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

Supplementary Material

6. Further implementation details

In this section, we provide further branch-specific implementation details, about the real-world setting.

- 2D Branch: Same configuration as in [15], with visual augmentations only.
- 3D Branch: The input frontal projection is resized to (368 × 1248) and min-max normalized using global dataset-wide minimum and maximum values. The 3D student model receives sequences of images augmented using *data-jittering* with a probability 0.4, while the teacher model receives the original projected image. For the scene completion task, 20% of the projected points are dropped from the input and the 3D student is trained to recover them.

7. Detailed Multi-modal Object Discovery results

The results in Table 7 indicate that the 3D baseline model (DIOD-3D) exhibits relatively low recall, particularly in real-world data (KITTI), where it achieves 13.2, highlighting challenges in object localization within sparse 3D data. Recall improves when the 3D branch is trained alongside RGB images via cross-modal training, as seen with xMOD (3D), which increases recall to 16.9 on KITTI. This improvement suggests that the model benefits from clearer object patterns in RGB images. On the other hand, the 2D only method presents low precision due to the noise and less homogeneous textures compared to 3D data, resulting in lower precision in 2D (*e.g*. 17.8 for DIOD on KITTI). 2D precision is enhanced through cross-modal training (from 17.8 to 22.8), possibly due to inconsistencies between RGB noise and point cloud data. The proposed late fusion in xMOD (2D + 3D) significantly boosts precision while maintaining high recall, achieving a better balance between both, as reflected by higher F1 scores of 27.4 for KITTI and 42.5 for TRI-PD. In Table 8 we see that, across specific depth ranges (10-30m), the late fusion preserves high recall while maintaining strong precision, demonstrating reduced trade-offs and improved performance.

8. Definition of new benchmarks with available 3D data.

As stated in the main paper, experiments that include 3D data are evaluated on new test sets for TRI-PD and KITTI, with available 3D data.

- The test set for TRI-PD was derived from the old training set (with these scenes excluded during training). It consists of the first 17 non-banned scenes, each including views from cameras 1, 5, and 6 :

  scene_000003, scene_000007, scene_000010,
  scene_000011, scene_000013, scene_000014,
  scene_000017, scene_000020, scene_000022,
  scene_000023, scene_000024, scene_000025,
  scene_000026, scene_000027, scene_000030,
  scene_000031, scene_000033

- The new test-set for KITTI includes 142 scenes, selected from the original 200 scenes in the previous test set, where LiDAR data is available. The scenes names are listed below:
  000163, 000107, 000031, 000090, 000053,
  000069, 000119, 000070, 000075, 000043,
  000042, 000015, 000086, 000048, 000011,
  000109, 000157, 000142, 000037, 000033,
  000019, 000098, 000039, 000016, 000145,
  000114, 000025, 000095, 000113, 000010,
  000076, 000110, 000038, 000018, 000160,
  000044, 000040, 000027, 000034, 000045,
  000029, 000093, 000147, 000122, 000128,
  000067, 000143, 000141, 000047, 000002,
  000052, 000158, 000149, 000020, 000079,
  000032, 000055, 000036, 000097, 000074,
  000051, 000066, 000089, 000092, 000009,
  000077, 000028, 000162, 000115, 000124,
  000085, 000108, 000054, 000080, 000123,
  000126, 000088, 000148, 000094, 000159,
  000132, 000017, 000155, 000078, 000072,
  000105, 000081, 000168, 000073, 000116,
  000164, 000112, 000199, 000056, 000106,
  000050, 000129, 000024, 000068, 000169,
  000059, 000003, 000130, 000065, 000146,
  000064, 000023, 000131, 000144, 000117,
  000013, 000058, 000062, 000049, 000012,
  000121, 000026, 000091, 000150, 000041,
  000071, 000022, 000060, 000046, 000096,
  000030, 000007, 000161, 000111, 000118,
  000084, 000014, 000127, 000008, 000063,
  000125, 000120, 000021, 000057, 000035,
  000061

9. Comparison with Clusternet

In the ClusterNet [37] average precision (AP) computation, a subtle yet significant implementation nuance in the handling of predictions below the Intersection over Union

	KITTI				TRI-PD			
	all-ARI	F1@50	Precision	Recall	all-ARI	F1@50	Precision	Recall
DIOD [15]	62.8	18.7	17.8	19.7	**66.1**	30,6	22,4	**48,1**
xMOD (2D)	69.7	22.3	22.8	**21.8**	64.7	35.5	30.4	42.8
DIOD-3D	51.6	15.5	18.9	13.2	65.1	39.6	**47.0**	34.3
xMOD (3D)	58.8	18.9	21.6	16.9	65.0	37.5	32.4	44.6
xMOD (2D + 3D)	**75.8**	**27.4**	**56.9**	18.0	64.8	**42.5**	42.9	42.0

Table 7. **Multi-modal Object Discovery** evaluated on the new KITTI and TRI-PD testsets with available 3D data (see section 8). The models resulting from our proposed approach are presented in blue. Parentheses indicate the modality used during inference.

	0-10			10-30			30-70		
	F1@50	Precision	Recall	F1@50	Precision	Recall	F1@50	Precision	Recall
DIOD [15]	15.3	11.6	25.4	26.2	21.6	34.8	12.7	8.5	21.1
xMOD (2D)	20.6	30.8	15.6	32.5	31.1	34.3	16.0	15.0	16.5
DIOD-3D	15.2	30.3	10.7	22.9	21.9	24.0	8.9	20.0	5.5
xMOD (3D)	20.2	60.7	12.3	28.7	25.2	32.8	10.1	9.4	12.1
xMOD (2D + 3D)	21.7	68.2	12.9	46.4	85.7	31.8	7.2	29.5	4.1

Table 8. **Multi-modal Object Discovery on KITTI for different subsets of object defined by their distance to the camera**. The models resulting from our proposed approach are presented in blue. Parentheses indicate the modality used during inference.

(IoU) threshold between predicted and ground truth masks is present. The original code lacks an explicit ”else” clause when evaluating instance matches, effectively omitting predictions with IoU values below the specified threshold. This results in an incomplete categorization of predictions, where instances not meeting the IoU criterion are neither classified as true positives (TP) nor false positives (FP). Consequently, the lists tracking false positives (FP), true positives (TP), and prediction scores (scores) become inconsistent, potentially introducing computational errors in metric calculations. The original paper (main or supplementary material) did not provide any precision regarding this customized AP computation.

Furthermore, regarding the implementation of the method itself, key components such as the 3D Instance Segmentation module are absent from the repository, thus we were not able to reproduce it in order to estimate their performances with a standard AP formula. To facilitate comparison, we report in Table 9 the performances of our approach with an implementation of the AP that is similar to that of ClusterNet [37].

Modality	Method	WOD
Multi	Clusternet [37]	**26.2**
	xMOD (2D+3D)	**30.0**

Table 9. **Multi-modal Object Discovery evaluated on the dataset WOD in AP@70**. The models resulting from our proposed approach are presented in blue. .

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, 4, 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. To-

wards 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. 3

[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. 3

[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 fine-tuning 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. 3

[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, 1

[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