PAT 464/564 (Winter 2026)

Generative AI for Music & Audio Creation

Lecture 14: Diffusion Models

Instructor: Hao-Wen Dong

UNIVERSITY OF MICHIGAN

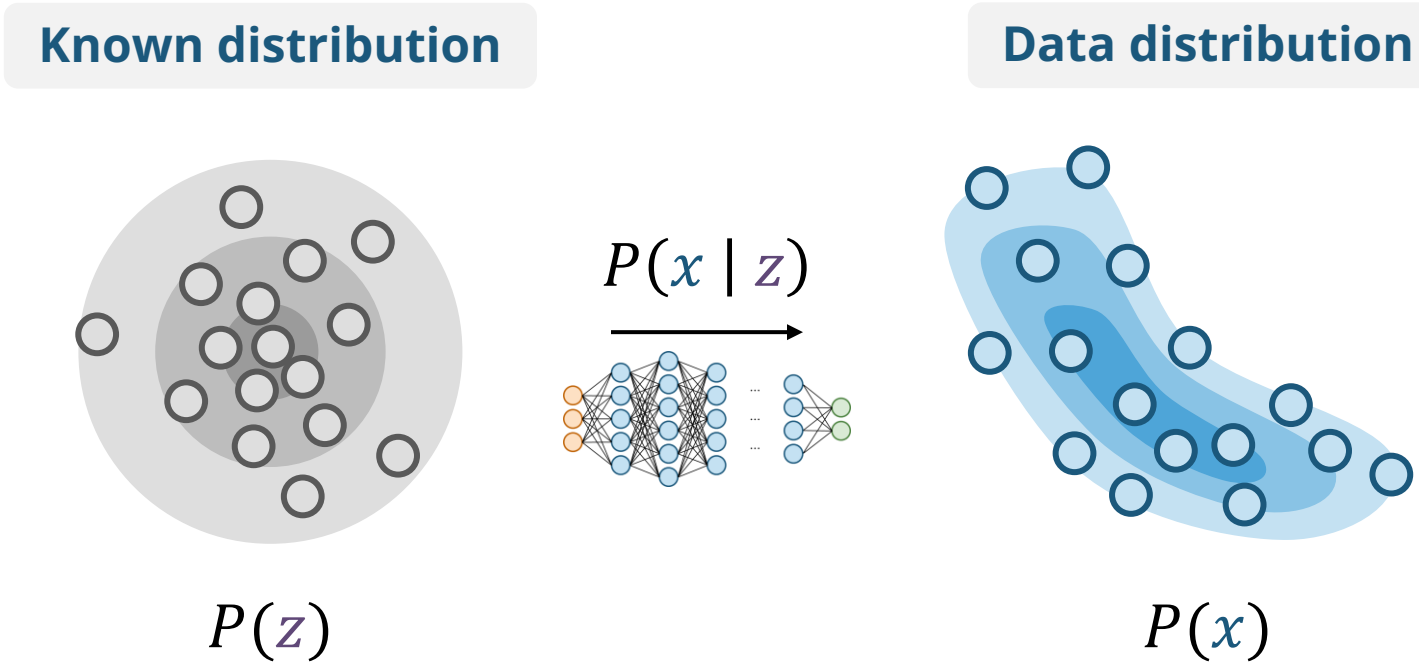
Representative Types of Deep Generative Models

- **Deep autoregressive models**
  - Recurrent neural network (RNN)
  - Long short-term memory (LSTM)
  - Transformer model
- **Deep latent variable models**
  - Variational autoencoder (VAE)
  - Generative adversarial network (GAN)
  - Diffusion model **Today's topic!**
  - Flow-based model
- *And many others…*

Deep Latent Variable Models

Deep Latent Variable Models

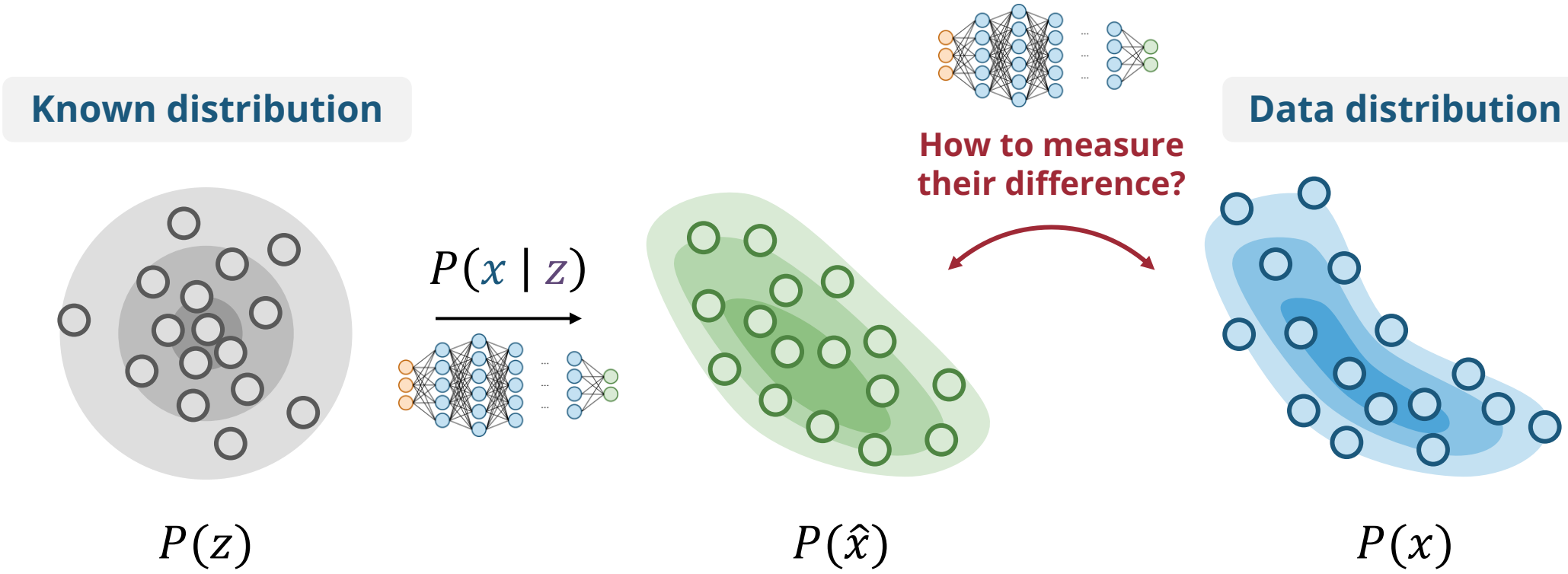
- **Intuition**: Learn to map a known distribution to the data distribution



$$P(x) = P(z)\ P(x \mid z)$$

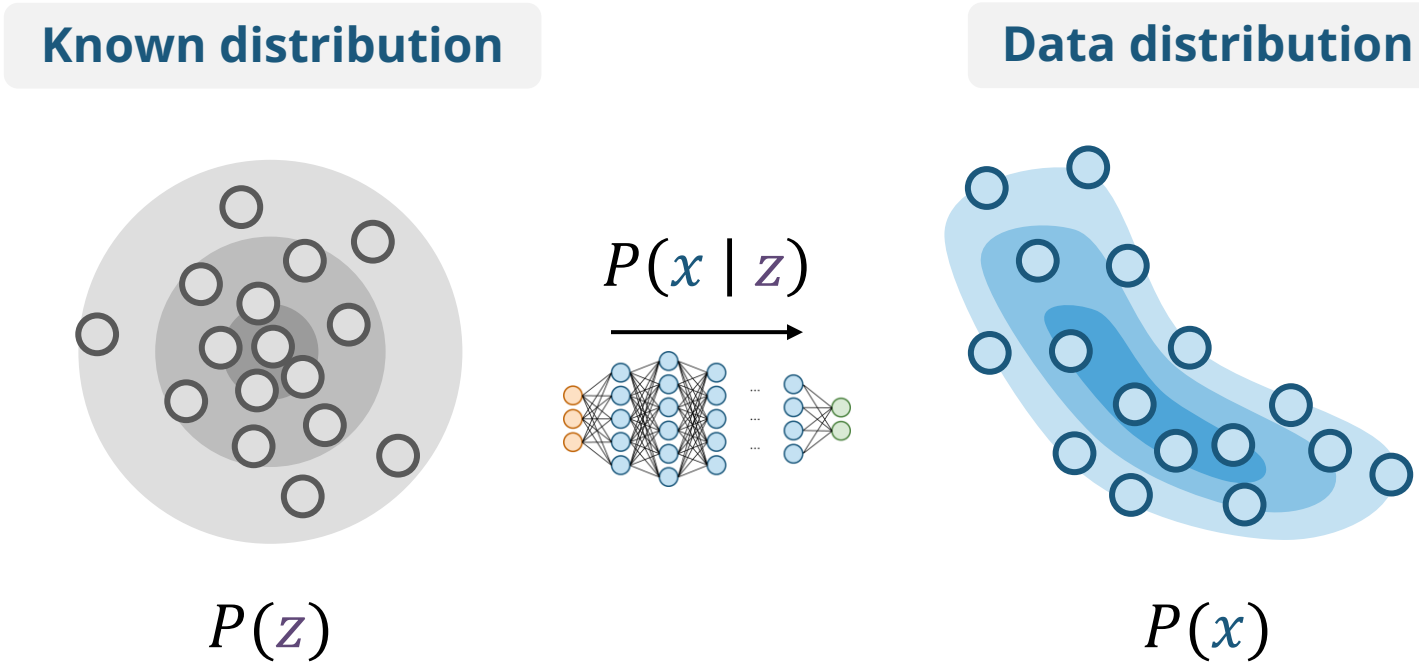
Deep Latent Variable Models

- **Intuition**: Learn to map a known distribution to the data distribution



Deep Latent Variable Models

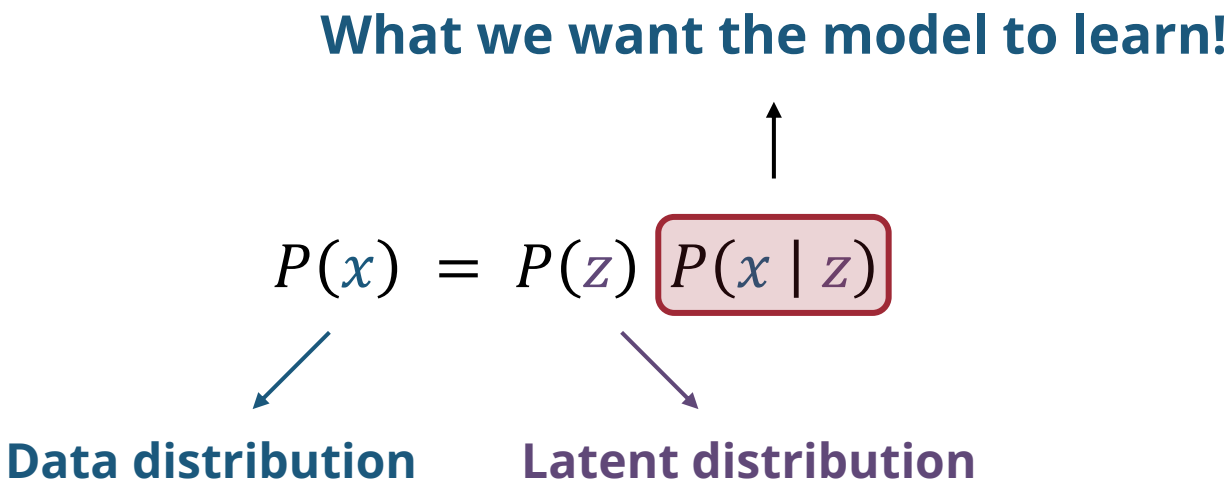
- **Intuition**: Learn to map a known distribution to the data distribution



$$P(x) = P(z)\ P(x \mid z)$$

Deep Latent Variable Models

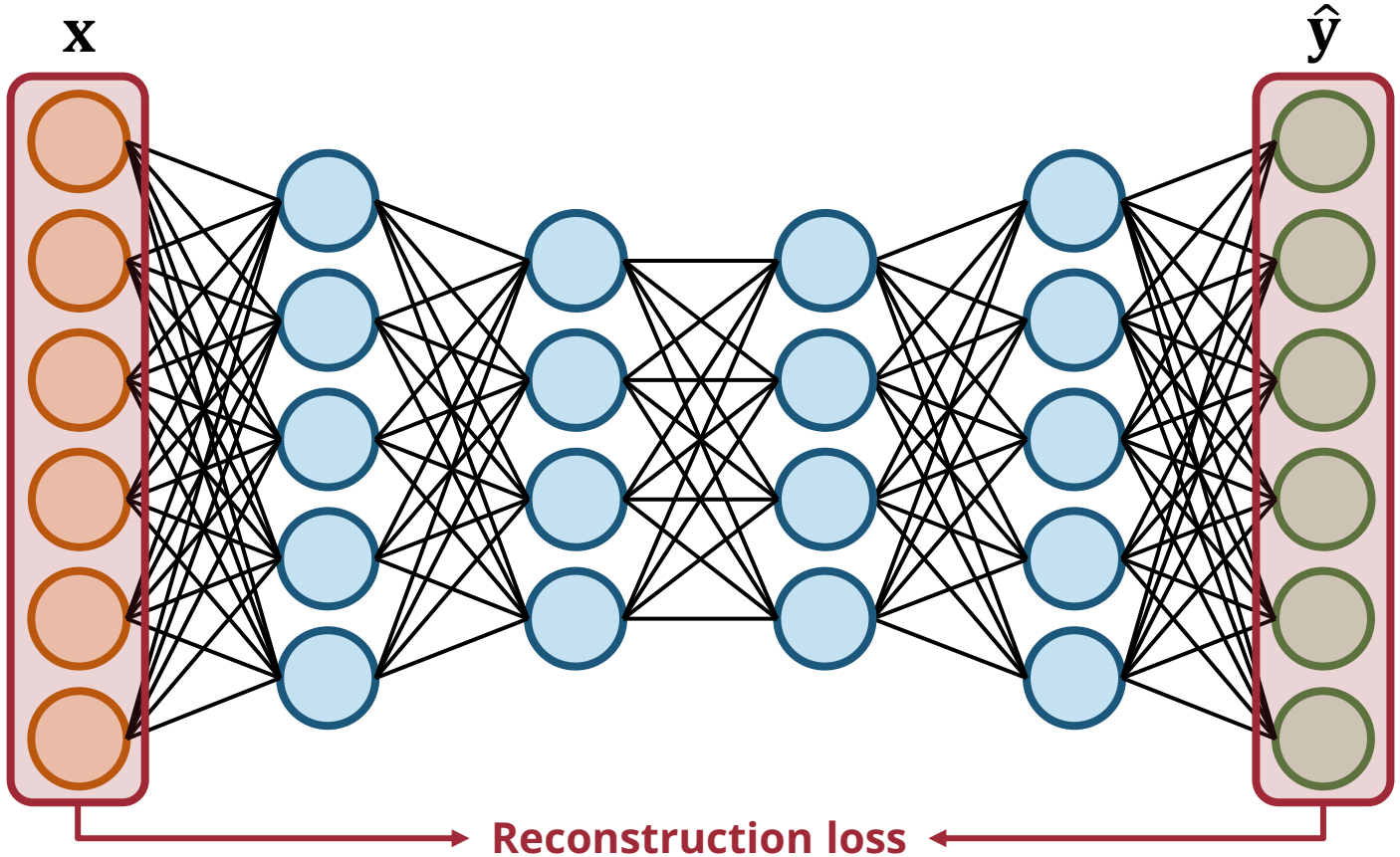
- **Intuition**: Learn to map a known distribution to the data distribution



Diffusion Models

Autoencoders

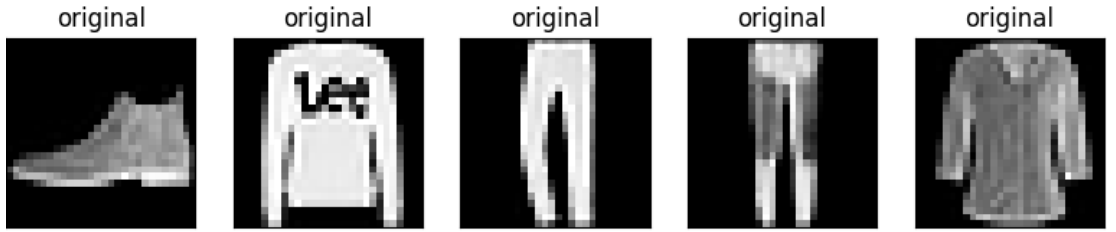
- A neural network where the **input and output are the same**



Autoencoders: Reconstruction Examples

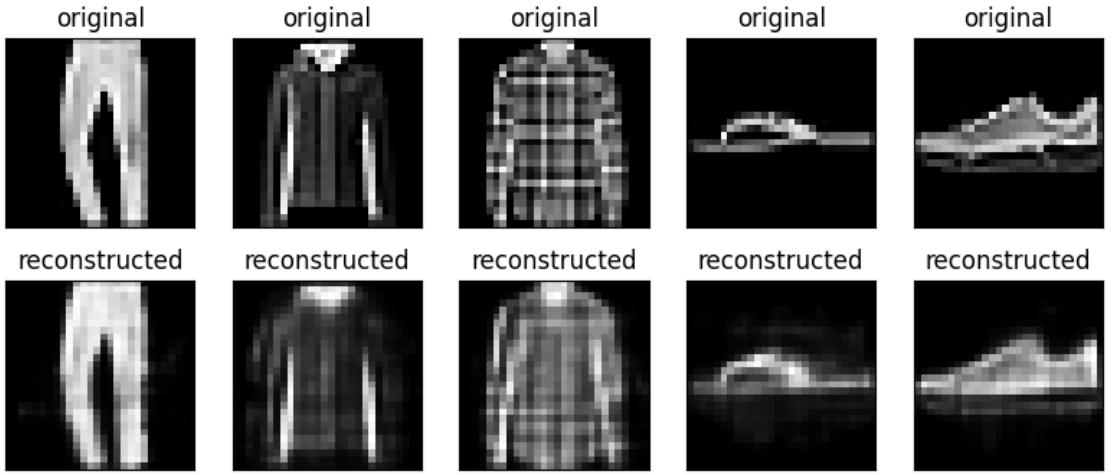
Original

Reconstructed



Original

Reconstructed



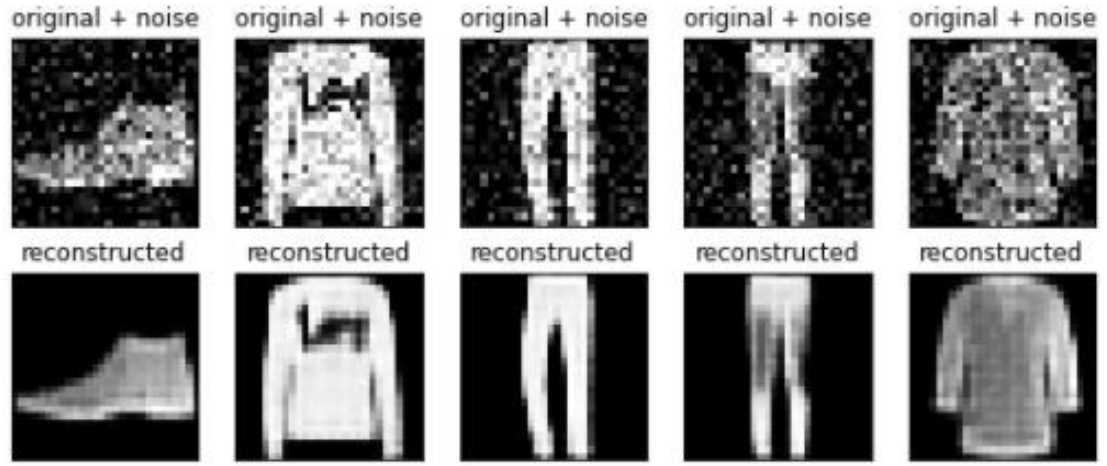
(Source: tensorflow.org)

tensorflow.org/tutorials/generative/autoencoder

Denoising Autoencoders

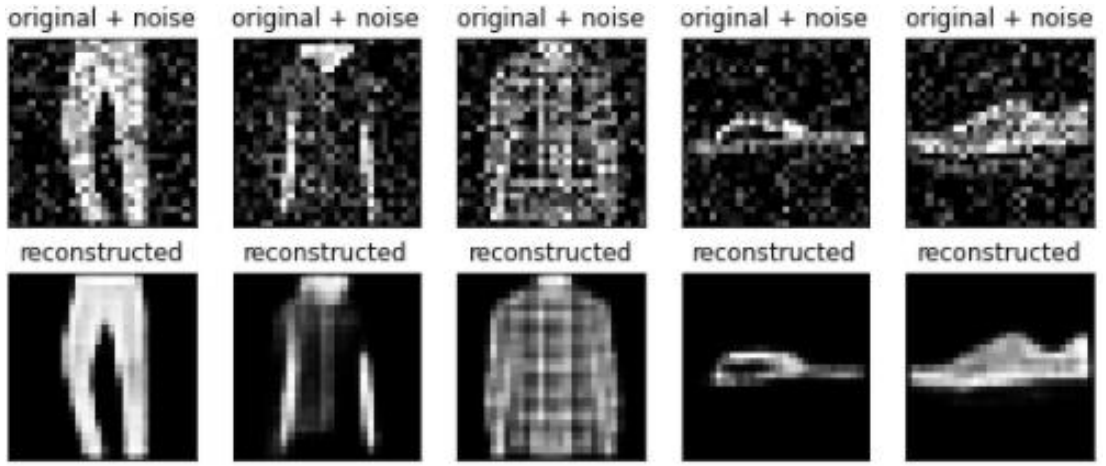
Noisy

Reconstructed



Noisy

Reconstructed



Denoising autoencoders learn to reconstruct noisy inputs

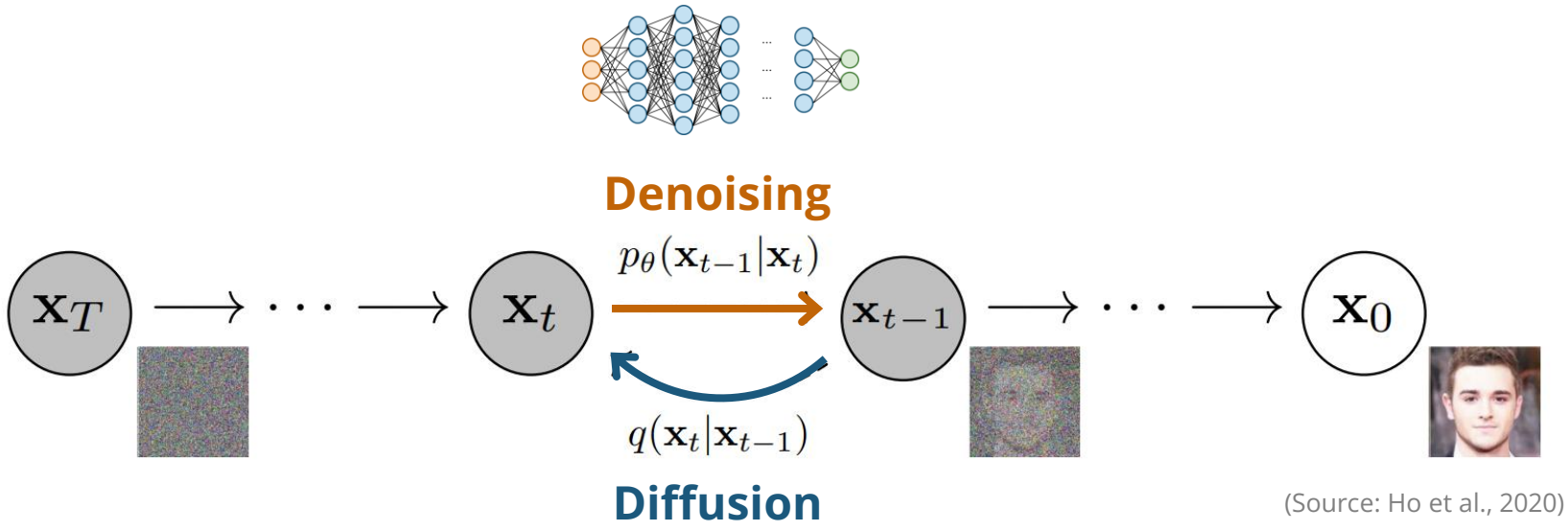
(Source: tensorflow.org)

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.

Diffusion Models (Ho et al., 2020)

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

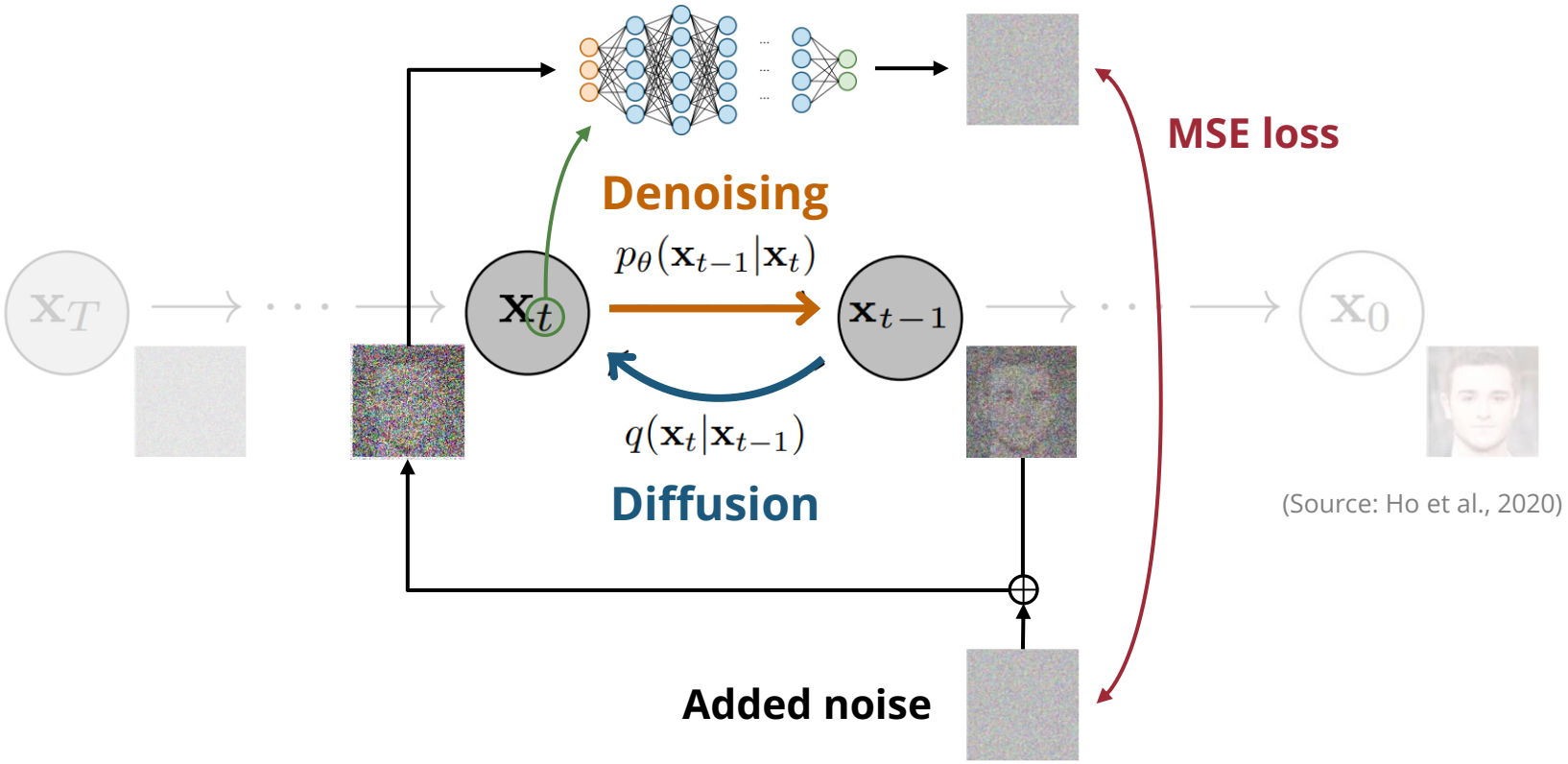


(Source: Ho et al., 2020)

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

Diffusion Models: Training

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

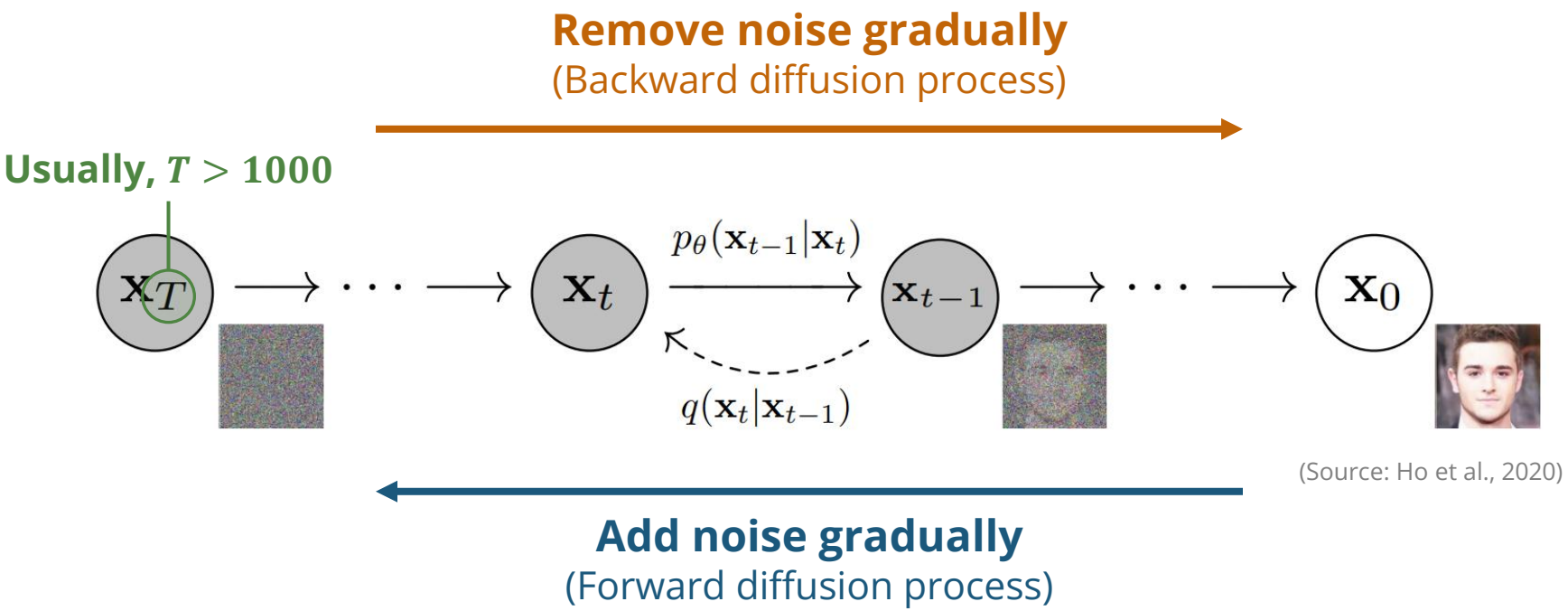


(Source: Ho et al., 2020)

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

Diffusion Models (Ho et al., 2020)

- **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: Generation

Remove noise gradually
(Backward diffusion process)

Input → **Output**

Coarse shapes
(low-frequency components)

Fine details
(high-frequency components)

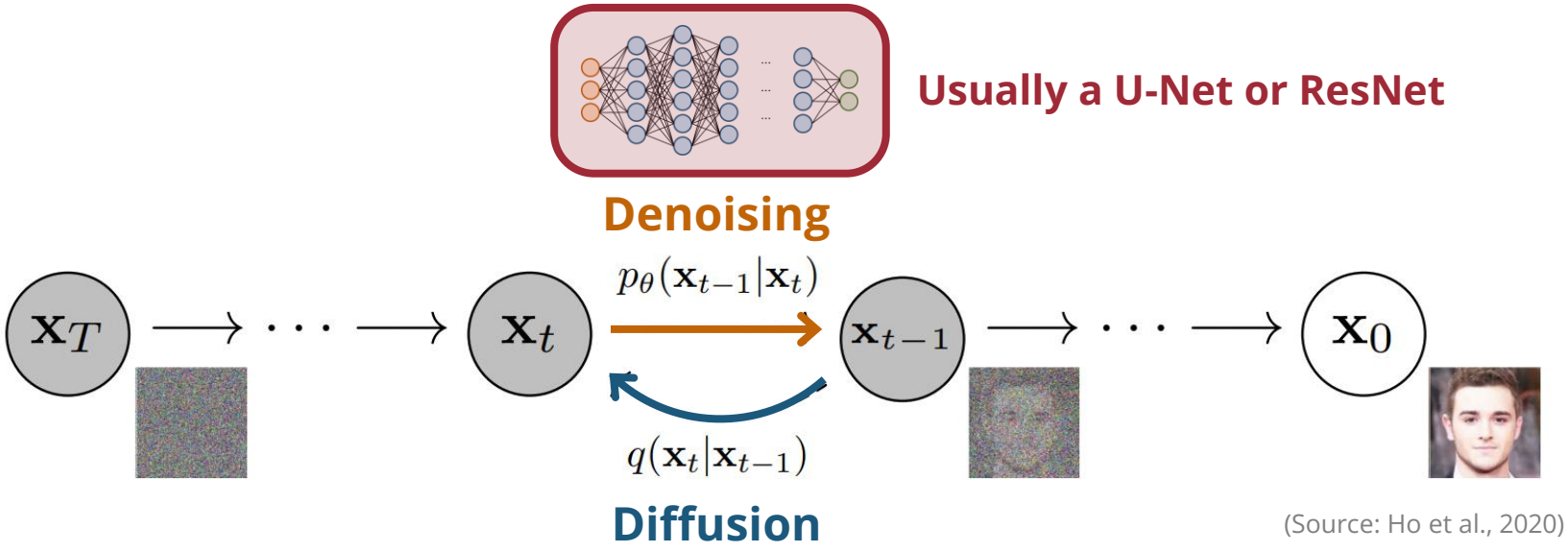
(Source: Ho et al., 2020)

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

U-Nets & ResNets

Diffusion Models (Ho et al., 2020)

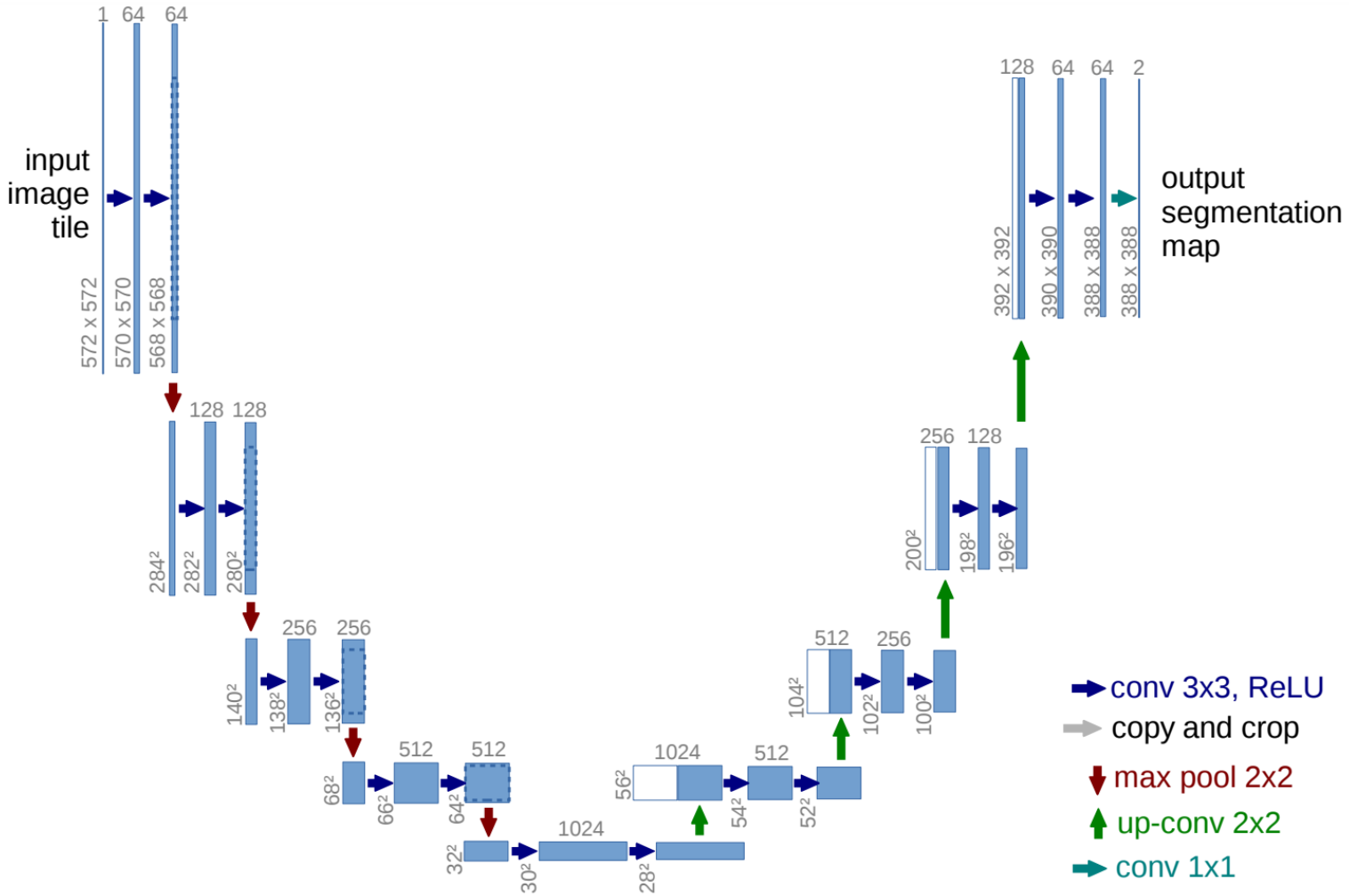
- **Intuition**: Many denoising autoencoders stacked together



(Source: Ho et al., 2020)

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

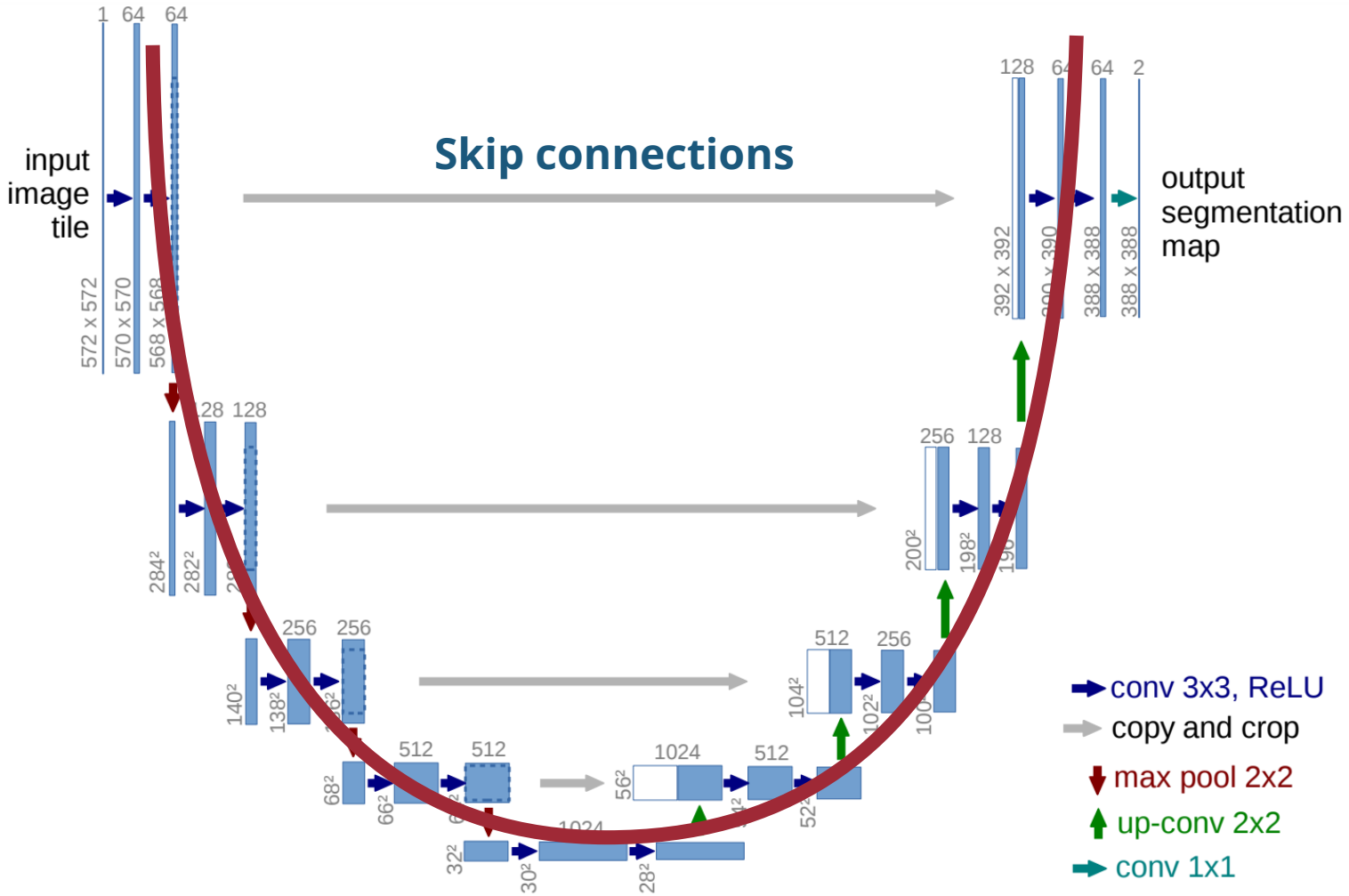
U-Net (Ronneberger et al., 2015)



(Source: Ronneberger et al., 2015)

Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *MICCAI*, 2015.

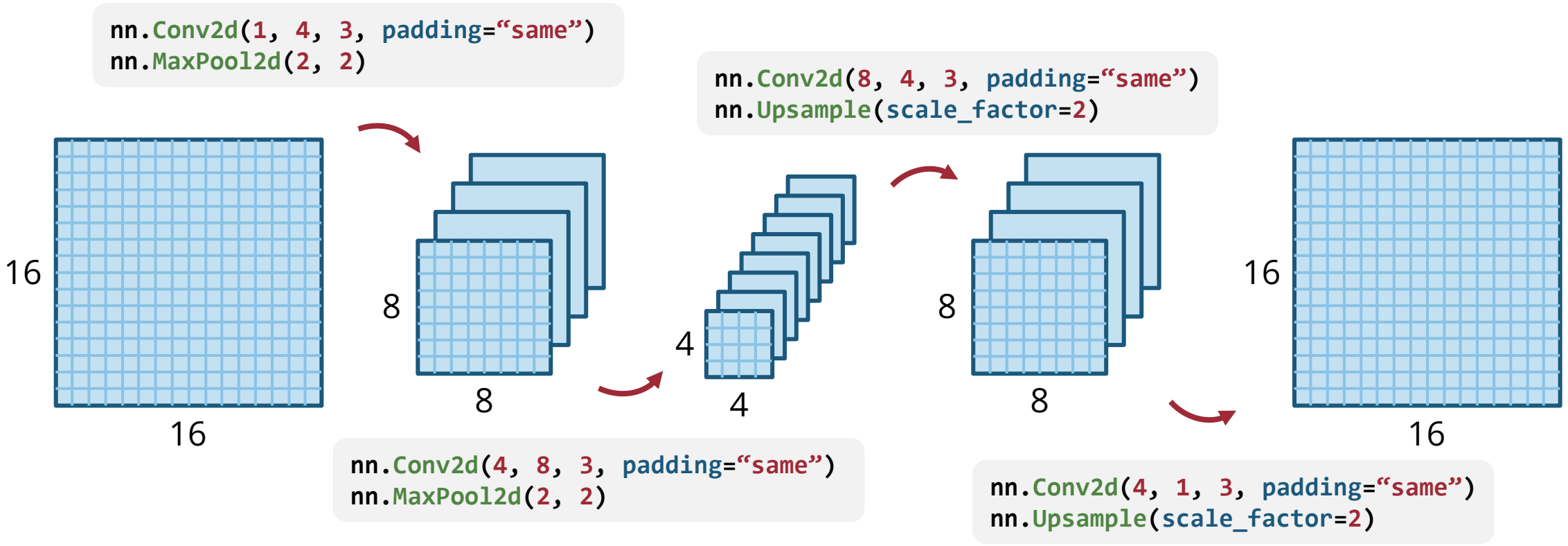
U-Net (Ronneberger et al., 2015)



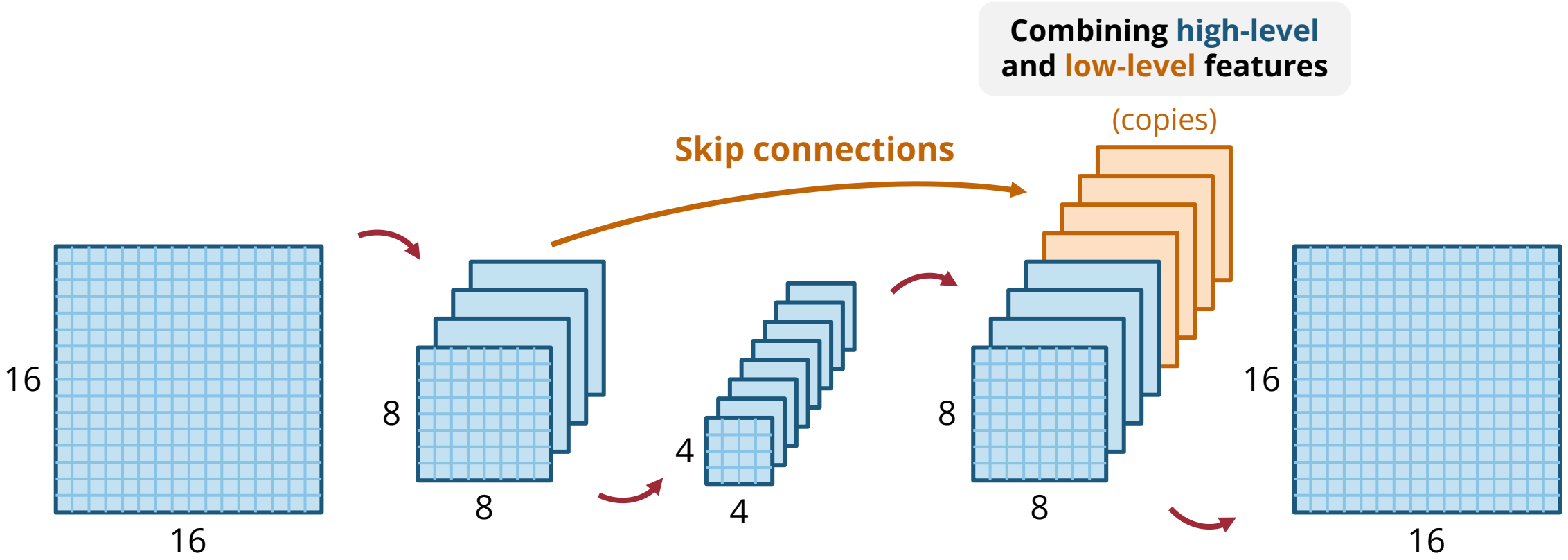
(Source: Ronneberger et al., 2015)

Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *MICCAI*, 2015.

A Toy Example of U-Net



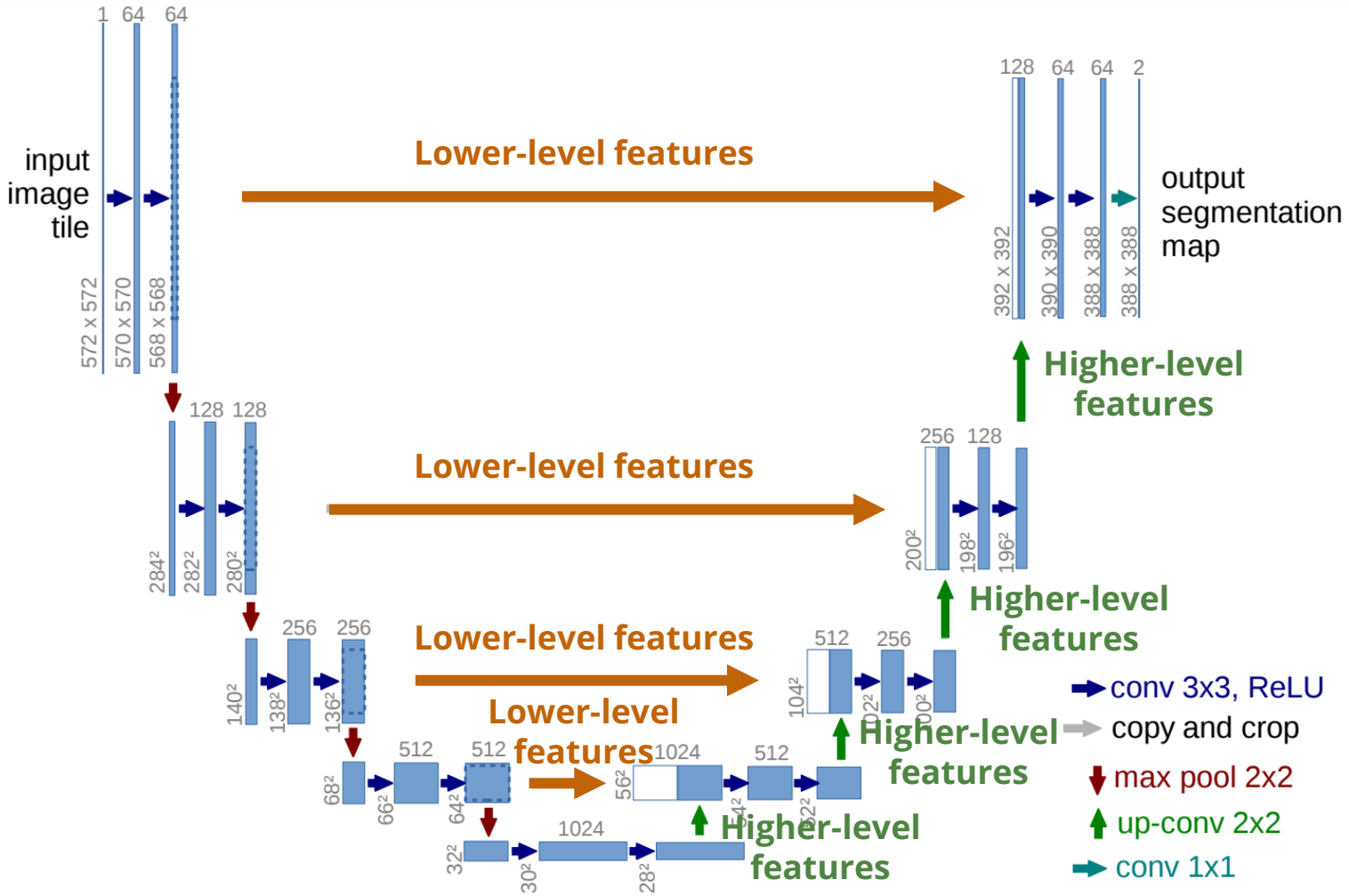
A Toy Example of U-Net




```
nn.Conv2d(8, 1, 3, padding="same")
nn.Upsample(scale_factor=2)
```

U-Nets are useful when the inputs and outputs have the same shape!

U-Net (Ronneberger et al., 2015)

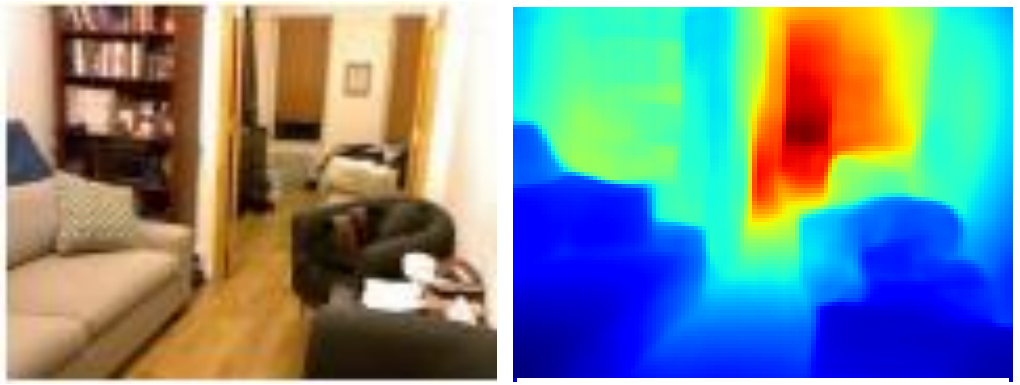


(Source: Ronneberger et al., 2015)

Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *MICCAI*, 2015.

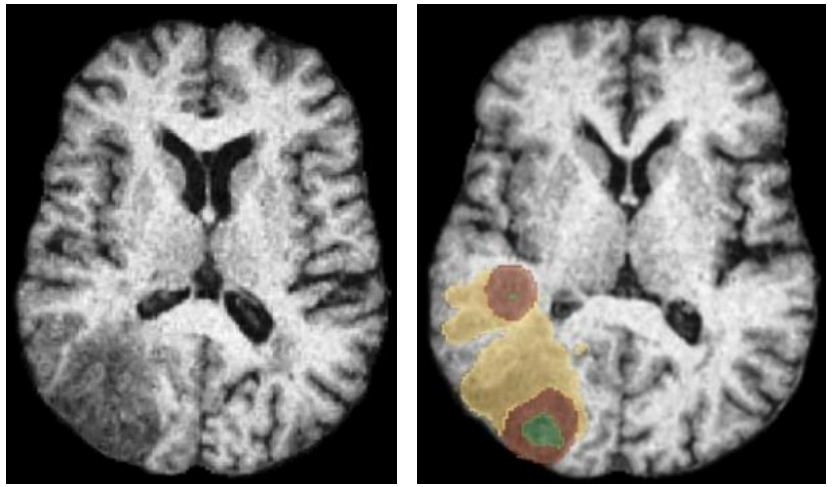
Applications of U-Nets

Depth Estimation



(Source: Barakat, 2018)

Tumor Segmentation



(Source: Kharaji et al., 2024)

Image Segmentation

(Source: Kirillov et al., 2023)

Omar Barakat, "Depth estimation with deep Neural networks part 1," *Medium*, January 11, 2018

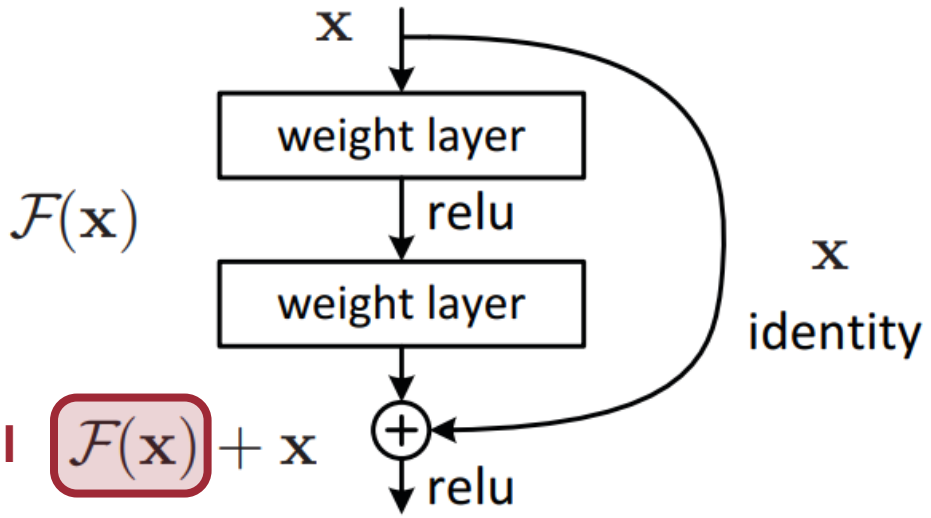
Mona Kharaji, Hossein Abbasi, Yasin Orouskhani, Mostafa Shomalzadeh, Foad Kazemi, and Maysam Orouskhani, "nnU-Net for Brain Tumor Segmentation," *Neuroscience Informatics*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick, "Segment Anything," *ICCV,* 2023.

ResNet: Residual Neural Network (He et al., 2016)

- **Intuition**: Learn how to **update** the input x to $x + \Delta x$

residual

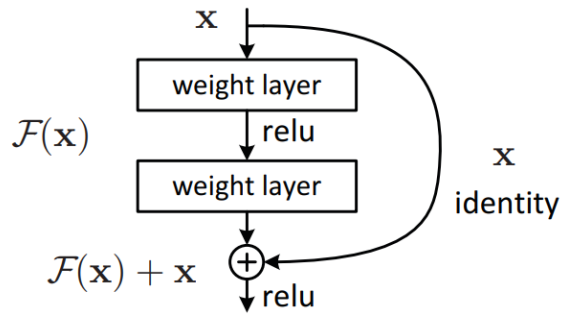


(Source: He et al., 2016)

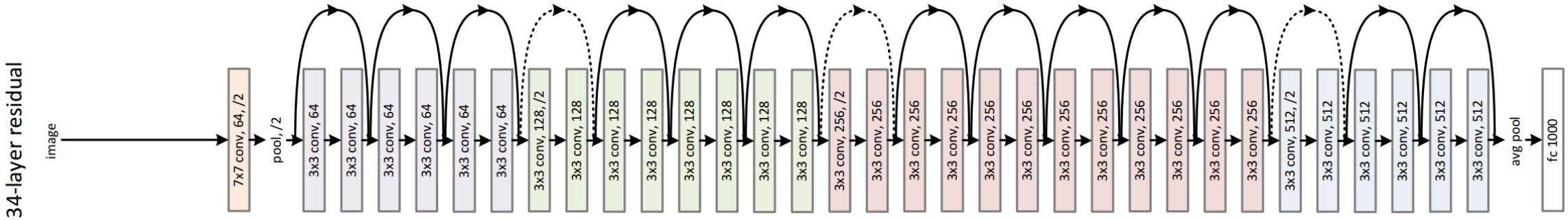
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition," *CVPR*, 2016.

ResNet: Residual Neural Network (He et al., 2016)

- **Intuition**: Learn how to **update** the input x to $x + \Delta x$



(Source: He et al., 2016)

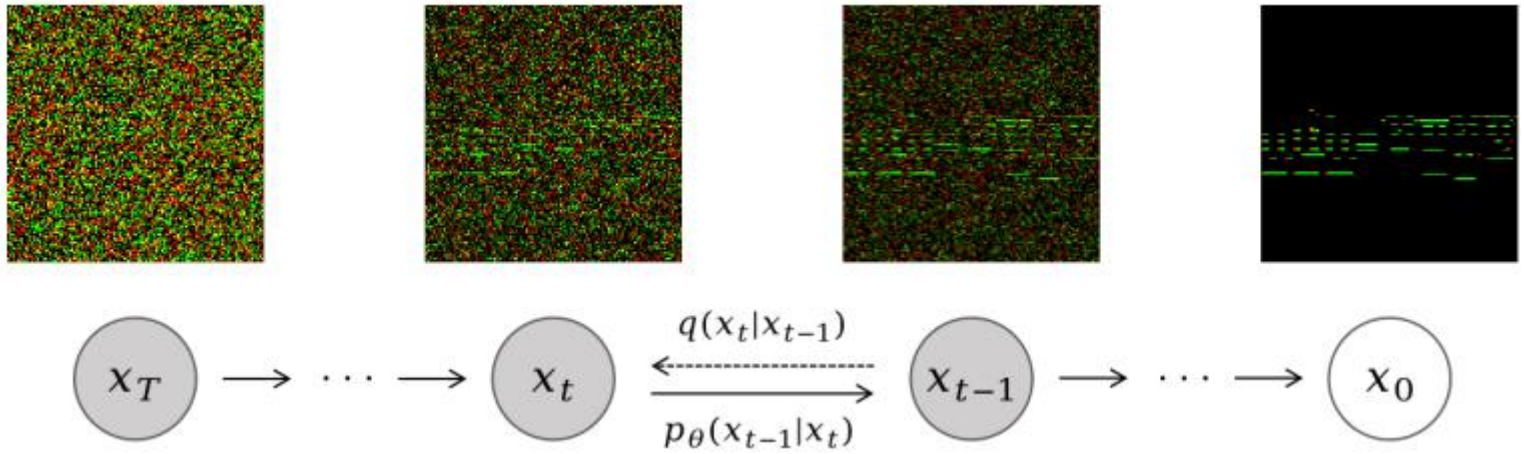


(Source: He et al., 2016)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition," *CVPR*, 2016.

Generating Music using Diffusion Models

Polyffusion (Min et al., 2023)

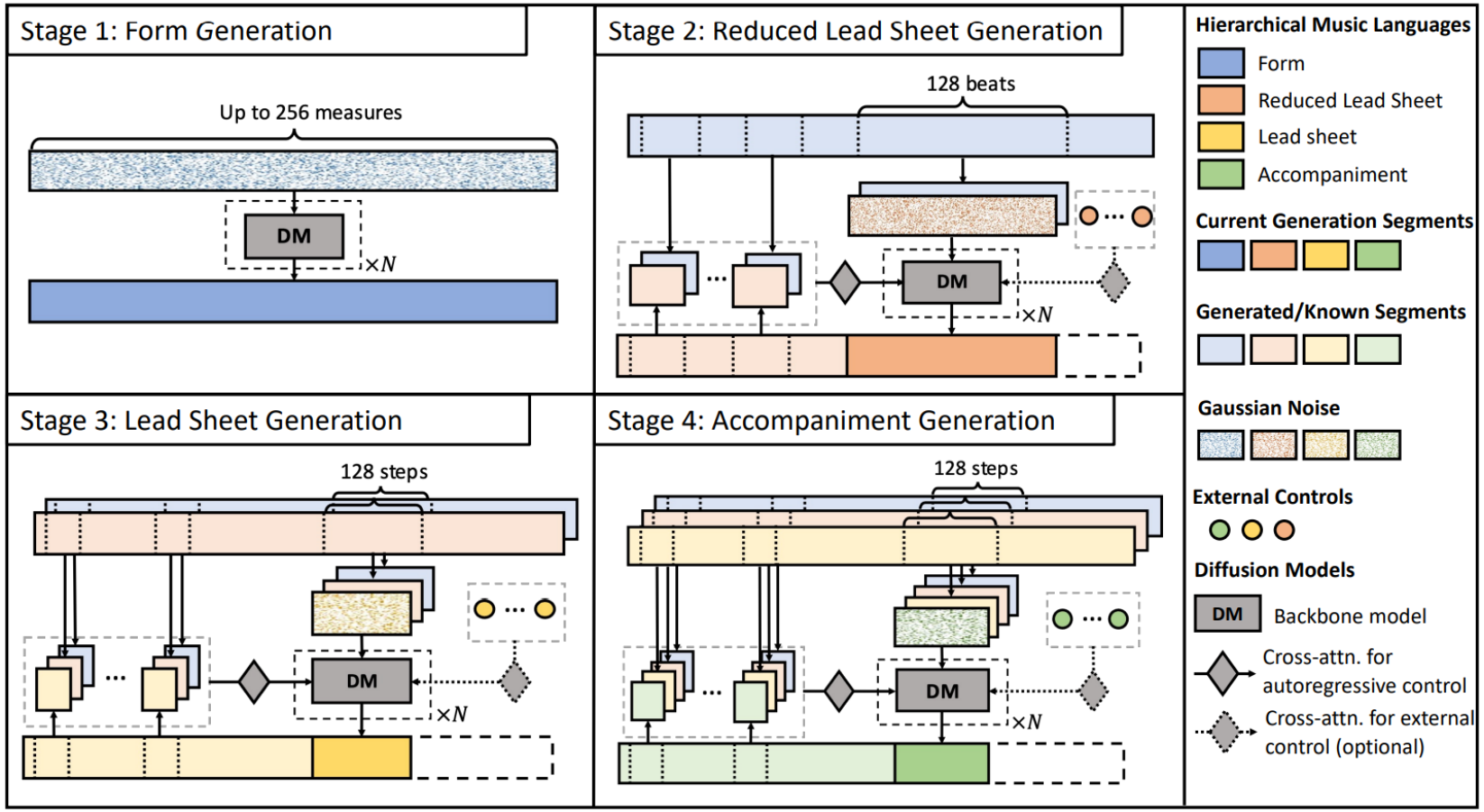


(Source: Min et al., 2023)

polyffusion.github.io

Lejun Min, Junyan Jiang, Gus Xia, and Jingwei Zhao, "Polyffusion: A Diffusion Model for Polyphonic Score Generation with Internal and External Controls," *ISMIR*, 2023.

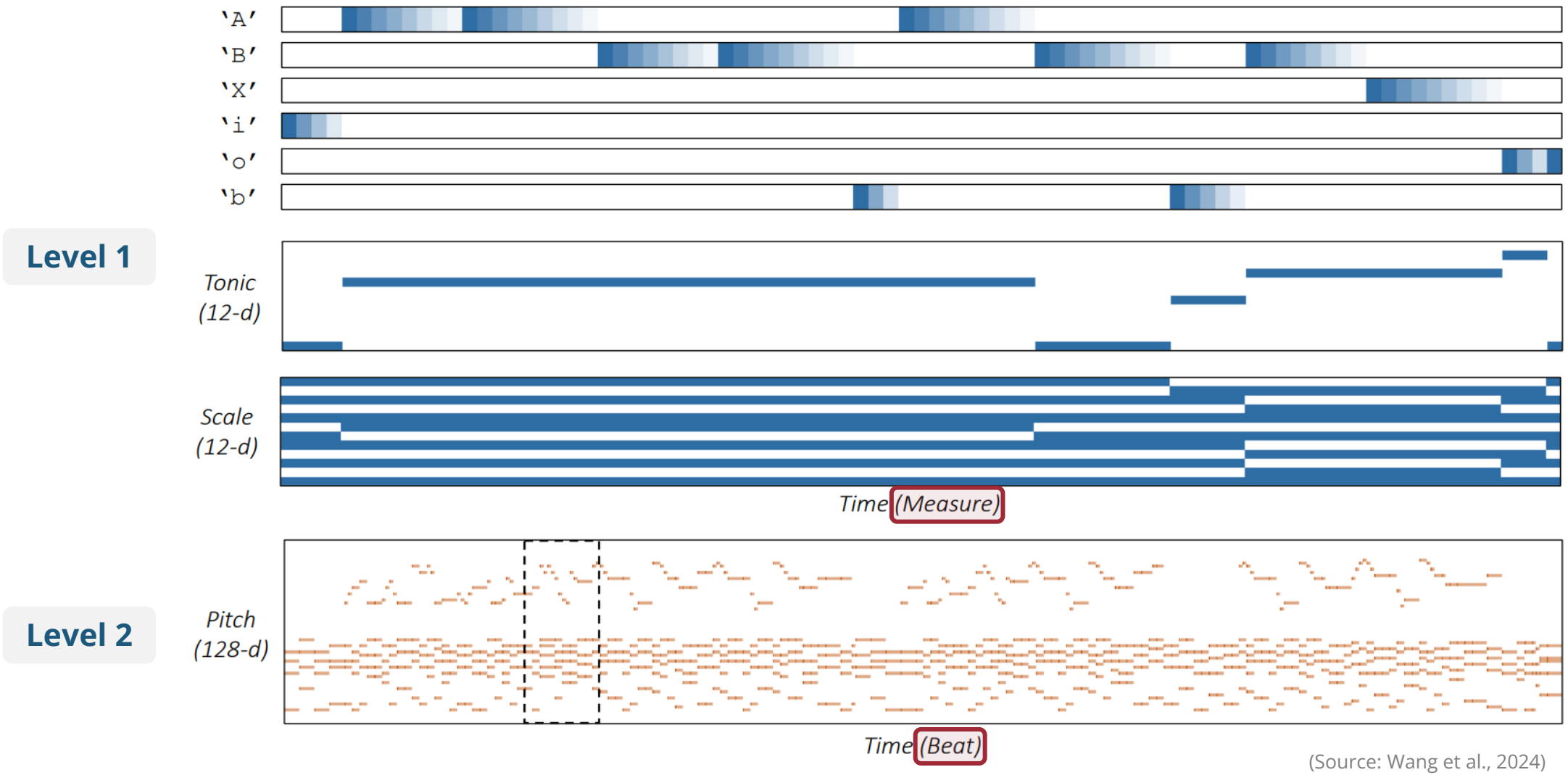
Cascaded Diffusion Models (Wang et al., 2024)



(Source: Wang et al., 2024)

Ziyu Wang, Lejun Min, and Gus Xia, "Whole-Song Hierarchical Generation of Symbolic Music Using Cascaded Diffusion Models," *ICLR*, 2024.

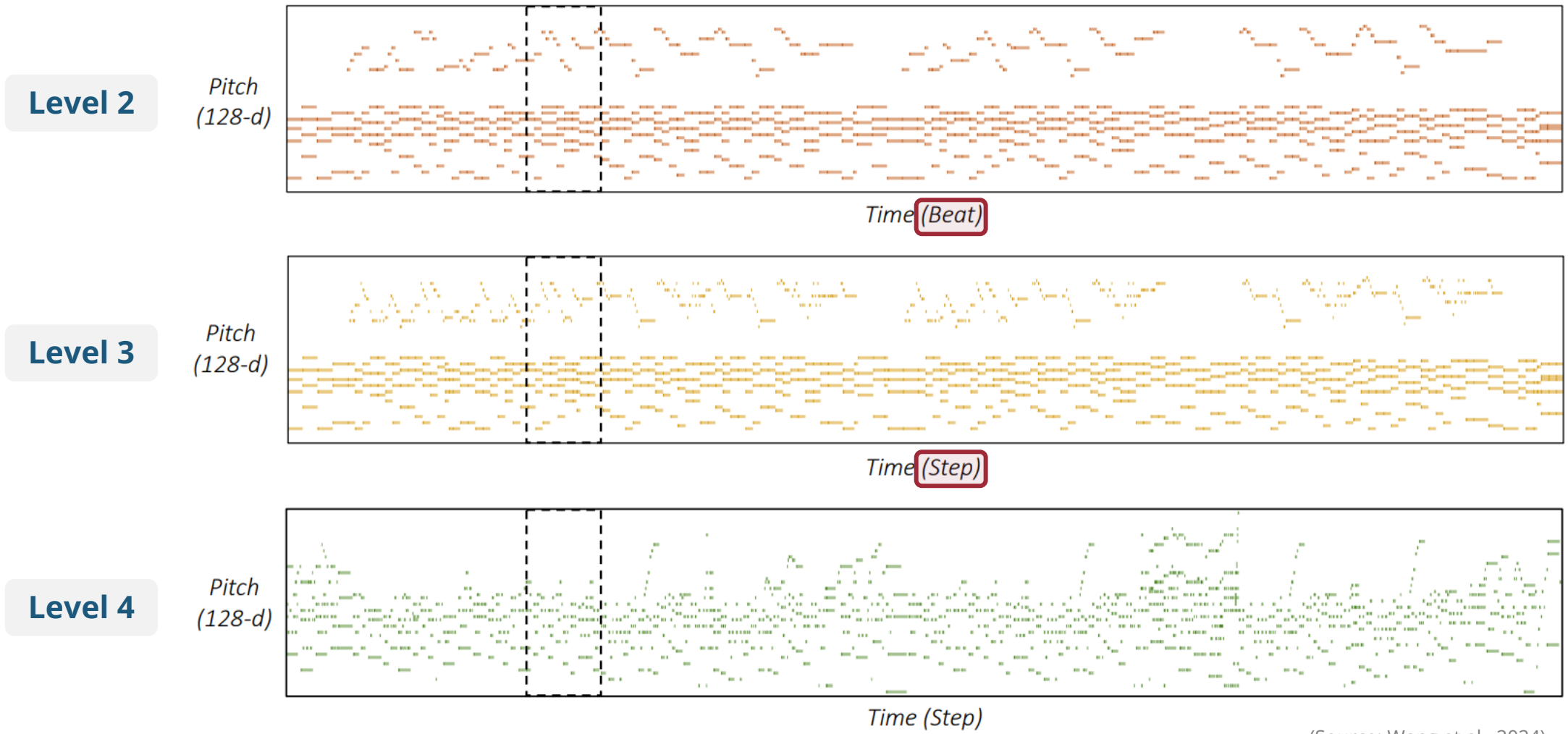
Cascaded Diffusion Models (Wang et al., 2024)



(Source: Wang et al., 2024)

Ziyu Wang, Lejun Min, and Gus Xia, "Whole-Song Hierarchical Generation of Symbolic Music Using Cascaded Diffusion Models," *ICLR*, 2024.

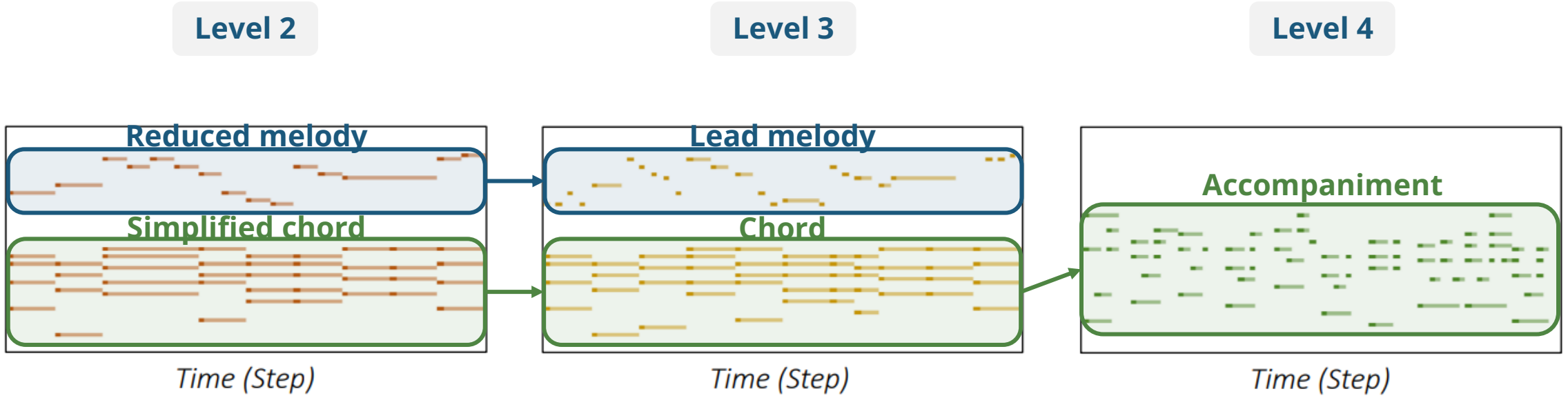
Cascaded Diffusion Models (Wang et al., 2024)



(Source: Wang et al., 2024)

Ziyu Wang, Lejun Min, and Gus Xia, "Whole-Song Hierarchical Generation of Symbolic Music Using Cascaded Diffusion Models," *ICLR*, 2024.

Cascaded Diffusion Models (Wang et al., 2024)



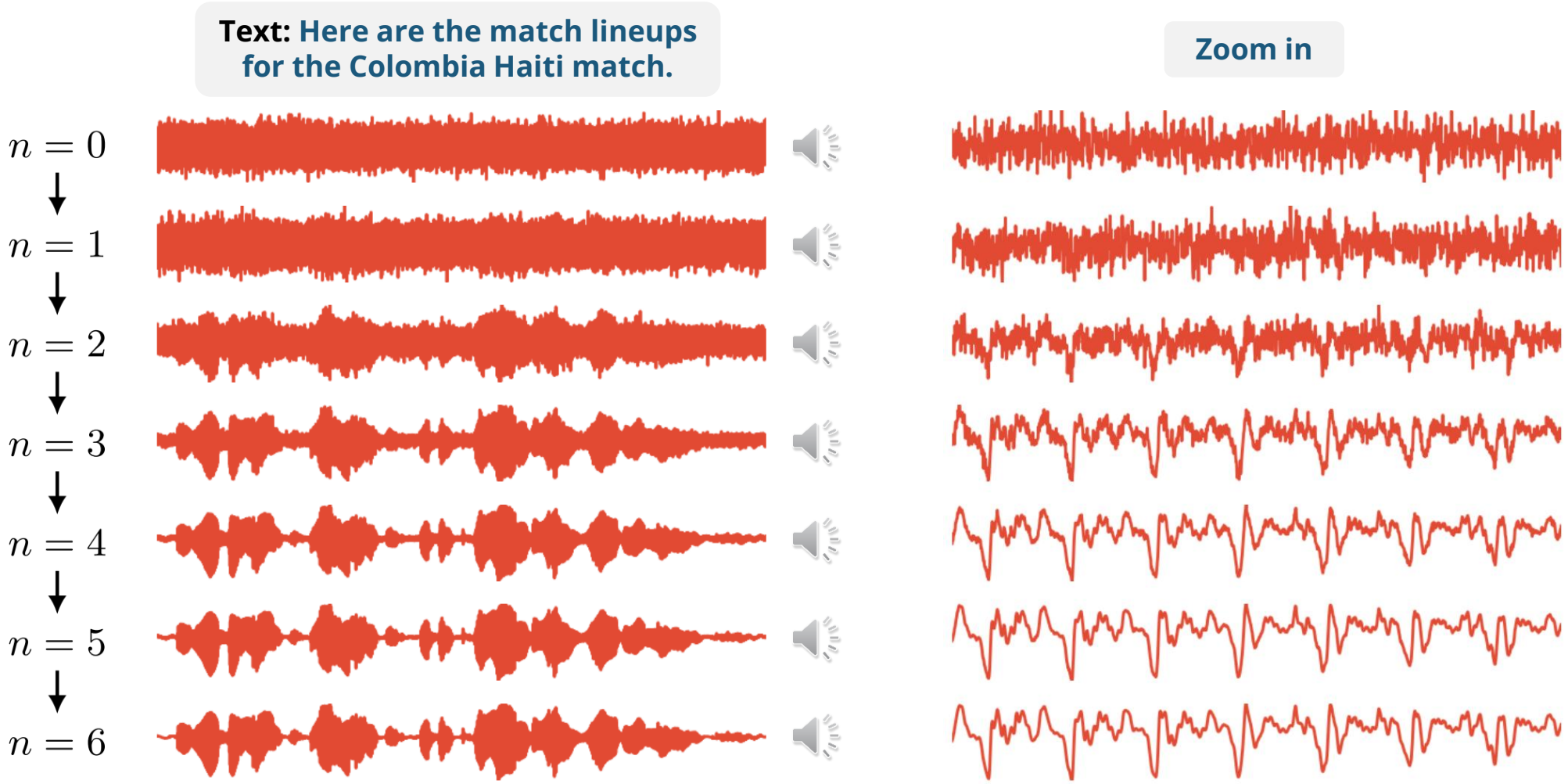
(Source: Wang et al., 2024)

wholesonggen.github.io

Ziyu Wang, Lejun Min, and Gus Xia, "Whole-Song Hierarchical Generation of Symbolic Music Using Cascaded Diffusion Models," *ICLR*, 2024.

Diffusion Models for Audio

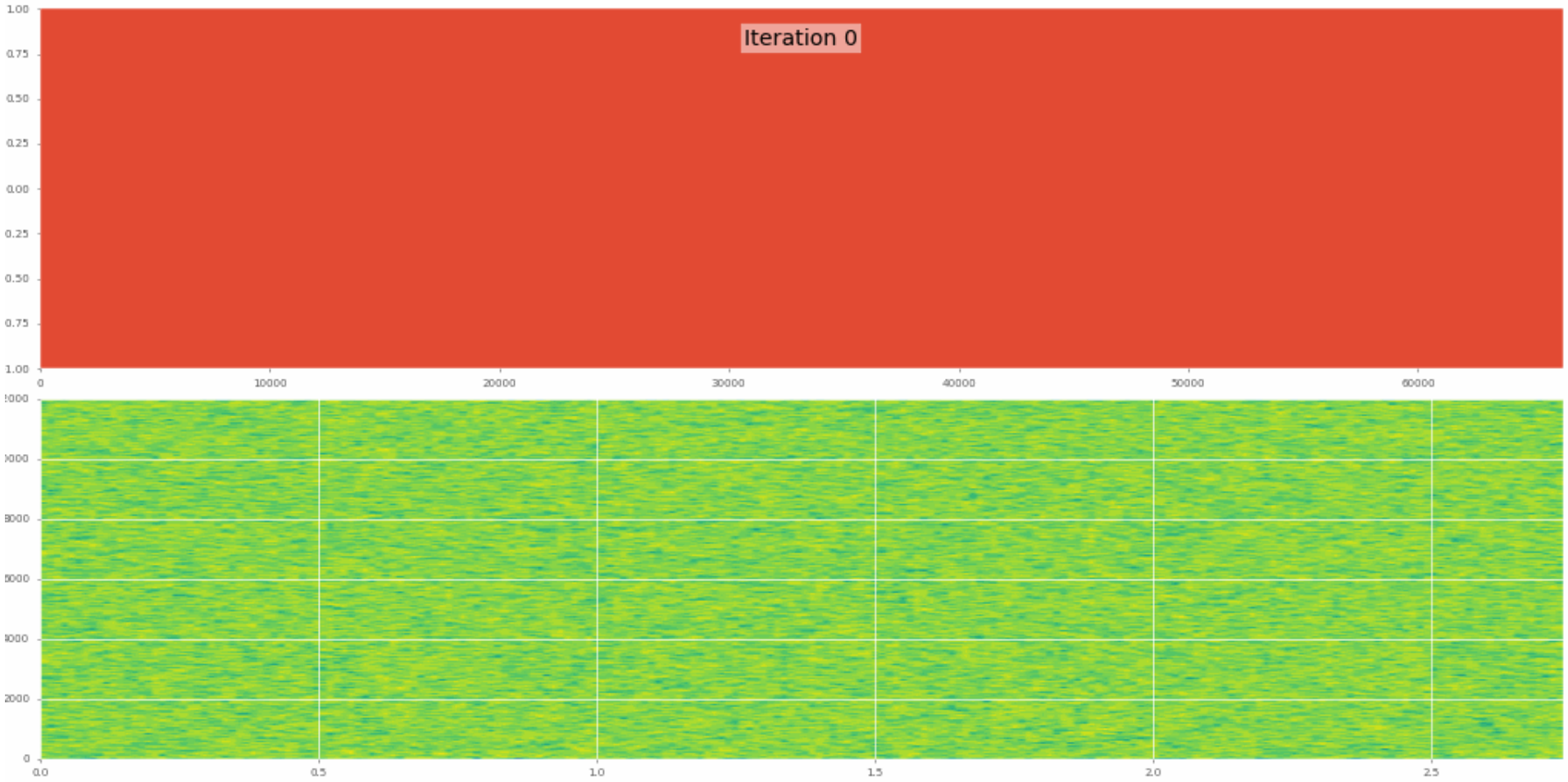
WaveGrad: Diffusion for Waveforms (Chen et al., 2021)



(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.

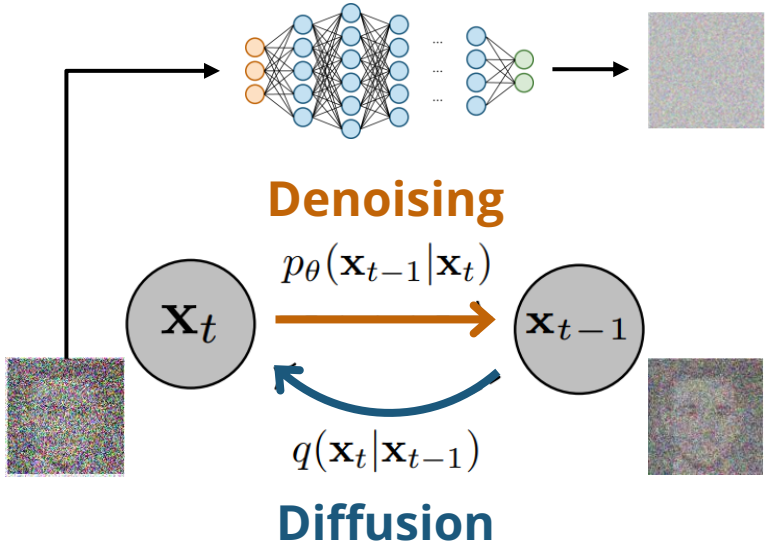
WaveGrad: Diffusion for Waveforms (Chen et al., 2021)



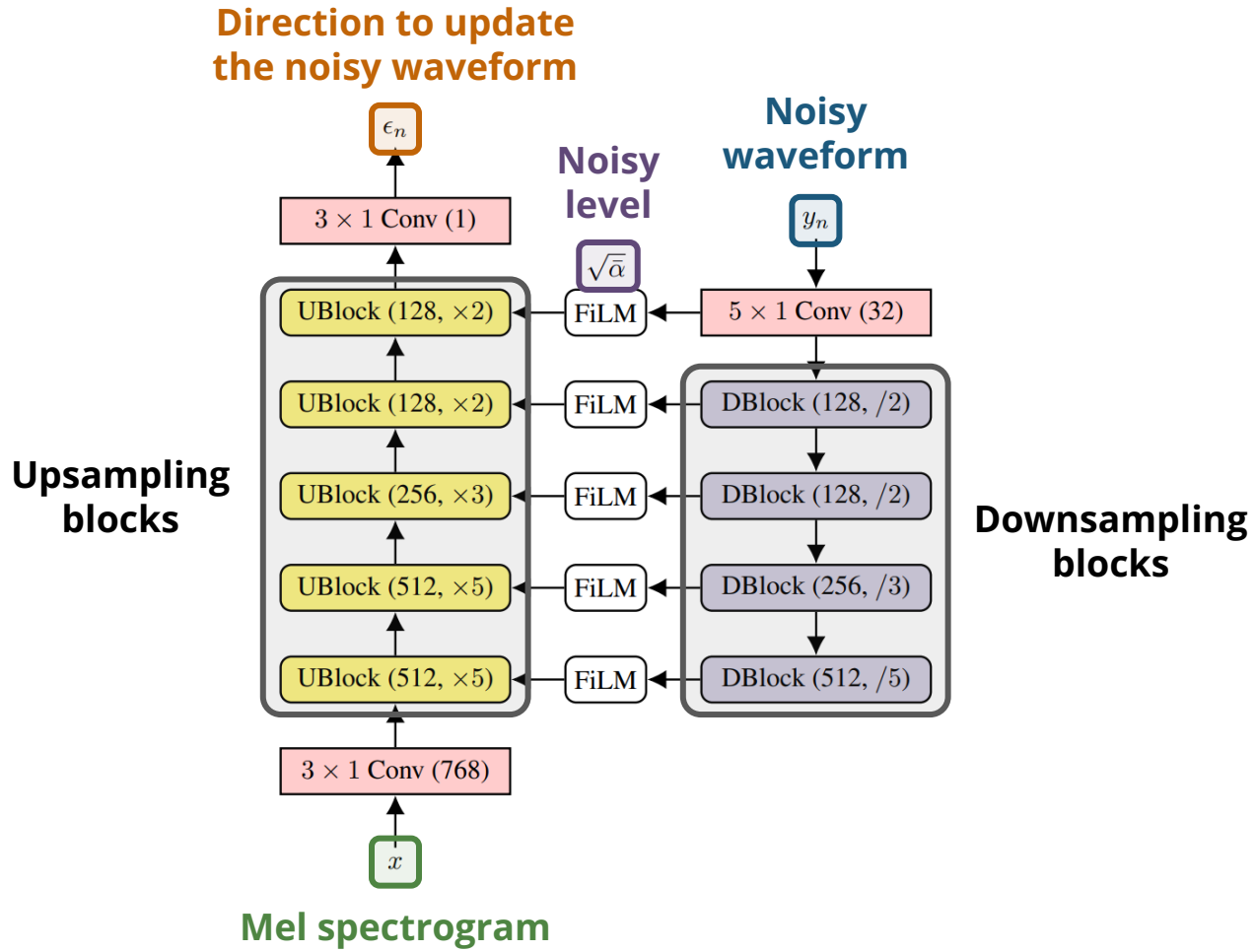
(Source: Chen et al., 2021)

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 (Chen et al., 2021)



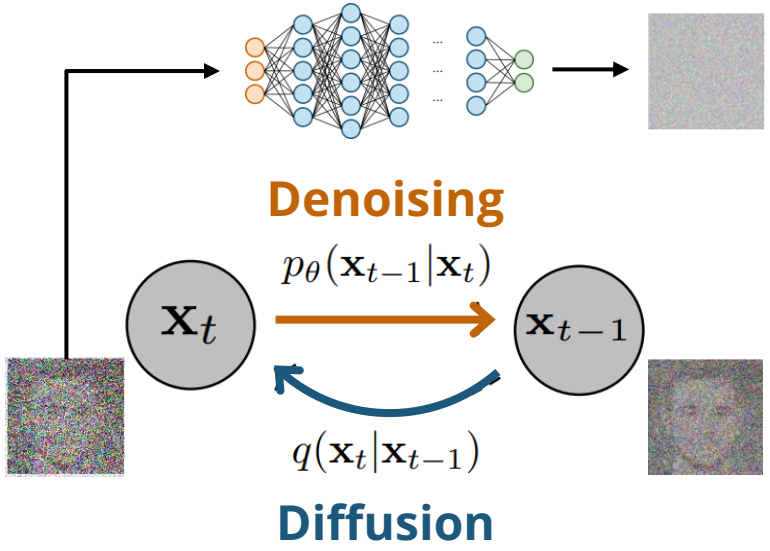
(Source: Ho et al., 2020)



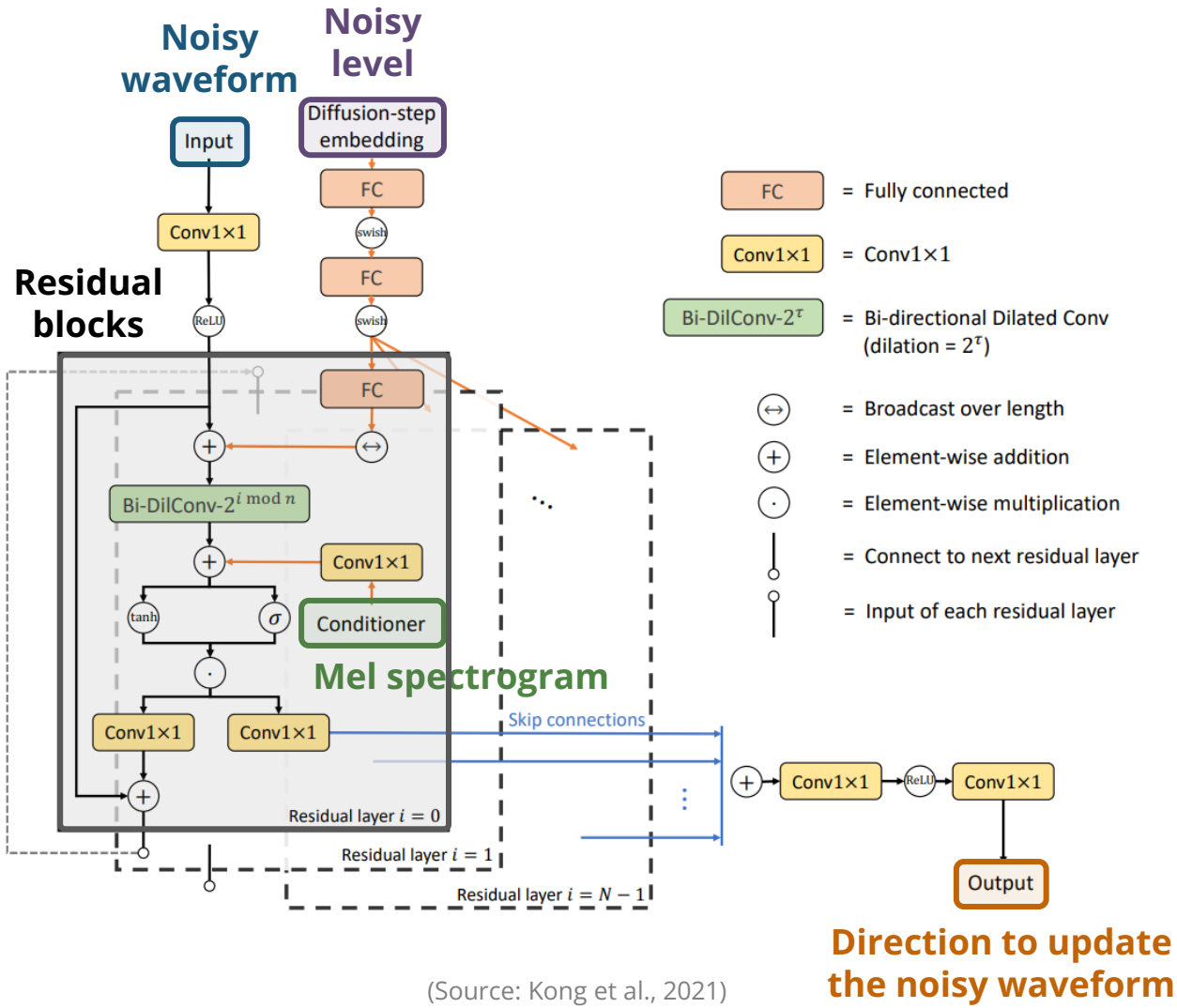
(Source: Chen et al., 2021)

Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.
Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan, "WaveGrad: Estimating Gradients for Waveform Generation," *ICLR*, 2021.

DiffWave: Diffusion Model for Waveforms (Kong et al., 2021)



(Source: Ho et al., 2020)



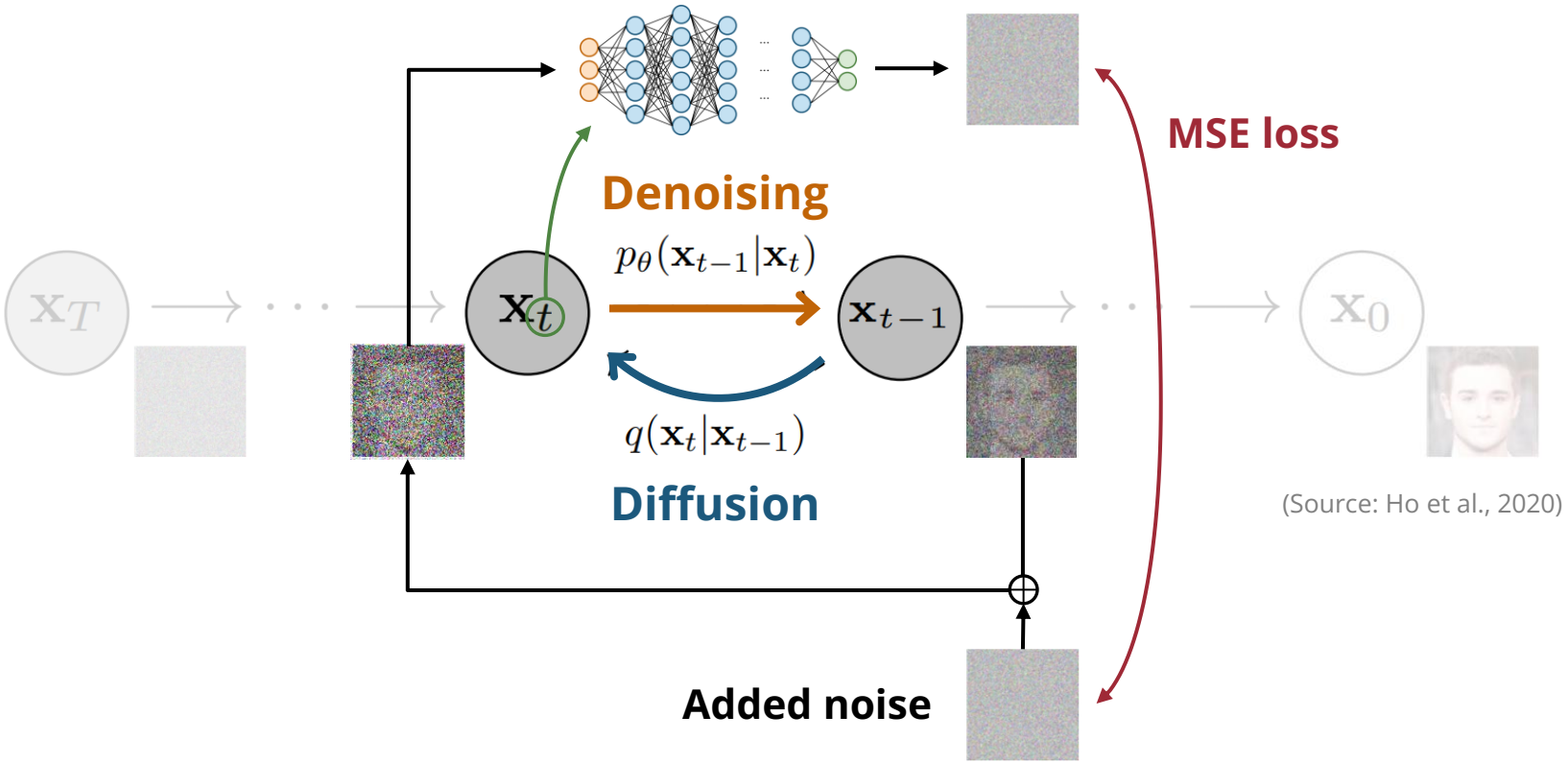
(Source: Kong et al., 2021)

Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.
Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan, "WaveGrad: Estimating Gradients for Waveform Generation," *ICLR*, 2021.

Efficient Diffusion Models

Diffusion Models: Training

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

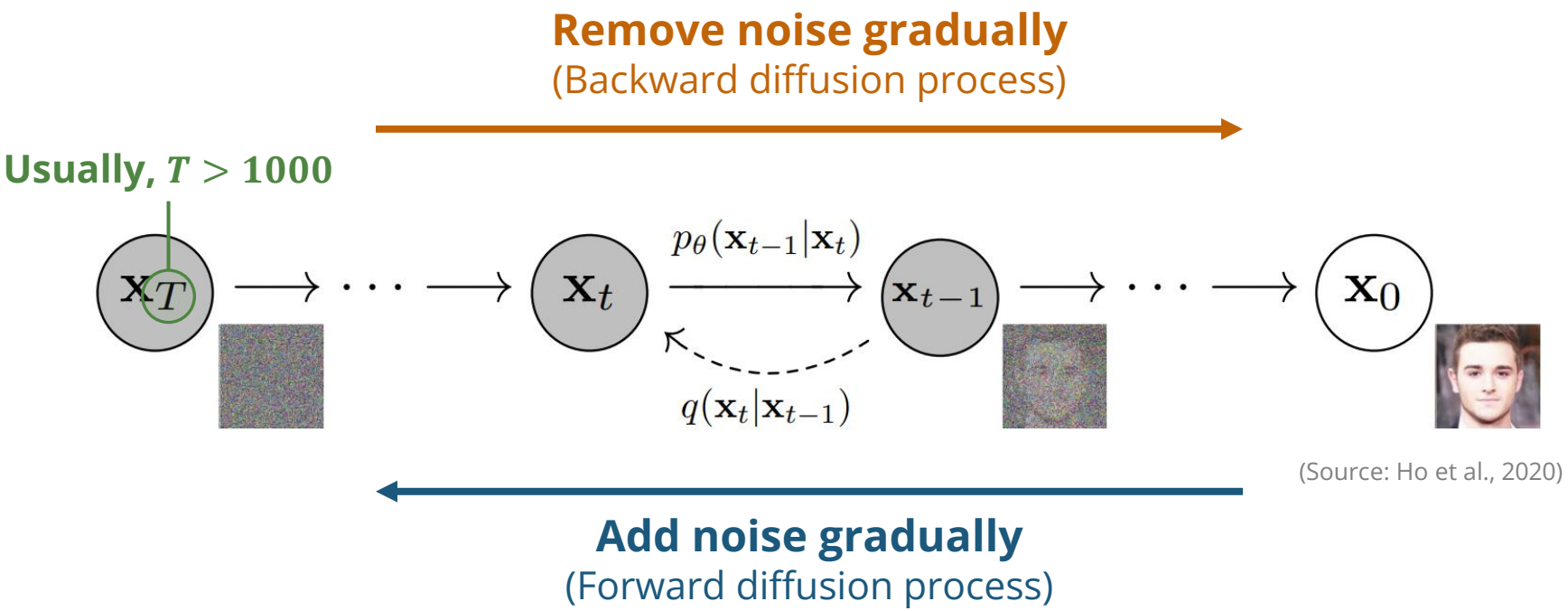


(Source: Ho et al., 2020)

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

Diffusion Models (Ho et al., 2020)

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

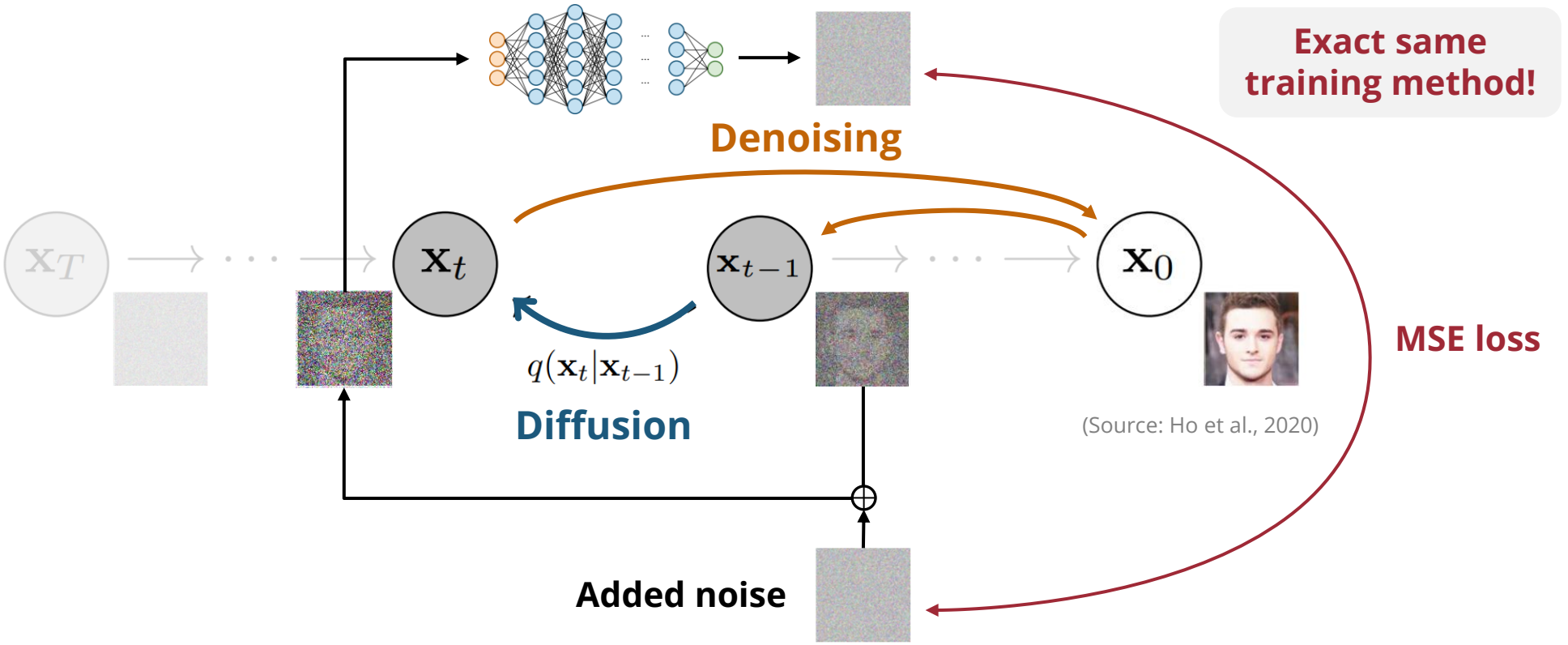


(Source: Ho et al., 2020)

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

Fast Sampling for Diffusion Models

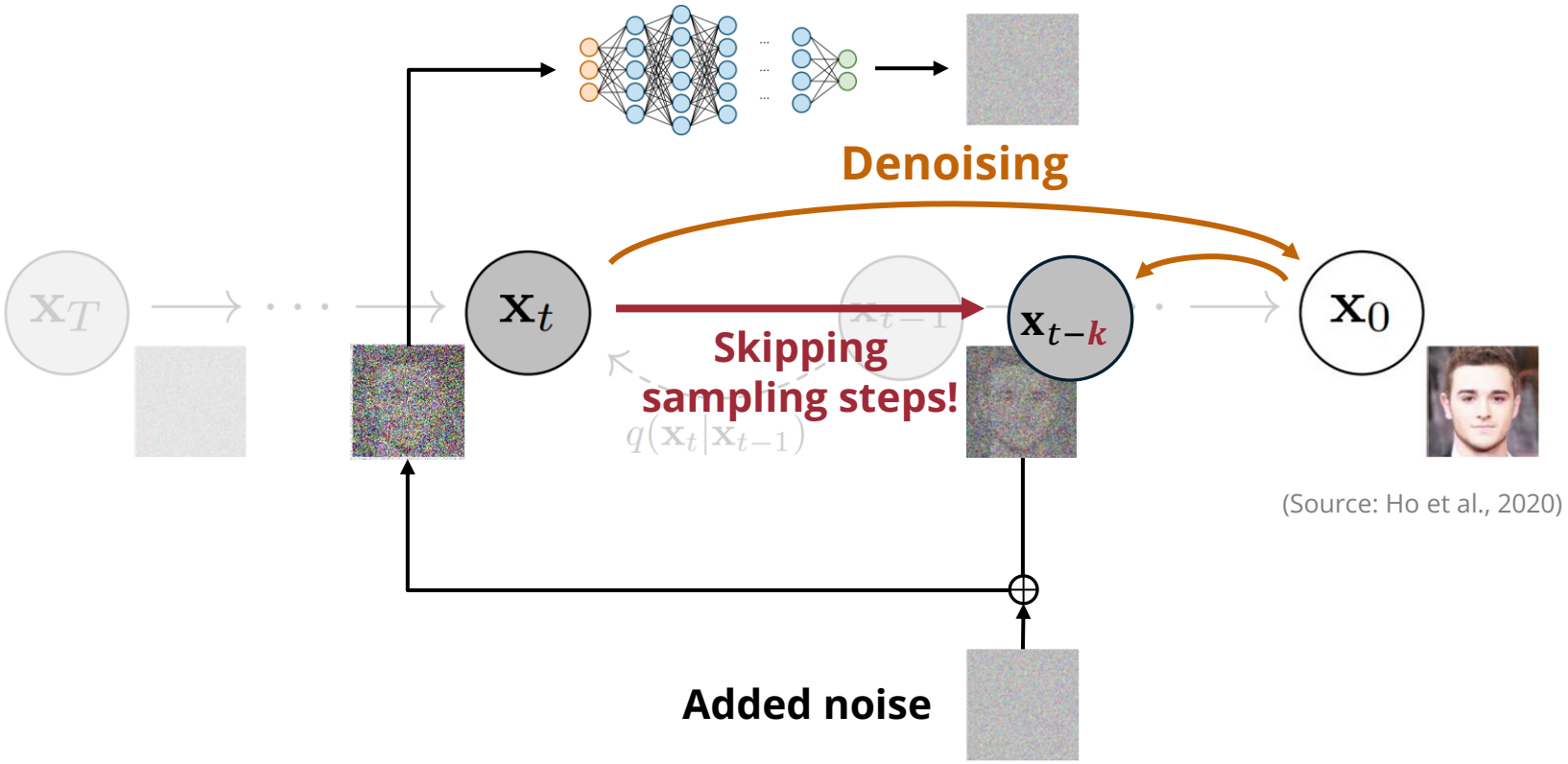
- **Intuition**: Skip some sampling steps



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

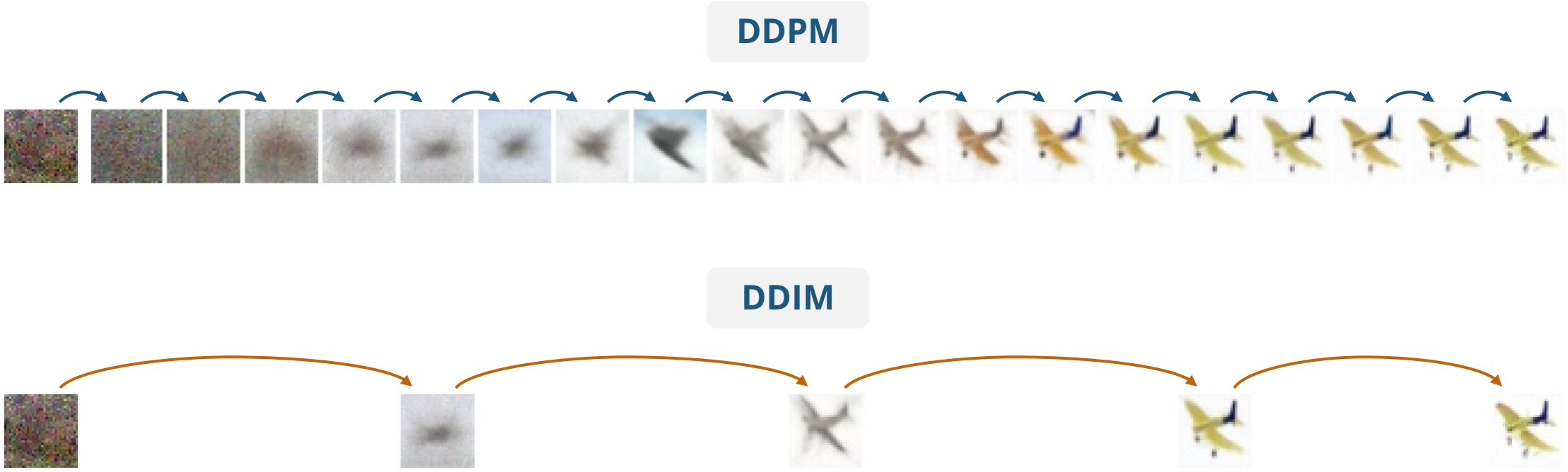


(Source: Ho et al., 2020)

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

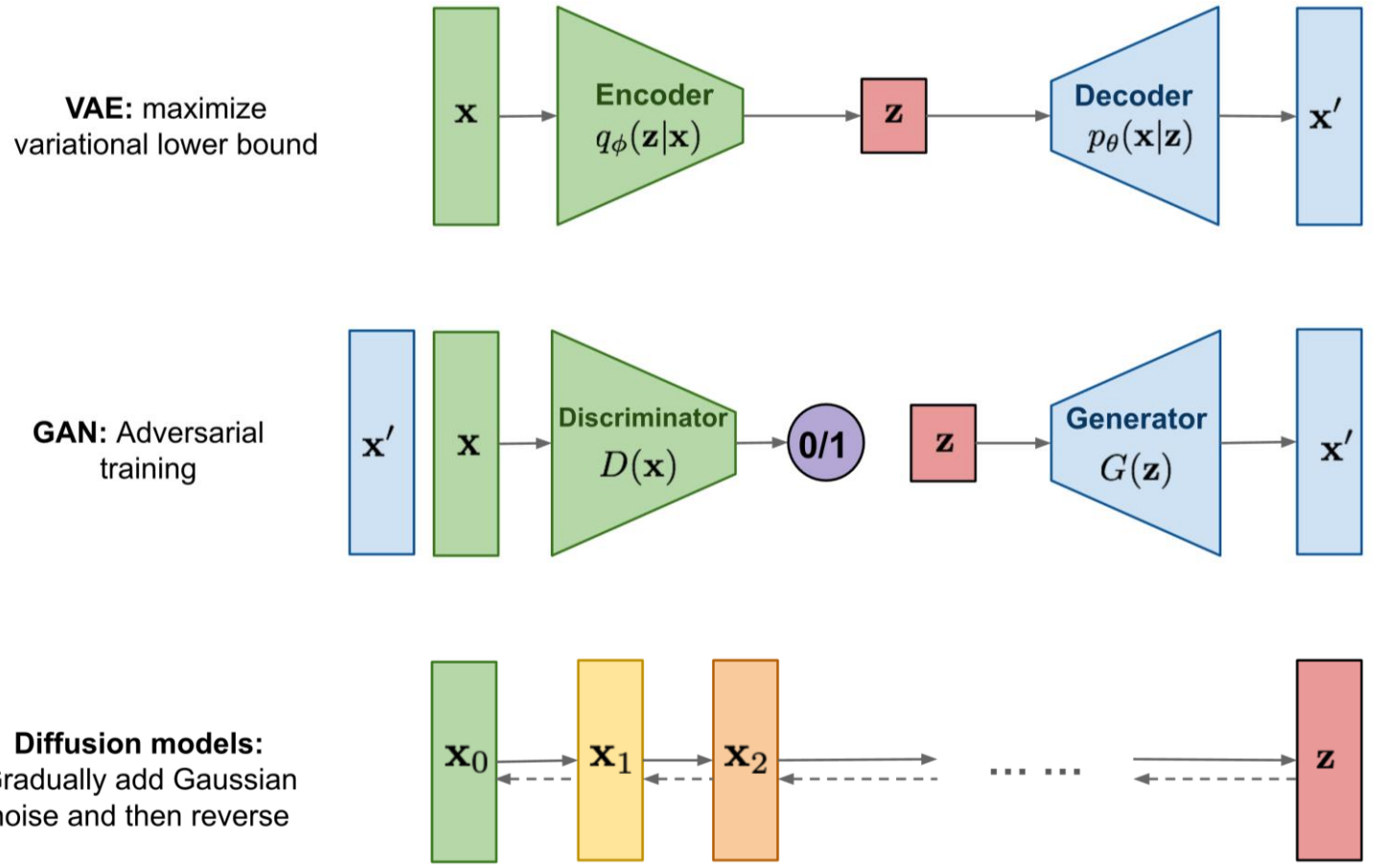


(Source: Ho et al., 2020)

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

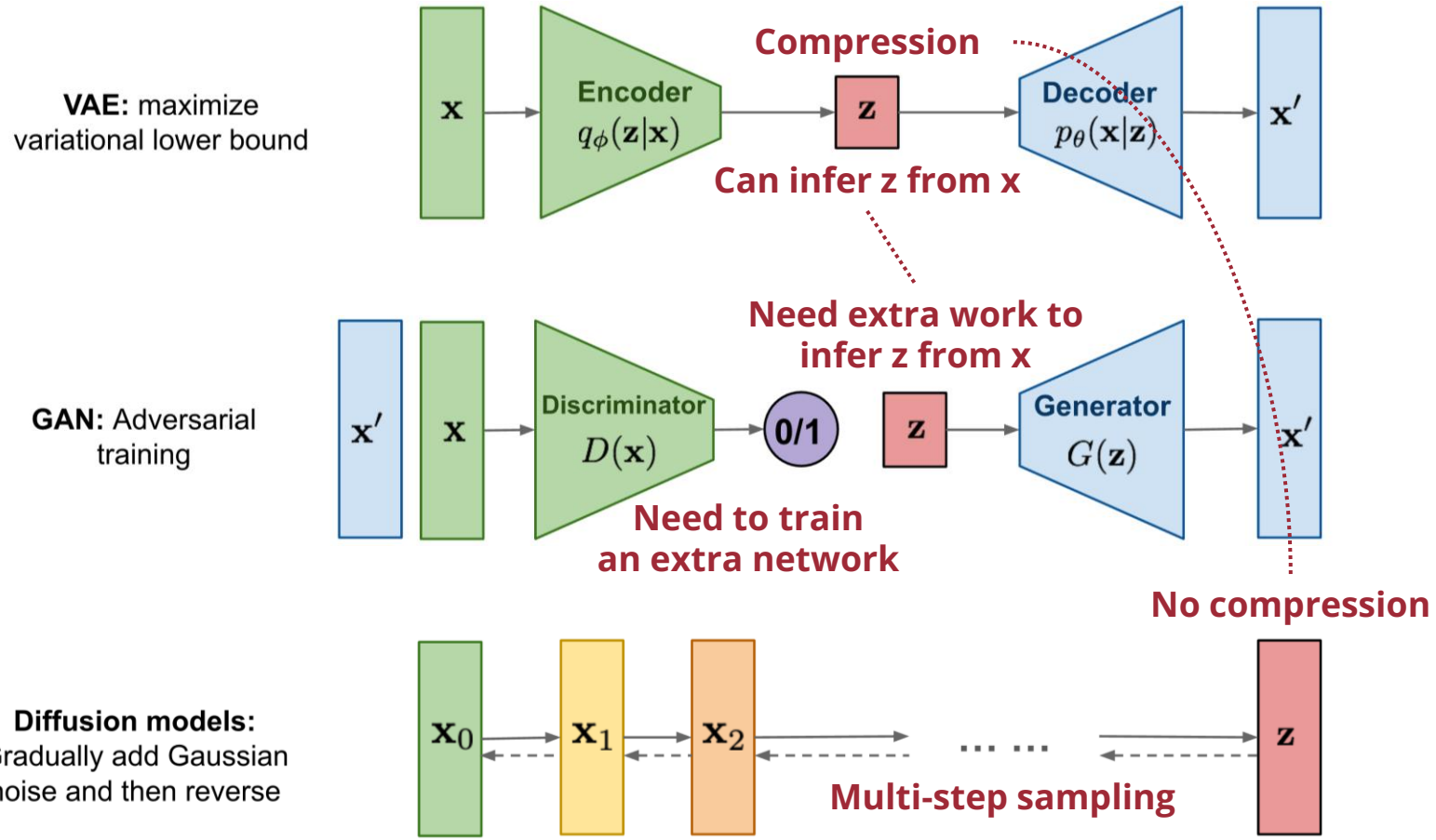
Comparison of Deep Latent Variable Models



(Source: Weng, 2021)

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

Comparison of Deep Latent Variable Models



(Source: Weng, 2021)

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

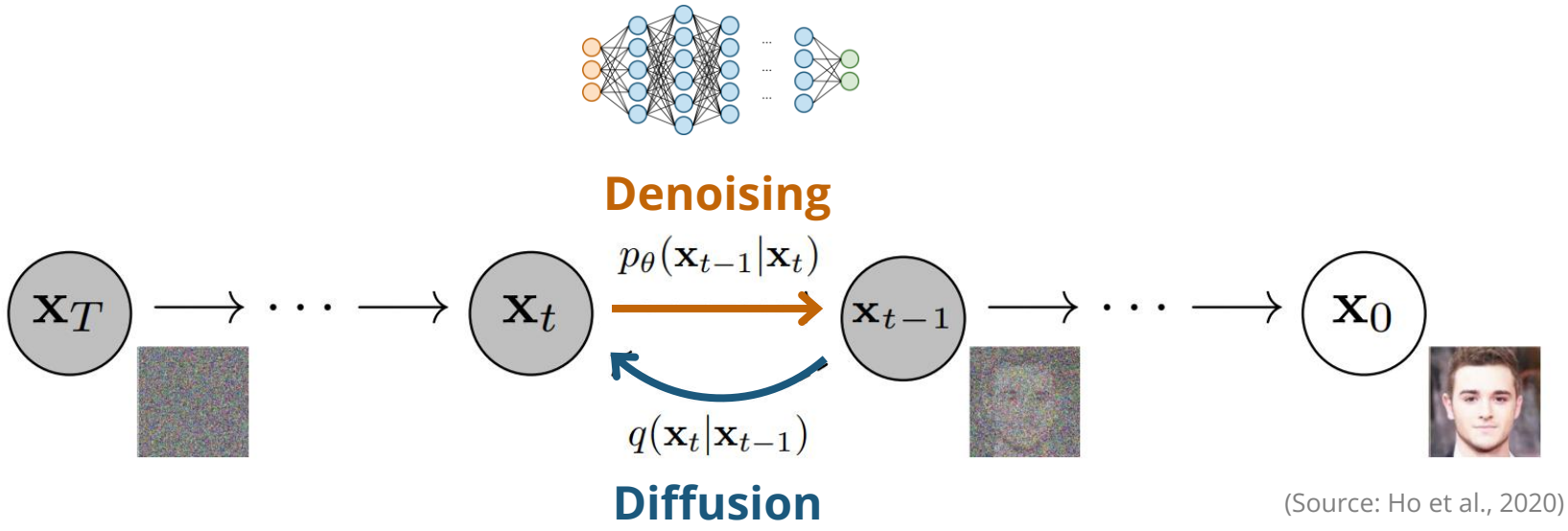
Network Architectures vs. Training Frameworks

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
⋮	⋮

Recap

Diffusion Models (Ho et al., 2020)

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

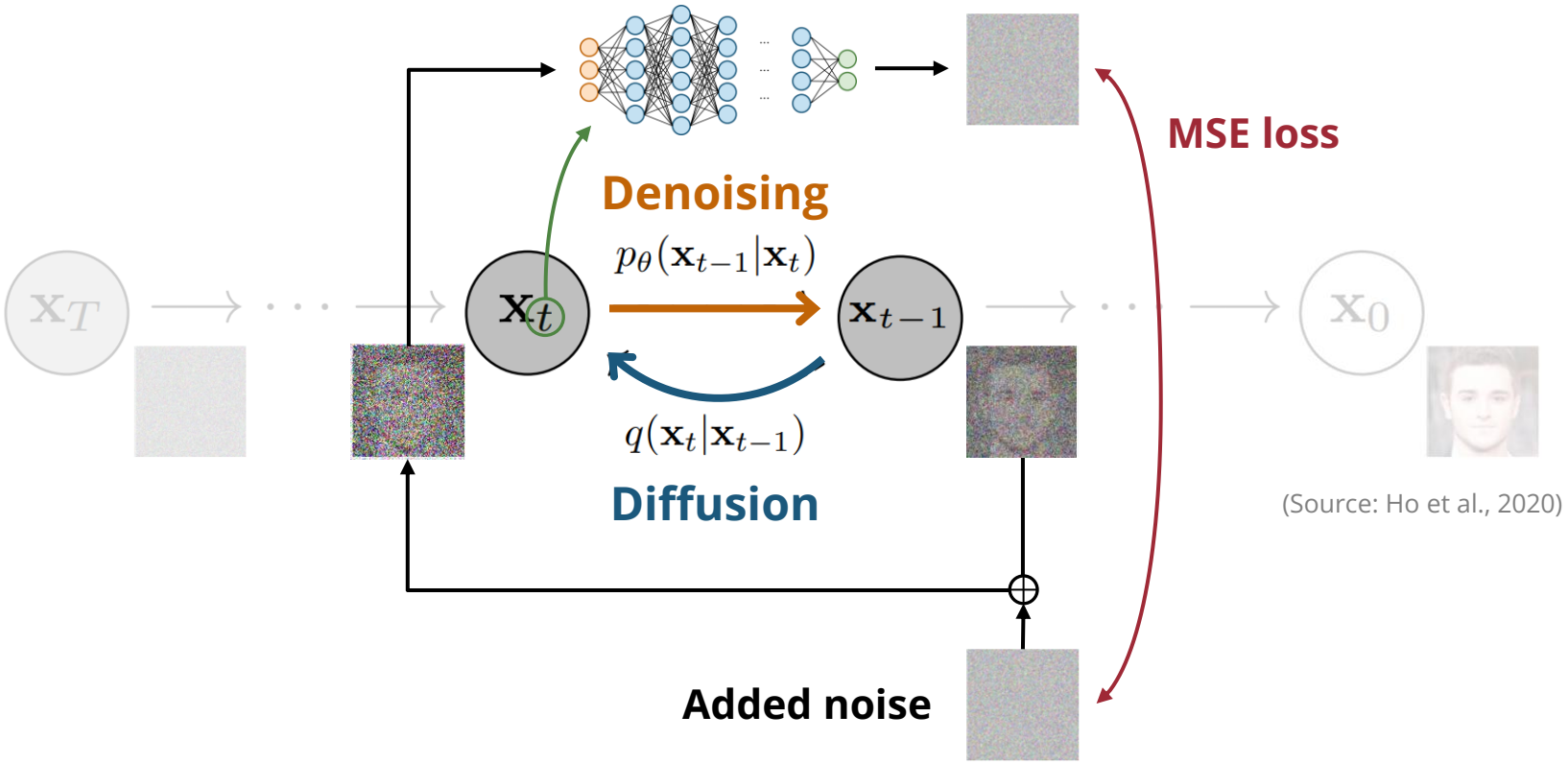


(Source: Ho et al., 2020)

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

Diffusion Models: Training

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

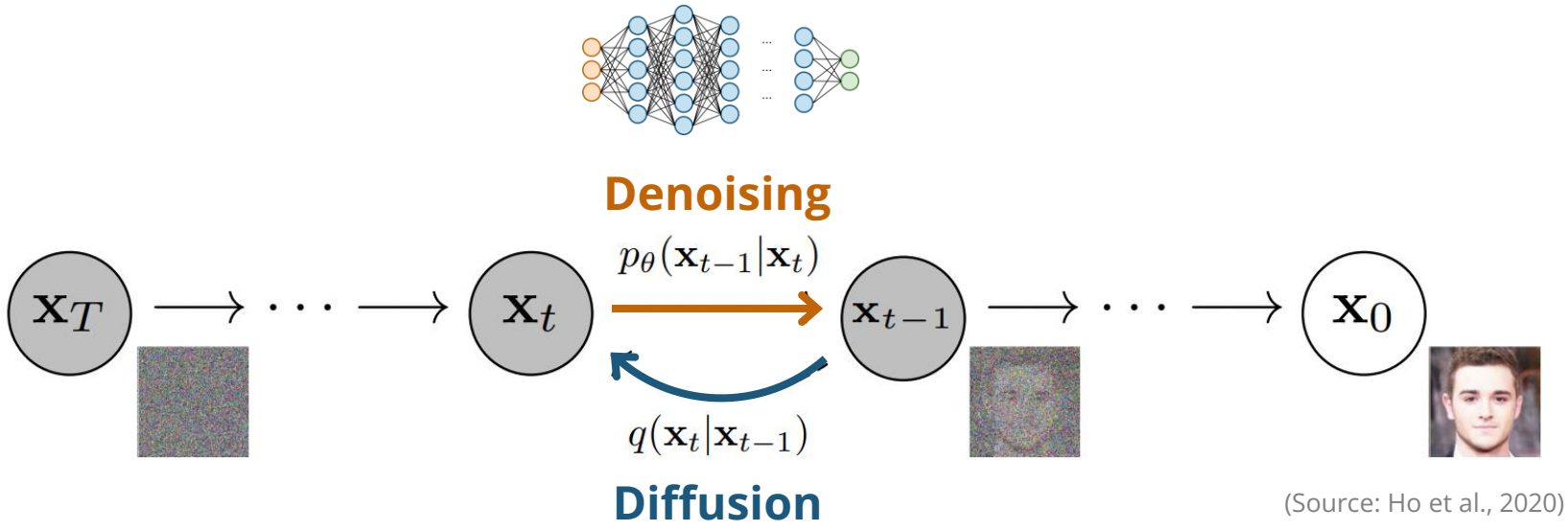


(Source: Ho et al., 2020)

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

Diffusion Models (Ho et al., 2020)

- **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: Generation

Remove noise gradually
(Backward diffusion process)

Input → **Output**

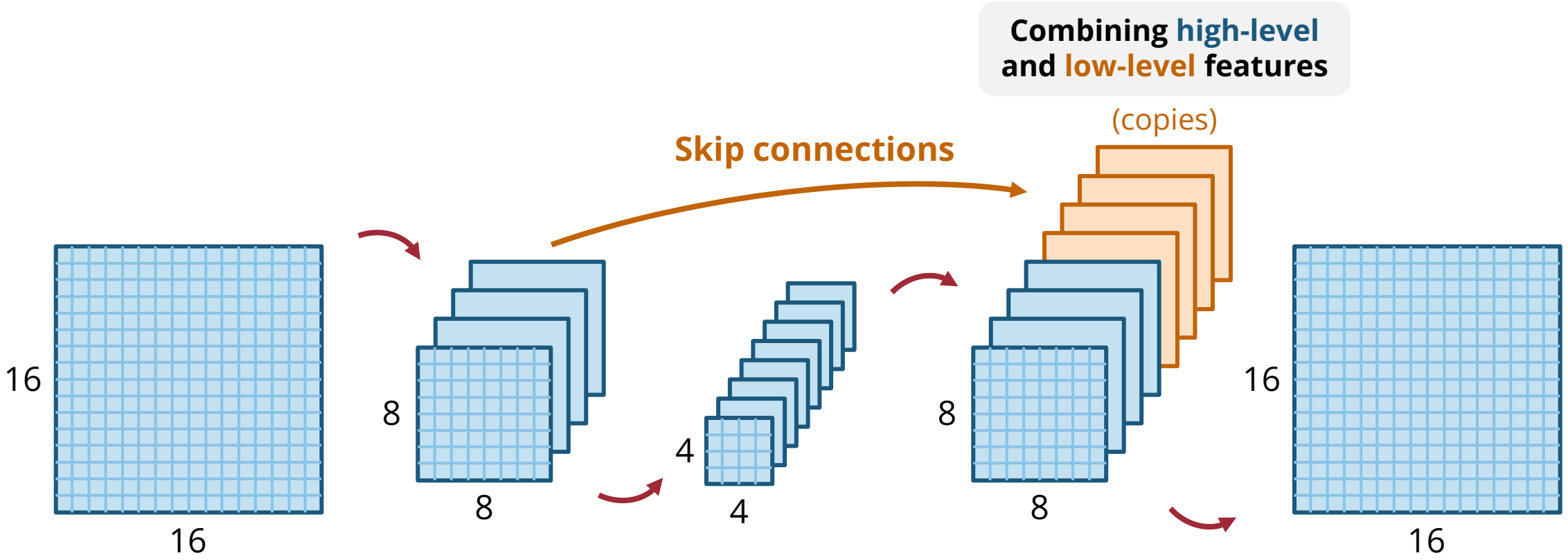
Coarse shapes
(low-frequency components)

Fine details
(high-frequency components)

(Source: Ho et al., 2020)

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

A Toy Example of U-Net

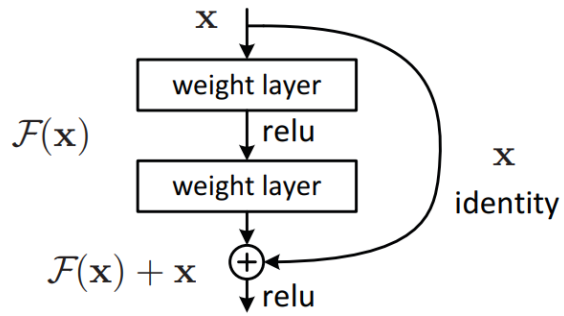



```
nn.Conv2d(8, 1, 3, padding="same")
nn.Upsample(scale_factor=2)
```

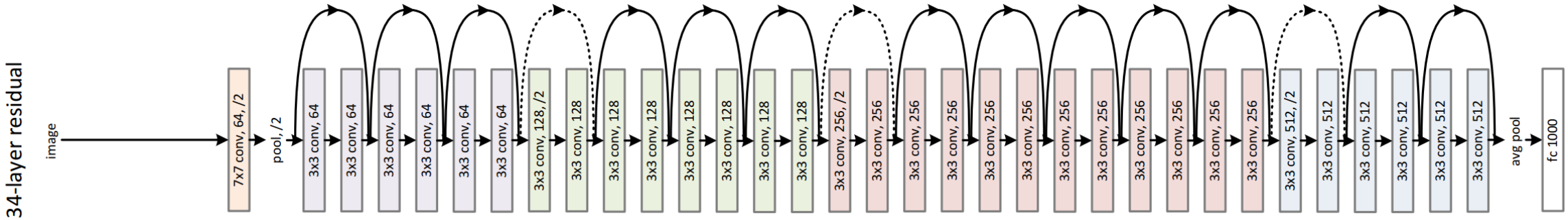
U-Nets are useful when the inputs and outputs have the same shape!

ResNet: Residual Neural Network (He et al., 2016)

- **Intuition**: Learn how to **update** the input x to $x + \Delta x$



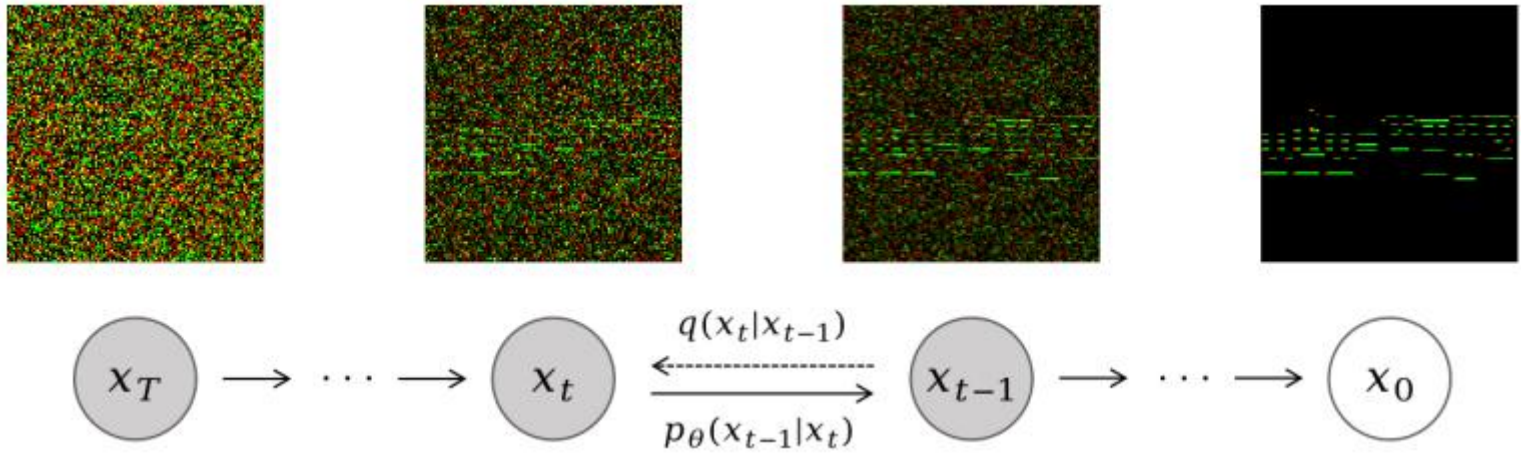
(Source: He et al., 2016)



(Source: He et al., 2016)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition," *CVPR*, 2016.

Polyffusion (Min et al., 2023)

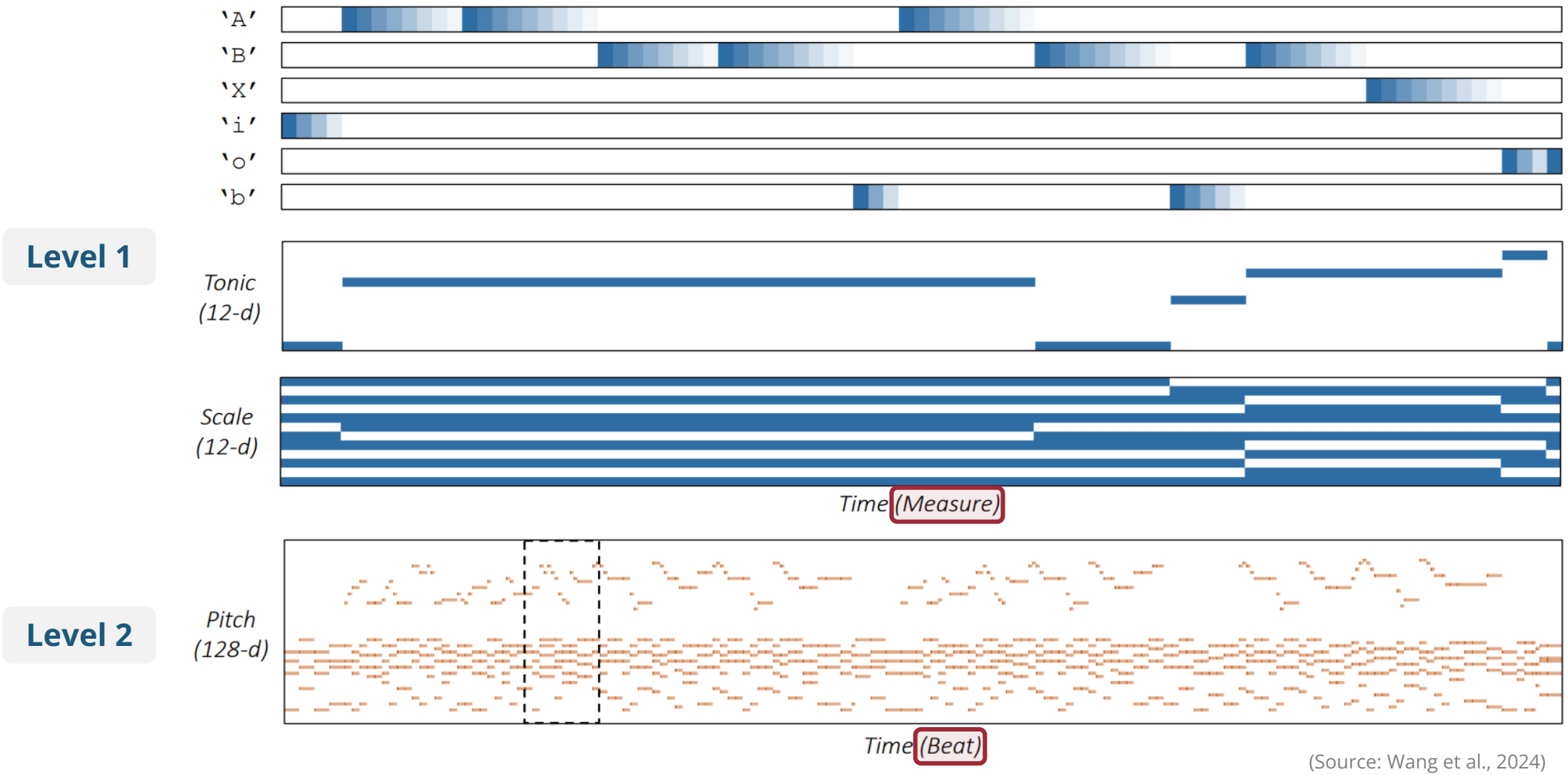


(Source: Min et al., 2023)

polyffusion.github.io

Lejun Min, Junyan Jiang, Gus Xia, and Jingwei Zhao, "Polyffusion: A Diffusion Model for Polyphonic Score Generation with Internal and External Controls," *ISMIR*, 2023.

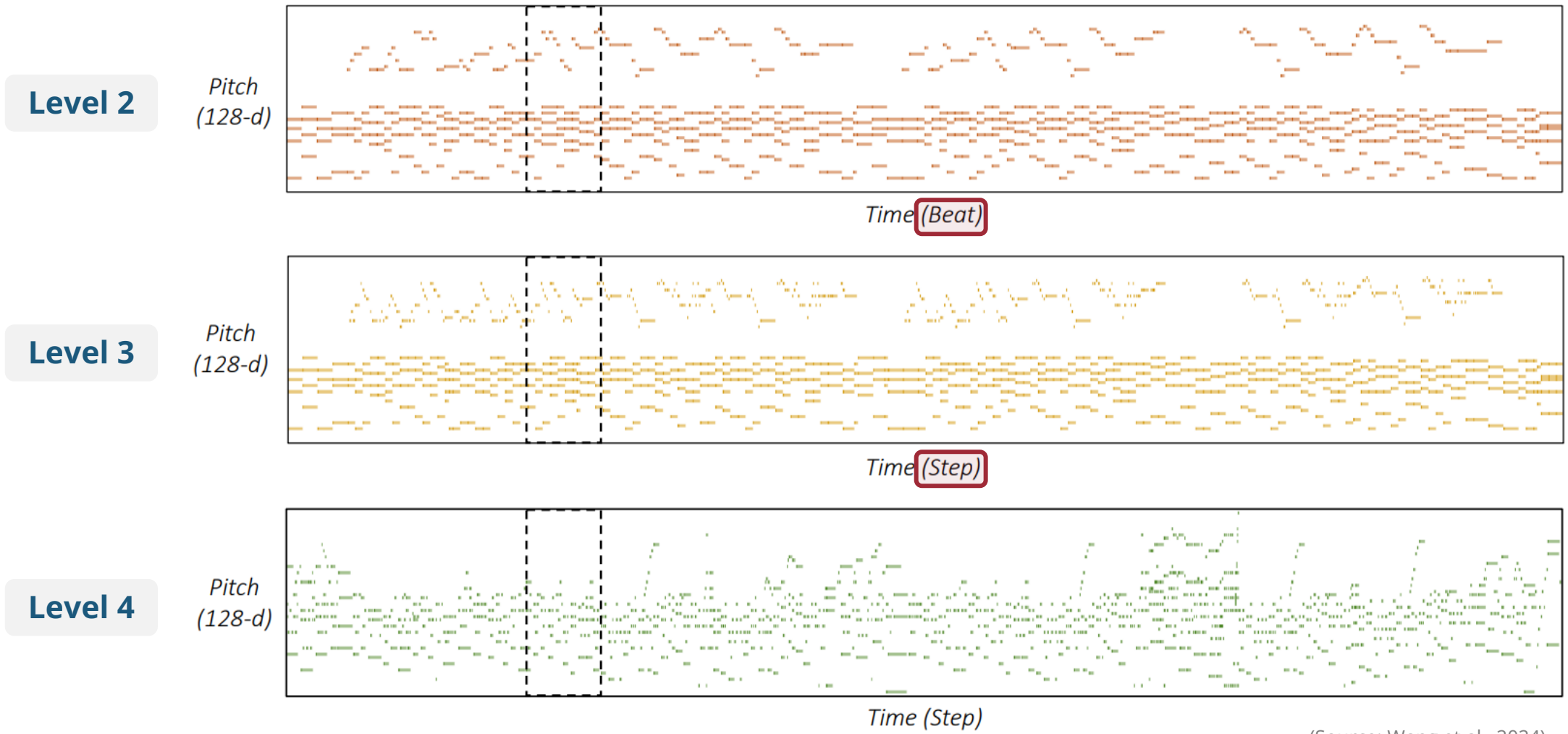
Cascaded Diffusion Models (Wang et al., 2024)



(Source: Wang et al., 2024)

Ziyu Wang, Lejun Min, and Gus Xia, "Whole-Song Hierarchical Generation of Symbolic Music Using Cascaded Diffusion Models," *ICLR*, 2024.

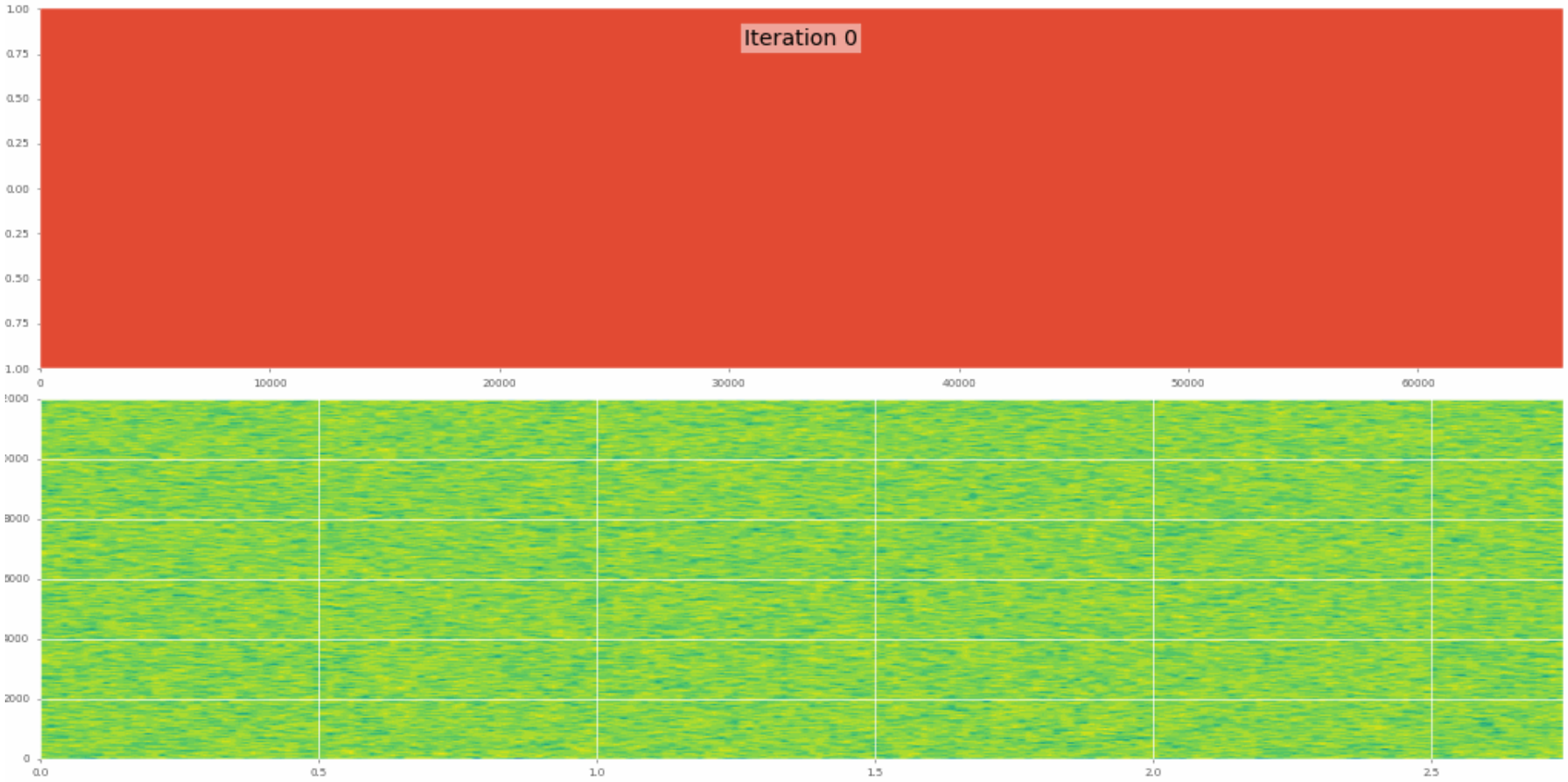
Cascaded Diffusion Models (Wang et al., 2024)



(Source: Wang et al., 2024)

Ziyu Wang, Lejun Min, and Gus Xia, "Whole-Song Hierarchical Generation of Symbolic Music Using Cascaded Diffusion Models," *ICLR*, 2024.

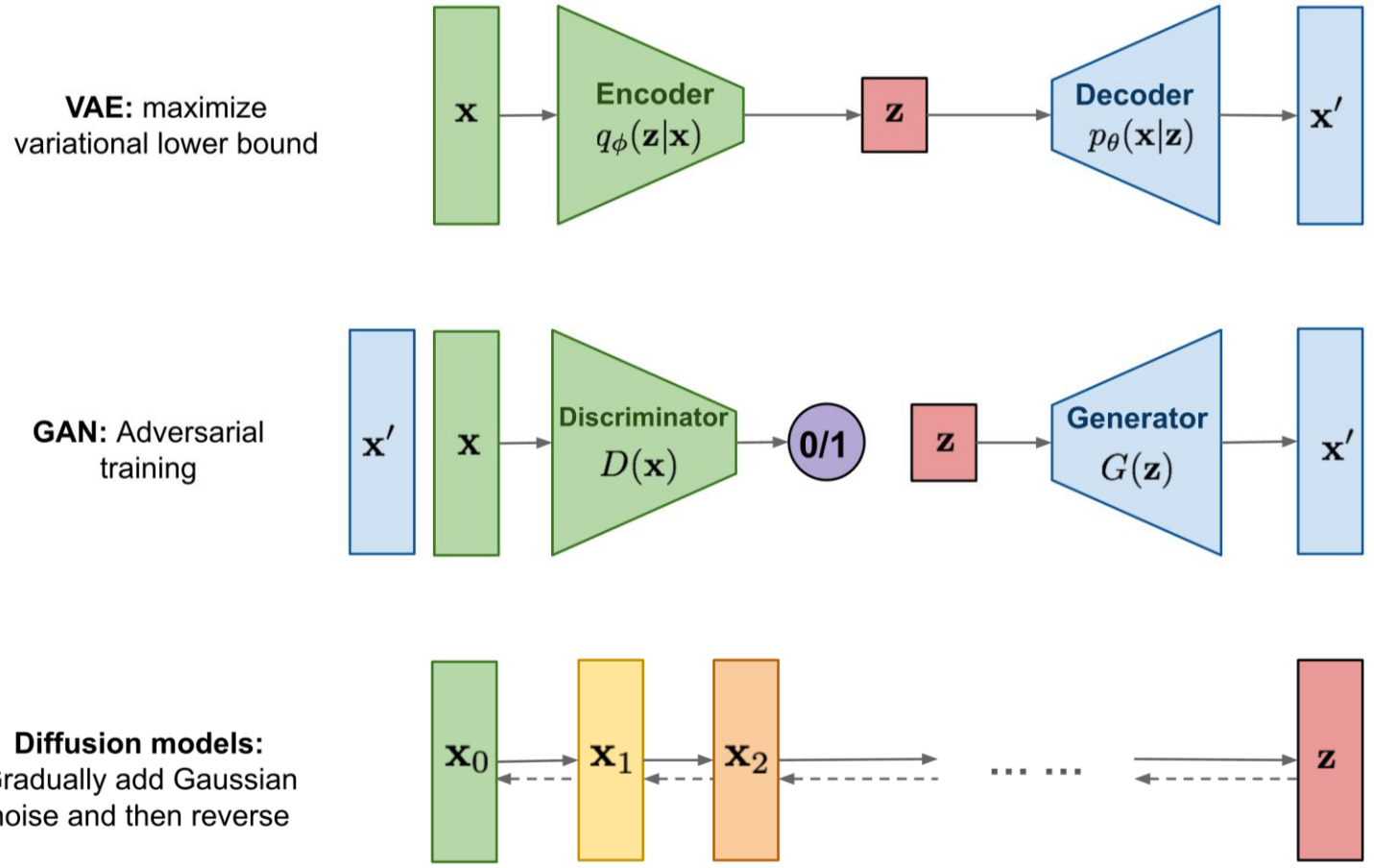
WaveGrad: Diffusion for Waveforms (Chen et al., 2021)



(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.

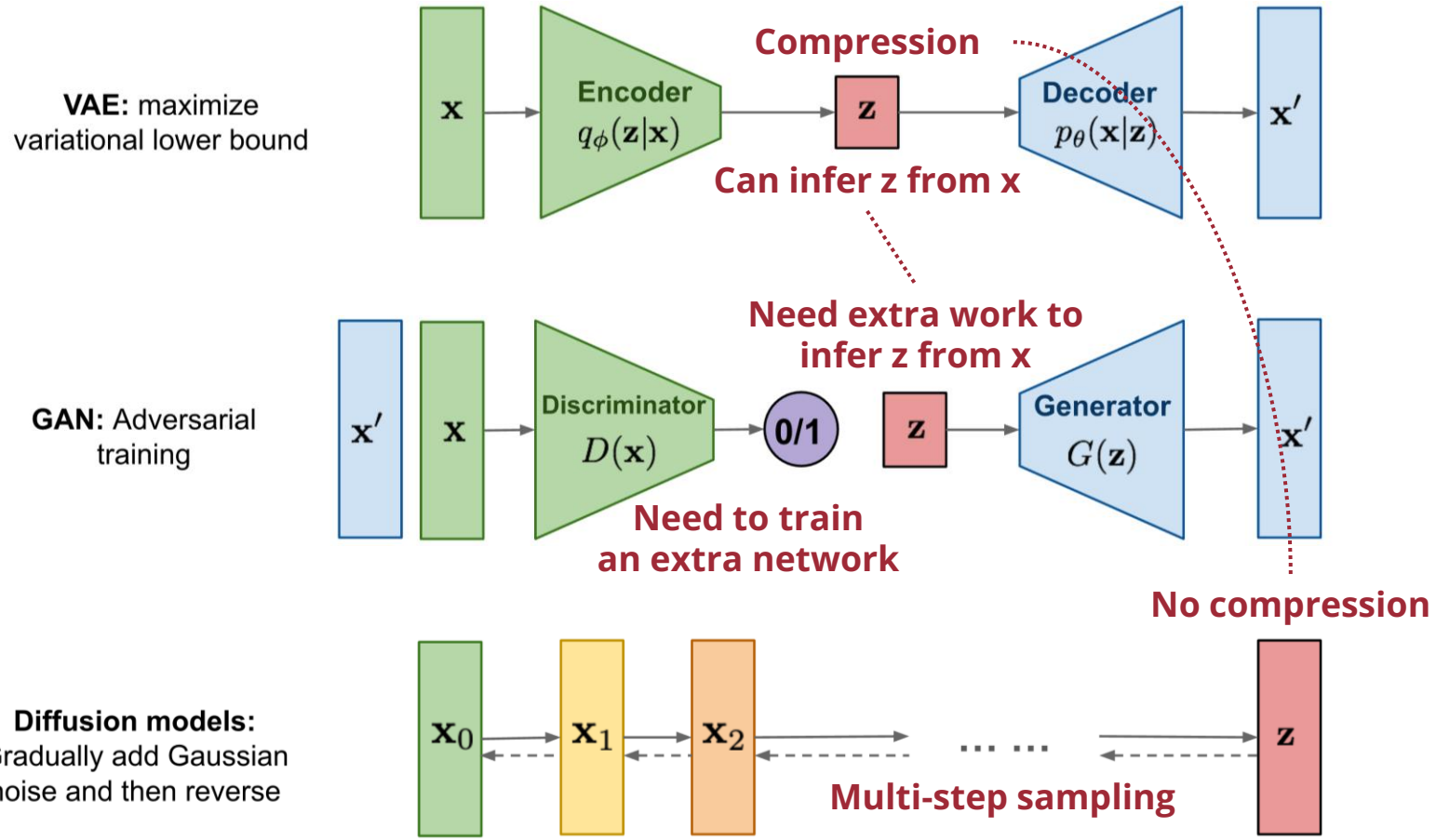
Comparison of Deep Latent Variable Models



(Source: Weng, 2021)

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

Comparison of Deep Latent Variable 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
⋮	⋮

Next Lecture

Latent Diffusion Models