How2comm: Communication-Efficient and Collaboration-Pragmatic Multi-Agent Perception

Dingkang Yang[1,2†] Kun Yang[1†] Yuzheng Wang[1] Jing Liu[1]
Zhi Xu[1] Rongbin Yin[5] Peng Zhai[1,2*] Lihua Zhang[1,2,3,4*]

[1]Academy for Engineering and Technology, Fudan University
[2]Cognition and Intelligent Technology Laboratory (CIT Lab)
[3]Engineering Research Center of AI and Robotics, Ministry of Education
[4]AI and Unmanned Systems Engineering Research Center of Jilin Province
[5]FAW (Nanjing) Technology Development Company Ltd
{dkyang20, kungyang20, pzhai, lihuazhang}@fudan.edu.cn

1 Highlights of Our Contribution

We propose *How2comm*, a collaborative perception framework that seeks a trade-off between perception performance and communication bandwidth. The contributions of our How2comm are highlighted as follows:

- We propose a Mutual Information-aware Communication (MIC) mechanism to address the **communication redundancy** issue. On the one hand, MIC directs collaborators to share spatial-channel sparse, yet perceptually critical information through spatial-channel message filtering. On the other hand, the mutual information maximization supervision is introduced to mitigate the depletion of locally valuable semantics in the transmitted features.
- We design a Flow-guided Delay Compensation (FDC) strategy to address the **transmission delay** issue. FDC eliminates two-sided fusion error at the feature level due to the temporal asynchrony by predicting the future features of the collaborators. FDC provides a promising solution to mitigate the inevitable transmission delay in collaborative perception by a self-supervised motion estimation paradigm.
- We present a Spatio-Temporal Collaboration transFormer (STCFormer) to address the **collaboration heterogeneity** issue. STCFormer delivers a unified structure for pragmatic and robust collaboration by simultaneously modelling the temporal dynamics and spatial semantics among heterogeneous agents.

2 Details about the System Pipeline

Alg. 1 presents the system pipeline of the proposed multi-agent collaborative perception framework How2comm. We assume that there is a data transmission delay τ (*i.e.*, $t = t_0 + \tau$) between the ego agent (i) and the collaborators. The point clouds $\{\boldsymbol{X}_j^{t_0-k}, ..., \boldsymbol{X}_j^{t_0} \mid 1 \leq j \leq N, j \neq i\}$ of $N-1$ collaborators and the ego agent's historical frames $\{\boldsymbol{X}_i^{t_0}, ..., \boldsymbol{X}_i^{t}\}$ are defined as the system inputs, which are synchronized to the current coordinate system of the ego agent. As stated in the *main paper*, the primary goal of How2comm is to maximize the 3D object detection performance of the ego agent by aggregating the semantic information from the above system inputs under a communication budget B. Specifically, **FFN** is the feed-forward network, **TCA**, **ESA**, **CSA**, and **ALF** denote the

[†]Equal contributions.
[*]Corresponding authors.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Algorithm 1: System pipeline of How2comm

```
Input  :$\{\boldsymbol{X}_j^{t_0-k}, ..., \boldsymbol{X}_j^{t_0} \mid 1 \leq j \leq N, j \neq i\}$, $\{\boldsymbol{X}_i^{t_0}, ..., \boldsymbol{X}_i^{t} \mid t = t_0 + \tau\}$
# Each agent processes independently
# 1. Mutual Information-aware Communication
for each agent $j = 1, 2, ..., N$ do
    $\boldsymbol{F}_j^{t_0} = f_{enc}(\boldsymbol{X}_j^{t_0}) \in \mathbb{R}^{H\times W\times C}$;
    Obtain attention queries: $\mathcal{A}_{i,s}^{t_0} \in \mathbb{R}^{H\times W\times 1}$, $\mathcal{A}_{i,c}^{t_0} \in \mathbb{R}^{1\times 1\times C}$;
    if $j == i$ then
        Broadcast request queries: $\mathcal{R}_{i,s}^{t_0} = 1 - \mathcal{A}_{i,s}^{t_0}$, $\mathcal{R}_{i,c}^{t_0} = 1 - \mathcal{A}_{i,c}^{t_0}$;
    else
        $\mathcal{M}_j^{t_0} =$ Eq. (2)* $\in \{0, 1\}^{H\times W\times C}$;
        $\mathcal{O}_j^{t_0}$, $\mathcal{S}_j^{t_0} = f_{flow}([\boldsymbol{F}_j^{t_0-k}, ..., \boldsymbol{F}_j^{t_0}])$;
        Send $\tilde{\boldsymbol{F}}_j^{t_0} = \boldsymbol{F}_j^{t_0} \odot \mathcal{M}_j^{t_0}$ and $\{\mathcal{O}_j^{t_0}, \mathcal{S}_j^{t_0}\}$ to the $i$-th agent;
# For ego agent
# 2. Flow-guided Delay Compensation
$\boldsymbol{Z}_j^t = f_{warp}(\tilde{\boldsymbol{F}}_j^{t_0}, (t - t_0) \cdot \mathcal{O}_j^{t_0}) \odot \mathcal{S}_j^{t_0}$;
# 3. Spatio-Temporal Collaboration Transformer
$\mathcal{I}_i^t / \mathcal{I}_j^t = \sigma(\varphi_m(f_{gen}(\boldsymbol{F}_i^t / \boldsymbol{Z}_j^t))) \in [0, 1]^{H\times W}$;
$\boldsymbol{Z}_{j,e}^t = f_{sel}((1 - \mathcal{I}_i^t) \odot \mathcal{I}_j^t) \odot \boldsymbol{Z}_j^t$, $\boldsymbol{Z}_{j,c}^t = f_{sel}(\mathcal{I}_i^t \odot \mathcal{I}_j^t) \odot \boldsymbol{Z}_j^t$;
$\boldsymbol{H}_i^t = \mathbf{TCA}([\tilde{\boldsymbol{F}}_j^{t_0}, \boldsymbol{F}_i^{t_0}, ..., \boldsymbol{F}_i^t])$;
$\boldsymbol{F}_{i,e}^t = \mathbf{ESA}(\boldsymbol{H}_i^t, \{\boldsymbol{Z}_{j,e}^t \mid 1 \leq j \leq N, j \neq i\}) + \boldsymbol{H}_i^t$, $\boldsymbol{F}_{i,e}^t = \mathbf{FFN}(\boldsymbol{F}_{i,e}^t) + \boldsymbol{F}_{i,e}^t$;
$\boldsymbol{F}_{i,c}^t = \mathbf{CSA}(\boldsymbol{H}_i^t, \{\boldsymbol{Z}_{j,c}^t \mid 1 \leq j \leq N, j \neq i\}) + \boldsymbol{H}_i^t$, $\boldsymbol{F}_{i,c}^t = \mathbf{FFN}(\boldsymbol{F}_{i,c}^t) + \boldsymbol{F}_{i,c}^t$;
$\mathcal{F}_i^t = \mathbf{ALF}(\boldsymbol{F}_{i,e}^t, \boldsymbol{F}_{i,c}^t) \in \mathbb{R}^{H\times W\times C}$;
# 4. Detection Decoders
$\mathcal{Y}_{i,r}^t = f_{dec}^r(\mathcal{F}_i^t) \in \mathbb{R}^{H\times W\times 7}$, $\mathcal{Y}_{i,c}^t = f_{dec}^c(\mathcal{F}_i^t) \in \mathbb{R}^{H\times W\times 2}$.
```

proposed temporal cross-attention, exclusive spatial attention, common spatial attention, and adaptive late fusion components, respectively. **Eq.** $(\cdot)$* represents the equation in the *main paper*, σ and $\varphi_m(\cdot)$ denote the sigmoid activation and max pooling function, respectively. $\odot$ and $\cdot$ denote the point-wise multiplication and scalar multiplication, respectively. As Alg. 1 shows, How2comm leverages two tailored components to facilitate an efficient communication pattern and eliminate temporal asynchrony, respectively. Moreover, the proposed STCFormer component effectively aggregates the historical context clues and comprehensive collaborator information, leading to a more pragmatic collaboration. The output prediction results of the ego agent are produced based on the refined visual representation $\mathcal{F}_i^t$. Note that we present the processing of each agent in an iteration format only for clarity; these processes are conducted in parallel during the implementation.

3 Experiments

3.1 Implementation Details

The height and width resolution of the feature encoder $f_{enc}(\cdot)$ is set to 0.4 m. The extracted BEV feature map size is $(100,\ 252,\ 64)$ on the DAIR-V2X [11] dataset and $(100,\ 352,\ 64)$ on the V2XSet [9] and OPV2V [10] datasets.

Communication Volume. As stated in the *main paper*, the j-th agent, as a collaborator, employs the learned binary message filtering matrix $\mathcal{M}_j^{t_0} \in \{0, 1\}^{H\times W\times C}$ to obtain the sparse feature as $\tilde{\boldsymbol{F}}_j^{t_0} = \boldsymbol{F}_j^{t_0} \odot \mathcal{M}_j^{t_0} \in \mathbb{R}^{H\times W\times C}$ and sends $\tilde{\boldsymbol{F}}_j^{t_0}$ to the ego agent. Moreover, the feature flow $\mathcal{O}_j^{t_0} \in \mathbb{R}^{H\times W\times 2}$ and scale matrix $\mathcal{S}_j^{t_0} \in \mathbb{R}^{H\times W\times 1}$ are also transmitted. We count the message size by byte in the log scale with base 2 by formulating the communication volume as follows:

$$\log_2[(|\mathcal{M}_j^{t_0}| + 3HW) \times 32/8], \tag{1}$$

where $|\cdot|$ is the $L0$-norm operation counting non-zero elements. 32 represents that each element of the transmitted feature is the float32 data type, and 8 represents that byte is used as the metric. The vanilla calculation format of communication volume as [1] is applied for other feature compression or spatial filtering methods.

Bandwidth Limitation. The existing commercial V2V communication products based on 802.11p achieve transmission rates close to 6 MBps, while LTE-V2V channels can reach approximately 10 MBps [5]. Given the capturing frequency of 10 Hz for commonly used LiDAR [10], LTE-V2V communication products can transmit message sizes of about 1 MB in 100 ms. Therefore, we set a basic bandwidth limitation of ≈ 1 MB to simulate the restricted communication resources in real V2X/V scenarios and evaluate the performance of all models under this limitation setting.

3.2 Experimental Setup

Communication Graph. Following the settings of existing works [3, 4, 6, 9, 10], we identify an agent (*e.g.*, vehicle) as the ego agent and set other agents within the communication range as collaborators during the collaboration. The communication range is set to 70 m [1] in practice. With this design, our primary goal is to improve the perception ability of the ego agent by aggregating complementary information from its local historical observations and collaborator-shared features. In practice, the ego agent can broadcast its metadata (*e.g.*, poses, extrinsic, and sensor type) using the existing Vehicle-to-Vehicle/Everything communication channel. The transmission delay of the metadata is negligible due to the small message size. Upon receiving the metadata of the ego agent, the collaborators will first project their local point cloud observations under the ego coordinate system through the coordinate transformation. Moreover, to extract historical context clues efficiently, the preceding point cloud frames considered in our framework are synchronized to the current coordinate frame via the ego-motion compensation module [7].

Feature Encoder. For optimizing processing latency and memory usage, we employ the anchor-based method PointPillar [2] as the feature encoder $f_{enc}(\cdot)$ to extract visual representations from the captured point cloud frames. The encoder converts the raw point clouds into the stacked pillars and scatters them into a 2D pseudo-image. A feature pyramid network processes the pseudo-image and outputs the final Bird's-Eye-View (BEV) features. Given the point cloud $\boldsymbol{X}_i^t$ of i-th agent at timestamp t, the corresponding extracted BEV feature map is $\boldsymbol{F}_i^t = f_{enc}(\boldsymbol{X}_i^t) \in \mathbb{R}^{H\times W\times C}$, where H, W, C stand for the height, width, and channel. The network parameters of the feature encoder $f_{enc}(\cdot)$ are shared among all agents.

Selection Function. As shown in Eq. (2) in *the main paper*, the selection function $f_{sel}(\cdot)$ is introduced to select the salient spatial-channel elements and obtain the binary message filtering matrix $\mathcal{M}_j^{t_0}$. Specifically, the elements of the input matrix are first arranged in descending order. Considering the dynamic communication resources of the realistic V2V/X channel, $f_{sel}(\cdot)$ leverages the given bandwidth limitation to set a communication threshold, which determines the spatial-channel elements that will be transmitted. The matrix $\mathcal{M}_j^{t_0}$ is obtained by a filtering operation, where the corresponding elements in the input matrix greater than the communication threshold are set to 1 and 0 verses.

Time Embedding. Time embedding is presented to explicitly introduce the temporal relationships of input BEV features into the temporal cross-attention module. Following [12], we apply a linear layer to encode the relative time $[t-t_0, t-t_0, ..., 0] \in \mathbb{R}^{1\times(N\tau+1)}$ of the input features $[\tilde{\boldsymbol{F}}_j^{t_0}, \boldsymbol{F}_i^{t_0}, ..., \boldsymbol{F}_i^t]$ as a time embedding, whose shape is $(N\tau+1,\ 1,\ 1,\ C)$. Then, the input features are reshaped to the size of $(N\tau+1,\ H,\ W,\ C)$ and added with the learned time embedding across all spatial locations.

3.3 Qualitative Evaluation

Figure 1 shows more detection results to verify the effectiveness of How2comm in high-density traffic flow scenarios. We show the visual comparisons between How2comm and current SOTA methods, including V2X-ViT [9], CoBEVT [8], and Where2comm [1], under the collaboration noises and communication bandwidth constraints aligned with the *main paper*. Unsurprisingly, How2comm has fewer regression misalignments and fewer undetected objects, resulting in a more robust performance. In comparison, previous methods fail to tackle fast-moving objects and perceive remotely occluded regions. The indisputable evidence is that some ground truth bounding boxes are not well matched with their detection results.

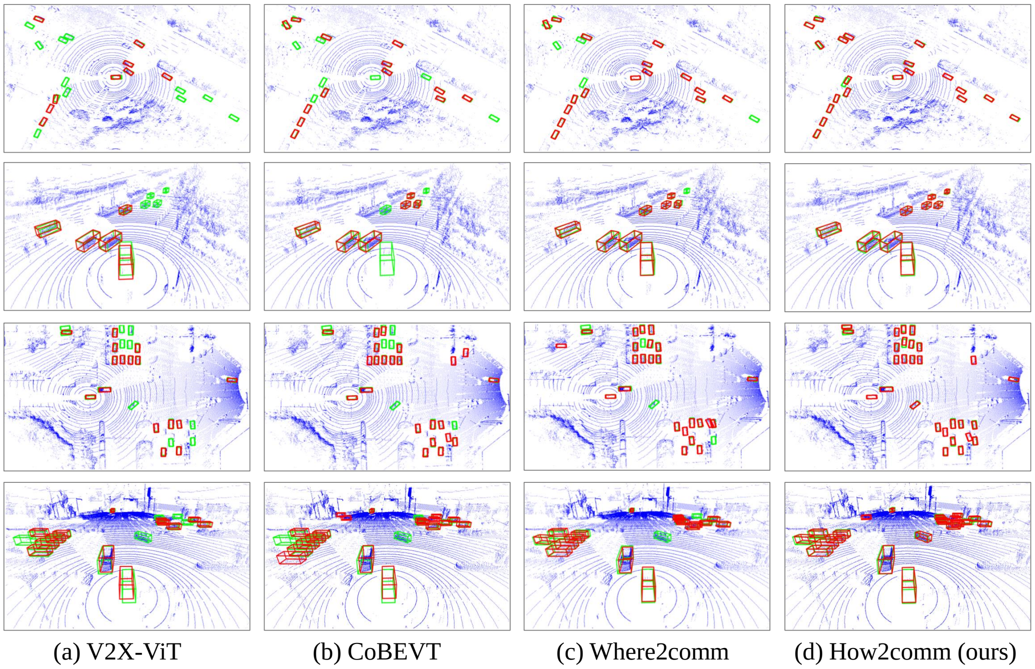

Figure 1: Detection visualization comparison in real-world scenarios. Green and red boxes denote the ground truth and detection results, respectively.

4 Broader Impacts

► The positive impact of this work is that How2comm can leverage vehicular communication technology to facilitate information sharing and enhance the vehicles' perception ability of the surrounding environment, reducing the incidence of traffic accidents.

► Additionally, the collaborative perception ability provided by How2comm can be used to monitor real-time traffic conditions and improve traffic efficiency.

► Collaborative perception systems may leak vehicle-related information (*e.g.*, locations and driving speed), violating the owner's privacy. Meanwhile, illegal personnel may invade the vehicle communication channel to attack the collaborative perception system, leading to the vehicles' misjudgment of the surrounding environment or malicious remote control.

► In summary, this work facilitates Vehicle-to-Vehicle/Everything (V2V/X)-oriented multi-agent communication applications in autonomous driving.

5 Discussion on the Realistic Limitations

As an emerging field, there are many challenges in building a robust multi-agent collaborative perception system. This paper addresses three of the biggest challenges, including communication redundancy, transmission delay, and collaboration heterogeneity, through tailored components. Although previous works have appropriately focused on these dilemmas, collaborative perception systems still have considerable room for improvements in the trade-off between perception performance and communication bandwidth. How2comm achieves superior performance with lower communication overheads in real V2V/X scenarios with limited communication capacity. We further discuss other realistic limitations and indicate future research directions to optimize our system.

► For the attack issue, How2comm could improve the robustness against potential adversarial attacks by introducing an adversarial training paradigm. In addition, by focusing on decoupled spatial regions and discovering perceptually critical information, How2comm is relatively less likely to be attacked.

► For the data bias issue, How2comm has demonstrated its preliminary generalization on real-world and simulated datasets with different configurations and sensor facilities. In future work, How2comm could disentangle causal relationships among variables in the collaborative perception pipeline with the assistance of causal-effect techniques, further enhancing its ability to generalize to more complex real-world scenarios.

► For the domain gap issue, How2comm could hopefully achieve sim2real adaptation and inter-agent adaptation by introducing an unsupervised sim2real domain bridging mechanism.

References

[1] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:4874–4886, 2022. 3

[2] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12697–12705, 2019. 3

[3] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:29541–29552, 2021. 3

[4] Guiyang Luo, Hui Zhang, Quan Yuan, and Jinglin Li. Complementarity-enhanced and redundancy-minimized collaboration network for multi-agent perception. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 3578–3586, 2022. 3

[5] Hang Qiu, Pohan Huang, Namo Asavisanu, Xiaochen Liu, Konstantinos Psounis, and Ramesh Govindan. Autocast: Scalable infrastructure-less cooperative perception for distributed collaborative driving. 2022. 3

[6] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 605–621. Springer, 2020. 3

[7] Pengxiang Wu, Siheng Chen, and Dimitris N Metaxas. Motionnet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11385–11395, 2020. 3

[8] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. In *Conference on Robot Learning (CoRL)*, 2022. 3

[9] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 107–124. Springer, 2022. 2, 3

[10] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022. 2, 3

[11] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui 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 (CVPR)*, pages 21361–21370, 2022. 2

[12] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh. Centerformer: Center-based transformer for 3d object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 496–513, 2022. 3