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:

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, 19, 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 \( L_i \in \mathbb{R}^{n \times 3} \) 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 \( F_i \in \mathbb{R}^{n \times Y} \) from \( L_i \) using an encoder. Then, all feature and pose pairs \((F_i, \xi_i)\) for neighboring agents are transmitted to i-th agent, where \( N_i \) is the set of neighboring agents communicated with i-th agent. Next, each extracted feature \( F_j \) is aligned with the feature \( F_i \) based on their 6 DoF poses \( \xi_i \) and \( \xi_j \). 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 \( \sum_i g(Y_i, Y_i) \), 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 \( F_i \in \mathbb{R}^{n \times Y} \) into windows to reduce huge computation and memory cost. When the size...
of window is \( s \times s \), we obtain the new feature \( F'_i \in \mathbb{R}^{ \frac{h}{p_i} \times \frac{w}{p_i} \times s \times s } \). Then, we generate feature tokens to suppress relative pose error by using diverse size of pooling kernels. When distance \( Dist(i, i) \leq Dist(i, k) \leq Dist(i, l) \) \( \ldots \), we have pooling kernel \( p_i \leq p_k \leq p_l \) \( \ldots \). The BEV feature set \( F = \{ F_i, F_{k \rightarrow i}, F_{l \rightarrow i}, \ldots \} \) are downsampled using these pooling kernels and then processed by separate fully connected layers as \( F^p_i = \{ F^p_{i \rightarrow k}, F^p_{i \rightarrow l}, \ldots \} \), the size of them is \( \frac{h}{p_i} \times \frac{w}{p_i}, j \in i, k, l, \ldots \). Finally, we flatten and concatenate these BEV feature tokens as \( F'_i = \text{concat}( F'_i, F^p_{i \rightarrow k}, F^p_{i \rightarrow l}, \ldots ) \), which covers multi-agent multi-granularity BEV perceptual information.

Mine useful feature and filter out more error information. First, we rewrite \( F_i \) as \( F^0_i \) 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 \( Q^t = f_{FC}(F^t_{i \rightarrow -1}), K^t = f_{FC}(F^t_i), V^t = f_{FC}(F^t_i) \) 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 \( F'_i = f_{FC}(\text{concat}(F'_i, f_{softmax}(Q^t \times K^t \sqrt{d}))) \in \mathbb{R}^{ \frac{h}{p_i} \times \frac{w}{p_i} \times s \times s } \), where \( f_{concat}(\cdot) \) is concatenate operation and \( d \) is the scale factor.

**Performance problem brought by communication setup.**
An upperbound \( N^w \) 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 cannot 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 \( N^w \). The smaller kernels are allocated for the \( i \)-th agent and the copied agents, and the larger kernels are allocated for neighboring agents.
REFERENCES


