BM2CP: Efficient Collaborative Perception with LiDAR-Camera Modalities

Anonymous Author(s)
Affiliation
Address
email

Abstract: Collaborative perception enables agents to share complementary perceptual information with nearby agents. This can significantly benefit the perception performance and alleviate the issues of single-view perception, such as occlusion and sparsity. Most proposed approaches mainly focus on single modality (especially LiDAR), and not fully exploited the superiority of multi-modal perception. We propose an collaborative perception paradigm, *BM2CP*, which employs LiDAR and camera to achieve efficient multi-modal perception. *BM2CP* utilizes LiDAR-guided modal fusion, cooperative depth generation and modality-guided intermediate fusion to acquire deep interactions between modalities and agents. Moreover, it is capable to cope with the special case that one of the sensors is unavailable. Extensive experiments validate that it outperforms the state-of-the-art methods with $100\times$ lower communication volumes in real-world autonomous driving scenarios. Our code is available at *supplementary materials*.

Keywords: Multi-Agent Perception, Multi-Modal Fusion, Vehicle-to-Everything (V2X) Application

1 Introduction

Collaborative perception enables agents to share complementary perceptual information with their nearby agents. It is raised to fundamentally alleviate the issues of single-agent perception, such as occlusion and sparsity in raw observations. Recent works have proposed different strategies to realize collaborative perception, which could be divided in three categories: LiDAR based collaboration [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15], camera based collaboration [16, 17, 18, 19] and multi-modal based collaboration [20].

An intuition is that different types of sensors can provide heterogeneous information at different levels, and thus more accurate perception could be achieved through multi-modal analysis. However, most of recent works are not multi-modal based approaches and they present better performance only using LiDAR. It is also demonstrated by experiments in Section 4.4 that fusing different modalities straightforward brings negative profits. Perhaps we could get answers from the properties of sensors. LiDAR captures accurate spatial information, which provides information of shapes and depths in low resolution. Thus LiDAR can meet the fundamental demands of task-specific collaborative perception. Whereas camera specializes in rich semantics and contexts with fixed view. This demonstrates the domination of LiDAR and the role of camera, which is used to complement the information where point clouds are absent or sparse.

Therefore, it poses challenges to build a well-behaved collaborative perception method with LiDAR-camera modalities: (a) How to fuse the LiDAR data and camera data effectively? (b) How to collaborate with multi-modal data? (c) How to deal with the case that one of the sensor is missing?

In this paper, we answer the aforementioned questions by proposing a Biased Multi-modal Collaborative Perception(*BM2CP*) method including three components: (a) LiDAR-camera fusion, which

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

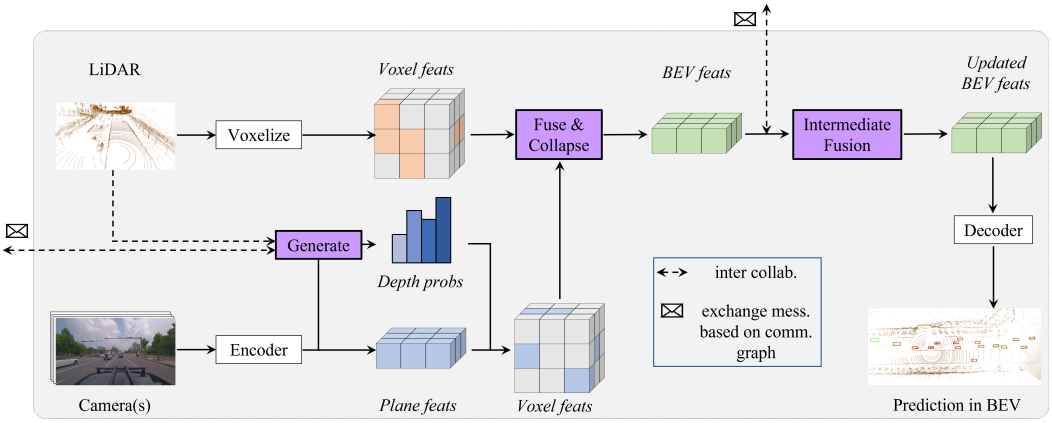


Figure 1: Overall framework of the proposed *BM2CP*. We build multi-agent collaborative perception with LiDAR-camera modalities in voxel space. *BM2CP* includes biased multi-modal fusion, cooperative depth generation and modality-guided collaborative fusion, which are colored in purple. Object detection task is taken as example.

achieves sufficient multi-modal fusion guided by LiDAR; (b) cooperative depth generation, which provides reliable depth distribution for camera modality branch; (c) modality-guided collaborative fusion, which achieves multi-view multi-modal critical feature learning. Compared to recent LiDAR based methods, *BM2CP* exploits deeper interactions between modalities and agents. Meanwhile, it has the capability to deal with the case that one of the sensor is unavailable.

In summary, our main contributions are threefold:

- We propose a novel framework for multi-modal collaborative perception, where modal fusion is guided by LiDAR and collaborative fusion is guided by modality. Besides, it can handle most cases without fine-tuning when one of the modalities is present.
- We design LiDAR-guided depth generation and modal fusion, which achieves deeper interactions between LiDAR and camera modalities. And we exchange the message of depths and features across agents, which achieves better modal feature learning and efficient communication. These designs encourage sufficient feature fusion in both modality aspect and collaboration aspect.
- Extensive experimental results and ablation studies on the real-world public dataset DAIR-V2X show the performance and efficiency of *BM2CP*. It achieves superior performances with 64.03% AP at IoU0.5 and 48.99% AP at IoU0.7. When all camera sensors are missing, the performance is still comparable to other state-of-the-art LiDAR based methods.

2 Related Works

Multi-modal perception in single agent. Single-agent perception mainly focuses on three types of commonly equipped sensors: camera, LiDAR and RADAR. Similar to collaborative perception, most 3D object detectors only use LiDAR as input and achieve state-of-the-art (SOTA) [21]. But fusing LiDAR and camera signals still attracts increasing interest recently and is expected to outperform LiDAR based perception. To collect useful semantics, one fusion paradigm associate camera and LiDAR by projecting 3D queries to 2D planes [22, 23, 24, 25, 26]. However, such association is quite sparse, since few 3D points could be projected and most 2D semantics would be wasted. Another recent paradigm learns joint representations by transforming dense image elements and sparse point cloud elements to a shared space [27, 28, 21, 29, 30], which generally expands on voxel space or bird's-eye view (BEV) space. But the fusion strategies are often simple or brute without extra interactions.

In this paper, we strive to build deep interactions between LiDAR and camera by depth guidance and special-designed fusion strategy to provide an effective solution. Considering the advantage in global, the two deep interactions are dominant by LiDAR modality.

Collaborative perception. Most collaborative methods solve diverse issues using LiDAR such as performance [1, 3, 6, 8, 12], bandwidth trade-off [11, 18], communication interruption [7], latency [13] and pose error correction [5, 15]. For camera based category, Xu et al. [18] propose the first attention-based multi-camera cooperative framework, which merely shares BEV features with Transformer [31] and generates BEV predictions. Hu et al. [19] not only exchange BEV features but also estimated depths to reduce the impact of erroneous depth estimates. However, depth ground truth is used and therefore need to be further collected using depth sensor, which limits the generalization of their approach.

For LiDAR-camera based category, Zhang et al. [20] first generates virtual 3D points with limited number of RGB image pixels where locate in 2D predicted bounding box. It does not change the conventional framework that perform collaborative detection with single modality (LiDAR cloud points). Xiang et al. [32] construct communications with BEV features, while it mainly focus on the design of attention blocks. And their method only covers the case that each agent is only equipped with single modality. Compared with aforementioned methods, neither extra ground truth like depth is needed, nor attention mechanism is the focus in our method. Each agent is equipped with both LiDAR and camera, and only task-related ground truth is used.

Special case in multi-modal perception that one of the sensors is missing. Few works discuss the special case that one of the modality is absent, and none in multi-agent collaborative perception. To the best of our knowledge, Li et al. [33] give a solution to single-agent detection when RADAR or LiDAR is missing. They treat missing sensor as clear (data full of zero value), and use a teacher model with exponential moving average (EMA) to learn equally from both modalities. To avoid the same problem, we decide only to use present data. In this situation, the main problem is to work with an unfixed number of modalities. Finally, we adjust the workflows of multi-modal fusion and collaborative fusion, which make the model fit the modality-absent cases without fine-tuning.

It is worth emphasizing that camera cannot finish the perception independently especially when agents are only equipped with one camera. In this case, reliable depth generation becomes impossible and thus the final results is badly poor. It can also be inferred based on the experiments on DAIR-V2X dataset [34] conducted by Hu et al. [19], which additionally uses depth ground truth. Therefore, we propose LiDAR-guided modal fusion, and experiment results that camera is only used should not be expected.

3 Efficient Collaborative Perception with LiDAR-Camera Modalities

Our motivation comes from two aspects. First, depth is the major gap for RGB image to lift up to voxels, but part of correct depths could be provided by LiDARs of its own and nearby. Thus, we could transmit depth information between modalities and agents. This allows compensation from LiDAR in different views, and reduces the uncertainty of the infinite depth prediction. Second, the transmitted message should include detection clues to provide refined and complementary information, which can fundamentally overcome inevitable limitations detected by single modality or single agent.

Based on the above rationals, our designs include (a) *biased multi-modal fusion*, which achieves sufficient single-view local feature learning; (b) *cooperative depth generation*, which is guided by LiDAR and shares depth information across agents; and (c) *modality-guided collaborative fusion*, which shares critical BEV detection features to improve representation and performance.

3.1 Problem formulation

To the best of our knowledge, no work gives the formulation about multi-agent collaborative perception with LiDAR-camera modal fusion. Here, we provide an accessible and reasonable definition

which fuses the data of LiDAR and camera in voxel space. Considering N agents travel in the scenario, let I_i be the RGB image and P_i be the point cloud collected by camera and LiDAR of the i-th agent, and Y_i be the corresponding ground truth detection. First of all, agents exchange their positions and relative poses to build a communication graph. In order to fuse features in voxel space, the point cloud and the image need to be voxelized. Note that in the camera branch, 2D plane features and categorical depth distribution are generated based on raw image. Then voxel sampling points are generated at the center of each voxel, and are transformed to the frustum grid [35].

$$\mathbf{V}_i^l = f_{voxelize}(P_i), \mathbf{V}_i^c = f_{img_enc}(I_i) \otimes f_{dep_est}(I_i) \tag{1}$$

Then we fuse point cloud features and image features by cell and collapse the fused voxel features $\mathbf{V}_i^f$ to bird's-eys-view (BEV). Note that voxel features in different branches are shaped to the same spatial resolution, which is prepared for modal fusion in the next step.

$$\mathbf{V}_i^f = f_{modal_fuse}(\mathbf{V}_i^l, \mathbf{V}_i^c), \mathbf{F}_i = f_{collap}(\mathbf{V}_i^f) \tag{2}$$

After collapsing, message packing and passing will be performed based on communication graph, which is used to exchange useful and compact intermediate BEV features. And each agent aggregates the received features with its own BEV features to conduct intermediate fusion. Finally, a decoder is implemented to predict the results for a specific task such as object detection and scenario segmentation.

$$\tilde{\mathbf{F}}_i = f_{feat_fuse}(\mathbf{F}_i, \{\mathbf{F}_{j\to i}\}_{j\in N_i}), \hat{Y}_i = f_{decoder}(\tilde{\mathbf{F}}_i) \tag{3}$$

where $\mathbf{F}_{j\to i}$ is the warped features based on relative poses of the j-th agent and the i-th agent. And N_i is the nearby agents that the i-th agent can communicate with. The objective of *BM2CP* is to minimize the distance between predicted and the ground truth detection.

$$\min \sum_i g(\hat{Y}_i, Y_i) \tag{4}$$

where $g(\cdot,\cdot)$ is the detection evaluation metric. The overall framework is also illustrated in Fig 1.

3.2 Biased multi-modal fusion

After voxelization, we propose a LiDAR-guided modal fusion to achieve interactions between modalities, which is shown in Fig 2a. We first group the voxels according to their modalities: LiDAR voxels (colored in orange), camera voxels (colored in blue), LiDAR-camera hybrid voxels (colored in green), and normal voxels (colored in gray). Motivated by the fact that LiDAR-based detectors usually surpass camera-based counterparts, we take LiDAR as the guiding modality to select useful information from camera features.

For LiDAR-camera hybrid voxels, the fusion result $\tilde{v}_H$ is updated by LiDAR-conditioned gate, which is formulated as:

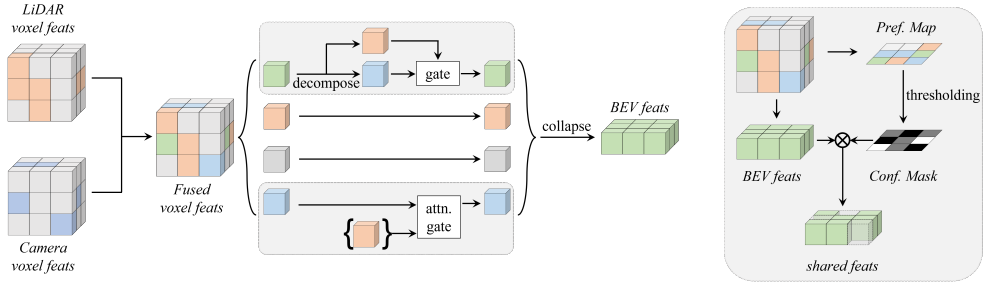
$$gate = ReLU(Conv(v_L)), \tilde{v}_H = Conv([gate * v_C, v_L]) \tag{5}$$

where v_L and v_C denote voxel features from LiDAR and camera, respectively.

For camera voxels, the LiDAR-conditioned gate cannot be implemented since camera voxels do not contain features from LiDAR. Instead, we conduct an attention gate which uses the global LiDAR voxel features to filter camera voxels. The attention gate and fusion result $\tilde{v}_C$ are formulated as:

$$gate = Sigmoid(Linear(MHA(\{v_L\}, v_C, v_C))) > threshold, \tilde{v}_C = gate * v_C \tag{6}$$

where $MHA(\cdot)$ is the multi-head attention which outputs the scaled dot-product attention weight. The threshold is empirically set as 0.5.



(a) (b)

Figure 2: LiDAR-guided modal fusion (a) and modality-guided masked feature selection (b).

For LiDAR voxels and normal voxels, they are directly used as the fusion result. Since they are not affected by camera voxel features. Finally, biased modal fusion is achieved and the fine-grained voxel features are collapsed to BEV features.

3.3 Cooperative depth generation

To lift a pixel in the 2D space to the voxel space, the depth of each pixel should be first estimated. To reduce the erroneous of estimation and obtain a reliable depth distribution, we use a hybrid strategy, termed *prediction&projection*. *Prediction* is to predict pixel-wise depth using neural network. Following Reading et al. [35], the predicted depths are classified to a series of discrete values. The number of classes corresponds to the number of discretized depth bins. *Projection* is to transform LiDAR point clouds to RGB image coordinates. Let $P_i = \{(x_i, y_i, z_i, 1)\}$ be the point clouds, where (x_i, y_i, z_i) is the 3D coordinate. Let $T_{ldr2cam}$ be the mapping from LiDAR sensor to camera sensor, and $T_{cam2img}$ represent the mapping from camera sensor to the RGB image. The overall mapping from LiDAR to RGB image is

$$I_i^{'} = T_{ldr2img} P_i = T_{ldr2cam} T_{cam2img} P_i \tag{7}$$

where $I_i^{'} = \{(u_i, v_i, d_i)\}$ is the projected image with depth information. (u_i, v_i) is the 2D coordinate and d_i is the corresponding depth of each pixel. To match the idea of Reading et al. [35], the projected depths are also classified.

Note that some pixels may have no projected depth and some have more than one, the hybrid strategy is implemented as follows: pixel with no projected depth uses the predicted depth from *prediction*, and pixel with more than one projected depth select the minimized one. For pixel with only one projected depth, no extra operation is needed.

On the other hand, LiDAR point clouds of nearby agents contains different depth information. The intuition is, the 3D location of a correct depth candidate is spatially consistent through viewpoints of multiple agents. Therefore, depth information could be exchanged to obtain more reliable depth distribution and promotes more accurate 3D representations. Only the depths of part pixels are shared, since the association between LIDAR and camera data is sparse. Thus, the communication is efficient and less time-consuming.

3.4 Multi-agent collaborative perception

To promote efficient intermediate communication, we design a modality-guided collaboration by selecting the most critical spatial features, which is shown as Fig 2b. we first generate a 2D preference map based on the fused voxel features $\mathbf{V}_i^f \in \mathcal{R}^{X \times Y \times Z}$. For each cell of preference map, we consider a set of voxels which can be collapsed to the cell in Z-axis. The preferred order is *hybrid* $>$ *LiDAR* $>$ *camera* $>$ *normal*. For example, if a hybrid voxel is contained in the voxel set, we mark the corresponding cell of preference map as hybrid.

Then we generate the confidence mask based on preference map, which decides whether to communicate each cell of BEV features to nearby agents. Once a cell is larger than a threshold, the cell of the same location in BEV features is regarded as critical and need to be broadcast. Since the hybrid cell is guided by LiDAR and contains sufficient multi-modal information, the threshold for hybrid cell is set as 0. For a better performance-bandwidth trade-off, the threshold for other types of cell is set as 0.5.

To fuse the features from nearbys, multi-head attention is also implemented to generate attention weight for each agent. The query, key and value for the i-th agent is

$$q = \mathbf{F}_i, k = v = \{\mathbf{F}_{j\to i}\}_{j\in N_i} \tag{8}$$

After collaboration, the updated BEV features are used to generate the final detection with task-specific decoder.

3.5 Robustness against the case of missing sensor

The proposed framework is capable to deal with the case that one of sensors is missing. This is done by adjusting the workflow at the steps of modal fusion and collaborative fusion. In modal fusion step, there are less types of voxel features when point cloud or RGB image is absent. If the RGB image is absent, the LiDAR voxels and normal voxels are directly used as fusion result. And the case is similar when the point cloud is absent. In collaborative fusion step, alternative types of voxels in Z-axis reduces, and no hybrid voxels exist. We set unified threshold as 0.5 for all cells. By conducting this, *BM2CP* adapts well to the modality-absent cases. The experimental results are collected in Sec 4.2.

4 Experimental Results

4.1 Experimental setup

Dataset. We conduct experiments of 3D object detection task on DAIR-V2X dataset [34]. DAIR-V2X dataset is the first public real-world collaborative perception dataset. Each sample contains a vehicle and an infrastructure, and they are equipped with a LiDAR and a front-view camera. The perception range is $201.6m \times 80m$.

Compared methods. We consider comparisons with following LiDAR-based methods: *No Fusion*, *Late fusion*, *When2com* [4], *V2VNet* [3], *DiscoNet* [18], *V2X-ViT* [12] and *Where2comm* [11]. Among these methods, *No Fusion* is considered as the baseline which only uses individual observation. *Late fusion* shares the detected 3D bounding boxes with nearby agents. Rest are state-of-the-art (SOTA) LiDAR-based collaborative perception algorithms.

Implementation details. We re-implement all methods based on the pyTorch [36] framework and OpenCOOD [9] codebase with Adam [37] optimizer and multi-step learning rate scheduler. In order to compare communication volumes fairly, all compared methods use the same architectures and follows PointPillar [38] with no feature compression. The transmitted BEV intermediate representation has a size of $200 \times 504 \times 64$, and the resolution is 0.4m per pixel in length and width. Weighted cross entropy loss is used. The detection results are evaluated by Average Precision (AP) at Intersection-over-Union (IoU) threshold of 0.50 and 0.70.

4.2 Quantitative Evaluation

Comparison with baseline and SOTA methods. Table 1 shows the comparisons with recent methods in terms of communication volumes and detection performance. Mathematically, the communication volume is calculated in log-scale by $log_2(|\mathbf{F}_{i\to j}|)$, where $|\cdot|$ is the $L0$ norm which counts the non-zero elements in BEV features. It is observed that *BM2CP*: (a) achieves a superior perception-communication trade-off; (b) achieves significant improvements over previous SOTA methods, which improves the performance by 5.2% on AP@IoU0.5 and 5.4% on AP@IoU0.7; (c)

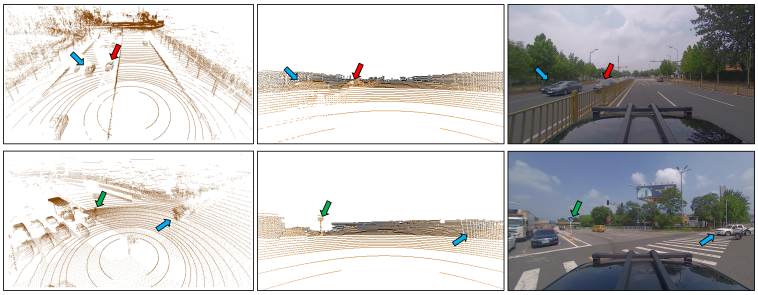


Figure 3: Two visualizations of projected depths from LiDAR coordinates to image coordinates. Paired arrows in colors indicate the same objects including car and sign in LiDAR data (left), camera data (right) and projected depth map (middle), respectively.

Table 1: 3D detection quantitative results on DAIR-V2X dataset. *Comm* is the communication volumes in log-scale.

Method	Comm	AP@0.5	AP@0.7
No Fusion	0	52.75	45.63
Late Fusion	18.62	53.88	38.12
When2com (CVPR'20)	20.31	51.88	37.05
V2VNet (ECCV'20)	22.90	58.80	43.75
DiscoNet (NeurIPS'21)	21.85	55.17	39.84
CoBEVT (CoRL'22)	21.31	50.70	39.39
V2X-ViT (ECCV'22)	20.45	56.75	39.90
Where2comm (NeurIPS'22)	16.54	60.85	46.48
BM2CP	11.01	64.03	48.99

Table 2: Evaluation results against the issue of missing sensor. *Camera* or *LiDAR* denotes the sensor that the agent is equipped with.

Ego	Nearby	AP@0.5	AP@0.7
Camera	LiDAR	19.21	5.74
Camera	Both	20.41	6.05
LiDAR	LiDAR	61.59	47.30
LiDAR	Both	62.71	47.64
Both	Camera	53.31	43.86
Both	LiDAR	63.54	48.38
Both	Both	64.03	48.99

achieves a better performance with extremely less communication volume, which is 105 times less than baseline and 46 times less than *Where2comm*.

Robustness against the absence of modality. Table 2 shows the results with the setting of missing sensor. It is observed that: (a) the performance degrades severely when one of the agents is only equipped with camera sensor. This is consistent with the conclusion that the performance is poor when only camera sensor works; (b) The performance (61.59% on AP@IoU0.5 and 47.30% on AP@IoU0.7) is comparable with SOTA methods when agents only use LiDAR for perception; (c) The performance drop slightly when one of the agents miss the camera sensor.

4.3 Qualitative Evaluation

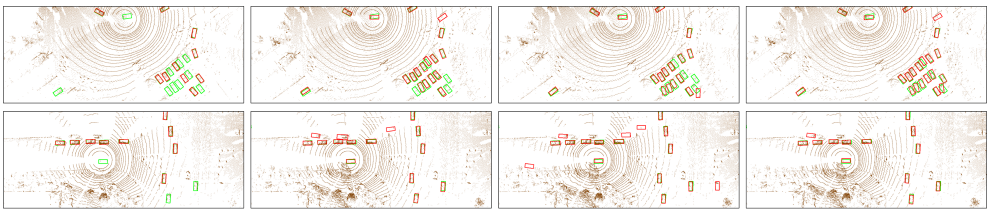
Visualization of projected depths. Fig 3 shows how depth distribution is empowered by LiDAR. In the scene, objects show evident contour against the background with lighter coloring. And more depths from nearby agents will further fill the depth in empty (white pixel in image). Therefore, *projection* promotes more reliable depth generation, and benefits multi-modal fusion.

Comparison with SOTA methods. Fig 4 shows the comparison with *No Fusion*, *V2X-ViT* and *Where2comm*. *BM2CP* achieves more complete and less false negative detection results. The reason is that *BM2CP* leverages cell-level multi-modal feature fusion in critical voxels and employs modality-guided cell-level confidence mask to achieve more comprehensive fusion.

4.4 Ablation Study

Ablation study is conducted to investigate the effectiveness of the main components in our method. The overall results are presented in Table 3.

LiDAR-guided modal fusion. As shown in Table 3a, equally multi-modal fusion can result in a minor drop in performance. It proves the importance of the guidance by LiDAR.



(a) No Fusion (b) V2X-ViT (c) Where2comm (d) BM2CP

Figure 4: Qualitative results in DAIR-V2X dataset. Ground truths are colored in green and predictions are colored in red. *BM2CP* detects more objects than *V2X-ViT* and *Where2comm* in the first row, and detects less false positive objects in the second row.

Table 3: Component ablation studies.

(a) Modal fusion strategy.

Strategy	AP@0.5/0.7
No	50.17/37.33
Equal	58.33/42.13
Biased	60.31/44.42

(b) Depth projection.

Projection	AP@0.5/0.7
No	52.84/38.07
Ego	60.31/44.42
All	62.25/46.46

(c) Collaborative strategy.

Strategy	AP@0.5/0.7
Max	61.33/46.02
Concat	62.25/46.46
Attention	63.65/48.42

LiDAR-based depth generation. We investigate the effect of using projected depths from ego and nearbys. Results in Table 3b indicate LiDAR-based cooperative depth generation is essential.

Modality-guided collaborative fusion. We compare our masking-attention strategy with strategies of max fusion and concatenate fusion. Table 3c show that both max fusion and concatenate fusion lead to performance degradation and increase the demand of bandwidth.

5 Conclusion and Limitations

In this paper, we propose a novel framework, termed *BM2CP*, for multi-agent collaborative perception, which focuses on fusion with LiDAR-camera modalities. We adopt LiDAR-guided modal fusion to achieve reasonable and comprehensive modal feature learning, apply cooperative depth generation to enhance the modal fusion with more reliable depth information, and propose modality-guided collaborative fusion for more efficient and critical feature fusion. Extensive experiments demonstrate the superior performance and the effectiveness of designed components.

Limitations and future works. We also evaluate the robustness against the localization noise following V2VNet [3] and V2X-ViT [12]. Gaussian noise with a mean of 0m and a standard deviation of 0m-0.6m is used. Unfortunately, performance degrading happens when localization noise increases. The reason comes from the *cooperative depth generation*. The shared depth distributions are broadcast based on the localization and relative pose of each agent. Therefore, it intensifies the errors and provides wrong depth information for images. This is listed as one of the future works. A possible solution is to use the powerful transformer architecture, which can attentively select the most reliable depths and project them to the correct cells; Another possible solution is to build the global consensus about the localization of agents according to Vadivelu et al. [5] and Lu et al. [15].

DAIR-V2X dataset only contains two agents, and each agent is equipped with one camera. This makes some studies impossible such as the robustness against camera dropout, the number of agents, etc. Besides, validations on more datasets are critical to prove its generalization.

References

[1] Q. Chen, S. Tang, Q. Yang, and S. Fu. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 514–524. IEEE, 2019.

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

[3] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 605–621. Springer, 2020.

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

[5] N. Vadivelu, M. Ren, J. Tu, J. Wang, and R. Urtasun. Learning to communicate and correct pose errors. In *Conference on Robot Learning*, pages 1195–1210. PMLR, 2021.

[6] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34:29541–29552, 2021.

[7] S. Ren, Z. Lei, Z. Wang, S. Chen, and W. Zhang. Robust collaborative perception against communication interruption. *the 2nd IJCAI Workshop on Artificial Intelligence for Autonomous Driving*, 2022.

[8] Y. Yuan, H. Cheng, and M. Sester. Keypoints-based deep feature fusion for cooperative vehicle detection of autonomous driving. *IEEE Robotics and Automation Letters*, 7(2):3054–3061, 2022.

[9] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022.

[10] J. Cui, H. Qiu, D. Chen, P. Stone, and Y. Zhu. Coopernaut: end-to-end driving with cooperative perception for networked vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17252–17262, 2022.

[11] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Advances in Neural Information Processing Systems*, volume 35, pages 4874–4886, 2022.

[12] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 107–124. Springer, 2022.

[13] Z. Lei, S. Ren, Y. Hu, W. Zhang, and S. Chen. Latency-aware collaborative perception. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 316–332. Springer, 2022.

[14] Y. Li, J. Zhang, D. Ma, Y. Wang, and C. Feng. Multi-robot scene completion: Towards task-agnostic collaborative perception. In *Proceedings of The 6th Conference on Robot Learning*, pages 2062–2072. PMLR, 2022.

[15] Y. Lu, Q. Li, B. Liu, M. Dianati, C. Feng, S. Chen, and Y. Wang. Robust collaborative 3d object detection in presence of pose errors. In *IEEE International Conference on Robotics and Automation (ICRA)*, May 2023.

[16] N. Glaser, Y.-C. Liu, J. Tian, and Z. Kira. Overcoming obstructions via bandwidth-limited multi-agent spatial handshaking. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2406–2413. IEEE, 2021.

[17] Y. Zhou, J. Xiao, Y. Zhou, and G. Loianno. Multi-robot collaborative perception with graph neural networks. *IEEE Robotics and Automation Letters*, 7(2):2289–2296, 2022.

[18] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. In *Proceedings of The 6th Annual Conference on Robot Learning*, pages 989–1000. PMLR, 2022.

[19] Y. Hu, Y. Lu, R. Xu, W. Xie, S. Chen, and Y. Wang. Collaboration helps camera overtake lidar in 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2023.

[20] H. Zhang, G. Luo, Y. Cao, Y. Jin, and Y. Li. Multi-modal virtual-real fusion based transformer for collaborative perception. In *2022 IEEE 13th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP)*, pages 1–6, 2022.

[21] Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17182–17191, 2022.

[22] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.

[23] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018.

[24] T. Huang, Z. Liu, X. Chen, and X. Bai. Epnet: Enhancing point features with image semantics for 3d object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 35–52. Springer, 2020.

[25] S. Vora, A. H. Lang, B. Helou, and O. Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020.

[26] C. Wang, C. Ma, M. Zhu, and X. Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11794–11803, 2021.

[27] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *arXiv preprint arXiv:2205.13790*, 2022.

[28] Y. Li, X. Qi, Y. Chen, L. Wang, Z. Li, J. Sun, and J. Jia. Voxel field fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1120–1129, 2022.

[29] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1090–1099, 2022.

[30] Y. Hong, H. Dai, and Y. Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 87–104. Springer, 2022.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[32] H. Xiang, R. Xu, and J. Ma. Hm-vit: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer. *arXiv preprint arXiv:2304.10628*, 2023.

[33] Y.-J. Li, J. Park, M. O'Toole, and K. Kitani. Modality-agnostic learning for radar-lidar fusion in vehicle detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 918–927, 2022.

[34] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21361–21370, 2022.

[35] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564, 2021.

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[37] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[38] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019.