Distance-Aware Attentive Framework for Multi-Agent Collaborative Perception in Presence of Pose Error

Extended Abstract

Binyu Zhao
Harbin Institute of Technology
Harbin, China
byzhao@stu.hit.edu.cn

Wei Zhang
Harbin Institute of Technology
Harbin, China
weizhang@hit.edu.cn

Zhaonian Zou
Harbin Institute of Technology
Harbin, China
znzou@hit.edu.cn

ABSTRACT

Multi-agent collaborative perception exchanges information to promote holistic perception, especially for remote and invisible areas that are limited by detection range and occlusion. Due to imperfect localization in practice, it usually suffers from pose estimation error, which can cause spatial message misalignment and performance degradation. Unlike most existing methods using additional module or procedure to correct pose error, we propose a novel framework, *DistAtt*, to suppress pose error and mine useful information simultaneously. It mainly consists of distance-aware feature sampling and cross-agent feature aggregation. The former utilizes diverse pooling kernels to downsample the intermediate features to different multiple granularities, and the latter utilizes specially designed attention mechanism to learn the most critical information. Furthermore, it adopts compensation strategy for more stable optimization. Experimental results show that *DistAtt* significantly suppresses the effect of localization noise and achieves outperformed performance when pose error exists.

KEYWORDS

Multi-Agent Perception; Vehicle-to-Everything (V2X) Application; Pose Error Suppression

ACM Reference Format:
Binyu Zhao, Wei Zhang, and Zhaonian Zou. 2024. Distance-Aware Attentive Framework for Multi-Agent Collaborative Perception in Presence of Pose Error: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Multi-agent collaborative perception aims at sharing complementary perceptual information with neighboring agents to overcome limitations in single-agent view and promote holistic scene comprehension, which attracts continuous attention in recent years. With high-quality supporting datasets emerging [19–22], different methods have been proposed to handle various problems such as performance [1, 9, 16, 18, 18, 20, 23], bandwidth trade-off [7, 15, 24], pose error [10, 13], latency [8] and communication interruption [12]. Among these issues, localization is usually imperfect and therefore produces unwanted relative pose error, which is a grand challenge. To address this problem, previous works often design additional module or procedure to correct relative pose error [5, 10, 13]. However, they increase model complexity and might be inconvenient to follow during inference.

In this paper, we use two alternatives to suppress pose errors without additional module: **i)** Reduce the quantity of received perceptual information. It is believed that the pose error is relevant to relative distance. The uncertainty of pose error increases with the distance between paired agents. **ii)** Apply attention mechanism. Self-attention is a popular choice to model global relationships [3, 14, 17] with the drawbacks of computation and modeling redundancy [2, 4, 6]. Based on these considerations, we propose a novel Distance-aware Attentive framework, *DistAtt*, to suppress the effect of pose error and mine useful information for more accurate and robust collaborative perception. Specifically, we first use distance-aware feature sampling (DFS) to reduce the quantity of collaborative features by pooling based on the distance with ego agent. Since the more distant perceptual information contains more uncertainty and larger relative pose error. Then we apply cross-agent feature aggregation (CFA) to assemble and aggregate lower-resolution spatial features, attentively filtering out the most suitable features and further reducing the ratio of noisy features. Furthermore, we adopt compensation strategy (CS) to stabilize the total number of communicated agents in temporal, which benefits the network optimization and improves the final performance.

2 METHODOLOGY

Considering N agents travels in the scene, let $L_i \in \mathcal{R}^{n\times 3}$ be the raw LiDAR data of the i-th agent, and Y_i be the corresponding ground truth detection. Firstly, agent extracts bird's-eyes-view (BEV) feature $\mathbf{F}_i \in \mathcal{R}^{X\times Y}$ from L_i using an encoder. Then, all feature and pose pairs $\{(\mathbf{F}_j, \xi_j)\}_{j\in\mathcal{N}_i}$ of neighboring agents are transmitted to i-th agent, where $\mathcal{N}_i$ is the set of neighboring agents communicated with i-th agent. Next, each extracted feature $\mathbf{F}_j$ is aligned with the feature $\mathbf{F}_i$ based on their 6 DoF poses ξ_i and ξ_j. After transformation, the i-th agent aggregates the received features with its own to conduct intermediate fusion. Finally, a decoder is implemented to predict the results for a specific task. The objective of collaborative perception is $\min \sum_i g(\hat{Y}_i, Y_i)$, where $g(\cdot,\cdot)$ is the evaluation metric of the specific task. The overview of *DistAtt* is illustrated in Fig 1.

Reduce the effect of relative pose error based on relative distance. *First*, we split the BEV feature $\mathbf{F}_i \in \mathcal{R}^{X\times Y}$ into windows to reduce huge computation and memory cost. When the size

Wei Zhang is the Corresponding Author.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

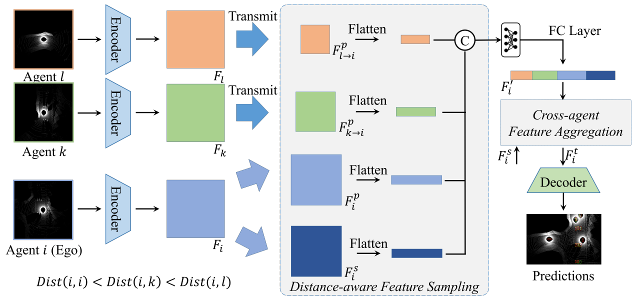

Figure 1: The workflow of *DistAtt* (3-agent scenario, object detection task). Distance-aware Feature Sampling (DFS) suppresses the effect of pose error by using different pooling kernels to downsample the ego and received features. Then features are flattened and gathered to conduct Cross-agent Feature Aggregation (CFA) for further error suppression.

of window is $s \times s$, we obtain the new feature $\mathbf{F}_i^s \in \mathcal{R}^{\frac{h}{s} \times \frac{w}{s} \times s \times s}$. *Then*, we generate feature tokens to suppress relative pose error by using diverse size of pooling kernels. When distance $Dist(i,i) \leq Dist(i,k) \leq Dist(i,l) \leq ...$, we have pooling kernel $p_i \leq p_k \leq p_l \leq ....$ The BEV feature set $\mathbf{F} = \{\mathbf{F}_i, \mathbf{F}_{k\to i}, \mathbf{F}_{l\to i}, ...\}$ are downsampled using these pooling kernels and then processed by separate fully connected layers as $\mathbf{F}_i^p = \{\mathbf{F}_i^p, \mathbf{F}_{k\to i}^p, \mathbf{F}_{l\to i}^p, ...\}$, the size of them is $\frac{h}{p_j} \times \frac{w}{p_j}, j \in i, k, l, ...$ *Finally*, we flatten and concatenate these BEV feature tokens as $\mathbf{F}_i^{'} = f_{concat}(\mathbf{F}_i^s, \mathbf{F}_i^p, \mathbf{F}_{k\to i}^p, \mathbf{F}_{l\to i}^p, ...)$, which covers multi-agent multi-granularity BEV perceptual information.

Mine useful feature and filter out more error information. *First*, we rewrite $\mathbf{F}_i^{'}$ as $\mathbf{F}_i^0$ for convenience. For the t-th layer of CFA, the query Q^t, key K^t, and value V^t are computed using fully connected layers $Q^t = f_{FC}(\mathbf{F}_i^{t-1}), K^t = f_{FC}(\mathbf{F}_i^{'}), V^t = f_{FC}(\mathbf{F}_i^{'})$ where $t \in [1, N]$ and N is the total number of attention layers. *Then*, we conduct multi-head attention mechanism to continuously exploit information and suppress error. The updated BEV feature $F_i^t = f_{FC}(f_{concat}(\{F_i^{t-1} + f_{softmax}(\frac{Q^t \times K^t}{\sqrt{d}}) \times V^t\})) \in \mathcal{R}^{\frac{h}{s} \times \frac{w}{s} \times s \times s}$, where $f_{concat}(\cdot)$ is concatenate operation and d is the scale factor.

Performance problem brought by communication setup. An upperbound N^u is practically set to limit the number of communicated agents and the quantity of message. Correspondingly, the tensor size applying attention mechanism (including element-wise addition and matrix product) in CFA and the total number of used fully connected layers need to be fixed. However, part of the networks can not be fully optimized under this circumstance and *DistAtt* would achieve sub-optimal performance during inference. To solve this problem, we complement the number of agents so that the total number of communicated agents always maintains at the upperbound. The agents to be complemented are the replications of ego agent. This compensation solution introduces no error information for the ego agent comparing with copying information from neighboring agents. And it enhances the exploration of useful information from its own. The number of elements in pooling kernel set is also fixed as N^u. The smaller kernels are allocated for the i-th agent and the copied agents, and the larger kernels are allocated for neighboring agents.

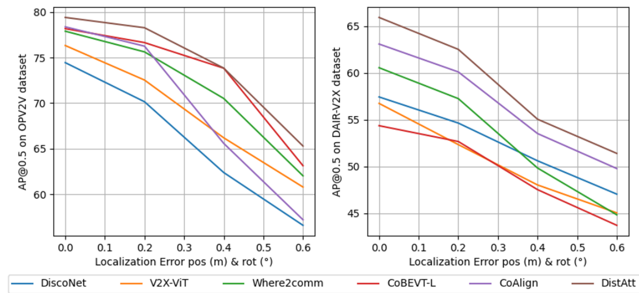

Figure 2: Robustness to localization error. The robustness to localization noise follows the setting in *V2VNet* [16] and *V2X-ViT* [20]. Gaussian noise with a mean of $0m$ and a standard deviation of $0m - 0.6m$ is used.

Table 1: Ablation study on the OPV2V dataset.

	DFS	CFA	CS	AP@0.5	AP@0.7
(a)	✗	✗	✗	76.12	57.17
(b)	✗	✗	✓	76.37	60.33
(c)	✓	✗	✓	77.96	64.86
(d)	✓	✓	✗	77.55	64.67
(e)	✓	✓	✓	79.40	68.17

3 EXPERIMENT

We conduct experiments of 3D object detection task on OPV2V dataset [21] and DAIR-V2X dataset [22]. We implement the method based on the pyTorch [11] framework and OpenCOOD [21] codebase. Weighted cross entropy loss is used for optimization. The detection results are evaluated by Average Precision (AP) at Intersection-over-Union (IoU) threshold of 0.50 and 0.70.

Performance Comparison and robustness to localization noise. Fig 2 shows the detection performance comparisons with state-of-the-art (SOTA) methods [7, 9, 10, 18, 20] and the robustness to localization noise. It is observed that *DistAtt* outperforms previous SOTA models in both simulated and real-world scenarios, which proves the effectiveness and rationality of simultaneously considering the noise suppression and collaboration. Besides, it can be seen that the performance of all methods consistently deteriorates as localization error increases continuously. Noticeably, our method consistently surpasses the previous SOTA models under all noise levels, which evidently demonstrates the robustness of *DistAtt* against pose error.

Ablation study. We also conduct ablation study to investigate the effectiveness of the main components and the necessity of the different designs in our method. The overall results are presented in Table 1. The results of (b)/(c), (c)/(e), and (a)/(b) (or (d)/(e)) reveal their capability of suppressing localization noises especially from distant agents and providing more valuable and less error information for prediction and optimization.

ACKNOWLEDGMENTS

This work was partially supported by the National Natural Science Foundation of China (No. 62072138).

REFERENCES

[1] Qi Chen, Sihai Tang, Qing Yang, and Song Fu. 2019. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, IEEE, Richardson, TX, United states, 514–524.

[2] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. 2021. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems* 34 (2021), 9355–9366.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*. International Conference on Learning Representations, Virtual, Online, 12.

[4] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. IEEE, Virtual, Online, Canada, 6824–6835.

[5] Nathaniel Glaser, Yen-Cheng Liu, Junjiao Tian, and Zsolt Kira. 2021. Overcoming obstructions via bandwidth-limited multi-agent spatial handshaking. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, IEEE, Prague, Czech republic, 2406–2413.

[6] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. 2021. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Virtual, Online, Canada, 11936–11945.

[7] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. 2022. Where2comm: Communication-Efficient Collaborative Perception via Spatial Confidence Maps. In *Advances in Neural Information Processing Systems*, Vol. 35. Neural information processing systems foundation, New Orleans, LA, United states, 4874–4886.

[8] Zixing Lei, Shunli Ren, Yue Hu, Wenjun Zhang, and Siheng Chen. 2022. Latency-aware collaborative perception. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*. Springer, Springer, Tel Aviv, Israel, 316–332.

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

[10] Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, and Yanfeng Wang. 2023. Robust collaborative 3d object detection in presence of pose errors. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, IEEE, London, United kingdom, 4812–4818.

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019), 12.

[12] Shunli Ren, Zixing Lei, Zi Wang, Siheng Chen, and Wenjun Zhang. 2022. Robust Collaborative Perception against Communication Interruption. *the 2nd IJCAI Workshop on Artificial Intelligence for Autonomous Driving* (2022), 9.

[13] Nicholas Vadivelu, Mengye Ren, James Tu, Jingkang Wang, and Raquel Urtasun. 2021. Learning to communicate and correct pose errors. In *Conference on Robot Learning*. PMLR, PMLR, London, United Kingdom, 1195–1210.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017*. Advances in Neural Information Processing Systems, Long Beach, CA, USA, 5998–6008.

[15] Binglu Wang, Lei Zhang, Zhaozhong Wang, Yongqiang Zhao, and Tianfei Zhou. 2023. CORE: Cooperative Reconstruction for Multi-Agent Perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Paris, France, 8710–8720.

[16] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. 2020. 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*. Springer, Springer, Glasgow, United kingdom, 605–621.

[17] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Salt Lake City, UT, USA, 7794–7803.

[18] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. 2022. CoBEVT: Cooperative Bird's Eye View Semantic Segmentation with Sparse Transformers. In *Proceedings of The 6th Annual Conference on Robot Learning*. PMLR, PMLR, Auckland, New zealand, 989–1000.

[19] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. 2023. 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*. IEEE, Vancouver, British Columbia, Canada, 13712–13722.

[20] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. 2022. 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*. Springer, Springer, Tel Aviv, Israel, 107–124.

[21] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. 2022. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, IEEE, Philadelphia, PA, United states, 2583–2589.

[22] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. 2022. 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*. IEEE, New Orleans, LA, United states, 21361–21370.

[23] Yunshuang Yuan, Hao Cheng, and Monika Sester. 2022. Keypoints-based deep feature fusion for cooperative vehicle detection of autonomous driving. *IEEE Robotics and Automation Letters* 7, 2 (2022), 3054–3061.

[24] Binyu Zhao, Wei Zhang, and Zhaonian Zou. 2023. BM2CP: Efficient Collaborative Perception with LiDAR-Camera Modalities. In *Proceedings of The 7th Conference on Robot Learning*. PMLR, Atlanta, GA, United states, 1022–1035.