Physically Interpretable Object Detection in Foggy Point Cloud

Jinbiao Zhao, Zeyu Xiao, *Member, IEEE*, Zhao Zhang, *Senior Member, IEEE*, Haijun Zhang, *Senior Member, IEEE*, Yi Yang, *Senior Member, IEEE*, and Meng Wang, *Fellow, IEEE*

***Abstract*—Under ideal conditions, 3D object detection (3DOD) provides accurate spatial position and object classification information. However, in adverse weather conditions such as foggy, LiDAR signals suffer from severe scattering, resulting in abnormal point cloud density and distance shifts. This significantly affects the accuracy and reliability of 3DOD. Existing methods in foggy conditions are mainly limited by two factors: first, the scarcity of annotated point cloud data in foggy weather; and second, the lack of foggy-aware guidance mechanisms in existing deep learning frameworks. To address these issues, we propose physically Interpretable Object Detection in Foggy Point Cloud (IFOD), a novel architecture that employs physically-motivated design principles through an Atmospheric Scattering Model (ASM) to achieve robust perception in foggy conditions. The IFOD architecture includes two key innovative components: the Focus Sparse Convolution (FSC) module and the ASM-Guided Enhancement (AGE) module. FSC improves information sparsity issues by adapting the position of output elements. AGE, in addition to maintaining the original 3D spatial branch, generates a 2D pseudo-image feature space branch to extract key fog features for encoding 3D features, and this consistency checking mechanism helps to reduce false predictions and improve the model's ability to recognize object in foggy conditions. Furthermore, to address the data labelling scarcity problem, we introduce FogKITTI, an extended dataset simulated through physical scattering models, providing valuable resources for 3D point cloud perception research in foggy environments. Experimental results show that IFOD achieves superior detection performance on both real-world and simulated datasets.**

***Index Terms*—Point Cloud, Foggy Weather, 3D Object Detection, LiDAR.**

I. Introduction

LIDAR is crucial for autonomous driving, but its performance degrades severely in foggy weather due to signal scattering and point cloud sparsity [1]–[3]. Therefore, using methods based on non-fog data can lead to suboptimal detection performance. This makes achieving reliable 3DOD on foggy a critical requirement. Furthermore, while 4D millimeter-wave radar performs well in foggy conditions, many deployed autonomous vehicles are only equipped with LiDAR [4], [5]. In this situation, directly replacing LiDAR with 4D radar is impractical. Therefore, improving the robustness of LiDAR in foggy conditions through algorithmic means is of significant practical importance.

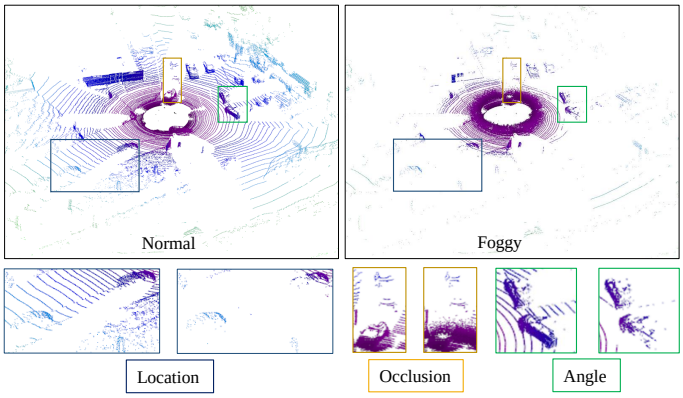

Fig. 1. **Normal vs. foggy LiDAR point cloud.** Visualization of LiDAR point cloud under normal (top left) and foggy (top right) weather conditions, where fog is simulated using an optical system. The bottom row displays a three-fold magnified view of the selected area, highlighting the local structural differences between normal and foggy conditions regarding scanning location, angle, and occurrence. Through detailed observation, the LiDAR returns become significantly denser in proximal areas and notably sparser at longer ranges under foggy conditions. This phenomenon illustrates the severe signal attenuation and spatial sparsity introduced by atmospheric scattering, which compromises the geometric structure of distant object and presents critical challenges for 3DOD in degraded environments.

In practical outdoor autonomous driving scenarios, as shown in Fig. 1, short-range objects yield dense reflection point cloud, whereas long-range object produce sparse reflections. This density imbalance, coupled with LiDAR limitations and adverse conditions, aggravates sparsity and feature extraction. To improve this problem, existing 3D object detection algorithms typically combine 2D image texture information to derive 3D bounding boxes [6]–[8]. For example, VirConv [6] applies depth estimation to generate dense pseudo points and filters redundant points via a distance-based strategy. However, the accuracy of such image-based methods is inherently constrained by depth estimation quality. To overcome the limitations of image-based depth estimation, recent works shift to LiDAR designs [9]–[11]. Adv3D [11] crafts sparse, physically-plausible adversarial point clouds to spoof obstacles, exposing the heightened vulnerability of voxel-based 3D detectors in practical deployment.

With the growing focus on environmental perception in adverse weather, related research continues to advance, yet existing benchmark datasets exhibit notable limitations [12], [13]. For instance, NuScenes-C [14] introduces multiple interfering factors based on NuScenes [15] such as rain, snow, motion blur, and spatio-temporal misalignment. Such coupled interferences not only amplify cross-domain discrepancies but

Jinbiao Zhao and Meng Wang are with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China (e-mail: 2023010067@mail.hfut.edu.cn; eric.mengwang@gmail.com).

Zeyu Xiao is with the Department of Electrical and Computer Engineering, National University of Singapore, 117583, Singapore (e-mail: zeyuxiao1997@163.com).

Zhao Zhang is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China, and also with the Yunnan Key Laboratory of Software Engineering, Yunan 650051, China (e-mail: cszzhang@gmail.com).

Haijun Zhang is with the Department of Computer Science, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: hjzhang@hit.edu.cn).

Yi Yang is with the College of Computer Science and Technology Zhejiang University, Hangzhou 310027, China (e-mail: yangyics@zju.edu.cn).

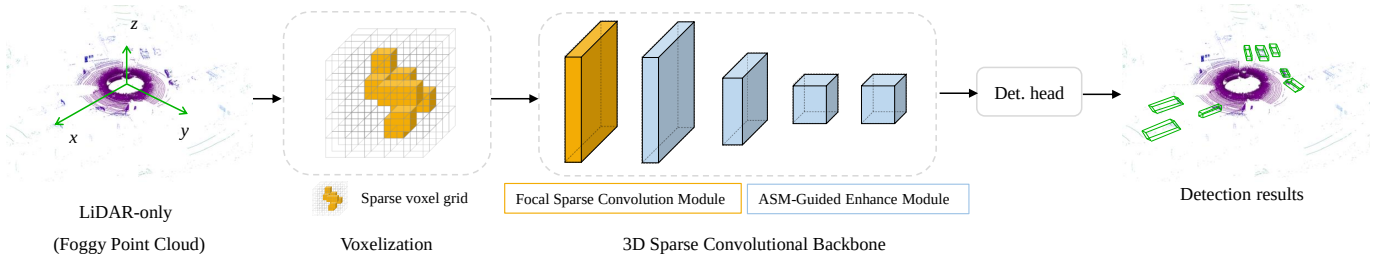

Fig. 2. **IFOD architecture.** The LiDAR data captured in foggy weather is first voxelized by dividing the point cloud space into uniform cubes for point grouping, reducing data complexity and improving processing efficiency. The voxelized data is then processed by a 3D sparse convolutional backbone comprising one FSC module and four AGE modules. The resulting features are fed to a detection head, which performs object detection, including object classification and estimation of position and size.

also hinder fair evaluation of individual interference types within a unified framework. Moreover, the complexity of these composite datasets may exceed actual operational demands. In practical scenarios, such as ship navigation, homogeneous dense fog formed by the continuous condensation of water vapour is the primary threat. Therefore, *Constructing a LiDAR dataset dedicated to foggy weather and conducting in-depth analysis of its characteristics is of great importance for enhancing perception under adverse conditions*.

Unlike existing methods that treat fog as generic noise or rely on image inputs that fail in dense fog, we explicitly analyze two fundamental degradation effects of fog on LiDAR: (i) foreground sparsity inhomogeneity, and (ii) signal scattering. To address these two issues respectively, we propose a physically interpretable architecture, IFOD, guided by the Atmospheric Scattering Model (ASM) [16]. Specifically, the **Focus Sparse Convolution (FSC)** module targets sparsity inhomogeneity by introducing a two-stage residual importance prediction branch, which expands the receptive field to stabilize voxel importance estimation. The **ASM-Guided Enhancement (AGE)** module targets signal scattering by embedding ASM as a structural prior into a LiDAR-only pseudo-image branch, without requiring image inputs. Due to the scarcity of annotated foggy data, we construct FogKITTI by extending the classic KITTI [17] dataset. The fog simulation process adopts a widely recognized physical synthesis method to ensure high realism, providing a reliable and valuable benchmark for advancing LiDAR-based perception in real outdoor foggy scenarios. Our main contributions are summarized as follows:

- We propose a novel 3DOD architecture, IFOD, which achieves physically interpretable perception in foggy weather through design principles guided by the Atmospheric Scattering Model (ASM). IFOD benefits from two specially-designed modules: the FSC module, which addresses fog-induced sparsity inhomogeneity via a two-stage residual importance branch, and the AGE module, which embeds ASM priors into a LiDAR-only pseudo-image branch to handle signal scattering. Thus, it shows robustness in foggy weather.
- We introduce FogKITTI, a dataset generated through optical system-based model simulations to ensure the realism of fog effects. This dataset addresses the lack of annotated foggy data and serves as a controllable research benchmark for analyzing LiDAR signal degradation and evaluating the robustness of 3D detectors in foggy.
- Extensive experiments show that IFOD performs better on real-world and simulated foggy datasets.

II. Related Work

LiDAR-only 3D object detection. The main methods that rely only on LiDAR as input data are mainly divides into two categories: Voxel-based and Point-based [18]–[20]. The voxel-based method realizes a dense representation of the point cloud by dividing the sparse point cloud into regular 3D voxel grids and storing the sampled point cloud data in each grid [21], [22]. This method not only solves the problems of sparsity and disorder in point cloud data but also significantly improves data computational efficiency. Point-based methods can efficiently retain key features and avoid information loss by directly sampling the original point cloud for dimensionality reduction [10], [23]. However, the quadratic computational complexity of the transformer remains a challenge.

Defog in 2D images. Image defogging is an image enhancement technology [24], [25], which aims to restore clear images from blurred images caused by atmospheric scattering. Some work is considered from the perspective of optimizing the atmospheric model. For example, SSID [26] proposes a self-supervised defog framework that generates diverse hazy images via a redefined transmission function, enabling training with only clear images and enhancing robustness on real-world hazy images. There are also some works based on the CNN (Convolutional Neural Network) technique to train the network to learn the mapping function directly from the fog image to the fog-free image to achieve defogging.

Adverse conditions dataset. Due to the high cost of collecting data covering various corruption phenomena in the real-world, a potential direction is to use computer technology to simulate or generate various corruption phenomena on clean datasets with existing labels, in order to reduce data collection costs. For example, the SnowKITTI [27] is based on the KITTI odometer benchmark dataset and simulates snowfall weather conditions through synthesis techniques.

III. Method

A. Voxelization

Given a point cloud of size N, we represent it as an indexed set $x_p = \{x_{p_i}, y_{p_i}, z_{p_i}, r_{p_i} | i = 1, \cdots N\}$, where $x_{p_i}, y_{p_i}, z_{p_i} \in \mathbb{R}^3$ are the Cartesian coordinates along the X, Y and Z axes respectively, and $r_{p_i} \in \mathbb{R}_{\geq 0}$ is generally used to represent the intensity of each point. We subdivide the 3D space into equally

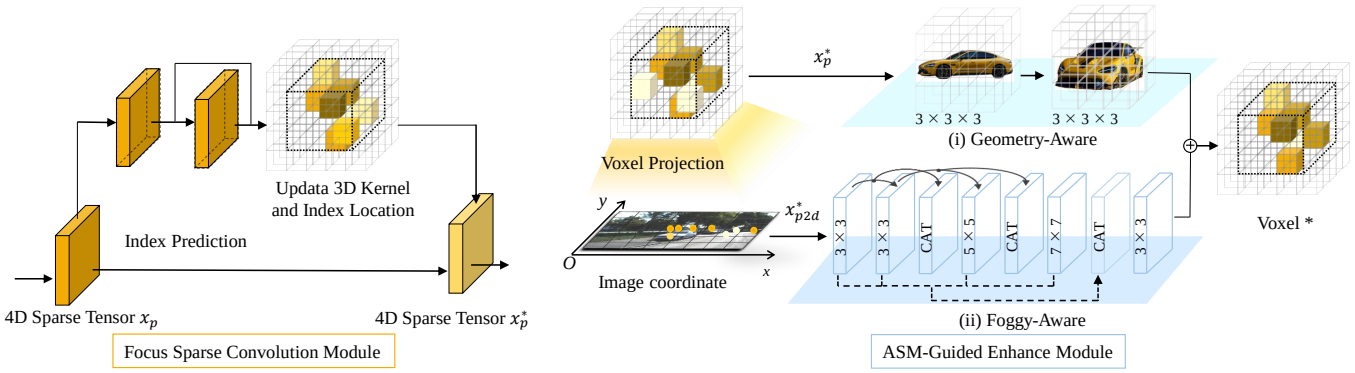

Fig. 3. **Focus Sparse Convolution (FSC) Module** uses two sequential sparse submanifold convolutions to obtain voxel importance indices, then reassembles the convolutions to update the sparse tensors. **ASM-Guided Enhancement (AGE) Module** consists of a 3D branch and a 2D pseudo-image branch derived from voxel projection and mapping. The 2D branch uses ASM priors to perceive fog features and guide the 3D branch. Note: No image inputs.

spaced voxels (basic units of volume in a 3D space) as shown in Figure 2. Here, we assume that the size of each voxel unit is V_d, V_h, and V_w; the range of values for x_p on the X, Y, and Z axes is W, H and D respectively. So the number of grids in the 3D coordinate axis direction is $D^{'} = \left\lceil \frac{D}{V_d} \right\rceil$, $H^{'} = \left\lceil \frac{H}{V_h} \right\rceil$ and $W^{'} = \left\lceil \frac{W}{V_w} \right\rceil$ respectively. The $\lceil \cdot \rceil$ represents the ceiling function, which rounds up to the nearest integer. Then, based on the coordinates of the 8 corner points of the single voxel, it can be determined whether x_{p_i} belongs to itself. Therefore, this type of grouping can easily lead to a zero or overly dense number of points. In fact, LiDAR data consists of about 100k points. If the computer directly processes all points, it increases the memory overhead and the highly variable point density in the entire voxel space, which may cause detection bias. To alleviate this problem, Voxelnet [21] proposes to conduct a fixed number of samples within each voxel, which can reduce sampling bias and computational complexity. As of now, this voxel-based method is still the mainstream design. We also use the advantages of this technology.

After voxelization, x_p is transformed into a 4D sparse tensor representation, denoted as $x_p \in C \times D^{'} \times H^{'} \times W^{'}$, where C represents the dimension of the voxel-wise. The distribution of 4D sparse tensors is uneven, leading to low efficiency when processed with uniform grids.

B. Focus Sparse Convolution Module

To effectively address this issue of uneven spatial distribution and computational inefficiency inherent in the 4D sparse tensor x_p, we introduce a dynamic voxel attention mechanism to adaptively reallocate computational focus toward high-value foreground regions, as shown in Figure 3. Specifically, To determine the significance of each voxel, FSC employs a dual-stage refinement branch consisting of two sequential submanifold convolutions. The residual connection between these two stages expands the effective receptive field from $3 \times 3 \times 3$ to $5 \times 5 \times 5$. This multi-scale design fuses local and broader spatial context for robust importance estimation. The process can be formulated as

$$\mathcal{P}_{\text{out}} = \mathcal{P}^* \left(p, K_1(x_p)\right) \cup \mathcal{P}^* \left(p, K_2(K_1(x_p))\right), \tag{1}$$

where K_1 and K_2 denote two sparse submanifold convolutions (both with kernel size $3 \times 3 \times 3$, stride 1). K_1 processes the input sparse tensor to produce an initial importance map focusing on local patterns; K_2 then takes the output of K_1 and further aggregates a broader spatial context, effectively expanding the receptive field from 3^3 to 5^3 via the residual connection. $\mathcal{P}_{\text{in}}$ and $\mathcal{P}_{\text{out}}$ denote the set of input and output voxel positions, respectively. $p \in \mathcal{P}_{\text{in}}$. Unlike single-stage sparse importance prediction [28], our dual-stage FSC is better suited for foggy scenes (see Table VI, K_1 vs. $K_1 + K_2$).

Then, we sort the features to identify key features and obtain their index information I_p. $I_p \in [0,1]^{|K^d|}$ is the cubic importance map sharing the same spatial shape as the convolution kernel K^d (*e.g.*, $3 \times 3 \times 3$ for kernel size 3). Based on the sorting results, FSC empirically [28] sets the threshold τ to 0.5 to obtain the most significant input positions $\mathcal{P}^*$ and adjusts the kernel shape $\mathcal{K}^*$. By applying this thresholding mechanism, the module effectively prunes unpromising background voxels and adaptively concentrates the network's receptive field on salient foreground regions.

Update $\mathcal{P}^*$:

$$\mathcal{P}^* = \{p \mid I_p \geq \tau, p \in \mathcal{P}_{\text{in}}\}, \tag{2}$$

Update $\mathcal{K}^*$:

$$\mathcal{K}^* = \left\{k \mid p + k \in \mathcal{P}_{\text{in}}, I_p^k \geq \tau, k \in K^d\right\}, \tag{3}$$

where k is an offset distance from p. I_p^k is location in kernel space. These refinements alleviate neighborhood disconnections due to limited receptive fields in standard sparse submanifold convolutions.

Finally, we combine $\mathcal{P}^*$ and $\mathcal{K}^*$ to automatically reassemble a new sparse submanifold convolution operator, which is then applied to the initial 4D sparse tensor x_p to produce the output x_p^*. By reconfiguring the convolution to operate exclusively on the updated spatial set x_p^*. FSC dynamically adjusts the model's focus on foreground regions in the data, making it more efficient in handling sparse data. Additionally, inspired by the architecture design of the YOLOX [29] series, the FSC is activated only once within IFOD.

C. ASM-Guided Enhancement Module

As shown in Figure 3, the AGE consists of two branches: (i) one that processes point cloud information in a natural 3D space, and (ii) one that operates in a 2D space derived from voxel features through projection and mapping calculations.

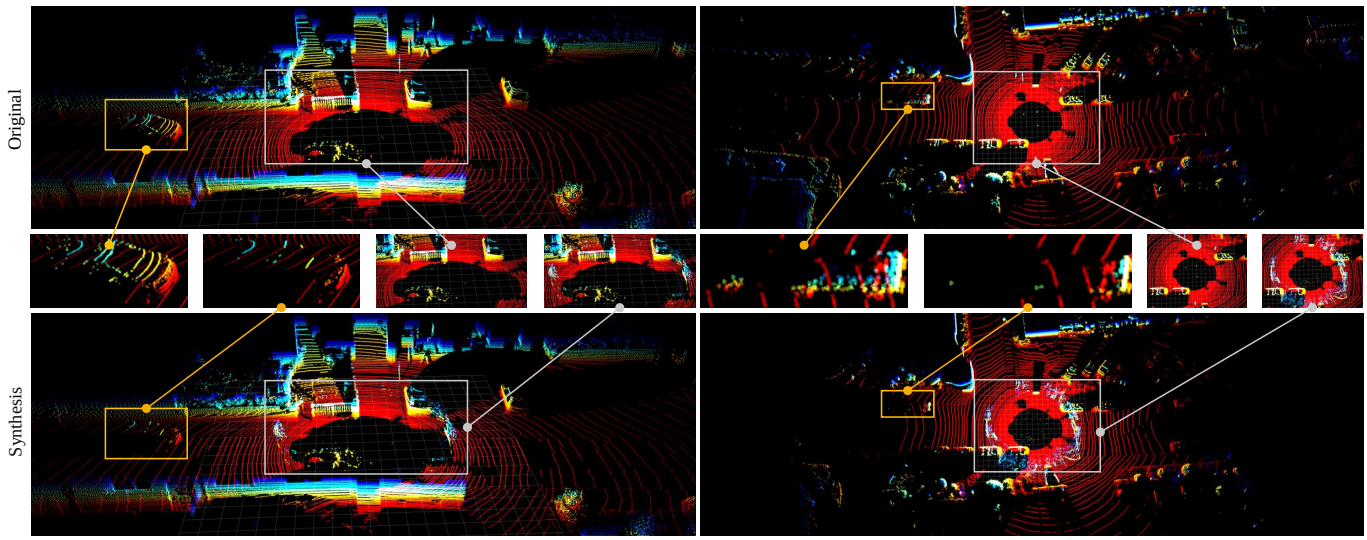

Fig. 4. **Visualization of the FogKITTI dataset construction process.** From top to bottom: the original KITTI scenes, the locally zoomed-in regions for detailed comparison, the foggy scenes that are simulated using the atmospheric scattering model (ASM), and the final synthesized foggy point cloud.

The 2D branch is responsible for perceiving the foggy feature through physics-guided priors derived from ASM, and using them to guide the 3D branch to achieve more accurate distance regression. Unlike existing multi-modal fusion methods (e.g., RoboFusion [30], GraphAlign++ [31], ViKIENet [32]) that rely on real camera images, our AGE module operates on a pseudo-image space derived purely from LiDAR voxel projections. This design ensures robustness even when camera sensors fail in dense fog. The ASM prior is embedded as a feature-level regularization, guiding the 2D branch to learn fog-robust representations without requiring paired image data.

Encoding geometry features in the 3D space. In the 3D branch, to simplify computations and improve efficiency, we apply the sparse convolution of the 3D submanifold. As shown in Figure 3, x_p^* is fed into the 3D branch. The process analyzes key geometric features: shape, size, spatial position, and their relationships, which are physical attributes. Concurrently, this also leverages the unique advantages of the point cloud.

Encoding foggy-aware features in 2D image space. Under the influence of atmospheric scattering effects, ASM accurately reveals the generation mechanism of foggy images. Consequently, ASM has become a classic theoretical model in this field of research. ASM is formulated as

$$I\left(x\right)=J\left(x\right)t\left(x\right)+A\left(1-t\left(x\right)\right), \tag{4}$$

where $I\left(x\right)$ denote the observed fog image, $J\left(x\right)$ denote the restored clean image. A denote the global atmospheric light, and $t\left(x\right)$ denote the transmission matrix.

In foggy weather, although 2D images experience a reduction in contrast, a considerable portion of texture information can still be preserved through ASM-based pixel alignment and grayscale distribution, thereby maintaining relatively strong robustness. In comparison, point clouds exhibit even more pronounced limitations in complex texture analysis due to their inherent sparsity and non-uniformity. This deficiency in explicit texture cues is the fundamental reason why point clouds are highly sensitive to fog-induced sparsity, whereas the 2D pseudo-image branch can leverage grayscale distributions and pixel alignment to maintain structural robustness. Foggy conditions further amplify the disparity in texture information processing between 2D image pixels and point cloud, underscoring the need for tailored cross-modal strategies.

In order to compensate for the lack of texture information in the point cloud, we have designed a branching structure, as shown in Figure 3. The 2D branch is responsible for perceiving the foggy feature and using it as a priori information to guide the 3D branch to achieve more accurate distance regression. This effectively improves the 3DOD performance of the model in complex environments such as foggy weather. Next, we will provide a detailed introduction to branch (ii).

First, we convert the index of x_p^* into the coordinates of the corresponding grid. Then, based on the calibration parameters of the LiDAR to camera project, the grid points are projected onto the 2D image plane to obtain a 2D sparse tensor x_{p2d}^*.

Second, in order to better integrate into the 3D branch network, Eq. (4) needs reformulated as

$$\begin{cases} J\left(x\right)=K\left(x\right)I\left(x\right)-K\left(x\right)+b \\ K\left(x\right)=\frac{\frac{1}{t(x)}(I(x)-A)+(A-b)}{I(x)-1} \end{cases}, \tag{5}$$

$K(x)$ contains some physical parameters [33], such as the gamma correction, and b is a constant and is set to 1. Next, based on Eq. (5), we design a fog-aware branch. Specifically, the 2D branch applies five 2D submanifold convolutional layers with varying filter sizes to extract multi-scale feature representations. This configuration enables the branch to perceive spatial structures at different resolutions, thereby capturing both fine-grained details and coarse contextual patterns. To maintain information integrity throughout convolution, intermediate skip connections are introduced between layers, mitigating the loss of critical features during deep transformations. Finally, the 2D branch features are fused into the 3D branch outputs to produce robust voxel representations.

By projecting voxels onto a 2D pseudo-image plane, AGE leverages the robustness of grayscale distribution and the physical principles of ASM to extract complementary multi-scale representations. This design contributes to mitigating visibility-induced ambiguities, potentially offering more robust

TABLE I
3D DETECTION RESULTS ON THE FOGKITTI VALIDATION SET.

Method	Times(ms)	Data	Car 3D AP (R40)			Pedestrian 3D AP (R40)			Cyclist 3D AP (R40)		
			Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
PillarNet [34]	33	L	88.6840	77.9533	75.3279	45.4910	42.0234	39.0678	83.2726	56.5423	52.4081
VoxelSeT [35]	35	L	90.2738	79.1324	76.1979	55.0862	49.8556	45.4381	88.4541	64.4059	60.0470
PV-RCNN [36]	110	L	90.8294	80.3819	78.5539	64.8123	57.1358	52.7293	88.8947	62.5982	58.5351
VoxelNeXt [37]	35	L	87.8663	77.5695	75.3037	58.5709	53.9718	49.5832	88.3770	60.8011	56.3868
DSVT-Voxel [38]	62	L	87.6817	67.0953	66.2064	62.5703	55.8523	50.6005	84.3842	56.2443	52.5076
SEED [39]	65	L	76.8267	67.2117	64.8117	38.6698	33.5821	29.9865	55.2820	40.3254	37.1907
LION [23]	139	L	90.0521	79.0088	77.5261	65.0829	58.7226	53.4986	81.0174	58.0715	54.2095
Voxel-Mamba [40]	132	L	88.0536	77.3783	74.9371	63.8114	58.5620	53.4260	90.0599	62.9944	59.1423
FSHNet [41]	-	L	-	-	-	-	-	-	-	-	-
GraphBEV [42]	-	L+C	-	-	-	-	-	-	-	-	-
RoboFusion [30]	201	L+C	91.8636	82.0190	**81.9645**	64.9697	**58.8795**	**53.9657**	**92.3489**	**71.0836**	**67.1979**
IFOD(Ours)	101	L	**92.7060**	**82.6434**	80.4353	**65.4331**	58.4810	53.5103	90.5963	65.1276	61.8475

Note: "-": non-square grid feature incompatible with square-transformer head; **bold** = best, underline = second best.

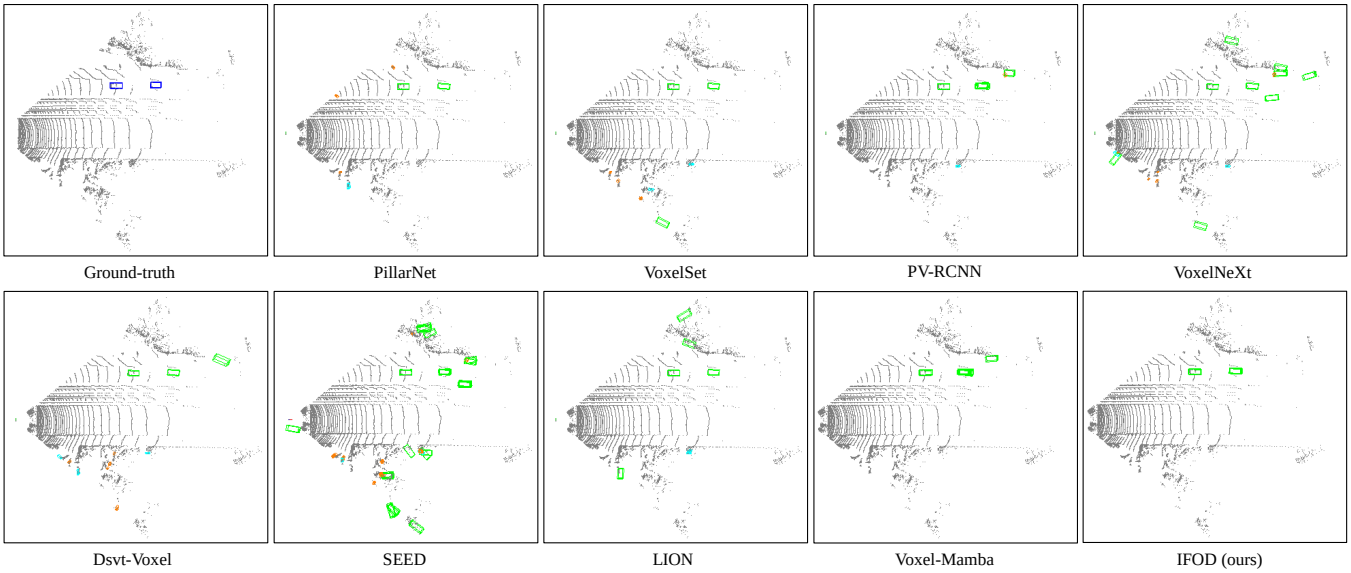

Fig. 5. The visualization results of IFOD to compare IFOD with the latest and most advanced on FoggKITTI. ground-truth has Car, and the predicted object by the method are Car, $Pedestrian$ and $Cyclist$.

representations than conventional 3D-only detectors under certain foggy conditions. AGE effectively extracts and integrates scene information from both 2D and 3D spaces, thereby enhancing detection performance. It significantly reduces misjudgment rates, particularly in challenging scenarios with long detection distances and minimal scanned LiDAR data. This capability plays a crucial role in improving the robustness and reliability of detection algorithms, especially in environments with limited or degraded sensory input.

At the training phase, the detection head configuration of IFOD follows Voxel R-CNN [43], and is supervised by the weighted cross-entropy loss $\mathcal{L}^{\text{wce}}$ and the Lovász-Softmax loss $\mathcal{L}^{\text{lovasz}}$ [44]. $\mathcal{L}_{\text{total}}$ is formulated as

$$\mathcal{L}_{\text{total}} = \mathcal{L}^{\text{lovasz}} + \mathcal{L}^{\text{wce}}. \tag{6}$$

IV. EXPERIMENTS

A. Dataset

KITTI and FogKITTI. The KITTI dataset [20] is extended to FogKITTI via a physically-accurate fog simulation [57] based on ASM (Fig. 4), providing a benchmark for foggy LiDAR detection. Details are in the Supplementary Material.

Seeing Through Fog (STF). For real-world evaluation, we use the STF dataset [58], training on its clear-weather daytime subset (filtering low-quality frames) and testing on light fog, dense fog, and snowy subsets.

B. Experimental Settings

Network details. IFOD consists of one level of FCS and four levels of AGE blocks with feature sizes of 4, 8, 16, 32, and 32, respectively. Voxel size is [0.05, 0.05, 0.05].

Data augmentation. We adopt the widely used local and global data augmentation, including ground-truth sampling, and global transformations such as rotation and flipping.

Evaluation Metrics. We adopted 3D Average Precision (AP) under 40 recall thresholds (R40). The IoU thresholds in this metric are 0.7, 0.5, and 0.5 for Car, $Pedestrian$, and $Cyclist$, respectively. Additionally, for each of these classes, difficulty levels (Easy, Moderate (Mod.) and Hard) are set.

Training and inference details. All three detectors are trained on single 4090 GPU. We use a learning rate of 0.01 with a one-cycle learning rate strategy and trained the IFOD for 50 epochs. Our method is implemented based on the open-source framework OpenPCDet [45].

TABLE II
3D DETECTION RESULTS ON THE KITTI VALIDATION SET.

Method	Data	Car 3D AP (R40)			Pedestrian 3D AP (R40)			Cyclist AP (R40)		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
PillarNet [34]	L	86.0226	76.9047	73.8289	33.8919	27.1914	24.8334	73.0211	51.8965	48.5194
VoxelSeT [35]	L	88.2187	78.8008	75.4201	26.6152	19.8979	17.8557	74.5118	53.9405	50.2065
PV-RCNN [36]	L	89.3149	80.3691	78.4160	27.9779	22.5660	20.2987	83.5761	62.8348	58.4931
VoxelNeXt [37]	L	86.1829	77.4516	74.8445	41.7772	33.6074	29.8064	70.0207	47.3284	44.0532
DSVT-Voxel [38]	L	87.2974	68.2204	65.9974	37.7098	31.6986	28.1978	79.9860	51.7373	48.1637
SEED [39]	L	73.7358	65.5766	62.5184	3.3838	1.8266	1.5200	7.8142	10.8722	9.9201
LION [23]	L	88.1417	79.7811	77.4050	39.3333	33.3911	29.7057	74.3075	53.5889	49.0633
Voxel-Mamba [40]	L	88.3982	77.9773	75.2562	46.2959	39.8956	35.7026	83.0477	54.7166	51.1792
FSHNet [41]	L	-	-	-	-	-	-	-	-	-
GraphBEV [42]	L+C	-	-	-	-	-	-	-	-	-
RoboFusion [30]	L+C	91.4335	**83.4827**	**81.8487**	48.5794	42.3708	38.1816	**88.3565**	**69.3890**	**65.3382**
IFOD(Ours)	L	**92.7446**	83.2610	80.9052	**62.9030**	**56.2446**	**50.0685**	88.1665	67.3777	62.8184

Note: "-": non-square grid feature incompatible with square-transformer head; **bold** = best, underline = second best.

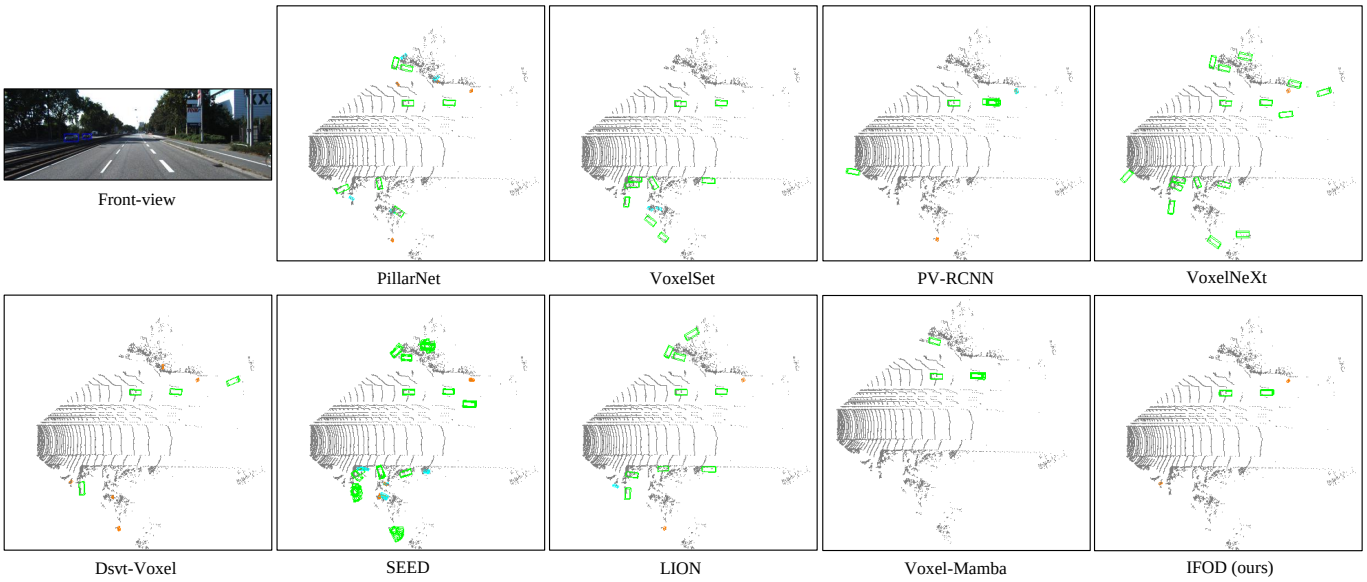

Fig. 6. The visualization results of IFOD to compare IFOD with the latest and most advanced on KITTI. Ground-truth has Car, and the predicted object by the method are Car, $Pedestrian$ and $Cyclist$.

C. Comparison with State-of-the-art Methods

We compare IFOD with the latest and most advanced 3DOD methods as shown in Table I (FogKITTI), Table II (KITTI validation), Table V (KITTI test), and Table VI (STF).

Results on FogKITTI validation set. The Table I compares the LiDAR-based detectors listed in the table across most object classes and difficulty levels, showing strong resilience to visibility degradation. For Car detection, IFOD achieves 92.7060%, 82.6434%, and 80.4353% AP(R40) under Easy, Mod. and Hard settings, respectively. It is higher than PV-RCNN 1.8766, 2.2615, and 1.8814, which have the second-best performance. The performance gap becomes more pronounced under the Hard setting, where fog density and range attenuation severely impact object geometry. This highlights the effectiveness of the FSC module in improving the saliency of local features under sparse or degraded conditions. For $Pedestrian$, LION and Voxel-Mamba perform competitively with IFOD due to their spatial context reasoning, which benefits real-world small objects.

Among multi-modal methods, RoboFusion enhances reliable perception capabilities using LiDAR and camera data (L+C). In easy settings, IFOD outperforms RoboFusion on *Car* and *Pedestrian*, demonstrating the robustness of our physics-based spatial priors. However, in moderate and hard, the geometric information from LiDAR becomes weaker, and semantic information from the image takes precedence, thus RoboFusion consistently ranks first.

Results on KITTI validation set. Table II shows that IFOD outperforms all LiDAR-based methods in every class and difficulty level of the KITTI validation set. Notably, several comparative methods exhibit significant performance degradation in the $Pedestrian$ and $Cyclist$.

Furthermore, the inherent severe performance gap between foggy and clear-weather conditions is a direct consequence of the physical limitations of LiDAR sensors, as atmospheric scattering inevitably leads to signal power attenuation and point cloud sparsity. In this context, the state-of-the-art results achieved by IFOD on the original clear-weather KITTI dataset validate the superior generalization capability of our ASM-guided design. Even when compared with the multi-modal baseline RoboFusion, IFOD delivers highly competitive

TABLE III
3D CAR DETECTION RESULTS ON THE STF DATASET.

Method	Data	Light Foggy			Dense Foggy			Snowy		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
PillarNet [34]	L	32.8995	31.3624	29.7094	25.2356	23.1198	20.8543	31.5218	29.2257	25.6542
VoxelSeT [35]	L	30.7725	29.2577	27.7680	22.5438	20.8611	18.4443	35.5874	32.9812	28.5726
PV-RCNN [36]	L	31.5631	30.6889	29.0871	24.2642	22.7863	20.9528	34.5059	32.4001	28.7519
VoxelNeXt [37]	L	17.3975	18.5277	17.9467	8.2531	8.5127	7.5120	18.0855	18.4011	16.7843
DSVT-Voxel [38]	L	35.0752	33.4100	29.9332	24.1055	22.5761	20.3077	33.4217	31.5699	26.9766
SEED [39]	L	31.0969	29.7114	27.7796	25.8847	24.3957	22.2540	29.6268	27.7908	23.9868
LION [23]	L	32.6389	31.4097	29.6311	24.7756	22.8386	20.8425	33.4461	31.5104	27.5324
Voxel-Mamba [40]	L	30.1087	28.5672	26.6003	16.6202	15.4841	14.1491	29.4128	27.4644	23.6720
FSHNet [41]	L	-	-	-	-	-	-	-	-	-
GraphBEV [42]	L+C	-	-	-	-	-	-	-	-	-
RoboFusion [30]	L+C	29.7203	28.7191	27.2177	21.6743	20.6297	18.3718	27.7063	27.2411	24.4308
IFOD(Ours)	L	**35.1540**	**33.7920**	**31.2927**	**28.7391**	**27.4321**	**24.8616**	**35.6823**	**34.2481**	**30.2529**

Note: "-": non-square grid feature incompatible with square-transformer head; **bold** = best, underline = second best.

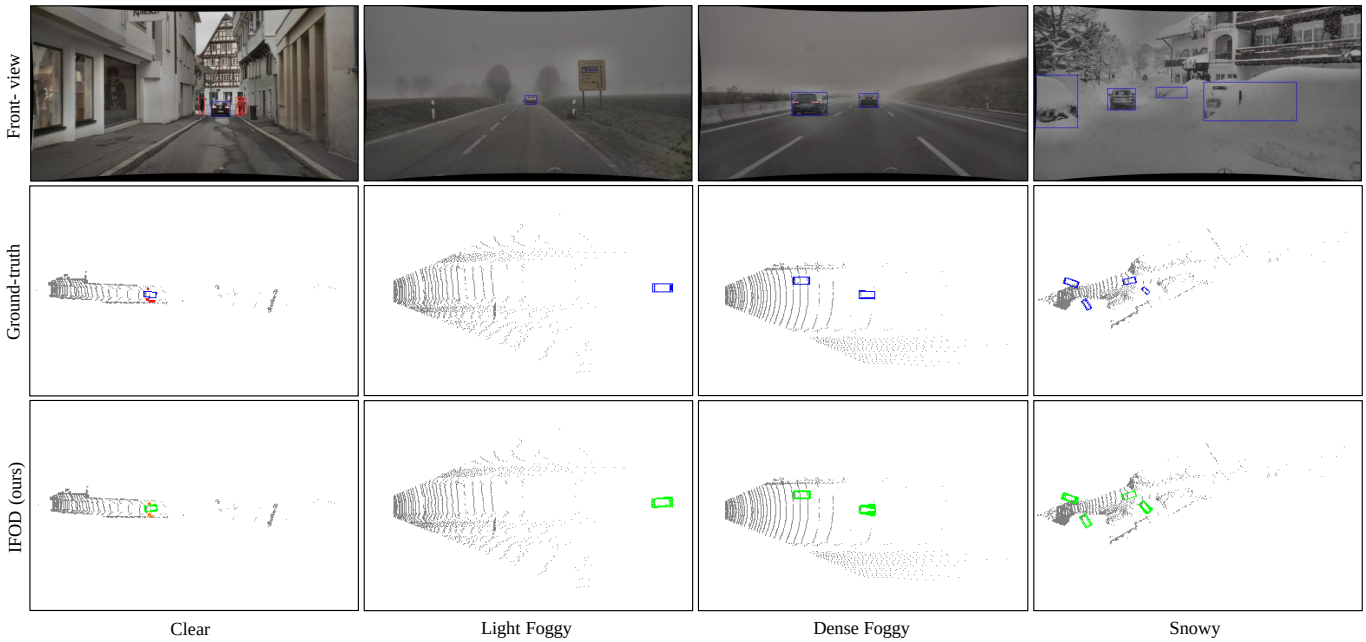

Fig. 7. The visualization results of IFOD on STF. Ground-truth has Car and $Pedestrian$. The predicted object by the method are Car and $Pedestrian$.

results. Specifically, it surpasses RoboFusion in all difficulty levels of the Pedestrian class on FogKITTI.

Results on KITTI test set. To further validate the generalization of IFOD on real-world clean-weather scenes, we submit the model to the official KITTI test server, and the results are shown in Table V. On the Car, IFOD achieves a Mod. 3D AP of 81.51, which is competitive with LiDAR-only baseline methods including PV-RCNN (81.88) and VoxSeT (82.06). The test set contains unseen scene sequences, leading to a performance gap of approximately -1.8 between validation and test results for Car. IFOD also demonstrates competitive performance compared to several recently proposed methods, such as DSFNet (70.44), Fde3D (75.57), and MSFASA-3DNet (77.21). However, multi-modal fusion methods such as GraphAlign++, ViKIENet, and VirConv achieve superior detection performance by leveraging complementary information from image sensors. Overall, IFOD delivers stable and reliable results on the official test set.

Results on STF dataset. Table III shows the Car detection performance of IFOD and other state-of-the-art latest and most advanced methods on the real-world STF dataset, covering LF, DF and challenging Snowy scenarios.

In LF, IFOD achieves 35.1540%, 33.7920%, and 31.2927% AP (Easy/Mod./Hard), leading on Mod./Hard and matching DSVT-Voxel on Easy. In DF, IFOD outperforms all compared methods. In Snowy, IFOD also achieves the best results (35.6823%, 34.2481%, and 30.2529%). In conclusion, IFOD is not only effective in foggy conditions but also robust to other weather conditions.

However, RoboFusion fails to achieve the highest score on all three test sets, and its performance is even worse than methods using LiDAR-only. This may be because when the model processes egraded images and degraded point clouds simultaneously, the corrupted camera data not only fails to provide complementary information but also reduces the reliability of multi-modal fusion.

It is worth noting that the overall lower performance on the real-world STF dataset compared to synthetic FogKITTI is a

TABLE VI
CROSS-DATASET EVALUATION RESULTS FOR 3D CAR DETECTION BETWEEN FOGKITTI AND STF.

Method	Data	Light Foggy ($F \rightarrow S$)			Dense Foggy ($F \rightarrow S$)			Snowy ($F \rightarrow S$)			Foggy ($S \rightarrow F$)		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
PillarNet [34]	L	0.8754	0.7617	0.7445	3.9258	3.4268	3.2126	0.0100	0.0100	0.0100	18.4431	13.4534	11.7719
VoxelSeT [35]	L	0.7353	0.7353	0.7353	0.0373	0.0373	0.0373	0.0000	0.0000	0.0000	4.8817	4.6571	4.4384
PV-RCNN [36]	L	0.2006	0.2287	0.2287	3.7495	3.6212	3.2691	0.1276	0.0266	0.0295	24.9584	17.7055	15.8075
VoxelNeXt [37]	L	0.0481	0.0517	0.0553	0.6412	0.4744	0.3877	0.0000	0.0012	0.0012	4.9092	2.4303	2.3292
DSVT-Voxel [38]	L	0.1835	0.2143	0.1937	0.7765	0.8088	0.4501	0.0079	0.0118	0.0131	14.8390	11.8801	11.3091
SEED [39]	L	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
LION [23]	L	1.4576	1.5890	1.3485	2.3198	2.3435	1.6069	0.2624	0.2603	0.2507	30.9282	22.1604	19.7895
Voxel-Mamba [40]	L	0.0979	0.0428	0.0428	0.8687	0.9259	0.5906	0.2542	0.2542	0.2542	15.6263	10.2530	8.3976
RoboFusion [30]	L+C	0.43010	0.7353	0.7353	0.0000	0.6250	0.6250	0.1256	0.1991	0.2586	32.6687	22.5274	19.3546
IFOD(Ours)	L	**7.9317**	**8.4795**	**8.5035**	**12.0603**	**11.4231**	**10.2422**	**9.7097**	**10.0357**	**9.0919**	**54.1730**	**40.5307**	**39.6228**

TABLE V
PERFORMANCE COMPARISON ON THE KITTI TEST SET.

Method	Data	Car AP		
		Easy	Mod.	Hard
PillarNet [34]	L	89.65	81.06	76.67
VoxSeT [35]	L	88.53	82.06	77.46
PV-RCNN [36]	L	90.25	81.43	76.82
Fde3D [46]	L	86.08	75.57	70.47
MSFASA-3DNet [47]	L	86.06	77.21	71.89
GraphAlign++ [31]	L+C	90.98	83.76	80.16
ViKIENet-R [32]	L+C	91.20	86.04	81.18
VirConv-S [6]	L+C	**92.48**	**87.20**	**82.45**
IFOD(Ours)	L	88.37	81.51	76.87

direct consequence of the substantial domain gap, characterized by STF's lower point density and higher sensor noise. Rather than simulator flaws, this trend confirms the extreme difficulty of real-world visibility-induced ambiguities.

Moreover, the performance of our IFOD combined with the Segment Anything Model (SAM) [48] component from RoboFusion is provided in the supplementary material.

Inference Efficiency. We evaluate the computational performance of various latest and most advanced detection methods and report their inference time, as shown in Table I. While our method does not achieve the highest inference speed, it offers a favourable trade-off. Specifically, IFOD's performance surpasses that of PillarNet, VoxelSet, DSVT-Voxel, and SEED. Furthermore, it exhibits a more comprehensive advantage over LION and Voxel-Mamba.

D. Cross-Dataset Validation on FogKITTI and STF

We evaluate the cross-dataset generalization capability and practical value of FogKITTI by conducting experiments between FogKITTI and STF (Table VI). Two transfer directions are considered: $F \rightarrow S$ (training on FogKITTI, testing on STF) and $S \rightarrow F$ (vice versa).

The results show a significant performance asymmetry. Under $F \rightarrow S$, all methods suffer severe degradation; IFOD achieves the relatively optimal results (e.g., on dense fog Hard: 10.24% AP vs. PV-RCNN 3.2691% and RoboFusion 0.6250%), demonstrating its physical modeling in mitigating the domain gap. Under $S \rightarrow F$, IFOD reaches 39.6228% AP, nearly twice the second-ranked method, indicating that real fog data generalizes more easily to simulated fog. This asymmetric phenomenon stems from intrinsic differences: FogKITTI uses a controllable physical fog simulation dominated by atmospheric scattering, while STF contains real-world sensor noise and non-uniform point cloud loss. Thus, FogKITTI serves as a physically controllable research benchmark to isolate core fog degradation mechanisms, and the results confirm IFOD's relative robustness under domain shifts.

TABLE VI
ABLATION RESULTS ON THE FOGKITTI VALIDATION SET USING DIFFERENT DESIGN COMPONENTS.

FSC		AGE	Car 3D AP		
K_1	K_2		Easy	Mod.	Hard
w/o	w/o	w/o	92.0533	80.6395	78.2279
✓	w/o	w/o	92.2447	82.3454	80.0340
✓	✓	w/o	92.4494	82.5600	80.3575
w/o	w/o	✓	92.2993	82.5901	80.1959
✓	✓	✓	**92.7060**	**82.6434**	**80.4353**

E. Ablation Study

Effectiveness of FSC. As shown in Table VI, the removal of either FSC leads to a notable performance drop. When FSC is removed, the Car AP for Moderate cases drops from 82.7600% to 82.4268%, highlighting the importance of dynamic voxel attention in focusing on informative sparse features. Furthermore, the performance of the dual-stage K_1+K_2 is superior to that of the single-stage K_1, confirming the necessity of a larger receptive field.

Effectiveness of AGE. When the AGE module is removed, the Mod. AP for the Car decreases from 82.8015% to 82.1098%, indicating that fog-aware features extracted by AGE from the 2D pseudo-image branch enhance the model's robustness. However, when the two modules were integrated together, the final IFOD model achieved a cumulative moderate 3D AP of 82.6434%, which is 2% higher than the baseline total gain. This is consistent with the logic of IFOD design, where the FSC module focuses on enhancing foreground voxel saliency and trimming redundant background noise, while the AGE module utilizes ASM guided physical priors to refine geometric features and mitigate fog induced degradation.

Ablation on the number of FSC and AGE blocks. To validate the design choice of one FSC and four AGE blocks, we further conduct a joint accuracy-efficiency analysis across different module counts on the FogKITTI validation set, as reported in Table VII. Three observations are noteworthy. First, increasing the FSC from 1 to 2 achieves a Mod. AP gain of +1.74, but increases the runtime by 21ms. Adding more feature space convolution modules (3 or 4) results in diminishing returns and increases latency. Second, when FSC is fixed at 4, increasing the number of AGE blocks does not bring a stable improvement in accuracy, and additional AGE blocks

TABLE VII
ABLATION STUDY ON THE NUMBER OF FSC AND AGE BLOCKS ON THE FOGGYKITTI VALIDATION SET.

FSC	AGE	Runtime (ms)	Car AP		
			Easy	Mod.	Hard
1	4	101	92.7060	82.6434	80.4353
2	4	122	92.9774	84.3760	82.2977
3	4	135	92.4099	82.9972	82.2319
4	4	155	92.3097	82.7904	81.6890
4	1	117	93.2040	83.2502	82.4690
4	2	126	92.3816	82.9178	82.3632
4	3	132	92.8091	83.3201	82.5622

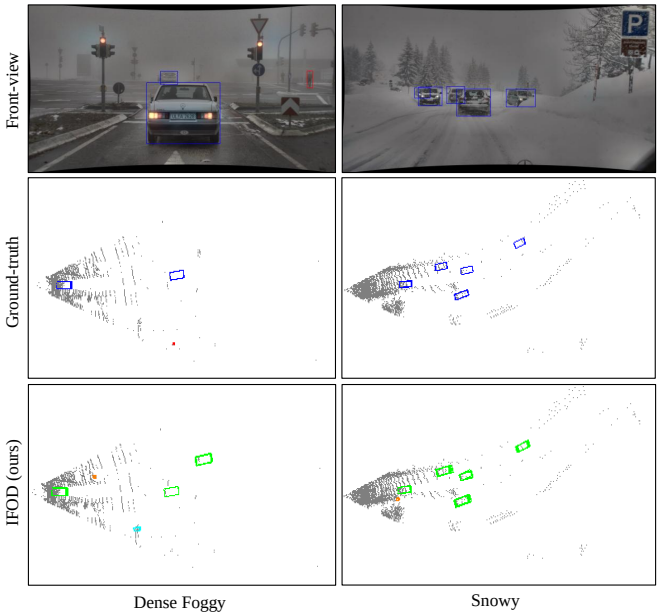

Fig. 8. The visualization of the failure case in IFOD. Ground-truth has classes Car and $Pedestrian$. The predicted object by the method are Car, $Pedestrian$ and $Cyclist$.

may introduce redundancy. Third, the chosen 1 FSC + 4 AGE configuration achieves the lowest runtime (101 ms) while preserving competitive accuracy, embodying the lightweight philosophy of YOLOX.

F. Visualization Analysis

Furthermore, as shown in Figure 5, we visualized the detection results and conducted a comparison with the ground-truth. For the convenience of observation and class measurement, we provide a front-view image in Figure 6, which corresponds to the LiDAR scan. All methods yield accurate measurements consistent with the ground truth. However, PillarNet, VoxelSet, VoxelNeXt, and DSVT-Voxel predict the presence of cars, cyclists, and pedestrians at additional locations. PV-RCNN and LION also predict cars and cyclists. Specifically, although the ground truth's real class is Car, the prediction method incorrectly identifies the object as $Cyclist$. This difference may be due to the model not fully learning the feature differences between $Cyclist$ and Car during the training process, or the presence of different distributions in the test data compared to the training data. Only Voxel-Mamba and our method accurately predict a single class, although Voxel-Mamba also exhibits a single incorrect prediction.

The Figure 6 shows the detection results of the clean scene corresponding to 5 in the KITTI dataset. It is observed that due to enhancements in data quality, some methods exhibit improved performance, resulting in detections that extend beyond the labelled annotations, similar to the observations in Figure 5. Upon magnifying the front-view images, it becomes evident that in addition to the labelled vehicles, there are also ambiguous object within the shadowed regions. LiDAR scans indeed capture reflective points; however, due of the long distances involved, the quantity of acquired scan data is minimal, which can easily lead to erroneous predictions by these methods. Consequently, many methods exhibit high error rates. IFOD accounts for associations with pseudo-image features, effectively predicting unlabeled information and filtering out object with higher error rates. This is the advantage of AGE, which not only maintains accurate judgments under foggy conditions but also enhances the performance of detecting object in normal environments.

The Figure 7 shows the detection performance of IFOD under various weather conditions, including clear, LF, DF and Snowy scenarios. Due to space limitations, the visualization results of other comparison methods can be found in the supplementary materials. The comparison shows that IFOD still achieves the best detection results.

G. Limitations

As shown in Figure 8, IFOD still encounters failure cases under DF and Snowy. Specifically, in DF scenes, the model occasionally misjudges object distances, resulting in inaccurate bounding box localization. In addition, small object-like noise caused by snowflakes is sometimes misclassified as valid object, leading to false positives. These observations suggest that the current fog-aware guidance may be misled by non-structural noise or insufficient depth cues in extreme conditions. In future work, we plan to incorporate image-based semantic features to assist 3D detection, especially for small object such as pedestrians and cyclists. By compensating for missing fine-grained information through semantic priors, the model is expected to improve its accuracy and robustness in detecting small occluded object under adverse conditions.

V. CONCLUSION

In this paper, we proposed IFOD, an physically interpretable 3DOD framework for foggy LiDAR point cloud. Guided by the atmospheric scattering model, IFOD integrates the FSC and AGE modules to enhance robustness under foggy. We further constructed FogKITTI, a physically simulated dataset derived from an optical system model, providing realistic fog conditions for evaluation. The performance of IFOD suggests that its architecture could potentially be adapted for broader perception tasks in terrestrial and maritime navigation where fog is a primary concern. However, weaknesses remain in extreme conditions, such as distance misjudgment in dense fog and false positives induced by snowflake noise. Future work will explore multi-modal image fusion and integration of 4D millimeter wave radar to improve the detection of small occluded objects under extreme visibility limitations.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (62472137), CCF-NetEase ThunderFire Innovation Research Funding (CCF-Netease 202611), Open Foundation of Yunnan Key Laboratory of Software Engineering (2023SE103), and the Key Scientific Research Platform and Project of Guangdong Provincial Education Department (2022ZDZX1038). Zhao Zhang is the corresponding author of this paper.

REFERENCES

[1] C. Chen, A. Jin, Z. Wang, Y. Zheng, B. Yang, J. Zhou, Y. Xu, and Z. Tu, "Sgsr-net: Structure semantics guided lidar super-resolution network for indoor lidar slam," *IEEE Trans. Multimedia.*, vol. 26, p. 1842–1854, 2024.

[2] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *CVPR*, 2020, pp. 11 682–11 692.

[3] S. Wang, K. Lu, J. Xue, and Y. Zhao, "Da-net: Density-aware 3d object detection network for point clouds," *IEEE Trans. Multimedia.*, vol. 27, pp. 665–678, 2025.

[4] X. Huang, Z. Xu, H. Wu, J. Wang, Q. Xia, Y. Xia, J. Li, K. Gao, C. Wen, and C. Wang, "L4dr: Lidar-4dradar fusion for weather-robust 3d object detection," in *AAAI*, vol. 39, no. 4, 2025, pp. 3806–3814.

[5] L. Zhuang, Y. Yao, and T. Zhang, "Rcoc-distill: Boosting 4d radar-camera online calibration with knowledge distillation from lidar features," *IEEE Trans. Multimedia.*, 2026.

[6] H. Wu, C. Wen, S. Shi, X. Li, and C. Wang, "Virtual sparse convolution for multimodal 3d object detection," in *CVPR*, 2023, pp. 21 653–21 662.

[7] Z. Liu, J. Cheng, J. Fan, S. Lin, Y. Wang, and X. Zhao, "Multi-modal fusion based on depth adaptive mechanism for 3d object detection," *IEEE Trans. Multimedia.*, vol. 27, pp. 707–717, 2025.

[8] L. Zhao, H. Zhou, X. Zhu, X. Song, H. Li, and W. Tao, "Lif-seg: Lidar and camera image fusion for 3d lidar semantic segmentation," *IEEE Trans. Multimedia.*, vol. 26, pp. 1158–1168, 2024.

[9] A. A. Khan, J. Shao, Y. Rao, L. She, and H. T. Shen, "Lrdnet: Lightweight lidar aided cascaded feature pools for free road space detection," *IEEE Trans. Multimedia.*, vol. 27, p. 652–664, 2025.

[10] 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 Trans. Multimedia.*, vol. 26, pp. 2795–2808, 2024.

[11] J. Wang, F. Li, X. Zhang, and H. Sun, "Adversarial obstacle generation against lidar-based 3d object detection," *IEEE Trans. Multimedia.*, vol. 26, pp. 2686–2699, 2024.

[12] Y. Gong, N. Wang, J. Lu, X. Zhang, Y. Gao, J. Zhao, Z. Huang, H. Bai, N. Zeng, N. Su *et al.*, "Progressive bird's eye view perception for safety-critical autonomous driving: A comprehensive survey," *arXiv preprint arXiv:2508.07560*, 2025.

[13] N. Wang, D. Shang, Y. Gong, X. Hu, Z. Song, L. Yang, Y. Huang, X. Wang, and J. Lu, "Collaborative perception datasets for autonomous driving: A review," *arXiv preprint arXiv:2504.12696*, 2025.

[14] Y. Dong, C. Kang, J. Zhang, Z. Zhu, Y. Wang, X. Yang, H. Su, X. Wei, and J. Zhu, "Benchmarking robustness of 3d object detection to common corruptions," in *CVPR*, 2023, pp. 1022–1032.

[15] 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 *CVPR*, 2020, pp. 11 621–11 631.

[16] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," *Int. J. Comput. Vision*, vol. 48, pp. 233–254, 2002.

[17] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*. IEEE, 2012, pp. 3354–3361.

[18] Z. Song, L. Liu, F. Jia, Y. Luo, C. Jia, G. Zhang, L. Yang, and L. Wang, "Robustness-aware 3d object detection in autonomous driving: A review and outlook," *IEEE Trans. Intell. Transp. Syst.*, 2024.

[19] R. Qian, X. Lai, and X. Li, "3d object detection for autonomous driving: A survey," *Pattern Recognit.*, vol. 130, p. 108796, 2022.

[20] J. Mao, S. Shi, X. Wang, and H. Li, "3d object detection for autonomous driving: A comprehensive survey," *Int. J. Comput. Vision*, vol. 131, no. 8, pp. 1909–1963, 2023.

[21] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *CVPR*, 2018, pp. 4490–4499.

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

[23] Z. Liu, J. Hou, X. Wang, X. Ye, J. Wang, H. Zhao, and X. Bai, "Lion: Linear group rnn for 3d object detection in point clouds," in *NeurIPS*, vol. 37, 2024, pp. 13 601–13 626.

[24] Z. Wang, H. Zhao, J. Peng, L. Yao, and K. Zhao, "Odcr: Orthogonal decoupling contrastive regularization for unpaired image dehazing," *CVPR*, pp. 25 479–25 489, 2024.

[25] Y. Yang, C.-L. Guo, and X. Guo, "Depth-aware unpaired video dehazing," *IEEE Trans. Image Process.*, vol. 33, pp. 2388–2403, 2024.

[26] J. Chen, W. Ren, H. Zhao, Q. Xia, and G. Yang, "You only need clear images: Self-supervised single image dehazing," *IEEE Trans. Multimedia.*, pp. 1–16, 2025.

[27] A. Seppänen, R. Ojala, and K. Tammi, "4denoisenet: Adverse weather denoising from adjacent point clouds," *IEEE Rob. Autom. Lett.*, vol. 8, no. 1, pp. 456–463, 2022.

[28] Y. Chen, Y. Li, X. Zhang, J. Sun, and J. Jia, "Focal sparse convolutional networks for 3d object detection," in *CVPR*, 2022, pp. 5428–5437.

[29] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *Advances in arXiv preprint arXiv:2107.08430*, 2021.

[30] Z. Song, G. Zhang, L. Liu, L. Yang, S. Xu, C. Jia, F. Jia, and L. Wang, "Robofusion: Towards robust multi-modal 3d object detection via sam," in *IJCAI*, 2024, pp. 1272–1280.

[31] Z. Song, C. Jia, L. Yang, H. Wei, and L. Liu, "Graphalign++: An accurate feature alignment by graph matching for multi-modal 3d object detection," *TCSVT*, vol. 34, no. 4, pp. 2619–2632, 2023.

[32] Z. Yu, B. Qiu, and A. W. Khong, "Vikienet: Towards efficient 3d object detection with virtual key instance enhanced network," in *CVPR*, 2025, pp. 11 844–11 853.

[33] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *ICCV*, 2017, pp. 4770–4778.

[34] G. Shi, R. Li, and C. Ma, "Pillarnet: Real-time and high-performance pillar-based 3d object detection," in *ECCV*, 2022, pp. 35–52.

[35] C. He, R. Li, S. Li, and L. Zhang, "Voxel set transformer: A set-to-set approach to 3d object detection from point clouds," in *CVPR*, 2022, pp. 8417–8427.

[36] 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 *CVPR*, 2020, pp. 10 529–10 538.

[37] Y. Chen, J. Liu, X. Zhang, X. Qi, and J. Jia, "Voxelnext: Fully sparse voxelnet for 3d object detection and tracking," in *CVPR*, 2023, pp. 21 674–21 683.

[38] 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 *CVPR*, 2023, pp. 13 520–13 529.

[39] Z. Liu, J. Hou, X. Ye, T. Wang, J. Wang, and X. Bai, "Seed: A simple and effective 3d detr in point clouds," in *ECCV*, 2024, pp. 110–126.

[40] G. Zhang, L. Fan, C. He, Z. Lei, Z.-X. ZHANG, and L. Zhang, "Voxel mamba: Group-free state space models for point cloud based 3d object detection," in *NeurIPS*, vol. 37, 2024, pp. 81 489–815 093.

[41] S. Liu, M. Cui, B. Li, Q. Liang, T. Hong, K. Huang, and Y. Shan, "Fshnet: Fully sparse hybrid network for 3d object detection," in *CVPR*, 2025, pp. 8900–8909.

[42] Z. Song, L. Yang, S. Xu, L. Liu, D. Xu, C. Jia, F. Jia, and L. Wang, "Graphbev: Towards robust bev feature alignment for multi-modal 3d object detection," in *ECCV*, 2024, pp. 347–366.

[43] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel r-cnn: Towards high performance voxel-based 3d object detection," in *AAAI*, vol. 35, no. 2, 2021, pp. 1201–1209.

[44] M. Berman, A. R. Triki, and M. B. Blaschko, "The lovasz-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *CVPR*, 2018, pp. 4413–4421.

[45] O. D. Team, "Openpcdet: An open-source toolbox for 3d object detection from point clouds," https://github.com/open-mmlab/OpenPCDet, 2020.

[46] W. Ye, Q. Xia, H. Wu, Z. Dong, R. Zhong, C. Wang, and C. Wen, "Fade3d: Fast and deployable 3d object detection for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 9, pp. 12 934–12 946, 2025.

[47] T. Prasanth, R. P. Padhy, and B. Sivaselvan, "Msfasa-3dnet: Multi-scale feature aggregation and spatial attention for 3d object detection," *IEEE Trans. Artif. Intell.*, pp. 1–13, 2026.

[48] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *ICCV*, 2023, pp. 4015–4026.