A Proposal-based Paradigm for Self-supervised Sound Source Localization in Videos

Hanyu Xuan[1], Zhiliang Wu[1], Jian Yang[1], Yan Yan[2]*, Xavier Alameda-Pineda[3]†
[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2] Department of Computer Science, Illinois Institute of Technology, USA
[3] Inria, Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
{xuanhanyu,wu_zhiliang,csjyang}@njust.edu.cn, yyan34@iit.edu, xavier.alameda-pineda@inria.fr

Abstract

Humans can easily recognize where and how the sound is produced via watching a scene and listening to corresponding audio cues. To achieve such cross-modal perception on machines, existing methods only use the maps generated by interpolation operations to localize the sound source. As semantic object-level localization is more attractive for potential practical applications, we argue that these existing map-based approaches only provide a coarse-grained and indirect description of the sound source. In this paper, we advocate a novel proposal-based paradigm that can directly perform semantic object-level localization, without any manual annotations. We incorporate the global response map as an unsupervised spatial constraint to weight the proposals according to how well they cover the estimated global shape of the sound source. As a result, our proposal-based sound source localization can be cast into a simpler Multiple Instance Learning (MIL) problem by filtering those instances corresponding to large sound-unrelated regions. Our method achieves state-of-the-art (SOTA) performance when compared to several baselines on multiple datasets.

1. Introduction

In human multi-modal perception [35], the natural co-occurrence of audio-visual events in time provides potential cues for better perception [4]. Such temporal co-occurrence comes from the fact that "Sound is produced by the vibration of objects". By such inherent and universal correspondence, we can distinguish and associate diverse visual appearances from their produced sounds, i.e., the sound source can be visually localized and detected.

Machine intelligence should also have the human-like ability of sound source localization in visual scenes, i.e., the sound source can be localized once they appear in the images. Such ability can play an important role in many practical applications. For example, the rescue robots can find people who are calling for help after a disaster [42], and the sound can help photographers focus better, since the sound source is usually more interesting subjects in photos [19].

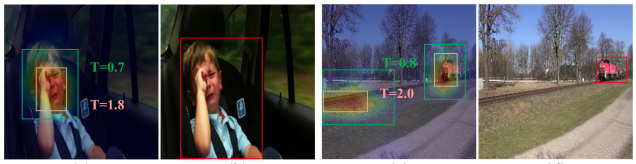

Figure 1. Map-based paradigm uses the maps (Fig.(a) and Fig.(c)) generated by interpolation operations to perform pseudo pixel-level localization or indirect detection by a post-processing step. The typical issues of such maps: 1) tend to highlight the most discriminative regions; 2) may encounter a pathological bias, e.g., rail for "train" and 3) it is non-trivial to choose an optimal threshold since different thresholds T can result in the bounding boxes with different sizes and locations, which are marked in green and pink. In this paper, we advocate a novel proposal-based paradigm, which directly performs semantic object-level localization of class-agnostic sound source (Fig.(b) and Fig.(d)).

For this purpose, some works [29, 38, 44] exploit the attention map to localize the sound source in a weakly-supervised manner. Other methods modify the CNN frameworks to generate Class Activation Maps [48] for sound source localization. Nevertheless, these methods can only perform class-specific, rather than class-agnostic localization, owing to the usage of image-level annotations. Collecting these manual annotations is not only labor-intensive and time-consuming, but also often subjective and error-prone. As a promising approach to solve this problem, self-supervised learning [25] was introduced into this task for its excellent data efficiency and generalization ability. For example, some works [1, 6, 20, 21, 31, 47] localize the class-agnostic sound source through the confidence score map learnt from massive unlabeled videos.

*Corresponding author
†This research was supported by ANR ML3RI (19-CE33-0008-01).

Such maps are not only an effective way to explain the network's decision, but also are utilized to localize the sound source. As shown in Fig.1(a), the long-standing weakness for these map-based paradigm is that it often results in *highlighting the most discriminative parts*, and thereby covering partial regions rather than the entire extent of objects. Moreover, map-based paradigm may *encounter a pathological bias* in the training samples when the foreground objects incidentally always correlate with the same background object, e.g., the rail for "train" in Fig.1(c).

Furthermore, an up-sampling operation (i.e., interpolation function) is required to generate such maps, and introduces numerous uncertainties. We argue that the maps generated in map-based paradigm not only perform coarse-grained (pseudo pixel-level) localization, but also are too smooth to indicate the accurate boundaries of the sound source. For practical applications, the semantic object-level localization of the sound source is more attractive. Although the bounding box of the sound source can also be inferred from such maps by a post-processing step with fixed thresholds, as shown in the green and pink bounding boxes in Fig.1(a) and Fig.1(c), *it is non-trivial to choose an optimal threshold*, since different thresholds can result in bounding boxes with different sizes and locations.

Despite the fact that visual object detection [10, 22, 24, 50] has been widely studied and has achieved promising results, these current prevailing object detection frameworks cannot be utilized for object-level sound source localization. On the one hand, this task involves both audio and vision rather than only vision. On the other hand, all objects in videos could produce sound, including the objects whose annotations are difficult to obtain due to their large variations. In other words, sound source localization must be addressed in a class-agnostic fashion, and therefore location-, frame- or even video-level annotations are extremely difficult to obtain, if even possible at all.

In this paper, we advocate for a paradigm shift on sound source localization and propose a novel *proposal-based method*, which performs semantic object-level localization of class-agnostic sound source. Mainly, we aspire to make the next logical step in this task. Motivated by the great success of Weakly-Supervised Object Detection (*WSOD*) [28,36,46], we pose the proposal-based sound source localization problem as a Multiple Instance Learning (*MIL*) [11] problem, where the sound source detectors (instance classifiers) are put into the network as hidden nodes to learn in an end-to-end manner. The audio information, viewed as a "*free*" source of weak annotations, is introduced into the traditional *WSOD* framework to establish associations between visible objects and corresponding audio contents.

In our setting, the temporal co-occurrence of audio-visual signals is employed as the only available supervision. In such weak supervision, only considering the features of local instances without looking at a global scope of all instances may make our model difficult to train. Furthermore, there are a large number of sound-unrelated objects in the instances, which introduces distractions during training. To this end, we incorporate the Global Response Map (*GRM*), which reveals the coarse location of the sound source and provides information about its global shape, as an unsupervised spatial constraint to weight the local instances according to how well they cover the estimated *GRM*. Our motivation is that, although some instances only capture the sound source partially, the instances having high spatial overlap with the *GRM* may cover the entire sound source, or at least contain a larger portion. Intuitively, such proposals only cover a partial region of the entire frame and ignore the backgrounds where no objects exist, making our model pay more attention to the regions where the sound source maybe exists since the sound source cannot exist in the backgrounds.

Our main contributions can be summarized as follows:

- A novel proposal-based solution for sound source localization, which directly performs semantic object-level localization;
- An unsupervised spatial constraint is introduced to build an effective three-stream framework, which simplifies our problem and improves training efficiency;
- Alleviating the shortcomings of map-based methods, resulting in SOTA performance.

2. Related Works

Audio-visual Representation Learning aims to learn an universal and effective pattern representation from data automatically. It is motivated by the fact that the quality of the representation usually determines the success of machine learning algorithms [5]. Aytar et al. [3] used the temporal co-occurrence between audio-visual signals to learn powerful audio representations. A student-teacher training process was adopted, which exploits unlabeled videos as a bridge to transfer knowledge from vision to audio. Owens et al. [27] attempted to learn enhanced visual representations considering the sound as supervisory signals. Although these approaches manage to learn enhanced audio or visual representations, they do not address the issue of learning joint audio-visual representations.

For this reason, Chung et al. [7] employed a CNN network called SyncNet to learn joint audio-visual representations. Halperin et al. [15] used the joint representations learned by SyncNet to achieve dynamic temporal alignment between speech and video, synchronising the re-recorded speech clips with the pre-recorded videos. These methods exploited the pairwise loss based on binary matching,

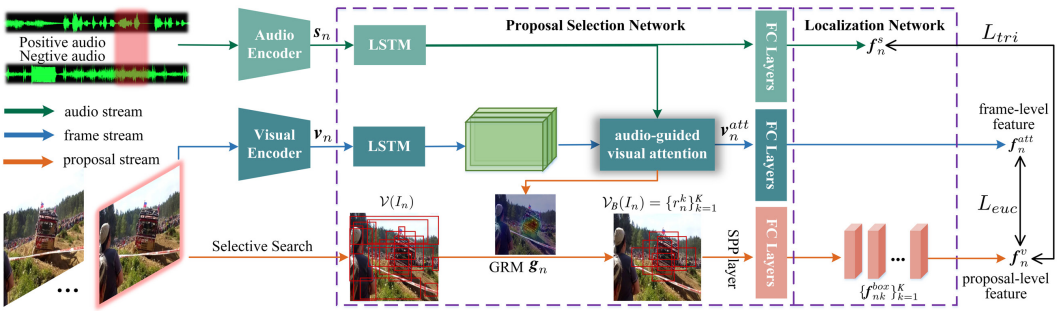

Figure 2. The diagram of the proposal-based sound source localization advocated by us, where different streams are represented by different colors, the *GRM* $\boldsymbol{g}_n$ is exploited as a reliable prior to cast the proposal-based sound source localization problem defined by Eq.1 into a simpler *MIL* problem formulated by Eq.2.

ignoring context information. Chung et al. [8] applied a multi-way matching strategy to train the model. Motivated by these works, we exploit a similar pretext task for self-supervised training.

Weakly-Supervised Object Detection aims to detect the objects through only using image-level instead of location-level annotations. Many methods [9,28,33,46] followed the *MIL* pipeline have acquired promising results. They treat the image as a bag and the region proposals captured by proposal extractors as instances to train the instance classifiers under the *MIL* constraint[1]. A classification loss is usually employed to select the most confident positive proposals. Follow-up works are further improved by leveraging spatial relations [36,37], better optimization [2,23,40], and segmentation collaboration [12,43,45]. Motivated by these methods, we pose the proposal-based sound source localization problem as a *MIL* problem. While *WSOD* accepts well-curated and "*clean*" annotations (correct categories for all samples), our problem involves "*noisy*" annotations, i.e., self-annotated audio-visual pairs may be irrelevant, such as out-screen sound.

Sound Source Localization aims to localize the sound source once they appear in the image. Some works [29, 38, 44] used the attention map to localize the class-specific sound source in a weakly-supervised manner, where Tian et al. [38] introduced an audio-visual distance learning network to achieve sound source localization, Qian et al. [29] performed cross-modal feature alignment in a coarse-to-fine manner and Xuan et al. [44] used cross-modal attention to explore potential hidden correlations of same-modal and cross-modal signals. Some methods [1, 6, 20, 21, 31, 47] localized the class-agnostic sound source by the confidence score map. Arandjelović et al. [1] projected audio and visual features onto common space. Chen et al. [6] introduced an automatic background mining technique. Hu et al. [20] exploited Deep Multi-modal Clustering (DMC) to perform efficient audio-visual matching. Hu et al. [21] employed a two-stage learning strategy. Arda Senocak et al. [31] applied the attention mechanism. Zhao et al. [47] used a mix-then-separate learning strategy.

These existing works can be summarized into a map-based paradigm, which only performs coarse-grained localization or indirect detection of the sound source. Different from them, we propose a proposal-based method that can localize semantic object-level rather than pseudo pixel-level, class-agnostic rather than class-specific sound source.

3. Method

3.1. Problem Formulation

Let N denote the number of frames in a video sequence $\{I_n\}_{n=1}^{N}$, where $I_n \in \mathbb{R}^{W\times H}$, W and H indicate the width and height of the frame I_n respectively. We extract the audio clip A_n of fixed duration centered at the corresponding frame I_n. Our goal is to detect the object that produces the sound A_n in each frame I_n. We denote the set of all class-agnostic proposals in the frame I_n as $\mathcal{V}(I_n)$, which can be easily obtained using an off-the-shelf proposal extractor such as Selective Search [39] or EdgeBox [49]. Our task is to find a subset containing only the sound source:

$$\mathcal{V}_S(I_n) \subseteq \mathcal{V}(I_n). \tag{1}$$

At first glance, it seems to be similar to the classic *WSOD* [32, 36], but there is an essential difference between *WSOD* and ours. Since the image-level annotations provided in *WSOD* satisfy the *MIL* constraint[1], the subset $\mathcal{V}_S$ acquired by *WSOD* is non-empty. However, the set $\mathcal{V}_S$ in our task may be an ***empty*** subset of $\mathcal{V}$ since there is no manual annotation available. $\mathcal{V}_S(I_n) = \varnothing$ means that the sound source may be out of the screen so that we cannot see them in the corresponding video frames. For example, the voice-over of a photographer.

[1]Positive bags have at least one instance corresponding to instance of a certain category, while no instances from the negative bags contain this category.

In this paper, we propose a novel three-stream framework consisting of an audio stream, a frame stream and a proposal stream, aiming to establish the associations between visible objects and audio content. We employ the natural co-occurrence of audio-visual samples in time as the only available supervision. Similar to other end-to-end *WSOD* frameworks [28, 36, 46], our method is built on top of the Region of Interest (RoI) pooling layer [14], RoI align layer [16] or Spatial Pyramid Pooling (SPP) layer [17] to share convolutional (Conv) computations among different proposals for model acceleration.

3.2. Notations and Overview

The frame I_n is fed into a visual encoder to produce a fixed-size feature map $\boldsymbol{v}_n = [\boldsymbol{v}_n^1, \ldots, \boldsymbol{v}_n^c] \in \mathbb{R}^{d_v \times c}$. Correspondingly, the audio clip A_n is fed into an audio encoder to produce a high-level feature $\boldsymbol{s}_n \in \mathbb{R}^{d_s}$. Here, c denotes the vectorized spatial dimension of each visual feature map, d_v indicates the number of visual feature maps and d_a represents the dimension of each audio feature. Respectively, these audio-visual features $\{\boldsymbol{s}_n, \boldsymbol{v}_n\}_{n=1}^N$ are utilized as inputs to modality-specific *LSTMs* to model the temporal dependency. Then, the audio-guide visual attention module is introduced for calculating the global feature $\boldsymbol{v}_n^{att}$ of current frame I_n and generating the *GRM* $\boldsymbol{g}_n$.

For the proposal stream, we first utilize Selective Search as a proposal extractor. Proposal Selection Network exploits the generated *GRM* with coarse-grained location and shape of the sound source to select sound-related proposals. The K most confident proposals are preserved, denoted as $\mathcal{V}_B(I_n) = \{r_n^k\}_{k=1}^K$. Then, these preserved proposals $\mathcal{V}_B$ are fed into the SPP layer to extract their corresponding features $\boldsymbol{B}_n = \{\boldsymbol{b}_n^k\}_{k=1}^K$, where $\boldsymbol{b}_n^k \in \mathbb{R}^{d_b}$, d_b is the dimension of each proposal feature.

Due to the huge gaps between samples in heterogeneous data from audio-visual modalities, we employ some fully-connected (FC) layers to transform the output of three streams into the same dimension d, which makes them comparable. Respectively, we denote the outputs of audio stream, frame stream and proposal stream as $\boldsymbol{f}_n^s, \boldsymbol{f}_n^{att}$ and $\{\boldsymbol{f}_{nk}^{box}\}_{k=1}^K$, where $\boldsymbol{f}_n^s, \boldsymbol{f}_n^{att}, \boldsymbol{f}_{nk}^{box} \in \mathbb{R}^d$. Localization Network exploits these features that have the same dimension in a common comparable space to perform proposal-based sound source localization. The diagram of our method is shown in Fig.2 and will be elaborated on below.

3.3. Proposal Selection Network

As sound implies abundant information about its source and location, some existing map-based methods [1, 20, 21, 44] inspire us to utilize audio signals as a mean of guidance when searching for the location of the sound source. Similar to them, the global feature $\boldsymbol{v}_n^{att} \in \mathbb{R}^{d_v}$ can be obtained in a weighted-average manner, i.e., $\boldsymbol{v}_n^{att} = \sum_{i=1}^c \boldsymbol{w}_n^i \boldsymbol{v}_n^i$. Here, $\boldsymbol{w}_n \in \mathbb{R}^c$ is the spatial attention weight, denoted as $\boldsymbol{w}_n = \phi(\boldsymbol{U}_v \boldsymbol{v}_n)^T \phi(\boldsymbol{U}_s \boldsymbol{s}_n)$, where matrices $\boldsymbol{U}_v$ and $\boldsymbol{U}_s$ are used to project $\boldsymbol{v}_n$ and $\boldsymbol{s}_n$ to the same dimension. ϕ denotes the *softmax* function, used for normalization. Intuitively, such inner product operation measures the cosine similarity between audio and visual features.

After reshaping the weight vector $\boldsymbol{w}_n$ into the spatial dimension c of each visual feature map, the location map $\boldsymbol{m}_n \in \mathbb{R}^{W \times H}$ can be obtained by up-sampling the reshaped $\boldsymbol{w}_n$ to the image size using a interpolation function. We argue that the map-based methods perform indistinct localization and indirect detection of the sound source though such location map. On the one hand, the interpolation operation attempts to utilize the low-dimensional vector $\boldsymbol{w}_n$ to supplement and generate the high-dimensional vector $\boldsymbol{m}_n$, due to $c \ll W \times H$. Inevitably, such operation introduces numerous uncertainties, resulting in ambiguous localization. On the other hand, the location map $\boldsymbol{m}_n$ is too smooth to indicate the accurate boundaries of the sound source.

Although the obtained $\boldsymbol{m}_n$ only reflects a coarse-grained description of the sound source, it provides reliable spatial priors about its location and shape. For this purpose, we normalize the $\boldsymbol{m}_n$ to generate the *GRM* $\boldsymbol{g}_n$, i.e., $\boldsymbol{g}_n = (\boldsymbol{m}_n - min(\boldsymbol{m}_n))/max(\boldsymbol{m}_n)$, which can be regarded as the probability distribution over each pixel in the frame I_n where the sound A_n is produced. The generated *GRM* $\boldsymbol{g}_n$ is used as an unsupervised spatial constraint to weight all extracted proposals $\mathcal{V}$, according to how well they cover the *GRM*. We assume that those proposals with sufficiently high values in the *GRM* should include at least parts of the sound source, and other proposals with enough low values contain sound-unrelated regions (e.g., backgrounds). The confidence of each proposal can be calculated by simply accumulating the corresponding pixel value in the *GRM* $\boldsymbol{g}_n$.

At this point, our proposal-based method defined by Eq. 1 can be transformed into:

$$\mathcal{V}_S \subseteq \mathcal{V} \overset{\mathcal{V}_B \subseteq \mathcal{V}}{\Longrightarrow} \mathcal{V}_S \subseteq \mathcal{V}_B, \tag{2}$$

since $\mathcal{V}_B \subseteq \mathcal{V}$. It means that we only need to find the sound source in the set $\mathcal{V}_B$. Using the set $\mathcal{V}_B$ containing fewer proposals simplifies our problem. We treat the $\mathcal{V}_B$ rather than $\mathcal{V}$ as a bag to detect the sound source by using MIL strategy. Intuitively, it is easier to find the sound source in specific regions of the frame rather than the entire frame. Moreover, though the *GRM* may only capture the most discriminative parts of the sound source, these preserved proposals $\mathcal{V}_B$ having the highest spatial overlaps with the *GRM* may cover the whole sound source, or at least contain a larger portion, resulting in more accurate localization.

3.4. Localization Network

The semantic object-level sound source can be localized by evaluating the association between the preserved pro-

posals $\mathcal{V}_B$ and the corresponding audio clip A_n. In order to cope with the rare case where the number of extracted proposals $\mathcal{V}(I_n)$ is less than K, we design a binary mask $\boldsymbol{e} = [e_1, e_2, \ldots, e_K] \in \mathbb{R}^K$, where the k-th element e_k indicates whether the corresponding proposal exists or not. We employ the cosine similarity to measure the association:

$$d_n^k = ReLU\left(\frac{\boldsymbol{f}_n^{s\,T} \boldsymbol{f}_{nk}^{box}}{\|\boldsymbol{f}_n^s\|_2 \|\boldsymbol{f}_{nk}^{box}\|_2}\right) \cdot e_k, \tag{3}$$

where the $ReLU$ function is employed to suppress negative correlation values, which has been discussed in [31,41]. The semantic object-level sound source $\mathcal{V}_S$ can be localized by these calculated similarities $\{d_n^k\}_{k=1}^K$, which will be detailed in Sect.4.3.

For the proposal stream, we can obtain a proposal-level feature $\boldsymbol{f}_n^v$ through using these similarities to weight corresponding proposals, i.e., $\boldsymbol{f}_n^v = \sum_{k=1}^K \alpha_n^k \cdot \boldsymbol{f}_{nk}^{box}$, where α_n^k is the normalized $\{d_n^k\}_{k=1}^K$. It combines both the regions of each frame and corresponding audio to generate a new representation related to the sound source, resulting in paying more attention to the high probability regions where the sound is produced.

3.5. Loss Function

The pseudo-annotations for our pretext task can be obtained automatically based on whether audio-visual clips co-occur in time. We employ $\boldsymbol{f}_n^v$ from a video as a query, and its positive pair $\boldsymbol{f}_n^{s+}$ is corresponding audio clip from the same video, while negative one $\boldsymbol{f}_n^{s-}$ is the audio clip from another randomly selected video. The distance ratio loss $\mathcal{L}_n^{tri}$[2] [31] makes positive pair $(\boldsymbol{f}_n^v, \boldsymbol{f}_n^{s+})$ close to each other, but negative pair $(\boldsymbol{f}_n^v, \boldsymbol{f}_n^{s-})$ far from each other.

In addition, we apply the Euclidean distance loss $\mathcal{L}_n^{edu}$ [41] to make a connection between the frame stream and the proposal stream to generate the *GRM*. The $\mathcal{L}_n^{edu}$ term encourages the distance between the frame-level feature $\boldsymbol{f}_n^{att}$ and the proposal-level feature $\boldsymbol{f}_n^v$ to be as small as possible, i.e., $\mathcal{L}_n^{edu} = \|\boldsymbol{f}_n^v - \boldsymbol{f}_n^{att}\|_2^2$. Notably, although both $\boldsymbol{f}_n^v$ and $\boldsymbol{f}_n^{att}$ are utilized to represent the frame I_n, their focus is various. As shown in Fig.2, the former, output by the frame stream, concentrates on the approximate regions where the sound is produced, while the latter, output by the proposal stream, pays more attention to those proposals related to the sound source. The final loss function $\mathcal{L}$ of our model for whole video is represented as $\mathcal{L} = \sum_{n=1}^N (\mathcal{L}_n^{tri} + \lambda \mathcal{L}_n^{edu})$, where λ is a balancing parameter.

[2]See supplementary materials for specific expression.

Table 1. The accuracy comparison of on-screen/out-screen sound recognition based on different criteria, i.e., the distance metric m_{on} and the global activation value d_{on}.

	AVE	**SSL**
$m_{on}.$	64.6	71.8
$d_{on}.$	66.2	70.5
$m_{on}. \parallel d_{on}.$	**71.2**	**78.5**

4. Experiments

4.1. Datasets and Implementation Details

Sound Source Localization[3] (SSL) dataset [31] holds 2786 sound-image pairs selected from Flickr-SoundNet dataset, where *Flickr-SoundNet dataset* [3] contains over 2 million unconstrained videos from Flickr. There are three annotations for each sample, by listening the sound for 20s and draw a bounding box on the regions where the dominant sound comes from in the frame. Notably, we employ randomly selected sound-image pairs from Flickr-SoundNet dataset as training samples. And the results of sound source localization are quantitatively evaluated on SSL dataset.

MUSIC[3] dataset [47] comprises 685 untrimmed videos, 536 solo and 149 duet, covering 11 classes of musical instruments. Note that, some videos are now not available on YouTube, we finally get 489 solo and 141 duet videos. For quantitative evaluation, we utilize the bounding boxes provided in [21] as the ground truth on the test set, where the bounding boxes are generated by a well-trained Faster RCNN detector w.r.t 11 instruments.

MUSIC-Synthetic[3] dataset [21] is synthesized by randomly selecting four 1s solo audio-visual pairs of different instruments, then mixing random two of the four audio clips with jittering as the multi-source audio waveform, and concatenating four frames of these clips as the multi-source video frame. It means that there are two instruments producing sound while the other two are silent in the synthesized audio-visual samples.

Audio-Visual Event[3] (AVE) dataset [38] includes 4143 samples covering a wide range of real-life scenes, including music performances, main street, public speaking and so on. Each sample contains 10s audio-visual signals. These samples are filled with abrupt view changes and different types of noise. Notably, although the AVE dataset provides second-level and video-level event category annotations, we do not utilize any annotations during training. Since no ground truth is provided, we present some visualized localization results on AVE dataset.

Implementation details[4]. Given a series of audio-frame pairs $\{A_n, I_n\}_{n=1}^N$ in a video, the VGGish [18] pre-trained

[3]All the datasets we use are available online. The website for downloading these datasets can be found in the supplementary materials.

[4]All the codes we use are under MIT license.

Table 2. Performance comparison of sound source localization with the SOTA map-based methods [1, 6, 20, 21, 31, 47] on three datasets, i.e., MUSIC-solo, MUSIC-Synthetic and SSL. *IoU@0.5*, *AUC* and *cIoU* are reported.

	MUSIC-solo		**MUSIC-Synthetic**		**SSL**	
	IoU@0.5	*AUC*	*IoU@0.5*	*AUC*	*cIoU*	*AUC*
PixelPlayer [47]	40.5	43.3	35.5	11.8	60.4	50.3
Objects-that-Sound [1]	26.1	35.8	20.3	10.2	63.6	48.2
Attention [31]	37.2	38.7	25.9	12.3	66.0	55.8
DMC [20]	29.1	38.0	31.6	16.3	67.1	56.8
AV-Matching [21]	51.4	43.6	40.0	23.5	68.7	56.9
SS-Hard-Way [6]	52.3	44.1	40.6	23.6	69.9	57.3
Ours	***58.0***	***46.3***	***46.7***	***30.1***	***72.8***	***61.4***

on AudioSet [13] and the VGG16 [34] pre-trained on ImageNet [30] are employed to initialize the weights of audio encoder and visual encoder, respectively. In addition, we replace the last max-pooling layer of VGG16 with SPP layer to extract the proposals' feature. To increase the size of each visual feature map, we replace the penultimate max-pooling layer and its subsequent Conv layers with the dilated Conv layers. During the training phase, we employ the Adam optimizer with default parameters. Empirically, the regularization weight λ is set to 0.5. Considering the amount of memory available on the GPU, we set the batch size to 128.

Evaluation metric. In order to quantitatively evaluate the proposed method, we use the evaluation metrics adopted in [6, 21, 31]. For both MUSIC-Synthetic dataset and MUSIC dataset, Intersection over Union (*IoU*) and Area Under Curve (*AUC*) are adopted as evaluation metrics, which are calculated using the predicted regions of the sound source and annotated bounding boxes. For SSL dataset, since multiple annotations are provided, the consensus IoU (*cIoU*) [2] is adopted as an evaluation metric.

4.2. On/out-screen Sound Recognition

As mentioned in Sect.3.1, the sound source may not be on the screen in videos, which will result in $\mathcal{V}_S = \varnothing$. As a result, we first need to recognize whether the sound is on the screen or out of the screen, i.e., on-screen or out-screen sound recognition.

Map-based methods [20, 26, 31] exploit a distance metric m_{on} to recognise on-screen or out-screen sound, where the distance metric m_{on} is usually set to the average of the median of the Euclidean distance of all positive and negative samples in the comparable common space. Compared with map-based methods, our proposal-based method can also perform on/out-screen sound recognition by using the similarities $\{d_n^k\}_{k=1}^K$ calculated in Eq.3. Specifically, we can obtain a global activation value d_{on} to measure the correlation between the frame I_n and the sound A_n, where d_{on} is defined as the maximum value of these similarities. When d_{on} is large enough, i.e., $d_{on} > 0.2$, we consider that the sound is produced by the corresponding object in the corresponding video clip, that is, the sound is on-screen.

To sum up, the distance metric m_{on} and the global activation value d_{on} can be used as a judgment criterion for on-screen/out-screen sound recognition. We argue that using only m_{on} or d_{on} as a criterion is too strict. Therefore, we define a looser criterion, i.e., we consider the sound is on-screen as long as one of the two criteria is met. As shown in the last row of Tab.1, adding a new criterion (i.e., d_{on}) improves the performance of on/out-screen sound recognition on both AVE dataset and SSL dataset.

4.3. Quantitative Comparisons with the SOTA

After judging the sound on the screen, i.e., $\mathcal{V}_S \neq \varnothing$, we localized the semantic object-level sound source based on the similarities $\{d_n^k\}_{k=1}^K$ calculated by Eq.3, elaborated as follows. We compare our method with some recent SOTA map-based methods [1, 6, 20, 21, 31, 47] in three different scenarios. The results of the quantitative comparisons with these methods are presented in Tab.2.

Single Sound Source Scenario First of all, we focus on a simple scenario where there is at most one sound source in the video. In this case, we make use of the proposal with global activation value d_{on} as the final localization result. The performances are evaluated on the MUSIC-solo dataset. Compared with map-based methods, our proposal-based method achieves competitive performance. Map-based methods tend to localize the most discriminative regions, resulting in undesirable results.

Multiple Sound Sources Scenario Multiple sound sources may appear in the video at the same time. For the MUSIC-Synthetic dataset, there are two sound sources in each video. To deal with this situation, the two non-overlapping proposals with the highest confidence are output as the final localization results.

As shown in Tab.2, our proposal-based method significantly outperforms the map-based methods. In this scenario, the map-based methods often obtain a coarse-grained

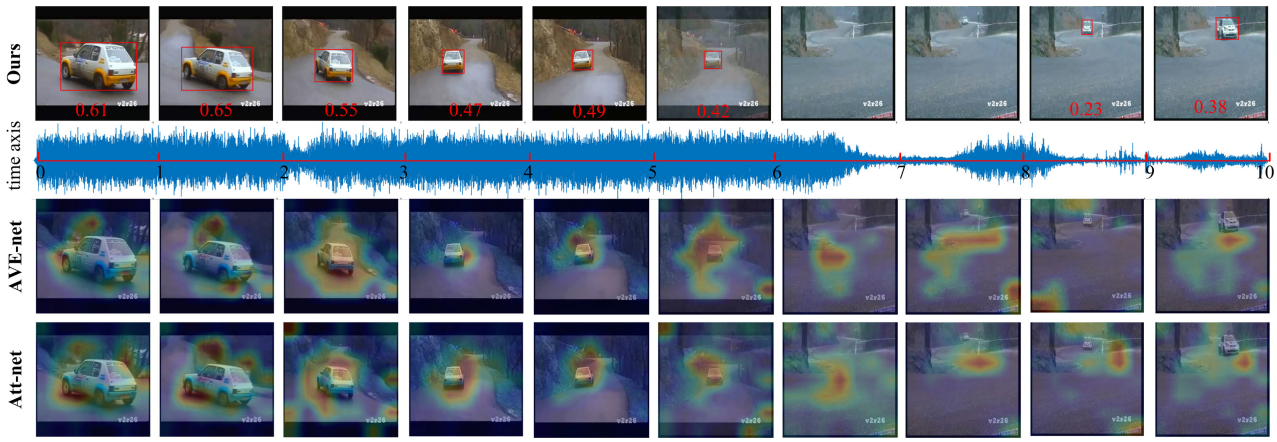

Figure 3. Qualitative comparisons with two recent map-based weakly-supervised methods [38, 44] in this scenario: a car is moving away and then coming in the video. Best viewed in color.

Table 3. Ablation studies on SSL dataset, where the minus sign "−" denotes to ignore specific module in our method.

	cIoU	*AUC*
Ours(K=10)	68.9	56.2
Ours(K=20)	***72.8***	***61.4***
Ours(K=30)	73.1	57.3
Ours − ProSelectNet	58.9	48.5
The location map $\boldsymbol{m}_n$	*69.2*	*59.3*

Table 4. Performance evaluation on SSL dataset using various numbers of training samples from the Flickr-SoundNet dataset.

	cIoU		*AUC*	
	Attention [31]	**Ours**	Attention [31]	**Ours**
10k	43.6	45.4	44.9	51.3
50k	—	55.6	—	55.4
90k	—	63.7	—	58.7
144k	66.0	***72.8***	55.8	***61.4***

map to cover the entire sounding areas without distinguishing each sound source, resulting in low *AUC*. Notably, PixelPlayer [47] could effectively correlate specific visual region with audio in the simple scene with single sound source, but suffers from the noisy scenario with multiple sound sources, as it highly dependent on the quality of the input audio during training.

Complex Real-life Scenario For real-life videos with complex scenario, the sound source localization is challenging for machines. Compared with both MUSIC dataset and MUSIC-Synthetic dataset, the Flickr-SoundNet dataset covers more complex scenarios. As shown in Tab.2, our proposal-based method outperforms other map-based methods. Notably, AV-Matching [21] uses a two-stage learning strategy, which extremely depends on the quality of pre-learned knowledge in the first stage. In contrast, our method does not require to construct a complex learning strategy.

4.4. Qualitative Comparisons

To further validate the effectiveness of our method, we visualize some localization results compared with two recent map-based weakly-supervised methods, i.e., AVE-net [38] and Att-net [44], in two representative scenes. The visualization results of qualitative comparisons are shown in Fig.3 and Fig.4, where the final localization results output by our proposal-based method are marked in red. Notably, both employ video-level annotations, while we did not use any manual annotations.

As shown in Fig.3, in the first 6s, when the car is moving away and the sound is getting weaker and weaker, our method localizes the car with a lower and lower confidence. In the last 2s, when the car is coming closer and closer and the sound is more and more apparent, our method captures the car with a higher and higher confidence. As shown in Fig.4, a percussion instrument is being played in the first 4s, while a ukulele is being played in the next 6s. We mark these two probable sound sources and corresponding confidence in each frame. With the switching of two instruments, the corresponding confidence is also changing.

To sum up, the map-based methods, whether it is based on weakly-supervised or self-supervised setting, suffer from the shortcomings mentioned in Sect.1. Our proposal-based method alleviates such shortcomings to some extent, resulting in more accurate localization.

More visualisation results of our method on MUSIC-Synthetic dataset and SSL dataset are available in the supplementary materials. In addition, we also analyse ***some failures and limitations*** about our method in some extreme

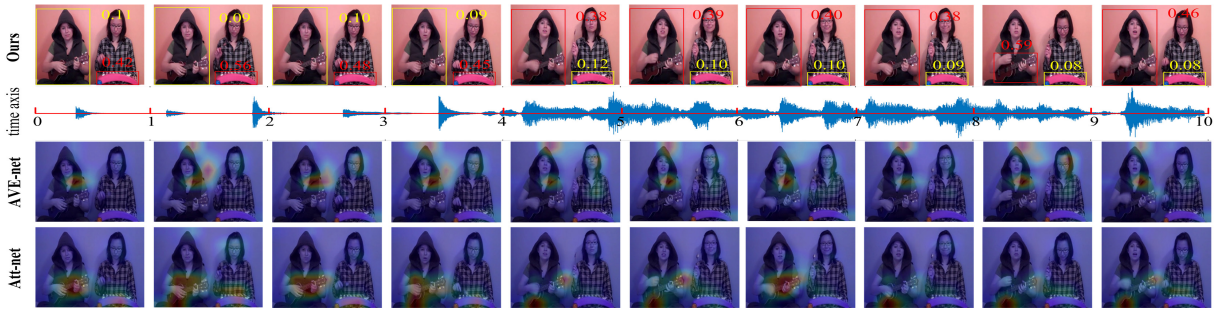

Figure 4. Qualitative comparisons with two recent map-based weakly-supervised methods [38, 44] in this scenario: two instruments are being played in the video, where two main probable sound sources are marked with orange and red bounding box. Best viewed in color.

(a) (b)

Figure 5. The visualization of the location map, where Fig.(a) is generated by our proposal-based method, Fig.(b) is generated by the map-based Attention [31].

scenarios. The failure results and limitation analyses are presented as supplementary materials.

4.5. Ablation Study & Validity Verification

Ablation Study First of all, we explore the impact of the value of K in Sect.3.4 on the localization performance of the sound source. As shown in the first-three rows of Tab.3, when K is set to 20, our method achieves the best performance on both *cIoU* and *AUC*.

In addition, the *GRM* in Sect.3.3 is employed as a spatial constraint for selecting sound-related proposals. To evaluate its necessity, we ignore our Proposal Selection Network and only employ audio stream and proposal stream, i.e., train our model with the distance ratio loss $\sum_{n=1}^{N} \mathcal{L}_n^{tri}$. It means that we need to find the sound source $\mathcal{V}_S$ in all proposals $\mathcal{V}$ extracted by Selective Search. As shown in the 5th row of Tab.3, the performance drops significantly. Such result indicates that using selected proposals is better than using all proposals, as described in Sect.3.3.

Validity Verification Tab.4 shows the performance evaluation on SSL dataset when using various numbers of training samples from the Flickr-SoundNet dataset. As expected, both Attention [31] and ours are gradually improving performance with the increase of training samples. Compared with Attention [31], our method achieves better performance on SSL dataset.

Similar to map-based methods, our location map $\boldsymbol{m}_n$ in Sect.3.3 can be used not only to generate the *GRM* for proposal selection, but also directly to perform the sound source localization. We evaluate the localization performance of our location map on SSL dataset. The result is presented in the last row of Tab.3. Compared with map-based methods (see Tab.2), our location map achieves the competitive results. In addition, we also visualize our location map compared with the location map generated by map-based Attention [31]. A visual comparison can be found in Fig.5. Our location map tends to capture larger and more complete shapes of the sound source thanks to the introduction of proposal stream.

4.6. Discussion

It is not difficult to find that the map-based methods only employ frame-level features, while our proposal-based method makes also use of proposal-level features in our proposal stream. The introduction of proposal-level features makes our location map can alleviate the shortcomings of the location map generated by map-based methods to some extent, resulting in more precise localization. On the other hand, the addition of a new criterion allows our model to more accurately recognize on/out-screen sound. Furthermore, the usage of proposal extractor makes our method to prevent from an non-trivial post-processing step.

5. Conclusion

In this paper, we advocate a novel proposal-based solution for sound source localization, which performs semantic object-level localization of class-agnostic sound source. First of all, we give a formal definition of our proposal-based method. The global response map is incorporated as an unsupervised spatial constraint to filter the proposals with massive sound-unrelated objects and backgrounds. These preserved proposals are treated as a MIL problem with "*noisy*" annotations. The quantitative and qualitative comparisons with map-based methods on multiple datasets demonstrate the effectiveness of our method. We also compare some differences between the map-based method and our proposal-based method.

References

[1] Relja Arandjelovìc and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 3, 4, 6

[2] A. Arun, C. Jawahar, and M. Kumar. Dissimilarity coefficient based weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9424–9433, 2019. 3

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 892–900, 2016. 2, 5

[4] Pascal Belin, Shirley Fecteau, and Catherine Bedard. Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences*, 8(3):129–135, 2004. 1

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(8):1798–1828, 2013. 2

[6] 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. 1, 3, 6

[7] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Proceedings of the Asian Conference on Computer Vvision (ACCV)*, pages 251–263, 2016. 2

[8] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3965–3969, 2019. 3

[9] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 39(1):189–203, 2016. 3

[10] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4013–4022, 2020. 2

[11] Zhouyu Fu, Antonio Robles-Kelly, and Jun Zhou. Milis: Multiple instance learning with instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(5):958–977, 2010. 2

[12] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1277–1286, 2018. 3

[13] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 6

[14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 1440–1448, 2015. 4

[15] Tavi Halperin, Ariel Ephrat, and Shmuel Peleg. Dynamic temporal alignment of speech to lips. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3980–3984, 2019. 2

[16] K. He, G. Gkioxari, P. Doll¨¢r, and R. Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 42(2):386–397, 2020. 4

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 37(9):1904–1916, 2015. 4

[18] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017. 5

[19] Ala' Abu Hilal and Thabet Mismar. Drone positioning system based on sound signals detection for tracking and photography. In *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0008–0011, 2020. 1

[20] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9248–9257, 2019. 1, 3, 4, 6

[21] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 3, 4, 5, 6, 7

[22] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu. A survey of deep learning-based object detection. *IEEE Access*, 7:128837–128868, 2019. 2

[23] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4294–4302, 2017. 3

[24] Wei Ke, Tianliang Zhang, Zeyi Huang, Qixiang Ye, Jianzhuang Liu, and Dong Huang. Multiple anchor learning for visual object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10206–10215, 2020. 2

[25] Yann LeCun. Self-supervised learning: The plan to make deep learning data-efficient. https://bdtechtalks.com/2020/03/23/yann-lecun-self-supervised-learning/. 1

[26] 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. 6

[27] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–816, 2016. 2

[28] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10598–10607, 2020. 2, 3, 4

[29] Qian Rui, Hu Di, Dinkel Heinrich, Wu Mengyue, Xu Ning, and Lin Weiyao. Multiple sound sources localization from coarse to fine. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 292–308, 2020. 1, 3

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 6

[31] 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 (CVPR)*, pages 4358–4366, 2018. 1, 3, 5, 6, 7, 8

[32] Yunhang Shen, Rongrong Ji, Kuiyuan Yang, Cheng Deng, and Changhu Wang. Category-aware spatial constraint for weakly supervised detection. *IEEE Transactions on Image Processing (TIP)*, 29:843–858, 2020. 3

[33] Miaojing Shi, Holger Caesar, and Vittorio Ferrari. Weakly supervised object localization using things and stuff transfer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3381–3390, 2017. 3

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2014. 6

[35] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial Life*, 11(1-2):13–29, 2005. 1

[36] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan L Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 42(1):176–191, 2020. 2, 3, 4

[37] P. Tang, X. Wang, X. Bai, and W. Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3059–3067, 2017. 3

[38] 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. 1, 3, 5, 7, 8

[39] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013. 3

[40] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[41] Feng Wang, Di Guo, Huaping Liu, Junfeng Zhou, and Fuchun Sun. Sound-indicated visual object detection for robotic exploration. In *International Conference on Robotics and Automation (ICRA)*, pages 8070–8076, 2019. 5

[42] Lin Wang, Ricardo Sanchez-Matilla, Andrea Cavallaro, et al. Audio-visual sensing from a quadcopter: dataset and baselines for source localization and sound enhancement. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 5320–5325, 2019. 1

[43] Li Xiaoyan, Shan Meina, Kanand Shiguang, and Chen Xilin. Weakly supervised object detection with segmentation collaboration. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9735–9744, 2019. 3

[44] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 279–286, 2020. 1, 3, 4, 7, 8

[45] Gao Yan, Boxiao Liu, Nan Guo, Xiaochun Ye, and Dongrui Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[46] Ke Yang, Peng Zhang, Peng Qiao, Zhiyuan Wang, Huadong Dai, Tianlong Shen, Dongsheng Li, and Yong Dou. Rethinking segmentation guidance for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 946–947, 2020. 2, 3, 4

[47] 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 587–604, 2018. 1, 3, 5, 6, 7

[48] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2921–2929, 2016. 1

[49] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 391–405, 2014. 3

[50] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019. 2