When the City Teaches the Car: Label-Free 3D Perception from Infrastructure

Anonymous CVPR submission

Paper ID ****

Abstract

Building robust 3D perception for self-driving still relies heavily on large-scale data collection and manual annotation, yet this paradigm becomes impractical as deployment expands across diverse cities and regions. Meanwhile, modern cities are increasingly instrumented with roadside units *(RSUs), static sensors deployed along roads and at intersections to monitor traffic. This raises a natural question:* can the city itself help train the vehicle? *We propose* infrastructure-taught, label-free *3D perception, a paradigm in which RSUs act as stationary, unsupervised teachers for ego vehicles. Leveraging their fixed viewpoints and repeated observations, RSUs learn local 3D detectors from unlabeled data and broadcast predictions to passing vehicles, which are aggregated as pseudo-label supervision for training a standalone ego detector. The resulting model requires no infrastructure or communication at test time. We instantiate this idea as a fully* label-free three-stage *pipeline and conduct a concept-and-feasibility study in a CARLA-based multi-agent environment. With CenterPoint, our pipeline achieves 82.3% AP for detecting vehicles, compared to a fully supervised ego upper bound of 94.4%. We further systematically analyze each stage, evaluate its scalability, and demonstrate complementarity with existing egocentric label-free methods. Together, these results suggest that city infrastructure itself can potentially provide a* scalable supervisory signal *for autonomous vehicles, positioning infrastructure-taught learning as a promising* orthogonal *paradigm for reducing annotation cost.*

1. Introduction

Developing robust perception for self-driving is inherently challenging: the task is safety-critical, and driving scenes are highly diverse. Conventional practice scales ego-view data collection and annotation to improve generalization across deployment environments. However, traffic distributions vary substantially across cities and regions, reflecting differences in road layout, infrastructure, and traffic patterns. As deployment expands to new areas, repeated collect–label–retrain cycles are necessary, and simply scaling this process becomes infeasible for addressing the structural mismatch between finite labeled data and open-world complexity.

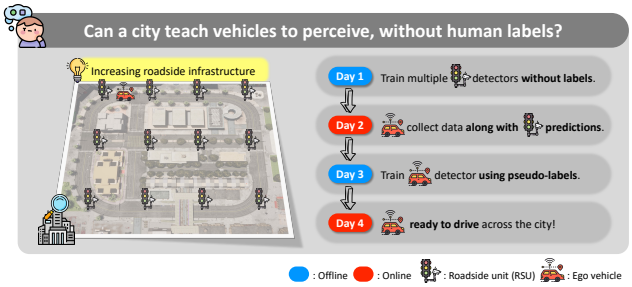

Figure 1. **Can city infrastructure teach vehicles to perceive?** We explore a new paradigm where roadside infrastructure acts as distributed teachers, providing supervision to train ego perception models without manual annotations.

Meanwhile, modern cities are increasingly deploying roadside units (RSUs)—static sensors installed at intersections and along roads—to monitor traffic and support connected and cooperative intelligent transportation systems. For instance, transportation agencies in the United States and Europe have already begun deploying such infrastructure as part of large-scale connected-vehicle initiatives [10, 47]. Unlike ego sensors, RSUs observe traffic from fixed vantage points with stable geometry and wide fields of view.

This raises a compelling possibility: *what if the city's own sensing infrastructure (*e.g.*, RSUs) could help train autonomous vehicles?* More concretely, we propose turning increasingly prevalent RSUs into *local perception experts* that provide supervision to passing ego vehicles. In principle, this could reduce the repeated burden of manual ego-side annotation and enable geographically localized supervision wherever such infrastructure is available. At a broader scale, one can envision a city-wide collection of RSUs serving as *distributed annotators*, allowing vehicles

to acquire supervision simply by driving through the city.

At first glance, this idea may seem overly ambitious: training a large number of RSUs into reliable perception systems could itself require prohibitively expensive labeled data. In that case, the burden of annotations would merely shift from the vehicle to the city, rather than being fundamentally reduced. Our key insight is that RSUs possess a property that makes this circle breakable: *stationarity*. Because they repeatedly observe the same physical scene from fixed viewpoints over time, RSUs are more amenable to label-free learning based on temporal persistence, background stability, and recurring traffic patterns. Motivated by this insight, as shown in Fig. 1, we investigate infrastructure perception as a *self-learning* teacher, leading to a fully label-free ecosystem for 3D perception.

In this paper, we instantiate this new learning paradigm as *infrastructure-taught, label-free 3D perception* and conduct a *concept-and-feasibility* study. Because existing real-world multi-agent datasets remain limited in geographic diversity and coverage [34, 41] and do not fully support controlled evaluation of our setting, we prepare a simulated dataset CIVET (**CI**ty-as-a-teacher for ego-**VE**hicle **T**raining), collected in a multi-agent environment built on CARLA [8] and V2XVerse [29]—following a common first step in V2X research, where simulation provides a reproducible testbed for principled study [21, 63]. In a nutshell, CIVET spans four towns, each instrumented with 12 RSUs, covering both urban and rural environments (*cf*. Sec. B.1). Two towns provide contrasting traffic characteristics, while the remaining towns are incorporated to study the scalability of the proposed paradigm.

With this setup, we make the problem concrete and testable through a three-stage pipeline. In the first stage, each stationary RSU learns its own detector from unlabeled observations by exploiting temporal consistency in dynamic traffic scenes. In the second stage, their predictions are broadcast to passing ego vehicles and aggregated as pseudo-label supervision for ego training. In the third stage, the ego detector is trained *offline* and operates independently at test time, carrying distilled knowledge from the city.

We systematically evaluate the feasibility of this paradigm through baselines, bottleneck analysis, and upper-bound comparisons. Within a town, our fully label-free pipeline achieves 82.3% AP for detecting vehicles with CenterPoint [71], compared to a fully supervised ego upper bound of 94.4% (*cf*. Tab. 2). This observation is consistent across different detectors [18, 71] and geolocations. We further investigate scalability and geographic generalization by jointly training across four towns using all their RSUs (48 in total), and evaluating on the combined test set *with larger diversity*, achieving 82.7% AP for vehicles compared to an upper bound of 91.0% (*cf*. Tab. A3). Finally, we demonstrate complementarity with existing *ego-centric* label-free methods. Combining pseudo-labels from a representative ego-centric approach (Oyster [78]) with our infrastructure-generated supervision yields an additional ∼10 AP improvement over using either source alone (*cf*. Tab. A5).

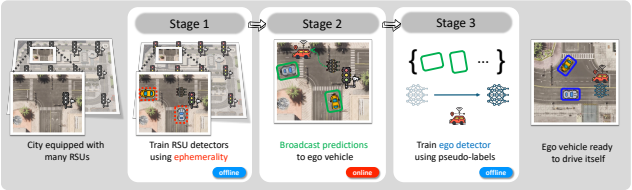

Figure 2. **Overview of infrastructure-taught, label-free 3D perception.** Stage 1: each RSU learns a location-specialized detector in an unsupervised manner by exploiting temporal consistency from its stationary viewpoint. Stage 2: trained RSUs broadcast their predicted 3D bounding boxes to nearby ego vehicles when their fields of view overlap. Stage 3: the ego vehicle aggregates these predictions as pseudo-labels to train its own detector offline, producing a standalone ego model that no longer requires infrastructure at deployment time.

Remark. Our setting is distinct from prior V2X and collaborative perception works [24, 41, 64, 75, 84], which perform online multi-agent fusion at inference time and rely on labeled training data. We use infrastructure only during training: RSUs act as unsupervised, stationary teachers whose predictions are distilled into a single ego-only model. Training is label-free, inference is ego-only, and no V2X link is needed at test time. Moreover, our goal is not to replace existing *ego-centric* unsupervised approaches [2, 19, 52, 74, 78], but to *complement* them with a new supervision source. Rather than proposing another architecture within the standard ego-only regime, our paradigm changes *where the training signals originate*, offering a complementary path toward more label-efficient learning. Please see Sec. A.2 for additional discussion on related work.

2. Label-Free 3D Perception

We instantiate *infrastructure-taught, label-free 3D perception* as a fully label-free framework for training an ego 3D detector using roadside infrastructure as supervision. At a high level, the framework operates in three stages (see Fig. 2): (1) each RSU learns a detector from its own unlabeled observations, (2) trained RSUs broadcast predictions to passing ego vehicles, and (3) the ego detector is trained using these infrastructure-generated pseudo-labels.

2.1. Problem Setup

We consider a geo-fenced city instrumented with a set of RSUs, denoted as $\mathcal{R} = \{R_1, R_2, \ldots, R_M\}$. Each R_i is equipped with sensors (*e.g.*, cameras and LiDAR) and a local 3D detector f_R^i. Leveraging their static viewpoints, RSUs train these detectors in a label-free manner using tem-

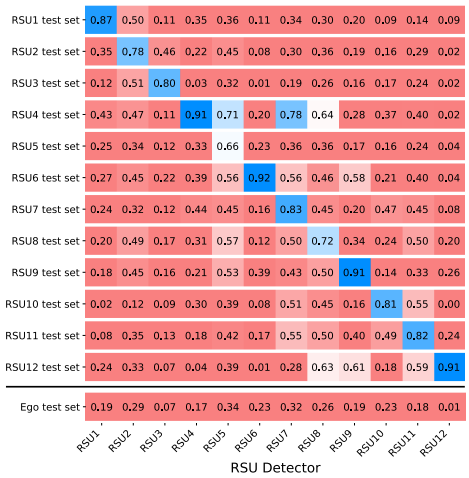

Figure 3. **Cross-evaluation across RSUs.** Each RSU detector is trained at its own location and evaluated across all other RSU viewpoints. The heatmap shows that detectors perform strongly within their own trained view but degrade substantially when evaluated at different locations, indicating that RSU detectors are location-specific. Furthermore, detectors trained at RSU viewpoints fail to generalize to the ego viewpoint (last row).

poral consistency. The resulting detectors produce predictions $Y_R^i = f_R^i(X_R^i)$ from sensory inputs X_R^i.

An ego vehicle E drives through the same city, collecting its own sensory data X_E. Whenever it is within the communication range of an RSU, it receives predictions from nearby RSUs, which are transformed into ego-centric coordinates using the known rigid transformation $\mathcal{T}_{R_i \to E}$. The ego aggregates valid predictions from nearby RSUs to obtain pseudo-labels $\mathcal{Y}_E$. Together with X_E, these form the training set $\mathcal{D}_E = \{(X_E, \mathcal{Y}_E)\}$ used to train the ego's own detector f_E offline.

Our goal is to train an ego-centric detector f_E using only label-free supervision transferred from the RSUs, such that it can operate independently at deployment without manual labels or infrastructure communication.

2.2. Overall Pipeline

Stage 1: Unsupervised RSU Training. Manual annotation may be feasible for a handful of RSUs, but it becomes impractical as infrastructure scales. Thus, the first stage of our pipeline enables each RSU to train its own detector f_R^i in a fully label-free manner, using only its continuous local observations.

We exploit the *static nature of RSUs*: their viewpoints remain constant over time, allowing them to repeatedly observe the same environment under varying traffic dynamics. In such settings, background regions remain temporally consistent, while dynamic objects—such as vehicles, cyclists, and pedestrians—appear and disappear across frames. This temporal contrast provides a natural self-supervised cue for discovering moving objects without human labels.

Table 1. **Quality of broadcasted pseudo-labels for ego training in Stage 2.** Pseudo-labels generally achieve high recall, while communication noise degrades localization quality.

Annotation source	Comm. Noise	Car Prec./Rec.	Ped. Prec./Rec.	Cyc. Prec./Rec.
PP score	✓	92.4/83.8	22.7/19.7	29.5/45.3
PP score	–	98.3/89.1	77.9/67.6	54.9/84.3
PP score + tracking	✓	91.0/83.6	23.8/23.6	33.9/52.9
PP score + tracking	–	97.3/89.5	74.3/73.7	60.7/94.7
Supervised RSUs	✓	92.7/94.2	23.7/17.2	50.8/51.6
Supervised RSUs	–	97.2/98.8	90.2/65.5	93.7/95.2

Concretely, for each RSU $R_i \in \mathcal{R}$ equipped with a LiDAR sensor, we adopt the *persistence point score* (PP score) proposed in MODEST [74] to quantify the temporal stability of points across repeated observations of the same scene. Points with low persistence are identified as likely dynamic and are used as seeds for object proposals. We apply DBSCAN [9] to cluster these transient points into coarse pseudo-labels for each frame.

To further improve the spatial and temporal consistency of the generated boxes, we incorporate a tracking method [78] that smooths trajectories and suppresses spurious initial boxes. Specifically, we track each object and select its most reliable annotation—the frame in which the object is closest to the RSU in the x–y plane. The box size from this frame is then used to adjust the sizes of all remaining boxes corresponding to the same object. The resulting pseudo-labels Y_R^i serve as supervision to train each RSU detector f_R^i.

Through this process, each RSU becomes specialized to its own fixed field of view and produces local detections that later serve as supervision for the ego detector. As shown in Fig. 3, these unsupervised RSU detectors perform strongly within their own views, achieving 82.8% AP on average across all RSUs. This local specialization forms the basis for the next RSU-to-ego supervision stage.

Stage 2: Label Transfer and Aggregation. After the RSUs have been trained, the ego vehicle traverses the town to collect sensory data for training its own detector. Instead of relying on manual annotations, we transfer supervision from infrastructure: each RSU, equipped with its self-learned detector, broadcasts its predictions to the ego vehicle in an online manner.

Specifically, when the ego enters the field of view of an RSU $R_i \in \mathcal{R}$, the RSU transmits its predictions $Y_R^i = f_R^i(X_R^i)$, which are transformed into the ego's coordinate using $\mathcal{T}_{R_i \to E}$. The ego aggregates all such predictions as pseudo-labels aligned with its own sensory data, thereby collecting supervision as it drives.

To improve pseudo-label quality, boxes that are too dis-

tant or sparsely populated by LiDAR points are filtered out. In regions with overlapping RSU coverage, multiple predictions for the same object may occur. We address this with distance-weighted non-maximum suppression (NMS), which scores each box by the inverse distance between the RSU and the detected object, thereby prioritizing closer and typically more reliable observations. This yields a unified pseudo-label set $\mathcal{Y}_E$. We note that more sophisticated aggregation strategies are possible; however, we keep the aggregation module simple yet workable, as our goal is to study the feasibility of infrastructure-taught supervision rather than to optimize this component.

By repeating this process, the ego vehicle constructs its own pseudo-labeled dataset $\mathcal{D}_E$ without any manual annotation. As shown in Tab. 1, the RSUs collaboratively provide high-quality pseudo-labels for the ego vehicle, achieving recall of 89.5%, 73.7%, and 94.7% for cars, pedestrians, and cyclists, respectively.

Stage 3: Ego Detector Training. With pseudo-labels $\mathcal{D}_E = \{(X_E, \mathcal{Y}_E)\}$ gathered across the city, the ego vehicle trains its own 3D detector f_E entirely offline. Unlike the RSU specialists, which are tied to fixed viewpoints, the ego detector integrates supervision transferred from multiple regions into a single model. Once trained, it operates independently at test time without requiring RSUs or manual annotations.

2.3. CIVET Dataset

To study the proposed paradigm under controlled conditions, we construct a simulated multi-agent dataset, CIVET, built on CARLA [8] using V2XVerse [29]. Our goal is not merely to collect multi-agent data, but to create a controlled setting that enables systematic analysis under varying RSU configurations and geographic layouts. Details are provided in Suppl. Sec. C.

3. Main Experiments and Analysis

We primarily report results on Town 10 (urban), which exhibits higher object density; additional results are provided in the Suppl. Sec. D. We evaluate multiple configurations of the proposed pipeline to understand the key factors governing infrastructure-taught ego learning. The main quantitative results are reported in Tab. 2, and we elaborate key findings below.

First, comparing ③❸ with the ego upper bound (⋆), we find that the fully label-free pipeline is *feasible* in practice, achieving competitive performance across object categories for both PointPillars [18] and CenterPoint [71]. This indicates that infrastructure-taught supervision can recover a substantial portion of the supervised upper bound.

Second, comparing ①❶, ③❸, and ⑥❻, we find that pseudo-labels generated by unsupervised RSU detectors provide effective supervision for ego training, approaching the performance of pseudo-labels produced by *supervised RSU detectors*. Incorporating tracking consistently improves ego performance.

Table 2. **Systematic study of infrastructure-taught ego training.** We vary RSU supervision and communication noise to analyze their impact on downstream ego detector performance. "Refine" denotes applying heuristic box refinement [33] to improve pseudo-label quality before ego training.

	Stage 1 (RSU training)		Stage 2 (Broadcasting)		Stage 3 (Ego testing)			
ID	Annotation source	Method	Comm. noise	Refine	Car	Ped.	Cyc.	Avg.
Ego detector: PointPillars [18]								
①	12 RSUs (unsup.)	PP score	–	–	74.3	77.0	68.0	73.1
②	12 RSUs (unsup.)	PP score	✓	–	77.1	49.4	57.0	61.2
③	12 RSUs (unsup.)	PP score + tracking	–	–	79.3	77.5	79.0	78.6
④	12 RSUs (unsup.)	PP score + tracking	✓	–	80.3	49.1	69.6	66.3
⑤	12 RSUs (unsup.)	PP score + tracking	✓	✓	76.5	52.7	79.8	69.7
⑥	12 RSUs (sup.)	–	–	–	92.1	82.2	92.8	89.0
⑦	12 RSUs (sup.)	–	✓	–	90.1	40.3	86.2	72.2
⋆	Ego ground-truth	–	–	–	94.4	88.0	93.7	92.0
Ego detector: CenterPoint [71]								
❶	12 RSUs (unsup.)	PP score	–	–	78.7	78.3	61.1	72.7
❷	12 RSUs (unsup.)	PP score	✓	–	79.2	45.2	49.5	58.0
❸	12 RSUs (unsup.)	PP score + tracking	–	–	82.3	79.3	68.5	76.7
❹	12 RSUs (unsup.)	PP score + tracking	✓	–	80.4	52.3	66.4	66.4
❺	12 RSUs (unsup.)	PP score + tracking	✓	✓	77.7	56.0	75.7	69.8
❻	12 RSUs (sup.)	–	–	–	93.9	84.4	93.4	90.6
❼	12 RSUs (sup.)	–	✓	–	92.4	35.9	87.3	71.9
⋆	Ego ground-truth	–	–	–	94.4	91.8	96.6	94.3

Third, comparing ①❶ with ②❷, and ③❸ with ④❹, we observe that simulated communication imperfections degrade ego detector performance, with the largest drop occurring for pedestrians. Since pedestrians are smaller objects, they are more sensitive to RSU-to-ego misalignment, which is also reflected in the pseudo-label quality reported in Tab. 1. This highlights pseudo-label refinement as an important direction for future work.

Finally, we explore a baseline strategy to mitigate communication noise. Inspired by [33], we apply a heuristic box refinement method that samples candidate boxes around each pseudo-label and selects the highest-scoring candidate using heuristics (*e.g.*, the number of points inside the box and boundary alignment; details in Suppl. Sec. B). Comparing ④❹ with ⑤❺, we observe consistent average AP improvements. This suggests that improved alignment and calibration techniques from V2X could further benefit our infra-taught ego learning.

4. Conclusion

We formulate *infrastructure-taught, label-free* 3D perception, a new paradigm in which stationary RSUs learn from their own observations and supervise ego vehicles. Through a controlled concept-and-feasibility study, we show that distributed RSUs can provide useful supervision for training standalone ego detectors, highlighting the potential for city infrastructure itself to provide a *scalable supervisory signal* for autonomous vehicles. We hope this work provides a concrete foundation for further exploration of infrastructure-taught ego learning.

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *CVPR*, 2022. 8

[2] Stefan Baur, Frank Moosmann, and Andreas Geiger. Liso: Lidar-only self-supervised 3d object detection. In *ECCV*, 2024. 2, 8

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 8

[4] Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In *ACM/IEEE Symposium on Edge Computing*, 2019. 8

[5] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *CVPR*, 2017. 8

[6] Xiangyu Chen, Zhenzhen Liu, Katie Z Luo, Siddhartha Datta, Adhitya Polavaram, Yan Wang, Yurong You, Boyi Li, Marco Pavone, Wei-Lun Chao, et al. Diffubox: Refining 3d object detection with point diffusion. In *NeurIPS*, 2024. 8

[7] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 8

[8] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *CoRL*, 2017. 2, 4, 8

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996. 3

[10] European Commission and CINEA. C-roads platform: Cooperative intelligent transport systems deployment in europe, 2023. European Cooperative Intelligent Transport Systems (C-ITS) deployment platform. 1

[11] Xiangbo Gao, Runsheng Xu, Jiachen Li, Ziran Wang, Zhiwen Fan, and Zhengzhong Tu. Stamp: Scalable task and model-agnostic collaborative perception. In *ICLR*, 2025. 8

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 8

[13] Ruiyang Hao, Siqi Fan, Yingru Dai, Zhenlin Zhang, Chenxi Li, Yuntian Wang, Haibao Yu, Wenxian Yang, Jirui Yuan, and Zaiqing Nie. Rcooper: A real-world large-scale dataset for roadside cooperative perception. In *CVPR*, 2024. 8

[14] Shixin Hong, Yu Liu, Zhi Li, Shaohui Li, and You He. Multi-agent collaborative perception via motion-aware robust communication network. In *CVPR*, 2024. 8

[15] Yue Hu, Juntong Peng, Sifei Liu, Junhao Ge, Si Liu, and Siheng Chen. Communication-efficient collaborative perception via information filling with codebook. In *CVPR*, 2024. 8

[16] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 8

[17] Wentao Jiang, Hao Xiang, Xinyu Cai, Runsheng Xu, Jiaqi Ma, Yikang Li, Gim Hee Lee, and Si Liu. Optimizing the placement of roadside lidars for autonomous driving. In *ICCV*, 2023. 15

[18] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019. 2, 4, 8, 9, 15, 17

[19] In-Jae Lee, Mungyeom Kim, Kwonyoung Ryu, Pierre Musacchio, and Jaesik Park. Openbox: Annotate any bounding boxes in 3d. In *NeurIPS*, 2025. 2

[20] Ted Lentsch, Holger Caesar, and Dariu Gavrila. Union: Unsupervised 3d object detection using object appearance-based pseudo-classes. In *NeurIPS*, 2024. 8

[21] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *RAL*, 2022. 2, 8

[22] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *AAAI*, 2023. 8

[23] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *AAAI*, 2023. 8

[24] Yiming Li, Zhiheng Li, Nuo Chen, Moonjun Gong, Zonglin Lyu, Zehong Wang, Peili Jiang, and Chen Feng. Multiagent multitraversal multimodal self-driving: Open mars dataset. In *CVPR*, 2024. 2, 8

[25] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *TPAMI*, 2024. 8

[26] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022. 8

[27] Chuandong Liu, Chenqiang Gao, Fangcen Liu, Jiang Liu, Deyu Meng, and Xinbo Gao. Ss3d: Sparsely-supervised 3d object detection from point cloud. In *CVPR*, 2022. 8

[28] Chang Liu, Xiaoyan Qian, Binxiao Huang, Xiaojuan Qi, Edmund Lam, Siew-Chong Tan, and Ngai Wong. Multimodal transformer for automatic 3d annotation and object detection. In *ECCV*, 2022. 8

[29] Genjia Liu, Yue Hu, Chenxin Xu, Weibo Mao, Junhao Ge, Zhengxiang Huang, Yifan Lu, Yinda Xu, Junkai Xia, Yafei Wang, et al. Towards collaborative autonomous driving: Simulation platform and end-to-end system. *TPAMI*, 2025. 2, 4, 8

[30] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *ICCV*, 2023. 8

[31] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 8

[32] Yifan Lu, Yue Hu, Yiqi Zhong, Dequan Wang, Siheng Chen, and Yanfeng Wang. An extensible framework for open heterogeneous collaborative perception. In *ICLR*, 2024. 8

[33] Katie Z 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. In *NeurIPS*, 2023. 4, 8, 9, 17

[34] Katie Z Luo, Minh-Quan Dao, Zhenzhen Liu, Mark Campbell, Wei-Lun Chao, Kilian Q Weinberger, Ezio Malis, Vincent Fremont, Bharath Hariharan, Mao Shan, et al. Mixed signals: A diverse point cloud dataset for heterogeneous lidar v2x collaboration. In *ICCV*, 2025. 2, 8

[35] Cong Ma, Lei Qiao, Chengkai Zhu, Kai Liu, Zelong Kong, Qing Li, Xueqi Zhou, Yuheng Kan, and Wei Wu. Holovic: Large-scale dataset and benchmark for multi-sensor holographic intersection and vehicle-infrastructure cooperative. In *CVPR*, 2024. 8

[36] Tao Ma, Xuemeng Yang, Hongbin Zhou, Xin Li, Botian Shi, Junjie Liu, Yuchen Yang, Zhizheng Liu, Liang He, Yu Qiao, et al. Detzero: Rethinking offboard 3d object detection with long-term sequential point clouds. In *ICCV*, 2023. 8

[37] 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 *ECCV*, 2022. 8

[38] Tai-Yu Pan, Sooyoung Jeon, Mengdi Fan, Jinsu Yoo, Zhenyang Feng, Mark Campbell, Kilian Q Weinberger, Bharath Hariharan, and Wei-Lun Chao. Transfer your perspective: Controllable 3d generation from any viewpoint in a driving scene. In *CVPR*, 2025. 8

[39] Charles R Qi, Yin Zhou, Mahyar Najibi, Pei Sun, Khoa Vo, Boyang Deng, and Dragomir Anguelov. Offboard 3d object detection from point cloud sequences. In *CVPR*, 2021. 8

[40] Jenny Seidenschwarz, Aljoša Ošep, Francesco Ferroni, Simon Lucey, and Laura Leal-Taixé. Semoli: What moves together belongs together. In *CVPR*, 2024. 8

[41] Karthikeyan Chandra Sekaran, Markus Geisler, Dominik Rößle, Adithya Mohan, Daniel Cremers, Wolfgang Utschick, Michael Botsch, Werner Huber, and Torsten Schön. Urbaning-v2x: A large-scale multi-vehicle, multi-infrastructure dataset across multiple intersections for cooperative perception. In *NeurIPS*, 2025. 2, 8

[42] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*, 2019. 8, 15

[43] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *CVPR*, 2020. 8

[44] Zhiying Song, Lei Yang, Fuxi Wen, and Jun Li. Traf-align: Trajectory-aware feature alignment for asynchronous multi-agent perception. In *CVPR*, 2025. 8

[45] 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 *CVPR*, 2020. 8

[46] Zongheng Tang, Yi Liu, Yifan Sun, Yulu Gao, Jinyu Chen, Runsheng Xu, and Si Liu. Cost: Efficient collaborative perception from unified spatiotemporal perspective. In *ICCV*, 2025. 8

[47] U.S. Department of Transportation, Intelligent Transportation Systems Joint Program Office. Connected vehicle pilot deployment program phase 3: Final report. Technical report, USDOT ITS JPO, 2023. 1

[48] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *CVPR*, 2021. 8

[49] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *ICCV*, 2023. 8

[50] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *ECCV*, 2020. 8

[51] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Train in germany, test in the usa: Making 3d object detectors generalize. In *CVPR*, 2020. 8

[52] Hai Wu, Shijia Zhao, Xun Huang, Chenglu Wen, Xin Li, and Cheng Wang. Commonsense prototype for outdoor unsupervised 3d object detection. In *CVPR*, 2024. 2, 8, 9

[53] Hai Wu, Hongwei Lin, Xusheng Guo, Xin Li, Mingming Wang, Cheng Wang, and Chenglu Wen. Motal: Unsupervised 3d object detection by modality and task-specific knowledge transfer. In *ICCV*, 2025.

[54] Hai Wu, Shijia Zhao, Xun Huang, Qiming Xia, Chenglu Wen, Li Jiang, Xin Li, and Cheng Wang. Unsupervised 3d object detection by commonsense clue. *TPAMI*, 2025. 8

[55] Qiming Xia, Jinhao Deng, Chenglu Wen, Hai Wu, Shaoshuai Shi, Xin Li, and Cheng Wang. Coin: Contrastive instance feature mining for outdoor 3d object detection with very limited annotations. In *ICCV*, 2023. 8

[56] Qiming Xia, Wei Ye, Hai Wu, Shijia Zhao, Leyuan Xing, Xun Huang, Jinhao Deng, Xin Li, Chenglu Wen, and Cheng Wang. Hinted: Hard instance enhanced detector with mixed-density feature fusion for sparsely-supervised 3d object detection. In *CVPR*, 2024. 8

[57] Qiming Xia, Wenkai Lin, Haoen Xiang, Xun Huang, Siheng Chen, Zhen Dong, Cheng Wang, and Chenglu Wen. Learning to detect objects from multi-agent lidar scans without manual labels. In *CVPR*, 2025. 8

[58] Hao Xiang, Zhaoliang Zheng, Xin Xia, Runsheng Xu, Letian Gao, Zewei Zhou, Xu Han, Xinkai Ji, Mingxi Li, Zonglin Meng, et al. V2x-real: a largs-scale dataset for vehicle-to-everything cooperative perception. In *ECCV*, 2024. 8

[59] Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection. In *ICCV*, 2023. 8

[60] Junhao Xu, Yanan Zhang, Zhi Cai, and Di Huang. Cosdh: Communication-efficient collaborative perception

via supply-demand awareness and intermediate-late hybridization. In *CVPR*, 2025. 8

[61] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. In *CoRL*, 2022.

[62] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *ECCV*, 2022. 8

[63] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *ICRA*, 2022. 2, 8, 9

[64] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *CVPR*, 2023. 2, 8

[65] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 2018. 8

[66] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *CVPR*, 2023. 8

[67] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *CVPR*, 2021. 8

[68] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d++: Denoised self-training for unsupervised domain adaptation on 3d object detection. *TPAMI*, 2022. 8

[69] Yuxue Yang, Lue Fan, and Zhaoxiang Zhang. Mixsup: Mixed-grained supervision for label-efficient lidar-based 3d object detection. *arXiv preprint arXiv:2401.16305*, 2024. 8

[70] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Ipod: Intensive point-based object detector for point cloud. *arXiv preprint arXiv:1812.05276*, 2018. 8

[71] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, 2021. 2, 4, 8, 9, 12, 15, 16, 17

[72] Jinsu Yoo, Zhenyang Feng, Tai-Yu Pan, Yihong Sun, Cheng Perng Phoo, Xiangyu Chen, Mark Campbell, Kilian Q. Weinberger, Bharath Hariharan, and Wei-Lun Chao. Learning 3d perception from others' predictions. In *ICLR*, 2025. 8

[73] Yurong You, Katie Z Luo, Xiangyu Chen, Junan Chen, Wei-Lun Chao, Wen Sun, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Hindsight is 20/20: Leveraging past traversals to aid 3d perception. In *ICLR*, 2022. 8

[74] Yurong You, Katie Z Luo, Cheng Perng Phoo, Wei-Lun Chao, Wen Sun, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Learning to detect mobile objects from lidar scans without labels. In *CVPR*, 2022. 2, 3, 8, 9, 10, 15

[75] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *CVPR*, 2022. 2, 8

[76] Yunshuang Yuan, Yan Xia, Daniel Cremers, and Monika Sester. Sparsealign: A fully sparse framework for cooperative object detection. In *CVPR*, 2025. 8

[77] Dingyuan Zhang, Dingkang Liang, Zhikang Zou, Jingyu Li, Xiaoqing Ye, Zhe Liu, Xiao Tan, and Xiang Bai. A simple vision transformer for weakly semi-supervised 3d object detection. In *ICCV*, 2023. 8

[78] 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 *CVPR*, 2023. 2, 3, 8, 10, 14, 15

[79] Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. Harnessing uncertainty-aware bounding boxes for unsupervised 3d object detection. In *ICCV*, 2025. 8

[80] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021. 8

[81] Shijia Zhao, Qiming Xia, Xusheng Guo, Pufan Zou, Maoji Zheng, Hai Wu, Chenglu Wen, and Cheng Wang. Sp3d: Boosting sparsely-supervised 3d object detection via accurate cross-modal semantic prompts. In *CVPR*, 2025. 8

[82] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, 2018. 8

[83] Zewei Zhou, Seth Z Zhao, Tianhui Cai, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Turbotrain: Towards efficient and balanced multi-task learning for multi-agent perception and prediction. In *ICCV*, 2025. 8

[84] Walter Zimmer, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois C Knoll. Tumtraf v2x cooperative perception dataset. In *CVPR*, 2024. 2, 8

When the City Teaches the Car: Label-Free 3D Perception from Infrastructure

Supplementary Material

In this supplementary material, we provide additional details and experimental results that complement the main paper. Specifically, we include:

- Sec. A: Additional discussion.
- Sec. B: Additional experimental setups.
- Sec. C: Further details on the curated dataset.
- Sec. D: Expanded experimental results.

A. Additional Discussion

A.1. Disclosure of LLM Usage

Portions of this manuscript were edited for clarity and readability using a large language model (LLM). The LLM was not used to generate research ideas, design experiments, analyze data, or draw conclusions. All scientific content, methodologies, and results are the authors' original work.

A.2. Related Work

3D Object Detection. Accurate 3D detection is a fundamental building block for reliable autonomous driving. With the release of large-scale benchmarks [3, 12, 45], researchers have developed a wide range of approaches, including LiDAR-based [18, 42, 43, 65, 70, 71, 80, 82], camera-based [16, 22, 23, 25, 26, 30, 49, 66], and fusion-based [1, 5, 31, 59] methods. These models typically rely on large-scale labeled training data and often suffer from domain shifts when deployed in environments that differ from their training domains. Our proposed paradigm facilitates scaling to new environments by providing supervision without manual annotation.

Label-Efficient Learning. To alleviate the high cost of manual annotations—particularly for 3D point cloud data—numerous label-efficient methods have been explored. Representative directions include unsupervised learning [2, 7, 20, 33, 37, 40, 52–54, 57, 67, 73, 74, 78, 79], semi-supervised learning [27, 28, 48, 55, 56, 69, 77, 81], and domain adaptation [6, 51, 67, 68], as well as offboard detection [36, 39]. Despite remarkable progress, these approaches remain fundamentally *ego-centric*: supervision originates from the ego vehicle's own observations. Recent work [72] has shown that leveraging perception from a nearby expert vehicle can reduce ego annotation cost. We extend this idea to a city-wide, fully label-free ecosystem by treating distributed *self-learning* RSUs as complementary teachers that collectively supervise the ego vehicle.

Multi-Agent Collaborative Perception. Collaborative perception has gained attention as a means to mitigate occlusion and range limitations inherent to single-agent sensing. Typical approaches enable agents (*e.g.*, vehicles and infrastructures) to exchange raw sensory data, intermediate neural features, or predicted objects to achieve better scene understanding [4, 11, 14, 15, 32, 38, 44, 46, 50, 60–62, 64, 75, 76, 83]. In contrast, our work uses infrastructure during training rather than for inference-time fusion. The resulting ego model remains standalone at deployment, without requiring communication at test time.

Datasets for Collaborative Driving. Recent progress has been driven largely by community efforts to collect V2X datasets—covering vehicle-to-vehicle [24, 63, 64], vehicle-to-infrastructure [21, 29, 34, 35, 41, 58, 62, 75, 84], and infrastructure-to-infrastructure [13] settings. However, even recent real-world datasets remain geographically limited—often covering only a handful of intersections and RSUs [34, 41]—due to hardware deployment cost, maintenance overhead, and regulatory constraints. Such limitations make it difficult to systematically study our proposed paradigm, in which a city-wide collection of RSUs can serve as teachers for label-free ego training. Following the common simulation-to-real trajectory in V2X research, we begin by studying the proposed paradigm in a controlled simulation environment to assess its feasibility.

A.3. Future Opportunities

Our study opens several promising directions for future work. First, the current label-free RSU training primarily focuses on dynamic objects by leveraging temporal cues from traffic scenes. Extending the paradigm to static categories is an important next step and may require incorporating additional signals such as scene-level context. Second, while our current pipeline focuses on LiDAR-based detection, integrating multimodal inputs (*e.g.*, camera and LiDAR) could further improve the robustness of infrastructure-taught supervision. Third, our experiments are conducted in a controlled simulation environment, which enables systematic analysis of the proposed paradigm. A natural next step is to validate the pipeline in real-world deployments as connected infrastructure becomes increasingly available.

B. Additional Experimental Setups

B.1. Data Collection

Town Selection. Data are collected in CARLA [8] Town 1, 2, 7, and 10. Among them, Town 7 and 10 pro-

vide complementary rural and urban environments used in our core experiments. To better reflect diversity in deployment conditions, these two towns differ not only in layout but also in object scale and traffic composition: Town 10 contains relatively smaller cars and denser pedestrian and cyclist traffic, whereas Town 7 features larger vehicle models and sparser human activity. Town 1 and 2 are additionally included to study scalability.

Sensor Configuration. We deploy 12 RSUs in each town at diverse locations to enable multi-RSU supervision under varying spatial relationships with the ego vehicle. Each RSU is equipped with four RGB cameras, four depth cameras, and a 128-beam LiDAR. The simulator provides 3D bounding box annotations for three object categories: *Car*, *Pedestrian*, and *Cyclist*, with consistent object IDs across frames. The ego vehicle uses a similar sensor suite—four RGB cameras, four depth cameras, and a 64-beam LiDAR—and additionally has access to a top-down map view and the same box annotations.

Dataset Size. For RSU training, we collect 7k frames per RSU (5k for training and 2k for validation), resulting in 84k RSU frames *per town* across 12 RSUs. For RSU-to-ego broadcasting (ego training), we collect an additional 5k synchronized frames *per town* in which the ego vehicle and multiple RSUs share overlapping fields of view, enabling pseudo-label transfer. A separate set of 3k ego frames *per town* is reserved for evaluating the ego detector. In total, the dataset contains 608k frames across four towns.

Since the data are collected in simulation, larger-scale datasets can be generated if needed. In practice, the current scale already enables systematic evaluation of the proposed paradigm and its components. We therefore keep the dataset size fixed in this study, while noting that scaling up data collection is straightforward in our reproducible simulation environment.

B.2. Implementation Details

All experiments are conducted using LiDAR-based 3D detectors. For unsupervised RSU training, we divide 5k training frames into 10 temporally equidistant segments and compute the PP score for each segment. Following [74], we remove estimated ground points before obtaining initial pseudo-labeled bounding boxes. Each RSU detector adopts CenterPoint [71] and is trained for 20 epochs using an initial learning rate of 5×10^{-4}, which is reduced by a factor of 0.1 at epochs 10 and 15.

During RSU-to-ego broadcasting, we set the maximum communication range to 160m. Following [63], we simulate realistic communication noise by adding Gaussian perturbations to transmitted 3D positions ($\sigma = 0.2\,\mathrm{m}$) and yaw angles ($\sigma = 0.05\,\mathrm{rad}$), together with a fixed 100ms communication delay.[1] To fuse overlapping boxes from multiple RSUs, we apply distance-weighted NMS with an IoU threshold of 0.1. Since each RSU detector outputs class-agnostic boxes, we further assign semantic categories using a template-based class matching scheme that compares box sizes to known object prototypes, inspired by [52]. For ego training, we use two 3D detectors, PointPillars [18] and CenterPoint [71], trained with the same learning schedule and number of epochs as the RSU detectors. All experiments are conducted on two NVIDIA A100 GPUs.

B.2.1. Unsupervised RSU Training

Due to computational constraints, we did not perform dedicated hyperparameter tuning for each unsupervised RSU detector. For pseudo-label generation using MODEST [74], we first calculate PP score using a neighborhood radius of 0.01 across the frames, we then adopt DBSCAN clustering with a neighborhood radius of 0.3 and require at least 2 points to form a valid cluster. After clustering, we apply common-sense filtering rules: each box must contain at least 5 points, its volume must fall between 0.5 and 20, and its vertical extent must be reasonable—specifically, the maximum height must be at least 0.5m and the minimum height must not exceed 2.5m. For tracking refinement, we use the official module provided in [52]. While our study uses a unified set of hyperparameters across all RSUs, we expect that tuning these values individually for each RSU could further improve the quality of unsupervised detection. We leave such fine-grained optimization to future work.

B.2.2. Broadcasted Pseudo-Label Refinement

In real-world deployments, communication imperfections during RSU prediction broadcasting—such as localization drift or temporal asynchrony—can introduce noise into the transmitted pseudo-labels. To mitigate these effects, we adopt a heuristic-based box refinement method [33] as a simple baseline module. Specifically, for each broadcast pseudo-label box, we generate a small set of candidate boxes by translating the box center over a discrete grid of horizontal offsets $(\Delta x, \Delta y)$ while keeping its size and orientation unchanged. Each candidate is scored based on the number of LiDAR points that fall inside the box as well as the geometric tightness of the enclosed points. The candidate with the highest score is selected as the refined box for each actor. If the support is too weak—*i.e.*, the candidate contains fewer than a predefined minimum number of LiDAR points—we retain the original broadcast box instead of replacing it.

[1] Communication imperfections are *shared challenges* across V2X systems rather than being specific to our formulation. We therefore simulate representative noise conditions and explore a simple refinement module (*cf*. Tab. 2 ⑤❺). Advances in synchronization and calibration would directly benefit this setting.

B.2.3. Combining Ego-Centric and Infrastructure Pseudo-Labels

Broadcasted pseudo-labels from infrastructure may occasionally exhibit inaccurate sizes or orientations due to sparse LiDAR observations at long range or communication noise during broadcasting. In contrast, ego-centric pseudo-labeling methods [74, 78] can observe denser LiDAR measurements when objects are closer to the ego vehicle and are not affected by communication noise. As a result, ego-centric approaches may produce more accurate bounding boxes in such cases. Motivated by this insight, we combine pseudo-labels from both sources when generating the pseudo-label supervision used for Tab. A5. For each infrastructure-provided pseudo-label, we check whether an overlapping box exists in the ego-centric pseudo-label set. If both candidates are present, we select the one containing a larger number of ego LiDAR points, which typically indicates better alignment with the ego observation. If no overlapping ego-centric pseudo-label exists, the infrastructure-provided pseudo-label is retained. We see that this simple fusion strategy can effectively leverage the complementary strengths of infrastructure-based supervision and ego-centric observations.

B.3. Evaluation

For evaluation of the unsupervised RSU detectors, we follow MODEST [74] and merge the *Car*, *Pedestrian*, and *Cyclist* categories into a single foreground class. For ego detector evaluation, unless otherwise specified, we report bird's-eye-view (BEV) average precision (AP) at IoU thresholds of 0.5 for cars and 0.3 for pedestrians and cyclists. The detection region is defined as $[-80, 80]$m longitudinally for both RSU and ego detectors, while the lateral range is $[-80, 80]$m for RSUs and $[-40, 40]$m for the ego detector.

C. Additional Details on CIVET

C.1. Data Curation

Town environments. Our main study focuses on Town 7 and Town 10, which capture distinct rural and urban driving conditions, differing in road layout, traffic density, and vehicle composition (Tab. A1). Town 7 represents a low-density rural setting with larger vehicles, whereas Town 10 features dense urban traffic with smaller and more diverse actors. The RSU deployment for both towns is shown in Fig. A1, where each RSU's field of view is visualized; one coverage region is shaded slightly darker for clarity. As illustrated, the RSUs collectively cover most of the drivable area, enabling comprehensive supervision throughout the geo-fenced space. In addition, we collect data from Town 1 and Town 2 to study scalability. The RSU placement for these environments is shown in Fig. A2.

Algorithm 1 Aggregating Knowledge from Many RSUs

Require: Trained RSU detectors $\{f_R^i\}_{i=1}^M$, ego trajectory frames $\{t = 1..T\}$, maximum RSU distance $d_{\max}$

Initialize ego pseudo-label dataset $\mathcal{D}_E = \emptyset$
for each time step $t = 1$ to T **do**
  Acquire ego sensory data $X_E(t)$
  Initialize temporary collection $\mathcal{C} = \emptyset$
  for each RSU R_i such that $\text{dist}(E_t, R_i) < d_{\max}$ **do**
    Obtain RSU prediction $Y_R^i = f_R^i(X_R^i(t))$
    Transform to ego frame $\tilde{Y}_R^i = \mathcal{T}_{R_i \to E}(Y_R^i)$
    $\mathcal{C} \leftarrow \mathcal{C} \cup \tilde{Y}_R^i$
  end for
  Aggregate bounding boxes $\mathcal{C}$ to obtain $\mathcal{Y}_E(t)$
  Store $(X_E(t), \mathcal{Y}_E(t))$ into $\mathcal{D}_E$
end for
return $\mathcal{D}_E$

Figure A1. **RSU deployment and coverage in Town 7 and Town 10.** Each shaded region denotes the field of view of an RSU. Together, the RSUs provide complementary coverage of the geo-fenced environment, enabling infrastructure-taught supervision across most of the drivable area. One region is shaded slightly darker for clearer visualization.

C.1.1. World initialization

We establish a CARLA client and initialize the simulator in synchronous mode with a fixed simulation time step of 1/20 seconds (20 Hz). The Traffic Manager is also configured in synchronous mode to ensure deterministic behavior of background traffic.

C.1.2. Route and scenario loading

For each simulation route, the system first loads the target map and then configures traffic-light behavior depending on the stage of the pipeline. In Stage 1, all traffic lights are disabled to prevent prolonged stopping at intersections, ensuring continuous traffic flow. In Stage 2 and Stage 3, normal traffic-light control is enabled, introducing realistic driving patterns and dynamics for RSU-to-ego broadcasting and ego evaluation.

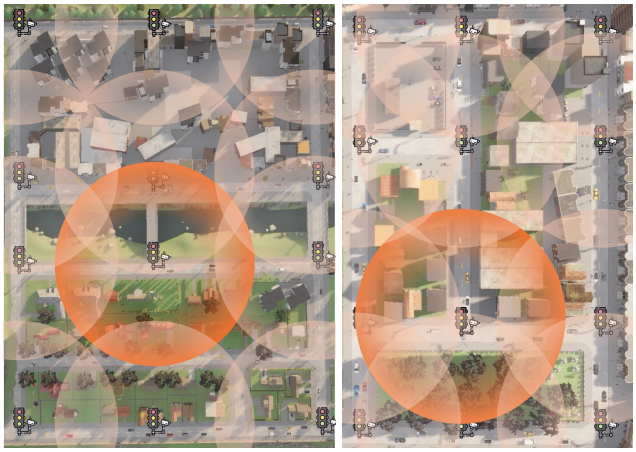

Figure A2. **RSU deployment and coverage in Town 1 and Town 2.** One region is shaded slightly darker for clearer visualization.

Figure A3. **Reference driving paths in Town 7 and Town 10.** The ego vehicle follows high-level reference routes designed to traverse most of the drivable area, ensuring broad coverage during data collection.

C.1.3. Agent initialization

The ego vehicle is initialized at a fixed starting pose and is assigned a high-level reference route for each scenario (Fig. A3 and Fig. A4). Note that this route specifies the intended global path but does not prescribe fine-grained control. As a result, the ego's actual trajectory varies across simulation runs due to interactions with surrounding agents and dynamic traffic conditions. In contrast, all non-ego actors—including cars, pedestrians, and cyclists—are randomly spawned throughout the map at the beginning of each episode. Their motion plans are sampled from CARLA's navigation graph, introducing diverse and unpredictable traffic behaviors. This randomized initialization ensures that, even with the same high-level route, the ego experiences different encounters and visibility patterns in every run, increasing the diversity of collected data and mitigating overfitting to any single traffic configuration.

C.1.4. Closed-loop scenario execution

At each simulation tick, the system runs in a fully closed-loop setup. All ego onboard sensors (RGB cameras, LiDAR, GNSS, IMU) and all 12 RSUs (mounted at 7.5m) produce synchronized observations, which are delivered through CARLA's synchronous `SensorInterface`. The autonomous agent fuses ego-vehicle and RSU observations and outputs control commands (steering, throttle, brake). Although observations are generated every tick, we record them at a downsampled rate of once every two frames to reduce storage and computation.

Figure A4. **Reference driving paths in Town 1 and Town 2.**

C.1.5. Environment and vehicle assets

We intentionally differentiate the domains of Town 7 and Town 10 to study the general applicability of our proposed paradigm. Detailed environmental statistics of Town 7 and Town 10 are summarized in Tab. A1. Town 7 features larger vehicles and sparser traffic, whereas Town 10 contains smaller urban vehicles along with a higher density of pedestrians and cyclists. The physical dimensions of all vehicle models are provided in Tab. A2. Each town uses a distinct subset of vehicles to reflect realistic rural and urban traffic patterns—for example, Town 7 includes sedans, pickups, and trucks, while Town 10 primarily uses compact cars and microcars suited for dense urban streets.

Overall metadata. Each ego frame includes synchronized RGB, depth, LiDAR, semantic LiDAR, segmentation maps, bird's-eye views, camera parameters, and navigation states such as speed, waypoints, and IMU. RSU frames contain multi-view RGB, depth, high-resolution LiDAR, semantic LiDAR, and calibration data. These modalities are retained for all collected frames and used to support different stages of the pipeline and subsequent analysis.

C.2. Sample Data

Fig. A5 presents samples from our curated dataset, with each column showing paired examples from the urban and rural environments. As illustrated, the two towns exhibit clearly distinct visual and semantic characteristics:

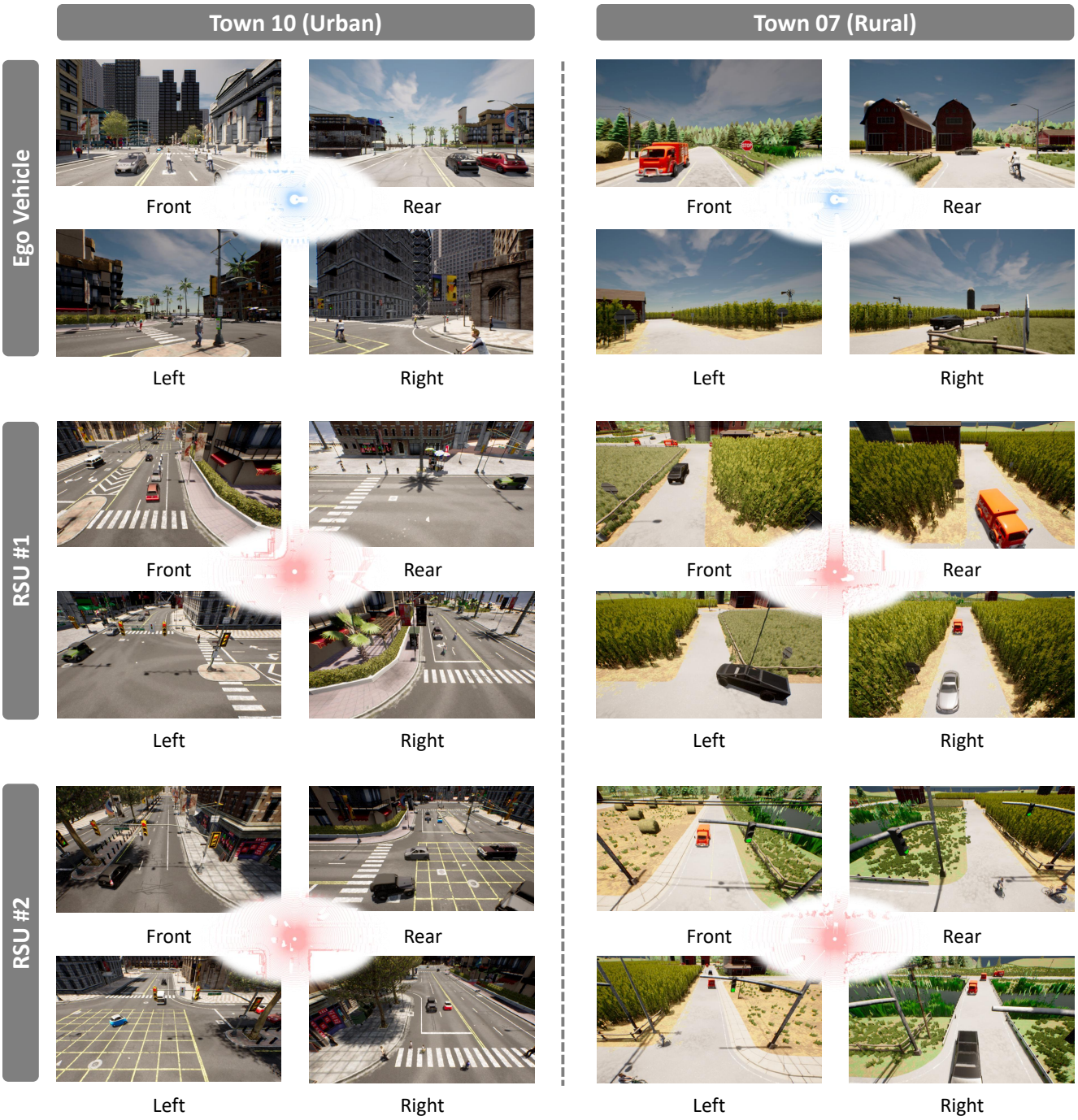

Figure A5. **Sample data from the CIVET dataset (urban and rural towns).** Each column shows paired examples from the urban and rural environments. The two towns exhibit distinct semantics—urban scenes contain smaller vehicles and denser pedestrian/cyclist activity, while rural scenes feature larger vehicles and sparser non-vehicle traffic.

urban scenes contain smaller vehicles and denser pedestrian/cyclist traffic, whereas rural scenes feature larger vehicles and sparser non-vehicle participants. Fig. A6 shows sample scenes from Town 1 and Town 2 used for the scalability study.

D. Additional Experimental Results

D.1. Further Analysis and Insights

We primarily use CenterPoint [71] and urban town as the ego detector in the following analyses.

Figure A6. **Sample data from the CIVET dataset (additional towns).** These environments are used to evaluate the scalability of the proposed framework.

Do static views enable reliable object discovery? As shown in Fig. A7, the fixed viewpoints of RSUs allow them to observe the same scene repeatedly over time, producing highly consistent background points and low-persistence dynamic objects. This results in a discriminative PP-score distribution, enabling clean dynamic-object proposals and reliable pseudo-labels for unsupervised training. Furthermore, as illustrated in Fig. A8, incorporating tracking consistently improves recall, which further improves pseudo-label quality.

Is the paradigm geographically scalable? We conduct an experiment to examine whether the proposed pipeline scales

Table A1. **Comparison of environment statistics between the two towns.** The rural town contains larger vehicles and fewer pedestrians/cyclists, whereas the urban town features smaller vehicles and denser non-vehicle traffic. Actor counts are reported as car/pedestrian/cyclist, and the average car size is given in (length × width × height).

Scene type	# Actors	Avg. car size
Rural	50 / 25 / 25	5.22 × 2.18 × 1.72
Urban	50 / 100 / 50	3.99 × 1.89 × 1.59

Table A2. **Vehicles and their dimensions for each town.** ✓ indicates that the vehicle model is used in the corresponding town.

Vehicle	L (m)	W (m)	H (m)	Rural	Urban
mercedesccc (**ego**)	4.67	2.00	1.44	✓	✓
chevrolet.impala	5.36	2.03	1.41	✓	–
chargercop2020	5.24	2.10	1.64	✓	–
benz.coupe	5.03	2.15	1.64	✓	–
charger2020	5.01	2.10	1.53	✓	–
dodge.charger.police	4.97	2.04	1.55	✓	–
tesla.cybertruck	6.27	2.39	2.10	✓	–
carlacola	5.20	2.61	2.47	✓	–
audi.tt	4.18	1.99	1.39	–	✓
seat.leon	4.19	1.82	1.47	–	✓
toyota.prius	4.51	2.01	1.52	–	✓
nissan.patrol	4.60	1.93	1.85	–	✓
volkswagen.t2	4.48	2.07	2.04	–	✓
citroen.c3	3.99	1.85	1.62	–	✓
mini.cooperst	3.81	1.97	1.48	–	✓
audi.a2	3.71	1.79	1.55	–	✓
nissan.micra	3.63	1.85	1.50	–	✓
bmw.isetta	2.21	1.48	1.38	–	✓
jeep.wrangler.rubicon	3.87	1.91	1.88	–	✓

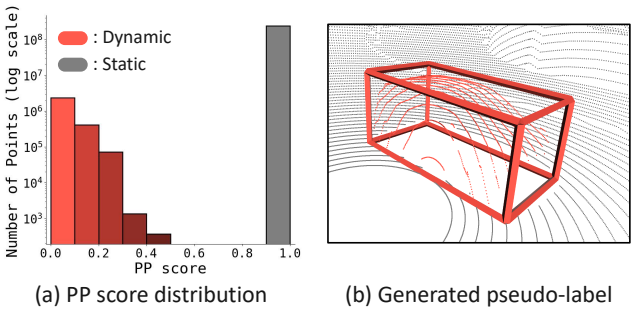

Figure A7. **Effectiveness of PP scores for RSU.** (a) Discriminative distribution allows a clear separation between static background and objects. (b) Pseudo-labels exhibit high localization quality.

to larger environments. Since existing datasets do not provide sufficient geographic coverage for such analysis, we combine our data from all four towns and treat them as a *unified* environment. In this setup, we train RSU detectors in an unsupervised manner across all 48 RSUs (12 per town) using PP-score proposals with tracking, and use their predictions to broadcast pseudo-labels for ego training. A single ego detector is then trained using the aggregated supervision and evaluated on the test splits of all towns. As shown in Tab. A3, the results exhibit trends similar to those observed in Tab. 2, suggesting that the proposed paradigm can naturally extend to larger-scale deployments.

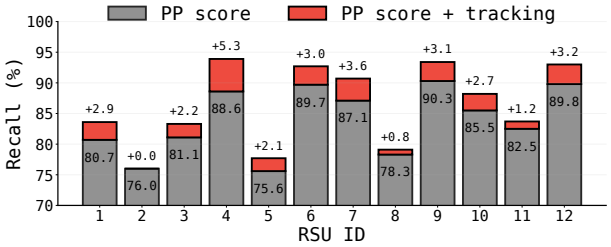

Figure A8. **Effect of tracking refinement [78].** Incorporating tracking improves pseudo-label recall, yielding stronger supervision for unsupervised RSU training.

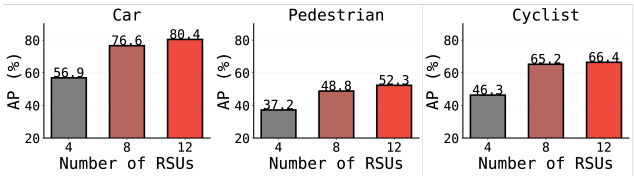

Figure A9. **Effect of the number of RSUs.** Increasing the number of available RSUs improves ego detector performance by providing more diverse and complementary pseudo-label supervision.

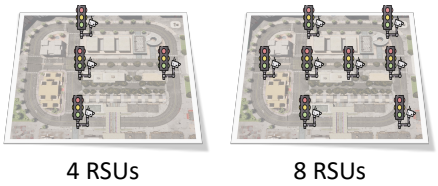

Figure A10. **RSU placement used in the analysis of the number of RSUs.**

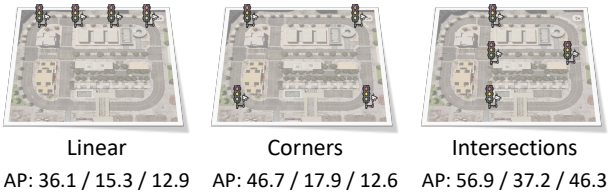

Figure A11. **Effect of RSU placement.** Intersection placement provides more balanced coverage and higher-quality pseudo-labels than linear or corner layouts. APs: *Car*, *Pedestrian*, and *Cyclist*.

Does increasing the number of RSUs help? We investigate how the number of RSUs affects ego detector per-

Table A3. **Scalability study.** We synthesize a unified environment by combining data from four towns and train 48 RSU detectors to generate pseudo-labels for ego training. A single ego detector is then trained using the aggregated supervision. Unlike Tab. 2, evaluation here is performed on the *combined test sets* from all four towns.

Stage 1 (RSU training)	Stage 2 (Broadcasting)		Stage 3 (Ego testing)			
Annotation source	Comm. noise	Refine	Car	Ped.	Cyc.	Avg.
Ego detector: PointPillars [18]						
48 RSUs (unsup.)	–	–	84.1	78.4	80.2	80.9
48 RSUs (unsup.)	✓	–	80.4	56.8	77.8	71.7
48 RSUs (unsup.)	✓	✓	81.7	63.9	81.7	75.8
Ego ground-truth	–	–	92.4	89.3	94.6	92.1
Ego detector: CenterPoint [71]						
48 RSUs (unsup.)	–	–	82.7	81.0	78.3	80.7
48 RSUs (unsup.)	✓	–	78.2	53.5	72.5	68.1
48 RSUs (unsup.)	✓	✓	80.5	66.4	81.5	76.1
Ego ground-truth	–	–	91.0	90.6	95.3	92.3

Table A4. **Adapting a pretrained ego detector using infra-taught pseudo-labels.** A detector trained in a rural town is evaluated in an urban town. Fine-tuning with pseudo-labels from the new town improves performance under this domain shift.

Train domain	Car	Ped.	Cyc.	Avg.
Urban	82.3	79.3	68.5	76.7
Rural	59.2	66.2	67.5	64.3
Rural→Urban	81.9	79.9	76.1	79.3

formance within our experimental setting. For fair comparison, we collect additional data for configurations with 4 and 8 RSUs while keeping the total number of training frames same as the 12-RSU setup. As shown in Fig. A9, performance improves as the ego receives pseudo-labels from more RSUs. This trend suggests that additional RSUs provide more diverse and complementary supervision during ego training. Fig. A10 provides RSU placement for Fig. A9.

Where should RSUs be placed? We next examine how RSU placement affects ego detector performance under a limited-budget scenario where only four RSUs can be deployed. Unlike [17], which studies optimal placement for *online* multi-RSU collaborative perception, our focus is on how placement influences the *offline* training quality of the ego detector. For fair comparison, we collect training data for each placement strategy while keeping the total number of training frames same. As shown in Fig. A11, we compare three layouts: *linear*, *corner*, and *intersection*. Among these, intersection placement yields the strongest performance, suggesting that locations with high traffic flow provide richer pseudo-labels and more diverse supervision for ego training.

Does the pipeline help adapt detectors across domains?

Table A5. **Complementarity with ego-centric unsupervised methods.** Combining our pseudo-labels with existing approaches further improves performance. Evaluation follows [74] with PointRCNN [42] as the detector. "Dyn" denotes dynamic objects.

Anno. source	Noise	Dyn.
MODEST [74]	–	18.0
Oyster [78]	–	39.8
Ours	–	62.0
Ours	✓	49.9
Oyster + Ours	–	64.9
Oyster + Ours	✓	59.5

We study whether infrastructure-taught pseudo-labels can adapt an ego detector trained in a different geographic domain. Specifically, we take CenterPoint [71] trained on a rural town (Town 7) and evaluate it directly on an urban town (Town 10). As shown in Tab. A4, the model performs poorly under this domain shift (*e.g.*, only 59.2% Car AP). We note that this setting is intentionally designed to induce a meaningful domain shift. Fine-tuning the model previously trained in a rural environment with infrastructure-provided pseudo-labels from the current urban town leads to consistent improvements, and in some cases even surpasses training the ego detector from scratch (*e.g.*, Cyclist AP increases to 76.1%, compared to 68.5%). These results indicate that infra-taught supervision can potentially serve as an effective signal for *label-free domain adaptation*.

Can our pseudo-labels complement ego-centric pseudo-labels? Although our goal is not to compete with existing ego-centric unsupervised methods, but rather to explore a new source of supervision, we compare the effectiveness of our pseudo-labels with representative ego-centric approaches: MODEST [74] and Oyster [78]. To isolate the effect of the supervision source, we disable self-training for all methods and tune hyperparameters to the best of our ability. As shown in Tab. A5, training an ego detector using our pseudo-labels alone already outperforms the baselines. More importantly, combining both sources yields further improvements, indicating that infrastructure-taught pseudo-labels and ego-centric learning provide *complementary signals*.

D.2. Unsupervised RSU Training

Pseudo-label quality across depth ranges. To better understand the quality of infrastructure-provided pseudo-labels, we analyze RSU detection performance as a function of the distance to the RSU. Specifically, we break down RSU AP (for all classes) by distance-to-RSU, averaged over all 12 RSUs. As shown in Tab. A6, pseudo-label

Table A6. **Pseudo-label quality across RSU distance ranges.** We report RSU detection AP averaged over all RSUs. Pseudo-label quality is higher for objects closer to the RSU sensor.

Range	0–30m	30–80m
AP	86.5	67.9

Table A7. **Shared vs. independent RSU detectors.** We compare detectors trained independently for each RSU with a single detector trained using data from all RSUs. The shared detector performs slightly better, suggesting that a general RSU model is feasible *when data from multiple RSUs can be combined*.

RSU	Independent	Shared
AP	84.5	85.5

quality is higher for objects located closer to the RSU sensor. This trend is expected, as nearby objects are observed with denser LiDAR measurements and suffer less from sparsity. These observations suggest that incorporating distance-aware strategies when aggregating infrastructure-generated pseudo-labels could further improve the supervision quality.

Can RSU detector be shared? In the main paper, we assume that each RSU is trained independently using its own local observations, without sharing data across different sensors. This assumption reflects a realistic deployment scenario in which infrastructure sensors operate independently. However, in practice, it may be possible for RSUs to share their data, enabling the training of a *shared* detector across multiple RSUs. To investigate this possibility, we train a single detector using the combined data from all RSUs and compare it with detectors trained independently for each RSU. As shown in Tab. A7, the shared detector performs slightly better than independently trained RSU detectors. This result suggests that a general RSU detector is feasible *when data collected from multiple RSUs can be combined together*.

D.3. Ego Detector

Different distance ranges. To further understand the effect of infrastructure-taught supervision, we analyze ego detector performance across different distance ranges. We use CenterPoint [71] trained on the urban town under the ideal communication setting and evaluate Car AP at IoU 0.5 for near (0–30 m), mid-range (30–50 m), and far (50–80 m) distances. As shown in Tab. A8, our method maintains strong performance even for distant objects, recovering a substantial fraction of the supervised upper bound. This highlights an important advantage of infrastructure-taught supervision: RSUs can provide pseudo-labels for objects that are far away or poorly visible from the ego viewpoint, enabling more complete training coverage.

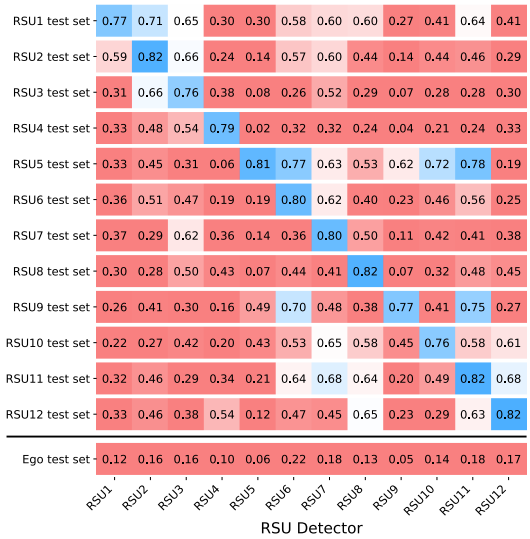

Figure A12. **Cross-evaluation across RSUs in the rural town.** Compared to the urban setting, RSU detectors exhibit slightly better cross-location generalization due to the simpler traffic semantics in rural environments. Nevertheless, each RSU performs best within its own trained field of view and degrades when evaluated elsewhere. The last row further shows that no single RSU detector transfers well to the ego viewpoint.

Table A8. **Ego detector performance across distance ranges.** Reported numbers represent Car AP@0.5 for CenterPoint [71] on the urban town under noise-free communication. "Relative performance" denotes the percentage of the supervised upper-bound AP recovered by our method. Even for the challenging 50–80 m range, the ego detector retains a strong fraction of the upper-bound performance, demonstrating that RSUs can provide high-quality pseudo-labels for distant objects.

Method	0–30m	30–50m	50–80m
Our pipeline	85.0	81.4	76.0
Upper bound	97.1	94.7	86.5
Relative performance (%)	*87.5*	*86.0*	*87.9*

Results on the rural town. As shown in Fig. A12, unsupervised RSUs in the rural town exhibit slightly better cross-location generalization than those in the urban setting, likely due to the simpler scene structure. However, this generalization remains limited: each RSU performs best within its own fixed field of view and degrades noticeably when evaluated elsewhere. This observation highlights the inherently limited spatial coverage of individual infrastructure detectors and motivates collaborative infrastructure supervision. Consistent with the observations from the urban

Table A9. **Infrastructure-taught ego training on the *rural* town.** We analyze the impact of RSU supervision, communication noise, and pseudo-label refinement on ego detector performance. "Refine" denotes applying heuristic box refinement [33] before ego training.

Stage 1 (RSU training)		Stage 2 (Broadcasting)		Stage 3 (Ego testing)			
Annotation source	Method	Comm. noise	Refine	Car	Ped	Cyc	Avg.
Ego detector: PointPillars [18]							
12 RSUs (unsup.)	PP score + tracking	–	–	81.9	76.6	58.7	72.4
12 RSUs (unsup.)	PP score + tracking	✓	–	81.7	23.7	30.7	45.4
12 RSUs (unsup.)	PP score + tracking	✓	✓	83.0	63.1	31.3	59.1
12 RSUs (sup.)	–	–	–	91.0	81.6	76.5	83.0
12 RSUs (sup.)	–	✓	–	92.4	35.0	46.5	58.0
Ego ground-truth	–	–	–	90.8	84.9	77.2	84.3
Ego detector: CenterPoint [71]							
12 RSUs (unsup.)	PP score + tracking	–	–	79.8	75.6	62.4	72.6
12 RSUs (unsup.)	PP score + tracking	✓	–	76.8	25.2	19.1	40.4
12 RSUs (unsup.)	PP score + tracking	✓	✓	78.6	58.2	35.8	57.5
12 RSUs (sup.)	–	–	–	89.7	70.9	63.4	74.7
12 RSUs (sup.)	–	✓	–	89.3	25.6	41.5	52.1
Ego ground-truth	–	–	–	86.5	83.0	69.2	79.6

scene, we also find that no RSU detector generalizes well to the ego viewpoint (last row of Fig. A12).

Quantitative results for the ego detector are reported in Tab. A9. Due to computational constraints, we report only the core evaluations for the rural setting using two ego detectors [18, 71]. Consistent with the findings from the urban setting in the main paper, the ego detector trained with infrastructure-generated pseudo-labels remains close to the supervised upper bound under ideal, noise-free communication for both detectors (*e.g.*, 83.0% AP for detecting vehicles compared to the upper bound of 90.8% on PointPillars [18]). When communication noise is introduced, performance degrades, particularly for pedestrians and cyclists, reflecting the smaller number of pedestrian and cyclist instances available in the rural town for training. Nevertheless, simple heuristic-based box refinement [33] effectively mitigates such communication noise (*e.g.*, over 10 AP improvement on average and over 30 AP improvement for pedestrian detection for both detectors). Overall, these results further demonstrate the feasibility and effectiveness of infrastructure-taught ego training.

D.4. Qualitative Results

Qualitative results in Fig. A13 show that the ego detector [71] produces accurate 3D bounding boxes. Note that these results are obtained under simulated communication noise during RSU broadcasting, which better demonstrates the robustness of our proposed learning pipeline. Additional uncurated examples are provided in Fig. A15 and Fig. A16. We further analyze failure cases in Fig. A14. While both RSU and ego detectors perform well in many scenarios,

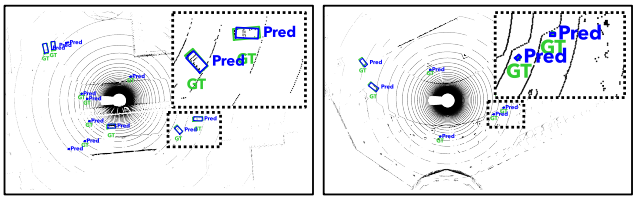

Figure A13. **Qualitative results of the ego detector with CenterPoint [71].** Despite being trained without human labels, the ego detector generates accurate 3D bounding boxes, demonstrating the feasibility of our infrastructure-taught pipeline.

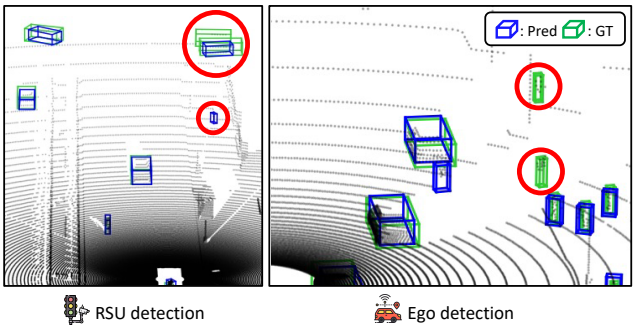

Figure A14. **Failure cases.** While both RSU and ego detectors perform well in many cases, RSUs often miss distant objects and the ego detector struggles with small objects.

RSUs may miss distant objects and the ego detector may struggle with small objects. Overall, these qualitative examples further support the feasibility of infrastructure-taught, label-free 3D perception and highlight promising directions for future improvements.

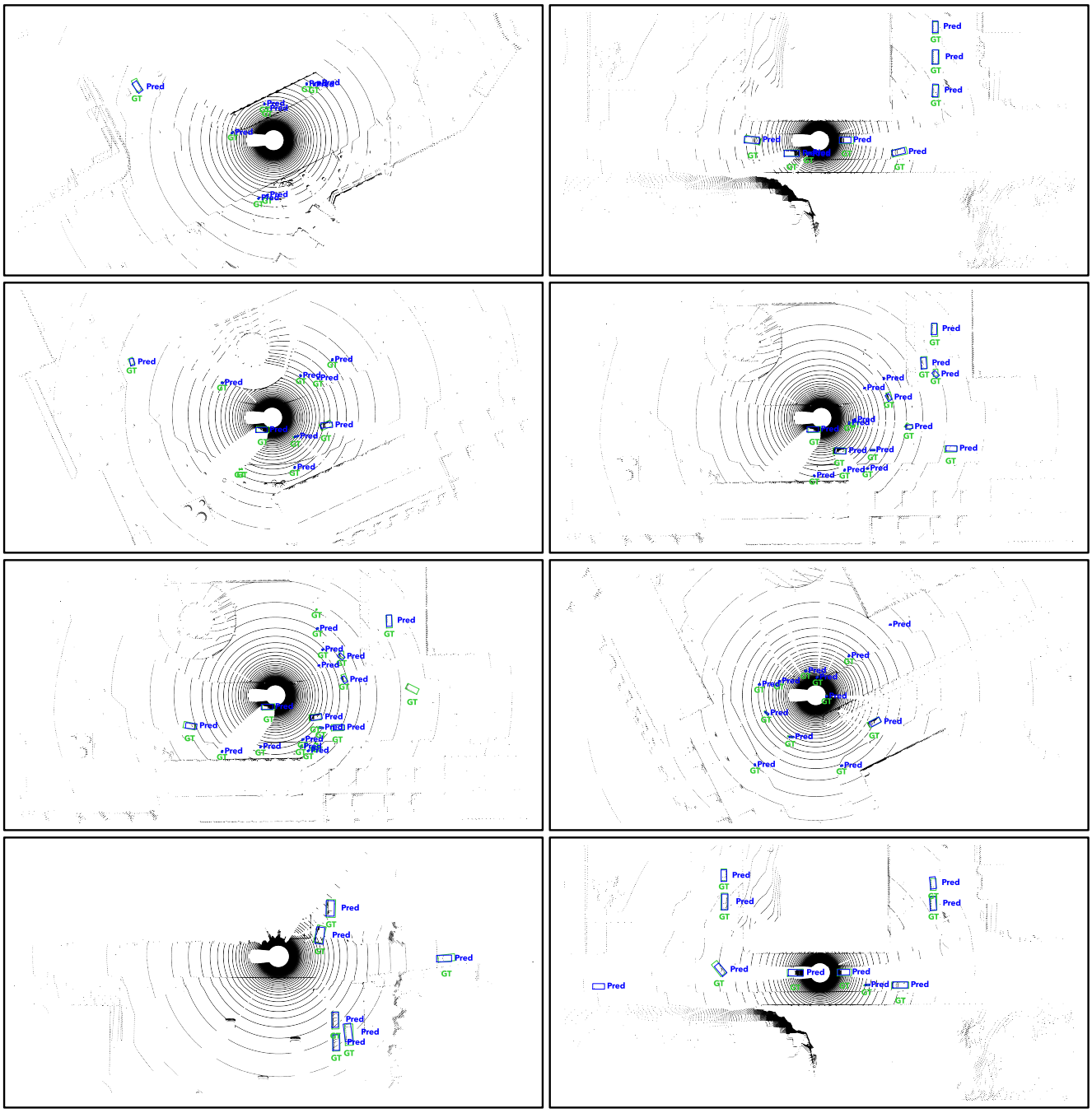

Figure A15. **Additional qualitative results of the ego detector. [1/2]**

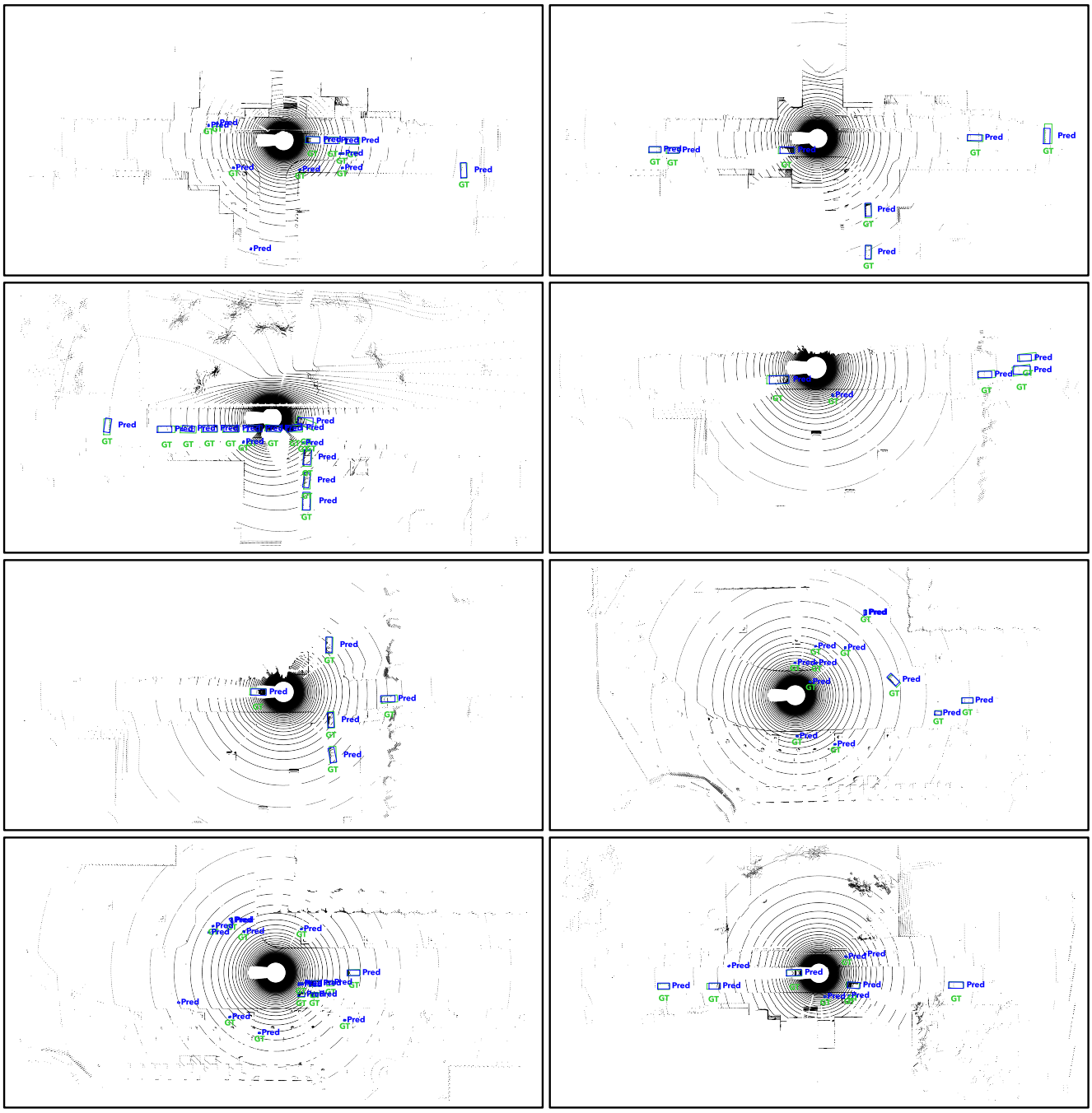

Figure A16. **Additional qualitative results of the ego detector. [2/2]**