Sparse LiDAR and Binocular Stereo Fusion Network for 3D Object Detection

Weiqing Yan[1], Kaiqi Su[1], Jinlai Ren[1(✉)], Runmin Cong[2], Shuai Li[3], and Shuigen Wang[4]

[1] Yantai University, Yantai 264005, Shandong, China
qfrenjinlai@126.com
[2] Beijing Jiaotong University, Beijing 100044, China
[3] Shandong University, Jinan 250100, Shandong, China
[4] Yantai IRay Technologies Ltd., Co., Yantai 264006, China

Abstract. 3D object detection is an essential task in autonomous driving and virtual reality. Existing approaches largely rely on expensive LiDAR sensors for accurate depth information to have high performance. While much lower-cost stereo cameras have been introduced as a promising alternative, there is still a notable performance gap. In this paper, we explore the idea to leverage sparse LiDAR and stereo images obtained by low-cost sensors for 3D object detection. We propose a novel multi-modal attention fusion end-to-end learning framework for 3D object detection, which effectively integrate the complementarities of sparse LiDAR and stereo images. Instead of directly fusing LiDAR and stereo modalities, we introduce a deep attention feature fusion module, which enables interactions between intermediate layers of LiDAR and stereo image paths by exploring the interdependencies of channel features. These fused features connect higher layer features after upsampling and lower layer features from the stereo image pathway and sparse LiDAR pathway. Hence, the fused features have high-level semantics with higher resolution, which is beneficial for the following object detection network. We provide detailed experiments on KITTI benchmark and achieve state-of-the-art performance compared with the low-cost based methods.

Keywords: 3D object detection · Sparse LiDAR · Stereo images · Low cost

1 Introduction

Autonomous driving is receiving more and more attention from the industry and the research community, the requirements for 3D object detection are also getting more and higher. Besides autonomous driving, 3d object detection has

This work was supported by the National Natural Science Foundation of China under Grants 61801414, 62072391, Natural Science Foundation of Shandong Province under Grants ZR2020QF108.

S. Yu et al. (Eds.): PRCV 2022, LNCS 13536, pp. 41–55, 2022.
https://doi.org/10.1007/978-3-031-18913-5_4

been applied to many other fields, such as virtual reality and medical simulation. It is one of the most important tasks in the field of computer vision.

Different from 2D object detection, 3D object detection can estimate depth and orientation of bounding boxes of objects by input sensor data. Depending on the different type of sensor, 3D object detection can be divided into LiDAR-based methods(point cloud-based methods) [6,9,12,16,20,36,40] , monocular image-based methods [1,18,21,27,32], and binocular stereo image-based methods [5,17,24,30,31,35,39]. Existing LiDAR-based methods provide accurate depth information by 3D point clouds. Although highly precise and reliable, LiDAR sensors are notoriously expensive: a 64-beam model can cost around $75,000 (USD). Compared with LiDAR and binocular stereo cameras, monocular cameras provide the cheapest data for 3D detection. However, a single image lacks reliable depth information, which results in low precision for 3D detection. Compared to monocular cameras, binocular stereo cameras can provide absolute depth information. And it is not expensive and can provide denser image information for small objects. While much lower-cost stereo cameras have been introduced as a promising alternative, there is still a notable performance gap compared with the results of LiDAR. All these sensors have their own advantage and disadvantage, in which none of them complete well on all practical scenarios. Some works [7,19,25,33,37] have researched how to fuse multiple sensors information so that improve the performance of 3D object detection. However, these methods take LiDAR data with 32 or 64 beams as input, which are very expensive. LiDAR sensors with 4 beams are cheaper compared with 64 beams and thus it is easily affordable. However, it cannot be used to detect small 3D object only by themselves, since 4 beams LiDAR data are very sparse. As aforementioned, stereo images can provide denser information for small objects. Therefore, we consider the fusion of binocular stereo camera and 4 beams LiDAR sensor for 3D object detection, which is a more practical choice. Depending on sparse LiDAR and stereo images, You et al. [41] proposed Pseudo-LiDAR++ method for 3D object detection. In this method, they first generate a dense depth map by stereo images, and then correct depth map by using sparse LiDAR information. However, in process of generating depth map, they need 64-beams LiDAR supervision.

In this paper, we propose a novel multi-modal fusion architecture that make full use of the advantages from both sparse LiDAR and stereo image feature fusion. It is worth noting that our proposed architecture is designed from low-cost sensors. Since 4-beam LiDAR information is extremely sparse, the fusion is from LiDAR feature to image feature, which augments image features with information accuracy of LiDAR features.

Different from the previous fusion methods based on LiDAR and images, we take a sparse 4-beam LiDAR and stereo images to detect 3D object by using the complementary information between both. In the proposed framework, we first take a sparse 4-beam LiDAR and make it dense image coordinate by using a simple and fast depth completion method. And then, the feature of stereo images and sparse LiDAR depth maps is extracted respectively, an feature attention

fusion module is proposed to integrate the feature information from two pathways. Next, this network takes Stereo RPN [17] to output corresponding left and right RoI proposals. Left and right feature maps are fed into two different branches. One is the stereo regression branch to regress accurate 2D stereo boxes, dimensions, viewpoint angle and 2D center. Another is the depth prediction branch employed to predict the single-variable depth z of the 3D bounding box's center.

Our main contributions of this paper are summarized as follows:

- We propose a novel multi-modal fusion end-to-end learning framework for 3D object detection, which effectively integrate the complementarities of sparse LiDAR and stereo images.
- An deep attention feature fusion module is proposed to explore the interdependencies of channel features in the sparse LiDAR and stereo images while fusing the significant multi-modality spatial features.
- The proposed method achieves state-of-the-art performance compared with the low-cost sensor based methods without depth map supervision.

2 Related Work

2.1 LiDAR-Based 3D Object Detection

Since LiDAR sensors can provide the more accurate 3D information, most 3D detection approaches [6,9,12,16,20,36,40] utilize LiDAR data as input to obtain the best performance. Current LiDAR data can be processed into different representations for input to 3D object detection, including raw point clouds [20,36], volumetric forms [9,40], and 2D projection [16]. The representations of raw point clouds and volumetric forms can make full use of the 3D information of the object. However, they improve the computation cost drastically, especially for large-scale datasets. To improve the efficiency of 3D representations, the 3D point clouds are projected into a 2D image to utilize standard 2D object detection networks for predicting 3D bounding boxes.

2.2 Monocular-Based 3D Object Detection

Some works focus on 3D object detection using monocular cameras due to its low cost and convenient use. MonoGRNet [27] utilizes instance-level depth estimation to obtain a coarse 3D location, which is then refined by combining early features. M3D-RPN [1] proposes a standalone 3D region proposal network for joint prediction of 2D and 3D boxes. RTM3D [18] first predicts nine perspective keypoints of the 3D bounding box and then leverages geometric constraints of perspective projection to optimize 3D object information. SMOKE [21] uses the prediction information of a single key point paired with each object and the 3D regression information to predict a 3D bounding box. M3DSSD [22] solves the feature mismatching problem based on anchor-based methods by feature alignment and extracts depth-wise features for accurate depth prediction.

2.3 Stereo-Based 3D Object Detection

With the improvement of 3D object detection performance based on stereo vision, the gap with LiDAR-based methods is narrowing. Stereo-RCNN [17] extends Faster RCNN [29] to match objects in stereo images and utilizes dense alignment to refine the center depth of 3D bounding boxes. Disp R-CNN [31] and ZoomNet [39] share a similar idea that constructing the instance point cloud to improve detection quality. Pseudo-LiDAR [35] first converts the depth map from stereo vision to pseudo-LiDAR representation and then applies existing LiDAR-based algorithms to detect 3D bounding boxes. DSGN [5] transforms 2D feature to differentiable volumetric representation for encoding 3D geometry structure in 3D regular space. IDA-3D [24] proposes an IDA module for accurate the depth predicted of objects center to have high performance.

2.4 Multi-modal 3D Object Detection

Recently, some techniques [7,19,25,33,37] are proposed to improve 3D object detection performance by exploiting multiple sensors(e.g. 64 beams LiDAR and camera). Though LiDAR sensors with 64 beams are notoriously expensive, LiDAR sensors with only 4 beams are cheaper and easily affordable. In this respect, Pseudo-LiDAR++ [41] proposes a propagation algorithm to integrate the two data modalities, which takes advantage of sparse LiDAR to de-bias the 3D point cloud converted by the depth map from stereo vision. It is complex since it incorporates several independent networks. SLS-Fusion [23] proposes a approach to fuse sparse LiDAR and stereo camera for depth estimation, which is then converted to Pseudo-LiDAR for 3D object detection. However, it fuse the two data modalities by adding directly, not a weighted, which may lead to non-discriminative depth estimation. Inspired by the above approaches, we propose a novel attention network with fusing sparse LiDAR and binocular stereo images to accurately predict the information of 3D bounding box.

3 Proposed Method

In this section, we introduce the proposed 3D object detection architecture by using binocular stereo images and 4-beam sparse LiDAR information in detail. Our detection architecture consists of three stages: we first extract feature for input binocular stereo images respectively by weight-shared Resnet network and extract feature for sparse LiDAR in the same way as stereo images. And then different modal features are fused by attention fusion module. Finally, stereo information and single depth is obtained by regression network to predict 3D bounding boxes. Our architecture is shown in Fig. 1. In this architecture, we fuse LiDAR point cloud information to stereo images feature to augment image features with geometry information accuracy of LiDAR features. However, instead of directly using a 3D point cloud from 4-beam LiDAR, we form two sparse LiDAR depth maps corresponding to stereo images by reprojecting the 4-beam

LiDAR to both left and right image coordinates using the calibration parameters. LiDAR can provide accurate 3D information for 3D object detection. However, the observation is sparse, especially 4-beam LiDAR. Here, we perform depth completion of sparse LiDAR depth maps to produce dense depth maps, similar to the approach in [14]. The holes in the sparse LiDAR depth image are filled by morphological operations and Gaussian blurring operations using nearby valid depth values. The filled depth image is then normalized by the maximum depth value in the dataset, resulting in depth values between 0 and 1. Next, we present each component in detail.

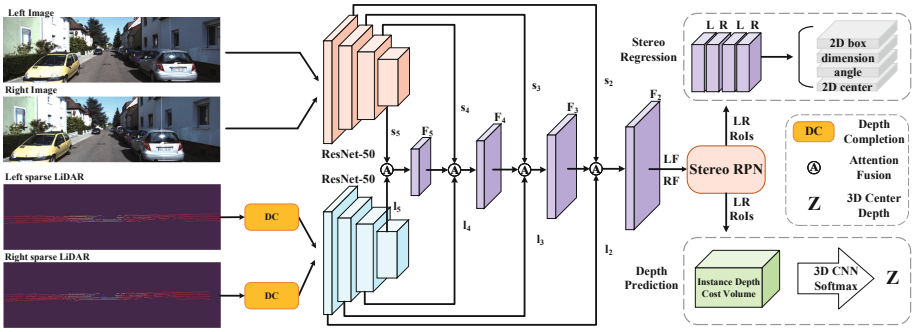

Fig. 1. Network architecture. Our network has three stages. First, sparse LiDAR and stereo RGB images use ResNet-50 as encoder to extract features respectively. Next, stereo images features and their corresponding sparse LiDAR features are fused by attention mechanism. After left and right features (LF,RF) passed through Stereo RPN, we obtain rough alignment region of interest of left and right view (LR RoIs). Finally, we predict position, dimensions and orientation of 3D bounding box.

3.1 Feature Extraction

The stereo images and sparse LiDAR use identical feature encoder architectures, one for each input sensor information. Both encoders for stereo images and sparse LiDAR consist of a series of ResNet blocks. By convolution with stride and downsampling operation, the feature resolution eventually is 1/16 of the input. Each feature encoder weight are shared with left and right input.

We propose a deep fusion approach to fuse sparse LiDAR and stereo image features hierarchically. Specifically, we fuse left sparse LiDAR with corresponding left feature maps in this module, which is the same way for the right. For a network with L layers in encoder stage, early fusion [15,34] combines features from multiple views in the input stage:

$$F_L = D_L(D_{L-1}(\cdots D_l(D_1(F_0^s \oplus F_0^l)))) \tag{1}$$

where D_l is feature transformation function, $\oplus$ is a join operation (e.g., summation [15], concatenation [34]), F_0^s, F_0^l are the input information of stereo images

and sparse LiDAR data respectively. Recently, [26] uses separate subnetworks to learn feature transformation independently and combines their outputs in the prediction stage:

$$\begin{aligned} F_L = D_L^s(D_{L-1}^s(\cdots D_1^s(F_0^s))) \\ \oplus D_L^l(D_{L-1}^l(\cdots D_1^l(F_0^l))) \end{aligned} \tag{2}$$

where D^s, D^l are the separate feature transformation function of stereo images and LiDAR data respectively.

To make more interactions among features from different modalities, the following deep feature fusion process is presented as:

$$\begin{aligned} &F_{i+1} = F_i \oplus F_j^s \oplus F_j^l \\ &i\epsilon\{l+1, \cdots, L\}; j\epsilon\{L-l+1, \cdots, 2\} \end{aligned} \tag{3}$$

where F_i represents the fused feature, F_g^s, F_g^l are the feature activations output of stereo images and sparse LiDAR by each stage's last block in encoder, F_l^s, F_l^l refer to the last left view feature output in encoder. By this fusion, higher resolution features are produced by upsampling feature obtained by Resnet network in higher layers of stereo image pathway and LiDAR pathway, which are spatially coarser, but semantically stronger feature. These features are then enhanced with lower layer features from the stereo image pathway and sparse LiDAR pathway via connections. Moreover, These lower layer features from the stereo image pathway and LiDAR pathway are of lower-level semantics, but its activations are more accurately localized due to its higher resolution. Therefore, the fused features have high-level semantics with higher resolution, which is beneficial for object detection. In our network, F_1^r and F_1^s aren't added to the fusion module due to its large memory footprint.

Since the input sparse depth is strongly related with the decoder output(the prediction depth of object Z), features from the sparse depth should contribute more in the decoder. As such, we add the features from the sparse depth onto the stereo features in decoder instead of concatenation. As the summation favors the features on both sides in the same domain [4], the decoder is encouraged to learn features more related to depth, which keep consistent with the feature from the sparse depth. However, the 4 laser beams LiDAR are too sparse to alone provide sufficient information for 3D detection. Therefore, the fusion is directed from LiDAR steam to image steam to augment image features.

3.2 Attention Fusion

As Eq. 3 indicates that features of different models are fused equally, not a weighted, which may lead to the different importance of each models cannot be correctly reflected. To solve this problem, we employ an attention mechanism to add sparse LiDAR feature into image feature, which sets the weight w_i for each feature level. Since the depth information in sparse LiDAR is accurate, we hope to capture the depth information from sparse LiDAR map to stereo

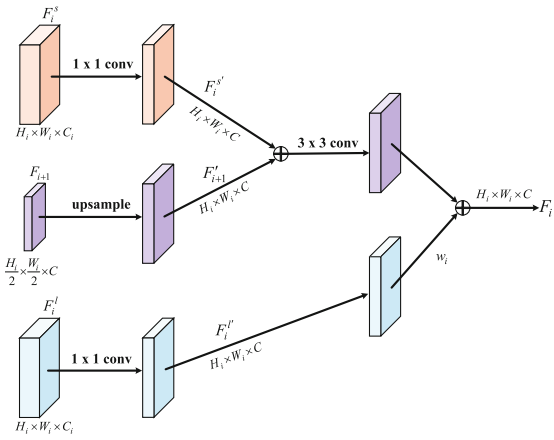

Fig. 2. Illustration of attention fusion module.

images. Therefore, the w_i is obtained by calculating the correlation between sparse LiDAR and corresponding stereo feature maps on each level. It is defined as:

$$w_i = \cos < F_i^s, F_i^l >= \frac{F_i^{s\top} \cdot F_i^l}{\parallel F_i^s \parallel \cdot \parallel F_i^l \parallel}, i = 2 \ldots, 5 \tag{4}$$

where $F_i^s, F_i^l \in \mathbb{R}^{(H_i \times W_i \times C_i) \times 1}$ are the i^{th} stereo images and sparse LiDAR feature maps in the feature extraction. Technically, as shown in Fig. 2, we first upsample F_{i+1} by a factor of 2 into $F_{i+1}^{'} \in R^{H_i \times W_i \times C}$ (using nearest neighbor upsampling for simplicity). Next, we apply 1×1 convolution operation to F_i^s and F_i^l to reduce channel dimensions. The process can be described as:

$$\begin{aligned} F_{i+1}^{'} &= upsample(F_{i+1}) \\ F_i^{s'} &= f_{1\times1}(F_i^s) \\ F_i^{l'} &= f_{1\times1}(F_i^l) \end{aligned} \tag{5}$$

where *upsample* is the up-sampling operation via nearest neighbor interpolation, and $f_{1\times1}$ refers to the 1×1 convolution layer.

Further, we fuse the upsampled feature map $F_{i+1}^{'}$ and the corresponding $F_i^{s'}$ feature map by element-wise addition. Here, a 3×3 convolution is appended on each merged feature map to reduce the aliasing effect of upsampling. Finally, the merged feature is added to the sparse LiDAR feature $F_i^{l'}$, which applies the w_i. The output feature is computed as follow:

$$F_i = f_{3\times3}(F_i^{s'} + F_{i+1}^{'}) + w_i \cdot F_i^{l'} \tag{6}$$

where $f_{3\times3}$ represent the 3×3 convolution layer. The fusion result F_i is exactly the higher level feature of the next fusion stage. This process is iterated until

the final feature map is generated. To start the iteration, we simply to produce the init fusion feature map F_5, which can be formulated as:

$$F_5 = f_{3\times3}(F_i^{s'} + w_5 \cdot F_i^{l'}) \tag{7}$$

where $F_i^{s'}, F_i^{l'}$ are the 5^{th} feature level of the stereo image and sparse LiDAR, respectively, which is used in the decoder stage.

3.3 3D Object Information Regression Prediction

After feature extraction and fusion, we employ stereo Region Proposal Network (RPN) module [17] to generate some pairs of Regions of Interest (RoI) in the left and right images. Different of RPN, the stereo RPN produces an union RoI for left and right images in order to ensure the starting points of each pair of RoIs, and then six regressing terms are used to predict the offsets of anchor box in left and right images. The six regressing terms include the offsets of horizontal and vertical coordinates, the offsets of width and height of the 2D box in left image, and the offsets of horizontal coordinate and width in right image. After stereo RPN, we can obtain corresponding feature maps in left-right proposal pairs by applying RoI Align [10] on the left and right feature maps respectively at appropriate pyramid level. The left and right RoI features are concatenated and fed into the stereo regression branch, which includes four sub-branches to predict 2D box, dimension, and viewpoint angle, 2D center respectively. In addition to the stereo regression branch, we predict the 3D depth of object center in the depth prediction branch. Instead of predicting the depth information of each pixel, we only compute the depth of instance object between left and right images. In our network, we takes Instance-Depth-Aware (IDA) module [24] to predict the depth of instance object center.

Finally, 3D bounding box can be represented by 2D box, dimension, orientation, and depth information.

3.4 Implementation Details

Our loss function can be formulated as:

$$\begin{aligned} L = {} & w_1 L_{cls}^s + w_2 L_{reg}^s + w_3 L_{box}^r + w_4 L_{\dim}^r \\ & + w_5 L_{\alpha}^r + w_6 L_{ctr}^r + w_7 L_z^d \end{aligned} \tag{8}$$

where we use $(\cdot)^s$,$(\cdot)^r$ and $(\cdot)^d$ for representing the loss in Stereo RPN module [17], Stereo Regression module [10], and Depth Prediction module [24]respectively. L_{cls}^s and L_{reg}^s denote the loss of classification and regression on stereo RPN module respectively. L_{box}^r,$L_{\dim}^r$, L_{α}^r, L_{ctr}^r are the loss of stereo boxes, dimension, viewpoint, 2D center on stereo Regression respectively. L_z^d is the loss of depth on Depth Precision module. Each loss is weighted to balance the whole loss following [13]. In our experiment, the weight is 1,1,1,3,0.1,2,0.2 separately.

Two weight-shared ResNet-50 [11] architecture are treated as the feature encoder for stereo images and sparse LiDAR, respectively. For data augmentation, we flip and exchange the left and right image in the training set and mirror the image information. For sparse LiDAR information, we first project it on image planes using the calibration parameters and then apply the same flipping strategy as previous stereo images. Our model is implemented under PyTorch 1.1.0, CUDA 10.0. By default, we train our network with batch-size 4 on 4 NVIDIA Tesla V100 GPUs for 65000 iterations, and the overall training time is about 26 h. We apply stochastic gradient descent(SGD) optimizer with initial learning rate 0.02.

4 Experiments

Table 1. 3D object detection results evaluated on the KITTI object validation set. We report average precision of bird's eye view (AP_{bev}) and 3D boxes (AP_{3d}) for the car category. PL(AVOD) is reported by [5] without LiDAR supervision. We use original KITTI evaluation metric here.

Method	AP_{bev}(IoU = 0.5)			AP_{3d}(IoU = 0.5)		
	Easy	Moderate	Hard	Easy	Moderate	Hard
MonoGRNet [27]	54.21	39.69	33.06	50.51	36.97	30.82
M3D-RPN [1]	55.37	42.49	35.29	48.96	39.57	33.01
RTM-3D [18]	57.47	44.16	42.31	54.36	41.90	35.84
Decoupled-3D [2]	73.22	54.31	45.97	69.40	50.50	42.46
MLF [38]	-	53.56	-	-	19.54	-
3DOP [3]	55.04	41.25	34.55	46.04	34.63	30.09
TL-Net [28]	62.46	45.99	41.92	59.51	43.71	37.99
PL(AVOD) [35]	76.8	65.1	56.6	75.6	57.9	49.3
Stereo R-CNN [17]	87.13	74.11	58.93	85.84	66.28	57.24
IDA-3D [24]	88.05	76.69	67.29	87.08	74.57	60.01
Ours	88.58	77.70	68.15	87.92	75.32	66.27

4.1 KITTI Dataset

Our method is evaluated on the challenging KITTI object detection dataset [8], which provides 7481 training images and 7581 testing images. In this paper, the 4-beam LiDAR signal on KITTI benchmark is simulated by sparsifying the original 64-beam signal as the way of [8]. Following [3], the training data is divided into roughly the same amount of training set and validation set. The ground-truth of Car, Pedestrian and Cyclist is provided by annotations in the training set. Following the KITTI settings, each category is divided into three regimes: easy, moderate, and hard, depending on the occlusion/truncation and the size of 2D box height.

Table 2. 3D object detection results evaluated on the KITTI object validation set. We report average precision of bird's eye view (AP_{bev}) and 3D boxes (AP_{3d}) for the car category. PL(AVOD) is reported by [5] without LiDAR supervision. We use original KITTI evaluation metric here.

Method	AP_{bev}(IoU = 0.7)			AP_{3d}(IoU = 0.7)		
	Easy	Moderate	Hard	Easy	Moderate	Hard
MonoGRNet [27]	24.97	19.44	16.30	13.88	10.19	7.62
M3D-RPN [1]	25.94	21.18	17.90	20.27	17.06	15.21
RTM-3D [18]	25.56	22.12	20.91	20.77	16.86	16.63
Decoupled-3D [2]	44.42	29.69	24.60	26.95	18.68	15.82
MLF [38]	-	47.42	-	-	9.80	-
3DOP [3]	12.63	9.49	7.59	6.55	5.07	4.10
TL-Net [28]	29.22	21.88	18.83	18.15	14.26	13.72
PL(AVOD) [35]	60.7	39.2	37.0	40.0	27.4	25.3
Stereo R-CNN [17]	68.50	48.30	41.47	54.11	36.69	31.07
IDA-3D [24]	70.68	50.21	42.93	54.97	37.45	32.23
Ours	71.62	52.15	44.6	56.00	39.77	33.64

4.2 Evaluation Metrics

We use average precision of 3D detection (AP_{3d}) and average precision of bird's-eye-view (BEV) detection (AP_{bev}) to evaluate the performance of our method. The results of AP_{3d} and AP_{bev} on the validation set are reported on the car's category. It is worth noting that the Intersection over Union (IoU) thresholds are set at 0.5 and 0.7, following previous works [17,24]. In order to compare with previous approaches fairly, our validation results are evaluated using the original evaluation code, which calculates AP with 11 recall positions instead of 40 recall positions.

4.3 Main Results

The main results as shown in Table 1, 2 (IoU = 0.5,0.7), where we compare the proposed method with previous state-of-the-art approaches from low-cost sensors (monocular to binocular). Our method obtains a significant improvement in comparison to previous monocular-based methods in all cases across all IoU thresholds. Comparing with binocular-based methods, our method gains the highest performance at 0.5 IoU and 0.7 IoU. Specifically, our approach outperforms previous state-of-art IDA-3D [24] by 1.94% and 1.67% in AP_{bev} across moderate and hard sets at 0.7 IoU, respectively. The similar improvement trends can be obverse in AP_{3d}, which manifest that our approach achieves consistent improvement compared with other approaches. The results of our approach in the moderate and hard set on the most metric AP_{3d} (IoU = 0.7) outperform

IDA-3D by over 2.32% and 1.41%. Although only a small margin of our approach outperforms IDA-3D (IoU = 0.7) in the easy set,the proposed method gain significant improvement over 6.26% on AP_{3d} (IoU = 0.5) in the hard set. The reason is that our method fuses sparse LiDAR information to extract feature, which provides more accurate depth.

In addition to the aforementioned comparison methods, we also compare with the current multi-modality based method. Since these methods [19,25,33,37] use 64-beams LiDAR information as input or intermediate supervision, we only compare the proposed method to the Pseudo-LiDAR++ (PL++) [41], which takes L#+S as input. PL++ produced dense depth map with 64-beams LiDAR supervision, however, the proposed method only use 4-beam LiDAR. We show the reproduced result of PL++ without 64-beam LiDAR supervision (PL++*(AVOD)) in Table 3. The experimental results in Table 3 demonstrate that our approach outperforms PL++*(AVOD) approach on some metrics. Specifically, we achieve 11.3% improvement for AP_{3d} using IoU = 0.7 in the easy set. For AP_{bev}, our method gains over 7.82% improvements. The reason is the proposed network pays more attention to nearby objects, while the 3D point cloud is projected onto the front-viewing image. In addition, Table 3 also reports the running time comparison between PL++*(AVOD) method and the proposed method. Our approach has a high speed of 0.116 s per frame at inference time, which far exceeds PL++*(AVOD) method. The efficiency is attributed to our network, which is an end-to-end architecture with light weight modules compared to the network of PL++ method.

Table 3. AP_{bev} and AP_{3d} of IoU = 0.7 on KITTI validation set.

Method	Running time (s/frame)	AP_{bev}		AP_{3d}	
		Easy	Moderate	Easy	Moderate
PL++*(AVOD) [41]	0.519	63.8	57.2	44.7	38.9
Ours	0.116	71.62	52.15	56	39.77

In addition to the above quantitative analysis, we also show the qualitative detection results of several scenarios in the KITTI validation set in Fig. 3. It can be observed that the proposed method can accurately detect objects in these scenarios, and the detected 3D box are well aligned on the vertical view and front view point cloud.

4.4 Ablation Study

In this section, we analyze the effectiveness of Sparse LiDAR, Depth Completion, and Attention Fusion components in our method. Results are shown in Table 4. In condition of just using Sparse LiDAR, we directly add the sparse LiDAR features into their corresponding stereo images features at appropriate level in

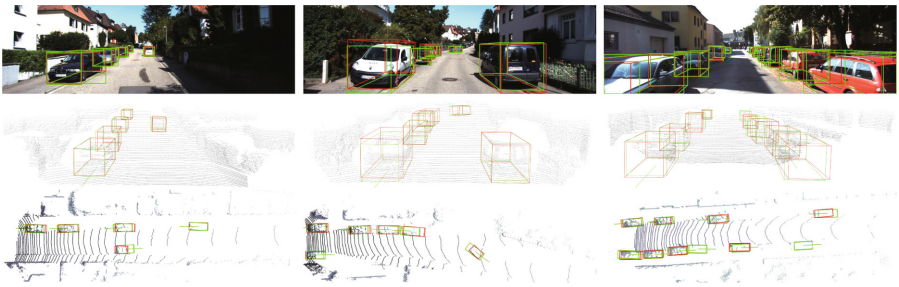

Fig. 3. 3D object detection results on the KITTI validation set. The predicted results are shown in green box and the ground truth are shown in red box. In order to facilitate observation, the detection results are shown on the vertical and front view point cloud. (Color figure online)

the decoder. In condition of not using Depth Completion, we regard the sparse LiDAR depth maps as the depth feature extractor input. In condition of not using Attention Fusion module, the sparse LiDAR feature maps are added directly to their corresponding image feature maps.

From Table 4, we can see that the performance achieves significant improvement, when sparse LiDAR is only applied, which demonstrates that sparse LiDAR is crucial for high-quality 3D detection. The absence of Depth Completion makes the percentage drop from 38.83% to 37.31% on AP_{3d} with a threshold IoU = 0.7 in the moderate set. Besides, the performance of our AP_{bev} has a drop of 1.87% at 0.7 IoU in the easy set when Attention Fusion is removed. Large improvements can be observed on all metrics by using these three key components together, and results surpass almost all prior low-cost based methods.

Table 4. Ablation studies on the KITTI validation set.

Sparse LiDAR	Attention fusion	Depth completion	AP_{bev}(IoU = 0.7)			AP_{3d}(IoU = 0.7)		
			Easy	Moderate	Hard	Easy	Moderate	Hard
			67.66	48.74	41.73	53.35	36.49	31.26
✓			70.06	50.47	42.86	55.77	37.31	31.6
✓		✓	69.75	51.52	44.22	55.93	38.83	33.05
✓	✓		71.35	51.46	43.87	55.8	37.91	32.7
✓	✓	✓	71.62	52.15	44.6	56	39.77	33.64

5 Conclusion

In this paper, we take 4-beam sparse LiDAR and stereo images as input for 3D object detection. The key idea is that a deep fusion module combines features

across multiple modalities by utilizing an attention mechanism. Our deep attention feature fusion module explores the interdependencies of channel features in the sparse LiDAR and stereo images while fusing the significant multi-modality spatial features. Experimental results show higher 3D detection performance of our proposed method compared with other low-cost sensor based method.

References

1. Brazil, G., Liu, X.: M3d-RPN: Monocular 3d region proposal network for object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9287–9296 (2019)
2. Cai, Y., Li, B., Jiao, Z., Li, H., Zeng, X., Wang, X.: Monocular 3d object detection with decoupled structured polygon estimation and height-guided depth estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10478–10485 (2020)
3. Chen, X., Kundu, K., Zhu, Y., Ma, H., Fidler, S., Urtasun, R.: 3d object proposals using stereo imagery for accurate object class detection. IEEE Trans. Pattern Anal. Mach. Intell. **40**(5), 1259–1272 (2017)
4. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1907–1915 (2017)
5. Chen, Y., Liu, S., Shen, X., Jia, J.: DSGN: deep stereo geometry network for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12536–12545 (2020)
6. Cheng, B., Sheng, L., Shi, S., Yang, M., Xu, D.: Back-tracing representative points for voting-based 3d object detection in point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8963–8972 (2021)
7. Choi, C., Choi, J.H., Li, J., Malla, S.: Shared cross-modal trajectory prediction for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 244–253 (2021)
8. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361. IEEE (2012)
9. He, C., Zeng, H., Huang, J., Hua, X.S., Zhang, L.: Structure aware single-stage 3d object detection from point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11873–11882 (2020)
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
12. He, Y., et al.: DVFENet: dual-branch voxel feature extraction network for 3d object detection. Neurocomputing **459**, 201–211 (2021)
13. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7482–7491 (2018)
14. Ku, J., Harakeh, A., Waslander, S.L.: In defense of classical image processing: fast depth completion on the CPU. In: 2018 15th Conference on Computer and Robot Vision (CRV), pp. 16–22. IEEE (2018)

15. Ku, J., Mozifian, M., Lee, J., Harakeh, A., Waslander, S.L.: Joint 3d proposal generation and object detection from view aggregation. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1–8. IEEE (2018)
16. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12697–12705 (2019)
17. Li, P., Chen, X., Shen, S.: Stereo R-CNN based 3d object detection for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7644–7652 (2019)
18. Li, P., Zhao, H., Liu, P., Cao, F.: RTM3D: real-time monocular 3d detection from object keypoints for autonomous driving. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12348, pp. 644–660. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58580-8_38
19. Liang, M., Yang, B., Chen, Y., Hu, R., Urtasun, R.: Multi-task multi-sensor fusion for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7345–7353 (2019)
20. Liu, Y., Fan, B., Xiang, S., Pan, C.: Relation-shape convolutional neural network for point cloud analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8895–8904 (2019)
21. Liu, Z., Wu, Z., Tóth, R.: Smoke: Single-stage monocular 3d object detection via keypoint estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 996–997 (2020)
22. Luo, S., Dai, H., Shao, L., Ding, Y.: M3DSSD: monocular 3d single stage object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6145–6154 (2021)
23. Mai, N.A.M., Duthon, P., Khoudour, L., Crouzil, A., Velastin, S.A.: Sparse lidar and stereo fusion (SLS-fusion) for depth estimation and 3d object detection. arXiv preprint arXiv:2103.03977 (2021)
24. Peng, W., Pan, H., Liu, H., Sun, Y.: IDA-3D: instance-depth-aware 3d object detection from stereo vision for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13015–13024 (2020)
25. Prakash, A., Chitta, K., Geiger, A.: Multi-modal fusion transformer for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7077–7087 (2021)
26. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum PointNets for 3d object detection from RGB-D data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 918–927 (2018)
27. Qin, Z., Wang, J., Lu, Y.: MonoGRNet: a geometric reasoning network for monocular 3d object localization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8851–8858 (2019)
28. Qin, Z., Wang, J., Lu, Y.: Triangulation learning network: from monocular to stereo 3d object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7607–7615. IEEE (2019)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. Adv. Neural. Inf. Process. Syst. **28**, 91–99 (2015)
30. Shi, Y., Guo, Y., Mi, Z., Li, X.: Stereo centerNet-based 3d object detection for autonomous driving. Neurocomputing **471**, 219–229 (2022)

31. Sun, J., Chen, L., Xie, Y., Zhang, S., Jiang, Q., Zhou, X., Bao, H.: DISP R-CNN: stereo 3d object detection via shape prior guided instance disparity estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10548–10557 (2020)
32. Tang, Y., Dorn, S., Savani, C.: Center3D: Center-based monocular 3d object detection with joint depth understanding. In: Akata, Z., Geiger, A., Sattler, T. (eds.) DAGM GCPR 2020. LNCS, vol. 12544, pp. 289–302. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-71278-5_21
33. Vora, S., Lang, A.H., Helou, B., Beijbom, O.: Pointpainting: Sequential fusion for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4604–4612 (2020)
34. Wang, T.H., Hu, H.N., Lin, C.H., Tsai, Y.H., Chiu, W.C., Sun, M.: 3d lidar and stereo fusion using stereo matching network with conditional cost volume normalization. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5895–5902. IEEE (2019)
35. Wang, Y., Chao, W.L., Garg, D., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8445–8453 (2019)
36. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph CNN for learning on point clouds. ACM Trans. Graph. (TOG) **38**(5), 1–12 (2019)
37. Xiao, Y., Codevilla, F., Gurram, A., Urfalioglu, O., López, A.M.: Multimodal end-to-end autonomous driving. IEEE Trans. Intell. Transp. Syst. (2020)
38. Xu, B., Chen, Z.: Multi-level fusion based 3d object detection from monocular images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2345–2353 (2018)
39. Xu, Z., et al.: ZoomNet: part-aware adaptive zooming neural network for 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12557–12564 (2020)
40. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11784–11793 (2021)
41. You, Y., et al.: Pseudo-lidar++: accurate depth for 3d object detection in autonomous driving. In: ICLR (2020)