Where2comm: Efficient Collaborative Perception via Spatial Confidence Maps

Anonymous Author(s)
Affiliation
Address
`email`

1 Appendix

1.1 Highlights of our contribution

To sum up, our contributions are:

• We propose a novel fine-grained spatial-aware communication strategy, where each agent can decide where to communicate and pack messages only related to the most perceptually critical spatial areas. This strategy not only enables more precise support for other agents, but also more targeted request from other agents in multi-round communication.

• We propose `Where2comm`, a novel collaborative perception framework based on the spatial-aware communication strategy. With the guidance of the proposed spatial confidence map, `Where2comm` leverages novel message packing and communication graph learning to achieve lower communication bandwidth, and adopts confidence-aware multi-head attention to reach better perception performance.

• We conduct extensive experiments to validate `Where2comm` achieves state-of-the-art performance-bandwidth trade-off on multiple challenging datasets across views and modalities.

1.2 Detailed information about the system pipeline

Alg. 1 presents the pipeline of our multi-round spatial confidence-aware collaborative perception system.

1.3 Detailed information about the module design

Spatial confidence-aware message packing. Fig. 1 presents the detail about the spatial confidence-aware message packing module. For the message from agent i to agent j at kth communication round, the module takes the spatial confidence map $\mathbf{C}_i^{(k)}$ of agent i and the request map $\mathbf{R}_j^{(k-1)}$ of agent j as input, and outputs the message $\mathcal{P}_{i\to j}^{(k)}$ including the masked feature map $\mathcal{Z}_{i\to j}^{(k)}$ and the request map of agent i.

Spatial confidence-aware communication graph construction. Fig. 2 presents the comparisons on the communication graph with previous works. *Fully connected* versus *agent-level partially connected* versus ours *spatial-decouple partially connected* communication. *Fully connected* communication results in a large amount of bandwidth usage, growing on the order of $O(N^2)$, where N is the number of agents in a network. *Agent-level partially connected* communication prune irrelevant connections between agents while may erroneously sever the information connection. *Spatial-decouple partially connected* communication could further flexibly prune irrelevant connections per-location and can substantially reduce the overall network complexity.

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

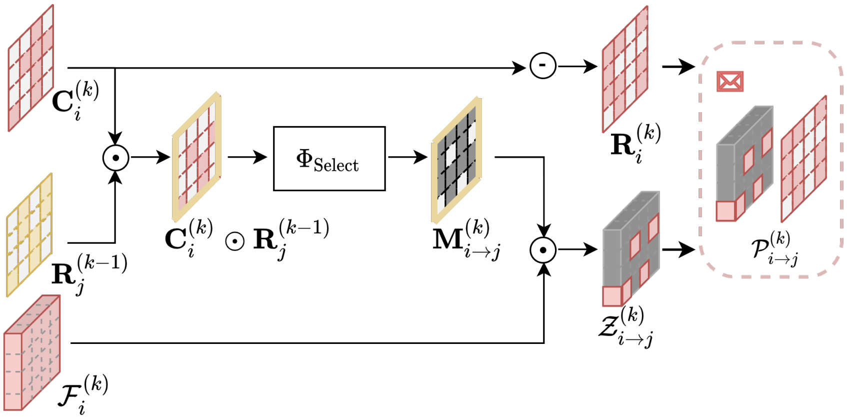

$\mathbf{C}_i^{(k)}$
$\mathbf{R}_j^{(k-1)}$
$\mathcal{F}_i^{(k)}$
$\mathbf{C}_i^{(k)} \odot \mathbf{R}_j^{(k-1)}$
Φ_{Select}
$\mathbf{M}_{i\to j}^{(k)}$
$\mathbf{R}_i^{(k)}$
$\mathcal{Z}_{i\to j}^{(k)}$
$\mathcal{P}_{i\to j}^{(k)}$

Figure 1: Spatial confidence-aware message packing module. $\odot$ denotes point-wise multiplication, $\ominus$ denotes point-wise minus by a matrix with the same shape as the input and filled with 1. Best viewed in color. Grey denotes the location being filled with zeros for the binary selection matrix $\mathbf{M}_{i\to j}^{(k)}$ and the feature map $\mathcal{Z}_{i\to j}^{(k)}$.

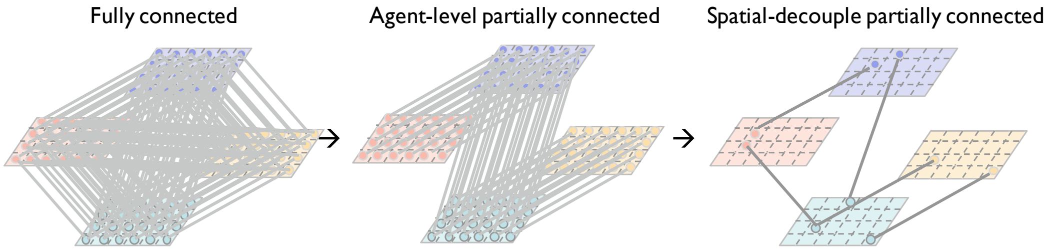

Fully connected
Agent-level partially connected
Spatial-decouple partially connected

Figure 2: Spatial confidence-aware communication graph construction module. We spatially decouple the full feature map, and could flexibly involve the informative spatial areas in the communication. This *Spatial-decouple partially connected* communication could further flexibly prune irrelevant connections per-location and is more bandwidth-efficient.

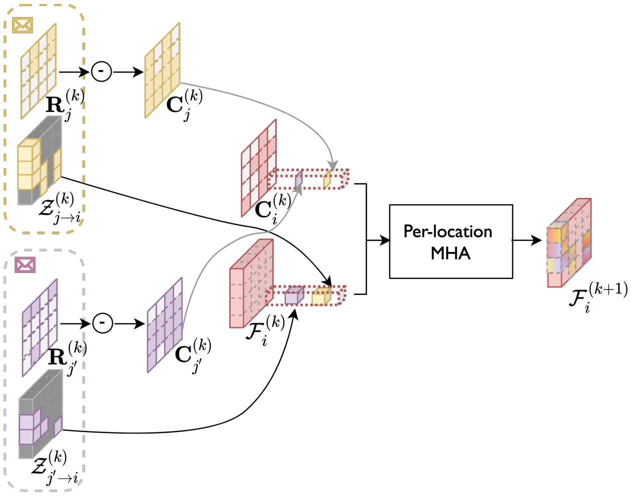


Figure 3: Spatial confidence-aware message fusion module. Each agent attentively augments the features with the received messages at each location. And the per-location multi-head attention are separately operated at each location in parallel, it takes the features and the corresponding confidence scores as input, and outputs the augmented features.

Algorithm 1 Multi-round spatial confidence-aware collaborative perception system

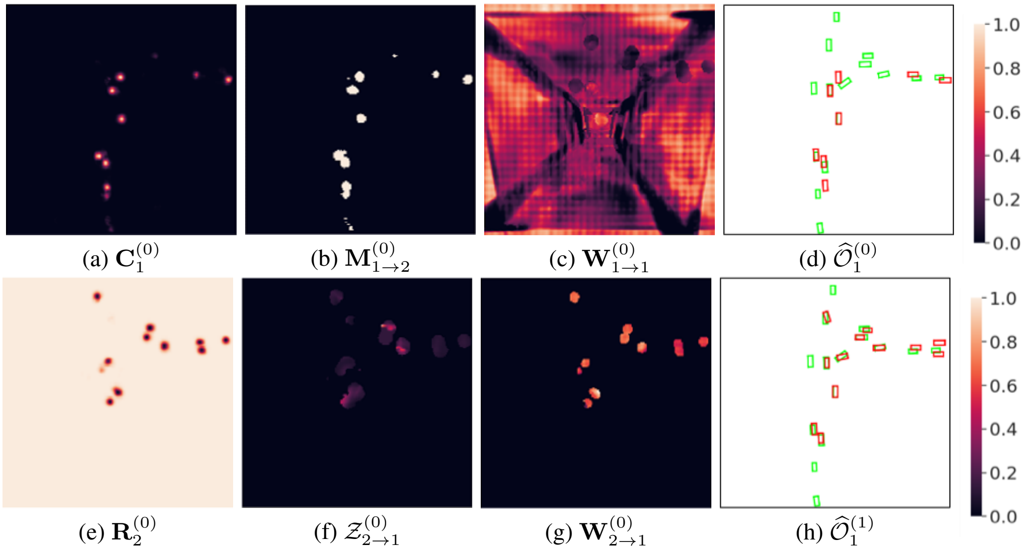
```
Define $N$ as the number of agents , $K$ as communication round
 # Initialization
for $i = 1, 2, \dots, N$, do
    $\mathcal{F}_i^{(0)} = \Phi_{\text{enc}}(\mathcal{X}_i) \in \mathbb{R}^{H\times W\times D}$    ▷ Extract intermediate feature
end for
for $k = 0, 1, \dots, K-1$, do
    for $i = 1, 2, \dots, N$, do  #  Each agent is computing individually
        $\mathbf{C}_i^{(k)} = \Phi_{\text{generator}}(\mathcal{F}_i^{(k)}) \in \mathbb{R}^{H\times W}$    ▷ Generate spatial confidence map
        for $j = 1, 2, \dots, N$, do
             # Message packing
            $\mathbf{R}_i^{(k)} = 1 - \mathbf{C}_i^{(k)} \in \mathbb{R}^{H\times W}$    ▷ Pack request map
            if $k = 0$ then
                $\mathbf{M}_{i\to j}^{(k)} = \Phi_{\text{select}}(\mathbf{C}_i^{(k)}) \in \{0,1\}^{H\times W}$    ▷ Select critical areas
            else
                $\mathbf{M}_{i\to j}^{(k)} = \Phi_{\text{select}}(\mathbf{C}_i^{(k)} \odot \mathbf{R}_j^{(k-1)}) \in \{0,1\}^{H\times W}$    ▷ Select requested areas
            end if
            $\mathcal{Z}_{i\to j}^{(k)} = \mathbf{M}_{i\to j}^{(k)} \odot \mathcal{F}_i^{(k)} \in \mathbb{R}^{H\times W\times D}$    ▷ Pack spatially sparse features
             # Communication graph learning
            if $k = 0$ then
                $\mathbf{A}_{i\to j}^{(k)} = 1$    ▷ Broadcast critical features and request
            else
                $\mathbf{A}_{i\to j}^{(k)} = \max_{h,w}\left(\mathbf{M}_{i\to j}^{(k)}\right)_{h,w} \in \{0,1\}$    ▷ Communicate only when necessary
            end if
        end for
         # Communication
        Send $P_{i\to j} = \left(\mathcal{Z}_{i\to j}^{(k)}, \mathbf{R}_i^{(k)}\right)$ to other agents
        Receive $\{P_{j\to i} = \left(\mathcal{Z}_{j\to i}^{(k)}, \mathbf{R}_j^{(k)}\right), j \neq i\}$ from other agents
         # Message fusion
        $\mathcal{F}_i^{(k+1)} = f_{\text{fuse}}\left(\mathcal{F}_i^{(k)}, \{(\mathcal{Z}_{j\to i}^{(k)}, \mathbf{R}_j^{(k)}), j = 1, 2, ..., N\}\right) \in \mathbb{R}^{H\times W\times D}$
    end for
    Store $\mathcal{F}_i^{(k+1)}$ and $\{\mathbf{R}_j^{(k)}, j \neq i\}$ for the next round
end for
$\mathcal{O}_i^{(K)} = \Phi_{\text{dec}}(\mathcal{F}_i^{(K)})$    ▷ Output the final detections
```

Spatial confidence-aware message fusion. Fig. 3 presents the detail about the spatial confidence-aware message fusion module. Given the received messages $\{\mathcal{P}_{j\to i}^{(k)}, j \in \mathcal{N}_i\}$, each agent i attentively augments the features with the received messages at each location. And the request map $\mathbf{R}_j^{(k)}$ in the received message is firstly decoded to the confidence map $\mathbf{C}_j^{(k)}$ via a point-wise minus. Then the per-location multi-head attention are separately operated at each location in parallel, it takes the features and the corresponding confidence scores as input, and outputs the augmented features.

1.4 Experimental settings

Implementation details. For camera-only 3D object detection task on OPV2V, we implement the detector following CADDN [1]. The model is trained 100 epoch with initial learning rate of 1e-3, and decay by 0.1 at epoch 80. For LiDAR-based 3D object detection task, our detector follows MotionNet [2]. We train 120 epoch with learning rate 1e-3. For the camera-only 3D object detection task on CoPerception-UAVs, our detector follows the CenterNet [3] with DLA-34 [4] backbone. The model is trained 140 epoch with learning rate 5e-4.

Inference strategy in multi-round setting. For the single-round communication, all the communication budget are used in this broadcast communication round. For the two-round communication, a small bandwidth (about 20%) is allocated to activate the collaboration; for the next round, the remained relatively large (about 80%) bandwidth is allocated to transmit the targeted information to



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

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

Figure 4: Visualization of collaboration between Vehicle 1 and Vehicle 2 on OPV2V 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 Vehicle 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 can be detected through transmitting spatially sparse, yet perceptually critical message.

meet agents' request. For more than two rounds communication setting, we strategically allocate communication budget across multiple communication rounds. For the initial broadcast round, a small bandwidth (about 20%) is allocated to activate the collaboration; for the next round, a relatively large (about 60%) bandwidth is allocated to transmit the targeted information to meet agents' request; then, the bandwidth is gradually reduced, accounting for the communication degradation with the increasing rounds.

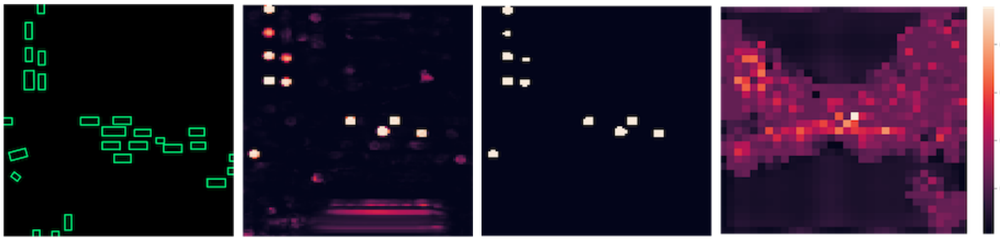
1.5 Visualization of spatial confidence map

Visualization of collaboration in OPV2V. Fig. 4 illustrates how `Where2comm` is empowered by the proposed spatial confidence map. In the scene, with Vehicle 2's help, Vehicle 1 is able to detect the missed objects in the single view. Fig. 4 (a-d) shows Vehicle 1's spatial confidence map, binary selection matrix, ego attention weight, and the detection results by its own observation. Fig. 4 (e-f) shows Vehicle 2's message sent to Drone 1, including the request map (opposite of confidence map) and the sparse feature map, achieving efficient communication. Fig. 4 (g) shows the attention weight for Vehicle 1 to fuse Vehicle 2's messages, which is sparse, yet highlights the objects' positions. Fig. 4 (d) and (h) compares the detection results before and after the collaboration with Vehicle 2. We see that the proposed spatial confidence map contributes to spatially sparse, yet perceptually critical message, which effectively helps Vehicle 1 detect occluded objects.

Visualization of spatial confidence map on V2X-Sim. Fig. 5 illustrates how `Where2comm` is empowered by the proposed spatial confidence map on V2X-Sim dataset. We see that: i) the confidence map is extremely sparse and highlights the spatial regions with objects; ii) the constructed binary communication graph promotes similar sparsity as the spatial confidence map; and iii) among the communicating spatial areas, the regions with objects have higher fusion weights than background areas.

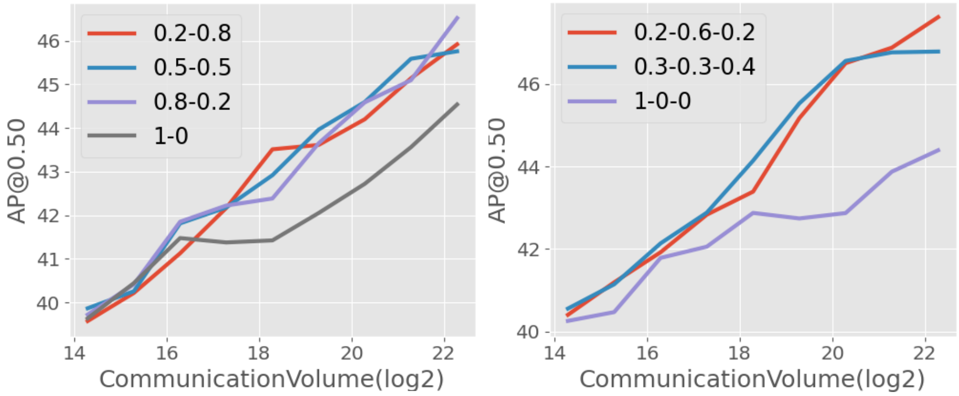
1.6 Ablation on bandwidth allocation

Fig. 6 shows the bandwidth allocation ablation study in multi-round communication setting. We see that allocating more bandwidth in the second and subsequent communication rounds achieves a better performance-bandwidth trade-off than allocating all bandwidth in the initial communication round, and the gain is stable for different bandwidth allocation strategies. The reason is that multi-round



(a) GTs in BEV.
(b) Confidence map.
(c) Selection matrix.
(d) Fusion weight.

Figure 5: Visualization of V2X-Sim dataset. The spatial confidence map is extremely sparse and the spatial regions with objects are highlighted. The constructed binary communication graph promotes similar sparsity as the spatial confidence map. And among the communicating spatial areas, the regions with objects have higher fusion weights than background areas.



(a) 2-round communication.
(b) 3-round communication.

Figure 6: Bandwidth allocation ablation study in multi-round communication. (a-b) shows the perception performance and communication bandwidth trade-offs for 2- and 3-round communication using different bandwidth allocation strategies on the OPV2V dataset. The legend shows the bandwidth ratio from the initial communication round to the entire communication round. Allocating more bandwidth in the second and subsequent communication rounds achieves a better performance-bandwidth trade-off than allocating all bandwidth in the initial communication round.

communication employs a request map in the second and subsequent communication rounds to denote the spatial area where each agent needs more information, which enables more targeted and efficient communication.

References

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

[2] 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.

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

[4] 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.