PAT 498/598 (Fall 2024)

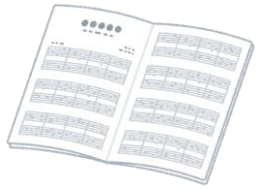
Special Topics: Generative AI for Music and Audio Creation

Lecture 15: Time-domain Audio Synthesis

Instructor: Hao-Wen Dong

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

(Recap) Four Paradigms



Symbolic music generation

- **Text-based**
  - MIDI

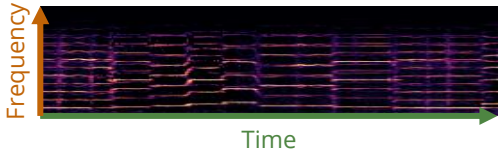
```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67,
...
```

- **Image-based**
  - Piano roll

Audio-domain music generation

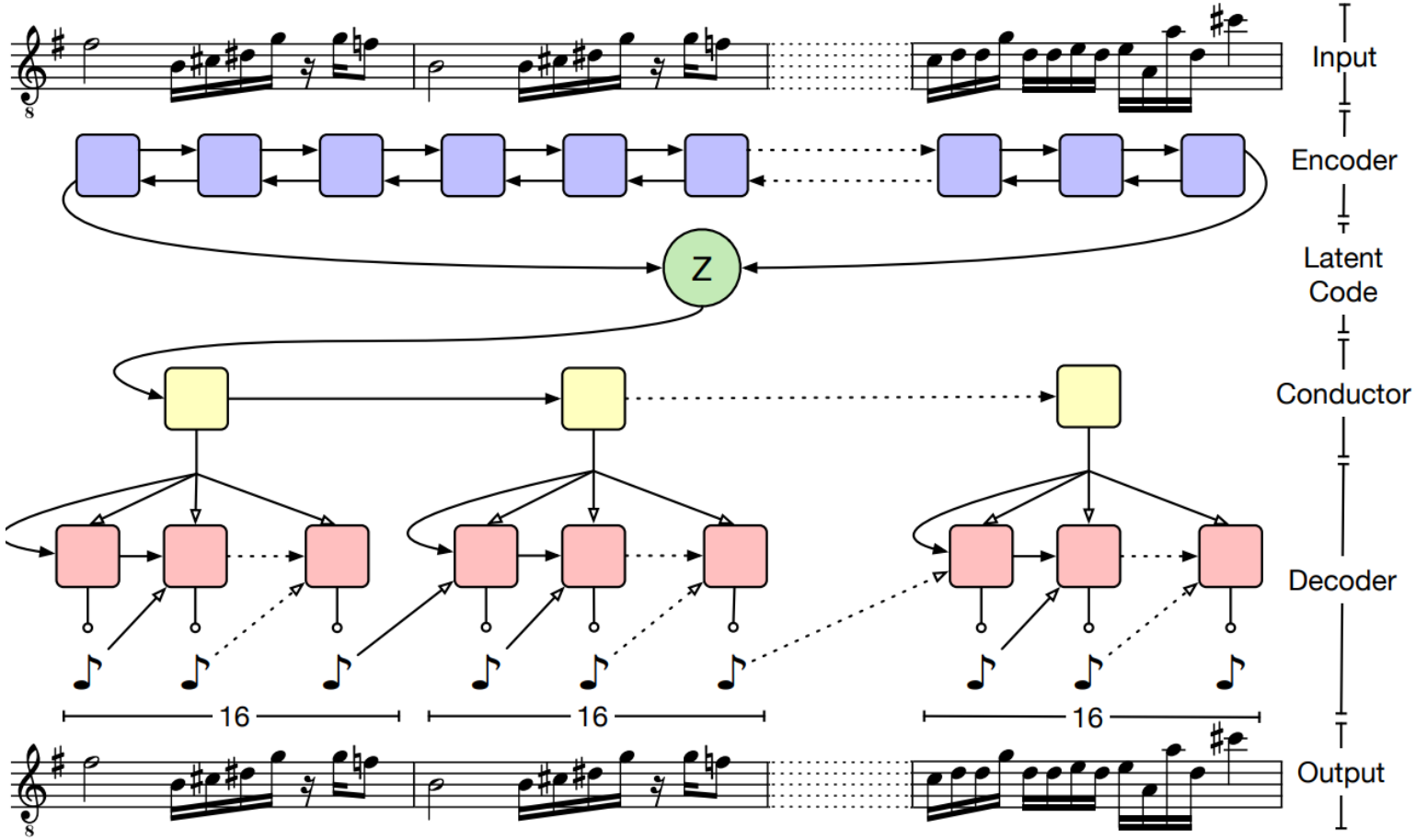
- **Time series-based**
  - Waveform

- **Image-based**
  - Spectrogram



Today, we also have many latent-space based systems!

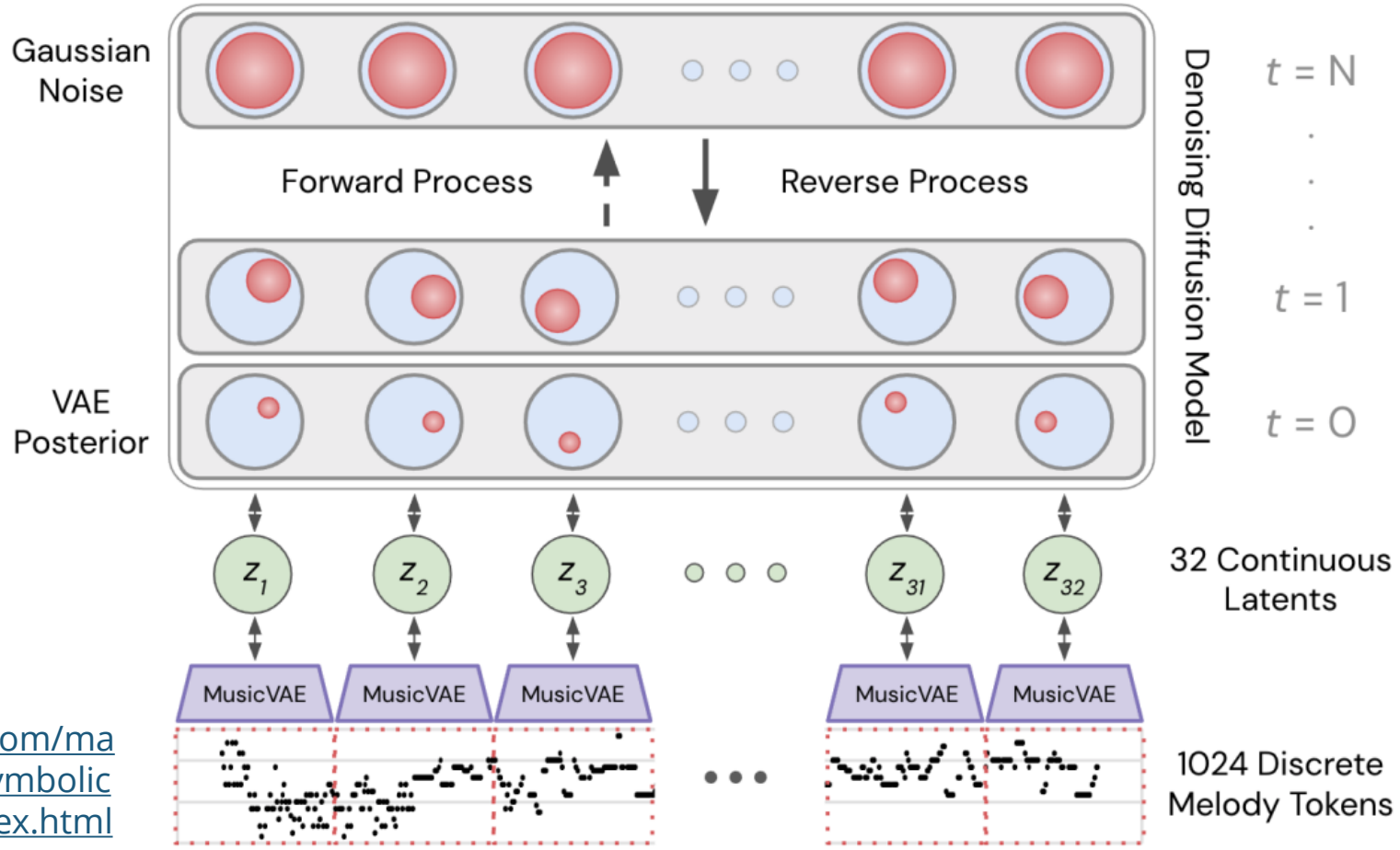
(Recap) Example: MusicVAE (Roberts et al., 2018)



(Source: Roberts et al., 2018)

Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck, "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music," *ICML*, 2018.

(Recap) Example: Latent Diffusion (Mittal et al., 2021)

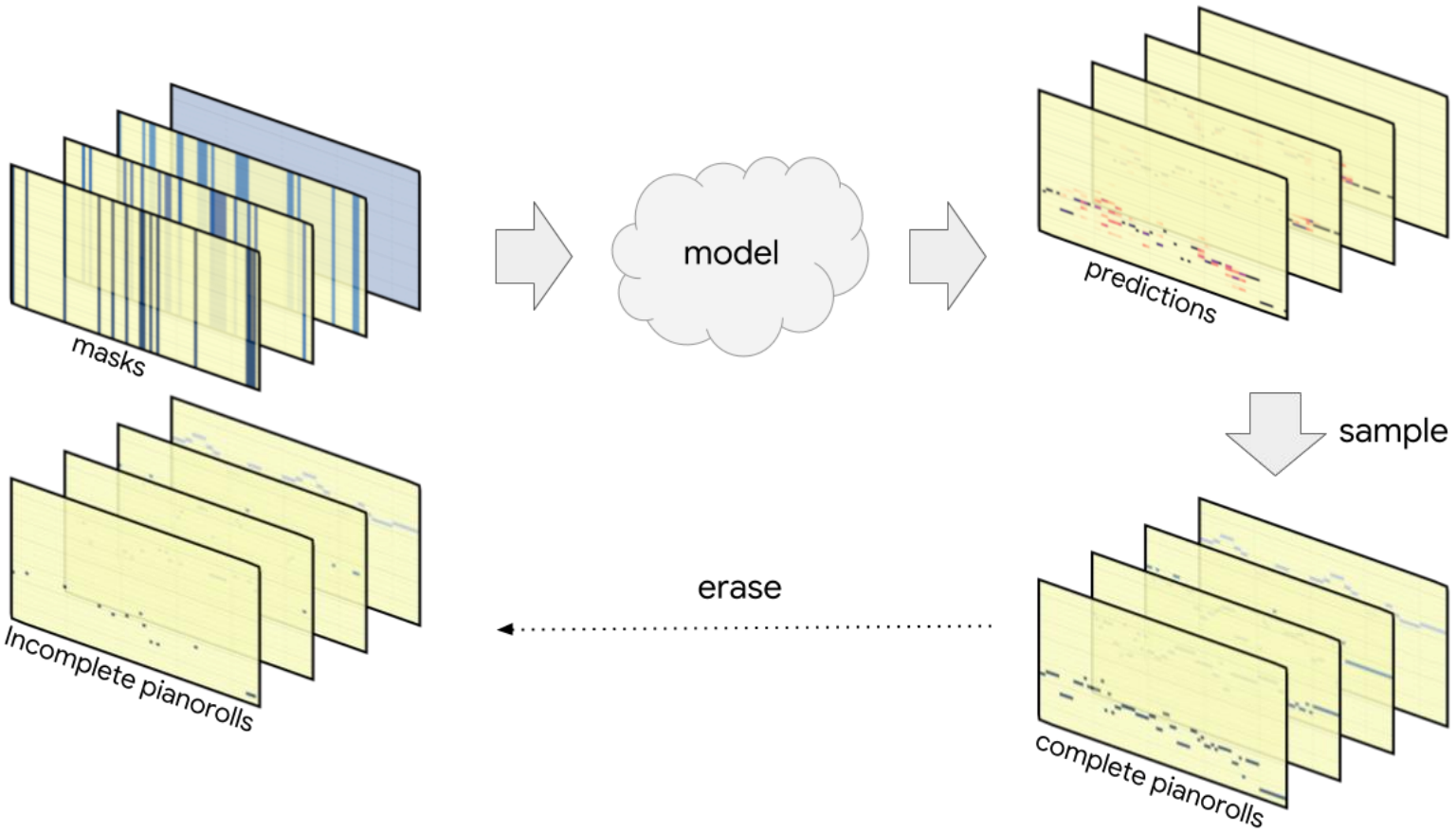


storage.googleapis.com/magentadata/papers/symbolic-music-diffusion/index.html

(Source: Mittal et al., 2021)

Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon, "Symbolic Music Generation with Diffusion Models," *ISMIR*, 2021.

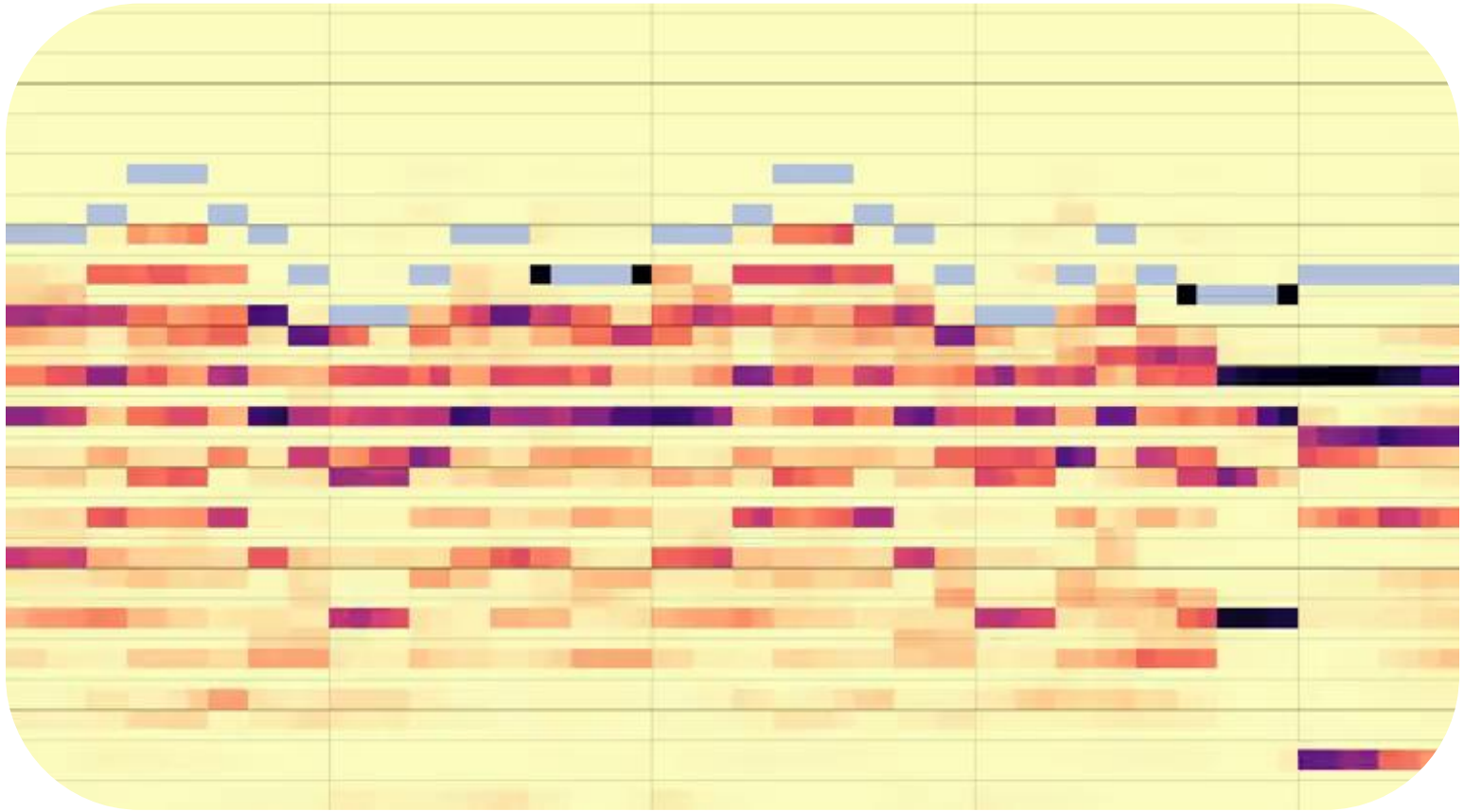
(Recap) Example: Coconet (Huang et al., 2017)



(Source: Huang et al., 2019)

Cheng-Zhi Anna Huang, Tim Cooijmans, Adam Roberts, Aaron Courville, and Douglas Eck, "Counterpoint by Convolution," *ISMIR*, 2017.
Cheng-Zhi Anna Huang, Tim Cooijmans, Monica Dinculescu, Adam Roberts, and Curtis Hawthorne, "Coconet: the ML model behind today's Bach Doodle," *Magenta Blog*, 2019.

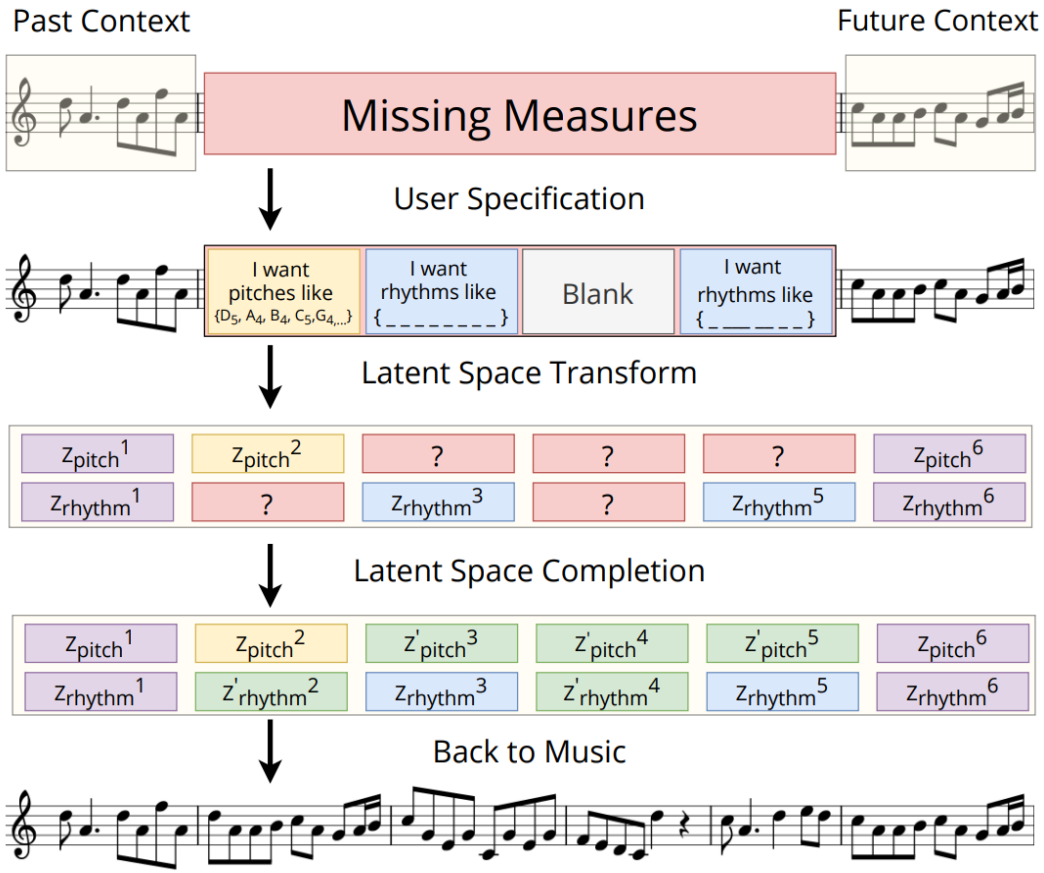
(Recap) Example: Coconet (Huang et al., 2017)

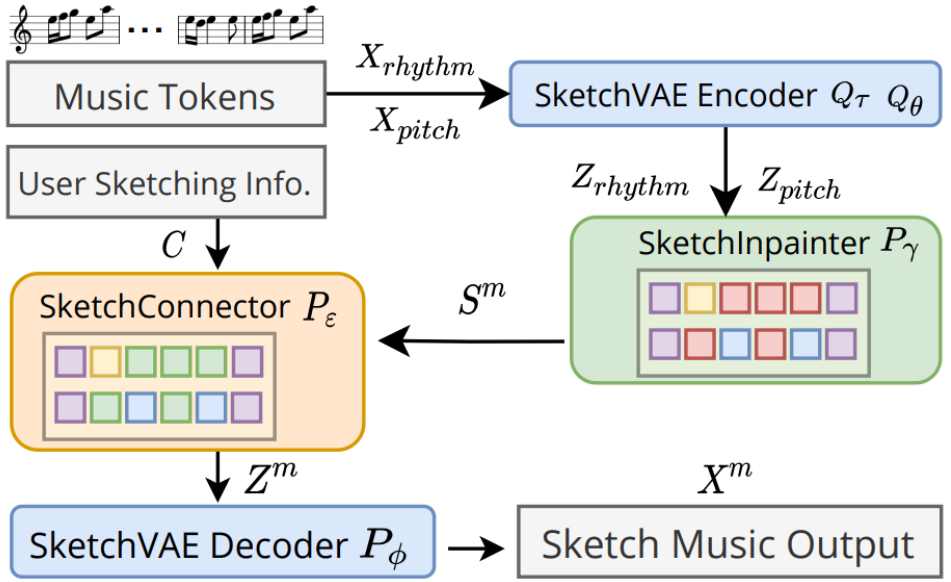


(Source: Huang et al., 2017)

Cheng-Zhi Anna Huang, Tim Cooijmans, Adam Roberts, Aaron Courville, and Douglas Eck, "Counterpoint by Convolution," *ISMIR*, 2017.

(Recap) Example: Music SketchNet (Chen et al., 2020)

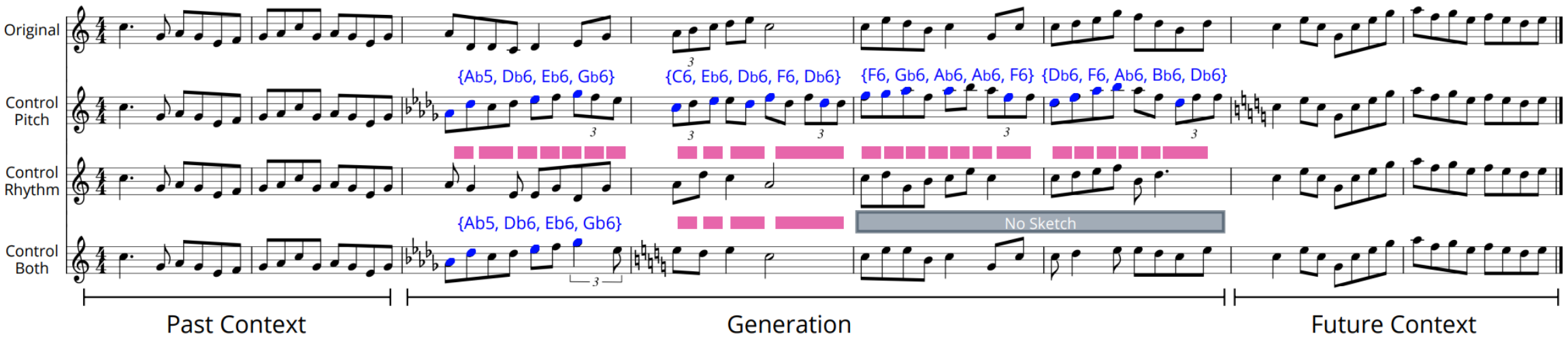




(Source: Chen et al., 2020)

Ke Chen, Cheng-i Wang, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling," *ISMIR*, 2020.

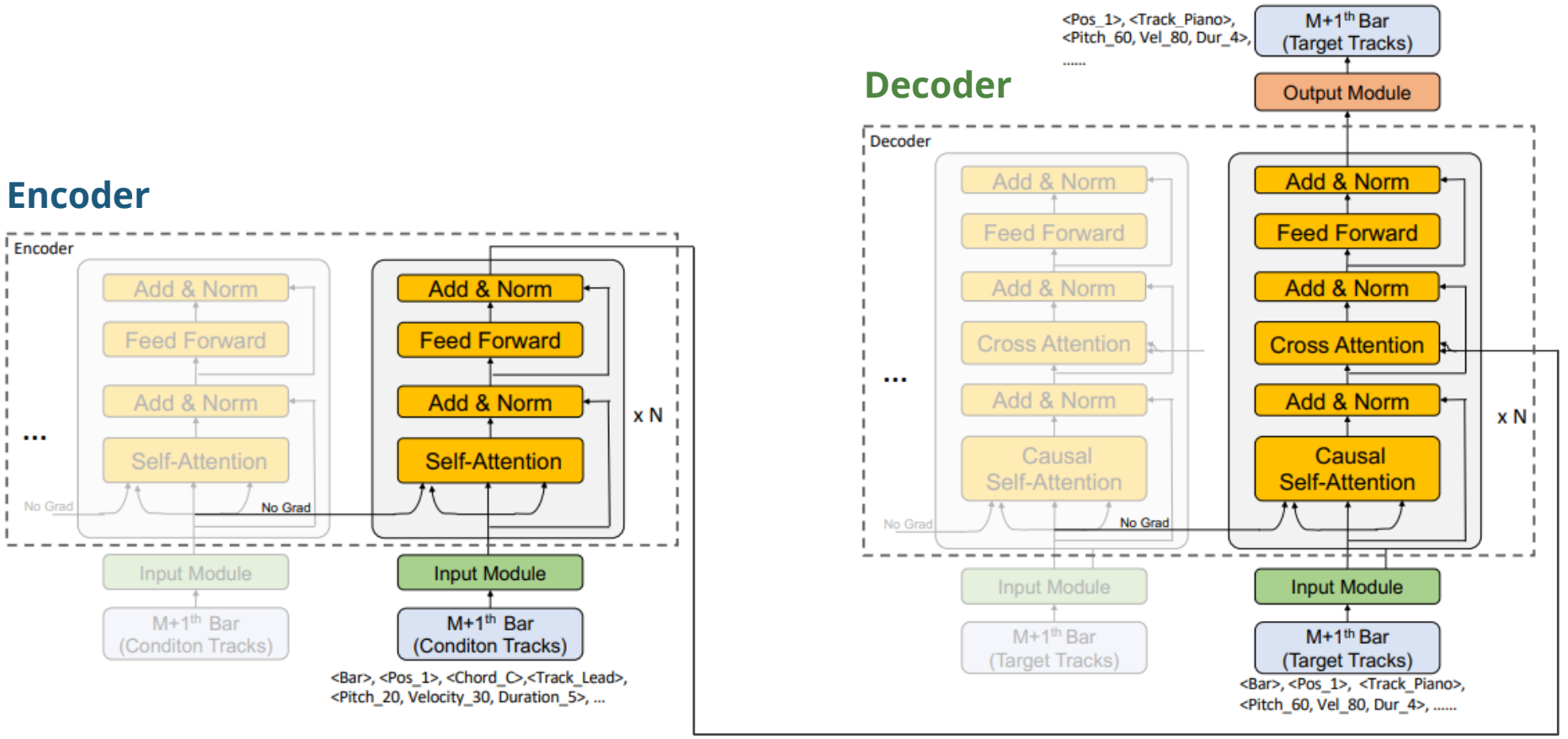
(Recap) Example: Music SketchNet (Chen et al., 2020)



(Source: Chen et al., 2020)

Ke Chen, Cheng-i Wang, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling," *ISMIR*, 2020.

(Recap) Example: PopMAG (Ren et al., 2020)

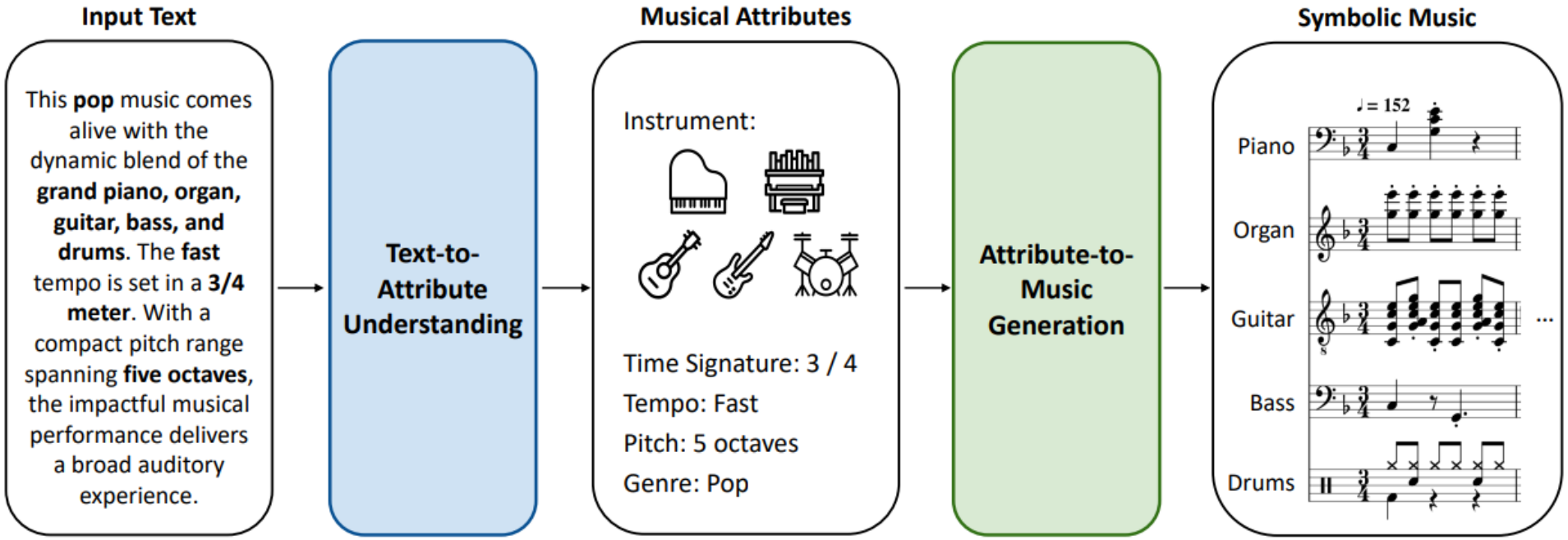


(Source: Ren et al., 2020)

ai-muzic.github.io/popmag

Yi Ren, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu, "PopMAG: Pop Music Accompaniment Generation," *MM*, 2020.

(Recap) Example: MuseCoco (Lu et al., 2023)

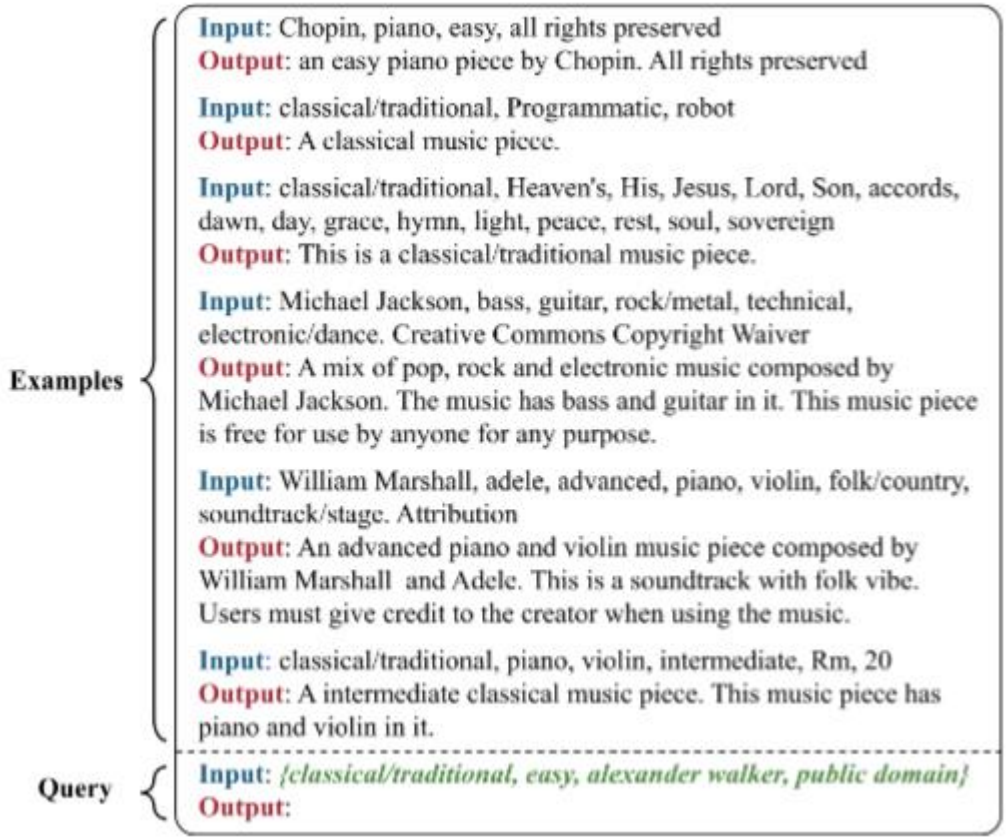


(Source: Lu et al., 2023)

ai-muzic.github.io/musecoco

Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, and Jiang Bian, "MuseCoco: Generating Symbolic Music from Text," *arXiv preprint arXiv:2306.00110*, 2023.

(Recap) Example: MetaScore Transformer (Xu et al., 2024)



(Source: Xu et al., 2024)

Prompt: This is an easy classical piano piece composed by wolfgang amadeus mozart.

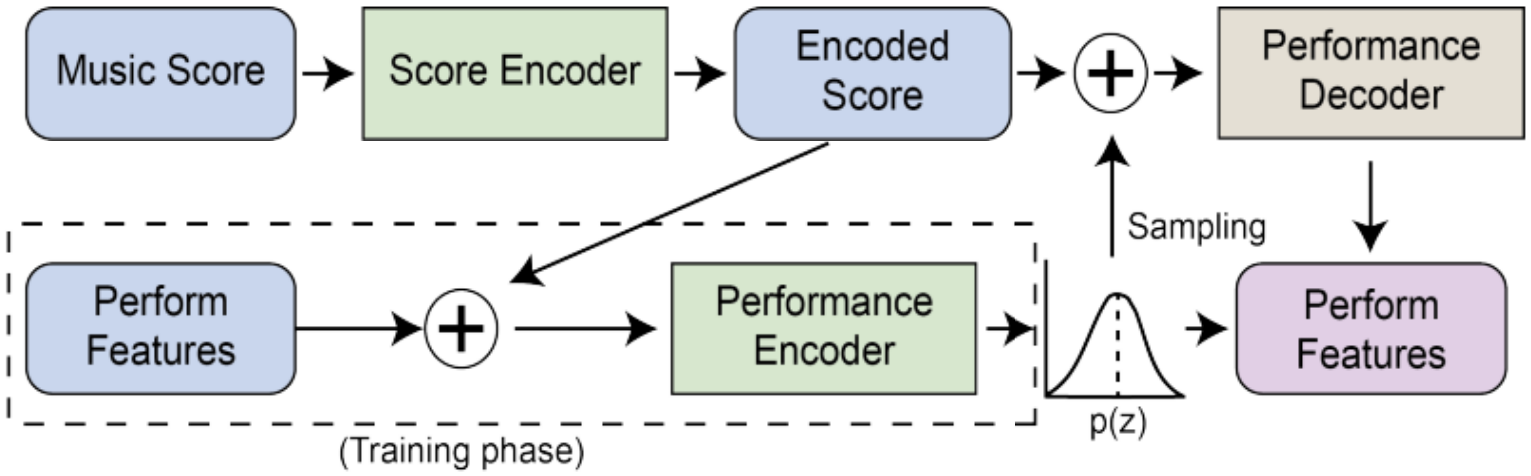
Prompt: A short and emotional music piece inspired by an anime scene.

Prompt: A powerful orchestral music piece.

Weihan Xu, Julian McAuley, Taylor Berg-Kirkpatrick, Shlomo Dubnov, and Hao-Wen Dong, "Generating Symbolic Music from Natural Language Prompts using an LLM-Enhanced Dataset," *arXiv preprint arXiv:2410.02084*, 2024.

(Recap) Example: VirtuosoNet (Jeong et al., 2019)

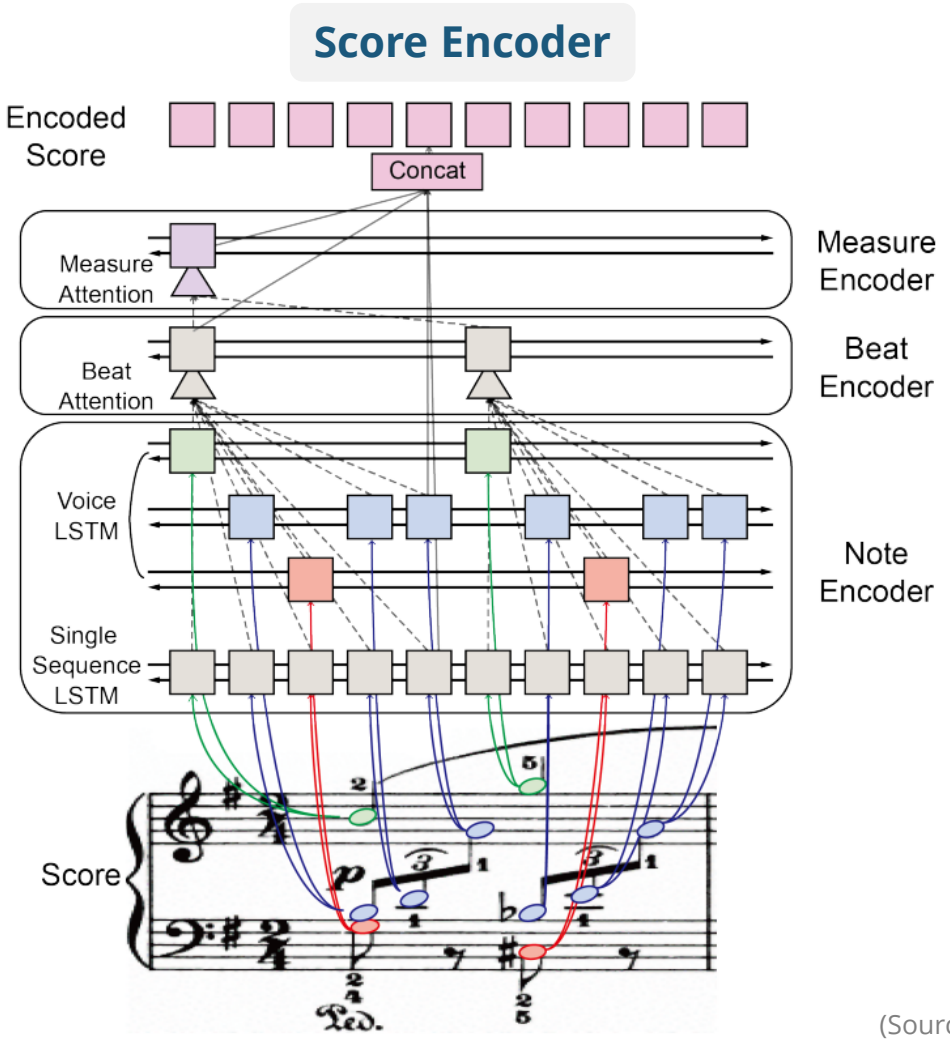
- **Input**: pitch, duration, articulation marking, slur and beam status, tempo marking, and dynamic marking, etc.
- **Output**: absolute tempo, velocity, onset deviation, articulation, pedal usages

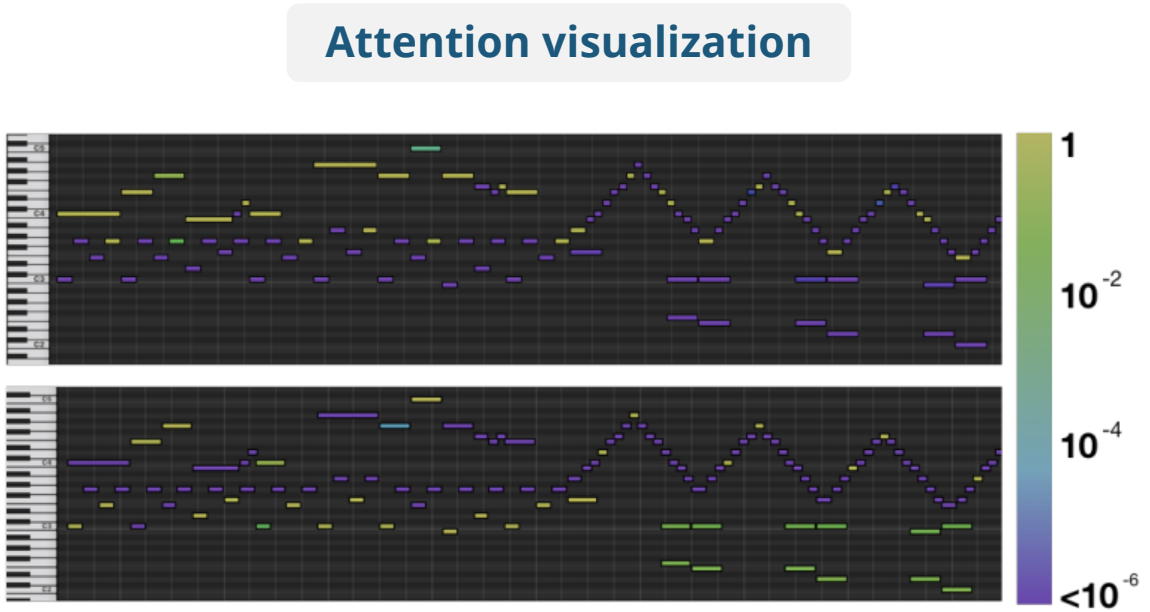


(Source: Jeong et al., 2019)

Dasaem Jeong, Taegyun Kwon, Yoojin Kim, Kyogu Lee, and Juhan Nam, "VirtuosoNet: A Hierarchical RNN-based system for modeling expressive piano performance," *ISMIR*, 2019.

(Recap) Example: VirtuosoNet (Jeong et al., 2019)

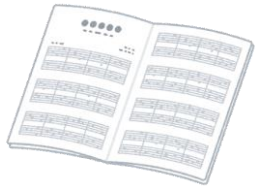




(Source: Jeong et al., 2019)

Dasaem Jeong, Taegyun Kwon, Yoojin Kim, Kyogu Lee, and Juhan Nam, "VirtuosoNet: A Hierarchical RNN-based system for modeling expressive piano performance," *ISMIR*, 2019.

(Recap) Four Paradigms



Symbolic music generation

Text-based

```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67,
...
```

MIDI

Image-based

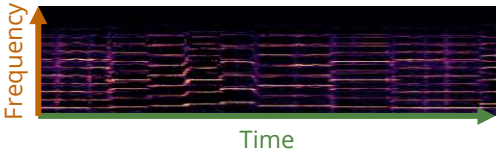
Piano roll

Audio-domain music generation

Time series-based

Waveform

Image-based

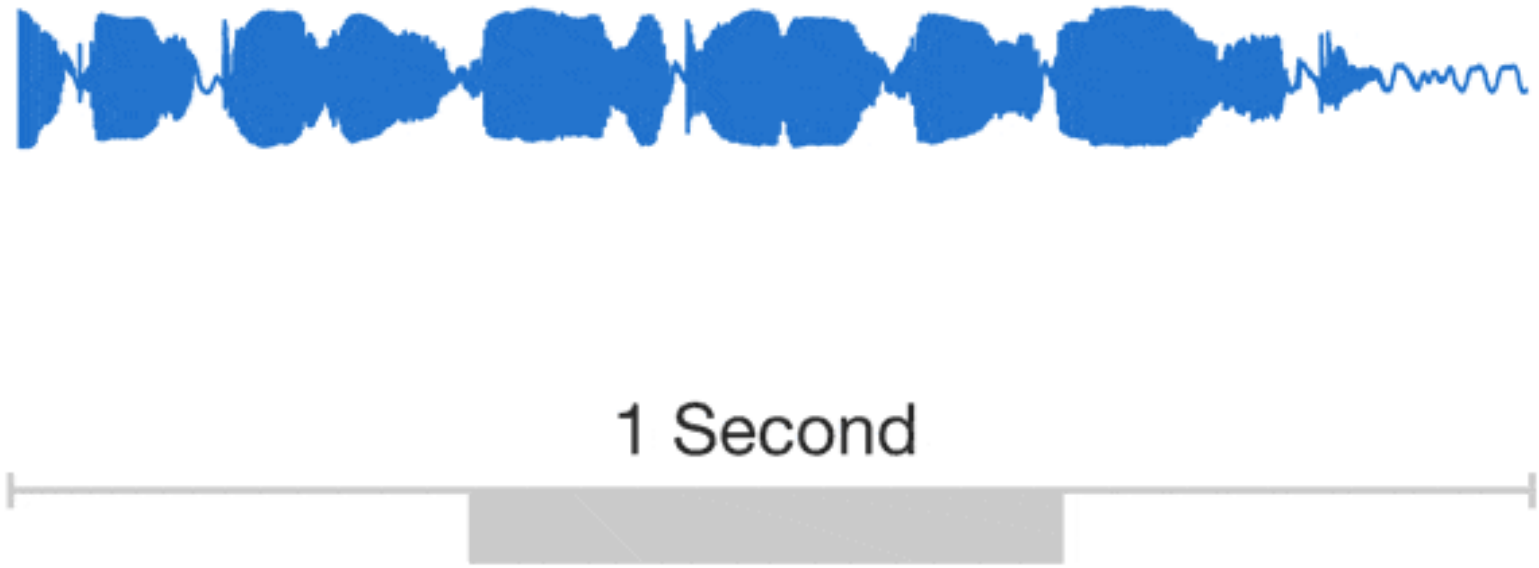


Spectrogram

Today, we also have many latent-space based systems!

Autoregressive Waveform Synthesis

Generating Waveforms using a Neural Network

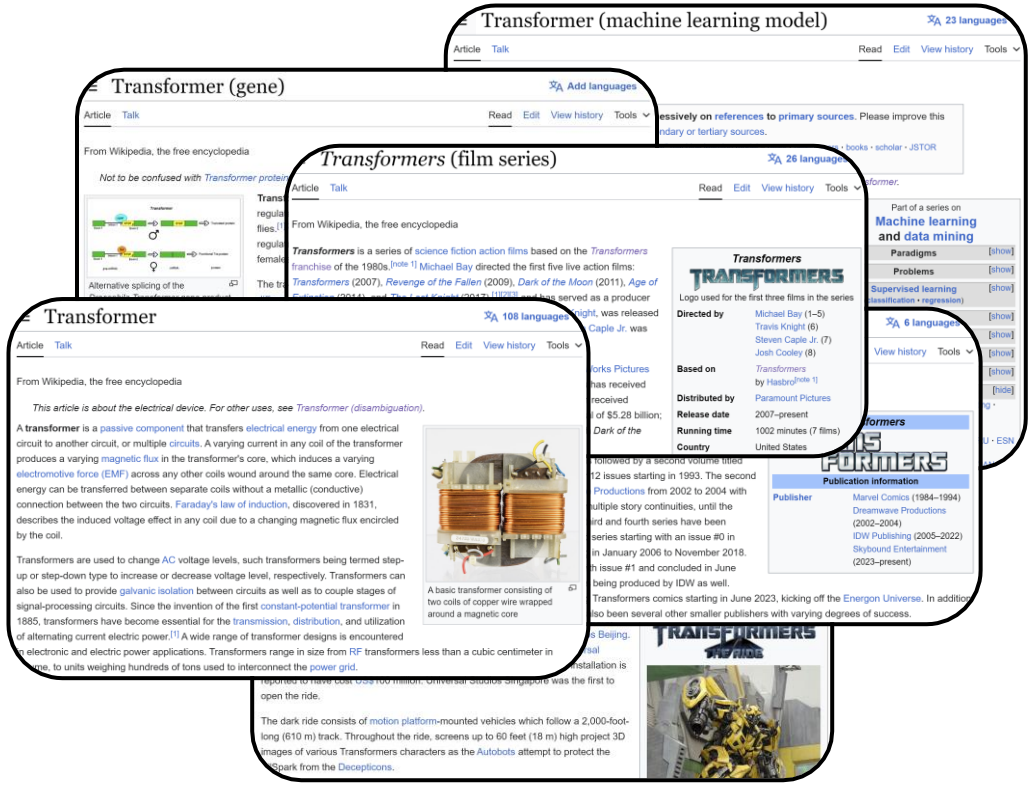


(Source: van den Oord et al., 2016)

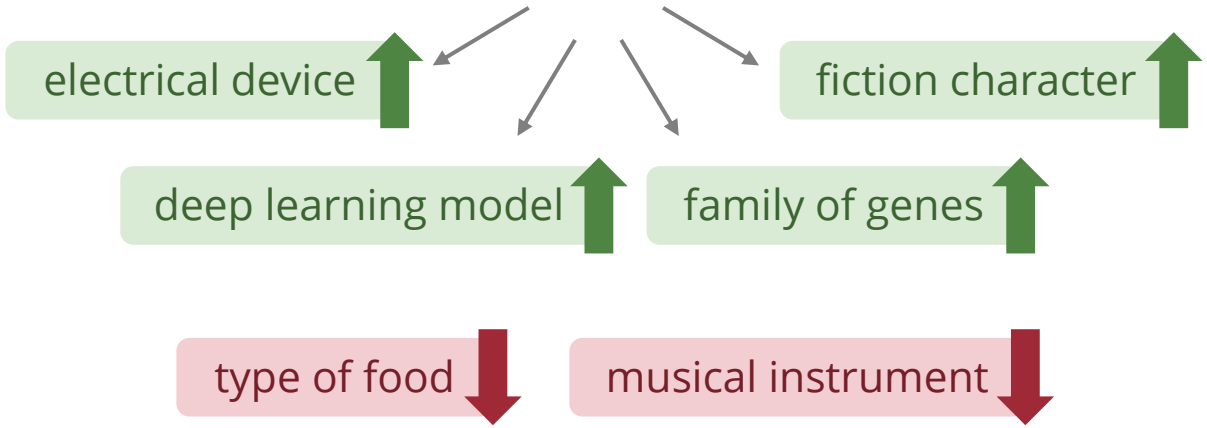
Aaron van den Oord and Sander Dieleman, "WaveNet: A generative model for raw audio," *DeepMind Blog*, September 8, 2016.

(Recap) Language Models

- Predicting the next word **given the past sequence of words**



A transformer is a _____



(Recap) Language Models (Mathematically)

- A class of machine learning models that learn the next word probability

$$P(\ x_i \mid \underbrace{x_1, x_2, \ldots, x_{i-1}}\ )$$

Next word — Previous words

$P($ electrical $|$ A transformer is a $)$ ↑

$P($ character $|$ A transformer is a $)$ ↑

$P($ gene $|$ A transformer is a $)$ ↑

$P($ model $|$ A transformer is a $)$ ↑

$P($ food $|$ A transformer is a $)$ ↓

$P($ musical $|$ A transformer is a $)$ ↓

Autoregressive Models (Mathematically)

- A class of machine learning models that learn the probability of the next value given previous values

$$P(\ x_i \mid \underbrace{x_1, x_2, \ldots, x_{i-1}}\ )$$

Next number Previous numbers

$$P(\ 0.1 \mid 0.5, 0.4, 0.3, 0.2)$$ ↑

$$P(\ 0.09 \mid 0.5, 0.4, 0.3, 0.2)$$ ↑

$$P(\ 0.11 \mid 0.5, 0.4, 0.3, 0.2)$$ ↑

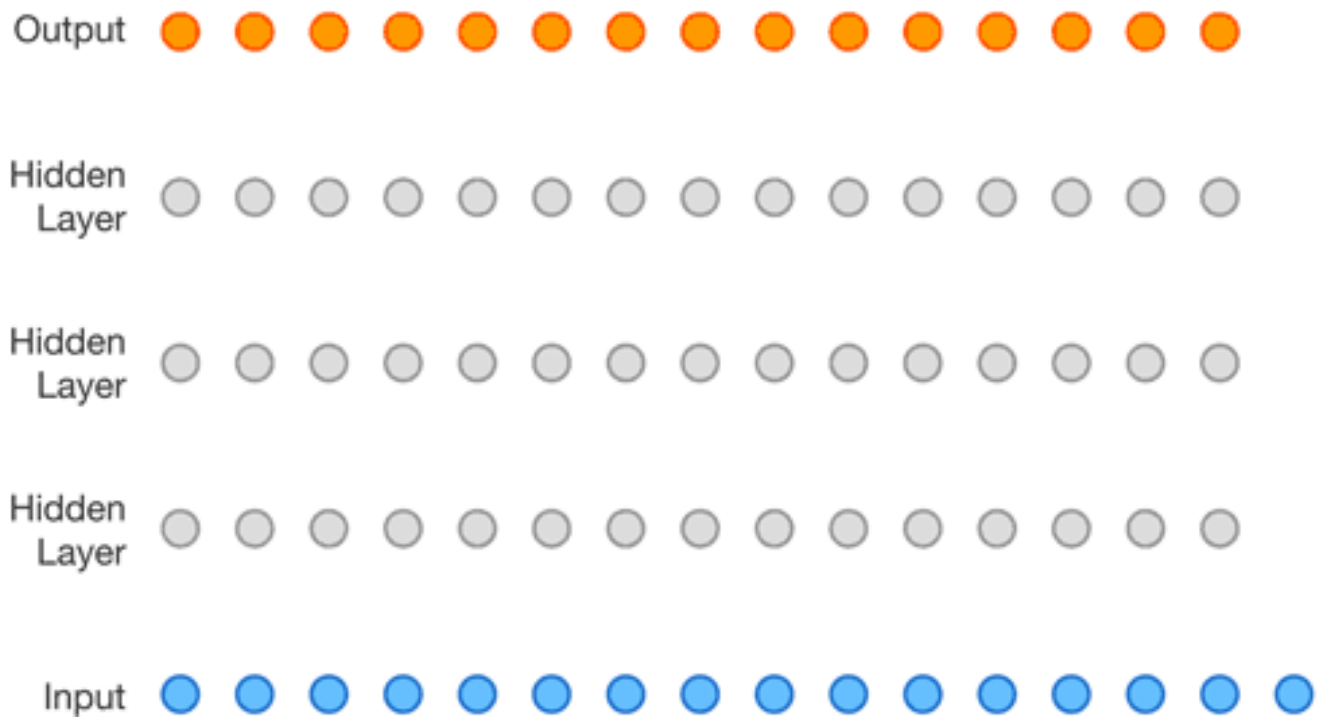
$$P(\ 99 \mid 0.5, 0.4, 0.3, 0.2)$$ ↓

$$P(\ -1 \mid 0.5, 0.4, 0.3, 0.2)$$ ↓

The term "autoregressive" has different definitions in machine learning and signal processing.
In signal processing, an autoregressive model needs to be a linear model.

Unconditional Audio Synthesis using CNNs

Example: WaveNet (van den Oord et al., 2016)



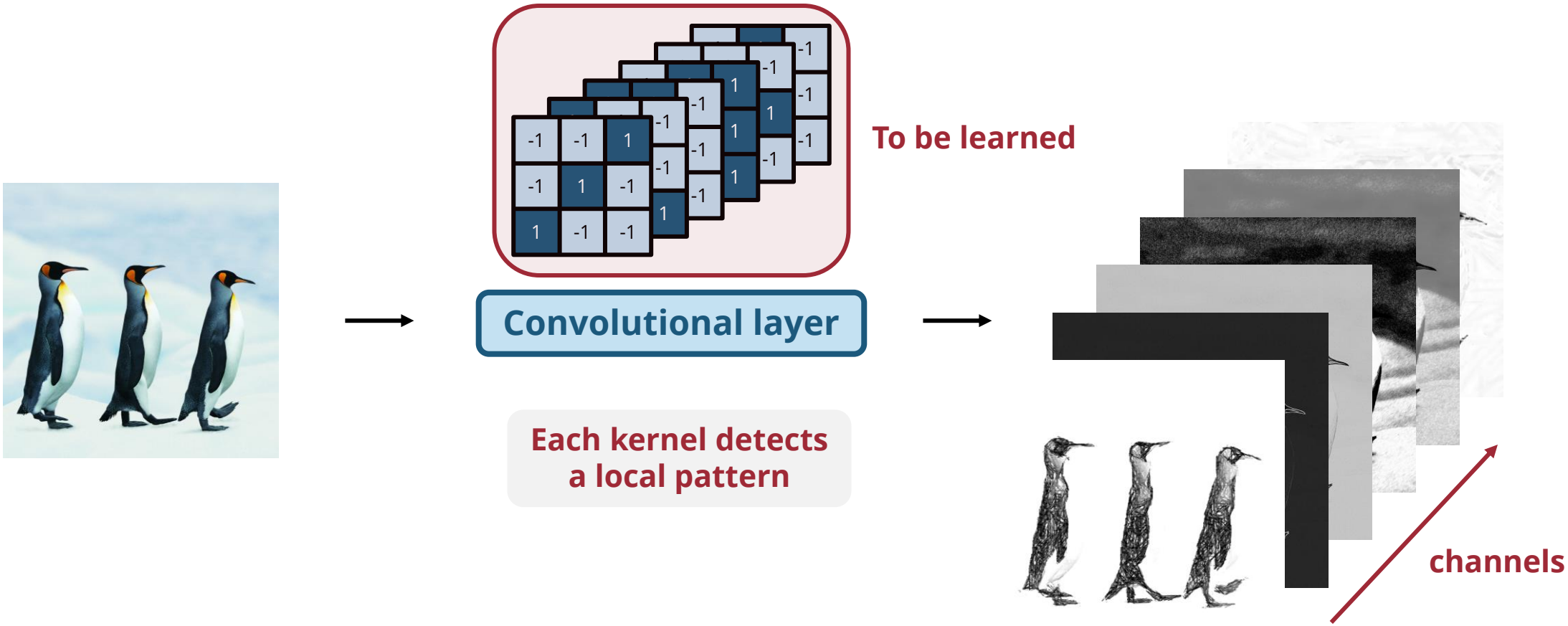
(Source: van den Oord et al., 2016)

A convolutional neural network for raw waveform generation

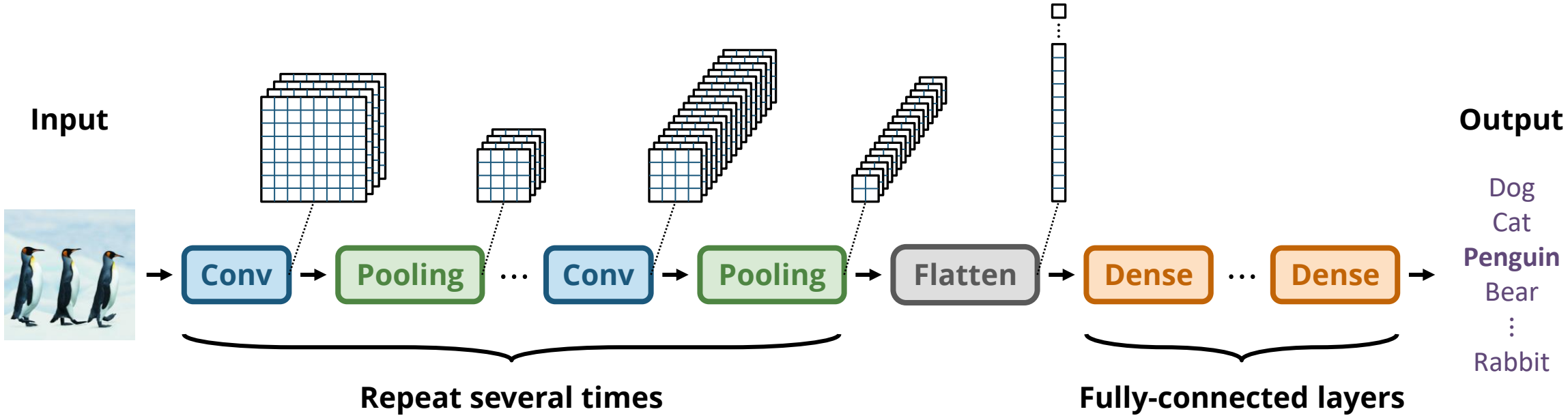
Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *ICML*, 2016.

(Recap) Convolutional Layer

- A convolutional layer consists of many **learnable kernels** (channels)

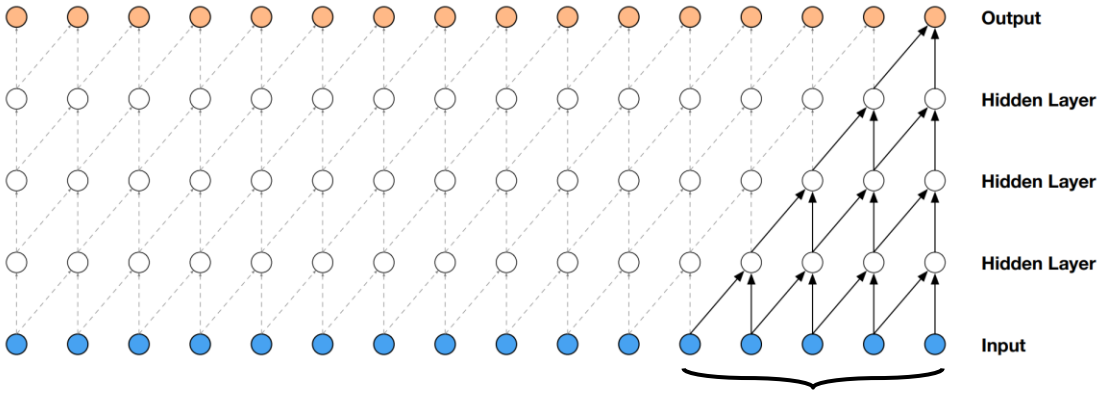


(Recap) Convolutional Neural Network (CNNs)



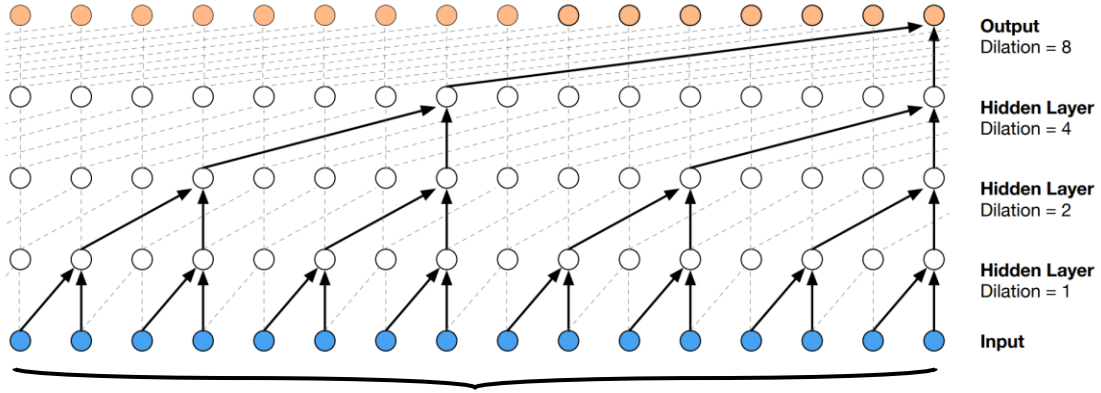
Example: WaveNet (van den Oord et al., 2016)

Standard convolution



Smaller "receptive field"

Dilated convolution



Larger "receptive field"

deepmind.google/discover/blog/wavenet-a-generative-model-for-raw-audio

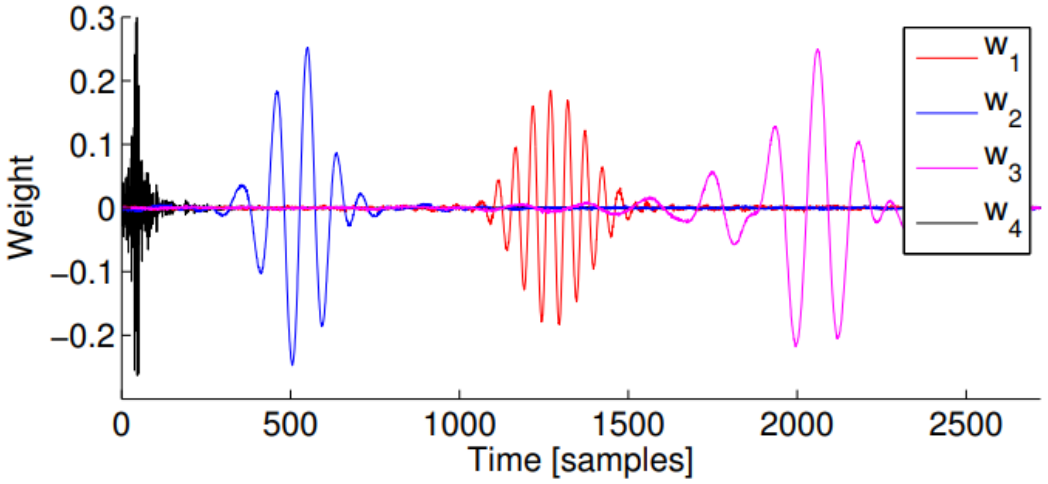
Example of generated music

(Source: van den Oord et al., 2016)

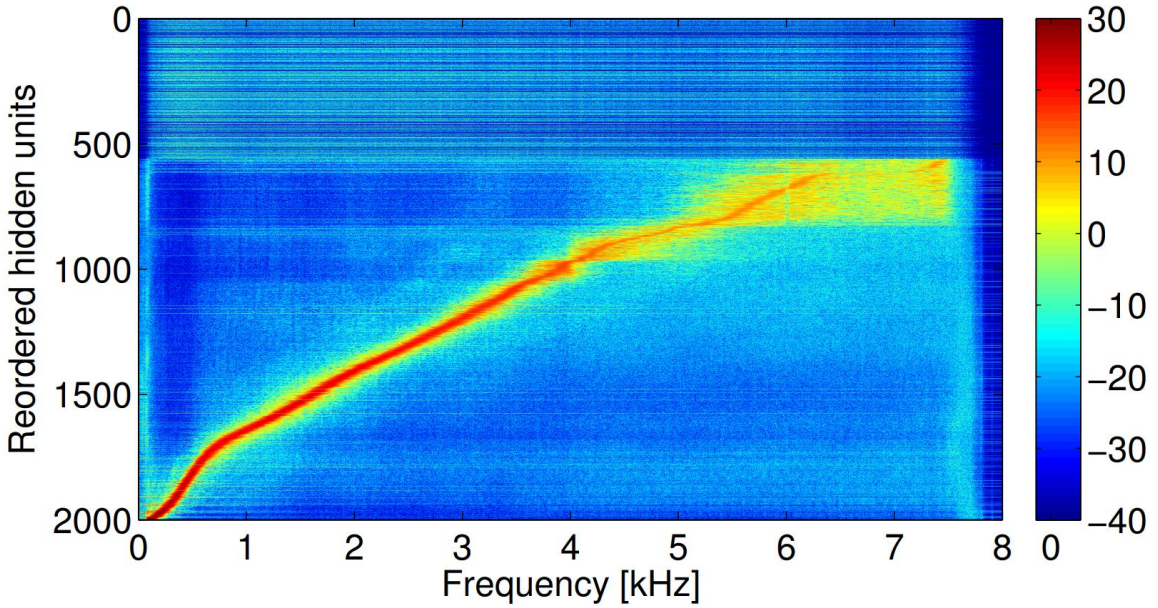
Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *ICML*, 2016.

1D CNNs & Fourier Transform

Convolution kernels learned



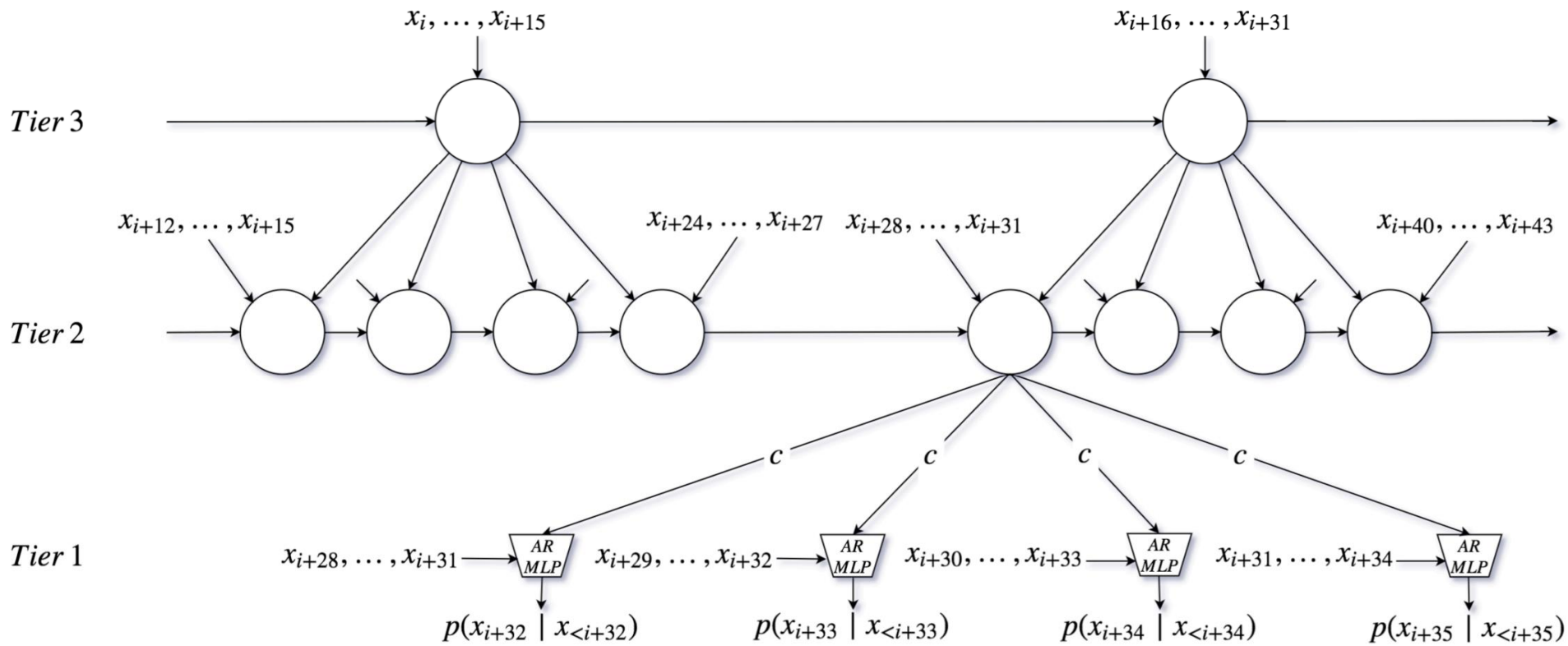
Peak frequency detected by the learned kernels



Pavel Golik, Zoltán Tüske, Ralf Schlüter, and Hermann Ney, "Convolutional Neural Networks for Acoustic Modeling of Raw Time Signal in LVCSR," *INTERSPEECH*, 2015.

Unconditional Audio Synthesis using RNNs

Example: SampleRNN (Mehri et al., 2017)

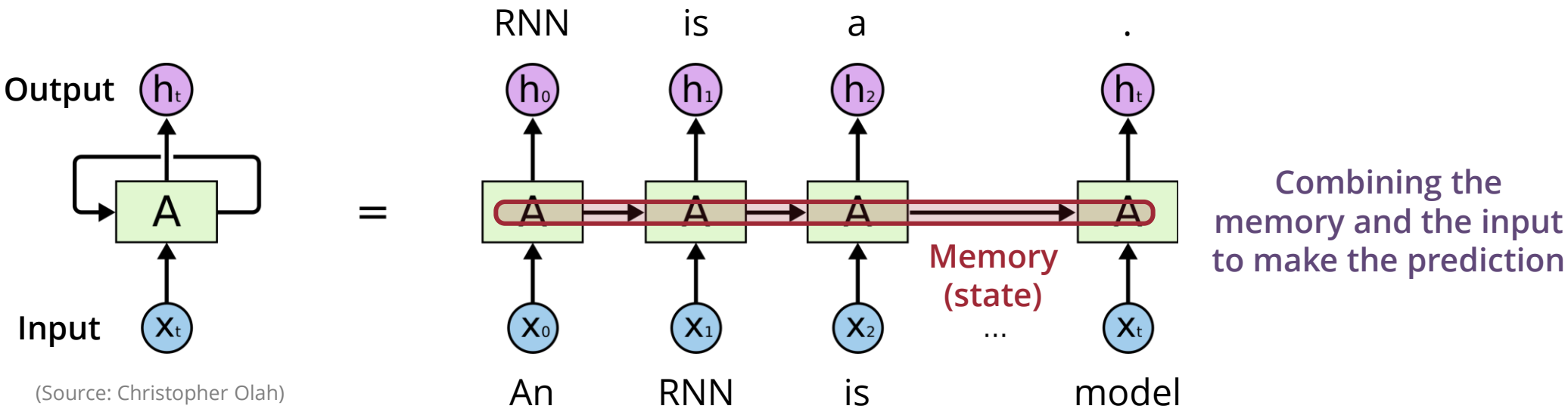


(Source: Mehri et al., 2017)

Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio, "SampleRNN: An Unconditional End-to-End Neural Audio Generation Model," *ICLR*, 2017.

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

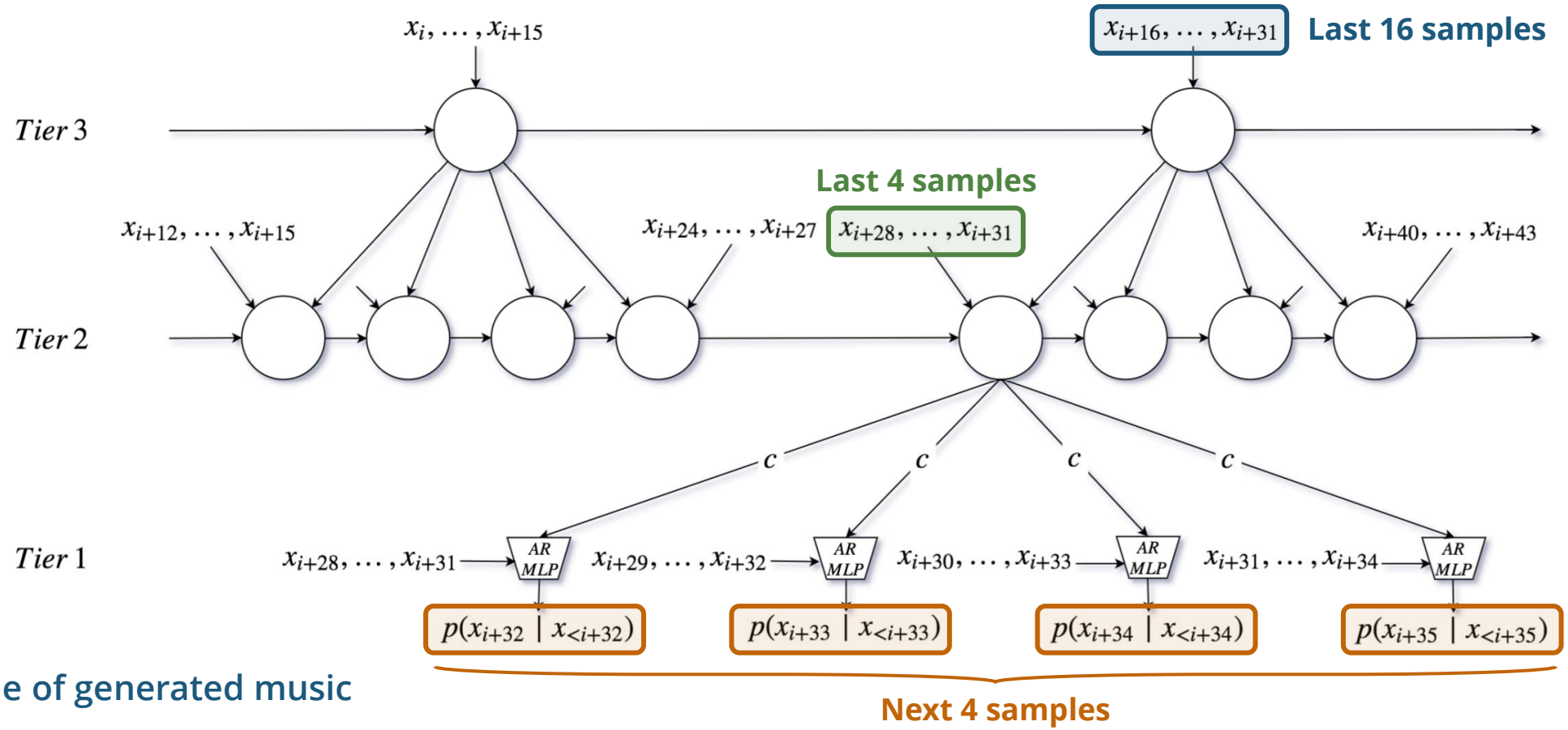
- A type of neural networks that have **loops**
- Widely used for **modeling sequences** (e.g., in natural language processing)



(Source: Christopher Olah)

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

Example: SampleRNN (Mehri et al., 2017)



Example of generated music

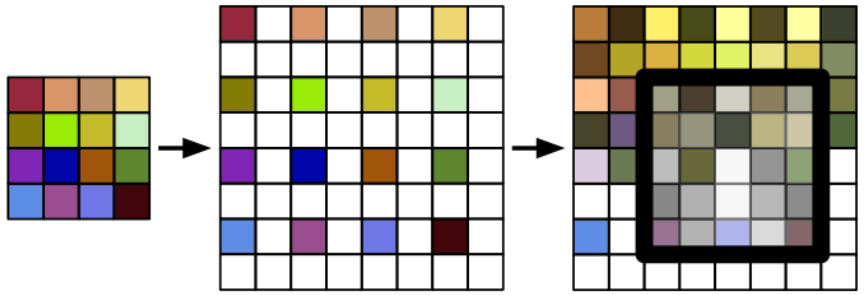
(Source: Mehri et al., 2017)

Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio, "SampleRNN: An Unconditional End-to-End Neural Audio Generation Model," *ICLR*, 2017.

Unconditional Audio Synthesis using GANs

Example: WaveGAN (Donahue et al., 2019)

DCGAN for images

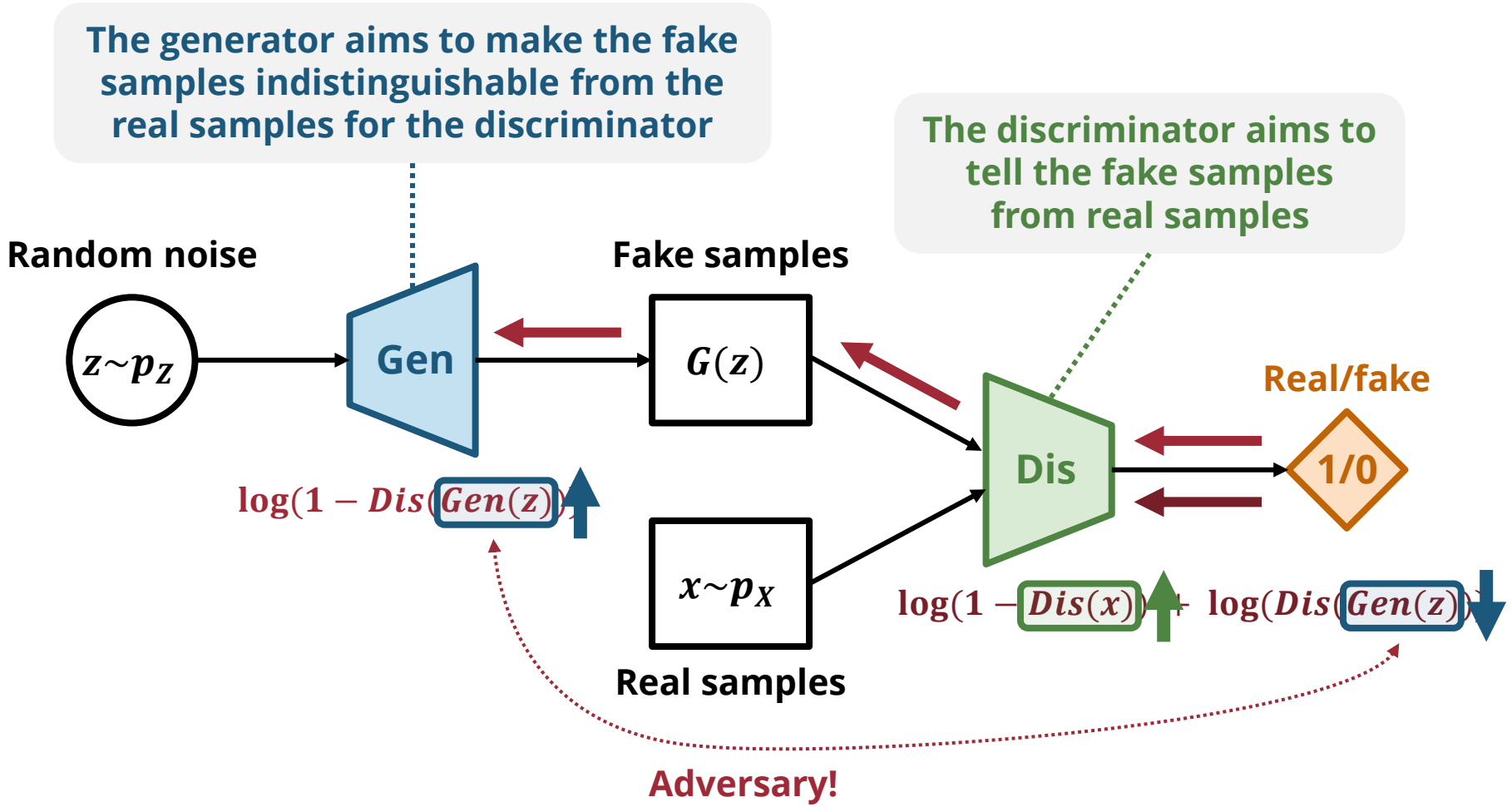


5x5 kernels

(Source: Donahue et al., 2019)

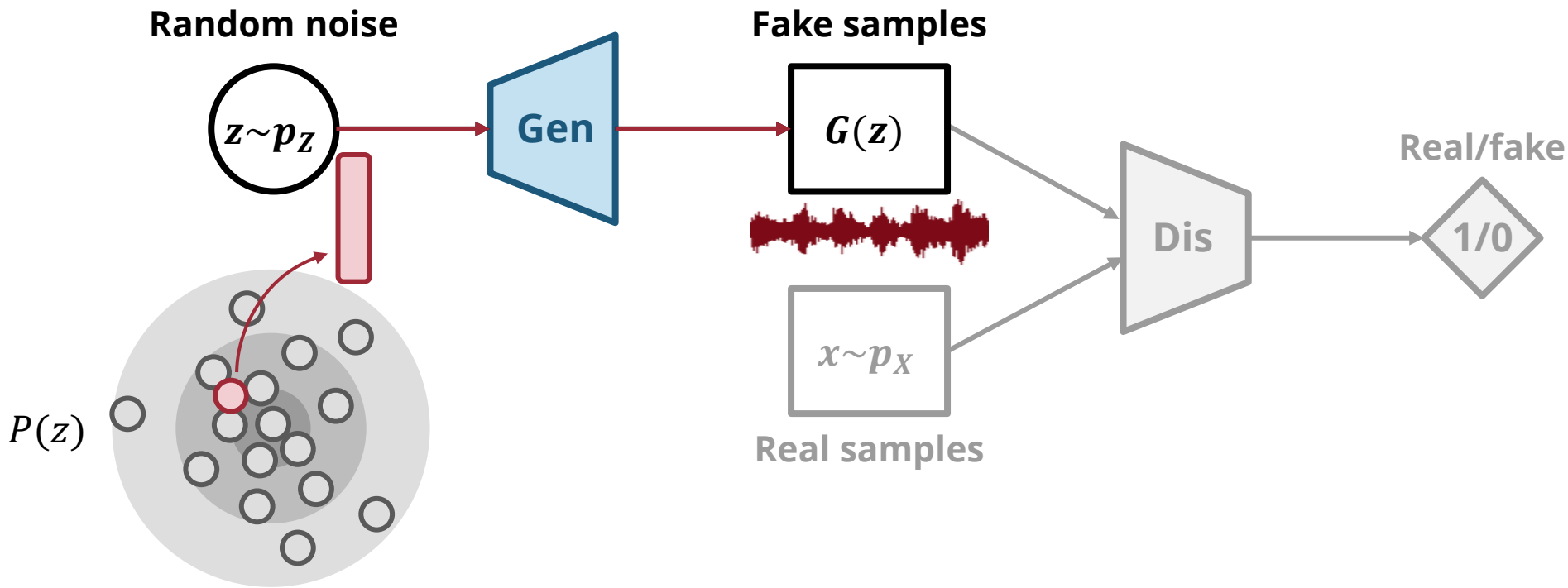
Chris Donahue, Julian McAuley, and Miller Puckette, "Adversarial Audio Synthesis," *ICLR*, 2019.

(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.

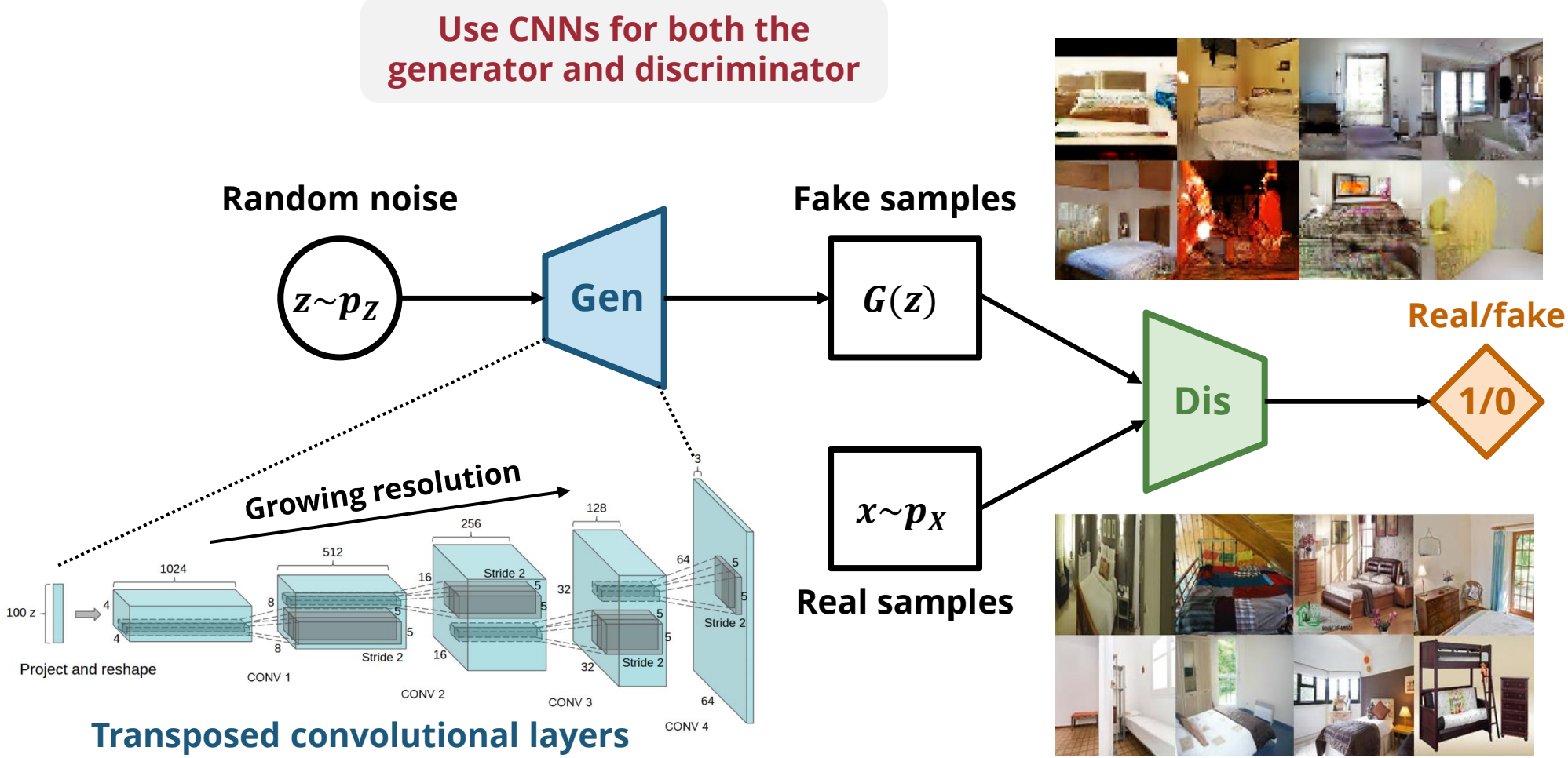
(Recap) Generative Adversarial Nets (GANs) – Generation



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) Deep Convolutional GANs (DCGANs)

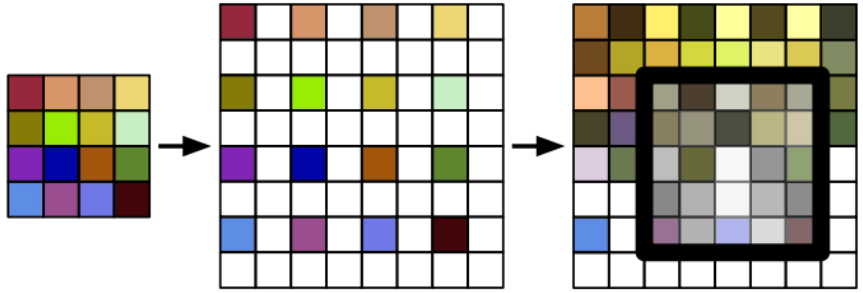
Use CNNs for both the generator and discriminator



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

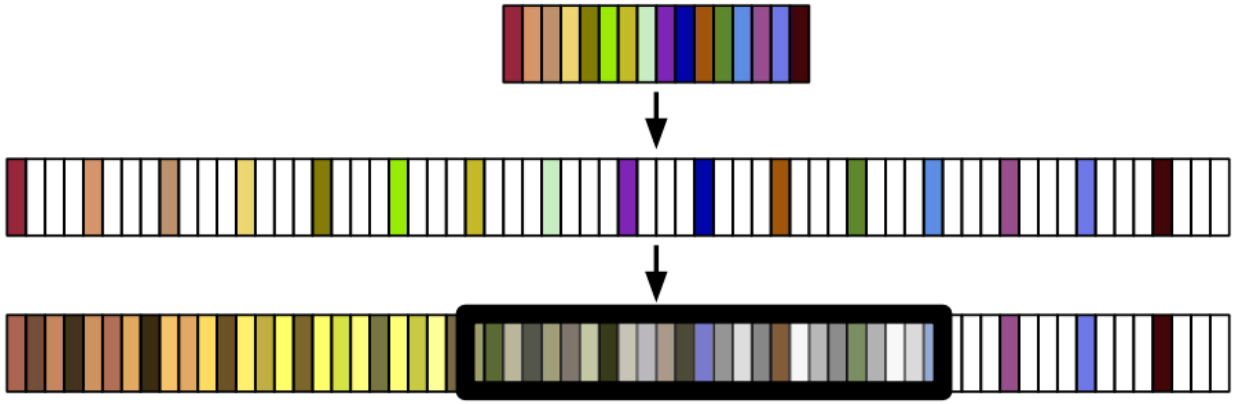
Example: WaveGAN (Donahue et al., 2019)

DCGAN for images



5x5 kernels

WaveGAN for audio



25x1 kernels

(Source: Donahue et al., 2019)

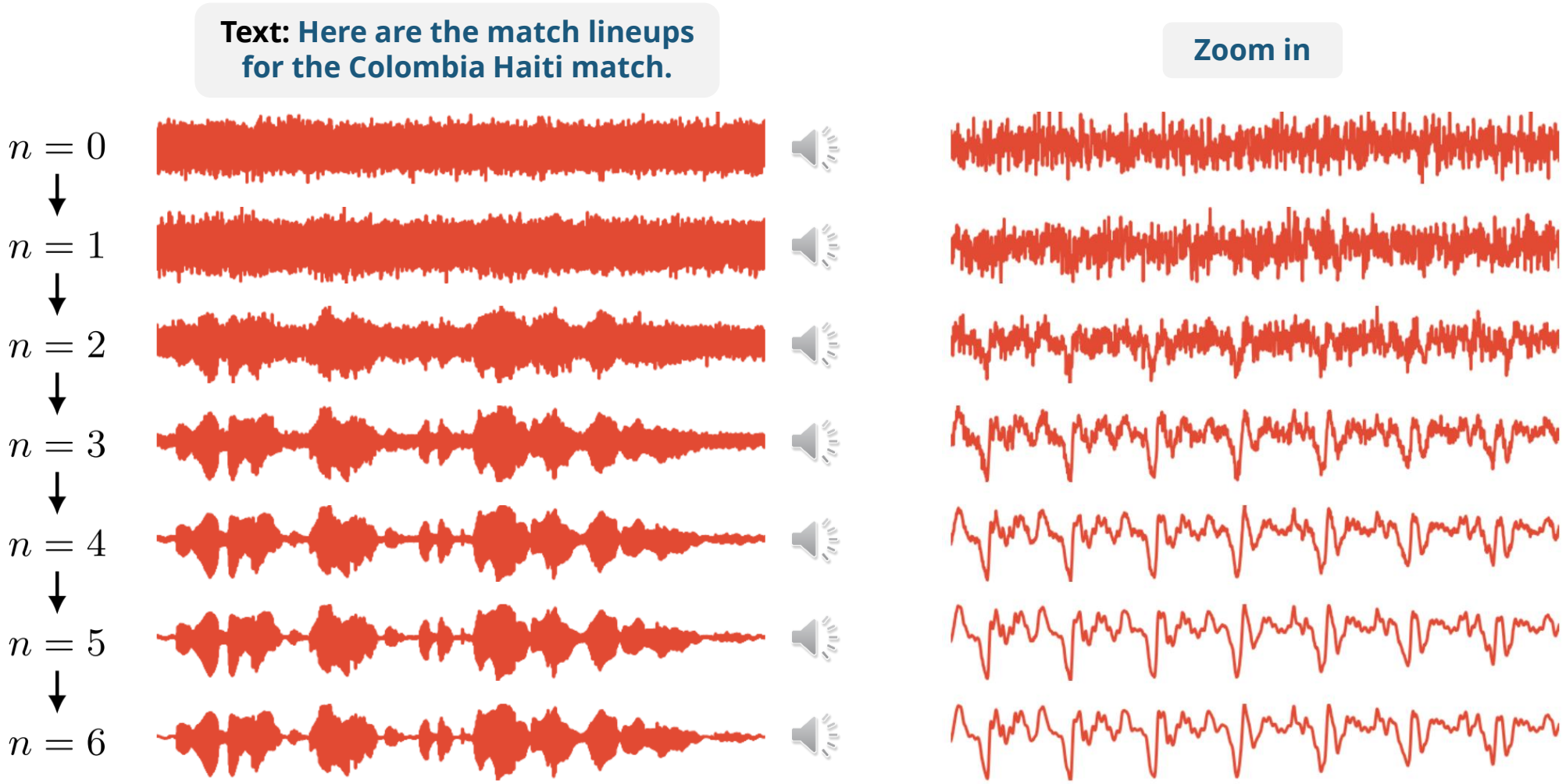
chrisdonahue.com/wavegan_examples
chrisdonahue.com/wavegan

Example of generated music

Chris Donahue, Julian McAuley, and Miller Puckette, "Adversarial Audio Synthesis," *ICLR*, 2019.

Unconditional Audio Synthesis using Diffusions

Example: WaveGrad (Chen et al., 2021)

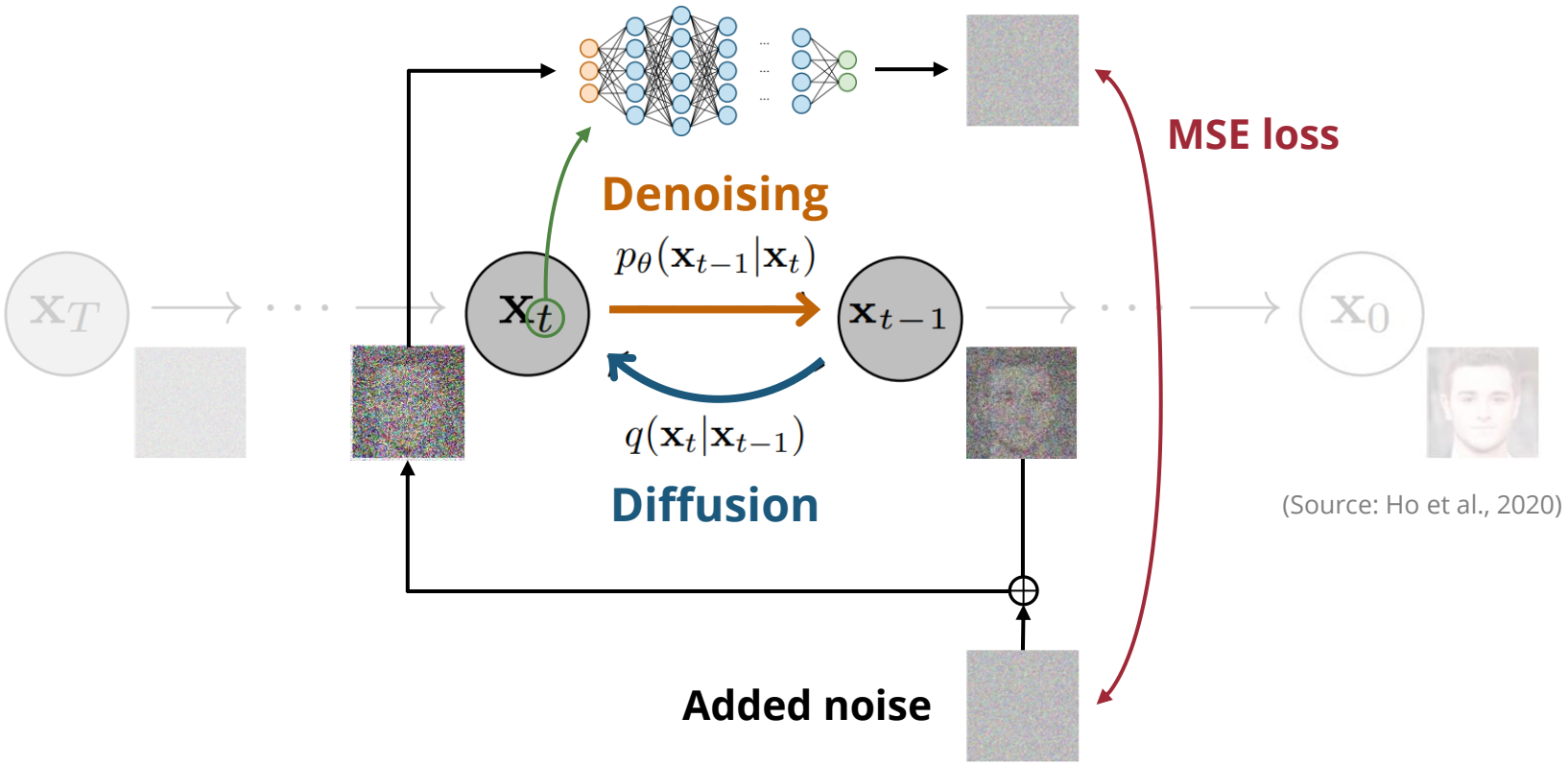


(Source: Chen et al., 2021)

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

(Recap) Diffusion Models – Training

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

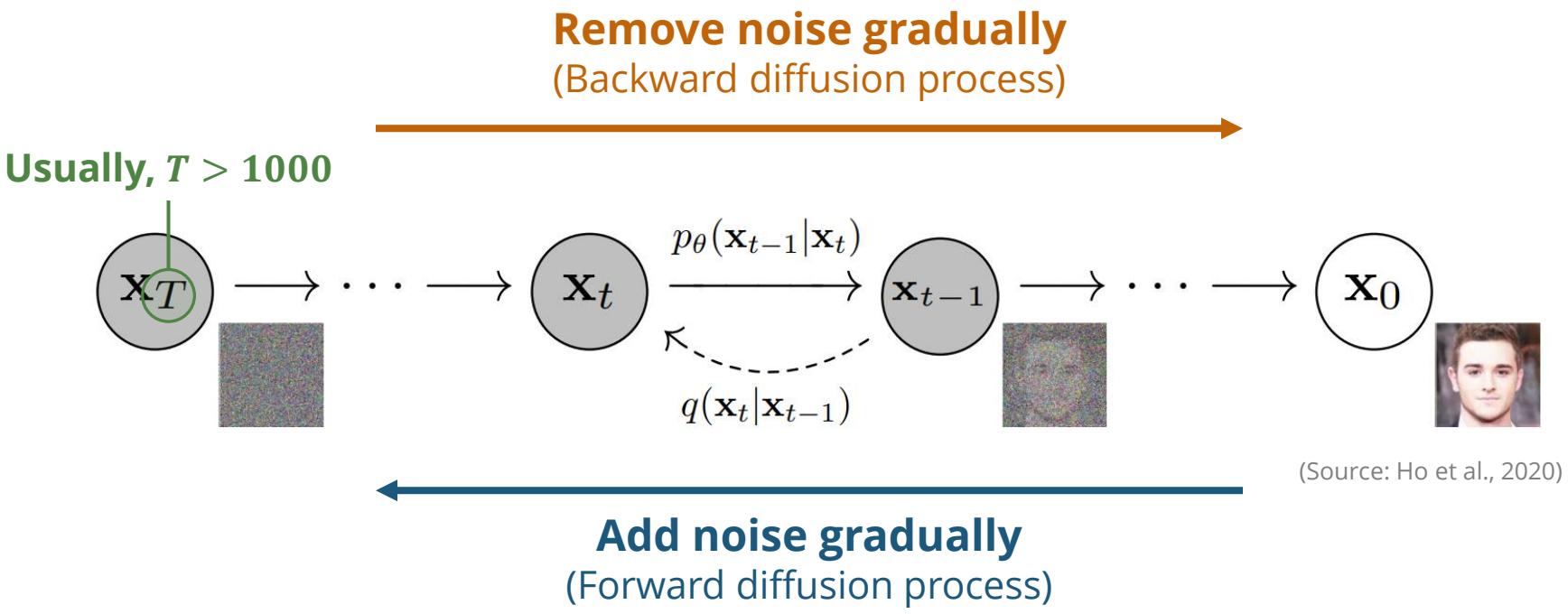


(Source: Ho et al., 2020)

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

(Recap) Diffusion Models

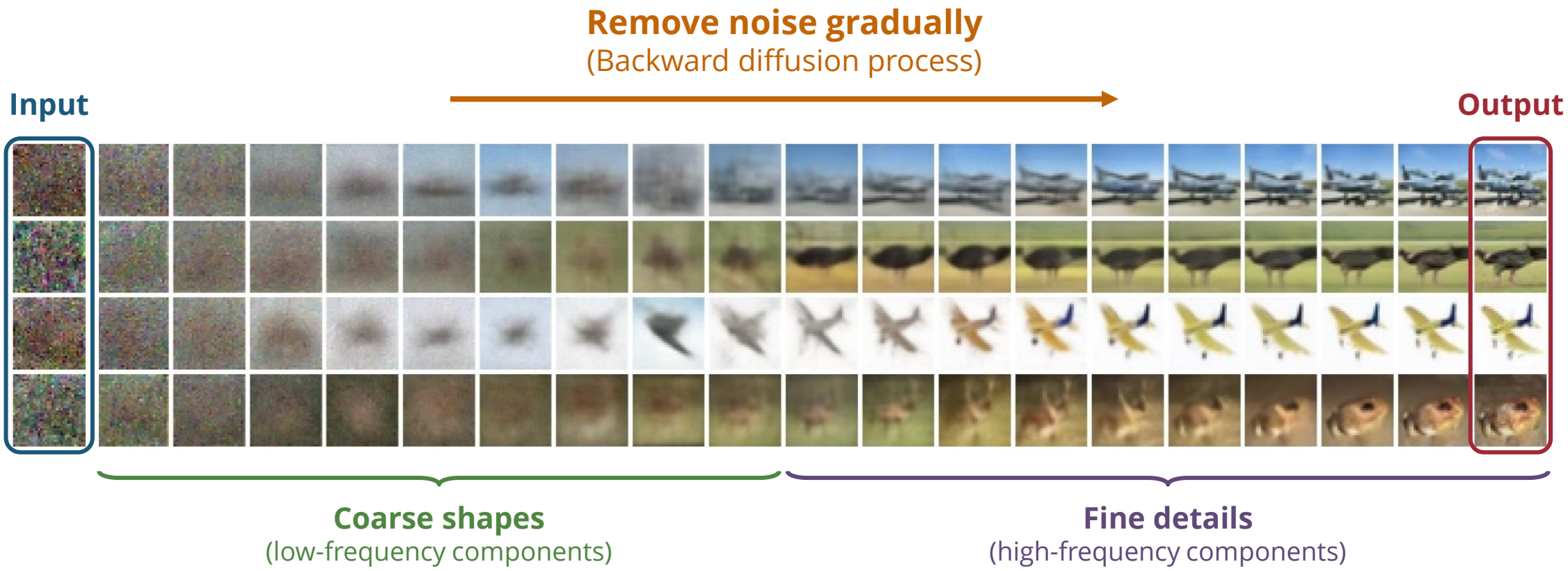
- **Intuition**: Many denoising autoencoders stacked together



(Source: Ho et al., 2020)

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

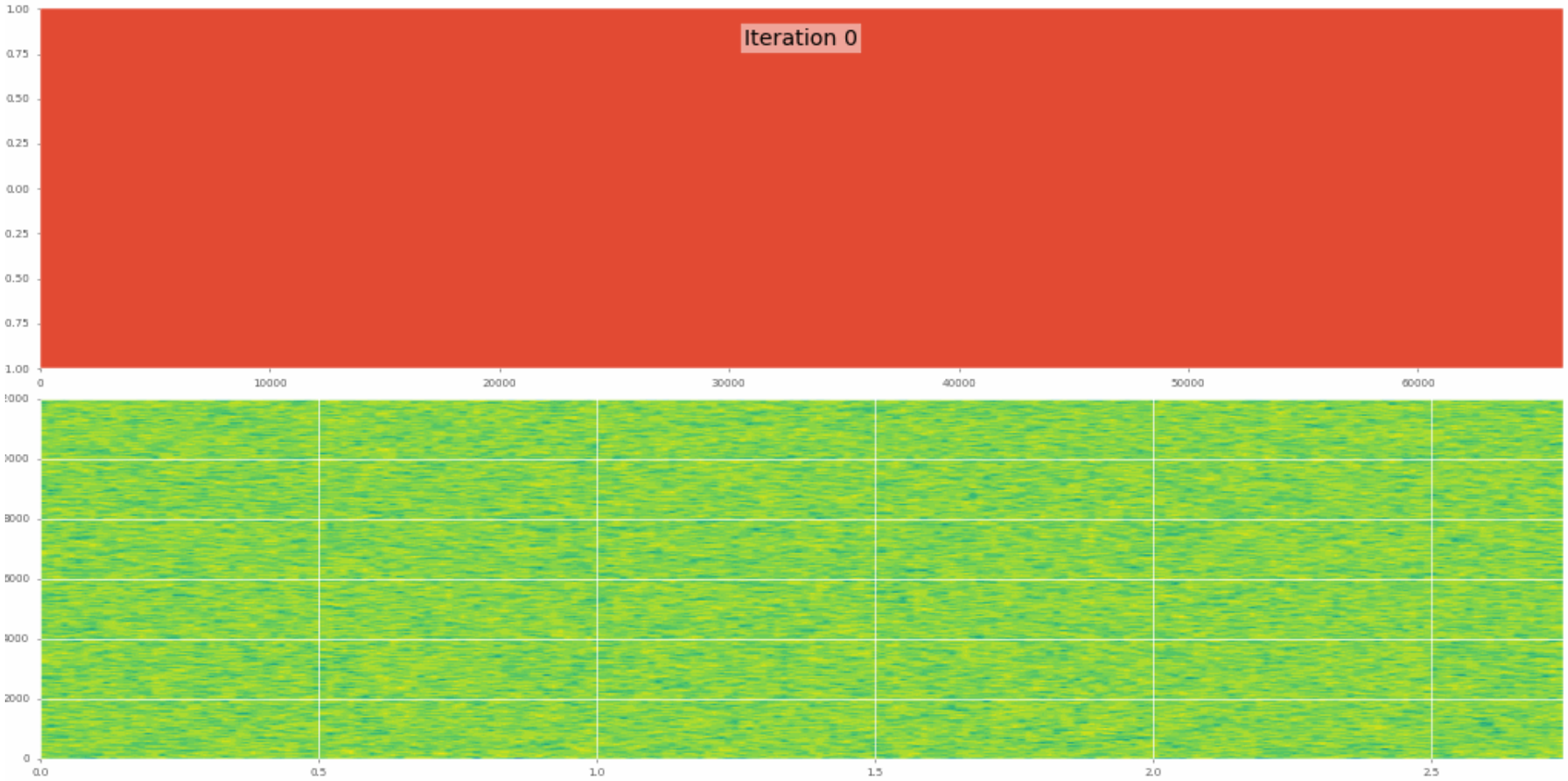
(Recap) Diffusion Models – Generation



(Source: Ho et al., 2020)

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

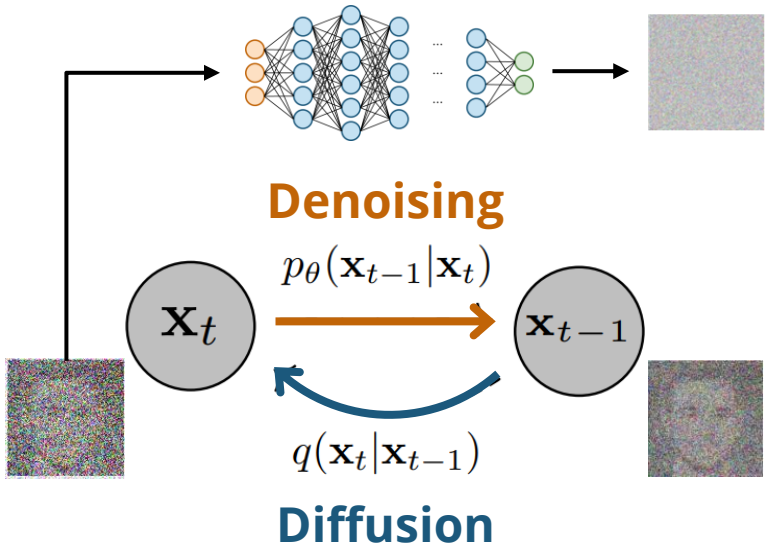
Example: WaveGrad (Chen et al., 2021)



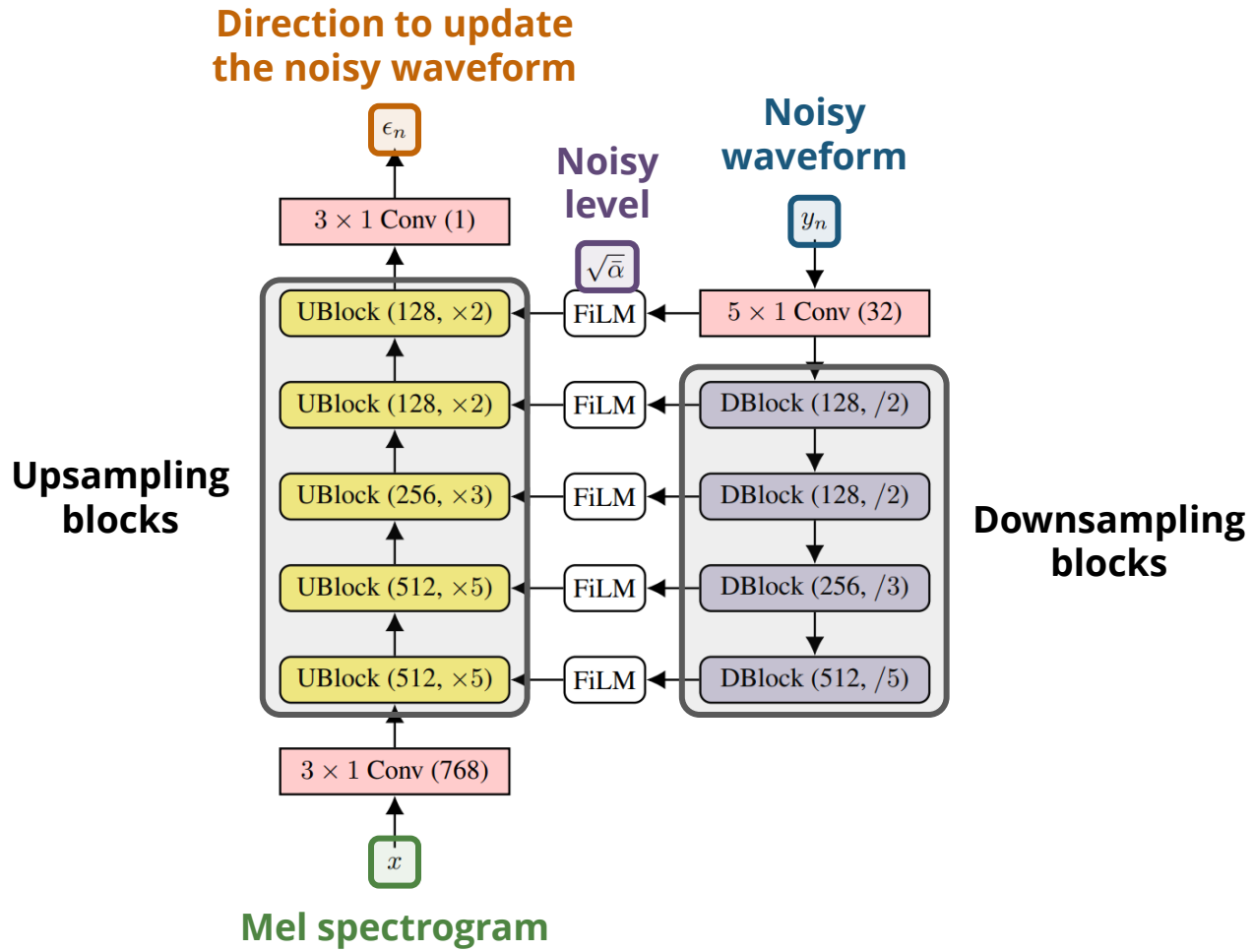
(Source: Chen et al., 2021)

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

Example: WaveGrad (Chen et al., 2021)



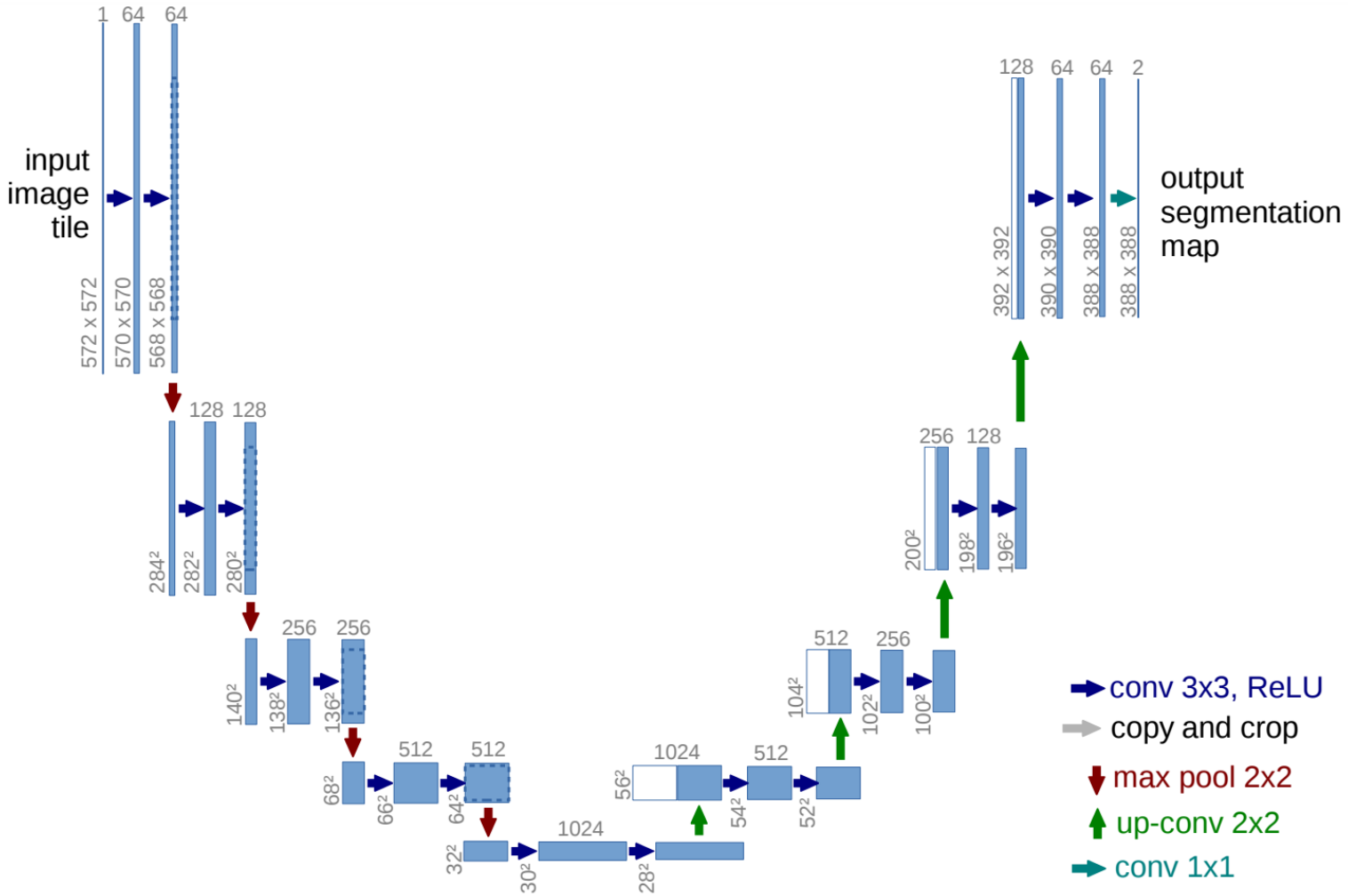
(Source: Ho et al., 2020)



(Source: Chen et al., 2021)

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

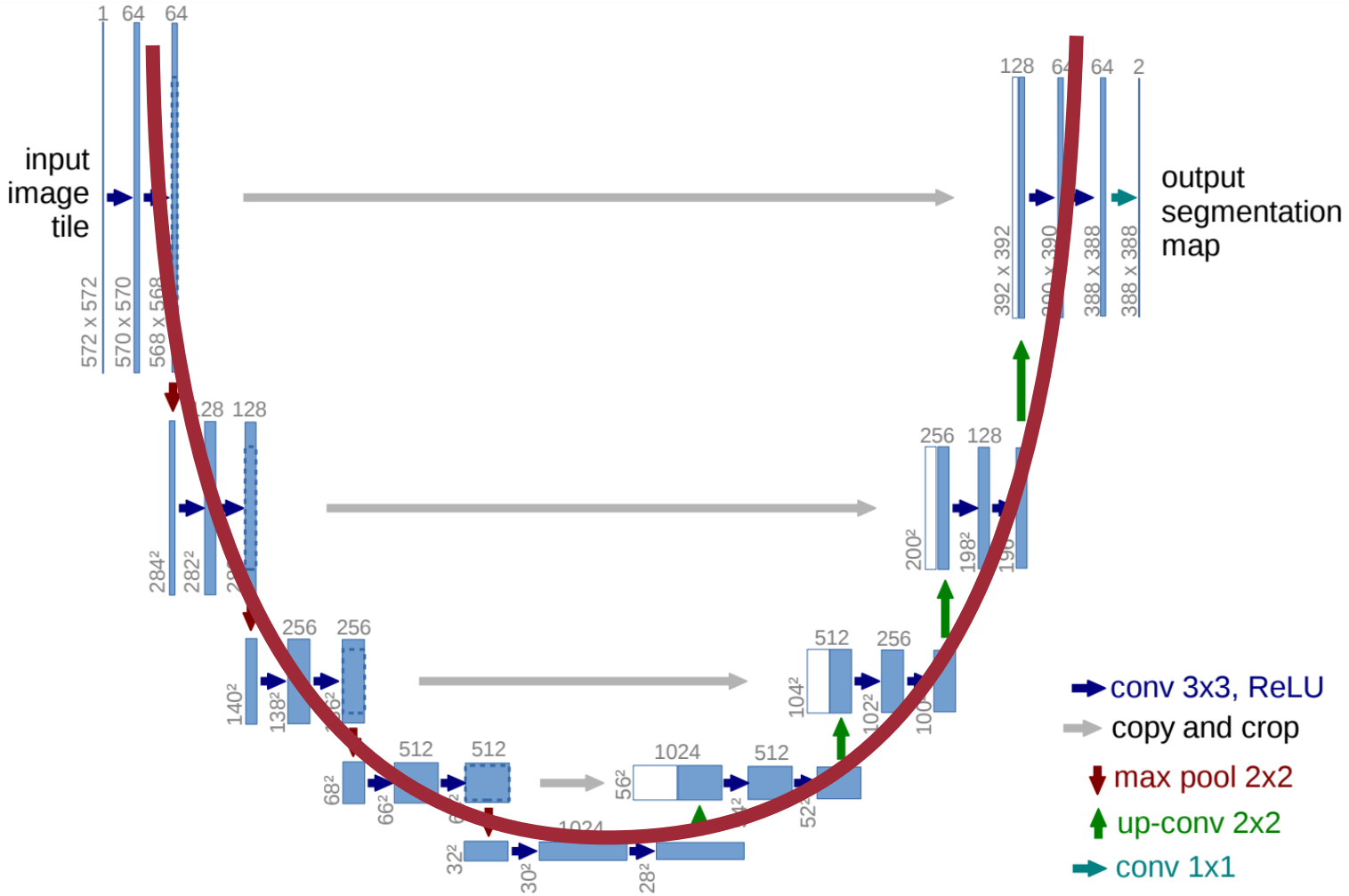
U-Net (Ronneberger et al., 2015)



(Source: Ronneberger et al., 2015)

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

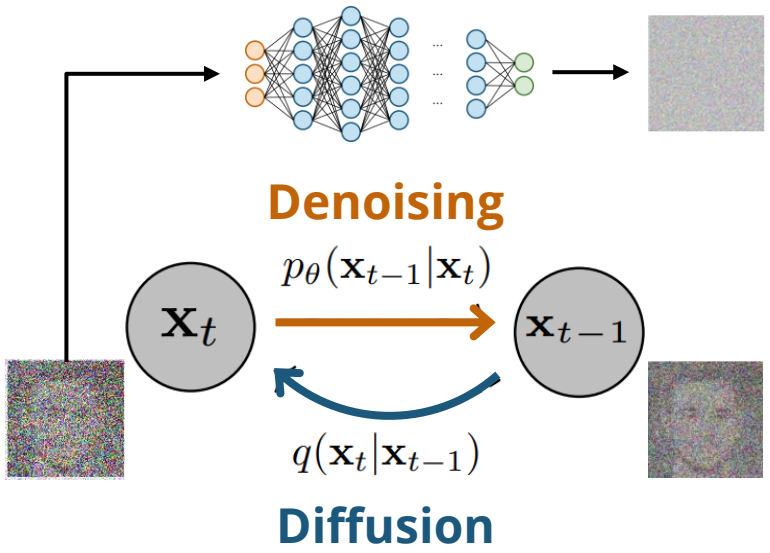
U-Net (Ronneberger et al., 2015)



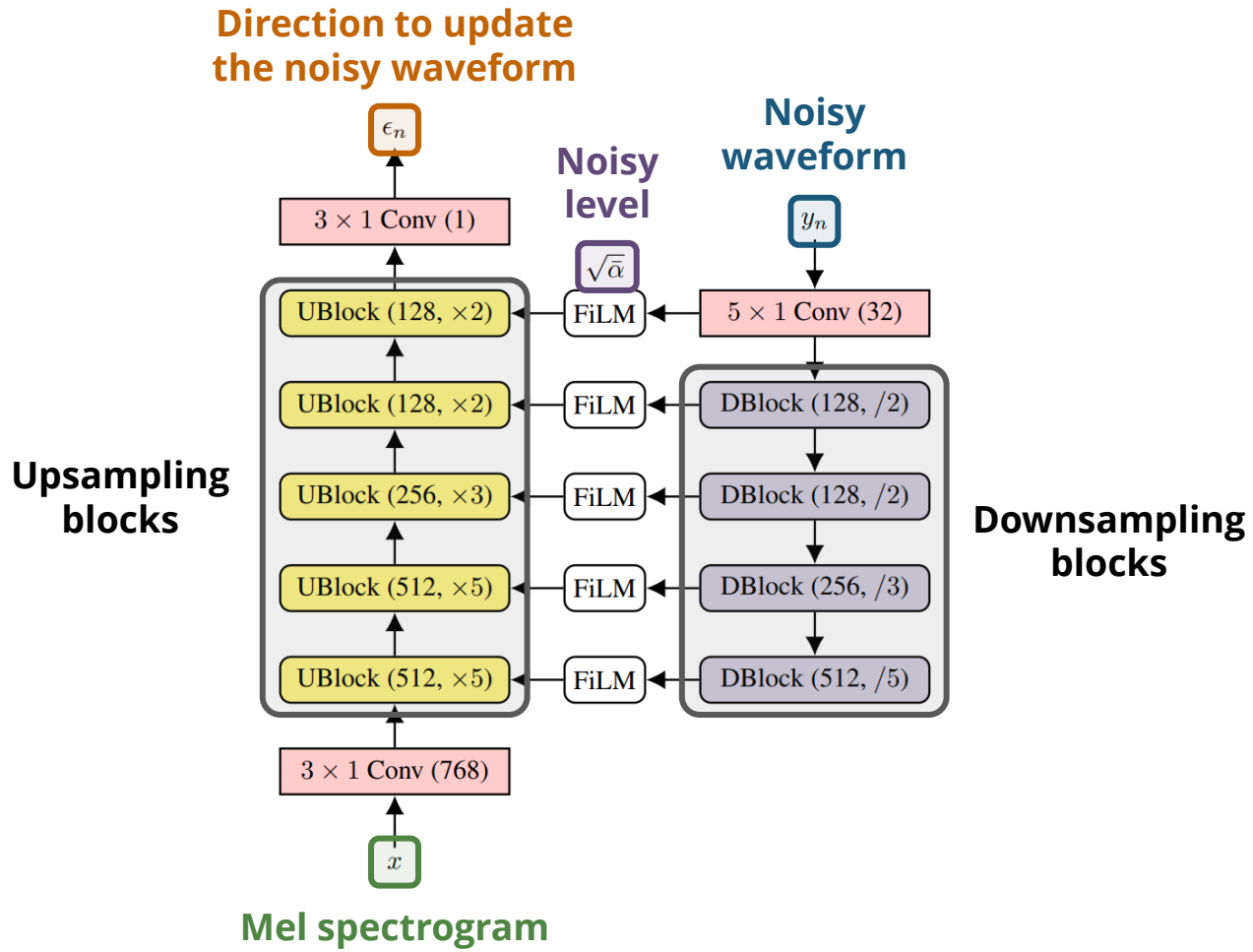
(Source: Ronneberger et al., 2015)

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

Example: WaveGrad (Chen et al., 2021)



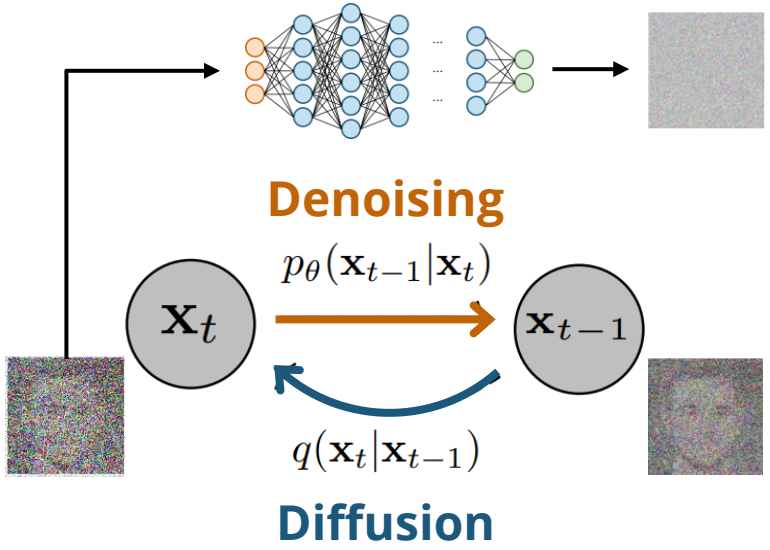
(Source: Ho et al., 2020)



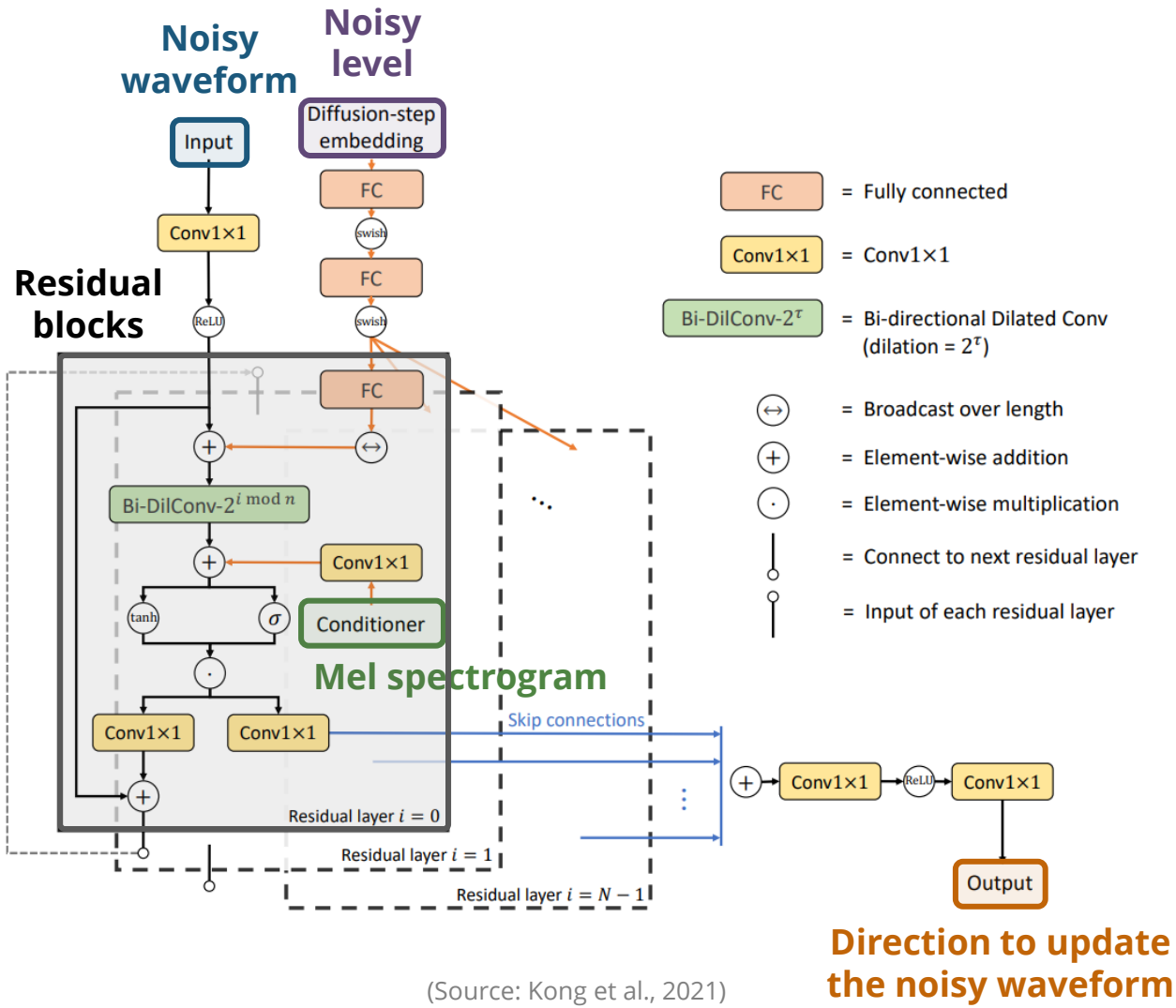
(Source: Chen et al., 2021)

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

Example: DiffWave (Kong et al., 2021)



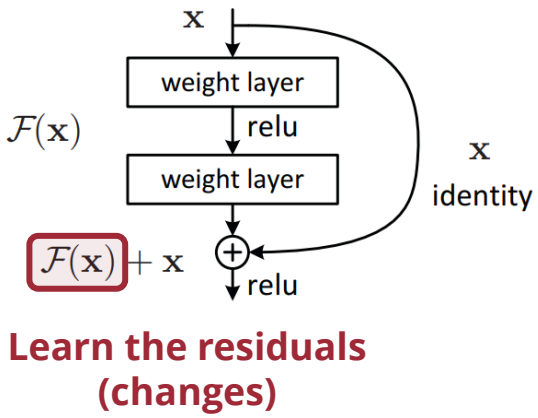
(Source: Ho et al., 2020)



(Source: Kong et al., 2021)

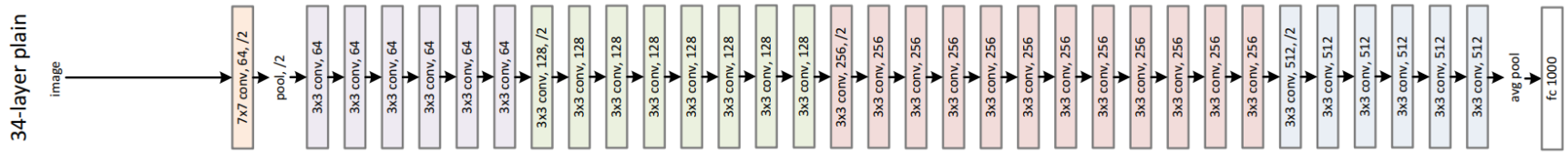
Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, Bryan Catanzaro, "DiffWave: A Versatile Diffusion Model for Audio Synthesis," *ICLR*, 2021.

Deep Residual Nets (ResNets) (He et al., 2016)

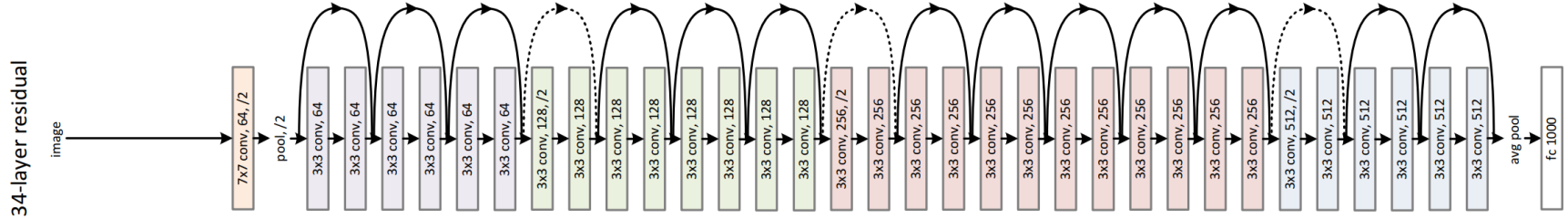


Learn the residuals (changes)

Without skip connections



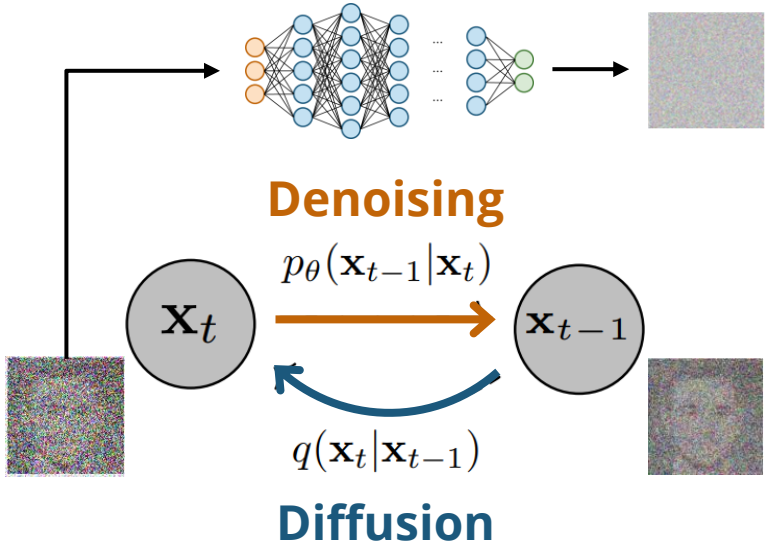
ResNet



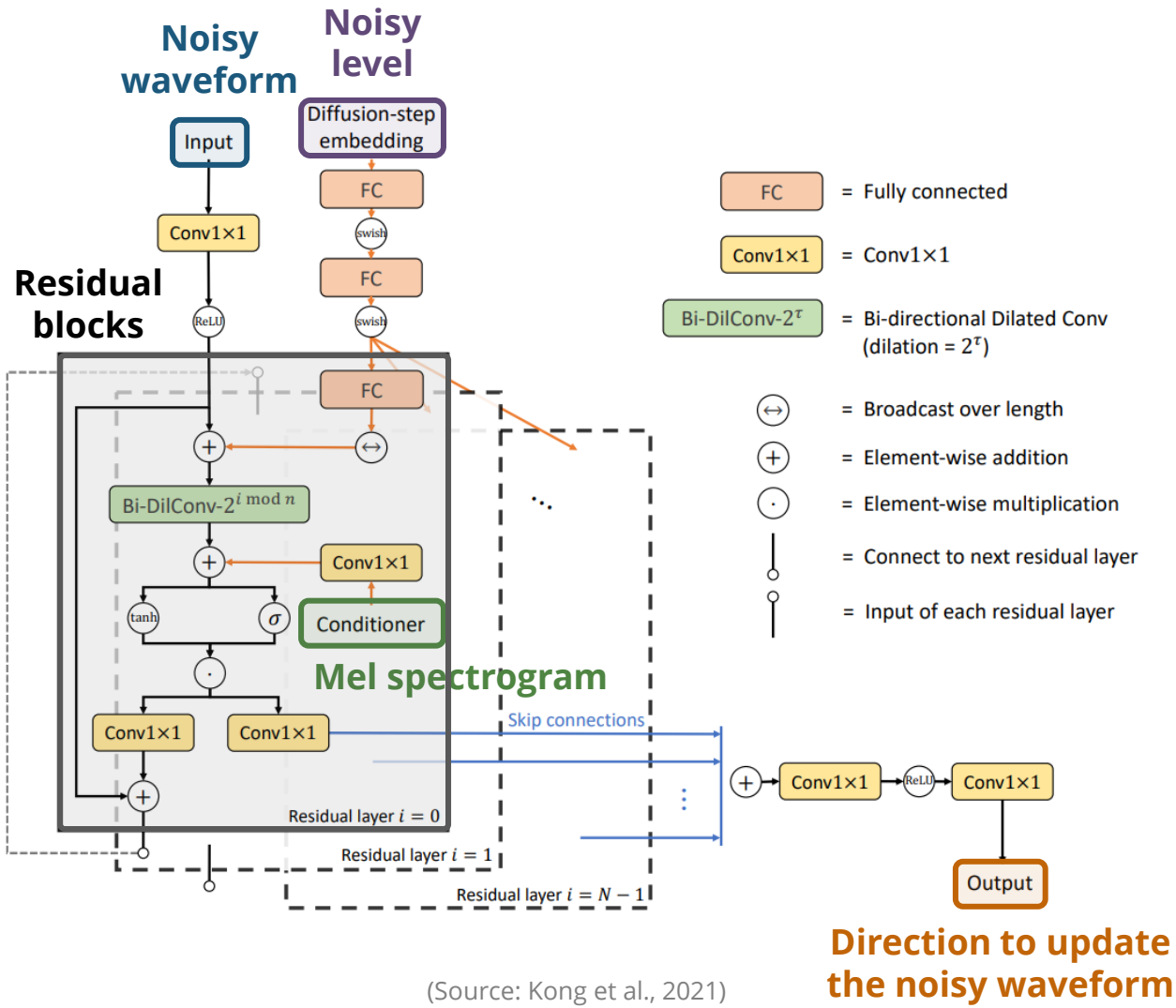
(Source: He et al., 2016)

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

Example: DiffWave (Kong et al., 2021)



(Source: Ho et al., 2020)

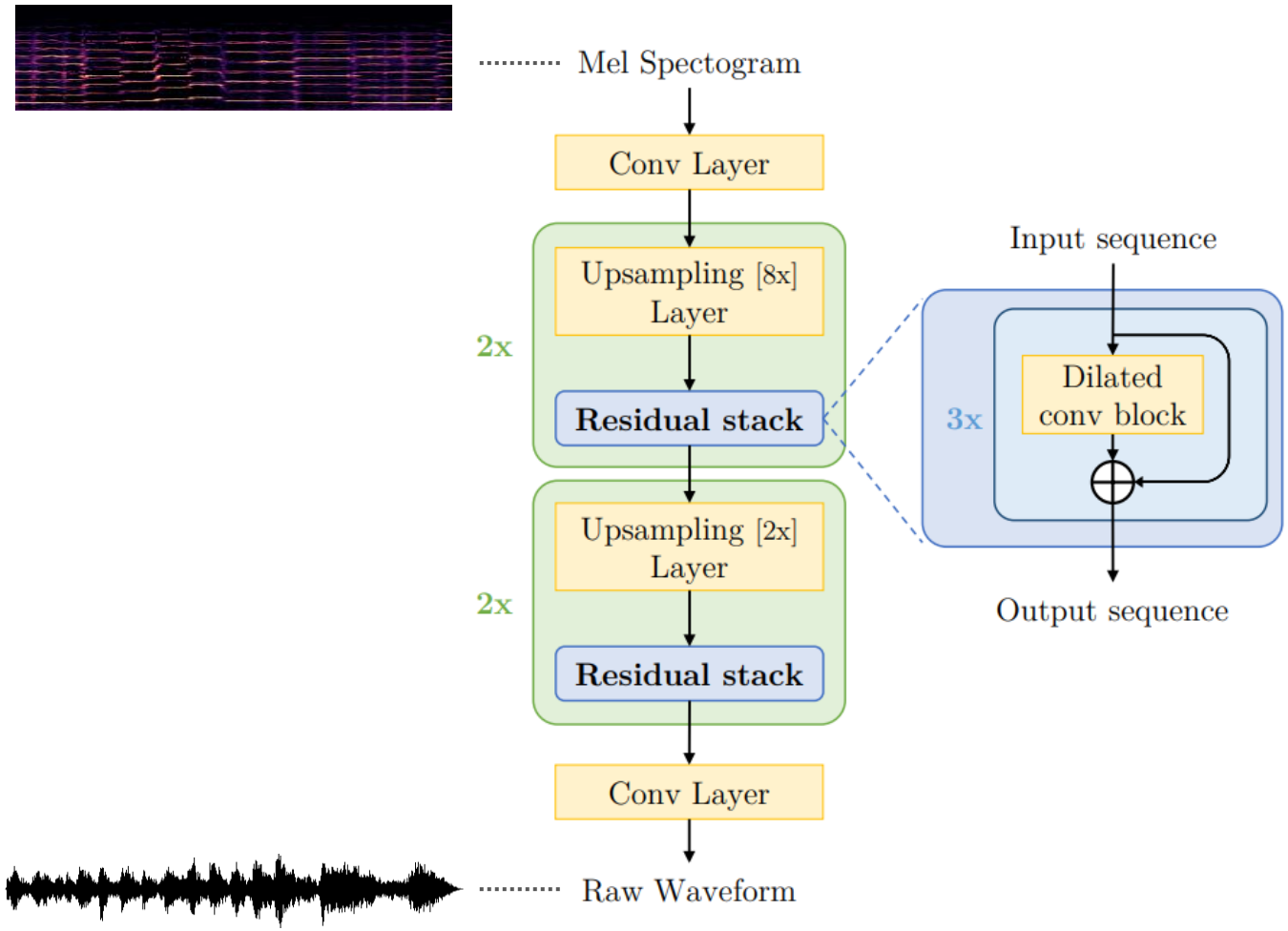


(Source: Kong et al., 2021)

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

Conditional Audio Synthesis

Example: MelGAN (Kumar et al., 2019)

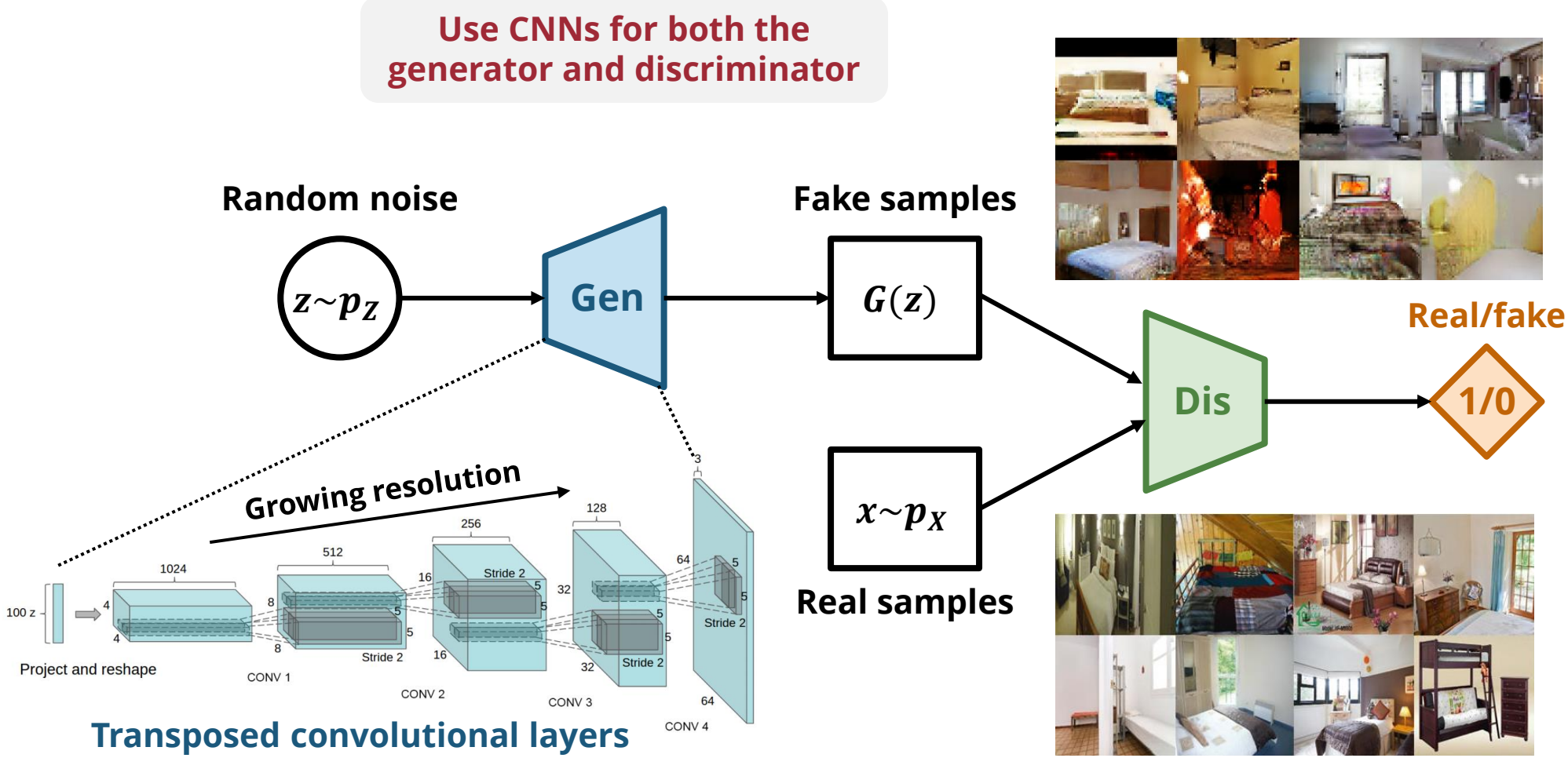


(Source: Kumar et al., 2019)

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," *NeurIPS*, 2019.

(Recap) Deep Convolutional GANs (DCGANs)

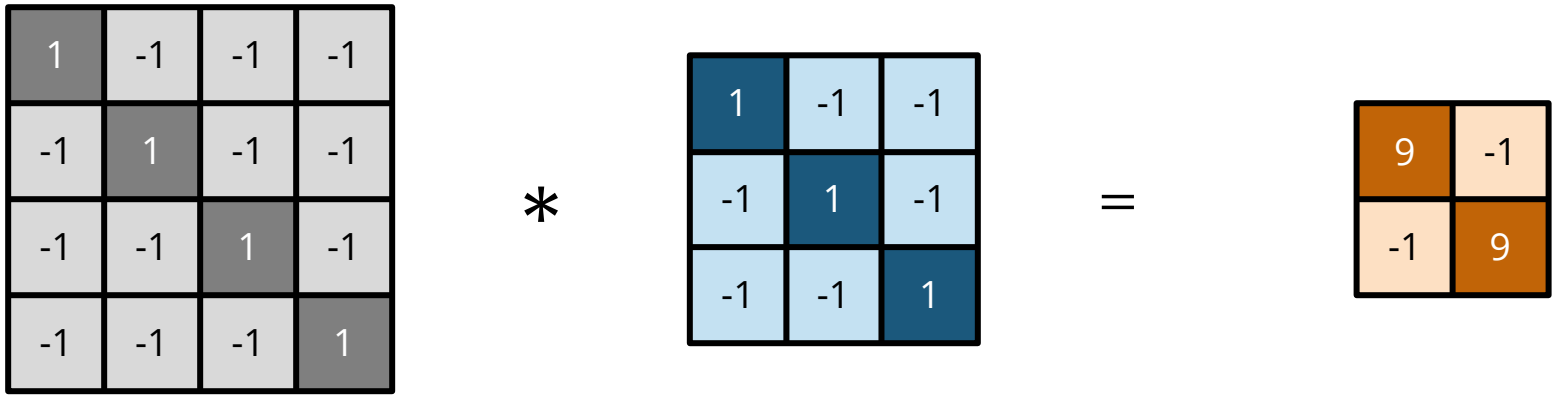
Use CNNs for both the generator and discriminator



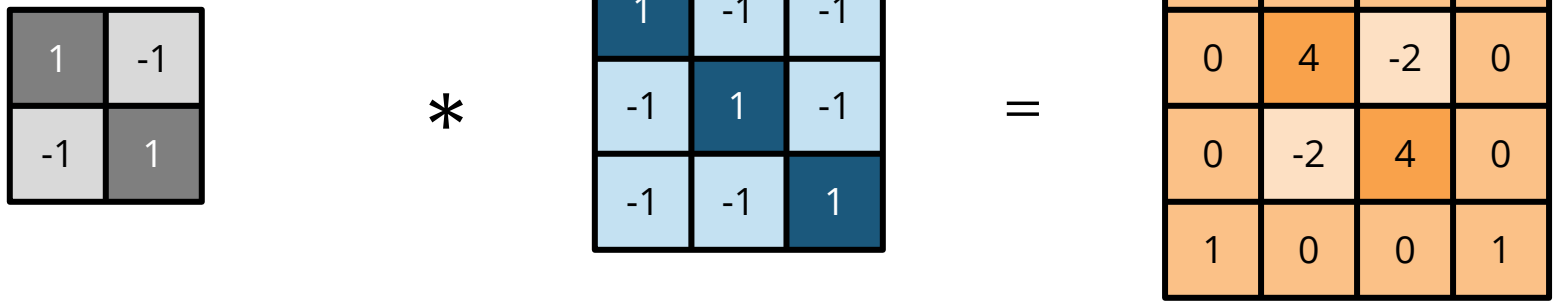
Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *ICLR*, 2016.

(Recap) Transposed Convolution

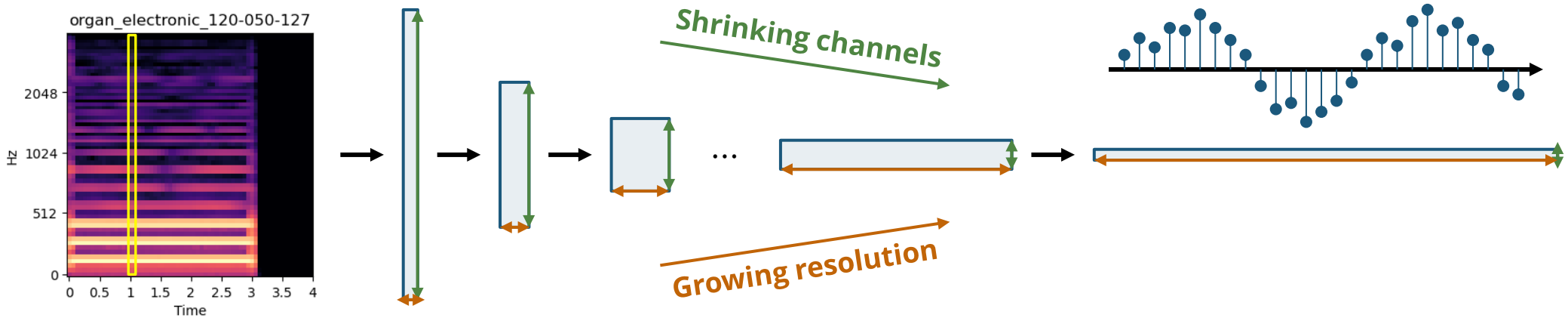
Convolution



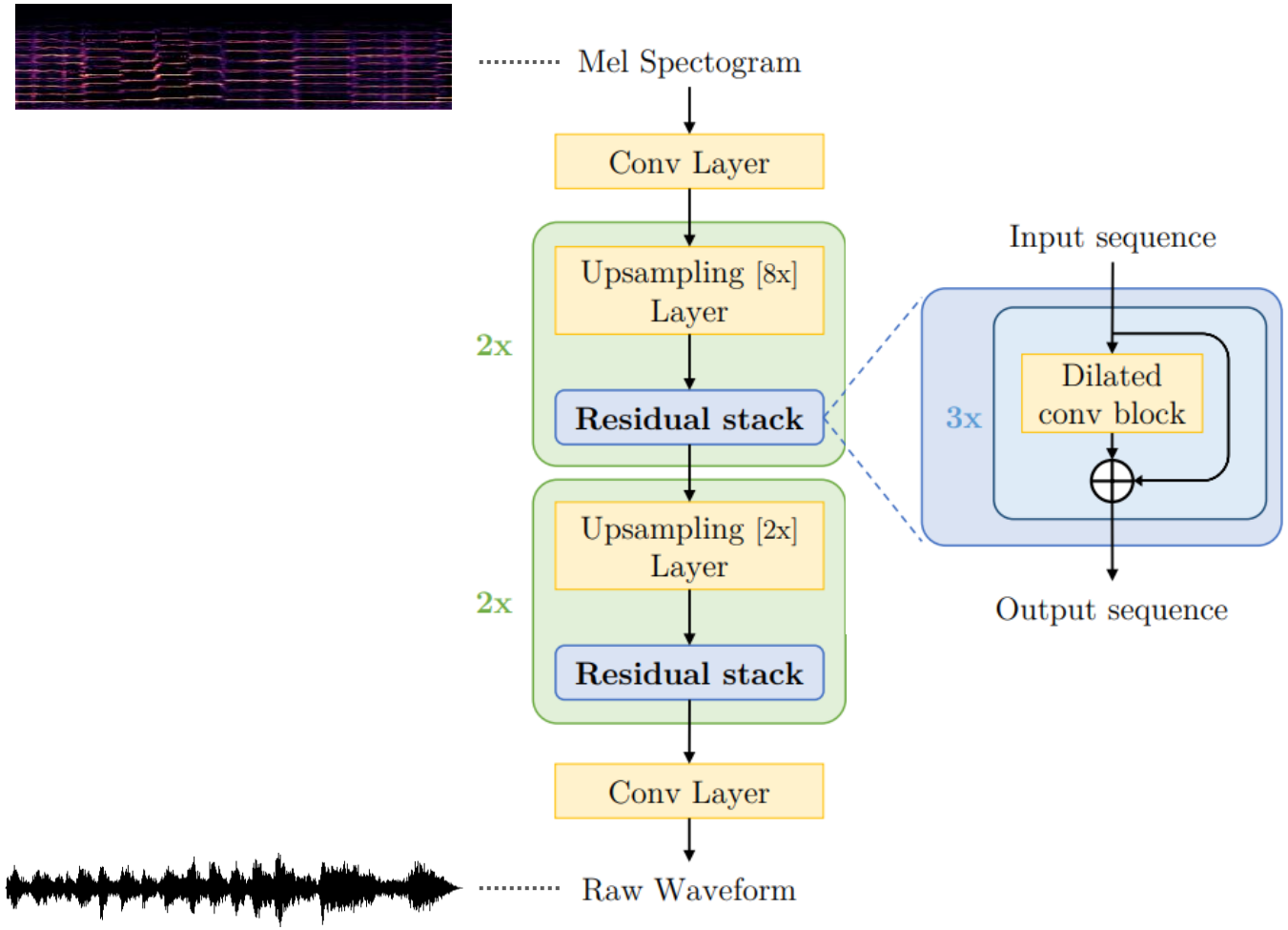
Transposed convolution



Transposed Convolution for Vocoders



Example: MelGAN (Kumar et al., 2019)

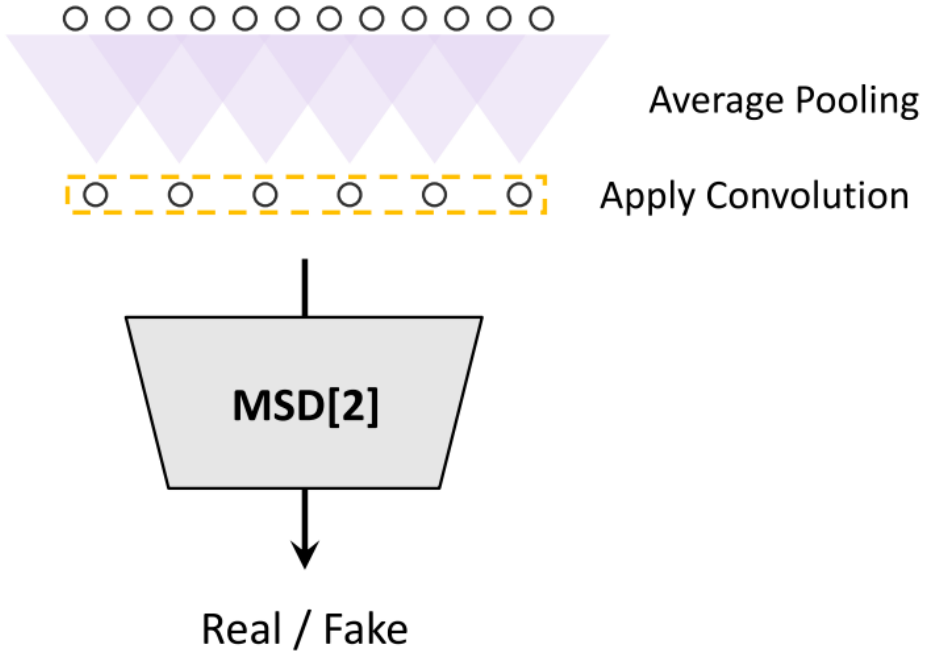


(Source: Kumar et al., 2019)

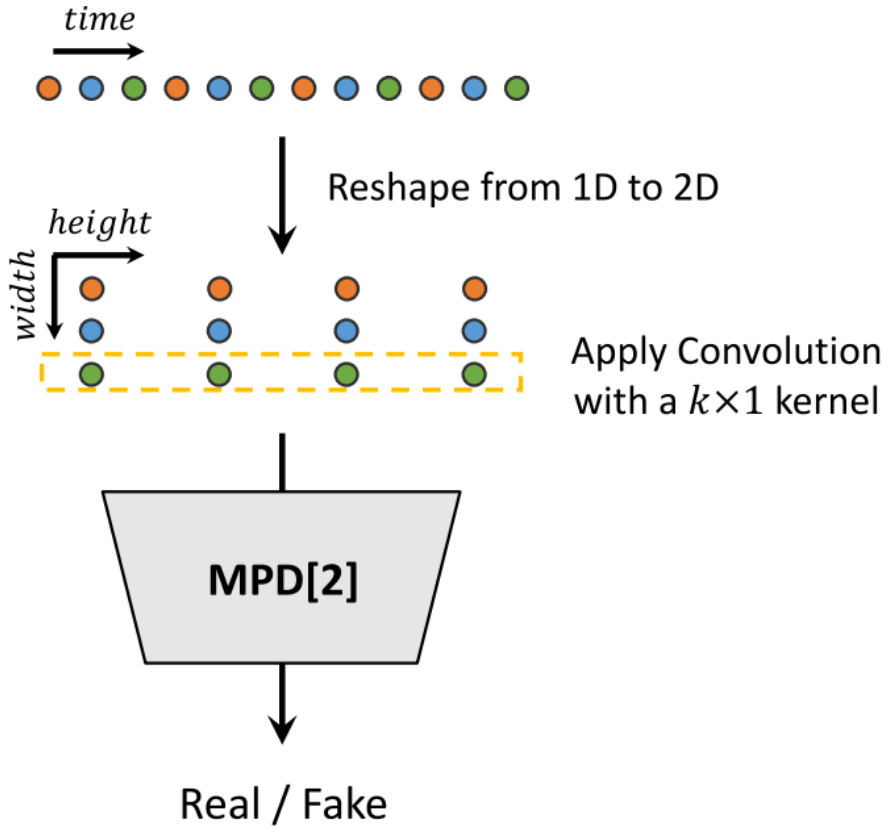
Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," *NeurIPS*, 2019.

Example: MelGAN (Kumar et al., 2019)

Multi-scale discriminator



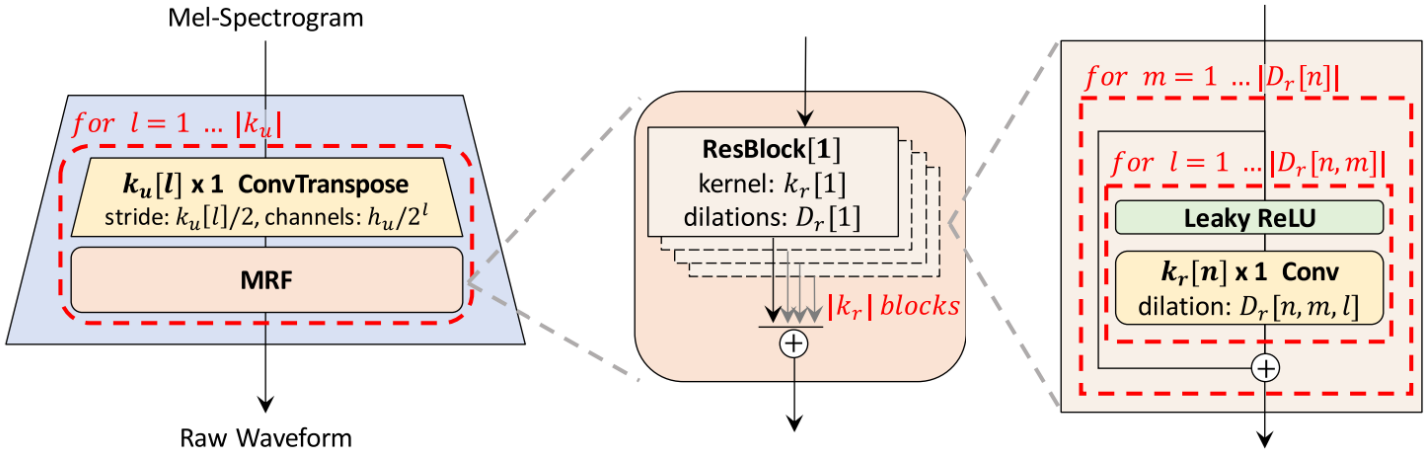
Multi-period discriminator

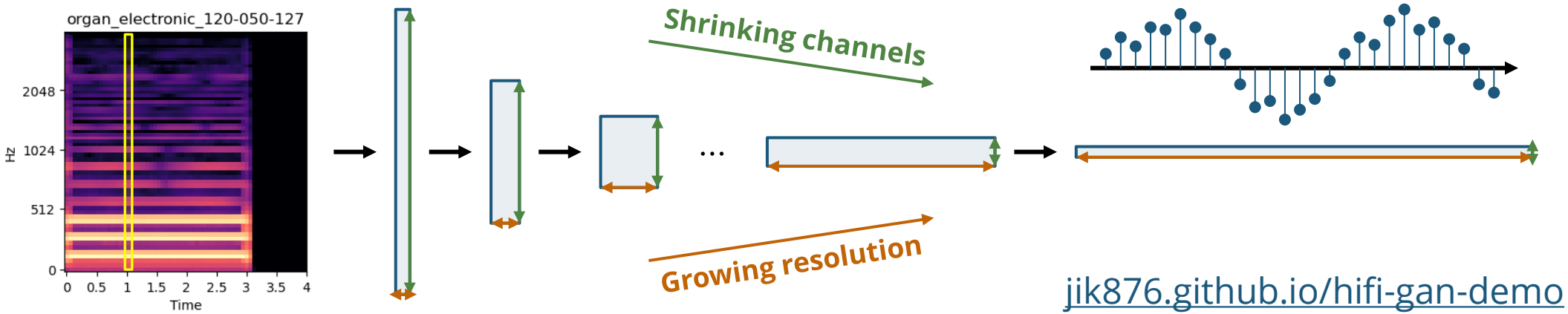


(Source: Kong et al., 2019)

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," *NeurIPS*, 2019.
Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *NeurIPS*, 2020.

Example: Hifi-GAN (Kong et al., 2020)





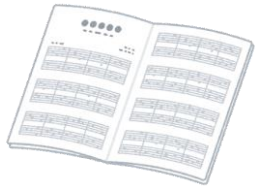
jik876.github.io/hifi-gan-demo

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

Optional Reading

- A very nice blog on “**Generating music in the waveform domain**” by Sander Dieleman: sander.ai/2020/03/24/audio-generation

(Recap) Four Paradigms



Symbolic music generation

Text-based

Image-based

Audio-domain music generation

Time series-based

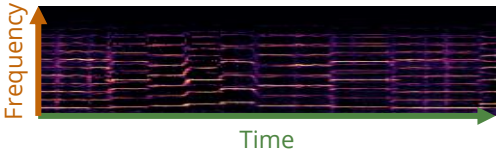
Image-based

```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67,
...
```

MIDI

Piano roll

Waveform



Spectrogram

Today, we also have many latent-space based systems!