
- GAN: Generative adversarial networks
- Diffusion models

2. Vision & language

- Text-to-image retrieval
- Text-to-image generation
- Image captioning

Parenthesis:
(Language Models

(Large) Language models (LLMs)

[GPT] Radford, Narasimhan, Salimans, Sutskever, Improving Language Understanding by Generative Pre-Training, 2018

Before: RNNs, Supervised *GPT: Transformers, Unsupervised*

[BERT] Devlin, Chang, Lee, Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL 2019

Before: Autoregressive encoder-decoder generation *BERT: Non-autoregressive, encoder-only, masked modeling*

[GPT-2] Radford, Wu, Child, Luan, Amodei, Sutskever, Language Models are Unsupervised Multitask Learners, 2019

***1.5B parameter** Transformer + a new dataset of millions of webpages (WebText), SOTA zero-shot results on 7/8 datasets, still underfits WebText*

[T5] Raffel, Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li, Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, JMLR 2020.

***11 billion parameters,** survey-like controlled study, CommonCrawl data*

[GPT-3] Brown, Mann, Ryder, Subbiah, … Radford, Sutskever, Amodei, Language Models are Few-Shot Learners, NeurIPS 2020

***175 billion parameters**, 10x more than any previous non-sparse language model, **trained on 400B tokens** from CommonCrawl data*

[GPT-4] [LlaMa] [LlaMa-2] …

(Large) Language models (LLMs)

[GPT] Radford, Narasimhan, Salimans, Sutskever, Improving Language Understanding by Generative Pre-Training, 2018

Before: RNNs, Supervised *GPT: Transformers, Unsupervised*

Transformers

[BERT] Devlin, Chang, Lee, Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL 2019

Before: Autoregressive encoder-decoder generation *BERT: Non-autoregressive, encoder-only, masked modeling*

[GPT-2] Radford, Wu, Child, Luan, ... pervised Multitask Learners, 2019

***1.5B parameter** Transformer + a new* ... *-shot results on 7/8 datasets, still underfits WebText*

Mostly unsupervised, e.g., next word prediction

[T5] Raffel, Shazeer, Roberts, Lee, N... ts of Transfer Learning with a Unified Text-to-Text Transformer, JMLR 2020.

***11 billion parameters**, survey-like controlled study, CommonCrawl data*

[GPT-3] Brown, Mann, Ryder, Subbiah, ... Radford, Sutskever, Amo... rners, NeurIPS 2020

***175 billion parameters**, 10x more than any previous non-sparse language m*... *Crawl data*

Models & data getting bigger

[GPT-4] [LlaMa] [LlaMa-2] ...

(a) Datasets used to train GPT-3

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

"Datasets for language models have rapidly expanded, culminating in the Common Crawl dataset [RSR+19] constituting nearly a trillion words. This size of dataset is sufficient to train our largest models ***without ever updating on the same sequence twice****."*

(b) Total Compute Used During Training

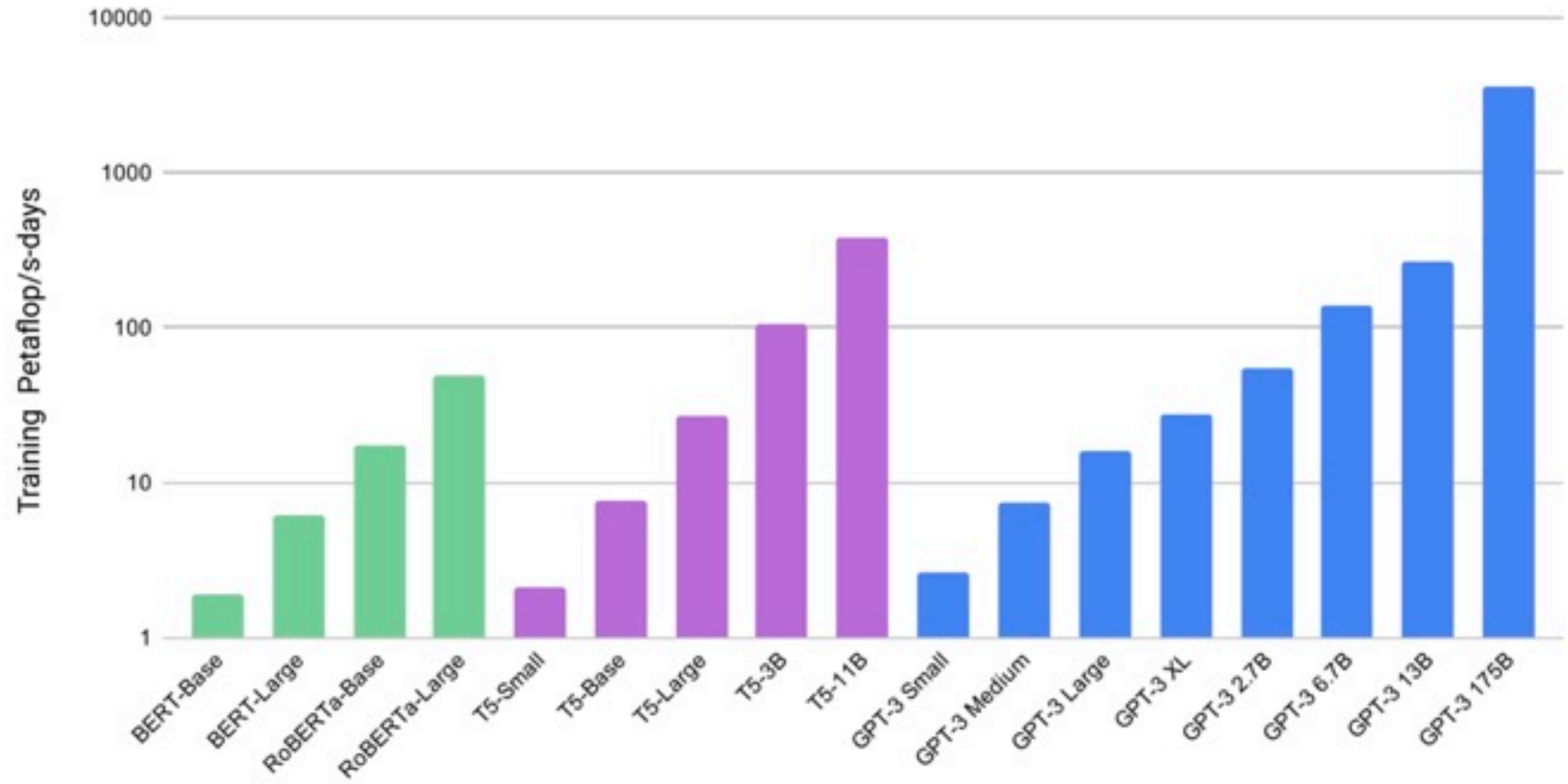


Figure 2.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH+20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in Appendix D.

Model	Total train compute (PF-days)	Total train compute (flops)	Params (M)	Training tokens (billions)
T5-Small	2.08E+00	1.80E+20	60	1,000
T5-Base	7.64E+00	6.60E+20	220	1,000
T5-Large	2.67E+01	2.31E+21	770	1,000
T5-3B	1.04E+02	9.00E+21	3,000	1,000
T5-11B	3.82E+02	3.30E+22	11,000	1,000
BERT-Base	1.89E+00	1.64E+20	109	250
BERT-Large	6.16E+00	5.33E+20	355	250
RoBERTa-Base	1.74E+01	1.50E+21	125	2,000
RoBERTa-Large	4.93E+01	4.26E+21	355	2,000
GPT-3 Small	2.60E+00	2.25E+20	125	300
GPT-3 Medium	7.42E+00	6.41E+20	356	300
GPT-3 Large	1.58E+01	1.37E+21	760	300
GPT-3 XL	2.75E+01	2.38E+21	1,320	300
GPT-3 2.7B	5.52E+01	4.77E+21	2,650	300
GPT-3 6.7B	1.39E+02	1.20E+22	6,660	300
GPT-3 13B	2.68E+02	2.31E+22	12,850	300
GPT-3 175B	3.64E+03	3.14E+23	174,600	300

[GPT-3] Brown, Mann, Ryder, Subbiah, Kaplan, … Radford, Sutskever, Amodei, Language Models are Few-Shot Learners

Reviews of the GPT-3 paper

https://papers.nips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Review.html

Review 4

Summary and Contributions: The paper shows scaling up language models can achieve task-agnostic few-shot performances on various NLP tasks. Besides other promising results on various tasks and examples, this paper has a clear contribution to the community; industry-level, heavy engineering efforts, and their analyses on various aspects. I do appreciate such efforts and empirical findings described in the paper.

Strengths: The paper has the following strengths: (1) A comprehensive analysis has been made to evaluate the model in terms of task-agonistic behavior, memorization, and weakness based on the prior works/critiques made on the previous version of this work. (2) The empirical observations about the few-shot models' capabilities are made (Figure 1.1 and Section 3.4), showing the scaling effects, limitations, and upper-bound of few-shot settings. (3) Experiments on different tasks in various applications indicate how much existing datasets/tasks are getting benefit from these huge-size language models and what kinds of operations (e.g., bidirectional, reasoning, external knowledge) should be done in the future toward that direction. (4) I pretty much enjoyed reading the Broader Impact section, which tries to adopt the feedbac
fairness and bias, energy usage, and potential misuse of the model. (5) For a perspective of usefulness, I think this c

Weaknesses: Improvements from few shot LMs are not that surprising because it is mainly because the model uses
not fair to compare with relatively smaller models, the improvements themselves may not be the main contribution of
engi
coul
few
artifi
rathe
Compared to GPT2, the only difference made in this work is scaling up the training in terms of data size and compari
one-shot, and zero-shot. What scientific values does this paper bring to our community except for empirical observa
on NaturalQS shows GPT3 mean that it does not include any external knowledge in Wikipedia and their appropriate r
Figure 1.1., let's say you use 175B x 1000 parameters, do you think the improvement from {zero,one,few}-shots still li
of contexts increases, the degree of improvements from the few-shot GPT3 seems to be not that steep. Does this in
around K=100 or something? Also, please show me the zero/one-short cases as well.

Correctness: Please see many comments above.

Clarity: Yes, the paper is written well and easy to follow.

Relation to Prior Work: Yes.

Reproducibility: Yes

> Improvements … not that surprising … mainly because
> … more training data/parameters/computing resources.
> … an empirical paper with huge engineering efforts

Review 2

Summary and Contributions: In this paper, the authors empirically demonstrate that increasing the model size -- in term of depth and width, and thus number of parameters -- of language models (LM) result in a better task-agnostic learner, which can zero/one/few-shots multiple well-know NLP tasks. - The authors use the same transformer base architecture as GPT-2, except for the Sparse attention (Child, et.al. 2019), which improve the model efficiency. They trained 8 models from 125 M to 175 B parameters to study the effect of the model size in the zero/one/few-shots settings. - The authors train the LM using 300 billion tokens from 5 sources (i.e., Common Crawl, WebText2, Book Collection 1 and 2, Wikipedia). - The authors evaluate the models' performance in a zero/one/few-shot setting on a large variety of NLP tasks such as LM perplexity, QA, CQA, SuperGLUE, MT, etc. Importantly, the zero/one/few-shots is done without fine-tuning the model, but by providing as context --priming-- the task-description (i.e., for zero-shot) or pairs of examples (one/few-shots), and making the model auto-regressively generate the response. As also clearly stated by the authors, this approach is not novel, since also GPT-2 used the same mechanism, but in this paper, the author extended the evaluation to way more tasks and showed that by increasing the model size the few-shot ability of the model greatly increases. - The authors compare the performance of the model to the current state-of-the-art and they highlight the advantages and disadvantages of the proposed model. I really appreciated the openness of the
authors when they described their results, avoiding st … and clear for a large
audience and the broader impact of the paper is clea … model. I am personally
not an expert in the bias/ethical ML, but in my perspe … race, gender).

Strengths: - the zero/one/few-shots methodology is … arios. - the evaluation includes a large variety of NLP tasks and SOTA baselines, and support the claim of the paper - The paper include an human evaluation over news article generation showing that human found hard (52% accuracy) to recognise which article is written by humans or the GPT-3 model (175B)

Weaknesses: - the authors already discussed most of the limitation of the current model (e.g. missing of bi-directional attection etc.). I found that one limitation could be the length of the context when increasing the number of shots. To elaborate, in some tasks (e.g. QA, summurization) where the input are entire articles, going beyond the 25/30 shots would be very challenging. GPT-3 already double (2048 tokens) the context size compare to GPT-2, but scaling very long inputs remains challenging, both in term of memory consumption and models inference (although the authors already use Sparse Transformer).

Correctness: Yes.

Clarity: Yes.

Relation to Prior Work: yes, to the best they could do in 8 pages. I think the citation format is not the NeurIPS 2020 template, but this can be easily change in the camera ready.

Reproducibility: Yes

Additional Feedback:

> although not novel per se (i.e. GPT-2),
> … can have a big impact

[GPT-3] Brown, Mann, Ryder, Subbiah, Kaplan, … Radford, Sutskever, Amodei, Language Models are Few-Shot Learners

Parenthesis Closed:
Language Models)

Image Captioning: Image in, Text out

A politician receives a gift from politician.

A collage of different colored ties on a white background.

Silhouette of a woman practicing yoga on the beach at sunset.

Aerial view of a road in autumn.

Mokady et al., [ClipCap: CLIP Prefix for Image Captioning](), arXiv 2021

ClipCap: CLIP Prefix for Image Captioning

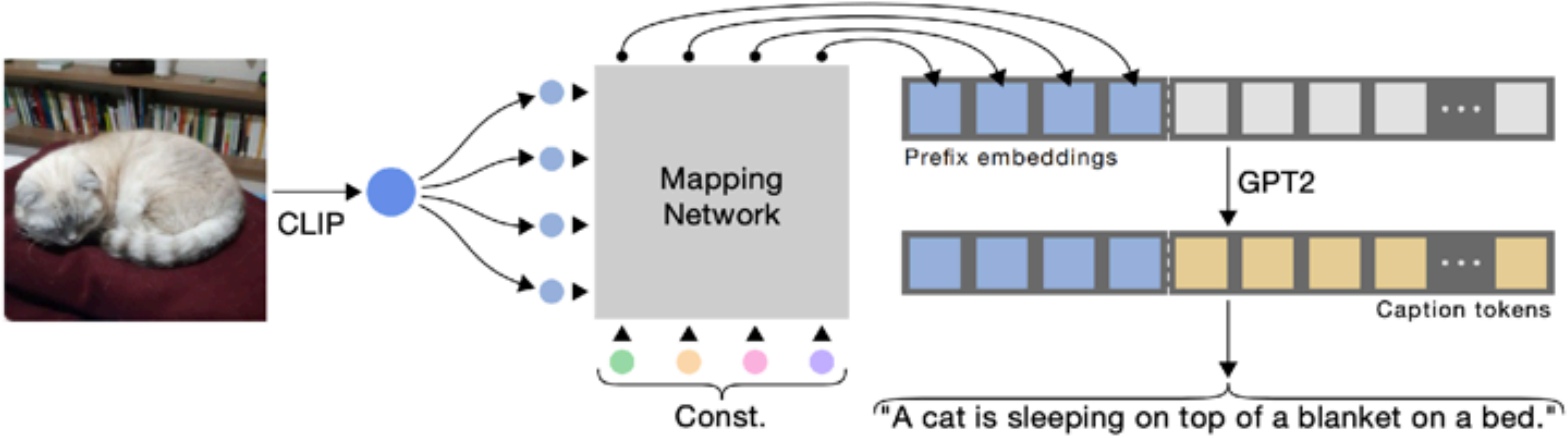
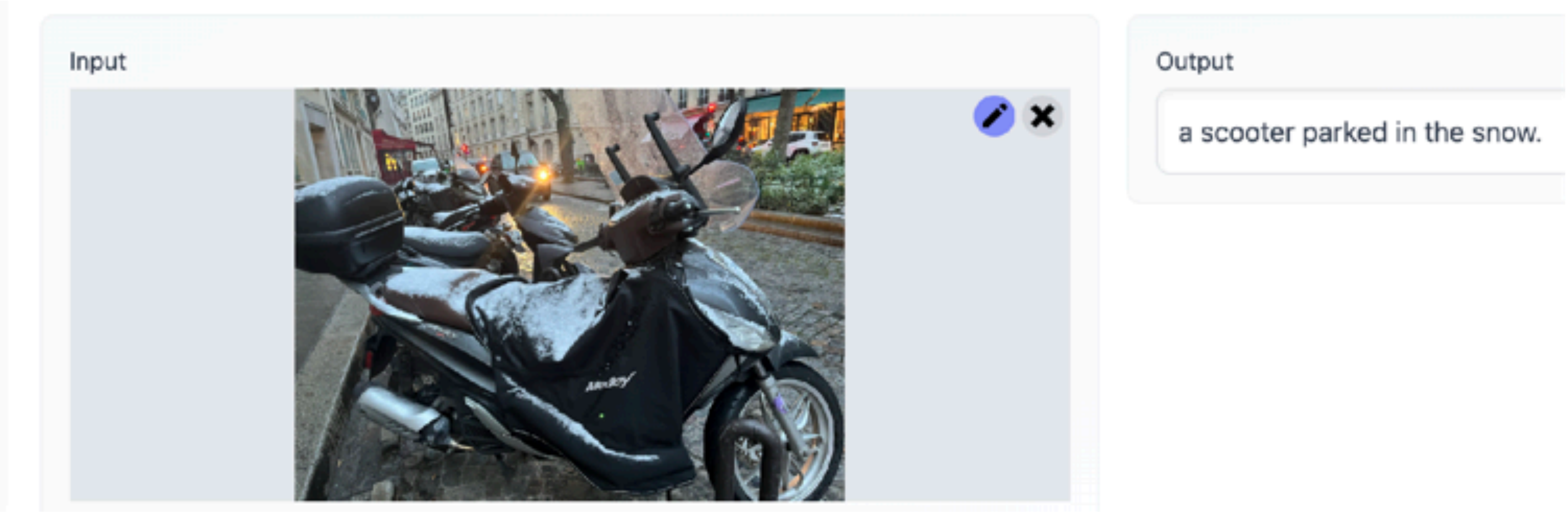


Figure 2. Overview of our transformer-based architecture, enabling the generation of meaningful captions while both CLIP and the language model, GPT-2, are frozen. To extract a fixed length prefix, we train a lightweight transformer-based mapping network from the CLIP embedding space and a learned constant to GPT-2. At inference, we employ GPT-2 to generate the caption given the prefix embeddings. We also suggest a MLP-based architecture, refer to Sec. 3 for more details.

Mokady et al., [ClipCap: CLIP Prefix for Image Captioning](), arXiv 2021

ClipCap: CLIP Prefix for Image Captioning

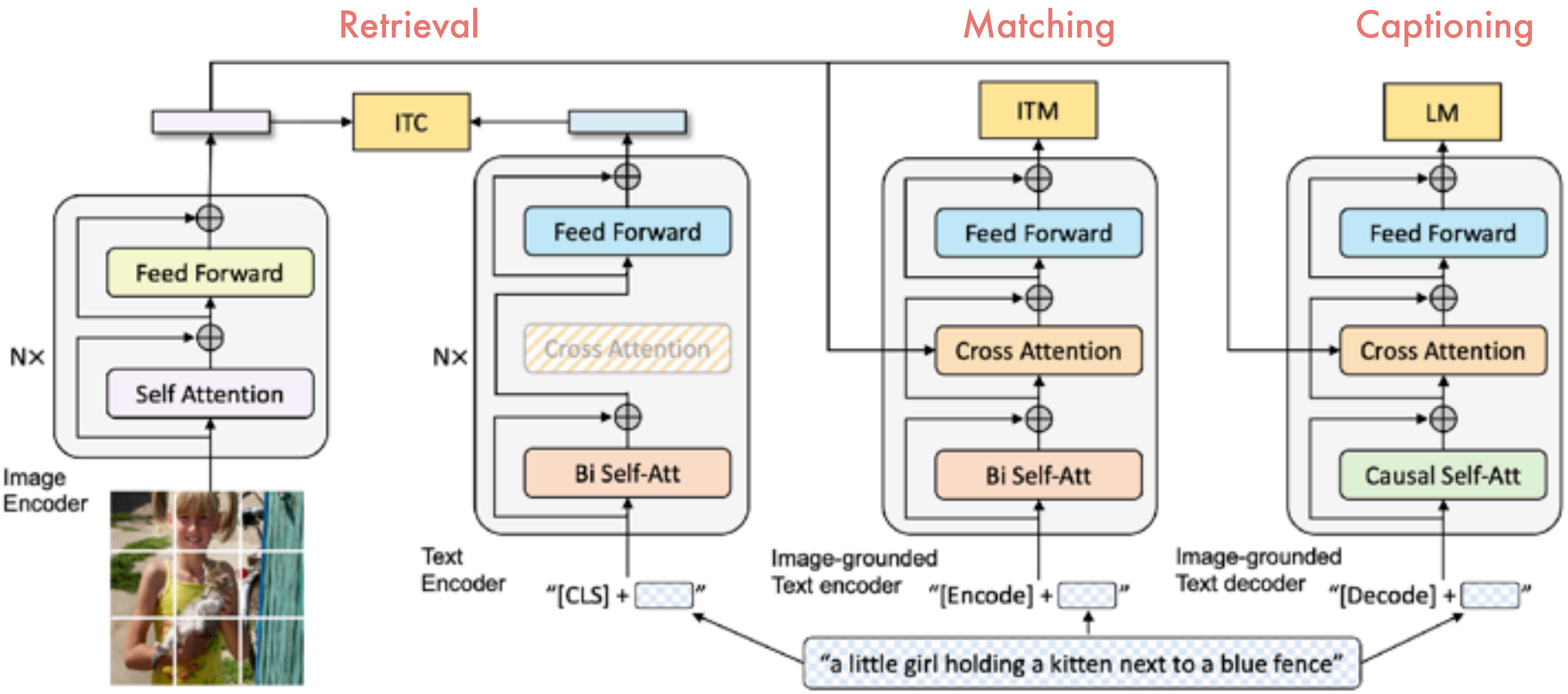
- Demo: https://huggingface.co/spaces/akhaliq/CLIP_prefix_captioning



Mokady et al., [ClipCap: CLIP Prefix for Image Captioning](), arXiv 2021

BLIP

Retrieval + Captioning

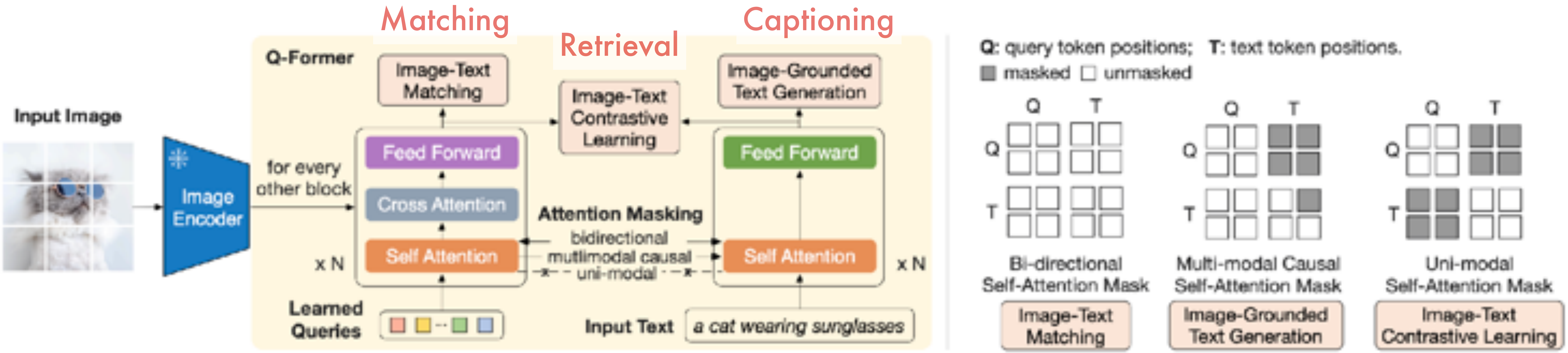


Li et al., [BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation](), ICML 2022

BLIP-2

Retrieval + Captioning with "Q-Former" architecture

Stage 1 training (similar to BLIP-1): **vision-text representation learning**



BLIP-2

Retrieval + Captioning with "Q-Former"

Stage 2 training with a frozen LLM to **generate text**

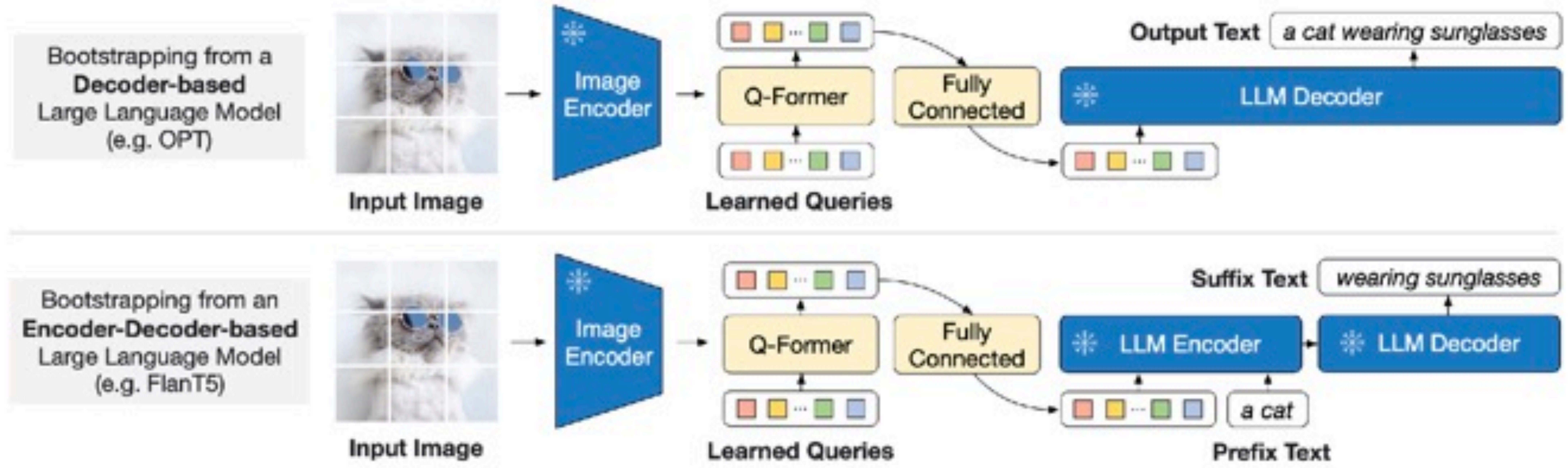


Figure 3. BLIP-2's second-stage vision-to-language generative pre-training, which bootstraps from frozen large language models (LLMs). (**Top**) Bootstrapping a decoder-based LLM (e.g. OPT). (**Bottom**) Bootstrapping an encoder-decoder-based LLM (e.g. FlanT5). The fully-connected layer adapts from the output dimension of the Q-Former to the input dimension of the chosen LLM.

Li et al., BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, ICML 2023

BLIP1 & BLIP2 Training Data

Similar to CLIP
129M image-text pairs
with automatic captioning + filtering

- Visual Genome
- CC3M*
- CC12M*
- COCO
- LAION400M (115M)*
- SBU*

*web datasets

Li et al., BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, ICML 2022
Li et al., BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, ICML 2023

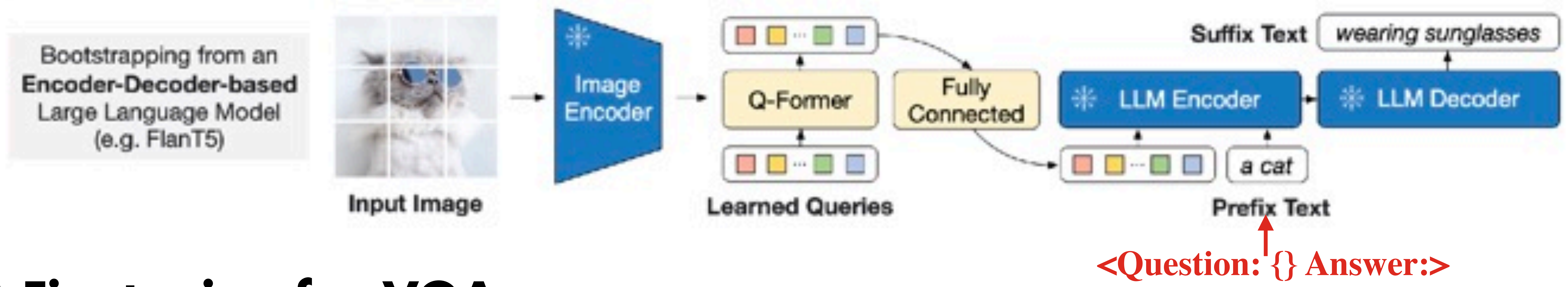
Bonus: VQA in 1 slide

Vision & language

- Text-to-image retrieval
- Text-to-image generation
- Image captioning
- Visual question answering (VQA)

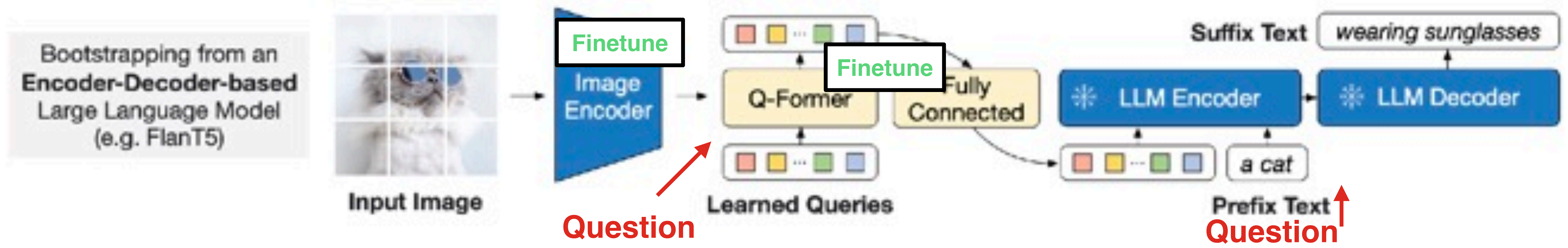
BLIP2 for Visual Question Answering

(a) Zero-shot (no finetuning for VQA)



(b) Finetuning for VQA

"Given annotated VQA data, we finetune the parameters of the Q-Former and the image encoder while keeping the LLM frozen. … LLM receives Q-Former's output and the question as input, and is asked to generate the answer.
In order to extract image features that are more relevant to the question, we additionally condition Q-Former on the question."



Li et al., [BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models](), ICML 2023

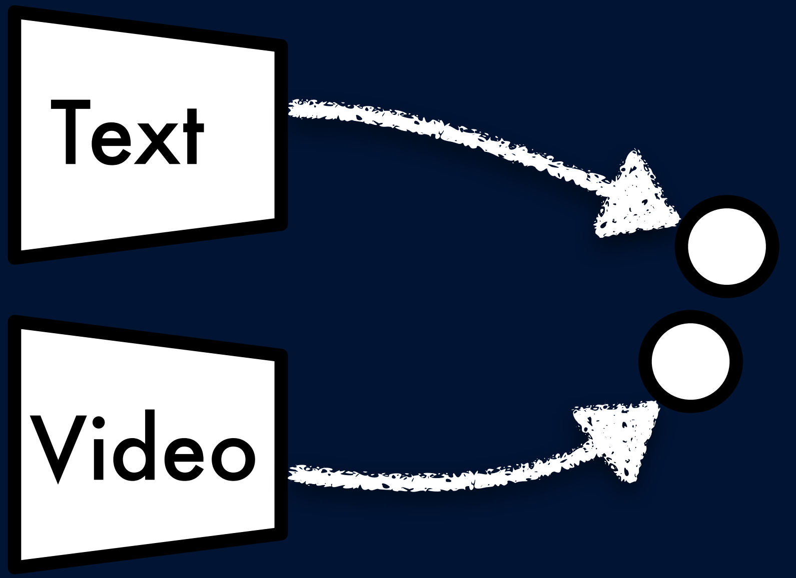
Bonus: Examples from our works

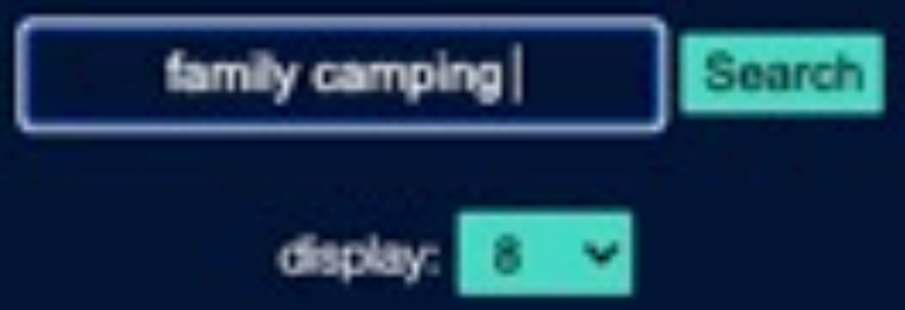
Retrieval tasks

Text Queries for Search:

[Bain, Nagrani, Varol, Zisserman, "Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval", ICCV 2021]

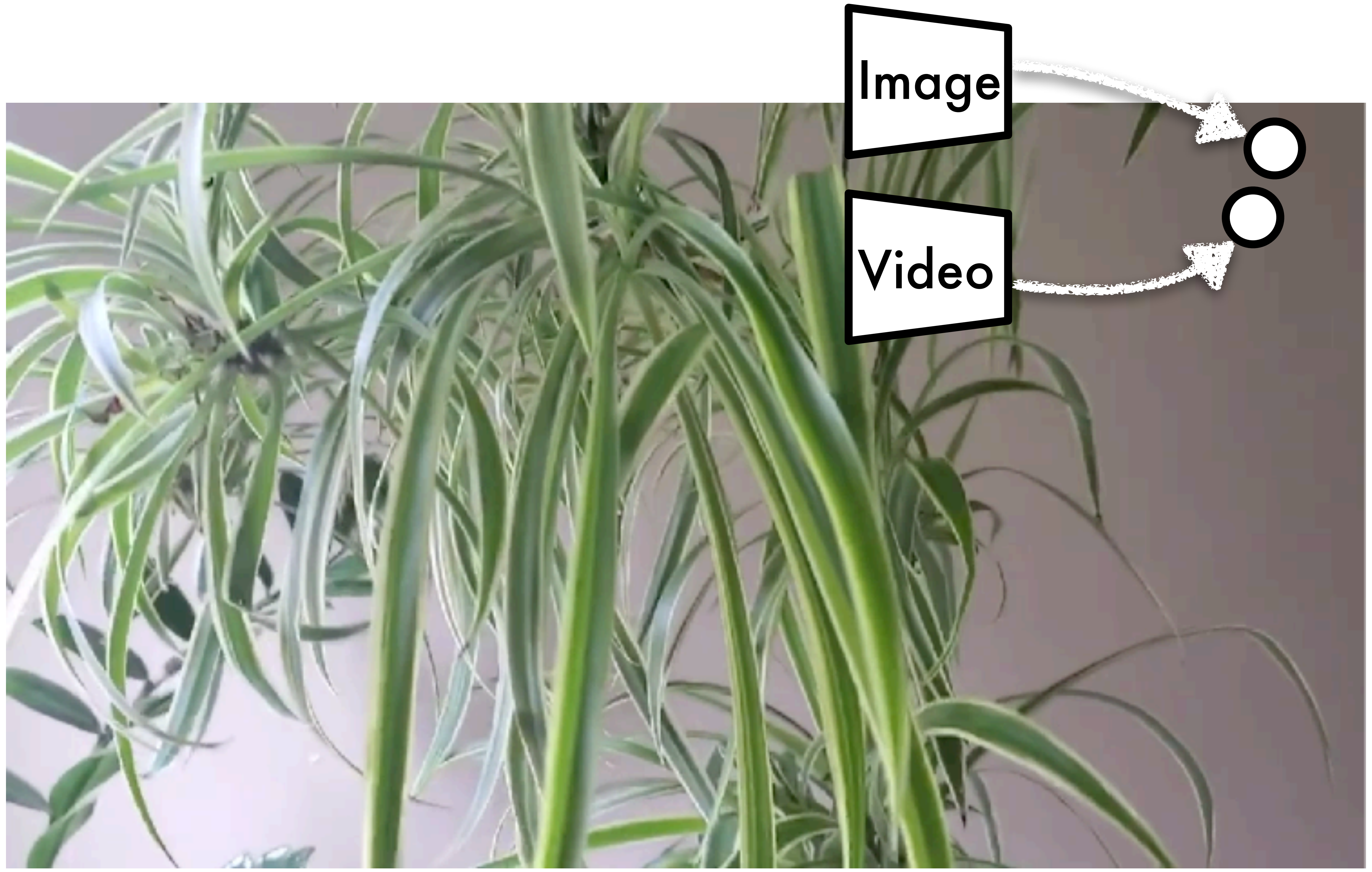
https://www.robots.ox.ac.uk/~vgg/research/frozen-in-time/





Efficient search with https://github.com/facebookresearch/faiss

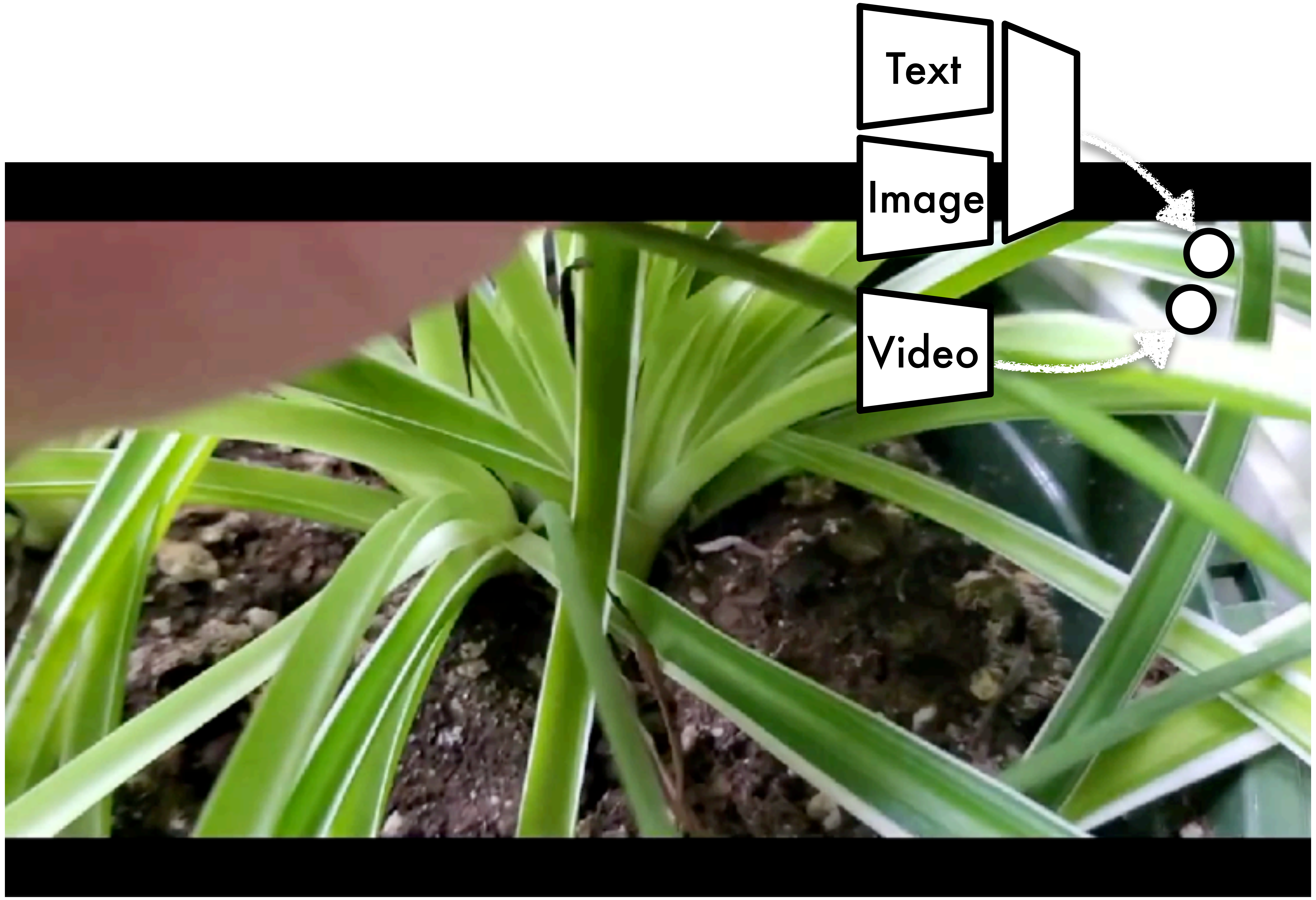
Image Queries for Search:



Spoiler: *Chlorophytum comosum* (aka "spider plant")

Image+Text Queries for Search:

"prune this plant"

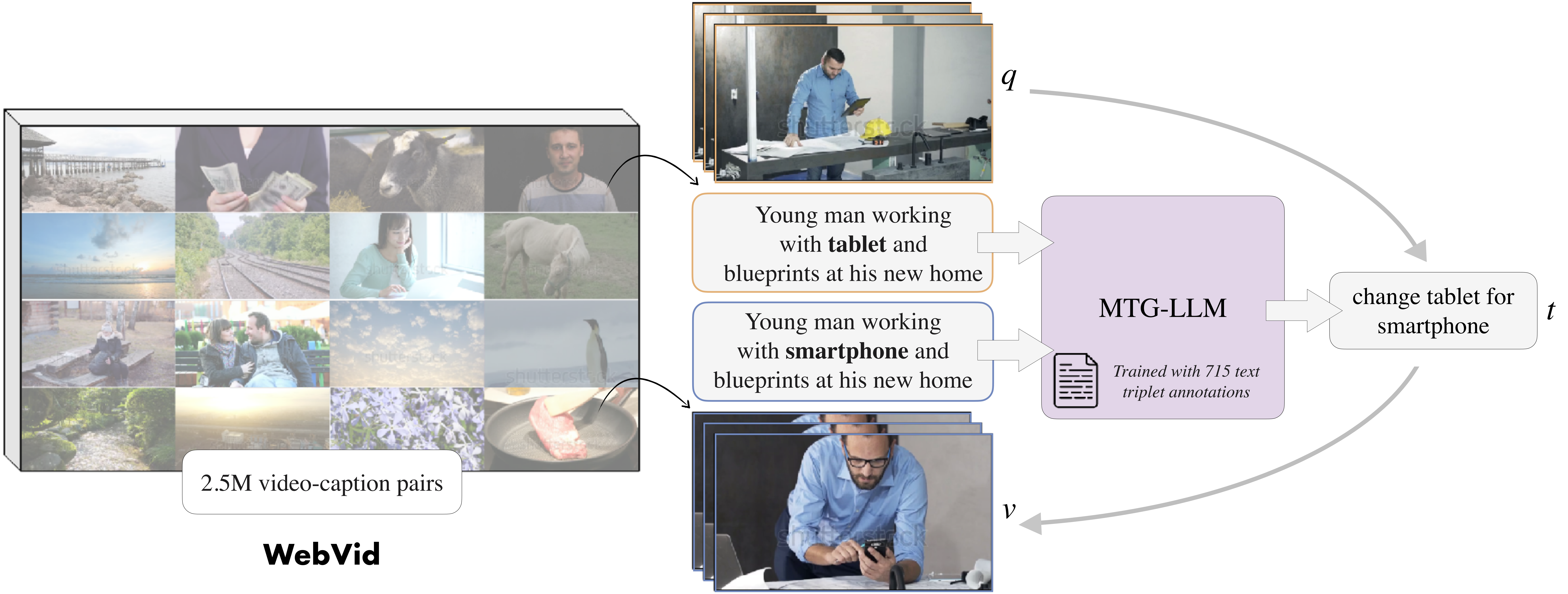


Ventura, Yang, Schmid, Varol, [CoVR: Learning Composed Video Retrieval from Web Video Captions](), 2023

Problem: Data? No annotated image-text-video triplets.

Ventura, Yang, Schmid, Varol, CoVR: Learning Composed Video Retrieval from Web Video Captions, 2023

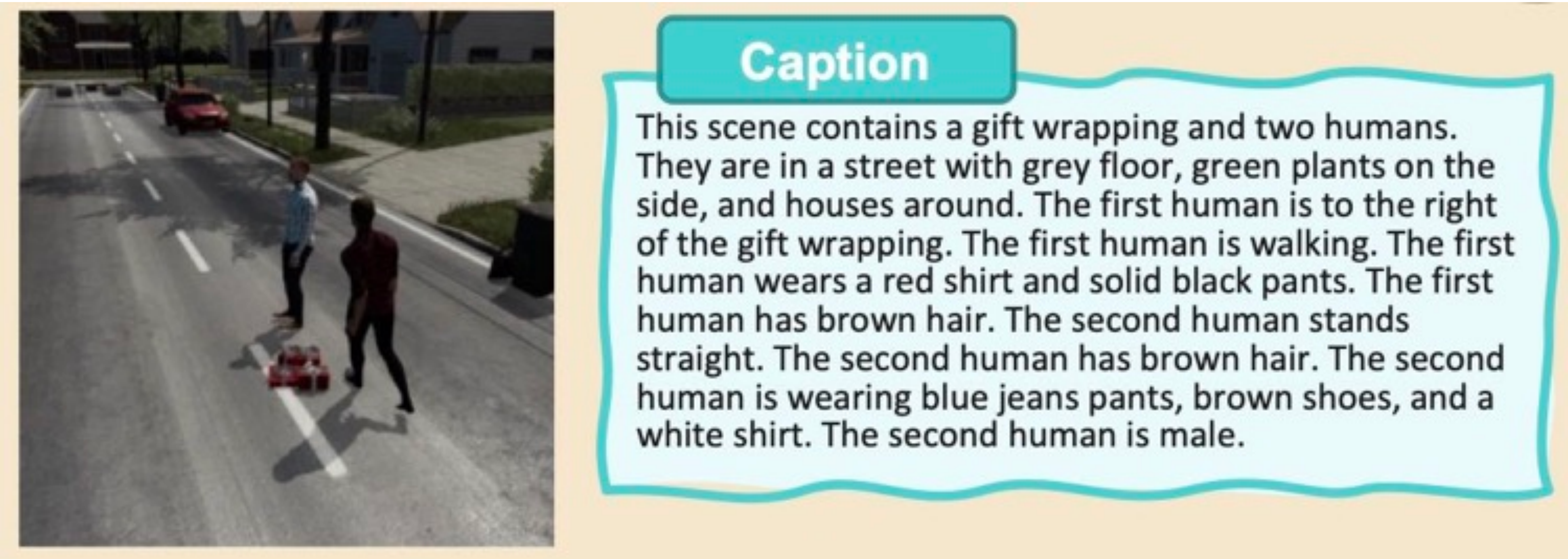
Mining similar caption pairs & Modification text generation (MTG)



["WebVid", Bain, Nagrani, Varol, Zisserman, ICCV 2021]

Ventura, Yang, Schmid, Varol, [CoVR: Learning Composed Video Retrieval from Web Video Captions](), 2023

Training Data Augmentation with Synthetic Renderings



Cascante-Bonilla, Shehada, Smith, Doveh, Kim, Panda, Varol, Oliva, Ordonez, Feris, Karlinsky,
Going Beyond Nouns With Vision & Language Models Using Synthetic Data, ICCV 2023

This scene contains a car tire, a television set, a stool, a microwave, a car tire, and one human. They are in a street with grey floor, green plants on the side, and houses around. The human is to the right of the stool. The car tire is in front of the television set. The microwave is to the left of the television set. The television set is behind the stool. The human is to the right of the car tire. The car tire is behind the stool. The car tire is in front of the television set. The stool is in front of the microwave. The human is behind the stool. The microwave is to the left of the stool. The car tire is to the right of the microwave. The human is to the right of the television set. The car tire is to the left of the human. The stool is in front of the car tire. The car tire is to the right of the car tire. The human is in front of the television set. The television set is to the right of the car tire. The car tire is to the left of the stool. The microwave is in front of the television set. The human is to the right of the microwave. The car tire is to the right of the stool. The human straight jump with full twist. The human is bald. The human is male. The human wears dark blue jeans. The human is clothed in a blue hoodie with a white logo on the front.

Vision-to-text tasks

Movie Audio Description Generation

Audio Description (AD) = Narration describing visual elements in the movie to aid the visually impaired

Movie clip from `Out of Sight' (1998) **with Audio Description**

Han, Bain, Nagrani, Varol, Xie, Zisserman, [AutoAD: Movie Description in Context](), CVPR 2023

Han, Bain, Nagrani, Varol, Xie, Zisserman, [AutoAD II: The Sequel – Who, When, and What in Movie Audio Description](), ICCV 2023

Sign Language Translation: Video in, Text out

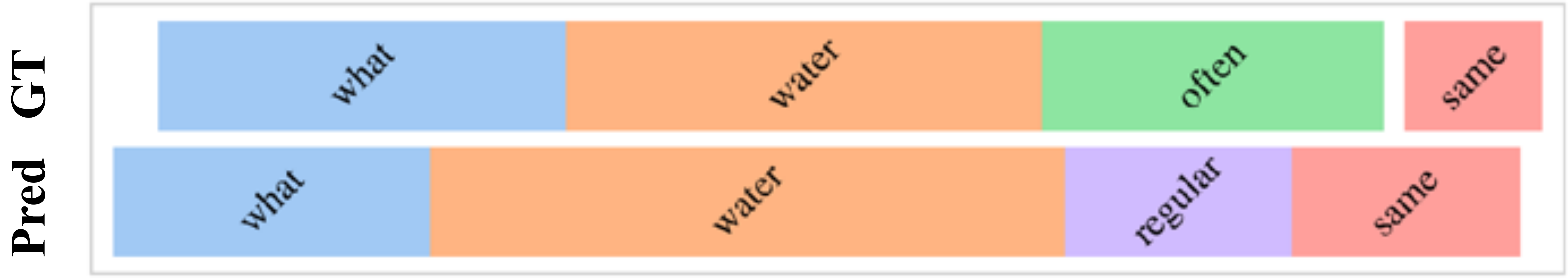
Doesn't really work yet

"Every Spring, our planet is transformed"

with signs:

(EVERY; SPRING; OUR; PLANET; **HAPPEN; WHAT;** *TRANSFORM)*

Sign Language Transcription (aka Continuous SL Recognition)



Word Error Rate: 25.0

Subtitle (unused): The majority of sweat is just straightforward water.

Word Error Rate: 42.9

Subtitle (unused): We were worried that it might happen again, falling on a primary school, someone's home, or a playground.

Raude, Prajwal, Momeni, Bull, Albanie, Zisserman, Varol, A Tale of Two Languages: Large-Vocabulary Continuous Sign Language Recognition from Spoken Language Supervision (WIP)

Text-to-vision tasks

Text-to-3D human motion

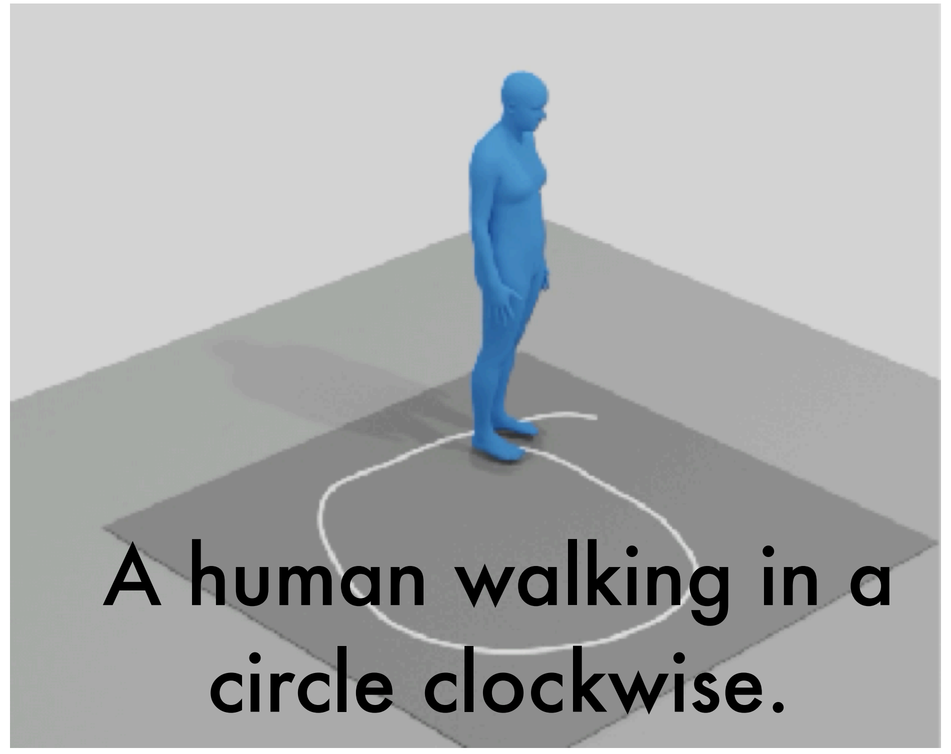
Motion synthesis conditioned on:

Actions

✦ **ACTOR**: <u>Action</u>-conditioned synthesis

[Petrovich, Black, Varol. ICCV'21]

Pickup



Descriptions

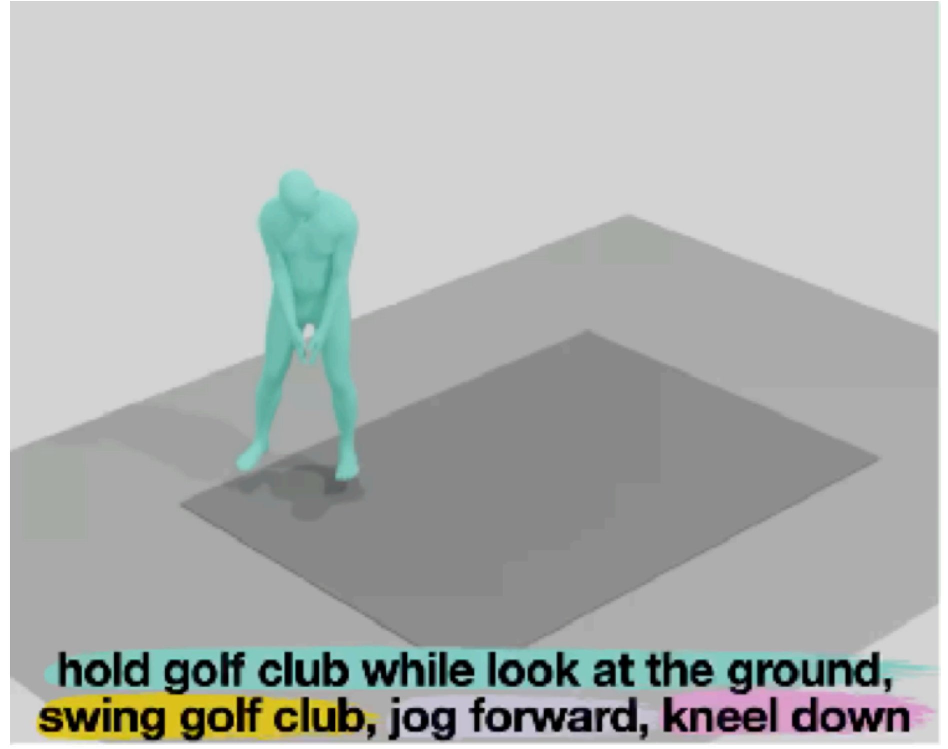
✦ **TEMOS**: <u>Text</u>-conditioned synthesis

[Petrovich, Black, Varol. ECCV'22]

Sequence of descriptions

✦ **TEACH**: <u>Temporal compositionality</u> for sequential actions

[Athanasiou, Petrovich, Black, Varol. 3DV'22]



Set of descriptions

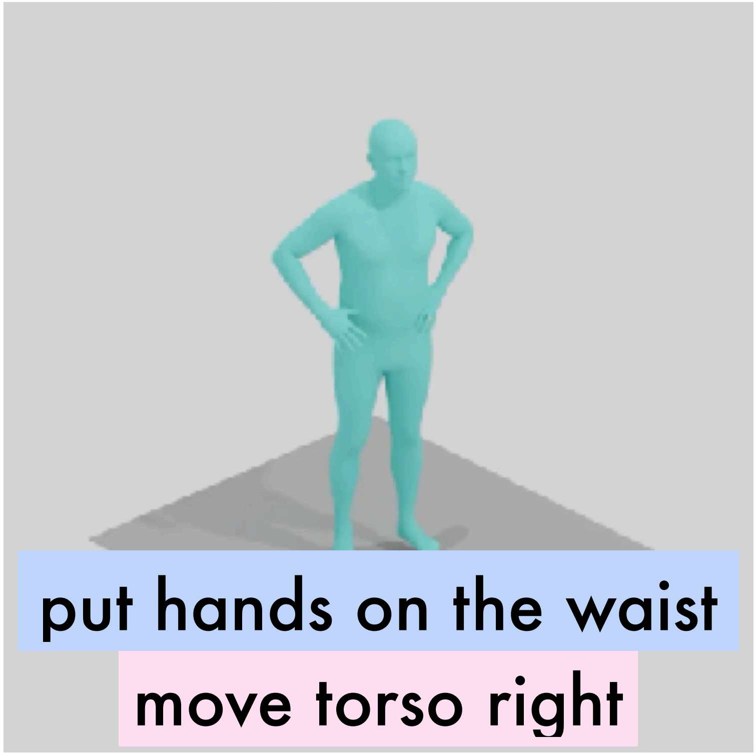
✦ **SINC**: <u>Spatial compositionality</u> for simultaneous actions

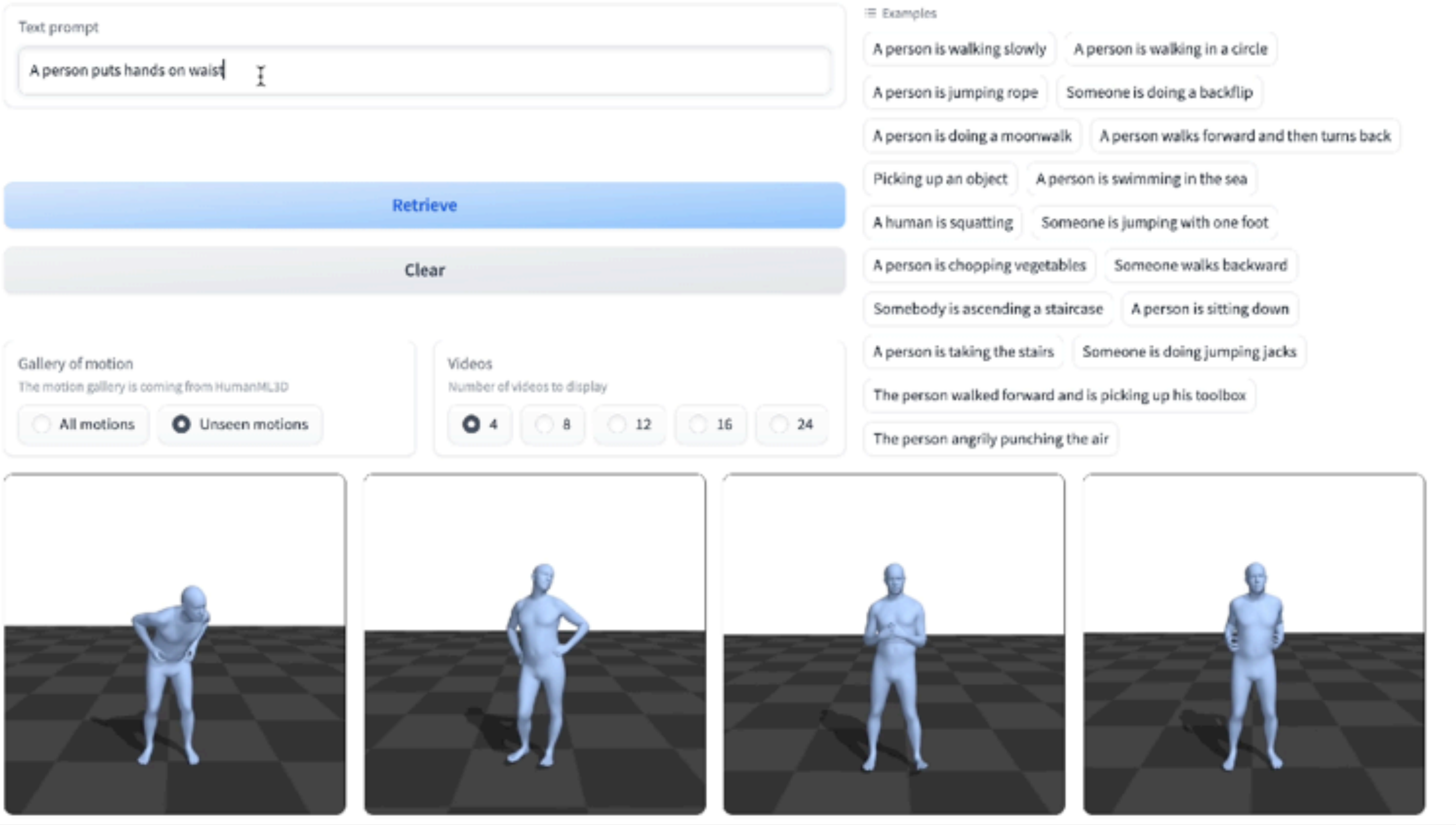
[Athanasiou*, Petrovich*, Black, Varol. ICCV'23]

Motion retrieval

✦ **TMR**: Text-to-motion <u>retrieval</u>, i.e., CLIP for motion-text

[Petrovich, Black, Varol. ICCV'23]





Motion retrieval

✦ **TMR**: Text-to-motion retrieval, i.e., CLIP for motion-text

[Petrovich, Black, Varol. ICCV'23]

https://mathis.petrovich.fr/tmr/

Agenda

1. Generative neural networks

- VAE: Variational autoencoders
- GAN: Generative adversarial networks
- Diffusion models

2. Vision & language

- Text-to-image retrieval
- Text-to-image generation
- Image captioning
- Bonus: Visual question answering (VQA)
- Bonus: Examples from our works