sensors

MDPI

Article

LiDAR-Based 3D Temporal Object Detection via Motion-Aware LiDAR Feature Fusion

Gyuhee Park [1], Junho Koh [1], Jisong Kim [1], Jun Moon [1] and Jun Won Choi [2,*]

[1] Department of Electrical Engineering, Hanyang University, Seoul 04763, Republic of Korea; ghpark@spa.hanyang.ac.kr (G.P.); jhkoh@spa.hanyang.ac.kr (J.K.); jskim@spa.hanyang.ac.kr (J.K.); junmoon@hanyang.ac.kr (J.M.)

[2] Department of Electrical and Computer Engineering, College of Liberal Studies, Seoul National University, Seoul 08826, Republic of Korea

* Correspondence: junwchoi@snu.ac.kr; Tel.: +82-02-880-9527

Abstract: Recently, the growing demand for autonomous driving in the industry has led to a lot of interest in 3D object detection, resulting in many excellent 3D object detection algorithms. However, most 3D object detectors focus only on a single set of LiDAR points, ignoring their potential ability to improve performance by leveraging the information provided by the consecutive set of LIDAR points. In this paper, we propose a novel 3D object detection method called temporal motion-aware 3D object detection (TM3DOD), which utilizes temporal LiDAR data. In the proposed TM3DOD method, we aggregate LiDAR voxels over time and the current BEV features by generating motion features using consecutive BEV feature maps. First, we present the temporal voxel encoder (TVE), which generates voxel representations by capturing the temporal relationships among the point sets within a voxel. Next, we design a motion-aware feature aggregation network (MFANet), which aims to enhance the current BEV feature representation by quantifying the temporal variation between two consecutive BEV feature maps. By analyzing the differences and changes in the BEV feature maps over time, MFANet captures motion information and integrates it into the current feature representation, enabling more robust and accurate detection of 3D objects. Experimental evaluations on the nuScenes benchmark dataset demonstrate that the proposed TM3DOD method achieved significant improvements in 3D detection performance compared with the baseline methods. Additionally, our method achieved comparable performance to state-of-the-art approaches.

Keywords: 3D object detection; LiDAR; temporal; motion-aware aggregation; autonomous driving

Citation: Park, G.; Koh, J.; Kim, J.; Moon, J.; Choi, J.W. LiDAR-Based 3D Temporal Object Detection via Motion-Aware LiDAR Feature Fusion. *Sensors* **2024**, *24*, 4667. https://doi.org/10.3390/s24144667

Academic Editor: Ram M. Narayanan

Received: 20 May 2024
Revised: 25 June 2024
Accepted: 9 July 2024
Published: 18 July 2024

1. Introduction

The demand for accurate 3D object detection in robotics and autonomous driving has spurred the development of various LiDAR-based methods [1–11]. However, most existing approaches focus on processing individual LiDAR point sets without considering the temporal data. In real-world scenarios, LiDAR sensors generate a continuous stream of point cloud data. This temporal sequence provides valuable spatiotemporal information which can improve detection performance. Neglecting this temporal aspect limits the ability to leverage sequential data and handle challenges such as occlusion, missing points, and irregular sampling. In order to address these limitations, it is essential to develop methods that effectively utilize the temporal structure of LiDAR data. By considering the temporal order and incorporating motion-aware techniques, we can enhance the accuracy and robustness of 3D object detection, advancing the capabilities of robotics and autonomous driving systems.

In order to address these limitations, recent advancements have extended video object detection techniques to the domain of 3D object detection, and the various techniques are shown in [12–17]. The authors of [12,14,16,17] proposed harnessing the temporal aspect of LiDAR data. These include recurrent neural networks (RNNs), such as long short-term memory

(LSTM), which capture temporal dependencies and learn the evolution of object features over time. Transformer models, originally developed for natural language processing, have also been adapted to model the temporal relationships in LiDAR data effectively. Furthermore, graph neural networks (GNNs) have shown promise in capturing the spatial and temporal dependencies among LiDAR points, enabling accurate detection and object tracking. These methods utilize graph representations to model point cloud data and exploit the interrelations between points within and across frames.

In this paper, we present a new 3D object detection architecture, referred to as TM3DOD. The key focus of TM3DOD is to leverage the temporal aspects of data to enhance the voxel encoder and aggregate the current BEV features by generating motion features. The proposed TM3DOD method consists of two main components, which are detailed below.

First, TM3DOD takes into account the temporal relationships between LiDAR scans within a set of LiDAR points. This set consists of subsets of LiDAR points collected from temporal laser sweeps. TM3DOD recognizes the distinct subsets when encoding the points within each 3D voxel. The underlying concept is that points obtained from temporal laser sweeps may influence voxel encoding in different ways. In order to address this, we introduce a temporal voxel encoder (TVE), which applies varying attention weights to the LiDAR points in each sweep.

Second, we aim to enhance the current BEV feature map by leveraging sequential LiDAR data. We extract individual BEV feature maps from the consecutive LiDAR voxel and calculate the motion feature by measuring the temporal variation between two adjacent BEV feature maps. Subsequently, we add each motion feature with the corresponding BEV feature and concatenate the aggregated features together. This process ensures that the temporal information is effectively incorporated into the BEV feature map, enabling improved 3D object detection performance.

In order to verify the effectiveness of TM3DOD, we conducted experiments on the widely used nuScenes dataset [18]. We applied TM3DOD to two baseline methods: PointPillar [2] and CenterPoint [19]. The results of our experiments demonstrate significant performance improvements. Specifically, TM3DOD achieved a 4.18% increase in mean average precision (mAP) compared with PointPillar in an ablation study and a 3.61% increase compared with CenterPoint in the nuScenes test benchmark. These performance gains highlight the effectiveness of TM3DOD in enhancing 3D object detection accuracy. Additionally, our proposed TM3DOD method achieved comparable performance to state-of-the-art methods.

The key contributions of this paper can be summarized as follows:

- In this study, we propose a novel 3D object detection architecture called TM3DOD which leverages the temporal aspects of LiDAR data to improve detection performance. TM3DOD focuses on enhancing the voxel encoder and aggregating the current BEV features through the generation of motion features.
- TM3DOD improves the voxel encoder by considering the temporal relationships between LiDAR scans. It also enhances the BEV feature maps by incorporating sequential LiDAR data and calculating motion features. These enhancements enable TM3DOD to better utilize the temporal information for more accurate 3D object detection.
- Our experiments on the nuScenes benchmark dataset confirm the effectiveness of TM3DOD, as it achieved substantial enhancements in 3D detection performance compared with the baseline methods. Moreover, our method achieved comparable performance to state-of-the-art approaches.

2. Related Works

2.1. 3D LiDAR Object Detection Based on a Single Frame

Overall, 3D object detection techniques using a single snapshot of sensor measurements have undergone rapid advancement and have played a crucial role in autonomous driving. The existing 3D object detectors can be grouped into three categories according

to the data used, which are camera-based [20–25], LiDAR-based [1–11,26,27], and fusion-based methods [28–33]. In this work, we focus on LiDAR-based methods since they are suitable for obtaining accurate depth information and are less sensitive to illumination and weather conditions.

Existing works on single-frame LiDAR-based 3D object detection can be roughly categorized into three approaches: voxel-based methods [1–4,26], point-based methods [5–7], and hybrid methods which combine both approaches [8–11]. Grid-based methods project the irregular point clouds onto regular grid representations, such as voxels [3] or pillars [2], to extract point features using 2D or 3D convolutional neural networks (CNNs). These methods are more computationally efficient compared with point-based approaches. However, they suffer from quantization loss due to the limited resolution of the grid representation. Examples of grid-based methods include SECOND [1], VoxelNet [3], and PartA2 [26]. Grid-based methods are widely used in autonomous driving applications due to their efficiency. Point-based methods, on the other hand, retain the geometric information of the point cloud by directly processing the unordered points using PointNet [34] or PointNet++ [35]. These methods have the potential to achieve better performance than grid-based methods since they do not suffer from the quantization loss caused by grid representations. However, point-based methods require additional computation costs for sampling and grouping points, which can be computationally expensive. PointRCNN [5], 3DSSD [6], and Point-GNN [7] are examples of point-based methods. Hybrid methods aim to combine the advantages of both grid-based and point-based approaches. They attempt to fuse the merits of grid-based feature extraction and point-based geometric information. Although hybrid methods often achieve better performance than individual approaches, they tend to be more time-consuming compared with grid-based methods. Some examples of hybrid methods include FastPointR-CNN [8], STD [9], PV-RCNN [10], and SA-SSD [11].

2.2. 3D LiDAR Object Detection Based on Multiple Frames

Despite the rapid advancements in 3D object detection using single LiDAR point sets, these methods have limitations due to the lack of utilization of temporal information in sequence data. In order to overcome the limitations of single point-based 3D object detection, recent methods [12–17,27,36] have focused on utilizing sequence data to exploit temporal information. To capture a spatiotemporal representation of point clouds, some approaches [12,13] employ recurrent neural networks (RNNs) such as LSTM [37] and GRU [38]. VelocityNet [15] employs deformable convolution to align temporal features. Other methods [16,17] utilize transformer models known for their superior performance in sequential tasks such as natural language processing. Transformers excel at capturing long-range dependencies and complex temporal relationships. As LiDAR point clouds are affected by egomotion, which can degrade 3D object detection performance, most approaches mitigate this influence by incorporating egomotion transformation using GPS sensors. However, accurately detecting dynamic objects which undergo siginificant amounts of motion remains a challenge despite aligning static objects across frames. In order to address these challenges, 3DVID [13] introduces a temporal transformer attention scheme to align moving objects. TCTR [14] also has a similar architecture, using temporal information based on a transformer.

In our proposed method, TM3DOD, we further enhance the temporal aspect by utilizing temporal motion information in both the voxel and BEV domains. Previous research has suggested feature aggregation in the BEV domain. However, this has a limitation; the BEV feature does not contain all of the data information of LiDAR due to encoding. In order to overcome this, we aim to improve feature representation over time by aggregating LiDAR voxels and current BEV features through motion feature generation. By incorporating these enhancements, TM3DOD aims to enhance the accuracy and robustness of 3D object detection. It effectively leverages temporal information and utilizes motion-aware feature aggregation to improve object detection across frames.

3. Proposed Method

In this section, we present details of the structure of the proposed method. We first provide a comprehensive overview of the proposed architecture for 3D object detection, referred to as TM3DOD. Afterward, a temporal voxel encoder (TVE) is presented which considers the temporal correlations between LiDAR data in a voxel domain. Finally, we explain the motion-aware feature aggregation network (MFANet), which combines BEV features by using an attention mechanism.

3.1. Overall Architecture

The overall architecture of our proposed TM3DOD method is illustrated in Figure 1. TM3DOD comprises two main modules: a temporal voxel encoder (TVE) and motion-aware feature aggregation network (MFANet). These modules are designed to effectively incorporate the temporal aspects of LiDAR data and enhance the performance of 3D object detection.

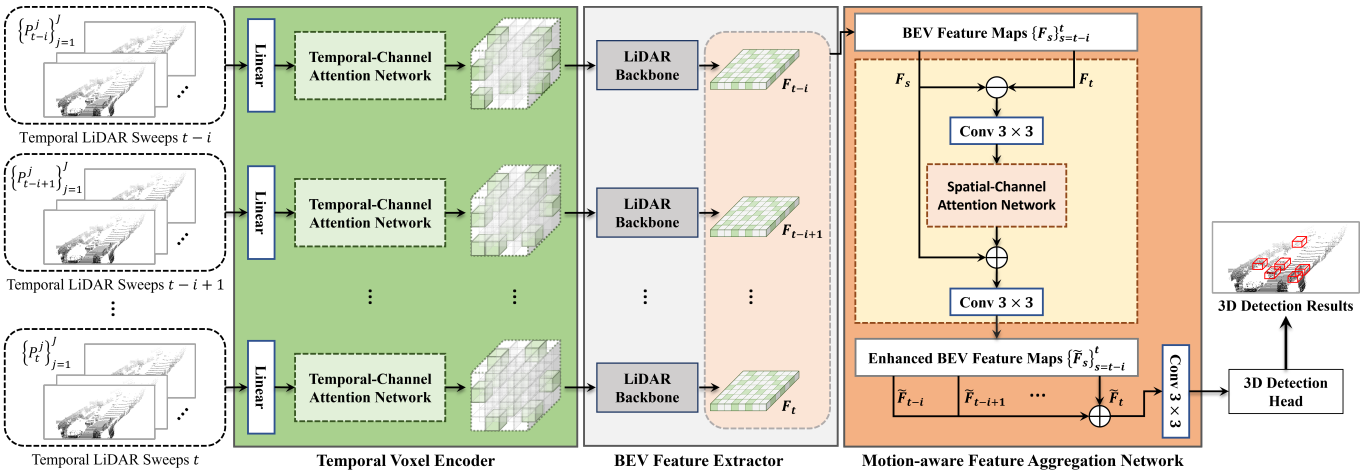

Figure 1. Overall architecture of the proposed TM3DOD method.

In the TVE module, the input consists of consecutive LiDAR points obtained from temporal laser sweeps. The goal of the TVE module is to produce voxel representative features over time. In order to achieve this, the TVE module employs a temporal channel attention network, which assigns varying attention weights to the LiDAR points within each voxel based on their temporal relationships. This attention mechanism enables the encoder to focus on the important channels and timestamps of point features while being robust against irrelevant or noisy points within the voxel. By leveraging the temporal relationships between LiDAR scans, the TVE module effectively encodes the points within each voxel, taking into account their varying influences across different laser sweeps.

Once the LiDAR points have been organized into voxels using the TVE module, shared-weight LiDAR backbone networks are utilized to process N sequential point sets and generate BEV spatial feature maps. This backbone network plays a crucial role in extracting discriminative features from the BEV representations, which are essential for accurate 3D object detection.

Following utilization of the TVE module and the LiDAR backbone networks, the MFANet module is employed for motion-aware feature aggregation. This module aims to enhance the current feature map by considering both the current feature map and the previously generated feature maps. It involves generating motion features by quantifying the temporal variation between two consecutive visual feature maps. By capturing the changes in feature representations over time, the motion features provide valuable information about object dynamics and improve the discriminative power of the feature maps. In order to incorporate the motion features into the current feature map, each motion feature is added to the corresponding BEV feature, and the resulting features are added. This enables

the TM3DOD method to effectively aggregate and integrate the temporal information into the feature maps, leading to enhanced 3D object detection performance.

In summary, the proposed TM3DOD method leverages the TVE module to capture the temporal relationships between LiDAR scans, the LiDAR backbone networks for feature extraction from BEV representations, and the MFANet module for motion-aware feature aggregation. By effectively incorporating the temporal aspects of LiDAR data, TM3DOD aims to improve the accuracy and robustness of 3D object detection in dynamic environments.

Please refer to Figure 1 for a visual representation of the overall TM3DOD architecture.

3.2. Temporal Voxel Encoder (TVE)

Point Feature Extraction: To encode point features in the temporal domain, we divided the points within the temporal LiDAR sweeps into fixed time intervals. This process is illustrated in Figure 2. For each voxel, let P_{t-i}^{j} represent the $(t-i)^{th}$ point in the j^{th} time interval. To capture the temporal relationships, we calculated the relative position of each point P_{t-i}^{j} with respect to the mean position of the points within the voxel at time j, denoted as $(x - x_d, y - y_d, z - z_d)$. Here, (x, y, z) represents the position of P_{t-i}^{j}, and (x_d, y_d, z_d) represents the mean position of the points at time j. This distance computation was performed for all points within the voxel and across all voxels. The resulting point features were then processed through a linear layer to extract relevant information.

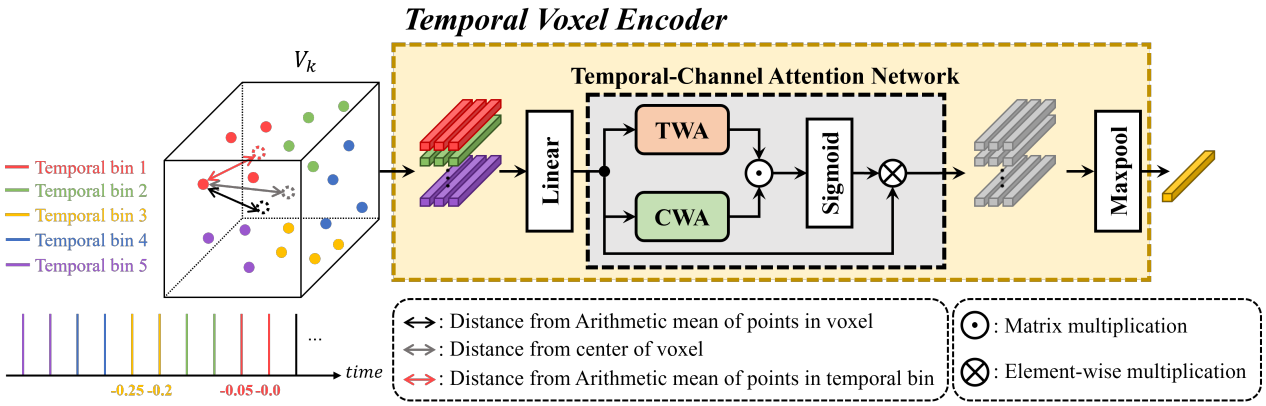

Figure 2. The proposed network of the temporal voxel encoder (TVE).

Temporal Channel Attention Network (TCANet): In TCANet, we generated attentive voxel features by focusing on the crucial channels and timestamps of the point features. Let $V^k \in \mathbb{R}^{N^k \times C}$ be a set of point features of the k^{th} voxel and $v_j^k \in \mathbb{R}^{N_j^k \times C}$ be a set of point features of the j^{th} sweep in the k^{th} voxel (i.e., $N^k = \sum_{j=1}^{J} N_j^k$). Note that N is the number of LiDAR points in a voxel. Then, we applied temporal-wise attention (TWA) and channel-wise attention (CWA) to extract the temporal-wise response $U^k \in \mathbb{R}^J$ and channel-wise response $E^k \in \mathbb{R}^C$, respectively. First, in TWA, the temporal response $H^k \in \mathbb{R}^J$ is produced through max pooling across the point-wise dimension and channel-wise dimension on each v_j^k, which is a subset of V^k. Similarly, the channel response $G^k \in \mathbb{R}^C$ is produced through max pooling across the point-wise dimension of V^k. In order to explore temporal- and channel-wise correlation, following the attention mechanism called SENet [39], TWA and CWA extract the temporal-wise attention $U^k \in \mathbb{R}^{J \times 1}$ and $E^k \in \mathbb{R}^{1 \times C}$ as follows:

$$U^k = W_{TWA}^2 \delta(W_{TWA}^1 H^k) \tag{1}$$

$$E^k = W_{CWA}^2 \delta(W_{CWA}^1 G^k) \tag{2}$$

where $W_{\cdot}^1$ and $W_{\cdot}^2$ are the weight parameters of two fully connected layers in each attention module and δ is the ReLU function. Next, we obtain the attention matrix $M^k \in \mathbb{R}^{J \times C}$ as follows:

$$M^k = \sigma(U^k \odot E^k) = [m_1^k, ..., m_J^k] \in \mathbb{R}^{J \times C} \tag{3}$$

where $\odot$ denotes the matrix multiplication operation and σ denotes the sigmoid function, which is employed to normalize the values of the attention matrix to the range of [0,1]. The temporal-wise weighted feature Z^k for k^{th} voxel is calculated as follows:

$$Z^k = maxpool[m_1^k \otimes v_1^k, ..., m_J^k \otimes v_J^k] \in \mathbb{R}^C \tag{4}$$

where $\otimes$ denotes the element-wise multiplication operation and max pooling is performed across the voxel-wise dimension.

3.3. Motion-Aware Feature Aggregation Network (MFANet)

The goal of MFANet is to leverage the temporal information from consecutive LiDAR sweeps and incorporate it into the current feature map to improve the accuracy of 3D object detection. MFANet is illustrated in Figure 3. The BEV features, denoted as F_s from times $t - i$ to t, are extracted from the LiDAR backbone network and passed to MFANet. Note that F_t indicates the BEV feature at time t, and this is the current BEV feature.

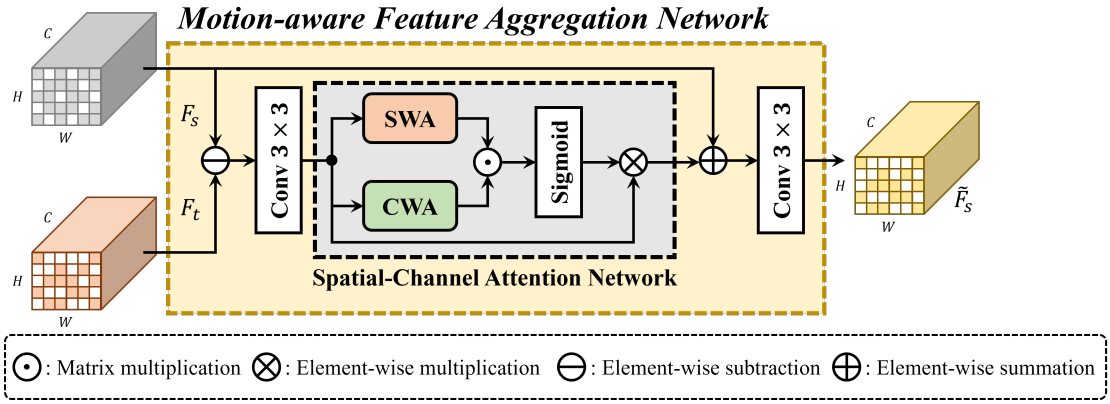

Figure 3. Proposed network of motion-aware feature aggregation network (MFANet).

In order to generate each motion feature map in the MFANet, a two-step process is followed. First, the differences between two adjacent BEV feature maps are calculated. These difference maps are then passed through a convolutional network to extract the motion features M_t:

$$M_s = (\text{conv}_{3\times3}(F_s - F_t)), \tag{5}$$

In order to further enhance the motion features and enable the network to focus on important spatial and channel-wise information, the spatial-wise attention (SWA) and channel-wise attention (CWA) mechanisms are employed. In SWA, the spatial-wise attention weights are obtained by performing max pooling across the channel-wise dimension of the motion feature maps. This allows the network to highlight important spatial regions in the motion features. Similarly, in CWA, the channel-wise attention weights are obtained by performing max pooling across the pixel-wise dimension of the motion feature maps. This enables the network to emphasize important channels in the motion features. In order to generate the attention weights, we used a sigmoid function which normalized the feature maps between 0 and 1, similar to the TCANet approach. In order to incorporate the generated attention weights into the current feature map, they were multiplied element-wise with the corresponding difference map of two adjacent BEV feature maps:

$$\tilde{M}_s = \sigma(CWA(M_s) \odot SWA(M_s)) \otimes M_s, \tag{6}$$

where $\tilde{M}_s$ denotes the enhanced motion feature from CWA and SWA. Finally, the enhanced motion feature was then added to the corresponding BEV feature map and passed into the convolution network to generate the final aggregated feature maps $\tilde{F}_s$:

$$\tilde{F}_s = \text{conv}_{3\times3}(\tilde{M}_s + F_s), \tag{7}$$

4. Experiments

4.1. nuScenes

The nuScenes dataset is a popular benchmark for 3D object detection in the field of autonomous driving. It comprises a large collection of annotated driving scenes specifically designed for evaluating 3D object detection algorithms. The dataset consists of more than 1000 driving scenes, with 700 scenes designated for training, 150 for validation, and 150 for testing. In the context of 3D object detection, the nuScenes dataset provides annotated point cloud data captured by LiDAR sensors. The LiDAR point clouds are represented by coordinates (x, y, z), a reflectance value (r), and the time lag (Δ t) between each point and the corresponding keyframe. In order to enhance the density of the dataset, the point clouds from the previous 10 sweeps were utilized, resulting in approximately 300,000 point clouds per scene. This allowed for a comprehensive representation of the scene and enabled accurate 3D object detection. The dataset includes annotations for various object categories relevant to autonomous driving, such as bicycles, buses, cars, motorcycles, pedestrians, trailers, trucks, construction vehicles, and traffic cones. For the 3D detection performance metrics, we used the mAP and nuScenes detection score (NDS) [18] as suggested for the nuScenes dataset.

4.2. Implementation Details

Architecture: We utilized two widely used 3D LiDAR object detectors as single-frame baseline detectors. First, we adopted PointPillar [2] in our framework since it has the advantage of low computation with 2D BEV representation. Second, we chose CenterPoint [19] with a VoxelNet backbone as a baseline so that we could produce a detailed 3D representation. The range of the point cloud was within $[-51.2, 51.2] \times [-51.2, 51.2] \times [-5.0, 3.0]$ m on the (x, y, z) axis, which was voxelized, with each voxel size being $0.2 \times 0.2 \times 8.0$ m in the PointPillar backbone. On the other hand, we considered the points located within $[-54.0, 54.0] \times [-54.0, 54.0] \times [-5.0, 3.0]$ m on the (x, y, z) axis in the VoxelNet backbone. The grid size for grouping points was set to $0.075 \times 0.075 \times 0.2$ m, and this partitioning led to a voxel structure $1440 \times 1440 \times 40$ in size. To implement the TVE module, we used the input features which contained original geometric point information and the distance from the mean of the points in the temporal bin. We set bin periods of 2 and 5 to extract motion information at the point level.

Training and Inference: PointPillar [2] and CenterPoint [19] were used as single-frame baselines to implement the proposed TM3DOD method. We trained the entire network using a one-cycle learning rate policy for 50 epochs, with a maximum learning rate of 0.003 for PointPillar and 0.001 for CenterPoint, using the whole nuScenes training set. The sequence length was set to 2 s, which contained 4 keyframes. The proposed model was trained using data augmentation methods, including the random flipping, rotation, scaling, and ground truth box sampling used in [1]. We utilized the loss function suggested for the anchor-based [2] or anchor-free baseline [19] to train the proposed model. An Adam optimizer was used to optimize the loss function. We trained our proposed method on 4 NVIDIA RTX 2080 Ti GPUs with a batch size of 8 in the PointPillar-based model and a batch size of 4 in the CenterPoint-based model.

4.3. Performance Analysis

In Table 1, we present a comparison between the performance of the proposed TM3DOD method and several state-of-the-art 3D object detectors [2,6,19,33,40–47] on the nuScenes test benchmark.

Table 1. Performance comparison on nuScenes *test* benchmark.

Method	Input	Car	Truck	Bus	Trailer	C.V	Ped.	Motor.	Bicycle	T.C	Barrier	NDS	mAP
InfoFocus [40]	Single frame	77.9	31.4	44.8	37.3	10.7	63.4	29.0	6.1	46.5	47.8	39.5	39.5
PointPillars [2]	Single frame	68.4	23.0	28.2	23.4	4.1	59.7	27.4	1.1	30.8	38.9	45.3	30.5
3DSSD [6]	Single frame	81.2	47.2	61.4	30.5	12.6	7.2	36.0	8.6	31.1	47.9	56.4	42.6
PointPainting [41]	Single frame	77.9	35.8	36.2	37.3	15.8	73.3	41.5	24.1	62.4	60.2	58.1	46.4
PointPillars_DSA [42]	Single frame	81.2	43.8	57.2	47.8	11.3	73.3	32.1	7.9	60.6	55.3	59.2	47.0
SSN V2 [43]	Single frame	82.4	41.8	46.1	48.0	17.5	75.6	48.9	24.6	60.1	61.2	61.6	50.6
3DCVF [33]	Single frame	83.0	45.0	48.8	49.6	15.9	74.2	51.2	30.4	62.9	65.9	62.3	52.7
CBGS [44]	Single frame	81.1	48.5	54.9	42.9	10.5	80.1	51.5	22.3	70.9	65.7	63.3	52.8
CVCNet [45]	Single frame	82.7	46.1	45.8	46.7	20.7	81.0	61.3	34.3	69.7	69.9	64.2	55.8
HotSpotNet [46]	Single frame	83.1	50.9	56.4	53.3	23.0	81.3	63.5	36.6	73.0	**71.6**	66.6	59.3
CyliNet [47]	Single frame	85.0	50.2	56.9	52.6	19.1	84.3	58.6	29.8	**79.1**	69.0	66.1	58.5
CenterPoint [19]	Single frame	85.2	53.5	63.6	56.0	20.0	84.6	59.5	30.7	78.4	71.1	67.3	60.3
3DVID [13]	Multi-frame	79.7	33.6	47.1	43.0	18.1	76.5	40.7	7.9	58.8	48.8	-	45.4
TCTR [14]	Multi-frame	83.2	51.5	63.7	33.0	15.6	74.9	54.0	22.6	52.5	53.8	-	50.5
Method of [16]	Multi-frame	86.2	57.2	**67.2**	35.5	14.6	**85.5**	58.1	37.0	71.3	66.2	66.7	59.0
Our TM3DOD (with PP)	Multi-frame	78.8	43.1	52.1	48.1	15.0	71.8	50.2	15.8	59.2	50.3	60.1	48.44
Our TM3DOD (with CP)	Multi-frame	**86.7**	**56.0**	64.6	**58.3**	**27.7**	81.2	**72.7**	**45.2**	77.4	69.3	**69.94**	**63.91**

The results demonstrate that the proposed TM3DOD method achieved superior performance compared with the existing 3D object detection methods. Specifically, when compared with the CenterPoint baseline, TM3DOD exhibited a notable improvement of 2.64% in terms of NDS and 3.61% in terms of mAP. These improvements highlight the effectiveness and robustness of the proposed method in accurately detecting 3D objects in LiDAR-based systems.

Overall, the experimental results presented in Table 1 demonstrate that the proposed TM3DOD method outperformed the existing state-of-the-art approaches and established a new benchmark for 3D object detection in LiDAR-based systems, specifically on the nuScenes dataset.

4.4. Ablation Study

In this section, we present some ablation studies to verify the effect of the ideas in the proposed TM3DOD method. Note that our ablation study was performed using PointPillar [2] as a single-frame baseline. Since nuScenes has many training samples compared with KITTI (28,130 versus 3712), the multiple models were trained with time efficiency using the downsampled training set (namely the mini-training set) and validated using the fully valid set. The mini-training set contained about 4000 samples uniformly sampled from the fully trained set.

First, Table 2 demonstrates the effectiveness of the proposed modules used in TM3DOD. We trained the model in combination with the baseline and proposed modules (i.e., TVE and MFANet) while using the mini-training set to validate each of the proposed modules. When MFANet was added to the baseline, it offered 2.19% and 1.32% performance gains over the baseline in the mAP and NDS metrics, respectively. For both TVE and MFANet, it yielded a 4.18% gain in mAP and a 2.40% gain in NDS over the single-frame baseline. Regarding computational complexity aspects, the runtime increased by 13 ms and 8 ms for the MFANet and TVE modules, respectively.

Table 2. The ablation study conducted on the nuScenes *valid* set.

	Motion-Aware Feature Aggregation	Temporal Voxel Encoder	mAP (%)	NDS (%)	Runtime
Single-Frame Baseline			41.18	55.53	31 ms
Our TM3DOD	✓		$43.37_{\uparrow 2.19}$	$56.85_{\uparrow 1.32}$	44 ms
	✓	✓	$45.36_{\uparrow 4.18}$	$57.93_{\uparrow 2.40}$	52 ms

In our analysis, we investigated the impact of the components of the TVE module on the performance of the model. We conducted experiments and summarized the results in Table 3. To ensure a fair comparison, we trained the single-frame baseline model using the mini-training set. We conducted an ablation study on the TVE module using two steps. First, we verified the effectiveness of temporal channel attention. Second, we analyzed the performance according to the temporal bin period in the TVE module. When we added the temporal channel attention mechanism to the voxel encoder, we observed an improvement in the mean average precision (mAP) compared with the baseline model. It improved the mAP by 1.25%. We also show the contribution of the temporal bin period in Table 3. The performance increased by 1.54% when using a temporal bin period of two and 2.67% when using a temporal bin period of five over the baseline. Furthermore, we explored the effect of different temporal bin periods on performance. By setting the temporal bin period to two and five, we observed additional improvements in the mAP by 3.29% over the baseline. These results demonstrate the effectiveness of incorporating the temporal channel attention mechanism and optimizing the temporal bin period in the TVE module, leading to improved performance in 3D object detection tasks.

Table 3. The ablation study of TCANet conducted on the nuScenes *valid* set.

Modules		Performance	
Temporal Bin Period	**Temporal-Channel Attention**	**mAP(%)**	**Δ**
-	-	37.6	-
-	✓	38.85	+1.25
2	✓	39.25	+1.65
5	✓	40.27	+2.67
2 and 5	✓	40.89	+3.29

For further analysis, we compared the qualitative results between the single-frame baseline method and the proposed method on the nuScenes validation dataset. In Figure 4, the red boxes represent the ground truth, and the green boxes represent the predictions. In order to verify the effectiveness of the proposed method, we compared the results for a frame that had both static objects and moving objects. In the baseline method, the static objects (blue boxes) were detected well, while the moving objects (orange boxes) were not detected. In contrast, our proposed method detected both static and moving objects.

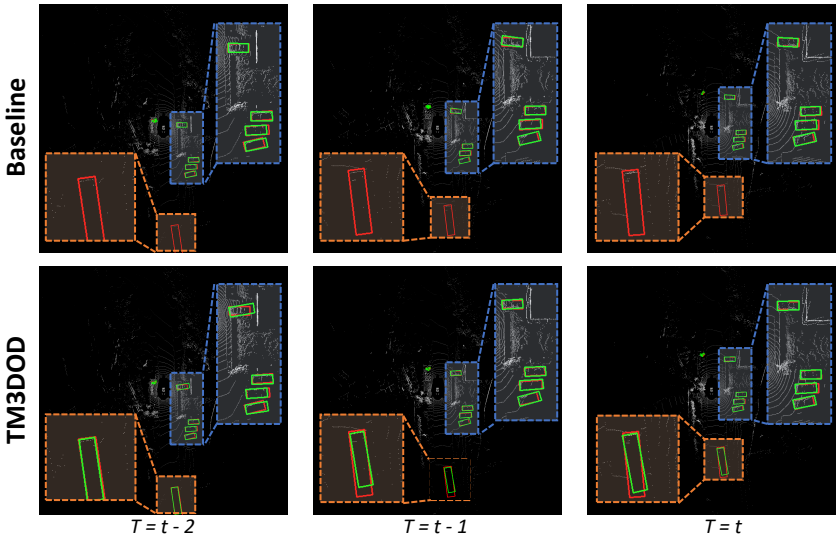

Figure 4. Comparison of qualitative results between the single-frame baseline method and the proposed method on the nuScenes validation dataset.

5. Conclusions and Future Work

In conclusion, the TM3DOD framework presents a novel approach to 3D object detection in LiDAR-based systems. By incorporating the temporal voxel encoder (TVE) and motion-aware feature aggregation network (MFANet) modules, TM3DOD effectively utilized the temporal information from consecutive LiDAR sweeps to improve the accuracy and performance of object detection. The TVE module captured motion over time with its temporal channel attention network, enhancing the encoder's ability to focus on important points and improve robustness. The MFANet module further improved the detection results by aggregating the motion features using the current feature map. Our experiments on the nuScenes dataset demonstrated the effectiveness of TM3DOD compared with LiDAR-only baselines. The proposed method achieved significant improvements in accuracy and showed comparable performance to state-of-the-art methods. Nevertheless, there were some performance decreases for small or stationary objects such as pedestrians, traffic cones, and barriers. Even if we use the multi-frame LiDAR point cloud data, it is difficult to improve performance for such categories, which show similar characteristics in terms of their LiDAR data aspects. As a future work, we need to research 3D object detection with sensor fusion to overcome these difficult cases.

Author Contributions: Conceptualization, G.P. and J.W.C.; software G.P. and J.K. (Junho Koh); validation, G.P., J.K. (Junho Koh) and J.K. (Jisong Kim); writing, G.P., J.K. (Junho Koh), J.K. (Jisong Kim), J.M. and J.W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00421129).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: nuScenes dataset at https://www.nuscenes.org/.

Acknowledgments: This research was supported by the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00421129).

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

References

1. Yan, Y.; Mao, Y.; Li, B. Second: Sparsely embedded convolutional detection. *Sensors* **2018**, *18*, 3337. [CrossRef] [PubMed]
2. 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.
3. 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.
4. Deng, J.; Shi, S.; Li, P.; Zhou, W.; Zhang, Y.; Li, H. Voxel r-cnn: Towards high performance voxel-based 3D object detection. *arXiv* **2020**, arXiv:2012.15712.
5. 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.
6. Yang, Z.; Sun, Y.; Liu, S.; Jia, J. 3dssd: Point-based 3D single stage object detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11040–11048.
7. Shi, W.; Rajkumar, R. Point-gnn: Graph neural network for 3D object detection in a point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1711–1719.
8. Chen, Y.; Liu, S.; Shen, X.; Jia, J. Fast point r-cnn. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9775–9784.
9. Yang, Z.; Sun, Y.; Liu, S.; Shen, X.; Jia, J. Std: Sparse-to-dense 3D object detector for point cloud. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1951–1960.
10. 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.
11. He, C.; Zeng, H.; Huang, J.; Hua, X.S.; Zhang, L. Structure aware single-stage 3D object detection from point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11873–11882.

12. Huang, R.; Zhang, W.; Kundu, A.; Pantofaru, C.; Ross, D.A.; Funkhouser, T.; Fathi, A. An lstm approach to temporal 3D object detection in lidar point clouds. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 266–282.
13. Yin, J.; Shen, J.; Guan, C.; Zhou, D.; Yang, R. Lidar-based online 3D video object detection with graph-based message passing and spatiotemporal transformer attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11495–11504.
14. Yuan, Z.; Song, X.; Bai, L.; Wang, Z.; Ouyang, W. Temporal-channel transformer for 3D lidar-based video object detection for autonomous driving. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 2068–2078. [CrossRef]
15. Emmerichs, D.; Pinggera, P.; Ommer, B. VelocityNet: Motion-Driven Feature Aggregation for 3D Object Detection in Point Cloud Sequences. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 13279–13285.
16. Xiong, Z.; Ma, H.; Wang, Y.; Hu, T.; Liao, Q. LiDAR-based 3D Video Object Detection with Foreground Context Modeling and Spatiotemporal Graph Reasoning. In Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 2994–3001.
17. Yang, Z.; Zhou, Y.; Chen, Z.; Ngiam, J. 3D-MAN: 3D multi-frame attention network for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1863–1872.
18. 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.
19. 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.
20. Chen, Y.; Tai, L.; Sun, K.; Li, M. Monopair: Monocular 3D object detection using pairwise spatial relationships. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12093–12102.
21. Ma, X.; Wang, Z.; Li, H.; Zhang, P.; Ouyang, W.; Fan, X. Accurate monocular 3D object detection via color-embedded 3d reconstruction for autonomous driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6851–6860.
22. Ku, J.; Pon, A.D.; Waslander, S.L. Monocular 3D object detection leveraging accurate proposals and shape reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11867–11876.
23. Li, P.; Chen, X.; Shen, S. Stereo r-cnn based 3D object detection for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7644–7652.
24. Chen, Y.; Liu, S.; Shen, X.; Jia, J. Dsgn: Deep stereo geometry network for 3D object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12536–12545.
25. Wang, Y.; Chao, W.L.; Garg, D.; Hariharan, B.; Campbell, M.; Weinberger, K.Q. Pseudo-lidar from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8445–8453.
26. Shi, S.; Wang, Z.; Shi, J.; Wang, X.; Li, H. From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 2647–2664. [CrossRef] [PubMed]
27. Hu, P.; Ziglar, J.; Held, D.; Ramanan, D. What You See Is What You Get: Exploiting Visibility 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. 11001–11009.
28. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-view 3D object detection network for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1907–1915.
29. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S.L. Joint 3D proposal generation and object detection from view aggregation. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 1–8.
30. Liang, M.; Yang, B.; Chen, Y.; Hu, R.; Urtasun, R. Multi-task multi-sensor fusion for 3D object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7345–7353.
31. Liang, M.; Yang, B.; Wang, S.; Urtasun, R. Deep continuous fusion for multi-sensor 3D object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 641–656.
32. Qi, C.R.; Liu, W.; Wu, C.; Su, H.; Guibas, L.J. Frustum pointnets for 3D object detection from rgb-d data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 918–927.
33. Yoo, J.H.; Kim, Y.; Kim, J.; Choi, J.W. 3D-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3D object detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 8–14 September 2020; pp. 720–736.
34. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
35. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.

36. Ngiam, J.; Caine, B.; Han, W.; Yang, B.; Chai, Y.; Sun, P.; Zhou, Y.; Yi, X.; Alsharif, O.; Nguyen, P.; et al. Starnet: Targeted computation for object detection in point clouds. *arXiv* **2019**, arXiv:1908.11069.
37. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
38. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In Proceedings of the NIPS 2014 Workshop on Deep Learning, Montreal, BC, Canada, 13 December 2014.
39. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
40. Wang, J.; Lan, S.; Gao, M.; Davis, L.S. Infofocus: 3D object detection for autonomous driving with dynamic information modeling. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 8–14 September 2020; pp. 405–420.
41. Vora, S.; Lang, A.H.; Helou, B.; Beijbom, O. Pointpainting: Sequential fusion for 3D object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4604–4612.
42. 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.
43. Zhu, X.; Ma, Y.; Wang, T.; Xu, Y.; Shi, J.; Lin, D. Ssn: Shape signature networks for multi-class object detection from point clouds. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 8–14 September 2020; pp. 581–597.
44. Zhu, B.; Jiang, Z.; Zhou, X.; Li, Z.; Yu, G. Class-balanced grouping and sampling for point cloud 3D object detection. *arXiv* **2019**, arXiv:1908.09492.
45. Chen, Q.; Sun, L.; Cheung, E.; Yuille, A.L. Every view counts: Cross-view consistency in 3D object detection with hybrid-cylindrical-spherical voxelization. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21224–21235.
46. Chen, Q.; Sun, L.; Wang, Z.; Jia, K.; Yuille, A. Object as hotspots: An anchor-free 3D object detection approach via firing of hotspots. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 8–14 September 2020; pp. 68–84.
47. Rapoport-Lavie, M.; Raviv, D. It's All Around You: Range-Guided Cylindrical Network for 3D Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2992–3001.