THE ANDREW & ERNA VITERBI FACULTY OF ELECTRICAL ENGINEERING

SIPL
Signal and Image Processing Lab

TECHNION
Israel Institute of Technology

Work in Progress

Seeing Sound: Estimating Image From Sound

Sagy Gersh and Yahav Vinokur,
Supervised by Tamar Rott-Shaham and Idan Kligvasser

Introduction

- Use of neural networks for different tasks is a growing field.
- Image and sound Classification are common areas of research.
- Recently, Improving in image generating technologies.
- Several attempts to create image generator from sound have been made with little to no success, using old technologies.
- Using the state of the art technologies might produce better results.

Understanding that using auto-encoders have been tried with little success, we decided to go with GANs, and created our system – SoundGAN.

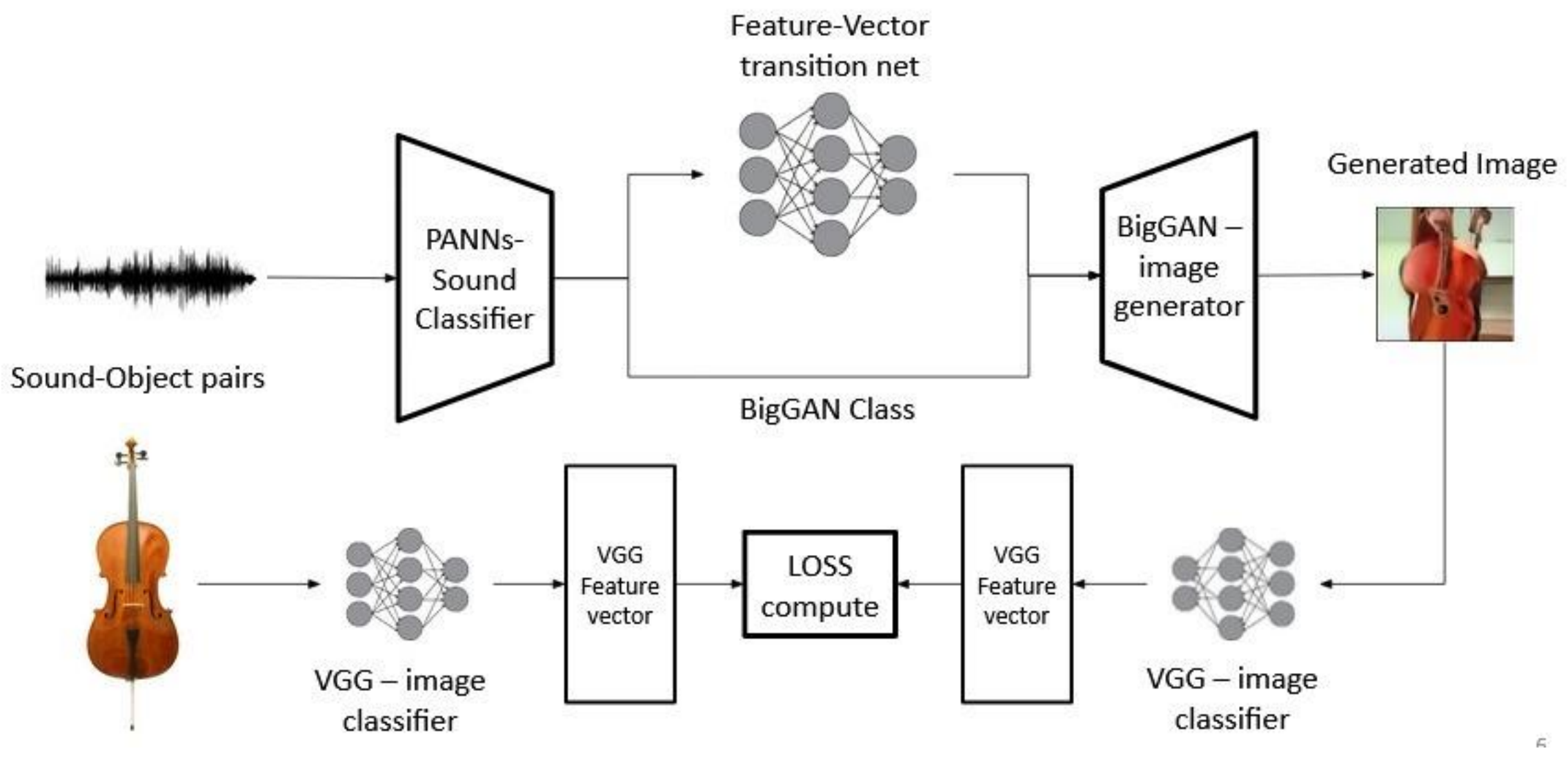
Goals

- Create an estimated image of a given sound source
- Show that manipulating the sound source affect the output image in a sensible way

Challenges

- No significant datasets for image-sound pairs
  - Existing databases are small and have very low variance
- Relatively bad result using previous methods in the literatures
  - Using auto-encoder over GANs enables better influence on the created image but has significantly lower visual performance than GANs.

Chosen Solution

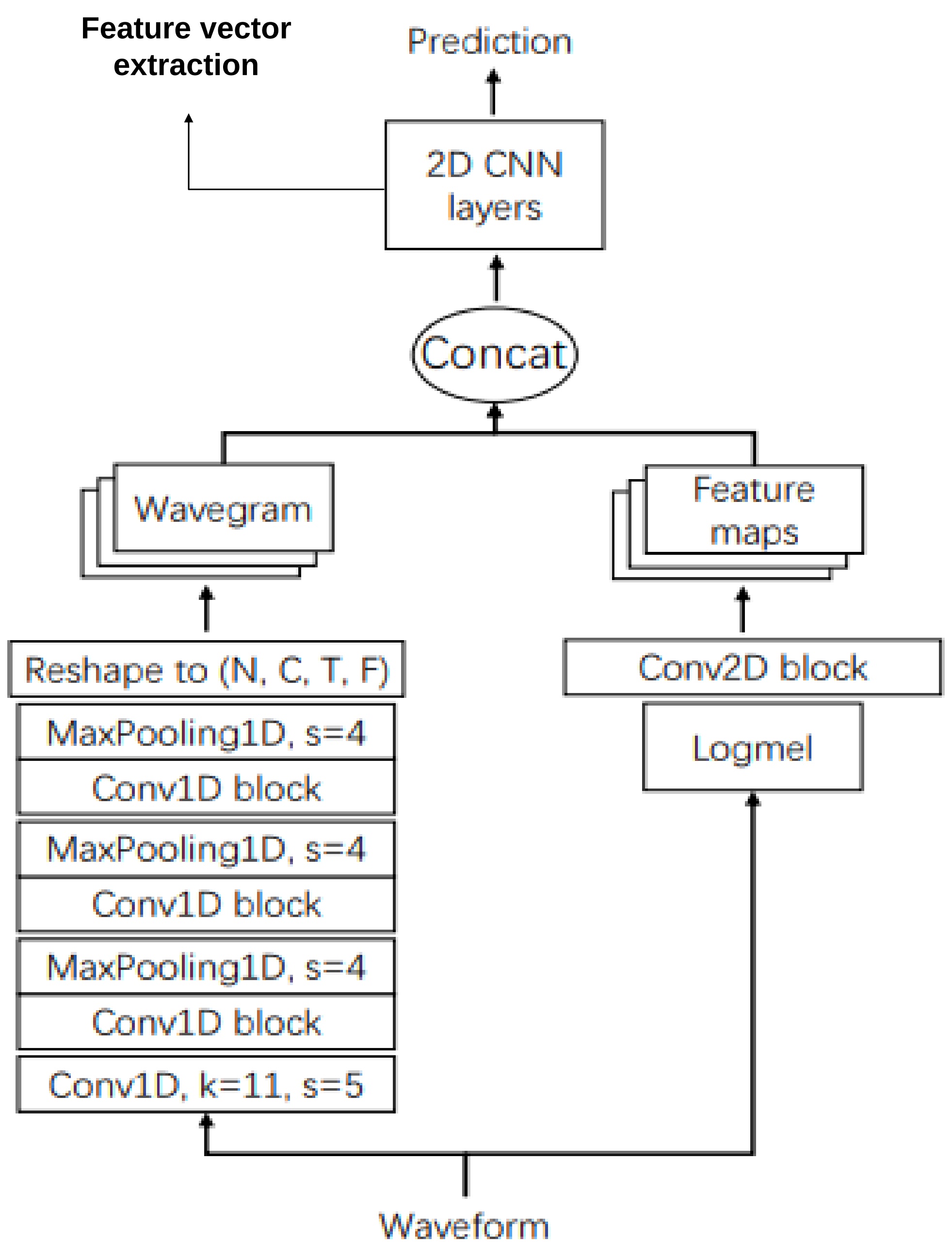


Our chosen solution connects sound classifier and image generator using a Transition-Net which we trained using sound-image pairs.

- Train transition-net between sound classifier feature vector and image generator input vector
- Using state of the art sound classifier named PANNs
- Using state of the art image generator named BigGAN

Sound Classifier

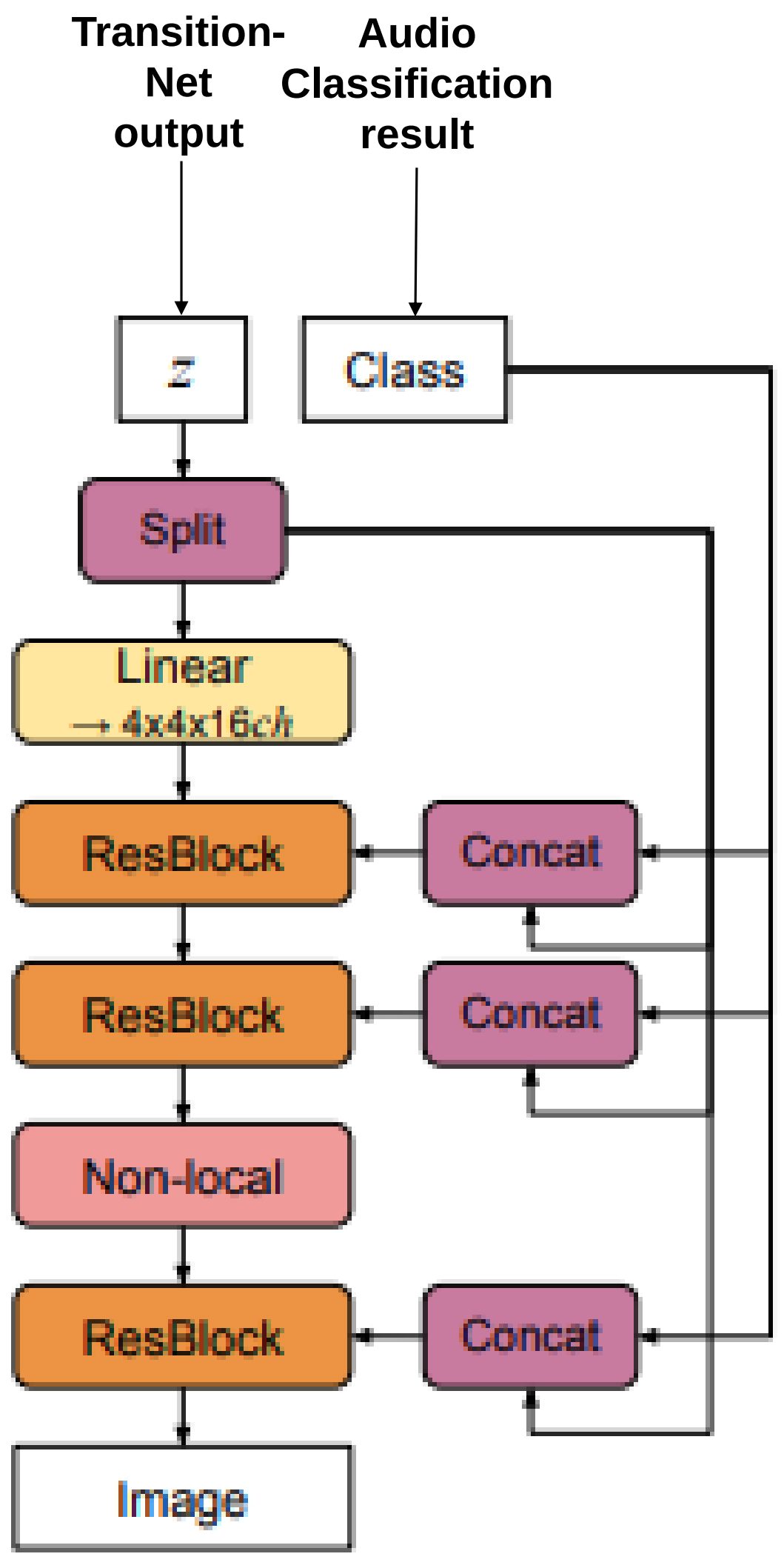
- PANNs - State of the art pre-trained audio neural-network
- Trained on large-scale ‘AudioSet’ dataset.
- By using log-mel spectrogram and waveform as input feature.
- Added a of feature vector branch-off before prediction



Architecture of Wavegram-Logmel-CNN

Image Generator

- BigGAN - State of the art pre-trained Generative Adversarial Network.
- Class-conditional image synthesis.
- Trained on a large scale – ImageNet with orthogonal regularization.



A typical architectural layout for BigGAN’s

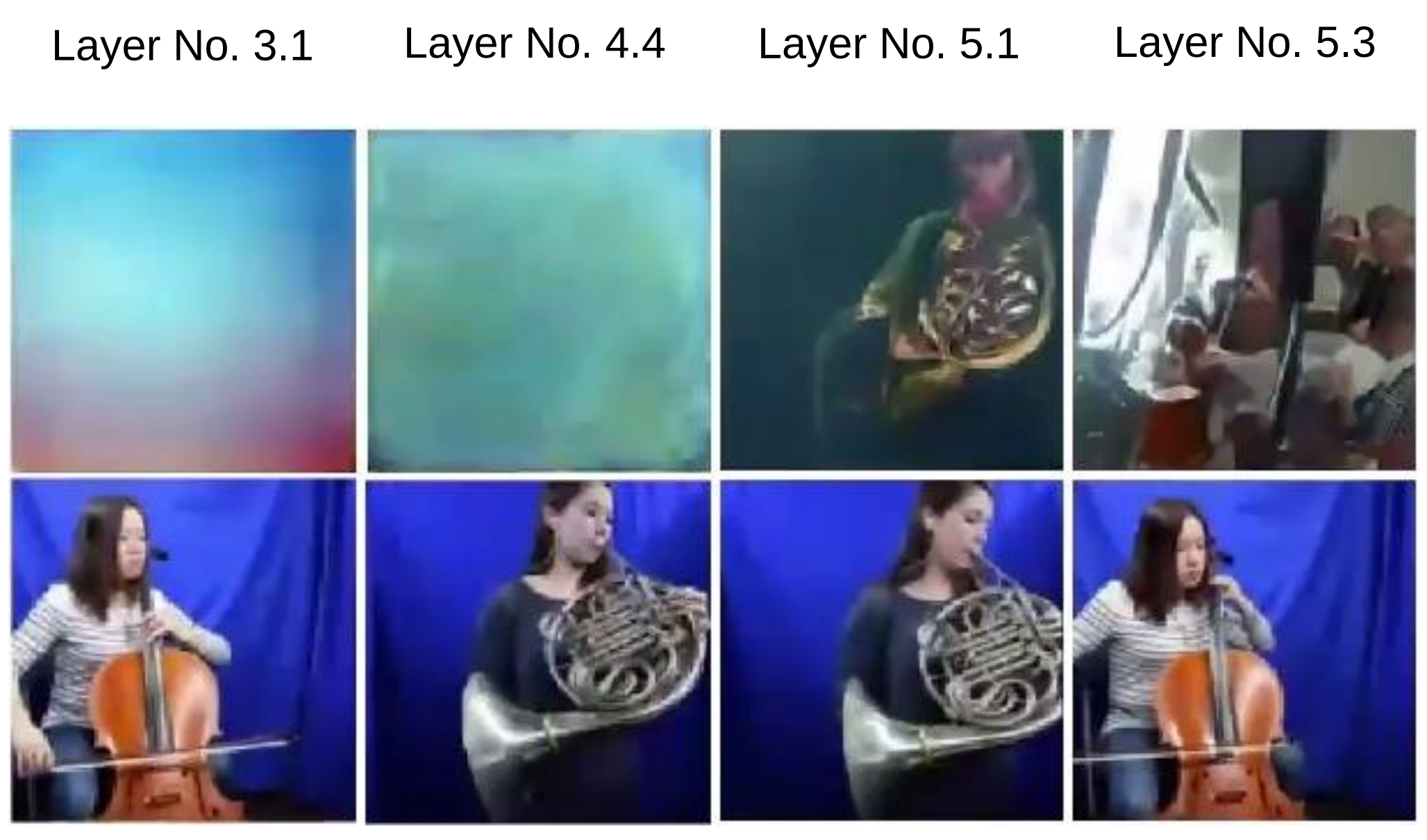
Class-conditional samples generated by BigGAN

Transition-Net

- Fully connected network
- Trained by us on AudioSet’s sub-set, using perceptual LOSS between the original image and BigGAN’s generated image.
- Subset created by running AudioSet with PANNs, and VGG19 to create image-audio pairs dataset.

Perceptual LOSS

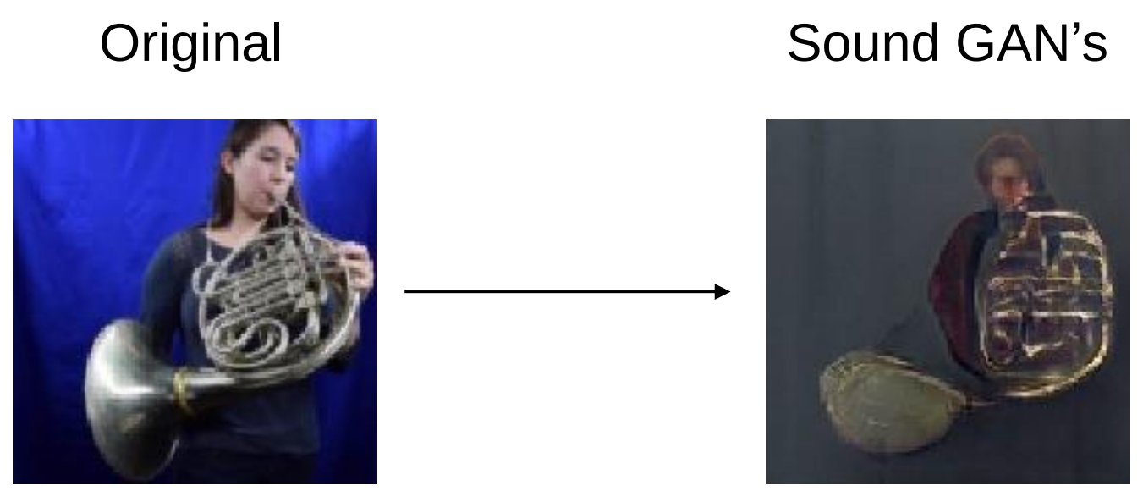
- Finding layers in VGG19 which contains object features.
- We ran diagnostic on VGG19’s layer-blocks 3,4,5 to determine the best layer for the task.
- Using L1Loss between feature-vector from that specific layer.



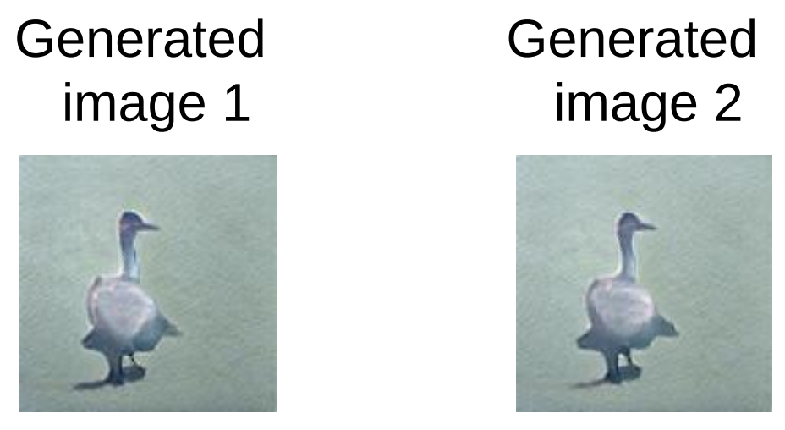
LOSS layer diagnostics

Current results

- We succeeded in creating a picture from audio that is similar in meaning to the original image



- We started by using pre-existing smaller dataset, resulting in very-low-variance output



Gooses created from different origins of sound

Conclusions

- We could see from our results that there is enough data in the sound in order to create an image.
- We couldn’t affect the images by changing the sound because of the low-variance dataset we trained on.
  - Next step – move to AudioSet’s sub-set

June 2021