ECE 188 Final:
Music Video Generation

By Mikhail Kardash and Sean Liu

Does music influence the music video?

Our goal

Cole Bennett is a sought after director for many rap music video productions. This makes him perfect for analysis since we can look for a relationship between rap music and music videos of one certain director.

Our goal is to train our network on a variety of music videos from this director to see if music influences the aesthetic of the music video.

Hypothesis

While there may be aesthetic variation in music and the music videos, we expect a highly biased network with variations in some shapes and color.

We do not expect the network to create a setting or even a person.

We based this on the results of our previous network.

Our Previous Work

Methodology

So much data processing...

- Download music videos
- Process the music videos
- Create spectrograms for the sound input
- Train Pix2Pix network on the spectrogram-image pairs
- Create a series of images based on another song

Pre-Processing the Data

1. Rip every 45th frame from the music video between 30-90 second marks. Resize to 256x256.
2. For each frame, take the previous 1.5 seconds of music and create a spectrogram image.
    a. Get 44100*1.5 values, then use scipy.spectrogram with 44100 sample frequency and with binnings such that output is 256x256.
3. Repeat this process for 30 music videos.
4. Do a separate video where we grab every frame for 10 seconds and the corresponding sound samples for input to the final model.

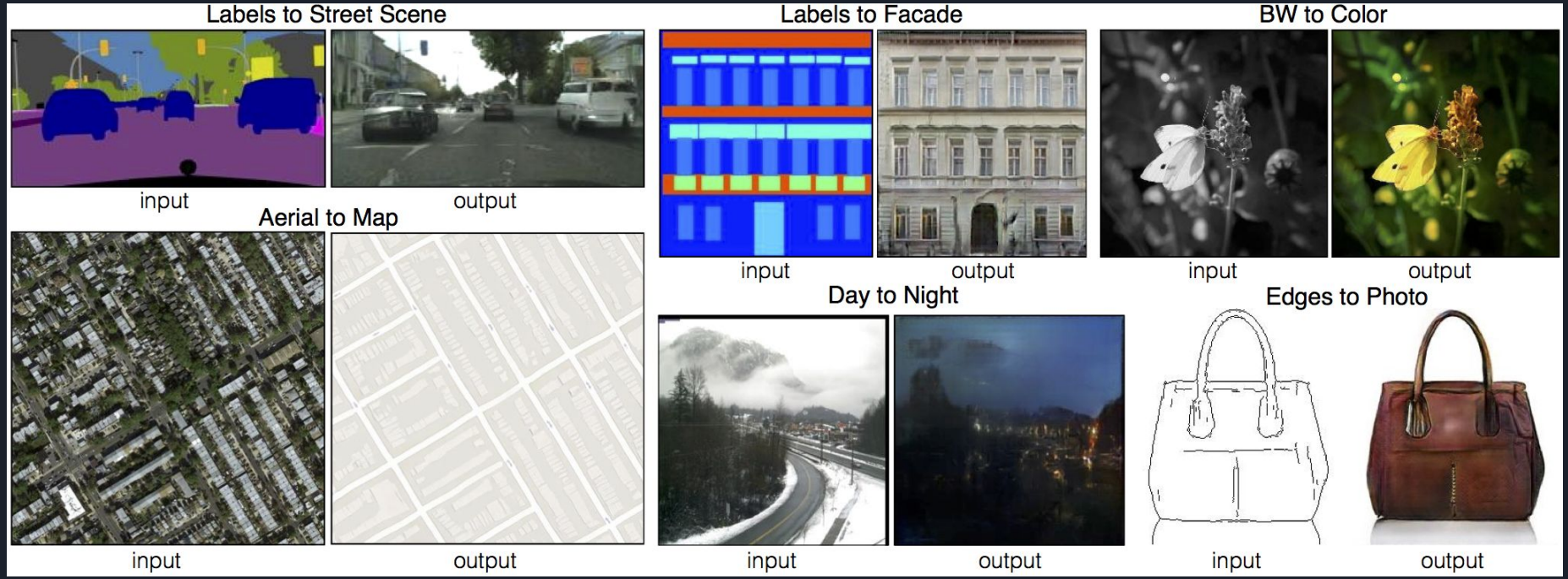
Pix2Pix

Pix2Pix is a conditional GAN that is designed to do any general image to image translation.

Only condition on training data is that inputs and outputs have same dimensionality.

Link: https://phillipi.github.io/pix2pix/

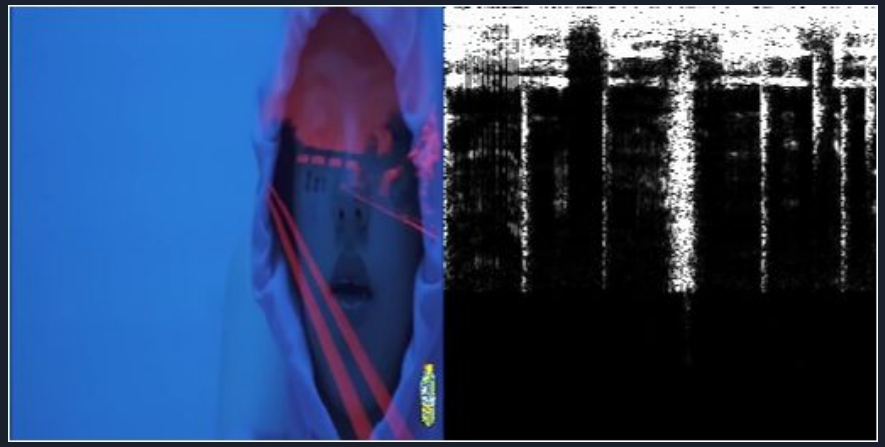
Pix2Pix Previous Works



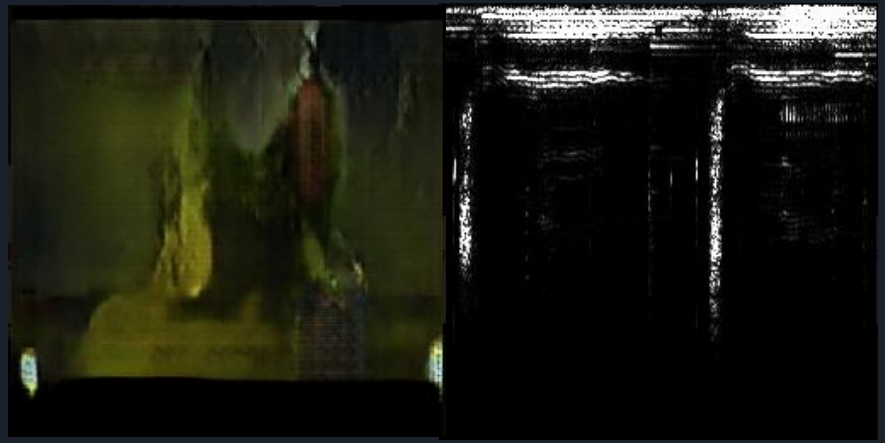
Implementation and Results

1. Here are examples of input and output image-data pairs.
2. We trained for 200 epochs. This took about 10 hours.
3. The individual image results are as we expected, but putting them into a video format tells a different story.

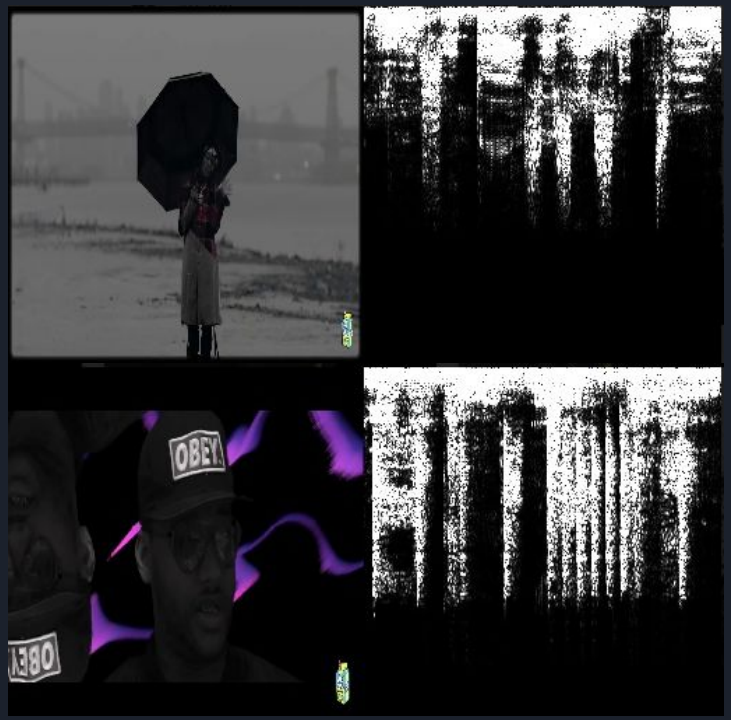
Input



Output



More Input Image Pairs



Video Generation

1. Overall shape and color are not only varying but transitioning smoothly in ways that make sense.
2. The bias actually provides context to the output image.
3. Transitioning shapes express ideas
4. Boundaries are ambiguous and undefined
5. Contrasting usage of light and shadow possibly reflective of its original musical medium
6. Texture is Decorative and Spontaneous

Input: https://www.youtube.com/watch?v=tc-v8MVw0S8&t=30s

Video Generation (cont.)

180 Epoch Result:

95 Epoch Result:

Video Generation (cont.)

Another 200 epoch result

180 epoch result

Conclusion

The network was actually able to produce color transitions that made sense and was clearly communicating aesthetic ideas through the way it morphed shapes.

The bias in the network actually helped with our overall video result even though it may not have seemed that way when each image was taken in isolation.

More heavily trained data ensures better overall output.

While we weren't able to prove that music influences the aesthetic of the music videos themselves, we can nonetheless conclude that music can have legitimate influence in creating abstract visual media.a