



3D object detection and knowledge distillation in autonomous driving: A survey

Weiying Yan^{a,*}, Changhong Bu^a, Yuanyang Zhang^b, Guanghui Yue^c, Wei Zhou^d,
Xiatian Zhu^e, Jungong Han^f

^a School of Computer and Control Engineering, Yantai University, Yantai, 261400, China

^b School of Computer Science and Engineering, Southeast University, Nanjing, 210096, China

^c School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, 518060, China

^d School of Computer Science and Informatics, Cardiff University, Cardiff, CF24 4AG, UK

^e Surrey Institute for People-Centred Artificial Intelligence (PAI), Centre for Vision, Speech and Signal Processing (CVSSP), School of Computer Science and Electronic Engineering, University of Surrey, Guildford, UK

^f Department of Automation, Tsinghua University, China

ARTICLE INFO

Keywords:

Autonomous driving
3D object detection
Multi-modal fusion
Model compression
Knowledge distillation

ABSTRACT

Recent advances in autonomous driving have greatly promoted the development of 3D object detection. Nevertheless, state-of-the-art 3D detectors usually require large-scale networks and heavy computational costs, which severely restrict their deployment on resource-limited onboard platforms. Knowledge distillation (KD), which transfers knowledge from a high-capacity teacher model to a compact student model, has become an effective solution for improving the efficiency of 3D detection while maintaining high accuracy. This paper presents a comprehensive review of knowledge distillation methods for 3D object detection in autonomous driving. We first summarize the mainstream paradigms of 3D object detection, including LiDAR-based, camera-based, and multi-modal fusion frameworks, together with the fundamental categories of knowledge distillation. Then, existing KD-based 3D detection methods are systematically reviewed from three representative perspectives: LiDAR-to-LiDAR, LiDAR-to-camera, and multi-modal-to-single-modal distillation. The characteristics, advantages, and limitations of different distillation strategies are analyzed and compared in detail. Finally, we discuss the major challenges and open issues in current KD-based 3D object detection, and outline several promising directions for future research. This survey aims to provide a concise yet comprehensive reference for the design and deployment of efficient 3D perception systems in autonomous driving. We also provide a continuously updated project homepage to support ongoing research in this area [here](#).

1. Introduction

Autonomous driving has drawn substantial interest from both academia and industry in recent years. Among its core perception modules, 3D object detection [1,2] plays a pivotal role, as illustrated in Fig. 1. By integrating data from multiple sensors, including LiDAR, cameras, radar, GPS, and IMU, autonomous driving systems can accurately estimate the three-dimensional position, size, and orientation of surrounding objects. Specifically, LiDAR provides high-precision 3D point clouds to capture spatial structures, cameras supply rich texture and semantic information to enhance object recognition, and radar excels at detecting moving objects over long distances and under adverse conditions. GPS (Global

Positioning System) [3] offers global positioning to support localization and multi-frame alignment, while IMU (Inertial Measurement Units) [4] provides vehicle pose and motion compensation to improve temporal detection and tracking stability. Through complementary multi-sensor fusion [5,6], the system can achieve robust and high-accuracy 3D object detection in complex scenarios, providing reliable geometric and dynamic cues for downstream tasks such as localization, tracking, motion prediction, and planning.

Driven by the rapid development of deep learning and multi-sensor perception technologies, remarkable progress has been achieved in LiDAR-based [7,8], camera-based [9,10], Radar-based [11,12], and multi-modal fusion [13,14] 3D object detection. Although techniques such as data augmentation [15] can further improve detection accuracy,

* Corresponding author.

E-mail addresses: wqyan@tju.edu.cn (W. Yan), bch88888@s.ytu.edu.cn (C. Bu), zhangyuanyang@seu.edu.cn (Y. Zhang), yueguanghui@szu.edu.cn (G. Yue), zhouwujie@163.com (W. Zhou), xiatian.zhu@surrey.ac.uk (X. Zhu), jungonghan77@gmail.com (J. Han).

<https://doi.org/10.1016/j.inffus.2026.104706>

Received 16 March 2026; Received in revised form 1 July 2026; Accepted 12 August 2026

Available online 14 August 2026

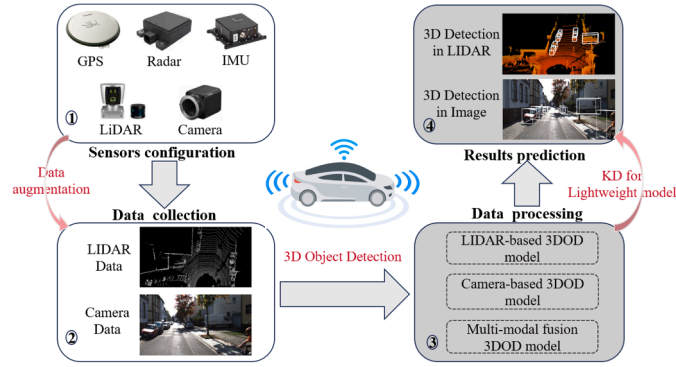


Fig. 1. The process of 3D object detection (3DOD) in autonomous driving scenarios based on different sensors.

the pursuit of higher precision often results in increasingly complex network architectures, consuming substantial computational and memory resources for only marginal performance gains. This significantly limits the feasibility of deploying these models for efficient real-time inference on resource-constrained in-vehicle platforms. As a result, how to achieve an effective trade-off between detection accuracy and computational efficiency has become a fundamental challenge for practical autonomous driving systems. In this context, knowledge distillation [16–18] has emerged as a promising model compression paradigm, enabling lightweight 3D detectors to inherit the representational power of large-scale teacher models while maintaining high efficiency, thereby motivating a comprehensive review of its recent advances in 3D object detection.

Knowledge Distillation (KD), originally introduced by Hinton et al. [19] in 2015 for image classification [20,21], has since developed into a versatile framework for transferring knowledge across models. Beginning in 2017, KD was extended to 2D object detection tasks [22–26], and in 2020, it was first applied to 3D object detection [27]. Since then, an increasing number of KD-based methods have been proposed for 3D detection [27–29], as illustrated in Fig. 2. Researchers have further extended KD to a wide range of emerging detection tasks, including small-object detection [30,31], remote sensing object detection [32,33], incremental object detection [34–36], weakly supervised detection [37,38], and 3D object detection [39,40]. To tackle the unique challenges in these domains, various innovative mechanisms have been proposed, such as incorporating task-specific priors [41,42]; leveraging complementary information across different modalities [43,44]; adopting multi-teacher frameworks [45,46]; and designing self-supervised feature extraction strategies [47,48]. KD has proven particularly effective in promoting the design of lightweight models [29,49], enabling compact student networks to retain the representational power of large teacher models while significantly reducing computational and memory requirements.

Overall, knowledge distillation (KD) has been widely adopted in 3D object detection for designing lightweight models and has gained increasing attention. With advances in algorithms and sensor technologies, KD has emerged as a key paradigm to tackle the inherent challenges of 3D detection in autonomous driving, including limited computational resources, real-time processing, and multi-modal data fusion. Nevertheless, a comprehensive review of this emerging field is still lacking. To address this, this survey, for the first time, systematically presents KD-based 3D detection methods, summarizes their evolution and contributions, and highlights critical challenges and promising directions for future research.

The main contents of this paper are as follows. Section 3 introduces the datasets related to 3D object detection in autonomous driving. In Section 4, we present an overview of the paradigms in 3D object detection, including LiDAR-based 3D object detection (Section 4.1), camera-based object detection (Section 4.2), LiDAR-camera fusion approaches

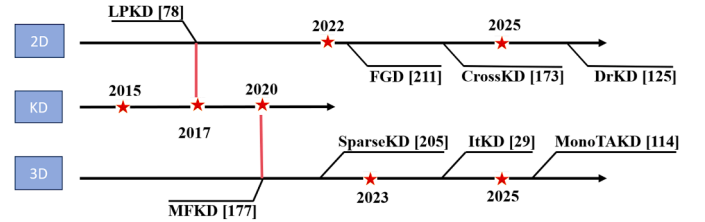


Fig. 2. The temporal intersection of knowledge distillation with 2D and 3D object detection.

(Section 4.3) and end-to-end autonomous driving (Section 4.4). Sections 3 and 4 together provide readers with a comprehensive understanding of 3D object detection before introducing knowledge distillation. Next, in Section 5, we provide a detailed overview of the types of knowledge distillation in 3D object detection. Subsequently, we review and analyze knowledge distillation methods for 3D object detection from three perspectives: LiDAR-to-LiDAR (Section 6), LiDAR-to-Camera (Section 7), and LiDAR-based multi-modal to single-modal (Section 8). Section 9 then offers a comprehensive discussion of the current challenges and prospective research directions in 3D object detection. To better highlight the novelty and unique contributions of this survey, we present a systematic comparison with existing related surveys across several key dimensions, including research scope, taxonomy structure, datasets, inclusion of knowledge distillation, model efficiency and memory consumption, and future research directions, as shown in Table 1. The structural arrangement of this paper is illustrated in Fig. 3.

2. Datasets and metrics

3. Datasets and metrics

In deep learning research, high-quality datasets have always played a foundational role, just as large-scale image datasets such as ImageNet [71], COCO [72], Cityscapes [73], and others [74,75] have significantly advanced the fields of image recognition and understanding. The advancement of 3D object detection technologies similarly relies on larger-scale, more diverse, and more precisely annotated datasets [54–60,64–66,68,70,76], which are essential for addressing the complex and dynamic perception challenges encountered in real-world applications such as autonomous driving. Against this backdrop, this subsection provides a systematic overview of the mainstream benchmark datasets for 3D object detection widely used in the autonomous driving domain, with the aim of offering a solid data foundation and a valuable reference for subsequent methodological innovations and performance evaluation.

Table 1

A comprehensive comparison of autonomous driving surveys across 2DOD (2D Object Detection), 3DOD (3D Object Detection), BEV (Bird's Eye View), robustness, end-to-end systems, VLA (Vision-Language-Action), CP (Collaborative Perception), AD (Autonomous Driving), data Cov. (Datasets coverage), KD (Knowledge Distillation), Eff./ Mem. (Efficiency / Memory Analysis), and F.Work (Future Work).

Title	Year	Scope	Taxonomy	Data Cov.	KD	Eff. / Mem.	F. Work
Multi-modal 3D Object Detection in Autonomous Driving: A Survey and Taxonomy [50]	TIV2023	3DOD	Task-based	Low	No	Yes/No	Yes
3D Object Detection for Autonomous Driving: A Comprehensive Survey [1]	IJCV2023	3DOD,Fusion	Task-based	Extensive	No	Yes/No	Yes
Multi-modal 3d object detection in autonomous driving: a survey [13]	IJCV2023	3DOD, Fusion	Modality-based	Low	No	Yes/No	Yes
When object detection meets knowledge distillation: A survey [51]	TPAMI2023	3DOD,2DOD	System-level	–	Yes	No/No	Yes
Robustness-aware 3d object detection in autonomous driving: A review and outlook [2]	TITS2024	3DOD, Robustness	Task-based	Low	No	No/No	Yes
A survey on vision-language-action models for autonomous driving [52]	ICCV2025	End-to-End AD					
		VLA	Method-based	Low	No	No/No	Yes
Progressive Bird's Eye View Perception for Safety-Critical Autonomous Driving: A Comprehensive Survey [53]	Arxiv2025	BEV, CP					
		Safety-critical AD	System-level	Extensive	No	No/No	Yes
Ours	–	3DOD, BEV End-to-End AD,VLA	System-level	Extensive	Yes	Yes/Yes	Yes

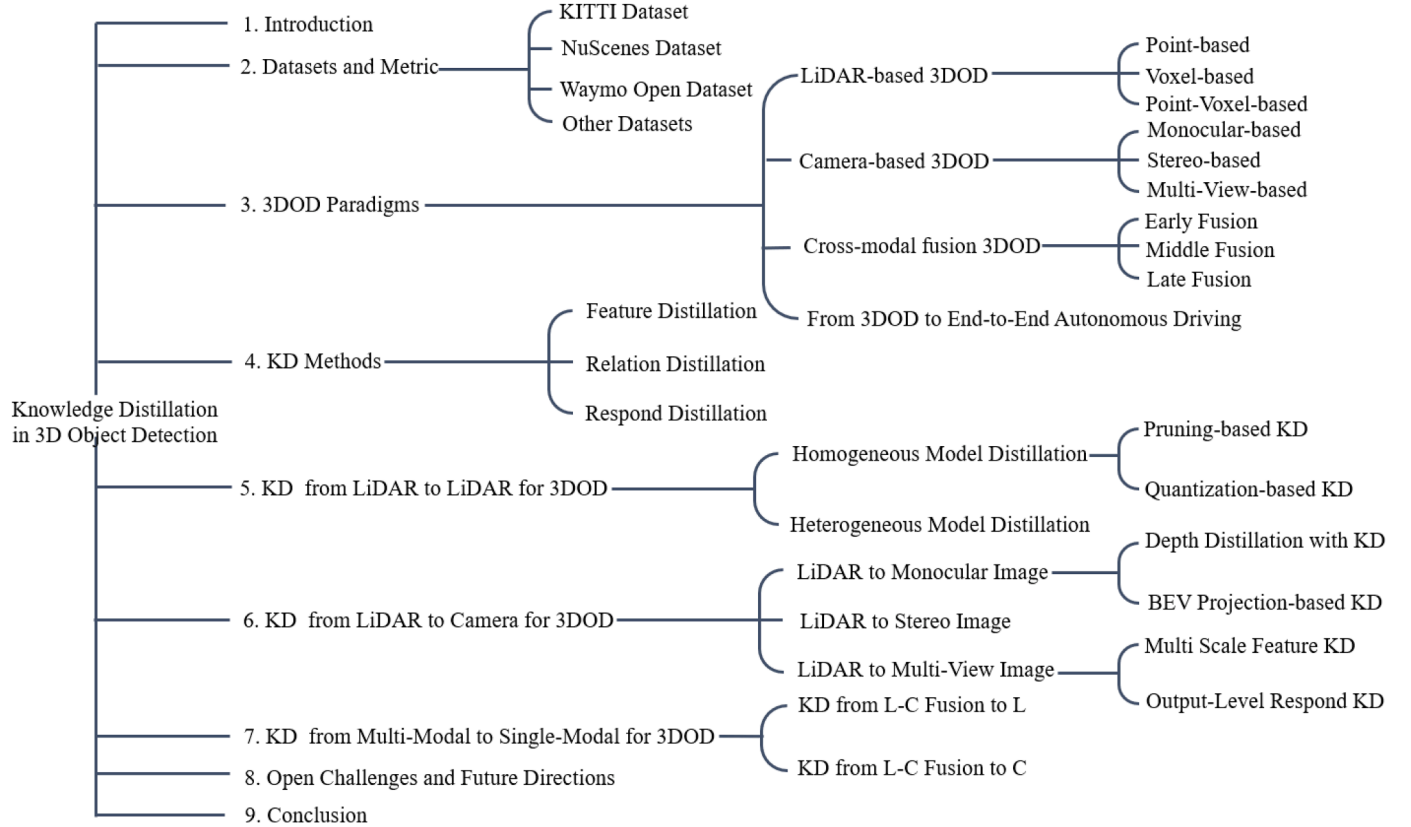


Fig. 3. Structural framework of the paper. ‘3DOD’: 3D Object Detection, ‘L’: LiDAR, ‘C’: Camera.

3.1. KITTI

The KITTI [54] dataset is a milestone multi-modal benchmark in autonomous driving. It was collected using a vehicle equipped with a 64-beam LiDAR, four cameras, and a GPS/IMU system, and provides synchronously captured stereo images. The dataset was collected using professional sensor platforms across more than 20 urban scenarios and contains 7481 training frames and 7518 testing frames. Each frame

is equipped with precise calibration parameters and annotated 2D/3D bounding boxes covering eight object categories, including vehicles and pedestrians. Detection difficulty is further categorized into Easy, Moderate, and Hard levels based on object occlusion and size.

In terms of task design, KITTI defines three sub-tasks: 2D Detection (2D bounding boxes in images), Bird's Eye View (BEV) Detection, and 3D Detection (3D bounding boxes), with each task evaluated using Average Precision (AP). During evaluation, KITTI applies category-specific

Table 2

The widely used datasets for 3D object detection in driving scenarios. 'L', 'C' and 'R' denote LiDAR, Camera and Radar, respectively. We provide a detailed overview of the datasets, including the year of collection, number of LiDAR channels, number of cameras, number of frames, number of scenes, scene duration (s), adverse Weather / Night Scenes, number of object categories, and collection locations.

Dataset	Sensor	Year	LiDAR-C	Camera-N	Frame	Scenes	Duration	Weather Diversity	Classes	Location
KITTI [54]	L + C	2012	64	4	15k	50	–	No/No	8	Germany
ApolloScape [55]	L + C	2018	1	6	20k	53	–	No/No	6	China
H3D [56]	L + C	2019	64	3	27k	160	60–120	No/No	8	USA
PandaSet [57]	L + C	2020	64	6	14k	179	10–15	Yes/Yes	28	USA
KITTI-360 [58]	L + C	2020	64	6	80k	11	–	No/No	19	Germany
A2D2 [59]	L + C	2020	64	6	12.5k	361	–	No/No	38	Germany
OpenLane [60]	L + C	2022	32	1	200k	1000	20	No/No	14	China
V2V4Real [61]	L + C	2023	32	7	20k	67	–	No/No	5	USA
TUMTraf-V2X [62]	L + C	2024	64	5	7k	10	–	No/No	8	Germany
AGC-Drive [63]	L + C	2025	128	5	120k	14	10	No/No	13	China
NuScenes [64]	L + C + R	2019	32	6	40k	1000	20	Yes/Yes	23	USA, Sg
Waymo [65]	L + C + R	2019	64	5	230k	1150	20	Yes/Yes	4	USA
Argoverse [66]	L + C + R	2019	32	9	46k	366	15–20	Yes/Yes	15	USA
Cirrus [67]	L + C + R	2021	128	3	6.2k	–	–	No/No	8	USA
ONCE [68]	L + C + R	2021	40	7	15k	1000	20	Yes/Yes	5	China
ZOD [69]	L + C + R	2023	128	1	100k	1473	20	Yes/Yes	16	Europe
Argoverse2 [70]	L + C + R	2023	32	7	250k	1000	15–30	Yes/Yes	30	USA

IoU thresholds to determine predictions: a predicted box is considered a True Positive (TP) if its IoU with a ground-truth (GT) box exceeds the threshold and the GT has not yet been matched; otherwise, it is counted as a False Positive (FP), while undetected GTs are counted as False Negatives (FN). This strict evaluation protocol establishes a standardized benchmark for comparing subsequent 3D detection methods. Precision and recall are defined as follows:

$$precision = \frac{TP}{TP + FP}, \quad (1)$$

$$recall = \frac{TP}{TP + FN}, \quad (2)$$

$$AP = \int_0^1 p(r) dr. \quad (3)$$

3.2. NuScenes

The NuScenes [64] dataset, released by Motional (formerly nuTonomy), is a large-scale multi-modal benchmark for autonomous driving, collected in Boston and Singapore under diverse urban and weather conditions. NuScenes is currently one of the most widely used dataset in autonomous drive. It contains 1000 driving scenes of 20 seconds each, with keyframes at 2 Hz, and integrates data from six surround-view cameras, one 32-beam LiDAR, five millimeter-wave radars, and IMU/GPS, enabling comprehensive multi-sensor perception.

NuScenes provides 3D bounding box and attribute annotations for 23 object classes. Unlike KITTI, it evaluates detection performance using both accuracy and multiple error metrics, including mATE (mean Average Translation Error), mAOE (mean Average Orientation Error), mASE (mean Average Scale Error), mAVE (mean Average Velocity Error), and mAAE (mean Average Attribute Error), which are further combined into the NDS (NuScenes Detection Score). Mean Average Precision (mAP) remains the fundamental metric for quantifying the trade-off between detection precision and recall, and, together with multi-dimensional error measures, forms a more comprehensive evaluation framework for assessing 3D object detection performance. Owing to its rich annotations and rigorous evaluation protocol, NuScenes has become a widely used benchmark for 3D object detection and multi-modal perception research.

$$mAP = \frac{1}{N_{classes}} \sum_{c=1}^{N_{classes}} \frac{1}{N_{dist}} \sum_{d=1}^{N_{dist}} AP(c, d), \quad (4)$$

$$NDS = \frac{1}{10} \left[5mAP + \sum_{mTP \in TP} (1 - \min(1, mTP)) \right], \quad (5)$$

where $N_{classes}$ denotes the twenty-three evaluation object classes and N_{dist} represents the four distance thresholds used to assess the average precision under different levels of strictness.

3.3. Waymo open dataset

The Waymo Open Dataset [65], collected and annotated by Waymo autonomous vehicles, is a high-quality, large-scale multi-modal dataset covering diverse driving scenarios from urban to suburban environments. It contains 798 training scenes and 202 validation scenes, each lasting 20 seconds, captured using five LiDAR sensors and five pinhole cameras. The dataset provides 2D and 3D bounding box annotations for four object categories: vehicles, pedestrians, cyclists, and traffic signs.

In terms of evaluation, Waymo adopts Average Precision (AP) as the core metric, similar to KITTI, but further introduces APH (Average Precision weighted by Heading), which accounts for heading prediction accuracy, providing a more comprehensive assessment of detection performance. In addition, Waymo supports research on domain adaptation, enabling knowledge transfer from source domains to target domains where annotations are limited or unavailable, thereby alleviating the challenge of insufficient large-scale labeling. Owing to its scale, sensor configuration, and evaluation framework, the Waymo Open Dataset is among the largest and most influential benchmarks for 3D object detection research.

3.4. Other datasets

In addition to the aforementioned mainstream datasets, a number of emerging autonomous driving datasets [55–62,66,68,70,76] have recently gained recognition in the research community. For example, Argoverse [66] provides rich annotations, including 3D bounding boxes, object trajectories, and map-based contextual information, along with multi-modal sensor data; it covers diverse urban driving scenarios with varying weather, lighting, and traffic conditions, providing a challenging and realistic environment for autonomous driving research. ONCE [68] is a large-scale multi-modal perception dataset designed for autonomous driving research, covering over 100 hours of real-world driving scenarios across urban roads, highways, and complex traffic environments, with diverse traffic conditions and weather. PandaSet [57], developed for autonomous driving and robotic perception research, is a multi-modal open 3D dataset collected with the Panda automotive platform. It covers diverse scenarios including urban roads, highways, and suburban areas, covering approximately 100 km of driving trajectories.

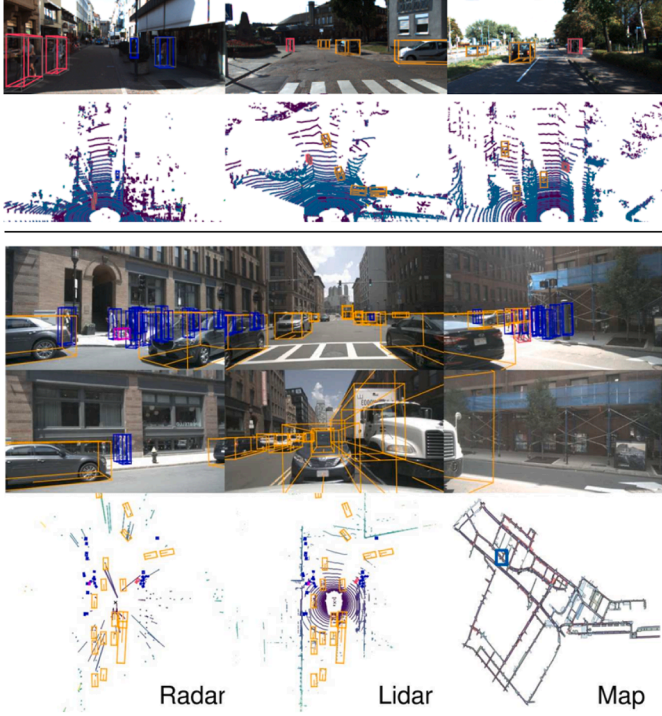


Fig. 4. Examples of image-point cloud pairs and ground-truth 3D bounding boxes in KITTI [54] and nuScenes [64].

TUMTraF-V2X [62] is a cooperative perception dataset from the Technical University of Munich, supporting 3D object detection and tracking in V2X scenarios. It includes synchronized multi-sensor data from vehicle-mounted and roadside LiDAR, cameras, GPS, and IMU, with over 30,000 annotated 3D bounding boxes across diverse traffic conditions, enabling evaluation of cooperative perception algorithms for autonomous driving. OpenLane [60] provides multi-view camera data with accurate 3D lane annotations under diverse driving conditions. Owing to its realistic scenes and high-quality annotations, it has become an important benchmark for 3D lane perception in autonomous driving. To illustrate the differences in information captured by various sensors, Fig. 4 presents the monocular images and corresponding LiDAR point clouds from the KITTI dataset, multi-view images and point clouds from NuScenes, radar data, as well as high-definition (HD) map information. These datasets jointly promote the development of autonomous driving. Table 2 summarizes the currently available open-source autonomous driving datasets along with their key attributes.

3.5. Safety and system-level evaluation metrics

Beyond conventional detection accuracy metrics such as AP, mAP, and NDS, evaluating autonomous driving systems in real-world scenarios requires a more comprehensive set of metrics that capture system-level safety, robustness, and deployment efficiency. First, robustness-related metrics [77] are essential for assessing model performance under challenging conditions, including sensor corruption, adverse weather, and domain shifts. These metrics evaluate the stability of perception systems when inputs are degraded due to noise, occlusion, or environmental variations, as well as their generalization ability under sim-to-real transfer settings. Second, for multi-agent and infrastructure-assisted scenarios, V2X cooperative perception gain has become an important metric, measuring the performance improvement achieved through information sharing among vehicles and roadside units. In addition, perception-to-planning consistency is critical for evaluating whether perception outputs can reliably support downstream decision-making in closed-loop driving systems. Third, system-level efficiency

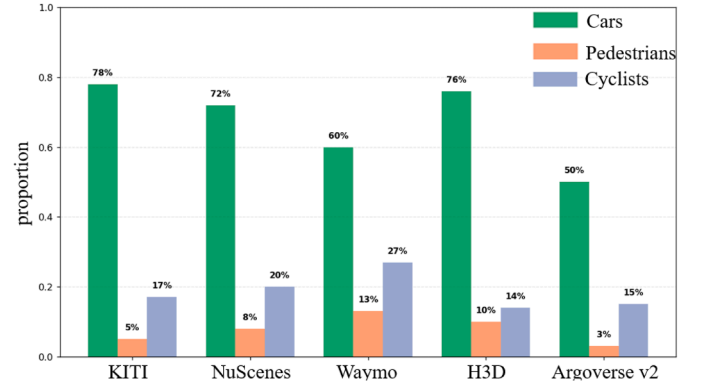


Fig. 5. Development of the proportion of the main categories in autonomous driving datasets.

metrics such as latency, frames per second (FPS), memory consumption, and energy cost are crucial for real-world deployment, especially in resource-constrained onboard systems. These metrics directly reflect the practicality of deploying large-scale perception models in real-time autonomous driving applications. Finally, safety-critical evaluation [78], including failure case analysis and closed-loop driving performance, is increasingly important to ensure that perception errors do not propagate into unsafe planning behaviors. Overall, these metrics provide a more holistic evaluation framework that goes beyond detection accuracy and better reflects real-world autonomous driving requirements.

3.6. Discussion

The development of autonomous driving datasets exhibits three notable characteristics. First, dataset sizes have grown rapidly, with the number of samples in mainstream benchmarks rising from approximately 15,000 frames [54] to over 230,000 frames [65]. Second, 3D datasets provide far fewer annotated instances compared to 2D datasets. For example, the 2D dataset ImageNet [71] contains 1.4 million annotated images, whereas widely used 3D datasets such as ONCE [68] and Argoverse [66] contain only 417k and 993k annotations, respectively, highlighting a substantial gap between 2D and 3D data resources. Third, data collection is increasingly oriented toward multi-modal [76], high-density [63], and long-sequence recordings [69]; under such data characteristics, the spatiotemporal alignment [79] across different modalities becomes more challenging, and the substantial redundancy in continuous scenes further complicates the understanding of complex driving environments.

Moreover, the object classes in these datasets are limited and often imbalanced. As shown in Fig. 5, the proportions of vehicles, pedestrians, and cyclists vary substantially across datasets. Small-object categories are typically underrepresented, whereas large-object categories, such as cars, constitute the majority of annotations. Consequently, current 3D object detection models tend to exhibit suboptimal performance on small objects, largely due to these inherent dataset limitations.

4. 3D object detection paradigms

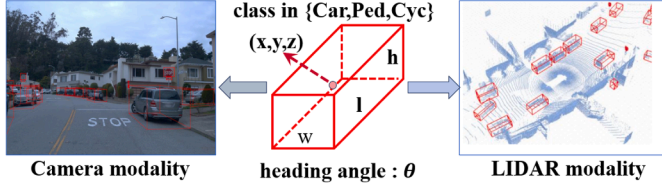
The core task of 3D object detection can be formally defined as follows: given a sensor or multi-modal sensor input stream S from an autonomous driving scenario—typically comprising LiDAR point clouds, RGB images, or other complementary modalities—the goal is to learn a mapping function $f_{det} : S \rightarrow B$, where B denotes the set of predicted 3D bounding boxes along with their corresponding semantic class labels [80,81].

$$B = f_{det}(S), \quad (6)$$

Table 3

LiDAR sensor vs. Camera sensor. "FOV" stands for field of view, and "FPS" represents frames per second.

Type	LiDAR/Image	Format	FOV	FPS	Cost	Advantage	Disadvantage
Velodyne HDL-64		binary float matrix	360°	10Hz	-	Precise 3D geometric information	Sensitive to bad weather
Color camera		png format	60° – 90°	15Hz	\$255	low cost Rich semantic information	Lack of depth information
Gray camera		png format	60° – 90°	20Hz	\$134	lowest cost High frame rate	Lack of depth information

**Fig. 6.** 3D detection results in autonomous driving. The camera and LiDAR samples are from [1].

where $B = \{x, y, z, l, w, h, \theta, class\}$ is the collection of 3D objects in N scenarios, f_{det} is a 3D object detection model, $\{x, y, z\}$ is the 3D center coordinate of the predicted detection box. $\{l, w, h\}$ denote the length, width, and height of the cuboid, respectively, while θ represents the heading angle of the cuboid on the ground plane. The variable *class* indicates the category of the 3D object, such as car, pedestrian, or cyclist. As shown in Fig. 6, sensor inputs can be broadly categorized into LiDAR point clouds and images. Table 3 provides a detailed comparison between LiDAR sensor and Camera sensor. Notably, due to the large number of LiDAR sensor brands—such as Velodyne, Ouster, Hesai, RoboSense, and Livox—their performance and prices vary significantly; therefore, Only the price of the camera sensor is reported, which is sourced from [13]. In this section, we provide a brief overview of commonly used algorithms [82–87] corresponding to these two types of sensors.

4.1. LiDAR-based 3D object detection

LiDAR-based 3D object detection [88–93] has long been a key research focus in autonomous driving. Below, we briefly review existing LiDAR-based 3D detection methods, including point-based (Section 4.1.1), voxel-based (Section 4.1.2), and point-voxel fusion approaches (Section 4.1.3). A chronological summary of these methods is illustrated in Fig. 7.

4.1.1. Point-based 3D object detection

Traditional point-based methods generally consist of three main stages: sampling, feature learning, and final prediction, as illustrated in Fig. 8(a). The sampling stage aims to select representative subsets of points from raw point clouds, with commonly used strategies including farthest point sampling (FPS) [114], random sampling (RS) [93], density-based sampling (DS) [118], and voxel-based sampling (VS) [7]. The sampled points are subsequently fed into a feature learning network—such as multi-layer perceptrons (MLPs) [119], transformers [120], or graph convolutional networks (Graph CNNs) [121]—to extract discriminative feature representations. Finally, a prediction head [108] is employed to perform tasks such as 3D object detection and object localization.

Notable examples include PointNet [80], which first enabled direct deep learning modeling of unordered point cloud data; its enhanced version, PointNet++ [119], effectively captures local geometric structures through hierarchical feature extraction. SFD [122] converts sparse point

clouds into denser feature representations, achieving efficient and accurate 3D detection. Pointformer [97] innovatively introduces the Transformer mechanism, providing long-range dependency modeling capabilities for point cloud processing. 3DSSD [95] designs a fusion-based sampling strategy that combines feature and geometric space, better preserving representative points while avoiding interference from excessive background points. Point-GNN [96] models point clouds as graph structures and employs Graph Neural Networks (GNNs) to enhance feature, addressing the limitations of traditional point cloud methods in local neighborhood modeling.

Due to the need to process a large amount of unordered point cloud data, point-based methods typically require substantial computational resources, which poses significant challenges when handling large-scale datasets.

4.1.2. Voxel-based 3D object detection

To reduce computational cost and memory consumption while enabling efficient large-scale 3D scene processing, voxel-based 3D object detection methods provide a structured representation for feature extraction and bounding box prediction by discretizing unordered point clouds into regular 3D voxel grids, as shown in Fig. 8(b). Two common types of voxel-based methods are voxel grid-based [89,102] and pillar grid-based [123,124] approaches. After voxelization, the encoded voxels are fed into feature learning networks [116] for representation extraction, and the learned features are typically projected onto the bird's-eye view (BEV) [89,91] plane for final prediction.

The pioneering work VoxelNet [125] first enabled end-to-end voxel feature learning, significantly improving detection performance. Subsequently, PointPillars [123] introduced a pillar-based voxel representation, converting 3D convolutions into efficient 2D convolutions and greatly accelerating inference. SECOND [88] effectively reduced the computational cost of conventional 3D convolutions by employing sparse convolutions. To further optimize both accuracy and efficiency, Voxel R-CNN [7] designed a voxel RoI pooling operation, which substantially enhanced detection precision while maintaining real-time performance. In recent years, driven by breakthroughs in Transformers for visual tasks and the demand for lightweight designs, several novel approaches have emerged. CenterPoint [102] represents objects as center points (keypoints) rather than traditional boxes or voxel sets, simplifying modeling complexity and improving efficiency. FocalFormer3D [126] incorporates the identification and mining of hard-to-detect samples into a multi-stage detection training process, enhancing the model's attention to challenging instances and recall.

Voxel-based methods effectively reduce computational cost while enhancing the ability to model local point cloud structures, and have become the mainstream approach in 3D object detection.

4.1.3. Point-voxel-based 3D object detection

Hybrid point-voxel-based [8,94,114–117] 3D object detection achieves significant performance improvements by leveraging the complementary properties of point clouds and voxels, the framework is illustrated in Fig. 8(c).

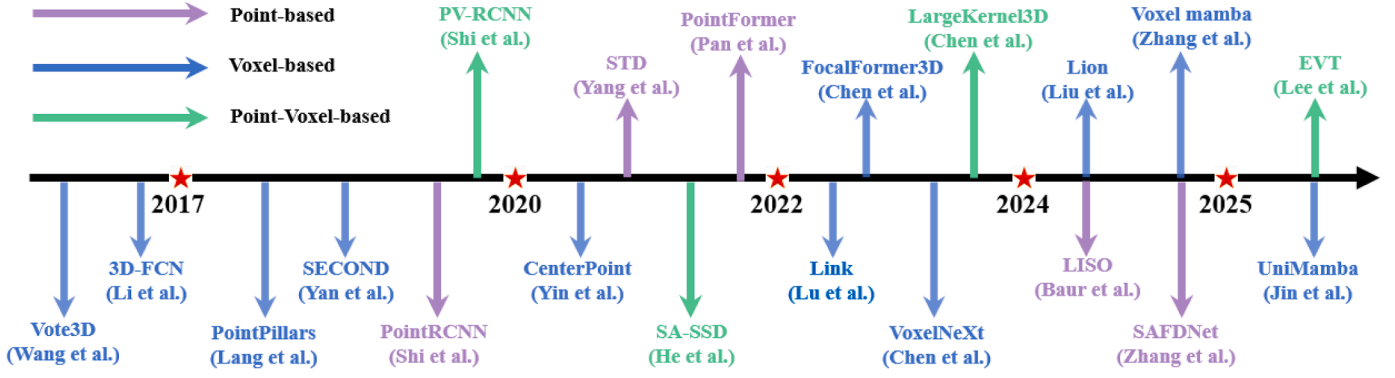


Fig. 7. Chronological overview of the LiDAR-based 3D object detection methods.

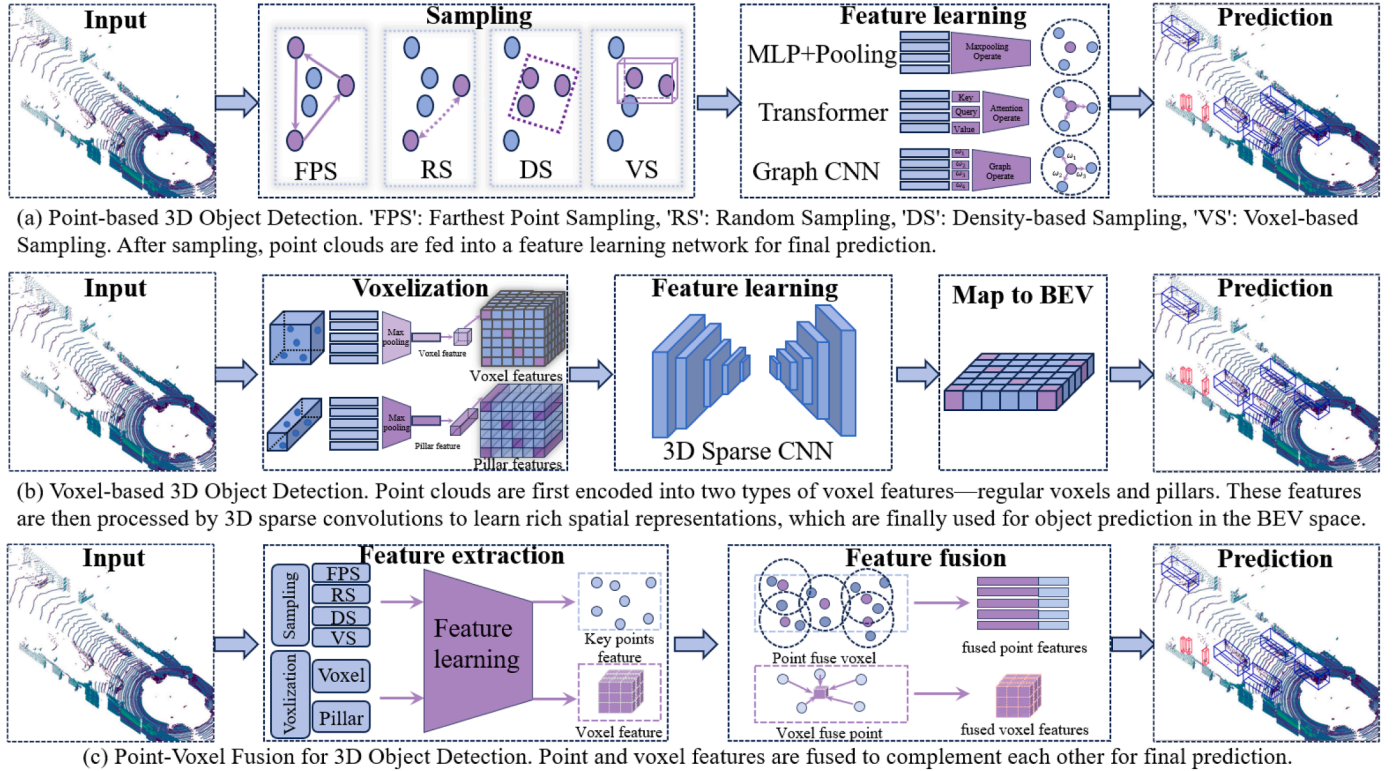


Fig. 8. LiDAR-based 3D detection architectures include point-based, voxel-based, and point-voxel fusion-based methods.

For example, STD [94] proposes an innovative two-stage detection framework: the first stage generates initial candidate boxes from point features, while the second stage refines these boxes using voxel features. PV-RCNN [8] combines the computational efficiency of voxel-based CNNs with the geometric fidelity of point features through a carefully designed multi-scale feature fusion mechanism, significantly improving detection performance. SA-SSD [113] introduces an efficient deformable operation that aligns the classification confidence with the predicted bounding boxes, enhancing the fusion efficiency and thereby improving detection accuracy. VP-Net [116] treats voxel representations as points and effectively integrates the structural information of voxels with point-level features, enabling efficient fusion of point cloud and voxel features. This allows the network to better capture fine-grained spatial information. Although point-voxel fusion methods achieve superior performance, they often suffer from high training costs and strong dependencies, which has led to a decreasing number of studies in this direction. LargeKernel3D [104] decomposes large convolutional kernels into multiple small spatial regions with shared

weights, reducing parameter count. Unlike conventional channel-wise weight sharing, spatial weight sharing better captures spatial relationships. During inference, these spatially partitioned convolutions are recombined into standard convolutions, maintaining computational efficiency. Point-Voxel-based 3D detection combines point-level geometric detail with voxel-based structural efficiency, achieving a balance between accuracy and efficiency. However, it introduces additional computational complexity and potential quantization errors due to voxelization.

4.1.4. Analysis

LiDAR-based 3D object detection methods can be categorized into point-based, voxel-based, and point-voxel-based approaches. Point-based methods preserve fine geometric details and achieve high accuracy but suffer from high computational cost and limited scalability due to irregular data processing. Voxel-based methods improve efficiency through structured representation and convolutional operations, but introduce quantization errors and lose fine spatial information. Point-

Table 4

A comprehensive performance analysis of LiDAR-based 3D object detection methods across multiple datasets. We report the inference time as originally reported in the respective papers, along with $AP|_{R_{0.05}}$ (%) for 3D car detection on the KITTI test benchmark, mAP (%) and NDS scores on the NuScenes test set, and Level 1 (L1) and Level 2 (L2) mAP on the Waymo validation set. We sort these methods by publication year and group them according to the type of sensor input.

Method	Type	Year	Inference	NuScenes		Waymo Vehicle		Contribution	L1	L2	
				Easy	Mod.	Hard	mAp				
PointRCNN [92]	Point	CVPR2019	100	85.94	75.76	68.32	–	–	–	–	3D proposals from point
STD [94]	Point	ICCV2019	80	86.61	77.63	76.06	–	–	–	–	Proposal Generation
3D-SSD [95]	Point	CVPR2020	38	88.36	79.57	74.55	42.6	56.4	–	–	Single stage frame
Point-GNN [96]	Point	CVPR2020	640	88.33	79.47	72.29	–	–	–	–	Introduce graph structure
Pointformer [97]	Point	CVPR2021	–	87.13	77.06	69.25	53.6	–	–	–	Take transformer into 3D
IA-SSD [98]	Point	CVPR2022	12	88.34	80.13	75.04	–	–	–	–	Not all points are equal
SVGA-Net [99]	Point	AAAI2022	–	87.33	80.47	75.91	–	–	–	–	Spherical Voxel Grouping
MSF [100]	Point	CVPR2023	8	–	–	–	–	–	–	–	Motion-guided Sequential Fusion
Ptt [101]	Point	CVPR2024	64	–	–	–	–	–	81.32	75.71	Trajectory with point
SECOND [88]	Voxel	Sensors2018	50	83.13	73.66	66.20	–	–	–	–	3D sparse convolution
CenterPoint [102]	Voxel	CVPR2021	70	–	–	–	58.0	65.5	76.7	68.8	Center-based framework
Voxel R-CNN [7]	Voxel	AAAI2021	40	90.90	81.62	77.06	–	–	75.5	66.5	Voxel RoI Pooling
PillarNet [103]	Voxel	ECCV2022	–	–	–	–	–	–	74.2	65.5	Efficient voxel feature encoder
VoxelNeXt [89]	Voxel	CVPR2023	66	89.10	80.35	76.99	66.0	71.4	83.2	76.0	Completely sparse voxel processing
LargeKernel3D [104]	Voxel	CVPR2023	14	–	–	–	–	–	78.0	69.8	Spatial-wise Partition Convolution
SAFDNet [105]	Voxel	CVPR2024	–	–	–	–	66.3	71.0	81.9	76.5	Adaptive Feature Diffusion
SEED [106]	Voxel	ECCV2024	–	–	–	–	66.2	71.2	79.7	71.8	Deformable Grid Attention
VoxelMamba [91]	Voxel	NeIPS2024	12	–	–	–	67.5	71.9	80.8	72.6	Group-Free serialization strategy
LION [107]	Voxel	NeIPS2024	–	–	–	–	68.0	72.1	80.3	72.0	Linear RNN
Re-VoxelDet [108]	Voxel	WACV2024	–	–	–	–	64.6	70.5	78.7	73.3	Multi-view Voxel Backbone
SAFDNet [105]	Voxel	CVPR2024	–	–	–	–	68.3	72.3	81.9	76.5	Fully Sparse Architecture
SparseDet [109]	Voxel	TGRS2024	–	90.79	81.17	78.11	66.7	71.9	–	–	End-to-End Sparse Framework
LISO [110]	Voxel	ECCV2024	–	–	–	–	–	–	–	–	Self-supervised 3D detection framework
UniMamba [90]	Voxel	CVPR2025	–	–	–	–	68.5	72.6	80.6	72.2	Voxel serialization
PV-RCNN [8]	P. + V.	CVPR2020	80	90.25	81.43	76.82	–	–	77.5	68.9	Combine point with voxel
SPVNAS [111]	P. + V.	ECCV2020	–	87.81	78.43	74.85	–	–	–	–	Sparse voxels + lightweight point branches
CT3D [112]	P. + V.	ICCV2021	–	87.83	81.77	77.16	–	–	76.3	69.0	Channel-Wise Decoding Module
SA-SSD [113]	P. + V.	CVPR2020	40	88.75	79.79	74.16	–	–	–	–	Structure-Aware Single-Stage detection
PV-RCNN + + [114]	P. + V.	IJCV2023	–	90.14	81.88	77.15	–	–	79.2	70.6	Sectorized Proposal-Centric Sampling
APVR [115]	P. + V.	TAI2023	–	91.45	82.17	78.08	58.6	65.9	–	–	Lightweight Point-Voxel characterization
VP-Net [116]	P. + V.	TGRS2023	59	90.46	82.03	79.65	–	–	–	–	Internal association modeling
Pvtransformer [117]	P. + V.	ICRA2023	–	–	–	–	–	–	84.8	77.3	Transformer in pooling

voxel-based methods balance accuracy and efficiency by combining both representations, while increasing model complexity. Overall, these methods reflect a trade-off between representation fidelity and computational efficiency. Table 4 provides a comprehensive summary of existing representative LiDAR-based 3D detection methods. For each method, we also highlight its most notable contribution.

4.2. Camera-based 3D object detection

Cameras are among the most widely used sensors in autonomous vehicles. With the rapid advancement of 2D object detection techniques, researchers have progressively extended these methods to 3D perception, leading to a variety of image-based approaches [82,127,129,132,133,137] for 3D object detection. As a result, camera-only 3D detection can achieve competitive performance at a relatively low cost, benefiting from high-resolution images and advanced neural network architectures. In some certain tasks [86,143,147], it can even outperform LiDAR-based methods. In the following, we explore 3D object detection algorithms under different camera configurations, including monocular-based (Section 4.2.1), stereo-based (Section 4.2.2), and multi-view-based (Section 4.2.3) methods. Meanwhile, Fig. 9 presents a chronolog-

ical overview of representative camera-based 3D object detection methods.

4.2.1. Monocular-based 3D object detection

Monocular cameras provide dense semantic information in the form of pixel intensities, capturing rich shape and texture cues within a scene. Such visual cues can be further leveraged for 3D object detection. Although monocular systems inherently lack explicit depth information, they are inexpensive, easy to deploy, and have become indispensable sensors in autonomous driving. Consequently, an increasing number of studies have focused on advancing monocular 3D object detection [82,83,127–130,132–134] in recent years.

For instance, Mono3D [148] leverages geometric priors, especially ground truth information, to generate class-specific 3D candidate regions. These candidates are then evaluated using semantic and positional features and subsequently fed into a CNN for classification and bounding box regression. Mono3D + + [149] utilizes a deformable wireframe frame to precisely characterize vehicle shape and pose. Simultaneously, a 3D bounding box model is constructed as a coarse shape representation to improve robustness against unstable or inaccurate 2D features, achieving a balance between precision and robustness.

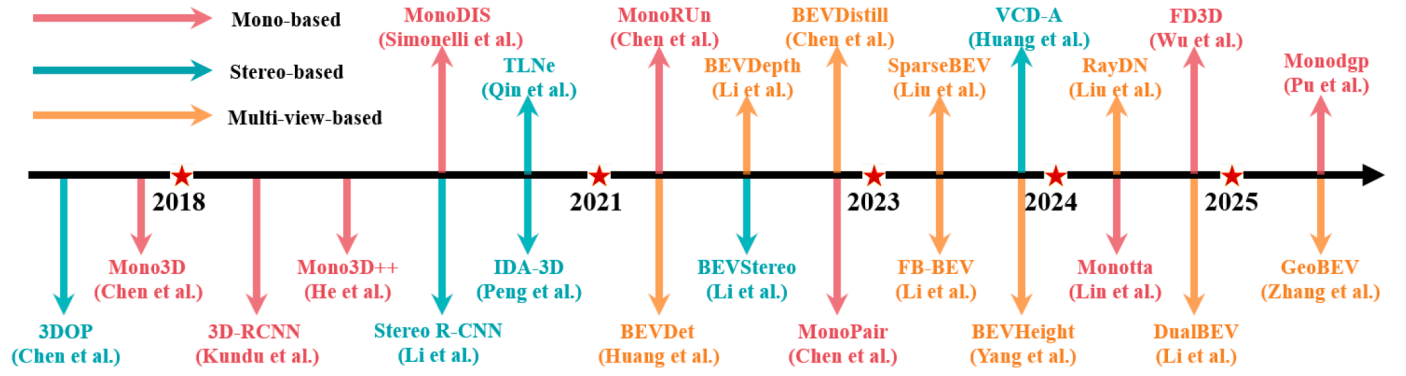


Fig. 9. Chronological overview of the Camera-based 3D object detection methods.

Table 5

A comprehensive performance analysis of camera-based 3D object detection methods across different datasets. The inference time and performance accuracy as originally reported in the respective paper.

Method	Type	Year	Inference	NuScenes				
				Easy	Mod.	Hard	mAP	NDS
MonoDIS [127]	Mon.	ICCV2019	100	10.3	7.9	6.4	30.4	38.4
MonoPair [82]	Mon.	CVPR2020	60	13.0	9.9	8.6	–	–
PatchNet [128]	Mon.	ECCV2020	400	15.6	11.1	10.1	–	–
MonoDTR [129]	Mon.	CVPR2022	37	21.9	15.3	12.7	–	–
MonoDETR [130]	Mon.	ICCV2023	38	25.0	16.4	13.5	–	–
OBMO [131]	Mon.	TIP2023	–	22.7	15.7	13.2	–	–
MonoCD [132]	Mon.	CVPR2024	36	25.5	16.5	14.5	–	–
Monowad [133]	Mon.	ECCV2024	–	28.1	20.3	16.9	–	–
FD3D [83]	Mon.	AAAI2024	40	25.3	17.1	14.5	43.1	48.5
MonoDGP [134]	Mon.	CVPR2025	–	26.3	18.7	15.9	–	–
MonoTAKD [28]	Mon.	CVPR2025	–	27.9	19.4	16.5	–	–
Stereo R-CNN [135]	Ste.	CVPR2019	420	47.5	30.2	23.7	–	–
OC-Stereo [136]	Ste.	ICRA2020	350	55.1	37.6	30.2	–	–
LIGA-Stereo [84]	Ste.	ICCV2021	350	81.3	64.6	57.2	–	–
SNVC [137]	Ste.	AAAI2022	–	78.5	61.3	54.2	–	–
StereoDistill [138]	Ste.	AAAI2023	–	81.6	66.3	57.3	–	–
DETR3D [139]	M.V.	PMLR2022	–	–	–	–	41.2	47.9
BEVDepth [140]	M.V.	AAAI2023	–	–	–	–	52.0	60.9
BEVStereo [141]	M.V.	AAAI2023	–	–	–	–	52.5	61.0
FB-BEV [85]	M.V.	CVPR2023	225	–	–	–	53.7	62.4
BEVNeXt [142]	M.V.	CVPR2024	227	–	–	–	55.7	64.2
RayDN [143]	M.V.	ECCV2024	–	–	–	–	56.5	64.5
DualBEV [86]	M.V.	ECCV2024	–	–	–	–	55.2	63.4
G-DETR4D [144]	M.V.	TIP2024	–	–	–	–	54.1	62.0
CorrBEV [145]	M.V.	CVPR2025	–	–	–	–	56.2	64.0
DySS [87]	M.V.	CVPR2025	–	–	–	–	57.4	65.3
GeoBEV [146]	M.V.	AAAI2025	–	–	–	–	57.9	66.2

MonoDIS [127] decouples different parameter groups in 2D and 3D detection losses, thereby avoiding interference between multiple regression tasks and complex weight tuning, which enhances training stability and convergence speed. MonoPair [82] introduces pairwise spatial relationships, leveraging adjacent objects to assist the detection of occluded targets and thereby improving robustness. Monotta [150] proposes a novel dynamic threshold mechanism to adaptively screen trusted targets, maintaining the flexibility and robustness of the adaptation process. However, it is fundamentally limited by depth ambiguity and lack of explicit geometric cues, resulting in lower accuracy and weaker robustness compared to LiDAR-based methods, especially in complex or occluded environments.

4.2.2. Stereo-based 3D object detection

Compared with monocular cameras, stereo cameras provide more accurate depth estimation, thereby enabling higher detection accuracy. However, this advantage comes at the cost of substantially increased computational complexity and system overhead. Despite these chal-

lenges, some studies continue to explore stereo-based 3D object detection [84,136–138,151], aiming to balance depth accuracy with efficiency and practicality.

For example, 3DOP [81] formulates 3D proposal generation as an energy minimization problem, producing diverse and accurate candidate boxes while effectively handling occlusions and scale variations. Stereo R-CNN [135] enhances proposal quality by linking multi-scale features from the left and right images and aligning RoIs across views, enabling precise 3D box reconstruction. TLNet [152] establishes geometric correspondences between stereo image regions through 3D anchors, allowing efficient 3D localization without relying on dense depth maps. IDA-3D [151] further simplifies the stereo-based pipeline by predicting the depth of 3D box centers in an end-to-end manner using only RGB inputs and 3D labels. BEVStereo [141] improves depth estimation for moving objects via dynamically selected matching intervals and a temporal-stereo module. Finally, VCD-A [153] incorporates LiDAR as a depth prior and adopts a trajectory distillation strategy to refine object motion, benefiting both static and dynamic target detection. Cur-

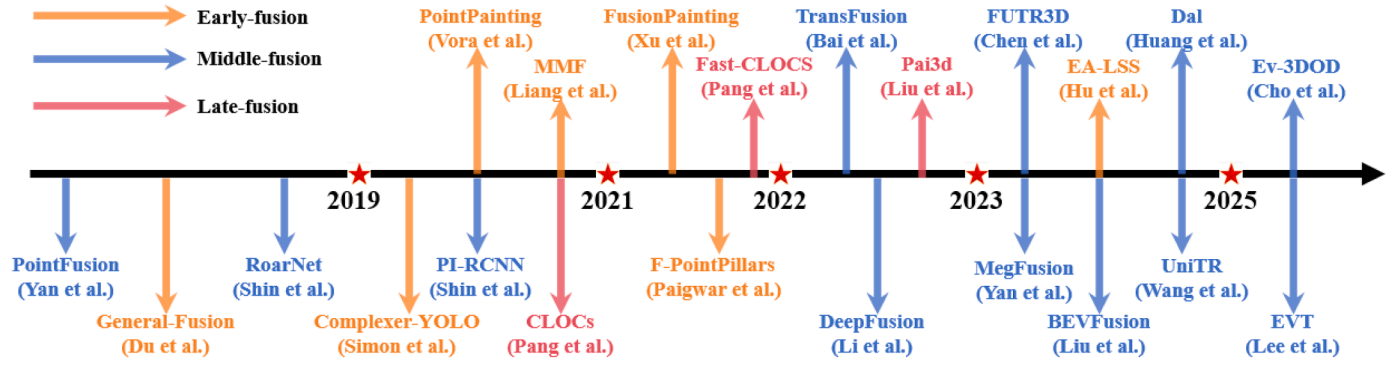


Fig. 10. Chronological overview of the LiDAR-camera fusion 3D object detection methods.

rently, research on stereo-based 3D object detection remains relatively limited. And although it incurs substantially increased computational cost, its performance still falls behind that of multi-view-based 3D detection methods.

4.2.3. Multi-view-based 3D object detection

Multi-view data, arranged in a surround-view configuration, provide reliable depth information, enabling multi-view-based methods to achieve significantly superior performance compared with other approaches. Significantly, LSS [9] introduces a pioneering “Lift-Splat-Shoot” paradigm to address BEV perception tasks using multi-view cameras. This approach consists of three core steps: Lift—performing depth prediction at the pixel level to project multi-view image features into a 3D frustum with discretized depth bins; Splat—unprojecting the 3D frustum into a unified bird’s-eye view (BEV) plane, thereby achieving an end-to-end transformation from image features to BEV features; and Shoot—conducting downstream perception tasks directly on the generated BEV feature map. This paradigm has since become the foundation for numerous subsequent studies.

For example, BEVDet [10] adopts the LSS paradigm, incorporating task-specific data augmentation and optimized NMS to enhance generalization and detection accuracy. BEVDepth [140] leverages LiDAR-provided ground-truth depth to supervise the depth prediction module, improving depth estimation. BEVDistill [154] addresses the challenge of cross-modal feature transfer by unifying LiDAR and camera features in the BEV space and employing a teacher-student framework for high-precision localization. SparseBEV [147] introduces an end-to-end sparse query detector with single-stage decoding and self-attention for efficient multi-scale feature aggregation, eliminating the need for dense BEV maps. FB-BEV [85] combines forward and backward projection to lift 2D features to 3D and refine BEV features via depth, enhancing feature density and accuracy. BEVHeight [155] regresses pixels to ground height rather than depth, achieving distance-independent detection and simplifying optimization. DualBEV [86] proposes a unified 2D–3D feature transformation framework with probabilistic correspondence metrics to improve efficiency and accuracy. GeoBEV [87] enhances BEV fidelity through radial–Cartesian sampling, improving spatial resolution and geometric structure preservation.

Multi-view-based 3D object detection [85,87,139,141–143,146] has emerged as a major research direction. By exploiting geometric constraints across different views, these methods effectively reconstruct the spatial structure and achieve detection accuracy comparable to LiDAR-based approaches.

4.2.4. Analysis

In summary, Camera-based 3D object detection can be categorized into monocular, stereo, and multi-view methods. Monocular approaches are cost-efficient but suffer from depth ambiguity and limited robustness. Stereo methods improve depth estimation using binocular geometry, yet rely on accurate calibration and degrade in challenging vi-

sual conditions. Multi-view methods provide better spatial coverage and higher accuracy but incur higher computational cost. Overall, these methods reflect a trade-off between cost, accuracy, and efficiency. Table 5 systematically summarizes the progress of these methods across different datasets.

4.3. LiDAR-camera fusion 3D object detection

LiDAR–camera fusion for multi-modal 3D object detection has become a key component of modern autonomous driving perception systems. LiDAR point clouds provide precise 3D geometric information, enabling accurate estimation of object size, position, and orientation. In contrast, camera images offer rich color, texture, and semantic cues, which effectively complement LiDAR data—particularly in scenarios with sparse point clouds or geometrically similar objects. By integrating these complementary modalities, fusion-based methods [15,157–163] significantly enhance detection accuracy and robustness. Fusion strategies are typically categorized into early fusion (Section 4.3.1), middle fusion (Section 4.3.2), and late fusion (Section 4.3.3). A chronological overview of representative LiDAR–camera fusion-based multi-modal 3D detection methods is presented in Fig. 10. We illustrate the specific implementations of these three fusion strategies in Fig. 11.

4.3.1. LiDAR–camera early-fusion for 3D object detection

Early fusion [120,157,158,164–166] refers to the integration of raw or minimally preprocessed data from different modalities—typically LiDAR point clouds and camera images—at the initial stage of sensor data processing. This approach generates a unified, enriched data representation, which is then fed into a single detection network.

For example, Complexer-YOLO [167] extends the YOLO-style detector to LiDAR point clouds, achieving a balance between accuracy and real-time performance by using semantically segmented points to reduce redundancy and eliminate costly proposal-generation procedures. PointPainting [14] adopts a sequential fusion strategy that projects LiDAR points onto the image plane and appends semantic probability vectors from image segmentation, forming a semantically enriched point cloud for downstream 3D detectors. MMF [168] integrates LiDAR geometry with image semantics to improve detection accuracy, while FusionPainting [120] performs independent 2D/3D semantic segmentation and fuses the resulting features at the voxel level. F-PointPillars [164] applies a multi-stage fusion pipeline that uses 2D detections to generate frustums and refines them with LiDAR points, preserving both semantic cues and geometric details. EA-LSS [169] generates sparse multi-view depth maps from point clouds, densifies them via expansion, and fuses them with edge maps under weighted supervision, leading to more accurate depth estimation and more effective multi-modal fusion.

In recent years, early fusion-based methods have been increasingly supplanted by middle- and late-fusion strategies due to their limited flexibility in handling heterogeneous modalities and high computational

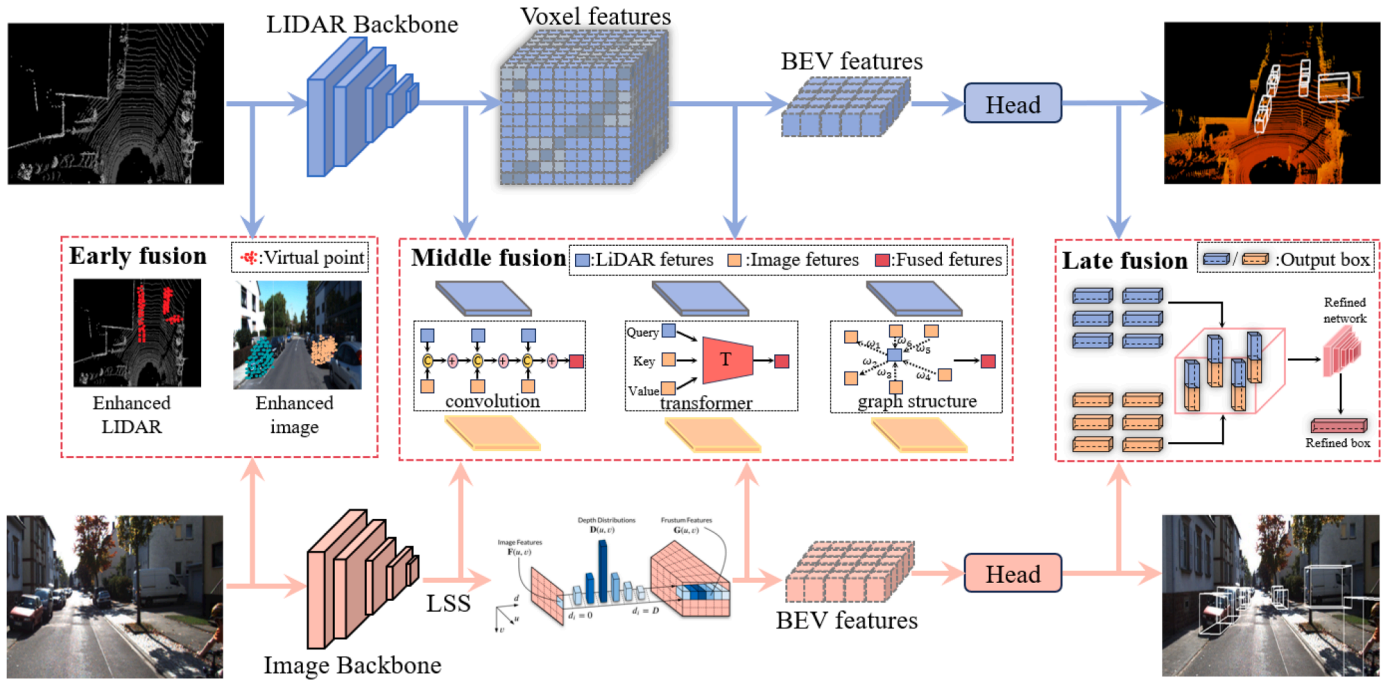


Fig. 11. Architecture of LiDAR-camera fusion 3D detection. LSS module is from [156].

cost arising from operating directly on raw or low-level feature alignment. In particular, early fusion requires strict temporal and spatial synchronization and often suffers from noise amplification when combining unprocessed sensor inputs, which makes it less robust in complex real-world driving scenarios.

4.3.2. LiDAR-camera middle-fusion based 3D object detection

Middle fusion [158,159,165,173,175,176,187–192] is the most widely adopted and flexible multi-modal fusion strategy. Its core idea is to first process each modality using its dedicated feature extraction network to obtain high-level feature representations. These modality-specific features are then deeply fused through various mechanisms—such as learnable fusion convolution [193], graph structure [144], or cross-modal transformers [187]—to achieve effective information integration. By combining features at a semantically rich stage while preserving modality uniqueness, middle-level fusion offers a favorable balance between robustness and fine-grained interaction, making it a dominant paradigm in multi-sensor perception.

For example, PointFusion [194] proposes a general multi-modal framework in which image features extracted by a CNN and point cloud features obtained via PointNet are jointly fused for 3D bounding box prediction. TransFusion [175] introduces a Transformer-based fusion mechanism with soft association, reducing the dependence on strict calibration and improving robustness under degraded image quality or sensor misalignment. DeepFusion [195] ensures accurate spatial alignment between LiDAR and image features during data augmentation by reversing geometric transformations, making it compatible with most detection architectures while boosting performance. BEVFusion [176] unifies LiDAR and camera features in a shared bird's-eye view (BEV) space with minimal architectural changes, enabling seamless multi-modal fusion. This approach preserves complementary geometric and semantic information from both modalities and supports a range of 3D perception tasks, including detection and segmentation, while maintaining computational efficiency. GraphAlign [196] and GraphAlign++ [197] formulate multi-modal feature fusion as a graph matching problem for 3D object detection, where cross-modal features are modeled as graph nodes to explicitly capture correspondences for more accurate alignment. GraphAlign++ further enhances fusion quality by improv-

ing graph construction and matching strategies, resulting in more robust and discriminative feature integration in complex driving scenarios. GaussianPretrain [198] introduces a unified 3D Gaussian representation for visual pre-training in autonomous driving, which encodes multi-view inputs into a compact 3D Gaussian scene model to enhance spatial understanding and downstream perception performance.

Currently, most methods employ Transformers [199], attention mechanisms [112], or BEV alignment [156] to facilitate deep feature interaction [130], with a particular emphasis on spatial alignment [79] and adaptive weighting [120] to improve detection performance for small or occluded objects. Consequently, mid-level fusion has become the dominant paradigm for multi-modal 3D detection, finding widespread application in autonomous driving and robotic perception.

4.3.3. LiDAR-camera Late-fusion based 3D object detection

Late fusion [179,185,186] processes each modality through an independent branch and combines their outputs using a fusion network at the prediction stage. This strategy better leverages the strengths of modality-specific architectures while avoiding the alignment challenges arising from cross-sensor discrepancies.

CLOCs [162] performs fusion at the detection-result level, offering high flexibility and compatibility. It can be integrated with nearly any existing network architecture without retraining or structural modification, while incurring negligible computational overhead—making it well suited for practical deployment. Fast-CLOCs [163] builds upon CLOCs by introducing a more efficient fusion pipeline that further improves both 3D detection accuracy and runtime performance. Pai3D [184] utilizes an image instance segmentation model to obtain 2D masks and semantic labels, which are projected onto the point cloud to generate initial 3D instance priors. Each LiDAR point is assigned an instance ID and semantic label, after which the method refines projection errors and extracts more precise geometric cues to estimate accurate 3D instance centers.

Due to their strong robustness, late-fusion methods continue to serve as the mainstream paradigm in multi-sensor fusion frameworks. However, their ability to capture fine-grained feature interactions is limited, and they are often combined [179,186] with middle fusion ap-

Table 6

A comprehensive performance analysis of LiDAR-Camera fusion 3D object detection methods across different datasets. We categorize the fusion methods into three types and report the GPU versions used for training and the inference time and performance accuracy as originally reported in their respective papers.

Method	fusion Stage		Year	Inference		Waymo Vehicle			mAp	NDS	L1	L2
	Time (ms)	GPU	KITTI Car	NuScenes		Easy	Mod.	Hard				
		Early	Middle	Late								
F-PointNet [170]	✓				CVPR2018	–	–	81.20	70.39	62.19	–	–
F-ConvNet [171]	✓				IROS2019	–	–	85.88	76.51	68.08	–	–
RoarNet [172]	✓				IROS2019	100	–	83.95	75.79	67.88	–	–
PointPainting [14]	✓				CVPR2020	–	–	82.11	71.70	67.08	46.4	58.1
PointAugmenting [15]	✓				NeIPS2021	542	V100	–	–	–	66.8	71.0
MVP [157]	✓				NeIPS2021	–	V100	–	–	–	66.4	70.5
EPNet + [158]	✓				TPAMI2022	–	V100	91.37	81.96	76.71	–	–
PPF-Det [166]	✓				TITS2024	29	Titan X	89.51	84.46	78.91	–	–
MSF [165]	✓				TGRS2024	63	V100	–	–	–	68.2	71.6
AVOD [159]		✓			IROS2018	100	Titan X	81.94	71.88	66.38	–	–
ContFuse [173]		✓			ECCV2018	60		82.54	66.22	64.04	–	–
MMF [168]		✓			CVPR2019	80		86.81	76.75	68.41	–	–
3D-CVF [174]		✓			ECCV2020	75	1080Ti	89.20	80.05	73.11	52.7	62.3
TransFusion [175]		✓			CVPR2022	265	V100	–	–	–	68.9	71.7
BEVFusion [176]		✓			ICRA2023	119	3090	–	–	–	70.2	72.9
LoGoNet [177]		✓			CVPR2023	–	–	92.04	85.04	84.31	–	–
FSF [178]		✓			TPAMI2024	141	3090	–	–	–	70.6	74.0
GraphBEV [79]		✓			ECCV2024	141	A100	–	–	–	71.7	73.6
SparseLIF [179]		✓			ECCV2024		A100	–	–	–	71.2	74.6
DAL [180]		✓			ECCV2024		3090	–	–	–	71.5	74.0
IS-Fusion [181]		✓			CVPR2024			–	–	–	73.0	75.2
EVT [182]		✓			ICCV2025	–	3090	–	–	–	72.1	74.6
MambaFusion [183]		✓			ICCV2025	–	A100	–	–	–	72.7	75.0
CLOCs [162]			✓		IROS2020	150	1080 Ti	88.94	80.67	77.15	–	–
Fast-CLOCs [163]			✓		WACV2022	125	3080	89.11	80.34	76.98	63.1	68.7
Pai3d [184]			✓		ECCV2022	–	Tesla P40	–	–	–	64.7	68.4
LT3D			✓		ICLR2023	–	–	–	–	–	64.7	68.4
Sparsefusion [185]	✓		✓		ICCV2023	–	A6000	–	–	–	72.8	70.4
SparseLIF [179]			✓	✓	ECCV2024	–	–	–	–	–	71.2	74.6
ViKLENet [186]		✓	✓	✓	CVPR2025	15	3090	91.20	86.04	81.18	68.7	–

proaches to enhance perception performance rather than being used in isolation.

4.3.4. Analysis

Currently, Multi-modal 3D object detection can be categorized into early, middle, and late fusion. Early fusion enables strong cross-modal interaction but is sensitive to noise and computationally expensive. Late fusion is efficient and flexible but fails to fully exploit cross-modal correlations. Middle fusion achieves a better balance by integrating intermediate representations and has become the dominant paradigm in BEV-based methods due to its superior trade-off between accuracy and efficiency. Table 6 provides a summary of representative LiDAR-camera fusion methods for multi-modal 3D object detection.

4.4. From 3D object detection to end-to-end autonomous driving, world models, and VLA

Traditional autonomous driving systems are built upon modular perception pipelines, where 3D object detection serves as a key component for scene understanding, typically followed by separate prediction and planning stages relying on intermediate representations such as bounding boxes or BEV features. Recently, a paradigm shift toward end-to-end autonomous driving has emerged [200–208], enabling joint optimization of perception and decision-making and reducing error accumulation across modules. Within this trend, world models learn predictive representations of environment dynamics for simulation and long-horizon planning, while VLA models further unify perception, reasoning, and action within a single framework. Representative works include GuideFlow [202], which introduces constraint-guided flow matching for safe

and feasible trajectory generation; GAIA-2 [209], a controllable multi-view diffusion-based world model for realistic driving scene simulation; and failure-aware VLA learning [203], which improves robustness through explicit learning from failure cases.

4.4.1. Analysis

Overall, this evolution from modular perception to end-to-end driving, world models, and VLA reflects a shift toward unified, foundation-model-driven autonomous driving systems. Compared with traditional modular pipelines, which decompose perception, prediction, and planning into independent stages and rely heavily on intermediate representations such as bounding boxes or BEV features, unified paradigms reduce error accumulation across stages and enable joint optimization of perception and decision-making. In addition, they allow richer multimodal reasoning and global trajectory optimization, improving generalization in complex and long-tail driving scenarios. However, these advantages come at the cost of significantly increased data requirements, as large-scale multimodal pretraining becomes essential for robust performance. Moreover, the high computational and memory overhead of foundation models limits real-time deployment on onboard systems, while their end-to-end nature reduces interpretability and makes safety validation more challenging compared to modular systems with explicit intermediate states.

4.5. Discussion

Camera-based methods provide low-cost and high-resolution semantic information but suffer from depth ambiguity and environmental sensitivity, while LiDAR-based methods offer accurate 3D geometry but are limited by sparsity and high cost. Multi-modal fusion addresses

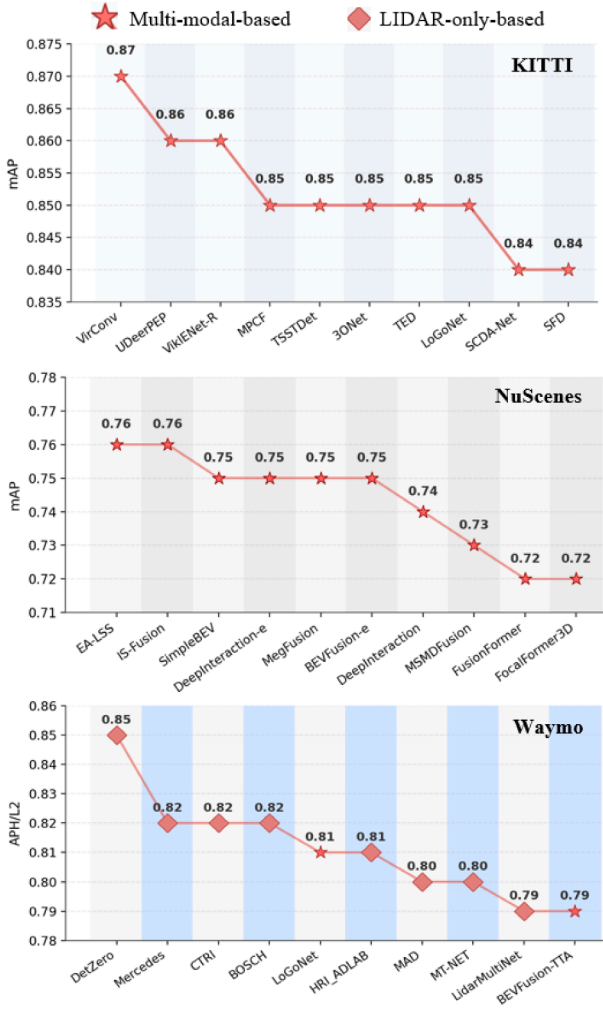


Fig. 12. Top-10 methods on the KITTI, NuScenes and Waymo dataset. The horizontal axis represents the specific methods, while the vertical axis represents the mean average precision.

these limitations by exploiting complementary information, achieving improved robustness and accuracy. Beyond fusion, recent end-to-end autonomous driving methods jointly optimize perception, prediction, and planning, reducing reliance on intermediate representations but introducing challenges in interpretability and deployment. More recently, VLA systems further integrate perception, reasoning, and action into a unified framework, enhancing decision-making capability while requiring large-scale data and high computational resources. Overall, autonomous driving paradigms are evolving from sensor-level fusion to end-to-end systems and VLA, reflecting a trade-off between performance, efficiency, and deployability.

Fig. 12 presents the top-performing methods on the three benchmark datasets, and we only include methods that provide links to accessible paper codes. Notably, the top 10 entries on both the KITTI [54] and NuScenes[64] leaderboards are all multi-modal fusion approaches, and two fusion-based methods also appear among the leading approaches on the Waymo [65] leaderboard. This highlights the substantial performance gains enabled by cross-sensor fusion. However, these improvements come at the cost of increased computational demands and greater deployment complexity. This trend has driven the development of lightweight designs, making knowledge distillation [29,210–212] a promising strategy for balancing accuracy and efficiency in future 3D detection frameworks.

5. Knowledge distillation methods

Knowledge distillation [129,213,214] is regarded as a typical paradigm of “Have both fish and bear’s paw.” Its core idea lies in transferring the knowledge in a large-scale teacher model to a lightweight student model, enabling the student to achieve performance comparable to, or even surpassing, the teacher while maintaining low computational and storage costs.

This property is particularly crucial for LiDAR-based 3D object detection, where models typically rely on high-resolution voxelization [8], complex 3D backbone networks [175], and dense multi-scale feature interactions [149], resulting in excessive parameter redundancy, high inference latency, and prohibitive deployment costs on resource-constrained platforms such as in-vehicle edge devices. Compared with 2D vision tasks [22], 3D object detection [102] additionally requires the regression of high-dimensional geometric attributes and must cope with challenges including severe point cloud sparsity, non-uniform spatial distributions, and substantial variations across sensor configurations, which impose more stringent demands on feature representation capacity and training stability. By leveraging knowledge distillation, the structured spatial perception ability [211], global contextual information [25], and inter-class discriminative [215] knowledge embedded in teacher models can effectively guide the learning process of student models, thereby mitigating the performance degradation of low-capacity networks [49,216] in complex 3D scenes. Moreover, knowledge distillation can be combined with other model compression techniques, such as pruning [49] and quantization [217], to further enhance the performance and deployment efficiency of student models. Consequently, developing effective knowledge distillation strategies tailored to 3D object detection has become a critical research direction for improving detection accuracy, robustness, and practical deployability. Knowledge distillation in 3D object detection is generally categorized into three major paradigms: feature distillation (Section 5.1), relation distillation (Section 5.2), and response distillation (Section 5.3). In this chapter, we provide a detailed exposition of the distinctions among these paradigms and present extensive illustrative figures to facilitate a clear understanding for readers.

5.1. Feature distillation

Feature distillation [16,28,129,193,218–222] is among the most widely adopted knowledge distillation strategies in 3D object detection. Unlike output-level distillation [213,223], which only aligns the final predictions of the student and teacher models, feature distillation operates on intermediate representations, allowing the student to capture richer, more detailed supervisory signals. Specifically, feature distillation guides the student model to align with the teacher’s feature maps across multiple levels of abstraction, including semantic [224], spatial [18], and geometric cues [129]. This multi-level alignment enables the student to learn fine-grained structural and contextual information, improving the capacity of its feature representations.

In 3D object detection tasks, feature distillation techniques can be applied to various representation spaces, such as voxel grids [225], bird’s-eye view (BEV) feature maps [28], and multi-scale features [222]. By transferring knowledge at the feature level, student models can effectively capture the spatial structures and semantic characteristics of objects, leading to significant improvements in detection accuracy. For example, BEV feature distillation [193,226] allows the structured geometric information to be conveyed directly in the task-critical bird’s-eye view space, enabling student models to more efficiently learn scene layout and geometric cues. Multi-scale [227,228] feature distillation enables students to simultaneously acquire semantic and fine-grained information across different feature levels, enhancing their ability to represent objects of varying sizes and improving robustness in complex scenarios. Masked feature distillation [221,229] enables explicit focus on task-critical regions, thereby avoiding interference from redundant or

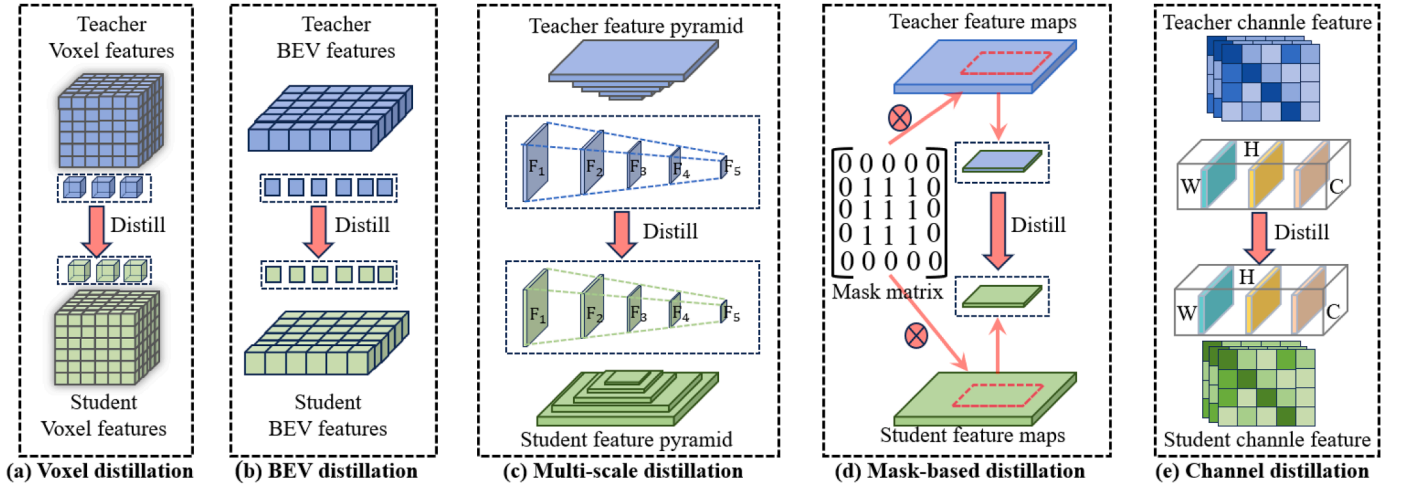


Fig. 13. Different types of feature distillation.

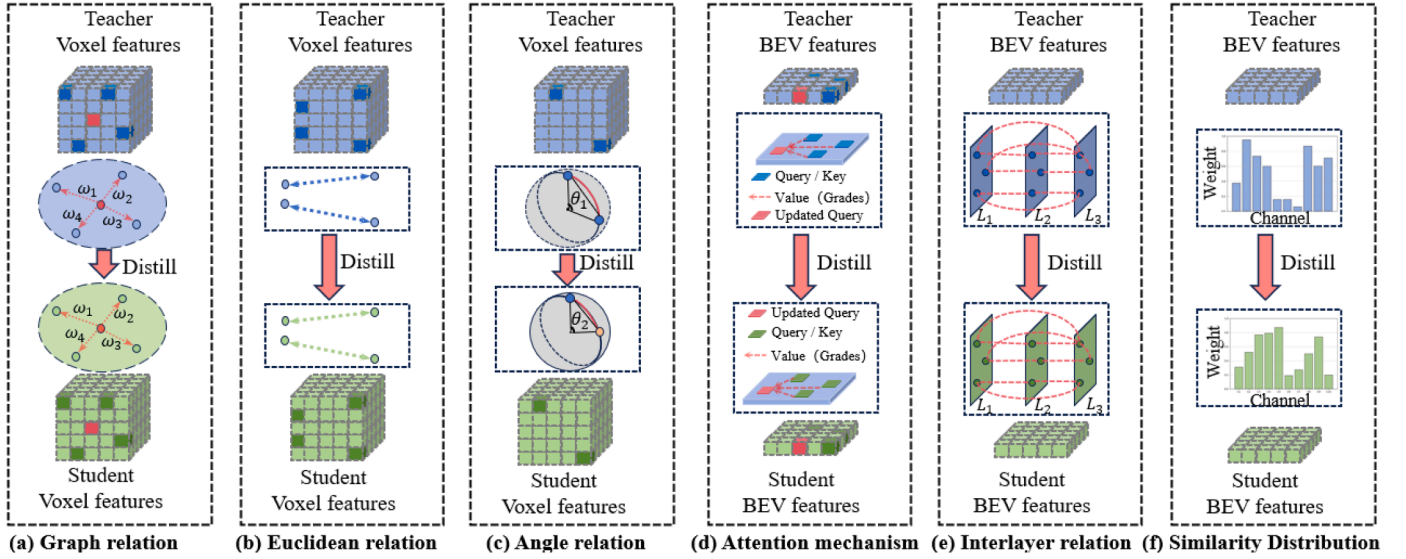


Fig. 14. Different types of relational distillation.

noisy information in global features. This allows the student network to prioritize learning the most discriminative structural representations within its limited feature capacity. Overall, feature distillation is effective because it aligns intermediate representations between teacher and student models, enabling the student to learn rich semantic and structural information beyond final output supervision, thereby improving representation quality, generalization ability, and robustness, especially in complex scenarios. Fig. 13 illustrates five representative feature distillation techniques commonly used in 3D object detection frameworks.

5.2. Relation distillation

Relation distillation [17,82,213,221,222,230,230,231][17,82,82,213,221,222,230,230,231] is a knowledge transfer strategy that extends beyond direct feature imitation. Its objective is not only to reproduce the teacher model's intermediate representations, but also to explicitly guide the student in capturing the structured relationships [232] among objects or spatial locations within a scene. By emphasizing the relational information embedded in feature maps, relation distillation enables the student to model the interdependencies between entities—an ability that is essential for understanding complex environments.

Specifically, relation distillation typically aligns similarity matrices [213,231], adjacency graphs [211] or linear correlation [82,230] between the teacher and student models. For example, graph relation distillation [79,211] explicitly models the structured dependencies among objects or feature elements, leveraging graph-topological constraints to enhance the student network's ability to understand global scene relationships. Angle relation distillation [216,230] constrains the angular distribution between feature vectors, enabling the student to better preserve the teacher's discriminative geometric structure and thereby improve inter-class separability and intra-class consistency. Interlayer relation distillation [222,228] captures the cross-scale dependencies across different levels of the teacher network, allowing the student to jointly learn shallow and deep associations, which leads to richer hierarchical semantic representations. Overall, relation distillation is effective because it captures structural dependencies between features or instances, allowing the student model to learn higher-order contextual relationships beyond individual feature alignment, thereby improving its ability to model global structure, inter-object interactions, and scene-level consistency. We introduce six commonly used relational distillation strategies, as illustrated in Fig. 14, each of which models the relationships between teacher and student features or outputs from a different perspective.

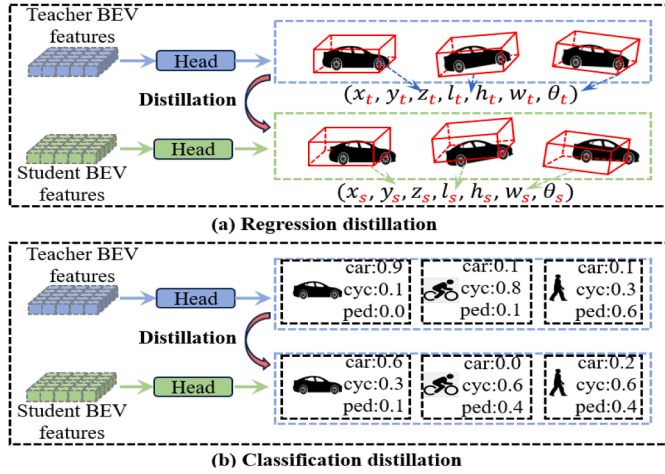


Fig. 15. Various distillation strategies for response distillation.

5.3. Respond distillation

In 3D object detection, response distillation [16,17,28,213,219,221–223] primarily focuses on transferring knowledge from the output layer of the teacher model to the student model. By guiding the student to align with the teacher's predictions on classification and bounding box regression. Response distillation provides task-specific supervision that directly reflects the detection objectives. This allows the student model to learn how to produce more accurate and reliable outputs, effectively improving detection performance. Response distillation generally comprises two types: regression distillation and classification distillation, as illustrated in Fig. 15.

In classification distillation [219,230,233], the student model is trained to mimic the soft label predictions of the teacher model, which typically contain more nuanced information than hard one-hot labels. Hard labels [234] are the conventional one-hot annotations, where each sample is assigned a single class with a probability of 1, and all other classes are assigned 0, providing a clear but coarse supervision signal. In contrast, soft labels [219], typically generated by a teacher model in knowledge distillation, encode a probability distribution over all classes, capturing the relative similarities and uncertainties among categories. By leveraging soft labels, student models can learn richer inter-class relationships and subtle patterns absent in hard labels, thereby improving generalization and enabling more stable training, as illustrated in Fig. 15(a). In contrast, regression distillation [26,49,210,235] focuses on transferring knowledge related to the positional offsets of bounding boxes, including object locations, dimensions, and orientations. By guiding the student to minimize the discrepancies between its predicted offsets and those of the teacher, regression distillation enables the student model to achieve more precise 3D localization, thereby improving detection accuracy, as illustrated in Fig. 15(b), where the decimals indicate the predicted scores for each class (Car, Pedestrian, Cyclist).

5.4. Discussion

In summary, knowledge distillation leverages complementary strategies, including feature, relation, and response distillation, to enhance student model performance. Feature distillation aligns intermediate representations to capture semantic and structural information, relation distillation models dependencies among features or instances to learn contextual relationships, and response distillation transfers teacher outputs to guide decision-level learning. Together, these strategies form a comprehensive framework for 3D object detection, often used in combination to enable lightweight models to achieve performance close to teacher models while significantly reducing computational cost. From an information fusion perspective, knowledge distillation can be viewed

as an implicit fusion mechanism that integrates heterogeneous knowledge within a unified learning process. Different distillation levels correspond to hierarchical information aggregation, from representation alignment to decision fusion. Unlike explicit multimodal fusion operating on structured inputs, distillation performs fusion in latent space via teacher-student transfer, enabling more flexible and efficient integration and improving generalization and robustness in 3D perception tasks.

6. LiDAR-to-LiDAR knowledge distillation for 3D object detection

In LiDAR-to-LiDAR knowledge distillation, the teacher model is typically a parameter-heavy network [8,14,114,236] that achieves high performance at the cost of increased inference time, whereas the student model is usually a more compact network [88,89,91,102,123] with faster inference but lower accuracy. The goal of knowledge distillation is to transfer as much knowledge as possible from the teacher to the student, improving the student's performance while maintaining a reduced parameter count. Depending on whether the teacher and student share the same architecture, the distillation process can be broadly categorized into homogeneous model distillation (Section 6.1) and heterogeneous model distillation (Section 6.2). A chronological overview of LiDAR-to-LiDAR knowledge distillation methods for 3D detection is illustrated in Fig. 16.

6.1. Homogeneous model distillation

In homogeneous distillation, the teacher and student networks typically share the same architecture. For example, SECOND [88] → lightweight SECOND or PV-RCNN [8] → lightweight PV-RCNN. In such cases, the design of distillation strategies is relatively straightforward, as the teacher and student maintain identical architectural configurations at each stage, with the primary difference lying in the depth or width of the networks. Depending on how the student model is constructed, homogeneous distillation can be categorized into pruning-based distillation [29,80,125,211,212,216,237,238] and quantization-based distillation [217,239,240].

6.1.1. Pruning-based distillation

The pruning-based distillation paradigm represents a typical two-stage model compression strategy. In the first stage, the original high-performance teacher model undergoes random [211] or conditional pruning [212], in which redundant channels [216], convolutional kernels [29], or even entire network layers [49] are removed to derive a lightweight student model architecture. While pruning substantially reduces model complexity and accelerates inference, it often results in a noticeable drop in detection performance. To mitigate this issue, the second stage incorporates knowledge distillation [233], where the unpruned full teacher model serves as the knowledge source to guide the pruned student model in recovering performance. This two-stage process enables achieving high compression ratios and significant inference speedups while preserving as much detection accuracy as possible, as illustrated in Fig. 17.

For example, in random pruning, PointDistiller [211] and H2RKD [216] reduces the number of channels in the teacher's feature maps to one-half or one-quarter of the original size, substantially decreasing the network parameters and FLOPs, and then employs knowledge distillation to recover performance. InstKD [233] further demonstrates this idea on the KITTI dataset: by reducing the channel width of SECOND [88] by half, the model size is reduced by 75.5% and the FLOPs by 74.5%, while the distilled student still achieves 66.83% mAP, surpassing its teacher model. RDD [235] reduces the number of network layers while simultaneously minimizing the differences between teacher and student features and logits in the proposal regions, effectively narrowing the representation gap. This significantly decreases the performance gap and improves the accuracy of lightweight models, achieving notable

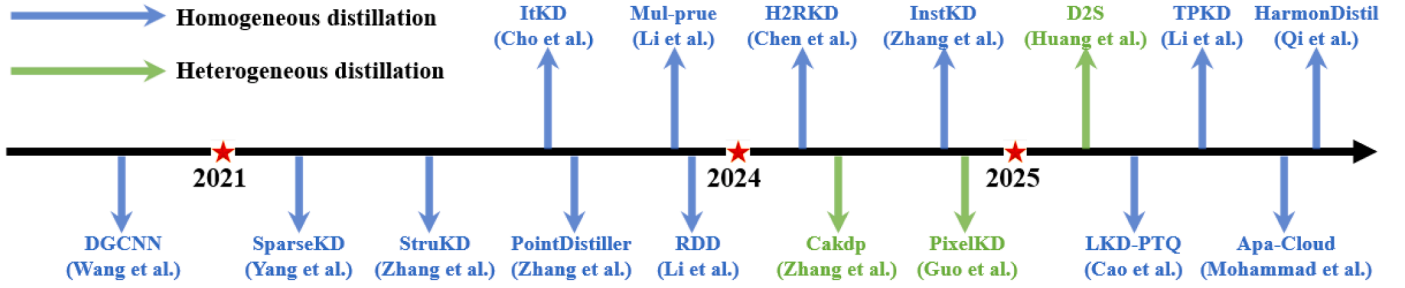


Fig. 16. Chronological overview of the LiDAR-to-LiDAR knowledge distillation for 3D object detection.

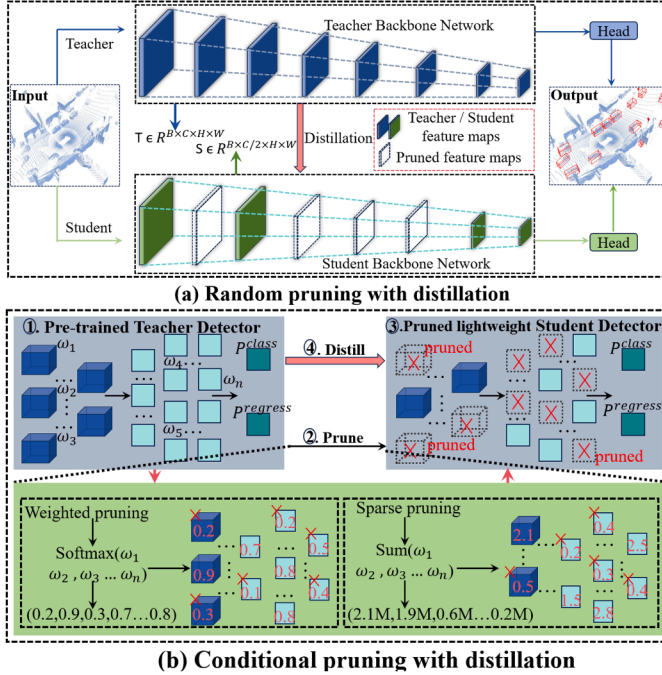


Fig. 17. Architectures of two widely used pruning-based distillation methods for 3D object detection.

mAP gains on benchmarks such as nuScenes while substantially reducing FLOPs.

In conditional pruning, the student model selectively removes certain network layers based on predefined criteria—for example, pruning layers with smaller weight magnitudes or discarding layers with larger parameter counts. Specifically, Mul-prune [212] estimates the importance of each voxel and employs an importance-sampling module to discard less informative voxels when constructing the student model. The distilled knowledge from the teacher is then transferred to the student, effectively eliminating redundant computation while preserving essential structural features. EFD [238] computes the importance of network channels via a softmax-based weighting scheme and progressively adjusts a pruning threshold to reduce the number of parameters.

Pruning-based distillation combines network pruning with knowledge transfer to build lightweight 3D detection models. By removing low-contribution weights or channels, these methods reduce parameters and speed up inference. As shown in Table 7, a well-designed teacher–student framework can even enable the student to outperform its teacher. However, their effectiveness depends on carefully designed pruning and distillation strategies, as overly aggressive pruning [29] can degrade feature representation and performance.

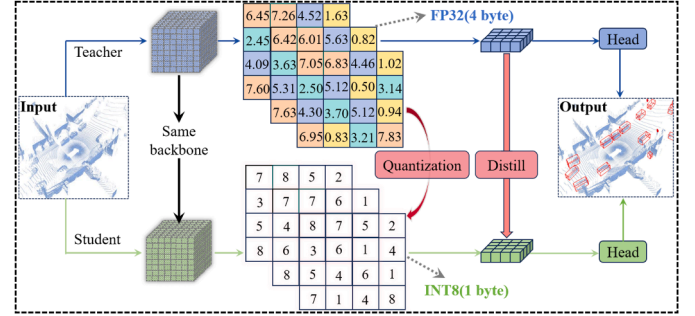


Fig. 18. Architectures of quantization-based distillation methods for 3D object detection.

6.1.2. Quantization-based distillation

The core idea of quantization is to reduce the numerical precision of a neural network, thereby decreasing model size, accelerating computation, and lowering energy consumption. Specifically, the original 32-bit floating-point (FP32) representations of weights and activation tensors are converted to lower-precision formats, such as 8-bit integers (INT8), or even more aggressive 4-bit and 2-bit representations. To mitigate the accuracy degradation caused by quantization, knowledge distillation is often employed, wherein the student network learns from the teacher's knowledge to recover performance, as illustrated in Fig. 22.

LiDAR-PTQ [239] presents a post-training quantization frame for LiDAR-based 3D object detection. It employs a customized loss function to measure the discrepancy between predictions before and after quantization, ultimately selecting the most suitable quantization thresholds and incorporating knowledge distillation to enhance the accuracy of the quantized model. QD-BEV [217] proposes a view-guided distillation framework that integrates quantization during training to simulate low-precision computation and mitigate performance degradation. Dqs3D [240] utilizes voxel-level dense matching to align teacher and student predictions, generating richer pseudo-labels and stronger training signals. Additionally, it compensates for misalignment caused by quantization through a point-to-voxel discretization closure rule, ensuring teacher-student consistency in the quantized view and improving both efficiency and accuracy.

Quantization-based distillation [217,239–241] accelerates 3D object detection while maintaining accuracy. Quantization reduces numerical precision—for instance, by using lower-bit weights and activations—to decrease memory usage and computational cost, whereas distillation transfers knowledge from a high-precision teacher to guide the low-precision student. Through this combination, efficient deployment on resource-constrained devices is achieved while minimizing performance loss. However, its effectiveness depends on careful calibration of the quantization levels and distillation design, as overly aggressive quantization can degrade feature representations and detection accuracy Fig. 18.

Table 7

Evaluation of model efficiency analysis of pruning-based distillation on KITTI dataset. ‘T’: Teacher model, ‘S’: Student model, ‘F’: FLOPs, ‘P’: Parameters, ‘mAP’: mean average precision..

	Model	F(G)	P(M)	Car	Ped.	Cys.	mAP(↑)
T	SECOND [88]	69.8	5.3	88.9	58.1	69.9	72.3
S	+ PointDistiller [211]	17.8	1.4	88.6	58.1	72.9	73.2
	+ H2RKD [216]	17.8	1.4	88.0	59.2	71.8	73.0
T	PointRCNN [92]	104.9	4.1	86.7	64.5	74.6	75.3
S	+ PointDistiller [211]	13.7	0.5	85.7	67.2	72.3	75.1
	+ H2RKD [216]	4.6	0.4	86.7	68.2	71.8	75.8
T	Pointpillars [123]	34.3	4.8	88.1	51.8	65.0	68.3
S	+ PointDistiller [211]	17.8	1.4	89.0	52.8	65.8	69.2
	+ H2RKD [216]	4.6	0.4	88.4	53.0	67.7	69.6

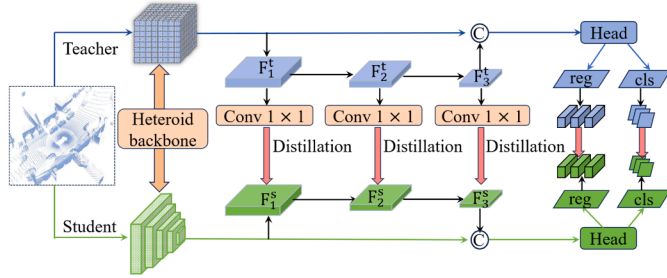


Fig. 19. A universal framework for heterogeneous distillation in 3D object detection.

6.2. Heterogeneous model distillation

This represents a specialized form of knowledge distillation. In conventional homogeneous distillation, the teacher and student models typically share similar network architectures [211]. Heterogeneous distillation, on the other hand, refers to scenarios where the teacher and student networks have fundamentally different structures. When knowledge needs to be transferred from one architecture (e.g., a large transformer-based model [117] or a two-stage detector [114]) to a substantially different architecture (e.g., a voxel-based sparse model [102] or a single-stage detector [113]), homogeneous distillation often fails, making heterogeneous distillation necessary. However, due to architectural differences between the teacher and student, the intermediate feature maps differ significantly in dimensions, scale, and levels of semantic abstraction, rendering direct alignment infeasible. This constitutes the core challenge of heterogeneous knowledge distillation.

For example, the teacher model PV-RCNN [8] generates multi-scale voxel features and keypoint features, whereas the student model PointPillars [164] produces only single-scale pillar feature maps. In this scenario, applying L1 or L2 losses directly between these features is infeasible due to the lack of feature correspondence. A common solution is to use convolutions [242] to align the intermediate feature maps before applying the loss for knowledge transfer. Alternatively, output-based distillation [49] transfers knowledge at the final detection results, avoiding the discrepancies in intermediate network structures, since the final outputs are consistent, as illustrated in Fig. 19.

Specifically, CaKDP [49] proposes a novel approach that introduces a category-aware distillation strategy by analyzing the discrepancies between teacher and student models. This method adjusts the distillation targets and transfers the knowledge from a two-stage detector to a single-stage student model, thereby alleviating the limitations of the teacher. It not only enhances the performance of the student model but also opens up new research directions for compression and distillation across heterogeneous architectures. D2S [242] introduces a dense-to-sparse distillation strategy that transfers the teacher model's knowledge obtained under dense point cloud inputs to the student model, enabling

it to maintain efficient detection performance even with sparse point cloud inputs. Furthermore, it incorporates multi-level feature alignment, which aligns the feature representations of teacher and student models across different layers, thereby strengthening the student's ability to learn from the teacher. PixelKD [243] proposes an innovative distillation paradigm, termed Pixel Distillation, which addresses the limitations of traditional distillation methods in handling input size variations and architectural heterogeneity. It extends knowledge distillation to the input level, while breaking architectural constraints and enabling flexible cost control. As shown in Table 8, heterogeneous distillation demonstrates excellent performance both in terms of model compression and accuracy improvement.

6.3. Discussion

In 3D object detection, LiDAR-to-LiDAR knowledge distillation has emerged as a crucial approach for enhancing the performance of lightweight models. Current studies primarily focus on two scenarios: homogeneous distillation [211,216] and heterogeneous distillation [49,242]. Homogeneous distillation transfers intermediate features and detection responses between structurally similar networks, offering relatively stable and straightforward strategies. In contrast, heterogeneous distillation involves teacher and student models with distinct architectures, requiring feature alignment, projection transformations, or relational modeling to bridge representational gaps. Although heterogeneous distillation faces greater challenges in loss design and knowledge transfer, it demonstrates stronger potential for performance improvement. Looking forward, with the rapid development of Transformer-based models [117], sparse convolution networks [178,185], and multi-modal fusion [179] frameworks in 3D detection, LiDAR-to-LiDAR distillation is expected to become increasingly diverse. Future research will not only need to address cross-representation alignment but also integrate emerging techniques such as self-distillation [228], contrastive learning [246], and semi-supervised learning [110] to further improve the accuracy and generalization of lightweight models. We have collected representative knowledge distillation (KD) methods applied in LiDAR-to-LiDAR scenarios and briefly summarize their corresponding contributions, as shown in Table 9.

7. LiDAR-to-camera knowledge distillation for 3D object detection

In 3D object detection, LiDAR-based detectors [89,91,102] typically achieve higher performance than image-based methods [16,156,247] because they have access to dense and highly accurate geometric information. Camera sensors, however, are more practical in real-world applications due to their low cost, low power consumption, and ease of deployment. Yet, image data inherently lacks reliable depth cues, which limits the accuracy of camera-based 3D detectors. To bridge this performance gap, extensive research has explored cross-modal knowledge distillation [17,154,219,221,226,248], where a LiDAR detector serves

Table 8

Evaluation of model efficiency analysis of heterogeneous distillation methods on KITTI dataset. ‘T’: Teacher model, ‘S’: Student model..

	Scheme	F(G)	P(M)	Car	Ped.	Cys.	mAP(↑)
T	Voxel-RCNN [7]	81.6	11.0	84.9	57.7	73.7	72.1
	SECOND [88]	80.7	5.3	81.3	52.9	65.8	66.7
S	+ PointDistiller [211]	45.7	3.0	81.3	50.6	66.4	66.1
	+ CaKDP [49]	30.2	1.5	82.7	60.4	72.1	71.7
T	PV-RCNN [8]	93.1	13.1	84.2	57.6	72.3	71.4
	Centerpoint [102]	97.9	5.8	83.4	57.5	70.8	70.6
S	+ PointDistiller [211]	43.2	3.1	86.5	58.5	69.3	68.7
	+ CaKDP [49]	39.5	1.8	80.5	59.1	70.2	69.9

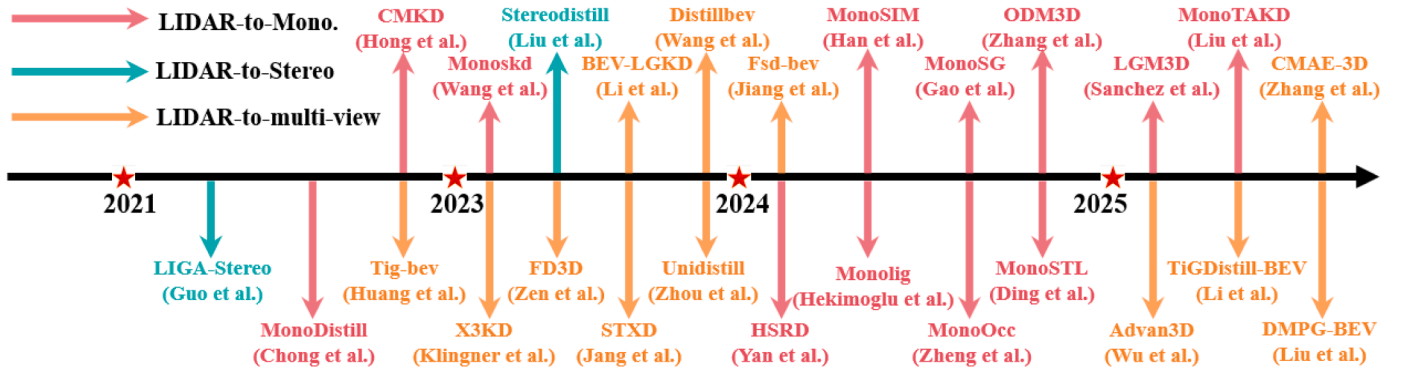


Fig. 20. Chronological overview of the LiDAR-to-camera knowledge distillation for 3D object detection.

Table 9

Performance of different methods from LiDAR-to-LiDAR knowledge distillation for 3D object detection. We sort these methods by publication year and present the float operations (FLOPs) and number of model parameters (Para.) for each, along with a brief summary of their respective contributions.

Method (G) (M)	Type Para.	Year	KITTI Car		NuScenes		Contribution		Hard	mAP	NDS	
			FLOPs	Para.	Easy	Mod.	Hard	Mod.				
SparseKD [29]	homo.	NeIPS2022	20.5	1.4	–	64.21	–	53.0	40.2	–	–	Key position distillation and teacher-guided initialization
ItKD [210]	homo.	CVPR2023	–	–	–	–	–	53.9	41.3	–	–	Interchange distillation mechanism compression
PointDistiller [211]	homo.	CVPR2023	45.6	4.8	88.9	76.9	73.6	56.4	65.1	–	–	Local graph convolution modeling compression
Mul-prue [212]	homo.	ICIP2023	–	–	–	78.5	–	–	–	–	–	Multi-dimensional pruning strategy compression
RDD [235]	homo.	ICCV2023	36.0	2.8	–	68.2	–	54.0	62.1	–	–	Representation disparity-aware distillation
H2RKD [216]	homo.	MM2024	45.8	4.6	87.9	76.8	73.9	56.3	65.7	–	–	Combine homogeneity and heterogeneity
InstKD [233]	homo.	TIV2024	39.0	3.3	89.5	78.3	75.1	52.7	61.8	–	–	Instance-Aware Knowledge Distillation
Apa-Cloud [244]	homo.	RA-L2025	–	–	–	80.8	–	–	–	–	–	Adaptive hierarchical framework compression
CaKDP [49]	hete.	CVPR2024	30.2	1.5	–	82.4	–	–	–	–	–	Category-Aware knowledge distillation
CLEAN [245]	hete.	TPAMI2025	27.3	1.3	–	83.1	–	–	–	–	–	Category-Aware Network Pruning

as the teacher network to transfer structured geometric information to a camera-based student network. Camera-captured image data can be categorized into monocular, stereo, and multi-view images. Although the core idea of distillation remains the same across these settings, the distillation strategies differ. Therefore, we review LiDAR-to-monocular (Section 7.1), LiDAR-to-stereo (Section 7.2), and LiDAR-to-multi-view (Section 7.3) distillation methods separately. Fig. 20 presents a detailed chronological overview of the various LiDAR-to-Camera knowledge distillation methods that have been proposed for 3D object detection.

7.1. LiDAR to monocular image

In LiDAR-to-monocular image 3D object detection, the core objective of distillation is to transfer depth information from the LiDAR model to the monocular model, thereby compensating for the inherent limitations of depth perception in monocular images. Common distillation strategies include using depth distillation via LiDAR depth maps [156,218,249,250] and BEV projection distillation [213,226,230,251].

7.1.1. Depth distillation via LiDAR depth maps

Monocular images inherently lack explicit depth cues, making monocular 3D object detection highly ill-posed and susceptible to depth ambiguity. In contrast, LiDAR sensors provide precise and reliable geometric and depth measurements. Using LiDAR as the teacher in a distillation framework enables the transfer of depth-aware structural knowledge, thereby substantially enhancing the monocular detector's capability to infer accurate 3D geometry.

For example, MonoDistill [218] projects LiDAR points onto the image plane to produce image-like feature maps, and a LiDAR-based detector with the same architecture serves as the teacher. Subsequently, knowledge distillation is employed to guide the student model in learning depth-aware spatial representations, thereby enhancing its detection capabilities. MonoSTL [249] tackles the negative transfer issue in cross-modal distillation, particularly under severe depth ambiguity in monocular images. It adaptively samples contextual nodes and dynamically predicts hybrid depth-aware filter weights and affinity matrices to fuse multi-scale depth cues with image context. CaDDN [156] proposes the categorical depth distribution approach, which predicts a discrete depth probability distribution for each pixel instead of directly regressing con-

tinuous depth. This method enables more robust handling of depth uncertainty and spatial geometry, allowing monocular 3D detection networks to leverage reliable depth information and thereby significantly improve 3D object localization performance.

7.1.2. BEV projection-based distillation

In 3D object detection, existing knowledge distillation strategies primarily focus on the Bird's-Eye-View (BEV) representation [28,213,219,226,230,252], as BEV naturally places all objects within a unified and well-structured Euclidean coordinate system, providing an ideal and efficient intermediate representation for knowledge transfer. Within the BEV space, object scale, location, and spatial relationships are accurately preserved, enabling the precise geometric and semantic priors learned by the teacher model to be conveyed to the student model in a clearer and more structured manner.

MonoSKD [213] introduces a BEV-level distillation scheme based on the Spearman rank correlation coefficient to alleviate cross-modal discrepancies in monocular 3D detection. By focusing on the relative ordering of features rather than their absolute values, it achieves more robust cross-modal alignment and avoids the pitfalls of strict feature matching. A relaxed distillation strategy further tolerates moderate discrepancies between teacher and student feature spaces, improving training stability and robustness. SHRD [230] introduces a cross-modal distillation framework that transfers high-order geometric structures such as points, lines, and planes from LiDAR to the monocular detector in BEV space. By conveying these structural cues, SHRD substantially strengthens spatial perception, particularly improving performance on non-rigid categories such as cyclists. X3KD [226] reduces the depth and spatial representation gap between camera-only students and LiDAR-based teachers. After projecting features into BEV space, the student is guided by LiDAR-derived representations, and adversarial learning is incorporated to enforce BEV consistency. This combination leads to more depth-aware and LiDAR-aligned BEV features for monocular 3D detection.

7.1.3. Analysis

LiDAR-to-monocular knowledge distillation improves monocular 3D detection by transferring geometric priors from LiDAR teachers. Existing methods mainly adopt depth-based or BEV-based distillation. Depth distillation provides direct geometric supervision but is limited by local consistency and depth discretization noise. BEV distillation enforces global spatial alignment in a unified top-down space, but introduces additional projection complexity and potential information loss. Overall, depth distillation emphasizes local accuracy, while BEV distillation focuses on global consistency, reflecting a trade-off between simplicity and structural completeness. As shown in Table 10, the camera-based monocular 3D object detection methods achieve substantial performance improvements after applying knowledge distillation.

7.2. LiDAR to stereo image

LiDAR-to-stereo knowledge distillation [84,138] is a technique that enhances stereo 3D perception by combining binocular visual information with accurate geometric priors provided by LiDAR. The core idea is similarly to leverage a pretrained LiDAR teacher model to provide reliable depth supervision and spatial structural knowledge to the stereo image student model. However, stereo 3D detection relies on high-quality calibration of the dual cameras, including precise baseline measurements, intrinsic and extrinsic parameters, and synchronized image frames. In public datasets such as KITTI [54] and NuScenes [64], annotated stereo data are relatively scarce, whereas monocular and LiDAR data are more abundant. Consequently, research on stereo-based 3D object detection has received comparatively less attention, though some studies have addressed this area.

For example, LIGA-Stereo [84] extracts high-level geometry-aware features from a LiDAR teacher—such as object boundaries and surface normals—to enhance the stereo student's geometric representation. To

Table 10

Performance improvement of camera-based detectors via knowledge distillation on the KITTI dataset: average BEV and 3D detection at IoU 0.7.

Method	Modality	Teacher	AP_{BEV}	AP_{3D}
CaDDN [156]	C	–	18.91	13.41
+ CMKD [219]	C	L	23.10	16.99
+ ADD [16]	C	L	23.58	16.81
+ Monocd [132]	C	L	22.81	16.59
+ MonoTAKD [28]	C	L	27.76	19.43

compensate for the limited semantic supervision inherent in indirect 3D guidance, an auxiliary 2D detection head is introduced to provide explicit semantic cues, thereby improving overall performance. Stereodistill [138] narrows the performance gap between stereo and LiDAR by introducing a cross-modal distillation framework. The teacher's 3D bounding boxes are decomposed into sub-components, and distillation is applied only when they align with ground truth, increasing valid positive samples and simplifying student learning.

LiDAR-to-stereo distillation transfers geometric priors from LiDAR to stereo representations, but suffers from modality gap and difficult cross-modal alignment. Moreover, stereo inputs are sensitive to calibration errors and low-texture regions, further limiting effectiveness. Consequently, this paradigm remains relatively under-explored compared to LiDAR-to-monocular or BEV-based distillation methods.

7.3. LiDAR to multi-view image

LiDAR to multi-view distillation [17,18,18,140,169,193,217,222,223,226,227,232,253,254] is a key technique in 3D object detection. It leverages a LiDAR teacher model with rich geometric features to provide cross-modal geometric and semantic supervision for the multi-view image student model. Current approaches typically employ view-transformation modules—such as LSS-based methods [9] or Transformer-based BEV generators [255]—to encode multi-view images into a unified BEV representation, followed by multi-level alignment with the BEV features produced by the LiDAR teacher. This alignment mainly involves multi-scale feature distillation [222,230] and output-level respond distillation [221].

7.3.1. Multi-scale feature distillation

Multi-scale distillation is a hierarchical knowledge transfer mechanism that guides a lightweight student model to mimic the feature representations of a larger teacher model across multiple levels—including low-level details, middle-level structures, and high-level semantics—thereby enabling more comprehensive and fine-grained knowledge transfer. This strategy proves particularly effective in the context of multi-view data.

For instance, DistillBEV [222] presents a cross-modal distillation framework where the student learns multi-scale BEV features from a LiDAR teacher, improving geometric perception under image-only input. Distillation is performed across scales and with temporal fusion, enabling the student to capture essential spatio-temporal cues from multi-view BEV representations. FSD-BEV [227] eliminates the need for a pre-trained teacher by employing a joint teacher–student backbone. The teacher branch, guided by LiDAR hard labels, focuses on foreground feature learning to bridge modality gaps. Sparse point clouds are supplemented with pseudo-points generated from ground-truth boxes, and multi-scale foreground features are fused via a feature pyramid for enhanced detection. CMAE-3D [228] introduces a self-supervised approach combining contrastive learning and masked autoencoding to address foreground–background imbalance and uneven point cloud density. Representative regions are selectively masked to enhance attention, while multi-scale latent feature reconstruction captures high-level semantics and suppresses redundant low-level details, improving representation efficiency.

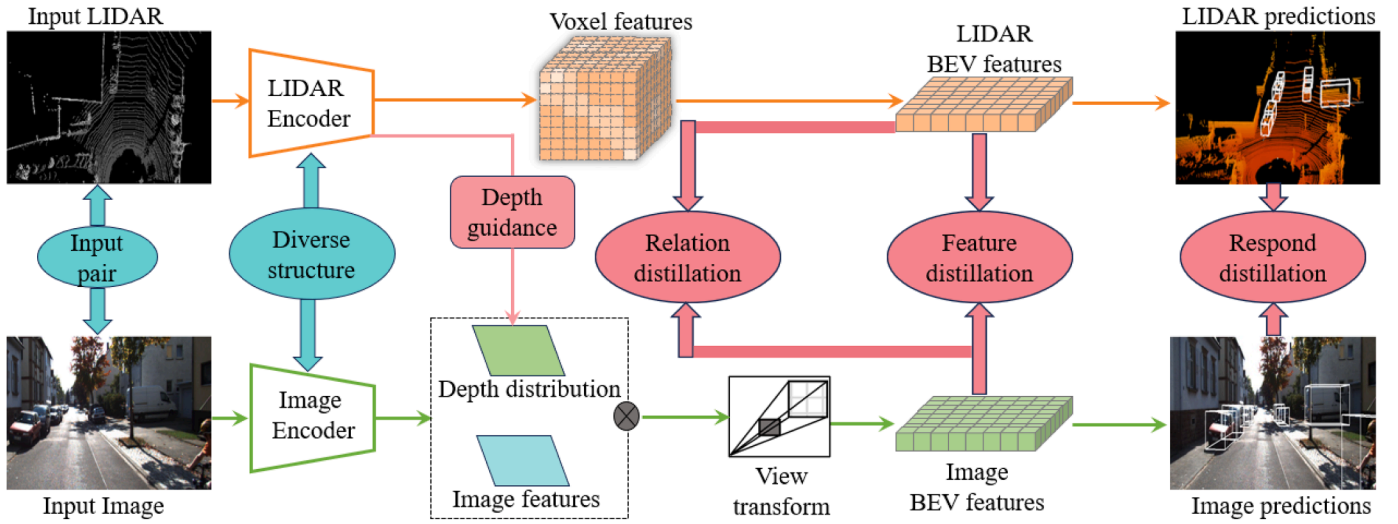


Fig. 21. Frame of knowledge distillation from LiDAR to camera for 3D object detection.

7.3.2. Output-level respond distillation

Output-level distillation refers to training the student model using the final predictions of the teacher model as supervision, including class probabilities [17] and bounding box locations [221]. By aligning the outputs of the teacher and student, the student model can learn the teacher's decision patterns and spatial distribution information, thereby improving prediction accuracy.

FD3D [220] introduces learnable queries to locate key regions in the teacher model's predictions and generate masked features, addressing noise and misleading teacher signals in multi-camera BEV-based 3D detection. This enhances the student's representations in critical regions, improving both accuracy and efficiency. STXD [17] proposes a structure- and temporal-aware cross-modal distillation framework that maximizes classification similarity by correlating teacher and student predictions across channels, thereby strengthening the student's structural representation capability. UniDistil [221] leverages shared detection exemplars in the BEV space to enhance single-view detectors. It computes the linear relationships between teacher and student BEV features relative to multiple levels of ground truth, promoting consistent and accurate knowledge transfer and improving single-view 3D detection performance.

7.3.3. Analysis

In the field of 3D object detection, LiDAR to multi-view image knowledge distillation has emerged as an important approach for enhancing the performance of vision-based multi-view detectors. Compared to monocular or stereo methods, multi-view images provide richer spatial information and broader viewpoint coverage. By introducing a LiDAR teacher model, depth estimation and spatial understanding can be further improved. Nevertheless, this research direction still faces challenges such as data scarcity and high computational complexity. For instance, almost all multi-view image-based 3D detection methods are evaluated on the NuScenes [64] dataset, as other sensors do not provide multi-view image data.

7.4. Discussion

In 3D object detection, knowledge distillation from LiDAR to camera models has become a key approach for enhancing image-based perception. Existing methods primarily focus on guidance with LiDAR depth information and aligning image features with LiDAR features in the BEV space. These approaches benefit from the strong geometric supervision provided by LiDAR, enabling more accurate depth reasoning and improved spatial consistency in image-based representations. We summa-

rize these methods in Tables 11 and 12 and categorize the knowledge transfer approaches into four types: depth guidance, feature distillation, relation distillation, and response distillation. This taxonomy helps better illustrate the focus of each method, as shown in Fig. 21.

8. Multi-modal to single-modal knowledge distillation for 3D object detection

In 3D object detection, knowledge distillation from a multi-modal teacher [168,176]—typically fusing LiDAR and camera features—to a single-modal student (e.g., camera-only [128,129] or LiDAR-only [89,125]) has emerged as an effective approach to enhance the student model's perception capabilities. By leveraging the multi-modal teacher, the student can simultaneously acquire complementary knowledge, including visual semantics, precise geometry, and depth priors. This allows the student to overcome the inherent limitations of single-modal inputs in spatial understanding and depth perception, significantly improving accuracy and robustness in 3D detection tasks. Common approaches distill a LiDAR-camera fused multi-modal teacher into a LiDAR-based student (Section 8.1) or camera-based (Section 8.2), enabling the student to utilize multi-modal knowledge to optimize single-modal perception performance.

8.1. LiDAR-camera fusion to LiDAR

Although LiDAR-based [114,116] 3D object detection inherently achieves strong performance, it still lags behind multi-modal fusion approaches [14,171]. However, multi-modal methods incur higher floating-point computation and introduce additional redundant model parameters. Therefore, some studies focus on distilling knowledge from multi-modal fusion models into single-modal LiDAR models for 3D object detection.

VeXKD [248] focuses knowledge distillation on BEV feature maps, efficiently transferring semantic features from multi-modal teachers (e.g., LiDAR-Camera fusion) to single-modality student models (e.g., image-only). This avoids complex cross-modal feature-level alignment, simplifying the distillation process and making it easily adaptable to different tasks and model architectures. These methods demonstrate the effectiveness of BEV-level knowledge distillation in multi-modal 3D detection. MonoTAKD [28] designs a powerful camera-based temporal attention model and introduces cross-modal residual distillation to transfer spatial information from a LiDAR teacher model to the student. Through this dual-modality distillation, the camera-based student simultaneously ac-

Table 11

Performance of different methods from LiDAR-to-Camera knowledge distillation for 3D object detection. We sort these methods by publication year and group them according to the type of sensor input. The performance metrics on different datasets are reported based on the camera-based detectors as presented in the original papers.

Method	Publish Year	Teacher Modality	Student Modality	Backbone (Camera)	Depth guidance	Distillation category			KITTI Car			NuScenes		Waymo	
						Feature	Relation	Respond	Easy	Mod.	Hard	mAP	NDS	L1	L2
Monodtr [129]	CVPR2022	LiDAR	Mono.	GUPNet [256]	✓	✓		✓	21.9	15.3	12.7	–	–	–	–
MONODISTILL [218]	ICLR2022	LiDAR	Mono.	GUPNet		✓		✓	22.9	16.0	13.6	–	–	–	–
DID-M3D [257]	ECCV2022	LiDAR	Mono.	GUPNet	✓				24.4	16.2	13.7	–	–	20.6	19.3
CMKD [219]	ECCV2022	LiDAR	Mono.	CaDDN		✓		✓	25.0	16.9	15.3	–	–	14.5	12.9
MonoSKD [213]	ECAI2023	LiDAR	Mono.	GUPNet			✓	✓	28.4	17.3	15.0	–	–	–	–
MonoNeRD [258]	ICCV2023	LiDAR	Mono.	CaDDN [156]	✓				22.7	17.1	15.6	–	–	31.1	29.2
ADD [16]	AAAI2023	LiDAR	Mono.	CaDDN		✓		✓	25.6	16.8	13.7	–	–	–	–
OccupancyM3D [252]	CVPR2024	LiDAR	Mono.	DLA34 [259]	✓				25.5	17.0	14.7	–	–	28.9	27.2
MonoTAKD [28]	CVPR2025	LiDAR	Mono.	CaDDN		✓		✓	27.9	19.4	16.5	–	–	–	–
Tig-BEV [193]	Arxiv2022	LiDAR	Mul-view.	ResNet-101	✓	✓			–	–	–	44.0	54.4	–	–
Distillbev [222]	ICCV2023	LiDAR	Mul-view.	ResNet-101		✓		✓	–	–	–	45.0	54.7	–	–
STXD [17]	NeIPS2023	LiDAR	Mul-view.	BEVFormer [255]			✓	✓	–	–	–	46.5	55.5	–	–
FD3D [220]	CVPR2023	LiDAR	Mul-view.	BEVFormer		✓	✓		–	–	–	31.0	43.0	–	–
Unidistill [221]	CVPR2023	LiDAR	Mul-view.	ResNet-50		✓	✓	✓	–	–	–	29.6	39.3	–	–
EA-LSS [169]	CVPR2023	LiDAR	Mul-view.	ResNet-50	✓				–	–	–	33.4	44.1	–	–
SA-BEV [260]	ICCV2023	LiDAR	Mul-view.	BEVDepth [261]	✓				–	–	–	36.5	48.3	–	–
BEV-LGKD [18]	T-IV2024	LiDAR	Mul-view.	ResNet-18		✓			–	–	–	30.5	42.5	–	–
Fsd-bev [227]	ECCV2024	LiDAR	Mul-view.	ResNet-101	✓	✓			–	–	–	48.8	58.9	–	–
LgGp [223]	IJCV2025	LiDAR	Mul-view.	ResNet-50		✓		✓	–	–	–	38.4	49.1	–	–
MemDistill [253]	ICCV2025	LiDAR	Mul-view.	ResNet-50		✓			–	–	–	42.5	53.1	–	–

Table 12

A multi-dimensional comparison of knowledge distillation methods for 3D object detection in autonomous driving perception.

Method	Publish Year	Teacher Modality	Student Modality	Student Backbone	Deployment cost			Distillation space				Distillation target				Robustness setting		
					Para.	FLOPS	FPS	image	point	voxel	BEV	query	Feature	Relation	Logit	geometry	temporal	Sensor Fail. Noise
ItKD [210]	CVPR2023	LiDAR	LiDAR	Centerpoint [102]	20.5	1.4	42.7	✓	✓		✓		✓		✓			
Point-Distiller [211]	CVPR2023	LiDAR	LiDAR	Centerpoint	45.6	4.8	–		✓	✓				✓		✓		✓
RDD [235]	ICCV2023	LiDAR	LiDAR	Centerpoint	36.0	2.8	–	✓					✓		✓			✓
H2RKD [216]	MM2024	LiDAR	LiDAR	Centerpoint	45.8	4.6	–		✓			✓						✓
InstKD [233]	TIV2024	LiDAR	LiDAR	Centerpoint	39.0	3.3	–	✓			✓		✓		✓			
CaKDP [49]	CVPR2024	LiDAR	LiDAR	Centerpoint	30.2	1.5	–	✓				✓	✓		✓			✓
CLEAN [245]	TPAMI2025	LiDAR	LiDAR	Centerpoint	20.5	1.0	–	✓		✓		✓	✓		✓		✓	✓
Mono-TAKD [28]	CVPR2025	LiDAR	Camera	CaDDN [156]	45.1	41.3	–	✓			✓		✓		✓			✓
MemDistill [253]	ICCV2025	LiDAR	Camera	BEVDepth [140]	–	–	–		✓	✓	✓	✓			✓		✓	

quires visual knowledge and 3D spatial cues, significantly enhancing its detection performance.

8.2. LiDAR-camera fusion to camera

In 3D object detection, using a LiDAR-camera fused model [158, 165, 175, 176, 179, 180] as the teacher for a single-camera [131, 144–146, 179] student model is a key strategy for enhancing image-based percep-

tion. Through the multi-modal teacher, the student can learn not only visual semantic features but also precise geometric and depth information provided by LiDAR, thereby compensating for the shortcomings of monocular or multi-view images in spatial perception and depth understanding.

For instance, SimDistill [254] proposes a multi-modal knowledge distillation framework for BEV 3D object detection, where a LiDAR-Camera fused teacher guides a single-modality student. By leveraging

the rich geometric and semantic information encoded in the multi-modal teacher, the student learns to replicate both spatial structure and object-level semantics in BEV space using intra-modal, cross-modal, and fusion distillation strategies. VeXKD [248] introduces a flexible framework in which a multi-modal teacher fuses LiDAR and camera features in BEV space and transfers this knowledge to a single-sensor student. A modality-general fusion module bridges modality gaps, and a BEV query-guided mask focuses distillation on key regions, enabling the student to capture rich geometric and semantic cues without additional inference cost.

8.3. Discussion

Looking ahead, LiDAR-based multi-modal knowledge distillation to other modalities is expected to become increasingly important for 3D object detection, particularly as autonomous driving systems seek to balance perception performance with deployment efficiency. Future research may focus on more effective cross-modal feature alignment, robust distillation strategies for sparse or distant point clouds, and adaptive knowledge transfer that accounts for varying sensor quality and environmental conditions. Overall, this direction holds promise for bridging the gap between high-performance multi-modal perception and practical, low-cost sensor deployment [262].

9. Open challenges and future directions

In this section, we discuss the challenges faced by knowledge distillation and autonomous driving, as well as potential solutions, aiming to provide useful guidance for designing more effective distillation strategies and lightweight model architectures.

9.1. Feature differences of cross-modal distillation

Currently, knowledge distillation is mainly performed between LiDAR-LiDAR or LiDAR-Camera modalities. As multi-modal fusion becomes mainstream, distillation scenarios are expected to grow more complex, involving heterogeneous sensors such as LiDAR, Camera, Radar, and Event Cameras. For instance, RadarDistill [11] uses LiDAR-derived depth to mitigate radar limitations, while Ev-3DOD [189] fuses LiDAR with event camera data to address temporal gaps. However, the intrinsic modality gap—determined by sensor hardware—poses a fundamental challenge, making effective cross-modal knowledge transfer an ongoing problem. Two promising strategies may help overcome this challenge. First, instead of directly eliminating modality discrepancies, they can be circumvented by projecting all modalities into a unified representation space—such as BEV [217], voxel [211], or other intermediate feature spaces [216]—and then performing knowledge distillation efficiently within this shared space. Second, if modality differences fundamentally limit distillation effectiveness, the modalities themselves can be unified. This can be achieved via generative distillation approaches, for example, by using diffusion models [263] to synthesize pseudo-LiDAR point clouds from images, or by leveraging NeRF [264] or 3D Gaussian Splatting [265] to reconstruct multi-view representations from LiDAR data.

9.2. Limited capacity of the student model during distillation

In 3D object detection, teacher models typically use larger, more expressive architectures, such as multi-modal fusion networks [248] or two-stage detectors [49], while student models are often lightweight one-stage networks [233] with reduced capacity. This architectural gap makes it challenging for the student to fully absorb the teacher's high-dimensional or multi-modal knowledge. Moreover, limited student capacity can lead to rapid early decreases in distillation loss—"surface convergence"—before effectively learning from the teacher.

Table 13

The influence of adverse weather conditions on detection [7] performance on Waymo Open dataset. Precision and Recall are reported at IoU thresholds of 0.3, 0.5, and 0.7.

Weather	Precision			Recall		
	0.3	0.5	0.7	0.3	0.5	0.7
Sunny	46.8	44.8	34.6	87.3	84.1	69.7
Rainy	33.6	32.3	27.3	83.2	80.0	66.7
Degradation	-13.2	-12.5	-7.3	-4.1	-4.1	-3.0



Fig. 22. Various factors affecting autonomous driving in real-world environments.

To address this, selective distillation [266] transmits only critical channels or regions, enabling focused learning of the most informative knowledge. Dynamic distillation [234] adaptively weights transferred knowledge at each iteration, mitigating surface convergence. Progressive distillation [262] decomposes the teacher's knowledge by difficulty and importance, allowing the student to gradually learn representations and decision-making capabilities from simple to complex and from lower- to higher-level features.

9.3. Robust perception under adverse weather conditions

Most existing 3D object detection methods are developed and evaluated under relatively favorable conditions, such as clear weather and sufficient illumination. However, in real-world autonomous driving scenarios, perception systems are frequently challenged by adverse conditions, including rain, snow, fog, nighttime low-light environments, strong illumination variations (e.g., glare and overexposure), and dense traffic congestion, all of which degrade perception reliability through increased noise, occlusions, and incomplete geometric structures as shown in Fig. 22.

As summarized in Table 13, LiDAR and camera modalities are highly sensitive to weather and illumination variations, whereas radar-based sensing offers a more robust alternative due to its invariance to lighting conditions and resilience to atmospheric interference. In particular, 4D radar [267–270] provides both spatial and velocity information, making it well-suited for long-range and low-visibility scenarios. Consequently, radar-camera and radar-LiDAR fusion have emerged as effective directions for improving robustness in multi-modal 3D perception. Within this context, knowledge distillation [271] offers a lightweight strategy for enhancing robustness by transferring knowledge from teacher models trained under adverse or multi-condition settings.

Overall, extending current multi-modal fusion frameworks to radar-enhanced and collaborative perception settings holds significant potential for improving robustness and reliability in complex real-world driving environments, where radar provides complementary geometric and motion cues, and collaborative perception enables multi-agent information sharing to mitigate individual sensor limitations.

9.4. Toward more interpretable and theory-grounded distillation

Although current knowledge distillation methods for 3D object detection have achieved significant progress, their interpretability remains limited. Future research should focus on developing a more rigorous and systematic theoretical foundation for 3D spatial knowledge transfer, particularly in cross-modal and cross-domain distillation scenarios. This includes quantitatively analyzing the knowledge representations of teacher and student models in high-dimensional feature spaces, theoretically modeling cross-modal feature alignment, and mathematically characterizing information loss and model capacity constraints during distillation. Establishing such a comprehensive theoretical framework can not only improve training stability but also enhance the student model's ability to understand and absorb knowledge from the teacher, thereby substantially increasing the interpretability of the distillation process.

9.5. Distillation from large language models

Large language models (LLMs) [272,273] provide strong capabilities in semantic reasoning, world knowledge encoding, and high-level scene understanding, offering new opportunities for 3D object detection in autonomous driving. Beyond traditional visual or geometric supervision, LLMs can serve as external knowledge sources that enrich perception models with contextual reasoning, such as scene-level semantics, object relationships, and driving intent understanding. In future research, a key direction is to effectively distill such high-level semantic and reasoning knowledge from LLMs into LiDAR-, camera-, and multi-modal perception models. This is fundamentally different from conventional feature-level distillation, as it requires bridging heterogeneous representations between language-based reasoning spaces and geometric perception spaces. To achieve this, it is necessary to develop effective cross-modal alignment strategies that can map language-driven semantic knowledge into structured 3D representations such as BEV or object-level embeddings. In addition, selective distillation mechanisms will be critical to extract task-relevant reasoning signals while filtering out irrelevant or overly abstract knowledge from LLMs. Overall, distillation from LLMs represents a promising direction toward integrating semantic reasoning with 3D perception, enabling autonomous driving systems to move beyond pure geometric understanding toward more human-like scene interpretation and decision-aware perception.

9.6. Distillation for efficient deployment

Knowledge distillation plays a critical role in bridging the gap between high-performance models and real-world autonomous driving deployment. By transferring knowledge from large teacher models to compact student networks, it enables significant reductions in computational cost, memory footprint, and energy consumption, while retaining competitive perception performance. More importantly, knowledge distillation transforms the deployment paradigm of 3D perception systems. Instead of relying on heavy models that are difficult to deploy on edge devices, distilled student models can achieve real-time inference under strict latency constraints, making them suitable for onboard autonomous driving platforms. From a system perspective, this enables a seamless transition from algorithm development to practical deployment, where high-capacity models guide the design of lightweight yet effective perception networks. As a result, knowledge distillation provides a scalable pathway toward real-time, energy-efficient, and safety-critical autonomous driving systems operating reliably in complex real-world environments.

9.7. World models and vision-language-action systems for autonomous driving

Traditional autonomous driving systems are built on modular pipelines, where perception, prediction, and planning are decoupled and rely on intermediate representations such as bounding boxes, BEV features, or trajectories. Such systems are largely reactive and limited in modeling long-term action consequences. In contrast, future autonomous driving shifts toward more holistic intelligence. World models [203] enable prediction of future environment dynamics, allowing the system to simulate possible outcomes before action execution, thereby supporting proactive planning in complex scenarios. Building on this, Vision-Language-Action (VLA) systems [206] integrate semantic reasoning and natural language understanding into decision-making, enabling direct mapping from high-level instructions to executable actions. Complementary foundation models further enhance generalization through large-scale multimodal pretraining. From a knowledge distillation perspective, these paradigms shift the focus from low-level feature transfer toward high-level reasoning, planning, and policy transfer for more adaptive autonomous driving.

10. Conclusion

This paper provides a systematic review of 3D object detection techniques in autonomous driving and their integration with knowledge distillation. We begin by introducing the core sensors in autonomous driving perception systems and their significance for 3D detection tasks, analyzing the strengths and limitations of different modalities, including LiDAR, camera, and LiDAR-camera fusion. Next, we focus on three main directions of knowledge distillation in 3D object detection: first, distillation between LiDAR-to-LiDAR models, including both homogeneous and heterogeneous distillation, which serves as an effective means for model compression and enhancing the performance of lightweight models; second, distillation from LiDAR models to camera models, representing an important approach to significantly improve detection performance while reducing sensor costs; and third, distillation from LiDAR-camera fused models to single-modal (LiDAR or camera) models, providing a key method to further optimize single-modal perception capabilities. Finally, we review the open challenges in applying knowledge distillation to 3D object detection and provide an outlook on its future development directions.

CRediT authorship contribution statement

Weiqing Yan: Writing - review & editing, Validation, Supervision; **Changhong Bu:** Writing - original draft, Visualization, Conceptualization; **Yuanyang Zhang:** Writing - review & editing, Formal analysis; **Guanghui Yue:** Writing - review & editing; **Wei Zhou:** Writing - review & editing; **Xitian Zhu:** Writing - review & editing; **Jungong Han:** Writing - review & editing.

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.

Acknowledgment

This work was supported by the project of **National Natural Science Foundation of China** No. 62471426, the project No. ZR2024MF060 supported by Shandong Provincial Natural Science Foundation, the project of Youth Innovation Team Project of Higher Education Institutions in Shandong Province No. 2022KJ268, the Science and Technology Innovation Development Plan Projects of Yantai No. 2023JCYJ046.

Data availability

No data was used for the research described in the article.

References

- [1] J. Mao, S. Shi, X. Wang, H. Li, 3D object detection for autonomous driving: a comprehensive survey, *Int. J. Comput. Vis.* 131 (8) (2023) 1909–1963.
- [2] Z. Song, L. Liu, F. Jia, Y. Luo, C. Jia, G. Zhang, L. Yang, L. Wang, Robustness-aware 3d object detection in autonomous driving: a review and outlook, *IEEE Trans. Intell. Transp. Syst.* 25 (11) (2024) 15407–15436.
- [3] L. Gao, X. Xia, Z. Zheng, J. Ma, GNSS/IMU/LiDAR fusion for vehicle localization in urban driving environments within a consensus framework, *Mech. Syst. Signal Process.* 205 (2023) 110862.
- [4] Y. Zhao, Y. Liang, Z. Ma, L. Guo, H. Zhang, Localization and mapping algorithm based on lidar-imu-camera fusion, *J. Intell. Connect. Veh.* 7 (2) (2024) 97–107.
- [5] Z. Yang, J. Chen, Z. Miao, W. Li, X. Zhu, L. Zhang, Deepinteraction: 3D object detection via modality interaction, *Adv. Neural Inf. Process. Syst.* 35 (2022) 1992–2005.
- [6] Z. Yang, N. Song, W. Li, X. Zhu, L. Zhang, P.H.S. Torr, Deepinteraction++: multi-modality interaction for autonomous driving, *IEEE Trans. Pattern Anal. Mach. Intell.* 47 (8) (2025) 6749–6763.
- [7] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, H. Li, Voxel r-cnn: towards high performance voxel-based 3d object detection, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 35, 2021, pp. 1201–1209.
- [8] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, H. Li, PV-RCNN: point-voxel feature set abstraction for 3d object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10529–10538.
- [9] J. Philion, S. Fidler, Lift, split, shoot: encoding images from arbitrary camera rigs by implicitly unprojecting to 3d, in: *European Conference on Computer Vision*, Springer, 2020, pp. 194–210.
- [10] J. Huang, G. Huang, Z. Zhu, Y. Ye, D. Du, Bevdet: High-performance multi-camera 3d object detection in bird-eye-view, 2112.11790 (2021).
- [11] G. Bang, K. Choi, J. Kim, D. Kum, J.W. Choi, Radardistill: boosting radar-based object detection performance via knowledge distillation from lidar features, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15491–15500.
- [12] P. Palmer, M. Krüger, S. Schütte, R. Altendorfer, G. Adam, T. Bertram, LEROJD: lidar extended radar-only object detection, in: *European Conference on Computer Vision*, Springer, 2024, pp. 379–396.
- [13] Y. Wang, Q. Mao, H. Zhu, J. Deng, Y. Zhang, J. Ji, H. Li, Y. Zhang, Multi-modal 3D object detection in autonomous driving: a survey, *Int. J. Comput. Vis.* 131 (8) (2023) 2122–2152.
- [14] S. Vora, A.H. Lang, B. Helou, O. Beijbom, Pointpainting: sequential fusion for 3d object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4604–4612.
- [15] C. Wang, C. Ma, M. Zhu, X. Yang, Pointaugmenting: cross-modal augmentation for 3d object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11794–11803.
- [16] Z. Wu, Y. Wu, J. Pu, X. Li, X. Wang, Attention-based depth distillation with 3d-aware positional encoding for monocular 3d object detection, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 37, 2023, pp. 2892–2900.
- [17] S. Jang, D.U. Jo, S.J. Hwang, D. Lee, D. Ji, STXD: structural and temporal cross-modal distillation for multi-view 3D object detection, *Adv. Neural Inf. Process. Syst.* 36 (2023) 29323–29342.
- [18] J. Li, M. Lu, J. Liu, Y. Guo, Y. Du, L. Du, S. Zhang, BEV-LGKD: a unified lidar-guided knowledge distillation framework for multi-view bev 3d object detection, *IEEE Trans. Intell. Veh.* 9 (1) (2023) 2489–2498.
- [19] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, 1503.02531 (2015).
- [20] D. Walawalkar, Z. Shen, M. Savvides, Online ensemble model compression using knowledge distillation, in: *European Conference on Computer Vision*, Springer, 2020, pp. 18–35.
- [21] J. Gou, L. Sun, B. Yu, S. Wan, D. Tao, Hierarchical multi-attention transfer for knowledge distillation, *ACM Trans. Multimed. Comput. Commun. Appl.* 20 (2) (2023) 1–20.
- [22] S.S.S. Kruthiventi, P. Sahay, R. Biswal, Low-light pedestrian detection from RGB images using multi-modal knowledge distillation, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 4207–4211.
- [23] G. Chen, W. Choi, X. Yu, T. Han, M. Chandraker, Learning efficient object detection models with knowledge distillation, *Adv. Neural Inf. Process. Syst.* 30 (2017) 742–751.
- [24] J. Wang, Y. Chen, Z. Zheng, X. Li, M.-M. Cheng, Q. Hou, CrossKD: cross-head knowledge distillation for object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16520–16530.
- [25] Z. Yang, Z. Li, X. Jiang, Y. Gong, Z. Yuan, D. Zhao, C. Yuan, Focal and global knowledge distillation for detectors, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4643–4652.
- [26] Y. Lv, Y. Cai, Y. He, M. Li, DrKD: decoupling response-based distillation for object detection, *Pattern Recognit.* 161 (2025) 111275.
- [27] Y. Wang, A. Fathi, J. Wu, T. Funkhouser, J. Solomon, Multi-frame to single-frame: Knowledge distillation for 3D object detection, 2009.11859 (2020).
- [28] H.-I. Liu, C. Wu, J.-H. Cheng, W. Chai, S.-Y. Wang, G. Liu, H. Latapie, J.-C. Wu, J.-N. Hwang, H.-H. Shuai, et al., MonoTAKD: teaching assistant knowledge distillation for monocular 3D object detection, in: *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22266–22275.
- [29] J. Yang, S. Shi, R. Ding, Z. Wang, X. Qi, Towards efficient 3D object detection with knowledge distillation, *Adv. Neural Inf. Process. Syst.* 35 (2022) 21300–21313.
- [30] C. Deng, M. Wang, L. Liu, Y. Liu, Y. Jiang, Extended feature pyramid network for small object detection, *IEEE Trans. Multimed.* 24 (2021) 1968–1979.
- [31] C. Li, X. Qu, A. Gnanasambandam, O.A. Elgandy, J. Ma, S.H. Chan, Photon-limited object detection using non-local feature matching and knowledge distillation, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3976–3987.
- [32] J. Chen, S. Wang, L. Chen, H. Cai, Y. Qian, Incremental detection of remote sensing objects with feature pyramid and knowledge distillation, *IEEE Trans. Geosci. Remote Sens.* 60 (2020) 1–13.
- [33] Y. Yang, X. Sun, W. Diao, H. Li, Y. Wu, X. Li, K. Fu, Adaptive knowledge distillation for lightweight remote sensing object detectors optimizing, *IEEE Trans. Geosci. Remote Sens.* 60 (2022) 1–15.
- [34] L. Chen, C. Yu, L. Chen, A new knowledge distillation for incremental object detection, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–7.
- [35] D. Yang, Y. Zhou, D. Wu, C. Ma, F. Yang, W. Wang, Two-level residual distillation based triple network for incremental object detection, 2007.13428 (2020).
- [36] K.J. Joseph, J. Rajasegaran, S. Khan, F.S. Khan, V.N. Balasubramanian, Incremental object detection via meta-learning, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (12) (2021) 9209–9216.
- [37] J. Zhang, H. Su, W. Zou, X. Gong, Z. Zhang, F. Shen, CADN: a weakly supervised learning-based category-aware object detection network for surface defect detection, *Pattern Recognit.* 109 (2021) 107571.
- [38] Z. Zeng, B. Liu, J. Fu, H. Chao, L. Zhang, WSOD2: learning bottom-up and top-down objectness distillation for weakly-supervised object detection, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8292–8300.
- [39] H. Felix, W.M. Rodrigues, D. Macêdo, F. Simoes, A.L.L. Oliveira, V. Teichrieb, C. Zanchettin, Squeezed deep 6dof object detection using knowledge distillation, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–7.
- [40] Z. Qin, J. Wang, Y. Lu, Weakly supervised 3d object detection from point clouds, in: *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4144–4152.
- [41] Y. Zhu, C. Zhao, C. Han, J. Wang, H. Lu, Mask guided knowledge distillation for single shot detector, in: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2019, pp. 1732–1737.
- [42] P. Chen, S. Liu, H. Zhao, J. Jia, Distilling knowledge via knowledge review, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5008–5017.
- [43] Y. Piao, Z. Rong, M. Zhang, W. Ren, H. Lu, A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9060–9069.
- [44] F.R. Valverde, J.V. Hurtado, A. Valada, There is more than meets the eye: self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11612–11621.
- [45] M. Kang, J. Mun, B. Han, Towards oracle knowledge distillation with neural architecture search, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 2020, pp. 4404–4411.
- [46] X. Chen, J. Su, J. Zhang, A two-teacher framework for knowledge distillation, in: *International Symposium on Neural Networks*, Springer, 2019, pp. 58–66.
- [47] Z. Wu, Z. Hu, Object detection based on self feature distillation, in: *Journal of Physics: Conference Series*, 1982, IOP Publishing, 2021, p. 012081.
- [48] K. Kim, B. Ji, D. Yoon, S. Hwang, Self-knowledge distillation with progressive refinement of targets, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6567–6576.
- [49] H. Zhang, L. Liu, Y. Huang, Z. Yang, X. Lei, B. Wen, Cakdp: category-aware knowledge distillation and pruning framework for lightweight 3d object detection, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2024, pp. 15331–15341.
- [50] L. Wang, X. Zhang, Z. Song, J. Bi, G. Zhang, H. Wei, L. Tang, L. Yang, J. Li, C. Jia, et al., Multi-modal 3D object detection in autonomous driving: a survey and taxonomy, *IEEE Trans. Intell. Veh.* 8 (7) (2023) 3781–3798.
- [51] Z. Li, P. Xu, X. Chang, L. Yang, Y. Zhang, L. Yao, X. Chen, When object detection meets knowledge distillation: a survey, *IEEE Trans. Pattern Anal. Mach. Intell.* 45 (8) (2023) 10555–10579.
- [52] S. Jiang, Z. Huang, K. Qian, Z. Luo, T. Zhu, Y. Zhong, Y. Tang, M. Kong, Y. Wang, S. Jiao, et al., A survey on vision-language-action models for autonomous driving, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 4524–4536.
- [53] Y. Gong, N. Wang, J. Lu, X. Zhang, Y. Gao, J. Zhao, Z. Huang, H. Bai, N. Zeng, N. Su, et al., Progressive bird's eye view perception for safety-critical autonomous driving: a comprehensive survey, 2508.07560 (2025).
- [54] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3354–3361.
- [55] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, R. Yang, The apollocape dataset for autonomous driving, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 954–960.
- [56] A. Patil, S. Malla, H. Gang, Y.-T. Chen, The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 9552–9557.

- [57] P. Xiao, Z. Shao, S. Hao, Z. Zhang, X. Chai, J. Jiao, Z. Li, J. Wu, K. Sun, K. Jiang, et al., PandaSet: advanced sensor suite dataset for autonomous driving, in: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), IEEE, 2021, pp. 3095–3101.
- [58] Y. Liao, J. Xie, A. Geiger, Kitti-360: a novel dataset and benchmarks for urban scene understanding in 2d and 3d, *IEEE Trans. Pattern Anal. Mach. Intell.* 45 (3) (2022) 3292–3310.
- [59] J. Geyer, Y. Kassahun, M. Mahmudi, X. Ricou, R. Durgesh, A.S. Chung, L. Hauswald, V.H. Pham, M. Mühlegg, S. Dorn, et al., A2d2: audi autonomous driving dataset, 2004.06320 (2020).
- [60] L. Chen, C. Sima, Y. Li, Z. Zheng, J. Xu, X. Geng, H. Li, C. He, J. Shi, Y. Qiao, et al., Persformer: 3D lane detection via perspective transformer and the open-lane benchmark, in: *European Conference on Computer Vision*, Springer, 2022, pp. 550–567.
- [61] R. Xu, X. Xia, J. Li, H. Li, S. Zhang, Z. Tu, Z. Meng, H. Xiang, X. Dong, R. Song, et al., V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13712–13722.
- [62] W. Zimmer, G.A. Wardana, S. Sritharan, X. Zhou, R. Song, A.C. Knoll, Tumtraf v2x cooperative perception dataset, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22668–22677.
- [63] Y. Hou, B. Zou, M. Zhang, R. Chen, S. Yang, Y. Zhang, J. Zhuo, S. Chen, J. Chen, H. Ma, AGC-Drive: a large-scale dataset for real-world aerial-ground collaboration in driving scenarios, 2506.16371 (2025).
- [64] H. Caesar, V. Bankiti, A.H. Lang, S. Vora, V.E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, Nusenes: a multimodal dataset for autonomous driving, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11621–11631.
- [65] P. Sun, H. Kretschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al., Scalability in perception for autonomous driving: Waymo open dataset, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2446–2454.
- [66] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, et al., Argoverse: 3D tracking and forecasting with rich maps, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8748–8757.
- [67] Z. Wang, S. Ding, Y. Li, J. Fenn, S. Roychowdhury, A. Wallin, L. Martin, S. Rylvola, G. Sapiro, Q. Qiu, Cirrus: a long-range bi-pattern lidar dataset, in: 2021 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2021, pp. 5744–5750.
- [68] J. Mao, M. Niu, C. Jiang, H. Liang, J. Chen, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li, et al., One million scenes for autonomous driving: Once dataset, 2106.11037 (2021).
- [69] M. Alibeigi, W. Ljungbergh, A. Tonderski, G. Hess, A. Lilja, C. Lindström, D. Motorniu, J. Fu, J. Widahl, C. Petersson, Zenseact open dataset: a large-scale and diverse multimodal dataset for autonomous driving, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20178–20188.
- [70] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J.K. Pontes, et al., Argoverse 2: next generation datasets for self-driving perception and forecasting, 2301.00493 (2023).
- [71] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- [72] T.Y. Lin, M. Maire, S. Belongie, J. Hays, C.L. Zitnick, Microsoft COCO: common objects in context, *Springer Int. Publ.* 8693 (Part V) (2014) 740–755.
- [73] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [74] M. Everingham, L.V. Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, *Int. J. Comput. Vis.* 88 (2) (2010) 303–338.
- [75] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene parsing through ADE20K dataset, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [76] X. Huang, J. Wang, Q. Xia, S. Chen, B. Yang, X. Li, C. Wang, C. Wen, V2X-R: cooperative LiDAR-4D radar fusion with denoising diffusion for 3D object detection, in: *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27390–27400.
- [77] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A.S. Ecker, M. Bethge, W. Brendel, Benchmarking robustness in object detection: Autonomous driving when winter is coming, 1907.07484 (2019).
- [78] N. Karnchanachari, D. Geromichalos, K.S. Tan, N. Li, C. Eriksen, S. Yaghoubi, N. MehdiPour, G. Bernasconi, W.K. Fong, Y. Guo, et al., Towards learning-based planning: the nuplan benchmark for real-world autonomous driving, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 629–636.
- [79] Z. Song, L. Yang, S. Xu, L. Liu, D. Xu, C. Jia, F. Jia, L. Wang, Graphbev: towards robust bev feature alignment for multi-modal 3d object detection, in: *European Conference on Computer Vision*, Springer, 2024, pp. 347–366.
- [80] C.R. Qi, H. Su, K. Mo, L.J. Guibas, PointNet: deep learning on point sets for 3d classification and segmentation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 652–660.
- [81] X. Chen, K. Kundu, Y. Zhu, A.G. Berneshawi, H. Ma, S. Fidler, R. Urtasun, 3D object proposals for accurate object class detection, *Adv. Neural Inf. Process. Syst.* 28 (2015) 424–432.
- [82] Y. Chen, L. Tai, K. Sun, M. Li, Monopair: monocular 3d object detection using pairwise spatial relationships, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12093–12102.
- [83] Z. Wu, Y. Gan, Y. Wu, R. Wang, X. Wang, J. Pu, Fd3D: exploiting foreground depth map for feature-supervised monocular 3D object detection, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 38, 2024, pp. 6189–6197.
- [84] X. Guo, S. Shi, X. Wang, H. Li, LIGA-STEREO: learning lidar geometry aware representations for stereo-based 3d detector, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3153–3163.
- [85] Z. Li, Z. Yu, W. Wang, A. Anandkumar, T. Lu, J.M. Alvarez, FB-Bev: bev representation from forward-backward view transformations, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6919–6928.
- [86] P. Li, W. Shen, Q. Huang, D. Cui, DualBEV: unifying dual view transformation with probabilistic correspondences, in: *European Conference on Computer Vision*, Springer, 2024, pp. 286–302.
- [87] J. Zhang, Y. Zhang, Y. Qi, Z. Fu, Q. Liu, Y. Wang, GeoBEV: learning geometric bev representation for multi-view 3d object detection, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 39, 2025, pp. 9960–9968.
- [88] Y. Yan, Y. Mao, B. Li, Second: sparsely embedded convolutional detection, *Sensors* 18 (10) (2018) 3337.
- [89] Y. Chen, J. Liu, X. Zhang, X. Qi, J. Jia, Voxelnex: fully sparse voxelnet for 3d object detection and tracking, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21674–21683.
- [90] X. Jin, H. Su, K. Liu, C. Ma, W. Wu, F. Hui, J. Yan, UniMamba: unified spatial-channel representation learning with group-efficient mamba for LiDAR-based 3D object detection, in: *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1407–1417.
- [91] G. Zhang, L. Fan, C. He, Z. Lei, Z.-X. ZHANG, L. Zhang, Voxel Mamba: group-free state space models for point cloud based 3d object detection, *Adv. Neural Inf. Process. Syst.* 37 (2024) 81489–81509.
- [92] S. Shi, X. Wang, H. Li, Pointcnn: 3d object proposal generation and detection from point cloud, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 770–779.
- [93] W. Zheng, W. Tang, S. Chen, L. Jiang, C.-W. Fu, CIA-SSD: confident iou-aware single-stage object detector from point cloud, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 35, 2021, pp. 3555–3562.
- [94] Z. Yang, Y. Sun, S. Liu, X. Shen, J. Jia, Std: sparse-to-dense 3D object detector for point cloud, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1951–1960.
- [95] Z. Yang, Y. Sun, S. Liu, J. Jia, 3Dssd: point-based 3d single stage object detector, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11040–11048.
- [96] W. Shi, R. Rajkumar, Point-Gnn: graph neural network for 3d object detection in a point cloud, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1711–1719.
- [97] X. Pan, Z. Xia, S. Song, L.E. Li, G. Huang, 3D object detection with pointformer, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7463–7472.
- [98] Y. Zhang, Q. Hu, G. Xu, Y. Ma, J. Wan, Y. Guo, Not all points are equal: learning highly efficient point-based detectors for 3d lidar point clouds, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18953–18962.
- [99] Q. He, Z. Wang, H. Zeng, Y. Zeng, Y. Liu, Svga-Net: sparse voxel-graph attention network for 3d object detection from point clouds, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, 2022, pp. 870–878.
- [100] C. He, R. Li, Y. Zhang, S. Li, L. Zhang, Msf: motion-guided sequential fusion for efficient 3d object detection from point cloud sequences, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5196–5205.
- [101] K.-C. Huang, W. Lyu, M.-H. Yang, Y.-H. Tsai, Ptt: point-trajectory transformer for efficient temporal 3d object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14938–14947.
- [102] T. Yin, X. Zhou, P. Krahenbuhl, Center-based 3D object detection and tracking, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11784–11793.
- [103] G. Shi, R. Li, C. Ma, PillarNet: real-time and high-performance pillar-based 3d object detection, in: *European Conference on Computer Vision*, Springer, 2022, pp. 35–52.
- [104] Y. Chen, J. Liu, X. Zhang, X. Qi, J. Jia, LargeKernel3d: scaling up kernels in 3d sparse cnns, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13488–13498.
- [105] G. Zhang, J. Chen, G. Gao, J. Li, S. Liu, X. Hu, SafdNet: a simple and effective network for fully sparse 3d object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14477–14486.
- [106] Z. Liu, J. Hou, X. Ye, T. Wang, J. Wang, X. Bai, Seed: a simple and effective 3d detr in point clouds, in: *European Conference on Computer Vision*, Springer, 2024, pp. 110–126.
- [107] Z. Liu, J. Hou, X. Wang, X. Ye, J. Wang, H. Zhao, X. Bai, Lion: linear group rnn for 3d object detection in point clouds, *Adv. Neural Inf. Process. Syst.* 37 (2024) 13601–13626.
- [108] J.-K. Lee, J.-H. Lee, J. Lee, S. Kwon, H. Jung, Re-VoxelDet: rethinking neck and head architectures for high-performance voxel-based 3d detection, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7503–7512.

- [109] L. Liu, Z. Song, Q. Xia, F. Jia, C. Jia, L. Yang, Y. Gong, H. Pan, SparseDet: a simple and effective framework for fully sparse lidar-based 3D object detection, *IEEE Trans. Geosci. Remote Sens.* 62 (5707114) (2024) 1–14.
- [110] S.A. Baur, F. Moosmann, A. Geiger, Liso: lidar-only self-supervised 3d object detection, in: *European Conference on Computer Vision*, Springer, 2024, pp. 253–270.
- [111] H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, S. Han, Searching efficient 3d architectures with sparse point-voxel convolution, in: *European Conference on Computer Vision*, Springer, 2020, pp. 685–702.
- [112] H. Sheng, S. Cai, Y. Liu, B. Deng, J. Huang, X.-S. Hua, M.-J. Zhao, Improving 3D object detection with channel-wise transformer, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2743–2752.
- [113] C. He, H. Zeng, J. Huang, X.-S. Hua, L. Zhang, Structure aware single-stage 3D object detection from point cloud, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11873–11882.
- [114] S. Shi, L. Jiang, J. Deng, Z. Wang, C. Guo, J. Shi, X. Wang, H. Li, PV-RCNN++: point-voxel feature set abstraction with local vector representation for 3D object detection, *Int. J. Comput. Vis.* 131 (2) (2023) 531–551.
- [115] J. Cao, C. Tao, Z. Zhang, Z. Gao, X. Luo, S. Zheng, Y. Zhu, Accelerating point-voxel representation of 3-D object detection for automatic driving, *IEEE Trans. Artif. Intell.* 5 (1) (2023) 254–266.
- [116] Z. Song, H. Wei, C. Jia, Y. Xia, X. Li, C. Zhang, VP-Net: voxels as points for 3-D object detection, *IEEE Trans. Geosci. Remote Sens.* 61 (2023) 1–12.
- [117] Z. Leng, P. Sun, T. He, D. Anguelov, M. Tan, Pvrtransformer: point-to-voxel transformer for scalable 3d object detection, in: *2024 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2024, pp. 4238–4244.
- [118] J.S.K. Hu, T. Kuai, S.L. Waslander, Point density-aware voxels for lidar 3d object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8469–8478.
- [119] C.R. Qi, L. Yi, H. Su, L.J. Guibas, PointNet++: deep hierarchical feature learning on point sets in a metric space, *Adv. Neural Inf. Process. Syst.* 30 (2017) 5099–5108.
- [120] S. Xu, D. Zhou, J. Fang, J. Yin, Z. Bin, L. Zhang, Fusionpainting: multimodal fusion with adaptive attention for 3d object detection, in: *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, IEEE, 2021, pp. 3047–3054.
- [121] Y. Wang, J.M. Solomon, Object dgcnn: 3D object detection using dynamic graphs, *Adv. Neural Inf. Process. Syst.* 34 (2021) 20745–20758.
- [122] X. Wu, L. Peng, H. Yang, L. Xie, C. Huang, C. Deng, H. Liu, D. Cai, Sparse fuse dense: towards high quality 3d detection with depth completion, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5418–5427.
- [123] A.H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, 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. 12697–12705.
- [124] J. Zhang, J. Liu, Voxel-to-Pillar: knowledge distillation of 3d object detection in point cloud, in: *Proceedings of the 4th European Symposium on Software Engineering*, 2023, pp. 99–104.
- [125] Y. Zhou, O. Tuzel, VoxelNet: end-to-end learning for point cloud based 3d object detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499.
- [126] Y. Chen, Z. Yu, Y. Chen, S. Lan, A. Anandkumar, J. Jia, J.M. Alvarez, FocalFormer3d: focusing on hard instance for 3d object detection, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8394–8405.
- [127] A. Simonelli, S.R. Bulo, L. Porzi, M. López-Antequera, P. Kotschieder, Disentangling monocular 3d object detection, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1991–1999.
- [128] X. Ma, S. Liu, Z. Xia, H. Zhang, X. Zeng, W. Ouyang, Rethinking pseudo-lidar representation, in: *European Conference on Computer Vision*, Springer, 2020, pp. 311–327.
- [129] K.-C. Huang, T.-H. Wu, H.-T. Su, W.H. Hsu, Monodtr: monocular 3d object detection with depth-aware transformer, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4012–4021.
- [130] R. Zhang, H. Qiu, T. Wang, Z. Guo, Z. Cui, Y. Qiao, H. Li, P. Gao, MonoDETR: depth-guided transformer for monocular 3D object detection, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9155–9166.
- [131] C. Huang, T. He, H. Ren, W. Wang, B. Lin, D. Cai, OBMO: One bounding box multiple objects for monocular 3D object detection, *IEEE Trans. Image Process.* 32 (2023) 6570–6581.
- [132] L. Yan, P. Yan, S. Xiong, X. Xiang, Y. Tan, MonoCD: monocular 3d object detection with complementary depths, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10248–10257.
- [133] Y. Oh, H.-I. Kim, S.T. Kim, J.U. Kim, MonoWAD: weather-adaptive diffusion model for robust monocular 3d object detection, in: *European Conference on Computer Vision*, Springer, 2024, pp. 326–345.
- [134] F. Pu, Y. Wang, J. Deng, W. Yang, MonoDGP: monocular 3D object detection with decoupled-query and geometry-error priors, in: *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6520–6530.
- [135] P. Li, X. Chen, S. Shen, Stereo r-cnn based 3D object detection for autonomous driving, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7644–7652.
- [136] A.D. Pon, J. Ku, C. Li, S.L. Waslander, Object-centric stereo matching for 3d object detection, in: *2020 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2020, pp. 8383–8389.
- [137] S. Li, Z. Liu, Z. Shen, K.-T. Cheng, Stereo neural vernier caliper, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, 2022, pp. 1376–1385.
- [138] Z. Liu, X. Ye, X. Tan, E. Ding, X. Bai, Stereodistill: pick the cream from lidar for distilling stereo-based 3D object detection, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 37, 2023, pp. 1790–1798.
- [139] Y. Wang, V.C. Guizilini, T. Zhang, Y. Wang, H. Zhao, J. Solomon, Detr3D: 3D object detection from multi-view images via 3d-to-2d queries, in: *Conference on Robot Learning*, PMLR, 2022, pp. 180–191.
- [140] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, Z. Li, BEVDDepth: acquisition of reliable depth for multi-view 3d object detection, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 37, 2023, pp. 1477–1485.
- [141] Y. Li, H. Bao, Z. Ge, J. Yang, J. Sun, Z. Li, Bevstereo: enhancing depth estimation in multi-view 3d object detection with temporal stereo, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 37, 2023, pp. 1486–1494.
- [142] Z. Li, S. Lan, J.M. Alvarez, Z. Wu, BEVNEXt: reviving dense bev frameworks for 3d object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20113–20123.
- [143] F. Liu, T. Huang, Q. Zhang, H. Yao, C. Zhang, F. Wan, Q. Ye, Y. Zhou, Ray denoising: depth-aware hard negative sampling for multi-view 3d object detection, in: *European Conference on Computer Vision*, Springer, 2024, pp. 200–217.
- [144] Z. Chen, Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, F. Wu, F. Zhao, Graph-Detr4d: spatio-temporal graph modeling for multi-view 3d object detection, *IEEE Trans. Image Process.* 33 (2024) 4488–4500.
- [145] Z. Xue, M. Guo, H. Fan, S. Zhang, Z. Zhang, CorrBEV: multi-view 3D object detection by correlation learning with multi-modal prototypes, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 27413–27423.
- [146] R. Yasarla, S. Han, H. Cai, F. Porikli, DySS: dynamic queries and state-space learning for efficient 3D object detection from multi-camera videos, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025, pp. 2535–2544.
- [147] H. Liu, Y. Teng, T. Lu, H. Wang, L. Wang, Sparsebev: high-performance sparse 3D object detection from multi-camera videos, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18580–18590.
- [148] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, R. Urtasun, Monocular 3D object detection for autonomous driving, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2147–2156.
- [149] T. He, S. Soatto, Mono3D++: monocular 3d vehicle detection with two-scale 3d hypotheses and task priors, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 2019, pp. 8409–8416.
- [150] H. Lin, Y. Zhang, S. Niu, S. Cui, Z. Li, MonoTTA: fully test-time adaptation for monocular 3d object detection, in: *European Conference on Computer Vision*, Springer, 2024, pp. 96–114.
- [151] W. Peng, H. Pan, H. Liu, Y. Sun, Ida-3D: instance-depth-aware 3d object detection from stereo vision for autonomous driving, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13015–13024.
- [152] Z. Qin, J. Wang, Y. Lu, Triangulation learning network: from monocular to stereo 3d object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7615–7623.
- [153] L. Huang, Z. Li, C. Sima, W. Wang, J. Wang, Y. Qiao, H. Li, Leveraging vision-centric multi-modal expertise for 3d object detection, *Adv. Neural Inf. Process. Syst.* 36 (2023) 38504–38519.
- [154] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, F. Zhao, Bevdistill: cross-modal bev distillation for multi-view 3d object detection, 2211.09386 (2022).
- [155] L. Yang, K. Yu, T. Tang, J. Li, K. Yuan, L. Wang, X. Zhang, P. Chen, BEVHeight: a robust framework for vision-based roadside 3d object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21611–21620.
- [156] C. Reading, A. Harakeh, J. Chae, S.L. Waslander, Categorical depth distribution network for monocular 3d object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8555–8564.
- [157] T. Yin, X. Zhou, P. Krähenbühl, Multimodal virtual point 3d detection, *Adv. Neural Inf. Process. Syst.* 34 (2021) 16494–16507.
- [158] Z. Liu, T. Huang, B. Li, X. Chen, X. Wang, X. Bai, EPNet++: cascade bi-directional fusion for multi-modal 3D object detection, *IEEE Trans. Pattern Anal. Mach. Intell.* 45 (7) (2022) 8324–8341.
- [159] J. Ku, M. Mozifian, J. Lee, A. Harakeh, S.L. Waslander, Joint 3d proposal generation and object detection from view aggregation, in: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 1–8.
- [160] G. Zhang, Z. Song, L. Liu, Z. Ou, FGU3R: fine-grained fusion via unified 3D representation for multimodal 3D object detection, in: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.
- [161] G. Zhang, J. Xie, L. Liu, Z. Wang, K. Yang, Z. Song, Urformer: unified representation lidar-camera 3D object detection with transformer, in: *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, Springer, 2023, pp. 401–413.
- [162] S. Pang, D. Morris, H. Radha, CLOCs: Camera-LiDAR object candidates fusion for 3D object detection, in: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2020, pp. 10386–10393.
- [163] S. Pang, D. Morris, H. Radha, Fast-CLOCs: fast camera-LiDAR object candidates fusion for 3D object detection, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 187–196.
- [164] A. Paigwar, D. Sierra-Gonzalez, Ö. Erkent, C. Laugier, Frustum-pointpillars: a multi-stage approach for 3d object detection using rgb camera and lidar, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2926–2933.

- [165] S. Xu, F. Li, Z. Song, J. Fang, S. Wang, Z.-X. Yang, Multi-sem fusion: multimodal semantic fusion for 3-D object detection, *IEEE Trans. Geosci. Remote Sens.* 62 (2024) 1–14.
- [166] G. Xie, Z. Chen, M. Gao, M. Hu, X. Qin, PPF-Det: Point-pixel fusion for multi-modal 3D object detection, *IEEE Trans. Intell. Transp. Syst.* 25 (6) (2024) 5598–5611.
- [167] M. Simon, K. Amende, A. Kraus, J. Honer, T. Samann, H. Kaulbersch, S. Milz, H. Michael Gross, Complexer-yolo: real-time 3D object detection and tracking on semantic point clouds, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0.
- [168] M. Liang, B. Yang, Y. Chen, R. Hu, R. Urtaun, Multi-task multi-sensor fusion for 3D object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7345–7353.
- [169] H. Hu, F. Wang, J. Su, Y. Wang, L. Hu, W. Fang, J. Xu, Z. Zhang, Ea-lss: Edge-aware lift-splat-shot framework for 3d bev object detection, 2303.17895 (2023).
- [170] C.R. Qi, W. Liu, C. Wu, H. Su, L.J. Guibas, Frustum pointnets for 3d object detection from rgb-d data, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 918–927.
- [171] Z. Wang, K. Jia, Frustum convnet: sliding frustums to aggregate local point-wise features for amodal 3D object detection, in: *2019IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2019, pp. 1742–1749.
- [172] K. Shin, Y.P. Kwon, M. Tomizuka, RoarNet: a robust 3D object detection based on region approximation refinement, in: *2019 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2019, pp. 2510–2515.
- [173] M. Liang, B. Yang, S. Wang, R. Urtaun, Deep continuous fusion for multi-sensor 3d object detection, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 641–656.
- [174] J.H. Yoo, Y. Kim, J. Kim, J.W. Choi, 3D-cvf: generating joint camera and lidar features using cross-view spatial feature fusion for 3D object detection, in: *European Conference on Computer Vision*, Springer, 2020, pp. 720–736.
- [175] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, C.-L. Tai, TransFusion: robust lidar-camera fusion for 3d object detection with transformers, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1090–1099.
- [176] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, S. Han, Bevfusion: multi-task multi-sensor fusion with unified bird's-eye view representation, 2205.13542 (2022).
- [177] X. Li, T. Ma, Y. Hou, B. Shi, Y. Yang, Y. Liu, X. Wu, Q. Chen, Y. Li, Y. Qiao, et al., LogoNet: towards accurate 3d object detection with local-to-global cross-modal fusion, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17524–17534.
- [178] Y. Li, L. Fan, Y. Liu, Z. Huang, Y. Chen, N. Wang, Z. Zhang, Fully sparse fusion for 3D object detection, *IEEE Trans. Pattern Anal. Mach. Intell.* 46 (11) (2024) 7217–7231.
- [179] H. Zhang, L. Liang, P. Zeng, X. Song, Z. Wang, SparseLIF: high-performance sparse LiDAR-camera fusion for 3D object detection, in: *European Conference on Computer Vision*, Springer, 2024, pp. 109–128.
- [180] J. Huang, Y. Ye, Z. Liang, Y. Shan, D. Du, Detecting as labeling: rethinking lidar-camera fusion in 3d object detection, in: *European Conference on Computer Vision*, Springer, 2024, pp. 439–455.
- [181] J. Yin, J. Shen, R. Chen, W. Li, R. Yang, P. Frossard, W. Wang, Is-Fusion: instance-scene collaborative fusion for multimodal 3d object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14905–14915.
- [182] Y. Lee, H.-M. Jeong, Y. Jeon, S. Kim, EVT: efficient view transformation for multi-modal 3D object detection, 2411.10715 (2024).
- [183] H. Wang, J. Gao, W. Hu, Z. Zhang, Height-fidelity dense global fusion for multi-modal 3D object detection, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 26664–26674.
- [184] H. Liu, Z. Xu, D. Wang, B. Zhang, G. Wang, B. Dong, X. Wen, X. Xu, Pai3D: painting adaptive instance-prior for 3d object detection, in: *European Conference on Computer Vision*, Springer, 2022, pp. 459–475.
- [185] Y. Xie, C. Xu, M.-J. Rakotosaona, P. Rim, F. Tombari, K. Keutzer, M. Tomizuka, W. Zhan, Sparsefusion: fusing multi-modal sparse representations for multi-sensor 3D object detection, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17591–17602.
- [186] Z. Yu, B. Qiu, A.W.H. Khong, ViKIENet: towards efficient 3D object detection with virtual key instance enhanced network, in: *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11844–11853.
- [187] J. Yan, Y. Liu, J. Sun, F. Jia, S. Li, T. Wang, X. Zhang, Cross modal transformer: towards fast and robust 3d object detection, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18268–18278.
- [188] Y. Jiang, L. Zhang, Z. Miao, X. Zhu, J. Gao, W. Hu, Y.-G. Jiang, Polarformer: multi-camera 3D object detection with polar transformer, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 37, 2023, pp. 1042–1050.
- [189] H. Cho, J.-y. Kang, Y. Kim, K.-J. Yoon, Ev-3DOD: pushing the temporal boundaries of 3d object detection with event cameras, in: *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27197–27210.
- [190] Z. Song, G. Zhang, L. Liu, L. Yang, S. Xu, C. Jia, F. Jia, L. Wang, RoboFusion: towards robust multi-modal 3D object detection via SAM, 2401.03907 (2024).
- [191] Y. Chen, G. Cai, Z. Song, Z. Liu, B. Zeng, J. Li, Z. Wang, LVP: leverage virtual points in multimodal early fusion for 3-D object detection, *IEEE Trans. Geosci. Remote Sens.* 63 (2024) 1–15.
- [192] F. Jia, C. Jia, A. Liu, S. Xu, Q. Xia, L. Liu, L. Yang, Y. Gong, Z. Song, DGFusion: dual-guided fusion for robust multi-modal 3D object detection, *IEEE Trans. Circuits Syst. Video Technol.* 36 (4) (2026) 4956–4970.
- [193] P. Huang, L. Liu, R. Zhang, S. Zhang, X. Xu, B. Wang, G. Liu, Tig-Bev: multi-view bev 3D object detection via target inner-geometry learning, 2212.13979 (2022).
- [194] D. Xu, D. Anguelov, A. Jain, PointFusion: deep sensor fusion for 3d bounding box estimation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 244–253.
- [195] Y. Li, A.W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q.V. Le, et al., DeepFusion: lidar-camera deep fusion for multi-modal 3d object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17182–17191.
- [196] Z. Song, H. Wei, L. Bai, L. Yang, C. Jia, GraphAlign: enhancing accurate feature alignment by graph matching for multi-modal 3D object detection, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3358–3369.
- [197] Z. Song, C. Jia, L. Yang, H. Wei, L. Liu, GraphAlign++: an accurate feature alignment by graph matching for multi-modal 3D object detection, *IEEE Trans. Circuits Syst. Video Technol.* 34 (4) (2023) 2619–2632.
- [198] S. Xu, F. Li, S. Jiang, Z. Song, L. Liu, Z.-x. Yang, Gaussianpretrain: a simple unified 3D Gaussian representation for visual pre-training in autonomous driving, 2411.12452 (2024).
- [199] L. Fan, Z. Pang, T. Zhang, Y.-X. Wang, H. Zhao, F. Wang, N. Wang, Z. Zhang, Embracing single stride 3D object detector with sparse transformer, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8458–8468.
- [200] B. Sun, Y. Cao, Y. Wang, R. Wang, J. Shang, X. Feng, J. Lu, J. Shi, S. Yang, X. Yan, et al., MindDrive: an all-in-one framework bridging world models and vision-language model for end-to-end autonomous driving, 2512.04441 (2025).
- [201] Y. Luo, F. Li, Q. Chen, S. Xu, J. Liu, Z. Song, Z.-x. Yang, F. Wen, Unleashing via potentials in autonomous driving via explicit learning from failures, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 24833–24842.
- [202] L. Liu, C. Jia, G. Yu, Z. Song, J. Li, F. Jia, P. Wu, X. Hao, Y. Luo, GuideFlow: constraint-guided flow matching for planning in end-to-end autonomous driving, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 3719–3728.
- [203] X. Sun, S. Wang, F. Zhang, L. Liu, C. Jia, Z. Song, Z. Huang, Y. Luo, Vggt-world: transforming vggt into an autoregressive geometry world model, 2603.12655 (2026).
- [204] Z. Song, C. Jia, L. Liu, H. Pan, Y. Zhang, J. Wang, X. Zhang, S. Xu, L. Yang, Y. Luo, Don't shake the wheel: momentum-aware planning in end-to-end autonomous driving, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 22432–22441.
- [205] S. Zuo, Z. Xie, W. Zheng, S. Xu, F. Li, H. Li, L. Chen, Z.-X. Yang, J. Lu, DVGT-2: Vision-geometry-action model for autonomous driving at scale, 2604.00813 (2026).
- [206] Y. Luo, F. Li, S. Xu, Y. Ji, Z. Zhang, B. Wang, Y. Shen, J. Cui, L. Chen, G. Chen, et al., LAST-VLA: Thinking in latent spatio-temporal space for vision-language-action in autonomous driving, 2603.01928 (2026).
- [207] Y. Yang, Z. Duan, T. Xie, F. Cao, P. Shen, P. Song, C. Zhao, P. Jin, G. Sun, S. Xu, et al., FPC-VLA: a vision-language-action framework with a supervisor for failure prediction and correction, *Expert Syst. Appl.* 316 (2026) 131742.
- [208] S. Zuo, Z. Xie, W. Zheng, S. Xu, F. Li, S. Jiang, L. Chen, Z.-X. Yang, J. Lu, Dvgt: driving visual geometry transformer, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 14658–14668.
- [209] L. Russell, A. Hu, L. Bertoni, G. Fedoseev, J. Shotton, E. Arani, G. Corrado, Gaia-2: a controllable multi-view generative world model for autonomous driving, 2503.20523 (2025).
- [210] H. Cho, J. Choi, G. Baek, W. Hwang, Itkd: interchange transfer-based knowledge distillation for 3d object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13540–13549.
- [211] L. Zhang, R. Dong, H.-S. Tai, K. Ma, Pointdistiller: structured knowledge distillation towards efficient and compact 3d detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21791–21801.
- [212] L. Li, X. Yue, Z. Xu, S. Xie, Multi-dimensional pruned sparse convolution for efficient 3D object detection, in: *2023 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2023, pp. 3190–3194.
- [213] S. Wang, J. Zheng, Monoskd: general distillation framework for monocular 3d object detection via spearman correlation coefficient, 2310.11316 (2023).
- [214] T. Dai, R. Huang, H. Guo, J. Wang, B. Li, Z. Zhu, MFD-KD: multi-scale frequency-driven knowledge distillation, *Big Data Min. Anal.* 8 (3) (2025) 563–574.
- [215] F. Qian, J. Hong, H. Yan, H. Chen, C. Zhang, H. Su, S. Zhao, Empowering object detection: unleashing the potential of decoupled and interactive distillation, *IEEE Trans. Intell. Transp. Syst.* (2024).
- [216] S. Chen, L. Wei, L. Liang, C. Lang, Joint homophily and heterophily relational knowledge distillation for efficient and compact 3D object detection, in: *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 2127–2135.
- [217] Y. Zhang, Z. Dong, H. Yang, M. Lu, C.-C. Tseng, Y. Du, K. Keutzer, L. Du, S. Zhang, QD-BEV: quantization-aware view-guided distillation for multi-view 3D object detection, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3825–3835.
- [218] Z. Chong, X. Ma, H. Zhang, Y. Yue, H. Li, Z. Wang, W. Ouyang, Monodistill: learning spatial features for monocular 3d object detection, 2201.10830 (2022).
- [219] Y. Hong, H. Dai, Y. Ding, Cross-modality knowledge distillation network for monocular 3d object detection, in: *European Conference on Computer Vision*, Springer, 2022, pp. 87–104.

- [220] J. Zeng, L. Chen, H. Deng, L. Lu, J. Yan, Y. Qiao, H. Li, Distilling focal knowledge from imperfect expert for 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 992–1001.
- [221] S. Zhou, W. Liu, C. Hu, S. Zhou, C. Ma, Unidistill: a universal cross-modality knowledge distillation framework for 3D object detection in bird's-eye view, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5116–5125.
- [222] Z. Wang, D. Li, C. Luo, C. Xie, X. Yang, Distillbev: boosting multi-camera 3d object detection with cross-modal knowledge distillation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8637–8646.
- [223] L. Huang, H. Wang, J. Zeng, S. Zhang, L. Cao, J. Yan, H. Li, LiDAR-guided geometric pretraining for vision-centric 3D object detection, *Int. J. Comput. Vis.* 133 (7) (2025) 3877–3890.
- [224] Z. Wen, J. Liu, H. Zhao, Q. Wang, A triple semantic-aware knowledge distillation network for industrial defect detection, *Comput. Ind.* 166 (2025) 104252.
- [225] J. Zhang, J. Liu, Y. Pei, J. Zhang, X. Zhao, Learn from voxels: knowledge distillation for pillar-based 3D object detection with LiDAR point clouds in autonomous driving, *IEEE Trans. Intell. Veh.* (2024) 1–11. <https://doi.org/10.1109/TIV.2024.3397617>
- [226] M. Klingner, S. Borse, V.R. Kumar, B. Rezaei, V. Narayanan, S. Yogamani, F. Porikli, X3kd: knowledge distillation across modalities, tasks and stages for multi-camera 3D object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13343–13353.
- [227] Z. Jiang, J. Zhang, Y. Zhang, Q. Liu, Z. Hu, B. Wang, Y. Wang, FSD-BEV: foreground self-distillation for multi-view 3d object detection, in: European Conference on Computer Vision, Springer, 2024, pp. 110–126.
- [228] Y. Zhang, J. Chen, D. Huang, Cmae-3D: contrastive masked autoencoders for self-supervised 3d object detection, *Int. J. Comput. Vis.* 133 (5) (2025) 2783–2804.
- [229] L. Zhao, J. Song, K.A. Skinner, Crkd: enhanced camera-radar object detection with cross-modality knowledge distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 15470–15480.
- [230] W. Yan, L. Xu, H. Liu, C. Tang, W. Zhou, High-order structural relation distillation networks from lidar to monocular image 3D detectors, *IEEE Trans. Intell. Veh.* 9 (2) (2023) 3593–3604.
- [231] W. Cao, Y. Zhang, J. Gao, A. Cheng, K. Cheng, J. Cheng, Pkd: general distillation framework for object detectors via pearson correlation coefficient, *Adv. Neural Inf. Process. Syst.* 35 (2022) 15394–15406.
- [232] L. Zhang, Y. Shi, H.-S. Tai, Z. Zhang, Y. He, K. Wang, K. Ma, Structured knowledge distillation towards efficient and compact multi-view 3d detection, 2211.08398 (2022).
- [233] H. Zhang, L. Liu, Y. Huang, X. Lei, L. Tong, B. Wen, InstKD: towards lightweight 3D object detection with instance-aware knowledge distillation, *IEEE Trans. Intell. Veh.* (2024).
- [234] J. Li, Z. Liu, J. Hou, D. Liang, Dds3D: Dense pseudo-labels with dynamic threshold for semi-supervised 3d object detection, 2303.05079 (2023).
- [235] Y. Li, S. Xu, M. Lin, J. Yin, B. Zhang, X. Cao, Representation disparity-aware distillation for 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 6715–6724.
- [236] C. Bu, W. Yan, Z. Liu, Y. Song, C. Tang, Dynamic riemannian geometry relational distillation for lightweight 3D object detection, *IEEE Trans. Artif. Intell.* (2026).
- [237] F. Li, C. Min, L. Xiao, D. Zhao, S. Si, H. Xue, Y. Nie, B. Dai, Pre-ruined distillation for point cloud-based 3D object detection, in: 2024 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2024, pp. 3192–3198.
- [238] F. Li, C. Min, J. Wang, L. Xiao, D. Zhao, Y. Nie, B. Dai, Efficient distillation using channel pruning for point cloud-based 3D object detection, *IEEE Trans. Intell. Transp. Syst.* 26 (10) (2025) 16144–16158.
- [239] S. Zhou, L. Li, X. Zhang, B. Zhang, S. Bai, M. Sun, Z. Zhao, X. Lu, X. Chu, LIDAR-PTQ: Post-training quantization for point cloud 3d object detection, 2401.15865 (2024).
- [240] H.-a. Gao, B. Tian, P. Li, H. Zhao, G. Zhou, Dqs3D: densely-matched quantization-aware semi-supervised 3D detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21905–21915.
- [241] T. Cao, K. Hu, Q. Xu, S. Chen, Post-training quantization based on knowledge distillation for point cloud 3D object detection, in: 2025 IEEE 14th International Conference on Communications, Circuits and Systems (ICCCAS), IEEE, 2025, pp. 423–429.
- [242] Y. Huang, L. Liu, Y. Gao, H. Zhang, H. Li, D2S: towards efficient sparse 3D object detection via dense to sparse knowledge distillation, in: ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2025, pp. 1–5.
- [243] G. Guo, D. Zhang, L. Han, N. Liu, M.-M. Cheng, J. Han, Pixel distillation: cost-flexible distillation across image sizes and heterogeneous networks, *IEEE Trans. Pattern Anal. Mach. Intell.* 46 (12) (2024) 9536–9550.
- [244] M.S. Mohammad, U. Kamal, S. Mukhopadhyay, Adaptive-cloud: dynamic computation control for 3D object detection from LIDAR point clouds, *IEEE Robot. Autom. Lett.* 10 (6) (2025) 6512–6519.
- [245] H. Zhang, L. Liu, F. Hui, B. Zhang, H. Zhang, Z. Zha, CLEAN: category knowledge-driven compression framework for efficient 3D object detection, *IEEE Trans. Pattern Anal. Mach. Intell.* 47 (10) (2025) 8740–8755.
- [246] Y. Zhang, Y. Lin, W. Yan, L. Yao, X. Wan, G. Li, C. Zhang, G. Ke, J. Xu, Incomplete multi-view clustering via diffusion contrastive generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 39, 2025, pp. 22650–22658.
- [247] Y. Zheng, X. Li, P. Li, Y. Zheng, B. Jin, C. Zhong, X. Long, H. Zhao, Q. Zhang, MONOCC: digging into monocular semantic occupancy prediction, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 18398–18405.
- [248] J.I. Yuzhe, Y. Chen, L. Yang, D. Rui, M. Yang, X. Zheng, VeXKD: the versatile integration of cross-modal fusion and knowledge distillation for 3D perception, *Adv. Neural Inf. Process. Syst.* 37 (2024) 125608–125634.
- [249] R. Ding, M. Yang, N. Zheng, Selective transfer learning of cross-modality distillation for monocular 3d object detection, *IEEE Trans. Circuits Syst. Video Technol.* 34 (10) (2024) 9925–9938.
- [250] W. Zhang, D. Liu, C. Ma, W. Cai, Alleviating foreground sparsity for semi-supervised monocular 3d object detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 7542–7552.
- [251] H. Gao, X. Yu, Y. Xu, Q. Ran, W. Hussain, MonoFG: monocular 3D object detection with knowledge distillation for human-centric autonomous driving systems, *ACM Trans. Adapt. Syst.* (2024) <https://doi.org/10.1145/3703458>.
- [252] L. Peng, J. Xu, H. Cheng, Z. Yang, X. Wu, W. Qian, W. Wang, B. Wu, D. Cai, Learning occupancy for monocular 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 10281–10292.
- [253] D. Kwon, Y. Yoon, H. Son, S. Kwak, MemDistill: distilling LiDAR knowledge into memory for camera-only 3D object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 6828–6838.
- [254] H. Zhao, Q. Zhang, S. Zhao, Z. Chen, J. Zhang, D. Tao, Simdistill: simulated multi-modal distillation for bev 3d object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 38, 2024, pp. 7460–7468.
- [255] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, J. Dai, Bevformer: learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arxiv* 2022, 2203.17270 (2026).
- [256] Y. Lu, X. Ma, L. Yang, T. Zhang, Y. Liu, Q. Chu, J. Yan, W. Ouyang, Geometry uncertainty projection network for monocular 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3111–3121.
- [257] L. Peng, X. Wu, Z. Yang, H. Liu, D. Cai, Did-m3d: decoupling instance depth for monocular 3d object detection, in: European Conference on Computer Vision, Springer, 2022, pp. 71–88.
- [258] J. Xu, L. Peng, H. Cheng, H. Li, W. Qian, K. Li, W. Wang, D. Cai, Mononerf: nerf-like representations for monocular 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 6814–6824.
- [259] F. Yu, D. Wang, E. Shelhamer, T. Darrell, Deep layer aggregation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2403–2412.
- [260] J. Zhang, Y. Zhang, Q. Liu, Y. Wang, Sa-Bev: generating semantic-aware bird's-eye-view feature for multi-view 3D object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3348–3357.
- [261] K. Ning, Y. Liu, Y. Su, K. Jiang, Diversity knowledge distillation for lidar-based 3-D object detection, *IEEE Sens. J.* 23 (11) (2023) 11181–11193.
- [262] Y. Wei, Z. Wei, Y. Rao, J. Li, J. Zhou, J. Lu, Lidar distillation: bridging the beam-induced domain gap for 3d object detection, in: Proceedings of the European Conference on Computer Vision, Springer, 2022, pp. 179–195.
- [263] C.-J. Ho, C.-H. Tai, Y.-Y. Lin, M.-H. Yang, Y.-H. Tsai, Diffusion-SS3D: diffusion model for semi-supervised 3d object detection, *Adv. Neural Inf. Process. Syst.* 36 (2023) 49100–49112.
- [264] A. Malik, P. Mirdehghan, S. Niousias, K. Kutulakos, D. Lindell, Transient neural radiance fields for lidar view synthesis and 3d reconstruction, *Adv. Neural Inf. Process. Syst.* 36 (2023) 71569–71581.
- [265] S. Zhou, H. Chang, S. Jiang, Z. Fan, Z. Zhu, D. Xu, P. Chari, S. You, Z. Wang, A. Kadambi, Feature 3DGS: supercharging 3d gaussian splatting to enable distilled feature fields, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 21676–21685.
- [266] Z. Liu, H. Mu, X. Zhang, Z. Guo, X. Yang, K.-T. Cheng, J. Sun, MetaPruning: meta learning for automatic neural network channel pruning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3296–3305.
- [267] S. Xu, S. Jiang, F. Li, L. Liu, Z. Song, B. Yang, Z.-x. Yang, Sparseinteraction: sparse semantic guidance for radar and camera 3D object detection, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 9224–9233.
- [268] X. Zhang, L. Wang, J. Chen, C. Fang, G. Yang, Y. Wang, L. Yang, Z. Song, L. Liu, X. Zhang, et al., Dual radar: a multi-modal dataset with dual 4D radar for autonomous driving, *Sci. Data* 12 (1) (2025) 439.
- [269] L. Wang, G. Yang, L. Yang, X. Zhang, Z. Song, Y. Chen, L. Liu, J. Gao, Z. Li, Q. Yang, et al., S2R-bench: a sim-to-real evaluation benchmark for autonomous driving, *Sci. Data* 12 (1) (2025) 2006:1–15.
- [270] J. Chen, Y. Kang, J. Kang, M. Xu, Y. Tong, F. Wu, D. Niyato, Hierarchical control multi-agent DRL for vehicle twin migration with workload prediction in UAV-assisted vehicular metaverses, *IEEE Trans. Mob. Comput.* (2026) <https://doi.org/10.1109/TMC.2026.3674825>.
- [271] X. Huang, H. Wu, X. Li, X. Fan, C. Wen, C. Wang, Sunshine to rainstorm: cross-weather knowledge distillation for robust 3d object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 38, 2024, pp. 2409–2416.
- [272] J. Libberstedt, E.R. Guerrero, N. Uhlemann, M. Lienkamp, V3lma: Visual 3d-enhanced language model for autonomous driving, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 4769–4778.
- [273] R. Zhao, A. Du, M. Cai, Z. Zhu, B. He, Vision-language model-driven human-vehicle interaction for autonomous driving: status, challenge, and innovation, *Big Data Min. Anal.* 9 (2) (2026) 425–447.