MOD-UV: Learning Mobile Object Detectors from Unlabeled Videos

Yihong Sun and Bharath Hariharan

Cornell University, Ithaca NY 14853, USA
{yihong, bharathh}@cs.cornell.edu

Abstract. Embodied agents must detect and localize objects of interest, *e.g.* traffic participants for self-driving cars. Supervision in the form of bounding boxes for this task is extremely expensive. As such, prior work has looked at unsupervised instance detection and segmentation, but in the absence of annotated boxes, it is unclear how pixels must be grouped into objects and which objects are of interest. This results in over-/under-segmentation and irrelevant objects. Inspired by human visual system and practical applications, we posit that the key missing cue for unsupervised detection is *motion*: objects of interest are typically *mobile objects* that frequently move and their motions can specify separate instances. In this paper, we propose MOD-UV, a **M**obile **O**bject **D**etector learned from **U**nlabeled **V**ideos only. We begin with instance pseudo-labels derived from motion segmentation, but introduce a novel training paradigm to progressively discover small objects and static-but-mobile objects that are missed by motion segmentation. As a result, though only learned from unlabeled videos, MOD-UV can detect and segment mobile objects from a single static image. Empirically, we achieve state-of-the-art performance in unsupervised mobile object detection on Waymo Open, nuScenes, and KITTI Datasets without using any external data or supervised models. Code is available at `github.com/YihongSun/MOD-UV`.

Keywords: Unsupervised · Mobile Object · Object Detection · Videos

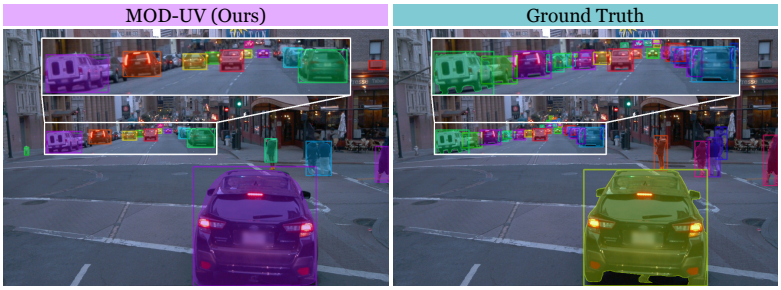

Fig. 1: Our approach, MOD-UV, learns from unlabeled videos in Waymo Open [46] only and can reliably detect and segment mobile objects from a single input image.

1 Introduction

Embodied agents such as self-driving cars must detect and localize objects of interest such as traffic participants to operate safely and effectively. Today, building such a detector requires the expensive and laborious annotation of millions of boxes over thousands of images. This process is so expensive that the largest detection dataset is orders of magnitude smaller than classification datasets and has much fewer classes. The limited set of classes further runs the risk of missing important object categories (*e.g.* snowplows for self-driving applications).

These concerns have motivated research into *unsupervised* object detection techniques that automatically discover objects from unlabeled data [6, 54, 55]. Under the hood, these techniques use self-supervised features to segment unlabeled images and produce candidate object annotations which are then used to train a detector. However, while promising, these approaches often produce many uninteresting and irrelevant "objects" in cluttered scenes (*e.g.* buildings and roads) and over- or under-segment objects of interest (*e.g.* multiple detections partitioning a large bus or a single detection grouping a row of parked cars). These failures shouldn't be surprising: after all, how can a completely unsupervised feature representation encode which assortment of windows, doors and wheels belongs together as an object, and which objects are of interest?

In this paper, we argue that a key missing cue for addressing the aforementioned issues in unsupervised instance detection is *motion*. In a practical sense, objects of interest are commonly *mobile* objects that frequently move. For example, robots performing navigation tasks must plan their trajectories carefully around objects that *can move* of their own volition. Thus, we argue that if we see similar groups of pixels frequently move of their own volition in unlabeled videos (*e.g.* vehicles and pedestrians in driving videos), this is sufficient information for building a *mobile object detector* that can detect such instances in static frames.

The importance of motion as a perceptual cue (the Gestalt principle of *common fate*) is well known [35]. Indeed, motion-based grouping is one of the first forms of grouping to appear developmentally in human infants [45] and can bootstrap other grouping cues [34]. There is also some prior work on using motion-based grouping in computer vision to produce (pseudo) ground-truth for feature learning [36] and discovering isolated, salient objects [10, 11]. However, there is a big-gap between motion segmentation and the kind of ground-truth we need for training a full-fledged instance-level object detector. First, motion segmentation produces a binary segmentation; this must be resolved into individual instances. Second, it only identifies moving objects and does not include objects (*e.g.* parked cars) that are static but mobile. Finally, motion segmentation only identifies nearby objects, since the pixel motion of faraway objects is too subtle to discern. Thus motion segmentation alone will still under-segment and miss many mobile objects: a problem for building object detectors.

Here we propose a new training scheme to address these challenges. Our approach (MOD-UV, a **M**obile **O**bject **D**etector learned from **U**nlabeled **V**ideos only; Figure 1) trains on unlabeled videos alone and produces a *mobile object detector* that can run on static frames. We first generate pseudo training la-

bels from motion segmentation estimated by our prior unsupervised framework Dynamo-Depth [47]. We then propose a new training scheme to address the challenges above, resulting in a final mobile object detector that detects 12× more mobile objects than the initial motion segmentation.

We test MOD-UV on self-driving scenes but evaluate on a variety of datasets. Specifically, we compare to recent state-of-the-art unsupervised object detectors and demonstrate improvements across the board, with notable improvements in Box AR by 6.6 on Waymo Open [46], 4.9 on nuScenes [4] and 6.2 on KITTI [17].

In sum, our contributions are:

1. We argue that motion as a cue is sufficient for unsupervised training of instance-level object detectors.
2. We propose a new training scheme that trains on unlabeled videos to produce a mobile object detector that can run on static images.
3. We demonstrate marked improvements over unsupervised object detection baselines across a range of datasets and metrics.

2 Related Work

Unsupervised Object Detection/Discovery from Images. Learning to identify and localize objects from unlabeled images is a challenging task, since object information must be obtained without any explicit human annotations. A long line of work seeks to discover prominent objects in large image collections [9, 51–53]. However, these approaches are fundamentally limited by the quality of the object proposals. More recently, Locatello *et al*. [30] and DINOSAUR [38] consider object discovery as object-centric learning [18] and decompose a complex scene into independent objects. Nevertheless, the reconstruction objective is difficult to scale and likely to discover irrelevant patches as well.

More recent work has relied on the fact that bottom-up segmentation algorithms when applied to self-supervised pretrained representations yield good object proposals. Specifically, pseudo mask labels can be generated from DINO [7] features to train downstream object detectors [41, 42, 54]. MaskDistill [50] extends upon this by distilling from affinity graph produced by DINO [7] features, while TokenCut [56] and CutLER [55] use Normalized Cuts [39]. HASSOD [6] leverages hierarchical adaptive clustering, which improves the detection of small objects and object parts. This line of work now produces detectors that can run on static images, similar to our work. However, the detected objects can often be irrelevant (*e.g*. buildings and road) or over-/under-segment objects of interest. In contrast, our proposed MOD-UV discovers and detects a more meaningful and practical set of mobile objects instead, and can learn from their apparent motion in unlabeled videos only without relying on any additional datasets.

Unsupervised Object Detection/Discovery from 3D. In addition to unlabeled images, 3D information is also useful for discovering objects. Herbst *et*

al. [21] and MODEST [62] discover non-persistent objects via multiple traversals with 3D sensors. Garcia *et al*. [16] discovers salient objects by late-fusing color and depth segmentation from RGB-D inputs, while Tian *et al*. [49] generates candidate segments from LiDAR 3D point clouds. In comparison, MOD-UV does not require any additional sensors or modalities beyond unlabeled videos, which allows our method to work in more general settings.

Unsupervised Object Detection/Discovery from Videos. Inspired by Gestalt principle of common fate [35], another class of related work discovers objects via their apparent motions observed in videos [31, 59, 61]. By leveraging optical flow information from an input video, a binary segmentation of the moving objects can be extracted [25, 26, 36, 43, 57, 60, 63, 64]. Lian *et al*. [28] proposes further improvements for cases of articulated/deformable objects and shadow/reflections by relaxing the common fate assumption. Another line of work uses a reconstruction objective to identify the moving object [1, 48]. Du *et al*. [13] models explicit object geometry and physical dynamics by exploiting motion cues. Bao *et al*. [2] improves training for object-centric representation via an additional motion segmentation regularization, while SAVi++ [14] incorporates LiDAR data when training an object-centric video model. Unlike MOD-UV, these approaches do not build a static image detector. However, the output segmentation can be used as an initialization for our approach.

Closer to our work, Pathak *et al*. [36], Croitoru *et al*. [11] and Choudhury *et al*. [10] train a single-frame binary segmentation network on video frames as input and leverage object motion as supervision. Furthermore, LOCATE [44] applies graph-cut to obtain binary motion mask from DINO [7] and optical flow feature similarities, which in turn is treated as pseudo-labels for bootstrapped self-training of a downstream segmentation network. However, these techniques can only detect a *single* salient object per frame. In contrast, MOD-UV generalizes to multi-object detection beyond single-object saliency detection.

3 Method

Problem setup: We assume an uncurated collection of unlabeled videos as input. In particular, we assume that these videos are obtained by an embodied agent observing, and optionally acting in the world. Solely from the unlabeled videos, the goal is to learn a detector that operates from a single frame and can detect and segment all mobile objects that can move of their own volition.

3.1 Initialization with Unsupervised Motion Segmentation

A key insight in MOD-UV is that if an object *can move*, it is likely that it *does move* many times in the collected data. Thus, we start by identifying moving objects in the videos; they can be initial seeds for learning about mobile objects.

Fortunately, the task of identifying independently moving pixels from unlabeled videos is a well-studied one [3, 37, 40, 58]. In particular, many recent

techniques have been proposed that learn motion segmentation without supervision from unlabeled videos. Many of these techniques also produce depth and camera motion [23,27,32,37]. Here, we use our prior work, Dynamo-Depth [47]. Dynamo-Depth trains on unlabeled videos and learns both a monocular depth estimator as well as a motion segmentation network. We use the outputs of these trained networks on our unlabeled videos as a starting point. Concretely, we denote the input set of unlabeled videos as $\{v_i\}$, with each video v_i containing consecutive frames $I_1, \ldots, I_n$ and known camera intrinsics. For each frame I_i, we obtain its estimated motion mask m_i and estimated monocular depth d_i.

With the given binary motion mask m_i, we first need to partition the moving pixels into instance-level labels. While disjoint moving regions can be easily separated, multiple moving objects in the same region would require additional information (*e.g.* 3D information) to separate. Therefore, for each image I_i, we project the corresponding moving pixels in m_i into pseudo 3D point clouds P_i via the estimated monocular depth d_i and inverse camera intrinsics K^{-1}.[1]

$$P_i = \{d_i(p)K^{-1}\overrightarrow{p} \quad | \quad m_i(p) = 1\} \tag{1}$$

Then, we cluster P_i via DBSCAN [15] to get a pseudo depth-aware partition of the motion mask m_i, which we treat as the initial pseudo-labels, $L_i^{(0)}$.

We evaluate the quality of these pseudo-labels qualitatively in Figure 2 and quantitatively in the top rows of Table 4. We find that these pseudo-labels have high precision, but have two severe limitations. First, they only identify moving objects, so they miss objects that are static but can move (*e.g.* parked cars) Second, they miss almost all faraway objects which tend to be small. This is because the apparent pixel motion of faraway objects is very hard to detect.

To tackle this issue of limited recall, we propose two self-training stages, *Moving2Mobile* and *Large2Small*, that progressively recover more mobile objects in the scene by aligning the training distribution of *static* and *small* objects with the available large moving objects in the initial pseudo-labels, respectively.

3.2 Self-Training for Unsupervised Mobile Object Detection.

Here, we describe each self-training stage of MOD-UV, as we progressively discover mobile objects to train the final mobile object detector.

***Moving2Mobile*: Learning to Detect Static Objects.** On the left of Figure 2, the initial pseudo-labels $L_i^{(0)}$, while having high precision, fail to capture the large static objects, *e.g.* the black sedan in the bottom. However, a parked black sedan looks the same as a moving black sedan if all one has is a single frame. In other words, moving and static objects are indistinguishable when observed from a single frame.

Thus, in *Moving2Mobile*, we simply train a detector to reproduce the pseudo-labeled instances in $L_i^{(0)}$, but with only a single frame as input. Since object

[1] $\overrightarrow{\cdot}$ denotes the conversion to homogeneous coordinates

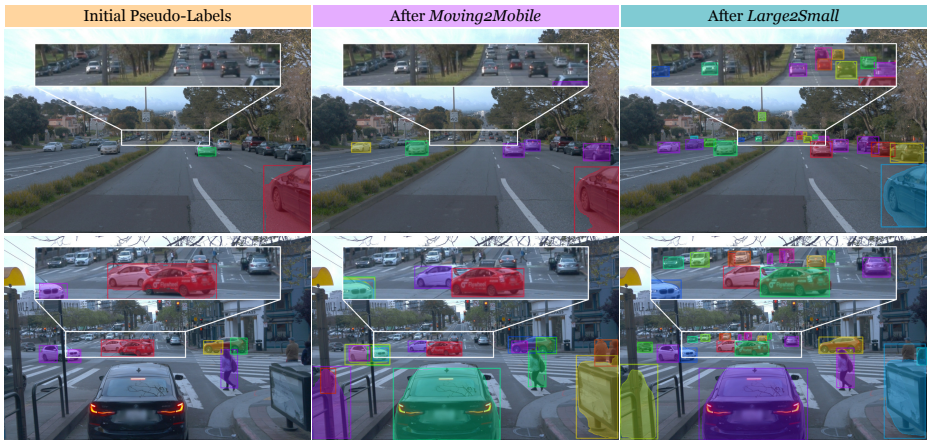

Fig. 2: Visualization of pseudo-labels at each stage of our self-training paradigm. From the initial pseudo-labels $L_i^{(0)}$ generated from motion mask, $L_i^{(1)}$ retrieves the large static objects after *Moving2Mobile* and $L_i^{(2)}$ recovers the small objects after *Large2Small*.

motion is not apparent in a single frame, this detector cannot distinguish moving objects from static ones and thus is forced to detect anything that share the appearance of moving objects, thus detecting static mobile objects as well.

However, there may exist domain-specific statistical regularities that give hints to object motion even in a single frame. For example, lit-up tail-lights might indicate that the car is stopped, while a highway background might suggest that the cars are moving. To prevent the detector from overfitting to these priors, we stop training early. Afterwards, we treat the high confidence predictions by the detector as the pseudo-labels for the next round $L_i^{(1)}$.

***Large2Small*: Learning to Detect Small Objects.** As shown in Figure 2, $L_i^{(1)}$, pseudo-labels after *Moving2Mobile*, appropriately recovers the large static objects, however, the smaller objects remain absent. Intuitively, faraway objects have much smaller apparent pixel motion (and thus are absent from $L_i^{(0)}$) and also look different from large moving objects (and thus are absent from $L_i^{(1)}$).

To learn to detect small objects, we create a new training dataset by scaling down both the image and the pseudo-labels (while also padding the image to maintain image size). We then train a separate "small object" detector by training on this new dataset. Intuitively, by training on the scaled down training pair, the output detector would need to detect the same object at a much lower scale, directly promoting the extension to small objects.

Also, since $L_i^{(1)}$ came from a heavily-regularized detector, we maintain and finetune the pseudo-labels for larger objects by training a second detector from scratch in parallel, on the training pair $(I_i, L_i^{(1)})$ without down-scaling or padding. Notably, this is different from traditional scale jittering, since the singular de-

tector would be discouraged from detecting small objects at larger scales due to the limitations in $L_i^{(1)}$. Upon convergence, we have a Large-object detector trained at original scale and another Small-object detector trained at a reduced scale. After aggregating their predictions and resolving conflicting proposals, we obtain the final pseudo-labels, $L_i^{(2)}$. We note that separating out large and small object detectors in this way has been explored in supervised face detection [22].

***Final Round* of Self-Training.** As shown in the right of Figure 2, $L_i^{(2)}$, the final pseudo-labels after *Large2Small*, successfully recovers both static and small objects without introducing excessive false-positives. From here, we train the final detector from scratch, on the training pair $(I_i, L_i^{(2)})$ to convergence.

3.3 Implementation details

We follow the official code release by Dynamo-Depth [47] and train the system on Waymo Open [46]. During initial pseudo-label generation, we binarize the estimated motion mask via a threshold of 0.1 and cluster the pseudo 3D points P_i via DBSCAN [15] using a 10-by-10 local pixel neighborhood connectivity.

We adopt Mask R-CNN [19] with a ResNet-50 [20] backbone as the detector architecture. We initialize the backbone via two strategies, namely MoCo v2 [8] on randomly sampled Waymo [46] patches and MoCo v2 on ImageNet [12], denoted as MOD-UV‡ and MOD-UV, respectively. We use Adam optimizer [24] with initial learning rate of 1e-4 and decay by $\frac{1}{2}$ after 10 epochs.

During *Moving2Mobile*, we train a detector for 3 epochs, with scale jittering from 0.5 to 1.0. Since early-stopping is applied, we adopt a lower confidence threshold of 0.5 to compute the next round pseudo-labels $L_i^{(1)}$. During *Large2Small*, we train both the large and small detectors for 20 epochs, with fixed scaling at 1.0 and 0.25, respectively. As both are trained to convergence, we adopt a higher confidence threshold of 0.9 and 0.8, respectively, to compute the next round pseudo-labels $L_i^{(2)}$ with aggregation. For *Final round*, we train the detector from scratch for 20 epochs, with scale jittering from 0.5 to 1.0. The self-training in MOD-UV takes 27 hours on 1 NVIDIA A6000 GPU.

4 Experiments

4.1 Experimental Setup

Datasets. For evaluation, we focus our attention to self-driving datasets since uncurated, unlabeled video data is available and detecting mobile objects is of interest for autonomous vehicles. We train both MOD-UV‡ and MOD-UV on Waymo [46] and compare our performance with baselines on Waymo. We then also evaluate generalization to nuScenes [4], KITTI [17], and COCO [29].

We split the 798 sequences from Waymo train set into 762 for training and 36 for validation. After method development concludes, we evaluate on the held-out

1,881 test images (averaging 28.4 mobile instances per image) from Waymo val set [33]. Additionally, nuScenes, KITTI, and COCO are only used for evaluating generalization. We test on 3,249 front-camera images (average of 8.2 mobile instances per image) from the nuImage validation set for nuScene and 7,481 images (average of 6.9 mobile instances per image) in the 2D Detection training set for KITTI. For COCO, we evaluate on 870 images (average of 3.8 mobile instances per image) in COCO val 2017 that contain ground vehicles.

Baselines. For the task of unsupervised mobile object detection, there is no directly comparable baselines to the best of our knowledge. Therefore, we consider methods for unsupervised object detection, namely CutLER [55] and HASSOD [6], as the closest points of comparison.

CutLER [55] uses normalized cuts on DINO features (trained on ImageNet) to generate pseudo labels that are used to train a Cascade Mask R-CNN [5] with ResNet-50 backbone. We evaluate the official checkpoint. Furthermore, we consider an additional baseline where $L_i^{(1)}$ is directly predicted via CutLER, which effectively ablates the use of motion cues, as denoted by CutLER^{L2S}.

HASSOD [6] is a follow-up to *CutLER*. It discovers objects on COCO [29] via a hierarchical adaptive clustering of DINO features (trained from ImageNet). The hierarchical clustering yields three "levels": objects, object parts, and object sub-parts. For consistency, we consider all three hierarchical levels for evaluation. As before, we use its official released checkpoint. Since MOD-UV‡ is trained on Waymo Open [46], we also consider a version of HASSOD solely trained on Waymo, which we denote as HASSOD†.

CutLER [55] and HASSOD [6] are trained to detect all objects in an image regardless of their ability to move. However, our evaluation and approach is focused on mobile objects. We therefore also consider an "oracle" version of these baselines where we additionally remove any CutLER and HASSOD predictions that overlap by less than 0.1 in IoU with the ground truth instances. These oracles, namely CutLER*, CutLER^{L2S*}, HASSOD*, and HASSOD$^{\dagger *}$, are grayed out in the tables to indicate additional ground-truth-based filtering.

We also consider a fully-supervised Mask R-CNN (Sup. Mask R-CNN) trained on COCO [29] for an oracle comparison, marked in gray.

Metrics. We evaluate both Average Recall ($\text{AR}^{\text{Box}}_{100}$ and $\text{AR}^{\text{Mask}}_{100}$) and Average Precision (AP^{Box} and AP^{Mask}). Since the task is unsupervised and no semantic information is given during training, we follow prior arts [6, 55] and evaluate class-agnostic AR and AP by treating all predicted and ground truth instances as a single class of "foreground" or "mobile" objects.

4.2 Unsupervised Mobile Object Detection and Segmentation

In-Domain Performance on Waymo. Table 1 compares MOD-UV‡ and MOD-UV against CutLER, HASSOD, and HASSOD† in unsupervised mobile object detection on Waymo. We also report performance for a MaskR-CNN trained on COCO as an oracle in the first row of Table 1.

Recall. We report $\text{AR}^{0.5}$ (average recall with IoU= 0.5), AR, AR_{S}, AR_{M}, and AR_{L} for both box and mask predictions. MOD-UV significantly outperforms

Table 1: Unsupervised Mobile Object Detection on Waymo Open Dataset [46]. We report detection and segmentation metrics and note the training data (Train) and backbone initialization data (Init.), including ImageNet (IN), COCO, and Waymo Open (W). † indicates manual replication with official released code, and * indicates the removal of proposals with <0.1 IoU overlap with ground truth instances.

Method	Train	Init.	Box					Mask				
			$AR^{0.5}$	AR	AR_S	AR_M	AR_L	$AR^{0.5}$	AR	AR_S	AR_M	AR_L
Sup. Mask R-CNN [19]			54.3	31.9	13.4	53.0	79.8	48.9	27.5	9.2	46.3	78.9
CutLER [55]	IN	IN	20.9	11.7	1.3	17.3	54.1	20.0	10.7	0.9	14.9	52.7
CutLER^{L2S} [55]	IN+W	IN	29.3	15.0	4.6	24.3	48.2	28.1	14.4	4.2	22.9	47.8
HASSOD [6]	COCO	IN	21.9	12.7	1.8	17.6	**59.5**	20.7	11.0	1.2	14.2	**55.6**
HASSOD† [6]	W	IN	15.3	8.3	0.5	11.0	43.7	14.6	7.2	0.2	7.9	42.6
MOD-UV‡ (ours)	W	W	**40.0**	17.5	8.4	25.7	46.4	35.4	14.6	6.8	21.2	40.4
MOD-UV (ours)	W	IN	39.9	**19.3**	**8.6**	**28.5**	54.1	**35.8**	**16.4**	**7.2**	**23.8**	47.3
			AP_{50}	AP	AP_S	AP_M	AP_L	AP_{50}	AP	AP_S	AP_M	AP_L
Sup. Mask R-CNN [19]			46.1	25.7	8.0	42.2	72.2	41.9	22.4	5.3	36.2	73.0
CutLER [55]	IN	IN	8.8	5.0	0.5	3.9	32.0	9.1	5.2	0.0	3.4	34.6
CutLER^{L2S} [55]	IN+W	IN	9.6	4.3	0.9	9.6	16.9	9.6	4.4	0.8	9.8	17.7
HASSOD [6]	COCO	IN	5.0	3.1	0.1	2.5	28.4	5.0	2.8	0.0	2.0	27.7
HASSOD† [6]	W	IN	3.7	2.2	0.0	1.1	17.2	3.9	2.0	0.0	0.9	18.3
MOD-UV‡ (ours)	W	W	26.1	10.9	4.2	16.2	32.0	25.0	9.5	3.7	14.1	28.7
MOD-UV (ours)	W	IN	**26.3**	**12.6**	**4.9**	**17.9**	**41.0**	**25.1**	**11.1**	**4.5**	**15.6**	**36.3**
CutLER* [55]	IN	IN	14.3	7.4	1.3	10.7	38.4	14.3	7.4	0.9	9.6	41.0
CutLER^{L2S}* [55]	IN+W	IN	23.2	10.1	3.3	16.4	32.6	22.8	10.2	3.1	16.3	34.1
HASSOD* [6]	COCO	IN	15.3	8.5	1.4	10.9	42.6	15.0	7.7	1.0	8.9	41.3
HASSOD†* [6]	W	IN	9.3	4.7	0.6	6.5	25.2	9.6	4.5	0.1	4.9	27.0

prior arts across all recall metrics except the recall for large objects where it is comparable. The improvement is especially large for small objects (**4.7**× higher AR_S^{Box} than the nearest competitor). Compared to a supervised Mask R-CNN trained on COCO [29], MOD-UV closes the gap in Box AR_S **from 11.6 to 4.8**. Our gains are also much larger on the $AR^{0.5}$ metric (nearly **2**× prior state-of-the-art). This suggests that we detect significantly more objects than prior work, but their localization can be improved. Even so, we still show a 6-point improvement on overall AR. We also found HASSOD to underperform when trained solely on Waymo, which we suspect is due to the uncurated nature of self-driving scenes.

Precision. We report AP at an overlap threshold of 0.5, as well as AP, AP_S, AP_M and AP_L. Since CutLER and HASSOD are trained to detect all objects in an image regardless of their ability to move, we also compare to oracle versions of these techniques with ground-truth-based filtering. On Waymo, MOD-UV significantly outperforms prior arts across **all** precision metrics. Even with ground-truth filtering (in gray), MOD-UV still consistently outperforms baselines on all precision metrics (except for larger objects where it is comparable). Specifically, MOD-UV outperforms the nearest competitor (with ground-truth filtering) by

Table 2: Zero-shot Unsupervised Mobile Object Detection on nuScenes [4]. We report detection and segmentation metrics and note the training data (Train) and backbone initialization data (Init.), including ImageNet (IN), COCO, and Waymo Open (W). † indicates manual replication with official released code, and * indicates the removal of proposals with <0.1 IoU overlap with ground truth instances.

Method	Train	Init.	Box					Mask				
			$AR^{0.5}$	AR	AR_S	AR_M	AR_L	$AR^{0.5}$	AR	AR_S	AR_M	AR_L
CutLER [55]	IN	IN	28.9	16.0	3.8	23.4	56.1	27.8	14.3	3.0	20.4	53.0
HASSOD [6]	COCO	IN	30.9	17.0	5.0	24.0	**56.8**	29.7	14.7	3.9	20.1	**53.6**
HASSOD† [6]	W	IN	24.3	12.7	2.7	18.0	48.2	23.0	10.7	1.9	14.0	45.7
MOD-UV‡ (ours)	W	W	42.1	17.3	8.7	24.2	39.8	36.4	13.9	8.2	18.1	29.7
MOD-UV (ours)	W	IN	**48.9**	**21.9**	**12.0**	**29.8**	48.2	**42.3**	**18.3**	**10.7**	**24.0**	39.1
			AP_{50}	AP	AP_S	AP_M	AP_L	AP_{50}	AP	AP_S	AP_M	AP_L
CutLER [55]	IN	IN	6.0	3.7	0.3	3.1	23.7	5.9	3.5	0.1	2.8	22.7
HASSOD [6]	COCO	IN	3.9	2.2	0.1	2.1	20.5	3.8	2.0	0.1	1.8	19.5
HASSOD† [6]	W	IN	3.6	2.2	0.0	1.7	15.2	3.6	1.8	0.0	0.9	14.6
MOD-UV‡ (ours)	W	W	18.8	7.3	2.6	10.4	22.3	17.1	6.0	2.4	8.0	17.2
MOD-UV (ours)	W	IN	**23.6**	**10.7**	**4.3**	**14.9**	**31.5**	**21.8**	**9.0**	**3.8**	**12.2**	**25.6**
CutLER* [55]	IN	IN	15.6	8.2	2.3	12.6	38.4	15.2	7.7	1.7	11.4	37.0
HASSOD* [6]	COCO	IN	18.6	9.5	2.3	13.3	38.2	17.9	8.6	1.8	11.5	36.2
HASSOD†* [6]	W	IN	13.0	6.3	1.2	9.1	27.9	12.8	5.6	1.1	7.5	27.2

4.1 on Box AP and 3.4 on Mask AP, and is **4.5**× higher on AP_S^{Mask}. Intriguingly, compared to a supervised Mask R-CNN trained on COCO [29], MOD-UV closes the gap in Mask AP_S **from 4.3 to 0.8 points**.

Generalization to Out-of-Domain Data. We next take our detector trained on Waymo, and apply it *out of the box* on nuScenes, KITTI, and COCO.

Recall. As shown in Table 2 and Table 3, on nuScenes and KITTI, MOD-UV consistently outperforms prior arts across all AR metrics except for large objects, achieving a more than **1.5**× improvement on $AR^{0.5}$ over the nearest competitor on nuScenes. MOD-UV also shows large gains on small objects, improving AR_S^{Box} by **2.4**× on nuScenes and over **1.7**× on KITTI.

Finally, we evaluate on COCO, which is in-domain for HASSOD and a big domain shift for MOD-UV. Notably, MOD-UV maintains superiority on AR_S and AR at IoU= 0.5, while being comparable to HASSOD on AR and AR_M.

Precision. This improvement is also seen in AP. On both nuScenes and KITTI, MOD-UV consistently outperforms baseline on all AP metrics except being comparable for $AP_L$with prior arts with ground-truth filtering. Specially, MOD-UV improves upon HASSOD*(with ground truth filtering) on Box AP by 1.2 on nuScenes and 2.6 on KITTI, with notable improvements on AP_S^{Box} by over **1.8**× on nuScenes and **2.6**× on KITTI. Even on COCO, MOD-UV outperforms prior arts on all metrics without ground-truth filtering except AP_L.

Table 3: Unsupervised Mobile Object Detection on KITTI [17] and COCO [29]. We report Average Recall (AR^{Box}_{100}) and Average Precision (AP^{Box}). Manual replication with official released code is indicated by †, and * indicates the removal of proposal with less than 0.1 IoU overlap with all ground truth instances.

Method	Train	Init.	KITTI [17]					COCO [29]				
			$AR^{0.5}$	AR	AR_S	AR_M	AR_L	$AR^{0.5}$	AR	AR_S	AR_M	AR_L
CutLER [55]	IN	IN	50.9	24.4	11.2	23.0	42.9	49.0	**27.9**	9.8	33.9	**61.6**
HASSOD [6]	COCO	IN	52.4	26.4	13.4	23.9	**47.1**	51.2	**27.9**	12.5	**36.2**	51.7
HASSOD† [6]	W	IN	49.9	23.7	15.0	22.1	37.4	48.8	26.1	12.5	31.2	50.7
MOD-UV‡ (ours)	W	W	66.1	29.3	21.2	29.0	39.6	55.9	23.3	20.1	26.9	25.5
MOD-UV (ours)	W	IN	**68.1**	**32.6**	**23.7**	**32.0**	44.3	**58.3**	26.6	**22.3**	29.8	32.0
			AP_{50}	AP	AP_S	AP_M	AP_L	AP_{50}	AP	AP_S	AP_M	AP_L
CutLER [55]	IN	IN	18.6	8.6	0.4	5.6	24.0	9.8	5.6	0.4	2.4	**21.6**
HASSOD [6]	COCO	IN	14.3	7.2	0.3	5.1	**29.5**	3.1	1.4	0.1	1.9	7.6
HASSOD† [6]	W	IN	16.7	7.5	0.6	6.8	19.3	4.7	2.7	0.3	3.1	10.0
MOD-UV‡ (ours)	W	W	38.5	15.8	9.8	15.3	26.2	14.1	5.8	5.1	7.6	5.9
MOD-UV (ours)	W	IN	**38.9**	**18.0**	**11.7**	**18.0**	28.4	**14.2**	**6.6**	**5.6**	**9.0**	6.6
CutLER* [55]	IN	IN	28.4	12.3	3.5	9.9	28.8	24.3	12.8	4.3	12.1	40.1
HASSOD* [6]	COCO	IN	33.4	15.4	4.4	11.5	34.9	21.3	10.1	5.4	13.8	18.9
HASSOD$^{\dagger *}$ [6]	W	IN	30.8	12.7	5.7	11.1	23.3	22.8	10.6	5.8	12.7	21.3

4.3 Qualitative Results

In Figure 3, we show qualitative examples of MOD-UV against CutLER and HASSOD after ground truth filtering, which we denote by CutLER* and HASSOD*, respectively. In addition, we highlight the regions containing small objects with an additional zoom-in.

Without using any annotations, MOD-UV detects mobile objects accurately, especially recovering many more small and faraway objects compared to prior arts. In contrast, due to the reliance on image features from static images, both CutLER and HASSOD tend to group multiple objects into a single proposal (seen in the second row in Waymo).

Notably, MOD-UV reliably detects static and small mobile objects in the scene without excess amount of false positives. This improvement mostly originates from the proposed *Moving2Mobile* and *Large2Small* (see ablation below).

Beyond accurate detection and segmentation on Waymo, MOD-UV demonstrates impressive generalization when applied on nuScenes, KITTI, and COCO.

4.4 Ablation Study

We conduct ablation studies with MOD-UV‡ trained solely from Waymo [46] to understand the effects of each proposed component, including pseudo-label generation, static object discovery, small object discovery, and final round training.

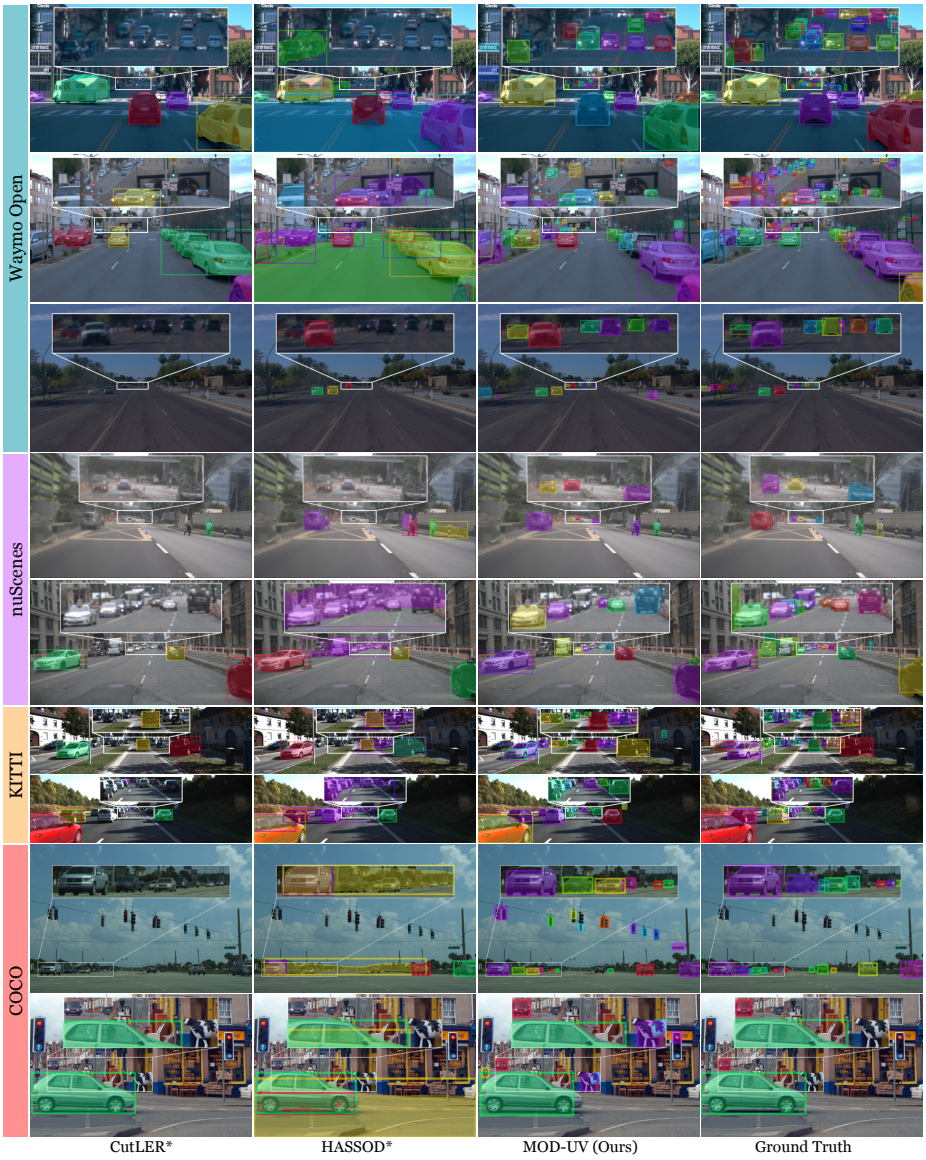

Fig. 3: Qualitative Results on Waymo Open, nuScenes, KITTI, and COCO, where all proposals with over 0.5 confidence are visualized. For CutLER and HASSOD, we apply an additional filtering that removes any proposals with <0.1 IoU with ground truth mobile objects, as denoted by CutLER* and HASSOD*, respectively.

Motion Cues for Initial Pseudo-Labels. As shown in Table 1, there is a consistent AR improvement for small and medium objects in CutLER^{L2S} over CutLER. This highlights the effectiveness of our *Large2Small* strategy in

Table 4: Ablation Study on the processing of pseudo-labels with MOD-UV‡. We report pseudo-label mask quality in terms of AR^{Mask} on the training set of Waymo Open [46].

Pseudo Masks	# Epochs in *M2M*	All				Static				Moving			
		AR	AR_S	AR_M	AR_L	AR	AR_S	AR_M	AR_L	AR	AR_S	AR_M	AR_L
Contour	✗	0.9	0.0	0.3	6.4	0.6	0.0	0.0	2.0	3.8	0.0	1.0	14.4
Depth	✗	1.5	0.0	0.5	10.0	0.7	0.0	0.0	2.4	6.3	0.0	1.5	23.8
Depth	1	4.4	0.0	3.2	25.9	7.1	0.0	2.2	21.1	11.4	0.0	6.7	34.7
Depth	3	5.3	0.0	4.6	29.0	9.1	0.0	3.5	25.8	12.4	0.0	8.7	35.1
Depth	5	4.3	0.0	3.2	24.7	6.7	0.0	1.8	20.2	11.1	0.0	7.1	32.9
Depth	20	4.1	0.0	2.3	25.6	6.6	0.0	1.3	20.8	10.6	0.0	5.2	34.4

Table 5: Ablation Study on the proposed self-training scheme involving *Moving2Mobile*, *Large2Small*, and *Final Round* with MOD-UV‡. We report Box AR_{100} and Box AP for Unsupervised Mobile Object Detection on the Waymo Open [46].

Stage 1 *M2M*	→	Stage 2 *L2S*	→	Stage 3 *Final*	Box AR					Box AP				
					$AR^{0.5}$	AR	AR_S	AR_M	AR_L	AP_{50}	AP	AP_S	AP_M	AP_L
✗		✗		✓	18.0	7.8	0.3	9.7	43.3	10.5	4.2	0.4	3.5	25.3
✗		$\boldsymbol{L}$ only		✗	12.9	5.6	0.0	4.5	38.3	7.6	2.9	0.4	1.1	22.3
✗		$\boldsymbol{S}$ only		✗	23.0	10.2	3.0	18.3	28.3	11.8	5.1	1.6	8.9	15.0
✗		$\boldsymbol{L}+\boldsymbol{S}$		✓	28.4	12.6	2.3	20.8	47.7	16.4	7.0	1.6	10.6	32.8
✓		✗		✗	23.2	10.1	1.0	16.0	45.1	13.7	5.6	0.5	7.8	28.1
✓		✗		✓	28.1	12.5	2.2	20.3	48.4	16.7	7.1	1.4	10.8	33.5
✓		$\boldsymbol{L}$ only		✗	21.3	9.4	0.2	14.9	45.2	12.8	5.7	0.5	7.2	31.2
✓		$\boldsymbol{S}$ only		✗	31.0	13.4	5.8	22.4	32.5	17.1	7.2	2.9	13.6	15.0
✓		$\boldsymbol{L}+\boldsymbol{S}$		✓	40.0	17.5	8.4	25.7	46.4	26.1	10.9	4.2	16.2	32.0

improving detector performance on small objects, which is a challenge for all unsupervised detectors/object discovery techniques because of the limited signal on small objects. Despite of this gain, MOD-UV still outperforms CutLER^{L2S} because motion offers a stronger cue to separate small objects that appear close to each other in pixel space, as shown in Figure 3.

Pseudo-Label Generation. In Table 4, we measure the quality of the pseudo-labels in terms of AR for All, the Static, and the Moving instances.

In the top of Table 4, compared to using 2D contours for generate pseudo-labels, clustering pseudo 3D points from monocular depth estimations improves Moving AR_L by 9.4 and Moving AR by 2.5. This underlines the benefit in leveraging 3D information for partitioning moving instances. Nevertheless, the Static AR is notably smaller, with Static AR_L only up to **10%** of Moving AR_L. Also, small and medium objects are almost entirely missed, with All AR_S at 0.0 and All AR_M at 0.5, compared to All AR_L at 10.0. These observations further verify

the bias pointed out in Section 3.1, where static and small objects are mostly absent in the initial pseudo-labels generated from motion segmentation.

Furthermore, in the bottom half of Table 4, we demonstrate the effectiveness of self-training in the *Moving2Mobile* stage in recovering static objects from the initial pseudo-labels with varying number of training epochs. It is worth noting that regardless of convergence, self-training successfully improves Static AR, with improvements as high as 23.4 on Static AR_L. Interestingly, as the initial pseudo-labels contain mostly moving objects, additional training beyond 3 epochs shows clear degradation in performance (reducing Static AR_L by 5.0 while retaining Moving AR_L), as the trained detector overfits to moving instances by exploiting contextual priors. In addition to improvements on All AR_L, the *Moving2Mobile* stage is also able to slightly lift up All AR_M by the highest at 4.1. Nevertheless, with no improvements on AR_S, it is clear that the *Moving2Mobile* stage alone cannot alleviate the negative bias from motion segmentation.

Self-Training Pipeline. In Table 5, we evaluate every combination of the 3-stage self-training pipeline with MOD-UV‡ to evaluate the effectiveness of each. Here, the *Moving2Mobile* stage is again shown to be essential for recovering static objects from the initial pseudo-labels. When *Moving2Mobile* is ablated, performance decrease across all combinations, with notable reductions in AR by 4.9 and AP by 3.9 when solely ablated from MOD-UV‡.

Additionally, when *Large2Small* is ablated, AR_S reduces by nearly $\mathbf{4}\times$ and AP_S by $\mathbf{3}\times$, underlining its importance for small object detection. Lastly, the final self-training round effectively learns the aggregated proposals from *Large2Small*, leading to an improved Box AR by 8.1 from ***L***arge-object detector trained at original scale and by 4.1 from ***S***mall-object detector trained at reduced scale.

5 Conclusion

We argue that motion is an important cue for unsupervised object detection, and propose the task of unsupervised mobile object detection. We propose a new training pipeline, MOD-UV, that bootstraps from motion segmentation but removes its bias by discovering static and small objects. MOD-UV achieves significant improvement over prior self-supervised detectors on multiple datasets.

Limitations. Our work makes an assumption that all mobile objects would often move in the given unlabeled video dataset. Although MOD-UV can ideally learn and detect all mobile objects, in practice, the learning-based framework can only learn and detect things that frequently move in the videos. That said, for general applications where the autonomous agents can manipulate their surroundings, the agent can still learn to detect rarely moving objects by interacting with their environment, *e.g.* poking at static objects within reach.

Societal Impact. Being an unsupervised detection framework, our work does not include any negative social impacts beyond object detection itself.

Acknowledgement

This research is based upon work supported in part by the National Science Foundation (IIS-2144117 and IIS-2107161). Yihong Sun is supported by an NSF graduate research fellowship.

References

1. Aydemir, G., Xie, W., Guney, F.: Self-supervised object-centric learning for videos. Advances in Neural Information Processing Systems **36** (2024)
2. Bao, Z., Tokmakov, P., Jabri, A., Wang, Y.X., Gaidon, A., Hebert, M.: Discovering objects that can move. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11789–11798 (2022)
3. Bouthemy, P., François, E.: Motion segmentation and qualitative dynamic scene analysis from an image sequence. International Journal of Computer Vision **10**(2), 157–182 (1993)
4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
5. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
6. Cao, S., Joshi, D., Gui, L.Y., Wang, Y.X.: Hassod: Hierarchical adaptive self-supervised object detection. arXiv preprint arXiv:2402.03311 (2024)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
8. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
9. Cho, M., Kwak, S., Schmid, C., Ponce, J.: Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1201–1210 (2015)
10. Choudhury, S., Karazija, L., Laina, I., Vedaldi, A., Rupprecht, C.: Guess what moves: Unsupervised video and image segmentation by anticipating motion. arXiv preprint arXiv:2205.07844 (2022)
11. Croitoru, I., Bogolin, S.V., Leordeanu, M.: Unsupervised learning from video to detect foreground objects in single images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4335–4343 (2017)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
13. Du, Y., Smith, K., Ulman, T., Tenenbaum, J., Wu, J.: Unsupervised discovery of 3d physical objects from video. arXiv preprint arXiv:2007.12348 (2020)
14. Elsayed, G., Mahendran, A., van Steenkiste, S., Greff, K., Mozer, M.C., Kipf, T.: Savi++: Towards end-to-end object-centric learning from real-world videos. Advances in Neural Information Processing Systems **35**, 28940–28954 (2022)
15. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: kdd. vol. 96, pp. 226–231 (1996)

16. García, G.M., Potapova, E., Werner, T., Zillich, M., Vincze, M., Frintrop, S.: Saliency-based object discovery on rgb-d data with a late-fusion approach. In: 2015 IEEE International Conference on Robotics and Automation (ICRA). pp. 1866–1873. IEEE (2015)
17. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
18. Greff, K., Van Steenkiste, S., Schmidhuber, J.: On the binding problem in artificial neural networks. arXiv preprint arXiv:2012.05208 (2020)
19. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
21. Herbst, E., Ren, X., Fox, D.: Rgb-d object discovery via multi-scene analysis. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4850–4856. IEEE (2011)
22. Hu, P., Ramanan, D.: Finding tiny faces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 951–959 (2017)
23. Hui, T.W.: Rm-depth: Unsupervised learning of recurrent monocular depth in dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1675–1684 (2022)
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
25. Kwak, S., Cho, M., Laptev, I., Ponce, J., Schmid, C.: Unsupervised object discovery and tracking in video collections. In: Proceedings of the IEEE international conference on computer vision. pp. 3173–3181 (2015)
26. Lamdouar, H., Yang, C., Xie, W., Zisserman, A.: Betrayed by motion: Camouflaged object discovery via motion segmentation. In: Proceedings of the Asian Conference on Computer Vision (2020)
27. Li, H., Gordon, A., Zhao, H., Casser, V., Angelova, A.: Unsupervised monocular depth learning in dynamic scenes. In: Conference on Robot Learning. pp. 1908–1917. PMLR (2021)
28. Lian, L., Wu, Z., Yu, S.X.: Bootstrapping objectness from videos by relaxed common fate and visual grouping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14582–14591 (2023)
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
30. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. Advances in Neural Information Processing Systems **33**, 11525–11538 (2020)
31. Lu, X., Wang, W., Shen, J., Tai, Y.W., Crandall, D.J., Hoi, S.C.: Learning video object segmentation from unlabeled videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8960–8970 (2020)
32. Luo, C., Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R., Yuille, A.: Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2624–2641 (2019)

33. Mei, J., Zhu, A.Z., Yan, X., Yan, H., Qiao, S., Chen, L.C., Kretzschmar, H.: Waymo open dataset: Panoramic video panoptic segmentation. In: European Conference on Computer Vision. pp. 53–72. Springer (2022)
34. Ostrovsky, Y., Meyers, E., Ganesh, S., Mathur, U., Sinha, P.: Visual parsing after recovery from blindness. Psychological Science **20**(12), 1484–1491 (2009)
35. Palmer, S.E.: Visual Perception of Objects, chap. 7, pp. 177–211. John Wiley & Sons, Ltd (2003). https://doi.org/https://doi.org/10.1002/0471264385.wei0407
36. Pathak, D., Girshick, R., Dollár, P., Darrell, T., Hariharan, B.: Learning features by watching objects move. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2701–2710 (2017)
37. Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M.J.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12240–12249 (2019)
38. Seitzer, M., Horn, M., Zadaianchuk, A., Zietlow, D., Xiao, T., Simon-Gabriel, C.J., He, T., Zhang, Z., Schölkopf, B., Brox, T., et al.: Bridging the gap to real-world object-centric learning. arXiv preprint arXiv:2209.14860 (2022)
39. Shi, J.: Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence **22**(8), 888–905 (2000)
40. Shi, J., Malik, J.: Motion segmentation and tracking using normalized cuts. In: Sixth international conference on computer vision (IEEE Cat. No. 98CH36271). pp. 1154–1160. IEEE (1998)
41. Siméoni, O., Puy, G., Vo, H.V., Roburin, S., Gidaris, S., Bursuc, A., Pérez, P., Marlet, R., Ponce, J.: Localizing objects with self-supervised transformers and no labels. arXiv preprint arXiv:2109.14279 (2021)
42. Siméoni, O., Sekkat, C., Puy, G., Vobeckỳ, A., Zablocki, É., Pérez, P.: Unsupervised object localization: Observing the background to discover objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3176–3186 (2023)
43. Singh, S., Deshmukh, S., Sarkar, M., Jain, R., Hemani, M., Krishnamurthy, B.: Fodvid: Flow-guided object discovery in videos. arXiv preprint arXiv:2307.04392 (2023)
44. Singh, S., Deshmukh, S., Sarkar, M., Krishnamurthy, B.: Locate: Self-supervised object discovery via flow-guided graph-cut and bootstrapped self-training. arXiv preprint arXiv:2308.11239 (2023)
45. Spelke, E.S.: Principles of object perception. Cognitive science **14**(1), 29–56 (1990)
46. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
47. Sun, Y., Hariharan, B.: Dynamo-depth: Fixing unsupervised depth estimation for dynamical scenes. Advances in Neural Information Processing Systems **36** (2024)
48. Tangemann, M., Schneider, S., Von Kügelgen, J., Locatello, F., Gehler, P., Brox, T., Kümmerer, M., Bethge, M., Schölkopf, B.: Unsupervised object learning via common fate. arXiv preprint arXiv:2110.06562 (2021)
49. Tian, H., Chen, Y., Dai, J., Zhang, Z., Zhu, X.: Unsupervised object detection with lidar clues. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5962–5972 (2021)

50. Van Gansbeke, W., Vandenhende, S., Van Gool, L.: Discovering object masks with transformers for unsupervised semantic segmentation. arXiv preprint arXiv:2206.06363 (2022)
51. Vo, H.V., Bach, F., Cho, M., Han, K., LeCun, Y., Pérez, P., Ponce, J.: Unsupervised image matching and object discovery as optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8287–8296 (2019)
52. Vo, H.V., Pérez, P., Ponce, J.: Toward unsupervised, multi-object discovery in large-scale image collections. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16. pp. 779–795. Springer (2020)
53. Vo, V.H., Sizikova, E., Schmid, C., Pérez, P., Ponce, J.: Large-scale unsupervised object discovery. Advances in Neural Information Processing Systems **34**, 16764–16778 (2021)
54. Wang, X., Yu, Z., De Mello, S., Kautz, J., Anandkumar, A., Shen, C., Alvarez, J.M.: Freesolo: Learning to segment objects without annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14176–14186 (2022)
55. Wang, X., Girdhar, R., Yu, S.X., Misra, I.: Cut and learn for unsupervised object detection and instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3124–3134 (2023)
56. Wang, Y., Shen, X., Yuan, Y., Du, Y., Li, M., Hu, S.X., Crowley, J.L., Vaufreydaz, D.: Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
57. Xie, C., Xiang, Y., Harchaoui, Z., Fox, D.: Object discovery in videos as foreground motion clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9994–10003 (2019)
58. Yan, J., Pollefeys, M.: A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In: Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV 9. pp. 94–106. Springer (2006)
59. Yan, P., Li, G., Xie, Y., Li, Z., Wang, C., Chen, T., Lin, L.: Semi-supervised video salient object detection using pseudo-labels. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7284–7293 (2019)
60. Yang, C., Lamdouar, H., Lu, E., Zisserman, A., Xie, W.: Self-supervised video object segmentation by motion grouping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7177–7188 (2021)
61. Yao, R., Lin, G., Xia, S., Zhao, J., Zhou, Y.: Video object segmentation and tracking: A survey. ACM Transactions on Intelligent Systems and Technology (TIST) **11**(4), 1–47 (2020)
62. You, Y., Luo, K.Z., Phoo, C.P., Chao, W.L., Sun, W., Hariharan, B., Campbell, M., Weinberger, K.Q.: Learning to detect mobile objects from lidar scans without labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2022)
63. Zhang, K., Zhao, Z., Liu, D., Liu, Q., Liu, B.: Deep transport network for unsupervised video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8781–8790 (2021)
64. Zhou, T., Li, J., Li, X., Shao, L.: Target-aware object discovery and association for unsupervised video multi-object segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6985–6994 (2021)