Complementary Cues from Audio Help Combat Noise in Weakly-Supervised Object Detection

Cagri Gungor[1] and Adriana Kovashka[1,2]
[1]Intelligent Systems Program, [2]Department of Computer Science
University of Pittsburgh
cagri.gungor@pitt.edu, kovashka@cs.pitt.edu
https://cagrigungor.github.io/AudioVisualWSOD/

Abstract

We tackle the problem of learning object detectors in a noisy environment, which is one of the significant challenges for weakly-supervised learning. We use multimodal learning to help localize objects of interest, but unlike other methods, we treat audio as an auxiliary modality that assists to tackle noise in detection from visual regions. First, we use the audio-visual model to generate new "ground-truth" labels for the training set to remove noise between the visual features and noisy supervision. Second, we propose an "indirect path" between audio and class predictions, which combines the link between visual and audio regions, and the link between visual features and predictions. Third, we propose a sound-based "attention path" which uses the benefit of complementary audio cues to identify important visual regions. We use contrastive learning to perform region-based audio-visual instance discrimination, which serves as an intermediate task and benefits from the complementary cues from audio to boost object classification and detection performance. We show that our methods, which update noisy ground truth and provide indirect and attention paths, greatly boosting performance on the AudioSet and VGGSound datasets compared to single-modality predictions, even ones that use contrastive learning. Our method outperforms previous weakly-supervised detectors for the task of object detection by reaching the state-of-art on AudioSet, and our sound localization module performs better than several state-of-art methods on AudioSet and MUSIC.

1. Introduction

There has recently been a rise in interest in multimodal learning, where multiple channels are available to help detect the presence of objects or events. In particular, several tasks exist that use both audio and visual information for prediction. Examples include multimodal pretraining,

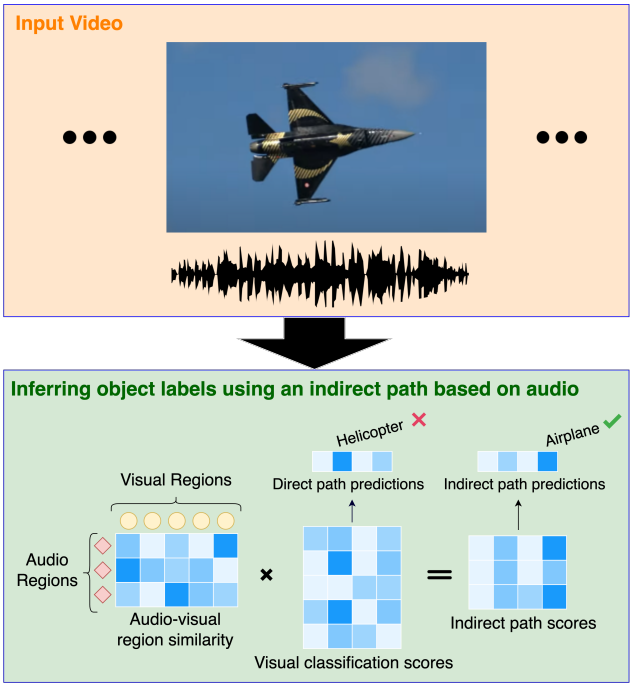

Figure 1. Illustration of one of our contributions which utilizes complementary audio cues. Our method includes a region-based audio-visual instance discrimination module which produces an audio-visual region similarity, in conjunction with the visual classification scores, to create an indirect visual path through audio and improve the accuracy of predicted object labels.

where visual representations are learned jointly with the aid of audio ones (either using raw sound or extracting any present speech from it), and sound localization, where the locations of objects that produce sounds are inferred based on visual features. Learning from multimodal data offers new opportunities: the supervision obtained from such data can be considered "free", however, when labels (e.g. object categories) are extracted from such data, these labels may be noisy. With video data specifically, even if a label (e.g. an

object label extracted from the accompanying speech) applies to a video clip as a whole, the object may not manifest in all frames of that clip.

In this paper, we explore the potential of treating audio as an auxiliary modality that helps deal with noise and errors in the predictions of the model that is based on visual features. Our method is related to sound localization [2, 21, 4, 44, 28, 22, 24] because we learn corresponding visual regions to audio similar to prior work. However, it is different because our main task is object detection, where objects can be detected whether or not they make a sound. We show that if there is relevant audio accompanying an image, complementary cues from the audio improve detection performance. We believe our research has significant potential in industry, especially for autonomous cars that frequently detect objects that make noise.

We focus on the task of weakly-supervised object detection in video, where object annotations are available only at the video level. We consider a noisy setting, where not all frame-level object labels are correct, due to the challenge of extracting those from the multimodal data. We then propose three mechanisms for audio to serve as a helper modality to help cope with the noise and provide a measure of confidence for predictions from the visual channel.

Our *first innovation* relies on the intuition that the noise will affect the visual and audio-visual predictions in a different way because audio provides complementary cues. Thus, to deal with the noise in the visual predictions, we compute a new "ground-truth" training set based on the predictions from the audio-visual model. We then retrain all models with this new ground-truth set.

Our *second innovation* uses the associations between the visual frames and corresponding audio (obtained through region-based audio-visual instance discrimination) to provide secondary, additional evidence to predict an object label in a given frame (or region of a frame). In particular, we look for audio as an intermediate link between the visual features and object predictions, and make predictions using this indirect path. This and the next contribution significantly improve classification performance, whether or not the labels in the original dataset are clean or noisy. This part of our contribution is illustrated in Fig. 1.

Our *third innovation* uses sound as an attention mechanism to determine important visual regions, as a further technique for dealing with noisy predictions. In particular, given an association matrix of region and sound tokens, we compute aggregated importance for each region, and use this to weight our region-level object predictions.

We conduct experiments on two datasets, VGGSound [9] and AudioSet [18]. Additionally we use the MUSIC [44] dataset to compare with prior sound localization methods. VGGSound and AudioSet match the characteristics of multimodal learning we are interested in, namely that they contain rich complementary information from audio and visual features. However, these datasets contain relatively clean object category labels, which does not match our envisioned setting because providing such clean labels requires human effort. To make the setting more realistic, we artificially introduce noise in the labels in these datasets, at the video clip level, by flipping a small fraction of clip labels to mimic the natural noise that is present on the internet—for example, noisy supervision in the form of natural language descriptions that web users provide when uploading their videos to social media sites. We show that our methods very successfully help cope with the noise in the training set, and we achieve comparable results to those using the expensive noise-free training set. Further, we show the individual contribution of each of our method components. In particular, our sound-based indirect and attention paths boost results over just using contrastive learning and instance discrimination in the sound localization module. *This improvement holds both in the clean and noisy video clip label settings.*

Our method performs better than other weakly-supervised object detectors on AudioSet [18]. Furthermore, our sound localization module outperforms several recent methods on the AudioSet [18] and MUSIC datasets [44], although sound localization is not our main task.

To summarize, our contributions are: (1) a method that handles noise in the object labels, by inferring additional ground-truth labels from the audio channel and retraining the visual channel with those; (2) a method that uses audio as an extra link between the visual input and predicted labels (i.e., an object label should be inferrable from the visual channel alone, as well as indirectly: through sound tokens associated with the label, which themselves are related to the visual input); and (3) a method that uses sound to infer which visual regions are important for predicting object labels.

2. Related Work

Weakly-supervised object detection (WSOD). WSOD is the task of learning to predict categories and locations of objects, from only image-level labels available at training time. The problem implies a multiple-instance learning framework, where the regions in the image are considered a "bag", and the image-level object label suggests that at least one of the items in the bag contains the object. Thus, the image-level prediction can be computed as a (weighted) summation over region-level scores for the object of interest, and then a loss can be formulated over this image-level prediction. Example approaches include [6, 36, 39, 17, 35]. Some approaches rely on an iterative improvement where high-scoring proposals are treated as pseudo ground-truth [36, 40, 43, 31, 33, 35].

There have also been works that alleviate the need for image-level labels by extracting noisy label information

from caption or subtitle data [41, 10, 42, 38, 14]. In contrast to these works, we perform WSOD using visual and audio data, by using audio to provide confidence in the visual predictions, and we deal with noise in the extracted labels through this additional modality.

Multimodal pretraining. Researchers have proposed learning visual representations in a joint multimodal context, through techniques such as contrastive learning. The modalities are often images, video, text, and sound. For example, Miech et al. [26] learn to project video and temporally co-occurring narrations close together in a learned space, in contrast to non-co-occurring video and narrations, which should be far. Alayrac et al. [3] learn how to best fuse the visual, audio, and language modalities. Chen et al. [11] ensure cooperation between image, video, and sound features through distillation. Bertasius et al. [5] and Zareian et al. [42] obtain representations with contrastive learning for object detection tasks, but require some manually labeled bounding boxes. Morgado et al. [27] only consider visual and audio features (no speech), and contrast representations both within and across modalities. Representations have also been learned using transformer architectures, for example in the context of joint image-text representations learned on massive datasets (CLIP [29], UNITER [12], LXMERT [34], Vilbert [25], etc.) or even smaller datasets [13].

These works perform pretraining, where some amount of noise in terms of lack of semantic matches between the different co-occurring modalities is tolerable. We instead focus on a downstream object detection task, trained in a supervised way with labels whose purity is important. We thus propose how to use sound as a helper modality to help cope with noise in the visual predictions.

Afouras et al. [1] extract supervisory signal from audio-visual data to teach object detector in a self-supervised manner. In the first stage, their method learns pseudo labels and boxes in a contrastive sound localization network. In the second stage, pseudo labels and boxes are used to train Faster-RCNN [30]. They also experiment with a weakly-supervised version using ground truth labels rather than pseudo labels. Unlike Afouras et al. [1], we propose a novel end-to-end network that trains object detector and sound localization modules together. We use audio signals during object detection training and inference so audio modality has a direct effect on performance on test data. For example, during the detection of a car object in a test image, the audio of the car is used to improve the detection. Further, we use visual region proposals, while they use spatial visual features to localize visual regions with the help of audio. While we train our methods on AudioSet for 5 epochs on 2 GPUs in less than a day, Afouras [1] train their methods on AudioSet for 230 epochs on 64 GPUs over 3 days.

Sound localization and separation. Sound localization [4, 8, 28, 24] is to find the sounding region in the visual scene. [4, 8] calculate similarity between audio and spatial visual features to produce a heatmap. [16, 44, 15] perform separation of sound mixtures by estimating spectrogram masks based on visual signals. [37, 2, 32] propose audio-guided attention mechanisms. [2] utilize audio-visual concurrency to train a video model capable of distinguishing and grouping occurrences of the same category. [4, 37, 8] use contrastive learning to link audio and visual information for localization and separation. [21, 22] propose using an object dictionary and training a model using category-level audio-visual distribution matching to understand the category of sound sources.

Prior works do not have a special object detection module and do not use any object detection labels, nor do they do training for detection. According to our knowledge, our method is the first method that uses sound in an object detection network in an end-to-end manner. Our attention approach detects important visual regions with the guidance of audio, similar to prior works [37, 2, 32] but findings are used to enhance the predictions from visual features in object detection module. In prior sound localization works, audio is an indispensable modality to localize the visual regions. However, our object detection module detects all target objects not considering whether they produce sound. In case the audio is not existing or is unrelated, our method still detects objects. Different than [4, 37, 8], we use region proposals rather than spatial visual features in contrastive learning. While prior works use the metrics such as IoU, CIoU and AUC, we use mAP metric from object detection literature. While [21, 22] produces class pseudo labels, we use image-level labels in weakly-supervised object detection module. However, our sound localization module do not use labels. It produces class-agnostic predictions.

3. Method

Our goal is to learn visually-based object detectors in a weakly-supervised manner with the help of the audio modality. We set our work in the classic weakly-supervised object detection (WSOD) setting. However, we use region-based audio-visual instance discrimination to define a sound indirect path to object predictions, and sound-based attention mechanism for visual regions. Our approach consists of three stages described in Fig. 2: visual detection module for weakly-supervised detection, audio detection module, and region-based audio-visual instance discrimination module.

3.1. Visual detection module

The visual module closely follows prior work in weakly-supervised object detection [6, 41, 38]. We extract visual proposals with their accompanying features. An image is fed into the visual convolutional layers. Then, ROIAlign is used for cropping the proposals, and visual regions are

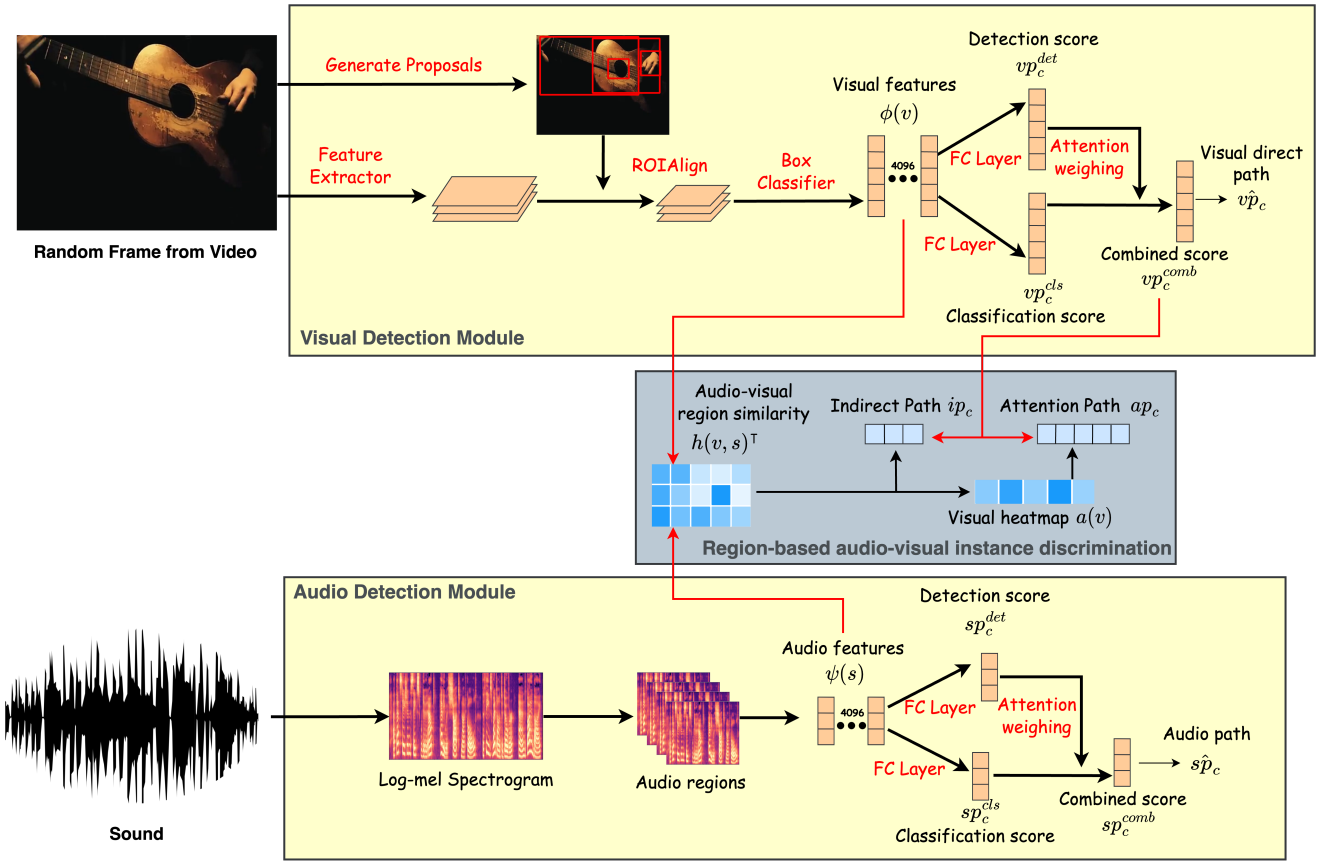

Figure 2. We propose a region-based audio-visual instance discrimination module whose resulting similarity (middle) is computed from both visual (top) and audio (bottom) modules, and combined with the visual module to use sound as auxilary to improve visual detection module performance.

generated by Edge Boxes [45], resulting in fixed-sized convolutional feature maps. Finally, a box feature extractor is applied to extract a fixed-length feature for each visual region. We use v_i where $i \in \{1, ..., M\}$ to denote the visual regions $\mathbf{v}$ of a given frame. This process results in visual region feature vectors $\phi(v_i) \in \mathbb{R}^d$ ($d = 4096$).

Because no region-level labels are available, during training we optimize visual predictions of image-level labels $\hat{vp}_c$ where $c \in \{1, ..., C\}$ and C is the number of classes. The proposal features $\phi(v_i)$ are fed into two parallel fully-connected layers to compute the visual detection scores $v_{i,c}^{det} \in \mathbb{R}^1$ and classification scores $v_{i,c}^{cls} \in \mathbb{R}^1$:

$$v_{i,c}^{det} = w_c^{det\intercal}\phi_i(v) + b_c^{det}, \quad v_{i,c}^{cls} = w_c^{cls\intercal}\phi_i(v) + b_c^{cls} \quad (1)$$

These classification and detection scores are converted to probabilities such that $vp_{i,c}^{cls}$ is the probability that class c is in present proposal v_i, and $vp_{i,c}^{det}$ is the probability that v_i is important for predicting image-level label y_c. Element-wise multiplication of classification and detection score probabilities $vp_{i,c}^{comb}$ is used to compute the loss, and in inference to compute mAP results.

$$vp_{i,c}^{det} = \frac{exp(v_{i,c}^{det})}{\sum_{k=1}^{M} exp(v_{k,c}^{det})}, \quad vp_{i,c}^{cls} = \frac{exp(v_{i,c}^{cls})}{\sum_{k=1}^{C} exp(v_{i,k}^{cls})} \quad (2)$$

Finally, visual aggregated image-level predictions $\hat{vp}_c$ are computed as follows, where greater values of $\hat{vp}_c \in [0, 1]$ mean higher likelihood that c is present in the image.

$$vp_{i,c}^{comb} = vp_{i,c}^{det} vp_{i,c}^{cls}, \quad \hat{vp}_c = \sigma\left(\sum_{i=1}^{M} vp_{i,c}^{comb}\right) \quad (3)$$

Assuming the label $y_c = 1$ if and only if class c is present, the visual classification loss used for training the model is defined as follows. Again, since no region-level labels are provided, we must derive region-level scores indirectly, by optimizing this loss.

$$\mathcal{L}_{\mathbf{v}} = -\sum_{c=1}^{C} [y_c \log \hat{vp}_c + (1 - y_c) \log(1 - \hat{vp}_c)] \quad (4)$$

3.2. Audio detection module

We extract audio features for each region from log-mel spectrograms. Let s_j where $j \in \{1, ..., N\}$ denote the audio regions $\mathbf{s}$ in a video, and N is the number of audio regions which is variable and dependent on the duration of the audio. We split the audio into regions for each second of the audio. Because each audio region contains information of different intensities, we believe that using audio regions rather than using spectrogram as a whole improves performance in audio-visual instance discrimination. This process results in audio region feature vectors $\psi(s_j) \in \mathbb{R}^d$ ($d = 4096$).

The aggregated video-level sound prediction $\hat{sp}_c$ is computed similarly to the visual detection module, and the audio classification loss is defined as:

$$\mathcal{L}_{\mathbf{s}} = -\sum_{c=1}^{C} [y_c \log \hat{sp}_c + (1 - y_c) \log(1 - \hat{sp}_c)] \quad (5)$$

3.3. Region-based audio-visual instance discrimination

Our region-based audio-visual instance discrimination module is trained using a contrastive learning framework where audio representations are contrasted with those of negative video representations or vice versa, inspired by [27]. The purpose of our method is to learn many-to-many relations between visual region features $\phi(v_i)$ and audio region features $\psi(s_j)$. In other words, our method learns which visual region is related to which sound region and to what extent.

The visual region features $\phi(v_i)$ and the audio region features $\psi(s_j)$ share the same d-dimensional embedding space so they can be contrasted. We further L2-normalize the $\phi(v_i)$ and $\psi(s_j)$ vectors. The cosine similarity of these feature vectors is computed to obtain a similarity of audio and visual regions, with the expectation that the visual region showing an object is correlated with the audio region having the sound of the corresponding object. The similarity is given by:

$$\mathrm{h}(\mathrm{v}_i, s_j) = \langle \phi(v_i), \psi(s_j) \rangle / \rho, \quad i \in \{1, ..., M\}, j \in \{1, ..., N\} \quad (6)$$

where ρ, is a learnable temperature parameter.

We next compute an aggregated visual similarity $a(v_i) \in \mathcal{R}^1$ that is part of our attention path. This visual similarity indicates the relation of each visual region with the corresponding audio set $\mathbf{s}$ in the video clip. Our attention path and the most strongly attended visual region, based on the audio set $\mathbf{s}$, can be computed as:

$$a(v_i) = \sum_{j=1}^{N} h(v_i, s_j), \quad S(\mathbf{v}, \mathbf{s}) = \max_i a(v_i) \quad (7)$$

We use noise contrastive estimation (NCE) [19] to define the contrastive learning by considering image and audio pairs $(\mathbf{v}, \mathbf{s}) \in \mathcal{B}$ where $\mathcal{B}$ is a image-audio pair batch. Pairs are defined as a randomly sampled frame from a video, and the audio channel of that video. The first component of the NCE loss contrasts an image with negative audio samples to measure how closely the image matches with its audio pair among the others in the batch:

$$\mathrm{L}_{\mathbf{s}\to\mathbf{v}} = -\frac{1}{|\mathcal{B}|} \sum_{(\mathbf{v},\mathbf{s})\in\mathcal{B}} log \frac{exp(S(\mathbf{v}, \mathbf{s}))}{exp(S(\mathbf{v}, \mathbf{s})) + \sum_{(\mathbf{v}',\mathbf{s}')\in\mathcal{B}} exp(S(\mathbf{v}, \mathbf{s}'))} \quad (8)$$

The second component of the NCE loss contrasts an audio with negative image samples to measure how closely the audio matches with its image pair among the others in the batch:

$$\mathrm{L}_{\mathbf{v}\to\mathbf{s}} = -\frac{1}{|\mathcal{B}|} \sum_{(\mathbf{v},\mathbf{s})\in\mathcal{B}} log \frac{exp(S(\mathbf{v}, \mathbf{s}))}{exp(S(\mathbf{v}, \mathbf{s})) + \sum_{(\mathbf{v}',\mathbf{s}')\in\mathcal{B}} exp(S(\mathbf{v}', \mathbf{s}))} \quad (9)$$

These two components are summed to obtain the NCE loss:

$$\mathcal{L}_{NCE} = \mathcal{L}_{\mathbf{s}\to\mathbf{v}} + \mathcal{L}_{\mathbf{v}\to\mathbf{s}} \quad (10)$$

We jointly optimize our framework with the three defined losses, and the final loss is given by:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{NCE} + \lambda_2 \mathcal{L}_{\mathbf{v}} + \lambda_3 \mathcal{L}_{\mathbf{s}} \quad (11)$$

where λ_1, λ_2, and λ_3 are weighting hyperparameters.

3.4. Sound as Indirect Path

We next describe how to use sound as a helper modality to provide confidence for or to adjust the visual-only predictions. We define an indirect path to link between visual frames and audio to the predict object label in a given frame. We use the region-based audio-visual instance discrimination module to make this association. The similarity of audio and visual regions $h(\mathbf{v}, \mathbf{s})$, and the combined classification and detection score probabilities from the visual detection module vp^{comb}, are matrix-multiplied. Finally, the aggregated image-level indirect path prediction is computed as follows:

$$ip_{j,c} = h(\mathbf{v}, s_j)^\intercal \, vp_c^{comb}, \quad \hat{ip}_c = \sigma\left(\sum_{j=1}^{N} ip_{j,c}\right) \quad (12)$$

where $h(\mathbf{v}, s_j)^\intercal \in \mathcal{R}^{1\times M}$ ($\intercal$ is transpose) and $vp_c^{comb} = [vp_{1,c}^{comb}, \ldots, vp_{M,c}^{comb}] \in \mathcal{R}^{M\times 1}$. The greater values of $\hat{ip}_c \in [0, 1]$ mean higher likelihood that class c is present in the image. This means that a class c will have a strong prediction probability if a visual region strongly indicates

it, but that visual region is strongly related to a sound region. In other words, we only make confident predictions when there is evidence of some object producing a sound. Furthermore, $ip_c = [ip_{1,c}, \ldots, ip_{N,c}] \in \mathcal{R}^{N\times 1}$ represents the indirect path in Fig. 2. The indirect path is only used for classification because there are no scores computed at the visual region level.

3.5. Sound as Attention Path

As our second contribution, we define an attention path from the sound modality, which indicates the importance of visual regions. This auxiliary path helps improve both classification and detection performance. The visual similarity $a(v_i)$ (Eq. 7) is used to weight the visual region scores $vp_{i,c}^{comb}$ by performing element-wise multiplication. Thus, corresponding visual regions that include more information about the object gain priority. The attention path $ap_{i,c}$ for detection and the aggregated attention path for classification are computed as follows:

$$ap_{i,c} = a(v_i)\ vp_{i,c}^{comb}, \quad \hat{ap}_c = \sigma\left(\sum_{i=1}^{M} ap_{i,c}\right) \tag{13}$$

where $\hat{ap}_c \in [0,1]$. Furthermore, $ap_c = [ap_{1,c}, \ldots, ap_{M,c}] \in \mathcal{R}^{Mx1}$ represents the attention path in Fig. 2 and is also used in the combination path (Sec. 3.6). The attention path is used for both detection and classification, since both visual-region ($ap_{i,c}$) and frame scores ($\hat{ap}_c$) are available.

3.6. Sound as the Combination of Paths

The indirect and attention paths are different paths that provide complementary cues to the visual detector from audio, to help combat noise. We combine these paths to benefit from both cues. We input the attention path ap_c into the indirect path computation to make them cooperate with the similarity of audio and visual regions $h(\mathbf{v}, s)$ as follows:

$$\hat{cp}_c = \sigma\left(\sum_{j=1}^{N} h(\mathbf{v}, s_j)^\intercal\ ap_c\right) \tag{14}$$

where $\hat{cp}_c \in [0,1]$. The combination of paths is only used for classification.

3.7. Sound to Update Training Set Labels

Deep neural networks, rather than simply memorizing noise, can generalize after training on noisy data. Moreover, we expect that noisy labels have impact on visual and audio-visual predictions in a different way. Based on this information, as our final contribution, we define new ground-truth (GT) training set labels using the predictions of sound as the combination of paths model. This method uses the generalization ability of neural networks and varied effect of noise on different modalities.

We first train the audio-visual model with noisy labels, then we use the model (Sec. 3.6) to make prediction on training set. If the model prediction $\hat{cp}_c$ is different from the noisy label and $\hat{cp}_c > 0.7$ (to be sure that it is a strong enough prediction to clean noise), we change the noisy label with the prediction as the new GT label:

$$y_k = 1(k = \text{argmax}\ \hat{cp}_c) * 1(\hat{cp}_k > 0.7) \tag{15}$$

where $1(\cdot)$ denotes the indicator function. We follow this procedure for the whole training set and generate new GT labels. Then we retrain all model variations with the new labels.

4. Experiments

We evaluate the components of our method on classification and detection tasks. We test the following methods:

- The visual-only direct path (VISUAL-ONLY, Sec. 3.1) and audio-only direct path (SOUND-ONLY, Sec. 3.2);
- The same paths but trained with audio-visual instance discrimination and contrastive learning (Sec. 3.3), resulting in VISUAL-ONLY-CONT. and SOUND-ONLY-CONT.;
- Our method contributions: sound as indirect path (SOUND-INDIRECT, Sec. 3.4), sound as attention path (SOUND-ATTENTION, Sec. 3.5) and sound as the combination of paths (SOUND-COMBINATION, Sec. 3.6).

We evaluated all methods under three different settings: Clean (where labels are expected to be clean at the video level, but some noise still persists in that not all frames in a video exhibit the objects mentioned in the label set), Noisy (where we flip 20% of the video labels to obtain a more realistic scenario), and New GT. This last setting uses the labels obtained using our SOUND-UPDATE method (Sec. 3.7).We compare our detection performance with state-of-art detectors Afouras [1] and PCL [35] in Table 2. Furthermore, we assess the performance of audio-visual instance discrimination method with state-of-art sound localization papers [44, 4, 32, 20, 21, 1] in Table 4.

4.1. Experimental Setup

4.1.1 Data

AudioSet [18] is a large audio-visual dataset consisting of 10-second videos from YouTube. During training, we use the "unbalanced" split of AudioSet-Instruments [4] used by [21] spanning 15 instrument classes. We use the "balanced" split of AudioSet-Instruments for evaluation on the annotations provided by [21]. The full AudioSet-Instruments is used for class-agnostic single object localization in Table 4.
VGGSound [9] is an audio-visual correspondent dataset consisting of 10-second clips, extracted from videos up-

Method	Clean			Noisy			New GT (SOUND-UPDATE)		
	mAP_{30}	mAP_{50}	$mAP_{[50:95:5]}$	mAP_{30}	mAP_{50}	$mAP_{[50:95:5]}$	mAP_{30}	mAP_{50}	$mAP_{[50:95:5]}$
VISUAL-ONLY	48.3	26.7	10.1	30.2	14.8	4.2	41.2	21.7	6.7
VISUAL-ONLY-CONT.	52.9	30.4	11.3	31.8	15.5	4.6	43.5	23.9	7.9
SOUND-ATTENTION	**53.3**	**30.8**	**11.6**	**33.1**	**16.0**	**4.9**	**44.3**	**25.6**	**8.6**
VISUAL-ONLY	36.3	17.6	5.8	27.1	13.9	3.8	30.3	15.3	4.8
VISUAL-ONLY-CONT.	38.9	20.0	6.4	30.2	14.3	4.9	33.1	16.7	5.3
SOUND-ATTENTION	**41.8**	**21.4**	**7.0**	**32.8**	**15.1**	**5.2**	**36.4**	**18.1**	**5.7**

Table 1. mAP (%) results of visual methods with clean, noisy, and new GT labels on AudioSet (top) and VGGSound (bottom). The best performer per column is in **bold**.

Method	mAP_{30}	mAP_{50}	$mAP_{[50:95:5]}$
PCL [35]	39.0	17.5	4.4
AFOURAS - SELF SUP. [1]	44.3	28.0	9.6
AFOURAS - WEAK SUP. [1]	50.6	**30.9**	10.3
SOUND-ATTENTION (OURS)	**53.3**	30.8	**11.6**

Table 2. Comparison to detection methods on AudioSet. Baselines' numbers taken from [1]. The best performer per column is in **bold**.

Method	AudioSet			VGGSound		
	Clean	Noisy	New GT	Clean	Noisy	New GT
VISUAL-ONLY	60.1	58.7	59.8	79.5	77.0	82.7
VISUAL-ONLY-CONT.	62.2	59.4	60.4	82.9	77.8	81.5
SOUND-INDIRECT	62.9	59.6	60.7	85.4	78.4	83.3
SOUND-ATTENTION	63.1	59.6	60.8	85.5	78.5	83.5
SOUND-COMBINATION	**63.5**	**59.9**	**61.1**	**86.3**	**79.4**	**84.3**
SOUND-ONLY	71.7	69.8	70.4	76.7	74.3	75.9
SOUND-ONLY-CONT.	**73.2**	**71.1**	**71.9**	**78.6**	**75.8**	**76.5**

Table 3. Accuracy (%) of visual (top five lines) and audio (bottom two lines) methods with clean, noisy, and new GT labels on AudioSet and VGGSound. The best performer per column is in **bold**, and all of our proposed methods that outperform the VISUAL-ONLY-CONT. are underlined.

loaded to YouTube. We chose a subset that includes 10k training and 2k test videos with 13 classes (guitar, car, dog, train, violin, keyboard, motorboat, drum, airplane, helicopter, trombone, motorcycle and saxophone). We manually annotated 800 test frames for ground truth boxes. We produce our mAP results according to these annotations.

The **MUSIC** dataset [44] contains 685 videos, including 536 solo and 149 duet. There are 11 different types of musical instruments. We use the first five/two videos in each instrument category in solo/duet during test. The remaining videos are used for training. We use this dataset to only compare our sound localization module with other papers and we use the annotations provided by [21]. While we perform class-agnostic localization on MUSIC-solo, class-aware localization is performed on MUSIC-dual following the prior literature.

Noisy Labels Our motivation to generate new GT labels is similar to [41] that reduces noise to improve detection, but our novelty is to benefit from sound. [41] uses the COCO dataset [23] that includes noisy captions. However, it does not contain sound and we could not find any dataset that includes sounding objects and noisy supervision. Thus, we choose commonly used audio-visual datasets, AudioSet and VGGSound, and artificially create noise to mimic the natural noise. To create noise, we randomly change 20 percent of the labels. We observe new GT labels reduce the noise to 4 percent for AudioSet and 5 percent for VGGSound.

4.1.2 Implementation Details

Before training the visual detector, we extract at most 1000 proposals using Edge Boxes [45], commonly used in weakly-supervised detection [6, 35], from OpenCV [7]. We use the indirect path, attention path and the combination of paths only during inference. They are not part of the training, but the audio-visual similarity that they rely on is learned in training. Further implementation details can be found in the supplementary file.

4.2. Sound as auxiliary modality

Our proposed SOUND-INDIRECT and SOUND-ATTENTION methods that benefit from complementary audio cues outperforms the VISUAL-ONLY and VISUAL-ONLY-CONT. methods that only use the visual signals for classification (Table 3) in all noise and dataset settings. Note that VISUAL-ONLY-CONT. also benefits from sound (through contrastive learning), but only during representation learning, while our proposed methods directly affect object prediction results and thus outperform VISUAL-ONLY-CONT.

SOUND-INDIRECT and SOUND-ATTENTION link between audio and visual in a different way providing distinct complementary cues. Thus, we observe that SOUND-COMBINATION outperforms indirect and attention paths and reaches the best results in all settings (Table 3).

Note that we use SOUND-INDIRECT, SOUND-ATTENTION and SOUND-COMBINATION paths only in inference, which means they are obtained from the same trained network as VISUAL-ONLY-CONT., and there is no randomness effect of training in our results.

In the edge case that there is only audio signal, only the audio recognition can be performed, as in SOUND-ONLY and SOUND-ONLY-CONT (Table 3).

Importantly, we use SOUND-ATTENTION for detection (note SOUND-INDIRECT cannot be evaluated in this setting), and it considerably outperforms the VISUAL-ONLY and VISUAL-ONLY-CONT. results in all noise, dataset and mAP settings in Table 1. This shows that using audio cues improves the performance in object detection.

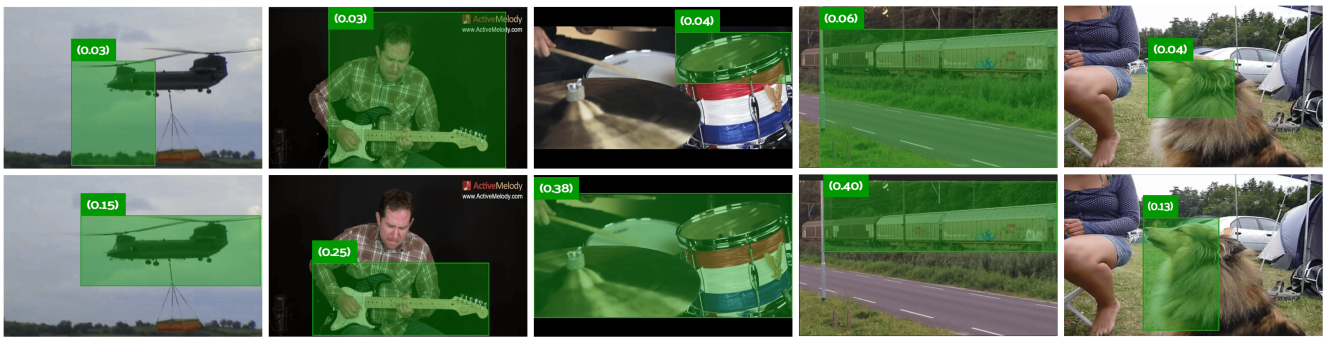

Figure 3. Qualitative comparison of VISUAL-ONLY (top) and our proposed SOUND-ATTENTION (bottom) with clean labels on VGGSound. We show boxes with the highest confidence for each image. The ground-truth objects are helicopter, guitar, drum, train, dog in this order.

Method	MUSIC-solo		MUSIC-dual		AudioSet	
	IoU@0.5	AUC	CIoU@0.3	AUC	IoU@0.5	AUC
SOUND-OF-PIXELS [44]	40.5	43.3	16.8	16.8	38.2	40.6
OBJECT-THAT-SOUND [4]	26.1	35.8	13.2	18.3	32.7	39.5
ATTENTION [32]	37.2	38.7	21.5	19.4	36.5	39.5
DMC [20]	29.1	38.0	17.3	21.1	32.8	38.2
DSOL [21]	**51.4**	43.6	30.2	22.1	38.9	40.9
AFOURAS [1]	-	-	-	-	**50.6**	**47.5**
OURS	50.8	**46.2**	**41.1**	**26.0**	43.4	40.3

Table 4. Comparison to sound localization methods on the MUSIC-solo, MUSIC-dual and AudioSet datasets. The best performer per column is in **bold**.

4.3. Comparison of clean and noisy environments, contribution of SOUND-UPDATE

We experiment with our methods in three different noise settings which are clean, noisy, and new GT. We use noisy labels to obtain new GT labels with the help of SOUND-COMBINATION, resulting in SOUND-UPDATE. We observe the results in the new GT settings are superior to the results in the noisy setting showing that using audio to help clean up the label set and improve visual predictions (SOUND-UPDATE), is quite effective.

4.4. Qualitative analysis

We visualize the object detection performance of VISUAL-ONLY and SOUND-ATTENTION in Fig. 3. SOUND-ATTENTION is more successful in detecting different objects than VISUAL-ONLY that detects only some part of the object including unrelated regions for the examples. Moreover, the detected boxes of SOUND-ATTENTION have higher confidence scores (between 0.13 and 0.40) in the examples than the detected boxes of VISUAL-ONLY (between 0.03 and 0.06) The confidence scores are shown in the upper-left of the boxes in the examples. The confidence score is defined as $vp_{i,c}^{comb}$ in Eq. 3.

4.5. Comparison to detection methods

The main comparison is presented in Table 2 because we propose object detection as the main task. Our SOUND-ATTENTION method clearly outperforms PCL [35] and AFOURAS - SELF SUPERVISED [1] on AudioSet in Table 2. Even though AFOURAS - WEAK SUPERVISED [1] performs slightly better than our method using the mAP@50 metric, SOUND-ATTENTION method clearly outperforms it in the more relaxed and stricter mAP metrics, and we reach state-of-art on AudioSet.

4.6. Comparison to sound localization methods

The comparison in Table 4 is a supportive that shows although the localization module aims to assist the detection module, it also performs competitively on an individual basis, compared to state-of-the-art methods on the MUSIC and AudioSet datasets. Sound localization is a task in which the sound modality is essential in inference. Only target objects that produce sound should be localized. Thus, we use audio-visual region similarity (Eq. 6) rather than proposed visual detection module to produce localization results. We use the union of bounding boxes having audio-visual similarity more than a threshold to obtain a heatmap following the sound localization literature. We use class predictions for each visual region in the detection module to perform class-aware localization in the MUSIC-dual dataset. Our method outperforms most localization methods.

5. Conclusion

We have demonstrated how sound can help to tackle noise in weakly-supervised object detection. Our method created new GT labels to reduce noise in supervision. SOUND-UPDATE successfully handled the noise and improved classification and detection results. The indirect path (SOUND-INDIRECT) provides additional evidence through audio as an intermediate connection. Furthermore, we presented the attention path (SOUND-ATTENTION) that determines relevant visual regions based on sound.

Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant No. 2046853. CG was also supported by a University of Pittsburgh Intelligent Systems Program fellowship.

References

[1] Triantafyllos Afouras, Yuki M Asano, Francois Fagan, Andrea Vedaldi, and Florian Metze. Self-supervised object detection from audio-visual correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10575–10586, 2022.

[2] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 208–224. Springer, 2020.

[3] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:25–37, 2020.

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018.

[5] Gedas Bertasius and Lorenzo Torresani. Cobe: Contextualized object embeddings from narrated instructional video. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:15133–15145, 2020.

[6] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2846–2854, 2016.

[7] Gary Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000.

[8] 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 (CVPR)*, pages 16867–16876, 2021.

[9] 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.

[10] Kai Chen, Hang Song, Chen Change Loy, and Dahua Lin. Discover and learn new objects from documentaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] Yanbei Chen, Yongqin Xian, A Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7016–7025, 2021.

[12] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*, pages 104–120. Springer, 2020.

[13] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11162–11173, 2021.

[14] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. In *European Conference on Computer Vision (ECCV)*, 2022.

[15] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018.

[16] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3879–3888, 2019.

[17] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, 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.

[18] 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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[19] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[20] 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 (CVPR)*, pages 9248–9257, 2019.

[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. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:10077–10087, 2020.

[22] Di Hu, Yake Wei, Rui Qian, Weiyao Lin, Ruihua Song, and Ji-Rong Wen. Class-aware sounding objects localization via audiovisual correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[24] Xian Liu, Rui Qian, Hang Zhou, Di Hu, Weiyao Lin, Ziwei Liu, Bolei Zhou, and Xiaowei Zhou. Visual sound localization in the wild by cross-modal interference erasing. In *AAAI Conference on Artificial Intelligence*, 2022.

[25] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 32:13–23, 2019.

[26] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[27] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021.

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

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 28, 2015.

[31] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[32] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4358–4366, 2018.

[33] Yunhang Shen, Rongrong Ji, Zhiwei Chen, Yongjian Wu, and Feiyue Huang. Uwsod: Toward fully-supervised-level capacity weakly supervised object detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:7005–7019, 2020.

[34] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5103–5114, 2019.

[35] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(1):176–191, 2018.

[36] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[37] 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.

[38] Mesut Erhan Unal, Keren Ye, Mingda Zhang, Christopher Thomas, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Learning to overcome noise in weak caption supervision for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022.

[39] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] Ke Yang, Dongsheng Li, and Yong Dou. Towards precise end-to-end weakly supervised object detection network. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[41] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 9686–9695, 2019.

[42] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14393–14402, 2021.

[43] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[44] 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.

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