

EPNet++: Cascade Bi-directional Fusion for Multi-Modal 3D Object Detection

Zhe Liu, Tengting Huang, Bingling Li, Xiwu Chen, Xi Wang, Xiang Bai

Abstract—Recently, fusing the LiDAR point cloud and camera image to improve the performance and robustness of 3D object detection has received more and more attention, as these two modalities naturally possess strong complementarity. In this paper, we propose EPNet++ for multi-modal 3D object detection by introducing a novel Cascade Bi-directional Fusion (CB-Fusion) module and a Multi-Modal Consistency (MC) loss. More concretely, the proposed CB-Fusion module enhances point features with plentiful semantic information absorbed from the image features in a cascade bi-directional interaction fusion manner, leading to more powerful and discriminative feature representations. The MC loss explicitly guarantees the consistency between predicted scores from two modalities to obtain more comprehensive and reliable confidence scores. The experimental results on the KITTI, JRDB and SUN-RGBD datasets demonstrate the superiority of EPNet++ over the state-of-the-art methods. Besides, we emphasize a critical but easily overlooked problem, which is to explore the performance and robustness of a 3D detector in a sparser scene. Extensive experiments present that EPNet++ outperforms the existing SOTA methods with remarkable margins in highly sparse point cloud cases, which might be an available direction to reduce the expensive cost of LiDAR sensors. Code is available at: <https://github.com/happinessiz/EPNetV2>.

Index Terms—3D Object Detection, Multi-Modal Fusion, Cascade Bi-directional, Consistency.



1 INTRODUCTION

3D object detection serves as one fundamental technique in self-driving, whose main research can be classified into three categories based on the input modalities: image [1–11], LiDAR [12–21] and multi-sensor fusion [22–34]. In general, LiDAR-based methods [12–21] usually achieve superior 3D object detection performance over image-based [1–11], as image-based methods can not obtain the accurate localization of 3D objects due to the lack of depth and 3D structural information. However, LiDAR-based methods often produce false-positive detection results, especially when the geometric structure of a background object is similar to the 3D objects to be detected. Moreover, LiDAR-based methods may fail to detect distant or small objects, where the point clouds are too sparse even if collected by a high-quality LiDAR. Yet, distant or small objects are easier to be detected with the help of appearance cues in 2D camera images. Motivated by these observations, our intuition is to design an effective deep learning framework for fully exploiting the complementary semantics of camera images and point clouds. However, integrating the cues of camera images and point clouds for 3D object detection is non-trivial due to their quite different data representations.

The previous works on multi-sensor fusion follow the pipeline of sequential fusion [26, 32, 33] or multi-sensor

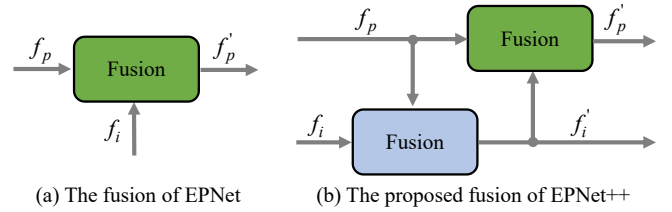


Fig. 1. Comparison of the fusion approach in EPNet and that in this paper. f_i and f_p represent the image feature and point feature, while f'_i and f'_p denote the enhanced image feature and enhanced point feature.

input reasoning [23, 25, 34]. The former approaches train the networks with images and point clouds separately and implement the multi-modal fusion by connecting the two stages. The final detection performance is highly dependent on the performance of each stage, posing extra optimization difficulty for the fusion module to coordinate both modalities. The latter methods [23, 25, 34] alternatively study the fusion of camera images and Bird’s-Eye View (BEV) images, which are generated from point clouds through the operations of perspective projection and voxelization. However, the 3D information of point clouds on the height dimension is usually lost when performing BEV projection. Besides, a voxel feature usually comes from the aggregation of multiple point features. Therefore, the correspondence between voxels and image features is coarser than that based on the raw point cloud. These two disadvantages might limit their upper bound on detection tasks.

Different from previous methods [23, 25, 26, 32–34], our preliminary work named EPNet [27] proposed a *LiDAR-guided Image Fusion* (LI-Fusion) module to enrich the semantic information of point clouds for improving 3D detection. The main characteristic is that EPNet fuses the deep features

- Z. Liu, B. Li and X. Chen are with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, 430074, China.
E-mail: {zheliu1994, blli, xitouchen}@hust.edu.cn
- T. Huang is with Megvii (Face++) Inc., Beijing, 100190, China.
E-mail: tengtinghuang@foxmail.com.
- X. Wang is the founder and CEO of CalmCar, Suzhou, 215168, China.
E-mail: xi.wang@calmcar.com.
- X. Bai is with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, 430074, China.
E-mail: xbai@hust.edu.cn.

from raw point clouds and camera images at different levels of resolution in a fine-grained point-wise paradigm, which can be trained in an end-to-end manner. Though effective, EPNet simply utilizes image features for enhancing point features in a point-wise manner as shown in Fig. 1 (a), which may ignore the benefits of the bi-directional feature interaction of both modalities. Intuitively, point features can provide important geometric cues (*e.g.*, the contour and depth of objects) for image features, which can assist the image network to achieve more accurate object segmentation and in turn further promote the point cloud network.

In this paper, we investigate a bi-directional fusion mechanism for multi-sensor 3D object detection, which is named *Cascade Bi-directional Fusion* (CB-Fusion) module as illustrated in Fig. 1 (b). We adopt point features to enhance image features first and then reinforce point features. Although the manner of feature enhancement in a reversed order is also an option, we select the order in Fig. 1 (b). Since we use point features instead of image features to predict 3D boxes, it is more reasonable to aggregate the enhanced image features to point features. We argue that the proposed CB-Fusion is more effective than the direct fusion in EPNet, as it allows more interaction between the two modalities. More concretely, the geometric information from LiDAR points is helpful to obtain a more discriminative image representation that is robust to illumination change, and the enhanced image representation in turn brings more plentiful semantics to LiDAR points. Besides, CB-Fusion is not limited to point-based detectors [15, 35], but can also be applied to popular voxel-based [17, 36, 37] and transformer-based detectors [38] to boost the detection performance.

Besides studying multi-sensor fusion at the feature level, we propose a training strategy for further exploring the complementarity of point clouds and images. Our assumption is that the prediction results based on a point cloud should be consistent with those based on its corresponding camera image. Specifically, given a point that is predicted to be from a foreground object, its projection on an image should be predicted to be from the same object as well. Based on this observation, we adopt the proposed *Multi-Modal Consistency loss* (MC loss) to guarantee that the network outputs based on different modalities are consistent with each other during the training process.

Finally, in this paper, we emphasize a critical but easily overlooked problem: handling 3D object detection in highly sparse point cloud scenes, which might be an effective strategy to reduce the cost of the LiDAR sensor. Extension experiments demonstrate that our EPNet++ significantly outperforms these existing SOTA methods in highly sparse point cloud cases with the aid of the novel CB-Fusion module and effective MC loss.

In summary, the main contributions in EPNet++ are as follows: 1) The proposed CB-Fusion module establishes the point-wise feature correspondence between two modalities at different levels of resolution and achieves the cascade bidirectional feature interaction, leading to more comprehensive feature representations for 3D detection tasks. Besides, CB-Fusion shows superior generalization capability to multiple architectures, including point-based [15], voxel-based [17, 36, 37] and transformer-based [38] detectors. 2) The MC loss encourages the consistency of the predicted

confidences of two modalities, which can alleviate the ambiguity, especially when there is a huge gap between the confidence scores of two modalities. 3) The proposed EPNet++ achieves the competitive even SOTA results on three common 3D object detection benchmark datasets, *i.e.*, the KITTI dataset [39], SUN-RGBD dataset [40] and JRDB dataset [41]. Moreover, EPNet++ achieves promising performance even for highly sparse point cloud scenes, which alleviates the demanding requirements of expensive high-resolution LiDAR sensors.

2 RELATED WORKS

2.1 3D Object Detection with Camera Images

Achieving 3D object detection at a lower cost is still an important goal for autonomous driving technology. Compared with LiDAR sensors, the price of camera sensors is much lower. Therefore, current 3D object detection methods pay much attention to camera images, such as monocular [1–6] and stereo images [7, 8, 42]. Specifically, Chen *et al.* [9] obtain 2D bounding boxes with a CNN-based object detector and infer their corresponding 3D bounding boxes with semantic, context, and shape information. Deep3DBox [11] estimates localization and orientation from 2D bounding boxes of objects by exploiting the constraint of projective geometry. However, these methods based on the camera image have difficulty generating accurate 3D bounding boxes due to the lack of depth information. Compared with these image-based methods, our proposed EPNet++ achieves much better detection performance through properly fusing the camera image and the highly sparse point cloud (*e.g.*, 2048 points) provided by a LiDAR sensor.

2.2 3D Object Detection with Point Clouds

According to data representations of point clouds, we further divide the 3D object detection with point clouds into voxel-based and point-based methods.

voxel-based Methods. Since point clouds are irregular data formats, some existing methods attempt to convert a point cloud into a regular voxel grid representation and then achieve 3D object detection via 2D/3D convolutional neural network (CNN). In detail, VoxelNet [13] divides a point cloud into voxels and employs stacked voxel feature encoding layers to extract voxel features. SECOND [36] further introduces a sparse convolution operation to improve the computational efficiency of [13]. To get rid of time-consuming 3D convolution operations, The following work PointPillars [16] converts the point cloud to a pseudo-image and applies 2D CNN to produce the final detection results. Some other works [37, 43–45] extract the voxel features via 3D convolutional operations and obtain more accurate predictions in a coarse-to-refine two-stage manner. These methods are effective, but they are sensitive to the parameter of voxel size in terms of detection performance and speed.

Point-based Methods. PointRCNN [15] is a pioneering point-based two-stage detector, which directly produces 3D proposals based on PointNets [46, 47] from the whole point cloud and refines these coarse proposals via 3D ROI pooling operation on the RCNN stage. Then, STD [48] applies

a point-based proposal generation to achieve high recall through spherical anchors. The following work 3DSSD [49] proposes a combination of F-FPS and D-FPS to improve the STD [48] in terms of accuracy and efficiency. In addition, VoteNet [35] presents a novel deep hough voting for 3D object detection. These methods can directly process the raw point cloud. Therefore, to establish the accurate correspondence between the camera image and LiDAR point cloud, we utilize the popular point-based detector PointRCNN [15] as our backbone in this paper.

2.3 3D Object Detection with Multiple Sensors

Recently, much progress has been made in exploiting multiple sensors, such as camera images and LiDAR point clouds. We roughly divide these existing fusion methods into three categories: multi-view fusion, voxel-based and image fusion, and point-based and image fusion.

Multi-View Fusion Methods. As the information of a single view (*e.g.*, the front view image or BEV point cloud) is usually not sufficient for understanding real scenes, some researchers try to explore multi-view fusion to improve the performance of 3D object detection tasks. MV3D [12] and AVOD [50] refine the detection box via fusing BEV and camera feature maps for each ROI region. Although these multi-view approaches usually outperform single-view-based methods, they still suffer from information loss due to the process of converting point clouds to a specific view. In contrast, our proposed fusion module directly operates on the LiDAR point cloud and thus better retains the geometric structure information.

Voxel & Image Fusion Methods. Many recent LiDAR-only methods convert the raw LiDAR point cloud to regular voxel grids for 3D object detection, thanks to its effectiveness and efficiency. To further improve the robustness of 3D detectors, some researchers [22, 23, 25, 51] devote their efforts to the voxel-based and camera image fusion methods. Specifically, ConFuse [25] proposes a novel continuous fusion layer, which not only achieves the voxel-wise alignment between BEV and image feature maps but also captures local information to improve the detection performance. MVX-Net [22] proposes to enhance the voxel feature representations with semantic image features by fusing the features of the camera image and LiDAR point cloud in the early stage. 3D-CVF [23] effectively fuses the spatial feature from both camera image and LiDAR point cloud via utilizing a cross-view spatial feature fusion strategy. Due to the quantized error brought from the voxelization operation, these methods have limitations in establishing the accurate corresponding relationship between the camera image and the LiDAR point cloud. Compared with them, our approach achieves fine-grained correspondence by directly dealing with raw point clouds.

Raw Point Cloud & Image Fusion Methods. Considering that the point cloud possesses rich geometric structure information but lacks plentiful semantic information, some researchers [26, 27, 29] have tried to fuse the raw point cloud and the camera image. Specifically, PointFusion [29] and SIFRNet [33] first extract semantic features and produce 2D proposals from camera images using off-the-shelf 2D

detectors [52–55]. The extracted semantic features are then combined with the point features extracted from the corresponding frustum to generate 3D bounding boxes. Point-Painting [26] enriches each point feature with the corresponding output class scores predicted by a pre-trained image semantic segmentation network. PI-RCNN [28] employs a segmentation sub-network to extract full-resolution semantic feature maps from images and then fuses the multi-sensor features via a powerful PACF attention module. ImVoteNet [24] further improves the detection performance of VoteNet [35] by lifting 2D camera images votes as well as geometric, semantic, and texture cues from an off-the-shelf 2D detector and then combining them with 3D votes in point clouds. DenseFusion [56] combines cropped images of interesting objects and corresponding masked point clouds at an object level in a dense pixel-wise fusion, with the help of a segmentation network [57]. Although these methods are effective, their performance still depends on pre-trained 2D detectors or 2D image segmentation networks. Compared with them, our preliminary work EPNet [27] enhances the semantic information of point features with image features at different levels of resolution in a point-wise manner, which is fully end-to-end trainable. In this paper, we further extend EPNet [27] to EPNet++, which leverages a cascade bi-directional fusion paradigm to allow more information exchange between two modalities and obtain a more comprehensive and robust point feature representation for the subsequent estimation of 3D boxes.

3 METHODS

Exploiting the complementary information of multiple sensors is crucial for improving the accuracy and robustness of 3D object detectors, especially for sparse scenes. In this paper, we design a novel multi-modal fusion framework for 3D object detection named EPNet++, which is illustrated in Fig. 2. EPNet++ is trainable in an end-to-end manner, which involves a two-stream RPN for 3D proposal generation and a refinement network to produce more precise 3D bounding boxes. To effectively explore the complementarity of different sensors, we design the fusion mechanism from both the feature representations and loss constraints. First, we propose two feature fusion modules, including a *LiDAR-guided Image Fusion* (LI-Fusion) module to unidirectionally enhance the semantic information of point features with image features and a *Cascade Bi-directional Fusion* (CB-Fusion) module to enable bidirectional feature enhancement between LiDAR and camera images. Second, we employ the *Multi-Modal Consistency loss* (MC loss) to promote consistency of the foreground/background classification confidence predicted from the LiDAR point cloud and camera image. In the following, we present the technical details of EPNet++.

3.1 Two-Stream RPN

As shown in Fig. 2, the two-stream RPN consists of an image stream and a geometric stream, each of which includes an encoder and a decoder. These two streams can be seamlessly bridged through the proposed LI-Fusion/CB-Fusion modules. Concretely, we employ several CB-Fusion modules in the encoder at multiple scales for early fusion, which enables instant and effective information exchanges between

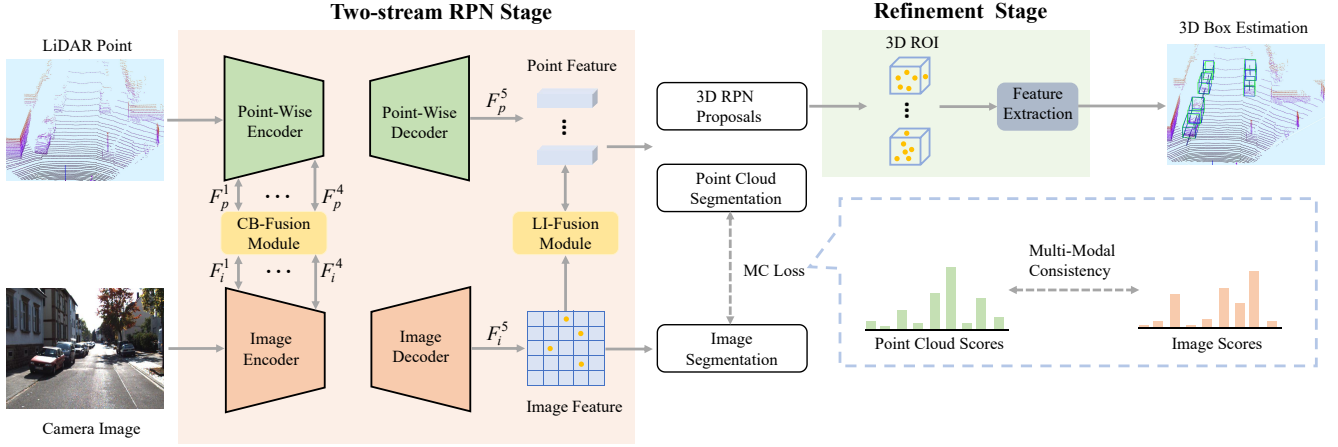


Fig. 2. Illustration of the overall pipeline of EPNet++ for 3D object detection, which includes a two-stream RPN stage and a refinement stage. In the two-stream RPN stage, several *Cascade Bi-directional Fusion* (CB-Fusion) modules are employed to establish bi-directional information exchange pathways between two streams in multiple scales, leading to enhanced feature representations. Besides, we utilize the *Multi-Modal Consistency loss* (MC loss) to encourage the consistency of the segmentation confidences predicted by these two streams, which is helpful for selecting high-quality 3D boxes. Finally, the high-quality 3D proposals are fed into the refinement network to produce the final detection results.

the two streams and produces enhanced features with richer semantic and geometry information. Note that we do not employ CB-Fusion modules in the decoder for two reasons. First, multiple transposed convolution layers (resp. Feature Propagation layers [47]) are used in the image stream (resp. geometric stream) to recover the resolution, which further leads to more serious misalignment. Second, the architecture with the CB-Fusion removed from the decoder is more efficient and thus friendly for deployment. Besides, at the end of the two-stream RPN, we simply adopt an LI-Fusion module to enhance the point features for the following detection task. Although CB-Fusion is also an alternative, we prefer LI-Fusion here for its efficiency when processing image feature maps with high resolution. In the following, we will present more details about each component of the two-stream RPN.

3.1.1 Image Stream

The image stream takes camera images as input and extracts the semantic image information with a set of convolution operations. We adopt a simple structure composed of an image encoder and an image decoder, which is illustrated at the bottom of Fig. 2. The image encoder contains four lightweight convolutional blocks. Each convolutional block consists of two 3×3 convolution layers followed by a batch normalization layer [58] and a ReLU activation function. We set the second convolution layer in each block with a stride of 2 to enlarge the receptive field and save GPU memory. F_i^k ($k=1, 2, 3, 4$) denotes the outputs of these four convolutional blocks. F_i^k provides sufficient semantic image information to enrich the LiDAR point features in different scales. The image decoder includes four parallel transposed convolution layers with different strides to recover the image resolution, leading to feature maps of the same resolution as the original image. Then, we combine them in a concatenation manner and obtain a more representative feature map F_i^5 , which contains rich image semantic information. As is shown later, the feature map F_i^5 is also employed to enhance the LiDAR point features to generate more reliable 3D proposals.

3.1.2 Geometric Stream

As shown in the top branch of Fig. 2, the geometric stream takes LiDAR point cloud as input and generates 3D proposals for the subsequent refinement stage. We take the popular PointNet++ [47] as the backbone of the geometric stream, which is composed of a point-wise encoder for feature aggregation and a point-wise decoder for feature restoration. The encoder consists of four Set Abstraction (SA) layers and the decoder is composed of four corresponding Feature Propagation (FP) layers. For the convenience of description, the outputs of four SA layers and the last FP layer are denoted as F_p^k ($k=1, 2, 3, 4$) and F_p^5 , respectively. We combine the point features F_p^k with the corresponding semantic image features F_i^k with the aid of our CB-Fusion module. Besides, we further enrich the decoded point feature F_p^5 with the high-resolution image feature F_i^5 through an LI-Fusion module to obtain a compact and discriminative feature representation, which is then fed to the RPN head for foreground point segmentation and 3D proposal generation.

3.1.3 LI-Fusion Module

The LI-Fusion module aims at enhancing the point cloud feature with rich semantic information from the image stream. As is illustrated in Fig. 3, the LI-Fusion module involves two core components, *i.e.*, a point-wise correspondence operation and an LI-Fusion layer.

Point-wise Correspondence establishes the point-wise correspondence between 2D location and 3D point cloud through a known projection matrix M . For a particular point p in the point set P , its corresponding position in the camera image is $p' = Mp$. Then we use bilinear interpolation to get the corresponding image feature at the continuous coordinates, which well leverages the neighboring context information around the projected position. This process can be formulated as follows:

$$F_i^{(p)} = \mathcal{B}(F_i^{\mathcal{N}(p')}), \quad (1)$$

where $F_i^{(p)}$ is the corresponding image feature for a specific point p , \mathcal{B} denotes the bilinear interpolation function, and

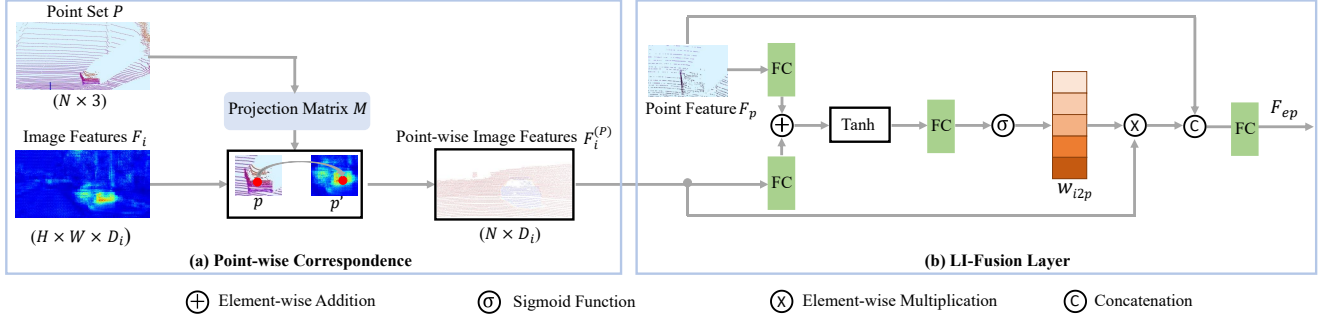


Fig. 3. Illustration of the structure of the LI-Fusion module. FC represents the fully connected layer.

$F_i^{\mathcal{N}(p')}$ represents the image features of the neighboring pixels for the sampling position p' . Finally, we can obtain the total point-wise image features $F_i^{(P)}$ for a point set P .

LI-Fusion Layer. Although concatenation is a simple and common method for feature fusion, it introduces noisy and harmful information when the raw data deteriorates. For example, the sensor outputs low-quality camera images when it is under bad conditions, such as bad illumination, occlusion, extreme weather (rainy and snowy), and many other extreme conditions. To overcome this issue, we design an attention mechanism to adaptively attend to reliable features while suppressing harmful ones by computing the relevance of image and point features. As illustrated in Fig. 3 (b), we first feed the point cloud feature F_p and the point-wise image feature $F_i^{(P)}$ into a fully connected layer and map them into the same channel dimension. Then we add them together to form a compact feature representation, which is then compressed into a weight map with a single channel through another fully connected layer. Subsequently, a sigmoid activation function is used to normalize the weight map into the range of $[0, 1]$. Thus, we can obtain the *Image-to-Point* (I2P) attention gate weight w_{i2p} , which indicates the relevance between the image feature and the point feature. w_{i2p} can be calculated as follows:

$$w_{i2p} = \sigma(W_1 \tanh(W_2 F_p + W_3 F_i^{(P)})), \quad (2)$$

where W_1, W_2, W_3 denote the learnable weight matrices in our LI-Fusion layer, and σ represents the sigmoid function.

After obtaining the I2P attention gate weight w_{i2p} , we combine the point cloud feature F_p and the point-wise semantic image feature $F_i^{(P)}$ in a concatenation manner, which can be formularized as follows:

$$F_p' = F_p \parallel w_{i2p} \cdot F_i^{(P)}. \quad (3)$$

Finally, to make the LI-Fusion module more flexible, we compress the channel dimension of the combined feature F_p' to be the same as the input point cloud feature F_p via a fully connected layer and obtain the enhanced point cloud feature F_{ep} .

3.1.4 CB-Fusion Module

The above LI-Fusion module only utilizes the image feature to enhance the semantics information of the point cloud. Although effective, the LI-Fusion module only achieves a unidirectional fusion pathway from image to point cloud,

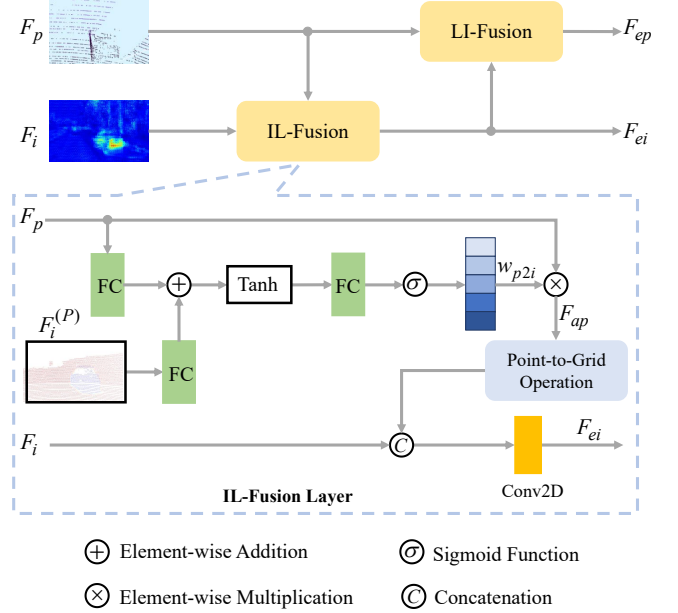


Fig. 4. Illustration of the structure of the CB-Fusion module, which enables a bi-directional fusion pathway and enhances the image features and the point features in sequential order.

which may not sufficiently exploit the complementarity between these two modalities. We argue that the reversed pathway from the point cloud to the image is also valuable. The reason is twofold. First, the geometric information from the LiDAR sensor is beneficial for understanding the object contour information in the camera images. Second, the depth information collected from a LiDAR sensor is more robust to the change of illumination and provides cues for better localizing objects in the camera images. Motivated by these observations, we propose an *Image-guided LiDAR Fusion* (IL-Fusion) module to enable the feature enhancement pathway from LiDAR to image. The combination of IL-Fusion and LI-Fusion forms our novel *Cascaded Bi-directional Multi-Modal Fusion* (CB-Fusion) module. As is shown in Fig. 4, the CB-Fusion module first generates more informative image features with point cloud depth information embedded through IL-Fusion, and then outputs a more robust point cloud feature fused with image semantic information through LI-Fusion. In addition to employing the same point-wise correspondence operation mentioned in the part of the LI-Fusion module to obtain the point-wise

image feature $F_i^{(P)}$, the IL-Fusion module involves an IL-Fusion layer, which is detailed in the following.

IL-Fusion layer. Similar to LI-Fusion, IL-Fusion is also designed to improve the robustness of fusion, when the LiDAR sensor is under bad conditions, such as extreme weather (rainy and snowy), etc. To alleviate this issue, we introduce the *Point-to-Image* (P2I) attention gate weight w_{p2i} to adaptively compute the importance of the point cloud features to the image features. As shown in the bottom of Fig. 4, the P2I attention gate weight w_{p2i} is calculated in a similar mechanism with the I2P attention gate weight w_{i2p} by only swapping the input features of F_p and $F_i^{(P)}$ in the formula (2). Subsequently, we multiply the point cloud feature F_p with w_{p2i} and obtain the point cloud feature with attention enhancement F_{ap} .

To combine the enhanced point feature F_{ap} of shape $N \times D_p$ and the image feature F_i of shape $H \times W \times D_i$, we need to convert them into the same shape. Here, D_p and D_i represent the channel dimension for the features of F_{ap} and F_i , respectively. The point-to-grid operation scatters the point features into the image grid based on the point-wise correspondence and produces the projected point feature with the same shape of F_i . However, the projected location of a point is rarely on the image grid and usually between pixels. Thus, we restore the feature with an integer grid by applying bilinear interpolation to its adjacent point features, leading to a grid-like feature map.

After converting F_{ap} into the grid-like point feature map, we combine it with the image feature F_i in a concatenation manner. Finally, we feed the combined features to a 2D convolution layer with the kernel size of 3 and the stride of 1, leading to the enhanced image feature F_{ei} .

3.2 Refinement Stage

As shown in Fig. 2, we employ the NMS procedure to keep the high-quality proposals predicted by the two-stream RPN and feed them into the refinement network. Similar to PointRCNN [15], we randomly select 512 points inside a proposal as its 3D RoI feature descriptor. For those proposals with less than 512 points, we simply sample the points with replacements. The refinement network includes three SA layers [47] and two detection heads consisting of two cascaded 1×1 convolution layers. Specifically, three SA layers are used to extract a compact global descriptor for each 3D ROI, and two detection heads are employed to classify and regress the final 3D objects.

3.3 Multi-Modal Consistency Loss

The two-stream RPN in EPNet++ generates a corresponding proposal for each foreground point, which is similar to PointRCNN [15]. In this process, how to obtain reliable confidence scores of foreground points is a crucial role in selecting high-quality boxes in the process of RPN. However, the image stream and the geometric stream might produce distinct confidence for the foreground object prediction. Therefore, it indicates that there is some room for improvement by properly using the valuable confidence difference. For example, compared with the predicted image scores from the image stream, the scores from the geometric stream may be more uncertain for recognizing the foreground or

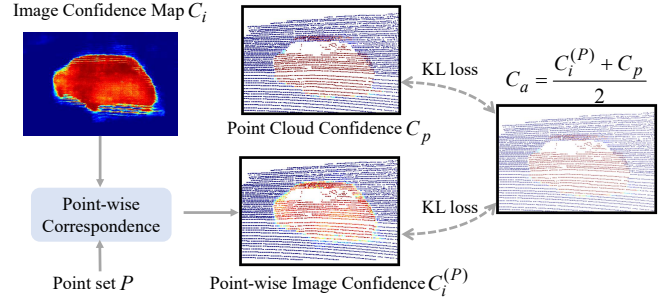


Fig. 5. Illustration of the process of the MC loss.

background in a highly sparse point cloud scene. However, in a denser case, the point cloud scores from the geometric stream are more reliable in dealing with occluded objects and distinguishing the boundary of objects than the camera image scores. Thus, it naturally motivates us to leverage the complementarity of different modalities by aligning the prediction confidence of the two streams.

We propose the *Multi-Modal Consistency* (MC) loss to guarantee the consistency between the confidence scores predicted by the two streams. Concretely, we adopt a simple yet effective strategy, that is, pulling the confidence scores of both streams to their average value. Fig. 5 illustrates the details of our MC loss. Given a certain point from the input point set P , we find its corresponding pixel from the image confidence map C_i through the point-wise correspondence process. C_p and $C_i^{(P)}$ denote the confidence of the LiDAR points and the corresponding pixels. MC loss pulls both C_p and $C_i^{(P)}$ to their average value $C_a = (C_i^{(P)} + C_p) / 2$ through KL loss [59]. The averaged confidence score can avoid the overconfident prediction from one of the modalities. Since when there is a huge gap between the confidence scores of two modalities, this may bring ambiguity in judging whether an object actually exists. Besides, when both C_p and $C_i^{(P)}$ are smaller than a predefined threshold score τ for a specific spatial position, we regard it as a background point, which is not calculated in KL loss. Finally, the MC loss can be formularized as follows:

$$\mathbb{I}_m = \begin{cases} 1 & \text{Max}(C_i^{(P)}, C_p) > \tau \\ 0 & \text{Max}(C_i^{(P)}, C_p) \leq \tau \end{cases} \quad (4)$$

$$L_{i2p} = \mathbb{I}_m \cdot KL(C_i^{(P)} || C_a), \quad (5)$$

$$L_{p2i} = \mathbb{I}_m \cdot KL(C_p || C_a), \quad (6)$$

$$L_{mc} = \frac{1}{N} \sum_{j=0}^N (\lambda_1 L_{i2p}^{(j)} + \lambda_2 L_{p2i}^{(j)}), \quad (7)$$

where \mathbb{I}_m is an indicator function, which outputs ones for foreground positions while zeros for background positions. τ is the score threshold for differentiating foreground and background. L_{i2p} represents the MC loss for pulling the $C_i^{(P)}$ to C_a , while L_{p2i} pulls C_p . N is the total number of input points. λ_1 and λ_2 represent the balancing weights of L_{i2p} and L_{p2i} . More discussions about the effects of λ_1 and λ_2 on the detection performance are in Section 4.4.6.

3.4 Total Loss Function

We utilize a multi-task loss function for jointly optimizing the two-stream RPN and the refinement network. The total loss can be formulated as:

$$L_{total} = L_{rpn} + L_{rcnn}, \quad (8)$$

where L_{rpn} and L_{rcnn} denote the training objectives for the two-stream RPN and the refinement network, both of which adopt a similar optimizing goal, including a classification loss, a regression loss and a *Consistency Enforcing loss* (CE loss) [27]. More concretely, we adopt the focal loss [60] as our classification loss to balance the positive and negative samples with the setting of $\alpha = 0.25$ and $\gamma = 2.0$. For a bounding box, the network needs to regress its center point (x, y, z) , size (l, h, w) , and orientation θ . The CE loss is employed to encourage the consistency of the classification and regression confidence scores, which was first proposed in our preliminary work EPNet [27]. As the effectiveness of CE loss for improving the detection performance has been verified, we adopt the CE loss L_{ce} as the default option in our EPNet++. The details of CE loss will be introduced at the end of this part.

Since the range of the Y-axis (the vertical axis) is relatively small, we directly calculate its offset to the ground truth with a smooth L1 loss [61]. Similarly, the size of the bounding box (h, w, l) is also optimized with a smooth L1 loss. As for the X-axis, the Z-axis and the orientation θ , we adopt a bin-based regression loss [15, 62]. For each foreground point, we split its neighboring area into several bins. The bin-based loss first predicts which bin b_u the center point falls in, and then regresses the residual offset r_u within the bin. In addition to the common loss mentioned above, the two-stream RPN loss includes an image segmentation loss L_{ims} and an MC loss L_{mc} . Finally, we formulate the loss functions in the two-stream RPN stage as follows:

$$L_{rpn} = L_{cls} + L_{reg} + L_{ims} + L_{mc} + \beta L_{ce}, \quad (9)$$

$$L_{cls} = -\alpha(1 - C_p)^\gamma \log C_p, \quad (10)$$

$$L_{ims} = -\alpha(1 - C_i)^\gamma \log C_i, \quad (11)$$

$$L_{reg} = \sum_{u \in x, z, \theta} E(b_u, \hat{b}_u) + \sum_{u \in x, y, z, h, w, l, \theta} S(r_u, \hat{r}_u), \quad (12)$$

where E and S denote the cross-entropy loss and the smooth L1 loss, respectively. β is the balance weight for CE loss. \hat{b}_u and \hat{r}_u denote the ground truth of the bins and the residual offsets. Similarly, L_{rcnn} can be computed by only removing L_{mc} and L_{ims} in the formula (9).

CE loss. In general, we assume that the classification confidence can serve as an agent for the real IoU between the bounding and the ground truth, *i.e.*, the localization confidence. However, the classification confidence and the localization confidence are often inconsistent, leading to sub-optimal performance. Thus, we use a CE loss to ensure the consistency between the localization and classification confidence so that boxes with high localization confidence possess high classification confidence and vice versa. The CE loss can be written as follows:

$$L_{ce} = -\log(C_c \cdot \frac{Area(B_p \cap B_g)}{Area(B_p \cup B_g)}) \quad (13)$$

where B_p and B_g represent the predicted bounding box and the ground truth. $Area(\star)$ means to calculate the area of \star . And C_c denotes the classification confidence. In particular, C_c is equal to C_a when MC loss is enabled, otherwise $C_c = C_p$. Towards optimizing this loss function, the classification confidence and localization confidence (*i.e.*, the IoU) are encouraged to be as high as possible jointly. Hence, boxes with large overlaps will possess high classification possibilities and be kept in the NMS procedure.

4 EXPERIMENTS

4.1 Experimental Datasets and Evaluation Metrics

KITTI Dataset [39] is a popular benchmark dataset for autonomous driving, which is collected by a 64-beam LiDAR sensor and two camera sensors. The dataset includes 7,481 training frames and 7,518 testing frames for 3D object detection. Following the same dataset split protocol as [15, 62], the 7,481 frames are further split into 3,712 frames for training and 3,769 frames for validation. In addition, for the image segmentation task, we obtain the mask annotations of Cars, Pedestrians and Cyclists from the KINS dataset [85]. In our experiments, we provide the results on both the validation and the testing set for three difficulty levels, *i.e.*, *Easy*, *Moderate*, and *Hard*. Objects are categorized into different difficulty levels according to their sizes, occlusion, and truncation. Recently, the KITTI dataset adopts a better evaluation protocol [86] which computes the mean Average Precision (mAP) using 40 recall positions instead of 11 as before. We compare our methods with state-of-the-art methods under this new evaluation protocol.

JRDB Dataset [41] is a large-scale multi-modal dataset collected from a social mobile manipulator JackRabbit. The dataset is designed for facilitating perceptual tasks necessary for a robot to understand a scene and human behavior. The dataset provides stereo cylindrical 360° RGB video streams, continuous 3D point clouds scanned from two 16-channel Velodyne LiDARs, and 360° spherical image from a fisheye camera. There are 54 sequences collected in both indoor and outdoor environments. The sequences are divided into 27 training sequences and 27 testing sequences for the 3D detection task. Following the official suggestion, we select 7 out of 27 training sequences as the validation split. JRDB provides both the 2D and 3D bounding box annotations for Pedestrians in the full 360° scenes, different from the KITTI dataset which annotates the front view scenes. Note that the 2D dense segmentation annotations for our image stream are not available. However, we can obtain sparse segmentation masks. Specifically, we consider the 2D projection pixels as foreground if the corresponding LiDAR points are inside the 3D bounding boxes, otherwise as background. The official evaluation metrics are similar to KITTI, while the 3D IoU threshold for Pedestrians is set to 0.3 instead of 0.5 in KITTI.

SUN-RGBD Dataset [40] is an indoor benchmark dataset for 3D object detection. The dataset is composed of 10,335 images with 700 annotated object categories, including 5,285 images for training and 5,050 images for testing. We report results on the testing set for ten main object categories following previous works [24, 35] since the number of these object categories on the whole dataset is relatively

TABLE 1

Quantitative comparisons with the state-of-the-art 3D object detection methods on the KITTI test benchmark. For *Modality* column, *P* and *I* denote the point cloud and the camera image, respectively. The same denotations are used in the following Table 2, Table 3 and Table 4.

Method	Conference	Modality	Cars(Recall40)			Pedestrians(Recall40)			Cyclists(Recall40)			
			Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard	
MonoGRNet [63]	TPAMI 2021	I	9.61	5.74	4.25	-	-	-	-	-	-	
M3D-RPN [64]	ICCV 2019		14.76	9.71	7.42	4.92	3.48	2.94	0.94	0.65	0.47	
MonoPair [65]	CVPR 2020		13.04	9.99	8.65	10.02	6.68	5.53	3.79	2.12	1.83	
RTM3D [66]	ECCV 2020		14.41	10.34	8.77	-	-	-	-	-	-	
PatchNet [67]	ECCV 2020		15.68	11.12	10.17	-	-	-	-	-	-	
CaDDN [68]	CVPR 2021		19.17	13.41	11.46	12.87	8.14	6.76	7.00	3.41	3.30	
GUP Net [69]	ICCV 2021		20.11	14.20	11.77	14.72	9.53	7.87	4.18	2.65	2.09	
DD3d [70]	ICCV 2021		23.22	16.34	14.20	-	-	-	-	-	-	
PointPillars [16]	CVPR 2019		P	82.58	74.31	68.99	51.45	41.92	38.89	77.10	58.65	51.92
PointRCNN [15]	CVPR 2019			86.96	75.64	70.70	47.98	39.37	36.01	74.96	58.82	52.53
TANet [71]	AAAI 2020	84.39		75.94	68.82	53.72	44.34	40.49	75.70	59.44	52.53	
Pointformer [38]	CVPR 2021	87.13		77.06	69.25	50.67	42.43	39.60	75.01	59.80	53.99	
Fast PointRCNN [72]	ICCV 2019	85.29		77.40	70.24	-	-	-	-	-	-	
HotSpotNet [73]	ECCV 2020	87.60		78.31	73.34	53.10	45.37	41.47	82.59	65.95	59.00	
PartA ² [44]	TPAMI 2020	87.81		78.49	73.51	53.10	43.35	40.06	79.17	63.52	56.93	
SERCNN [74]	CVPR 2020	87.74		78.96	74.30	-	-	-	-	-	-	
SECOND [36]	Sensors 2018	87.44		79.46	73.97	-	-	-	-	-	-	
Point-GNN [75]	CVPR 2020	88.33		79.47	72.29	51.92	43.77	40.14	78.60	63.48	57.08	
MGAFF-3DSSD [76]	ACM MM 2021	88.16		79.68	72.39	-	-	-	-	-	-	
3DSSD [48]	CVPR 2020	88.36		79.57	74.55	54.64	44.27	40.23	82.48	64.10	56.90	
STD [48]	ICCV 2019	87.95		79.71	75.09	53.29	42.47	38.35	78.69	61.59	55.30	
SA-SSD [77]	CVPR 2020	88.75		79.79	74.16	-	-	-	-	-	-	
CIA-SSD [78]	AAAI 2021	89.59		80.28	72.87	-	-	-	-	-	-	
PV-RCNN [43]	CVPR 2020	90.25		81.43	76.82	52.17	43.29	40.29	78.60	63.71	57.65	
FromVoxelToPoint [79]	ACM 2021	88.53		81.58	77.37	-	-	-	-	-	-	
BANet [80]	Arxiv 2021	89.28		81.61	76.58	-	-	-	-	-	-	
Voxel R-CNN [37]	AAAI 2021	90.90		81.62	77.06	-	-	-	-	-	-	
SIENet [81]	Arxiv 2021	88.22		81.71	77.22	-	-	-	-	-	-	
M3DETR [21]	WACV 2022	90.28		81.73	76.96	45.70	39.94	37.66	83.83	66.74	59.03	
MV3D [12]	CVPR 2017	P + I		74.97	63.63	54.00	-	-	-	-	-	-
Confuse [25]	ECCV 2018			83.68	68.78	61.67	-	-	-	-	-	-
F-Pointnet [62]	CVPR 2018			82.19	69.79	60.59	50.53	42.15	38.08	72.27	56.12	49.01
PointPainting [26]	CVPR 2020			82.11	71.70	67.08	50.32	40.97	37.87	77.63	63.78	55.89
AVOD-FPN [50]	IROS 2018			83.07	71.76	65.73	50.46	42.27	39.04	63.76	50.55	44.93
PI-RCNN [28]	AAAI 2020			84.37	74.82	70.03	-	-	-	-	-	-
F-Convnet [82]	IROS 2019			87.36	76.39	66.69	52.16	43.38	38.80	81.98	65.07	56.54
MMF[34]	CVPR 2019			88.40	77.43	70.22	-	-	-	-	-	-
CLOCs_SecCas [83]	IROS 2020			86.38	78.45	72.45	-	-	-	-	-	-
MVAF [84]	Arxiv 2020		87.87	78.71	75.48	-	-	-	-	-	-	
3D-CVF [23]	ECCV 2020		89.20	80.05	73.11	-	-	-	-	-	-	
CLOCs_PVCas [83]	IROS 2020		88.94	80.67	77.15	-	-	-	-	-	-	
EPNet [27]	ECCV 2020		89.81	79.28	74.59	-	-	-	-	-	-	
EPNet++ (Ours)	-		91.37	81.96	76.71	52.79	44.38	41.29	76.15	59.71	53.67	

enormous. Following [40], we adopt mAP as the evaluation metric to measure the 3D object detection performance by setting the value of the 3D IoU threshold as 0.25.

4.2 Implementation Details

On the large-range outdoor scenes datasets of JRDB and KITTI, we adopt the identical network architecture for EPNet++, as is shown in Fig. 2. For the indoor SUN-RGBD dataset, we integrate the proposed CB-Fusion and LI-Fusion modules into different backbones, *e.g.*, VoteNet [35] and ImVoteNet [24], to verify the generalization capability of our method. We carefully follow their settings on SUN-RGBD referring to the original papers. In the following, we mainly focus on the technical details of our EPNet++.

Network Settings. EPNet++ takes both the LiDAR point cloud and the camera image as inputs. For each 3D scene on the KITTI dataset, the range of the LiDAR point cloud is set to [-40, 40], [-1, 3], [0, 70.4] meters along the X (right), Y (down), Z (forward) axis in the camera coordinate, respectively. On the JRDB dataset, we only limit the horizontal range of the point cloud within a radius of 26 meters, without restricting the vertical direction. The heading orientation θ is in the range of $[-\pi, \pi]$ on both datasets. We subsample

16,384 and 32,768 points from the raw LiDAR point cloud as the input of the geometric stream for the KITTI dataset and the JRDB dataset, respectively. The image stream takes images with a resolution of 1280×384 as input for the KITTI dataset. While on the JRDB dataset, we scale the original image to a smaller size of 1888×240 for saving GPU memory. Four set abstraction layers are employed to downsample the input LiDAR point cloud to the resolution of 4096 (8192), 1024 (2048), 256 (512), and 64 (128) on the KITTI (JRDB) dataset, respectively. Four feature propagation layers are used to gradually recover the original resolution of the point cloud for the following foreground segmentation and 3D proposal generation. Similar to the geometric stream, we use four convolution blocks with the same stride of 2 to downsample the input image. Besides, we employ four parallel transposed convolutions with the strides of 2, 4, 8, and 16 to recover the original image resolution from feature maps of different scales. In the NMS process, we select the top 8000 boxes generated by the two-stream RPN according to the comprehensive confidence score C_a . After that, we filter redundant boxes with the IoU threshold of 0.8 and obtain 100 positive candidate boxes, which will be refined by the refinement network.

TABLE 2

Quantitative comparisons with SOTA methods for 3D object detection task on the SUN-RGBD validation set. * means the reproduced results.

Method	Conference	Modality	bathtub	bed	bookshelf	chair	desk	dresser	nightstand	sofa	table	toilet	mAP
MLCVNet [87]	CVPR 2020	P	79.2	85.8	31.9	75.8	26.5	31.3	61.5	66.3	50.4	89.1	59.8
H3DNet [88]	ECCV 2020	P	73.8	85.6	31.0	76.7	29.6	33.4	65.5	66.5	50.8	88.2	60.1
HGNet [89]	CVPR 2020	P	78.0	84.5	35.7	75.2	34.3	37.6	61.7	65.7	51.6	91.1	61.6
Pointformer [38]	CVPR 2021	P	80.1	84.3	32.0	76.2	27.0	37.4	64.0	64.9	51.5	92.2	61.1
Group-Free-3D [90]	ICCV 2021	P	80.0	87.8	32.5	79.4	32.6	36.0	66.7	70.0	53.8	91.1	63.0
DSS [91]	CVPR 2016	P + I	44.2	78.8	11.9	61.2	20.5	6.4	15.4	53.5	50.3	78.9	42.1
2d-driven [92]	ICCV 2017	P + I	43.5	64.5	31.4	48.3	27.9	25.9	41.9	50.4	37.0	80.4	45.1
COG [93]	CVPR 2016	P + I	58.3	63.7	31.8	62.2	45.2	15.5	27.4	51.0	51.3	70.1	47.6
PointFusion [32]	CVPR 2018	P + I	37.3	68.6	37.7	55.1	17.2	24.0	32.3	53.8	31.0	83.8	44.1
F-PointNet [62]	CVPR 2018	P + I	43.3	81.1	33.3	64.2	24.7	32.0	58.1	61.1	51.1	90.9	54.0
EPNet [27]	ECCV 2020	P + I	75.4	85.2	35.4	75.0	26.1	31.3	62.0	67.2	52.1	88.2	59.8
VoteNet [35]	ICCV 2019	P	74.4	83.0	28.8	75.3	22.0	29.8	62.2	64.0	47.3	90.1	57.7
VoteNet* [35]	ICCV 2019	P	75.5	85.6	31.9	77.4	24.8	27.9	58.6	67.4	51.1	90.5	59.1
+LI-Fusion (EPNet)	ECCV 2020	P + I	73.8	86.4	35.6	80.1	27.4	31.8	63.0	67.3	54.9	88.8	60.9
+CB-Fusion (Ours)	-	P + I	80.3	84.7	33.1	79.7	27.2	36.2	63.1	69.9	53.5	86.9	61.5
ImVoteNet [24]	CVPR 2020	P + I	75.9	87.6	41.3	76.7	28.7	41.4	69.9	70.7	51.1	90.5	63.4
ImVoteNet* [24]	CVPR 2020	P + I	77.3	88.2	41.4	79.5	30.8	39.9	68.0	72.4	51.5	91.5	64.0
+LI-Fusion (EPNet)	ECCV 2020	P + I	72.8	88.4	47.6	80.9	32.3	42.5	64.5	70.8	53.6	93.1	64.6
+CB-Fusion (Ours)	-	P + I	76.3	89.1	47.1	80.2	32.5	45.2	67.4	71.9	51.3	92.4	65.3

Training Scheme. Our two-stream RPN and refinement network are jointly optimized in an end-to-end manner. The regression loss L_{reg} and the CE loss L_{ce} are only applied to positive candidates, *i.e.*, proposals generated by foreground points in the RPN stage, and boxes sharing the IoU larger than 0.55 with the ground truth in the RCNN stage. The balancing weight β for the CE loss is set to 5. For the MC loss L_{mc} , we simply set the default values of λ_1 and λ_2 to 0.5 and 0.5. The score threshold τ in the indicator function is set to 0.2. We optimize the model by Adaptive Moment Estimation (Adam) [94] with the initial learning rate, weight decay, and momentum factor set to 0.002, 0.001, and 0.9, respectively. We train the model with a batch size of 8 for around 50 epochs on the KITTI dataset and 20 epochs on the JRDB dataset. All the experiments are conducted on a machine with four Titan V GPUs using the popular deep learning framework Pytorch [95].

Data Augmentation. Three common data augmentation strategies are adopted to prevent over-fitting, including rotation, flipping, and scale transformations. First, we randomly rotate the point cloud along the vertical axis within the range of $[-\pi/18, \pi/18]$. Then, the point cloud is randomly flipped along the forward axis. Besides, each ground truth box is randomly scaled following the uniform distribution of $[0.95, 1.05]$. It is noted that many LiDAR-based methods utilize the GT Sampling data augmentation [36] to improve the diversity of objects, which samples ground truth boxes from the whole dataset and places them into the raw 3D frames to simulate real scenes with crowded objects. Although effective, this data augmentation needs the prior information of road planes, which is usually difficult to acquire for kinds of real scenes. In addition, the strategy may lead to the misalignment of the camera image and point cloud, which is not so suitable for multi-modal fusion methods. Hence, we do not utilize this augmentation mechanism in our framework for applicability and generality.

4.3 Comparisons with State-of-the-art Methods

4.3.1 Evaluation on KITTI Dataset

Table 1 presents the quantitative comparison with state-of-the-art 3D object detection methods on the KITTI test

benchmark. As is shown, EPNet [27] and EPNet++ outperform our LiDAR-based baseline PointRCNN by a large margin, yielding an improvement of 3.64% and 6.32% on the moderate Cars, respectively. These results demonstrate the effectiveness of the proposed fusion mechanism. Besides, EPNet++ achieves better or comparable performance over the state-of-the-art methods on Cars and Pedestrians for all difficulty levels. For Cyclists, we find that their proportion of instances on the training KITTI dataset is about 4.67%, which is much smaller than that of Pedestrians (12.87%) and Cars (82.46%). In this case, constructing more instances with the GT Sampling strategy can effectively improve the detection performance for Cyclists. Hence, our method achieves relatively lower detection performance on Cyclists compared to a few methods [43, 48, 49] using GT Sampling. However, EPNet++ achieves consistent performance improvement over the baseline PointRCNN on all difficulty levels for Cyclists, further demonstrating the effectiveness of our approach.

4.3.2 Evaluation on JRDB Dataset

In Table 3, we compare our EPNet++ with the available published methods F-PointNet [62] and TANet [71] provided by the JRDB 3D detection benchmark. We also present our baseline configuration named EPNet (LiDAR-only) and EPNet [27]. F-PointNet achieves a lower detection performance than other detection methods. The main reason is that F-PointNet predicts 3D boxes from point cloud frustums, which are highly dependent on the performance of the 2D detector. Compared with the LiDAR-only method TANet [71], our fusion-based methods EPNet++ and EPNet respectively obtain performance improvement with the mAP of 11.69% and 8.65%, which illustrates the superiority of our fusion manner. Besides, EPNet++ outperforms the EPNet and EPNet (LiDAR-only) methods with the mAP of 3.04% and 7.38%, respectively, which further verifies the effectiveness of our method.

4.3.3 Evaluation on SUN-RGBD Dataset

To further verify the effectiveness and generalization capability of our methods, we integrate the LI-Fusion module and the CB-fusion module into another two rep-

TABLE 3

Quantitative comparisons on JRDB test set for 3D object detection. Ped. is short for Pedestrians. EPNet (LiDAR-only) removes the image stream and LI-Fusion module of the original EPNet [27].

Method	Conference	Class	Modality	mAP
F-PointNet [62]	CVPR 2018	Ped.	P + I	38.21
TANet [71]	AAAI 2020	Ped.	P	54.94
EPNet (LiDAR-only) [27]	ECCV 2020	Ped.	P	59.25
EPNet [27]	ECCV 2020	Ped.	P + I	63.59
EPNet++ (Ours)	-	Ped.	P + I	66.63

representative 3D object detectors, including LiDAR-based VoteNet [35] and fusion-based ImVoteNet [24]. We implement our method based on the open-source mmdetection3d codebase¹. As is shown in Table 2, LI-Fusion and CB-Fusion modules bring an obvious improvement of 1.8% and 2.4% mAP, when using VoteNet as the backbone. ImVoteNet [24] boosts the detection performance of VoteNet [35] via introducing geometric cues, semantic cues, and texture cues from the 2D camera images to point cloud features. Although effective, ImVoteNet [24] simply combines these image features with point cloud features in a concatenation manner at the end of the network, without exploring the complementarity between two modalities at a deep feature level. Adding our LI-Fusion and CB-fusion modules into ImVoteNet [24] leads to an improvement of mAP 0.6% and 1.3%, respectively, which indicates that our fusion mechanism produces more informative feature representations. Besides, CB-Fusion leads to consistent improvements over LI-Fusion, demonstrating the effectiveness of the bi-directional fusion strategy. It is worth noting that ImVoteNet [24] with CB-Fusion achieves a new state-of-the-art result on the SUN-RGBD dataset and outperforms all published point-based and fusion-based methods, which clearly illustrates the superiority of the CB-Fusion module.

4.4 Ablation Studies

In this section, we conduct all the experiments on the KITTI validation dataset. We provide results under different LiDAR beam settings (64 beams, 16 beams and 8 beams) to investigate the generalization capability of our model to sparse scenes. To simulate real-world sparse scenes, we generate 8-beam and 16-beam LiDAR points following the open-source code² in Pseudo-Lidar-V2 [96]. Fig. 6 illustrates a visualization result of generated 16-beam and 8-beam LiDAR points using the algorithm above. In the following, we take the 64-beam LiDAR points as input, unless otherwise stated. Since CE loss is well verified in our preliminary work EPNet [27], we add it to the optimization goal by default and do not discuss the effect of CE loss here. For more details about CE loss, please refer to EPNet [27].

4.4.1 Comparisons under Different Point Cloud Densities

To comprehensively evaluate the detection performance under dense and sparse scenes, we compare our method with several state-of-the-art detectors under three different LiDAR beam settings. Specifically, we select two LiDAR-based methods (Voxel RCNN [37]³, PV-RCNN [43]³), and

1. <https://github.com/open-mmlab/mmdetection3d>
 2. https://github.com/mileyan/Pseudo_Lidar_V2
 3. <https://github.com/open-mmlab/OpenPCDet>

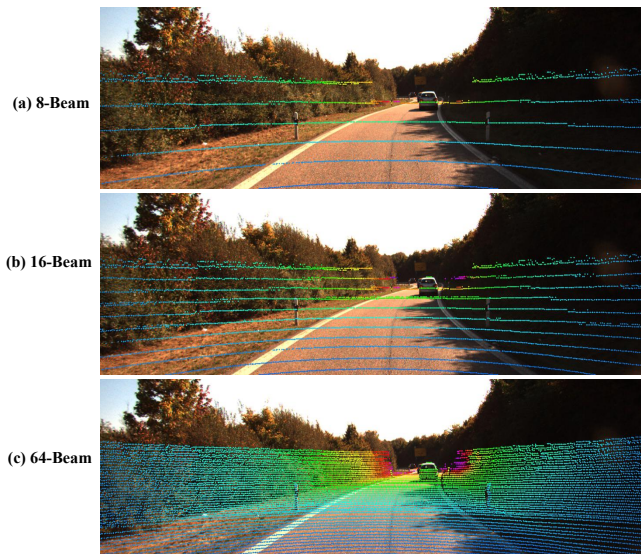


Fig. 6. Visualization of different beam LiDAR points projected on the image plane.

two fusion-based methods (CLOCs [83]⁴, 3D-CVF [23]⁵). Since the GT Sampling data augmentation usually effectively boosts the performance of LiDAR-based detectors while not applicable for fusion-based methods, we provide the results of LiDAR-based detectors with and without GT Sampling for fair comparison in Table 4.

At the top of Table 4, we first compare EPNet++ with state-of-the-art LiDAR-based methods which leverage the powerful GT Sampling augmentation. Although EPNet++ produces worse performance on Cars and Cyclists in dense 64-beam LiDAR setting, our method demonstrates clear superiority on Cars for sparse 8-beam LiDAR, outperforming Voxel R-CNN and PV-RCNN on the moderate difficulty level by 3.42% and 5.78% in terms of mAP. Besides, the performance gap on the Cyclists is also reduced. When GT Sampling is not employed, EPNet++ consistently outperforms previous methods for all the different settings of categories, difficulty levels and the number of beams, as is shown at the bottom of Table 4. Specifically, for sparse scenes with 8-beam LiDAR points, EPNet++ yields reliable improvements with mAP of 2.13% and 8.44% over the fusion-based methods CLOCs and 3D-CVF on Cars of moderate difficulty level, respectively. For more challenging Pedestrians in the case of 8-beam LiDAR points as input, EPNet++ attains the obvious gains with the mAP of 10.21%, 16.41% and 9.63% over these LiDAR-based methods Voxel-RCNN, PV-RCNN and EPNet (LiDAR-only) without GT Sampling on the moderate difficulty level. Moreover, EPNet++ achieves consistent and remarkable improvements over EPNet, which reveals the effectiveness of our new design for multi-modal 3D object detection tasks.

4.4.2 Ablation Studies on Different Components

In this part, we validate the effectiveness of core components of our EPNet++, including LI-Fusion with/without

4. <https://github.com/pangsu0613/CLOCs>
 5. <https://github.com/rasd3/3D-CVF>

TABLE 4

Comparisons with state-of-the-art approaches under different point cloud densities (8-beam, 16-beam and 64-beam) on the KITTI validation dataset. P and I are short for point cloud and camera image.

Method	GT Sampling	Input	LiDAR	Cars			Pedestrians			Cyclists		
				Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Voxel RCNN [37]	✓	P	8-beam	69.10	51.69	46.44	40.71	35.21	30.48	38.48	22.96	21.66
			16-beam	86.55	68.70	63.93	57.07	51.26	46.28	66.87	43.34	40.46
			64-beam	92.57	84.49	82.35	68.63	60.95	55.67	93.16	76.52	72.57
PV RCNN [43]		P	8-beam	68.30	49.33	44.03	42.30	37.13	32.40	40.29	23.61	22.64
			16-beam	85.60	66.96	62.78	59.92	52.79	48.27	68.64	45.67	42.94
			64-beam	92.35	84.50	82.47	66.39	60.34	55.70	93.71	73.99	69.53
Voxel RCNN [37]	×	P	8-beam	65.30	48.03	43.18	42.03	35.58	31.74	25.60	14.39	13.24
			16-beam	85.18	66.35	61.54	58.64	53.34	47.40	55.03	33.83	31.93
			64-beam	92.04	82.64	80.23	70.94	64.55	58.35	85.12	61.53	58.93
PV RCNN [43]		P	8-beam	62.55	46.77	41.08	34.53	29.38	26.11	22.95	12.80	12.52
			16-beam	83.73	65.36	60.80	54.32	46.74	42.68	47.37	31.71	29.97
			64-beam	91.85	82.68	80.20	67.49	62.21	57.91	80.04	58.72	56.36
CLOCs [83]		P+I	8-beam	67.35	52.98	46.02	-	-	-	-	-	-
			16-beam	85.41	67.26	62.02	-	-	-	-	-	-
			64-beam	92.48	82.79	77.71	-	-	-	-	-	-
3D-CVF [23]		P+I	8-beam	61.95	46.67	40.50	-	-	-	-	-	-
			16-beam	81.82	63.12	57.19	-	-	-	-	-	-
			64-beam	91.97	82.87	80.36	-	-	-	-	-	-
EPNet (LiDAR)	P	8-beam	64.55	47.10	42.32	41.81	36.16	31.15	33.07	18.88	17.99	
		16-beam	84.47	63.30	59.59	58.47	52.56	45.73	55.94	33.80	31.59	
		64-beam	90.87	81.15	79.59	70.26	61.30	54.16	84.49	61.50	57.86	
EPNet [27]	P+I	8-beam	69.11	53.90	49.06	46.65	40.62	35.36	35.41	20.73	19.13	
		16-beam	85.11	67.72	63.32	61.08	54.16	47.71	58.01	34.53	32.47	
		64-beam	92.28	82.59	80.14	72.56	63.11	56.32	84.88	62.43	58.65	
EPNet++ (Ours)	P+I	8-beam	69.20	55.11	50.29	51.31	45.79	39.78	41.34	23.31	21.68	
		16-beam	87.08	69.12	64.90	66.77	59.86	53.07	61.37	36.93	34.95	
		64-beam	92.51	83.17	82.27	73.77	65.42	59.13	86.23	63.82	60.02	

TABLE 5

Ablation experiments on the effects of different components of the proposed EPNet++ on 16-beam and 64-beam LiDAR points. LI (CB) wo att represents the LI-Fusion (CB-Fusion) without attention mechanism. MC is the MC loss.

LiDAR	LI wo Att	LI	CB wo Att	CB	MC	Cars			Pedestrians			Cyclists		
						Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
16-beam	-	-	-	-	-	84.47	63.30	59.59	58.47	52.56	45.73	55.94	33.80	31.59
16-beam	✓	-	-	-	-	84.60	67.63	63.20	59.47	53.26	46.87	57.45	33.93	31.96
16-beam	-	✓	-	-	-	85.11	67.72	63.32	61.08	54.16	47.71	58.01	34.53	32.47
16-beam	-	-	✓	-	-	86.10	68.86	63.84	61.86	54.24	47.56	58.12	35.35	32.99
16-beam	-	-	-	✓	-	85.69	68.83	64.22	65.73	57.69	51.13	58.72	37.25	34.47
16-beam	-	-	-	-	✓	84.50	63.69	60.22	61.95	55.90	49.02	60.71	35.96	33.61
16-beam	-	-	-	✓	✓	87.08	69.12	64.90	66.77	59.86	53.07	61.37	36.93	34.95
64-beam	-	-	-	-	-	90.99	81.72	79.75	70.26	61.30	54.16	84.49	61.50	57.86
64-beam	✓	-	-	-	-	91.92	82.42	80.16	71.02	61.56	54.49	84.79	62.08	58.24
64-beam	-	✓	-	-	-	92.28	82.59	80.14	72.56	63.11	56.32	84.88	62.43	58.65
64-beam	-	-	✓	-	-	92.45	82.73	80.44	72.49	61.96	54.52	84.69	62.16	58.19
64-beam	-	-	-	✓	-	92.54	82.93	80.56	73.08	64.03	57.65	85.57	62.92	58.70
64-beam	-	-	-	-	✓	91.69	82.12	80.19	70.71	62.76	55.74	85.46	62.34	58.25
64-beam	-	-	-	✓	✓	92.51	83.17	82.27	73.77	65.42	59.13	86.23	63.82	60.02

attention, CB-Fusion with/without attention and MC loss. We build our baseline by removing the image stream and these components from EPNet++ for both 64-beam and 16-beam LiDAR points. As is shown in Table 5, CB-Fusion (or LI-Fusion) without attention brings general and consistent improvements over the baseline on Cars, Pedestrians and Cyclists under all the difficulty settings on different beam LiDAR points, which demonstrates the effectiveness of the cascade bi-directional feature enhancement pathway. Then, introducing attention operations into CB-Fusion (or LI-Fusion) brings a remarkable gain with mAP of 3.45% (or 0.90%) on Pedestrians under the case of 16-beam LiDAR points, which reveals the importance of enhancing reliable features and suppressing harmful ones by computing the relevance of image and point features. Moreover, the CB-Fusion with attention module yields an improvement of 1.11%, 3.53% and 2.72% over the LI-Fusion module in terms of mAP on Cars, Pedestrians and Cyclists for the moderate difficulty level under the case of 16-beam LiDAR points,

which further illustrates the superiority of adopting feature interaction in a bi-directional manner. Therefore, CB-Fusion is the most indispensable component in EPNet++ for promising performance especially under the sparse setting. Beyond the feature fusion, MC loss leads to consistent improvements across all the categories and difficulty levels on different beam LiDAR points. The reason is that MC loss effectively aligns the confidence scores from two streams and the aligned score is a better criterion for evaluating the quality of 3D proposals in the RPN stage. Combining CB-Fusion with MC loss, EPNet++ achieves significant gains with mAP of 5.82%, 7.30%, 3.13% on Cars, Pedestrians and Cyclists at the moderate difficulty level over the baseline for 16-beam LiDAR points. It is worth mentioning that MC loss is easy to implement without introducing extra computation cost in the inference stage, which is friendly to a more computing-sensitive application.

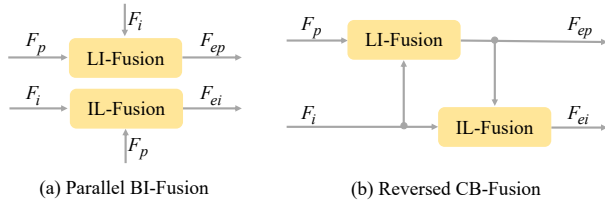


Fig. 7. The structure of two alternative fusion mechanisms: the *parallel bi-directional interaction fusion* (Parallel BI-Fusion) in (a) and the *reversed CB-Fusion* in (b).

TABLE 6

Comparisons with different bi-directional fusion mechanisms. For each category, we report the average value of mAP under three difficulty levels for 64-beam LiDAR points.

Methods	Cars	Pedestrians	Cyclists	mAP
EPNet++ w/o CB-Fusion	84.67	63.07	68.68	72.14
Parallel BI-Fusion	85.49	64.47	69.78	73.25
Reversed CB-Fusion	85.41	65.07	69.69	73.39
CB-Fusion	85.98	66.11	70.02	74.04

TABLE 7

Comparisons with different fusion methods. For each category, we report the average value of mAP under three difficulty levels for 64-beam LiDAR points.

Methods	Cars	Pedestrians	Cyclists	mAP
LiDAR-only EPNet	83.87	61.91	67.95	71.24
PointPainting [26]	84.48	63.37	68.58	72.14
Pointaugmenting [30]	84.13	63.60	69.12	72.28
DenseFusion [56]	84.94	63.35	68.80	72.36
Transformer-based CB-Fusion	84.52	64.54	68.43	72.50
CB-Fusion	85.34	64.92	69.07	73.11
Pointaugmenting+CB-Fusion	84.56	66.30	71.95	74.27

4.4.3 Different Fusion Paradigms

Different Bi-Directional Fusion Mechanisms. In Table 6, we provide two alternative CB-Fusion methods, namely, a parallel BI-Fusion shown in Fig. 7(a) and a reversed CB-Fusion shown in Fig. 7(b). EPNet++ without the CB-Fusion module is regarded as our baseline model. The core difference between parallel BI-Fusion and (reversed) CB-Fusion is whether to use the enhanced features for the attention operation. Parallel BI-Fusion conducts two attention operations parallelly using the original image/point features. However, CB-Fusion and reversed CB-Fusion work in a cascaded manner, and the second attention operation takes as input the enhanced feature (i.e., the output of the first attention operation). Intuitively, the enhanced feature, which has already combined image semantic information with geometric information, will be more beneficial. Experimental results in Table 6 demonstrate the clear superiority of using the enhanced feature. CB-Fusion (resp. reversed CB-Fusion) outperforms parallel BI-Fusion by 0.79% (resp. 0.14%) in terms of mAP. Reversed CB-Fusion differs from CB-Fusion in the fusion order. Since the geometric stream (point feature) is finally applied for predicting 3d boxes, CB-Fusion is more reasonable and leads to better performance than reversed CB-Fusion (74.04% v.s. 73.39% in terms of mAP).

Comparisons with different fusion methods. In Table 7, we provide the quantitative comparisons for CB-Fusion, transformer-based CB-Fusion, PointPainting-based methods [26, 30] and DenseFusion [56]. The baseline model is LiDAR-only EPNet for a fair comparison. First, we in-

tegrate the popular PointPainting-based fusion methods PointPainting [26] and Pointaugmenting [30] into our baseline. More concretely, we first use Mask-RCNN [55] to train a strong segmentation network. For PointPainting [26], we concatenate the output segmentation scores and input LiDAR points to achieve multi-modal fusion. For Pointaugmenting [30], the input LiDAR points are decorated with the corresponding image features from the first output layer after the backbone of ResNet-50 [97] from Mask-RCNN in a concatenation manner. As shown in Table 7, our CB-Fusion outperforms these methods with input fusion on average. Besides, integrating CB-Fusion into the input-feature fusion PointAugmenting produces a performance gain of 1.99% mAP on average, which further demonstrates the effectiveness of this bidirectional feature interaction between LiDAR points and images. Then, we conduct the pixel-wise dense fusion similar to DenseFusion[56] at object levels in the RCNN stage of EPNet++ by combining 3D RoI features and the corresponding projected 2D boxes feature in a concatenation manner. As shown in Table 7, DenseFusion brings a reliable improvement of 1.12% mAP (72.36 v.s. 71.24) on average over the baseline, but is still inferior to CB-Fusion (73.11%). Finally, we provide the transformer-based CB-Fusion through a global cross-attention operation (refer to Section 4.4.7 for details) instead of attention operation in LI-Fusion and IL-Fusion, which even produces a slight performance drop. This reveals effectively utilizing the one-to-one correspondence between two modalities is crucial in computing attention weights.

4.4.4 Effect of the Attention Gate Weights

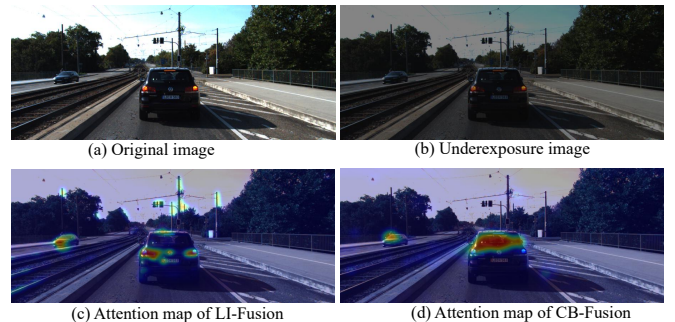


Fig. 8. Visualization of the attention map for LI-Fusion and CB-Fusion, respectively.

TABLE 8

Quantitative analysis of the effect of I2P attention gate w_{i2p} and P2I attention gate w_{p2i} . For each category, we report the average value of mAP under three difficulty levels.

Method	w_{i2p}	w_{p2i}	Cars	Pedestrians	Cyclist	mAP
Baseline	-	-	83.00	58.28	62.39	67.89
LI-Fusion	-	-	82.63	57.73	62.95	67.77
LI-Fusion	✓	-	83.03	59.19	66.66	69.63
CB-Fusion	-	-	83.50	57.17	65.92	68.86
CB-Fusion	✓	✓	83.83	60.98	67.89	70.90

In a real scene, the camera image usually suffers from underexposure and overexposure due to the change of illumination. Furthermore, the point cloud collected by a LiDAR sensor is easily disturbed by severe weather (e.g.,

fog, rain) conditions, resulting in noise points. It poses challenges for 3D detectors to produce accurate predictions under these scenes. We argue that the attention gate weights employed in our fusion module can effectively alleviate the issue of noisy images and point clouds. To investigate the effect of the attention gate weights, we simulate the real noising scenes by changing the illumination of the camera image and adding inference points into the raw LiDAR point cloud. Specifically, for each image in the KITTI dataset, we simulate the illumination variance through the transformation $y = a * x + b$, where x and y denote the original and transformed RGB value for a pixel. The coefficient a is randomly sampled from a uniform distribution in the range of $[0.5, 1.5]$. The offset b is set as a fixed value of 5. For a raw point cloud on the KITTI dataset, we generate 100 noise points around each GT 3D object following TANet [71].

The experimental results are presented in Table 8. We first present a LiDAR-based baseline result by removing the image stream of EPNet, which produces 67.89% mAP averaged across Cars, Pedestrians and Cyclists. Then we introduce the LI-Fusion module without I2P attention gate weight into the baseline, leading to slight performance degradation of 0.12% on average, which indicates that combining noisy sensor data is harmful to 3D object detection. In contrast, with the guidance of the I2P attention gate weight w_{i2p} , the LI-Fusion module yields a gain of 1.74% mAP over the baseline. Similarly, integrating the attention gate weights w_{i2p} and w_{p2i} into the CB-Fusion module brings a remarkable performance improvement of 2.04% mAP, which further demonstrates the importance of the attention gate weights for dealing with the challenging scenes especially when the sensor data is disturbed. Besides, as shown in Fig. 8, we visualize the attention map of the block from the image stream in the case of illumination interference (e.g., $a = 0.5$ and $b = 5$ for underexposure image in Fig. 8 (b)) through Grad-CAM [98]. The attention maps from LI-Fusion in Fig. 8 (c) and CB-Fusion in Fig. 8 (d) mainly focus on the region of foreground objects. Besides, compared with LI-Fusion, CB-Fusion pays more attention to more discriminative foreground regions and ignores irrelevant objects (e.g., traffic signs).

4.4.5 Effect of MC Loss on Score Alignment

TABLE 9

Statistical analysis of the L1 error of confidence scores from two modalities in terms of mean and variance. *Var.* is short for Variance.

LiDAR	MC Loss	Cars		Pedestrians		Cyclists	
		Mean	Var.	Mean	Var.	Mean	Var.
16-beam	-	4.70	9.50	4.19	14.76	4.50	22.12
	✓	4.49	0.62	3.24	0.95	3.27	0.79
64-beam	-	4.51	15.48	3.41	6.84	3.27	28.33
	✓	1.52	0.51	0.64	0.14	1.63	1.13

To analyze the effect of MC loss on promoting the consistency of confidence scores from two modalities, we compute their L1 error in terms of the mean and variance of confidence scores with or without MC loss. Table 9 presents the L1 error for three categories and different beam LiDAR points. It is obvious that MC loss narrows the difference in confidence scores predicted by two streams, which indicates

that MC loss is beneficial for generating more stable and reliable scores.

4.4.6 Influence of λ_1 and λ_2

TABLE 10

Quantitative analysis of the hyperparameter λ_1 and λ_2 . For each category, we report the average value of mAP under three difficulty levels.

LiDAR	λ_1	λ_2	Cars	Pedestrians	Cyclists	mAP
16-beam	0.0	0.0	73.16	58.00	44.43	58.53
16-beam	0.0	1.0	72.41	58.69	45.35	58.82
16-beam	0.5	0.5	73.70	59.90	44.42	59.34
16-beam	1.0	0.0	74.40	56.24	44.30	58.31
64-beam	0.0	0.0	85.11	65.16	69.19	73.15
64-beam	0.0	1.0	85.26	65.75	70.57	73.86
64-beam	0.5	0.5	85.98	66.11	70.02	74.04
64-beam	1.0	0.0	85.29	65.00	66.74	72.34

As shown in Tab. 10, we conduct extensive experiments by varying the values of λ_1 and λ_2 in MC loss to analyze their impact on the detection performance under different densities of points. Note that the average confidence score of the image and geometric stream is used as the default threshold in the nms procedure. Setting both λ_1 and λ_2 to 0 means the MC loss is not employed, which can serve as the baseline. MC loss outperforms the baseline by 0.81% and 0.89% in terms of mAP under 16-beam and 64-beam LiDAR setting respectively, demonstrating the effectiveness of the MC loss in selecting high-quality boxes. Besides, for sparse scenes with 16-beam LiDAR points, setting $\lambda_1 = 0.0$ and $\lambda_2 = 1.0$ produces better detection results on Pedestrians and Cyclists than the setting of $\lambda_1 = 1.0$ and $\lambda_2 = 0.0$. It indicates that the confidence scores predicted by the geometric stream are less reliable especially for these objects with fewer collected points in sparse scenes, and thus more attention should be paid to the scores generated by the image stream.

4.4.7 Scalability of CB-Fusion

TABLE 11

Result of Cars for three difficulty levels on KITTI validation. [†] represents the results with GT sampling strategy [36] enabled.

3D Backbone	Methods	Easy	Moderate	Hard
Transformer-Based	Pointformer [38]	91.06	81.20	79.74
	+CB-Fusion	92.52	82.74	80.39
	+Trans. CB-Fusion	91.99	81.47	79.68
Voxel-Based	SECOND [36]	88.72	78.13	74.03
	+CB-Fusion	89.13	79.47	74.88
	Voxel-RCNN [37]	92.04	82.64	80.23
	+CB-Fusion	92.64	83.51	80.77
Voxel-RCNN [37] [†]		92.57	84.49	82.35
	+CB-Fusion	92.89	85.20	83.06

Transformer-based CB-Fusion. Recent approaches [38, 99–101] have shown the effectiveness of transformers for dealing with point clouds. In this section, we take Pointformer [38] as the backbone network to investigate the application of our CB-Fusion in transformer-based architecture. As shown in the top of Table 11, integrating CB-Fusion into Pointformer leads to a performance gain with mAP of 1.54% on the moderate Cars of the KITTI validation set, which demonstrates the effectiveness of CB-Fusion on the transformer-based 3D backbone. Moreover, motivated by

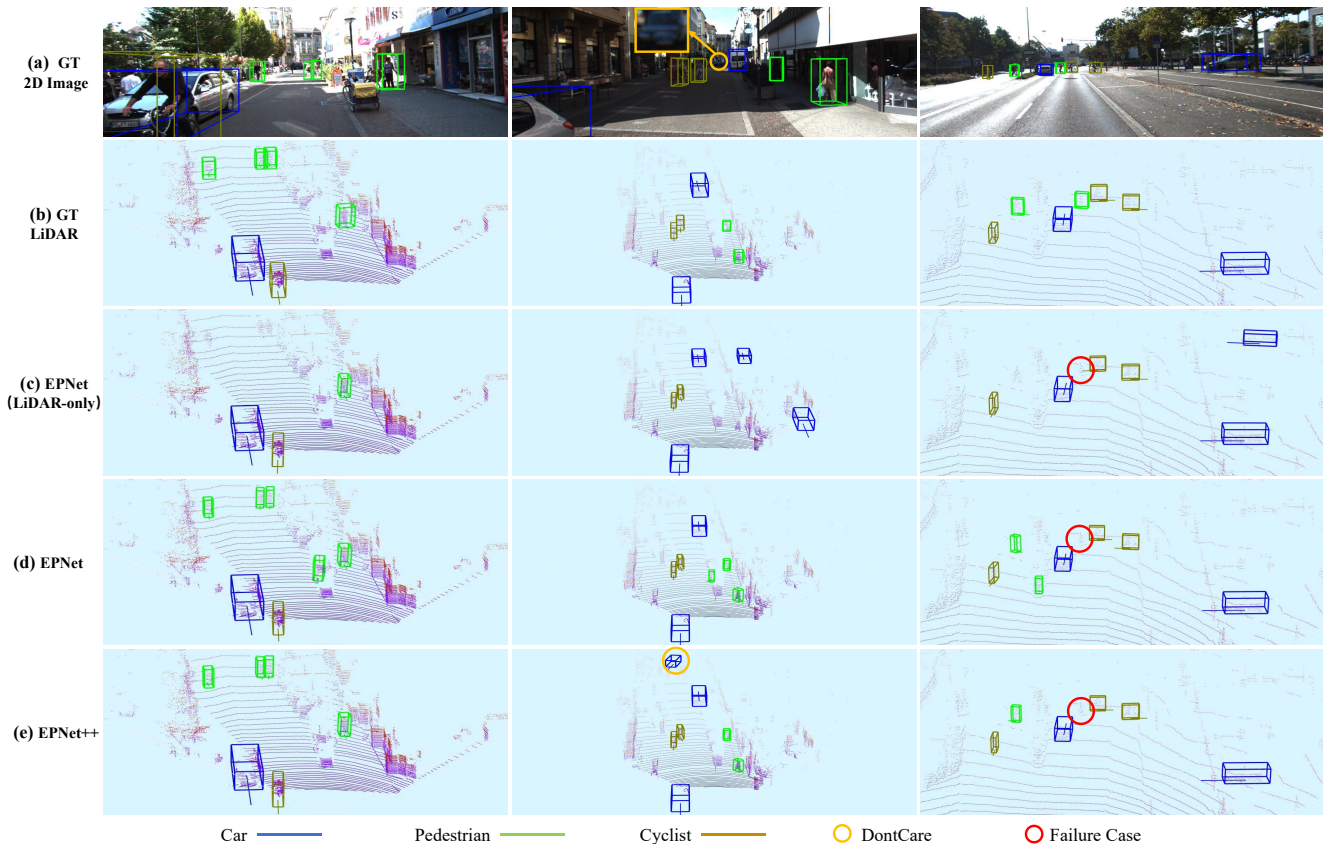


Fig. 9. The visualization on the KITTI dataset. We localize the objects of Cars, Pedestrians and Cyclists with the blue, green and brown 3D bounding boxes, respectively. Besides, we highlight the DontCare object and the failure case (e.g., false negative) with the yellow and red circle box.

the transformer-based sensor fusion [102], we implement a variant of LI-Fusion and IL-Fusion modules by replacing the original attention mechanism with cross-attention. However, we notice that cross-attention even leads to slightly worse performance. We assume the reason is that cross-attention neglects the valuable preliminary information of one-to-one correspondence provided by a projection matrix.

Voxel-based CB-Fusion. We also investigate the application of CB-Fusion in the popular voxel-based frameworks [17, 36, 37]. Concretely, We insert the CB-Fusion module after the first 3D convolution layer of the 3D backbone, since the high-resolution voxel feature is more valuable for reducing the misalignment between two modalities. We select the one-stage detector of SECOND [36] and the two-stage detector of Voxel-RCNN [37] as LiDAR-only baseline. As shown in Table 11, CB-Fusion brings noticeable improvement on Cars for all difficulty levels on the KITTI dataset, which illustrates the effectiveness of CB-Fusion. Besides, we provide the results when the GT sampling strategy [36] is employed for reference inspired by MoCa [51]. Under this setting, CB-Fusion also leads to consistent improvements and boosts the performance of Voxel-RCNN from 84.49% to 85.20% on moderate Cars.

CB-Fusion on Large-scale Dataset. To further illustrate the scalability of our method, we apply our CB-Fusion module to the challenging large-scale Waymo dataset [103]. As shown in Table 12, CB-Fusion leads to consistent improvements across different backbones, categories, and difficulty levels. Moreover, we notice that a deeper 2D feature ex-

TABLE 12
Results on the Waymo validation dataset with a single frame of 20% training samples. We use the official mAP/mAPH [103] as the evaluation metric. Vec-(L1/L2) and Ped-(L1/L2) means the Vehicle and Pedestrians on the difficulty of LEVEL1/LEVEL2. † means using a deeper feature extractor (Mask-RCNN [55] based on ResNet-50 [97]) in the image stream.

Methods	Vec-L1	Vec-L2	Ped-L1	Ped-L2
CenterPoint [17]	71.29/70.72	63.19/62.68	71.61/64.75	63.70/57.45
+ CB-Fusion	71.67/71.09	63.61/63.08	71.97/65.35	63.97/57.93
+ CB-Fusion [†]	71.84/71.31	63.82/63.33	72.68/65.99	64.83/58.70
Voxel RCNN [37]	76.14/75.67	67.74/67.31	78.31/71.22	69.72/63.19
+ CB-Fusion	76.33/75.85	67.91/67.48	78.75/72.09	70.17/64.02
+ CB-Fusion [†]	76.57/76.10	68.29/67.86	79.30/72.83	70.78/64.77

tractor is important for better performance. As shown in Table 12, replacing the default tiny image backbone with Mask-RCNN [55] based on ResNet-50 [97], CB-Fusion leads to noticeable improvements of 0.55% and 1.58% over Voxel-RCNN on Vehicles and Pedestrians in terms of L2 mAPH. In this paper, we mainly focus on the design of the fusion mechanism and leave the usage of a large image backbone as future work.

4.5 Analysis of Visualization

4.5.1 Visualization of Detection Results

KITTI Dataset. Fig. 9 presents qualitative comparison for LiDAR-only EPNet (removing the image stream), EPNet, and EPNet++ on the KITTI dataset. As is shown in the first column, LiDAR-only EPNet fails to detect the pedestrians

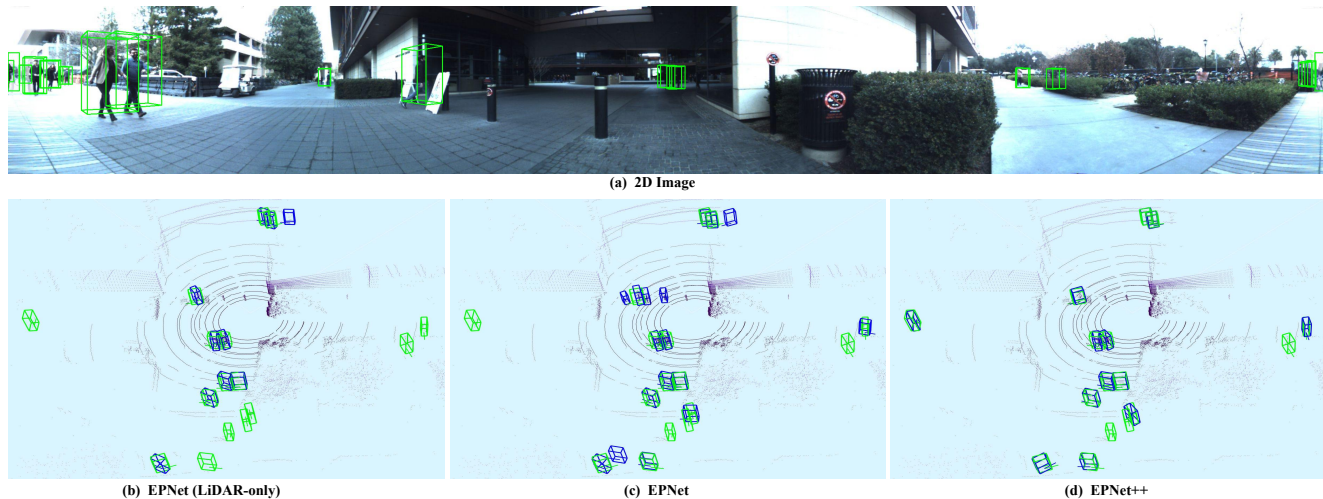


Fig. 10. The visualization results on the JRDB dataset. The predicted boxes and GT boxes are highlighted in blue and green, respectively. In the first row, we project the GT 3D bounding boxes onto the 360° cylindrical image. The second row shows the GT boxes and the prediction results of EPNet (LiDAR-only), EPNet and the proposed EPNet++, respectively.

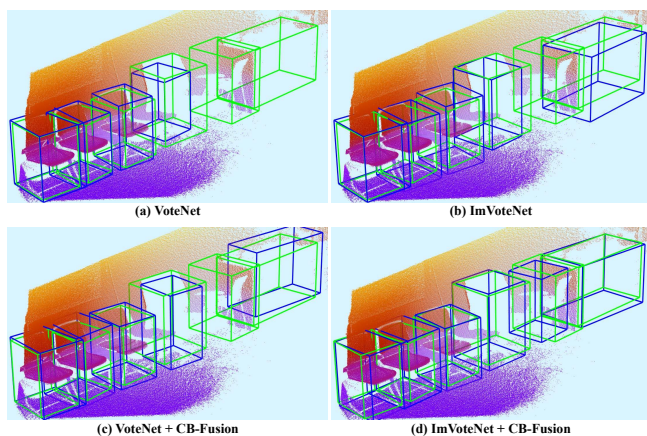


Fig. 11. The visualization results on the SUN-RGBD dataset. The first row displays the 3D detection boxes of VoteNet [35] and ImVoteNet [24]. The second row presents the results of integrating our CB-Fusion module into these two approaches. The predicted boxes and GT boxes are highlighted in blue and green, respectively.

far away with few points. With the aid of multi-modal fusion, EPNet and EPNet++ detect these pedestrians by leveraging the semantic information from images. In the second column, our EPNet++ even detects a 3D object that is labeled as *DontCare* [39]⁶ in GT label. In the last column, our methods produce a failure case, where a long-distance pedestrian is not recognized by our detector. The possible reason is that this distant pedestrian with only occupied few points and maybe be interfered by its nearby objects.

JRDB Dataset. As is shown in Fig. 10, EPNet detects more pedestrians than LiDAR-only EPNet with the help of the LI-Fusion module. However, EPNet also produces several false positives for challenging objects with occlusion, and/or few points. In contrast, EPNet++ successfully filters these false positives thanks to both the more discriminative feature

6. KITTI Dataset uses “DontCare” to denote regions in which objects have not been labeled, which are usually too far away from the laser scanner.

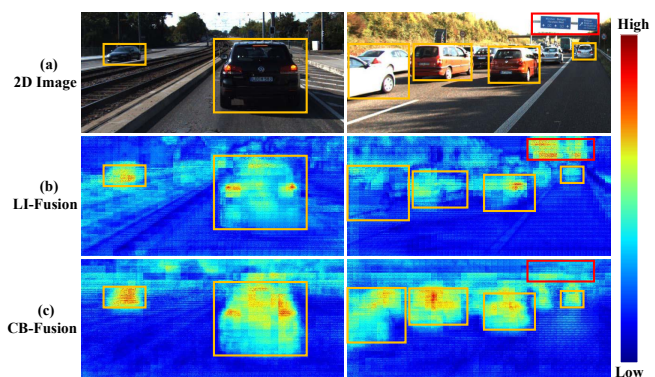


Fig. 12. Visualization of the learned semantic image feature. Foreground objects are highlighted with yellow rectangle boxes. The red rectangle box marks the bad region learned from the image stream.

representations generated by superior CB-Fusion and more reliable confidence scores optimized by MC loss.

SUN-RGBD Dataset. The Fig. 11 illustrates the qualitative results of two representative detectors VoteNet [35], ImVoteNet [24] under two settings, *i.e.*, with and without our CB-Fusion module. It can be observed that CB-Fusion leads to more precise boxes and higher recall, further demonstrating the effectiveness and generalization of our CB-Fusion module.

4.5.2 Visualization of Learned Image Semantic Features

To investigate the image semantic features learned by the LI-Fusion and CB-Fusion modules, we remove the explicit supervision information (*i.e.*, 2D semantic segmentation annotations) from the two-stream RPN. It means that the image stream is optimized with only the implicit supervision from the geometric stream of the two-stream RPN. We present the visualization of the learned image feature map by summing all the channel maps, as shown in Fig. 12. Although no explicit supervision is applied, surprisingly, the image stream learns well to differentiate the foreground objects

and extracts rich semantic features from camera images. It demonstrates that the proposed LI-Fusion and CB-Fusion modules accurately establish the correspondence between LiDAR point cloud and camera image, thus can provide the complement semantic image information to enhance the point features. Besides, in the second column, compared with the LI-Fusion module, the learned semantic feature through the CB-Fusion module is more plentiful and can effectively suppress the bad region.

5 CONCLUSION

This paper has presented a novel end-to-end trainable framework for multi-modal 3D object detection named EP-Net++, consisting of a two-stream RPN and a refinement network. Specifically, we propose a novel CB-Fusion module to enable bi-directional feature enhancement pathways, leading to more discriminative and comprehensive feature representations. The MC loss is utilized to promote the consistency of the confidence scores between the image and geometric streams, which is beneficial for selecting high-quality proposals. Extensive ablation studies are conducted to validate the effectiveness of the proposed CB-Fusion module and the MC loss. Our method achieves competitive and even SOTA detection performance on the KITTI, JRDB, and SUN-RGBD datasets, which demonstrates the superiority of EPNet++. Moreover, EPNet++ outperforms the SOTA methods with remarkable margins in highly sparse point cloud scenes, benefiting from the powerful CB-Fusion module and MC loss. In the future, we would like to investigate more efficient architecture for multi-modal fusion through unifying different sensors into one stream.

6 ACKNOWLEDGEMENT

This work was supported by the National Science Foundation of China for Distinguished Young Scholars (No. 62225603).

REFERENCES

- [1] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang, and X. Fan, "Accurate monocular object detection via color-embedded 3d reconstruction for autonomous driving," in *Proc. ICCV*, 2019.
- [2] Z. Qin, J. Wang, and Y. Lu, "Monogrnnet: A geometric reasoning network for monocular 3d object localization," in *Proc. AAAI*, vol. 33, no. 01, 2019, pp. 8851–8858.
- [3] J. Ku*, A. D. Pon*, and S. L. Waslander, "Monocular 3d object detection leveraging accurate proposals and shape reconstruction," in *Proc. CVPR*, 2019.
- [4] B. Li, W. Ouyang, L. Sheng, X. Zeng, and X. Wang, "Gs3d: An efficient 3d object detection framework for autonomous driving," in *Proc. CVPR*, 2019.
- [5] L. Liu, J. Lu, C. Xu, Q. Tian, and J. Zhou, "Deep fitting degree scoring network for monocular 3d object detection," in *Proc. CVPR*, 2019, pp. 1057–1066.
- [6] Z. Liu, D. Zhou, F. Lu, J. Fang, and L. Zhang, "Autoshape: Real-time shape-aware monocular 3d object detection," in *Proc. ICCV*, 2021, pp. 15 641–15 650.
- [7] P. Li, X. Chen, and S. Shen, "Stereo r-cnn based 3d object detection for autonomous driving," in *Proc. CVPR*, 2019.
- [8] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *Proc. CVPR*, 2019.
- [9] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *Proc. CVPR*, 2016.
- [10] X. Shi, Q. Ye, X. Chen, C. Chen, Z. Chen, and T.-K. Kim, "Geometry-based distance decomposition for monocular 3d object detection," *arXiv:2104.03775*, 2021.
- [11] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3d bounding box estimation using deep learning and geometry," in *Proc. CVPR*, 2017.
- [12] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proc. CVPR*, 2017, pp. 1907–1915.
- [13] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proc. CVPR*, 2018.
- [14] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3d object detection from point clouds," in *Proc. CVPR*, 2018.
- [15] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proc. CVPR*, 2019, pp. 770–779.
- [16] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proc. CVPR*, 2019.
- [17] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proc. CVPR*, 2021, pp. 11 784–11 793.
- [18] J. Mao, Y. Xue, M. Niu, H. Bai, J. Feng, X. Liang, H. Xu, and C. Xu, "Voxel transformer for 3d object detection," in *Proc. ICCV*, 2021, pp. 3164–3173.
- [19] W. Zheng, W. Tang, L. Jiang, and C.-W. Fu, "Se-ssd: Self-ensembling single-stage object detector from point cloud," in *Proc. CVPR*, 2021, pp. 14 494–14 503.
- [20] G. P. Meyer, A. Laddha, E. Kee, C. Vallespi-Gonzalez, and C. K. Wellington, "LaserNet: An efficient probabilistic 3D object detector for autonomous driving," in *Proc. CVPR*, 2019.
- [21] T. Guan, J. Wang, S. Lan, R. Chandra, Z. Wu, L. Davis, and D. Manocha, "M3detr: Multi-representation, multi-scale, mutual-relation 3d object detection with transformers," in *Proc. WACV*, 2022, pp. 772–782.
- [22] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvx-net: Multi-modal voxelnet for 3d object detection," in *Intl. Conf. on Robotics and Automation*. IEEE, 2019, pp. 7276–7282.
- [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 *Proc. ECCV*. Springer, 2020, pp. 720–736.
- [24] C. R. Qi, X. Chen, O. Litany, and L. J. Guibas, "Imvotenet: Boosting 3d object detection in point clouds with image votes," in *Proc. CVPR*, 2020, pp. 4404–4413.
- [25] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proc. ECCV*, 2018, pp. 641–656.
- [26] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Point-painting: Sequential fusion for 3d object detection," in *Proc. CVPR*, 2020, pp. 4604–4612.
- [27] T. Huang, Z. Liu, X. Chen, and X. Bai, "Epnet: Enhancing point features with image semantics for 3d object detection," in *Proc. ECCV*. Springer, 2020, pp. 35–52.
- [28] 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 *Proc. AAAI*, vol. 34, no. 07, 2020, pp. 12 460–12 467.
- [29] D. Xu, D. Anguelov, and A. Jain, "Pointfusion: Deep sensor fusion for 3d bounding box estimation," in *Proc. CVPR*, 2018, pp. 244–253.
- [30] C. Wang, C. Ma, M. Zhu, and X. Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in

- Proc. CVPR*, 2021, pp. 11 794–11 803.
- [31] A. Piergiovanni, V. Casser, M. S. Ryoo, and A. Angelova, “4d-net for learned multi-modal alignment,” in *Proc. ICCV*, 2021, pp. 15 435–15 445.
- [32] D. Xu, D. Anguelov, and A. Jain, “Pointfusion: Deep sensor fusion for 3d bounding box estimation,” in *Proc. CVPR*, 2018.
- [33] X. Zhao, Z. Liu, R. Hu, and K. Huang, “3d object detection using scale invariant and feature reweighting networks,” in *Proc. AAAI*, vol. 33, no. 01, 2019, pp. 9267–9274.
- [34] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, “Multi-task multi-sensor fusion for 3d object detection,” in *Proc. CVPR*, 2019, pp. 7345–7353.
- [35] C. R. Qi, O. Litany, K. He, and L. J. Guibas, “Deep hough voting for 3d object detection in point clouds,” *Proc. ICCV*, 2019.
- [36] Y. Yan, Y. Mao, and B. Li, “Second: Sparsely embedded convolutional detection,” *Sensors*, vol. 18, no. 10, p. 3337, 2018.
- [37] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, “Voxel r-cnn: Towards high performance voxel-based 3d object detection,” in *Proc. AAAI*, vol. 35, no. 2, 2021, pp. 1201–1209.
- [38] X. Pan, Z. Xia, S. Song, L. E. Li, and G. Huang, “3d object detection with pointformer,” in *Proc. CVPR*, 2021, pp. 7463–7472.
- [39] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *Proc. CVPR*. IEEE, 2012, pp. 3354–3361.
- [40] S. Song, S. P. Lichtenberg, and J. Xiao, “Sun rgb-d: A rgb-d scene understanding benchmark suite,” in *Proc. CVPR*, 2015.
- [41] R. Martin-Martin*, M. Patel*, H. Rezatofighi*, A. Sheno, J. Gwak, E. Frankel, A. Sadeghian, and S. Savarese, “JRDB: A dataset and benchmark of egocentric robot visual perception of humans in built environments,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.
- [42] Y. Chen, S. Liu, X. Shen, and J. Jia, “Dsgn: Deep stereo geometry network for 3d object detection,” in *Proc. CVPR*, 2020, pp. 12 536–12 545.
- [43] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, “Pv-rcnn: Point-voxel feature set abstraction for 3d object detection,” in *Proc. CVPR*, 2020.
- [44] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, “From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.
- [45] Z. Li, F. Wang, and N. Wang, “Lidar r-cnn: An efficient and universal 3d object detector,” in *Proc. CVPR*, 2021, pp. 7546–7555.
- [46] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proc. CVPR*, 2017, pp. 652–660.
- [47] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” in *Proc. NeurIPS*, 2017, pp. 5099–5108.
- [48] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, “Std: Sparse-to-dense 3d object detector for point cloud,” in *Proc. CVPR*, 2019, pp. 1951–1960.
- [49] Z. Yang, Y. Sun, S. Liu, and J. Jia, “3dssd: Point-based 3d single stage object detector,” in *Proc. CVPR*, 2020, pp. 11 040–11 048.
- [50] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, “Joint 3d proposal generation and object detection from view aggregation,” in *IEEE/RSJ Intl. Conf. on Intell. Robots and Systems*. IEEE, 2018, pp. 1–8.
- [51] W. Zhang, Z. Wang, and C. Change Loy, “Multi-modality cut and paste for 3d object detection,” *arXiv e-prints*, pp. arXiv–2012, 2020.
- [52] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *Proc. ECCV*. Springer, 2016, pp. 21–37.
- [53] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proc. CVPR*, 2016, pp. 779–788.
- [54] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *Proc. NeurIPS*, vol. 28, pp. 91–99, 2015.
- [55] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proc. ICCV*, 2017, pp. 2961–2969.
- [56] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, “Densefusion: 6d object pose estimation by iterative dense fusion,” in *Proc. CVPR*, 2019, pp. 3343–3352.
- [57] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, “Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes,” *arXiv:1711.00199*, 2017.
- [58] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. ICML*, 2015.
- [59] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, “Bounding box regression with uncertainty for accurate object detection,” in *Proc. CVPR*, 2019, pp. 2888–2897.
- [60] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proc. ICCV*, 2017.
- [61] R. Girshick, “Fast r-cnn,” in *Proc. ICCV*, 2015, pp. 1440–1448.
- [62] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, “Frustum pointnets for 3d object detection from rgb-d data,” in *Proc. CVPR*, 2018.
- [63] Z. Qin, J. Wang, and Y. Lu, “Monogrnet: A general framework for monocular 3d object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.
- [64] G. Brazil and X. Liu, “M3d-rpn: Monocular 3d region proposal network for object detection,” in *Proc. ICCV*, 2019, pp. 9287–9296.
- [65] Y. Chen, L. Tai, K. Sun, and M. Li, “Monopair: Monocular 3d object detection using pairwise spatial relationships,” in *Proc. CVPR*, 2020, pp. 12 093–12 102.
- [66] P. Li, H. Zhao, P. Liu, and F. Cao, “Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving,” in *Proc. ECCV*. Springer, 2020, pp. 644–660.
- [67] X. Ma, S. Liu, Z. Xia, H. Zhang, X. Zeng, and W. Ouyang, “Rethinking pseudo-lidar representation,” in *Proc. ECCV*. Springer, 2020, pp. 311–327.
- [68] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, “Categorical depth distribution network for monocular 3d object detection,” in *Proc. CVPR*, 2021, pp. 8555–8564.
- [69] Y. Lu, X. Ma, L. Yang, T. Zhang, Y. Liu, Q. Chu, J. Yan, and W. Ouyang, “Geometry uncertainty projection network for monocular 3d object detection,” in *Proc. ICCV*, 2021, pp. 3111–3121.
- [70] D. Park, R. Ambrus, V. Guizilini, J. Li, and A. Gaidon, “Is pseudo-lidar needed for monocular 3d object detection?” in *Proc. ICCV*, 2021, pp. 3142–3152.
- [71] Z. Liu, X. Zhao, T. Huang, R. Hu, Y. Zhou, and X. Bai, “Tanet: Robust 3d object detection from point clouds with triple attention,” in *Proc. AAAI*, vol. 34, no. 07, 2020, pp. 11 677–11 684.
- [72] Y. Chen, S. Liu, X. Shen, and J. Jia, “Fast point r-cnn,” in *Proc. ICCV*, 2019, pp. 9775–9784.
- [73] Q. Chen, L. Sun, Z. Wang, K. Jia, and A. Yuille, “Object as hotspots: An anchor-free 3d object detection approach via firing of hotspots,” in *Proc. ECCV*. Springer, 2020, pp. 68–84.
- [74] D. Zhou, J. Fang, X. Song, L. Liu, J. Yin, Y. Dai, H. Li, and R. Yang, “Joint 3d instance segmentation and object detection for autonomous driving,” in *Proc. CVPR*, 2020,

- pp. 1839–1849.
- [75] W. Shi and R. R. Rajkumar, “Point-gnn: Graph neural network for 3d object detection in a point cloud,” in *Proc. CVPR*, 2020.
- [76] J. Li, H. Dai, L. Shao, and Y. Ding, “Anchor-free 3d single stage detector with mask-guided attention for point cloud,” in *ACM Intl. Conf. on Multimedia*. ACM, 2021.
- [77] C. He, H. Zeng, J. Huang, X.-S. Hua, and L. Zhang, “Structure aware single-stage 3d object detection from point cloud,” in *Proc. CVPR*, 2020.
- [78] W. Zheng, W. Tang, S. Chen, L. Jiang, and C.-W. Fu, “Cia-ssd: Confident iou-aware single-stage object detector from point cloud,” in *Proc. AAAI*, 2021.
- [79] J. Li, H. Dai, L. Shao, and Y. Ding, “From voxel to point: Iou-guided 3d object detection for point cloud with voxel-to-point decoder,” in *ACM Intl. Conf. on Multimedia*. ACM, 2021.
- [80] R. Qian, X. Lai, and X. Li, “Badet: Boundary-aware 3d object detection from point clouds,” *Pattern Recognition*, vol. 125, p. 108524, 2022.
- [81] Z. Li, Y. Yao, Z. Quan, W. Yang, and J. Xie, “Sienet: spatial information enhancement network for 3d object detection from point cloud,” *arXiv:2103.15396*, 2021.
- [82] Z. Wang and K. Jia, “Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection,” in *IEEE/RSJ Intl. Conf. on Intell. Robots and Systems*. IEEE, 2019, pp. 1742–1749.
- [83] S. Pang, D. Morris, and H. Radha, “Clocs: Camera-lidar object candidates fusion for 3d object detection,” in *IEEE/RSJ Intl. Conf. on Intell. Robots and Systems*. IEEE, 2020, pp. 10 386–10 393.
- [84] G. Wang, B. Tian, Y. Zhang, L. Chen, D. Cao, and J. Wu, “Multi-view adaptive fusion network for 3d object detection,” *arXiv:2011.00652*, 2020.
- [85] L. Qi, L. Jiang, S. Liu, X. Shen, and J. Jia, “Amodal instance segmentation with kins dataset,” in *Proc. CVPR*, 2019, pp. 3014–3023.
- [86] A. Simonelli, S. R. Bulo, L. Porzi, M. López-Antequera, and P. Kontschieder, “Disentangling monocular 3d object detection,” in *Proc. ICCV*, 2019, pp. 1991–1999.
- [87] Q. Xie, Y.-K. Lai, J. Wu, Z. Wang, Y. Zhang, K. Xu, and J. Wang, “Mlcvnet: Multi-level context votenet for 3d object detection,” in *Proc. CVPR*, 2020, pp. 10 447–10 456.
- [88] Z. Zhang, B. Sun, H. Yang, and Q. Huang, “H3dnet: 3d object detection using hybrid geometric primitives,” in *Proc. ECCV*, 2020.
- [89] J. Chen, B. Lei, Q. Song, H. Ying, D. Z. Chen, and J. Wu, “A hierarchical graph network for 3d object detection on point clouds,” in *Proc. CVPR*, 2020, pp. 392–401.
- [90] Z. Liu, Z. Zhang, Y. Cao, H. Hu, and X. Tong, “Group-free 3d object detection via transformers,” *arXiv:2104.00678*, 2021.
- [91] S. Song and J. Xiao, “Deep sliding shapes for amodal 3d object detection in rgb-d images,” in *Proc. CVPR*, 2016.
- [92] J. Lahoud and B. Ghanem, “2d-driven 3d object detection in rgb-d images,” in *Proc. ICCV*, 2017.
- [93] Z. Ren and E. B. Sudderth, “Three-dimensional object detection and layout prediction using clouds of oriented gradients,” in *Proc. CVPR*, 2016.
- [94] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv:1412.6980*, 2014.
- [95] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Proc. NeurIPS*, vol. 32, pp. 8026–8037, 2019.
- [96] Y. You, Y. Wang, W.-L. Chao, D. Garg, G. Pleiss, B. Hariharan, M. Campbell, and K. Q. Weinberger, “Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving,” in *Proc. ICLR*, 2020.
- [97] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, 2016, pp. 770–778.
- [98] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proc. CVPR*, 2017, pp. 618–626.
- [99] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, “Point transformer,” in *Proc. CVPR*, 2021, pp. 16 259–16 268.
- [100] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, “Pct: Point cloud transformer,” *Computational Visual Media*, vol. 7, no. 2, pp. 187–199, 2021.
- [101] X.-F. Han, Y.-J. Kuang, and G.-Q. Xiao, “Point cloud learning with transformer,” *arXiv:2104.13636*, 2021.
- [102] A. Prakash, K. Chitta, and A. Geiger, “Multi-modal fusion transformer for end-to-end autonomous driving,” in *Proc. CVPR*, 2021, pp. 7077–7087.
- [103] 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 *Proc. CVPR*, 2020, pp. 2446–2454.