sustainability

MDPI

Article

Center-Aware 3D Object Detection with Attention Mechanism Based on Roadside LiDAR

Haobo Shi [1,2,3], Dezao Hou [1,2,3,*] and Xiyao Li [1,2,3]

[1] Research Institute of Highway, Ministry of Transport, Beijing 100088, China
[2] Key Laboratory of Intelligent Transportation Technology and Transportation Industry, Beijing 100088, China
[3] National Intelligent Transport Systems Center of Engineering and Technology, Beijing 100088, China
* Correspondence: dz.hou@rioh.cn

Abstract: Infrastructure 3D Object Detection is a pivotal component of Vehicle-Infrastructure Cooperated Autonomous Driving (VICAD). As turning objects account for a high proportion of traffic at intersections, anchor-free representation in the bird's-eye view (BEV) is more suitable for roadside 3D detection. In this work, we propose CetrRoad, a simple yet effective **ce**nter-aware detector with **tr**ansformer-based detection head for **road**side 3D object detection with single LiDAR (Light Detection and Ranging). CetrRoad firstly utilizes a voxel-based roadside LiDAR feature encoder module that voxelizes and projects the raw point cloud into BEV with dense feature representation, following a one-stage center proposal module that initializes center candidates of objects based on the top N points in the BEV target heatmap with unnormalized 2D Gaussian. Then, taking attending center proposals as query embedding, a detection head with multi-head self-attention and multi-scale multi-head deformable cross attention can refine and predict 3D bounding boxes for different classes moving/parked at the intersection. Extensive experiments and analyses demonstrate that our method achieves state-of-the-art performance on the DAIR-V2X-I benchmark with an acceptable training time cost, especially for Car and Cyclist. CetrRoad also reaches comparable results with the multi-modal fusion method for Pedestrian. An ablation study demonstrates that center-aware query as input can provide denser supervision than a purified feature map in the attention-based detection head. Moreover, we were able to intuitively observe that in complex traffic environment, our proposed model could produce more accurate 3D detection results than other compared methods with fewer false positives, which is helpful for other downstream VICAD tasks.

Keywords: vehicle-infrastructure cooperative autonomous driving; roadside 3D detection; LiDAR-based detection; central point representation; deformable attention

Citation: Shi, H.; Hou, D.; Li, X. Center-Aware 3D Object Detection with Attention Mechanism Based on Roadside LiDAR. *Sustainability* **2023**, *15*, 2628. https://doi.org/10.3390/su15032628

Academic Editors: Shaopeng Zhong and Hongmei Zhou

Received: 21 December 2022
Revised: 19 January 2023
Accepted: 30 January 2023
Published: 1 February 2023

1. Introduction

In recent years, the driving situation perception for autonomous vehicles has become one of the most important applications of computer vision. Accurate detection results can effectively improve the safety of autonomous vehicles as well as transportation efficiency. The vast majority of current research is based on vehicle-side sensors, such as monocular or binocular cameras, LiDAR, etc. Due to the inherent disadvantage of installation height for on-board sensors, however, even well-performing object detection algorithms cannot effectively identify occluded objects from the ego-vehicle perspective. Conversely, data captured from roadside sensors has the intrinsic advantages of occlusion robustness and object detection for dense traffic flow, as they are collected by sensors installed on traffic poles with a certain height. Intelligent transportation systems using external infrastructure with modern sensors could offer great potential and the possibility to support connected vehicles and autonomous driving [1].

Roadside perception is crucial to the successful operation of Vehicle-Infrastructure Cooperated Autonomous Driving (VICAD), which is an advanced form of Autonomous Driving (AD). With the development of roadside perception, automated vehicles will no longer need as many perception sensors and high-performance computing units as before. Object detection based on Deep Learning (DL) has been developed for a long time, and the mean average precision (mAP) from the perspective of the vehicle on the nuScenes [2] dataset has reached more than 75%. Until now, most research has mainly focused on the ego-vehicle perspective, and current detection algorithms from the roadside perspective are generally based on typical machine vision. The traditional pipeline of roadside perception with LiDAR sensors usually follows three steps: background filtering, feature clustering and object classification. This requires many steps of calculation and analysis, and cannot satisfy the real-time requirements in mixed-traffic environments. Although general 3D object detection has gone through an era of rapid development, research based on roadside LiDAR is still an emerging topic and has the potential to provide detection results complementary to the on-board sensors for VICAD.

Both anchor-based and transformer-based detectors have been widely used in 3D object detection for autonomous driving. However, their performance gap and applications based on roadside sensors remain to be studied. As one of the most popular anchor-free detection baselines, center-based representation has several key advantages for roadside detection. Firstly, left- or right-turning objects generally account for a high proportion at an intersection. Although both the anchor-based and center-based method could detect vehicles going straight accurately, bounding boxes of rotated objects are difficult to refine based on axis-aligned anchors. Secondly, centerpoint-based representation will considerably reduce the search space and training cost of detection models and allows the LiDAR backbone to learn the rotational invariance and equivalence of objects [3]. Thirdly, center-based detection results with voxel feature extraction enable more effective multi-sensor fusion in the shared bird's-eye view (BEV) representation space. Furthermore, as a kind of middle-feature representation, center candidates of traffic participants could provide more accurate query embedding for attention-based blocks.

Taking the aforementioned aspects into consideration, we propose CetrRoad, a simple yet effective **ce**nter-aware detector with deformable **tr**ansformer for **road**side Single-View 3D (SV3D) Object Detection. A simplified overview of CetrRoad is shown in Figure 1. Overall, our contributions could be summarized as follows: (a) To the best of our knowledge, motivated by real-world roadside datasets, CetrRoad is the first transformer-based detector with center-aware proposals for infrastructure LiDAR-only 3D object detection; (b) the proposed center-aware query embedding could provide dense supervision to bounding box refinement and prediction, which makes the performance of our model outstanding among the compared anchor-based methods and transformer-based detectors without dense query input; (c) the designed detection head with multi-head deformable cross-attention blocks was able to efficiently aggregate sampled features of center candidates on multi-scale BEV feature maps; (d) extensive experiments and analyses demonstrate that CetrRoad achieves a new state-of-the-art performance on the DAIR-V2X-I dataset, with fewer false positive predictions in *Car* and *Cyclist*, as well as results comparable to a multi-modal model in *Pedestrian*, which is helpful for other downstream VICAD tasks.

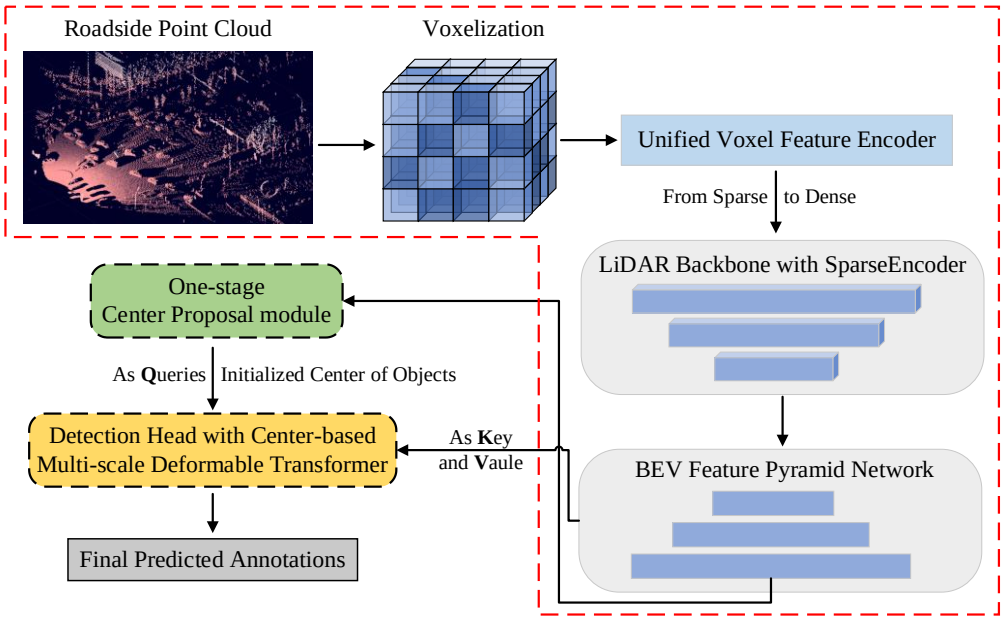

Figure 1. The pipeline of our CetrRoad framework. CetrRoad is a simple yet effective center-aware detector with deformable cross-attention for LiDAR-only 3D object detection from the roadside perspective. CetrRoad firstly utilizes a voxel-based roadside LiDAR feature encoder module that voxelizes and projects the raw point cloud into BEV with dense feature representation, following a one-stage center proposal module that initializes center candidates of objects based on the top N points in the BEV target heatmap with unnormalized 2D Gaussian. Then, taking attending center proposals as query embedding, a detection head with multi-head self-attention and multi-scale multi-head deformable cross attention could refine and predict 3D bounding boxes for different classes moving/parked at the intersection. More details about submodules can be found in Section 3.

2. Related Work

2.1. Camera-Based Roadside Detection

Cameras are the most common sensor at the roadside due to their low cost and ease of deployment. For cooperative vehicle infrastructure systems (CVIS), a monocular 3D vehicle detection method [4] was proposed without the need for 3D labels in the contour of the vehicle. It consists of three steps: (1) clustering arbitrary object contours into linear equations with instance segmentation and image gradients; (2) estimating the position, orientation and dimensions of the vehicle regardless of it being moving or stationary by applying a K-means-like method; (3) fine-tuning the final 3D object detection results by maximizing the posterior probability of previous 2D results.

Optical and thermal cameras installed on the roadside could be utilized to establish a roadside detection system for cooperative autonomous driving, including 3D object detection, tracking, and camera data fusion [5]. The training of this model only relies on 2D ground-truth annotations with a localization strategy motivated by landmark reference. Then, 2D predictions would be transformed into 3D with the landmark unicity and multi-camera intrinsic. Due to the efficient support of MobileNet-v2 [6], the whole framework could operate timely on the roadside computing unit with a transmission delay of less than 20 ms.

2.2. LiDAR-Based Roadside Detection

Due to a lack of large real-world roadside datasets with annotations, most previous studies still follow a typical pipeline to realize LiDAR-based object detection, which mainly consists of three steps: (1) Background Filtering to purify the LiDAR points reflected from the road surface or buildings by applying filtering methods, such as 3D density statistic filtering (3D-DSF) for both statistic and actional background [7]; (2) clustering to generate

clusters for the laser points by implementing clustering methods, such as Density-Based Spatial Clustering Applications with Noise (DBSCAN) [8] for large spatial LiDAR data and Multi-Rectified DBSCAN to identify traffic markings; and (3) classification to predict different labels for vehicles and pedestrians in traffic scenes by back propagation (BP) neural networks [9]. This pipeline is logically clear and applicable for implementation on the roadside. But the inference speed cannot meet real-time requirements and the performance of detection has a large gap compared with deep-learning-based paradigms. For pedestrian detection based on roadside LiDAR, Gong et al. proposed a real-time pedestrian detection algorithm by combining traditional and deep learning algorithms with high reliability in practical application [10]. To satisfy the real-time requirement, Octree with region-of-interest (ROI) selection and an improved Euclidean clustering algorithm with adaptive search radius were introduced. The total process of background filtering and clustering takes 88.7 ms per frame, and the final inference time reaches 110 ms per frame.

As a novel real-time traffic surveillance system for exploring the potential of roadside sensors for enabling CDA in the real world, Cyber Mobility Mirror (CMM) [11] can utilize a roadside 3D LiDAR for data collection and 3D object perception. The roadside point clouds should be first transformed into a vehicle-side coordinate system by Roadside Point-cloud Encoder and Decoder (RPEaD). Then the detection network will be trained with open-source onboard dataset configurations, e.g., nuScenes [2], while inferencing on the roadside. To eliminate the threats of large shifting along the z-axis, only voxelization on the x–y plane was performed to produce point cloud pillars following the strategy applied in PointPillars [12]. Next, the aggregated features will be sent to Feature Pyramid Network (FPN) followed by Single Shot multi-box Detector (SSD) [13], a 3D anchor-based detection head, to generate predicted bounding boxes. Similarly, DASE-ProPillars [14], a single-stage LiDAR-only detector, utilizing PointPillars [12] as the baseline model with additional Attentive Hierarchical modules to improve the 3D detection performance within roadside LiDARs. The pillar feature net (PFN) takes voxelized pillars as the input, extracts pillar features, and transforms pillars back to a pseudo-image for 2D convolution operations in the middle layers. The post-training is motivated by the shape-aware data augmentation and self-assembling training framework [15], where the predictions of the pre-trained model can be used as soft supervision and ground truth as hard supervision to simplify the handling of partial occlusions, sparsity and different shapes of objects in the same class.

Arnold et al. proposed a cooperative detection system which consists of different roadside sensors, such as LiDAR and depth cameras, with positional correction for multi-view 3D object detection simultaneously [16]. Their study mainly concentrated on early, late or mixed fusion strategies at intersections. The result shows that early fusion has the highest communication transmission with good detection performance, and the hybrid fusion outperforms late fusion with a lower cost than early fusion, but performs worse than early fusion due to the loss of crucial original points. The experiment indicates that a large number of roadside sensors at the intersection is highly profitable in preventing occlusion and a limited field-of-view with spatially various and repetitive observations in complex traffic scenes. As the first deep-learning-based cooperative object detection method which integrated point cloud data from both on-board and roadside LiDARs, PillarGrid [17] proposed a cooperative-feature fusion module named Grid-wise Feature Fusion (GFF). After converting the feature data into the grid plane, each grid will include specific hidden features representing original point cloud data at the particular spatial location. Then, a CNN backbone and an anchor-based 3D detection head were applied to predict the oriented 3D bounding boxes for vehicles and pedestrians, respectively.

2.3. Transformer-Based 3D Detection

Transformer [18] was originally designed for natural language processing (NLP) and achieved excellent results in machine translation. Inspired by the huge success of attention mechanism in image classification [19,20] and 2D object detection [21], transformer-based 3D object detection has recently become a mainstream research direction. Without predicted

depth maps or any post-processing, DETR3D [22] uses backward geometric projection to connect 2D feature extraction and 3D bounding box prediction, fusing information from multiple camera views in each computation layer. M3DETR [23] simultaneously models multi-representation, multi-scale, mutual-relation features of point clouds with transformer blocks and is robust with regard to the hyper-parameter tuning of transformer architectures. To acquire global information for capturing long-range interactions, Full Self Attention (FSA) and Deformable Self-Attention (DSA) [24] was proposed in parallel to convolution networks for the augmentation of standard convolutional features. These two modules can be applied across a range of modern point-cloud based detector architectures and systematically improves their original performance, including BEV- [12], voxel- [25], point- [26] and point-voxel-based [27] 3D detectors.

Point Cloud Transformer (PCT) [28] aims to encode the original points into a new higher-dimensional feature space, by embedding the input coordinates. The design philosophy is almost the same as for the original Transformer, while discarding the positional embedding. The proposed offset-attention module serves as a self-attention block, which produces the offset between the off-attention features and the input queries by element-wise sampling. PCT has an invariant composition and is suitable for research on unordered point clouds with irregular territory, but the performance gap in larger real-world datasets still remains to be studied.

For a more effective and distinguishing feature representation, the Stacked Triple Attention (TA) module in TANet [29] has been proposed to strengthen the deficient learning of moving objects and perform better with irrelevant points, which can be applied to roadside LiDAR with both voxel- and pillar-based feature encoders [14]. The TA module extracts features in each pillar grid with a channel-wise, pillar-based and voxel-wise attention mechanism. The final output feature is integrated with the combination of all three attention scores. To further utilize the middle feature of the attention block, the input of each triple attention module is integrated or summed as the output to fuse more information, similar to the residual concatenations in ResNet [30]. The inner attention mechanism simply re-weights the features, without increasing their dimensions, as the following fully connected forward network. Taking multiple cameras as inputs, BEVFormer [31] is a spatial-temporal transformer model which utilizes both transformer (spatial) and temporal modules to generate bird's-eye-view (BEV) features. BEVFormer employs learnable BEV features as queries, along with a spatial cross-attention block and a temporal self-attention block, to search spatial features from overlapping cameras and relevant temporal features from previous proposals, respectively, and then integrate them into uniform BEV features. This dense representation is a universal 2D feature map that can be used for various autonomous driving detection heads with tiny modifications, such as 3D object detection and instance segmentation.

Attention mechanism could also be used for 3D object detection based on multi-modality sensor fusion, i.e., high-resolution LiDARs, low-resolution LiDARs, cameras and radars. FUTR3D [32], a unified sensor fusion framework with any sensor configuration for 3D detection, employs a query-based Modality-Agnostic Feature Sampler (MAFS), together with a transformer decoder with a set-to-set loss to avoid exploiting late fusion heuristics and post-processing tricks. FUTR3D also exploits the advantages of multi-sensor fusion, where lower cost sensor configurations (4-beam LiDAR and camera) could achieve performance comparable to 32-beam LiDAR.

Specifically, Transfusion [33] is a multi-modal fusion method for 3D detection with a soft attention mechanism to adaptively select what information should be taken from multiple sensors under weak conditions. The model includes standard 2D and 3D convolutional backbones to separately extract a middle feature map of image and a LiDAR BEV feature map following two transformer decoder blocks in sequence as the detection head. An image-based query initialization module is designed to dispose small objects on the LiDAR BEV feature with the guidance of images. This module includes a query and corresponding position providing the localization of the object, and a query proposal

encoding bounding box annotation, such as three-dimensional size and orientation. The first transformer decoder block predicts initialized 3D bounding boxes by a sparse set of queries, and the second intensively fuses previous object queries through initial candidates from the first stage with the image features, predicting dense texture and color information for better detection results.

To address large inter-modal discrepancies of LiDAR point-clouds and RGB images, Contrastively Augmented Transformer for multimodal 3D object Detection (CAT-Det) [34] was proposed. The whole framework consists of three main modules: (1) Two-stream Pointformer and Imageformer (TPI), (2) Cross-Modal Transformer (CMT), and (3) One-way Multi-modal Data Augmentation (OMDA). Hierarchical contrastive learning at both the point and object levels allowed improving accuracy only by augmenting point-clouds with GT-Paste [35], which pastes extra 3D objects from other LiDAR frames without spatial collision, and is thus free from a complex generation of paired samples of the two modalities.

2.4. Roadside Dataset for Object Detection

Owing to emerging needs in surrounding perception for autonomous vehicles, as Figure 2 shows, most existing datasets for object detection are collected from on-board sensors, such as KITTI [36], NuScenes [2] and Waymo [37]. However, comprehensively improving the perception ability of roadside infrastructure is the core focus and primary task for the development of VICAD. BAAI-VANJEE [38], published in 2021, is the first real-world roadside dataset, which includes 5000 frames of RGB images and 2500 frames of LiDAR data with 74 k 3D object annotations for 12 classes. In 2022, Rope3D [39] was proposed to accelerate the progress of camera-only roadside perception, and it contains 50 k images and more than 1.5 M 3D annotations of objects in various traffic conditions. Particularly, LiDAR equipped on a moving/parked vehicle was adopted to obtain matched 3D point clouds for reliable ground-truth 2D–3D joint annotation.

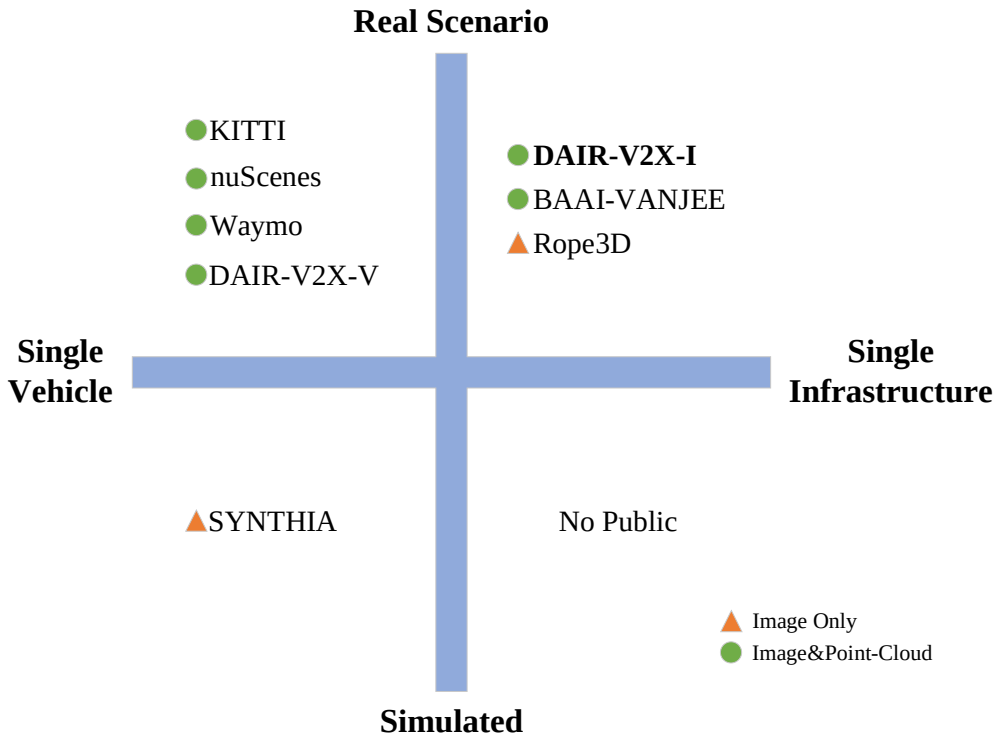

Figure 2. Datasets available for Single-View 3D (SV3D) Object Detection in autonomous driving. DAIR-V2X-I is the first real-world multi-modal dataset with 3D joint annotation of images and point clouds for infrastructure 3D detection.

Furthermore, DAIR-V2X [40] serves as the first and only large-scale, multi-modality, multi-view dataset for VICAD at this time. It consists of 71254 LiDAR frames and 71254 camera frames, which are all captured from real scenarios with 3D annotations. As a subset of DAIR-V2X, DAIR-V2X-I concentrates on roadside-centric 3D object detection, such as monocular 3D detection, LiDAR 3D detection and multi-modal 3D detection. More details on DAIR-V2X-I can be found in Section 4.

With more and more roadside datasets available, the innovation research and achievement transformation in the field of VICAD will be actively assisted and accelerated. The major problem of limited perception range in the ego-vehicle perspective will also be resolved in future.

2.5. Problems in Previous Work

In summary, LiDAR-based detectors perform better than camera-based detectors. However, most of the existing perception methods based on roadside LiDAR are traditional machine vision models. The application of deep-learning tricks for more effective object perception methods based on roadside LiDAR deserves serious research in the future. Moreover, transformer-based 3D detection models have achieved impressive results with on-board sensors, whose applications in the roadside perspective still show a lot of potential for study.

However, due to the lack of sufficient data collected from the real world, the aforementioned deep-learning-based detectors with LiDAR utilize either simulated datasets generated by the CARLA simulator [41] or customized datasets without multi-sensor calibration and ground-truth labels. This will significantly prevent the accuracy improvement of some optimization strategies for roadside detection. The previous object detection models mainly studied in simulated scenarios are difficult to apply to real roads or intersections.

3. Approach

Our research mainly focuses on how to detect more traffic participants with true positive annotations based on a limited amount of roadside LiDAR data. In this section, we present the proposed model, CetrRoad, a simple yet effective center-aware detector with deformable cross-attention for LiDAR-only 3D object detection from the roadside perspective. The overall network of CetrRoad is illustrated in Figure 1, consisting of three main blocks: (1) a voxel-based roadside LiDAR feature encoder module that voxelizes and projects the raw point cloud into BEV with dense feature representation (Section 3.1); (2) a one-stage center proposal module that initializes center candidates of objects based on the top N points in the BEV target heatmap with unnormalized 2D Gaussian (Section 3.2); (3) a detection head with multi-head self-attention and multi-scale deformable cross attention, taking outstanding center proposals as query embedding, that can predict 3D bounding boxes for different classes moving/parked at the intersection (Section 3.3). For the integrity and coherence of our presentation, a set-to-set loss function between one prediction and its corresponding ground-truth will also be introduced in Section 3.4.

3.1. Roadside LiDAR Feature Encoder

The aim of our model is to predict a set of 3D bounding boxes $B = \{b_i\}$ in the BEV as good as the ground truth, where i denotes the number of real annotations in each frame of the roadside LiDAR point cloud. Each bounding box b_i consists of a center location $\{x, y, z\}$ relative to the virtual LiDAR coordinate system, 3D size $\{w, l, h\}$, and yaw angle α. Modern LiDAR-based detectors usually utilize a 3D encoder to quantize the point-cloud into regular voxels or pillars. Then a voxel or point-based backbone, where most of the computation and quantized operation happens, will extract primary features from all points inside these bins. Common LiDAR-based 3D backbones include VoxelNet [42] and PointPillars [12]. For better performance of the final predictions, we use SECOND [25] as the 3D backbone to voxelize roadside LiDAR point clouds, following a sparse convolution encoder to pool these features into the major feature representation. The input channel of the sparse encoder

is set to 4 in our model, which depends on the storage format of raw LiDAR data. The outputs of a previous backbone with diverse channels will then be imported to Feature Pyramid Network (FPN) [43] for feature upsampling through three layers.

3.2. One-Stage Center Proposal Module

With a BEV overhead feature map, a modern detection head will produce object detections with predefined anchors. As the turning objects account for a high percentage at the intersection, 3D bounding boxes of the ground-truth usually come with various sizes and orientations. Anchor-based 3D detectors have difficulty refining an axis-aligned 2D box into a 3D object which turns right or left randomly. Actually, distances between any traffic participant are absolute in BEV, and these are distorted by perspective in the image-view. Compared with hand-designed anchors, center-based representation in BEV is more efficient and suitable for turning-objects recognition.

Since the center part of traffic participants is usually the highest at the z-axis, we could use a target heatmap to describe the center of positive locations. Following previous work [3,44], we propose a one-stage center proposal module based on the last BEV feature map Z from BEV Feature Pyramid Network (BEV FPN). Taking the projection of 3D centers of ground-truth bounding boxes as the input of an unnormalized 2D Gaussian distribution, an l channel heatmap H of object centers can be predicted during training. Each channel contains a heatmap score of one class. To reduce the penalty to negative locations and strengthen the positive supervision of the ground-truth object center, we set the smallest allowable Gaussian radius as 1.

The center candidate generation in this module is under denser supervision from nearby pixels, as Figure 3 shows. The location of the top N heatmap scores will be extracted as the center proposals. We determined N = 100 in our model empirically for a trade-off between performance and training cost.

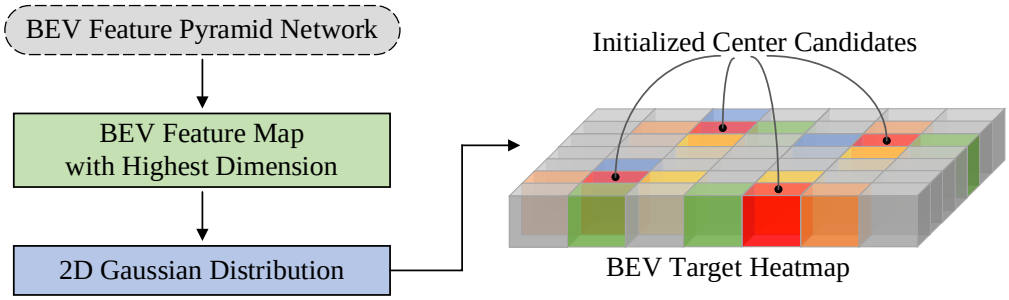

Figure 3. Simple illustration of one-stage center proposal module. Taking the projection of 3D centers of annotated bounding boxes and nearby pixels into BEV as input, a 2D unnormalized Gaussian is trained for initializing center candidates within a limited Gaussian radius. Note that red color blocks represent selected center candidates, and surrounding colors serve as referenced pixels.

3.3. Detection Head with Center Proposal and Deformable Attention

Attention mechanism has been widely used in object detection after the proposal of DETR [21]. Previous work usually takes the whole BEV feature map as the input query of the self-attention block, following the cross-attention block to refine bounding boxes and regress final predictions. However, after downsampling and upsampling, even the highest-dimensional BEV feature map is much sparser than the raw point cloud data, which means that there are much more zeros in the query vectors at the same input scale. Spatial sparsity in the transformer will lead to an unnecessary consumption of computing resources. More false positive predictions will potentially be proposed due to sparse supervision.

In our proposed detection head, taking previous initial center-aware query embedding as the input, the multi-head self-attention block firstly calculates the similarity between each query and other queries one by one. The similarity matrix, together with original input query embedding, will be normalized into a list of weight vectors by the *softmax*

function, which will then be multiplied by original center query embedding to purify initial center proposals. Close center proposals will be integrated as a whole or partially discarded. The residual connection in attention blocks guarantees that the output of the network will never be zero. In other words, the output of the self-attention block could quantify the importance of each initial center position, as a guideline for the further accurate generation of bounding boxes.

Inspired by FUTR3D [32] and Deformable DETR [45], we designed a multi-scale multi-head deformable cross-attention block within the detection head of CetrRoad to sample the outstanding center candidates automatically based on the output of the multi-head self-attention block, which serves as the input query of multi-head deformable cross-attention. As Figure 4 shows, the multi-scale feature maps from the BEV FPN module act as key and value simultaneously. Similarly to a deformable convolutional network, multi-head deformable cross-attention could learn 2D offsets Δq of purified center proposals q at each scale of BEV feature maps in all M subheads. Each subhead of the cross-attention block comparably acts as a deformable convolution layer, where K points will be randomly sampled around a reference center point. As with the self-attention block, the relevance of the updated center query q + Δq to multi-scale BEV feature maps will be calculated after bilinear interpolation and sampling. The output of the deformable cross attention module can be produced as follows:

$$DCAM(c) = \sum_{m=1}^{M} W_m [\sum_{k=1}^{K} \sigma(W_{mk} C(p)) F_B(p + \Delta p_{mk})] \quad (1)$$

where W_m represents the attention weight of each subhead in the deformable cross attention block, F_B is the multi-scale feature maps from the BEV Feature Pyramid Network, $C(p)$ is the purified center-aware feature from the self-attention block, and $\sigma(W_{mk}C(p))$ is the attention weight of the k-th sampling point in the m-th deformable attention subhead. K is the total number of sampled key points in each subhead. We used M = 8 and K = 10 in our experiments.

3.4. Label Assignment and Matching Losses

Playing the same role as the heuristic assignment strategy used to match region proposals [46] or hand-designed anchors [43] to the ground truth in typical detectors, we treat the final prediction generation as a pair-wise set matching problem without any complicated processing components, such as multiple anchors or non-maximal suppression (NMS). For a group of N predictions through the decoder layer, a set-to-set matching cost between prediction $\hat{y}_{\sigma(i)}$ with index $\sigma(i)$ and its corresponding ground-truth y_i is defined as follows,

$$C_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) = -1_{\{c_i \neq \varnothing\}} \hat{p}_{\sigma(i)}(c_i) + 1_{\{c_i = \varnothing\}} L_{box}(b_i, \hat{b}_{\sigma(i)}) \quad (2)$$

where $\hat{p}_{\sigma(i)}(c_i)$ denotes the probability of class c_i for the prediction with index $\sigma(i)$, and $b_i, \hat{b}_{\sigma(i)} \in \mathbb{R}^{N \times 8}$ is a vector with normalized coordinates that contains sine and cosine values of rotation, center coordinates (x, y, z) of the ground-truth bounding box and prediction, respectively, and its height, width and length. Since the number of predictions N is generally larger than the ground truth, we pad the set of ground truth up to N with zeros for brief computation, represented as $\varnothing$ (negative sample or background). The goal of optimization is to minimize the matching cost as follows:

$$\hat{\sigma} = \text{argmin} \sum_{i}^{N} C_{match}(y_i, \hat{y}_{\sigma(i)}) \quad (3)$$

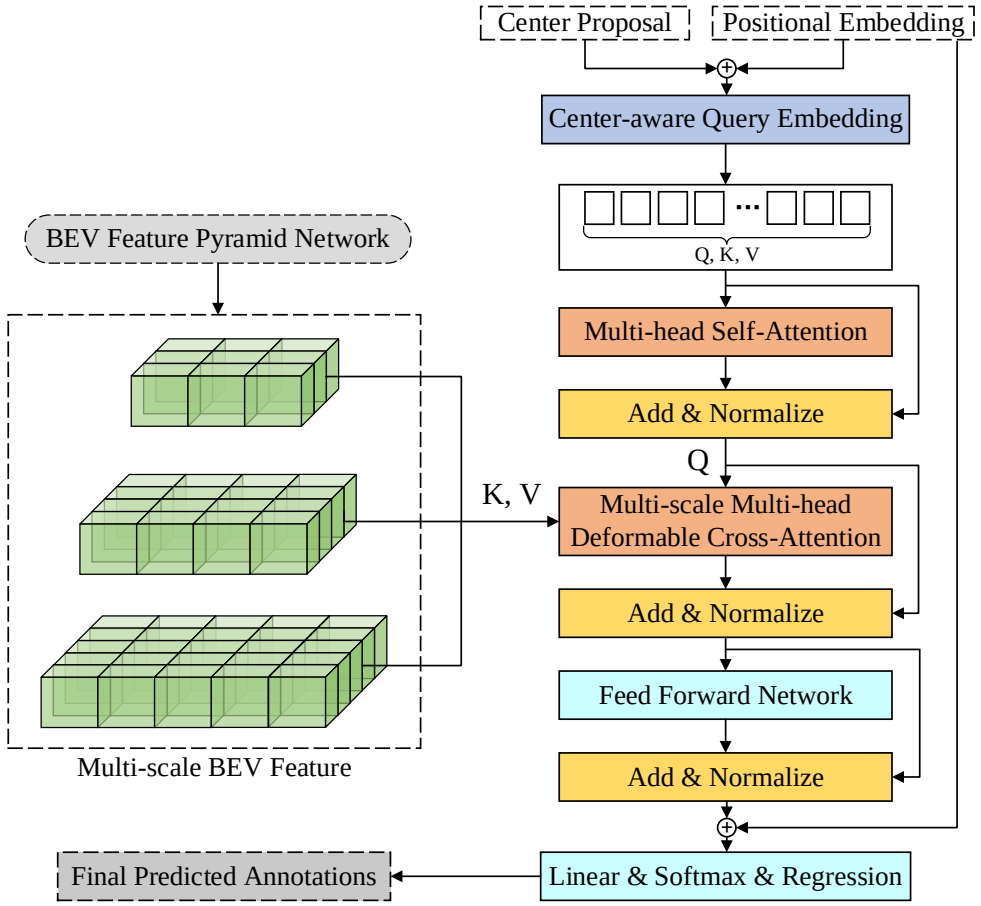

Figure 4. An overview of detection head with center proposal and deformable attention. The multi-head self-attention block could quantify the importance of each initial center position. Based on purified center queries, each subhead of the multi-head cross attention block will learn 2D offsets at each scale of BEV feature maps in all subheads, where K points will be randomly sampled around a reference center point. The output of self-attention and cross attention will be aggregated and sent to a feed-forward network, following a linear layer and softmax function to regress the final predictions. After the deformable cross attention block, $DCAM(c)$, together with the output of multi-head self-attention, will be aggregated and sent to a feed-forward network, following a linear layer and softmax function to regress the final predictions, which is a dictionary including the annotations of all predictions. The evaluation results, with more details, can be found in Section 4.3.

Motivated by [21,22,32,33,40], we use the Hungarian algorithm [47] to search for better bipartite matching results between roadside LiDAR-based predictions and the ground truth of objects. Similar to the losses of common object detectors, the Hungarian loss for 3D detection is defined as a linear combination of Focal Loss [48] for the class label predictions and a L1 loss for the bounding box parameters:

$$L_{Hungarian}(y, \hat{y}) = \sum_{1}^{N} [\lambda_1 L_{cls}(p_i, \hat{p}_{\sigma(i)}) + \lambda_2 L_{bbox}(b_i, \hat{b}_{\sigma(i)})] \quad (4)$$

where $\lambda_1 = 2.0$ and $\lambda_2 = 0.25$. For notational convenience, we define L_{cls} as follows:

$$L_{cls}(p_i, \hat{p}_{\sigma(i)}) = C_{cls}(p_i, \hat{p}_{\sigma(i)})_{c_i \neq \varnothing} - C_{cls}(p_i, \hat{p}_{\sigma(i)})_{c_i = \varnothing} \quad (5)$$

In detail, $C_{cls}(p_i, \hat{p}_{\sigma(i)})$ is represented as:

$$C_{cls}(p_i, \hat{p}_{\sigma(i)}) = \begin{cases} -\alpha(1-\hat{p}_{\sigma(i)})^{\gamma}\log(\hat{p}_{\sigma(i)}+\varepsilon)\,, c_i \neq \varnothing \\ -(1-\alpha)(\hat{p}_{\sigma(i)})^{\gamma}\log(1-\hat{p}_{\sigma(i)}+\varepsilon)\,, c_i = \varnothing \end{cases} \quad (6)$$

where $\alpha = 0.25$, $\gamma = 2$ and $\varepsilon = 1 \times 10^{-12}$ as default. The loss of predicted bounding box L_{bbox} is defined as L1 norm between b_i and $\hat{b}_{\sigma(i)}$:

$$L_{bbox}(b_i, \hat{b}_{\sigma(i)}) = \| b_i - \hat{b}_{\sigma(i)} \|_1 \quad (7)$$

4. Experiments

CetrRoad is evaluated on the challenging DAIR-V2X-I benchmark. We first introduce the main characteristics of the DAIR-V2X-I dataset and a quantitative evaluation metric for the comparison of detection performance in Section 4.1. Then, we present critical information about model training and evaluation in Section 4.2 from beginning to end. We compare CetrRoad with three official DAIR-V2X-I baselines and two other representative LiDAR-only models on the DAIR-V2X-I validation set. Furthermore, the quantitative and qualitative analysis are comprehensively presented in Section 4.3, and visualization of some example predicted results are shown in Figure 5.

4.1. Experimental Setup

Dataset. We evaluated our method on the DAIR-V2X [40] roadside 3D detection dataset (**DAIR-V2X-I**). The DAIR-V2X-I dataset is the first large-scale roadside multimodal dataset with 3D joint annotation of images and point clouds, including 10,084 frames of roadside image data and 10,084 frames of point cloud data. All the data were captured by a pair of RGB cameras and LiDAR, which are installed in the same azimuth and calibrated at the same time, and the images are undistorted. Since there is a pitch angle between the roadside LiDAR and the ground, for the convenience of research, the roadside LiDAR coordinate system and corresponding point clouds are uniformly transferred to the virtual LiDAR coordinate system through the roadside LiDAR external parameter matrix. Specifically, all images and point cloud frames include exhaustive annotations of the 10 object classes with their category attributes, occlusion states, truncated states, and precise seven-dimensional cuboids modeled as x, y, z, *width*, *length*, *height*, and *yaw angle*.

Evaluation metrics. Following the PASCAL VOC criteria [49], we evaluated the 3D object detection performance on the DAIR-V2X-I dataset by AP (Average Precision), which is defined as the area under the Precision-Recall curve. Specially, Precision (also called positive predictive value) is the proportion of ground truth among all predicted annotations, while Recall is defined as the number of ground truth predicted successfully divided by the total number of ground truth. According to object size, occlusion and truncation levels, the ground-truth labels are categorized into Easy, Moderate and Hard for evaluation. For impartial comparison, we set [0.7, 0.5, 0.5] as the uniform IoU (Intersection over Union) threshold for Car and [0.5, 0.25, 0.25] for Pedestrian and Cyclist, respectively.

4.2. Implementation Details

Preprocessing. First of all, we preprocessed the raw DAIR-V2X-I dataset under the open-source OpenDAIRV2X, which supports the Vehicle-Infrastructure Cooperative 3D Object Detection (VIC3D) task and two Single-View (SV) 3D detection benchmarks. To facilitate the effective training and compare the performance with the official benchmark, we converted the original data into KITTI [36] format, as Figure 6 shows. We divided DAIR-V2X-I into a training set, validation set and testing set according to 5:2:3. Note that the testing folder is empty because the *testing* set of the DAIR-V2X-I dataset has not been released at this point. The DAIR-V2X-I available dataset contains 7058 frames of roadside image data and 7058 frames of point cloud data, which are completely utilized by the methods presented in this paper.

Figure 5. Visualization of some example qualitative results on the DAIR-V2X-I dataset. The ground-truth bounding boxes of input images and LiDAR point cloud are shown in orange and blue, respectively. The predicted results on the raw point cloud are showed in green. It can be seen that CetrRoad predicts results much better, with fewer false positives than other methods, especially for pedestrians and cyclists across the zebra crossing.

```
└── single-infrastructure-side
    ├── calib
    ├── data_info.json
    ├── image
    ├── ImageSets
    ├── kitti_dbinfos_train.pkl
    ├── kitti_gt_database
    ├── kitti_infos_test_mono3d.coco.json
    ├── kitti_infos_test.pkl
    ├── kitti_infos_train_mono3d.coco.json
    ├── kitti_infos_train.pkl
    ├── kitti_infos_trainval_mono3d.coco.json
    ├── kitti_infos_trainval.pkl
    ├── kitti_infos_val_mono3d.coco.json
    ├── kitti_infos_val.pkl
    ├── label
    ├── testing
    ├── training
    └── velodyne
```

Figure 6. The file structure of DAIR-V2X-I after data preprocessing. OpenDAIRV2X preprocessed the raw data of DAIR-V2X-I for training and evaluation. More details can be found at https://github.com/AIR-THU/DAIR-V2X (accessed on 12 July 2022).

Configuration. Our implementation is based on the public codebase MMDetection3D (v0.17.1) [50]. For fair comparison of all methods presented in Section 4, we uniformly constrained the detection range in every point cloud frame to [0, 70.4 m] for the x-axis, [−40 m, 40 m] for the y-axis, and [−3 m, 1 m] for the z-axis, respectively. Our model consists of three parts: a voxel-based roadside LiDAR feature encoder that projects the raw point cloud into BEV features with dense representation; the size of each voxel is set to [0.05 m, 0.05 m, 0.1 m] and the total number of voxels in each single frame is 1600×1408. CetrRoad adopts SECOND [25] as a backbone and SparseConv [51] with sparse shape [41, 1600, 1408] as the middle feature extractor. SparseConv can flatten the 3D feature tensors into the BEV plane by simply collapsing the z-axis, which consists of four blocks of [2, 3, 3, 3] 3D sparse convolutional layers with the dimensions [16, 32, 64, 128]. The input features of the LiDAR backbone will be downsampled to $1/2$ and $1/4$ of the original feature map respectively, which is the opposite of the BEV FPN layers.

3D Data Augmentation. Inspired by previous work [25], several new objects from ground-truth and corresponding points in other frames were partially pasted into the current training LiDAR frame except specific bounding boxes overlapping with some boxes in the present frame. The number of sampled points is flexible; the setting is 10 for Cyclist, 10 for Pedestrian and 12 for Car in our experiment. Then, global rotation, translation and scaling were applied to the whole point cloud in three-dimensional space, where the probability of flipping in each frame is set to 0.5. The scale of rotation angle is random from $[-\pi/4, \pi/4]$, with the scaling ratio between 0.95 and 1.05. Random flip of bounding boxes was also applicable in the horizontal direction of BEV, where the probability of rotation is 0.5.

Training & Evaluation. AdamW [52] was employed to train our model without any pretrained assistance. The weight decay for AdamW is 10^{-2}. The learning rate was 10^{-3} initially and finally decreased to 10^{-7} following a cyclic schedule. We did not use any post-processing such as NMS. All experimental results presented in this paper were produced based on the public DAIR-V2X-I dataset, whose training and evaluation were both implemented on our own devices. In consideration of cost and the lifespan of devices, multiple computing units in the single Road Side Unit (RSU) are less likely to be deployed. Therefore, we assumed that there is only one GPU (Graphics Processing Unit) available for training and evaluation on the DAIR-V2X-I dataset. A total of 24 GB memory of

the NVIDIA RTX3090 GPU was applicable for reimplementing DAIR-V2X-I detection baselines, including PointPillars [12], SECOND [25] and MVX-NET [53]. However, Object DGCNN [54] and FSD [55] need more memory for smooth training. We adopted a single NVIDIA A100 GPU with 40 GB memory instead to produce optimal detection results for Object DGCNN, FSD and CetrRoad. To achieve the best performance with limited resources in the shortest time, PointPillars [12] was trained for 160 epochs and the others for 80 epochs with maximized batch size. More training details and settings are reported in Table 1. 160 training epochs are sufficient for PointPillars, which is the official setting in MMDetection 3D and OpenDAIRV2X. For a fair comparison, we also trained CetrRoad with the same 80 epochs as other models. The evaluation pipeline is consistent with training, where we used the toolkit provided by MMDetection3D to show the quantitative and qualitative results clearly.

Table 1. Some detailed training settings for all the models mentioned in our experiment.

Method	Environment	Memory Cost	Training Time	Average Iteration Time	Training Epochs
PointPillars	Single NVIDIA RTX3090 GPU	6.7 GB (batch size = 6)	24 h	0.2782 s/iter	160
SECOND		6.0 GB (batch size = 6)	8 h	0.3616 s/iter	80
MVX-NET		3.9 GB (batch size = 1)	30 h	0.2502 s/iter	80
Object DGCNN(voxel)	Single NVIDIA A100 GPU 40G	13.7 GB (batch size = 4)	17 h	0.5813 s/iter	80
FSD		7.4 GB (batch size = 2)	26 h	0.4851 s/iter	80
Ours		5.3 GB (batch size = 4)	16.5 h	0.5917 s/iter	80

4.3. Performance Comparison on DAIR-V2X-I

Quantitative Analysis. We compared the performance of CetrRoad with the following categories of methods on the DAIR-V2X-I *validation* set, including (a) PointPillars [12] and SECOND [25], the anchor-based LiDAR-only detectors without attention mechanism; (b) Object DGCNN, the representative LiDAR-only detector with attention-based detection head; (c) FSD [55], the state-of-the-art approach with a fully sparse detector on the Waymo dataset [37], which utilized five LiDAR sensors simultaneously for 3D object detection; and (d) MVX-NET, the modern multi-modal (LiDAR + image) fusion method [53]. It is worth noting that (a) and (d) are three DAIR-V2X-I detection baselines. Table 2 summarizes the detailed results of BEV Average Precision on the DAIR-V2X-I validation set over CetrRoad and all aforementioned methods, whose perspective of observation is bird's-eye view. Similarly, Table 3 reports more results of 3D Average Precision with the same evaluation pipeline as Table 2. The observation perspective in Table 3 is a general three-dimensional view. AP (Average Precision of each class) and mAP (mean Average Precision of all classes) of each method are both reported for complete evaluation. Tables 2 and 3 present AP of three classes, showing the gap in quantitative results between different methods more concisely. By virtue of the specially designed center proposal structure with deformable attention mechanism and effective transformer-based detector, as Tables 2 and 3 show, our method outperforms all other approaches for *Car* and *Cyclists* on DAIR-V2X-I, becoming the new state of the art. For the challenging class of *Pedestrian*, CetrRoad also reaches comparable scores to MVX-NET [53], which is a single stage detector for combining images and point cloud frames with more training data. Object DGCNN [54] with sparse convolution backbone adopts the multi-scale deformable attention structure as dense head. However, its performance on DAIR-V2X-I is not as good as that of non-transformer counterparts. FSD [55] is a fully sparse 3D object detector for enabling efficient long-range

LiDAR-based object detection, which predicts comparable quantitative results as CetrRoad in all three classes. It is difficult to compare the computational and spatial cost of FSD and our model over various batch sizes. With the same training epochs and a single GPU fully operational, as Table 1 shows, the training time of FSD is much greater than for CetrRoad, which is the only thing we could confirm.

Table 2. Performance comparison of BEV Average Precision (BEV AP) on the DAIR-V2X-I *validation* set. Here 'L' denotes LiDAR input and 'I' denotes RGB image. We set [0.7, 0.5, 0.5] as the IoU threshold for Car and [0.5, 0.25, 0.25] for Pedestrian and Cyclist, respectively. The bold results denote the best of all methods based on the DAIR-V2X-I official leaderboard.

Method	Modality		Car BEV AP (%)			Pedestrian BEV AP (%)			Cyclist BEV AP (%)		
	L	I	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
PointPillars [12]	√		63.58	54.49	54.50	50.36	44.58	44.56	47.10	24.86	27.08
SECOND [25]	√		63.60	54.51	54.51	70.23	67.18	67.25	60.07	33.29	33.41
Object DGCNN(voxel) [54]	√		61.81	52.58	52.79	64.91	62.02	62.53	58.46	32.17	32.53
FSD [55]	√		69.63	**54.51**	60.61	70.68	69.59	69.77	67.37	35.65	36.70
MVX-NET [53]	√	√	63.54	54.45	54.46	71.59	**71.17**	**71.21**	63.42	34.27	34.43
Ours	√		**70.97**	54.23	**61.96**	**74.43**	70.79	70.86	**67.85**	**35.94**	**38.32**

Table 3. Performance comparison of 3D Average Precision (3D AP) over CetrRoad and all aforementioned methods on DAIR-V2X-I *validation* split. Best in bold.

Method	Modality		Car 3D AP (%)			Pedestrian 3D AP (%)			Cyclist 3D AP (%)		
	L	I	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
PointPillars [12]	√		63.57	54.49	54.49	50.23	44.52	44.51	47.08	24.85	27.07
SECOND [25]	√		63.59	54.51	54.51	70.06	67.05	67.12	60.05	33.29	33.41
Object DGCNN (voxel) [54]	√		61.53	52.34	52.56	64.15	61.01	61.52	58.33	32.14	32.49
FSD [55]	√		69.60	**54.51**	54.51	70.54	69.38	69.61	67.29	35.64	36.68
MVX-NET [53]	√	√	63.54	54.46	54.46	71.39	**70.89**	**70.97**	63.34	34.25	34.41
Ours	√		**70.82**	54.19	**61.83**	**74.11**	70.45	70.50	**67.73**	**35.91**	**37.28**

Qualitative Analysis. Some example qualitative results of CetrRoad and compared methods on the *validation* set of the DAIR-V2X-I dataset are visualized in Figure 5 for unambiguous comparison. To verify the robustness of our proposed model, we selected four representative scenes for infrastructure 3D detection from day to night, where CetrRoad performed better than others. As Scenario 1 shows, CetrRoad can produce more accurate 3D detection results of *Car* at the intersection, especially for the occluded car under roadside LiDAR. Moreover, our model almost detected all pedestrians and cyclists across the zebra crossing, with fewer false positives in Scenario 2. Finally, we can intuitively observe that the predictions of FSD and Object DGCNN are both worse than CetrRoad because they generate many unnecessary bounding boxes at non-target locations, which may causes more trouble for other downstream VICAD tasks.

5. Discussion

Both anchor-based and transformer-based detectors have been widely used in 3D object detection for autonomous driving. But their performance gap based on roadside sensors remains to be studied. As the ablation study shows in Table 4, we find that the input query type of transformer-based detection head is a major factor that greatly influences the accuracy of predictions. Without dense representation of query, an anchor-based detection head is likely to be better than an attention-based one. Object DGCNN [54]

is a popular LiDAR-only detector with transformer-based detection head, whose input query is sparse BEV feature maps. The bounding box predictions and visualized results of Object DGCNN both demonstrate that a BEV feature map of LiDAR is not a good type of input query, leading to many false positive predictions with approximately the same time cost as for our model. It is worth noting that center-aware proposals could provide dense query embedding for multi-head self-attention. The output of the previous self-attention block, together with purified multi-scale BEV feature map, also serves as a strong supervision to the multi-head deformable cross-attention block for bounding box generation and refinement.

Table 4. Ablation study to evaluate the sub-modules of CetrRoad. Attention-based detection head is worse than anchor-based if there is no dense query as input.

Method	Input Query Type		Detection Head		Cyclist 3D AP (%)		
	BEV Feature Map	Center-Aware Proposal	Anchor-Based	Transformer-Based	Easy	Moderate	Hard
Baseline	√		√		60.05	33.29	33.41
	√			√	58.33	32.14	32.49
CetrRoad		√		√	67.73	35.91	37.28

However, similarly to other compared methods, CetrRoad still cannot effectively detect long-distance objects at intersections. The main reason is that fewer reference pixels can be used to supervise the center point generation due to the sparser point cloud in the far distance. There were widely distributed cameras at the intersection. Supplementing additional information from cameras and exploring more efficient multi-modal data fusion algorithms based on roadside sensors may be a solution for this issue. Finally, the resolution of images is usually much denser than LiDAR point, which could easily generate dense middle features of images accessible for roadside multi-sensor fusion.

6. Conclusions

In this work, we propose CetrRoad, a simple yet effective center-aware detector with deformable cross-attention for LiDAR-only 3D object detection from the roadside perspective. Our model provides a solution for the dilemma of detecting turning and occluded objects at an intersection. CetrRoad utilizes a voxel-based roadside LiDAR feature encoder module that voxelizes and projects the raw point cloud into BEV with dense feature representation, following a one-stage center proposal module that initializes center candidates of objects based on the top N points in the BEV target heatmap with unnormalized 2D Gaussian. Then, taking previous center proposals as query embedding, a detection head with multi-head self-attention and multi-scale multi-head deformable cross attention can predict 3D bounding boxes for different classes moving/parked at the intersection.

Quantitative studies show that our method outperforms various strong baselines and achieves state-of-the-art performance on the DAIR-V2X-I benchmark with an acceptable training time cost, especially in *Car* and *Cyclist*. CetrRoad also reaches comparable detection results with the multi-modal fusion method in *Pedestrian*. An ablation study demonstrates that the attention-based detection head with center-aware proposals could predict more accurate results than an anchor-based head, as well as a transformer-based head with single BEV feature map as input query. Moreover, we could intuitively observe that in a complex traffic environment, our proposed model was able to produce more accurate 3D detection results than other methods compared, with fewer false positives, which is helpful for other downstream tasks of Vehicle-Infrastructure Cooperated Autonomous Driving.

Author Contributions: Conceptualization, H.S. and D.H.; methodology, H.S.; software, H.S.; validation, H.S.; formal analysis, H.S.; investigation, H.S. and X.L.; data curation, H.S. and X.L.; writing—original draft preparation, H.S.; writing—review and editing, H.S. and D.H.; visualization, H.S.; supervision, D.H.; project administration, D.H.; funding acquisition, D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Joint Funds of the National Natural Science Foundation of China, grant number U21B2089.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and pre-trained model for evaluation presented in this study are available on request from the corresponding author.

Acknowledgments: We sincerely acknowledge the Beijing Super Cloud Computing Center (BSCC) for providing HPC resources that have contributed to the research results reported within this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Creß, C.; Knoll, A.C. Intelligent Transportation Systems With The Use of External Infrastructure: A Literature Survey. *arXiv* **2021**, arXiv:2112.05615.
2. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11621–11631.
3. Yin, T.; Zhou, X.; Krahenbuhl, P. Center-based 3d object detection and tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11784–11793.
4. Guo, E.; Chen, Z.; Rahardja, S.; Yang, J. 3D Detection and Pose Estimation of Vehicle in Cooperative Vehicle Infrastructure System. *IEEE Sens. J.* **2021**, *21*, 21759–21771. [CrossRef]
5. Zou, Z.; Zhang, R.; Shen, S.; Pandey, G.; Chakravarty, P.; Parchami, A.; Liu, H.X. Real-time full-stack traffic scene perception for autonomous driving with roadside cameras. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; pp. 890–896.
6. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
7. Wu, J.; Xu, H.; Zheng, J. Automatic background filtering and lane identification with roadside LiDAR data. In Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, 16–19 October 2017; pp. 1–6.
8. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd* **1996**, *96*, 226–231.
9. Li, J.; Cheng, J.-h.; Shi, J.-y.; Huang, F. Brief introduction of back propagation (BP) neural network algorithm and its improvement. In *Advances in Computer Science and Information Engineering*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 553–558.
10. Gong, Z.; Wang, Z.; Zhou, B.; Liu, W.; Liu, P. Pedestrian Detection Method Based on Roadside Light Detection and Ranging. *SAE Int. J. Connect. Autom. Veh.* **2021**, *4*, 413–422. [CrossRef]
11. Bai, Z.; Nayak, S.P.; Zhao, X.; Wu, G.; Barth, M.J.; Qi, X.; Liu, Y.; Oguchi, K. Cyber Mobility Mirror: Deep Learning-based Real-time 3D Object Perception and Reconstruction Using Roadside LiDAR. *arXiv* **2022**, arXiv:2202.13505. [CrossRef]
12. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. Pointpillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12697–12705.
13. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
14. Zimmer, W.; Grabler, M.; Knoll, A. Real-Time and Robust 3D Object Detection Within Road-Side LiDARs Using Domain Adaptation. *arXiv* **2022**, arXiv:2204.00132.
15. Zheng, W.; Tang, W.; Jiang, L.; Fu, C.-W. SE-SSD: Self-ensembling single-stage object detector from point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14494–14503.
16. Arnold, E.; Dianati, M.; de Temple, R.; Fallah, S. Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 1852–1864. [CrossRef]
17. Bai, Z.; Wu, G.; Barth, M.J.; Liu, Y.; Sisbot, E.A.; Oguchi, K. Pillargrid: Deep learning-based cooperative perception for 3d object detection from onboard-roadside lidar. In Proceedings of the 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), Macau, China, 8–12 October 2022; pp. 1743–1749.

18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
19. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
20. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
21. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
22. Wang, Y.; Guizilini, V.C.; Zhang, T.; Wang, Y.; Zhao, H.; Solomon, J. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In Proceedings of the Conference on Robot Learning, Auckland, New Zealand, 14–18 December 2022; pp. 180–191.
23. Guan, T.; Wang, J.; Lan, S.; Chandra, R.; Wu, Z.; Davis, L.; Manocha, D. M3detr: Multi-representation, multi-scale, mutual-relation 3d object detection with transformers. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 772–782.
24. Bhattacharyya, P.; Huang, C.; Czarnecki, K. Sa-det3d: Self-attention based context-aware 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3022–3031.
25. Yan, Y.; Mao, Y.; Li, B. Second: Sparsely embedded convolutional detection. *Sensors* **2018**, *18*, 3337. [CrossRef]
26. Shi, S.; Wang, X.; Li, H. Pointrcnn: 3d object proposal generation and detection from point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 770–779.
27. Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; Li, H. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10529–10538.
28. Guo, M.-H.; Cai, J.-X.; Liu, Z.-N.; Mu, T.-J.; Martin, R.R.; Hu, S.-M. Pct: Point cloud transformer. *Comput. Vis. Media* **2021**, *7*, 187–199. [CrossRef]
29. Liu, Z.; Zhao, X.; Huang, T.; Hu, R.; Zhou, Y.; Bai, X. Tanet: Robust 3d object detection from point clouds with triple attention. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11677–11684.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
31. Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Yu, Q.; Dai, J. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. *arXiv* **2022**, arXiv:2203.17270.
32. Chen, X.; Zhang, T.; Wang, Y.; Wang, Y.; Zhao, H. Futr3d: A unified sensor fusion framework for 3d detection. *arXiv* **2022**, arXiv:2203.10642.
33. Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; Tai, C.-L. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1090–1099.
34. Zhang, Y.; Chen, J.; Huang, D. CAT-Det: Contrastively Augmented Transformer for Multi-modal 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 908–917.
35. Xu, D.; Anguelov, D.; Jain, A. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 244–253.
36. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
37. Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B. Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2446–2454.
38. Yongqiang, D.; Dengjiang, W.; Gang, C.; Bing, M.; Xijia, G.; Yajun, W.; Jianchao, L.; Yanming, F.; Juanjuan, L. BAAI-VANJEE Roadside Dataset: Towards the Connected Automated Vehicle Highway technologies in Challenging Environments of China. *arXiv* **2021**, arXiv:2105.14370.
39. Ye, X.; Shu, M.; Li, H.; Shi, Y.; Li, Y.; Wang, G.; Tan, X.; Ding, E. Rope3D: The Roadside Perception Dataset for Autonomous Driving and Monocular 3D Object Detection Task. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 21341–21350.
40. Yu, H.; Luo, Y.; Shu, M.; Huo, Y.; Yang, Z.; Shi, Y.; Guo, Z.; Li, H.; Hu, X.; Yuan, J. DAIR-V2X: A Large-Scale Dataset for Vehicle-Infrastructure Cooperative 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 21361–21370.
41. Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; Koltun, V. CARLA: An open urban driving simulator. In Proceedings of the Conference on robot learning, Mountain View, CA, USA, 13–15 November 2017; pp. 1–16.

42. Zhou, Y.; Tuzel, O. Voxelnet: End-to-end learning for point cloud based 3d object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4490–4499.
43. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
44. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
45. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J.F. DD Deformable transformers for end-to-end object detection. In Proceedings of the 9th International Conference on Learning Representations, Virtual Event, Austria, 3–7 May 2021.
46. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards Real-Time Object Detection with Region Proposal Networks. Part of Advances in Neural Information Processing Systems 28 (NIPS 2015). 2015, Volume 28. Available online: https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf (accessed on 14 July 2021).
47. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [CrossRef]
48. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 2980–2988.
49. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
50. Contributors, M.D. MMDetection3D: OpenMMLab Next-Generation Platform for General 3D Object Detection. 2020. Available online: https://github.com/open-mmlab/mmdetection3d (accessed on 10 December 2021).
51. Graham, B.; Engelcke, M.; Van Der Maaten, L. 3d semantic segmentation with submanifold sparse convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9224–9232.
52. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
53. Sindagi, V.A.; Zhou, Y.; Tuzel, O. Mvx-net: Multimodal voxelnet for 3d object detection. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 7276–7282.
54. Wang, Y.; Solomon, J.M. Object dgcnn: 3d object detection using dynamic graphs. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 20745–20758.
55. Fan, L.; Wang, F.; Wang, N.; Zhang, Z. Fully Sparse 3D Object Detection. *arXiv* **2022**, arXiv:2207.10035.