FESFN: Frequency-Enhanced Supervised Fusion Networks for 3D object detection

1st Shengfa Pei
School of Artificial Intelligence and Computer Science
North China University of Technology
Beijing, China
pizzafa@163.com

2nd Yuanyao Lu
School of Artificial Intelligence and Computer Science
North China University of Technology
Beijing, China
luyy@ncut.edu.cn

3rd Kamoliddin Shukurov
Artificial Intelligence Department
Tashkent University of Information Technology named after Muhammad al-Khwarizmi
Tashkent, Uzbekistan
keshukurov@gmail.com

***Abstract*—In recent years, BEV-based 3D object detection has seen substantial progress in the context of perception for autonomous driving. To tackle the issues of inaccurate depth estimation, redundant fusion strategies, and imprecise classification supervision in existing Lift, Splat, Shoot (LSS) methods, we propose a Frequency-Enhanced Supervised Fusion Network (FESFN) for LiDAR-camera fusion detection. Specifically, a Frequency-Supervised Fusion Module (FSFM) is introduced to incorporate frequency information from LiDAR point clouds to guide the image-based LSS process, thereby improving depth estimation accuracy. A Progressive Fusion Module (PFM) is then designed to enhance cross-modal feature collaboration and reduce redundancy by combining dense and sparse feature fusion strategies. Additionally, an Instance-Aware Classification Head (IACH) is proposed to supervise classification only on predicted instances, enhancing the precision of classification guidance. Experimental results on the nuScenes dataset demonstrate that FESFN achieves 73.6% mAP and 75.5% NDS, outperforming current state-of-the-art methods.**

***Index Terms*—Autonomous driving, 3D object detection, LiDAR-Camera fusion, Frequency Supervision**

I. Introduction

As a fundamental component of autonomous driving technology, the precision of 3D object detection directly determines the reliability of subsequent decision-making tasks. Current in-vehicle sensing systems still heavily rely on a single type of sensor, such as a camera or LiDAR [1], [2]. Cameras provide rich semantic information and high-resolution images, but remain vulnerable to environmental disturbances [3]; Variations in lighting and occlusions caused by rain or fog can significantly degrade image quality and detection performance. Although LiDAR offers more accurate structural and depth information, its point clouds are often sparse and lack semantic richness. To improve detection accuracy and improve the robustness of autonomous driving systems, multimodal fusion methods that leverage the complementary strengths of both sensors [4], [5] have become a key research focus in both academia and industry.

Existing multimodal 3D object detection methods can be categorized into three types: data-based, proposal-based, and BEV-based fusion. Data-based fusion directly combines raw image and point cloud data, as in PointPainting [6], but often suffers from high computational cost due to redundant operations. Proposal-based fusion merges detection results or proposals from each modality, e.g., CLOCs [7], but cannot fully exploit cross-modal features, limiting performance. In contrast, BEV-based fusion integrates LiDAR and camera features in a unified Bird's Eye View (BEV) space [4], [8], balancing efficiency and accuracy. While direct alignment methods require camera intrinsics/extrinsics and are sensitive to calibration [9], BEV-based fusion has gained popularity for its robustness and effectiveness.

Despite their promise, BEV-based methods face three key challenges: reliance on accurate depth estimation, limitations in fusion strategies, and inefficiencies in loss functions. First, transforming multi-view image features into BEV space requires depth estimation [10], [11], which is easily compromised by poor image quality, lighting, or complex scenes, degrading performance [12]. Second, current LiDAR-image fusion methods typically concatenate features and apply attention modules [4], [13], potentially over-relying on one modality or introducing redundancy, which increases overfitting risk and ignores modality-specific cues like frequency information [14]. Lastly, conventional loss functions compute over all annotated objects, including irrelevant or low-quality ones, adding computational overhead and reducing generalization.

To deal with the challenges, we propose a 3D detection model called FESFN. By applying Discrete Cosine Transform (DCT) [15] to the BEV features, we extract both high- and low-frequency components, and use LiDAR-derived frequency features to supervise the image-based LSS process. This helps improve depth accuracy and embeds frequency-aware information into BEV features. Furthermore, we design the PFM, which separately performs dense and sparse feature fusion to fully exploit complementary modality features while reducing redundancy. Finally, we introduce IACH that computes loss only on valid instances, improving detection stability and accuracy.

Our contributions can be summarized as follows:

1) We propose FESFN for 3D object detection, which leverages LiDAR BEV frequency features to supervise the generation of image BEV features, leading to more accurate depth estimation and deeper feature extraction by integrating frequency information.

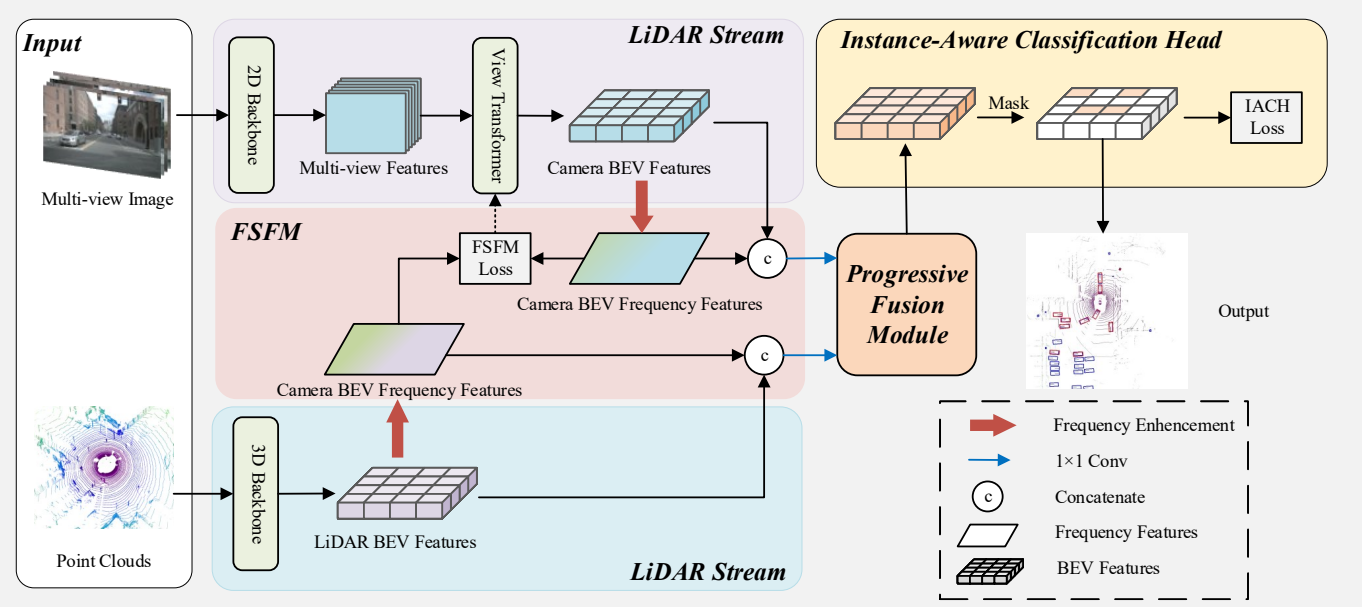

Fig. 1. Overview of the proposed multi-modal 3D object detection framework integrating LiDAR and camera streams. Feature extraction is first performed using 2D and 3D backbones, followed by BEV transformation via LSS. During this process, FSFM guides depth information supervision and embeds frequency-aware cues into BEV features. PFM then integrates LiDAR and image features at multiple stages to enhance semantic consistency. Finally, IACH processes the fused features to output object masks and classes. The framework emphasizes effective cross-modal interaction, hierarchical fusion, and adaptive supervision for robust detection in complex scenes.

2) We design the PFM that effectively improves fusion efficiency and fully utilizes multimodal cues, while avoiding over-reliance on a single modality or overfitting issues.

3) To the best of our knowledge, this is the first 3D object detection framework that introduces frequency information to supervise the LSS process and fuses frequency-aware features across modalities.

II. Related Work

A. Single-Modality 3D Detection

Some methods use only one sensor [2], such as a camera or LiDAR. Early camera-based approaches follow 2D detection frameworks: Mono3D [2] detects 2D boxes and generates 3D proposals, while MonoDepth [16] predicts depth maps from a single image. These methods often yield inaccurate depth due to limited visual information.

Multi-camera systems capture more views, supported by datasets like nuScenes [17], KITTI [18], and Waymo [19]. Methods such as BEVFormer [20] and DETR3D [21] project BEV queries onto images and leverage attention or temporal features. Many follow LSS [10], transforming multi-view features into BEV space. These methods are more accurate than monocular ones but still lack reliable depth and are sensitive to image quality.

LiDAR-based methods provide accurate 3D geometry and handle adverse weather. Point-based methods, like PointNet [22], use raw point clouds but are often slow. Voxel-based methods [23], such as VoxelNet [24], convert points to 3D grids for convolution. PointPillars [25] simplifies this with 2D CNNs by collapsing the height axis. These methods maintain spatial accuracy but require high computation.

B. Multi-Modal 3D Detection

Recent studies have widely adopted camera-LiDAR fusion to enhance 3D detection [6], [26]. These approaches combine visual and geometric information and are generally categorized as point-based, proposal-based, or BEV-based. The first two typically map image features onto point clouds or candidate proposals to enrich semantics, but their interaction remains limited, restricting the exploitation of complementary advantages.

The mainstream trend unifies image and point cloud features in BEV space. Methods such as LSS project multi-view images into BEV, while BEVFormer [20] and DETR3D [21] further incorporate Transformers and temporal modeling to enhance global representation and robustness. BEVFusion [4] and TransFusion [5] align LiDAR and image BEV features using cross-attention, and recent works like SparseFusion [27] and ReliFusion [28] explore efficient alignment and long-term temporal modeling, achieving stronger performance in large-scale dynamic environments.

Frequency-domain methods have proven effective in image restoration and depth estimation, enhancing fine structures and maintaining depth consistency, e.g., Dcdepth [29] and VistaDepth [30]. However, frequency information is rarely explored in multimodal 3D detection. This work introduces the FSFM module, which incorporates LiDAR frequency features

to guide image-to-BEV transformation, supplementing depth cues during fusion and improving detection accuracy.

III. Method

A. Network Architecture

An overview of FESFN is provided in this section, with its structure shown in Fig. 1. We first use an image encoder and a point cloud encoder to extract features. Next, the LSS module transforms image features into the BEV space. During the BEV feature transformation, FSFM provides auxiliary supervision by introducing frequency information, which enhances the depth modeling capability and subsequently embeds fine-grained spatial cues into the BEV features generated from images.

The BEV features enriched with frequency cues are then fed into the PFM for multi-stage fusion. This process progressively integrates BEV features from both LiDAR and image modalities. The hierarchical design promotes semantic consistency and strengthens the complementary information exchange across modalities. Finally, IACH focuses detection and classification only on instance-relevant regions, effectively avoiding interference from irrelevant areas. In this way, the framework achieves accurate 3D object detection through LiDAR-camera fusion.

B. Frequency-Supervised Fusion Module

Frequency information plays a crucial role in capturing different aspects of scene representation. Low-frequency components primarily encode structural characteristics such as object positions in the BEV, while high-frequency components preserve edges and textures that refine object details. Building on this insight, we propose the FSFM, which leverages BEV frequency features extracted from point clouds to guide the image-based LSS process. By doing so, it enables more accurate depth estimation, enhances the quality of image BEV features, and incorporates complementary frequency information to improve overall detection accuracy.

As shown in Fig. 2, the FSFM consists of two submodules: the Frequency Supervision (FS) module and the Frequency Fusion (FF) module. In the FS module, the input BEV feature map $F \in \mathbb{R}^{B\times C\times H\times W}$ is first split into multiple parts, denoted as $[F_0, F_1, \cdots, F_{n-1}]$, where B is the batch size, C is the number of channels, and H and W are the height and width of the feature map, respectively. Each part F_k is then processed using a 2D Discrete Cosine Transform (2D-DCT) as defined below:

$$\begin{aligned} F'_k &= \mathrm{2DDCT}(F_k) \\ &= \sum_{w=0}^{W-1}\sum_{h=0}^{H-1} F_k \cdot \cos\left(\frac{\pi(2h+1)u}{2H}\right) \\ &\quad \cdot \cos\left(\frac{\pi(2w+1)v}{2W}\right) \end{aligned} \tag{1}$$

where $h \in \{0, 1, \cdots, H-1\}$, $w \in \{0, 1, \cdots, W-1\}$, and $F'_k \in \mathbb{R}^{C'}$ represents the DCT-transformed vector for each split. Next, we concatenate $F'_1, F'_2, \cdots, F'_{n-1}$ along the channel dimension to obtain the full frequency feature representation:

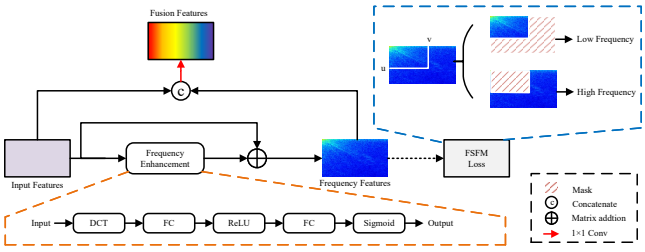

Fig. 2. The architecture of the FSFM. The FS module transforms features into the frequency domain and computes losses on high- and low-frequency components, while the FF module fuses frequency-enhanced an d original features to generate more discriminative representations.

$$F' = \mathrm{cat}[F'_1, F'_2, \cdots, F'_{n-1}] \tag{2}$$

where $\mathrm{cat}(\cdot)$ denotes concatenation along the channels, and F' represents the final frequency feature after DCT transformation.

To better capture high-level semantic details within frequency features, we pass F' through two fully connected layers with non-linear activations to generate the frequency attention weights:

$$W_{\text{weight}} = \sigma\left(\mathrm{FC}\left(\delta\left(\mathrm{FC}(F')\right)\right)\right) \tag{3}$$

where W_{weight} represents the learned frequency weights, $\sigma(\cdot)$ is the Sigmoid function, $\delta(\cdot)$ is the ReLU function, and FC denotes a fully connected layer.

We apply these weights to the original BEV feature map F via element-wise multiplication to obtain the final frequency-weighted BEV feature:

$$F'' = F \otimes W_{\text{weight}} \tag{4}$$

where F'' denotes the BEV feature modulated by frequency information.

Specifically, we feed the LiDAR BEV features $F_{\text{LIDAR bev}}$ and image BEV features $F_{\text{Image bev}}$ into the FS module to obtain their respective frequency-enhanced features $F''_{\text{LIDAR bev}}$ and $F''_{\text{Image bev}}$. Since we aim to use these frequency features to supervise depth estimation in the LSS process, low-frequency components mainly encode structural information, we further split the features by frequency coordinates (u, v): the region within a threshold $(u, v) \le \tau$ is regarded as the low-frequency component F''_{low}, while the region outside is treated as the high-frequency component F''_{high}. The threshold τ and the weights W_{high} and W_{low} are empirically chosen to balance structural consistency and fine-grained detail alignment, and are kept fixed across all experiments. When computing the loss, a mask is applied to suppress the other part, and different weights are assigned to the two components, which is defined as:

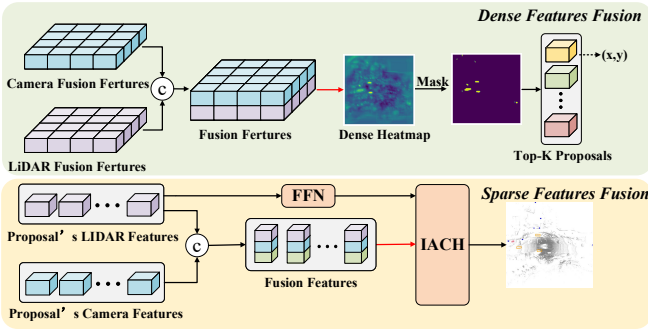

Fig. 3. The architecture of the PFM. The Dense Fusion stage generates proposals from a heatmap using fusion features, while the Sparse Fusion stage further refines proposal-specific features through FFN and IACH to obtain final discriminative representations.

$$\mathrm{MSE}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{5}$$

$$\begin{aligned} L_{\mathrm{FSFM}} = {} & W_{\mathrm{high}} \cdot \mathrm{MSE}(F''_{\mathrm{LIDAR\ high}}, F''_{\mathrm{Image\ high}}) \\ & + W_{\mathrm{low}} \cdot \mathrm{MSE}(F''_{\mathrm{LIDAR\ low}}, F''_{\mathrm{Image\ low}}) \end{aligned} \tag{6}$$

Here, $\mathrm{MSE}(\cdot)$ is the Mean Squared Error function, L_{FSFM} is the loss of the FS module, and $W_{\mathrm{high}}, W_{\mathrm{low}}$ are the weights of the high-frequency and low-frequency components, respectively. By incorporating this loss into backpropagation, we achieve supervision of the LSS process using frequency features.

Ultimately, the frequency-aware features were fused with the original inputs, producing enriched BEV representations that carry enhanced information for downstream tasks.

C. Progressive Fusion Module

Existing BEV-based multimodal detectors typically fuse point cloud and image features in the BEV space and leverage attention mechanisms to facilitate cross-modal interactions before applying the fused features to downstream detection tasks. While these approaches effectively enhance multimodal collaboration, they often introduce redundant information, which can lead to overfitting and degrade detection accuracy. To address these limitations, we propose a PFM, which implements a progressive fusion strategy consisting of dense fusion and sparse fusion, as illustrated in Fig. 3.

In the Dense Feature Fusion stage, BEV features extracted from point clouds and images via the FSFM module are concatenated and subsequently processed through a residual convolutional block. This block explicitly captures multi-scale spatial dependencies and rich semantic representations, generating a dense heatmap $H \in \mathbb{R}^{C\times X\times Y}$, where C denotes the number of object classes and X, Y represent the spatial dimensions. Each spatial location corresponds to the likelihood of a specific class, and based on this dense heatmap, we select the top-K proposals with the highest confidence scores, thereby focusing computation on the most promising object regions for subsequent refinement.

In the Sparse Feature Fusion stage, point cloud features corresponding to the coordinates of these top-K proposals are first extracted. Each feature is refined through a feedforward network, implemented as the prediction heads in the model, which predict key attributes including object center, height, size, and orientation. Subsequently, the corresponding image and point-cloud features for each proposal are fused using 1D convolution followed by normalization, which facilitates deeper semantic integration while maintaining feature balance across modalities. The resulting refined representations are then utilized to predict object classes with high accuracy.

D. Instance-Aware Classification Head

Inspired by previous works, we propose an instance-aware classification head to compute the loss. Traditional approaches typically calculate the loss based on all annotated targets, which can introduce a large number of irrelevant or unnecessary instances, thereby affecting the model's training efficiency and generalization ability. To mitigate this issue, our method computes the loss only for the predicted valid instances, enhancing both efficiency and accuracy.

As shown in Fig. 1 and Fig. 3,we first compute the centroid and bounding box of each valid instance, and use this information to extract features from the corresponding regions to generate a heatmap. By introducing Gaussian Focal Loss, we compute the heatmap loss L_{heatmap}, which helps the model to focus more accurately on the regions that contain actual instances. During this process, all non-instance regions are masked to prevent interference from irrelevant areas during loss computation.

Next, the generated heatmap is reshaped into the format $[\mathrm{batch\_size} \times \mathrm{num\_target}, \mathrm{num\_cls}]$, and focal loss is applied to quantify the difference between predictions and ground-truth labels, resulting in the classification loss L_{cls}.

Finally, the total loss for the instance-aware classification module is obtained by summing the two components:

$$L_{\mathrm{aux}} = L_{\mathrm{heatmap}} + L_{\mathrm{cls}} \tag{7}$$

Through this module, the model receives more precise supervision, leading to improved detection accuracy and enhanced training stability.

E. Loss function

Most components of our architecture are inherited from BEVFusion, with modifications made only to critical parts. Accordingly, our model's loss function is based on the BEVFusion loss, with the addition of the FSM module loss and the instance-aware auxiliary classification module loss. The overall loss is as follows:

$$L_{\mathrm{total}} = L_{\mathrm{BEVFusion}} + \lambda_1 L_{\mathrm{FSFM}}(x, y) + \lambda_2 L_{\mathrm{aux}}(p, \hat{p}) \tag{8}$$

where λ_1 and λ_2 are the weights for each component of the loss, which can be set within the range of 0.2-0.5. $L_{\mathrm{BEVFusion}}$ denotes the original loss function used in BEVFusion, L_{FSFM}

TABLE I
COMPARISON WITH SOTA METHODS ON THE NUSCENES TEST SET. C.V.: CONSTRUCTION VEHICLE, PED.: PEDESTRIAN, T.C.: TRAFFIC CONE. BOLD VALUES INDICATE THE BEST PERFORMANCE, AND UNDERLINED VALUES INDICATE THE SECOND-BEST PERFORMANCE.

Method	Modality	mAP	NDS	Car	Truck	Bus	Trailer	C.V.	Ped.	Motor.	Bike	T.C.	Barrier
Link [31]	L	66.3	71.0	86.1	55.7	65.7	62.1	30.9	85.8	73.5	47.5	80.4	75.5
SAFDNet [32]	L	68.3	72.3	87.3	57.3	68.0	63.7	37.3	89.0	71.1	44.8	84.9	79.5
LION [33]	L	69.8	73.9	87.2	61.1	68.9	65.0	36.3	90.0	74.0	49.2	87.3	79.5
SEGT [34]	L	70.1	73.9	87.3	60.7	69.3	66.7	37.8	88.5	77.2	49.7	85.7	77.8
TransFusion [5]	L+C	68.9	71.7	87.1	60.0	68.3	60.8	33.1	88.4	73.6	52.9	86.7	78.1
BEVFusion [4]	L+C	70.2	72.9	88.6	60.1	69.8	63.8	39.3	89.2	74.1	51.0	86.5	80.0
DeepInteraction-base [35]	L+C	70.8	73.4	87.9	60.2	70.8	63.8	37.5	90.3	75.4	54.5	87.0	80.4
UniTR [36]	L+C	70.9	74.5	87.9	60.2	72.2	65.1	39.2	89.4	75.8	52.2	**89.7**	76.8
ContrastAlign [37]	L+C	71.5	73.7	**90.1**	**67.1**	**77.2**	46.4	33.1	90.6	81.2	**68.4**	83.5	77.5
SparseFusion [13]	L+C	72.0	73.8	88.0	60.2	72.0	64.9	38.7	**90.9**	78.5	59.8	87.9	79.2
FusionFormer [38]	L+C	72.6	75.1	89.8	62.0	69.8	63.0	40.3	90.4	**83.0**	64.6	86.0	77.4
FESFN	L+C	**73.6**	**75.5**	89.7	64.3	75.7	**65.8**	37.7	88.9	82.7	59.6	88.6	**82.6**

represents the loss from the FSFM, and L_{aux} corresponds to the loss from the IACH. In this formulation, x and y represent the BEV features from the LiDAR and image modalities, while p and $\hat{p}$ denote the predicted and ground-truth class labels.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

The nuScenes [17] dataset is a large-scale dataset specifically designed to support autonomous driving. It contains multimodal data collected from multiple sensors. In total, it consists of 1,000 recorded scenes, with each scene capturing 20 seconds of data, including LiDAR point clouds, images, GPS, and other sensor information. This dataset includes 23 common categories, such as pedestrians, cars, trucks, buses, and bicycles. These objects are accurately annotated with spatial and dynamic information. The main evaluation metrics of nuScenes are mean Average Precision (mAP), mean true positive (mTP), and nuScenes Detection Score (NDS). NDS is computed from mTP and mAP according to a specific formula. The evaluation also includes additional metrics such as average translation error (ATE), average scale error (ASE), average orientation error (AOE), and average velocity error (AVE), to comprehensively assess the model's performance from multiple perspectives.

B. Implementation Details

Our model is primarily implemented and deployed based on MMDetection3D [39], an open-source codebase built on PyTorch. For the LiDAR branch, we adopt SECOND [40] as the backbone network and equip it with an FPN [41] neck to generate multi-scale feature maps, enhancing the model's ability to detect objects of varying sizes. For the camera branch, we use ResNet-50 as the backbone, also combined with an FPN for multi-scale feature extraction.We define the point cloud range as follows: the X-axis and Y-axis in $[-54.0\,\text{m}, 54.0\,\text{m}]$, and the Z-axis in $[-3.0\,\text{m}, 5.0\,\text{m}]$. The voxel size is set to $0.075\,\text{m} \times 0.075\,\text{m} \times 0.2\,\text{m}$.In terms of data augmentation, we apply random scaling in the range $[-0.06, 0.44]$ and random rotation within $[-5.4^\circ, 5.4^\circ]$ to the input images, simulating different viewing angles and improving the robustness of the model. All experiments are conducted using four NVIDIA 3090 GPUs with a batch size of 12. Unlike many existing methods that require stage-wise training, our model can be trained end-to-end in just 20 epochs.

TABLE II
RESULTS ON THE WAYMO OPEN DATASET

Method	ALL	Vehicle	Pedestrian	Cyclist
DeepFusion [42]	81.89	83.25	84.63	77.81
BEVFusion [8]	82.72	84.97	84.72	78.49
LoGoNet [43]	83.13	86.51	**86.84**	76.06
FESFN	**83.84**	**86.86**	86.03	**78.64**

C. Main Results

We evaluate our proposed FESFN against various state-of-the-art (SOTA) fusion-based detection methods on the nuScenes test benchmark, as shown in Table I. The results demonstrate that FESFN achieves the highest overall performance, reaching the best NDS and a competitive mAP. Notably, FESFN achieves clear accuracy gains for large-scale objects such as trailers and trucks, as well as small-scale objects like barriers, indicating its stronger ability to detect objects across different scales and depths. This advantage primarily stems from the proposed FSFM module, which introduces frequency-aware supervision for more precise depth estimation, and the carefully designed progressive fusion strategy that facilitates efficient cross-modal feature interaction.

We also conducted evaluations on the Waymo Open Dataset. As shown in Table II, FESFN achieves an mAP of 83.84, which is 0.71% higher than the current SOTA model LoGONet, demonstrating the strong robustness of our method.

D. Ablation Study

In this section, we perform ablation studies on the nuScenes validation set to assess the contribution of each component in FESFN to the overall detection performance.

1) FESFN: As shown in Table III, we adopt BEVFusion-MIT as the baseline method. Firstly, after introducing the FSFM module, the mAP and NDS increase by 1.9% and 1.3%, respectively. This indicates that supervising the depth

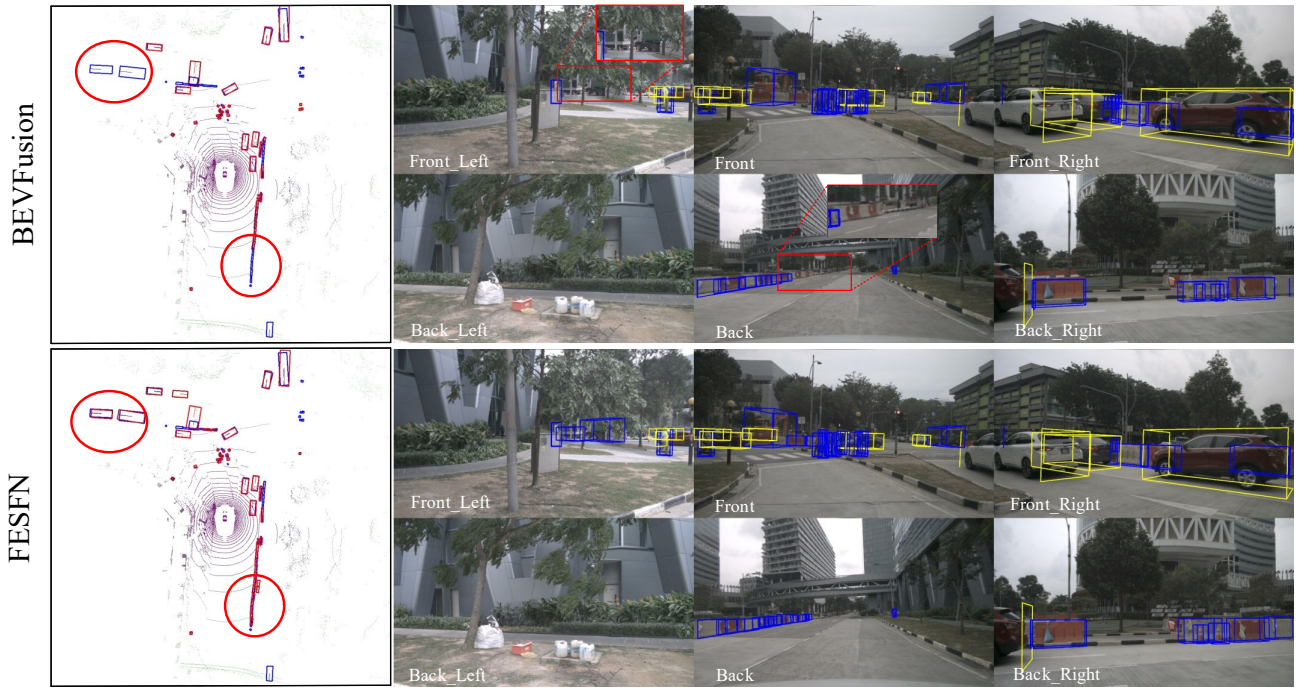

Fig. 4. Visualization of BEVFusion and FESFN. The blue boxes represent the Ground Truths, and the red boxes denote the predictions.The red circles and boxes show the detection ability of FESFN for small and occluded objects

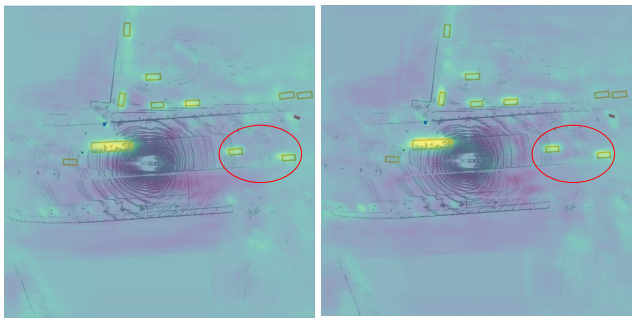

Fig. 5. The heatmap is overlaid with the ground truth, and the red-circled areas indicate that FSFM produces more accurate localization.

information and effectively integrating frequency information through the FSFM module significantly improves overall detection performance. With the further addition of the PFM module, mAP and NDS are improved by 1.0% and 0.8%, demonstrating its notable effectiveness in optimizing multi-modal feature representation. Finally, when the IACH module is incorporated, both mAP and NDS achieve an additional 0.5% improvement. This validates the effectiveness of the auxiliary classification module in enhancing instance localization accuracy and suppressing potential false positives.

TABLE III
ABLATION STUDIES FOR FESFN

ID	Baseline	FSFM	PFM	IACH	mAP	NDS
(a)	✓				70.2	72.9
(b)	✓	✓			72.1	74.2
(c)	✓	✓	✓		73.1	75.0
(d)	✓	✓	✓	✓	**73.6**	**75.5**

2) FSFM: This module can be further divided into two components: FS and FF. To analyze the impact of depth information and frequency information on detection results, we conducted ablation experiments on these two components individually. As shown in Table IV, we added the FS and FF modules separately into the baseline model. The results indicate that the FS module contributes a 0.9% improvement in mAP, while the FF module contributes a 1.3% increase. This demonstrates that frequency information can effectively provide richer detail. Moreover, high-frequency information in the BEV feature map—such as edges and contours—can be utilized for more accurate depth estimation. As shown in Fig.5, incorporating the FSFM module results in more precise localization of potential targets in the subsequent heatmap compared to the model without FSFM, further confirming the effectiveness of FSFM for depth estimation. When both modules are integrated into the model simultaneously, mAP significantly improves by 2.4%, fully demonstrating the strong complementarity between the two modules in processing BEV

TABLE IV
ABLATION STUDY OF FS AND FF MODULES

ID	FS	FF	mAP	NDS
(a)			70.2	72.9
(b)	✓		71.1	73.3
(c)		✓	71.5	73.7
(d)	✓	✓	**72.1**	**74.2**

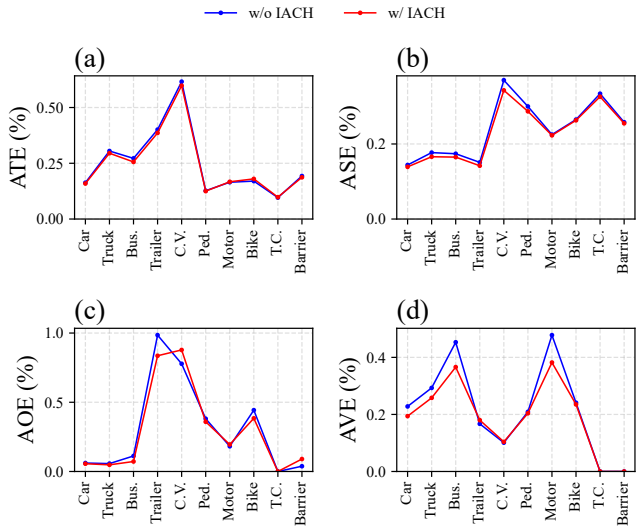

Fig. 6. Performance comparison of networks with and without IACH across different object classes.

feature maps. Their combined usage leads to better overall performance.

3) PFM: We conducted a series of ablation experiments to evaluate the effectiveness of the proposed components. As shown in Table V, Experiment 1 represents the baseline model without any enhancement strategies. Experiment 2 introduces frequency-aware LiDAR BEV features fused with image BEV features. Experiment 3 applies the progressive fusion strategy directly on the original BEV features. Experiment 4 integrates the complete PFM module, which combines both frequency features and the progressive fusion design. The results demonstrate the effectiveness of the PFM module.

4) IACH: We designed two sets of comparative experiments. In Fig. 6, Experiment 1 is the baseline model without the IACH module, while Experiment 2 is the improved model with the IACH module integrated. By comparing the performance of the two models in terms of ATE, AOE, ASE, and AVE metrics, it is evident that the model equipped with the IACH module outperforms the baseline in all error measures. This demonstrates that the IACH module has a significant effect in enhancing target localization, size estimation, and orientation prediction.

TABLE V
ABLATION STUDY OF PFM

ID	mAP	NDS
1	70.2	72.9
2	70.6	73.2
3	70.9	73.4
4	71.4	73.7

E. Visualization Analysis

As shown in Fig. 4, our FESFN model yields qualitative results in the nuScenes validation set, demonstrating its strong detection capability.

F. Latency Analysis

We compare the proposed model with several baselines in terms of both mAP and inference latency to evaluate the trade-off between accuracy and speed.As shown in the Table VI, FESFN demonstrates a favorable balance between inference latency and mAP.Although the proposed design introduces additional computation, the latency remains within an acceptable range while achieving improved detection accuracy. Future work will explore lightweight variants for real-time deployment.

TABLE VI
LATENCY ANALYSIS FOR MODELS. LATENCY IS MEASURED ON A RTX3090 GPU FOR REFERENCE.

Models	mAP	Latency(ms)
BEVFusion [4]	70.2	119.2
UniTR [36]	70.9	**107.5**
SparseFusion [27]	72.0	228.3
FESFN	**73.6**	192.4

V. CONCLUSION

This paper proposes the FESFN network, introducing three key modules—FSFM, PFM, and IACH—that enhance LiDAR-camera fusion in 3D object detection. FSFM supervises the LSS process with high- and low-frequency information, improving depth modeling of image BEV features. PFM employs progressive dense-to-sparse fusion, enhancing multi-modal feature interactions. IACH refines classification loss via instance selection, boosting accuracy and robustness. Ablation studies and comparisons on mainstream datasets show that FESFN outperforms existing methods across multiple metrics, demonstrating strong generalization and practical potential. Future work will explore combining frequency modeling with temporal fusion for more accurate dynamic scene perception.

REFERENCES

[1] T. Wang, J. Pang, and D. Lin, "Monocular 3d object detection with depth from motion," in *European conference on computer vision*. Springer, 2022, pp. 386–403.

[2] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2147–2156.

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

[4] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," *arXiv preprint arXiv:2205.13542*, 2022.

[5] 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.

[6] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.

[7] S. Pang, D. Morris, and H. Radha, "Clocs: Camera-lidar object candidates fusion for 3d object detection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 386–10 393.

[8] 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.

[9] L. Zhou, G. Wu, Y. Zuo, X. Chen, and H. Hu, "A comprehensive review of vision-based 3d reconstruction methods," *Sensors*, vol. 24, no. 7, p. 2314, 2024.

[10] 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.

[11] 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.

[12] K. Wang, Z. Zhang, Z. Yan, X. Li, B. Xu, J. Li, and J. Yang, "Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 16 055–16 064.

[13] Z. Zhou and S. Tulsiani, "Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 12 588–12 597.

[14] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 783–792.

[15] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE transactions on Computers*, vol. 100, no. 1, pp. 90–93, 2006.

[16] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 270–279.

[17] 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.

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

[19] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[20] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[21] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *Conference on robot learning*. PMLR, 2022, pp. 180–191.

[22] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[23] S. Song and J. Xiao, "Deep sliding shapes for amodal 3d object detection in rgb-d images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 808–816.

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

[25] 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.

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

[27] Y. Xie, C. Xu, M.-J. Rakotosaona, P. Rim, F. Tombari, K. Keutzer, M. Tomizuka, and W. Zhan, "Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 591–17 602.

[28] R. Sadeghian, N. Hooshyaripour, C. Joslin, and W. Lee, "Reliability-driven lidar-camera fusion for robust 3d object detection," *arXiv preprint arXiv:2502.01856*, 2025.

[29] K. Wang, Z. Yan, J. Fan, W. Zhu, X. Li, J. Li, and J. Yang, "Dcdepth: Progressive monocular depth estimation in discrete cosine domain," *Advances in Neural Information Processing Systems*, vol. 37, pp. 64 629–64 648, 2024.

[30] M. Zhan, L. Zhang, X. Chu, and B. Wang, "Vistadepth: Frequency modulation with bias reweighting for enhanced long-range depth estimation," *arXiv preprint arXiv:2504.15095*, 2025.

[31] T. Lu, X. Ding, H. Liu, G. Wu, and L. Wang, "Link: Linear kernel for lidar-based 3d perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1105–1115.

[32] G. Zhang, J. Chen, G. Gao, J. Li, S. Liu, and X. Hu, "Safdnet: A simple and effective network for fully sparse 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 477–14 486.

[33] 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," *Advances in Neural Information Processing Systems*, vol. 37, pp. 13 601–13 626, 2024.

[34] C. Mei, H. He, Y. Liu, and Z. Guo, "Segt: A general spatial expansion group transformer for nuscenes lidar-based object detection task," *arXiv preprint arXiv:2412.09658*, 2024.

[35] Z. Yang, J. Chen, Z. Miao, W. Li, X. Zhu, and L. Zhang, "Deepinteraction: 3d object detection via modality interaction," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1992–2005, 2022.

[36] H. Wang, H. Tang, S. Shi, A. Li, Z. Li, B. Schiele, and L. Wang, "Unitr: A unified and efficient multi-modal transformer for bird's-eye-view representation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 6792–6802.

[37] Z. Song, H. Pan, F. Jia, Y. Zhang, L. Liu, L. Yang, S. Xu, P. Wu, C. Jia, Z. Zhang *et al.*, "Contrastalign: Toward robust bev feature alignment via contrastive learning for multi-modal 3d object detection," *arXiv preprint arXiv:2405.16873*, 2024.

[38] Y. Cai, W. Zhang, Y. Wu, and C. Jin, "Fusionformer: A concise unified feature fusion transformer for 3d pose estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 900–908.

[39] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

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

[41] 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 conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[42] Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le *et al.*, "Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 182–17 191.

[43] X. Li, T. Ma, Y. Hou, B. Shi, Y. Yang, Y. Liu, X. Wu, Q. Chen, Y. Li, Y. Qiao *et al.*, "Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 524–17 534.