mmCooper: A Multi-agent Multi-stage Communication-efficient and Collaboration-robust Cooperative Perception Framework

Supplementary Material

1. Overview

The supplementary material is organized into the following sections:

- Sec. 2: The System Pipeline of mmCooper
- Sec. 3: Details of Multi-stage Fusion Method
- Sec. 4: Additional Experimental Results
  - Sec. 4.1: Implementation Details
  - Sec. 4.2: Supplements on Localization Errors
  - Sec. 4.3: Supplements on Transmission Delays
  - Sec. 4.4: Robustness to Heading Errors
  - Sec. 4.5: More Ablation on V2VSet Dataset
  - Sec. 4.6: Performance on V2V4Real Dataset
  - Sec. 4.7: Computation Costs
  - Sec. 4.8: Ablation of Deformable BBox Attention (DBA)
  - Sec. 4.9: Impact of Varying the Number of Agents
- Sec. 5: Additional Qualitative Evaluation Results
  - Sec. 5.1: Visualization of Detection Result
  - Sec. 5.2: Visualization of Multi-stage Fusion

2. The System Pipeline of mmCooper

The proposed system pipeline of mmCooper is illustrated in Algorithm 1. Note that the following pipeline is executed in parallel across all agents. For simplicity, we describe the pipeline from the perspective of the ego agent. The ego agent is represented by i, while the collaborative agents are denoted by j. First, the agent generates BEV features F^i and F^j through the Observation Encoding. In the Information Broadcasting, agents generate initial coarse bounding boxes B^i and B^j, where N_b^i and N_b^j represent the number of bounding boxes predicted by the ego agent and the collaborating agents, respectively. The collaborative agents then package and broadcast the filtered BEV features along with the coarse bounding boxes. Specifically, $f_{gaussian}(\cdot)$ denotes a Gaussian filter, and $\{\hat{F}^j, \hat{\mathcal{B}}^j\}$ represents the filtered features and bounding boxes from the collaborative agents.

Subsequently, in the Intermediate-stage Fusion, the ego agent performs feature fusion using the Multi-scale Offset-aware Attention module. The outputs from the fused features at different scales, after undergoing upsampling, are concatenated to obtain the final fused features. Here, $f_{up2}(\cdot)$ and $f_{up3}(\cdot)$ denote the upsampling operations, while $\{F_{sc1}^i, F_{sc2}^i, F_{sc3}^i\}$ represent the feature fusion outputs at three different scales.

In the later-stage fusion, the BBox Filtering & Calibration Module is employed to learn the bounding box offsets and scores. Specifically, DBA refers to the Deformable Bounding Boxes Attention Module, while FFN denotes the feed-forward network, $f_{enc,b}(\cdot)$ represents the feature extractor for bounding boxes, $f_{off}(\cdot)$ is the offset mapping function, and $f_{score}(\cdot)$ is the score mapping function. $\varphi(\cdot)$ represents the filtering and calibration through scores and offsets.

Finally, the fused features are input into a detection head to predict the fused bounding box $\mathcal{B}_{fused}$. The final bounding box is composed of the bounding box predicted from the fused features $\mathcal{B}_{fused}$ and the bounding box obtained after filtering and calibration $\hat{\mathcal{B}}_m^j$ from collaborative agents. $f_{post}(\cdot)$ represents the combination of all bounding boxes.

3. Details of Multi-stage Fusion Method

Note that our proposed model requires only a single round of processing and communication, the same as other single-stage methods. As shown in Fig. 1, both intermediate-stage features and late-stage bboxes are handled within a single inference cycle and jointly transmitted in one communication round. While mmCooper includes two detection heads, they share parameters, and each head introduces only 2.71ms of processing time. Notably, mmCooper is more efficient than ERMVP, a single-stage intermediate fusion method, with a lower total runtime of 45.41ms compared to ERMVP's 55.30ms. Combined with its superior performance, this confirms that mmCooper remains both practically feasible and system-efficient, without the overhead suggested in the comment.

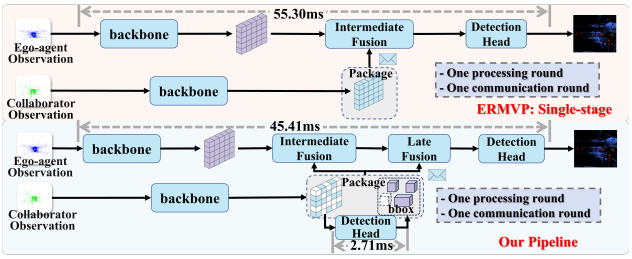

Figure 1. Comparison of single-stage method and our method.

4. Additional Experimental Results

In this section, we provide additional experiments to supplement the results on the datasets OPV2V [7], DAIR-V2X [9], V2XSet [6] and V2V4Real [8] presented in the *main paper*.

Algorithm 1 System Pipeline of the Proposed mmCooper

Define $N = \{1, ..., n\}$ as the agent set, $i \in N$ represents the ego agent, while $j \in N$ denotes the collaborative agents. $x_{i/j}$ serves as the input point cloud.

—

 # For collaborative agents.

Observation Encoding.

for each agent $j \in N$, **do**

$F^j = \psi_{encoder}(x^j) \in \mathbb{R}^{C \times H \times W}$

end for

Information Broadcasting.

for each $j \in N$, **do**

$\mathcal{B}^j = f_{dec}(F^j) \in \mathbb{R}^{N_b^j \times 7}$

$C_f^j, C_b^j = \phi_{conf}\left(F^j\right)$

$C_f^j, C_b^j = f_{gaussian}\left(C_f^j, C_b^j\right)$

$M_{f,t}^j, M_{b,t}^j = f_{top}\left(C_f^j, C_b^j\right)$

$M_{f,g}^j, M_{b,g}^j = f_{max}\left(\frac{exp\left((logC_{f/b}^j + g_f)/\tau\right)}{\sum_s exp\left((logC_s^j + g_s)/\tau\right)}\right)$

$\mathcal{M}_{f/b}^j = M_{f/b,g}^j \odot M_{f/b,t}^j$

$\hat{F}^j = \mathcal{M}_f^j \odot F^j; \hat{\mathcal{B}}^j = \mathcal{M}_b^j \odot \mathcal{B}^j$

broadcast $\{\hat{F}^j, \hat{B}^j\}$ to other agent

end for

—

 # For ego agent.

Observation Encoding.

$F^i = \psi_{encoder}(x^i) \in \mathbb{R}^{C \times H \times W}$

Information Broadcasting.

$\mathcal{B}^i = f_{dec}(F^i) \in \mathbb{R}^{N_b^i \times 7}$

Intermediate-stage Fusion.

Receive $\{\hat{F}^j, \hat{\mathcal{B}}^j\}$ sent by collaborative agents.

Encode the feature set $\{F^i, \hat{F}^j\}$ into three scales $S = \{sc1, sc2, sc3\}$.

For each $sc \in S$, **do**

$\mathcal{F}_{sc,(r,c)}^i = CrossAttn(MLP(F_{sc,(r,c)}^i), \hat{F}_{sc,(r,c)}^{j,s\times s}, \hat{F}_{sc,(r,c)}^{j,s\times s})$

end for

$\mathcal{F}^i = concat(\mathcal{F}_{sc1}^i, f_{up2}(\mathcal{F}_{sc2}^i), f_{up3}(\mathcal{F}_{sc3}^i))$

Late-stage Fusion.

$F_b(q) = f_{enc,b}(\hat{\mathcal{B}}^j)$

$DBA(q) = \sum_{\alpha=1}^{A} W_\alpha [\sum_{n=1}^{N} \sum_{m=1}^{M} \omega(W_\beta F_b(q)) \mathcal{F}^i(q + \triangle q_m)] + F_b(q)$

$F_{DBA} = FFN(F_{DBA}) + F_{DBA}$

$off, score = f_{off}(F_{DBA}), f_{score}(F_{DBA})$

$\hat{\mathcal{B}}_m^j = \varphi(off, score, \hat{\mathcal{B}}^j)$

Detection Decoders

$\mathcal{B}_{fused} = f_{dec}(\mathcal{F}^i)$

$\mathcal{B}_{final}^i = f_{post}(\mathcal{B}^i, \hat{\mathcal{B}}_m^j, \mathcal{B}_{fused})$

4.1. Implementation Details

On the OPV2V, DAIR-V2X and V2XSet datasets, the dimensions of the voxels encoded by the encoder are $0.4 \times 0.4 \times 4$. The shape of the shared BEV features among agents is (64, 100, 252) for DAIR-V2X and (64, 100, 352) for OPV2V and V2XSet. The shared bounding boxes among agents are represented by their center coordinates, dimensions (length, width, height), and heading angle. The detection decoder consists of two distinct 1×1 convolutional layers.

4.2. Supplements on Localization Errors

The localization errors on the OPV2V, DAIR-V2X and V2XSet datasets are sampled from a Gaussian distribution with a mean of 0 m and a standard deviation $\sigma \in \{0.0, 0.1, 0.2, 0.3, 0.4\}m$. The experimental results, as shown in Fig. 2, demonstrate that our proposed mmCooper outperforms the existing state-of-the-art methods [1, 3, 5, 6, 10] across different levels of localization error, highlighting its robustness to such errors.

4.3. Supplements on Transmission Delays

We evaluated the performance of the models on the OPV2V, DAIR-V2X and V2XSet datasets under transmission delays for $\{0, 100, 200, 300, 400\}ms$. As shown in Fig. 3, our proposed mmCooper consistently outperforms the existing state-of-the-art methods across all transmission delay conditions, demonstrating the superiority of our approach in scenarios with transmission delays.

4.4. Robustness to Heading Errors.

Fig. 4 demonstrates the performance of our proposed mmCooper method compared to other baseline methods under varying levels (i.e., $\{0.0, 0.2, 0.4, 0.6, 0.8\}^\circ$) of heading error noise on the OPV2V, DAIR-V2X and V2XSet datasets. As illustrated in Fig. 4, the performance of all models decreases as the heading errors increase. However, mmCooper consistently outperforms the current state-of-the-art models across all error levels, highlighting the advantages of our designed Multi-Scale Offset-Aware Fusion Module and BBox Filtering & Calibration Module in enhancing system robustness.

4.5. More Ablation on V2XSet Dataset

We supplement the *main paper* with ablation experiments conducted on the V2XSet dataset. As shown in Tab. 1, the results align with those presented in the *main paper*, demonstrating that the absence of any key component leads to performance degradation. Moreover, downgrading mmCooper to a single-stage approach also results in decreased performance.

4.6. Performance on V2V4Real Dataset

To further evaluate the performance of our model on more datasets, we also report the results on the V2V4Real dataset. Our experimental setup on the V2V4Real dataset is consistent with that of the DAIR-V2X dataset. As shown in Tab. 2, mmCooper still demonstrates outstanding performance on the V2V4Real dataset.

4.7. Computation Costs

Tab. 3 presents the overall computation time of our model as well as the computation time of each module. Although both our intermediate-stage and late-stage fusion modules utilize attention-based mechanisms, due to the sparsity of intermediate-stage features from collaborators and the sparsity of reference points used in deformable attention during late-stage fusion, our model achieves inference time comparable to other models.

Table 1. Ablation study results of different designs in mmCooper on the V2XSet datasets. CFG: Confidence-based Filter Generation Module; MOF: Multi-scale Offset-aware Fusion; BFC: BBox Filtering & Calibration Module; LF: Late-stage Fusion; IF: Intermediate-stage Fusion.

CFG	MOF	BFC	LF	IF	**AP@0.7/0.5**(↑)
					V2XSet
✔	✔	✔	✔	✔	**65.86/84.40**
Importance of Core Components					
✘	✔	✔	–	–	62.78/82.47
✔	✘	✔	–	–	64.33/84.04
✔	✔	✘	–	–	64.70/83.59
Results of Single-stage Fusion					
–	–	–	✘	✔	64.19/83.73
–	–	–	✔	✘	54.36/72.14

Table 2. Collaborative perception performance on the V2V4Real dataset with a time delay of 100 ms, localization errors of 0.2 m, and heading errors of 0.2° using Average Percision(AP)@0.7/0.5 as metrics.

Models	**V2V4Real**
	AP@0.7/0.5
No Fusion* [2]	24.72/43.52
Late Fusion* [2]	17.79/39.91
Intermediate Fusion* [2]	28.77/49.19
Where2comm [1]	29.83/49.43
V2X-ViT [6]	25.08/47.85
ERMVP [10]	26.81/44.57
Ours	**31.47/50.21**

4.8. Ablation of Deformable BBox Attention (DBA)

To further demonstrate the necessity of the Deformable BBox Attention (DBA) module, we replace it with the standard cross-attention mechanism. As shown in Tab. 4, due to the lack of focus on key reference points, replacing DBA with standard cross-attention leads to performance drops of 4.07%/0.07%, 0.58%/1.97%, and 1.44%/0.69% in AP@0.7/0.5 on OPV2V, DAIR-V2X, and V2XSet respectively.

4.9. Impact of Varying the Number of Agents

Fig. 5 illustrates the impact of varying the number of agents on detection performance in the OPV2V dataset. It can be observed that as the number of collaborating agents increases, the detection performance also improves.

Models	OPV2V
	Inference Time(ms)
Intermediate Fusion*	10.07
Where2comm [1]	17.05
V2X-ViT [6]	114.11
Select2col [4]	27.00
ERMVP [10]	55.30
Ours	45.41

Module	OPV2V
	Inference Time(ms)
Observation Encoding	12.26
Information Broadcasting	14.37
Intermediate-stage Fusion	13.54
Late-stage Fusion	4.06
mmCooper	45.41

Table 3. The left table presents the inference time of different models on the OPV2V dataset, while the right table shows the inference time of different components in mmCooper.

Table 4. Comparison of detection performance when using DBA or cross attention in the BBox Filtering & Calibration (BFC) module in AP@0.7/0.5 on OPV2V, DAIR-V2X, and V2XSet respectively.

Module	OPV2V	DAIR-V2X	V2XSet
Cross Attn	74.04/88.86	47.69/63.15	64.42/83.71
DBA	**78.11/88.93**	**48.27/65.12**	**65.86/84.40**

5. Additional Qualitative Evaluation Results

5.1. Visualization of Detection Result

We provide additional qualitative results on DAIR-V2X datasets. Fig. 7 present visualizations of different road scenarios. Our proposed mmCooper method can detect almost all ground truths without any false positives. The results demonstrate that our proposed mmCooper achieves outstanding detection performance across various scenarios.

5.2. Visualization of Multi-stage Fusion

We provide more visualizations for multi-stage fusion in different scenarios. As shown in the Fig. 6, our mmCooper effectively performs dynamic allocation of intermediate-stage features and late-stage bounding boxes across different scenarios while also refining the received bounding boxes.

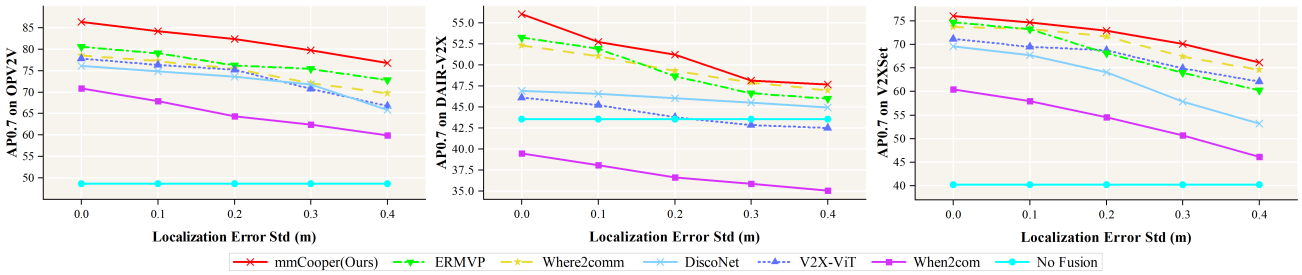

Figure 2. Robustness to the localization error on the OPV2V, DAIR-V2X and V2XSet datasets.

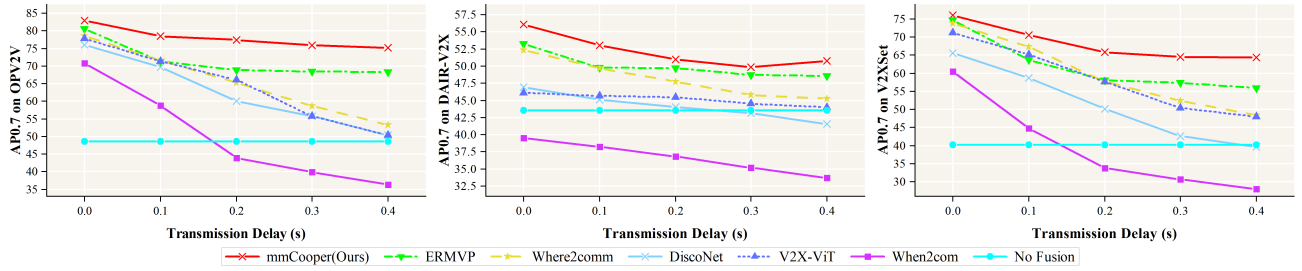

Figure 3. Robustness to the transmission delay on the OPV2V, DAIR-V2X and V2XSet datasets.

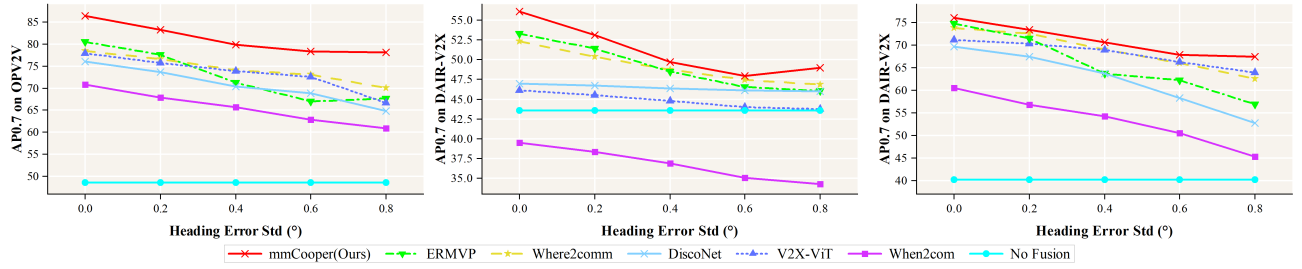

Figure 4. Robustness to the heading error on the OPV2V, DAIR-V2X and V2XSet datasets.

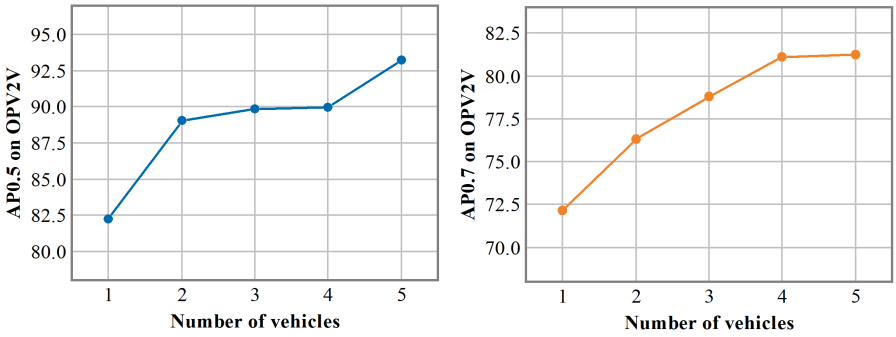

Figure 5. Impact of Varying the Number of Vehicle.

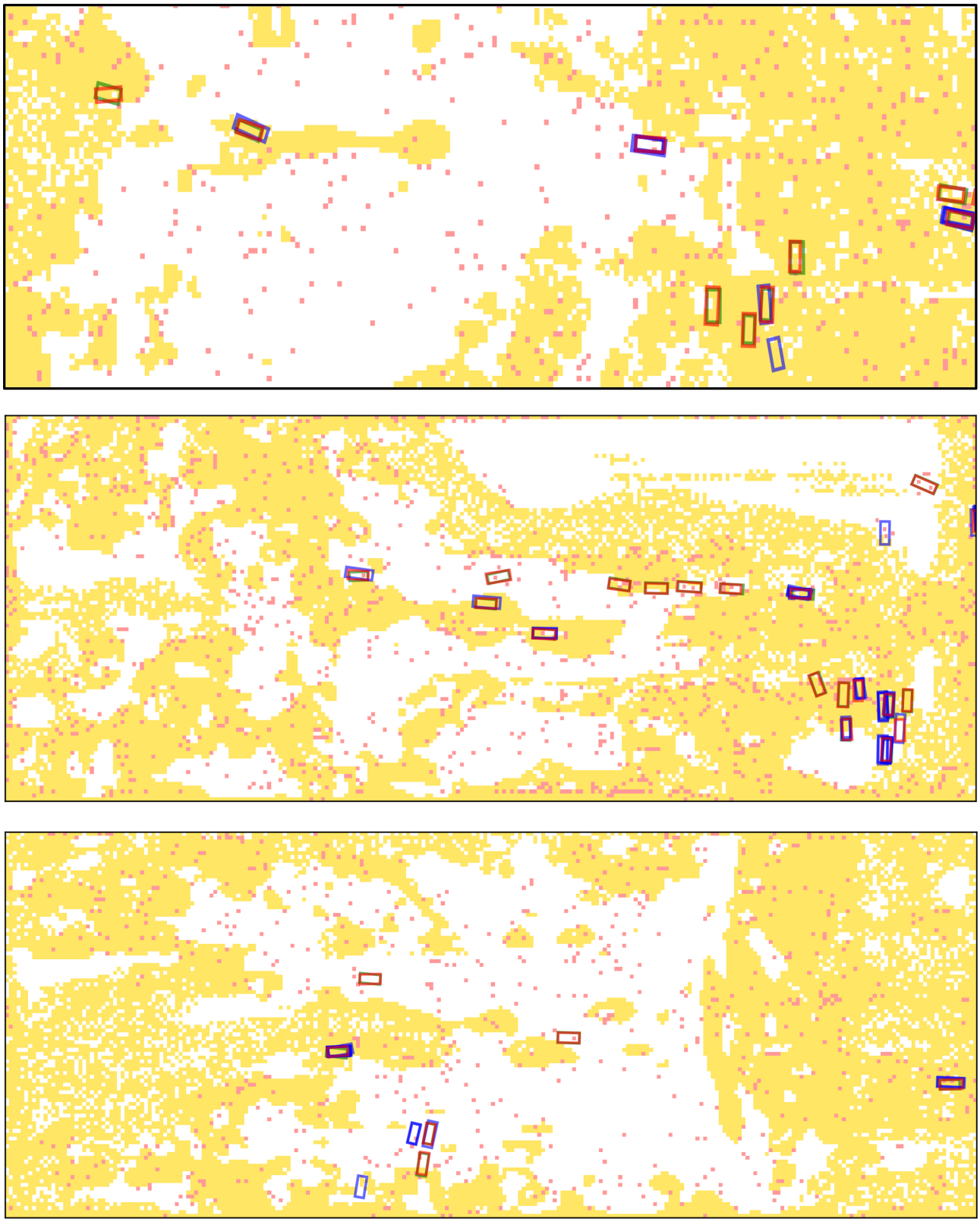

Figure 6. Visualization of two-stage fusion on DAIR-V2X. We show the results from the Confidence-based Filtering Generation Module, including discarded information (white background), BBoxes for transmission (red dots), and features for transmission (yellow background). We also show the BFC module results, including uncalibrated BBoxes (blue boxes), calibrated BBoxes (red boxes), and ground truth BBoxes (green boxes).

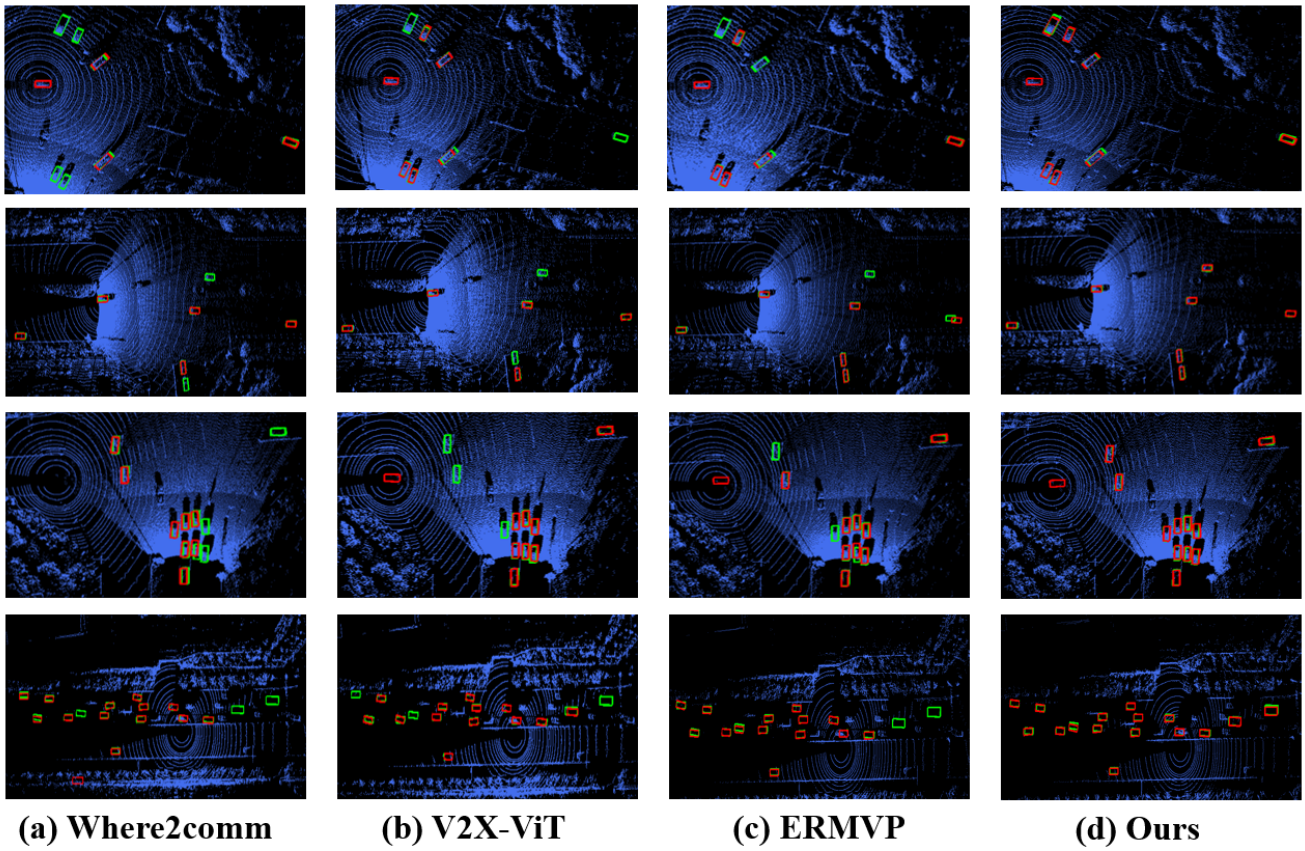

Figure 7. More visualization comparison of detection results on the DAIR-V2X dataset. Green and red boxes represent the ground truth and the model-predicted bounding boxes, respectively.

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*, 35:4874–4886, 2022. 3, 4

[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*, 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*, 34:29541–29552, 2021. 3

[4] Yuntao Liu, Qian Huang, Rongpeng Li, Xianfu Chen, Zhifeng Zhao, Shuyuan Zhao, Yongdong Zhu, and Honggang Zhang. Select2col: Leveraging spatial-temporal importance of semantic information for efficient collaborative perception. *IEEE Transactions on Vehicular Technology*, 2024. 4

[5] 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. 3

[6] 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 *European conference on computer vision*, pages 107–124. Springer, 2022. 1, 3, 4

[7] 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 *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022. 1

[8] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui 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*, pages 13712–13722, 2023. 1

[9] 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*, pages 21361–21370, 2022. 1

[10] Jingyu Zhang, Kun Yang, Yilei Wang, Hanqi Wang, Peng Sun, and Liang 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*, pages 12575–12584, 2024. 3, 4