Self-supervised object detection from audio-visual correspondence

Triantafyllos Afouras1+† Yuki M. Asano2* Francois Fagan3 Andrea Vedaldi3 Florian Metze3
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.

*Joint first authors.
†Work done during an internship at FAIR.
jects in images. On the other hand, sound source localisation \cite{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 \cite{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 \cite{21} and AudioSet \cite{31}, as well as testing only on a subset of OpenImages \cite{52}.

2. Related Work

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

The work most similar to ours is \cite{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 \cite{10, 18, 46, 89, 94, 97}, or clustering with segmentation \cite{90}. However, there is less work for clustering audio-visual data. In \cite{5} the authors extend Deep Cluster \cite{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 \cite{75} combines clustering with audio-visual co-segmentation achieving combined audio-visual source separation. In \cite{9}, the authors extend the self-labelling method of \cite{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 \(f^v_\psi(v) \in \mathbb{R}^C\) to denote the feature vector associated to location \(u \in \psi = \{1, \ldots, h\} \times \)
\{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 \( f^a_u(v) \) 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:

\[
h_u(v, a) = \langle f^u_u(v), f^a(a) \rangle / \rho, \quad u \in \psi,
\]

where \( \rho \) 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 \((v, a) \in B\) in a batch \( B \). This comprises two terms. The first tests how well a video frame matches with its specific audio among the ones available in the batch:

\[
\mathcal{L}_{a \rightarrow v}(B) = \frac{1}{|B|} \sum_{(v, a) \in B} \log \frac{\exp S(v, a)}{\sum_{(v', a') \in B} \exp S(v', a')}.
\]

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

\[
\mathcal{L}_{v \rightarrow a}(B) = \frac{1}{|B|} \sum_{(v, a) \in B} \log \frac{\exp S(v, a)}{\sum_{(v', a') \in B} \exp S(v', a')}.
\]

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

\[
\mathcal{L}_{NC}(B) = \frac{1}{|B|} \sum_{(v, a) \in B} \log \frac{\exp S(v, a)}{\sum_{(v', a') \in B} \exp S(v', a')}.
\]

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:

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

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:

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

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:

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

where \( D \) 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 \( D \), we optimize stochastically w.r.t. random batches \( B \) the loss:

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

where \( \lambda \) 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 \( f^v = f^v \circ q^v, g^v = g^v \circ q^v, f^a = f^a \circ q^a \) and \( g^a = g^a \circ q^a \).

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

10578
box \( t^* (v, a) \in \Omega \) 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 \( \beta \):

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

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:

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

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 \( y(m) = f^y_{\text{det}}(m|v) \in \{1, \ldots, K, \text{bkg}\} \) and \( t(m) = f^t_{\text{det}}(m|v) \in \mathbb{R}^4 \) 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 \( \text{bkg} \), 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:

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

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 \( \text{bkg} \). 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, analyze 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 localization methods that only produce heatmaps (not detections).
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 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 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...
number of self-labelling clusters. The performance is fairly stable when \( K \) is more than the ground truth classes, and gradually decreases when fewer clusters are used.

### 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...
investigate how general and robust our proposed method when applied on a far larger scale. For this, we increase our pretraining dataset by approximately $10 \times$, 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 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 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


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