Self-supervised object detection from audio-visual correspondence

Afouras, T.; Asano, Y.M.; Fagan, F.; Vedaldi, A.; Metze, F.

DOI
10.48550/arXiv.2104.06401
10.1109/CVPR52688.2022.01032

Publication date
2022

Document Version
Author accepted manuscript

Published in
2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition

Self-supervised object detection from audio-visual correspondence

Triantafyllos Afouras[1*†] Yuki M. Asano[2*] Francois Fagan[3] Andrea Vedaldi[3] Florian Metze[3]

[1] University of Oxford [2] University of Amsterdam [3] Meta AI

afourast@robots.ox.ac.uk

Abstract

We tackle the problem of learning object detectors without supervision. Differently from weakly-supervised object detection, we do not assume image-level class labels. Instead, we extract a supervisory signal from audio-visual data, using the audio component to "teach" the object detector. While this problem is related to sound source localisation, it is considerably harder because the detector must classify the objects by type, enumerate each instance of the object, and do so even when the object is silent. We tackle this problem by first designing a self-supervised framework with a contrastive objective that jointly learns to classify and localise objects. Then, without using any supervision, we simply use these self-supervised labels and boxes to train an image-based object detector. With this, we outperform previous unsupervised and weakly-supervised detectors for the task of object detection and sound source localization. We also show that we can align this detector to ground-truth classes with as little as one label per pseudo-class, and show how our method can learn to detect generic objects that go beyond instruments, such as airplanes and cats.

1. Introduction

While recent progress in learning image and video representations has been substantial [22, 33, 38, 93], this has not yet translated into an ability to learn interpretable and actionable concepts automatically. By that, we mean that some manual labels are still required in order to map unsupervised representations to useful concepts such as image classes or object detections. In this paper, we thus consider the problem of learning interpretable concepts without any manual supervision. In particular, we focus on a problem that has not been explored extensively in the literature: learning to simultaneously detect and classify objects with no manual labels at all.

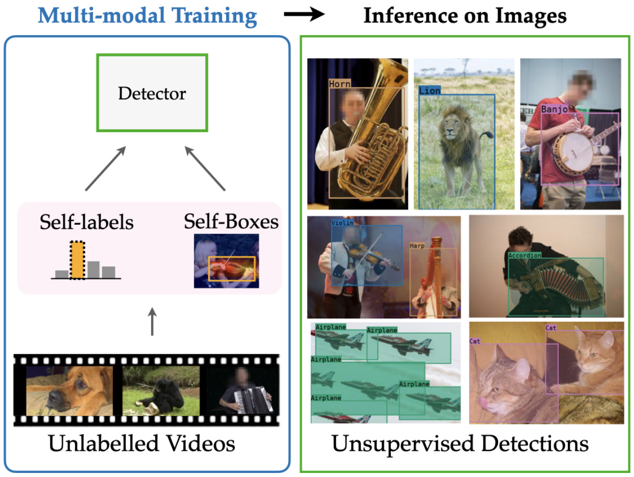

Figure 1. **We train an object detector simply by watching videos.** Without using any manual annotations, we learn to detect different objects in images, by first self-labelling boxes and object categories and then using those as targets to teach a detector. The detection results shown are outputs from our trained model; for visualisation purposes we show Hungarian-matched labels.

This problem is related to weakly supervised object detection (WSOD [16, 61]), with the difference that, in WSOD, the learning algorithm is given image-level labels telling it whether the image contains an occurrence of a given object type or not. Inspired by recent work in self-supervised learning, we seek to replace this source of external supervision with an internal supervisory signal extracted from the observation of video data. Videos are far richer than images, for example because they contain motion. Here, we focus on the multi-modal aspect of videos and use sound as a weak and noisy cue to learn about objects in the visual component of the data.

The power of multi-modal self-supervision has been demonstrated before in self-supervised representation learning, and, closely related, in *video clustering* [9]. However, while video clustering can provide an interpretation of the data in terms of discrete classes, it does not provide any information about the location of the relevant ob-

[*] Joint first authors.
[†] Work done during an internship at FAIR.

jects in images. On the other hand, *sound source localisation* [7, 11, 49, 65] has considered precisely the problem of localizing the source of sounds in images. It is therefore tempting to trivially combine image classification and sound source localisation in the hope of learning the type and location of objects automatically.

Unfortunately, such an approach does *not* lead to a satisfactory object detector. To understand why, it is important to note that the goal of sound source localisation is to *localize the sound while it is being heard*. This is insufficient for a detector because many objects emit sounds only occasionally and they become invisible to source localisation when they are silent. Instead, a detector that works in the visual domain should be responsive even when the object cannot be heard. Furthermore, source localisation methods generally only extract a heatmap giving the distribution of possible object locations; in contrast, a detector solves the much harder problem of enumerating all individual object instance that occur in an image by outputting instance-specific bounding boxes.

In order to solve these issues, we should treat the sound component as a useful cue to *learn* an object detector, but not as a cue which is *necessary for detection*. Instead, we consider the problem of taking as input a collection of raw videos and producing a list of object classes and locations, in order to train an image-based detector.

On a high level, our method is based on the following observation: we can use a sound source localisation network to learn about possible locations of sounding objects in videos. From this, we can extract a collection of bounding box pseudo-annotations for the objects and use those to learn a standard object detector. Because the latter only uses the visual modality, it immediately transfers to the detection of objects even when no relevant sound is present.

However, one challenge is that sound source localisation does not provide the necessary class information to train class-specific detectors, effectively resulting in only learning a region proposal network for generic objects, with high rates of false positives. To this end, we note that most sound source localizers are based on noise-contrastive formulations that, together with clustering-based approaches, are currently state-of-the-art in self-supervised representation learning. From this, we derive a joint formulation that can simultaneously benefit from and learn to localize sound sources and classify them without *any* supervision. The resulting output can then be used to train any off-the-shelf object detector such as a Faster-RCNN [72] to learn an object detector without any supervision, as shown in Fig. 1.

Empirically, we test our method by training and testing on VGGSound [21] and AudioSet [31], as well as testing only on a subset of OpenImages [52].

2. Related Work

Audio-Visual Sound Source Localisation. Early work in sound-source localisation includes probabilistic models for localisation [26, 39, 49] and segmentation [45], but more recently the focus has shifted to dual-stream neural networks. For example, [7, 37, 77] propose a contrastive learning approach that matches the visual and audio components of the data. The work of [40, 44] instead clusters visual and audio features, associating to them centroids by means of a contrastive loss. Other works [2, 63] learn heatmaps by exploiting audio-visual synchronization in the same video, used previously for lip-to-mouth synchronization and active-speaker detection [23, 56], or by leveraging explicit attention modules [48]. Zhao et al. [102, 103] learn to associate pixels with audio sources by training with a mix-and-separate objective. Others [69] combine activation maps learned from class labels [19, 76] with a contrastive objective, use different levels of supervision and fusion techniques [70], or improve heatmaps by mining hard negative locations [20].

The work most similar to ours is [42], who first train a source localisation model with a contrastive objective and then use the learned heatmaps to extract object representations that are clustered using K-means. The cluster assignments are then used to train classifiers on top of the audio and video encoders. The paper proceeds to use these learned representations to discriminatively localize sound sources while suppressing quiet objects in ‘cocktail party’ scenarios.

Compared to our work, none of the above can detect and thus enumerate individual object occurrences because they produce heatmaps. Furthermore, they all require audio during inference, and therefore cannot be used on individual images or to detect silent objects.

Audio-visual category discovery. Learning visual categories is usually cast as image clustering, for which there is abundant prior work, such as recent ‘deep clustering’ methods [10, 18, 46, 89, 94, 97], or clustering with segmentation [90]. However, there is less work for clustering audio-visual data. In [5] the authors extend Deep Cluster [18] to the video domain by constructing two sets of labels from opposing modalities, which are used for cross-modal representation learning. The work of [75] combines clustering with audio-visual co-segmentation achieving combined audio-visual source separation. In [9], the authors extend the self-labelling method of [10] to multi-modal data by learning a shared set of labels between the two modalities. This work builds on the latter to complement and boost sound source localisation in a joint learning framework.

Weakly Supervised Object Detection (WSOD). Weakly supervised detection uses (manual) image-level category labels without bounding box annotations. Many approaches

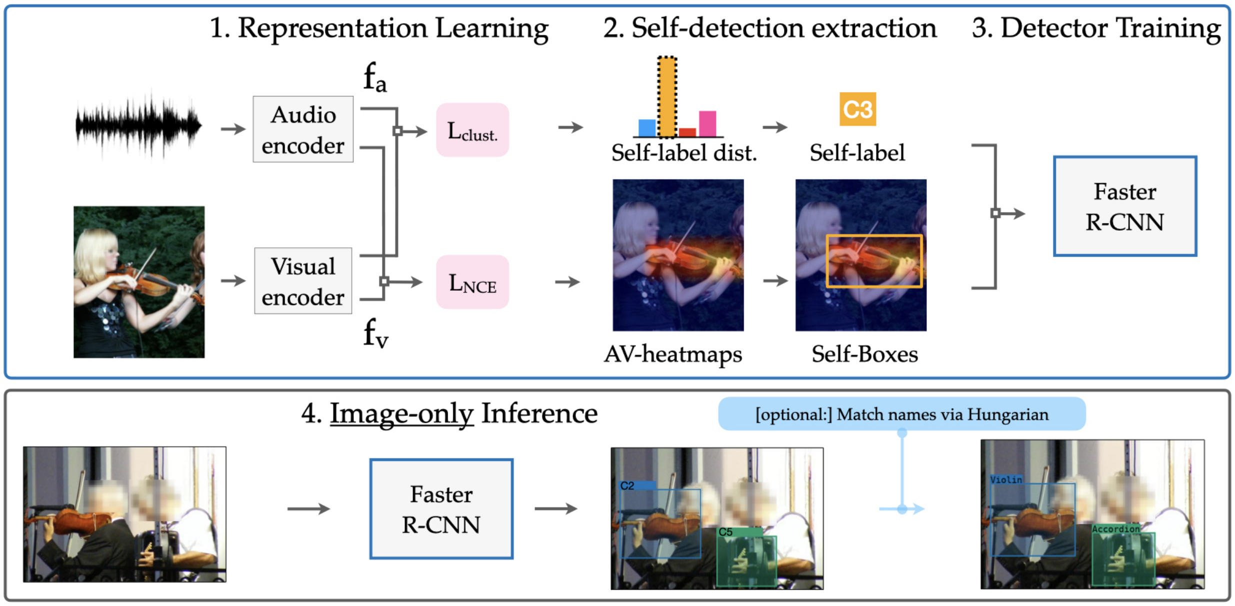

Figure 2. Self-supervised object detection from audio-visual correspondence: We combine noise-constrastive and clustering-based self-supervised learning to generate self-detections (boxes and labels) and use those as targets to train a detector. The trained detector can be used to detect objects from many categories on images without requiring audio.

are based on a form of multiple-instance learning [17, 32, 57, 85, 91, 95, 96, 98, 99, 101], or proposal clustering [84]. Recent works in the area [47, 73] combine a variety of ideas, such as self-training [106] and spatial dropout [92] or explore the use of mixed annotations [74]. Other works obtain improvements by adding curriculum learning [100], using motion cues in videos [81], adversarial training [79], combining segmentation and detection [30, 54, 78], or modelling the uncertainty of object locations [8]. Other methods use technique such as CAM or analogous techniques [15, 19, 27, 76, 80, 105] as a form of weakly supervised saliency or localisation maps. Recent works have suggested that saliency methods can also be applied to self-supervised networks [13, 34], *e.g.*for object co-localisation [13].

Self-supervised multi-modal learning. Our work is also related to methods that use multiple modalities for representation learning [5, 6, 10, 12, 59, 66, 67] and synchronization [24, 50, 64]. A number of recent papers have leveraged speech as a weak supervisory signal to train video representations [53, 58, 60, 82, 83] whereas [3] uses speech, audio and video. Some works distil knowledge learned from one modality into another [1, 4, 28, 104]. Other works incorporate optical flow and other modalities [35, 36, 68, 102] to learn representations. For instance, the work of [87] learns to temporally localize audio events through audio-visual attention. CMC [86] learns representations that are invariant to multiple views of the data such as different color channels. Multi-modal self-supervision is also used to learn sound source separation in [29], albeit they assume to have pre-trained detectors.

3. Method

Our goal is to learn object detectors using only unlabeled videos, simultaneously learning to enumerate, localize and classify objects. Our approach consists of three stages summarized in Fig. 2: first, we learn useful representations using clustering and contrastive learning; second, we extract bounding boxes and class categories by combining the trained localisation and classification networks; third, we train an off-the-shelf object detector by using these self-extracted labels and boxes as targets. Next, we explain each stage and refer the reader to the the arXiv version for further architecture and training details.

3.1. Representation Learning

Sound source spatial localisation. Our method starts by training a sound source localisation network (SSLN) using a contrastive learning formulation inspired by [7]. The SSLN is learned from pairs (v, a), where $v \in \mathbb{R}^{3\times H\times W}$ is a video frame (*i.e.*, a still image) and $a \in \mathbb{R}^{T\times F}$ is the spectrogram of the audio captured in a temporal window centered at that particular video frame.

We consider a pair of deep neural networks. The first network $f^v(v) \in \mathbb{R}^{C\times h\times w}$ extracts from the video frame a field of C-dimensional feature vectors, one per spatial location. We use the symbol $f^v_u(v) \in \mathbb{R}^C$ to denote the feature vector associated to location $u \in \psi = \{1, \dots, h\} \times$

$\{1, \ldots, w\}$. Here $h \times w$ is the resolution at which the spatial features are computed and is generally a fraction of the video frame resolution $H \times W$. The second network $f^a(a) \in \mathbb{R}^C$ extracts instead a feature vector for the audio signal.

Importantly, the spatial and audio features share the same C-dimensional embedding space and can thus be contrasted. We further assume that the vectors $f^v_u(v)$ and $f^a(a)$ are L^2 normalized (this is obtained by adding a normalization layer at the end of the corresponding networks). The cosine similarity of the two feature vectors is then used to compute a heatmap of spatial locations, with the expectation that objects that are correlated with the sounds would respond more strongly. This heatmap is given by:

$$h_u(v, a) = \langle f^v_u(v), f^v(a) \rangle / \rho, \quad u \in \psi,$$

where ρ, is a learnable temperature parameter.

For the multi-modal contrastive learning formulation [25, 62, 67], the heatmap is converted in an overall score that the video v and audio a are in correspondence. This is done by taking the maximum of the response:

$$S(v, a) = \max_{u \in \psi} h_u(v, a).$$

The contrastive learning objective is defined by considering videos $(v, a) \in \mathcal{B}$ in a batch $\mathcal{B}$. This comprises two terms. The first tests how well a video frame matches with its specific audio among the ones available in the batch:

$$\mathcal{L}_{a \to v}(\mathcal{B}) = -\frac{1}{|\mathcal{B}|} \sum_{(v,a) \in \mathcal{B}} \log \frac{\exp S(v, a)}{\sum_{(v',a') \in \mathcal{B}} \exp S(v, a')}.$$

The second is analogous, testing how well an audio matches with its specific video frame:

$$\mathcal{L}_{v \to a}(\mathcal{B}) = -\frac{1}{|\mathcal{B}|} \sum_{(v,a) \in \mathcal{B}} \log \frac{\exp S(v, a)}{\sum_{(v',a') \in \mathcal{B}} \exp S(v', a)}.$$

These two losses are averaged in the *noise-contrastive* loss:

$$\mathcal{L}_{\text{NC}}(\mathcal{B}) = (\mathcal{L}_{a \to v}(\mathcal{B}) + \mathcal{L}_{v \to a}(\mathcal{B}))/2 \tag{1}$$

Category self-labeling. Spatial localisation does not provide any class information, whereas our goal is to also associate 'names' to the different objects in the dataset. To this end, we consider the self-labelling approach of [9]. To briefly explain the formulation, let $y(v, a) \in \mathcal{Y} = \{1, \ldots, K\}$ be a label associated to the training pair (v, a). We also consider two classification networks. The first maps a video v to class scores $g^v(v) \in \mathbb{R}^K$ and is optimized by minimizing the standard cross-entropy loss:

$$\mathcal{L}_v(\mathcal{B}|y) = -\frac{1}{|\mathcal{B}|} \sum_{(v,a) \in \mathcal{B}} \log \operatorname{softmax}(y(v, a) \,|\, g^v(v)).$$

Note that this classification loss is equivalent to a contrastive loss on the cluster indices (as opposed to image indices) without normalization: As the last classification layer can be viewed as computing dot-products with each corresponding cluster's feature, it pushes the representation towards the feature of the corresponding cluster and away from the other clusters. The other network $g^a(a)$ is analogous, but uses the audio signal:

$$\mathcal{L}_a(\mathcal{B}|y) = -\frac{1}{|\mathcal{B}|} \sum_{(v,a) \in \mathcal{B}} \log \operatorname{softmax}(y(v, a) \,|\, g^a(a)).$$

As noted in [9], the crucial link between the two losses is that the labels y are shared between modalities. This is obtained by averaging the two losses:

$$\mathcal{L}_{\text{clust}}(\mathcal{B}|y) = (\mathcal{L}_v(\mathcal{B}|y) + \mathcal{L}_a(\mathcal{B}|y))/2. \tag{2}$$

Note that the labels y are unknown; following [9] these are learned in an alternate fashion with the classification networks, minimizing the same loss (2). In order to avoid degenerate solutions, the labels' marginal distribution must be specified, e.g. using a simple equipartitioning constraint:

$$\frac{1}{|\mathcal{D}|} \sum_{(v,a) \in \mathcal{D}} 1_{\{y(v,a)=k\}} = \frac{1}{K} \quad \text{for all } k = 1, \ldots, K \tag{3}$$

where $\mathcal{D}$ denotes the entire dataset (union of all batches). Optimizing y can be done efficiently by using the SK algorithm as in [9].

Joint training. To summarize, given the dataset $\mathcal{D}$, we optimize stochastically w.r.t. random batches $\mathcal{B}$ the loss:

$$\mathcal{L}(\mathcal{B}|y) = \lambda \mathcal{L}_{\text{NC}}(\mathcal{B}) + (1 - \lambda) \mathcal{L}_{\text{clust}}(\mathcal{B}|y) \tag{4}$$

where λ is a balancing hyperparameter.

The loss is optimized with respect to the localisation networks f^v and f^a and the classification networks g^v and g^a. These networks share common backbones q^v and q^a and differ only in their heads, so they can be written as $f^v = \hat{f}^v \circ q^v$, $g^v = \hat{g}^v \circ q^v$, $f^a = \hat{f}^a \circ q^a$ and $g^a = \hat{g}^a \circ q^a$.

The model is trained by alternating between updating the labels y with eq. (2) under constraint (3) and updating the networks by optimizing eq. (4).

3.2. Extraction of Self-labels for Detection

Once the localisation and classification networks have been trained, they can be used to extract self-annotations for training a detector. This is done in two steps: extracting object bounding boxes and finding their class labels.

Box extraction. To obtain "self-bounding boxes" for the objects, we use the simple heuristic suggested by [105]: the heatmap $h(v, a)$ is thresholded at a value $\epsilon(h)$, the largest connected component is identified, and a tight bounding box $t^*(v,a) \in \Omega^2$ around that component is extracted (the notation means that the box is specified by the location of the top-left and bottom-right corners).

The threshold is determined dynamically as a convex combination of the maximum and average responses of the heatmap, controlled by hyperparameter β:

$$\epsilon(h) = \beta \max_{u \in \psi} h_u + (1-\beta) \frac{1}{|\psi|} \sum_{u \in \psi} h_u. \tag{5}$$

Class labelling. As noted above, we only extract a single object from each frame for the purpose of training the detector. Likewise, we only need to extract a single class label for the frame. This is done by taking the maximum response of the visual and audio-based classification networks:

$$y^*(v,a) = \operatorname*{arg\,max}_{y \in \mathcal{Y}} [g^v_y(v) + g^a_y(a)]. \tag{6}$$

Filtering the annotations. The assumption that frames contain a dominant object introduces noise but simplifies the problem and gives us the ability to use the audio to obtain purer clusters. Notably, we do not require the method above to work for *all* frames but instead rely on our detector to smooth over the specific and noisy self-annotations to learn a holistic detection.

3.3. Training the Object Detector

The process described above results in a shortlist of training triplets $(v, t^*, y^*) \in \mathcal{D}_{\text{det}}$, where v is a video frame (an image), t^* is the extracted bounding box and y^* is its class label. We use this dataset to train an off-the-shelf detector, in particular Faster R-CNN [71] for its good compromise between speed and quality.

Recall that, given an image v, Faster R-CNN detector considers a set of bounding box proposals $m \in M(v) \subset \Omega^2$. It then trains networks $y(m) = f^{\text{cls}}_{\text{det}}(m|v) \in \{1,\dots,K,\text{bkg}\}$ and $t(m) = f^{\text{loc}}_{\text{det}}(m|v) \in \mathbb{R}^4$ inferring, respectively, the class label $y(m)$ and a refined full-resolution bounding box $t(m)$ for the box proposal m. The label space is extended to also include a *background class* bkg, which is required as most proposals do not land on any object.

The detector is trained by finding an association between proposals and annotations. To this end, if $m^* = \arg\max_{m \in M(v)} \text{IoU}(m, t^*)$ is the proposal that matches the pseudo-ground truth bounding box t^* the best, one optimizes:

$$\begin{aligned}\mathcal{L}_{\text{det}}(v, t^*, y^*) &= \mathcal{L}_{\text{reg}}(t(m^*), t^*) + \mathcal{L}_{\text{cls}}(y(m^*), y^*) \\ &+ \sum_{m \in M(v) : \text{IoU}(m,t^*) < \tau} \mathcal{L}_{\text{cls}}(y(m), \text{bkg}).\end{aligned}$$

Here $\mathcal{L}_{\text{reg}}$ is the L^1 loss for the bounding box corner coordinates and $\mathcal{L}_{\text{cls}}$ the standard cross-entropy loss. Intuitively, this loss requires the best proposal m^* to match the pseudo-ground truth class y^* and bounding box t^* of, while mapping proposal m that are a bad match ($\tau \leq 0.7$) to class bkg. Further details, including how the region-proposal network that generates the proposals is trained, are given in the the arXiv version

Discussion. Training a detector is obviously necessary to solve the problem we set out to address. However, it can also be seen as a way of extracting 'clean' information from the noisy self-annotations. Specifically: (i) the noise in individual annotations is smoothed over the entire dataset; (ii) because of the built-in NMS step, the detector still learns to extract multiple objects per image even though a single self-annotation is given for each training image; (iii) by learning to reject a large number of false bounding box proposals, the detector learns to be more precise than the self-annotations.

4. Experiments

We first introduce the datasets, experimental setup and relevant baselines; we then test our method against those, analyse it further via ablations and its capacity to generalize.

4.1. Datasets

AudioSet-Instrument. AudioSet [31] is a large scale audio-visual dataset consisting of 10-second video clips originally from YouTube. For training we use the *AudioSet-Instruments* [7] subset of the "unbalanced" split, containing 110 sound source classes as well as its more constrained subset used by [43] spanning 13 instrument classes. Following previous work we use the "balanced" subset for evaluation on the annotations provided by [43].

VGGSound. VGGSound contains over 200K 10-second clips from YouTube spanning 309 categories of objects where there is some degree of correlation between the audio and the video. We create one subset by keeping only the 50 musical instrument categories yielding around 54K training videos, and one other subset, by keeping from those only the 39 categories that can be roughly mapped to the test-set annotations (details in the arXiv version). For VGGSound pseudo-ground truth test-set annotations are obtained using a supervised detector from [29], following [43].

OpenImages. For evaluation, we also use the subset of the OpenImages [52] dataset containing musical instruments, which spans 15 classes.

4.2. Baselines

There is currently no prior work on learning an object detector for multiple object classes without any supervision. Instead, we compare against weakly-supervised detectors (hence using image-level labels) and unsupervised localisation methods that only produce heatmaps (not detections).

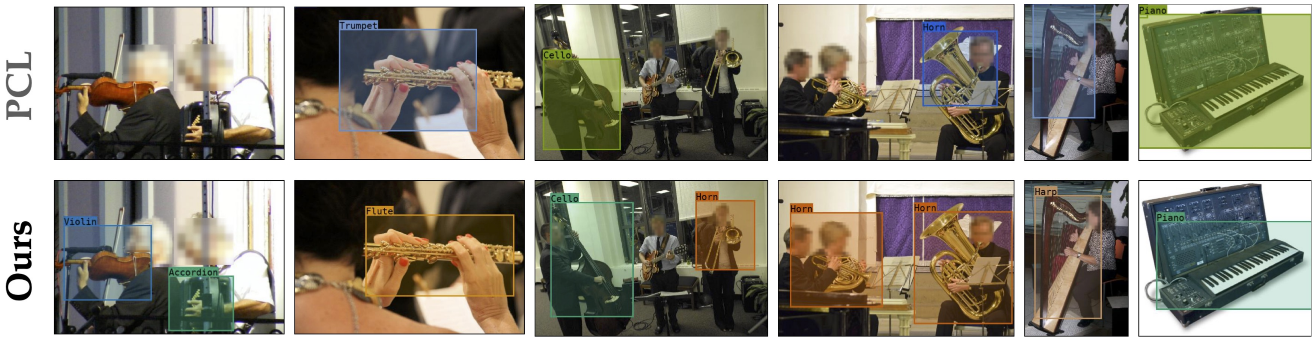

Figure 3. **Qualitative results and comparison** with a weakly supervised object detection method, PCL [84], on the OpenImages test set. Our method accurately detects objects, capturing their boundaries, even though it has been trained without *any* supervision. For visualisation purposes, we show the labels obtained from matching with the Hungarian method. More qualitative results provided in the arXiv version.

For **weakly-supervised detection**, we consider PCL [84], the strongest such baseline for which we could find an implementation.

For our second baseline, we consider **heatmap-based localisation methods** that, similarly to us, use cross-modal self-supervised learning. Here we compare against the state-of-the art DSOL method of [43] that produces a heatmap roughly localizing the objects and producing class pseudolabels.

Finally, we also compare against **other baselines** such as simply predicting a large centered box and class-agnostic region proposal methods such as selective search, and using a RPN obtained from supervised training on COCO [55]. Further details are provided in the the arXiv version.

4.3. Implementation Details.

Assessing class pseudolabels. Since class pseudolabels do not come with the name of the class (they are just cluster indices), they must be put in correspondence with human-labelled classes for evaluation. Following prior work in unsupervised image clustering [10, 14, 46, 89], we apply Hungarian matching [51] to the learned clusters and the ground truth classes. Importantly, the matching is done *after* the detector is trained and only done for assessment; meaning that the detector does not use any manual label.

Detector training. If not stated otherwise, the localizer and detector are trained on VGGSound and AudioSet whereas OpenImages are only used for evaluation. We do not have any information on the number of instruments in VGGSound and use all videos with no single/multi-object curation. For a fair comparison with DSOL, and only for the relevant experiment in Table 2, we train on AudioSet using the single-instrument subset for learning the localizer.

Number of clusters. For VGGSound training we use $K = 39$ if not stated otherwise, matching the 39 object types in the training set. Since the dataset is roughly balanced, uniform marginals are used as described in [9]. For AudioSet training we use $K = 30$ and Gaussian marginals. Further implementation details can be found in the the arXiv version

4.4. Results

Self-supervised object detection. We summarise the results of our evaluations on the three test sets that we consider in Table 1. Following the image object detection literature, we use mAP at different IoU thresholds as the evaluation metric. Our method clearly outperforms the PCL baseline even though it uses no manual annotations at all during training. PCL outperforms our approach in some of the datasets only if the IoU threshold used for mAP computation is relaxed substantially (0.3 IoU). However, for stricter thresholds our approach works better, which suggests that our detections have a relatively high spatial accuracy.

To understand the impact of the noisy class self-labels, we also train and test a detector (Ours - weak sup.) with the bounding box labels from our localisation network, but utilising the ground truth video categories. The resulting performance difference is modest, resulting in a 3% AP50 drop in VGGSound and AudioSet. This further demonstrates the accuracy of our class self-labels, but also shows the potential of our model to leverage weak supervision if available.

Per-class performance breakdown. To better understand the strengths and weaknesses of our method, we report a performance breakdown by object class in Table 3. We observe that the model obtains good results consistently for classes of large objects with a distinctive appearance (e.g. accordions and harps), while it is weaker for smaller objects such as oboes, or for objects that appear closely in numbers, like drums.

Comparison to audio-visual heatmaps. In Table 2 we compare the performance of our method trained on AudioSet to state-of-the-art sound source localisation methods. For a fair comparison to these methods, we convert the union of our predicted bounding boxes with confidence

		VGGSound			Audioset			OpenImages		
Method	No labels?↘	mAP$_{30}$	mAP$_{50}$	mAP	mAP$_{30}$	mAP$_{50}$	mAP	mAP$_{30}$	mAP$_{50}$	mAP
PCL (WSOD) [84]	✗	54.9	27.7	7.6	39.0	17.5	4.4	37.9	14.5	3.5
Ours - weak sup.	✗	67.6	42.9	14.2	50.6	30.9	10.3	48.9	33.7	9.5
Center Box*	✓	29.6	5.6	1.5	15.1	3.5	0.7	20.7	4.2	0.8
Selective Search* [88]	✓	5.2	1.1	0.4	2.8	0.4	0.1	7.4	2.1	0.7
COCO-trained RPN*	✗	33.4	7.5	1.6	19.0	4.1	0.8	24.4	11.1	2.6
Ours - self-boxes*	✓	48.1	29.6	10.0	27.8	14.1	4.8	NA	NA	NA
Ours - full	✓	**52.3**	**39.4**	**14.7**	**44.3**	**28.0**	**9.6**	**39.9**	**28.5**	**7.6**

Table 1. **Self-supervised object detection.** We report object detection metrics across three test datasets and find our method is far superior to other unsupervised approaches and outperforms even the weakly supervised baseline in most metrics. For methods denoted by *, we report class-agnostic evaluation numbers. The class-agnostic performance of the self-boxes that are used to train the detector reveals that the latter greatly outperforms them, which highlights the benefit of our approach.

	single-instr.		multi-instr.
Method	IoU-0.5	AUC	cIoU-0.3
Sound of pixels [103]	38.2	40.6	39.8
Object t. Sound [7]	32.7	39.5	27.1
Attention [77]	36.5	39.5	29.9
DMC [41]	32.8	38.2	32.0
DSOL [43]	38.9	40.9	48.7
Ours	**50.6**	**47.5**	**52.4**

Table 2. **Comparison to sound localisation methods.** Since our detector does not require audio, we obtain detections on the video frames directly. Our model outperforms the baselines. Baselines numbers taken from [42].

Dataset	**mAP$_{50}$**	Accordion	Cello	Drum	Flute	Horn	Guitar	Harp	Piano	Saxophone	Violin	Banjo	Trombone	Trumpet	Oboe
OpenImages	28.5	75.3	30.2	6.6	6.5	15.0	14.5	80.4	28.8	22.5	28.8	57.0	9.7	18.1	6.3
Audioset	28.0	41.3	44.9	0.9	5.5	21.7	39.5	82.6	52.7	2.5	17.4	46.7	8.0	-	-
VGGSound	39.4	88.6	39.4	1.8	50.0	3.4	34.9	95.6	50.2	14.4	56.3	100.0	2.2	11.0	3.8

Table 3. **Per-class mAP breakdown** For entries with '–' the test set does not contain any samples for that class.

		mAP$_{50}$	
# GT-cls.	K	VGGS	O.Images
39	20	34.4	24.4
39	30	35.1	25.1
39	39	39.4	**28.5**
39	50	**41.0**	27.5

Table 4. **Number of clusters K.** Our method is relatively robust ($<$ 5% decrease in AP) to the number of self-labelling clusters.

	mAP$_{50}$	
Matching	VGGS	O.Images
Hung.	39.4	28.5
Argmax	39.6	**30.1**
Manual	**41.0**	29.5
1-shot	36.4	25.1
10-shot	37.1	25.8

Table 5. **Matching strategies.** Even with 39 labels, our method performs accurately.

above a set threshold into a binary map, and use the latter as a pseudo-heatmap to use the same evaluation code. Our approach outperform others for both class-agnostic single object localisation and for class-aware multi-object localisation, *without* using audio signals during inference.

We note however that cIOU is not a very reliable metric for evaluating a detector (or even sound localizer) as it favours high recall over precision: by averaging this metric over all classes the most frequent ones (e.g. drums, guitars, pianos) dominate the metric. We therefore propose to the research community – and report in this paper – mean average precision (mAP) values as a more indicative metric.

Ablation: Number of clusters K. In Table 4, we perform an experiment varying the number of clusters K, and as a consequence the number of object categories that the detector learns, while keeping the test-set (containing 15 classes) fixed. We observe that out method achieves reasonable performance for a wide range of K. The performance is fairly stable when K is more than the ground truth classes, and gradually decreases when fewer clusters are used.

Data-efficient detector alignment. In Table 5 we conduct an investigation into the matching of the clusters to the ground truth labels. Recall that, for evaluation, we assign clusters to ground truth labels using the Hungarian algorithm. We can improve the computational efficiency of this step by using majority voting instead, which results in the same accuracy on VGGSound and a gain of 1.6% on OpenImages. Using a manual grouping strategy (see the arXiv version for details) yields another small boost. We can also improve the statistical efficiency as follows. While the evaluation protocol computes the optimal assignments assuming that all videos are labelled, in a real-world application we are interested in naming clusters with as little labelled data as possible. To do this, we still use majority voting, but we assume that only the top m videos per cluster (based on the strength of association) are labelled. We find that even by just using $m = 1$ (*i.e.*, a *total* number of 39 annotations), our method still achieves 37.1% and 25.3% on VGGSound and OpenImages. This is a 3% drop compared to the Hungarian algorithm and can be further reduced to 2.3% by using $m = 10$ (*i.e.*, 390 annotations).

Qualitative analysis. We show examples of successfully detected objects in challenging images in Fig. 3, where we also include the outputs of the PCL baseline. Although our model has not been manually shown any objects boundaries during training we see that it can learn very accurate boxes around them and that it can successfully identify multiple objects in complicated scenes. We provide further examples in the arXiv version.

Towards general object detection. The results presented thus far have focused on subsets of common datasets with instruments solely to ensure comparability with prior works. Since one main goal of self-supervised learning is to leverage the vast amount of unlabelled data, we wish to

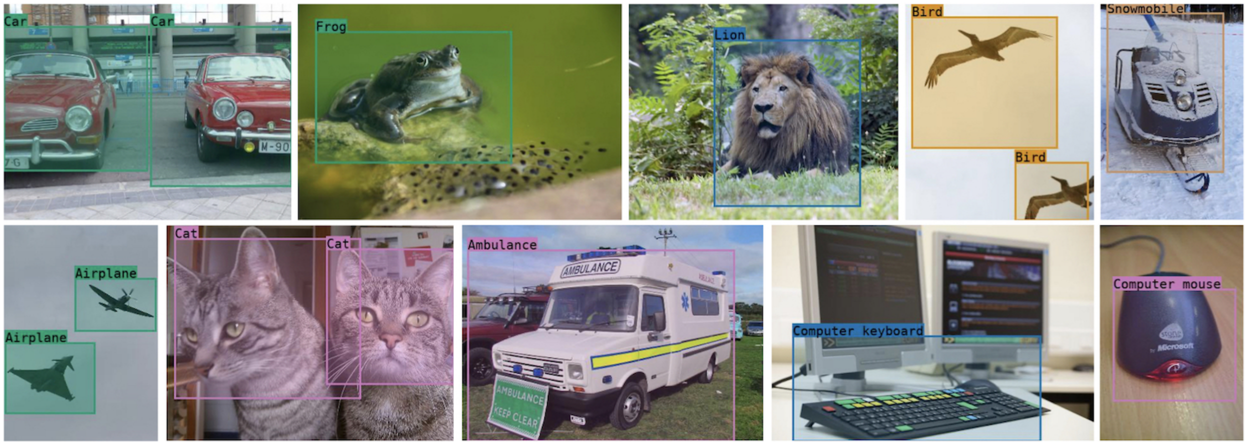

Figure 4. **Object detection beyond musical instruments**. Our proposed method can learn to accurately detect objects from more general categories, as long as they can be associated with a characteristic sound. The results shown here are from a model trained without labels directly on the full VGGSound dataset which includes 309 different video classes. Our method successfully learns to detect non-instrument objects, even in difficult multi-instance cases.

Class	AP_{30}	AP_{50}	$\text{AP}_{[50:95:5]}$
Mean	45.6	24.4	6.5
Airplane	62.7	27.0	6,5
Ambulance	56.9	30.9	7.1
Bird	26.5	15.8	3.7
Car	29.8	18.4	5.1
Cat	67.7	28.0	7.7
Comp. Keyboard.	53.3	42.6	12.9
Comp. Mouse	35.9	25.4	8.8
Frog	43.5	19.5	4.7
Lion	34.1	22.2	4.9
Snowmobile	64.3	14.3	3.5

Table 6. **Results on general object categories.**

investigate how general and robust our proposed method when applied on a far larger scale. For this, we increase our pretraining dataset by approximately 10×, simply by taking the whole of the VGGSound dataset, without any filtering. We set K to 300 and keep all training parameters the same; the result is an unsupervisedly trained object detector that can classify 300 pseudo-classes. As before, we match these to the VGGSound labels with the Hungarian algorithm and select ten categories for which we have annotations in the OpenImages dataset (details in the arXiv version).

In Fig. 4 we show qualitative results of some detections on OpenImages. The numerical results are given in Table 6. We find that even for objects that are deformable, such as cats, we get high AP_{30} values of 67.7% and that even objects that vary in shape, such as airplanes (see Fig. 4, bottom-right), we achieve a good performances 62.7%. While the results for the $\text{AP}_{50:95:5}$ metric indicate that there is still room for improvement, these initial results show that leveraging larger and more diverse video datasets for self-supervisedly learning object detectors is a promising avenue. We note that, since minimal curation is performed on the training data, and we use a large number of different object categories in a noisy dataset, this training setting is very challenging. These results further highlight the potential of our proposed method.

5. Discussion

Limitations & societal impact. We refer the reader to the arXiv version for an extensive examination of failure cases. Detection failures manifest as multiple instance grouping, missing instances or part-of-object detection, all well documented in the WSOD literature [73]. Another error is detecting wrong objects that often appear together with the objects of interest due to biases in the data (*e.g.* mouth regions as wind instruments). Wrong classification occurs often because i) visually similar classes are confused - *e.g.* Horn and Trumpet; ii) of incorrect semantic matching; iii) confusion due to the object's orientation – *e.g.* a vertical violin confused for a cello. Regarding training, there are likely further optimizations possible, such as a more end-to-end design. As for most unsupervised methods, a downside of our approach compared to supervised detection is the reduced human control on the learned concepts, which may warrant additional manual validation before deployment.

Conclusion. We have presented a method for training strong object detectors purely with self-supervision by watching unlabelled videos. We demonstrated that our best models perform better than a heatmap-based methods while not requiring and audio and better than weakly supervised baselines, even after curating the dataset to filter out noisy samples for training the latter. We have also addressed one shortcoming of using the Hungarian algorithm for evaluation by showing that data-efficient alignment of self-supervised detectors is possible with as little as one image per pseudo-label. Finally, we applied our method to domains beyond musical instruments and found that it can learn reasonable detectors in this much less curated setting, paving the way to general self-supervised object detection.

Acknowledgements

Y.M.A is thankful for MLRA funding from AWS.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. ASR is all you need: Cross-modal distillation for lip reading. In *Proc. ICASSP*, 2020. 3

[2] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Proc. ECCV*, 2020. 2

[3] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. In *Proc. NeurIPS*, 2020. 3

[4] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. In *Proc. ACMM*, 2018. 3

[5] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS*, 2020. 2, 3

[6] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proc. ICCV*, 2017. 3

[7] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proc. ECCV*, 2018. 2, 3, 5, 7

[8] Aditya Arun, C. V. Jawahar, and M. Pawan Kumar. Dissimilarity coefficient based weakly supervised object detection. In *CVPR*, 2019. 3

[9] Yuki M. Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *NeurIPS*, 2020. 1, 2, 4, 6

[10] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *Proc. ICLR*, 2020. 2, 3, 6

[11] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. SoundNet: Learning sound representations from unlabeled video. In *Proc. NeurIPS*, 2016. 2

[12] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *NeurIPS*, 2016. 3

[13] Kyungjune Baek, Minhyun Lee, and Hyunjung Shim. Psynet: Self-supervised approach to object localization using point symmetric transformation. In *Proc. AAAI*, 2020. 3

[14] Miguel A Bautista, Artsiom Sanakoyeu, Ekaterina Tikhoncheva, and Bjorn Ommer. Cliquecnn: Deep unsupervised exemplar learning. In *NeurIPS*, pages 3846–3854, 2016. 6

[15] Loris Bazzani, Alessandra Bergamo, Dragomir Anguelov, and Lorenzo Torresani. Self-taught object localization with deep networks. In *Proc. WACV*, pages 1–9. IEEE, 2016. 3

[16] Hakan Bilen, Vinay P. Namboodiri, and Luc Van Gool. Object and action classification with latent variables. In *Proc. BMVC*, 2011. 1

[17] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, 2016. 3

[18] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018. 2

[19] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proc. WACV*, pages 839–847. IEEE, 2018. 2, 3

[20] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proc. CVPR*, 2021. 2

[21] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *Proc. ICASSP*, 2020. 2

[22] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ICML*, 2020. 1

[23] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Proc. ACCV*, 2016. 2

[24] Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *ACCV Workshop on Multi-view Lip-reading*, 2016. 3

[25] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *Proc. ICASSP*, pages 3965–3969. IEEE, 2019. 4

[26] John W Fisher III, Trevor Darrell, William T Freeman, and Paul A Viola. Learning joint statistical models for audio-visual fusion and segregation. In *NeurIPS*, 2000. 2

[27] Ruth Fong and Andrea Vedaldi. Explanations for attributing deep neural network predictions. In *Proc. ICCV*, 2019. 3

[28] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proc. ICCV*, 2019. 3

[29] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proc. ICCV*, 2019. 3, 5

[30] 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 *CVPR*, 2018. 3

[31] Jort F. Gemmeke, Daniel P. W. 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 *Proc. ICASSP*, 2017. 2, 5

[32] Nicolas Gonthier, Yann Gousseau, Said Ladjal, and Olivier Bonfait. Weakly supervised object detection in artworks. *Computer Vision – ECCV 2018 Workshops*, page 692–709, 2019. 3

[33] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *NeurIPS*, 2020. 1

[34] Shir Gur, Ameen Ali, and Lior Wolf. Visualization of supervised and self-supervised neural networks via attribution guided factorization. In *arXiv:2012.02166*, 2020. 3

[35] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation

learning. In *Proc. ECCV*, 2020. 3

[36] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *Proc. NeurIPS*, 2020. 3

[37] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *Proc. ECCV*, pages 649–665, 2018. 2

[38] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *CVPR*, 2020. 1

[39] J Hershey and JR Movellan. Audio-vision: Locating sounds via audio-visual synchrony. In *NeurIPS*, volume 12, 1999. 2

[40] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proc. CVPR*, June 2019. 2

[41] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proc. CVPR*, 2019. 7

[42] 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. *arXiv.cs*, abs/2010.05466, 2020. 2, 7

[43] 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 *NeurIPS*, volume 33, pages 10077–10087, 2020. 5, 6, 7

[44] Di Hu, Zongge Wang, Haoyi Xiong, Dong Wang, Feiping Nie, and Dejing Dou. Curriculum audiovisual learning. *ArXiv*, abs/2001.09414, 2020. 2

[45] Hamid Izadinia, Imran Saleemi, and Mubarak Shah. Multimodal analysis for identification and segmentation of moving-sounding objects. *IEEE Transactions on Multimedia*, 15(2):378–390, 2012. 2

[46] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proc. ICCV*, pages 9865–9874, 2019. 2, 6

[47] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *CVPR*, 2017. 3

[48] Naji Khosravan, Shervin Ardeshir, and Rohit Puri. On attention modules for audio-visual synchronization. *arXiv preprint arXiv:1812.06071*, 1, 2018. 2

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

[50] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018. 3

[51] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6

[52] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018. 2, 5

[53] Tianhao Li and Limin Wang. Learning spatiotemporal features via video and text pair discrimination. *arXiv.cs*, abs/2001.05691, 2020. 3

[54] Xiaoyan Li, Meina Kan, Shiguang Shan, and Xilin Chen. Weakly supervised object detection with segmentation collaboration. In *Proc. CVPR*, 2019. 3

[55] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. ECCV*, 2014. 6

[56] Etienne Marcheret, Gerasimos Potamianos, Josef Vopicka, and Vaibhava Goel. Detecting audio-visual synchrony using deep neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015. 2

[57] Meng Meng, Tianzhu Zhang, Qi Tian, Yongdong Zhang, and Feng Wu. Foreground activation maps for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3385–3395, 2021. 3

[58] 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. *arXiv.cs*, abs/1912.06430, 2019. 3

[59] Pedro Morgado, Yi Li, and Nuno Vasconcelos. Learning representations from audio-visual spatial alignment. In *Proc. NeurIPS*, 2020. 3

[60] Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman. Speech2action: Cross-modal supervision for action recognition. In *Proc. CVPR*, 2020. 3

[61] Minh Hoai Nguyen, Lorenzo Torresani, Lorenzo de la Torre, and Carsten Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *Proc. ICCV*, 2009. 1

[62] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[63] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. *Proc. ECCV*, 2018. 2

[64] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proc. ECCV*, 2018. 3

[65] Andrew Owens, Phillip Isola, Josh H. McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. Visually indicated sounds. In *Proc. CVPR*, 2016. 2

[66] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Proc. ECCV*, 2016. 3

[67] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. *Proc. ICCV*, 2020. 3, 4

[68] A. J. Piergiovanni, Anelia Angelova, and Michael S. Ryoo. Evolving losses for unsupervised video representation learning. In *Proc. CVPR*, 2020. 3

[69] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu,

and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Proc. ECCV*, 2020. 2

[70] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *Proc. WACV*, March 2020. 2

[71] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *arXiv:1506.01497*, 2015. 5

[72] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proc. NeurIPS*, 2016. 2

[73] 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 *Proc. CVPR*, 2020. 3, 8

[74] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Alexander G. Schwing, and Jan Kautz. UFO ^{2}: A unified framework towards omni-supervised object detection. In *Proc. ECCV*, 2020. 3

[75] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised audio-visual co-segmentation. In *Proc. ICASSP*, pages 2357–2361. IEEE, 2019. 2

[76] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. ICCV*, pages 618–626, 2017. 2, 3

[77] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proc. CVPR*, 2018. 2, 7

[78] Y. Shen, R. Ji, Y. Wang, Y. Wu, and L. Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *Proc. CVPR*, pages 697–707, 2019. 3

[79] Y. Shen, R. Ji, S. Zhang, W. Zuo, and Y. Wang. Generative adversarial learning towards fast weakly supervised detection. In *Proc. CVPR*, pages 5764–5773, 2018. 3

[80] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proc. ICCV*, pages 3544–3553. IEEE, 2017. 3

[81] Krishna Kumar Singh and Yong Jae Lee. You reap what you sow: Using videos to generate high precision object proposals for weakly-supervised object detection. In *CVPR*, 2019. 3

[82] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *arXiv.cs*, abs/1906.05743, 2019. 3

[83] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proc. ICCV*, 2019. 3

[84] 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. *CoRR*, abs/1807.03342, 2018. 3, 6, 7

[85] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proc. CVPR*, 2017. 3

[86] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *Proc. ECCV*, 2020. 3

[87] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proc. ECCV*, 2018. 3

[88] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders. Selective search for object recognition. *IJCV*, 2013. 7

[89] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 6

[90] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. *ICCV*, 2021. 2

[91] 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 *Proc. CVPR*, 2019. 3

[92] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-Fast-RCNN: Hard positive generation via adversary for object detection. In *Proc. CVPR*, 2017. 3

[93] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proc. CVPR*, 2018. 1

[94] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proc. ICML*, pages 478–487, 2016. 2

[95] Jinheng Xie, Cheng Luo, Xiangping Zhu, Ziqi Jin, Weizeng Lu, and Linlin Shen. Online refinement of low-level feature based activation map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 132–141, 2021. 3

[96] G. Yan, B. Liu, N. Guo, X. Ye, F. Wan, H. You, and D. Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *Proc. ICCV*, pages 9833–9842, 2019. 3

[97] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *CVPR*, pages 5147–5156, 2016. 2

[98] Zhenheng Yang, Dhruv Mahajan, Deepti Ghadiyaram, Ram Nevatia, and Vignesh Ramanathan. Activity driven weakly supervised object detection. In *Proc. CVPR*, 2019. 3

[99] 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 *Proc. ICCV*, 2019. 3

[100] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag learning for weakly supervised object detection. In *CVPR*, 2018. 3

[101] Y. Zhang, Y. Bai, M. Ding, Y. Li, and B. Ghanem. W2f: A weakly-supervised to fully-supervised framework for object detection. In *Proc. CVPR*, pages 928–936, 2018. 3

[102] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. *Proc. ICCV*, 2019. 2, 3

[103] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proc. ECCV*, 2018. 2, 7

[104] M. Zhao, T. Li, M. A. Alsheikh, Y. Tian, H. Zhao, A. Tor-

ralba, and D. Katabi. Through-wall human pose estimation using radio signals. In *Proc. CVPR*, pages 7356–7365, 2018. 3

[105] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proc. CVPR*, 2016. 3, 4

[106] Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proc. ICCV*, 2019. 3