Co-Separating Sounds of Visual Objects

Ruohan Gao
UT Austin
rhgao@cs.utexas.edu

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

Abstract

Learning how objects sound from video is challenging, since they often heavily overlap in a single audio channel. Current methods for visually-guided audio source separation sidestep the issue by training with artificially mixed video clips, but this puts unwieldy restrictions on training data collection and may even prevent learning the properties of "true" mixed sounds. We introduce a co-separation *training paradigm that permits learning object-level sounds from unlabeled multi-source videos. Our novel training objective requires that the deep neural network's separated audio for similar-looking objects be consistently identifiable, while simultaneously reproducing accurate video-level audio tracks for each source training pair. Our approach disentangles sounds in realistic test videos, even in cases where an object was not observed individually during training. We obtain state-of-the-art results on visually-guided audio source separation and audio denoising for the MUSIC, AudioSet, and AV-Bench datasets.*

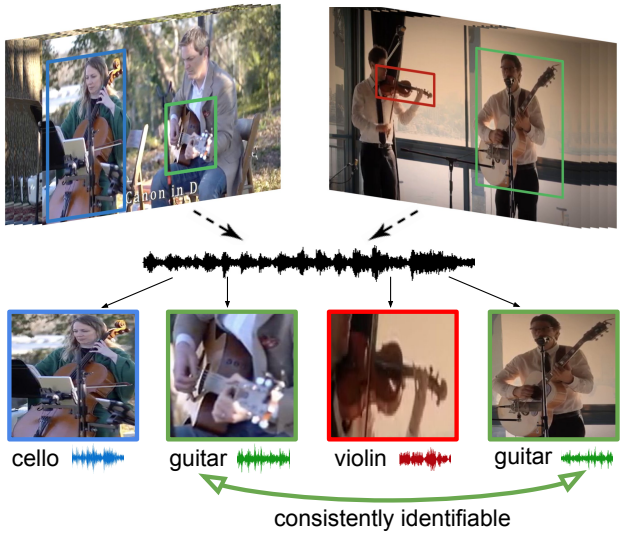

Figure 1: We propose a *co-separation* training objective to learn audio-source separation from unlabeled video containing multiple sound sources. Our approach learns to associate consistent sounds to similar-looking objects across pairs of training videos. Then, given a single novel video, it returns a separate sound track for each object. Image credit: [1, 2].

1. Introduction

Multi-modal perception is important to capture the richness of real-world sensory data for objects, scenes, and events. The sounds made by objects, whether actively generated or incidentally emitted, offer valuable signals about their physical properties and spatial locations—the cymbals crash on stage, the bird tweets up in the tree, the truck revs down the block, the silverware clinks in the drawer.

Objects often generate sounds while coexisting or interacting with other surrounding objects. Thus, rather than observe them in isolation, we hear them intertwined with sounds coming from other sources. Likewise, a realistic video records the various objects with a single audio channel that mixes all their acoustic frequencies together. Automatically *separating* the sounds of each object in a video is of great practical interest, with applications including audio denoising, audio-visual video indexing, instrument equalization, audio event remixing, and dialog following.

Whereas traditional methods assume access to multiple microphones or carefully supervised clean audio samples [22, 48, 9], recent methods tackle the audio(-visual) source separation problem using a "mix-and-separate" paradigm to train deep neural networks in a self-supervised manner [41, 50, 8, 31, 52]. Namely, such methods randomly mix audio/video clips, and the learning objective is to recover the original unmixed signals. For example, one can create "synthetic cocktail parties" that mix clean speech with other sounds [8], add pseudo "off-screen" human speakers to other real videos [31], or superimpose audio from clips of musical instruments [52].

There are two key limitations with this current training strategy. First, it implicitly assumes that the original real training videos are dominated by single-source clips containing one primary sound maker. However, gathering a large number of such clean "solo" recordings is impractical and will be difficult to scale beyond particular classes

like human speakers and musical instruments. Second, it implicitly assumes that the sources in a recording are independent. However, it is precisely the correlations *between real sound sources* (objects) that make the source separation problem most challenging at test time. Such correlations can go uncaptured by the artificially mixed training clips.

Towards addressing these shortcomings, we introduce a new strategy for learning to separate audio sources. Our key insight is a novel *co-separation* training objective that learns from naturally occurring multi-source videos.[1] During training, our co-separation network considers pairs of training videos and, rather than simply separate their artificially mixed soundtracks, it must also generate audio tracks that are *consistently identifiable at the object level* across all training samples. In particular, using noisy object detections from the unlabeled training video, we devise a loss requiring that within an individual training video, each separated audio track should be distinguishable as its proper object. For example, when two training instances both contain a guitar plus other instruments, there is pressure to make the separated guitar tracks consistently identifiable. See Fig. 1.

We call our idea "co-separation" as a loose analogy to image co-segmentation [38], whereby jointly segmenting two related images can be easier than segmenting them separately, since it allows disentangling a shared foreground object from differently cluttered backgrounds. Note, however, that our co-separation operates during training only; unlike co-segmentation, at test time our method performs separation on an individual video input.

Our method design offers the following advantages. First, co-separation allows training with "in the wild" sound mixes. It has the potential to benefit from the variability and richness of unlabeled multi-source video. Second, it enhances the supervision beyond "mix-and-separate". By enforcing separation *within a single video* at the object-level, our approach exposes the learner to natural correlations between sound sources. Finally, objects with similar appearance from different videos can partner with each other to separate their sounds jointly, thereby regularizing the learning process. In this way, our method is able to learn well from multi-source videos, and it can successfully separate an object sound in a test video even if the object has never been observed individually during training.

We experiment on three benchmark datasets and demonstrate the advantages discussed above. Our approach yields state-of-the-art results on separation and denoising. Most notably, it outperforms the prior methods and baselines by a large margin when learning from noisy AudioSet [14] videos. Overall co-separation is a promising direction to learn audio-visual separation from multi-source videos.

[1] Throughout, we use "multi-source video" as shorthand for video containing multiple sounds in its single-channel audio.

2. Related Work

Audio-Only Source Separation Audio source separation has a rich history in signal processing. While many methods assume audio captured by multiple microphones, some tackle the "blind" separation problem with single-channel audio [22, 7, 48, 9], most recently with deep learning [20, 18, 44]. Mix-and-separate style training is now commonly used for audio-only source separation to create artificial training examples [21, 18, 50]. Our approach adapts the mix-and-separate idea. However, different from all of the above, we leverage visual object detection to guide sound source separation. Furthermore, as discussed above, our co-separation framework is more flexible in terms of training data and can generalize to multi-source videos.

Audio-Visual Source Separation Early methods for audio-visual source separation focus on mutual information [10], subspace analysis [42, 34], matrix factorization [33, 39], and correlated onsets [5, 27]. Recent methods leverage deep learning for separating speech [8, 31, 3, 11], musical instruments [52, 13, 51], and other objects [12]. Similar to the audio-only methods, almost all use a "mix-and-separate" training paradigm to perform *video-level* separation by artificially mixing training videos. In contrast, we perform source separation at the *object level* to explicitly model sounds coming from visual objects, and our model enforces separation *within* a video during training.

Most related to our work are the "sound of pixels" (SoP) [52] and multi-instance learning (AV-MIML) [12] approaches. AV-MIML [12] also focuses on learning object sound models from unlabeled video, but its two-stage method relies on NMF to perform separation, which limits its performance and practicability. Furthermore, whereas AV-MIML simply uses image classification to obtain weak labels on video frames, our approach detects localized objects and our end-to-end network learns visual object representations in concert with the audio streams. SoP [52] outputs a sound for each pixel, whereas we predict sounds for visual objects with the help of a pre-trained object detector. More importantly, SoP works best when clean solo videos are available to perform video-level "mix-and-separate" training. Our method instead disentangles mixed sounds of objects within an individual training video, allowing more flexible training with multi-source data (though unlike [52] we do require an object detection step).

Localizing Sounds in Video Frames Localization entails identifying the pixels where the sound of a video comes from, but not separating the audio [25, 19, 10, 4, 40, 45]. Different from all these methods, our goal is to *separate* the sounds of multiple objects from a single-channel signal. We localize potential sound sources via object detection, and use the localized object regions to guide the separation learning process.

Generating Sounds from Video Sound generation methods synthesize a sound track from a visual input [32, 54, 6]. Given both visual input and monaural audio, recent methods generate spatial (binaural or ambisonic) audio [13, 30]. Unlike any of the above, our work aims to *separate* an existing real audio track, not synthesize plausible new sounds.

3. Approach

Our approach leverages localized object detection to visually guide audio source separation. We first formalize our object-level audio-visual source separation task (Sec. 3.1). Then we introduce our framework for learning object sound models from unlabeled video and our CO-SEPARATION deep network architecture (Sec. 3.2). Finally, we present our training criteria and inference procedures (Sec. 3.3).

3.1. Problem Formulation

Given an unlabeled video clip V with accompanying audio $x(t)$, we denote the set of N objects detected in the video frames as $\mathcal{V} = \{O_1, \ldots, O_N\}$. We treat each object as a potential sound source, and $x(t) = \sum_{n=1}^{N} s_n(t)$ is the observed single-channel linear mixture of these sources, where $s_n(t)$ are time-discrete signals responsible for each object. Our goal of object-level audio-visual source separation is to separate the sound $s_n(t)$ for each object O_n from $x(t)$.

Following [21, 18, 50, 52, 31, 13, 8], we start with the commonly adopted "mix-and-separate" idea to self-supervise source separation. Given two training videos V_1 and V_2 with corresponding audios $x_1(t)$ and $x_2(t)$, we use a pre-trained object detector to find objects in both videos. Then, we mix the audios of the two videos and obtain the mixed signal $x_m(t) = x_1(t) + x_2(t)$. The mixed audio $x_m(t)$ is transformed into a magnitude spectrogram $X^M \in \mathbb{R}_+^{F \times N}$ consisting of F frequency bins and N short-time Fourier transform (STFT) [15] frames, which encodes the change of a signal's frequency and phase content over time.

Our learning objective is to separate the sound each object makes from $x_m(t)$ conditioned on the localized object regions. For example, Fig. 3 illustrates a scenario of mixing two videos V_1 and V_2 with two objects O_1, O_2 detected in V_1 and one object O_3 detected in V_2. The goal is to separate $s_1(t)$, $s_2(t)$, and $s_3(t)$ for objects O_1, O_2, and O_3 from the mixture signal $x_m(t)$, respectively. To perform separation, we predict a spectrogram mask $\mathcal{M}_n$ for each object. We use real-valued ratio masks and obtain the predicted magnitude spectrogram by soft masking the mixture spectrogram: $X_n = X^M \times \mathcal{M}_n$. Finally, we use the inverse short-time Fourier transform (ISTFT) [15] to reconstruct the waveform sound for each object source.

Going beyond video-level mix-and-separation, the key insight of our approach is to simultaneously enforce separation *within a single video* at the object level. This enables our method to learn object sound models even from multi-source training videos. Our new co-separation framework can capture the correlations between sound sources and is able to learn from noisy Web videos, as detailed next.

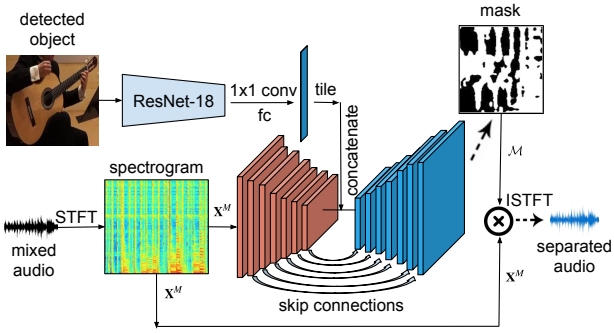

Figure 2: Our audio-visual separator network takes a mixed audio signal and a detected object from its accompanying video as input, and performs joint audio-visual analysis to separate the portion of sound responsible for the input object region.

3.2. Co-Separation Framework

Next we present our CO-SEPARATION training framework and our network architecture to perform separation.

Object Detection Firstly, we train an object detector for a vocabulary of C objects. In general, this detector should cover any potential sound-making object categories that may appear in training videos. Our implementation uses the Faster R-CNN [36] object detector with a ResNet-101 [17] backbone trained with Open Images [26]. For each unlabeled training video, we use the pre-trained object detector to automatically[2] find objects in all video frames. Then, we gather all object detections across frames to obtain a video-level pool of objects. See Supp. for details.

Audio-Visual Separator We use the detected object regions to guide the source separation process. Fig. 2 illustrates our audio-visual separator network that performs audio-visual feature aggregation and source separation. A related design for multi-modal feature fusion is also used in [13, 30, 31] for audio spatialization and separation. However, unlike those models, our separator network combines the visual features of a localized object region and the audio features of the mixed audio to predict a magnitude spectrogram mask for source separation.

The network takes a detected object region and the mixed audio signal as input, and separates the portion of the sound responsible for the object. We use a ResNet-18 network to extract visual features after the 4^{th} ResNet block with size $(H/32) \times (W/32) \times D$, where H, W, D denote the frame and channel dimensions. We then pass the visual feature through a 1×1 convolution layer to reduce the channel di-

[2] No manual object annotations are used for co-separation train-/testing.

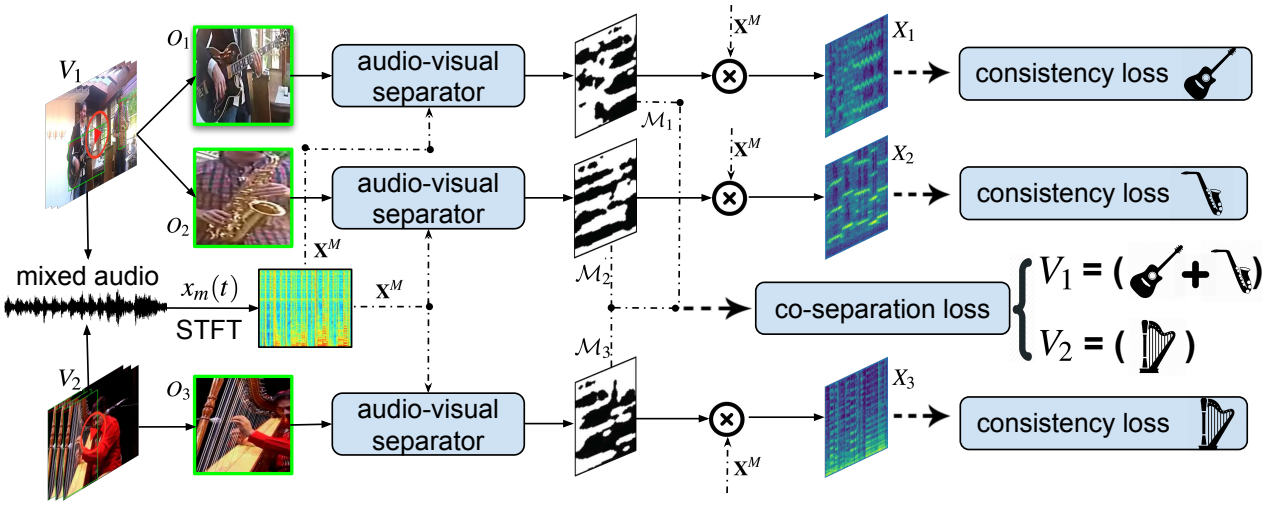

Figure 3: Co-separation training pipeline: our object-level co-separation framework first automatically detects objects in a pair of videos, then mixes the audios at the video-level, and separates the sounds for each visual object. The network is trained by minimizing the combination of the co-separation and object-consistency losses defined in Sec. 3.2.

mension, and use a fully-connected layer to obtain an aggregated visual feature vector.

On the audio side, we adopt a U-NET [37] style network for its effectiveness in dense prediction tasks, similar to [52, 31, 13]. The network takes the magnitude spectrogram X^M as input and passes it through a series of convolution layers to extract an audio feature of dimension $(T/128) \times (F/128) \times D$. We replicate the visual feature vector $(T/128) \times (F/128)$ times, tile them to match the audio feature dimension, and then concatenate the audio and visual feature maps along the channel dimension. Then a series of up-convolutions are performed on the concatenated audio-visual feature map to generate a multiplicative spectrogram mask $\mathcal{M}$. We find spectrogram masks to work better than direct prediction of spectrograms or raw waveforms for source separation, confirming reports in [49, 8, 13]. The separated spectrogram for the input object is obtained by multiplying the mask and the spectrogram of the mixed audio: $X = X^M \times \mathcal{M}$. Finally, ISTFT is applied to the spectrogram to produce the separated real-time signal.

Co-Separation Our proposed CO-SEPARATION framework first detects objects in a pair of videos, then mixes their audios at the video level, and finally separates the sounds for each detected object class. As shown in Fig. 3, for each video pair, we randomly sample a high confidence object window for each class detected in either video, and use the localized object region to guide audio source separation using the audio-visual separator network. For each object O_n, we predict a mask $\mathcal{M}_n$, and then generate the corresponding magnitude spectrogram.

Let $\mathcal{V}_1$ and $\mathcal{V}_2$ denote the set of objects for the two videos. We want to separate the sounds of their corresponding objects from the audio mixture of V_1 and V_2. For each video, summing up the separated sounds of all objects should ideally reconstruct the audio signal for that video. Namely,

$$x_1(t) = \sum_i^{|\mathcal{V}_1|} s_i(t) \quad \text{and} \quad x_2(t) = \sum_i^{|\mathcal{V}_2|} s_i(t), \tag{1}$$

where $|\mathcal{V}_1|$ and $|\mathcal{V}_2|$ are the number of detected objects for V_1 and V_2. For simplicity of notation, we defer presenting how we handle background sounds (those unattributable to detected objects) until later in this section. Because we are operating in the frequency domain, the above relationship will only hold approximately due to phase interference. As an alternative, we approximate Eq. (1) by enforcing the following relationship on the predicted magnitude spectrograms:

$$X^{V_1} \approx \sum_i^{|\mathcal{V}_1|} X_i \quad \text{and} \quad X^{V_2} \approx \sum_i^{|\mathcal{V}_2|} X_i, \tag{2}$$

where X^{V_1} and X^{V_2} are the magnitude spectrograms for $x_1(t)$ and $x_2(t)$. Therefore, we minimize the following *co-separation loss* over the separated magnitude spectrograms:

$$L_{co\text{-}separation\_spect} = ||\sum_{i=1}^{|\mathcal{V}_1|} X_i - X^{V_1}||_1 + ||\sum_{i=1}^{|\mathcal{V}_2|} X_i - X^{V_2}||_1, \tag{3}$$

which approximates to minimizing the following loss function over their predicted ratio masks:

$$L_{co\text{-}separation\_mask} = ||\sum_{i=1}^{|\mathcal{V}_1|} \mathcal{M}_i - \mathcal{M}^{V_1}||_1 + ||\sum_{i=1}^{|\mathcal{V}_2|} \mathcal{M}_i - \mathcal{M}^{V_2}||_1, \tag{4}$$

where $\mathcal{M}^{V_1}$ and $\mathcal{M}^{V_2}$ are the ground-truth spectrogram ratio masks for the two videos, respectively. Namely,

$$\mathcal{M}^{V_1} = \frac{X^{V_1}}{X^{V_1} + X^{V_2}} \quad \text{and} \quad \mathcal{M}^{V_2} = \frac{X^{V_2}}{X^{V_1} + X^{V_2}}. \tag{5}$$

In practice, we find that computing the loss over masks (vs. spectograms) makes the network easier to learn. We hypothesize that the sigmoid after the last layer of the audio-visual separator bounds the masks, making them more constrained and structured compared to spectrograms. In short, the proposed co-separation loss provides supervision to the network to only separate the audio portion responsible for the input visual object, so that the corresponding audios for each of the pair of input videos can be reconstructed.

In addition to the co-separation loss that enforces separation, we also introduce an *object-consistency loss* for each predicted audio spectrogram. The intuition is that if the sources are well-separated, the predicted "category" of the separated spectrogram should be consistent with the category of the visual object that initially guides its separation. Specifically, for the predicted spectrogram of each object, we introduce another ResNet-18 *audio* classifier[3] that targets the weak labels of the input visual objects. We use the following cross-entropy loss:

$$L_{\textit{object-consistency}} = \frac{1}{|\mathcal{V}_1|+|\mathcal{V}_2|} \sum_{i=1}^{|\mathcal{V}_1|+|\mathcal{V}_2|} \sum_{c=1}^{C} -y_{i,c} \log(p_{i,c}), \tag{6}$$

where C is the number of classes, $y_{i,c}$ is a binary indicator on whether c is the correct class for predicted spectrogram X_i, and $p_{i,c}$ is the predicted probability for class c. We stress that these audio "classes" are *discovered* during training; we have no pre-trained sound models for different objects.

Not all sounds in a video will be attributable to a visually detected object. To account for ambient sounds, off-screen sounds, and noise, we incorporate a $C+1^{st}$ "adaptable" audio class, as follows. During training, we pair each video with a visual scene feature in addition to the detected objects from the pre-trained object detector. Then an additional mask $\mathcal{M}_{\text{adapt}}$ responsible for the scene context is also predicted in Eq. (4) for both V_1 and V_2 to be optimized jointly. This step arms the network with the flexibility to assign noise or unrelated sounds to the "adaptable" class, leading to cleaner separation for sounds of the detected visual objects. These adaptable objects (ideally ambient sounds, noise, *etc*.) are collectively designated as having the "extra" $C+1^{st}$ audio label. The separated spectrograms for these adaptable objects are also trained to match their category label by the object-consistency loss in Eq. (6).

Putting it all together, during training the network needs to discover separations for the multi-source videos that 1) minimize the co-separation loss, such that the two source videos' object sounds reassemble to produce their original video-level audio tracks, respectively, while also 2) minimizing the object consistency loss, such that separated sounds for *any* instances of the same visual object are reliably identifiable as that sound. We stress that our model achieves the latter *without* any pre-trained audio model and *without* any single-source audio examples for the object class. The object consistency loss only knows that same-object sounds should be similar after training the network—not what any given object is expected to sound like.

[3]The ResNet-18 audio classifier is ImageNet pre-trained to accelerate convergence, but *not* pre-trained for audio classification. Our co-separation training aims to automatically discover the audio classes.

3.3. Training and Inference

We minimize the following combined loss function and train our network end to end:

$$L = L_{\textit{co-separation\_mask}} + \lambda L_{\textit{object-consistency}}, \tag{7}$$

where λ is the weight for the object-consistency loss.

We use per-pixel $L1$ loss for the co-separation loss, and weight the gradients by the magnitude of the spectrogram of the mixed audio. The network uses the weighted gradients to perform back-propagation, thereby emphasizing predictions on more informative parts of the spectrogram.

During testing, our model takes a *single* realistic multi-source video to perform source separation. Similarly, we first detect objects in the video frames by using the pre-trained object detector. For each detected object class, we use the most confident object region(s) as the visual input to separate the portion of the sound responsible for this object category from its accompanying audio. We use a sliding window approach to process videos segment by segment with a small hop size, and average the audio predictions on all overlapping parts.

We perform audio-visual source separation on video clips of 10s, and we pool all the detected objects in the video frames. Therefore, our approach assumes that each detected object within this period of 10s is a potential sound source, although it may only sound in some of the frames. For objects that are detected but do not make sound at all, we treat it as learning noise and expect our deep network to adapt by learning from large-scale training videos. We leave it as future work to explicitly model silent visual objects.

4. Experiments

We now validate our approach for audio-visual source separation and compare to existing methods.

4.1. Datasets

MUSIC This MIT dataset contains YouTube videos crawled with keyword queries [52]. It contains 685 untrimmed videos of musical solos and duets, with 536 solo videos and 149 duet videos. The dataset is relatively clean and collected for the purpose of training audio-visual source separation models. It includes 11 instrument categories: accordion, acoustic guitar, cello, clarinet, erhu, flute, saxophone, trumpet, tuba, violin and xylophone. Following the authors' public dataset file of video IDs, we hold

out the first/second video in each category as validation/test data, and the rest as training data. We split all videos into 10s clips during both training and testing, for a total of 8,928/259/269 train/val/test clips, respectively.

AudioSet-Unlabeled AudioSet [14] consists of challenging 10s video clips, many of poor quality and containing a variety of sound sources. Following [12], we filter the dataset to extract video clips of 15 musical instruments. We use the videos from the "unbalanced" split for training, and videos from the "balanced" split as validation/test data, for a total of 113,756/456/456 train/val/test clips, respectively.

AudioSet-SingleSource A dataset assembled by [12] of AudioSet videos containing only a single sounding object. We use the 15 videos (from the "balanced" split) of musical instruments for evaluation only.

AV-Bench This dataset contains the benchmark videos (Violin Yanni, Wooden Horse, and Guitar Solo) used in previous studies [12, 34] on visually-guided audio denoising.

On both MUSIC and AudioSet, we compose the test sets following standard practice [5, 52, 31, 12]—by mixing the audio from two single-source videos. This ensures the ground truth separated sounds are known for quantitative evaluation. There are 550 and 105 such test pairings for MUSIC and AudioSet, respectively (the result of pairwise mixing 10 random clips per the 15 classes for MUSIC and pairwise mixing all 15 clips for AudioSet). For qualitative results (Supp.), we apply our method to real multi-source test videos. In either case, we *train* our method with multi-source videos, as specified below.

4.2. Implementation Details

Our CO-SEPARATION deep network is implemented in PyTorch. For all experiments, we sub-sample the audio at 11kHz, and the input audio sample is approximately 6s long. STFT is computed using a Hann window size of 1022 and a hop length of 256, producing a 512×256 Time-Frequency audio representation. The spectrogram is then re-sampled on a log-frequency scale to obtain a $T \times F$ magnitude spectrogram of $T = 256, F = 256$. The settings are the same as [52] for fair comparison.

Our object detector is trained on images of $C = 15$ object categories from the Open Images dataset [26]. We filter out low confidence object detections for each video, and keep the top two[4] detected categories. See Supp. for details. During co-separation training, we randomly sample 64 pairs of videos for each batch. We sample a confident object detection for each class as its input visual object, paired with a random scene image sampled from the ADE dataset [53] as the adaptable object. The object window is resized to 256×256, and a randomly cropped 224×224 region is used as the input to the network. We use horizontal flipping, color and intensity jittering as data augmentation. λ is set to 0.05 in Eq. (7). The network is trained using an Adam optimizer with weight decay 1×10^{-4} with the starting learning rate set to 1×10^{-4}. We use a smaller starting learning rate of 1×10^{-5} for the ResNet-18 visual feature extractor because it is pre-trained on ImageNet.

[4]This is the number of objects detected in most training videos; relaxing this limit does not change the overall results (see Supp.).

4.3. Quantitative Results on Source Separation

We compare to the following baselines:

- **Sound-of-Pixels [52]:** We use the authors' publicly available code[5] to train 1-frame based models with ratio masks for fair comparison. Default settings are used for other hyperparameters.
- **AV-Mix-and-Separate:** A "mix-and-separate" baseline using the same audio-visual separation network as our model to do *video*-level separation. We use multi-label hinge loss to enforce video-level consistency, i.e., the class of each separated spectrogram should agree with the objects present in that training video.
- **AV-MIML [12]:** An existing audio-visual source separation method that uses audio bases learned from unlabeled videos to supervise an NMF separation process. The audio bases are learned from a deep multi-instance multi-label (MIML) learning network. We use the results reported in [12] for AudioSet and AV-Bench; the authors do not report results in SDR and do not report results for MUSIC.
- **NMF-MFCC [43]:** An off-the-shelf audio-only method that performs NMF based source separation using Mel frequency cepstrum coefficients (MFCC). This non-learned baseline is a good representation of a well established pipeline for audio-only source separation [47, 23, 24, 16].
- **AV-Loc [34], JIVE [28], Sparse CCA [25]:** We use results reported in [12] to compare to these methods for the audio denoising benchmark AV-Bench.

We use the widely used mir eval library [35] to evaluate the source separation and report the standard metrics: Signal-to-Distortion Ration (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR).

Separation Results. Tables 1 and 2 show the results for the MUSIC and AudioSet datasets, respectively.

Table 1 presents results on MUSIC as a function of the training source: single-source videos (solo) or multi-source videos (solo + duet). Our method consistently outperforms all baselines in separation accuracy, as captured by the SDR and SIR metrics.[6] While the SoP method [52] works well

[5]https://github.com/hangzhaomit/Sound-of-Pixels

[6]Note that SAR measures the *artifacts* present in the separated signal, but not the separation accuracy. So, a less well-separated signal can achieve high(er) SAR values. In fact, naively copying the original input twice (i.e., doing no separation at all) results in SAR ≈ 80 in our setting.

	Single-Source			Multi-Source		
	SDR	SIR	SAR	SDR	SIR	SAR
Sound-of-Pixels [52]	7.30	11.9	**11.9**	6.05	9.81	**12.4**
AV-Mix-and-Separate	3.16	6.74	8.89	3.23	7.01	9.14
NMF-MFCC [43]	0.92	5.68	6.84	0.92	5.68	6.84
CO-SEPARATION (Ours)	**7.38**	**13.7**	10.8	**7.64**	**13.8**	11.3

Table 1: Average audio source separation results on a held out MUSIC test set. We show the performance of our method and the baselines when training on only single-source videos (solo) and multi-source videos (solo + duet). NMF-MFCC is non-learned, so its results do not vary across training sets. Higher is better for all metrics. 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). Standard error is approximately 0.2 for all metrics.

	SDR	SIR	SAR
Sound-of-Pixels [52]	1.66	3.58	11.5
AV-MIML [12]	1.83	-	-
AV-Mix-and-Separate	1.68	3.30	12.2
NMF-MFCC [43]	0.25	4.19	5.78
CO-SEPARATION (Ours)	**4.26**	**7.07**	**13.0**

Table 2: Average separation results on AudioSet test set. Standard error is approximately 0.3.

	Sound-of-Pixels [52]			CO-SEPARATION (Ours)		
	SDR	SIR	SAR	SDR	SIR	SAR
Violin/Saxophone	1.52	1.48	12.9	8.10	11.7	11.2
Violin/Guitar	6.95	11.2	15.8	10.6	16.7	12.3
Saxophone/Guitar	0.57	0.90	16.5	5.08	7.90	9.34

Table 3: Toy experiment to demonstrate learning to separate sounds for objects never heard individually during training.

when training only on solo videos, it fails to make use of the additional duets, and its performance degrades when training on the multi-source videos. In contrast, our method actually *improves* when trained on a combination of solos and duets, achieving its best performance. This experiment highlights precisely the limitation of the mix-and-separate training paradigm when presented with multi-source training videos, and it demonstrates that our co-separation idea can successfully overcome that limitation.

Our method also outperforms all baselines, including [52], when training on solos. Our better accuracy versus the AV-Mix-and-Separate baseline and [52] shows that our object-level co-separation idea is essential. The NMF-MFCC baseline can only return ungrounded separated signals. Therefore, we evaluate both possible matchings and take its best results (to the baseline's advantage). Also, our gains are similar even if we give [52] the advantage of temporal pooling over 3 frames. Overall our method achieves large gains, and also has the benefit of matching the separated sounds to semantically meaningful visual objects in the video.

Table 2 shows the results when training on AudioSet-Unlabeled and testing on mixes of AudioSet-SingleSource. Our method outperforms all prior methods and the baselines by a large margin on this challenging dataset. It demonstrates that our framework can better learn from the noisy and less curated "in the wild" videos of AudioSet, which contains many multi-source videos. See Supp. for additional results on removing the limit of two objects per video.

Next we devise an experiment to test explicitly how well our method can learn to separate sound for objects it has not observed *individually* during training. We train our model and the best baseline [52] on the following four categories: violin solo, saxophone solo, violin+guitar duet, and violin+saxophone duet, and test by randomly mixing and separating violin, saxophone, and guitar test solo clips. Table 3 shows the results. We can see that although our system is not trained on any guitar solos, it can learn better from multi-source videos that contain guitar and other sounds. Our method consistently performs well on all three combinations, while [52] performs well only on the violin+guitar mixture. We hypothesize the reason is that it can learn by mixing the large quantity of violin solos and the guitar solo moments *within* the duets to perform separation, but it fails to disentangle other sound source correlations. Our method scores worse in terms of SAR, which again measures artifacts, but not separation quality.

See Supp. for additional experiments where we train *only* on duets as well as an ablation study to isolate the impact of each loss term.

Denoising Results. As a side product of our audio-visual source separation system, we can also use our model to perform visually-guided audio denoising. As mentioned in Sec. 3.3, we use an additional scene image to capture ambient/unseen sounds and noise, *etc*. Therefore, given a test video with noise, we can use the top detected visual object in the video to guide our system to separate out the noise.

Table 4 shows the results on AV-Bench [34, 12]. Though our method learns only from unlabeled video and does not explicitly model the low-rank nature of noise as in [34], we obtain state-of-the-art performance on 2 of the 3 videos. The method of [34] uses motion in manually segmented regions, which may help on Guitar Solo, where the hand's motion strongly correlates with the sound.

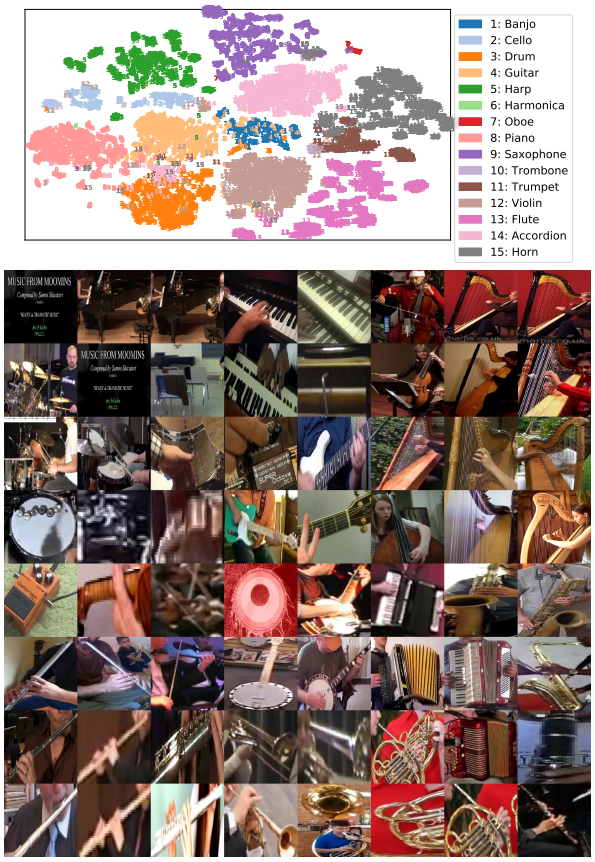

Figure 4: Embedding of separated sounds in AudioSet visualized with t-SNE in two ways: (top) categories are color-coded, and (bottom) visual objects are shown at their sound's embedding.

	Wooden Horse	Violin Yanni	Guitar Solo
Sparse CCA [25]	4.36	5.30	5.71
JIVE [28]	4.54	4.43	2.64
AV-Loc [34]	8.82	5.90	**14.1**
AV-MIML [12]	12.3	7.88	11.4
Ours	**14.5**	**8.53**	11.9

Table 4: Visually-assisted audio denoising on AV-Bench, in terms of NSDR (in dB, higher is better).

4.4. Qualitative Results

Audio-Visual Separation Video Examples. Our video results (see Supp.) show qualitative separation results. We use our system to discover and separate object sounds for realistic multi-source videos. They lack ground truth, but the results can be manually inspected for quality.

Learned Audio Embedding. To *visualize* that our CO-SEPARATION network has indeed learned to separate sounds of visual objects, Fig. 4 displays a t-SNE [29] embedding of the discovered sounds for various input objects in 20K AudioSet clips. We use the features extracted at the last layer of the ResNet-18 audio classifier as the audio representation for the separated spectrograms. The sounds our method learned from multi-source videos tend to cluster by object category, demonstrating that the separator discovers sounds characteristic of the corresponding objects.

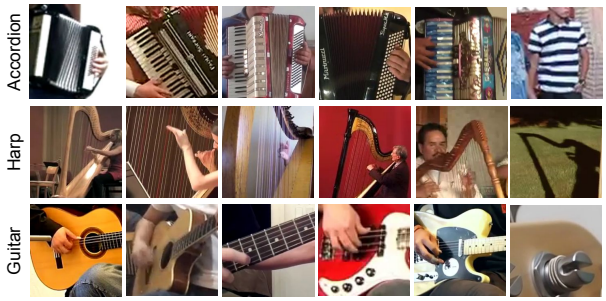

Figure 5: Top object proposals according to our discovered *audio* classifier. Last column shows typical failure cases.

Using Discovered Sounds to Detect Objects. Finally, we use our trained audio-visual source separation network for *visual object discovery* using 912 noisy unseen videos from AudioSet. Given the pool of videos, we generate object region proposals using Selective Search [46]. Then we pass these region proposals to our network together with the audio of its accompanying video, and retrieve the *visual* proposals that achieve the highest *audio* classification scores according to our object consistency loss.

Fig. 5 shows the top retrieved proposals for several categories after removing duplicates from the same video. We can see that our method has learned a good mapping between the visual and audio modalities; the best visual object proposals usually best activate the audio classifier. The last column shows failure cases where the wrong object is detected with high confidence. They usually come from objects of similar texture or shape, like the stripes on the man's t-shirt and the shadow of the harp.

5. Conclusion

We presented an object-level audio-visual source separation framework that associates localized object regions in videos to their characteristic sounds. Our CO-SEPARATION approach can leverage noisy object detections as supervision to learn from large-scale unlabeled videos. We achieve state-of-the-art results on visually-guided audio source separation and audio denoising. As future work, we plan to explore spatio-temporal object proposals and incorporate object motion to guide separation, which may especially benefit object sounds with similar frequencies.

Acknowledgements: Thanks to Dongguang You for help with experiments setup, and Yu-Chuan Su, Tushar Nagarajan, Santhosh Ramakrishnan and Xingyi Zhou for helpful discussions and reading paper drafts. UT Austin is supported in part by DARPA Lifelong Learning Machines.

References

[1] Bryan lucas band - acoustic guitar and cello wedding ceremony music. https://www.youtube.com/watch?v=_oRFIsHn2ek. 1

[2] Shape of my heart violin and guitar wedding ceremony sydney manly q station. https://www.youtube.com/watch?v=u12_qzOriVg. 1

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. In *Interspeech*, 2018. 2

[4] Relja Arandjelović and Andrew Zisserman. Objects that sound. In *ECCV*, 2018. 2

[5] Zohar Barzelay and Yoav Y Schechner. Harmony in motion. In *CVPR*, 2007. 2, 6

[6] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep cross-modal audio-visual generation. In *on Thematic Workshops of ACM Multimedia*, 2017. 3

[7] Daniel Patrick Whittlesey Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, Massachusetts Institute of Technology, 1996. 2

[8] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. In *SIGGRAPH*, 2018. 1, 2, 3, 4

[9] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation*, 2009. 1, 2

[10] John W Fisher III, Trevor Darrell, William T Freeman, and Paul A Viola. Learning joint statistical models for audio-visual fusion and segregation. In *NeurIPS*, 2001. 2

[11] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. In *Interspeech*, 2018. 2

[12] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018. 2, 6, 7, 8

[13] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *CVPR*, 2019. 2, 3, 4

[14] 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 *ICASSP*, 2017. 2, 6

[15] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984. 3

[16] Xin Guo, Stefan Uhlich, and Yuki Mitsufuji. Nmf-based blind source separation using a linear predictive coding error clustering criterion. In *ICASSP*, 2015. 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[18] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *ICASSP*, 2016. 2, 3

[19] John R Hershey and Javier R Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *NeurIPS*, 2000. 2

[20] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Deep learning for monaural speech separation. In *ICASSP*, 2014. 2

[21] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015. 2, 3

[22] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 2000. 1, 2

[23] Satoshi Innami and Hiroyuki Kasai. Nmf-based environmental sound source separation using time-variant gain features. *Computers & Mathematics with Applications*, 2012. 6

[24] Rajesh Jaiswal, Derry FitzGerald, Dan Barry, Eugene Coyle, and Scott Rickard. Clustering nmf basis functions using shifted nmf for monaural sound source separation. In *ICASSP*, 2011. 6

[25] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *CVPR*, 2005. 2, 6, 8

[26] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: 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. 3, 6

[27] Bochen Li, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. See and listen: Score-informed association of sound tracks to players in chamber music performance videos. In *ICASSP*, 2017. 2

[28] Eric F Lock, Katherine A Hoadley, James Stephen Marron, and Andrew B Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, 2013. 6, 8

[29] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 8

[30] Pedro Morgado, Nono Vasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360° video. In *NeurIPS*, 2018. 3

[31] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 1, 2, 3, 4, 6

[32] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *CVPR*, 2016. 3

[33] Sanjeel Parekh, Slim Essid, Alexey Ozerov, Ngoc QK Duong, Patrick Pérez, and Gaël Richard. Motion informed audio source separation. In *ICASSP*, 2017. 2

[34] Jie Pu, Yannis Panagakis, Stavros Petridis, and Maja Pantic. Audio-visual object localization and separation using low-rank and sparsity. In *ICASSP*, 2017. 2, 6, 7, 8

[35] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. mir_eval: A transparent implementation of common mir metrics. In *ISMIR*, 2014. 6

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 3

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 4

[38] Carsten Rother, Tom Minka, Andrew Blake, and Vladimir Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In *CVPR*, 2006. 2

[39] Farnaz Sedighin, Massoud Babaie-Zadeh, Bertrand Rivet, and Christian Jutten. Two multimodal approaches for single microphone source separation. In *24th European Signal Processing Conference*, 2016. 2

[40] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018. 2

[41] Andrew JR Simpson, Gerard Roma, and Mark D Plumbley. Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In *International Conference on Latent Variable Analysis and Signal Separation*, 2015. 1

[42] Paris Smaragdis and Michael Casey. Audio/visual independent components. In *International Conference on Independent Component Analysis and Signal Separation*, 2003. 2

[43] Martin Spiertz and Volker Gnann. Source-filter based clustering for monaural blind source separation. In *12th International Conference on Digital Audio Effects*, 2009. 6, 7

[44] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Adversarial semi-supervised audio source separation applied to singing voice extraction. In *ICASSP*, 2018. 2

[45] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018. 2

[46] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *IJCV*, 2013. 8

[47] Tuomas Virtanen. Sound source separation using sparse coding with temporal continuity objective. In *International Computer Music Conference*, 2003. 6

[48] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 2007. 1, 2

[49] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018. 4

[50] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *ICASSP*, 2017. 1, 2, 3

[51] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *ICCV*, 2019. 2

[52] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018. 1, 2, 3, 4, 5, 6, 7

[53] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 6

[54] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*, 2018. 3