Collaborative Multi-Object Tracking with Conformal Uncertainty Propagation

Sanbao Su[1], Songyang Han[3], Yiming Li[2], Zhili Zhang[1], Chen Feng[2], Caiwen Ding[1], and Fei Miao[1]

***Abstract*— Object detection and multiple object tracking (MOT) are essential components of self-driving systems. Accurate detection and uncertainty quantification are both critical for onboard modules, such as perception, prediction, and planning, to improve the safety and robustness of autonomous vehicles. Collaborative object detection (COD) has been proposed to improve detection accuracy and reduce uncertainty by leveraging the viewpoints of multiple agents. However, little attention has been paid to how to leverage the uncertainty quantification from COD to enhance MOT performance. In this paper, as the first attempt to address this challenge, we design an uncertainty propagation framework called MOT-CUP. Our framework first quantifies the uncertainty of COD through direct modeling and conformal prediction, and propagates this uncertainty information into the motion prediction and association steps. MOT-CUP is designed to work with different collaborative object detectors and baseline MOT algorithms. We evaluate MOT-CUP on V2X-Sim, a comprehensive collaborative perception dataset, and demonstrate a 2% improvement in accuracy and a $2.67\times$ reduction in uncertainty compared to the baselines, e.g. SORT and ByteTrack. In scenarios characterized by high occlusion levels, our MOT-CUP demonstrates a noteworthy 4.01% improvement in accuracy. MOT-CUP demonstrates the importance of uncertainty quantification in both COD and MOT, and provides the first attempt to improve the accuracy and reduce the uncertainty in MOT based on COD through uncertainty propagation. Our code is public on `https://coperception.github.io/MOT-CUP/`.**

I. Introduction

Object detection [1] and multiple object tracking (MOT) [2] represent crucial steps of self-driving, and their accuracy and uncertainty quantification (UQ) are important to facilitate various onboard modules including perception, prediction and planning, to improve the safety and robustness of the autonomous systems [3], [4], [5]. Multi-agent collaborative object detection (COD) has been proposed to leverage the viewpoints of multiple agents to enhance detection accuracy compared with individual viewpoints [6], [7]. Numerous studies have demonstrated the advantages of COD in enhancing the detection accuracy [8], [9], [10], [11] and reducing the uncertainty [12]. Currently, Tracking-by-Detection is considered as one of the most effective paradigms [13], using Kalman Filter to predict the next location based on the previous detection results and then performing data association [14], [15], [16], [17].

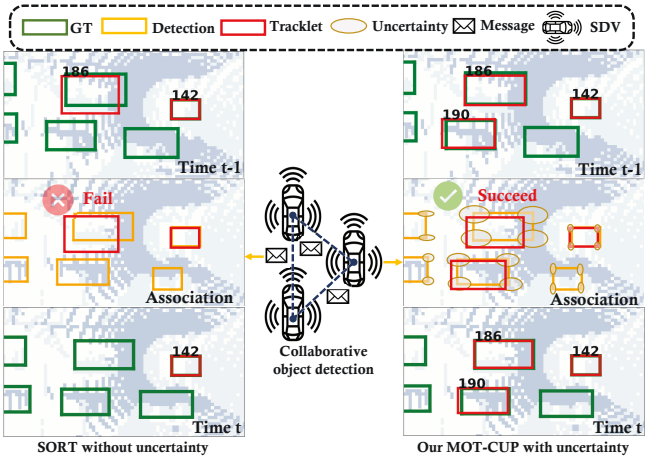

Fig. 1. **Difference in data association for MOT with and without considering uncertainty**. Ground truth bounding boxes are in green, detected bounding boxes in orange, and tracklets' bounding boxes in red, labeled with object IDs. Shadow ellipses indicate uncertainty of the detected bounding box. SORT [14], which doesn't consider uncertainty, is on the left side of the figure, while our MOT-CUP framework, which incorporates uncertainty, is on the right side. At time (t-1), both MOT algorithms output tracklet ID 186. However, at time t, SORT fails to associate the low-quality detection with tracklet 186 due to a large IoU distance. Thus, SORT removes the tracklet. In contrast, our MOT-CUP framework quantifies the uncertainty of COD with a larger shadow ellipse to represent the uncertainty of the bounding box for tracklet 186, and successfully associates the low-quality detection by considering the uncertainty of COD.

However, limitations exist in the methods mentioned above. Existing Kalman Filter (KF) algorithms for motion prediction typically use a fixed measured uncertainty for all detections instead of rigorously calculated uncertainty. Moreover, employing the Intersection-over-Union (IoU) association metric without considering uncertainty in the Hungarian algorithm might not suit poorly detected results due to occlusion. Hence, it remains challenging to rigorously quantify and propagate the uncertainty from COD to MOT to improve the accuracy, for both KF and association steps. For instance, Fig. 1 illustrates how our framework outperforms SORT [14] in associating tracklet 186 (red box) with the low-quality object detection (orange box) at time t. The IoU metric in SORT fails to match them due to the poor detection quality; whereas our framework, by incorporating detection uncertainty (shadow ellipses), effectively associates tracklet 186 even with the low-quality detection. It demonstrates that integrating uncertainty into MOT can improve tracking performance, especially for low-quality detection scenarios.

In this paper, we propose a novel uncertainty propagation framework to improve the performance of these Tracking-

Sanbao Su, Songyang Han, Zhili Zhang and Fei Miao are supported by the National Science Foundation under Grants CNS-1952096 and CNS-2047354. Chen Feng is supported by the National Science Foundation under Grants FRR-2238968, and CPS-2121391. [1]Sanbao Su, Zhili Zhang, Caiwen Ding, and Fei Miao are with the Department of Computer Science and Engineering, University of Connecticut, Storrs Mansfield, CT, USA 06268 {sanbao.su, zhili.zhang, caiwen.ding, fei.miao}@uconn.edu. [2]Yiming Li and Chen Feng are with Tandon School of Engineering, New York University, Brooklyn, NY, USA 11201 {yimingli, cfeng}@nyu.edu. [3]Songyang Han is now with Sony AI, songyang.han@sony.com and this work was done when he was a PhD candidate at the University of Connecticut.

by-Detection algorithms, called MOT-CUP (Multiply Object Tracking with Conformal Uncertainty Propagation). Specifically, our proposed MOT-CUP involves uncertainty quantification of collaborative object detection (COD) via direct modeling and conformal prediction techniques. The uncertainty obtained from the COD stage is subsequently incorporated into both the Kalman Filter and the association procedure of MOT. In particular, we define a new association metric with Negative Log Likelihood (NLL) considering the uncertainty of COD and potential low-quality detection results. Through extensive experiments on V2X-SIM [6] and a series of Tracking-by-Detection MOT algorithms, we show that MOT-CUP framework improves accuracy with up to 2% and reduces uncertainty with up to . In high occlusion-level scenarios, our MOT-CUP achieves a improvement in accuracy. This outcome underscores the effectiveness of our MOT-CUP in challenging scenarios with poor detection. Our results also provide strong validation for the effectiveness of rigorous conformal prediction-based uncertainty quantification in MOT. Overall, our findings highlight the potential benefits of propagating uncertainty quantification into MOT algorithms.

The main contributions of this work are as follows:

1) To the best of our knowledge, our MOT-CUP framework is the first attempt to leverage quantified uncertainty from collaborative object detection to improve MOT performance. This framework can be applied to most object detection models and MOT algorithms.
2) In the collaborative object detection stage, we employ direct modeling and conformal prediction techniques to rigorously quantify the uncertainty.
3) For MOT, we further improve the original MOT algorithm by designing two novel methods that effectively leverage uncertainty information for both the Kalman Filter and association.

II. Related Work

a) Collaborative Object Detection (COD): Collaborative Object Detection (COD) surpasses traditional Single-Agent Object Detection (SOD) by harnessing information from multiple agents or sensors, elevating detection accuracy [6], [7], [18] and mitigating uncertainty [12]. Multi-camera object detection (MOD) [19] methods use strategically positioned cameras of a single agent to improve performance [19]. In complex scenarios such as low-light conditions, occlusions, and adverse weather, COD outperforms MOD for sharing complementary information by communication between multiple agents [20]. Its dynamic integration of insights from various sources enables effective adaptation to changing environments and scene dynamics. Hence, COD harmonizes viewpoints and enables innovative occlusion handling. Moreover, COD extends coverage and precision, with well-designed orchestration to avert redundancy or misalignment [6]. In summary, COD redefines object detection, leveraging collaboration to tackle challenges and chart the future of detection [21]. In this study, we want to show that even though object detection performance has already improved, uncertainty propagation from an advanced model such as COD is still important to enhance the overall performance of subsequent modules.

b) Uncertainty Quantification and Propagation: Uncertainty Quantification (UQ) is vital for the collaborative perception of safety-critical systems such as robots [22], [23] and connected and autonomous vehicles [24], [25], [26]. In self-driving tasks, UQ from COD could improve trajectory prediction [3] and motion planning [5]. However, there is no research on how to leverage COD uncertainty to enhance tracking performance. Several methods for UQ in object detection (OD) require multiple inference runs, not designed for real-time tasks like COD, such as Monte-Carlo dropout [27] and deep ensemble [28], [29].

Direct modeling (DM) [1] methods for OD have been proposed [12], [30]. DM is promising for real-time computer vision tasks, as only requires a single inference pass. However, DM lacks rigorous UQ as the model may easily overfit the training dataset. The work in [12] proposes the bootstrap calibrated DM method for COD. However, the bounding box definition it presented, which relies on corner coordinates, is not congruent with tracking algorithms. Conformal prediction (CP) [31] is a statistical method that converts any heuristic notion of uncertainty (e.g. standard deviation estimations) into rigorous UQ. To rigorously quantify the uncertainty in COD and propagate the uncertainty to MOT, our MOT-CUP framework leverages CP to calibrate the uncertainty estimation from DM.

c) Multiple Object Tracking (MOT): Several recent MOT algorithms [13], [15], [17], [32], [33] use motion models based on Bayesian estimation [34] to predict states by maximizing the posterior estimation. Kalman Filter (KF) [35], a widely used motion model, operates as a recursive Bayes filter that follows a standard predict-update cycle. Current KF-based algorithms typically use a fixed measured uncertainty for all detections without considering rigorously estimated uncertainty to improve the prediction accuracy. In contrast, we propose a rigorous UQ of the COD process based on CP, and integrating it into KF for enhanced accuracy and uncertainty estimation of output.

Data association is a crucial step of MOT, which involves computing the similarity between tracklets and detected objects and utilizing various strategies to match them based on their similarity. The SORT algorithm [14] uses the Intersection over Union (IoU) between predicted and detected boxes to determine their similarity. This approach has proven to be highly competitive on a variety of MOT benchmarks, and serves as a strong baseline for more sophisticated tracking methods. With the similarity, matching strategies assign identities to the objects. This can be achieved through the Hungarian Algorithm [36] or greedy assignment [33]. ByteTrack [13] utilizes similarities between the low-quality detections and tracklets to recover accurate object identities, improving data association performance. However, using IoU distance as a similarity metric may not be appropriate for low-quality detections as explained in Fig. 1. The Mahalanobis distance is also a widely used similarity score by

quantifying the dissimilarity between the detected objects and the distributions of tracklets from the KF model [37]. Nonetheless, substantial uncertainty could result in minimal distances, leading to erroneous matching [38], [17]. The work in [38] proposes the association log-likelihood distance to overcome this problem by computing the logarithmized association probability. Unlike conventional similarity scores, we propose the Negative Log Likelihood (NLL) similarity score, which computes the NLL between the detected distribution of the objects and the mean of the tracklets to focus on the distribution of detections. So it can ignore the large uncertainty and facilitate more accurate associations for low-quality detections.

III. Methodology

A. Approach Overview

We design a novel framework for uncertainty propagation of collaborative object detection (COD) to MOT, named Multiple Object Tracking with Conformal Uncertainty Propagation (MOT-CUP). Fig. 2 presents the methodology overview. The major novelties are: *(1)* MOT-CUP framework first rigorously quantifies the uncertainty in the COD stage based on direct modeling and conformal prediction. *(2)* The uncertainty information is leveraged in the motion prediction stage of MOT, where a Standard Deviation-based Kalman Filter (SDKF) takes the uncertainty quantification (UQ) of COD as its input to improve the predicted precision of location. *(3)* We utilize the Negative Log Likelihood as the similarity metric for the association step, called NLLAI, to improve the accuracy and reduce the uncertainty of MOT.

In this section, we first introduce the conformal prediction in Subsection III-B as preliminary literature of UQ and a useful method to construct predicted uncertainty. We describe our proposed MOT-CUP (Multiply Object Tracking with Conformal Uncertainty Propagation) method as shown in Algorithm 1 and Subsection III-C, followed by the detailed process of UQ of COD based on direct modeling and conformal prediction in Subsection III-D, and uncertainty propagation to MOT in Subsection III-E.

B. Preliminary

Conformal prediction (CP) [31] is a statistical method to generate prediction sets for any model. It is a method to convert any heuristic notion of uncertainty (e.g. an estimate of the standard deviation) to rigorous UQ. For example, we assume that an uncertain scalar follows Gaussian distribution and train a model to output the mean and standard deviation. To be precise, we choose to model $\mathcal{N}$, where is a testing data, and is the corresponding label. We train and to maximize the likelihood of the data. Conformal prediction can turn this heuristic uncertainty notion into rigorous prediction intervals of the form , where is a quantile found by CP.

Consider the validation data with data points that are never seen during training, the CP for input and output includes the following steps: *(1)*

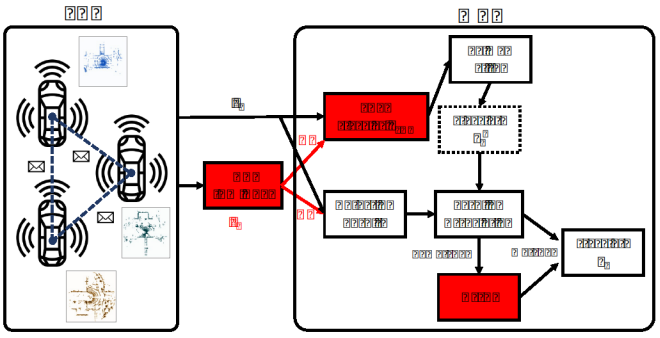

Fig. 2. **Overview of our MOT-CUP framework.** The red color highlights the novelties and important techniques in our MOT-CUP framework. In the collaborative object detection (COD) stage, we rigorously calculate uncertainty quantification (UQ) of each object detection via direct modeling (DM) and conformal prediction (CP). In the motion prediction stage of MOT, we adopt a Standard Deviation-based Kalman Filter (SDKF) to enhance the Kalman Filter process, that leverages the UQ results and predicts the locations of the objects in the next time step with higher precision. In the association step, we first apply the baseline association method and then associate the unmatched detections and tracklets with the Negative Log Likelihood similarity metric, called NLLAI.

Define the score function . (Smaller scores encode better agreement between and). *(2)* Compute as the ———— quantile of the validation scores , where is a user-chosen error rate. *(3)* Use this quantile to form the prediction sets $\mathcal{C}$ for new examples:

$\mathcal{C}$ (1)

Note that is a fresh test point from the same distributions of the validation data. The CP provides a coverage guarantee, as stated in the following lemma.

*Lemma 3.1 (**Conformal Coverage Guarantee** [31]):* Suppose and are , then the following holds:

$\mathcal{C}$ ——— (2)

In other words, the probability that the prediction set contains the correct label is almost exactly .

C. MOT-CUP Algorithm

The detail of MOT-CUP is presented in Algorithm 1. For each frame in the point cloud sequence , there are objects. For each frame, the trained collaborative object detector with direct modeling would generate a set of detected objects $\mathcal{D}$ (Line 3). The set includes the predicted classification probability and the location of each object . The location of each object is represented by random variables parameterized by where is the mean and is the standard deviation for -th variable. This object detector not only predicts the location of each object but also provides a measure of uncertainty.

To provide more accurate measures of uncertainty, we leverage the quantiles computed by CP to adjust the standard deviation (see Lines 4-8). Then, to track the detected objects across multiple frames, we employ a Kalman Filter to predict the current state of the tracklets, which is commonly used in MOT [14] (see Lines 10-12). In the association step,

we first apply the origin association method (see Line 13) and store all matched pairs in $\mathcal{A}$ for $\mathcal{D}$ and $\mathcal{T}$. Then we associate the unmatched detections and tracklets with the Negative Log Likelihood similarity metric for lower-quality detected objects (see Line 14). The detail of Negative Log Likelihood-based Association Improvement (NLLAI) will be introduced in Algorithm 2. To update the tracklets with the matched detections, we go beyond the traditional MOT algorithms by incorporating the detected standard deviation in addition to the detected mean (see Lines 16-18). This allows us to more accurately model the uncertainty associated with each detection and incorporate it into the tracklet. The detected standard deviations are also applied to generating new tracklets with unmatched detections (see Lines 20-22).

Algorithm 1: Multiple Object Tracking with Conformal Uncertainty Propagation (MOT-CUP)

Data: input point cloud sequence , the trained collaborative object detector , NLL threshold , baseline's Kalman Filter , baseline's association method , quantile of CP
Result: Tracklet list

```
1  Initialization:
2  for point cloud frame in in time sequence do
3  |                                  , where each location
   |   contains
4  |  for each object do
5  |  |  for i from 1 to I do
6  |  |  |
7  |  |  end
8  |  end
9  |  /* Adjust standard deviation by CP  */
10 |  for in do
11 |  |  Apply Kalman Filter (KF) [35]
12 |  end
13 |
14 |
   |   NLLAI
15 |  /* NLLAI is Algorithm 2              */
16 |  for in do
17 |  |  Apply KF with updated standard deviation
18 |  end
19 |
20 |  for in do
21 |  |                where
22 |  end
23 end
```

D. UQ on Collaborative Object Detection

We use direct modeling [12], [39] to estimate the standard deviation of each variable of the COD stage. We assume that all variables are independent and the distribution of each variable is a single-variate Gaussian distribution. For the distribution of each variable of the ground truth, we assume it as a Dirac delta function [39]. Then we define the regression loss function for the -th variable as the Kullback-Leibler (KL) divergence between the single-variate Gaussian distribution and the Dirac delta function [40]:

$\mathcal{L}$ ———— (3)

where is the ground-truth value for -th variable. An additional regression header is incorporated to forecast all standard deviations , with a comparable structure as the regression header for . This is accomplished based on the original collaborative object detector where no alterations have been made to the remaining components.

After we have the trained object detection model, we compute the quantile for the standard deviation of each variable by CP [31] based on the validation dataset, as introduced in Subsection III-B.

We define the score function for the -th variable as:

———— (4)

where is the point cloud input and it can be comprehended as a multiplicative correction factor applied to the standard deviation where .
After testing the detection model on the validation dataset and calculating the score function, we obtained a set of scores for the -th variable where is the number of all detected objects in all the frames in the validation set. Given an error rate , we select the quantile as the ———— quantile of the score set. The prediction set for is constructed following the proposition.

Proposition 1: When we assume the uncertain scalar for -th variable follows the Gaussian distribution with mean and standard deviation and select the score function as ———— in CP, the prediction set for -th variable is $\mathcal{C}$.

Proof: From Lemma 3.1, for a test point , it holds that

$\mathcal{C}$

Eq. (1)

———— Eq. (4)

$\mathcal{C}$ (5)

■

Then we adjust the standard deviation by to achieve rigorously estimated uncertainty.

E. Uncertainty Propagation to MOT

After obtaining the corrected standard deviation for each variable of detected objects, how to utilize and propagate it into the MOT stage remains a significant challenge. Here, we propose SDKF and NLLAI methods to leverage the uncertainty in both the motion prediction and association which are the primary steps of MOT.

Standard Deviation-based Kalman Filter (SDKF): As shown in Section II, Kalman Filter (KF) [41] is one important step for motion prediction. The inputs of KF encompass the observed state and measurement uncertainty. Compared to the existing MOT algorithm, we leverage our rectified standard deviation as the measurement uncertainty in place

of the pre-established values. By taking into account both the mean and standard deviation of the detections, we are able to better account for the uncertainty of objects and provide more robust tracklets over time. SDKF does not significantly impact the time complexity of algorithms, as it only modifies the measurement uncertainty input from fixed values to rigorously estimated ones.

Algorithm 2: NLL-based Association Improvement method (NLLAI)

Data: Matched detection and tracklet list $\mathcal{A}_{matched}$, unmatched detection list $\mathcal{D}_{unmatched}$, unmatched tracklet list $\mathcal{T}_{unmatched}$, NLL threshold τ
Result: New association results $\mathcal{A}'_{matched}$, $\mathcal{D}'_{unmatched}$, $\mathcal{T}'_{unmatched}$

1 Similarity matrix $SNLL = NLL(\mathcal{D}_{unmatched}, \mathcal{T}_{unmatched})$
2 $\mathcal{A}'_{matched} \leftarrow$ assoicate $\mathcal{D}_{unmatched}$ and $\mathcal{T}_{unmatched}$ by Hungarian Algorithm with $SNLL$
3 $\mathcal{D}'_{unmatched} \leftarrow \emptyset, \mathcal{T}'_{unmatched} \leftarrow \emptyset$
4 **for** (d, t) *in* $\mathcal{A}'_{matched}$ **do**
5 **if** $SNLL(d, t) > \tau$ **then**
6 $\mathcal{A}'_{matched} = \mathcal{A}'_{matched} \backslash \{(d, t)\}$
7 $\mathcal{D}'_{unmatched} = \mathcal{D}'_{unmatched} \cup \{d\}$
8 $\mathcal{T}'_{unmatched} = \mathcal{T}'_{unmatched} \cup \{t\}$
9 **end**
10 **end**
11 $\mathcal{A}'_{matched} = \mathcal{A}_{matched} \cup \mathcal{A}'_{matched}$

Negative Log Likelihood-based Association Improvement (NLLAI): Using Intersection over Union (IoU)-based similarity score cannot match low-quality detection results as shown in Section I, which poses a significant challenge during the association stage. To address this issue, we propose the NLLAI technique as shown in Algorithm 2. We first define Negative Log Likelihood (NLL) between the predicted locations of tracklets and the detected locations as a novel similarity score:

$$snll = -\frac{1}{I}\sum_{i=1}^{I} \log P(\dot{y}_i | \hat{y}_i, \hat{\sigma}_i), \tag{6}$$

where $\dot{y}_i$ is the predicted value for the i-th variable of the tracklet from the motion prediction model such as KF. As Subsection III-D, the distribution of each i-th variable for detected objects is a single-variate Gaussian distribution where $\hat{y}_i$ is the mean and $\hat{\sigma}_i$ is the standard deviation. Given the set of unmatched detections and unmatched tracklets after the original association method, we compute the NLL similarity matrix $SNLL$ with Equation 6 (see Line 1).

Then we utilize the Hungarian algorithm to establish associations between unmatched detections and unmatched tracklets based on $SNLL$. To eliminate matched pairs with high NLL scores, we introduce a hyperparameter denoted by τ as the NLL score threshold. Specifically, any matched pairs with $s_{NLL} > \tau$ shall be deemed ineligible for further consideration (see Lines 4-10).

The time complexity of NLLAI depends on the number of input unmatched detections N_D and the number of input unmatched tracklets N_T. The time complexity of computing NLL can be optimized to be $O(1)$ [42], so computing the similarity matrix needs $O(N_D N_T)$ time. Assuming $N_D > N_T$, the time complexity of associating with the Hungarian Algorithm can be $O(N_D^3)$ [36]. Thus the time complexity of our NLLAI is $O(N_D^3)$ which is polynomial.

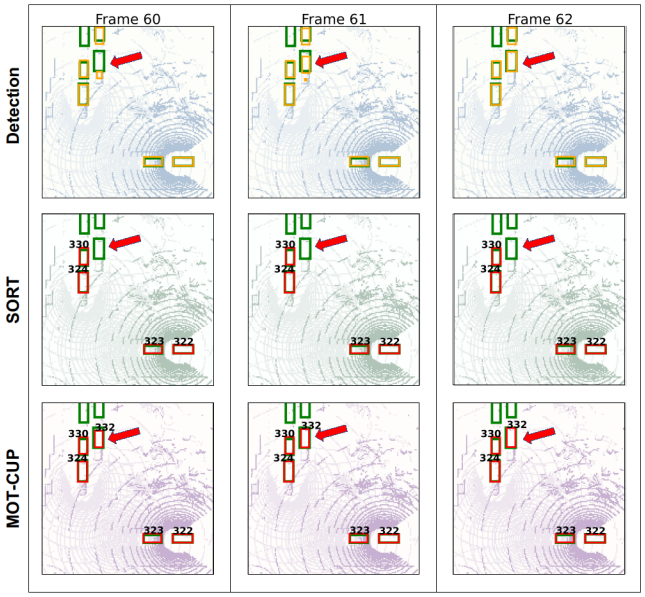

Fig. 3. **Visualization of results of the detection, original SORT, and our MOT-CUP framework over consecutive three frames.** The collaborative object detector here is Upper-bound. Green boxes are ground truth bounding boxes, orange boxes are detected bounding boxes, and red boxes are tracklets' bounding boxes as the output of MOT. The numbers next to the red boxes indicate object IDs. We observe that our MOT-CUP outperforms the original SORT algorithm in tracking object 332, as indicated by the red arrow. Furthermore, MOT-CUP improves the accuracy of location, compared with the object detector, such as object 332 in frame 60. Overall, our results demonstrate the importance of considering uncertainty in MOT.

IV. Experiment

A. Experimental Setups

We evaluate the uncertainty propagation framework MOT-CUP using the V2X-Sim dataset [6], which comprises 80 scenes for training, 10 scenes for validation, and 10 scenes for testing. V2X-Sim was generated using the CARLA simulation [43]. Each scene includes 100 time-series frames and features 2-5 connected vehicles, from which 3D point clouds are collected using LiDAR sensors. Except for V2X-Sim, there are currently no other open-source datasets tailored explicitly to support COD and MOT. Therefore, our experiments focus solely on utilizing the V2X-Sim dataset. The host machine is a server with Intel Core i9-10900X processors and four NVIDIA Quadro RTX 6000 GPUs.

MOT methods use the tracking-by-detection framework. Object detection quality has a significant impact on tracking performance [14]. We consider three collaborative object detectors for all tracking approaches as follows:

Lower-bound (LB) [6]: The single-agent object detector, which operates independently by utilizing point cloud data from one single LiDAR sensor without the need for collaboration with other detectors.

DiscoNet (DN) [10]: The intermediate collaborative object detector employs a directed graph with matrix-valued edge weight to dynamically aggregate features from various

agents. It demonstrates a favorable trade-off between performance and bandwidth.

Upper-bound (UB) [6]: The early collaborative object detector employs raw point cloud data from all connected vehicles to facilitate collaboration. This detector achieves high performance, while retaining lossless information. However, the approach often requires high communication bandwidth.

We apply our uncertainty propagation framework to two tracking baselines, SORT [14] and ByteTrack [13], and compare their performance in accuracy and uncertainty. SORT [14] is a pragmatic approach with simple, effective algorithms by using the Kalman Filter for estimation and the Hungarian algorithm for data association. Instead of only associating detection boxes with high scores, ByteTrack [13] also utilizes similarities between the low score detection boxes and tracklets to improve the performance on data association. In our MOT-CUP, we select the NLL threshold for SORT and for ByteTrack. Other hyperparameters such as the IoU threshold, are directly inherited from the original designs of [6], [13], [14], [31].

B. Evaluation Metrics

Accuracy Metrics:

Higher Order Tracking Accuracy (HOTA) [44]: captures the effect of accurate object detection, association, and localization in a well-balanced way. Such a unified measure captures the synergistic impact of these critical aspects and most comprehensively assesses the algorithm's effectiveness.

Multiple Object Tracking Accuracy (MOTA) [45]: quantifies missed detections, false positives and false negatives for detection, and identity switches for the association.

Multiple Object Tracking Precision (MOTP) [45]: measures the ability to estimate precise object locations.

Frames Per Second (FPS): refers to the number of frames processed per second, and measures the time complexity.

It is important to note that higher values of the aforementioned performance metrics indicate better performance in the context of MOT evaluation. When assessing the performance of MOT algorithms on the same object detector, even slight improvements in HOTA, MOTA, and MOTP mean good progress, as reported by [13], [28].

Uncertainty Metrics:

Negative Log Likelihood (NLL) [46]: a prevalent metric employed to assess the level of uncertainty in the predicted probability distribution of a given test dataset [3], [28], [30].

Continuous Ranked Probability Score (CRPS) [47]: measures the discrepancy between predicted and ground-truth probability distributions [48], [49].

Lower values indicate more precise uncertainty estimation.

C. Accuracy Evaluation

The outcomes of our framework on the V2X-SIM dataset with three distinct object detectors and two diverse MOT baselines are presented in Table I. The results indicate that our MOT-CUP framework is capable of leveraging quantified uncertainty from COD to enhance the performance of all original MOT algorithms, with up to improvement on HOTA, up to improvement on MOTA and up to improvement on MOTP. The performance of object detectors can significantly impact the performance of MOT. Specifically, when the object detector is capable of detecting more objects, such as Upper-bound, our framework can significantly enhance the performance of MOT algorithms.

TABLE I

PERFORMANCE EVALUATION OF OUR UNCERTAINTY PROPAGATION FRAMEWORK ON DIFFERENT MOT BASELINES AND OBJECT DETECTORS

Base	Detector	Method	HOTA ↑	MOTA ↑	MOTP ↑	FPS↑
SORT [14]	UB	Base	41.34	52.60	85.42	1026
		Our	42.19	53.72	86.07	877
	DN	Base	41.80	50.79	85.42	1052
		Our	42.49	51.73	85.85	885
	LB	Base	31.28	27.09	85.62	1568
		Our	31.69	27.70	85.69	1317
Byte-Track [13]	UB	Base	42.05	52.90	84.47	1251
		Our	42.56	53.77	85.49	1067
	DN	Base	42.64	51.48	84.01	1153
		Our	43.14	52.28	84.94	1074
	LB	Base	32.27	29.15	84.57	1637
		Our	32.55	29.65	84.97	1457

TABLE II

PERFORMANCE EVALUATION OF OUR MOT-CUP FRAMEWORK ON VARIOUS OCCLUSION-LEVEL SCENARIOS.

Scenario	Method	HOTA ↑	MOTA ↑	MOTP ↑
High OCL	Base	29.00	34.57	68.13
	Our	30.16 (4.01%)	36.14 (4.52%)	68.85 (1.06%)
Low OCL	Base	43.51	57.24	85.51
	Our	44.25 (1.71%)	58.03 (1.39%)	86.07 (0.67%)

MOT-CUP on high occlusion-level scenarios: We divide the entire test dataset into two subsets: one with high occlusion scenarios and the other with low occlusion scenarios. We conduct experiments using our MOT-CUP with SORT and Upper-bound, as in Table II. The results demonstrate that our MOT-CUP exhibits superior improvements in high occlusion-level scenarios, with a notable 4.01% enhancement in HOTA compared to a 1.71% improvement in HOTA for low occlusion-level scenarios. In high occlusion-level scenarios, the presence of poorly detected objects caused by occlusion leads to high uncertainty, which our SDKF and NLLAI utilize to enhance the tracking performance.

Fig. 3 presents visualizations of Upper-bound, original SORT, and our MOT-CUP framework's results over three consecutive frames. Our MOT-CUP outperforms the original SORT in tracking object 332, as indicated by the red arrow. Moreover, our MOT-CUP improves location accuracy, as shown for object 332 in frame 60, compared to the object detector. These results showcase the effectiveness of our approach in accurately tracking objects, even in challenging scenarios with poor detection or occlusion. Additionally, incorporating uncertainty into the Kalman Filter and association step enables better tracking performance over time.

D. Uncertainty Evaluation

We use Negative Log Likelihood (NLL) [46] and Continuous Ranked Probability Score (CRPS) [47] at IoU thresholds of 0.5 and 0.7 as the uncertainty measurement. We compare the uncertainty results of different UQ methods on object detection and MOT-CUP, including dropout (DO), deep ensemble (DE), and our conformal prediction (CP) in Table III.

TABLE III

NLL & CRPS COMPARISONS ON DETECTION AND MOT-CUP WITH DIFFERENT UNCERTAINTY QUANTIFICATION METHODS: DROPOUT (DO), DEEP ENSEMBLE (DE) AND CONFORMAL PREDICTION (CP). THE BEST RESULTS ARE SHOWN IN **BOLD**.

Base	Method	DO	DE	CP	NLL @IoU=0.5 ↓			NLL @IoU=0.7 ↓			CRPS @IoU=0.5 ↓			CRPS @IoU=0.7 ↓		
					UB	DN	LB	UB	DN	LB	UB	DN	LB	UB	DN	LB
SORT [14]	Detection				193	222	335	96	147	166	0.453	0.498	0.693	0.392	0.459	0.514
		✓			81.01	52.97	263	36.05	26.98	124	0.512	0.554	0.745	0.453	0.517	0.572
			✓		44.01	46.98	128	23.64	29.30	63.80	0.482	0.518	0.703	0.423	0.480	0.531
				✓	25.70	26.21	25.17	14.54	19.50	13.04	0.424	0.466	0.652	0.364	0.427	0.475
	MOT-CUP				9.61	12.09	14.45	7.87	11.21	11.06	0.312	0.355	0.463	0.297	0.347	0.409
		✓			6.54	3.48	18.34	4.38	3.44	11.40	0.345	0.379	0.489	0.331	0.374	0.436
			✓		2.56	2.31	6.15	2.17	2.23	3.82	0.338	0.360	0.483	0.324	0.354	0.431
				✓	**0.94**	**0.95**	**1.30**	**0.74**	**0.90**	**1.06**	**0.301**	**0.336**	**0.444**	**0.286**	**0.329**	**0.392**
Byte-Track [13]	Detection				1801	540	461	870	198	121	0.392	0.391	0.597	0.338	0.357	0.358
		✓			101	35.08	269	41.17	24.86	101	0.524	0.561	0.792	0.468	0.531	0.550
			✓		72.56	38.41	156.59	35.99	30.51	57.10	0.492	0.523	0.753	0.436	0.493	0.518
				✓	43.65	32.50	68.94	23.47	19.25	11.78	0.381	0.376	0.596	0.328	0.343	0.358
	MOT-CUP				29.49	25.30	57.23	17.67	17.57	11.27	0.302	0.318	0.412	0.276	0.308	0.310
		✓			24.35	7.66	26.43	6.97	7.37	21.70	0.385	0.436	0.532	0.359	0.432	0.437
			✓		23.52	11.61	24.62	6.20	9.89	20.77	0.362	0.406	0.512	0.336	0.402	0.421
				✓	**20.01**	**6.94**	**4.05**	**2.41**	**1.99**	**0.99**	**0.280**	**0.286**	**0.376**	**0.254**	**0.275**	**0.276**

TABLE IV

ABLATION STUDY ON MOT WITH THE UPPER-BOUND AND DISCONET DETECTORS. THE BEST RESULTS ARE SHOWN IN **BOLD**.

Base	CP	SDKF	NLLAI	Upper-bound				DiscoNet			
				HOTA ↑	MOTA ↑	MOTP ↑	FPS ↑	HOTA ↑	MOTA ↑	MOTP ↑	FPS ↑
SORT [14]				41.34	52.60	85.42	**1026**	41.80	50.79	85.42	**1052**
		✓		41.67	52.35	86.25	962	42.23	50.79	86.15	985
			✓	41.34	52.60	85.42	865	41.80	50.79	85.42	873
	✓	✓		41.80	52.65	86.20	991	42.23	50.93	86.09	993
	✓		✓	41.73	53.50	84.81	911	41.98	51.32	84.70	899
		✓	✓	41.67	52.35	**86.25**	841	42.23	50.80	**86.15**	852
	✓	✓	✓	**42.19**	**53.72**	86.07	877	**42.49**	**51.73**	85.85	885
Byte-Track [13]				42.05	52.90	84.47	**1251**	42.64	51.48	84.01	**1153**
		✓		42.49	53.63	85.82	1195	42.85	52.13	85.29	1078
			✓	42.06	52.97	84.45	1072	42.67	51.61	83.98	1010
	✓	✓		42.62	53.85	85.54	1219	43.09	52.21	85.04	1141
	✓		✓	42.12	53.21	84.30	1057	42.68	51.65	83.89	1029
		✓	✓	42.50	53.68	**85.80**	1091	42.89	52.26	**85.26**	1042
	✓	✓	✓	**42.56**	**53.77**	85.49	1067	**43.14**	**52.28**	84.94	1074

The vanilla baseline only utilizes direct modeling (DM). The implementations of DO and DE are as same as [1]. The representation formats of bounding boxes utilized by SORT and ByteTrack diverge, necessitating the training of distinct detection models. Consequently, the results on uncertainty are dissimilar between SORT and ByteTrack.

For NLL, CP outperforms all baselines, with up to improvement. In particular, CP achieves up to 95% reduction compared to DO and DE. Furthermore, in comparison to object detection, our MOT-CUP framework with CP produces precise uncertainty estimation, with up to improvement. The vanilla object detection shows a significantly large NLL for the DM overfits the training dataset and overestimates the uncertainty of the test dataset.

Compared with all baselines, CP can effectively reduce the CRPS, with up to 37% reduction. Specially, it achieves up to 31%, 37% and 35% reduction compared with the vanilla baseline, DO and DE. The reason that DO and DE increase the CRPS might be they cannot fully capture the entire distribution of possible values while CRPS requires the entire predicted distribution to be considered [48].

E. Ablation Study on Accuracy

We conduct an ablation study to evaluate the contributions of each proposed technique in our MOT-CUP framework as shown in Table IV with two different detectors and two diverse MOT baselines. CP is shown to contribute significantly to both SDKF and NLLAI. NLLAI, which focuses on refining the association step with a new SNLL metric, yields marked improvements in metrics capturing associations such as HOTA and MOTA. However, an increase in matching potential may lead to a decline in the precision of object localization, as reflected by the decrease in MOTP metric. In contrast, SDKF, where the Kalman Filter takes the COD uncertainty information as its input, primarily enhances metrics measuring localization, such as HOTA and MOTP, thereby improving the accuracy of object localization estimates. Notably, our proposed framework combined with diverse collaborative object detectors and MOT baselines always achieves the optimal performance outcomes.

a) Limitation: In terms of FPS, our framework results in an average decrease of 13.2%, yet it does not affect the real-time capacity of the MOT algorithms. It is noteworthy that the increase in time incurred by our framework is polynomial, as discussed in Section III. Furthermore, we have not implemented any specific strategies aimed at optimizing the quality of code with respect to running time. Therefore, the computational overhead of our framework is acceptable.

V. CONCLUSION

This paper presents the first attempt to leverage uncertainty quantification from collaborative object detection (COD) to

enhance the performance of multiple object tracking (MOT). Our proposed framework, MOT-CUP, employs direct modeling and conformal prediction techniques to quantify the uncertainty in COD. The uncertainty of COD is propagated to the Kalman Filter and the Negative Log Likelihood-based Association Improvement (NLLAI) procedure of MOT. We evaluate MOT-CUP on various CODs and MOT baselines, and demonstrate that our framework significantly improves both the accuracy and uncertainty of the original MOT. Our findings highlight and validate the benefits of incorporating COD uncertainty quantification into MOT algorithms. In future work, we plan to extend our method to popular single-agent object detection and MOT benchmarks, such as KITTI and nuScenes, and more MOT baselines.

REFERENCES

[1] D. Feng, A. Harakeh, S. L. Waslander, and K. Dietmayer, "A review and comparative study on probabilistic object detection in autonomous driving," *IEEE TITS*, 2021.

[2] C. Luo, X. Yang, and A. Yuille, "Exploring simple 3d multi-object tracking for autonomous driving," in *ICCV*, 2021, pp. 10 488–10 497.

[3] I. Boris, L. Yifeng, S. Shubham, C. Punarjay, and P. Marco, "Propagating state uncertainty through trajectory forecasting," in *ICRA*, 2022.

[4] S. Han, S. Zhou, J. Wang, L. Pepin, C. Ding, J. Fu, and F. Miao, "A multi-agent reinforcement learning approach for safe and efficient behavior planning of connected autonomous vehicles," *arXiv:2003.04371*, 2022.

[5] W. Xu, J. Pan, J. Wei, and J. M. Dolan, "Motion planning under uncertainty for on-road autonomous driving," in *ICRA*, 2014.

[6] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE RA-L*, 2022.

[7] 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 *ICRA*, 2022.

[8] E. Arnold, M. Dianati, R. de Temple, and S. Fallah, "Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors," *IEEE TITS*, 2020.

[9] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds," in *ICDCS*, 2019, pp. 514–524.

[10] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," *NeurIPS*, pp. 29 541–29 552, 2021.

[11] Y. Li, J. Zhang, D. Ma, Y. Wang, and C. Feng, "Multi-robot scene completion: Towards task-agnostic collaborative perception," in *CoRL*, 2022.

[12] S. Su, Y. Li, S. He, S. Han, C. Feng, C. Ding, and F. Miao, "Uncertainty quantification of collaborative detection for self-driving," in *ICRA*, 2023.

[13] Y. Zhang, P. Sun, Y. Jiang, D. Yu, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in *arXiv preprint arXiv:2110.06864*, 2021.

[14] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *ICIP*, 2016, pp. 3464–3468.

[15] J. Cao, X. Weng, R. Khirodkar, J. Pang, and K. Kitani, "Observation-centric sort: Rethinking sort for robust multi-object tracking," *arXiv preprint arXiv:2203.14360*, 2022.

[16] J. Pang, L. Qiu, X. Li, H. Chen, Q. Li, T. Darrell, and F. Yu, "Quasi-dense similarity learning for multiple object tracking," in *CVPR*, 2021.

[17] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *ICIP*, 2017.

[18] X. Cai, W. Jiang, R. Xu, W. Zhao, J. Ma, S. Liu, and Y. Li, "Analyzing infrastructure lidar placement with realistic lidar," *arXiv preprint arXiv:2211.15975*, 2022.

[19] Y. Jiang, L. Zhang, Z. Miao, X. Zhu, J. Gao, W. Hu, and Y.-G. Jiang, "Polarformer: Multi-camera 3d object detection with polar transformer," in *AAAI*, vol. 37, no. 1, 2023, pp. 1042–1050.

[20] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," *NeurIPS*, vol. 35, pp. 4874–4886, 2022.

[21] 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 *ECCV*, 2022.

[22] A. M. Z. Jasour and B. C. Williams, "Risk contours map for risk bounded motion planning under perception uncertainties," *Robotics: Science and Systems XV*, 2019.

[23] S. Dean, A. Taylor, R. Cosner, B. Recht, and A. Ames, "Guaranteeing safety of learned perception modules via measurement-robust control barrier functions," in *CoRL*, 2021, pp. 654–670.

[24] S. Han, S. Su, S. He, S. Han, H. Yang, and F. Miao, "What is the solution for state adversarial multi-agent reinforcement learning?" *arXiv:2212.02705*, 2022.

[25] S. Han, H. Wang, S. Su, Y. Shi, and F. Miao, "Stable and efficient shapley value-based reward reallocation for multi-agent reinforcement learning of autonomous vehicles," in *ICRA*. IEEE, 2022.

[26] S. He, S. Han, S. Su, S. Han, S. Zou, and F. Miao, "Robust multi-agent reinforcement learning with state uncertainty," *TMLR*, 2023.

[27] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf, "Dropout sampling for robust object detection in open-set conditions," in *ICRA*. IEEE, 2018, pp. 3243–3249.

[28] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *NeurIPS*, vol. 30, 2017.

[29] Z. Lyu, N. Gutierrez, A. Rajguru, and W. J. Beksi, "Probabilistic object detection via deep ensembles," in *ECCV*. Springer, 2020, pp. 67–75.

[30] G. P. Meyer and N. Thakurdesai, "Learning an uncertainty-aware object detector for autonomous driving," in *IROS*, 2020.

[31] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv preprint arXiv:2107.07511*, 2021.

[32] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in *ICCV*, 2015, pp. 3029–3037.

[33] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *ECCV*. Springer, 2020, pp. 474–490.

[34] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 2006.

[35] R. E. Kalman *et al.*, "Contributions to the theory of optimal control," *Bol. soc. mat. mexicana*, vol. 5, no. 2, pp. 102–119, 1960.

[36] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[37] M. Bertozzi, A. Broggi, A. Fascioli, A. Tibaldi, R. Chapuis, and F. Chausse, "Pedestrian localization and tracking system with kalman filtering," in *IV*. IEEE, 2004, pp. 584–589.

[38] R. Altendorfer and S. Wirkert, "Why the association log-likelihood distance should be used for measurement-to-track association," in *2016 IEEE intelligent vehicles symposium (IV)*. IEEE, 2016, pp. 258–265.

[39] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding box regression with uncertainty for accurate object detection," in *CVPR*, 2019, pp. 2888–2897.

[40] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT Press, 2012.

[41] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, 1960.

[42] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, 2007.

[43] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *CoRL*, 2017, pp. 1–16.

[44] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe, "Hota: A higher order metric for evaluating multi-object tracking," *IJCV*, vol. 129, no. 2, pp. 548–578, 2021.

[45] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *JIVP*, vol. 2008, pp. 1–10, 2008.

[46] A. Harakeh and S. L. Waslander, "Estimating and evaluating regression predictive uncertainty in deep object detectors," in *ICLR*, 2021.

[47] V. V. V'yugin and V. G. Trunov, "Online learning with continuous ranked probability score," in *COPA*, 2019, pp. 163–177.

[48] Y. Chung, I. Char, H. Guo, J. Schneider, and W. Neiswanger, "Uncertainty toolbox: an open-source library for assessing, visualizing, and improving uncertainty quantification," *arXiv preprint arXiv:2109.10254*, 2021.

[49] A. Korotin, V. V'yugin, and E. Burnaev, "Mixing past predictions," in *COPA*, 2020, pp. 171–188.