Co-Separating Sounds of Visual Objects

Ruohan Gao¹ Kristen Grauman¹,²
¹The University of Texas at Austin, ²Facebook AI Research

Audio-Visual Source Separation

Goal: audio-visual object source separation in videos

Current approaches: Mix-and-Separate

Motivation: Image Co-Segmentation

Jointly segmenting two related images can be easier than segmenting them separately

Our Idea: Co-Separation

Co-separation: separate sounds for pairs of training videos

Training paradigm:

We detect objects in a pair of videos, and require separated sounds from detected objects to be consistently identifiable.

At test time, input is single video.

Co-separation: separate the sounds in multi-source training samples by learning to associate similar sounds with detected objects.

Experimental Results

Datasets:

MUSIC (Zhao et al. 2018, 536 solos and 149 duet videos, 11 categories)
AudioSet-Unlabeled (Gemmeke et al. 2017, >100k clips of 15 categories)

What if we train with only duets?

Co-separation overcomes the limitation of mix-and-separate when presented with multi-source training videos.

Discover object sounds:

Trained with multi-source videos, our learned audio embedding discovers object sounds in AudioSet.

Localize what is heard:

Object proposals associated with highest confidence scores

Please visit our project page for video results:

vision.cs.utexas.edu/projects/coseparation/