electronics

MDPI

Article

Ego-Motion-Aware Temporal Fusion in BEV Space for Multi-Modal 3D Object Detection

Na Zhang *, Edmundo Guerra and Antoni Grau

School of Industry Engineering, Polytechnic University of Catalonia, 08028 Barcelona, Spain; edmundo.guerra@upc.edu (E.G.); antoni.grau@upc.edu (A.G.)
* Correspondence: na.zhang@upc.edu

Abstract

Multi-modal 3D object detection is critical for autonomous driving perception. While Bird's Eye View (BEV) fusion methods effectively integrate LiDAR and camera features, they primarily focus on single-frame fusion and neglect temporal context. We propose CamT-BEV, a camera-temporal-enhanced BEV fusion framework for improved multi-modal 3D object detection. Our key insight is that temporal modeling is particularly critical for the camera branch to resolve monocular depth ambiguity and object occlusion, while single-frame LiDAR representation already provides accurate instantaneous geometry. We thus propose a camera-centric temporal enhancement module via ego-motion warping and ConvLSTM temporal encoding. Extensive experiments on the nuScenes dataset demonstrate that CamT-BEV achieves competitive perception performance, attaining 0.6971 NDS and 0.6683 mAP, with notable relative AP gains on challenging categories such as bicycles (+27.3%) and motorcycles (+7.66%) evaluated under category-level mAP (averaged across 0.5 m to 4.0 m distance thresholds). Furthermore, evaluations under fog and miss-beam conditions in nuScenes-C confirm its improved robustness against specific visual and sensor degradations. Crucially, these gains are achieved with low additional computational and memory overhead, demonstrating that targeted camera-temporal fusion is a practical solution for 3D perception.

Keywords: computer vision; intelligent transportation systems; autonomous driving; multi-sensor fusion

1. Introduction

Accurate 3D object detection is essential for autonomous driving systems, enabling vehicles to perceive and localize objects in three-dimensional space for safe navigation and path planning. Modern autonomous vehicles typically employ multiple sensors, including LiDAR and cameras, each with complementary strengths and limitations.

LiDAR sensors provide precise depth measurements and geometric structures but lack rich semantic information. Cameras capture dense visual information with strong semantic understanding capabilities but struggle with accurate depth estimation. These complementary characteristics motivate multi-modal fusion approaches that leverage the strengths of different sensors for robust 3D object detection.

Recent Bird's Eye View (BEV)-based methods have shown promising results by transforming multi-modal features into a unified top-down representation. Methods such as BEVFusion [1] and TransFusion [2] achieve highly competitive performance through effective BEV space fusion. However, these methods primarily focus on single-frame fusion and do not fully exploit temporal information available in sequential driving scenarios.

Academic Editor: Aryya Gangopadhyay

Received: 11 August 2026
Revised: 12 September 2026
Accepted: 14 September 2026
Published: 16 September 2026

While some works have explored temporal modeling, existing approaches typically apply uniform temporal fusion across all modalities. This overlooks the distinct temporal characteristics of different sensors: LiDAR provides consistent geometric measurements across frames, while camera observations are more sensitive to lighting and viewpoint variations. Different modalities may therefore benefit from different temporal aggregation strategies.

Motivated by these intrinsic sensor differences, we focus on bolstering the camera branch using temporal context, while using LiDAR as a reliable per-frame geometric anchor. Specifically, camera projections suffer heavily from depth ambiguities and transient occlusions—weaknesses that multi-frame temporal consensus can effectively alleviate.

The main contributions of this work are:

- We propose CamT-BEV, a camera-temporal-enhanced BEV fusion framework specifically designed to enhance camera-side representations for multi-modal 3D object detection.
- We design the CameraTemporalFusion module for CamT-BEV, which combines ego-motion-based alignment with ConvLSTM temporal encoding for efficient multi-frame camera feature aggregation.
- We conduct extensive experiments on nuScenes, demonstrating that CamT-BEV achieves superior performance over BEVFusion (yielding +0.0154 NDS and +0.0302 mAP improvements), with significant gains on challenging object categories while maintaining an identical memory footprint.
- We provide comprehensive analyses validating the effectiveness of CamT-BEV's camera-temporal fusion for robust 3D object detection in autonomous driving scenarios.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed CamT-BEV method. Section 4 describes experiments and results. Section 5 concludes the paper.

2. Related Work

2.1. 3D Object Detection

3D object detection is a fundamental perception task in autonomous driving systems, aiming to accurately localize and classify objects in three-dimensional space. While early research laid the foundation using basic point cloud processing, recent paradigms have shifted towards computational efficiency and semantic scalability in complex, long-tail driving environments [3].

LiDAR-based methods directly process point cloud data to extract geometric features. Building upon established voxelization techniques, PointPillars [4] and CenterPoint [5] remain standard benchmarks due to their real-time performance and efficient anchor-free detection heads. However, as LiDAR resolution increases to 128 or 256 lines, the computational burden of processing raw points grows exponentially. Consequently, recent edge-assisted frameworks have emerged to optimize the delay–accuracy trade-off through dynamic offloading algorithms [6], allowing vehicles to delegate heavy geometric computations to roadside units (RSUs). Furthermore, research in 2025 has explored Mamba-based point cloud encoders [7], which leverage linear-complexity state-space models to process ultra-dense sequences without the quadratic computational overhead of standard Transformers. These models excel in capturing long-range spatial dependencies in unstructured point clouds. Despite their geometric precision, LiDAR systems remain inherently limited by data sparsity at long ranges and a lack of fine-grained semantic understanding (e.g., distinguishing between a traffic sign and a human-shaped billboard).

Camera-based approaches leverage rich visual semantics from RGB images. The dominance of spatial–temporal Transformers, exemplified by BEVFormer [8], has paved the way for unified Bird's Eye View (BEV) representations, which project perspective-view features into a top-down spatial grid. Entering 2025–2026, the research focus has evolved toward Open-Vocabulary 3D Detection (OVD) [9] and Foundation Models for Perception [10]. These systems utilize vision–language pre-training (e.g., aligning image features with text embeddings from CLIP-like models) to recognize long-tail object categories beyond fixed label sets, such as rare construction vehicles or specific animal species, significantly enhancing the safety of autonomous systems in unmapped urban scenarios.

2.2. Multi-Modal Sensor Fusion

To take advantage of complementary strengths, LiDAR precise depth detection and Camera's multi-modal dense semantic information fusion have become the standard paradigm. Recent trends have moved beyond simple feature concatenation toward sophisticated alignment and noise-resilient reasoning [11].

Intermediate and BEV-based fusion methods, such as BEVFusion, have established a unified spatial representation by transforming both modalities into the BEV space. However, traditional fusion often suffers from "modality collapse" when one sensor provides degraded data, such as camera overexposure or LiDAR interference in heavy rain. To address this, SeBFusion [12] introduced a Semantic-Enhanced Bidirectional framework that adaptively reweights modality confidence in real time, preventing noisy sensors from polluting the latent feature space. Additionally, DeepInteraction [13] proposes a scene-aware interaction mechanism that adaptively prioritizes dynamic multi-modal representations, reducing computational redundancy by focusing fusion on critical target-relevant regions. Furthermore, the UniTrans framework [14] has demonstrated that cross-modal alignment can be achieved with a parameter-efficient design, utilizing shared weight modules for heterogeneous sensors, which drastically reduces training costs while maintaining highly competitive accuracy on global benchmarks like nuScenes.

2.3. Temporal Modeling and 4D Perception

Integrating temporal context is critical for mitigating occlusion artifacts and capturing dynamic motion patterns in autonomous driving perception. Modern BEV detection methods have evolved from naive frame stacking to sophisticated spatiotemporal modeling for comprehensive scene understanding. Camera-only temporal methods such as BEVDet4D [15] and BEVFormer [8] adopt temporal self-attention or cross-attention to enforce cross-frame consistency. Beyond vision-based solutions, diffusion-based temporal modeling and 4D occupancy prediction [16] further advance scene motion estimation and holistic environmental representation. Despite these progressions, multi-modal temporal fusion still suffers from inherent sensor heterogeneity, including mismatched frame rates and asynchronous sampling between cameras and LiDAR. Existing LiDAR-camera-temporal methods, such as LIFT [17] and BEVFusion4D, adopt symmetric temporal aggregation on both sensor branches. However, temporal modeling on LiDAR features easily yields dynamic ghost artifacts and increases computational overhead.

To address these camera-specific limitations, this work proposes a dedicated camera-temporal fusion module. We perform ego-motion alignment and ConvLSTM-based temporal encoding exclusively on the camera branch, leveraging long-range visual temporal context to compensate for monocular depth uncertainty and visual occlusions, while keeping the LiDAR branch lightweight and instantaneous.

3. Proposed Method

3.1. CamT-BEV

Figure 1 presents the overall architecture of the proposed CamT-BEV framework. As illustrated, CamT-BEV follows a BEV-centric multi-modal fusion design and explicitly incorporates temporal modeling in the camera branch via a dedicated CameraTemporalFusion module.

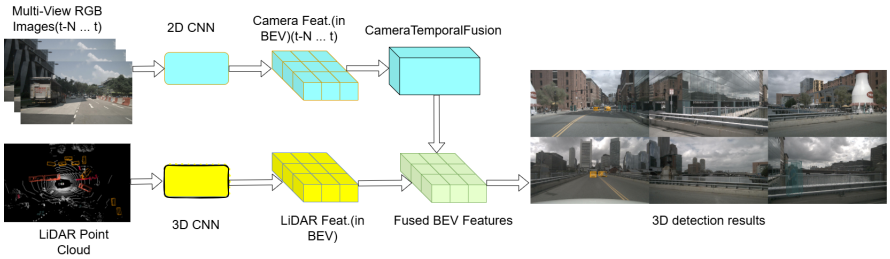


Figure 1. The overall structure of the CamT-BEV framework. Cyan blocks represent the camera processing stream and temporal fusion module; yellow blocks denote the LiDAR point cloud feature extraction stream; and the light green block represents the fused BEV features used for predicting 3D detection results.

As shown in Figure 1, the camera branch takes multi-view RGB images from the current frame and several preceding frames as input. After feature extraction by a shared 2D convolutional backbone, image features are lifted and projected into the BEV space, resulting in a sequence of camera BEV feature maps across time. These multi-frame BEV features are then fed into the proposed CameraTemporalFusion module, which aligns historical features with the current frame and aggregates temporal information to produce a temporally enhanced camera BEV representation.

In parallel, the LiDAR branch encodes point cloud data into BEV feature maps using a 3D convolutional network, providing precise geometric cues. The output of the CameraTemporalFusion module is subsequently fused with LiDAR BEV features in the BEV space to form unified fused BEV representations, which are finally passed to a BEV-based detection head for 3D object prediction.

3.2. CameraTemporalFusion

Given a sequence of multi-view images, our goal is to exploit temporal information to enhance BEV representations for 3D object detection. As illustrated in Figure 2, each frame is first independently transformed into the BEV space. To enable effective temporal fusion, we explicitly align historical BEV features to the current frame using ego-motion. A lightweight BEV feature alignment module is further applied to handle dynamic objects.

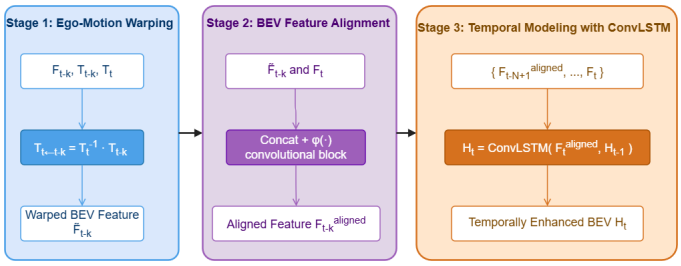


Figure 2. Pipeline of the CameraTemporalFusion module. Stage 1 (blue) calculates ego-motion warping; Stage 2 (purple) aligns the BEV features; Stage 3 (orange) models temporal context with ConvLSTM.

Finally, a ConvLSTM-based temporal encoder aggregates multi-frame BEV features to produce a temporally enhanced BEV representation.

3.2.1. Ego-Motion-Based BEV Warping

Let $\mathbf{F}_{t-k} \in \mathbb{R}^{C \times H \times W}$ denote the BEV feature at time $t-k$, and $\mathbf{F}_t$ denote the current BEV feature. Given the ego-to-global transformation matrices $\mathbf{T}_{t-k}$ (ego at $t-k$ to global) and $\mathbf{T}_t$ (ego at t to global), the relative transformation mapping points from the previous ego coordinate frame to the current ego frame is computed as follows:

$$\mathbf{T}_{t \leftarrow t-k} = \mathbf{T}_t^{-1} \mathbf{T}_{t-k}. \tag{1}$$

Each BEV grid point $\mathbf{p}_{t-k} = (x, y, 0, 1)^\top$ defined in the previous ego coordinate system is transformed into the current vehicle coordinate system:

$$\mathbf{p}_t = \mathbf{T}_{t \leftarrow t-k} \mathbf{p}_{t-k}. \tag{2}$$

We adopt the vehicle-centric coordinate system where x points forward, y points left, and z points upward. Equation (1) is strictly consistent with our implementation. The transformed coordinates are normalized to $[-1, 1]$ and used to sample historical BEV features via bilinear interpolation with `align_corners = False`. This ego-motion-based warping explicitly enforces geometric consistency across frames without introducing additional learnable parameters.

3.2.2. BEV Feature Alignment

Although ego-motion warping accurately aligns static regions, misalignment may still occur in dynamic objects. To mitigate this issue, we introduce a BEV feature alignment module. Specifically, the warped historical feature $\tilde{\mathbf{F}}_{t-k}$ is concatenated with the current feature $\mathbf{F}_t$ and processed by a convolutional block:

$$\mathbf{F}_{t-k}^{\text{aligned}} = \phi\big([\tilde{\mathbf{F}}_{t-k}, \mathbf{F}_t]\big), \tag{3}$$

where $\phi(\cdot)$ denotes a convolution layer followed by normalization. This module enables adaptive fusion of temporal information and improves robustness to dynamic scenes.

3.2.3. Temporal Modeling with ConvLSTM

After alignment, a sequence of BEV features from multiple frames is fed into a ConvLSTM-based temporal encoder. ConvLSTM preserves spatial structure by replacing fully connected operations with convolutions, making it suitable for BEV feature maps.

Given aligned features $\{\mathbf{F}_{t-N+1}^{\text{aligned}}, \ldots, \mathbf{F}_t\}$, the hidden state is updated recurrently:

$$\mathbf{H}_t = \text{ConvLSTM}(\mathbf{F}_t^{\text{aligned}}, \mathbf{H}_{t-1}). \tag{4}$$

The final hidden state $\mathbf{H}_t$ serves as the temporally enhanced BEV representation for downstream 3D detection. Compared with attention-based temporal fusion methods, our approach achieves efficient temporal modeling with significantly lower computational complexity.

3.3. Motivation and Design Rationale for Camera-Side Temporal Fusion

The design rationale for prioritizing temporal fusion on the camera branch stems from the fundamental physical differences between modalities:

Camera branch vulnerabilities: perspective-to-BEV transformation in monocular/multi-view cameras inherently suffers from depth ambiguity, visual occlusions, and sudden lighting changes. Multi-frame temporal consensus effectively provides implicit constraints for depth lifting and accumulates weak visual cues over time.

LiDAR branch characteristics: point clouds supply precise, explicit 3D geometry per frame. Standard ego-motion compensation on LiDAR BEV features often introduces motion blur or dynamic ghost artifacts for moving obstacles without fine-grained flow estimation. Therefore, keeping the LiDAR branch per-frame avoids excessive computational overhead and ghosting risks, allowing it to act as an instantaneous geometric anchor.

4. Experiment and Analysis

4.1. Hardware and Software Environment

All experiments are conducted on a Linux server equipped with two NVIDIA H100 GPUs (64 GB VRAM each), an Intel Xeon CPU, and 128 GB system memory. The proposed model is implemented based on Python 3.8.20, PyTorch 2.0.1, CUDA 11.8, and cuDNN 8.7. Multi-GPU training is accelerated via PyTorch Distributed Data Parallel (DDP) with the NCCL backend. All experiments are built on MMEngine 0.10.7 and OpenCV 4.12.0. To account for training variability and ensure statistical reproducibility, experiments are conducted across three independent random seeds (42, 100, and 2024), with the mean and standard deviation reported.

4.2. Network and Input Configuration

The model adopts Swin-Tiny as the camera backbone and a SECOND sparse encoder combined with a GeomAware voxel feature encoder as the LiDAR backbone. The input image resolution is fixed to 256×704. The LiDAR voxel size is set to $[0.075\,\text{m}, 0.075\,\text{m}, 0.2\,\text{m}]$, and the point cloud and BEV spatial range is $[-54.0\,\text{m}, -54.0\,\text{m}, -5.0\,\text{m}, 54.0\,\text{m}, 54.0\,\text{m}, 3.0\,\text{m}]$, resulting in a BEV feature resolution of 360×360.

4.3. Temporal Module Settings

Our temporal fusion module aggregates temporal features from four historical frames ($N = 4$) following the native timestamp interval of the nuScenes dataset. The embedded ConvLSTM is configured with 80 hidden dimensions ($C = 80$) and two layers.

For the BEV feature alignment module, historical BEV features are first warped into the current ego coordinate system via ego-motion transformation. To refine dynamic misalignments, the warped feature $\tilde{F}_{t-k}$ and current feature F_t are concatenated along the channel dimension, yielding an input dimension of $2C$ (160 channels). The alignment network $\phi(\cdot)$ consists of two sequential 3×3 convolutional layers: the first convolution projects the 160-channel concatenated feature map to an intermediate dimension of 128 channels, and the second convolution reduces it back to $C = 80$ channels. Each convolutional layer is followed by Synchronized Batch Normalization (SyncBatchNorm) and a ReLU activation function.

4.4. Normalization and Training Hyperparameters

Images are normalized with mean $[123.675, 116.28, 103.53]$ and std $[58.395, 57.12, 57.375]$. The LiDAR branch adopts BN1d and BN2d normalization layers. The model is optimized by AdamW with an initial learning rate of 2×10^{-4} and a weight decay of 0.01. The learning rate first undergoes a linear warm-up for 500 iterations and then decays via cosine annealing. The total batch size is 32, and all models are trained for six epochs with gradient clipping (max_norm = 35).

4.5. Dataset and Evaluation Indicators

We conduct experiments on the nuScenes [18] dataset, a large-scale autonomous driving benchmark featuring a full sensor suite including one LiDAR, five radars, six cameras, IMU, and GPS. The dataset consists of 1000 scenes of 20 s each, comprising approximately

1,400,000 camera images with comprehensive 3D object annotations. In this study, we utilize the LiDAR and camera modalities for multi-modal 3D object detection across 10 categories: car, truck, construction vehicle, bus, trailer, barrier, motorcycle, bicycle, pedestrian, and traffic cone. The perception system employs six surround-view cameras providing 360-degree coverage and a 32-beam top-mounted LiDAR sensor.

To systematically evaluate the robustness of our model under real-world sensor corruptions, we further extend our experiments to the nuScenes-C [19] dataset. Specifically, we focus on two representative corruption scenarios: fog, which degrades camera visual quality, and miss-beam, which simulates point cloud sparsity caused by LiDAR sensor beam dropouts.

The nuScenes dataset employs a comprehensive evaluation framework consisting of two primary metrics and five error-based metrics. The mean Average Precision (mAP) measures detection accuracy across all 10 object categories at four distance thresholds (0.5 m, 1.0 m, 2.0 m, and 4.0 m), while the nuScenes Detection Score (NDS) serves as the official ranking metric, combining mAP with five True Positive (TP) error metrics:

$$\mathrm{NDS} = \frac{1}{10}\left[5 \times \mathrm{mAP} + \sum(1 - \min(1, \mathrm{mTP}))\right] \tag{5}$$

The TP metrics include mATE (mean Average Translation Error) measuring localization accuracy as the Euclidean distance between predicted and ground truth centers, mASE (mean Average Scale Error) computed as $1 - \mathrm{IoU}$ after alignment to evaluate bounding box size accuracy, mAOE (mean Average Orientation Error) assessing heading estimation through the smallest yaw angle difference, mAVE (mean Average Velocity Error) calculating the L_2 norm of velocity error for moving objects, and mAAE (mean Average Attribute Error) measuring classification error for object attributes. All metrics are computed independently for each category and then averaged, with lower values indicating better performance for error metrics and higher values preferred for mAP and NDS.

4.6. Comparison and Analysis of Results

Table 1 compares CamT-BEV with representative multi-modal baselines on the nuScenes validation set, using BEVFusion as a primary baseline. To evaluate training stability, results for both BEVFusion and CamT-BEV are reported as the mean ± standard deviation across three independent random seeds. Compared to BEVFusion (0.6817 ± 0.0003 NDS and 0.6381 ± 0.0004 mAP), CamT-BEV achieves superior overall detection performance, yielding an NDS of 0.6971 ± 0.0003 (+0.0154/2.3% relative gain) and an mAP of 0.6683 ± 0.0006 (+0.0302/4.7% relative gain). Crucially, the standard deviations across independent runs remain very low ($\leq$0.0006), confirming that the observed performance gains significantly exceed training variance.

Table 1. Comparison of overall performance indicators on the nuScenes validation set.

METHOD	Sensors	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
Pointpillars [4]	LiDAR	0.49078	0.34327	0.4239	0.2844	0.5293	0.3773	0.1936
SSN [20]	LiDAR	0.49761	0.35167	0.42438	0.28495	0.49674	0.38269	0.19396
BEVFormer [8]	Camera	0.5690	0.4810	0.5820	0.2560	0.3750	0.3780	0.1260
PGD [21]	Camera	0.39335	0.31736	0.7636	0.2668	0.4572	1.2849	0.1658
PointPainting [22]	LiDAR+Camera	0.5810	0.4640	0.3877	0.2712	0.4958	0.2466	0.1114
3D-CVF [23]	LiDAR+Camera	0.4978	0.4217	0.3001	0.2455	0.4576	0.2795	0.12225
IS-Fusion [24]	LiDAR+Camera	0.6849	0.6647	0.2922	0.2708	0.3180	0.3873	0.2060
BEVFusion [1]	LiDAR+Camera	0.6817 ± 0.0003	0.6381 ± 0.0004	0.2830	0.2523	0.3535	0.2967	0.1880
CamT-BEV	**LiDAR+Camera**	$\mathbf{0.6971 \pm 0.0003}$	$\mathbf{0.6683 \pm 0.0006}$	**0.2830**	**0.2559**	**0.3558**	**0.2918**	**0.1842**

Note: ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

Meanwhile, CamT-BEV maintains an identical translation error (mATE of 0.2830). Regarding attribute estimation, CamT-BEV reduces the velocity error (mAVE) from 0.2967 to 0.2918. As a result, the mean Average Attribute Error (mAAE)—a composite metric evaluating non-geometric attributes such as velocity and state—decreases from 0.1880 to 0.1842. Conversely, orientation accuracy is measured by mAOE, where a minor regression is observed (0.3535 to 0.3558), alongside a slight increase in scale error (mASE from 0.2523 to 0.2559). Nevertheless, CamT-BEV achieves the highest NDS and mAP among the compared methods.

To contextualize the benefits of multi-modal integration, we also compare CamT-BEV against single-modality baselines. CamT-BEV noticeably outperforms the camera-only BEVFormer (NDS: 0.5690, mATE: 0.5820) and the LiDAR-only SSN (NDS: 0.4976, mAP: 0.3517). These comparisons illustrate the structural complementarity between modalities, where LiDAR provides strong localization cues and cameras supply visual semantics, which CamT-BEV effectively combines through temporal feature aggregation.

As presented in Table 2, CamT-BEV achieves competitive AP performance compared to the BEVFusion baseline and single-modality approaches across representative object categories and distance thresholds (0.5 m, 1.0 m, 2.0 m, and 4.0 m). Integrating temporal feature alignment provides notable benefits, particularly for small or dynamic objects. For instance, at the 0.5 m distance threshold, bicycle detection improves from 0.4273 to 0.5439 AP (+27.3% relative gain over BEVFusion), and reaches 0.6076 AP at 4.0 m. Similarly, motorcycle detection rises from 0.5822 to 0.6173 AP at 0.5 m, reaching 0.7870 AP at 4.0 m. These gains demonstrate that temporal aggregation captures fine-grained motion patterns of vulnerable road users.

Table 2. AP at different distance thresholds for representative categories.

Category	Distance (m)	Pointpillars	SSN	BEVFormer	PGD	PointPainting	3D-CVF	BEVFusion	CamT-BEV
Car	0.5	0.530	0.681	0.362	0.204	0.664	0.742	0.7932	**0.8059**
	1.0	0.696	0.799	0.645	0.487	0.790	0.839	0.8872	**0.8947**
	2.0	0.741	0.840	0.817	0.718	0.822	0.863	0.9164	**0.9238**
	4.0	0.769	0.856	0.883	0.834	0.841	0.875	0.9274	**0.9341**
Bicycle	0.5	0.050	0.104	0.165	0.097	0.195	0.275	0.4273	**0.5439**
	1.0	0.120	0.125	0.370	0.262	0.244	0.303	0.4599	**0.5840**
	2.0	0.130	0.129	0.519	0.404	0.258	0.317	0.4623	**0.5932**
	4.0	0.160	0.133	0.576	0.491	0.265	0.322	0.4689	**0.6076**
Pedestrian	0.5	0.499	0.686	0.162	0.116	0.642	0.720	0.8474	**0.8596**
	1.0	0.589	0.702	0.461	0.350	0.722	0.735	0.8592	**0.8714**
	2.0	0.633	0.718	0.721	0.580	0.769	0.748	0.8697	**0.8811**
	4.0	0.668	0.744	0.832	0.719	0.796	0.766	0.8803	**0.8907**
Truck	0.5	0.097	0.162	0.102	0.050	0.207	0.289	**0.4204**	0.4194
	1.0	0.222	0.339	0.316	0.204	0.353	0.450	0.5937	**0.6048**
	2.0	0.290	0.427	0.520	0.397	0.425	0.521	0.6714	**0.6845**
	4.0	0.312	0.450	0.632	0.543	0.447	0.540	0.7069	**0.7229**
Construction vehicle	0.5	0.000	0.001	0.003	0.000	0.018	0.009	**0.0530**	0.0527
	1.0	0.012	0.061	0.080	0.030	0.113	0.101	0.2166	**0.2245**
	2.0	0.059	0.150	0.316	0.166	0.230	0.227	0.3426	**0.3808**
	4.0	0.094	0.179	0.517	0.341	0.270	0.298	0.4802	**0.5082**
Bus	0.5	0.074	0.197	0.049	0.033	0.147	0.284	0.4740	**0.5266**
	1.0	0.280	0.419	0.239	0.186	0.362	0.515	0.7263	**0.7615**
	2.0	0.371	0.518	0.491	0.369	0.452	0.567	0.8472	**0.8805**
	4.0	0.403	0.544	0.648	0.553	0.486	0.587	0.8790	**0.8998**
Trailer	0.5	0.004	0.071	0.021	0.000	0.040	**0.164**	0.1316	0.1073
	1.0	0.106	0.301	0.205	0.085	0.286	**0.464**	0.4203	0.4089
	2.0	0.353	0.484	0.584	0.375	0.531	**0.646**	0.5798	0.5867
	4.0	0.471	0.537	**0.772**	0.602	0.633	0.709	0.6705	0.6849
Barrier	0.5	0.091	0.257	0.365	0.268	0.407	0.495	0.5675	**0.6044**
	1.0	0.383	0.507	0.614	0.544	0.613	0.670	0.6699	**0.7014**
	2.0	0.506	0.623	0.738	0.690	0.675	0.724	0.7125	**0.7417**
	4.0	0.574	0.661	0.781	0.742	0.713	0.747	0.7279	**0.7547**
Motorcycle	0.5	0.171	0.308	0.165	0.113	0.325	0.460	0.5822	**0.6173**
	1.0	0.285	0.376	0.427	0.339	0.433	0.521	0.7051	**0.7565**
	2.0	0.313	0.392	0.617	0.524	0.447	0.532	0.7214	**0.7732**
	4.0	0.327	0.396	0.708	0.612	0.457	0.535	0.7310	**0.7870**
Traffic cone	0.5	0.210	0.440	0.449	0.326	0.555	0.609	0.7462	**0.7641**
	1.0	0.289	0.457	0.698	0.581	0.608	0.619	0.7594	**0.7754**
	2.0	0.334	0.480	0.810	0.726	0.640	0.630	0.7784	**0.7934**
	4.0	0.399	0.535	0.855	0.788	0.693	0.659	0.8095	**0.8166**

For larger vehicle categories and static objects, CamT-BEV maintains consistent advantages across varying distance settings. At the 1.0 m threshold, car detection reaches 0.8947 AP, outperforming BEVFusion (0.8872 AP), camera-only BEVFormer (0.6450 AP), and LiDAR-only SSN (0.7990 AP). On challenging classes such as construction vehicles and buses, temporal fusion facilitates better feature continuity, with construction vehicle AP increasing from 0.3426 to 0.3808 at 2.0 m, and bus AP reaching 0.8998 at 4.0 m. For small or occluded stationary objects like traffic cones, CamT-BEV reaches 0.7641 AP at 0.5 m, outperforming BEVFusion (0.7462 AP).

We observe that performance gains are less pronounced in a small subset of sparse object categories under tight distance conditions (e.g., trailers at 0.5 m), where temporal warping may introduce minor noise if historical feature cues are insufficient. Nevertheless, these category-wise evaluations across distance thresholds confirm that CamT-BEV effectively leverages cross-modal complementarity and temporal context for robust 3D object detection.

4.7. Ablation Experiment

To quantitatively validate the effectiveness of the key components in our CamT-BEV model, including the ego-warp transformation, FeatureAlign module, and ConvLSTM temporal modeling branch, we conduct extensive ablation experiments on the nuScenes validation set, as presented in Table 3. Baseline A is the original BEVFusion model, serving as the benchmark for all comparisons.

Table 3. Ablation studies of camera-temporal modules on nuScenes val set.

Method	Ego-Warp	FeatureAlign	ConvLSTM	NDS	mAP
A: BEVFusion Baseline	—	—	—	0.6817	0.6384
B: +History w/o ego-warp	×	✓	✓	0.6742	0.6392
C: ego-warp only	✓	×	×	0.6808	0.6474
D: ego-warp + FeatureAlign	✓	✓	×	0.6809	0.6467
E: w/o FeatureAlign	×	×	✓	0.6892	0.6549
F: CamT-BEV (Ours)	✓	✓	✓	**0.6971**	**0.6683**

Note: ✓ denotes that the module is enabled; × denotes that the module is disabled; — denotes not applicable (baseline setting). Bold text indicates the best performance.

Model B incorporates historical temporal information and the FeatureAlign module while disabling ego-warp; its lower NDS score demonstrates that ego-warp effectively eliminates ego-motion interference and stabilizes temporal fusion. To isolate the individual contribution of FeatureAlign, Model C evaluates an ego-warp-only configuration. Compared to Baseline A, Model C provides a moderate gain (+0.0090 mAP) solely through rigid geometric alignment. Model D further adds FeatureAlign to Model C; however, without the temporal ConvLSTM branch, its gain remains limited. Model E adopts ConvLSTM without spatial and feature alignments, yielding suboptimal performance due to unaligned cross-frame features.

In contrast, our full CamT-BEV model (Model F) integrates all three components. Benefiting from rigid pose correction via ego-warp, fine-grained feature alignment via FeatureAlign, and temporal dependency modeling via ConvLSTM, Model F achieves the best performance with an NDS of 0.6971 and an mAP of 0.6683. These results verify the necessity and complementary synergy of all proposed modules.

We further analyze the computational overhead of our CamT-BEV, and the efficiency metrics are summarized in Table 4. Model weight size is calculated from the saved checkpoint on disk, and peak VRAM is recorded during inference with a batch size of 1 per GPU using standard PyTorch CUDA memory tracking functionality. Compared with BEVFusion,

CamT-BEV brings a minor increase of 0.46 M parameters. Although four-frame temporal processing is introduced, the model weight size and peak VRAM consumption remain virtually identical to the baseline after standard rounding. This is because 0.46 M parameters account for only approx 1.8 MB (in FP32) or 0.9 MB (in FP16), which is negligible relative to the overall model size, while temporal feature caching incurs minimal dynamic memory overhead compared to the primary feature extraction backbones and activation tensors. Benefiting from the lightweight design of our temporal modules, the additional latency is merely 4.25 ms. On 4060Ti, the FPS drops slightly by 0.0407, while the inference speed on H100 is unchanged. Overall, CamT-BEV achieves considerable accuracy gains with marginal computational overhead, demonstrating good practical deployment potential.

Table 4. Comparison of computational efficiency on the nuScenes validation set.

METHOD	Params (M)	Weight Size (MB)	Peak VRAM (GB)	Latency (ms)	FPS (4060Ti)	FPS (H100)
BEVFusion	40.80	475	4.89	321.08	3.1145	6.9
CamT-BEV	**41.26**	**475**	**4.89**	**325.33**	**3.0738**	**6.9**
Improvement	+0.46	0.00	0.00	+4.25	−0.0407	0.00

Note: Bold values represent our proposed method (CamT-BEV); italic row indicates the numerical improvement relative to BEVFusion.

We investigate the influence of the number of input historical frames N, and the quantitative results are listed in Table 5. BEVFusion ($N = 1$) serves as the baseline without temporal information. When increasing historical frames from $N = 1$ to $N = 2$, CamT-BEV achieves moderate gains on both NDS and mAP. Further enlarging the history length to $N = 4$ yields the optimal overall performance, bringing +0.0154 NDS and +0.0302 mAP improvements compared with the single-frame baseline. Most categories obtain consistent accuracy improvements, especially for Bicycle (+0.1276), Motor (+0.0486) and Bus (+0.0355), which benefit significantly from temporal cues. Only the Trailer category suffers a tiny performance drop of −0.0036. Note that these category-level improvements are evaluated based on the mAP averaged over four distance thresholds (0.5 m, 1.0 m, 2.0 m, and 4.0 m). These observations demonstrate that multi-frame temporal modeling provides complementary motion information, which is particularly helpful for small-size and moving objects. We adopt $N = 4$ as the default setting for our full model.

Table 5. Detection performance on the nuScenes validation set across different historical frames (N).

METHOD	NDS	mAP	Car	Bicycle	Pedes	Truck	Con.V	Bus	Trailer	Barrier	Motor	Tra.C
BEVFusion ($N = 1$)	0.6817	0.6381	0.8811	0.4546	0.8642	0.5981	0.2731	0.7316	**0.4506**	0.6695	0.6849	0.7734
CamT-BEV ($N = 2$)	0.6860	0.6493	0.8868	0.5171	0.8714	0.5999	0.2669	0.7359	0.4393	0.6835	0.7089	0.7830
CamT-BEV ($N = 4$)	**0.6971**	**0.6683**	**0.8896**	**0.5822**	**0.8757**	**0.6079**	**0.2916**	**0.7671**	0.4470	**0.7006**	**0.7335**	**0.7874**
Improvement ($N = 4$ vs. $N = 1$)	+0.0154	+0.0302	+0.0085	+0.1276	+0.0115	+0.0098	+0.0185	+0.0355	−0.0036	+0.0311	+0.0486	+0.0140

Note: Bold text indicates the best performance across settings; italic text indicates the comparison metric row; blue text highlights the performance differences between $N = 4$ and $N = 1$.

We further conduct robustness experiments on nuScenes-C under the miss-beam corruption, as shown in Table 6. Miss-beam corruption simulates LiDAR beam loss in real-world conditions, which brings severe point cloud degradation. Compared with BEVFusion, CamT-BEV achieves +0.0168 NDS and +0.0327 mAP gains under this noisy setting. Consistent performance improvements can be observed across all object categories. Notably, Bicycle (+0.1346) and Motor (+0.0529) obtain prominent gains, demonstrating that temporal modeling compensates for missing LiDAR observations by leveraging historical motion cues. The results verify that our method possesses stronger robustness against LiDAR beam dropout corruption.

Table 6. Robustness evaluation on nuScenes-C under the miss-beam corruption scenario.

METHOD	NDS	mAP	Car	Bicycle	Pedes	Truck	Con.V	Bus	Trailer	Barrier	Motor	Tra.C
BEVFusion	0.6765	0.6296	0.8717	0.4415	0.8563	0.5916	0.2726	0.7247	0.4432	0.6642	0.6662	0.7645
CamT-BEV	**0.6933**	**0.6623**	**0.8816**	**0.5761**	**0.8693**	**0.6040**	**0.2900**	**0.7592**	**0.4464**	**0.6968**	**0.7191**	**0.7806**
Improvement	+0.0168	+0.0327	+0.0099	+0.1346	+0.0130	+0.0124	+0.0174	+0.0345	+0.0032	+0.0326	+0.0529	+0.0161

Note: Bold values represent the proposed method (CamT-BEV); italic row indicates the numerical improvement relative to BEVFusion.

We further evaluate the robustness under fog corruption from nuScenes-C, as presented in Table 7. Fog corruption simulates adverse weather conditions, which degrades camera imaging quality and interferes with LiDAR point cloud acquisition. Compared with BEVFusion, CamT-BEV obtains +0.0172 NDS and +0.0324 mAP improvements. All categories achieve consistent performance gains. Significant improvements are observed on Bicycle (+0.1280) and Motor (+0.0494), indicating that temporal information can mitigate the adverse impact brought by foggy weather. The results demonstrate that our temporal-enhanced method exhibits better adaptability to bad-weather corruption.

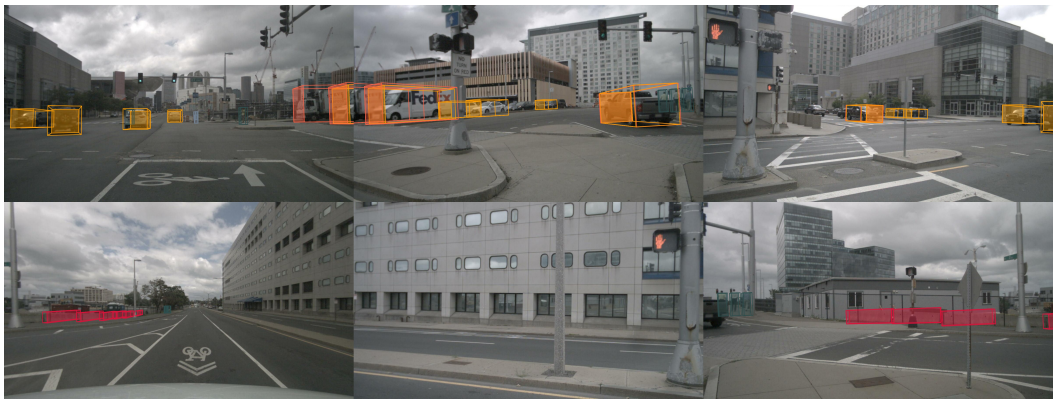
Table 7. Robustness evaluation on nuScenes-C under the fog corruption scenario.

METHOD	NDS	mAP	Car	Bicycle	Pedes	Truck	Con.V	Bus	Trailer	Barrier	Motor	Tra.C
BEVFusion	0.6737	0.6254	0.8615	0.4379	0.8536	0.5836	0.2628	0.7150	0.4383	0.6754	0.6561	0.7704
CamT-BEV	**0.6909**	**0.6578**	**0.8762**	**0.5659**	**0.8671**	**0.6034**	**0.2793**	**0.7507**	**0.4454**	**0.7001**	**0.7055**	**0.7845**
Improvement	+0.0172	+0.0324	+0.0147	+0.1280	+0.0135	+0.0198	+0.0165	+0.0357	+0.0071	+0.0247	+0.0494	+0.0141

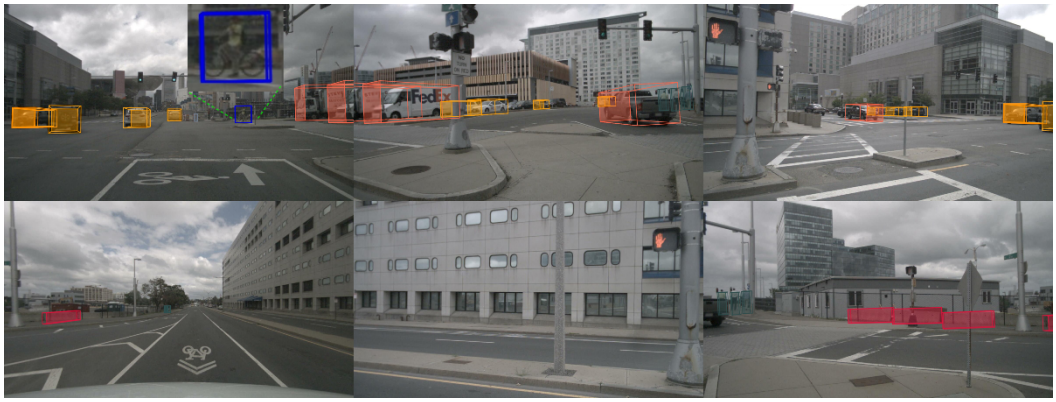
Note: Bold values represent the proposed method (CamT-BEV); italic row indicates the numerical improvement relative to BEVFusion.

4.8. Visualization of Test Results

Figure 3 presents a qualitative comparison between BEVFusion and CamT-BEV detection results in a multi-camera LiDAR fusion system, illustrating the perception results of our proposed method across surround-view scenarios. The visualization displays detection outputs from six surround-view cameras fused with LiDAR data, providing 360-degree environmental coverage for autonomous driving. A critical observation is CamT-BEV's enhanced capability in distinguishing fine-grained object categories that BEVFusion struggles with across this multi-sensor setup. Most notably, as highlighted by the blue bounding box in the upper-left camera view, while BEVFusion detects this object only as a pedestrian, CamT-BEV correctly identifies it as a cyclist with bicycle. This distinction is particularly significant in autonomous driving scenarios, as cyclists and pedestrians exhibit fundamentally different motion patterns and require different prediction strategies for safe navigation. The accurate bicycle detection exemplifies how CamT-BEV's camera-temporal fusion strategy effectively integrates information from multiple camera perspectives and LiDAR point clouds across temporal sequences, enabling the model to discern subtle but critical differences between object categories. By incorporating temporal information across the six-camera and LiDAR fusion framework, our method can leverage motion patterns and sequential appearance changes that are characteristic of cycling behavior, which single-frame fusion approaches like BEVFusion cannot fully exploit. Furthermore, across the complete surround-view panorama, CamT-BEV demonstrates more comprehensive scene understanding with consistent detection quality in all camera fields of view. The barrier detections shown in red across multiple camera views exhibit improved localization accuracy and completeness. These qualitative results corroborate our quantitative findings in Table 2, where CamT-BEV achieved a significant 27.3% relative improvement over BEVFusion for bicycle detection, and validate that the integration of temporal context within multi-camera LiDAR fusion significantly enhances the model's capability to accurately perceive and classify vulnerable road users in complex urban environments.



(**a**) **BEVFusion**



(**b**) **CamT-BEV (Ours)**

Figure 3. Qualitative comparison of surround-view 3D detection results across six cameras on the nuScenes validation set: (**a**) BEVFusion and (**b**) CamT-BEV. The projected 3D bounding boxes are color-coded by category: yellow boxes represent cars, orange boxes represent trucks, blue boxes denote cyclists, and red boxes denote barriers.

Figure 4 further demonstrates CamT-BEV's superior detection capability in challenging scenarios. A striking example appears in the bottom-left camera view, where CamT-BEV successfully detects a distant car that BEVFusion completely misses, as highlighted by the orange bounding box. This detection is particularly challenging due to the car's considerable distance, small appearance in the image, and partial occlusion by trees and buildings in the urban environment. The successful detection showcases CamT-BEV's enhanced long-range perception capability, which is critical for providing earlier awareness of potential traffic participants and enabling better motion planning. This improvement stems from our camera-temporal fusion strategy, which aggregates information across temporal frames to reinforce weak detection signals that single-frame analysis might miss. By leveraging historical observations from the six-camera LiDAR system, CamT-BEV accumulates evidence over time to confidently detect small or partially occluded objects. These qualitative results align with the quantitative improvements in Table 2, particularly the enhanced car detection performance at various distance thresholds, validating that temporal information integration substantially enhances the robustness and completeness of multi-modal 3D object detection in real-world driving scenarios.

(**a**) **BEVFusion**

(**b**) **CamT-BEV (Ours)**

Figure 4. Qualitative comparison of detection performance in challenging long-range and partially occluded scenarios on the nuScenes validation set: (**a**) BEVFusion and (**b**) CamT-BEV. The projected 3D bounding boxes are color-coded by category: yellow for cars, orange for trucks, blue for cyclists, and red for barriers.

4.9. Discussion and Limitations

We further evaluate model robustness under sensor corruptions using the nuScenes-C benchmark. Supplementary tests are carried out on *miss-beam* for simulating LiDAR point sparsity and *fog* for adverse-weather simulation. Results show that our camera-temporal fusion maintains steady performance gains under these corruptions, benefiting from multi-frame visual evidence aggregation. Note that motion blur and low-light night corruption are not provided in nuScenes-C, and robustness evaluation for these scenarios is left for future research.

This work follows the original nuScenes setup and assumes strict frame-level time synchronization between cameras and LiDAR. In practical deployment, timestamp offset and sampling asynchrony can introduce geometric misalignment during ego-motion warping and impair temporal fusion performance. Since no ready-to-use public benchmark supports such evaluation, systematic robustness tests and compensation design for time-asynchrony are reserved for future research.

5. Conclusions

In this paper, we presented CamT-BEV, a camera-temporal-enhanced BEV fusion framework for multi-modal 3D object detection. By incorporating a dedicated camera-side temporal fusion module, CamT-BEV effectively reinforces visual feature representation via ego-motion-based alignment and ConvLSTM aggregation via ego-motion-based alignment

and ConvLSTM aggregation, leaving the LiDAR representation untainted. Comprehensive evaluations on the nuScenes benchmark demonstrate consistent perception gains across challenging categories such as bicycles and motorcycles. Moreover, validation under fog and miss-beam scenarios in nuScenes-C highlights its enhanced robustness against specific visual and sensor corruptions, all while maintaining low additional computational and memory overhead. These findings validate that tailored, modality-specific temporal integration offers an effective and practical paradigm for 3D perception in autonomous vehicles. Future work will explore adaptive temporal modeling based on dynamic scene context, incorporate additional modalities such as radar, and address real-world sensor time-asynchrony.

Author Contributions: N.Z.: Writing—original draft, Validation, Software, Methodology, Investigation; E.G.: Writing—review editing, Conceptualization; A.G.: Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research project was funded by the China Scholarship Council (CSC) under Grant No. 202408440115, and supported by computational resources from the Barcelona Supercomputing Center.

Data Availability Statement: The data presented in this study are openly available in nuScenes at https://www.nuscenes.org/.

Acknowledgments: We acknowledge the computing resources and support provided by the Barcelona Supercomputing Center.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D.L.; Han, S. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 2774–2781.
2. Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; Tai, C.L. TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2022; pp. 1090–1099.
3. Contreras, M.; Jain, A.; Bhatt, N.P.; Banerjee, A.; Hashemi, E. A survey on 3D object detection in real time for autonomous driving. *Front. Robot. AI* **2024**, *11*, 1212070. [CrossRef] [PubMed]
4. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. PointPillars: Fast Encoders for Object Detection from Point Clouds. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12689–12697. [CrossRef]
5. Yin, T.; Zhou, X.; Krahenbuhl, P. Center-Based 3D Object Detection and Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2021; pp. 11784–11793.
6. Zhao, K.; Gou, Z.; Liu, H. Edge-assisted adaptive offloading algorithm for 3D object detection tasks. *PLoS ONE* **2026**, *21*, 0345876. [CrossRef] [PubMed]
7. Liang, D.; Zhou, X.; Xu, W.; Zhu, X.; Zou, Z.; Ye, X.; Tan, X.; Bai, X. Pointmamba: A simple state space model for point cloud analysis. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 32653–32677. [CrossRef]
8. Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Yu, Q.; Dai, J. BEVFormer: Learning Bird's-Eye-View Representation from LiDAR-Camera via Spatiotemporal Transformers. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 2020–2036. [CrossRef] [PubMed]
9. Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T. OpenScene: 3D Scene Understanding with Open Vocabularies. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 815–824. [CrossRef]
10. Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; Li, H. Drivelm: Driving with graph visual question answering. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 256–274.
11. Ye, L. Exploring The Current State of Multimodal Alignment and Fusion. In *Proceedings of the ITM Web of Conferences*; EDP Sciences: Les Ulis, France, 2025; Volume 78, p. 04036.

12. Jiao, T.; Chen, Y.; Feng, X.; Guo, C.; Song, J. Semantic-Enhanced Bidirectional Multimodal Fusion for 3D Object Detection Under Adverse Weather. *Appl. Sci.* **2026**, *16*, 2943. [CrossRef]
13. Yang, Z.; Chen, J.; Miao, Z.; Li, W.; Zhu, X.; Zhang, L. Deepinteraction: 3d object detection via modality interaction. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 1992–2005. [CrossRef]
14. Sun, J.; Chen, K.; He, X.; Liu, X.; Li, K.; Peng, C. Unitrans: Unified parameter-efficient transfer learning and multimodal alignment for large multimodal foundation model. *Comput. Mater. Contin.* **2025**, *83*, 219–238. [CrossRef]
15. Huang, J.; Huang, G. BEVDet4D: Exploit Temporal Cues in Multi-Camera 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1012–1021.
16. Ma, J.; Chen, X.; Huang, J.; Xu, J.; Luo, Z.; Xu, J.; Gu, W.; Ai, R.; Wang, H. Cam4DOcc: Benchmark for Camera-Only 4D Occupancy Forecasting in Autonomous Driving Applications. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 21486–21495. [CrossRef]
17. Zeng, Y.; Ma, C. LIFT: Learning 4D LiDAR-Image Fusion Transformer for 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 17151–17160.
18. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A Multimodal Dataset for Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11618–11628.
19. Dong, Y.; Kang, C.; Zhang, J.; Zhu, Z.; Wang, Y.; Yang, X.; Su, H.; Wei, X.; Zhu, J. Benchmarking Robustness of 3D Object Detection to Common Corruptions in Autonomous Driving. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 1022–1032. [CrossRef]
20. Zhu, X.; Ma, Y.; Wang, T.; Xu, Y.; Shi, J.; Lin, D. Ssn: Shape signature networks for multi-class object detection from point clouds. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 581–597.
21. Wang, T.; Xinge, Z.; Pang, J.; Lin, D. Probabilistic and geometric depth: Detecting objects in perspective. In *Proceedings of the Conference on Robot Learning*; PMLR: New York, NY, USA, 2022; pp. 1475–1485.
22. Vora, S.; Lang, A.H.; Helou, B.; Beijbom, O. PointPainting: Sequential Fusion for 3D Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4603–4611. [CrossRef]
23. Yoo, J.H.; Kim, Y.; Kim, J.; Choi, J.W. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 720–736.
24. Yin, J.; Shen, J.; Chen, R.; Li, W.; Yang, R.; Frossard, P.; Wang, W. IS-Fusion: Instance-Scene Collaborative Fusion for Multimodal 3D Object Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 14905–14915. [CrossRef]