

SupFusion: Supervised LiDAR-Camera Fusion for 3D Object Detection

Yiran Qin^{1*}, Chaoqun Wang^{1*}, Zijian Kang², Ningning Ma², Zhen Li³, Ruimao Zhang^{1†}

¹School of Data Science, Shenzhen Research Institute of Big Data,

The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China ²NIO

³School of Science and Engineering, Future Intelligent Network Research Institute,

The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China

{yiranqin@link, chaoqunwang@link, lizhen@, ruimaozhang@}{cuhk.edu.hk, {zijiankang, ningningma}@nio.com}

Abstract

LiDAR-Camera fusion-based 3D detection is a critical task for automatic driving. In recent years, many LiDAR-Camera fusion approaches sprung up and gained promising performances compared with single-modal detectors, but always lack carefully designed and effective supervision for the fusion process. In this paper, we propose a novel training strategy called SupFusion, which provides an auxiliary feature level supervision for effective LiDAR-Camera fusion and significantly boosts detection performance. Our strategy involves a data enhancement method named Polar Sampling, which densifies sparse objects and trains an assistant model to generate high-quality features as the supervision. These features are then used to train the LiDAR-Camera fusion model, where the fusion feature is optimized to simulate the generated high-quality features. Furthermore, we propose a simple yet effective deep fusion module, which contiguously gains superior performance compared with previous fusion methods with SupFusion strategy. In such a manner, our proposal shares the following advantages. Firstly, SupFusion introduces auxiliary feature-level supervision which could boost LiDAR-Camera detection performance without introducing extra inference costs. Secondly, the proposed deep fusion could continuously improve the detector's abilities. Our proposed SupFusion and deep fusion module is plug-and-play, we make extensive experiments to demonstrate its effectiveness. Specifically, we gain around 2% 3D mAP improvements on KITTI benchmark based on multiple LiDAR-Camera 3D detectors. Our code is available at <https://github.com/IranQin/SupFusion>.

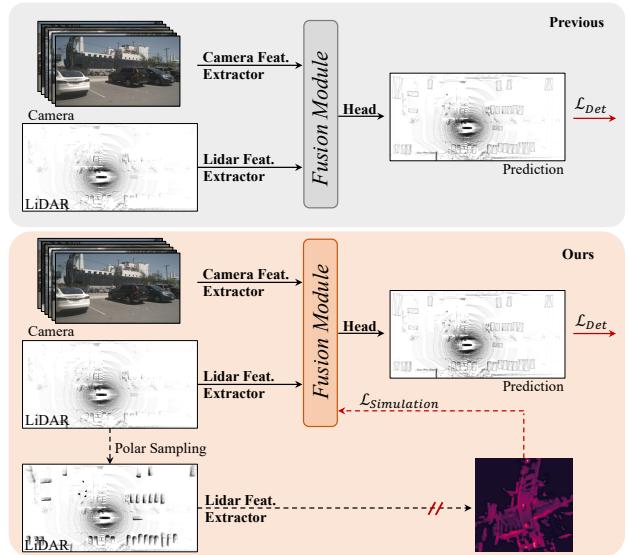


Figure 1: Top: The previous LiDAR-Camera 3D detector, the fusion module is optimized by detection loss. Bottom: Our proposed SupFusion, we introduce auxiliary supervision via high-quality features provided by an assistant model.

1. Introduction

LiDAR-Camera fusion-based 3D object detection is a critical and challenging task for autonomous driving and robotics [32, 35]. Previous approaches [20, 22, 34, 19] always project the camera inputs to LiDAR BEV or voxel space via intrinsic and extrinsic to align the LiDAR and camera features. Then, simple concatenation or summation is employed to obtain the fusion features for final detection. Also, some deep learning-based fusion approaches [1, 21] sprung up and gained promising performances. However, the previous fusion approaches always directly optimized the 3D/2D feature extraction as well as the fusion module by the

*Equal contribution. Work done during an internship at NIO.

†Corresponding author.

detection loss, which lacks carefully designed and effective supervision at the feature level and limits its performance.

In recent years, distillation manners have shown great improvements in feature-level supervision for 3D detection. For instance, some approaches [5, 6] provide LiDAR features to guide the 2D backbone in estimating the depth information from camera inputs. Also, some approaches [19] provide LiDAR-Camera fusion features to supervise LiDAR backbone learns global and contextual presentation from LiDAR inputs. Introducing feature-level auxiliary supervision via simulating more robust and high-quality features, the detectors could boost marginal improvements. Motivated by this, a natural solution to handle LiDAR-Camera feature fusion is *providing stronger and high-quality features and introducing auxiliary supervision for LiDAR-Camera 3D detection*.

To this end, in this paper we propose a supervised LiDAR-Camera fusion method named SupFusion, to generate high-quality features and provide effective supervision for the fusion as well as feature extraction process, and further boost the LiDAR-Camera fusion-based 3D detection performances. Specifically, we **first train an assistant model to provide high-quality features**. To achieve it, different from previous approaches, which leverage larger models or extra data, we proposed a novel data enhancement method named Polar Sampling. The polar sampling could dynamically enhance the object’s density from sparse LiDAR data, which is easier to detect and improve features quality, *e.g.*, the feature could conclude accuracy detection results. Then, we **simply train the LiDAR-Camera fusion-based detector with introduced auxiliary feature-level supervision**. In this step, we feed the raw LiDAR and camera inputs into 3D/2D backbones and fusion modules to obtain fusion features. In one aspect, the fusion features are fed into the detection head for final prediction, which is decision-level supervision. In another aspect, the auxiliary supervision mimics the fusion feature to high-quality features, which are obtained via the pretrained assistant model and enhanced LiDAR data. In such a manner, the proposed feature-level supervision could lead the fusion module to generate more robust features, and further boost the detection performance. To better fuse the LiDAR and camera features, we propose a simple yet effective deep fusion module, which consists of stacked MLP blocks and dynamic fusion blocks. SupFusion could sufficiently excavate the deep fusion module capacity, and continuously improve detection accuracy.

Compared with previous LiDAR-Camera 3D detectors, this is a brand new attempt to improve the effectiveness of the LiDAR-Camera fusion by introducing feature-level supervision, which consistently boosts 3D detection accuracy. The main contributions are summarized as follows.

- We propose a novel supervised fusion training strategy named SupFusion, which mainly consists of a high-quality feature generation process and **firstly propose**

auxiliary feature-level supervision loss for robust fusion feature extraction and accuracy 3D detection to our knowledge.

- To obtain high-quality features in SupFusion, we propose a data enhancement method named Polar Sampling to densify the sparse objects. Furthermore, we propose an efficient deep fusion module to contiguously boost detection accuracy.
- We make extensive experiments based on multiple detectors with different fusion strategies, and gain around 2% mAP improvements on KITTI benchmark. The source code will be released after the blind review.

2. Related Work

2.1. 3D Object Detection

Camera-based 3D Object Detection methods perform 3D detection on single- or multi-view images. Many methods are proposed to address monocular 3D detection[24, 23, 17, 51, 53, 27, 37, 36] since KITTI benchmark[9] only contains one from the camera. With the development of autonomous driving datasets which have more cameras covering a wider range of angles(nuScenes[2] and Waymo[33]), more and more methods[38, 40, 42] are developed that take multi-view images as input and get surprising performance compared to monocular methods.

LiDAR-based 3D Object Detection methods perform 3D detection on a point cloud scene. Depending on the three kinds of point cloud data pretreatments, 3D detectors are also classified into three mainstream. Point-based method [31, 46, 50, 4] use row point cloud data, leveraging pointnet[25] and pointnet++[26] as 3D backbone networks to get point-wised features which are sent to point-wise 3D detection head. Pillar-based method [18, 41, 29] pretreat point cloud to pillars, gain speed up while still maintaining a good performance. Voxel-based methods[54, 45, 8] are widely used, point clouds are voxelized to voxel and sent to a 3D sparse convolutional[10] based 3D backbone to get bird’s-eye-view features. Some methods[30] also try to gain more improvement by using both the voxel feature and the point feature.

2.2. LiDAR-Camera 3D Object Detection

LiDAR-camera fusion is a technique that combines data from LiDAR and camera sensors to obtain a more complete understanding of a scene in 3D. Due to the complementary nature of the features produced by these two modalities, researchers have developed methods that can be jointly optimized on both sensors. These methods can be classified into two categories based on their fusion mechanism. The first category is point-level fusion[14, 35, 47], where the image features are queried via the raw LiDAR points and then

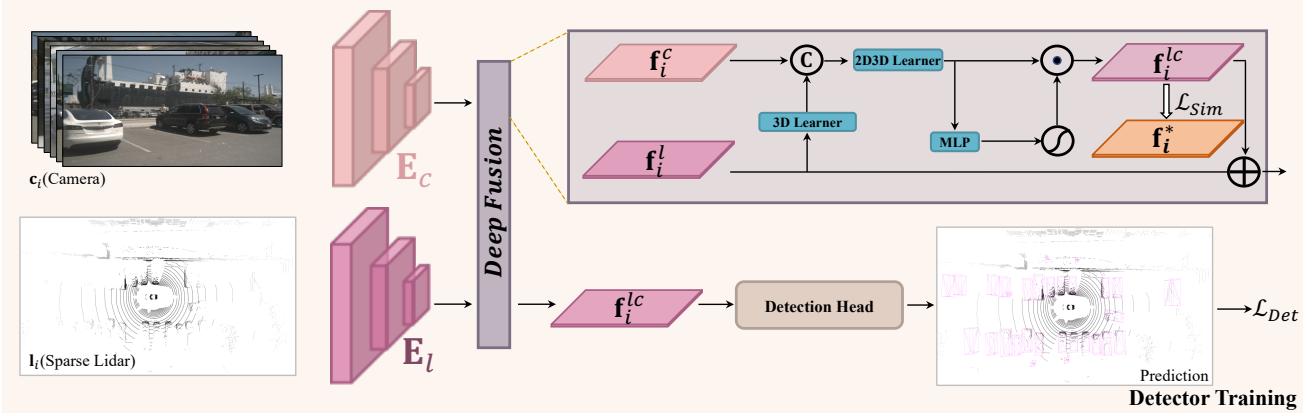


Figure 2: The detector training procedure. For a given LiDAR-Camera sample, we extract the feature presentation via the corresponding 3D/2D encoders, and the proposed deep fusion module is employed to fuse the multi-modal features. Besides the detection loss \mathcal{L}_{Det} , we introduce the auxiliary feature-level supervision \mathcal{L}_{Sim} , to mimic the fusion feature \mathbf{f}_i^{lc} to the high-quality features \mathbf{f}_i^* , which is generated by the assistant model and enhanced data. Here the \odot , \oplus , and $\odot\odot$ indicate the element-wise multiplication, summation, and channel-wise concatenation. More details could be referred in Sec. 3.1

concatenated back as additional point features. The second category is feature-level fusion, where the LiDAR points are projected into a specific feature space[48, 20, 19] or proposals level feature fusion[3, 16, 1, 43]. There are also methods that combine these two fusion manners, such as MVX-Net [32]. Recently, attention-based fusion methods have sprung up. TransFusion[1] uses the bounding box prediction of LiDAR features as a proposal to query the image feature, and then a Transformer-like architecture fuses the information back to LiDAR features. Similarly, DeepFusion[21] projects LiDAR features on each view image as queries and leverages cross-attention for the two modalities. Bird’s-eye-view space feature fusion[22, 34] is proposed recently and gained great performance with different tasks.

2.3. Knowledge Distillation for 3D Object Detection

Object detection tasks differ from classification tasks in that they require consideration of both global and local features. While numerous distillation methods have been proposed to address this challenge, such as those developed by[39, 7, 11, 49], many of these methods require the creation of specific distillation regions or objects based on ground-truth data, bounding box predictions, attention, or gradients to mitigate the imbalance between foreground and background. In addition to 2D detection, there has been growing interest in cross-model knowledge distillation methods between RGB and LiDAR in 3D object detection. For example, [12, 28, 6, 5] have developed methods that transfer knowledge from LiDAR detectors to image detectors. [52, 15] has trained a teacher model similar to that proposed by[35] and developed a multilevel distillation framework to transfer semantic segmentation information to a student model. In this

work, we want to further boost LiDAR-Camera detectors with an auxiliary feature level supervision.

3. Methodology

In this section, we first overview the overall SupFusion pipeline in Sec. 3.1, and then detail introduces the main components, including the proposed Polar Sampling in Sec. 3.2, and deep fusion module in Sec. 3.3.

3.1. Overall SupFusion Pipeline

For a given LiDAR-Camera fusion-based 3D detector D , which consists of a 3D Encoder E_l , a 2D Encoder E_c , a fusion module F , and a detection head. Denotes $\mathcal{D} = \{(\mathbf{l}_1, \mathbf{c}_1), (\mathbf{l}_2, \mathbf{c}_2), \dots, (\mathbf{l}_N, \mathbf{c}_N)\}$ and $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}$ are the given dataset with N samples, where $(\mathbf{l}_i, \mathbf{c}_i)$ are the paired LiDAR and camera inputs. The previous approaches feed the paired $(\mathbf{l}_i, \mathbf{c}_i)$ into 3D/2D encoders E_l and E_c to capture corresponding feature presentation \mathbf{f}_i^l and \mathbf{f}_i^c . Then, the paired features presentation is fed into the fusion module to obtain the fusion feature \mathbf{f}_i^{lc} , which is used for the final prediction. To better leverage the LiDAR and camera data, we propose a supervised LiDAR-Camera Fusion training strategy named SupFusion, which provides high-quality features for feature-level supervision. To end this, different from previous approaches, which adopt larger scale models or extra data input, we propose a data enhancement method named polar sampling to densify the sparse objects in LiDAR data. In such a manner, the data is easier and could easily generate high-quality features (*e.g.*, the feature could conclude accuracy detection results) with lightweight models. The proposed polar sampling mainly

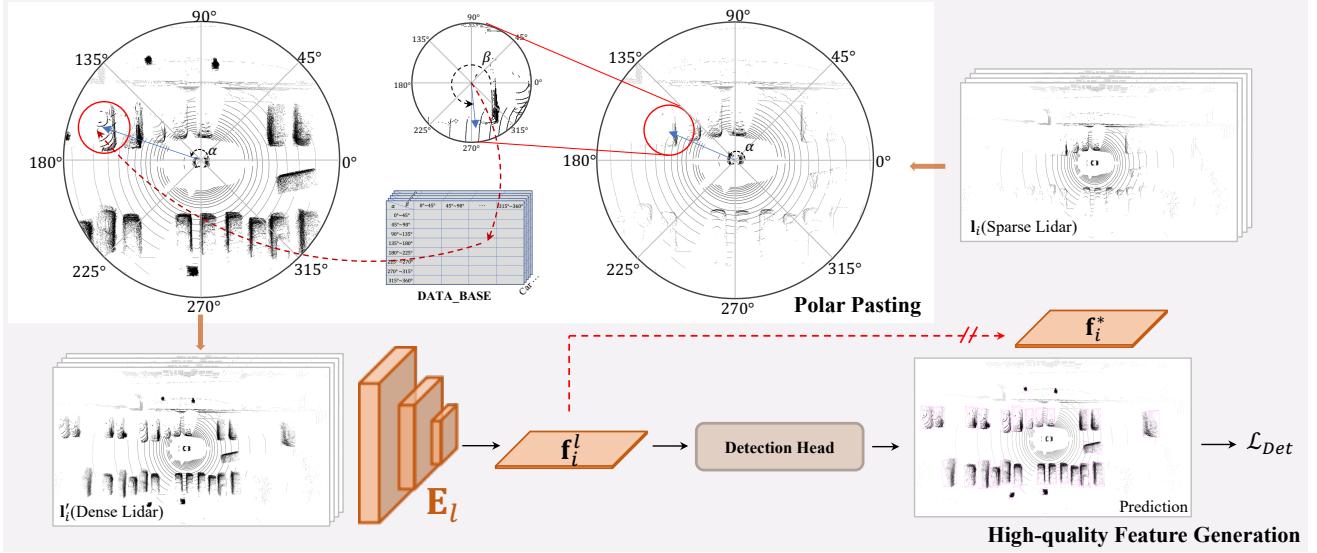


Figure 3: High-quality feature generation process. For any given LiDAR sample, we densify the sparse objects via polar pasting, which calculate the direction and rotation to query dense object from the database, and pasting to add extra points to sparse objects. We first train the assistant model via the enhanced data and feed the enhanced LiDAR data into the assistant model to generate high-quality features \mathbf{f}^* after it is convergent. More details could be referred in Sec. 3.1.

densifies the sparse object points in LiDAR data. Specifically, for any sample $(\mathbf{l}_i, \mathbf{c}_i)$, where $\mathbf{l}_i = \{l_i^1, l_i^2, \dots, l_i^{n_i}, l_i^s\}$ contains n_i objects points sets and one background points set. In the polar sampling, we add extra points to each object point set from raw LiDAR data, to obtain enhanced LiDAR data $\mathbf{l}'_i = \{(l_i^1, \hat{l}_i^1), (l_i^2, \hat{l}_i^2), \dots, (l_i^{n_i}, \hat{l}_i^{n_i}), l_i^s\}$ in polar pasting, where \hat{l}_i^k is the added points set for the k -th objects in i -th LiDAR sample and generated in polar grouping. We will detail introduce the polar sampling in Sec. 3.2.

High-quality Feature Generation. To provide feature-level supervision in our SupFusion, we adopt an assistant model to capture high-quality features from the enhanced data as shown in Fig. 3. Firstly, we train an assistant model to provide high-quality features. Specifically, for any sample in \mathcal{D} , we enhance the sparse LiDAR data \mathbf{l}_i to obtain enhanced data \mathbf{l}'_i via the polar pasting, which densifies the spare objects with added points set generated in polar grouping. Then, after the assistant model convergence (often with few epochs), we feed the enhanced sample \mathbf{l}'_i into the optimized assistant model, to capture high-quality features \mathbf{f}_i^* , which is used to train the LiDAR-Camera 3D detector. To better apply to the given LiDAR-Camera detector and easier implement, we simply adopt the LiDAR branch detector as the assistant model.

Detector Training. For any given LiDAR-Camera detector, we train the model with the proposed auxiliary supervision at the feature level. Specifically, given sample $(\mathbf{l}_i, \mathbf{c}_i) \in \mathcal{D}$, we first feed the LiDAR and camera into the 3D

and 2D Encoders \mathbf{E}_l and \mathbf{E}_c to capture the corresponding features \mathbf{f}_i^l and \mathbf{f}_i^c , which are fed into the fusion model to generate fusion features \mathbf{f}_i^{lc} , and flow into the detection head for final prediction. Furthermore, the proposed auxiliary supervision is employed to mimic the fusion feature with high-quality features, which is generated by the pre-trained assistant model and enhanced LiDAR data. The above process can be formulated as:

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{(\mathbf{l}_i, \mathbf{c}_i), \mathbf{y}_i \sim \mathcal{D}} \mathcal{L}_{det}(\mathbf{D}(\mathbf{l}_i, \mathbf{c}_i, \theta), \mathbf{y}_i) \quad (1)$$

$$+ \lambda \mathcal{L}_{Sim}(\mathbf{f}_i^{lc}, \mathbf{f}_i^*), \quad (2)$$

$$s.t. \quad \mathbf{f}_i^{lc} = \mathbf{F}(\mathbf{E}_l(\mathbf{l}_i, \theta_l), \mathbf{E}_c(\mathbf{c}_i, \theta_c), \theta_f), \quad (3)$$

$$\mathbf{f}_i^* = \mathbf{E}_l(\mathbf{l}'_i, \theta_l^*), \quad (4)$$

$$\mathcal{L}_{Sim} = \|\psi(\mathbf{f}_i^{lc}) - \mathbf{f}_i^*\|_2^2 \quad (5)$$

where θ^* is the optimized parameters of detector \mathbf{D} , \mathbf{f}_i^* is the high-quality features, which is generated from pre-trained assistant model LiDAR encoder \mathbf{E}_l with parameters θ_l^* , ψ is a simple convolutional layer with kernel size 1. λ is the balanced weight and we set it as 1.0.

3.2. Polar Sampling

To provide high-quality features via the assistant model in our proposed SupFusion, we propose a novel data enhancement method named Polar Sampling, to ease the sparse problem which frequently leads to the detection failers. To end this, we densify the sparse objects in LiDAR data with

a similar dense object. Specifically, the polar sampling consists of two parts, including polar grouping and polar pasting. In polar grouping, we mainly build a database to store the dense objects, which is used in polar pasting to densify the sparse objects.

Considering the characteristics of LiDAR sensors, the collected point cloud data naturally exists a particular density distribution, *e.g.*, the objects have more points on the surface towards the LiDAR sensor, while few points on the opposite sides. The density distribution is mainly influenced by the direction and rotation, while the points' density mainly depends on the distances, *e.g.* the objects closer to the LiDAR sensor have denser points. Motivated by this, we aim to densify the sparse objects in longer distances with the dense objects in short distances, according to their direction and rotation to keep the density distributions. Specifically, we build polar coordinate systems for the whole scene as well as the objects based on the center of the scene and specific objects and define the positive direction of the LiDAR sensor as 0 degrees to measure the corresponding directions and rotations. Then, we collect the objects with similar density distribution (*e.g.* with similar direction and rotation) and generate one dense object for each group in the polar grouping, and used it to densify the sparse objects in polar pasting.

Polar Grouping. As shown in Fig. 4, we build a database \mathcal{B} to store the generated dense object point set \hat{l} according to their directions and rotations in the polar grouping, which is noted as α and β in Fig. 4.

Firstly, we search the whole dataset and calculated their polar angles for all objects via the locations and provided rotations in the benchmarks. Secondly, we split the objects into groups according to their polar angles. Specifically, we manually divide the direction and rotation into N groups. For any object point set l , we could put it into the corresponding group according to the index:

$$\alpha'_l = \lfloor \frac{\alpha_l \cdot 2\pi}{N} \rfloor, \quad \beta'_l = \lfloor \frac{\beta_l \cdot 2\pi}{N} \rfloor, \quad (6)$$

where $\alpha'_l, \beta'_l \in \{1, 2, \dots, N\}$ are the direction and rotation index of object l , and we could put the points set into \mathcal{B} as: $\mathcal{B}_C^{\alpha'_l, \beta'_l} \leftarrow l$, where C is the label of object l . Finally, for each object group $\mathcal{B}_C^{\alpha', \beta'} = \{l_1, l_2, \dots, l_n\}$ contains n similar objects, we select the densest k objects and mix up to generate one dense and complete object \hat{l} , and $\mathcal{B}_C^{\alpha', \beta'} = \{\hat{l}\}$ is used in polar pasting phase. To keep the density and avoid too much calculation in the model, we randomly select partial points in the dense object. Here we set $k = 10$ and select 5000 points for all objects.

Polar Pasting. As shown in Fig. 2, the polar pasting is utilized to enhance the sparse LiDAR data to train the assistant model and generate high-quality features. Concretely, given a LiDAR sample $\mathbf{l}_i = \{l_i^1, l_i^2, \dots, l_i^{n_i}, l_i^s\}$ contains n_i objects, for any object l_i^k , we can calculate the direction

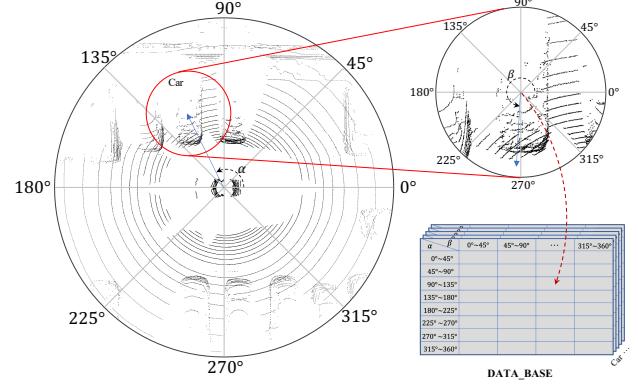


Figure 4: Polar grouping mainly builds a database to store the dense objects of each group in polar coordinate systems of direction and rotations for each class. More details could be referred to in Sec. 3.2.

and rotation the same as the grouping process and query the dense object \hat{l}_i^k from \mathcal{B} according to the label and indexes, which could be obtained from E.q. 6 In such a manner, we enhance all the objects in the sample and obtain the enhanced data: $\mathbf{l}'_i = \{(l_i^1, \hat{l}_i^1), (l_i^2, \hat{l}_i^2), \dots, (l_i^{n_i}, \hat{l}_i^{n_i}), l_i^s\}$

3.3. Deep Fusion

To mimic the high-quality features generated from enhanced LiDAR data, the fusion model aims to extract the missing information of the sparse objects from the rich color and contextual features in camera input. To end this, we propose the deep fusion module, to leverage image features and complete the LiDAR presentation. The proposed deep fusion mainly consists of a 3D learner and a 2D3D learner. The 3D learner is a simple convolutional layer, to transfer the 3D presentation into 2D space. Then, concatenating the 2D features and 3D presentation (*e.g.* in 2D space), the 2D3D learner is employed to fuse the LiDAR-Camera features. Finally, we weighed the fusion feature via an MLP and an activate function, adding back to the raw LiDAR features as the output of the deep fusion module. The 2D3D learner consists of stacked MLP blocks with depth K , learning to leverage camera features to complete the LiDAR representation from the sparse objects, to simulate the high-quality features of dense LiDAR objects.

4. Experiments

4.1. Experiments Settings

Datasets and Evaluation Metrics. KITTI [9] is a large-scale public benchmark for autonomous driving, including 3712 training, 3769 validation, and 7518 testing samples with LiDAR and camera pair data. All objects in KITTI are divided into 'easy', 'moderate', and 'hard' difficulties ac-

Table 1: The experiment results (mAP@R40 %) on KITTI validation set. We list the easy, moderate (mod.), and hard cases for three classes, and the overall performances. Here, L, LC, LC* indicate the results of the corresponding LiDAR detector, LiDAR-Camera fusion detectors, and our proposal. Δ indicates the improvements. The best results are shown in **bold** except L † , which is the assistant model, and test on the enhanced validation set. MVXNet is re-conduct based on mmdetection3d. PV-RCNN-LC and Voxel-RCNN-LC are re-conduct based on the open-source code of VFF [20].

Detector	Modality	Overall			Pedestrian			Cyclist			Car		
		easy	mod.	hard									
MVXNet	L †	88.44	90.93	88.58	87.33	87.61	80.15	80.09	86.01	86.19	97.89	99.17	99.41
	L	73.46	58.64	55.15	57.54	50.10	45.56	74.63	51.81	48.92	88.21	73.99	70.98
	LC	74.99	63.00	59.32	64.46	57.76	53.05	72.69	55.27	51.65	87.82	75.98	73.26
	LC*	75.37	64.73	60.52	63.11	58.13	53.73	74.79	57.60	53.82	88.20	78.46	74.01
	Δ	+0.38	+1.73	+1.20	-1.35	+0.37	+0.68	+2.10	+2.33	+2.17	+0.38	+2.48	+0.75
PV-RCNN-LC	L †	94.90	95.20	95.10	86.86	86.78	87.07	98.90	99.33	98.62	98.95	99.50	99.62
	L	81.37	71.55	68.08	67.80	59.56	54.40	85.06	70.83	67.96	91.26	84.25	81.88
	LC	84.16	74.54	70.86	70.45	64.16	59.50	90.47	74.46	70.25	91.56	85.01	82.83
	LC*	85.77	76.00	72.00	74.29	66.67	61.38	91.00	76.13	71.66	92.02	85.20	82.94
	Δ	+1.61	+1.46	+1.14	+3.84	+2.51	+1.88	+0.53	+1.67	+1.41	+0.46	+0.19	+0.11
Voxel-RCNN-LC	L †	99.26	99.66	99.75	97.84	98.05	98.17	94.36	94.60	94.85	99.90	99.90	99.91
	L	82.01	71.69	68.40	67.84	60.95	56.03	86.10	71.39	67.05	92.09	82.74	82.12
	LC	84.36	74.46	70.80	71.39	64.89	60.20	89.64	72.94	69.03	92.05	85.55	83.17
	LC*	86.13	75.90	72.17	73.95	66.06	61.83	92.02	76.18	71.83	92.42	85.47	82.83
	Δ	+1.77	+1.44	+1.37	+2.56	+1.17	+1.63	+2.38	+3.24	+2.80	+0.37	-0.08	-0.34

cording to the object scales and occlusion ratio. We evaluate the detection performance by the mean Average Precision under 40 recall thresholds (mAP@R40) the same as the official benchmark and evaluation for 3 classes including Pedestrian, Cyclist, and Car. **nuScenes** [2] is a large-scale autonomous driving dataset for 3D detection and tracking tasks, including 700 training, 150 validation, and 150 testing sequences, with each 20s long and annotated with 3D bounding boxes. For 3D object detection, the official evaluation metrics include the mean Average Precision (mAP) and nuScenes detection score (NDS). mAP measures the localization precision using a threshold base on the birds-eye view. NDS is a weighted combination of mAP and regression accuracy of other object attributes, including box size, orientation, and translation.

Detectors. To demonstrate the effectiveness of our proposed method, we apply the SupFusion training strategy and deep fusion module on variance LiDAR-Camera detectors including MVXNet [32], PV-RCNN-LC [30], and Voxel-RCNN-LC [8] on KITTI, and SECOND-LC [45] on nuScenes benchmark. MVXNet is a voxel-based LiDAR-Camera 3D detector based on VoxelNet [54]. The camera features project to LiDAR space and concatenates with points features. Voxel-RCNN and PV-RCNN are LiDAR-

based detectors, we re-conduct the LiDAR-Camera fusion approach following VFF [20] based on the open-source code, which is a previous state-of-the-art fusion approach based on the two detectors, mainly projects the camera features into voxel presentation and summarized fusion with the voxel features of LiDAR input. SECOND-LC introduces camera inputs based on the LiDAR-based SECOND detector, the same as BEVFusion [22], we concatenate the camera and LiDAR features in birds-eys-view spaces.

Implement Details. We conduct experiments on MVXNet, PV-RCNN-LC and Voxel-RCNN-LC, SECOND-LC following the open-sourced code and mmdetection3d on 8 Tesla A100 GPUs. All the architecture, configuration, and optimizer follow the settings reported in papers including voxel size, point cloud range, batch size, and learning rate. We modify the fusion method (*e.g.* concatenation, summarized) by our proposed deep fusion module and applied our SupFusion training strategy. For the deep fusion module, we set depth $K = 3$. For the auxiliary supervision, we utilize simple L2 loss and set $\lambda = 1.0$. More details about how to apply SupFusion and deep fusion modules on different detectors are demonstrated in the Appendix.

Table 2: NDS and mAP(%) results on nuScenes benchmarks.

Modality	3D	2D	NDS	MAP
L [†]	SECOND	-	90.38	94.44
L	SECOND	-	60.52	52.28
LC	SECOND	Swin-T	62.57	55.26
LC*	SECOND	Swin-T	64.58 _(+2.01)	56.64 _(+1.38)

Table 3: Results comparison with different fusion methods. Sup. indicates our proposed feature-level supervision. Ped., Cyc., and O.a. are short of Pedestrian, Cyclist, and Overall.

Method	Sup.	Ped.	Cyc.	Car	O.a.	Δ
Sum	✗	64.89	72.94	85.55	74.46	-
	✓	67.03	74.61	85.11	75.59	+1.13
Concat	✗	65.90	72.80	83.37	74.02	-
	✓	66.23	74.95	85.42	75.52	+1.50
Ours	✗	63.18	72.84	85.58	73.86	-
	✓	66.06	76.18	85.47	75.90	+2.04

4.2. Main Results

Overall Performances. The 3D mAP@R40 comparison based on three detectors is shown in Tab. 1 for three classes and overall performances for each difficulty split. We can clearly observe that the LiDAR-Camera approaches (LC) are superior to the LiDAR-based detectors (L) by introducing extra camera inputs. By introducing polar sampling, the assistant model (L^{\dagger}) could gain admirable performances (e.g. over 90% mAP) on the enhanced validation set. With the auxiliary supervision from high-quality features and our proposed deep fusion module, our proposal could contiguously boost detection accuracy. For instance, compared with the baseline (LC) model, our proposal could gain +1.54% and +1.24% 3D mAP improvements for the moderate and hard objects on average for the three detectors. Further, we also conduct experiments on nuScenes benchmark based on SECOND-LC as shown in Tab. 2, which gain +2.01% and +1.38% improvements for NDS and mAP. Also, we will prune the SupFusion training strategy and deep fusion module in Sec. 4.3 to detail discuss their effectiveness.

Class-Aware Improvements Analysis. Compared with the baseline model, our SupFusion and deep fusion could boost the detection performances not only for the overall performances but also for each class including Pedestrian, Cyclist, and Car. Comparing the average improvements (e.g. moderate cases) for three classes, we can obtain the following observation: the cyclist gains the largest improvements (+2.41%) while pedestrian and car gains +1.35% and +0.86% increases respectively. The reason is obvious: (1) The car is easier to detect and gains the best results compared with pedestrians and cyclists, hence it is more difficult to improve. (2) Cyclist gains more improvements against

Table 4: Results comparison with different supervision loss. \mathcal{L}_{Sim} indicates the auxiliary supervision loss in Eq. 1.

\mathcal{L}_{Sim}	Ped.	Cyc.	Car	O.a.	Δ
None	64.89	72.94	85.55	74.46	-
L1	65.76	75.57	85.42	75.58	+1.12
L2	66.06	76.18	85.47	75.90	+1.44

Table 5: Results with different depth K in deep fusion.

K	1	2	3	4
LC	74.23	74.00	73.86	73.22
LC*	75.64	75.75	75.90	75.88

Table 6: Results with different N in polar sampling.

K	1	4	8	16
L [†]	99.80	99.65	99.66	99.43
LC*	75.68	75.72	75.90	75.85

Pedestrian because pedestrians are non-grid, the generated dense object is not so well compared with Cyclist, hence gains lower performance improvements.

4.3. Ablation Study

To analyze the effectiveness of our proposed SupFusion and deep fusion module, we conduct ablation experiments based on Voxel-RCNN-LC on the KITTI benchmark.

Supervision. In SupFusion training strategy, we mainly introduce a high-quality guided feature-level supervision and simply employ L2 loss as the auxiliary loss. Here we also adopt L1 loss and the results are shown in Tab. 4, from which we can observe that L1 loss gains comparable but slight improvements (e.g. +1.12% versus +1.44%). To mimic the high-quality features by the fusion features, utilizing L1 and L2 to measure the feature distance is acceptable [13, 44].

Fusion Strategy. We conduct experiments with different fusion methods, including the concatenation and summation approaches employed in MVXNet and PV-RCNN-LC/Voxel-RCNN-LC, as well as our proposed deep fusion methods. As shown in Tab. 3, without auxiliary supervision in SupFusion, our proposed deep fusion and concatenation fusion gain comparable but lower performances, demonstrating introducing extra parameters in the fusion model would lead the model difficult to train and hard to optimize. By introducing auxiliary supervision, our proposal could gain the best performance and improvements compared with the baseline model. This demonstrates that the SupFusion could effectively promote the learning process, guiding the 3D/2D backbone and fusion model extracts meaningful robust fusion features from LiDAR and camera inputs. Furthermore, we also manually set variance depth in deep fusion, to rescale the fusion module size. The results are shown in Tab. 5, in which we can observe that by increasing the fusion depth,

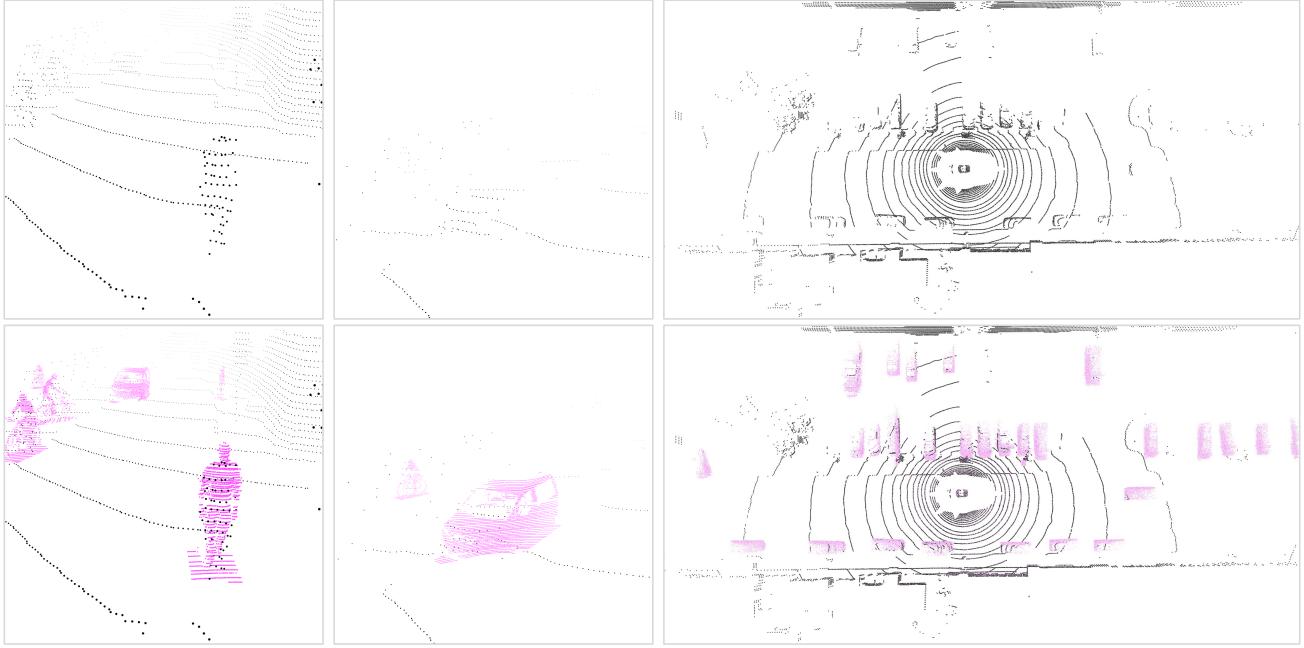


Figure 5: Visualization results. The top row shows the raw sparse LiDAR data and the bottom row displays the enhanced data. Particularly, the points in magenta are the added points and row points are shown in black. We can densify the sparse objects with few points via the dense objects in database while keeping the density distributions.

the baseline gets worse while our proposal gains better performances, demonstrating the effectiveness of our proposed SupFusion training strategy and deep fusion module again.

Polar Sampling. In the proposed polar sampling data enhancement, we split the objects into $N \times N$ groups according to their directions and rotations. To explore the influence of the split number N , we manually set $N = \{1, 4, 8, 16\}$ to conduct experiments, and the results are shown as Tab. 6, from which we can observe that the split number N in polar sampling affect the performances of assistant and our model in an extent. For instance, when setting $N = 1$, we only generate one object for each class, the object is complete but destroys the density distribution, which boosts the assistant model performances but decreases the improvements in an extent (e.g. +1.22% versus +1.44%), when setting N larger, the objects in each object set may get fewer, because we split all the objects in the dataset into more groups, which may influence the quality of the dense object in the database in the polar grouping.

Assistant Model. We utilize the enhanced dense LiDAR data to generate high-quality features. By densifying the sparse objects, the assistant model with enhanced data could avoid detection failers easily and provide high-quality features. To better apply to the given LiDAR-Camera detector and easier implement, we simply adopt the LiDAR branch detector as the assistant model. Here, we also conduct experiments to keep the assistant model sharing the same archi-

Table 7: Results with different teacher models in SupFusion. LC* and LC** indicate utilize LiDAR and LiDAR-Camera based assistant model.

Model	Ped.	Cyc.	Car	O.a.	Δ
LC	64.89	72.94	85.55	74.46	-
LC*	66.06	76.18	85.47	75.90	+1.44
LC**	66.39	75.71	85.47	75.86	+1.40

tecture with the given detector, which is a LiDAR-Camera detector. The results are shown as Tab. 7, from which we can observe that utilizing LiDAR-base and LiDAR-Camera based detectors as the assistant model, the SupFusion gains comparable performances (e.g. LC* versus LC**).

4.4. Qualitative Results

Here we visualize the raw sparse LiDAR data and enhanced dense LiDAR in Fig. 5. In the left 2 columns, we can observe the added points in magenta could effectively densify the sparse points in black points for all classes including pedestrians, cyclists, and cars. In the right column, we can find that the polar sampling densifies the sparse objects with dense ones while keeping the objects' density distribution. As we could see, enhanced point cloud data have high integrity and obvious class representation which could be utilized to generate high-quality features as the fusion supervision. To better demonstrate the quality of feature-level

supervision, we visualize the feature of assistant, baseline, and our model in Appendix.

4.5. Information leak

There is no information leak if we are not misunderstanding. In practice, we build the database in Fig. 4 only using the training set in the polar grouping phase, the validation set is not used in polar sampling, as well as the model training process.

5. Conclusion

In this paper, we propose a novel training strategy named SupFusion, to introduce auxiliary feature-level supervision for effective LiDAR-Camera fusion, and further boost 3D detection performances. Specifically, we introduce high-quality features to the LiDAR-Camera fusion-based 3D detector, which is used to provide auxiliary supervision for the 3D/2D feature extraction and multi-modal feature fusion. To achieve this, we propose a novel data enhancement method named polar sampling to densify the sparse objects, which is utilized to generate high-quality features via an assistant model. Furthermore, we proposed a simple yet effective deep fusion module, which leverages meaningful information from camera input and contiguously boosts the LiDAR-Camera fusion-based 3D detectors. This is the first approach to introducing feature-level supervision for LiDAR-Camera 3D detection and we conduct extensive experiments to demonstrate its effectiveness.

6. Acknowledge

The work is partially supported by the Young Scientists Fund of the National Natural Science Foundation of China under grant No. 62106154, by the Natural Science Foundation of Guangdong Province, China (General Program) under grant No.2022A1515011524, and by Shenzhen Science and Technology Program JCYJ20220818103001002 and by Shenzhen Science and Technology Program ZDSYS20211021111415025.

References

- [1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *CVPR*, pages 1090–1099, 2022.
- [2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giacarlo Baldan, and Oscar Beijbom. nuscenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.
- [3] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *CVPR*, pages 1907–1915, 2017.
- [4] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Fast point r-cnn. In *ICCV*, pages 9775–9784, 2019.
- [5] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qin-hong Jiang, and Feng Zhao. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. *arXiv preprint arXiv:2211.09386*, 2022.
- [6] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zihui Wang, and Wanli Ouyang. Monodistill: Learning spatial features for monocular 3d object detection. *arXiv preprint arXiv:2201.10830*, 2022.
- [7] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *CVPR*, pages 7842–7851, 2021.
- [8] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1201–1209, 2021.
- [9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE, 2012.
- [10] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, pages 9224–9232, 2018.
- [11] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *CVPR*, pages 2154–2164, 2021.
- [12] Xiaoyang Guo, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector. In *ICCV*, pages 3153–3163, 2021.
- [13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.
- [14] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. Epnet: Enhancing point features with image semantics for 3d object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 35–52. Springer, 2020.
- [15] Bo Ju, Zhikang Zou, Xiaoqing Ye, Minyue Jiang, Xiao Tan, Errui Ding, and Jingdong Wang. Paint and distill: Boosting 3d object detection with semantic passing network. In *ACMM*, pages 5639–5648, 2022.
- [16] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018.
- [17] Abhinav Kumar, Garrick Brazil, and Xiaoming Liu. Groomed-nms: Grouped mathematically differentiable nms for monocular 3d object detection. In *CVPR*, pages 8973–8983, 2021.
- [18] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, pages 12697–12705, 2019.
- [19] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. In *Advances in Neural Information Processing Systems*, 2022.
- [20] Yanwei Li, Xiaojuan Qi, Yukang Chen, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Voxel field fusion for 3d object detection. In *CVPR*, pages 1120–1129, 2022.
- [21] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *CVPR*, pages 17182–17191, 2022.
- [22] Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.
- [23] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. Autoshape: Real-time shape-aware monocular 3d object detection. In *ICCV*, pages 15641–15650, 2021.
- [24] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. In *ICCV*, pages 3111–3121, 2021.
- [25] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017.
- [26] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.
- [27] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *CVPR*, pages 8555–8564, 2021.
- [28] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *CVPR*, pages 9891–9901, 2022.
- [29] Guangsheng Shi, Ruifeng Li, and Chao Ma. Pillarnet: Real-time and high-performance pillar-based 3d object detection. In *ECCV*, pages 35–52. Springer, 2022.
- [30] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-

- voxel feature set abstraction for 3d object detection. In *CVPR*, pages 10529–10538, 2020.
- [31] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*, pages 770–779, 2019.
- [32] Vishwanath A Sindagi, Yin Zhou, and Oncel Tuzel. Mvx-net: Multimodal voxelnet for 3d object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7276–7282. IEEE, 2019.
- [33] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pages 2446–2454, 2020.
- [34] Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, Tingting Liang, Hongwei Xie, and Zhi Tang. BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework. In *Neural Information Processing Systems (NeurIPS)*, 2022.
- [35] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *CVPR*, pages 11794–11803, 2021.
- [36] Li Wang, Liang Du, Xiaoqing Ye, Yanwei Fu, Guodong Guo, Xiangyang Xue, Jianfeng Feng, and Li Zhang. Depth-conditioned dynamic message propagation for monocular 3d object detection. In *CVPR*, pages 454–463, 2021.
- [37] Li Wang, Li Zhang, Yi Zhu, Zhi Zhang, Tong He, Mu Li, and Xiangyang Xue. Progressive coordinate transforms for monocular 3d object detection. *Advances in Neural Information Processing Systems*, 34:13364–13377, 2021.
- [38] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022.
- [39] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *CVPR*, pages 4933–4942, 2019.
- [40] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *ICCV*, pages 913–922, 2021.
- [41] Yue Wang, Alireza Fathi, Abhijit Kundu, David A Ross, Caroline Pantofaru, Tom Funkhouser, and Justin Solomon. Pillar-based object detection for autonomous driving. In *ECCV*, pages 18–34. Springer, 2020.
- [42] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022.
- [43] Xiaopei Wu, Liang Peng, Honghui Yang, Liang Xie, Chenxi Huang, Chengqi Deng, Haifeng Liu, and Deng 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*, pages 5418–5427, 2022.
- [44] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, pages 9653–9663, 2022.
- [45] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.
- [46] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse-to-dense 3d object detector for point cloud. In *ICCV*, pages 1951–1960, 2019.
- [47] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Multi-modal virtual point 3d detection. *Advances in Neural Information Processing Systems*, 34:16494–16507, 2021.
- [48] Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *ECCV*, pages 720–736. Springer, 2020.
- [49] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *ICLR*, 2021.
- [50] Yifan Zhang, Qingyong Hu, Guoquan Xu, Yanxin Ma, Jianwei Wan, and Yulan Guo. Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In *CVPR*, pages 18953–18962, 2022.
- [51] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *CVPR*, pages 3289–3298, 2021.
- [52] Wu Zheng, Mingxuan Hong, Li Jiang, and Chi-Wing Fu. Boosting 3d object detection by simulating multimodality on point clouds. In *CVPR*, pages 13638–13647, 2022.
- [53] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monocular 3d object detection: An extrinsic parameter free approach. In *CVPR*, pages 7556–7566, 2021.
- [54] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, pages 4490–4499, 2018.