Mix and Localize: Localizing Sound Sources in Mixtures

Xixi Hu[1,2*] Ziyang Chen[1*] Andrew Owens[1]

University of Michigan[1] The University of Texas at Austin[2]

Abstract

We present a method for simultaneously localizing multiple sound sources within a visual scene. This task requires a model to both group a sound mixture into individual sources, and to associate them with a visual signal. Our method jointly solves both tasks at once, using a formulation inspired by the contrastive random walk of Jabri et al. We create a graph in which images and separated sounds correspond to nodes, and train a random walker to transition between nodes from different modalities with high return probability. The transition probabilities for this walk are determined by an audio-visual similarity metric that is learned by our model. We show through experiments with musical instruments and human speech that our model can successfully localize multiple sounds, outperforming other self-supervised methods.

1. Introduction

Humans have the remarkable ability to localize many sounds at once [20]. Existing audio-visual sound localization methods, by contrast, are generally trained with the assumption that only a single sound source is present at a time, and largely lack mechanisms for grouping the contents of a scene into multiple audio-visual events.

This problem is often addressed through contrastive learning [2–4, 32, 39]. These methods generally produce a single embedding for the audio, representing the sound source, and an embedding for each image patch, representing the sound's possible locations. They then learn cross-modal correspondences, such that image patches and sounds that co-occur within the same scene are brought close together, and pairings that do not co-occur are pulled apart. Extending this approach to multiple sound sources seemingly requires solving two different problems: separating the sources from a sound mixture, and localizing them within an image.

We propose a simple model that jointly addresses both of these problems. Our model uses cycle consistency to group a scene into sound sources, inspired by the contrastive random walk [25]. It produces multiple embedding vectors for a sound mixture, each representing a different sound source, and an audio-visual similarity metric that associates them with their corresponding image content. This similarity metric defines the transition probabilities for a random walk on a graph whose nodes correspond to images and predicted sound sources. Our model performs a random walk that transitions from the audio to images, and then back. We learn a similarity metric that maximizes the probability of *cycle-consistency* (*i.e*., return probability) for the walk. After training, we create an attention map between an image and each sound source by estimating the similarity score between the audio and visual embeddings.

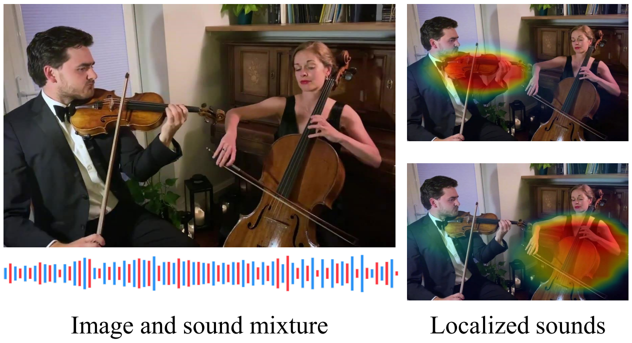


Figure 1. **Cycle-consistent multi-source localization.** Our model jointly learns to separate a sound mixture into sound sources, and to localize these sources within an image. To do this, we use a self-supervised grouping method based on cycle-consistent random walks. We show sound source localization results from our model.

Obtaining a cycle-consistent walk requires extracting sound sources from the mixture, and associating each one with distinct image content. Consequently, this formulation has several advantages over other self-supervised audio-visual localization methods. It separates sounds from a mixture, and explicitly groups the scene into discrete sound-making objects. The model is also simple, and can be implemented as a straightforward extension of previous contrastive methods.

We evaluate our model on synthetic and real-world sound mixtures containing musical instruments [10, 53] and human speakers [13]. We find that, in comparison to other self-

[*] Indicates equal contribution.

supervised localization methods, our model is more accurate in localizing sounds in multi-source mixtures.

2. Related Work

Sound source separation. There has been a long history of methods for separating monaural sound mixtures. Early work addressed this problem with probabilistic models [15, 38] and recent work has tackled it via deep neural networks [22,51]. These often use a "mix-and-separate" [53] training procedure [26, 41]. Other work combines source separation with visual cues. Zhao *et al*. [53] proposed to separate different musical instruments by associating separated audio sources with pixels in the video, and later used optical flow [52] to provide motion cues. Gao *et al*. [18] jointly solve audio-visual speech separation with a multi-task learning framework by incorporating cross-modal face-voice attributes and lip motion. Tian *et al*. [42] jointly learn sound separation and sounding object visual grounding, using an approach they term cyclic co-learning. Chatterjee *et al*. [8] model visual signals into scene graphs and learn to separate sounds by co-segmenting subgraphs and associated audio. Majumder *et al*. [29] introduce the active audio-visual source separation task that an agent learns movement policies to improve sound separation qualities. Like these works, we jointly localize and separate sound. However, we do not generate separated audio: we obtain embeddings that represent the separated sound sources using contrastive learning.

Audio-visual sound source localization. The co-occurrence of audio and visual cues in videos has been leveraged for sound source localization [19, 23, 35, 39, 43]. Researchers exploit audio-visual correspondence by matching audio and visual signals from the same video. Arandjelović and Zisserman [3, 4] measure the similarity between learned image region and audio representations and use multi-instance learning to localize sound sources. Owens and Efros [32] use class activation maps [54] to visualize the area contributing to solving audio-visual synchronization tasks. Chen et al. [9] mine hard negative image locations in cross-modal contrastive learning to obtain better sound localization results. Hu *et al*. [24] extend [4] and use clustering to generate pseudo-class labels, achieving class-aware sound source localization with mixed sounds. While we have a similar goal, our approach is entirely unsupervised, and does not require semi-supervised learning with labels, either at training or test time. Our work is motivated by them and aims to localize different sound sources in multi-source sound mixtures.

Audio-visual self-supervision. Apart from sound source localization and separation, many recent works have proposed to use paired audio-visual data for representation learning and other tasks as well. Owens *et al*. [33] learned visual representations for materials from impact sounds. Other work has learned features, scene structure, and geometric properties from sounds [11, 16, 34], or learns multisensory representations for both audio and vision [28, 32, 49]. Asano *et al*. [5] proposed self-supervised clustering and representation learning approach for providing labels to multimodal data. Other work has learned active speaker detection [2, 14], up-mixing the mono audio [17, 36, 50], cross-modal distillation [6, 44]. We take inspiration from them and learn the representation of mixture sounds.

Graph-based representation learning. A number of recent works use graphs to learn image and video segmentation [40] and space-time correspondence [25, 45, 46]. Jabri *et al*. [25] propose to use graphs to learn the visual correspondence between several frames clipped from video, and the graph is constructed by connecting patches in spatio-temporal neighborhoods. Bian *et al*. [7] propose multi-scale contrastive random walks to obtain pixel-level correspondence between frames. We extend this approach to a multi-modal learning domain, rather than to video representation learning.

3. Method

Our goal is to perform *multi-source* audio-visual sound localization. Given audio $\mathbf{a}$ and corresponding image $\mathbf{v}$, we will parse the scene into discrete sound sources and localize them within an image. We frame this as a representation learning problem. We produce embedding vectors $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_k$ from $\mathbf{a}$, representing the k visible sound sources, and associate them with visual embeddings for image regions $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m$.

If we were given ground-truth correspondences between sources $\mathbf{s}_i$ and image regions $\mathbf{x}_i$, then contrastive localization methods [3, 4, 39] could straightforwardly be applied to this problem. However, the sound sources are *latent* and must be estimated from the audio. We propose to jointly solve both problems: we produce audio embeddings from a mixture that provide cycle consistent matches with the image regions.

Single-source localization as a random walk. As a preliminary step toward solving this problem, we start by considering the simpler single-source localization problem. As in previous work [2,4,39], we can address this problem using contrastive learning. We learn an embedding $f(\mathbf{x}) \in \mathbb{R}^C$ for each image region and another embedding $g(\mathbf{a})$ for audio. In practice, we compute the image embeddings using a fully convolutional network that operates on full images. We define a cross-modal similarity metric:

$$\phi(\mathbf{v}, \mathbf{a}) = \max_{\mathbf{x}_i} f(\mathbf{x}_i)^\top g(\mathbf{a}), \tag{1}$$

where the pooling is performed over all image regions $\mathbf{x}_i$ in $\mathbf{v}$. Following [4], we summarize the similarity of the

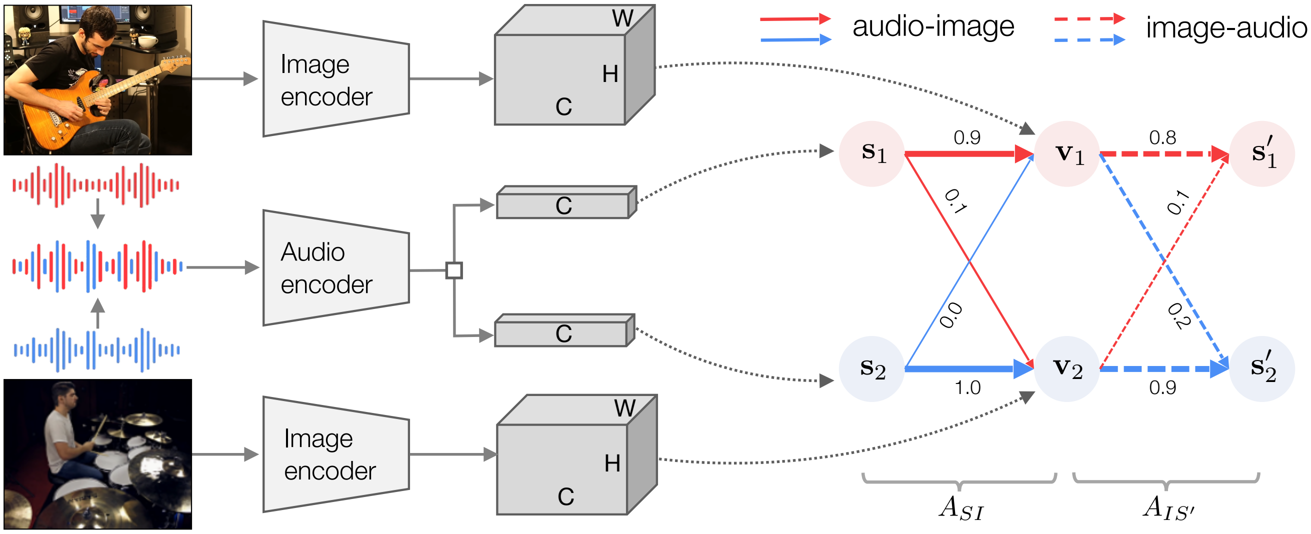


Figure 2. **Audio-visual random walks.** We learn representations for separating and localizing sounds. We generate a synthetic mixture by summing waveforms from multiple videos (we show $k = 2$ videos). Our model estimates embedding vectors from the audio mixture, representing each sound source, and learns an audio-visual similarity metric that associates image regions with the extracted sources. We solve a cycle-consistency problem in a graph. Edges connect each audio node to nodes that represent each image. A random walker is trained to walk from each audio node to an image node, then back to the audio. Our model learns to guide the random walker to the node it began its walk (*i.e*. to maximize its cycle consistency) using transition probabilities derived from the similarity function.

whole image by taking the max over all the image regions, under the assumption that the sound source occupies a small portion of the image. We can learn audio and visual representations using contrastive learning: $\mathbf{v}_i$ should be more similar to its corresponding audio track $\mathbf{a}_i$ than to $n-1$ other audio examples. If $A_{IS}(i,j)$ is the similarity between $\mathbf{v}_i$ and $\mathbf{a}_j$, these similarities can be formulated as:

$$A_{IS}(i,j) = \frac{\exp(\phi(\mathbf{v}_i, \mathbf{a}_j)/\tau)}{\sum_{k=1}^{n} \exp(\phi(\mathbf{v}_i, \mathbf{a}_k)/\tau)}, \tag{2}$$

where τ is a temperature hyperparameter [48]. The summation in the denominator iterates over both $\mathbf{a}_j$ and the $n-1$ other audio examples, and $A_{IS} \in \mathbb{R}^{n \times n}$. We maximize the diagonal of A_{IS} using the InfoNCE loss [31]:

$$\mathcal{L}_{\texttt{corresp}} = -\frac{1}{n}\,\mathrm{tr}(\log(A_{IS})), \tag{3}$$

where the log is performed elementwise. After training, the dot product between the image and audio embeddings $f(\mathbf{x}_i)^\top g(\mathbf{a})$ can be interpreted as the likelihood of $\mathbf{x}_i$ being the location of a sound source, since this conceptually represents the correlation between the visual and audio signals.

One can interpret A_{IS} as the transition matrix of a random walk that moves from images to sounds, and $\mathcal{L}_{\texttt{corresp}}$ as a penalty for transitioning to an incorrect sound. One could also compute an analogous matrix A_{SI} by matching from audio $\mathbf{a}_i$ to the image $\mathbf{v}_j$ with softmax normalization (equivalent to normalizing columns in Eq. 2, rather than rows, and then transposing).

Random walk with cycle-consistency. Now, suppose that we do not know the ground-truth correspondence between images and audio, but merely that there is an unknown, one-to-one relationship between the separated audio embeddings and images. We use a *cycle consistent* random walk to jointly learn the audio embeddings and associate them with images.

We are given a synthetic sound mixture containing k components, created by summing k different waveforms, along with the corresponding k images they were taken from (we use $k = 2$ in our experiments). We construct a directed graph containing nodes for each sound source and each image. Its edges lead from sound sources to images and back (Figure 2), with transition probabilities determined by audio-visual similarity. A random walker in this graph starts from an audio node $\mathbf{s}_i$, travels to an image node $\mathbf{v}_j$ and arrives at another audio node $\mathbf{s}_t$.

Inspired by recent works on visual correspondence [25, 46], we use a *cycle-consistency* loss to guide the random walker. While we do not know whether a given $\mathbf{s}_i$ and $\mathbf{v}_j$ pair belong to the same audio event, we do know that there should be a one-to-one relationship between images and sounds. As in Eq. 2, we compute a matrix $A_{IS} \in \mathbb{R}^{k \times k}$ such that $A_{IS}(i,j)$ measures the similarity between image embedding $\mathbf{v}_i$ and audio embedding $\mathbf{v}_j$. We encourage audio to return to itself with high probability in the random walk:

$$\mathcal{L}_{\texttt{cyc}} = -\frac{1}{k}\,\mathrm{tr}(\log(A_{SI} A_{IS})). \tag{4}$$

Under this loss, the model is encouraged to maximize the

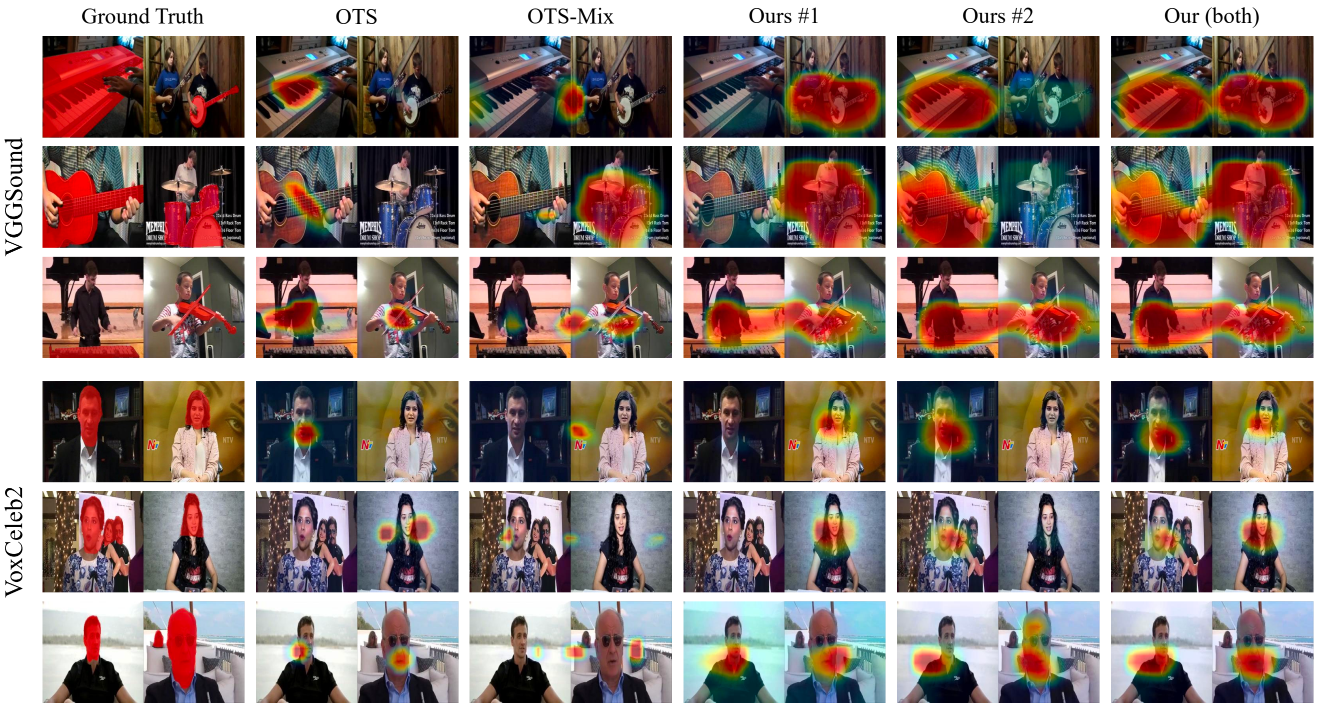


Figure 3. Multi-source localization results on synthetic mixtures from the VGGSound-Instruments and VoxCeleb2 datasets. We provide a comparison of localization maps generated by different methods. We show the two localization maps generated by our model's two embeddings. The color of the image region indicates its localization score, with the red regions having higher scores. We show failure cases in the last row of each dataset.

probability of associating a sound to highly discriminative image regions—those that can successfully select this sound over all others. Since the embeddings are produced from a sound mixture, a natural choice is to represent its sound sources. After training, the dot products $f(\mathbf{x}_i)^\top g(\mathbf{s}_j)$ convey the location of each of the k sound sources.

Data augmentation. We found that our models could often quickly drop the cycle consistency loss (Eq. 4) to low values, since high-dimensional embedding vectors are often cycle-consistent by chance. We encourage the model to learn other useful invariances by using randomly-shifted versions of the audio when computing the transition matrix A_{IS}, resulting in a matrix we call $A_{IS'}$ (Fig. 2).

4. Experiments

We evaluate our model on single- and multi-source sound localization for scenes containing musical instruments and human speech.

4.1. Implementation

Image encoder. We use ResNet-18 [21] as the backbone for the image encoder. During the training, each frame is randomly cropped and resized to 224 × 224. During inference, we directly resize the images without cropping. We encode the image such that the feature map will be downsampled to $\frac{W}{16} \times \frac{H}{16} \times C$ dimensional embedding vectors. We l_2 normalize them along the channel axis, following [48]. During testing, for the synthetic VGGSound and VoxCeleb2 datasets, we concatenate two images so that input of the image encoder is 448×224 and the output score map is 28×14. This will keep the aspect ratio of images similar during training and testing. For all other experiments, we use 224×224 images and 14×14 score maps. We apply bilinear interpolation to upsample score maps for all the methods.

Audio encoder. We use a ResNet-18 network [21] to extract k different l_2-normalized C-dimensional feature vectors from a 0.96s of sound, using a log-mel spectrogram input representation. We compute different embedding vectors for audio nodes by applying different fully connected layers to the final pooled convolutional features. We use dot product between image and audio features to calculate the similarity score between image regions and audio nodes.

Hyperparameters. During training, we use the Adam optimizer [27] with a learning rate of 10^{-4} on MUSIC and VGGSound dataset, and a learning rate of 10^{-5} on VoxCeleb dataset. We use a batch size of 128 and set the temperature $\tau = 0.07$ following [48]. We set the feature dimension $C = 128$. When processing the audio, the sounds are resam-

		Single sound source			**Multiple sound sources**							
		MUSIC-Solo			MUSIC-Synthetic				MUSIC-Duet			
		AP	AUC	IoU@0.5	CAP	PIAP	AUC	CIoU@0.3	CAP	PIAP	AUC	CIoU@0.3
Semi-supervised	Hu *et al.* [24]	-	43.6	51.4	-	-	23.5	32.3	-	-	22.1	30.2
	Sound of Pixels + Matching [24, 53]	-	43.3	40.5	-	-	11.8	8.1	-	-	16.8	16.8
Self-supervised	OTS [4]	**69.3**	35.8	26.1	11.4	17.6	10.2	3.7	35.4	42.8	18.3	13.2
	OTS-mix [4]	53.8	33.9	17.5	16.9	27.4	7.3	0	23.8	28.2	12.0	2.0
	Attention [39]	-	38.7	**37.2**	-	-	12.3	6.4	-	-	19.4	21.5
	DMC [23]	-	38.0	29.1	-	-	16.3	7.0	-	-	21.1	17.3
	Ours	67.9	**40.6**	29.2	**34.0***	**39.7**	**20.2***	**21.3***	**47.4***	**53.9**	**21.2***	**26.3***

Table 1. Sound source localization performance on MUSIC dataset. Following previous work [24, 39], IoU@0.5 measures ratio of successful samples at 0.5 threshold. Similarly, CIoU@0.3 measures the ratio of successful samples at 0.3 threshold. * indicates that the method might benefit from taking the best matching pairs (see Section 4.4).

pled to 16kHz. The 0.96s audio clips are converted to mel spectrograms of size 193 × 64 via the Short-Time Fourier Transform (STFT) using 64 mel filter banks, a window size of 160, and a hop length of 80.

4.2. Dataset

MUSIC. The MUSIC dataset [53] contains 11 instruments, in both solos and duets. We use the same training/testing splits as in Hu *et al.* [24]. The MUSIC-Synthetic dataset [24] contains concatenated images with frames from four videos, and the audio is synthesized such that there are two instruments making sounds while the other two are silent. The MUSIC-Duet dataset is a subset of MUSIC that contains duets videos with two instruments playing sounds. We use the solo videos, MUSIC-Solo, to evaluate the performance of single sound source localization, and use MUSIC-Duet and MUSIC-Synthetic datasets when evaluating multiple sound source localization, using the same annotations as [24].

VGGSound-Instruments. We also evaluate on VGGSound [10]. Each video in VGGSound only has a single category label. Analogous to [4], we filtered and sampled 37 video classes of musical instruments with 32k video clips of 10s length, and we call this subset VGGSound-Instruments. The category list is provided in the supplementary material. For evaluation, we filtered and annotated segmentation masks for 446 high-quality video frames[1]. When evaluating multi-source localization, we randomly concatenate two frames, resulting in 448 × 224 input images, and obtain sound mixtures by summing their waveforms.

Human speech. VoxCeleb2 dataset [13] is a large-scale audio-visual speaker recognition dataset containing over 1.5k hours of video for 6,112 celebrities. For evaluation, we use a face detector to annotate face segmentation masks for 1k random samples in the test set. We follow the same strategy as in VGGSound-Instruments to create a synthetic multi-speaker synthetic evaluation set.

4.3. Evaluated methods

We compare our model with several other audio-visual learning methods. When re-implementing, we use ResNet-18 as our backbone architecture to ensure fair comparisons.

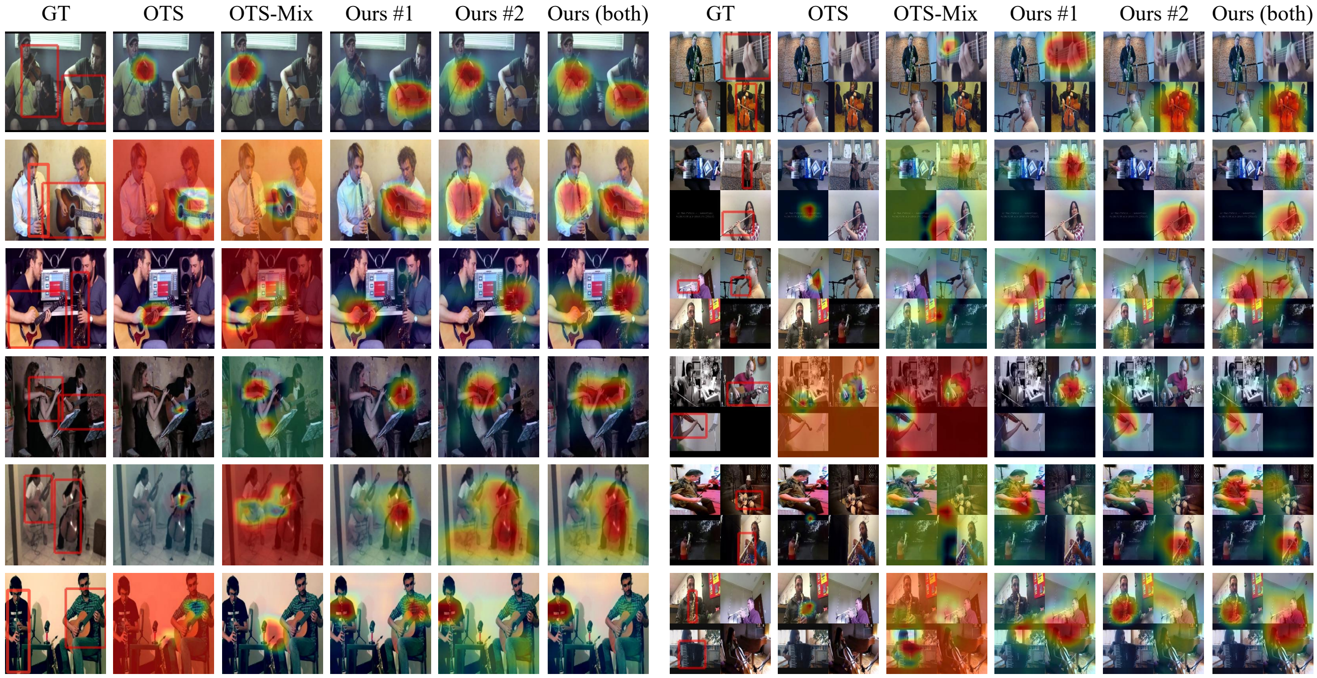
Self-supervised methods. We compare our model with several variants of the model from Arandjelović and Zisserman [4], which we call *OTS*. We follow the model architecture of [24] to implement the methods, which uses ResNet-18 to extract the features and global max pooling layer for the fused attention map. We keep the data prepossessing and network architectures the same as our method when re-implementing them. We also created a variation of OTS that is trained on synthetic mixtures and concatenated video frames, following [24]. We call this method *OTS-mix*. In contrast to approaches that use multiple frames from videos [1, 2, 37, 53] or require extra manual labels [24, 35], these two methods only use one frame and are trained in a fully self-supervised manner.

Semi-supervised methods. We also consider the semi-supervised, multi-source methods proposed for musical instrument localization in Hu *et al.* [24]. At training time, these methods cluster the features of training data and match the clusters with ground-truth labels. The sounding and silent visual areas can be obtained by retrieving similarity maps corresponding to the clusters. Since this method uses ground-truth labels to match the clustering result with the classes, we consider it to be semi-supervised. Additionally, we compare with a variant of Sound of Pixels [53] from [24], in which the model predicts 11 different score maps, and uses the training set labels to match the predictions with different classes. We call this method *Sound of Pixels + Matching*.

4.4. Evaluation of sound source localization

We evaluate these methods on both single- and multi-source sound localization. Unlike methods that produce only one localization map for all the sounding objects in the scene,

[1] These annotations can be found at https://web.eecs.umich.edu/~ahowens/mix-localize/



(a) MUSIC-Duet

(b) MUSIC-Synthetic

Figure 4. Multi-source sound localization on the MUSIC dataset [52, 53]. We show failure cases in the last two rows.

		Single source			Multiple sources			
		AP	AUC	IoU@0.3	CAP	PIAP	AUC	CIoU@0.1
VGGSound	OTS [4]	**47.3**	24.5	25.7	**23.2**	**37.6**	10.8	51.1
	OTS-mix [4]	37.0	20.9	24.9	18.1	30.7	10.8	50.7
	Ours	44.7	**32.1**	**49.6**	21.5*	37.4	**15.5***	**73.1***
VoxCeleb2	OTS [4]	43.9	23.5	6.2	**20.4**	32.6	7.0	15.8
	OTS-mix [4]	21.4	6.4	6.2	10.7	18.2	4.1	15.8
	Ours	**46.1**	**27.7**	**35.4**	20.1*	**35.4**	**14.2***	**17.4***

Table 2. Sound source localization performance on VGGSound-Instruments and VoxCeleb2 datasets.

our method localizes multiple objects (by producing k localization maps). We therefore expect our method to perform approximately as well when compared with other methods on single sound-source localization, and to outperform these methods on multiple sound localization.

Evaluation of single sound source localization. We evaluate the single sound source localization performance following [24, 39]. Given the ground truth bounding boxes or object segmentation mask, we compute the *Intersection over Union* (IoU) and *Area Under Curve* (AUC) according to the predicted sounding area. For methods such as Objects that Sound [4] (OTS) that produce one output, we take the produced single sounding area to compute the score. For our method, since there are 2 sounding area maps (corresponding to 2 audio nodes), we take the average of these maps as the final sounding area. When computing the scores, we use a threshold of 0.4 for our method in all experiments. For other methods, we choose the best threshold for each method according to the performance on validation set. Additionally, to avoid judging the methods based on a fixed threshold, we also use pixel-wise average precision (**AP**) [12].

Evaluation of multiple sound source localization. We follow [24] and use Class-aware IoU (CIoU) when evaluating multiple sound source localization. Analogously, we propose to use class-aware average precision (CAP) to provide a threshold-less evaluation metric. The CAP score is calculated as:

$$CAP = \frac{\sum_{k=1}^{K} \delta_k AP_k}{\sum_{k=1}^{K} \delta_k}, \tag{5}$$

where AP_k measures the pixel-wise average precision for the class k. The indicator δ_k indicates whether an object of class k is making sound. Since our method does not have class labels (*i.e.*, we do not know which localization map corresponds to which object), we use a modified version of CAP for our method, where we evaluate both pairings of the predicted and ground-truth labels, and report the best. Since this provides a potential unfair advantage to our method, we also introduce another metric called permutation-invariant average precision (PIAP). When computing this score, we take the average of the sounding area maps and compute the average precision using the ground truths of all sounding objects.

4.4.1 Single-source localization

We evaluate the performance of single sound source localization on MUSIC-Solo, VGGSound-Instruments and VoxCeleb2. Under this evaluation setup, the input audio to the models is original unmixed audio instead of mixture sounds. The results are shown in Table 1 and Table 2. It can be seen that our method performs approximately equally well on single sound source localization when compared with other methods. This suggests that our method is capable of localizing a single sound source. We find that when the input audio is from a single source (i.e., unmixed), the model tends to predict two similar localization maps. We note that these single-source sounds were not explicitly provided during training.

4.4.2 Multi-source localization

Quantitative results. We evaluate the multiple sound source localization performance on MUSIC-Duet, MUSIC-Synthetic, VGGSound-Instruments and VoxCeleb2 datasets in Table 1 and Table 2. The comparison with other work shows that our proposed method achieves better performance on the multiple sound source localization task. We note that our method does not use labels (unlike [24]), or multiple frames (unlike [53]). Instead of taking synthetic data as input as [24], we use the unmodified images in the dataset, which might be the reason why our method performs better on MUSIC-Duet than MUSIC-Synthetic. The experiment results indicate that the proposed cycle-consistency approach leads to improvements in multi-source sound localization.

Qualitative results. In Figure 3 and Figure 4, we visualize the localization maps generated by these methods. It can be seen that models based on Objects that Sound (OTS) [4] mainly focuses on one of the sounding objects, rather than all sounding objects, while our method spreads the probability to all objects. In particular, the audio features obtained by our approach group the visual regions corresponding to both sound sources. For the qualitative results on VoxCeleb2, we found that the model fails more often when the gender of the two speakers is the same.

4.5. Ablation study

We also explore a number of variants of our model for training the self-supervised audio-visual system. We compare our model with several other designs. We keep all the settings the same except the loss function.

Image-audio-image cycle. We consider cycles that start from image nodes, rather than audio nodes. This walk starts at an image, goes to the audio nodes, and finally cycles back to the image nodes. The two image nodes here are sampled from the same video, such that the semantic meaning of the nodes does not change. The similarity between two images, therefore, is evaluated by the probability that they reach each other on a cross-modal random walk. This model minimizes the loss:

$$\mathcal{L}_{\texttt{ISI}} = \frac{1}{k}\operatorname{tr}(\log(A_{IS}A_{SI})). \tag{6}$$

We call this model the *ISI* model, due to the image-sound-image path that the random walker takes.

	MUSIC-Synthetic				MUSIC-Duet			
	CAP	PIAP	AUC	CIoU@0.3	CAP	PIAP	AUC	CIoU@0.3
Corre	12.6	16.0	7.3	0.0	19.3	21.1	17.8	7.7
ISI	11.0	16.4	7.3	0.0	19.7	24.9	17.5	7.6
Permute	18.2*	24.4	9.1*	0.4*	24.0*	28.1	19.5*	12.4*
Ours	**34.0***	**39.7**	**20.2***	**21.3***	**47.4***	**53.9**	**21.2***	**26.3***

Table 3. **Ablation study.** We evaluate the sound source localization performance on MUSIC-Synthetic and MUSIC-Duet datasets for each ablation model. *Corre* denotes the mixed correspondence model while *Permute* represents the model with permutation invariant loss.

Mixed correspondence loss. To test whether the model benefits from cycle-based training (rather than other model differences), we compared with a model trained with the InfoNCE [31] loss with exactly the same input (a single frame and mixed audio). Since we do not know the association between audio nodes and image nodes, we modify Eq. (2) to account for it, *i.e.*,

$$A_{IS}(i,j) = \frac{\sum_{t=1}^{k}\exp(\phi(\mathbf{v}_i,\mathbf{s}_t^{(j)})/\tau)}{\sum_{\mathbf{s}_t\in S}\exp(\phi(\mathbf{v}_i,\mathbf{s}_t)/\tau)}, \tag{7}$$

where $\mathbf{s}_t^{(j)}$ is one of the k audio embeddings generated by the mixed audio for example j, S is the set of all audio embeddings in the batch, and ϕ is defined as in Eq. (2). In contrast to our method, this loss obtains significantly more negative samples by obtaining them from other examples in the batch.

Permutation invariant loss. Inspired by audio source separation methods [22, 47], we ask whether the association between audio and images can be learned from a *permutation invariant* loss. We consider all possible pairings of image and audio embeddings, and select the one with the largest total similarity. For $k = 2$, this loss is:

$$\begin{aligned}\mathcal{L}_{\texttt{PIT}} = -\max(&\phi(\mathbf{v}_i,\mathbf{s}_1)+\phi(\mathbf{v}_j,\mathbf{s}_2),\\ &\phi(\mathbf{v}_i,\mathbf{s}_2)+\phi(\mathbf{v}_j,\mathbf{s}_1)),\end{aligned} \tag{8}$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ are a pair of images used to create a synthetic mixture, and $\mathbf{s}_1$ and $\mathbf{s}_2$ are the audio embeddings

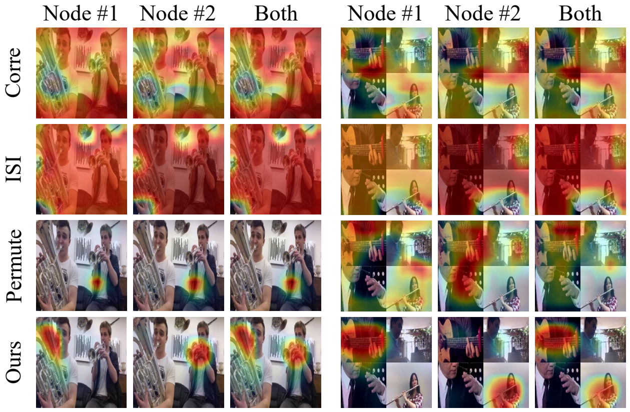


Figure 5. **Ablation study.** We visualize the score maps for these methods on MUSIC-Synthetic and MUSIC-Duet dataset. *Corre* denotes the mixed correspondence model while *Permute* represents the model with permutation invariant loss.

produced from their mixed sound. While this loss is similar to $\mathcal{L}_{\texttt{cyc}}$ (Eq. 4), it creates a "hard" association between images and audio embeddings via the $\max$ operation. By contrast, the random walk model makes "soft" assignments between the embeddings that during learning provides a gradient signal for both possible pairings.

Results. We evaluated multi-source localization on MUSIC-Synthetic and MUSIC-Duet datasets in Table 3. In Figure 5, we visualize the score maps predicted by these methods. These methods fail to produce different and correct localization maps for the two audio nodes, indicating they fail to produce different audio embeddings for each sound source. It can be seen that our method outperforms all these variants.

Although the ISI model is also based on cycle consistency, it does not learn to explicitly separate the two audio nodes. By contrast, our model needs to create two distinct embeddings for the audio nodes in order to successfully complete a cycle. Compared to mixed correspondence loss, which uses other images and audio in the batch as the negative samples for contrastive learning, our method instead takes advantage of other audio nodes derived from the same mixed audio. This will also encourage the model to separate different audio nodes. Moreover, unlike the permutation invariant model that requires a "hard" correspondence for every pair of images and audio, our method allows the model to assign a probability to the graph edge.

5. Discussion

In this paper, we have proposed a simple, self-supervised method for visually localizing sounds in audio mixtures. Our approach is based on learning a cycle-consistent random walk on a graph that connects nodes defined by the images and sound. We showed that our method identifies and segments multiple sound sources more accurately than other self-supervised methods based on traditional audio-visual correspondence learning.

Our results suggest that cycle-consistent random walks can be used to successfully group the contents of a multimodal scene into distinct objects. We hope that these techniques can be combined with tracking-based cycle consistency [25, 46] to group the scene contents over time as well. We also hope this approach provides further directions for the study of cross-modal learning. One such direction is to directly combine explicit source separation [22] with the "implicit" source separation we obtain through contrastive learning.

Limitations. Our released models are limited in scope to the benchmark video datasets they were trained on. Since these are popular datasets, information about their biases is publicly available. As in other audio-visual speech work [30] the learned models may learn to correlate visual properties of a speaker with their voice, making them susceptible to bias.

Acknowledgements. We thank Di Hu and Yake Wei for the help with the experimental setup. We would also like to thank Linyi Jin for suggestions on figure design. This work was funded in part by DARPA Semafor and Cisco Systems. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*, 2018. 5

[2] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. *arXiv preprint arXiv:2008.04237*, 2020. 1, 2, 5

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 1, 2

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *ECCV*, pages 435–451, 2018. 1, 2, 5, 6, 7

[5] Yuki M Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *arXiv preprint arXiv:2006.13662*, 2020. 2

[6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29:892–900, 2016. 2

[7] Zhangxing Bian, Allan Jabri, Alexei A Efros, and Andrew Owens. Learning pixel trajectories with multiscale contrastive random walks. *arXiv preprint arXiv:2201.08379*, 2022. 2

[8] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1204–1213, 2021. 2

[9] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021. 2

[10] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1, 5

[11] Ziyang Chen, Xixi Hu, and Andrew Owens. Structure from silence: Learning scene structure from ambient sound. In *5th Annual Conference on Robot Learning*, 2021. 2

[12] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3133–3142, 2020. 6

[13] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 1, 5

[14] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 2

[15] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shunichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009. 2

[16] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *European Conference on Computer Vision*, pages 658–676. Springer, 2020. 2

[17] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 2

[18] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. *arXiv preprint arXiv:2101.03149*, 2021. 2

[19] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European conference on computer vision (ECCV)*, pages 649–665, 2018. 2

[20] Monica L Hawley, Ruth Y Litovsky, and H Steven Colburn. Speech intelligibility and localization in a multi-source environment. *The Journal of the Acoustical Society of America*, 105(6):3436–3448, 1999. 1

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[22] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016. 2, 7, 8

[23] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019. 2, 5

[24] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 5, 6, 7

[25] Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. *Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 3, 8

[26] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep u-net convolutional networks. 2017. 2

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[28] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *arXiv preprint arXiv:1807.00230*, 2018. 2

[29] Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. Move2hear: Active audio-visual source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 275–285, 2021. 2

[30] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8427–8436, 2018. 8

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 7

[32] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 1, 2

[33] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016. 2

[34] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pages 801–816. Springer, 2016. 2

[35] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. *arXiv preprint arXiv:2007.06355*, 2020. 2, 5

[36] Kranthi Kumar Rachavarapu, Vignesh Sundaresha, AN Rajagopalan, et al. Localize to binauralize: Audio spatialization from visual sound source localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1930–1939, 2021. 2

[37] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised audio-visual co-segmentation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2357–2361. IEEE, 2019. 5

[38] Sam T Roweis. One microphone source separation. In *NIPS*, volume 13, 2000. 2

[39] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 1, 2, 5, 6

[40] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000. 2

[41] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*, 2018. 2

[42] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2745–2754, 2021. 2

[43] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 2

[44] Francisco Rivera Valverde, Juana Valeria Hurtado, and Abhinav Valada. There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11612–11621, 2021. 2

[45] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 391–408, 2018. 2

[46] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019. 2, 3, 8

[47] Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron J Weiss, Kevin Wilson, and John R Hershey. Unsupervised sound separation using mixture invariant training. *arXiv preprint arXiv:2006.12701*, 2020. 7

[48] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 3, 4

[49] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 2

[50] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9932–9941, 2020. 2

[51] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE, 2017. 2

[52] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1735–1744, 2019. 2, 6

[53] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 1, 2, 5, 6, 7

[54] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2