# Self-Supervised Audio-Visual Representation Learning for in-the-wild Videos

Zishun Feng*†, Ming Tu‡, Rui xia‡, Yuxuan Wang‡ and Ashok Krishnamurthy†§
†Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
§Renaissance Computing Institute, Chapel Hill, NC, USA
‡Bytedance Inc., Mountain View, CA, USA

***Abstract*—Humans understand videos from both the visual and audio aspects of the data. In this work, we present a self-supervised cross-modal representation approach for learning audio-visual correspondence (AVC) for videos in the wild. After the learning stage, we explore retrieval in both cross-modal and intra-modal manner with the learned representations. We verify our experimental results on the VGGSound dataset [1], and our approach achieves promising results.**

***Index Terms*—self-supervised learning, multimodal representation learning, large scale video understanding**

## I. Introduction

The Audio-visual correspondence task, first introduced by Arandjelovic et al. [2], aims to predict whether the input image corresponds to the input audio and simultaneously learn good visual and audio representations.

Previous works on audio-visual learning mainly focus on musical instruments data, which consists of more temporally continuous and spatially structured sounds compared to general audio. In this work, we explore the AVC performance not only on musical instrument data but also on videos in-the-wild. Moreover, these videos make the task more challenging due to complex audio spectral information, background noise, invisible audio sources, and video frames unrelated to sound, etc.

## II. Proposed Approach

From an intuitive concept, the visual information and audio information that stand for the same event should show up simultaneously in a video. We sample video frames and 1-second audio segments from a given dataset of videos. We assume the video frame is corresponding to the audio segments only if the video frame is sampled at the middle of the audio segments in the same video.

The proposed model is shown in Figure 1. The input to our network is a pair of a single video frame and an audio segment, which may or may not correspond to each other. We use two ResNet18 [3] networks to extract visual features and audio features, respectively. Then we activate the distance between visual and audio features with a sigmoid function to predict the correspondence of the input video frame and audio segment. Note that our method trains model in a self-supervised manner and does not require any category information for the training data, but only the correspondence between the input audio segments and video frame, which can be obtained at the time of sampling the input data from videos.

In this work, the embedding dimensions are 128 for both image and audio branches, and the distance metric is Euclidean distance.

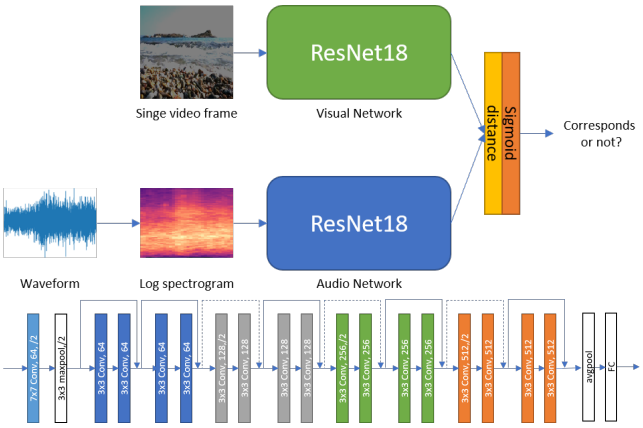



Fig. 1. Architecture of proposed network for audio-visual correspondence.Top: architecture of the entire network. Bottom: architecture of ResNet18. Note that the dimension of ResNet18's output is 128 in this work. Dotted arrows denote skip connections. "/2" denotes that the stride of the layer is 2.

## III. Experiments and Results

We trained and tested proposed approach on the VGGSound dataset [1]. We dynamically sampled audio-image pairs from over 200,000 videos with a positive rate of 0.5 for training the model. An audio segments and an image correspond only if they are sampled at the same time point within the same video. The model was trained for 400k steps with a batch size of 256, so over 100 million sampled audio-image pairs were used in total for training. The parameters of the model were optimized by Adam optimizer with a learning rate of 0.001.

### A. AVC Results

We compared our AVC performance with previous works [2], [4]. Since these methods were evaluated on different datasets, we have listed the methods and their corresponding datasets and results in the table below. As shown in Table I ,our approach achieves competitive results compared to other methods, despite the small size and large variety of training data.

*Work done partially during internship at Bytedance Inc.

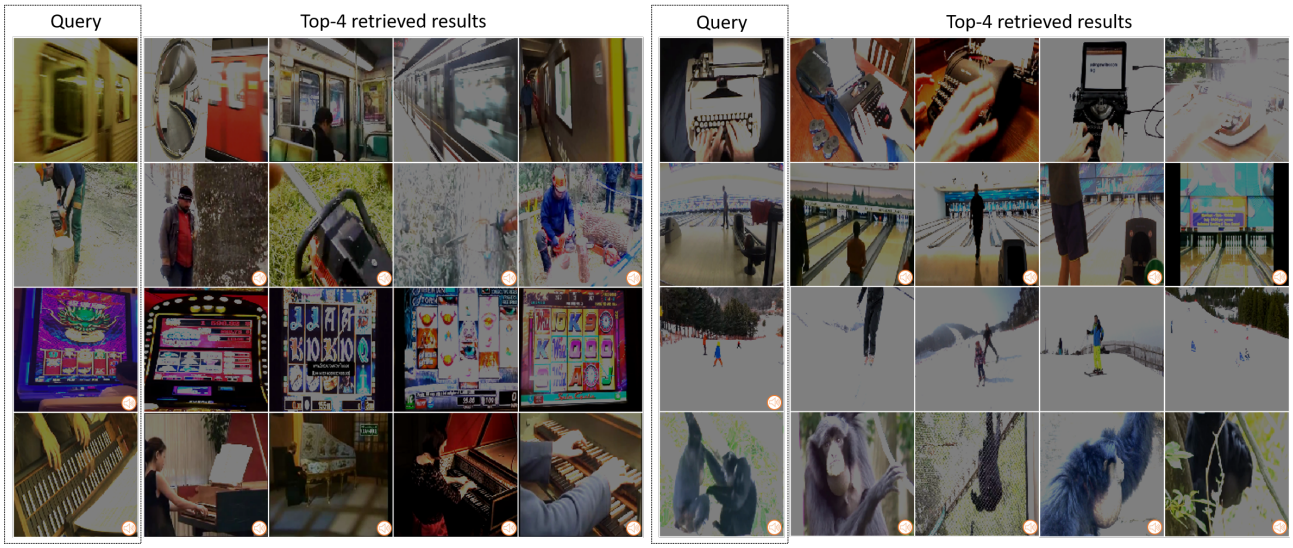



Fig. 2. Retrieval results. Each row contains query images (or audio), listed in the the dotted boxes, and retrieved results. From top to bottom: image to image retrieval, image to audio retrieval, audio to image retrieval and audio to audio retrieval. For the visualization purpose, each image with an icon indicates the sound corresponding to that image.

TABLE I
COMPARISION OF AVC PERFORMANCE.

| Method | Dataset | #Video | #Class | AVC |
|---|---|---|---|---|
| [2] | Flickr-SoundNet | 500k | unlabeled | 78% |
| [4] | AudioSet-Instruments | 300k | 110 | 81.9% |
| Ours | VGGSound | 200k | 309 | 79.8% |

### B. Intra-modal and Cross-modal Retrieval

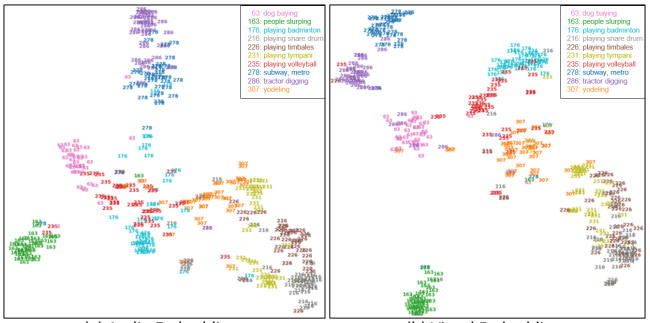
We show our qualitative retrieval results in both intra-modal and cross-modal manners in Figure 2. As shown in Figure 2, our method can be applied to obtain good retrieval results for video in various scenarios, both indoors and outdoors, at high or low volume, and with high or low noise.

### C. Visualization of Embeddings

Besides the retrieval results, we also show the visualization of the learned audio and visual representations using t-SNE method. As shown in Figure 3 (a) and (b), videos with the same or similar label tend to cluster together in both audio and visual space. For example, because snare drum, timbales, tympani are all drums and played in a similar way, they mix together in visual space, but are farther apart in audio space due to their tonal differences; the sound of both subway and tractor are loud and similar, so their clusters are very close in audio space; most of the videos of people playing volleyball and playing badminton are collected in the gymnasium, so they are clustered nearby in the video space.

## IV. CONCLUSION

In this work, we present a self-supervised learning approach to learn both audio and visual representations at the same time via audio-visual correspondence tasks. The experimental results show that our approach can learn good representations on a wide variety of videos and provide competitive AVC results in terms of videos in-the-wild.




(a) Audio Embeddings (b) Visual Embeddings

Fig. 3. Examples showing t-SNE visualization of learned audio (a) and visual (b) representations. To show the embeddings clearly, only 10 classes are sampled randomly and plotted in this figure Note that the labels shown in the figures are purely for visualization purposes and were not used during the training stage.

## REFERENCES

[1] Chen, Honglie, et al. "Vggsound: A Large-Scale Audio-Visual Dataset." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
[2] Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." Proceedings of the IEEE International Conference on Computer Vision. 2017.
[3] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[4] Arandjelovic, Relja, and Andrew Zisserman. "Objects that sound." Proceedings of the European Conference on Computer Vision (ECCV). 2018.