DeHCP: A Decoupled Framework for Scalable Open Heterogeneous Collaborative Perception

SUN Yuan[1], *CAO Jurui*[1], *LIU Yuying*[1], *ZHANG Dong*[1], *LI Zuoyong*[2], *DU Shaoyi*[1*]

1. State Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, P. R. China; 2. Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, School of Computer and Data Science, Minjiang University, Fuzhou 350121, P. R. China

(Received 22 May 2026; revised 16 June 2026; accepted 1 July 2026)

Abstract: The existing open heterogeneous collaborative perception systems typically align diverse features to a fixed semantic space. This alignment process inevitably leads to information loss and suppresses modality-specific characteristics. To address this limitation, this paper introduces a novel decoupled heterogeneous collaborative perception (DeHCP) framework, which decouples the intermediate feature space into a shared branch and a specific branch. The shared branch extracts a modality-agnostic common representation, while the specific branch preserves modality-specific characteristics conditioned on learnable modality embeddings. A dynamic gate aggregation mechanism adaptively integrates these specific features with the common representation, based on the global semantic context and ego-agent identity. Furthermore, a decoupled supervision strategy with backward alignment enables the integration of unseen heterogeneous agents by updating only local encoders and lightweight adapters, thereby avoiding collective retraining. Extensive experiments on collaborative benchmarks demonstrate that DeHCP achieves improved three-dimensional object detection accuracy while maintaining the extensibility of open heterogeneous collaborative perception through backward alignment. The proposed approach supports the development of highly scalable autonomous driving systems. Code is available at https://github.com/jurui-cloud/DeHCP.

Key words: collaborative perception; heterogeneous agent; sensor fusion; 3D object detection; autonomous vehicles

CLC number: TP391.4 **Document code:** A **Article ID:** 1005-1120(2026)04-0517-12

0 Introduction

Multi-agent collaborative perception has emerged as a promising approach for overcoming the inherent limitations of single-agent perception, including occlusion and restricted sensing range[1-2]. By exchanging intermediate bird's-eye-view (BEV) features via vehicle-to-everything (V2X) communication networks, connected autonomous vehicles can achieve a more comprehensive understanding of complex driving environments[3]. Multi-agent collaboration is not confined to autonomous-driving perception. It has also been extended to UAV swarm decentralized control inspired by starling flocking behavior[4] and UAV swarm task scheduling under limited communication ranges[5]. However, as illustrated in Fig.1, most existing collaborative perception frameworks rely on a strict homogeneous assumption, assuming identical sensor modalities and perception models across all agents. In Fig.1, idealized homogeneous and realistic open heterogeneous scenarios are compared. The existing unified-protocol paradigms force heterogeneous features into a fixed semantic space, inevitably causing information loss. The proposed decoupled heterogeneous collaborative perception (DeHCP) explicitly decouples the feature space into a modality-agnostic shared branch and a modality-specific branch. Unseen agents are efficiently integrated via backward alignment without collective retraining. Finally, a dynamic gate adaptively fuses these branches into a more effective representation.

*Corresponding author, E-mail address: dushaoyi@xjtu.edu.cn.

How to cite this article: SUN Yuan, CAO Jurui, LIU Yuying, et al. DeHCP: A decoupled framework for scalable open heterogeneous collaborative perception[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2026, 43(4): 517-528.

http://dx.doi.org/10.16356/j.1005-1120.2026.04.003

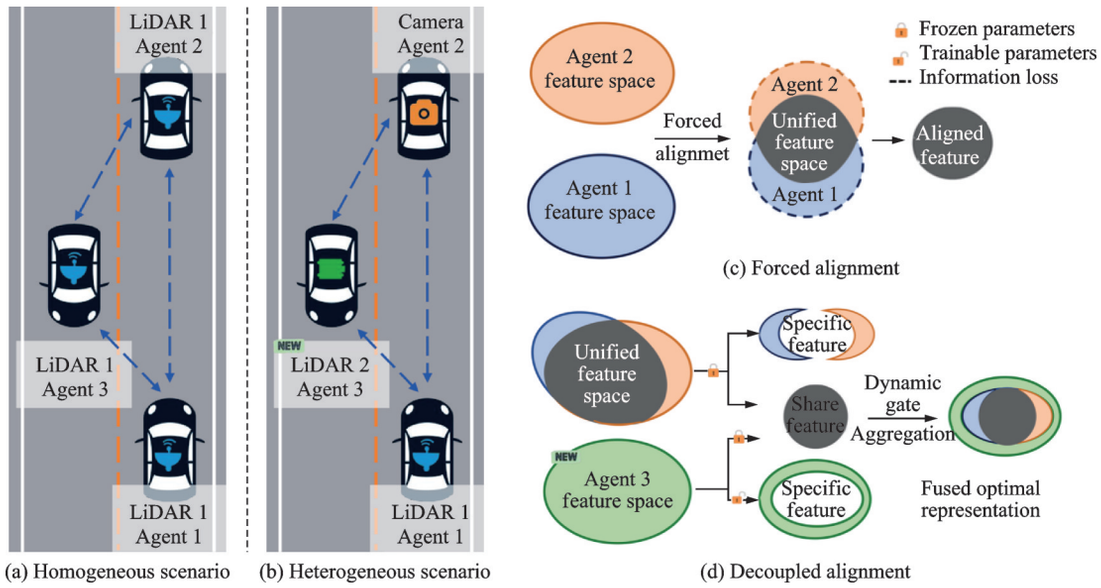


Fig.1 Comparison of feature alignment mechanisms for open heterogeneous collaborative perception

In real-world deployments, however, such a homogeneous assumption is often unrealistic. Vehicles from different manufacturers are typically equipped with diverse model architectures and sensor configurations, such as LiDAR sensors with different numbers of channels and RGB cameras[6]. Moreover, practical systems must dynamically accommodate new and diverse agents that continually join the collaborative effort. Thus, effectively integrating these emerging heterogeneous agents into an existing collaborative system without retraining the entire network remains a critical open challenge.

To address the challenge of open heterogeneous collaboration, several approaches have been proposed[7]. Although collective end-to-end retraining can narrow domain gaps between modalities, it introduces prohibitive computational overhead whenever a new agent type is added. To avoid collective retraining, recent methods, such as STAMP[8], employ lightweight adapter-reverter pairs to transform features between agent-specific domains and a shared protocol domain. Similarly, extensible frameworks like HEAL[9] establish a unified feature space for the initial agents and subsequently integrate new agents through a backward alignment strategy.

Despite these advances in extensibility, these frameworks predominantly align diverse feature representations into a fixed or unified semantic space. This alignment process inevitably leads to information loss by discarding modality-specific characteristics inherent in their original sensors. For instance, the aggregated representations may lose the fine-grained textures captured by cameras or the precise geometric structures provided by LiDAR.

To balance unified feature alignment with the preservation of modality-specific characteristics, we propose DeHCP, as shown in Fig.1. Building upon an extensible architecture, DeHCP explicitly decouples the intermediate feature space into a modality-agnostic shared branch and a modality-specific complementary branch. The shared branch extracts a modality-agnostic common representation, while the specific branch preserves modality-specific characteristics conditioned on learnable modality embeddings. Furthermore, we introduce a dynamic gate aggregation mechanism that adaptively fuses specific features based on the global semantic context and the ego agent's modality identity. To ensure high extensibility, DeHCP adopts a decoupled supervision strategy during the backward alignment stage, allowing each new agent to be trained locally without requiring full-system retraining.

Our main contributions are summarized as follows:

(1) We propose DeHCP, an open heterogeneous collaborative perception framework that mitigates information loss by balancing the unified feature alignment with the preservation of modality-specific characteristics.

(2) We design a decoupled shared-specific ar-

chitecture with a dynamic gate aggregation mechanism to adaptively integrate common representations and modality-specific features.

(3) We conduct extensive experiments on the OPV2V-H dataset, showing that DeHCP achieves competitive performance.

1 Related Work

1.1 Collaborative perception

Collaborative perception enhances scene understanding by enabling multiple agents to share complementary perceptual information, alleviating the limitations caused by occlusion and restricted sensing range. The existing methods are typically categorized into three classes according to their information sharing strategies: Early fusion, late fusion, and intermediate fusion. Early fusion directly shares raw sensor data before feature extraction[10]. Prior studies have shown that exchanging raw LiDAR point clouds across connected agents can effectively extend the sensing range and improve detection performance. However, this strategy demands substantial communication bandwidth, as raw data are costly to transmit[11]. In contrast, late fusion allows each agent to perform perception independently and to share only the final detection results. While this design is highly bandwidth-efficient, the lack of shared spatial context causes information loss, often resulting in suboptimal accuracy.

Intermediate fusion has become the most widely adopted choice, as it offers a better trade-off between communication bandwidth and detection accuracy. Instead of transmitting raw data or final outputs, agents exchange intermediate feature representations that retain critical semantic and spatial information. Existing studies have continuously improved this paradigm by enhancing cross-agent feature interactions[12], introducing adaptive fusion mechanisms[13-14], and reducing communication overhead through selective communication[15-16] and feature compression[17-20]. Despite this progress, most existing methods still rely on homogeneous sensor and model settings, leaving a substantial domain gap when applied to practical systems with heterogeneous agents.

1.2 Heterogeneous collaborative perception

Heterogeneous collaborative perception aims to bridge this domain gap, which arises from diverse sensor modalities and model architectures. To this end, recent research primarily focuses on feature alignment and domain adaptation. A mainstream solution is to transform heterogeneous features into a unified protocol domain using lightweight adapters[9] or graph-based message passing mechanisms[21]. To address modality heterogeneity, techniques such as multi-scale feature distillation and cross-modality dual-attention[22], or length-adaptive fusion modules are frequently employed to capture spatial correlations and maintain semantic consistency.

Beyond static network configurations, practical systems must dynamically accommodate unseen heterogeneous agents. To avoid the reliance on predefined protocol spaces, recent studies explore generative communication mechanisms[23] and dynamic common representation negotiation[24]. Particularly in open heterogeneous scenarios, extensible frameworks like HEAL[9] and fast adaptation strategies using parameter-efficient fine-tuning[25] have been introduced. By utilizing a backward alignment strategy, HEAL enables new agents to map their features into a unified feature space via local encoder updates, significantly reducing the computational cost of retraining the entire fusion network. However, most open heterogeneous frameworks still rely on a unified feature space for compatibility, which may suppress sensor-specific details and complementary modality cues from the original data.

DeHCP differs from existing open heterogeneous collaborative perception methods, such as STAMP and HEAL, primarily in terms of representation design. STAMP[8] focuses on transforming heterogeneous features into a unified feature space through lightweight adaptation modules. HEAL enables open heterogeneous extensibility through backward alignment, allowing newly introduced agents to be incorporated without collective retraining. In contrast, DeHCP adopts a similar extensible training paradigm to HEAL but introduces an explicit decoupling of the collaborative representation into a modality-agnostic shared branch and a modality-specific branch. This design allows the

model to preserve complementary modality-specific cues instead of forcing all heterogeneous information into a single unified feature space. Moreover, DeHCP introduces modality-aware gate aggregation to adaptively control the contribution of modality-specific information to the final fused representation. Therefore, the novelty of DeHCP lies in representation decoupling and adaptive fusion, rather than in redesigning the open-agent training protocol.

2 Methodology

Motivated by this trade-off between feature compatibility and modality-specific information preservation, we propose DeHCP. Following the extensible training paradigm of HEAL[9], DeHCP decouples the intermediate feature space into a modality-agnostic shared branch and a modality-specific complementary branch[26]. The overall framework is illustrated in Fig.2.

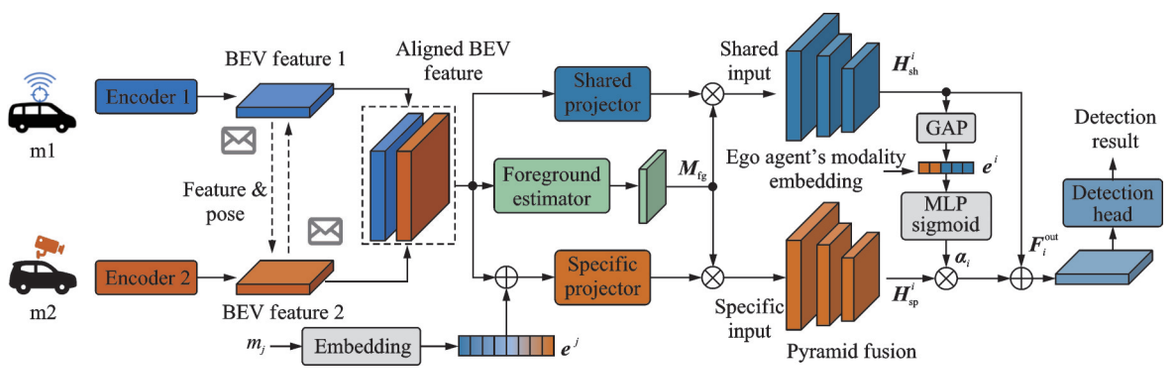


Fig.2 DeHCP framework overview

2.1 Prerequisites

Considering a heterogeneous collaborative perception system consisting of an ego agent i and a set of connected neighbors N_i. Each agent $k \in \{i\} \cup N_i$ is characterized by a specific sensor modality (e.g., LiDAR or camera) and a perception model, jointly denoted by a modality identity $m_k \in M$. Given raw sensor data $\boldsymbol{X}_k$, the local encoder E_{m_k} extracts a BEV feature $\boldsymbol{F}_k = E_{m_k}(\boldsymbol{X}_k) \in \boldsymbol{R}^{C \times H \times W}$.

After sharing these intermediate features via V2X communication, the ego agent i warps the received features $\boldsymbol{F}_j$ into its ego coordinate system using the relative 6-DoF spatial transformation $\Gamma_{j \to i}$, shown as

$$\tilde{\boldsymbol{F}}_{j \to i} = \Gamma_{j \to i}(\boldsymbol{F}_j) \quad \forall j \in N_i \tag{1}$$

The core challenge in open heterogeneous collaboration is effectively fusing the aligned feature set $\tilde{\boldsymbol{F}}_{j \to i}$ despite severe domain gaps arising from distinct modalities and model architectures $m_i \neq m_j$. To bridge these gaps, DeHCP processes the aligned features through a decoupled shared-specific fusion mechanism.

2.2 Modality-aware decoupled feature extraction

To avoid the loss of modality-specific characteristics caused by aligning diverse features into a single protocol space, DeHCP further decomposes this fusion process into a shared branch and a specific branch.

(1) Foreground estimation

Domain gaps in background regions often exacerbate spatial misalignment and localization noise during heterogeneous collaboration. To mitigate this, a lightweight foreground estimator f_{fg} predicts a spatial foreground mask $\boldsymbol{M}_{fg} \in [0,1]^{1 \times H \times W}$ from the aligned features, enabling subsequent networks to focus on perceptually critical objects

$$\boldsymbol{M}_{fg} = f_{fg}(\tilde{\boldsymbol{F}}_{j \to i}) \tag{2}$$

(2) Shared protocol branch

To establish a robust, modality-agnostic common representation, a shared projector P_{sh} processes the aligned features $\tilde{\boldsymbol{F}}_{j \to i}$, which are subsequently weighted by the foreground mask $\boldsymbol{M}_{fg}$ via element-wise multiplication. A shared pyramid fusion network U_{sh} then aggregates these filtered features from all agents to generate the unified common fea-

ture $\boldsymbol{H}_{\mathrm{sh}}^{i}$, shown as

$$\boldsymbol{H}_{\mathrm{sh}}^{i}=U_{\mathrm{sh}}\left(\{P_{\mathrm{sh}}(\tilde{\boldsymbol{F}}_{j\to i})\otimes\boldsymbol{M}_{\mathrm{fg}}\}_{j\in\{i\}\cup N_i}\right) \tag{3}$$

(3) Specific protocol branch

Parallel to the shared branch, the specific branch explicitly preserves modality-specific characteristics of heterogeneous sensors, particularly fine-grained camera-derived textures. A learnable modality embedding maps the discrete modality identity m_j to a continuous dense vector $\boldsymbol{e}^{j}$. This vector is spatially expanded and injected into the aligned features via element-wise addition as a modality prompt. The prompted features are processed by a specific projector P_{sp}, weighted by $\boldsymbol{M}_{\mathrm{fg}}$, and finally fused by a specific pyramid fusion network U_{sp} to extract the modality-specific complementary feature $\boldsymbol{H}_{\mathrm{sp}}^{i}$, shown as

$$\boldsymbol{H}_{\mathrm{sp}}^{i}=U_{\mathrm{sp}}\left(\{P_{\mathrm{sp}}(\tilde{\boldsymbol{F}}_{j\to i}\oplus\boldsymbol{e}^{j})\otimes\boldsymbol{M}_{\mathrm{fg}}\}_{j\in\{i\}\cup N_i}\right) \tag{4}$$

Both U_{sh} and U_{sp} follow the multi-scale foreground-aware pyramid fusion design of HEAL[9]. $\{\cdot\}_{j\in\{i\}\cup N_i}$ denotes the set of aligned features collected from the ego agent and its collaborating neighbors, while the actual multi-agent aggregation is performed inside U_{sh} and U_{sp}. Specifically, at each pyramid scale, the aligned features from different agents are first processed by the corresponding pyramid blocks. The foreground estimator comprises three scale-specific 1×1 convolutional heads, each attached to the shared pyramid features at a different pyramid level. These heads operate on feature maps with input channel dimensions of 64, 128 and 256, respectively, and predict a single-channel confidence map at each level. The resulting confidence maps are then normalized across agents via a softmax operation to obtain fusion weights. The agent features are subsequently aggregated by weighted summation in the ego frame. The shared and specific branches follow the same pyramid fusion design, while maintaining separate learnable parameters.

Collectively, the foreground estimator, together with the shared and specific pyramid fusion networks, constitutes the foreground-aware shared-specific pyramid fusion module, which forms the core feature extraction backbone of DeHCP.

2.3 Gate aggregation

To integrate the modality-agnostic common representation $\boldsymbol{H}_{\mathrm{sh}}^{i}$ and the modality-specific complementary feature $\boldsymbol{H}_{\mathrm{sp}}^{i}$, DeHCP introduces a dynamic gate aggregation mechanism. This enables the ego agent to adaptively evaluate the importance of specific features based on the global semantic context and its modality identity, aggregating beneficial complementary information while suppressing modality-specific noise.

(1) Gate generation

We apply global average pooling to the shared feature $\boldsymbol{H}_{\mathrm{sh}}^{i}$ to extract a channel-wise global semantic descriptor. To incorporate the ego agent's modality prior, this descriptor is concatenated with the modality embedding $\boldsymbol{e}^{i}$. The concatenated vector is then processed by a multi-layer perceptron and a sigmoid activation function σ to generate the channel-wise gate vector $\boldsymbol{\alpha}_i\in(0,1)^{C}$, C is the number of channels in $\boldsymbol{H}_{\mathrm{sh}}^{i}$ and $\boldsymbol{H}_{\mathrm{sp}}^{i}$

$$\boldsymbol{\alpha}_i=\sigma\left(\mathrm{MLP}([\mathrm{GAP}(\boldsymbol{H}_{\mathrm{sh}}^{i});\boldsymbol{e}^{i}])\right) \tag{5}$$

(2) Adaptive feature fusion

The gate vector $\boldsymbol{\alpha}_i$ modulates the modality-specific feature $\boldsymbol{H}_{\mathrm{sp}}^{i}$ via channel-wise multiplication, highlighting informative modality-specific characteristics, particularly fine-grained camera textures. A residual-like fusion strategy then combines the gated specific feature with the common representation to construct the final fused representation $\boldsymbol{F}_{i}^{\mathrm{out}}$, shown as

$$\boldsymbol{F}_{i}^{\mathrm{out}}=\boldsymbol{H}_{\mathrm{sh}}^{i}\oplus(\boldsymbol{\alpha}_i\otimes\boldsymbol{H}_{\mathrm{sp}}^{i}) \tag{6}$$

This asymmetric design uses $\boldsymbol{H}_{\mathrm{sh}}^{i}$ as a robust domain-invariant baseline, while $\boldsymbol{H}_{\mathrm{sp}}^{i}$ serves as an adaptive supplement. Finally, $\boldsymbol{F}_{i}^{\mathrm{out}}$ is passed to the detection head for 3D bounding box prediction.

2.4 Decoupled supervision and backward alignment

To integrate continually emerging heterogeneous agents without retraining the entire network, DeHCP adopts a two-stage training paradigm. This strategy ensures high extensibility and stability of the unified feature space.

Stage 1 Collaboration base training. In the initial phase, a set of homogeneous agents is used to

establish the collaboration base. To ensure that the shared branch captures a modality-agnostic common representation while the specific branch preserves complementary details, we propose a decoupled multi-task loss function. The total loss for Stage 1 is formulated as

$$L_{\text{total}} = L_{\text{det}}(F_i^{\text{out}}) + \lambda_{\text{aux}} L_{\text{det}}(H_{\text{sh}}^i) + \lambda_{\text{fg}} L_{\text{fg}} + \lambda_{\text{gate}} L_{\text{gate}} \tag{7}$$

where λ_{aux}, λ_{fg}, and λ_{gate} are the hyperparameters to balance the corresponding loss terms; $L_{\text{det}}(\cdot)$ represents the standard detection loss, combining focal loss for classification and smooth L_1 loss for bounding box regression. An auxiliary detection loss, $L_{\text{det}}(H_{\text{sh}}^i)$, is directly applied to the output of the shared branch. This decoupled supervision enables the shared branch to maintain robust detection capabilities independently, preventing over-reliance on the specific branch. L_{fg} denotes the binary cross-entropy loss for the foreground mask M_{fg}, and $L_{\text{gate}} = \frac{1}{C}\| \alpha_i \|_1$ applies an l_1 sparsity constraint on the gate vector, encouraging only informative modality-specific channels to remain active and suppressing redundant noise.

Stage 2 New agent type training. To support open heterogeneous collaboration, DeHCP accommodates unseen agent types via a backward alignment strategy. When a new modality m_{new} joins the system, we freeze the parameters of the shared branch P_{sh} and U_{sh}, the foreground estimator, the gate aggregation MLP, and the detection head. The training is conducted locally on each new agent of this type, updating only its local encoder $E_{m_{\text{new}}}$, the specific projector P_{sp}, and the newly initialized modality embedding e^{new}, shown as

$$\min_{E_{\text{new}}, P_{\text{sp}}, e^{\text{new}}} L_{\text{det}}(F_i^{\text{out}}) + \lambda_{\text{aux}} L_{\text{det}}(H_{\text{sh}}^i) \tag{8}$$

By fixing the shared branch and the detection back-end, the new encoder is optimized to align its feature distribution with the pre-established unified feature space, thereby supporting the auxiliary detection task. Simultaneously, the trainable specific projector and modality embedding act as lightweight adapters to retain modality-specific characteristics. This decoupled training mechanism minimizes the integration cost by avoiding collective retraining.

3 Experimental Results

3.1 Open heterogeneous settings

(1) OPV2V-H

To evaluate the proposed method in more realistic heterogeneous scenarios, we utilize OPV2V-H[9], an extension of the OPV2V dataset[13] designed specifically for heterogeneous collaborative perception. It provides diverse multi-agent sensor configurations, including 16/32/64-channel LiDARs and cameras in the same driving scenarios.

(2) Sequential integration

Rather than assuming a static network, we simulate a dynamic environment where unseen heterogeneous agents continually join an established system. Following the protocol in HEAL, the collaboration base is initialized with 64-channel LiDAR agents using a PointPillars encoder $L_{\text{P}}^{(64)}$. After establishing the unified feature space, three distinct heterogeneous agent types are sequentially integrated via backward alignment: a camera agent with EfficientNet $C_{\text{E}}^{(384)}$, a 32-channel LiDAR agent with SECOND $L_{\text{S}}^{(32)}$, and a camera agent with ResNet50 $C_{\text{R}}^{(336)}$. This process demonstrates DeHCP's capability to integrate new feature distributions while adaptively preserving modality-specific characteristics.

(3) Implementation details

For a fair comparison, we align our experimental settings with those of prior work[9]. The input data are encoded with a [0.4 m, 0.4 m] grid size, and intermediate features are compressed to 64 channels to reduce bandwidth consumption.

In DeHCP, both the shared and specific branches adopt a three-level multi-scale architecture, with the three levels processed by ResNeXt modules containing 4, 5 and 6 blocks, respectively. In our method, the modality embedding is represented by a 16D learnable vector. The network is optimized with Adam using an initial learning rate of 0.002, which is reduced by a factor of 0.1 at epoch 15. In the Pyramid fusion module, we adopt a three-level pyramid structure with channel dimensions of 64, 128 and 256, where the three levels are processed by ResNeXt modules containing 3, 5, and 8 blocks, respectively.

The loss weights for foreground supervision, auxiliary shared-branch detection, and gate sparsity regularization are set to $\lambda_{\text{fg}}=1.0$, $\lambda_{\text{aux}}=1.0$, and $\lambda_{\text{gate}}=0.05$, respectively. For the foreground supervision loss L_{fg}, the scale-specific weights are set to $\alpha_l=\{0.4, 0.2, 0.1\}$ for $l\in\{1, 2, 3\}$. The model is trained for 25 epochs to establish the collaboration base, followed by another 25 epochs for each subsequent integration stage.

3.2 Quantitative results

To quantitatively evaluate collaborative perception performance, we compare the proposed DeHCP framework with representative baselines on the OPV2V-H dataset. These baselines include No fusion, Late fusion, and intermediate fusion methods such as F-Cooper[27], V2X-ViT[3], CoBEVT[14], HM-ViT[28], and HEAL[9]. We employ average precision (AP) at intersection-over-union (IoU) thresholds of 0.5 and 0.7 as our primary evaluation metrics, denoted as AP0.5 and AP0.7, respectively. Table 1 presents the detection performance as three new heterogeneous agent types are sequentially integrated into the initial LiDAR-based collaboration base.

As shown in Table 1, DeHCP consistently outperforms HEAL across all stages of sequential integration. When the final agent type $C_R^{(336)}$ is added to the system, DeHCP achieves an AP0.5 of 0.899 and an AP0.7 of 0.822, compared with 0.894 and 0.813 for HEAL, respectively. Although the numerical gains are modest in some settings, the improvement trend remains consistent across all sequential integration stages and becomes more evident under the stricter AP0.7 metric, which is more sensitive to localization quality.

Table 1 Quantitative comparison of 3D object detection performance on OPV2V-H dataset

Method	AP0.5			AP0.7		
	$+C_{\text{E}}^{(384)}$	$+L_{\text{S}}^{(32)}$	$+C_{\text{R}}^{(336)}$	$+C_{\text{E}}^{(384)}$	$+L_{\text{S}}^{(32)}$	$+C_{\text{R}}^{(336)}$
No fusion	0.748	0.748	0.748	0.606	0.606	0.606
Late fusion	0.775	0.833	0.834	0.599	0.685	0.685
F-Cooper[27]	0.778	0.742	0.761	0.628	0.517	0.494
DiscoNet[11]	0.798	0.833	0.830	0.653	0.682	0.695
AttFusion[13]	0.796	0.821	0.813	0.635	0.685	0.659
V2X-ViT[3]	0.822	0.888	0.882	0.655	0.765	0.753
CoBEVT[14]	0.822	0.885	0.885	0.671	0.742	0.742
HM-ViT[28]	0.813	0.871	0.876	0.646	0.743	0.755
HEAL[9]	0.826	0.892	0.894	0.726	0.812	0.813
DeHCP(Ours)	**0.833**	**0.895**	**0.899**	**0.733**	**0.817**	**0.822**

DeHCP benefits primarily from the proposed foreground-aware shared-specific pyramid fusion module. Previous extensible frameworks, such as HEAL, typically project heterogeneous features into a single unified feature space. While this design facilitates compatibility, it may suppress valuable modality-specific details, such as rich appearance cues from cameras and precise geometric information from LiDAR. By explicitly decoupling the feature representation, DeHCP enables the shared branch to learn modality-agnostic features for robust spatial alignment, while the specific branch preserves complementary modality-specific details. Furthermore, the gate aggregation mechanism helps prevent modality-specific noise from dominating the fusion process, allowing the ego agent to adaptively select more informative features. This enables the system to effectively integrate additional heterogeneous agents while better leveraging their distinct perceptual strengths.

3.3 Robustness analysis

To further evaluate the robustness of DeHCP in practical collaborative perception scenarios, we conduct curve-based comparisons under two common deployment factors: Pose noise and feature compression. Pose noise reflects the localization un-

certainty during multi-agent feature alignment, while feature compression simulates the communication bandwidth constraint in V2X collaboration. As shown in Fig.3, we report AP0.5 and AP0.7 curves under different pose noise levels and compression ratios to compare the performance stability of different methods.

As the pose noise increases, all methods show performance degradation because inaccurate relative poses disturb BEV feature alignment among collaborative agents. Nevertheless, DeHCP maintains better robustness and shows a slower performance decline compared with representative baselines. Under feature compression, DeHCP also preserves stable detection accuracy, indicating that the proposed shared-specific decoupled representation can retain effective collaborative information even when communication is constrained. These results further demonstrate the robustness of DeHCP in open heterogeneous collaborative perception.

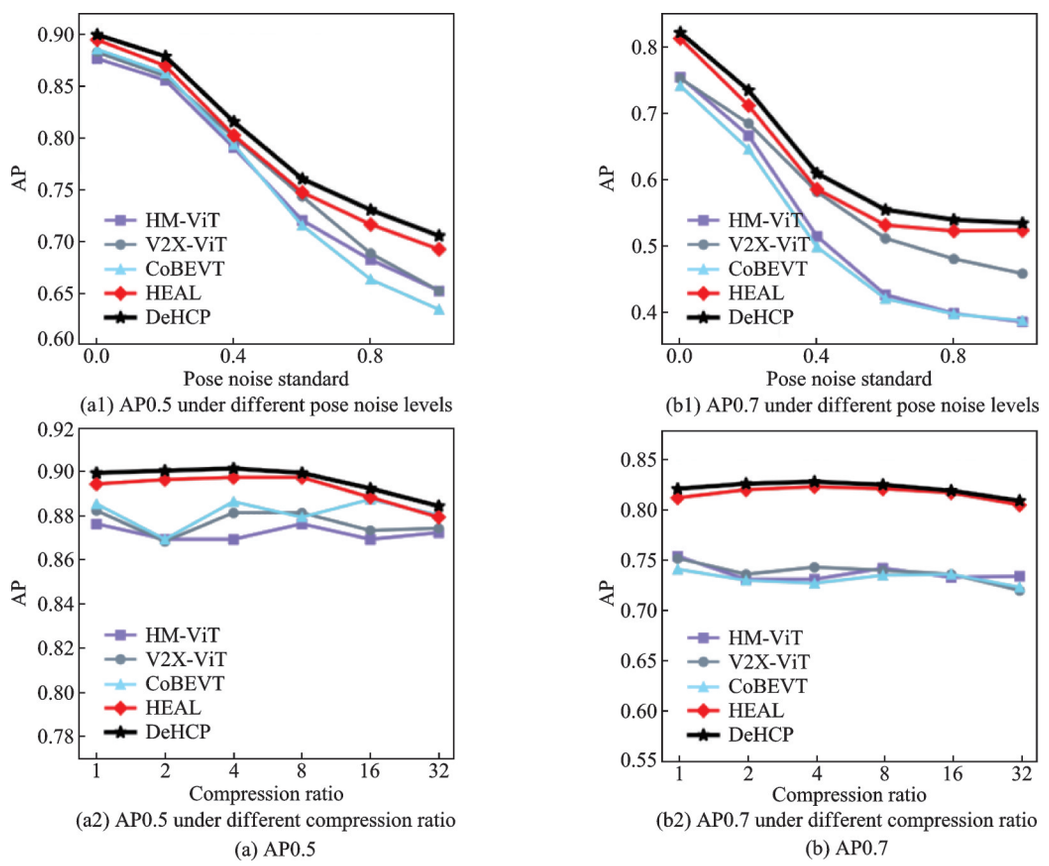


Fig.3 Robustness comparison under pose noise and feature compression on the OPV2V-H dataset

3.4 Ablation study

To quantitatively evaluate the contribution of each core component, we present a component-wise ablation study in Table 2, covering the shared branch, the specific branch, the foreground estimator, and the gate aggregation mechanism. All variants are evaluated under the final open heterogeneous setting, in which $C_{\mathrm{E}}^{(384)}$, $L_{\mathrm{S}}^{(32)}$, and $C_{\mathrm{R}}^{(336)}$ are sequentially integrated into the initial $L_{\mathrm{P}}^{(64)}$ collaboration base. As shown in Table 2, the full DeHCP achieves the best overall performance, demonstrating the effectiveness of the proposed decoupled shared-specific representation design.

Removing the shared branch results in the largest performance drop, with AP0.5 and AP0.7 dropping to 0.843 and 0.762, respectively. This indicates that the shared branch plays a crucial role in maintaining a stable modality-agnostic collaborative representation. Without such a shared representation, heterogeneous features induced by different sensors and model architectures remain difficult to align and aggregate consistently.

Removing the specific branch also leads to performance degradation, with AP0.5/AP0.7 decreasing from 0.899/0.822 to 0.886/0.811. This indicates that preserving modality-specific complementa-

Table 2 Component ablation study of shared branch, specific branch, gate, and foreground mask on OPV2V-H dataset

Variant	Agent types: $L_{\mathrm{P}}^{(64)}+C_{\mathrm{E}}^{(384)}+L_{\mathrm{S}}^{(32)}+C_{\mathrm{R}}^{(336)}$					
	Shared branch	Specific branch	Gate	Foreground mask	AP0.5	AP0.7
w/o shared	×	√	√	√	0.843	0.762
w/o specific	√	×	√	√	0.886	0.811
w/o gate	√	√	×	√	0.886	0.810
w/o FG mask	√	√	√	×	0.884	0.809
Full DeHCP	√	√	√	√	**0.899**	**0.822**

ry information is important for open heterogeneous collaboration. Compared with relying solely on a shared representation, the specific branch helps preserve modality- and architecture-specific information that may be suppressed when heterogeneous features are mapped into a single unified space.

The variant without the dynamic gate achieves an AP0.5 of 0.886 and an AP0.7 of 0.810, demonstrating the necessity of adaptive gated aggregation. Without dynamic modulation, simply merging shared and specific features reduces the model's capability to filter out noisy or redundant modality-specific responses. This result supports the design of using the ego-agent modality and global semantic context to control the injection of specific features.

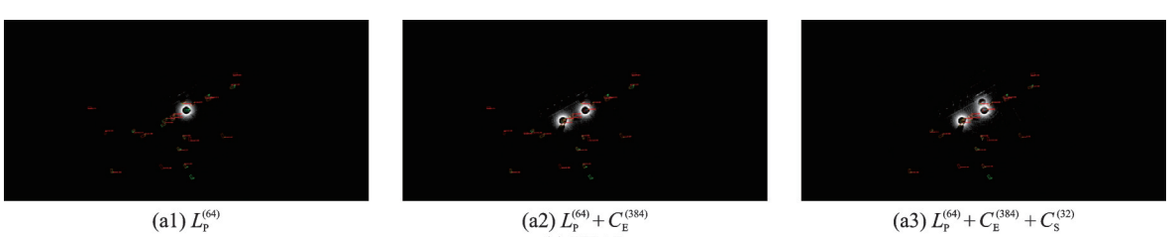
Finally, the variant without the foreground mask achieves an AP0.5 of 0.884 and an AP0.7 of 0.809, indicating a clear performance decline. This performance drop demonstrates that foreground-aware weighting helps the model focus on object-centric foreground regions while mitigating the interference caused by background domain gaps during heterogeneous feature fusion. Overall, these ablation results demonstrate that each component, including the shared branch, specific branch, dynamic gate, and foreground mask, contributes to the overall performance of DeHCP. Their joint design enables more effective and robust open heterogeneous collaborative perception.

3.5 Qualitative analysis

Fig. 4 illustrates the feature alignment process of DeHCP, while Fig. 5 presents its final 3D detection results.

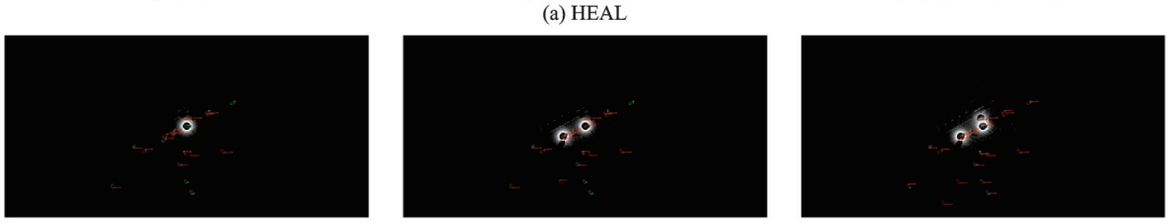
(a) $L_{\mathrm{S}}^{(32)}$ feature space before backward alignment (b) $L_{\mathrm{S}}^{(32)}$ feature space after backward alignment

Fig.4 Visualization of DeHCP backward alignment for the newly introduced $L_{\mathrm{S}}^{(32)}$ agent



(a1) $L_{\mathrm{P}}^{(64)}$ (a2) $L_{\mathrm{P}}^{(64)}+C_{\mathrm{E}}^{(384)}$ (a3) $L_{\mathrm{P}}^{(64)}+C_{\mathrm{E}}^{(384)}+C_{\mathrm{S}}^{(32)}$

(a) HEAL



(b1) $L_{\mathrm{P}}^{(64)}$ (b2) $L_{\mathrm{P}}^{(64)}+C_{\mathrm{E}}^{(384)}$ (b3) $L_{\mathrm{P}}^{(64)}+C_{\mathrm{E}}^{(384)}+C_{\mathrm{S}}^{(32)}$

(b) DeHCP

Fig.5 Visualization of collaborative perception results

Specifically, backward alignment maps the newly introduced $L_S^{(32)}$ agent into the pre-established $L_P^{(64)}$-based shared representation space. Before alignment, as shown in Fig.4(a), the feature distribution of the $L_S^{(32)}$ agent exhibits dispersed activations and clear structural misalignment with the collaboration base due to heterogeneous domain gaps. After backward alignment, as shown in Fig.4(b), the activations become more concentrated and spatially aligned with the modality-agnostic shared representation, enabling the new agent to collaborate effectively with the existing collaboration base.

Fig.5(a) shows the HEAL method, while Fig. 5(b) shows our proposed DeHCP method in the same scenarios. From left to right, we progressively add new agents to demonstrate performance gains. DeHCP outperforms HEAL by generating predictions closer to the ground truth.

As unseen heterogeneous agents are sequentially integrated, DeHCP consistently produces bounding-box predictions that are better aligned with the ground truth than those of the HEAL baseline. These qualitative results suggest that DeHCP can better bridge heterogeneous domain gaps while preserving modality-specific characteristics, which is consistent with the quantitative improvements observed in Table 1. We note that the current evaluation is limited to the heterogeneity range covered by OPV2V-H. Extending the analysis to more extreme modality gaps, such as millimeter-wave radar or event cameras, remains an important direction for future work.

4 Conclusions

This paper introduces DeHCP, a novel framework that mitigates information loss by explicitly decoupling collaborative representations into a modality-agnostic common representation and adaptive modality-specific characteristics. Furthermore, through a two-stage decoupled supervision strategy with backward alignment, the framework efficiently integrates unseen heterogeneous agents without requiring collective retraining. Consequently, DeHCP provides an effective and scalable solution for open heterogeneous collaborative perception, improving 3D detection accuracy and supporting the flexible expansion of collaborative autonomous driving systems. Future work will explore extending DeHCP to more diverse sensor modalities and real-world deployment scenarios.

References

[1] HAN Y, ZHANG H, LI H, et al. Collaborative perception in autonomous driving: Methods, datasets, and challenges[J]. IEEE Intelligent Transportation Systems Magazine, 2023, 15(6): 131-151.

[2] BAI Z, WU G, BARTH M J, et al. A survey and framework of cooperative perception: From heterogeneous singleton to hierarchical cooperation[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(11): 15191-15209.

[3] XU R, XIANG H, TU Z, et al. V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer[C]//Proceedings of the 2022 European Conference on Computer Vision. [S.l.]: Springer, 2022: 107-124.

[4] MEI Y, DUAN H. Decentralized cooperative control of unmanned aerial vehicles based on emergent mechanism of starlings[J]. Transactions of Nanjing University of Aeronautics & Astronautics, 2026, 43(2): 145-157.

[5] ZHENG J, ZHANG S, ZHANG D, et al. Task scheduling for UAV swarms with limited communication range[J]. Transactions of Nanjing University of Aeronautics & Astronautics, 2025, 42(6): 852-864.

[6] MA Y, LU J, CUI C, et al. MACP: Efficient model adaptation for cooperative perception[C]//Proceedings of the 2024 IEEE/CVF Conference on Applications of Computer Vision. [S.l.]: IEEE, 2024: 3361-3370.

[7] XU R, LI J, DONG X, et al. Bridging the domain gap for multi-agent perception[C]//Proceedings of the 2023 IEEE International Conference on Robotics and Automation. [S.l.]: IEEE, 2023: 6035-6042.

[8] GAO X, XU R, LI J, et al. STAMP: Scalable task and model-agnostic collaborative perception[C]//Proceedings of the 2025 International Conference on Learning Representations. [S.l.]:[s.n.], 2025.

[9] LU Y F, HU Y, ZHONG Y Q, et al. An extensible framework for open heterogeneous collaborative perception[C]//Proceedings of the 2024 International Conference on Learning Representations. [S.l.]:[s.n.], 2024.

[10] CHEN Q, TANG S, YANG Q, et al. Cooper: Co-

operative perception for connected autonomous vehicles based on 3D point clouds[C]//Proceedings of the 2019 IEEE International Conference on Distributed Computing Systems. [S.l.]: IEEE, 2019: 514-524.

[11] LI Y M, REN S L, WU P X, et al. Learning distilled collaboration graph for multi-agent perception[C]//Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021). [S.l.]: [s.n.],2021: 29541-29552.

[12] WANG T H, MANIVASAGAM S, LIANG M, et al. V2VNet: Vehicle-to-vehicle communication for joint perception and prediction[C]//Proceedings of the 2020 European Conference on Computer Vision. [S.l.]: Springer, 2020: 605-621.

[13] XU R, XIANG H, XIA X, et al. OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication[C]//Proceedings of the 2022 IEEE International Conference on Robotics and Automation. [S.l.]: IEEE, 2022: 2583-2589.

[14] XU R, TU Z, XIANG H, et al. CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers[C]//Proceedings of the 2023 Conference on Robot Learning. Atlanta, GA: [s.n.], 2023: 989-1000.

[15] HU Y, FANG S, LEI Z, et al. Where2comm: Communication-efficient collaborative perception via spatial confidence maps[C]//Proceedings of Advances in Neural Information Processing Systems (NeurIPS 2022). New Orleans: [s.n.], 2022: 4874-4886.

[16] WANG T, CHEN G, CHEN K, et al. UMC: A unified bandwidth-efficient and multi-resolution based collaborative perception framework[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2023: 8153-8162.

[17] WEI Q, DAI P, LI W, et al. InfoCom: Kilobyte-scale communication-efficient collaborative perception with information bottleneck[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2026: 29731-29739.

[18] DING Z, FU J, LIU S, et al. Point cluster: A compact message unit for communication-efficient collaborative perception[C]//Proceedings of the 2025 International Conference on Learning Representations. Singaore: [s.n.], 2025.

[19] HU Y, PENG J, LIU S, et al. Communication-efficient collaborative perception via information filling with codebook[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 15481-15490.

[20] WANG B, ZHANG L, WANG Z, et al. CORE: Cooperative reconstruction for multi-agent perception [C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2023: 8676-8686.

[21] SHAO C, LUO G, YUAN Q, et al. Hetecooper: Feature collaboration graph for heterogeneous collaborative perception[C]//Proceedings of the 2024 European Conference on Computer Vision. [S.l.]: Springer, 2024: 162-178.

[22] LI Z Y, CHENG G, LIN L L, et al. CD-HCP: A heterogeneous cooperative perception framework utilizing cross-modality dual-attention[J]. IEEE Transactions on Intelligent Transportation Systems, 2026, 27 (6): 6787-6799.

[23] ZHOU J F, DAI P L, WEI Q M, et al. Pragmatic heterogeneous collaborative perception via generative communication mechanism[C]//Proceedings of the Advances in Neural Information Processing Systems, 38 Main Conference (NeurIPS 2025). San Diego: [s.n.],2025.

[24] SHAO C, YUAN Q, LUO G, et al. NegoCollab: A common representation negotiation approach for heterogeneous collaborative perception[C]//Proceedings of the Advances in Neural Information Processing Systems, 38 Main Conference (NeurIPS 2025). San Diego: [s.n.],2025.

[25] WEI Q, DAI P, LI W, et al. CoPEFT: Fast adaptation framework for multi-agent collaborative perception with parameter-efficient fine-tuning[C]//Proceedings of the 2025 AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania: AAAI Press, 2025.

[26] YANG K, YANG D, ZHANG J, et al. What2comm: Towards communication-efficient collaborative perception via feature decoupling[C]//Proceedings of the 31st ACM International Conference on Multimedia. [S.l.]: ACM, 2023.

[27] CHEN Q, MA X, TANG S, et al. F-Cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds [C]//Proceedings of the 4th ACM/IEEE Symposium on Edge Computing. [S.l.]: ACM, 2019.

[28] XIANG H, XU R, MA J. HM-ViT: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2023.

Acknowledgements This work was supported by the National Key Research and Development Program of China (No.2025ZD0124204), the National Natural Science Foundation of China (No.U24A20252), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No.JYB2025XDXM504), the Fuzhou Artificial Intelligence "List-Bidding & Task-Commissioning" Project (No.2025-ZD-038), the Natural Science Basic Research Plan in Shaanxi Province of China (No. 2026JC-YBQN-0779), and the Fundamental Research Fund for the Central Universities (No.xzy012026035).

Authors

The first author Mr. **SUN Yuan** received the B.E. and M.S. degrees from College of Mechanical Engineering, Xi'an University of Technology, China. He is currently pursuing a Ph.D. degree in artificial intelligence at Xi'an Jiaotong University, Xi'an. His research interests include collaborative perception, multi-modal fusion, and object detection.

The corresponding author Prof. **DU Shaoyi** received the double B.S. degrees in computational mathematics and in computer science in 2002, the M.S. degree in applied mathematics in 2005 and the Ph.D. degree in pattern recognition and intelligence system from Xi'an Jiaotong University, China, in 2009. He is a professor with Xi'an Jiaotong University. His research interests include computer vision, machine learning and pattern recognition.

Author contributions Mr. **SUN Yuan** conceived the method, conducted the experiments, and drafted the manuscript. Mr. **CAO Jurui** contributed to method design and implementation. Dr. **LIU Yuying** assisted with experiment setup, data analysis, and result interpretation. Dr. **ZHANG Dong** participated in model development and experimental validation. Prof. **LI Zuoyong** contributed to technical discussion. Prof. **DU Shaoyi** supervised the project and provided overall guidance. All authors reviewed and approved the final manuscript.

Competing interests The authors declare no competing interests.

(Production Editor: SUN Jing)

DeHCP: 面向可扩展开放异构协同感知的解耦框架

孙 远[1], 曹俱睿[1], 刘宇颖[1], 张 栋[1], 李佐勇[2], 杜少毅[1]

(1.西安交通大学人工智能与机器人研究所人机混合增强智能全国重点实验室,西安 710049, 中国;
2.闽江大学计算机与大数据学院福建省信息处理与智能控制重点实验室,福州 350121, 中国)

摘要:现有开放式异构协同感知方法大多采用统一映射策略,将来自不同传感器和模型架构的特征投影至同一语义空间。然而,该过程不可避免地引起信息损失,并削弱各模态所具有的特有细节。针对这一问题,本文提出一种解耦式异构协同感知(Decoupled heterogeneous collaborative perception, DeHCP)框架,将中间特征空间划分为共享分支和特定分支:共享分支用于提取与模态无关的公共表征,特定分支则在可学习模态嵌入的引导下保留模态特有信息。进一步地,本文设计动态门控聚合机制,依据全局语义上下文和自车模态身份,自适应融合模态特定特征与公共表征。同时,提出一种结合反向对齐的解耦监督训练策略,使新加入的异构智能体仅需更新本地编码器和轻量化适配模块即可接入既有协同系统,无需对整个协同网络进行联合重训练。在协同感知基准数据集上的大量实验表明,DeHCP在保持开放式异构协同感知良好可扩展性的同时,进一步提升了三维目标检测精度,为构建具备跨模态兼容能力的高可扩展自动驾驶协同系统提供了有效方案。代码已开源:https://github.com/jurui-cloud/DeHCP。

关键词:协同感知;异构智能体;传感器融合;三维目标检测;自动驾驶车辆

研究亮点:

(1) 提出面向开放异构协同感知的DeHCP解耦框架,将协同特征显式划分为模态无关的共享表征和模态特定的互补表征,在保证跨模态兼容性的同时减少因统一特征空间造成的特有信息损失。

(2) 设计动态门控聚合机制,联合全局语义信息与自车模态先验,自适应调节模态特定特征的通道贡献,从而增强有效互补信息并抑制冗余或噪声响应。

(3) DeHCP在连续引入异构智能体的开放场景下能够稳定提升三维目标检测性能,并兼顾异构协同感知的精度与可扩展性。