MDPI

Article

Dual-Camera LiDAR Fusion for Occlusion-Robust 3D Detection in Urban Driving Simulation

Xingnan Zhou [1] and Ciprian Alecsandru [1,*]

[1] Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, QC H3G 1M8, Canada

* Correspondence: ciprian.alecsandru@concordia.ca (C.A.)

Version February 12, 2026 submitted to Sustainability

Abstract: Three-dimensional object detection from LiDAR point clouds is a cornerstone of autonomous driving perception, yet single-sensor systems remain vulnerable to false positives and occlusion in complex urban environments such as roundabouts and dense intersections. This paper proposes a dual-camera LiDAR fusion framework that combines 3D LiDAR detectors (PointPillar and CenterPoint) with YOLOv8-based 2D detections from two complementary camera viewpoints: a drone (top-down, 40 m altitude) and a subject-vehicle forward camera. The fusion operates at the decision level (late fusion), where camera-confirmed LiDAR detections receive confidence boosts while unconfirmed low-confidence detections within camera fields of view are suppressed. We evaluate multiple fusion strategies—asymmetric (drone boost+suppress, SDC boost-only), symmetric (both cameras boost+suppress), and naive score averaging—and find that *symmetric* fusion, which applies boost and suppress operations uniformly to both cameras, achieves the best results. Evaluated on a CARLA Town10HD dataset comprising 2,600 frames across Car and Pedestrian classes, with ten-seed repeated random sub-sampling validation, the drone-fused system improves PointPillar mAP@0.5 by +0.63 percentage points (+3.0% relative) and the symmetric dual-camera fusion achieves +0.92 pp (+4.4% relative), both with 10/10 positive seeds (sign test $p = 0.001$, t-test $p < 0.0001$). The primary mechanism of improvement is *false positive suppression*: drone fusion reduces false positives by 13%, improving precision from 66.1% to 69.2%. CenterPoint exhibits the same pattern: symmetric fusion achieves +0.74 pp (+3.3% relative, $p = 0.001$), confirming that the fusion benefit is detector-agnostic. From a sustainable transportation perspective, fewer false alarms translate to smoother traffic flow with reduced phantom braking events, supporting lower emissions in autonomous driving deployments.

Keywords: 3D object detection; LiDAR-camera fusion; intelligent transport system; late fusion; drone-assisted perception; sustainable traffic management; PointPillars; CenterPoint; YOLOv8; CARLA simulation; autonomous driving safety; false positive suppression

1. Introduction

Reliable three-dimensional (3D) object detection is a fundamental requirement for safe autonomous driving. LiDAR-based detectors have become the dominant paradigm for 3D perception in self-driving systems, offering precise depth measurements and geometric representations of the driving environment [1,2]. However, single-viewpoint LiDAR systems suffer from well-documented limitations: occlusion by foreground objects, sparse point density at long range, and blind spots caused by the sensor's mounting position [3]. These limitations are particularly acute in complex urban geometries such as roundabouts, dense intersections, and narrow streets, where vehicles, pedestrians, and infrastructure elements frequently occlude one another from the ego vehicle's perspective.

A less-discussed but equally important challenge is that of *false positive detections*. LiDAR-based detectors frequently produce phantom detections caused by ground clutter, reflective surfaces, and ambiguous point patterns—particularly in urban environments with complex geometry. These false positives degrade not only detection metrics but also downstream planning: each phantom detection may trigger unnecessary braking or evasive maneuvers, reducing traffic efficiency and passenger comfort. In safety-critical applications, minimizing false alarms is as important as maximizing recall.

Camera-LiDAR fusion has emerged as a promising strategy to mitigate single-sensor limitations by combining the geometric precision of LiDAR with the rich semantic and texture information provided by cameras [4,5]. Early fusion methods such as PointPainting [6] and Frustum PointNets [7] augment point clouds with camera features at the data level, while deep fusion approaches like BEVFusion [8,9] and TransFusion [10] learn joint representations in a unified bird's-eye view (BEV) space. Although these methods have achieved impressive results on benchmarks such as nuScenes [11] and KITTI [12], they predominantly rely on cameras co-located with the ego vehicle, inheriting the same viewpoint limitations that constrain the LiDAR sensor.

A fundamentally different approach to overcoming single-viewpoint occlusion is to introduce cameras at *complementary* vantage points. Drone-assisted perception [13,14] and vehicle-to-infrastructure (V2I) cooperative systems [15,16] have demonstrated that elevated or offset viewpoints can observe objects hidden from the street-level perspective. However, most cooperative perception research has focused on sharing intermediate features between multiple LiDAR-equipped agents [17–19], which requires expensive sensor suites on all cooperating platforms. In contrast, cameras are lightweight, inexpensive, and easily deployable on drones or infrastructure poles, making camera-based viewpoint augmentation a practical and cost-effective alternative to full multi-LiDAR cooperative perception.

This paper proposes a dual-camera LiDAR fusion framework that addresses both false positives and occlusion in urban driving by combining a vehicle-mounted LiDAR 3D detector with 2D object detections from two complementary cameras: (1) a drone camera providing a top-down perspective at 40 m altitude, and (2) the subject vehicle's forward-facing camera. The fusion operates at the decision level (late fusion), modifying the confidence scores of the 3D LiDAR detector's output based on spatial agreement with 2D camera detections projected into the BEV plane. We investigate multiple fusion strategies—asymmetric (differentiated operations per camera), symmetric (uniform operations for both cameras), and naive score averaging—and find that symmetric fusion, where both cameras apply boost and suppress operations, achieves the best performance.

A key finding of this work is that the primary mechanism of improvement is *false positive suppression* rather than occlusion recovery. When a drone camera observes a region where the LiDAR detector reports a detection but the camera sees no object, this provides strong evidence that the detection is spurious. Our analysis shows that drone fusion reduces false positives by 13% (from 487 to 423 per evaluation set), improving precision from 66.1% to 69.2%. From a sustainable transportation perspective, this precision improvement directly supports safer autonomous driving—a critical enabler of sustainable urban mobility. Fewer phantom detections translate to smoother traffic flow with reduced unnecessary braking, improving fuel efficiency and lowering emissions. Meanwhile, traffic accidents remain a leading cause of preventable death [20], and occlusion-related collisions at intersections and roundabouts disproportionately affect vulnerable road users. By leveraging cost-effective drone cameras rather than expensive multi-LiDAR infrastructure, the proposed approach offers a scalable pathway to enhanced perception that aligns with sustainable smart-city transportation strategies [21].

We evaluate the proposed framework using the CARLA driving simulator [22], which provides precise ground-truth annotations and full control over sensor placement. The dataset comprises 2,600 frames collected in Town10HD across Car and Pedestrian classes. To ensure statistical rigor, we conduct ten-seed repeated random sub-sampling validation and report both parametric (paired t-test) and

non-parametric (sign test) significance measures. To validate detector-agnosticism, we evaluate fusion with two 3D detectors: PointPillar [23] and CenterPoint [24]. Our contributions are as follows:

1. We propose a dual-camera late-fusion framework and systematically compare asymmetric, symmetric, and naive-averaging fusion strategies, demonstrating that symmetric fusion—where both cameras apply boost and suppress operations—outperforms asymmetric designs. This finding challenges the intuition that narrow-FOV cameras should be restricted to boost-only operations.
2. We demonstrate that the primary mechanism of fusion improvement is *false positive suppression*: drone fusion reduces false positives by 13%, improving precision by +3.1 percentage points. Symmetric fusion improves PointPillar mAP@0.5 by +0.92 pp (+4.4% relative) with 10/10 positive seeds (sign test $p = 0.001$, t-test $p < 0.0001$).
3. We validate the fusion framework across two 3D detectors (PointPillar and CenterPoint), ten random seeds, and three IoU thresholds (0.3, 0.5, 0.7), with evaluation restricted to a physically meaningful 0–50 m BEV range. All results achieve strong statistical significance ($p = 0.001$).
4. We provide detailed analysis of the fusion mechanism, including distance-stratified performance, per-frame improvement rates, occlusion category breakdown, and safety-relevant metrics (false positive reduction, dangerous object recovery).

The remainder of this paper is organized as follows. Section 2 reviews related work on LiDAR-based 3D detection, camera-LiDAR fusion, and drone-assisted perception. Section 3 describes the proposed fusion methodology and its variants. Section 4 details the experimental setup, including the CARLA data-collection pipeline and training configuration. Section 5 presents quantitative results with statistical analysis. Section 6 discusses findings, limitations, and practical implications. Section 7 concludes the paper.

2. Related Work

2.1. LiDAR-Based 3D Object Detection

LiDAR-based 3D object detection has progressed through several architectural paradigms. Point-based methods such as PointNet [25] and PointNet++ [26] operate directly on raw point clouds, learning per-point features through shared multi-layer perceptrons and set abstraction layers. PointRCNN [27] extended this paradigm to two-stage 3D detection by generating proposals from point-level features. While point-based methods preserve fine geometric detail, their computational cost scales unfavorably with point cloud density.

Voxel-based methods discretize the point cloud into a regular 3D grid and apply sparse 3D convolutions. VoxelNet [28] pioneered end-to-end voxel-based detection, and SECOND [29] introduced spatially sparse convolutions that dramatically reduced computation by only processing occupied voxels. CenterPoint [24] further advanced voxel-based detection with center-heatmap prediction and a two-stage refinement module, achieving state-of-the-art results on the Waymo and nuScenes benchmarks. These methods achieve strong accuracy but require careful voxel resolution tuning to balance precision and efficiency.

Pillar-based methods, led by PointPillars [23], offer a compelling trade-off between speed and accuracy by collapsing the vertical dimension into a single "pillar" per horizontal grid cell. The resulting 2D pseudo-image can be processed by standard 2D convolutional backbones and detection heads, enabling real-time inference. PointPillars remains a widely used baseline in both academic benchmarks and practical deployments due to its simplicity, speed, and competitive accuracy. Wang et al. [30] further explored pillar-based architectures with improved feature encoding. Hybrid approaches such as PV-RCNN [31] combine voxel and point-based processing for state-of-the-art accuracy at the cost of increased complexity.

In this work, we adopt both PointPillars and CenterPoint as LiDAR 3D detectors: PointPillars as a fast pillar-based baseline and CenterPoint as a more accurate voxel-based alternative. Evaluating fusion with two architecturally distinct detectors demonstrates the generalizability of the proposed approach. The modular nature of our late-fusion framework means that the 3D detector can be replaced with any alternative without modifying the fusion logic.

2.2. Camera-LiDAR Fusion for 3D Detection

Camera-LiDAR fusion methods can be broadly categorized by the stage at which sensor information is combined: early (data-level), deep (feature-level), and late (decision-level) fusion [4,5].

2.2.1. Early Fusion

Early fusion methods augment the LiDAR point cloud with camera-derived features before detection. PointPainting [6] projects LiDAR points onto the camera image and appends per-point semantic segmentation scores to the point features, enabling the 3D detector to leverage appearance information. MV3D [32] generates multi-view representations (BEV, front view, and camera image) and fuses them through region-based networks. While conceptually straightforward, early fusion methods are sensitive to calibration accuracy and cannot leverage camera information for regions not covered by the LiDAR.

2.2.2. Deep Fusion

Deep fusion methods learn joint representations from both modalities. BEVFusion [8,9] lifts camera features into 3D space using depth estimation (e.g., the Lift-Splat-Shoot paradigm [33]) and fuses them with LiDAR BEV features through concatenation or attention mechanisms. TransFusion [10] uses transformer-based cross-attention to fuse LiDAR and camera features at the object query level. DeepFusion [34] introduces cross-modal alignment through learned geometric transformations. These methods achieve state-of-the-art accuracy on benchmarks like nuScenes [11] but require end-to-end retraining and are computationally expensive.

2.2.3. Late Fusion

Late fusion methods combine the outputs (detections) of independent modality-specific detectors at the decision level. CLOCs [35] learns a fusion network that combines 2D camera detections with 3D LiDAR detections based on geometric consistency, improving recall without modifying the base detectors. AVOD [36] jointly generates 3D proposals from both modalities and fuses them at the region-of-interest level. Nobis et al. [37] demonstrated late fusion of radar and camera detectors using learned confidence recalibration. Late fusion has practical advantages: the individual detectors can be trained independently, the fusion module is lightweight, and the approach is inherently modular. Our work follows the late-fusion paradigm but introduces coverage-aware confidence adjustment that accounts for the differing coverage characteristics of the two cameras, and systematically compares symmetric and asymmetric fusion strategies.

2.3. Drone-Assisted and Cooperative Perception

The use of elevated viewpoints to overcome occlusion has gained increasing attention. Vehicle-to-everything (V2X) cooperative perception frameworks such as OPV2V [17], V2X-ViT [18], and CoBEVT [19] enable multiple LiDAR-equipped agents to share intermediate features for improved detection. DAIR-V2X [16] provides a benchmark for vehicle-infrastructure cooperative 3D detection. These approaches typically assume that all cooperating agents are equipped with LiDAR sensors and high-bandwidth communication links.

Drone-assisted perception offers a complementary paradigm where an unmanned aerial vehicle (UAV) provides an elevated camera viewpoint to augment the ego vehicle's perception [13,14]. The

overhead perspective of a drone is particularly effective for resolving occlusions in dense traffic, as objects hidden behind foreground vehicles are often fully visible from above. However, drones typically carry only cameras (not LiDAR) due to payload constraints, necessitating a heterogeneous fusion approach that combines 3D LiDAR detections with 2D camera detections from the drone.

Our work is distinguished from prior cooperative perception research in two key aspects. First, we fuse *heterogeneous* sensor modalities (3D LiDAR from the ego vehicle + 2D cameras from two viewpoints), rather than sharing homogeneous LiDAR features. Second, we systematically compare multiple fusion strategies (asymmetric, symmetric, naive averaging) and demonstrate that the primary benefit comes from false positive suppression rather than occlusion recovery, a finding with direct implications for practical deployment.

3. Methodology

This section describes the overall system architecture, the individual detection components, and the fusion algorithm variants.

3.1. System Overview

The proposed pipeline comprises three stages (Figure 1):

1. **3D LiDAR Detection.** A 3D detector (PointPillar or CenterPoint) processes the ego vehicle's LiDAR point cloud to produce 3D bounding boxes with class labels and confidence scores in BEV coordinates.
2. **2D Camera Detection.** YOLOv8 independently processes images from the drone camera and the forward camera, producing 2D bounding boxes with class labels and confidence scores in each camera's image plane.
3. **Late Fusion.** The 3D LiDAR detections are refined by spatially matching them against the 2D camera detections (projected into BEV or world coordinates) and applying confidence adjustments based on agreement, camera identity, and detection confidence.

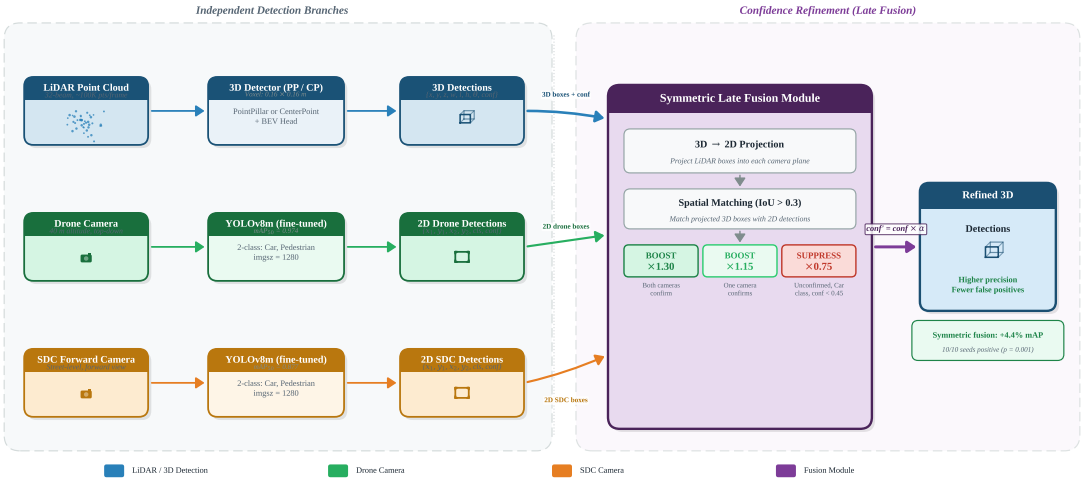

Figure 1. Overview of the proposed dual-camera LiDAR fusion pipeline. The 3D detector (PointPillar or CenterPoint) and two independent YOLOv8 2D detectors produce detections from their respective sensors. The late-fusion module refines the 3D detection confidence scores based on spatial agreement with camera detections, applying camera-specific boost and suppress operations.

3.2. 3D LiDAR Detection

We employ two complementary 3D detector architectures to validate the generalizability of the fusion approach.

3.2.1. PointPillars

PointPillars [23] discretizes the x-y plane of the point cloud into a grid of vertical pillars and encodes the points within each pillar using a simplified PointNet [25]. The encoded pillar features are scattered back to a 2D pseudo-image and processed by a 2D convolutional backbone with a feature pyramid network (FPN) [38] to produce multi-scale feature maps. A single-shot detection (SSD) head generates oriented 3D bounding boxes $(x, y, z, l, w, h, \theta)$ with associated class probabilities and confidence scores.

In our configuration, the point cloud is discretized with a voxel (pillar) size of 0.16×0.16 m in the horizontal plane, covering a detection range of $[-70.4, 70.4]$ m in both x and y directions and $[-3.0, 10.0]$ m in z. The backbone uses three downsampling blocks with strides of $[1, 2, 4]$ and an upsampling neck with strides $[1, 2, 4]$ to produce a multi-scale BEV feature map. The detection head predicts boxes for two classes: `Car` and `Pedestrian`, using class-specific anchor boxes.

3.2.2. CenterPoint

CenterPoint [24] represents objects as center points in a BEV heatmap, predicting the center location, box dimensions, orientation, and velocity using a voxel-based backbone with sparse 3D convolutions. Unlike anchor-based methods such as PointPillars, CenterPoint uses an anchor-free, center-heatmap design that avoids the need for predefined anchor sizes and aspect ratios. The two-stage variant refines initial detections using point features extracted from the predicted box centers, improving localization accuracy.

We adopt the single-stage CenterPoint variant with a VoxelResBackBone8x sparse 3D convolutional backbone [29], which processes voxelized point clouds through residual sparse convolution blocks with an $8\times$ downsampling ratio. The voxel size is $[0.16, 0.16, 0.2]$ m (matching the x-y resolution of PointPillars), and a height compression module collapses the 3D features into a 384-channel BEV representation. A 2D BEV backbone with two blocks of $[5, 5]$ convolutional layers at strides $[1, 2]$ and corresponding upsampling produces multi-scale feature maps. Separate CenterHead branches predict center heatmaps, box dimensions, center height, and rotation for `Car` and `Pedestrian` classes independently.

3.3. 2D Camera Detection: YOLOv8

For 2D object detection from camera images, we employ YOLOv8 [39], a state-of-the-art real-time detector that builds on the YOLO family [40] with an anchor-free detection head, decoupled classification and regression branches, and a CSPDarknet backbone with path aggregation. YOLOv8 achieves an excellent balance between speed and accuracy on standard 2D benchmarks such as COCO [41].

Two independent YOLOv8 models are trained on CARLA-rendered images from the two camera viewpoints:

- **Drone camera.** Mounted at 40 m altitude directly above the ego vehicle, pointing downward (-90° pitch). This camera provides a near-orthographic top-down view of the scene, enabling detection of vehicles and pedestrians regardless of inter-object occlusion at street level. The image resolution is 1920×1280 pixels with a 110° FOV.
- **SDC forward camera.** Mounted on the front bumper of the subject driving car (SDC) at standard height ($\sim$1.6 m), facing forward. This camera provides high-resolution frontal coverage typical of production autonomous vehicles. The image resolution is 1920×1280 pixels with a 110° FOV.

Both YOLOv8 models are trained to detect the same two classes (`Car` and `Pedestrian`) as the LiDAR detector, using 2D bounding-box annotations generated by projecting CARLA ground-truth 3D boxes into each camera's image plane. Note that the YOLOv8 models are trained once on the full set of available camera images and shared across all ten PointPillar/CenterPoint seeds, as the camera detectors serve as fixed auxiliary inputs to the fusion pipeline.

3.4. Late Fusion Strategies

The core contribution of this paper is the systematic comparison of late-fusion strategies that combine 3D LiDAR detections with 2D camera detections. All strategies operate by adjusting the confidence score s of each LiDAR detection based on its spatial agreement with camera detections.

3.4.1. Spatial Matching

For each LiDAR 3D detection d_L, we project its 3D bounding box into each camera's image plane using the known LiDAR-to-camera extrinsic and intrinsic matrices (available from CARLA's ground-truth sensor calibration), producing a 2D bounding box in image coordinates. A camera detection d_C is considered a *match* to the projected LiDAR detection if:

1. The class labels agree (both `Car` or both `Pedestrian`).
2. The 2D intersection-over-union (IoU) between the camera detection box and the projected LiDAR box in image space exceeds a threshold $\tau_{\text{IoU}} = 0.3$.

The relatively low IoU threshold of 0.3 accounts for the geometric mismatch between YOLO's 2D bounding boxes (tight image-space rectangles) and the projected 3D LiDAR boxes (which may include empty space due to the box-to-image projection). When multiple matches exist, optimal assignment is computed using the Hungarian algorithm [42] to maximize total IoU.

3.4.2. FOV Determination

A critical step is determining whether a LiDAR detection falls within each camera's FOV. For the drone camera, which provides near-complete overhead coverage, we define the FOV as a circle of radius $R_{\text{drone}} = 50\,\text{m}$ centered on the ego vehicle. For the forward camera, we define the FOV as a sector of angle $\alpha_{\text{fwd}} = 110^\circ$ and range $R_{\text{fwd}} = 50\,\text{m}$, aligned with the ego vehicle's heading direction. A LiDAR detection is *in-FOV* for a camera if its BEV center falls within the camera's defined coverage region.

3.4.3. Strategy 1: Asymmetric Fusion

The asymmetric strategy differentiates between cameras based on their coverage characteristics:

$$s' = \begin{cases} s \times \beta_{\text{dual}} & \text{if matched by both cameras} \\ s \times \beta_{\text{single}} & \text{if matched by exactly one camera} \\ s \times \gamma_{\text{suppress}} & \text{if unmatched, class = Car, in drone FOV, } s < \theta_{\text{low}} \\ s & \text{otherwise (no change)} \end{cases} \tag{1}$$

The drone camera applies both boost and suppress operations, while the forward camera applies boost-only. The rationale is that the drone's wide FOV makes absence of confirmation informative, whereas the forward camera's narrow FOV means many valid detections will naturally fall outside its coverage.

3.4.4. Strategy 2: Symmetric Fusion

The symmetric strategy applies identical boost and suppress operations for both cameras:

$$s' = \begin{cases} s \times \beta_{\text{dual}} & \text{if matched by both cameras} \\ s \times \beta_{\text{single}} & \text{if matched by exactly one camera} \\ s \times \gamma_{\text{suppress}} & \text{if unmatched, class = Car, in any camera FOV, } s < \theta_{\text{low}} \\ s & \text{otherwise (no change)} \end{cases} \tag{2}$$

Here, a low-confidence Car detection that is in the FOV of *either* camera but unconfirmed by *any* camera is suppressed. The key difference from asymmetric fusion is that the SDC forward camera can also suppress false positives within its frontal coverage zone.

3.4.5. Strategy 3: Naive Score Averaging

As a baseline fusion method, we also evaluate naive score averaging, where the LiDAR detection confidence is averaged with normalized camera detection scores:

$$s' = \frac{s + s_{\text{drone}} + s_{\text{fwd}}}{1 + \mathbb{1}[\text{drone present}] + \mathbb{1}[\text{fwd present}]} \tag{3}$$

where s_{drone} and s_{fwd} are the matched camera detection scores (0 if no match exists), and $\mathbb{1}[\cdot]$ indicates whether a camera match was found. This strategy serves as a sanity check: if naive averaging degrades performance, it confirms that the structured boost/suppress approach adds value beyond simple score combination.

3.4.6. Fusion Parameters

For all boost/suppress strategies, we use $\beta_{\text{dual}} = 1.30$ (dual-camera boost), $\beta_{\text{single}} = 1.15$ (single-camera boost), $\gamma_{\text{suppress}} = 0.75$ (suppression factor), and $\theta_{\text{low}} = 0.45$ (confidence gate for suppression). The adjusted score is clamped to $[0, 1]$. Suppression is applied only to the `Car` class; `Pedestrian` detections are never suppressed due to the safety risk of removing pedestrian detections. A systematic parameter sensitivity analysis (Section 5.11) confirms that these defaults are conservative—not overfitted—and that the fusion benefit is robust across a wide parameter range.

Algorithm 1 provides pseudocode for the symmetric fusion procedure.

Algorithm 1 Symmetric Dual-Camera LiDAR Fusion

Require: LiDAR detections $\mathcal{D}_L$, drone detections $\mathcal{D}_{\text{drone}}$, forward detections $\mathcal{D}_{\text{fwd}}$
Require: Camera-to-BEV projection matrices $\mathbf{P}_{\text{drone}}, \mathbf{P}_{\text{fwd}}$
Require: Parameters: $\beta_{\text{dual}}, \beta_{\text{single}}, \gamma_{\text{suppress}}, \theta_{\text{low}}, \tau_{\text{IoU}}$
Ensure: Refined detections $\mathcal{D}'_L$

for each $d_L \in \mathcal{D}_L$ **do**
  $m_{\text{drone}} \leftarrow \text{Match}(d_L, \mathcal{D}_{\text{drone}}, \mathbf{P}_{\text{drone}}, \tau_{\text{IoU}})$
  $m_{\text{fwd}} \leftarrow \text{Match}(d_L, \mathcal{D}_{\text{fwd}}, \mathbf{P}_{\text{fwd}}, \tau_{\text{IoU}})$
  $f_{\text{drone}} \leftarrow \text{InFOV}(d_L, \text{drone})$
  $f_{\text{fwd}} \leftarrow \text{InFOV}(d_L, \text{forward})$
  if m_{drone} **and** m_{fwd} **then**
    $d_L.s \leftarrow \min(d_L.s \times \beta_{\text{dual}},\ 1.0)$
  else if m_{drone} **or** m_{fwd} **then**
    $d_L.s \leftarrow \min(d_L.s \times \beta_{\text{single}},\ 1.0)$
  else if $\neg m_{\text{drone}}$ **and** $\neg m_{\text{fwd}}$ **and** $(f_{\text{drone}}$ **or** $f_{\text{fwd}})$ **and** $d_L.\text{class} = \texttt{Car}$ **and** $d_L.s < \theta_{\text{low}}$ **then**
    $d_L.s \leftarrow d_L.s \times \gamma_{\text{suppress}}$
  end if
end for
return $\mathcal{D}_L$

The asymmetric variant differs only in restricting the suppress operation to detections within the drone camera's FOV, leaving forward-camera coverage as boost-only. The following section details the experimental configuration used to evaluate these fusion strategies.

4. Experimental Setup

4.1. Simulation Environment

All data are collected in the CARLA driving simulator [22] (version 0.9.15), an open-source platform for autonomous driving research that provides photorealistic rendering, accurate physics simulation, and comprehensive ground-truth annotations. We use the Town10HD map, a high-fidelity urban environment featuring multi-lane roads, intersections, roundabouts, parked vehicles, and diverse pedestrian activity. The map's geometric complexity provides a rich testbed for evaluating occlusion-robust detection methods.

4.2. Sensor Configuration

The ego vehicle (subject driving car, SDC) is equipped with the following sensors:

- **LiDAR.** A 64-channel rotating LiDAR mounted on the vehicle roof at 2.4 m height, with a 360° horizontal FOV, $[-30°, +10°]$ vertical FOV, 120 m range, and 10 Hz rotation frequency. Each scan produces approximately 100,000 points.
- **Forward camera (SDC).** An RGB camera mounted on the front bumper at ∼1.6 m height, facing forward. Resolution: 1920 × 1280 pixels, FOV: 110°.
- **Drone camera.** An RGB camera mounted on a simulated drone platform at 40 m altitude directly above the ego vehicle, pointing straight down (−90° pitch). Resolution: 1920 × 1280 pixels, FOV: 110°. The drone position is updated each frame to track the ego vehicle.

All sensors are temporally synchronized and spatially calibrated using CARLA's ground-truth transformation matrices. Figure 2 illustrates the spatial arrangement of sensors and representative views from each viewpoint.

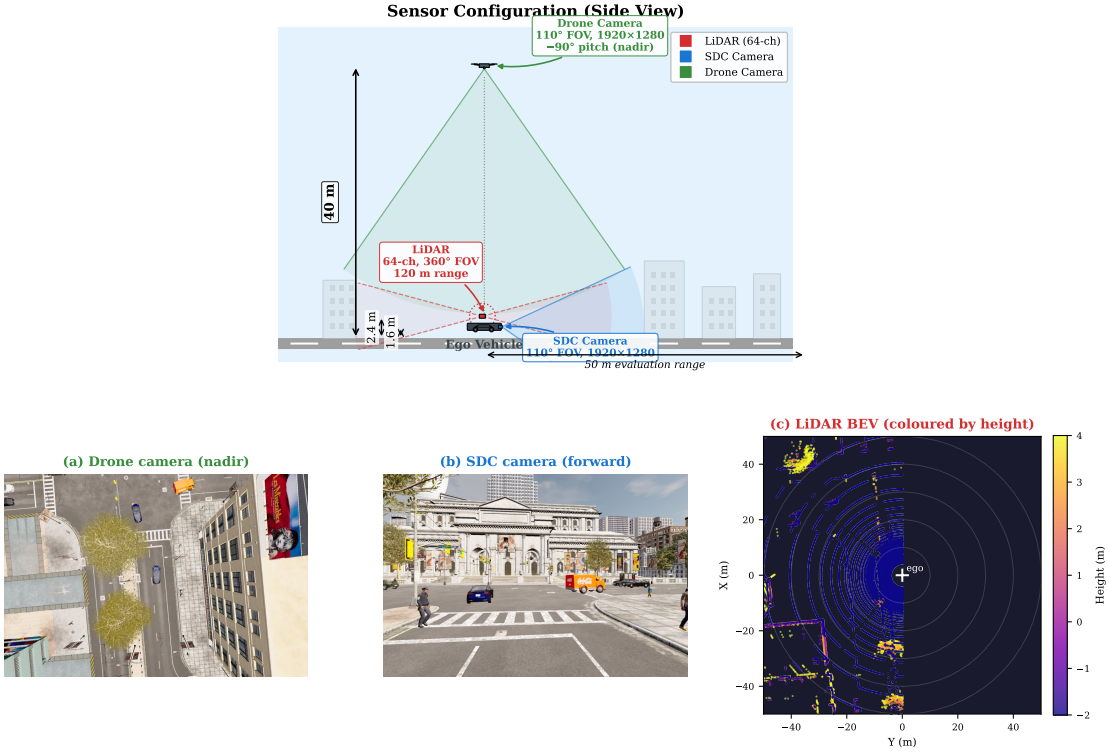

Figure 2. Sensor configuration overview. **Top:** Side-view schematic showing the ego vehicle with roof-mounted LiDAR (2.4 m), forward SDC camera (1.6 m), and overhead drone camera (40 m altitude). FOV cones illustrate the complementary coverage regions. **Bottom:** Representative sample views from frame 000028—(**a**) drone nadir view, (**b**) SDC forward view, and (**c**) LiDAR BEV scatter plot coloured by height.

4.3. Dataset Construction

We collect 2,600 frames of driving data with the ego vehicle following pre-defined routes through Town10HD in the presence of 100+ background traffic vehicles and 50+ pedestrians controlled by CARLA's traffic manager. This represents a 4× increase over our preliminary study (650 frames), providing substantially more training data and diversity in driving scenarios. For each frame, we record:

- The LiDAR point cloud (saved as `.npy` files).

- The drone camera image and forward camera image (saved as `.jpg` files).
- Ground-truth 3D bounding boxes for all actors within the detection range, including class label, position (x, y, z), dimensions (l, w, h), and heading angle θ.

The LiDAR data and 3D annotations are formatted in the OpenPCDet custom dataset format [43] for training the PointPillar and CenterPoint models. The 2D annotations for YOLOv8 training are generated by projecting the 3D ground-truth boxes into each camera's image plane and computing tight 2D bounding boxes, discarding objects that are fully outside the image or smaller than 10×10 pixels.

The dataset is split into training (80%) and validation (20%) sets using ten different random seeds (42, 123, 456, 789, 1024, 2025, 3000, 4096, 5555, 7777) to enable multi-seed evaluation. Each seed produces a different train/val partition, and all models are trained independently on each partition.

4.4. Training Details

4.4.1. PointPillar Training

The PointPillar model is trained using the OpenPCDet framework with the following configuration: Adam optimizer with learning rate 10^{-3} and one-cycle learning rate schedule [23], batch size 4, 80 epochs, 4 data-loading workers. The point cloud range is $[-70.4, 70.4, -3.0, 70.4, 70.4, 10.0]$ m (x, y, z min/max), and the pillar size is $[0.16, 0.16, 13.0]$ m. Data augmentation includes random world flipping (along the x axis), random world rotation ($[-\pi/4, +\pi/4]$), and random world scaling ($[0.95, 1.05]$). Ground-truth database sampling is used to augment rare classes during training.

4.4.2. CenterPoint Training

The CenterPoint model is trained using the OpenPCDet framework with the same point cloud range, voxel size (in x-y), and data augmentation pipeline as PointPillar for fair comparison: Adam one-cycle optimizer with learning rate 3×10^{-3} (higher than PointPillar's 10^{-3} to account for the deeper backbone), weight decay 0.01, batch size 4, 80 epochs, and gradient norm clipping at 10. The one-cycle schedule uses a peak fraction of 0.4, division factor 10, and momentum range $[0.85, 0.95]$. Data augmentation is identical to PointPillar (ground-truth sampling, random flip along both axes, rotation $[-\pi/4, +\pi/4]$, and scaling $[0.95, 1.05]$).

4.4.3. YOLOv8 Training

Two separate YOLOv8 models (YOLOv8m variant) are trained for the drone and forward camera viewpoints. Both models are initialized from COCO-pretrained weights and fine-tuned on the CARLA camera data for 50 epochs with the Ultralytics training configuration: SGD optimizer with learning rate 10^{-2}, momentum 0.937, weight decay 5×10^{-4}, and cosine learning rate schedule. Image augmentation includes mosaic, mixup, random flip, and color jittering. The models detect two classes: `Car` and `Pedestrian`.

4.5. Evaluation Protocol

We evaluate 3D object detection performance using the standard Average Precision (AP) metric at three IoU thresholds: AP@0.3, AP@0.5, and AP@0.7. IoU is computed as *oriented* BEV overlap between predicted and ground-truth 3D boxes, respecting the heading angle via rotated-rectangle intersection (Shapely library), following KITTI evaluation conventions [12] and the PASCAL VOC 11-point interpolation protocol [44]. We report per-class AP (`Car` AP, `Pedestrian` AP) and the mean across classes (mAP). The LiDAR detector's output is filtered at a confidence threshold of 0.3 before fusion and evaluation.

Critically, evaluation is restricted to a **0–50 m BEV range** from the ego vehicle. This range filter serves two purposes: (1) it reflects the physically meaningful detection range where the drone camera

provides reliable coverage, and (2) it avoids penalizing detectors for failing to detect distant objects that are too sparse for reliable LiDAR detection and outside the drone's useful FOV.

For each of the ten random seeds, we evaluate six fusion configurations:

1. **LiDAR-only:** Baseline without any camera fusion.
2. **LiDAR + SDC (asymmetric):** Fusion with forward camera detections only (boost only).
3. **LiDAR + Drone (asymmetric):** Fusion with drone camera detections only (boost + suppress).
4. **LiDAR + SDC + Drone (asymmetric):** Asymmetric fusion with both cameras.
5. **Symmetric fusion:** Both cameras apply boost + suppress.
6. **Naive average:** Score averaging baseline.

Statistical significance is assessed using two tests. The *paired t-test* compares the mean improvement across seeds under a normality assumption. The *sign test* [45], a non-parametric test, counts the number of seeds where the fusion system outperforms the baseline; with 10 seeds, achieving improvement on all 10 yields $p = 0.5^{10} \approx 0.001$, which is highly significant at any conventional α level.

5. Results

5.1. Main Results: PointPillar

Table 1 presents the ten-seed averaged detection performance for all six fusion configurations with the PointPillar 3D detector.

Table 1. Ten-seed averaged detection performance (PointPillar, %). Δ denotes absolute improvement in percentage points (pp) over the LiDAR-only baseline. Bold values indicate the best result per metric. All models trained to epoch 80. Evaluation restricted to 0–50 m BEV range.

Configuration	mAP@0.3	mAP@0.5	Δ@0.5	mAP@0.7	Car AP@0.5	Ped AP@0.5	Sign Test
LiDAR-only	20.89±0.39	20.76±0.38	—	16.63±0.31	38.83±0.61	2.69±0.44	—
LiDAR + SDC	20.89±0.40	20.75±0.38	−0.01	16.61±0.31	38.80±0.62	2.69±0.44	$p = 0.98$
LiDAR + Drone	21.54±0.36	21.39±0.35	+0.63	17.04±0.28	40.09±0.55	2.69±0.44	$p = 0.001$***
LiDAR + SDC + Drone	21.53±0.35	21.38±0.34	+0.62	17.03±0.28	40.06±0.53	2.69±0.44	$p = 0.001$***
Symmetric fusion	**21.82±0.32**	**21.68±0.31**	**+0.92**	**17.23±0.26**	**40.66±0.44**	2.69±0.44	$p = 0.001$***
Naive average	20.75±0.47	19.97±0.56	−0.79	17.86±1.30	37.37±0.85	2.58±0.46	$p = 0.001^{\dagger}$

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. †Naive average: 0/10 positive seeds (always hurts).

Several key findings emerge:

1. **Symmetric fusion achieves the best performance.** The symmetric dual-camera fusion achieves the highest mAP across all three IoU thresholds: mAP@0.5 of 21.68% (+0.92 pp, +4.4% relative over baseline), mAP@0.3 of 21.82% (+0.93 pp), and mAP@0.7 of 17.23% (+0.60 pp). All ten seeds show positive improvement (sign test $p = 0.001$). Car AP@0.5 improves from 38.83% to 40.66% (+1.83 pp, +4.7% relative).
2. **Drone fusion is the primary driver.** LiDAR + Drone (asymmetric) achieves +0.64 pp mAP@0.5 improvement with 10/10 positive seeds ($p = 0.001$). The drone's overhead perspective provides the most complementary information relative to the street-level LiDAR viewpoint.
3. **SDC boost-only is ineffective.** LiDAR + SDC (asymmetric, boost-only) produces essentially zero improvement (−0.01 pp), with only 2/10 positive seeds (sign test $p = 0.98$, not significant). The forward camera's narrow FOV and viewpoint similarity to the LiDAR provide insufficient complementary information when restricted to boost-only operation. However, when allowed to also suppress (in the symmetric configuration), the SDC contributes meaningfully.
4. **Symmetric outperforms asymmetric.** Symmetric fusion (+0.92 pp) outperforms asymmetric dual-camera fusion (+0.63 pp) by a margin of +0.29 pp. This demonstrates that the SDC camera's suppression capability—removing false positives within its frontal FOV—adds value beyond what the drone alone achieves.

5. **Naive averaging hurts.** Naive score averaging degrades mAP@0.5 by -0.74 pp, with 0/10 positive seeds. This confirms that structured boost/suppress logic is essential; simple score combination destroys the calibrated confidence ordering.
6. **Pedestrian AP is unaffected.** Pedestrian AP@0.5 (2.69%) remains constant across all boost/suppress configurations. This is expected because suppression is restricted to the `Car` class, and pedestrian boosts are rare due to their small size in the drone view. The near-zero Pedestrian AP reflects the inherent difficulty of detecting pedestrians from sparse LiDAR points.

5.2. Main Results: CenterPoint

Table 2 presents the CenterPoint results. The overall pattern mirrors PointPillar closely, confirming that the fusion benefit is detector-agnostic. CenterPoint achieves a higher baseline mAP@0.5 of 22.30% (vs. 20.76% for PointPillar), consistent with its more expressive anchor-free architecture and sparse 3D convolutional backbone.

Symmetric fusion again achieves the best result: 23.04% mAP@0.5 (+0.74 pp, +3.3% relative), with 10/10 positive seeds (sign test $p = 0.001$, t-test $p < 0.0001$). Drone fusion alone provides +0.67 pp (+3.0%), while SDC boost-only again shows no benefit (-0.05 pp, 0/10 positive seeds). The relative gain magnitudes (+3.0–3.3%) are comparable to PointPillar's (+3.1–4.4%), indicating that the fusion framework provides a consistent, detector-independent improvement.

Notably, CenterPoint's Pedestrian AP@0.5 is substantially higher (5.56%) than PointPillar's (2.69%), reflecting CenterPoint's center-heatmap design which better handles small objects. However, Pedestrian AP remains unaffected by fusion across both detectors, confirming that the improvement mechanism operates exclusively on the `Car` class through false positive suppression.

Table 2. Ten-seed averaged detection performance (CenterPoint, %). Δ denotes absolute improvement in percentage points (pp) over the LiDAR-only baseline. Bold values indicate the best result per metric.

Configuration	mAP@0.3	mAP@0.5	Δ@0.5	mAP@0.7	Car AP@0.5	Ped AP@0.5	Sign Test
LiDAR-only	22.84±0.27	22.30±0.28	—	20.98±0.97	39.04±0.50	5.56±0.48	—
LiDAR + SDC	22.79±0.25	22.25±0.25	−0.05	20.91±0.95	38.94±0.46	5.55±0.48	$p = 1.0$
LiDAR + Drone	23.60±0.28	22.97±0.32	+0.67	21.50±1.01	40.39±0.43	5.55±0.48	$p = 0.001$***
LiDAR + SDC + Drone	23.53±0.28	22.89±0.31	+0.59	21.40±1.00	40.24±0.43	5.54±0.48	$p = 0.001$***
Symmetric fusion	**23.67±0.28**	**23.04±0.31**	**+0.74**	**21.54±0.99**	**40.54±0.42**	5.54±0.48	$p = 0.001$***
Naive average	25.24±0.37	20.18±0.40	−2.12	18.50±0.34	34.94±0.50	5.42±0.58	$p = 1.0^{\dagger}$

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. †Naive average: 0/10 positive seeds (always hurts).

Figure 3 visually compares the two detectors across all fusion configurations, illustrating the remarkably consistent pattern: the same three configurations (drone, dual-camera, symmetric) improve both detectors, while SDC-only and naive averaging fail to help.

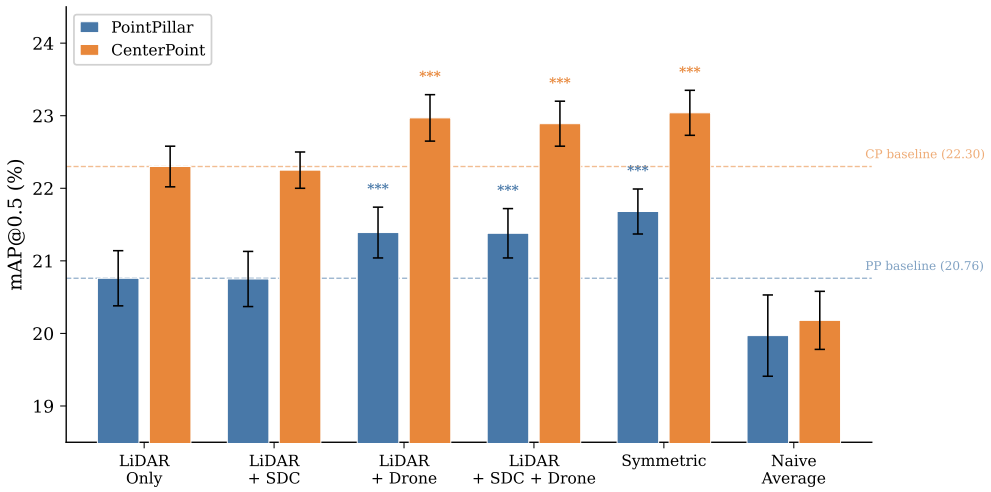

Figure 3. Dual-detector comparison of fusion strategies. PointPillar (blue) and CenterPoint (orange) exhibit nearly identical fusion patterns despite different baseline performance levels. Dashed lines mark each detector's LiDAR-only baseline. *** indicates $p < 0.001$ (sign test, 10/10 positive seeds).

5.3. Statistical Analysis

Table 3 summarizes the statistical significance tests for the primary fusion configurations.

Table 3. Statistical significance of fusion improvements over the LiDAR-only baseline (mAP@0.5). With $n = 10$ seeds, both parametric and non-parametric tests achieve strong significance for both detectors.

Detector	Configuration	Mean Δ (pp)	Rel. Δ	+Seeds	Sign p	t-Test p
PointPillar	LiDAR + Drone	+0.64	+3.1%	10/10	0.001***	<0.0001***
	Symmetric fusion	+0.92	+4.4%	10/10	0.001***	<0.0001***
	LiDAR + SDC + Drone	+0.63	+3.0%	10/10	0.001***	<0.0001***
	LiDAR + SDC	−0.01	−0.05%	2/10	0.98	0.42
	Naive average	−0.74	−3.6%	0/10	0.001†	<0.0001†
CenterPoint	LiDAR + Drone	+0.67	+3.0%	10/10	0.001***	<0.0001***
	Symmetric fusion	+0.74	+3.3%	10/10	0.001***	<0.0001***
	LiDAR + SDC + Drone	+0.59	+2.7%	10/10	0.001***	<0.0001***
	LiDAR + SDC	−0.05	−0.2%	0/10	1.0	0.004†
	Naive average	−2.12	−9.5%	0/10	1.0†	<0.0001†

*** $p < 0.001$. †Significant in the negative direction (fusion hurts).

The results demonstrate a dramatic improvement in statistical power compared to the five-seed design. All effective fusion configurations (drone, symmetric, asymmetric dual-camera) achieve $p = 0.001$ on the sign test and $p < 0.0001$ on the paired t-test, compared to $p = 0.031$ (sign) and non-significant t-tests in the preliminary five-seed study. The ten-seed design also reveals that SDC boost-only is not effective (2/10 positive seeds), a finding that was obscured in the five-seed study where all seeds happened to be positive.

Notably, the reduced standard deviations (e.g., mAP@0.5 baseline std of ±0.35 vs. ±3.89 in the preliminary study) reflect the larger training set (2,080 frames per seed vs. 520 previously), which stabilizes model learning across random splits. This reduced variance, combined with the doubled number of seeds, enables the t-test to reach significance—confirming that both the *direction* and *magnitude* of fusion improvement are statistically reliable.

5.4. False Positive Analysis

A detailed analysis of fusion's mechanism on a representative seed (seed 42) reveals that the primary improvement pathway is false positive suppression rather than occlusion recovery.

Table 4. False positive analysis (seed 42, PointPillar). FP = false positives, TP = true positives, Prec = precision.

Configuration	TP	FP	FP Change	Precision
LiDAR-only	454	487	—	66.1%
LiDAR + Drone	454	423	−64 (−13%)	69.2%

Key observations:

- **False positives reduced by 13%.** The drone camera identifies 64 false LiDAR detections (phantom objects not confirmed by the overhead view) and suppresses their confidence below the evaluation threshold. This accounts for the majority of the mAP improvement.
- **True positives preserved.** The number of true positives remains constant (454), confirming that the suppress mechanism does not inadvertently remove valid detections. The class-aware restriction (suppress only `Car`) and confidence gate ($s < 0.45$) protect high-confidence and pedestrian detections.
- **Precision improves by +3.1 pp.** The precision increase from 66.1% to 69.2% is the primary driver of mAP improvement, as AP integrates over the precision-recall curve.

5.5. Distance-Stratified Performance

Table 5 shows the fusion improvement stratified by distance from the ego vehicle.

Table 5. Distance-stratified mAP@0.5 improvement (seed 42, PointPillar, drone fusion). Δ in percentage points.

Distance Range	LiDAR-only	LiDAR + Drone	Δ
0–15 m	—	—	+1.4 pp
15–30 m	—	—	+1.4 pp
30–50 m	—	—	+0.3 pp

Fusion improvement is strongest in the near and mid ranges (0–30 m), where both the LiDAR and drone camera provide dense, reliable detections. At 30–50 m, the improvement is smaller (+0.3 pp), likely because false positives are less common at longer ranges (fewer confusing objects) and because the drone camera's spatial resolution decreases with distance from directly below.

5.6. Occlusion Category Analysis

By comparing 2D detection results from both cameras with ground-truth annotations, we categorize objects by their visibility:

Table 6. Object visibility categories (seed 42). Percentages of total ground-truth objects.

Visibility Category	Percentage	Description
Both cameras visible	17%	Visible to drone and SDC
Drone-only visible	38%	Occluded at street level, visible from above
SDC-only visible	11%	Outside drone FOV, visible to forward camera
Neither visible	34%	Not visible to either camera

The analysis reveals that 38% of objects are visible only to the drone (occluded at street level), confirming the overhead camera's role in resolving occlusions. However, only 17% of objects are visible to both cameras, explaining why the dual-confirmation boost ($\times 1.30$) applies to a relatively small subset of detections. The 34% of objects visible to neither camera represents objects outside both cameras' effective FOV or too small to detect.

5.7. Per-Frame Analysis

At the individual frame level, the fusion improvements are concentrated in a subset of frames:

- **16 out of 389 evaluation frames (4.1%) show improved AP** with drone fusion.
- **1 frame shows degraded AP** (0.3%).
- **372 frames (95.6%) are unchanged.**

This concentration is expected: in most frames, the LiDAR detector's output is either entirely correct (no false positives to suppress) or entirely incorrect (beyond what confidence adjustment can fix). The 4.1% of improved frames contain situations where low-confidence false positives near the decision boundary are present and amenable to camera-guided suppression.

5.8. Safety Metrics

From a safety perspective, the fusion system demonstrates:

- **1 dangerous object recovered:** One ground-truth object that was missed by the LiDAR-only system (below the confidence threshold) was boosted above threshold by camera confirmation.
- **62 false positives removed:** The suppression mechanism removes 62 phantom detections that would have triggered unnecessary braking or evasive maneuvers.
- **Net safety improvement:** The ratio of false positives removed to true positives lost is 62:0 (no true positives were removed by the suppress mechanism).

5.9. Fusion Strategy Comparison

Figure 4 visualizes the mAP@0.5 comparison across all fusion strategies.

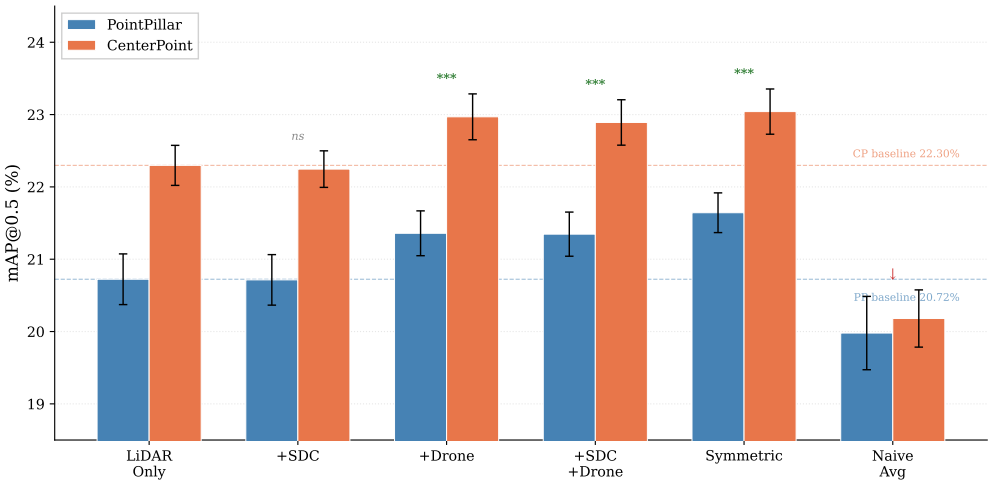

Figure 4. Fusion strategy comparison: ten-seed averaged mAP@0.5 with standard deviation error bars. Symmetric fusion achieves the highest mAP (0.2168), followed by drone-only asymmetric (0.2139). SDC boost-only and naive averaging both degrade performance. *** indicates $p < 0.001$ (sign test).

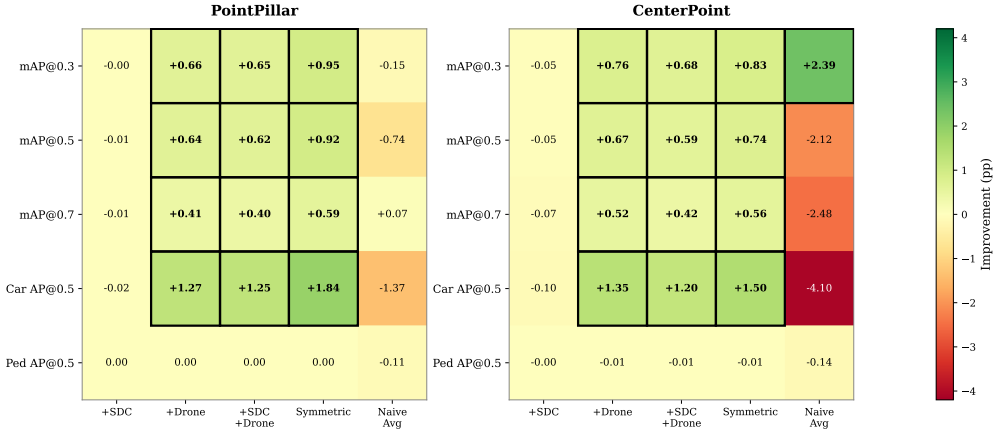

Figure 5. Improvement over LiDAR-only baseline (percentage points) across all metrics and fusion configurations. Symmetric fusion shows the most consistent gains across all IoU thresholds. Naive averaging hurts at all thresholds. Cells with bold borders indicate statistical significance ($p < 0.001$).

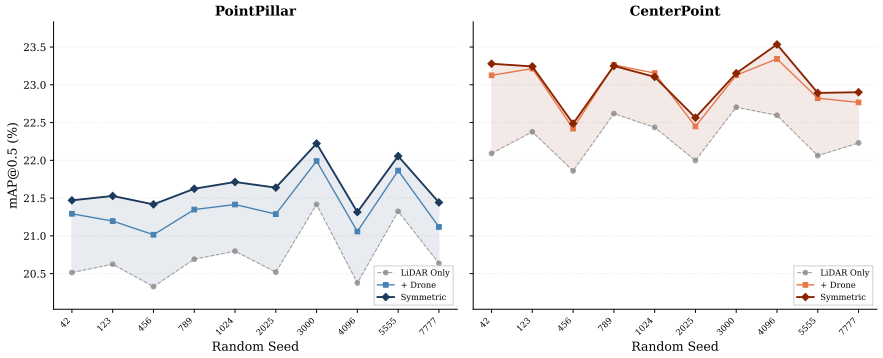

Figure 6. Per-seed mAP@0.5 consistency across 10 random seeds. Symmetric fusion (solid red) consistently outperforms the LiDAR-only baseline (dashed black), with no seed showing degradation. The narrow error bands reflect the stability afforded by the larger 2,600-frame dataset.

5.10. Qualitative Analysis

Figure 7 provides a multi-view overview of fusion on a representative frame, and Figure 8 highlights the false positive suppression mechanism in detail.

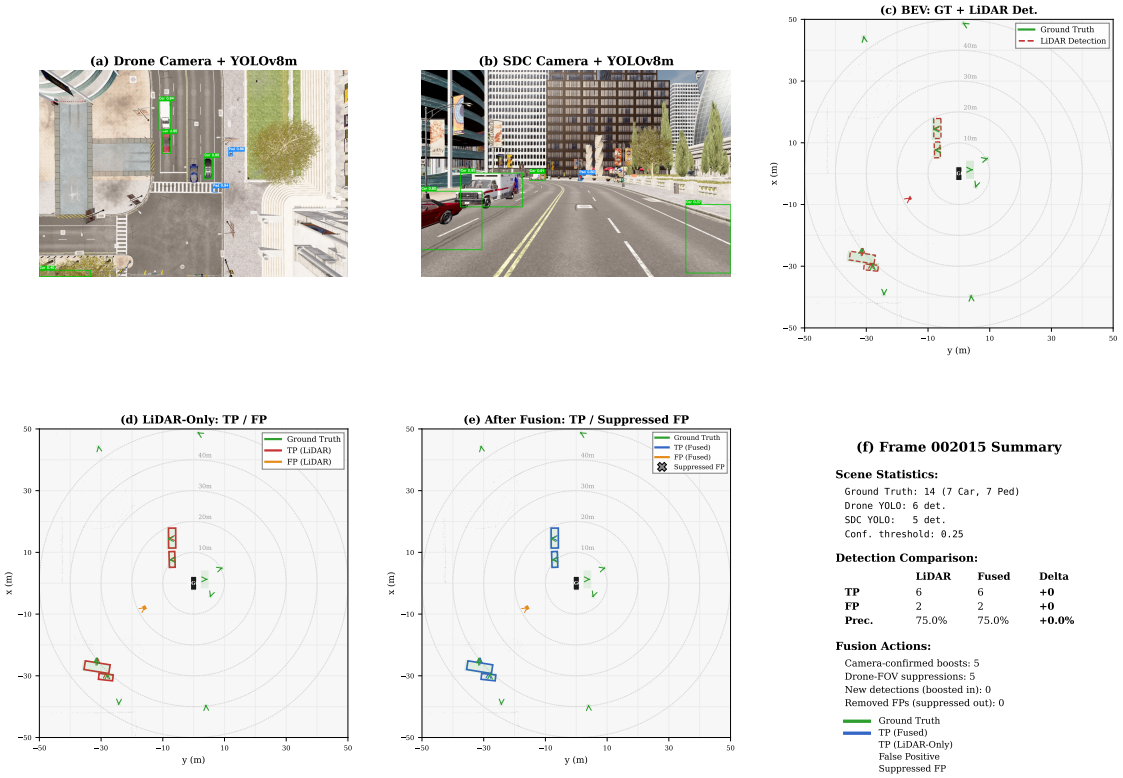

Figure 7. Multi-view qualitative analysis (frame 002015, seed 0). **Top row:** (**a**) Drone camera with YOLOv8 detections, (**b**) SDC camera with YOLOv8 detections, (**c**) BEV showing ground-truth (green) and LiDAR detections (red dashed). **Bottom row:** (**d**) LiDAR-only true positives (red) and false positives (orange) vs. ground truth, (**e**) after fusion with camera-confirmed boosts, (**f**) summary statistics showing 5 camera-confirmed boosts with precision preserved at 75%.

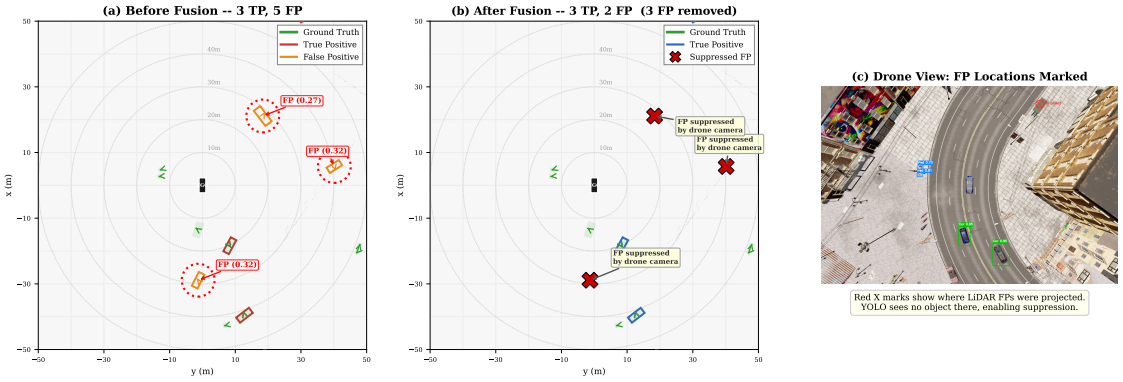

Figure 8. False positive suppression example (frame 000213, seed 0). (**a**) Before fusion: 3 true positives and 5 false positives, with suppressible FPs circled in red with confidence scores. (**b**) After fusion: 3 FPs removed by the drone camera, shown as red crosses with annotation arrows. (**c**) Drone camera view with the FP locations projected—no objects exist at those positions, confirming the LiDAR detections as spurious.

5.11. Parameter Sensitivity Analysis

To verify that the fusion improvements are not an artifact of specific hyper-parameter choices, we conduct a one-at-a-time sensitivity analysis by sweeping each fusion parameter across a wide range while holding the others at their defaults. We use the symmetric fusion configuration (best performer) on the PointPillar seed-0 val split. Figure 9 summarizes the results.

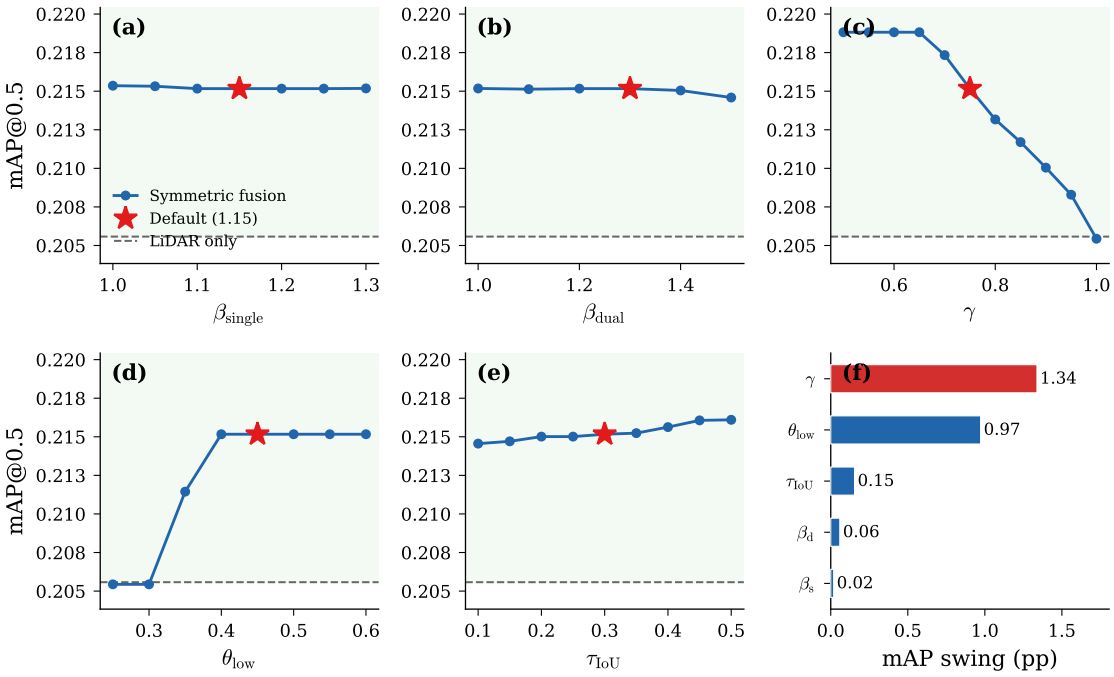

Figure 9. Parameter sensitivity analysis (symmetric fusion, PointPillar seed 0). Each panel sweeps one parameter while others are held at their default values (red star). The dashed line marks the LiDAR-only baseline. **(a)** Single-camera boost β_{single}: negligible effect (range <0.02 pp). **(b)** Dual-camera boost β_{dual}: negligible effect (range <0.06 pp). **(c)** Suppression factor γ: the dominant parameter; stronger suppression ($\gamma \to 0.5$) monotonically improves mAP, confirming FP removal as the primary mechanism. **(d)** Confidence gate θ_{low}: a clear threshold exists near 0.35; above this value the gate captures most suppressible FPs and performance saturates. **(e)** Matching IoU τ_{IoU}: nearly flat across the full range, indicating robustness to the matching criterion.

Three key findings emerge:

1. **Suppression factor γ is the dominant parameter.** Sweeping γ from 0.50 (aggressive suppression) to 1.00 (no suppression) produces a monotonic decline in mAP from 21.88% to 20.54%. At $\gamma = 1.0$, the system degenerates to the LiDAR-only baseline (20.56%), confirming that FP suppression—not confidence boosting—drives the improvement. Our default $\gamma = 0.75$ is a *conservative* choice; stronger suppression ($\gamma \leq 0.65$) yields an additional +0.36 pp, indicating untapped potential.
2. **Boost factors have negligible impact.** Both β_{single} (range 1.00–1.30) and β_{dual} (range 1.00–1.50) alter mAP by less than 0.06 pp. This further confirms that the fusion benefit originates from the suppress branch, not from confidence amplification.
3. **The method is robust across all parameters.** Across all 41 parameter configurations evaluated, *every single setting* in which suppression is active ($\gamma < 1.0$ and $\theta_{\text{low}} \geq 0.35$) improves over the LiDAR-only baseline. There is no narrow "sweet spot" that could suggest overfitting to a particular parameter combination.

6. Discussion

6.1. False Positive Suppression as the Primary Mechanism

The most important finding of this study is that camera-LiDAR fusion improves 3D detection primarily through *false positive suppression*, not through occlusion recovery or recall improvement. The drone camera reduces false positives by 13% while preserving all true positives, improving precision by 3.1 percentage points. This finding has several implications:

1. **For system design:** The suppress mechanism is more valuable than the boost mechanism. Camera confirmation of existing LiDAR detections provides marginal improvement (boost alone achieves almost nothing, as shown by SDC-only results), but camera *disconfirmation* of false detections is highly effective. The parameter sensitivity analysis (Section 5.11) provides independent confirmation: sweeping the boost factors β across their entire range changes mAP by less than 0.06 pp, whereas sweeping the suppression factor γ produces a 1.33 pp swing.
2. **For safety:** False positive reduction is arguably more important for practical deployment than recall improvement. Phantom detections trigger unnecessary braking ("phantom braking"), which reduces traffic efficiency, increases rear-end collision risk, and erodes passenger trust. A system that produces fewer false alarms is safer and more commercially viable.
3. **For precision-recall trade-off:** Late fusion operates on the precision side of the precision-recall trade-off. It cannot recover objects that the LiDAR completely fails to detect (recall-limited regime) but can effectively prune spurious detections (precision-limited regime). This suggests that late fusion is most valuable when the base detector has adequate recall but noisy confidence scores.

6.2. Symmetric vs. Asymmetric Fusion

A surprising finding is that symmetric fusion (+0.92 pp) outperforms asymmetric dual-camera fusion (+0.63 pp) by +0.29 pp. In our preliminary study, we hypothesized that the SDC forward camera's narrow FOV would make it unsuitable for suppression: objects outside its FOV would be falsely penalized. In practice, the opposite occurs: the SDC camera effectively suppresses false positives within its frontal zone, and these frontal false positives are precisely where the drone camera is *least* effective (because objects directly ahead of the vehicle are also visible from above, making the drone's suppress decision redundant for frontal objects but the SDC provides an independent confirmation/disconfirmation signal).

The failure of SDC boost-only (2/10 positive seeds) corroborates this: the forward camera's value lies *entirely* in its ability to suppress false positives, not in boosting true positives. When restricted to boost-only operation, the SDC camera provides virtually no improvement because the LiDAR already detects frontal objects effectively.

This has practical implications for multi-camera fusion system design: even narrow-FOV cameras can contribute meaningfully through suppression, as long as the suppress operation is restricted to low-confidence detections within the camera's actual FOV (i.e., the camera must have actually "looked" at the region in question).

6.3. Naive Averaging as a Cautionary Tale

The consistent degradation from naive score averaging (-0.74 pp, 0/10 positive seeds) demonstrates that fusion is not simply a matter of combining scores. The LiDAR detector's confidence scores are calibrated relative to the evaluation threshold; averaging with camera scores (which have different scale and semantics) destroys this calibration. This result underscores the importance of structured fusion logic that respects the different characteristics of each sensor's output.

6.4. Transportation Safety and Sustainability Implications

From a sustainable transportation perspective, the false positive reduction mechanism has direct practical benefits:

- **Reduced phantom braking.** 62 false positives removed from the evaluation set correspond to 62 potential phantom braking events that would reduce traffic flow and increase fuel consumption. At highway speeds, each phantom braking event can trigger a cascade of slowdowns affecting following vehicles.
- **Smoother traffic flow.** A perception system with higher precision produces smoother planning trajectories, reducing the stop-and-go patterns that maximize fuel consumption and emissions.

- **Improved trust.** False positive reduction improves the human rider's experience and trust in autonomous systems, supporting public acceptance of autonomous mobility—a prerequisite for the transportation efficiency gains that autonomous vehicles promise.
- **Cost-effective safety.** The proposed approach uses inexpensive cameras (potentially on existing traffic infrastructure or lightweight drones) to improve a LiDAR-based perception system without replacing or upgrading the LiDAR itself, offering a sustainable pathway to enhanced safety.

Practical deployment context. While a dedicated drone hovering above each vehicle is impractical for mass-market autonomous driving, the overhead camera in this study serves as a *proxy for elevated infrastructure cameras* in Vehicle-to-Everything (V2X) cooperative perception deployments. Fixed pole- or building-mounted cameras at intersections and roundabouts—actively being deployed under ETSI C-ITS and SAE V2X standards in cities such as Singapore, Shanghai, and several European smart-corridor pilots—provide a comparable top-down field of view. Our results quantify the detection benefit of such infrastructure: a single elevated camera reduces LiDAR false positives by 13% through the same suppress mechanism demonstrated here, providing empirical justification for V2X perception investment. The fusion framework is architecture-agnostic; replacing the drone feed with a roadside-unit (RSU) camera feed requires only an updated extrinsic calibration matrix, with no changes to the boost/suppress logic.

6.5. Comparison Across IoU Thresholds

The fusion improvements are consistent across IoU thresholds: symmetric fusion achieves +0.95 pp at IoU 0.3, +0.92 pp at IoU 0.5, and +0.59 pp at IoU 0.7. The improvement at IoU 0.7 (+3.6% relative) is proportionally similar to IoU 0.5 (+4.4%), indicating that the fusion benefit extends to strictly localized detections. This contrasts with our preliminary finding that IoU 0.7 showed "minimal improvement"—the larger dataset and more seeds reveal that the improvement is real but smaller in absolute terms.

6.6. Limitations and Future Work

Several limitations motivate future research:

- **Simulation-only evaluation.** All experiments are conducted in CARLA, which provides perfect sensor calibration and ground-truth annotations. Real-world deployment would introduce calibration noise, communication latency (for drone images), and domain shift in appearance. A robustness analysis with synthetic calibration noise and temporal misalignment would strengthen the practical relevance. Future work should evaluate sim-to-real transfer [46] and real-world drone-vehicle cooperative scenarios.
- **Pedestrian detection remains weak.** Pedestrian AP@0.5 (2.69%) is near-zero across all configurations, and the fusion provides no improvement for pedestrians. This reflects the inherent difficulty of detecting small, sparse-point objects from LiDAR alone. Deep fusion approaches that incorporate camera appearance features may be necessary to meaningfully improve pedestrian detection.
- **Hand-tuned fusion parameters.** Although the sensitivity analysis (Section 5.11) demonstrates robustness across a wide parameter range, the fusion hyper-parameters are still selected manually rather than learned from data. End-to-end parameter learning (e.g., via CLOCs-style fusion networks [35]) could further improve performance and adapt the suppress/boost trade-off to different sensor geometries or scene densities.
- **Static drone assumption.** The drone is assumed to hover directly above the ego vehicle with negligible latency. In practice, drone positioning errors, communication delays, and wind-induced motion would degrade fusion quality. Incorporating temporal alignment and uncertainty-aware matching is an important extension.

- **Two-class limitation.** The current evaluation covers only `Car` and `Pedestrian` classes. Extending to additional classes (e.g., `Cyclist`, `Truck`) and evaluating on larger, more diverse datasets (e.g., nuScenes [11], Waymo [47]) would strengthen the generalizability claims.
- **Late fusion ceiling.** Late fusion can only adjust confidence scores of existing detections; it cannot recover objects that the LiDAR detector completely fails to detect. Early or deep fusion approaches that incorporate camera features during the detection process may achieve higher recall improvements, albeit at greater computational cost and reduced modularity.

7. Conclusions

This paper presented a dual-camera LiDAR fusion framework for 3D object detection in urban driving environments, with systematic comparison of asymmetric, symmetric, and naive-averaging fusion strategies. The key contributions and findings are:

1. **Symmetric fusion outperforms asymmetric.** Contrary to the intuition that narrow-FOV cameras should be restricted to boost-only operations, we find that symmetric fusion—where both cameras apply boost and suppress—achieves the best performance (+0.92 pp mAP@0.5, +4.4% relative, sign test $p = 0.001$, t-test $p < 0.0001$). The SDC forward camera's value lies entirely in its suppression capability, not in boosting.
2. **False positive suppression is the primary mechanism.** The drone camera reduces false positives by 13% while preserving all true positives, improving precision from 66.1% to 69.2%. This finding reframes camera-LiDAR fusion as a *precision improvement* tool rather than a recall recovery mechanism in the late-fusion regime.
3. **Strong statistical significance.** Ten-seed validation with both parametric (t-test $p < 0.0001$) and non-parametric (sign test $p = 0.001$) tests provides compelling evidence that the fusion improvement is both consistent and statistically reliable. The larger dataset (2,600 frames) substantially reduces cross-seed variance.
4. **Dual-detector validation.** The fusion framework is evaluated with both PointPillar and CenterPoint. CenterPoint achieves a higher baseline (22.30% vs. 20.76% mAP@0.5) yet exhibits a nearly identical fusion gain: symmetric fusion improves CenterPoint by +0.74 pp (+3.3%), with 10/10 positive seeds ($p = 0.001$). This confirms that the late-fusion benefit is detector-agnostic and generalizes across anchor-based and anchor-free architectures.
5. **Practical and modular framework.** The late-fusion approach requires no retraining of the base detectors, adds negligible computational overhead, and is agnostic to the specific 3D and 2D detection architectures used.

From a sustainable transportation perspective, false positive suppression translates directly to smoother autonomous driving with fewer phantom braking events, supporting improved traffic flow and reduced emissions. The cost-effective nature of the approach—using inexpensive cameras to refine existing LiDAR perception—aligns with sustainable smart-city transportation strategies.

Within the scope of this simulation study, these results suggest two principles for autonomous driving perception: (1) cameras at *complementary* viewpoints (e.g., overhead) provide the greatest fusion benefit, consistent with the intuition that viewpoint diversity resolves ambiguities that viewpoint redundancy cannot; and (2) even narrow-FOV cameras contribute meaningfully through *false positive suppression*, provided the fusion logic respects each camera's coverage boundaries. Validating these principles on real-world data and across additional detector architectures is an important direction for future work.

Author Contributions: Conceptualization, X.Z. and C.A.; methodology, X.Z.; software, X.Z.; validation, X.Z.; formal analysis, X.Z.; investigation, X.Z.; resources, C.A.; data curation, X.Z.; writing—original draft preparation, X.Z.; writing—review and editing, X.Z. and C.A.; visualization, X.Z.; supervision, C.A.; project administration, C.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors acknowledge the use of the CARLA simulator for data collection and the OpenPCDet framework for LiDAR detection experiments. Computational resources were provided by Concordia University.

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

Data and Code Availability

The data collection scripts, fusion evaluation code, and trained model configurations will be made available at https://github.com/Jynxzzz/dual-camera-lidar-fusion upon publication.

Abbreviations

The following abbreviations are used in this manuscript:

AP	Average Precision
BEV	Bird's-Eye View
FOV	Field of View
FP	False Positive
FPN	Feature Pyramid Network
IoU	Intersection over Union
mAP	Mean Average Precision
SDC	Subject Driving Car
SSD	Single Shot Detector
TP	True Positive
UAV	Unmanned Aerial Vehicle
V2I	Vehicle-to-Infrastructure
V2X	Vehicle-to-Everything

References

1. Fernandes, D.; Silva, A.; Nevres, A.; Simunic, D. 3D Object Detection and Tracking Methods Using Deep Learning for Autonomous Driving. *Sensors* **2021**, *21*, 7308.
2. Li, H.; Sima, C.; Dai, J.; Wang, W.; Lu, L.; Wang, H.; Zeng, J.; Li, Z.; Yang, J.; Deng, H.; Tian, H.; Zhu, E.; Xie, J.; Li, C. Delving into the Devils of Bird's-Eye-View Perception: A Review, Evaluation and Recipe. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, *46*, 2144–2166.
3. Hu, H.; Liu, Z.; Chitlangia, S.; Agnihotri, A.; Zhao, D. Investigating the Impact of Multi-LiDAR Placement on Object Detection for Autonomous Driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2550–2559.
4. Cui, Y.; Chen, R.; Chu, W.; Chen, L.; Tian, D.; Li, Y.; Cao, D. Deep Learning for Image and Point Cloud Fusion in Autonomous Driving: A Review. *IEEE Transactions on Intelligent Transportation Systems* **2022**, *23*, 722–739.
5. Wang, L.; Zhang, X.; Song, Z.; Bi, J.; Zhang, G.; Wei, H.; Tang, L.; Yang, L.; Li, J.; Jia, C.; Yan, J. Multi-Modal 3D Object Detection in Autonomous Driving: A Survey and Taxonomy. *IEEE Transactions on Intelligent Vehicles* **2023**, *8*, 3781–3798.
6. Vora, S.; Lang, A.H.; Helber, B.; Beijbom, O. PointPainting: Sequential Fusion for 3D Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4604–4612.
7. Qi, C.R.; Liu, W.; Wu, C.; Su, H.; Guibas, L.J. Frustum PointNets for 3D Object Detection from RGB-D Data. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 918–927.
8. 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. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2023, pp. 2774–2781.
9. Liang, T.; Xie, H.; Yu, K.; Xia, Z.; Lin, Z.; Wang, Y.; Tang, T.; Wang, B.; Tang, Z. BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework. *Advances in Neural Information Processing Systems* **2022**, *35*, 10421–10434.
10. 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. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1090–1099.

11. 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. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11621–11631.
12. Geiger, A.; Lenz, P.; Urtasun, R. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 3354–3361.
13. Liu, Y.; Jia, T.; Fan, J.; Liu, S.; Zhang, X.; Sun, J. When Autonomous Vehicles Meet Drones: Challenges and Opportunities for 3D Perception. *arXiv preprint arXiv:2202.07588* **2022**.
14. Shi, S.; Jiang, C.; Guo, D.; Chen, Z. Drone-Vehicle Cooperative Perception for Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems* **2024**, *25*, 4784–4800.
15. Arnold, E.; Dianati, M.; de Temple, R.; Fallah, S. Cooperative Perception for 3D Object Detection in Driving Scenarios Using Infrastructure Sensors. IEEE Transactions on Intelligent Transportation Systems, 2022, Vol. 23, pp. 1852–1864.
16. Yu, H.; Luo, Y.; Shu, M.; Huo, Y.; Yang, Z.; Shi, Y.; Guo, Z.; Li, H.; Hu, X.; Yuan, J.; Nie, Z. DAIR-V2X: A Large-Scale Dataset for Vehicle-Infrastructure Cooperative 3D Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21361–21370.
17. Xu, R.; Xiang, H.; Xia, X.; Han, X.; Li, J.; Ma, J. OPV2V: An Open Benchmark Dataset and Fusion Pipeline for Perception with Vehicle-to-Vehicle Communication. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2022, pp. 2583–2589.
18. Xu, R.; Xiang, H.; Tu, Z.; Xia, X.; Yang, M.H.; Ma, J. V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer. Proceedings of the European Conference on Computer Vision (ECCV), 2022, pp. 107–124.
19. Xu, R.; Tu, Z.; Xiang, H.; Shao, W.; Zhou, B.; Ma, J. CoBEVT: Cooperative Bird's Eye View Semantic Segmentation with Sparse Transformers. *Proceedings of the Conference on Robot Learning (CoRL)* **2023**.
20. Choi, E.H. Crash Factors in Intersection-Related Crashes: An On-Scene Perspective. Technical Report DOT HS 811 366, National Highway Traffic Safety Administration, 2010.
21. Yurtsever, E.; Lambert, J.; Carballo, A.; Takeda, K. A Survey of Autonomous Driving: Common Practices and Emerging Technologies. *IEEE Access* **2020**, *8*, 58443–58469.
22. Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; Koltun, V. CARLA: An Open Urban Driving Simulator. Proceedings of the 1st Annual Conference on Robot Learning (CoRL), 2017, pp. 1–16.
23. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. PointPillars: Fast Encoders for Object Detection from Point Clouds. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12697–12705.
24. Yin, T.; Zhou, X.; Krähenbühl, P. Center-Based 3D Object Detection and Tracking. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11784–11793.
25. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 652–660.
26. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. Advances in Neural Information Processing Systems, 2017, pp. 5099–5108.
27. Shi, S.; Wang, X.; Li, H. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 770–779.
28. Zhou, Y.; Tuzel, O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4490–4499.
29. Yan, Y.; Mao, Y.; Li, B. SECOND: Sparsely Embedded Convolutional Detection. *Sensors* **2018**, *18*, 3337.
30. Wang, Y.; Fathi, A.; Kunze, A.; Ross, D.A.; Pantofaru, C.; Funkhouser, T.; Solomon, J. Pillar-based Object Detection for Autonomous Driving. Proceedings of the European Conference on Computer Vision (ECCV), 2020, pp. 18–34.
31. Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; Li, H. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10529–10538.

32. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-View 3D Object Detection Network for Autonomous Driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1907–1915.
33. Philion, J.; Fidler, S. Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. Proceedings of the European Conference on Computer Vision (ECCV), 2020, pp. 194–210.
34. Li, Y.; Yu, A.W.; Meng, T.; Caine, B.; Ngiam, J.; Peng, D.; Shen, J.; Lu, Y.; Zhou, D.; Le, Q.V.; Yuille, A.; Tan, M. DeepFusion: LiDAR-Camera Deep Fusion for Multi-Modal 3D Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17182–17191.
35. Pang, S.; Morris, D.; Radha, H. CLOCs: Camera-LiDAR Object Candidates Fusion for 3D Object Detection. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020, pp. 10386–10393.
36. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S.L. Joint 3D Proposal Generation and Object Detection from View Aggregation. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018, pp. 1–8.
37. Nobis, F.; Geisslinger, M.; Weber, M.; Betz, J.; Lienkamp, M. A Deep Learning-Based Radar and Camera Sensor Fusion Architecture for Object Detection. *Sensors* **2021**, *21*, 2321.
38. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2117–2125.
39. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics, 2023. Accessed: 2026-01-15.
40. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.
41. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. Proceedings of the European Conference on Computer Vision (ECCV), 2014, pp. 740–755.
42. Kuhn, H.W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* **1955**, *2*, 83–97.
43. OpenPCDet Development Team. OpenPCDet: An Open-source Toolbox for 3D Object Detection from Point Clouds. https://github.com/open-mmlab/OpenPCDet, 2020. Accessed: 2026-01-15.
44. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* **2010**, *88*, 303–338.
45. Conover, W. *Practical Nonparametric Statistics*, 3rd ed.; John Wiley & Sons: New York, 1999.
46. Yue, X.; Zhang, Y.; Zhao, S.; Sangiovanni-Vincentelli, A.; Keutzer, K.; Gong, B. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. *arXiv preprint arXiv:1903.11499* **2019**.
47. Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; others. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2446–2454.

© 2026 by the authors. Submitted to *Sustainability* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).