2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

WaveComm: Lightweight Communication for Collaborative Perception via Wavelet Feature Distillation

Erdemt Bao[1*†], Jin Yang[2*‡]

***Abstract*— In multi-agent collaborative sensing systems, substantial communication overhead from information exchange significantly limits scalability and real-time performance, especially in bandwidth-constrained environments. This often results in degraded performance and reduced reliability. To address this challenge, we propose WaveComm, a wavelet-based communication framework that drastically reduces transmission loads while preserving sensing performance in low-bandwidth scenarios. The core innovation of WaveComm lies in decomposing feature maps using Discrete Wavelet Transform (DWT), transmitting only compact low-frequency components to minimize communication overhead. High-frequency details are omitted, and their effects are reconstructed at the receiver side using a lightweight generator. A Multi-Scale Distillation (MSD) Loss is employed to optimize the reconstruction quality across pixel, structural, semantic, and distributional levels. Experiments on the OPV2V and DAIR-V2X datasets for LiDAR-based and camera-based perception tasks demonstrate that WaveComm maintains state-of-the-art performance even when the communication volume is reduced to 86.3% and 87.0% of the original, respectively. Compared to existing approaches, WaveComm achieves competitive improvements in both communication efficiency and perception accuracy. Ablation studies further validate the effectiveness of its key components.**

I. INTRODUCTION

Collaborative perception enables multiple agents to overcome inherent limitations of single viewpoints (e.g., occlusion, restricted sensing range) by exchanging complementary information through communication. This paradigm has been widely applied in connected and automated vehicles (CAVs), supported by Vehicle-to-Vehicle (V2V) and vehicle-to-infrastructure (V2I) technologies, significantly enhancing safety and efficiency in complex traffic environments [1], [2], [3].

However, the benefits of collaboration come at the cost of substantial communication overhead. Raw sensor sharing or dense feature map exchange generates extremely high bandwidth demands, which are impractical in real-world deployments [4], [5]. Even compression-oriented strategies (e.g., feature sparsification [6], learned transmission [7]) face trade-offs. They often lead to excessive bandwidth consumption under dense traffic, and transmit redundant or noisy information that does not contribute to downstream tasks, thereby reducing system efficiency.

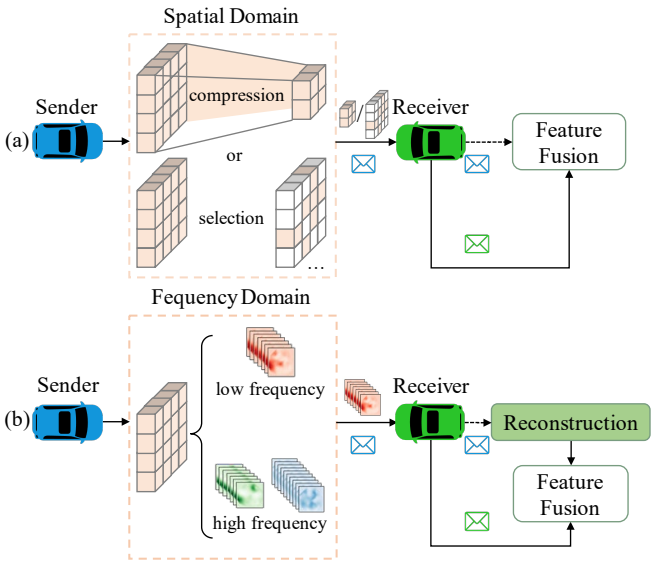

Fig. 1. Feature transmission methods in collaborative perception. (a) Spatial domain methods: The sender vehicle encodes data, applies compression or selection, and transmits it to the receiver vehicle, which performs feature fusion and decoding. (b) Frequency domain methods: The sender encodes data into low- and high-frequency components, primarily transmitting low-frequency data. The receiver reconstructs the original features from these components, followed by feature fusion and decoding.

A key limitation of existing approaches [6], [7], [8], [9], [10] is that they operate solely in the spatial domain, where redundancy is reduced by compressing or selecting feature maps, as shown in Figure 1(a). While such strategies can reduce redundancy, they overlook the inherent frequency structure of perceptual signals. In the frequency domain, signals naturally decompose into complementary components: low-frequency bands that capture global semantics and structural contexts, and high-frequency bands that provide some details such as edges and textures. This inherent separation offers a principled way to design communication schemes that are both compact and information-preserving.

Based on this insight, we propose WaveComm, a wavelet-based collaborative perception framework that achieves communication efficiency by explicitly operating in the spatial frequency domain, similar in spirit to JPEG-style frequency-domain processing, as illustrated in Figure 1(b). Specifically, WaveComm decomposes intermediate features using the Discrete Wavelet Transform (DWT) and transmits only the compact low-frequency components. Although high-frequency

[1]School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China. baoerdemt366@gmail.com
[2]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Shaanxi 710049, China. jin.y.hust@gmail.com
*These authors contributed equally to this work.
†Corresponding author. ‡Project Leader.
GitHub page: https://github.com/erdemtbao/WaveComm

components encode edges, textures, and contour sharpness, their incremental contribution to collaborative perception is small relative to their communication cost. We therefore omit them from transmission and instead reconstruct their effects at the receiver side. At the receiver side, instead of applying the Inverse Discrete Wavelet Transform (IDWT) to recover the original features, we design a lightweight generator that reconstructs full feature maps directly from the transmitted low-frequency components. The generator is optimized using a combination of reconstruction loss, perceptual loss, structural similarity loss, and adversarial loss, ensuring that the reconstructed features remain both semantically consistent and task-aligned.

Through the above design, WaveComm realizes a principled frequency-aware communication scheme that reduces bandwidth while preserving task-critical semantics. Moreover, the decomposition into low-frequency components is compatible with existing compression and feature selection strategies, making it easy to integrate WaveComm with prior frameworks for further improvements. Furthermore, we extend our design with multi-level wavelet decomposition, which enables progressive transmission and reconstruction at different resolutions, further enhancing flexibility under varying bandwidth constraints.

We summarize the contributions of this work as follows:

- We introduce a wavelet-based framework that reduces communication cost by selectively transmitting low-frequency components while retaining sufficient information for downstream tasks.
- We design a Wavelet Feature Distillation Module to reconstruct full features from transmitted low-frequency components and optimize it with a hybrid objective.
- Extensive experiments show that WaveComm achieves a superior efficiency–accuracy trade-off. Ablation studies further confirm the effectiveness of its modules.

II. RELATED WORK

A. Communication-Efficient Cooperative Perception

To reduce communication costs, a variety of communication-efficient methods have recently been proposed. Methods like Where2comm [11] and CodeFilling [12] selectively transmit critical features using spatial confidence or codebook representations, balancing efficiency and performance, though requiring prior single-agent perception that increases latency. CMiMC [13] maximizes mutual information to preserve discriminative data. How2Comm [14], ERMVP [9], and FFNet [15] exploit spatial sparsity of foreground objects or temporal correlations for downsampling and flow-based transmission. Fast2comm [16] uses bounding box priors to minimize noise and adapt to localization errors. DiscoNet [17] applies matrix-valued weights and teacher-student frameworks for interactions. Transformer methods include V2X-ViT [8] for V2X cooperation, CoBEVT [18] for BEV segmentation, and HM-ViT [19] with sparse heterogeneous attentions for multi-agent multi-camera 3D detection and heteromodal perception. Recent approaches like CoCMT [20] and CoopDETR [21] leverage object queries for efficient transmission, while Which2comm [22] integrates semantic detection boxes for sparse, object-level sharing.

Current research on communication-efficient cooperative perception focuses heavily on spatial domain techniques, often overlooking the frequency domain's potential. Frequency-based methods allow precise control over data transmission, prioritizing key components to reduce redundancy while improving computational efficiency.

B. Wavelet Transforms in Feature Processing

Wavelet transforms are widely applied in image processing to enhance visual quality, computational efficiency, and feature representation. In generative models, SWAGAN [23] integrates wavelets into GANs to improve image quality and performance, while Wavelet-srnet [24] predicts wavelet coefficients for super-resolution image reconstruction. Wavelet-based methods like CWNN-MRF [25] and Wavelet-Pooling [26] replace pooling operators in CNNs, using dual-tree complex wavelet transforms or learned wavelet bases to enhance performance without compression. For compression, wavelet-based approaches optimize neural network efficiency. Efficient Wavelet-Based Linear Layers [27] learn wavelet basis functions to compress linear layer weights, unlike activation compression. WCC [28] uses Haar-wavelet transforms to compress feature maps, integrating with pointwise convolutions to reduce costs in image-to-image tasks. Masked Wavelet NeRF [29] applies wavelets to grid-based neural fields for efficient parameter compression, maintaining data structure benefits. HL-RSCompNet [30] employs Discrete Wavelet Transform (DWT) to split features into high- and low-frequency components, enhancing compression via frequency domain encoding-decoding. Similarly, UGDiff [31] uses wavelets for high-frequency compression in diffusion models, predicting high frequencies and compressing residuals to improve fidelity.

Despite these advances, the use of frequency domain wavelet transforms in multi-agent collaborative perception remains limited. To address this gap, this work explores feature processing in the frequency domain and proposes a novel decomposition and reconstruction mechanism.

III. PROBLEM FORMULATION

We investigate a cooperative perception framework comprising N agents, each undertaking distinct detection tasks. In this setup, each agent simultaneously serves as a contributor, sharing perceptual data with others, and as a recipient, leveraging data received from peers. Let $\mathcal{X}_i$ denote the sensory input (e.g., from LiDAR or cameras) collected by the i-th agent, and let $\mathcal{G}_i^0$ represent the corresponding ground-truth annotations for 3D object detection. The goal is to optimize the detection model parameters to maximize the aggregate detection performance, subject to a total communication budget C. Specifically, the optimization problem is formulated as:

$$\underset{\theta,\mathcal{P}}{\operatorname{argmax}} \sum_{i=1}^{N} h\left(\Phi_\theta\left(\mathcal{X}_i, \{\mathcal{P}_{j\to i}\}_{j=1}^{N}\right), \mathcal{G}_i^0\right), \quad \text{s.t.} \quad \sum_{\substack{i,j=1 \\ j\neq i}}^{N} c(\mathcal{P}_{j\to i}) \leq C \tag{1}$$

where Φ_θ is the 3D object detection model parameterized by θ, $\mathcal{P}_{j\to i}$ represents the message transmitted from agent j to agent i, $h(\cdot,\cdot)$ is the evaluation metric for detection performance, and $c(\cdot)$ quantifies the communication cost of the messages. The primary challenge lies in designing the messages $\mathcal{P}_{j\to i}$ to be both informative for enhancing detection and compact to satisfy the communication constraint.

Communication volume. We follow the communication volume definition from Where2comm [11], but use float32 or float16 instead of a fixed 32-bit representation. For the message sent from the i-th to the j-th agent, the binary selection matrix $\mathbf{M}_{i\to j} \in \mathbb{R}^{H\times W}$ represents the spatial grid, with H and W as height and width. The communication volume is:

$$\log_2\left(|\mathbf{M}_{i\to j}| \times D \times n_L/8\right) \tag{2}$$

where $|\cdot|$ denotes the L0 norm, indicating the count of non-zero entries in the binary selection matrix (i.e., the total spatial grids transmitted), D represents the channel dimension, n_L corresponds to the float32 or float16 data type, and dividing by 8 converts the result to bytes.

IV. METHODOLOGY

A. Overview of WaveComm Architecture

The architecture of WaveComm is shown in Figure 2. Blue represents the ego vehicle, green denotes collaborative Connected Autonomous Vehicles (CAVs) or Road-Side Units (RSUs). Multiple agents uniformly process their individual observations, denoted as $\mathcal{X}_i$, through a Bird's-Eye-View (BEV) feature encoder to extract corresponding feature maps $\mathcal{F}_i$. These feature maps $\mathcal{F}_i$ are further processed to generate the target features $\mathcal{Z}_i$ for transmission. The features $\mathcal{Z}_i$ are decomposed using DWT into low-frequency and high-frequency components. To enhance communication efficiency in collaborative perception, only the low-frequency components are transmitted across the collaboration link.

At the receiving end, the Wavelet Feature Distillation module reconstructs the original $\mathcal{Z}_i$ by recovering the missing high-frequency details. This process begins with the compressed low-frequency component processed through IDWT to yield restored features $\mathcal{Z}_i'$, which serve as a supervisory signal. Next, a Generator module, using data priors and adversarial learning with a Discriminator, produces reconstructed features $\hat{\mathcal{Z}}_i$. The Discriminator, combined with a Multi-Scale Distillation (MSD) Loss, evaluates the authenticity of the reconstructed features, ensuring high-fidelity reconstruction and close resemblance to the original features. All reconstructed $\hat{\mathcal{Z}}_i$ from participating agents are then fused by the Pyramid Fusion network [32]. During feature fusion, each agent's reconstructed BEV feature $\hat{\mathcal{Z}}_i$ is warped into the ego coordinate frame using pairwise affine transformations, and a per-cell softmax-weighted sum across agents is computed to obtain the fused BEV feature, which is finally fed into the detection head.

B. BEV Feature Encoder

For each agent, the input data $\mathcal{X}_i$, which may consist of RGB images or 3D point clouds, is converted into a BEV feature representation. This method allows all agents to project their individual sensory inputs into a unified global coordinate system, streamlining collaboration by removing the need for intricate coordinate transformations. The BEV encoder processes the input to produce a feature map $\mathcal{F}_i \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the height, width, and number of channels, respectively. By sharing a common BEV coordinate system, agents can efficiently integrate and process data from either RGB images or point clouds, fostering seamless multi-agent cooperation. We use the term "feature" to refer to these learned CNN feature maps rather than raw occupancy counts, and in this work the BEV features are computed per frame without temporal aggregation.

C. Wavelet Feature Distillation Module

The Wavelet Feature Distillation Module consists of two parts: feature decomposition and feature reconstruction. We use DWT for feature decomposition, breaking down BEV features into low-frequency and high-frequency components. Low-frequency components retain most semantic information and global structure, and are the core basis for perception tasks. High-frequency components, on the other hand, primarily contain local details such as edges, textures, and contours. While they offer limited gains in perception accuracy, they incur additional communication overhead. Therefore, we only transmit low-frequency components during inter-vehicle communication to reduce bandwidth and computing resource consumption. Distinct from naive spatial downsampling which uniformly discards information, DWT preserves the global spatial structure while removing high-frequency details.

For feature reconstruction, the most direct approach is to use IDWT. However, since missing high-frequency components must be zeroed out, the resulting features are often overly smooth and fail to meet the discriminative feature requirements of downstream detection networks. To address this, we propose a Wavelet Generator that directly generates complete features from low-frequency components. Using data priors to infer missing high-frequency information, the resulting features are optimized through a series of loss functions. This ensures that the generated features maintain global consistency while recovering richer details and semantics, effectively improving downstream perception performance while ensuring communication efficiency. The Wavelet Generator is divided into three main modules: the Decoder, Upsample, and Output stages, as illustrated in Figure 3. The Decoder consists of two 3×3 convolutional layers with batch normalization (BN) and ReLU activation,

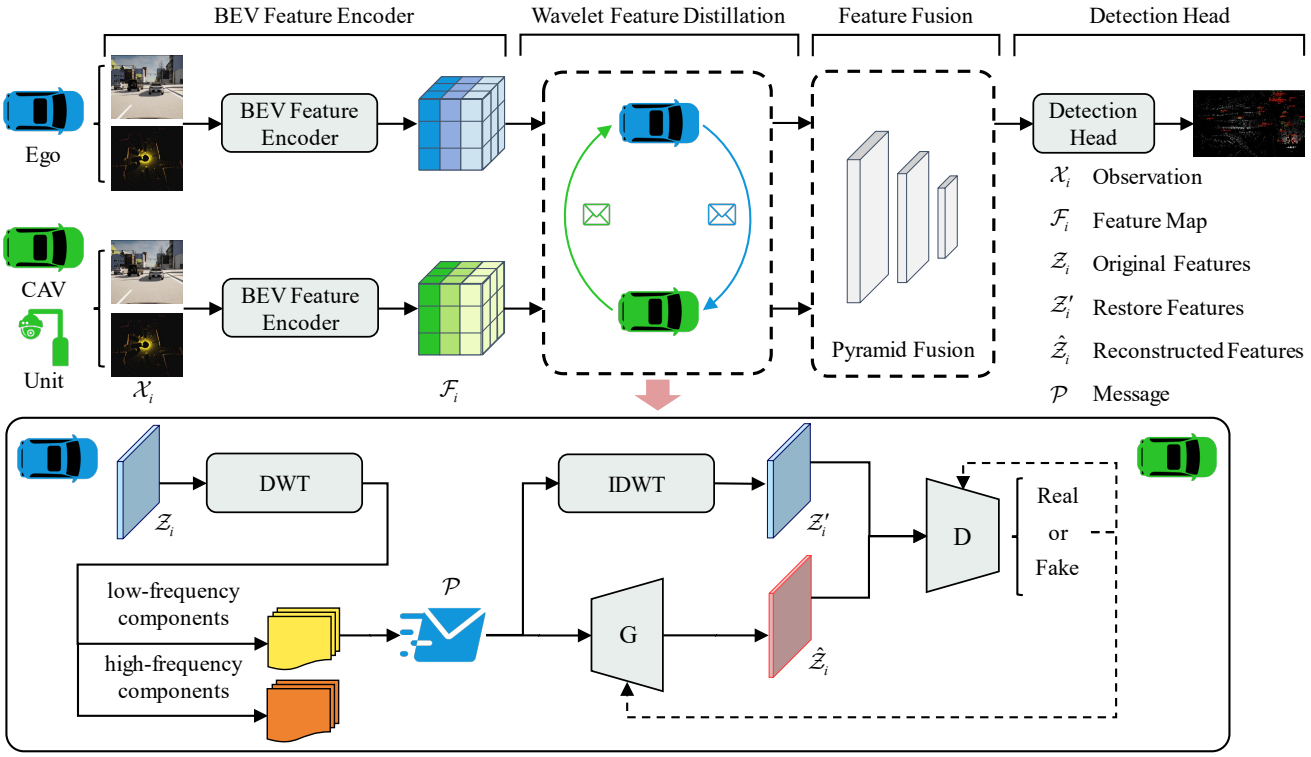

Fig. 2. Overview of WaveComm. WaveComm enables efficient information exchange among intelligent agents to support collaborative autonomous driving. (a) BEV Feature Encoder, which converts agent observations into BEV feature maps. (b) Wavelet Feature Distillation, which employs DWT to decompose features into low- and high-frequency components, followed by a Wavelet Generator and Wavelet Discriminator for efficient feature reconstruction using IDWT. (c) Feature Fusion, which integrates features from multiple agents to enhance the overall feature representation effectively. (d) Detection Head, which produces final detection outputs based on the fused features.

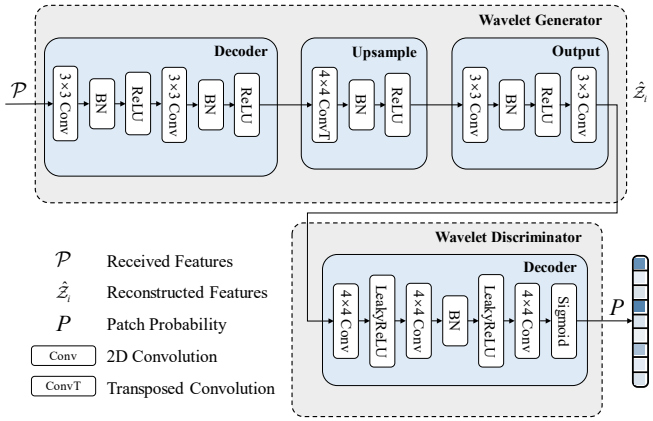

Fig. 3. Architecture of Wavelet Generator and Wavelet Discriminator. The Wavelet Generator uses Decoder, Upsample, and Output modules to reconstructed features $\hat{\mathcal{Z}}_i$ from the transmitted low-frequency component. The Wavelet Discriminator employs Sigmoid to generate a probability map P like PatchGAN [33].

processing the input low-frequency component into intermediate features. The Upsample module employs a 4×4 transposed convolutional layer (ConvTranspose) with stride 2, BN, and ReLU to upscale the features, doubling the spatial dimensions. The Output stage includes two additional 3×3 convolutional layers with BN and ReLU, producing the reconstructed features $\hat{\mathcal{Z}}_i$.

In addition, we further introduce the Wavelet Discriminator, which constrains the generator output through adversarial learning, making it not only globally consistent with the low-frequency input but also closer to the true features in terms of distribution. Specifically, the discriminator is trained to distinguish features recovered by the generator from true features, while the generator tries to fool the discriminator by minimizing the adversarial loss, thereby learning a sharper and more semantically consistent representation. The Wavelet Discriminator comprises a single Decoder module with three 4×4 convolutional layers, each with stride 2, interspersed with LeakyReLU activation and BN, as illustrated in Figure 3, culminating in a Sigmoid activation to output a probability map P for authenticity assessment.

D. Multi-Scale Distillation Loss

During training, we use features derived from low-frequency components via an IDWT as the generator's supervisory signal. While the IDWT reconstruction results appear too smooth due to the lack of high-frequency components and are therefore unsuitable as the final output for direct perception tasks, they are physically identical to the transmitted low-frequency components, providing a stable and reasonable learning target for the generator. Directly using the original features as supervision would force the model to recover high-frequency information that was never

transmitted during communication, making the training objective unattainable. In contrast, the IDWT results provide the generator with a baseline that complies with frequency domain constraints. While maintaining low-frequency consistency, the generator further recovers detailed and discriminative features through Multi- Scale Distillation (MSD) Loss combined with the following four levels of loss, thereby reconstructing features that are both physically reasonable and useful for downstream tasks.

- Pixel-Level, Reconstruction Loss: Ensures element-by-element numerical similarity through L1 loss.
- Structural-Level, SSIM Loss: Focuses on structural, brightness, and contrast similarity of feature maps through SSIM loss.
- Semantic-Level, Perceptual Loss: Emphasizes high-level semantic similarity through perceptual loss.
- Distributional-Level, Adversarial Loss: Ensures that the distribution of the generated features is close to the real features through adversarial loss.

The "multi-scale" nature of the MSD loss is evident in its hierarchical constraints across low- to high-level features, potentially supporting applications such as feature compression and cooperative perception.

For a given agent, we denote its IDWT-restored features $\mathcal{Z}'$ as f and its generator-reconstructed features $\hat{\mathcal{Z}}$ as $\hat{f}$ for brevity. The subscript n indexes individual spatial elements after flattening these feature tensors, with N denoting the total number of elements.

Reconstruction Loss. The low-frequency component of the wavelet decomposition is reconstructed into the original information to recover one of its original representations. The element-by-element difference between the reconstructed features and the restored features is measured using the L1 loss, which encourages accurate reconstruction at the pixel level, with the reconstructed features being as close as possible to the restored features in terms of value.

$$\mathcal{L}_{\text{Recon}} = \frac{1}{N}\sum_{n=1}^{N}\left|\hat{f}_n - f_n\right| \tag{3}$$

Perceptual Loss. During the reconstruction process, the goal is to ensure that the reconstructed features closely resembles the restored features in terms of high-level semantic content. The perceptual loss measures the feature similarity by computing the mean square error (MSE) between the normalized reconstructed features and the normalized restored features. This encourages semantic consistency in the high-dimensional feature space, beyond mere pixel-level accuracy. The perceptual loss function is defined as:

$$\mathcal{L}_{\text{Percep}} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\hat{f}_n}{\|\hat{f}\|_2} - \frac{f_n}{\|f\|_2}\right)^2 \tag{4}$$

where:

- $\|\hat{f}\|_2$, $\|f\|_2$: The L2 norm of the reconstructed and restored feature vectors, respectively, used for normalization.
- $\frac{\hat{f}_n}{\|\hat{f}\|_2}$, $\frac{f_n}{\|f\|_2}$: The normalized n-th elements of the reconstructed and restored features, respectively.

Structural Similarity Loss. This loss calculates the structural similarity index (SSIM), which is used to measure the similarity between two feature maps in terms of brightness, contrast, and structure. The structural similarity of feature maps is emphasized to make up for the insensitivity of L1 loss to structural information.

$$\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}(\hat{f}, f) \tag{5}$$

Adversarial Loss. The generator creates reconstructed features to mimic restored features, aiming to fool the discriminator. The discriminator distinguishes restored features from reconstructed ones, maximizing classification accuracy. The discriminator loss enhances this accuracy, while the generator loss optimizes the generator to produce reconstructed features indistinguishable from restored features.

$$\begin{aligned}\mathcal{L}_{\text{D}} &= \text{BCE}(D(f),1) + \text{BCE}(D(\hat{f}),0)\\ &= -\frac{1}{N}\sum_{n=1}^{N}\Big[\log(D(f_n)) + \log(1 - D(\hat{f}_n))\Big]\end{aligned} \tag{6}$$

$$\mathcal{L}_{\text{G}} = \text{BCE}(D(\hat{f}),1) = -\frac{1}{N}\sum_{n=1}^{N}\log(D(\hat{f}_n)) \tag{7}$$

where:

- $D(\cdot)$: The discriminator function, outputting a probability that the input feature is real.
- BCE: The binary cross-entropy loss function.

Total Loss. The total loss is obtained by taking the weighted sum of the aforementioned losses.

$$\mathcal{L}_{\text{MSD}} = \begin{cases}\mathcal{L}_{\text{ReconTotal}} & = \lambda_{\text{recon}} \cdot \big(\alpha \cdot \mathcal{L}_{\text{Recon}} + \beta \cdot \mathcal{L}_{\text{SSIM}} \\ & \quad +\gamma \cdot \mathcal{L}_{\text{Percep}}\big) \\ \mathcal{L}_{\text{Adv}} & = \lambda_{\text{adv}} \cdot \mathcal{L}_{\text{G}} \\ \mathcal{L}_{\text{D}} & = \text{BCE}(D(f),1) + \text{BCE}(D(\hat{f}),0)\end{cases} \tag{8}$$

where:

- $\mathcal{L}_{\text{Recon}}$, $\mathcal{L}_{\text{SSIM}}$, $\mathcal{L}_{\text{Percep}}$: The reconstruction, structural similarity, and perceptual loss components, respectively.
- $\mathcal{L}_{\text{G}}$, $\mathcal{L}_{\text{D}}$: The generator and discriminator losses in the adversarial framework, respectively.
- λ_{recon}, λ_{adv}, α, β, γ: Weighting coefficients to balance the contributions of each loss term.

V. EXPERIMENTAL RESULTS

A. Dataset and Evaluation Metrics

Datasets. OPV2V [34] is a simulated dataset designed for V2V collaborative perception. It is constructed using the CARLA simulator [35] and the OpenCDA framework [36], generating diverse driving scenarios with a particular emphasis on V2V communication for perception tasks. The dataset comprises 12,000 frames spanning eight different towns in CARLA and a digital replica of Culver City, Los

TABLE I

PERFORMANCE COMPARISON ON OPV2V AND DAIR-V2X. COMM DENOTES THE COMMUNICATION VOLUME CALCULATED WITH EQ. (2).

Dataset	OPV2V						DAIR-V2X					
Method	Camera-based			LiDAR-based			Camera-based			LiDAR-based		
	AP50 ↑	AP70 ↑	Comm	AP50 ↑	AP70 ↑	Comm	AP30 ↑	AP50 ↑	Comm	AP30 ↑	AP50 ↑	Comm
No Collaboration	0.405	0.216	0.0	0.782	0.634	0.0	0.014	0.004	0.0	0.421	0.405	0.0
F-Cooper (2019)	0.469	0.219	22.0	0.763	0.481	24.0	0.115	0.026	23.0	0.723	0.620	23.0
DiscoNet (2021)	0.517	0.234	22.0	0.882	0.737	24.0	0.083	0.017	23.0	0.746	0.685	23.0
AttFusion (2022)	0.529	0.252	22.0	0.878	0.751	24.0	0.094	0.021	23.0	0.738	0.673	23.0
V2X-ViT (2022)	0.603	0.289	22.0	0.917	0.790	24.0	0.198	0.057	23.0	0.785	0.521	23.0
CoBEVT (2022)	0.571	0.261	22.0	0.935	0.821	24.0	0.182	0.042	23.0	0.787	0.692	23.0
HM-ViT (2023)	0.643	0.370	22.0	0.950	0.873	24.0	0.163	0.044	23.0	0.818	0.761	23.0
WaveComm (ours)	**0.681**	**0.451**	**19.0**	**0.965**	**0.926**	**21.0**	**0.274**	**0.123**	**20.0**	**0.831**	**0.790**	**20.0**

Angeles, containing over 232,913 3D vehicle bounding boxes in total. It is designed to replicate complex real-world driving situations, including varying traffic densities and dynamic driving behaviors. In contrast, DAIR-V2X [5] is a real-world collaborative perception dataset consisting of 9,000 frames. Each frame includes synchronized data from one vehicle and one RSU, both equipped with a LiDAR sensor and a 1920×1080 camera. The LiDAR on the RSU has 300 channels, while the vehicle-mounted LiDAR is 40-channel.

Evaluation Metrics. We use Average Precision (AP) at Intersection-over-Union (IoU) thresholds of 0.3, 0.5, and 0.7 to evaluate 3D object detection performance. To evaluate the transmission cost, we follow the definition of communication volume in Eq. (2), which calculates the communication volume by measuring the message size in bytes, expressed in a logarithmic scale with base 2. In other methods, n_L is 32, while in our method, n_L is 16.

B. Experimental Setup

The experiments were conducted using the OpenCOOD framework on the OPV2V and DAIR-V2X datasets.

For the LiDAR-based setup, we employ PointPillars [37] as the feature encoder, which takes 64-channel LiDAR data as input. The detection range is set to $x, y \in [-102.4\text{m}, +102.4\text{m}]$. The feature map is downsampled by a factor of 2 and reduced to 64 dimensions to enable efficient communication. The multi-scale feature dimensions used in Pyramid Fusion are [64, 128, 256]. For the camera-based setup, the detection range is defined as $x, y \in [-51.2\text{m}, 51.2\text{m}]$. The BEV feature map is subsequently downsampled by a factor of 2 and compressed to 64 dimensions to facilitate message transmission. The parameters of MSD Loss are $\alpha = 1.0,\ \beta = 1.0,\ \gamma = 0.1$.

For OPV2V, we trained end-to-end for 30 epochs using the Adam optimizer (learning rate 0.002) with multistep learning rate scheduling, a batch size of 1, on 8 NVIDIA GeForce RTX 4090 GPUs. For DAIR-V2X, we trained for 40 epochs with the same optimizer and scheduler, using a batch size of 2 on 8 RTX 4090 GPUs.

C. Quantitative Results

We evaluate the performance of various collaborative 3D object detection models in homogeneous settings using the OPV2V and DAIR-V2X datasets. Figure 4 presents visualization results for both datasets under camera-based and LiDAR-based conditions, highlighting WaveComm's ability to maintain high recall rates with minimal bandwidth usage by transmitting only the low-frequency component. Additionally, WaveComm exhibits superior accuracy, particularly in LiDAR experiments, where it achieves more precise object detections compared to baselines.

Table I provides a comprehensive quantitative comparison, demonstrating that our proposed ***WaveComm*** outperforms state-of-the-art methods, including F-Cooper [38], DiscoNet [17], AttFusion [34], V2X-ViT [8], CoBEVT [18], and HM-ViT [19], across both LiDAR-based and camera-based homogeneous collaboration tasks. Specifically, on the OPV2V dataset for camera-based perception, WaveComm achieves an AP50 of 0.681 and AP70 of 0.451, surpassing HM-ViT's 0.643 and 0.370 by approximately 5.9% and 21.9%, respectively, while reducing communication volume (Comm) from 22.0 to 19.0. In LiDAR-based OPV2V, WaveComm reaches an AP50 of 0.965 and AP70 of 0.926, improving over HM-ViT's 0.950 and 0.873 by 1.6% and 6.1%, with Comm lowered from 24.0 to 21.0.

On the DAIR-V2X dataset, for camera-based tasks, WaveComm attains AP30 of 0.274 and AP50 of 0.123, outperforming V2X-ViT's 0.198 and 0.057 by 38.4% and 115.8%, respectively, alongside a Comm reduction from 23.0 to 20.0. For LiDAR-based DAIR-V2X, it achieves AP30 of 0.831 and AP50 of 0.790, exceeding HM-ViT's 0.818 and 0.761 by 1.6% and 3.8%. These results underscore WaveComm's effectiveness in enhancing detection accuracy while significantly minimizing communication overhead, making it particularly suitable for resource-constrained collaborative perception scenarios.

D. Ablation Studies

To evaluate the effectiveness and synergy of the proposed method, we conducted three ablation experiments using the baseline model on the DAIR-V2X dataset in the LiDAR-based setting.

IDWT and Generator. From the results of the comparative experiments in Table II, it can be seen that the performance of using the wavelet generator to reconstruct is better than that of using only the IDWT module.

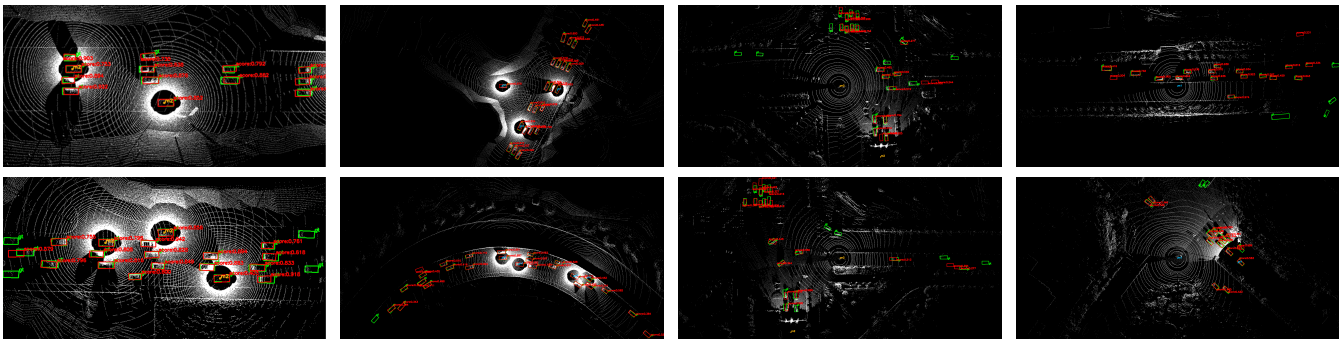

Fig. 4. Visualization of detection results on the OPV2V and DAIR-V2X datasets under both LiDAR-based and camera-based configurations. Green represents ground truth box, and Red represents predicted box.

TABLE II
ABLATION STUDY RESULTS OF THE IDWT AND GENERATOR

IDWT	Generator	AP30 ↑	AP50 ↑
✓		0.826	0.783
✓	✓	0.831	0.790

The generator enhances feature learning by mapping from a latent space and optimizing a loss function to produce richer, more expressive feature representations tailored to the target task, rather than merely restoring input features like the IDWT. While the IDWT's restored features serve as a supervisory signal, the generator does not fully rely on it; instead, by optimizing its loss function alongside this signal, it effectively recovers high-frequency information lost in compression, resulting in more accurate and robust feature representations.

Different Component. Ablation studies investigate the impact of combined transmission of different frequency components on the performance of collaborative 3D detection. In the Base setup, only the low-frequency component is input to the decoder for reconstruction. In the Add-Fuse setup, high-frequency components, processed through convolutional and linear layers, are added to the low-frequency component, and the fused result is then input to the decoder. In the Concat-Fuse setup, the low-frequency component is concatenated with the processed high-frequency components in the channel dimension, followed by fusion through an additional convolutional layer before being input to the decoder.

TABLE III
ABLATION STUDY RESULTS OF THE DIFFERENT COMPONENT

Component	AP30 ↑	AP50 ↑
Base	0.826	0.783
Add-Fuse	0.831	0.788
Concat-Fuse	0.829	0.783

Table III shows that both the Add-Fuse and Concat-Fuse methods achieve slight performance improvements over the Base method. However, these two methods introduce higher network complexity and memory usage. To strike a balance between performance improvement and resource efficiency, this paper chooses to adopt the Base method to avoid additional resource overhead while maintaining reasonable performance.

Multilevel Wavelet Transform. This ablation study evaluates the impact of multi-level wavelet transform on the performance of collaborative 3D detection. As shown in Table IV, varying the number of wavelet transform levels significantly affects detection accuracy. The 1-level transform achieves the highest performance, whereas performance progressively declines with higher levels. This is likely due to increased feature loss caused by excessive decomposition.

TABLE IV
ABLATION STUDY RESULTS OF THE MULTILEVEL WAVELET TRANSFORM

Multilevel	AP30 ↑	AP50 ↑	Comm
1-level	0.831	0.790	20.0
2-level	0.825	0.768	18.0
3-level	0.801	0.721	18.0

VI. CONCLUSIONS

In this paper, we propose WaveComm, a lightweight collaborative perception framework leveraging wavelet feature distillation to address excessive communication overhead. We design a wavelet-based compression–reconstruction mechanism: feature maps are decomposed into low- and high-frequency components, only compact low-frequency parts are transmitted, and a lightweight generator reconstructs missing details at the receiver. A multi-loss optimization further improves reconstruction fidelity under low communication cost. Experiments on multiple datasets show that WaveComm achieves strong detection accuracy with significantly reduced bandwidth. Future work will study WaveComm's robustness under more realistic V2X conditions, including heterogeneous agents and unreliable communication.

REFERENCES

[1] E. Arnold, M. Dianati, R. De Temple, and S. Fallah, "Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 1852–1864, 2020.

[2] 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 *European conference on computer vision*. Springer, 2020, pp. 605–621.

[3] S. Liu, C. Gao, Y. Chen, X. Peng, X. Kong, K. Wang, R. Xu, W. Jiang, H. Xiang, J. Ma *et al.*, "Towards vehicle-to-everything autonomous driving: A survey on collaborative perception," *arXiv preprint arXiv:2308.16714*, 2023.

[4] R. Xu, X. Xia, J. Li, H. Li, S. Zhang, Z. Tu, Z. Meng, H. Xiang, X. Dong, R. Song *et al.*, "V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 13 712–13 722.

[5] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan *et al.*, "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 361–21 370.

[6] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2020, pp. 4106–4115.

[7] R. Mao, H. Wu, Y. Jia, Z. Nan, Y. Sun, S. Zhou, D. Gündüz, and Z. Niu, "Diffcp: Ultra-low bit collaborative perception via diffusion model," *arXiv preprint arXiv:2409.19592*, 2024.

[8] 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 *European conference on computer vision*. Springer, 2022, pp. 107–124.

[9] J. Zhang, K. Yang, Y. Wang, H. Wang, P. Sun, and L. Song, "Ermvp: Communication-efficient and collaboration-robust multi-vehicle perception in challenging environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 575–12 584.

[10] D. Jin, Y. Zeng, and Y. Gong, "Bandwidth-efficient communication modelling for autonomous vehicle collaborative perception," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 6146–6155.

[11] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," *Advances in neural information processing systems*, vol. 35, pp. 4874–4886, 2022.

[12] Y. Hu, J. Peng, S. Liu, J. Ge, S. Liu, and S. Chen, "Communication-efficient collaborative perception via information filling with codebook," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 481–15 490.

[13] W. Su, L. Chen, Y. Bai, X. Lin, G. Li, Z. Qu, and P. Zhou, "What makes good collaborative views? contrastive mutual information maximization for multi-agent perception," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 16, 2024, pp. 17 550–17 558.

[14] 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 *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[15] H. Yu, Y. Tang, E. Xie, J. Mao, P. Luo, and Z. Nie, "Flow-based feature fusion for vehicle-infrastructure cooperative 3d object detection," *Advances in Neural Information Processing Systems*, vol. 36, pp. 34 493–34 503, 2023.

[16] Z. Zhang, Y. Wu, and H. Zhang, "Fast2comm: Collaborative perception combined with prior knowledge," *arXiv preprint arXiv:2505.00740*, 2025.

[17] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 541–29 552, 2021.

[18] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers," *arXiv preprint arXiv:2207.02202*, 2022.

[19] H. Xiang, R. Xu, and J. Ma, "Hm-vit: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 284–295.

[20] R. Wang, X. Gao, H. Xiang, R. Xu, and Z. Tu, "Cocmt: Communication-efficient cross-modal transformer for collaborative perception," *arXiv preprint arXiv:2503.13504*, 2025.

[21] Z. Wang, S. Xu, X. Zhuang, T. Xu, Y. Wang, J. Liu, Y. Chen, and Y.-Q. Zhang, "Coopdetr: A unified cooperative perception framework for 3d detection via object query," *arXiv preprint arXiv:2502.19313*, 2025.

[22] D. Yu, J. You, X. Pei, A. Qu, D. Wang, and S. Jia, "Which2comm: An efficient collaborative perception framework for 3d object detection," *arXiv preprint arXiv:2503.17175*, 2025.

[23] R. Gal, D. C. Hochberg, A. Bermano, and D. Cohen-Or, "Swagan: A style-based wavelet-driven generative model," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–11, 2021.

[24] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1689–1697.

[25] Y. Duan, F. Liu, L. Jiao, P. Zhao, and L. Zhang, "Sar image segmentation based on convolutional-wavelet neural network and markov random field," *Pattern Recognition*, vol. 64, pp. 255–267, 2017.

[26] T. Williams and R. Li, "Wavelet pooling for convolutional neural networks," in *International conference on learning representations*, 2018.

[27] M. Wolter, S. Lin, and A. Yao, "Neural network compression via learnable wavelet transforms," in *International Conference on Artificial Neural Networks*. Springer, 2020, pp. 39–51.

[28] S. E. Finder, Y. Zohav, M. Ashkenazi, and E. Treister, "Wavelet feature maps compression for image-to-image cnns," *Advances in Neural Information Processing Systems*, vol. 35, pp. 20 592–20 606, 2022.

[29] D. Rho, B. Lee, S. Nam, J. C. Lee, J. H. Ko, and E. Park, "Masked wavelet representation for compact neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 680–20 690.

[30] S. Xiang and Q. Liang, "Remote sensing image compression based on high-frequency and low-frequency components," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[31] J. Song, J. He, L. Yang, M. Feng, and K. Wang, "High frequency matters: Uncertainty guided image compression with wavelet diffusion," *arXiv preprint arXiv:2407.12538*, 2024.

[32] Y. Lu, Y. Hu, Y. Zhong, D. Wang, Y. Wang, and S. Chen, "An extensible framework for open heterogeneous collaborative perception," *arXiv preprint arXiv:2401.13964*, 2024.

[33] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[34] 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 *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2583–2589.

[35] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[36] R. Xu, Y. Guo, X. Han, X. Xia, H. Xiang, and J. Ma, "Opencda: an open cooperative driving automation framework integrated with co-simulation," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 1155–1162.

[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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[38] Q. Chen, X. Ma, S. Tang, J. Guo, Q. Yang, and S. Fu, "F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds," in *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 2019, pp. 88–100.