PAT 498/598 (Winter 2025)

Music & AI

Lecture 16: Audio-domain Music Generation

Instructor: Hao-Wen Dong

Homework 5: AI Song Contest

- **Q1**: Which is your favorite song?
- **Q2**: Following Q1, what did they do well?
- **Q3**: Following Q1, what can be improved?
- **Q4**: Based on the ten finalists, what tasks are easy for current AI in music production?
- **Q5**: Based on the ten finalists, what tasks are difficult for current AI in music production?

Homework 5: Favorite Songs

- **4 votes** for **Genre Cannon** by **Dadabots**
- **3 votes** for **Sudamérica** by **Onda Corta**
- **3 votes** for **Heart Not Found** by **Error 305**
- **3 votes** for **Echoes of the Synthetic Forest** by **KeRa**
- **1 vote** for **One Mantra** by **DJ Swami**
- **1 vote** for **binary b1o0d** by **HEL9000**

Homework 5: What did they do well?

- **Genre Cannon** by **Dadabots**
  - *“It felt like it **made use of the advantages we get with AI** instead of relying on it to generate everything and replace creativity.”*
  - *“They had **varying musical styles that transitioned into one another very seamlessly**.”*
  - *“A standout moment occurred when the song **transitioned from a “Star Wars cantina”-style piece into rock, then quickly pivoted to a more abstract percussive section**.”*

Homework 5: What did they do well?

- **Sudamérica** by **Onda Corta**
  - *"What I like more about this song than the others is the fact that* ***it makes an actual cohesive song to my ears****."*
  - *"Their production shows a thoughtful balance, using* ***AI tools to craft intricate soundscapes while retaining a warm, organic quality****."*
  - *"As an Argentinian, I really liked Sudamérica because* ***it surprisingly captured the essence of Latin American pride and culture in a way that felt genuine****."*

Homework 5: What did they do well?

- **Heart Not Found** by **Error 305**
  - *"… this song really exemplified **how you can make a real song using AI without it sounding like it was AI**."*
  - *"Other songs also played with genre, but I thought this piece did this in a less experimental, obvious way and **more to serve an artistic message within their song**."*

Homework 5: What did they do well?

- **Echoes of the Synthetic Forest** by **KeRa**
  - *"By using their own compositions as the primary training data, they ensured the **AI's output stayed true to their unique sound and artistic vision**."*
  - *"Unlike most AI-generated music, which often leans toward electronic, black metal, or hard rock, this piece **explores ambient music, rich with sonic textures that challenged my expectations of AI composition**."*

Homework 5: What did they do well?

- **One Mantra** by **DJ Swami**
    - *"There is a nice blend of timbres.* ***Each synth sound is distinct and is so amorphous that it isn't initially obvious AI constructed parts of the composition****."*
- ***binary b1o0d*** *by* ***HEL9000***
    - *"This is one of the only submissions that actually* ***accurately recreated it's source material****."*

Homework 5: What are easy?

- ”It seems like **getting musical ideas** is good for current AI.”
- “I also think that the current AI does a good job of **creating lyrics that are relevant to the subject of the song**.”
- “One of the easiest tasks for AI is **generating loops**, such as drum patterns, chord progressions, and melodic lines.”
- “**Synthesizing instrument sounds**, especially electronic instruments seems fairly easy.”
- “The technology can also **generate innovative sound effects and synthetic timbres** that enrich the overall production.“
- It appears that **making a solid beat or groove** is somewhat easy for current AI.

Homework 5: What are difficult?

- *“I think **voice AI can be improved a lot**, because based on the finalists, the ones that used AI for voice were quite poor.“*
- *“I also noticed **AI struggling to produce expressive and natural vocals** in many of the finalist pieces.”*
- *“On some tracks, **they [AI voice] were warpy and the lyrics were essentially unintelligible**. On the tracks where **the lyrics were perceptible and the vocals sounded realistic**, they completely stuck out in the mix.”*
- *“One dimension that most of the AI-generated songs seemed to lack was emotional depth, particularly in the vocal performances.”*

Homework 5: What are difficult?

- *"Despite its impressive capabilities, current AI still faces challenges when it comes to* ***capturing the nuanced emotional expression*** *that is in human performances."*
- *"AI often struggles to convey emotional depth and nuance, resulting in compositions that feel technically proficient but lack soul."*
- *"I feel a difficult task for current AI in music production is developing the emotional depth and nuance in the music that a human artist can create."*

Homework 5: What are difficult?

- *“To me, this indicates that there is a quality to music created by a human that* ***carries certain emotional weight that cannot be mimicked by AI*** *no matter how well it can copy humans.* ***Human intent cannot be faked, and was often added in.”***
- *“…, but when it comes to* ***building a fan base and fleshing out emotionally impactful songs,*** *I have a hard time seeing them succeed on their own.”*
- *“For example, it struggles to create songs that* ***have a clear emotional journey or tell a story from start to finish****.”*

Homework 5: What are difficult?

- *"AI **struggles to generate a coherent and creative long-form composition with a fully developed storyline**, so human guidance is essential in shaping the overall structure and artistic intent."*
- *"A common theme I noticed was that sometimes songs **have trouble keeping consistency**."*
- *"I also feel another challenge is effectively **integrating all the elements generated by AI into a cohesive final product**."*
- *"I would say that the current AI has the most difficulty **creating music that flows well together**."*

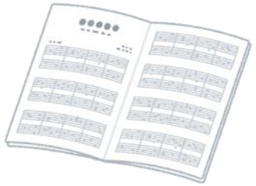
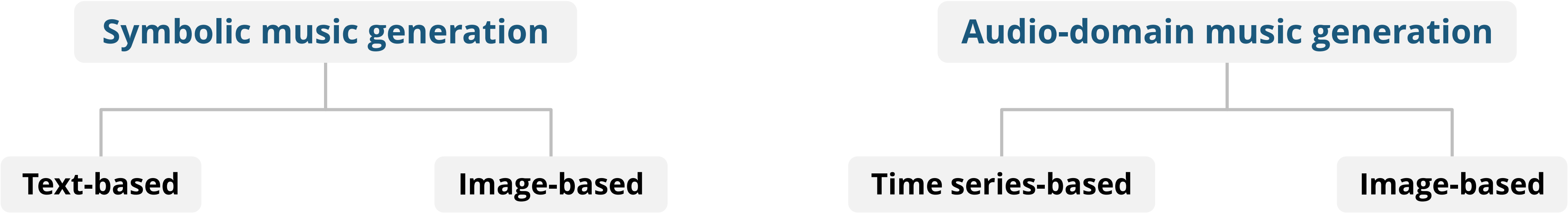
Homework 5: What are difficult?

- I think that it is evident that it is difficult to **one shot create AI music that fits what the user wants**, despite models like Suno being so good.
- **Phrasing is often a big issue**, with melodic lines often sounding like run-on sentences instead of melodic phrases which feel good to humans.
- A lot of the pieces, while super interesting musically, felt like they were **missing some of the details that make a track really come to life**.

Discussions

- **To what extent of human involvements** can a song still be called AI music?
- **Shall we intervene** if AI-generated material doesn't sound polished?
- What is the **goal of AI music**?

> "Whatever you now find weird, ugly, uncomfortable and nasty about a new medium will surely **become its signature**."
>
> – Brian Eno, 1996

(Recap) Four Paradigms of Music Generation



Symbolic music generation

Text-based

Image-based

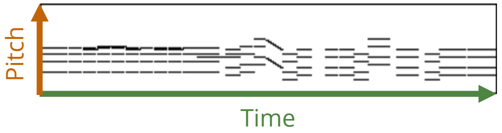
Audio-domain music generation

Time series-based

Image-based

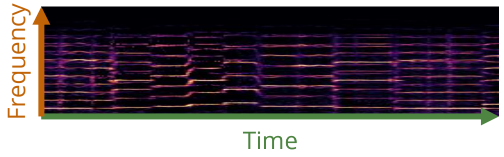
```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67,
...
```

MIDI



Piano roll

Waveform

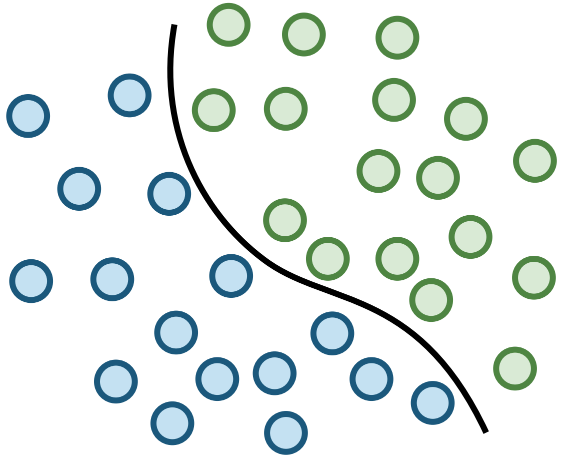


Spectrogram

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

(Recap) Discriminative vs Generative Models

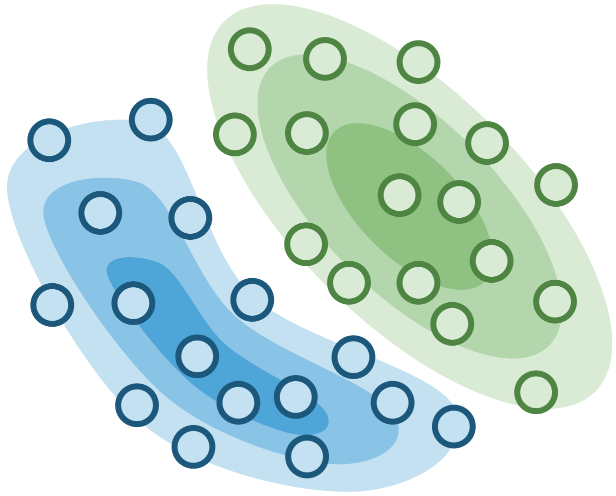
Discriminative



Discriminative models learn the decision boundary

$P(y|x)$

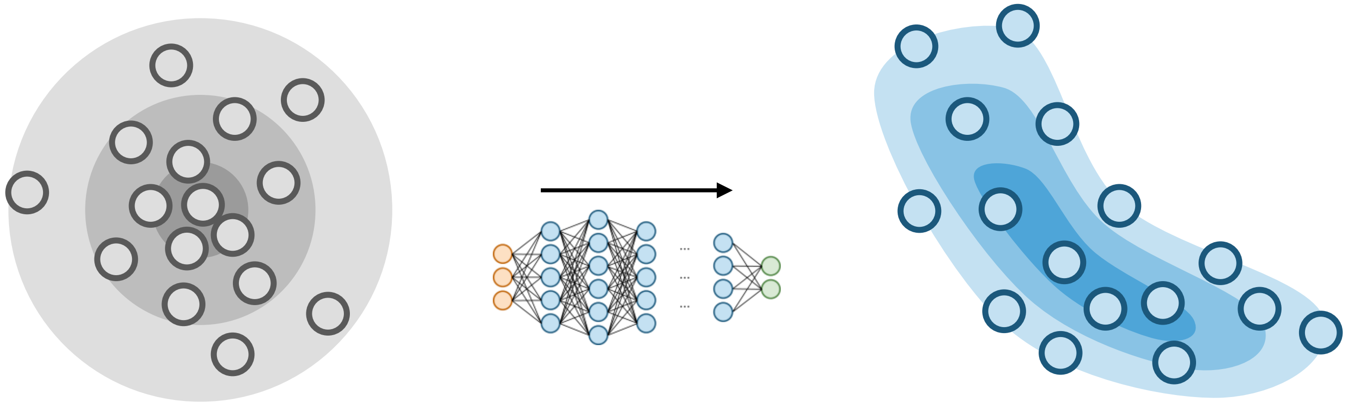
Generative



Generative models learn the underlying distribution

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

(Recap) Generating Data from a Random Distribution



Random distribution

Data distribution

$P(z)$

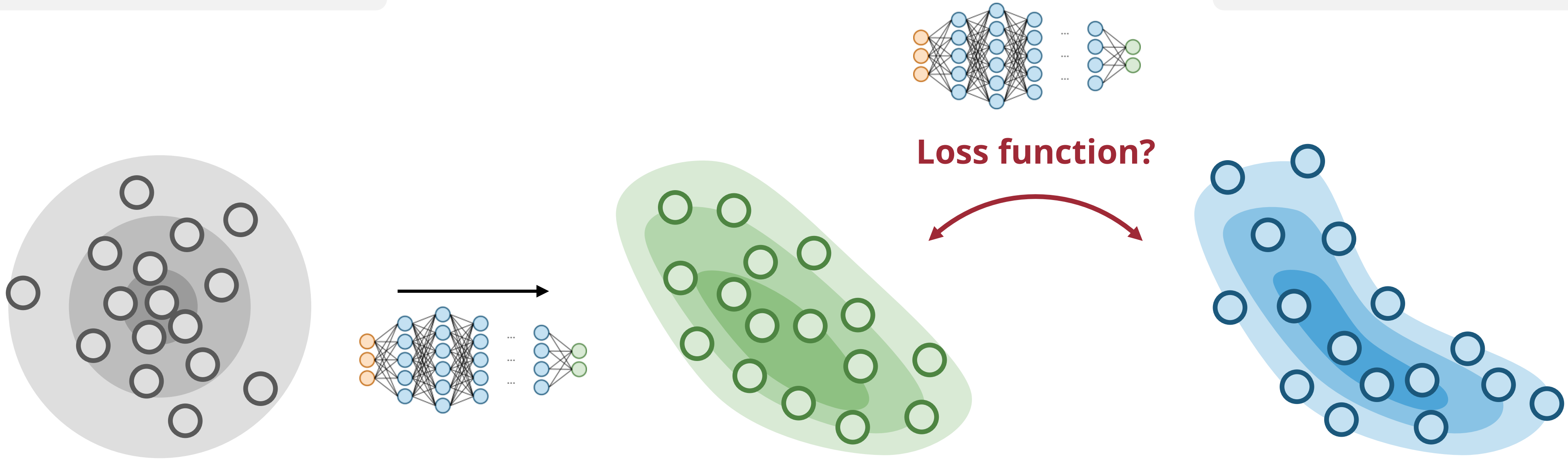
$P(x)$

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

(Recap) A Loss Function for Distributions

Random distribution

Data distribution



Loss function?

$P(z)$

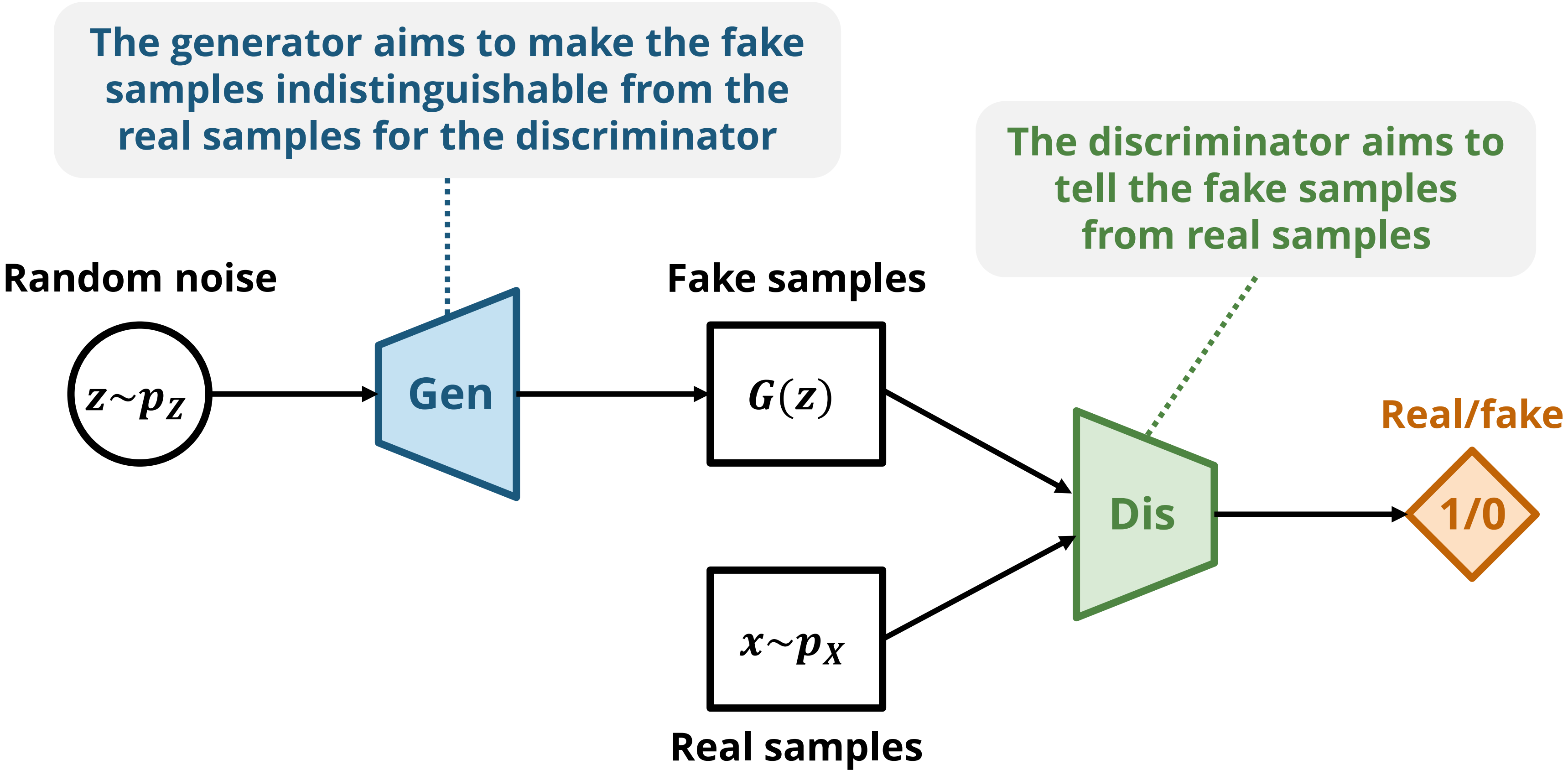
$P(\hat{x})$

$P(x)$

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

But what about another neural network!?

(Recap) Generative Adversarial Nets (GANs) (Goodfellow et al., 2014)

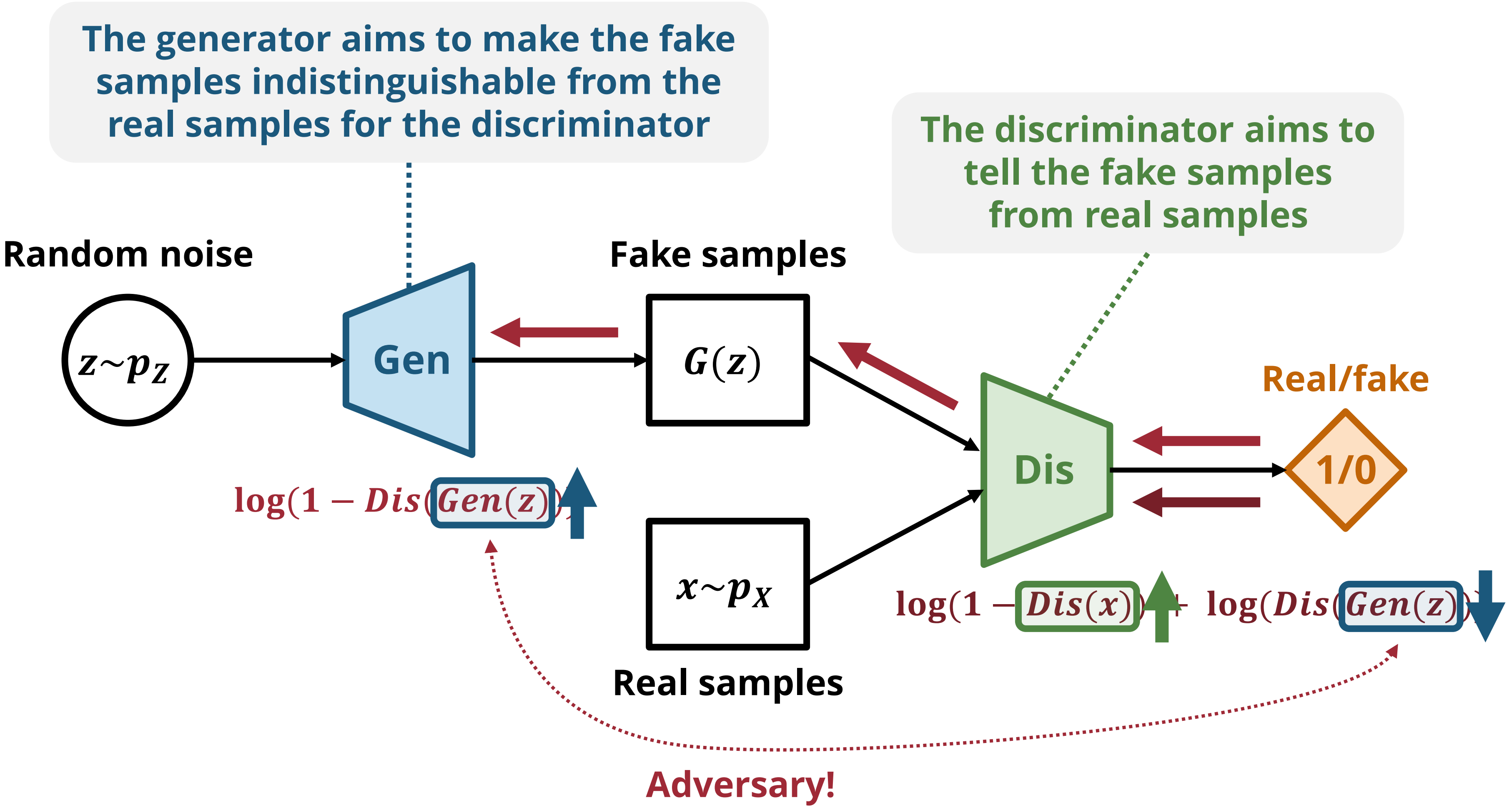


The generator aims to make the fake samples indistinguishable from the real samples for the discriminator

The discriminator aims to tell the fake samples from real samples

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) – Training



The generator aims to make the fake samples indistinguishable from the real samples for the discriminator

The discriminator aims to tell the fake samples from real samples

Random noise

$z \sim p_Z$

Gen

Fake samples

$G(z)$

Dis

Real/fake

1/0

$x \sim p_X$

Real samples

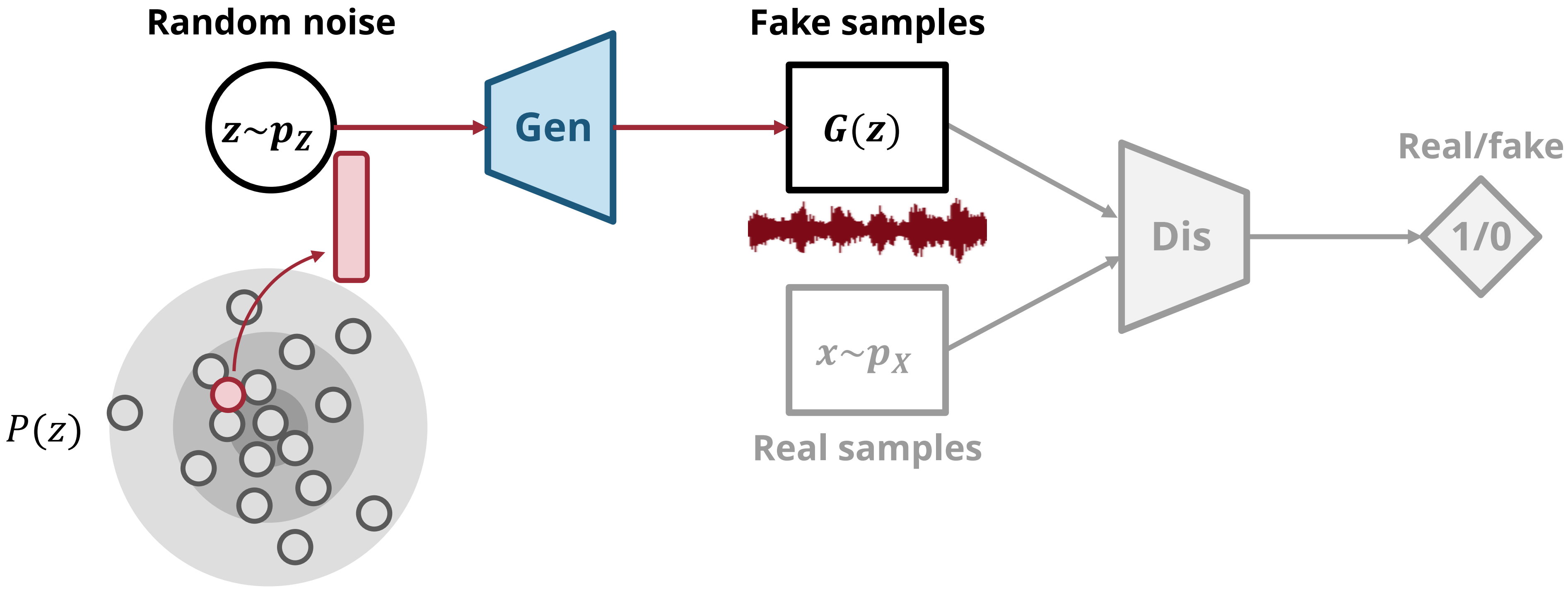
$\log(1 - Dis(\boxed{Gen(z)})) \uparrow$

$\log(1 - \boxed{Dis(x)}) \uparrow + \log(Dis(\boxed{Gen(z)})) \downarrow$

Adversary!

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.

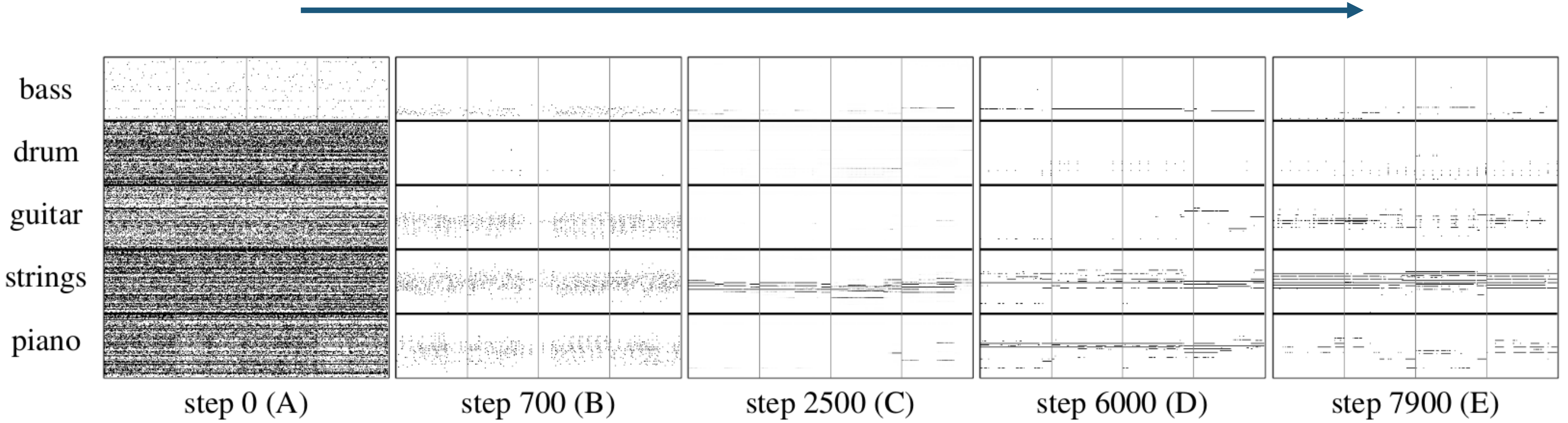
(Recap) Generative Adversarial Nets (GANs) – Generation



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) MuseGAN – A GAN for Pianorolls (Dong et al., 2018)

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

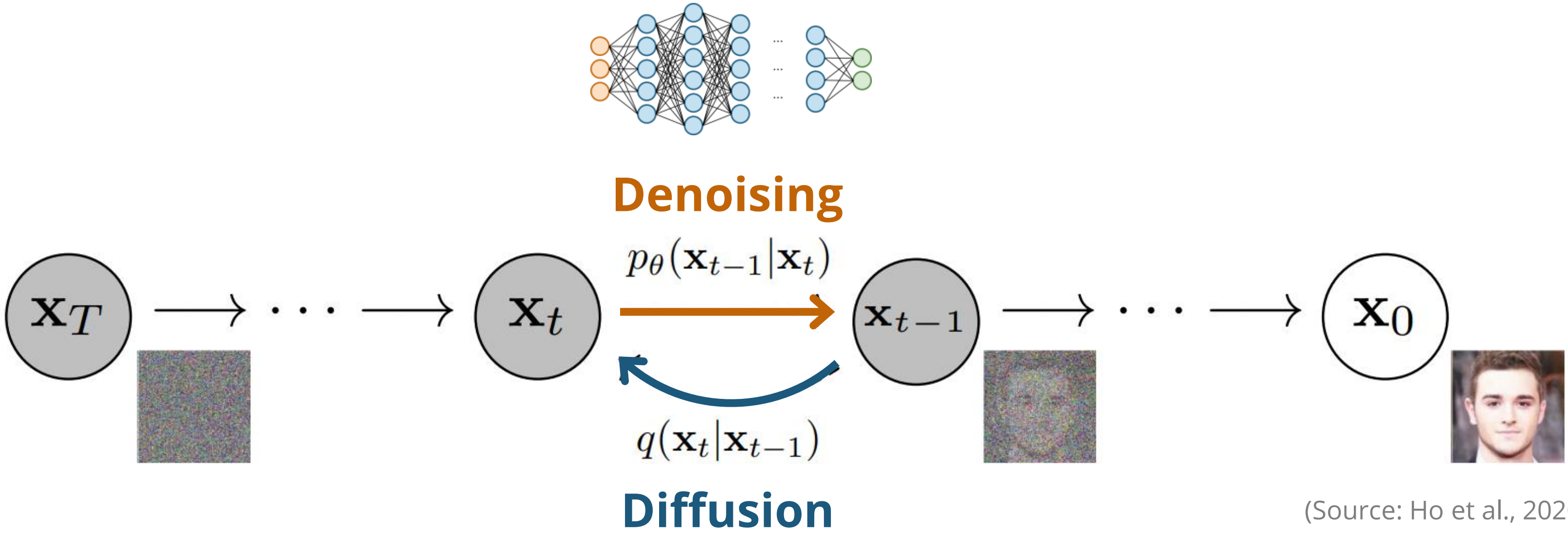


(Source: Dong et al., 2018)

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.

(Recap) Diffusion Models (Ho et al., 2020)

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

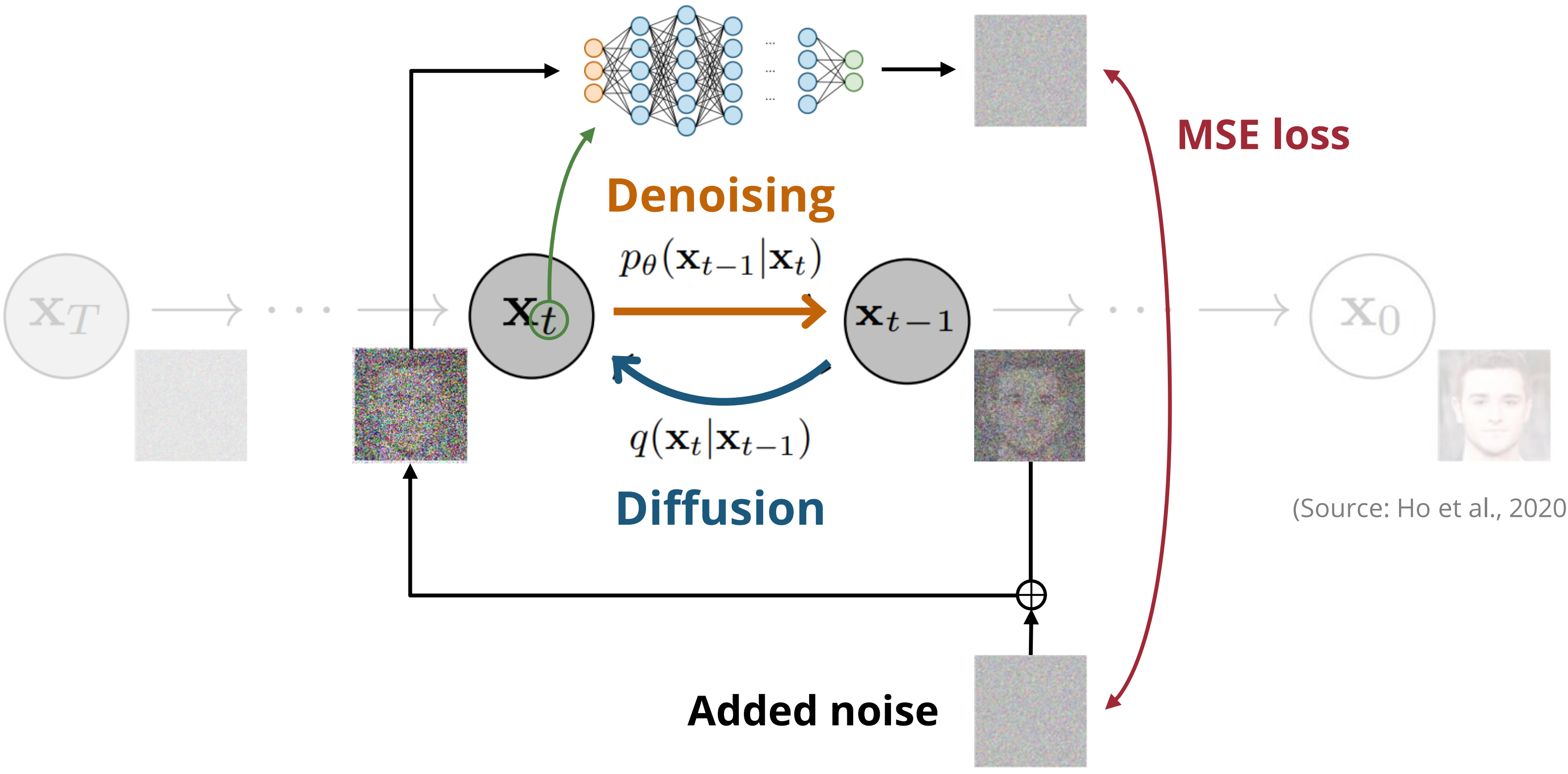


(Source: Ho et al., 2020)

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

(Recap) Diffusion Models – Training

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

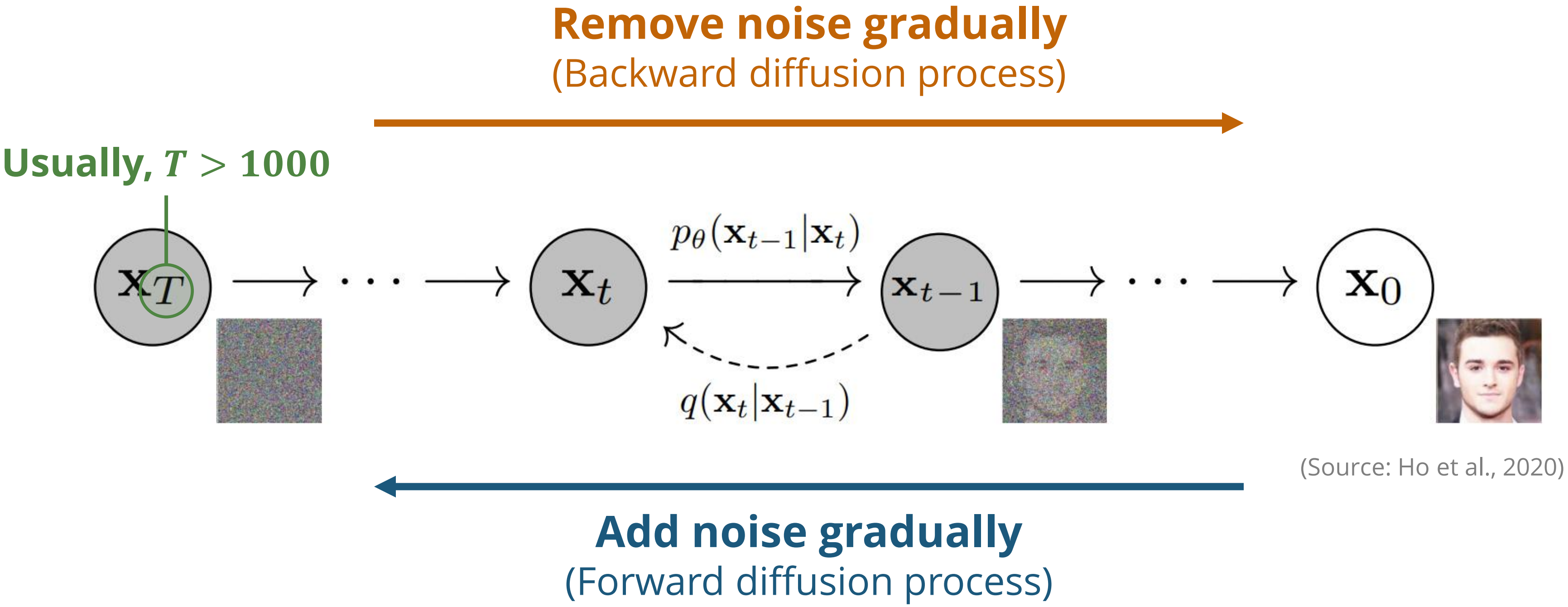


(Source: Ho et al., 2020)

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

(Recap) Diffusion Models (Ho et al., 2020)

- **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)

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

(Recap) Diffusion Models – Generation

Remove noise gradually
(Backward diffusion process)

Input

Output

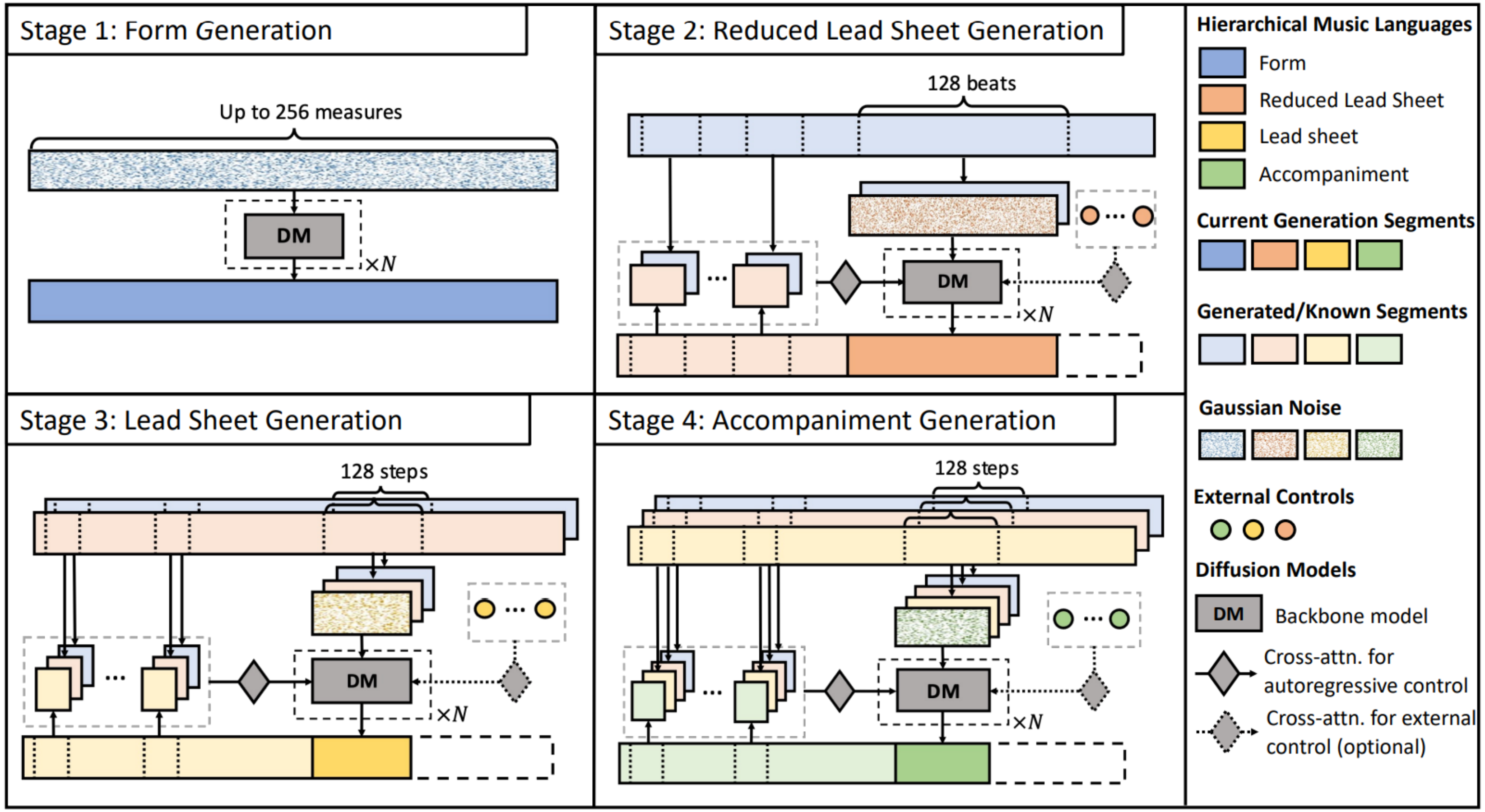
Coarse shapes
(low-frequency components)

Fine details
(high-frequency components)

(Source: Ho et al., 2020)

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

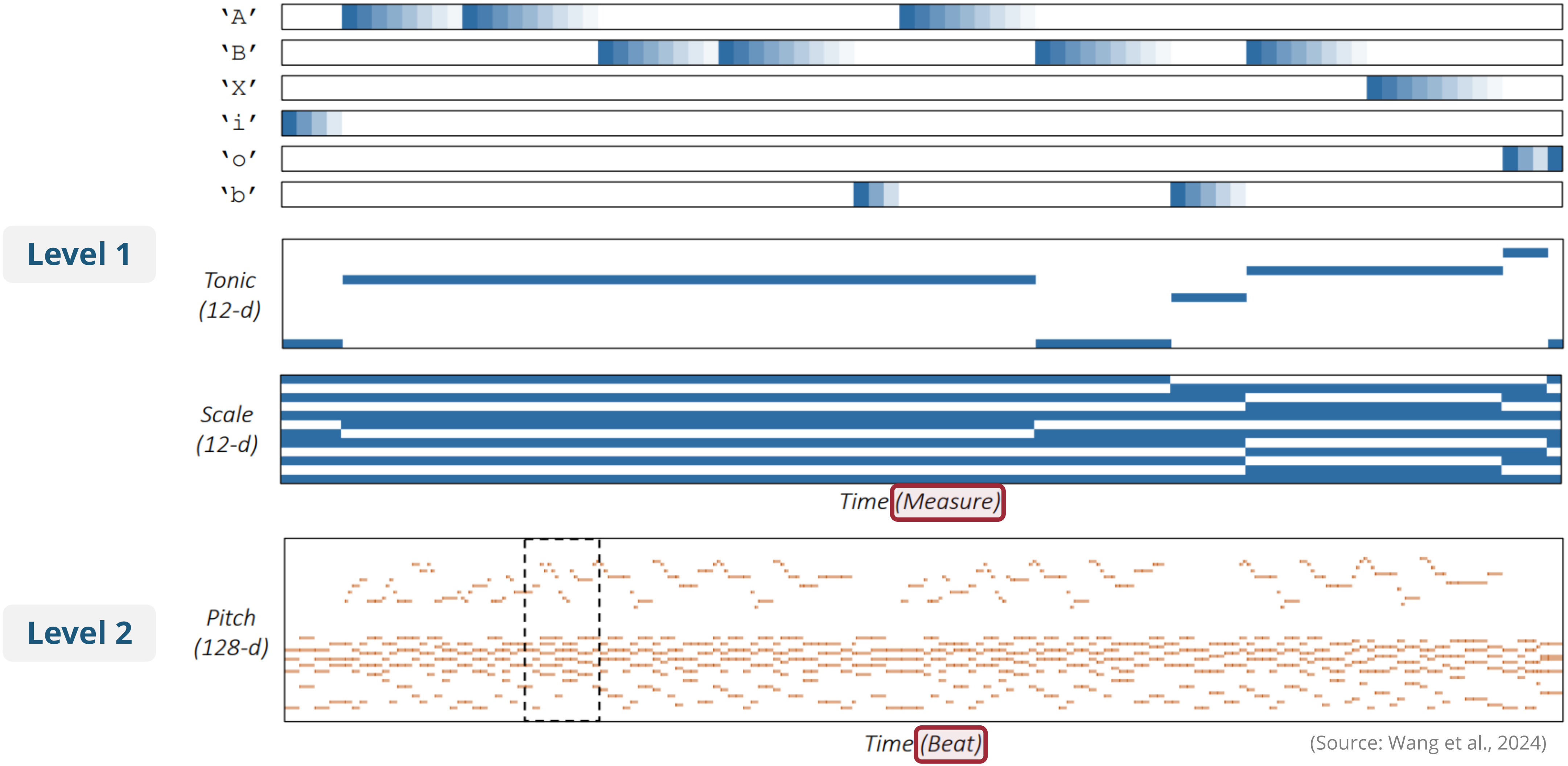
(Recap) Example: Cascaded Diffusion Models (Wang et al., 2024)



(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.

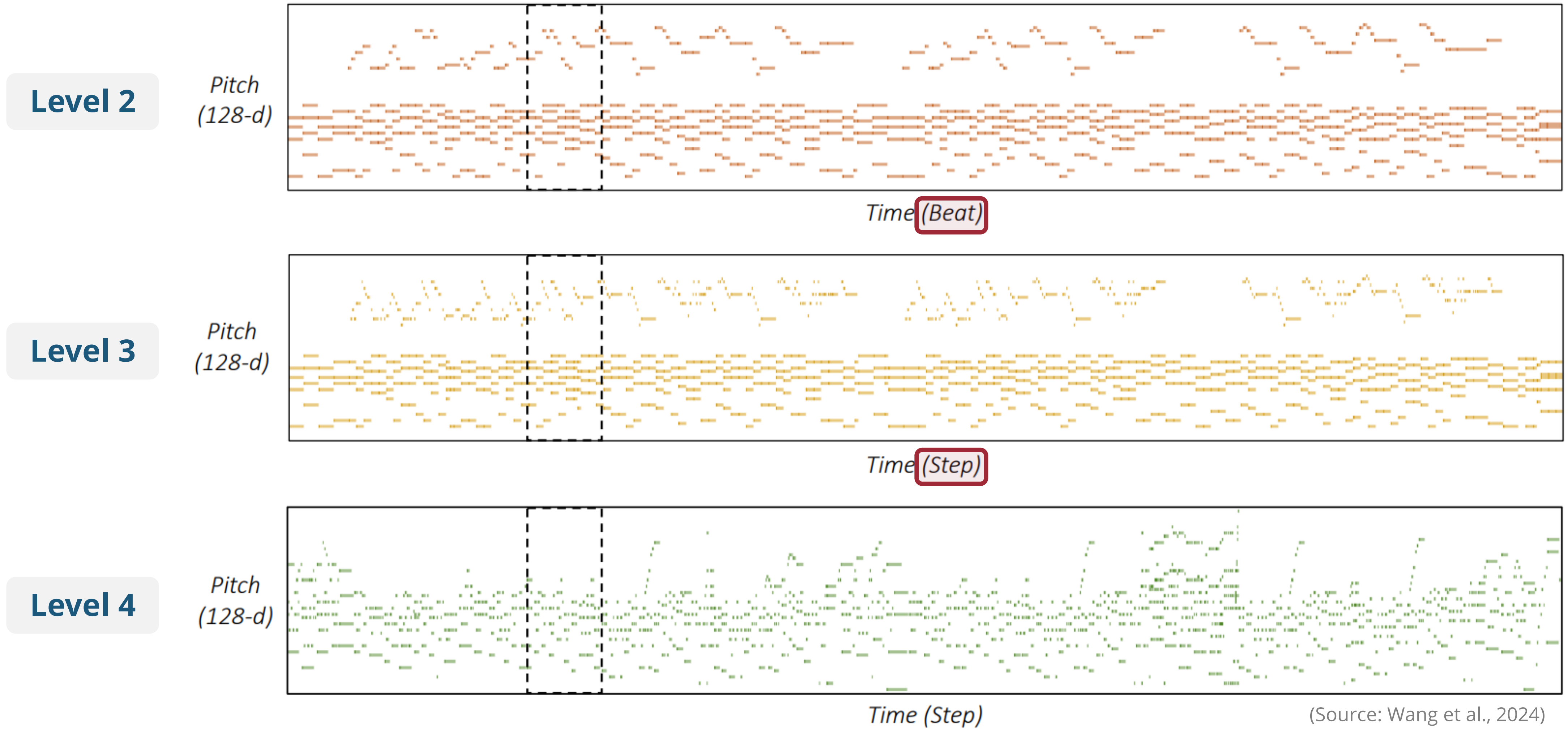
(Recap) Example: Cascaded Diffusion Models (Wang et al., 2024)



(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.

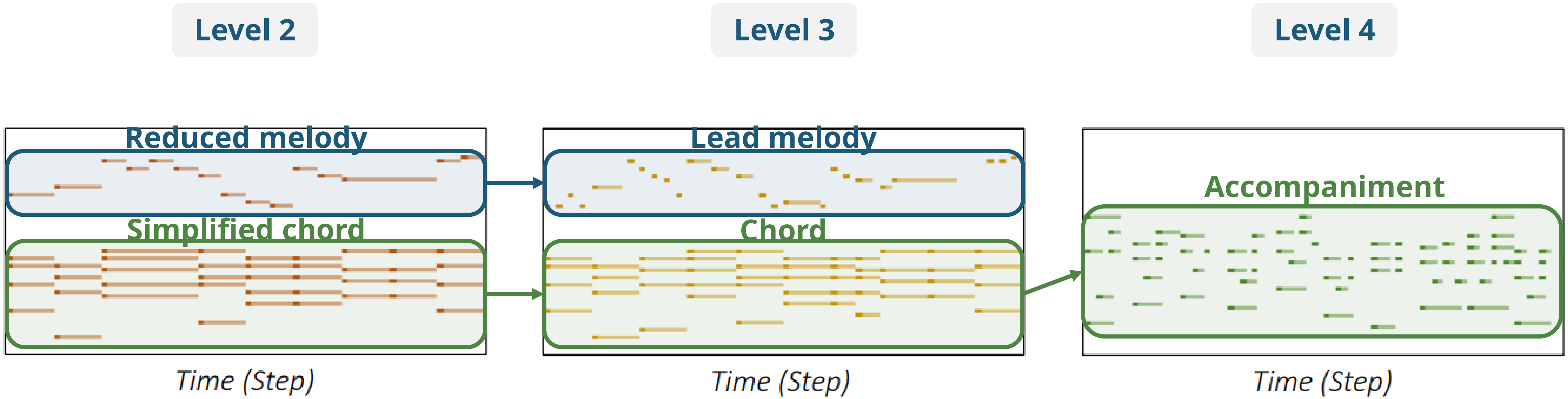
(Recap) Example: Cascaded Diffusion Models (Wang et al., 2024)



(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.

(Recap) Example: Cascaded Diffusion Models (Wang et al., 2024)



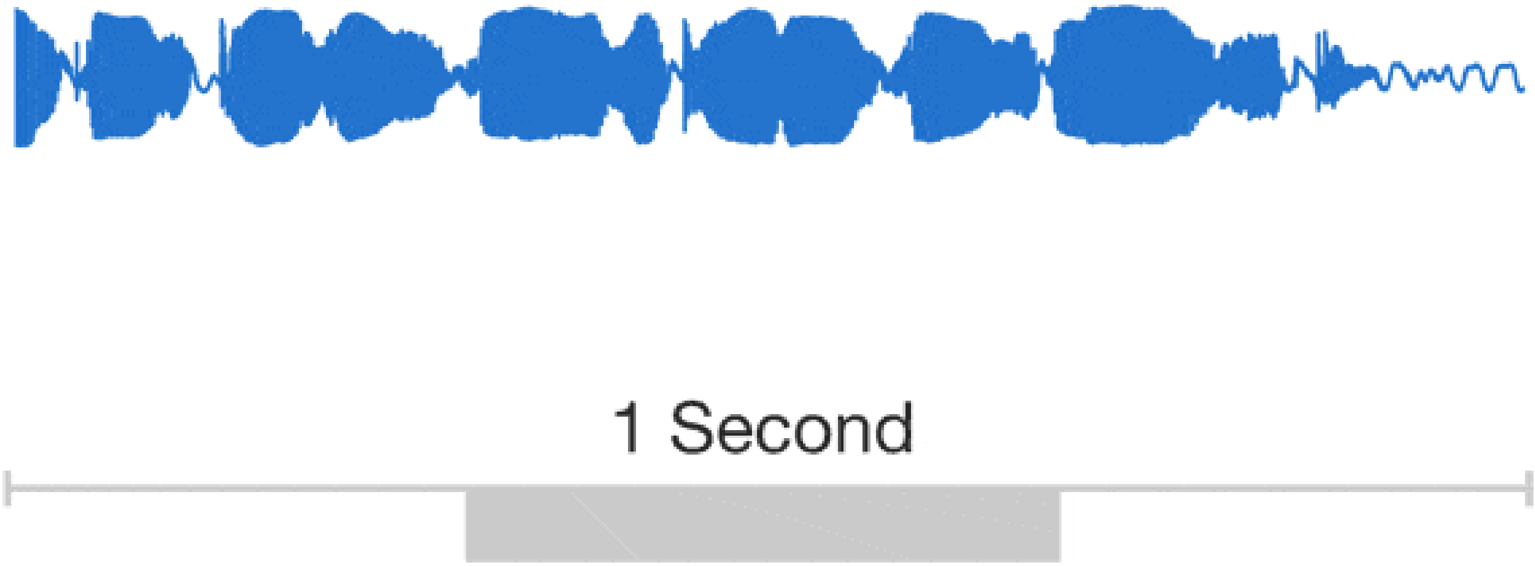
(Source: Wang et al., 2024)

wholesonggen.github.io

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

Autoregressive Waveform Synthesis

Generating Waveforms using a Neural Network

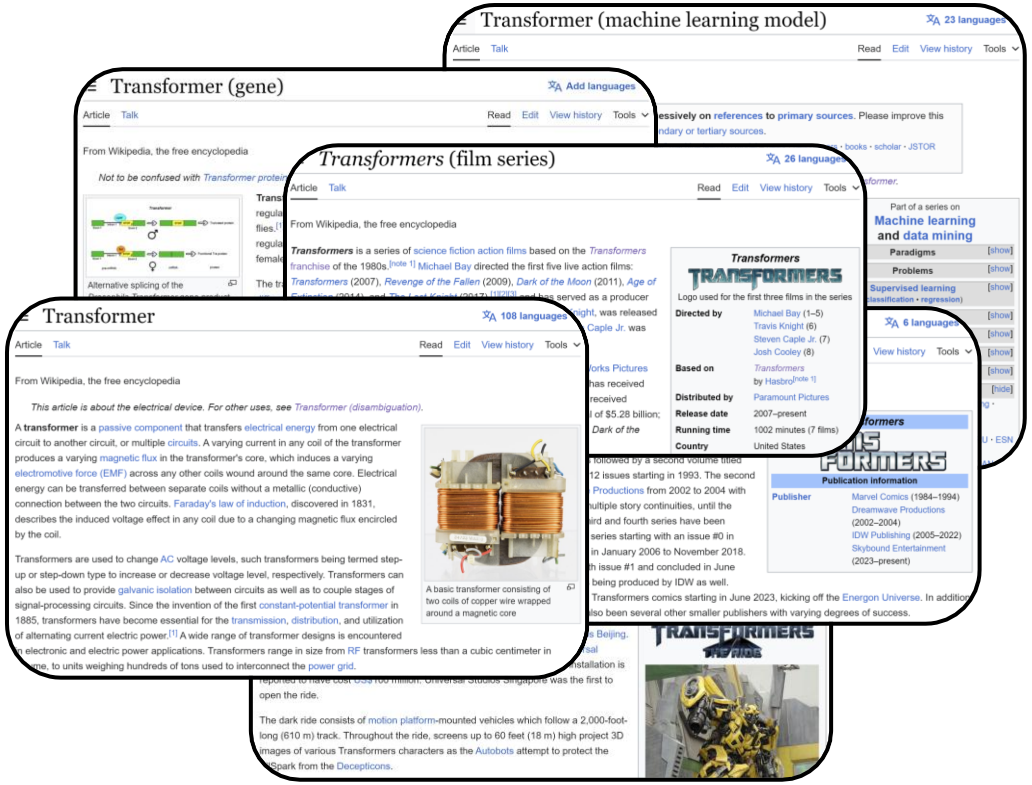


(Source: van den Oord et al., 2016)

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**



Transformer (machine learning model)

Transformer (gene)

Transformers (film series)

Transformers is a series of science fiction action films based on the *Transformers* franchise of the 1980s.[note 1] Michael Bay directed the first five live action films: *Transformers* (2007), *Revenge of the Fallen* (2009), *Dark of the Moon* (2011), …

Transformer

From Wikipedia, the free encyclopedia

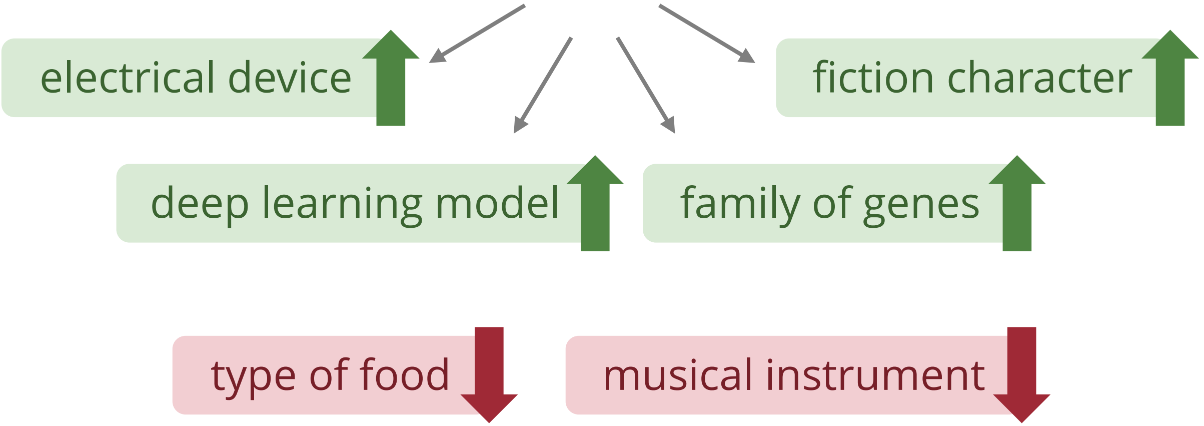
This article is about the electrical device. For other uses, see Transformer (disambiguation).

A **transformer** is a passive component that transfers electrical energy from one electrical circuit to another circuit, or multiple circuits. A varying current in any coil of the transformer produces a varying magnetic flux in the transformer's core, which induces a varying electromotive force (EMF) across any other coils wound around the same core. Electrical energy can be transferred between separate coils without a metallic (conductive) connection between the two circuits. Faraday's law of induction, discovered in 1831, describes the induced voltage effect in any coil due to a changing magnetic flux encircled by the coil.

Transformers are used to change AC voltage levels, such transformers being termed step-up or step-down type to increase or decrease voltage level, respectively. Transformers can also be used to provide galvanic isolation between circuits as well as to couple stages of signal-processing circuits. Since the invention of the first constant-potential transformer in 1885, transformers have become essential for the transmission, distribution, and utilization of alternating current electric power.[1] A wide range of transformer designs is encountered in electronic and electric power applications. Transformers range in size from RF transformers less than a cubic centimeter in volume, to units weighing hundreds of tons used to interconnect the power grid.

A basic transformer consisting of two coils of copper wire wrapped around a magnetic core

A transformer is a _________



electrical device ↑

fiction character ↑

deep learning model ↑

family of genes ↑

type of food ↓

musical instrument ↓

(Recap) Language Models (Mathematically)

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

$$P(\underset{\text{Next word}}{x_i} \mid \underbrace{x_1, x_2, \ldots, x_{i-1}}_{\text{Previous words}})$$

$P($ electrical $\mid$ A transformer is a $)$ ⬆

$P($ character $\mid$ A transformer is a $)$ ⬆

$P($ gene $\mid$ A transformer is a $)$ ⬆

$P($ model $\mid$ A transformer is a $)$ ⬆

$P($ food $\mid$ A transformer is a $)$ ⬇

$P($ musical $\mid$ A transformer is a $)$ ⬇

Autoregressive Models (Mathematically)

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

$$P(\underset{\text{Next number}}{x_i} \mid \underbrace{x_1, x_2, \dots, x_{i-1}}_{\text{Previous numbers}})$$

$P(\ 0.1 \mid 0.5, 0.4, 0.3, 0.2)$ ↑

$P(\ 0.09 \mid 0.5, 0.4, 0.3, 0.2)$ ↑

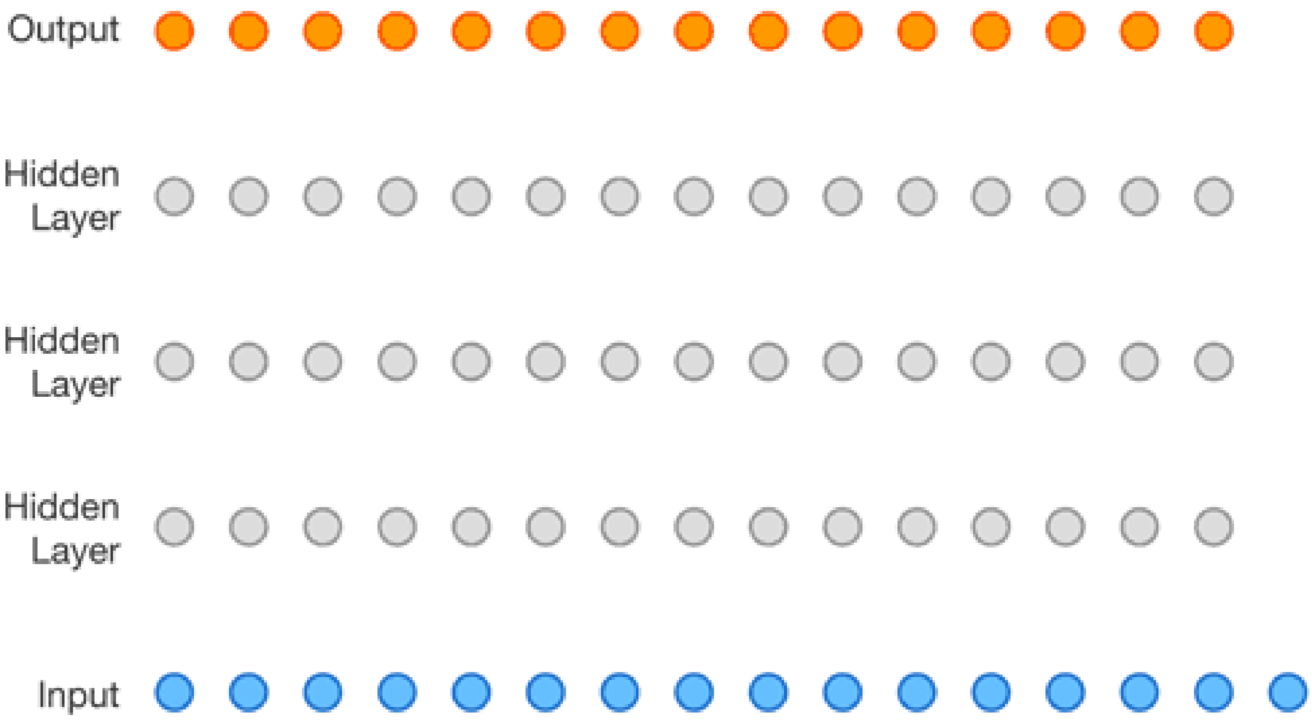
$P(\ 0.11 \mid 0.5, 0.4, 0.3, 0.2)$ ↑

$P(\ 99 \mid 0.5, 0.4, 0.3, 0.2)$ ↓

$P(\ -1 \mid 0.5, 0.4, 0.3, 0.2)$ ↓

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

WaveNet (van den Oord et al., 2016)



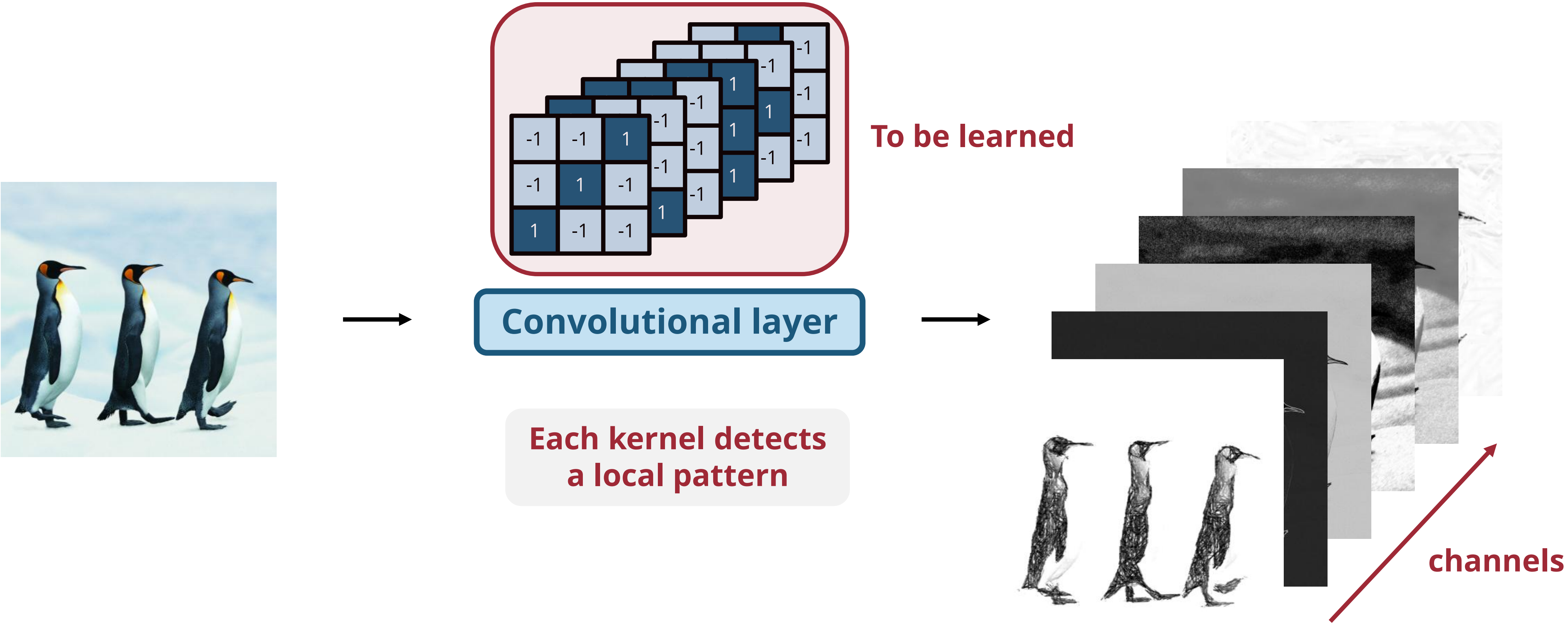
(Source: van den Oord et al., 2016)

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)



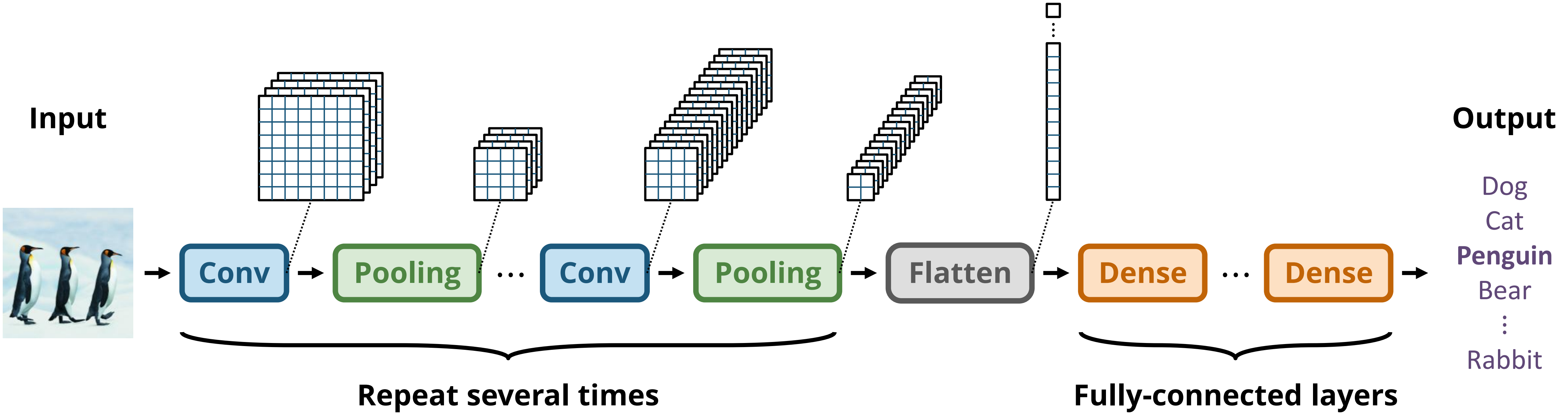
To be learned

Convolutional layer

Each kernel detects a local pattern

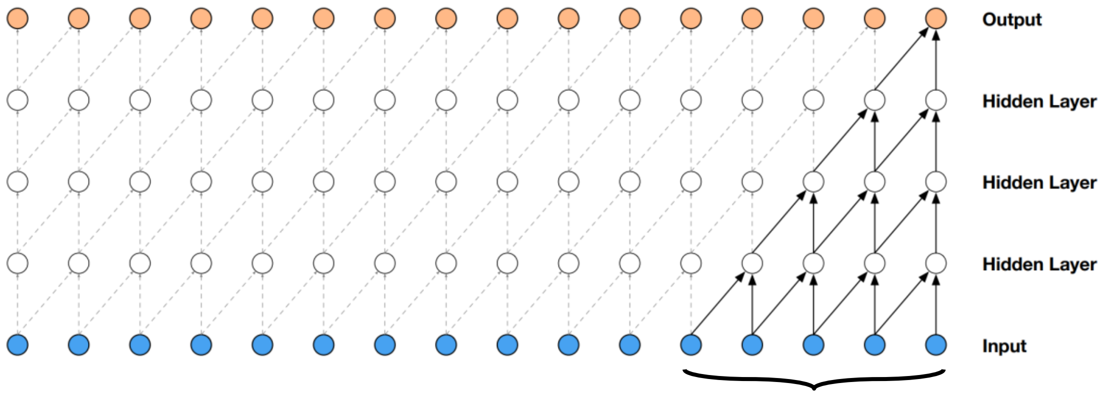
channels

(Recap) Convolutional Neural Network (CNNs)



WaveNet (van den Oord et al., 2016)

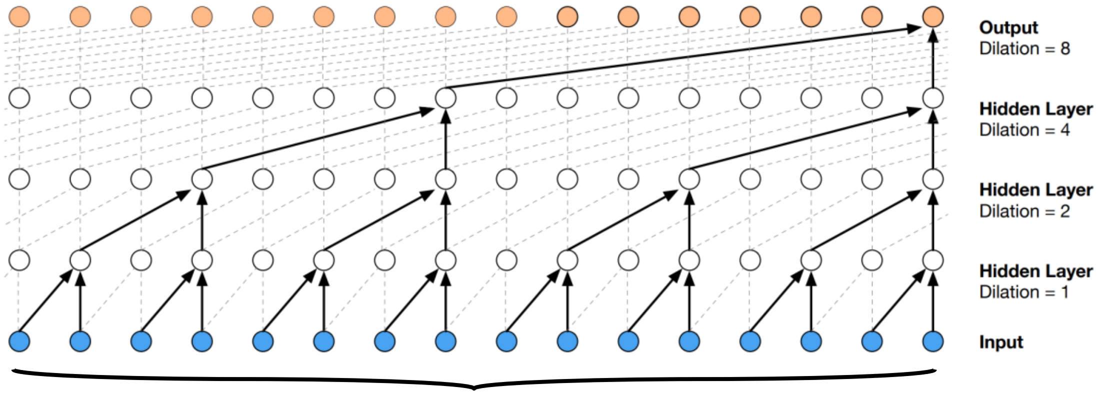
Standard convolution



Smaller "receptive field"

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

Dilated convolution



Larger "receptive field"

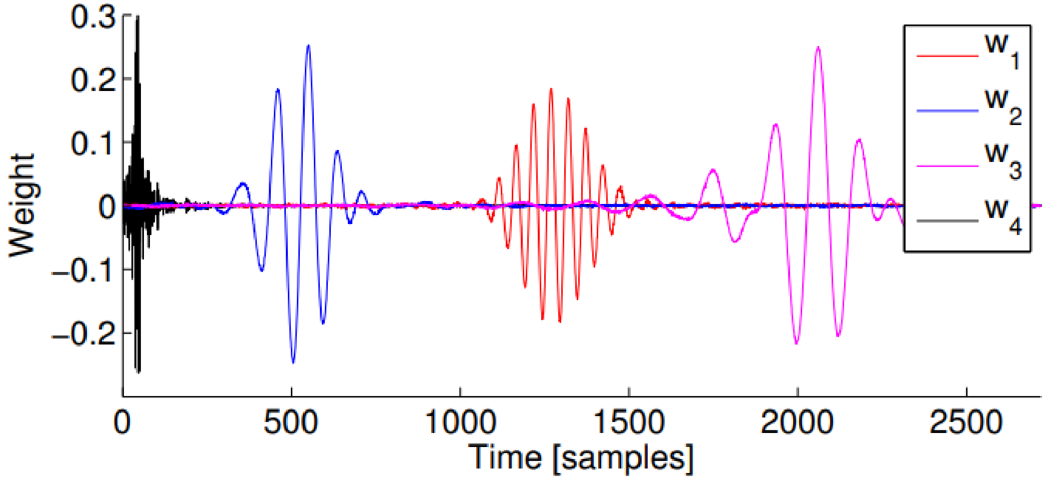
Example of generated music

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

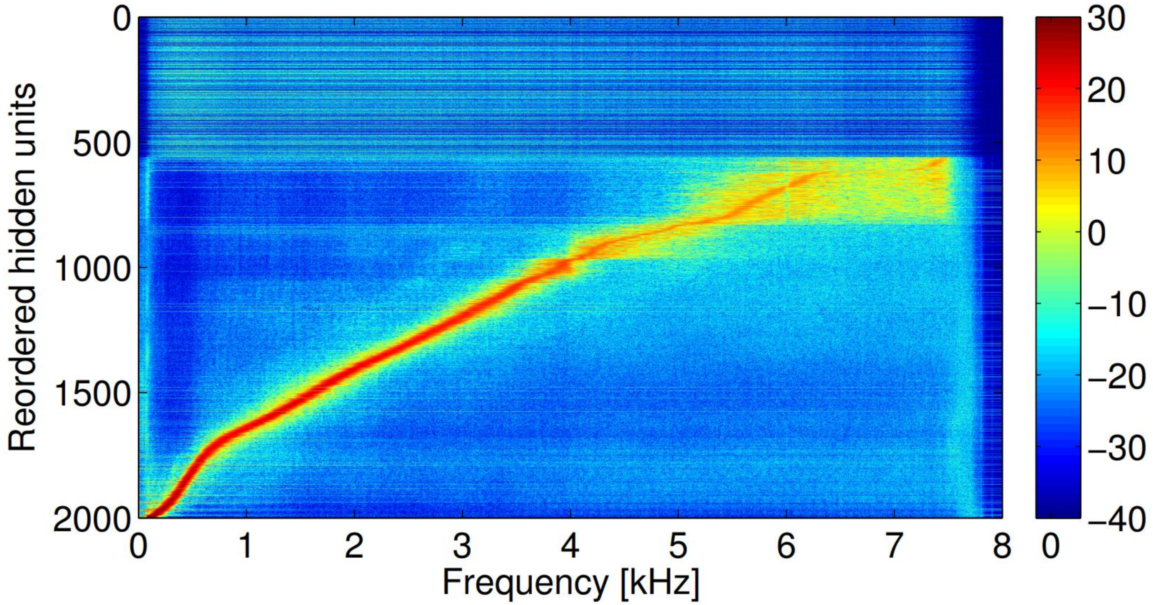
Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *ICML*, 2016.

1D CNNs & Fourier Transform

Convolution kernels learned



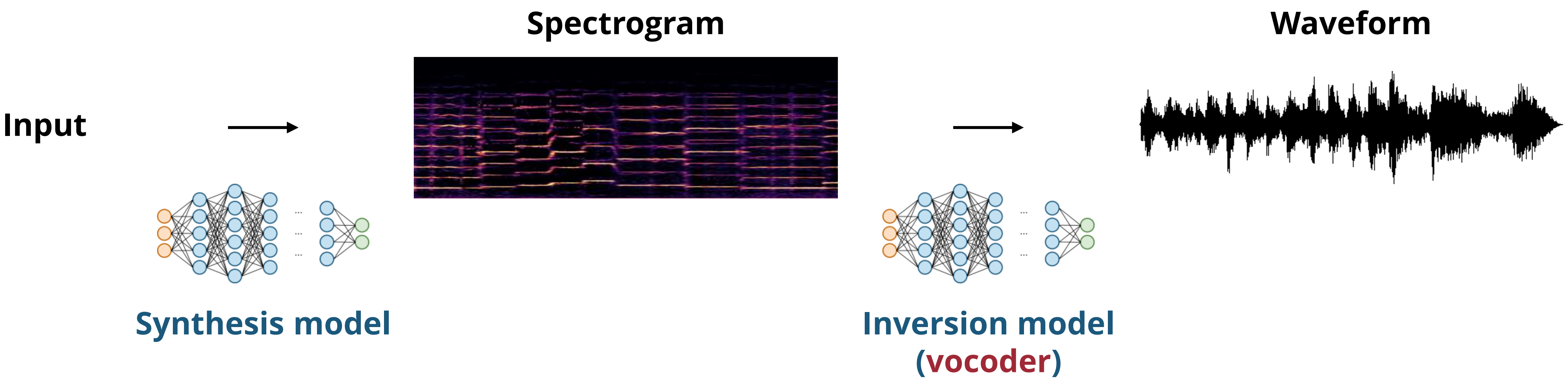
Peak frequency detected by the learned kernels



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

Frequency-domain Audio Synthesis

Frequency-domain Audio Synthesis



Spectrogram

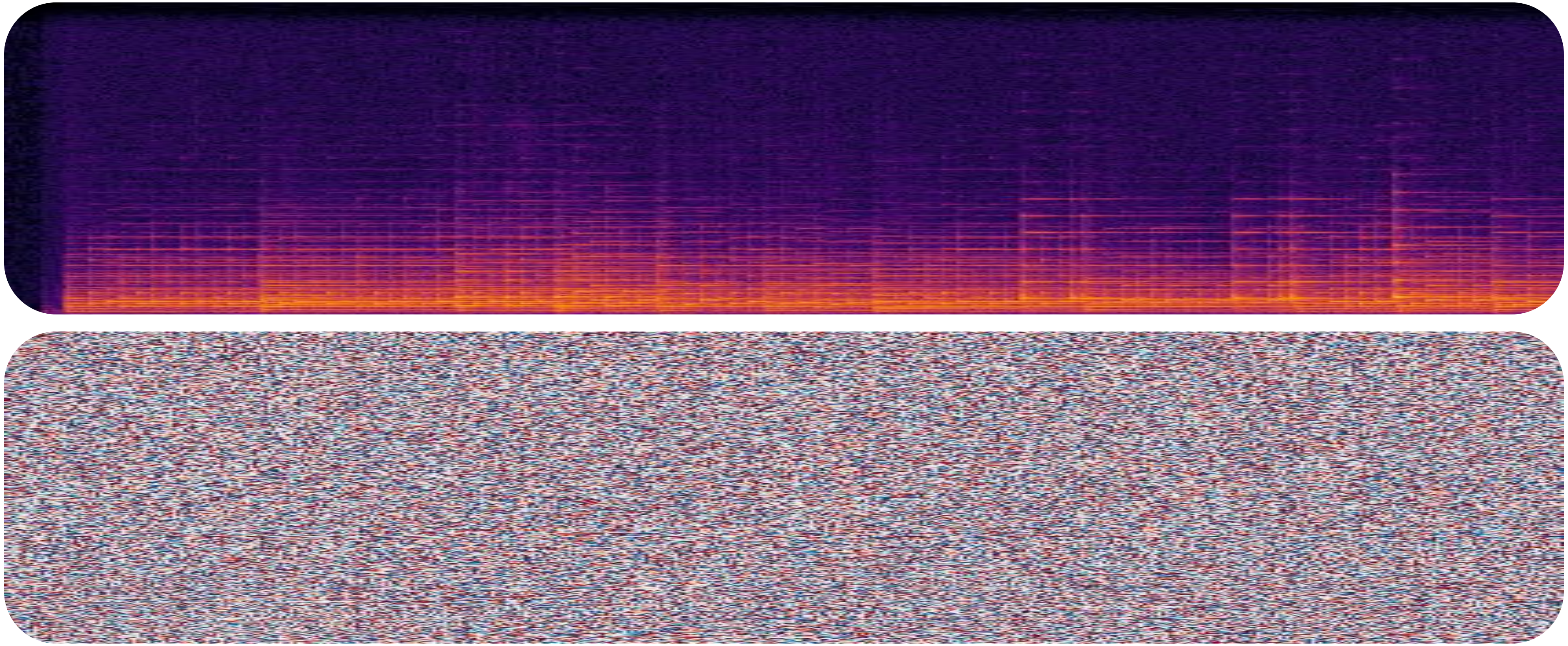
Waveform

Input

Synthesis model

Inversion model (vocoder)

Importance of the Phase Information

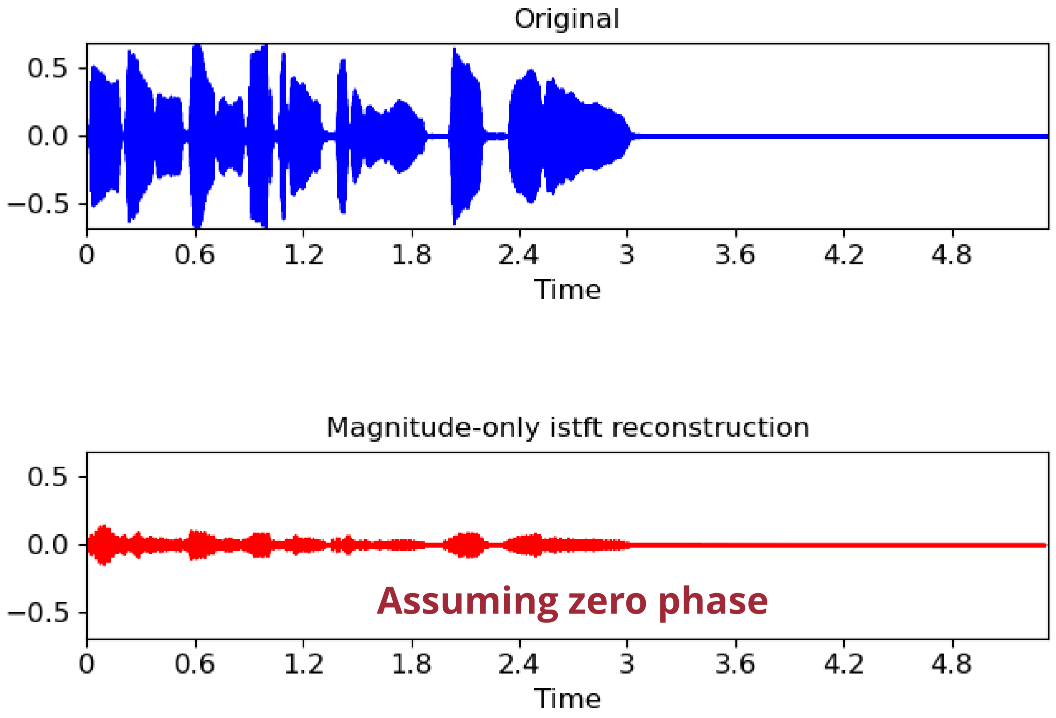


(Source: Dieleman et al., 2020)

Real phase **Random phase**

Sander Dieleman, "Generating music in the waveform domain," *Sander Dieleman's Blog*, March 24, 2020.

Inverse STFT without Phase Information



Assuming zero phase

(Source: librosa documentation)

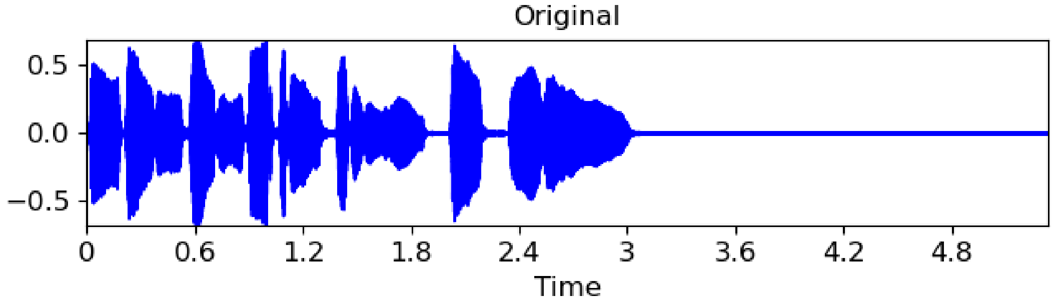
Complex-valued STFT matrix

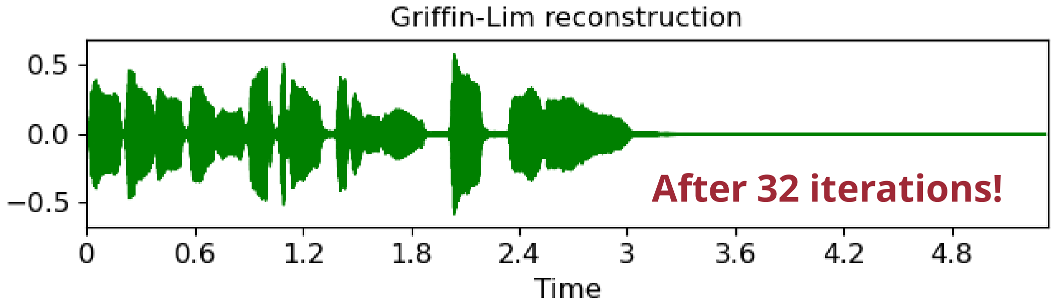
$$\mathrm{ISTFT}(\boxed{M}) = \arg\min_{y}\ (M - \mathrm{STFT}(y))^2$$

Find the signal y that minimize the MSE between the input and STFT(y)

librosa.org/doc/main/generated/librosa.griffinlim.html

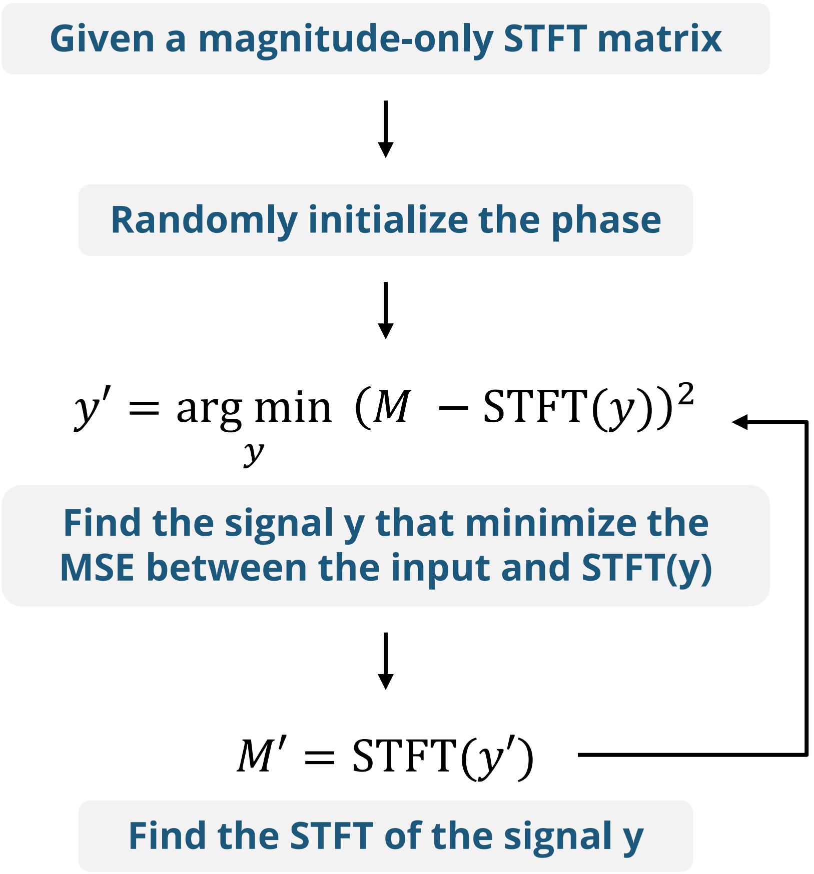
Griffin-Lim Algorithm (Griffin & Lim, 1984)


Original


Griffin-Lim reconstruction

After 32 iterations!

(Source: librosa documentation)



Given a magnitude-only STFT matrix

Randomly initialize the phase

$$y' = \arg\min_{y} \; (M \; - \mathrm{STFT}(y))^2$$

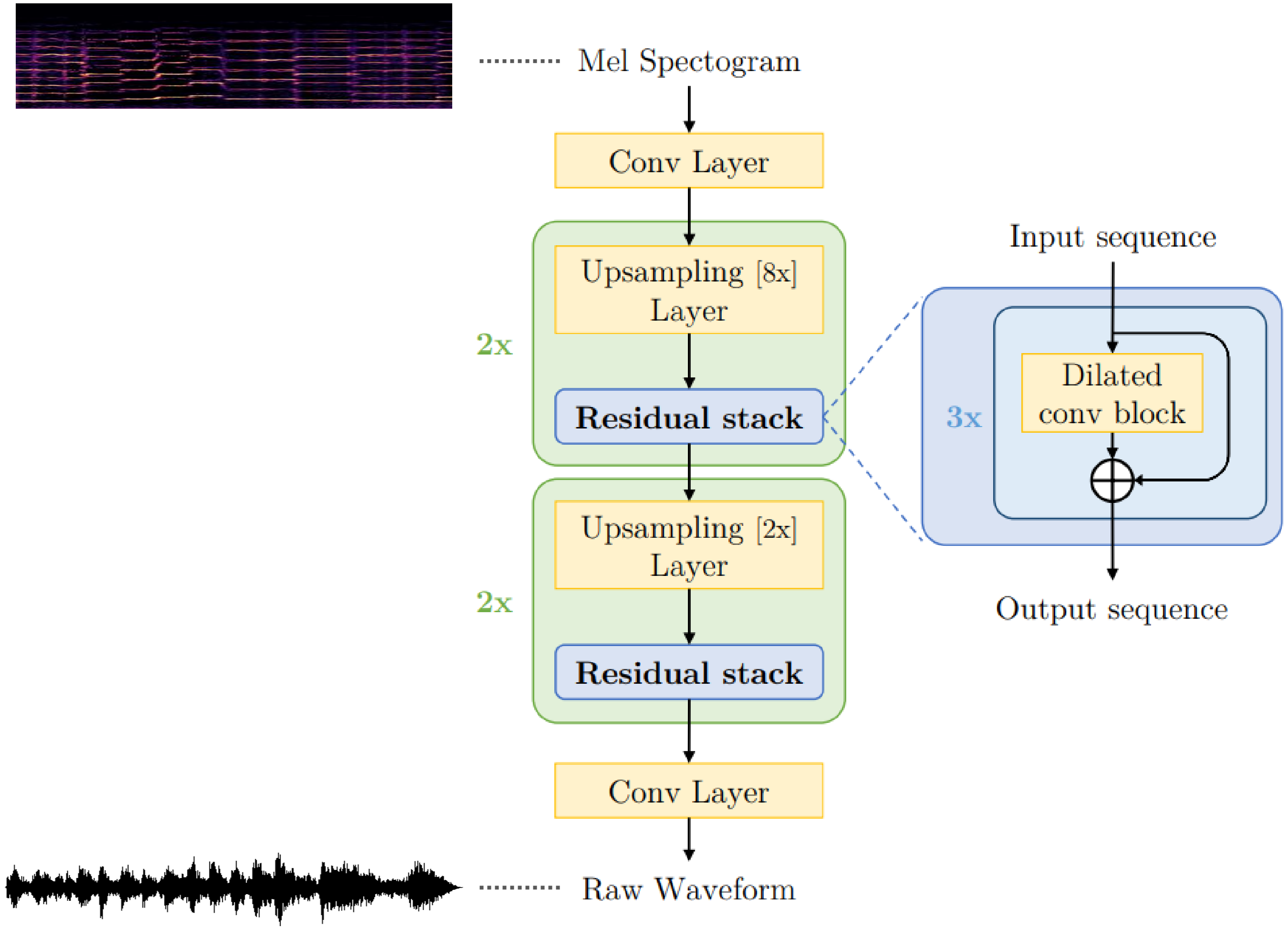
Find the signal y that minimize the MSE between the input and STFT(y)

$$M' = \mathrm{STFT}(y')$$

Find the STFT of the signal y

Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE TASSP*, 1984.
librosa.org/doc/main/generated/librosa.griffinlim.html

MelGAN (Kumar et al., 2019)

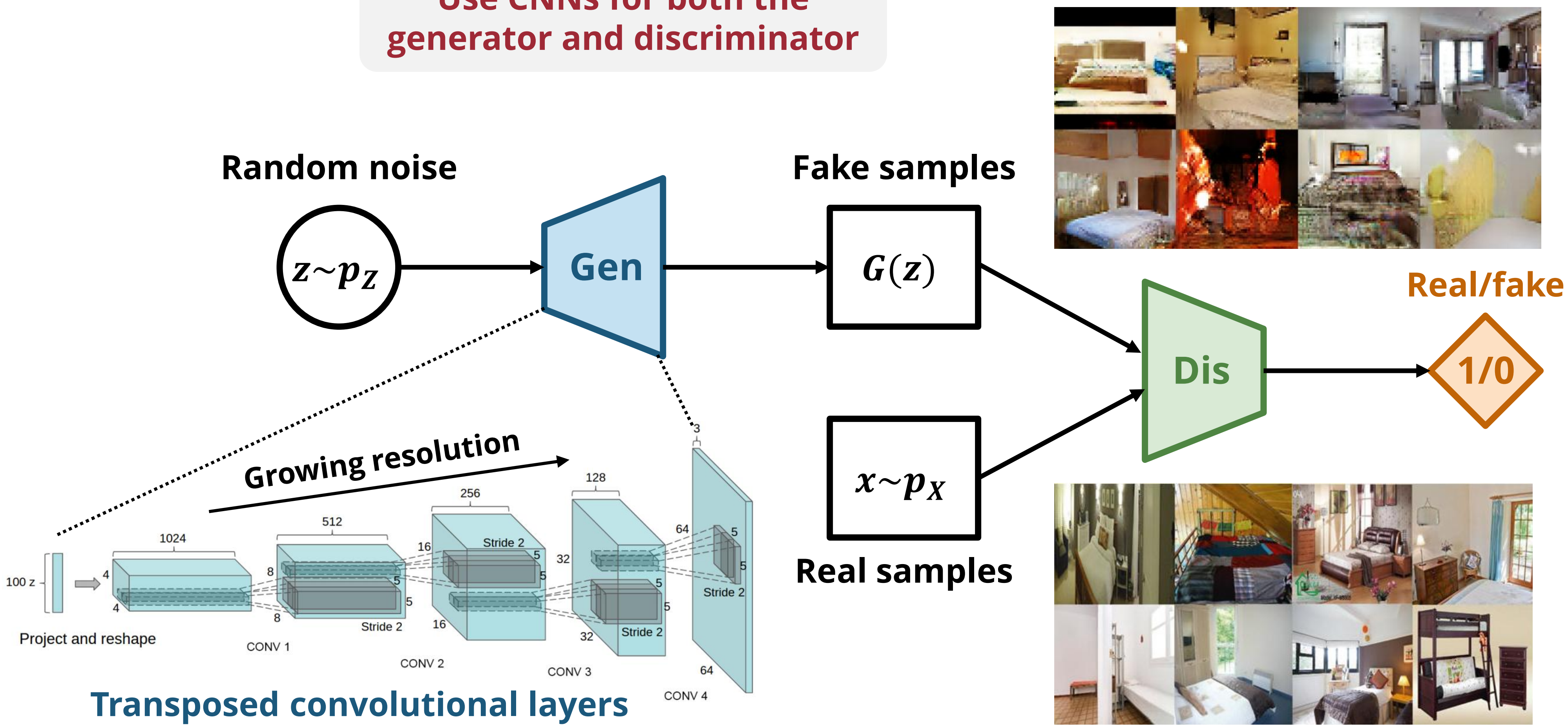


(Source: Kumar et al., 2019)

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.

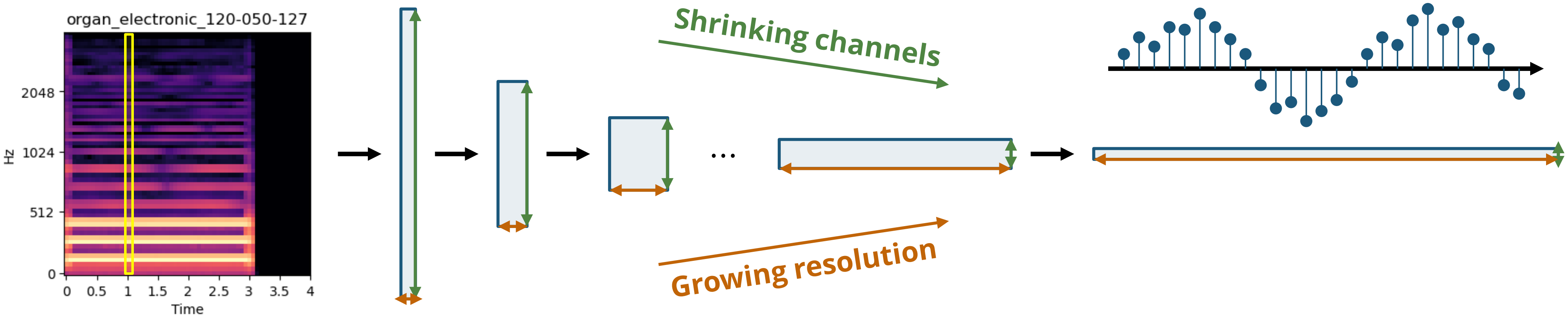
(Recap) Deep Convolutional GANs (DCGANs)

Use CNNs for both the generator and discriminator

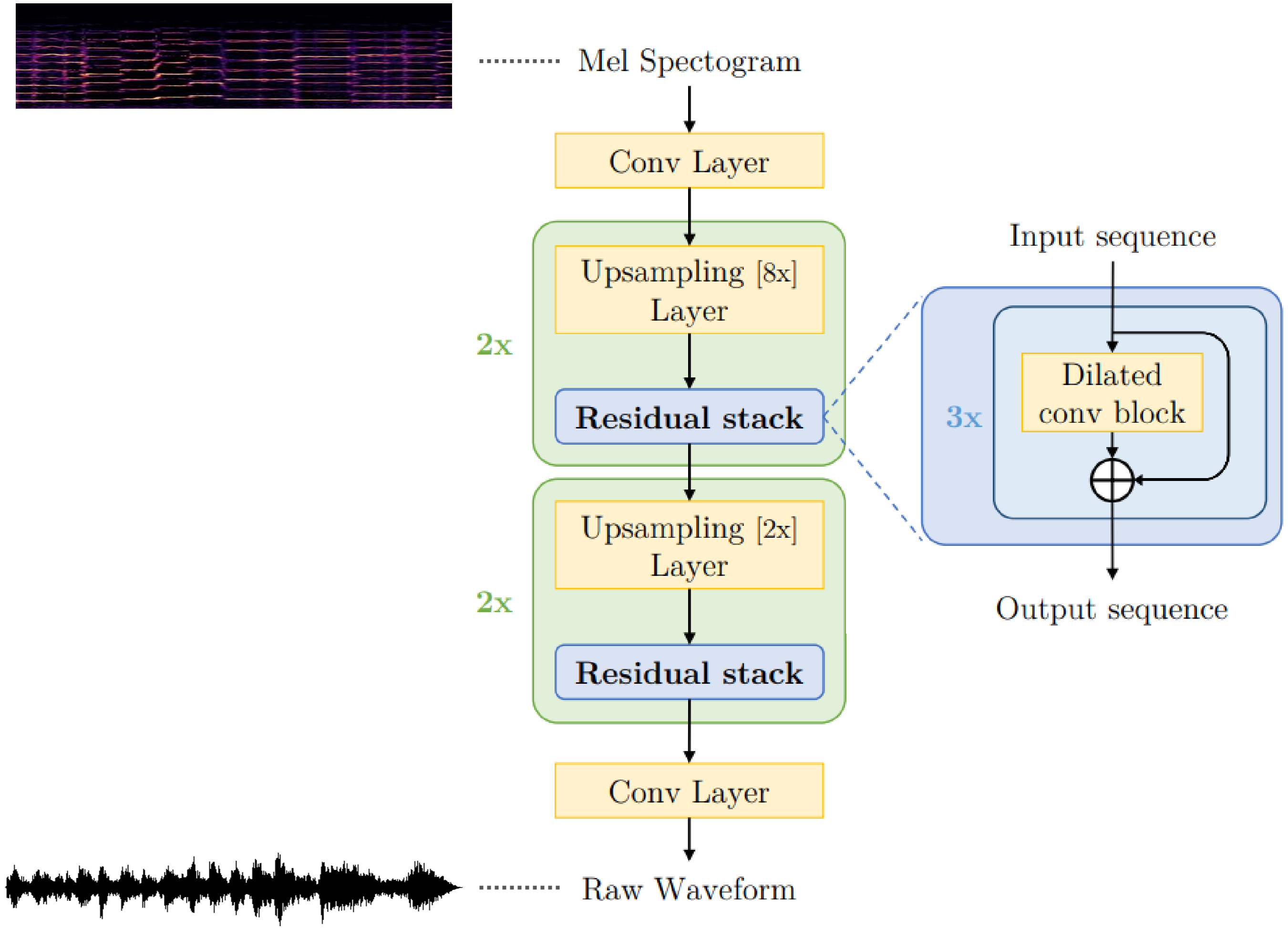


Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *ICLR*, 2016.

Upsampling for Vocoders



MelGAN (Kumar et al., 2019)

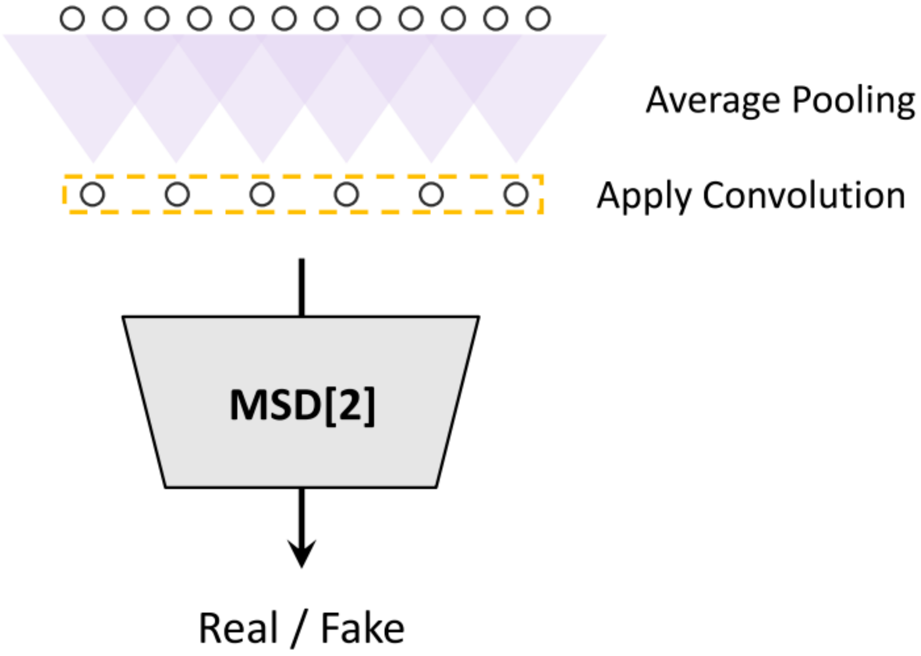


(Source: Kumar et al., 2019)

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.

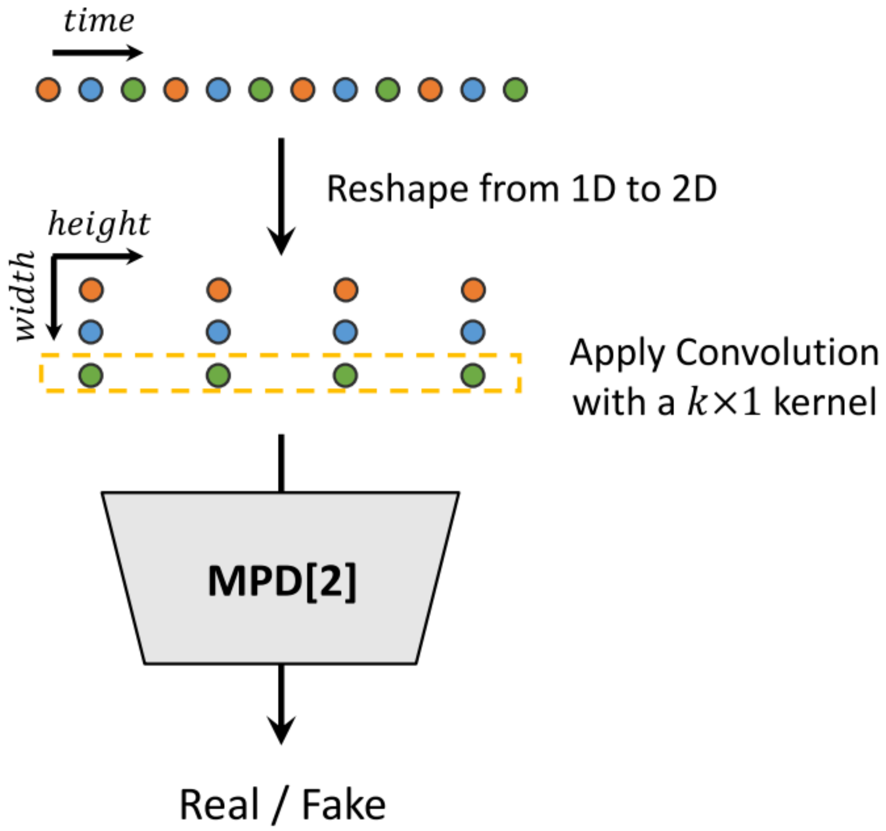
MelGAN (Kumar et al., 2019)

Multi-scale discriminator



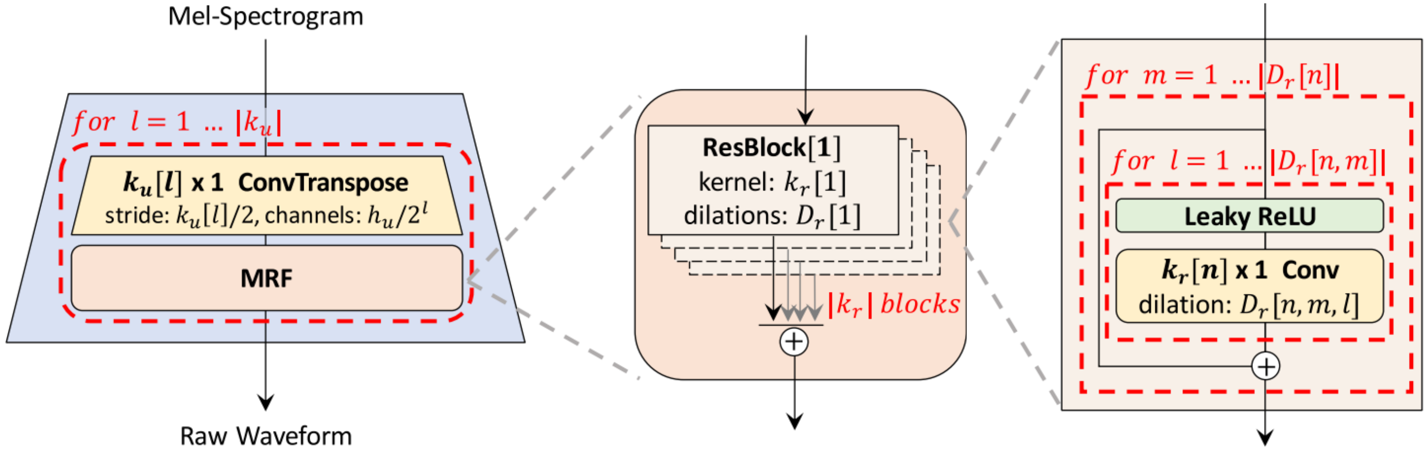
(Source: Kong et al., 2019)

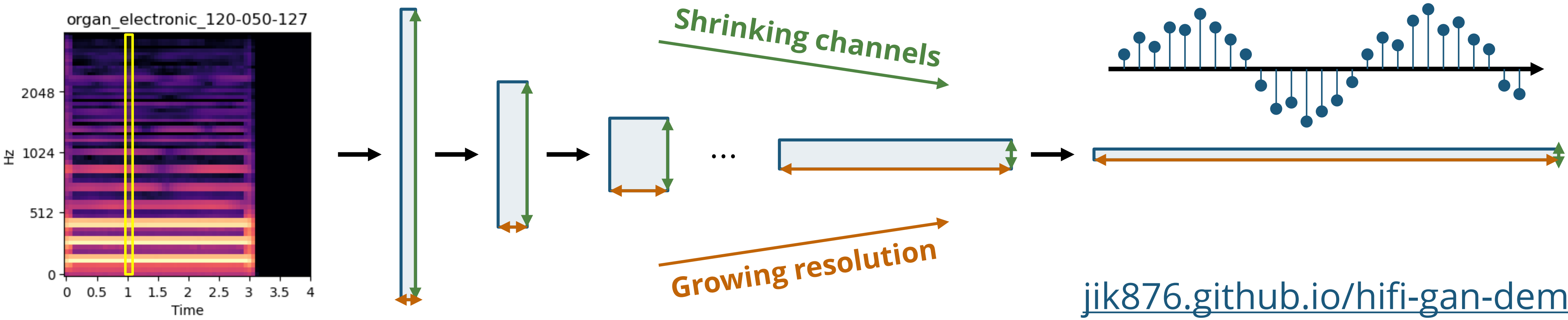
Multi-period discriminator



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.

Hifi-GAN (Kong et al., 2020)



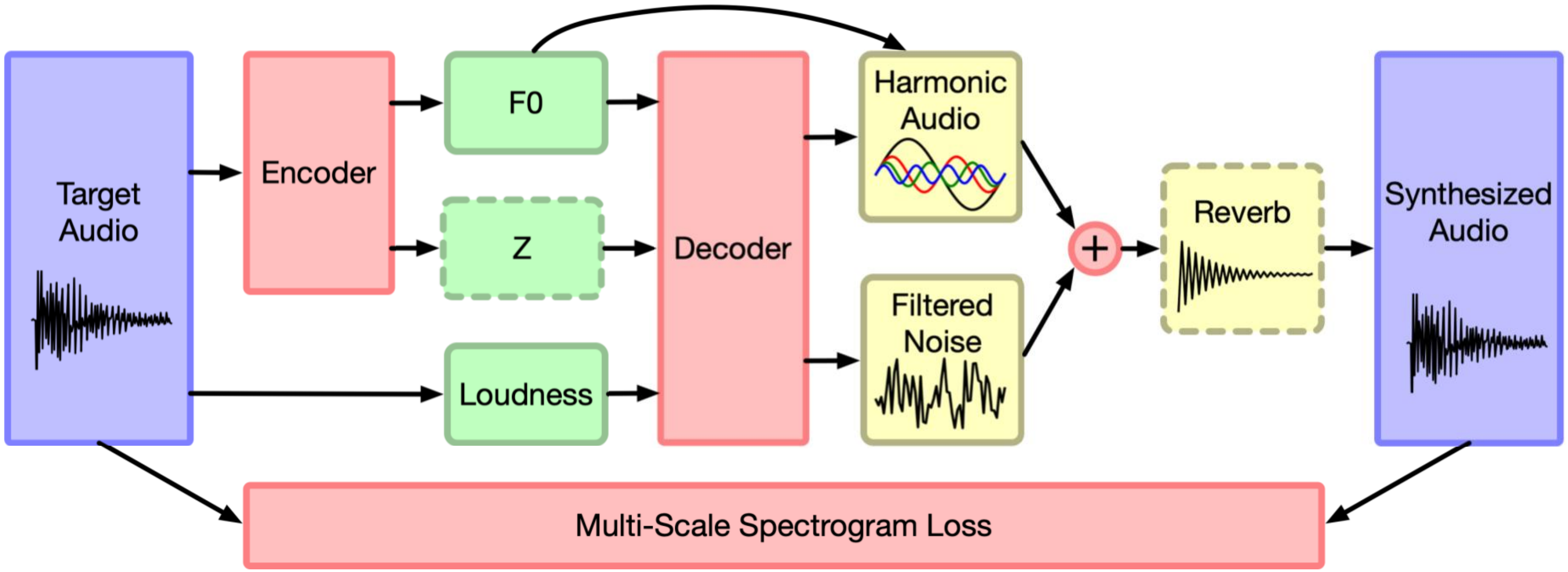


jik876.github.io/hifi-gan-demo

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

Differentiable DSP

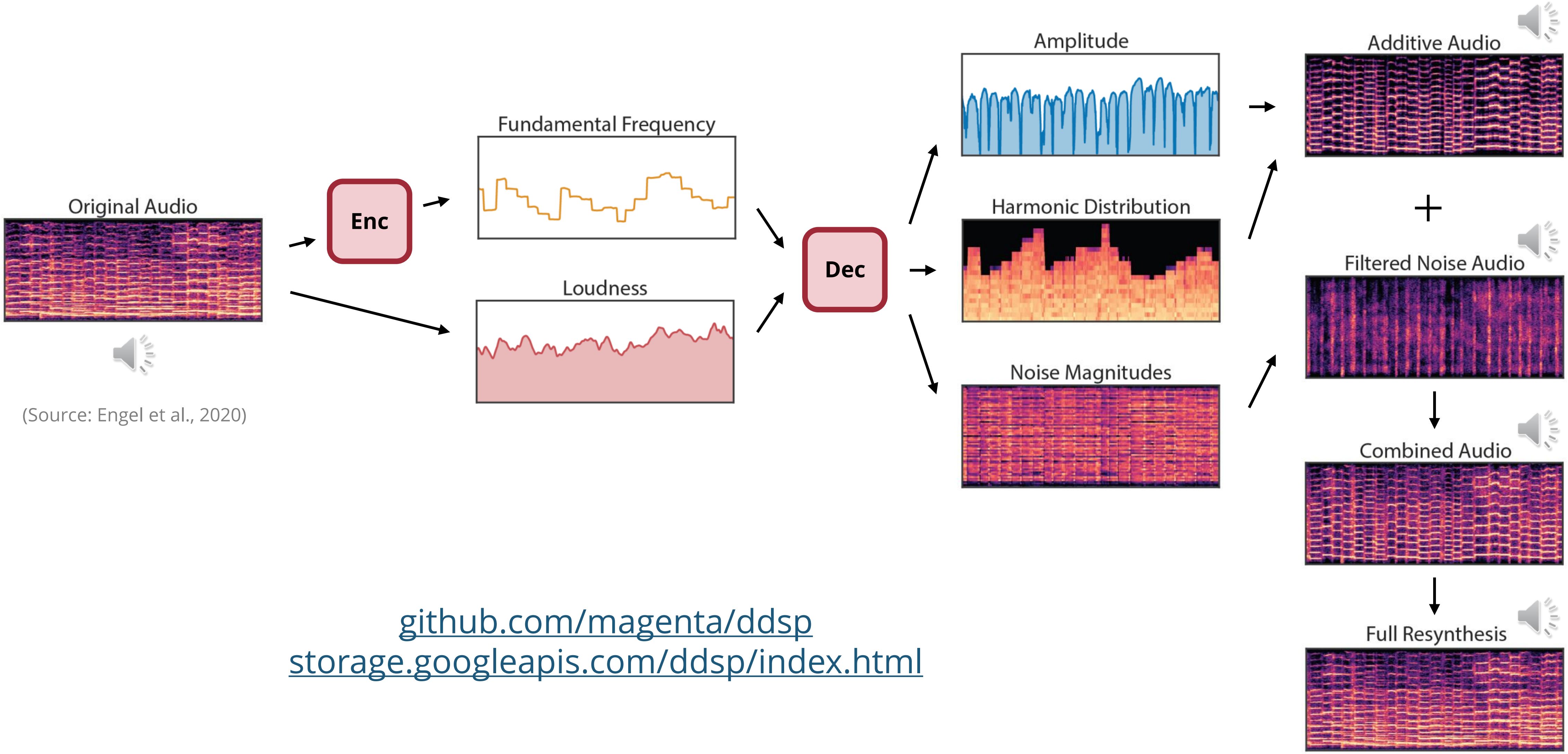
Differentiable DSP (DDSP) (Engel et al., 2020)



(Source: Engel et al., 2020)

Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, Adam Roberts, “DDSP: Differentiable Digital Signal Processing,” *ICLR*, 2020.

Differentiable DSP (DDSP) (Engel et al., 2020)



(Source: Engel et al., 2020)

github.com/magenta/ddsp
storage.googleapis.com/ddsp/index.html

Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, Adam Roberts, "DDSP: Differentiable Digital Signal Processing," *ICLR*, 2020.

Yaboi Hanoi: Entering Demons & Gods (2022)

youtu.be/PbrRoR3nEVw

soundcloud.com/yaboihanoi/enter-demons-and-gods

Optional Reading

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