PAT 498/598 (Fall 2024)

Special Topics:
Generative AI for Music and Audio Creation

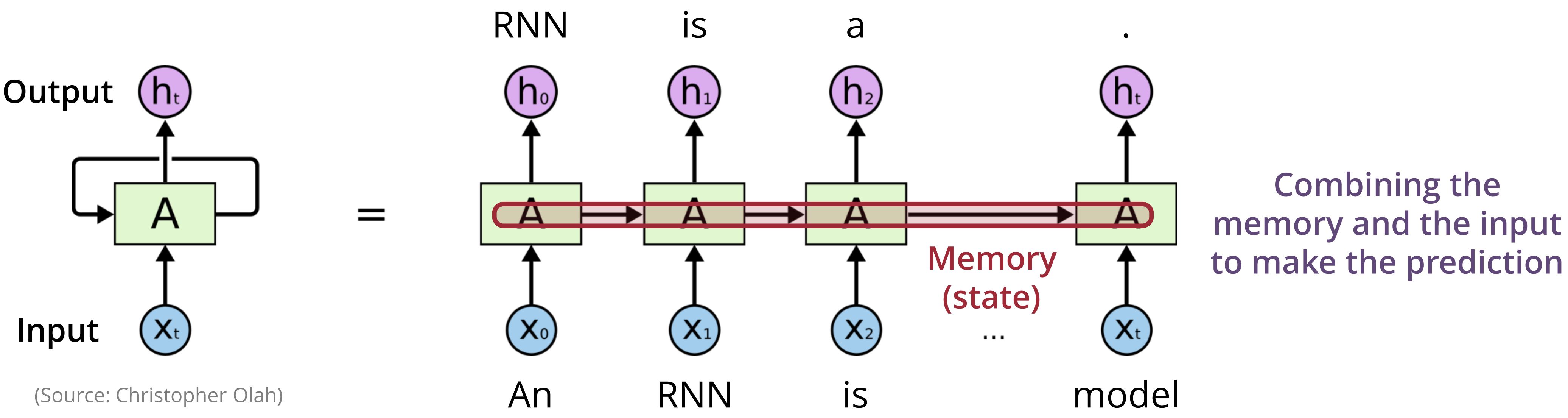
Lecture 9: VAEs & GANs

Instructor: Hao-Wen Dong

SCHOOL OF MUSIC, THEATRE & DANCE
PERFORMING ARTS TECHNOLOGY
UNIVERSITY OF MICHIGAN

(Recap) What is an RNN (Recurrent Neural Network)?

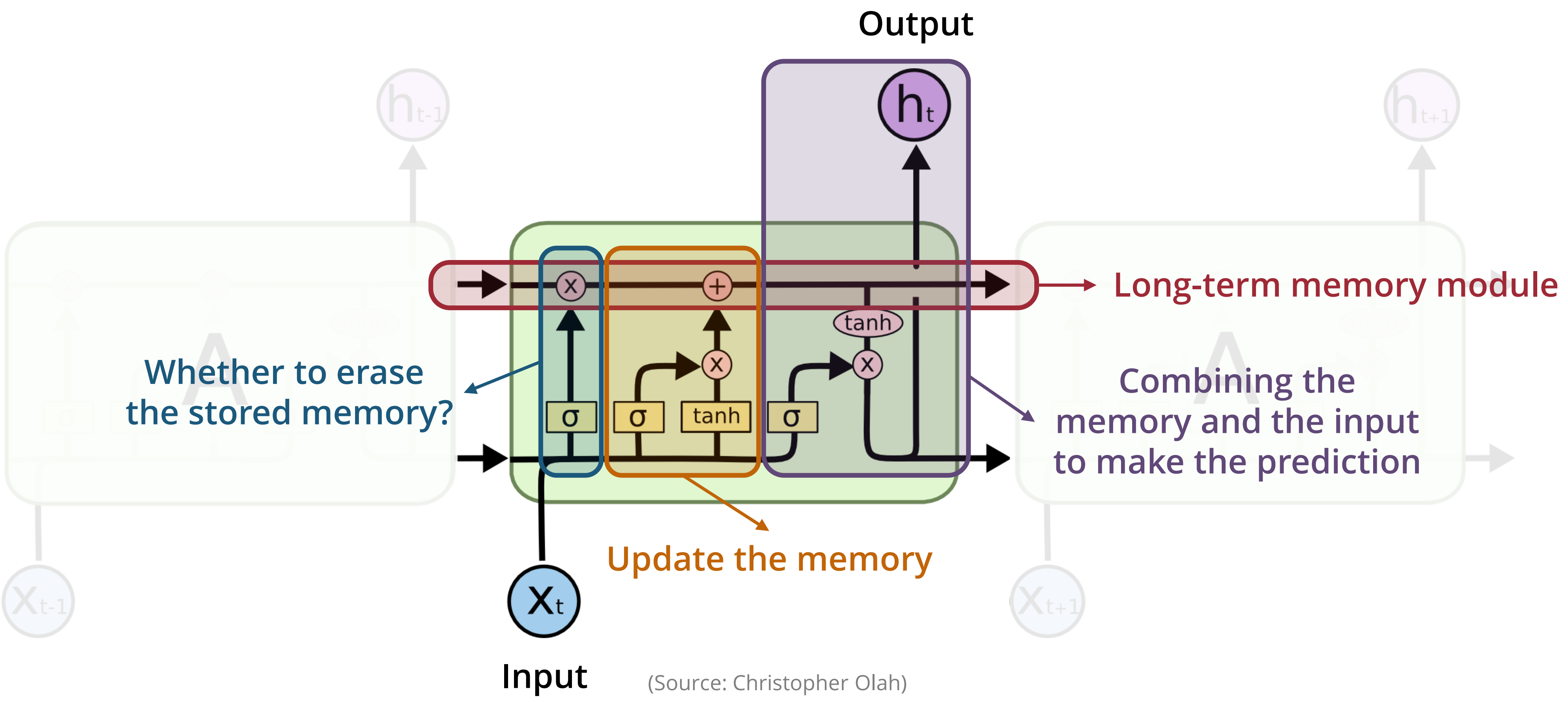
- A type of neural networks that have **loops**
- Widely used for **modeling sequences** (e.g., in natural language processing)



(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs/

(Recap) Demystifying LSTMs



(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs/

(Recap) Demystifying Transformers

A transformer is a ________

electrical device

fiction character

deep learning model

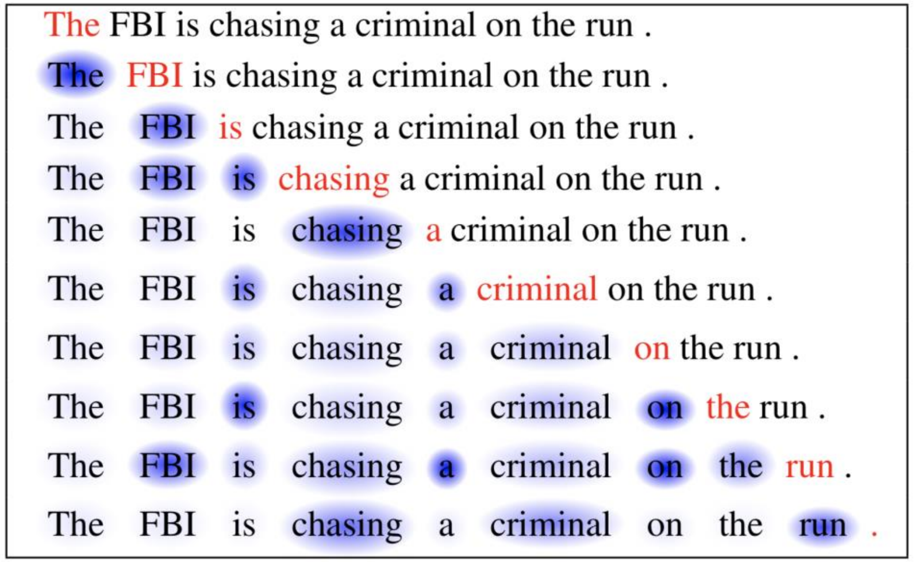
family of genes

Uniform attention: A | transformer | is | a | ?

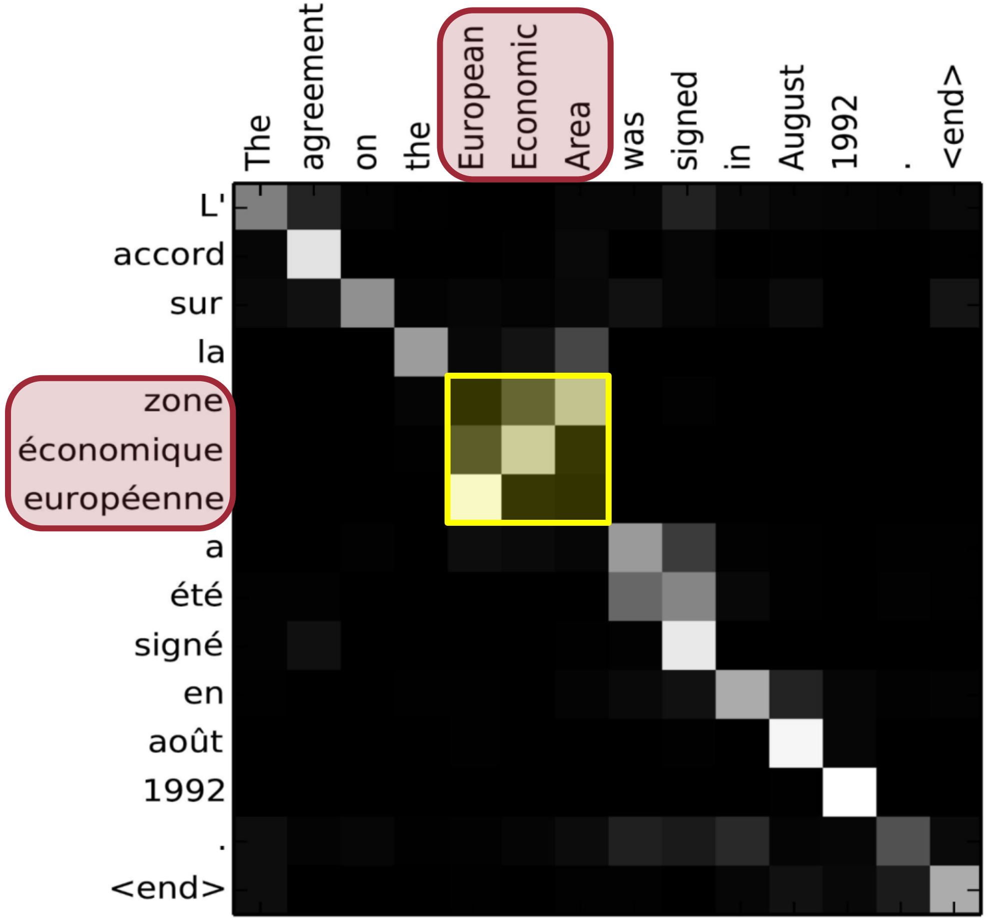
Variable attention: A | transformer | is | a | ?

Transformers learn what to attend to from big data!

(Recap) What does a Transformer Learn?



(Source: Cheng et al., 2016)

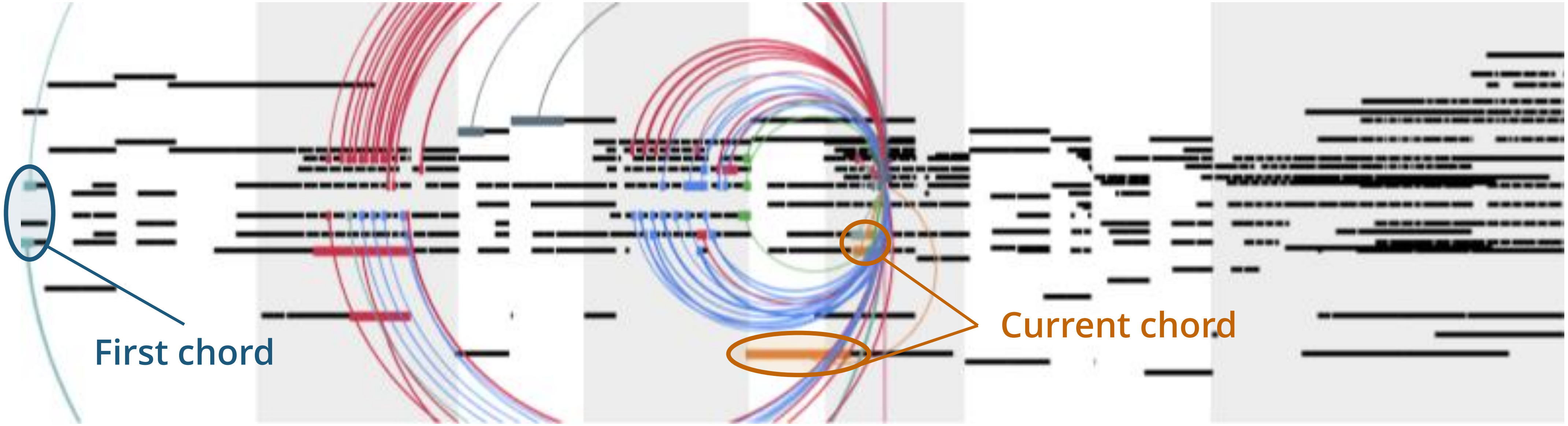


(Source: Bahdanau et al., 2015)

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *ICLR*, 2015.
Jianpeng Cheng, Li Dong, and Mirella Lapata, "Long Short-Term Memory-Networks for Machine Reading," *EMNLP*, 2016.

(Recap) What does a Transformer Learn?

(Each color represents an attention head)

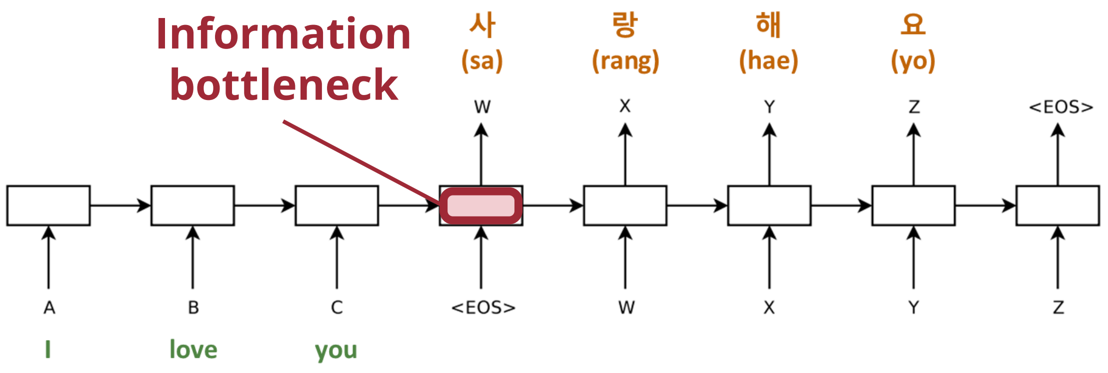


(Source: Huang et al., 2018)

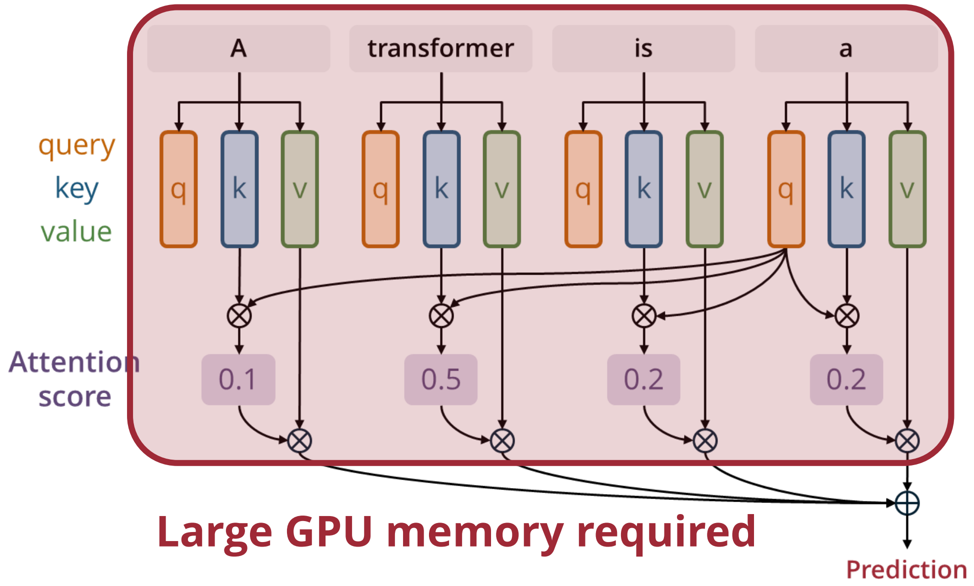
Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *Magenta Blog*, December 13, 2018.

(Recap) Seq2seq vs Transformers

Seq2seq



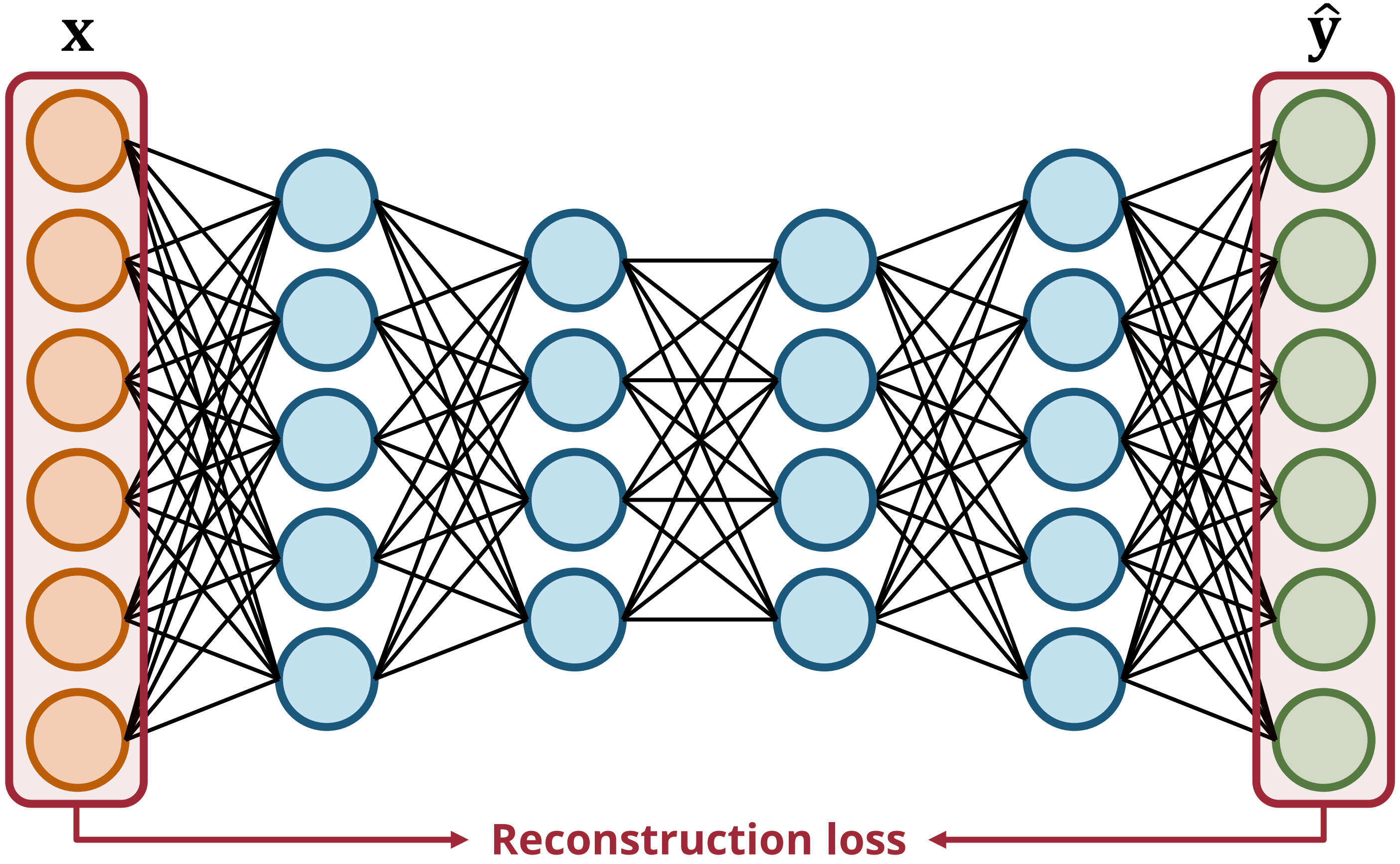
Transformers



Autoencoders

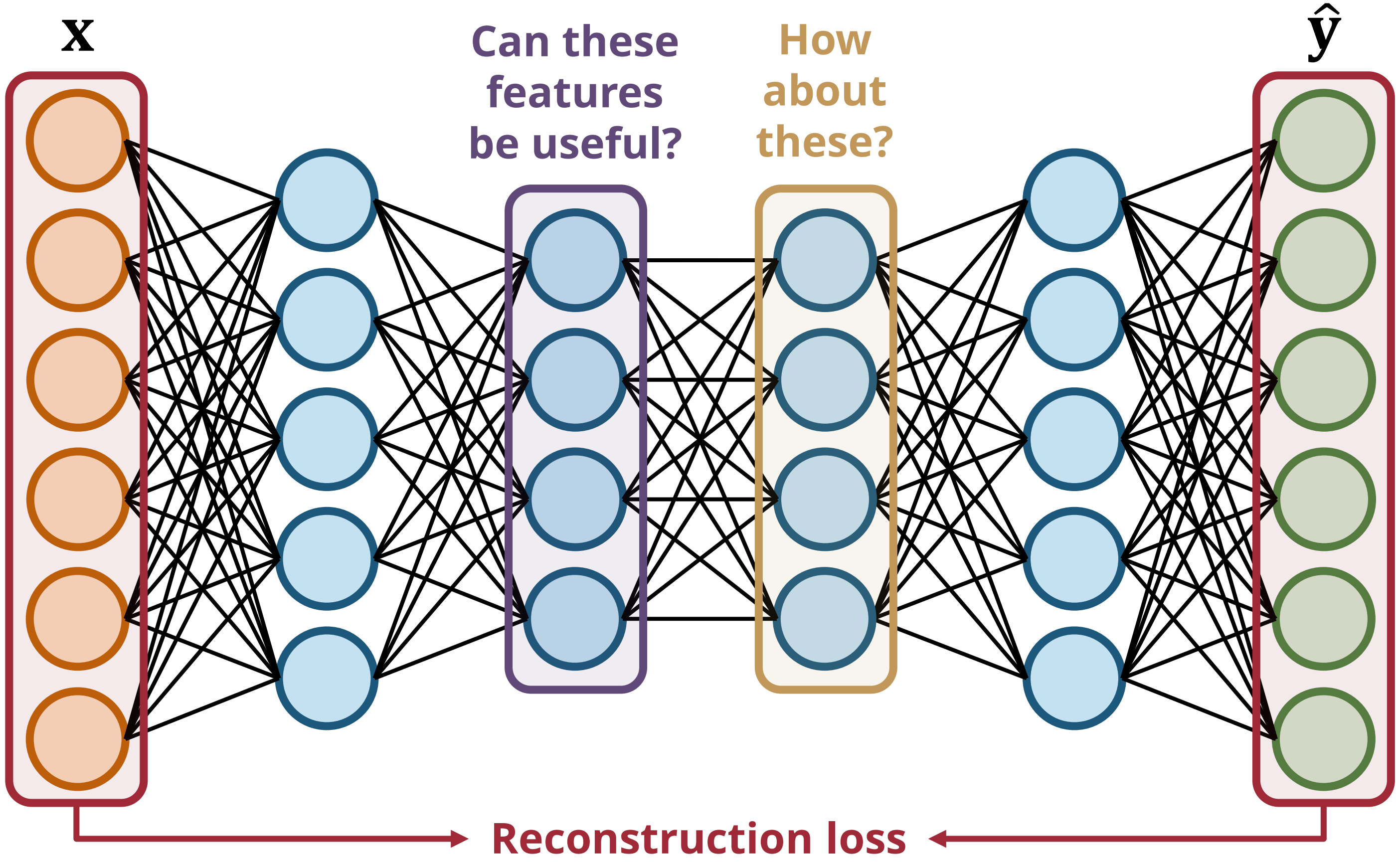
Autoencoders

- A neural network where the **input and output are the same**



Autoencoders

- A neural network where the **input and output are the same**



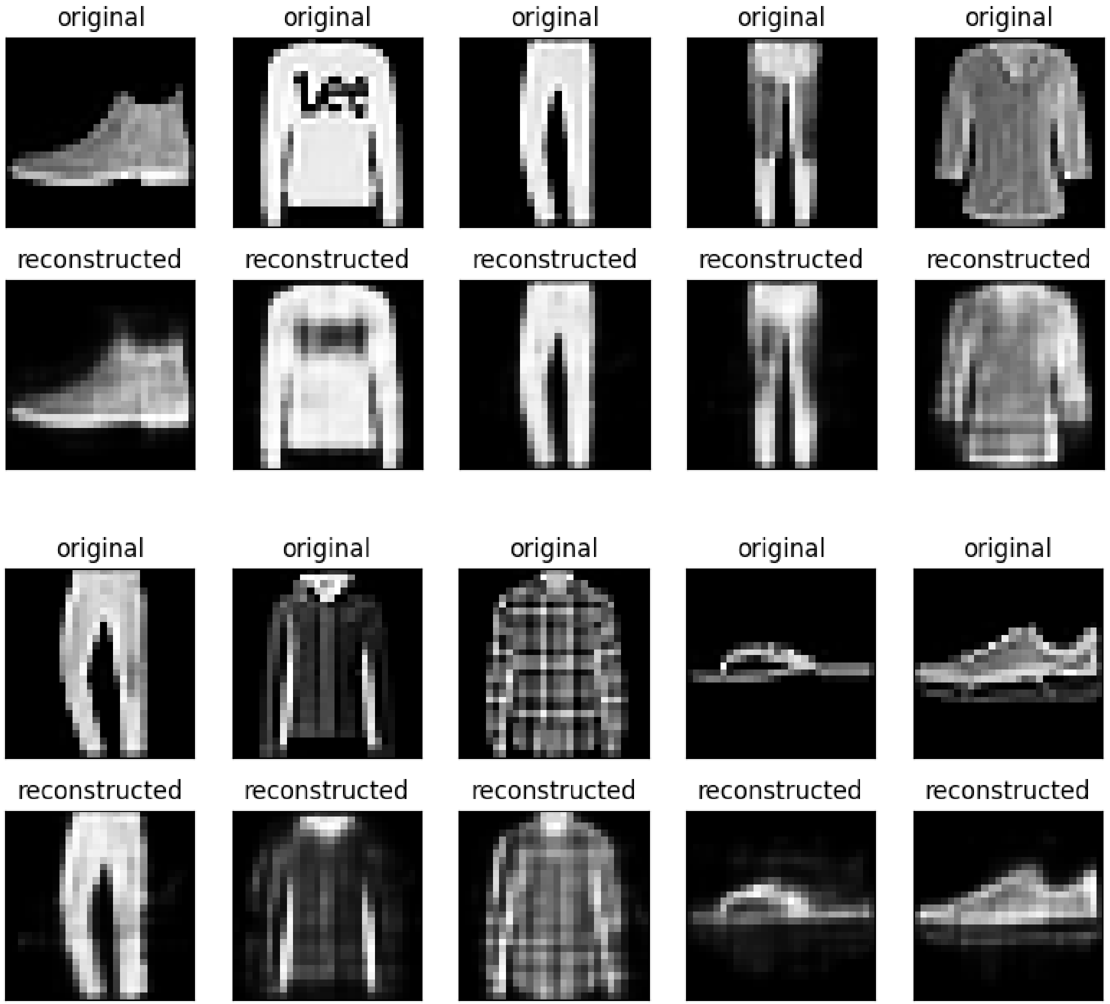
Autoencoders – Reconstruction Examples

Original

Reconstructed

Original

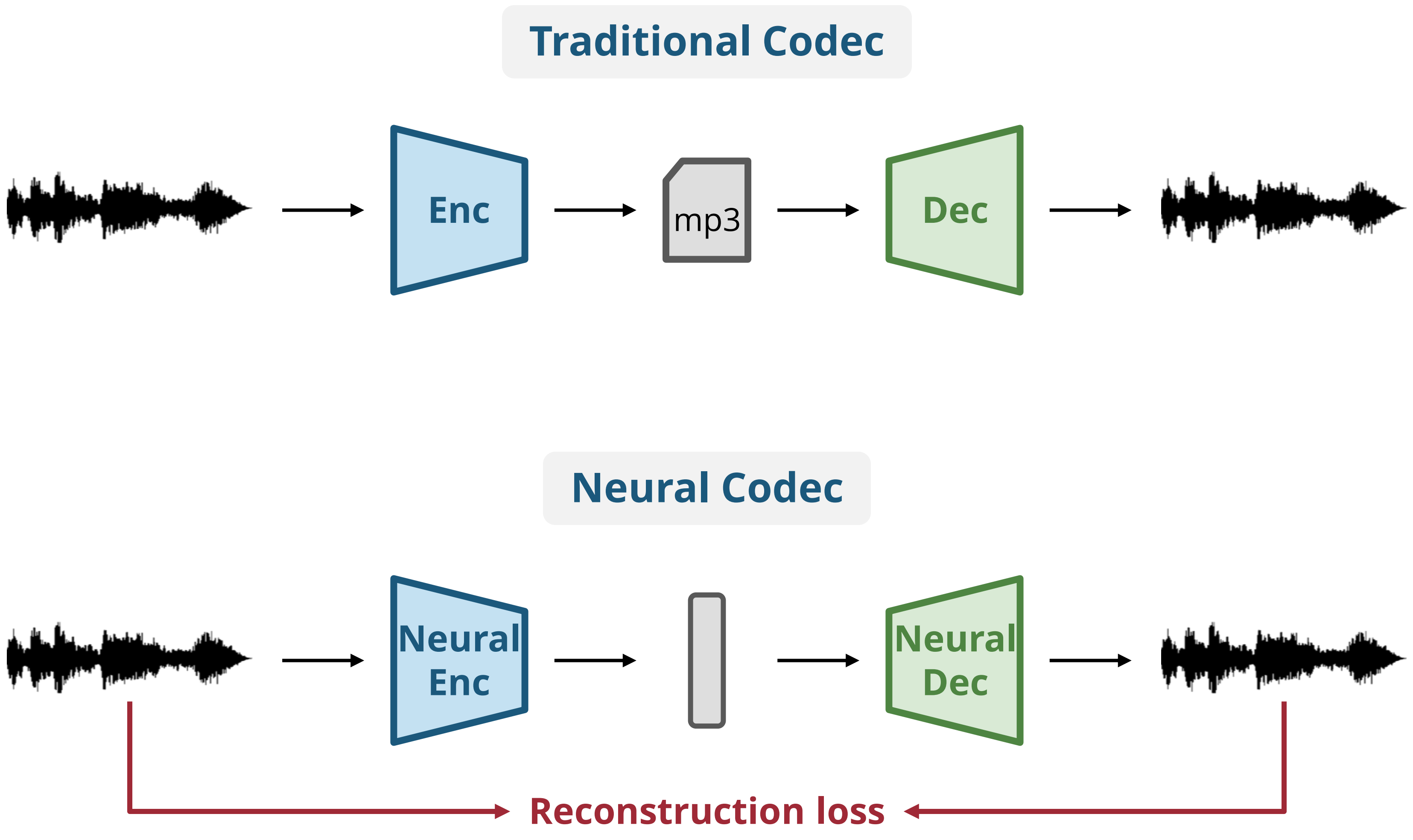
Reconstructed



(Source: tensorflow.org)

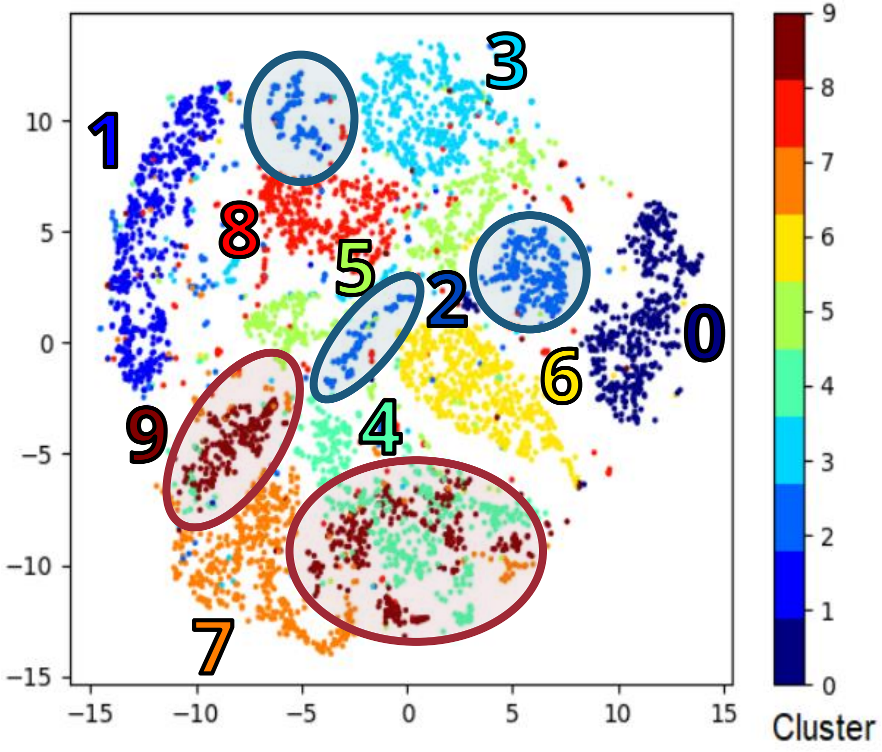
tensorflow.org/tutorials/generative/autoencoder

Codec is an Autoencoder

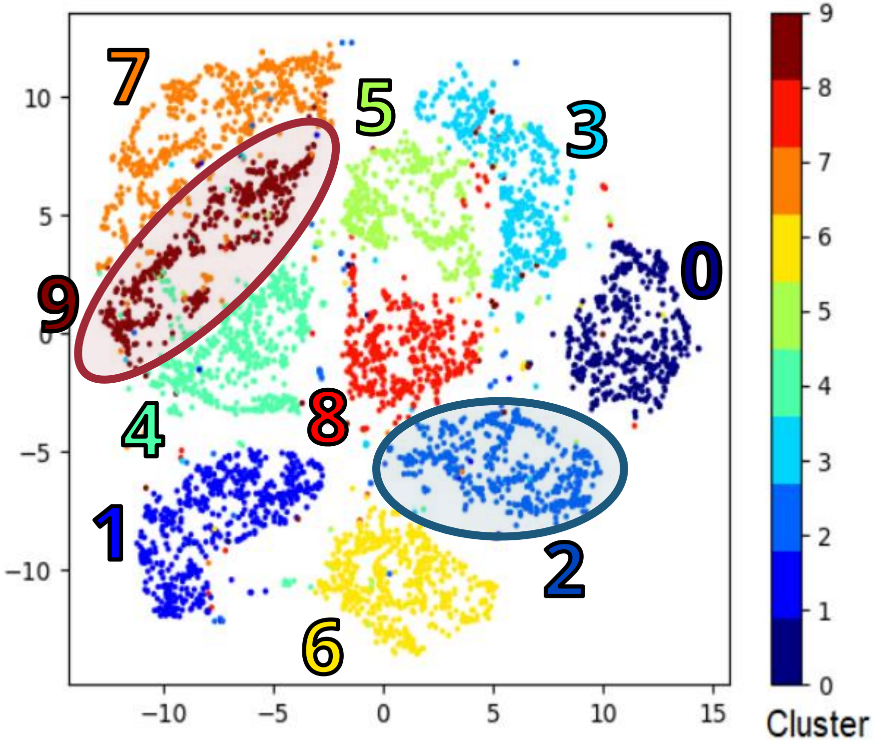


Unsupervised Clustering with an Autoencoder

Raw pixels



AE-encoded

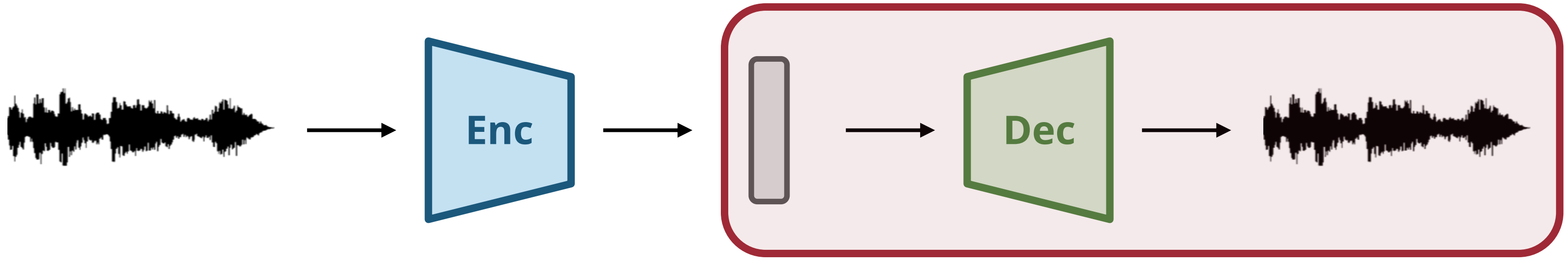


(Source: Aljalbout et al., 2020)

Elie Aljalbout, Vladimir Golkov, Yawar Siddiqui, Maximilian Strobel, and Daniel Cremers, "Clustering with Deep Learning: Taxonomy and New Methods," arXiv preprint arXiv:1801.07648, 2018.

Variational Autoencoders (VAEs)

Autoencoders

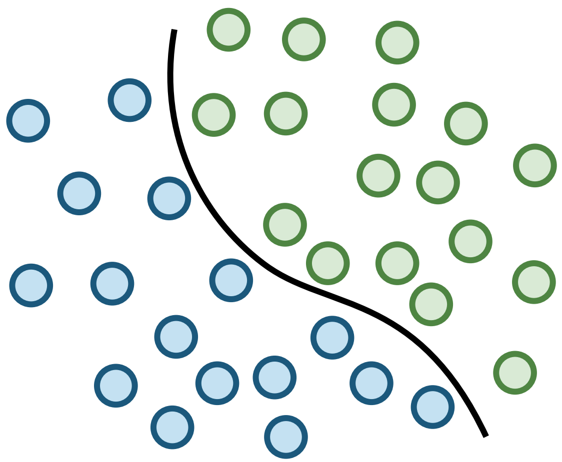


Isn't this like a generative model?

What exactly is a generative model?

Discriminative vs Generative Models

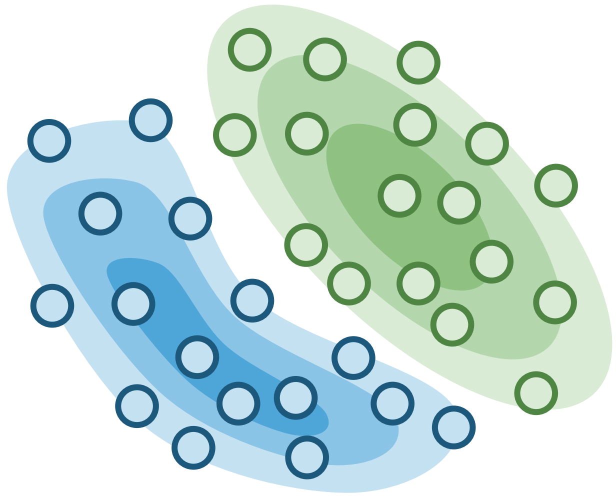
Discriminative



Discriminative models learn the decision boundary

$P(y|x)$

Generative

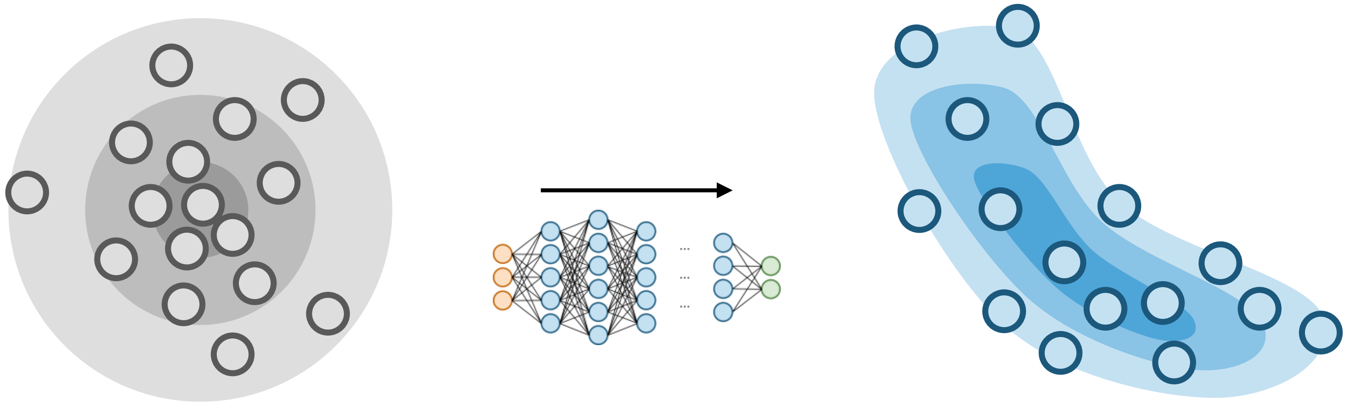


Generative models learn the underlying distribution

$P(x)$ or $P(x|y)$

Learning to Generate Data from Random Distributions

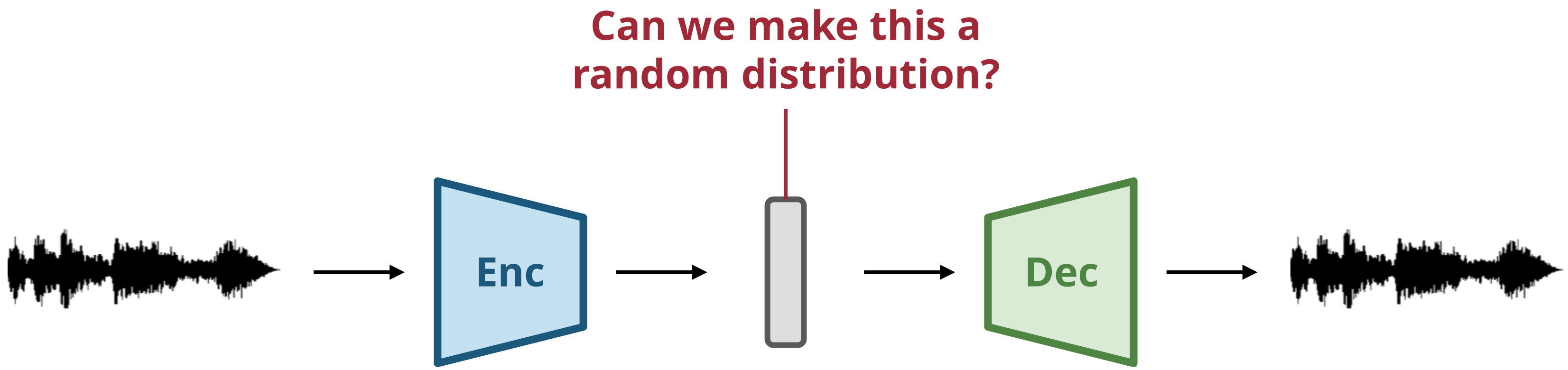
Random distribution **Data distribution**



$P(z)$ $P(x)$

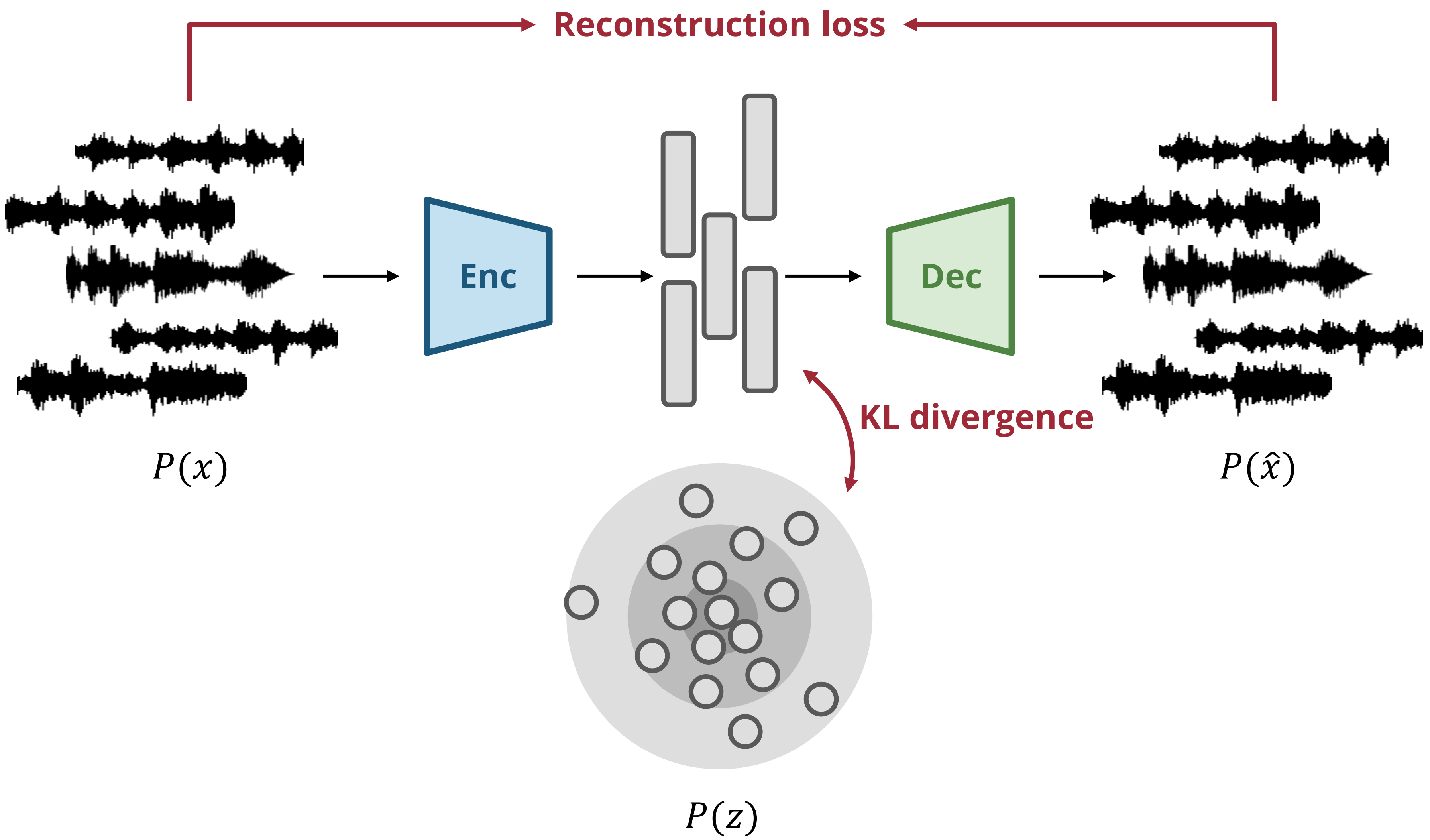
If we can learn this mapping, we can easily generate new samples from the data distribution

Variational Autoencoders (VAEs)



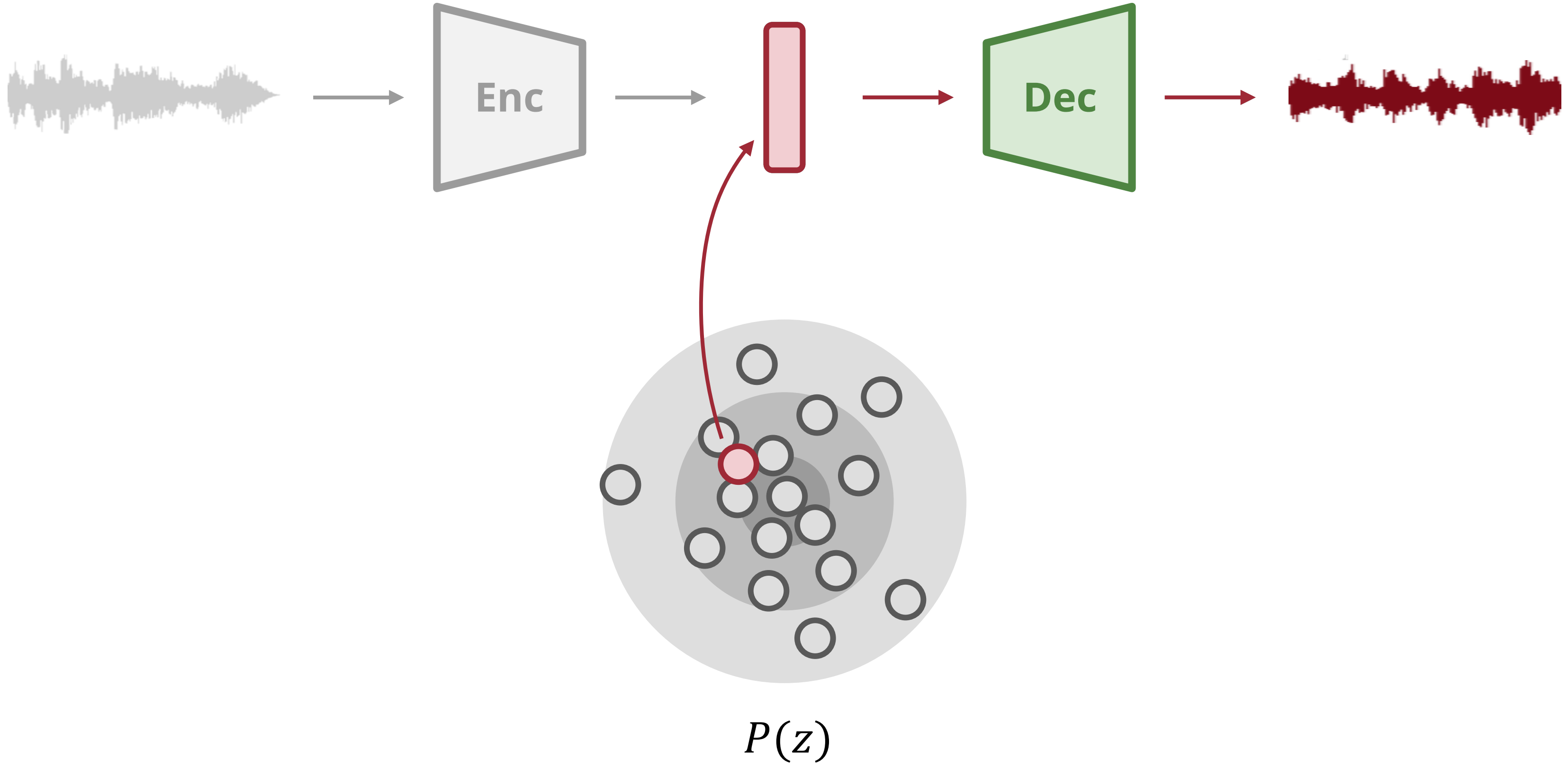
Diederik P Kingma and Max Welling, "Auto-Encoding Variational Bayes," *ICLR*, 2014.

Variational Autoencoders (VAEs) – Training



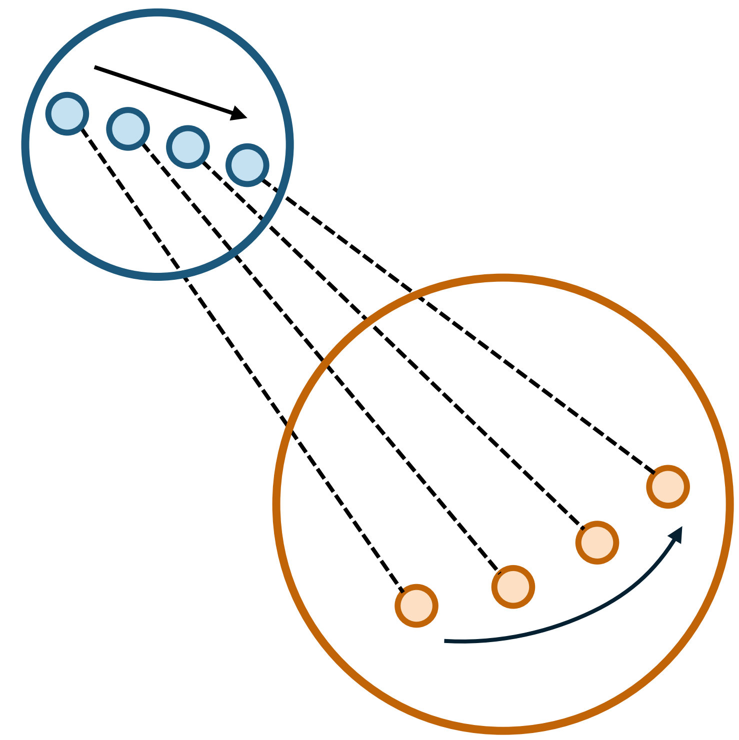
Diederik P Kingma and Max Welling, "Auto-Encoding Variational Bayes," *ICLR*, 2014.

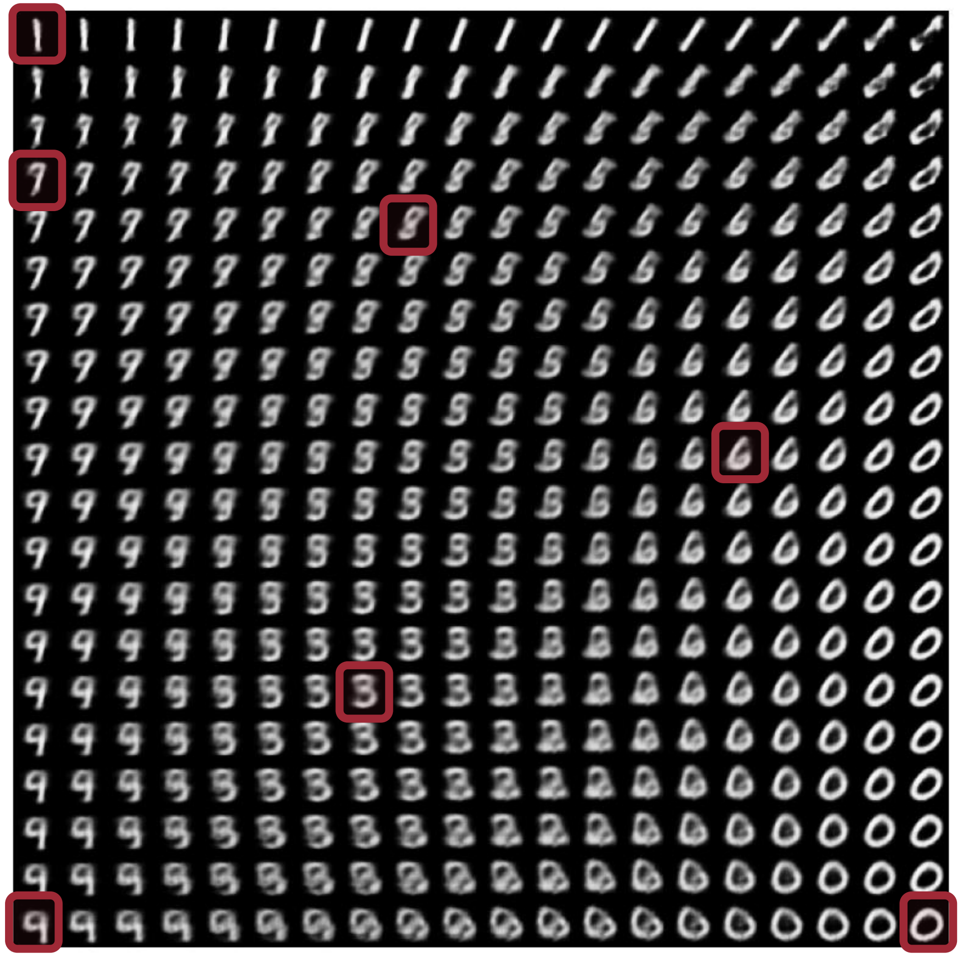
Variational Autoencoders (VAEs) – Generation



Diederik P Kingma and Max Welling, “Auto-Encoding Variational Bayes,” *ICLR*, 2014.

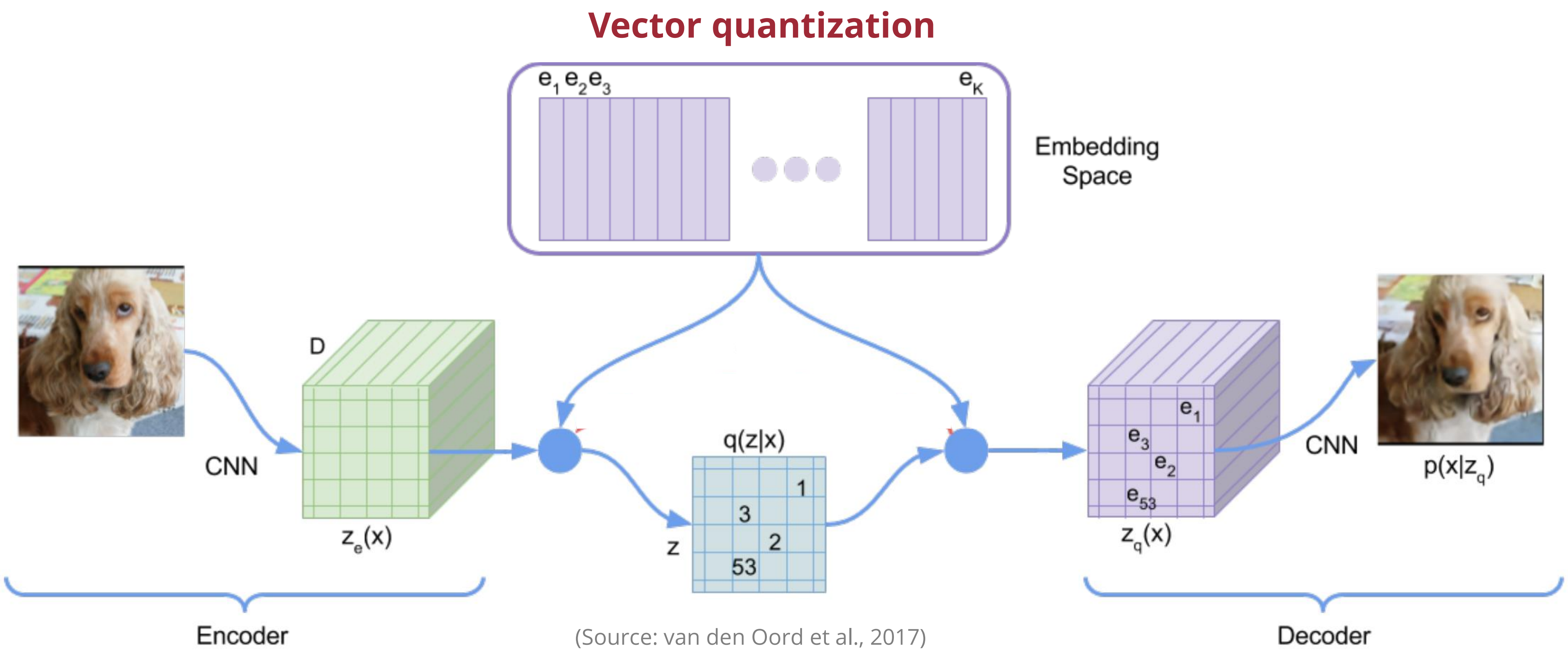
Decoding the Latent Space of a VAE





(Source: tensorflow.org)

Vector-Quantized VAEs (VQVAEs)



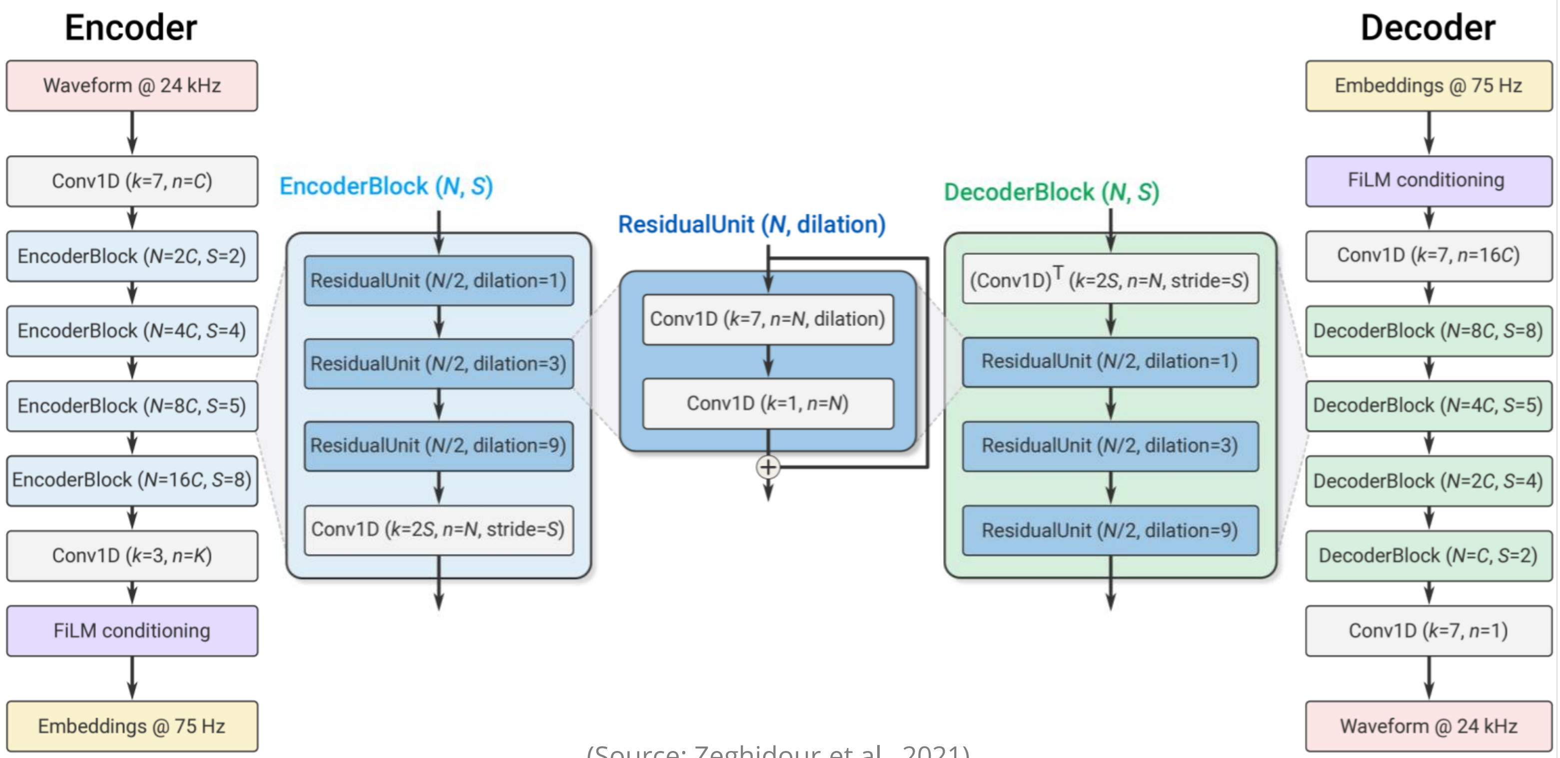
(Source: van den Oord et al., 2017)

Allow only a fixed number of vectors to be used in the bottleneck layer

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu, "Neural Discrete Representation Learning," *NeurIPS*, 2017.

SoundStream

- **Fully-convolutional autoencoder** for audio
- Follow-up work: Encodec & Descript Audio Codec



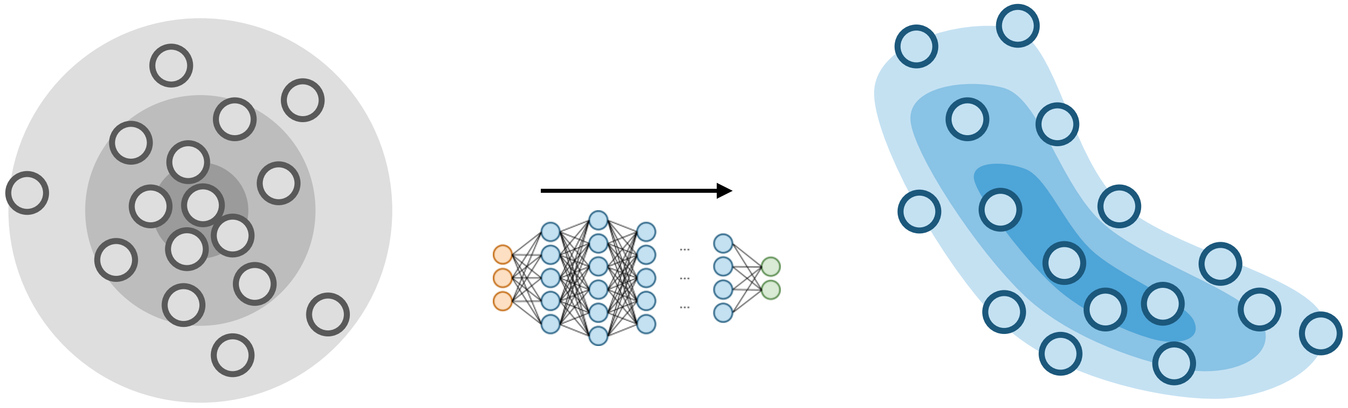
(Source: Zeghidour et al., 2021)

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "SoundStream: An End-to-End Neural Audio Codec," *TASLP*, 2021.

Generative Adversarial Nets (GANs)

Generating Data from a Random Distribution

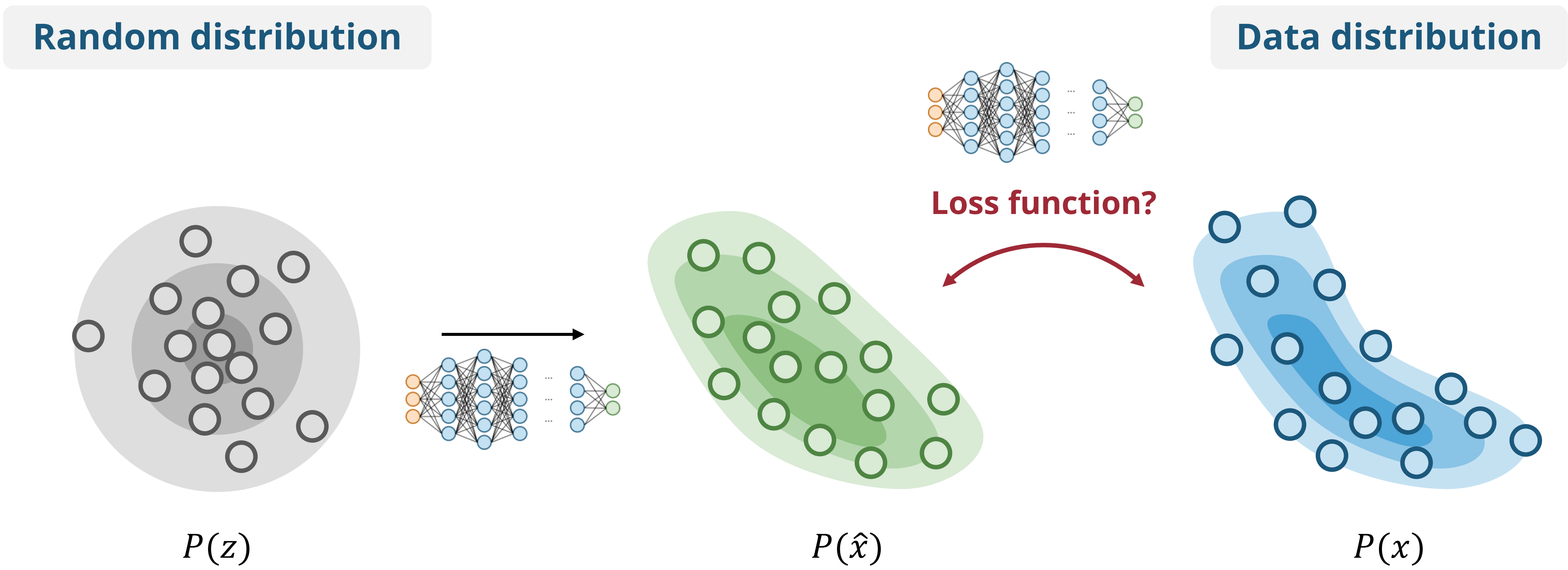
Random distribution | **Data distribution**



$P(z)$ $P(x)$

If we can learn this mapping, we can easily generate new samples from the data distribution

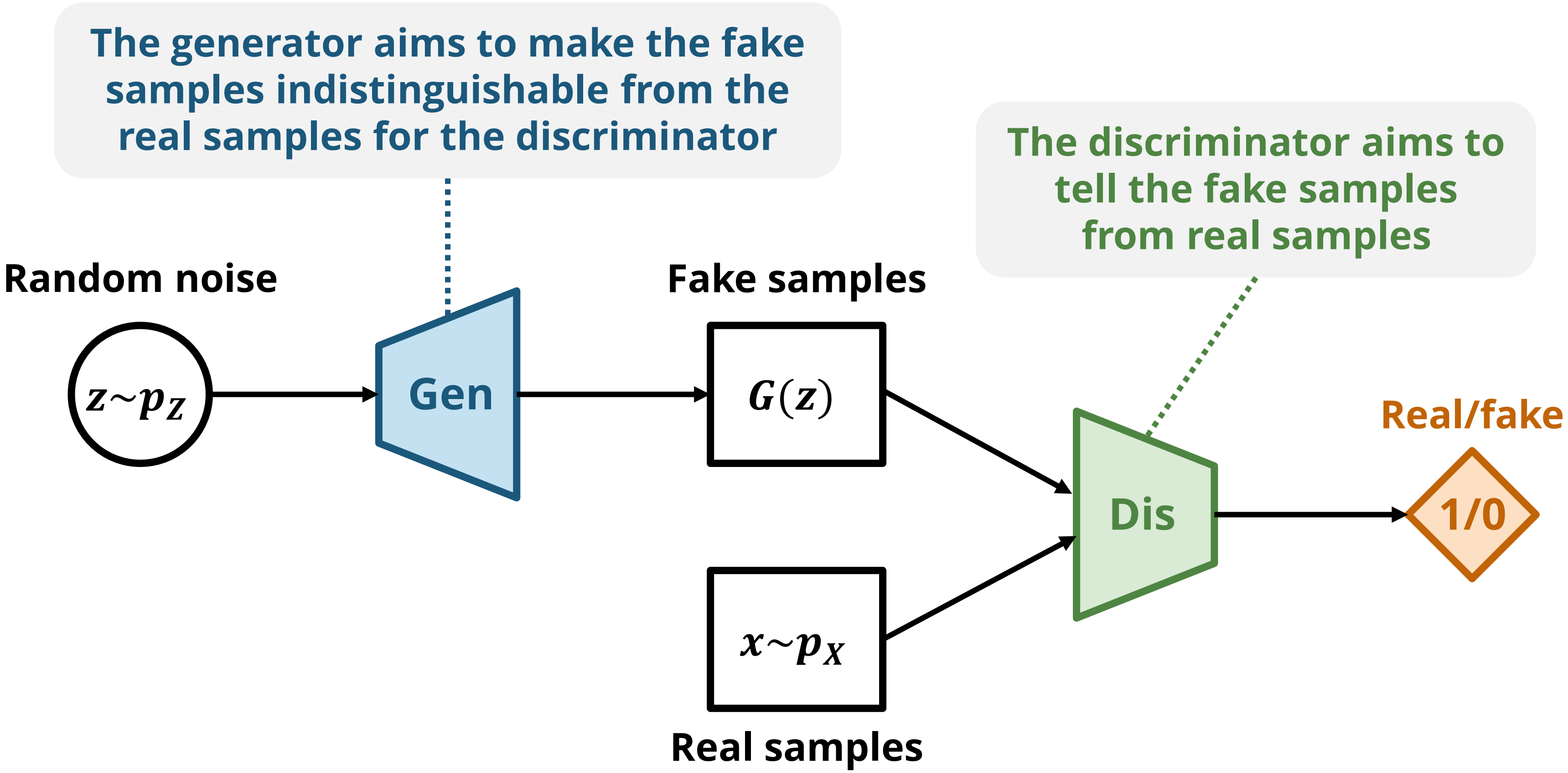
A Loss Function for Distributions



Unfortunately, no easy way to measure the difference between two distributions

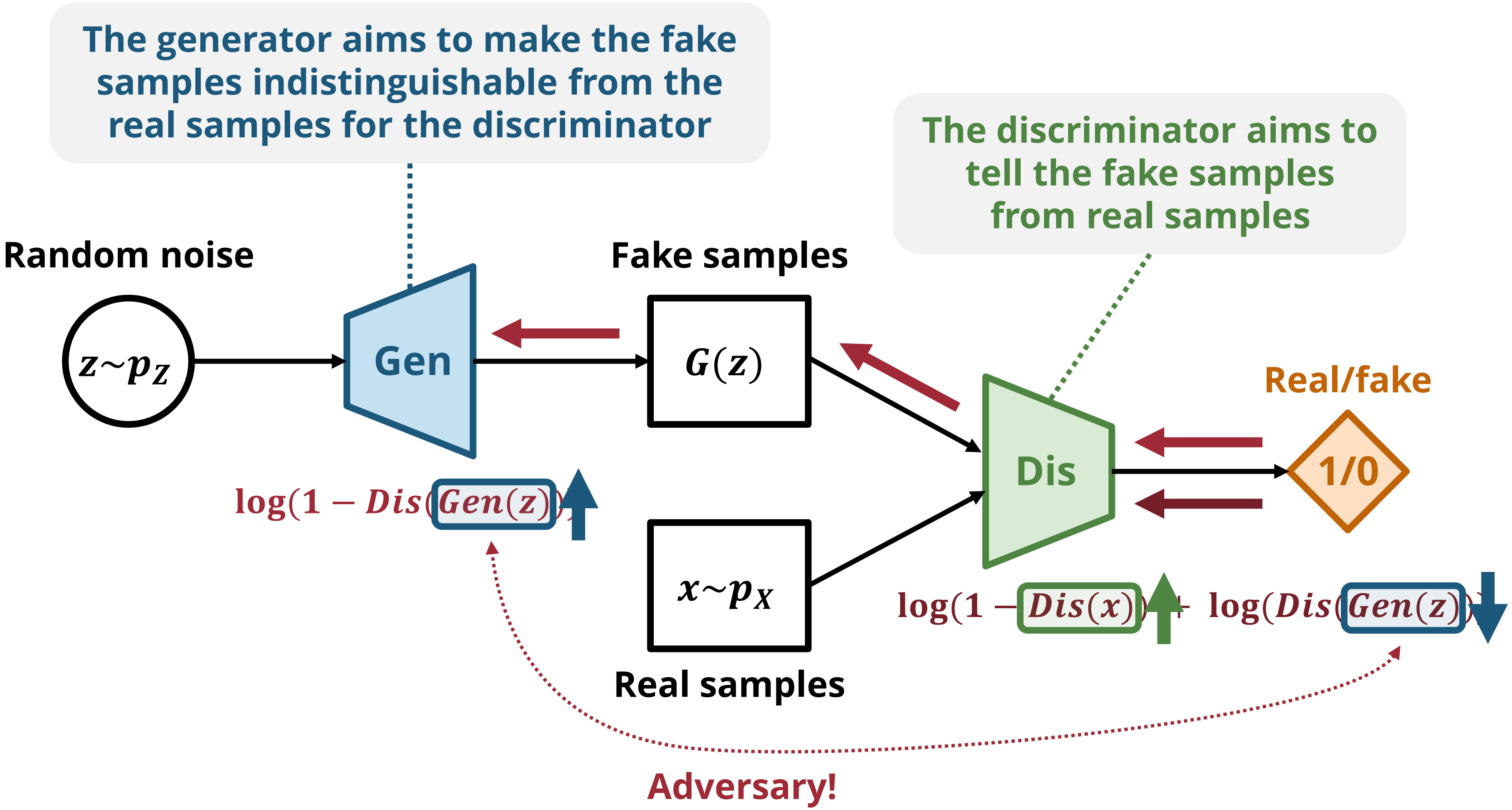
But what about another neural network!?

Generative Adversarial Nets (GANs)



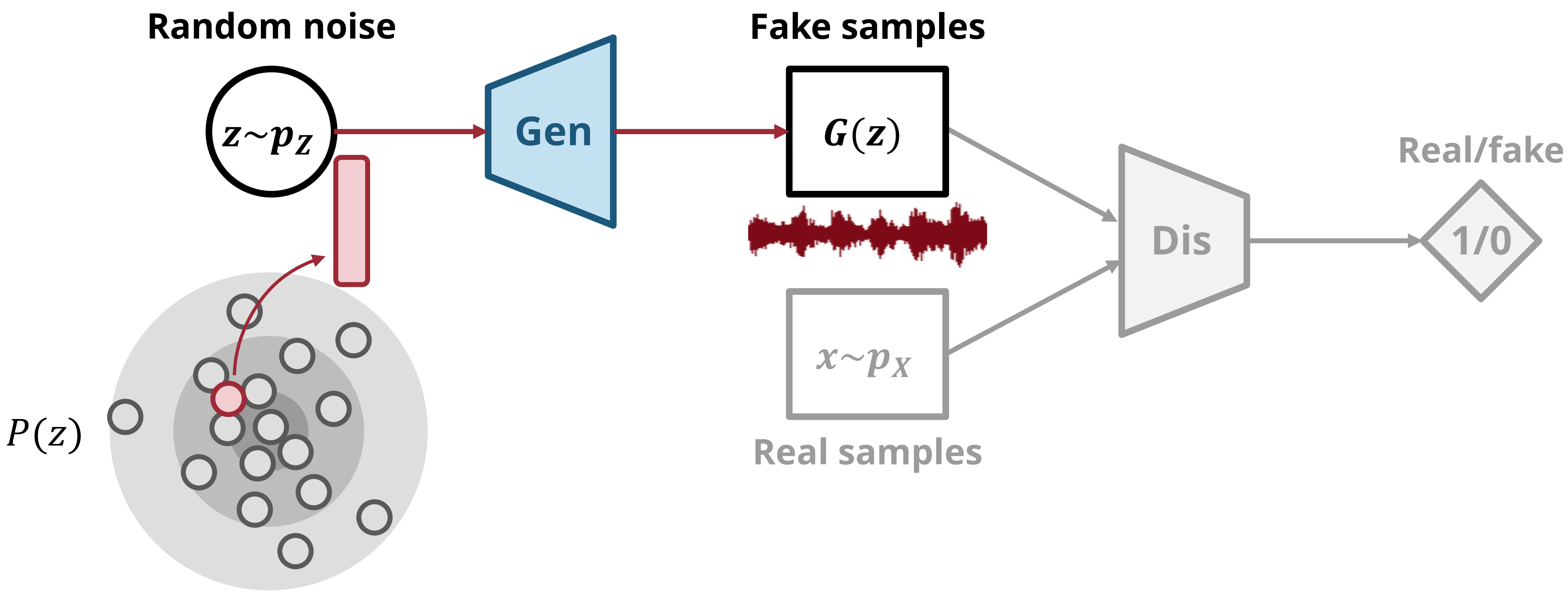
Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Networks," *NeurIPS*, 2014.

Generative Adversarial Nets (GANs) – Training



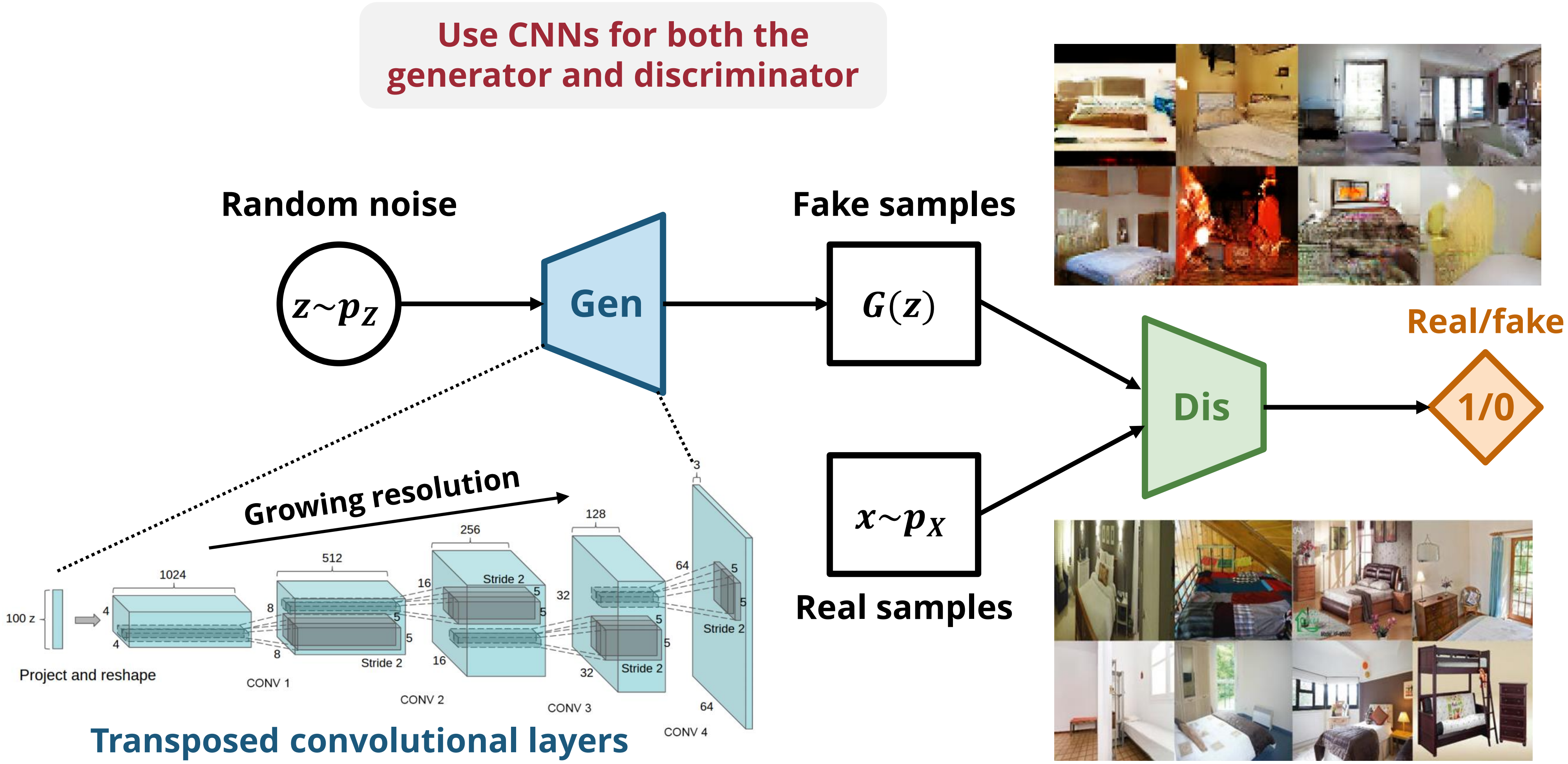
Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Networks," *NeurIPS*, 2014.

Generative Adversarial Nets (GANs) – Generation



Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Networks," *NeurIPS*, 2014.

Deep Convolutional GANs (DCGANs)



Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *ICLR*, 2016.

Transposed Convolution

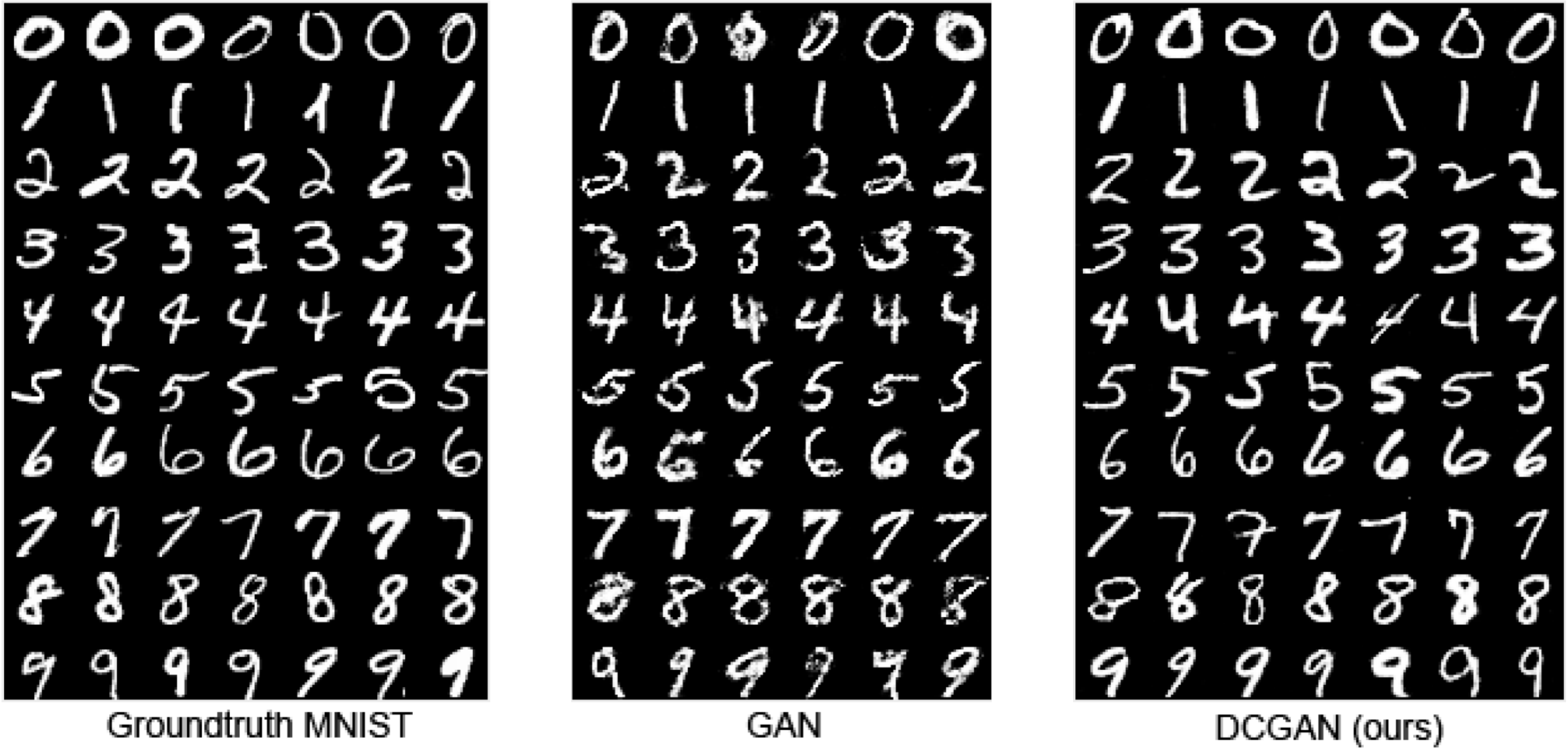
Convolution

$$\begin{bmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 \end{bmatrix} * \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 9 & -1 \\ -1 & 9 \end{bmatrix}$$

Transposed convolution

$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} * \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 4 & -2 & 0 \\ 0 & -2 & 4 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

DCGAN – Examples

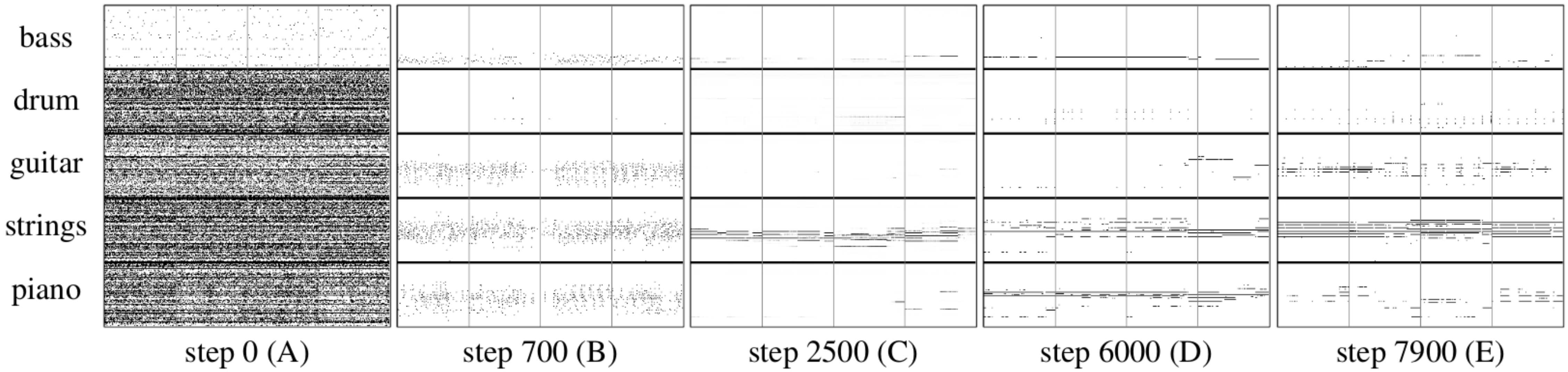


(Source: Radford et al., 2016)

Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *ICLR*, 2016.

MuseGAN – A GAN for Pianorolls

The generator improves over time **So does the discriminator!**



(Source: Dong et al., 2018)

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang, "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment," *AAAI*, 2018.

GigaGAN: Scaling up GANs

A portrait of a human growing colorful flowers from her hair. Hyperrealistic oil painting. Intricate details.

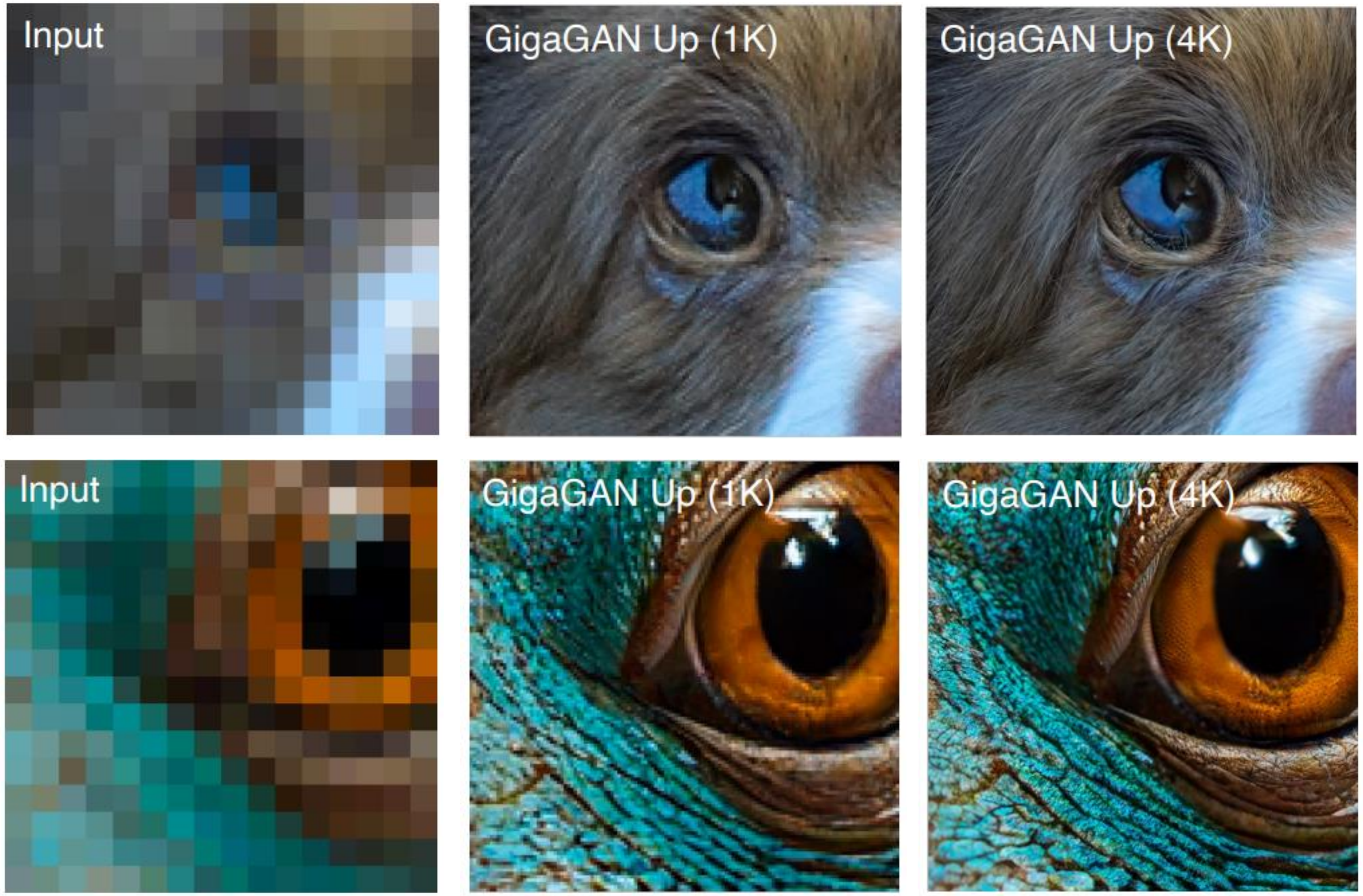
A golden luxury motorcycle parked at the King's palace. 35mm f/4.5.

a cute magical flying maltipoo at light speed, fantasy concept art, bokeh, wide sky

(Source: Kang et al., 2023)

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park, "Scaling up GANs for Text-to-Image Synthesis," *CVPR*, 2023.

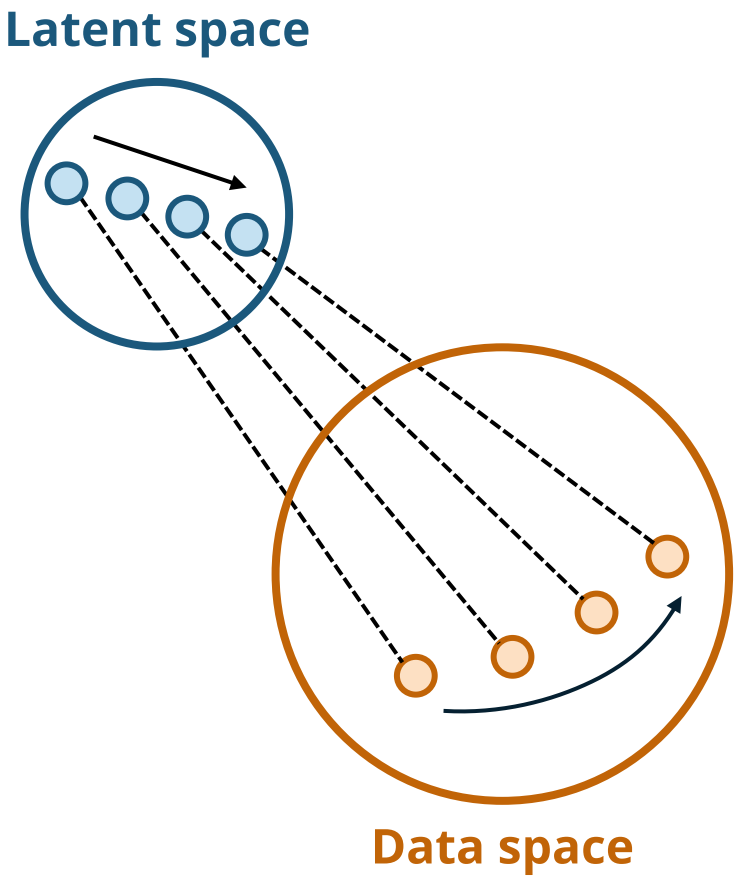
GigaGAN for Image Super-resolution

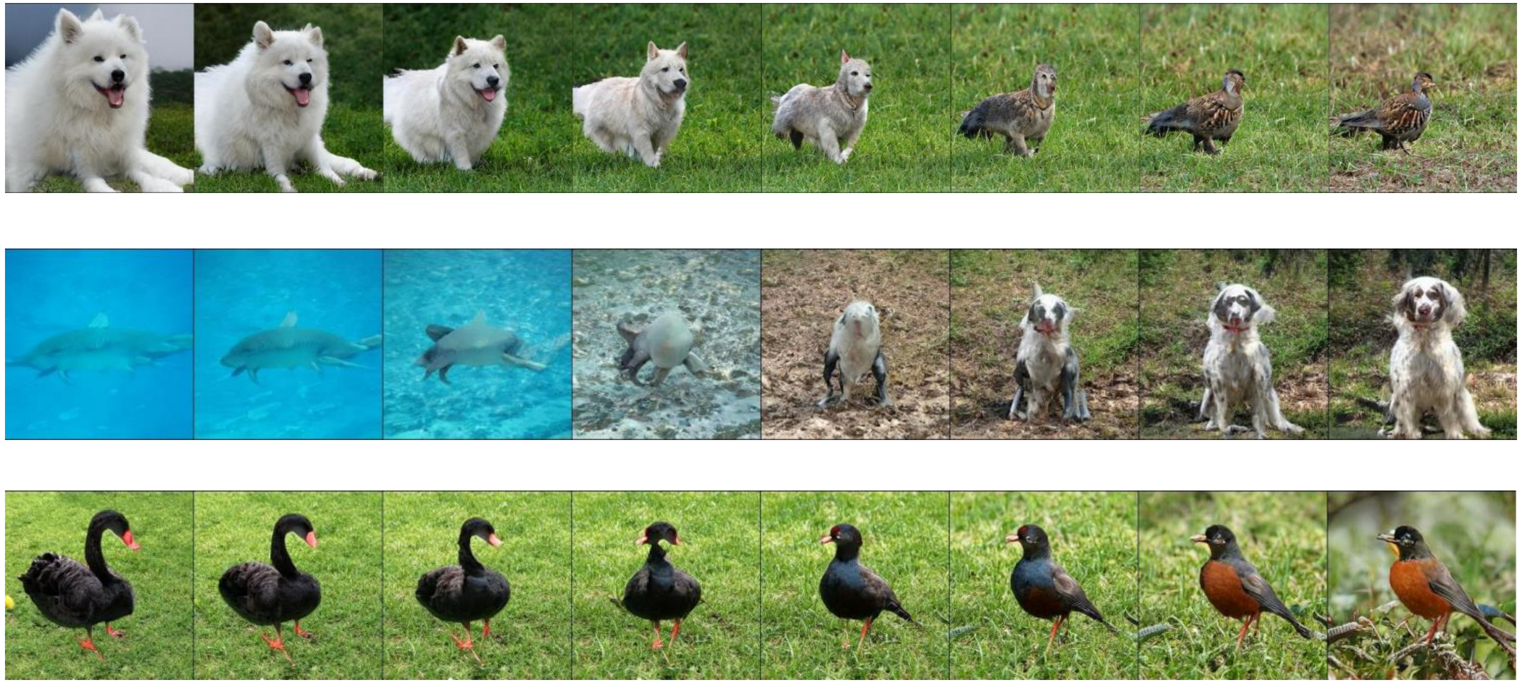


(Source: Kang et al., 2023)

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park, "Scaling up GANs for Text-to-Image Synthesis," *CVPR*, 2023.

Interpolation on the Latent Space

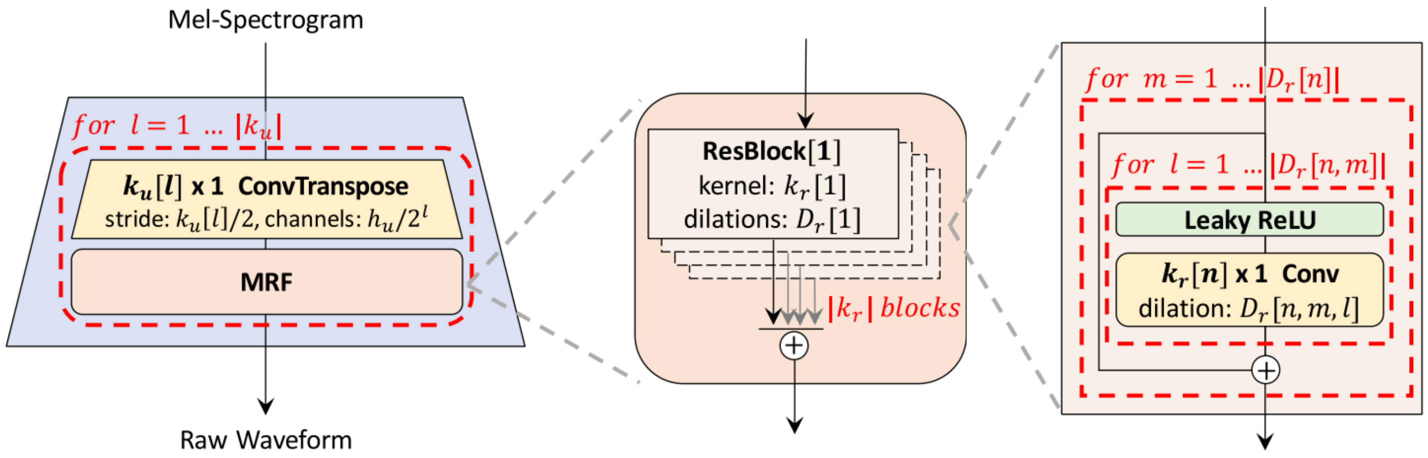


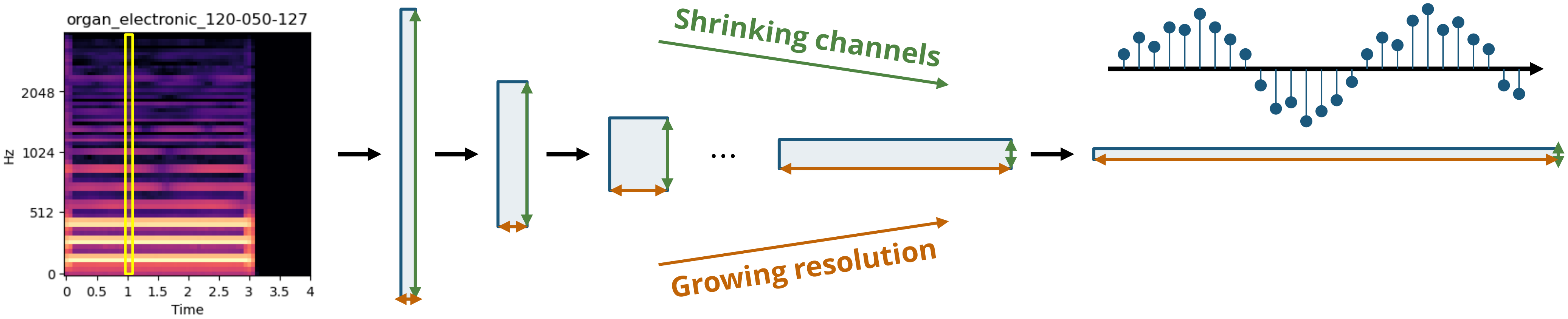


(Source: Brock et al., 2019)

Andrew Brock, Jeff Donahue, and Karen Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," *ICLR*, 2019.

Hifi-GAN – Generator

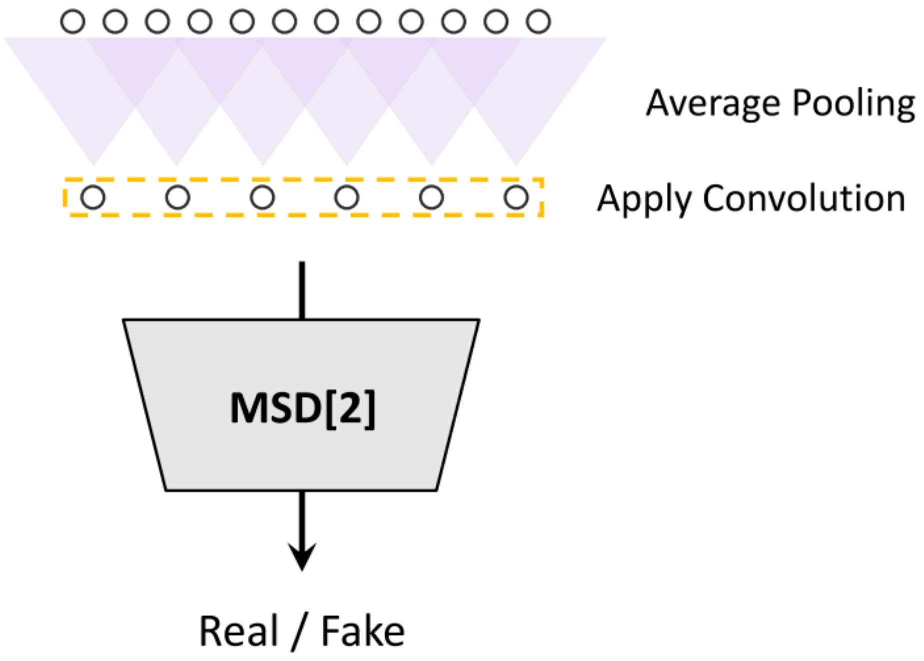




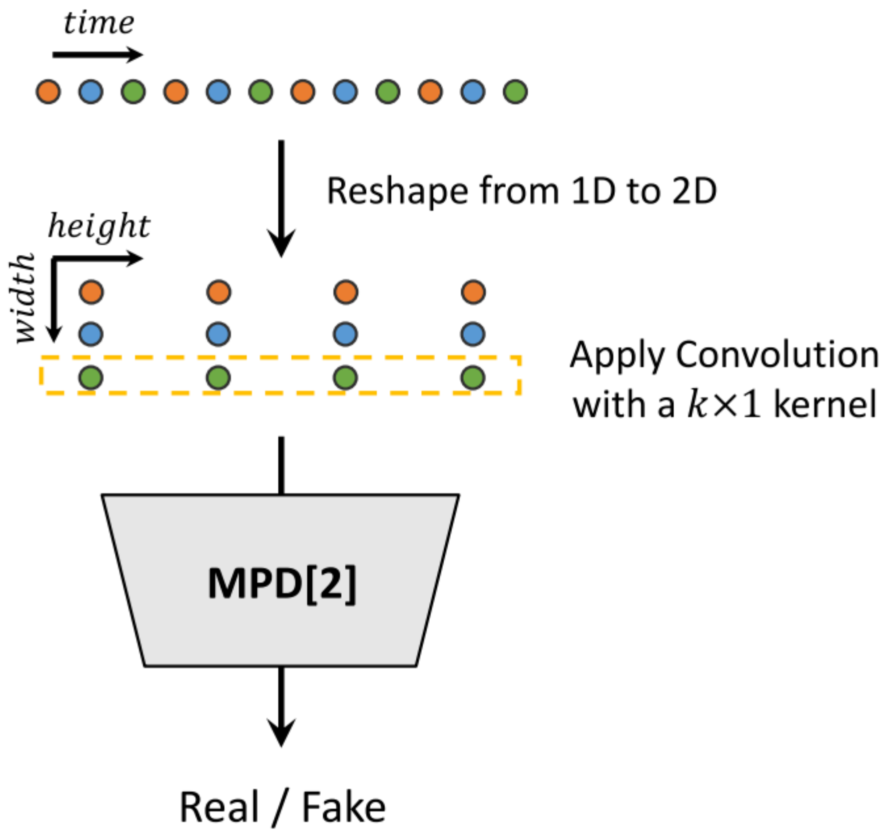
Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *NeurIPS*, 2020.

Hifi-GAN – Discriminator

Multi-scale discriminator



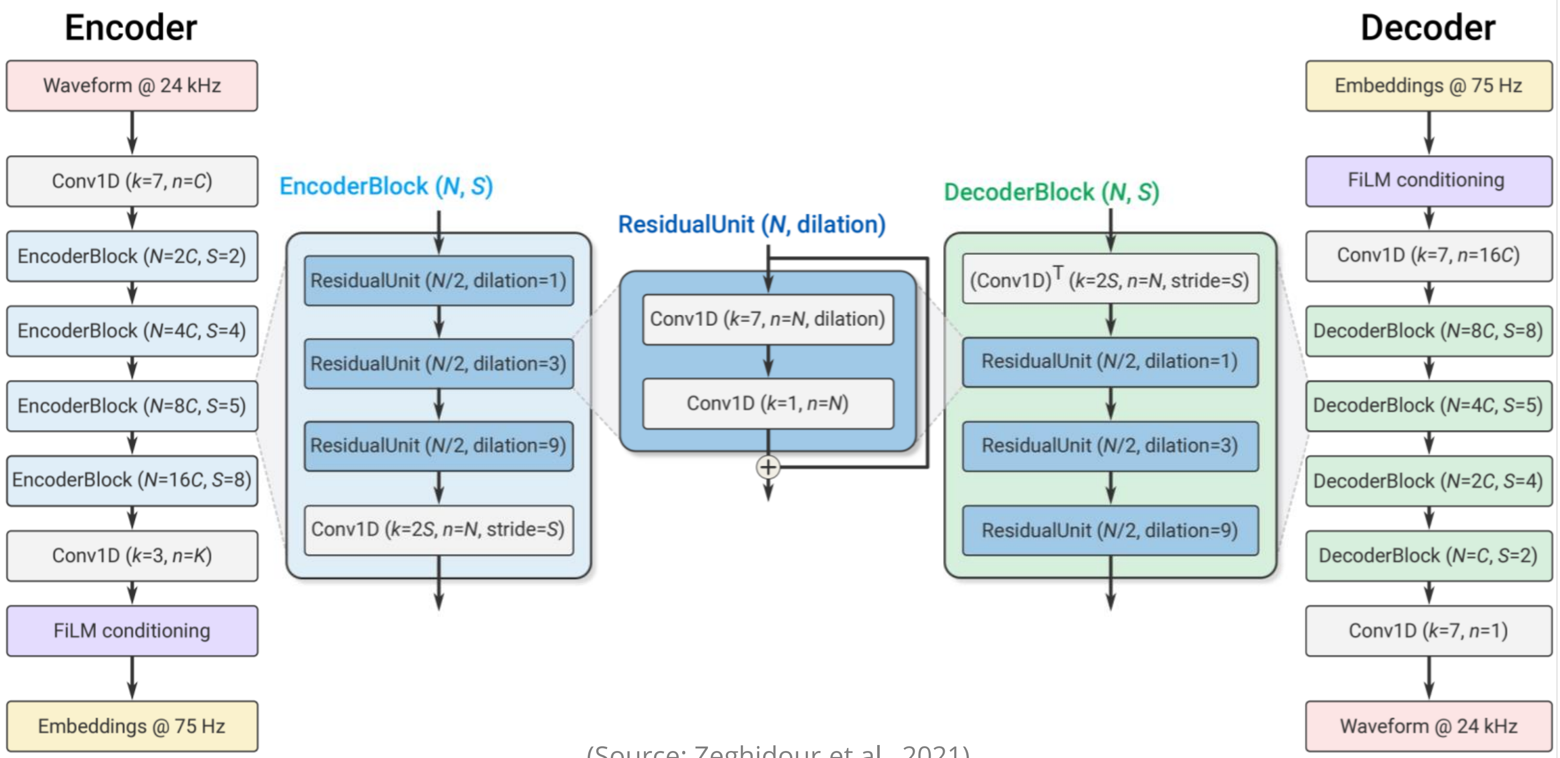
Multi-period discriminator



Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *NeurIPS*, 2020.

(Recap) SoundStream

- **Fully-convolutional autoencoder** for audio
- Follow-up work: Encodec & Descript Audio Codec

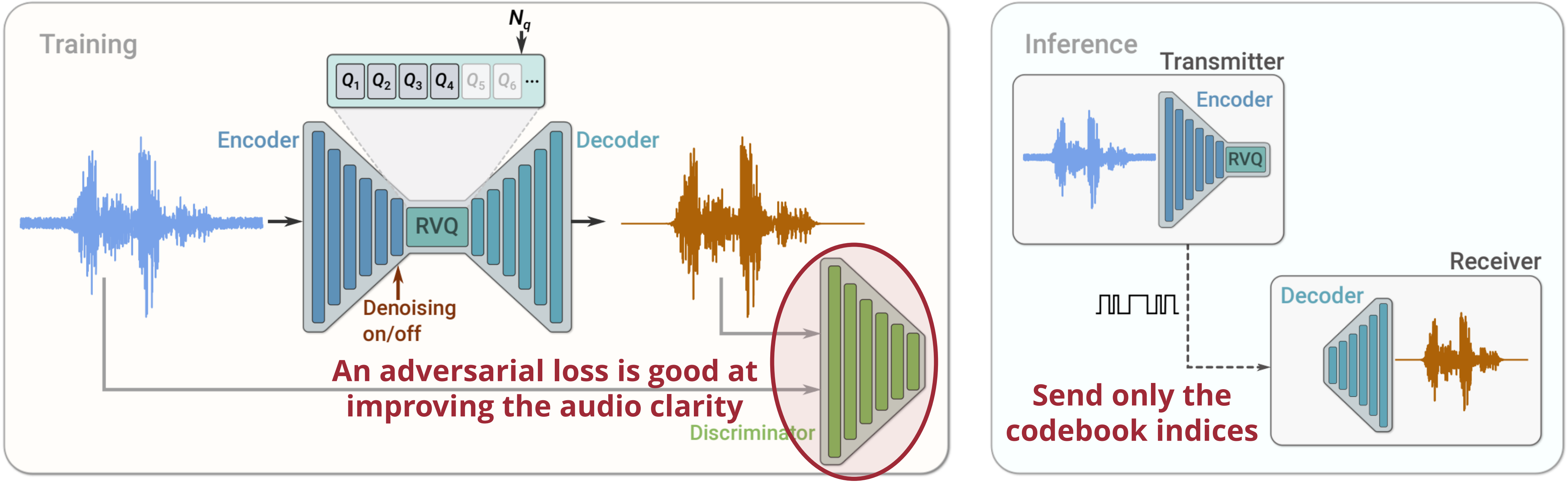


(Source: Zeghidour et al., 2021)

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "SoundStream: An End-to-End Neural Audio Codec," *TASLP*, 2021.

(Recap) SoundStream

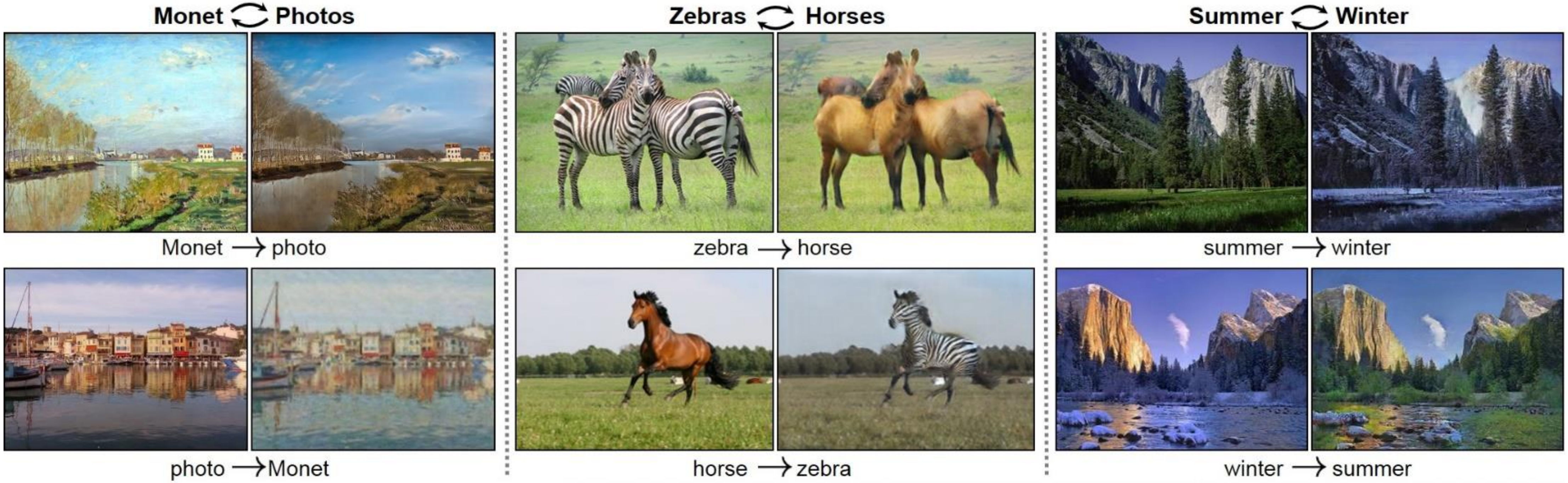
- **Fully-convolutional autoencoder** for audio
- Follow-up work: Encodec & Descript Audio Codec



(Source: Zeghidour et al., 2021)

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "SoundStream: An End-to-End Neural Audio Codec," *TASLP*, 2021.

CycleGAN – Examples



(Source: Zhu et al., 2017)

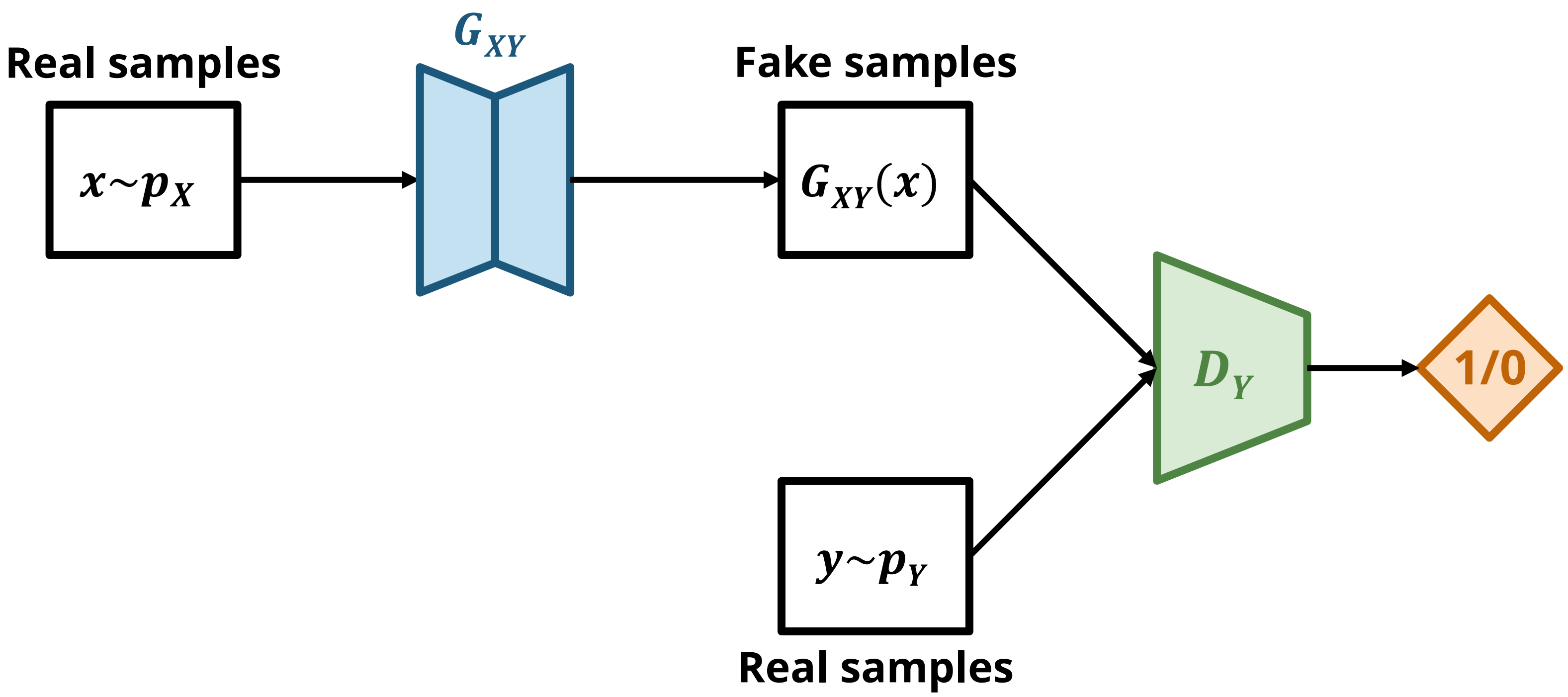
Work without paired data!
For example, we only need **a collection of photos** and **a collection of Monet's paintings**

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *ICCV*, 2017.

CycleGAN – Goal

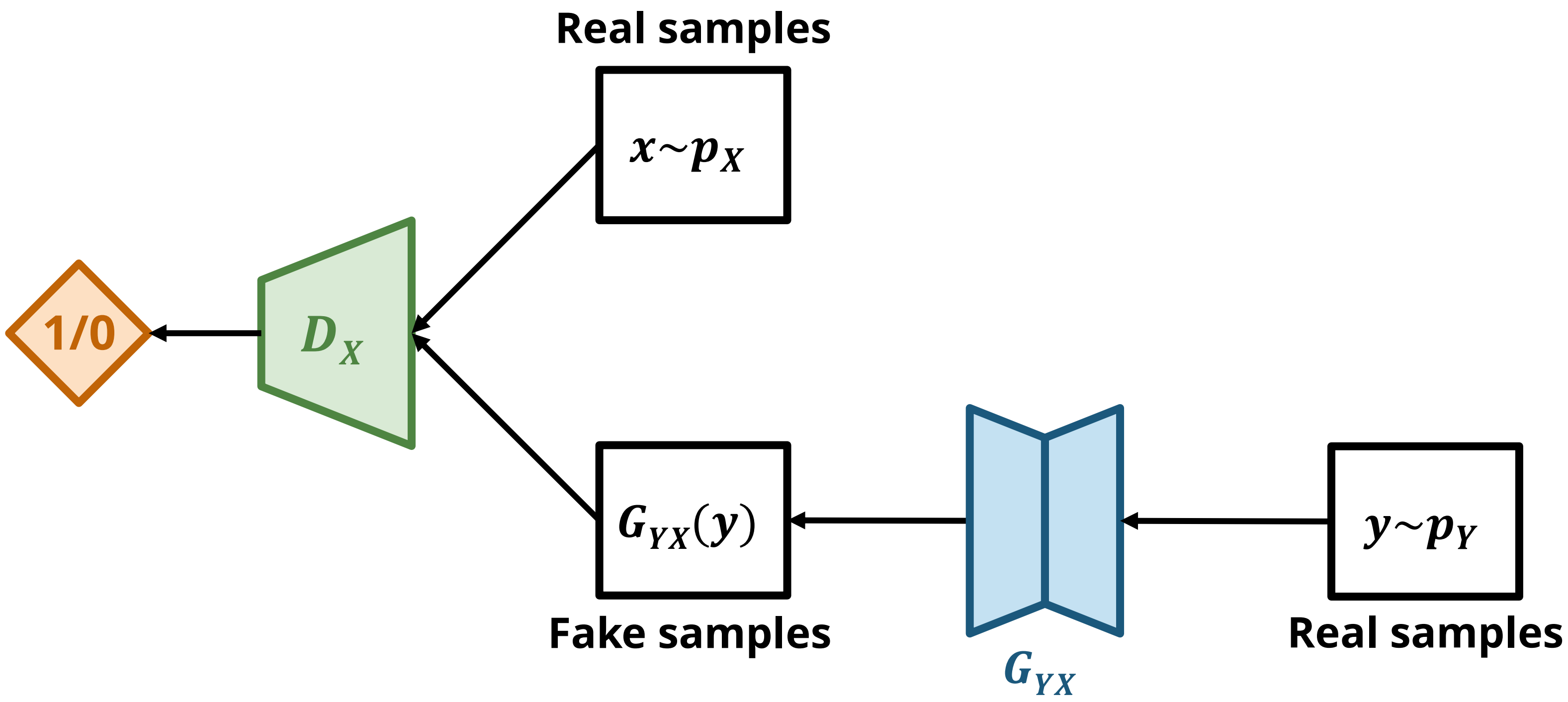
Can we learn the mapping without paired data?

Cycle-consistent GAN (CycleGAN)



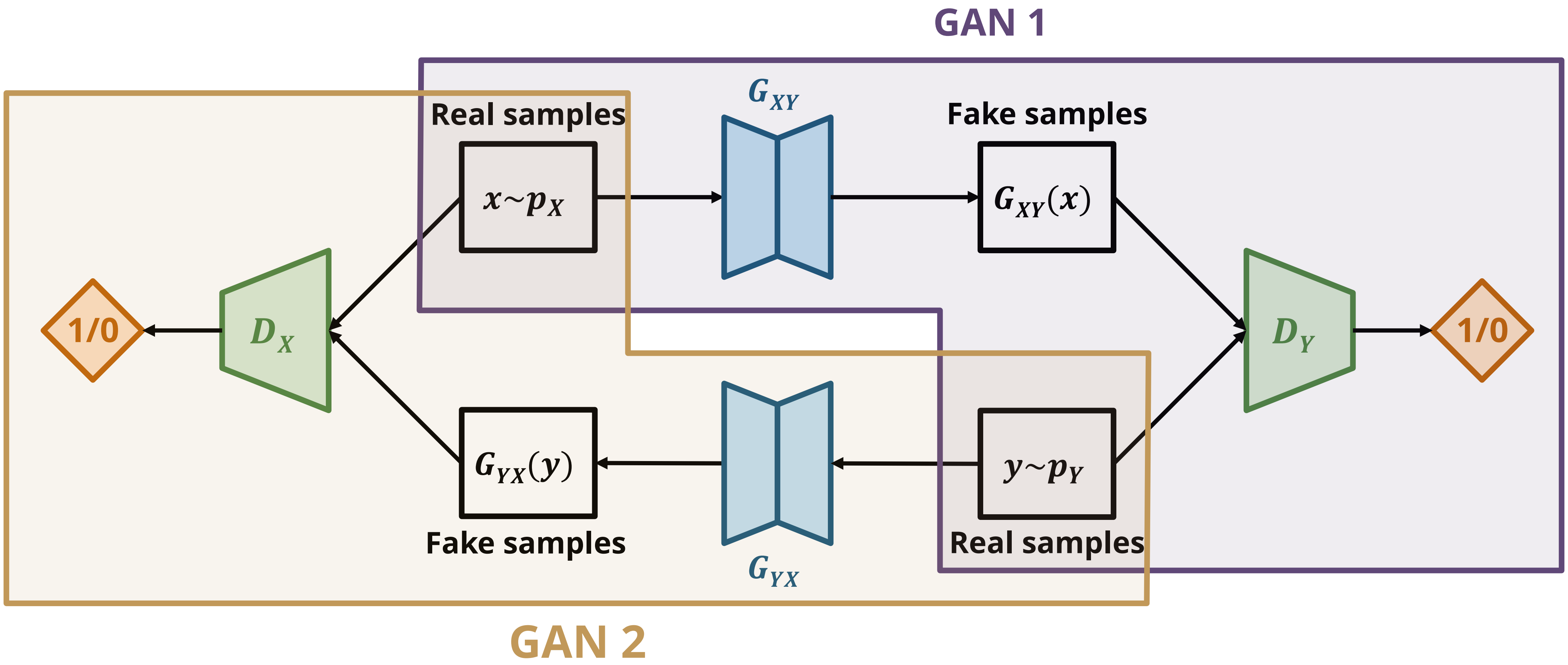
Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *ICCV*, 2017.

Cycle-consistent GAN (CycleGAN)



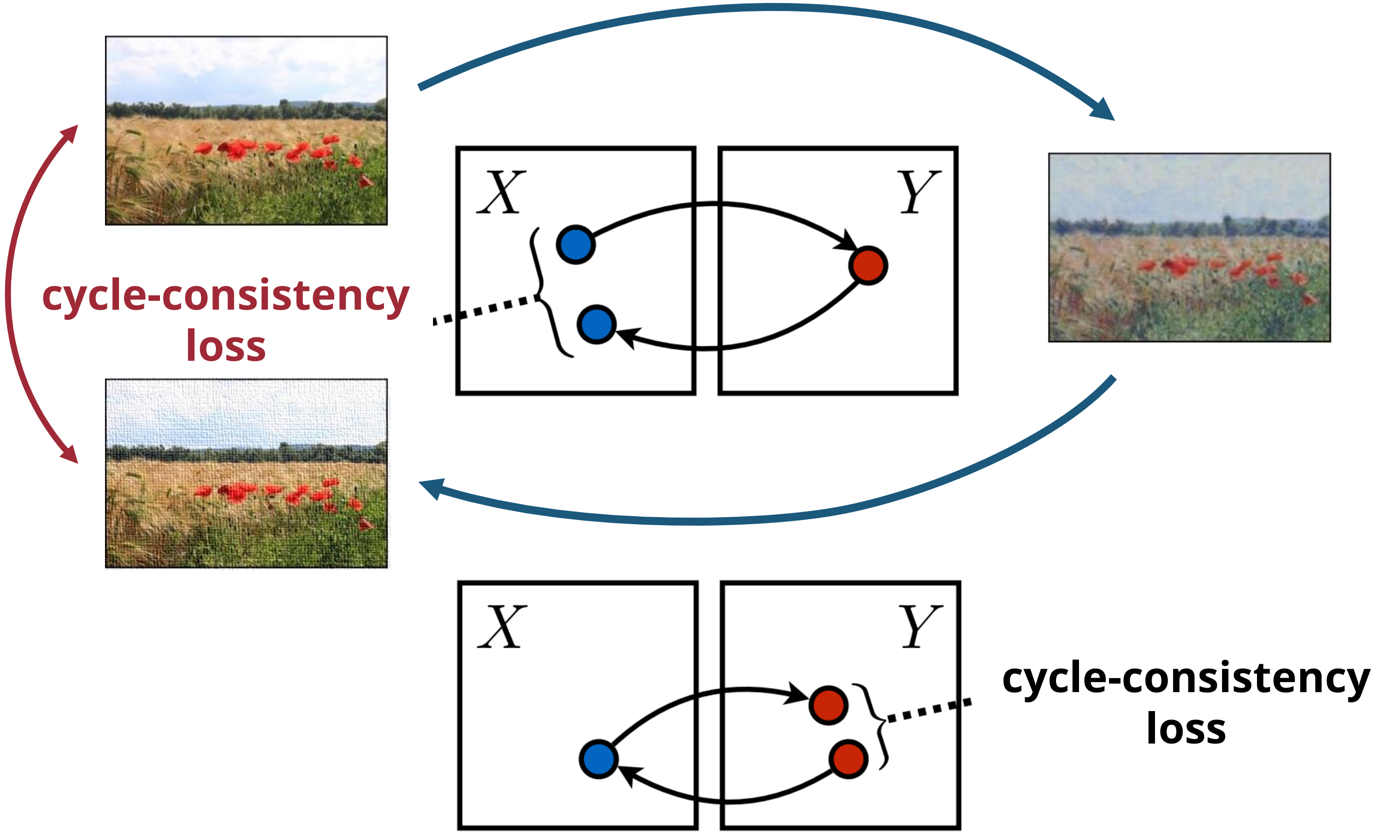
Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *ICCV*, 2017.

Cycle-consistent GAN (CycleGAN)



Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *ICCV*, 2017.

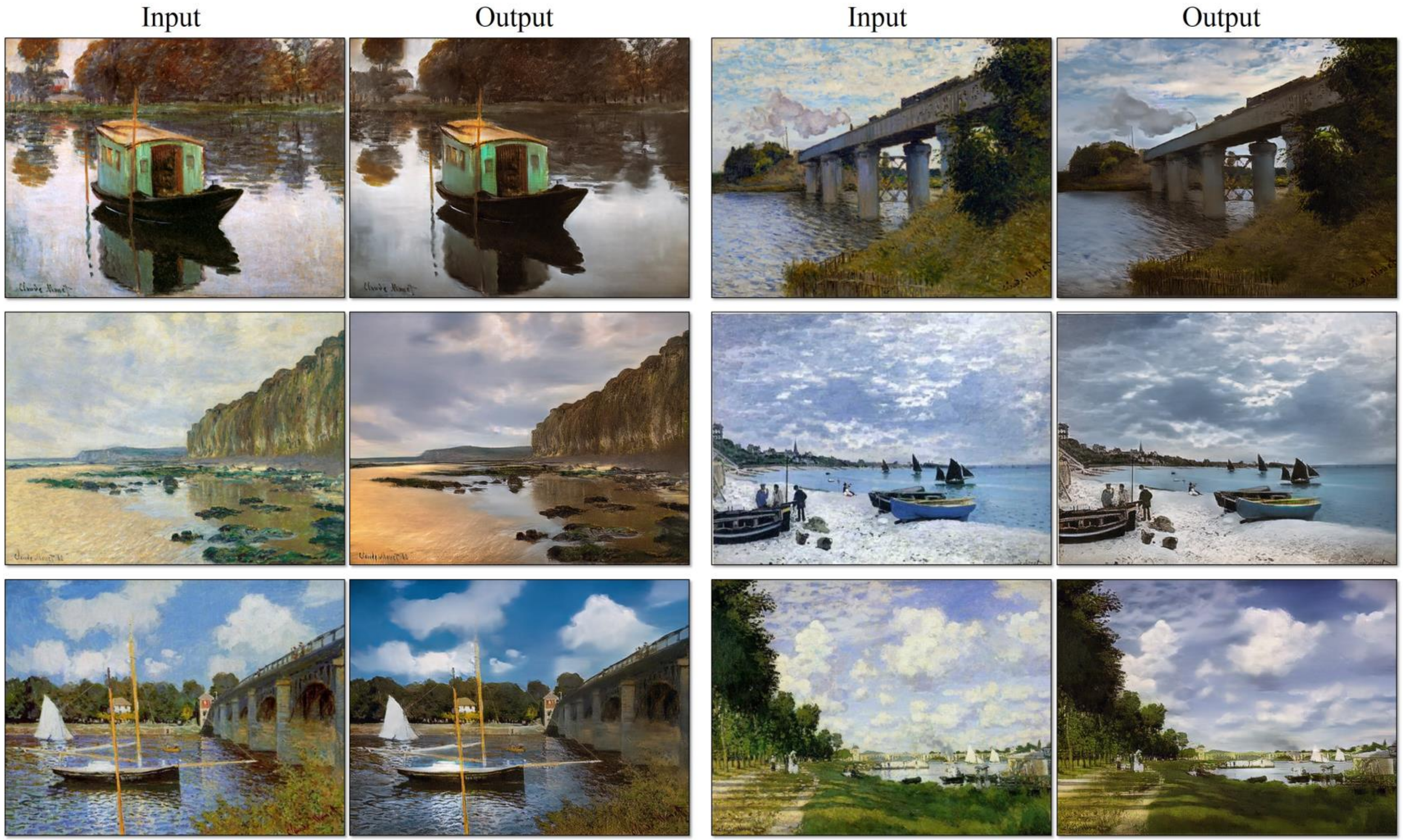
Cycle-consistency Loss



We only need unpaired samples in two domains

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *ICCV*, 2017.

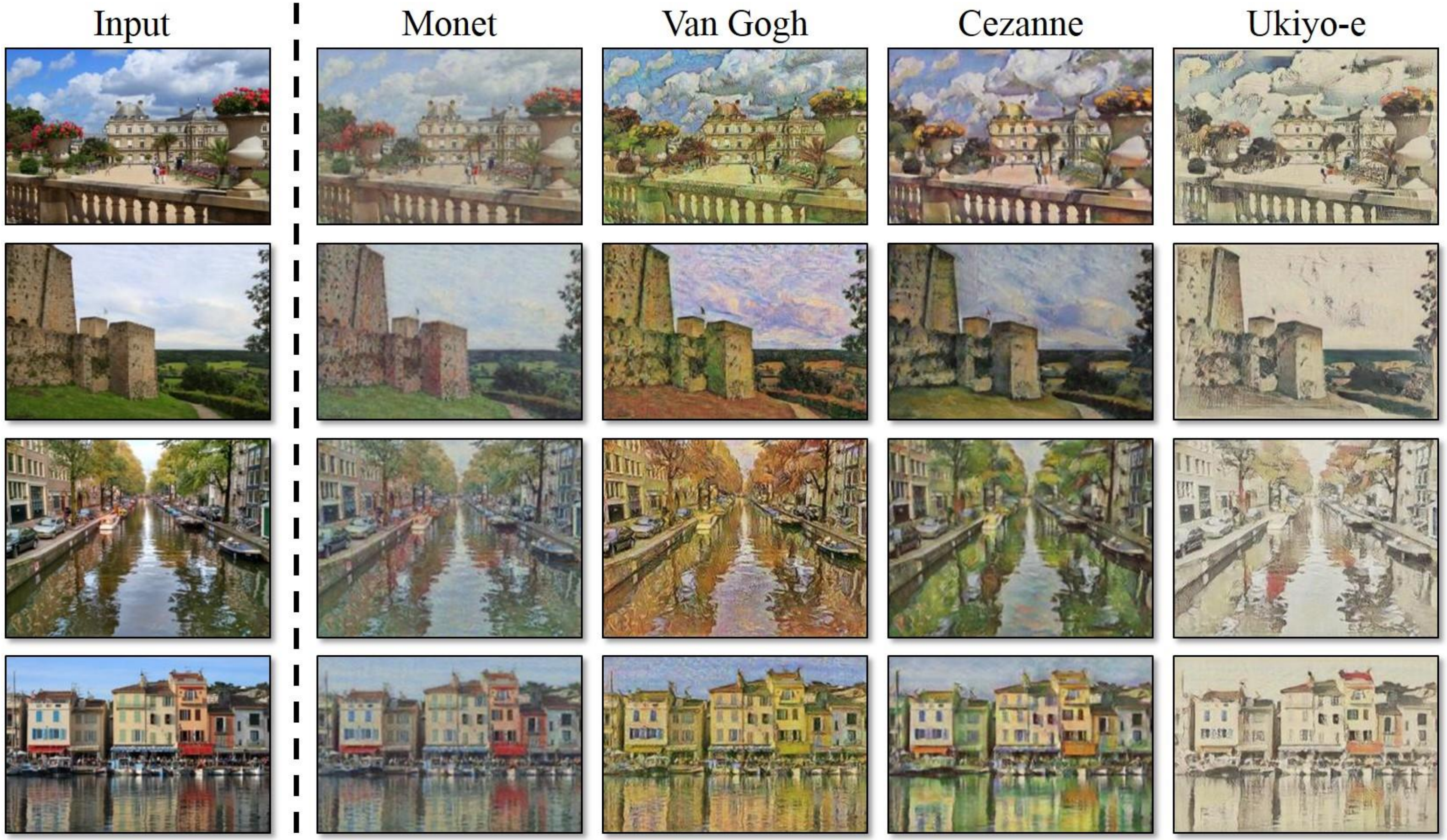
CycleGAN Examples – Monet to Photo



(Source: Zhu et al., 2017)

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *ICCV*, 2017.

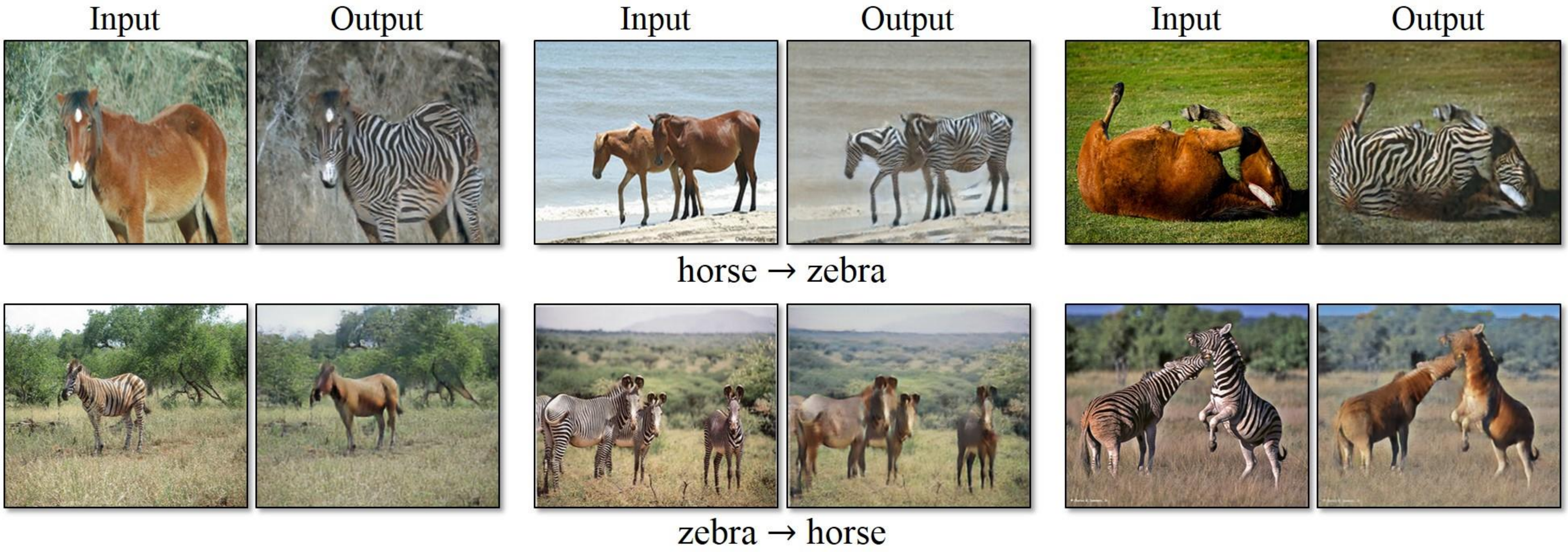
CycleGAN Examples – Artistic Style Transfer



(Source: Zhu et al., 2017)

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *ICCV*, 2017.

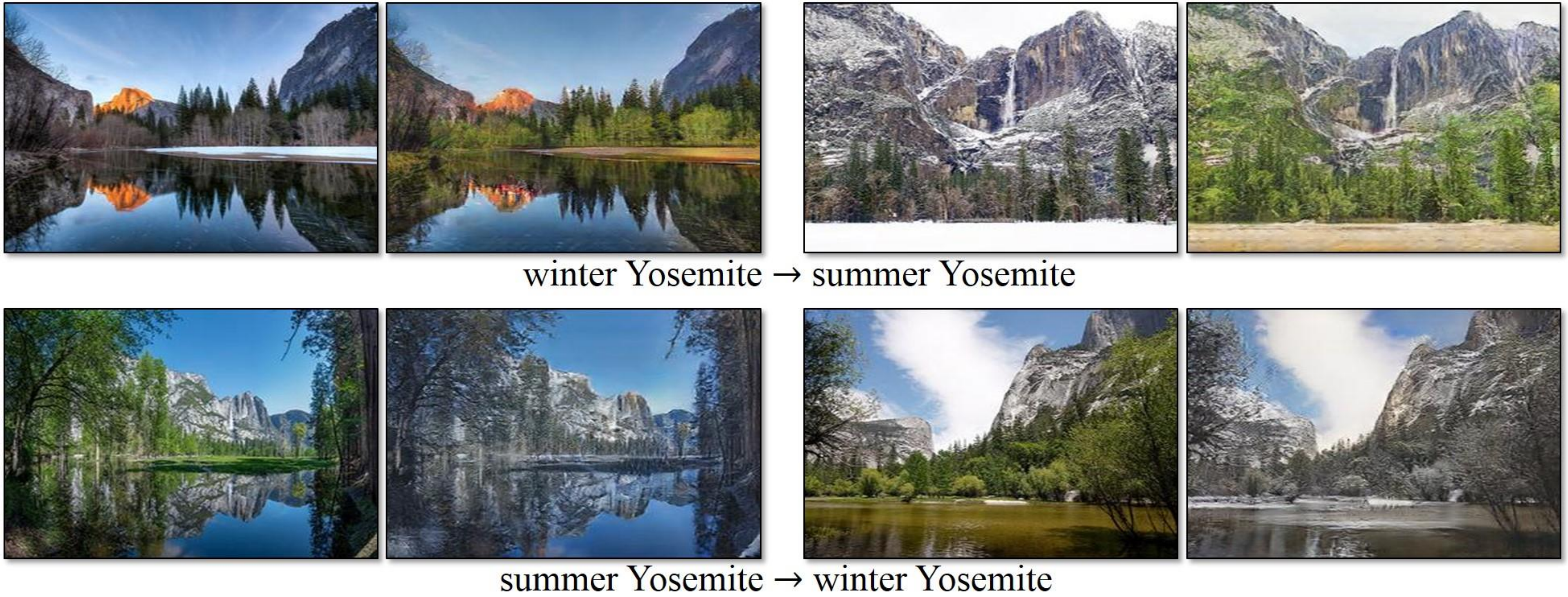
CycleGAN Examples – Object Transfer



(Source: Zhu et al., 2017)

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” *ICCV*, 2017.

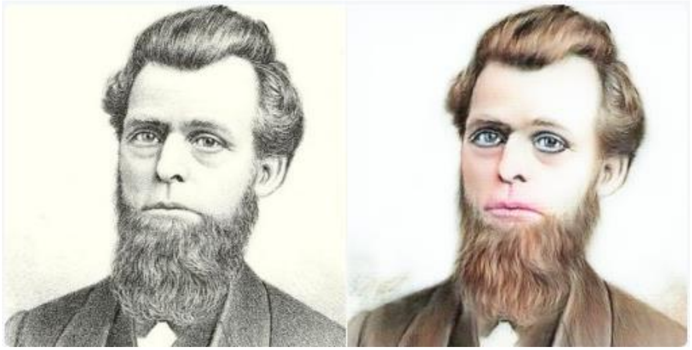
CycleGAN Examples – Season Transfer



(Source: Zhu et al., 2017)

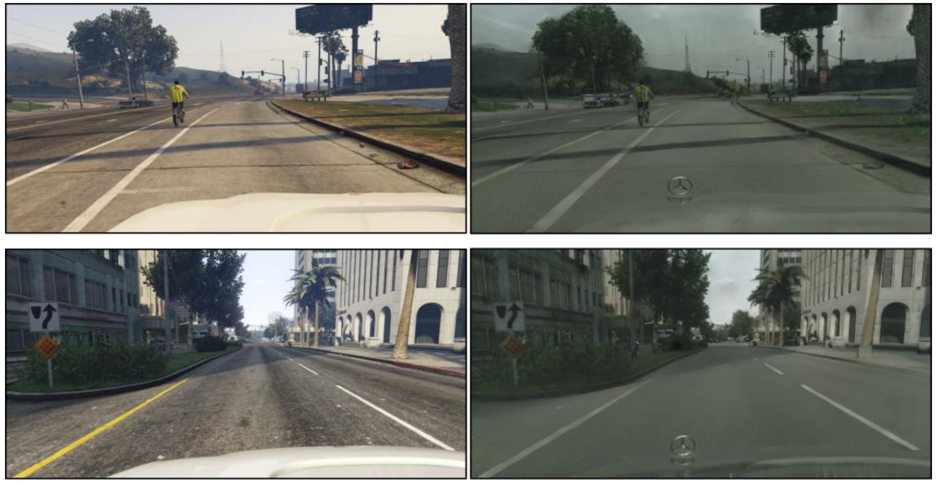
Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *ICCV*, 2017.

CycleGAN Examples – And Many More!



B&W → Color

Old city plan → Map → Satellite



GTA screenshot → Cityscape

Dog → Cat

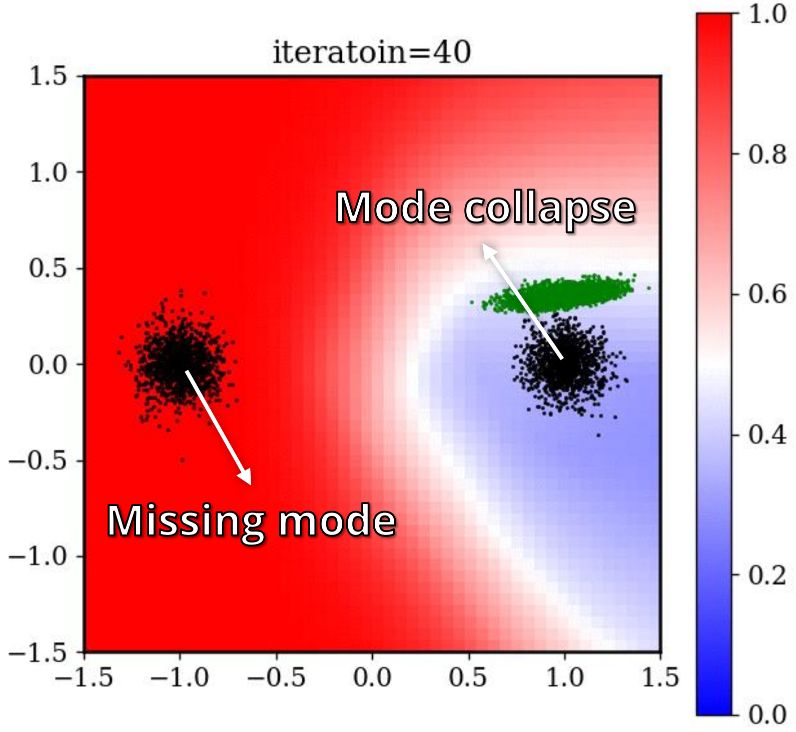
Cat → Dog

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *ICCV*, 2017.

Problems of Unregularized GANs

- **Key**—discriminator provides generator with gradients as a **guidance for improvement**
  - Discriminator has an easier job than the generator
  - Discriminator tends to provide large gradients
  - Results in unstable training of the generator
- Common failure cases
  - **Mode collapse**
  - **Missing modes**

A steep color change indicates a large gradient



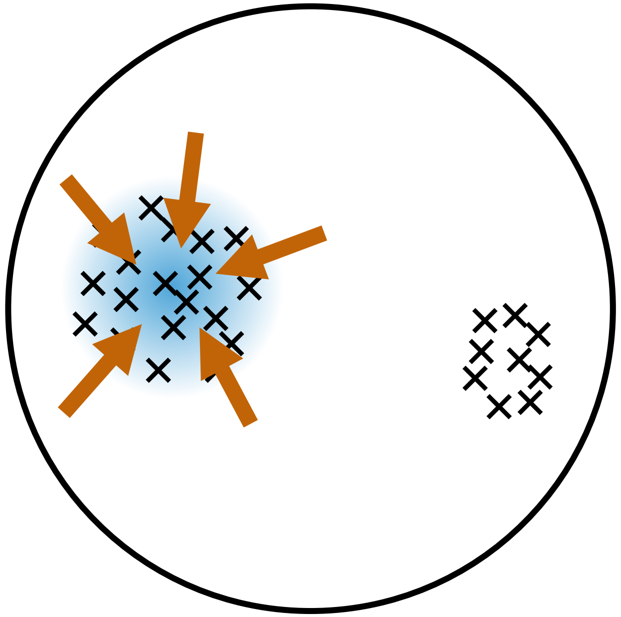
(Colors show the discriminator outputs)

Regularizing GANs

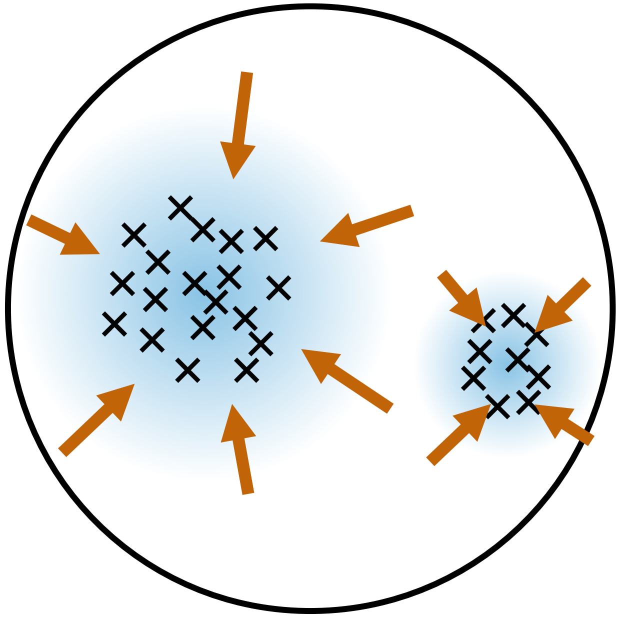
Advantages of gradient regularization

- **Provide a smoother guidance to the generator**
- **Alleviate mode collapse and missing modes issues**

Unregularized



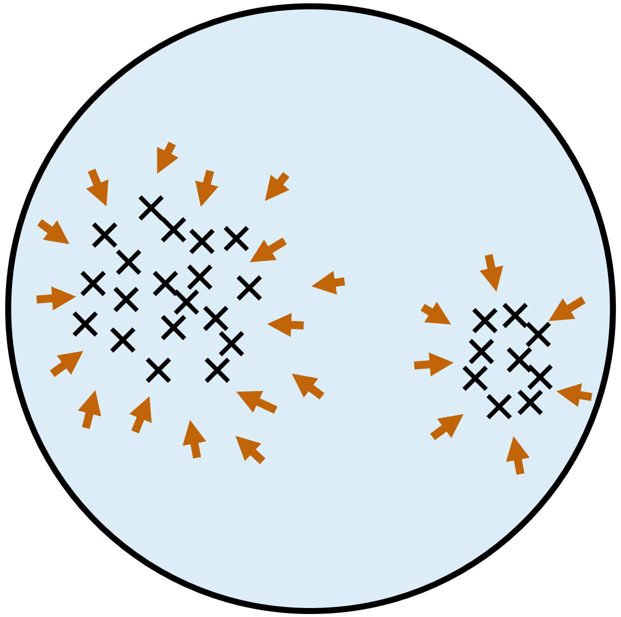
Locally regularized



Gradient clipping [1]

Gradient penalties [2,3]

Globally regularized



Spectral normalization [4]

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein Generative Adversarial Networks," *ICML,* 2017.
[2] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville, "Improved Training of Wasserstein GANs," *NeurIPS,* 2017.
[3] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira, "On Convergence and Stability of GANs," *arXiv preprint arXiv:1705.07215,* 2017.
[4] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida, "Spectral Normalization for Generative Adversarial Networks," *ICLR,* 2018.