FREQLIFT: BEV FUSION AND CROSS-FREQUENCY STATE SPACE MODELING FOR LIDAR-CAMERA 3D OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, significant progress has been made in BEV-based 3D object detection. However, processes such as voxelization, viewpoint transformation, and downsampling can still compromise local structural details. Camera features, when projected into the BEV space, are affected by depth estimation and projection errors, and direct fusion may introduce these errors into the LiDAR-camera BEV representation. To address these issues, this paper proposes FreqLift, a LiDAR-camera 3D detection framework that combines BEV fusion with cross-frequency state-space modeling. FreqLift first refines the camera BEV using CameraBEVRefiner and fuses it with the LiDAR BEV. Subsequently, WaveLift uses multi-level wavelet decomposition and Wavelet Cross-Gate to model high- and low-frequency interactions; Freq-Mamba then uses low-frequency structural differences and BEV position coordinates to modulate the state-space model, thereby enhancing the structural modeling capabilities of the fused BEV. On the nuScenes validation set, FreqLift achieves 71.2% mAP and 73.1% NDS, whereas its LiDAR-only variant, FreqLift-L, achieves 67.1% mAP and 71.0% NDS. An ablation shows that this +4.1 mAP gain comes from direct camera fusion (+1.4 mAP) and CameraBEVRefiner (+2.7 mAP).

1 INTRODUCTION

3D object detection aims to identify object categories and estimate their 3D spatial locations, serving as a crucial component of autonomous driving perception systems. LiDAR provides accurate geometric information, whereas cameras offer rich semantic and local visual details. Since their representations in Bird's-Eye-View (BEV) space differ, fusing them in BEV space yields richer feature information.

Traditional BEV fusion methods (Liu et al., 2023b) typically project image features into BEV space using LSS (Philion & Fidler, 2020) before fusing them with LiDAR BEV. However, depth estimation and view transformation can introduce projection errors into the camera BEV. Meanwhile, voxelization and BEV downsampling may degrade local spatial structures within the features (Zhang et al., 2024). Existing wavelet studies (Liu et al., 2018; Li et al., 2020) indicate that low-frequency components primarily preserve global structure, while high-frequency components capture local variations. A key challenge addressed in this paper is how to refine the camera BEV and jointly utilize information across different frequencies after fusion.

To this end, we propose FreqLift. This framework first employs a CameraBEVRefiner to predict and gate residual correction features, refining the camera BEV via a residual approach before fusing it with the LiDAR BEV. Subsequently, WaveLift performs multi-scale frequency decomposition and state-space modeling on the fused features. Our main contributions are as follows:

- We propose FreqLift, a BEV fusion framework for LiDAR-camera 3D detection that does not rely on an external, pre-trained camera-only 3D detector.
- We propose CameraBEVRefiner, in which an SDR Block separately processes smooth structures and detail residuals; a spatial residual gating map controls the injection inten-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sity of correction features at different BEV locations, thereby refining the camera BEV in a residual manner.

- We design WaveLift to model high- and low-frequency interactions using multi-layer 2D discrete wavelet transforms and a Wavelet Cross-Gate, while utilizing cross-scale low-frequency differences and BEV spatial coordinates within Freq-Mamba to modulate state-space parameters.

2 Related Works

2.1 3D Detection Method

Traditional LiDAR-based 3D detection methods can be broadly categorized into three types. Point-based methods, such as PointNet (Qi et al., 2017a) and PointNet++ (Qi et al., 2017b), extract features directly from raw point clouds. Subsequent methods like 3DSSD (Yang et al., 2020) and PointRCNN (Shi et al., 2019) further improved detection accuracy, though they often entail high computational costs. In contrast, voxel-based methods such as SECOND (Yan et al., 2018), VoxelNet (Zhou & Tuzel, 2018), and VoxelNeXt (Chen et al., 2023b) leverage sparse convolution to enhance feature extraction efficiency. Voxel-based methods have now become mainstream, giving rise to approaches like UniMamba (Jin et al., 2025), VoxelMamba (Zhang et al., 2024), and LION (Liu et al., 2024). However, the inherent sparsity of voxels can lead to the loss of certain object information. This issue is particularly pronounced in BEV frameworks, as information lost during the BEV stage is difficult to recover in subsequent processing steps.

Recently, LiDAR-centric fusion methods such as BEVDilation (Zhang et al., 2026a) have adopted LiDAR as the primary representation to mitigate the impact of noise from image depth estimation on the fusion process. In contrast, FreqLift employs CameraBEVRefiner to perform residual refinement on the camera BEV before the fusion stage, thereby supplementing the LiDAR BEV features.

2.2 Camera in 3D Detection

Traditional BEV fusion methods typically rely on a strong camera branch. For instance, BEVFusion (Liu et al., 2023b) builds its camera branch upon BEVDet (Huang et al., 2021) and incorporates a pooling operation. Camera features are projected via LSS (Philion & Fidler, 2020) and fused with LiDAR BEV, thereby supplementing the LiDAR branch with semantic and geometric depth information. Recently, methods such as MPCF (Gao et al., 2026) have utilized pseudo-points to achieve spatial alignment between camera and LiDAR data. However, the distribution of pseudo-points is dense and irregular—differing significantly from sparse point clouds—which can lead to the loss or redundancy of object features.

The aforementioned methods suggest that the performance of both traditional BEV fusion and pseudo-point strategies is closely tied to the quality of the camera branch. We argue, however, that effective BEV fusion depends on the reliability of the camera BEV representation rather than only on the strength of the camera detection branch. To address this, we propose CameraBEVRefiner to refine camera BEV features.

2.3 SSM in Computer Vision

In recent years, State Space Models (SSMs) have been employed in 3D detection as efficient alternatives to attention mechanisms. VoxelMamba (Zhang et al., 2024) designed a grouping-free state space module to expand the receptive field with linear computational cost, while LION (Liu et al., 2024) extended linear grouped modeling to large-scale point clouds. Beyond spatial domain modeling, SSMs have also been introduced into the frequency domain, and WaveMamba (Zhu et al., 2025) is the most closely related method. It first decomposes RGB and infrared features using DWT and then fuses the two modalities sub-band by sub-band, employing a Mamba-based module for the low-frequency sub-bands; subsequently, an IDWT-based detection head reconstructs the fused features to output detection results. Like WaveMamba, FreqLift combines DWT and SSM, but they operate at different stages: WaveMamba (Zhu et al., 2025) applies them to unimodal features to achieve cross-modal fusion, whereas FreqLift employs them after LiDAR-camera BEV fusion to model cross-frequency structures within the fused representation. Regarding the detection head,

FreqLift projects high-frequency responses as heatmap offsets rather than reconstructing features via IDWT.

Yu & Erichson (2026) introduced channel-wise biases to the B matrix in a blocked variant of Mamba-1 to enhance the model's generalized approximation capability; similarly, Mamba-3 (Lahoti et al., 2026) built upon Mamba-2 (Dao & Gu, 2024) by introducing learnable biases after BC normalization. Inspired by these approaches, this paper designs Freq-Mamba based on Mamba-1, incorporating BEV coordinates as positional biases for B and C and utilizing cross-scale low-frequency residuals to modulate the state update step size Δ. Unless otherwise specified, "Mamba" hereafter refers to Mamba-1.

3 METHOD

This section will introduce the FreqLift model we have proposed, a BEV-based multimodal detection method that utilises LiDAR as the primary sensor and a camera as the secondary sensor. We will first provide an overview of the overall network architecture, followed by a detailed description of the two core modules in FreqLift: CameraBEVRefiner and WaveLift.

3.1 PRELIMINARIES

3.1.1 MAMBA

MambaGu & Dao (2023) is essentially a state-space model of a continuous-time system. It is particularly well-suited to processing one-dimensional sequential signals. It takes a one-dimensional input signal $x(t) \in \mathbb{R}^L$, processes it through a hidden layer $h(t) \in \mathbb{R}^N$, and ultimately produces an output $y(t) \in \mathbb{R}^L$. Specifically, it can be represented in the following form:

$$\begin{aligned} h'(t) &= \mathbf{A}\, h(t) + \mathbf{B}\, x(t), \\ y(t) &= \mathbf{C}\, h(t). \end{aligned} \tag{1}$$

Where A denotes the state transition matrix, with $A \in \mathbb{R}^{N\times N}$, and B and C denote two projection matrices, with $B \in \mathbb{R}^{N\times 1}$, $C \in \mathbb{R}^{1\times N}$, and Δ represents the time scale parameter. During the discretisation process, MambaGu & Dao (2023) employs Zero-Order Hold (ZOH) to transform A and B into $\bar{A}$ and $\bar{B}$, which incorporate the time scale parameter. $\bar{A}$ and $\bar{B}$ can be expressed in the following form:

$$\begin{aligned} \bar{A} &= \exp(\Delta A), \\ \bar{B} &= (\Delta A)^{-1}\left(\exp(\Delta A) - I\right) \cdot \Delta B. \end{aligned} \tag{2}$$

Therefore, the discrete form of 1 can be expressed as:

$$h_t = \bar{A}\, h_{t-1} + \bar{B}\, x_t, \tag{3}$$

$$y_t = C h_t. \tag{4}$$

3.1.2 2D DISCRETE WAVELET TRANSFORM

The 2D Discrete Wavelet Transform (2D-DWT) decomposes a spatial feature map into one low-frequency sub-band and three high-frequency sub-bands by applying low-pass and high-pass filters separately in the horizontal and vertical directions (Mallat, 2002):

$$\{LL', LH', HL', HH'\} = \mathrm{DWT}(X). \tag{5}$$

In this paper, we employ the Haar wavelet, specifically Daubechies-1 (db1) and use reflection padding at the feature boundaries. For feature maps with odd spatial dimensions, a single pixel is padded at the bottom or right edge prior to decomposition.

3.2 OVERALL

Figure 1 illustrates our proposed FreqLift, a framework for BEV fusion and cross-frequency state-space modeling for LiDAR-camera object detection. Similar to existing methods (Liu et al., 2023b) (Yin et al., 2021) (Xu et al., 2021), LiDAR point clouds and multi-view images from camera sides are fed separately into their respective backbone networks for processing, in order to extract

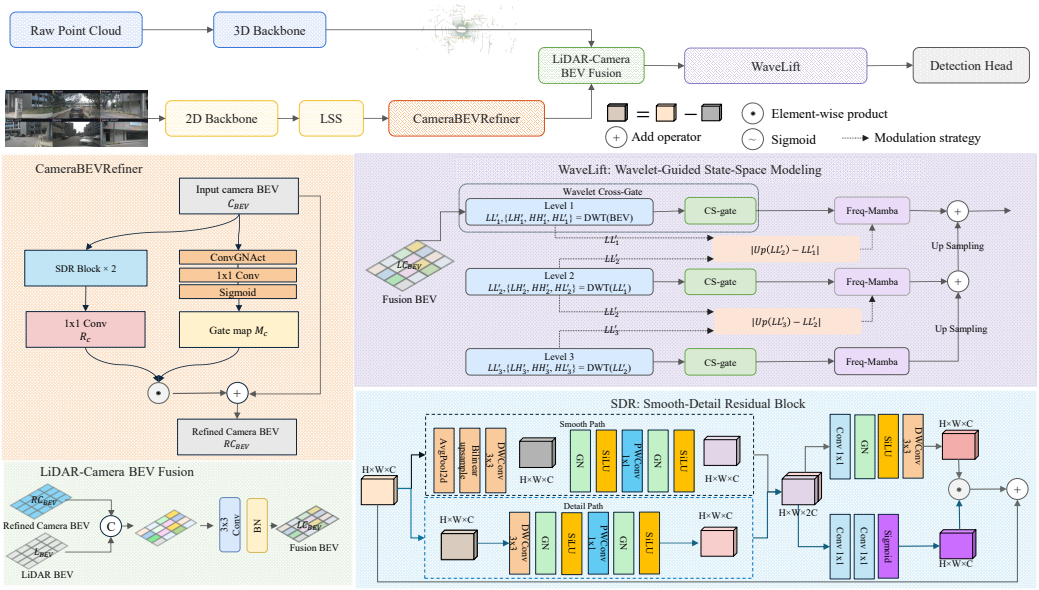

Figure 1: Overall architecture of FreqLift. Multi-view images and LiDAR point clouds are separately encoded into BEV features. The CameraBEVRefiner first refines the camera BEV, then fused with the LiDAR BEV. The fused BEV features undergo cross-frequency state-space modeling via WaveLift and are fed into the detection head to perform 3D object detection.

BEV features. Unlike traditional methods, in the camera branch, we do not directly utilise the camera BEV feature maps generated by LSS. Instead, we further refine the camera BEV maps with CameraBEVRefiner before fusion. Subsequently, the LiDAR BEV and the refined camera BEV are fused. Finally, the fused BEV features are fed into the WaveLift module, where cross-frequency modeling is employed to further enhance and extract BEV features. Detection boxes are then output via our TransFusion-type detection head, which utilises high-frequency bias modulation.

3.3 CameraBEVRefiner

In the BEV space, camera features are susceptible to depth estimation and projection errors, resulting in inherent spatial uncertainty. Since these errors are specific to the camera modality and become entangled with LiDAR features once the two BEVs are fused, we refine the camera BEV before fusion. Rather than directly replacing or suppressing the original Camera BEV, this paper refines it through gated residual correction. To this end, we propose the CameraBEVRefiner module.

As shown in the CameraBEVRefiner section of Figure 1, this module comprises three paths. The residual refinement branch consists of two SDR Blocks and a 1×1 convolution, serving to predict residual correction features denoted as $R_c \in \mathbb{R}^{B\times C\times H\times W}$ from the input Camera BEV. The residual gating branch, composed of a ConvGNAct layer, a 1×1 convolution, and a Sigmoid function, generates spatial gating weights $M_c \in \mathbb{R}^{B\times 1\times H\times W}$ to control the injection strength of residual correction features across different BEV locations. The identity branch preserves the original input features C_{BEV}. Consequently, the refined Camera BEV can be expressed as:

$$RC_{BEV} = C_{BEV} + M_c \odot R_c, \tag{6}$$

where RC_{BEV} denotes the refined Camera BEV, and $\odot$ represents the Hadamard product.

3.3.1 SDR: Smooth-Detail Residual Block

In the residual refinement branch, the SDR blocks are designed to separate smooth structural information from local detail within the camera BEV. Each SDR block splits its input X (for the first block, $X = C_{BEV}$) into a smooth component (average pooling followed by bilinear upsampling) and a detail residual (X minus the smooth component), processes them in separate lightweight convolutional paths, and concatenates the results. A gated residual derived from this joint representation

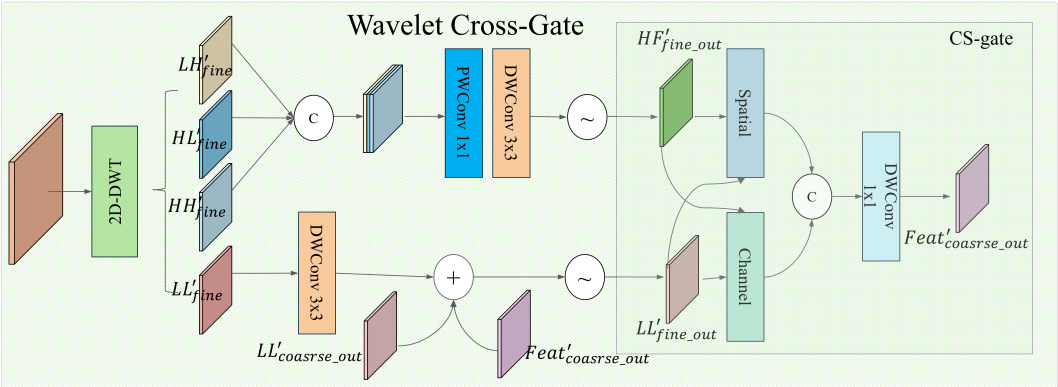

Figure 2: Wavelet Cross-Gate structure. This module obtains high-frequency activations and low-frequency representations via 2D-DWT and utilises spatial and channel gating to achieve cross-frequency feature interaction.

is then added back to X, so that the block corrects rather than overwrites the camera BEV. Details are shown in the SDR panel of Figure 1.

3.3.2 LiDAR-Camera BEV Fusion

Subsequently, the refined camera BEV RC_{BEV} and LiDAR BEV L_{BEV} are concatenated along the channel dimension and passed through convolution and normalization layers to obtain the fused BEV representation:

$$LC_{BEV} = \mathrm{BN}\left(\mathrm{Conv}_{3\times3}\left([RC_{BEV}; L_{BEV}]\right)\right). \tag{7}$$

This fusion process incorporates semantic and local detail information from the camera modality into the BEV representation, providing multi-modal input for the subsequent WaveLift module.

3.4 WaveLift: Wavelet-Guided State-Space Modeling

After obtaining the fused BEV features LC_{BEV}, we feed them into the WaveLift module to model multi-scale spatial structures and frequency information. WaveLift recursively performs three levels of 2D Discrete Wavelet Transform (2D-DWT) decomposition along the low-frequency branch. At the n-th scale, the input features are decomposed into one low-frequency sub-band LL'_n and three high-frequency sub-bands $\{LH'_n, HL'_n, HH'_n\}$.

The scales are not independent of one another. The 2D-DWT decomposition proceeds from fine to coarse along the low-frequency path, progressively obtaining structural representations at lower resolutions. Subsequently, low-level low-frequency activation features and CS-gate outputs are upsampled stage-by-stage and fused with higher-level low-frequency features, establishing a coarse-to-fine cross-scale information flow, as illustrated in Figure 1.

3.4.1 Wavelet Cross-Gate

At each wavelet level shown in Figure 2, we employ a Wavelet Cross-Gate (WCG) to model the interaction between low-frequency structural information and high-frequency detail information. The three high-frequency sub-bands are first concatenated along the channel dimension and transformed via lightweight convolution to obtain a high-frequency representation, while the low-frequency sub-band passes through a low-frequency activation branch to yield a structural representation. WCG then combines spatial and channel gating to model the interaction between the low- and high-frequency representations, generating cross-frequency features for the current scale. The WCG outputs from different scales are fed into the Freq-Mamba module and fused via progressive upsampling to produce the final WaveLift output features.

3.4.2 FREQ-MAMBA

To apply Mamba to 2D BEV features, we convert the BEV features at each scale into four directional 1D sequences: row-major scan, reverse row-major scan, column-major scan, and reverse column-major scan. These four directions are modeled by independent Mamba branches and then fused using weighted aggregation based on directional gating weights generated from the input tokens. This design enables Freq-Mamba to capture both horizontal and vertical dependencies within the BEV spatial domain. Regarding state-space parameters, standard Mamba generates the state update step size Δ, input projection parameter B, and output projection parameter C from the input sequence. In WaveLift, we further modulate Δ by leveraging the structural differences between adjacent low-frequency sub-bands. For the first two levels within the initial three layers of WaveLift, the modulation signal is defined as:

$$G_n^{\Delta} = \mathcal{P}_{\Delta}\left(\left|\mathrm{Up}(LL'_{n+1}) - LL'_n\right|\right), \quad n \in \{1, 2\}, \tag{8}$$

where $\mathrm{Up}(\cdot)$ denotes bilinear upsampling to the current scale, and $\mathcal{P}_{\Delta}(\cdot)$ represents a modulation mapping composed of a linear projection and a Sigmoid function. The modulated state update step size is:

$$\Delta'_n = \Delta_n \odot G_n^{\Delta}. \tag{9}$$

Since there is no lower-resolution LL'_4 for the coarsest third layer, this layer does not employ low-frequency difference modulation; instead, it utilises the unmodulated Freq-Mamba. It should be noted that G_n^{Δ} is generated via Sigmoid; its function is to preserve or attenuate the state update step size based on cross-scale structural differences, rather than explicitly amplifying Δ_n.

Furthermore, we introduce 2D positional biases for B and C. For each BEV token, the normalized coordinate is denoted as $p_i = (x_i, y_i)$. This coordinate undergoes linear projection to generate position-dependent bias terms:

$$[\beta_B(p_i), \beta_C(p_i)] = \mathcal{P}_{BC}(p_i), \tag{10}$$

where $\beta_B(p_i), \beta_C(p_i) \in \mathbb{R}^{d_{state}}$. For the i-th token, the position-modulated parameters are:

$$B'_{n,i} = B_{n,i} + \beta_B(p_i), \quad C'_{n,i} = C_{n,i} + \beta_C(p_i). \tag{11}$$

$\mathcal{P}_{BC}$ is employed independently within each Freq-Mamba branch and initialized to zero. Consequently, the additional positional biases for B and C are inactive at initialization. Meanwhile, the Δ modulation gate is initialized close to 1, so the additional parameter modulation starts close to the standard Mamba formulation in each directional branch, reducing its initial disturbance to the original Mamba dynamics.

3.5 DETECTION HEAD

For the detection head, we retain the existing designs for candidate generation, matching, and regression, leaving the detection loss and box decoder unchanged. The key difference is that the fine-scale, high-frequency response output by the WCG is projected into the category heatmap space to serve as a high-frequency bias for candidate region generation. This design is intended solely to enhance the initialization of candidate queries, enabling the detection head to leverage boundary and local structural cues during candidate region generation, without altering the subsequent box regression, matching, or loss formulations.

4 EXPERIMENTS

4.1 DATASETS AND IMPLEMENTATION DETAILS

Our experiments were conducted on the nuScenes dataset (Caesar et al., 2020), which comprises 700 training scenes and 150 validation scenes. Each scene includes data from a 32-beam LiDAR and six cameras, featuring annotations across ten categories. Performance is officially evaluated using mean average precision (mAP) and the nuScenes detection score (NDS).

We train our model using PyTorch and the open-source MMDetection3D (Contributors, 2020) framework. We have chosen a ResNet-50 (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015) as the 2D backbone network for our image branch. For the LiDAR branch, we use VoxelNet as the sparse 3D backbone, setting the voxel size to (0.075m, 0.075m, 0.2m) and the point cloud range to [-54m, 54m] for the X and Y axes and [-5m, 3m] for the Z axis.

Table 1: Comparison with state-of-the-art methods on the nuScenes **validation** set. "–" denotes results not reported. Best results are in **bold**.

Method	Modality	mAP	NDS
UVTR (Li et al., 2022)	C	33.3	41.9
PETRv2 (Liu et al., 2023a)	C	34.9	45.6
BEVFusion-MIT-C (Liu et al., 2023b)	C	35.5	41.2
TransFusion-L (Bai et al., 2022)	L	65.1	70.1
BEVFusion-MIT-L (Liu et al., 2023b)	L	64.7	69.2
FocalFormer3D (Chen et al., 2023a)	L	66.5	71.1
BEVFix (Li et al., 2025)	L	65.5	70.0
TransFusion (Bai et al., 2022)	LC	67.5	71.3
BEVFusion-PKU (Liang et al., 2022)	LC	67.9	71.0
BEVFusion-MIT (Liu et al., 2023b)	LC	68.5	71.4
GraphBEV (Song et al., 2024)	LC	70.1	72.9
DAL-Base (Huang et al., 2024)	LC	70.0	73.0
BEVFix (Li et al., 2025)	LC	69.1	72.1
BEVDiffuser (Ye et al., 2025)	LC	69.2	71.9
UGFusion (Zhang et al., 2026b)	LC	71.0	73.1
LBYF (Li et al., 2026)	LC	71.5	73.6
ISFusion (Yin et al., 2024)	LC	**72.8**	**74.0**
FreqLift-C (Ours)	C	26.0	31.5
FreqLift-L (Ours)	L	67.1	71.0
FreqLift (Ours)	LC	71.2	73.1

4.1.1 Training and Inference

We employ a multi-stage training strategy and do not use CBGS (Zhu et al., 2019) in any of the stages. The AdamW optimizer (Loshchilov & Hutter, 2017) is used across all stages with a weight decay of 0.01 and a cosine learning-rate schedule. For the LiDAR-only branch, inspired by BEV-Fix (Li et al., 2025), we adopt a two-stage training process. In the first stage, the model is trained for 20 epochs using a single-layer heatmap and a single-layer decoder with a base learning rate of 1×10^{-4}, ground-truth (GT) sample augmentation is applied during all but the final 5 epochs. In the second stage, training continues for another 6 epochs again with a base learning rate of 1×10^{-4}, while GT sample augmentation is used only during the first epoch. In each stage, the batch size is set to 4. Additionally, the LiDAR backbone features are frozen to focus on refining the detection head.

For the camera branch, to avoid confusing factors, such as additional training data, model capacity, or specific optimization strategies, introduced by using an externally pre-trained camera-only detector, we do not employ off-the-shelf camera detection models. Instead, we independently train a model using ResNet-50 as the backbone for 20 epochs with a base learning rate of 1.5×10^{-4}, a batch size of 4, and a camera input image resolution of 384 × 1056 to obtain a camera BEV representation better suited to our framework.

During the fusion stage, we first freeze the LiDAR backbone, allowing the camera BEV branch to learn fusion relationships based on the stable LiDAR BEV representation, while training the camera and fusion modules at a learning rate of 5×10^{-5} and a batch size of 4. Subsequently, we unfreeze both branches and perform joint fine-tuning at a learning rate of 2×10^{-5}. All experiments were conducted on a single RTX PRO 4000 Blackwell or RTX PRO 5000 Blackwell GPU.

4.2 Overall Results

To evaluate the detection performance of FreqLift, we compare it with representative camera-only, LiDAR-only, and LiDAR-camera fusion methods on the nuScenes validation set. The results reported here for FreqLift were obtained without using test-time augmentation or CBGS.

Table 2: Comparison with LiDAR-camera fusion methods using different camera branches on the nuScenes validation set. "C" denotes camera-only detection

Method	Image Backbone	Camera Branch Detector	mAP(C)
TransFusion (Bai et al., 2022)	ResNet-50	pretrained	–
BEVFusion-MIT (Liu et al., 2023b)	Swin-T	pretrained	35.5
SparseFusion (Xie et al., 2023)	Swin-T	pretrained	–
FocalFormer3D (Chen et al., 2023a)	ResNet-50	pretrained	30.0
CMT (Yan et al., 2023)	VoV-99	pretrained	40.0
FreqLift (Ours)	ResNet-50	self-trained	26.0

In the camera-only setting, FreqLift-C achieved 26.0% mAP and 31.5% NDS, trailing behind all the strong camera-based detection methods listed in the table. Nevertheless, when this camera branch is used for LiDAR-camera fusion, the complete model achieves competitive performance compared to other state-of-the-art methods.

In the LiDAR-only setting, FreqLift-L achieves 67.1% mAP and 71.0% NDS, outperforming TransFusion-L by +2.0 mAP / +0.9 NDS and BEVFusion-MIT-L by +2.4 mAP / +1.8 NDS.

In the LiDAR-camera fusion setting, FreqLift achieved 71.2% mAP and 73.1% NDS. While it may slightly trail SOTA methods, FreqLift nonetheless delivers competitive overall results.

Table 2 further compares the camera-side pretraining used by different LiDAR-camera fusion methods. While most existing approaches rely on pre-trained camera-only detectors or pre-trained camera BEV branches, FreqLift simply utilizes an ImageNet-pre-trained ResNet50 as the image backbone to train a camera branch in isolation. Although this independently trained camera branch achieves an mAP of only 26%, the complete LiDAR-camera fusion model attains an mAP of 71.2%, +4.1 mAP over the LiDAR-only baseline.

4.3 Ablation Study

The following ablation experiments were conducted under the same validation protocol to analyze the impact of design choices within FreqLift. Furthermore, Table 3 primarily compares internal variants of WaveLift.

4.3.1 WaveLift Architecture and High-Frequency Bias Ablation

Table 3 analyzes the internal design of WaveLift in a LiDAR-only setting. All variants employ Freq-Mamba at each layer and retain the wavelet decomposition and WCG-based frequency interaction mechanisms, while varying the number of layers, high-frequency bias, and Δ-modulation settings. The full "three-layer" configuration achieves an mAP of 67.1% and an NDS of 71.0%. Reducing WaveLift from three layers to two results in the mAP dropping to 66.4%, the largest decline among the variants, indicating that increasing the number of frequency layers is beneficial. Removing the high-frequency bias or Δ-modulation causes the mAP to drop to 66.5% and 66.7%, respectively, demonstrating that each component contributes to performance improvement in the current setup.

Table 3: LiDAR-only ablation on WaveLift design. FM denotes Freq-Mamba. Blocks are arranged from fine to coarse scales.

#WCG	Blocks	HF Bias	Δ Mod.	mAP / NDS
3	FM-FM-FM	✓	✓	**67.1** / 71.0
3	FM-FM-FM	–	✓	66.5 / 70.6
3	FM-FM-FM	✓	–	66.7 / 70.5
2	FM-FM	✓	✓	66.4 / 70.3

4.3.2 Ablation of Camera BEV Refinement in Camera-Only and LiDAR-Camera Fusion Settings

Table 4 evaluates the role of CameraBEVRefiner in both camera-only and LiDAR-camera fusion settings. In the camera-only branch, adding CameraBEVRefiner improves mAP/NDS from 18.8%/24.6% to 26.0%/31.5%. In the fusion setting, directly fusing the unrefined camera BEV improves FreqLift-L from 67.1%/71.0% to 68.5%/71.7% (+1.4 mAP), and enabling CameraBEVRefiner further raises performance to 71.2%/73.1%, i.e., +2.7 mAP over direct fusion and +4.1 mAP over FreqLift-L. These results indicate that the camera BEV provides complementary information for LiDAR-based detection and that refining the camera BEV before fusion improves fusion performance. CameraBEVRefiner and WaveLift contain approximately 0.26M and 1.60M parameters, respectively, adding about 1.86M parameters to the model in total.

Table 4: Ablation on camera BEV refinement in camera-only and LiDAR-camera fusion settings. All fusion variants use the same LiDAR checkpoint, WaveLift configuration, detection head, training schedule, and data augmentation.

Setting	CameraBEVRefiner	mAP / NDS
Camera-only branch		
w/o Refiner	–	18.8 / 24.6
w/ Refiner	✓	26.0 / 31.5
LiDAR-Camera fusion		
LiDAR-only	–	67.1 / 71.0
w/o Refiner	–	68.5 / 71.7
w/ Refiner	✓	**71.2 / 73.1**

5 Conclusion and Future Work

This paper introduces FreqLift, a method designed to enhance 3D object detection performance in LiDAR-camera BEV fusion. To address the loss of local details caused by BEV downsampling and the depth noise potentially introduced by direct camera-to-BEV fusion, we design WaveLift and CameraBEVRefiner for cross-frequency modeling and camera BEV residual refinement, respectively. Experiments demonstrate that FreqLift achieves competitive performance on the nuScenes validation set. Notably, it effectively improves LiDAR-camera fusion results even when utilizing a ResNet-50-based camera branch, without relying on a heavy camera detection branch. However, the proposed method has certain limitations: the multi-stage training process increases reproduction costs, gains across object categories are not entirely uniform, and the specific roles of different WaveLift levels and Freq-Mamba combinations require further analysis. Given that submissions to the nuScenes test set often involve training on the combined train-val sets, test-time augmentation, or stronger backbones, this paper focuses on analyzing the contributions of the proposed modules within a controlled validation set setting.

This work focuses on a setting where the camera branch does not utilize external camera pre-training. Evaluating CameraBEVRefiner on publicly pre-trained camera BEV branches allows for a clearer understanding of the relationship between its performance gains and the quality of the camera branch; furthermore, this approach extends the analysis of WaveLift to full LiDAR-camera models across multiple random seeds, enables robustness evaluation under depth and calibration perturbations, and facilitates the exploration of refinement on the LiDAR side.

AI use statement

In this work, we used generative AI tools to assist with translation between Chinese and English and to provide feedback on the presentation of ablation studies. We have not used generative AI tools to generate synthetic data sets, formulate mathematical claims, or implement the proposed methods, and the remaining required disclosure tasks (proofs, dataset cleaning, and qualitative data analysis) are not applicable to this work. Additionally, we used generative AI tools to draft and revise parts

of the paper, suggest its structure, and improve readability. We have reviewed all AI-assisted work. All experiments were designed and conducted by the authors, all reported numbers were checked by the authors against the experiment logs, and descriptions of related work were verified against the original papers. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1090–1099, 2022.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Yilun Chen, Zhiding Yu, Yukang Chen, Shiyi Lan, Anima Anandkumar, Jiaya Jia, and Jose M Alvarez. Focalformer3d: focusing on hard instance for 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8394–8405, 2023a.

Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21674–21683, 2023b.

MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. `https://github.com/open-mmlab/mmdetection3d`, 2020.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Pan Gao, Yuling Fan, Shuai Ren, Xiwen Ren, and Ping Zhang. Mpcf: Multi-phase consolidated fusion for multi-modal 3d object detection with pseudo point cloud. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

Junjie Huang, Yun Ye, Zhujin Liang, Yi Shan, and Dalong Du. Detecting as labeling: Rethinking lidar-camera fusion in 3d object detection. In *European Conference on Computer Vision*, pp. 439–455. Springer, 2024.

Xin Jin, Haisheng Su, Kai Liu, Cong Ma, Wei Wu, Fei Hui, and Junchi Yan. Unimamba: Unified spatial-channel representation learning with group-efficient mamba for lidar-based 3d object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1407–1417, 2025.

Aakash Lahoti, Kevin Y Li, Berlin Chen, Caitlin Wang, Aviv Bick, J Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. *arXiv preprint arXiv:2603.15569*, 2026.

Qiufu Li, Linlin Shen, Sheng Guo, and Zhihui Lai. Wavelet integrated cnns for noise-robust image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7245–7254, 2020.

Wenxuan Li, Jian Zhou, Chi Chen, Hongkai Yu, Bo Du, and Qin Zou. Bevfix: Deep feature enhancement for robust 3d object detection. *Neural Networks*, 190:107675, 2025.

Xiang Li, Zhangchi Hu, Xu Xiao, and Bin Kong. Look before you fuse: 2d-guided cross-modal alignment for robust 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11589–11598, June 2026.

Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *Advances in Neural Information Processing Systems*, 35:18442–18455, 2022.

Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in neural information processing systems*, 35:10421–10434, 2022.

Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 773–782, 2018.

Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3262–3272, 2023a.

Zhe Liu, Jinghua Hou, Xinyu Wang, Xiaoqing Ye, Jingdong Wang, Hengshuang Zhao, and Xiang Bai. Lion: Linear group rnn for 3d object detection in point clouds. *Advances in Neural Information Processing Systems*, 37:13601–13626, 2024.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 2774–2781. ieee, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 2002.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pp. 194–210. Springer, 2020.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas 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*, pp. 652–660, 2017a.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 770–779, 2019.

Ziying Song, Lei Yang, Shaoqing Xu, Lin Liu, Dongyang Xu, Caiyan Jia, Feiyang Jia, and Li Wang. Graphbev: Towards robust bev feature alignment for multi-modal 3d object detection. In *European conference on computer vision*, pp. 347–366. Springer, 2024.

Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17591–17602, 2023.

Shaoqing Xu, Dingfu Zhou, Jin Fang, Junbo Yin, Zhou Bin, and Liangjun Zhang. Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In *2021 IEEE international intelligent transportation systems conference (ITSC)*, pp. 3047–3054. IEEE, 2021.

Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 18268–18278, 2023.

Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11040–11048, 2020.

Xin Ye, Burhaneddin Yaman, Sheng Cheng, Feng Tao, Abhirup Mallik, and Liu Ren. Bevdiffuser: Plug-and-play diffusion model for bev denoising with ground-truth guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1495–1504, 2025.

Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14905–14915. IEEE, 2024.

Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Multimodal virtual point 3d detection. *Advances in Neural Information Processing Systems*, 34:16494–16507, 2021.

Annan Yu and N Benjamin Erichson. Block-biased mamba for long-range sequence processing. *Advances in Neural Information Processing Systems*, 38:120541–120577, 2026.

Guowen Zhang, Lue Fan, Zhen Lei, Zhaoxiang Zhang, Lei Zhang, et al. Voxel mamba: Group-free state space models for point cloud based 3d object detection. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024.

Guowen Zhang, Chenhang He, Liyi Chen, and Lei Zhang. Bevdilation: Lidar-centric multi-modal fusion for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12448–12456, 2026a.

Ziyang Zhang, Hegui Zhu, and Libo Zhang. Ugfusion: Uncertainty-guided lidar-camera fusion for 3-d object detection. *IEEE Sensors Journal*, 2026b.

Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4490–4499, 2018.

Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019.

Haodong Zhu, Wenhao Dong, Linlin Yang, Hong Li, Yuguang Yang, Yangyang Ren, Qingcheng Zhu, Zichao Feng, Changbai Li, Shaohui Lin, et al. Wavemamba: Wavelet-driven mamba fusion for rgb-infrared object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11219–11229, 2025.