Where2comm: Efficient Collaborative Perception via Spatial Confidence Maps

Anonymous Author(s)
Affiliation
Address
email

Abstract

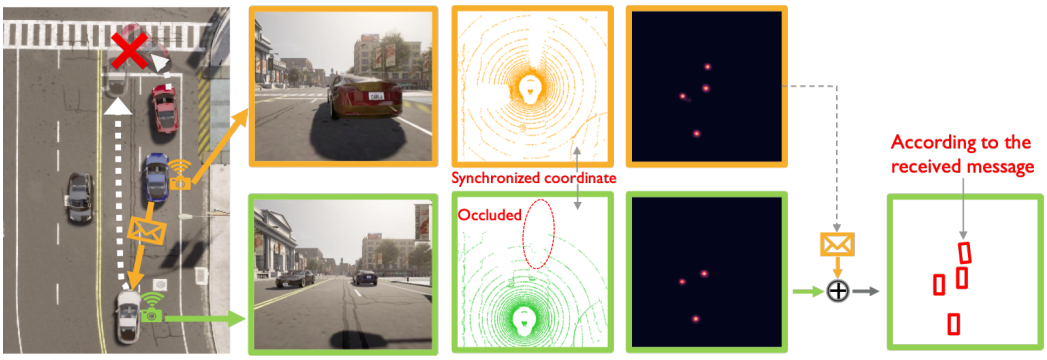
Multi-agent collaborative perception could significantly upgrade the perception performance by enabling agents to share complementary information with each other through communication. It inevitably results in a fundamental trade-off between perception performance and communication bandwidth. To tackle this bottleneck issue, we propose a spatial confidence map, which reflects the spatial heterogeneity of perceptual information. It empowers agents to only share spatially sparse, yet perceptually critical information, contributing to where to communicate. Based on this novel spatial confidence map, we propose `Where2comm`, a multi-round multi-agent collaborative perception framework. `Where2comm` has two distinct advantages: i) it uses less communication to achieve higher perception performance by focusing on perceptually critical areas; and ii) it can handle varying communication bandwidth by dynamically adjusting spatial areas involved in communication. To evaluate `Where2comm`, we consider 3D object detection tasks with two modalities (camera/LiDAR) and two agent types (cars/drones) on three datasets: OPV2V, V2X-Sim, and our original CoPerception-UAVs. `Where2comm` consistently achieves significant improvements over previous methods; for example, it achieves more than $10,000\times$ lower communication volume and still outperforms DiscoNet and V2VNet on OPV2V.

1 Introduction

Multi-agent collaborative perception enables multiple agents to share complementary perceptual information with each other, promoting more holistic perception. It provides a new direction to fundamentally overcome a number of inevitable limitations of single-agent perception, such as occlusion and long-range issues. Related methods and systems are desperately needed in a broad range of real-world applications, such as vehicle-to-everything-communication-aided autonomous driving [1, 2], multi-robot warehouse automation system [3, 4] and multi-UAVs (unmanned aerial vehicles) for search and rescue [5, 6]. To realize collaborative perception, recent works have contributed high-quality datasets [7, 8] and effective collaboration methods [9, 10, 2, 11].

In this emerging field, the current biggest challenge is how to optimize the trade-off between perception performance and communication bandwidth. Communication systems in real-world scenarios are always constrained that they can hardly afford huge communication consumption in real-time, such as passing complete raw observations or a large volume of features. Therefore, we cannot solely promote the perception performance without evaluating the expense of every bit of precious communication bandwidth. To achieve a better trade-off between performance and bandwidth, previous works put forth solutions from several perspectives. For example, When2com [9] considers a handshake mechanism which selects the most relevant collaborators; V2VNet [1] considers end-to-end-learning-based source coding; and DiscoNet [2] uses 1D convolution to achieve message compression. However, all previous works make a plausible, yet unnecessary assumption: once two

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) Scene in BEV (b) Image view (c) LiDAR view (d) Confidence in BEV (e) Detection in BEV

Figure 1: Collaborative perception could contribute to safety-critical scenarios, where the white car and the red car may collide due to occlusion. This collision could be avoided when the blue car can share a message about the red car's position. Such a message is spatially sparse, yet perceptually critical. Considering the precious communication bandwidth, each agent needs to speak to the point!

agents collaborate, they are obligated to share the perceptual information of all spatial areas *equally*. This unnecessary assumption can hugely waste the bandwidth as a large proportion of spatial areas may only contain irrelevant information for the perception task. Figure 1 illustrates such a spatial heterogeneity of perceptual information.

To fill this gap, we consider a novel spatial-confidence-aware communication strategy. The core idea is to enable a spatial confidence map for each agent, where each element reflects the perceptually critical level of a corresponding spatial area. Based on this map, agents decide which spatial area (where) to communicate about. That is, each agent offers spatially sparse, yet critical features to support other agents, and meanwhile requests complementary information from others through multi-round communication to perform efficient and mutually beneficial collaboration.

Following this strategy, we propose `Where2comm`, a novel multi-round multi-agent collaborative perception framework with the guidance of spatial confidence maps; see Fig. 2. `Where2comm` includes three key modules: i) a spatial confidence generator, which produces a spatial confidence map to indicate perceptually critical areas; ii) a spatial confidence-aware communication module, which leverages the spatial confidence map to decide *where* to communicate via novel message packing, and *who* to communicate via novel communication graph construction; and iii) a spatial confidence-aware message fusion module, which uses novel confidence-aware multi-head attention to fuse all messages received from other agents, upgrading the feature map for each agent.

`Where2comm` has two distinct advantages. First, it uses less communication to achieve higher perception performance by focusing on perceptually critical areas. Second, it adapts to various communication bandwidths and rounds, while previous models only handle one predefined communication bandwidth and a fixed number of rounds. To evaluate `Where2comm`, we consider the collaborative 3D object detection task on three datasets: V2X-Sim [7], OPV2V [8] and our original dataset CoPerception-UAVs. Our experiments cover two types of agents (cars and drones) and sensors (LiDAR and cameras). Results show that i) the proposed `Where2comm` consistently and significantly outperforms previous works in the performance-bandwidth trade-off across multiple datasets and modalities; and ii) `Where2comm` achieves better trade-off when the communication round increases.

2 Related Works

Multi-agent communication. The communication strategy in multi-agent systems has been widely studied [12]. Early works [13, 14, 15] often use predefined protocols or heuristics to decide how agents communicate with each other. However, it is difficult to generalize those methods to complex tasks. Recent works, thus, explore learning-based methods for complex scenarios. For example, CommNet [16] learns continuous communication in the multi-agent system. Vain [17] adopts the attention mechanism to help agents selectively fuse the information from others. Most of these previous works consider decision-making tasks and adopt reinforcement learning due to the lack of explicit supervision. In this work, we focus on the perception task. Based on direct perception supervision, we apply supervised learning to optimize the communication strategy in both trade-off perception ability and communication cost.

Table 1: Major components comparisons of collaborative perception systems.

Method	Venue	Message packing	Communication graph construction	Message fusing
When2com [9]	CVPR 2020	Full feature map	Handshake-based sparse graph	Attention per-agent
V2VNet [1]	ECCV 2020	Full feature map	Fully connected graph	Average per-agent
DiscoNet [2]	NeurIPS 2021	Full feature map	Fully connected graph	MLP-based attention per-location
OPV2V [8]	ICRA 2022	Full feature map	Fully connected graph	Self-attention per-location
Where2comm		*Confidence*-aware sparse feature map + request map	*Confidence*-aware sparse graph	*Confidence*-aware multi-head attention per-location

Collaborative perception. As a recent application of multi-agent communication systems to perception tasks, collaborative perception is still immature. To support this area of research, there is a surge of high-quality datasets (e.g., V2X-Sim [7], OpenV2V [8], Comap[18] and Dair-V2X[19]), as well as collaboration methods aimed for better performance-bandwidth trade-off (see comparisons in Table 1). When2com [9] proposes a handshake communication mechanism to decide *when* to communicate and create sparse communication graph. V2VNet [1] proposes multi-round message passing based on graph neural networks to achieve better perception and prediction performance. DiscoNet [2] adopts knowledge distillation to take the advantage of both early and intermediate collaboration. OPV2V [8] proposes a graph-based attentive intermediate fusion to improve perception performances. V2X-ViT [20] introduces a novel heterogeneous multi-agent attention module to fuse information across heterogeneous agents. In this work, we leverage the proposed spatial confidence map to promote more compact messages, more sparse communication graphs, and more comprehensive fusion, resulting in efficient and effective collaboration.

3 Problem Formulation

Consider N agents in the scene. Let $\mathcal{X}_i$ and $\mathcal{Y}_i$ be the observation and the perception supervision of the ith agent, respectively. The objective of collaborative perception is to achieve the maximized perception performance of all agents as a function of the total communication budge B and communication round K; that is,

$$\xi_\Phi(B,K) = \arg\max_{\theta,\mathcal{P}} \sum_{i=1}^{N} g\left(\Phi_\theta\left(\mathcal{X}_i, \{\mathcal{P}_{i\to j}^{(K)}\}_{j=1}^{N}\right), \mathcal{Y}_i\right), \ \text{ s.t. } \sum_{k=1}^{K}\sum_{i=1}^{N} |\mathcal{P}_{i\to j}^{(k)}| \leq B,$$

where $g(\cdot,\cdot)$ is the perception evaluation metric, Φ is the perception network with trainable parameter θ, and $\mathcal{P}_{i\to j}^{(k)}$ is the message transmitted from the ith agent to the jth agent at the kth communication round. Note that i) when $B = K = 0$, there is no collaboration and $\xi_\Phi(0,0)$ reflects the single-agent perception performance; ii) through optimizing the communication strategy and the network parameter, collaborative perception should perform well consistently at any communication bandwidth or round; and iii) we consider multi-round communication, where each agent serves as both a supporter (offering message to help others) and a requester (requesting messages from others).

In this work, we consider the perception task of 3D object detection and present three contributions: i) we make communication more efficient by designing compact messages and sparse communication graphs; ii) we boost the perception performance by implementing more comprehensive message fusion; iii) we enable the overall system to adapt to varying communication conditions by dynamically adjusting where and who to communicate.

4 `Where2comm`: Spatial Confidence-Aware Collaborative Perception System

This section presents `Where2comm`, a multi-round, multi-modality, multi-agent collaborative perception framework based on a spatial-confidence-aware communication strategy; see the overview in Fig. 2. `Where2comm` includes an observation encoder, a spatial confidence generator, the spatial confidence-aware communication module, the spatial confidence-aware message fusion module and a detection decoder. Among five modules, the proposed spatial confidence generator generates the spatial confidence map. Based on this spatial confidence map, the proposed spatial confidence-aware communication generates compact messages and sparse communication graphs to save communication bandwidth; and the proposed spatial confidence-aware message fusion module leverages informative spatial confidence priors to achieve better aggregation.

4.1 Observation encoder

The observation encoder extracts feature maps from the sensor data. `Where2comm` accepts single/multi-modality inputs, such as RGB images and 3D point clouds. This work adopts the feature representations in bird's eye view (BEV), where all agents project their individual perceptual information

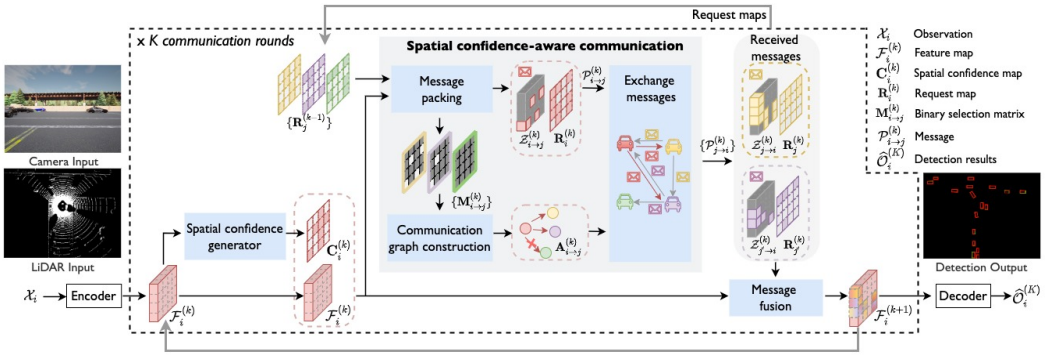


Figure 2: System overview. In `Where2comm`, spatial confidence generator enables the awareness of spatial heterogeneous of perceptual information, spatial confidence-aware communication enables efficient communication, and spatial confidence-aware message fusion boosts the performance.

to the same global coordinate system, avoiding complex coordinate transformations and supporting better shared cross-agent collaboration. For the ith agent, given its input $\mathcal{X}_i$, the feature map is $\mathcal{F}_i^{(0)} = \Phi_{\text{enc}}(\mathcal{X}_i) \in \mathbb{R}^{H\times W\times D}$, where $\Phi_{\text{enc}}(\cdot)$ is the encoder, the superscript 0 reflects that the feature is obtained before communication and H, W, D are its height, weight and channel. All agents share the same BEV coordinate system. For the image input, $\Phi_{\text{enc}}(\cdot)$ is followed by a warping function that transforms the extracted feature from front-view to BEV. For 3D point cloud input, we discretize 3D points as a BEV map and $\Phi_{\text{enc}}(\cdot)$ extracts features in BEV. The extracted feature map is output to the spatial confidence generator and the message fusion module.

4.2 Spatial confidence generator

The spatial confidence generator generates a spatial confidence map from the feature map of each agent. The spatial confidence map reflects the perceptually critical level of various spatial areas. Intuitively, for object detection task, the areas that contain objects are more critical than background areas. During collaboration, areas with objects could help recover the miss-detected objects due to the limited view; and background areas could be omitted to save the precious bandwidth. So we represent the spatial confidence map with the detection confidence map, where the area with high perceptually critical level is the area that contains an object with a high confidence score.

To implement, we use a detection decoder structure to produce the detection confidence map. Given the feature map at the kth communication round, $\mathcal{F}_i^{(k)}$, the corresponding spatial confidence map is

$$\mathbf{C}_i^{(k)} = \Phi_{\text{generator}}(\mathcal{F}_i^{(k)}) \in [0,1]^{H\times W}, \tag{1}$$

where the generator $\Phi_{\text{generator}}(\cdot)$ follows a detection decoder. Since we consider multi-round collaboration, `Where2comm` iteratively updates the feature map by aggregating information from other agents. Once $\mathcal{F}_i^{(k)}$ is obtained, (1) is triggered to reflect the perceptually critical level at each spatial location. The proposed spatial confidence map answers a crucial question that was ignored by previous works: for each agent, information at which spatial area is worth sharing with others. By answering this, it provides a solid base for efficient communication and effective message fusion.

4.3 Spatial confidence-aware communication

With the guidance of spatial confidence maps, the proposed communication module packs compact messages with spatially sparse feature maps and transmits messages through a sparsely-connected communication graph. Most existing collaboration perception systems [1, 2, 20] considers full feature maps in the messages and fully-connected communication graphs. To reduce the communication bandwidth without affecting perception, we leverage the spatial confidence map to select the most informative spatial areas in the feature map (where to communicate) and decide the most beneficial collaboration partners (who to communicate).

Message packing. Message packing determines what information should be included in the to-be-sent message. The proposed message includes: i) a request map that indicates at which spatial areas the agent needs to know more; and ii) a spatially sparse, yet perceptually critical feature map.

The request map of the ith agent is $\mathbf{R}_i^{(k)} = 1 - \mathbf{C}_i^{(k)} \in \mathbb{R}^{H\times W}$, negatively correlated with the spatial confidence map. The intuition is, for the locations with low confidence score, an agent is hard to

tell if there is really no objects or it is just caused by the limited information (e.g. occlusion). Thus, the low confidence score indicates there could be missing information at that location. Requesting information at these locations from other agents could improve the current agent's detection accuracy.

The spatially sparse feature map are selected based on each agent's spatial confidence map and the received request maps from others. Specifically, a binary selection matrix is used to represent each location is selected or not, where 1 denotes selected, and 0 elsewhere. For the message sent from the ith agent to the jth agent at the kth communication round, the binary selection matrix is

$$\mathbf{M}_{i\to j}^{(k)} = \begin{cases} \Phi_{\text{select}}(\mathbf{C}_i^{(k)}) \in \{0,1\}^{H\times W}, & k=0; \\ \Phi_{\text{select}}(\mathbf{C}_i^{(k)} \odot \mathbf{R}_j^{(k-1)}), \in \{0,1\}^{H\times W}, & k>0; \end{cases} \tag{2}$$

where $\odot$ is the element-wise multiplication, $\mathbf{R}_j^{(k-1)}$ is the request map from the jth agent received at the previous round, $\Phi_{\text{select}}(\cdot)$ is the selection function which targets to select the most critical areas conditioned on the input matrix, which represents the critical level at the certain spatial location. We implement $\Phi_{\text{select}}(\cdot)$ by selecting the locations where the largest elements at in the given input matrix conditioned on the bandwidth limit; optionally, a Gaussian filter could be applied to filter out the outliers and introduce some context. In the initial communication round, each agent selects the most critical areas from its own perspective as the request maps from other agents are not available yet; in the subsequent rounds, each agent also takes the partner's request into account, enabling more targeted communication. Then, the selected feature map is obtained as $\mathcal{Z}_{i\to j}^{(k)} = \mathbf{M}_{i\to j}^{(k)} \odot \mathcal{F}_i^{(k)} \in \mathbb{R}^{H\times W\times D}$, which provides spatially sparse, yet perceptually critical information.

Overall, the message sent from the ith agent to the jth agent at the kth communication round is $\mathcal{P}_{i\to j}^{(k)} = (\mathbf{R}_i^{(k)}, \mathcal{Z}_{i\to j}^{(k)})$. Note that i) $\mathbf{R}_i^{(k)}$ provides spatial priors to request complementary information for the ith agent's need in the next round; the feature map $\mathcal{Z}_{i\to j}^{(k)}$ provides supportive information for the ith agent's need in the this round. They together enable mutually beneficial collaboration; ii) since $\mathcal{Z}_{i\to j}^{(k)}$ is sparse, we only transmit non-zero features and corresponding indices, leading to low communication cost; and iii) the sparsity of $\mathcal{Z}_{i\to j}^{(k)}$ is determined by the binary selection matrix, which dynamically allocates the communication budget at various spatial areas based on their perceptual critical level, adapting to various communication conditions.

Communication graph construction. Communication graph construction targets to identify when and who to communicate to avoid unnecessary communication that wastes the bandwidth. Most previous works [1, 2, 8] consider fully-connected communication graphs. When2com [9] proposes a handshake mechanism, which uses similar global features to match partners. This is hard to interpret because two agents, which have similar global features, do not necessarily need information from each other. Different from all previous works, we provide an explicit design rationale: the necessity of communication between the ith and the jth agents is simply measured by the overlap between the information that the ith agent has and the information that the jth agent needs. With the help of the spatial confidence map and the request map, we construct a more interpretable communication graph.

For the initial communication round, every agent in the system is not aware of other agents yet. To activate the collaboration, we construct a fully-connected communication graph. Every agent will broadcast its message to the rest of the system. For the subsequent communication rounds, we examine if the communication between agent i and agent j is necessary based on the maximum value of the binary selection matrix $\mathbf{M}_{i\to j}^{(k)}$, i.e. if there is at least one patch is activated, then we regard the connection is necessary. Formally, let $\mathbf{A}^{(k)}$ be the adjacency matrix of the communication graph at the kth communication round, whose (i,j)th element is

$$\mathbf{A}_{i,j}^{(k)} = \begin{cases} 1, & k=0; \\ \max_{h\in\{0,1,..,H-1\}, w\in\{0,1,...,W-1\}} \left(\mathbf{M}_{i\to j}^{(k)}\right)_{h,w} \in \{0,1\}, & k>0; \end{cases}$$

where h, w index the spatial area, reflecting message passing from the ith agent to the jth agent. Given this sparse communication graph, agents can exchange messages with selected partners.

4.4 Spatial confidence-aware message fusion

Spatial confidence-aware message fusion targets to augment the feature of each agent by aggregating the received messages from the other agents. To achieve this, we adopt a transformer architecture, which leverages multi-head attention to fuse the corresponding features from multiple agents at each

individual spatial location. The key technical design is to include the spatial confidence maps of all the agents to promote cross-agent attention learning. The intuition is that, the spatial confidence map could explicitly reflect the perceptually critical level, providing a useful prior for attention learning.

Specifically, for the ith agent, after receiving the jth agent's message $\mathcal{P}_{j\to i}^{(k)}$, it could unpack to retrieve the feature map $\mathcal{Z}_{j\to i}^{(k)}$ and the spatial confidence map $\mathbf{C}_j^{(k)} = 1 - \mathbf{R}_j^{(k)}$. We also include the ego feature map in fusion and denote $\mathcal{Z}_{i\to i}^{(k)} = \mathcal{F}_i^{(k)}$ to make the formulation simple and consistent, where $\mathcal{Z}_{i\to i}^{(k)}$ might not be sparse. To fuse the features from the jth agent at the kth communication round, the cross-agent/ego attention weight for the ith agent is

$$\mathbf{W}_{j\to i}^{(k)} = \mathrm{MHA}_{\mathrm{W}}\left(\mathcal{F}_i^{(k)}, \mathcal{Z}_{j\to i}^{(k)}, \mathcal{Z}_{j\to i}^{(k)}\right) \odot \mathbf{C}_j^{(k)} \in \mathbb{R}^{H\times W}, \tag{3}$$

where $\mathrm{MHA}_{\mathrm{W}}(\cdot)$ is a multi-head attention applied at each individual spatial location, which outputs the scaled dot-product attention weight. Note that i) the proposed spatial confidence maps contributes to the attention weight, as the features with higher perceptually critical level are more preferred in the feature aggregation; ii) the cross-agent attention weight models the collaboration strength with a $H \times W$ spatial resolution, leading to more flexible information fusion at various spatial regions. Then, the feature map of the ith agent after fusing the messages in the kth communication round is

$$\mathcal{F}_i^{(k+1)} = \mathrm{FFN}\left(\sum_{j\in\mathcal{N}_i \bigcup \{i\}} \mathbf{W}_{j\to i}^{(k)} \odot \mathcal{Z}_{j\to i}^{(k)}\right) \in \mathbb{R}^{H\times W\times D},$$

where $\mathrm{FFN}(\cdot)$ is the feed-forward network and $\mathcal{N}_i$ is the neighbors of the ith agent defined in the communication graph $\mathbf{A}^{(k)}$. The fused feature $\mathcal{F}_i^{(k+1)}$ would serve as the ith agent's feature in the $(k+1)$th round. In the final round, we output $\mathcal{F}_i^{(k+1)}$ to the detection decoder to generate detections.

Sensor positional encoding. Sensor positional encoding represents the physical distance between each agent's sensor and its observation. It adopts a standard positional encoding function conditioned on the sensing distance and feature dimension. The features are summed up with the positional encoding of each location before inputting to the transformer.

Compared to existing fusion modules that do not use attention mechanism [1] or only use agent-level attentions [9], the per-location attention mechanism adopted by the proposed fusion emphasizes the location-specific feature interactions. It makes the feature fusion more targeted. Compared to the methods that also use the per-location attention-based fusion module[2, 8, 20], the proposed fusion module leverages multi-head attention with two extra priors, including spatial confidence map and sensing distances. Both assist attention learning to prefer high quality and critical features.

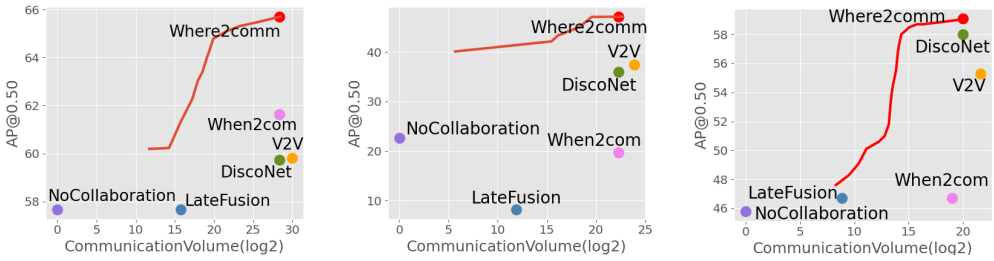
4.5 Detection decoder

The detection decoder decodes features into objects, including class and regression output. Given the feature map at the kth communication round $\mathcal{F}_i^{(k)}$, the detection decoder $\Phi_{\mathrm{dec}}(\cdot)$ generate the detections of ith agent by $\widehat{\mathcal{O}}_i^{(k)} = \Phi_{\mathrm{dec}}(\mathcal{F}_i^{(k)}) \in \mathbb{R}^{H\times W\times 7}$, where each location of $\widehat{\mathcal{O}}_i^{(k)}$ represents a rotated box with class $(c, x, y, h, w, \cos\alpha, \sin\alpha)$, denoting class confidence, position, size and angle. The objects are the final output of the proposed collaborative perception system. Note that $\widehat{\mathcal{O}}_i^{(0)}$ denotes the detections without collaboration.

4.6 Training details and loss functions

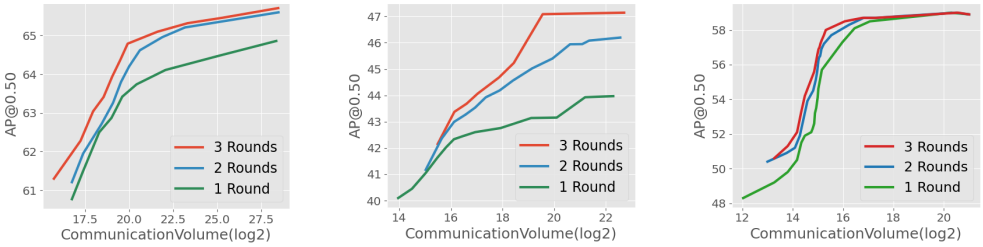
To train the overall system, we supervise two tasks: spatial confidence generation and object detection at each round. As mentioned before, the functionality of the spatial confidence generator is the same as the classification in the detection decoder. To promote parameter efficiency, our spatial confidence generator reuses the parameters of the detection decoder. For the multi-round settings, each round is supervised with one detection loss, the overall loss is $L = \sum_{k=0}^{K}\sum_{i}^{N} L_{\mathrm{det}}\left(\widehat{\mathcal{O}}_i^{(k)}, \mathcal{O}_i\right)$, where $\mathcal{O}_i$ is the ith agent's ground-truth objects, L_{det} is the detection loss [21].

Training strategy for multi-round setting. To adapt to multi-round communication and dynamic bandwidth, we train the model under various communication settings with curriculum learning strategy [22]. We first gradually increase the communication bandwidth and round; and then, randomly sample bandwidth and round to promote robustness. Through this training strategy, a single model can perform well at various communication conditions.



(a) CoPerception-UAVs (b) OPV2V (c) V2X-Sim

Figure 3: `Where2comm` achieves consistently superior performance-bandwidth trade-off on all the three collaborative perception datasets, e.g, `Where2comm` achieves *5,000* times less communication volume and still outperforms When2com on CoPerception-UAVs dataset. The entire red curve comes from a single `Where2comm` model evaluated at varying bandwidths.



(a) CoPerception-UAVs (b) OPV2V (c) V2X-Sim

Figure 4: More communication rounds continuously improve performance-bandwidth trade-off.

5 Experimental Results

Our experiments covers three datasets, two types of agents (cars and drones) and two types of sensors (LiDAR and cameras). Specifically, we conduct camera-only 3D object detection in the setting of V2X-communication aided autonomous driving on OPV2V dataset [8], camera-only 3D object detection in the setting of drone swarm on the proposed CoPerception-UAVs dataset, and LiDAR-based 3D object detection on V2X-Sim dataset [7]. The detection results are evaluated by Average Precision (AP) at Intersection-over-Union (IoU) threshold of 0.50 and 0.70. The communication results count the message size by byte in log scale with base 2. To compare communication results straightforward and fair, we do not consider any extra data/feature/model compression.

5.1 Datasets and experimental settings

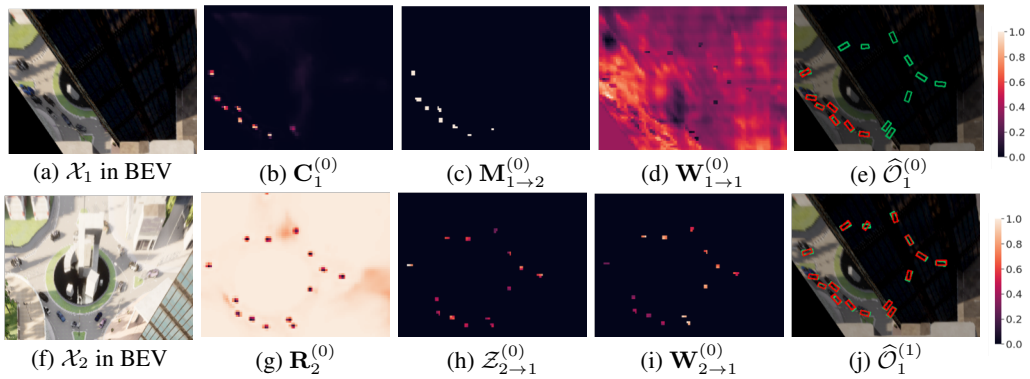
OPV2V. OPV2V [8] is a vehicle-to-vehicle collaborative perception dataset, co-simulated by OpenCDA [8] and Carla [23]. It includes 12K frames of 3D point clouds and RGB images with 230K annotated 3D boxes. The perception range is 40m$\times$40m. For camera-only 3D object detection task on OPV2V, we implement the detector following CADDN [24]. The input front-view image size is $(416, 160)$. The front-view input feature map is transformed to BEV with resolution 0.5m/pixel.

V2X-Sim. V2X-Sim [7] is a vehicle-to-everything collaborative perception dataset, co-simulated by SUMO [25] and Carla, including 10K frames of 3D LiDAR point clouds and 501K 3D boxes. The perception range is 64m$\times$64m. For LiDAR-based 3D object detection task, our detector follows MotionNet [26]. We discretize 3D points into a BEV map with size $(256, 256, 13)$ and the resolution is 0.4m/pixel in length and width, 0.25m in height.

CoPerception-UAVs. To enrich the collaborative perception datasets, we consider the swarm of unmanned aerial vehicles (UAV) and propose a UAV-swarm-based collaborative perception dataset: CoPerception-UAVs, co-simulated by AirSim [27] and Carla [23], including 131.9K aerial images and 1.94M 3D boxes. The perception range is 200m$\times$350m. For the camera-only 3D object detection task on CoPerception-UAVs, our detector follows the CenterNet [21] with DLA-34 [28] backbone. The input aerial image size is $(800, 450)$. The aerial-view input feature map is transformed to BEV with the resolution of 0.25m/pixel, and the size is $(192, 352)$; see more details in the appendix.

5.2 Quantitative evaluation

Benchmark comparison. Fig. 3 compares the proposed `Where2comm` with the previous methods in terms of the trade-off between detection performance (AP@IoU=0.50) and communication bandwidth.



(a) $\mathcal{X}_1$ in BEV (b) $\mathbf{C}_1^{(0)}$ (c) $\mathbf{M}_{1\to2}^{(0)}$ (d) $\mathbf{W}_{1\to1}^{(0)}$ (e) $\widehat{\mathcal{O}}_1^{(0)}$

(f) $\mathcal{X}_2$ in BEV (g) $\mathbf{R}_2^{(0)}$ (h) $\mathcal{Z}_{2\to1}^{(0)}$ (i) $\mathbf{W}_{2\to1}^{(0)}$ (j) $\widehat{\mathcal{O}}_1^{(1)}$

Figure 5: Visualization of collaboration between Drone 1 and Drone 2 on CoPerception-UAVs dataset, including spatial confidence map ($\mathbf{C}_1^{(0)}$), selection matrix ($\mathbf{M}_{1\to2}^{(0)}$), message ($\{\mathbf{R}_2^{(0)}, \mathcal{Z}_{2\to1}^{(0)}\}$) in the communication module, attention weight in the fusion module ($\mathbf{W}_{1\to1}^{(0)}$,$\mathbf{W}_{2\to1}^{(0)}$), and Drone 1's detection results before ($\widehat{\mathcal{O}}_1^{(0)}$) and after ($\widehat{\mathcal{O}}_1^{(1)}$) collaboration. Green and red boxes denote ground-truth and detection, respectively. The objects occluded by a tall building can be detected through transmitting spatially sparse, yet perceptually critical message.

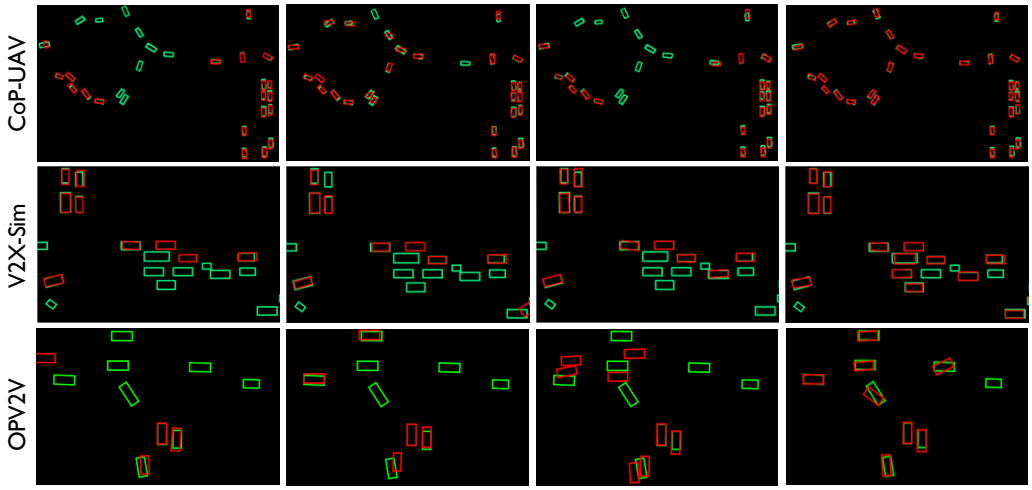
We consider single-agent detection without collaboration ($\widehat{\mathcal{O}}_i^{(0)}$), When2com [9], V2VNet [1], DiscoNet [2] and late fusion, where agents directly exchange the detected 3D boxes. The red curve comes from a single `Where2comm` model evaluated at varying bandwidths. We see that the proposed `Where2comm`: i) achieves a far-more superior perception-communication trade-off across all the communication bandwidth choices and various collaborative perception tasks, including camera-only 3D object detection from aerial view and car front view, and LiDAR-based 3D object detection; ii) outperforms previous state-of-the-arts, improves the SOTA performance by 6.62% on CoPerception-UAVs, 25.81% on OPV2V, and 1.9% on V2X-Sim; iii) achieves the same detection performance of previous state-of-the-arts with extremely less communication volume: 5128 times less on CoPerception-UAVs, more than 10,000 times less on OPV2V, and 55 times less on V2X-Sim.

Multi-round evaluation. Fig. 4 presents the performances of `Where2comm` at communication rounds ranging from 1 to 3. Each curve comes from a single `Where2comm` model with a certain communication round evaluated at varying bandwidths. Results show that 1 communication round is good, more rounds are even better. Multi-round communication steadily improves the performance-bandwidth trade-off across all three datasets, reflecting its effectiveness and robustness. This encourages the agents to actively collaborate without worrying the performance degradation. This also validates that `Where2comm` can well work at various communication bandwidths and rounds.

5.3 Qualitative evaluation

Visualization of spatial confidence map. Fig. 5 illustrates how `Where2comm` is empowered by the proposed spatial confidence map. In the scene, Drone 1's view is occluded by a tall building. With Drone 2's help, Drone 1 is able to detect through occlusion. Fig. 5 (a-d) shows Drone 1's observation, spatial confidence map (1), binary selection matrix (2), and ego attention weight (3). Fig. 5 (f-h) shows Drone 2's observation and message sent to Drone 1, including the request map (opposite of confidence map) and the sparse feature map, achieving efficient communication. Fig. 5 (i) shows the attention weight for Drone 1 to fuse Drone 2's messages, which is sparse, yet highlights the objects' positions. Fig. 5 (e) and (j) compares the detection results before and after the collaboration with Drone 2. We see that the proposed spatial confidence map contributes to spatially sparse, yet perceptually critical message, which effectively helps Drone 1 detect occluded objects.

Visualization of detection results. Fig. 6 shows that compared to *No Collaboration*, *When2com* and *DiscoNet*, `Where2comm` is able to achieves more complete and accurate detection results. Especially, *When2com* precisely detects extremely small objects in aerial view on CoPerception-UAVs dataset. The reason is that *When2com* employs a scalar to denote the agent-to-agent attention, which cannot distinguish which spatial area is more informative; *DiscoNet* employs a MLP-based fusion weight learning, which cannot well capture the complex collaboration attention; while `Where2comm` can zoom in to critical spatial areas in a cell-level resolution and leverage the spatial confidence map and sensing distances as priors to achieve more comprehensive fusion.



(a) No Collaboration (b) When2com (c) DiscoNet (d) Where2comm

Figure 6: `Where2comm` qualitatively outperforms the state-of-the-art methods in CoPerception-UAVs, V2X-Sim and OPV2V datasets. Green and red boxes denote ground-truth and detection, respectively.

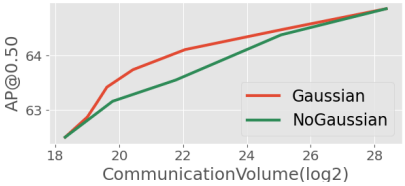


Figure 7: Selection matrix ablation study. Applying Gaussian filter improves performance.

Table 2: Fusion component ablation study. Multi-head attention (MHA), sensor positional encoding (SPE) and spatial confidence map (SCM) all improves the performances. Results are reported in AP@0.50/AP@0.70.

MHA	SPE	SCM	OPV2V	CoPerception-UAVs	V2X-Sim
			34.96/13.92	63.48/44.23	51.2/45.7
✓			38.75/13.28	63.99/44.46	57.3/50.8
✓	✓		39.82/16.43	64.34/46.86	59.1/52.0
✓	✓	✓	**47.30/19.30**	**64.83/47.62**	**59.1/52.2**

5.4 Ablation studies

Effect of Gaussian filter in perceptually critical area selection. Fig. 7 compares two versions of the selection matrix (2) with and without Gaussian filter. We see that applying Gaussian filter improves the overall performance. The reason is that: i) Gaussian filter could help filter out the outliers in the input map, selecting more robust critical regions; ii) it considers the context, benefiting the independent feature selection at each certain location by providing more information.

Effect of components in spatial confidence-aware message fusion. Tab. 2 assesses the effectiveness of the proposed fusion with two priors. We see that: i) per-location multi-head attention (MHA) outperforms the vanilla attention by 10.84% on OPV2V on AP@0.50, because MHA leverages information from multiple heads, better capturing cross-agent attention; and ii) As two informative priors, both sensing position encoding (SPE) and spatial confidence map (SCM) can consistently improve the performance. Especially, the version with all three designs improves the detection performance by 22.06% on OPV2V on AP@0.50.

6 Conclusion and limitation

We propose a novel multi-round collaborative perception framework with the guidance of spatial confidence maps, named `Where2comm`. The core idea is to enable a spatial confidence map for each agent, so that they can decide where to communicate with whom, and whose information to aggregate. Each agent offers spatially sparse, yet critical features to support other agents, and meanwhile, request the complementary information from others in multi-round communication, promoting bandwidth-efficiency. Comprehensive experiments covering multi-type agents and multi-modality inputs show that `Where2comm` achieves far superior performance-bandwidth trade-off.

Limitation and future work. The current work focuses on perceptually critical spatial areas. In future, we plan to expand a similar idea to the temporal dimension and determine critical time stamps. We will further reduce the cost by exploring when to communicate.

References

[1] 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 *European Conference on Computer Vision*, pages 605–621. Springer, 2020.

[2] 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*, 34, 2021.

[3] Zhi Li, Ali Vatankhah Barenji, Jiazhi Jiang, Ray Y Zhong, and Gangyan Xu. A mechanism for scheduling multi robot intelligent warehouse system face with dynamic demand. *Journal of Intelligent Manufacturing*, 31(2):469–480, 2020.

[4] Michela Zaccaria, Mikhail Giorgini, Riccardo Monica, and Jacopo Aleotti. Multi-robot multiple camera people detection and tracking in automated warehouses. In *2021 IEEE 19th International Conference on Industrial Informatics (INDIN)*, pages 1–6. IEEE, 2021.

[5] Jürgen Scherer, Saeed Yahyanejad, Samira Hayat, Evsen Yanmaz, Torsten Andre, Asif Khan, Vladimir Vukadinovic, Christian Bettstetter, Hermann Hellwagner, and Bernhard Rinner. An autonomous multi-uav system for search and rescue. In *Proceedings of the First Workshop on Micro Aerial Vehicle Networks, Systems, and Applications for Civilian Use*, pages 33–38, 2015.

[6] Ebtehal Turki Alotaibi, Shahad Saleh Alqefari, and Anis Koubaa. Lsar: Multi-uav collaboration for search and rescue missions. *IEEE Access*, 7:55817–55832, 2019.

[7] Yiming Li, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: A virtual collaborative perception dataset for autonomous driving. *ArXiv*, abs/2202.08449, 2022.

[8] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Liu, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. *ICRA*, 2022.

[9] Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 4106–4115, 2020.

[10] Yen-Cheng Liu, Junjiao Tian, Chih-Yao Ma, Nathan Glaser, Chia-Wen Kuo, and Zsolt Kira. Who2com: Collaborative perception via learnable handshake communication. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6876–6883. IEEE, 2020.

[11] Yang Zhou, Jiuhong Xiao, Yue Zhou, and Giuseppe Loianno. Multi-robot collaborative perception with graph neural networks. *IEEE Robotics and Automation Letters*, 2022.

[12] Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. Learning when to communicate at scale in multiagent cooperative and competitive tasks. *arXiv preprint arXiv:1812.09755*, 2018.

[13] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337, 1993.

[14] Faisal Qureshi and Demetri Terzopoulos. Smart camera networks in virtual reality. *Proceedings of the IEEE*, 96(10):1640–1656, 2008.

[15] Yiming Li, Bir Bhanu, and Wei Lin. Auction protocol for camera active control. In *2010 IEEE International Conference on Image Processing*, pages 4325–4328. IEEE, 2010.

[16] Sainbayar Sukhbaatar, Rob Fergus, et al. Learning multiagent communication with backpropagation. *Advances in neural information processing systems*, 29, 2016.

[17] Yedid Hoshen. Vain: Attentional multi-agent predictive modeling. *Advances in Neural Information Processing Systems*, 30, 2017.

[18] Y Yuan and M Sester. Comap: A synthetic dataset for collective multi-agent perception of autonomous driving. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:255–263, 2021.

[19] 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. *arXiv preprint arXiv:2204.05575*, 2022.

[20] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. *arXiv preprint arXiv:2203.10638*, 2022.

[21] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019.

[22] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[23] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[24] Cody Reading, Ali Harakeh, Julia Chae, and Steven L. Waslander. Categorical depth distributionnetwork for monocular 3d object detection. *CVPR*, 2021.

[25] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker. Recent development and applications of sumo-simulation of urban mobility. *International journal on advances in systems and measurements*, 5(3&4), 2012.

[26] Pengxiang Wu, Siheng Chen, and Dimitris N. Metaxas. Motionnet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11382–11392, 2020.

[27] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and service robotics*, pages 621–635. Springer, 2018.

[28] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 5.1 and the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5.1 and the supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We have not repeated experiments many times to get error bars since experiments of 3d object detection on large scale datasets is time-consuming.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 5.1 and the supplemental material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 5.1.
   (b) Did you mention the license of the assets? [Yes] See Section 5.1.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Section 5.1 and the supplemental material.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We generate the data using the open-sourced tool and the owners consent to all the public use for research purposes.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Our data is synthesized using the open-sourced tool, so there is no real-world personally identifiable information, nor offensive content.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]