UECP: Uncertainty-Enhanced Collaborative Perception

Kang Yang[1], Tianci Bu[2], Peng Wang[1], Deying Li[1], Wen Jie[3], and Yongcai Wang[1]*

[1] Renmin University of China, Beijing, China
{yangkang1205,peng.wang,deyingli,ycw}@ruc.edu.cn
[2] College of Systems Engineering, National University of Defense Technology, Changsha, China
btc010001@gmail.com
[3] Intelligent Shipping Center, China Waterborne Transport Research Institute, Beijing, China
wenjie@wti.ac.cn

Abstract. Collaborative perception serves as a pivotal solution to enhance the perception capability of individual agents in autonomous driving, where a core challenge lies in seeking reliable evidence to quantify and weight the contribution of each participating agent. Existing methods typically rely on a confidence map (co-trained with the detection head), which is, however, inherently correlated with the detection results and thus fails to provide unbiased physical evidence. Furthermore, how to deeply integrate evidence into the cooperative fusion process remains an open question. To address these issues, this paper first proposes *uncertainty map*, a physically grounded and unambiguous metric for evaluating perception quality. This map is directly supervised by real-time sensor signals (i.e., LiDAR point density), ensuring decoupling from detection noise and thereby providing physical scenario-aware evidence for weighting agent contribution. Based on this map, we develop the Uncertainty-Enhanced Collaborative Perception (UECP) framework, centered on the Uncertainty-Aware Pyramid Fusion (UAPF) module. UAPF uses a coarse-to-fine strategy, with two key components: Uncertainty-Weighted Downsampling (UWD) for high-fidelity feature preservation, and Uncertainty-Guided Residual Fusion (UGRF) to reinforce ego features, suppressing noise and ensuring robust fusion. Extensive experiments on real-world datasets show UECP outperforms SOTA methods in effectiveness and robustness by embedding the uncertainty map into fusion. Code will be publicly available.

Keywords: Collaborative Perception · Uncertainty Map · Feature Fusion · Autonomous Driving

* Corresponding author.

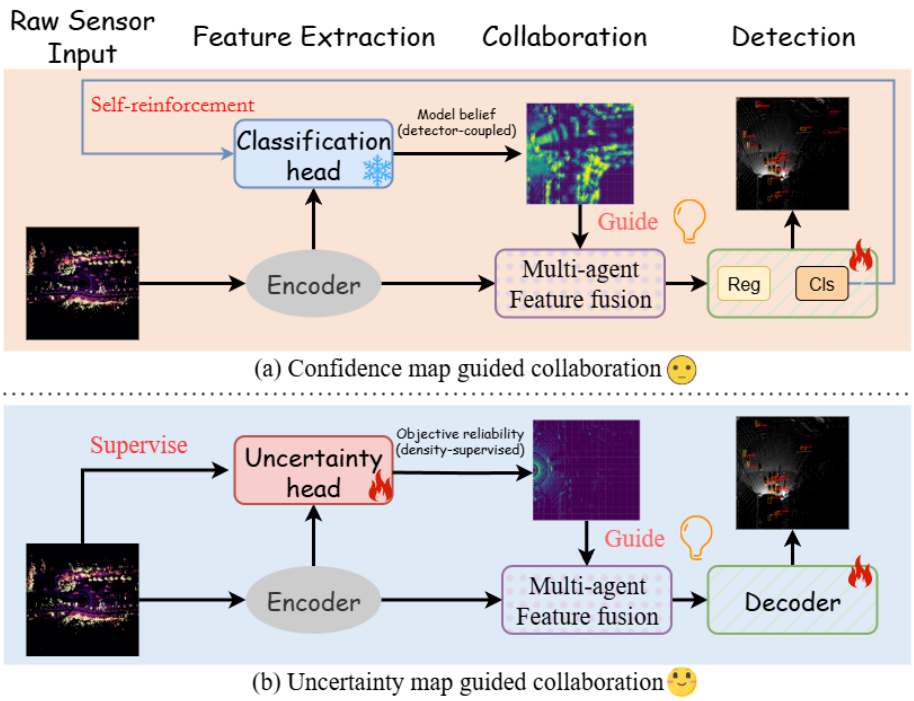


Fig. 1: The difference between traditional confidence map and the proposed uncertainty map. The confidence map is co-learned with the classification head, while the uncertainty map is supervised by the LiDAR point density.

1 Introduction

Collaborative perception is critical for autonomous driving, as it enables the ego agent to perceive non-line-of-sight hazards and thereby enhance driving safety [1, 13, 16, 27, 35, 51, 53, 57, 58]. A core challenge in collaborative perception lies in designing effective multi-agent information fusion strategies. Existing fusion approaches are generally categorized into early fusion [3, 29], intermediate fusion [16, 30, 46, 53, 54], and late fusion [19], among which intermediate fusion—conducting feature-level fusion via compressed feature communication—achieves an optimal balance between communication efficiency and detection performance. Within intermediate fusion, Bird's-Eye-View (BEV) feature fusion has become the dominant and most promising paradigm [50, 52]. However, mainstream BEV fusion methods [5, 46, 53] often suffer from significant collaborative noise, which stems from view discrepancies, spatial misalignment, and other practical challenges. Developing a fusion mechanism that effectively mitigates such noise is therefore essential for achieving robust and high-precision perception.

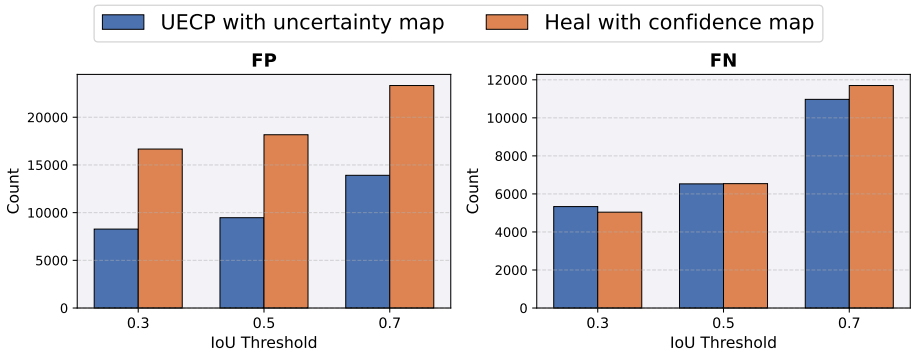


Fig. 2: Comparison of false positives (FP) and false negatives (FN) between the proposed UECP and HEAL on the DAIR-V2X dataset under different IoU thresholds (0.3/0.5/0.7).

To address this noise issue, existing methods typically adopt confidence maps derived from the detector's classification head to adaptively weight features during fusion [7,16,18,33,49,55]. Nevertheless, these confidence maps are inherently coupled with detection scores, failing to provide unbiased, physically grounded guidance for fusion. For example, they often struggle to suppress false positives, as demonstrated in Fig. 2. Additionally, how to deeply integrate such guidance information into the fusion process remains an under-explored problem.

To tackle these limitations, seeking physically grounded evidence is an important solution. For this purpose, *uncertainty map* is proposed in this paper. Fig. 1 illustrates the fundamental difference between the traditional confidence map and the proposed uncertainty map. Uncertainty map is supervised by raw data feature, i.e., local LiDAR point cloud density. Because in LiDAR detections, diverse uncertainty sources (e.g., occlusion, truncation, specular surfaces) all physically manifest as reduced point returns, making point density a reasonal physical proxy of observation confidence.

With the uncertianty informaiton, in subsequent fusion, high uncertainty in collaborative perception does not mean "undetectable" but rather "in need of neighboring information". This characteristic enables the fusion module to incooperate the uncertainty information into fusion process. More importantly, the uncertainty map is decoupled from the detection head training, making it more complementary to guiding fusion and improving detection performance.

Based on the uncertainty map, we further propose a novel multi-scale, uncertainty-involved fusion framework named Uncertainty-Enhanced Collaborative Perception (UECP). The core of UECP is an Uncertainty-Aware Pyramid Fusion (UAPF) module, which maximizes the utility of the uncertainty map through two key mechanisms. First, we employ Uncertainty-Weighted Downsampling (UWD) to preserve high-quality information during the uncertainty-weighted pyramid construction, enhancing robustness to noise. Second, and most criti-

cally, we introduce an Uncertainty-Guided Residual Fusion (UGRF) module at each pyramid scale. The UGRF generates a comprehensive fused feature representation for the ego-agent via an uncertainty-guided weighted summation, while its residual design stabilizes training by mitigating large variations in ego-centric features.

We perform extensive experiments to evaluate UECP on the real-world DAIR-V2X [57] and V2V4REAL [51] datasets. Our results demonstrate that UECP consistently outperforms state-of-the-art methods, including HEAL [33] and CoBEVT [50]. Furthermore, our method retains its superior performance even under noisy conditions underscoring its inherent robustness. In summary, our main contributions are three-fold:

- We propose uncertainty map, supervised by physical sensor signals, that acts as a robust and precise guidance signal for fusion in contrast to traditional confidence map.
- We propose UECP, a comprehensive and novel fusion framework, which deeply involves uncertainty in the fusion process to overcome the limitations of confidence maps and the challenges posed by collaborative noise.
- Extensive experiments on public real-world datasets validate that our proposed methods attain state-of-the-art (SOTA) performance across various precision metrics.

2 Related Work

2.1 Cooperative Perception by Intermediate Fusion

To overcome the perceptual limitations of individual agents, cooperative perception enables sensory information sharing across a network. The prevailing paradigm has gravitated toward intermediate fusion, which exchanges BEV feature maps to strike a balance between communication efficiency and fusion effectiveness. Intermediate fusion avoids the prohibitive bandwidth costs of early fusion (raw data) [3, 6] and outperforms the performance ceiling of late fusion. The core of this paradigm lies in the feature interaction mechanism. Initial works explored spatially-aware message passing using Graph Neural Networks (GNNs) [46], while AttnFuse [53] first introduced attention to dynamically model inter-agent relationships. This line of research was significantly advanced by Transformer-based architectures like V2X-ViT [52], which leverage powerful self-attention to learn global context from all collaborators. Concurrently, methods such as DiscoNet [29] have explored knowledge distillation to enrich individual agents' features.

2.2 Efficiency and Robustness in Fusion

Despite their effectiveness, the dense feature exchange in these models poses substantial computational and bandwidth challenges. This has spurred a parallel research direction focused on sparse and efficient fusion. Where2comm [16]

pioneered this area by learning communication priorities, selecting only the most salient feature regions. Subsequent works advanced this paradigm by focusing on instance-level features: TransIFF [7] adopted a transformer-based approach, QUEST [9] proposed interpretable query cooperation, and ACCO [55] introduced sparse anchor collaboration—all to further reduce bandwidth overhead. Beyond performance and efficiency, a key focus of recent research is ensuring the robustness and versatility required for real-world deployment. A large body of work has targeted enhancing robustness against systemic uncertainties inherent in cooperative systems, including specialized methods to mitigate pose error impacts [19, 34, 43, 44, 59] and compensate for communication latency [25, 28, 42, 47]. Building on these foundational robustness efforts, a further layer of complexity arises when collaborating agents are heterogeneous. Research in this space has explored interoperability across different sensor modalities: CoBEVT [50] and CoCa3D [17] extended the paradigm to camera-based systems, while V2VFormer++ [56], CoCMT [45], and BM2CP [60] targeted multimodal suites. Concurrently, another research line, including HEAL [33], HM-VIT [48], and STAMP [12], focuses on ensuring feature compatibility when agents employ disparate internal models or data types. In this paper, we facilitate essential and supportive information exchange among agents guided by the uncertainty map.

2.3 Uncertainty Modelling

Uncertainty plays a critical role in safe autonomous driving. We distinguish between aleatoric uncertainty, stemming from sensor noise, environmental variability, and measurement errors, and epistemic uncertainty, which reflects model ignorance due to limited data or coverage gaps [10, 20]. To estimate epistemic uncertainty, methods such as Bayesian neural networks, Monte Carlo dropout, and deep ensembles approximate predictive distributions via sampling or ensembling [4, 11, 22]. In contrast, aleatoric uncertainty is often captured in a single forward pass through learned variance or estimated using evidential and conformal prediction techniques, which offer calibration during training or post hoc [2, 39]. Although these approaches have proven effective in standard 3D detection and autonomous driving tasks [8, 14, 21, 24, 36], their application to cooperative perception remains limited. Many existing methods rely on learned confidence scores [7, 16, 18, 33, 49, 55], typically derived from model posteriors $p(\text{object} \mid \text{features})$. However, confidence-based fusion has inherent limitations in collaborative settings, as model-derived confidence often fails to reflect the reliability of shared sensor evidence. This typically leads to two failure modes. First, when observations are sparse, each agent's confidence may be low, and fusion discards weak yet consistent signals via thresholding—causing missed detections. Second, fusion can amplify false positives, where a spurious high-confidence activation dominates message weighting. These issues underscore a mismatch between confidence-based fusion and the demands of robust collaboration, motivating evidence-driven uncertainty representations that better reflect sensor quality.

3 Method

3.1 Collaborative Perception With Uncertainty

We consider a collaborative perception system consisting of N agents, whose interactions are modeled by a communication graph G. Each agent n is a vertex equipped with a unique sensor setup (e.g., LiDAR or camera) and captures local sensory input $\mathcal{X}_n$. While any agent can function as both an information sender and receiver, we focus on a common setting where a single agent i (the "ego-agent") acts as the receiver, and all other agents j serve as senders. The goal is to enhance the ego-agent's 3D object detection performance by aggregating information from its collaborators. We denote $\mathcal{O}_i$ the ground-truth objects surrounding agent i. Our objective is to learn a collaborative perception model Φ_θ based on $\{\mathcal{X}_i, \mathcal{U}_i, \{\mathcal{M}_{j\to i}\}_{j\in\mathcal{N}(i)}\}$ to maximize the detection gain $g(\text{prediction}, \text{groundtruth})$ for the ego-agent. The optimization problem with consideration of the observation uncertainty is formulated as:

$$\max_{\theta} \quad g\left(\Phi_\theta\left(\mathcal{X}_i, \mathcal{U}_i, \{\mathcal{M}_{j\to i}\}_{j\in\mathcal{N}(i)}\right), \mathcal{O}_i\right), \tag{1}$$

where $\mathcal{U}_i$ is the uncertainty clue of $\mathcal{X}_i$; $\mathcal{M}_{j\to i} = \mathcal{P}(\mathcal{X}_j, \mathcal{U}_j)$ denotes the message transmitted from agent j to agent i with the uncertainty information. $\mathcal{N}(i)$ is the set of neighbors of agent i in the graph G. The key challenge is to optimize the perception model Φ_θ and to properly design the uncertainty clues $\mathcal{U}_i$ to effectively fuse information from multiple agents.

3.2 Uncertainty Map Generation

To quantify the uncertainty of each agent's local perception, unlike existing methods that typically adopt a confidence map co-learned with the classification head, we propose a novel uncertainty map. Specifically, the uncertainty map is learned via a dedicated uncertainty head, supervised directly by raw sensor data, to provide physical grounding for feature fusion. We demonstrate that this physically grounded uncertainty information enhances model robustness and mitigates collaborative noise. We next describe how the raw observation data is processed to build the supervision signal and how the uncertainty map is learned.

BEV encoder. For the i-th agent, given its input $\mathcal{X}_i$, we extract the BEV feature $\mathcal{F}_i = \Phi_{\text{enc}}(\mathcal{X}_i) \in \mathcal{R}^{H\times W\times C}$ using a BEV encoder Φ_{enc}, where H, W and C denote the height, width, and channel dimensions, respectively.

Uncertainty Model. The uncertainty head $\Phi_{\text{unc}}(\cdot)$ predicts the spatial distribution of sensing reliability from each agent's BEV feature. Given the extracted BEV representation $\mathcal{F}_i$, it produces a dense uncertainty map:

$$\mathcal{U}_i = \Phi_{\text{unc}}(\mathcal{F}_i) \in [0,1]^{H\times W}, \tag{2}$$

, where $\mathcal{U}_i[x, y]$ indicates the uncertainty of grid $[x, y]$, with higher values indicating lower sensing reliability. Importantly, the uncertainty map captures *sensing*

reliability rather than semantic objectness. Semantic relevance is handled by the detector's learned features; the uncertainty map only modulates how much to trust each agent's contribution during fusion. High-density background regions thus contribute reliable negative evidence (i.e., "no object here"), which is equally valuable for suppressing false positives.

Uncertainty Head Training. To train the uncertainty model $\Phi_{\text{unc}}(\cdot)$, raw LiDAR scans are projected onto a BEV grid with resolution $\Delta x = \Delta y = 0.4$m to compute a density map $\mathcal{D}_i$ by counting points per cell. The map is normalized to $[0, 1]$, and ground-truth uncertainty is defined as:

$$\mathcal{U}_i^{\text{gt}} = \mathbf{1} - \text{norm}(\mathcal{D}_i), \tag{3}$$

, where sparser regions indicate higher uncertainty. The uncertainty $\mathcal{U}_i$ is trained toward $\mathcal{U}_i^{\text{gt}}$ using a hybrid loss that combines a continuous focal-style regression term to emphasize difficult regions and a Sobel-gradient consistency term to preserve structural boundaries:

$$\begin{aligned}\mathcal{L}_{\text{un}} = \alpha \frac{1}{HW} \sum_{x,y} |e_i[x,y]|^{\rho+1} \\ + (1-\alpha) w_g \left(\|\nabla_x \mathcal{U}_i - \nabla_x \mathcal{U}_i^{\text{gt}}\|_1 + \|\nabla_y \mathcal{U}_i - \nabla_y \mathcal{U}_i^{\text{gt}}\|_1 \right),\end{aligned} \tag{4}$$

, where $e_i = \mathcal{U}_i - \mathcal{U}_i^{\text{gt}}$, ρ is the focal exponent, and ∇_x, ∇_y are fixed Sobel filters [41]. This learned uncertainty map subsequently serves as a guidance signal for collaborative fusion.

3.3 Uncertainty Enhanced Collaborative Perception

Building upon the proposed uncertainty map, we further present the Uncertainty Enhanced Collaborative Perception (UECP) framework, as depicted in Fig. 3. UECP first feeds raw sensor data $\mathcal{X}_i$ into a BEV Encoder to produce feature $\mathcal{F}_i$, and subsequently outputs the uncertainty map $\mathcal{U}_i$ via the dedicated uncertainty head. Thereafter, $\mathcal{F}_i$, $\mathcal{U}_i$, alongside the features $\mathcal{F}_j$ and uncertainty maps $\mathcal{U}_j$ of neighboring agents $j \in \mathcal{N}_i$, where $\mathcal{N}_i$ denotes the neighborhood of agent i, are transmitted via suitable neighborhood communication protocols [16,33,50], a standard operation in multi-agent collaborative perception to ensure efficient and reliable feature fusion. These aggregated inputs are then fed into UECP's core component, namely the Uncertainty-Aware Pyramid Fusion (UAPF) module, for final processing.

Uncertainty-Aware Pyramid Fusion (UAPF). UAPF performs multi-scale collaborative fusion to facilitate effective multi-agent interaction. UAPF integrates two key submodules: (1) the Uncertainty-Weighted Downsampling (UWD) module to enable the model to capture rich collaborative features across scales, from coarse to fine; and (2) the Uncertainty-Guided Residual Fusion (UGRF) block. UWD extracts high-quality features while enhancing robustness through uncertainty-guided downsampling. Meanwhile, UGRF refines the receiving agent's

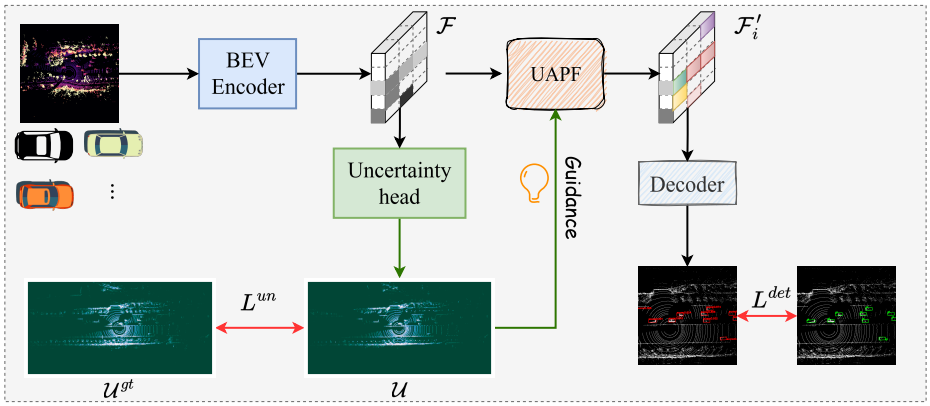


Fig. 3: The overall pipeline of our proposed framework, UECP. The framework begins by processing raw sensor data from each agent independently through a shared BEV Encoder to extract initial features. In parallel, an uncertainty head predicts a physically-grounded uncertainty map for each agent. This map serves as the key input to our Uncertainty-Aware Pyramid Fusion (UAPF) module. The resulting deeply-fused feature map is finally passed to a detection head to generate the cooperative predictions.

features via an uncertainty-aware enhancement process, incorporating an ego-centric residual structure. A detailed architectural illustration of UAPF is provided in Fig. 4. The process within the UAPF module unfolds as follows.

As a prerequisite for fusion, all uncertainty maps and BEV features are first transformed into a unified ego-coordinate frame [19]. The UAPF module then takes the aligned uncertainty map $\mathcal{U}_i$, BEV feature $\mathcal{F}_i$, and $\mathcal{M}_{j\to i}$ that encapsulates neighbor agents' feature and uncertainty $\{\mathcal{F}_j, \mathcal{U}_j\}, j \in \mathcal{N}(i)$ as input to produce the final collaborative feature $\mathcal{F}_i'$.

The feature data and uncertainty map are firstly downsampled to multiple scales. At each scale level s, we employ an Uncertainty-Weighted Downsampling (UWD) module, Φ_{uwd}, to process the input feature map $\mathcal{F}$ and uncertainty map $\mathcal{U}$. This operation generates the downsampled outputs $(\mathcal{F}_s, \mathcal{U}_s) = \Phi_{\text{uwd}}(\mathcal{F}, \mathcal{U})$. Subsequently, an Uncertainty-Guided Pyramid Fusion (UGRF) module processes these inputs to generate a fused feature map $\hat{\mathcal{F}}_{i,s}$ for the ego agent i. This fused feature is then post-processed by a convolutional refinement block to enhance its quality and informativeness. To achieve multi-scale fusion, the refined feature $\hat{\mathcal{F}}_{i,s}$ is stored as a prior and upsampled to the next scale, where it is added element-wise to the newly fused feature. The final output $\mathcal{F}_i'$ maintains the input BEV resolution.

Uncertainty-Weighted Downsampling (UWD). In a multi-scale feature pyramid, constructing lower-resolution representations that retain critical information from high-resolution inputs is essential for robust fusion. Conven-

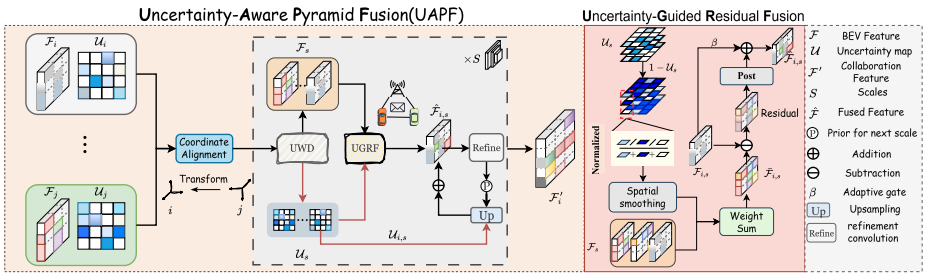


Fig. 4: The overall pipeline of UAPF. Within the UAPF module, our Uncertainty-Weighted Downsampling (UWD) mechanism first constructs a high-fidelity feature pyramid. Then, at each scale, an Uncertainty-Guided Residual Fusion (UGRF) block robustly integrates collaborative information.

tional downsampling methods such as max-pooling or average-pooling often suffer from noise sensitivity and information dilution [15, 38]. To overcome these limitations, we propose Uncertainty-Weighted Downsampling (UWD), which adaptively weighs feature contributions based on their reliability to preserve informative regions during the downsampling process, as illustrated in Fig. 5. UWD takes a feature map $\mathcal{F}$ and an uncertainty map $\mathcal{U}$ as input and produces their downsampled counterparts $\mathcal{F}_s$ and $\mathcal{U}_s$. It is implemented as the ratio of two average-pooling operations, enabling efficient, uncertainty-aware aggregation with minimal overhead. Given an uncertainty map $\mathcal{U}$, the downsampled feature is computed as:

$$\begin{aligned} \mathcal{F}_s &= \frac{\mathrm{AvgPool}_s(\mathcal{F} \odot (1-\mathcal{U}))}{\mathrm{AvgPool}_s(1-\mathcal{U}) + \epsilon}, \\ &\text{i.e., for each location } (j,k): \\ (\mathcal{F}_s)_{jk} &= \frac{\frac{1}{|\mathcal{W}_{jk}|}\sum_{p\in\mathcal{W}_{jk}}(1-\mathcal{U}_{(p)})\cdot\mathcal{F}_{(p)}}{\frac{1}{|\mathcal{W}_{jk}|}\sum_{p\in\mathcal{W}_{jk}}(1-\mathcal{U}_{(p)}) + \epsilon}. \end{aligned} \tag{5}$$

, where $\mathrm{AvgPool}_s(\cdot)$ denotes average pooling with a given scale and stride, $\mathcal{W}_{jk}$ represents the pooling window in the original feature map corresponding to location (j,k) in the downsampled feature $\mathcal{F}_s$, and $\odot$ is element-wise multiplication. We set $\epsilon = 10^{-8}$ to avoid unstable small denominators. This design leverages standard pooling operations to enable reliable, quality-aware downsampling with minimal modification.

Uncertainty-Guided Residual Fusion (UGRF). To minimize collaborative fusion noise and ensure stable, consistent feature integration, we design the Uncertainty-Guided Residual Fusion (UGRF) module, a simple yet effective approach centered on residual fusion. For simplicity, the scale index is omitted. Specifically, UGRF first generates a fusion weight λ_n for each agent n, guided

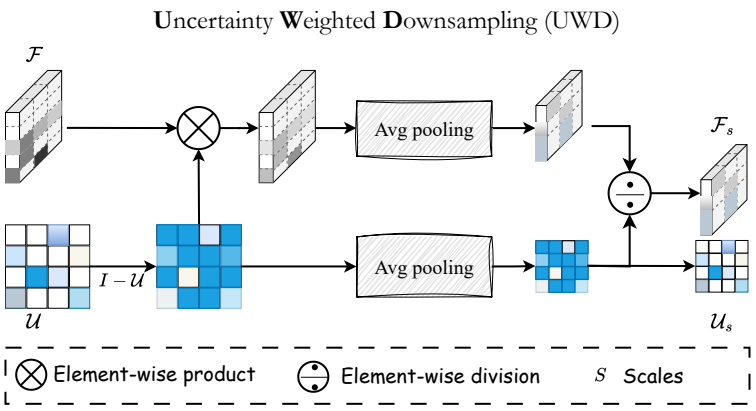


Fig. 5: Uncertainty-Weighted Downsampling module.

by its uncertainty map $\mathcal{U}_n$. The fusion weight is then computed as

$$\lambda_n = \frac{(\epsilon + \gamma \cdot (1 - \mathcal{U}_n)) \odot v_n}{\sum_{m=1}^{N} (\epsilon + \gamma \cdot (1 - \mathcal{U}_m)) \odot v_m + \epsilon}, \tag{6}$$

where $\gamma > 0$ is a learnable scalar (enforced positive via $\gamma = \text{softplus}(\tilde{\gamma})$), $v_n \in \{0,1\}^{H \times W}$ denotes the visibility mask after ego-frame alignment and FOV cropping, and $\epsilon = 10^{-8}$ ensures numerical stability. Finally, a 3×3 Gaussian blur [37] is applied to λ to promote spatial smoothness.

Using the adaptive fusion weights λ_n, the ego agent's fused representation is obtained through a weighted summation operation followed by a residual update:

$$\hat{\mathcal{F}}_i = \mathcal{F}_i + \beta \cdot \phi_{\text{post}} \left(\underbrace{(\sum_{n=1}^{N} \lambda_n \odot \mathcal{F}_n) - \mathcal{F}_i}_{\text{Residual}} \right), \tag{7}$$

where $\mathcal{F}_i$ denotes the ego agent's own feature, and $\mathcal{F}_n$ represents the spatially aligned feature from the n-th collaborator. ϕ_{post} is a shallow convolutional refinement network that enhances the collaborative residual, and $\beta \in (0,1)$ is a learnable gate controlling the strength of the update. This residual formulation preserves the ego feature as the primary representation, while the aggregated information from collaborators acts as a controlled refinement signal. Consequently, UGRF mitigates feature contamination from unreliable collaborators, stabilizes training, and ensures a more robust and reliable fusion outcome.

3.4 Loss function

Finally, the training loss of the model is simply the sum of the regression loss L_{reg}, classification loss L_{cls}, direction loss L_{dir} and uncertainty map loss L_{un}:

$$L_{\text{total}} = \lambda_{\text{reg}} L_{\text{reg}} + \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{dir}} L_{\text{dir}} + \lambda_{\text{un}} L_{\text{un}}, \tag{8}$$

, where the λ are the weighting factors of the different losses used in the optimization process, and L_{un} follows the uncertainty-head training objective in Sec. 3.2.

4 Experiments

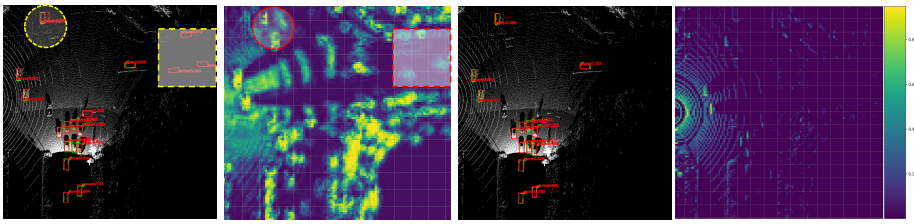
Experimental setup. We evaluate UECP on two real-world cooperative perception benchmarks, DAIR-V2X [57] and V2V4REAL [51]. DAIR-V2X targets vehicle-infrastructure cooperative detection, while V2V4REAL focuses on vehicle-to-vehicle cooperation in urban driving. All methods use PointPillar [23] as the LiDAR backbone under a unified BEV feature-fusion protocol. Following the benchmark settings, point clouds are voxelized with a 0.4 m × 0.4 m BEV grid, and detection AP is reported at IoU thresholds of 0.3, 0.5, and 0.7. Models are trained within the same codebase for 60 epochs using AdamW [32] and a one-cycle learning-rate schedule [40]. We compare against single-agent perception and representative intermediate-fusion baselines, including F-Cooper [5], Attn [53], V2VNet [46], V2X-ViT [52], CoBEVT [50], Where2comm [16], Who2comm [31], CodeFilling [18], and HEAL [33].

4.1 Effectiveness of the uncertainty map

Table 1: Replacing confidence map with uncertainty map consistently improves AP. Δ = Unc − Conf (pp).

Method	**AP@30**			**AP@50**			**AP@70**			**Avg** Δ
	Conf	Unc	Δ	Conf	Unc	Δ	Conf	Unc	Δ	
HEAL	79.11	**80.60**	+1.49	73.95	**75.20**	+1.25	55.56	**58.93**	+3.37	+2.03
Where2comm	74.92	**76.63**	+1.71	67.90	**71.48**	+3.58	48.86	**54.72**	+5.86	+3.71

To validate the generality of our uncertainty map, we replace the confidence map with our uncertainty map in Where2comm and HEAL (Tab. 1). The improvement is especially pronounced at IoU@0.7, indicating that the uncertainty map more effectively suppresses false positives. We attribute this to the confidence self-reinforcement problem: confidence maps are co-learned with the detector, so spurious high-confidence activations dominate fusion weighting and



(a) HEAL with conf (b) Confidence map (c) Heal with uncertainty map (d) 1 – uncertainty map

Fig. 6: A comparative analysis of HEAL's performance guided by confidence versus that guided by uncertainty maps.

amplify false positives across agents. In contrast, the uncertainty map is decoupled from detection predictions and breaks this self-reinforcing cycle, providing an independent and physically grounded fusion signal. Fig. 6 further confirms this with qualitative comparisons, demonstrating consistent gains across different baseline architectures.

Table 2: Comparison of mainstream works on the V2V4REAL and DAIR-V2X dataset. The best performance is highlighted in **bold**, and the second-best performance is marked in blue.

Method	V2V4REAL			DAIR-V2X		
	AP@30	AP@50	AP@70	AP@30	AP@50	AP@70
No Coll	46.12	44.03	27.71	59.97	54.23	41.27
Fcooper	47.97	40.83	30.64	76.94	70.57	54.65
Attn	48.18	40.06	28.75	69.83	64.06	48.98
V2VNet	45.12	43.24	31.75	69.25	65.52	47.84
V2X-VIT	40.11	37.18	24.53	77.28	72.47	57.97
CoBEVT	63.13	60.10	33.94	75.67	69.34	49.77
Where2comm	49.61	45.10	34.45	74.92	67.90	48.86
Who2comm	52.99	46.80	18.14	77.08	71.22	52.17
HEAL	51.57	48.70	35.24	79.11	73.95	55.56
CodeFilling	60.15	59.78	35.07	79.85	75.18	57.26
UECP	**65.69** (+2.56)	**63.38** (+3.28)	**38.90** (+3.66)	**81.78** (+1.93)	**77.59** (+2.41)	**61.12** (+3.15)

4.2 Comparison with state-of-the-art

Detection Performance. Tab. 2 presents a comprehensive comparison of detection performance on the DAIR-V2X and V2V4REAL datasets. Our proposed method, UECP, achieves new state-of-the-art results across all metrics on both benchmarks. Notably, our approach demonstrates a substantial advantage at high-precision thresholds, achieving an AP@70 of 61.12% on DAIR-V2X and

38.90% on V2V4REAL. On DAIR-V2X, UECP outperforms the second-best method by +1.93, +2.41, and +3.15 in AP@30, AP@50, and AP@70, respectively. The performance gains are even more pronounced on V2V4REAL, with improvements of +2.56, +3.28, and +3.66 across the same metrics.

Robustness. We evaluate robustness under pose noise (σ from 0 to 1 m, following CoAlign [34]) and communication latency (0–500 ms, following SyncNet [26]). As shown in Fig. 7, UECP maintains the highest accuracy under both disturbances. Communication-heavy methods degrade rapidly at high delay, whereas UECP's uncertainty-aware fusion preserves stable performance even under large temporal misalignment. These two stress tests target different failure modes: pose noise perturbs cross-agent spatial alignment, while latency makes shared observations temporally stale. UECP uses the same reliability weighting for both cases, suggesting that the uncertainty map captures a general feature-trust signal rather than a disturbance-specific correction.

Efficiency and communication. The robustness gain does not require a heavy auxiliary message. UECP transmits only one additional uncertainty-map channel alongside BEV features, adding about $1/256 \approx 0.39\%$ bandwidth when $C = 256$ feature channels are used. This extra reliability cue is much smaller than the shared BEV tensor itself and remains compatible with sparse communication policies such as Where2comm [16] and Who2comm [31], since it guides feature trust rather than replacing the communication selection rule. The model also stays within a practical operating point, with 8.34 M parameters, about 100 G FLOPs, and 9 FPS in our benchmark. These efficiency measurements follow the same inference protocol as the main comparison, so the added uncertainty branch does not change the evaluation pipeline. Thus, the gains in Fig. 7 come from more reliable feature weighting rather than simply increasing communication or computation.

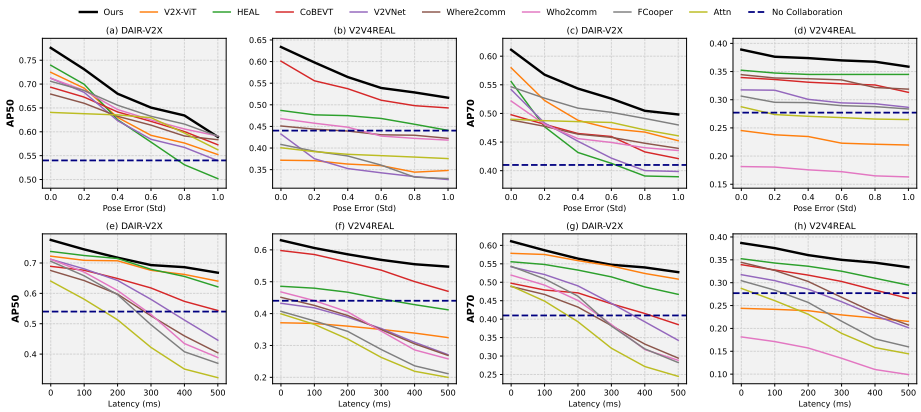


Fig. 7: Robustness analysis under pose error and latency error on DAIR-V2X and V2V4REAL datasets.

4.3 Ablation studies

Table 3: Ablation studies for proposed components.

Components				Average Precision (AP)		
UM	MS	UWD	UGRF	AP@0.3	AP@0.5	AP@0.7
				76.94	70.57	54.65
✓				78.52	72.39	56.92
✓	✓			79.98	75.73	59.74
✓	✓	✓		80.08	76.05	60.63
✓	✓	✓	✓	**81.78**	**77.59**	**61.12**

All ablations use the DAIR-V2X validation set: Tab. 3 isolates component accumulation, and Tab. 4 studies UAPF design choices.

Effectiveness of our key components. We conduct an ablation study to validate the effectiveness of our key components, with detailed results reported in Tab. 3. Our baseline, shown in the first row, emulates the F-Cooper setting by directly applying max fusion for collaboration. The subsequent rows demonstrate the progressive benefits of our contributions. First, incorporating our proposed uncertainty map yields a significant performance gain over the baseline. Building on this, the addition of the multi-scale (MS) pyramid architecture provides another substantial improvement, particularly at the stringent IoU@0.7 threshold. Finally, the results confirm that our two core modules, Uncertainty-Weighted Downsampling (UWD) and Uncertainty-Guided Residual Fusion (UGRF), each consistently and incrementally enhance the final detection performance.

Table 4: Design choice ablation of pyramid scales, map types, and fusion operations in the fusion module.

(a) Effect of pyramid scales.

Scales	AP@30	AP@50	AP@70
(1)	80.03	75.43	59.03
(2,1)	81.00	76.12	59.79
(4,2,1)	**81.78**	**77.59**	**61.12**

(b) Effect of map type.

Map	AP@30	AP@50	AP@70
No	77.89	72.63	55.91
Confidence	78.68	73.60	56.47
Density-direct	81.73	77.28	60.77
Uncertainty	**81.78**	**77.59**	**61.12**

(c) Effect of fusion op.

Operation	AP@30	AP@50	AP@70
Max	80.93	76.01	59.86
Mean	80.33	75.13	57.66
Sum	**81.78**	**77.59**	**61.12**

Fine-grained ablation. Tab. 4 further studies UAPF design choices. The multi-scale setting (4, 2, and 1) gives the best results in Tab. 4(a). In Tab. 4(b), learned uncertainty outperforms no guidance, confidence, and direct use of the density-based ground truth at inference, indicating that training converts density supervision into a feature-level ambiguity cue. Tab. 4(c) shows that additive fusion (`sum`) gives the highest AP.

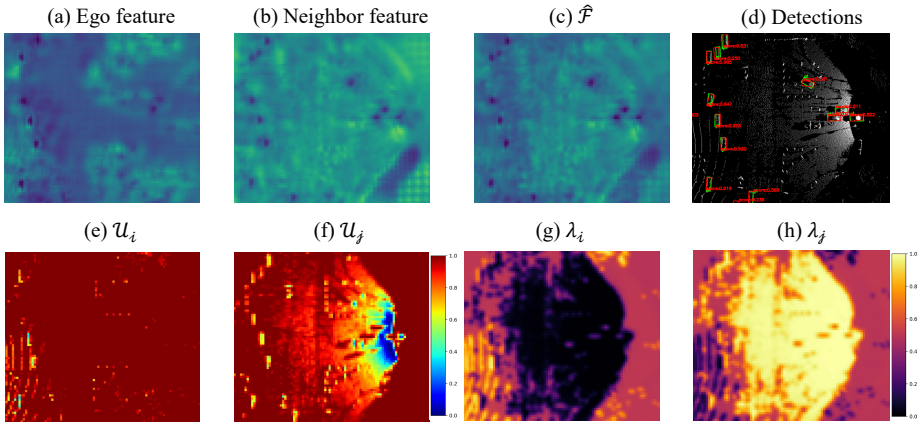


Fig. 8: Visualization of collaboration in UECP. Green and red denote ground truth and detection, respectively.

4.4 Qualitative evaluation

Visualization of information selection and representation. Fig. 8 illustrates UECP's collaborative process. Subfigures (a)–(d) show the ego-coordinate fused feature map $\hat{\mathcal{F}}$ and detections, where neighbor cues complement ego features. Subfigures (e)–(h) show ego/neighbor uncertainty maps and the derived weights λ from Eq. (6), highlighting how uncertainty redistributes trust across agents.

5 Conclusion

UECP is a physical evidence-guided collaborative 3D detection system built on uncertainty-aware pyramid fusion, uncertainty-weighted downsampling, and uncertainty-guided residual fusion. By learning an uncertainty map from density-based physical evidence and using it to guide feature trust, UECP achieves state-of-the-art perception performance and remains robust under pose errors and communication latency on real-world benchmarks.

References

1. Alotaibi, E.T., Alqefari, S.S., Koubaa, A.: Lsar: Multi-uav collaboration for search and rescue missions. Ieee Access **7**, 55817–55832 (2019)
2. Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 (2021)
3. 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 **23**(3), 1852–1864 (2020)

4. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: International conference on machine learning. pp. 1613–1622. PMLR (2015)
5. Chen, Q., Ma, X., Tang, S., Guo, J., Yang, Q., Fu, S.: F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In: Proceedings of the 4th ACM/IEEE Symposium on Edge Computing. pp. 88–100 (2019)
6. Chen, Q., Tang, S., Yang, Q., Fu, S.: Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). pp. 514–524. IEEE (2019)
7. Chen, Z., Shi, Y., Jia, J.: Transiff: An instance-level feature fusion framework for vehicle-infrastructure cooperative 3d detection with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18205–18214 (2023)
8. Durasov, N., Mahmood, R., Choi, J., Law, M.T., Lucas, J., Fua, P., Alvarez, J.M.: Uncertainty estimation for 3d object detection via evidential learning. arXiv preprint arXiv:2410.23910 (2024)
9. Fan, S., Yu, H., Yang, W., Yuan, J., Nie, Z.: Quest: Query stream for practical cooperative perception. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 18436–18442. IEEE (2024)
10. Feng, D., Rosenbaum, L., Dietmayer, K.: Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection. In: 2018 21st international conference on intelligent transportation systems (ITSC). pp. 3266–3273. IEEE (2018)
11. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
12. Gao, X., Xu, R., Li, J., Wang, Z., Fan, Z., Tu, Z.: Stamp: Scalable task and model-agnostic collaborative perception. arXiv preprint arXiv:2501.18616 (2025)
13. Guo, Y., Ma, J.: Leveraging existing high-occupancy vehicle lanes for mixed-autonomy traffic management with emerging connected automated vehicle applications. Transportmetrica A: Transport Science **16**(3), 1375–1399 (2020)
14. He, J.J., Hu, P., Laungani, D., Anguelov, D.: Modeling confidence in lidar-based bev perception. In: Conference on Robot Learning (CoRL). pp. 330–340 (2022)
15. Honari, S., Yosinski, J., Vincent, P., Pal, C.: Recombinator networks: Learning coarse-to-fine feature aggregation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5743–5752 (2016)
16. Hu, Y., Fang, S., Lei, Z., Zhong, Y., Chen, S.: Where2comm: Communication-efficient collaborative perception via spatial confidence maps. Advances in neural information processing systems **35**, 4874–4886 (2022)
17. Hu, Y., Lu, Y., Xu, R., Xie, W., Chen, S., Wang, Y.: Collaboration helps camera overtake lidar in 3d detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9243–9252 (June 2023)
18. Hu, Y., Peng, J., Liu, S., Ge, J., Liu, S., Chen, S.: Communication-efficient collaborative perception via information filling with codebook. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15481–15490 (2024)
19. Huang, Z., Wang, S., Wang, Y., Li, W., Li, D., Wang, L.: Roco: Robust cooperative perception by iterative object matching and pose adjustment. In: ACM Multimedia 2024 (2024)

20. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)
21. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)
22. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30 (2017)
23. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
24. Lee, Y., Chen, C.H., Cheng, H.C., Lin, W.C.: What you see is what you get: A probabilistic volumetric fusion for environment perception. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 8126–8132 (2022)
25. Lei, Z., Ren, S., Hu, Y., Zhang, W., Chen, S.: Latency-aware collaborative perception. In: European Conference on Computer Vision. pp. 316–332. Springer (2022)
26. Lei, Z., Ren, S., Hu, Y., Zhang, W., Chen, S.: Latency-aware collaborative perception. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII. p. 316–332. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-19824-3_19, https://doi.org/10.1007/978-3-031-19824-3_19
27. Li, J., Xu, R., Liu, X., Li, B., Zou, Q., Ma, J., Yu, H.: S2r-vit for multi-agent cooperative perception: Bridging the gap from simulation to reality. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 16374–16380. IEEE (2024)
28. Li, J., Xu, R., Liu, X., Ma, J., Chi, Z., Ma, J., Yu, H.: Learning for vehicle-to-vehicle cooperative perception under lossy communication. IEEE Transactions on Intelligent Vehicles (2023)
29. Li, Y., Ren, S., Wu, P., Chen, S., Feng, C., Zhang, W.: Learning distilled collaboration graph for multi-agent perception. Advances in Neural Information Processing Systems **34**, 29541–29552 (2021)
30. Liu, Y.C., Tian, J., Glaser, N., Kira, Z.: When2com: Multi-agent perception via communication graph grouping. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 4106–4115 (2020)
31. Liu, Y.C., Tian, J., Ma, C.Y., Glaser, N., Kuo, C.W., Kira, Z.: Who2com: Collaborative perception via learnable handshake communication. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 6876–6883 (2020). https://doi.org/10.1109/ICRA40945.2020.9197364
32. Loshchilov, I.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
33. Lu, Y., Hu, Y., Zhong, Y., Wang, D., Chen, S., Wang, Y.: An extensible framework for open heterogeneous collaborative perception. arXiv preprint arXiv:2401.13964 (2024)
34. Lu, Y., Li, Q., Liu, B., Dianati, M., Feng, C., Chen, S., Wang, Y.: Robust collaborative 3d object detection in presence of pose errors. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 4812–4818. IEEE (2023)
35. Ma, J., Leslie, E., Ghiasi, A., Huang, Z., Guo, Y.: Empirical analysis of a freeway bundled connected-and-automated vehicle application using experimental data. Journal of Transportation Engineering, Part A: Systems **146**(6), 04020034 (2020)
36. Michelmore, R., Wicker, M., Laurenti, L., Cardelli, L., Gal, Y., Kwiatkowska, M.: Uncertainty quantification with statistical guarantees in end-to-end autonomous

driving control. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 7344–7350 (2020). https://doi.org/10.1109/ICRA40945.2020.9196844
37. Park, N., Kim, S.: Blurs behave like ensembles: Spatial smoothings to improve accuracy, uncertainty, and robustness. In: International Conference on Machine Learning. pp. 17390–17419. PMLR (2022)
38. Saeedan, F., Weber, N., Goesele, M., Roth, S.: Detail-preserving pooling in deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9108–9116 (2018)
39. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 31 (2018)
40. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multi-domain operations applications. vol. 11006, pp. 369–386. SPIE (2019)
41. Sobel, I., Feldman, G., et al.: A 3x3 isotropic gradient operator for image processing. a talk at the Stanford Artificial Project in **1968**, 271–272 (1968)
42. Song, Z., Yang, L., Wen, F., Li, J.: Traf-align: Trajectory-aware feature alignment for asynchronous multi-agent perception. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12048–12057 (2025)
43. Su, S., Han, S., Li, Y., Zhang, Z., Feng, C., Ding, C., Miao, F.: Collaborative multi-object tracking with conformal uncertainty propagation. IEEE Robotics and Automation Letters **9**(4), 3323–3330 (2024)
44. Vadivelu, N., Ren, M., Tu, J., Wang, J., Urtasun, R.: Learning to communicate and correct pose errors. In: Conference on Robot Learning. pp. 1195–1210. PMLR (2021)
45. Wang, R., Gao, X., Xiang, H., Xu, R., Tu, Z.: Cocmt: Communication-efficient cross-modal transformer for collaborative perception. arXiv preprint arXiv:2503.13504 (2025)
46. Wang, T.H., Manivasagam, S., Liang, M., Yang, B., Zeng, W., Urtasun, R.: V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 605–621. Springer (2020)
47. Wei, S., Wei, Y., Hu, Y., Lu, Y., Zhong, Y., Chen, S., Zhang, Y.: Asynchrony-robust collaborative perception via bird's eye view flow. Advances in Neural Information Processing Systems **36**, 28462–28477 (2023)
48. Xiang, H., Xu, R., Ma, J.: Hm-vit: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 284–295 (2023)
49. Xu, J., Zhang, Y., Cai, Z., Huang, D.: Cosdh: communication-efficient collaborative perception via supply-demand awareness and intermediate-late hybridization. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6834–6843 (2025)
50. Xu, R., Tu, Z., Xiang, H., Shao, W., Zhou, B., Ma, J.: Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers (2022)
51. Xu, R., Xia, X., Li, J., Li, H., Zhang, S., Tu, Z., Meng, Z., Xiang, H., Dong, X., Song, R., et al.: V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13712–13722 (2023)
52. Xu, R., Xiang, H., Tu, Z., Xia, X., Yang, M.H., Ma, J.: V2x-vit: Vehicle-to-everything cooperative perception with vision transformer (2022)

53. 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: 2022 International Conference on Robotics and Automation (ICRA). pp. 2583–2589. IEEE (2022)
54. 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. Advances in Neural Information Processing Systems **36**, 25151–25164 (2023)
55. Yang, K., Bu, T., Li, L., Li, C., Wang, Y., Li, D.: Is discretization fusion all you need for collaborative perception? In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 9590–9596 (2025). https://doi.org/10.1109/ICRA55743.2025.11128776
56. Yin, H., Tian, D., Lin, C., Duan, X., Zhou, J., Zhao, D., Cao, D.: V2vformer++: Multi-modal vehicle-to-vehicle cooperative perception via global-local transformer. IEEE Transactions on Intelligent Transportation Systems **25**(2), 2153–2166 (2023)
57. Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., et al.: Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21361–21370 (2022)
58. Yu, H., Yang, W., Ruan, H., Yang, Z., Tang, Y., Gao, X., Hao, X., Shi, Y., Pan, Y., Sun, N., et al.: V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5486–5495 (2023)
59. Yuan, Y., Xia, Y., Cremers, D., Sester, M.: Sparsealign: A fully sparse framework for cooperative object detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22296–22305 (2025)
60. Zhao, B., Zhang, W., Zou, Z.: Bm2cp: Efficient collaborative perception with lidar-camera modalities. arXiv preprint arXiv:2310.14702 (2023)