HMS-SCP: Task-Oriented Multi-Scale Semantic Communication for V2X Cooperative Perception

Chun-Yeow Yeoh, *Senior Member, IEEE,* Chee Keong Tan, *Senior-Member, IEEE,* Joanne Mun-Yee Lim, *Senior-Member, IEEE,* and Heng-Siong Lim, *Senior Member, IEEE,*

***Abstract*—Cooperative perception enables vehicles and infrastructure to exchange sensor data via Vehicle-to-Everything (V2X) communication, extending sensing coverage beyond occlusions and mitigating blind spots. While critical for autonomous driving and safety, practical deployments often rely on bandwidth-efficient late fusion. Recently, intermediate fusion has emerged as a promising approach for an optimal bandwidth-accuracy trade-off. However, in dense urban environments, cumulative bandwidth demands can overwhelm network capacity, potentially compromising safety-critical Cooperative Intelligent Transport Systems (C-ITS) functions. To alleviate these problems, this paper proposes Hierarchical Multi-Scale Semantic-Aware Cooperative Perception (HMS-SCP), a robust noise-resilient and bandwidth-efficient framework for task-oriented semantic communication in cooperative perception. HMS-SCP employs a spatial importance predictor to identify task-relevant grid elements at each scale, which are then directly mapped into complex-valued symbols for Joint Source-Channel Coding (JSCC). Unlike prior methods that rely on high-dimensional symbol projections for robustness, HMS-SCP exploits structural semantic redundancy across multiple scales to enhance resilience against channel noise, while maintaining an ultra-low symbol rate. This design significantly reduces bandwidth consumption and mitigates network congestion in high-density vehicular environments. Extensive evaluations on the simulated OPV2V and real-world DAIR-V2X datasets demonstrate that HMS-SCP effectively prevents performance collapse under severe Rayleigh fading and extreme compression ratio, maintaining high-confidence far-field detection with a real-time latency of below 16 ms, well within the safety-critical thresholds for dynamic V2X environments.**

***Index Terms*—Cooperative perception, semantic communication, 3D object detection, Vehicle-to-Everything (V2X) safety, autonomous driving.**

I. Introduction

ROAD traffic accidents claim approximately 1.19 million lives globally each year, making them the leading cause of death among children and young adults aged 5 to 29 years [1], according to World Health Organization (WHO). Addressing this urgent global transportation safety challenge has long motivated the International Telecommunication Union (ITU) to promote Intelligent Transport Systems (ITS) [2]. ITS have evolved from basic traffic monitoring using induction loops to situational roadway analysis and interactive driver assistance. Beyond these driver assistance functions, the next generation of ITS places greater emphasis on cooperative perception, prediction, decision making, and control through V2X communication to enable autonomous driving. Conventional vehicular perception remains fundamentally constrained by their dependence on ego-centric sensor suites, including cameras, millimeter-wave (mmWave) radar, and Light Detection and Ranging (LiDAR). Although LiDAR has become a cornerstone technology for high-precision three-dimensional (3D) localization and environmental perception, single-vehicle sensing systems remain inherently vulnerable to critical performance degradation in complex driving environments. These limitations arise primarily from line-of-sight (LoS) occlusions, diminishing point-cloud density at extended sensing ranges, and the inability to perceive objects beyond physical obstructions. To address these challenges, recent advances in V2X communication, aligned with the C-ITS paradigm [3], have enabled the integration of heterogeneous sensor observations from distributed agents and roadside infrastructure to enhance sensing robustness, mitigates occlusion-induced failures, and extends the effective perception horizon beyond the physical constraints of individual vehicles [4], [5].

State-of-the-art (SOTA) V2X technologies primarily include

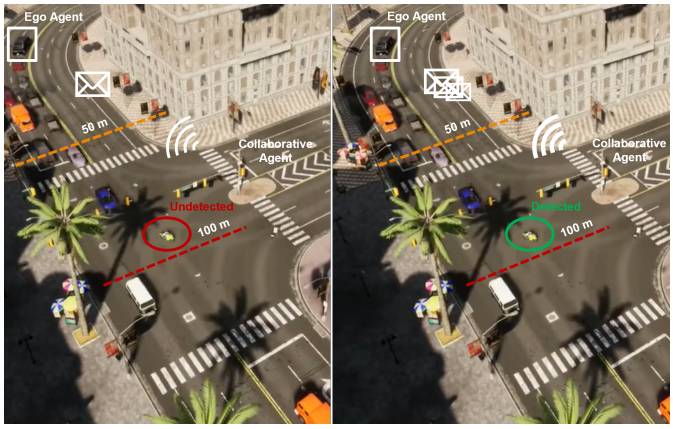


Fig. 1. Cooperative perception performance at extreme compression ratio ($\eta = 0.01$) and severe Rayleigh channel noise ($SNR = 0$ dB). The left panel illustrates the single-scale baseline which fails to resolve distant hazards beyond the 50 m mark. The right panel demonstrates our proposed hierarchical HMS-SCP framework, which successfully detects the 100 m hazard by preserving hierarchical features. This structural resilience effectively extends the ITS safety horizon, ensuring reliable perception under V2X communication constraints.

Chun-Yeow Yeoh and Chee Keong Tan are with the School of Information Technology, Monash University Malaysia, Malaysia (e-mail: chun-yeow.yeoh@monash.edu; tan.cheekeong@monash.edu).

Joanne Mun-Yee Lim is with the Department of Electrical and Robotics Engineering, School of Engineering, Monash University Malaysia, Malaysia (e-mail: Joanne.Lim@monash.edu).

Heng Siong Lim is with the Faculty of Engineering and Technology, Multimedia University, Malaysia (e-mail: hslim@mmu.edu.my).

ETSI ITS-G5 and Cellular V2X (C-V2X). ETSI ITS-G5, based on IEEE 802.11p, enables low-latency ad hoc vehicular communication in the 5.9 GHz ITS spectrum. In practice, it typically operates over 10 MHz channels to achieve nominal physical-layer data rates up to 6 Mbps under favorable link conditions [6]. In parallel, C-V2X that standardized under 3GPP Rel. 14/16, extends vehicular connectivity through direct sidelink (PC5) and network-assisted modes [7]. Supporting flexible 10–20 MHz bandwidths, it achieves throughput of 6–12 Mbps depending on channel conditions, modulation, and resource allocation. These communication constraints impose a critical bottleneck for cooperative perception, where multiple agents must exchange perception information in real time under highly dynamic traffic conditions. Consequently, recent studies [8] have focused on reducing transmission overhead through spatial sparsification, feature compression, and channel-efficient intermediate fusion strategies [9] to better align cooperative perception workloads with practical V2X bandwidth limitations. However, most existing frameworks assume idealized or quasi-static communication links, overlooking perception degradation from physical-layer noise, multipath fading, and packet corruption. This disconnect between algorithmic design and realistic wireless channel behavior remains a primary barrier to real-world deployment [10]–[12].

To bridge this gap, recent advances in semantic communication have progressively transformed cooperative perception from raw data exchange to task-oriented semantic transmission [13], offering a more communication-efficient and channel-resilient solution for vehicular networks affected by severe and time-varying wireless impairments [14]–[16]. Rather than conveying dense feature tensors, existing approaches prioritize importance-aware feature selection and the exchange of compact intermediate representations, such as learned latent features, object-level descriptors, or scene semantics, that are most relevant to downstream perception and cooperative fusion tasks. In this context, a deep Joint Source-Channel Coding (JSCC)-based architecture has emerged as a promising framework to improve robustness under bandwidth constraints and low signal-to-noise ratio (SNR) conditions. By jointly optimizing source compression and channel transmission in an end-to-end manner, JSCC-based systems exhibit graceful performance degradation as channel quality deteriorates, in sharp contrast to the abrupt cliff effect commonly observed in conventional separate source channel coding [17]–[19].

However, based on our observation, the existing semantic encoding process is generally restricted to a single-scale feature map, as exemplified by recent studies [14], [15], [20]. As shown in Fig. 1, this design introduces a fundamental limitation to the perception performance, particularly for distant actors that appear at smaller scales in the feature hierarchy [21]–[23]. This results in a shorter safety horizon, which is particularly dangerous at high speeds. In an ITS context, safe navigation at 100 km/h requires a perception horizon beyond 250 m [5] to maintain safe reaction times of approximately 6 to 9 seconds. Ego-centric perception becomes fundamentally insufficient at higher speeds, suffering a precipitous performance collapse beyond a 50 m threshold. To bridge this critical 50-250 m safety gap within stringent 6–12 Mbps V2X bandwidth constraints, we propose the collaborative HMS-SCP framework. By extracting and jointly optimizing hierarchical semantic representations, our approach replaces traditional channel coding redundancy [15] with multi-scale semantic redundancy. This paradigm shift improves transmission efficiency and noise resilience, achieving the optimal bandwidth-accuracy trade-off necessary for high-speed autonomous navigation.

The main contributions of this work are as follows:

[1] We pioneer a hierarchical multi-scale semantic communication paradigm for cooperative perception, moving beyond the conventional single-scale transmission framework. By jointly exploiting three complementary feature levels through deep JSCC, the proposed architecture preserves both fine-grained spatial details and high-level contextual semantics, substantially extending the reliable perception horizon under severe occlusion and wireless impairments.

[2] We introduce a fundamentally new structural-redundancy principle for channel robustness to replace the brute-force high-dimensional symbol expansion adopted in prior semantic communication methods. Instead of relying on excessive symbol repetition, our framework leverages cross-scale semantic complementarity to achieve strong resilience against Rayleigh fading and low-SNR conditions while maintaining ultra-low transmission overhead.

[3] We develop an importance-aware rate-controlled semantic transmission mechanism that dynamically selects the most task-critical spatial features across multiple scales and maps each selected location into only a single complex-valued symbol. This enables an ultra-efficient transmission ratio of $R = 0.01$, delivering up to 256× higher symbol efficiency than SOTA baselines without sacrificing perception accuracy.

[4] We establish a practical modular training strategy for next-generation V2X perception systems, allowing the proposed semantic communication layers to be seamlessly integrated with pretrained cooperative perception backbones and detection heads without full retraining. This significantly reduces the complexity of the deployment while preserving pretrained perception capability.

[5] We provide the comprehensive evidence that hierarchical semantic communication can simultaneously achieve near-upper-bound perception accuracy, real-time latency, and extreme bandwidth efficiency under adverse channel conditions. Extensive experiments on the simulated OPV2V and real-world DAIR-V2X benchmarks demonstrate superior robustness under Additive White Gaussian Noise (AWGN) and Rayleigh fading, with particularly large gains in the safety-critical range (50–100m) where existing methods collapse.

II. Related Works

A. V2X based Cooperative Perception

While single-vehicle Bird's-Eye-View (BEV) perception provides a convenient unified representation for downstream detection, these ego-centric models [24], [25] remain fundamentally constrained by LoS occlusions, sparse long-range

sensing, and viewpoint limitations. To overcome these deficiencies, multi-agent cooperative perception frameworks leveraging Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communications have emerge as a robust alternative, enabling the aggregation of complementary spatial observations across distributed agents. By fusing sensor data from across-agents, these systems significantly mitigate the unreliability of individual sensors in challenging environments. From a system design perspective, cooperative perception frameworks can be broadly categorized into early, intermediate, and late fusion [26], [27]. Early fusion provides a theoretical performance upper bound through raw data exchange but remains generally impractical due to extreme bandwidth demands. Conversely, late fusion minimizes network load by sharing only final outputs, such as bounding boxes, though it often lacks the semantic depth required to resolve complex occlusions.

More recently, intermediate fusion for cooperative perception has emerged as a dominant paradigm due to its optimal trade-off between perception accuracy and communication efficiency [9]. Hu et al. proposed Where2comm [11], an intermediate fusion framework that introduces a spatial confidence map to selectively transmit perceptually critical regions, significantly reducing communication bandwidth while maintaining high perception performance. Building on the principle of selective communication, Yang et al. proposed How2comm [28], which employs a mutual-information-aware communication mechanism with spatial–channel feature filtering to sparsify transmitted features, enabling agents to transmit only informative perception features. In a different direction, Hu et al. proposed CodeFilling [29], which reduces communication bandwidth by encoding perception features using a shared vector-quantized codebook and transmitting compact code indices for informative tokens, thereby enabling highly compressed communication for cooperative perception. Subsequently, Tao et al. [30] proposed enabling an ego vehicle to proactively prioritize specific directions of interest and intelligently reallocate limited bandwidth to those vital areas using a direction-aware selective attention mechanism to enhance local directional perception.

Despite the success of intermediate fusion in reducing bandwidth [10]–[12], [28], [29], [31], [32], existing frameworks remain fundamentally limited in three key aspects. First, they largely assume ideal or quasi-static communication channels, making them vulnerable to performance degradation under realistic wireless impairments such as fading, noise, and packet corruption. Second, they treat feature transmission primarily as a spatial pruning or compression problem, which reduces bandwidth but still transmits raw feature bits that may contain high-frequency noise or task-irrelevant background. Third, current designs lack a principled mechanism for hierarchical bandwidth allocation. In most cases, all transmitted semantic components are treated uniformly, without distinguishing between features that contribute to the global structural context required for scene stability and the high-resolution details necessary for safety-critical long-range detection. This uniform allocation fails to exploit the multi-scale importance distribution inherent in BEV representations, where different feature levels encode complementary spatial and semantic information. Consequently, these methods fail to fully exploit the intrinsic hierarchy of BEV representations to transmit only the minimal sufficient representation and lack robustness in safety-critical scenarios.

Semantic communication has been recently introduced into cooperative perception systems [14], [15], [20] to improve communication efficiency by prioritizing the transmission of task-relevant information. For instances, Sheng et al. [14] proposed a semantic cooperative perception framework based on JSCC and an importance map [11] that extracts critical semantic features from LiDAR feature maps before transmission. By selectively transmitting informative semantic components, the framework significantly reduces communication overhead while maintaining perception accuracy under various wireless channel conditions. Based on this idea, Gan et al. [15] introduced SComCP, a task-oriented semantic communication framework designed specifically for cooperative perception. In SComCP, an importance-aware feature selection network identifies and prioritizes semantic features that contribute most to the perception objective, enabling selective transmission among cooperating agents. In addition, a JSCC-based semantic codec [33] is employed to directly map semantic features into noise-resistant channel symbols, thereby improving robustness against channel impairments and facilitating reliable communication over noisy and dynamic wireless environments.

Despite these advances, existing semantic communication-based cooperative perception frameworks still face several limitations. First, most current architectures are inherently single-scale, operating on a unified singular feature representation extracted from a specific stage of the perception backbone. Although existing models [14], [15] effectively prioritize task-relevant regions within this representation, they overlook the hierarchical nature of modern perception networks. This single-scale constraint is fundamentally insufficient for the long-range perception required in high-speed ITS scenarios, as distant actors often manifest as small-scale features that are lost or poorly resolved in a unified feature map, degrading long-range perception reliability. Second, many existing frameworks [15] often rely on brute-force symbol redundancy to ensure channel robustness. For instance, projecting a minimal set of semantic features into high-dimensional symbol spaces (e.g., 256 symbols per dimension) inherently limits the volume of transmittable spatial information. This creates a sparsity bottleneck where the system may effectively transmit 12 isolated points, but remains blind to the broader semantic context of the environment, leading to a loss of global scene structure and contextual continuity.

III. Methodology

A. Semantic Communication based Cooperative Perception

In this section, the proposed HMS-SCP framework is presented for robust and communication-efficient cooperative perception under practical V2X constraints. First, the task objective is formulated, followed by an overview of the HMS-SCP architecture, as illustrated in Fig. 2. In this architecture, multi-scale backbone features are selectively encoded and transmitted through a semantic communication pipeline. In

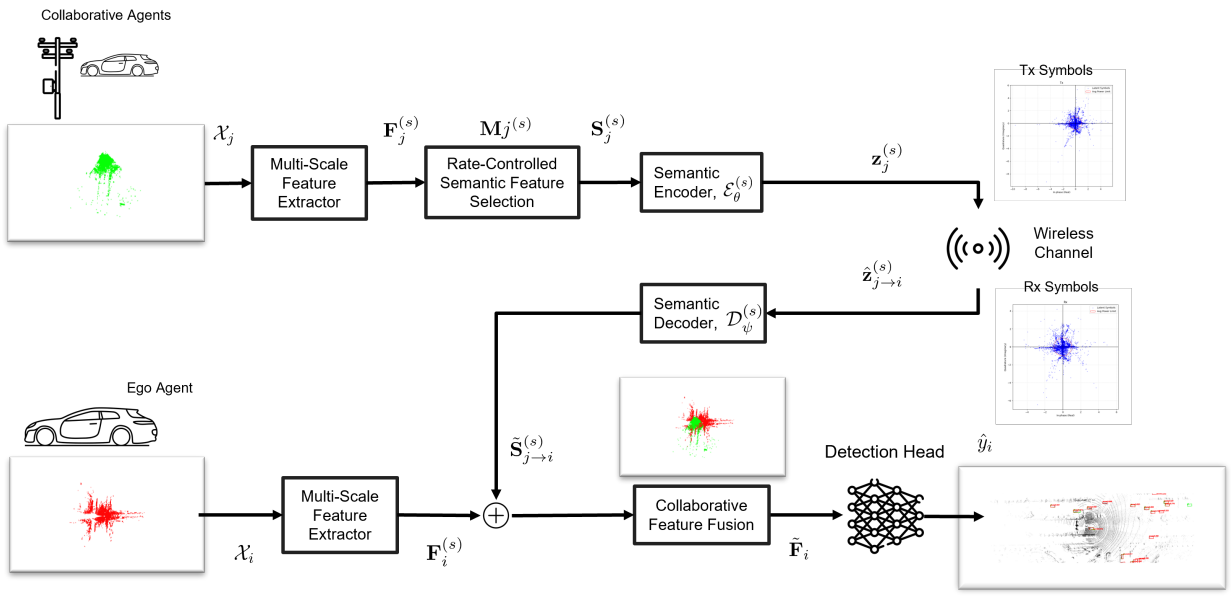


Fig. 2. Overview of Hierarchical Multi-Scale Semantic-Aware Cooperative Perception (HMS-SCP) architecture

particular, the proposed method introduces a rate-controlled semantic feature selection mechanism that identifies task-critical spatial regions across different scales. These selected features are then processed by a JSCC-based semantic codec, which directly maps them into channel symbols to ensure robustness against wireless fading and noise. Next, the noise-resilient architectural design is discussed with particular emphasis on its semantic codec and the integration with rate-controlled semantic feature selection. Finally, a task-oriented training strategy is presented to align semantic transmission with perception performance, balancing communication efficiency and detection accuracy.

B. Cooperative Perception Problem Formulation

Cooperative perception with N agents is considered which may include connected automated vehicles (CAVs) and intelligent infrastructure units under bandwidth and wireless channel constraints. The ego-agent i observes the surrounding environment and cooperates with neighboring agents $j \in \{1, \ldots, N\}, j \neq i$. to perform a perception task. Let X_i and y_i denote the observation and corresponding ground-truth label of ego-agent i, respectively. Rather than transmitting raw observations, each neighboring agent j extracts a set of hierarchical feature maps from its perception backbone. $\{\mathbf{F}_j^{(s)}\}_{s=1}^S$ denote the set of feature maps extracted from S hierarchical levels of the perception backbone.

To satisfy bandwidth constraints, each agent employs a spatial importance predictor $\mathcal{M}^{(s)}(\cdot)$ to identify task-relevant grid elements at each scale. The predictor generates a binary importance mask $\mathbf{M}_j^{(s)} \in \{0,1\}^{H\times W}$, which identifies the top-k most salient spatial indices at scale s. The sparsified semantic representation is then formulated as:

$$\mathbf{S}_j^{(s)} = \mathbf{M}_j^{(s)} \odot \mathbf{F}_j^{(s)}, \tag{1}$$

where $\odot$ denotes the element-wise product with broadcasting across the channel dimension C. The resulting sparse features are then mapped to the latent space via the semantic encoder:

$$\mathbf{z}_j^{(s)} = \mathcal{E}_\theta^{(s)}(\mathbf{S}_j^{(s)}), \tag{2}$$

where $\mathcal{E}_\theta^{(s)}(\cdot)$ denotes the semantic encoder at scale s, and $\mathbf{S}_j^{(s)}$ represents the semantic feature at scale s. The multi-scale representation captures both fine-grained local information and high-level contextual semantics that are important for robust perception. The received signal at agent i is modeled as:

$$\hat{\mathbf{z}}_{j\to i}^{(s)} = \mathcal{W}\left(\mathbf{z}_j^{(s)}; \gamma\right), \tag{3}$$

where $\mathcal{W}(\cdot;\gamma)$ represents the wireless channel transfer function (e.g., AWGN or Rayleigh fading) parameterized by the instantaneous SNR γ. Upon reception, the ego-agent reconstructs the corresponding feature maps using a semantic decoder:

$$\tilde{\mathbf{S}}_{j\to i}^{(s)} = \mathcal{D}_\psi^{(s)}\left(\hat{\mathbf{z}}_{j\to i}^{(s)}\right), \tag{4}$$

The reconstructed features are subsequently fused with the local features of ego-agent through a collaborative perception network $\Psi_\Theta(\cdot)$ to produce the final perception output:

$$\hat{y}_i = \Psi_\Theta\left(\{\mathbf{F}_i^{(s)}\}_{s=1}^S, \{\tilde{\mathbf{S}}_{j\to i}^{(s)}\}_{j\neq i,\, s=1}^S\right), \tag{5}$$

where $\hat{y}_i$ denotes the predicted perception output for agent i, $\{\mathbf{F}_i^{(s)}\}$ represents the ego-agent's local multi-scale features, and $\{\tilde{\mathbf{S}}_{j\to i}^{(s)}\}$ denotes the reconstructed multi-scale features received from collaborating agents.

To ensure that the collaborative perception system remains viable under limited wireless bandwidth, the semantic transmission is required to satisfy a global resource constraint

$$\sum_{j\neq i}\sum_{s=1}^{S} R_{j\to i}^{(s)} \leq B, \tag{6}$$

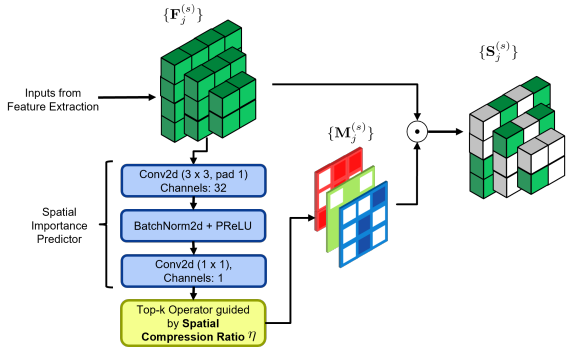


Fig. 3. Rate-controlled semantic feature selection process.

where

$$R_{j\to i}^{(s)} = \frac{n_{j\to i}^{(s)}}{d_j^{(s)}}, \tag{7}$$

denotes the transmission ratio or spatial sampling ratio at scale s. Here, $n_{j\to i}^{(s)}$ corresponds to the number of channel uses, while $d_j^{(s)}$ represents the dimensionality of the semantic representation, specifically the spatial grid size. The transmission ratio R serves as a tunable parameter that enables the framework to dynamically adapt to varying levels of channel congestion characterized by the available bandwidth B.

The system is trained in an end-to-end manner to maximize perception performance under varying wireless channel conditions:

$$\max_{\theta,\psi,\Theta} \sum_{i=1}^{N} \mathbb{E}_{\gamma\sim p(\gamma)} \Big[g\big(\hat{y}_i, y_i\big) \Big], \tag{8}$$

subject to (6) where $g(\cdot)$ denotes the evaluation metric of the perception task, and the expectation over $p(\gamma)$ enforces robustness to stochastic channel variation. In this context, the evaluation is done using average precision (AP), which is a commonly used metric to evaluate 3D object detection performance [34]. The multi-scale nature of the framework further allows for dynamic bit-allocation, where finer scales handling nearby objects and coarser scales providing global context for distant objects are assigned different channel uses $n_{j\to i}^{(s)}$ based on their contribution to the perception metric $g(\cdot)$. This allows the framework to dynamically balance communication efficiency and perception accuracy according to their contribution to the overall task objective $g(\cdot)$.

C. HMS-SCP Framework

The overall architecture of the proposed HMS-SCP framework as illustrated in Fig. 2, consists of five major components: multi-scale feature extraction, rate-controlled semantic feature selection, semantic codec, collaborative feature fusion and object detection. The transmission of ego agent's metadata, including poses and extrinsics, is assumed to be well synchronized, providing the necessary spatial information for all collaborative agents to project their LiDAR point clouds into the coordinate frame of the ego-agent prior to feature extraction. In practical C-ITS deployments, such spatial metadata can be reliably disseminated through standardized protocols such as the Collective Perception Message (CPM) [35], [36] defined in the ETSI Collective Perception Service (CPS), facilitating seamless multi-agent coordination.

Multi-Scale Feature Extraction: The feature extraction module processes LiDAR point cloud data as the primary sensory input for cooperative perception. For each neighboring agent j, the raw point cloud $\mathcal{X}_j$ is first encoded using a Pillar Feature Encoding (PFE) layer [37], which transforms sparse 3D points into a structured pillar-based representation. Through a pillar scattering operation, the encoded features are projected onto a 2D BEV pseudo-image, yielding a spatial representation map $\mathbf{F}_j \in \mathbb{R}^{H\times W\times C}$. The resulting BEV feature map is then processed by a hierarchical convolutional backbone to extract spatial features at multiple resolutions [31]. Specifically, the backbone progressively downsamples the feature maps through a stacked of convolutional layers, producing a set of multi-scale feature representations $\{\mathbf{F}_j^{(1)}, \mathbf{F}_j^{(2)}, \mathbf{F}_j^{(3)}\}$ corresponding to $2\times$, $4\times$, and $8\times$ downsampled resolutions. These hierarchical features provide complementary information across different receptive fields, encoding both fine-grained spatial details and high-level contextual semantics, and serve as inputs to the subsequent rate-controlled semantic feature selection module.

Rate-Controlled Semantic Feature Selection: To satisfy communication constraints, each agent j employs a learnable spatial importance predictor to estimate the contribution of each BEV location to the downstream perception task. For each scale s, the module takes the feature map $\mathbf{F}_j^{(s)}$ as input to generate an importance logit map, which is transformed via a sigmoid activation to produce a probabilistic saliency map:

$$\mathbf{P}_j^{(s)} \in (0,1)^{H_s\times W_s}. \tag{9}$$

This map represents the pixel-wise probability of a spatial location that contributes to the perception task. As shown in Fig. 3, a spatial compression ratio $\eta \in (0,1)$ is then introduced to directly govern the transmission budget. By selecting the top-k spatial indices corresponding to the k largest values in the saliency map $\mathbf{P}_j^{(s)}$, a fixed sparsity constraint is enforced as:

$$k = \lfloor \eta \cdot H_s W_s \rceil. \tag{10}$$

This operation yields a binary selection mask $\mathbf{M}_j^{(s)} \in \{0,1\}^{H_s\times W_s}$, where only the most informative spatial indices corresponding to the k highest-scoring positions in $\mathbf{P}_j^{(s)}$ are set to 1. In the forward pass, the importance predictor $\mathcal{M}^{(s)}(\cdot)$ applies this discrete mask to the feature maps $\mathbf{F}_j^{(s)}$, resulting in the masked multi-scale semantic representation $\mathbf{S}_j^{(s)}$ defined in (1). The physical mapping of this representation using spatial compression ratio η to directly dictate the communication overhead is further elaborated in the subsequent subsection.

Semantic Codec: As shown in Fig. 4, the semantic codec pipeline is designed to compress and reconstruct the masked multi-scale semantic features $\mathbf{S}_j^{(s)} \in \mathbb{R}^{H_s\times W_s\times C_s}$ to enable communication-efficient feature. For each scale $s \in \{1,\dots,S\}$, the feature map is transformed by a scale-specific semantic encoder $\mathcal{E}_\theta^{(s)}$ to produce a compact latent representation $\mathbf{z}_j^{(s)}$ as defined in (2). This scale-specific encoding

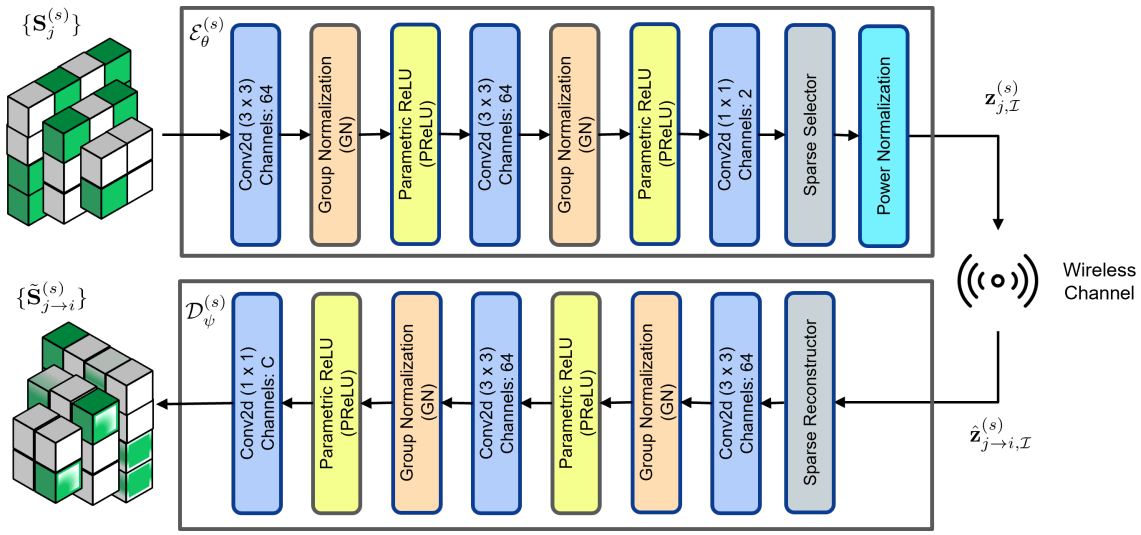


Fig. 4. The architecture of our proposed semantic codec

design allows each hierarchical feature level to be compressed according to its own spatial resolution and semantic characteristics, rather than forcing all features into a unified single-scale representation. The dimension of $\mathbf{z}_j^{(s)}$ determines the number of symbols for transmission. The encoder $\mathcal{E}_\theta^{(s)}$ utilizes a sequential convolutional architecture designed to compress spatial redundancy while preserving critical semantic information. Under JSCC, the encoded latent representation is directly mapped into channel symbols and transmitted over a wireless channel, resulting in the received signal $\hat{\mathbf{z}}_{j\to i}^{(s)}$ as defined in (3). At the receiver, the ego-agent employs a semantic decoder $\mathcal{D}_\psi^{(s)}$ to reconstruct the feature maps $\tilde{\mathbf{S}}_{j\to i}^{(s)}$ as defined in (4). The detailed noise-resilient architectural design of the codec is further discussed in the subsequent subsection.

Collaborative Feature Fusion: As illustrated in Fig. 5, the ego-agent i aggregates the reconstructed multi-scale semantic representations $\{\tilde{\mathbf{S}}_{j\to i}^{(s)}\}$ received from all neighboring agents $j \in \{1,\ldots,N\}, j \neq i$. In contrast to single-scale methods, the fusion process is performed independently for each of the S hierarchical scales, thereby preserving scale-specific spatial and semantic characteristics throughout the aggregation process. For each scale s, a dedicated fusion block is employed to perform the following operations: **Spatial Alignment**, which applies a warp affine transformation derived from pairwise spatial metadata to align neighbor features with the ego agent's BEV coordinate system; and **Attentional Aggregation**, which employs a a Scaled Dot-Product Attention mechanism [38] to fuse the aligned features from all neighboring agents with the ego agent's own feature representation $\mathbf{F}_i^{(s)}$. For each spatial location, the module computes a weighted sum of agent-wise features as follows:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \qquad (11)$$

where attention is performed across the agent dimension. This operation transforms multiple per-agent feature tensors into a unified ego-centric scene representation, enabling the model to emphasize reliable complementary observations while suppressing features corrupted by occlusion, misalignment, or channel-induced distortion.

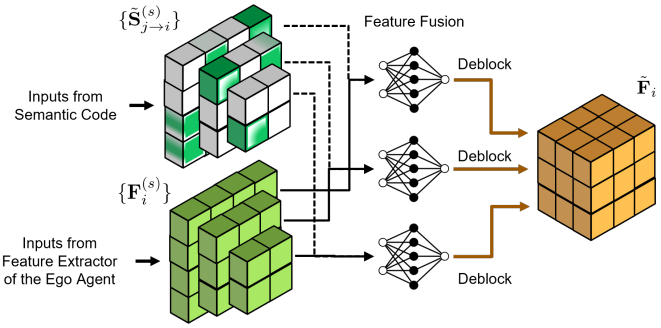


Fig. 5. Collaborative feature fusion process.

Importantly, the spatial dimensions and channel depths of each scale are preserved during this stage. Following per-scale fusion, a multiscale decoding module upsamples each fused representation to a common high-resolution spatial grid. Specifically, the fused features are processed by scale-specific deconvolutional blocks resulting in a unified set of feature maps and then concatenated along the channel dimension. Finally, the channel dimension is further reduced to produce the final comprehensive BEV representation $\tilde{\mathbf{F}}_i \in \mathbb{R}^{H\times W\times C}$, where C represents the channel dimension set to 256 in our implementation for computational efficiency. This fused volume captures a hierarchical range of geometric and semantic information, serving as the input for the downstream 3D object detection head.

Object Detection and Multi-Task Optimization: The detection head is optimized using a composite loss $\mathcal{L}_{\text{det}}$, comprising classification, regression, and orientation components. For classification, the *Focal Loss* [39] formulation is adopted to mitigate the inherent class imbalance between sparse foreground objects and the dominant background regions. The regression branch supervises the 3D bounding box parameters $(x, y, z, w, l, h, \theta)$, where (x, y, z) denotes the center position,

(w, l, h) represents the physical dimensions, and θ indicates the yaw angle. The *Smooth L1 loss* [40] is employed for bounding-box regression, together with an auxiliary direction loss [37] to resolve the 180° orientation ambiguity.

A multi-scale JSCC reconstruction loss is integrated to preserve the semantic fidelity of communicated features under channel impairments. Since the proposed semantic codec processes representations across S hierarchical resolutions, the reconstruction error is calculated independently at each scale by comparing the masked sender feature map before transmission, $\mathbf{S}_j^{(s)}$, with its reconstructed counterpart at the receiver, $\tilde{\mathbf{S}}_{j\to i}^{(s)}$. The overall reconstruction loss is obtained by averaging the per-scale losses:

$$\mathcal{L}_{\text{rec}} = \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}\left[\|\tilde{\mathbf{S}}_{j\to i}^{(s)} - \mathbf{S}_j^{(s)}\|_2^2\right]. \tag{12}$$

The final stage of the pipeline evaluates perception performance by applying a multi-task loss function to the fused BEV feature maps. A joint optimization strategy has been adopted to balance the primary detection objective with semantic reconstruction fidelity across all scales:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}}(\hat{y}_i, y_i) + \mu \cdot \mathcal{L}_{\text{rec}}, \tag{13}$$

where μ is a scaling hyperparameter that controls the contribution of the reconstruction term. By minimizing $\mathcal{L}_{\text{rec}}$ alongside the detection loss $\mathcal{L}_{\text{det}}$, the network learns to prioritize the preservation of task-critical semantic information across all hierarchical levels. This ensures that the multi-scale decoder can faithfully reconstruct high-dimensional features from the sparse symbol stream. This joint optimization enables the detection head to leverage both high-resolution geometric details and global context for robust 3D bounding box prediction, even under stringent bandwidth constraints and imperfect wireless channel conditions.

D. Noise-Resilient Architectural Design

To ensure high-fidelity reconstruction under imperfect channel conditions, a noise-resilient architecture for the semantic codec $(\mathcal{E}_\theta, \mathcal{D}_\psi)$ as illustrated in Fig. 4 has been proposed. The encoder $\mathcal{E}_\theta$ utilizes a sequential multi-scale convolutional design, where two stacked 3×3 are first employed to capture local spatial correlations followed by a 1×1 bottleneck layer that projects features into the latent space. This stacked design enables the encoder to capture richer local spatial dependencies and perform deeper nonlinear feature transformation prior to the latent projection, which helps produce a more robust semantic representation under channel distortion.

To further improve reconstruction stability, Group Normalization (GN) [41] and Parametric ReLU (PReLU) activations [33], [42] is integrated throughout the encoder-decoder pipeline. Compared with Batch Normalization (BN), GN provides stable feature normalization independent of batch size, which is important in decentralized collaborative perception, where the number of communicating neighbors and transmitted features can vary across scenes. Meanwhile, PReLU preserves negative responses via a learnable slope, adaptively reshaping feature distributions and avoiding information loss from hard zero-clipping. This is particularly beneficial in semantic communication, where weak yet informative features aid robust reconstruction under noise and fading. On the decoder, PReLU retains negative-valued signals to enhance recovered BEV feature fidelity for downstream tasks.

To satisfy the physical constraints of the wireless transmitter, the latent representation $\mathbf{z}_j^{(s)}$ is mapped into a complex-valued signal space. Specifically, the output channel dimension is interpreted as a paired set of real and imaginary components, such that

$$\mathbf{z}_j^{(s)} = [\mathbf{z}_{\text{re}}, \mathbf{z}_{\text{im}}], \tag{14}$$

where $\mathbf{z}_{\text{re}}, \mathbf{z}_{\text{im}} \in \mathbb{R}^{H_s\times W_s\times C\text{out}}$, with C_{out} denoting the number of complex-valued symbols assigned to each spatial location. Unlike conventional dense transmission, our framework performs sparse latent transmission. Following the rate-controlled selection mechanism, only the latent vectors at spatial indices $u \in \mathcal{I}_j^{(s)}$, defined by the mask $\mathbf{M}_j^{(s)}$, are transmitted. By setting $C_{out} = 1$, each selected location is mapped to a single complex-valued symbol $\mathbf{z}_{j,u}^{(s)} \in \mathbb{C}$. The resulting serialized sequence $\mathbf{z}_{j,\mathcal{I}}^{(s)} = (\mathbf{z}_{j,u}^{(s)})_{u\in\mathcal{I}_j^{(s)}}$ ensures the transmission volume is exactly $K_s = |\mathcal{I}_j^{(s)}| = k$. Consequently, the total channel uses are governed by $n_j = \sum_{s=1}^{S}|\mathbf{M}_j^{(s)}|_0$, and the normalized transmission ratio simplifies to $R = \eta$. This establishes a direct correspondence between the spatial sparsity budget and actual channel usage, enabling precise and interpretable control over communication overhead.

To regulate the average transmit power to a fixed constraint P, a power normalization operation $\phi(\cdot)$ is applied to the sparse latent sequence prior to channel perturbation. The normalized transmit symbols are expressed as

$$\mathbf{z}_{j\to i,\mathcal{I}}^{(s)} = \mathbf{z}_{j,\mathcal{I}}^{(s)} \cdot \sqrt{\frac{P}{\frac{1}{K_s}\sum_{u\in\mathcal{I}_j^{(s)}}\left\|\mathbf{z}_{j,u}^{(s)}\right\|^2 + \epsilon}}, \tag{15}$$

where ϵ is a small constant introduced for numerical stability. To prevent gradient explosion particularly during the initial stages of training, a scale clipping mechanism that upper-bounds the normalization factor at a maximum value $\lambda_{\max}$ is employed. This prevents the transmission of excessively high-energy symbols when the encoder produces latent representations with small magnitudes, which is particularly important during early training or under highly sparse feature activation. The normalized sparse symbols are then transmitted over a wireless channel. Let $\mathbf{z}_{j\to i,\mathcal{I}}^{(s)}$ denote the symbols received at agent i. The channel input-output relationship is modeled as:

$$\hat{\mathbf{z}}_{j\to i,\mathcal{I}}^{(s)} = h_{j\to i}^{(s)}\mathbf{z}_{j\to i,\mathcal{I}}^{(s)} + \mathbf{n}_{j\to i,\mathcal{I}}^{(s)}, \tag{16}$$

where $h_{j\to i}^{(s)} \in \mathbb{C}$ is the channel coefficient and $\mathbf{n}_{j\to i}^{(s)} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2\mathbf{I})$ denotes additive circularly symmetric complex Gaussian noise. For the AWGN setting, $h_{j\to i}^{(s)}$ is set to 1, such that the received signal is corrupted by additive noise only. For the Rayleigh fading setting, $h_{j\to i}^{(s)}$ is sampled from a zero-mean complex Gaussian distribution, i.e.,

$$h_{j\to i}^{(s)} \sim \mathcal{CN}(0, 1), \tag{17}$$

which models small-scale multipath fading with Rayleigh-distributed amplitude. Specifically, the effective average signal-to-noise ratio for both channel settings is defined as:

$$\gamma = \frac{P}{\sigma_n^2} \tag{18}$$

where P is the average power constraint per complex-valued symbol (set to 1 in our implementation without loss of generality) and σ_n^2 denotes the noise variance.

After transmission, the received sparse symbols $\hat{\mathbf{z}}_{j\to i,\mathcal{I}}^{(s)}$ are re-mapped to their original spatial coordinates based on the index set $\mathcal{I}_j^{(s)}$. Locations not selected for transmission are zero-padded, forming a dense latent feature map for decoding. The decoder $\mathcal{D}_\psi$, which mirrors the encoder structure, progressively upsamples these symbols through a symmetric set of GN-stabilized convolutional blocks to restore the feature map $\tilde{\mathbf{S}}_{j\to i}^{(s)}$. By modeling the system as an end-to-end differentiable pipeline, the proposed framework abstracts the packet-level protocols while strictly adhering to the power and bandwidth constraints. This formulation yields semantic representations inherently resilient to stochastic wireless impairments, enabling robust cooperative perception in safety-critical V2X environments.

E. Task-Oriented Training Strategy

As detailed in Algorithm 1, a two-stage optimization strategy is adopted to ensure stable convergence and align semantic transmission with the downstream perception task. Training is initialized from a pre-trained model without semantic codec and spatial importance predictor. In the first stage, the semantic codec $(\mathcal{E}_\theta, \mathcal{D}_\psi)$ and the spatial importance predictor are jointly trained to minimize the multi-scale JSCC reconstruction loss as defined in (12). This stage aims to establish a reliable semantic reconstruction capability before task-level fine-tuning. To improve robustness under realistic non-LoS V2X conditions, the training process is conducted over a wide range of Rayleigh fading channels with the SNR γ sampled from $\mathcal{U}(0, 20)$ dB. In parallel, the spatial compression ratio η is randomly sampled from $\mathcal{U}(0.01, 0.20)$, allowing the model to learn stable feature selection and reconstruction across different bandwidth budgets. Specifically, the Straight-Through Estimator (STE) technique is employed to enable end-to-end backpropagation through the non-differentiable Top-k selection mask, allowing the importance predictor to learn a task-relevant ranking of spatial features.

In the second stage, the optimization objective shifts from signal-level reconstruction to task-oriented perception accuracy. The feature extraction backbone initialized from the pre-trained model is frozen to preserve learned geometric representation and prevent the catastrophic forgetting of geometric features. The joint fine-tuning of the remaining architecture, encompassing the hierarchical multi-scale semantic codec, the spatial importance predictor, the downstream fusion and the detection head, is performed. By minimizing the composite loss function $\mathcal{L}_{\text{total}}$ in (13), the framework learns to prioritize semantic primitives that are most critical for 3D object detection under constrained communication. This selective optimization allows the fusion and neck layers to adaptively compensate for channel-induced distortions, ensuring that the synthesized BEV representation $\tilde{\mathbf{F}}_i$ is optimally conditioned for perception accuracy rather than merely structural similarity. Specifically, Rayleigh fading channels and spatial compression ratios in this training stage are incorporated using a methodology consistent with the first stage.

Algorithm 1: Two-Stage Training Strategy

Input: Dataset $\mathcal{D} = \{X_j, y_j\}_{j=1}^N$, Spatial compression ratio $[\eta_{\min}, \eta_{\max}]$, SNR $[\gamma_{\min}, \gamma_{\max}]$

Let : $\mathcal{E}_\theta, \mathcal{D}_\psi$: Codec; $\mathcal{W}$: Wireless Channel; Ψ_Θ: Collaborative Perception Network; $\mathcal{M}^{(s)}$: Spatial Importance Predictor.

Stage 1: Feature Reconstruction;
while *not converged* **do**
  Sample $\mathcal{B} \subset \mathcal{D}$;
  foreach $(X_j, y_j) \in \mathcal{B}$ **do**
    $\{\mathbf{F}_j^{(s)}\} \leftarrow \text{Backbone}(X_j)$;
    for $s = 1$ **to** S **do**
      $\mathbf{M}_j^{(s)} \leftarrow \text{Top-}k(\mathcal{M}^{(s)}(\mathbf{F}_j^{(s)}), \eta)$;
      $\hat{\mathbf{z}}_{j\to i}^{(s)} \leftarrow \mathcal{W}(\text{Norm}(\mathcal{E}_\theta^{(s)}(\mathbf{F}_j^{(s)} \odot \mathbf{M}_j^{(s)})); \gamma)$;
      $\tilde{\mathbf{S}}_{j\to i}^{(s)} \leftarrow \mathcal{D}_\psi^{(s)}(\hat{\mathbf{z}}_{j\to i}^{(s)})$;
  Compute reconstruction loss: $\mathcal{L}_{\text{rec}}$, defined as (12);
  Update $\{\theta, \psi, \mathcal{M}\}$ via STE to minimize $\mathcal{L}_{\text{rec}}$;

Stage 2: Task-Oriented Joint Optimization;
Freeze: Backbone;
Train: $\{\theta, \psi, M, \Theta\}$;
while *not converged* **do**
  Sample $\mathcal{B} \subset \mathcal{D}$;
  foreach *ego* i *and collaborators* $j \in \mathcal{B}$ **do**
    for $s = 1$ **to** S **do**
      $\mathbf{M}_j^{(s)} \leftarrow \text{Top-k}(\mathcal{M}^{(s)}(\mathbf{F}_j^{(s)}), \eta)$;
      $\hat{\mathbf{z}}_{j\to i}^{(s)} \leftarrow \mathcal{W}(\text{Norm}(\mathcal{E}_\theta^{(s)}(\mathbf{F}_j^{(s)} \odot \mathbf{M}_j^{(s)})); \gamma)$;
      $\tilde{\mathbf{S}}_{j\to i}^{(s)} \leftarrow \mathcal{D}_\psi^{(s)}(\hat{\mathbf{z}}_{j\to i}^{(s)})$;
    $\hat{y}_i \leftarrow \Psi_\Theta\left(\{\mathbf{F}_i^{(s)}\}_{s=1}^S, \{\tilde{\mathbf{S}}_{j\to i}^{(s)}\}_{j\neq i, s=1}^S\right)$;
  Compute total loss: $\mathcal{L}_{\text{total}}$, defined as (13);
  Jointly update $\{\theta, \psi, \mathcal{M}, \Theta\}$ to minimize $\mathcal{L}_{\text{total}}$;

IV. Experiments

A. Experimental Setup

Datasets: The proposed HMS-SCP framework is evaluated on two different cooperative perception datasets: the simulated OPV2V [43] and real-world DAIR-V2X [44] datasets covering both V2V and V2I collaboration respectively. The evaluation on the OPV2V dataset is conducted within a spatial range of $x \in [-140.8, 140.8]$ m and $y \in [-38.4, 38.4]$ m with a maximum collaboration radius of 70 m. The HMS-SCP framework is also further evaluated on the DAIR-V2X dataset within a spatial range of $x \in [-100.8, 100.8]$ m and $y \in [-40, 40]$ m with a maximum collaboration radius of 100 m.

Implementation: The HMS-SCP framework is implemented based on the OpenCOOD [43] framework, using PointPillar [37] adopted as the LiDAR feature encoder. The voxel grid size is set to (0.4 m, 0.4 m), and a maximum of 32

TABLE I
COMPARISON OF DIFFERENT SEMANTIC-AWARE COOPERATIVE PERCEPTION FRAMEWORKS.

Method	Ch.	Imp.	JSCC	Sym.	Scale
Where2Comm (SS)	×	✓	×	–	SS
Where2Comm (MS)	×	✓	×	–	MS
SComCP	✓	✓	✓	256	SS
HMS-SCP (SS)	✓	✓	✓	**1**	SS
HMS-SCP	✓	✓	✓	**1**	MS

Ch.: Channel-resilient; Imp.: Importance predictor; Sym.: Complex-valued symbols per selected feature dimension; SS: Single-scale; MS: Multi-scale.

points per Pillar. The number of hierarchical scales, S is set to 3 to provide a balanced multi-resolution representation for cooperative perception. The finest scale preserves high-resolution geometric details that are important for detecting small or distant objects, the intermediate scale captures object-level semantic structures, and the coarsest scale encodes broader contextual information for scene-level reasoning. Adding more scales incurs higher complexity in semantic coding, feature fusion, and bandwidth scheduling, while excessively coarser features offer marginal perceptual gains. Therefore, $S = 3$ offers an effective trade-off between semantic richness, communication efficiency, and real-time inference feasibility.

To evaluate performance under different communication budgets, the bandwidth constraints are configured by setting a target transmission ratio of $R \in \{0.01, 0.05, 0.10\}$. This corresponds to an extreme $100\times$, significant $20\times$, and moderate $10\times$ spatial reduction regime, respectively. For each configuration, the importance predictor is calibrated to select the top k percentile of spatial features according to the prescribed transmission ratio, while maintaining the single complex-valued symbol bottleneck per feature. To emulate practical wireless environments, the model is trained under Rayleigh fading channels with the SNR uniformly sampled from $[0, 20]$ dB, and evaluated under both AWGN and Rayleigh fading conditions. In addition, the spatial compression ratio η is dynamically sampled from $[0.01, 0.20]$ during both training stages to improve robustness across varying communication budgets. This strategy encourages the importance predictor to learn a generalized spatial ranking rather than overfitting to a fixed transmission ratio.

The Adam optimizer is employed for training, with learning rates set to 0.0002 for stage 1 and 0.0001 for stage 2. The loss balancing coefficient in (13) is set to $\mu = 0.05$ for stage 2, prioritizing task-specific detection loss $\mathcal{L}_{\text{det}}$ while using reconstruction loss $\mathcal{L}_{\text{rec}}$ as a mild regularizer. This configuration preserves physically meaningful latent representations without sacrificing the precision required for 3D object detection under low-SNR conditions. Unless otherwise specified, all remaining experimental settings follow the OpenCOOD configuration. All models are trained and evaluated on a workstation equipped with an NVIDIA RTX A5000 GPUs and an Intel i9-13900 CPU, with latency measured on the same platform.

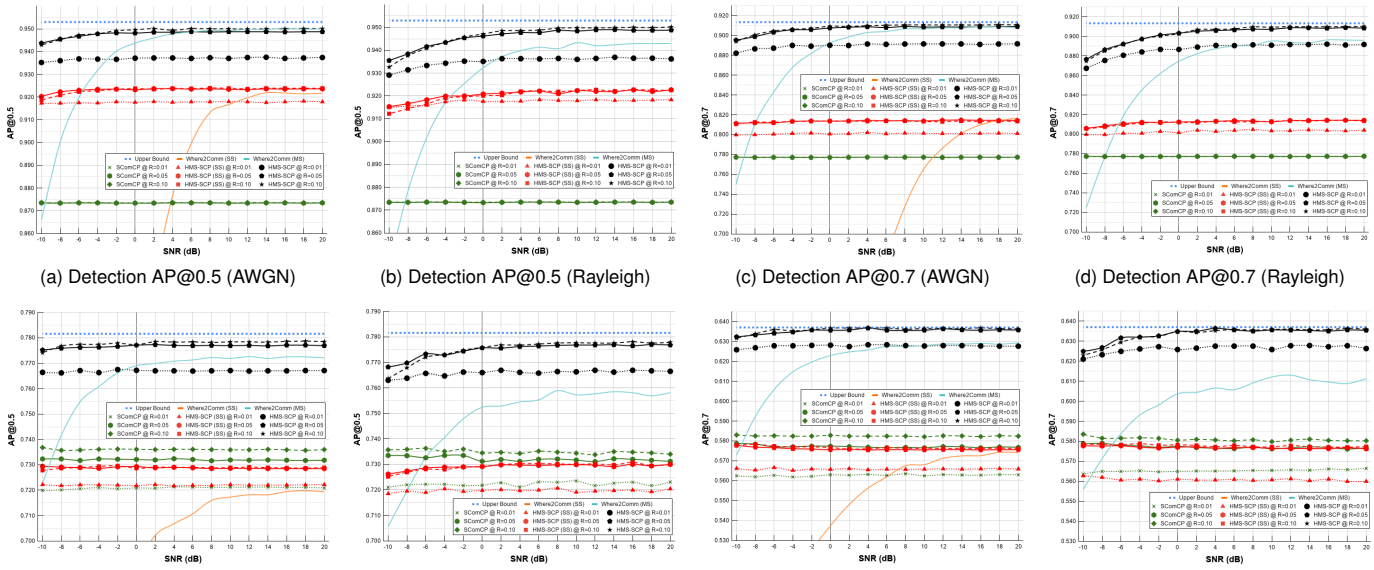
B. Comparative Performance Analysis

In this context, AP@0.5 and AP@0.7 are adopted as the primary evaluation metrics. The proposed HMS-SCP framework is compared against two representative SOTA cooperative perception methods, namely Where2Comm [11] and SComCP [15]. As the official implementation of SComCP is not publicly available, a baseline implementation is developed following the architectural descriptions and specifications reported in the original work. This includes the use of PointPillar encoder for feature extraction, importance-aware feature selection, single-scale attention-based fusion and a projection to 256 symbols per feature. Previous studies [15] indicate that Where2Comm exhibits performance degradation when the importance map is enabled. Consequently, its compression mechanism is disabled in the main comparison to assess its baseline perception capability, as Where2Comm is not inherently designed for channel-resilient transmission. In addition, a multi-scale variant of Where2Comm is implemented to examine whether hierarchical fusion alone can improve robustness beyond the single-scale configurations commonly considered in prior semantic communication studies [14], [15]. To further isolate the contribution of the proposed hierarchical semantic communication design, an additional single-scale (SS) variant of HMS-SCP is evaluated. In this configuration, the multi-scale features $\mathbf{F}_j^{(s)}$ are first passed through deconvolutional blocks and then concatenated at the transmitter to form a unified feature maps $\mathbf{S}_j$. A single importance mask is generated for this concatenated volume, and the selected features are transmitted in a single pass. Finally, the attention-based fusion mechanism is applied to the reconstructed single-scale feature map $\tilde{\mathbf{S}}_{j\to i}$. Table I compares our framework with existing baselines.

TABLE II
PERFORMANCE GAP (%) OF HMS-SCP RELATIVE TO THE UPPER BOUND AT SNR $= 0$ dB.

Dataset	Channel	Metric	$R = 0.01 \downarrow$	$R = 0.05 \downarrow$	$R = 0.10 \downarrow$
OPV2V	Rayleigh	AP@0.5	1.88	0.71	0.62
		AP@0.7	2.92	1.11	1.17
	AWGN	AP@0.5	1.67	0.51	0.36
		AP@0.7	2.55	0.65	0.45
DAIR-V2X	Rayleigh	AP@0.5	1.99	0.75	0.73
		AP@0.7	1.76	0.35	0.35
	AWGN	AP@0.5	1.84	0.58	0.55
		AP@0.7	1.40	0.22	0.05

Quantitative Evaluation: The performance of the proposed framework under varying SNR conditions on the OPV2V and DAIR-V2X datasets is illustrated in Fig. 6. In both figures, the upper bound denotes the maximum achievable performance obtained using the complete semantic feature map under an ideal physical channel. A primary observation is the superior robustness of HMS-SCP under adverse channel conditions compared with non-channel-resilient baselines. Traditional non-channel-resilient single-scale methods like Where2Comm (SS) suffer from a catastrophic cliff effect as SNR drops below 10 dB under the AWGN channel. The degradation becomes more severe under Rayleigh fading, where deep fad-



(a) Detection AP@0.5 (AWGN) (b) Detection AP@0.5 (Rayleigh) (c) Detection AP@0.7 (AWGN) (d) Detection AP@0.7 (Rayleigh)

(e) Detection AP@0.5 (AWGN) (f) Detection AP@0.5 (Rayleigh) (g) Detection AP@0.7 (AWGN) (h) Detection AP@0.7 (Rayleigh)

Fig. 6. Performance comparison across (a)-(d) OPV2V and (e)-(h) DAIR-V2X datasets under varying SNR conditions.

ing and burst-like channel distortions substantially impair the transmitted features. Interestingly, although the uncompressed multi-scale methods like Where2Comm (MS) is capable of achieving high performance under relatively favorable channel conditions, particularly within the 6 to 20 dB range, it still exhibits a pronounced cliff effect at lower SNR levels, typically below 6 dB. This indicates that multi-scale feature representation alone is insufficient to ensure reliable perception unless channel-induced distortions are explicitly considered.

In contrast, other channel-resilient methods, such as SComCP and our HMS-SCP, maintain a more stable AP performance and exhibit graceful degradation even in challenging Rayleigh fading environments at 0 dB. This noise resilience is primarily attributed to the integration of deep JSCC with task oriented training strategy, which enables semantic features to be optimized jointly for compression, transmission, and perception. Notably, HMS-SCP consistently outperforms both the single-scale baseline SComCP and the single-scale HMS-SCP variant across the two datasets. Despite assigning only one complex-valued symbol per spatial location, resulting in a $256\times$ reduction in overhead compared with SComCP, HMS-SCP achieves a higher perception performance across all ranges for both simulated and real-world datasets. This suggests that hierarchical mapping provides a more effective semantic redundancy than the high-dimensional projection used in previous SOTA methods. Furthermore, the results highlight the significant performance gain provided by the multi-scale representation. At SNR above 10 dB, HMS-SCP nearly converges to the upper-bound performance, especially using the transmission ratio of $R = 0.05$ and $R = 0.1$, confirming its ability to achieve a favorable balance between communication efficiency and perception accuracy.

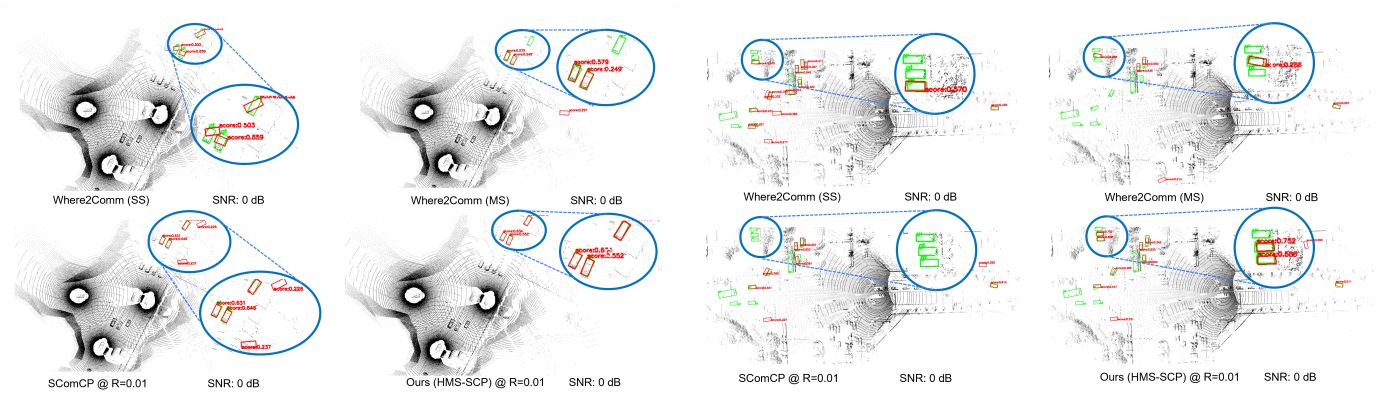
As reported in Table II, HMS-SCP achieves near-upper-bound detection accuracy on the DAIR-V2X dataset even under the challenging 0 dB SNR condition. At the transmission ratio of $R \in \{0.01, 0.05, 0.10\}$, the AP@0.5 performance gaps relative to the upper bound are within 1.99%, 0.75%, and 0.73%, respectively. These results demonstrate that the proposed framework can preserve near-optimal cooperative perception capability while consuming only a small fraction of the available communication budget. Notably, as the transmission ratio increases to $R = 0.05$, the performance gap narrows to a negligible 0.75%, corresponding to a recovery of 99.25% of the upper-bound detection accuracy. Furthermore, doubling the budget to $R = 0.10$ yields a marginal gain of only 0.02%, indicating that task-relevant semantic information is largely captured within the first 5% of the transmitted spatial features. This efficiency is even more pronounced on stricter AP@0.7 metric, where the performance gap remains within 1.76%, 0.35%, and 0.35% of the upper bound, respectively. The saturation of the AP@0.7 gap at $R = 0.05$ further suggests that the core geometric features required for high-precision localization are fully captured within the initial 5% of the transmitted semantic payload. The performance gains are even more prominent under AWGN conditions. A similar comparative analysis for the OPV2V dataset is also provided in Table II. Again, HMS-SCP exhibits remarkable resilience to the challenging 0 dB SNR environment, restricting the AP@0.5 performance gap to within 1.88% of the upper bound even at extreme sparsity ($R = 0.01$), demonstrating a near-lossless 0.36% gap under AWGN conditions as the transmission ratio reaches $R = 0.10$. These results confirm that HMS-SCP achieves strong robustness across both simulated and real-world datasets, while maintaining high communication efficiency under severe channel constraints.

As shown in Table III, the single-scale baselines experience catastrophic performance collapse in the far-field (50–100m) under severe channel noise, whereas the proposed HMS-SCP architecture maintains a significant perceptual margin. On the DAIR-V2X dataset, the HMS-SCP method outperforms the

TABLE III
RELIABILITY ANALYSIS UNDER CHALLENGING CHANNEL CONDITIONS (RAYLEIGH FADING, $SNR = 0\,\text{dB}$)

	OPV2V Dataset						DAIR-V2X Dataset					
	AP@0.5 ↑			AP@0.7 ↑			AP@0.5 ↑			AP@0.7 ↑		
Method	0–30m	30–50m	50–100m	0–30m	30–50m	50–100m	0–30m	30–50m	50–100m	0–30m	30–50m	50–100m
Where2Comm (SS)	0.6036	0.4695	0.3099	0.4417	0.2718	0.1245	0.6383	0.5532	0.3133	0.5112	0.4255	0.1949
Where2Comm (MS)	0.9856	0.9459	0.8002	0.9660	0.8924	0.6775	0.8502	0.7912	0.6109	0.7449	0.6309	0.4416
SComCP @ $R = 0.01$	0.9559	0.8989	0.6828	0.9194	0.7819	0.4932	0.8304	0.7690	0.5629	0.7080	0.6160	0.3821
HMS-SCP (SS) @ $R = 0.01$	0.9795	0.9347	0.7909	0.9483	0.8189	0.5393	0.8342	0.7675	0.5584	0.7000	0.6202	0.3765
HMS-SCP @ $R = 0.01$	**0.9877**	**0.9475**	**0.8170**	**0.9725**	**0.9066**	**0.7041**	**0.8518**	**0.8067**	**0.6344**	**0.7562**	**0.6546**	**0.4726**

Note: Bold values indicate the best performance for the specific range. $R = 0.01$ indicates 100× compression (1 complex-valued symbol per feature).



(a) OPV2V Dataset

(b) DAIR-V2X Dataset

Fig. 7. Visualization of detection results beyond 50 m under severe Rayleigh fading (SNR = 0 dB) at a 1% transmission ratio ($R = 0.01$). Green: ground truth; Red: prediction. Our HMS-SCP framework effectively preserves far-field semantic details that are otherwise lost or corrupted in the baseline methods.

closest competing method, SComCP, by more than 12.7% in AP@0.5 and a remarkable 23.7% in AP@0.7 at the 100 m horizon, despite utilizing an identical 1% spatial sampling ratio. This performance gain is even more pronounced on the OPV2V dataset, where HMS-SCP exceeds SComCP by 19.7% in AP@0.5 and 42.7% in AP@0.7. These results suggest that hierarchical semantic representations provide a more resilient information bottleneck against channel-induced distortions than traditional high-dimensional projections. In particular, the multi-scale design preserves complementary fine-grained and contextual features that are critical for long-range object detection, thereby improving perception reliability in safety-critical V2X scenarios.

Qualitative Evaluation: Fig. 7 presents the visualization of detection results for far-field objects beyond 50 m for for both the OPV2V (Fig. 7a) and DAIR-V2X (Fig. 7b) datasets. Under severe Rayleigh fading ($SNR = 0$ dB), all baseline methods exhibit noticeable performance degradation in far-field perception, primarily due to corruption of transmitted semantic features. Specifically, for the DAIR-V2X dataset, SComCP fails to detect any of the three highlighted distant vehicles, whereas HMS-SCP is capable to identify vehicles beyond 50 m with significantly higher confidence score for the true positive predictions. For the OPV2V dataset, HMS-SCP also accurately detects clusters of far-field objects that are either missed or poorly localized by the Where2Comm baselines. Furthermore, HMS-SCP has no false positive predictions compared to SComCP, indicating improved robustness under noisy channel conditions. These qualitative results provide clear evidence that hierarchical semantic representations are more resilient to channel-induced distortions, ensuring perceptual fidelity in safety-critical long-range V2X scenarios.

C. Ablation Studies

To examine scale-wise contributions of hierarchical feature levels to cooperative perception, a series of ablation experiments are conducted on both datasets under a challenging 0 dB SNR Rayleigh fading channel and an extreme transmission ratio of $R = 0.01$. The far-field results are reported in Table IV. The ablation settings are organized into two categories: transmission suppression, where semantic features from a specific neighboring agent's scale are removed, and fusion ablation, where the model is restricted to ego-only processing at a given scale without using collaborative information from neighboring agents.

The first analysis investigates the importance of each hierarchical scale in the neighbor-to-ego communication channel, with particular emphasis on the 50–100 m horizon, where local sensor limitations are most pronounced. In these configurations, the transmitted feature representation at a specific scale feature s from neighboring agent j is deactivated, forcing the ego agent to rely solely on its local features at that scale. The most significant performance degradation is observed when the neighbor suppresses the transmission of Scale 1. In this

TABLE IV
ABLATION ANALYSIS: IMPACT OF HIERARCHICAL SCALES AND FUSION UNDER RAYLEIGH FADING ($SNR = 0\,\mathrm{dB}$, $R = 0.01$)

	OPV2V (50–100m)			DAIR-V2X (50–100m)		
Configuration	**AP@0.5 ↑**	**Precision ↑**	**Recall ↑**	**AP@0.5 ↑**	**Precision ↑**	**Recall ↑**
HMS-SCP (Full)	**0.8170**	0.7683	**0.8629**	**0.6344**	0.5526	**0.6945**
Tx Suppression						
w/o $\mathbf{z}_j^{(1)}$	0.6078	0.6856	0.6607	0.5037	0.4905	0.5525
w/o $\mathbf{z}_j^{(2)}$	0.8035	**0.7986**	0.8463	0.6230	**0.6031**	0.6760
w/o $\mathbf{z}_j^{(3)}$	0.8167	0.7715	0.8624	0.6346	0.5420	0.6941
Fusion Ablation						
Local-only $\mathbf{F}_i^{(1)}$	0.5932	0.7921	0.6356	0.4987	0.4789	0.5481
Local-only $\mathbf{F}_i^{(2)}$	0.8185	0.6852	0.8697	0.6231	0.5406	0.6800
Local-only $\mathbf{F}_i^{(3)}$	0.8161	0.7688	0.8621	0.6347	0.5435	0.6941
HMS-SCP (SS)	0.7942	0.6865	0.8629	0.5590	0.5024	0.6421

case, the AP@0.5 for far-field drops from 0.8170 to 0.6078 for OPV2V and from 0.6344 to 0.5037 for DAIR-V2X. This consistent performance collapse across both datasets confirms that high-resolution semantic data from neighbors is critical for detecting occluded or distant objects that cannot be reliably resolved by the ego agent's own local Scale 1 features. Interestingly, suppressing Scale 3 transmission has a negligible impact on far-field performance in both datasets. This suggests that the coarse global semantic context encoded by Scale 3 is either spatially redundant across agents or can be sufficiently captured by the ego agent's local representation.

Next, multi-scale fusion is evaluated by disabling self-attention at target scales, forcing the ego vehicle to rely strictly on local features at those hierarchical levels. On the DAIR-V2X dataset, the transition from local-only $\mathbf{F}_i^{(1)}$ and $\mathbf{F}_i^{(2)}$ configurations with full HMS-SCP yields substantial absolute gains of 13.6% and 1.13% in far-field AP@0.5, respectively. This gain is even more evident on the OPV2V dataset, where disabling Scale 1 fusion causes a notable decline in far-field AP@0.5 to 0.5932. These results indicate that cooperative fusion is particularly critical at high-resolution feature levels, where fine-grained spatial cues are essential for detecting distant objects. Although the local backbone provides a strong baseline representation, the scale-wise fusion mechanism substantially enhances long-range perception by integrating complementary observations from neighboring agents.

Lastly, far-field precision and recall metrics are analyzed to evaluate performance from a safety-critical perspective. On both datasets, suppressing Scale 2 transmission produces the highest far-field Prec@0.5, reaching 0.7986 for OPV2V and 0.6031 for DAIR-V2X. However, this gain in precision is achieved at the expense of Rec@0.5, which drops to 0.8463 and 0.6760, respectively. In C-ITS context, false negatives, such as missed vehicles, are generally more safety-critical than marginal gains in bounding-box precision, as they directly affect hazard awareness and reaction time. Thus, the full hierarchical design remains preferred, achieving the highest far-field Rec@0.5 (0.8629 on OPV2V, 0.6945 on DAIR-V2X). This demonstrates superior target coverage in the safety-critical 50–100 m range, offering an optimal precision-recall balance for long-range object awareness.

Overall, the consistency in hierarchical dependencies observed on both the simulated OPV2V and real-world DAIR-V2X datasets confirms that the proposed HMS-SCP is broadly effective across diverse cooperative perception settings. When trained and evaluated under the standard per-dataset protocol, the framework delivers stable performance gains on both datasets, indicating strong architectural adaptability and consistent effectiveness. These results show that the hierarchical multi-scale approach can balance semantic richness and communication overhead across different data distributions and sensor configurations, highlighting its versatility for various V2X scenarios.

D. Bandwidth and Communication Efficiency Analysis

In this section, the communication efficiency of the proposed HMS-SCP is quantified to evaluate its feasibility for real-world V2X deployment. HMS-SCP applies a fixed transmission ratio $R = 0.01$, corresponding to a 1% spatial sampling rate, to the spatial grid points of each hierarchical scale. Each selected spatial location is then mapped to exactly one complex-valued symbol, effectively compressing the corresponding multi-channel semantic latent into a single transmission element. As detailed in Table V, this design results in a lightweight footprint of only 331 symbols for the DAIR-V2X dataset and 444 symbols for OPV2V dataset per perception cycle. The hierarchical structure normally maintains a 16:4:1 symbol ratio across scales, with the majority of the bandwidth dedicated to high-resolution Scale 1 features, which constitutes 76.2% of the total bandwidth budget. This allocation is consistent with the ablation findings, where Scale 1 features are shown to be most critical for long-range perception. In contrast, Scale 3 requires only 16 to 21 symbols, reflecting its role in providing compact global structural context with minimal communication overhead.

TABLE V
LIGHTWEIGHT HIERARCHICAL SYMBOL FOOTPRINT OF HMS-SCP ($R = 0.01$)

Scale	**OPV2V**		**DAIR-V2X**	
	Grid ($H \times W$)	**Symbols**	**Grid ($H \times W$)**	**Symbols**
Scale 1	96×352	338	100×252	252
Scale 2	48×176	85	50×126	63
Scale 3	24×88	21	25×63	16
Total	—	**444**	—	**331**

The proposed HMS-SCP offers a more communication-efficient architecture than the SComCP baseline [15]. SComCP transmits an average of 12 selected semantic features protected by a high symbol expansion with 256 symbols per feature, resulting in approximately 3,072 symbols per perception cycle in each communication round. While SComCP consumes nearly 7 times more bandwidth than the proposed HMS-SCP, it substantially limits the number of spatial locations that can be communicated, leading to a highly sparse spatial

representation. In contrast, HMS-SCP adopts a high-efficiency 1-to-1 semantic mapping strategy by relocating approximately 37 times (444 vs. 12) or 27 times (331 vs. 12) more spatial anchors in the environment. This spatially-augmented approach extending the grid density allowing the model to better capture distant vehicles that sparse models may overlook. As a result, this provides an optimal balance between transmission volume and perceptual reliability. This remarkably small footprint of 331 or 444 symbols per perception cycle underlines the practical feasibility of HMS-SCP for real-world V2X deployment. Considering that 3GPP C-V2X operates in the ITS 5.9 GHz spectrum, the required transmission payload constitutes significantly less than 1% of the available capacity per perception cycle. Under a standard 10 MHz channel and a typical LiDAR operating at 10 Hz, the proposed symbol footprint constitutes merely 0.04% of the available channel capacity per cycle. This efficiency ensures that HMS-SCP can scale to dense traffic without overloading the channel, while meeting the low-latency demands of safety-critical cooperative perception.

E. Complexity and Latency Analysis

To evaluate the practical feasibility of the proposed framework for real-time V2X applications, a comprehensive complexity and latency analysis is conducted. The computational overhead is quantified in terms of learnable parameters, while inference latency is measured on the OPV2V and DAIR-V2X benchmarks. The results are summarized in Table VI. The proposed HMS-SCP architecture comprises 8.70M learnable parameters. As shown in Table VI, the overall complexity of the architecture is primarily dominated by the feature extraction module, which accounts for 68.72%, corresponding to 5.98M parameters. Notably, the additional modules introduced by the HMS-SCP framework remain highly lightweight; the feature selection mechanism implemented through the spatial importance predictor requires only 129K parameters (1.49%), and the semantic codec contributes an additional 514K parameters (5.90%). This limited parameter overhead shows that the proposed modules introduced a minimal memory footprint, making the framework suitable for deployment on resource-constrained V2X edge devices.

In addition to detection accuracy, real-time computational efficiency is essential for safety-critical cooperative perception [45]. The proposed HMS-SCP framework achieves an end-to-end inference latency of 15.73 ms on the OPV2V dataset and 12.17 ms on the DAIR-V2X dataset, which translates to approximately 63 and 82 frames per second (FPS), respectively. Detailed profiling of the pipeline runtime reveals that the feature fusion module remains the primary computational bottleneck, requiring 4.95–6.73 ms due to intensive spatial alignment and coordinate transformation operations performed across multiple agents. Critically, the newly introduced HMS-SCP components, such as feature selection and semantic codec, introduce only marginal overhead, collectively contributing approximately 4–5 ms to the overall pipeline. Given that total processing time remains significantly below the 100 ms latency threshold mandated for safety-critical V2X applications [46], [47], these results confirm that HMS-SCP provides sufficient throughput for high-speed collaborative 3D detection in dynamic and complex urban environments.

TABLE VI
COMPLEXITY ANALYSIS WITH LATENCY

Components	**Params**	**%**	**OPV2V (ms)**	**DAIR-V2X (ms)**
Feature Extraction	5.98M	68.72	3.43	3.30
Feature Selection	129K	1.49	1.26	0.90
Semantic Codec	514K	5.90	4.13	2.88
Feature Fusion	2.07M	23.83	6.73	4.95
Detection Head	5.1K	0.06	0.18	0.14
Total	**8.70M**	**100**	**15.73**	**12.17**

V. CONCLUSION

In this paper, we propose HMS-SCP, a hierarchical multi-scale semantic-aware cooperative perception framework for task-oriented semantic communication in V2X collaborative 3D object detection. Unlike existing frameworks that typically operate on a single intermediate feature map, HMS-SCP exploits a spatial importance predictor that dynamically identifies critical semantic features across hierarchical feature levels, which are then directly mapped to complex-valued symbols via a JSCC codec to streamline the communication bottleneck. This multi-scale selection mechanism enables the system to dynamically allocate bandwidth across different semantic levels, prioritizing fine-grained details or global context depending on their contribution to the perception task. To address the high uncertainty of V2X communication in dense traffic, HMS-SCP leverages structural multi-scale redundancy across multiple feature scales rather than relying on high-dimensional symbol projections. This allows the framework to achieve an ultra-low symbol rate while mitigating network congestion and preserving perception reliability.

Extensive experiments on the simulated OPV2V dataset and the real-world DAIR-V2X dataset demonstrate that HMS-SCP consistently outperforms SOTA baselines under both AWGN and Rayleigh fading channels. Under severe Rayleigh fading, HMS-SCP outperforms the closest competing method by 23.7% on DAIR-V2X and 42.7% on OPV2V at the far-field range beyond 50 m horizon. This has confirmed the effectiveness of hierarchical semantic transmission for long-range cooperative perception. While the current framework employs a uniform spatial compression ratio across all hierarchical scales, future research will explore adaptive, scale-wise ratio allocation to further optimize the trade-off between communication efficiency and perceptual robustness. Furthermore, extending HMS-SCP to multi-modal cooperative perception by incorporating camera, radar, and LiDAR sensors represents a promising direction for achieving more accurate and resilient perception in highly dynamic vehicular environments.

REFERENCES

[1] World Health Organization. (2023) Global status report on road safety 2023. World Health Organization. Geneva, Switzerland. ISBN: 978-92-4-008651-7.

[2] M. Noor-A-Rahim, Z. Liu, H. Lee, M. O. Khyam, J. He, D. Pesch, K. Moessner, W. Saad, and H. V. Poor, "6g for vehicle-to-everything (v2x) communications: Enabling technologies, challenges, and opportunities," *Proceedings of the IEEE*, vol. 110, no. 6, pp. 712–734, 2022.

[3] M. Lu, O. Turetken, O. E. Adali, J. Castells, R. Blokpoel, and P. Grefen, "C-ITS (cooperative intelligent transport systems) deployment in europe: challenges and key findings," in *25th ITS World Congress, Copenhagen, Denmark*, 2018, pp. 17–21.

[4] C. Creß, Z. Bing, and A. C. Knoll, "Intelligent transportation systems using roadside infrastructure: A literature survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 6309–6327, 2024.

[5] E. Palladin, S. Brucker, F. Ghilotti, P. Narayanan, M. Bijelic, and F. Heide, "Self-supervised sparse sensor fusion for long range perception," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025.

[6] E. Moradi-Pari, D. Tian, M. Bahramgiri, S. Rajab, and S. Bai, "DSRC Versus LTE-V2X: Empirical performance analysis of direct vehicular communication technologies," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 5, pp. 4889–4903, 2023.

[7] M. Gonzalez-Martín, M. Sepulcre, R. Molina-Masegosa, and J. Gozalvez, "Analytical models of the performance of c-v2x mode 4 vehicular communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1155–1166, 2019.

[8] T. Huang, J. Liu, X. Zhou, D. C. Nguyen, M. Rahimi Azghadi, Y. Xia, Q.-L. Han, and S. Sun, "Vehicle-to-Everything cooperative perception for autonomous driving," *Proceedings of the IEEE*, vol. 113, no. 5, pp. 443–477, 2025.

[9] M. Yazgan, T. Graf, M. Liu, T. Fleck, and J. M. Zöllner, "A survey on intermediate fusion methods for collaborative perception categorized by real world challenges," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2024, pp. 2226–2233.

[10] J. Xu, Y. Zhang, Z. Cai, and D. Huang, "CoSDH: communication-efficient collaborative perception via supply-demand awareness and intermediate-late hybridization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2025, pp. 6834–6843.

[11] Y. Hu, Z. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 4874–4886.

[12] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 107–124.

[13] W. Weaver, "Recent contributions to the mathematical theory of communication," in *The mathematical theory of communication*. Urbana: University of Illinois Press, 1949, pp. 1–28.

[14] Y. Sheng, H. Ye, L. Liang, and S. Jin, "Semantic communication for cooperative perception based on the importance map," in *Proc. IEEE Int. Conf. Wireless Commun. Signal Process. (WCSP)*, 2023, pp. 177–182.

[15] J. Gan, Y. Sheng, H. Zhang, L. Liang, H. Ye, C. Guo, and S. Jin, "SComCP: Task-Oriented Semantic Communication for Collaborative Perception," *IEEE Trans. Veh. Technol.*, pp. 1–15, 2026.

[16] W. Zhang, C. Sun, H. Li, S. Wang, C. Yuen, and Y. Zhang, "Standardization view: Cross-domain semantic communications enabled real-time autonomous driving," *IEEE Commun. Stand. Mag.*, vol. 9, no. 4, pp. 79–85, 2025.

[17] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[18] L. X. Nguyen, A. D. Raha, P. S. Aung, D. Niyato, Z. Han, and C. S. Hong, "A Contemporary Survey on Semantic Communications: Theory of Mind, Generative AI, and Deep Joint Source-Channel Coding," *IEEE Commun. Surv. Tutor.*, vol. 28, pp. 2377–2417, 2026.

[19] X. Liu, H. Liang, Z. Bao, C. Dong, and X. Xu, "A semantic communication system for point cloud," *IEEE Trans. Veh. Technol.*, vol. 74, no. 1, pp. 894–910, 2025.

[20] Y. Liu, Q. Huang, R. Li, Z. Zhao, S. Zhao, Y. Liu, Y. Zhu, and H. Zhang, "Semantic communication empowered collaborative perception in constrained networks," *IEEE Wireless Commun. Lett.*, vol. 14, no. 3, pp. 701–705, 2025.

[21] Y. Zha, W. Shangguan, J. Chen, L. Chai, W. Qiu, and A. M. López, "Heterogeneous multiscale cooperative perception for connected autonomous vehicles via v2x interaction," *IEEE Internet of Things Journal*, vol. 12, no. 14, pp. 26 633–26 645, 2025.

[22] T. Zhou, J. Chen, Y. Shi, K. Jiang, M. Yang, and D. Yang, "Bridging the view disparity between radar and camera features for multi-modal fusion 3d object detection," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1523–1535, 2023.

[23] Y. Lu, Y. Hu, Y. Zhong, D. Wang, S. Chen, and Y. Wang, "An extensible framework for open heterogeneous collaborative perception," in *The Twelfth International Conference on Learning Representations*, 2024.

[24] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023, pp. 2774–2781.

[25] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "BEVFormer: Learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 3, pp. 2020–2036, 2025.

[26] E. Arnold, M. Dianati, R. de Temple, and S. Fallah, "Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 1852–1864, 2022.

[27] L. Wan, J. Zhao, A. Wiedholz, M. Bied, M. Martinez de Lucena, A. Dinkar Jagtap, A. Festag, A. Augusto Fröhlich, H. Ejaz Keen, and A. Vinel, "A systematic literature review on vehicular collaborative perception—a computer vision perspective," *IEEE Trans. Intell. Transp. Syst.*, vol. 27, no. 1, pp. 81–118, 2026.

[28] D. Yang, K. Yang, Y. Wang, J. Liu, Z. Xu, R. Yin, P. Zhai, and L. Zhang, "How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023, pp. 25 151–25 164.

[29] Y. Hu, J. Peng, S. Liu, J. Ge, S. Liu, and S. Chen, "Communication-efficient collaborative perception via information filling with codebook," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024, pp. 15 481–15 490.

[30] Y. Tao, S. Hu, Z. Fang, and Y. Fang, "Directed-CP: Directed collaborative perception for connected and autonomous vehicles via proactive attention," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2025, pp. 7004–7010.

[31] Y. Lu, Q. Li, B. Liu, M. Dianati, C. Feng, S. Chen, and Y. Wang, "Robust collaborative 3d object detection in presence of pose errors," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023, pp. 4812–4818.

[32] Z. Zhou, H. Xiang, Z. Zheng, S. Z. Zhao, M. Lei, Y. Zhang, T. Cai, X. Liu, J. Liu, M. Bajji *et al.*, "V2xpnp: Vehicle-to-everything spatio-temporal fusion for multi-agent perception and prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025, pp. 25 399–25 409.

[33] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019, pp. 4774–4778.

[34] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 3354–3361, 2012.

[35] ETSI, "Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Collective Perception Service; Release 2," Tech. Rep. ETSI TS 103 324 V2.1.1, Jun. 2023.

[36] J. Shi, J. Zhao, L. Zhuo, X. Wang, X. Zhan, and H. Liu, "V2V cooperative perception with adaptive communication loss for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 10, pp. 14 866–14 878, 2025.

[37] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019, pp. 12 689–12 697.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, p. 6000–6010.

[39] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2020.

[40] R. Girshick, "Fast R-CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1440–1448.

[41] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[42] D. B. Kurka and D. Gündüz, "DeepJSCC-f: Deep joint source-channel coding of images with feedback," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 178–193, 2020.

[43] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2022, pp. 2583–2589.

[44] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, and Z. Nie, "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022, pp. 21 329–21 338.

[45] G. Luo, C. Shao, N. Cheng, H. Zhou, H. Zhang, Q. Yuan, and J. Li, "EdgeCooper: Network-aware cooperative lidar perception for enhanced vehicular awareness," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 1, pp. 207–222, 2024.
[46] 3GPP, "Service requirements for enhanced V2X scenarios," 3rd Generation Partnership Project (3GPP), Tech. Rep. TS 22.186, June 2020.
[47] S.-C. Lin, Y. Zhang, C.-H. Hsu, M. Skach, M. E. Haque, L. Tang, and J. Mars, "The architectural implications of autonomous driving: Constraints and acceleration," *SIGPLAN Not.*, vol. 53, no. 2, p. 751–766, Mar. 2018.