Learning from Distributed Eyes: Leveraging Collaborative Perception for Automated Model Adaptation

Yanan Ma[1], Yihang Tao[1], Zhengru Fang[1], Zihan Fang[1], Yiqin Deng[2]
Xianhao Chen[3], and Yuguang Fang[1]

***Abstract*— In autonomous driving, perception models often struggle to generalize to new environments due to domain shifts. While unsupervised model adaptation offers a feasible solution without labor-intensive manual labeling, existing methods that rely solely on the ego-vehicle's data often lead to inferior pseudo-labeling performance. To address this critical issue, we propose `LDE`, Learning from Distributed "Eyes", a novel framework that transforms collaborative perception (CP) into a source of high-quality supervision for model adaptation. This pseudo-labeling approach is hyperparameter-insensitive and relatively reliable, assuming CP often outperforms single-agent's perception. However, naively implementing this approach encounters (1) the communication bottleneck of sharing rich features under time and bandwidth constraints, (2) the view discrepancy between the CP view and the learner's Field of View (FoV), and (3) the unreliability even in CP-generated labels. To address these issues, we design an adaptation-oriented feature sharing mechanism that selectively transmits the most critical information for adaptation, an FoV filtering method that meticulously eliminates mismatched labels, and a curriculum learning strategy to progressively exploit pseudo labels. Extensive experiments on 3D object detection tasks demonstrate that `LDE` consistently outperforms both the pre-trained models and state-of-the-art unsupervised adaptation methods.**

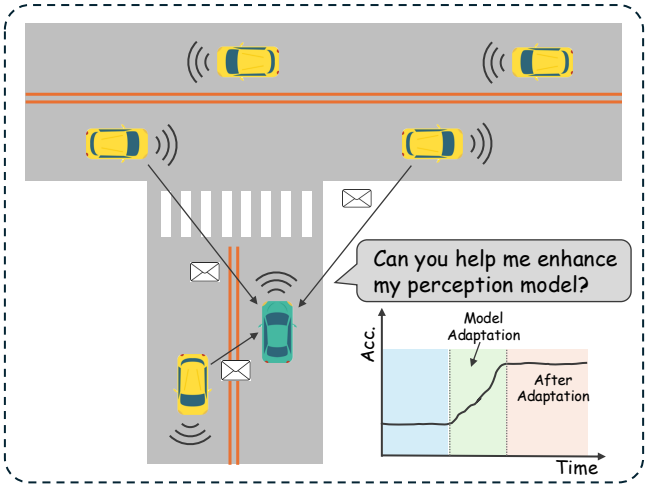

Fig. 1: Overview of leveraging CP for automated model adaptation. By leveraging features from surrounding CAVs, the ego vehicle generates high-quality pseudo-labels to improve its local perception model without manual annotation.

I. INTRODUCTION

While perception models in autonomous driving and robotics perform well in their training dataset [1], [2], they often struggle to generalize in the open real world [3]–[5]. This generalization gap necessitates continuous model adaptation to new environments. Nonetheless, traditional supervised model adaptation can be prohibitively expensive, requiring time-consuming and labor-intensive manual labeling, e.g., ground-truth bounding boxes and object classes in target scenes. Considering the case where models need to be adapted on an agent locally due to privacy and bandwidth constraints [6], [7], it is unrealistic to expect drivers or users, who are often not expert human annotators, to label data whenever they encounter new environments or changing conditions. Moreover, recent semi-automated annotation methods (e.g., leveraging powerful foundation models) may not be feasible either, as they typically require substantial computing/bandwidth resources, demand access to cloud APIs, and violate user privacy [8]–[10]. This raises a fundamental question: *Can we bypass manual labeling for perception model adaptation on autonomous agents?*

To answer this question, unsupervised model adaptation or test-time adaptation (TTA) enables pre-trained models to be adapted to entirely unlabeled data. However, existing TTA solutions face significant shortcomings when applied to autonomous driving. *First*, the quality of pseudo labels generated by TTA schemes can be unreliable as they are often highly sensitive to hyperparameter choices and exhibit significantly varying performance across different scenarios [11]–[13]. *Second*, the performance of these methods is fundamentally limited by the ego-centric views. Operating only on the ego vehicle's perspective makes it difficult to generate reliable labels for partially occluded objects or

The work was supported in part by the JC STEM Lab of Smart City funded by The Hong Kong Jockey Club Charities Trust under Contract 2023-0108, in part by the Research Grants Council of the Hong Kong SAR, China (Project No. CityU 11216324), and in part by the Hong Kong SAR Government under the Global STEM Professorship. The work of Y. Deng was supported in part by the National Natural Science Foundation of China under Grant No. 62301300 and in part by the Shandong Province Science Foundation under Grant No. ZR2023QF053. The work of X. Chen was supported in part by the Research Grants Council of Hong Kong under Grant 27213824 and CRS HKU702/24.

[1]Yanan Ma, Yihang Tao, Zhengru Fang, Zihan Fang, and Yuguang Fang are with the Hong Kong JC Lab of Smart City and the Department of Computer Science, City University of Hong Kong, Hong Kong, China. E-mail: yananma8-c@my.cityu.edu.hk, yihang.tommy@my.cityu.edu.hk, zhefang4-c@my.cityu.edu.hk, zihanfang3-c@my.cityu.edu.hk, my.fang@cityu.edu.hk.

[2]Yiqin Deng is with the School of Data Science, Lingnan University, Tuen Mun, Hong Kong, China. E-mail: yiqindeng@ln.edu.hk.

[3]Xianhao Chen is with the Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong, China. E-mail: xchen@eee.hku.hk.

resolve perceptual ambiguities. This limitation of these methods, due to *data quality*, inherently restricts their potential in pseudo-labeling.

Collaborative perception (CP) has emerged as a key technology for autonomous driving, which aggregates data from multiple vehicles to achieve perception accuracy far beyond that of a single agent [2], [6], [7], [14]–[18]. Vehicle-to-vehicle (V2V) systems for CP have been widely recognized in 5G systems and beyond (5G+) to enhance road safety and traffic efficiency [1], [19]. Crucially, we advocate that this high-fidelity output can naturally serve as "teacher" predictions to supervise the adaptation of an individual vehicle's model. For instance, as illustrated in Fig. 1, under a limited communication budget, an ego connected and autonomous vehicle (CAV) can proactively leverage CP to generate high-quality pseudo-labels for model adaptation. This pseudo-labeling approach is hyperparameter-insensitive and relatively reliable, assuming CP often outperforms a single vehicle's perception. This simple yet effective strategy not only integrates seamlessly with 5G V2V systems but also enables reliable model adaptation with relatively accurate pseudo-labels derived from the "combined wisdom" of CAV.

However, existing CP schemes never leverage these CP prediction results for model adaptation purposes. Attempting to do so presents challenges regarding communication constraints, view discrepancies between the learner and CP results, and the reliability of pseudo-labeling. In response, we propose `LDE`, Learning from Distributed "Eyes", a fully automated model adaptation approach for 3D object detection with a multi-stage pipeline. To tackle the communication bottleneck, we introduce a selective, adaptation-oriented feature sharing mechanism. Instead of naively sharing all features, this approach reformulates feature selection as a multiple-choice knapsack problem (MCKP), prioritizing features that yield the highest utility score within the learner's FoV under the transmission budget. To resolve the learner-CP view discrepancy, we develop a FoV filtering method. It first conservatively shrinks bounding boxes to account for shape and range uncertainty, and then retains only the pseudo labels that are visible from the learner's egocentric perspective. Finally, to mitigate inherent pseudo-label noise, we employ a confidence-based curriculum learning strategy, allowing the model to adapt by initially focusing on the most reliable pseudo-labels before progressively processing others.

The main contributions of this paper are summarized as follows.

- We propose a novel unsupervised model adaptation framework named `LDE`, which leverages CP as a source of "teacher" supervision to generate high-quality pseudo labels and overcome the unreliability of single-agent adaptation.
- We design an adaptation-oriented feature-sharing approach by considering practical communication constraints, utilize FoV filtering to eliminate out-of-range labels caused by view discrepancies, and employ curriculum learning to enhance adaptation reliability.
- We conduct extensive performance evaluations on both simulated and real-world datasets, i.e., V2X-Sim and DAIR-V2X datasets, to demonstrate that our `LDE` framework outperforms non-adaptive baselines and existing unsupervised benchmarks on 3D object detection tasks.

II. Related Work

A. Test-time Adaptation for Object Detection

Source-free unsupervised domain adaptation or TTA, which adapts models to unlabeled test data without requiring access to the source dataset [11], [20], is often demanded in autonomous driving and robotics, e.g., unsupervised 3D object detection tasks [11], [21]–[23]. Yoo *et al.* [21] developed an unsupervised adaptation scheme for 3D perception by learning from other predictions. However, this scheme can work only if others have better prediction quality than the learner. You *et al.* [22] exploited several repeated traversals of the same routes in the target domain to enhance unsupervised 3D object detection. Nonetheless, while a company can build a large-scale dataset through many repeated traversals, an individual vehicle (learner) may not be able to acquire such a local dataset. Xia *et al.* [23] devised DOtA, an automated pipeline for constructing object detection labels based on unlabeled CP datasets. However, this scheme exploits the shared pose and shape information of each CAV, which may not be accurately available and may not be effective for detecting objects other than CAVs, such as pedestrians and bicycles. In a nutshell, prior approaches, when applied to automated model adaptation in autonomous driving, have limited application scopes. More fundamentally, TTA methods often find it difficult to obtain reliable pseudo labels in the target domain [13]. This motivates us to investigate how to utilize CP results as pseudo-labels for unsupervised model adaptation.

B. Collaborative Perception

Benefiting from the combined information from multiple CAVs, CP is widely recognized for its superiority over single-agent perception. A significant portion of existing CP research focuses on enhancing communication efficiency while preserving prediction quality [14], [16]–[18], [24], [25]. For instance, Who2com [24] employs a multi-stage handshake mechanism to compress information via matching scores. V2VNet [25] uses graph neural networks to aggregate information from nearby CAVs, while Where2comm [16] utilizes the detection head to direct regions for sparse interactions. How2comm [26] proposes to use a mutual information-aware mechanism for feature sparsification and a flow-guided strategy to compensate for temporal asynchrony. Similarly, PACP [7] develops a BEV-match mechanism to prioritize vehicles and optimize transmissions. To address heterogeneity among agents, STAMP [27] introduces a scalable, task- and model-agnostic pipeline. It uses lightweight adapter-reverter pairs to transform BEV features between agent-specific models and a shared protocol domain, enabling efficient collaboration even when agents run different model architectures. Despite these advancements in communication

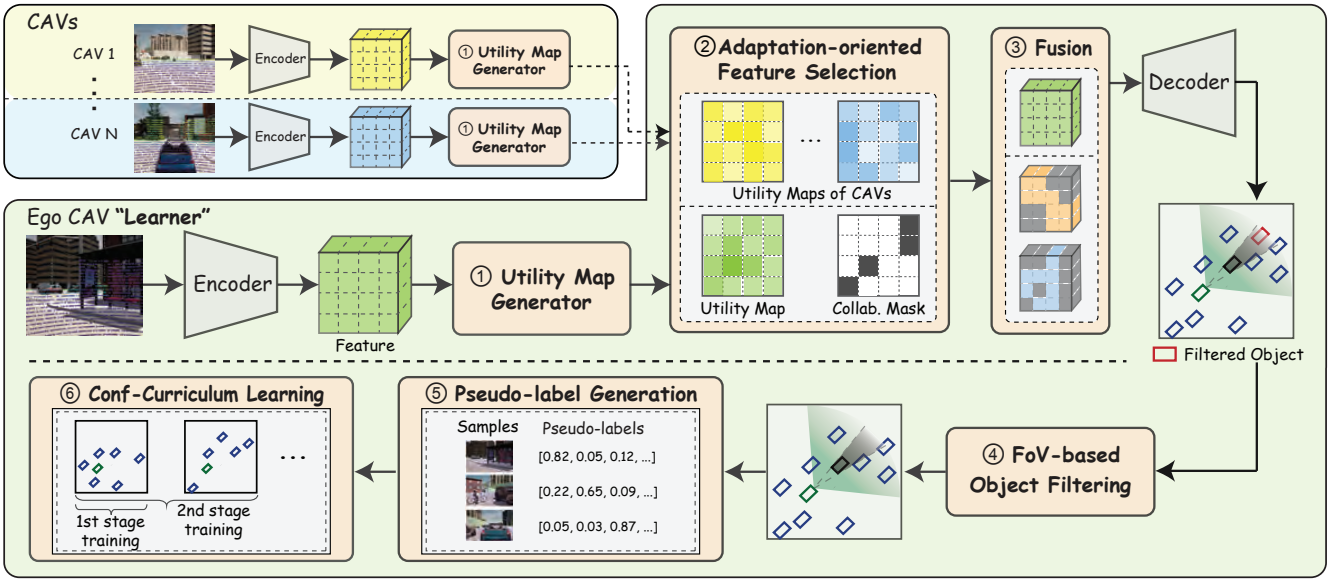

Fig. 2: Illustration of the LDE framework. The ego CAV (learner) collaborates with N surrounding CAVs to generate high-quality pseudo labels for model adaptation. The framework utilizes adaptation-oriented feature selection and fusion based on spatial utility maps, followed by FoV-based pseudo label filtering to remove out-of-range detections. Finally, confidence-based curriculum learning progressively adapts the learner's perception model to mitigate the impact of corrupted labels.

efficiency and model-agnostic fusion, none of these works employ CP for *unsupervised adaptation of a vehicle's individual model*.

When directly applied to model adaptation, the existing CP methods lead to three issues: 1) the transmitted data may not necessarily be the most valuable results for model adaptation; 2) the out-of-range detection results not in the ego view of the learner can mislead model adaptation due to the view discrepancy; 3) learning may overfit toward unreliable pseudo labels, even when generated by CP.

III. Problem Definition

Given the perception model of an ego vehicle (the *learner*) pre-trained on a source dataset, our goal is to adapt it to an unlabeled target dataset $\mathcal{D}$. Formally, the objective is to find the ego vehicle's predictor $f^\star$ that minimizes the empirical risk on samples drawn from the target dataset

$$f^\star = \arg\min_{f\in\mathcal{F}} \frac{1}{|\mathcal{D}|} \sum_{(X_{\text{ego}},y_t)\in\mathcal{D}} \ell\big(y_t, f(X_{\text{ego}})\big), \qquad (1)$$

where (X_{ego}, y_t) denotes a sample from $\mathcal{D}$, $|\mathcal{D}|$ is the total number of samples in $\mathcal{D}$, $\mathcal{F}$ is a hypothesis class of predictors, and $\ell(\cdot,\cdot)$ is the detection loss. However, in the target domain, the ground-truth labels y_t are unavailable, making it infeasible to optimize the above objective.

To address the dilemma, we consider the case where the learner collaborates with N surrounding CAVs in a vehicular network (N and the set of CAVs can vary as the learner moves). Let X_i denote the observations of the i-th CAV, which has many overlapped objects with X_{ego} in $\mathcal{D}$ from different perspectives and distances. By considering communication efficiency, each selected CAV transmits sparse intermediate feature P_i extracted from X_i to the learner. Through CP, the learner can obtain the fusion results by

$$\tilde{y} = g\left(X_{\text{ego}}, \{P_i\}_{i=1}^{N}\right), \qquad (2)$$

where $g(\cdot)$ is a function that maps the ego vehicle's data and received features to a pseudo label. Since CP results are generally more accurate than those of the learner alone, $\tilde{y}$ can be naturally used as pseudo labels to supervise the adaptation of the learner's model.

IV. Proposed LDE Framework

The proposed LDE framework, illustrated in Fig. 2, consists of five key steps: 1) Each CAV first extracts features with an encoder and then generates a *spatial utility map*. 2) The learner selectively requests relevant features from neighboring CAVs under communication constraints. 3) The feature fusion module aggregates these features. 4) A FoV filtering method removes out-of-range pseudo labels that are not visible to the learner. 5) A confidence-based curriculum learning approach is used for progressive model adaptation. In what follows, we introduce the framework step by step.

A. Feature and Utility Map Generation

Each CAV extracts feature maps from its raw sensor data (e.g., 3D point clouds) using an encoder $\Phi_{\text{Enc}}(\cdot)$, which can be represented in a bird's-eye view (BEV) and hence projected into a unified global coordinate system. For observation X_i of the i-th CAV, the extracted feature map is $\mathbf{F}_i = \Phi_{\text{Enc}}(X_i) \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote

the height, width, and number of channels, respectively. Similarly, given observation X_{ego}, the feature map of the learner is given by $\mathbf{F}_{\text{ego}} = \Phi_{\text{Enc}}(X_{\text{ego}}) \in \mathbb{R}^{H\times W\times C}$.

To select valuable features for sharing, each vehicle derives a spatial utility map for its feature map. For the i-th CAV, this is computed as $\mathbf{D}_i = \Phi_{\text{Util}}(\mathbf{F}_i) \in [0,1]^{H\times W}$, where the utility generator $\Phi_{\text{Util}}(\cdot)$ is implemented via a detection decoder. Similarly, the learner's utility map is $\mathbf{D}_{\text{ego}} = \Phi_{\text{Util}}(\mathbf{F}_{\text{ego}}) \in [0,1]^{H\times W}$.

B. Adaptation-oriented Feature Sharing and Fusion

Given utility maps, we aim to maximize collaborative gain under a constrained communication budget. To achieve this, we first model wireless channels to determine data rates and then formulate feature selection as a budget-constrained optimization problem.

Communication Modeling. We consider an orthogonal frequency division multiple access (OFDMA) transmission scheme between surrounding CAVs and the learner. According to Shannon's channel capacity, the data rate R_i between the i-th CAV and the learner is given by

$$R_i = B\log_2\left(1+\frac{p_i d_i^{-\alpha}|h_i|^2}{n_0}\right), \tag{3}$$

where B is the allocated bandwidth, p_i is the transmit power of CAV i, d_i is the distance between CAV i and the learner, α is the path loss exponent, $|h_i|^2$ is the small-scale fading gain, and n_0 is the noise power.

Feature Sharing Problem Formulation. Our objective is to maximize the total utility gain from feature sharing. Formally, the utility gap $[\mathbf{G}_i]_{j,k}$ is defined as:

$$[\mathbf{G}_i]_{j,k} = \max\{[\mathbf{D}_i]_{j,k} - [\mathbf{D}_{\text{ego}}]_{j,k}, 0\}, \tag{4}$$

where $[\mathbf{D}_i]_{j,k}$ denotes the (j,k)-th element of the utility matrix $\mathbf{D}_i$ for the i-th CAV. A large utility gap characterizes a spatial grid perceptible to neighboring CAVs but uncertain for the learner.

To determine which features to transmit, we introduce a binary selection matrix $\mathbf{S}_i \in \{0,1\}^{H\times W}$ for each CAV, where $[\mathbf{S}_i]_{j,k} = 1$ if the grid (j,k) from CAV i is selected, and 0 otherwise. The corresponding feature sharing optimization is formulated as

$$\max_{\mathbf{S}_i} \sum_{i=1}^{N}\sum_{j=1}^{H}\sum_{k=1}^{W}[\mathbf{S}_i]_{j,k} * [\mathbf{G}_i]_{j,k} \tag{5a}$$

$$\text{s.t.} \sum_{i=1}^{N}[\mathbf{S}_i]_{j,k} \le 1,\ \forall j,\forall k, \tag{5b}$$

$$\sum_{i=1}^{N}\frac{\sum_{j=1}^{H}\sum_{k=1}^{W}[\mathbf{S}_i]_{j,k} * \delta}{R_i} \le T, \tag{5c}$$

$$[\mathbf{S}_i]_{j,k} \in \{0,1\},\ \forall i,\forall j,\forall k, \tag{5d}$$

where δ denotes the data volume of a single feature grid, T is the communication latency requirement. Constraint (5b) enforces that each feature grid is selected at most once, and Constraint (5c) ensures that the total transmission latency does not exceed deadline T (which is subject to vehicle contact time and spectrum resources in a vehicular network).

Solution Approach. The feature-sharing problem we consider follows the structure of a multiple-choice knapsack problem (MCKP), which is NP-hard and highly challenging to solve [28]. We resort to a heuristic algorithm to obtain the solution efficiently. First, to reduce computational complexity and enhance robustness, we partition each high-resolution utility gap matrix $\mathbf{G}_i \in \mathbb{R}^{H\times W}$ into $C_h \times C_w$ non-overlapping cells of size $a_h \times a_w$, where $C_h = H/a_h$ and $C_w = W/a_w$, with index set of cell (p,q) being $\Omega_{p,q} = \{(j,k)|(p-1)a_h < j \le pa_h, (q-1)a_w < k \le qa_w\}$, for $p = 1,\dots,C_h$ and $q = 1,\dots,C_w$. The cell-level utility gap at (p,q) is then calculated by respecting the collaboration mask $\mathbf{S}$.

To facilitate solution finding, we define $\eta_{i,p,q} = [\mathbf{G}_i^C]_{p,q}/\Delta t_i$ as the utility-to-latency ratio, where $\Delta t_i = \delta/R_i$ is the transmission time for a single feature cell. We sort all candidate cells by $\eta_{i,p,q}$ in descending order and select them sequentially until either the latency budget T is exhausted or all candidates have been evaluated. Upon completing the selection, the learner reconstructs the binary mask $\mathbf{S}_i$ for each CAV by

$$[\mathbf{S}_i]_{j,k} = \begin{cases} 1, & (j,k) \in \Omega_{p,q} \text{ for the chosen cell } (p,q), \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The learner then transmits $\mathbf{S}_i$ to the corresponding CAV. Finally, each selected CAV packs and transmits the resulting sparse feature map $P_i = \mathbf{S}_i \odot \mathbf{F}_i$ to the learner, where $\odot$ denotes the Hadamard product.

Feature Fusion and Detection. Upon receiving the sparse features from collaborating CAVs, the learner applies a transformer-based fusion module that uses multi-head attention to aggregate spatially aligned features. For notational simplicity, we denote the learner as index $i = 0$ with $\mathbf{F}_0 = \mathbf{F}_{\text{ego}}$. The fused feature map is expressed as $\mathbf{F}_{\text{fuse}} = \text{FFN}\left(\sum_{i=0}^{N}\mathbf{W}_i \odot \mathbf{F}_i\right)$, where $\text{FFN}(\cdot)$ is a feedforward network, and $\mathbf{W}_i$ is the attention weight map. Subsequently, feeding the fused feature map $\mathbf{F}_{\text{fuse}}$ into the detection decoder yields a set of M predictions, denoted as $\{(b_i,c_i)\}_{i=1}^{M} = \Phi_{\text{Dec}}(\mathbf{F}_{\text{fuse}})$. These predictions serve as the adaptation-oriented collaborative perception results. Each prediction tuple consists of a 3D bounding box b_i and its associated confidence score $c_i \in [0,1]$. The geometric configuration of each bounding box is parameterized as $b_i = (x_i, y_i, z_i, h_i, w_i, l_i, \theta_i)$. Finally, confidence-threshold filtering is employed to discard detections whose predicted score falls below the predefined threshold τ to eliminate low-confidence outputs.

C. FoV-based Pseudo Label Filtering

Although the CP results can directly serve as pseudo-labels to supervise the learner's model adaptation, they may include objects outside the learner's effective FoV (e.g., in blind spots). Such out-of-view objects can mislead the adaptation process. To mitigate this view discrepancy, we introduce an

TABLE I: **Comparative results under different pre-trained model accuracy on V2X-Sim and DAIR-V2X datasets.** We report test accuracies AP@0.3 and AP@0.5 after 10 epochs with best in **bold**, second-best underlined, except for the fully-supervised (upper bound). ↓ represents performance degradation compared with the pretrained (non-adaptive) model. Setup 1 and Setup 2 represent adaptation cases from pre-trained models with different initial accuracies, respectively.

Method	V2X-Sim				DAIR-V2X			
	Setup 1		Setup 2		Setup 1		Setup 2	
	AP@0.3	AP@0.5	AP@0.3	AP@0.5	AP@0.3	AP@0.5	AP@0.3	AP@0.5
Pretrained	55.69	49.06	68.79	60.31	46.01	41.22	55.61	50.37
AdaBN	48.12↓	41.88↓	63.22↓	54.73↓	38.12↓	33.29↓	48.64↓	44.08↓
ST	51.66↓	45.79↓	67.71↓	58.43↓	40.81↓	35.26↓	50.12↓	46.77↓
SN	56.42	51.00	70.51	61.53	42.33↓	37.29↓	57.38	52.14
CPD	57.71	52.84	71.85	63.89	48.79	44.10	58.02	54.71
DOtA	59.17	55.37	74.18	68.01	50.82	45.79	62.67	58.84
CP4Adaptation	44.38↓	37.73↓	59.36↓	51.19↓	39.78↓	36.70↓	48.81↓	44.33↓
LDE (Ours)	63.33	55.24	77.49	71.27	55.60	49.11	67.28	63.41
LDE-Full (Ours)	**69.05**	**61.72**	**83.93**	**75.73**	**59.01**	**52.25**	**69.00**	**64.98**
Upper Bound	86.17	81.79	86.93	82.34	70.23	65.76	71.21	66.29

FoV-based label filtering mechanism that retains only those detections truly visible to the learner.

To enforce visibility consistency, we conduct a line-of-sight (LoS) analysis for LiDAR [29] on each candidate detection. This efficiently determines whether a detection lies along an unobstructed ray originating from the learner. Let d_i denote the distance from the learner's sensor to the centroid of bounding box b_i. We first sort the set of bounding boxes $\{b_i\}$ in ascending order of d_i to prioritize closer objects. Next, because cuboidal bounding boxes often overshoot the boundaries of real-world objects with curved surfaces, we conservatively shrink each box b_i by a scaling factor $\epsilon_i \in (0, 1]$. Specifically, we scale the spatial dimensions of the original bounding box to generate a shrunken counterpart $\hat{b}_i = (x_i, y_i, z_i, \epsilon_i h_i, \epsilon_i w_i, \epsilon_i l_i, \theta_i)$. To account for growing localization uncertainty at longer ranges, we define this distance-based ratio as

$$\epsilon_i = \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min})\, e^{-\lambda\, d_i}. \tag{7}$$

Consequently, more distant boxes (i.e., larger d_i) are shrunk more aggressively ($\epsilon_i \to \epsilon_{\min}$), thereby reducing false positives caused by noisy long-range detections. Ultimately, this conservative shrinkage mitigates potential occlusion errors arising from bounding-box inaccuracies and shape mismatches.

Next, we check *visibility* by ray-casting from the learner's sensor origin toward the centroid and vertices of $\hat{b}_i$. As long as at least one ray reaches $\hat{b}_i$ unobstructed by closer objects, this bounding box is deemed (partially) visible to the learner. Finally, we verify *non-occlusion* by ensuring that $\hat{b}_i$ does not overlap with any previously accepted box $\hat{b}_j$ with $d_j < d_i$ in the BEV projection, which otherwise may indicate misdetections. By retaining only those b_i that satisfy both visibility and non-occlusion, we obtain a filtered pseudo-label set aligned with the learner's ego-view, thereby facilitating reliable model adaptation.

D. Confidence-based Curriculum Learning

Upon model adaptation, we introduce a curriculum learning strategy to adapt the model in an "easy-to-hard" manner [30], [31]. Rather than relying on a single, fixed confidence threshold, we employ a dynamic threshold τ_c. Specifically, we initialize the adaptation process with a high confidence threshold to construct a high-quality subset of pseudo-labels, comprising only the most certain detections. Subsequently, we gradually decay this threshold to introduce more lower-confidence samples. This enables the model to smoothly propagate knowledge learned from the highly reliable initial set to broader, more challenging data distributions. Ultimately, this curriculum-based approach prevents early-stage model corruption and ensures a more stable and robust adaptation process.

V. Experiments

A. Experimental Setup

Datasets. We conduct experiments on both simulated and real-world datasets, i.e., V2X-Sim dataset [32] and DAIR-V2X dataset [33].

- **V2X-Sim dataset.** V2X-Sim [32] is a simulated V2X collaborative perception dataset simulated using SUMO and CARLA [3]. It comprises 10,000 frames of 3D LiDAR point clouds captured from 5 CAVs, alongside 501,000 annotated 3D bounding boxes. We discretize the 3D point clouds into a BEV map with a size of $(256, 256, 13)$, and the resolution is 0.4 m/pixel in both length and width.
- **DAIR-V2X dataset.** DAIR-V2X [33] is a real-world collaborative perception dataset wherein each sample captures synchronized data from a vehicle and an infrastructure node. The effective perception range spans 201.6 m $\times$ 80 m. We represent the BEV map with size of $(200, 504, 64)$ and the resolution is 0.4 m/pixel.

Implementation Details. For a LiDAR-based 3D object detection task, we conduct experiments with PointPillars [34]

TABLE II: **Comparison of label quality on V2X-Sim and DAIR-V2X datasets.** We report the Recall and Precision of IoU@0.3 and IoU@0.5, with the best results shown in **bold**, second-best underlined.

Method	V2X-Sim				DAIR-V2X			
	Recall		Precision		Recall		Precision	
	IoU@0.3	IoU@0.5	IoU@0.3	IoU@0.5	IoU@0.3	IoU@0.5	IoU@0.3	IoU@0.5
Pretrained	69.09	66.37	59.57	53.78	65.30	60.29	47.81	44.73
AdaBN	59.28↓	52.82↓	43.79↓	38.53↓	58.62↓	52.50↓	39.56↓	36.18↓
ST	67.39↓	63.02↓	42.39↓	37.95↓	63.42↓	57.10↓	36.56↓	34.86↓
SN	74.88	64.69↓	52.62↓	47.21↓	68.65	60.45	43.85↓	41.05↓
CPD	77.03	68.12	57.89↓	55.71	69.90	61.17	49.16	45.81
DOtA	78.51	69.62	62.43	58.78	70.80	61.97	50.56	47.71
CP4Adaptation	77.81	68.67	40.41↓	36.09↓	69.53	60.22	37.15↓	34.82↓
LDE (Ours)	79.20	71.79	64.72	60.78	71.20	63.01	53.48	50.65
LDE-Full (Ours)	**87.40**	**85.01**	**79.30**	**76.23**	**75.90**	**69.63**	**61.88**	**57.31**

as the default detector. The bandwidth is $B = 20$ MHz and the transmit power is 5 W. The path loss exponent α is set to 2.3, with a noise power of $n_0 = -120$ dBm. The time requirement is $T = 300$ ms. The data is split into model adaptation, validation, and test sets, with a ratio of 8:1:1. First, we pre-train the detector on the OPV2V dataset [15] to a predefined accuracy. Then, we perform inference on the adaptation dataset to obtain CP results by sharing the selected features with the learner under the communication constraint, which serve as pseudo-labels. Label filtering is conducted with $\epsilon_{\max} = 0.95$, $\epsilon_{\min} = 0.8$, and $\lambda = 1$. For our confidence-based curriculum learning, we adapt the model over 10 epochs. We start with a high confidence threshold of $\tau_c = 0.75$ for the first 3 epochs, then relax it to $\tau_c = 0.5$ for epochs 4-7, and finally use $\tau_c = 0.25$ for the remaining epochs. We adapt the model using Adam optimizer with lr=$1e^{-5}$. All experiments are conducted on a server with 2 Intel(R) Xeon(R) Silver 4410Y CPUs, 4 NVIDIA RTX A5000 GPUs, and 512 GB RAM. The performance of all methods is evaluated using average precision (AP) at 0.3 IoU (AP@0.3) and 0.5 IoU (AP@0.5) for detection accuracy.

B. Quantitative Evaluation

Baselines. We evaluate our `LDE` framework against other model adaptation baselines: **AdaBN** [35], **ST** (self-training) [36], **SN** (statistical normalization) [4], **CPD** [37], **DOtA** [23], and a fully supervised **Upper Bound**. For a fair comparison, all methods use PointPillars as the backbone. We also include two additional baselines: **CP4Adaptation** employs the CP results directly from the Where2comm framework [16] as pseudo labels, and **LDE-Full**, which is an implementation of **LDE** with an unlimited communication budget.

Baseline Comparison. The quantitative evaluations are detailed in Table I, which reports the adaptation performance across two distinct initial model accuracies (Setup 1 and Setup 2). We observe that existing TTA baselines exhibit a critical weakness: they exploit knowledge only from the ego-vehicle's captured data. Consequently, methods such as AdaBN, ST, and SN fail to adapt effectively to the open-world setting, often resulting in significant performance degradation (↓) compared to the non-adaptive pretrained model. While more advanced methods, such as CPD and DOtA, achieve modest performance gains, our proposed `LDE` framework consistently and substantially outperforms them across all setups on both the V2X-Sim and DAIR-V2X datasets. Specifically, on the real-world DAIR-V2X dataset under Setup 1, `LDE` achieves an AP@0.3 of 55.60% and an AP@0.5 of 49.11%, demonstrating robust real-world generalization. Furthermore, compared with CP4Adaptation, our approach yields striking absolute performance gains—such as an 18.95% increase in AP@0.3 and a 17.51% increase in AP@0.5 under Setup 1 on the V2X-Sim dataset. This demonstrates that naively implementing existing collaborative perception schemes for model adaptation is highly ineffective. Finally, our unconstrained `LDE-Full` variant consistently achieves the highest unsupervised accuracy, significantly narrowing the performance gap to the fully supervised Upper Bound.

Label Quality Analysis. To evaluate the quality of the generated pseudo-labels, we analyze the recall and precision across various approaches on both the V2X-Sim and DAIR-V2X datasets. As detailed in Table II, `LDE` consistently yields superior pseudo-labels compared to existing unsupervised methods. Notably, `LDE` maintains robust performance and achieves a high recall of 79.20% and a precision of 64.72% on the V2X-Sim dataset at IoU@0.3. This substantial enhancement is primarily attributable to the multi-vehicle collaborative architecture of `LDE`: by aggregating spatially diverse features, the framework captures richer, multi-view object representations, thereby mitigating occlusions and reducing false negatives.

C. Ablation Study

Effect of Key Modules. We first analyze the impact of our three core components, as shown in Table III. Replacing our adaptation-oriented feature sharing (**Adapt FS**) with a random selection strategy degrades performance, as this wastes bandwidth on irrelevant data and lowers pseudo-label quality. Similarly, removing the FoV-based pseudo-label fil-

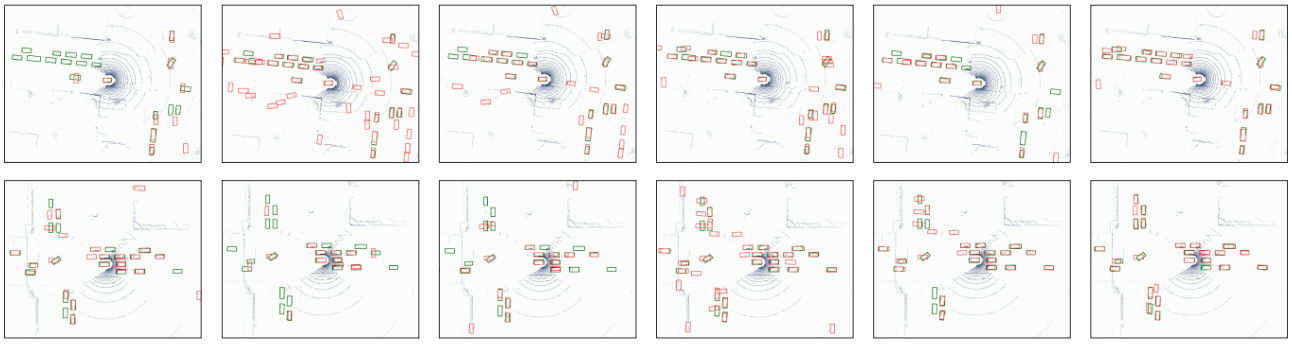

Fig. 3: Visualization of 3D detection results on the V2X-Sim dataset, comparing the performance of the adapted model with that of the original model and some baselines. Red bounding boxes denote the model predictions, while green ones represent the ground truth (GT).

TABLE III: Ablation study on key components of LDE on V2X-Sim dataset. Best in **bold**, second-best underlined.

Adapt FS	FoV-Filter	CCL	V2X-Sim	
			AP@0.3	AP@0.5
			68.79	60.31
✓	✓	×	75.92	69.75
✓	×	✓	63.86↓	55.45↓
×	✓	✓	74.37	67.78
✓	✓	✓	**77.49**	**71.27**

TABLE IV: Effects of the different communication constraints on V2X-Sim and DAIR-V2X datasets. Best in **bold**, second-best underlined.

Comm. Budget	V2X-Sim		DAIR-V2X	
	AP@0.3	AP@0.5	AP@0.3	AP@0.5
Pretrained	68.79	60.31	55.61	50.37
T = 100 ms	73.01	64.44	60.41	55.27
T = 300 ms	77.49	71.27	67.28	63.41
T = 500 ms	**78.92**	**72.87**	**68.92**	**64.33**

tering (**FoV-Filter**) is highly detrimental, as it introduces a severe perceptual mismatch by forcing the model to learn from misleading labels outside its FoV. Finally, removing the confidence-based curriculum learning (**CCL**) in favor of a single-shot adaptation also harms performance, confirming our curriculum learning is essential for stable adaptation by allowing the model to learn from the most reliable labels first and preventing early-stage error accumulation.

Effect of Communication Budget. We further investigate the impact of the communication time budget on the LDE framework, with results summarized in Table IV. As time allocation increases, the learner is able to aggregate richer collaborative information and improve the quality of generated pseudo-labels, thereby leading to superior model adaptation performance.

TABLE V: Effects of the different parameters for label filtering on the V2X-Sim dataset. Best in **bold**, second-best underlined.

Conf.	V2X-Sim			
	$\epsilon_{\max} = 0.95, \epsilon_{\min} = 0.8$		$\epsilon_{\max} = 1, \epsilon_{\min} = 1$	
	AP@0.3	AP@0.5	AP@0.3	AP@0.5
Pretrained	68.79	60.31	68.11	59.86
τ = 0.15	76.23	70.04	75.45	69.75
τ = 0.25	**77.49**	**71.27**	**77.00**	**70.78**
τ = 0.35	77.02	70.63	76.34	70.27

Effect of Parameters for Label Filtering. Table V presents ablation results for label-filtering parameters on model adaptation. Adjusting the bounding-box shrinkage factor has little effect on overall detection accuracy. By contrast, the confidence threshold exhibits a clear trade-off: setting it too low admits noisy pseudo labels and degrades adaptation, while setting it too high excludes informative labels and likewise reduces performance.

D. Qualitative Analysis

Visualization of Detection Results. To qualitatively demonstrate the effectiveness of our approach, Fig. 3 visualizes the 3D detection results of models adapted by our LDE framework alongside several competitive benchmarks. As observed, existing baselines frequently struggle in complex driving scenes, producing numerous false positives and suffering from poor bounding box localization. In contrast, our LDE framework successfully leverages distributed spatial features to significantly suppress erroneous detections, delivering more accurate and robust detection results.

VI. Conclusion

In this paper, we have addressed the challenge of adapting perception models to new environments without the need for manual labeling for autonomous driving. We have introduced the Learning from Distributed "Eyes" (LDE) framework,

which leverages collaborative perception to generate high-quality pseudo labels, enabling automated model adaptation. First, we have designed an adaptation-oriented feature-sharing mechanism to produce accurate predictions under communication constraints. Then, we have developed a field-of-view filtering scheme to mitigate view discrepancies among vehicles. Finally, we have introduced a confidence-based curriculum learning strategy to stabilize the adaptation process by managing inherent label noise. Extensive simulations have demonstrated that the proposed `LDE` framework delivers robust, consistent performance improvements over pre-trained models and existing unsupervised adaptation methods, paving the way for reliable perception in real-world autonomous driving scenarios.

While we choose 3D object detection in autonomous driving as the subject of study, our approach has the potential to benefit a wide range of vision tasks, such as depth estimation and segmentation, as well as other applications, such as drone and robotic systems.

REFERENCES

[1] X. Chen, Y. Deng, H. Ding, G. Qu, H. Zhang, P. Li, and Y. Fang, "Vehicle as a service (VaaS): Leverage vehicles to build service networks and capabilities for smart cities," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 3, pp. 2048–2081, 2024.

[2] S. Hu, Z. Fang, Y. Deng, X. Chen, and Y. Fang, "Collaborative perception for connected and autonomous driving: Challenges, possible solutions and opportunities," *IEEE Wireless Communications*, 2025.

[3] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on Robot Learning*. PMLR, 2017, pp. 1–16.

[4] Y. Wang, X. Chen, Y. You, L. E. Li, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. Chao, "Train in germany, test in the usa: Making 3d object detectors generalize," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 713–11 723.

[5] Y. Ma, S. Hu, Z. Fang, Y. Ji, Y. Deng, and Y. Fang, "Sense4FL: Vehicular crowdsensing enhanced federated learning for object detection in autonomous driving," *IEEE Transactions on Mobile Computing*, vol. 25, no. 8, pp. 13 004–13 018, 2026.

[6] Z. Fang, J. Wang, Y. Ma, Y. Tao, Y. Deng, X. Chen, and Y. Fang, "R-ACP: Real-time adaptive collaborative perception leveraging robust task-oriented communications," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2025.

[7] Z. Fang, S. Hu, H. An, Y. Zhang, J. Wang, H. Cao, X. Chen, and Y. Fang, "PACP: Priority-aware collaborative perception for connected and autonomous vehicles," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 15 003–15 018, 2024.

[8] K. G. Ince, A. Koksal, A. Fazla, and A. A. Alatan, "Semi-automatic annotation for visual object tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1233–1239.

[9] D. Bai, S. Wei, X. He, and Q. Yu, "Annotation methods for object detection: A comparative analysis from manual labeling to automated annotation technologies," in *2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)*. IEEE, 2025, pp. 1473–1479.

[10] M. A. Reza, E. Manley, S. Chen, S. Chaudhary, and J. Elafros, "Segbuilder: A semi-automatic annotation tool for segmentation," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 8494–8503.

[11] X. Ruan and W. Tang, "Fully test-time adaptation for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1038–1047.

[12] S. Cao, J. Zheng, Y. Liu, B. Zhao, Z. Yuan, W. Li, R. Dong, and H. Fu, "Exploring test-time adaptation for object detection in continually changing environments," *arXiv preprint arXiv:2406.16439*, 2024.

[13] M. Boudiaf, R. Mueller, I. Ben Ayed, and L. Bertinetto, "Parameter-free online test-time adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8344–8353.

[14] Y. Tao, S. Hu, Z. Fang, and Y. Fang, "Directed-CP: Directed collaborative perception for connected and autonomous vehicles via proactive attention," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 7004–7010.

[15] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2583–2589.

[16] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4874–4886, 2022.

[17] Y. Ma, Z. Fang, Y. Tao, Y. Guo, Y. Deng, X. Chen, and Y. Fang, "Birdcast: Interest-aware BEV multicasting for infrastructure-assisted collaborative perception," *arXiv preprint arXiv:2604.00701*, 2026.

[18] Y. Ma, Z. Zhao, Z. Fang, H. An, X. Chen, and Y. Fang, "Update the unseen only: Minimizing AoI for collaborative perception through online learning," *arXiv preprint arXiv:2607.20967*, 2026.

[19] S. Hu, Z. Fang, H. An, G. Xu, Y. Zhou, X. Chen, and Y. Fang, "Adaptive communications in collaborative perception with domain alignment for autonomous driving," in *GLOBECOM 2024-2024 IEEE Global Communications Conference*. IEEE, 2024, pp. 746–751.

[20] M. Zhang, S. Levine, and C. Finn, "MEMO: Test time robustness via adaptation and augmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 629–38 642, 2022.

[21] J. Yoo, Z. Feng, T.-Y. Pan, Y. Sun, C. P. Phoo, X. Chen, M. Campbell, K. Weinberger, B. Hariharan, and W.-L. Chao, "Learning 3D perception from others' predictions," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 82 610–82 630.

[22] Y. You, C. P. Phoo, K. Luo, T. Zhang, W.-L. Chao, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Unsupervised adaptation from repeated traversals for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 716–27 729, 2022.

[23] Q. Xia, W. Lin, H. Xiang, X. Huang, S. Chen, Z. Dong, C. Wang, and C. Wen, "Learning to detect objects from multi-agent lidar scans without manual labels," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1418–1428.

[24] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6876–6883.

[25] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2VNet: Vehicle-to-vehicle communication for joint perception and prediction," in *European Conference on Computer Vision*. Springer, 2020, pp. 605–621.

[26] D. Yang, K. Yang, Y. Wang, J. Liu, Z. Xu, R. Yin, P. Zhai, and L. Zhang, "How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 36, pp. 25 151–25 164, 2023.

[27] X. Gao, R. Xu, J. Li, Z. Wang, Z. Fan, and Z. Tu, "STAMP: Scalable task-and model-agnostic collaborative perception," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 54 656–54 676.

[28] H. Kellerer, U. Pferschy, and D. Pisinger, "Introduction to NP-completeness of knapsack problems," in *Knapsack problems*. Springer, 2004, pp. 483–493.

[29] S. Hagstrom and D. Messinger, "Line-of-sight analysis using voxelized discrete lidar," in *Laser Radar Technology and Applications XVI*, vol. 8037. SPIE, 2011, pp. 104–114.

[30] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 41–48.

[31] Z. Zhu, Q. Meng, X. Wang, K. Wang, L. Yan, and J. Yang, "Curricular object manipulation in lidar-based object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1125–1135.

[32] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 914–10 921, 2022.

[33] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan *et al.*, "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 361–21 370.

[34] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705.

[35] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou, "Revisiting batch normalization for practical domain adaptation," *arXiv preprint arXiv:1603.04779*, 2016.

[36] A. RoyChowdhury, P. Chakrabarty, A. Singh, S. Jin, H. Jiang, L. Cao, and E. Learned-Miller, "Automatic adaptation of object detectors to new domains using self-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 780–790.

[37] H. Wu, S. Zhao, X. Huang, C. Wen, X. Li, and C. Wang, "Commonsense prototype for outdoor unsupervised 3D object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 968–14 977.