PAT 498/598 (Fall 2024)

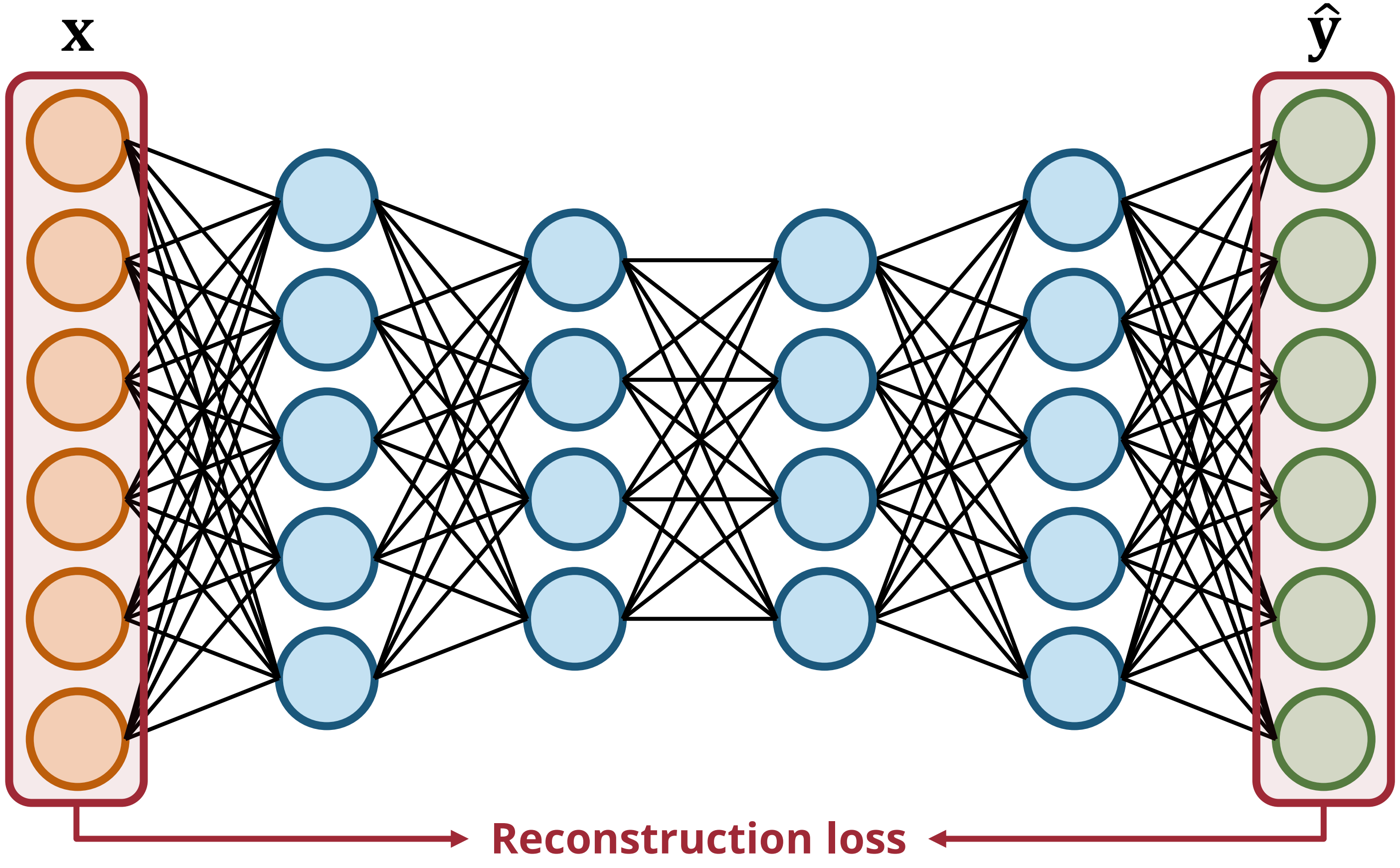
Special Topics:
Generative AI for Music and Audio Creation

Lecture 10: Diffusion Models

Instructor: Hao-Wen Dong

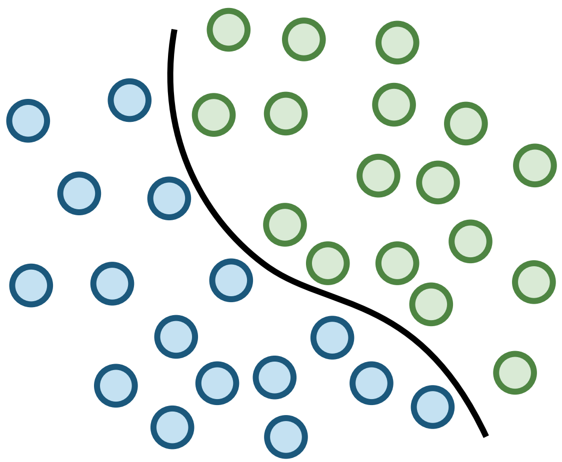
(Recap) Autoencoders

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



(Recap) Discriminative vs Generative Models

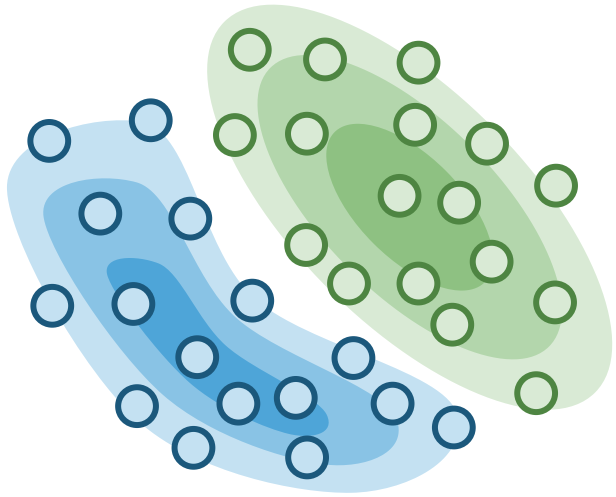
Discriminative



Discriminative models learn the decision boundary

$P(y|x)$

Generative



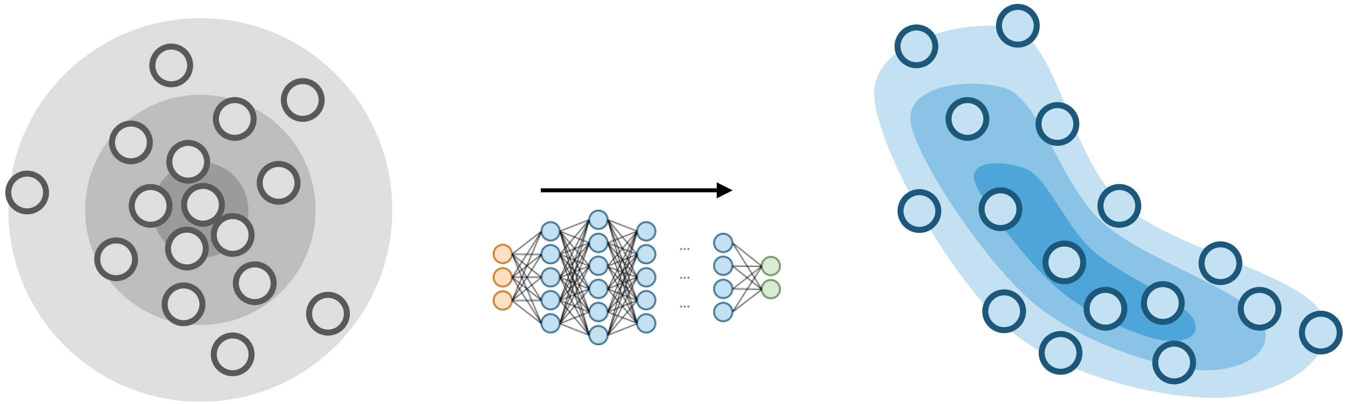
Generative models learn the underlying distribution

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

(Recap) Generating Data from a Random Distribution

Random distribution

Data distribution

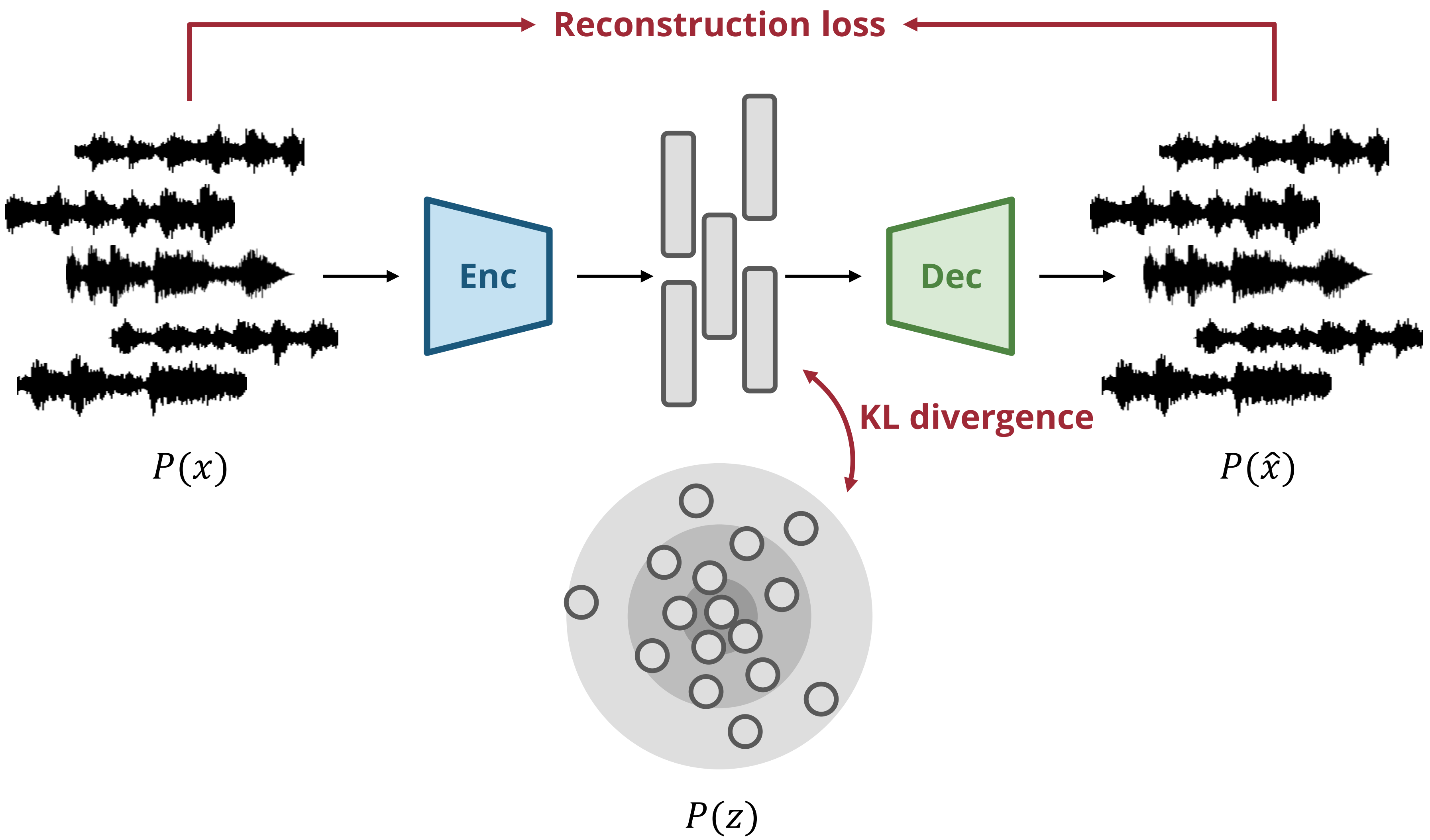


$P(z)$

$P(x)$

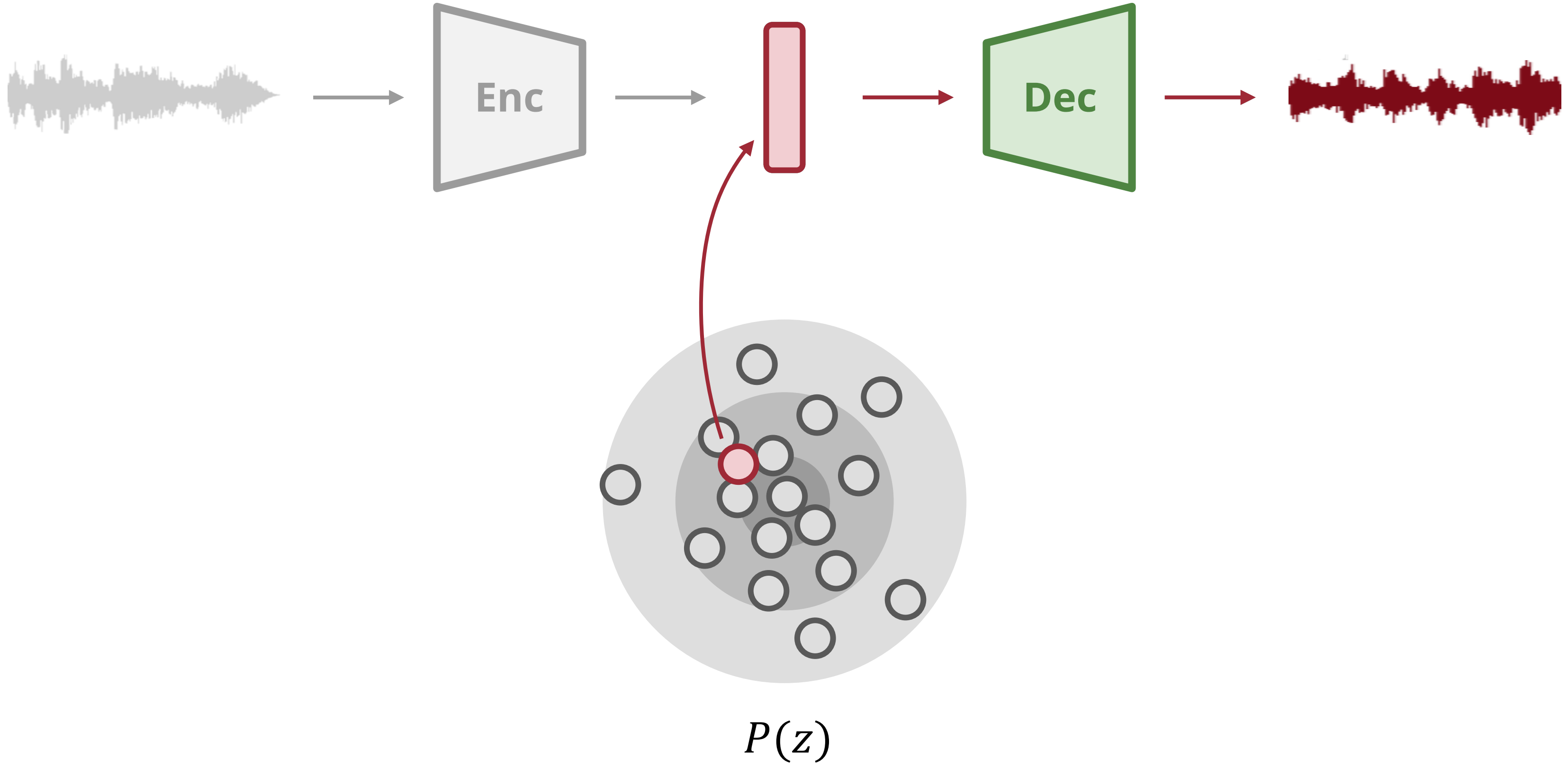
If we can learn this mapping, we can easily generate new samples from the data distribution

(Recap) Variational Autoencoders (VAEs) – Training



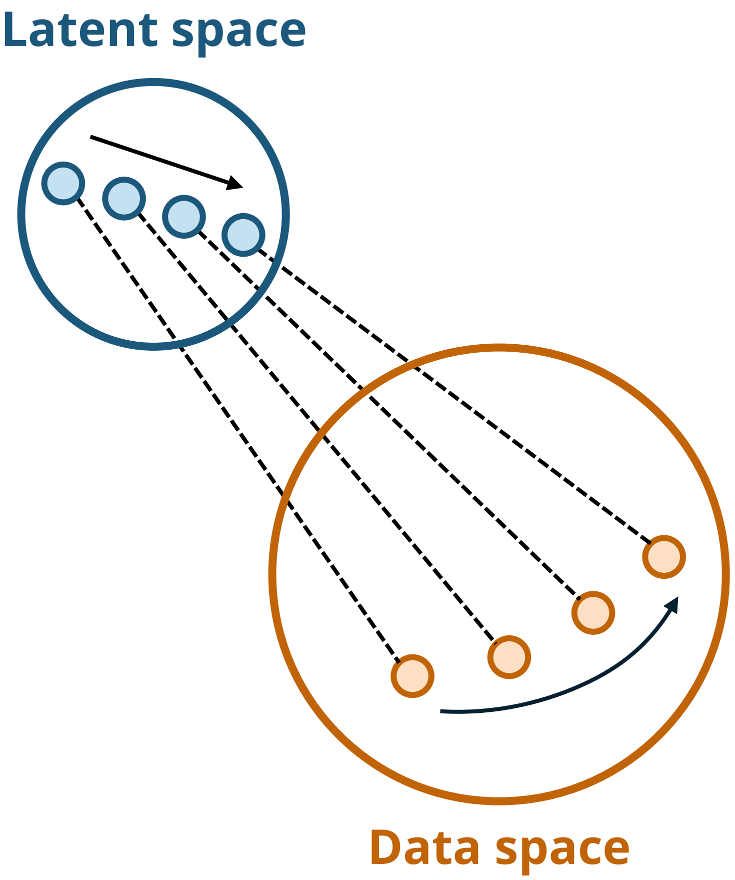
Diederik P Kingma and Max Welling, "Auto-Encoding Variational Bayes," *ICLR*, 2014.

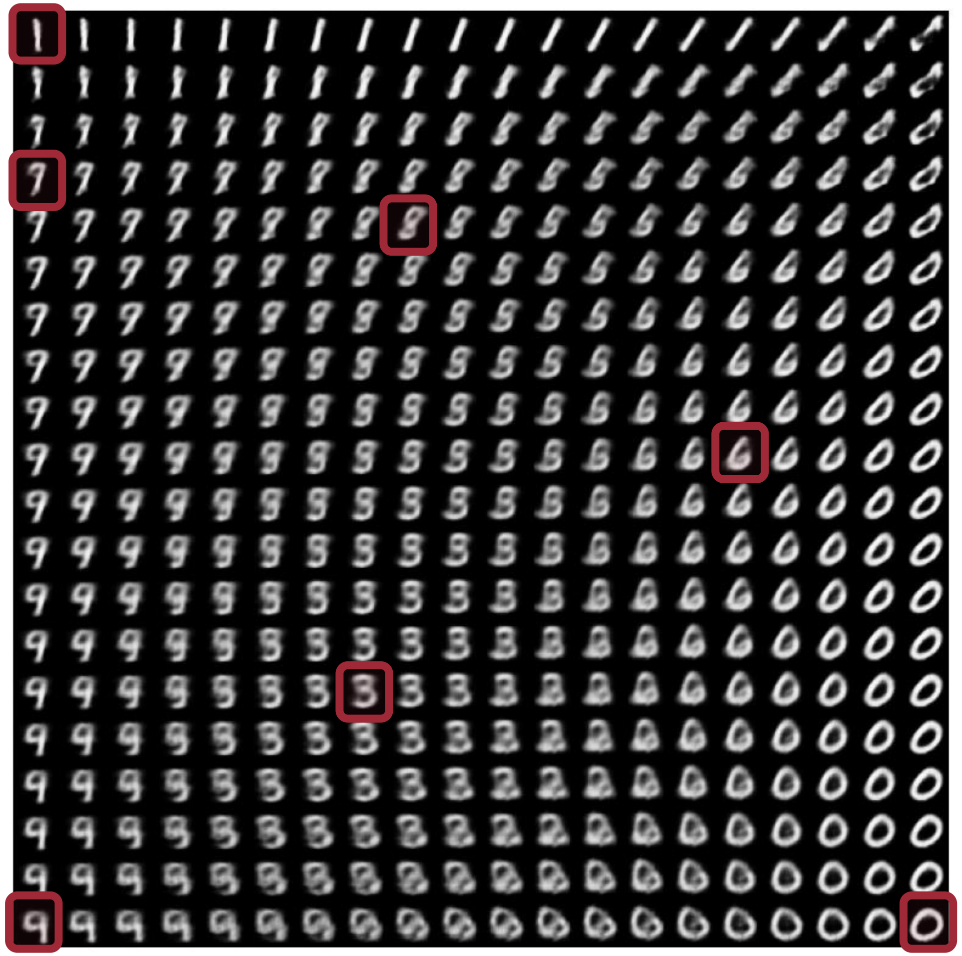
(Recap) Variational Autoencoders (VAEs) – Generation



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

(Recap) Decoding the Latent Space of a VAE

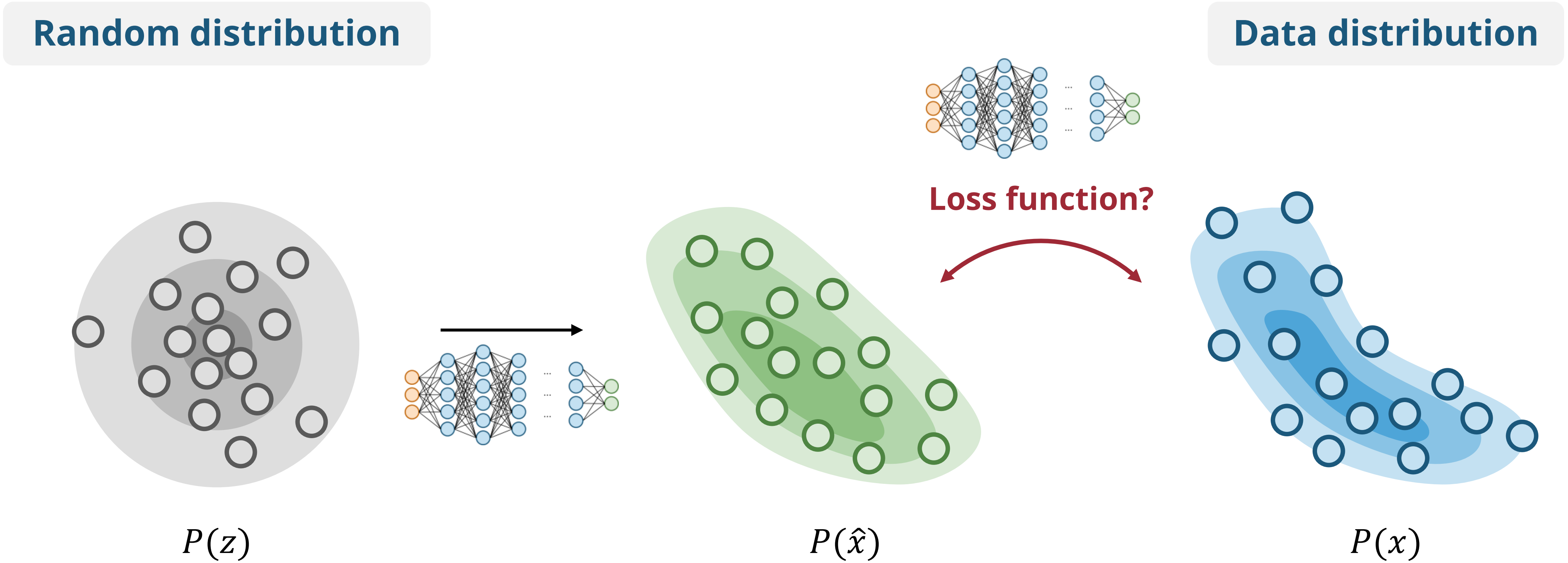




(Source: tensorflow.org)

tensorflow.org/tutorials/generative/autoencoder

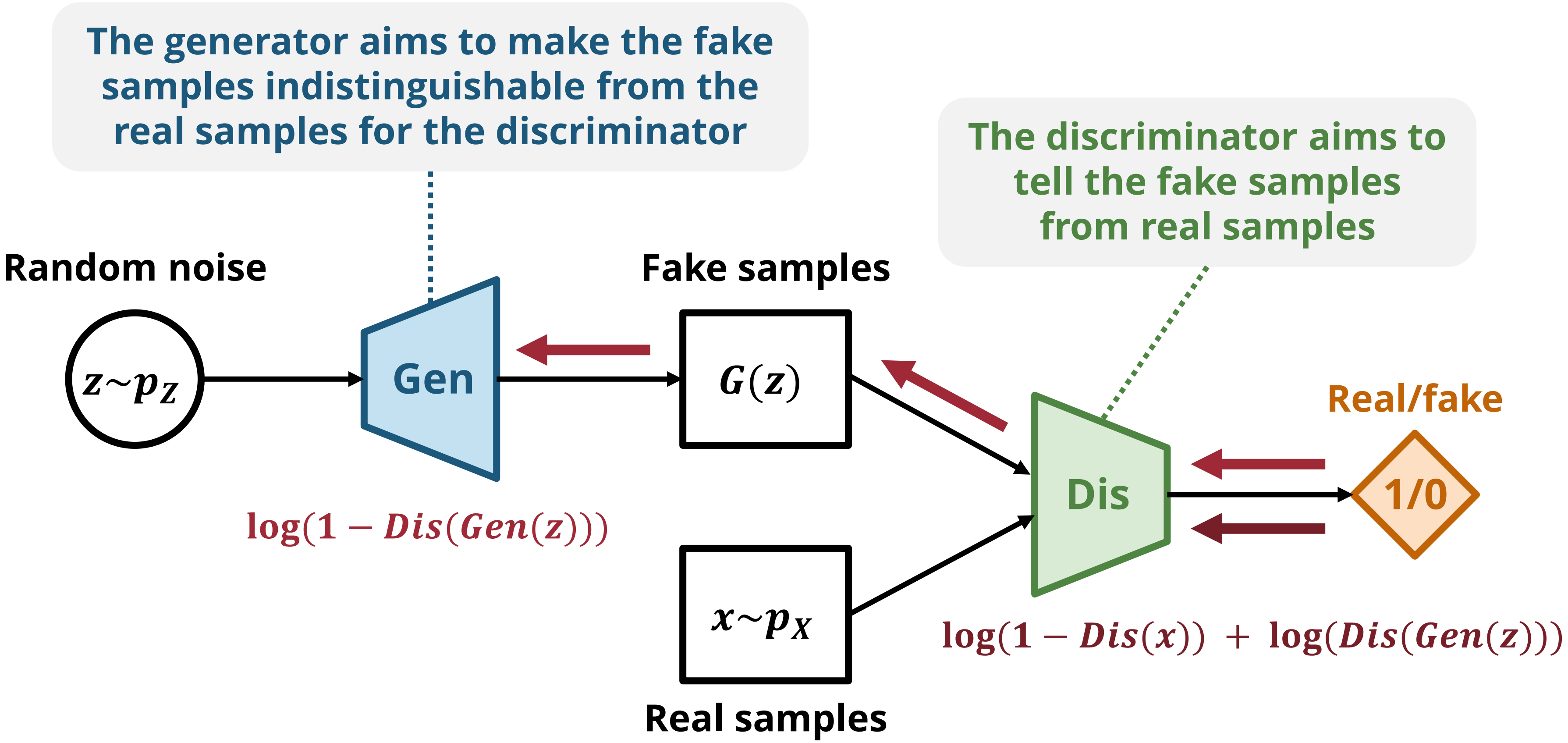
(Recap) A Loss Function for Distributions



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

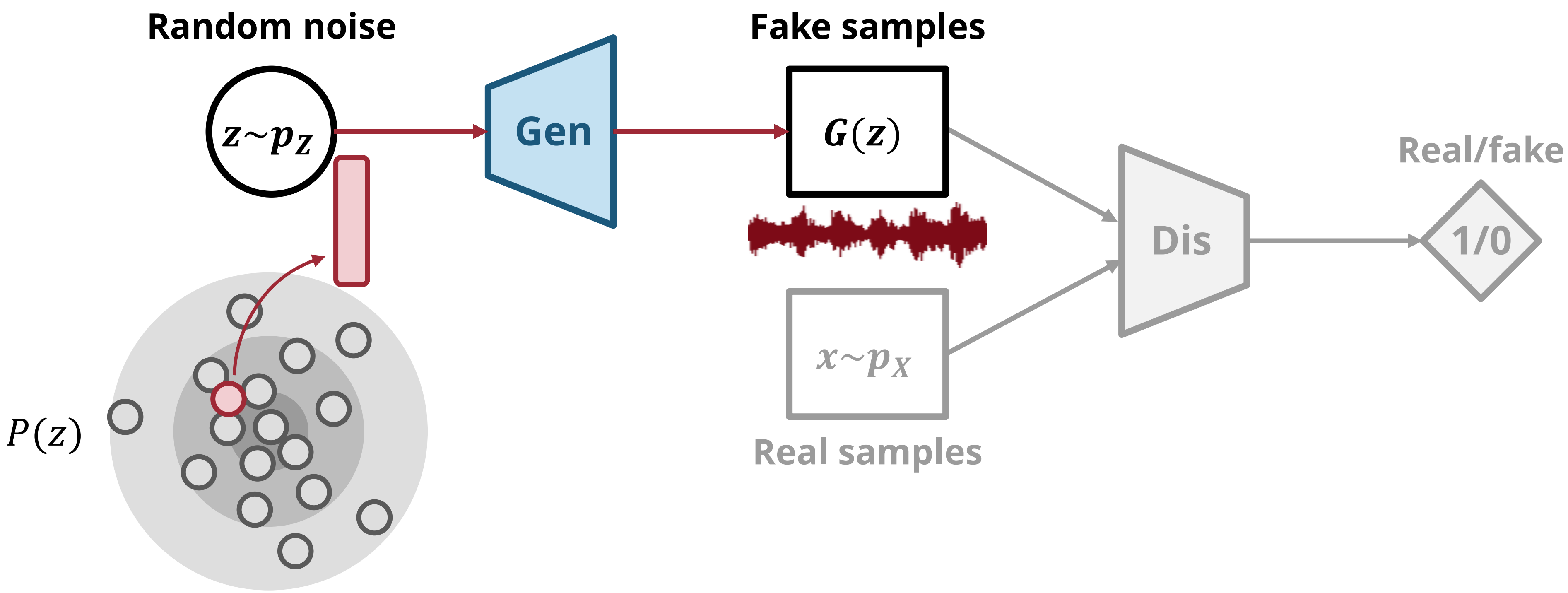
But what about another neural network!?

(Recap) Generative Adversarial Nets (GANs) – Training



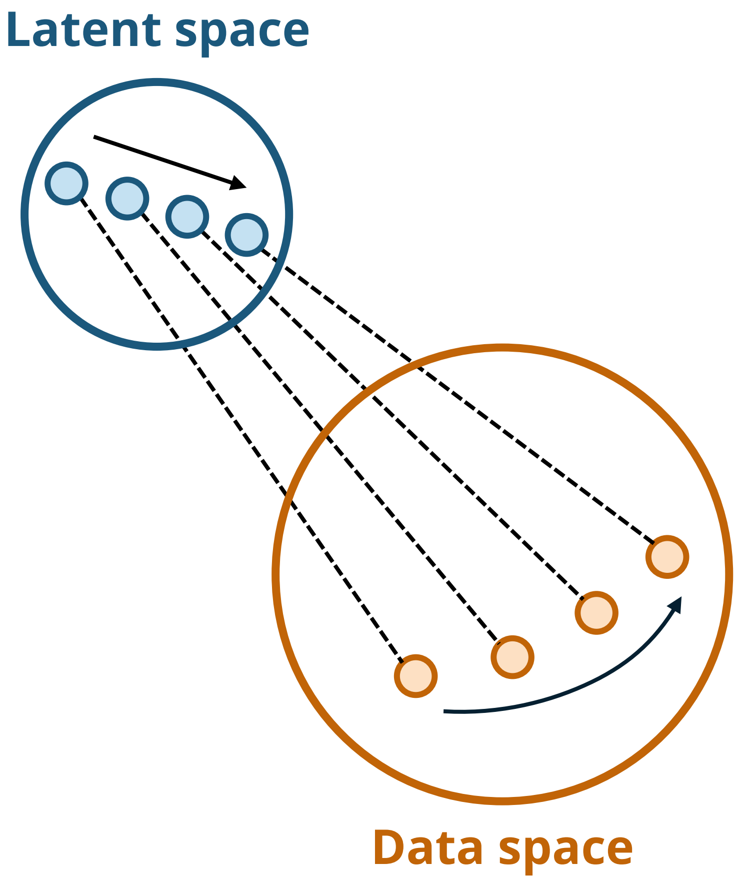
Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Networks," *NeurIPS*, 2014.

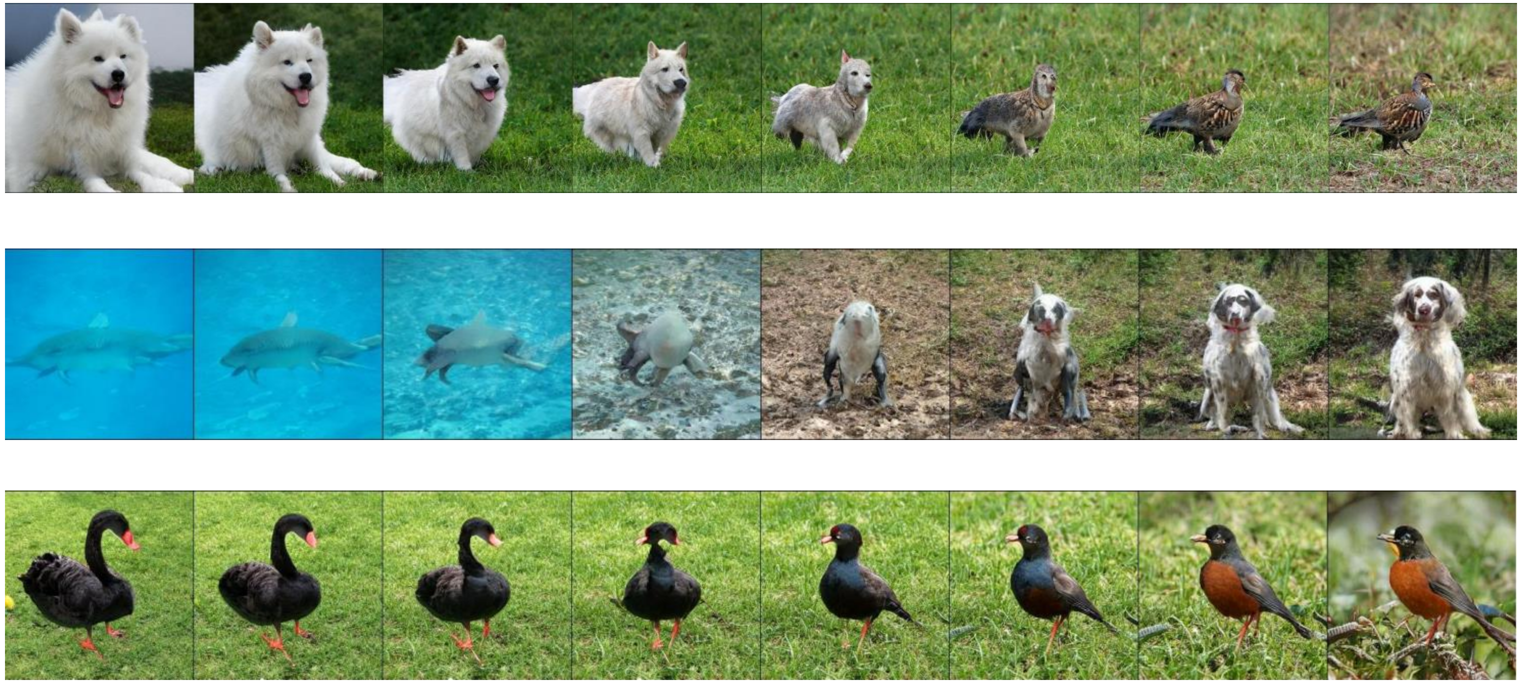
(Recap) Generative Adversarial Nets (GANs) – Generation



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.

(Recap) Interpolation on the Latent Space





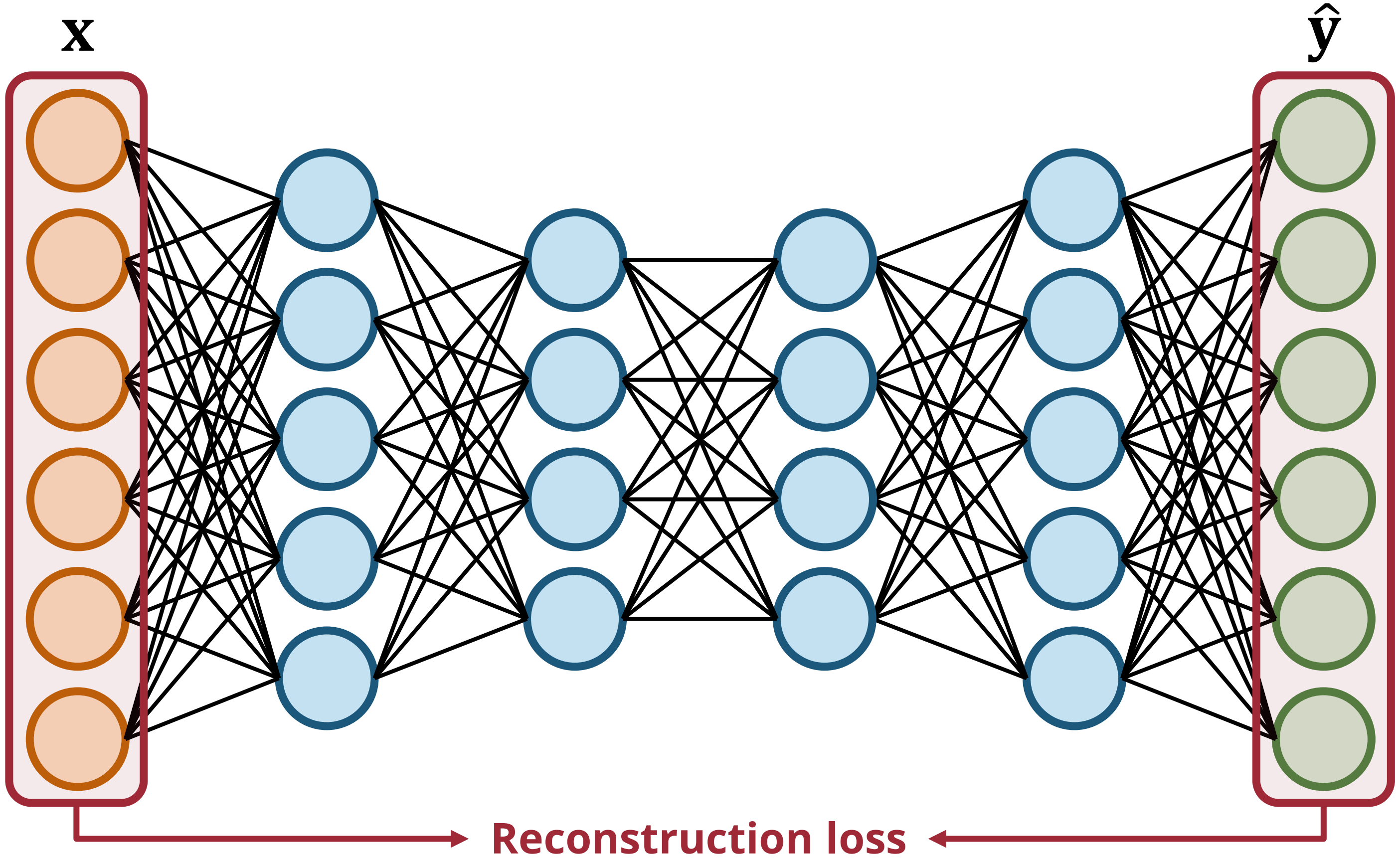
(Source: Brock et al., 2019)

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

Diffusion Models

(Recap) Autoencoders

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



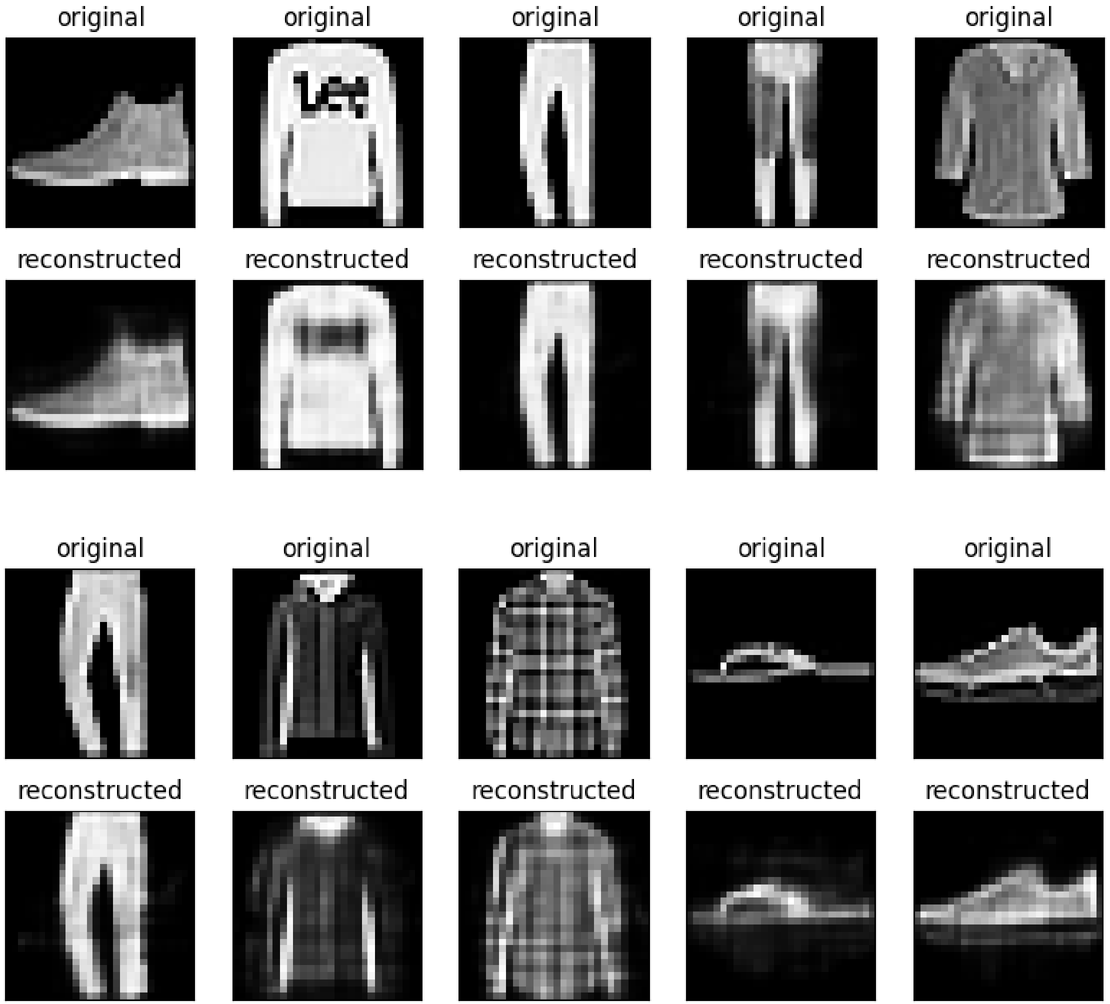
(Recap) Autoencoders – Reconstruction Examples

Original

Reconstructed

Original

Reconstructed



(Source: tensorflow.org)

tensorflow.org/tutorials/generative/autoencoder

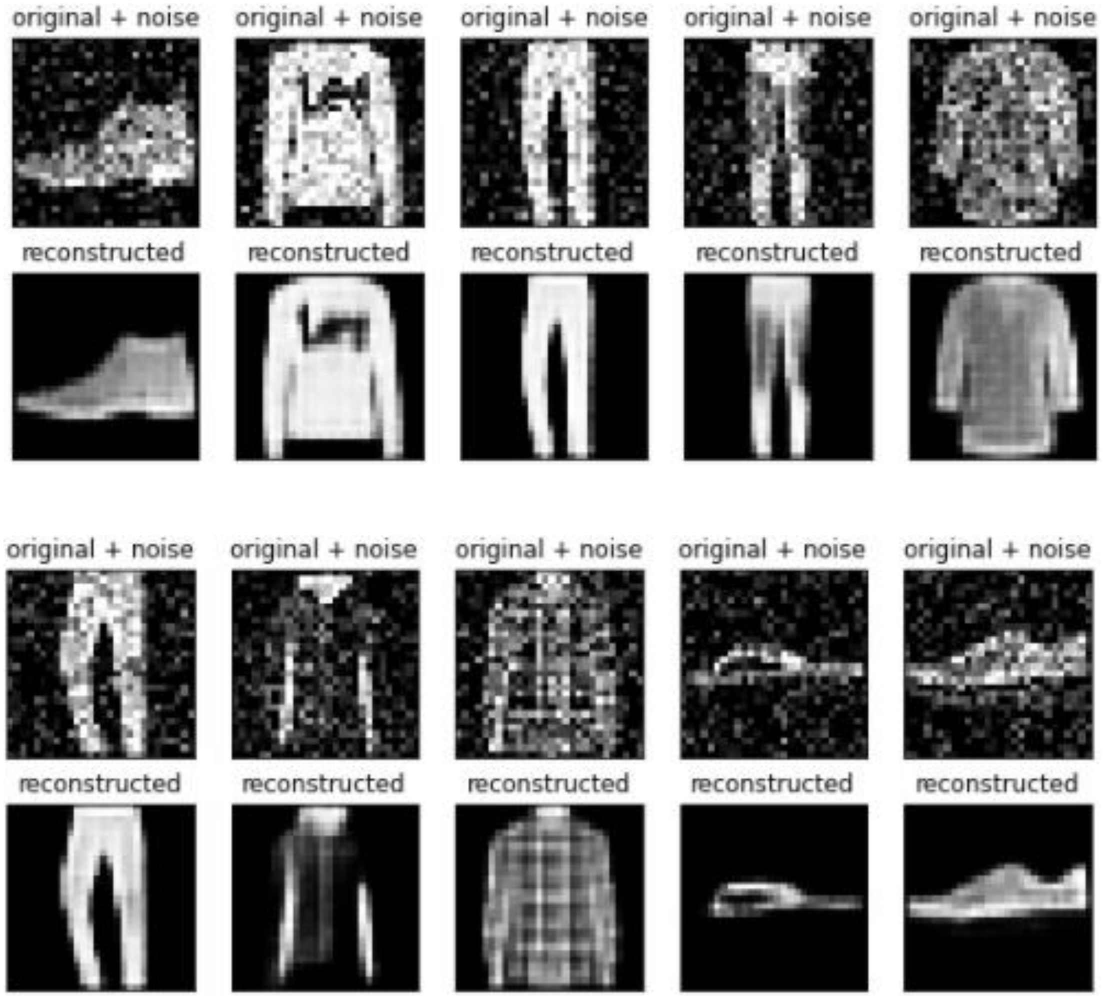
Denoising Autoencoders

Noisy

Reconstructed

Noisy

Reconstructed



(Source: tensorflow.org)

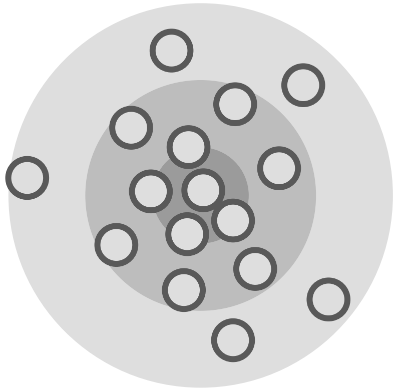
Denoising autoencoders learn to reconstruct noisy inputs

tensorflow.org/tutorials/generative/autoencoder

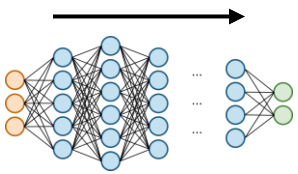
Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, "Extracting and Composing Robust Features with Denoising Autoencoders," *ICML*, 2008.

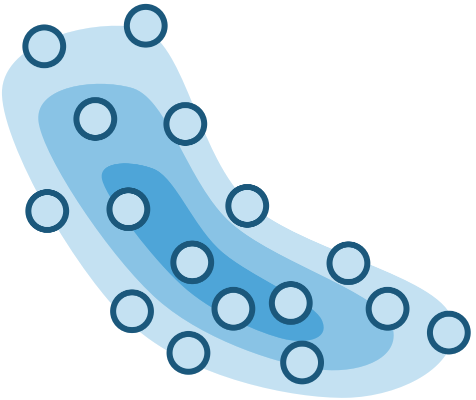
(Recap) Generating Data from a Random Distribution

Random distribution | **Data distribution**



$P(z)$



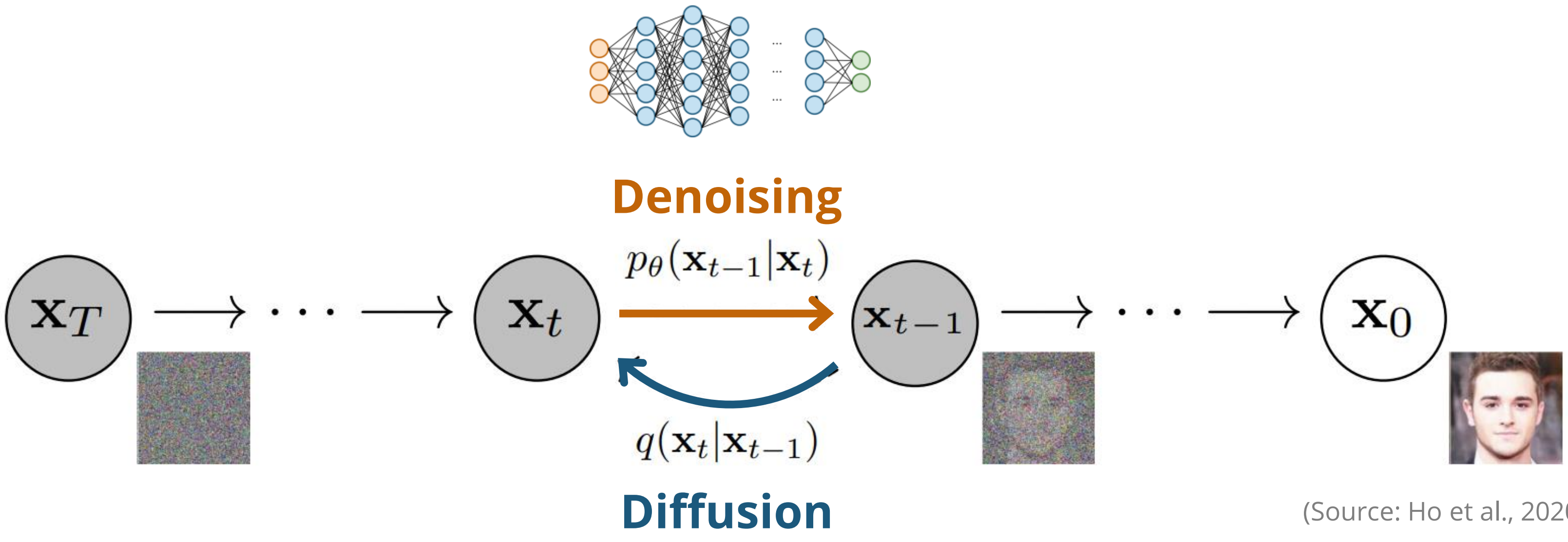


$P(x)$

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

Diffusion Models

- **Intuition**: Many denoising autoencoders stacked together

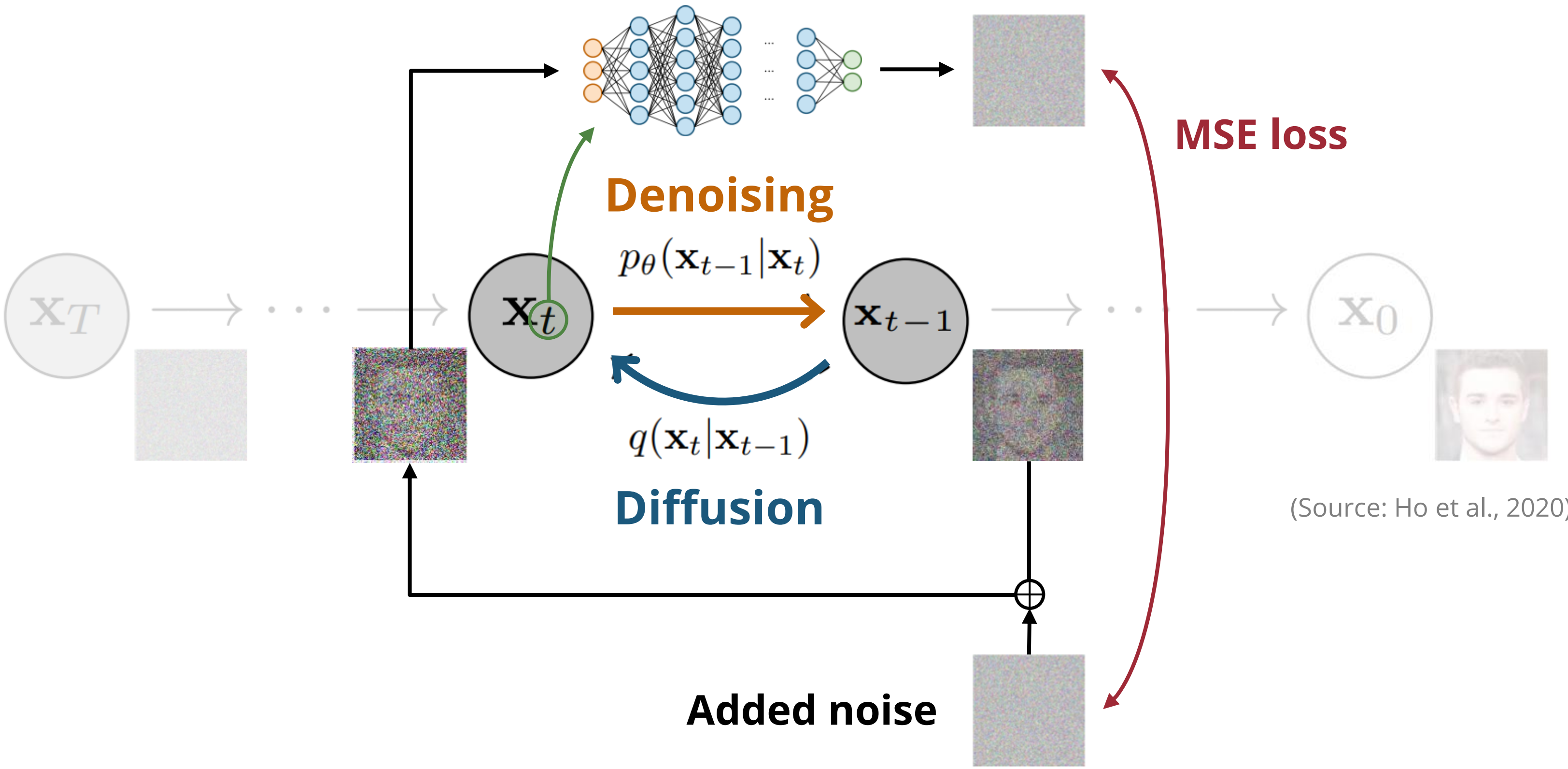


(Source: Ho et al., 2020)

Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.

Diffusion Models – Training

- **Intuition**: Many denoising autoencoders stacked together



(Source: Ho et al., 2020)

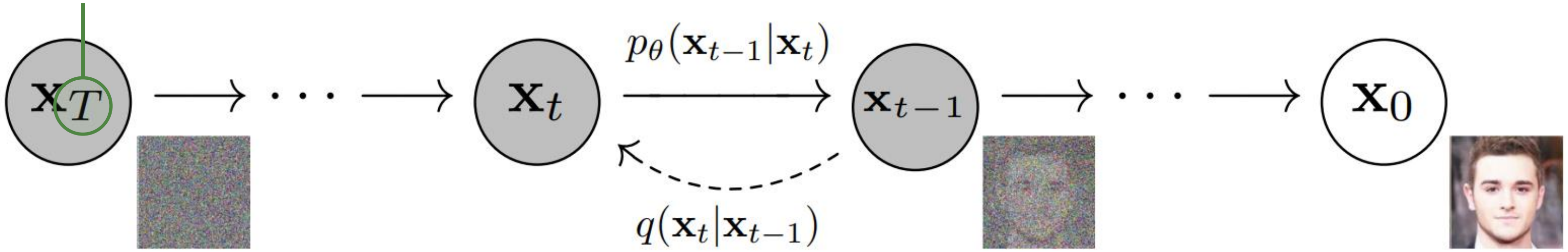
Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.

Diffusion Models

- **Intuition**: Many denoising autoencoders stacked together

Remove noise gradually
(Backward diffusion process)

Usually, $T > 1000$



(Source: Ho et al., 2020)

Add noise gradually
(Forward diffusion process)

Diffusion Models – Generation

Remove noise gradually
(Backward diffusion process)

Input **Output**

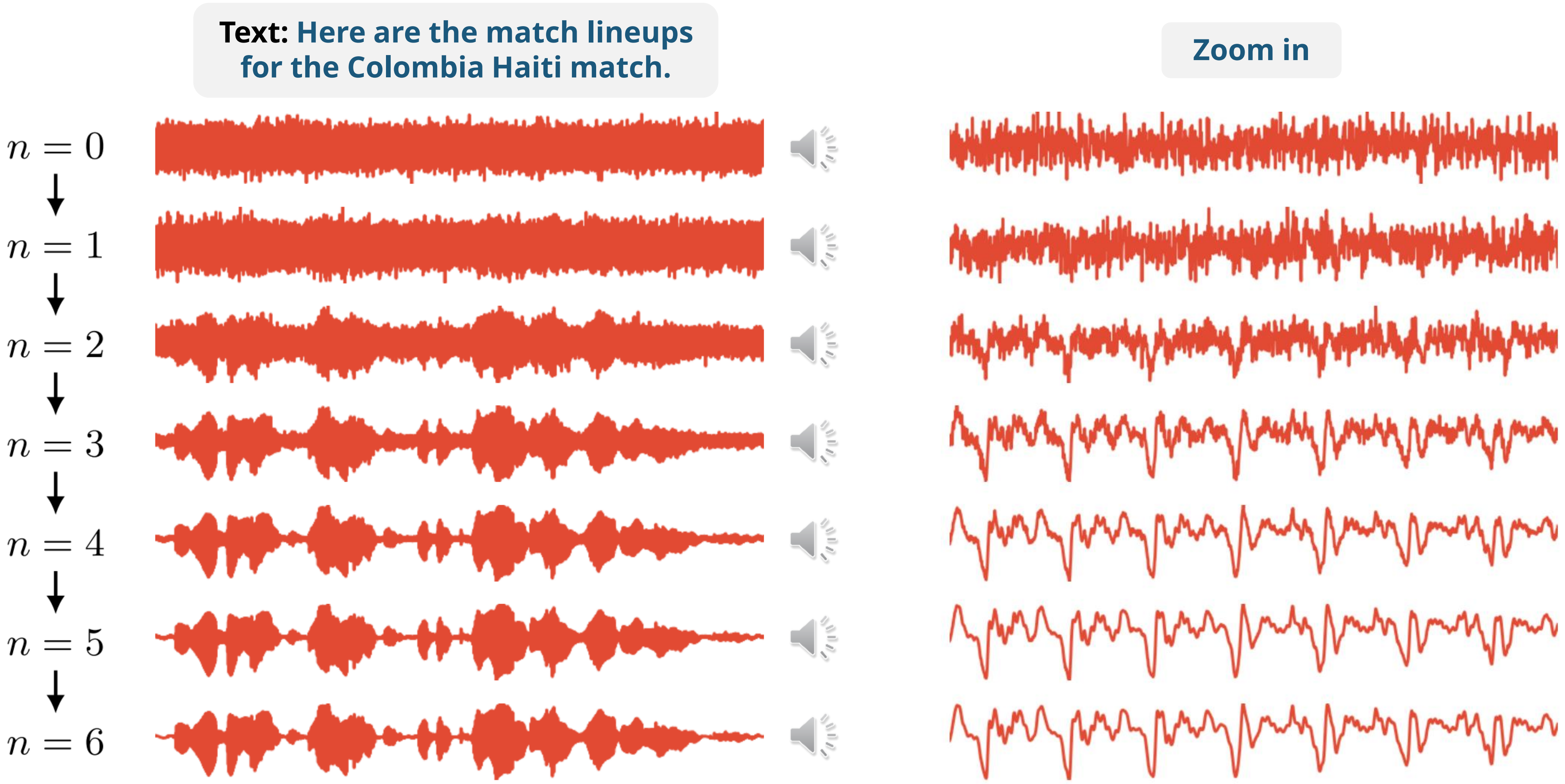
Coarse shapes
(low-frequency components)

Fine details
(high-frequency components)

(Source: Ho et al., 2020)

Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.

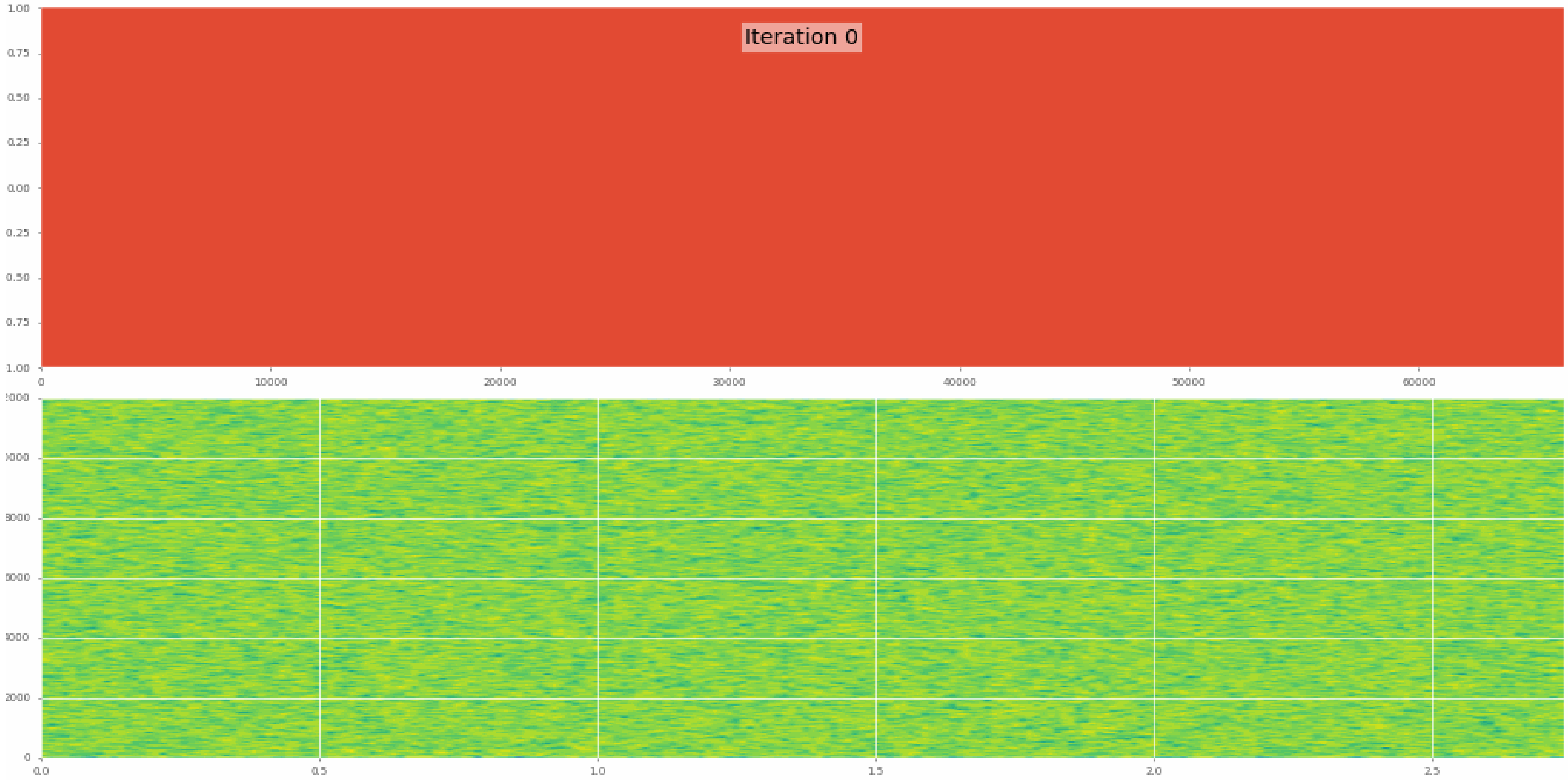
WaveGrad – Diffusion Model for Waveforms



(Source: Chen et al., 2021)

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan, "WaveGrad: Estimating Gradients for Waveform Generation," *ICLR*, 2021.

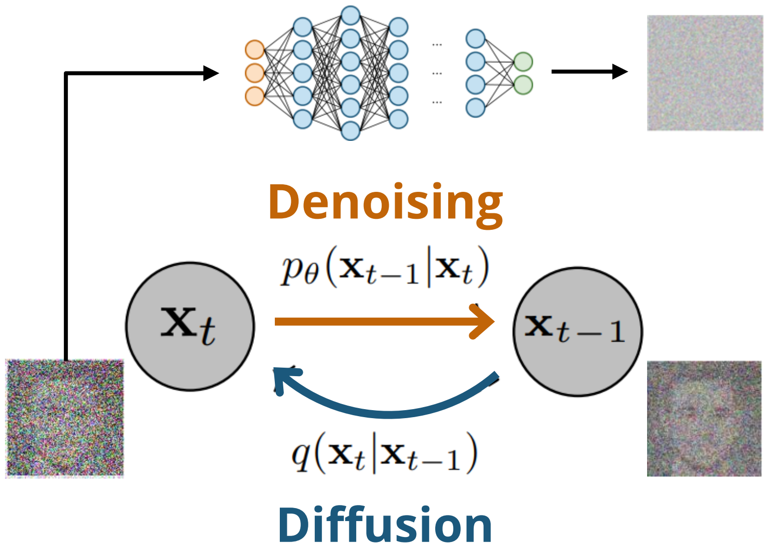
WaveGrad – Diffusion for Waveforms



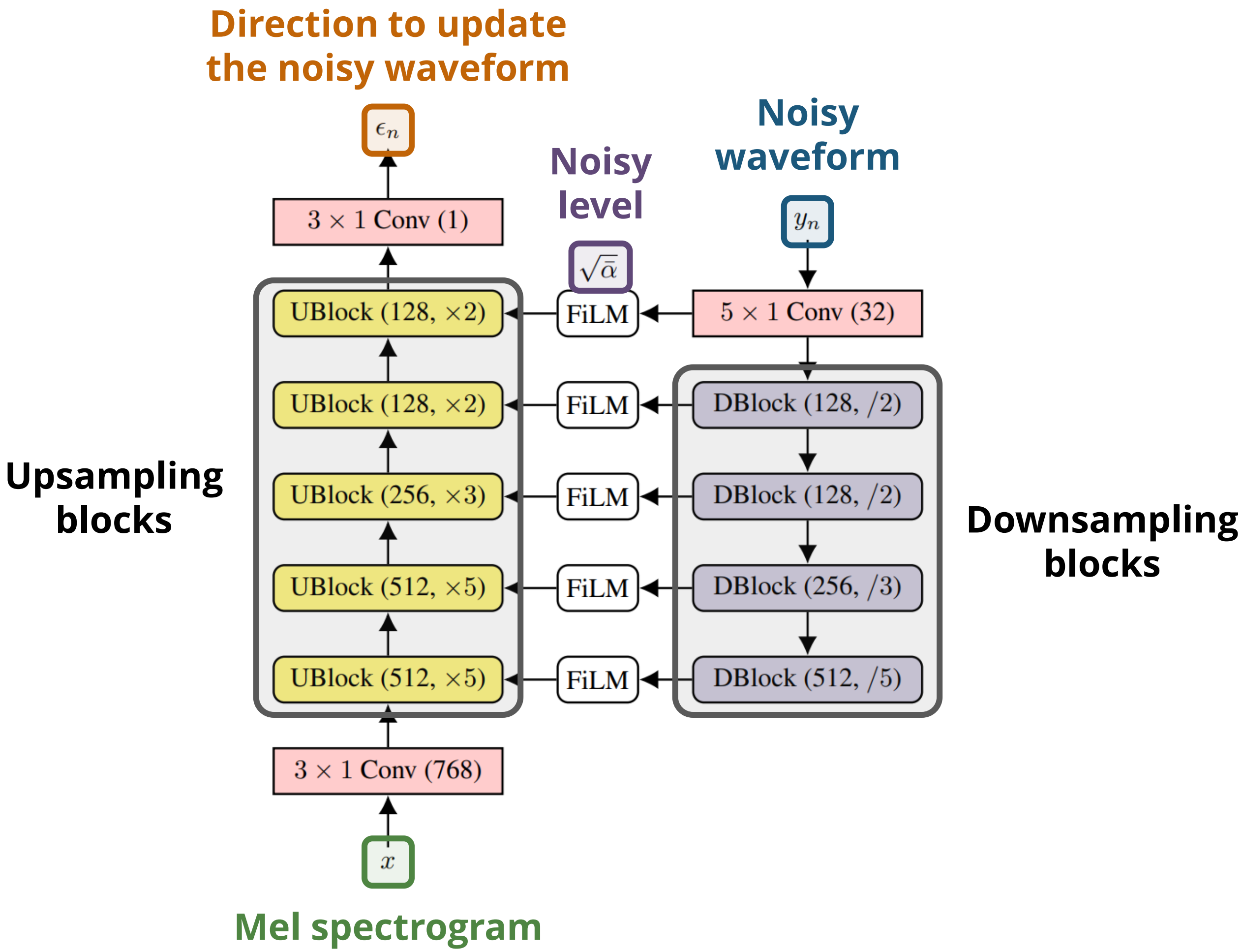
(Source: Chen et al., 2021)

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan, "WaveGrad: Estimating Gradients for Waveform Generation," *ICLR*, 2021.

WaveGrad – Diffusion for Waveforms



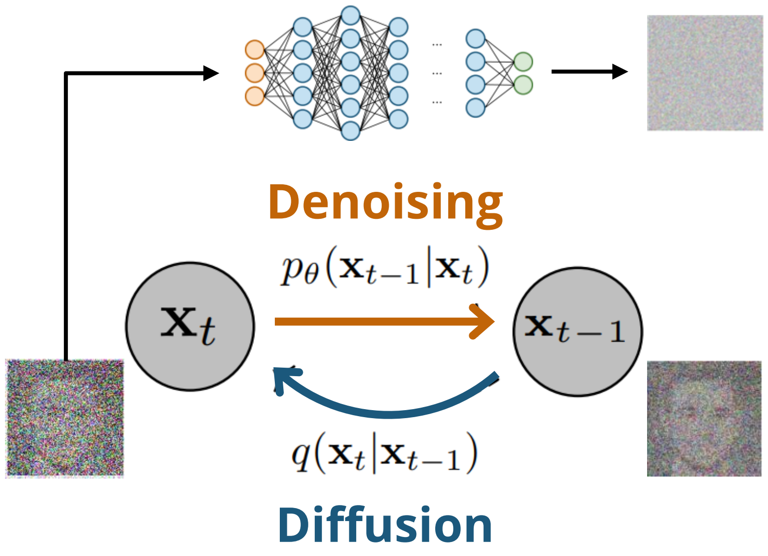
(Source: Ho et al., 2020)



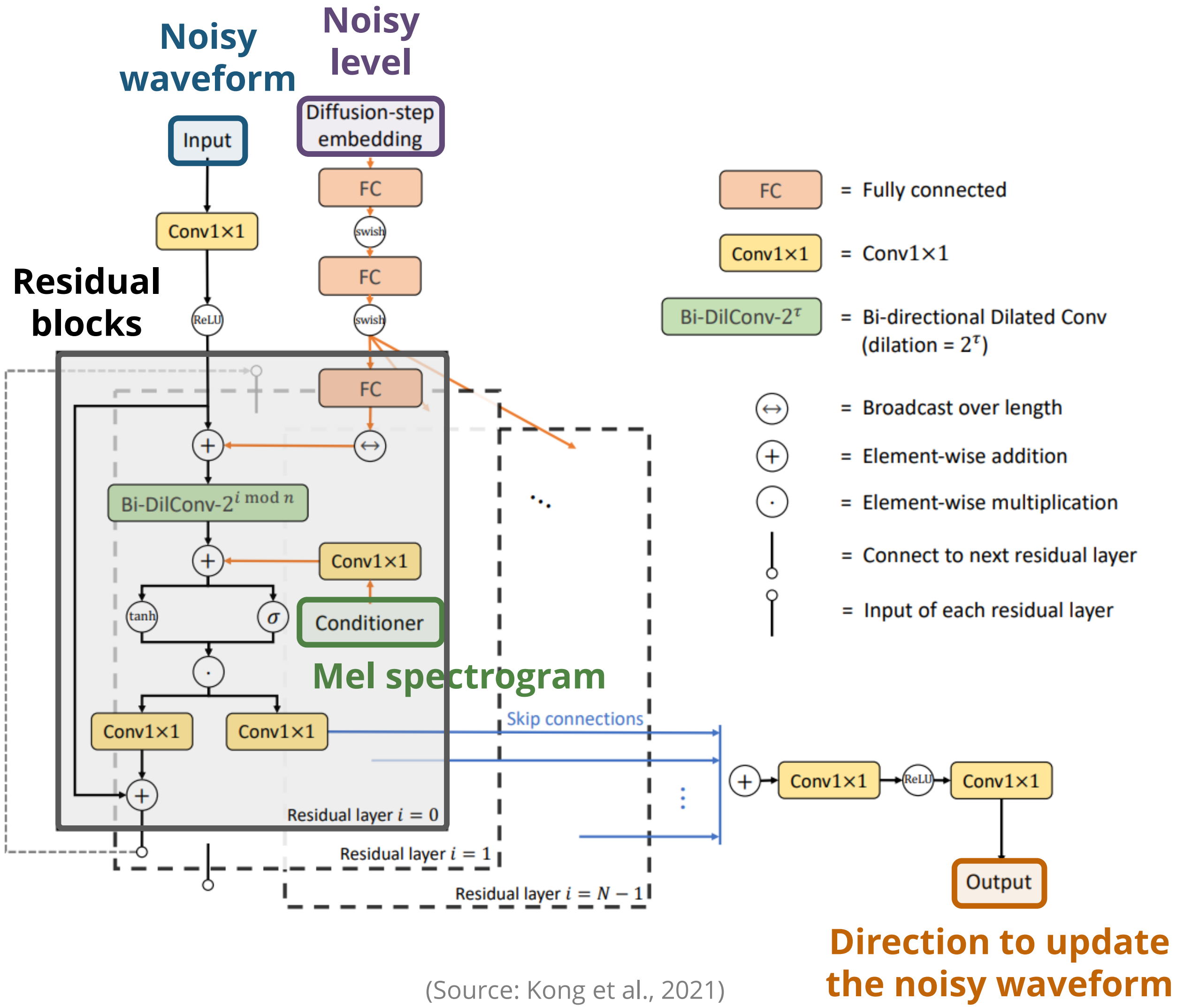
(Source: Chen et al., 2021)

Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.

DiffWave – Another Diffusion Model for Waveforms



(Source: Ho et al., 2020)

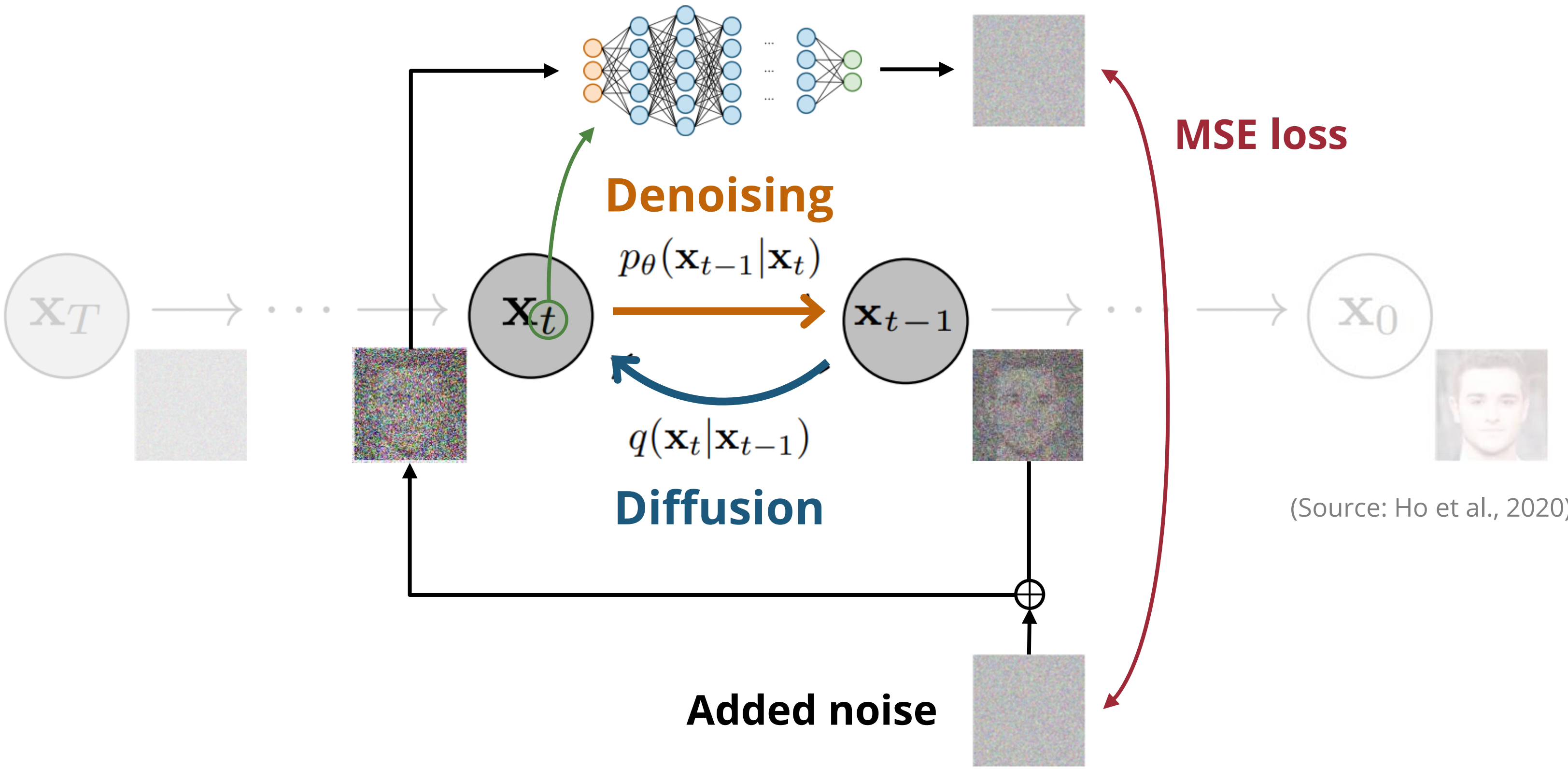


(Source: Kong et al., 2021)

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, Bryan Catanzaro, "DiffWave: A Versatile Diffusion Model for Audio Synthesis," *ICLR*, 2021.

(Recap) Diffusion Models – Training

- **Intuition**: Many denoising autoencoders stacked together

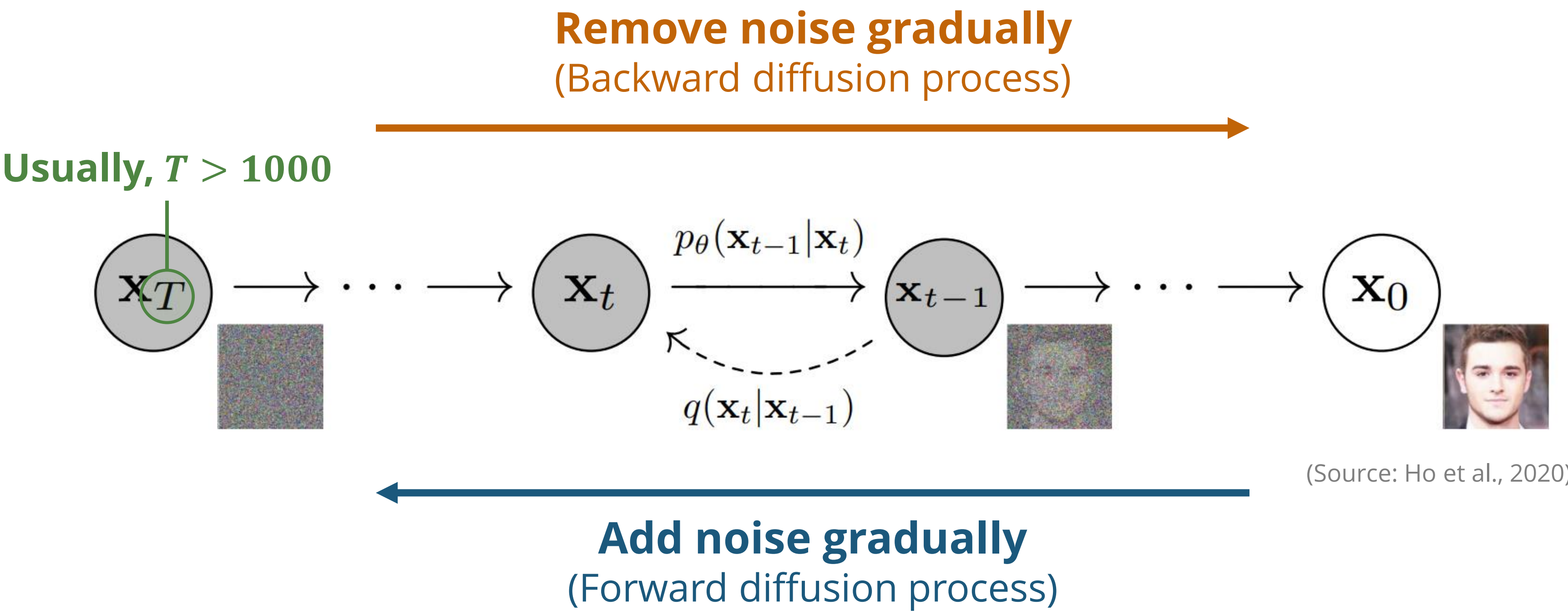


(Source: Ho et al., 2020)

Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.

(Recap) Diffusion Models

- **Intuition**: Many denoising autoencoders stacked together

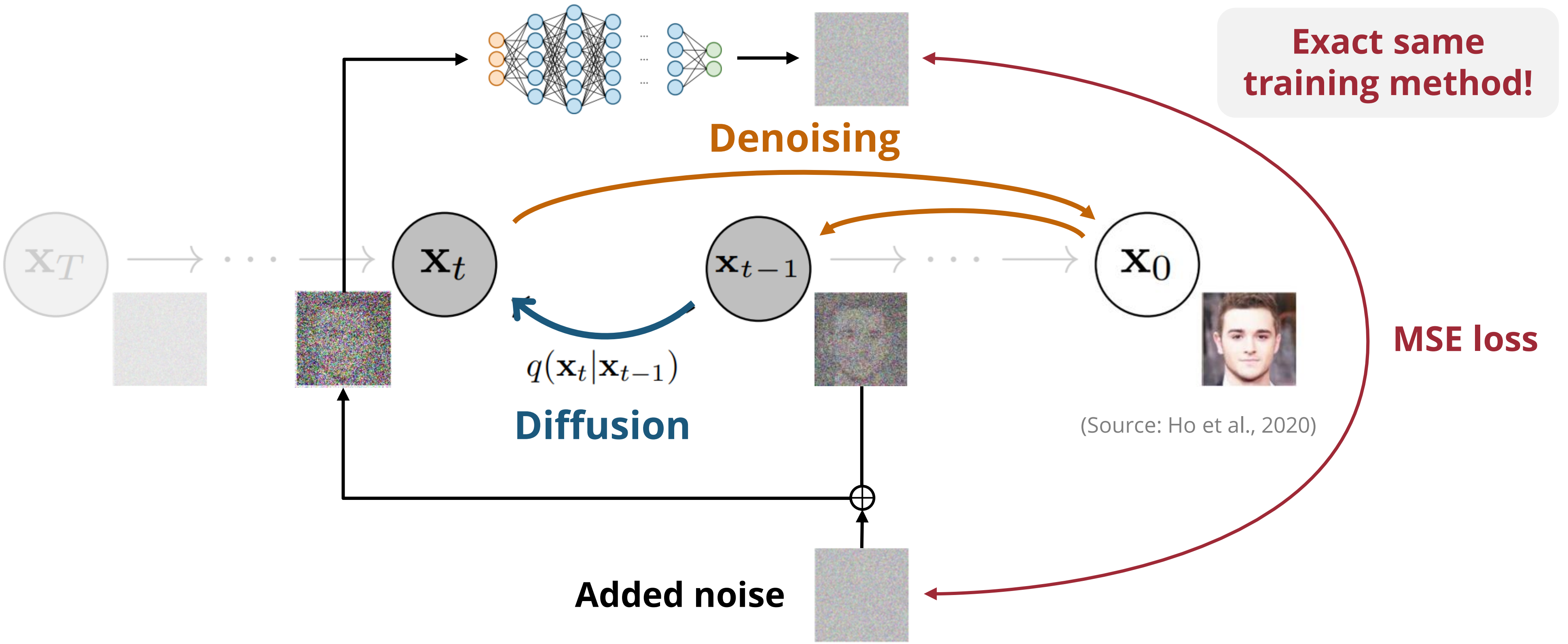


(Source: Ho et al., 2020)

Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.

Fast Sampling for Diffusion Models

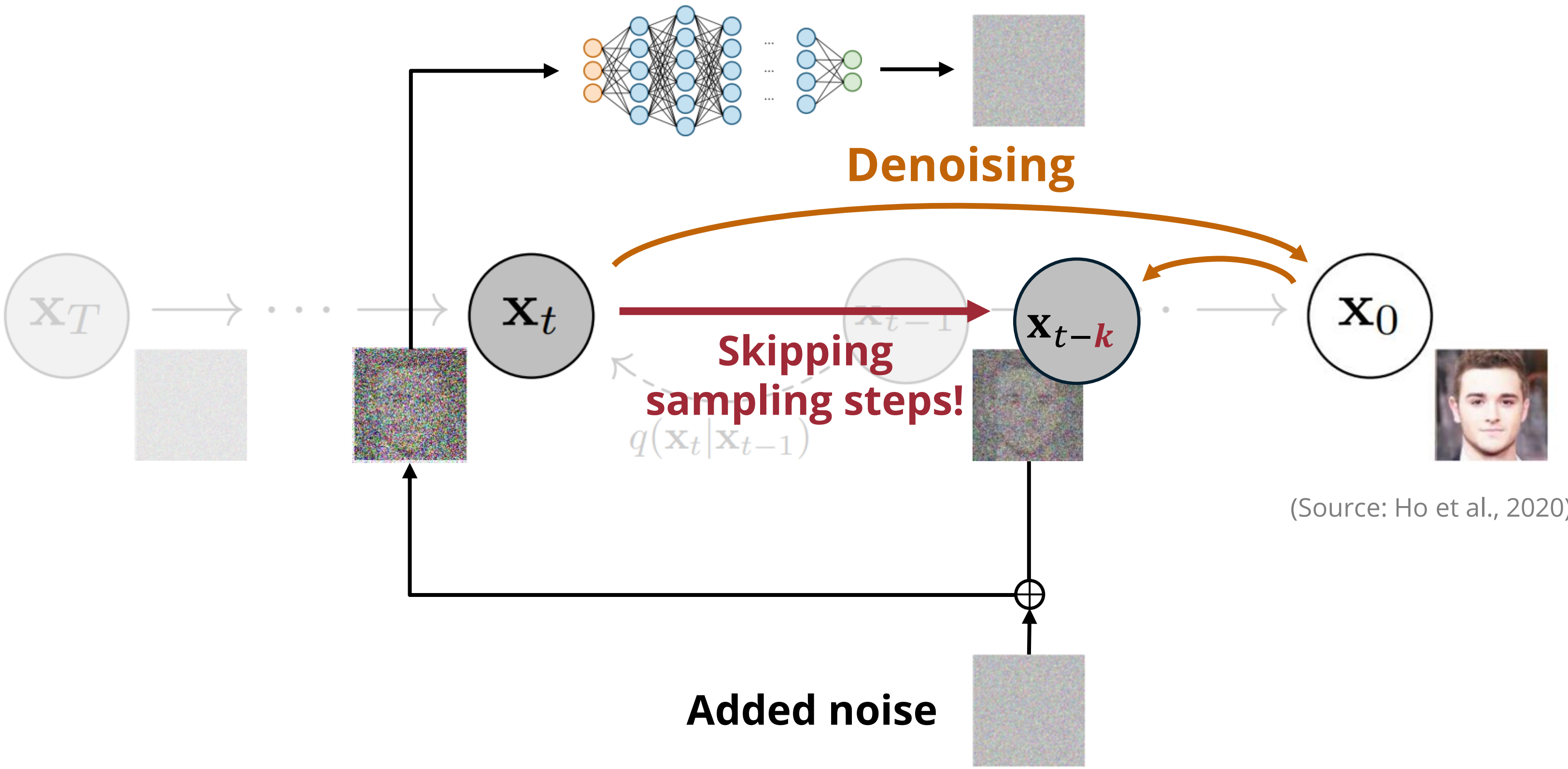
- **Intuition**: Skip some sampling steps



Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.
Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising Diffusion Implicit Models," *ICLR*, 2021.

Fast Sampling for Diffusion Models

- **Intuition**: Skip some sampling steps



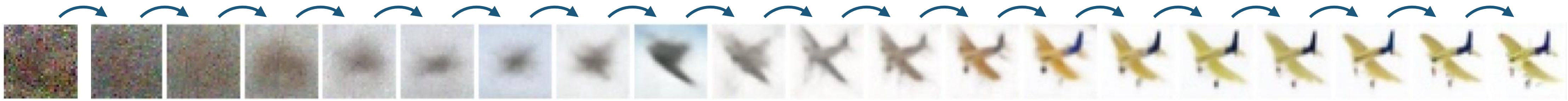
(Source: Ho et al., 2020)

Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.
Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising Diffusion Implicit Models," *ICLR*, 2021.

Fast Sampling for Diffusion Models

- **Intuition**: Skip some sampling steps

DDPM



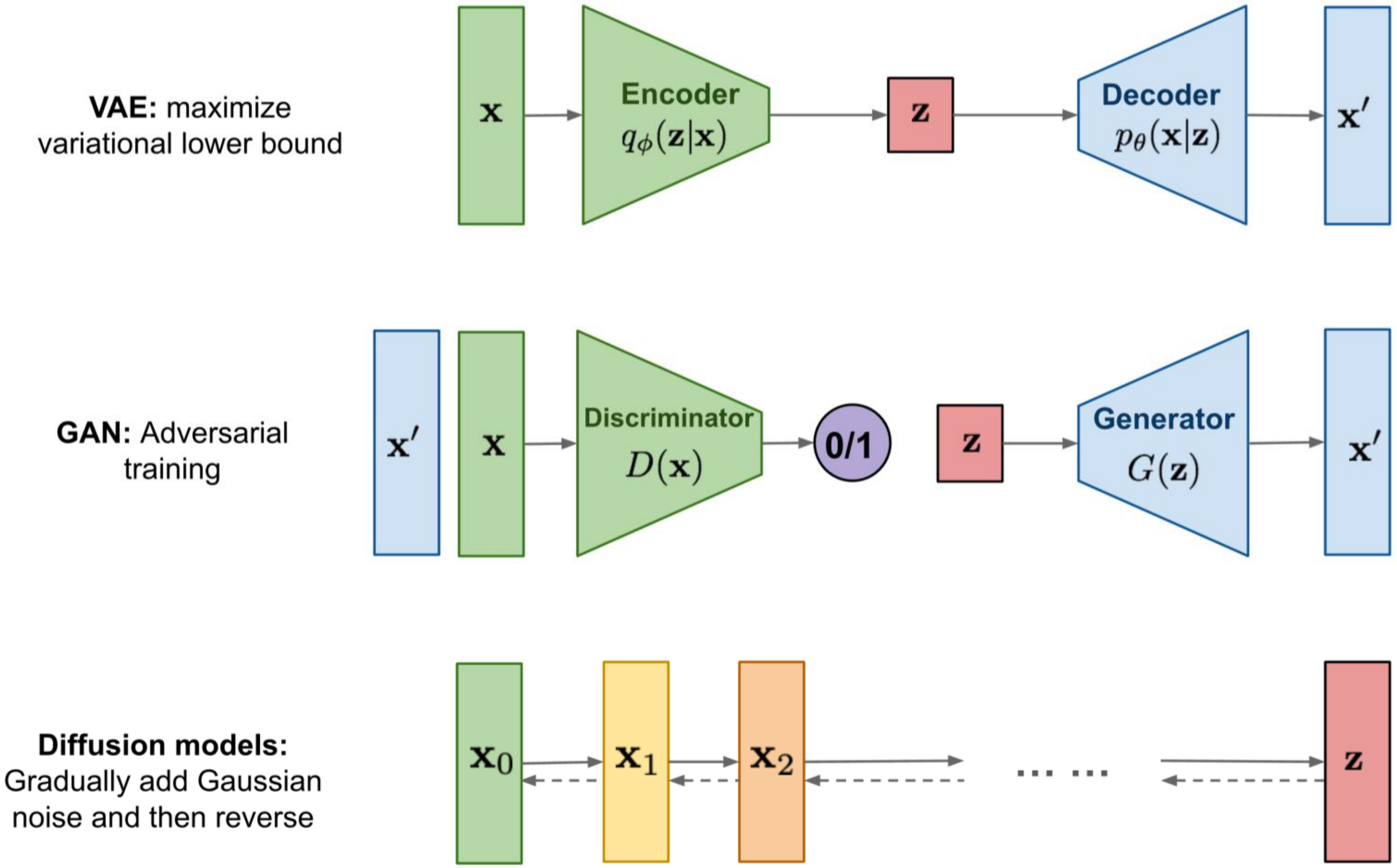
DDIM

(Source: Ho et al., 2020)

Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.
Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising Diffusion Implicit Models," *ICLR*, 2021.

Comparison of Deep Generative Models

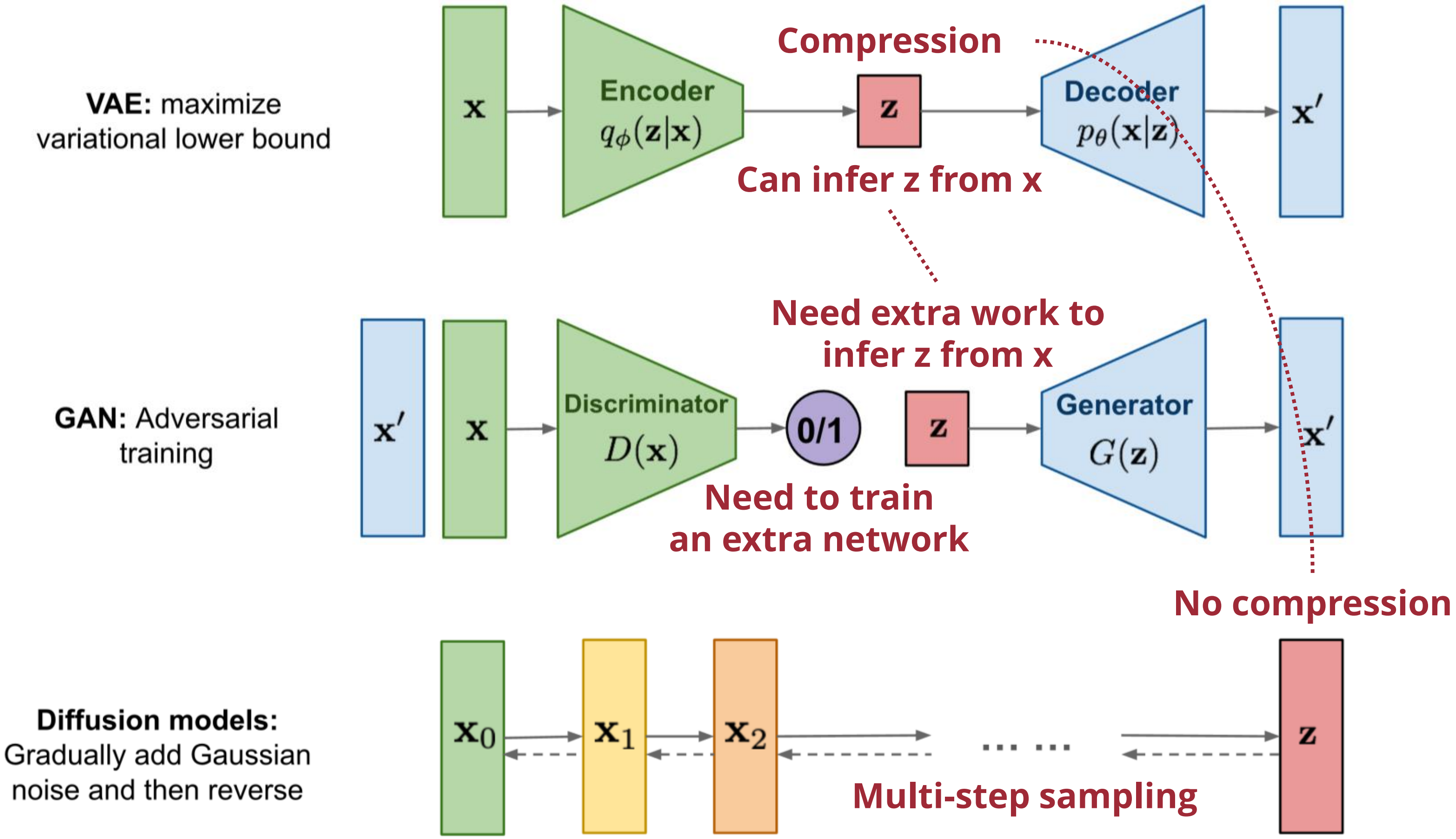
Comparison of Deep Generative Models



(Source: Weng, 2021)

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

Comparison of Deep Generative Models



(Source: Weng, 2021)

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

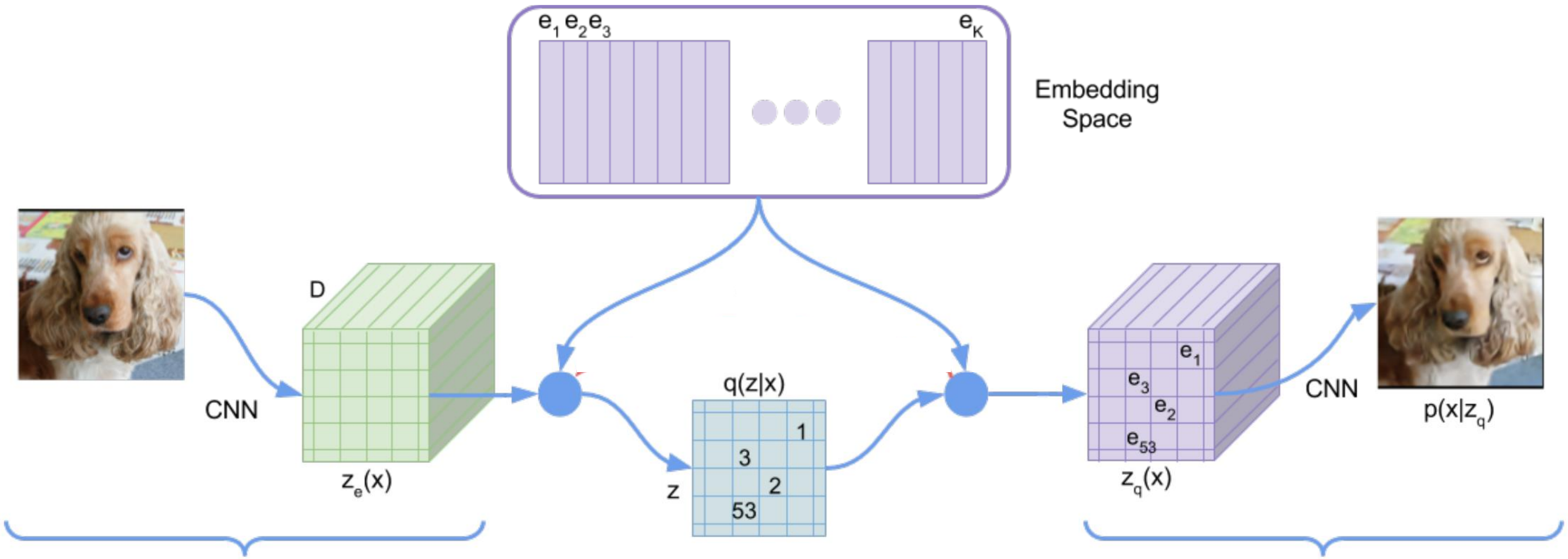
Network Architectures vs Training Frameworks

Network architectures	**Training frameworks**
Multilayer perceptron (MLP)	Autoregressive
Convolutional neural networks (CNNs)	Autoencoders
Recurrent neural networks (RNNs)	Variational autoencoders (VAEs)
Transformers	Generative adversarial networks (GANs)
ResNets	Diffusion models
U-Nets	Consistency models
⋮	⋮

Latent Diffusion Models

(Recap) Vector-Quantized VAEs (VQVAEs)

Vector quantization



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

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

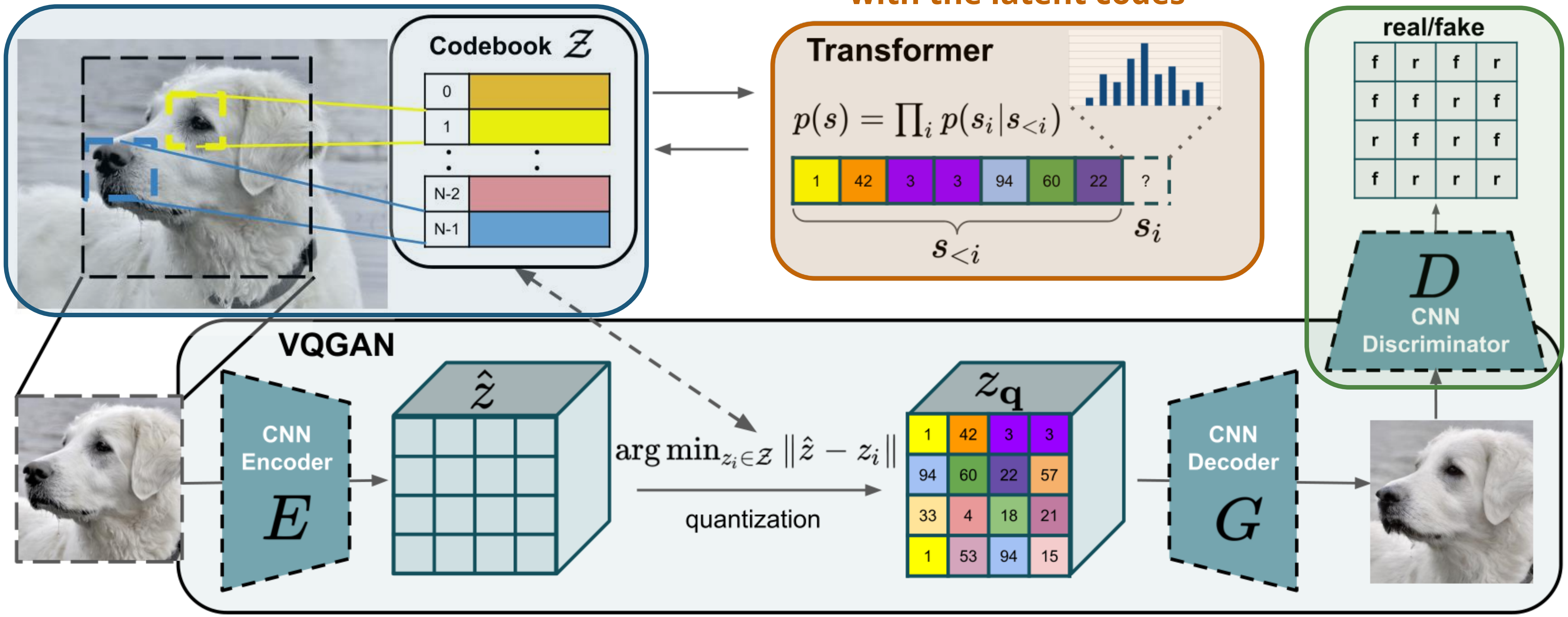
Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu, "Neural Discrete Representation Learning," *NeurIPS*, 2017.

VQGAN

Each path is encoded into a latent code

A transformer-based language model trained with the latent codes

Patch discriminator



(Source: Esser et al., 2021)

A VQGAN is a VQVAE equipped with adversarial loss

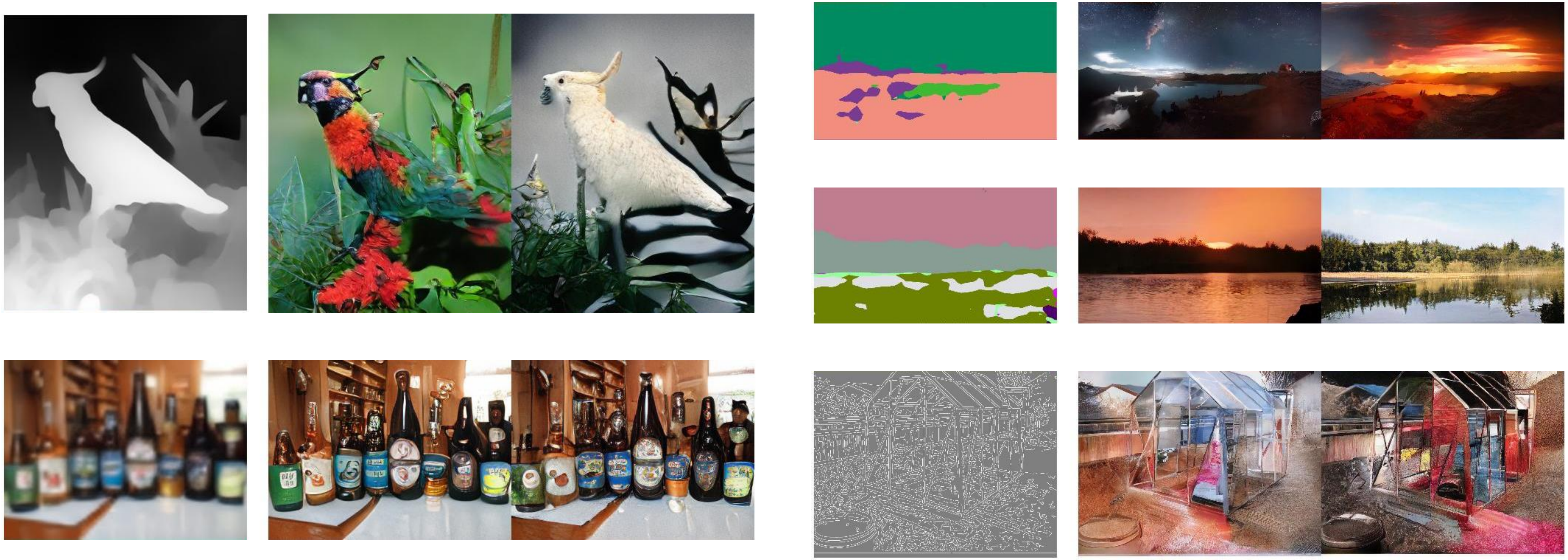
Patrick Esser, Robin Rombach, and Björn Ommer, "Taming Transformers for High-Resolution Image Synthesis," *CVPR*, 2021.

VQGAN – Conditional Generation

(Source: Esser et al., 2021)

Patrick Esser, Robin Rombach, and Björn Ommer, “Taming Transformers for High-Resolution Image Synthesis,” *CVPR*, 2021.

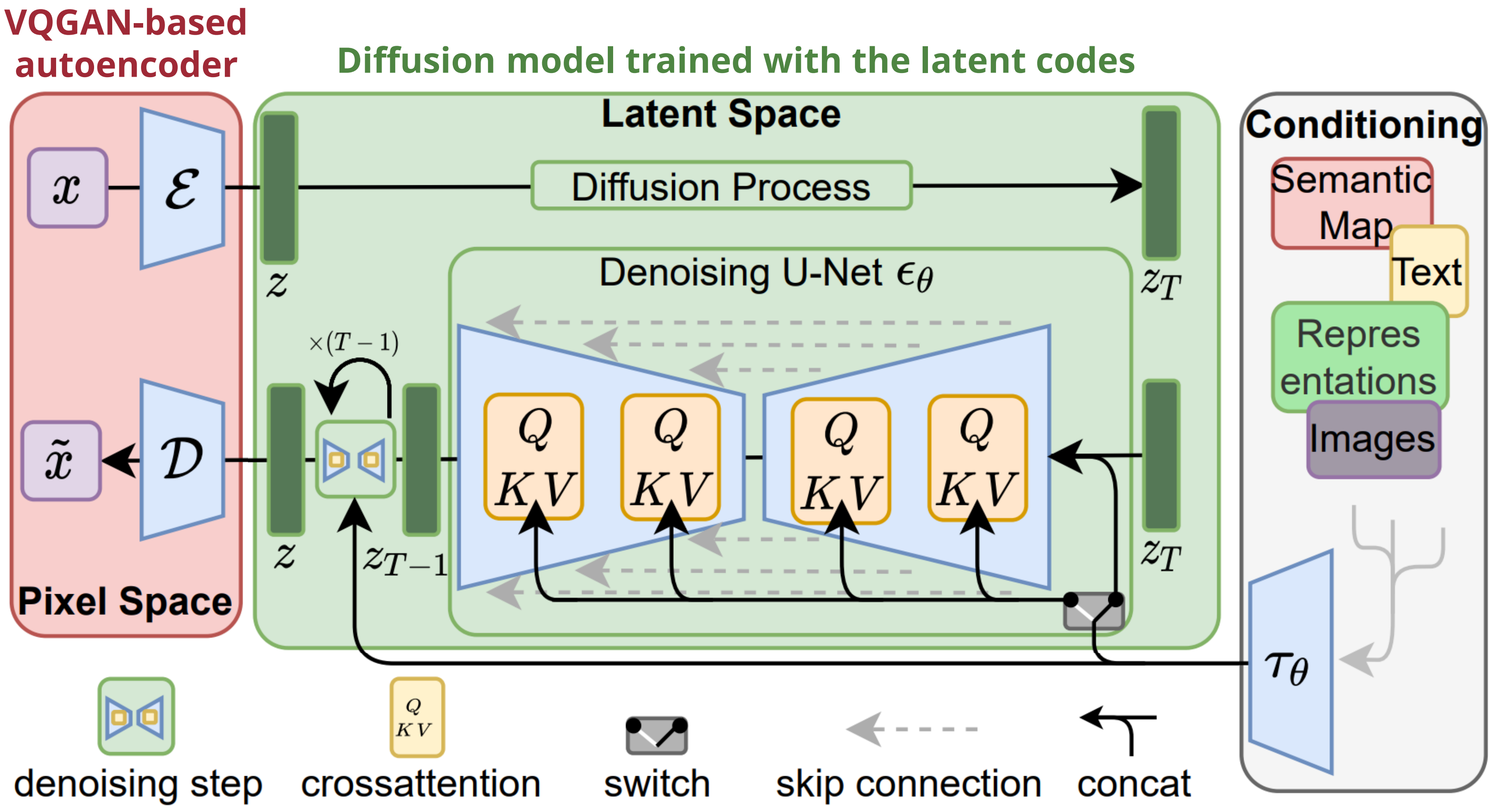
VQGAN – Conditional Generation



(Source: Esser et al., 2021)

Patrick Esser, Robin Rombach, and Björn Ommer, "Taming Transformers for High-Resolution Image Synthesis," *CVPR*, 2021.

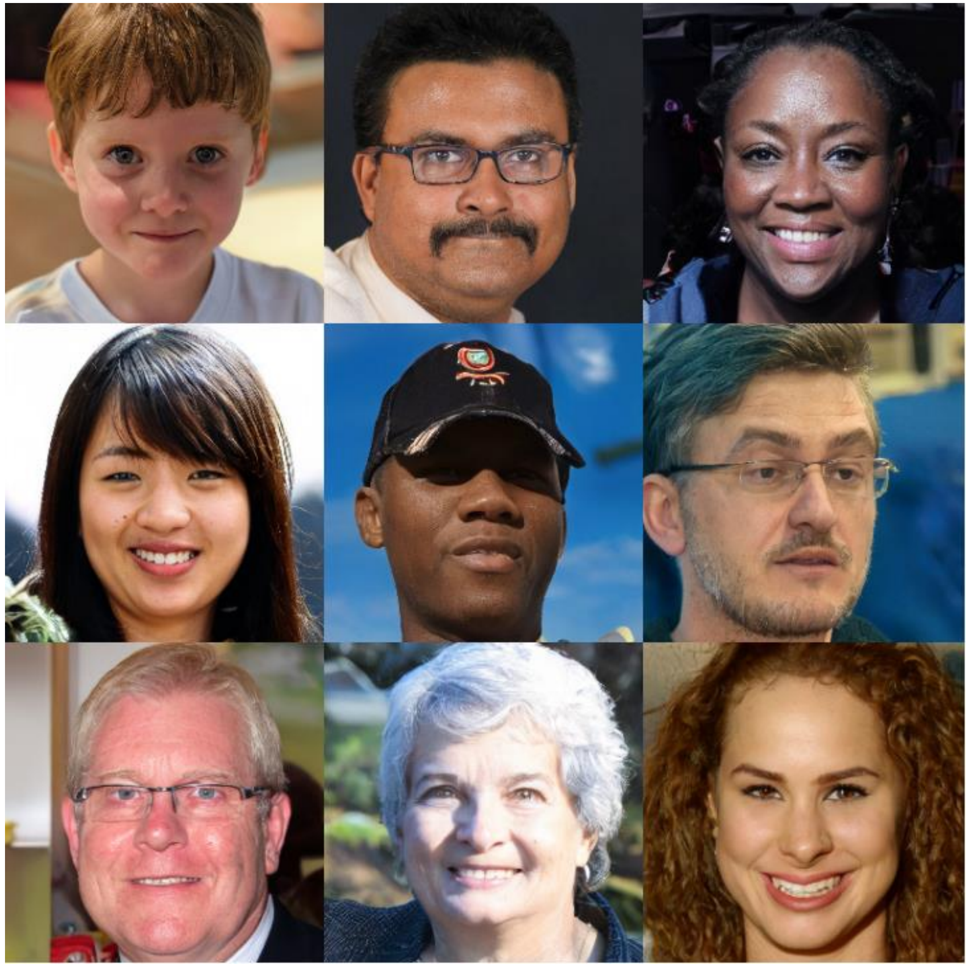
Latent Diffusion Models (LDMs)



(Source: Rombach et al., 2022)

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *CVPR*, 2022.

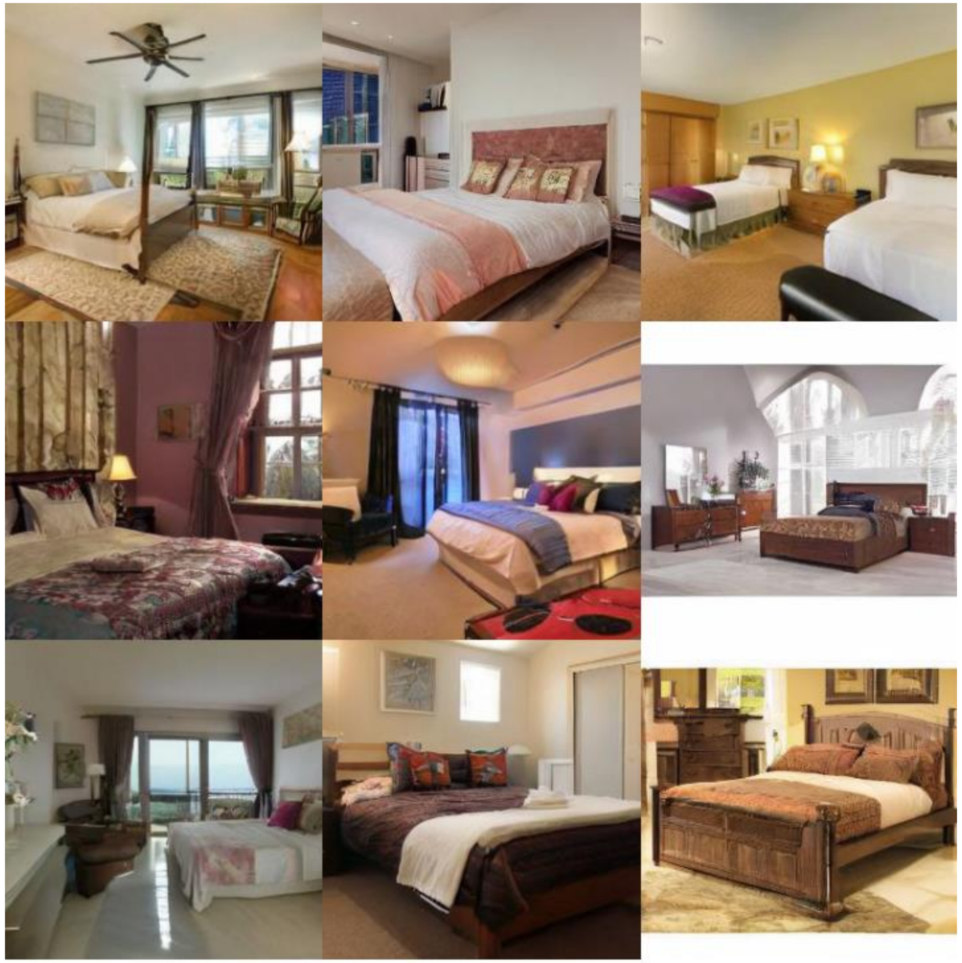
LDMs – Examples



(Source: Rombach et al., 2022)

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *CVPR*, 2022.

LDMs – Examples

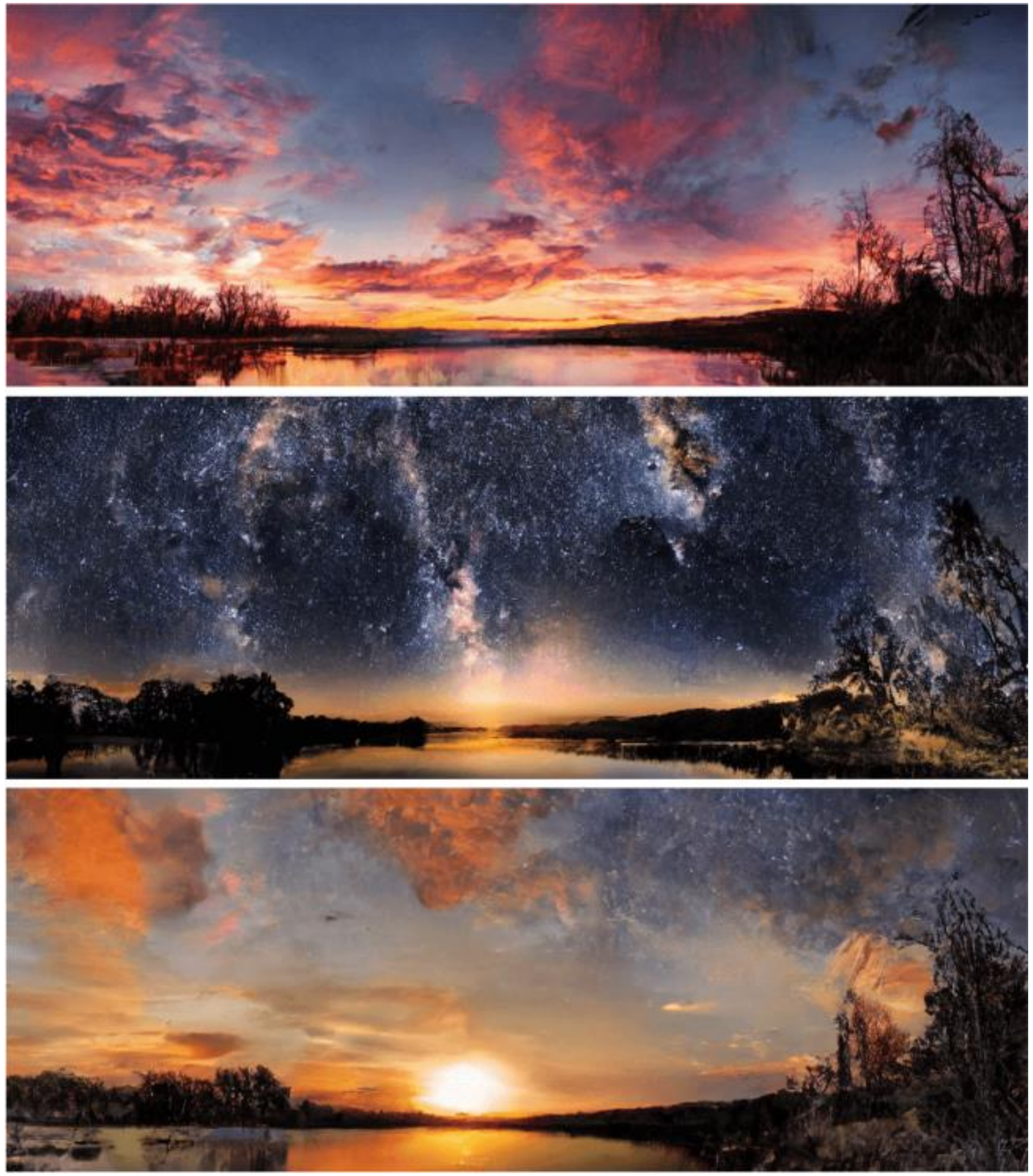


(Source: Rombach et al., 2022)

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” *CVPR*, 2022.

LDMs – Semantic Synthesis

(Source: Rombach et al., 2022)

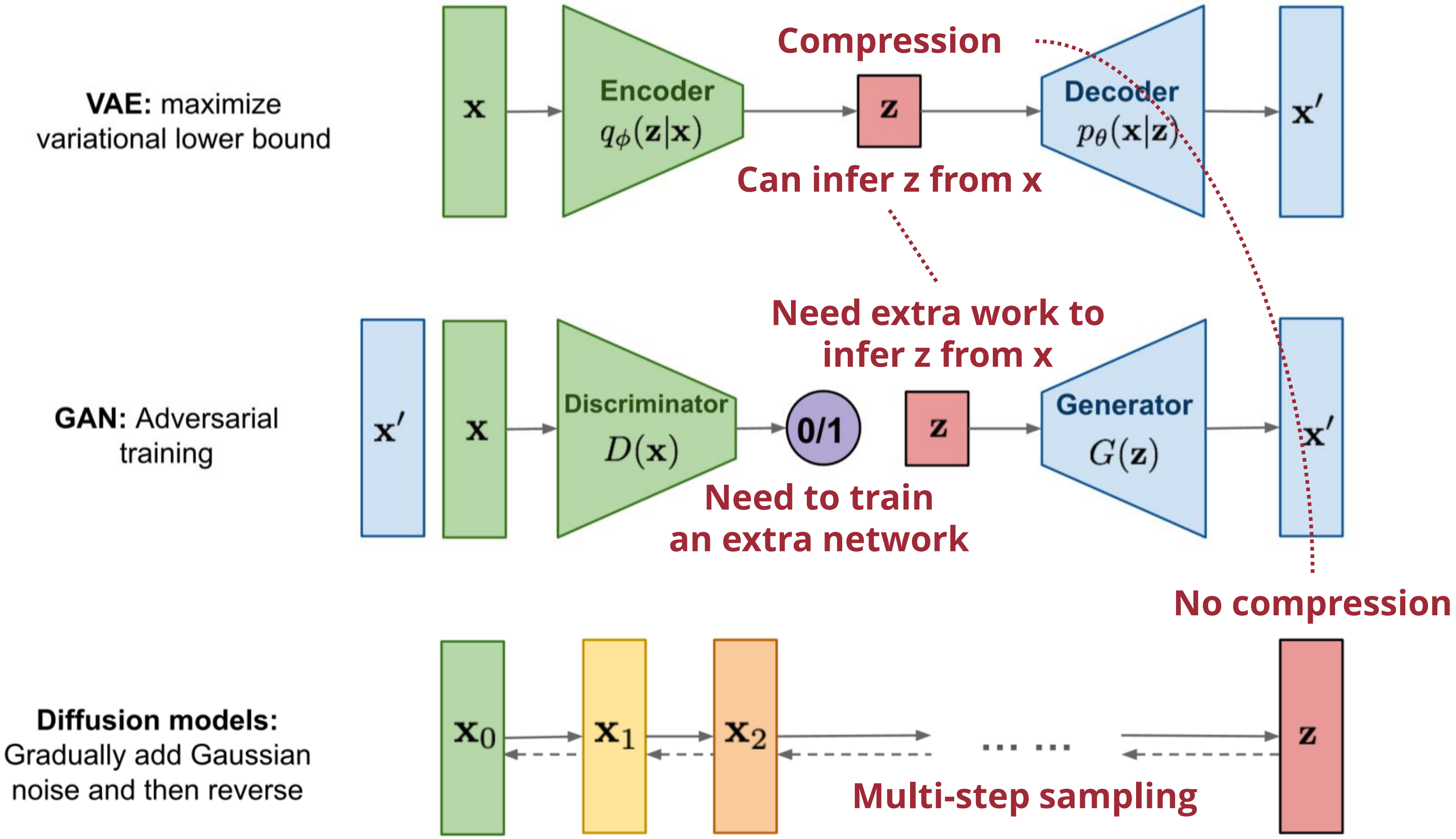


Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” *CVPR*, 2022.

(Recap) Network Architectures vs Training Frameworks

Network architectures	**Training frameworks**
Multilayer perceptron (MLP)	Autoregressive
Convolutional neural networks (CNNs)	Autoencoders
Recurrent neural networks (RNNs)	Variational autoencoders (VAEs)
Transformers	Generative adversarial networks (GANs)
ResNets	Diffusion models
U-Nets	Consistency models
⋮	⋮

(Recap) Comparison of Deep Generative Models



(Source: Weng, 2021)

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