Motal: Unsupervised 3D Object Detection by Modality and Task-specific Knowledge Transfer

Hai Wu[1,2] Hongwei Lin[1] Xusheng Guo[1] Xin Li[3]
Mingming Wang[4,5] Cheng Wang[1] Chenglu Wen[1*]
[1]Xiamen University [2] Pengcheng Laboratory [3]Texas A&M University
[4] Tsinghua University [5] Guangzhou Automobile Group Co. R&D Center

Method	Vehicle	Pedestrian	Cyclist
CenterPoint (Fully supervised)	63.16	64.27	66.11
CPD [5]	32.13	13.22	4.87
train by CPD pseudo labels (score>0.1)	33.93	7.09	0.46
train by CPD pseudo labels (score>0.3)	34.77	12.85	0.69
train by CPD pseudo labels (score>0.5)	33.15	10.14	1.65
train with image pseudo points	33.41	5.18	3.40
train with image RGB	33.98	5.97	7.47
`Motal` (CPD)	46.45	27.76	29.94

Table 1. 3D AP L2 on WOD validation set.

1. Why did we not train the detector network with initial pseudo labels directly?

As analyzed in our main paper, pseudo-labels generated by motion or geometry heuristics cannot be accurate in classification and regression at the same time. Directly using the pseudo labels to train the detector cannot attain desirable performance improvement. For example, we use the pseudo-labels output from CPD to train the VoxelRCNN [1]. The results are in Table 1. We observe that no matter what score threshold is used to select the pseudo-labels, the detection performance improves marginally or even decreases. The reasons are: (1) Using a low score threshold to select pseudo-labels introduces classification and regression errors, leading to decreased detection accuracy. (2) Using a high score threshold to select pseudo-labels disregards numerous useful supervision signals. Therefore, we designed the `Motal`, which well addressed this problem by a modality and task-specific knowledge transfer design.

2. Why did we not use the image as input directly?

A possible way is to directly feed the image to a multi-modal detector but this will introduce additional inference time. Moreover, since lots of false supervision signals exist in the pseudo labels, the detection performance cannot be enhanced. An example of using virtual points of VirConv [4] and pseudo labels from CPD is shown in Table 1, where the performance improves marginally. Some classes even become worse.

3. Ablation results for parameters.

The ablation results for $\hat{\eta}_{cg}$ and η_{cg} are shown in the Table 3. Since $\hat{\eta}_{cg} = 0.1$ and $\eta_{cg} = 0.7$, we obtain the best results, we use $\hat{\eta}_{cg} = 0.1$ and $\eta_{cg} = 0.7$ in this paper.

4. More comparison with AML.

Since AML is also based on the motion-based design, we provide comparison results with AML [3] using the metrics of AML. The results are shown in the Table 4. Our method outperforms the previous method greatly. These advancements come from the MLE and TMT designs, which better leverage the motion, image appearance, and geometry prior via a modality and task-specific knowledge transfer framework for unsupervised 3D object detection.

5. More examples of challenge objects

In our paper, we use the image to extend classification labels, as many objects cannot be distinguished on points but can be recognized on the image. Here, we present more examples in Fig. 1.

6. More detailed comparison results

We also provide more detailed comparison results including 3D AP / 3D APH / BEV AP on the WOD validation set (see Table 2). Our method shows better performance on all metrics, further demonstrating its effectiveness. We provide the training rounds vs. performances in Table 5. Compared with fully supervised method, OYSTER requires around 2× training time and MODEST requires around 10× training time. Our `Motal` also requires 2× training times, but its performance is significantly higher than MODEST and

*Corresponding author

Method	Metric	Average AP L1	Avarage AP L2	Veh. L1 $IoU_{0.5/0.7}$	Veh. L2 $IoU_{0.5/0.7}$	Ped. L1 $IoU_{0.3/0.5}$	Ped. L2 $IoU_{0.3/0.5}$	Cyc. L1 $IoU_{0.3/0.5}$	Cyc. L2 $IoU_{0.3/0.5}$
DBSCAN [2]	3D AP	0.57	0.43	2.32 / 0.29	1.94 / 0.25	0.51 / 0.00	0.19 / 0.00	0.28 / 0.03	0.20 / 0.00
DBSCAN* [6]		3.73	3.19	17.36 / 2.65	14.87 / 2.29	1.65 / 0.00	1.35 / 0.00	0.48 / 0.25	0.43 / 0.20
MODEST [6]		6.59	5.42	18.51 / 6.46	15.83 / 5.48	11.83 / 0.17	8.96 / 0.10	1.47 / 1.14	1.17 / 1.01
OYSTER [7]		8.54	7.58	30.48 / 14.66	26.21 / 14.10	4.33 / 0.18	3.52 / 0.14	1.27 / 0.33	1.24 / 0.32
CPD [5]		24.05	20.67	57.79 / 37.40	50.18 / 32.13	21.91 / 16.31	18.01 / 13.22	5.83 / 5.06	5.61 / 4.87
`Motal` (CPD)		44.89	39.74	75.07 / 53.61	66.17 / 46.45	43.79 / 33.26	36.79 / 27.76	33.71 / 29.94	32.49 / 28.83
DBSCAN [2]	3D APH	0.41	0.29	1.78 / 0.15	1.45 / 0.13	0.34 / 0.00	0.07 / 0.00	0.20 / 0.00	0.12 / 0.00
DBSCAN* [6]		3.14	2.63	15.31 / 2.12	12.84 / 1.64	1.12 / 0.00	1.02 / 0.00	0.23 / 0.11	0.21 / 0.08
MODEST [6]		4.71	3.73	16.43 / 4.25	14.04 / 3.63	5.59 / 0.11	4.18 / 0.05	1.07 / 0.82	0.45 / 0.07
OYSTER [7]		7.62	6.77	28.56 / 12.87	25.01 / 12.54	3.12 / 0.12	2.03 / 0.06	0.87 / 0.24	0.82 / 0.21
CPD [5]		19.55	16.91	54.19 / 34.97	46.99 / 30.09	12.01 / 9.24	10.06 / 7.68	3.68 / 3.26	3.55 / 3.14
`Motal` (CPD)		33.25	29.63	68.55 / 49.17	60.40 / 42.60	18.39 / 13.72	15.46 / 11.46	26.24 / 23.45	25.28 / 22.58
DBSCAN [2]	BEV AP	0.76	0.56	2.91 / 0.55	2.34 / 0.47	0.73 / 0.01	0.21 / 0.01	0.32 / 0.10	0.25 / 0.10
DBSCAN* [6]		7.41	6.68	22.33 / 13.30	20.60 / 11.95	7.21 / 0.23	6.49 / 0.10	1.03 / 0.39	0.73 / 0.24
MODEST [6]		10.51	8.51	27.16 / 16.58	21.13 / 13.31	15.98 / 0.34	14.06 / 0.13	1.73 / 1.27	1.38 / 1.07
OYSTER [7]		14.05	11.84	37.73 / 30.57	32.31 / 25.04	13.53 / 0.57	11.76 / 0.30	1.56 / 0.38	1.32 / 0.33
CPD [5]		27.93	24.91	60.81 / 53.01	53.66 / 47.48	22.96 / 19.31	20.21 / 17.26	5.98 / 5.56	5.68 / 5.22
`Motal` (CPD)		48.75	43.28	77.94 / 68.78	69.01 / 60.27	44.04 / 37.04	37.02 / 30.99	33.71 / 31.04	32.49 / 29.91

Table 2. Unsupervised 3D object detection results on WOD validation set. We report the 3D AP, 3D APH, and BEV AP using the official metric code of WOD with different IoU thresholds. * denotes initial training.

$\hat{\eta}_{cg}$	0.05	0.1	0.15	0.2
mAP L2	32.15	34.34	34.12	33.89
η_{cg}	0.65	0.7	0.75	0.8
mAP L2	33.78	34.34	33.92	32.45

Table 3. Ablation results for $\hat{\eta}_{cg}$ and η_{cg}.

Method	3D mAP		2D mAP	
	L1	L2	L1	L2
Unsup Flow + AML [3]	42.1	40.4	49.1	47.4
`Motal`	60.5	57.5	61.4	58.4

Table 4. Comparison with AML.

Method	Rounds	L1	L2
MODEST [6]	10	2.5	2.2
OYSTER [7]	2	7.4	6.4
CPD [5]	1	20.5	18.1
`Motal`	2	42.1	38.0

Table 5. Training rounds and performances.

OYSTER. We will investigate more efficient methods of speeding up training in the future.

7. More detailed IoU distribution after box propagation

To better understand how box propagation improves the box quality, we present a more detailed box IoU distribution of regression labels before and after box propagation in Fig 2. The IoU distribution between pseudo-label and ground truth becomes closer to 1, verifying its effectiveness.

8. More example of label extension on image

To better understand how CGloss discovers new objects, we present more heatmaps predicted by different methods in Fig. 3. We observe that, by using our CGloss, the generated heatmaps contain more generalized instances. These results further verified the effectiveness of our method.

References

[1] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wen gang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 1

[2] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining*, 1996. 2

[3] 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. 1, 2

[4] Hai Wu, Chenglu Wen, Shaoshuai Shi, Xin Li, and Cheng Wang. Virtual sparse convolution for multimodal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[5] Hai Wu, Shijia Zhao, Xun Huang, Chenglu Wen, Xin Li, and Cheng Wang. Commonsense prototype for outdoor unsupervised 3d object detection. In *Proceedings of the IEEE/CVF*

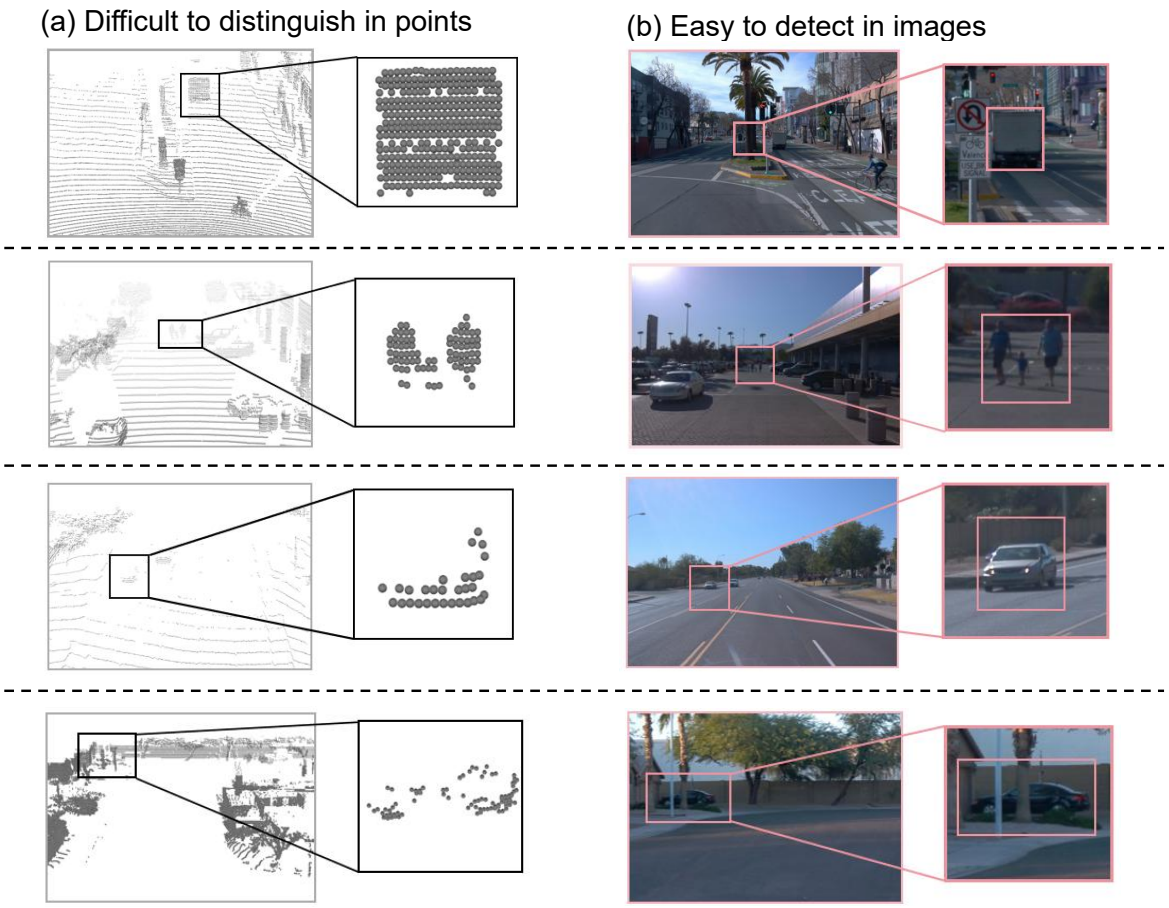

Figure 1. (a) It's difficult to distinguish objects in 3D points. (b) But it's easy to recognize objects in 2D images.

Conference on Computer Vision and Pattern Recognition, pages 14968–14977, 2024. 1, 2

[6] 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. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[7] 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

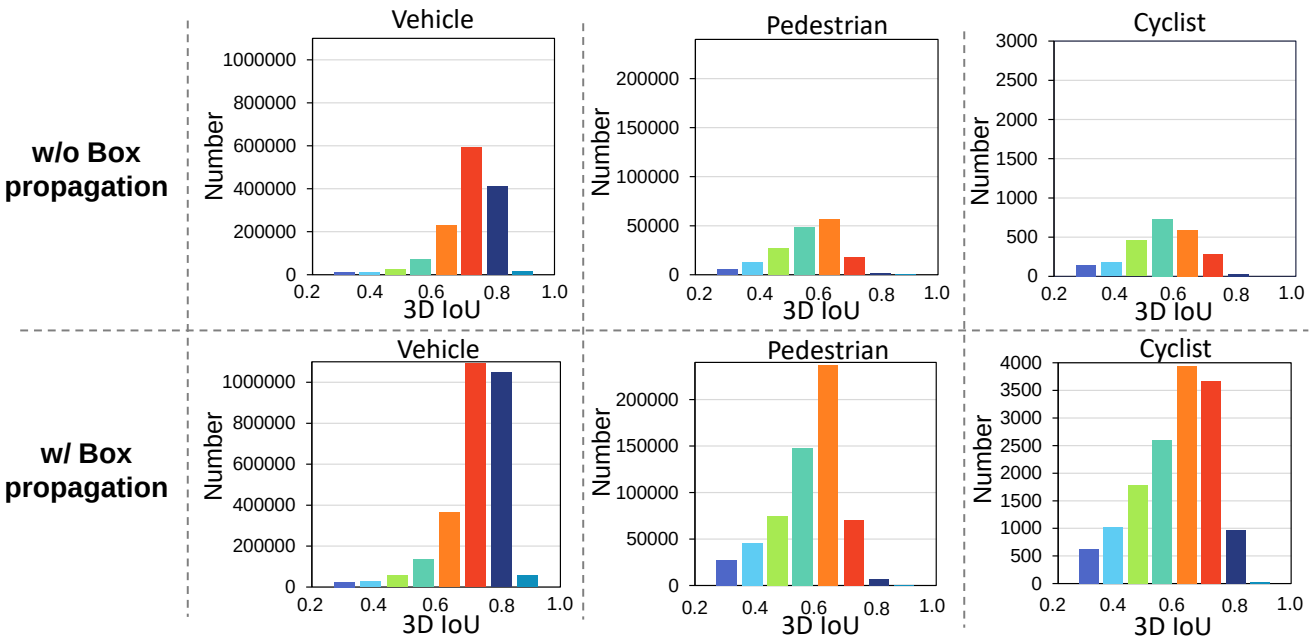

Figure 2. Label IoU distribution before and after box propagation.

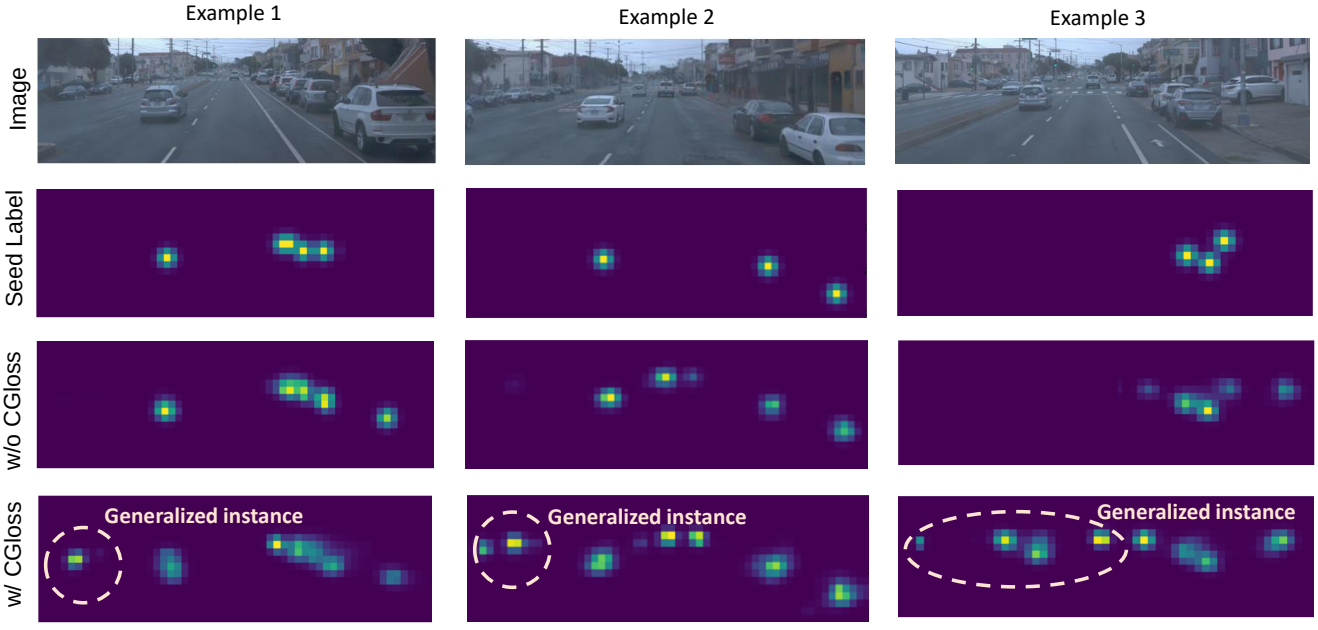

Figure 3. Heatmap predicted by different 2D networks.