# Learning to Separate Object Sounds by Watching Unlabeled Video (Supplementary Materials)

Ruohan Gao[1], Rogerio Feris[2], Kristen Grauman[3]

[1]The University of Texas at Austin, [2]IBM Research, [3]Facebook AI Research
rhgao@cs.utexas.edu, rsferis@us.ibm.com, grauman@fb.com**

The supplementary materials consist of:

A. Supplementary video.
B. Details on AudioSet-Unlabeled and merging of ImageNet synsets.
C. "Non-MIL-Pooling" variant
D. Details on the bases collection process.
F. Spectrogram visualization of audio source separation.

## A Supplementary video

In our supplementary video, we show (a) example audio source separation results on novel "in the wild" videos; (b) example unlabeled videos and their discovered audio basis-object associations; (c) visually-assisted audio denoising results on three benchmark videos (see Table 2 and Sec. 4.4 in the main paper for the quantitative results).

## B Details on AudioSet-Unlabeled and merging of ImageNet synsets

We use the publicly available AudioSet dataset [1] as our source of unlabeled videos. AudioSet offers noisy video-level audio class annotations. However, our method does not use any of its label information. We filter the dataset to those likely to display audio-visual events, including the following categories: Accordion; Acoustic Guitar; Banjo; Cello; Drum; Electric guitar; Flute; French horn; Harmonica; Harp; Marimba; Piano; Saxophone; Trombone; Violin; Dog; Cat; Frog; Chicken,rooster; Car; Motorcycle; Rail transport; Aircraft.

We merge similar ImageNet categories to roughly align with the AudioSet classes. We merge the categories by averaging the prediction scores of the corresponding ImageNet classes. Specifically, the merged categories are—cat: 'tabby, tabby cat', 'tiger cat', 'Persian cat', 'Siamese cat, Siamese', 'Egyptian cat'; dog: 'Chihuahua', ... , 'African hunting dog, hyena dog, Cape hunting dog, Lycaon pictus' (ImageNet label index 151-275); chicken: 'cock', 'hen'; frog: 'tree frog,

** *On leave from The University of Texas at Austin (*grauman@cs.utexas.edu*).*

tree-frog', 'tailed frog, bell toad, ribbed toad, tailed toad, Ascaphus trui', 'loggerhead, loggerhead turtle, Caretta caretta'; racing car: 'convertible', 'racer, race car, racing car', 'sports car, sport car'; train: 'bullet train, bullet', 'electric locomotive', 'passenger car, coach, carriage', 'steam locomotive'; plane: 'airliner', 'plane, carpenter's plane, woodworking plane', 'warplane, military plane'; motorbike: 'moped', 'motor scooter, scooter'.

## C "Non-MIL-Pooling" variant

The proposed MIML network obtains mAP of 0.418 on the validation set. To calibrate, we ran a "non-MIL-Pooling" variant of our network that removes the last two pooling operations and adds a FC layer ($K \times L \times M \Rightarrow L$), in order to see if traditional classification (using the same architecture otherwise) would be competitive. The new classifier obtains 0.375 mAP. Our MIML design not only offers a classification accuracy gain, but it also has the benefit of enabling basis-object relation discovery (Sec. 3.4).

## D Details on the bases collection process

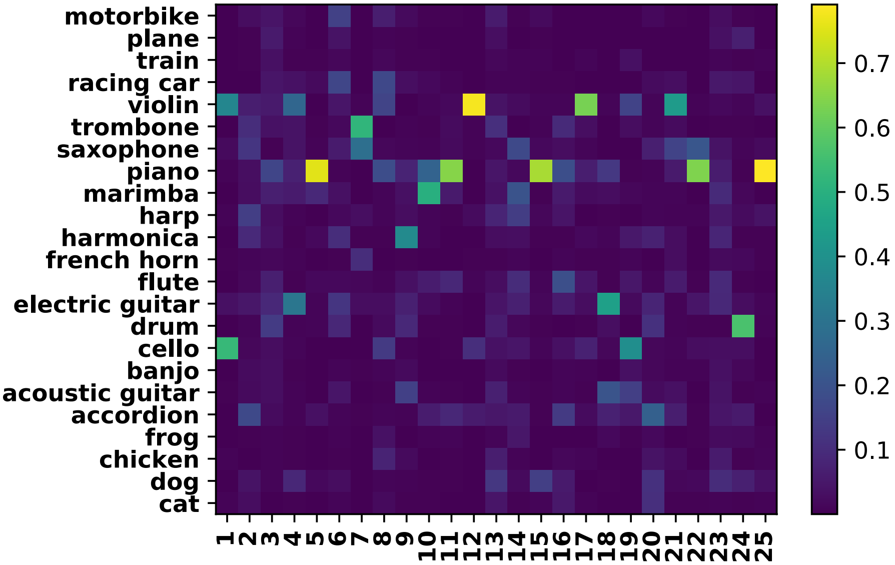

**Fig. 1.** A typical audio basis-object relation map, where the horizontal axis shows the indices of the basis vectors and the vertical axis shows the object classes. The color represents the probability of a basis vector belonging to a class as inferred by our deep MIML network.

The audio basis-object relation map after the first pooling layer of the MIML network produces matching scores across all basis vectors for all object labels. We perform a dimension-wise softmax over the basis dimension ($M$) to normalize object matching scores to probabilities along each basis dimension. Fig. 1 shows a typical audio basis-object relation map, with lightness indicating the probability of a basis belonging to a certain class.

By examining the probability map, we can discover links from audio bases to objects. We only collect the key bases that trigger the prediction of the correct objects (namely, the visually detected objects). Further, we only collect bases from an unlabeled video if multiple basis vectors strongly activate the correct object(s). For example, the visual predictions of the map shown in Fig. 1 identify piano and violin. In the basis-object relation map, we can see that five basis vectors strongly activate the label piano, indicating that the sound of piano is very likely to be contained in the corresponding audio track. Due to class imbalance of the training data, we collect a maximum of 3,000 basis vectors for each object category.

## E Spectrogram visualization of audio source separation

Fig. 2 shows the spectrograms of the mixed and separated audios as well as the ground-truth for the test data used in Sec.4.4 in the main paper. We can see that audios of different classes can have different spectral patterns. Our system leverages the learned prototypical basis vectors for each class to supervise source separation. Consistent with the SDR results quantified in Table 1 in the main paper, the separated spectrograms (last two columns) look close to the ground-truth (first two columns). Moreover, our system can match the separated signals with the correct meaningful acoustic objects present in the video. In the last row, we show a typical failure case of our system. The spectrogram of acoustic guitar and electric guitar are similar and heavily overlap, and the sources are not well separated.

## References

1. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: ICASSP. (2017) 1

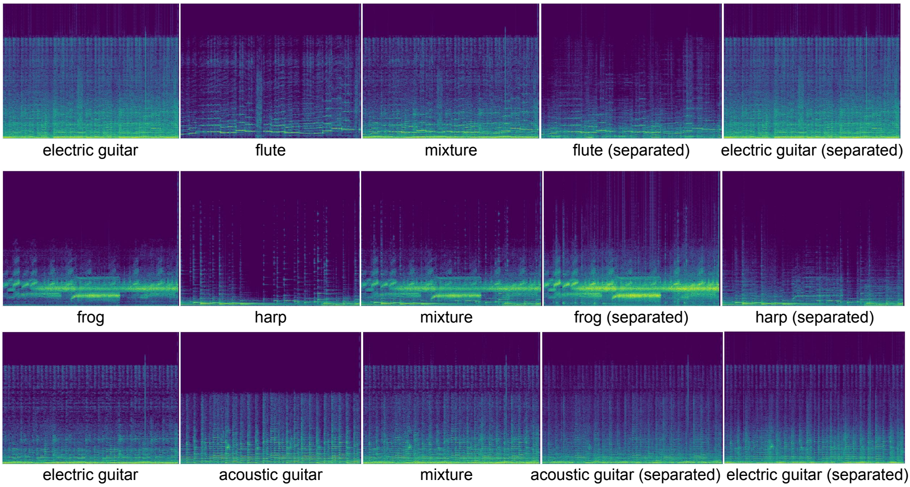

**Fig. 2.** Visualization of the spectrograms of the ground-truth single source audios (first two columns), the mixed audio (third columns), and the separated audio tracks (last two columns) using our system. The last row shows a typical failure case where signals of similar acoustic characteristics cannot be well separated.