

DOUGLAS REEVES

Drum Source Separation via Generative Adversarial Network

DSSGAN

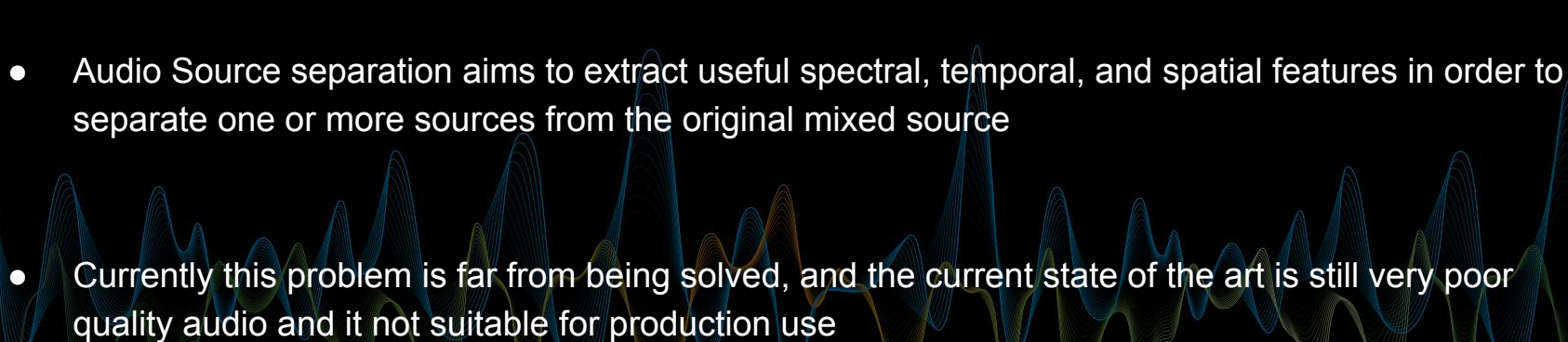


Design by Art Gravity

galvanize

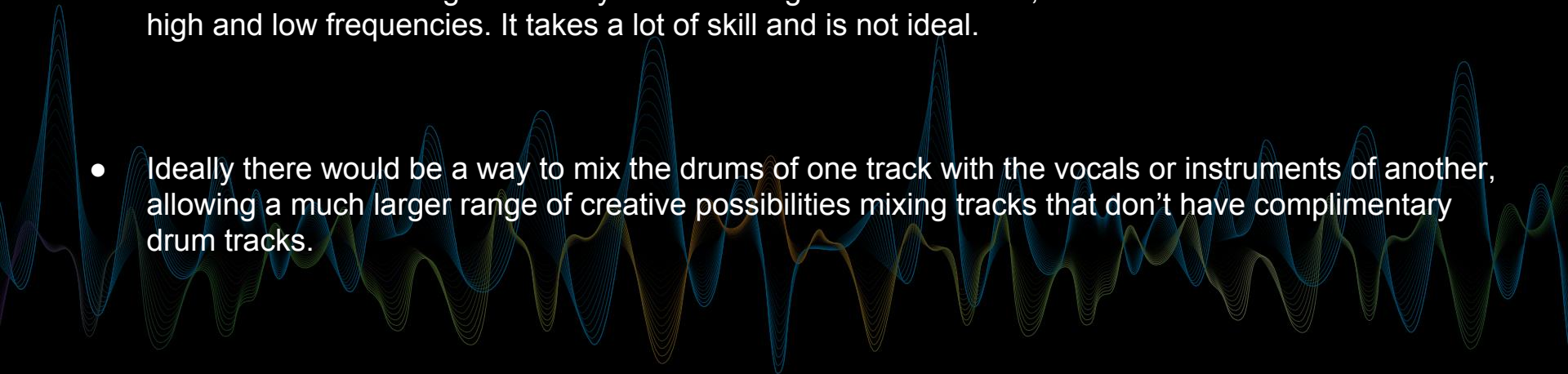
The Problem : Separating Professionally Produced Music

Most audio signals are mixtures of several audio sources:

- Vocals
 - Drums
 - Bass
 - Instruments
-
- Audio Source separation aims to extract useful spectral, temporal, and spatial features in order to separate one or more sources from the original mixed source
 - Currently this problem is far from being solved, and the current state of the art is still very poor quality audio and it not suitable for production use
- 
- A decorative graphic at the bottom of the slide consisting of multiple overlapping, semi-transparent waveforms in shades of blue and green, resembling an audio spectrogram or a complex waveform.

The Motivation

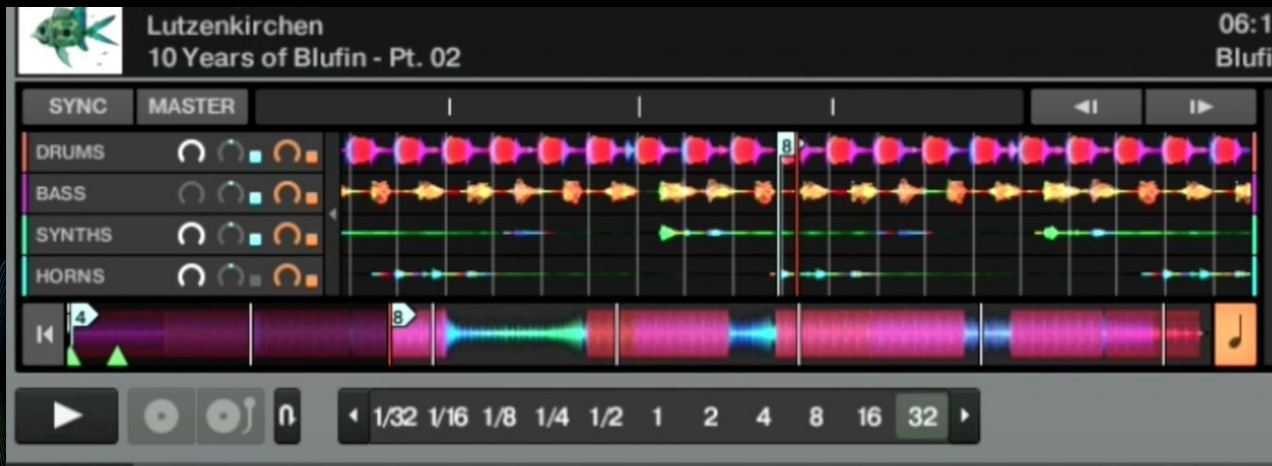
- DJ's and Producers traditionally have had to mix audio tracks together using only filters such as an equalizer. This means you can either cut the bass or the treble on each track. This limits which tracks can be mixed together smoothly, as the beats must be aligned in a complimentary way.
- Drums and percussion lie across the whole frequency spectrum so you end up having to blend the kick drums of one song with the cymbals and high hats of another, and often snares will be in both high and low frequencies. It takes a lot of skill and is not ideal.
- Ideally there would be a way to mix the drums of one track with the vocals or instruments of another, allowing a much larger range of creative possibilities mixing tracks that don't have complimentary drum tracks.



The Data

Native Instruments stems format

- This new format aims to address this problem for new music, and new artists are slowly starting to produce and sell stems, but it doesn't solve the problem for existing collections
- Contains 4 separated parts plus the mixture built on the mp4 video format
- Open source and python libraries available for manipulating the data easily

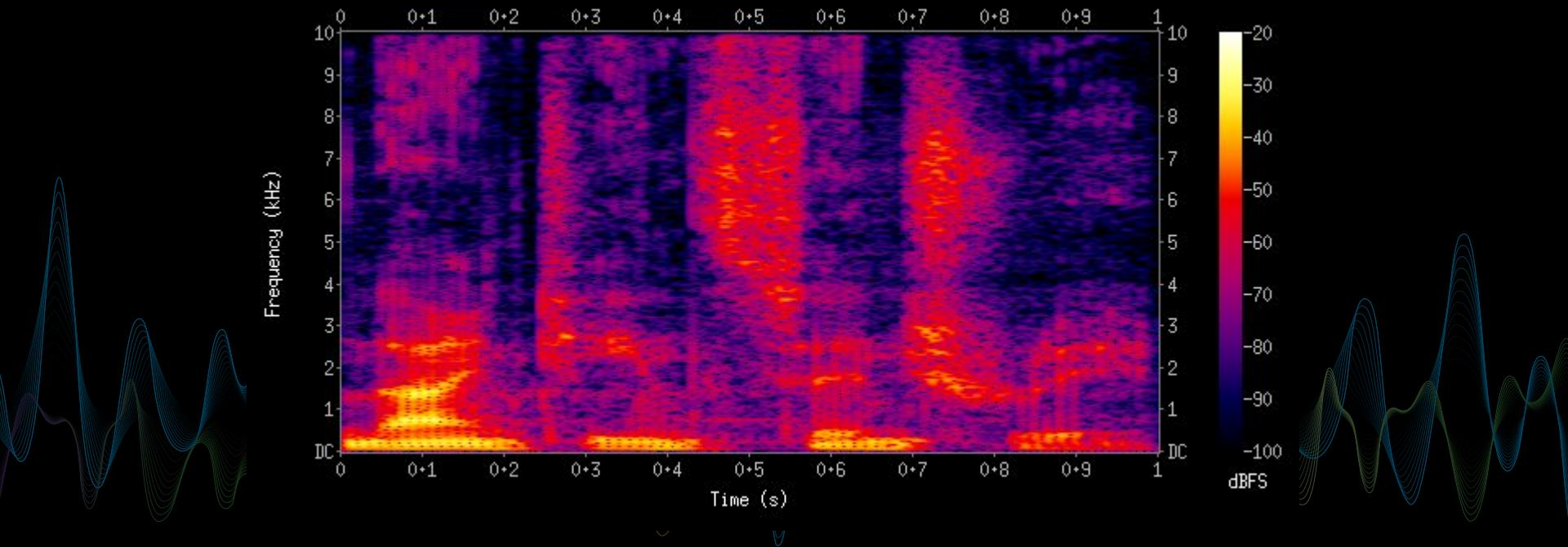


The Data Shortage

- MUSDB18 is a public dataset with only 150 multitracks, that means only a few songs of each genre and style, overfitting is a serious problem!
- I compiled about 100 hours of stem formatted music from my own collection of stems, professional sample libraries, and multitracks of remix projects to add to the supervised dataset
- Unsupervised data is much easier to find, as it does not have to be paired with a mixture or other source, and I was able to collect plenty of drums only tracks and music without drums for the unsupervised datasets
- I crowdsourced additional data from the music production and DJ community

Representing Audio Data

- Neural Networks work best with matrices as input, but raw audio data is normally a 1 dimensional array of amplitude values, so data is transformed using a STFT, or Short-Time Fourier Transform into a 2 dimensional matrix of Frequency vs Time, where magnitude is represented by color.



The Generative Adversarial Network (GAN)

- GAN's provide the best of both worlds in machine learning, combining supervised and unsupervised learning
- GAN's have been called the greatest advance in machine learning to date, but they are so difficult to fully understand that even the top researchers in the field don't always agree on how they work!
- Despite all this confusion, I will still try to explain it..

Generative Vs Discriminative Algorithms

- Discriminative algorithms try to classify data by learning the boundary between classes. So given the features of a sample of data, they predict probabilities that the data belongs could belong to a class or label. This is unsupervised learning trained on unlabeled data. Is this this drums or not drums?
- Generative algorithms do the opposite, so they predict the probability of features given a label. In other words they attempt to model the distribution of each class. This is supervised learning that requires labeled multitrack data. So given a source of mixed data, it outputs the parts of the data most likely to be drums.

Training the Model

- First the Generative Network is trained on the supervised dataset. The true drums source from the multitracks is compared to the predicted generated output, and the error is calculated using an MSE loss function and used to optimize the weights of the network
- Once the network is optimized, training is stopped and the model retains its weights for the next stage.
- The model can now produce separation predictions on it's own, but performance is not very good and for drum separation my results were of poor quality for data outside the training set. The next stage will attempt to improve performance by incorporating prior knowledge of the source distribution.

Adversarial Training

- The Generator is given a mixed audio source as input and generates a new spectrogram of audio with highest probability of being only real drums.
- The Discriminator takes this data and predicts whether it is true drum source data or if it is fake generated data.
- The result is back-propagated to both networks, providing a measure of error, and the process repeats.

The goal of each network is this:

- The generator is learning to create fake samples of separated drum audio that will not be detected as fake by the discriminator
- The discriminator is learning to detect the fake drums

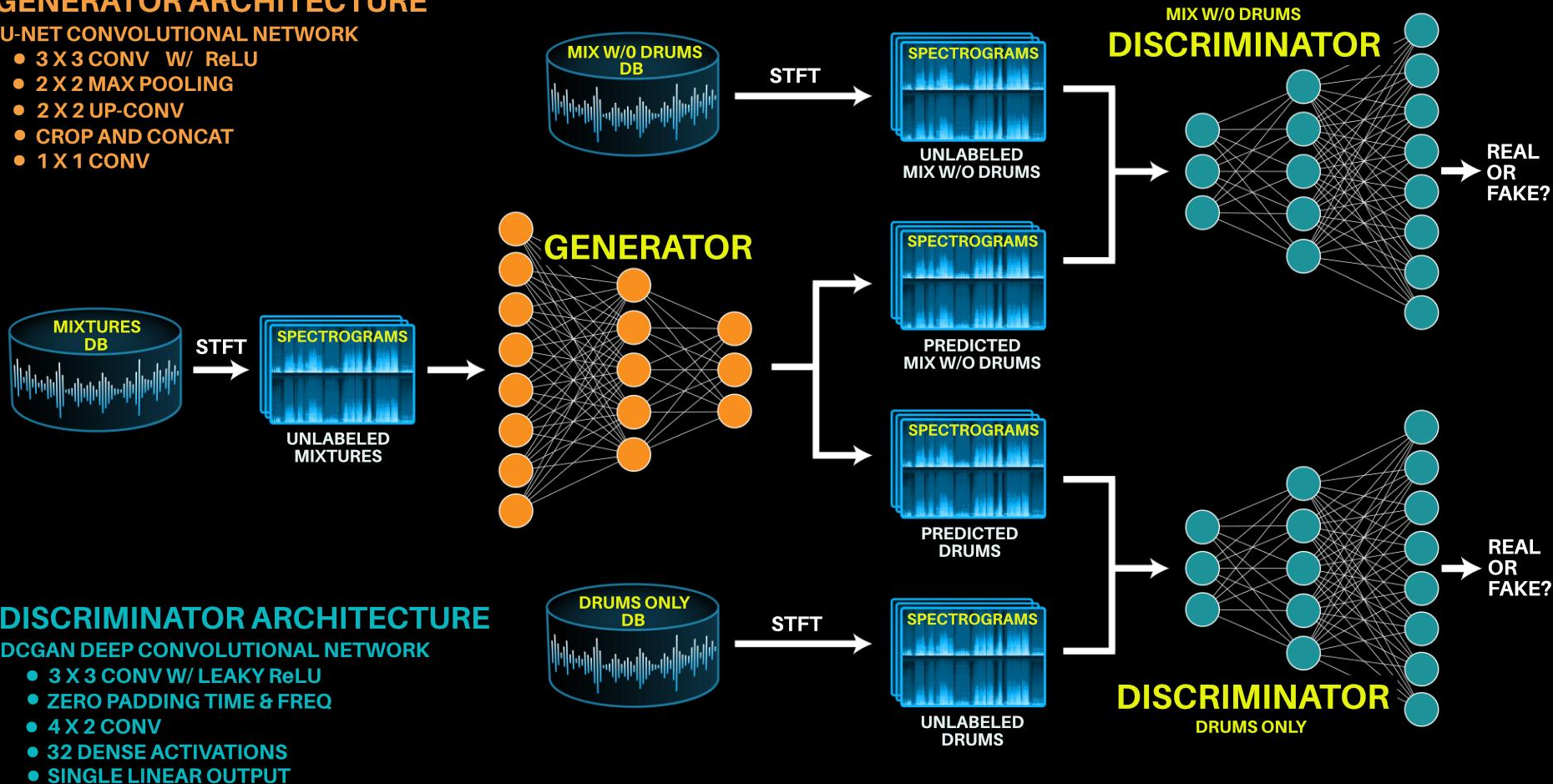
Both are in feedback loops with each other, and each is trying to optimize an opposing loss function, and as they each learn to adjust their behaviors to it becomes harder for either one to win.

ADVERSARIAL TRAINING

GENERATOR ARCHITECTURE

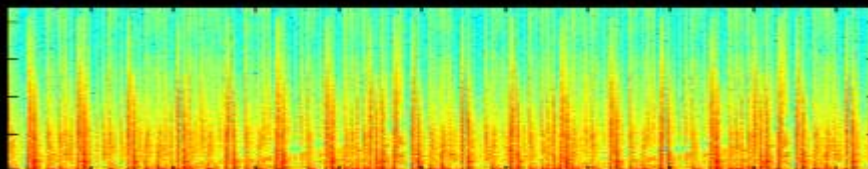
U-NET CONVOLUTIONAL NETWORK

- 3 X 3 CONV W/ ReLU
- 2 X 2 MAX POOLING
- 2 X 2 UP-CONV
- CROP AND CONCAT
- 1 X 1 CONV

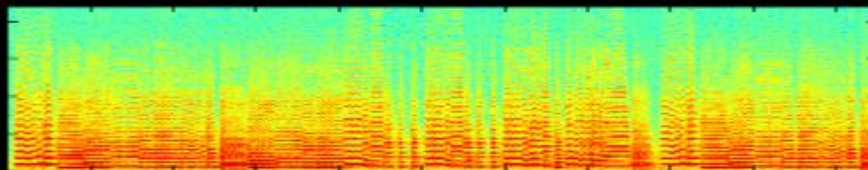


Results

- Results were on par with current methods for songs in the public dataset, but for music of different genres I got very poor results with the first trained model.
- I trained a second model specifically for electronic music which performed very well on similar music but very poor on rock, jazz, and other genres. This solidifies the idea that multiple models and classification of input signals is needed.
- This track had the best score!
- You can hear it on my iPad!



From_the_Heart_drums.wav



From_the_Heart_acc.wav

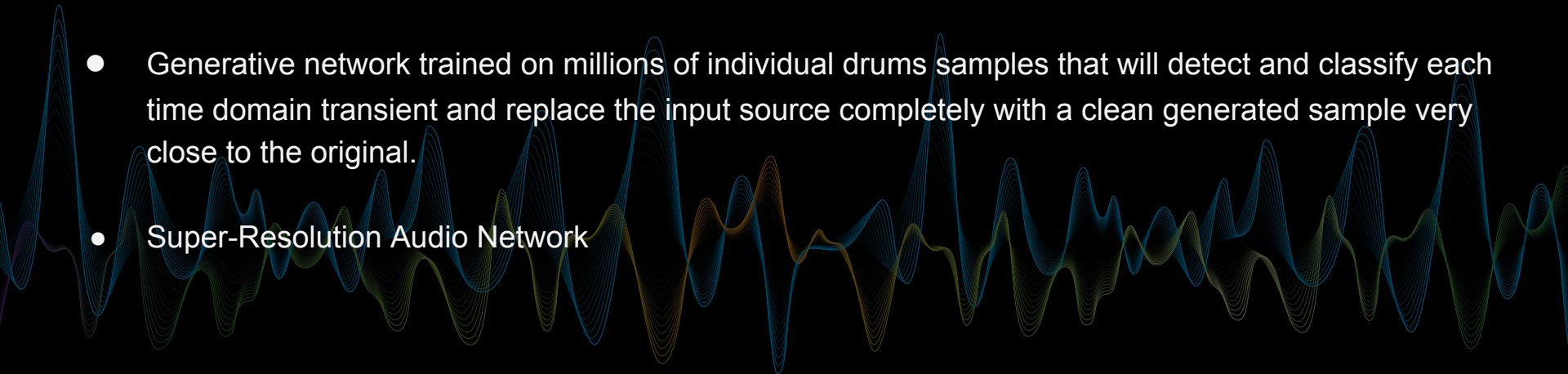
Next Steps

Add preprocessing steps:

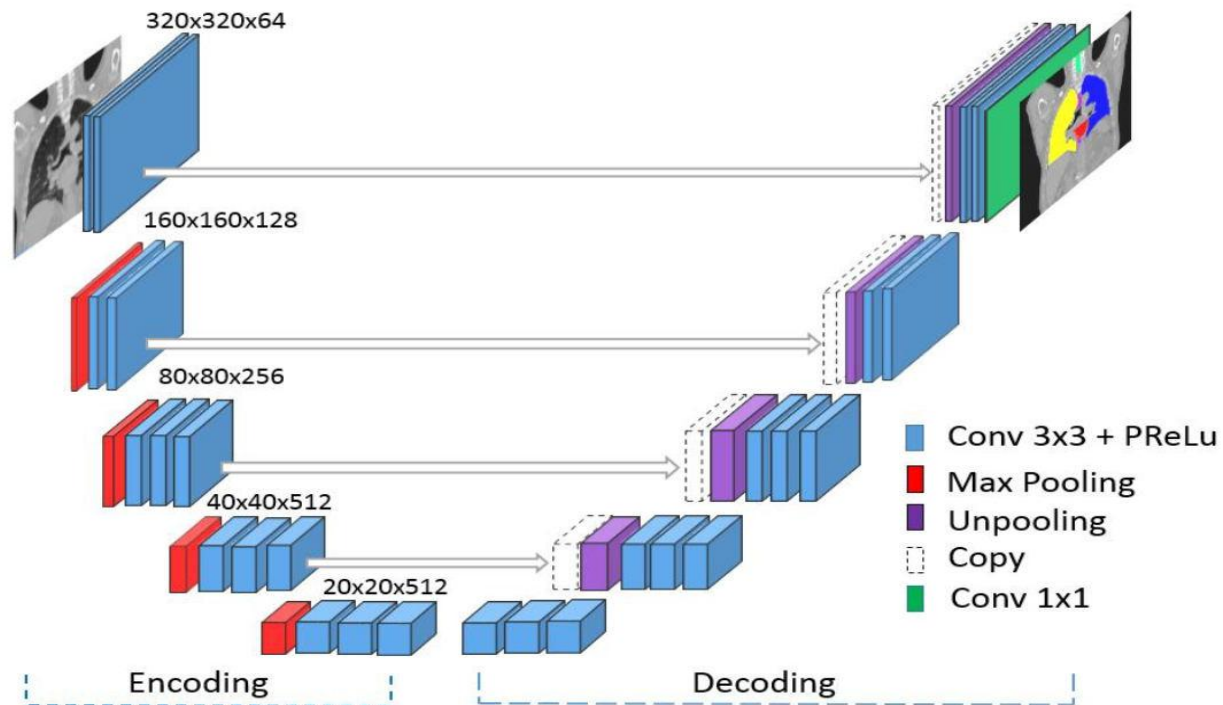
- harmonic-percussive source separation (HPSS)
- classification network that will utilize several GAN models trained on music of specific types and genres to account for the diverse characteristics of mixed audio sources.

Add post processing steps:

- Generative network trained on millions of individual drums samples that will detect and classify each time domain transient and replace the input source completely with a clean generated sample very close to the original.
- Super-Resolution Audio Network



Generator U-Net Architecture



How Adversarial Networks work

