Learning to Separate Object Sounds by Watching Unlabeled Video

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**

Abstract. Perceiving a scene most fully requires all the senses. Yet modeling how objects look and sound is challenging: most natural scenes and events contain multiple objects, and the audio track mixes all the sound sources together. We propose to learn audio-visual object models from unlabeled video, then exploit the visual context to perform audio source separation in novel videos. Our approach relies on a deep multi-instance multi-label learning framework to disentangle the audio frequency bases that map to individual visual objects, even without observing/hearing those objects in isolation. We show how the recovered disentangled bases can be used to guide audio source separation to obtain better-separated, object-level sounds. Our work is the first to learn audio source separation from large-scale "in the wild" videos containing multiple audio sources per video. We obtain state-of-the-art results on visually-aided audio source separation and audio denoising. Our video results: http://vision.cs.utexas.edu/projects/separating_object_sounds/

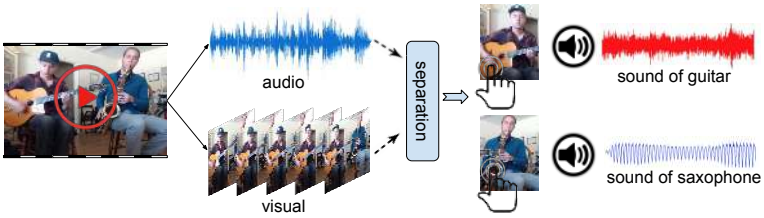

Fig. 1. Goal: Learn from unlabeled video to separate object sounds

1 Introduction

Understanding scenes and events is inherently a multi-modal experience. We perceive the world by both looking and listening (and touching, smelling, and tasting). Objects generate unique sounds due to their physical properties and interactions with other objects and the environment. For example, perception of a coffee shop scene may include seeing cups, saucers, people, and tables, but also hearing the dishes clatter, the espresso machine grind, and the barista shouting

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

an order. Human developmental learning is also inherently multi-modal, with young children quickly amassing a repertoire of objects and their sounds: dogs bark, cats mew, phones ring.

However, while recognition has made significant progress by "looking"—detecting objects, actions, or people based on their appearance—it often does not listen. Despite a long history of audio-visual video indexing [45,57,70,71,79], *objects* in video are often analyzed as if they were silent entities in silent environments. A key challenge is that in a realistic video, object sounds are observed not as separate entities, but as a *single audio channel* that mixes all their frequencies together. Audio source separation, though studied extensively in the signal processing literature [25, 40, 76, 86], remains a difficult problem with natural data outside of lab settings. Existing methods perform best by capturing the input with multiple microphones, or else assume a clean set of single source audio examples is available for supervision (e.g., a recording of only a violin, another recording containing only a drum, etc.), both of which are very limiting prerequisites. The blind audio separation task evokes challenges similar to image segmentation—and perhaps more, since all sounds overlap in the input signal.

Our goal is to learn how different objects sound by both looking at *and* listening to unlabeled video containing multiple sounding objects. We propose an unsupervised approach to disentangle mixed audio into its component sound sources. The key insight is that observing sounds in a variety of visual contexts reveals the cues needed to isolate individual audio sources; the different visual contexts lend weak supervision for discovering the associations. For example, having experienced various instruments playing in various combinations before, then given a video with a guitar and a saxophone (Fig. 1), one can naturally anticipate what sounds could be present in the accompanying audio, and therefore better separate them. Indeed, neuroscientists report that the mismatch negativity of event-related brain potentials, which is generated bilaterally within auditory cortices, is elicited only when the visual pattern promotes the segregation of the sounds [63]. This suggests that synchronous presentation of visual stimuli should help to resolve sound ambiguity due to multiple sources, and promote either an integrated or segregated perception of the sounds.

We introduce a novel audio-visual source separation approach that realizes this intuition. Our method first leverages a large collection of unannotated videos to discover a latent sound representation for each object. Specifically, we use state-of-the-art image recognition tools to infer the objects present in each video clip, and we perform non-negative matrix factorization (NMF) on each video's audio channel to recover its set of frequency basis vectors. At this point it is unknown which audio bases go with which visible object(s). To recover the association, we construct a neural network for multi-instance multi-label learning (MIML) that maps audio bases to the distribution of detected visual objects. From this audio basis-object association network, we extract the audio bases linked to each visual object, yielding its prototypical spectral patterns. Finally, given a novel video, we use the learned per-object audio bases to steer audio source separation.

Prior attempts at visually-aided audio source separation tackle the problem by detecting low-level correlations between the two data streams for the input video [8, 12, 15, 27, 52, 61, 62, 64], and they experiment with somewhat controlled domains of musical instruments in concert or human speakers facing the camera. In contrast, we propose to *learn object-level sound models* from hundreds of thousands of unlabeled videos, and generalize to separate new audio-visual instances. We demonstrate results for a broad set of "in the wild" videos. While a resurgence of research on cross-modal learning from images and audio also capitalizes on synchronized audio-visual data for various tasks [3, 4, 5, 47, 49, 59, 60], they treat the audio as a single monolithic input, and thus cannot associate different sounds to different objects in the same video.

The main contributions in this paper are as follows. Firstly, we propose to enhance audio source separation in videos by "supervising" it with visual information from image recognition results[1]. Secondly, we propose a novel deep multi-instance multi-label learning framework to learn prototypical spectral patterns of different acoustic objects, and inject the learned prior into an NMF source separation framework. Thirdly, to our knowledge, we are the first to study audio source separation learned from large scale online videos. We demonstrate state-of-the-art results on visually-aided audio source separation and audio denoising.

2 Related Work

Localizing sounds in video frames The sound localization problem entails identifying which pixels or regions in a video are responsible for the recorded sound. Early work on localization explored correlating pixels with sounds using mutual information [27, 37] or multi-modal embeddings like canonical correlation analysis [47], often with assumptions that a sounding object is in motion. Beyond identifying correlations for a single input video's audio and visual streams, recent work investigates learning associations from many such videos in order to localize sounding objects [4]. Such methods typically assume that there is *one* sound source, and the task is to localize the portion(s) of the visual content responsible for it. In contrast, our goal is to *separate* multiple audio sources from a monoaural signal by leveraging learned audio-visual associations.

Audio-visual representation learning Recent work shows that image and audio classification tasks can benefit from representation learning with both modalities. Given unlabeled training videos, the audio channel can be used as free self-supervision, allowing a convolutional network to learn features that tend to gravitate to objects and scenes, resulting in improved image classification [3, 60]. Working in the opposite direction, the SoundNet approach uses image classifier predictions on unlabeled video frames to guide a learned audio representation for improved audio scene classification [5]. For applications in cross-modal retrieval or zero-shot classification, other methods aim to learn aligned representations

[1] Our task can hence be seen as "weakly supervised", though the weak "labels" themselves are inferred from the video, not manually annotated.

across modalities, e.g., audio, text, and visual [6]. Related to these approaches, we share the goal of learning from unlabeled video with synchronized audio and visual channels. However, whereas they aim to improve audio or image classification, our method discovers associations in order to isolate sounds per object, with the ultimate task of audio-visual source separation.

Audio source separation Audio source separation (from purely audio input) has been studied for decades in the signal processing literature. Some methods assume access to multiple microphones, which facilitates separation [20, 56, 82]. Others accept a single monoaural input [39,69,72,76,77] to perform "blind" separation. Popular approaches include Independent Component Analysis (ICA) [40], sparse decomposition [86], Computational Auditory Scene Analysis (CASA) [22], non-negative matrix factorization (NMF) [25,26,51,76], probabilistic latent variable models [38,68], and deep learning [36,39,66]. NMF is a traditional method that is still widely used for unsupervised source separation [31,41,44,72,75]. However, existing methods typically require supervision to get good results. Strong supervision in the form of isolated recordings of individual sound sources [69,77] is effective but difficult to secure for arbitrary sources in the wild. Alternatively, "informed" audio source separation uses special-purpose auxiliary cues to guide the process, such as a music score [35], text [50], or manual user guidance [11,19,77]. Our approach employs an existing NMF optimization [26], chosen for its efficiency, but unlike any of the above we tackle audio separation informed by automatically detected visual objects.

Audio-visual source separation The idea of guiding audio source separation using *visual* information can be traced back to [15, 27], where mutual information is used to learn the joint distribution of the visual and auditory signals, then applied to isolate human speakers. Subsequent work explores audio-visual subspace analysis [62, 67], NMF informed by visual motion [61, 65], statistical convolutive mixture models [64], and correlating temporal onset events [8, 52]. Recent work [62] attempts both localization and separation simultaneously; however, it assumes a moving object is present and only aims to decompose a video into background (assumed low-rank) and foreground sounds/pixels. Prior methods nearly always tackle videos of people speaking or playing musical instruments [8, 12, 15, 27, 52, 61, 62, 64]—domains where salient motion signals accompany audio events (e.g., a mouth or a violin bow starts moving, a guitar string suddenly accelerates). Some studies further assume side cues from a written musical score [52], require that each sound source has a period when it alone is active [12], or use ground-truth motion captured by MoCap [61].

Whereas prior work correlates low-level visual patterns—particularly motion and onset events—with the audio channel, we propose to learn from video how different *objects* look and sound, whether or not an object moves with obvious correlation to the sounds. Our method assumes access to visual detectors, but assumes no side information about a novel test video. Furthermore, whereas existing methods analyze a single input video in isolation and are largely constrained to human speakers and instruments, our approach learns a valuable prior for audio separation from a large library of *unlabeled* videos.

Concurrent with our work, other new methods for audio-visual source separation are being explored specifically for speech [1, 23, 28, 58] or musical instruments [84]. In contrast, we study a broader set of object-level sounds including instruments, animals, and vehicles. Moreover, our method's training data requirements are distinctly more flexible. We are the first to learn from uncurated "in the wild" videos that contain multiple objects and multiple audio sources.

Generating sounds from video More distant from our work are methods that aim to *generate* sounds from a silent visual input, using recurrent networks [59, 85], conditional generative adversarial networks (C-GANs) [13], or simulators integrating physics, audio, and graphics engines [83]. Unlike any of the above, our approach learns the association between how objects look and sound in order to disentangle real audio sources; our method does not aim to synthesize sounds.

Weakly supervised visual learning Given unlabeled video, our approach learns to disentangle which sounds within a mixed audio signal go with which recognizable objects. This can be seen as a weakly supervised visual learning problem, where the "supervision" in our case consists of automatically detected visual objects. The proposed setting of weakly supervised audio-visual learning is entirely novel, but at a high level it follows the spirit of prior work leveraging weak annotations, including early "words and pictures" work [7,21], internet vision methods [9,73], training weakly supervised object (activity) detectors [2,10,14,16,78], image captioning methods [18,46], or grounding acoustic units of spoken language to image regions [32, 33]. In contrast to any of these methods, our idea is to learn *sound* associations for objects from unlabeled video, and to exploit those associations for audio source separation on new videos.

3 Approach

Our approach learns what objects sound like from a batch of unlabeled, multi-sound-source videos. Given a new video, our method returns the separated audio channels and the visual objects responsible for them.

We first formalize the audio separation task and overview audio basis extraction with NMF (Sec. 3.1). Then we introduce our framework for learning audio-visual objects from unlabeled video (Sec. 3.2) and our accompanying deep multi-instance multi-label network (Sec. 3.3). Next we present an approach to use that network to associate audio bases with visual objects (Sec. 3.4). Finally, we pose audio source separation for novel videos in terms of a semi-supervised NMF approach (Sec. 3.5).

3.1 Audio Basis Extraction

Single-channel audio source separation is the problem of obtaining an estimate for each of the J sources s_j from the observed linear mixture $x(t)$: $x(t) = \sum_{j=1}^{J} s_j(t)$, where $s_j(t)$ are time-discrete signals. The mixture signal can be transformed into a magnitude or power spectrogram $\mathbf{V} \in \mathbb{R}_{+}^{F \times N}$ consisting of F frequency bins and N short-time Fourier transform (STFT) [30] frames, which

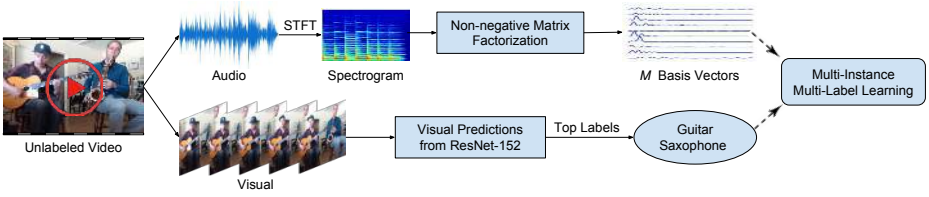

Fig. 2. Unsupervised training pipeline. For each video, we perform NMF on its audio magnitude spectrogram to get M basis vectors. An ImageNet-trained ResNet-152 network is used to make visual predictions to find the potential objects present in the video. Finally, we perform multi-instance multi-label learning to disentangle which extracted audio basis vectors go with which detected visible object(s).

encode the change of a signal's frequency and phase content over time. We operate on the frequency domain, and use the inverse short-time Fourier transform (ISTFT) [30] to reconstruct the sources.

Non-negative matrix factorization (NMF) is often employed [25,26,51,76] to approximate the (non-negative real-valued) spectrogram matrix $\mathbf{V}$ as a product of two matrices $\mathbf{W}$ and $\mathbf{H}$:

$$\mathbf{V} \approx \tilde{\mathbf{V}} = \mathbf{W}\mathbf{H}, \tag{1}$$

where $\mathbf{W} \in \mathbb{R}_+^{F \times M}$ and $\mathbf{H} \in \mathbb{R}_+^{M \times N}$. The number of bases M is a user-defined parameter. $\mathbf{W}$ can be interpreted as the non-negative audio spectral patterns, and $\mathbf{H}$ can be seen as the activation matrix. Specifically, each column of $\mathbf{W}$ is referred to as a *basis vector*, and each row in $\mathbf{H}$ represents the gain of the corresponding basis vector. The factorization is usually obtained by solving the following minimization problem:

$$\min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}|\mathbf{W}\mathbf{H}) \text{ subject to } \mathbf{W} \geq 0, \mathbf{H} \geq 0, \tag{2}$$

where D is a measure of divergence, e.g., we employ the Kullback-Leibler (KL) divergence.

For each unlabeled training video, we perform NMF independently on its audio magnitude spectrogram to obtain its spectral patterns $\mathbf{W}$, and throw away the activation matrix $\mathbf{H}$. M audio basis vectors are therefore extracted from each video.

3.2 Weakly-Supervised Audio-Visual Object Learning Framework

Multiple objects can appear in an unlabeled video at the same time, and similarly in the associated audio track. At this point, it is unknown which of the audio bases extracted (columns of $\mathbf{W}$) go with which visible object(s) in the visual frames. To discover the association, we devise a multi-instance multi-label learning (MIML) framework that matches audio bases with the detected objects.

As shown in Fig. 2, given an unlabeled video, we extract its visual frames and the corresponding audio track. As defined above, we perform NMF independently on the magnitude spetrogram of each audio track and obtain M basis

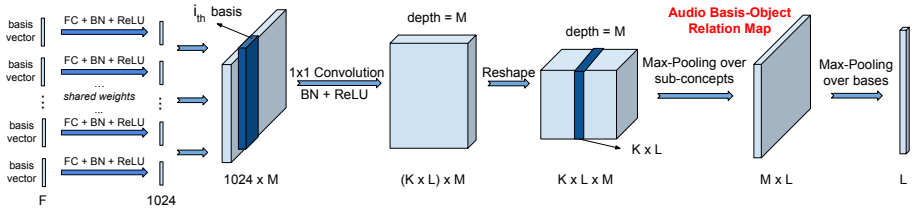

Fig. 3. Our deep multi-instance multi-label network takes a bag of M audio basis vectors for each video as input, and gives a bag-level prediction of the objects present in the audio. The visual predictions from an ImageNet-trained CNN are used as weak "labels" to train the network with unlabeled video.

vectors from each video. For the visual frames, we use an ImageNet pre-trained ResNet-152 network [34] to make object category predictions, and we max-pool over predictions of all frames to obtain a video-level prediction. The top labels (with class probability larger than a threshold) are used as weak "labels" for the unlabeled video. The extracted basis vectors and the visual predictions are then fed into our MIML learning framework to discover associations, as defined next.

3.3 Deep Multi-Instance Multi-Label Network

We cast the audio basis-object disentangling task as a multi-instance multi-label (MIML) learning problem. In single-label MIL [17], one has bags of instances, and a bag label indicates only that some number of the instances within it have that label. In MIML, the bag can have multiple labels, and there is ambiguity about which labels go with which instances in the bag.

We design a deep MIML network for our task. A bag of basis vectors $\{\mathbf{B}\}$ is the input to the network, and within each bag there are M basis vectors $\mathbf{B}_i$ with $i \in [1, M]$ extracted from one video. The "labels" are only available at the bag level, and come from noisy visual *predictions* of the ResNet-152 network trained for ImageNet recognition. The labels for each instance (basis vector) are unknown. We incorporate MIL into the deep network by modeling that there must be *at least one* audio basis vector from a certain object that constitutes a positive bag, so that the network can output a correct bag-level prediction that agrees with the visual prediction.

Fig. 3 shows the detailed network architecture. M basis vectors are fed through a Siamese Network of M branches with shared weights. The Siamese network is designed to reduce the dimension of the audio frequency bases and learns the audio spectral patterns through a fully-connected layer (FC) followed by batch norm (BN) [42] and a rectified linear unit (ReLU). The output of all branches are stacked to form a $1024 \times M$ dimension feature map. Each slice of the feature map represents a basis vector with reduced dimension. Inspired by [24], each label is decomposed to K sub-concepts to capture latent semantic meanings. For example, for drum, the latent sub-concepts could be different types of drums, such as bongo drum, tabla, and so on. The stacked output from the Siamese network is forwarded through a 1×1 Convolution-BN-ReLU mod-

ule, and then reshaped into a feature cube of dimension $K \times L \times M$, where K is the number of sub-concepts, L is the number of object categories, and M is the number of audio basis vectors. The depth of the tensor equals the number of input basis vectors, with each $K \times L$ slice corresponding to one particular basis. The activation score of the $(k, l, m)_{\text{th}}$ node in the cube represents the matching score of the k_{th} sub-concept of the l_{th} label for the m_{th} basis vector.

To get a bag-level prediction, we conduct two max-pooling operations. Max pooling in deep MIL [24,80,81] is typically used to identify the positive instances within an aggregated bag. Our first pooling is over the sub-concept dimension (K) to generate an audio basis-object relation map. The second max-pooling operates over the basis dimension (M) to produce a video-level prediction. We use the following multi-label hinge loss to train the network:

$$\mathcal{L}(A, \mathcal{V}) = \frac{1}{L} \sum_{i=1, i \neq \mathcal{V}_j}^{L} \sum_{j=1}^{|\mathcal{V}|} \max[0, 1 - (A_{\mathcal{V}_j} - A_i)], \tag{3}$$

where $A \in \mathbb{R}^L$ is the output of the MIML network, and represents the object predictions based on audio bases; $\mathcal{V}$ is the set of visual objects, namely the indices of the $|\mathcal{V}|$ objects predicted by the ImageNet-trained model. The loss function encourages the prediction scores of the correct classes to be larger than incorrect ones by a margin of 1. We find these pooling steps in our MIML formulation are valuable to learn accurately from the ambiguously "labeled" bags (i.e., the videos and their object predictions); see Supp.

3.4 Disentangling Per-Object Bases

The MIML network above learns from audio-visual associations, but does not itself disentangle them. The sounds in the audio track and objects present in the visual frames of unlabeled video are diverse and noisy (see Sec. 4.1 for details about the data we use). The audio basis vectors extracted from each video could be a component shared by multiple objects, a feature composed of them, or even completely unrelated to the predicted visual objects. The visual predictions from ResNet-152 network give approximate predictions about the objects that could be present, but are certainly not always reliable (see Fig. 5 for examples).

Therefore, to collect high quality representative bases for each object category, we use our trained deep MIML network as a tool. 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. By examining the normalized map, we can discover links from 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). See Supp. for details, and see Fig. 5 for examples of typical basis-object relation maps. In short, at the end of this phase, we have a set of audio bases for each visual object, discovered purely from unlabeled video and mixed single-channel audio.

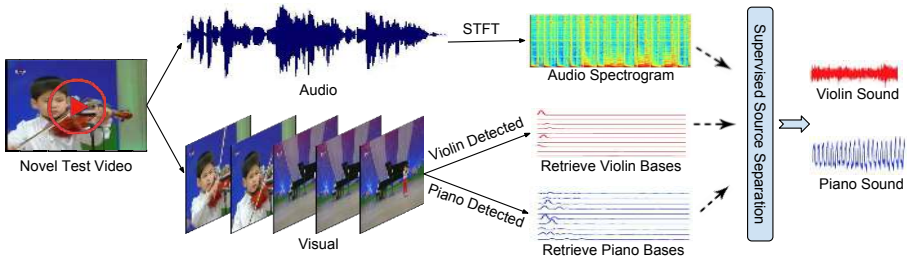

Fig. 4. Testing pipeline. Given a novel test video, we detect the objects present in the visual frames, and retrieve their learnt audio bases. The bases are collected to form a fixed basis dictionary $\mathbf{W}$ with which to guide NMF factorization of the test video's audio channel. The basis vectors and the learned activation scores from NMF are finally used to separate the sound for each detected object, respectively.

3.5 Object Sound Separation for a Novel Video

Finally, we present our procedure to separate audio sources in new videos. As shown in Fig. 4, given a novel test video q, we obtain its audio magnitude spectrogram $\mathbf{V}^{(q)}$ through STFT and detect objects using the same ImageNet-trained ResNet-152 network as before. Then, we retrieve the learnt audio basis vectors for each detected object, and use them to "guide" NMF-based audio source separation. Specifically,

$$\begin{aligned}\mathbf{V}^{(q)} \approx \tilde{\mathbf{V}}^{(q)} &= \mathbf{W}^{(q)}\mathbf{H}^{(q)} \\ &= \left[\mathbf{W}_1^{(q)} \cdots \mathbf{W}_j^{(q)} \cdots \mathbf{W}_J^{(q)}\right] \left[\mathbf{H}_1^{(q)} \cdots \mathbf{H}_j^{(q)} \cdots \mathbf{H}_J^{(q)}\right]^T,\end{aligned} \tag{4}$$

where J is the number of detected objects (J potential sound sources), and $\mathbf{W}_j^{(q)}$ contains the retrieved bases corresponding to object j in input video q. In other words, we concatenate the basis vectors learnt for each detected object to construct the basis dictionary $\mathbf{W}^{(q)}$. Next, in the NMF algorithm, we hold $\mathbf{W}^{(q)}$ fixed, and only estimate activations $\mathbf{H}^{(q)}$ with multiplicative update rules. Then we obtain the spectrogram corresponding to each detected object by $\mathbf{V}_j^{(q)} = \mathbf{W}_j^{(q)}\mathbf{H}_j^{(q)}$. We reconstruct the individual (compressed) audio source signals by soft masking the mixture spectrogram:

$$\mathbb{V}_j = \frac{\mathbf{V}_j^{(q)}}{\sum_{i=1}^{J} \mathbf{V}_i^{(q)}} \mathbb{V}, \tag{5}$$

where $\mathbb{V}$ contains both magnitude and phase. Finally, we perform ISTFT on $\mathbb{V}_j$ to reconstruct the audio signals for each detected object. If a detected object does not make sound, then its estimated activation scores will be low. This phase can be seen as a self-supervised form of NMF, where the detected visual objects reveal which bases (previously discovered from unlabeled videos) are relevant to guide audio separation.

4 Experiments

We now validate our approach and compare to existing methods.

4.1 Datasets

We consider two public video datasets: AudioSet [29] and the benchmark videos from [43, 53, 62], which we refer to as AV-Bench.

AudioSet-Unlabeled: We use AudioSet [29] as the source of unlabeled training videos[2]. The dataset consists of short 10 second video clips that often concentrate on one event. However, our method makes no particular assumptions about using short or trimmed videos, as it learns bases in the frequency domain and pools both visual predictions and audio bases from all frames. The videos are challenging: many are of poor quality and unrelated to object sounds, such as silence, sine wave, echo, infrasound, etc. As is typical for related experimentation in the literature [4,85], we filter the dataset to those likely to display audio-visual events. In particular, we extract musical instruments, animals, and vehicles, which span a broad set of unique sound-making objects. See Supp. for a complete list of the object categories. Using the dataset's provided split, we randomly reserve some videos from the "unbalanced" split as validation data, and the rest as the training data. We use videos from the "balanced" split as test data. The final AudioSet-Unlabeled data contains 104k, 2.9k, 1k / 22k, 1.2k, 0.5k / 58k, 2.4k, 0.6k video clips in the train, val, test splits, for the instruments, animals, and vehicles, respectively.

AudioSet-SingleSource: To facilitate quantitative evaluation (cf. Sec. 4.4), we construct a dataset of AudioSet videos containing only a single sounding object. We manually examine videos in the val/test set, and obtain 23 such videos. There are 15 musical instruments (accordion, acoustic guitar, banjo, cello, drum, electric guitar, flute, french horn, harmonica, harp, marimba, piano, saxophone, trombone, violin), 4 animals (cat, dog, chicken, frog), and 4 vehicles (car, train, plane, motorbike). Note that our method never uses these samples for training.

AV-Bench: This dataset contains the benchmark videos (Violin Yanni, Wooden Horse, and Guitar Solo) used in previous studies [43, 53, 62].

4.2 Implementation Details

We extract a 10 second audio clip and 10 frames (every 1s) from each video. Following common settings [3], the audio clip is resampled at 48 kHz, and converted into a magnitude spectrogram of size 2401×202 through STFT of window length 0.1s and half window overlap. We use the NMF implementation of [26] with KL divergence and the multiplicative update solver. We extract $M = 25$ basis vectors from each audio. All video frames are resized to 256×256, and 224×224 center crops are used to make visual predictions. We use all relevant ImageNet categories and group them into 23 classes by merging the posteriors of similar categories to roughly align with the AudioSet categories; see Supp. A softmax

[2] AudioSet offers noisy video-level audio class annotations. However, we do not use any of its label information.

is finally performed on the video-level object prediction scores, and classes with probability greater than 0.3 are kept as weak labels for MIML training. The deep MIML network is implemented in PyTorch with $F = 2,401$, $K = 4$, $L = 25$, and $M = 25$. We report all results with these settings and did not try other values. The network is trained using Adam [48] with weight decay 10^{-5} and batch size 256. The starting learning rate is set to 0.001, and decreased by 6% every 5 epochs and trained for 300 epochs.

4.3 Baselines

We compare to several existing methods [47, 55, 62, 72] and multiple baselines:

MFCC Unsupervised Separation [72]: This is an off-the-shelf unsupervised audio source separation method. The separated channels are first converted into Mel frequency cepstrum coefficients (MFCC), and then K-means clustering is used to group separated channels. This is an established pipeline in the literature [31, 41, 44, 75], making it a good representative for comparison. We use the publicly available code[3].

AV-Loc [62], JIVE [55], Sparse CCA [47]: We refer to results reported in [62] for the AV-Bench dataset to compare to these methods.

AudioSet Supervised Upper-Bound: This baseline uses AudioSet ground-truth labels to train our deep MIML network. AudioSet labels are organized in an ontology and each video is labeled by many categories. We use the 23 labels aligned with our subset (15 instruments, 4 animals, and 4 vehicles). This baseline serves as an upper-bound.

K-means Clustering Unsupervised Separation: We use the same number of basis vectors as our method to initialize the $\mathbf{W}$ matrix, and perform unsupervised NMF. K-means clustering is then used to group separated channels, with K equal to the number of ground-truth sources. The sound sources are separated by aggregating the channel spectrograms belonging to each cluster.

Visual Exemplar for Supervised Separation: We recognize objects in the frames, and retrieve bases from an exemplar video for each detected object class to supervise its NMF audio source separation. An exemplar video is the one that has the largest confidence score for a class among all unlabeled training videos.

Unmatched Bases for Supervised Separation: This baseline is the same as our method except that it retrieves bases of the wrong class (at random from classes absent in the visual prediction) to guide NMF audio source separation.

Gaussian Bases for Supervised Separation: We initialize the weight matrix $\mathbf{W}$ randomly using a Gaussian distribution, and then perform supervised audio source separation (with $\mathbf{W}$ fixed) as in Sec. 3.5.

4.4 Quantitative Results

Visually-aided audio source separation For "in the wild" unlabeled videos, the ground-truth of separated audio sources never exists. Therefore, to allow quantitative evaluation, we create a test set consisting of combined single-source videos,

[3] `https://github.com/interactiveaudiolab/nussl`

	Instrument Pair	Animal Pair	Vehicle Pair	Cross-Domain Pair
Upper-Bound	2.05	0.35	0.60	2.79
K-means Clustering	-2.85	-3.76	-2.71	-3.32
MFCC Unsupervised [72]	0.47	-0.21	-0.05	1.49
Visual Exemplar	-2.41	-4.75	-2.21	-2.28
Unmatched Bases	-2.12	-2.46	-1.99	-1.93
Gaussian Bases	-8.74	-9.12	-7.39	-8.21
Ours	**1.83**	**0.23**	**0.49**	**2.53**

Table 1. We pairwise mix the sounds of two single source AudioSet videos and perform audio source separation. Mean Signal to Distortion Ratio (SDR in dB, higher is better) is reported to represent the overall separation performance.

following [8]. In particular, we take pairwise video combinations from AudioSet-SingleSource (cf. Sec. 4.1) and 1) compound their audio tracks by normalizing and mixing them and 2) compound their visual channels by max-pooling their respective object predictions. Each compound video is a test video; its reserved source audio tracks are the ground truth for evaluation of separation results.

To evaluate source separation quality, we use the widely used BSS-EVAL toolbox [74] and report the Signal to Distortion Ratio (SDR). We perform four sets of experiments: pairwise compound two videos of musical instruments (Instrument Pair), two of animals (Animal Pair), two of vehicles (Vehicle Pair), and two cross-domain videos (Cross-Domain Pair). For unsupervised clustering separation baselines, we evaluate both possible matchings and take the best results (to the baselines' advantage).

Table 1 shows the results. Our method significantly outperforms the Visual Exemplar, Unmatched, and Gaussian baselines, demonstrating the power of our learned bases. Compared with the unsupervised clustering baselines, including [72], our method achieves large gains. It also has the capability to match the separated source to acoustic objects in the video, whereas the baselines can only return ungrounded audio signals. We stress that both our method as well as the baselines use no audio-based supervision. In contrast, other state-of-the-art audio source separation methods supervise the separation process with labeled training data containing clean ground-truth sources and/or tailor separation to music/speech (e.g., [36, 39, 54]). Such methods are not applicable here.

Our MIML solution is fairly tolerant to imperfect visual detection. Using weak labels from the ImageNet pre-trained ResNet-152 network performs similarly to using the AudioSet ground-truth labels with about 30% of the labels corrupted. Using the true labels (Upper-Bound in Table 1) reveals the extent to which better visual models would improve results.

Visually-aided audio denoising To facilitate comparison to prior audio-visual methods (none of which report results on AudioSet), next we perform the same experiment as in [62] on visually-assisted audio denoising on AV-Bench. Following the same setup as [62], the audio signals in all videos are corrupted with white noise with the signal to noise ratio set to 0 dB. To perform audio denoising, our method retrieves bases of detected object(s) and appends the same number of randomly initialized bases as the weight matrix $\mathbf{W}$ to supervise NMF. The

	Wooden Horse	Violin Yanni	Guitar Solo	Average
Sparse CCA (Kidron et al. [47])	4.36	5.30	5.71	5.12
JIVE (Lock et al. [55])	4.54	4.43	2.64	3.87
Audio-Visual (Pu et al. [62])	8.82	5.90	**14.1**	9.61
Ours	**12.3**	**7.88**	11.4	**10.5**

Table 2. Visually-assisted audio denoising results on three benchmark videos, in terms of NSDR (in dB, higher is better).

randomly initialized bases are intended to capture the noise signal. As in [62], we report Normalized SDR (NSDR), which measures the improvement of the SDR between the mixed noisy signal and the denoised sound.

Table 2 shows the results. Note that the method of [62] is tailored to separate noise from the foreground sound by exploiting the low-rank nature of background sounds. Still, our method outperforms [62] on 2 out of the 3 videos, and performs much better than the other two prior audio-visual methods [47, 55]. Pu et al. [62] also exploit motion in manually segmented regions. On Guitar Solo, the hand's motion may strongly correlate with the sound, leading to their better performance.

4.5 Qualitative Results

Next we provide qualitative results to illustrate the effectiveness of MIML training and the success of audio source separation. Here we run our method on the real multi-source videos from AudioSet. They lack ground truth, but results can be manually inspected for quality (see our video[4]).

Fig. 5 shows example unlabeled videos and their discovered audio basis associations. For each example, we show sample video frames, ImageNet CNN visual object predictions, as well as the corresponding audio basis-object relation map predicted by our MIML network. We also report the AudioSet audio ground truth labels, but note that they are never seen by our method. The first example (Fig. 5-a) has both piano and violin in the visual frames, which are correctly detected by the CNN. The audio also contains the sounds of both instruments, and our method appropriately activates bases for both the violin and piano. Fig. 5-b shows a man playing the violin in the visual frames, but both piano and violin are strongly activated. Listening to the audio, we can hear that an out-of-view player is indeed playing the piano. This example accentuates the advantage of learning object sounds from thousands of unlabeled videos; our method has learned the correct audio bases for piano, and "hears" it even though it is off-camera in this test video. Fig. 5-c/d show two examples with inaccurate visual predictions, and our model correctly activates the label of the object in the audio. Fig. 5-e/f show two more examples of an animal and a vehicle, and the results are similar. These examples suggest that our MIML network has successfully learned the prototypical spectral patterns of different sounds, and is capable of associating audio bases with object categories.

Please see our **video**[4] for more results, where we use our system to detect and separate object sounds for novel "in the wild" videos.

[4] `http://vision.cs.utexas.edu/projects/separating_object_sounds/`

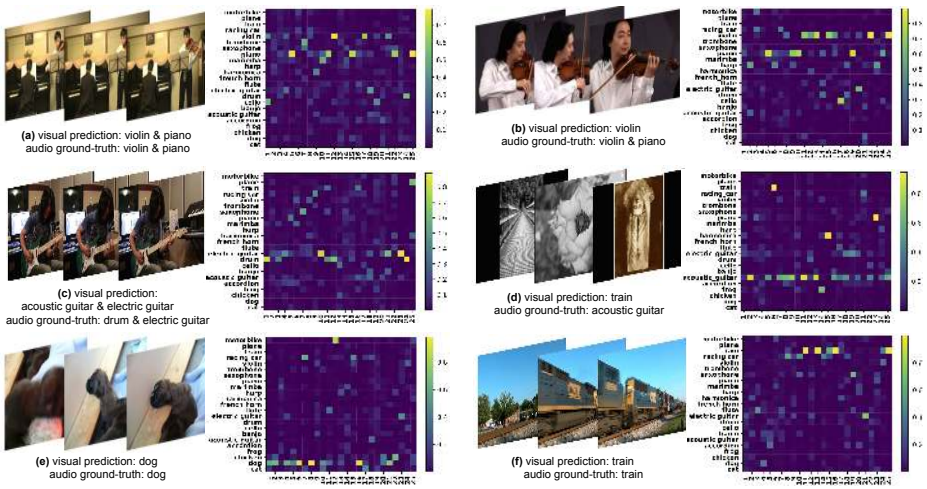

Fig. 5. In each example, we show the video frames, visual predictions, and the corresponding basis-label relation maps predicted by our MIML network. Please see our video[4] for more examples and the corresponding audio tracks.

Overall, the results are promising and constitute a noticeable step towards visually guided audio source separation for more realistic videos. Of course, our system is far from perfect. The most common failure modes by our method are when the audio characteristics of detected objects are too similar or objects are incorrectly detected (see Supp.). Though ImageNet-trained CNNs can recognize a wide array of objects, we are nonetheless constrained by its breadth. Furthermore, not all objects make sounds and not all sounds are within the camera's view. Our results above suggest that learning can be robust to such factors, yet it will be important future work to explicitly model them.

5 Conclusion

We presented a framework to learn object sounds from thousands of unlabeled videos. Our deep multi-instance multi-label network automatically links audio bases to object categories. Using the disentangled bases to supervise non-negative matrix factorization, our approach successfully separates object-level sounds. We demonstrate its effectiveness on diverse data and object categories. Audio source separation will continue to benefit many appealing applications, e.g., audio events indexing/remixing, audio denoising for closed captioning, or instrument equalization. In future work, we aim to explore ways to leverage scenes and ambient sounds, as well as integrate localized object detections and motion.

Acknowledgements: This research was supported in part by an IBM Faculty Award, IBM Open Collaboration Research Award, and DARPA Lifelong Learning Machines. We thank members of the UT Austin vision group and Wenguang Mao, Yuzhong Wu, Dongguang You, Xingyi Zhou and Xinying Hao for helpful input. We also gratefully acknowledge a GPU donation from Facebook.

References

1. Afouras, T., Chung, J.S., Zisserman, A.: The conversation: Deep audio-visual speech enhancement. arXiv preprint arXiv:1804.04121 (2018) 5
2. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and multiple instance learning. PAMI (2010) 5
3. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: ICCV (2017) 3, 10
4. Arandjelović, R., Zisserman, A.: Objects that sound. arXiv preprint arXiv:1712.06651 (2017) 3, 10
5. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: NIPS (2016) 3
6. Aytar, Y., Vondrick, C., Torralba, A.: See, hear, and read: Deep aligned representations. arXiv preprint arXiv:1706.00932 (2017) 4
7. Barnard, K., Duygulu, P., de Freitas, N., Blei, D., Jordan, M.: Matching words and pictures. JMLR (2003) 5
8. Barzelay, Z., Schechner, Y.Y.: Harmony in motion. In: CVPR (2007) 3, 4, 12
9. Berg, T., Berg, A., Edwards, J., Maire, M., White, R., Teh, Y., Learned-Miller, E., Forsyth, D.: Names and faces in the news. In: CVPR (2004) 5
10. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: CVPR (2016) 5
11. Bryan, N.: Interactive Sound Source Separation. Ph.D. thesis, Stanford University (2014) 4
12. Casanovas, A.L., Monaci, G., Vandergheynst, P., Gribonval, R.: Blind audiovisual source separation based on sparse redundant representations. IEEE Transactions on Multimedia (2010) 3, 4
13. Chen, L., Srivastava, S., Duan, Z., Xu, C.: Deep cross-modal audio-visual generation. In: on Thematic Workshops of ACM Multimedia (2017) 5
14. Cinbis, R., Verbeek, J., Schmid, C.: Weakly supervised object localization with multi-fold multiple instance learning. PAMI (2017) 5
15. Darrell, T., Fisher, J., Viola, P., Freeman, W.: Audio-visual segmentation and the cocktail party effect. In: ICMI (2000) 3, 4
16. Deselaers, T., Alexe, B., Ferrari, V.: Weakly supervised localization and learning with generic knowledge. IJCV (2012) 5
17. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence (1997) 7
18. Donahue, J., Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR (2015) 5
19. Duong, N.Q., Ozerov, A., Chevallier, L., Sirot, J.: An interactive audio source separation framework based on non-negative matrix factorization. In: ICASSP (2014) 4
20. Duong, N.Q., Vincent, E., Gribonval, R.: Under-determined reverberant audio source separation using a full-rank spatial covariance model. IEEE Transactions on Audio, Speech, and Language Processing (2010) 4
21. Duygulu, P., Barnard, K., de Freitas, N., Forsyth, D.: Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: eccv (2002) 5
22. Ellis, D.P.W.: Prediction-driven computational auditory scene analysis. Ph.D. thesis, Massachusetts Institute of Technology (1996) 4

23. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. arXiv preprint arXiv:1804.03619 (2018) 5
24. Feng, J., Zhou, Z.H.: Deep miml network. In: AAAI (2017) 7, 8
25. Févotte, C., Bertin, N., Durrieu, J.L.: Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. Neural computation (2009) 2, 4, 6
26. Févotte, C., Idier, J.: Algorithms for nonnegative matrix factorization with the β-divergence. Neural computation (2011) 4, 6, 10
27. Fisher III, J.W., Darrell, T., Freeman, W.T., Viola, P.A.: Learning joint statistical models for audio-visual fusion and segregation. In: NIPS (2001) 3, 4
28. Gabbay, A., Shamir, A., Peleg, S.: Visual speech enhancement using noise-invariant training. arXiv preprint arXiv:1711.08789 (2017) 5
29. 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) 10
30. Griffin, D., Lim, J.: Signal estimation from modified short-time fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing (1984) 5, 6
31. Guo, X., Uhlich, S., Mitsufuji, Y.: Nmf-based blind source separation using a linear predictive coding error clustering criterion. In: ICASSP (2015) 4, 11
32. Harwath, D., Glass, J.: Learning word-like units from joint audio-visual analysis. In: ACL (2017) 5
33. Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., Glass, J.: Jointly discovering visual objects and spoken words from raw sensory input. arXiv preprint arXiv:1804.01452 (2018) 5
34. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 7
35. Hennequin, R., David, B., Badeau, R.: Score informed audio source separation using a parametric model of non-negative spectrogram. In: ICASSP (2011) 4
36. Hershey, J.R., Chen, Z., Le Roux, J., Watanabe, S.: Deep clustering: Discriminative embeddings for segmentation and separation. In: ICASSP (2016) 4, 12
37. Hershey, J.R., Movellan, J.R.: Audio vision: Using audio-visual synchrony to locate sounds. In: NIPS (2000) 3
38. Hofmann, T.: Probabilistic latent semantic indexing. In: International ACM SIGIR Conference on Research and Development in Information Retrieval (1999) 4
39. Huang, P.S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P.: Deep learning for monaural speech separation. In: ICASSP (2014) 4, 12
40. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural networks (2000) 2, 4
41. Innami, S., Kasai, H.: Nmf-based environmental sound source separation using time-variant gain features. Computers & Mathematics with Applications (2012) 4, 11
42. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015) 7
43. Izadinia, H., Saleemi, I., Shah, M.: Multimodal analysis for identification and segmentation of moving-sounding objects. IEEE Transactions on Multimedia (2013) 10
44. Jaiswal, R., FitzGerald, D., Barry, D., Coyle, E., Rickard, S.: Clustering nmf basis functions using shifted nmf for monaural sound source separation. In: ICASSP (2011) 4, 11

45. Jhuo, I.H., Ye, G., Gao, S., Liu, D., Jiang, Y.G., Lee, D., Chang, S.F.: Discovering joint audio–visual codewords for video event detection. Machine vision and applications (2014) 2
46. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015) 5
47. Kidron, E., Schechner, Y.Y., Elad, M.: Pixels that sound. In: CVPR (2005) 3, 11, 13
48. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015) 11
49. Korbar, B., Tran, D., Torresani, L.: Co-training of audio and video representations from self-supervised temporal synchronization. arXiv preprint arXiv:1807.00230 (2018) 3
50. Le Magoarou, L., Ozerov, A., Duong, N.Q.: Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization. Journal of Signal Processing Systems (2015) 4
51. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems (2001) 4, 6
52. Li, B., Dinesh, K., Duan, Z., Sharma, G.: See and listen: Score-informed association of sound tracks to players in chamber music performance videos. In: ICASSP (2017) 3, 4
53. Li, K., Ye, J., Hua, K.A.: What's making that sound? In: ACMMM (2014) 10
54. Liutkus, A., Fitzgerald, D., Rafii, Z., Pardo, B., Daudet, L.: Kernel additive models for source separation. IEEE Transactions on Signal Processing (2014) 12
55. Lock, E.F., Hoadley, K.A., Marron, J.S., Nobel, A.B.: Joint and individual variation explained (jive) for integrated analysis of multiple data types. The annals of applied statistics (2013) 11, 13
56. Nakadai, K., Hidai, K.i., Okuno, H.G., Kitano, H.: Real-time speaker localization and speech separation by audio-visual integration. In: IEEE International Conference on Robotics and Automation (2002) 4
57. Naphade, M., Smith, J.R., Tesic, J., Chang, S.F., Hsu, W., Kennedy, L., Hauptmann, A., Curtis, J.: Large-scale concept ontology for multimedia. IEEE multimedia (2006) 2
58. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. arXiv preprint arXiv:1804.03641 (2018) 5
59. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: CVPR (2016) 3, 5
60. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: ECCV (2016) 3
61. Parekh, S., Essid, S., Ozerov, A., Duong, N.Q., Pérez, P., Richard, G.: Motion informed audio source separation. In: ICASSP (2017) 3, 4
62. Pu, J., Panagakis, Y., Petridis, S., Pantic, M.: Audio-visual object localization and separation using low-rank and sparsity. In: ICASSP (2017) 3, 4, 10, 11, 12, 13
63. Rahne, T., Böckmann, M., von Specht, H., Sussman, E.S.: Visual cues can modulate integration and segregation of objects in auditory scene analysis. Brain research (2007) 2
64. Rivet, B., Girin, L., Jutten, C.: Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures. IEEE transactions on audio, speech, and language processing (2007) 3, 4
65. Sedighin, F., Babaie-Zadeh, M., Rivet, B., Jutten, C.: Two multimodal approaches for single microphone source separation. In: 24th European Signal Processing Conference (2016) 4

66. Simpson, A.J., Roma, G., Plumbley, M.D.: Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In: International Conference on Latent Variable Analysis and Signal Separation. pp. 429–436. Springer (2015) 4
67. Smaragdis, P., Casey, M.: Audio/visual independent components. In: International Conference on Independent Component Analysis and Signal Separation (2003) 4
68. Smaragdis, P., Raj, B., Shashanka, M.: A probabilistic latent variable model for acoustic modeling. In: NIPS (2006) 4
69. Smaragdis, P., Raj, B., Shashanka, M.: Supervised and semi-supervised separation of sounds from single-channel mixtures. In: International Conference on Independent Component Analysis and Signal Separation (2007) 4
70. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: Proceedings of the 8th ACM international workshop on Multimedia information retrieval (2006) 2
71. Snoek, C.G., Worring, M.: Multimodal video indexing: A review of the state-of-the-art. Multimedia tools and applications (2005) 2
72. SPIERTZ, M.: Source-filter based clustering for monaural blind source separation. In: 12th International Conference on Digital Audio Effects (2009) 4, 11, 12
73. Vijayanarasimhan, S., Grauman, K.: Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization. In: CVPR (2008) 5
74. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE transactions on audio, speech, and language processing (2006) 12
75. Virtanen, T.: Sound source separation using sparse coding with temporal continuity objective. In: International Computer Music Conference (2003) 4, 11
76. Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE transactions on audio, speech, and language processing (2007) 2, 4, 6
77. WANG, B.: Investigating single-channel audio source separation methods based on non-negative matrix factorization. In: ICA Research Network International Workshop (2006) 4
78. Wang, L., Xiong, Y., Lin, D., Gool, L.V.: Untrimmednets for weakly supervised action recognition and detection. In: CVPR (2017) 5
79. Wang, Z., Kuan, K., Ravaut, M., Manek, G., Song, S., Yuan, F., Seokhwan, K., Chen, N., Enriquez, L.F.D., Tuan, L.A., et al.: Truly multi-modal youtube-8m video classification with video, audio, and text. arXiv preprint arXiv:1706.05461 (2017) 2
80. Wu, J., Yu, Y., Huang, C., Yu, K.: Deep multiple instance learning for image classification and auto-annotation. In: CVPR (2015) 8
81. Yang, H., Zhou, J.T., Cai, J., Ong, Y.S.: Miml-fcn+: Multi-instance multi-label learning via fully convolutional networks with privileged information. In: CVPR (2017) 8
82. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Transactions on signal processing (2004) 4
83. Zhang, Z., Wu, J., Li, Q., Huang, Z., Traer, J., McDermott, J.H., Tenenbaum, J.B., Freeman, W.T.: Generative modeling of audible shapes for object perception. In: ICCV (2017) 5
84. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. arXiv preprint arXiv:1804.03160 (2018) 5

85. Zhou, Y., Wang, Z., Fang, C., Bui, T., Berg, T.L.: Visual to sound: Generating natural sound for videos in the wild. arXiv preprint arXiv:1712.01393 (2017) 5, 10
86. Zibulevsky, M., Pearlmutter, B.A.: Blind source separation by sparse decomposition in a signal dictionary. Neural computation (2001) 2, 4