Co-Separating Sounds of Visual Objects (Supplementary Materials)

Ruohan Gao
UT Austin
rhgao@cs.utexas.edu

Kristen Grauman
UT Austin and Facebook AI Research
grauman@cs.utexas.edu

The supplementary materials for [1] consist of:

A. Supplementary video.

B. Details of using pre-trained object detector to find potential sound-making objects.

C. Details of Audio-Visual Separator network.

D. Additional experiments and ablation study.

A. Supplementary Video

In our supplementary video, we show example separation results. We use our system to discover and separate object sounds for realistic multi-source videos from AudioSet dataset and duets in MUSIC dataset. We compare to our best audio-visual baseline (Sound-of-Pixels, Zhao *et al*. ECCV18) and the audio-only baseline (Spiertz & Gnann *et al*. DAFx09). The AV-MIML baseline is trained on a different set of object categories and is therefore not available for comparison. The Sound-of-Pixels baseline originally performs video-level mix-and-separate source separation. To perform source separation at object-level for realistic videos during testing, we use the localized object region as the input to the visual stream and the multi-source audio as the input to the audio stream to separate the sound responsible for the input visual object. Therefore, we can then obtain the sounds grounded to each detected object as our method.

From the examples, we can see that our co-separation approach can discover and separate object sounds for realistic multi-source videos. Our method generates cleaner separation compared to the baseline methods, and it can also ground the separated sounds to the meaningful visual objects in the video. In the last separation example of piano and trumpet, the piano is silent in that video. Our model properly captures this in the separation, creating a "silent" separation track for the object even though it is visible and detected visually in the frame. In the last two failure cases, we show that our model can be constrained by the breadth of the pre-trained object detector. Furthermore, it finds difficult to perform separation in diverse scenes with unmodeled sounds such as human voice.

B. Details of Object Detection

We train an object detector on $\sim$30k images of 15 object categories from the Open Images dataset [3]. The 15 object categories include: Banjo, Cello, Drum, Guitar, Harp, Harmonica, Oboe, Piano, Saxophone, Trombone, Trumpet, Violin, Flute, Accordion, and Horn. We use the public PyTorch implementation[1] [4] of Faster R-CNN to train an object detector with a ResNet-101 [2] backbone.

Then we use our pre-trained object detector to find objects in video frames for the AudioSet-Unlabeled dataset. We extract 80 frames from each unlabeled 10s video clip, and perform object detection on each frame. We use the following filtering procedures to reduce the noise of the obtained detections: 1) We only keep object detections of confidence larger than 90%; 2) If two object detections of different class overlap more than 70%, we only keep the one with the larger confidence; 3) We only keep the top two detected categories of the largest confidence, because this agrees with the number of objects detected by our pre-trained detector in most training videos.

As mentioned in step 3) above, we limit the maximum number of object categories to be two in our co-separation training. The reason is that this matches with the number of objects detected in most training videos, but nothing fundamental in our model restricts it to two. To demonstrate our framework is flexible to the number of detected objects per video, we train our model on AudioSet without the limit of two, i.e., drop step 3). An AudioSet video typically has 1-5 detected object classes. Eliminating the hard limit, our results only slightly degrade: 3.46/6.55/12.7 (Table 2 in the main paper), which is still much stronger separation than the baselines.

C. Details of Audio-Visual Separator Network

Our audio-visual separator network consists of a visual branch and an audio branch. The visual branch takes images of dimension 224x224x3 as input, and extracts a fea-

[1] https://github.com/jwyang/faster-rcnn.pytorch

ture map of dimension 7x7x512 through ResNet-18 ImageNet pre-trained network. The visual feature map is then passed though a 1x1 convolution layer to reduce the channel dimension, and produces a feature map of dimension 7x7x128. The feature map is then flattened and passed through a fully-connected layer to produce an aggregated visual feature vector of dimension 512.

The audio branch is of a U-NET style architecture, namely an encoder-decoder network with skip connections. It consists of 7 convolution layers and 7 up-convolution layers. All convolutions and up-convolutions use 4 x 4 spatial filters applied with stride 2, and followed by a BatchNorm layer and a ReLU. After the last layer in the decoder, an up-convolution is followed by a Sigmoid layer to bound the values of the spectrogram mask. The encoder uses leaky ReLUs with a slope of 0.2, while ReLUs in the decoder are not leaky. Skip connections are added between each layer i in the encoder and layer $n-i$ in the decoder, where n is the total number of layers. The skip connections concatenate activations from layer i to layer $n-i$.

The U-NET produces an audio feature map of dimension $512 \times 2 \times 2$, with 512 the channel dimension, after the last convolution layer. The visual feature vector of dimension 512 is replicated 2×2 times to produce a $512 \times 2 \times 2$ visual feature map. Then we concatenate the audio and visual features along the channel dimension to produce an audio-visual feature map of dimension $(512+512) \times 2 \times 2$. The series of up-convolutions in U-NET is finally performed on the concatenated audio-visual feature map to generate a multiplicative spectrogram mask $\mathcal{M}$.

D. Additional experiments and ablation study

The Table 3 experiment in the main paper is designed to demonstrate our method can learn from duets only: Guitar only appears in duets, but our model successfully separates the sound of guitar during testing. We perform an additional experiment to train on *only* duets. We train with all duet types with at least 25 videos in MUSIC[2] and test on mixes of the solo videos of those instruments. The results for ours and SoP are 2.75/4.32/11.9 and 0.15/0.39/15.4, respectively. Our approach learns much better from duet videos only.

We also perform an ablation study to examine the impact of the key components of our CO-SEPARATION framework. Table 1 compares the source separation performance of several variants of our model on AudioSet dataset. We compare our model with one variant that only uses the co-separation loss; one variant that only uses the object-consistency loss; one variant that removes the "adaptable" class. We can see that object-consistency loss alone does not suffice to learn source separation, but together with the separation loss we obtain the best performance. The "adaptable" class is not essential to our system, but arms the network with the flexibility to assign noise or unrelated sounds to it, leading to better separation performance as shown in the table.

[2]Since this experiment requires training with much less data, the absolute performance is lower and the setup is not comparable with Table 1 in the main paper.

	SDR	SIR	SAR
co-separation loss only	3.65	6.13	13.2
object-consistency loss only	0.14	0.14	**45.0**
without "adaptable" class	3.70	5.30	14.4
CO-SEPARATION (Ours)	**4.26**	**7.07**	13.0

Table 1: Ablation study. Note that SDR and SIR capture separation accuracy; SAR captures only the absence of artifacts (and hence can be high even if separation is poor).

References

[1] R. Gao and K. Grauman. Co-separating sounds of visual objects. In *ICCV*, 2019. 1

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[3] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, S. Kamali, M. Malloci, J. Pont-Tuset, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy. Open-images: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017. 1

[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1