Motion-Refined DINOSAUR for Unsupervised Multi-Object Discovery

Supplementary Material

We first extend our approach to the synthetic multi-object video dataset MOVI-E [78]. Next, we evaluate the effectiveness of our quasi-static frames retrieval method used for pseudo-label generation. Additionally, we provide qualitative insights into our work, including failure cases, pseudo-label visualization, and MR-DINOSAUR results. We finish with details about the datasets employed, as well as the implementation details, to facilitate reproducibility.

A. MR-DINOSAUR on MOVI-E

We experiment on MOVI-E to broaden the range of datasets and methods for comparison. Because none of those methods report F1 or all-ARI on MOVI-E, we restrict our evaluation to fg-ARI here. MOVi-E [78] introduces a constant, artificial camera motion that violates our static-frame assumption for pseudo-labeling. Despite this disadvantage, our method achieves promising results in the ballpark of methods that explicitly deal with camera motion in the training data as shown in Tab. 6.

Table 6. **Unsupervised multi-object discovery on MOVI-E** using fg-ARI. * denotes using DINOv2. Underlined methods use supervision.

Method	fg-ARI
GWM [14] BMVC'22	42.5
SPOT [33] CVPR'24	59.9
PPMP [36] NeurIPS'22	63.1
DINOSAUR [56] ICLR'23	65.1
MoTok [5] CVPR'23	66.7
Safadoust et al. [54] ICCV'23	78.3
VideoSAUR [79] NeurIPS'23	78.4
SOLV* [3] NeurIPS'23	80.8
DIOD* [35] CVPR'24	82.2
DINOSAUR* [56] ICLR'23	66.2
MR-DINOSAUR* *(Ours)*	80.1

B. Quasi-static Frame Retrieval Analysis

We evaluate the effectiveness of our quasi-static frame retrieval method on the KITTI dataset, which includes detailed annotations of ground truth camera velocity for every video frame. Given that our motion segmentation approach used for pseudo-label generation relies on the static background assumption, the quasi-static frame retrieval plays an important role in achieving high-quality pseudo labels. We evaluate the quasi-static frame retrieval by comparing the set of frames our method retrieves from the training data to the ground-truth quasi-static frames. Ground-truth quasi-static frames are defined as frames with camera velocities below 0.2 m/s. Our method achieves an impressive 99.4 % accuracy, 99.2 % precision, and 96.6 % recall as shown in Tab. 7, confirming that thresholding the average flow magnitude at the image corners is a simple and effective way to retrieve quasi-static frames.

Table 7. **Quasi-static frame retrieval analysis** using accuracy, precision, recall (all in %) on the KITTI dataset. We compare the set of frames retrieved from the training videos by our method to the set of frames with a ground-truth velocity smaller than 0.2 m/s.

Ground-truth Velocity	Accuracy	Precision	Recall
< 0.2 m/s	99.4	99.2	96.6

C. More Qualitative Results

We provide additional qualitative visualizations of our pseudo-labels and our proposed method MR-DINOSAUR, as well as failure cases.

C.1. Qualitative pseudo-label examples

Fig. 6 shows additional visualizations of our pseudo-labels comparing to TSAM pseudo labels. Consistent with the analysis in Sec. 4.2, our pseudo-labels are precise and of high quality for both synthetic TRI-PD and real-world KITTI data. Although some samples exhibit motion artifacts introduced by the unsupervised optical flow from SMURF [60] (*e.g.*, the left KITTI image), we mainly observe accurate object masks. Compared to TSAM pseudo labels, our pseudo labels exhibit fewer artifacts.

C.2. Failure cases

We visualize representative failure cases for MR-DINOSAUR in Fig. 7. Occasionally, predictions cover non-object structures (*e.g.*, a tree in the left image) and over-segmentation occurs on large objects with intricate textures (*e.g.*, the truck in the right TRI-PD image). Precise segmentation of small, overlapping objects also remains challenging. Notably, similar issues—such as artifacts and missed small objects—are observed with the state-of-the-art DIOD [35] method.

C.3. Qualitative MR-DINOSAUR examples

Finally, Fig. 8 presents additional qualitative examples comparing our method, MR-DINOSAUR, to the baseline DINOSAUR [56] and DIOD [35]. While DINOSAUR establishes a solid foundation, it tends to undersegment and blur the distinction between objects and background. DIOD produces good qualitative results but often yields noisy

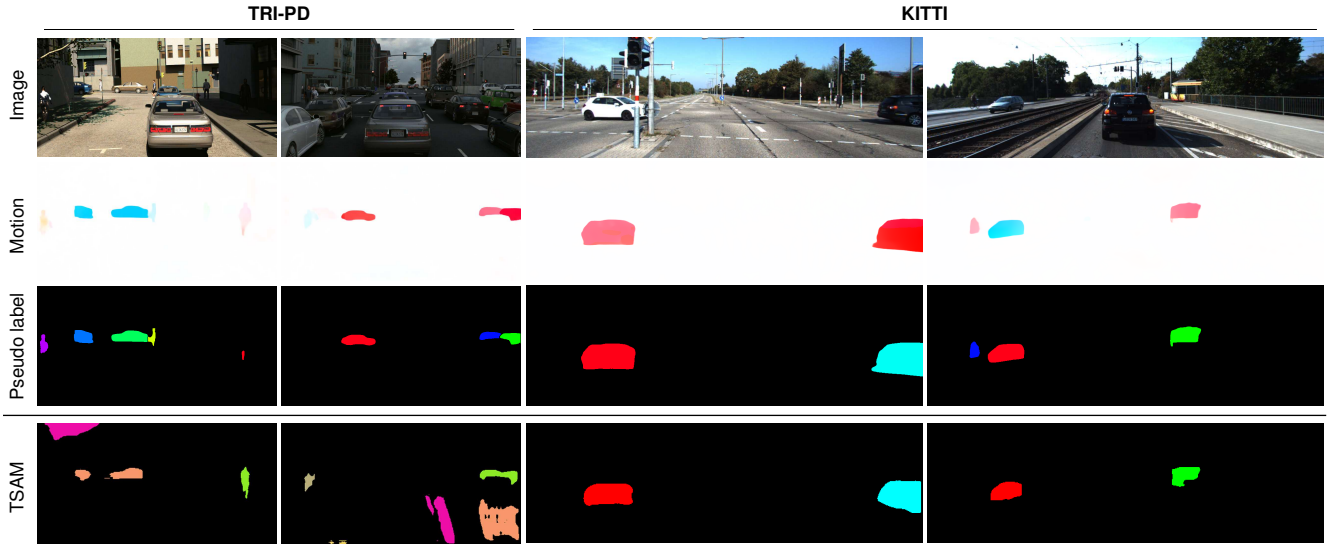

Figure 6. **Additional visualizations** of our pseudo-labels on the TRI-PD [4] and KITTI [24] dataset. We further visualize the respective TSAM pseudo labels used by DIOD [35]. Here we use random colors for different objects.

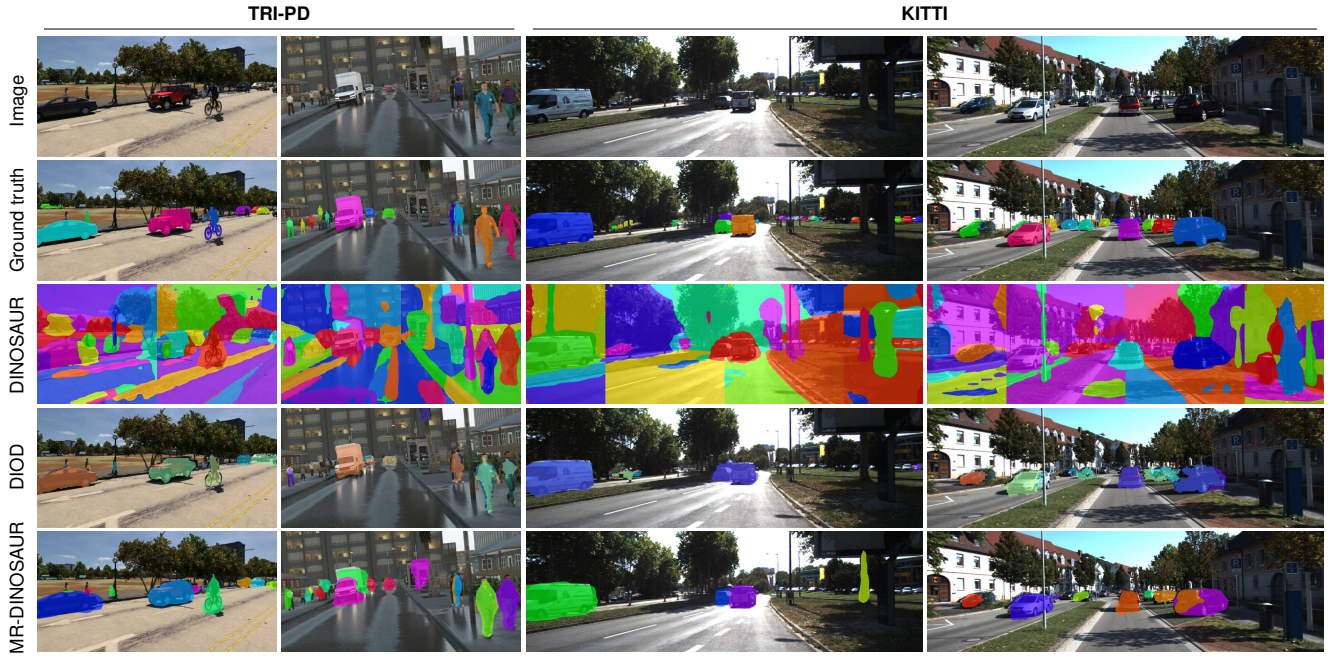

Figure 7. **Failure cases** of MR-DINOSAUR *(Ours)* comparing to DIOD [35] and our baseline DINOSAUR [56] on the TRI-PD [4] and KITTI [24] dataset. Here we use random colors for different objects.

masks by merging multiple objects into a single mask or missing objects entirely. In contrast, MR-DINOSAUR effectively differentiates foreground from background, resulting in fewer false positives and demonstrating superior capability in detecting small instances.

D. Reproducibility

To facilitate reproducibility, we elaborate on the technical and implementation details. Note that our code is available at https://github.com/visinf/mrdinosaur.

D.1. Datasets

TRI-PD [4] is a synthetic urban driving-scene dataset extracted from Parallel Domain [77]. It includes detailed

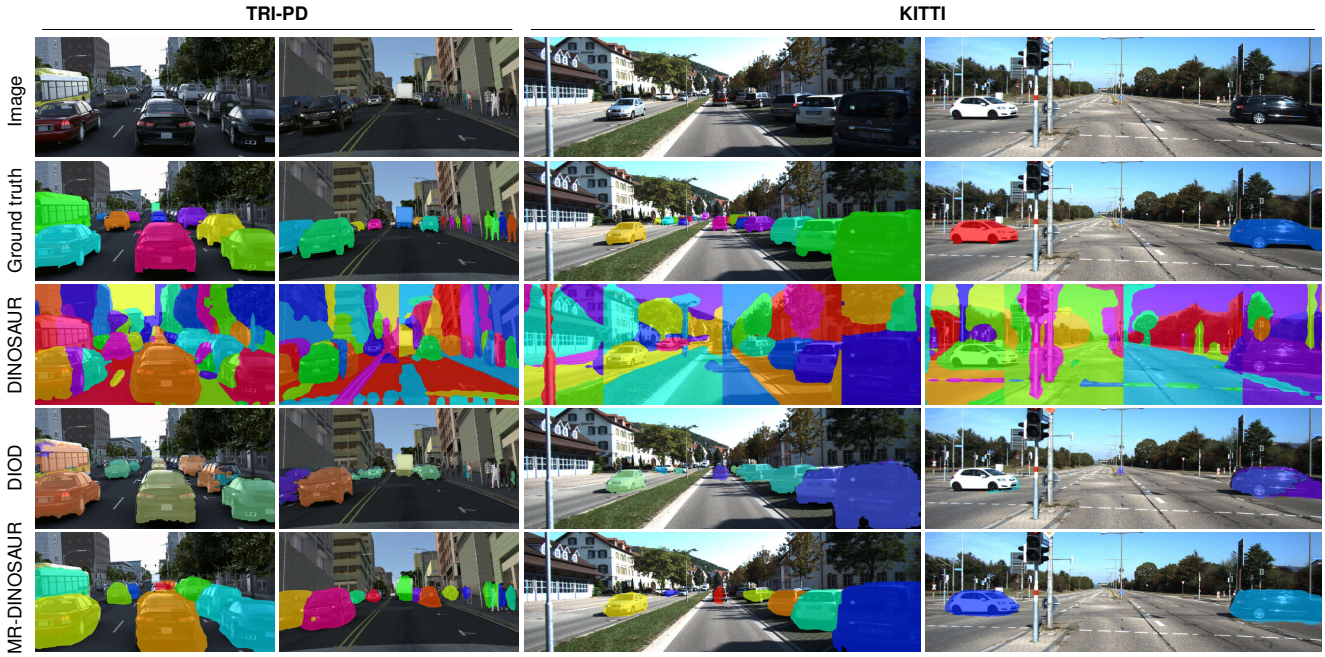

Figure 8. **Qualitative comparison** of our baseline DINOSAUR [56], DIOD [35], and MR-DINOSAUR *(Ours)* on the TRI-PD [4] datasets and KITTI [24]. Here we use random colors for different object instances.

annotations, including camera pose, calibration, depth, instance segmentation, semantic segmentation, 2D/3D bounding box, depth, forward/backward 2D motion vectors, and forward/backward 3D motion vectors. The training set consists of 200 photorealistic scenes captured by six cameras, each with 200 frames; the validation set comprises 17 scenes recorded by three cameras, totaling 10 200 frames. In this paper, we only use the three front-camera frames that align with the validation set for training. Following previous work [4, 5, 34, 35], we discard scenarios with low visibility (*e.g.*, foggy and dark scenes), resulting in 157 scenes and a total of 94 200 frames, for training DINOSAUR. From these, we extract 13 280 quasi-static frames for training MR-DINOSAUR. The resolution of all frames is 1216×1936. Following previous work [4, 5, 34, 35], we resize and crop the images to a resolution of 980×490 for pseudo-labeling and training. Training is performed on two non-overlapping square crops of size 490×490.

KITTI [24] is a widely used autonomous driving dataset. It includes various sensor data collected from various environments, *e.g.*, urban, rural, and highway scenes, offering extensive annotations and a diverse range of scenarios. We train DINOSAUR using all images provided in the raw data, resulting in 95 778 frames from 151 videos. We retrieve 12 526 quasi-static frames for training MR-DINOSAUR. For evaluation, we utilize the instance segmentation subset, which consists of 200 frames with a resolution of 375×1242. Each frame is an individual image, rather than part of a consecutive sequence. Also, following previous work [4, 5, 34, 35], we resize the images to a resolution of 378×1260 for pseudo-labeling and training. Training is performed on four non-overlapping square crops of size 378×378.

MOVI [78] is a synthetic video dataset comprising six sub-datasets (MOVI-A to MOVI-F) of increasing complexity. Each sub-dataset consists of generated scenes, with each scene representing a two-second rigid-body simulation of falling objects. The sub-datasets vary in object count and type, background, camera trajectory, and whether all objects are in motion or some remain stationary. We experiment on the MOVi-E dataset used by several previous works on multi-object discovery [3, 5, 35, 54, 79]. MOVi-E introduces simple camera movement, where the camera moves along a straight line at a random but constant velocity. Each video consists of 24 frames, with the training set containing 9749 videos (a total of 233 976 frames) and the validation set containing 250 videos (a total of 6000 frames). We randomly selected 9 frames from each video for training, resulting in a total of 87 741 images used for training DINOSAUR. We retrieve 84 831 quasi-static frames for training MR-DINOSAUR. Images originally at 256×256 are resized to 266×266 for training to account for the patch size of DINOv2.

Table 8. DINOSAUR and MR-DINOSAUR hyperparameters used for the results on the TRI-PD, KITTI, and MOVI-E datasets.

DINOSAUR				
Dataset		**TRI-PD**	**KITTI**	**MOVi-E**
Training steps		500k	500k	500k
Batch size		16	64	64
Optimizer		Adam	Adam	Adam
Number of warmup steps		10k	10k	10k
Peak learning rate		1e-4	4e-4	4e-4
Exponential decay half-life		100k	100k	100k
ViT architecture		DINOv2-ViT-B/14	DINOv2-ViT-B/14	DINOv2-ViT-B/14
Image/Crop size		490	378	266
Cropping strategy		Random	Random	Full
Augmentations		Random Horizontal Flip	Random Horizontal Flip	-
Decoder	Type	MLP	MLP	MLP
	Layers	4	4	4
	MLP hidden dimension	2048	2048	1024
Slot Attention	Number of slots	30	15	24
	Total number of slots	60	60	24
	Iterations	3	3	3
	Slot dimension D_{slots}	32	32	128
MR-DINOSAUR				
Pseudo label generation	Quasi-static frame retrieval threshold τ_{static}	0.5	1.7	1.7
	Foreground mask threshold τ_{fg}	2.5	2.5	2.5
	Flow-gradient threshold τ_{∇}	20	20	20
Training stage 1	Training epochs	15	15	15
	Batch size	8	8	8
	Learning rate	4e-06	4e-06	4e-06
Training stage 2	Training epochs	1	1	1
	Batch size	8	8	8
	Learning rate	4e-05	4e-05	4e-05
	Regularization term α	0.2	0.2	0.2
	Drop similarity τ_{drop}	0.99	0.99	0.99
Slot deactivation module	Layers	4	4	4
	MLP hidden dimension	2048	2048	2048

D.2. Computational Requirements

All experiments use a single NVIDIA RTX 6000 Ada Generation GPU (48 GB VRAM) in a workstation equipped with an AMD EPYC 7343 CPU (32 cores) and 512 GB RAM.

On TRI-PD, a full 500 000-step training of the DINOSAUR baseline takes approximately 267 h, involving 97.3 M parameters, of which 10.7 M are trainable. For MR-DINOSAUR, training stage 1 (15 epochs, batch size 8) takes approximately 11 h, utilizing the same 97.3 M parameters, of which 634 K are trainable. Training stage 2 (1 epoch, batch size 8) takes approximately 40 min and utilizes 105.8 M parameters with 8.5 M trainable. Peak memory usage reaches 40.2 GB. At inference, we process an image at a resolution of 490×980 in 740 ms.

D.3. Further Implementation Details

Finally, we provide an overview of all hyperparameters used for training the baseline DINOSAUR and our method MR-DINOSAUR in Tab. 8.

References

[77] Parallel Domain. https://paralleldomain.com/. June 2025. ii

[78] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam H. Laradji, Hsueh-Ti Derek Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, A. Cengiz Öztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong and Andrea Tagliasacchi. Kubric: A scalable dataset generator. In *CVPR*, pages 3739–3751, 2022. i, iii

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