

---

# Generating Images from Audio

---

**Chih Wen Lin** \*  
Electrical Engineering, Stanford University  
jennlinc@stanford.edu

**Ting-Wei Su** \*  
CCRMA, Stanford University  
twsu@ccrma.stanford.edu

## Abstract

We took the state-of-the-art audio signal encoder and deep generative model to achieve images from sounds. Our model, Audio2Img, is trained and evaluated on audio and images from a subset of AudioSet. This work shows preliminary results of generating images based on audio data and points out a need for higher quality audio-visual dataset to bridge the gap between audio and visual learning.

## 1 Introduction

Human perception of auditory and visual stimuli are shown to be strongly correlated. Given an audio clip, we can often imagine motions and contents associated with the sound. Despite recent advance in various generative models based on text descriptions, few audio-visual models have been successfully developed. In this paper, we explore an architecture that can map audio information into visual images.

Constructing an audio-visual generative model involves audio feature extraction and conditional image synthesis. Audio feature extraction is a commonly explored problem. Here, we simply take the log-mel-spectrogram of audio clips and convert to embedding vector via deep convolutional neural networks. To synthesize images from the given audio embeddings, we are inspired by text-to-image GANs [1, 2], where the generation of images is also conditioned on the given text embeddings.

## 2 Methods

### 2.1 Dataset

AudioSet [3] collects 10 seconds of footage from millions of videos on YouTube. We selected 10 classes of videos, 49,643 files in total, that have obvious correlations between audio and images. These 10 classes include: Cat, Helicopter, Train, Acoustic Guitar, Baby Crying, Firework, Dog, Race Car, Rooster Crowing, and Ocean. We randomly picked 3 images from each sample video, cropped them into squares and down-sampled them to  $64 \times 64$ . The sampling rate of the audio is 22,050Hz. Our evaluation set, totally 917 samples, is based on the evaluation sets of the same 10 classes in AudioSet.

### 2.2 Audio Feature Extraction

We use the whole 10 seconds of audio as our input data. The raw waveform is first averaged across two channels and then transformed to its Mel-spectrogram with 128 Mel-frequency bins. We used the default parameters in Librosa[4] to obtain this feature. To further extract latent embeddings from the Mel-spectrogram, we created an embedding encoder that contains several residual blocks of convolutional layers[5].

---

\*The two authors have equal contribution

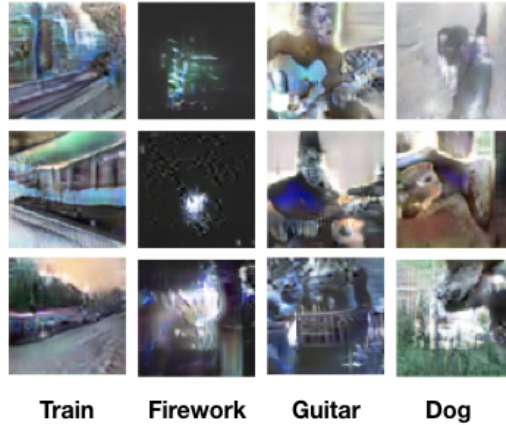


Figure 1: Images generated by Audio2Img. Each column represents a kind of label to the input audio.

### 2.3 Image Synthesis

Our generative model has a structure similar to the stage I of StackGAN[2], with text embeddings replaced by audio embeddings. The Conditioning Augmentation (CA) network re-samples the embedding by passing the audio latent vector  $\phi$  through a simple fully connected layer to generate  $\mu_0(\phi)$  and  $\sigma_0$  for the Gaussian distribution  $N(\mu_0(\phi))$ . Next, the CA net outputs the re-sampled audio feature by drawing a sample from the conditioned Gaussian. This feature is then concatenated with random noise and up-sampled to generate a  $64 \times 64$  image. The generator and the discriminator are both constructed with convolutional layers.

## 3 Discussion & Conclusion

Part of the results generated by our model are shown in Figure 1. The model was trained with 400 epochs. The quality of generated images is highly dependent on labels, and we concluded that there are two major difficulties in this task: the lack of temporal information in audio and the movement of target objects in images. Neither do we know where the most representative sound is within the 10-second audio file, nor do we have the position or the frame of the actual object. As a result, on the audio side, the most important features within the 10 seconds might not be captured by the embedding encoder, causing the generator to create an unrelated image. We observed that it is easier for the embedding encoder to discover sounds with longer and more obvious average appearance such as guitar and firework. Sounds of ocean and helicopter have long sound duration, but the training results for these samples are worse because the sound too similar to noise. Similarly, the sound of the animals are usually too short in duration (and sometimes even not obvious) that the embedding encoder can hardly capture the crucial features.

As for images, classes with higher similarity between sample images are usually guaranteed with clearer generated images of the objects, while images in other classes does not. These variations in correlations between images and audio features result in noticeable differences in generated images. For example, objects such as train, train track and firework are more observable since their representations are more static across most training data. We also found that guitar sounds usually result in images of human playing guitar, which may due to the fact that most videos with this label are stable recordings of someone playing a guitar. On the other hand, objects like dog, cat, and baby crying vary a lot between videos since they can be captured in many different directions. Furthermore, videos of these labels often contain highly dynamic background and don't necessarily contain the object itself. We acknowledged that AudioSet is not the perfect dataset for audio-visual model and hope this result can bring initiatives to construct a more suitable dataset for future audio-visual models.

Overall, despite the noisy training data, Audio2Img is able to generate recognizable images solely based on short clips of audio. We proved the possibility of conditioning GAN on audio features, marking a successful first step in bridging the audio and visual learning gap.

## References

- [1] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.
- [2] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. Stacked generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 4, 2017.
- [3] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 776–780. IEEE, 2017.
- [4] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015.
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.