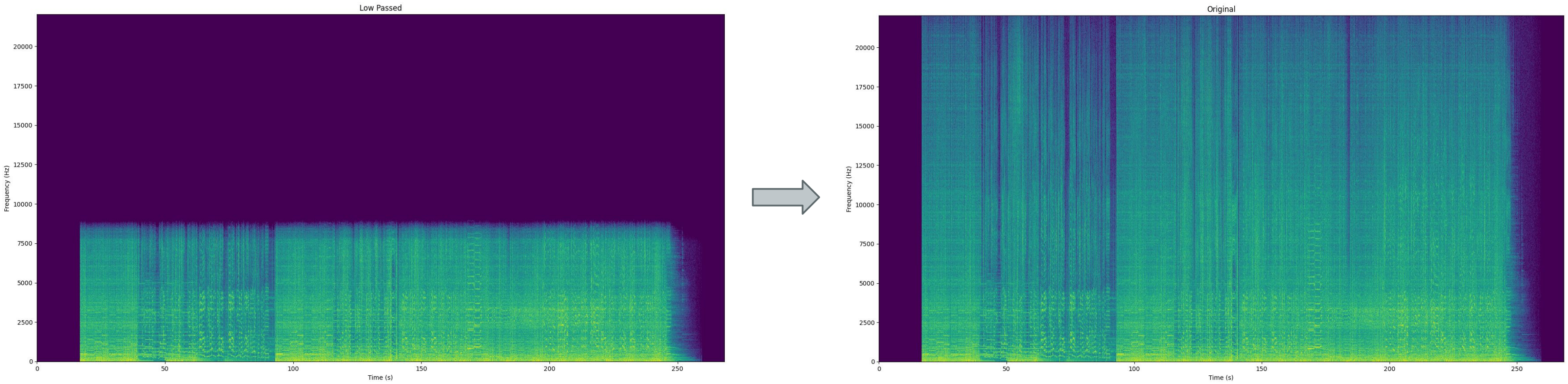
GAN-Based Bandwidth Extension for Music

Cassius Close

What is Bandwidth Extension (BWE)?

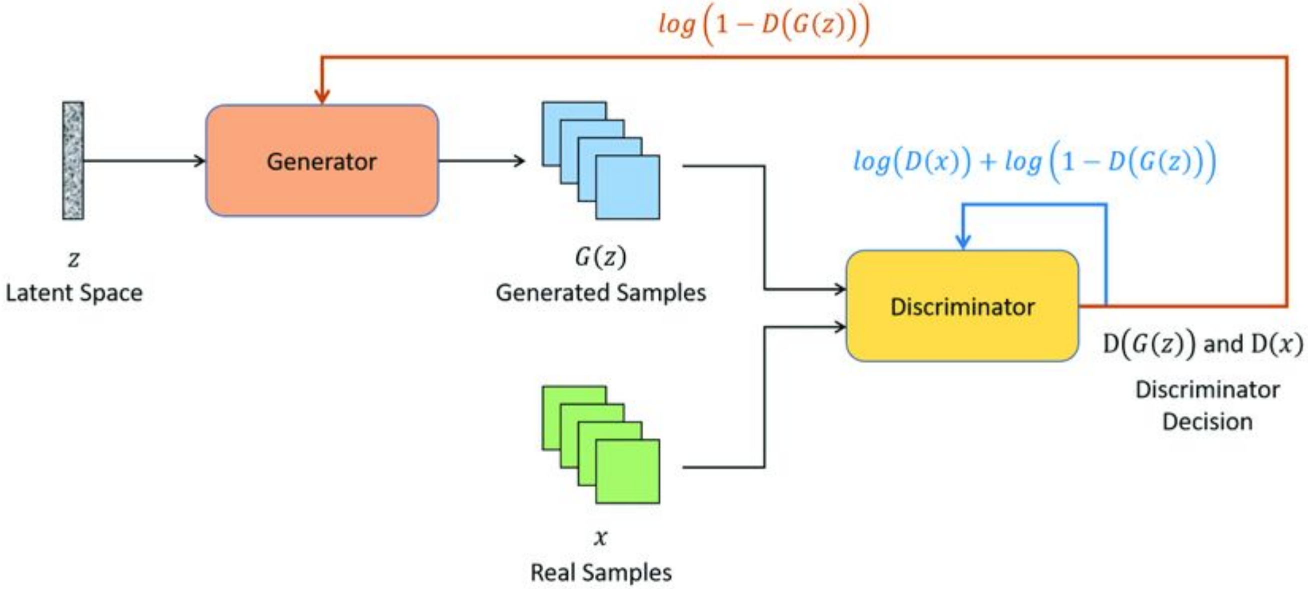
- Guess high frequency content from a low-resolution signal
- Applications
  - Telephone systems
  - Old recordings w/ missing high-freq content (music & speech)



Ideal Bandwidth Extension for a sample pop song

Generative Adversarial Networks

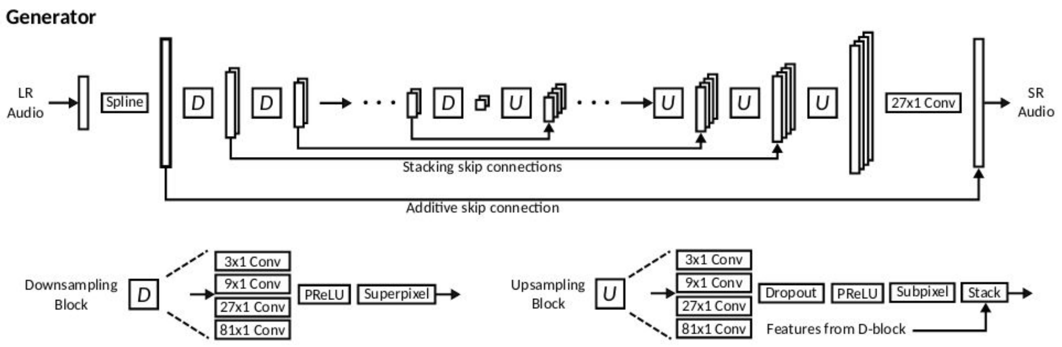
- Generator tries to maximize loss, discriminator tries to minimize loss
- More detailed output



https://www.researchgate.net/figure/Typical-Generative-Adversarial-Networks-GAN-architecture_fig2_349182009

Existing Methods

- Three recent methods work on music
  - All use U-Net architecture
- Recently, a HiFi-GAN based approach (BWE Is All You Need) has good results for speech
  - Uses WaveNet-based architecture for the generator



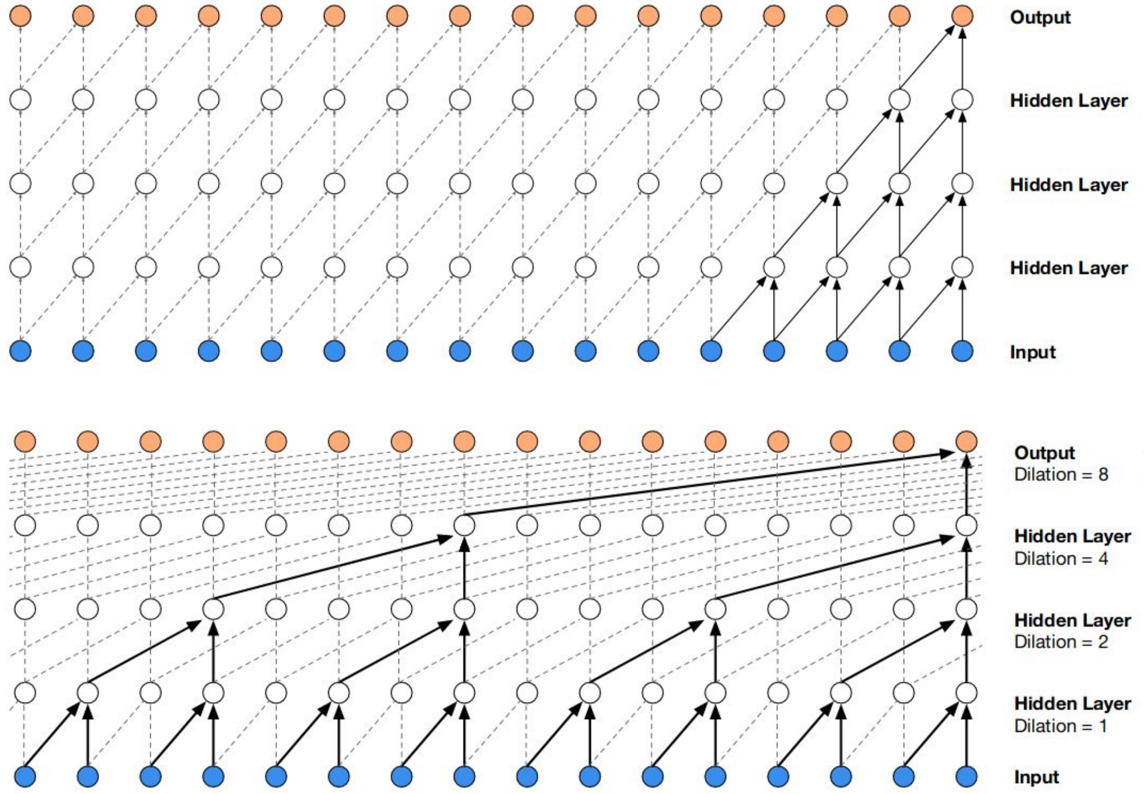
Kim, Sung, and Visvesh Sathe. "Bandwidth Extension on Raw Audio via Generative Adversarial Networks." arXiv, March 21, 2019. http://arxiv.org/abs/1903.09027.

My Method

- Try to apply the WaveNet architecture (dilated convolution) to BWE
- Mono input signal
- Using the time domain (phase is implicit)
- Upsample from 16kHz to 44.1kHz

Methods - Generator

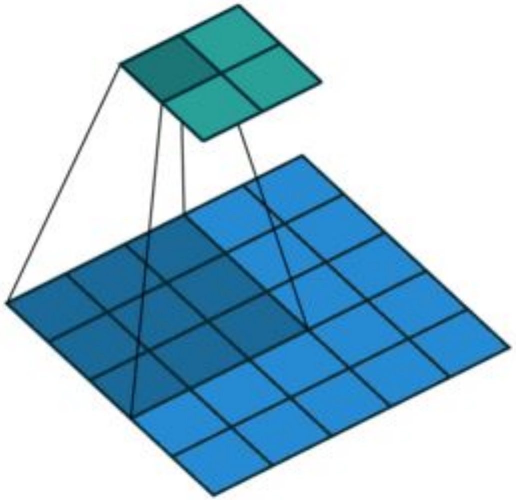
- Dilated convolution layers
  - similar to WaveNet architecture
  - Non-causal



https://www.deepmind.com/blog/wavenet-a-generative-model-for-raw-audio

Methods - Discriminators

- Strided convolutions
  - Reduce size of input



- Discriminator output is average value of last layer
  - Large number if real, small number if fake

Methods - Losses

- Training Discriminators
    - Loss ~= Discriminator(fake) + (1 - Discriminator(real))
- Training Generators
    - Adversarial loss
        - Loss ~= (1 - Discriminator(fake))
        - Spectrogram discriminators
        - 3 waveform discriminators (1x, 2x, and 4x downsampled)
    - L1 waveform loss
    - L1 spectrogram loss

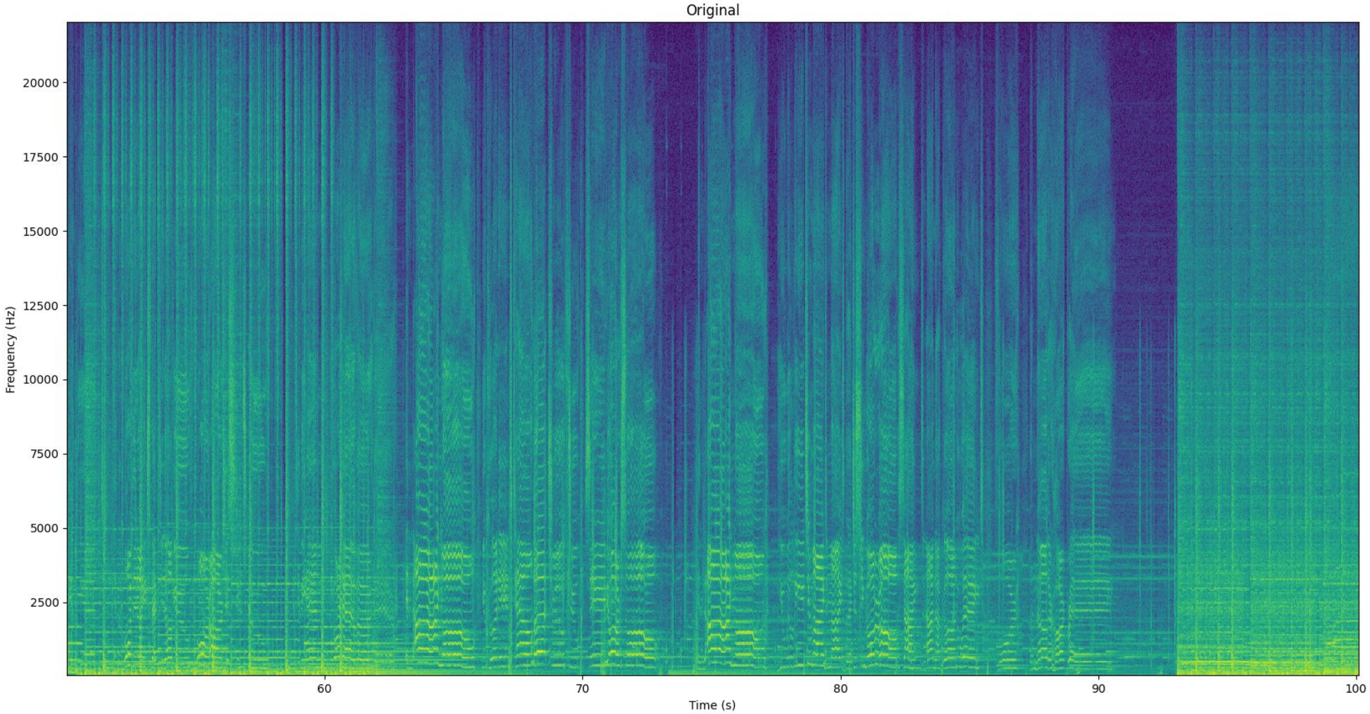
Objective Results

- Was looking at Frechet Audio Distance (FAD), but it's only trained with 16kHz data.
- Not accurate for perceptual quality

	Generator Only	Waveform Discriminators	Full Model
Signal-to-Distortion Ratio (SDR)	13.91 dB	15.35 dB	14.87 dB
Signal-to-Noise Ratio (SNR)	8.44 dB	7.45 dB	8.45 dB
Log-Spectral Distortion (LSD)	3.65 dB	2.64 dB	2.59 dB

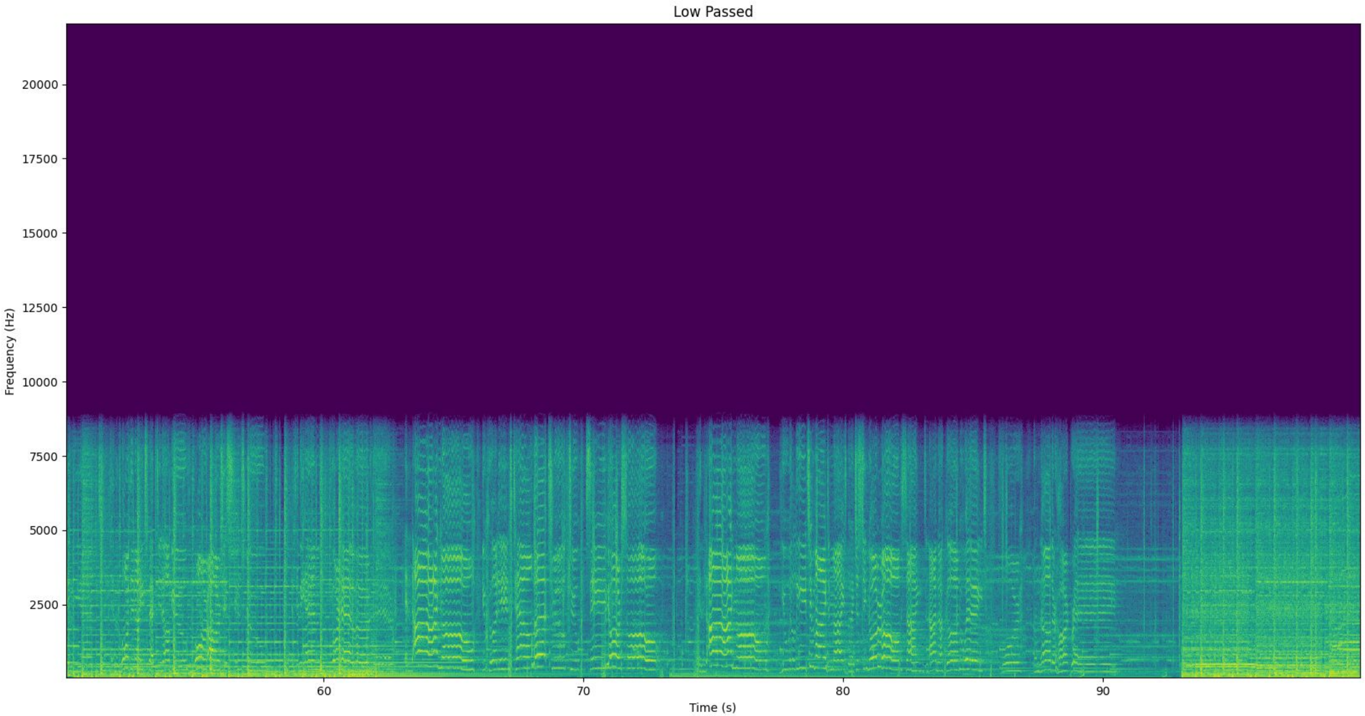
Visual Results: Original Spectrogram

-



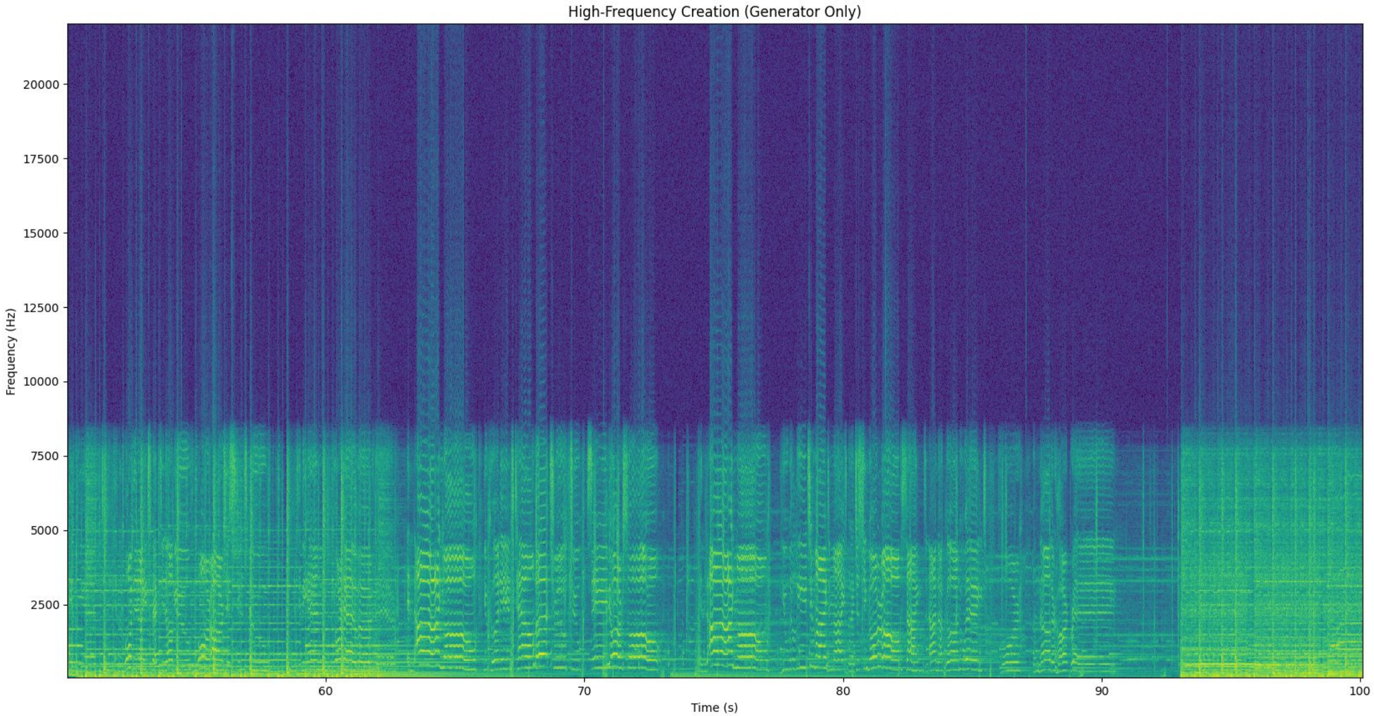
Visual Results: Input

-



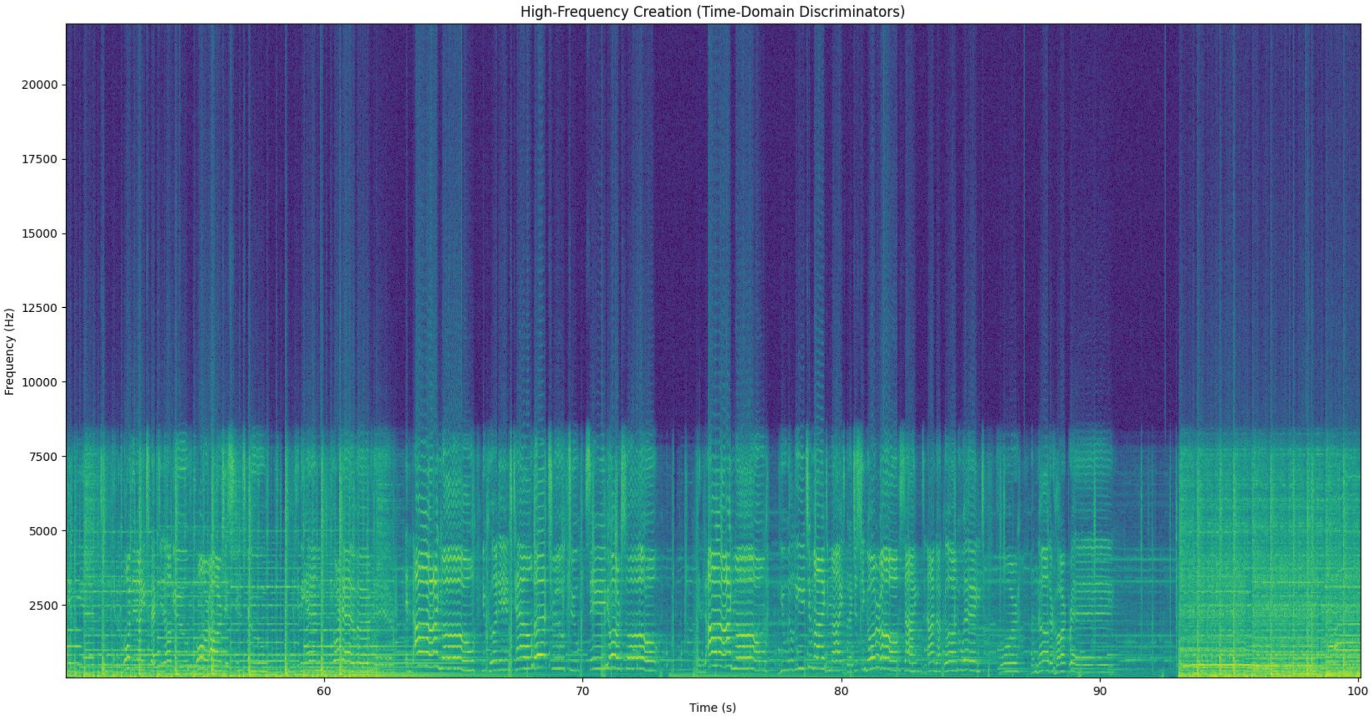
Visual Results: Generator Only

-

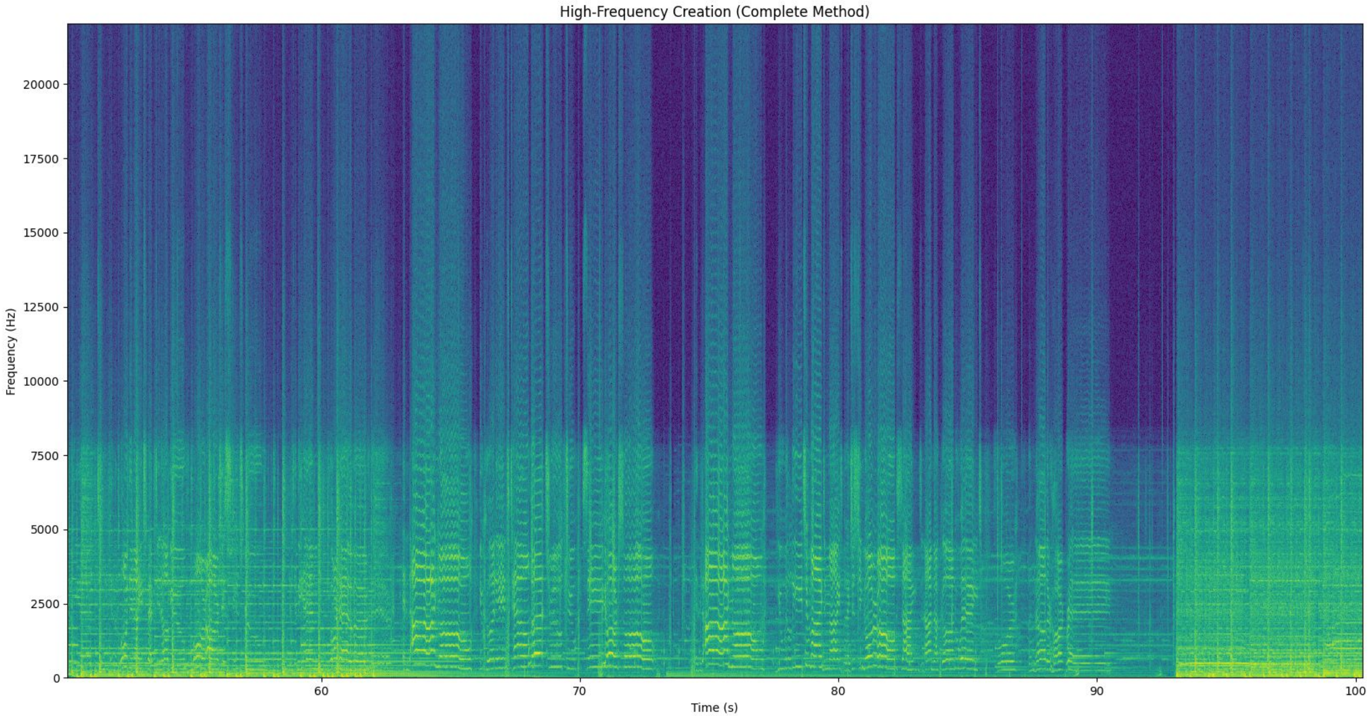


Visual Results: Waveform Discriminators

-

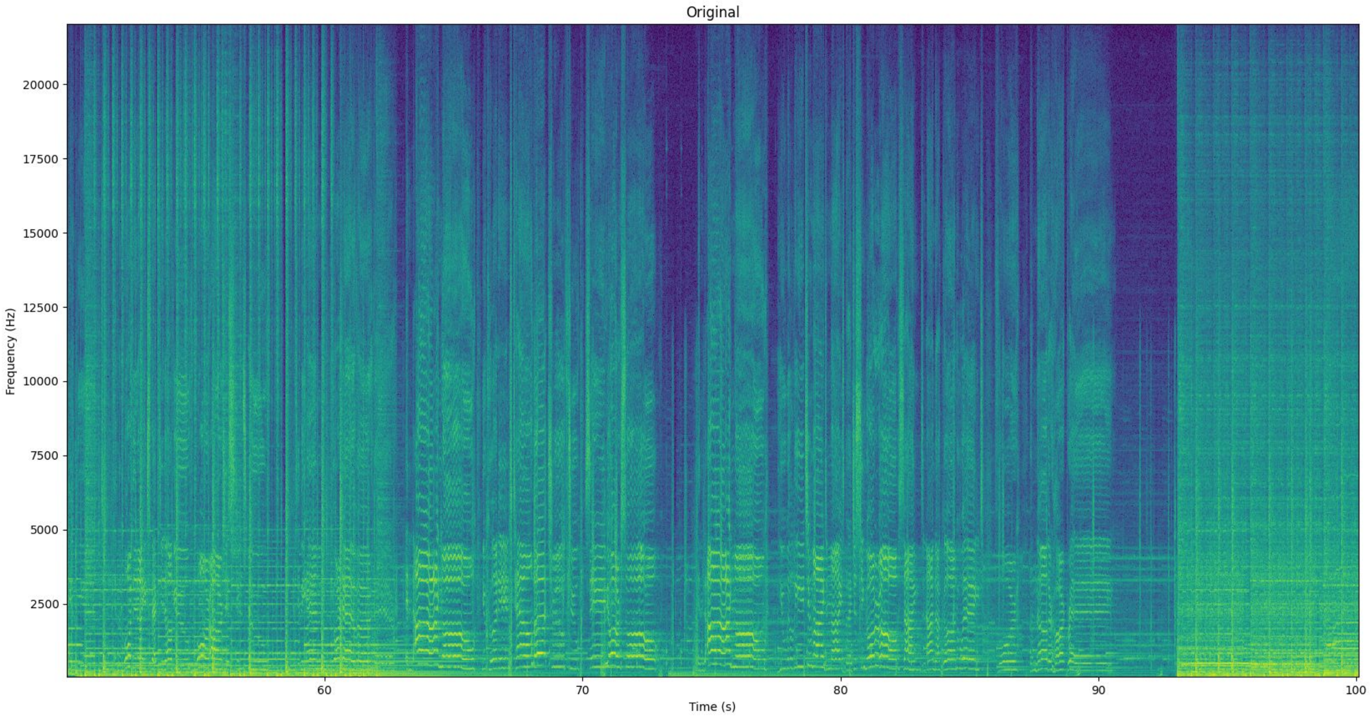


Visual Results: Waveform + Spect. Discriminators

- 

Visual Results: Original Spectrogram

-



Aural Results

Original:

Input:

Waveform + Spectrogram Discriminators:

Waveform Discriminators:

Conclusions/Limitations

- Results are passable, but certainly not wonderful

- Difficult domain, could have used more data
- Memory limitations (greater batch sizes, more layers)

Questions?