The Background Also Matters: Background-Aware Motion-Guided Objects Discovery

Sandra Kara Hejer Ammar Florian Chabot Quoc-Cuong Pham

Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
{firstname.lastname}@cea.fr

Abstract

Recent works have shown that objects discovery can largely benefit from the inherent motion information in video data. However, these methods lack a proper background processing, resulting in an over-segmentation of the non-object regions into random segments. This is a critical limitation given the unsupervised setting, where object segments and noise are not distinguishable. To address this limitation we propose BMOD, a Background-aware Motion-guided Objects Discovery method. Concretely, we leverage masks of moving objects extracted from optical flow and design a learning mechanism to extend them to the true foreground composed of both moving and static objects. The background, a complementary concept of the learned foreground class, is then isolated in the object discovery process. This enables a joint learning of the objects discovery task and the object/non-object separation. The conducted experiments on synthetic and real-world datasets show that integrating our background handling with various cutting-edge methods brings each time a considerable improvement. Specifically, we improve the objects discovery performance with a large margin, while establishing a strong baseline for object/non-object separation.

1. Introduction

Deep learning-based approaches have demonstrated significant success in addressing a wide array of computer vision tasks [5]. However, the high performance of these methods heavily relies on the availability of abundant labeled data: sparse labels or compromised label quality impairs the effectiveness of supervised approaches [37]. This limitation becomes challenging when tackling dense tasks like segmentation, where the acquisition of accurate labels requires considerable resources. This observation has motivated numerous studies to propose alternative architectures, including weakly supervised [19, 40], semi-supervised [24, 39], and unsupervised methods [27, 36], aiming to tackle vision tasks with minimal supervision.

In this work, we address the task of localizing objects in videos without the use of human annotations. This task, which is commonly approached as a segmentation problem [2, 8, 17], is particularly suited for video data due to its inherent advantages over static images. Specifically, the motion information derived from videos offers a means to obtain *free* pseudo-labels for moving objects localization. This makes the motivation even stronger to explore self-supervised methods, capable of leveraging motion cues. Moreover, the ambiguity surrounding the definition of objects, which remains a challenge for object discovery in images [15, 34], can be addressed in the video data. Specifically, relying on motion cues to localize objects provides, by design, a definition of what an object is: we consider as object any entity that could exhibit an independent motion. This definition is even in line with human perception as demonstrated in [30]: in our perception, we divide the observed scene into parts that are capable of moving while remaining connected. Some recent approaches draw inspiration from and build upon this result, to address moving objects localization [6].

Recently, object-centric learning architectures have demonstrated a significant potential for solving the object discovery task [21]. It emerges as a new deep learning based approach to decompose the input image into *meaningful* regions, in an unsupervised way. Although initially validated on simple synthetic image datasets, many subsequent works have proposed variants of this architecture to scale it to video data as well as to more complex scenarios. Those primarily concentrate on modifying the reconstruction space (optical flow [17], depth [8]) and enhancing the encoder's capability [29]. More recently, some works introduced the use of motion cues to direct the learning process of slots and provided evidence of the

Figure 1. **Illustration of the addressed problem.** Results from [1] showing background over-segmentation in both settings: unsupervised (middle) and using motion supervision (right). When the ground truth is not available, foreground objects cannot be automatically separated from the background.

effectiveness of this guidance signal in solving objects discovery in complex scenarios [1, 2].

Our work fits into this same line of research, but tackles a specific problem that is not covered by existing methods, namely the background control in the object discovery task. Previous works did not focus on learning the background pattern, which results in the background being split across the slots into *noise* regions, as illustrated in figure 1. This over-segmentation of the scene is not even penalized by the commonly used metrics, since the segmentation quality is evaluated on foreground regions only. However, in a real-world setting where ground truth is not available, it is impossible to distinguish between objects and background segments. The aim of this work is therefore to learn this object/non-object boundary, while solving the multiple objects discovery task.

We propose to leverage motion cues to jointly learn the multiple objects discovery and the objectness task (foreground/background separation). The motion cues are moving objects masks, extracted from optical flow. For the first task, each motion mask is used to guide one slot's attention. For the second, we propose to learn the generalization from the moving foreground (summed motion masks) to the *true* foreground containing both moving and static objects. The complementary mask, which is the background, is positioned within a specific slot, competing with all others, to isolate its distinct pattern.

Our contributions can be summarized as following:

- We propose BMOD (Background-aware Motion-guided Objects Discovery), a simple yet effective learning mechanism for modeling the background while solving the object discovery task. To the best of our knowledge, this is the first method that addresses these two tasks concurrently, without the need for human supervision.
- We demonstrate that modeling the background not only allows for a more precise objects discovery (automatic filtering of noise segments), but also improves the localization of foreground objects. This validates our insight that controlling the background reduces the amount of noise captured by the slots, making it easier for the model to learn the *object pattern*.
- We establish a new baseline for the objectness learning in the object discovery task. For the first time, we introduce the computation of suitable metrics for evaluating the objectness learning task (Jaccard score), or by evaluating the two tasks together (all-ARI).
- We demonstrate through comprehensive experiments the effectiveness of our method on the challenging TRI-PD dataset as well as the real-world dataset KITTI. The experiments show that multiple cutting-edge methods derive significant advantage from our objectness learning mechanism, without increasing architectural complexity. Moreover, we show that our method, when enhanced with rich features from the recent DINOv2 [23] pretraining, brings about a considerable performance leap. This provides evidence of the representation bottleneck in current methods, which is overcome through the use of improved features.

2. Related work

2.1. Objects discovery in images

Object discovery in images is the task of localizing objects without the use of human annotations. The inception of this task was marked by heuristic-based object proposal methods, which relied on an over-segmentation of the image and various similarity measures to merge *similar* regions hierarchically [25, 32, 41]. Due to their very low precision, utilizing these object candidates in an unsupervised setting has been challenging.

In the era of deep learning, object discovery has profited from deep features, either derived from CNNs learned through the ImageNet classification task [33–35], or from the more recent self-supervised pretraining, in particular of vision transformers (ViTs) [15, 27, 36]. In the first category, methods typically aim at discovering the dataset-structure, with the most connected/similar object proposals becoming the top object candidates. In the second category, methods are mostly motivated to investigate unsupervised clustering in the space of self-supervised features, given the segmentation properties exhibited by ViTs [4]. In both categories, the methods solely rely on the semantic information learned within the image modality, which limits their ability to separate object instances.

A recent group of methodologies, known as compositional generative models, has emerged as a deep learning-based alternative to the classical clustering methods [20]. Notably, MONet [3] employs an attention mechanism to focus on individual scene parts. Both the input image and

the attention map are then passed into a variational auto-encoder module, to only reconstruct the highlighted scene part/object in the corresponding mask. IODINE [10] replaced the one-pass attention mechanism with an iterative inference to refine the understanding of the image over multiple steps. In this same category, SCALOR [13] adapted the generative process to a larger number of objects, while Slot-Attention [21] proposed a more efficient object discovery architecture with a single image encoding step. [21] discovers objects by enforcing the disentanglement within the latent space of an auto-encoder architecture.

All previous methods, whether based on heuristics or deep learning, suffer from the ambiguity of object definition. This limitation prevents both the design of a definition-based method and the establishment of objective evaluation criteria. Efforts are now being directed towards video data, which provide the means for a more generalized object definition (see section 2.2). In this work, although we focus on the analysis of video data, we provide comparisons with the latest image-based methods, applied to individual frames.

2.2. Objects discovery in videos

In this work, we address the problem of discovering objects in videos, which is a distinct task from the video object segmentation (VOS). The latter is more about motion segmentation, with as objective to localize a salient moving object within a video [38]. The task we address, in contrast, consists in localizing objects that are capable of moving, even when they remain static in the analyzed sequence.

Object discovery in videos emerges as a promising research area, largely driven by the inherent motion information in videos, compared to static images. The motivation to exploit video data for localizing objects is not new; the earliest methods typically selected regions of interest from object candidates as spatio-temporal tubes, maximizing similarity across videos while maintaining temporal consistency [18].

The recent advent of the slot-attention architecture [21], recognised as a promising solution for object discovery, has motivated many efforts to scale it to video data. SAVI [17] and Karazija et al. [16] incorporated optical flow as a more task-appropriate reconstruction space for the targeted segmentation task. [17] also proposed the use of weak supervision on the initial frame, such as the centers of objects to be tracked throughout the sequence. By design, [17] presents the limitation of localizing moving objects only. SAVI++ [8], in contrast, also localizes static objects through the reconstruction of the more generic depth signal. [8] also demonstrated the potential of data augmentations, often under-explored in unsupervised settings. Among methods that utilize motion cues for object discovery, Bao et al. [1] introduced an explicit guidance for slots learning, using moving-object masks derived from optical flow. STEVE [29] a concurrent work, investigated the use of a more powerful transformer decoder. Building upon the findings of these preceding methods, MoTok [2] proposed a more powerful motion-guided slot attention architecture through a tokenized reconstruction space.

In the previous methods, attention was only allowed to the discovery of foreground objects, without considering a proper background modeling. Our insight, however, is that a proper background modeling prevents the presence of noise regions captured by each slot, which favors the learning of the object structure. We propose in this work a complete motion-guided object discovery architecture to jointly learn the multiple objects localization task and the foreground/background separation.

2.3. Unsupervised background segmentation

Early attempts to solve the task of foreground/background separation focused on the image modality. In simple scenarios with a *neat* background, methods typically relied on thresholding or binary clustering in the color space or other hand-crafted features [14].

A more recent category, known as saliency detection methods, aim to extract a salient foreground from the background in an unsupervised way. In particular, LOST and TokenCut [27, 36] used deep features from pre-trained vision transformers (ViTs) [4]. LOST defined object regions as the patches least correlated with the whole image, while TokenCut investigated applying spectral clustering [26] to self-supervised ViTs features. More recently, FOUND [28] proposed to discover the background as the class containing the least activated patch in ViT activation maps, and then refine it using a lightweight segmentation head. This method was presented as a way of overcoming the ambiguity of object definition. However, we believe that the problem of object definition remains valid for the background class, since the two are complementary semantic concepts.

In the video modality, the binary segmentation that received significant attention was motion segmentation. Active benchmarks on this subject have been established under the terminology of Video Object Segmentation (VOS) [38]. While this provides a well-defined criterion for object identification (i.e. moving objects), we believe this definition is restrictive. Indeed, without also localizing the static objects, we can only achieve a limited understanding of the scene.

Our method, in contrast, proposes by design a foreground/background separation, where the targeted foreground is composed of both moving and static objects. The robustness of our method in complex scenarios is ensured by the use of motion cues, extracted from optical flow, which is typically insensitive to an increasing background complexity in the color space. Moreover, we also decompose the foreground class into object instances, which is not covered by the previous background segmentation methods.

3. Method

3.1. Context: motion guided slot-attention for objects discovery

Since our method is based on a slot attention architecture [21] involving the use of motion information [1], we first briefly describe these two approaches below.

The slot attention architecture [21] has been proposed as a deep learning-based alternative for the classical unsupervised clustering methods [20]. It consists of an auto-encoder architecture with a latent space that is partitioned into embedding vectors called slots. The architecture competes among these slots to provide a comprehensive explanation for the input image. The mechanism for partitioning the image is encouraged by the use of a small decoder: each slot is individually passed through the decoder. The small decoder being unable to explain the whole scene from one slot, this compels the features to be split across the slots, encouraging image partitioning into *meaningful* regions.

In our method, we build upon a recent variant of slot attention that exploits motion cues to guide slots learning [1]. Concretely, the method receives as input a sequence of T video frames. Each frame is passed through a CNN encoder for features extraction. Features from the T frames are then combined using a convGRU module to get spatio-temporal information H^t, for each frame I^t. This representation is then assigned to K slots through the attention module. Specifically, given k, q, v three learnable linear projections, attentions between features H and slots S are computed as $W = \frac{1}{\sqrt{D}} k(H) \cdot q(S) \in \mathbb{R}^{N \times K}$, where N is the feature maps size and D the dimension of features after projection. Attentions are used to update the current slot state $S^t = W^t v(H^t)$, where W^t are computed using the slot state at frame I^{t-1}. [1] introduced the use of motion guidance by assuming access to M motion masks for the sequence of T frames. The masks are resized to match the dimensions of the attention maps W and subsequently paired with them via a bipartite matching algorithm. Motion supervision then occurs between these pairs of masks m and the learned attention maps W. The method shows that introducing motion cues replaces the initial inductive bias about individual slot decoding, which reduces memory demands. Although this architecture showed generalization ability to objects without corresponding masks, it still suffers from the object/non-object ambiguity, since slots with no supervision may contain either objects or background regions. This motivates our work which we describe in next sections.

3.2. Modeling the background class using motion cues

Illustrated in figure 2, our method receives as input T video frames $I^t \in \mathbb{R}^{h \times w \times 3}$. Spatio-temporal representation $H^t \in \mathbb{R}^{h' \times w' \times D'}$ for each frame is extracted following the process in [1], with D' the dimension of features output by the convGRU module. These features are then forwarded to the attention module where we propose to jointly learn the object discovery task and the background modeling. The objective is to force background regions to occupy one single slot's attention map, instead of being randomly split across multiple slots. We denote S_{bg} the slot dedicated to the background class and $W_{bg} \in \mathbb{R}^{h' \times w'}$ the corresponding attention map. It is important to note that the motion segments cannot directly provide information on the positions of background regions, since the complement of these masks also contains the static objects we aim to localize (see ablation study in section 5.1). Instead, these masks can be used as samples of what the object of interest looks like. Therefore, we propose to learn the background class by compelling its complementary mask $W_{fg} = 1 - W_{bg}$ to contain the *true* foreground class with both moving and static objects. The complementary background mask W_{bg} will thus contain all remaining, non-object regions. On the other hand, the softmax operation applied to the attention maps W ensures their complementarity, preventing the background class from appearing in other *object* slots. Note that W_{fg} is an auxiliary attention map only used in the training phase to help the background modeling. It is not involved in the object discovery nor the image reconstruction task.

We formulate the foreground modeling as one-class learning problem, since only the positive class is known (some moving objects masks). This paradigm is commonly used for binary classification tasks [12]. In our proposed foreground modeling, the samples to classify are the pixels positions in W_{fg}. The positive class corresponds to pixels in motion (e.g. moving cars), which we propagate to also capture static objects of the same semantic class (e.g. parked cars). For a given frame I^t, the corresponding moving foreground mask is denoted $m_{fg} = \sum_{c=1}^{C} m_c$, where C is the number of motion masks available for frame I^t. In order to compel all objects regions to be activated in W_{fg} , we use the following negative log likelihood (NLL) loss:

$$L_{NLL,reg}(m_{fg}, W_{fg}) = -\frac{1}{N} \sum_{i=1}^{N} m_{fg}(i) \log(W_{fg}(i)) + \frac{\alpha}{N_s} \sum_{j=1}^{N_s} W_{fg}(j) \tag{1}$$

where $N = h' \times w'$ is the size of W_{fg}, N_s the number of pixels with no motion information in m_{fg} and α a weighting hyper-parameter. The first term in equation 1 is the NLL loss that forces all motion segments to be contained in W_{fg}, encouraging generalization to visually similar regions (static objects). We can easily predict the collapse

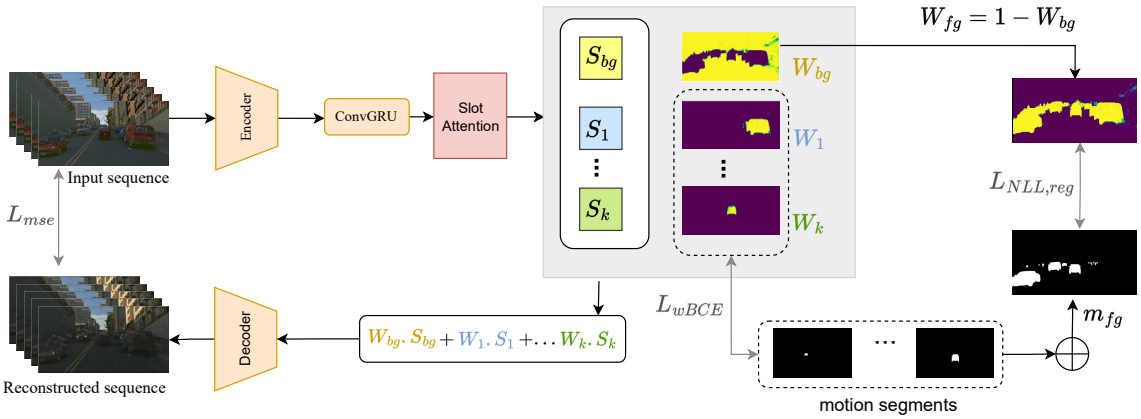

Figure 2. **Pipeline of the proposed method.** The input sequence is encoded into a spatio-temporal representation, which is then forwarded to the slot attention module. This produces a set of slots along with their respective attention maps. The separate motion masks individually supervise the attention of *objects* slots, while their sum (m_{fg}) is generalized to form the *true* foreground, using the $L_{NLL,reg}$ loss. The complement of the learned foreground class is assigned to a specific attention slot W_{bg} so as to isolate the background pattern. Finally, the sum of slots, weighted with their respective attention maps, is decoded to reconstruct the input sequence. B/W masks represent binary supervision masks derived from motion, while masks shown in the viridis colormap are learnable attention maps.

that would occur if only this first term is used: the model would converge towards a trivial solution by activating the entire map W_{fg}, which is not the desired behavior. We rather want the model to only activate objects regions (moving and static) in W_{fg}. For this, we add as a regularization term in 1 the average activation within the unlabeled regions of m_{fg} (i.e. where m_{fg} is 0), so as to constraint the model confidence in non-object regions. Given a batch size B and T frames per sequence, the final fg/bg loss is defined as follows:

$$L_{fg/bg} = \frac{1}{BT} \sum_{b=1}^{B} \sum_{t=1}^{T} L_{NLL,reg}(m_{fg}^{b,t}, W_{fg}^{b,t}) \quad (2)$$

3.3. Background-aware motion guided objects discovery

Similar to [1], the object discovery task is learned through the motion guidance of slots learning: a bipartite matching is performed to associate the motion masks to some of the K *objects* attention maps (excluding the slot assigned to the background). These receive a supervision using the corresponding mask and the Binary Cross-Entropy (BCE) objective function. Different from [1], our method includes a dedicated attention map specifically for the background class, which competes with other slots, while being predominant. This would bias the model towards activating most regions in the background slot, resulting in some objects being lost, specially small ones. To avoid this, we use a weighted BCE loss denoted L_{wBCE}, where the weights are set automatically and depend on the object size. Given a motion mask m containing one moving object, which matches the predicted attention map W, L_{wBCE} between the two is defined as follows (for the sake of simplification, we denote the i-th element of m and W as m_i and W_i respectively) :

$$L_{wBCE}(m, W) = \frac{1}{N} \sum_{i=1}^{N} (-(2 - r).m_i \log(W_i) - (1 - m_i) \log(1 - W_i)) \quad (3)$$

where r is the ratio of the number of object pixels to the whole mask size and is computed as $\frac{1}{N} \sum_i m_i$. In the above, the first loss term is assigned a dynamic weight which depends on the size of the object and varies between 1 and 2: the smaller the object in m , the more the model is encouraged to activate its corresponding pixels in W. This weighting has proven effective in maintaining the objects discovery performance, even in the presence of a predominant class (the background) competing with other *object* slots (see ablation study in section 5.1).

Finally, the learned slots are broadcasted into 2D maps. The sum of the slots, weighted each by its corresponding attention map, are decoded to reconstruct the input frame. This dense pretext task ensures the activation of all image regions, which further encourages the generalization to non-moving objects. It is learned using a mean squared error (mse) loss between the original and reconstructed video sequence. The final loss is defined as follows:

$$L = L_{mse} + L_{wBCE} + L_{fg/bg} \quad (4)$$

4. Experiments

We conduct our experiments on two video object discovery benchmarks: ParallelDomain (TRI-PD) [1] and KITTI [9]. We further demonstrate the generalizability of the proposed background learning mechanism by integrating it into another state-of-the-art method [2]. This comparison is conducted under two different settings on TRI-PD, which differ in the source of the motion masks. In the unsupervised setting, referred to as *estimated* in the results tables, these masks are derived from optical flow (see section 4.3). In the second setting denoted *gt*, ground-truth instance masks of moving objects are used as guidance signal. Results in this setting provide an upper-bound for the unsupervised one.

4.1. Datasets

ParallelDomain (TRI-PD): Introduced by [1], TRI-PD is a recent benchmark for objects discovery in urban driving scenarios. This is a challenging dataset, composed of dense, photo-realistic scenes, which also provides useful support for a variety of visual tasks, as it includes diverse semantic and instance-level annotations. Following [1], we train our object discovery models on a set of 924 video clips, each 200 frames long. Evaluations are performed on a separate test set of 51 video sequences.

KITTI is a real-world video dataset of urban scenes scenarios, and an active benchmark for various perception tasks. We use for training all raw-data from the KITTI benchmark (without annotations), totalling 147 videos. Following previous works [1, 2], evaluation is conducted on the instance segmentation subset of KITTI, composed of 200 frames.

4.2. Metrics

fg-ARI and **all-ARI**: The Adjusted Rand Index (ARI) is a measure used to quantify the similarity between two clusterings (*gt* and *predicted*) in a permutation-invariant way. In the literature of object discovery, studies typically compute the fg-ARI, which stands for ARI in foreground regions. This metric does not account for the segmentation quality in the background regions. We introduce in this paper the computation of the more suitable all-ARI metric, by also incorporating the background class into the ground-truth clusters. In all-ARI, both the foreground objects discovery and the quality of background segmentation are evaluated.

Jaccard score: The Jaccard score is calculated as the ratio of the size of the intersection to the size of the union of two label sets: the ground truth and the predicted labels. We use this metric to assess the quality of foreground/background classes separation. The final score for each of the two classes is the average Jaccard Score accross all frames.

4.3. Implementation details

Base setting (BMOD): In this setting, we use a resnet18 [11] encoder, without pre-training. For a fair comparison, we follow the same training schedule as the method in which we incorporate our background handling mechanism [1, 2]. Particularly, we use a batch size of 8 with input sequences of length $T = 5$. Frames are resized to (480×968) and (368×1248) for TRI-PD and KITTI respectively. The regularization strength α, involved in the background modeling, is set to 0.2 for TRI-PD and 0.4 in KITTI dataset. In the unsupervised setting, we use the same motion masks as previous methods [1, 2]. These are generated using the approach proposed in [7] which maps the optical flow to instance masks. The optical flow is computed using RAFT [31] and the mapping is learned on the synthetic dataset FlyingThings3D [22]. Following previous methods, models trained on KITTI are initialized with pretraining on TRI-PD dataset, using estimated motion masks. Evaluation on TRI-PD is conducted following the protocol in [1] where windows of size T frames (same size as during training) are successively passed to the model. In KITTI dataset, since the test frames are not temporally linked, evaluation is conducted on each frame individually.

Enhanced setting using self-supervised pretraining (BMOD*): In this setting, we replace the resnet18 encoder with a ViT-S/14 pretrained using the recent DINOv2 method [23]. For this, we resize the input frames dimensions to adapt to the ViT patch size. Input sizes become (490×980) and (378×1260) for TRI-PD and KITTI respectively. We tested integrating the multi-scale features described in DINOv2 paper. Specifically, we extract features from the last 4 layers of the ViT model, which we concatenate, getting new embedding vectors of size 384×4, with spatial dimensions down-sampled with a factor 14. Before passing these features maps to the convGRU, we upscale them to match the spatial dimensions yielded by the resnet18 encoder, resulting in a down-sampling factor of 4.

4.4. Unsupervised objects discovery

We recall that the primary objective of our work is to enable background handling, for a more precise objects discovery. The results from tables 1 and 2 show that this modeling also improves the localization of foreground objects (e.g. we observe $+3\%$ improvement in fg-ARI with BMOD([1]) in the unsupervised setting). This validates our assumption that isolating the background minimizes the presence of random segments in the learned attention maps, which favors a proper learning of the object structure. Although we observe in one test (BMOD([2]), *estimated*, table 1), a slight decrease in fg-ARI (-0.9), our approach brings a significant gain of performance on the more complete all-ARI metric ($+20.2$, table 3). Our results are further improved when using self-supervised pretraining (**BMOD***). This is well-justified given the rich semantic and depth information contained within these features [23]: objects are more easily captured as regions of independent motion, consistent

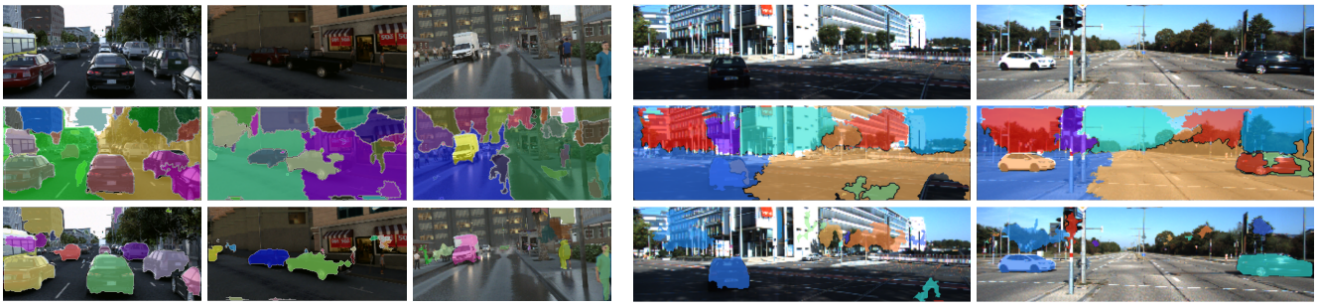

Figure 3. **Qualitative comparison under the unsupervised setting on TRI-PD (left) and KITTI dataset (right).** By row: the input frame, results of [1] showing both objects and noise segments (indiscernible given the lack of confidence criteria in the unsupervised setting), our segmentation result showing the noise reduction through background modeling.

Guidance signal	Method	fg-ARI
-	SlotAttention [1, 21]	10.2
-	MONet [1, 3]	11.0
-	SCALOR [1, 13]	18.6
-	IODINE [1, 10]	9.8
-	MCG [1, 25]	25.1
gt motion	Bao et al. [1]	59.6
	BMOD ([1])	<u>74.1</u>
	BMOD* ([1])	**83.0**
	MoTok [2]	76.3
	BMOD ([2])	<u>81.5</u>
	BMOD* ([2])	**87.5**
estimated motion	Bao et al. [1]	50.9
	BMOD ([1])	<u>53.9</u>
	BMOD* ([1])	**58.5**
	MoTok [2]	<u>55.1</u>
	BMOD ([2])	54.2
	BMOD* ([2])	**60.9**

Table 1. Evaluation of objects discovery performance on TRI-PD dataset under two settings. Best results are put in **bold**, second best <u>underlined</u>. BMOD([X]) stands for our method built upon the approach in [X] and BMOD* is our method enhanced with features from DINOv2 [23] pretraining.

depth, and similar semantics.

4.5. Background modeling using motion cues

In this section, we highlight our main contribution, namely learning the object/non-object boundary without human supervision. We use two distinct metrics to evaluate this task, all-ARI and Jaccard Score (see section 4.2). We recall that in our method, the background slot is known since it is constrained by design. Calculation of the Jaccard Score is therefore straightforward. For previous methods, however, no information of the background class is available. For a fair comparison, we consider in these methods as background the largest segment returned in all slots. Even so, the results in table 3 show the clear improvement brought by our method in the two tested settings: with motion supervision and unsupervised. Particularly, incorporating our training mechanism into [1] brings considerable all-ARI improvement of $+22.3$ on TRI-PD under the unsupervised setting, and a further enhancement of $+13.5$ on KITTI dataset. The other observation we can draw is that, for both fg-ARI and the fg/bg separation tasks, a wide gap remains between the two settings *gt* and *estimated* (the upper-bound results being very high), suggesting strong potential for improvement by addressing the quality of pseudo-labels.

Method	fg-ARI
SlotAttention [1, 21]	13.8
MONet [1, 3]	14.9
SCALOR [1, 13]	21.1
IODINE [1, 10]	14.4
MCG [1, 25]	40.9
SAVI [2, 17]	20.0
SAVI++ [2, 8]	23.9
STEVE [2, 29]	11.9
Karazija et al. [16]	50.8
Karazija et al. (WL) [16]	51.9
Bao et al [1]	47.1
BMOD ([1])	<u>54.7</u>
BMOD* ([1])	**60.8**

Table 2. Performance comparison of BMOD and previous methods for unsupervised object discovery on KITTI dataset.

5. Ablation and further analysis

5.1. Analysis of the composition of objective functions

In this section, we investigate the composition of our objective functions. First, we test a more *naive* way of isolating the background, by explicitly placing the **non-moving background** in one slot's attention map, using BCE loss. As expected, this method fails to capture most foreground objects, resulting in a critical degradation of fg-ARI. Indeed, the non-moving background in the estimated masks contains all static objects and a few moving but difficult-

Dataset	Guidance	Method	all-ARI	Jaccard score	
				fg-class	bg-class
TRI-PD	gt	Bao et al. [1]	18.1	19.3	46.2
		BMOD ([1])	79.7	73.0	95.8
		BMOD* ([1])	**84.9**	**78.2**	**97.6**
		MoTok [2]	25.2	26.5	64.3
		BMOD ([2])	81.7	75.7	96.5
		BMOD* ([2])	**84.0**	**77.0**	**97.5**
	est	Bao et al. [1]	6.3	15.0	33.4
		BMOD ([1])	28.6	**27.2**	77.5
		BMOD* ([1])	**29.1**	26.5	**78.7**
		MoTok [2]	4.7	14.8	28.5
		BMOD ([2])	24.9	25.1	73.2
		BMOD* ([2])	**26.7**	**25.7**	**75.2**
KITTI	est	Bao et al. [1]	4.2	9.1	39.3
		BMOD ([1])	17.8	13.7	**70.5**
		BMOD* ([1])	**21.7**	**14.9**	69.9

Table 3. Performance comparison of BMOD with previous methods on foreground/background separation.

to-capture instances. All these elements are considered as background in the previous test. Another test is to **apply regularization to the whole attention map**. One might be motivated to do this to attenuate the noisy regions contained in the estimated motion masks, but this is not optimal as it encourages the model to attenuate activation on object regions too. Finally, we test the variant of our proposed loss functions **without the dynamically weighted BCE** described in section 3.3, by setting a fixed weight of one, instead. As expected, not accounting for object size in our method, when one group/class is predominant (background), encourages the model to place more objects in that group, causing objects to be lost (see table 4).
all-ARI results are not reported here since they are not informative when there is a significant loss in fg-ARI. In this case, a high all-ARI means that objects have been falsely attributed to the predominant class (background). Our aim, however, is to handle noise in the background without compromising the ability to capture objects.

Method	fg-ARI
isolate only non-moving background	10.1
regularization on the whole map	48.4
w/o weighted BCE	45.0
Our full approach (BMOD)	**53.9**

Table 4. Ablation study on the design of the objective functions on TRI-PD under the unsupervised setting.

5.2. Enhancing the unsupervised setting performance: gains from noiseless pseudo-labels

In this section we investigate the upper-bound performance that can be achieved in the unsupervised setting, which corresponds to the use of pseudo-masks of moving objects, extracted from optical flow. It is important to note that these estimated labels are subject to a significant amount of noise arising from camera motion. This noise takes the form of random segments which are used to guide the slots learning. In this study, we apply a simple heuristic related to the nature of the analysed scenes, to filter out this noise. Since we're looking to localize objects of a driving scene, which are unlikely to lie at the top of the frame, we apply the heuristic to the position of the objects, filtering out any segment in the first upper tier of the image. As can be seen below, this simple heuristic provides a stronger baseline for all-ARI, while maintaining objects localization performance (equivalent fg-ARI). This indicates that the method's potential for improvement is related to the quality of the pseudo-labels, justifying further exploration in this area.

Method	fg-ARI	all-ARI	Jaccard score	
			Fg-class	Bg-class
Bao et al. [1]	50.9	6.3	15.0	33.4
BMOD ([1])	**53.9**	28.6	27.2	77.5
BMOD + noiseless pseudo labels	52.3	**58.8**	**51.5**	**91.6**

Table 5. Study of the impact of noise contained in pseudo-labels.

6. Conclusion

In this work, we present an objects discovery method that takes into account the particular semantic concept of the background, which is isolated while decomposing the scene into *objects* regions. We showed through the computation of adapted metrics the effectiveness of our method in separating object/non-object regions, without human supervision. In addition, objects localization was found to benefit considerably from the background modeling. This important result is justified by the noise reduction induced by our method, enabling better focusing on object regions. We hope the baseline proposed in this work will motivate further research on background modeling in object discovery. A further analysis showed the potential for scaling the performance of the method by improving the quality of the motion masks. We believe this deserves further exploration in future work. Finally, given the reduced amount of noise among our produced segments, this work opens up the perspective of re-using the discovered objects, for example, with a pseudo-labeling approach.

7. Acknowledgements

This work benefited from the FactoryIA supercomputer financially supported by the Ile-de-France Regional Council

References

[1] Zhipeng Bao, Pavel Tokmakov, Allan Jabri, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Discorying object that can move. In *CVPR*, 2022. 2, 3, 4, 5, 6, 7, 8

[2] Zhipeng Bao, Pavel Tokmakov, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Object discovery from motion-guided tokens. In *CVPR*, 2023. 1, 2, 3, 6, 7, 8

[3] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019. 2, 7

[4] 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, 3

[5] Junyi Chai, Hao Zeng, Anming Li, and Eric W.T. Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, 2021. 1

[6] Honglin Chen, Rahul Venkatesh, Yoni Friedman, Jiajun Wu, Joshua B Tenenbaum, Daniel LK Yamins, and Daniel M Bear. Unsupervised segmentation in real-world images via spelke object inference. In *European Conference on Computer Vision*, pages 719–735. Springer, 2022. 1

[7] A. Dave, P. Tokmakov, and D. Ramanan. Towards segmenting anything that moves. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1493–1502, Los Alamitos, CA, USA, oct 2019. IEEE Computer Society. 6

[8] 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, 3, 7

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

[10] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International Conference on Machine Learning*, pages 2424–2433. PMLR, 2019. 3, 7

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[12] Wenpeng Hu, Mengyu Wang, Qi Qin, Jinwen Ma, and Bing Liu. Hrn: A holistic approach to one class learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. 4

[13] Jindong Jiang, Sepehr Janghorbani, Gerard De Melo, and Sungjin Ahn. Scalor: Generative world models with scalable object representations. *arXiv preprint arXiv:1910.02384*, 2019. 3, 7

[14] Vijay Jumb, Mandar Sohani, and Avinash Shrivas. Color image segmentation using k-means clustering and otsu's adaptive thresholding. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 3(9):72–76, 2014. 3

[15] Sandra Kara, Hejer Ammar, Florian Chabot, and Quoc-Cuong Pham. Image segmentation-based unsupervised multiple objects discovery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3277–3286, 2023. 1, 2

[16] Laurynas Karazija, Subhabrata Choudhury, Iro Laina, C. Rupprecht, and Andrea Vedaldi. Unsupervised multi-object segmentation by predicting probable motion patterns. *ArXiv*, abs/2210.12148, 2022. 3, 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, 3, 7

[18] Suha Kwak, Minsu Cho, Ivan Laptev, Jean Ponce, and Cordelia Schmid. Unsupervised object discovery and tracking in video collections. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3173–3181, 2015. 3

[19] Qing Liu, Vignesh Ramanathan, Dhruv Mahajan, Alan Yuille, and Zhenheng Yang. Weakly supervised instance segmentation for videos with temporal mask consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13968–13978, June 2021. 1

[20] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. 2, 4

[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 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11525–11538. Curran Associates, Inc., 2020. 1, 3, 4, 7

[22] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society. 6

[23] 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, 6, 7

[24] Yassine Ouali, Celine Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[25] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):128–140, 2016. 2, 7

[26] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. 3

[27] 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, 2, 3

[28] Oriane Siméoni, Chloé Sekkat, Gilles Puy, Antonin Vobecky, Éloi Zablocki, and Patrick Pérez. Unsupervised object localization: Observing the background to discover objects. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2023. 3

[29] Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 18181–18196. Curran Associates, Inc., 2022. 1, 3, 7

[30] Elizabeth S Spelke. Principles of object perception. *Cognitive science*, 14(1):29–56, 1990. 1

[31] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, 2020. 6

[32] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 2

[33] Huy V Vo, Francis Bach, Minsu Cho, Kai Han, Yann LeCun, Patrick Pérez, and Jean Ponce. Unsupervised image matching and object discovery as optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8287–8296, 2019. 2

[34] Huy V Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *European Conference on Computer Vision*, pages 779–795. Springer, 2020. 1, 2

[35] Huy V. Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce. Large-scale unsupervised object discovery. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2021. 2

[36] 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, 2, 3

[37] Mengmeng Xu, Yancheng Bai, and Bernard Ghanem. Missing labels in object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 1

[38] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *CoRR*, abs/1809.03327, 2018. 3

[39] Yanzhao Zhou, Xin Wang, Jianbin Jiao, Trevor Darrell, and Fisher Yu. Learning saliency propagation for semi-supervised instance segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10304–10313, 2020. 1

[40] Yi Zhu, Yanzhao Zhou, Huijuan Xu, Qixiang Ye, David Doermann, and Jianbin Jiao. Learning instance activation maps for weakly supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[41] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 391–405, Cham, 2014. Springer International Publishing. 2