2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

ProEqBEV: Product Group Equivariant BEV Network for 3D Object Detection in Road Scenes of Autonomous Driving

Hongwei Liu[1*], Jian Yang[2*], Zhengyu Li[1*], Ke Li[2], Jianzhang Zheng[3], Xihao Wang[3], Xuan Tang[1], Mingsong Chen[1], Xiong You[2†], Xian Wei[1†]

***Abstract*— With the rapid development of autonomous driving systems, 3D object detection based on Bird's Eye View (BEV) in road scenes has witnessed great progress over the past few years. As a road scene exhibits a part-whole hierarchy between the within objects and the scene itself, simple parts (e.g., roads, lane lines, vehicles and pedestrians) can be assembled into progressively more complex shapes to form a BEV representation of the whole road scene. Therefore, a BEV often has multiple levels of freedom on motion, i.e., the rotation and the moving shift of the whole BEV, and the random movements of objects (e.g., pedestrians and vehicles) inside the BEV. However, most of the current single-sensor or multi-sensor fusion-based BEV object detection methods have not yet taken into account capturing such multi-level motion in a BEV. To address this problem, we propose a product group equivariant object detection network framework that is equivariant with respect to multiple levels of symmetry groups based on multi-sensor fusion. The proposed framework extracts local equivariant features of objects in point clouds, while global equivariant features are extracted in both point clouds and images. Furthermore, the network learns diverse rotation-equivariant features and mitigates a significant amount of detection errors caused by rotations of BEV and objects inside a BEV, thereby further enhancing the performance of object detection. The experiment results show that the network architecture significantly improves object detection on mAP and NDS, respectively. In addition, in order to demonstrate the effectiveness of the proposed local-multi-global equivariant components, we conduct sufficient ablation experiments. The results show that the individual components are indispensable for the object detection performance improvement of the overall network architecture.**

I. INTRODUCTION

3D object detection [1] is an essential task performed by perception systems of autonomous driving vehicles, which have witnessed continuous advancements in recent years, particularly in terms of network architecture and detection precision [2], [3]. Notably, Bird's Eye View (BEV)-based 3D object detection has further achieved even better performance [4]. The perception system utilizes on-board vehicle sensors to capture information about the surrounding environment, and multi-sensors [5] can significantly improve the efficiency of environmental perception tasks compared to relying solely on a single sensor. Recently, Liu et al. [6] and Liang et al. [7] unify the multi-modal information into a consistent form for feature fusion. In their approach, the BEV features are generated from the point cloud encoded by Voxel Feature Encoding(VFE) [8], which effectively preserves the geometric information. Furthermore, they utilize the images and projecting them as BEV through explicit depth estimation [9] and then fuse them directly with the point cloud BEV representation, thereby retaining both geometric priors and dense semantic information. Consequently, multi-sensor network architectures based on BEV fusion have gained popularity.

However, in real-world road scenarios, the freedom of BEV representations has a significant impact on environmental perception, where the freedom of the whole BEV is influenced by factors like the camera and LiDAR, while the freedom of objects inside the BEV encompasses rotation and translation. Therefore, the BEV tends to have multiple levels of freedom on the motion, i.e., the rotation and translation of the whole BEV, and the random movements of the objects inside the BEV (e.g., pedestrians and vehicles). Based on this concept, the whole BEV and the BEV interior constitute a local-global hierarchy and exhibit an inherent compositional nature. Most of the current multi-sensor fusion-based BEV object detection methods have not yet considered capturing the local-global hierarchy of BEV, i.e., the multiple motion freedom levels in BEV.

To solve the above problem, in this paper, we construct a multi-sensor fusion-based product group equivariant network, which decouples the freedom within BEV and the global BEV representations and realizes the rotation-equivariance of the network itself. Specifically, we construct a network to realize the extraction of product group equivariant features based on the fusion of camera and LiDAR, namely, *Product group Equivariant BEV object detection Network (ProEqBEVNet)*. Specifically, we first perform rotation-equivariant processing on the local object of the point cloud to realize the local equivariant of the global scene, and design the voxel distance coding to further extract the local geometric features. Furthermore, we separately extract the global rotation-equivariant feature of the point cloud and image to construct the product-equivariant relationship. By constructing a local-multi-global equivariant BEV object detection network through multi-sensor fusion, we successfully extract rich rotation-equivariant features. This enhancement significantly improves the object detection performance of the network. The experiment results show that the performance of object detection is improved by

*Equal technical contribution
[1] East China Normal University, hwhongwei.liu@foxmail.com, zyli@stu.ecnu.edu.cn, xtang@cee.ecnu.edu.cn, mschen@sei.ecnu.edu.cn, xian.wei@tum.de
[2] School of Geospatial Information, Information Engineering University, jian.yang@tum.de, like19771223@163.com, youarexiong@163.com
[3] Technical University of Munich, xihao.wang@tum.de, zhengjianzhang18@mails.ucas.edu.cn.
† Corresponding Author

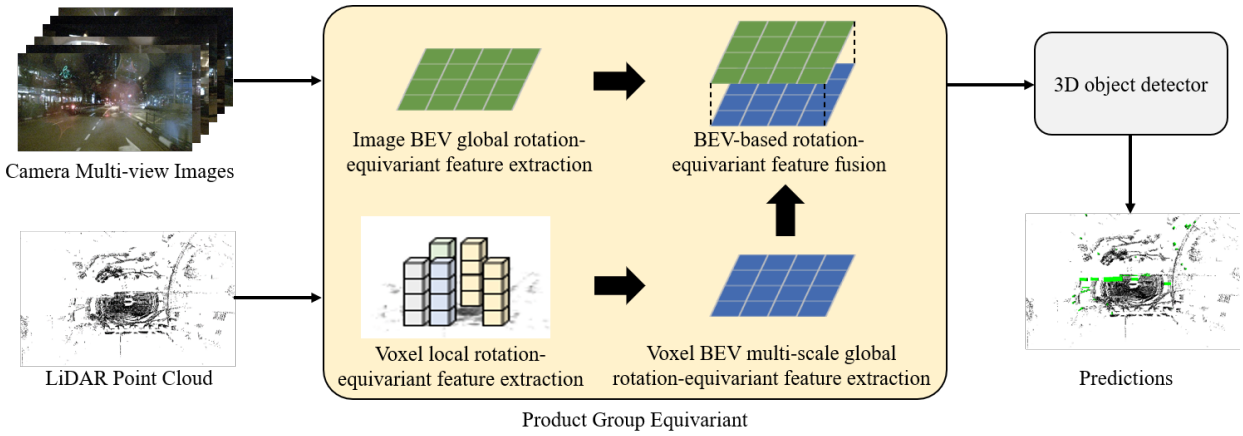


Fig. 1. The overall framework of ProEqBEVNet. The product group equivariance consists of the image BEV global rotation-equivariance, a voxel local rotation-equivariance, and a voxel BEV global rotation-equivariance.

capturing equivariant features in multiple motion freedom levels.

Moreover, our experimental results on the challenging autonomous driving public dataset nuScenes [10] demonstrate the effectiveness of ProEqBEVNet. It achieves an impressive average precision of 69.4% and a nuScenes detection score of 72.0. These results underscore the network's ability to maintain rotation-equivariance and its superior object detection performance compared to current mainstream object detection methods and architectures.

Our main contributions are listed below:

- Leveraging the inherent local-global equivariant hierarchy within the BEV representations of real road scenes and drawing upon equivariance theory, we introduce a novel product-equivariant network. This network not only achieves rotation-equivariance within its architecture but also comprehensively considers the multi-level motion relationships within BEV.
- We extract the local rotation-equivariant features of point cloud by encoding voxels and the voxel distance features among them, which helps to enhence the lcoal geometric features of surrounding objects in road scenes.
- In order to demonstrate the effectiveness of the proposed network architecture, we conduct comprehensive experiments on the nuScenes validation dataset. ProEqBEVNet achieves excellent performance compared to other mainstream network architectures.

II. RELATED WORK

A. BEV-based 3D Object Detection

The point cloud provided by LiDAR is a natural 3D space with accurate geometric information, and the accurate BEV can be obtained by utilizing the natural properties of the point cloud [11], [12]. However, the irregularity, sparsity and disorder of the raw point cloud render it difficult to be utilized directly. To facilitate the further processing of point cloud information, Zhou et al. [13] processed the raw point cloud into regular voxels, which can be directly used for 3D convolution and can be directly encoded into BEV. Unlike voxels, which are divided into three dimensions, Lang et al. [14] proposed PointPillars, which divides the point cloud into point pillars, which can be converted into BEV, and the method greatly improves the speed of the point cloud processing. However, it loses part of the information compared to the voxels.

Different from point clouds, 2D images can not provide 3D spatial information directly, which makes it difficult to generate accurate BEV. However, converting images to the BEV can greatly facilitate 3D object detection and can greatly alleviate the occlusion problem of object detection in the front view perspective [4]. The performance of object detection depends on the quality of BEV generation. Mallot et al. [15] proposed IPM, which directly uses internal and external parameters to convert BEV, provided all objects are on the plane. However, bumps can seriously affect the quality of BEV generation and cause distortion. LSS [9] avoided the weakness of IPM and generated BEV by depth prediction with powerful prior knowledge and obtained relatively accurate BEV. Huang et al. [16] proposed BEVDet, which follows the BEV generation paradigm of LSS and designs a camera-only BEV framework for 3D object detection. Li et al. [17] proposed BEVDepth, which introduces LiDAR's depth information to supervise the camera's depth prediction in the training stage, and the obtained depth information has higher accuracy. Different from BEVDepth, Huang et al. [18] introduced time as an information complement to connect the BEV of the previous moment with the current moment.

In addition, there are still other methods to generate the BEV without depth prediction, Lu et al. [19] utilized the MLP strategy to generate the BEV, which utilized a fixed matrix to transform the front view, which is only suitable for converting a single image. The BEV generated based on MLP strategy relies only on the fixed matrix and cannot rely on the input image for inference. In contrast, the following methods are different in that they utilize the input image for inference and are more suited for BEV generation. BEVFormer [20] considered the relationship between the

historical BEV and the current BEV using temporal self-attention and spatial self-attention to generate BEV features.

B. Group Equivariant Network

The convolution operation in the traditional Convolutional Neural Network (CNN) has translation-equivariance but not rotation-equivariance. Incorporating rotation-equivariance can improve the expressive ability of the model.

Group equivariant network is an equivariant network that uses group equivariant convolution to extract features with specific group properties. Cohen et al. [21] first proposed the group equivariant CNN, which lifted the traditional convolution to group with group properties. It was experimentally demonstrated that the group equivariant CNN has a higher degree of weight sharing and improves the network's performance without significantly increasing the number of network parameters.

Thomas and Smidt et al. [22] proposed Tensor Field Network (TFN) to realize the rotation and translation equivariance in 3D format. Esteves et al. [23] utilized the spherical function to model 3D data and introduced the equivariant convolutional network to process the spherical function output data. Compared to TFN, B. Fuchs et al. [24] introduced attention mechanism [25] instead of convolution to construct SE(3)-transformer equivariant attention network.

Despite the rapid progress in the study of group equivariant networks, their application in 3D object detection and object recognition remains relatively rare, and substantial challenges persist. Wang et al. [26] have taken a noteworthy step by incorporating equivariance as additional prior knowledge into networks during continuous learning [27]. This approach helps mitigate the sharp increase in task complexity that can result from point cloud augmentation. However, leveraging rotation-equivariance [28] to address the complexities of 3D object detection within multi-sensor fusion networks, especially when dealing with diverse data modalities, continues to pose a significant challenge.

III. The Proposed Method

A. Preliminary

Traditional CNN generally has only translational equivariant property, which means that it does not perform well in processing rotational orientation information. To realize the convolution operation to process rotational orientation information, Cohen et al. [21] introduced the concept of symmetric group $p4$. It realizes the rotation-equivariance of the convolution by lifting the input image to the symmetric group $p4$ to have the group property. The process can be represented by the following equation:

$$[f \star \psi](g) = \sum_{x \in \mathbb{Z}^2} \sum_{k} f_k(x) \psi_k \left(g^{-1} x\right) \tag{1}$$

where f is the image feature, $\star$ represents the convolution operation, ψ is the convolution kernel, which exists on the plane $\mathbb{Z}^2$. g represents the $p4$ transform, x is the image feature point, and k is the number of channels. The obtained image has specific properties of the symmetry group $p4$ and the convolution operation is performed on the symmetry group $p4$. This process can be defined as:

$$[f \star \psi](g) = \sum_{h \in G} \sum_{k} f_k(h) \psi_k \left(g^{-1} h\right) \tag{2}$$

where h is the image feature point on the group. Then, the output image features obtain both the translational equivariance and rotational equivariance.

Similarly, 3D data is subject to rotation transformations using the 3D rotation group $SO(3)$. This transformation is based on the work [22],[23],[24]. Hence, there are specific equivariant condition as follows[22]:

$$\mathcal{F} \circ \mathcal{P}_\sigma = \mathcal{P}_\sigma \circ \mathcal{F} \tag{3}$$

where $\mathcal{F}$ is a function, $\circ$ denotes the composite operation of the function, and $\mathcal{P}_\sigma\left(\vec{r}_a, x_a\right) := \left(\vec{r}_{\sigma(a)}, x_{\sigma(a)}\right)$, and $\sigma(a)$ is the indexed point.

Furthermore, the rotational equivariant condition is:

$$\mathcal{F} \circ \left[\mathcal{R}(g) \oplus D^{\mathcal{X}}(g)\right] = \left[\mathcal{R}(g) \oplus D^{\mathcal{Y}}(g)\right] \circ \mathcal{F} \tag{4}$$

where D^x represents the $SO(3)$ representation in the vector space X, and D^y is its counterpart in vector space Y. $\oplus$ denotes the concatenate, $\mathcal{R}(g)$ represents $g \in SO(3)$ acting on $\vec{r_a} \in \mathbb{R}^3$.

B. Product Group Equivariance

Taking into account the complexity of real road scenes, characterized by varying hierarchical equivariant relationships, each hierarchy requiring its unique equivariant feature extraction method, we have developed the ProEqBEVNet model. Its formal representation can be articulated as follows:

$$[(\mathcal{P}_{\mathcal{L}} \times \mathcal{P}_{\mathcal{G}}) \subseteq \mathcal{P}] \times (\mathcal{I}_{\mathcal{G}} \subseteq \mathcal{P})] \subseteq \mathcal{N}_{\mathcal{E}\mathrm{II}} \tag{5}$$

where $\mathcal{P}_{\mathcal{L}}$ denotes voxel local equivariance, $\times$ represents the product, $\mathcal{P}_{\mathcal{G}}$ signifies voxel global equivariance and is a subset of point cloud equivariance $\mathcal{P}$, while $\mathcal{I}_{\mathcal{G}}$ represents image global equivariance and is a part of image equivariance $\mathcal{P}$. By combining these equivariance components, we establish product group equivariance. The network architecture exhibits local-multi-global equivariance, as illustrated in Figure 1 (Fig. 1).

C. Voxel Local Equivariance

Traditional voxel encoding and its variants have only translation-equivariance and lack rotation-equivariance [29]. To address this issue, we modify the voxel encoder and achieve the equivariance of the local point cloud, thereby extending local object-level equivariance within a BEV to the global scene.

To realize the local equivariant encoding of the point cloud, inspired by the work [29], [30], we utilize the rotation group $SO(3)$ to act on the voxels, which captures the local information of the point cloud.

$$\mathcal{V}^{'}(p) = [\mathcal{V}(p) \oplus D^{\mathcal{X}}(p)] \tag{6}$$

Hence, pair-wise relationships are constructed to capture the Euclidean distances between points within the voxel.

Then, the pair-wise point distances are added as symmetric geometry feature information to the voxel encoding channels. Based on this process, the point-to-point relationship inside the voxel is strictly kept relative to each other, and we consider that it satisfies the local rotation-equivariant condition. The mathematical proof is shown as follows. Given that the original point cloud: $P = p_1, p_2, \ldots, p_N$, where $p_i = (x_i, y_i, z_i, r_i, n_i)$, (x_i, y_i, z_i) is the 3D coordinate of the point, (r_i, n_i) is the feature information of the point. After voxel encoding the voxel feature representation is obtained as $V = v_1, v_2, \ldots, v_M$, where v_i is the i voxel feature representation, the relative Euclidean distance feature of each point inside the voxel is $f_{dis}(p) = \|p_i - p_j\|, i = 1, 2, \ldots, N; j = 1, 2, \ldots, N$. Next, there exists R for the rotation transformation operator, and the local rotation-equivariant $R(p)$ acts on the original point cloud to obtain the voxel $\mathcal{V}^{'}(p)$. We have $f^{'}_{dis}(p) = \left\|p^{'}_i - p^{'}_j\right\|, i = 1, 2, \ldots, N; j = 1, 2, \ldots, N$. Since the distance between the points is strictly invariant, rotation-transformation is performed on the points inside the voxel, which satisfy the local rotation-equivariant condition, i.e: $f_{dis}(p) = f^{'}_{dis}(p)$.

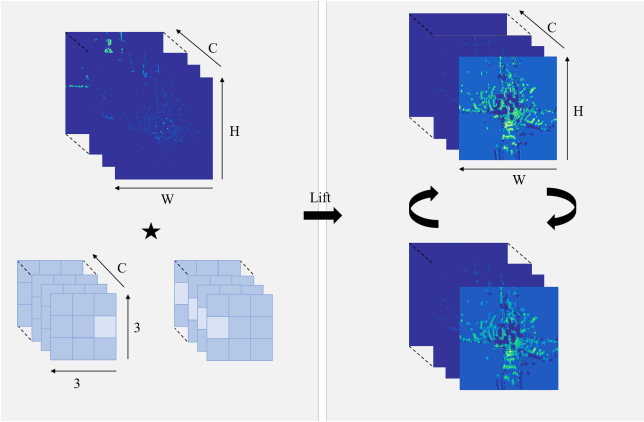


Fig. 2. Voxel Global Rotation-Equivariant Lifting. Where $\star$ represent the convolution operation.

D. Voxel Global Equivariance

Further, to extract the global rotation-equivariant features in the point cloud global scene, we combine the equivariance concept to improve SECOND [8] and FPN [31] to construct the voxel global rotation-equivariant.

To realize global rotation-equivariant, the output features of sparse coding are lifted to group and shown in Fig. 2. According to eq. (1), where lifting can be expressed in mathematical form as:

$$[f_{voxel} \star \psi](g_{c2}) = \sum_{x_p \in \mathbb{Z}^2} \sum_{k} f_{voxel_k}(x_p)\psi_k\left(g_{c2}^{-1}x_p\right) \quad (7)$$

where f_{voxel} represents the result of voxel encoding after the local otation-equivariant and x_p is the BEV feature point. we utilize the cyclic group $c2$ as a rotation-transformation operator in the voxel global equivariance. Then, to extract multi-scale features, we perform multiple convolutions on the group, namely, multi-scale group convolution, as shown

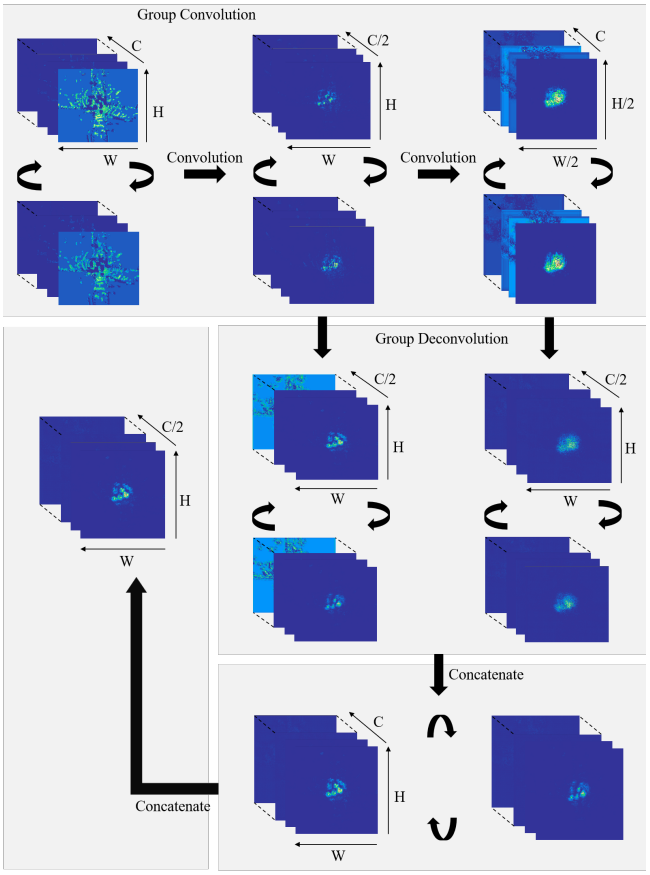


Fig. 3. Voxel Multi-scale Global Rotation-Equivariance

in Fig. 3. According to Eq. (2), the group convolution at this time can be expressed mathematically as:

$$[f_{group} \star \psi](g_{c2}) = \sum_{h \in \mathbb{G}} \sum_{k} f_{group_k}(h)\psi_k\left(g_{c2}^{-1}h\right) \quad (8)$$

where f_{group} represents the multi-scale features on the cyclic group $c2$.

To process multi-scale feature information with multiple directions, we combine the concept of equivariance to improve the FPN network, and thus perform the deconvolution operation on multi-scale information with multiple directions, as shown in Fig. 3. The above strategy further improves the ability of the network to learn features with different directions and scales.

E. Image Global Equivariance

Since the nuScenes dataset [10] is captured by six cameras at a fixed angle, rotation-equivariant at arbitrary angles cannot be realized in the similar way as point clouds. Therefore, we introduce the cyclic group $c2$ to capture global rotation-equivariant information for scenes with large angles. The strategy allows us to synthesize the information from multiple cameras and process the global information of the whole scene more comprehensively and efficiently, thus extracting the global rotation-equivariant features of the camera BEV.

TABLE I

COMPARISON WITH OTHER METHODS (MAP(%) AND NDS) ON NUSCENES VALIDATION SET.

Method	Modality	Per-class										Metric	
		Car	*Truck*	*Bus*	*Trailer*	*C.V.*	*Ped.*	*Motor.*	*Bicycle*	*T.C.*	*Barrier*	*mAP*↑	*NDS*↑
SECOND[8]	L	81.6	51.9	68.5	38.2	18.0	77.4	40.1	18.2	56.9	57.8	50.9	62.0
HotSpotNet[32]	L	84.0	56.2	67.4	38.0	20.7	82.6	66.2	49.7	65.8	64.3	59.5	66.0
UVTR-L[33]	L	85.2	52.8	68.8	41.3	**24.4**	83.5	69.7	53.6	67.3	62.0	60.9	67.7
ProEqBEVNet-L(our)	L	**86.6**	**61.6**	**74.5**	**42.9**	21.8	**87.1**	**71.0**	**55.8**	**76.4**	**69.1**	**64.7**	**69.7**
3D CVF[34]	L+C	83.0	45.0	48.8	49.6	15.9	74.2	51.2	30.4	62.9	65.9	52.7	62.3
FUTR3D[35]	L+C	86.3	61.5	71.9	42.1	26.0	82.6	73.6	63.3	70.1	64.4	64.2	68.0
UVTR-M[33]	L+C	87.2	60.7	71.3	41.1	27.7	85.1	74.1	**66.6**	73.4	66.4	65.4	70.2
MVP[36]	L+C	86.8	58.5	67.4	57.3	26.1	**89.1**	70.0	49.3	**85.0**	**74.8**	66.4	70.2
PointAugmenting[37]	L+C	87.5	57.3	65.2	**60.7**	**28.0**	87.9	74.3	50.9	83.6	72.6	66.8	71.0
ProEqBEVNet-M(our)	L+C	**89.0**	**66.7**	**78.1**	45.6	27.7	88.8	**78.5**	65.8	81.9	71.5	**69.4**	**72.0**

Class name abbreviations: Construction Vehicle (C.V.), Pedestrian (Ped.), Motorcycle(Motor.), Traffic Cone (T.C.), Camera (C), LiDAR (L).

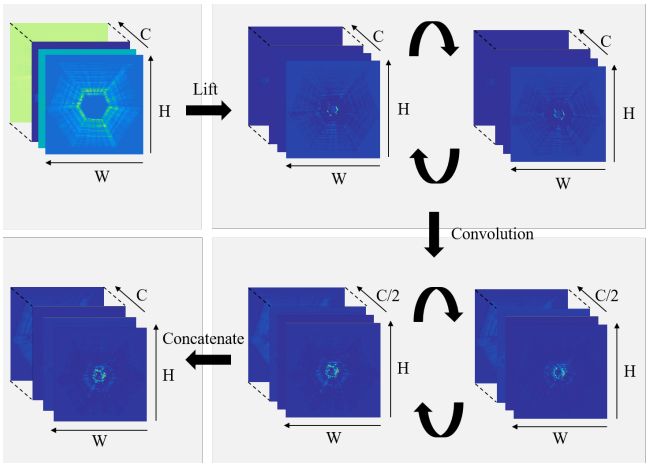


Fig. 4. Image Global Rotation-Equivariance

To realize the fusion of images and point clouds on BEV, we follow the classical work on BEV-based fusion [6], [7]. Specifically, we chose to use DB-Swin-Transformer [38] as the backbone network for the images and FPN [31] as the neck network to extract multi-scale features. Meanwhile, the LLS network [9] is utilized to encode the original multi-views into BEV. To introduce global rotation-equivariant into the BEV encoding and provide richer global rotation-equivariant information of the image to the network, we try to use the multi-scale rotation-equivariant convolution operation, and this process is shown in Fig. 4. According to (1), the multi-scale image feature maps are lifted into groups and the process can be expressed as:

$$[f_{cambev} \star \psi](g_{c2}) = \sum_{x_c \in \mathbb{Z}^2} \sum_{k} f_{cambev_k}(x_c)\psi_k\left(g_{c2}^{-1}x_c\right) \tag{9}$$

where f_{cambev} represents image BEV feature and x_c is image feature point. In order to alignment with the point cloud BEV features, the BEV features with multiple orientations are fused in the channel dimension.

F. Object Detection Head

To adequately capture multi-level degree-of-freedom object information in autonomous driving scenes, efficient object detection head is required to detect fused feature information with rotation-equivariant. meanwhile, following the classical work on BEVFusion [6],[7], we utilize transfusion head [39] as the final object detection head.

IV. EXPERIMENT AND RESULT ANALYSIS

Most of datasets cannot directly provide BEV information of the real road scenes, while BEV generation requires full around-vehicle view. The public dataset nuScenes [10] provides the complete scene information around the vehicle with 6 cameras, which becomes one of the popular public datasets for BEV generation. Moreover, the dataset collects vehicles equipped with one 32-line LiDAR and five millimeter-wave radars as well as IMU. In object detection, it supports the object detection of 10 classes, e.g., cars, trucks, and bicycles. And there are the following types of evaluation metrics: Average Precision (AP), Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), and Average Attribute Error (AAE). In addition, the nuScenes dataset presents the nuScenes Detection Score (NDS), which is calculated using the above six categories.

A. Experiment Setup

Our proposed method is trained and evaluated on the nuScenes validation set and focuses on two metrics: mAP and NDS.

In order to obtain more stable fusion performance, we perform rotation-equivariant pre-training on the camera BEV and the LiDAR BEV, where the camera BEV is trained 20 epochs, select the AdamW optimizer [40], the learning rate is set to 0.00001, the weight decay is set to 0.05. And the lidar BEV is pre-trained using the AdamW optimizer with 0.0001 learning rate, the weight decay is set to 0.01, and the range of the point cloud is set to $[-54.0, -54.0, -5.0, 54.0, 54.0, 3.0]$ and the voxel size is set to $[0.075, 0.075, 0.2]$. Furthermore, CBGS [41] is utilized for class-balanced grouping and sampling augmentation, and fade strategy [37] is used for point cloud training. Then, we train the fused global rotation-equivariant network of images and point clouds for 10 epochs, using the AdamW optimizer with learning rate set to 0.0001 and decay weight of 0.05. Furthermore, utilize Global-RotScaleTransBEV [6] and RandomFlip3DBEV [6]

for data augmentation, and finally inherit the pretrained weights with rotation-equivariance.

B. Experiment Results and Analysis

We will show the experiment results of our network framework on the nuScenes validation set and compare it with mainstream object detection frameworks. The results are shown in Table I. Compared with other mainstream methods, ProEqBEVNet has better results. Furthermore, we show the detection result using local-multi-global equivariant in Fig. 5, and ProEqBEVNet significantly reduces object leakage compared to the detection results without multi-level equivariant.

From the table, it can be found that ProEqBEVNet achieves 69.4% on mAP and 72.0 on NDS. The ProEqBEVNet achieves better results on Car, Truck, Bus and Motorcycle class compared to other mainstream methods. While in other classes besides Trailer, AP is close to other advanced methods, and we believe it benefits from local and global rotation-equivariant feature extraction to extract multi-level degree-of-freedom information, which improves the overall performance. Meanwhile, the product group equivariant network with only LiDAR as input has improved its whole performance due to the local-global equivariant feature extraction from the point cloud and has obtained the best AP in most class.

TABLE II

ABLATION EXPERIMENTS ON THE NUSCENES VALIDATION SET

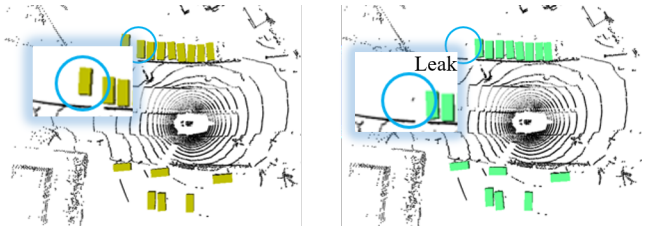
Experiments	LoEq	GloEq	mAP	NDS
1	✗	✗	68.3	70.9
2	✓	✗	69.0	71.3
3	✗	✓	68.9	71.7
4	✓	✓	69.4	72.0

Local-equivariance(LoEq), Global-Equivariance(GloEq)

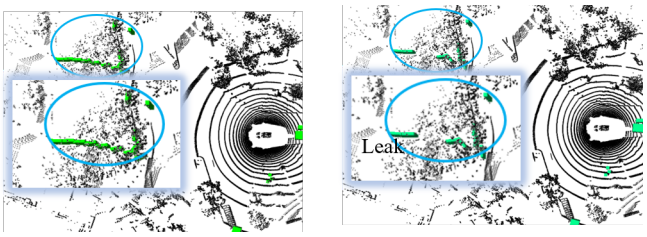
C. Ablation Experiments

To demonstrate the effectiveness of the local equivariant and multi-global equivariant components of our proposed network architecture, we conducted several ablation experiments on the nuScenes validation set, and the experiment results are shown in Table II.

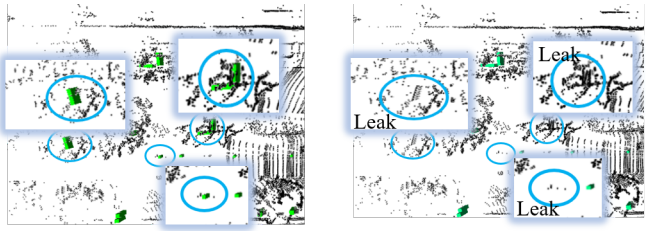
We perform ablation experiments based on a local-global hierarchy to validate the effectiveness of local and global equivariance on the network, respectively. Specifically, ProEqBEVNet achieves a significant improvement of 1.1% and 1.1 in both mAP and NDS compared to the baselines. In Experiments 2 and 3, the performance improvement of local and global equivariance on the overall network is verified, respectively. Particularly, the global equivariance has more significant NDS compared to the local equivariance. In addition, the product group equivariance consisting of local and global equivariance has a better performance enhancement for the network compared to the single equivariance.



(a) Road scene 1: Objects with lateral occlusion.



(b) Road scene 2: Objects with sequential occlusion.



(c) Road scene 3: Objects with dispersed distribution.

Fig. 5. Comparisons of object detection results w/o ProEqBEVNet in sample road scenes. The left figure shows the results of using local-multi-global equivariant, and the right figure shows the results without using the proposed method.

V. CONCLUSIONS

In this paper, leveraging insights from the hierarchical structure of local-global relationships in BEV representations for real-world road scenarios, and the multiple levels of freedom inherent in BEV, we propose a product group equivariant 3D object detection network architecture. The network extracts rich point cloud local rotation-equivariant features by constructing local rotation-equivariant, and further realizes the fusion of point cloud and image global equivariant on BEV. We realize the rotation-equivariant of the network architecture itself in the above way and achieve excellent experiment results on the nuScenes dataset. In addition, we further demonstrate the effectiveness of the local and multi-global equivariant components through ablation experiments.

Although our network architecture initially realizes multi-sensor based local-multi-global equivariant, there are still many factors that we have not taken into account. In the future, we will continue to explore the problems and consider more complex rotation scenes to develop a more efficient equivariant object detection network.

ACKNOWLEDGMENT

This research is supported by National Natural Science Foundation of China (No.42130112, 42371479, 41901335) and KartoBit Research Network(No.KRN2201CA)

REFERENCES

[1] J. Mao, S. Shi, X. Wang, and H. Li, "3d object detection for autonomous driving: A comprehensive survey," *International Journal of Computer Vision*, pp. 1–55, 2023.

[2] J. Lei, J. Guo, H. Yu, H. Lan, C. Li, and Z. Zhang, "Radar-rpn: Accurate region proposal with mmwave radar in 3d detection," in *2022 4th International Conference on Frontiers Technology of Information and Computer (ICFTIC)*. IEEE, 2022, pp. 1033–1037.

[3] 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*, 2021, pp. 11 784–11 793.

[4] Y. Ma, T. Wang, X. Bai, H. Yang, Y. Hou, Y. Wang, Y. Qiao, R. Yang, D. Manocha, and X. Zhu, "Vision-centric bev perception: A survey," *arXiv preprint arXiv:2208.02797*, 2022.

[5] K. Huang, B. Shi, X. Li, X. Li, S. Huang, and Y. Li, "Multi-modal sensor fusion for auto driving perception: A survey. arxiv 2022," *arXiv preprint arXiv:2202.02703*.

[6] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[7] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "Bevfusion: A simple and robust lidar-camera fusion framework," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 421–10 434, 2022.

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

[9] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *European Conference on Computer Vision*. Springer, 2020, pp. 194–210.

[10] 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*, 2020, pp. 11 621–11 631.

[11] Y. Li, L. Ma, Z. Zhong, F. Liu, M. A. Chapman, D. Cao, and J. Li, "Deep learning for lidar point clouds in autonomous driving: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 8, pp. 3412–3432, 2020.

[12] D. Fernandes, A. Silva, R. Névoa, C. Simões, D. Gonzalez, M. Guevara, P. Novais, J. Monteiro, and P. Melo-Pinto, "Point-cloud based 3d object detection and classification methods for self-driving applications: A survey and taxonomy," *Information Fusion*, vol. 68, pp. 161–191, 2021.

[13] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499.

[14] 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*, 2019, pp. 12 697–12 705.

[15] H. A. Mallot, H. H. Bülthoff, J. Little, and S. Bohrer, "Inverse perspective mapping simplifies optical flow computation and obstacle detection," *Biological cybernetics*, vol. 64, no. 3, pp. 177–185, 1991.

[16] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[17] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1477–1485.

[18] J. Huang and G. Huang, "Bevdet4d: Exploit temporal cues in multi-camera 3d object detection," *arXiv preprint arXiv:2203.17054*, 2022.

[19] C. Lu, M. J. G. van de Molengraft, and G. Dubbelman, "Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 445–452, 2019.

[20] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European Conference on Computer Vision*. Springer, 2022, pp. 1–18.

[21] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *International conference on machine learning*. PMLR, 2016, pp. 2990–2999.

[22] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley, "Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds," *arXiv preprint arXiv:1802.08219*, 2018.

[23] C. Esteves, C. Allen-Blanchette, A. Makadia, and K. Daniilidis, "Learning so (3) equivariant representations with spherical cnns," in *European Conference on Computer Vision*, 2018, pp. 52–68.

[24] F. Fuchs, D. Worrall, V. Fischer, and M. Welling, "Se (3)-transformers: 3d roto-translation equivariant attention networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1970–1981, 2020.

[25] 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.

[26] X. Wang and X. Wei, "Continual learning for pose-agnostic object recognition in 3d point clouds," *arXiv preprint arXiv:2209.04840*, 2022.

[27] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *arXiv preprint arXiv:2302.00487*, 2023.

[28] D. K. Gupta, D. Arya, and E. Gavves, "Rotation equivariant siamese networks for tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 362–12 371.

[29] H. Wu, C. Wen, W. Li, X. Li, R. Yang, and C. Wang, "Transformation-equivariant 3d object detection for autonomous driving," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 2795–2802.

[30] X. Wang, J. Lei, H. Lan, A. Al-Jawari, and X. Wei, "Dueqnet: Dual-equivariance network in outdoor 3d object detection for autonomous driving," *arXiv preprint arXiv:2302.13577*, 2023.

[31] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[32] 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*. Springer, 2020, pp. 68–84.

[33] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3d object detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 442–18 455, 2022.

[34] J. H. Yoo, Y. Kim, J. Kim, and J. W. Choi, "3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 720–736.

[35] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "Futr3d: A unified sensor fusion framework for 3d detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 172–181.

[36] T. Yin, X. Zhou, and P. Krähenbühl, "Multimodal virtual point 3d detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 494–16 507, 2021.

[37] C. Wang, C. Ma, M. Zhu, and X. Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 794–11 803.

[38] T. Liang, X. Chu, Y. Liu, Y. Wang, Z. Tang, W. Chu, J. Chen, and H. Ling, "Cbnet: A composite backbone network architecture for object detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 6893–6906, 2022.

[39] 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*, 2022, pp. 1090–1099.

[40] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[41] 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.