UniVerse: A Unified Voxel-Semantic Ensemble Framework for Multi-Modal 3D Object Detection

Yichen Yao[1], Xiang Qiu[1], and Sixian Chan[2*]

***Abstract*—As a core component of autonomous driving perception systems, 3D object detection faces a critical bottleneck: how to achieve effective joint representation of geometric features and semantic information across heterogeneous modalities (LiDAR point clouds and RGB images). Prevailing multi-sensor fusion methods are prone to cascading errors stemming from sequential optimization processes spanning geometric modeling, cross-modal alignment, and semantic integration. These errors ultimately lead to systematic failures in semantic understanding and undermine robustness. To address this limitation, we propose the UniVerse framework, which systematically cascades a progressive optimization paradigm through "feature disentanglement, alignment enhancement, and deep aggregation." Specifically, the UniVerse introduces a novel triple synergy mechanism: the unified dual-stream perception encoder (GeSe-Encoder) ensures high purity of geometric and semantic features from the source; the cross-modal hierarchical feature fusion module (CHFM) achieves precise alignment through bidirectional enhancement; and the semantic-aware global enhancement module (SAGE) realizes robust semantic understanding in multi-source information interaction, ultimately enabling effective joint modeling of geometric structures and semantic understanding. Finally, extensive experiments on the KITTI dataset validate the effectiveness of our method, achieving 65.63% 3D mAP and outperforming comparable baseline methods by 3.52%, confirming the effectiveness of this cascade optimization strategy.**

***Index Terms*—3D object detection, multi-modal fusion, hierarchical feature fusion, point cloud analysis**

I. INTRODUCTION

With the rapid advancement of sensor technologies and edge computing, autonomous driving is poised to transform transportation systems and improve road safety. Multi-modal 3D object detection is essential for autonomous driving, enabling vehicles to perceive and interact with complex environments [1]–[6]. Current approaches primarily fuse LiDAR and RGB-camera data, combining LiDAR's precise geometric measurements with cameras' semantic information to achieve accurate detection, which is widely applicable.

[1]Yichen Yao and Xiang Qiu are with the College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, Zhejiang, China. {221123030369, qiuxiang}@zjut.edu.cn

[2]Sixian Chan is with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, Zhejiang, China. sxchan@zjut.edu.cn

*Corresponding author. This work is partially supported by the National Key Research and Development Program of China (Grant No. 2024YFC3306902), the National Natural Science Foundation of China (Grant No. U24A20242), the Fundamental Research Funds for the Provincial Universities of Zhejiang (Grant No. RF-A2024013), the Zhejiang Provincial Natural Science Foundation of China (Grant No. LY23F020023, LDT23F02024F02, LZ26F020008), and the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant No. 2025C01152).

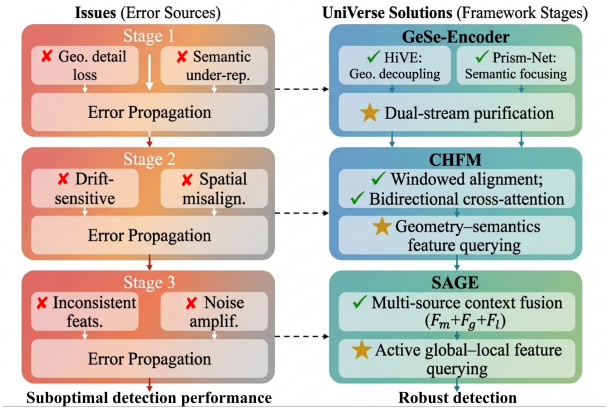

Fig. 1. High-level comparison between existing pipelines (✗) and **UniVerse** (✓, ★). ✓ denotes advantages; ✗ denotes limitations; ★ highlights newly introduced mechanisms. Left: conventional fusion pipelines suffer from error accumulation across feature extraction, alignment, and fusion stages. Right: **UniVerse** suppresses cascade errors via feature disentanglement (HiVE [10] /Prism-Net [8]), windowed bidirectional alignment with position-aware encoding (CHFM/CRAM), and deep aggregation with SAGE.

Despite this complementary strength, effective multi-modal data fusion remains challenging due to cascade error propagation. As shown in Fig. 1, errors originating from imperfect feature extraction propagate sequentially through subsequent processing stages, with each stage's deficiencies compounding the overall detection degradation. Specifically, *in the feature extraction stage*, voxel-based methods [2], [3] improve efficiency but inevitably lose geometric detail, while CNN-based image encoders [7], [8] often miss fine-grained semantics crucial for small objects. *In the alignment stage*, projection-based approaches [1] fail to handle dynamic scenes and calibration errors, causing spatial mismatches between features. *In the fusion stage*, such as EPNet [4] and TransFusion [9] typically use simple concatenation or unidirectional attention, which struggle with noisy features under occlusion and small-object conditions. These stage-wise deficiencies compound over the pipeline, severely limiting final detection performance.

To overcome these challenges, we propose Unified Voxel-Semantic Ensemble Network (**UniVerse**), a progressive 3D detection framework built on three core improvements: dual-stream purification, geometry–semantics feature querying, and active global–local feature querying. The details are as follows:

(1) **GeSe-Encoder** — a unified dual-stream perception architecture extracting Geometry-preserving features and Semantic-refining features, providing high-fidelity inputs for cross-modal alignment.

(2) **CHFM** — a Cross-modal Hierarchical Fusion Module using "divide-and-conquer" and "bidirectional enhancement". Its core unit, Cross-modal Reciprocal Attention Module (CRAM), performs fine-grained, two-way interaction between point cloud and image features within 3D windows, applied hierarchically to global and local features for precise alignment and mutual enhancement.
(3) **SAGE** — a Semantic-Aware Global Enhancement module where fused features actively query a knowledge base formed by global geometric (F_g) and local detail (F_l) features, iteratively refining semantic understanding via self-attention and cross-query mechanisms.

Our main contributions are as follows.

1) We propose a progressive systematic framework named UniVerse, with a "decoupling purification, alignment enhancement, and deep aggregation" cascade design aimed at fundamentally mitigating error propagation in multi-modal 3D detection.
2) We design collaboratively working modules: GeSe-Encoder for parallel purification of geometric and semantic features, CHFM for precise alignment through bidirectional reciprocal attention, and SAGE for robust context aggregation through multi-source interaction.
3) We conduct extensive experiments and ablation studies on the KITTI dataset, verifying the effectiveness of UniVerse and the rationality of its core components in achieving robust detection performance.

II. RELATED WORK

A. Feature Extraction in 3D Object Detection

Early *point-based* methods (PointNet [11]) process point clouds directly with permutation invariance and geometric modeling, but their quadratic cost and limited receptive fields make them impractical for dense urban scenes (>100K points/frame). Aggressive downsampling removes fine structures (*e.g.*, poles, wires, contours), harming small-object detection. Voxel-based methods (VoxelNet [3], SECOND [2]) partition points into regular 3D grids for parallel CNN processing, but fixed resolution causes quantization artifacts and localization drift. Hybrid designs (PV-RCNN [12], CenterPoint [5]) combine voxel efficiency with point precision, yet face feature inconsistency and high memory use. Attention-based models (Fast Point Transformer [13], Point-MAE [14]) improve adaptability but still trade efficiency for geometric fidelity. Our **GeSe-Encoder** extracts *geometry-preserving features* and *semantic-refining features*, providing high-fidelity detail without sacrificing efficiency for cross-modal alignment.

B. Cross-Modal Fusion Mechanism

Projection-based methods (PointPainting [1], PointFusion [15]) map image pixels to 3D points rigidly, offering simplicity but failing under calibration drift (*e.g.*, $2^\circ \Rightarrow \sim$1 m at 30 m), dynamic scenes, and occlusion boundaries. Attention-based fusion (TransFusion [9], EPNet [4]) improves robustness but is mostly unidirectional (image→point), missing reverse flow; global fusion incurs $\mathcal{O}(N^2)$ cost and overlooks local details critical for pedestrians/cyclists. Transformer-based approaches (BEVFusion [6], AutoAlign [16], PrimePSegter [17]) unify features in BEV space, while DeepInteraction [18], UniFormer [19], and SyNet [20] add dynamic interaction and adaptive weights, yet still struggle to balance alignment precision and efficiency. Our **CHFM module** applies *CRAM* within local 3D windows, enabling mutual querying and reducing complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(M^2)$ $(M \ll N)$, preserving both global context and local detail.

C. Semantic Integration Methods

Early integration methods (*e.g.*, MPPNet [21], VISTA [22]) concatenate or average aligned features, but suffer from noise, semantic ambiguity, and scale inconsistency; a distant car may be locally noisy yet clear in a global view. Hierarchical approaches (CrossDTR [23], Unifusion [24]) enforce multi-view consistency but remain passive, amplifying upstream errors under occlusion or poor visibility. Query-based detectors (DETR3D [25], PolarFormer [26]) and temporal models (BEVerse [27]) add learnable queries or multi-frame cues, yet robust fusion under noisy, contradictory evidence remains difficult. Our **SAGE module** builds an *active decision environment* where fused features query global geometric and local detail knowledge bases, turning passive aggregation into targeted acquisition for context-rich semantic representations.

III. METHODOLOGY

A. Overview

As shown in Fig.2, UniVerse employs a cascade progressive optimization strategy comprising three stages: feature purification (GeSe-Encoder), cross-modal alignment (CHFM), and semantic aggregation (SAGE), which systematically addresses error propagation in multi-modal 3D detection.

B. GeSe-Encoder: Geometric and Semantic Encoder

To mitigate error propagation from the source, we propose the **GeSe-Encoder**, a unified dual-stream perception architecture. It performs geometry-preserving feature extraction via the HiVE subnetwork [10] and semantic-refining feature extraction via the Prism-Net subnetwork [8], producing high-fidelity, modality-specific features for subsequent cross-modal alignment.

1) HiVE: Hierarchical Voxel Encoder: To efficiently process sparse, unordered point clouds, we introduce the Hierarchical Voxel Encoder (HiVE), whose core is a geometric decoupling convolution that factorizes standard 3D convolution into volumetric and planar kernels:

$$W_{\text{eff}} = \alpha_{\text{vol}} W_{\text{vol}} + \sum_{uv \in \{HW, HD, WD\}} \alpha_{uv} W_{uv} \qquad (1)$$

where, W_{vol} captures full 3D structures, while W_{uv} learn patterns on orthogonal planes. This design reduces parameters while enhancing fine-grained geometric modeling. With Concat+$1 \times 1 \times 1$ fusion, non-shared planar branches, and channel allocation $r_v + 3r_p = 1$, the parameter counts are $P_{\text{3D}} = C_{\text{in}} C_{\text{out}} K^3$ (standard 3D conv) and $P_{\text{HiVE}} =$

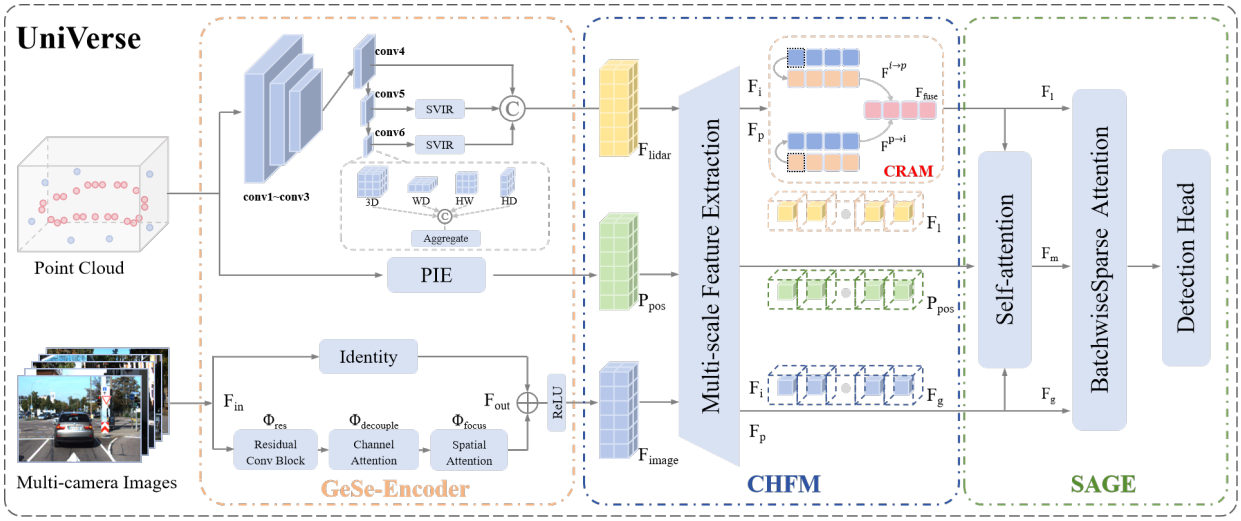

Fig. 2. The architecture of the proposed UniVerse, comprising three core components: GeSe-Encoder for dual-stream feature extraction, CHFM for cross-modal hierarchical fusion, and SAGE for semantic-aware global enhancement, enabling end-to-end 3D object detection from point clouds and images.

$C_{\text{in}}C_{\text{out}}(r_vK^3 + 3r_pK^2) + C_{\text{out}}^2$. This ratio, derived from theoretical analysis and ablation, balances efficiency and representational capacity. Inspired by VoxelNeXt [10], HiVE adopts a hierarchical structure with cross-layer feature propagation to preserve fidelity, and—unlike general factorized 3D convolutions—uses non-shared planar branches to better capture anisotropic patterns in sparse point clouds.

2) Prism-Net: Decoupling-Focusing Attention Network: In the image branch, Prism-inspired Decoupling-Focusing Attention Network (Prism-Net) extracts discriminative semantics F_{out} via a *Decoupling-Focus Residual Block*, inspired by optical prism dispersion and focusing.

$$F_{\text{out}} = \text{ReLU}\left(F_{\text{in}} + \Phi_{\text{focus}}\left(\Phi_{\text{decouple}}\left(\Phi_{\text{res}}\left(F_{\text{in}}\right)\right)\right)\right) \quad (2)$$

where, Φ_{res} is a residual convolution block that performs initial feature encoding, Φ_{decouple} implements a channel attention mechanism to isolate and reweight key channels, and Φ_{focus} applies a spatial attention mechanism to highlight salient regions. This sequential "channel-then-spatial" attention strategy, inspired by CBAM [8], allows global identification of critical categories followed by precise localization. As a result, Prism-Net produces highly sensitive semantic features to target categories and spatial positions, providing reliable priors for alignment with HiVE's geometric outputs.

C. CHFM: Cross-modal Hierarchical Fusion Module

After the GeSe-Encoder outputs high-purity dual-stream features, the next step is efficient, deep alignment of the two heterogeneous modalities. Global-only fusion in existing methods is computationally expensive and prone to missing local fine structures critical for small objects and occlusions. To address this, we propose the CHFM Module, which adopts a "divide-and-conquer" and "bidirectional enhancement" strategy. Its core unit, the CRAM module, performs fine-grained, two-way interaction between point cloud and image features,

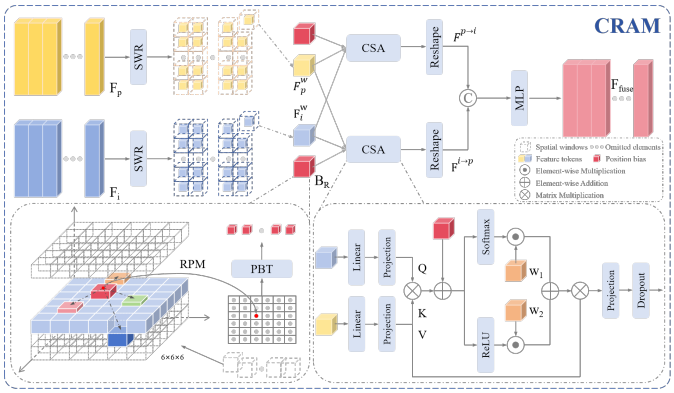

Fig. 3. The detail of the CRAM. Point cloud and image features are processed through spatial window reshaping (SWR) and fed into the CRAM module, which performs bidirectional enhancement ($p \to i$ and $i \to p$) with position-aware encoding. The Cross-modal Synergistic Attention (CSA) module then integrates the enhanced features to generate the final fused representation.

applied hierarchically to global and local features to exploit both large-scale context and local structural cues.

As the core of CHFM, the CRAM module enables efficient, robust cross-modal interaction by restricting feature fusion to local 3D windows, thus avoiding the high cost of global attention. The point cloud features F_p and the image features F_i are projected to a unified $d = 128$ space via independent transformation networks ψ_p and ψ_i. The Spatial Window Reshaping (SWR) operation $\Gamma(\cdot)$ then reorganizes the linear sequences into windows containing local spatial neighborhoods, reducing the attention complexity from $O(N^2)$ to $O(M^2)$, where $M \ll N$ and N is the total number of feature points.

$$F_p^w = \Gamma(F_p') \in \mathbb{R}^{B\times M\times d}, \quad F_i^w = \Gamma(F_i') \in \mathbb{R}^{B\times M\times d} \quad (3)$$

To inject precise 3D geometric priors into local windows, we adopt a two-stage position-aware encoding. Relative Posi-

tion Mapping (RPM) discretizes coordinate differences within each window into indices P_{idx}, which query an initial bias B_{init} from a learnable table T. A Position Bias Transformation (PBT) network then refines this into B_R, rich in geometric information for attention computation.

On this basis, the CRAM employs a dual attention fusion strategy: Softmax attention captures global dependencies, while ReLU-based attention emphasizes local saliency. We weight and sum both to model linear and nonlinear relationships, enabling bidirectional enhancement—point cloud features query images for semantics ($p \to i$), and image features query point clouds for geometry ($i \to p$)—with residual connections. Let $S = \frac{QK^\top}{\sqrt{d_k}} + B_R$, then the dual attention fusion is computed as:

$$A = w_1 \cdot \text{softmax}(S) + w_2 \cdot \text{ReLU}(S)^2 \tag{4}$$

where A is the dual attention matrix combining global and local patterns. Based on this attention matrix, we achieve bidirectional feature enhancement:

$$F^{p\to i} = \text{MLP}(A_p V_i) + F_i^w \tag{5}$$

$$F^{i\to p} = \text{MLP}(A_i V_p) + F_p^w \tag{6}$$

The enhanced features are restored to their original spatial layout via Γ^{-1} and passed to the Cross-modal Synergistic Attention (CSA) module for final integration:

$$F_{\text{fuse}} = \phi_{\text{CSA}}\left(\Gamma^{-1}\left([F^{p\to i}; F^{i\to p}]\right)\right) \tag{7}$$

This hierarchical, bidirectional fusion achieves precise cross-modal alignment and mutual enhancement with high computational efficiency.

D. SAGE: Semantic-Aware Global Enhancement

After high-fidelity feature extraction by the GeSe-Encoder and alignment by the CHFM, the resulting multi-modal features contain precise geometry and preliminary semantics. However, this marks the step from "observation" to "alignment," with residual noise and ambiguities still present. To address suboptimal semantic integration—the final stage of error accumulation, we propose the SAGE module. Unlike passive fusion, it actively constructs a multi-source "decision environment" and refines the aligned features into robust semantic understanding via an interactive "cross-query" mechanism.

The SAGE integrates three complementary sources: fused features F_m as current judgments, global geometric features F_g as unbiased references, and local detail features F_l preserving fine textures. Differentiated projections map them into a unified space, retaining their unique contributions while preserving long-range dependencies and enhancing nonlinear expressiveness.

At its core, the SAGE employs a deep fusion engine with six Transformer decoder layers. In each layer, the fused features F_m first undergoes self-attention to strengthen internal

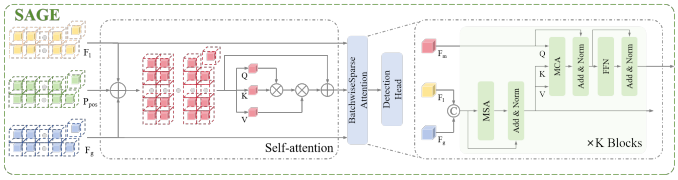

Fig. 4. The structure of the SAGE module. The module integrates local features F_l, position information P_{pos}, and global features F_g through self-attention and BatchwiseSparse Attention mechanisms to generate robust semantic representations for detection.

coherence and contextual awareness. The resulting enhanced features then act as queries in a cross-query process to a "knowledge base" formed by F_g and F_l:

$$\begin{aligned} Q &= \Psi_q(F_m), \\ K &= \Psi_k([F_g; F_l]), \\ V &= \Psi_v([F_g; F_l]) \end{aligned} \tag{8}$$

For efficient computation, we adopt batch-level cross-attention, constraining the scope within batches.

$$A_b = \text{softmax}\left(\frac{Q_b K_b^\top}{\sqrt{d_k}}\right) V_b, \quad b \in \{1, \ldots, B\} \tag{9}$$

where A_b is the attention output for batch b.

The enhanced information is integrated through a feed-forward network and fused with residual connections. This "self-enhancement - cross-query - deep integration" process iterates, ultimately outputting a robust unified feature representation that fuses multi-source advantages, providing a reliable foundation for the detection task.

E. Loss Function

To achieve end-to-end optimization of the entire UniVerse framework, we design a composite multi-task loss function L_{total}. This function consists of the standard cross-entropy classification loss L_{cls} and a composite regression loss L_{reg} weighted together. Among them, L_{reg} adopts weighted smooth L1 loss L_{box} following [12] to robustly regress bounding box parameters and introduces corner loss L_{corner} as proposed in [28] as an explicit geometric regularization term. The total loss function is defined as follows:

$$L_{\text{total}} = \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{reg}}(L_{\text{box}} + \lambda_{\text{corner}} L_{\text{corner}}) \tag{10}$$

where λ_{cls}, λ_{reg}, and λ_{corner} are hyperparameters.

IV. EXPERIMENTS

A. Experimental Settings

1) Datasets: To verify the effectiveness of the proposed method, we conducted systematic experimental studies on the representative KITTI benchmark dataset [29] in the autonomous driving field. The KITTI dataset is collected from

urban road driving scenarios, containing 7,481 training samples and 7,518 test samples, each equipped with precisely synchronized point cloud data and front-view RGB images. Following the standard experimental setup [30], we split the training set into a training set (3,712 samples) and a validation set (3,769 samples) in approximately a 1:1 ratio. Performance evaluation utilizes the official average precision (AP) metric, categorizing detection difficulty into easy, moderate, and hard levels based on the degree of occlusion, truncation, and object size in the images.

2) Implementation Details: To ensure fair comparison with existing methods, we optimized the proposed modules while maintaining standard KITTI [29] detection ranges: X-axis [0, 70.4m], Y-axis [-40m, 40m], Z-axis [-3m, 1m], with corresponding voxel sizes of (0.05m, 0.05m, 0.1m).

Network architecture. In GeSe-Encoder, HiVE fused feature maps from the last three layers (conv4–conv6) of the 3D backbone for enhanced multi-scale representation, while Prism-Net was built on ResNet-50+FPN. CHFM's CRAM employed $6 \times 6 \times 6$ 3D windows with 4 attention heads. SAGE used a 6-layer Transformer decoder. The image branch adopted ImageNet-pretrained ResNet-50+FPN. For efficiency, the input image resolution was halved, and all cross-modal feature dimensions were unified to 128.

Training. We used AdamW with an initial learning rate of 0.001 and cosine annealing, training for 80 epochs on a single NVIDIA A6000 GPU (batch size 4). Data augmentation included random X-axis flipping, global scaling ([0.95, 1.05]), and Z-axis rotation ($[-\pi/4, \pi/4]$). Post-processing applied the NMS with IoU thresholds of 0.7 and 0.55. To stabilize feature extraction and focus on fusion optimization, the image backbone (ResNet-50+FPN) parameters were frozen during training. For inference, we used the same hardware (single NVIDIA A6000 GPU) with batch size 2 and FP32 precision; the input image resolution was set to half of the original (approximately 704×384), consistent with the training configuration.

B. Comparison Results

KITTI Test Set. For fairness, we compared UniVerse with other advanced methods on the public KITTI test set leaderboard. As shown in Table I, our method achieved competitive performance across all categories. Compared to some classic multi-modal methods, UniVerse demonstrated its ability to handle complex scenarios in Car, Pedestrian, and Cyclist categories, ultimately achieving 65.63% 3D mAP, demonstrating the effectiveness and potential of our framework in overall performance. This result preliminarily validated the rationality of our framework design, namely: 1) GeSe-Encoder provides a high-quality geometric and semantic feature foundation for subsequent processing; 2) The CHFM module calibrated cross-modal features through bidirectional enhancement, helping to alleviate localization inaccuracies caused by projection errors.

Notably, UniVerse achieves its most prominent gains on Hard-level objects, ranking first across all three categories (Car, Pedestrian, Cyclist), while improvements on Easy and Moderate levels are comparatively modest. This is expected: Easy-level targets are typically close-range and densely observed, where LiDAR geometry alone suffices and strong single-modality baselines such as 3DSSD already achieve near-saturation performance, leaving limited room for cross-modal fusion to contribute. Hard-level targets, by contrast, involve sparse point returns, severe occlusion, or long range, where pure geometric evidence becomes insufficient. It is precisely in these challenging cases that SAGE's semantic compensation and CHFM's bidirectional alignment deliver the most meaningful gains—exactly where the proposed cascade-error mitigation is most needed.

KITTI Validation Set. We conducted a more detailed comparison with existing methods on the local validation set, with results shown in Table II. UniVerse achieved robust performance across all three categories (Car, Pedestrian, Cyclist), demonstrating its good generalization ability. Especially on the small object categories of Pedestrian and Cyclist, our method performed well, benefiting from the fine fusion at the local feature level in the CHFM module and the global context perception capability of the SAGE module.

C. Ablation Studies

Analysis of Module Performance. On the 3,769-sample KITTI validation set, GeSe-Encoder, CHFM, and SAGE improve the baseline by +1.29%, +1.79%, and +0.67% mAP, respectively, as shown in Table III. Combining GeSe-Encoder and CHFM yields a +2.64% mAP gain, highlighting their synergy. Integrating all modules achieves the best result of 74.74% mAP, confirming the effectiveness of the progressive optimization strategy.

Robustness of the GeSe-Encoder. Table IV evaluated the two subnetworks, HiVE and Prism-Net, by replacing them with SECOND and ResNet-50+FPN, respectively. Substituting SECOND with HiVE yielded +0.81% mAP, with notable gains in the structure-dependent Car (+1.31%) and Cyclist (+0.93%) categories. Replacing the standard image backbone with Prism-Net contributed +0.77% mAP, with the most pronounced improvement in Pedestrian detection (+1.39%), which requires fine semantic discrimination. These results confirmed the dual-stream encoder's effectiveness in decoupling and purifying geometric and semantic features.

Effectiveness of the CHFM. Table V examined key CHFM designs. Removing hierarchical fusion (*i.e.*, fusing only at the global level) degrades overall mAP by 0.94%, especially harming Pedestrian detection (-0.93%), underscoring the importance of multi-scale processing. Changing CRAM's bidirectional enhancement to unidirectional (point $\rightarrow$ image) reduced mAP by 1.09%, while replacing CRAM with simple concatenation causes a larger drop, highlighting the reciprocal attention mechanism's role in precise cross-modal alignment.

Validity of the SAGE. Table VI analyzed the impact of multi-source inputs. The SAGE augmented fused features F_m with raw global features F_g and local features F_l as a "knowledge base." Using only F_m lowered mAP by 1.42%, and removing F_g or F_l individually also harms performance.

TABLE I
COMPARISONS ON KITTI 3D TESTING SET. 3D MAP IS AVERAGED OVER ALL CATEGORIES AND DIFFICULTY LEVELS.

Method	Modality	3D Detection (Car)			3D Detection (Ped.)			3D Detection (Cyc.)			3D mAP
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	
VoxelNet [3]	LiDAR	77.47	65.11	57.73	39.48	33.69	31.51	61.22	48.36	44.37	50.99
Point RCNN [31]	LiDAR	86.96	75.64	70.70	47.98	39.37	36.01	74.96	58.82	52.53	60.33
TANet [32]	LiDAR	84.39	75.94	68.82	53.72	**44.34**	40.49	75.70	59.44	52.53	61.71
STD [33]	LiDAR	87.95	79.71	75.09	53.29	42.47	38.35	78.69	61.59	55.30	63.60
PV-RCNN [12]	LiDAR	90.25	81.43	76.82	52.17	43.29	40.29	78.60	63.71	57.65	64.91
3DSSD [34]	LiDAR	88.36	79.57	74.55	**54.64**	44.27	40.23	**82.48**	64.10	56.90	65.01
SegVoxelNet [35]	LiDAR	86.04	76.13	70.76	-	-	-	-	-	-	-
SE-SSD [36]	LiDAR	91.49	82.54	77.15	-	-	-	-	-	-	-
SFD [37]	LiDAR+RGB	**91.73**	**84.76**	77.92	-	-	-	-	-	-	-
F-PointNet [38]	LiDAR+RGB	82.19	69.79	60.59	50.53	42.15	38.08	72.27	56.12	49.01	57.86
PointPainting [1]	LiDAR+RGB	82.11	71.70	67.08	50.32	40.97	37.87	77.63	63.78	55.89	60.82
Faraway-Frustum [39]	LiDAR+RGB	87.45	79.05	76.14	46.33	38.58	35.71	77.36	62.00	55.40	62.00
BEVFusion [6]	LiDAR+RGB	80.23	78.45	74.01	50.97	43.12	39.01	80.10	63.90	55.88	62.42
EPNet [4]	LiDAR+RGB	80.05	78.12	74.01	50.97	42.15	39.01	80.10	63.78	55.89	62.66
F-ConvNet [40]	LiDAR+RGB	87.36	76.39	66.69	52.16	43.38	38.80	81.98	**65.07**	56.54	63.15
CLIX3D [41]	LiDAR+RGB	80.87	79.03	75.12	51.88	43.88	39.74	80.90	64.21	56.33	63.32
MuStD [42]	LiDAR+RGB	82.03	80.12	76.45	52.10	44.27	40.88	81.65	65.12	56.87	64.21
UniVerse (Ours)	LiDAR+RGB	89.86	81.87	**78.97**	52.17	44.03	**41.74**	80.28	64.03	**57.78**	**65.63**

TABLE II
COMPARISONS ON KITTI 3D VALIDATION SET. RESULTS ARE THE MEAN AP AVERAGED OVER EASY, MODERATE, AND HARD DIFFICULTIES. "L" AND "R" INDICATE USING LIDAR AND RGB IMAGES, RESPECTIVELY.

Method	Modality	Car	Ped.	Cyc.	3D mAP
SECOND [2]	L	81.48	52.42	70.28	68.06
Point RCNN [31]	L	81.72	55.60	74.68	70.67
PV-RCNN [12]	L	83.62	57.97	73.35	71.64
3DSSD [34]	L	81.43	54.02	74.02	69.82
SE-SSD [36]	L	85.23	-	-	-
SFD [37]	L+R	**87.35**	-	-	-
F-PointNet [38]	L+R	72.78	**61.64**	62.34	65.58
EPNet [4]	L+R	81.97	60.28	70.69	70.96
UniVerse(Ours)	L+R	84.03	61.16	**79.03**	**74.74**

TABLE III
EFFECT OF EACH MODULE ON KITTI 3D VALIDATION SET. VALUES ARE MAP FOR EACH CLASS (EASY/MODERATE/HARD); RIGHTMOST COLUMN IS OVERALL 3D MAP.

GeSe-Encoder	CHFM	SAGE	Car	Ped.	Cyc.	3D mAP
			80.51	57.13	76.02	71.22
✓			81.73	58.51	77.28	72.51
	✓		82.15	59.02	77.85	73.01
		✓	81.08	57.89	76.71	71.89
✓	✓		83.15	60.02	78.41	73.86
✓		✓	82.88	59.54	78.05	73.49
	✓	✓	82.25	59.11	77.63	73.00
✓	✓	✓	**84.03**	**61.16**	**79.03**	**74.74**

TABLE IV
ANALYSIS ON THE GESE-ENCODER

Backbone Networks		Category Performance			Overall
LiDAR	Image	Car	Ped.	Cyc.	3D mAP
SECOND	ResNet-50+FPN	82.11	58.93	77.65	72.90
HiVE	ResNet-50+FPN	83.42	59.14	78.58	73.71
SECOND	Prism-Net	82.75	60.32	77.93	73.67
HiVE	Prism-Net	**84.03**	**61.16**	**79.03**	**74.74**

TABLE V
ANALYSIS RESULTS OF THE CHFM MODULE

Method	Car	Ped.	Cyc.	3D mAP
Baseline (Concat)	81.73	58.51	77.28	72.51
UniVerse w/o Hierarchical	83.51	59.23	78.65	73.80
UniVerse w/o Bidirection	82.94	59.89	78.11	73.65
UniVerse (Full)	**84.03**	**61.16**	**79.03**	**74.74**

This demonstrated that deep interaction among F_m, F_g, and F_l is essential for robust, context-aware detection.

TABLE VI
ANALYSIS ON INPUTS FOR THE SAGE MODULE

Method	Car	Ped.	Cyc.	3D mAP
UniVerse w/o (F_g,F_l)	82.68	59.82	77.45	73.32
UniVerse w/o (F_g)	83.21	60.15	78.23	73.86
UniVerse w/o (F_l)	83.45	60.76	78.48	74.23
UniVerse (Full)	**84.03**	**61.16**	**79.03**	**74.74**

D. Qualitative Analysis

Figs. 5–8 show qualitative results across diverse scenarios: dense urban traffic, distant occlusions, complex crowded hubs, and rural roads. Compared to the baseline, which exhibits false positives, missed detections, and inaccurate boxes, UniVerse consistently produces complete and precise detections, complementing the quantitative findings.

E. Inference Speed Analysis

We evaluated UniVerse's computational efficiency under a standardized setting (single NVIDIA A6000 GPU, batch size 2, FP32 precision) for fair, reproducible comparison. Table VII reports inference speed on the KITTI validation set, with FPS values from official reports or our re-implementations where official data was unavailable.

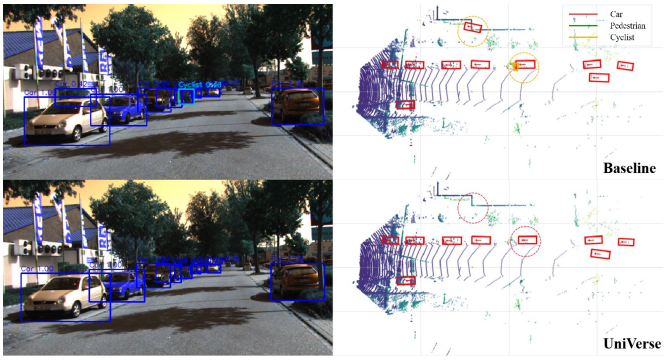

Fig. 5. Qualitative results in a dense urban scene. The baseline yields false positives (e.g., roadside flowerbeds as vehicles, distant blurry objects as cyclists) and misses small objects under clutter. UniVerse suppresses such errors and recovers valid targets with tighter boxes.

Fig. 6. Qualitative results for a distant, partially occluded vehicle. The baseline misses the target due to insufficient context reasoning, whereas UniVerse detects it with accurate localization.

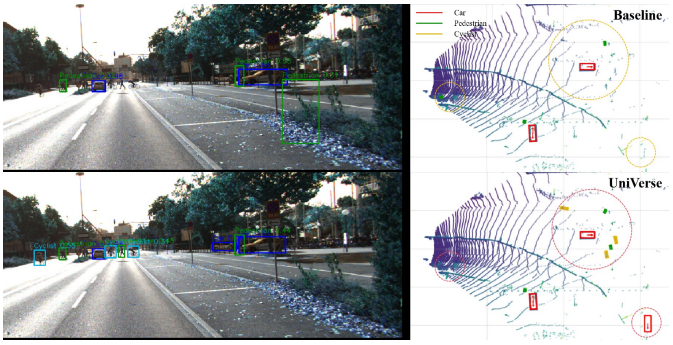

Fig. 7. Qualitative results in a complex urban hub. The baseline fails to detect multiple pedestrians and cyclists on a distant crosswalk, misses a distant occluded vehicle, and misclassifies nearby greenery as a vehicle. UniVerse correctly detects all valid objects while avoiding such false positives.

Among LiDAR+RGB methods, UniVerse achieved 7.5 FPS, comparable to BEVFusion† and CLIX3D, and faster than F-PointNet and EPNet. The lower speed relative to 3DSSD (25.8 FPS) is expected, as 3DSSD is a LiDAR-only method without cross-modal fusion stages. Within the multi-modal group, UniVerse's throughput reflects a deliberate accuracy–efficiency trade-off; windowed attention in CHFM reduces complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(M^2)$, partially offsetting the cost of bidirectional fusion. Mixed-precision inference and model pruning remain future directions.

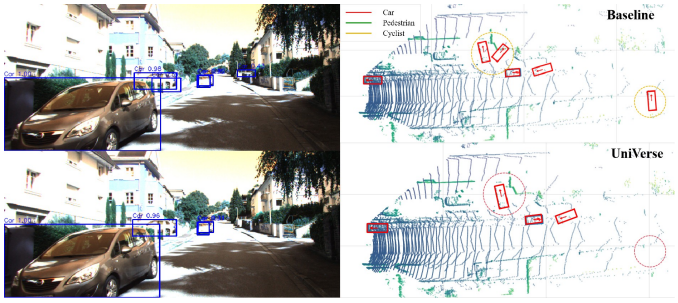

Fig. 8. Qualitative results in a simple rural road scene. The baseline erroneously splits a nearby occluded vehicle into two detections and misclassifies a distant low roadside wall as a vehicle. UniVerse provides correct and robust predictions.

TABLE VII
INFERENCE SPEED COMPARISON ON KITTI VALIDATION SET. † INDICATES OUR RE-IMPLEMENTATIONS.

Method	Modality	FPS
SE-SSD [36]	LiDAR	7.6
PV-RCNN [12]	LiDAR	8.1
PointRCNN [31]	LiDAR	9.7
3DSSD [34]	LiDAR	**25.8**
F-PointNet [38]	LiDAR+RGB	5.6
EPNet [4]	LiDAR+RGB	6.3
BEVFusion† [6]	LiDAR+RGB	8.6
CLIX3D [41]	LiDAR+RGB	9.8
UniVerse (Ours)	LiDAR+RGB	7.5

F. Discussion

The ablation study identifies cross-modal feature alignment as the primary bottleneck: CHFM alone yields the largest single-module gain (+1.79% mAP), and combining it with GeSe-Encoder produces a synergistic improvement of +2.64%, exceeding the sum of individual contributions. The most pronounced category-level gain is in Pedestrian detection (57.13%→61.16%), reflecting the complementary contributions of Prism-Net's semantic discrimination and SAGE's multi-source aggregation—both particularly beneficial for small, occluded objects. Gains in the Car category further confirm broad robustness. The full model consistently outperforms every partial configuration, validating that the three-stage cascade design addresses the targeted error sources in a coherent and non-redundant manner.

V. CONCLUSIONS

This paper presented UniVerse, a multi-modal 3D detection framework that addresses cascade error propagation through a three-stage design: GeSe-Encoder decouples geometric and semantic features at the source, CHFM achieves precise cross-modal alignment via bidirectional hierarchical attention, and SAGE consolidates multi-source information into a robust unified representation. Experiments on KITTI, supported by systematic ablation studies, confirm the contribution of each component and the overall superiority of the cascade paradigm. Limitations include single-dataset evaluation (KITTI was selected for its standardized protocol and broad

baseline coverage, though validation on nuScenes or Waymo remains necessary) and sub-real-time inference speed; cascade error reduction also lacks direct perturbation-based quantification. Future work will extend the framework to temporal multi-frame settings, validate generalization on nuScenes and Waymo, and investigate model compression to bridge the gap toward practical deployment.

REFERENCES

[1] S. Vora, A. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 4604–4612.

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

[3] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4490–4499.

[4] T. Huang, Z. Wang, L. Chen, W. Xu *et al.*, "EPNet: Enhancing point features with image semantics for 3D object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 35–52.

[5] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 11 784–11 793.

[6] Z. Liu, M. Chen, K. Li, P. Zhao *et al.*, "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023, pp. 2774–2781.

[7] C. Wu, Y. Zhu *et al.*, "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 2736–2746.

[8] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 3–19.

[9] X. Bai, Y. Zhu, J. Zhou, M. Li *et al.*, "TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1090–1099.

[10] Y. Chen, X. Liu, X. Qi, and J. Jia, "VoxelNeXt: Fully sparse VoxelNet for 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 21 674–21 683.

[11] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 652–660.

[12] S. Shi, Z. Wang, and H. Li, "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10 529–10 538.

[13] C. Park, J. Lee, Y. J. Jeong *et al.*, "Fast point transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 16 949–16 958.

[14] Z. Pang, Y. Wang, S. Liu *et al.*, "Masked autoencoders for point cloud self-supervised learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 604–621.

[15] D. Xu, D. Anguelov, and A. Jain, "PointFusion: Deep sensor fusion for 3D bounding box estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 244–253.

[16] Z. Chen *et al.*, "AutoAlign: Pixel-instance feature aggregation for multi-modal 3D object detection," *arXiv preprint arXiv:2201.06493*, 2022.

[17] Z. X. Yu, Hongqi and et al, "PrimePSegter: Progressively combined diffusion for 3D panoptic segmentation with multi-modal BEV refinement," *IEEE Transactions on Multimedia*, vol. 27, no. 9, Sep. 2025.

[18] J. Yang *et al.*, "DeepInteraction: 3D object detection via modality interaction," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 1992–2005.

[19] K. Li *et al.*, "UniFormer: Unifying convolution and self-attention for visual recognition," *arXiv preprint arXiv:2201.09450*, 2022.

[20] C. Li, R. Kumar, H. Wang, and A. Mian, "SyNet: A synergistic network for 3D object detection through geometric-semantic-based multi-interaction fusion," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2025, pp. 1–8.

[21] Q. Chen, H. Wang, L. Xu *et al.*, "MPPNet: Multi-frame feature intertwining with proxy points for 3D temporal object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 680–697.

[22] J. Deng, X. Li, R. Zhao *et al.*, "VISTA: Boosting 3D object detection via dual cross-view spatial attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 8438–8447.

[23] C.-Y. Tseng, W.-C. Huang, H.-K. Chen *et al.*, "CrossDTR: Cross-view and depth-guided transformers for 3D object detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023, pp. 6600–6606.

[24] Z. Qin, M. Zhou *et al.*, "UniFusion: Unified multi-view fusion transformer for spatial-temporal representation in Bird's-Eye-View," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 12 685–12 695.

[25] Y. Wang *et al.*, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. Conf. Robot Learn. (CoRL)*. PMLR, 2022, pp. 180–191.

[26] Y. Jiang *et al.*, "PolarFormer: Multi-camera 3D object detection with polar transformers," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2023, pp. 993–1002.

[27] Z. Zhu, W. Zheng *et al.*, "BEVerse: Unified perception and prediction in Bird's-Eye-View for vision-centric autonomous driving," *arXiv preprint arXiv:2205.09743*, 2022.

[28] Y. Zhou, Y. Sun, W. Liu, and J. Deng, "IoU loss for 2D/3D object detection," in *Proc. IEEE Int. Conf. 3D Vis. (3DV)*, 2019, pp. 85–94.

[29] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, 2012, pp. 3354–3361.

[30] H. Wu, C. Wen, W. Li, X. Li, R. Yang, and C. Wang, "Transformation equivariant 3D object detection for autonomous driving," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, Washington, DC, USA, 2023, pp. 2795–2802.

[31] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 770–779.

[32] Z. Liu, X. Zhao, T. Huang, R. Hu, Y. Zhou, and X. Bai, "TANet: Robust 3D object detection from point clouds with triple attention," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, New York, NY, USA, 2020, pp. 11 677–11 684.

[33] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "STD: Sparse-to-dense 3D object detector for point cloud," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, 2019, pp. 1951–1960.

[34] Z. Yang, Y. Sun, S. Liu, and J. Jia, "3DSSD: Point-based 3D single stage object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 11 040–11 048.

[35] H. Yi, S. Shi, M. Ding, J. Wang, and S. Bai, "SegVoxelNet: Exploring semantic context and depth-aware features for 3D vehicle detection from point cloud," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Paris, France, 2020, pp. 2274–2280.

[36] C. Zheng, W. Xu, H. Howard-Jenkins, C. Chen, A. Leonardis, and C. Wen, "SE-SSD: Self-ensembling single-stage object detector from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, 2021, pp. 14 494–14 503.

[37] X. Wu, L. Peng, H. Yang, L. Xie, C. Huang, C. Deng, H. Liu, and D. Cai, "Sparse fuse dense: Towards high quality 3D detection with depth completion," *arXiv preprint arXiv:2203.09780*, 2022.

[38] C. R. Qi *et al.*, "Frustum PointNets for 3D object detection from RGB-D data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 918–927.

[39] D. Yang *et al.*, "Faraway-frustum: Dealing with LiDAR sparsity for 3D object detection using fusion," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*. IEEE, 2021, pp. 2646–2652.

[40] Z. Wang and K. Jia, "Frustum ConvNet: Sliding frustums to aggregate local point-wise features for amodal 3D object detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*. IEEE, 2019, pp. 1742–1749.

[41] H. Li, Z. Wang, and Y. Liu, "CLIX3D: A language-image fusion framework for 3D object detection," *IEEE Trans. Multimedia*, vol. 27, no. 8, pp. 2154–2166, Aug. 2025.

[42] M. Ibrahim *et al.*, "Multistream network for LiDAR and camera-based 3D object detection in outdoor scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2025, pp. 1–8.