CL3D: Camera-LiDAR 3D Object Detection With Point Feature Enhancement and Point-Guided Fusion

Chunmian Lin, Daxin Tian, *Senior Member, IEEE*, Xuting Duan, *Member, IEEE*, Jianshan Zhou, Dezong Zhao, *Senior Member, IEEE*, and Dongpu Cao

***Abstract*— Camera-LiDAR 3D object detection has been extensively investigated due to its significance for many real-world applications. However, there are still of great challenges to address the intrinsic data difference and perform accurate feature fusion among two modalities. To these ends, we propose a two-stream architecture termed as CL3D, that integrates with point enhancement module, point-guided fusion module with IoU-aware head for cross-modal 3D object detection. Specifically, pseudo LiDAR is firstly generated from RGB image, and point enhancement module (PEM) is then designed to enhance the raw LiDAR with pseudo point. Moreover, point-guided fusion module (PFM) is developed to find image-point correspondence at different resolutions, and incorporate semantic with geometric features in a point-wise manner. We also investigate the inconsistency between localization confidence and classification score in 3D detection, and introduce IoU-aware prediction head (IoU Head) for accurate box regression. Comprehensive experiments are conducted on publicly available KITTI dataset, and CL3D reports the outstanding detection performance compared to both single- and multi-modal 3D detectors, demonstrating its effectiveness and competitiveness.**

***Index Terms*— 3D object detection, camera-LiDAR fusion, deep learning, autonomous driving, intelligent transportation systems.**

I. Introduction

RECENT years, 3D object detection has received more and more attention on both industry and academia, due to its various applications in many fields such as autonomous driving. With the advancement of deep learning and convolutional neural network (CNN), 2D object detection technique has achieved remarkable progress in efficient network architectures [1], [2], hierarchical feature presentations [3], [4] and outstanding performance [5], [6]. Therefore, numerous image-based object detection methods have been developed for 3D object localization [7]–[9], whereas monocular 3D object detector suffers from false positives owing to the lack of depth information of object. As an alternative, LiDAR sensor can provide more robust geometric feature that is applicable to describe the spatial structure of object in the 3D scene [10], [11]. Consequently, many LiDAR-based 3D detection methods are developed, and they can be roughly divided into two classes. On one hand, PointNet [10] architecture directly consumes the raw LiDAR for geometric feature learning, and spatial feature maps are further utilized for proposal generation and 3D bounding-box regression [12], [13]. On the other hand, researchers focus on regular representation by converting the raw point cloud into grid format (i.e. voxel), and thus efficient CNN architecture can be adopted for 3D object detection [14]–[16]. However, without the help of semantic feature, it is still difficult to distinguish the adjacent objects if they have similar geometric structure. Furthermore, because of the inherent sparisity and orderless of point cloud, LiDAR-only detectors easily suffer from false detection in far-away and small objects, as shown in Fig.1. Generally, neither image-based nor LiDAR-only methods present satisfactory performance in the challenging physical world.

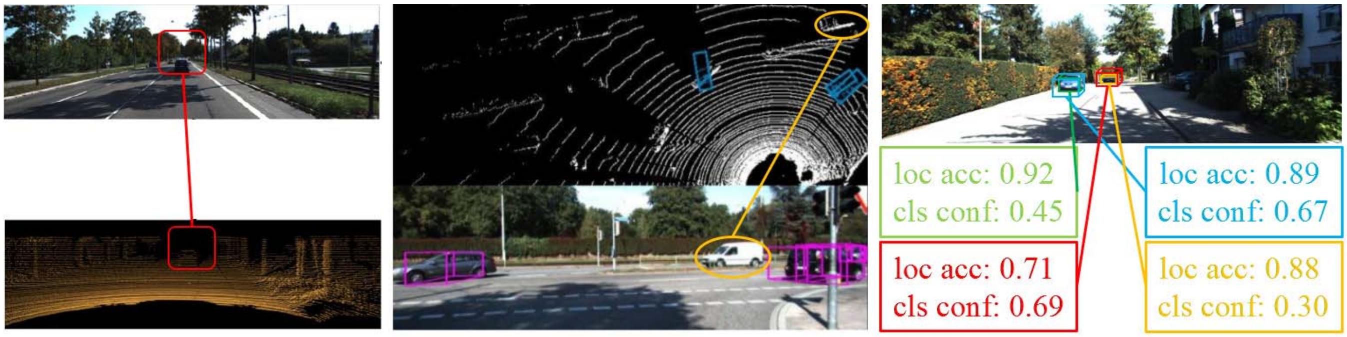
Therefore, researchers are increasingly interested in fusing multi-modal information and exploiting the complementary among different sensors for more robust and accurate detection performance. Camera and LiDAR are two widely used sensors in 3D object detection community. The former captures RGB image with dense object semantics, while the latter provides reliable point cloud information for describing the geometric structure of object. Combining these two modalities would capture both semantic and geometric feature representations, and can result in more promising detection results. Currently, there are several camera-LiDAR 3D object detection methods that perform image-point feature aggregation via various fusion strategies and architectures, such as multi-view feature fusion [17], [18], continuous convolution fusion [19], multi-task fusion [20], point-wise fusion [21], [22], cross-view spatial feature fusion [23], proposal candidate fusion [24], etc. As demonstrated in Fig.1, current works mostly perform

Manuscript received March 15, 2021; revised September 29, 2021, November 15, 2021, and January 27, 2022; accepted February 22, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62061130221, Grant U20A20155, and Grant 62173012; in part by the Beijing Municipal Natural Science Foundation under Grant L191001; in part by the Zhuoyue Program of Beihang University (Postdoctoral Fellowship); and in part by the China Postdoctoral Science Foundation under Grant 2020M680299. The Associate Editor for this article was R. Arghandeh. *(Corresponding author: Daxin Tian.)*

Chunmian Lin, Daxin Tian, Xuting Duan, and Jianshan Zhou are with the Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, Beijing Advanced Innovation Center for Big Data and Brain Computing, School of Transportation Science and Engineering, Beihang University, Beijing 100191, China (e-mail: dtian@buaa.edu.cn).

Dezong Zhao is with the James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, U.K.

Dongpu Cao is with the Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada.



(a) Sparsity of LiDAR (b) Imperfect Correspondence (c) Inconsistence in Object Detection

Fig. 1. The existing problems in 3D object detection. **(a) Sparisity of LiDAR.** the LiDAR point is too sparse for faraway car to distinguish. **(b) Imperfect Correspondence.** most fusion methods fail to find perfect image-point corrspondence, and thus result in false or missing detection. **(c) Inconsistency in Object Detection.** the misalignment between localization accuracy and classification confidence commonly exists in 3D object detection. Best viewed in color.

coarse fusion via spatial transformation, which leads to information loss and quantization error inevitably. And without accurate point-wise correspondences, these methods generally report marginal performance gains. Consequently, it is still of great challenge to develop an effective cross-modal fusion method for 3D object detection.

To solve these problems, we propose a camera-LiDAR 3D object detector named CL3D, that is a two-stream architecture with point enhancement module (PEM), point-guided fusion module (PFM) and IoU-aware head (IoU Head). On one hand, point enhancement module combines the pseudo representation generated from RGB image with the raw LiDAR for point feature enhancement. One the other hand, point-guided fusion module exploits image-point correspondences to aggregate multi-level semantic with geometric features, resulting in more representative cross-modal features under different resolutions. Furthermore, we investigate the inconsistence between localization confidence and classification score as illustrated in Fig.1, and develop a simple yet effective IoU-aware prediction head (IoU Head) for accurate 3D box regression. Extensive experiments are conducted on publicly available KITTI dataset [25], and our proposed CL3D presents competitive detection accuracy and significant performance gains over single- and multi-modal 3D detection methods.

Overall, the contributions in this work can be summarized as follows:

1) We generate pseudo point directly from RGB image, and fuse it with the raw LiDAR by point enhancement module (PEM), which can enhance point feature representation effectively.

2) Point-guided fusion module (PFM) is proposed to find perfect correspondence between image and point, and we perform point-wise feature aggregation to produce more discriminative multi-modal feature at various resolutions.

3) The misalignment between localization confidence and classification score is investigated, and we design IoU-aware prediction head (IoU Head) for the IoU calculation between each ground-truth and predicted box.

4) We integrate PEM, PFM and IoU Head into a two-stream architecture for camera-LiDAR 3D object detection, termed as CL3D. Extensive experimental results on KITTI dataset demonstrate the effectiveness and competitiveness of CL3D, with the promising detection performance and considerable improvements.

The remainder of this paper is organized as follows: we review related works in Section II, and introduce the proposed method CL3D in Section III; experimental analysis and conclusions are presented in Section IV and V, respectively.

II. Related Works

In this part, we would briefly review the development of 3D object detection and the inconsistency problem between localization and classification in object detection.

A. Camera-Based 3D Object Detection

Many researches extends the pipeline of 2D object detection to 3D detection task, that directly performs 3D bounding-box regression and confidence prediction from 2D camera image [26]–[31]. For instance, [26] infers monocular 3D bounding box with RGB semantics and contextual information (i.e. size, location, shape, etc.). Reference [8] imposes geometric constraints on the 2D bounding box, and produces 3D result with object pose from a single frame. Furthermore, [28], [29] defines the notion of pseudo LiDAR for mimicking the LiDAR signal, and generates a set of pseudo point from images by depth estimation algorithm and coordinate transformation. And LiDAR-based methods are further applied for 3D object localization. To further investigate the potential of pseudo-LiDAR representation, [30], [31] propose an end-to-end training pipeline with more accurate depth estimation method to facilitate monocular 3D detection performance. Although RGB image contains rich and dense channel features, it is still difficult to achieve satisfactory 3D detection accuracy due to the lack of reliable depth information.

B. LiDAR-Based 3D Object Detection

LiDAR-based 3D object detection have rapidly developed owing to the advantage of point cloud representation in describing the 3D structure of object. Recent works can be roughly classified into two folds. On one hand, PointNet architecture [10], [11] directly consumes the raw point cloud, and predicts 3D bounding box from pre-defined region proposals [13], [32]–[34]. Benefited by powerful spatial representation of raw LiDAR, STD [32] proposes a new spherical anchor, and converts point cloud from sparse to dense representation for robust 3D detection. 3DSSD [13] develops a novel 3D single-stage object detector with feature fusion sampling strategy and box prediction network. In [33], the authors designs triple attention modules for multi-level point feature fusion, and coarse-to-fine regression branch is further utilized for accurate 3D pedestrian localization. On the other hand, voxel representation is also extensively investigated, that quantifies point cloud into regular grid format (i.e. voxel) and adopts efficient CNN architecture for 3D object detection [34]–[36]. Inspired by VoxelNet [36], [37] designs part-aware and part-aggregation modules for intra-object point feature encoding and box refinement. Moreover, PV-RCNN [16] explicitly combines both 3D voxel convolution and keypoint representation to learn more discriminative feature maps to improve 3D detection performance. CenterPoint [38] is an anchor-free framework that adopts voxel-based feature encoder for simultaneous 3D object detection and tracking. These LiDAR-based approaches achieve state-of-the-art 3D detection performance, however, they easily suffer from semantic ambiguity and false positive results particularly in faraway or similar objects.

C. Camera-LiDAR 3D Object Detection

Multi-modal 3D detection is an emerging research interest, and various camera-LiDAR fusion methods are investigated for representation enhancement and performance gains [20], [22], [23], [39], [40]. To be specific, Frustum-PointNet [39] obtains 3D bounding-box prediction results by extruding a 3D viewing frustum from 2D object region proposal. MMF [20] introduces multi-task applications, and utilizes ground estimation and depth information to facilitates 3D detection performance. Inspired from continuous convolution [19], PI-RCNN [22] proposes point-based attentive continuous-convolution fusion (PACF) module to aggregate cross-modal features effectively. In [23], authors explore the importance of fusing camera-LiDAR feature map from different locations, and design cross-view spatial feature fusion and adaptive gated fusion modules for cross-modal 3D object detection. Nevertheless, these methods mostly align image-point feature via projection transformation, which leads to information loss and quantization error inevitably. And due to the sparsity of LiDAR signal, imperfect image-point correspondences would also degrade the 3D detection performance to some extent.

D. Inconsistence in Object Detection

The inconsistence in object detection implies the misalignment between localization accuracy and classification score, that is an essential issue widely explored in 2D detection community [41]–[45]. IoU-Net [43] calculates the IoU metric between each prediction and the matched ground-truth box, and formulate it as an objective for box refinement. In [44], a single-stage 2D detector with IoU-aware branch is designed to improve the localization accuracy. Moreover, the author investigates the effect of loss function for detection performance, and proposes IoU-balanced classification loss [45] to assign the positive example with better IoU and higher classification score simultaneously. However, this imbalanced problem has not received enough attention in 3D detection research yet. Recently, 3D IoU-Net [46] considers the box matching strategy and proposes the IoU alignment method to improve 3D-IoU prediction. CIA-SSD [47] presents IoU-aware rectification module to correct the prediction error between the localization confidence and classification score. In this work, we would also investigate this inconsistence between localization and classification, and provide a simple yet effective solution to alleviate this problem.

III. CL3D: Camera-LiDAR 3D Object Detection

As illustrated in Fig.2, the overall architecture of CL3D mainly contains point enhancement module, image and point backbone, point-guided fusion module and refinement network. Detailed information would be introduced as follows.

A. Point Enhancement Module

As mentioned above, pseudo LiDAR is an effective representation that improves 3D detection performance significantly. It generates directly from RGB image via depth estimation and spatial coordinate transformation, which indicates pseudo point would contains dense semantics from RGB channels. Therefore, we attempt to generate pseudo representation from RGB image and enhance the raw LiDAR feature via point enhancement module (PEM).

Specifically, we adopt the pretrained pyramid stereo matching network (PSMNet) [48], that takes a pair of input images to calculate the disparity map, with the size of 375 × 1242. And subsequently, 3D coordiantes of each pixel are derived from the left camera coordinate system via the following formulations Eq.1-Eq.3:

$$m = \frac{(u - c_u) \times b}{D(u, v)} \tag{1}$$

$$n = \frac{(v - c_v) \times t}{f_v} \tag{2}$$

$$t = \frac{f_h \times b}{D(u, v)} \tag{3}$$

where (m, n, t) denotes the 3D coordinate value corresponding to each pixel (u, v) in image plane; c_u and c_v define the pixel location of camera center; f_h and f_v are the horizontal and vertical focal length, respectively; $D(u, v)$ presents the disparity map generated by PSMNet, and b is the horizontal offset between a pair of images. To alleviate the noise interference, we further disregard the pseudo point with abnormal height, and set the reflectance to 1.0 for each point. As presented in Fig.3, the generated pseudo LiDAR can be denoted as $\{(m_i, n_i, t_i)_{i=(1,\dots,M)}\}$, where M denotes the number of effective pseudo point ($100k$∼$400k$).

As illustrated in Fig.4, point enhancement module (PEM) is further designed to enhance the raw LiDAR with pseudo representations. To be specific, the dense pseudo LiDAR

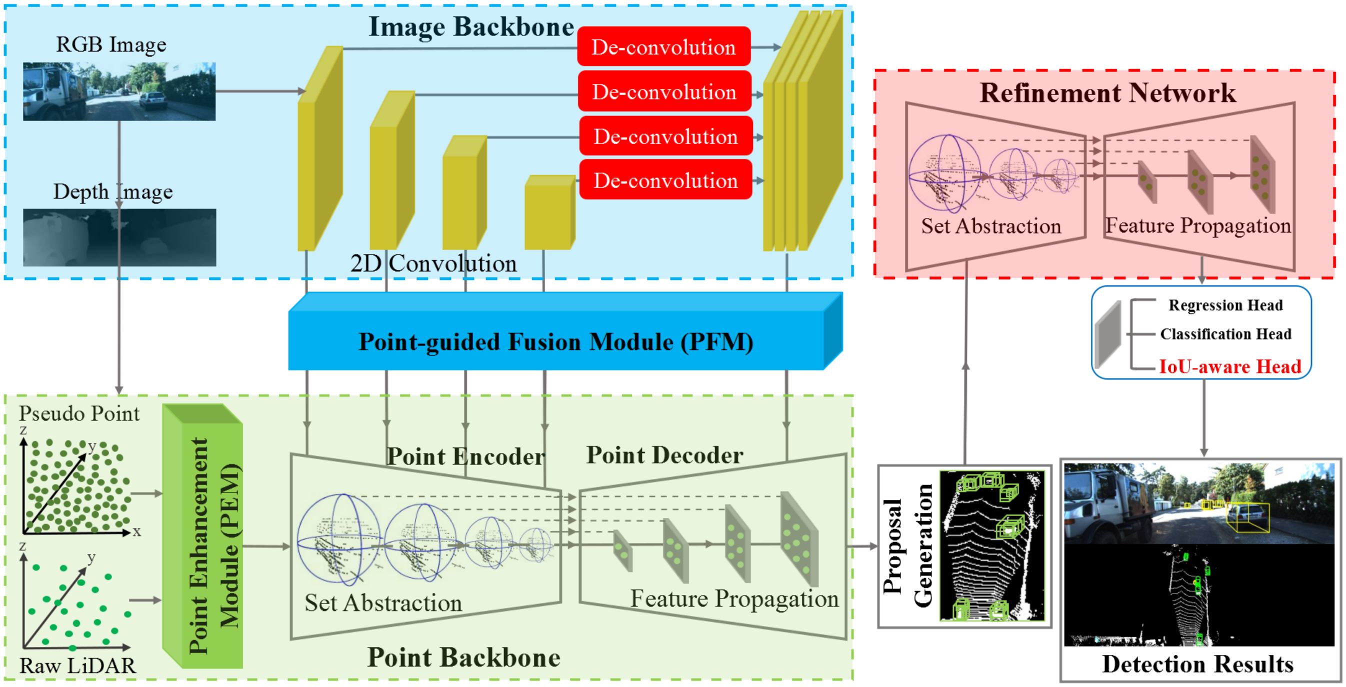


Fig. 2. Overview of CL3D architecture. **(1) Image Backbone:** adopts four convolutional layers to learn semantic feature map hierarchically, and simultaneously recovers the size of feature map via de-convolution for multi-scale semantic feature fusion. **(2) Point Enhancement Module (PEM):** combines pseudo point generated from RGB image and raw LiDAR signal to alleviate the sparisity of LiDAR and enhance point cloud feature. **(3) Point Backbone:** contains four set abstraction layers to aggregate point features in the neighboring region, followed by four feature propagation layers to project point cloud back into the original space. **(4) Point-guided Fusion Module (PFM):** finds image-point correspondence under different resolutions, and fuse semantic and geometric feature in a point-wise manner. **(5) Refinement Network:** utilizes pairs of set abstraction and feature propagation layers for proposal refinement. And multi-task head introduces IoU-aware Head (IoU Head) to calculate the IoU between each ground-truth and predict box. Best viewed in color.

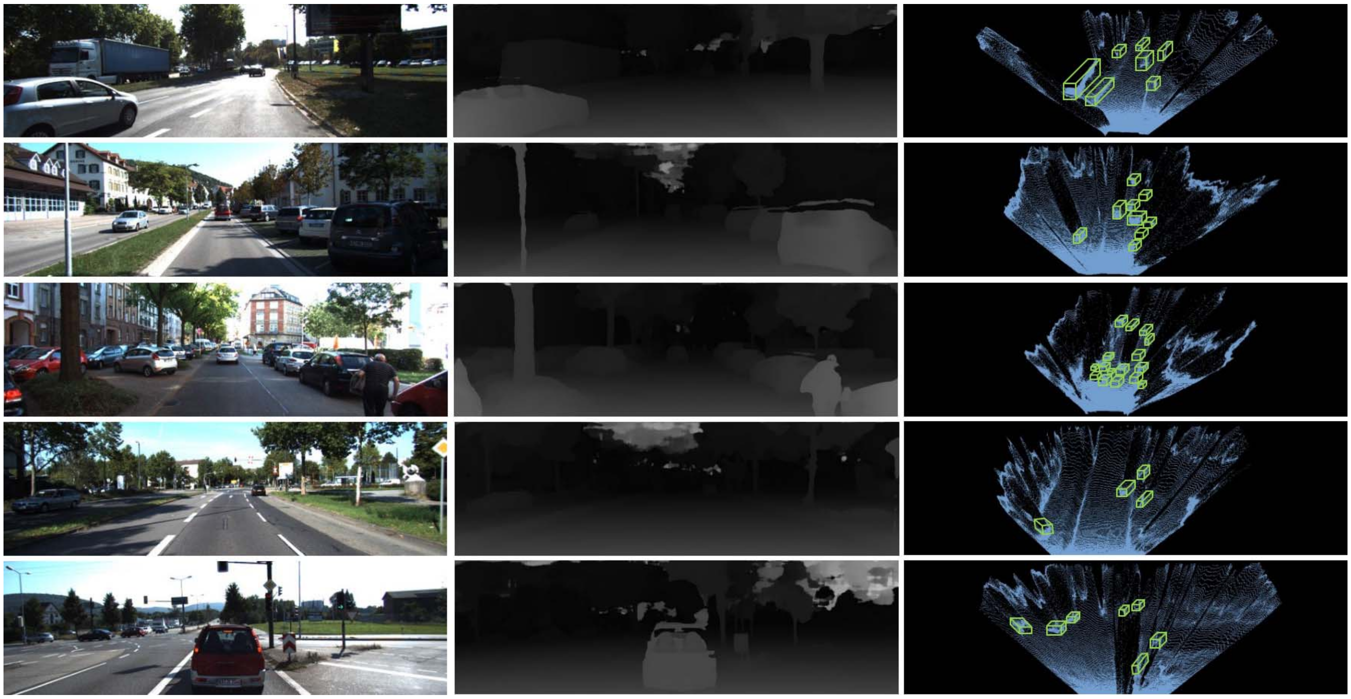


Fig. 3. The genereted examples of pseudo LiDAR according to the training data in KITTI. The original, depth image and pseudo LiDAR are listed in left, middle and right column, respectively. Best view in color.

is firstly sub-sampled according to the calibration matrix, and we concatenate it $\left(Q=(m,n,t)\in\mathbb{R}^{N\times 3}\right)$ with the raw LiDAR $P = (x,y,z) \in \mathbb{R}^{N\times 3}$ in a point-wise manner. Considering the differences in coordinate permutation, e.g. $Q\bigoplus P$ and $P\bigoplus Q$, we feed two $N\times 6$ point vectors into independent fully-connected architectures to capture global

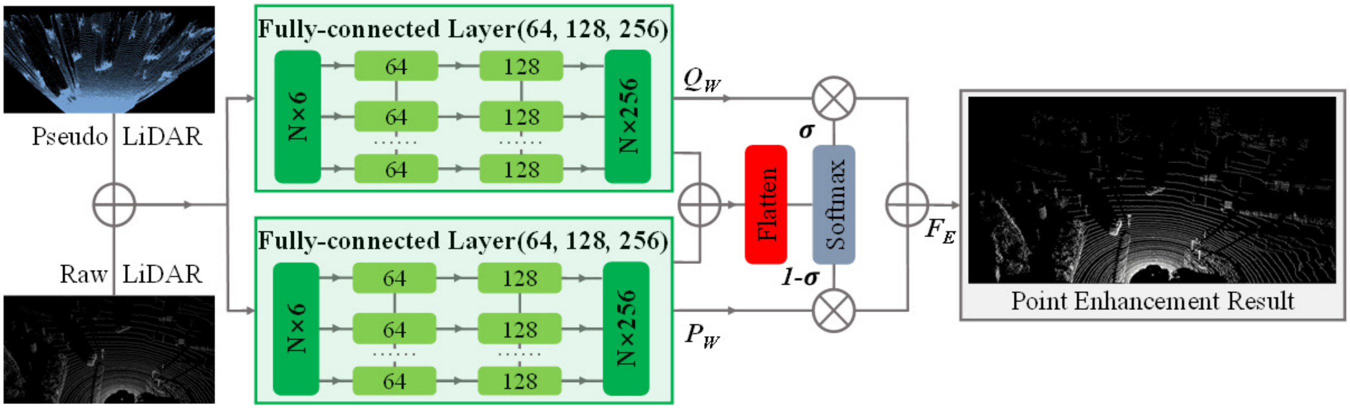


Fig. 4. Schematic diagram of point enhancement module (PEM). At first, the pseudo LiDAR is sub-sampled and combined with the raw LiDAR, with respect to the 3D coordinate. We then feed two $N \times 6$ vectors (N stands for the number of points), i.e. $Q \bigoplus P$ and $P \bigoplus Q$, into fully-connected blocks to obtain global feature responses Q_W and P_W, respectively. Through the softmax function, activation probability σ is produced to present the importance of feature channel. Finally, we reweight two feature branches by the element-wise product operation, and concatenate them to generate the point enhancement result. Best viewed in color.

responses Q_W and P_W in the high-dimensional feature space, respectively. To exploit the more significant feature information, two 256-dimensional representations are concatenated and flattened into one vector before the softmax function. The activation probability σ is regarded as a weighting parameter to evaluate the discriminability of feature channel. Finally, we reweight both point features by product operation, and aggregate them to obtain the point enhancement result in an element-wise concatenation. The whole process can be mathematically described as Eq.4-Eq.6.

$$Q_W = W_2^Q \Big(W_1^Q \Big(Q \qquad P \tag{4}$$

$$P_W = W_2^P \Big(W_1^P \Big(P \qquad Q \tag{5}$$

$$F_E = \sigma Q_W \qquad (1 - \sigma) P_W \tag{6}$$

where N is the number of point cloud, σ presents the softmax function, W_1^* and W_2^* are the weight parameters of fully-connected layers, $\bigoplus$ denotes the element-wise concatenation, and F_E implies the point enhancement result. The introduction of pseudo point provides dense RGB channel semantics for the raw LiDAR feature enhancement. More importantly, the PEM design can reweight the significance of different point channels adaptively, thus resulting in more robust and discriminative feature representations.

B. Image and Point Backbone

As depicted in Fig.2, we propose a two-stream architecture to encode image and point features, respectively. To be specific, image backbone has four convolutional blocks, each of which contains two 3×3 convolutions with residual connection, followed by batch normalization (BN) and ReLU activation function. In each block, the second convolution is with stride 2 to downsample the resolution of feature map and enlarge the receptive field simultaneously. And four de-convolution layers are further utilized to recover the object details, and multi-scale image feature maps are generated with dense semantics.

For point backbone, we adopt PointNet++ [11] architecture, that contains four set abstractions with a scale of 4096, 1024, 256 and 64 for adaptive point feature aggregation under increasing contextual scales. After that, four feature propagation layers project the sub-sampled points back into the original space. In this way, geometric correlation between local and global points can be explored, and we would perform multi-modal feature fusion under different resolution.

C. Point-Guided Fusion Module

As demonstrated in Fig.5, we propose point-guided fusion module (PFM) to find better image-point correspondence for perfect multi-modal fusion. Specifically, we project each point onto the image feature map via calibration matrix to obtain point-wise pixel correspondence. To consider the effect of adjacent pixels, bilinear interpolation is further utilized to capture local semantic features for each point. This process can be mathematically described as follows Eq.7: where I is the bilinear interpolation function, M is the calibration matrix, F_P is the point feature, F_I defines the point-wise correspondence on the image plane, and $\bigotimes$ denotes the element-wise multiplication.

$$F_I = I \Big(M \bigotimes F_P \tag{7}$$

After that, each point and its pixel correspondence are fed into fully-connected layers, respectively, and we adopt sigmoid function δ to compress the feature vector into the range of [0, 1]. This can be regarded as an attention mechanism that enforces the model to pay more attention to the significant image feature region. The joint feature map F_{joint} is finally obtained by concatenating two feature maps in an element-wise manner. We denote F_{add} as the feature addition result, and the whole process is formulated in Eq.8-Eq.9:

$$F_{add} = W_2^P \Big(W_1^P F_P \quad + W_2^I \Big(W_1^I F_I\Big) \tag{8}$$

$$F_{joint} = F_P \qquad \Big(F_I \bigotimes \delta\left(F_{add}\right) \tag{9}$$

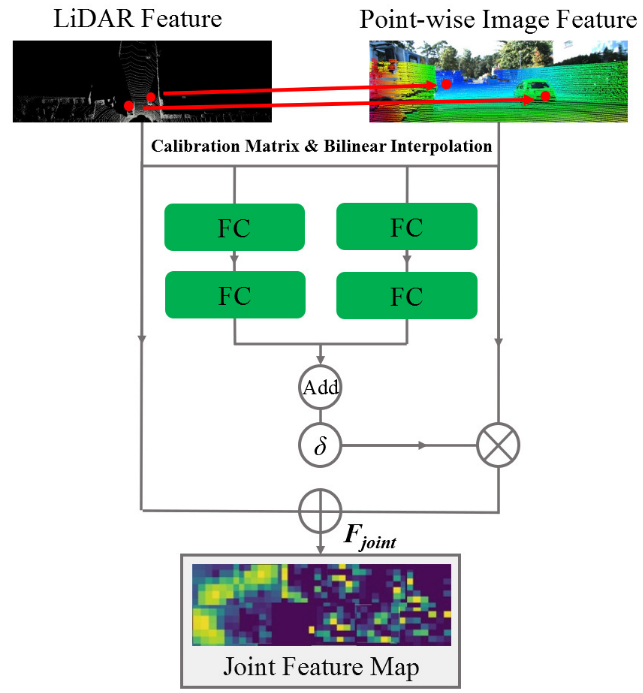


Fig. 5. Schematic diagram of point-guided fusion module (PFM). After finding point-wise image correspondences via calibration matrix and bilinear interpolation, LiDAR and point-wise image features are fed into two fully-connected layers respectively. We adopt sigmoid function to squeeze the feature vector into [0, 1], and the joint feature map is obtained by concatenating point and the updated image features in an element-wise manner. Noted that FC denotes fully-connected layer.

We totally introduce five PFMs into two-stream architecture at different resolutions: four PFMs are set between point and image backbones, to constructs the relationship of each pair of point abstraction and convolutional feature map; and the other PFM is utilized to fuse the final image and point feature representations.

D. Refinement Network

As the previous works [12], [40] described, we also apply non-maximum suppression (NMS) algorithm to choose high-quality proposals for the box refinement stage. To be specific, we keep 8000 proposals generated by our two-stream architecture according to the classification confidence, and randomly select 512 points from each proposal as the corresponding feature representation that feeds into box refinement network.

The refinement network consists of three set abstraction layers with a group size of 128, 64 and 32 to learn representative feature descriptor from each proposal, and three parallel branches with two cascaded 1 × 1 convolution layers for object classification, regression and IoU-aware prediction respectively. Commonly, classification branch predicts object confidence score, and regression branch is responsible for measuring localization accuracy. To consider the inconsistence problem between classification and localization, we additionally design IoU-aware prediction head (IoU Head) in parallel with box regression and classification branches, to calculate the IoU between each ground-truth and predicted box. Noted that a sigmoid activation function is utilized in the IoU Head to ensure the value is between 0 and 1. During training, we jointly optimize three branches; and in the inference, confidence score is multiplied by the IoU value of each predicted box for ranking all detections in NMS and average precision (AP) computation procedure. Therefore, this simple design can alleviate the misalignment between classification score and localization accuracy effectively, and we would describe its effect on performance improvement in the ablation study.

E. Loss Function

We adopt multi-task loss function to jointly optimize CL3D architecture. Overall, the total loss L can be formulated as Eq.10:

$$L = L_{reg} + L_{cls} + L_{dir} + L_{iou} \tag{10}$$

Specifically, we firstly parameterize the 3D ground-truth box as $(x_g, y_g, z_g, l_g, w_g, h_g, \theta_g)$, where (x_g, y_g, z_g) denote the center coordinate of bounding box in 3D space, (l_g, w_g, h_g) define the size of bounding box, and θ_g is the yaw rotation along the z-axis. Correspondingly, the 3D prior box can be described as $(x_a, y_a, z_a, l_a, w_a, h_a, \theta_a)$. Therefore, the residual vector Δr= (Δx, Δy, Δz, Δl, Δw, Δh, $\Delta\theta$) in 3D box regression box can be computed as Eq.11:

$$\begin{aligned} &\Delta x = \frac{x_g - x_a}{d_a}, \quad \Delta y = \frac{y_g - y_a}{d_a}, \quad \Delta z = \frac{z_g - z_a}{d_a}, \\ &\Delta l = \log(\frac{l_g}{l_a}), \quad \Delta w = \log(\frac{w_g}{w_a}), \quad \Delta h = \log(\frac{h_g}{h_a}), \\ &\Delta\theta = \theta_g - \theta_a, \quad d_a = \overline{w_a^2 + l_a^2} \end{aligned} \tag{11}$$

We utilize *Smooth-L*1 function to calculate regression loss L_{reg} for positive predictions N_{pos}, and the expression is described as Eq.12:

$$L_{reg} = \frac{1}{N_{pos}} {}_{i} \; Smooth_{L1}(\Delta r) \tag{12}$$

For classification loss L_{cls}, we adopt focal loss [49] to alleviate the foreground-background imbalance problem. The formulation is presented as Eq.13: where p_i denotes the confidence score for i-th box, hyperparameter $\alpha = 0.25$ and $\gamma = 2$.

$$L_{cls} = \frac{1}{N_{pos}} {}_{i} \; -\alpha(1 - p_i)^{\gamma} \log(p_i) \tag{13}$$

As for direction loss L_{dir}, we use bin-based loss following the PointRCNN [12], which predicts the centroid coordinate and regresses its offset for each bin. We also calculate IoU loss L_{iou} by introducing binary cross-entropy function, which can be mathematically described as Eq.14. The I_{gt} denotes target IoU that is computed between positive prediction and ground-truth box, and I_p is the predicted IoU for each detected box.

$$L_{iou} = \frac{1}{N_{pos}} {}_{i} \; -I_{gt} \log(I_p) \tag{14}$$

TABLE I

3D AP Performance of CL3D and the State-of-the-Art Detectors on KITTI Test Split. Note That All 3D Detection Methods in KITTI Leaderboard Are Ranked by the AP Performance at Moderate Level With the IoU Threshold of 0.7. the Abbreviate of 'M', 'L' and 'C+L' Denote the 'Modalities', 'LiDAR-only' and 'Camera-LiDAR', Respectively

M	Algorithms	Car (AP%)			Pedestrian (AP%)			Cyclist (AP%)			FPS
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard	
L	Pointpillar [15]	82.58	74.31	68.99	51.45	41.92	38.89	77.10	58.65	51.92	**62**
	PointRCNN [12]	85.94	75.76	68.32	49.43	41.78	38.63	73.93	59.60	53.59	12
	TANet [33]	84.39	75.94	68.82	**53.72**	**44.34**	**40.49**	75.70	59.44	52.53	*28*
	STD [32]	87.95	79.71	*75.09*	*53.29*	42.47	38.35	*78.69*	61.59	55.30	12
	Part-A^2 [36]	87.81	78.49	73.51	53.10	*43.35*	40.06	**79.17**	*63.52*	*56.93*	12
	PV-RCNN [16]	**90.25**	**81.43**	**76.82**	52.17	43.29	*40.29*	78.60	**63.71**	**57.65**	12
C+L	F-PointNet [39]	82.19	69.79	60.59	50.53	42.15	38.08	72.27	56.12	49.01	5
	AVOD-FPN [18]	83.07	71.76	65.73	50.46	42.27	39.04	63.76	50.55	44.93	10
	PI-RCNN [22]	84.37	74.82	70.03	-	-	-	-	-	-	10
	EPNet [40]	*89.81*	79.28	74.59	-	-	-	-	-	-	10
	3D-CVF [23]	89.20	80.05	73.11	-	-	-	-	-	-	12
	CL3D[1]	87.45	*80.28*	*76.21*	47.30	39.42	36.97	77.33	62.02	55.52	10

IV. Experiments

A. Dataset and Implementation Details

1) Dataset: KITTI dataset [25] is one of the most popular benchmark of 3D object detection for autonomous driving, which contains 7481 training and 7518 test samples, respectively. Here, we further divide training set into train split with 3712 examples and val split with 3769 examples, as commonly done in previous work [17]. For fair comparison, we also follow the official evaluation protocol, and adopt average precision from 40-point precision-recall (PR) curve as the evaluation metric. For model evaluation, We would provide the 3D detection performance of CL3D and the state-of-the-art 3D detectors under three difficulties (i.e. easy, moderate and hard) on both validation and test split.

2) Inplementation Details: For point backbone, the range of input LiDAR is limited to $[0, 70.4] \times [-40, 40] \times [-1, 3]m$ in LiDAR coordinate, and we randomly subsample 16384 points as input. Moreover, image backbone takes RGB image with a size of 1280×384 pixels as input. Based on NVIDIA TITAN RTX GPUs, the overall architecture is trained using ADAM optimizer with the batch size 12 from the initial learning rate 1e-3 for 80 epochs, and cosine annealing strategy is adopted to decay the learning rate. The weight decay and momentum factor are set to 0.002 and 0.9, respectively.

3) Data Augmentation: Considering the modality difference between image and point cloud, it is challenging to guarantee the precise pixel-point correspondence after spatial augmentated transformation [22]. Consequently, we do not perform data augmentation during training, which is different from many LiDAR-only or camera-LiDAR methods as mentioned above.

TABLE II

3D Car AP Performance of CL3D and the State-of-the-Art Detectors on KITTI Val Split. The Abbreviate of 'M', 'L' and 'C+L' Denote the 'Modalities', 'LiDAR-only' and 'Camera-LiDAR', Respectively

M	Algorithms	Car AP (%)		
		Easy	Moderate	Hard
L	PointPillars [15]	83.62	75.22	72.40
	PointRCNN [12]	87.07	77.83	72.18
	TANet [33]	85.27	77.64	72.13
	STD [32]	89.74	80.96	77.81
	Part-A^2 [36]	89.33	81.59	76.05
	PV-RCNN [16]	**92.95**	**84.26**	**79.32**
C+L	AVOD-FPN [18]	76.02	67.93	60.29
	F-PointNet [39]	84.69	73.41	64.27
	EPNet [40]	*91.96*	81.54	77.16
	CL3D	90.32	*83.20*	*78.91*

B. Evaluation Results on KITTI Dataset

We adopt PointRCNN [12] as the baseline, and compare our proposed CL3D to the state-of-the-art 3D detectors on test split of KITTI benchmark, as seen in Table I. Generally, CL3D achieves prominent 3D detection performance, and outperforms the baseline by a large margin. For instance, it improves the PointRCNN by $2\% - 8\%$ AP on car class and $2\% - 4\%$ AP on cyclist class at three different levels, respectively.

Compared to other camera-LiDAR 3D detectors, CL3D reports top 3D detection perofrmance particularly in car and cyclist classes, e.g., 80.28% and 76.21% car AP, 62.02% and 55.52% cyclist AP at moderate and hard levels, which surpasses over other multi-modal detectors substantially. Considering the difficulty level classified by the size, occlusion

[1] The method remarked by 'FusionDetv1' on the online KITTI leaderboard corresponds to the proposed 'CL3D' in this work.

TABLE III

The 3D Detection Performance Improvements Provided by Different LiDAR Signals and Feature Combination Methods on KITTI Val Split. Noted That the Accuracy Gains Are Drawn in Red Color in the Brackets

Components	Raw LiDAR		✓	✓		
	Pseudo LiDAR				✓	✓
	Concatenation		✓		✓	
	PEM			✓		✓
Car AP (%)	Easy	87.07	87.39(+0.32)	87.56(+0.49)	87.47(+0.40)	**87.69**(+0.57)
	Moderate	77.83	78.57(+0.74)	78.93(+1.10)	78.82(+0.99)	**79.24**(+1.35)
	Hard	72.18	73.10(+0.92)	73.47(+1.29)	73.37(+1.19)	**73.82**(+1.56)

TABLE IV

Ablation Studies of Each Component of CL3D on KITTI Val Split. Noted That the Performance Gains Provided by Each Component Are Highlighted in Red Color in the Brackets. PL Denotes the Pseudo LiDAR

Components	PEM (& PL)		✓	✓	✓
	PFM			✓	✓
	IoU Head				✓
Car AP (%)	Easy	87.07	87.69(+0.57)	89.21(+2.09)	**90.32**(+3.20)
	Moderate	77.83	79.24(+1.35)	81.78(+3.89)	**83.20**(+5.31)
	Hard	72.18	73.82(+1.56)	77.02(+4.76)	**78.91**(+6.65)

and truncation of object in KITTI dataset, remarkable 3D AP results indicate the significant detection performance of CL3D in such difficult scenes. Furthermore, CL3D runs at 10 FPS in an inference, which performs better or on par with these cross-modal methods in terms of detection speed. We assume that feature concatenation in PEM and point-wise projection in PFM might be time-consuming, and detailed runtime analysis would be explored to speed up 3D object detection in the future.

As for LiDAR-only detectors, PV-RCNN [16] is one of the state-of-the-art (SOTA) 3D detectors on KITTI benchmark, and our CL3D is slightly inferior to it by 1.15% and 0.61% car AP at moderate and hard levels, respectively. Moreover, CL3D still presents better performance than other LiDAR-only method, i.e., Part-A^2, with 1.79% and 2.70% AP gains on car class at middle and difficult levels. However, there is a large performance gap in pedestrian class between CL3D and other SOTA algorithms. Due to the feature ambiguity and similarity between pedestrian and other instances (e.g., traffic pole), our CL3D easily suffers from false positives in pedestrian detection. Moreover, it is assumed that inaccurate image-point correspondence occurred in certain location might confuse the model and lead to poor detection results. We would investigate these problems and provide a solution for more robust and accurate performance in the future work.

Furthermore, we also evaluate the 3D detection performance on val split of KITTI dataset. For simplicity, we just elaborate 3D AP performance on car class in Table II. It is clear that CL3D still offers remarkable performance gains over the baseline, and outperforms all cross-modal methods by a significant margin, demonstrating its effectiveness and competitiveness. Finally, we visualize the predicted boxes achieved by CL3D on KITTI test split in Fig.6, and accurate detection results can be observed even in occluded or crowded scenes.

C. Ablation Studies

Ablation studies are conducted to investigate the effect of each component for the final detection performance. All experiments are performed on KITTI val split, and we list the 3D car detection results in Table III and Table IV. Notably, the baseline PointRCNN [12] achieves 87.07%, 77.83% and 72.18% car AP in three difficulty levels. We would append pseudo LiDAR and point enhancement module (PEM), point-guided fusion module (PFM) and IoU head, to evaluate the contribution to 3D detection result respectively.

1) Pseuo LiDAR and Point Enhancement Module (PEM): We measure the effectiveness of various LiDAR signals and feature aggregation methods for 3D detection performance. As tabulated in Table III, the combination of pseudo LiDAR and PEM achieves the outstanding performance gains over other counterparts, which improves the baseline by 0.57%, 1.35% and 1.56% in easy, moderate and hard levels, respectively. When adopting the raw LiDAR with PEM, the performance is slightly inferior to that of our method, which implies the channel semantics within RGB image would be complementary with geometric information of the raw LiDAR, and the pseudo signal is preferable to 3D object detection. Moreover, incorporating the pseudo LiDAR with simple feature concatenation only brings in marginal improvements, and it demonstrates the significance reweighting mechanism in PEM can be indeed beneficial to learn more discriminative features and to facilitate the detection accuracy.

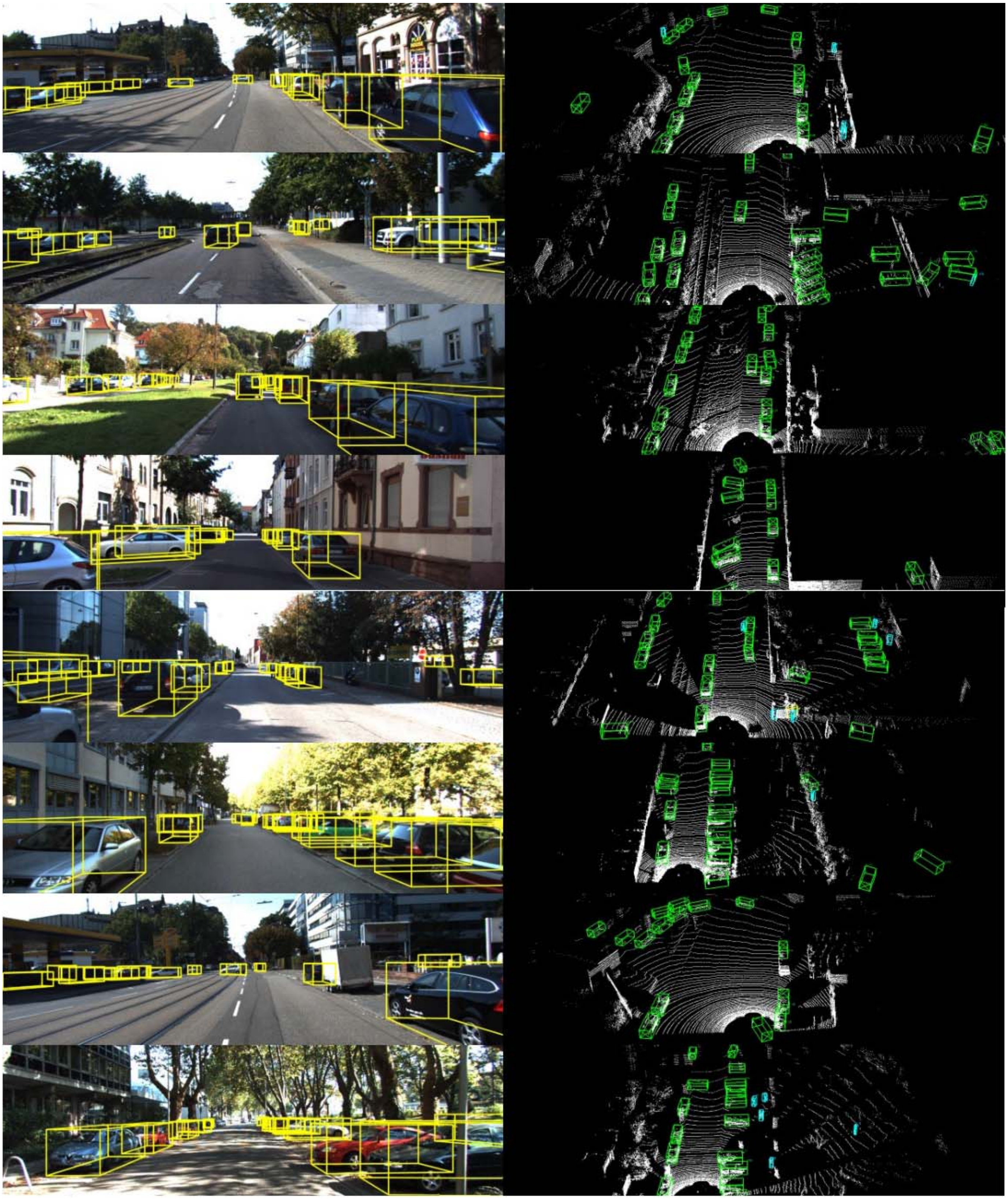


Fig. 6. The visualization results achieved by CL3D on KITTI test split. We list a pair of RGB image and its corresponding LiDAR in each row. Accurate and robust detection results in such occluded or crowded scenes demonstrate the effectiveness and competitiveness of CL3D.

2) Point-Guided Fusion Module (PFM): We additionally introduce PFM to aggregate multi-scale image and point features, and 2.09%, 3.89% and 4.76% AP gains in three difficulties can be seen in Table IV. The substantial improvements describe the effectiveness of PFM and the advantage of point-wise fusion manner. It can provide perfect image-point correspondence on the feature map, and combines semantic with geometric information to localize the object accurately in various scenes.

3) IoU-Aware Head (IoU Head): We finally add IoU Head in parallel with the localization and classification branches. Compared with the baseline, it totally leads to 3.20%, 5.31% and 6.65% boosts in AP_{Easy}, $AP_{Moderate}$ and AP_{Hard}, respectively. We argue that the design of IoU Head alleviates the misalignment problem between localization confidence and classification score, and simultaneously promotes to generate high-quality 3D detection boxes.

V. Conclusion

In this paper, we build up a novel camera-LiDAR 3D detection architecture termed as CL3D, with pseudo LiDAR and point enhancement module (PEM), point-guided fusion module (PFM) and IoU-aware prediction Head (IoU Head). Particularly, pseudo LiDAR is generated directly from RGB image, and we propose point enhancement module to incorporate it with the raw LiDAR, which alleviates the inherent sparsity of LiDAR and enhances point feature representation. Point-guided fusion module is designed to aggregate semantic and geometric information under different resolutions in a point-wise manner. To alleviate the inconsistence between localization accuracy and classification confidence, we further introduce IoU head to calculate the similarity between each ground-truth and predicted boxes. Comprehensive experiments are performed on the publicly available KITTI dataset, and our CL3D reports competitive 3D detection performance compared to the-state-of-the-art (SOTA) single- and cross-modal detectors. More importantly, CL3D presents great improvements over the baseline in ablation studies, demonstrating the effectiveness of each components.

However, there is a huge performance gap between CL3D and other SOTA methods particularly in pedestrian detection. We assume feature ambiguity or misalignment severely damages the detection accuracy, and in the future, an effective solution would be developed to eliminate this problem. Moreover, we also focus on designing more lightweight module for efficient multi-modal 3D detection task.

References

[1] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[2] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.

[3] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[4] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, and H. Ling, "M2det: A single-shot object detector based on multi-level feature pyramid network," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 9259–9266.

[5] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, May 2021.

[6] G. Song, Y. Liu, and X. Wang, "Revisiting the sibling head in object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10563–10572.

[7] X. Chen *et al.*, "3D object proposals for accurate object class detection," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 424–432.

[8] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3D bounding box estimation using deep learning and geometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7074–7082.

[9] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.

[10] C. R. Qi, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[11] 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. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5099–5108.

[12] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 770–779.

[13] Z. Yang, Y. Sun, S. Liu, and J. Jia, "3DSSD: Point-based 3D single stage object detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11040–11048.

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

[15] 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. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12698–12705.

[16] S. Shi *et al.*, "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10529–10538.

[17] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1907–1915.

[18] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. Waslander, "Joint 3D proposal generation and object detection from view aggregation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1–8.

[19] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3D object detection," in *Proc. Euporean Conf. Comput. Vis. (ECCV)*, 2018, pp. 641–656.

[20] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7345–7353.

[21] D. Xu, D. Anguelov, and A. Jain, "PointFusion: Deep sensor fusion for 3D bounding box estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 244–253.

[22] L. Xie *et al.*, "Pi-rcnn: An efficient multi-sensor 3D object detector with point-based attentive cont-conv fusion module," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 12467–12469.

[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. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 720–736.

[24] S. Pang, D. Morris, and H. Radha, "CLOCs: Camera-LiDAR object candidates fusion for 3D object detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 10386–10393.

[25] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[26] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3D object detection for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2147–2156.

[27] J. Ku, A. D. Pon, and S. L. Waslander, "Monocular 3D object detection leveraging accurate proposals and shape reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11867–11876.

[28] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8437–8445.

[29] X. Weng and K. Kitani, "Monocular 3D object detection with pseudo-LiDAR point cloud," 2019, *arXiv:1903.09847*.

[30] Y. You *et al.*, "Pseudo-LiDAR++: Accurate depth for 3D object detection in autonomous driving," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–22.

[31] R. Qian *et al.*, "End-to-end pseudo-LiDAR for image-based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5881–5890.

[32] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "STD: Sparse-to-dense 3D object detector for point cloud," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2019, pp. 1951–1960.

[33] 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 Conf. Artif. Intell. (AAAI)*, 2020, pp. 1–8.

[34] C. He, H. Zeng, J. Huang, X. Hua, and L. Zhang, “Structure aware single-stage 3D object detection from point cloud,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11873–11882.

[35] Y. Chen, S. Liu, X. Shen, and J. Jia, “Fast point R-CNN,” 2019, *arXiv:1908.02990*.

[36] 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.*, vol. 43, no. 8, pp. 2647–2664, Feb. 2020.

[37] Y. Zhou and O. Tuzel, “VoxelNet: End-to-end learning for point cloud based 3D object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4490–4499.

[38] T. Yin, X. Zhou, and P. Krahenbuhl, “Center-based 3D object detection and tracking,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11784–11793.

[39] 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. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 918–927.

[40] T. Huang, Z. Liu, X. Chen, and X. Bai, “Epnet: Enhancing point features with image semantics for 3d object detection,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 35–52.

[41] L. Tychsen-Smith and L. Petersson, “Improving object localization with fitness NMS and bounded IoU loss,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6877–6885.

[42] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, “Mask scoring R-CNN,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6409–6418.

[43] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, “Acquisition of localization confidence for accurate object detection,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 784–799.

[44] S. Wu, X. Li, and X. Wang, “IoU-aware single-stage object detector for accurate localization,” 2019, *arXiv:1912.05992*.

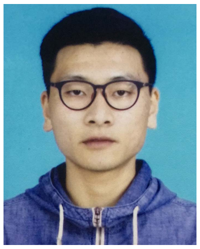
[45] S. Wu, J. Yang, X. Wang, and X. Li, “IoU-balanced loss functions for single-stage object detection,” 2019, *arXiv:1908.05641*.

[46] J. Li *et al.*, “3D IoU-Net: IoU guided 3D object detector for point clouds,” 2020, *arXiv:2004.04962*.

[47] W. Zheng, W. Tang, S. Chen, L. Jiang, and C.-W. Fu, “CIA-SSD: Confident IoU-aware single-stage object detector from point cloud,” 2020, *arXiv:2012.03015*.

[48] J. R. Chang and Y. S. Chen, “Pyramid stereo matching network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5410–5418.

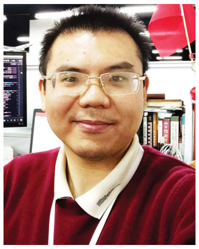
[49] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.



Chunmian Lin is currently pursuing the Ph.D. degree with the School of Transportation Science and Engineering, Beihang University, Beijing, China. His current research interests include autonomous driving, image processing, computer vision, artificial intelligence, and deep learning, particularly their applications in intelligent transportation systems.

Daxin Tian (Senior Member, IEEE) received the Ph.D. degree in computer application technology from Jilin University, Changchun, China, in 2007. He is currently a Professor with the School of Transportation Science and Engineering, Beihang University, Beijing, China. His current research interests include mobile computing, intelligent transportation systems, vehicular *ad hoc* networks, and swarm intelligent. He is a member of IEEE Intelligent Transportation Systems Society and IEEE Vehicular Technology Society. He was awarded the Changjiang Scholars Program (Young Scholar) of Ministry of Education of China in 2017, the National Science Fund for Distinguished Young Scholars in 2018, and the Distinguished Young Investigator of China Frontiers of Engineering in 2018. He served as the Technical Program Committee Member/the Chair/the Co-Chair for several international conferences, including EAI 2018, ICTIS 2019, IEEE ICUS 2019, IEEE HMWC 2020, and GRAPH-HOC 2020.

Xuting Duan (Member, IEEE) received the Ph.D. degree in traffic information engineering and control from Beihang University, Beijing, China. He is currently an Assistant Professor with the School of Transportation Science and Engineering, Beihang University. His current research interests include vehicular *ad hoc* networks, cooperative vehicle infrastructure systems, and the Internet of Vehicles.



Jianshan Zhou received the B.Sc., M.Sc., and Ph.D. degrees in traffic information engineering and control from Beihang University, Beijing, China, in 2013, 2016, and 2020, respectively. From 2017 to 2018, he was a Visiting Research Fellow with the School of Informatics and Engineering, University of Sussex, Brighton, U.K. He is currently a Post-Doctoral Research Fellow supported by the Zhuoyue Program, Beihang University, and the National Postdoctoral Program for Innovative Talents. He is the author or coauthor of more than 20 international scientific publications. His research interests include wireless communication, artificial intelligent systems, intelligent transportation systems, the modeling and optimization of vehicular communication networks and air–ground cooperative networks, the analysis and control of connected autonomous vehicles, and intelligent transportation systems. He was a recipient of the First Prize in the Science and Technology Award from the China Intelligent Transportation Systems Association in 2017, the First Prize in the Innovation and Development Award from the China Association of Productivity Promotion Centers in 2020, the National Scholarships in 2017 and 2019, the Outstanding Top-Ten Ph.D. Candidate Prize from Beihang University in 2018, the Outstanding China-SAE Doctoral Dissertation Award in 2020, and the Excellent Doctoral Dissertation Award from Beihang University in 2021. He was the Technical Program Session Chair with the IEEE EDGE 2020, the TPC Member with the IEEE VTC2021-Fall track, and the Youth Editorial Board Member of the Unmanned Systems Technology.

Dezong Zhao (Senior Member, IEEE) received the B.Eng. and M.S. degrees in control science and engineering from Shandong University, Jinan, China, in 2003 and 2006, respectively, and the Ph.D. degree in control science and engineering from Tsinghua University, Beijing, China, in 2010. He is currently a Senior Lecturer in autonomous systems with the School of Engineering, University of Glasgow, U.K. His research interests include connected and autonomous vehicles, machine learning, and control engineering. His work has been recognised by being awarded an EPSRC Innovation Fellowship in 2018 and a Royal Society-Newton Advanced Fellowship in 2020.

Dongpu Cao received the Ph.D. degree from Concordia University, Canada, in 2008. He is currently an Associate Professor and the Director of the Waterloo Cognitive Autonomous Driving (CogDrive) Laboratory, University of Waterloo, Canada. He is the Canada Research Chair in driver cognition and automated driving. His research interests include driver cognition, automated driving, and cognitive autonomous driving. He has contributed more than 200 papers and three books. He received the SAE Arch T. Colwell Merit Award in 2012, IEEE VTS 2020 Best Vehicular Electronics Paper Award, and three Best Paper Awards from the ASME and IEEE conferences. He serves as the Deputy Editor-in-Chief for *IET Intelligent Transport Systems* journal and an Associate Editor for IEEE Transactions on Vehicular Technology, IEEE Transactions on Intelligent Transportation Systems, IEEE/ASME Transactions on Mechatronics, IEEE Transactions on Industrial Electronics, IEEE/CAA Journal of Automatica Sinica, IEEE Transactions on Computational Social Systems, and *Journal of Dynamic Systems, Measurement and Control* (ASME). He was a Guest Editor of *Vehicle System Dynamics*, Systems, Man, and Cybernetics—Part A: Systems, and IEEE Internet of Things Journal. He serves on the SAE Vehicle Dynamics Standards Committee and acts as the Co-Chair of IEEE ITSS Technical Committee on Cooperative Driving. He is an IEEE VTS Distinguished Lecturer.