smart cities

MDPI

Article

Feature-Level Vehicle-Infrastructure Cooperative Perception with Adaptive Fusion for 3D Object Detection

Shuangzhi Yu, Jiankun Peng *, Shaojie Wang, Di Wu and Chunye Ma

School of Transportation, Southeast University, Nanjing 211119, China; yuyuyu@seu.edu.cn (S.Y.); wsj@seu.edu.cn (S.W.); wudidi@seu.edu.cn (D.W.); cma@seu.edu.cn (C.M.)

* Correspondence: jkpeng@seu.edu.cn

Highlights

What are the main findings?

- The proposed feature-level VICP framework consistently outperforms state-of-the-art baselines on the DAIR-V2X-C dataset, achieving higher AP_{3D} and AP_{BEV}.
- Experiments show that RFR delivers the largest gain, UWF improves robustness via adaptive uncertainty weighting, and CDCA enhances feature calibration.

What is the implication of the main finding?

- Cooperative perception effectively overcomes occlusion and blind-spot limitations of vehicle-centric systems.
- The proposed model provides a reference for scalable and generalizable deployment of cooperative perception within smart city infrastructure.

Abstract

As vehicle-centric perception struggles with occlusion and dense traffic, vehicle-infrastructure cooperative perception (VICP) offers a viable route to extend sensing coverage and robustness. This study proposes a feature-level VICP framework that fuses vehicle- and roadside-derived visual features via V2X communication. The model integrates four components: regional feature reconstruction (RFR) for transferring region-specific roadside cues, context-driven channel attention (CDCA) for channel recalibration, uncertainty-weighted fusion (UWF) for confidence-guided weighting, and point sampling voxel fusion (PSVF) for efficient alignment. Evaluated on the DAIR-V2X-C benchmark, our method consistently outperforms state-of-the-art feature-level fusion baselines, achieving improved AP_{3D} and AP_{BEV} (reported settings: 16.31% and 21.49%, respectively). Ablations show RFR provides the largest single-module gain +3.27% AP_{3D} and +3.85% AP_{BEV}, UWF yields substantial robustness gains, and CDCA offers modest calibration benefits. The framework enhances occlusion handling and cross-view detection while reducing dependence on explicit camera calibration, supporting more generalizable cooperative perception.

Keywords: vehicle-infrastructure cooperative perception; 3D object detection; smart city infrastructure; fusion-based perception

check for updates

Academic Editor: Pierluigi Siano

Received: 15 September 2025
Revised: 11 October 2025
Accepted: 12 October 2025
Published: 14 October 2025

Citation: Yu, S.; Peng, J.; Wang, S.; Wu, D.; Ma, C. Feature-Level Vehicle-Infrastructure Cooperative Perception with Adaptive Fusion for 3D Object Detection. *Smart Cities* **2025**, *8*, 171. https://doi.org/10.3390/smartcities8050171

1. Introduction

As autonomous driving advances toward higher levels of automation, monocular, vehicle-centric perception encounters inherent limitations in scenarios characterized by occlusions, dense traffic participants, and complex environments. Such limitations often result in incomplete environmental perception, which is crucial to address because undetected road users or obstacles can compromise the safety and reliability of autonomous driving systems. Vehicle-infrastructure cooperative perception (VICP) has thus emerged as a promising paradigm to overcome these challenges. The development of V2I communication provides the prerequisite infrastructure for high-level vehicle-infrastructure information fusion, enabling the timely and reliable exchange of data between vehicles and roadside units [1]. By integrating the macroscopic, global information acquired by roadside infrastructure sensors (e.g., fixed cameras and LiDAR) with the local, dynamic data provided by onboard sensors, VICP mitigates the blind spots inherent to single-vehicle systems [2]. This cooperative framework not only enhances the completeness of environmental perception but also fortifies the robustness and integrity of the overall sensing architecture, establishing VICP as a critical direction for transcending the constraints of standalone perception.

Fusion methodologies in vehicle-infrastructure collaborative perception are broadly classified into three categories: early, intermediate, and late fusion [3]. Early fusion operates at the data level by directly combining raw sensor inputs, such as images and point clouds, preserving maximal information content but imposing stringent demands on bandwidth and sensor synchronization [4]. Late fusion, in contrast, performs independent detection at each sensing endpoint and subsequently merges object-level outputs, which minimizes computational overhead but limits the exploitation of complementary sensor details [5,6]. Intermediate fusion occupies the middle ground: local feature extraction is performed onboard, followed by cross-domain feature alignment and fusion. By balancing representational richness with communication efficiency [7,8], intermediate fusion has become the prevailing focus of contemporary research.

However, implementing effective feature-level fusion in VICP faces several critical challenges, as illustrated in Figure 1. Precise alignment between heterogeneous sensors is often impeded by spatio-temporal asynchrony and viewpoint mismatch [9]. Meanwhile, the inherently multi-scale and complementary features from different sources can introduce semantic noise if fused improperly [10]. In addition, many existing approaches rely on external priors (e.g., geometric calibration or known camera poses), which compromises robustness and generalization in real-world deployments where strict sensor synchronization may not hold. Moreover, most static fusion schemes neglect source-dependent uncertainty, hindering truly reliable and adaptive information integration. Therefore, it is crucial to address these limitations to achieve a reliable and generalizable cooperative perception system.

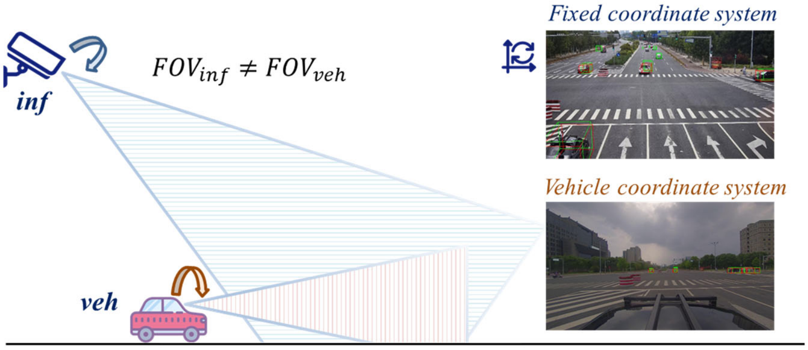

Figure 1. A visual example of VICP.

Recent advances in visual fusion and attention modeling have provided effective paradigms for enhancing perception reliability in complex multi-source environments. Specifically, some research has demonstrated that incorporating spatiotemporal interactions and adaptive attention can substantially enhance perception robustness under occlusion and dynamic illumination [11]. Similarly, multi-scale transformer-based fusion models [12,13] highlight the importance of scale-aware representation and hierarchical feature refinement, while context-aware recalibration strategies [14] have been shown to effectively suppress semantic noise in cross-domain or occluded scenarios. Collectively, these studies reveal three key principles: (1) adaptive feature alignment across heterogeneous domains, (2) context-driven attention to balance multi-scale semantics, and (3) confidence-aware fusion to mitigate uncertainty in dynamic conditions.

In contrast to prior cooperative perception frameworks, such as V2X-ViT [8], CoBEVT [15], and EMIFF [16], which primarily rely on multi-view attention or BEV-based feature alignment, our proposed framework removes dependence on explicit geometric calibration and fixed fusion heuristics. While V2X-ViT and CoBEVT enhance inter-view interaction through cross-attention and transformer structures, they still assume synchronized, high-quality inputs and lack adaptive mechanisms to handle feature-level uncertainty or degraded sensor observations. EMIFF improves efficiency through intermediate fusion but uses static weighting strategies that limit adaptability in dynamic, heterogeneous traffic conditions. Thus, existing frameworks have yet to fully resolve the issues of misalignment and uncertainty in cooperative perception, underscoring the necessity for a more adaptive and robust solution.

To address these limitations, this work integrates deformable attention, contextual channel adaptation, and uncertainty-driven fusion into a unified cooperative perception framework, thereby enhancing the robustness, adaptability, and interpretability of vehicle-infrastructure feature fusion. The main contributions can be summarized as follows:

- A feature-level vehicle-infrastructure perception framework is proposed and experimentally validated, demonstrating superior occlusion handling and adaptability.
- This study proposes a confidence-map-based regional feature reconstruction (RFR) that decouples roadside features and uses deformable attention to reconstruct and augment onboard features, thereby improving onboard representation and cross-view detection.
- A context-driven channel attention (CDCA) module is proposed to exploit global image-level context for adaptive, channel-wise recalibration of multi-scale features on both vehicle and roadside sensors, thereby eliminating reliance on external calibration parameters.
- An uncertainty-weighted fusion (UWF) mechanism is designed to estimate voxel-level uncertainty across heterogeneous feature sources and to allocate fusion weights dynamically based on confidence, significantly enhancing robustness under noise, occlusion, and projection errors.

The remainder of this paper is organized as follows. Section 2 reviews recent work on vision-based 3D object detection, vehicle-infrastructure cooperative perception, channel recalibration, and uncertainty-driven fusion. Section 3 presents the proposed methodology, including feature representation and regional feature reconstruction, the context-driven channel attention module, and the uncertainty-weighted fusion mechanism. Section 4 describes the experimental setup and reports comparative results on public datasets. Finally, Section 5 concludes the study and discusses directions for future research.

2. Related Work

2.1. Vision-Based 3D Object Detection

Convolutional neural networks (CNNs) have significantly advanced object detection by learning hierarchical feature representations that generalize across complex scenes. Traditional 2D detectors, such as R-CNN [17], Faster R-CNN [18], SSD [19], and YOLO [20], achieve high accuracy for planar localization but do not recover depth. In applications such as autonomous driving, accurate 3D localization in a real-world coordinate frame is essential for safe navigation. Monocular 3D detection addresses this need despite its inherent ill-posedness due to missing stereo information.

Monocular methods follow two main strategies. Anchor-based approaches regress object centers and dimensions against predefined 3D priors. For example, M3D RPN incorporates both 2D and 3D geometric constraints in a region proposal network to predict volumetric boxes [21]. Keypoint-based techniques estimate 2D projections of critical 3D points and predict depth offsets. SMOKE [22] predicts the projected center and depth offset of each 3D box, while CenterLoc3D [23] regresses object centroids and eight corners with multi-scale feature fusion to improve precision.

To mitigate depth ambiguity, depth augmented pipelines supply auxiliary spatial cues. D4LCN uses predicted depth maps and dilated convolutions to approximate 3D structure from 2D images [24]. Pseudo LiDAR converts depth estimates into pseudo point clouds so that LiDAR detectors can operate on monocular inputs, and later extensions enrich these clouds with RGB color data or semantic labels [25,26]. Prior-driven methods refine detection further by introducing shape and ground plane priors. Deep MANTA matches 2D detections with 3D CAD templates to reconstruct object geometry [27]. Mono3D++ jointly optimizes unsupervised monocular depth, ground plane constraints, and vehicle shape priors to enhance 3D box accuracy [28].

Bird's eye view (BEV) representations have emerged as an effective solution for monocular 3D detection. Transformer-based approaches, such as DETR3D, index 2D features with sparse 3D queries and apply camera transformations for 3D prediction [29]. Depth-based methods lift image features into BEV by predicting per-pixel depth values [30]. BEVDepth projects features into voxels and aggregates them into the BEV plane while learning camera parameters as attention weights [31]. Building on these foundations, the recently studied GraphBEV [32] introduces graph-structured reasoning to capture spatial relationships among projected BEV tokens across views. BEVFusion4D have further advanced multi-view feature alignment and temporal modeling [33], and BEVFusion4D extends BEV representation to the temporal domain for 4D scene understanding, these methods have further advanced multi-view feature alignment and temporal modeling. Collectively, these methods demonstrate that BEV transformation provides a robust intermediate space for multi-view stitching and long-range perception, which conceptually supports the voxel-based alignment strategy adopted in this study. Calibration-free methods, such as CBR, separate perspective features into front view and BEV representations with multi-layer perceptrons and enhance BEV through cross-view matching [34]. Despite these advances, reliance on depth estimation or calibration priors still limits robustness in heterogeneous real-world deployments.

2.2. Vehicle-Infrastructure Cooperative Perception

Advances in V2X communications enable vehicles to exchange raw sensor data with each other or with roadside units (RSUs), improving coverage and redundancy. Early V2V systems demonstrated that sharing camera or LiDAR streams extends sensing range but remains vulnerable to communication delays and occlusions [35]. Roadside infrastructure offers a complementary global context with stable power, wide fields of view, and robust environmental tolerance, significantly enhancing detection reliability.

Cooperative perception methods may be categorized according to the fusion stage. Data level fusion aligns and concatenates raw sensor streams from vehicles and infrastructure into a common reference frame, maximizing information retention but demanding high bandwidth and strict synchronization [36]. Object-level fusion merges only final detection outputs, such as bounding boxes and confidence scores, minimizing communication overhead but sacrificing complementary low-level features and suffering when individual detections are unreliable [37]. Feature-level fusion extracts intermediate representations locally and then aligns and merges them within a unified coordinate system, reducing transmission while preserving complementary multi-source information, and has become the prevailing paradigm [3].

Early feature-level fusion relied on simple concatenation or linear weighting. As feature diversity increased, these static rules proved inadequate. More recent methods exploit graph convolution and attention. EMIFF introduces camera-aware channel masking to weight multi-scale features [16]. V2X-ViT alternates heterogeneous multi-agent self-attention with multi-scale windowed attention to capture inter-agent and intra-agent interactions [8]. ViT-FuseNet applies cross-attention across point cloud and image modalities for global spatial coupling [38]. Region-based strategies address overlapping and non-overlapping fields of view, and auxiliary losses, such as mutual information minimization and information gain, have been introduced to supervise fusion [39–41]. These advances improve performance, but dependence on precise calibration or fixed fusion rules limits resilience under heterogeneous deployment conditions.

Prior works such as EMIFF and V2X-ViT primarily rely on either enhanced attention patterns to aggregate features across agents or static calibration priors to drive alignment. These approaches improve global information flow but can fail when cross-view semantics conflict, such as roadside-exclusive structures imprinted into fused features, or when calibration is unreliable.

2.3. Channel Recalibration and Uncertainty Driven Fusion

Channel attention enhances feature discrimination by adaptively weighting channel responses. SE models global channel importance using squeeze and excitation operations [42]; CBAM adds spatial attention to refine channel modulation [43]. Both are designed for single-view inputs and do not capture semantic dependencies across multiple viewpoints. Recent cross-view channel attention methods, such as ICAFusion [44], align inter-view features but still depend on rigid calibration priors, limiting generalization.

Uncertainty estimation is crucial when fusing multiple source features of varying reliability. Bayesian deep learning quantifies aleatoric and epistemic uncertainties to inform model confidence [45]. Voxel-level uncertainty has been used to adjust fusion weights dynamically in CenterFusion [46] and BEVFusion [47]. Bayesian deep learning captures both epistemic uncertainty, from model parameters, and aleatoric uncertainty, from data noise, yielding calibrated confidence estimates. Most fusion approaches, however, use these confidences only at the final decision stage and ignore them during feature fusion, preventing uncertainty from guiding intermediate representation learning [48].

Current channel calibration and uncertainty-aware methods remain inadequate for cooperative perception. Channel attention modules often ignore cross-view context and depend on fixed geometric priors, while uncertainty estimates are applied only after fusion, offering no guidance during feature integration. Effective fusion in heterogeneous, partially calibrated, and dynamic vehicle-infrastructure systems, therefore, requires embedding both semantic context and uncertainty guidance directly into feature-level processing.

Aligned with recent advances in visual fusion and attention modeling, which show that multi-scale transformer fusion emphasizes scale-aware representation and hierarchical refinement, context-aware recalibration effectively suppresses semantic noise in cross-domain settings. We aim to propose a calibration-agnostic, context-driven channel reweighting mechanism for cooperative perception by aggregating global descriptors from both infrastructure and vehicle streams and learning shared attention coefficients.

Furthermore, unlike geometry-guided fusion methods that rely on camera intrinsic or extrinsic calibration to establish spatial correspondence, our method is fully data-driven and calibration-free, making it inherently robust to spatial misalignment, synchronization error, and limited calibration diversity.

3. Methods

This work is inspired by the pipelines of EMIFF [12] and other state-of-the-art VICP studies. It aims to further improve detection accuracy while relaxing the reliance on burdensome roadside camera calibration and enhancing scene generalization. Leveraging V2X communication to fuse vehicle- and infrastructure-side features, the overall framework comprises five main modules: regional feature reconstruction (RFR), which spatially decouples roadside information and directionally augments onboard representations; context-driven channel attention (CDCA), which exploits global image-level context for adaptive channel-wise recalibration; uncertainty-weighted fusion (UWF), which estimates voxel-level uncertainty across heterogeneous feature sources and dynamically allocates fusion weights by confidence; and point sampling voxel fusion (PSVF), which efficiently aligns and merges sampled point features with voxel representations. The full architecture and module details are illustrated in Figure 2.

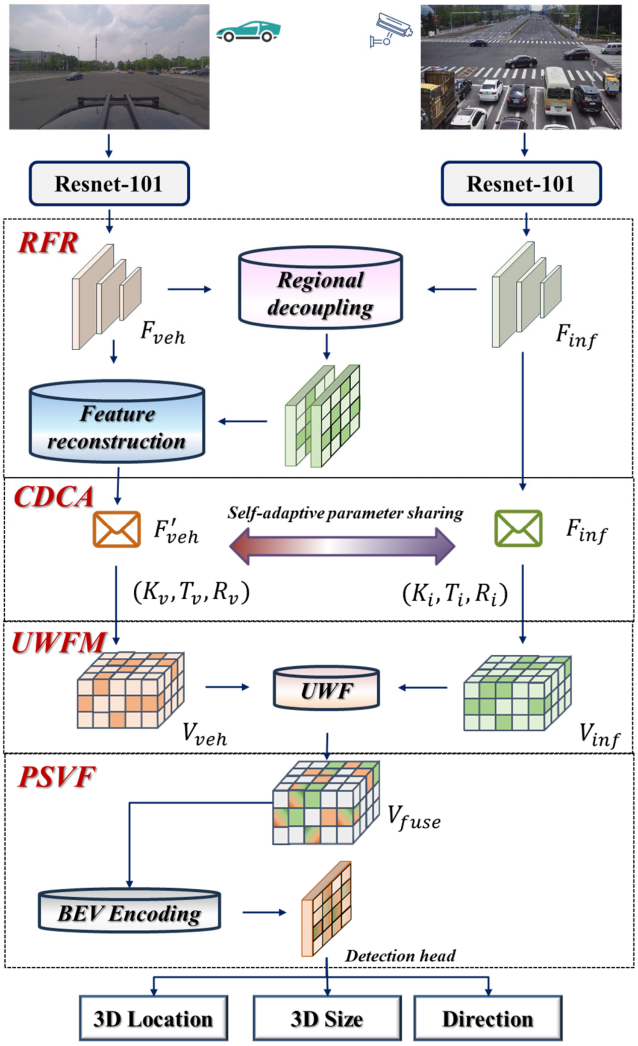

Figure 2. Framework of the proposed model.

To better illustrate the synergy of the proposed framework, we emphasize that the four modules are designed to operate in a complementary manner rather than as isolated components. The backbone and FPN establish a unified multi-scale representation that ensures consistent feature hierarchies across infrastructure and vehicle views. On this foundation, the RFR module performs spatial alignment and compensates for geometric discrepancies, providing spatially coherent intermediate features.

Building upon these aligned representations, the CDCA module adaptively recalibrates channel responses according to the global semantic context extracted from both streams. This process effectively suppresses redundant or misaligned feature activations and enhances semantic consistency prior to fusion. The refined contextual features are then passed to the UWF module, which estimates voxel-level confidence and assigns adaptive weights to each source. The synergy between CDCA's semantic calibration and UWF's uncertainty weighting allows the framework to jointly reduce fusion noise and mitigate spatial misalignment, yielding more stable and interpretable representations than the standalone attention or fusion modules used in prior studies.

3.1. Feature Representation and Regional Feature Reconstruction

Inspired by recent work such as SETR-Net [49], which highlights the effectiveness of enhancing a single feature representation while leveraging temporal recurrence for robust context modeling, we design our feature extraction pipeline to favor semantically rich, context-aware recalibration over purely multi-scale fusion strategies. During multi-scale feature extraction, vehicle-mounted and roadside images are processed through a shared ResNet-101 [50] backbone augmented with a Feature Pyramid Network (FPN), yielding three levels of pyramid features $\left\{f_{veh,s}, f_{inf,s}\right\}$ $(s = 3, 4, 5)$. Low-level features emphasize texture and edge information, facilitating fine-grained object localization, whereas high-level features capture global semantics, aiding robust detection in complex scenes. This multi-scale architecture balances representational richness with computational efficiency and provides diverse resolution cues for subsequent alignment and fusion. To address spatiotemporal asynchrony between the two streams, deformable convolutions perform pixel-level geometric correction, effectively mitigating projection shifts induced by timing offsets.

To deeply integrate heterogeneous perceptions from vehicle-mounted and roadside sensors while accounting for their spatial distributions, we introduce a region-based feature reconstruction method grounded in probabilistic confidence maps. First, confidence maps are generated for both streams to spatially decouple roadside features. Then, a deformable attention mechanism reconstructs vehicle-side features within these decoupled regions, substantially enhancing feature representation and adaptability under real-world deployment conditions.

As shown in Figure 3, considering that some regions in roadside features do not provide significant gains for improving vehicle-side perception performance, the regional decoupling module generates a corresponding confidence map $P \in \mathbb{R}^{H\times W\times 1}$ for the input vehicle-side or roadside image feature $F \in \mathbb{R}^{H\times W\times C}$ through 1×1 convolution and Sigmoid function mapping. Each value in the confidence map reflects the probability estimate of a target existing at the corresponding position in the scene:

$$P_{veh/inf} = \sigma\left(Conv\left(F_{veh/inf}\right)\right) \tag{1}$$

where F_{veh} and F_{inf} denote the vehicle-side and roadside feature maps, and F_{veh} and F_{inf} correspond to the vehicle-side and roadside probability confidence maps. $Conv(\cdot)$ represents a 1×1 convolutional layer; and $\sigma(\cdot)$ is the Sigmoid activation function.

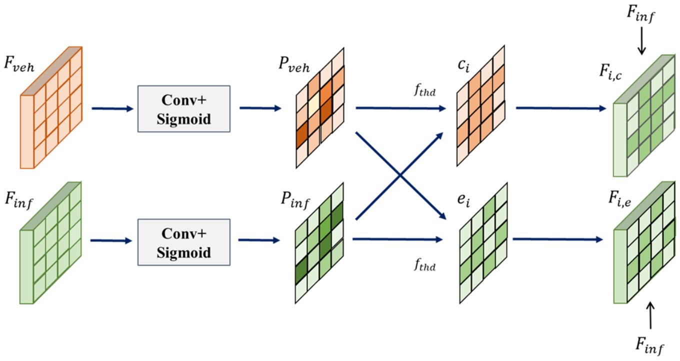

Figure 3. Regional decoupling structure.

In object detection tasks, regions containing potential targets typically carry higher information value and are critical for downstream perception and decision-making. The introduction of probabilistic confidence maps serves as a spatial attention mechanism, enabling the network to adaptively prioritize these high-confidence regions and enhance perception performance. Building upon this, the roadside features are spatially decoupled to generate a shared perception mask c_i and an exclusive perception mask e_i, both are defined with respect to the vehicle-centric coordinate system. These masks are designed to capture the common and distinctive components of vehicle-infrastructure perception, respectively. This process can be formally expressed as follows:

$$c_i = f_{thd}\left(P_{veh} \bigodot P_{inf}\right) \tag{2}$$

$$e_i = f_{thd}(1 - P_{veh}) \bigodot P_{inf} \tag{3}$$

where the thresholding function $f_{thd}(\cdot)$ binarizes the probabilistic confidence maps: it outputs 1 if the input probability exceeds the corresponding threshold τ_c or τ_e. Using this binary mask, the roadside features can be spatially decoupled as follows:

$$F_{i,c} = c_i \bigodot F_{inf} \tag{4}$$

$$F_{i,e} = e_i \bigodot F_{inf} \tag{5}$$

where $F_{i,c}$ and $F_{i,e}$ denote the shared-region and exclusive-region feature representations, respectively. This region-based decoupling strategy enables downstream vehicle-side perception to learn consistent embeddings from homologous features while exploiting complementary features to significantly amplify perceptual gains.

The binary decoupling in Equations (2)–(5) uses thresholds τ_c or τ_e to separate shared and exclusive perception regions. Conceptually, a lower threshold increases included region coverage, while a higher threshold enforces strict sparsity, which may risk omission of weak but informative cues. We selected the default values $\tau_c = \tau_e = 0.01$ as a conservative operating point that retains weak target-supporting features in low-contrast roadside views and suppresses homogeneous background activations produced by the Sigmoid output of the 1×1 confidence predictor. To validate robustness to threshold choice, we performed a structured sensitivity analysis on the validation split: thresholds were swept over $\{0.001, 0.005, 0.01, 0.02, 0.05\}$. For each threshold pair, we measured AP_{BEV}, AP_{3D} and the proportion of the feature map area selected by the binary mask. The analysis shows that model performance exhibits a broad plateau at approximately $\tau \in [0.005, 0.02]$, indicating insensitivity to small threshold perturbations; the chosen $\tau = 0.01$ represents a stable trade-off between recall of weak cues and background suppression.

To address feature misalignment caused by dynamic scene changes and object motion, a deformable attention-based feature reconstruction module is proposed in Figure 4. The module comprises two parallel branches, one aggregating exclusive-area features with vehicle-side features, and the other aggregating common-area features with vehicle-side features. Although both branches share an identical network topology, they learn independent parameter sets. The key components are referred to as exclusive-area feature aggregation (EFA) [51] and common-area feature aggregation (CFA) [52], respectively.

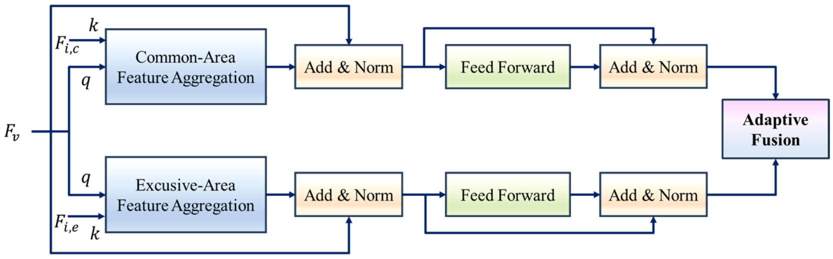

Figure 4. Regional reconstruction structure.

Taking the EFA module as an example, the vehicle-side feature map $F_{veh} \in \mathbb{R}^{H\times W\times C}$ is first projected through an embedding layer to generate the query vectors $Q \in \mathbb{R}^{N_q\times C}$, where N_q denotes the number of query tokens, determined by the positions in the probabilistic confidence map that exceed the confidence threshold.

A positional embedding is then generated for each query via a linear transformation. For every query position, the network predicts a set of sampling offsets through a linear layer to dynamically adjust the sampling locations. Based on the reference point positions and their learned offsets, the corresponding feature values $V \in \mathbb{R}^{N_q\times K\times C}$ are extracted from the roadside feature map $F_{inf} \in \mathbb{R}^{H\times W\times C}$, where K is the predefined number of sampling keys.

A cross-attention fusion is performed between the query vectors Q and the sampled feature values V to generate the enhanced feature representation F_{att}. Specifically, within the cross-attention block, the feature values V are linearly transformed into keys and values, and the similarity between Q and the keys is computed to produce attention weights through an activation function, which are then applied to the values. The output F_{att} of the cross-attention block is linearly transformed and integrated into the original vehicle-side feature F_{veh} to obtain the exclusive-area feature representation $F_{veh,e}$. The generation process of $F_{veh,e}$ follows the workflow illustrated in Figure 5.

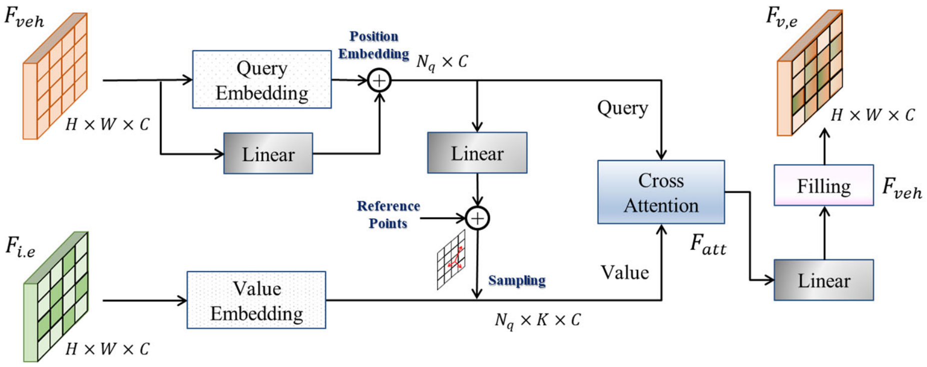

Figure 5. Exclusive-Area Feature Aggregation Structure.

Following EFA and CFA, their outputs $F_{v,e}$ and $F_{v,c}$ are each passed through a 3 × 3 convolution to produce weight maps A_e and A_c. These maps enable adaptive fusion of the two feature streams, effectively leveraging the complementary strengths of distinct perceptual regions to enhance vehicle-side feature representation:

$$F'_{veh} = A_e \bigodot F_{v,e} + A_c \bigodot F_{v,c} \tag{6}$$

3.2. Context-Driven Channel Attention

In previous studies, spatial geometric priors were typically injected into the network through the intrinsic and extrinsic parameters of cameras to guide feature recalibration. However, this method is highly dependent on the accuracy of external calibration, often leading to performance fluctuations in practical deployment due to calibration errors or synchronization issues. To address these shortcomings, this study proposes a CDCA module whose workflow comprises the following:

CDCA takes as input the infrastructure features F_{inf} and the preliminary enhanced vehicle features F'_{veh}. Global average pooling (GAP) and global max pooling (GMP) are applied on each feature tensor $F\left\{F_{inf}, F'_{veh}\right\}$ to obtain channel descriptors:

$$v_{avg} = GAP(F) \tag{7}$$

$$v_{max} = GMP(F) \tag{8}$$

where $\left[v_{avg}; v_{max}\right]$ are concatenated along the channel dimension, and then pass through a two-layer MLP, with a bottleneck ratio r, followed by ReLU and Sigmoid activations to produce the attention weights:

$$z = \sigma\left(W_2 \delta\left(W_1\left[v_{avg}; v_{max}\right]\right)\right) \in \mathbb{R}^C \tag{9}$$

Next, we proceed with the intelligent recalibration of the channels by performing element-wise multiplication on F; multiplying F with z, we obtain the recalibrated features:

$$\widetilde{F} = z \bigodot F \tag{10}$$

The above recalibration is independently applied to and F_{inf} and F'_{veh}, producing $\widetilde{F}_{inf}$ and $\widetilde{F}_{veh}$ for subsequent alignment and UWF processing.

The channel logits are computed from compact global descriptors obtained by spatial pooling: for a feature map $F \in \left\{F_{inf}, F'_{veh}\right\}$, we form descriptors by GAP and GMP and feed their concatenation to a small two-layer bottleneck MLP. Concretely, the following is obtained:

$$z = Sigmoid(MLP([GAP(F), GMP(F)])) \tag{11}$$

The MLP uses a bottleneck ratio ($r = 16$) with shared weights across streams; to mitigate overfitting we apply dropout (rate = 0.1) and weight decay. In practice the MLP can be implemented as 1×1 convolutions on the pooled descriptors for efficiency; CDCA is applied independently at each FPN level and yields per-stream attention vectors.

The CDCA module is completely driven by global contextual data, without the need for external parameters, and has stronger robustness and generalization ability under heterogeneous and asynchronous deployment conditions.

3.3. Uncertainty-Weighted Fusion Mechanism

To enable adaptive fusion of heterogeneous features in the voxel domain, we introduce an UWF mechanism. Multi-scale features recalibrated by CDCA are projected onto a common 3D voxel grid, yielding vehicle-side voxel features V_{veh} and infrastructure-side voxel features V_{inf}. In our implementation, selected pixels are inverse-projected via $X = \pi^{-1}(u, d(u))$ and assigned to voxel indices by $(i, j, k) = (x - x_{min})/v_x, (y - y_{min})/v_y, (z - z_{min})/v_z$ before per-voxel aggregation. Voxel-wise uncertainty σ_v is estimated either via variance across multiple frame projections or through a lightweight MLP regressor. This uncertainty is then mapped to a confidence weight:

$$w_v = exp(-\alpha\sigma_v),\ \alpha > 0 \tag{12}$$

which governs the relative contribution of each stream. The fused voxel representation is obtained by the following:

$$V_{fuse}(v) = w_v V_{veh}(v) + (1 - w_v)V_{inf}(v) \tag{13}$$

By dynamically modulating fusion weights according to feature reliability, UWF effectively mitigates the impact of noise, occlusion, and projection errors, thereby enhancing the robustness and discriminative power of the fused representation. The resulting voxel tensor V_{fuse} is subsequently processed by a 3D convolutional network and a BEV detection head to produce accurate 3D object predictions.

3.4. Point Sampling Voxel Fusion

To ensure reproducibility and computational efficiency in lifting image features into the shared voxel representation, we implement PSVF that concentrates representational capacity on pixels identified as informative by the RFR confidence map, which maps sampled pixels into a 3D voxel grid and emits compact per-voxel descriptors with sampling-confidence summaries consumed by the UWF stage. PSVF samples a bounded set of pixel locations from each selected feature map using a mixture strategy: most samples are drawn by importance sampling from the RFR confidence distribution $C(u)$, which is the probabilistic map produced by Equation (1), while a small fraction is drawn uniformly to preserve coverage of low-confidence areas. Concretely, importance sampling probabilities are formed as $p_{imp}(u) \propto C(u)^{\alpha}$, where the exponent α controls concentration on high-confidence pixels; a uniform fallback fraction β prevents complete neglect of low-confidence regions. Empirically, modest values of α and small β balance robustness and selectivity.

Each sampled pixel u is associated with a depth proxy $d(u)$, which is the predicted depth, stereo estimate, or scene prior, and is inverse-projected to a 3D point in the sensor frame via $X_{sensor} = \pi^{-1}(u, d(u))$. The point is transformed to the common/world frame by the sensor extrinsic $T_{sensor \to world}$, yielding $X_{sensor \to world}$. X_{sensor} Voxel indices are obtained with the convention.

$$i = \left[\frac{x - x_{min}}{v_x}\right], j = \left[\frac{y - y_{min}}{v_y}\right], k = \left[\frac{z - z_{min}}{v_z}\right] \tag{14}$$

where $(x, y, z) = X_{world}$, $(x_{min}, y_{min}, z_{min})$ denote the voxel grid origin and $\left(v_x, v_y, v_z\right)$ the voxel sizes. Points outside the grid bounds are discarded. The voxelization and filling process are illustrated in Figure 6. To bound memory and latency, each voxel collects up to a fixed cap N_v points. Within each voxel, collected point features $\{f_n\}$ are aggregated with a learnable per-point weighting. A shallow MLP produces scalar scores $s_n = MLP(F_n)$, which are normalized by a softmax to obtain attention weights a_n. The voxel descriptor is the weighted sum:

$$F_{voxel} = \sum_{n=1}^{N_v} a_n f_n \tag{15}$$

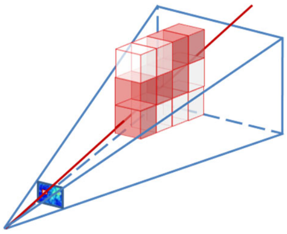

Figure 6. Voxel filling diagram.

PSVF also computes a concise confidence summary for each voxel, for example, $c = [max_n a_n, mean_n a_n]$, which quantifies sampling concentration and provides an evidential cue. The aggregated voxel descriptor and its confidence summary are concatenated and provided as inputs to the UWF uncertainty regressor, thereby enabling UWF to consider both feature content and sampling reliability when producing fusion weights. In practice, this design enables the fusion module to down-weight voxels with diffuse or noisy evidence even when scattered point activations exist. According to the experiment, the parameters are set to the following: per-frame sample budget $N_{max} = 2048$, per-voxel cap $N_v = 8$, confidence exponent $\alpha = 2$, and uniform fallback fraction $\beta = 0.1$. To evaluate sensitivity, we recommend sweeping $N_{max} \in \{1024,\ 2048,\ 4096\}$, $N_v \in \{4,\ 8,\ 16\}$, and $\beta \in \{0,\ 0.1,\ 0.2\}$, and recording AP_{3D} and AP_{BEV} on a per-frame latency to guide deployment tradeoffs.

4. Dataset Description and Experimental Setup

4.1. Dataset Description

DAIR-V2X constitutes China's inaugural public large-scale vehicle-infrastructure collaborative multi-modal dataset, acquired within Beijing's advanced autonomous driving demonstration zone. It encompasses complex urban environments, including intersections and roundabouts, under diverse meteorological conditions. Through rigorous spatiotemporal synchronization with microsecond-level precision, the dataset comprises 72,890 valid frames sampled at 10 Hz from 100 continuous 20 s sequences, partitioned into three subsets: the vehicle-infrastructure collaborative set (DAIR-V2X-C, comprising 40481 image-point cloud pairs), infrastructure-only set (DAIR-V2X-I), and vehicle-only set (DAIR-V2X-V).

Annotations employ LiDAR-coordinated 2D–3D joint labeling across 15 obstacle categories (e.g., pedestrians, vehicles). Each frame integrates 3D bounding boxes (position, dimensions, orientation), 2D image annotations, occlusion levels, truncation states, and cross-frame tracking identifiers. This protocol yields over 1.2 million high-quality 3D annotations spanning 230,000 unique instances, establishing foundational benchmarks for collaborative perception research.

4.2. Experimental Settings

In this study, all experiments were conducted on a single server equipped with 8 NVIDIA GeForce RTX 3090 GPUs and an Intel Xeon Gold 5320 CPU to ensure reproducibility and fair comparison. The software environment comprised Ubuntu 20.04, CUDA 11.6 with CuDNN 8.3, and Python 3.7. The model was implemented in pyTorch 1.9.1, with data preprocessing and visualization facilitated by OpenCV 4.7 and Matplotlib 3.5.2.

The experiments were implemented based on the MMDetection3D framework (version 1.4.0, Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong, China) with its default training configurations. The proposed model was trained for 20 epochs with a batch size of 8, using the AdamW optimizer with an initial learning rate of 1×10^{-3} and a weight decay of 0.0001. The RepeatDataset strategy in MMDetection3D was employed to maintain data balance, where each scene was repeated three times per epoch. The detection head adopted Rotated Non-Maximum Suppression (NMS) to filter redundant bounding boxes and generate the final predictions. It is worth noting that only vehicles were considered as the detection category in this study.

4.3. Evaluation Metrics

For the 3D object detection task, evaluation follows standard object detection protocols, using Precision, Recall, and Average Precision (AP) to quantify detection accuracy. AP measures the area under the Precision, Recall curve and is typically evaluated at specific Intersection over Union (IoU) thresholds, such as AP@0.5 or AP@0.7, to ensure sufficient spatial overlap between predictions and ground truth.

In 3D detection, two complementary AP metrics are commonly reported: AP_{BEV}, which assesses localization accuracy on the BEV plane, and AP_{3D}, which further incorporates the predicted bounding box's height and center position, providing a stricter evaluation of volumetric precision. The mean Average Precision (mAP) represents the overall performance averaged across all object categories, which is defined as the arithmetic mean of AP values over all classes and provides a holistic assessment of a detector's performance. It is calculated as follows:

$$mAP = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} AP_i \tag{16}$$

where N_{cls} represents the number of target categories, and AP_i is the AP value of the target in the ith category.

For transparency and to facilitate reproducibility, we provide here the mathematical definition and computational steps for the 3D IoU used in our experiments. A 3D bounding box B is parameterized as $(x_c,\ y_c,\ z_c,\ l,\ w,\ h,\ \theta)$, where $(x_c,\ y_c,\ z_c)$ is the box center in world coordinates, $l,\ w,\ h$ denote length, width, and height, and θ is the yaw around the vertical axis. The 3D IoU between two boxes, ground truth B^g and prediction B^p, is defined as

$$IoU_{3D}(B^g, B^p) = \frac{Vol(B^g \cap B^p)}{Vol(B^g \cup B^p)} \tag{17}$$

The intersection volume computation is implemented by decomposing each box into a vertical prism: (i) compute the 2D polygon intersection area $A_\cap$ between the yaw-rotated rectangles on the ground plane, and (ii) compute the overlapped height interval as follows:

$$H_\cap = \max\left(0, \min\left(z_{top}^g, z_{top}^p\right) - \max\left(z_{top}^g, z_{top}^p\right)\right) \tag{18}$$

where $z_{top} = z_c + h/2, z_{bot} = z_c - h/2$. Thus,

$$Vol(B^g \cap B^p) = A_\cap \cdot H_\cap \tag{19}$$

Because the 2D polygon intersection $A_\cap$ depends on the yaw θ, small heading errors $\Delta\theta$ can reduce $A_\cap$ nonlinearly for elongated objects. For small angular perturbations, a first-order approximation yields the following:

$$\Delta A_\cap \approx -\kappa(l,\ w, \theta) \cdot |\Delta\theta| \tag{20}$$

where $\kappa(\cdot)$ is a scene and geometry dependent factor larger for high aspect ratio boxes and for near-edge alignments; this explains why IoU drops faster for orientation-biased cases. Consequently, AP_{3D} is more sensitive to heading errors than AP_{BEV}, which ignores height and, to a lesser extent, vertical misalignment.

The BEV projection discards height information by evaluating the IoU of 3D bounding boxes projected onto the ground plane. Mathematically, this renders the BEV IoU invariant to pure vertical displacements, tolerating height errors that do not alter the 2D footprint. This makes AP_{BEV} a valid and efficient metric for ground-level path planning, where lateral and longitudinal extents are primary constraints. In contrast, the full 3D IoU is sensitive to orientation and height, making AP_{3D} indispensable for tasks requiring volumetric precision, such as vertical clearance assessment. We, therefore, report both metrics: AP_{BEV} to quantify planning-centric horizontal localization and AP_{3D} to quantify full spatial fidelity.

5. Results and Analysis

5.1. Analysis of Comparative Experimental Results

To rigorously evaluate the proposed method, we performed comparative experiments on the DAIR-V2X-C dataset against leading vision-based cooperative 3D object detection frameworks. In the first set of experiments, we assessed single-agent perception by adapting our network to vehicle-only and infrastructure-only workflows: we retained the feature-extraction and BEV-encoding modules while removing the region-level reconstruction branch, yielding the configurations denoted Ours_Veh and Ours_Inf, which we compared to the corresponding EMIFF_Veh and EMIFF_Inf baselines.

Next, decision-level fusion was implemented using the official DAIR-V2X fusion protocol: EMIFF_Veh and EMIFF_Inf were merged to form the baseline decision-level fusion (DL) configuration, and Ours_Veh and Ours_Inf were fused in the same manner to produce our decision-level fusion variant. Finally, we evaluated the full feature-level fusion (FL) capability of our model against representative feature-level fusion methods, including EMIFF [16] and QUEST [53]. All comparisons focus exclusively on the car category, with detection accuracy measured by mean Average Precision (mAP) in both 3D and BEV projections at IoU thresholds of 0.3 and 0.5. Communication overhead accompanying each setting is also reported. Quantitative results are summarized in Table 1.

Table 1. Detection results for different types of models.

Type	Model	AP_{3D}(%)		AP_{BEV}(%)	
		$IoU_{0.5}$	$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.3}$
Single_Veh	EMIFF_Veh	14.62	7.92	15.77	9.65
	Ours_Veh	14.78	8.26	15.99	9.61
Single_Inf	EMIFF_Inf	22.27	8.40	23.41	13.34
	Ours_Inf	21.34	8.28	23.12	13.05
DL	EMIFF_Veh and Inf	26.22	11.60	29.25	16.44
	Ours_Veh and Inf	26.65	12.08	29.86	16.73
FL	EMIFF	30.24	15.07	33.73	20.74
	QUEST	33.30	14.10	39.40	20.30
	Ours	31.01	15.76	33.38	21.02

Experimental results demonstrate that, for all evaluated models and IoU thresholds, the mean Average Precision in AP_{3D} is systematically lower than in AP_{BEV}. This disparity arises because 3D detection must recover object height information, substantially increasing model complexity and learning difficulty; in typical traffic scenarios, however, all objects lie on a common ground plane, so height contributes minimally to planning and decision-making.

Moreover, both the baseline EMIFF models and the proposed method exhibit inferior single-agent performance on the vehicle side compared to the infrastructure side. This degradation is primarily due to the vehicle-mounted camera's limited field of view and susceptibility to occlusion, which reduces data quality. Consequently, supplementing vehicle perception with infrastructure imagery yields significant practical benefits, as cooperative perception consistently outperforms vehicle-only detection under any fusion strategy.

Regarding fusion strategies, FL incurs higher communication overhead—owing to the transmission of full feature maps—than DL, which exchanges only detection outputs. Nevertheless, FL achieves superior accuracy by more effectively leveraging multi-source contextual semantics. Under FL, the proposed method surpasses both EMIFF and QUEST across all IoU thresholds. At $IoU = 0.5$, our model attains an AP_{3D} of 15.76%, compared to 15.07% for EMIFF and 14.10% for QUEST; similarly, AP_{BEV} reaches 21.02%, exceeding the baselines. These results validate that the introduced BEV-encoding strategy enhances height information utilization, thereby improving 3D detection performance.

To investigate the influence of backbone depth on object detection performance, ResNet-50 [54], ResNet-101 [50], and VGG-16 [55] were adopted as the feature extraction networks for both the EMIFF baseline and the proposed method. The mAP is evaluated at IoU thresholds of 0.3 and 0.5. The results are summarized in Table 2.

Table 2. Detection performance across different feature extraction backbones.

Backbone	AP_{3D}(%)		AP_{BEV}(%)		Param	GFLOPS
	$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.3}$	$IoU_{0.5}$		
VGG16_EMIFF	28.12	14.27	30.45	17.05	138.36M	152.30
VGG16_Ours	28.85	14.88	31.12	17.64	143.73M	135.98
Res50_EMIFF	30.24	15.07	33.38	20.74	49.33M	123.76
Res101_EMIFF	31.13	15.84	34.68	21.30	87.31M	201.46
Res50_Ours	31.01	15.76	33.73	21.02	54.72M	107.44
Res101_Ours	31.58	16.31	34.16	21.49	92.66M	184.93

In vehicle-infrastructure cooperative object detection tasks, the choice of backbone network exerts a profound influence on overall performance. Compared with the shallow VGG16, ResNet architectures leverage residual connections to deliver substantially richer feature representations. In particular, the deeper ResNet101 extracts more discriminative high-level semantic features, boosting the EMIFF baseline's average precision by 0.89% and its average recall by 0.77%, while our proposed model achieves corresponding gains of 0.57% and 0.55%. However, these accuracy improvements incur a marked computational overhead: our model's parameter count rises from 54.72 million to 92.66 million and its GFLOPs increase from 107.44 billion to 184.93 billion, whereas the EMIFF baseline grows from 49.33 million to 87.31 million parameters and from 123.76 to 201.46 GFLOPs. Such an almost two-fold expansion in model size and inference workload challenges the strict real-time and resource-constrained requirements of intelligent transportation systems. Accordingly, to strike the optimal balance between detection accuracy and computational efficiency, this study adopts ResNet-50 as the feature extraction backbone.

To further verify computational efficiency, local inference tests were conducted without deployment acceleration. All evaluations were performed on the workstation described in Section 4.2. Two input resolutions (1280 × 720 and 1920 × 1080) were tested, and the corresponding inference results are shown in Table 3. We report the FPS of the Backbone + FPN stage, the overall Total FPS, and the Total Latency (ms). Each measurement was averaged over 300 frames, after 100 warm-up frames, using CUDA events. The results show that Res50_Ours achieves near real-time performance on a single GTX 3090, while Res101_Ours maintains reasonable latency with improved accuracy.

Table 3. Inference results under different input resolutions.

Model	Input Resolution	Backbone + FPN FPS ↑	Total FPS ↑	Total Latency (ms) ↓
Res50_Ours	1280 × 720	42.6	34.1	29.3
	1920 × 1080	33.4	26.8	37.3
Res101_Ours	1280 × 720	31.7	25.5	39.2
	1920 × 1080	25.6	21.3	46.9

5.2. Ablation Study

To verify the effectiveness of each module in the proposed model, we designed ablation experiments to evaluate the impact of the RFR module, the camera parameter embedding module, its alternative based on global image-level context for adaptive recalibration (i.e., CDCA), and the UWF module on both 3D object detection and BEV object detection performance. The results are reported in Table 4.

Table 4. Ablation results on model architecture, where symbol √ indicates that the corresponding module is enabled.

RFR	Camera Parameter Embedding	CDCA	UWF	AP_{3D}(%)		AP_{BEV}(%)	
				$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.3}$	$IoU_{0.5}$
√	√			27.53	12.89	30.13	17.46
		√	√	27.86	13.04	30.39	17.64
√	√		√	31.03	15.18	33.85	21.07
√		√		29.69	14.84	32.72	19.30
√		√	√	31.58	16.31	34.16	21.49

At an IoU threshold of 0.5, replacing the camera parameter embedding module with the CDCA module yields relative improvements of 0.15% in AP_{3D} and 0.18% in AP_{BEV}. Although the gains are modest, they are notable given the simplicity of the module and its ability to eliminate the need for rule-based camera parameter settings across different scenes, thereby facilitating model transferability. The UWF module, by estimating voxel-level uncertainty across heterogeneous feature sources and dynamically allocating fusion weights, provides a more substantial benefit, improving AP_{3D} and AP_{BEV} by 1.47% and 2.19%, respectively. Most importantly, the RFR module achieves the largest improvements, with AP_{3D} and AP_{BEV} increasing by 3.27% and 3.85%, respectively, demonstrating that region-wise reconstruction of onboard features using roadside information significantly enhances model performance.

To further assess the role of the PSVF module, we compared it against two alternative voxelization schemes: (1) uniform voxel sampling, which randomly samples pixels with equal probability, and (2) full-grid voxelization, which converts all projected points into voxels without sampling.

As summarized in Table 5, PSVF achieves the highest efficiency and accuracy trade-off. Specifically, compared with full-grid voxelization, PSVF reduces total latency by 26%, while also slightly outperforming uniform sampling in both AP_{3D} and AP_{BEV} metrics. The results confirm that PSVF provides a compact yet informative voxel representation, effectively concentrating computational resources on salient high-confidence regions identified by the RFR confidence maps. This design preserves the structural integrity of the fusion pipeline while delivering significant gains in efficiency and detection performance.

Table 5. Comparison of different voxel construction strategies in the fusion stage.

Voxelization Strategy	Param	GFLOPS	Total FPS ↑	Total Latency (ms) ↓	$AP_{3D}(\%)$ $@IoU_{0.5}$	$AP_{BEV}(\%)$ $@IoU_{0.5}$
Full-grid voxelization	93.4 M	189.7	22.4	47.5	15.89	21.27
Uniform voxel sampling	92.5 M	186.2	25.1	42.8	16.02	21.33
PSVF (ours)	92.7 M	184.9	31.7	39.2	16.31	21.49

5.3. Visualization

To demonstrate the detection performance of the proposed model, we present its detection results across different scenarios from three perspectives: the onboard camera view, the roadside camera view, and the ego-centric BEV. We also provide a performance comparison under identical scene settings between a single-vehicle detector, baseline methods, and our proposed model; the results are shown in Figures 7–9. In the visualizations, green bounding boxes denote ground truth, red bounding boxes denote predicted detections, and the blue point indicates the ego vehicle's position.

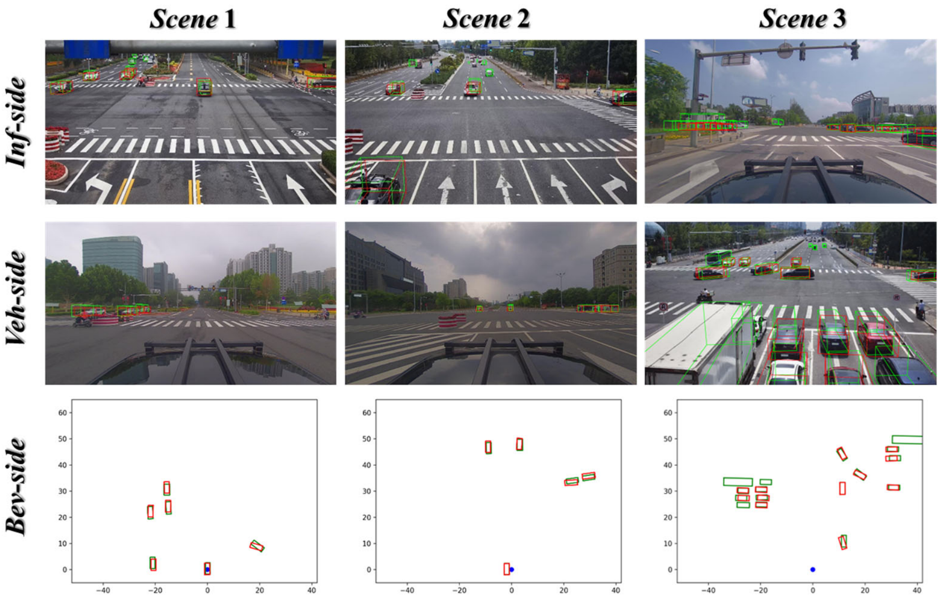

Figure 7. Visualization of results across different scenarios, where green bounding boxes represent ground truth and red bounding boxes denote the predicted results.

In sparse traffic-flow scenarios, the proposed model achieves accurate detection for the vast majority of vehicles. In dense traffic-flow scenarios, our model not only effectively identifies mutually occluded vehicles but also detects vehicles that lie outside the onboard camera's field of view, as indicated by the circled region in Scene 3 of Figure 7. However, there remain some missed detections in areas roughly 30m ahead of the ego vehicle and about 25 m laterally beyond the ego sensing range, indicating that long-range perception remains an area for improvement.

Across identical traffic scenes, observable differences exist between the baseline and our proposed model. As shown in Figure 8, EMIFF effectively fuses roadside perception data and partially mitigates the beyond-ego-field-of-view detection problem, substantially improving the detection of occluded vehicles and enabling more comprehensive perception in complex environments such as intersections. Nevertheless, EMIFF still exhibits small but systematic biases in orientation estimation and localization when processing occluded targets, which is likely attributable to insufficient modeling of voxel-level uncertainty across heterogeneous feature sources. By fully leveraging roadside augmentation and complementary information, our proposed model achieves significant improvements in both detection accuracy and spatial coverage.

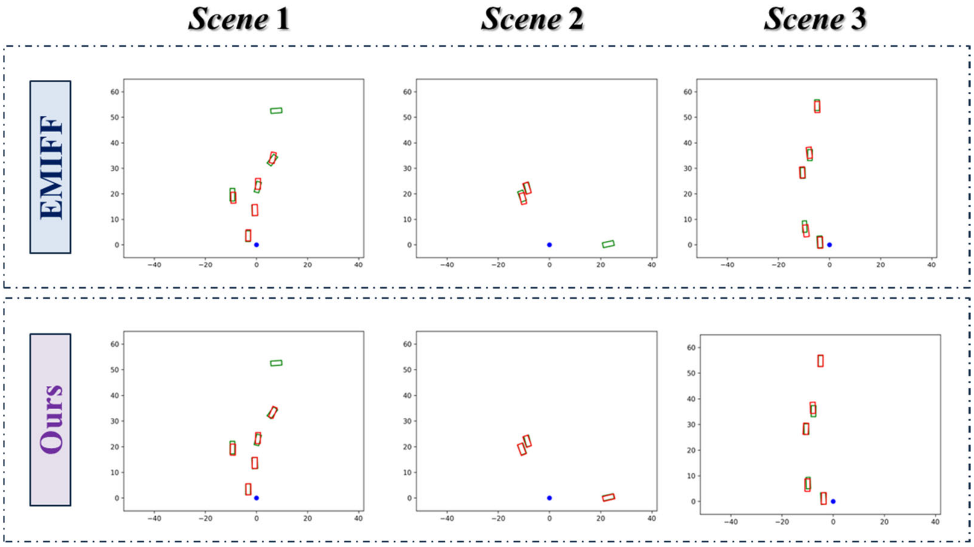

Figure 8. Visual comparison of proposed model and baselines, where green bounding boxes represent ground truth and red bounding boxes denote the predicted results.

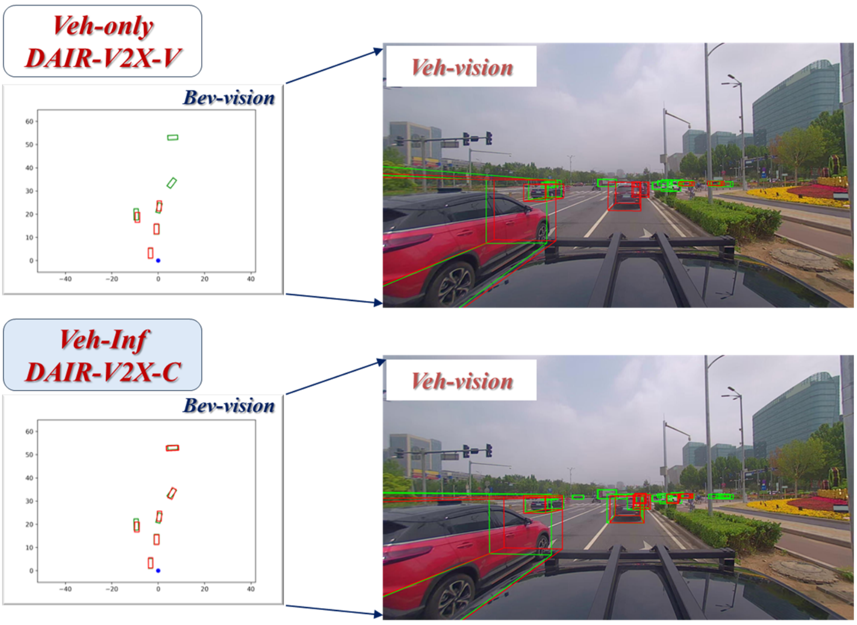

Figure 9. Visual comparison of DAIR-V2X-V and DAIR-V2X-C, where green bounding boxes represent ground truth and red bounding boxes denote the predicted results.

To further validate the generalizability of the proposed framework, we additionally compare the single-vehicle detection model trained on DAIR-V2X-V with the vehicle-infrastructure cooperative model trained on DAIR-V2X-C under identical driving scenes.

As shown in Figure 9, the single-vehicle model exhibits severe performance degradation when target vehicles are partially or fully occluded by leading traffic, resulting in clear missed detections within the blind area of the onboard camera. In contrast, the cooperative perception model effectively compensates for these occlusions by integrating roadside visual cues, enabling accurate long-range detection and complete spatial coverage. This comparison provides intuitive evidence that feature-level vehicle-infrastructure fusion significantly enhances robustness and perception completeness in complex traffic environments.

6. Conclusions

This study proposed a feature-level VICP framework that integrates vehicle- and roadside-derived features through V2X communication. The framework combines RFR, CDCA, UWF, and PSVF to jointly address occlusion, calibration dependency, and feature heterogeneity. Comprehensive experiments on the DAIR-V2X-C benchmark demonstrated consistent improvements over state-of-the-art baselines, while ablation studies confirmed the complementary contributions of each module. These findings establish VICP as an effective means to extend sensing coverage and enhance robustness, offering a practical reference for scalable deployment within smart city infrastructure.

While attention mechanisms and uncertainty weighting have been applied independently in prior VICP works (e.g., EMIFF uses multi-scale fusion and channel masking; V2X-ViT applies multi-agent attention), our contribution is twofold and integrative: (1) RFR performs explicit spatial decoupling between shared and exclusive perception regions, which prevents misleading roadside-exclusive semantics from contaminating vehicle-side embeddings during cross-view reconstruction; and (2) the combination of CDCA and UWF acts at complementary granularities, CDCA provides channel-wise semantic rebalancing conditioned on global context, whereas UWF controls voxel-level fusion strength according to estimated uncertainty. The synergy of these components yields robustness to viewpoint mismatch and noisy inputs that is not achieved by standalone attention modules or by post-hoc weighting alone.

Future research will extend this work in three directions: integrating additional sensing modalities such as radar and event cameras, improving real-time efficiency and communication adaptability for large-scale V2X networks, and exploring cross-domain generalization to ensure reliable operation under diverse environmental and traffic conditions. Despite the promising results, the current framework still relies on high-quality temporal synchronization in the DAIR-V2X-C dataset, which may constrain its scalability to heterogeneous infrastructures and complex traffic scenarios. Moreover, long-range object perception remains challenging due to reduced feature consistency and signal attenuation across sensing modalities. Furthermore, the robustness of the current framework could be affected under conditions of significant motion blur or temporal asynchrony, which were not the focus of this study. Enhancing temporal alignment and mitigating these effects represent key objectives for future work, along with improving generalization and deployment feasibility in real-world cooperative perception systems.

Author Contributions: Conceptualization, S.Y. and S.W.; methodology, S.Y. and S.W.; software, S.Y. and J.P.; validation, S.Y. and J.P.; formal analysis, S.Y.; investigation, S.Y. and; resources J.P. and C.M.; data curation, S.W. and D.W.; writing—original draft preparation, S.Y. and S.W.; writing—review and editing, S.Y., J.P. and S.W.; visualization, S.Y. and D.W.; supervision, C.M.; project administration, J.P.; funding acquisition, S.Y. and J.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key R&D Program of Shandong Province, China, grant number 2023CXGC010111; and the SEU Innovation Capability Enhancement Plan for Doctoral Students of southeast university, grant number CXJH_SEU 24180.

Data Availability Statement: The data used in this study are publicly available from the DAIR-V2X dataset at https://air.tsinghua.edu.cn/DAIR-V2X/index.html (accessed on 5 August 2025).

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

References

1. Clancy, J.; Molloy, D.; Hassett, S.; Leahy, J.; Ward, E.; Denny, P.; Jones, E.; Glavin, M.; Deegan, B. Evaluating the feasibility of intelligent blind road junction V2I deployments. *Smart Cities* **2024**, *7*, 973–990. [CrossRef]
2. Yang, L.; Zhang, X.; Yu, J.; Li, J.; Zhao, T.; Wang, L.; Huang, Y.; Zhang, C.; Wang, H.; Li, Y. MonoGAE: Roadside monocular 3D object detection with ground-aware embeddings. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 17587–17601. [CrossRef]
3. Han, Y.; Zhang, H.; Li, H.; Jin, Y.; Lang, C.; Li, Y. Collaborative perception in autonomous driving: Methods, datasets, and challenges. *IEEE Intell. Transp. Syst. Mag.* **2023**, *15*, 131–151. [CrossRef]
4. Arnold, E.; Dianati, M.; De Temple, R.; Fallah, S. Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 1852–1864. [CrossRef]
5. Wang, T.H.; Manivasagam, S.; Liang, M.; Yang, B.; Zeng, W.; Urtasun, R. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In Proceedings of the Computer Vision—ECCV 2020, Virtually, 23–28 August 2020; pp. 605–621.
6. Hu, Y.; Lu, Y.; Xu, R.; Xie, W.; Chen, S.; Wang, Y. Collaboration helps camera overtake lidar in 3d detection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 9243–9252.
7. Luo, G.; Zhang, H.; Yuan, Q.; Li, J. Complementarity-enhanced and redundancy-minimized collaboration network for multi-agent perception. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 3578–3586.
8. Xu, R.; Xiang, H.; Tu, Z.; Xia, X.; Yang, M.H.; Ma, J. V2x-ViT: Vehicle-to-everything cooperative perception with vision transformer. In Proceedings of the ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; pp. 107–124.
9. Cui, G.; Zhang, W.; Xiao, Y.; Yao, L.; Fang, Z. Cooperative perception technology of autonomous driving in the internet of vehicles environment: A review. *Sensors* **2022**, *22*, 5535. [CrossRef]
10. Liu, J.; Wang, P.; Wu, X. A Vehicle-Infrastructure Cooperative Perception Network Based on Multi-Scale Dynamic Feature Fusion. *Appl. Sci.* **2025**, *15*, 3399. [CrossRef]
11. Zhang, Y.; Zhang, Y.; Xiao, Y.; Wang, T. Spatiotemporal Dual-Branch Feature-Guided Fusion Network for Driver Attention Prediction. *Expert Syst. Appl.* **2025**, *292*, 128564. [CrossRef]
12. Zhang, Y.; Wang, S.; Zhang, Y.; Yu, P. Asymmetric light-aware progressive decoding network for RGB-thermal salient object detection. *J. Electron. Imaging* **2025**, *34*, 013005. [CrossRef]
13. Zhang, Y.; Yu, P.; Xiao, Y.; Wang, S. Pyramid-structured multi-scale transformer for efficient semi-supervised video object segmentation with adaptive fusion. *Pattern Recognit. Lett.* **2025**, *194*, 48–54. [CrossRef]
14. Zhang, Y.; Liu, T.; Zhen, J.; Kang, Y.; Cheng, Y. Adaptive downsampling and scale enhanced detection head for tiny object detection in remote sensing image. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 6003605. [CrossRef]
15. Xu, R.; Tu, Z.; Xiang, H.; Shao, W.; Zhou, B.; Ma, J. CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers. *arXiv* **2022**, arXiv:2207.02202.
16. Wang, Z.; Fan, S.; Huo, X.; Xu, T.; Wang, Y.; Liu, J.; Chen, Y.; Zhang, Y.Q. Emiff: Enhanced multi-scale image feature fusion for vehicle-infrastructure cooperative 3d object detection. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; pp. 16388–16394.
17. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015; p. 28.
19. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
20. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
21. Brazil, G.; Liu, X. M3d-rpn: Monocular 3d region proposal network for object detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9287–9296.
22. Liu, Z.; Wu, Z.; Tóth, R. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 996–997.
23. Tang, X.; Wang, W.; Song, H.; Zhao, C. CenterLoc3D: Monocular 3D vehicle localization network for roadside surveillance cameras. *Complex Intell Syst.* **2023**, *9*, 4349–4368. [CrossRef]
24. Ding, M.; Huo, Y.; Yi, H.; Wang, Z.; Shi, J.; Lu, Z.; Luo, P. Learning depth-guided convolutions for monocular 3d object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1000–1001.

25. Wang, Y.; Chao, W.L.; Garg, D.; Hariharan, B.; Campbell, M.; Weinberger, K.Q. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 8445–8453.
26. Ma, X.; Wang, Z.; Li, H.; Zhang, P.; Ouyang, W.; Fan, X. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27–28 October 2019; pp. 6851–6860.
27. Chabot, F.; Chaouch, M.; Rabarisoa, J.; Teuliere, C.; Chateau, T. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2040–2049.
28. He, T.; Soatto, S. Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 8409–8416. [CrossRef]
29. Wang, Y.; Guizilini, V.C.; Zhang, T.; Wang, Y.; Zhao, H.; Solomon, J. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In Proceedings of the 5th Conference on Robot Learning (CoRL 2021), London, UK, 8–11 November 2022; pp. 180–191.
30. Philion, J.; Fidler, S. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 194–210.
31. Li, Y.; Ge, Z.; Yu, G.; Yang, J.; Wang, Z.; Shi, Y.; Sun, J.; Li, Z. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 1477–1485. [CrossRef]
32. Song, Z.; Yang, L.; Xu, S.; Liu, L.; Xu, D.; Jia, C.; Jia, F.; Wang, L. Graphbev: Towards robust bev feature alignment for multi-modal 3d object detection. In Proceedings of the 2024 European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 347–366.
33. Cai, H.; Zhang, Z.; Zhou, Z.; Li, Z.; Ding, W.; Zhao, J. Bevfusion4d: Learning lidar-camera fusion under bird's-eye-view via cross-modality guidance and temporal aggregation. *arXiv* **2023**, arXiv:2303.17099.
34. Fan, S.; Wang, Z.; Huo, X.; Wang, Y.; Liu, J. Calibration-free bev representation for infrastructure perception. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 1–5 October 2023; pp. 9008–9013.
35. 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. In Proceedings of the 2022 IEEE International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; pp. 2583–2589.
36. Xiang, C.; Feng, C.; Xie, X.; Shi, B.; Lu, H.; Lv, Y.; Yang, M.; Niu, Z. Multi-sensor fusion and cooperative perception for autonomous driving: A review. *IEEE Intell. Transp. Syst. Mag.* **2023**, *15*, 36–58. [CrossRef]
37. Xu, R.; Chen, W.; Xiang, H.; Liu, L.; Ma, J. Model-agnostic multi-agent perception framework. *arXiv* **2022**, arXiv:2203.13168.
38. Zhou, Y.; Yang, C.; Wang, P.; Wang, C.; Wang, X.; Van, N.N. ViT-FuseNet: MultiModal fusion of vision transformer for vehicle-infrastructure cooperative perception. *IEEE Access* **2024**, *12*, 31640–31651. [CrossRef]
39. Liu, H.; Gu, Z.; Wang, C.; Wang, P.; Vukobratovic, D. A lidar semantic segmentation framework for the cooperative vehicle-infrastructure system. In Proceedings of the 2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall), Hong Kong, China, 10–13 October 2023; pp. 1–5.
40. Yang, D.; Yang, K.; Wang, Y.; Liu, J.; Xu, Z.; Yin, R.; Zhai, P.; Zhang, L. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), Orleans, LA, USA, 10–16 December 2023; Volume 36, pp. 25151–25164.
41. Zhou, L.; Gan, Z.; Fan, J. CenterCoop: Center-based feature aggregation for communication-efficient vehicle-infrastructure cooperative 3d object detection. *IEEE Robot. Autom. Lett.* **2023**, *9*, 3570–3577. [CrossRef]
42. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
43. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the Computer Vision—ECCV 2018: 15th European Conference, Munich, Germany, 8–14 September 2018; pp. 3–19.
44. Shen, J.; Chen, Y.; Liu, Y.; Zuo, X.; Fan, H.; Yang, W. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognit.* **2024**, *145*, 109913. [CrossRef]
45. Kendall, A.; Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 4–9 December 2017; p. 30.
46. Nabati, R.; Qi, H. Centerfusion: Center-based radar and camera fusion for 3d object detection. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 5–9 January 2021; pp. 1527–1536.
47. 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. In Proceedings of the 36th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 28 November–9 December 2022; Volume 35, pp. 10421–10434.

48. Li, G.; Yang, L.; Lee, C.G.; Wang, X.; Rong, M. A Bayesian deep learning RUL framework integrating epistemic and aleatoric uncertainties. *IEEE Trans. Ind. Electron.* **2020**, *68*, 8829–8841. [CrossRef]
49. Zhang, Y.; Xiao, Y.; Zhang, Y.; Zhang, T. Video saliency prediction via single feature enhancement and temporal recurrence. *Eng. Appl. Artif. Intell.* **2025**, *160*, 111840. [CrossRef]
50. Demir, A.; Yilmaz, F.; Kose, O. Early detection of skin cancer using deep learning architectures: Resnet-101 and inception-v3. In Proceedings of the 2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey, 3–5 October 2019; pp. 1–4.
51. Vangimalla, R.R.; Sreevalsan-Nair, J. A multiscale consensus method using factor analysis to extract modular regions in the functional brain network. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) in Conjunction with the 43rd Annual Conference of the Canadian Medical and Biological Engineering Society, Virtual, 20–24 July 2020; pp. 2824–2828.
52. Zhang, Y.; Wu, C.; Guo, W.; Zhang, T.; Li, W. CFANet: Efficient detection of UAV image based on cross-layer feature aggregation. *IEEE Trans. Geosci.* **2023**, *61*, 5608911. [CrossRef]
53. Fan, S.; Yu, H.; Yang, W.; Yuan, J.; Nie, Z. Quest: Query stream for vehicle-infrastructure cooperative perception. *arXiv* **2023**, arXiv:2308.01804.
54. Wen, L.; Li, X.; Gao, L. A transfer convolutional neural network for fault diagnosis based on ResNet-50. *Neural Comput. Appl.* **2019**, *32*, 6111–6124. [CrossRef]
55. Haque, M.F.; Lim, H.Y.; Kang, D.S. Object detection based on VGG with ResNet network. In Proceedings of the 2019 International Conference on Electronics, Information, and Communication (ICEIC), Auckland, New Zealand, 22–25 January 2019; pp. 1–3.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.