Temporal and Spatial Representation Learning for Multimodal Low-Beam 3D Object Detection

Lin Wang[1], **Shiliang Sun**[1,2,3,*], **Jing Zhao**[1]

[1]School of Computer Science and Technology, East China Normal University, Shanghai, China
[2]School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China
[3]Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, Shanghai, China
linwang_ecnu@163.com, shiliangsun@gmail.com, jzhao@cs.ecnu.edu.cn

Abstract

To facilitate the large-scale deployment of autonomous driving in real-world scenarios, developing low-cost and high-performance 3D object detection systems has become a critical technical challenge. Although high-beam LiDARs provide denser point cloud data, their prohibitive hardware cost and high power consumption limit their practicality. In contrast, low-beam LiDARs offer advantages in terms of affordability and energy efficiency, but often suffer from inadequate perception accuracy due to their sparser point cloud data. This paper focuses on the task of multimodal 3D object detection with low-beam LiDARs, and proposes a novel approach that integrates temporal and spatial representation learning to enhance detection accuracy under sparser sensor conditions. Specifically, our approach comprises: (1) a Temporal Feature Prediction Learning (TFPL) module, which predicts the current BEV representation based on a sequence of historical BEV features; (2) a Spatial Feature Observation Learning (SFOL) module, which aligns BEV (Bird's-Eye-View) features from high-beam and low-beam LiDAR to enforce the low-beam features to approximate high-beam representations; (3) an Uncertainty-Aware Fusion (UAF) strategy, which performs feature-wise weighting between the predicted and observed BEV features by leveraging channel-wise variances, effectively mitigating perturbations in the learned BEV representations. Extensive experiments on the KITTI and nuScenes 3D object detection datasets demonstrate that the proposed approach significantly improves detection performance under low-beam LiDAR configurations.

Introduction

Autonomous driving relies heavily on high-precision perception systems to ensure safety and robustness. To enhance perception performance, multimodal sensing technologies have attracted widespread attention, particularly approaches that fuse LiDAR point clouds with image data. LiDAR provides accurate three-dimensional structural information, while images contain rich semantic content. These two modalities are highly complementary, enabling more comprehensive and reliable scene understanding (Wang et al. 2023; Yu et al. 2023).

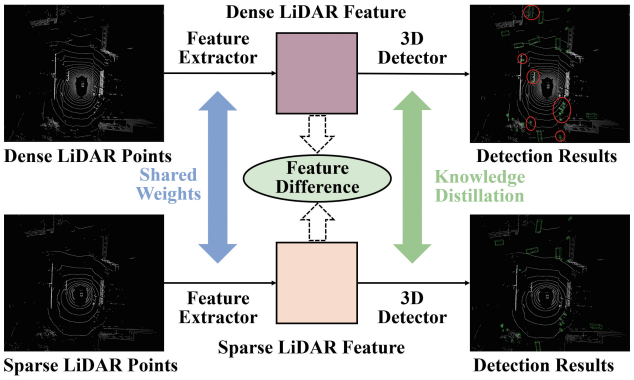

Figure 1: Schematic of existing approaches for learning feature representations from sparse low-beam LiDAR data. These methods directly distill high-beam features into the low-beam branch, which is overly restrictive and leads to unstable detection results in the low-beam branch model. Red regions highlight missed detections.

LiDAR sensors are commonly categorized as high-beam or low-beam based on the number of laser beams. High-beam LiDARs (e.g., 128/64-beam) offer higher spatial resolution and dense point clouds, thus supporting fine-grained 3D perception. However, their high cost limits deployment in mass-production scenarios (Wei et al. 2022; Shan et al. 2023). Although solid-state LiDARs have seen cost reductions, their inherently limited field-of-view (FOV) requires multiple sensor placements on the vehicle, leading to a multi-sensor fusion process to generate high-beam point clouds. This multi-sensor configuration introduces additional challenges related to power consumption and storage demands. In contrast, mechanical low-beam LiDARs (e.g., 32/16-beam) offer sparser point cloud data but present significant advantages in terms of cost efficiency and system integration, making them more amenable to large-scale deployment and resource-limited scenarios (Huang 2024). Therefore, achieving accurate perception with low-beam LiDAR is a key research challenge (Li, Ma, and Li 2024). Meanwhile, camera-only 3D object detection models (Lu et al. 2025; Zhang et al. 2025) have made progress in learning geometric representations from images alone,

*Corresponding author.

thereby reducing dependence on expensive high-beam LiDARs. However, images inherently suffer from limitations in depth perception, and the precision of LiDAR in depth measurement remains irreplaceable. Moreover, in adverse conditions such as nighttime or poor weather, camera-based methods often degrade due to limited visibility and illumination, whereas LiDAR remains robust to such changes.

The aforementioned studies motivate the current research on fusing images with low-beam LiDAR to achieve reliable and accurate 3D object detection in diverse real-world scenarios. To tackle this problem, several works (Wei et al. 2022; Hu, Liu, and Hu 2023) have proposed knowledge distillation frameworks tailored for 3D object detection, aiming to mitigate the performance degradation caused by the disparity in LiDAR beam density between high-beam and low-beam inputs. Similarly, Bi3D (Yuan et al. 2023) tackles cross-domain degradation by aligning source and target features through a bi-domain active learning strategy. In summary, as illustrated in Fig. 1, existing methods typically adopt a teacher-student distillation paradigm, where separate 3D feature extractors are employed for high-beam and low-beam LiDAR inputs. The student model is trained to mimic the feature representations extracted from dense point clouds by the teacher, enabling the transfer of representational capacity from high-resolution to low-resolution settings (Lu, Lin, and Hsu 2025).

Although existing methods have achieved promising results on large-scale autonomous driving datasets, such as KITTI (Geiger, Lenz, and Urtasun 2012) and nuScenes (Caesar et al. 2020), we argue that direct feature-level imitation learning can lead to unstable perception of local features. As illustrated in Fig. 1, the low-beam branch fails to detect certain targets that are successfully captured by the high-beam branch. These observations motivate us to move beyond feature-level imitation and leverage richer contextual information. To this end, we adopt a LiDAR–camera fusion approach, incorporating temporal features from consecutive frames as predictive cues and spatial features from high-beam and low-beam branches as observational cues, thereby enhancing the model's fine-grained feature representation of local regions. Building upon this, we introduce the concept of uncertainty modeling by first representing the observed and predicted BEV features as Gaussian distributions before fusion, effectively mitigating the perturbations arising from observation and prediction learning. Specifically, this paper proposes: (1) a Temporal Feature Prediction Learning (TFPL) module that predicts the current point cloud features based on the historical sequential BEV features; (2) a Spatial Feature Observation Learning (SFOL) module that enforces consistent BEV-space feature representations between high-beam and low-beam point clouds; and (3) an Uncertainty-Aware Fusion (UAF) strategy that estimates channel-wise uncertainty to fuse TFPL-derived predictions and SFOL-derived observations in the BEV feature space.

Our contributions can be summarized as follows:

- We propose a novel temporal and spatial representation learning framework for multimodal low-beam 3D object detection which explicitly captures both spatial observations and temporal predictions in BEV space to enhance robustness with low-beam inputs.
- We propose two key modules: the TFPL for forecasting current BEV representations from historical features, and the SFOL for reconstructing low-beam BEV features by leveraging spatial discrepancies. These are integrated through a UAF strategy that adaptively weights features according to their uncertainty.
- We conduct comprehensive experiments under low-beam LiDAR configurations on benchmark datasets. Results demonstrate that our method significantly outperforms state-of-the-art baselines under low-beam settings.

Related Work

LiDAR-Camera Fusion for Object Detection. Fusion perception methods combining images and point clouds have gained significant attention. For instance, F-PointNet (Qi et al. 2018) generates 2D image detections, projects them into 3D frustums to produce proposals, and refines 3D bounding boxes with point cloud data. However, its performance largely relies on the accuracy of 2D detections. In contrast, methods like PointPainting (Vora et al. 2020) and PointAugmenting (Wang et al. 2021) enhance point cloud representations by incorporating 2D semantic segmentation scores and image features, respectively, as additional inputs. TransFusion (Bai et al. 2022) proposes a unified framework for 3D object detection by effectively fusing LiDAR and camera features in the BEV space using a transformer-based fusion module. It leverages sparse LiDAR geometry to guide dense image feature aggregation, enabling precise and efficient multi-modal fusion for improved detection performance. In recent years, the LSS (Philion and Fidler 2020) method has pioneered a new paradigm for image-point cloud fusion based on BEV feature representation. Through the Lift-Splat-Shoot operation, camera data is transformed into BEV features. Building on this, BEVFusion (Liu et al. 2023) extracts BEV features separately from images and point clouds, and then aggregates them in a cascaded manner for downstream 3D segmentation and detection tasks. SimpleBEV (Zhao et al. 2024) performs 3D object detection in BEV space by encoding LiDAR and camera inputs through dual branches, enhancing LiDAR features via multi-level fusion and auxiliary camera supervision, and improving robustness through GT-Paste and multi-model ensembling with test-time augmentation.

Point Cloud Cross-Beam Learning. Despite significant progress in multimodal object detection, most existing works still face challenges of expensive annotation costs and poor cross-beam generalization caused by variations in LiDAR beam configurations (Griesbacher and Fruhwirth-Reisinger 2025). Recently, several studies have attempted to address cross-beam LiDAR object detection. LiDAR Distillation (Wei et al. 2022) employs spherical coordinate transformation and normalized sampling ranges to unify sampling between high-beam and low-beam LiDAR point clouds, enabling imitation learning within a single modality. BEVDistill (Chen et al. 2022) introduces a novel cross-modal knowledge distillation framework that aligns and

transfers knowledge between LiDAR-based and multi-view image-based detectors within the BEV space. It leverages a teacher-student paradigm to effectively unify heterogeneous representations without requiring additional inference cost. Bi3D (Yuan et al. 2023) proposes a dual-domain active learning framework for cross-domain 3D detection, employing domain-aware source sampling and diversity-based target sampling to select relevant and informative samples. It also integrates proposal-based feature distillation to enhance feature alignment across domains while working within limited annotation budgets. CLIX^{3D} (Hegde et al. 2025) addresses domain generalization in 3D object detection by leveraging multimodal LiDAR-image data to enhance robustness against unseen domain shifts. It combines multimodal fusion with supervised contrastive learning to align same-class features across domains while separating different classes. While the above methods alleviate cross-domain learning challenges to some extent, they either neglect the complementary role of image data or fail to exploit temporal information to enhance feature representation.

Fusion with Uncertainty. In autonomous driving perception, multi-modal and temporal feature fusion is essential for improving the accuracy and robustness of 3D scene understanding (Li et al. 2024). However, the fusion process involves challenges such as multimodal information heterogeneity, data noise, and feature noise, which make uncertainty-aware fusion even more critical. To address the uncertainty inherent in the fusion process, numerous studies have focused on uncertainty modeling and fusion strategies to effectively enhance the stability and reliability of perception systems. For example, CLR-BNN (Ravindran, Santora, and Jamali 2022) introduces a Bayesian neural network-based multi-sensor fusion framework that models uncertainty to improve multi-object detection accuracy and prediction reliability in autonomous driving, outperforming traditional deterministic fusion models. MOCT (Chiu et al. 2024) proposes a 3D multi-object cooperative tracking algorithm leveraging a differentiable multi-sensor Kalman Filter that learns per-detection measurement uncertainty, fully exploiting the optimality of the Kalman Filter to better handle uncertainty in vehicle-to-vehicle cooperative perception. UncertainBEV (Xu et al. 2025) designs the UncertainFuser module to model the uncertainty of camera and LiDAR features, dynamically adjusting fusion weights to effectively alleviate feature misalignment caused by sensor calibration errors, significantly improving the accuracy and robustness of multimodal perception. For the uncertainty-aware fusion of two sets of data or features, the formulation is expressed as follows:

$$\mathbf{F}^{(f)} = \frac{\sigma_1^{-2}\mathbf{F}^{(2)} + \sigma_2^{-2}\mathbf{F}^{(1)}}{\sigma_1^{-2} + \sigma_2^{-2}}, \tag{1}$$

where $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$ denote two different feature vectors. σ_1^2 and σ_2^2 represent the variances (uncertainties) of the corresponding features. The fused feature $\mathbf{F}^{(f)}$ is computed as a weighted sum of the two features, where the weights are inversely proportional to their variances, giving more importance to features with lower uncertainty.

Method

As illustrated in Fig. 2, both the Low-Beam Branch and High-Beam Branch share the same Camera BEV Encoder module, LiDAR BEV Encoder module, and BEV Feature Fusion module, which respectively extract features to obtain BEV_f^{low} and BEV_f^{high}. First, the historical temporal features $BEV_f^i \in \{BEV_f^{t-n}, \ldots, BEV_f^{t-2}, BEV_f^{t-1}\}$, where n denotes the number of historical frames, are fed into the Temporal Feature Prediction Learning module to predict the current BEV feature BEV_f^{pred}. This process is referred to as Learning from Predictions. Meanwhile, BEV_f^{low} from the Low-Beam Branch is passed into the Spatial Feature Observation Learning module to generate the observation BEV feature BEV_f^{obs}, which approximates BEV_f^{high} from the High-Beam Branch. This process is referred to as Learning from Observations. Finally, the features BEV_f^{obs} and BEV_f^{pred} obtained from the two processes are fused through a carefully designed Uncertainty-Aware Fusion strategy, resulting in the final fused feature at the current time step, denoted as BEV_f^t. The fused feature BEV_f^t is then used for the subsequent detection task.

Temporal Feature Prediction Learning

In this section, we propose a Temporal Feature Prediction Learning (TFPL) module by exploiting temporal continuity and spatial consistency across adjacent frames. Fig. 3 illustrates the architecture of the temporal BEV feature fusion module designed to aggregate information from multiple historical BEV feature maps. Given a sequence of BEV features $BEV_f^{i-n}, \ldots, BEV_f^{i-1}$ from the previous n frames, the module progressively integrates temporal information to generate a predictive BEV representation BEV_f^{pred} for the current frame. Each BEV feature first undergoes a series of residual processing blocks and normalization operations. Specifically, the earliest frame BEV_f^{i-n} is passed through a ResBlock to enhance its representation. Then, each BEV feature at time t is combined with the processed output from the previous time step via an Add & Norm operation, followed by another ResBlock. This recursive structure enables the network to gradually accumulate temporal context across frames. The intermediate outputs from all temporal steps are concatenated and passed through a Feed Forward network to produce the final predicted BEV feature BEV_f^{pred}.

This design facilitates the modeling of long-range temporal dependencies while maintaining spatial consistency across frames. To further constrain the predicted BEV features of the Low-Beam Branch, we also apply representation learning based on Temporal Feature Prediction Learning to the sequential BEV features from the High-Beam branch. Subsequently, the two branches align their predicted features, BEV_f^{pred} and $BEV_f^{pred(high)}$, using a consistency loss. Let the input BEV feature map be denoted as $\mathbf{X} \in \mathbb{R}^{W \times H \times C}$, where W and H are the spatial dimensions of the grid, and C is the number of feature channels per grid cell. To efficiently compute the consistency loss between the

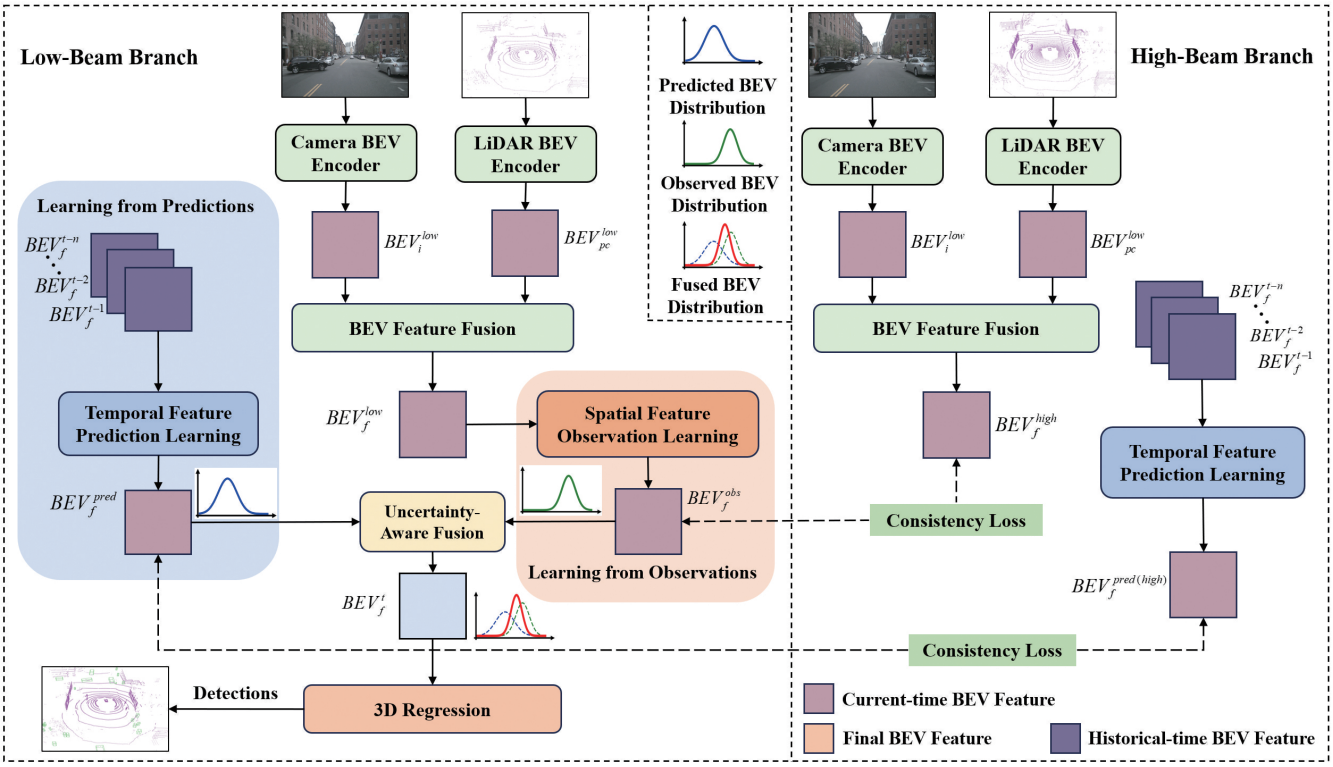

Figure 2: Overall Architecture. The diagram consists of two main parts: the Low Beam Branch on the left and the High Beam Branch on the right. The Camera BEV Encoder, LiDAR BEV Encoder, and BEV Feature Fusion modules share weights, which are pre-trained using the High Beam Branch. The core components include two representation learning modules: Temporal Feature Prediction Learning (TFPL) and Spatial Feature Observation Learning (SFOL), along with an Uncertainty-Aware Fusion (UAF) strategy. The High Beam Branch is used only during training and is removed at inference time.

two sets of BEV features, the BEV tensor is reshaped into a vector of size $(1, W \times H \times C)$. Here, $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors, and $\| \cdot \|$ represents the vector magnitude. The consistency loss is then computed as:

$$\mathcal{L}_{\text{cons}}^{p} = 1 - \frac{\langle BEV_f^{pred}, BEV_f^{pred(high)} \rangle}{\|BEV_f^{pred}\| \cdot \|BEV_f^{pred(high)}\|}. \quad (2)$$

Spatial Feature Observation Learning

In this section, we propose a Spatial Feature Observation Learning (SFOL) module to enhance low-beam feature representations. The method leverages a pre-trained high-beam 3D detector to extract features from both high-beam and low-beam inputs via a shared extractor. By aligning low-beam features (BEV_f^{low}) with high-beam features (BEV_f^{high}), the network transfers rich spatial knowledge from high-beam data to low-beam inputs. As shown in the Fig. 3, the SFOL module consists of a Self-Attention, an MLP (Multi-Layer Perceptron), a Feed Forward, and an Add & Norm submodule.

To model long-range dependencies within the BEV space, a self-attention mechanism is applied over the BEV feature map, enabling each spatial location to attend to all others through dynamically computed attention weights. The MLP with 3 layers is applied as a composition of linear transformations and nonlinear activations. Each hidden layer applies an affine transformation followed by a nonlinearity, while the final layer outputs a transformed feature representation without activation. To ensure stable training and effective information propagation, we incorporate an Add & Norm sublayer before feeding the Feed Forward stage. This process converts the residual-bearing BEV features into BEV_f^{obs}, whose shape matches that of the input BEV_f^{low}, enabling effective alignment with BEV_f^{high} from the high-beam branch. Furthermore, a consistency loss $\mathcal{L}_{\text{cons}}^{o}$ is introduced to quantify the discrepancy between the two BEV representations $\{BEV_f^{low}, BEV_f^{high}\}$, aiming to refine the feature representation of BEV_f^{low}. The computation of $\mathcal{L}_{\text{cons}}^{o}$ follows the same formulation as $\mathcal{L}_{\text{cons}}^{p}$ in Equation (2).

Uncertainty-Aware Fusion

Prior to modeling the observed and predicted BEV features with a Gaussian distribution, we note that the channel-wise feature vectors at each BEV location tend to exhibit properties of approximately Gaussian distributions. This is due to the fact that both observed and predicted BEV features are obtained through deep convolutional or transformer-based encoders, which involve multiple layers of linear transfor-

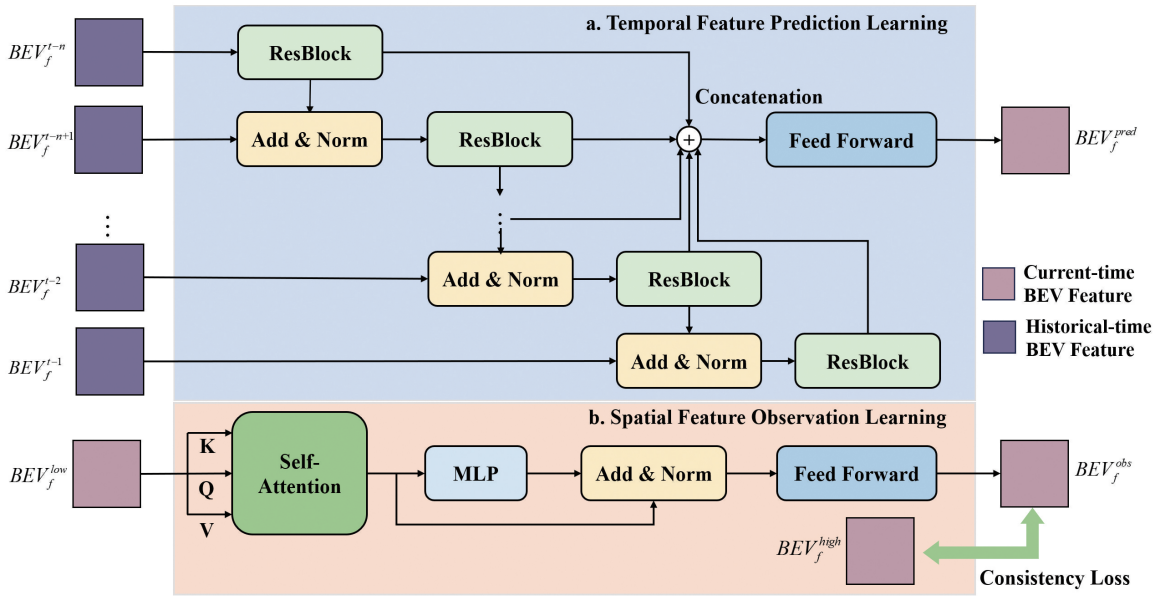

Figure 3: Structural diagram of the temporal and spatial representation learning module. The upper part of the figure illustrates the Temporal Feature Prediction Learning (TFPL) module, while the lower part depicts the Spatial Feature Observation Learning (SFOL) module.

mations, nonlinearities, and normalization operations. According to the Central Limit Theorem, such compositions of transformations over high-dimensional data tend to produce feature distributions that are approximately Gaussian, especially when aggregated across time, views, or sensor modalities. This observation justifies the use of Gaussian modeling for representing the uncertainty and variability of BEV features at each spatial location. Given two BEV feature maps ($\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{W\times H\times C}$), we perform Gaussian modeling along the channel dimension at each spatial location. Specifically, for every grid cell (i, j), we treat the corresponding feature vectors $\mathbf{x}_1^{(i,j)}, \mathbf{x}_2^{(i,j)} \in \mathbb{R}^C$ as samples to estimate a Gaussian distribution:

$$\begin{aligned} \mathbf{x}_1^{(i,j)} &\sim \mathcal{N}(\boldsymbol{\mu}_{i,j}^{pred}, \boldsymbol{\Sigma}_{i,j}^{pred}), \\ \mathbf{x}_2^{(i,j)} &\sim \mathcal{N}(\boldsymbol{\mu}_{i,j}^{obs}, \boldsymbol{\Sigma}_{i,j}^{obs}), \end{aligned} \tag{3}$$

where $\boldsymbol{\mu}_{i,j}^{pred}, \boldsymbol{\mu}_{i,j}^{obs} \in \mathbb{R}^C$ and $\boldsymbol{\Sigma}_{i,j}^{pred}, \boldsymbol{\Sigma}_{i,j}^{obs} \in \mathbb{R}^C$ denote the mean and covariance of the channel-wise feature distribution at position (i, j). This modeling captures the uncertainty and correlation across feature channels, providing a probabilistic representation that can be utilized for feature fusion or confidence-aware reasoning.

In this section, an Uncertainty-Aware Fusion (UAF) strategy is proposed by modeling BEV_f^{obs} and BEV_f^{pred} as Gaussian distributions, thereby explicitly capturing the uncertainty in each feature.. The UAF strategy first flattens the two BEV feature maps along the channel dimension (C) and computes the mean and variance for each channel, as shown in the Fig. 4. Then, based on Equation (5), a weight matrix (K) is constructed using the variances of the two features to guide the fusion process. To be specific, let

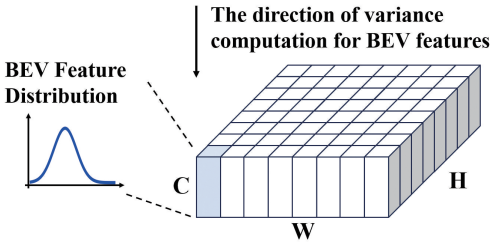

Figure 4: An overview of the BEV feature uncertainty modeling direction. The diagram illustrates how uncertainty estimates are computed.

$\sigma_{i,j}^{2(obs)}, \sigma_{i,j}^{2(pred)} \in \mathbb{R}^C$ denote the covariance (uncertainty) of two features (BEV_f^{obs} and BEV_f^{pred}) along the channel dimension. For each BEV grid location, an uncertainty coefficient can be computed based on the feature variance, which is then used to form the weight matrices for the two sets of BEV features. The calculation formula is as follows:

$$k_{i,j} = \frac{\sigma_{i,j}^{2(obs)}}{\sigma_{i,j}^{2(obs)} + \sigma_{i,j}^{2(pred)}} \in \mathbb{R}^C, \tag{4}$$

$$K = [k_{0,0}, ..., k_{i,j}, k_{W-1,H-1}]_{W\times H}, \tag{5}$$

where $k_{i,j}$ denotes the uncertainty coefficient at BEV position (i, j). The weight matrix K composed of elements $k_{i,j}$ forms a $W \times H$ matrix that serves as the weighting matrix for constructing both the predicted BEV features and the observed BEV features. Then, let BEV_f^t denote the fused BEV feature at time step t, based on which the uncertainty-aware

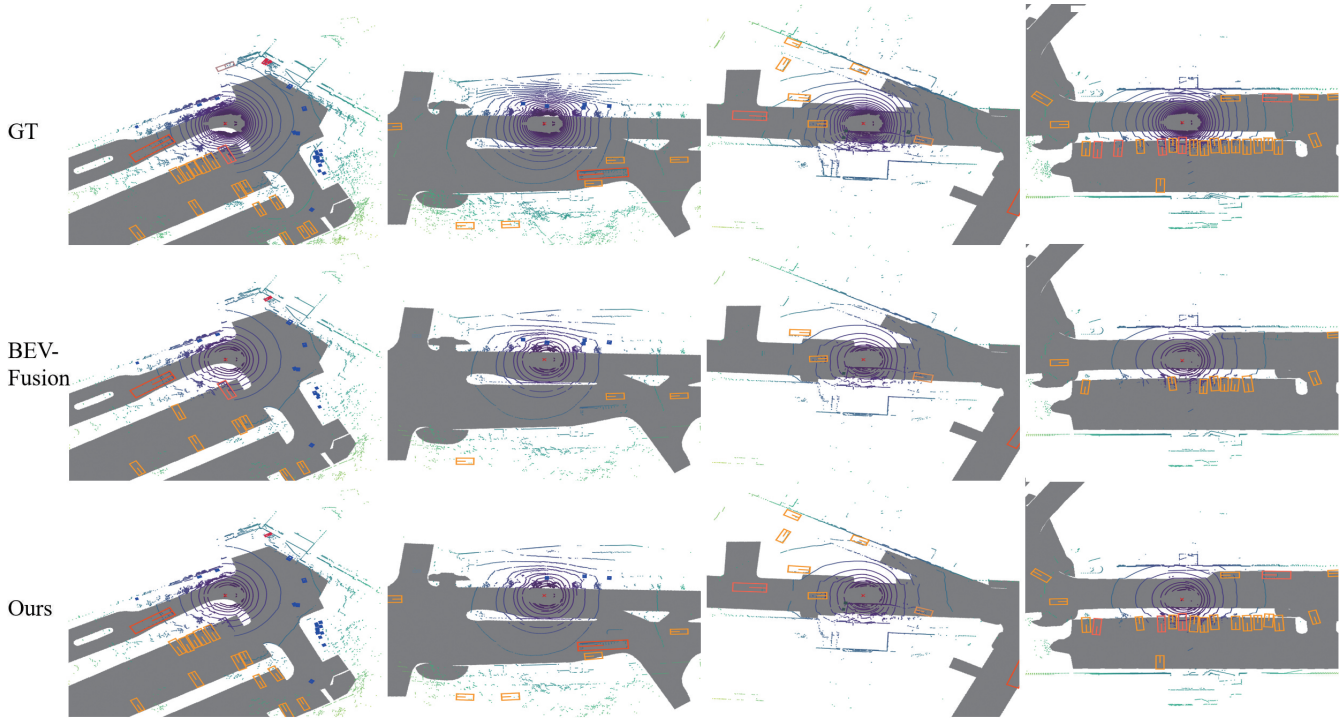

Figure 5: Visualization results on the nuScenes dataset. Top: The 32-beam point cloud and ground truth from the original dataset. Middle: Detection results of the baseline method on nuScenes-16. Bottom: Detection results of our method on nuScenes-16. The visualization results demonstrate that our method achieves more accurate detection results compared to the baseline.

fusion process is defined as:

$$BEV_f^t = BEV_f^{pred} + K \odot (BEV_f^{obs} - BEV_f^{pred}), \quad (6)$$

where $\odot$ denotes element-wise multiplication. Finally, BEV_f^t serves as the final feature representation and is utilized for the downstream 3D regression task.

Experiments

Datasets and Implementation Details

In this paper, we use two fundamental datasets: nuScenes and KITTI. nuScenes is a multimodal dataset for autonomous driving, providing 32-beam LiDAR and images from six cameras across 1,000 scenes (700 train / 150 val / 150 test), each lasting about 20 seconds with a 20 Hz LiDAR frequency. KITTI is a standard benchmark with 64-beam LiDAR and images from diverse driving scenarios. It contains 7,481 annotated training frames and 7,518 test frames with 3D bounding boxes for cars, pedestrians, and cyclists.

To simulate low-beam LiDAR inputs, we downsample the nuScenes and KITTI datasets following the LiDAR Distillation strategy. The original 32-beam nuScenes data is reduced to 16 beams (nuScenes-16), while KITTI data undergoes a two-stage downsampling from 64 to 32 beams (KITTI-32), and then to 16 beams (KITTI-16). The detailed dataset settings, evaluation metrics, and experimental results are provided in the supplementary materials.

Our approach builds upon the open-source BEVFusion framework licensed under MIT. In this setup, the image pipeline utilizes Swin-T (Liu et al. 2021) as the backbone for extracting image features, whereas VoxelNet (Zhou and Tuzel 2018) is adopted for point cloud feature extraction. Following the original design, multi-view images are resized to 256×704 pixels in the image branch, where pixel-level depth is predicted using the Lift-Splat-Shoot module. For the LiDAR branch, we maintain the detection area within $[-54m, 54m]$ along both the x and y directions, and between $[-5m, 3m]$ in the z direction, employing a voxel resolution of 0.075 m.

Methods	Modality	mAP ↑	NDS ↑
LiDAR Distillation (Wei et al. 2022)	L	61.5	65.4
Bi3D (Yuan et al. 2023)	L	62.9	66.2
TransFusion (Bai et al. 2022)	C+L	47.2	52.3
SimpleBEV (Zhao et al. 2024)	C+L	60.7	66.2
BEVDistill (Chen et al. 2022)	C+L	59.8	63.4
BEVFusion† (Liu et al. 2023)	C+L	62.9	67.6
CLIX^{3D} (Hegde et al. 2025)	C+L	65.8	68.4
Ours	C+L	**67.1**	**70.3**

Table 1: Experimental results on the nuScenes-16 dataset. † indicates the baseline model used in our experiments.

Comparative Experiments

We compare our proposed method with several representative approaches, including LiDAR Distillation, Bi3D, TransFusion, SimpleBEV, BEVDistill, BEVFusion, and CLIX^{3D}. As presented in Table 1 and Table 2, our method consistently

Methods	Modality	Beam	Easy	Moderate	Hard	Average ↑
LiDAR Distillation (Wei et al. 2022)	L	32	86.00	70.15	66.86	74.34
		16	80.21	59.87	55.32	65.13
Bi3D (Yuan et al. 2023)	L	32	87.45	71.33	67.10	75.29
		16	81.24	62.57	56.23	66.68
TransFusion (Bai et al. 2022)	C+L	32	72.45	60.30	50.46	61.07
		16	66.76	54.17	44.27	55.07
SimpleBEV (Zhao et al. 2024)	C+L	32	85.59	71.75	63.78	73.71
		16	78.08	54.33	55.64	62.68
BEVDistill (Chen et al. 2022)	C+L	32	78.92	66.58	59.20	68.23
		16	72.36	60.21	50.41	60.99
BEVFusion† (Liu et al. 2023)	C+L	32	81.20	66.58	59.43	69.07
		16	76.25	54.17	50.49	60.30
CLIX^{3D} (Hegde et al. 2025)	C+L	32	87.81	72.45	67.45	75.90
		16	82.16	63.16	56.98	67.43
Ours	C+L	32	89.32	73.30	68.96	**77.19**
		16	82.76	66.46	58.63	**69.28**

Table 2: Experimental results on the KITTI-32/KITTI-16 dataset. † indicates the baseline model used in our experiments.

Description	TFPL	SFOL	UAF	mAP	NDS
`using baseline`				62.9	67.6
`using pred`	✓			65.7	68.4
`using obs`		✓		65.8	68.7
`using obs+pred`	✓	✓		66.0	68.9
`sing uaf`	✓	✓	✓	67.1	70.3

Table 3: Ablation study on key components of the proposed framework.

Number of Historical Frames	mAP	NDS
n=2	66.2	68.8
n=3	66.7	69.6
n=4	67.1	70.3
n=5	67.0	70.1
n=6	66.8	69.5

Table 4: Performance impact of incorporating different numbers of historical-time BEV features.

outperforms all baselines on the nuScenes-16 and KITTI-16/KITTI-32 datasets. These results indicate that incorporating temporal-spatial features and uncertainty-aware fusion not only enhances fine-grained perception but also effectively mitigates the performance degradation commonly observed in previous distillation-based and fusion approaches.

To further demonstrate the effectiveness of our method, we provide qualitative visualizations on the nuScenes-16 dataset, as shown in Fig. 5. We adopt the state-of-the-art multi-modal 3D detection framework BEVFusion as our baseline. Compared to BEVFusion, our method exhibits significantly improved robustness and stability under the highly sparse nuScenes-16 setting.

Ablation Study

We conduct ablation studies on the nuScenes-16 dataset to evaluate the contributions of the TFPL module, the SFOL module, and the UAF strategy. As shown in Table 3, five configurations are tested: (1) `using baseline` directly adopts the low-beam feature BEV_f^{low}; (2) `Using pred` employs BEV_f^{pred} from the TFPL module; (3) `using obs` utilizes BEV_f^{obs} from the SFOL module; (4) `using obs+pred` fuses BEV_f^{obs} and BEV_f^{pred} via element-wise addition; and (5) `using uaf` applies the UAF strategy to adaptively combine both into BEV_f^t. The performance improves consistently with each component. Incorporating TFPL, SFOL, and the UAF strategy achieves the best results, reaching 67.1 mAP and 70.3 NDS. This improvement mainly comes from the complementary strengths of TFPL, SFOL, and the adaptive fusion of UAF, which together enhance feature representation and detection robustness.

To investigate the Learning from Predictions process, we incorporate historical-time BEV features as input to the Temporal Feature Prediction Learning (TFPL) module. The parameter n, denoting the number of historical frames, controls the temporal receptive field. To assess the influence of temporal history length on prediction accuracy, we perform ablation studies on the nuScenes-16 dataset, varying n from 2 to 6. As shown in Table 4, the best performance is achieved when n=4, indicating a balanced trade-off between temporal context and feature redundancy.

Conclusion

This paper presents a novel temporal and spatial representation learning framework for multimodal 3D object detection under low-beam LiDAR configurations. The proposed approach incorporates two key modules: TFPL, which captures temporal dynamics by forecasting current BEV representations from historical observations, and SFOL, which enhances spatial representations by aligning low-beam features with high-beam distributions. These modules are further integrated through a UAF strategy that models feature-level uncertainty and adaptively fuses temporal and spatial cues. Extensive experiments on KITTI and nuScenes with low-beam LiDAR demonstrate the framework's effectiveness and its potential as a robust, cost-effective solution for autonomous driving.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Projects 62576206 and 62476089, the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education, and the Fundamental Research Funds for the Central Universities.

References

Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; and Tai, C.-L. 2022. TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1090–1099.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11621–11631.

Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; and Zhao, F. 2022. BEVDistill: Cross-modal BEV distillation for multi-view 3D object detection. *arXiv preprint arXiv:2211.09386*.

Chiu, H.-K.; Wang, C.-Y.; Chen, M.-H.; and Smith, S. F. 2024. Probabilistic 3D multi-object cooperative tracking for autonomous driving via differentiable multi-sensor kalman filter. In *IEEE International Conference on Robotics and Automation*, 18458–18464.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3354–3361.

Griesbacher, C.; and Fruhwirth-Reisinger, C. 2025. An investigation of beam density on LiDAR object detection performance. *arXiv preprint arXiv:2503.15087*.

Hegde, D.; Lohit, S.; Peng, K.-C.; Jones, M.; and Patel, V. 2025. Multimodal 3D object detection on unseen domains. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2499–2509.

Hu, Q.; Liu, D.; and Hu, W. 2023. Density-insensitive unsupervised domain adaption on 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17556–17566.

Huang, M. 2024. Overview of solid-state LiDAR for robotics. In *Fourth International Conference on Optics and Communication Technology*, volume 13398, 81–85.

Li, S.; Ma, L.; and Li, X. 2024. Domain generalization of 3D object detection by density-resampling. In *European Conference on Computer Vision*, 456–473.

Li, Z.; Lan, S.; Alvarez, J. M.; and Wu, Z. 2024. BEVNeXt: Reviving dense bev frameworks for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20113–20123.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D. L.; and Han, S. 2023. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation*, 2774–2781.

Lu, H.; Zhang, Y.; Wang, G.; Lian, Q.; Du, D.; and Chen, Y.-C. 2025. Towards generalizable multi-camera 3D object detection via perspective rendering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5811–5819.

Lu, H.-C.; Lin, C.-Y.; and Hsu, W. H. 2025. Improving generalization ability for 3D object detection by learning sparsity-invariant features. *arXiv preprint arXiv:2502.02322*.

Philion, J.; and Fidler, S. 2020. Lift, Splat, Shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In *European Conference on Computer Vision*, 194–210.

Qi, C. R.; Liu, W.; Wu, C.; Su, H.; and Guibas, L. J. 2018. Frustum pointnets for 3D object detection from RGB-D data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 918–927.

Ravindran, R.; Santora, M. J.; and Jamali, M. M. 2022. Camera, LiDAR, and Radar sensor fusion based on Bayesian neural network (CLR-BNN). *IEEE Sensors Journal*, 22(7): 6964–6974.

Shan, J.; Zhang, G.; Tang, C.; Pan, H.; Yu, Q.; Wu, G.; and Hu, X. 2023. Focal distillation from high-resolution data to low-resolution data for 3D object detection. *IEEE Transactions on Intelligent Transportation Systems*, 24(12): 14064–14075.

Vora, S.; Lang, A. H.; Helou, B.; and Beijbom, O. 2020. PointPainting: Sequential fusion for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4604–4612.

Wang, C.; Ma, C.; Zhu, M.; and Yang, X. 2021. PointAugmenting: Cross-modal augmentation for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11794–11803.

Wang, L.; Zhang, X.; Song, Z.; Bi, J.; Zhang, G.; Wei, H.; Tang, L.; Yang, L.; Li, J.; Jia, C.; et al. 2023. Multi-modal 3D object detection in autonomous driving: A survey and taxonomy. *IEEE Transactions on Intelligent Vehicles*, 8(7): 3781–3798.

Wei, Y.; Wei, Z.; Rao, Y.; Li, J.; Zhou, J.; and Lu, J. 2022. LiDAR Distillation: Bridging the beam-induced domain gap for 3D object detection. In *European Conference on Computer Vision*, 179–195.

Xu, J.; Song, C.; Shi, C.; Liu, H.; and Wang, Q. 2025. UncertainBEV: Uncertainty-aware BEV fusion for roadside 3D object detection. *Image and Vision Computing*, 159: 105567.

Yu, K.; Tao, T.; Xie, H.; Lin, Z.; Liang, T.; Wang, B.; Chen, P.; Hao, D.; Wang, Y.; and Liang, X. 2023. Benchmarking the robustness of LiDAR-camera fusion for 3D Object

Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3188–3198.

Yuan, J.; Zhang, B.; Yan, X.; Chen, T.; Shi, B.; Li, Y.; and Qiao, Y. 2023. Bi3D: Bi-domain active learning for cross-domain 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15599–15608.

Zhang, J.; Zhang, Y.; Qi, Y.; Fu, Z.; Liu, Q.; and Wang, Y. 2025. GeoBEV: Learning geometric BEV representation for multi-view 3D object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9960–9968.

Zhao, Y.; Gong, Z.; Zheng, P.; Zhu, H.; and Wu, S. 2024. SimpleBEV: Improved LiDAR-camera fusion architecture for 3D object detection. *arXiv preprint arXiv:2411.05292*.

Zhou, Y.; and Tuzel, O. 2018. VoxelNet: End-to-end learning for point cloud based 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4490–4499.