Coupling and Decoupling: Towards Temporal Feedback for 3D Object Detection

Yubo Cui, Zhikang Zou, Xiaoqing Ye, Xiao Tan, Zhiheng Li, and Zheng Fang*, *Member, IEEE*

***Abstract*—3D object detection has garnered significant attention within the academic community, primarily due to its broad utility in domains such as autonomous driving and robotics. Prior research efforts have predominantly concentrated on leveraging temporal contextual information embedded within sequential data to enhance the current feature representations. However, a notable limitation of these endeavors lies in their inadequate treatment of the inherent noise present within historical sequences, thereby constraining the efficiency of fusion methods. In this paper, we propose a new temporal feedback network, named TFNet, to model and correct the temporal noise by designing a *coupling-decoupling* mechanism. Central to our approach are two distinct modules: (i) Foreground Feature Enhancement, which amplifies sparse instance details across temporal frames, thereby furnishing essential local information priors for subsequent fusion; and (ii) Coupling-Decoupling Feature Interaction, designed to first aggregate temporal contextual information and then disentangle fusion features into frame-specific representations. Leveraging a feedback strategy, this module can adaptively enhance useful information and eliminate noise within individual frame features. Empirical evaluations conducted on the nuScenes benchmark demonstrate the effectiveness of TFNet, achieving the new state-of-the-art performance without any bells and whistles.**

***Index Terms*—3D object detection, temporal feedback, coupling-decoupling.**

I. INTRODUCTION

RECENTLY, with the development of autonomous driving technology, 3D object detection [1]–[4] has received increasing attention as a crucial component. By directly outputting the 3D coordinates and bounding boxes of surrounding objects, 3D object detection provides more valuable information for downstream tasks such as object tracking and distance measurement. Moreover, compared to visual images, 3D point clouds contain richer geometric information, making them well-suited for 3D detection tasks. However, due to the sparsity and unordered characteristic of point clouds, LiDAR-based 3D object detection still faces challenges, such as difficulty in effectively detecting objects when they are too far away and the point cloud becomes too sparse.

To address these challenges, some methods try to fuse the 2D image to provide more rich appearance information. However, the misalignment between different modalities poses challenges in the fusion process. Meanwhile, in some poorly lit scenarios, image data may not provide sufficient useful information. Therefore, there is enormous research value in studying the single-modal approach of pure point clouds. To provide more additional LiDAR information, temporal 3D detection [5]–[8] has been proposed to compensate for the shortcomings of single-frame detection. By exploring the temporal relationships between frames, prior information from historical frames can be learned and utilized to compensate for insufficient information in the current frame, leading to the improvement of detection performance. For example, ConvLSTM [5] proposes an LSTM module to recursively combine the previous memory feature with the current to enhance the current representation. 3DVID [6] utilizes GRU to aggregate the previous temporal information to enhance the current frame feature. Furthermore, 3D-MAN [7] proposes an attention-based proposal-level fusion to fuse the previous frames' proposals into the current. MGTANet [8] utilizes motion information to align the other frame features to the

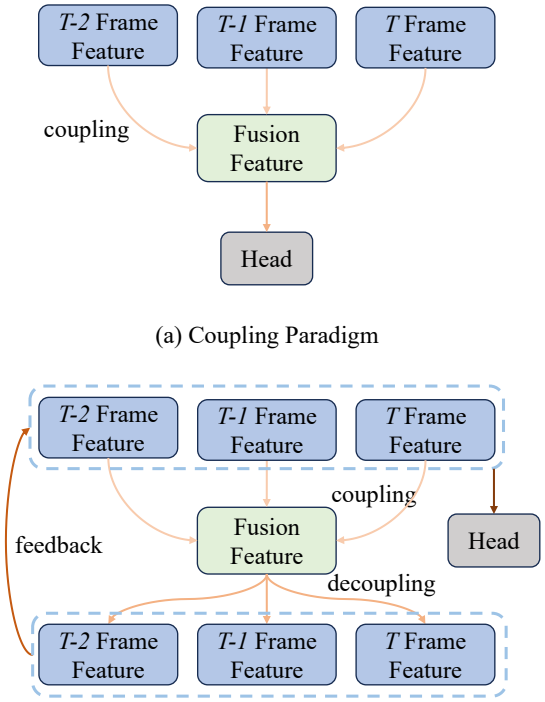

Fig. 1. (a). Previous coupling method. (b). Our proposed coupling-decoupling method. Compared to the coupling method, the coupling-decoupling method has a feedback mechanism to eliminate the noise. Here we use color to represent the feature flow.

Yubo Cui, Zhiheng Li and Zheng Fang are with the Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China; Zhikang Zou, Xiaoqing Ye and Xiao Tan are with the department of computer vision technology, Baidu Inc. Corresponding author: Zheng Fang, e-mail: fangzheng@mail.neu.edu.cn

This work was supported in part by the National Natural Science Foundation of China under Grants 62073066, in part by the Fundamental Research Funds for the Central Universities under Grant N2226001, and in part by 111 Project under Grant B16009.

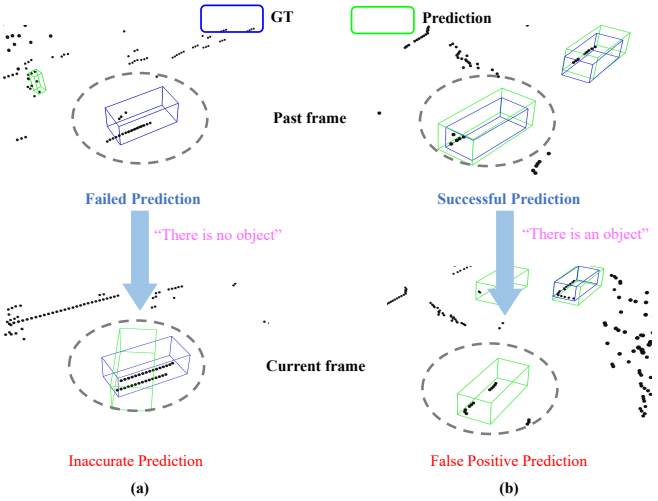

Fig. 2. The noise information from the past frame conflicts with the current frame, leading to inaccurate detection or false positive detection.

current frame and uses deformable attention [9] to fuse these aligned features.

Although these methods have achieved significant detection performance, they can be regarded as a feature coupling fusion paradigm that aggregates multiple features into single fused features, as shown in Figure 1(a). However, this paradigm lacks feedback correction for each single-frame feature based on the fusion feature, thus limiting the effectiveness of coupling. As shown in Figure 2(a), in autonomous driving, if the detector fails to detect the object in the past frame, the error information 'there is no object." within the past feature would be coupled and propagated to the current frame. This noise could lead to situations where objects that should have been detected are missed or inaccurately detected. Meanwhile, in Figure 2(b), when an object moves too much, information about "there is an object" from past features can actually become noise, leading to false positives in the current frame.

To overcome the aforementioned issue, in this paper, we propose a new temporal 3D object detector named *TFNet (**T**emporal **F**eedback **Net**work).* Different from previous methods, we introduce the feedback mechanisms by leveraging a "coupling-decoupling" strategy to adjust and correct the single frame features, as shown in Figure 1(b). After obtaining the temporal feature through feature coupling, we further decouple the temporal feature by combining it with each frame feature, completing the feedback of temporal information for each single frame feature. Through this temporal feedback mechanism, each frame can selectively adjust its features according to its own situation, correct interference and compensate for insufficiency, leading to better temporal fusion. For Figure 2(a), we could feedback the current frame information of the detected object to the past frame through our "coupling-decoupling" strategy. This temporal feedback process eliminates noise and enables the object to be detected in the past frame, enhancing the past frame feature and further improving the detection accuracy of the current frame through temporal fusion.

Following the guidance principle of the "coupling-decoupling" strategy, we propose a *Coupling-Decoupling Feature Interaction* module to fuse the multi-frame features in local and global feature scopes. In both scopes, the temporal information is first learned through multi-frame feature aggregation, and then decomposed back into multi-frame features. By feeding back the learned temporal information and combining it with the actual situation of each frame, the noise information could be eliminated and useful information could be enhanced. Meanwhile, the local-to-global fusion strategy also leads to more comprehensive multi-frame feature fusion. Moreover, to alleviate the sparsity of point clouds and better capture the local information, we also propose a *Foreground Feature Enhancement* module to enhance local foreground features across frames. Through proposing dynamic group attention, the proposed module enhances the local representation no matter the directions and size of objects.

Overall, our contributions could be summarized as follows:

- We propose a Foreground Feature Enhancement module to enhance sparse multi-frame foreground features through dynamic aggregation.
- We propose a "coupling-decoupling" fusion strategy and introduce the feedback mechanism for better multi-frame feature fusion in temporal detection.
- We propose several modules to fuse multi-frame features from local to global, and further propose a new temporal detector named TFNet.
- We achieve new state-of-the-art 3D object detection performance on the nuScenes benchmark. Besides, the proposed modules can be plug-and-play in single-frame frameworks, demonstrating the versatility ability of our method.

II. RELATED WORK

A. Single Frame 3D Object Detection

F-PointNet [10] is the first work attempting to detect 3D objects with point clouds in autonomous driving scenarios. However, they still depend on a 2D detector to obtain a local point cloud frustum. To get rid of images, PointRCNN [2] directly inputs all point clouds of the scenarios and detects 3D objects by a two-stage RCNN-like framework. Furthermore, STD [11] introduces spherical anchors and proposes PointsPool to generate compact representations for proposals. FARP-Net [12] adaptively aggregates local-global features and designs a relation-aware proposal network for high-quality 3D object detection. However, limited by the unordered structure of point clouds, point-based methods usually suffer from slow running speeds. Differently, VoxelNet [13] converts the points into 3D voxel grids and uses 3D CNN-based network to detect objects, leading to a faster running speed. Based on that, SECOND [3] further utilizes sparse convolution to improve running efficiency. Moreover, CenterPoint [14] uses the CenterHead-manner to regress with sparse convolution backbone and achieve a good balance between accuracy and speed. SP-Det [15] presents a saliency prediction-based 3D object detector to improve the robustness for the sparsity. VPFNet [16] additionally fuses the LiDAR and stereo data by designing virtual point to bridge the resolution gap between

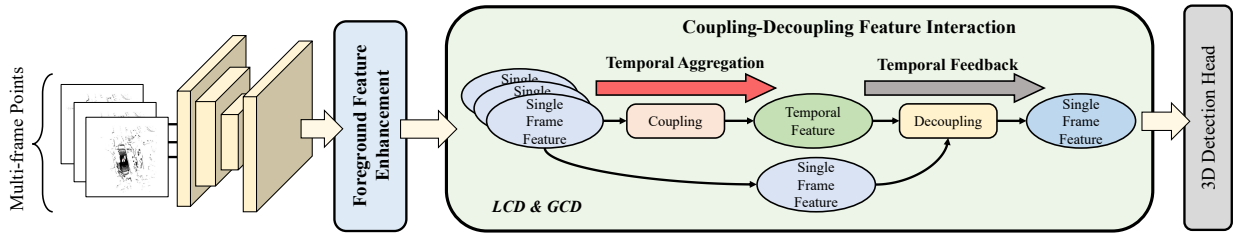

Fig. 3. Illustration of the proposed TFNet. After multi-frame feature extraction, we first sample and enhance the multi-frame foreground features in the Foreground Feature Enhancement (FFE) module. Then, in the Coupling-Decoupling Feature Interaction (CDFI) module, we perform multi-frame feature interaction in both local and global scopes by the proposed Local Coupling Decoupling (LCD) and Global Coupling Decoupling (GCD) respectively. We use the fused current frame feature for final detection.

the two modalities. However, compared to raw point clouds, the voxel representation inevitably loses some information. Therefore, PV-RCNN [4] takes both advantages of points and voxels, and proposes RoI-grid pooling to refine the proposals. In addition, some work focuses on improving the effectiveness of point or voxel representations. For example, IA-SSD [17] studies the efficiency problem of point-based methods and reduces the computational cost by sampling foreground points. Focal conv [18] enhances the capability of sparse convolution by making feature sparsity learnable with position-wise importance prediction. LargeKernel3D [19] proposes spatial-wise partition convolution for 3D large kernels to improve the efficiency of large-kernel convolution. Recently, with the rise of transformer [20], some works [21]–[24] try to use the attention mechanism to better capture information from point cloud. Inspired by Swin [25], SST [22] proposes a window-based attention backbone to replace the common used sparse convolution, achieving better performances. Moreover, VoTr [21] and VoxSet [23] integrate the attention into backbone with voxel manner. In this paper, we propose the coupling and decoupling modules that could be plug-and-play in most single-frame detection methods.

B. Multi Frames 3D Object Detection

Nowadays, several temporal-based 3D detectors have been proposed to improve detection performance by exploring temporal information from history frames. ConvLSTM [5] extracts each frame feature with a U-Net style 3D sparse convolution network and proposes an LSTM module to combine the current feature with the hidden features. To better model the spatial-temporal relationship, 3DVID [6] proposes a graph network for spatial feature encoding and utilizes GRU to aggregate spatio-temporal features. Furthermore, TCTR [26] proposes a temporal-channel transformer module to explore the spatial, temporal, and channel correlations among different frames and enhance the target frame, leading to better temporal modeling. Moreover, TransPillars [27] uses multi-head attention to fuse different frame features at different scales. Different from learning the spatial-temporal overall features, 3D-MAN [7] proposes an attention-based fusion strategy to match and fuse each frame proposals to refine current frame proposals. MGTANet [8] utilizes motion information to align the other features to the current frame and uses deformable attention [9] to fuse these aligned features. Different from the above coupling methods, we design a "coupling-decoupling" strategy to introduce the temporal feedback for each frame feature, leading to better temporal fusion and detection results.

III. METHODOLOGY

A. Problem Definition

For single frame LiDAR-based 3D object detection, we input current t frame point cloud $\mathcal{P}_t$ and predict current I 3D bounding boxes $\{\mathcal{B}_t^i\}_{i=1}^I$. Each 3D box could be represent as $\mathcal{B}^i = (x_i, y_i, z_i, w_i, l_i, h_i, \theta_i)$, where (x_i, y_i, z_i) is the 3D box center, (w_i, l_i, h_i) is the 3D box size, and θ_i is the heading angle. Therefore, the problem could be formulated as follows:

$$\mathcal{F}(\mathcal{P}_t) \mapsto \{\mathcal{B}_t^i\}_{i=1}^I \tag{1}$$

Different from the single frame detection, 3D temporal detection usually has N consecutive frames point cloud at t frame ranging from $\mathcal{P}_{t-N}$ to $\mathcal{P}_t$. Therefore, in addition to focusing on the detection part, the temporal detection methods usually also need to focus on the fusion of multi-frame features. Besides this, temporal 3D detection is similar to single-frame detection, thus it could be formulated as follows:

$$\mathcal{F}(\{\mathcal{P}_j\}_{j=t-N}^t) \mapsto \{\mathcal{B}_t^i\}_{i=1}^I \tag{2}$$

B. Framework Overview

The overall architecture of the proposed TFNet is shown in Figure 3. Following previous works [3], [13], [14], we also extract dense BEV (Bird's-Eye View) features for N frames of point clouds, *i.e.*, $\{P_n\}_{n=t-N}^t$. We first voxelize each frame of point cloud P_n into regular voxels and then use sparse 3D convolution [3] and 2D convolution to extract BEV features $V_n \in \mathbb{R}^{H\times W\times C_1}$. We further project each V_n to the size of $H \times W \times C_2$ for channel reduction. Two main modules are proposed to fuse and enhance the multi-frame features, including foreground feature enhancement (FFE) and coupling-decoupling feature interaction (CDFI). **In the FFE**, we perform foreground segmentation and sampling on each frame to gather M foreground voxel features $F_n \in \mathbb{R}^{M\times C_2}$. To better capture the temporal foreground information from the sampled features, we propose group attention to enhance information among the multi-frame foreground features $\{F_n\}_{n=t-N}^t$. This module not only enhances foreground information but also provides local information priors for subsequent fusion. Then, to perform temporal feedback at both local and global levels, the proposed **CDFI** consists of two sub-modules, the Local

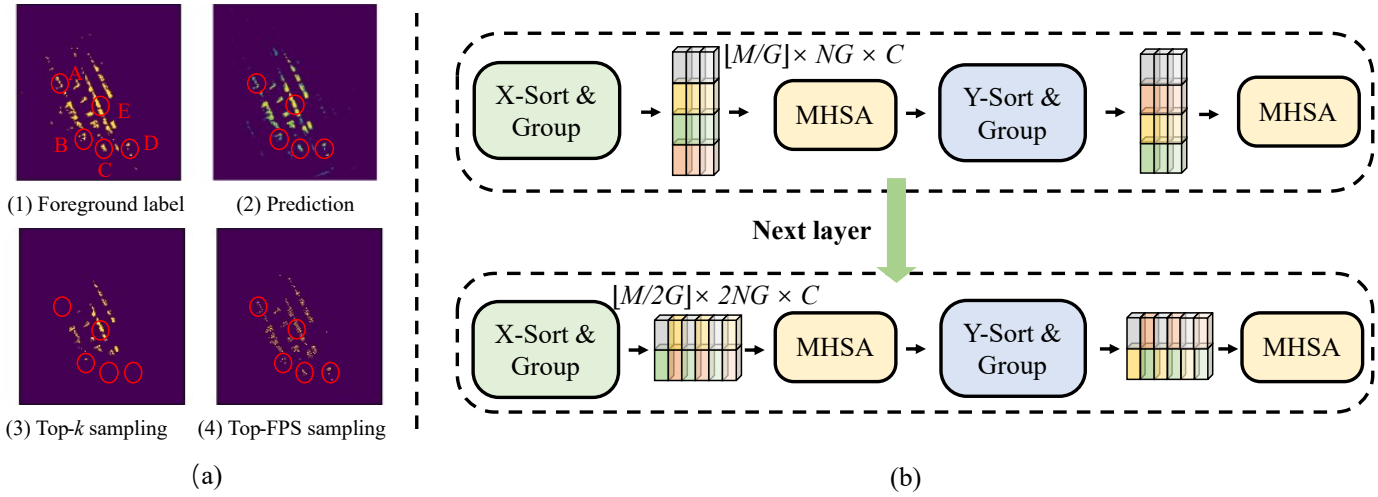

Fig. 4. (a). Different foreground sampling methods. (b). Details of the proposed Group Attention. Here we use $G = 3$ and $M = 12$ as an example. We first sort all voxels based on their X coordinates and then split them into equal-size groups. Multi-head self-attention (MHSA) is performed in each group to capture the local information. Subsequently, we repeat this process along the Y coordinate to capture additional features. Furthermore, we double the group size in the next layer to enlarge the scope of feature interaction, leading to a dynamic feature fusion and better representation.

Coupling Decoupling (**LCD**) module and Global Coupling and Decoupling (**GCD**) module. These modules are designed to facilitate the "coupling-decoupling" temporal interaction from local to global scopes, enabling a comprehensive exploration of multi-frame features. **Finally**, the decoupled current frame feature is used for detection.

C. Foreground Feature Enhancement

In order to effectively propagate temporal information, it is necessary to identify foreground instance information within each frame. However, due to the sparsity of point clouds, it is challenging to directly facilitate local feature interaction. To this end, we introduce the foreground feature enhancement module to extract and enhance the local feature representation. We first use two 2D Conv layers and sigmoid function to predict the foreground scores at each frame. Then, we sample M voxels in each frame to represent the foreground information. A common approach is to take top-k voxels based on their prediction scores. However, this approach does not guarantee the uniformity of sampling because it only samples voxels based on the score. As shown in Figure 4(a)(2), although regions (A)-(E) in the red circles are predicted as foregrounds, regions (A)-(D) have lower scores than region (E), leading to different sampling output. Because of the lower prediction scores or the fewer number of foreground voxels, these four regions have few foreground voxels after top-k sampling, as shown in Figure 4(a)(3). In contrast, region (E) has too many foreground voxels, leading to redundant information. Therefore, direct using top-k sampling leads to an imbalance results, including information loss (regions (A)-(D)) and information redundancy (region (E)).

To overcome this problem, we design a new top-FPS strategy based on the Farthest Point Sampling (FPS) [28] with voxels grid coordinates. Specifically, we first use top-k to sample M' foreground voxels based on their prediction scores, where $M' > M$, then utilize FPS on the M' voxels to further sample M voxels. As shown in Figure 4(a)(4), our sampling strategy preserves more information in the low-score regions while removing redundant information in the high-score regions, leading to a more balanced sampling results.

Group Attention. Given M sampled foreground voxel features in each frame, we further try to enhance the multi-frame foreground features $\{F_n\}_{n=t-N}^{t}$ representation. However, there are differences in the quantity and representation of features in different regions. For example, compared to features from cars, features from pedestrians have different semantics and may be fewer in number. Therefore, the global modeling from vanilla attention [20] may lead to wrong information interaction between different objects. Meanwhile, we notice that the voxels belonging to the same object tend to cluster together in a local region. Although there are differences in quantity, this clustering characteristic is clearly formed according to coordinates. Inspired by this observation, we propose Group Attention which divides all foreground features into different groups based on their coordinates and performs local information interaction. As shown in Figure. 4(b), we set the voxel number as G in each group, and then sort each frame foreground feature F_n based on their voxels coordinates (X-axis or Y-axis). Therefore, in each frame, we divide the $F_n \in \mathbb{R}^{M \times C_2}$ into grouped features $F_n \in \mathbb{R}^{\lfloor \frac{M}{G} \rfloor \times G \times C_2}$. Meanwhile, due to the short interval time between frames, adjacent groups between different frames have a temporal correlation. Therefore, we concatenate these multi-frame grouped features into $F_g \in \mathbb{R}^{\lfloor \frac{M}{G} \rfloor \times NG \times C_2}$, and apply self-attention within each group to extract temporal local information as follows:

$$F'_g = F + \text{MHSA}(\text{LN}(F_g)) \tag{3}$$

$$F''_g = F'_g + \text{FFN}(\text{LN}(F'_g)) \tag{4}$$

where MHSA$(\cdot)$, LN$(\cdot)$ and FFN$(\cdot)$ represent multi-head self-attention, layer normalization and feed-forward layer respectively. Since F'_g includes features from different frames, we

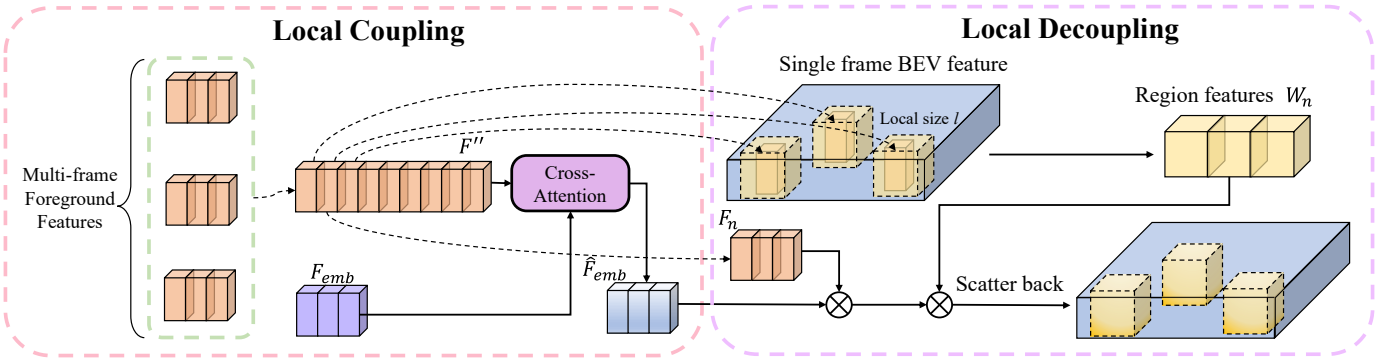

Fig. 5. Details of the proposed LCD. In the coupling process, to couple the temporal information, we perform cross-attention between a learnable foreground embedding F_{emb} with multi-frame foreground features. The multi-frame foreground features are first interacted with a learnable foreground embedding F_{emb} to couple the temporal foreground information. Then, to maintain the local consistency and expand the decoupling scopes, we decouple the temporal information to local regions centered at foreground voxels. Here we only show decoupling a single frame for clarity.

do not add position embedding in the attention.

However, as we mentioned above, objects in different categories usually occupy different sizes of space, and objects oriented in different directions also have different spaces in different directions. Consequently, fixed local settings cannot adapt to all objects in the scene, limiting the enhancement of features. To resolve this issue, we further introduce two dynamic operations to flexibly capture surrounding neighborhood information. Specifically, in response to the different directions between objects, we group voxels using the X-coordinate at the i-th layer while using the Y-coordinate at the $i+1$-th layer. This dynamic grouping direction can learn information from different directions without being limited to a specific direction. In order to cope with objects of different sizes, we use different numbers of groups in each layer to dynamically change the range of groups. If there are G voxels in a group at i-th layer, then there will be $2G$ voxels in a group at $i+1$ layer. This dynamic grouping size can learn information from different range neighborhoods without being limited to a specific size. Based on the semantic feature correlation, these two dynamic settings enable each voxel to adaptively select appropriate directions and sizes of other voxels for feature enhancement. Overall, different from the global dependence of vanilla attention, our proposed Group Attention performs attention in a dynamic local group, thereby enhancing the ability to capture the neighboring information of each foreground voxel more effectively.

D. Coupling-Decoupling Feature Interaction

In this module, we aim to interact the multi-frame features following the "coupling-decoupling" strategy. To explore multi-frame features more comprehensively and integrate them more deeply, we propose LCD and GCD for conducting feature coupling-decoupling at the local and global levels, respectively.

1) Local Coupling Decoupling: After the FFE mentioned above, we already have multi-frame foreground features that provide us with sufficient local information. In this section, based on these enhanced foreground features, we aim to perform a coupling-decoupling temporal fusion, thereby providing feedback correction for the foreground features. The detailed architecture is shown in Figure 5. Meanwhile, since the attention mechanism is only dependent on input and has strong feature interaction capabilities, we use attention to implement our coupling-decoupling approach.

We first use a learnable foreground embedding F_{emb} to couple with the enhanced multi-frame foreground features $\{F''_n\}^t_{n=t-N}$. The foreground embedding has the same shape as each frame foreground feature, *i.e.* $M \times C_2$. We also concatenate the multi-frame foreground features $\{F''_n\}^t_{n=t-N}$ to obtain the temporal foreground features $F'' \in \mathbb{R}^{NM \times C_2}$. We perform cross-attention between F_{emb} and F'' by transforming F_{emb} to query and transforming F'' to key and value, coupling the temporal foreground information into $\hat{F}_{emb}$.

$$\hat{F}_{emb} = \text{CrossAttn}(F_{emb}, F'', F'') \tag{5}$$

After aggregating the temporal foreground information in the embedding, we would like to decouple the temporal information into each frame for feature feedback. However, the group attention has performed information interaction for the foreground features. Meanwhile, only decoupling these individual features back to each frame feature may lead to information inconsistency in the local area. Considering the above two points, we incorporate the neighboring regions of each foreground feature during the decoupling process. Since the regional feature and central foreground feature are usually similar due to the local consistency, and aggregated temporal features are also learned based on central features, we decouple the foreground information into local regions by using central foreground features as a bridge. The local decoupling not only expands the fusion score but also maintains the feature consistency. We will show the effectiveness of the region-level decoupling in the experimental comparison.

As shown in Figure 5, inspired by the attention mechanism [20], we propose a similarity-based feature decoupling method. We first define a local region R centered on each foreground of interest and of size $l \times l$ in each frame, thus the size of local region features in each frame is $M \times l \times l \times C_2$. We then reshape it to $M \times L \times C$ for the following feature decoupling, where $L = l^2$. Next, for each local region, we compute the similarity between the centered foreground feature and the foreground embedding, which is also normalized in (0,

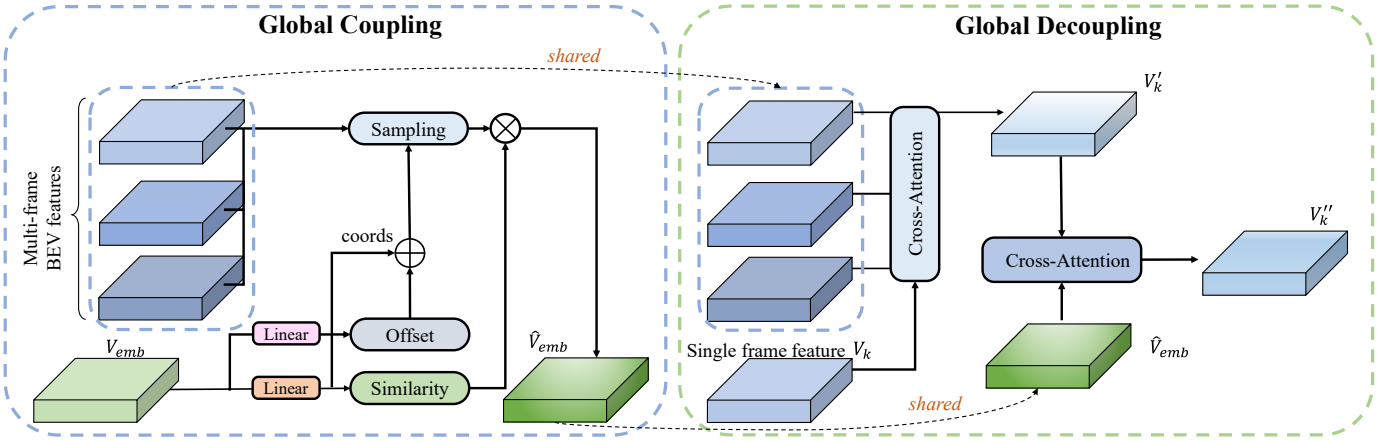

Fig. 6. Illustration of our GCD module. Similar to LCD, we first couple the global temporal information into the scene embedding. Then, to fully explore the temporal information, we design a two-stage decoupling process. Specifically, each frame feature is first directly enhanced by interacting with multiple frame features, forming explicit temporal feedback. Afterwards, the enhanced feature is fused with the temporal features from the coupling process for implicit temporal feedback. Here we only show decoupling a single frame for clarity.

1) by softmax. Furthermore, we multiply the similarity with the local region features to decouple the temporal contextual information to the neighboring region. The process could be formulated as follows:

$$A = \text{Softmax}\left(\frac{\hat{F}_{emb} F''_n}{\sqrt{d}}\right) \tag{6}$$

$$W'_n = A \cdot W_n \tag{7}$$

where W_n is the local region feature of each frame and d is the number of channels. Finally, the local region feature is reshaped back to $M \times l \times l \times C$ and is scattered back to the BEV feature. By decoupling within the local region, the aggregated temporal foreground information in F_{emb} could be feedback to eliminate the noise and augment the features of potential object regions.

2) Global Coupling Decoupling: After conducting multi-frame features interaction in the local scope, we further explore fusing scene information of multi-frame features in the global scope. As shown in Figure 6, similar to the local feature interaction, we also design a learnable embedding $V_{emb} \in \mathbb{R}^{H \times W \times C_2}$, named temporal scene embedding, to couple and decouple scene information.

Similar to the LCD, we also couple the multi-frame scene information through the attention mechanism. However, different from the LCD which only attends to the foreground features, GCD needs to process the overall BEV scene features. Therefore, directly applying vanilla attention [20] to the multi-frame BEV features would introduce a huge computational load. To avoid this problem, we utilize the deformable attention [9] to couple the temporal scene embedding with the multi-frame BEV features by limiting the receptive field of Value feature V. Compared to vanilla attention, for each query point, deformable attention generates a fixed number of similarity scores by applying a linear layer on the Q. Meanwhile, by predicting the same number of sampling offsets with another linear layer, the deformable attention sample value feature from the original value feature. Finally, by multiplying the predicted weight A with the sampled feature V, deformable attention outputs the fusion feature. By learning the weights of the attention linear layer and offset linear layer, the final fusion output not only reflects the relationship between the two features but also avoids too much computation cost. The process could be formulated as:

$$A = \text{Softmax}\left(\text{Linear}\left(V_{emb}\right)\right), \ \Delta u = \text{Linear}(V_{emb}) \tag{8}$$

$$V'_n = S(V_n, u + \Delta u) \tag{9}$$

$$\hat{V}_{emb} = \sum_{h=1}^{H} W_h \left(\sum_{n=1}^{NJ} A_k \cdot W'_h V'_{nk} \right) \tag{10}$$

where $S(\cdot)$ represents bilinear sampling, $W_h \in \mathbb{R}^{C_h \times C_2}$ and $W'_h \in \mathbb{R}^{C_2 \times C_h}$. C_h is the channels of each head, H is the number of heads. Although the attention weight A here is not directly calculated by multiplying Q and K, it is predicted by using a learnable linear layer on V_{emb}. Meanwhile, V_n is also sampled based on the offset predicted from V_{emb}, and the final output is to multiply the sampled V_n with the weight A. Therefore, by learning the weights of the attention linear layer and offset linear layer, the final output also reflect the relationship between the two features.

After coupling the temporal scene information into the temporal scene embedding, we would like to allocate the temporal scene-level information in V_{emb} back to each frame BEV feature, completing the global feature decoupling. Although we could directly fuse the learnable scene embedding with each frame BEV feature for feature decoupling, this approach cannot fully explore the temporal information because each single frame BEV feature only contains spatial information but not temporal information. For this purpose, we first need to embed the temporal information in individual frame features to enhance their temporal awareness. Similar to the FFE, we first use cross-attention to enhance each frame feature by explicitly interacting with all frames. By directly interacting with multi-frame features, we can effectively integrate the spatial-temporal information from multiple frames into individual single-frame features and make single-frame feature temporal-aware, preparing for the fusion with temporal embedding.

TABLE I
PERFORMANCE COMPARISON WITH OTHER METHODS ON THE NUSCENES *test* SET.

Method	NDS	mAP	Car	Truck	Bus	Trailer	C.V	Ped.	Motor.	Bicycle	T.C	Barrier
PointPillars [1]	45.3	30.5	68.4	23.0	28.2	23.4	4.1	59.7	27.4	1.1	30.8	38.9
WYSIWYG [30]	41.9	35.0	79.1	30.4	46.6	40.1	7.1	65.0	18.2	0.1	28.8	34.7
3DSSD [31]	56.4	42.6	81.2	47.2	61.4	30.5	12.6	7.2	36.0	8.6	31.1	47.9
SA-Det3D [32]	59.2	47.0	81.2	43.8	57.2	47.8	11.3	73.3	32.1	7.9	60.6	55.3
SSN V2 [33]	61.6	50.6	82.4	41.8	46.1	48.0	17.5	75.6	48.9	24.6	60.1	61.2
CBGS [34]	63.3	52.8	81.1	48.5	54.9	42.9	10.5	80.1	51.5	22.3	70.9	65.7
CVCNet [35]	64.2	55.8	82.7	46.1	45.8	46.7	20.7	81.0	61.3	34.3	69.7	69.9
HotSpotNet [36]	66.6	59.3	83.1	50.9	56.4	53.3	23.0	81.3	63.5	36.6	73.0	71.6
CyliNet [37]	66.1	58.5	85.0	50.2	56.9	52.6	19.1	84.3	58.6	29.8	79.1	69.0
CenterPoint [14]	67.3	60.3	85.2	53.5	63.6	56.0	20.0	84.6	59.5	30.7	78.4	71.1
AFDetV2 [38]	68.5	62.4	86.3	54.2	62.5	58.9	26.7	85.8	63.8	34.3	80.1	71.0
S2M2-SSD [39]	69.3	62.9	86.3	56.0	65.4	59.8	26.2	84.5	61.6	36.4	77.7	75.1
TransFusion-L [29]	70.2	65.5	86.2	56.7	66.3	58.8	28.2	86.1	68.3	44.2	82.0	**78.2**
VISTA [40]	70.4	63.7	84.7	54.2	64.0	55.0	29.1	83.6	71.0	45.2	78.6	71.8
Focals Conv [18]	70.0	63.8	86.7	56.3	67.7	59.5	23.8	87.5	64.5	36.3	81.4	74.1
VoxelNeXt [41]	70.0	64.5	84.6	53.0	64.7	55.8	28.7	85.8	73.2	45.7	79.0	74.6
LargeKernel3D [19]	70.6	65.3	85.5	53.8	64.4	59.5	29.7	85.9	72.7	46.8	79.9	75.5
Link [42]	71.0	66.3	86.1	55.7	65.7	62.1	30.9	85.8	73.5	47.5	80.4	75.5
FocalFormer3D [43]	72.6	68.7	87.2	57.1	**69.6**	**64.9**	34.4	88.2	76.2	49.6	82.3	77.8
FSTR [44]	71.5	67.2	86.5	54.1	66.4	58.4	33.4	88.6	73.7	48.1	84.4	78.1
HEDNet [45]	72.0	67.7	87.1	56.5	70.4	63.5	33.6	87.9	70.4	44.8	85.1	78.1
DSVT [46]	72.7	68.4	86.8	**58.4**	67.3	63.1	37.1	88.0	73.0	47.2	84.9	78.4
SAFDNet [47]	72.3	68.3	87.3	57.3	68.0	63.7	37.3	89.0	71.1	44.8	84.9	**79.5**
Voxel Mamba [48]	73.0	69.0	86.8	57.1	68.0	63.2	35.4	89.5	74.7	50.8	**86.9**	77.3
3DVID [6]	71.4	65.4	87.5	56.9	63.5	60.2	32.1	82.1	74.6	45.9	78.8	69.3
TCTR [26]	-	50.5	83.2	51.5	63.7	33.0	15.6	74.9	54.0	22.6	52.5	53.8
TransPillar [27]	-	52.3	84.0	52.4	62.0	34.3	18.9	77.9	55.2	27.6	55.4	55.1
MGTANet [8]	71.2	65.4	87.7	56.9	64.6	59.0	28.5	86.4	72.7	47.9	83.8	65.9
SUIT [49]	68.9	62.8	85.9	53.7	59.0	54.9	26.4	85.6	68.9	42.9	79.7	71.6
TFNet-Dense	72.8	68.8	87.6	57.4	66.4	64.2	34.1	87.0	77.8	**55.0**	82.2	76.2
TFNet-Sparse	**73.3**	**69.8**	**89.2**	58.0	65.0	64.0	**38.6**	**89.7**	**80.2**	52.0	86.2	75.3

Subsequently, we employ the deformable attention [9] to perform cross-attention between the enriched temporal-aware single-frame feature V_i and the temporal scene embedding V_{emb}. Given that each frame feature has gained temporal awareness by interacting with all frame features, linking it with the temporal scene embedding enables comprehensive exploration of aggregated temporal information, resulting in better feature decoupling. The decoupling process for $k-$th frame could be formulated as follows:

$$V_k' = \text{CrossAttn}(V_k, \{V_i\}_{i=t-N}^{t}, \{V_i\}_{i=t-N}^{t}), \quad k \in \{t-N, ..., t\} \tag{11}$$

$$V_k'' = \text{CrossAttn}(V_k', \hat{V}_{emb}, \hat{V}_{emb}), k \in \{t-N, ..., t\} \tag{12}$$

Finally, we use the decoupled current frame feature V_t'' to detect objects by the detection head.

E. Loss

Our detection losses are the same as the single-frame detector [14], [29]. Additionally, we use CrossEntropy loss in the FFE as follows:

$$L_{seg} = \frac{1}{N} \sum_{i} -[\hat{y}_i \log(y_i) + (1-\hat{y}_i) \log(1-y_i)] \tag{13}$$

Where $\hat{y}_i$ is the label and y_i is prediction. The total loss is:

$$L = L_{det} + L_{seg} \tag{14}$$

Where L_{det} represents the detection head loss of the detection method we build on.

IV. EXPERIMENTS

In this section, we compare our proposed method with the state-of-the-art methods on nuScenes datasets. Then, we conduct extensive ablation studies to demonstrate the effectiveness of our proposed modules of TFNet.

A. Experimental Setup

Dataset. Most of the 3D temporal detection methods [8], [27], [49], [51] uses the nuScenes dataset for comparison. Following these previous works, we also adopt this dataset in our comparison. nuScenes [52] is a large autonomous driving dataset for 3D object detection in urban scenes. The dataset contains 1000 driving sequences, and is officially split into 700, 150, and 150 scenes for training, validation, and testing respectively. The LiDAR data is collected at 20 Hz while the annotations are provided at 2 Hz. Moreover, we also conduct a comparison on the Waymo dataset. Compared to nuScene dataset, Waymo consists of more scenes and a larger LiDAR range. This dataset consists of 798 scenes for training and 202 scenes for validation. Each scene includes 200 frames at a frame of 10 HZ, with labeled 3D bounding boxes for vehicles, pedestrians, cyclists and signs.

Metric. For nuScene dataset, we use mean average precision (mAP) and nuScenes detection score (NDS) [52] as the evaluation metric. The mAP is defined by the BEV center distance instead of the 3D IoU, which is computed over the distance thresholds of 0.5 m, 1 m, 2 m, 4 m across 10 categories. NDS is a weighted sum of mAP and other true

TABLE II
PERFORMANCE COMPARISON WITH OTHER METHODS ON THE NUSCENES *valid* SET.

Method	NDS	mAP	Car	Truck	Bus	Trailer	C.V	Ped.	Motor.	Bicycle	T.C	Barrier
SECOND [3]	62.3	50.6	81.8	51.7	66.9	37.3	15.0	77.7	42.5	17.5	57.4	59.2
CBGS [34]	56.3	56.3	82.9	52.9	64.7	37.5	18.3	80.3	60.1	39.4	64.8	64.3
CVCNet [35]	65.5	54.6	83.2	50.0	62.0	34.5	20.2	81.2	54.4	33.9	61.1	65.5
HotSpotNet [36]	66.0	59.5	84.0	56.2	67.4	38.0	20.7	82.6	66.2	49.7	65.8	64.3
CenterPoint [14]	66.8	59.6	85.5	58.6	71.5	37.3	17.1	85.1	58.9	43.4	69.7	68.5
VoxelNeXt [41]	67.1	60.0	85.6	58.4	71.6	38.6	17.9	85.4	59.7	43.4	70.8	68.1
Focals Conv [18]	68.1	61.2	86.6	60.2	72.3	40.8	20.1	86.2	61.3	45.6	70.2	69.3
LargeKernel3D [19]	69.1	63.9	85.1	60.1	72.6	41.4	24.3	85.6	70.8	59.2	72.3	67.7
TransFusion-L* [29]	70.0	65.6	87.5	61.8	73.7	43.1	25.1	88.0	74.6	58.1	75.1	69.7
HEDNet [45]	71.4	66.7	87.7	60.6	77.8	50.7	**28.9**	87.1	74.3	56.8	76.3	66.9
DSVT [46]	71.1	66.4	87.4	62.6	75.9	42.1	25.3	88.2	74.8	58.7	77.8	70.9
Voxel Mamba [48]	**71.9**	67.5	87.9	**62.8**	76.8	45.9	24.9	**89.3**	77.1	58.6	**80.1**	71.5
ScatterFormer [50]	70.5	66.9	87.1	60.4	77.8	**48.7**	**28.9**	87.7	76.2	59.8	76.1	66.5
SAFDNet [47]	71.0	66.3	87.6	60.8	**78.0**	43.5	26.6	87.8	75.5	58.0	75.0	69.7
MGTANet [8]	70.6	64.8	87.9	61.3	73.2	40.2	23.0	86.6	74.1	59.9	75.7	66.1
TFNet-Dense	70.9	65.7	88.0	62.2	71.8	43.7	23.4	87.5	75.2	**64.0**	73.7	67.6
TFNet-Sparse	**71.9**	**67.6**	**89.5**	62.5	75.4	46.1	26.6	**89.3**	**78.0**	58.8	**77.2**	**72.7**

positive metrics, including translation, orientation, and other attributes. For the Waymo dataset, the official metrics are mean Average Precision (mAP) and mean Average Precision with Heading (mAPH). The two metrics are computed with 3D IoU threshold of 0.7 for vehicles.

B. Implementation Details

Data Process. Following previous works [6], [8], we also use 10 sweeps to aggregate the keyframe and use 3 keyframes as the network input. Meanwhile, for each LiDAR point cloud, we use the (x, y, z, r, t) as input, where (x, y, z) represent the 3D coordinate, r and t represents the reflectance and keyframe time respectively.

Network Architecture. Our TFNet has two versions: (a) TFNet-Dense: is built upon the dense detector CenterPoint [14]. (b) TFNet-Sparse: is built upon the sparse detector TransFusion-L [29]. We use an input point cloud range of [-54.0m, 54.0m] for the X-Y axes and [-5m, 3m] for the Z-axis. The voxel size is set to [0.075m, 0.075m, 0.2m] thus leading to $1440 \times 1440 \times 40$ voxel grid. In feature extraction, we also use sparse convolution [3] to extract sparse 3D voxel features and then convert them to 2D BEV features. The downsample rate is 8 and the final basic feature V_n is $180 \times 180 \times 512$. Before the proposed modules, we first use a convolution layer to reduce the channel number from 512 to 256. In the FFE, we sample 256 foreground voxels in each frame. The basic number for each group in the Group Attention is set as 8. All attention modules consist of 6 layers. In the LCD, we set the local region size as 5×5.

Training. Following [14], [29], we use Adam optimizer with one-cycle learning rate policy, the max learning rate is set as 1×10^{-3}, the weight decay is 0.01 and momentum ranges from 0.85 to 0.95. Our model is trained for 20 epochs with batch size of 16 using 8 NVIDIA V100 GPUs. We also apply data augmentations in the training stage, including global rotation, global scaling, random flipping and GT sampling. The rotation angle is [-$\frac{\pi}{4}$, $\frac{\pi}{4}$] along the z-axis. The scale factor is within [0.95, 1.05]. All the above augmentations are added to multi-frame point clouds in the same way. We also follow CBGS [34] to perform class-balanced sampling.

TABLE III
PERFORMANCE ON THE WAYMON *validation* SET FOR CLASS VEHICLE.

Methods	Frames	L2 AP	L2 APH
PointPillars [1]	1	55.18	54.69
SECOND [3]	1	63.90	63.30
CenterPoint [14]	1	-	66.20
CenterPoint [14]	2	-	67.30
3D-MAN [7]	16	67.61	67.14
TFNet-Dense	5	**71.11**	**70.66**

C. Comparison to the state-of-the-art

As shown in Table I, our proposed TFNet achieves new state-of-the-art detection performance. Compared to the two single-frame baseline methods, TransFusion [29] and CenterPoint [14], our methods improve the performance by 3.1 points and 5.5 points on NDS respectively. Meanwhile, compared to our baseline methods, some recent single-frame methods [44]–[47] also have more advanced point cloud feature extraction networks, resulting in better performance. However, through the proposed "coupling-decoupling" algorithm, we effectively fused multiple frames of information and outperforms these recent methods. Moreover, compared to the previous top-ranked temporal method MGTANet [8], our two versions still carry out superior results with 2.1 points and 1.6 points improvement on NDS respectively. It is worth noting that our method uses the exact same data settings as MGTANet [8], and our TFNet-Dense, which uses the same baseline network as MGTANet, also performs better than MGTANet. The result verifies the effectiveness of our proposed "coupling-decoupling" temporal fusion strategy.

We also report our performance in the nuScenes valid set, as shown in Table II. The results are consistent with our results on the test set, showing that our proposed method could achieve better performance. It is obvious that our method improves the overall accuracy compared to previous single-frame

Fig. 7. Visualization of the prediction results. The boxes represent the ground-truth boxes, the boxes, boxes and the boxes represent the predictions of ours, TransFusion-L [29] and CenterPoint [14], respectively.

methods. Compared to the baseline method TransFusion [29] and CenterPoint [14], our method improves the performance of NDS/mAP by 1.9/2.0 points and 4.1/6.1 points respectively. Moreover, compared to previous temporal methods, our proposed method also achieves higher detection results. The two results verify the effectiveness of our proposed temporal feedback mechanism in temporal fusion. We also visualize some prediction results in Figure 7.

In Waymo dataset, to have a fair comparison, we use the TFNet-Dense in the experiment. As Table III shows, our method shows better performance. Compared to the baseline model CenterPoint, our method improves the L2 APH by 3.36. Compared to 3D-MAN, we improve the L2 AP/APH by 3.50/3.52, verifying the effectiveness of the proposed method.

TABLE IV
ABLATION STUDY OF EACH COMPONENT ON NUSCENES *valid* SET.

Methods	Proposed Modules			Performance	
	FFE	LCD	GCD	NDS	mAP
Baseline				66.34	60.44
Ours	✓			67.18↑0.84	61.53↑1.09
	✓	✓		67.67↑1.33	62.45↑2.01
			✓	67.84↑1.50	62.84↑2.40
	✓		✓	68.68↑2.34	64.15↑3.71
	✓	✓	✓	69.50↑3.16	64.70↑4.26

D. Ablation Study

We conduct ablation studies on the nuScenes valid set to explore the effect of each proposed module. Following previous works [8], we also only use 1/8 of the training set

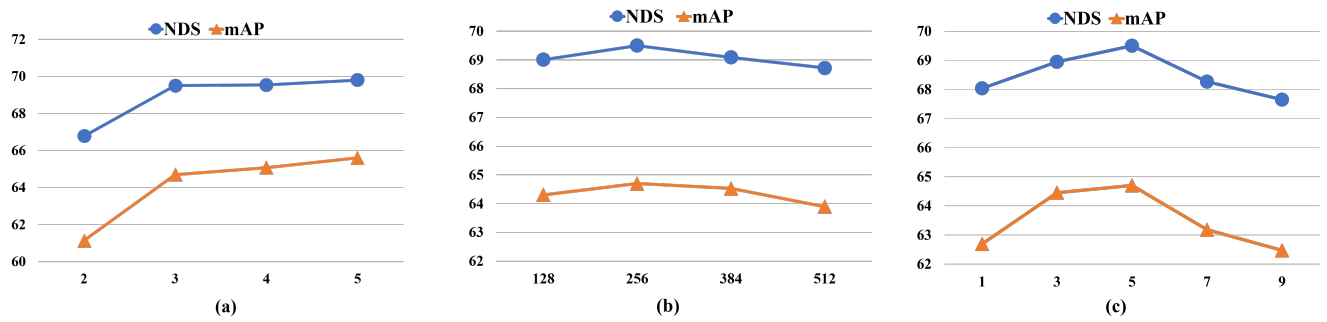

Fig. 8. (a) The comparisons of different frames. (b) The comparisons of different sampling numbers in FFE. (c) The comparisons of different region sizes in local decoupling.

in training and all valid set in the evaluation to reduce the required time of experiments. All ablation studies are built on Transfusion-L [29].

Effect of proposed three modules. We first conduct ablation studies on our proposed modules, as shown in Table IV. Compared to the baseline model which simply concatenates multi-frame features, the proposed FFE improves the performance by 0.84 points and 1.09 points in NDS and mAP. We further add LCD to this model, leading to improvements of 0.49 and 0.92 points in NDS and mAP. Based on the baseline model, we also only perform GCD and improve the results by 1.50 and 2.40 points. Compared to the above two modules, the overall BEV feature has a larger perceiving range and more interacted features, resulting in higher performance. Additionally, with the FFE, the GCD could better enhance each frame feature, further improving the NDS and mAP with 0.84 and 1.31 points. Finally, by equipping all proposed modules, we achieve the best results. Compared to the baseline model, we improved the NDS and mAP by 3.16 and 4.26 points. We also compare different fused frames, shown in Figure 8 (a). With the number of frames increasing, the performance also gets improved.

Comparison in FFE. As shown in Table V, we explore different components in the proposed FFE, including the sampling strategies and attention types. We first compare top-k sampling, setting threshold sampling, FPS and our top-FPS in the foreground sampling. For threshold sampling, we set the threshold as 0.5 and use repeatedly random sampling. Compared to top-k sampling, FPS provides more balanced sampling outcomes, leading to improved performance. Furthermore, by employing a top-k approach as an initial filter, our top-FPS can obtain more accurate and balanced results, achieving the best performance, verifying our hypothesis in Section III-C. We also compare different sampling numbers. As shown in Figure 8 (b), sampling 256 foreground voxels achieves the best performance. The results are consistent with the average number of foreground voxels in each point cloud scene in the nuScenes dataset. When sampling more voxels, too much background and noise will be introduced, resulting in performance degradation. Finally, we compare different attention methods with our proposed group attention in the foreground enhancement, including vanilla attention [20], trajectory attention [53] and KNN Attention [54]. The trajectory attention [53] is proposed in the video transformer to model the temporal correspondences. We follow the default setting in Point Transformer [54] as set $K = 16$ in the KNN Attention. As we discussed in Section III-C, the global dependency in vanilla attention introduces noise between different objects. Meanwhile, we also notice that the trajectory attention achieves lower performance than the other methods. We believe that the sparsity of points limits the temporal modeling of trajectory attention, which results in lower results. KNN Attention [54] also implements local attention by projecting neighboring points as key and value features, similar to ours. Nevertheless, our Group Attention has the dynamic setting to enlarge the receptive field, resulting in a better performance.

Comparison in Group Attention. We also conduct ablation studies on the proposed Group Attention. As shown in Ta-

TABLE V
ABLATION STUDY ON THE PROPOSED FFE.

FFE Modules	Methods	Performance	
		NDS	mAP
Sampling Strategies	Top-k	68.17	63.25
	Threshold	68.95	64.33
	FPS	69.09	64.53
	Top-FPS	**69.50**	**64.70**
Attention Types	Attention	68.98	63.80
	Traj-Attention	68.51	63.39
	KNN Attention	69.43	64.62
	Group Attention	**69.50**	**64.70**

TABLE VI
ABLATION STUDY ON PROPOSED GROUP ATTENTION.

Group Attention Modules	Methods	Performance	
		NDS	mAP
X-Y Alternation	✗	69.02	63.86
	✓	**69.50**	**64.70**
Double Number	✗	69.28	**64.73**
	✓	**69.50**	64.70
Basic Number	4	69.24	64.66
	8	**69.50**	**64.70**
	16	69.08	64.57

TABLE VII
COUPLING AND DECOUPLING ON PROPOSED LCD.

LCD	Performance	
	NDS	mAP
w/o LCD	68.68	64.15
+Coupling	68.37↓0.31	63.14↓1.01
+Coupling +Decoupling	69.50↑0.82	64.70↑0.55

TABLE VIII
COUPLING AND DECOUPLING ON PROPOSED GCD.

GCD	Performance	
	NDS	mAP
w/o GCD	67.67	62.45
+Coupling	68.57↑0.90	63.87↑1.42
+Coupling +Decoupling	69.50↑1.83	64.70↑2.25

TABLE IX
DECOUPLING ON PROPOSED LCD.

Methods	Performance	
	NDS	mAP
w/o Decoupling	68.37	63.14
Vanilla Attention	68.12↓0.25	63.24↑0.10
Depth-wise Correlation	68.77↑0.40	64.12↑0.98
Proposed Method	69.50↑1.13	64.70↑1.56

ble VI, removing the X-Y alternation sorting leads to a 0.48 and 0.84 points performance degradation in NDS and mAP. Meanwhile, removing the double-group setting between layers results in NDS decrease by 0.22 points but mAP increase by 0.03 points. The results prove that the two dynamic operations can provide sufficient neighborhood information to enhance local representation, which further boosts the performance of 3D object detection. Moreover, we compare different basic numbers of points within a group and find that setting $G = 8$ could achieve the best performance. We believe that too few points in a group result in insufficient information aggregations, while too many points may introduce noise.

Effects of Coupling and Decoupling. Here we explore the proposed "coupling-decoupling" temporal interaction strategy in both LCD and GCD. As shown in Table VII, in the local scope, only using feature coupling leads to a performance drop of 0.31 and 1.01 points in NDS and mAP. As mentioned in Group Attention, different regions typically correspond to different objects. Therefore, directly coupling these multi-frame foreground features into the F_{emb} leads to incorrect information interaction. Meanwhile, scattering back the temporal foreground feature to the frame feature also leads to information inconsistency in a local area. Based on the coupling, we add local decoupling and achieve the best performance. Compared to only performing feature coupling, a complete coupling-decoupling brings in an improvement of 1.13 and 1.56 points in NDS and mAP respectively. The result highlights the importance of decoupling at the local scope, which can effectively feedback temporal information and eliminate the noise caused by coupling. Meanwhile, compared to the baseline, the "coupling-decoupling" strategy brings in 0.82 and 0.55 improvements in NDS and mAP. We further conduct a similar ablation study in the global scope, as shown in Table VIII. Only coupling the multi-frame scene features could improve the performance by 0.90 and 1.42 points in NDS and mAP. Based on feature coupling, feature decoupling further improves the performance by 0.93 and 0.83 points in NDS and mAP, and brings in 1.83 and 2.25 points in NDS and mAP compared to the baseline. For temporal point cloud detection, most existing methods involve transferring past frames to the current frame and then fusing them through convolution or attention. This can be seen as the "coupling-only" strategy. However, as shown in Table VII and Table VIII, the "coupling-only" strategy achieves limited improvement. We believe that the "coupling-only" strategy overlooks temporal noise introduced during multi-frame fusion, where past frames may not always share consistent information with the current frame. In contrast, our proposed "coupling-decoupling" strategy introduces a temporal feedback mechanism through the decoupling process, which refines the aggregated features by feeding them back into the current frame, effectively mitigating temporal noise. Therefore, in both local and global feature scopes, the proposed "coupling-decoupling" strategy could further improve the detection performance.

Comparisons in Decoupling. We first analyze the decoupling methods in both LCD and GCD. As shown in Table IX, we compare vanilla attention, depth-wise correlation and our proposed similarity-based method in the decoupling of LCD. Specifically, in attention implementation, the query comes from the region features while the key and value come from the learnable foreground embedding F_{emb}. In correlation implementation, the learnable foreground embedding works as the correlation kernel. Compared to the two methods, our approach demonstrates superior performance, showing improvements of 1.38/1.46 points and 0.73/0.58 points in NDS/mAP, respectively. These results validate the importance of the centered foreground features, which serve as a bridge in fusing the region features and embedding features F_{emb} in our method. Moreover, we also compare different region sizes shown in Figure 8 (c). When the region size is small, the performance improves as the size increases until it reaches its peak at 5. Subsequently, the performance starts to decline with further increases in size. We believe that too small size limits the foreground feature decoupling, and too large size may introduce too many irrelevant regions and noise. Additionally, we compare different components in the decoupling of GCD. We divide the decoupling part into two parts: feature enhancement and temporal fusion, which perform the cross-attention fusion by using multi-frame features and temporal embedding respectively. As shown in Table X, the enhancement and temporal fusion bring in 0.15/0.70 points and 0.62/0.77 points improvements respectively. Finally, using both of them achieves the best performance. The results verify the effectiveness of the two parts in the decoupling of GCD.

Effect of Voxel Size. Finally, we explore the effectiveness of voxel size in our method, shown in Table XI. With the

TABLE X
DECOUPLING ON PROPOSED GCD.

Methods	Performance	
	NDS	mAP
w/o Decoupling	68.57	63.87
+ Enhance	68.72↑0.15	64.57↑0.70
+ Temporal	69.19↑0.62	64.64↑0.77
+ Enhance + Temporal	69.50↑0.93	64.70↑0.83

TABLE XI
ABLATION STUDY ON VOXEL SIZE.

	Voxel Size	Performance	
		NDS	mAP
V1	(0.15, 0.15, 0.2)	64.71	57.78
V2	(0.125, 0.125, 0.2)	66.32	60.23
V3	(0.075, 0.075, 0.2)	**69.50**	**64.70**

voxel size getting smaller, the performance gets better. We believe that a larger voxel size would lose detailed geometry information. The results are also consistent with previous detection methods [3], [29].

V. CONCLUSION

In this paper, we propose TFNet, a temporal feedback network for 3D temporal detection. Different from previous works, TFNet could feedback the temporal feature to eliminate the noise by the proposed "coupling-decoupling" strategy. The Foreground Feature Enhancement (FFE) module and the Coupling-Decoupling Feature Interaction (CDFI) module are proposed, which are designed to enhance the multi-frame foreground information and fuse the multi-frame features from local to global respectively. The experimental results on the nuScenes dataset show that our method achieves a new state-of-the-art performance, and the ablation study also verifies the effectiveness of the proposed temporal feedback mechanism. In the future, we would like to integrate the image information into temporal 3D detection.

REFERENCES

[1] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12697–12705, 2019.

[2] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 770–779, 2019.

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

[4] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10529–10538, 2020.

[5] R. Huang, W. Zhang, A. Kundu, C. Pantofaru, D. A. Ross, T. Funkhouser, and A. Fathi, "An lstm approach to temporal 3d object detection in lidar point clouds," in *European Conference on Computer Vision*, pp. 266–282, Springer, 2020.

[6] J. Yin, J. Shen, X. Gao, D. Crandall, and R. Yang, "Graph neural network and spatiotemporal transformer attention for 3d video object detection from point clouds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[7] Z. Yang, Y. Zhou, Z. Chen, and J. Ngiam, "3d-man: 3d multi-frame attention network for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1863–1872, 2021.

[8] J. Koh, J. Lee, Y. Lee, J. Kim, and J. W. Choi, "Mgtanet: Encoding sequential lidar points using long short-term motion-guided temporal attention for 3d object detection," in *2023 Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 37, pp. 1179–1187, 2023.

[9] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[10] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 918–927, 2018.

[11] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Std: Sparse-to-dense 3d object detector for point cloud," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1951–1960, 2019.

[12] T. Xie, L. Wang, K. Wang, R. Li, X. Zhang, H. Zhang, L. Yang, H. Liu, and J. Li, "Farp-net: Local-global feature aggregation and relation-aware proposals for 3d object detection," *IEEE Transactions on Multimedia*, pp. 1–15, 2023.

[13] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4490–4499, 2018.

[14] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11784–11793, 2021.

[15] P. An, Y. Duan, Y. Huang, J. Ma, Y. Chen, L. Wang, Y. Yang, and Q. Liu, "Sp-det: Leveraging saliency prediction for voxel-based 3d object detection in sparse point cloud," *IEEE Transactions on Multimedia*, pp. 1–14, 2023.

[16] H. Zhu, J. Deng, Y. Zhang, J. Ji, Q. Mao, H. Li, and Y. Zhang, "Vpfnet: Improving 3d object detection with virtual point based lidar and stereo data fusion," *IEEE Transactions on Multimedia*, vol. 25, pp. 5291–5304, 2023.

[17] Y. Zhang, Q. Hu, G. Xu, Y. Ma, J. Wan, and Y. Guo, "Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2022.

[18] Y. Chen, Y. Li, X. Zhang, J. Sun, and J. Jia, "Focal sparse convolutional networks for 3d object detection," IEEE, Jun 2022.

[19] Y. Chen, J. Liu, X. Zhang, X. Qi, and J. Jia, "Largekernel3d: Scaling up kernels in 3d sparse cnns," 2022.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] J. Mao, Y. Xue, M. Niu, H. Bai, J. Feng, X. Liang, H. Xu, and C. Xu, "Voxel transformer for 3d object detection," IEEE, Oct 2021.

[22] L. Fan, Z. Pang, T. Zhang, Y.-X. Wang, H. Zhao, F. Wang, N. Wang, and Z. Zhang, "Embracing single stride 3d object detector with sparse transformer," IEEE, Jun 2022.

[23] C. He, R. Li, S. Li, and L. Zhang, "Voxel set transformer: A set-to-set approach to 3d object detection from point clouds," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2022.

[24] Z. Zhou, X. Zhao, Y. Wang, P. Wang, and H. Foroosh, "Centerformer: Center-based transformer for 3d object detection," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pp. 496–513, Springer, 2022.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," IEEE, Oct 2021.

[26] Z. Yuan, X. Song, L. Bai, Z. Wang, and W. Ouyang, "Temporal-channel transformer for 3d lidar-based video object detection for autonomous driving," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2068–2078, 2022.

[27] Z. Luo, G. Zhang, C. Zhou, T. Liu, S. Lu, and L. Pan, "Transpillars: Coarse-to-fine aggregation for multi-frame 3d object detection," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan 2023.

[28] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[29] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with

transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1090–1099, 2022.
[30] P. Hu, J. Ziglar, D. Held, and D. Ramanan, "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*, pp. 11001–11009, 2020.
[31] Z. Yang, Y. Sun, S. Liu, and J. Jia, "3dssd: Point-based 3d single stage object detector," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11040–11048, 2020.
[32] P. Bhattacharyya, C. Huang, and K. Czarnecki, "Sa-det3d: Self-attention based context-aware 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3022–3031, 2021.
[33] X. Zhu, Y. Ma, T. Wang, Y. Xu, J. Shi, and D. Lin, "Ssn: Shape signature networks for multi-class object detection from point clouds," in *European Conference on Computer Vision*, pp. 581–597, Springer, 2020.
[34] B. Zhu, Z. Jiang, X. Zhou, Z. Li, and G. Yu, "Class-balanced grouping and sampling for point cloud 3d object detection," *arXiv preprint arXiv:1908.09492*, 2019.
[35] Q. Chen, L. Sun, E. Cheung, and A. L. Yuille, "Every view counts: Cross-view consistency in 3d object detection with hybrid-cylindrical-spherical voxelization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21224–21235, 2020.
[36] Q. Chen, L. Sun, Z. Wang, K. Jia, and A. Yuille, "Object as hotspots: An anchor-free 3d object detection approach via firing of hotspots," in *European conference on computer vision*, pp. 68–84, Springer, 2020.
[37] M. Rapoport-Lavie and D. Raviv, "It's all around you: Range-guided cylindrical network for 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2992–3001, 2021.
[38] Y. Hu, Z. Ding, R. Ge, W. Shao, L. Huang, K. Li, and Q. Liu, "Afdetv2: Rethinking the necessity of the second stage for object detection from point clouds," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 969–979, 2022.
[39] W. Zheng, M. Hong, L. Jiang, and C.-W. Fu, "Boosting 3d object detection by simulating multimodality on point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13638–13647, 2022.
[40] S. Deng, Z. Liang, L. Sun, and K. Jia, "Vista: Boosting 3d object detection via dual cross-view spatial attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8448–8457, 2022.
[41] Y. Chen, J. Liu, X. Zhang, X. Qi, and J. Jia, "Voxelnext: Fully sparse voxelnet for 3d object detection and tracking," 2023.
[42] T. Lu, X. Ding, H. Liu, G. Wu, and L. Wang, "Link: Linear kernel for lidar-based 3d perception," 2023.
[43] Y. Chen, Z. Yu, Y. Chen, S. Lan, A. Anandkumar, J. Jia, and J. M. Alvarez, "Focalformer3d: Focusing on hard instance for 3d object detection," pp. 8394–8405, 2023.
[44] D. Zhang, Z. Zheng, H. Niu, X. Wang, and X. Liu, "Fully sparse transformer 3-d detector for lidar point cloud," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.
[45] G. Zhang, C. Junnan, G. Gao, J. Li, and X. Hu, "HEDNet: A hierarchical encoder-decoder network for 3d object detection in point clouds," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.
[46] H. Wang, C. Shi, S. Shi, M. Lei, S. Wang, D. He, B. Schiele, and L. Wang, "Dsvt: Dynamic sparse voxel transformer with rotated sets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13520–13529, June 2023.
[47] G. Zhang, J. Chen, G. Gao, J. Li, S. Liu, and X. Hu, "Safdnet: A simple and effective network for fully sparse 3d object detection," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14477–14486, 2024.
[48] G. Zhang, L. Fan, C. He, Z. Lei, Z. Zhang, and L. Zhang, "Voxel mamba: Group-free state space models for point cloud based 3d object detection," *arXiv preprint arXiv:2406.10700*, 2024.
[49] Z. Zhou, J. Lu, Y. Zeng, H. Xu, and L. Zhang, "Suit: Learning significance-guided information for 3d temporal detection," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9399–9406, 2023.
[50] C. He, R. Li, G. Zhang, and L. Zhang, "Scatterformer: Efficient voxel transformer with scattered linear attention," in *Computer Vision – ECCV 2024*, pp. 74–92, Springer Nature Switzerland, 2025.
[51] J. Yin, J. Shen, C. Guan, D. Zhou, and R. Yang, "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*, pp. 11495–11504, 2020.
[52] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.
[53] M. Patrick, D. Campbell, Y. M. Asano, I. Misra, F. Metze, C. Feichtenhofer, A. Vedaldi, and J. F. Henriques, "Keeping your eye on the ball: Trajectory attention in video transformers," 2021.
[54] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point transformer," in *2021 ICCV*, IEEE, Oct. 2021.