This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3264658

Multi-modal 3D Object Detection in Autonomous Driving: A Survey and Taxonomy

Li Wang, Xinyu Zhang, Ziying Song, Jiangfeng Bi, Guoxin Zhang, Haiyue Wei, Liyao Tang, Lei Yang, Jun Li, Caiyan Jia, Lijun Zhao

***Abstract*—Autonomous vehicles require constant environmental perception to obtain the distribution of obstacles to achieve safe driving. Specifically, 3D object detection is a vital functional module as it can simultaneously predict surrounding objects' categories, locations, and sizes. Generally, autonomous vehicles are equipped with multiple sensors, including cameras and LiDARs. The fact that single-modal methods suffer from unsatisfactory detection performance motivates utilizing multiple modalities as inputs to compensate for single sensor faults. Although many multi-modal fusion detection algorithms exist, there is still a lack of comprehensive and in-depth analysis of these methods to clarify how to fuse multi-modal data effectively. Therefore, this paper surveys recent advancements in fusion detection methods. First, we present the broad background of multi-modal 3D object detection and identify the characteristics of widely used datasets along with their evaluation metrics. Second, instead of the traditional classification method of early, middle, and late fusion, we categorize and analyze all fusion methods from three aspects: feature representation, alignment, and fusion, which reveals how these fusion methods are implemented in an essential way. Third, we provide an in-depth comparison of their pros and cons and compare their performance in mainstream datasets. Finally, we further summarize current challenges and research trends for realizing the full potential of multi-modal 3D object detection.**

***Index Terms*—Autonomous driving, 3D object detection, multi-modal fusion.**

This work was supported by the National High Technology Research and Development Program of China under Grant No. 2018YFE0204300, the National Natural Science Foundation of China under Grant No. 62273198, U1964203, 62073101, the China Postdoctoral Science Foundation (No.2021M691780), and State Key Laboratory of Robotics and Systems (HIT) (SKLRS-2022-KF-12). *(Corresponding author: Xinyu Zhang.)*

Li Wang is with the State Key Laboratory of Automotive Safety and Energy, and the School of Vehicle and Mobility, Tsinghua University, Beijing 100084, and also with State Key Laboratory of Robotics and Systems (HIT), Harbin 150001, China (e-mail: wangli_thu@mail.tsinghua.edu.cn).

Xinyu Zhang is with the State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing, 100084, and the School of Transportation Science and Engineering, Beihang University, Beijing, 100191, China (e-mail: xyzhang@tsinghua.edu.cn).

Lei Yang and Jun Li are with the State Key Laboratory of Automotive Safety and Energy, and the School of Vehicle and Mobility, Tsinghua University, Beijing, 100084 China (e-mail: yanglei20@mails.tsinghua.edu.cn; lijun19580326@126.com).

Ziying Song, Caiyan Jia are with the School of Computer and Information Technology, Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China (e-mail: 22110110@bjtu.edu.cn; cyjia@bjtu.edu.cn).

Jiangfeng Bi, Guoxin Zhang, Haiyue Wei are with the School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China (e-mail: bjf19981227@163.com; zhangguoxincs@gmail.com; ezio59624@gmail.com).

Liyao Tang is with the school of computer science, the University of Sydney, Australia (e-mail: ltan9687@uni.sydney.edu.au).

Lijun Zhao is with the State Key Laboratory of Robotics and System at Harbin Institute of Technology, Harbin 150001, China (e-mail: zhaolj@hit.edu.cn).

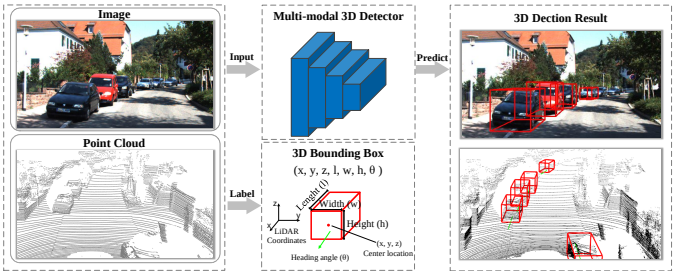


Fig. 1: An illustration of multi-modal 3D object detection in autonomous driving scenarios. Images have rich semantic information, and point clouds contain depth information. They have complementary characteristics, which can improve 3D object detection accuracy and help vehicles better perceive their surroundings.

I. Introduction

AUTONOMOUS driving, which is gradually attracting attention, is seen as a sanctuary for humans. According to the World Health Organisation (WHO) [1], traffic accidents kill approximately 1.35 million people each year, injure approximately 54 million and kill an average of 145 people every hour of every day, with the majority of people aged 5 to 29 years old. Therefore, autonomous driving is aim to improve safe driving, increase traffic efficiency and relieve drivers' burden [2]–[4]. In driving scenarios, self-driving vehicles require accurate and efficient perception operations to predict their driving environment at all times. In particular, the perception system, which transforms the various sensor data (**Fig.** 1) to semantic information, is a core and indispensable component for autonomous driving [5].

In pursuit of a comprehensive understanding of the driving environment, numerous computer vision tasks, *e.g.*, object detection, object tracking [6]–[10] *etc.*, are applied to the perception system. In particular, 3D object detection, which can intelligently predict locations, sizes, and categories of the important objects nearby vehicles, plays a fundamental role in the perception system, which provides object-level information for downstream perception tasks [11]–[13]. Moreover, 3D object detection has recently developed rapidly, so the perception accuracy obtained greatly improves. Specifically, numerous state-of-the-art methods have been proposed with the development of deep learning in computer vision [14]–[17]. Initially, the single-modal methods using only LiDAR point clouds or

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3264658

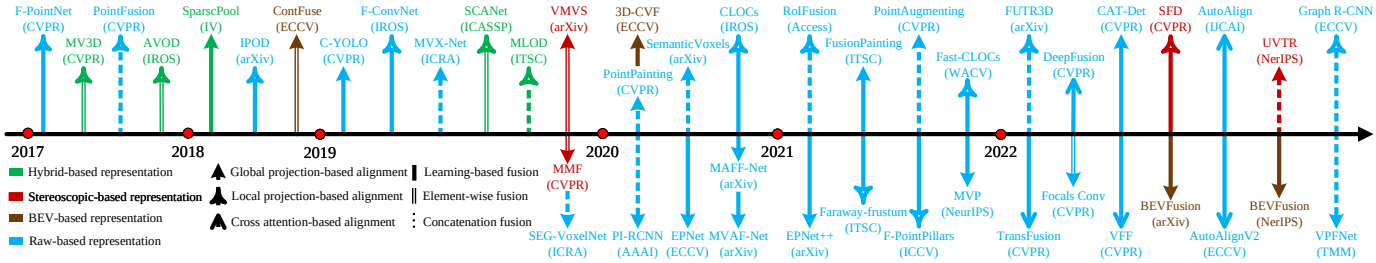


Fig. 2: Chronological overview of prestigious multi-modal 3D object detection works. The color, arrow, and line type represent the categories of representation, projection, and fusion, respectively.

images are developed rapidly. However, only one modality has its own defects [18]. Specifically, point-cloud-based methods [19]–[35], derived by LiDAR or RADAR, considerably limits their performances by sparsity and disorderliness of points, since the point cloud provides poor information in texture and occluded regions which is more serious at long distance. On the contrary, image-based methods [36]–[51] provides sufficient texture and context information but insufficiently geometrical information. Therefore, they both have certain drawbacks which degenerate detection performance.

To address the intrinsic limitation of single modality, multi-modal methods have been proposed, where the fusion from multiple sources is expected to provide a better and more complete perception of the 3D environment.

Multi-view methods [52], [53] propose to fuse inputs from different modalities into the same dimension. Furthermore, frustum-based models [54]–[57] provide a novel approach to combining heterogeneous features. Further, feature-wise fusion has received attention in multi-modal tasks, which has started a trend of feature-wise methods in multi-modal 3D object detection. Several methods [58]–[60] propose to transform heterogeneous modality to a unified representation, which can narrow the heterogeneity gap in a joint semantic subspace. Since different dimensions of features generate a lot of additional noise, more time consumption *etc.*, it isn't easy to leverage heterogeneous information with only a single model. However, numerous multi-modal methods are sophisticated for sundry variants. Therefore, we conduct a comprehensive survey of multi-modal 3D object detection. We hope such a systematic discussion on these recent advances could inspire fascinating future research [61]–[66]. In addition, recent research on collaborative control [67], [68] and multi-agent environment [69]–[73] perception are revolutionizing future transportation systems. Similarly, they require multi-modal perception as a foundation.

To this end, we propose a comprehensive review and provide an in-depth analysis of multi-modal 3D object detection. Compared with previous reviews [5], [12]–[16], [61]–[66], [74], [75] that focus on the task of 3D object detection, we specifically study the fusion of multi-model data for 3D object detection in general.

Furthermore, we introduce multi-modal methods from three insightful perspectives: representation, alignment, and fusion [**?**], [76], [77]. *First*, in multi-modal, data **representation** is a fundamental component since a key challenge is how to exploit and summarize the complementary of heterogeneous modality. *Second*, the multi-modal methods require ascertaining the relationship between two or multiple modalities, due to the heterogeneous coordinate system. Therefore, data **alignment** plays a crucial role in multi-modal. *Third*, another key challenge in multi-modal learning is to combine heterogeneous data into joint information. For that, **fusion** is vital for multi-modal 3D object detection. **Fig.** 2 is a chronological overview of prestigious multi-modal works. A comprehensive introduction of our taxonomy is shown in **Fig.** 3. Meanwhile, our paper presents an overview of advancements in this field and compares state-of-the-art methods comprehensively. The major contribution of this work can be summarized as follows:

- To the best of our knowledge, this is the *first* survey paper to comprehensively review multi-modal 3D object detection for autonomous driving rather than view it as a trivial subset of 3D object detection.
- We propose a *taxonomy* for multi-modal 3D object detection that exceeds the traditional early, middle, and late fusion split and consists of three aspects: *representation*, *alignment*, and *fusion*.
- This paper covers *the most recent and advanced progress* of multi-modal 3d object detection. We provide state-of-the-art methods for readers.
- We comprehensively compare existing methods on several publicly available datasets and provide insightful analysis.

II. Background

In this section, we introduce the general background of 3D object detection and the relationship between single-modal and multi-modal 3D object detection. We also introduce the common datasets and the evaluation metrics.

A. 3D Object Detection

Problem Definition. 3D object detection dedicates to predict the properties of object in 3D scenario, which includes locations, sizes, categories, *etc.* Generally, it can be represented as follows:

$$\beta(b_1, b_2, ..., b_N) = F_{detect}(\alpha), \tag{1}$$

where $\beta(b_1, b_2, ..., b_N)$ represents a set containing N object states in a frame scenario, and $(b_1, b_2, ..., b_N)$ is N 3D objects correspondingly. F_{detect} is the 3D detection function and α is

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3264658

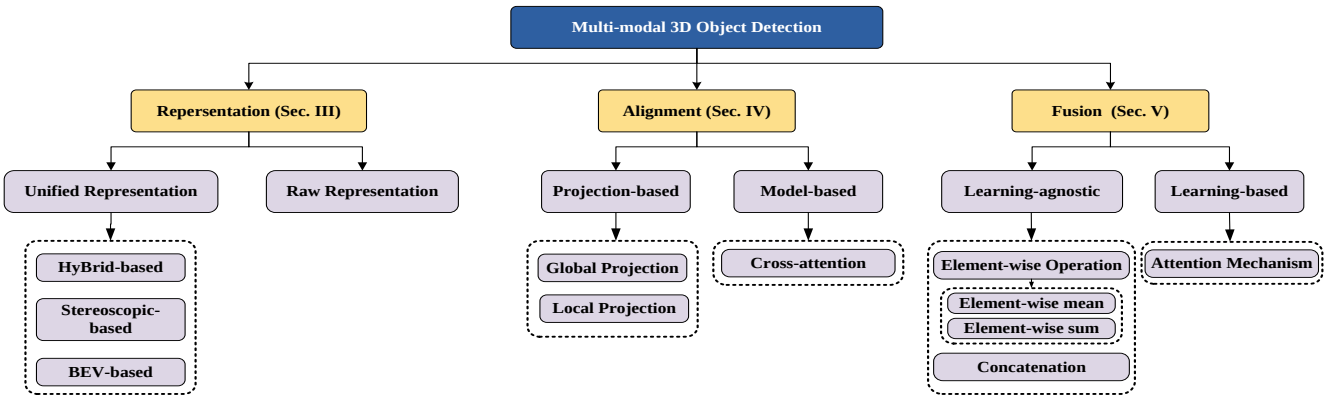


Fig. 3: Our proposed taxonomy with three components of multi-modal fusion for 3D object detection.

TABLE I: An analysis of common sensors in autonomous driving.

Sensor	Advantages	Disadvantages
Monocular Camera	Affordable price, simple construction, suitable for close-range measurement and robot vision	Limited depth perception, sensitive to lighting conditions, susceptible to distortion
Stereo Camera	Strong depth perception, insensitive to lighting conditions, suitable for long-range measurement and robot vision	Higher cost, requires calibration, susceptible to distortion, limited field of view, not suitable for transparent or reflective surfaces
LiDAR	High accuracy, suitable for outdoor environments, capable of working in low or no light conditions	Higher cost, larger size, limited resolution, may be affected by environmental factors such as rain, fog or snow
RADAR	Suitable for outdoor environments, capable of working in strong light or harsh weather, with long-range detection capability	Lower accuracy, sensitive to object shape and orientation, limited resolution

the input data, usually from the sensor. The b_i(*bounding box*) usually contains information such as location, size, category, *etc.*, and may contain more information depending on the model set.

Sensors. In 3D object detection, several popular sensors are shown in **Tab.** I including monocular camera, stero camera, LiDAR, and RADAR. Meanwhile, we compare the advantages and disadvantages.

Cameras capture images with rich color and texture properties and have the advantage of high frame rates and negligible cost. However, it lacks depth information and is vulnerable to illumination. On the other hand, point clouds are the data acquired by LiDAR or RADAR, which is a massive collection of points expressing the spatial distribution of the object and the spectral properties of the object surface in the same spatial reference system. LiDAR provides high-precision and high-density point cloud data with high resolution for object detection. However, its acquisition requires significant computational resources and is sensitive to adverse weather conditions. RADAR can measure point cloud data over a large range, unaffected by environmental conditions, and detect moving objects. However, its measurement precision and object resolution are relatively low and can be affected by reflection interference.

Single-Modal. In autonomous driving, utilizing single sensors performing object detection is unsatisfactory. Specifically, single-modal has inherent defects, which reasons insufficient environment perception in 3D scenarios. For example, camera-based 3D detectors achieve low-accuracy performance, as the image does not provide sufficient depth information. Although LiDAR-Based methods, such as [20], [29], [30], overcome the issue of poor depth information, they also suffer from the defects of LiDAR, *e.g.* low resolution, sparsity, poor texture, *etc.* Is there any resolution?

Multi-Modal. Multi-modal 3D object detection proposes integrating multiple sensors, combining the advantages of multiple modalities to achieve better performance. In contrast to single-modal, it can fully exploit the advantages of multiple modalities (e.g. depth information from point clouds, texture information from images), which brings great potential and enhancement to autonomous driving perception. However, this also brings numerous problems and challenges. For example, MV3D [52], a pioneer in multi-modal 3D object detection, strives to combine the data of two modalities applied together but ignores the gap of heterogeneous modalities. Meanwhile, the heterogeneous gap is a key challenge in multi-modal learning.

B. Datasets and Evaluation Metrics

In the past few years, numerous datasets for autonomous driving have been proposed, which can greatly facilitate research on 3D object detection. To further relieve the burden of investigation for readers, we have surveyed the available datasets (**Tab.** II) and reviewed the most popular ones, including KITTI [78], nuScenes [79], and Waymo [80].

KITTI Dataset. The KITTI [78] dataset, jointly constructed by Karlsruhe Institute of Technology and Toyota Institute of

TABLE II: Datasets for 3D object detection in autonomous driving. "-" means not mentioned.

Dataset	Year	Sensors	Frames	Annotated Frames	3D Boxes	LiDAR Coverage	Camera Coverage	Night/Rain
KITTI [78]	2012	1 LiDAR, 2 Cameras	15K	15K	80K	360°	90°	No/No
KAIST [81]	2018	1 LiDAR, 2 Cameras	8.9K	8.9K	-	360°	90°	Yes/No
ApolloScape [82]	2018	2 LiDARs, 2 Cameras	20K	20K	475K	360°	-	Yes/No
H3D [83]	2019	1 LiDAR, 2 Cameras	27K	27K	1M	360°	180°	No/No
Lyft L5 [84]	2019	3 LiDARs, 7 Cameras	46K	46K	1.3M	360°	360°	No/No
Argoverse [85]	2019	1 LiDAR, 7 Cameras	44K	-	993K	360°	360°	Yes/Yes
nuScencs [79]	2019	1 LiDAR, 6 Cameras	400K	40K	1.4M	360°	360°	Yes/Yes
Waymo [80]	2019	5 LiDARs, 5 Cameras	230K	230K	12M	360°	270°	Yes/Yes
A*3D [86]	2019	1 LiDAR, 2 Cameras	39K	39K	230K	360°	57.3°	Yes/Yes
PandaSet [87]	2020	2 LiDARs, 6 Cameras	8.2K	8.2K	-	360°	360°	Yes/No
Cirrus [88]	2021	2 LiDARs, 1 Camera	6.2K	6.2K	-	240°	90°	-/-
ONCE [89]	2021	1 LiDAR, 7 Cameras	1M	1M	417K	360°	360°	Yes/Yes

Technology in 2012, is the most commonly used dataset for 3D perception. The KITTI dataset acquires data by driving an acquisition vehicle, including data from urban, highway, and rural scenes. Each frame can contain up to 15 cars and 30 pedestrians, with varying degrees of occlusion and truncation. The dataset comprises 389 stereo image pairs, optical flow maps, 39.2 km visual range sequences, 15,000 point cloud frames, and 200,000 manually labeled 3D object frames. The data acquisition vehicle has two grayscale, two color cameras, a Velodyne 64-line LiDAR, four optical lenses, and a GPS navigation system. KITTI provides the raw data in the 3D perception benchmark, different evaluation metrics for each corresponding benchmark, and an online test platform to test and compare the performance of different methods.

KITTI uses the APR_{40} interpolation method as their official evaluation method. Two types are ranked in the official KITTI evaluation, including 3D object detection and bird's eye view (BEV) detection. In the evaluation of 3D object detection, the 3D intersection ratio (3D Intersection over Union, 3D IoU) is used as the detection threshold, and the intersection ratio between the predicted bounding box and the real bounding box is greater than the threshold value, which is considered as correct detection. It is calculated by projecting the 3D bounding box onto the ground for evaluation, etc. The KITTI dataset used APR_{11} until August 10, 2019, and the evaluation criteria for 3D object detection was adjusted from APR_{11} to APR_{40}. The KITTI dataset also classifies objects into three difficulties according to their degree of recognition: Easy, Moderate, and Hard.

For object orientation prediction, KITTI uses a new evaluation metric, AOS (Average Orientation Similarity):

$$AOS = \frac{1}{11} \sum_{r \in 0, 0.1, \ldots, 1} \max_{\tilde{r}: \tilde{r} \geq r} s(\tilde{r}), \tag{2}$$

where r denotes the recall in PASCAL object detection; s denotes the directional similarity, which takes values in the range [0,1]; $s(r)$ is a variant of cosine similarity, defined as:

$$s(r) = \frac{1}{\mid D(r) \mid} \sum_{i \in D(r)} \frac{1 + \cos \Delta_{\theta}^{(i)}}{2} \delta_i, \tag{3}$$

where $D(r)$ denotes the set of all object detection results under recall r; $\Delta_{\theta}^{(i)}$ is the angular difference between the predicted and true directions of detection object i.

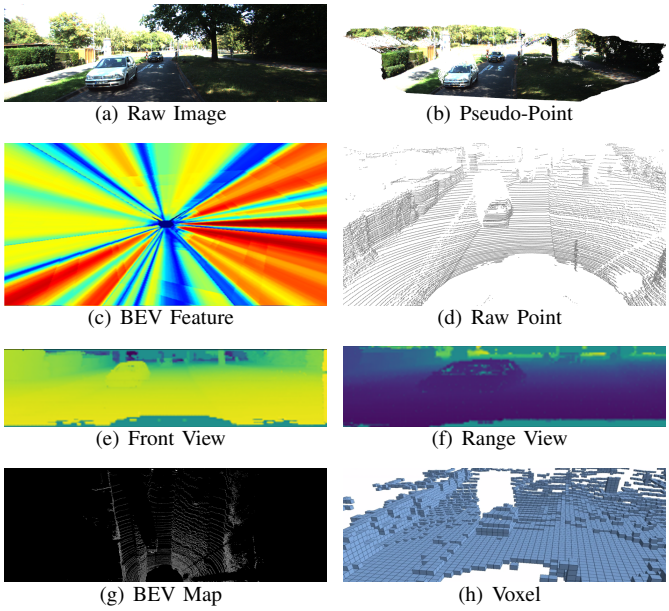
nuScenes Dataset. The nuScenes [79] dataset is a large-scale dataset for autonomous driving built by nuTonomy in 2019. The nuScenes provides several excellent benchmarks, e.g., 3D object detection and 3D tracking, and proposes a corresponding online benchmark for testing and comparing the performance of numerous works. To diversify the data, nuScenes conducted collection in Boston and Singapore, known for their complex traffic environments and challenging driving conditions. In both cities, 1,000 complete scenes are collected, each taking about 20 seconds and containing complex scenes such as sunny days, rainy days, and dark nights. The nuScenes includes a total of approximately 1.4 million camera images, 390,000 LiDAR scans, 1.3 million millimeter wave radar scans, and 1.4 million 3D object annotations. Compared to other datasets, all of its sensors provide a 360-degree view. Specifically, the nuScenes have six cameras, five RADARs, and one LiDAR. These multi-angle cameras and radars allow the nuScenes to obtain views from different angles, resulting in a decent 360-degree scene.

The nuScenes dataset utilizes the nuScenes Detection Score (NDS) as a metric to evaluate an assay:

$$NDS = \frac{1}{10}[5mAP + \sum_{mTP \in TP} (1 - min(1, mTP))]. \tag{4}$$

Among them, mAP represents the mean Average Precision; mTP (True Positive metrics) is composed of 5 metrics: (1) Average Translation Error (ATE); (2) Average Scale Error (ASE); (3) Average Orientation Error (AOE); (4) Average Velocity Error (AVE) ; (5) Average Attribute Error, (AAE);

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3264658



(a) Raw Image (b) Pseudo-Point (c) BEV Feature (d) Raw Point (e) Front View (f) Range View (g) BEV Map (h) Voxel

Fig. 4: An illustration of general representation of point and image. (b) and (c) come from (a); (e), (f), (g), and (h) come from (d). Specifically, (b) is a 3D representation similar to (d) but dense. (e), (f), and (g) are compact 2D representations encoded by the different views, and (h) is a dense 3D representation generated by (d).

mTP is averaged for each the mTP is averaged for each class (C represents the set of object classes):

$$mTP = \frac{1}{\mid C \mid} \sum_{c \in C} TP_c. \tag{5}$$

In summary, NDS is a composite metric that combines the attributes of predicted object location, size, direction, and velocity. As shown in **Eq.** 4, half of the weight in NDS comes from the detection performance, and the other half is the quality assessment of the prediction frame position, size, direction, velocity, and attributes. Since the values of $mAVE$, $mAOE$, and $mATE$ will be greater than 1, each value is constrained in the interval [0,1] in **Eq.** 4.

Waymo Dataset. The Waymo [80] Open Dataset, a large-scale autonomous driving dataset released in 2020, is now the largest and most diverse multi-modal dataset. The Waymo provides several 3D perception benchmarks, such as 3D object detection, and 3D semantic segmentation. It collects data from multiple cities, such as San Francisco, Phoenix, etc. These cities have a large geographic coverage, including different scenarios under various driving conditions such as day, night, dawn, dusk, and rain. Specifically, the entire dataset consists of 1,150 scenes, each taking approximately 20 seconds, with a total of 230,000 frames of data, of which approximately 12 million objects were manually labeled. The Waymo used a 10 Hz frequency to acquire multiple sensors simultaneously. Specifically, the acquisition equipment consisted of five high-resolution cameras and five high-quality LiDAR sensors.

The Waymo dataset uses Average Precision weighted by Heading (APH) as the main evaluation metric for the model. This metric is improved based on AP, and AP and APH are calculated as shown in **Eq.** 6 and **Eq.** 7.

$$AP = 100 \int_0^1 maxp(r') \mid r' >= r dr, \tag{6}$$

$$APH = 100 \int_0^1 maxh(r') \mid r' >= r dr, \tag{7}$$

where $h(r')$ is calculated similarly to the P/R curve, where $p(r)$ is the P/R curve; $h(r)$ is calculated similarly to $p(r)$, where the TP (True Positives) used is the value-weighted by the direction (Heading); r is denoted as the recall function, which uses 21 equally spaced recall intervals $r \in [0, 1/20, 2/20, ..., 1]$.

III. Representation

As a 3D environment is perceived by various sensors with different characteristics in multi-modal settings, the data representation becomes a critical design choice for fusing the information from different sensors. In the scenario of autonomous driving, the input data mainly consists of images and LiDAR point clouds. However, more representations are proposed for data fusion to exploit the different data modalities better.

Data representation, in multi-modal learning, is a vital part that determines the most critical stage of modeling task input.

To this end, we review popular representation in multi-modal 3D object detection, shown in **Tab.** III and an illustrate is in **Fig.** 4. To help understand the breadth of existing methods, we categorize them into two types: *Unified Representation* and *Raw Representation*. The *unified Representation* methods aim to transform heterogeneous data into the homology format and are arguably more challenging to build as they require the ability to construct a specific space for heterogeneity. *Raw Representation*, without any pre-processing, refers to the direct heterogeneous data used for prediction.

A. Unified Representation

Unified representation aims to process heterogeneous data (or features) in a coincident format, which can narrow the heterogeneous gap. According to the representation types, the methods can be divided into three categories: hybrid-based, 3D-based, and BEV-based.

Hybrid-based. Hybrid-based methods aim to combine heterogeneous information in a homogeneous format, for example, by converting a 3D point cloud into a 2D representation (the same as an image). Hybrid-based methods address the multi-modal detection problem from two aspects: designing new representations that can cope with heterogeneity and selecting an appropriate viewpoint for learning. As a pioneering work, MV3D [52] represents raw points in two different viewpoints, the range view, and brid's-eye view. Specifically, MV3D proposes an encoding method for the front view(similar to the range view) and brid's-eye view, which contains height, density, and intensity. In this way, a 3D representation can be converted to a 2D pseudo-image, allowing the network to extract geometric details using 2D convolutions. This design

TABLE III: A summary of various representations from image and point cloud. "Dim." represents a dimension.

Type	Dim.	References
Image		
Raw Image	2D	[25], [52]–[57], [90], [92]–[95], [99]–[125]
Pseudo-Point	3D	[58], [103], [126]
BEV Feature	2D	[23], [59], [60], [91]
Point cloud		
Raw Point	3D	[54], [56], [57], [99], [100], [105], [108]–[110], [112]–[115], [120], [122]
Front View	2D	[52]
Range View	2D	[94], [95]
BEV Map	2D	[52], [53], [90]–[93], [103]
Voxel	3D	[23], [25], [55], [58]–[60], [101], [102], [104], [106], [107], [111], [116]–[119], [121]–[125], [127]

philosophy has been followed by many works, [53], [90]–[93]. [94], [95] propose to address viewpoint misalignment by translating heterogeneous data into a dense 2D representation.

Stereoscopic-based. In contrast to hybrid-based methods, where the representation is densely distributed in a 2D space, stereoscopic-based methods aim to fuse heterogeneous representation in 3D space by converting 2D representation into 3D. There are several works [58] propose to translate an image from 2D space into a pseudo point, which contains geometry and texture information simultaneously. Since this way of generating a pseudo point requires each pixel's depth information, stereoscopic-based methods always exploit a depth estimate model, *e.g.* depth completion. SFD [58] proposes a basic pipeline for combining raw voxel and pseudo-point features, which relieves the original heterogeneous gap between data representations.

BEV-based. BEV representation, in 3D perception, is widely used because it is strongly interpretative and benefits expanding sensor modality and exploiting downstream tasks. BEV representation can address challenging problems that exist in autonomous driving scenarios, *e.g.,* vehicle occlusion, sparse representation. As for point clouds, changing the viewpoint is easy. In contrast, changing the camera's viewpoint requires laborious parameters and transformation decisions. Benefiting from advancements in BEV camera-only works [96]–[98], the development of *BEV-based* methods has been facilitated. The model proposed by [60] implements high-efficiency camera-to-BEV translation and effective semantic incorporation for BEV representation.

B. Raw Representation

An alternative to a unified multi-modal representation is the *Raw Representation* that aims to take no supererogatory representation translation or encoding for preserving maximum available information.

As most high-performance single-modal detectors are dominated by point clouds, in order to extend this superiority, several multi-modal methods propose to incorporate raw representations of other modalities, *e.g.* cameras, to decorate point clouds. For example, the model proposed by [104] presents a novel paradigm that decorates raw point clouds with semantic scores from semantic segmentation tasks. This extra multi-modal advantage benefits from strong 2D vision tasks for raw representation, such as 2D object detection or 2D semantic segmentation. To take full advantage of the raw representations, F-PointNet [54] uses 2D raw representation, with 2D object detection, to narrow the scope of 3D representations, resulting in accurate foreground information for prediction. This design paradigm has been followed in a lot of works [55]–[57], [100]. Although this way can alleviate gaps between features, they do not fully exploit the original information from heterogeneous data at the feature level.

Several methods have been proposed to exploit the complete original representation by the vanilla feature extractor. PointFusion [99] utilizes vanilla backbone, PointNet [27] for 3D, and ResNet [128] for 2D, to directly extract features from the raw representation. [101] followed it. [108] proposes a pillar-based encoding method to transform raw representation into pillar representation and handle pillar features with [30]. Unlike previous primitive feature extractors, [109], [129] proposes using encoder-decoder structures to enhance the interaction and fusion of heterogeneous representations. Due to the superiority of primitive 2D representation, more variants of 2D auxiliary tasks are allowed. [112] utilizes 2D detection for the raw image in the image branch, which achieves both 2D and 3D detection for ROI pooling (Region of interest pooling). In multi-modal methods, feature-wise fusion, which combines different features from the vanilla backbone and its variations, is becoming more common. Such as [117]–[120], [122]–[125]. This is mainly because raw representation can preserve more information from the original sensor, and their representation is more suitable for multi-modal inference.

C. Conclusion and Discussion

In this section, we identify two major categories of multi-modal representation in 3D object detection, unified representation, and raw representation. Unified representation projects multi-modal data (or feature) into a unified format (or space) and addresses the misalignment of representation or format. It has been extensively used in popular 3D detectors and has greatly improved performance and effectiveness, especially in the BEV-based paradigm. Raw representation, on the other hand, takes no transformation on original representation to preserve maximal raw information. Generally, it introduces auxiliary task, *e.g.*, semantic segmentation, and auxiliary object detection, for the original feature. A summary of their advantages and disadvantages is in **Tab.** IV. Finally, the multi-modal representation in 3D object detection is being developed, and we may see more high-efficiency representation in the future.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3264658

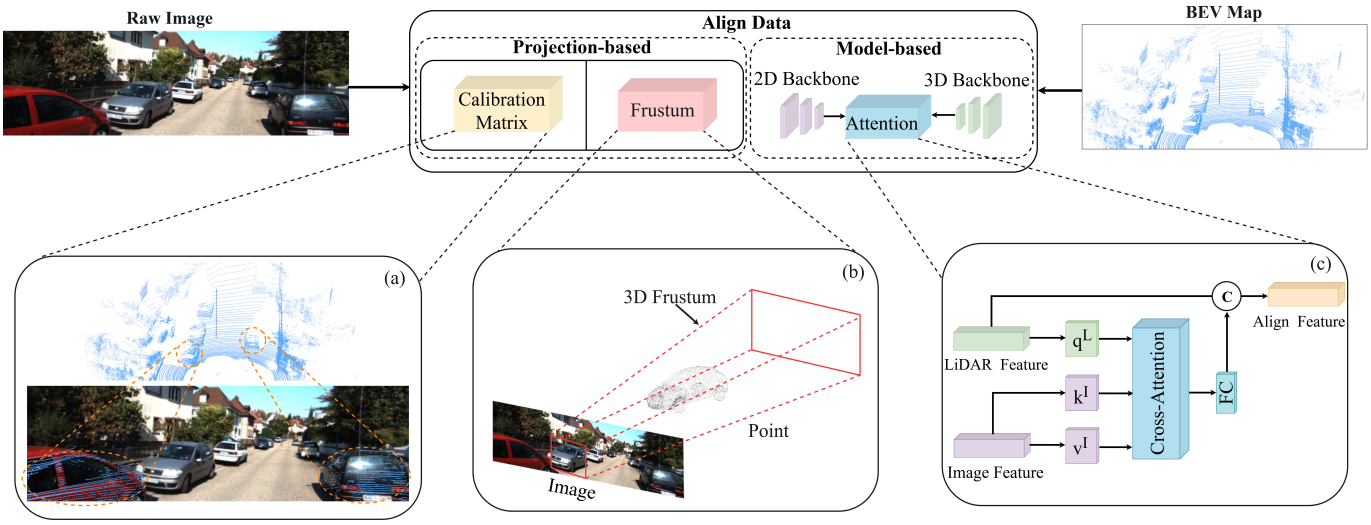


Fig. 5: An overview of alignment methods for multi-modal 3D object detection. We classify fusion methods into two broad categories: projection-based and model-based. Figures (a) and (b) are both projection-based methods. Figure (a) indicates that the point cloud is projected onto the image. Figure (b) indicates that the image is projected into the 3D space through the frustum of view. Figure (c) is a learnable way to achieve feature alignment through cross attention.

TABLE IV: An analysis of representation.

	Advantages	Disadvantages
Unified	+Small calculation volume +Fast calculation speed	-Loss of depth information -Loss of geometric relationships
Raw	+Retain depth information +Improve detection accuracy	-Large calculation volume -Large inference time

IV. Alignment

The input data of multi-modal fusion has different forms of feature representation, which are usually heterogeneous. Thus, it becomes an important step to construct the corresponding relationship between the data and different modalities. We propose to summarize this step as alignment because if directly using unaligned features from different modalities, it is probable that the gain of multi-modal data will be reduced or even counterproductive. Therefore, it is essential to consider feature alignment to construct the correspondence between different modal data.

Multi-modal feature alignment refers to constructing correspondences between different modal data features. In multi-modal 3D object detection, point cloud (as shown in **Fig.** 4) data provides accurate geometric information and depth information, but due to its inherent sparse and irregular distribution characteristics, point cloud lacks resolution and texture information. In contrast, images (as shown in **Fig.** 4) contain fine-grained texture and color information but lack depth information. The features extracted from the two heterogeneous data through the neural network are heterogeneous, and aligning the features of the two heterogeneous modalities is quite challenging.

The correspondence between LiDAR and camera is implemented by a projection matrix [130], [131], which consists of the intrinsic and extrinsic parameters to transform 3D world coordinate space into the 2D image coordinates. Multiple works utilize the calibration matrix to find the correspondence between 3D and 2D to achieve feature alignment. This method is effective, but it destroys the semantic information of the image. To better solve this problem, many researchers adopt deep learning techniques to achieve feature alignment. Based on these considerations, we divide feature alignment methods into two categories: 1) Projection-based, and 2) Model-based, as shown in **Fig.** 5 and **Tab.** V.

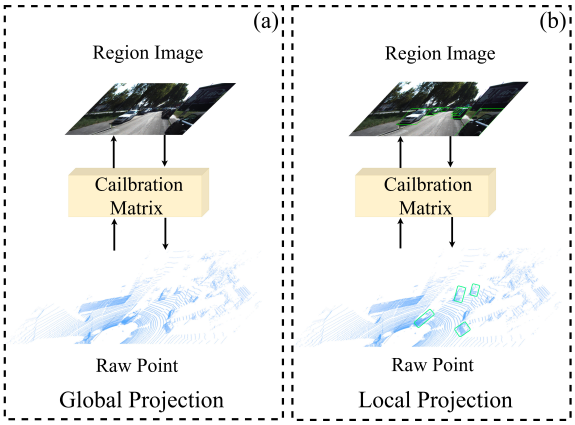


Fig. 6: Projection-based alignment in multi-modal 3D object detection. Figure (a) shows the global projection method, which uses the entire image as the search area of the point cloud. Figure (b) shows the method of local projection, and the search area of the point cloud is only the range marked by bounding boxes in the figure. The essential difference between the two methods is whether or not to reduce the scope of the point cloud through the 2D bounding box.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3264658

TABLE V: An overview of our taxonomy of multi-modal alignment.

Type	Advantages	Disadvantages	Reference
Projection-based			
Global projection	+Less calculation +Effective alignment	-Wide search space for point clouds	[23], [25], [59], [60], [91], [101]–[105], [107], [108], [108]–[111], [115], [120], [124]–[127]
Local projection	+Reduce point cloud search space	-Corrupts semantic information -Time consuming	[23], [25], [59], [60], [91], [101]–[105], [107], [108], [108]–[111], [115], [120], [124]–[127]
Model-based			
Attention	+Complete image semantic information	-High computational cost	[117]–[119], [121]

A. Projection-Based Feature Alignment

Previous work mainly utilizes camera projection matrices to align image and point features in a deterministic manner, which is efficient and fast and can maintain location consistency through projection matrices. Projection-based methods can be roughly divided into global projection and local projection, as shown in **Fig.** 6.

Global Projection. Global projection refers to taking the image features processed by the instance segmentation network or converting the image into a BEV as input, projecting the point cloud to the processed image, and inputting it into the 3D backbone for further processing.

For example, popular detection methods such as PointPainting [104] and PI-RCNN [105] fuse the image features in the image branch and semantic features in the raw LiDAR point cloud to enhance the point cloud with image-based semantic segmentation. Specifically, images are passed through a segmentation network to obtain pixel-level semantic labels, and then semantic labels are attached to 3D points by point-to-pixel projection. Complexer-yolo [102], [132], Seg-voxelnet [25], and FusionPainting [111] also follow this paradigm. MVP [110] draws on the idea of PointPainting, first uses image instance segmentation, and establishes the alignment relationship between the instance segmentation mask and the point cloud through the projection matrix, but the difference is that MVP randomly samples the pixels in each range, which is consistent with the point cloud The pixels on the point projection are connected by the nearest neighbor, and the depth of the laser point on the connection is taken as the depth of the current pixel. These points are then projected back to the laser coordinate system to obtain virtual LiDAR points. Mvx-Net [101] does not use the PointNet [27] network to extract point cloud features but preprocesses the original LiDAR point cloud into voxels to further use the backbone of more advanced single-mode 3D object detection, and pass the image feature vector of the corresponding pixel through the projected method attached to the voxel. This method appends the ROI image feature vector to the dense feature vector for each voxel in the LiDAR point cloud.

The three methods of Contfuse [91], BEVFusion [60], and 3D-CVF [23] express the data of the two modalities uniformly. The image features are transformed into a BEV representation by projection and aligned with the point cloud BEV representation. In Contfuse, the image features are projected into the BEV space through MLP learning. First, find each pixel's K neighborhood points in the image, then pass the projection matrix to the 3D space, and then project it into the image [133], [134]. The coordinate offsets of the feature and object pixels are input into the MLP. The image features of the object point are obtained. Then it is fused with the BEV feature map to form a dense feature map. Inspired by the LSS [134] algorithm, BEVFusion converts 3D ego-car coordinates to BEV representation by converting camera images to 3D ego-car coordinates and employing the BEV Encoder module. The 3D-CVF [23] transforms the 2D camera features into a smooth spatial feature map with maximum correspondence to the radar features in the BEV via self-calibration projection. This feature map also belongs to the BEV.

Local Projection. Local projection uses a 2D detection to extract knowledge from images to narrow down object candidate regions in 3D point clouds, transfers image knowledge to point clouds, and finally inputs the enhanced point clouds to LiDAR-based 3D object detectors.

Frustum-PointNet [54] proposes a frustum with predicted forward and backward truncated radial distances, extending the 2D box to 3D. First, an image is passed through a 2D object detector to generate a 2D bounding box around the object of interest. Then, the object within the 2D box is projected into a 3D frustum using a calibration matrix. The information in the 3D frustum is applied to the LiDAR point cloud to align the image and the point cloud. Some Works such as Frustum-ConvNet [56], Faraway-Frustum [57], Frustum-PointPillars [55], and Roarnet [100] have followed this setting. On this basis, corresponding innovations have been made. Specifically, Frustum-ConvNet aggregates point-wise into frustum-wise feature vectors. These feature vectors are combined into a feature map to use their fully convolutional network (FCN), which spatially fuses frustum feature

vectors and supports end-to-end and continuous estimation of oriented boxes in 3D space. Frustum-PointPillars adopts pillars to speed up calculations.

MV3D [52] transforms the LiDAR point cloud to BEV and front view (FV) through projection to generate proposals and then fuses the BEV, FV, and image features to predict the final 3D bounding box. In this process, a 3D proposal network is utilized to generate high-precision 3D candidate boxes, and 3D proposals are projected to feature maps in multiple views to achieve feature alignment between the two modalities. AVOD [53] also adopts the same idea, but unlike MV3D, AVOD removes the FV and proposes a more fine-grained region proposal.

PointAugmenting [115] does not use the features obtained from the image instance segmentation network but uses the feature map of the object detection network. This is mainly due to the fact that segmentation annotation is too expensive, while a 2D annotation is easy to implement.

SFD [58] proposes a method to employ pseudo point clouds, and the point cloud branch processes the raw point cloud to generate ROI regions of interest. The projection matrix is used to project the point cloud onto the image to generate a pseudo point cloud with color to achieve feature alignment of the two data. Finally, the search range of the point cloud is reduced by the generated ROI.

B. Model-Based Feature Alignment

Different from the previous method of aligning the two kinds of data using the camera projection matrix, some recent multi-modal 3D object detection methods propose aligning camera images and point clouds through a learning method mainly using attention. For example, both AutoAlign [117] and Deepfusion [119] employ a cross-attention mechanism to achieve feature alignment of two modalities. They convert voxels into query q and camera features, key k, and value v, respectively. For each query (i.e., voxel unit), we perform an inner product between the query and the key to obtain a matrix containing the correlations between the voxel and all its corresponding camera features. A softmax operator is employed to normalize, and then it is aggregated and weighted with a value v containing camera information. To reduce the amount of calculation, AutoAlignV2 [118] is inspired by Deformable DETR [135] and proposes a Cross-Domain DeformCAFA operation. DeformCAFA uses a deformable cross-attention mechanism in which the query q and key value k still adopt the settings in AutoAlign. Value v has a new change. First, the projection matrix is used to query the image features corresponding to the voxel features. Then, the offset is learned through MLP, and the image feature corresponding to the offset is extracted as the value v. The cross-attention enables each voxel to perceive the entire image, enabling feature alignment of the two modalities. Two transformer decoders are used in Transfusion [121]. The first decoder layer utilizes a sparse set of object queries to generate initial bounding boxes from LiDAR features. The second one adaptively fuses object queries with useful image features associated with spatial and contextual relations.

C. Conclusion and Discussion

It is efficient to apply the camera projection matrix to align both the image and point cloud. Although the aggregation of features is carried out at the fine pixel level, the point cloud is sparse, and the image is dense. Use the projection matrix to find the corresponding relationship between the LiDAR point and the image pixel. The point cloud feature aggregates the image information in a coarse-grained manner through such a hard association, which can destroy the semantic information in the image. For example, a car has 100 points in the point cloud, while the car may have thousands of pixels in the corresponding image. Each point is projected to the image plane through the projection matrix. Although the feature alignment is at the per-pixel level, due to the sparsity of the point cloud, the image features may still lose the semantic information of the context. With a soft association mechanism, this method uses a cross-attention mechanism to find the correspondence between LiDAR points and image pixels. It can dynamically focus on pixel-level information from images. The features of each point cloud query the entire image, enabling point cloud features to aggregate image information in a fine-grained manner to obtain a pixel-level semantic alignment map. Although this method can better obtain the semantic information in the image, due to the use of the attention mechanism, each pixel in the image will be matched, and the model has a large amount of calculation and consumes more time. AutoAlignV2 [118] uses a DeformCAFA module to reduce the number of query image features and the amount of calculation.

V. Fusion

In this section, we summarize the fusion methods for multi-modal 3D object detection, which is always considered the most important part of multi-modal methods. Based on the fusion methods, the purpose of enhanced 3D object detection can be better achieved. For now, the most dominant fusion methods for multi-modal 3D object detection are represented by complementation, i.e., augmentations of one modality to another. After analysis, it is found that the multi-modal methods are mainly feature complementation of image features to point cloud features. In the field of 3D object detection, the detection accuracy of point clouds is much higher than that of images, as shown in **Fig.** 8. The lack of depth information in images leads to low accuracy of 3D object detection. Meanwhile, the image information has rich semantic information, which can be a data complement to the point cloud information.

The current multi-modal methods of complementation are performed by different fusion methods. The main difference is whether learning is required in the multi-modal 3D object detection fusion process. To help understand the existing fusion methods, we categorize them into two categories: **learning-agnostic** and **learning-based**. The learning-agnostic methods perform arithmetic operations and splicing operations on features. These methods are simple to operate and easy to calculate but do not have good scalability and robustness. The learning-based methods utilize attention to fuse features,

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3264658

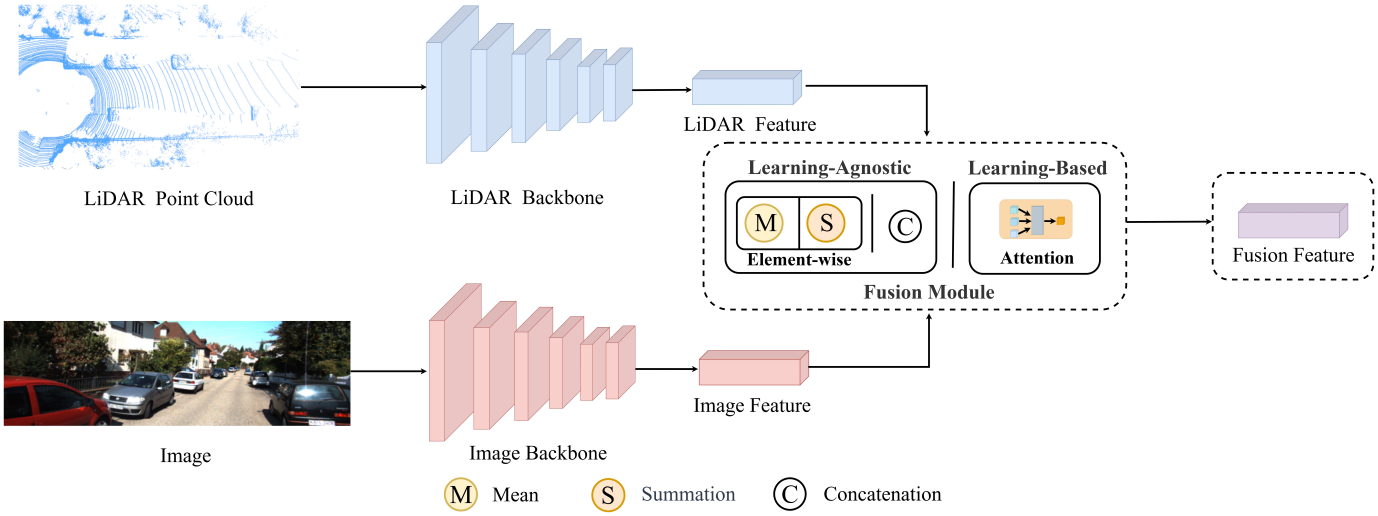


Fig. 7: An overview of the multi-modal 3D object detection fusion methods. We divide the fusion methods into two broad categories: Learning-agnostic and learning-based. Learning-agnostic methods adopt element-wise operations (mean, summation) and concatenation. Learning-based methods utilize the attention mechanism.

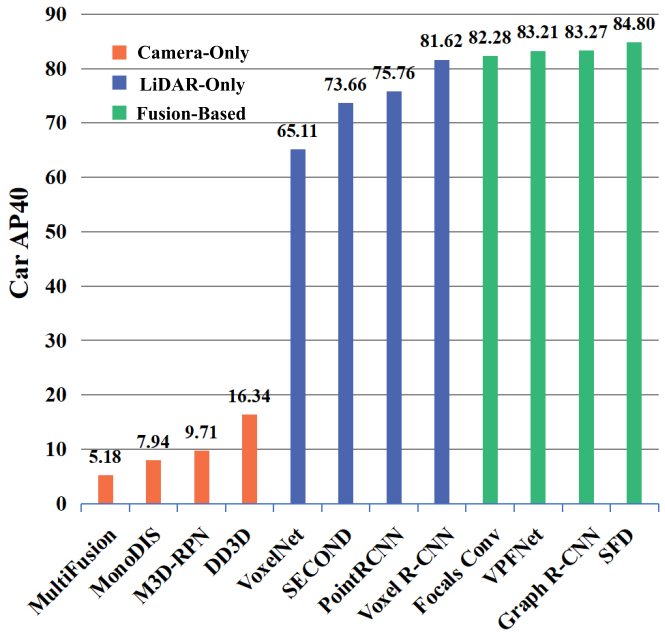


Fig. 8: Performance comparison of camera-only, LiDAR-only, and fusion-based methods. In the field of 3D object detection, the LiDAR-only detection methods have much higher detection accuracy than the camera-only detection methods. Images lack depth information, and point clouds can provide accurate 3D geometric information with higher detection accuracy. Therefore, a fusion of the two modal information can improve the detection accuracy of the model.

which is relatively complex and increases the number of parameters. However, the learning-based methods can focus on important information with high weights and ignore irrelevant information with low weights, so it has higher scalability and robustness. An overview of the multi-modal 3D object detection fusion methods is shown in **Fig.** 7 and **Tab.** VI.

A. Learning-Agnostic Fusion

Traditional fusion methods focus on arithmetic and concatenation operations on features. Learning-agnostic methods are one of these fusion methods using feature operations and concatenation. Learning-agnostic methods have two main types: element-wise operations (summation, mean) and concatenation.

Element-wise Operations. Element-wise operations utilize arithmetic operations to handle features of the same dimension (summation, mean). Element-wise operations are easy to parallel operation. It combines the two features into a composite vector. It has the advantage of simple computation and easy operation. Meanwhile, calculating mean values across different channels or getting the sum over increases the information on point cloud features, but the feature dimensionality does not increase. Only the amount of information under each dimension increases. The increase in the amount of information can improve detection accuracy.

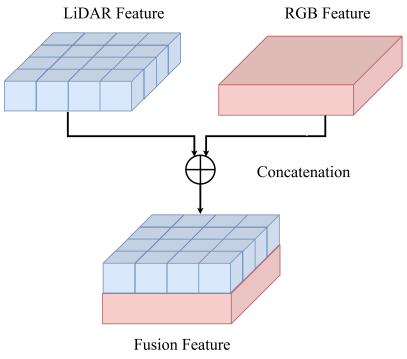


Fig. 9: An illustration of the concatenation fusion operation for multi-modal 3D object detection methods.

In early works, MV3D [52] is a pioneer in this way, which uses the mean value method to fuse the features of three dif-

TABLE VI: An overview of our taxonomy of multi-modal fusion."-" means not mentioned.

Type	Level	Advantages	Disadvantages	References
Learning-Agnostic				
Mean	Element-wise	+Simple calculation	-Data redundancies	[52], [53], [93], [126]
Summation	Element-wise	+Easy operation	-Loss of variability	[91], [103], [124]
Concatenation	Feature	+Effective retention of information	-High volume of calculations	[25], [92], [99], [101], [104], [105], [107], [114], [115], [122], [123], [125], [127]
Learning-Based				
Attention	-	+Get the global and local connections	-Large number of participants	[23], [58]–[60], [106], [108], [109], [111], [113], [117]–[121]

ferent views. The feature fusion process is easy to operate and simplifies the computation of the fusion process. AVOD [53] uses MV3D [52] as the baseline, which generates new fusion features from the feature maps of both views by element-wise mean. It inherits the advantage of small computation during MV3D fusion. This way, feature maps of the same shape can be fused efficiently. ContFuse [91] correlates features through sensor coordinate correspondence and uses element-wise sum to combine feature maps of the same dimensions element-wise to fuse different modality information. Element-wise fusion has only been adopted by a few methods in recent studies. This is mainly because element-wise does not accurately obtain the correct foreground information and usually carries noise. SCANet [93] and MMF [103], also adopt element-wise operations. However, unlike previous studies, MMF [103] utilizes multiple tasks to help detect and fuse features in the backbone. Focals Conv [124] proposes a lightweight fusion module that extracts the image feature using a semantic segmentation network and utilizes element-wise summation to aggregate images feature and voxel features.

Concatenation. Feature concatenation is to convert the transformed multi-modal features into the same feature vector size and then concatenate the image feature vector with the point cloud feature vector. An overview of the concatenation fusion methods is shown in **Fig.** 9. Unlike element-wise operations, concatenation operation is the merging of channel numbers which is more computationally intensive than element-wise operations. But it avoids the information loss caused by direct element-wise operations. Meanwhile, concatenation operation is not limited by the number of channels. Particularly, it is more popular among the multi-modal 3D object detection methods. PointFusion [99] is a pioneer in applying concatenation operations to multi-modal 3D object detection. The PointFusion approach is to concatenate the point-wise feature and the image feature to preserve the maximum information of each modality. VoxelNet [19] extends single-modal input to multi-modal input, enabling further performance improvements. MVX-Net [101] and SEG-VoxelNet [25] use the concatenation operation to complement the corresponding image features to the coordinates of the 3D points. Unlike element-wise operations, the concatenation operation can retain modal information to a greater extent and has shallow information loss. The PointPainting [104] method obtains the pixel segmentation score by a semantic segmentation network. This method uses a concatenation operation to fuse the segmented scores to complete the point cloud to preserve the point cloud information and the segmented scores. In the study of fusion methods, these previous multi-modal methods have been tried continuously, and it can be found that concatenation is simple to operate and can retain more feature information. Other multi-modal methods [105], [107], [114], [115], [122], [123], [125], [127] also utilize the concatenation operation to accomplish feature fusion.

B. Learning-Based Fusion

In 2020, DETR [98], [135]–[137] combines CNNs with attention for object detection tasks. DETR enables the whole network to achieve end-to-end object detection, significantly simplifying the object detection pipeline. Later DETR3D [98], [138] applies attention to 3D object detection. With the development of attention, cross-modal attention can provide a new way of fusion for multi-modal methods. The learning-based methods learn the weight distribution, where different parts of the input data or feature map have different weights. According to different weights, high weights are used to retain important information, while low weights ignore irrelevant information. The learning-based fusion methods have better robustness.

DETR is a landmark algorithm for attention-applied object detection. In the same year, some methods also try to apply attention to multi-modal 3D object detection in fusion methods,

TABLE VII: An overview of 3D object detection results on the nuScenes test set.

Model	NDS	mAP	FPS	Year	Publisher
Pointpainting [104]	58.1	46.4	-	2020	CVPR
MVP [110]	70.5	66.4	40	2021	NeurIPS
Fusionpainting [111]	70.4	66.3	-	2021	ITSC
BEVFusion [59]	**75.0**	**76.1**	8.4	2022	arXiv
BEVFusion [60]	60.4	53.5	-	2022	NerIPS
AutoAlign [117]	71.1	66.6	-	2022	IJCAI
AutoAlignV2 [118]	72.4	68.4	-	2022	ECCV
Transfusion [121]	71.7	68.9	-	2022	CVPR
FUTR3D [122]	68.0	64.2	-	2022	arXiv
VFF+CenterPoint [116]	72.4	68.4	-	2022	CVPR
UVTR [127]	71.1	67.1	-	2022	arXiv
PointAugmenting [115]	71.0	66.8	-	2021	CVPR
3D-CVF [23]	62.3	52.7	-	2020	ECCV
Fast-CLOCs [139]	68.7	63.1	-	2022	WACV

such as 3D-CVF [23], MVAF-Net [106], MAFF-Net [108], EPNet [109]. 3D-CVF proposes an adaptive gated fusion network, which generates attention significantly simplifying 3×3 convolutional layers and sigmoid functions. The attention mapping complements the projected image features into the point cloud features. This type of fusion allows better focusing of helpful information to be fused, making the fusion method learnable. The MVFF part of MVAF-Net [106], proposes to be combined with the APF module, adaptively fuses multi-task features using attention mechanisms. The MAFF-Net [108] model proposes the PointAttentionFusion (PAF) module. PAF fuses each 3D point using a fusion of one image feature and two attention features to achieve adaptive fusion features. Since the camera sensor is susceptible to the effects of lighting, occlusion, and other factors, under these influences, interference information is introduced in the process of complementing image features to point cloud features. To solve this problem, EPNet [109] uses the attention method to estimate the importance of images for fusion adaptively.

With the development of attention, many attention fusion methods have emerged in the field of multi-modal 3D object detection, such as FusionPainting [111], AutoAlign [117], AutoAlignV2 [118], DeepFusion [119], CAT-Det [120], BEVFusion [59], BEVFusion [60]. These models utilize attention fusion to fuse critical information with high weights and redundant information with low weights. This dramatically improves the fusion efficiency and prevents interference information from affecting the detection efficiency.

C. Conclusion and Discussion

In this chapter, we discuss multi-modal fusion methods and classify data fusion into learning-agnostic and learning-based categories. Learning-agnostic mainly consists of two operations, element-wise and concatenation operations, to estimate the importance of images for fusion adaptively.

Multi-modal fusion is an extensively researched topic. Many solutions have been proposed in this field, each with advantages and disadvantages. Learning-Agnostic methods are suitable for smaller datasets, while learning-based offer better robustness. Despite these advances, multi-modal fusion still faces the following challenges:

- The data information has different degrees of information loss in the feature transformation.
- The current fusion methods use image features to complement the point features, and the image features will have problems, such as domain gaps when using a point cloud baseline.
- Learning-agnostic methods need to consider the problem of fusion according to the importance of information
- Learning-based methods have many parameters and need to consider the problem of parameter number optimization.

VI. Challenges and Trends

Although many fusion methods have already been available, image and point cloud fusion algorithms in autonomous driving face many challenges due to the demands in accuracy, robustness, and real-time abilities. Besides, data alignment with the point cloud and the image is still under extensive exploration and is far from mature. This section discusses challenges and trends in multi-modal 3D object detection.

- *Data Noise.* How to effectively fuse multi-modal information has been the primary challenge in multi-modal learning. With various sensors, there is an information gap between the data from different modalities, leading to unsynchronized information. This issue introduces significant noise in feature fusion, which harms information representation learning. For example, the two-stage detector causes the incorporation of background features in the image due to the presence of ROI in different dimensions during fusion. Some recent works [59], [60] utilize BEV representation to unify different heterogeneous modalities, which provides a new perspective on solving this problem and is worth further exploration.
- *Limited Reception Field in Open-source Datasets.* Insufficient sensor coverage is detrimental to the performance of multi-modal detection. Recently, increasing multi-modal works have focused on the nuScenes [79] because of its excellent perceptual range (360 degrees for both point clouds and cameras). An excellent perceptual scope facilitates multi-modal learning, especially for autonomous driving perception tasks. Probably the utilization of sensors with excellent sensing range, such as nuScenes [79] and Waymo [80], could improve the coverage of multi-modal detection systems and enhance their performance in complex environments, which may provide a possible idea to solve the problem of limited reception field in open-source datasets.
- *Compact Representation.* Compact representations contain more information with fewer data scales. Although existing work [94], [95] attempts to encode sparse 3D representations into two-dimensional representations, there is a significant loss of information in the encoding procedure. The projection of a range image can cause multiple

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3264658

TABLE VIII: An overview of 3D object detection results for car, pedestrian, and cyclist the KITTI test set. "Mod." represents moderate. "-" means not mentioned.

Method	Publisher	Car			Pedestrian			Cyclist			FPS	GPU
		Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard		
MV3D [52]	2017 CVPR	74.97	63.63	54.00	-	-	-	-	-	-	2.8	TITAN X
AVOD [53]	2018 IROS	81.94	71.88	66.38	50.80	42.81	40.88	64.00	52.18	46.61	12.5	TITAN XP
ContFuse [91]	2018 ECCV	82.54	66.22	64.04	-	-	-	-	-	-	16.7	-
PointFusion [99]	2018 CVPR	77.92	63.00	53.27	33.36	28.04	23.38	49.34	29.42	26.98	-	-
IPOD [31]	2018 arXiv	79.75	72.57	66.33	56.92	44.68	42.39	71.40	53.46	48.34	-	Tesla P40
Frustum PointNets [54]	2018 CVPR	81.20	70.39	62.19	51.21	44.89	40.23	71.96	56.77	50.39	5.9	GTX 1080
Complexer-YOLO [102]	2019 CVPR	55.63	49.44	44.13	19.45	15.32	14.80	28.36	23.48	22.85	15.6	GTX 1080Ti
Mvx-Net(PF) [101]	2019 ICRA	83.20	72.70	65.20	-	-	-	-	-	-	6.0	-
RoarNet [100]	2019 IV	83.71	73.04	59.16	-	-	-	-	-	-	10.0	TITAN XP
MMF [103]	2019 CVPR	86.81	76.75	68.41	-	-	-	-	-	-	12.5	-
SCANet [93]	2019 ICASSP	76.09	66.30	58.68	-	-	-	-	-	-	11.0	GTX 1080Ti
VMVS [126]	2019 IROS	-	-	-	53.98	45.01	41.72	-	-	-	-	-
MLOD [92]	2019 ITSC	72.24	64.20	57.20	48.26	40.97	35.74	67.66	49.89	42.23	-	-
Frustum ConvNet [56]	2019 IROS	85.88	76.51	68.08	52.37	45.61	41.49	79.58	64.68	57.03	21.3	GTX 1080
Pointpainting [104]	2020 CVPR	82.11	71.70	67.08	50.32	40.97	37.87	77.63	63.78	55.89	2.50	-
PI-RCNN [105]	2020 AAAI	84.37	74.82	70.03	-	-	-	-	-	-	11.1	TITAN RTX
EPNet [109]	2020 ECCV	89.81	79.28	74.59	-	-	-	-	-	-	10.0	Titan XP
MAFF-Net [108]	2020 arXiv	85.52	75.04	67.61	-	-	-	-	-	-	24.0	Tesla V100
SemanticVoxcls [107]	2020 MFI	-	-	-	50.90	42.19	39.52	-	-	-	-	-
MVAF-Net [106]	2020 arXiv	87.87	78.71	75.48	-	-	-	-	-	-	15.0	-
3D-CVF [23]	2020 ECCV	89.20	80.05	73.11	-	-	-	-	-	-	13.3	GTX 1080Ti
CLOCs [140]	2020 IROS	88.94	80.67	77.15	-	-	-	-	-	-	6.7	-
PFF3D [141]	2021 Access	81.11	72.93	67.24	43.93	36.07	32.86	63.27	46.78	41.37	-	-
Frustum-pointpillars [55]	2021 ICCV	-	-	-	51.22	42.89	39.28	-	-	-	14.3	GTX 1080
MSF-MC [113]	2021 ICCV	89.63	80.06	75.83	**66.69**	**58.18**	**56.21**	**82.36**	59.17	58.71	-	-
RoIFusion [114]	2021 Access	88.32	79.54	74.47	42.22	35.14	32.92	80.84	64.05	58.37	-	GTX 1080Ti
Cross-Modality [112]	2021 WACV	87.22	77.28	72.04	-	-	-	-	-	-	-	RTX 2080Ti
Faraway-Frustum [57]	2021 ITSC	87.45	79.05	76.14	46.33	38.58	35.71	77.36	62.00	55.40	-	-
Autoalign [117]	2022 arXiv	86.84	77.58	73.23	53.99	44.08	40.82	80.41	64.36	56.88	-	-
CAT-Det [120]	2022 CVPR	89.87	81.32	76.68	54.26	45.44	41.94	83.68	**68.81**	**61.45**	-	GTX 1080Ti
VFF + Voxel R-CNN [116]	2022 CVPR	89.50	82.09	79.29	-	-	-	-	-	-	-	-
VFF + PV-RCNN [116]	2022 CVPR	89.58	81.97	79.17	-	-	-	-	-	-	-	-
Fast-CLOCs [139]	2022 WACV	89.11	80.34	76.98	52.10	42.72	39.08	82.83	65.31	57.43	8.0	RTX 3080
SFD [58]	2022 CVPR	91.73	**84.80**	**84.76**	-	-	-	-	-	-	10.2	RTX 2080Ti
Graph R-CNN [123]	2022 ECCV	**91.89**	83.27	77.78	-	-	-	-	-	-	-	-
Focals Conv [124]	2022 CVPR	90.50	82.28	77.59	-	-	-	-	-	-	8.0	-
VPFNet [125]	2022 TMM	91.02	83.21	78.20	-	-	-	-	-	-	15.0	RTX 2080Ti

points to fall into the same pixel, which results in a loss of information. Recently, the Waymo Open Dataset [80] has provided high-resolution range images, but only a small amount of work has examined them. High-quality representation still remains an open challenge. Maybe a more compact 3D representation can be achieved using advanced coding techniques, such as using deep learning-based autoencoders and generative adversarial networks to represent 3D features.

- *Information Loss.* How to maximize the retention of multi-modal information has been one of the critical challenges in multi-modal 3D object detection [142]–[144]. The fusion of information from multiple modalities can lead to the loss of information. For example, in the fusion stage, the image semantic information is lost when the image is complemented to the point cloud features. This causes the fusion process not to make better use of the image feature information, resulting in sub-optimal model performance. State-of-the-art models [109], [117] in multi-modal learning may prove beneficial for sensor fusion in 3D object detection, and new fusion methods and neural network architectures can be explored that

maximize the preservation of multi-modal information.

- *Unlabeled Data.* Unlabeled data is prevalent in autonomous driving scenarios, and unsupervised learning can provide more robust learning of representations, which has been studied to some extent in similar tasks, such as 2D object detection [145]–[153]. However, there is no convincing research on unsupervised representation in the current multi-modal 3D object detection. Especially in the field of multi-modal research, it is a challenging research topic to perform better-unsupervised learning of multi-modal representations. In future research, the difficulty of unsupervised learning characterization will revolve around multi-modal differences to characterize multi-modal data simultaneously.
- *High Computation Complexity.* One of the important challenges of multi-modal 3D object detection is to detect the object quickly and in real-time in autonomous driving scenarios. Since multi-modal methods need to process multiple information, which leads to increased parameters and computation, longer training time, and inference time, the application cannot meet the real-time performance. Recent multi-modal methods also consider real-time, e.g., MVP [110], BEVFusion [59] experiments on nuScenes dataset have used FPS as a model evaluation metric. As shown in **Tab.** VII. To mitigate the issue of high computational complexity, it would be encouraged future works to explore the model pruning and quantization techniques [154], [155]. These techniques aim to streamline the model structure and reduce the model parameters for efficient model deployment, which requires further study in autonomous driving scenarios.
- *Long Tail Effect.* How to solve the long-tail effect caused by varied performance is one of the important challenges of multi-modal 3D object detection. In the autonomous driving domain, most models are required to detect cars, but other objects, such as pedestrians, are also essential detection requirements. As shown in **Tab.** VIII, there are many categories in the autonomous driving scenario. The models which are efficient in detecting cars may be inefficient in detecting the pedestrian, for example, the SFD [58]. This leads to uneven category detection. In future work, it may be possible to explore using loss functions and sampling strategies as a potential solution to address the aforementioned problem.
- *Cross-Modal Data Augmentation.* Data augmentation is a key part of achieving competitive results in 3D object detection, but data augmentation is mostly applied in single-modality methods and rarely considered in multi-modal scenarios [156]–[161]. Since point clouds and images are two kinds of heterogeneous data, it is difficult to achieve cross-modal synchronous enhancement, which will lead to serious cross-modal misalignment. Applying gt-aug to point cloud and camera data without distortion is difficult. In some methods, only the point cloud part is enhanced, and the image part is ignored. There are also some methods to keep the original image unchanged and perform the reverse transformation in the point cloud to achieve the purpose of image point cloud correspondence. Pointaugmenting [115] proposes a more complex cross-modal data augmentation method, but uses additional mask annotations on the image branch and is prone to noise. None of these methods are good at solving the synchronous problem of crossmodal data augmentation. One potential solution to address this challenge is through representation reconstruction, which converts heterogeneous data into a unified representation and enables simultaneous data augmentation.
- *Temporal synchronization.* Temporal synchronization is a crucial issue in multi-modal 3D object detection. Due to the differences in the sampling rate, working mode, and acquisition speed of different sensors, there is a time deviation between the data collected by the sensors, which leads to the misalignment of multi-modal data, which in turn affects the accuracy and efficiency of multi-modal 3D object detection. First, there may be errors in the temporal stamps of different sensors. Even if the hardware is used for timing synchronization, it is difficult to guarantee the consistency of sensor temporal stamps fully. This method may require expensive equipment. Software synchronization methods can be used, such as the time stamp interpolation method, temporal synchronization algorithm based on the Kalman filter, and time synchronization algorithm based on deep learning, etc. Secondly, sensor data has frame loss or delay, which also affects the accuracy of multi-modal 3D object detection. The idea to solve the problem is to use a cache mechanism to deal with delayed or missing data and use data interpolation or extrapolation methods to fill in the vacant part of the data. Temporal synchronization in multi-modal 3D object detection is a complex problem that needs to be solved using various technical means.

VII. Conclusion

In this paper, we comprehensively review and analyze various aspects of multi-modal 3D object detection. We first analyze the reasons for the emergence of multi-modal 3D object detection, introduce existing datasets and evaluation metrics, and comprehensively compare datasets. We present a new classification method for multi-modal 3D object detection. Specifically, the existing methods are analyzed from three perspectives of data representation, feature alignment, and feature fusion. The advantages and disadvantages of classification methods from different perspectives are reviewed in detail. Finally, we summarize the recent trends, present challenges, and problems, and look forward to the future research direction of multi-modal 3D object detection.

References

[1] W. H. Organization *et al.*, "European regional status report on road safety 2019," 2020.
[2] J. Wang, L. Zhang, Y. Huang, and J. Zhao, "Safety of autonomous vehicles," *Journal of advanced transportation*, vol. 2020, 2020.
[3] R. Mariani, "An overview of autonomous vehicles safety," in *2018 IEEE International Reliability Physics Symposium (IRPS)*. IEEE, 2018, pp. 6A–1.

[4] P. Koopman, U. Ferrell, F. Fratrik, and M. Wagner, "A safety standard approach for fully autonomous vehicles," in *International Conference on Computer Safety, Reliability, and Security*. Springer, 2019, pp. 326–332.

[5] J. Mao, S. Shi, X. Wang, and H. Li, "3d object detection for autonomous driving: a review and new outlooks," *arXiv preprint arXiv:2206.09474*, 2022.

[6] K. Huang, B. Shi, X. Li, X. Li, S. Huang, and Y. Li, "Multi-modal sensor fusion for auto driving perception: A survey," *arXiv preprint arXiv:2202.02703*, 2022.

[7] X. Zhang, H. Yu, and Y. Zhuang, "A robust rgb-d visual odometry with moving object detection in dynamic indoor scenes," *IET Cyber-Systems and Robotics*, vol. 5, no. 1, p. e12079, 2023.

[8] K. Wang, Y. Wang, S. Zhang, Y. Tian, and D. Li, "Slms-ssd: Improving the balance of semantic and spatial information in object detection," *Expert Systems with Applications*, vol. 206, p. 117682, 2022.

[9] G. Guo and S. Zhao, "3d multi-object tracking with adaptive cubature kalman filter for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, 2022.

[10] Y. Ma, J. Zhang, G. Qin, J. Jin, K. Zhang, D. Pan, and M. Chen, "3d multi-object tracking based on dual-tracker and ds evidence theory," *IEEE Transactions on Intelligent Vehicles*, 2022.

[11] K. Wang, T. Zhou, X. Li, and F. Ren, "Performance and challenges of 3d object detection methods in complex scenes for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, 2022.

[12] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[13] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3d object detection methods for autonomous driving applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3782–3795, 2019.

[14] Y. Cui, R. Chen, W. Chu, L. Chen, D. Tian, Y. Li, and D. Cao, "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 722–739, 2021.

[15] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3d point clouds: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 12, pp. 4338–4364, 2020.

[16] Y. Wang, Q. Mao, H. Zhu, Y. Zhang, J. Ji, and Y. Zhang, "Multi-modal 3d object detection in autonomous driving: a survey," *arXiv preprint arXiv:2106.12735*, 2021.

[17] D. Luo, Y. Zhuang, and S. Wang, "Hybrid sparse monocular visual odometry with online photometric calibration," *The International Journal of Robotics Research*, vol. 41, no. 11-12, pp. 993–1021, 2022.

[18] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3d point clouds: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 12, pp. 4338–4364, 2020.

[19] Y. Zhou and 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.

[20] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[21] D. Z. Wang and I. Posner, "Voting for voting in online point cloud object detection." in *Robotics: Science and Systems*, vol. 1, no. 3. Rome, Italy, 2015, pp. 10–15.

[22] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, and I. Posner, "Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1355–1361.

[23] J. H. Yoo, Y. Kim, J. Kim, and 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.

[24] R. Ge, Z. Ding, Y. Hu, Y. Wang, S. Chen, L. Huang, and Y. Li, "Afdet: Anchor free one stage 3d object detection," *arXiv preprint arXiv:2006.12671*, 2020.

[25] J. Dou, J. Xue, and J. Fang, "Seg-voxelnet for 3d vehicle detection from rgb and lidar data," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4362–4368.

[26] Q. Chen, L. Sun, Z. Wang, K. Jia, and A. Yuille, "Object as hotspotcs: An anchor-free 3d object detection approach via firing of hotspots," in *European conference on computer vision*. Springer, 2020, pp. 68–84.

[27] C. R. Qi, H. Su, K. Mo, and 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.

[28] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[29] S. Shi, X. Wang, and H. Li, "Pointrcnn: 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.

[30] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[31] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Ipod: Intensive point-based object detector for point cloud," *arXiv preprint arXiv:1812.05276*, 2018.

[32] W. Shi and 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.

[33] J. Ngiam, B. Caine, W. Han, B. Yang, Y. Chai, P. Sun, Y. Zhou, X. Yi, O. Alsharif, P. Nguyen *et al.*, "Starnet: Targeted computation for object detection in point clouds," *arXiv preprint arXiv:1908.11069*, 2019.

[34] Z. Yang, Y. Sun, S. Liu, X. Shen, and 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.

[35] T. Gao, H. Pan, and H. Gao, "Monocular 3d object detection with sequential feature association and depth hint augmentation," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 2, pp. 240–250, 2022.

[36] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and 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.

[37] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3d bounding box estimation using deep learning and geometry," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 7074–7082.

[38] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teuliere, and T. Chateau, "Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2040–2049.

[39] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," in *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2017, pp. 924–933.

[40] A. Kundu, Y. Li, and J. M. Rehg, "3d-rcnn: Instance-level 3d object reconstruction via render-and-compare," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3559–3568.

[41] B. Xu and Z. Chen, "Multi-level fusion based 3d object detection from monocular images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2345–2353.

[42] T. He and S. Soatto, "Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8409–8416.

[43] J. Chang and G. Wetzstein, "Deep optics for monocular depth estimation and 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 193–10 202.

[44] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6569–6578.

[45] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3d object detection," *arXiv preprint arXiv:1811.08188*, 2018.

[46] C. Reading, A. Harakeh, J. Chae, and 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.

[47] T. Hoang, J. Lawall, Y. Tian, R. J. Oentaryo, and D. Lo, "Patchnet: Hierarchical deep learning-based stable patch identification for the linux kernel," *IEEE Transactions on Software Engineering*, vol. 47, no. 11, pp. 2471–2486, 2019.

[48] S. Luo, H. Dai, L. Shao, and Y. Ding, "M3dssd: Monocular 3d single stage object detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6145–6154.

[49] X. Shi, Q. Ye, X. Chen, C. Chen, Z. Chen, and T.-K. Kim, "Geometry-based distance decomposition for monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 172–15 181.

[50] H. Chen, Y. Huang, W. Tian, Z. Gao, and L. Xiong, "Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 379–10 388.

[51] L. Wang, L. Du, X. Ye, Y. Fu, G. Guo, X. Xue, J. Feng, and L. Zhang, "Depth-conditioned dynamic message propagation for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 454–463.

[52] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.

[53] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and 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.

[54] C. R. Qi, W. Liu, C. Wu, H. Su, and 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.

[55] A. Paigwar, D. Sierra-Gonzalez, Ö. Erkent, and 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.

[56] Z. Wang and K. Jia, "Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1742–1749.

[57] H. Zhang, D. Yang, E. Yurtsever, K. A. Redmill, and Ü. Özgüner, "Faraway-frustum: Dealing with lidar sparsity for 3d object detection using fusion," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 2646–2652.

[58] X. Wu, L. Peng, H. Yang, L. Xie, C. Huang, C. Deng, H. Liu, and 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.

[59] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," *arXiv preprint arXiv:2205.13542*, 2022.

[60] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "Bevfusion: A simple and robust lidar-camera fusion framework," *arXiv preprint arXiv:2205.13790*, 2022.

[61] K. Huang, B. Shi, X. Li, X. Li, S. Huang, and Y. Li, "Multi-modal sensor fusion for auto driving perception: A survey," *arXiv preprint arXiv:2202.02703*, 2022.

[62] Y. Wu, Y. Wang, S. Zhang, and H. Ogai, "Deep 3d object detection networks using lidar data: A review," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 1152–1171, 2020.

[63] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[64] M. Abdelfattah, "Adversarial attacks on multi-modal 3d detection models," Ph.D. dissertation, University of British Columbia, 2021.

[65] J. L. Gómez, G. Villalonga, and A. M. López, "Co-training for deep object detection: Comparing single-modal and multi-modal approaches," *Sensors*, vol. 21, no. 9, p. 3185, 2021.

[66] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[67] W. Liu, M. Hua, Z. Deng, Y. Huang, C. Hu, S. Song, L. Gao, C. Liu, L. Xiong, and X. Xia, "A systematic survey of control techniques and applications: From autonomous vehicles to connected and automated vehicles," *arXiv preprint arXiv:2303.05665*, 2023.

[68] L. Xiong, X. Xia, Y. Lu, W. Liu, L. Gao, S. Song, and Z. Yu, "Imu-based automated vehicle body sideslip angle and attitude estimation aided by gnss using parallel adaptive kalman filters," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 10 668–10 680, 2020.

[69] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2vnet: Vehicle-to-vehicle communication for joint perception and prediction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 605–621.

[70] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*. Springer, 2022, pp. 107–124.

[71] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 541–29 552, 2021.

[72] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers," *arXiv preprint arXiv:2207.02202*, 2022.

[73] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," *arXiv preprint arXiv:2209.12836*, 2022.

[74] R. Qian, X. Lai, and X. Li, "3d object detection for autonomous driving: a survey," *Pattern Recognition*, p. 108796, 2022.

[75] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3d object detection methods for autonomous driving applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3782–3795, 2019.

[76] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu *et al.*, "Multibench: Multiscale benchmarks for multimodal representation learning," *arXiv preprint arXiv:2107.07502*, 2021.

[77] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *Digital Signal Processing*, p. 103514, 2022.

[78] A. Geiger, P. Lenz, and 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.

[79] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[80] P. Sun, H. Kretzschmar, 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.

[81] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "Kaist multi-spectral day/night data set for autonomous and assisted driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 934–948, 2018.

[82] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 954–960.

[83] A. Patil, S. Malla, H. Gang, and 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.

[84] K. R., U. M., H. J., P. T., N. K., F. A., Y. M., L. B., J. A., O. P., O. S., S. S., K. A., K. A., T. C., P. L., J. W., and S. V. (2019) Lyft level 5 av dataset 2019. [Online]. Available: https://level5.lyft.com/dataset/

[85] 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.

[86] Q.-H. Pham, P. Sevestre, R. S. Pahwa, H. Zhan, C. H. Pang, Y. Chen, A. Mustafa, V. Chandrasekhar, and J. Lin, "A* 3d dataset: Towards autonomous driving in challenging environments," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2267–2273.

[87] 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.

[88] Z. Wang, S. Ding, Y. Li, J. Fenn, S. Roychowdhury, A. Wallin, L. Martin, S. Ryvola, G. Sapiro, and 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.

[89] 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," *arXiv preprint arXiv:2106.11037*, 2021.

[90] Z. Wang, W. Zhan, and M. Tomizuka, "Fusing bird's eye view lidar point cloud and front view camera image for 3d object detection," in *2018 IEEE intelligent vehicles symposium (IV)*. IEEE, 2018, pp. 1–6.

[91] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 641–656.

[92] J. Deng and K. Czarnecki, "Mlod: A multi-view 3d object detection based on robust feature fusion method," in *2019 IEEE intelligent transportation systems conference (ITSC)*. IEEE, 2019, pp. 279–284.

[93] H. Lu, X. Chen, G. Zhang, Q. Zhou, Y. Ma, and Y. Zhao, "Scanet: Spatial-channel attention network for 3d object detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1992–1996.

[94] Z. Zhang, Z. Liang, M. Zhang, X. Zhao, H. Li, M. Yang, W. Tan, and S. Pu, "Rangelvdet: Boosting 3d object detection in lidar with range image and rgb image," *IEEE Sensors Journal*, vol. 22, no. 2, pp. 1391–1403, 2022.

[95] X. Zhang, L. Wang, G. Zhang, T. Lan, H. Zhang, L. Zhao, J. Li, L. Zhu, and H. Liu, "Ri-fusion: 3d object detection using enhanced point features with range-image fusion for autonomous driving," *IEEE Transactions on Instrumentation and Measurement*, 2022.

[96] J. Huang, G. Huang, Z. Zhu, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[97] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," *arXiv preprint arXiv:2203.17270*, 2022.

[98] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and 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.

[99] D. Xu, D. Anguelov, and 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.

[100] K. Shin, Y. P. Kwon, and 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.

[101] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvx-net: Multimodal voxelnet for 3d object detection," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7276–7282.

[102] M. Simon, K. Amende, A. Kraus, J. Honer, T. Samann, H. Kaulbersch, S. Milz, and 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. 0–0.

[103] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "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.

[104] S. Vora, A. H. Lang, B. Helou, and 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.

[105] L. Xie, C. Xiang, Z. Yu, G. Xu, Z. Yang, D. Cai, and X. He, "Pi-rcnn: An efficient multi-sensor 3d object detector with point-based attentive cont-conv fusion module," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 460–12 467.

[106] G. Wang, B. Tian, Y. Zhang, L. Chen, D. Cao, and J. Wu, "Multi-view adaptive fusion network for 3d object detection," *arXiv preprint arXiv:2011.00652*, 2020.

[107] J. Fei, W. Chen, P. Heidenreich, S. Wirges, and C. Stiller, "Semanticvoxels: Sequential fusion for 3d pedestrian detection using lidar point cloud and semantic segmentation," in *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 2020, pp. 185–190.

[108] Z. Zhang, Y. Shen, H. Li, X. Zhao, M. Yang, W. Tan, S. Pu, and H. Mao, "Maff-net: Filter false positive for 3d vehicle detection with multi-modal adaptive feature fusion," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 369–376.

[109] T. Huang, Z. Liu, X. Chen, and X. Bai, "Epnet: Enhancing point features with image semantics for 3d object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 35–52.

[110] T. Yin, X. Zhou, and P. Krähenbühl, "Multimodal virtual point 3d detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 494–16 507, 2021.

[111] S. Xu, D. Zhou, J. Fang, J. Yin, Z. Bin, and 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.

[112] M. Zhu, C. Ma, P. Ji, and X. Yang, "Cross-modality 3d object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3772–3781.

[113] Z. Wang, Z. Zhao, Z. Jin, Z. Che, J. Tang, C. Shen, and Y. Peng, "Multi-stage fusion for multi-class 3d lidar detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3120–3128.

[114] C. Chen, L. Z. Fragonara, and A. Tsourdos, "Roifusion: 3d object detection from lidar and vision," *IEEE Access*, vol. 9, pp. 51 710–51 721, 2021.

[115] C. Wang, C. Ma, M. Zhu, and 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. 11 794–11 803.

[116] Y. Li, X. Qi, Y. Chen, L. Wang, Z. Li, J. Sun, and J. Jia, "Voxel field fusion for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1120–1129.

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

[118] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao, "Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3d object detection," *arXiv preprint arXiv:2207.10316*, 2022.

[119] 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. 17 182–17 191.

[120] Y. Zhang, J. Chen, and D. Huang, "Cat-det: Contrastively augmented transformer for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 908–917.

[121] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and 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.

[122] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "Futr3d: A unified sensor fusion framework for 3d detection," *arXiv preprint arXiv:2203.10642*, 2022.

[123] H. Yang, Z. Liu, X. Wu, W. Wang, W. Qian, X. He, and D. Cai, "Graph r-cnn: Towards accurate 3d object detection with semantic-decorated local graph," in *European Conference on Computer Vision*. Springer, 2022, pp. 662–679.

[124] Y. Chen, Y. Li, X. Zhang, J. Sun, and J. Jia, "Focal sparse convolutional networks for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5428–5437.

[125] H. Zhu, J. Deng, Y. Zhang, J. Ji, Q. Mao, H. Li, and Y. Zhang, "Vpfnet: Improving 3d object detection with virtual point based lidar and stereo data fusion," *IEEE Transactions on Multimedia*, 2022.

[126] J. Ku, A. D. Pon, S. Walsh, and S. L. Waslander, "Improving 3d object detection for pedestrians with virtual multi-view synthesis orientation estimation," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 3459–3466.

[127] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3d object detection," *arXiv preprint arXiv:2206.00630*, 2022.

[128] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[129] Z. Liu, B. Li, X. Chen, X. Wang, X. Bai *et al.*, "Epnet++: Cascade bi-directional fusion for multi-modal 3d object detection," *arXiv preprint arXiv:2112.11088*, 2021.

[130] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, 2021.

[131] T. Yin, X. Zhou, and P. Krähenbühl, "Multimodal virtual point 3d detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 494–16 507, 2021.

[132] J. Wang, Z. Wei, T. Zhang, and W. Zeng, "Deeply-fused nets," *arXiv preprint arXiv:1605.07716*, 2016.

[133] G. Larsson, M. Maire, and G. Shakhnarovich, "Fractalnet: Ultra-deep neural networks without residuals," *arXiv preprint arXiv:1605.07648*, 2016.

[134] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *European Conference on Computer Vision*. Springer, 2020, pp. 194–210.

[135] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[136] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 213–229.

[137] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[138] Y. Wang, T. Ye, L. Cao, W. Huang, F. Sun, F. He, and D. Tao, "Bridged transformer for vision and point cloud 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 114–12 123.

[139] S. Pang, D. Morris, and 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.

[140] S. Pang, D. Morris, and 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. 10 386–10 393.

[141] L.-H. Wen and K.-H. Jo, "Fast and accurate 3d object detection for lidar-camera-based autonomous vehicles using one shared voxel-based backbone," *IEEE Access*, vol. 9, pp. 22 080–22 089, 2021.

[142] W. G. Unruh and R. M. Wald, "Information loss," *Reports on Progress in Physics*, vol. 80, no. 9, p. 092002, 2017.

[143] S. Graham, H. Chen, J. Gamper, Q. Dou, P.-A. Heng, D. Snead, Y. W. Tsang, and N. Rajpoot, "Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images," *Medical image analysis*, vol. 52, pp. 199–211, 2019.

[144] K. Oono and T. Suzuki, "Graph neural networks exponentially lose expressive power for node classification," *arXiv preprint arXiv:1905.10947*, 2019.

[145] E. Xie, J. Ding, W. Wang, X. Zhan, H. Xu, P. Sun, Z. Li, and P. Luo, "Detco: Unsupervised contrastive learning for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8392–8401.

[146] Z. Dai, B. Cai, Y. Lin, and J. Chen, "Up-detr: Unsupervised pre-training for object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1601–1610.

[147] A. Bar, X. Wang, V. Kantorov, C. J. Reed, R. Herzig, G. Chechik, A. Rohrbach, T. Darrell, and A. Globerson, "Detreg: Unsupervised pre-training with region priors for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 605–14 615.

[148] Y. Yang, A. Loquercio, D. Scaramuzza, and S. Soatto, "Unsupervised moving object detection via contextual information separation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 879–888.

[149] S. Jin, A. RoyChowdhury, H. Jiang, A. Singh, A. Prasad, D. Chakraborty, and E. Learned-Miller, "Unsupervised hard example mining from videos for improved object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 307–324.

[150] D. Guan, J. Huang, A. Xiao, S. Lu, and Y. Cao, "Uncertainty-aware unsupervised domain adaptation in object detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 2502–2514, 2021.

[151] Q. Xu, Y. Zhou, W. Wang, C. R. Qi, and D. Anguelov, "Spg: Unsupervised domain adaptation for 3d object detection via semantic point generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 446–15 456.

[152] P. Oza, V. A. Sindagi, V. VS, and V. M. Patel, "Unsupervised domain adaptation of object detectors: A survey," *arXiv preprint arXiv:2105.13502*, 2021.

[153] Y. Wang, Y. Chen, and Z. Zhang, "4d unsupervised object discovery," *arXiv preprint arXiv:2210.04801*, 2022.

[154] M. Figurnov, M. D. Collins, Y. Zhu, L. Zhang, J. Huang, D. Vetrov, and R. Salakhutdinov, "Spatially adaptive computation time for residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1039–1048.

[155] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370–403, 2021.

[156] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[157] B. Zoph, E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," in *European conference on computer vision*. Springer, 2020, pp. 566–583.

[158] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6256–6268, 2020.

[159] T. Zhao, Y. Liu, L. Neves, O. Woodford, M. Jiang, and N. Shah, "Data augmentation for graph neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 11 015–11 023.

[160] Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, and H. Xu, "Time series data augmentation for deep learning: A survey," *arXiv preprint arXiv:2002.12478*, 2020.

[161] W. Lu, D. Zhao, C. Premebida, L. Zhang, W. Zhao, and D. Tian, "Improving 3d vulnerable road user detection with point augmentation," *IEEE Transactions on Intelligent Vehicles*, 2023.

Li Wang was born in Shangqiu, Henan Province, China in 1990. He received his Ph.D. degree in mechatronic engineering at State Key Laboratory of Robotics and System, Harbin Institute of Technology, in 2020.

He was a visiting scholar at Nanyang Technology of University for two years. Currently, he is a postdoctoral fellow in the State Key Laboratory of Automotive Safety and Energy, and the School of Vehicle and Mobility, Tsinghua University.

Dr. Wang is the author of more than 20 SCI/EI articles. His research interests include autonomous driving perception, 3D robot vision and multi-modal fusion.

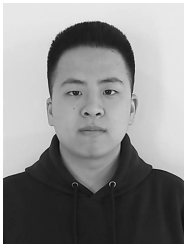
Xinyu Zhang was born in Huining, Gansu Province, and he received a B.E. degree from the School of Vehicle and Mobility at Tsinghua University, in 2001.

He was a visiting scholar at the University of Cambridge. He is currently a researcher with the School of Vehicle and Mobility, and the head of the Mengshi Intelligent Vehicle Team at Tsinghua University.

Dr. Zhang is the author of more than 30 SCI/EI articles. His research interests include intelligent driving and multi-modal information fusion.

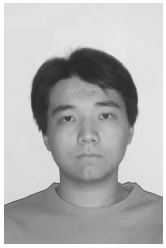
Ziying Song was born in Xingtai, Hebei Province, China, in 1997. He received his B.S. degree from Hebei Normal University of Science and Technology (China) in 2019. He received a master's degree from Hebei University of Science and Technology (China) in 2022. He is now a Ph.D. student majoring in Computer Science and Technology at Beijing Jiaotong University (China), with research focus on Computer Vision.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3264658



Jiangfeng Bi was born in Shijiazhuang, Hebei Province, China, in 1998. He received his B.S. degree from Hebei University of Science and Technology (China) in 2021. He is now a master's student majoring in Computer Science and Technology at the Hebei University of Science and Technology (China), mainly engaging in computer vision-related research during school.

Since Jul. 2022, he has been interning at the State Key Laboratory of Automotive Safety and Energy and the School of Vehicle and Mobility, Tsinghua University, responsible for developing 3D object detection based on the point cloud.



Guoxin Zhang was born in 1998 in Xingtai, Hebei Province, China. In 2021, he received his Bachelor's degree. He is now studying for his master's degree at the Hebei University of Science and Technology (China). His research interests are in computer vision.

Since May 2022, he has been interning at the State Key Laboratory of Automotive Safety and Energy and the School of Vehicle and Mobility, Tsinghua University, responsible for developing multi-sensor fusion 3D detection algorithms for smart vehicles.

Haiyue Wei was born in Shijiazhuang, Hebei Province, China, in 1997. He received his B.S. degree from the Hebei University of Science and Technology (China) in 2020. He is now a master's student majoring in Computer Science and Technology at the Hebei University of Science and Technology (China), mainly engaging in computer vision-related research.

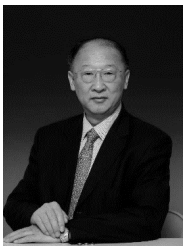
Since Jul. 2022, he has been interning at the State Key Laboratory of Automotive Safety and Energy and the School of Vehicle and Mobility, Tsinghua University, responsible for developing multi-sensor fusion 3D detection algorithms for smart vehicles.

Liyao Tang was born in Guangdong Province, China in 1996. He received his B.S. degree from the Beijing Institute of Technology and his master's degree from the University of Sydney.

He is now pursuing a Ph.D. degree in the School of Computer Science at the University of Sydney, and his research interests include computer vision, 3D scene understanding, and point cloud processing.

Lei Yang was born in DaTong, ShanXi Province, China in 1993. He received his master degree in robotics at Beihang University, in 2018. Then he joined the Autonomous Driving R&D Department of JD.COM as an algorithm researcher from 2018 to 2020.

He is now a PhD student in School of Vehicle and Mobility at Tsinghua University since 2020. His research interests are computer vision, autonomous driving and environmental perception.



Jun Li was born in Jilin Province, China in 1958. He received a Ph.D. degree in internal-combustion engineering at Jilin University of Technology, in 1989.

He has joined the China FAW Group Corporation in 1989 and currently works as a professor with the School of Vehicle and Mobility at Tsinghua University. Now he also serves as the chairman of the China Society of Automotive Engineers (SAE).

In these years, Dr. Li has presided over the product development and technological innovation of large-scale automobile companies in China. Dr. Li has many scientific research achievements in the fields of automotive powertrain, new energy vehicles, and intelligent connected vehicles. Dr. Li is the author of more than 98 papers. In 2013, Dr. Li was awarded an academician of Chinese Academy of Engineering (CAE) for contributions to vehicle engineering.

Caiyan Jia was born in 1976. She received her Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, PR China, in July 2004. She had been a postdoctoral fellow in the Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, PR China, in 2004–2007. She is now a professor in the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, PR China. Her current research interests include deep learning in computer vision, graph neural networks and social computing.

Lijun Zhao was born in Harbin, Heilongjiang Province, and he received a Ph.D degree from Robotics Institute of Technology, at Harbin Institute of Technology, China, in 2009. He currently works as a professor with Robotics Institute, State Key Laboratory of Robotics and System at Harbin Institute of Technology. Dr. Zhao is the author of more than 70 SCI/EI articles. His research interests include SLAM, environments perception and navigation of mobile robots.