International Journal of Applied Earth Observation and Geoinformation 155 (2026) 105678

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Hierarchical active learning for annotation-efficient 3D object detection in construction-scene LiDAR sensing

You Li [a,1], Jialu Sui [c,1], Yasong Shi [d], Xiaokang Zhang [e,a,*], Shengjun Tang [b,f,g,**], Ruisheng Wang [b,f,g], Renzhong Guo [b,f,g]

[a] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, 518060, China
[b] Research Institute for Smart Cities, School of Architecture and Urban Planning, Shenzhen University, Shenzhen, 518060, China
[c] Department of Geography, The University of Hong Kong, Hong Kong, China
[d] School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, 430081, China
[e] School of Artificial Intelligence, Wuhan University, Wuhan, 430072, China
[f] State Key Laboratory of Subtropical Building and Urban Science, Shenzhen, China
[g] Guangdong-Hong Kong-Macau Joint Laboratory for Smart Cities, Shenzhen, China

ARTICLE INFO

Dataset link: https://github.com/zxk688/ConVeh3D

Keywords:
LiDAR sensing
3D object detection
Active learning
Annotation-efficient learning
Construction-scene sensing

ABSTRACT

3D object detection in construction-scene LiDAR sensing remains challenging because sparse returns, irregular layouts, heavy occlusion, and diverse machinery make object representation and localization difficult. Dense manual annotation in such scenes is also inefficient and costly, creating a strong need for annotation-efficient learning. Existing active-learning methods often struggle in this setting because they tend to favor uncertain but redundant samples while overlooking representative scene coverage. To address this problem, we propose Hierarchical Clustering-enhanced Active Learning (HCAL), an active-learning framework for LiDAR-based 3D object detection in construction and roadwork environments. HCAL improves sample selection under limited labeling budgets by identifying informative candidate keyframes, suppressing redundant uncertain samples, and preserving representative coverage through entropy filtering, gradient-space clustering, and density-consistency selection. To support field-relevant evaluation, we further establish ConVeh3D as a construction-scene LiDAR benchmark for annotation-efficient 3D detection and unknown-category-aware evaluation in a controlled category-withheld setting. Experiments on ConVeh3D, with KITTI as secondary supporting evidence, show that HCAL achieves the best average detection accuracy across most tested labeling budgets. The results indicate that HCAL provides an effective framework for improving sample efficiency in construction-scene LiDAR-based 3D object detection under limited labeling budgets. The dataset and code will be made available at https://github.com/zxk688/ConVeh3D.

1. Introduction

Complex geospatial scenes like construction and roadwork environments pose a demanding problem for LiDAR sensing because they combine open work areas, irregular object placement, heavy clutter, and sparse long-range observations (Chen et al., 2017). In these scenes, 3D object detection is needed for scene perception, equipment localization, and safety management (Gong et al., 2020; Nath and Behzadan, 2020). However, reliable detection remains difficult when machinery and vehicles must be localized under sparse returns, heavy occlusion, and heterogeneous scene structure (Shi et al., 2020; Ding et al., 2024). This is a field-relevant sensing problem because the acquisition conditions directly affect target visibility, point density, and the cost of producing usable labels (Zhou and Tuzel, 2018; Zhang et al., 2022).

Dense point cloud annotation is particularly inefficient and expensive in these environments. (Lei et al., 2022; Zeng et al., 2022) LiDAR-based detectors still depend on large volumes of annotated data (Song et al., 2023c; C. Yu et al., 2022), but in construction and roadwork scenes annotators must localize objects in dynamic layouts with frequent occlusion, substantial appearance variation, and uneven point density (Shi et al., 2019). Annotation-efficient 3D object detection

* Corresponding author.
** Corresponding author at: Research Institute for Smart Cities, School of Architecture and Urban Planning, Shenzhen University, Shenzhen, 518060, China.
E-mail addresses: liyou@gml.ac.cn (Y. Li), jialusui99@connect.hku.hk (J. Sui), yasongs333@gmail.com (Y. Shi), zhangxiaokang@whu.edu.cn (X. Zhang), shengjuntang@szu.edu.cn (S. Tang), ruiswang@szu.edu.cn (R. Wang), guorz@szu.edu.cn (R. Guo).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jag.2026.105678
Received 1 June 2026; Received in revised form 22 July 2026; Accepted 6 November 2026
Available online 16 November 2026
1569-8432/© 2026 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

(Yang et al., 2022, 2025) is therefore both a practical requirement and a meaningful research problem in construction-scene LiDAR sensing.

Active learning offers a natural way to reduce this burden, but existing point cloud methods remain insufficient for the target setting (Shannon, 1948; Roth and Small, 2006; Sener and Savarese, 2017). Uncertainty-driven selection can over-select ambiguous but redundant frames, while diversity-based selection can be expensive in high-dimensional point cloud space and may still fail to preserve representative coverage across heterogeneous scene densities. The central methodological challenge is therefore to select samples that remain informative, nonredundant, and representative in the category-withheld evaluation setting.

Evaluation is constrained by benchmark mismatch as well. Public datasets such as KITTI (Geiger et al., 2012), Waymo (Sun et al., 2020), and SceneNN (Hua et al., 2016) are important for 3D scene understanding, but they do not adequately capture the heavy machinery, work-zone clutter, and irregular target distributions that characterize construction-oriented LiDAR sensing. A construction-scene LiDAR benchmark is therefore needed as the field-relevant basis for evaluating annotation-efficient detection under the intended sensing conditions.

To address these issues, we propose Hierarchical Clustering-enhanced Active Learning (HCAL), an active-learning framework for LiDAR-based 3D object detection in construction and roadwork environments. HCAL improves sample selection under limited labeling budgets through three coordinated stages: entropy filtering identifies informative candidate keyframes, gradient-space clustering suppresses redundant uncertain samples, and density-consistency selection preserves representative coverage across heterogeneous scene densities. ConVeh3D provides the primary construction-scene LiDAR benchmark for annotation-efficient 3D detection and unknown-category-aware evaluation under a controlled category-withheld setting, while KITTI serves as secondary supporting evidence.

The main contributions of this study are summarized as follows:

- We propose HCAL, a three-stage active-learning framework for LiDAR-based 3D object detection that improves annotation efficiency by selecting informative and representative samples through entropy filtering, gradient-space clustering, and density-consistency selection.
- We establish ConVeh3D as the construction-scene LiDAR benchmark that provides the field-relevant evaluation basis for annotation-efficient 3D detection in complex construction and roadwork environments.
- We evaluate HCAL primarily on ConVeh3D, with KITTI as secondary supporting evidence, and show that it achieves the best average detection accuracy across most tested labeling budgets in the category-withheld evaluation setting.

The rest of this article is organized as follows. In Section 2, we review related work on 3D object detection and active learning. Section 3 presents the ConVeh3D benchmark. Section 4 details the proposed HCAL strategy. Section 5 reports the experimental evaluation. Finally, Section 6 concludes the article.

2. Related work

2.1. Supervised 3D object detection

Supervised 3D object detection (Song et al., 2024, 2026; Chen et al., 2022) for point clouds has advanced rapidly, with methods tailored to the sparsity and irregular sampling of LiDAR data. PointNet (Qi et al., 2017) pioneered direct feature extraction from raw point clouds. PV-RCNN++ (Shi et al., 2023) introduced sectorized proposal-centric keypoint sampling together with feature pooling and aggregation to improve local representation. Video-based methods such as DetZero (Ma et al., 2023) and related temporal detectors (Yin et al., 2021) further exploit sequential LiDAR observations to refine 3D detection. Multi-modal methods, such as GraphAlign (Song et al., 2023b) and GraphAlign++ (Song et al., 2023a), further improve 3D object detection by enhancing feature alignment between LiDAR point clouds and RGB images through graph matching. Despite these advances, supervised 3D object detection models still rely heavily on large volumes of high-quality annotated data to achieve strong performance. However, the acquisition and annotation of point cloud data are highly resource-intensive, which limits the scalability and widespread deployment of such models in real-world scenarios (Chen et al., 2024).

2.2. Active learning for 3D object detection

Active learning is a key technique for reducing labeling demand while maintaining detection performance. It has shown promise in 3D point cloud object detection, particularly for mitigating high annotation costs and improving the handling of unknown categories in controlled evaluation settings (Greer et al., 2024). The central idea is to select the most informative samples for annotation, and existing methods are typically categorized into uncertainty-based, margin-based, diversity-based, hybrid, and consistency-based approaches. Uncertainty-based methods (Shannon, 1948; Roy et al., 2018) select samples with low model confidence, often using metrics such as entropy or uncertainty scores. Entropy sampling (Wu et al., 2022; Safaei et al., 2024) ranks samples based on the uncertainty of predicted distributions, while margin sampling (Roth and Small, 2006; Ducoffe and Precioso, 2018) focuses on samples with the smallest probability margins. Diversity-based approaches (Lin et al., 2024), such as core-set selection (Sener and Savarese, 2017; Lin et al., 2017; Zhang et al., 2024), maximize coverage across feature space to ensure representativeness. Hybrid strategies (Elhamifar et al., 2013; Wu et al., 2022) combine uncertainty and representativeness. Consistency-based methods (W. Yu et al., 2022) exploit agreement between raw and augmented data, while CRB (Chen et al., 2023) applies hierarchical heuristics to select concise, representative, and geometrically balanced unlabeled point clouds. Despite these advances, current active-learning methods face notable limitations. Single-strategy biases remain: uncertainty-based methods often overemphasize ambiguous samples at the expense of representativeness, resulting in redundant selections that offer limited gains under constrained labeling budgets. Additionally, consistency- and heuristic-based methods suffer from high computational costs, limiting scalability for larger point cloud datasets.

2.3. Public point cloud object detection datasets for construction scenes

In remote sensing and geospatial scene analysis, annotated 3D point clouds remain scarce for construction-oriented environments. Current 3D point cloud object detection models are typically developed and evaluated using public benchmarks that focus on autonomous driving (Geiger et al., 2012; Sun et al., 2020; Caesar et al., 2020; Mao et al., 2021), indoor environments (Dai et al., 2017; Song et al., 2015; Hua et al., 2016), and urban scenes (Behley et al., 2019; Ying et al., 2020; Roynard et al., 2018). However, these datasets inadequately represent construction sites and roadworks. While autonomous-driving datasets such as nuScenes (Caesar et al., 2020) and ONCE (Mao et al., 2021) include some vehicle categories such as trucks and buses, their overall object diversity is still insufficient to capture the complexity of construction environments. Autonomous-driving datasets are primarily collected on open urban roads, whereas construction sites are open and unstructured spaces where construction vehicles and machinery are irregularly distributed without a fixed spatial arrangement. These domain differences significantly limit the effectiveness of models trained on driving benchmarks when applied to real construction scenarios. As shown in Table 1, previous efforts have attempted to bridge this gap by using synthetic 3D models (Ding and Luo, 2024; Yan et al., 2025) to generate construction-scene data. However, such virtual datasets

Table 1
Comparison of 3D datasets for construction sites.

Name	Data type	No. of scenes	No. of samples	No. of categories
HFK	Multi-frame point clouds	1	5	17
VCVW-3D	Synthetic images	15	375,000	10
ConVeh3D	Single-frame point clouds	5	4195	4

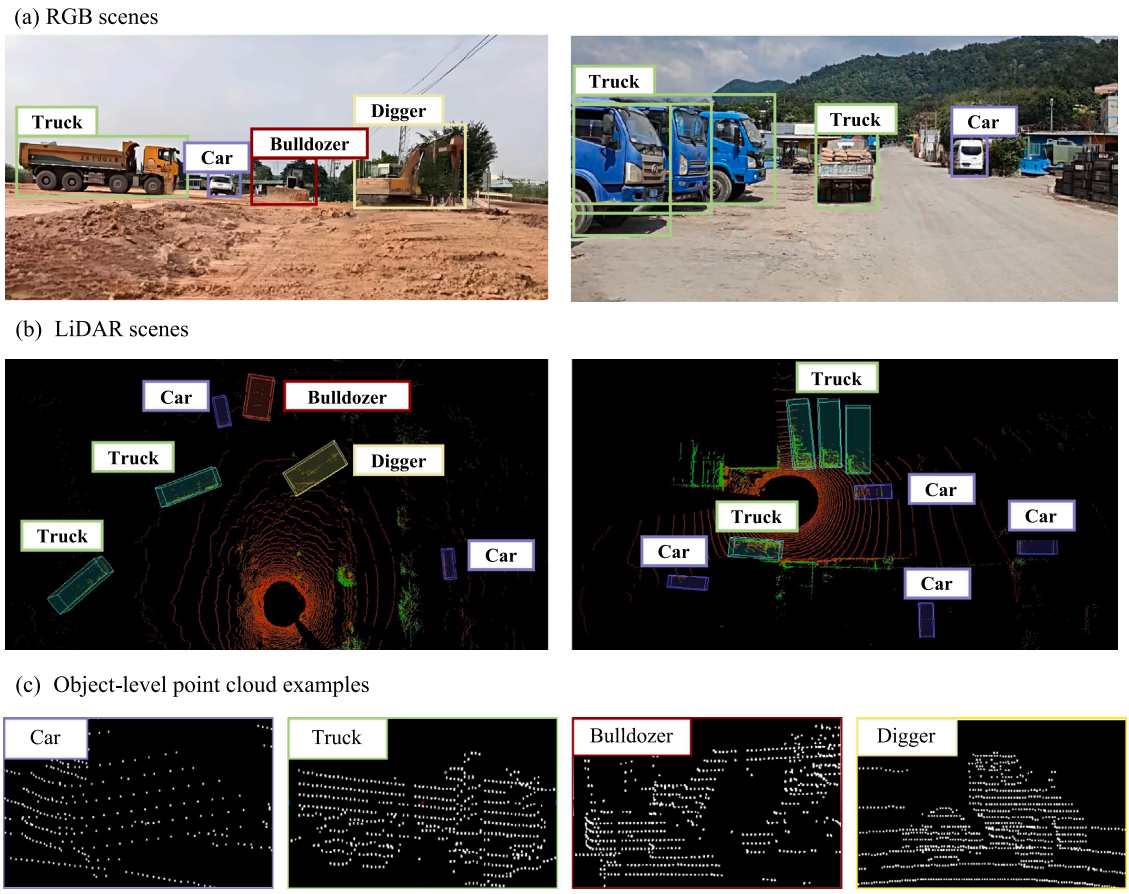


Fig. 1. Overview of the ConVeh3D dataset. (a) Representative RGB images of a construction-site scene. (b) Corresponding annotated LiDAR point cloud of the entire scenes. (c) Object-level point-cloud examples of the four annotated categories: Car, Truck, Bulldozer, and Digger.

suffer from domain and feature gaps caused by limited realism and diversity in construction-specific training data, which restricts their applicability in real-world construction environments. Huang et al. (2022) collected global 3D data for construction scenes using unmanned aerial vehicles, but this acquisition approach is costly and typically covers only limited areas. Other datasets (Xuehui et al., 2021) provide only construction-site images, requiring depth-prediction methods (Shen et al., 2024) to obtain 3D data, which often lacks sufficient accuracy. To overcome these limitations, we introduce ConVeh3D as a point cloud benchmark designed for construction scenes. ConVeh3D supports 3D object detection and scene understanding in complex construction environments.

3. ConVeh3D benchmark

ConVeh3D is the construction-scene LiDAR benchmark used in this paper to evaluate 3D object detection under sensing conditions that are weakly represented in conventional driving-oriented datasets. As illustrated in Fig. 1, the benchmark covers five construction and roadwork sites observed under daytime operating conditions, with data acquisition centered on the active work area so that machinery interactions, background clutter, and irregular scene layout are captured directly. Across the full corpus, ConVeh3D contains 4195 LiDAR frames and approximately 25 billion points. For object-detection benchmarking, 29,030 3D bounding boxes are annotated on benchmark frames across four categories: Car, Truck, Digger, and Bulldozer. Synchronized optical images are retained to support human interpretation and annotation quality control, while additional sequential LiDAR streams are kept without 3D boxes to support annotation-efficient learning on the same sensing benchmark.

3.1. Platform

A self-built sensor integration platform is used for synchronized field acquisition of LiDAR and optical data. As shown in Fig. 2(a) and (b), the platform integrates a high-resolution camera, an Ouster OS1-64 LiDAR sensor, and an NVIDIA Jetson AGX Orin edge-computing unit.

The LiDAR and camera are mounted at opposite ends of the platform to provide complementary 360-degree scene coverage, as illustrated in Fig. 2(c). The LiDAR is installed in an inverted configuration to reduce reflections from buildings, trees, and other tall structures and to improve robustness under dust, rain, and similar site conditions. Because field LiDAR acquisition produces a sustained high data rate, the Jetson AGX Orin is used for synchronized capture, buffering, transmission, and local storage. Detailed specifications of the sensing platform are listed in Table 2.

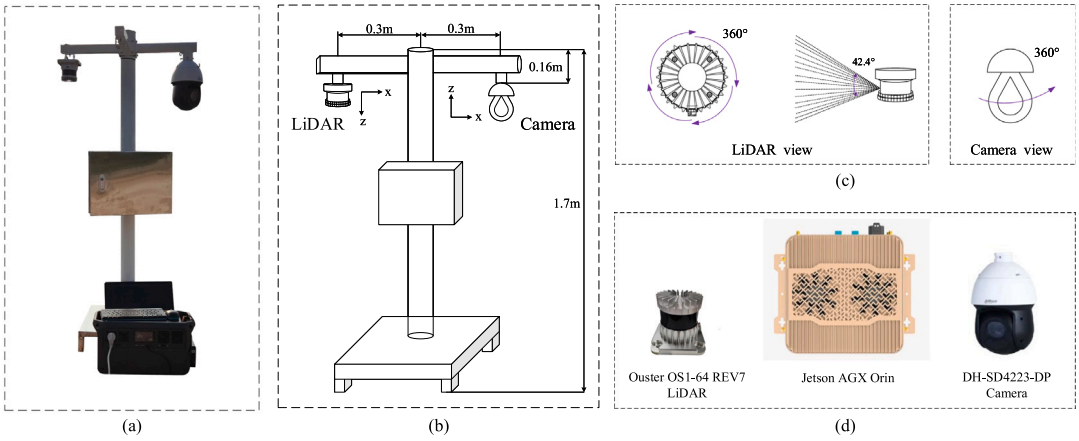


Fig. 2. The developed platform for real-time 3D point cloud data acquisition. (a) The developed LiDAR-camera acquisition platform. (b) Schematic illustration of the platform configuration and sensor layout. (c) Sensor placement and fields of view, where the LiDAR and camera are mounted at opposite ends of the platform to maximize scene coverage. (d) Main hardware components of the platform, including the Ouster OS1-64 REV7 LiDAR, Jetson AGX Orin computing unit, and DH-SD4223-DP camera.

Table 2
Details of the data-acquisition platform.

Camera	
Maximum Resolution	1920 × 1080
Field of View	H: 55.4°, V: 31°, D: 60.7°
Illumination Range	100 m
Lens Focal Length	4.8–110 mm
OUSTER OS1-64	
Number of Channels	64
Range	170 m
Field of View	H: 360°, V: 42.4°
Rotation Rate	10/20 Hz
Points Per Second	2,621,440
NVIDIA Jetson AGX Orin (64 GB)	
Storage	64 GB
CPU	12-core Arm Cortex-A78AE 64-bit
GPU	2048-core NVIDIA Ampere GPU
Memory	64 GB

Table 3
Annotated categories and default annotation box dimensions.

Category	Default dimensions (L/W/H)
Car	4.6/1.8/1.7 m
Truck	9.3/3.0/3.7 m
Digger	9.5/4.0/6.0 m
Bulldozer	6.3/3.0/3.3 m

3.2. Data annotation

Synchronized scene photographs are used as visual references during annotation, so the LiDAR point clouds are first transformed into the camera coordinate system. This is accomplished by aligning the X-axes of the two coordinate systems on the same horizontal plane and separating the coordinate origins by a fixed distance of 60 cm. For a LiDAR point $\mathbf{p}_L = [X_L, Y_L, Z_L]$, the inverted installation of the sensor requires the following coordinate reversal and transformation:

$$\mathbf{p}_C = \mathbf{R}\mathbf{p}'_L + \Delta\mathbf{t}, \tag{1}$$

$$\mathbf{R} = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}, \tag{2}$$

where $\mathbf{p}'_L = [X_L, Y_L, -Z_L]$ denotes the Z-axis-reversed LiDAR coordinates, $\mathbf{R}$ is the rotation matrix, $\Delta\mathbf{t} = [60, 0, 0]$ cm is the translation vector from the LiDAR coordinate system to the camera coordinate system, θ is the rotation angle between the two sensors, and $\mathbf{p}_C$ is the point transformed into the camera coordinate system. After this transformation, standard range and height filtering are applied before manual labeling.

All 3D bounding boxes were manually annotated in CVAT by a third-party annotation company (CVAT, 2021). The research team subsequently reviewed the delivered annotations, and boxes identified as inaccurate or inconsistent were returned to the company for correction. This review-and-revision cycle was repeated over multiple rounds. During both annotation and review, synchronized scene photographs were used as common visual references for object identity and box placement, while the category-specific default dimensions in Table 3 were used as shared references for bounding-box scale and consistency. The four labeled categories, Car, Truck, Digger, and Bulldozer, were selected because they represent the vehicles and machinery most frequently encountered in the target construction and roadwork scenes. The benchmark contains 29,030 manually annotated 3D boxes across these categories.

3.3. Category descriptions

Car denotes passenger vehicles commonly observed on access roads and in work areas. The default annotation dimensions are 4.6 m × 1.8 m × 1.7 m, which are consistent with a mid-sized sedan.

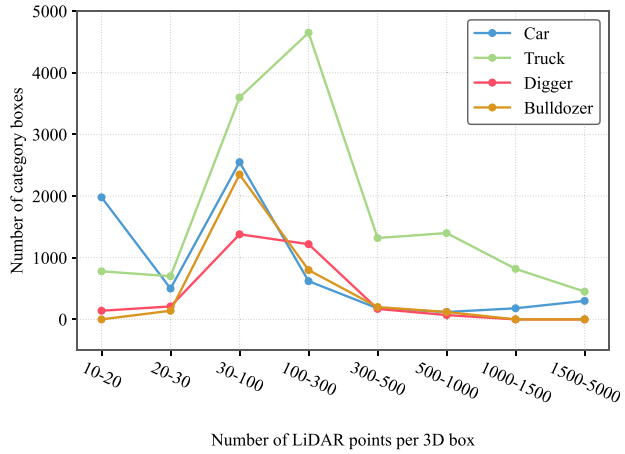
Truck denotes heavy vehicles used for construction transport and roadwork logistics. The default annotation dimensions are 9.3 m × 3.0 m × 3.7 m, reflecting the size of heavy-duty trucks in the captured scenes.

Digger denotes excavating machinery used for earthwork and material handling. The default annotation dimensions are 9.5 m × 4.0 m × 6.0 m, capturing the size envelope needed to label the excavator body and working structure.

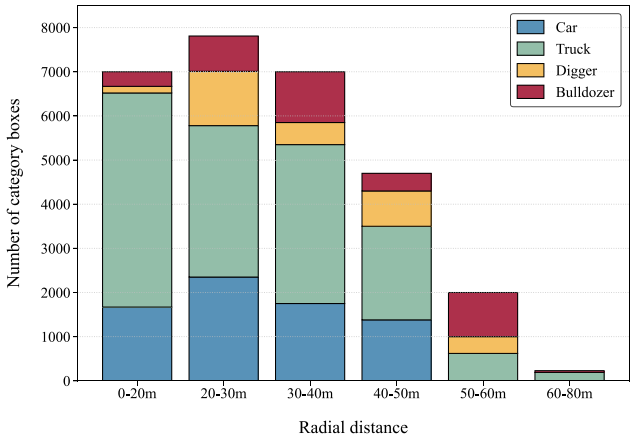
Bulldozer denotes tracked machinery used for pushing and grading material in unstructured terrain. The default annotation dimensions are 6.3 m × 3.0 m × 3.3 m, which reflect the typical geometry of bulldozers observed in the benchmark scenes.

3.4. Dataset analysis

Fig. 3(a) presents the distribution of point counts across object categories in ConVeh3D, and Fig. 3(b) shows the radial distance distribution of annotated 3D boxes relative to the sensor. Together, these



(a)



(b)

Fig. 3. Distribution of (a) per-object point counts and (b) radial distances across the annotated categories in ConVeh3D.

plots summarize the sensing characteristics that distinguish ConVeh3D from conventional driving-oriented point cloud benchmarks.

3.4.1. *Construction-oriented composition*

ConVeh3D emphasizes construction machinery and work-zone vehicles that are underrepresented in conventional autonomous-driving datasets. Trucks, diggers, and bulldozers appear throughout the benchmark rather than as rare edge cases, so the category mix places greater weight on heavy-equipment recognition and localization than car-dominated driving benchmarks.

3.4.2. *Scene diversity and sparsity*

ConVeh3D spans construction and roadwork scenes with open work areas, irregular machinery placement, and cluttered backgrounds. Compared with objects in structured urban-road scenes, many objects in ConVeh3D appear farther from the sensor, typically 20–30 m away, and are often partially occluded. These conditions yield sparse observations: most objects are represented by approximately 30–300 points. The corresponding point-count and radial-distance distributions for KITTI are provided in Figs. 4 and 5 for reference.

3.4.3. *Benchmark role*

ConVeh3D contains 4195 LiDAR frames, approximately 25 billion points, and 29,030 annotated 3D bounding boxes. Within this paper, it functions as the common construction-scene LiDAR benchmark for

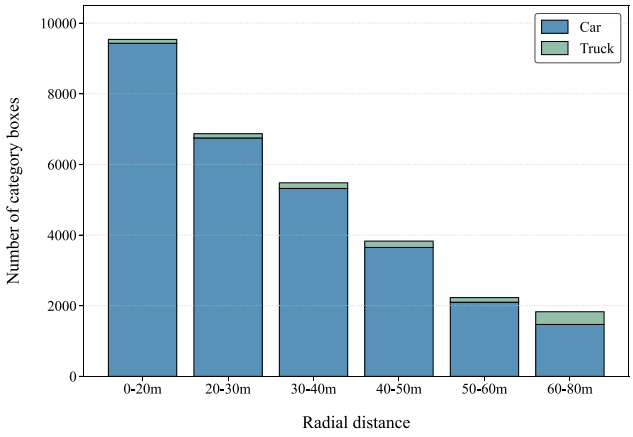
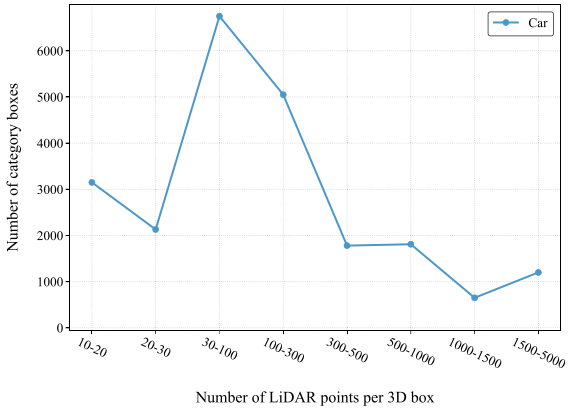
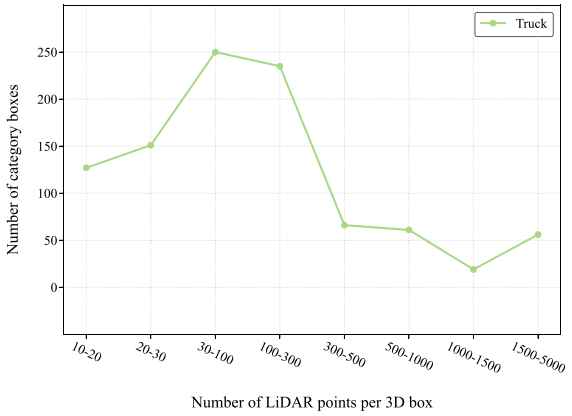


Fig. 4. Radial distance distributions of the Car and Truck categories in the KITTI dataset.



(a)



(b)

Fig. 5. Distribution of LiDAR point counts per 3D bounding box for (a) Car and (b) Truck in the KITTI dataset.

annotation-efficient learning, for evaluation in the category-withheld evaluation setting, and for fully supervised 3D detector comparison under the same field acquisition conditions.

4. Methodology

HCAL is an unknown-category-aware active-learning strategy for annotation-efficient LiDAR-based 3D object detection in the category-withheld evaluation setting used in this paper. The method is designed

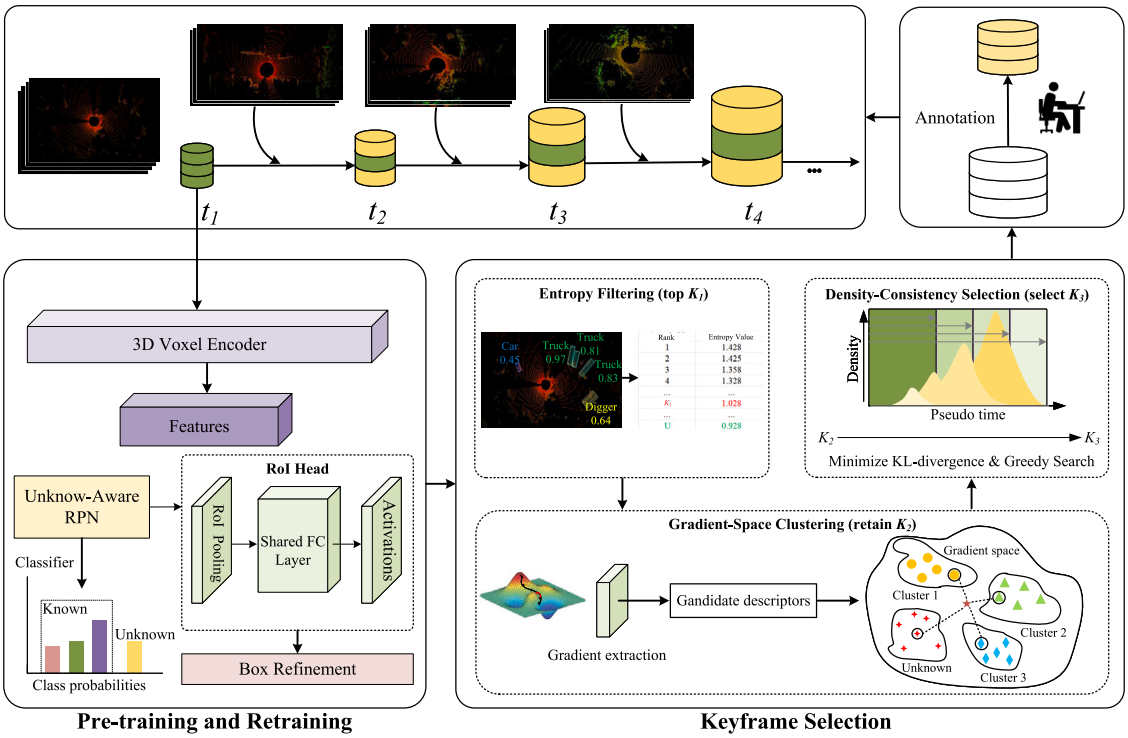


Fig. 6. Framework of the proposed HCAL. The detector is first pre-trained on the initial labeled set and then iteratively updated through active learning. In each round, entropy filtering, gradient-space clustering, and density-consistency selection are sequentially applied to identify informative keyframes for annotation.

for construction-scene LiDAR sensing, where point clouds are sparse, scene layouts are irregular, and manual 3D bounding-box annotation is expensive. Rather than relying on a single query criterion, HCAL performs sample selection in three consecutive stages: entropy filtering ϕ_{En}, gradient-space clustering ϕ_{G}, and density-consistency selection ϕ_{D}. This hierarchical design first identifies informative frames, then removes redundant high-uncertainty samples, and finally preserves coverage across different density conditions in the current unlabeled pool. The overall workflow is illustrated in Fig. 6.

4.1. Active-learning formulation

In this paper, each queried sample is a LiDAR keyframe. Let $\mathcal{L} = \{(x_l, y_l)\}_{l=1}^{L}$ denote the initial labeled set, $\mathcal{U} = \{x_u\}_{u=1}^{U}$ the unlabeled pool, B the total labeling budget measured in queried keyframes, and b the number of keyframes annotated per active-learning round. The cumulative queried set is denoted by $\mathcal{D}_{\mathrm{obj}}$, which is initialized as $\mathcal{D}_{\mathrm{obj}} = \emptyset$ and updated after each round.

$$T = \left\lfloor \frac{B}{b} \right\rfloor, \tag{3}$$

where T is the total number of sampling rounds allowed by the budget. In the nth round, HCAL selects a keyframe subset $\mathcal{S}_n \subset \mathcal{U}$ with $|\mathcal{S}_n| = b$. After expert annotation, the queried set is updated by:

$$\mathcal{D}_{\mathrm{obj}} \leftarrow \mathcal{D}_{\mathrm{obj}} \cup \mathcal{S}_n, \tag{4}$$

and the unlabeled pool is updated by:

$$\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{S}_n. \tag{5}$$

After each query round, the detector is retrained on the accumulated labeled data $\mathcal{L} \cup \mathcal{D}_{\mathrm{obj}}$. Within HCAL, the queried subset $\mathcal{S}_n$ is obtained through three successive candidate sets: $\mathcal{K}_1$, $\mathcal{K}_2$, and $\mathcal{K}_3$, with target sizes K_1, K_2, and K_3, respectively. The final queried subset is $\mathcal{S}_n = \mathcal{K}_3$, so $b = K_3$. This round-based formulation keeps the method focused on annotation-efficient sample selection in the category-withheld evaluation setting used in the experiments.

Algorithm 1 summarizes the operational HCAL loop. The key methodological point is that the three query stages are applied sequentially within each active-learning round: entropy filtering provides an uncertainty-focused candidate pool, gradient-space clustering suppresses redundant selections, and density-consistency selection preserves coverage of sparse and dense regions in the current unlabeled pool. This hierarchical coupling, rather than any single scoring component in isolation, defines the methodological contribution of HCAL.

4.2. Object detection model

HCAL is implemented on top of a SECOND-based detector (Yan et al., 2018), which provides the class probabilities, gradient features, and bounding-box predictions required by the three query stages. The detector uses sparse voxel convolution for point cloud feature extraction, an unknown-aware Region Proposal Network (RPN) for soft category prediction, and a Region of Interest (RoI) head for box refinement. Gradient descriptors are extracted from the shared fully connected layers of the RoI head and are later used in the gradient-space clustering stage. In this implementation, the detector serves as the scoring backbone, whereas the methodological novelty of HCAL lies in the hierarchical query strategy built on top of these signals.

The detector is first pre-trained on a randomly selected labeled subset to establish stable predictions on construction-scene LiDAR frames. During active learning, the detector is retrained after each query round using the cumulative labeled set $\mathcal{L} \cup \mathcal{D}_{\mathrm{obj}}$. Under the category-withheld evaluation setting adopted in this paper, this update cycle is intended to improve annotation efficiency while tracking performance on both the labeled categories and the withheld category used for evaluation.

Algorithm 1 Round-based HCAL query procedure.

Input:
  Data: labeled set $\mathcal{L}$, unlabeled pool $\mathcal{U}$, annotation function Annotate(·)
  Model: detector $f_\theta(\cdot)$
  Selection operators: ϕ_{En} (entropy filtering), ϕ_{G} (gradient-space clustering), ϕ_{D} (density-consistency selection)
  Budget: total budget B, round budget b, stage sizes K_1, K_2, and K_3
Pre-train f_θ on $\mathcal{L}$
$\mathcal{D}_{\text{obj}} \leftarrow \emptyset$
$T \leftarrow \lfloor B/b \rfloor$
for $n = 1, \ldots, T$ **do**
  Infer class probabilities, bounding boxes, and gradient descriptors for all samples in $\mathcal{U}$
  $\mathcal{K}_1 \leftarrow \phi_{\text{En}}(\mathcal{U}, K_1)$
  $\mathcal{K}_2 \leftarrow \phi_{\text{G}}(\mathcal{K}_1, K_2)$
  $\mathcal{K}_3 \leftarrow \phi_{\text{D}}(\mathcal{K}_2, K_3)$
  $S_n \leftarrow \text{Annotate}(\mathcal{K}_3)$
  $\mathcal{D}_{\text{obj}} \leftarrow \mathcal{D}_{\text{obj}} \cup S_n$
  $\mathcal{U} \leftarrow \mathcal{U} \setminus S_n$
  Retrain f_θ on $\mathcal{L} \cup \mathcal{D}_{\text{obj}}$
end for
return f_θ

4.3. Keyframe selection strategy

After the initial pre-training stage, HCAL applies a three-stage keyframe selection strategy within each active-learning round. A single criterion is insufficient for construction-scene LiDAR frames because sparse observations, long-range targets, and scene clutter can make uncertainty estimates noisy and can also cause redundant frames to dominate the candidate pool. HCAL therefore uses entropy filtering to identify informative frames, gradient-space clustering to retain representative frames, and density-consistency selection to preserve coverage across different point-density regions. The candidate sets produced by these three stages are denoted by $\mathcal{K}_1$, $\mathcal{K}_2$, and $\mathcal{K}_3$, respectively.

4.3.1. Entropy filtering

The first stage ranks unlabeled keyframes by predictive uncertainty. This is useful in construction-scene LiDAR because sparse returns and partial occlusions often lead to ambiguous box-level predictions, especially near the withheld category used in evaluation. For an unlabeled keyframe $x \in \mathcal{U}$, suppose that the detector predicts N bounding boxes. For the ith box, the classification head outputs the soft probability vector $\mathbf{p}_i = [p_{i,1}, p_{i,2}, \ldots, p_{i,C}, p_{i,u}]$, where C is the number of labeled categories and $p_{i,u}$ denotes the score of the unknown-category channel used in the category-withheld evaluation setting. The box-level entropy is defined as:

$$h_i = -\sum_{c=1}^{C} p_{i,c} \log p_{i,c} - p_{i,u} \log p_{i,u}, \tag{6}$$

$$H(x) = \frac{1}{N} \sum_{i=1}^{N} h_i, \tag{7}$$

$$\mathcal{K}_1 = \underset{x \in \mathcal{U}}{\text{TopK}} \left(H(x), K_1 \right), \tag{8}$$

where $H(x)$ denotes the average entropy of the predicted boxes in keyframe x, and $\mathcal{K}_1$ retains the K_1 most uncertain keyframes. This stage is used only to screen informative candidates and does not explicitly recognize categories beyond the category-withheld evaluation setting.

4.3.2. Gradient- space clustering

Entropy filtering can still return many redundant frames from visually similar work zones. HCAL therefore performs a second-stage selection in gradient space. For unlabeled keyframes, the classification branch uses the maximum-probability pseudo-label $\hat{y} = \arg\max_c p(c \mid x)$, and the regression branch uses MC-dropout to stabilize predicted boxes before gradient extraction. The resulting gradient descriptors from the shared RoI layers indicate how strongly each keyframe would affect parameter updates if it were annotated.

Let $G_{\mathcal{A}}$ denote the set of gradient descriptors extracted from a subset $\mathcal{A} \subseteq \mathcal{K}_1$, and let $G_{\mathcal{U}}$ denote the corresponding descriptor set estimated from the current unlabeled pool. We use the current unlabeled pool, rather than any held-out evaluation split, as the target distribution so that the query set remains representative of the data available for annotation in the active-learning round. The second-stage candidate set $\mathcal{K}_2$ is obtained by minimizing the KL divergence between these two gradient distributions:

$$\mathcal{K}_2 = \arg \min_{\mathcal{A} \subseteq \mathcal{K}_1,\ |\mathcal{A}|=K_2} D_{\text{KL}} \left(P_{G_{\mathcal{A}}} \parallel P_{G_{\mathcal{U}}} \right), \tag{9}$$

where $P_{G_{\mathcal{A}}}$ and $P_{G_{\mathcal{U}}}$ denote the probability distributions of the gradient descriptors for the selected subset and the current unlabeled pool, respectively. Assuming Gaussian statistics $G_{\mathcal{A}} \sim \mathcal{N}(\mu_{\mathcal{A}}, \sigma^2_{\mathcal{A}})$ and $G_{\mathcal{U}} \sim \mathcal{N}(\mu_{\mathcal{U}}, \sigma^2_{\mathcal{U}})$, the divergence is defined as:

$$D_{\text{KL}} \left(P_{G_{\mathcal{A}}} \parallel P_{G_{\mathcal{U}}} \right) = \frac{1}{2} \left[\frac{\sigma^2_{\mathcal{A}}}{\sigma^2_{\mathcal{U}}} + \frac{(\mu_{\mathcal{A}} - \mu_{\mathcal{U}})^2}{\sigma^2_{\mathcal{U}}} - 1 + \log \frac{\sigma^2_{\mathcal{U}}}{\sigma^2_{\mathcal{A}}} \right]. \tag{10}$$

In practice, hierarchical clustering is applied to the entropy-filtered set $\mathcal{K}_1$ to avoid repeatedly selecting near-duplicate high-entropy frames. Let $\{C_1, C_2, \ldots, C_m\}$ denote the resulting clusters. Each cluster is scored by how closely its gradient statistics match those of the unlabeled pool:

$$\text{score}(C_j) = \frac{\sigma^2_{C_j}}{\sigma^2_{\mathcal{U}}} + \frac{(\mu_{C_j} - \mu_{\mathcal{U}})^2}{\sigma^2_{\mathcal{U}}} + \log \frac{\sigma^2_{\mathcal{U}}}{\sigma^2_{C_j}}. \tag{11}$$

Representative keyframes are then drawn from the clusters with lower scores until $|\mathcal{K}_2| = K_2$. This stage keeps the selected set informative while reducing redundancy among uncertain construction-scene frames.

4.3.3. Density-consistency selection

The third stage refines $\mathcal{K}_2$ into the final query set by preserving density coverage. Construction-scene LiDAR data exhibit substantial density variation caused by range changes, occlusion, and object-size imbalance, so a representative query set should not focus exclusively on either dense or sparse observations. Let ρ denote the point density associated with a predicted bounding box. Because the density function is not available analytically, HCAL estimates it by kernel density estimation:

$$\hat{p}(\rho) = \frac{1}{N\tau} \sum_{i=1}^{N} K \left(\frac{\rho - \rho_i}{\tau} \right), \tag{12}$$

where N is the number of bounding boxes used for density estimation, τ is the kernel bandwidth, and $K(\cdot)$ is the Gaussian kernel. The target density distribution is estimated from the current unlabeled pool and denoted by $P_{\mathcal{U}}$.

Starting from $\mathcal{K}_3 = \emptyset$, HCAL greedily adds samples from $\mathcal{K}_2$ so that the density distribution of the selected set remains close to $P_{\mathcal{U}}$. The KL divergence used in this stage is defined as:

$$D_{\text{KL}} \left(P_{\mathcal{K}_3 \cup \{z\}} \parallel P_{\mathcal{U}} \right) = \sum_{\rho} P_{\mathcal{K}_3 \cup \{z\}}(\rho) \log \frac{P_{\mathcal{K}_3 \cup \{z\}}(\rho)}{P_{\mathcal{U}}(\rho)}, \tag{13}$$

where $P_{\mathcal{K}_3 \cup \{z\}}$ denotes the density estimate of the candidate set after adding candidate keyframe z.

At each step, the next query keyframe is selected as:

$$z^* = \arg \min_{z \in \mathcal{K}_2 \setminus \mathcal{K}_3} D_{\text{KL}} \left(P_{\mathcal{K}_3 \cup \{z\}} \parallel P_{\mathcal{U}} \right). \tag{14}$$

The selected keyframe is then added to $\mathcal{K}_3$, and the divergence is recomputed until $|\mathcal{K}_3| = K_3$. The resulting set $\mathcal{K}_3$ is sent for annotation in the current active-learning round.

Table 4
ConVeh3D average AP under the category-withheld evaluation setting, with Truck treated as the unknown category. The best and second-best are bolded and underlined, respectively.

Category	Labeled frames	Random	Cider	Entropy	GradNorm	React	Coreset	OpenCRB	Ours
Average	500	43.18	39.57	40.57	40.21	38.55	43.98	<u>46.03</u>	**46.53**
	700	56.57	51.31	48.51	56.24	50.94	<u>58.03</u>	57.71	**59.43**
	900	63.38	61.44	58.62	62.83	57.54	**66.12**	63.49	<u>64.75</u>
	1100	66.60	64.80	60.16	67.09	62.27	<u>68.42</u>	67.47	**69.17**
	1300	68.51	66.49	67.24	68.09	64.40	<u>69.55</u>	69.11	**69.86**
	1500	70.22	68.58	69.25	70.68	66.44	70.78	<u>71.29</u>	**71.57**

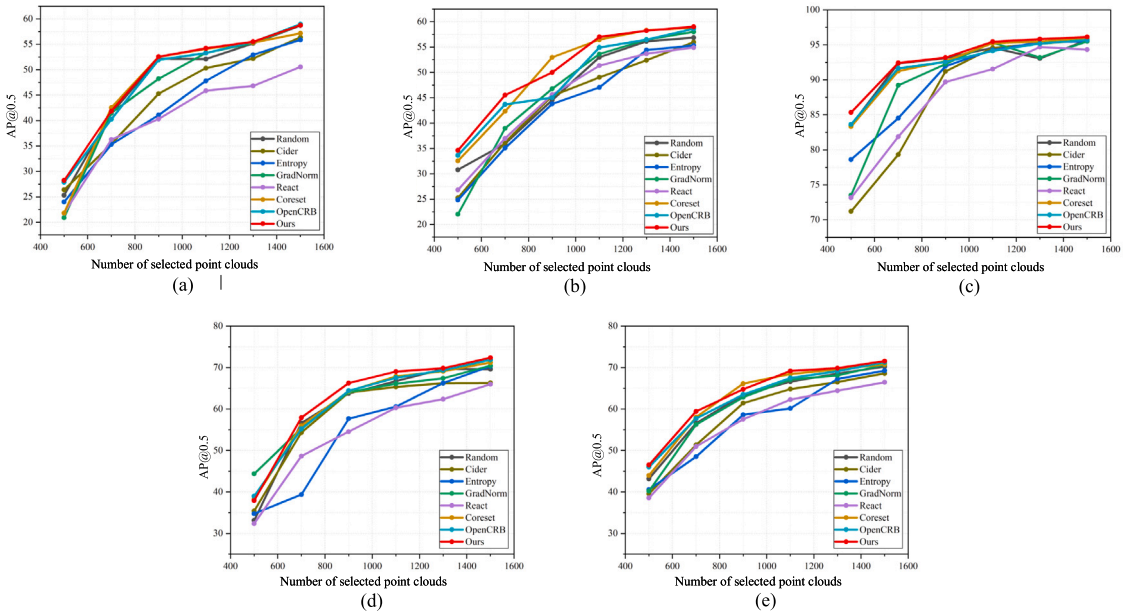


Fig. 7. Active-learning performance curves on ConVeh3D under the category-withheld evaluation setting, showing AP@0.5 for (a) Car, (b) Truck, (c) Digger, (d) Bulldozer, and (e) the average across categories.

5. Experiments

We evaluate HCAL as an unknown-category-aware active-learning method in a category-withheld evaluation setting on ConVeh3D and KITTI. ConVeh3D provides the primary construction-scene LiDAR evidence, while KITTI serves as secondary supporting evidence under the same evaluation setting. ConVeh3D is also the benchmark used for the retained fully supervised detector runs. In this paper, "fully supervised" refers to training other 3D detectors on labeled ConVeh3D scenes rather than to another active-learning setting. The active-learning results reported below focus on HCAL, six active-learning baselines, and random sampling.

5.1. Experimental protocol

ConVeh3D contains 4195 LiDAR frames in total, with 2546 frames used for training and 1649 for testing. In the category-withheld evaluation setting, Truck is treated as the unknown category, and AP@0.5 is reported for all categories. The initial labeled set contains 500 training frames. In each active-learning round, HCAL first forms $K_1 = 1000$ entropy-filtered candidates, then reduces them to $K_2 = 600$ and finally to $K_3 = 200$ queried frames, consistent with the round-based procedure in Section 4. These values implement a progressive 5:3:1 reduction: K_1 retains a broad entropy-ranked candidate pool, K_2 narrows this pool after gradient-space redundancy suppression, and K_3 equals the per-round annotation budget b. The same stage sizes are used throughout the ConVeh3D active-learning rounds. Table 4 and Fig. 7 therefore report cumulative labeled-set sizes from 500 to 1500 labeled frames. Table 4 presents the average results, while the detailed per-category results are provided in Appendix Table 8. The same labeled-sample checkpoints are used for the compared active-learning baselines.

For KITTI, 3712 frames are used for training and 3769 for testing following common practice. Pedestrian is treated as the unknown category in the category-withheld evaluation setting, and the reported results follow the Easy, Moderate, and Hard difficulty levels. AP@0.7 is reported for Car, while AP@0.5 is reported for Pedestrian and Cyclist. The initial labeled set contains 400 training frames. The reported active-learning checkpoints in Tables 5 and 6 and Fig. 8 are 400, 600, 800, 1,200, and 1,400 labeled frames. As in ConVeh3D, the compared active-learning methods are evaluated at matched labeled-sample checkpoints.

To ensure a fair comparison, all active-learning methods were evaluated using the same training/test split and exactly the same initial labeled set for each benchmark. The initial labeled set was randomly selected once and then fixed throughout all experiments, so every compared method started from identical labeled and unlabeled pools. This protocol controls variation from data initialization; consequently, performance differences at subsequent active-learning checkpoints mainly reflect the effectiveness of the sample-selection strategies rather than differences in the initial data pools.

5.2. Experimental settings

Both benchmarks adopt Average Precision (AP) calculated at 40 recall positions as the core evaluation metric, where point cloud is specified as the labeling-budget unit for all comparative experiments. The active-learning experiments are conducted on a single NVIDIA A40 GPU using the Adam optimizer with a batch size of 4. For ConVeh3D,

Table 5
KITTI Pedestrian AP under the category-withheld evaluation setting, with Pedestrian treated as the unknown category. The best and second-best are bolded and underlined, respectively.

Difficulty	Labeled frames	Random	Cider	Entropy	GradNorm	React	Coreset	OpenCRB	Ours
Easy	400	34.10	24.64	34.33	**34.67**	26.48	23.26	30.33	29.43
	600	**45.58**	45.01	45.23	39.25	43.82	39.44	44.25	45.33
	800	47.65	47.26	48.31	40.54	45.02	43.91	49.02	**50.13**
	1200	50.38	48.07	51.01	44.55	46.92	45.66	50.93	**53.91**
	1400	51.09	52.45	52.14	52.00	42.50	49.62	53.85	**59.01**
Moderate	400	31.32	23.24	31.25	**31.45**	24.60	28.20	27.00	26.58
	600	**40.11**	39.08	39.77	35.71	39.65	35.73	36.01	37.94
	800	40.53	41.89	41.35	39.33	40.32	39.88	42.12	**43.50**
	1200	44.87	44.91	45.12	42.54	42.92	42.61	45.90	**47.81**
	1400	45.35	46.06	45.72	42.63	37.19	42.95	46.96	**52.15**
Hard	400	28.33	21.82	28.34	**28.43**	22.05	24.12	27.83	24.41
	600	35.23	37.20	**38.23**	35.45	35.61	33.45	35.27	35.95
	800	36.25	37.98	37.55	36.95	36.74	35.60	39.33	**39.67**
	1200	39.89	38.16	40.89	40.33	38.09	37.83	41.01	**41.93**
	1400	40.71	41.57	41.12	44.63	33.45	39.07	45.20	**46.15**

Table 6
KITTI average AP under the same active-learning checkpoints used for the category-withheld evaluation setting. The best and second-best are bolded and underlined, respectively.

Difficulty	Labeled frames	Random	Cider	Entropy	GradNorm	React	Coreset	OpenCRB	Ours
Easy	400	45.39	35.92	54.60	46.32	35.76	**59.33**	54.00	48.33
	600	69.28	69.74	**70.17**	67.56	68.81	69.29	69.82	70.02
	800	71.18	70.53	70.48	70.46	70.39	70.01	70.88	**72.22**
	1200	72.69	71.42	72.34	71.93	72.15	71.33	71.95	**74.18**
	1400	73.22	74.26	73.44	73.89	68.51	73.22	75.45	**77.49**
Moderate	400	34.69	28.47	37.99	36.09	26.51	**42.33**	41.22	40.99
	600	54.57	54.32	**55.35**	53.89	54.74	52.27	51.20	54.66
	800	56.18	55.49	57.35	57.01	56.25	56.01	56.08	**57.89**
	1200	57.32	57.01	58.38	57.02	57.35	57.25	57.76	**59.35**
	1400	58.14	60.01	59.61	59.02	55.22	59.22	59.87	**62.82**
Hard	400	31.98	26.74	32.35	34.22	24.34	**41.02**	38.34	37.35
	600	50.08	50.08	**51.02**	49.10	50.43	49.41	49.77	50.50
	800	52.11	51.32	52.01	51.09	51.01	51.12	**52.16**	52.03
	1200	52.25	52.87	52.44	54.02	53.06	53.25	53.05	**54.19**
	1400	55.67	55.71	55.23	56.43	51.71	54.97	57.68	**57.86**

the primary evidence in this section comes from the active-learning setting, where the labeled set grows round by round according to the queried sample budget. The retained fully supervised detector runs on ConVeh3D use a separate training setting: a single RTX 3090 GPU, the Adam optimizer, a batch size of 2200 training epochs, cosine annealing learning-rate decay, and an initial learning rate of 0.001. For two-stage detectors, 128 candidate boxes are randomly selected from each scene for second-stage training, and 50% of them satisfy a minimum 3D overlap threshold of 0.55.

5.3. Active-learning results

For active-learning comparison, HCAL is evaluated against six active-learning baselines, namely Cider, Entropy, GradNorm, React, Coreset, and OpenCRB, as well as random sampling. (Ming et al., 2022; Wu et al., 2022; Huang et al., 2021; Liu et al., 2020; Sener and Savarese, 2017; Chen et al., 2023). ConVeh3D provides the primary task-validating evidence for construction-scene LiDAR sensing, while KITTI is used as a secondary benchmark under the same evaluation setting. The quantitative results are summarized in Tables 4–6.

5.3.1. Quantitative results on the ConVeh3D dataset

Table 4 and Fig. 7 report the ConVeh3D results with Truck withheld as the unknown category. Table 4 summarizes the average performance, while the detailed results for each category are provided in Appendix Table 8. Detection performance generally increases as the labeled set grows from 500 to 1,500 labeled frames, which is consistent with the round-based active-learning protocol. HCAL achieves the highest average AP at five of the six reported budgets and remains competitive at 900 labeled frames, where Coreset reaches 66.12 compared with 64.75 for HCAL. This budget-wise behavior is more informative than isolated best cells because it shows how the method uses additional annotation budget on the construction-scene LiDAR benchmark.

For the held-out Truck category, HCAL becomes the highest-scoring method from 700 labeled frames onward and reaches 72.37 AP at 1,500 labeled frames, compared with 71.88 for OpenCRB and 71.22 for Coreset. On the overall average at 1,500 labeled frames, the margin is smaller: 71.57 for HCAL versus 71.29 for OpenCRB. Accordingly, the ConVeh3D results support a controlled claim: HCAL improves unknown-category-aware sample selection under the tested construction-scene protocol, while the final average-AP advantage at the largest budget should be read as modest rather than as statistically established superiority.

Fig. 7 shows the same trend visually. The curves indicate that HCAL is especially competitive in the mid-to-high budget range for Truck, Digger, and the overall average, while several baselines remain strongest in isolated Car or Bulldozer settings. This mixed pattern is why the ConVeh3D evidence should be interpreted as more reliable use of the annotation budget under the category-withheld evaluation setting, not as uniform dominance in every category-budget pair.

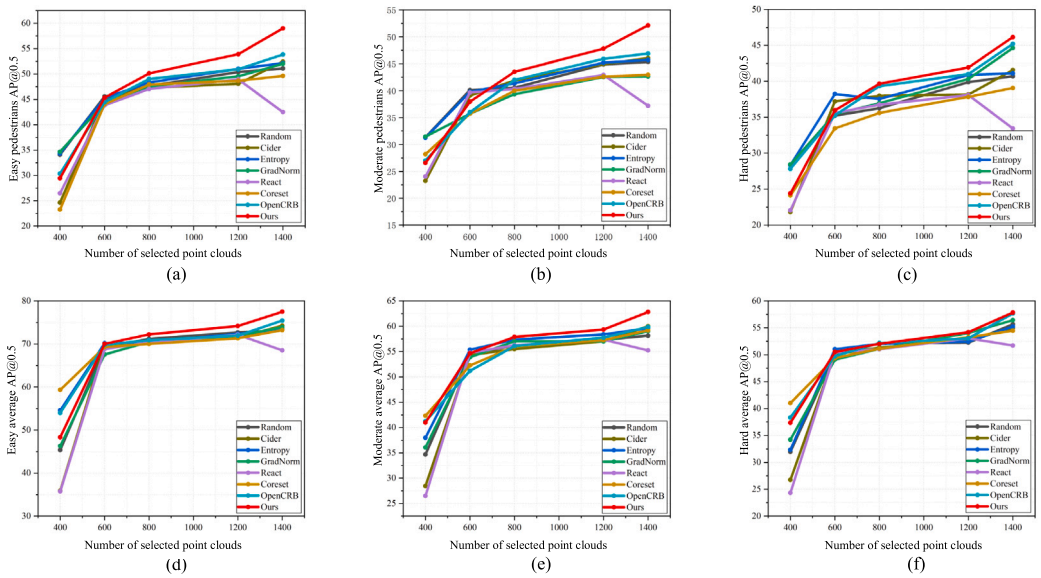


Fig. 8. KITTI active-learning performance curves under the category-withheld evaluation setting: (a) Easy Pedestrian AP, (b) Moderate Pedestrian AP, (c) Hard Pedestrian AP, (d) Easy average AP, (e) Moderate average AP, and (f) Hard average AP.

Table 7
Ablation of the HCAL query stages on ConVeh3D under the category-withheld evaluation setting. HGradNorm denotes hierarchical clustering in the gradient space.

Entropy	HGradNorm	Density	mAP
✓			69.25
	✓		70.68
✓	✓		71.22
✓	✓	✓	**71.57**

5.3.2. Quantitative results on the KITTI dataset

Tables 5 and 6, together with Fig. 8, evaluate the same category-withheld evaluation setting on KITTI, with Pedestrian withheld as the unknown category. This secondary benchmark is useful for testing whether the HCAL querying strategy transfers beyond ConVeh3D, but it does not by itself establish broad generalization beyond the reported settings. The results are mixed at the smaller budgets and become more favorable to HCAL once the labeled set reaches 800 to 1,400 labeled frames.

For the held-out Pedestrian category, HCAL achieves the highest AP at 800, 1,200, and 1,400 labeled frames across Easy, Moderate, and Hard, reaching 59.01, 52.15, and 46.15 at 1,400 labeled frames. For the average AP over categories, HCAL is highest at 1,400 labeled frames on Easy and Moderate, with a small numerical lead on Hard (57.86 versus 57.68 for OpenCRB). These results suggest that the hierarchical query strategy remains effective under a second benchmark and a different withheld category, but the evidence should still be read as performance under the tested settings rather than as a broad generalization claim.

Fig. 8 complements the tables by showing that several baselines fluctuate across difficulty levels and sample budgets, whereas HCAL tends to improve more steadily once the queried set becomes larger. This behavior is consistent with the method design in Section 4, where entropy filtering, gradient-space clustering, and density-consistency selection are intended to reduce redundant queries and preserve representative coverage.

5.4. Visual comparisons

Fig. 9 provides qualitative examples from ConVeh3D roadwork and building-construction scenes. These visualizations are intended as illustrative support for the quantitative results, not as primary evidence. In the scenes shown, HCAL yields fewer missed Truck detections and fewer false positives than several comparison strategies, especially in cluttered areas with partial occlusion. The figure is therefore most useful for interpreting failure modes under sparse construction-scene LiDAR sensing rather than for making a separate performance claim.

5.5. Ablation studies

Table 7 reports the ablation study on ConVeh3D. Using only entropy filtering yields 69.25 mAP, and using only HGradNorm yields 70.68. Combining the two raises mAP to 71.22, which indicates that uncertainty screening and gradient-space representativeness are complementary in the active-learning setting. Adding density-consistency selection further increases mAP to 71.57, a gain of 0.35 over the two-stage variant. The ablation therefore supports the full HCAL design while also showing that the density-consistency stage provides a smaller incremental contribution than the first two stages.

6. Conclusion

This article presented Hierarchical Clustering-enhanced Active Learning (HCAL) for annotation-efficient LiDAR-based 3D object detection in construction and roadwork environments. HCAL combines entropy filtering, gradient-space clustering, and density-consistency selection so that queried keyframes remain informative while covering the sparse and irregular sensing conditions characteristic of construction scenes. ConVeh3D serves as the primary construction-scene LiDAR benchmark for this study, while KITTI provides secondary supporting evidence under the same category-withheld evaluation setting. On ConVeh3D, HCAL achieves the highest average AP at five of the six reported labeling budgets, and the KITTI results further support the usefulness of the same query strategy under a second benchmark. Taken together, these results indicate that HCAL is effective for reducing annotation demand in construction-scene LiDAR sensing under the tested conditions and that ConVeh3D provides a field-relevant benchmark for studying this problem.

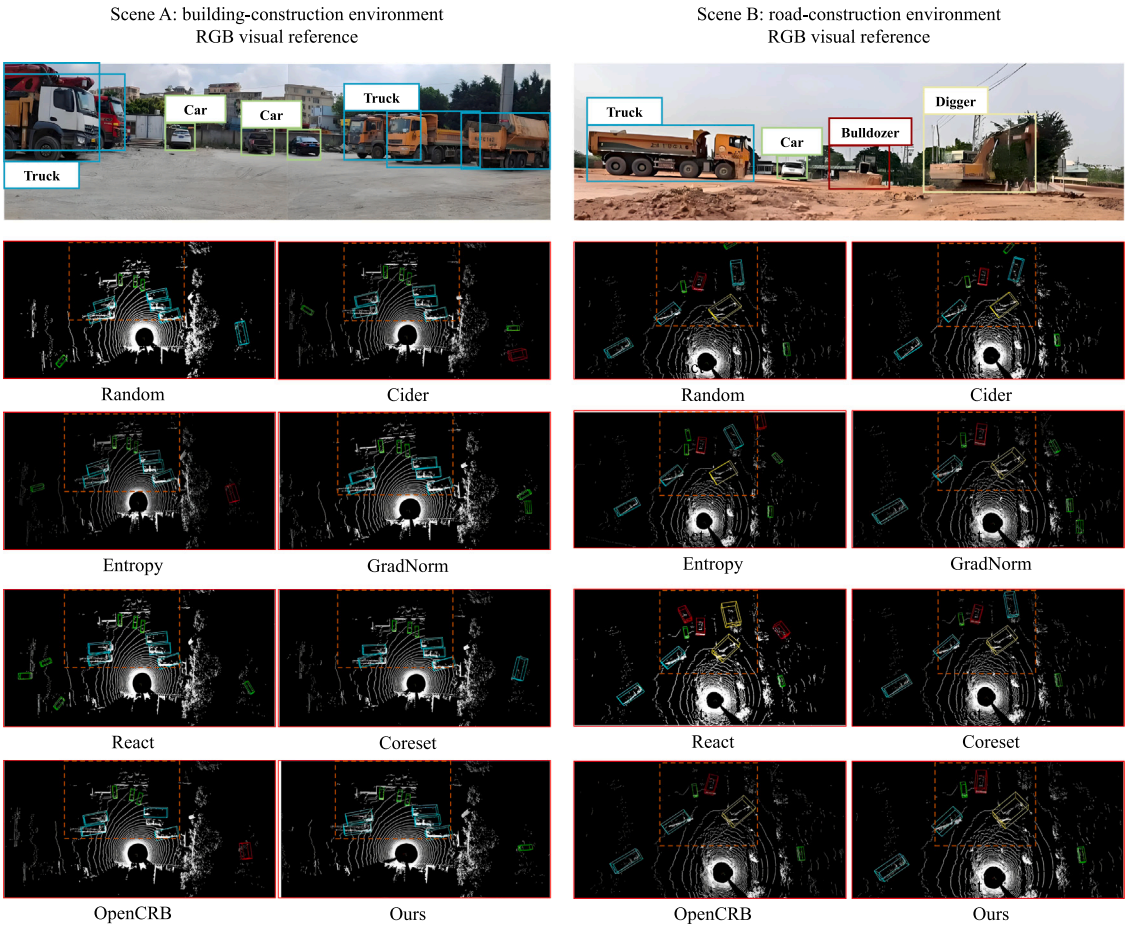


Fig. 9. Qualitative comparison of sampling strategies on ConVeh3D road-construction and building-construction scenes. The orange dashed boxes indicate the field of view of the RGB camera. Green, blue, yellow, and red boxes denote Car, Truck, Digger, and Bulldozer, respectively. Truck is the unknown category in the category-withheld evaluation setting.

CRediT authorship contribution statement

You Li: Writing – review & editing, Methodology, Conceptualization. **Jialu Sui:** Writing – original draft, Methodology. **Yasong Shi:** Data curation. **Xiaokang Zhang:** Writing – review & editing, Methodology. **Shengjun Tang:** Project administration, Investigation. **Ruisheng Wang:** Validation. **Renzhong Guo:** Supervision, Funding acquisition.

Funding

This work was supported by the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) under Grant No. GML-KF-24-28, Natural Science Foundation of Guangdong Province of China under Grant No. 2023A1515010717 and the Research Task Assignment Project from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) under Grant No.GML-26420009.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

See Table 8.

Table 8
Active-learning results for each category on ConVeh3D under the category-withheld evaluation setting, with Truck treated as the unknown category. The best and second-best are bolded and underlined, respectively.

Category	Labeled frames	Random	Cider	Entropy	GradNorm	React	Coreset	OpenCRB	Ours
Car	500	25.32	26.37	24.02	20.91	21.81	21.83	<u>27.88</u>	**28.24**
	700	41.47	35.42	35.32	41.51	36.25	**42.56**	40.28	<u>41.87</u>
	900	<u>52.17</u>	45.29	41.09	48.22	40.32	52.54	51.96	**52.58**
	1100	52.11	50.33	47.81	53.32	45.87	<u>54.11</u>	53.29	**54.21**
	1300	55.18	52.21	52.95	55.28	46.83	55.32	<u>55.48</u>	**55.51**
	1500	<u>58.69</u>	56.39	55.87	58.79	50.55	57.16	**58.95**	58.76

(*continued on next page*)

Table 8 (*continued*).

Category	Labeled frames	Random	Cider	Entropy	GradNorm	React	Coreset	OpenCRB	Ours
Truck	500	33.14	35.42	34.78	**44.37**	32.38	38.21	38.96	37.92
	700	56.62	54.32	39.37	55.31	48.66	55.97	55.21	**57.93**
	900	63.73	63.98	57.66	64.04	54.54	64.28	64.39	**66.24**
	1100	66.72	65.32	60.60	66.10	60.32	67.87	67.54	**69.01**
	1300	69.64	66.21	66.24	67.41	62.38	69.08	69.32	**69.83**
	1500	69.61	66.29	70.32	70.44	65.98	71.22	71.88	**72.37**
Digger	500	30.79	25.26	24.85	22.03	26.83	32.55	33.65	**34.63**
	700	35.94	36.19	35.09	38.95	36.98	42.32	43.69	**45.53**
	900	44.54	45.38	43.79	46.79	45.62	**52.98**	45.03	49.99
	1100	52.99	49.02	47.05	53.61	51.33	56.46	54.92	**57.01**
	1300	56.15	52.39	54.42	56.48	53.69	**58.32**	56.46	58.26
	1500	56.87	55.98	55.32	58.02	54.91	58.84	58.64	**59.01**
Bulldozer	500	83.48	71.21	78.61	73.51	73.16	83.33	83.64	**85.33**
	700	92.36	79.32	84.52	89.21	81.87	91.25	91.66	**92.41**
	900	93.09	91.21	91.95	92.27	89.69	92.66	92.56	**93.17**
	1100	94.57	94.55	94.26	95.32	91.55	95.23	94.12	**95.45**
	1300	93.04	95.16	95.35	93.18	94.68	95.47	95.18	**95.82**
	1500	95.69	95.64	95.49	95.48	94.32	95.93	95.68	**96.13**

Data availability

The ConVeh3D dataset supporting the findings of this study will be made publicly available through the project repository at https://github.com/zxk688/ConVeh3D.

References

Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J., 2019. SemanticKITTI: A dataset for semantic scene understanding of lidar sequences. In: IEEE/CVF International Conference on Computer Vision. pp. 9297–9307.

Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2020. nuScenes: A multimodal dataset for autonomous driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11631.

Chen, K., Guan, H., Wang, L., Yu, Y., Zang, Y., Qin, N., Liu, J., Li, J., 2024. A feature perturbation weakly supervised learning network for airborne multispectral LiDAR pointcloud classification. Int. J. Appl. Earth Obs. Geoinf. 127, 103683.

Chen, Z., Li, Z., Zhang, S., Fang, L., Jiang, Q., Zhao, F., Zhou, B., Zhao, H., 2022. Autoalign: Pixel-instance feature aggregation for multi-modal 3d object detection. arXiv preprint arXiv:2201.06493.

Chen, Z., Luo, Y., Wang, Z., Wang, Z., Yu, X., Huang, Z., 2023. Towards open world active learning for 3d object detection. arXiv preprint arXiv:2310.10391.

Chen, X., Ma, H., Wan, J., Li, B., Xia, T., 2017. Multi-view 3d object detection network for autonomous driving. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1907–1915.

CVAT, 2021. CVAT: Computer vision annotation tool. URL https://github.com/openvinotoolkit/cvat.

Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5828–5839.

Ding, Y., Luo, X., 2024. A virtual construction vehicles and workers dataset with three-dimensional annotations. Eng. Appl. Artif. Intell. 133, 107964.

Ding, Y., Zhang, M., Pan, J., Hu, J., Luo, X., 2024. Robust object detection in extreme construction conditions. Autom. Constr. 165, 105487.

Ducoffe, M., Precioso, F., 2018. Adversarial active learning for deep networks: a margin based approach. arXiv preprint arXiv:1802.09841.

Elhamifar, E., Sapiro, G., Yang, A., Sasrty, S.S., 2013. A convex optimization framework for active learning. In: IEEE International Conference on Computer Vision. pp. 209–216.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361.

Gong, Z., Lin, H., Zhang, D., Luo, Z., Zelek, J., Chen, Y., Nurunnabi, A., Wang, C., Li, J., 2020. A frustum-based probabilistic framework for 3D object detection by fusion of LiDAR and camera data. ISPRS J. Photogramm. Remote Sens. 159, 90–100.

Greer, R., Antoniussen, B., Andersen, M.V., Møgelmose, A., Trivedi, M.M., 2024. The why, when, and how to use active learning in large-data-driven 3D object detection for safe autonomous driving: An empirical exploration. arXiv preprint arXiv:2401.16634.

Hua, B.-S., Pham, Q.-H., Nguyen, D.T., Tran, M.-K., Yu, L.-F., Yeung, S.-K., 2016. Scenenn: A scene meshes dataset with annotations. In: International Conference on 3D Vision. pp. 92–101.

Huang, R., Geng, A., Li, Y., 2021. On the importance of gradients for detecting distributional shifts in the wild. Adv. Neural Inf. Process. Syst. 34, 677–689.

Huang, R., Xu, Y., Hoegner, L., Stilla, U., 2022. Semantics-aided 3D change detection on construction sites using UAV-based photogrammetric point clouds. Autom. Constr. 134, 104057.

Lei, X., Guan, H., Ma, L., Yu, Y., Dong, Z., Gao, K., Delavar, M.R., Li, J., 2022. WSPointNet: A multi-branch weakly supervised learning network for semantic segmentation of large-scale mobile laser scanning point clouds. Int. J. Appl. Earth Obs. Geoinf. 115, 103129.

Lin, J., Liang, Z., Deng, S., Cai, L., Jiang, T., Li, T., Jia, K., Xu, X., 2024. Exploring diversity-based active learning for 3d object detection in autonomous driving. IEEE Trans. Intell. Transp. Syst.

Lin, L., Wang, K., Meng, D., Zuo, W., Zhang, L., 2017. Active self-paced learning for cost-effective and progressive face identification. IEEE Trans. Pattern Anal. Mach. Intell. 40 (1), 7–19.

Liu, W., Wang, X., Owens, J., Li, Y., 2020. Energy-based out-of-distribution detection. Adv. Neural Inf. Process. Syst. 33, 21464–21475.

Ma, T., Yang, X., Zhou, H., Li, X., Shi, B., Liu, J., Yang, Y., Liu, Z., He, L., Qiao, Y., et al., 2023. DetZero: Rethinking offboard 3d object detection with long-term sequential point clouds. In: IEEE/CVF International Conference on Computer Vision. pp. 6736–6747.

Mao, J., Niu, M., Jiang, C., Liang, H., Chen, J., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., et al., 2021. One million scenes for autonomous driving: Once dataset. arXiv preprint arXiv:2106.11037.

Ming, Y., Sun, Y., Dia, O., Li, Y., 2022. How to exploit hyperspherical embeddings for out-of-distribution detection?. arXiv preprint arXiv:2203.04450.

Nath, N.D., Behzadan, A.H., 2020. Deep convolutional networks for construction object detection under different visual conditions. Front. Built Env. 6, 97.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. PointNet: Deep learning on point sets for 3d classification and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660.

Roth, D., Small, K., 2006. Margin-based active learning for structured output spaces. In: European Conference on Machine Learning. pp. 413–424.

Roy, S., Unmesh, A., Namboodiri, V.P., 2018. Deep active learning for object detection.. In: BMVC, Vol. 362, p. 91.

Roynard, X., Deschaud, J.-E., Goulette, F., 2018. Paris-lille-3D: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. Int. J. Robot. Res. 37 (6), 545–557.

Safaei, B., Vibashan, V., de Melo, C.M., Patel, V.M., 2024. Entropic open-set active learning. In: AAAI Conference on Artificial Intelligence, Vol. 38, (5), pp. 4686–4694.

Sener, O., Savarese, S., 2017. Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423.

Shen, J., Jiao, L., Zhang, C., Peng, K., 2024. Monocular 3D object detection for construction scene analysis. Comput.-Aided Civ. Infrastruct. Eng. 39 (9), 1370–1389.

Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H., 2020. PV-RCNN: Point-voxel feature set abstraction for 3d object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10529–10538.

Shi, S., Jiang, L., Deng, J., Wang, Z., Guo, C., Shi, J., Wang, X., Li, H., 2023. PV-rcnn++: Point-voxel feature set abstraction with local vector representation for 3D object detection. Int. J. Comput. Vis. 131 (2), 531–551.

Shi, S., Wang, X., Li, H., 2019. PointRCNN: 3d object proposal generation and detection from point cloud. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–779.

Song, Z., Jia, C., Liu, L., Xu, S., Yang, L., Luo, Y., 2026. GraphBEV++: Multi-modal feature alignment for autonomous driving. arXiv preprint arXiv:2606.16354.

Song, Z., Jia, C., Yang, L., Wei, H., Liu, L., 2023a. GraphAlign++: An accurate feature alignment by graph matching for multi-modal 3D object detection. IEEE Trans. Circuits Syst. Video Technol. 34 (4), 2619–2632.

Song, S., Lichtenberg, S.P., Xiao, J., 2015. SUN RGB-D: A rgb-d scene understanding benchmark suite. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 567–576.

Song, Z., Wei, H., Bai, L., Yang, L., Jia, C., 2023b. GraphAlign: Enhancing accurate feature alignment by graph matching for multi-modal 3D object detection. In: IEEE/CVF International Conference on Computer Vision. pp. 3358–3369.

Song, Z., Yang, L., Xu, S., Liu, L., Xu, D., Jia, C., Jia, F., Wang, L., 2024. Graphbev: Towards robust bev feature alignment for multi-modal 3d object detection. In: European Conference on Computer Vision. Springer, pp. 347–366.

Song, Z., Zhang, G., Xie, J., Liu, L., Jia, C., Xu, S., Wang, Z., 2023c. VoxelNextFusion: A simple, unified, and effective voxel fusion framework for multimodal 3-D object detection. IEEE Trans. Geosci. Remote Sens. 61, 1–12.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al., 2020. Scalability in perception for autonomous driving: Waymo open dataset. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2446–2454.

Wu, J., Chen, J., Huang, D., 2022. Entropy-based active learning for object detection with progressive diversity constraint. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9397–9406.

Xuehui, A., Li, Z., Zuguang, L., Chengzhi, W., Pengfei, L., Zhiwei, L., 2021. Dataset and benchmark for detecting moving objects in construction sites. Autom. Constr. 122, 103482.

Yan, Y., Mao, Y., Li, B., 2018. Second: Sparsely embedded convolutional detection. Sensors 18 (10), 3337.

Yan, T., Yin, J., Lang, X., Yang, R., Xu, C.-Z., Shen, J., 2025. Olidm: Object-aware LiDAR diffusion models for autonomous driving. In: AAAI Conference on Artificial Intelligence, Vol. 39, (9), pp. 9121–9129.

Yang, Z., Chen, J., Miao, Z., Li, W., Zhu, X., Zhang, L., 2022. DeepInteraction: 3d object detection via modality interaction. Adv. Neural Inf. Process. Syst. 35, 1992–2005.

Yang, Z., Song, N., Li, W., Zhu, X., Zhang, L., Torr, P.H., 2025. DeepInteraction++: Multi-modality interaction for autonomous driving. IEEE Trans. Pattern Anal. Mach. Intell.

Yin, J., Shen, J., Gao, X., Crandall, D.J., Yang, R., 2021. Graph neural network and spatiotemporal transformer attention for 3D video object detection from point clouds. IEEE Trans. Pattern Anal. Mach. Intell. 45 (8), 9822–9835.

Ying, Y., Koeva, M., Kuffer, M., Zevenbergen, J., 2020. Urban 3D modelling methods: a state-of-the-art review. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. 43, 699–706.

Yu, C., Lei, J., Peng, B., Shen, H., Huang, Q., 2022. SIEV-net: A structure-information enhanced voxel network for 3D object detection from LiDAR point clouds. IEEE Trans. Geosci. Remote Sens. 60, 1–11.

Yu, W., Zhu, S., Yang, T., Chen, C., 2022. Consistency-based active learning for object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3951–3960.

Zeng, Z., Xu, Y., Xie, Z., Tang, W., Wan, J., Wu, W., 2022. LEARD-net: Semantic segmentation for large-scale point cloud scene. Int. J. Appl. Earth Obs. Geoinf. 112, 102953.

Zhang, D., Lin, X., Yi, B., Wu, Y., 2024. Active learning with core-set sampling and scale-sensitive loss for 3D object detection. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3075–3079.

Zhang, L., Wang, J., Wang, Y., Sun, H., Zhao, X., 2022. Automatic construction site hazard identification integrating construction scene graphs with BERT based domain knowledge. Autom. Constr. 142, 104535.

Zhou, Y., Tuzel, O., 2018. VoxelNet: End-to-end learning for point cloud based 3d object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4490–4499.