Cross-Modal Distillation for 2D/3D Multi-Object Discovery from 2D motion

Anonymous CVPR submission

Paper ID 18736

Abstract

Object discovery, which refers to the task of localizing objects without human annotations, has gained significant attention in 2D image analysis. However, despite this growing interest, it remains under-explored in 3D data, where approaches rely exclusively on 3D motion, despite its several challenges. In this paper, we present a novel framework that leverages advances in 2D object discovery which are based on 2D motion to exploit the advantages of such motion cues being more flexible and generalizable and to bridge the gap between 2D and 3D modalities. Our primary contributions are twofold: (i) we introduce DIOD-3D, the first baseline for multi-object discovery in 3D data using 2D motion, incorporating scene completion as an auxiliary task to enable dense object localization from sparse input data; (ii) we develop xMOD, a cross-modal training framework that integrates 2D and 3D data while always using 2D motion cues. xMOD employs a teacher-student training paradigm across the two modalities to mitigate confirmation bias by leveraging the domain gap. During inference, the model supports both RGB-only and point cloud-only inputs. Additionally, we propose a late-fusion technique tailored to our pipeline that further enhances performance when both modalities are available at inference. We evaluate our approach extensively on synthetic (TRIP-PD) and challenging real-world datasets (KITTI and Waymo). Notably, our approach yields a substantial performance improvement compared with the 2D object discovery state-of-the-art on all datasets with gains ranging from $+8.7$ *to* $+15.1$ *in* $F1@50$ *score* [1].

1. Introduction

Object detection has been extensively explored, leading to fast, high-performance approaches [4, 26, 27]. However, these methods adopt a fully supervised setting that suffers from high annotation costs and makes them impractical for scaling with the increasing data needed for better generalization. Additionally, this setting is limited to detecting specific semantic categories, which poses challenges in identifying out-of-distribution instances and rare categories. Object discovery has thus emerged as an unsupervised alternative to the *localization* component of object detection. It focuses on localizing *objects* within images or videos without explicit prior knowledge provided by human annotations. Interest in this task continues to grow in the 2D modality [2, 15, 28, 30] driven by the presence of object patterns *for free* within low-level and automatically acquired modalities (motion [1, 17], depth [10], etc), resulting in interesting performances. Moreover, the class-agnostic nature of object discovery and its reliance on low-level signals allow for a broader application, built around general definitions of objects, such as salient objects [31, 38] and objects that can move [1]. These properties address the limitations of the fully supervised setting. In contrast, these advances are not mirrored enough in the 3D modality where only 3D motion cues are explored despite being sparse and demanding extensive fine-tuning with changing domains.

In this work, we show that 3D object discovery (3DOD) can largely benefit from advancements achieved in the 2D modality. Specifically, we adapt the recent motion-guided 2D object discovery (2DOD) approach in [15] to accommodate 3D data, using the same 2D motion masks. Object discovery being unsupervised, it typically includes a reconstruction pretext task as a powerful regularization method. In the real-world scenario, we discovered that the inherent sparsity in LiDAR data (*i.e*. missing data points and poor spatial resolution) makes 3DOD challenging, leading to incomplete object segments. We thus propose scene completion as a more suitable pretext task for 3DOD. Specifically, we encourage the prediction of a denser point cloud, which helps avoid propagating the input sparsity to the predicted object masks.

Subsequently, we aim to ensure that the transition to 3D is not disconnected from the 2D data, which is rich in complementary information such as colors and textures. To this end, we propose bridging the two modalities by jointly optimizing the tasks of 2D and 3D object discovery, while always using the same 2D motion cues. The effectiveness of

[1] Code available upon acceptance

distillation for object discovery has been demonstrated in an intra-modal setting [15], where it progressively reintegrates discovered objects into the supervision set and eliminates noisy pseudo-labels, enhancing robustness. In our work, we explore distillation in a cross-modal setting. To achieve this, we design two teacher-student systems, one for each modality, and establish interactions between the four models using objective functions that enable the student model of one domain to be supervised by the teacher model of the alternate domain. Advantageously, our approach increases the robustness of the system when a modality becomes inoperative due to a difficult environment, such as night scenes for a camera (*2D-blind*) or the absence of reflections for a 3D sensor (*3D-blind*). This process also leverages the domain gap between the student and teacher models, as each receives inputs from a distinct modality, reducing the risk of confirmation bias.

During inference, our method can accommodate 2D only, 3D only and multi-modal inputs, depending on the application and available sensors. In the multi-modal setting, we explore the consistency between both modalities as a source of reliability in multi-sensor applications, considering consistent predictions between the two modalities as the most reliable object candidates.

In summary, (i) we propose a first baseline to solve multiple object discovery from point clouds using 2D motion cues, with scene completion as a suitable pretext task for 3DOD; (ii) we design a cross-modal training framework, based on 2D motion information, that integrates 2D and 3D data to enable interaction between the two modalities, addressing modality-related difficult cases. Experiments conducted on three datasets demonstrate that each modality benefits significantly from cross-modal learning with the alternate modality, validating the effectiveness of the proposed approach.

2. Related Work

2.1. Object discovery in RGB images (2DOD)

Object discovery in 2D images/videos addresses the challenge of localizing instances of objects when human annotations are unavailable. In RGB images, this task has significantly benefited from advances in self-supervised learning [5, 25], which have led to the emergence of segmentation properties in learned representations [32]. Notably, DINOSAUR [30] demonstrated that reconstructing those prelearned features enables self-supervised scene decomposition into objects.

Recently, 2DOD has achieved greater success in video data, driven by the availability of motion information that serves as a cue for object localization. Motion information has been primarily incorporated into slot-attention-based approaches, with slot-attention being the mechanism that facilitates scene decomposition into objects within the latent space of an auto-encoder architecture [21]. Motion is integrated in various ways across different methods: SAVI [17] learns to predict optical flow, focusing particularly on the localization of moving objects. On the other hand, VideoSAUR [44], a video version of [30], exploits semantic similarity between image patches to predict their temporal displacement, thus incorporating motion implicitly into the learned representation. More explicitly, another research direction [1, 2, 14, 15] leverages motion-derived segments, highlighting moving objects, to guide slots' learning; some approaches also address noise in image backgrounds [14] and the generalization from moving to static objects [15].

Although these methods demonstrated interesting results for 2DOD, these advances along with the use of 2D motion cues have not been exploited yet for both 3DOD and corssmodal object discovery.

2.2. Object Discovery in 3D data (3DOD)

Compared to 2DOD, 3DOD is less explored [22, 24]. In single images, it is typically limited to single-object localization [39], while in sequential data, the primary approach leverages 3D motion cues to identify only moving objects [9, 24]. However, in LiDAR-based applications like road scenes, ignoring stationary objects, such as stopped vehicles, raises safety concerns. In an other category, Open-set detection [6] generalizes to unknown objects but primarily relies on highly-supervised closed-set detectors, while vision-language methods [11] assume known or describable classes of objects, which is more restrictive than general object discovery. Other approaches [22, 37, 43, 45], while unsupervised, mostly cluster 3D point clouds [22, 43] or scene flow cues [37], requiring intensive tuning and heuristic-based filtering of non-object regions [43]. Clustering in 3D data is further challenged by LiDAR's low resolution and sparse points on distant objects.

In this work, we aim to extend advancements from 2DOD (Section 2.1) to the 3D domain. Similar to how 2D object-centric learning offers a deep learning-based alternative to 2D clustering, this extension seeks to replace clustering methods for 3D point clouds, which are sensitive to parameters like object count and intra-object point density. Our hypothesis is also that 3D data can, in turn, enhance object discovery in 2D, thus the proposed cross-modal distillation framework.

2.3. Motion Cues for Object Localization

An important part of understanding a scene is modelling its dynamics. This has motivated many works on motion estimation both in RGB images through optical flow estimation (*i.e.* the pixel displacement between successive frames) [33, 35, 40] and in 3D by estimating 3D displacements of each point, known as scene flow [18, 20, 23]. Motion information has notably served as a cue for the presence of objects of interest: moving objects [8], objects capable of moving [1], *etc*. For instance, in [29] which addresses semi-supervised segmentation of moving objects in point clouds, scene flow is employed to localize mobile objects. Conversely, 2D methods utilize optical flow for scene analysis [41, 42].

In this work, the choice of using 2D-derived motion cues, instead of 3D scene flow offers several advantages: (i) It avoids the need for pre-processing steps like ground removal in point clouds, a common requirement in clustering-based methods [3] that entails additional hyper-parameter tuning. Recent advances in video object discovery (2.1) handle this automatically, even filtering out other permanently static regions such as buildings. (ii) Using the 2D domain as the source for pseudo-labels, rather than point clouds, helps reduce errors associated with the low resolution of LiDAR data, particularly on distant objects. (iii) Finally, leveraging 2D-derived supervision to solve 3DOD opens the perspective of using the vast resource of foundation models emerging rapidly in the 2D domain [7, 19], and transferring this knowledge into the 3D space.

3. Method

Our method consists of two main components (Figure 1). First, we introduce an approach for multi-object discovery from 3D data based on 2D motion, which we call DIOD-3D. Next, we design a cross-modal distillation framework (xMOD) that enables interaction between two branches, xMOD (2D) and xMOD (3D), which process 2D and 3D data, respectively, and generate pseudo-labels for the alternate modality. In the following sections, we first outline the 2DOD method that forms the foundation of our approach, before describing the two distinct aspects of our work.

3.1. Context: distilled motion-guided slot attention for 2D object discovery

The objective is to utilize automatically acquired motion information to localize mobile objects; and to generalize to other static objects within the same semantic category [1]. A recent approach specifically addresses the challenge of generalization by proposing a method that first uses as targets pseudo-labels for mobile objects, generated from optical flow. Leveraging these pseudo-labels, a distillation framework is employed to gradually expand the pseudo-labels set to include static objects identified by the model, thus covering all instances within the semantic category of interest [15]. Specifically, during an initial burn-in phase, the model processes a sequence of T frames to generate a video representation $H^t \in h \times w \times D$ at each timestep t. The features H are distributed across K slots (queries) through an attention module in two main steps: i) Attention weights W are computed between H^t and the set of slots from the previous timestep as $W^t = \frac{1}{\sqrt{D}} k(H^t) \cdot q(S^{t-1}) \in \mathbb{R}^{N \times K}$, ii) Slots S^t are updated as ${W^t}^\top v(H^t)$, where v, k, and q are three learnable projections and $N = h \times w$ [36]. To enable objects activation within the attention maps, these are supervised using a set $\mathcal{M}_{2D} = \{\mathbf{m}_l \in \{0,1\}^{h \times w} : l \in \{1, \ldots, L\}\}$ of pseudo-labels extracted from optical flow [1], with L being the number of pseudo labels available for a given image. These masks are aligned with the model-generated attention maps through Hungarian matching. The background class is isolated within a specific attention map W_{bg} using a negative log-likelihood loss function, as described in [14]. Following the burn-in, the model enters a distillation phase where it is duplicated into teacher and student models. The student model learns to discover objects through gradient back-propagation, while the teacher model is updated as an exponential moving average (EMA) of the student, ensuring gradual refinement of model capabilities. Notably, during the burn-in phase, the teacher model learns to generalize from moving objects to static objects within the same category through semantics. Distillation then allows both moving and static objects extracted from the teacher model to be presented as targets to the student model. Specifically, any connected region in one teacher's attention map $\overline{W}$ is identified as a candidate object and, if it passes a confidence test, is added to the targets for supervising the student model. For supervision, a weighted Binary Cross-Entropy (BCE) loss function is employed, where weighting is based on the confidence associated with each object segment. Alongside the teacher's predictions, the motion pseudo-labels continue to be used during the distillation phase and act as a regularization.

3.2. 3D Object Discovery

The inherent sparsity in 3D data is a challenge for tasks like object detection, in particular in the unsupervised setting of object discovery, where detailed and complete input information is crucial. Additionally, directly processing raw 3D data requires more complex and computationally intensive algorithms. To address this, 2D projections of point clouds are used to transfer data into a denser grid-structured space, manageable by efficient 2D models.

3.2.1. DIOD-3D: our approach for 3D Object Discovery

For each scene, the corresponding LiDAR-generated point cloud (*i.e.* a set of 3D points) is projected into 2D using a front-view projection, as shown in Figure 1. Let $\mathbf{I}_{\text{fv}} \in \mathbb{R}^{H' \times W' \times 4}$ be the projected 2D image matrix for a given scene. Each pixel in $\mathbf{I}_{\text{fv}}$ contains four channels: $\mathbf{I}_{\text{fv}}(i,j) = (X_{ij}, Y_{ij}, Z_{ij}, d_{ij})$ for $i \in \{1, \ldots, H'\}$ and $j \in \{1, \ldots, W'\}$, with d being the distance of the projected points from the RGB camera origin. The pixel (i,j) is assigned a fill-value vector (f, f, f, f), where f is set to 0 in this work to indicate the absence of

CVPR #18736

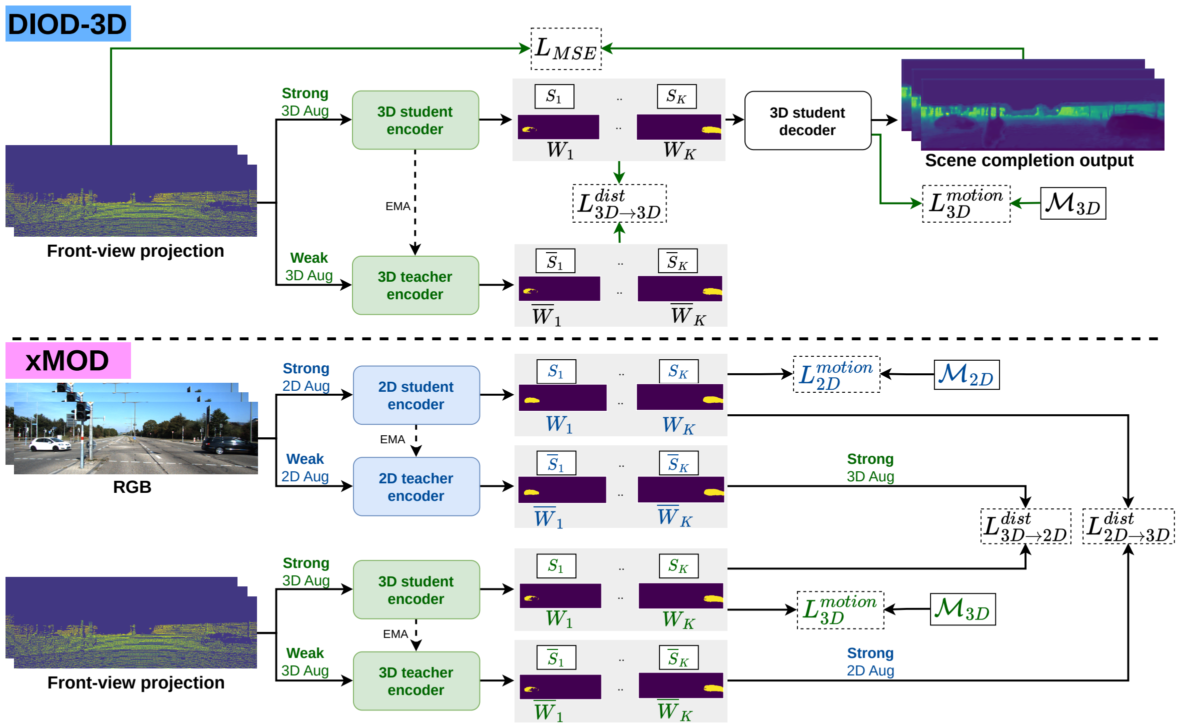

Figure 1. **Overview of the proposed approach**. **i) DIOD-3D**. At each iteration, a sequence front-view projections of point clouds is passed to the 3D teacher and student models. Attention maps from the teacher model are presented as targets to the student model through $L_{3D\to 3D}^{dist}$. An MSE objective is employed to predict the original scene from input with missing data, enabling 3D scene completion as an auxiliary task for 3DOD. **ii) Cross-modal distillation (xMOD)**. Alongside the 3D branch, sequences of RGB images are forwarded to the 2D teacher and student models. $L_{2D\to 3D}^{dist}$ means pseudo-labels from the 2D teacher model are aligned with the 3D student input and used for its supervision; $L_{3D\to 2D}^{dist}$ works similarly for 3D to 2D pseudo-labeling. Motion pseudo-labels $\mathcal{M}_{2D}$ and $\mathcal{M}_{3D}$ are used for regularization, with $\mathcal{M}_{3D}$ being the 2D motion segments with corresponding 3D points. We omit representing 2D reconstruction and 3D completion task for simplification.

an associated 3D point. This can occur either due to the LiDAR's lower resolution compared to the camera or because the camera's vertical field of view (FOV) is wider than that of the LiDAR.

Due to the inherent differences in the vertical FOV between the LiDAR and RGB camera, the motion pseudo-labels extracted from the optical flow (in the 2D domain) can occupy regions without any corresponding 3D information, particularly at the top of the projected image. This has been observed to cause model hallucinations in those regions, in the form of high-confidence noise segments. To address this, motion masks without corresponding 3D data are discarded in the motion guidance. We denote the new set of motion pseudo-labels as $\mathcal{M}_{3D}$. For each scene $\mathbf{I}_{\text{fv}}$, $\mathcal{M}_{3D}$ is a subset of the 2D pseudo-labels $\mathcal{M}_{2D}$ defined as:

$$\mathcal{M}_{\text{3D}} = \left\{ \mathbf{m}_l \in \mathcal{M}_{\text{2D}} \;\middle|\; \begin{array}{l} \exists (i,j) \text{ such that } \mathbf{m}_l(i,j) = 1 \\ \text{and } (X_{ij}, Y_{ij}, Z_{ij}, d_{ij}) \neq (f,f,f,f) \end{array} \right\}. \tag{1}$$

Let $m_{3D} \in \mathcal{M}_{\text{3D}}$ be a motion pseudo-label for the scene $\mathbf{I}_{\text{fv}}$, that matches the attention map W (Hungarian matching) learned by the student model. Motion supervision is applied via the following BCE loss:

$$\begin{split} L_{3D}^{motion}(m_{3D}, W) = -\frac{1}{N}\sum_{i=1}^{N}\Big[(1+s_{m_{3D}})m_{3D}(i)\log\big(W(i)\big) \\ + \big(1-m_{3D}(i)\big)\log\big(1-W(i)\big)\Big] \end{split} \tag{2}$$

where the confidence score $s_{m_{3D}}$ is computed as the average activation within the learned foreground map W_{fg} at the object's location in m_{3D}.

Similar to the 2D approach in [15], L_{3D}^{motion} is employed as the sole supervisory signal during the burn-in phase. During the distillation phase, each highly confident teacher-generated pseudo-label c is incorporated as a target using $L_{3D\to 3D}^{dist}(c, W)$ (same definition as L_{3D}^{motion}).

3.2.2. Scene Completion as a Pretext Task for 3DOD

Scene reconstruction has proven to be an effective pretext task in RGB images [2]. However, this conclusion does not hold for the task of 3DOD. Trying to reproduce the high and variable sparsity of LiDAR data makes scene understanding challenging, and results in sparse and less accurate predictions. Refer to the ablation study in subsection 4.6 for a quantitative evaluation of these limitations.

For this reason, we propose to rely on scene completion as a pretext task for 3DOD. Let $\mathcal{P}$ be the set of coordinates corresponding to valid projections of 3D points. We randomly remove a subset $\mathcal{P}_{\text{drop}} \subset \mathcal{P}$ from these coordinates. The objective is then to reconstruct the pixels at positions in $\mathcal{P}$ using the information from pixels at positions in $\mathcal{P} \setminus \mathcal{P}_{\text{drop}}$ (see Figure 1). The reconstruction is guided by a mean squared error loss, optimized only for valid projections of 3D points to avoid reproducing the input sparsity; and is defined as:

$$L_{\text{MSE}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left(\hat{\mathbf{I}}(i,j) - \mathbf{I}_{\text{fv}}(i,j)\right)^2 \tag{3}$$

where $\hat{\mathbf{I}}$ and $\mathbf{I}_{\text{fv}}$ are the reconstructed and original frontal projections. The previous objective enables scene completion behavior, which enhances scene understanding and segmentation.

3.3. Cross-Modal Distillation for Unsupervised 2D/3D Object Discovery

In the previous sections, we proposed a first method for real-world object discovery using LiDAR data. Our approach is based on intra-modal distillation, where both the student and teacher models receive the same 3D data. Even when these data are augmented differently, the gap between the two inputs remains limited, suggesting that the teacher's contribution to the student might be reduced in this setting. This assumption is confirmed in the ablation study presented in subsection 4.6.

In this section, we propose a cross-modal distillation framework that places the teacher and student models in two different domains: 2D and 3D modalities. Specifically, we jointly optimize two teacher-student systems, one in each modality, and enable pseudo-labeling from the teacher in one modality to the student in the alternate modality, as shown in Figure 1. Thus, the 3D teacher provides a guidance signal from the 3D domain, addressing the limitations of the 2D student in *2D-blind* scenarios (such as night scenes or fog). The 2D teacher, in turn, enhances the robustness of the 3D student in *3D-blind* scenarios such as objects with low reflectivity or highly cluttered environments.

Concretely, at each iteration, a video sequence of length T is passed through the four models (2D teacher, 2D student, 3D teacher, and 3D student), with strong modality-specific augmentations applied to the inputs of the student models. Attention maps are produced by the slot-attention module within each model and are involved in the cross-model supervision. The attention maps from the teacher models are binarized to generate object candidates as described in [15]. For simplicity, we will consider the case where $T = 1$ frame. Let D_1 and D_2 be the source and target domains, respectively, during the exchange of pseudo-labels. We denote c_{D_1} an object candidate proposed by the teacher model of domain D_1, which matches the attention map W_{D_2} of the student model from domain D_2. The inter-modal distillation objective function for the previous pair is defined as follows:

$$\begin{aligned} L^{dist}_{D_1\to D_2}(c_{D_1}, W_{D_2}) = -\frac{1}{N}\sum_{i=1}^{N}\big[(1+s_c)c_{D_1}(i)\log\left(W_{D_2}(i)\right) \\ + \left(1-c_{D_1}(i)\right)\log\left(1-W_{D_2}(i)\right)\big] \end{aligned} \tag{4}$$

D_1 and D_2 can be either 2D or 3D modalities, based on the direction of the pseudo-label exchange. Specifically:

- $L^{dist}_{2D\to 3D}(c, W)$ when the object candidate c is derived from the 2D teacher and W is a learned 3D student's attention map.
- $L^{dist}_{3D\to 2D}(c, W)$ when the object candidate c is derived from the 3D teacher and W is a learned 2D student's attention map.

The case where $D1 = D2 \in \{2D, 3D\}$ corresponds to intra-modal distillation, which is not utilized as an objective in our proposed training approach. The ablation study in section 4.6 demonstrates the ineffectiveness of this distillation compared to inter-modal pseudo-labelling.

Given the findings presented in section 3.2.2, we employ scene completion as a pretext task for the 3D branch, while the 2D branch continues to pursue a 2D scene reconstruction objective. Additionally, the motion masks $\mathcal{M}_{2D}$ and $\mathcal{M}_{3D}$ are still used as targets for the 2D and 3D branches, respectively, for regularization. Corresponding objective functions are denoted as L^{motion}_{2D} and L^{motion}_{3D}.

3.4. Late fusion of modalities

The 2D student and 3D student models, trained through cross-modal distillation, can be independently applied to a single sensor—either an RGB camera or LiDAR—depending on the specific application. To further enhance performance, we propose merging the predictions from both models for multi-sensor applications. The underlying assumption in our fusion method is that the pseudo-label exchange during cross-modal training should lead to consistent object regions between the two modalities. In contrast, inconsistent predictions are likely due to domain-specific noise. We therefore suggest using inter-domain consistency as a measure of confidence in the predictions. During inference, we propose a simple late fusion strategy by retaining the union of predictions from both models that overlap by at least a threshold value τ, while discarding predictions *unique* to only one modality.

4. Experiments

4.1. Datasets

We evaluate the proposed approach on TRI-PD [1], KITTI [12] and WOD [34] datasets. **TRI-PD** is a benchmark for 2DOD, comprising an extensive collection of highly realistic synthetic videos of driving environments. The benchmark's test set contains solely RGB images. To accommodate evaluations involving a 3D model, we introduce a new test set composed of 17 scenes with 3 camera views each, randomly extracted from the former TRI-PD training set. Point clouds for each image are computed using the GT dense depth and camera poses. In all our experiments, this test set is excluded from the training sequences. The list of KITTI frames used in 3D evaluation, as well as the list of scenes of the new TRI-PD test set, are provided in the appendix. **KITTI** is a set of benchmarks designed for computer vision tasks in road scene applications. The instance segmentation subset has been adopted in previous works as a benchmark for 2DOD. This subset includes 200 frames, of which only 142 have associated 3D information (LiDAR points). We use this new subset for evaluation in the multi-modal setting. During training, all raw-data are used without labels. **Waymo Open Dataset** (WOD) [34] is a large-scale dataset for autonomous driving, which includes 3D point clouds and 2D RGB images. Although WOD has not been traditionally used for 2DOD benchmarks, its complex, real-world scenarios are valuable for testing our unsupervised method. For our experiments, we use point clouds from the top-mounted 64-channel LiDAR, along with video frames from the front-facing camera. The training set includes approximately 800 sequences of 200 frames each, while the validation set contains 200 sequences of 200 frames each.

4.2. Metrics

Consistent with previous work on object discovery [15], we validate our approach using three metrics: foreground Adjusted Rand Index (fg-ARI), all-ARI, and F1@50. The **fg-ARI** measures the similarity between two clusterings by considering all pairs of points within the foreground area, counting pairs that are either assigned to the same cluster or different clusters in both the predicted and true clustering. Both metrics aim to evaluate the quality of the instance segmentation, considering only the foreground regions without relying on class labels. The **all-ARI** is a variation of ARI that accounts for the accurate segmentation of the image background. Both of these metrics are pixel-wise measures and do not normalize for the size of the objects, which tend to be biased toward correctly segmenting larger objects. [15] has addressed this bias by calculating an instance-wise metric, known in object detection as **F1@50**.

Modality	Method	TRI-PD		KITTI		WOD	
		all-ARI	F1	all-ARI	F1	all-ARI	F1
2D	DIOD	**66.1**	30.6	62.8	18.7	59.4	27.5
	xMOD (2D)	64.7	35.5	69.7	22.3	66.1	35.1
3D	DIOD-3D	65.1	39.6	51.6	15.5	55.3	25.6
	xMOD (3D)	65.0	37.5	58.8	18.9	62.3	31.0
Multi	xMOD (2D+3D)	64.8	**42.5**	**75.8**	**27.4**	**72.3**	**42.6**

Table 1. **Multi-modal Object Discovery**. The models resulting from our proposed approach are presented in blue. Parentheses indicate the modality used during inference. A comparison with ClusterNet [37] is provided in the supplementary materials.

4.3. Implementation details

Synthetic photo-realistic dataset (TRI-PD). Given the availability of dense depth maps, we used the camera poses to generate XYZd-formatted input and omitted the scene completion task. Both the RGB images and front-view projections were resized to (480×968). Images were augmented similarly to [15], while depth maps were transformed using data jittering, data drop, horizontal-flip and crop-resize, all with a probability 0.4. The model was trained for 300 epochs.

Real-world setting (KITTI and WOD). We forwarded RGB images to the 2D branch and front-projected 3D point clouds to the 3D branch, both using a ResNet18 [13] backbone without pre-training. The training was conducted for 100 epochs following a burn-in period, using batches of 8 input sequences of length $T = 5$. For each modality, the teacher parameters were computed as the EMA of the student with a keeping-rate 0.996. For KITTI, the motion segments used for guiding the slots' learning were extracted from RAFT optical flow [35], using the approach in [8]. For WOD, pseudo-motion segments are generated using xMOD trained on KITTI. Specific details for each branch are provided in the appendix.

4.4. Multi-modal Object Discovery

In Table 1, we present the quantitative results on the three datasets for the 2D and 3D object discovery tasks. On the TRI-PD dataset the point cloud data is very dense and contains less texture compared to RGB input, simplifying the task of object discovery. Consequently, the 3DOD baseline approach (DIOD-3D) achieves significantly higher performance than the 2DOD baseline (DIOD), with a 9-point increase in F1 score. Cross-modal training further enhances the 2D model's performance by 4.9 point, attributed to the 3D model, which experiences a 2.1-point decrease mainly due to lower precision. Detailed precision and recall results are provided in the appendix. Ultimately, late fusion of modalities yields the highest performance on this dataset, achieving an F1 score of 42.5. The sparsity of point cloud data in the KITTI and WOD datasets presents added challenges for the DIOD-3D baseline relative to the 2D baseline. Cross-modal training helps mitigate these challenges, boosting the $F1@50$ score of the 2D model by 3.6 and 7.6 points and the 3D model by 3.4 and 4.6 points on KITTI and WOD, respectively. In this context, late fusion proves highly beneficial, increasing performance by 5.1 points on KITTI and 7.5 points on WOD compared to the next best model, our xMOD (2D) branch. The discrepancy between all-ARI and F1 scores across datasets arises from the differing nature of these metrics: all-ARI is pixel-wise, while F1 score is instance-wise. This means that if the model detects a large, noisy segment, it minimally impacts the F1 score (counting as a single false positive) but lowers the all-ARI score due to many misclassified pixels. As a result, the model may perform better on TRIP-PD and Waymo in terms of F1 score, but achieve higher all-ARI on KITTI, where the effects of pixel-wise noise differ.

4.5. 2D Object Discovery

Guidance signal	Method	TRI-PD	KITTI
	DINOSAUR [30]	-	70.3
optical flow	PPMP [16]	-	51.9
flow + depth	SAVI++ [2, 10]	-	23.9
2D motion masks	Bao et al. [1]	50.9	47.1
	MoTok [2]	55.1	64.4
	BMOD [14]	53.9	54.7
	DIOD [15]	<u>66.1</u>	<u>73.5</u>
	xMOD (2D)	**68.0**	**75.5**
	BMOD* [14]	58.5	60.8
	DIOD* [15]	**69.7**	<u>72.3</u>
	xMOD* (2D)	<u>67.1</u>	**76.9**

Table 2. Evaluation of 2D object discovery in foreground regions using fg-ARI metric on the TRI-PD and KITTI test sets. Methods using an encoder pre-trained with DINOv2 [25] are marked with *.

In previous experiments, we introduced a baseline in 3DOD, which was enhanced through cross-modal training and late fusion during inference. We emphasize that these results were achieved on the new KITTI and TRI-PD test sets, with available 3D data (see section 4.1). In this section, for an objective comparison with previous methods in 2DOD, we evaluate the 2D branch of our model (xMOD (2D)) on the conventional test sets of the studied benchmarks. We use the most widely employed metric in 2DOD, ie. fg-ARI, for evaluation. The results in Table 2 show that xMOD (2D) branch also benefits from cross-modal training, exploiting *readily* available 3D data.

4.6. Ablation studies

Early fusion vs. late fusion. We explored two fusion strategies for integrating RGB images and front-projected point clouds. Early fusion combines the modalities at the input level with concatenation across the channel axis, while late fusion, as explained in subsection 3.4, refines segmentation by cross-examining predictions from the two modalities. With an overlap threshold of $\tau = 0.3$, late fusion significantly outperformed early fusion by 8 F1 points after cross-modal training such as shown in Table 3.

	Method	F1@50
end of burn-in	2DOD	9.3
	3DOD	8.6
	Early fusion	**12.8**
final setting	Early fusion	19.4
	Late fusion	**27.4**

Table 3. Early *vs.* late fusion.

Impact of scene completion. We evaluated using the pretext task of scene completion, where the model estimates point positions based on neighbors. As shown in Table 4, this task helped our method discover objects, resulting in a 7.7-point increase in F1 score.

Scene completion	all-ARI	F1@50
✗	63.7	19.7
✓	**75.8**	**27.4**

Table 4. Ablation study on the scene completion pretext task on KITTI dataset, using the late fusion strategy.

Impact of intra-modal distillation. Unlike prior work, we focused solely on cross-modality distillation losses, without applying intra-modality losses between the teacher and student of the same modality. Experiments showed (Table 5) that adding intra-modality losses decreased performance slightly by 0.6 F1 points. This suggests the intra-modality loss may act as a redundant constraint, hindering the model's ability to learn valuable features from the other modality through cross-modal losses.

Limitations on nearby and distant objects. From the qualitative analysis in Figure 2 and Figure 3, we observe that segmentation quality depends on the object's distance from the camera, affecting its size in the 2D image. Based

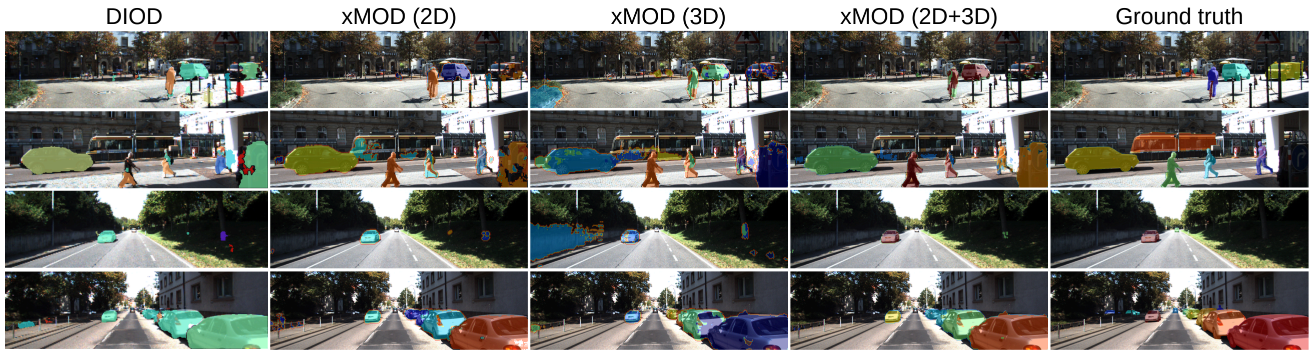

Figure 2. Qualitative comparison of our method with state-of-the-art approach DIOD [15], the cross-modal branches xMOD (2D), xMOD (3D) separately and the final result after fusion xMOD (2D+3D) in real-world scenes (KITTI [12]). Parentheses indicate the modality used during inference. Each colored mask represents the content of one slot. The segmentations are displayed above the RGB image for visualisation purposes only. Improvements in xMOD are especially evident in pedestrian detection and background noise suppression.

Losses		F1@50
Cross-modal	Intra-modal	
✓	✓	26.8
✓		**27.4**

Table 5. Analysis of the impact of intra-modal losses ($L^{dist}_{2D\to 2D}$ and $L^{dist}_{3D\to 3D}$) on object discovery in real-world (KITTI).

on this, we split the test set into three distance-based subsets and measure the F1 score for each in Table 6. For objects within 10 meters, which are usually cropped in images and front view projections (see the example of the red car at the top right of Figure 3), the F1 score decreases. Mid-distance objects (10-30 meters), which are clearly visible and densely represented in the point cloud, achieve a higher F1 score. Beyond 30 meters, objects are small and LiDAR data is sparse, dropping the F1 score to 7.2 due to low recall. A potential solution is re-injecting object instances from the high-confidence range into the two other ranges to enhance model sensitivity in these areas.

Distance (m)	AvgPts/Obj	F1@50	Precision	Recall
0-10	2640	21.7	68.2	12.9
10-30	941	46.4	85.7	31.8
30-70	134	7.2	29.5	4.1
0-70	1105	27.4	56.9	18.0

Table 6. Object discovery performance on KITTI on 3 subsets of objects defined by their distance to the camera. AvgPts/Obj is the average number of points per object in the subset.

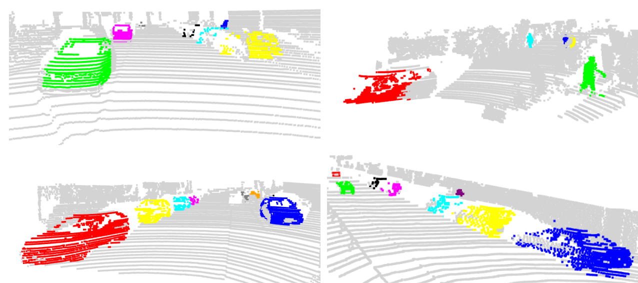

Figure 3. 3D visualization of predictions produced by xMOD (2D+3D). The background is displayed in gray and each colored mask represents the content of a distinct slot.

5. Conclusion

In this work, we first presented a method for discovering multiple objects in 3D data. Our approach builds on the latest advancements in motion-guided object discovery in images and introduces necessary adjustments to handle sparse 3D point cloud data from LiDAR sensors. In particular, we found that scene completion is a well-suited pretext task for 3DOD, as scene understanding is critical in this unsupervised setting. We also proposed a cross-modal distillation training method, where two branches, each processing a distinct modality—2D or 3D—exchange pseudo-labels during training. The experiments showed advantages for both modalities, which can be attributed to the limitations of each sensor when used independently. To further investigate the multi-modal setting, we proposed a late fusion strategy during inference, using multi-modal consistency as a confidence criterion. The high precision of this approach at medium distances opens perspective for instance injection methods to improve the model reliability in more challenging conditions. Future work could also explore the use of multi-scale supervision—beyond the latent space—to address the reduced sensitivity observed for small objects.

References

[1] Zhipeng Bao, Pavel Tokmakov, Allan Jabri, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Discorying object that can move. In *CVPR*, 2022. 1, 2, 3, 6, 7

[2] Zhipeng Bao, Pavel Tokmakov, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Object discovery from motion-guided tokens. In *CVPR*, 2023. 1, 2, 5, 7

[3] Igor Bogoslavskyi and Cyrill Stachniss. Efficient online segmentation for sparse 3d laser scans. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 85:41–52, 2017. 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2

[6] Jun Cen, Peng Yun, Junhao Cai, Michael Wang, and Ming Liu. Open-set 3d object detection. pages 869–878, 2021. 2

[7] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16901–16911, 2024. 3

[8] Achal Dave, Pavel Tokmakov, and Deva Ramanan. Towards segmenting anything that moves. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019. 3, 6

[9] Ayush Dewan, Tim Caselitz, Gian Diego Tipaldi, and Wolfram Burgard. Motion-based detection and tracking in 3d lidar scans. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 4508–4513. IEEE, 2016. 2

[10] Gamaleldin F. Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael C. Mozer, and Thomas Kipf. SAVi++: Towards end-to-end object-centric learning from real-world videos. In *Advances in Neural Information Processing Systems*, 2022. 1, 7

[11] Christian Fruhwirth-Reisinger, Wei Lin, Duvsan Mali'c, Horst Bischof, and Horst Possegger. Vision-language guidance for lidar-based unsupervised 3d object detection. *ArXiv*, abs/2408.03790, 2024. 2

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 6, 8

[13] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 6

[14] Sandra Kara, Hejer Ammar, Florian Chabot, and Quoc-Cuong Pham. The background also matters: Background-aware motion-guided objects discovery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1216–1225, 2024. 2, 3, 7

[15] Sandra Kara, Hejer Ammar, Julien Denize, Florian Chabot, and Quoc-Cuong Pham. Diod: Self-distillation meets object discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3975–3985, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[16] Laurynas Karazija, Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised multi-object segmentation by predicting probable motion patterns. *Advances in Neural Information Processing Systems*, 35:2128–2141, 2022. 7

[17] Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional Object-Centric Learning from Video. In *International Conference on Learning Representations (ICLR)*, 2022. 1, 2

[18] Yancong Lin and Holger Caesar. Icp-flow: Lidar scene flow estimation with icp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15501–15511, 2024. 2

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[20] Xingyu Liu, Charles R Qi, and Leonidas J Guibas. Flownet3d: Learning scene flow in 3d point clouds. *CVPR*, 2019. 2

[21] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, pages 11525–11538. Curran Associates, Inc., 2020. 2

[22] Katie Luo, Zhenzhen Liu, Xiangyu Chen, Yurong You, Sagie Benaim, Cheng Perng Phoo, Mark Campbell, Wen Sun, Bharath Hariharan, and Kilian Q Weinberger. Reward finetuning for faster and more accurate unsupervised object discovery. *Advances in Neural Information Processing Systems*, 36:13250–13266, 2023. 2

[23] Himangi Mittal, Brian Okorn, and David Held. Just go with the flow: Self-supervised scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[24] Mahyar Najibi, Jingwei Ji, Yin Zhou, Charles R Qi, Xinchen Yan, Scott Ettinger, and Dragomir Anguelov. Motion inspired unsupervised perception and prediction in autonomous driving. In *European Conference on Computer Vision*, pages 424–443. Springer, 2022. 2

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 7

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 1

[28] Sadra Safadoust and Fatma Güney. Multi-object discovery by low-dimensional object motion. In *ICCV*, pages 734–744, 2023. 1

[29] Jenny Seidenschwarz, Aljosa Osep, Francesco Ferroni, Simon Lucey, and Laura Leal-Taixé. Semoli: What moves together belongs together. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14685–14694, 2024. 3

[30] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Scholkopf, Thomas Brox, and Francesco Locatello. Bridging the gap to real-world object-centric learning. *ArXiv*, abs/2209.14860, 2022. 1, 2, 7

[31] Oriane Sim'eoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick P'erez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *BMVC*, 2021. 1

[32] Oriane Siméoni, Éloi Zablocki, Spyros Gidaris, Gilles Puy, and Patrick Pérez. Unsupervised object localization in the era of self-supervised vits: A survey. In *IJCV*, 2024. 2

[33] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 2

[34] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 6

[35] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2, 6

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 3

[37] Yuqi Wang, Yuntao Chen, and ZHAO-XIANG ZHANG. 4d unsupervised object discovery. *Advances in Neural Information Processing Systems*, 35:35563–35575, 2022. 2, 6

[38] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022. 1

[39] Yuang Wang, Xingyi He, Sida Peng, Haotong Lin, Hujun Bao, and Xiaowei Zhou. Autorecon: Automated 3d object discovery and reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21382–21391, 2023. 2

[40] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. *arXiv preprint arXiv:2405.14793*, 2024. 2

[41] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. In *Advances in Neural Information Processing Systems*, 2022. 3

[42] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, 2021. 3

[43] Yurong You, Katie Luo, Cheng Perng Phoo, Wei-Lun Chao, Wen Sun, Bharath Hariharan, Mark E. Campbell, and Kilian Q. Weinberger. Learning to detect mobile objects from lidar scans without labels. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1120–1130, 2022. 2

[44] Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Object-centric learning for real-world videos by predicting temporal feature similarities. In *NeurIPS*, 2023. 2

[45] Lunjun Zhang, Anqi Joyce Yang, Yuwen Xiong, Sergio Casas, Bin Yang, Mengye Ren, and Raquel Urtasun. Towards unsupervised object detection from lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9317–9328, 2023. 2