Semantic Communication Empowered Collaborative Perception in Constrained Networks

Yuntao Liu, Qian Huang, Rongpeng Li, *Senior Member, IEEE*, Zhifeng Zhao, *Senior Member, IEEE*, Shuyuan Zhao, Yuan Liu, Yongdong Zhu, and Honggang Zhang, *Fellow, IEEE*

***Abstract*—Traditional collaborative perception methods typically focus on optimizing perception performance in ideal wireless transmission conditions. However, in real-world constrained networks, it is crucial for collaborative perception schemes to alleviate network load while ensuring performance robustness. To address this challenge, this letter introduces S2CP, a Semantic Communication empowered Collaborative Perception framework. Within the S2CP, we propose the Multi-scale Dilated Cross-Attention (MDCA) module to effectively extract task-oriented valuable semantic features for transmission, thereby minimizing data transmission overhead and improving perception performance. Furthermore, to mitigate feature distortion during wireless transmission, we develop a pre-training strategy utilizing Masked AutoEncoders (MAE) to enhance the robustness of S2CP. Experimental results demonstrate that S2CP significantly enhances perception performance while substantially reducing network transmission volume.**

***Index Terms*—Collaborative perception, semantic communication, pre-training, masked autoencoders.**

I. Introduction

ACQUIRING precise perception results is crucial for agents. However, single-agent perception alone encounters difficulties in real-world scenarios, especially with occluded or distant objects [1]. Fig. 1 depicts a situation where the occlusion caused by vehicle *B* poses a detection challenge for the *ego* in identifying vehicle *A*. By leveraging the perception information from the collaborator vehicle *C*, the *ego* can successfully detect vehicle *A*, thereby preventing potential collisions. Therefore, collaborative perception emerges as a viable solution to enhance perception performance [2]. Mainstream research in collaborative perception primarily focuses on effectively leveraging shared data from a computer vision perspective. In that regard, studies such as V2VNet [3], V2X-Vit [4] and Select2Col [5] have made significant advancements. In particular, V2X-Vit enhances feature fusion by employing a vision transformer architecture, while Select2Col utilizes multi-scale and short-term attention mechanisms to fuse features based on the spatiotemporal importance of semantic information. Nevertheless, these works often assume ideal network transmission conditions, raising concerns about their effectiveness in practical constrained environments.

(a) Ego's view (b) Vehicle C 's view (c) Bird's eye view

Fig. 1. Collaborative perception scenario.

In addition, traditional communication strategy struggles to meet the demands of collaborative perception due to encountered network load pressure [2] and the cliff effect in low signal-to-noise ratio (SNR) regimes [6]. In contrast, semantic communication, as a new paradigm, offers a promising solution to the aforementioned limitations [7]. Particularly in the context of multimodal applications for future 6G networks [8], semantic communication holds significant importance. Recent studies [9], [10], [11] have proposed various semantic communication systems for enhancing point cloud transmission in perception tasks. Notably, the research most relevant to this letter is presented in [11], which introduces a collaborative perception scheme that integrates semantic communication. This scheme employs a Deep Joint Source-Channel Coding (D-JSCC) framework while leveraging the PointPillars [12] for feature extraction and a convolutional neural network (CNN)-based importance map to identify and transmit significant semantic information. Although this scheme is concise and effective, there remains considerable potential for improving the efficiency of semantic feature extraction and transmission,

Received 9 October 2024; revised 18 November 2024; accepted 14 December 2024. Date of publication 20 December 2024; date of current version 10 March 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFE0200600, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LR23F010005. The associate editor coordinating the review of this article and approving it for publication was S. Dang. *(Corresponding author: Rongpeng Li.)*

Yuntao Liu and Zhifeng Zhao are with Zhejiang Lab, Hangzhou 311121, China, and also with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310058, China (e-mail: liuyt@zhejianglab.org; zhaozf@zhejianglab.org).

Qian Huang, Shuyuan Zhao, Yuan Liu, and Yongdong Zhu are with Zhejiang Lab, Hangzhou 311121, China (e-mail: huangq@zhejianglab.com; zhaosy@zhejianglab.com; liuyuan@zhejianglab.com; zhuyd@zhejianglab.com).

Rongpeng Li is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: lirongpeng@zju.edu.cn).

Honggang Zhang is with the Faculty of Data Science, City University of Macau, Macau, China, and also with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: hgzhang@cityu.edu.mo).

Digital Object Identifier 10.1109/LWC.2024.3520660

as well as reducing the overall network transmission volume. Prominently, in real constrained networks, the objectives of collaborative perception should not only maximize perception performance but also minimize network impact, such as reducing network load and alleviating transmission pressure. Nevertheless, achieving these goals remains challenging. On one hand, reducing data transmission volume often leads to increased complexity and difficulty in recovering distorted information, potentially diminishing perception performance. On the other hand, designing and training robust network models without adding complexity also poses notable challenges.

To tackle these challenges, this letter presents a Semantic Communication empowered Collaborative Perception framework, named S2CP, aiming to minimize network transmission volume while maximizing perception performance. Within this framework, we design MDCA (Multi-scale Dilated Cross-Attention), a semantic feature extraction module that significantly reduces network transmission data while effectively retaining valuable feature information for perception tasks. Furthermore, we develop a pre-training strategy utilizing Masked AutoEncoders (MAE) [13] to mitigate potential feature information distortion, thereby improving the robustness of S2CP without increasing complexity.

The remainder of this letter is organized as follows. Section II formulates the collaborative perception system model. Section III elaborates on the proposed S2CP framework. We conduct experiments to verify S2CP in Section IV while concluding this letter with future research directions in Section V.

II. System Model and Problem Formulation

Consistent with prior works [4], [5] and [11], we focus on sharing semantic feature information derived from LiDAR point clouds for a downstream 3D object detection task. Meanwhile, we consider a semantic communication-based collaborative perception system within which a receiver capably leverages the delivered feature information from its collaborators.

A. System Model

For clarity of representation, we denote the receiver as agent i and the sender as agent j with $j = 1, 2, \ldots, M$, where M indicates the total number of agent i's collaborators. Without loss of generality, agent j captures a raw point cloud frame X_j and then employs a semantic encoder S_α with parameters α to extract semantic information F_j as

$$F_j = S_\alpha(X_j). \tag{1}$$

In addition, agent j utilizes a channel encoder C_β with parameters β to map F_j into Z_j for transmission, that is

$$Z_j = C_\beta(F_j). \tag{2}$$

We assume that the semantic transmission occurs on a single communication link. Due to impairments in wireless transmission, such as noise and channel fading, the received $\widehat{Z_j}$ by agent i can be modeled as

$$\widehat{Z_j} = h \cdot Z_j + \eta, \tag{3}$$

where $\eta \sim \mathcal{CN}(0, \sigma^2)$ indicates Gaussian noise with a variance of σ^2, and h denotes the channel link from agent j to agent i. Specifically, for the AWGN channel, h is set to 1. For the Rayleigh fading channel, the probability density function of $|h|$ is given by $\frac{r}{\sigma_h^2}\exp(-\frac{r^2}{2\sigma_h^2})$, where r and σ_h^2 indicate the magnitude and variance of channel gain, respectively. For the Rician fading channel, $h = \sqrt{\frac{K}{K+1}}h_{\mathrm{L}} + \sqrt{\frac{1}{K+1}}h_{\mathrm{S}}$, where h_{L} and h_{S} represent the line-of-sight and scattered path gains, respectively, and K denotes the K-Factor [14]. After receiving the $\widehat{Z_j}$, agent i recovers the features to $\widehat{F_j}$ using a channel decoder C_γ^{-1} with parameters γ as

$$\widehat{F_j} = C_\gamma^{-1}\left(\widehat{Z_j}\right). \tag{4}$$

Moreover, agent i may receive semantic features from multiple collaborators during its perceptual cycle. To effectively utilize these features, agent i utilizes a semantic decoder S_δ^{-1} with parameters δ to fuse the received features with its own perception information F_i and jointly decode them as

$$\widehat{Y_i} = S_\delta^{-1}\left(\left\{\widehat{F_j}\right\}_{j\in\{1,2\cdots M\}}, F_i\right), \tag{5}$$

where $\widehat{Y_i}$ denotes the perception results of agent i.

Aligned with previous studies [4], [5] and [11], we employ the average precision (AP) metric to evaluate perception performance. That is,

$$\mathrm{AP} = f_{\mathrm{eva}}\left(\widehat{Y_i}, Y_i, \mathrm{IoU}_{\mathrm{threshold}}\right), \tag{6}$$

where Y_i denotes the ground truth of agent i's perception result, and $\mathrm{IoU}_{\mathrm{threshold}}$ represents the Intersection over Union (IoU) threshold. The function $f_{\mathrm{eva}}(\cdot)$ denotes the standard evaluation method that measures the accuracy of the perception result at specific IoU thresholds, such as 0.5 and 0.7, which are commonly employed [4], [5].

B. Problem Description

Given the constraints of network capacity and potential information distortion in challenging network environments, the objective of collaborative perception is to maximize perception performance while minimizing network transmission volume. Thus, this objective can be formulated as

$$\max_{\alpha,\beta,\gamma,\delta} \mathrm{AP} + \lambda * \exp\left(\frac{-1}{M}\sum_{j=1}^{M}\mathrm{size}(Z_j)\right), \tag{7}$$

where $\mathrm{size}(\cdot)$ measures the volume of transmitted feature information in terms of bits, and λ serves as a scaling factor that balances the tradeoff between perception performance and transmitted data volume.

III. S2CP: A Semantic Communication Empowered Collaborative Perception Framework

This section introduces our proposed collaborative perception framework S2CP. Based on Select2Col [5], S2CP strives to enhance perception performance while reducing transmission volume by integrating the MDCA module and the MAE-based pre-training strategy.

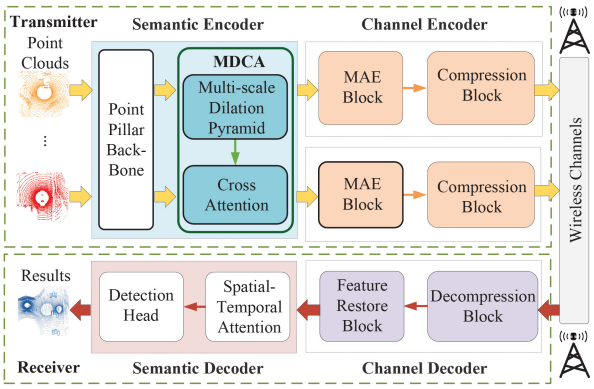


Fig. 2. The framework of our proposed S2CP.

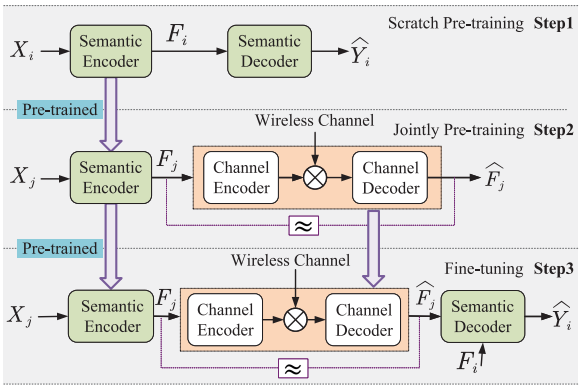


Fig. 3. The MAE-based pre-training strategy.

A. Model Overview

As illustrated in Fig. 2, our proposed S2CP framework comprises components such as semantic encoder S_α, channel encoder C_β, channel decoder C_γ^{-1}, and semantic decoder S_δ^{-1}. Consistent with [11], we employ the D-JSCC strategy and utilize an end-to-end neural network model to effectively integrate and train these components. The detailed design of each component is presented below.

1) Semantic Encoder: In line with [4], [5] and [11], we initially employ the PointPillar [12] backbone to extract rough object features from the raw point cloud frame X_j. To optimize (7), we further add a lightweight feature refinement module MDCA to enhance the precision and relevance of the feature information. The MDCA module is described in Section III-B.

2) Channel Encoder: To enhance the robustness of S2CP and reduce network overhead, we integrate an MAE block [13] and a CNN-based compression block. Initially, the MAE block randomly masks a specified percentage (50% in our experiments) of features in F_j during the training process, while this percentage is set to 0 (i.e., no masking) during the inference process. Subsequently, the compression block downsamples the masked features along the channel dimension to generate Z_j at a specified compression ratio r, where r serves as a hyperparameter.

3) Channel Decoder: To align with the channel encoder, we utilize a similar CNN-based decompression block to upsample the received $\widehat{Z_j}$. Considering the potential distortion of feature information during wireless transmission, a CNN-based feature restore block with a convolutional stride of 1 is employed to restore the received feature information to $\widehat{F_j}$, aiming to mitigate the adverse effects of wireless transmission.

4) Semantic Decoder: We integrate a spatial-temporal attention mechanism to fuse and decode the received semantic features with the receiver's features, inspired by Select2Col [5]. However, compared to the computation-intensive multi-scale attention operation utilized in Select2Col, a simpler single-scale attention-based MDCA achieves comparable functionality with significantly reduced computations. The fused features are then processed by a detection head [12] to generate the perception results $\widehat{Y_i}$.

B. MDCA Module

The objective of the MDCA module is to refine feature information that is closely related to downstream perception tasks, thereby enhancing feature effectiveness while minimizing network transmission volume. As illustrated in Fig. 2, the MDCA module primarily consists of a multi-scale dilation pyramid block and a cross-attention block.

Initially, the multi-scale dilation pyramid block is utilized to refine the rough object features extracted by the PointPillar network across various spatial dimensions. This block consists of three dilated convolutions with dilation rates of 3, 2, and 1, respectively. Each dilated convolution employs a stride value of 3, and the padding value is set as the same one as its dilation rate. The number of input and output channels for the dilated convolutions remains consistent. The multi-scale dilation pyramid block produces three new features. Subsequently, the cross-attention block further refines these features through concatenation and attention-based aggregation [15], thus obtaining a single refined feature F_j. Moreover, to further reduce the feature dimensions, we leverage a dual convolution operation for downsampling the F_j, consistent with [4].

This design offers dual advantages. First, dilated convolutions effectively expand the receptive field by incorporating zeros into the convolution kernel, without increasing the computational complexity. By utilizing varying dilation rates, MDCA can capture receptive fields of different sizes, thereby acquiring multi-scale contextual information. Second, the cross-attention mechanism effectively associates the features with different dilation rates to obtain essential task-relevant cross-scale information while filtering out irrelevant data, thus leading to more precise feature extraction.

C. MAE-Based Pre-Training

As depicted in Fig. 3, our MAE-based Pre-training strategy comprises three consecutive training stages.

1) Step 1: We focus on optimizing perception performance solely through pre-training the semantic encoder S_α and decoder S_δ^{-1}. The perception loss L_{obj} is defined as

$$L_{\text{obj}} = L_{\text{cls}}\left(Y_i, \widehat{Y_i}\right) + \eta L_{\text{reg}}\left(Y_i, \widehat{Y_i}\right), \tag{8}$$

where $L_{\text{cls}}(\cdot)$, $L_{\text{reg}}(\cdot)$ and η represent the object detection classification loss function, regression loss function, and the scaling factor, respectively, consistent with [5].

2) Step 2: We conduct joint pre-training with self-supervised MAE on both the channel encoder C_β and decoder C_γ^{-1}. The transmission loss L_{trans} is formulated as

$$L_{\text{trans}} = \frac{1}{M}\sum_{j=1}^{M}\left|F_j - \widehat{F_j}\right|. \tag{9}$$

3) Step 3: We perform the fine-tuning of the entire model to maximize (7). And the total loss L_{total} is represented as

$$L_{\text{total}} = L_{\text{obj}} + L_{\text{trans}} + \frac{1}{\lambda}\exp\left(\frac{1}{M}\sum_{j=1}^{M}\text{size}(Z_j)\right). \tag{10}$$

In this strategy, Step 2 is crucial. During this phase, the input feature information F_j undergoes random masking using the MAE block, which simulates the distortion of feature information caused by noise interference and channel fading in wireless transmission environments. Subsequently, a coding and decoding-based training is applied to learn the reconstruction of the original feature information F_j from simulated distorted features. Notably, during the inference phase, the MAE block is bypassed.

Such a strategy contributes to effectively recovering distorted features and helps improve the robustness of the S2CP model in tackling feature distortion caused by noise interference and channel fading.

IV. Experimental Evaluation

This section provides a comprehensive performance evaluation of our proposed S2CP.

A. Experimental Settings

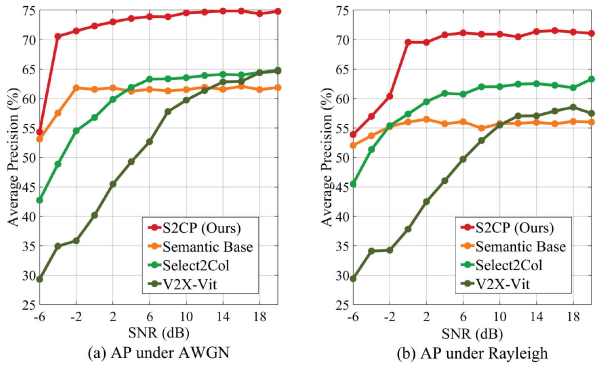
We utilize three state-of-the-art methods, V2X-Vit [4], Select2Col [5], and Semantic Base [11] as our comparative baselines. Both V2X-Vit and Select2Col are collaborative perception methods that rely on ideal network conditions, while Semantic Base belongs to the latest scheme designed for collaborative perception in constrained environments.

We utilize the AP values at IoU thresholds of 0.5 and 0.7 to assess perception performance. Furthermore, given the varying number of collaborators in different scenarios, we measure the network load in terms of the average data transmission per collaborator. We evaluate our S2CP overall performance on two well-known large-scale open-source datasets in this field, namely V2XSet [4] and V2V4Real [16].

All experiments are performed on an X86 workstation equipped with an Intel i7-11700 processor running at 2.50 GHz, a 128-cores CPU, 256 GB RAM, and an NVIDIA RTX3090 GPU. During training, we set the batch size N and epoch size P are 4 and 60, respectively. To mitigate the impact of randomness, we use the average results from 702 test samples in V2XSet and 1, 993 test samples in V2V4Real as the experimental outcomes. Each sample is assigned a uniformly random SNR value within the range of -6 dB to 20 dB to simulate constrained network environments. In addition, we employ additive white Gaussian noise (AWGN), Rician and Rayleigh fading channel models to evaluate the performance of S2CP under varying network conditions. Based on empirical findings, the hyperparameter λ is set to 0.1, while the hyperparameter r is configured to 256, 256, and 128 for the three respective channel models.

TABLE I
Overall Performance Comparison

Dataset	Methods	AP at IoU 0.5 (%)			Tx data (Mbit)
		AWGN	Rician	Rayleigh	
V2V4Real	V2X-Vit	46.09	43.80	39.58	2.16
	Select2Col	49.36	48.48	47.69	1.08
	Semantic Base	50.54	49.71	46.10	0.69
	S2CP (Ours)	**58.14**	**55.23**	**51.59**	**0.27**
V2XSet	V2X-Vit	58.40	53.42	52.73	2.16
	Select2Col	80.28	80.26	80.25	1.08
	Semantic Base	82.92	81.15	80.31	0.69
	S2CP (Ours)	**91.08**	**87.51**	**86.97**	**0.27**



(a) AP under AWGN (b) AP under Rayleigh

Fig. 4. AP under varying SNRs with IoU threshold 0.7 on the V2XSet dataset.

B. Overall Performance Evaluation

Table I and Fig. 4 provide a comparative analysis of the overall performance of our proposed S2CP against other methods across different datasets and settings. The experimental results indicate that S2CP significantly outperforms others in terms of AP and transmission volume and exhibits remarkable robustness. For example, in the AWGN channel model on V2XSet, compared to Select2Col and Semantic Base, S2CP demonstrates substantial improvements of 10.80% and 8.16% in AP performance, while achieving impressive reductions of 75.00% and 60.87% in transmission volume, respectively.

Remark: The impressive results achieved by S2CP can be attributed to the adoption of the MDCA module and the MAE-based pre-training strategy. The MDCA refines features through dilated convolutions and cross-attention operations, enhancing perception performance while reducing network transmission volume. Moreover, the MAE-based pre-training strategy trains and optimizes S2CP by randomly masking feature information at the sender and recovering the original data at the receiver to effectively mitigate feature distortion caused by wireless channels, thereby improving the robustness of S2CP.

TABLE II
COMPUTATIONAL COMPLEXITY COMPARISON

Method	Number of Parameters (M)	MACs (G)	Interference Time (ms)
V2X-Vit	12.4554	260.9526	126.97
Selec2Col	8.2875	349.0030	55.53
Semantic Base	8.1933	107.7043	21.47
S2CP (Ours)	**9.3234**	**287.2713**	**45.09**

TABLE III
ABLATION STUDY ON V2XSET DATASET

IoU Threshold	MDCA	MAE	AP (%)	
			AWGN	Rayleigh
0.5	✘	✘	80.28	80.25
	✔	✘	84.12	83.21
	✘	✔	87.81	85.91
	✔	✔	91.08	86.97
0.7	✘	✘	60.76	59.68
	✔	✘	67.66	66.72
	✘	✔	66.86	62.05
	✔	✔	72.54	68.86

C. Complexity Evaluation

Table II illustrates the complexity comparison between S2CP and other methods in terms of neural network parameters, computational overhead (measured in standard Multiply-ACcumulate operations (MACs)), and inference time. While delivering outstanding performance, S2CP does not introduce notable complexity.

Remark: The MAE block is employed only during the training process, allowing it to be bypassed during the inference process. Consequently, the MAE module does not contribute to the actual complexity of S2CP. In addition, the MDCA module comprises three dilated convolutions and a cross-attention. The dilated convolutions operate similarly to standard 2D convolutions, while the cross-attention is a well-established module. Consequently, the MAE module does not significantly increase the complexity of S2CP. Although S2CP is more complex than the Semantic Base, S2CP significantly outperforms Semantic Base. Given that the perception cycle of LiDAR operates at 10 Hz and the inference time of S2CP is well below 100 ms, the computational performance of S2CP is entirely acceptable. In conclusion, S2CP achieves superior perception performance with high computational efficiency.

D. Ablation Studies

Table III presents the results of our ablation studies, clearly demonstrating that each innovative technique significantly contributes to enhancing perception performance.

Remark: The MDCA module improves perception performance by refining shared features, while the MAE-based pre-training strategy enhances performance by optimizing the S2CP's ability to recover distorted features. Intuitively, a higher IoU threshold indicates greater detection difficulty, leading to more significant benefits from effective feature extraction. Conversely, at a lower IoU threshold, MAE achieves greater gains by correcting feature distortions. Overall, both techniques effectively enhance perception performance.

V. CONCLUSION

This letter presents S2CP, a collaborative perception framework empowered by semantic communication. By introducing semantic communication and developing the MDCA, as well as the MAE-based pre-training strategy, S2CP not only enhances perception performance but also reduces the transmission volume.

Despite its advantages, S2CP still has certain limitations. First, it requires effectively accommodating heterogeneous sensors in real-world scenarios. Second, there is a need to explore adaptive feature extraction and compression techniques to enhance its adaptability. Third, investigating the integration of S2CP with edge computing, federated learning, multi-modal learning, and aggregation techniques for specific applications in Beyond 5G (B5G) and 6G network environments represents a valuable research direction.

REFERENCES

[1] L. Chen et al., "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1046–1056, Feb. 2023.
[2] H. Ngo, H. Fang, and H. Wang, "Cooperative perception with V2V communication for autonomous vehicles," *IEEE Trans. Veh. Technol.*, vol. 72, no. 9, pp. 11122–11131, Sep. 2023.
[3] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2VNet: Vehicle-to-vehicle communication for joint perception and prediction," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 605–621.
[4] 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 *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, Oct. 2022, pp. 107–124.
[5] Y. Liu et al., "Select2Col: Leveraging spatial-temporal importance of semantic information for efficient collaborative perception," *IEEE Trans. Veh. Technol.*, vol. 73, no. 9, pp. 12556–12569, Sep. 2024.
[6] X. Liu, H. Liang, Z. Bao, C. Dong, and X. Xu, "A semantic communication system for point cloud," *IEEE Trans. Veh. Technol.*, early access, Sep. 12, 2024, doi: 10.1109/TVT.2024.3456099.
[7] D. Gündüz et al., "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, Jan. 2023.
[8] J. Du, T. Lin, C. Jiang, Q. Yang, C. F. Bader, and Z. Han, "Distributed foundation models for multi-modal learning in 6G wireless networks," *IEEE Wireless Commun.*, vol. 31, no. 3, pp. 20–30, Jun. 2024.
[9] Z. Wang, Y. Deng, and A. H. Aghvami, "Goal-oriented semantic communications for avatar-centric augmented reality," *IEEE Trans. Commun.*, vol. 72, no. 12, pp. 7982–7995, Dec. 2024.
[10] T. Han, K. Chi, Q. Yang, and Z. Shi, "Semantic-aware transmission for robust point cloud classification," in *Proc. IEEE Global Commun. Conf.*, Kuala Lumpur, Malaysia, Dec. 2023, pp. 7617–7622.
[11] Y. Sheng, H. Ye, L. Liang, S. Jin, and G. Y. Li, "Semantic communication for cooperative perception based on importance map," *J. Frankl. Inst.*, vol. 361, no. 6, 2024, Art. no. 106739.
[12] 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 Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 12697–12705.
[13] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, Jun. 2022, pp. 16000–16009.
[14] X. Yang and A. O. Fapojuwo, "Coverage probability analysis of heterogeneous cellular networks in Rician/rayleigh fading environments," *IEEE Commun. Lett.*, vol. 19, no. 7, pp. 1197–1200, Jul. 2015.
[15] A. Vaswani et al., "Attention is all you need," in *Proc. 31 Conf. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017, pp. 1–11.
[16] R. Xu et al., "V2V4Real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, Jun. 2023, pp. 13712–13722.