

SOTIF Entropy: Online SOTIF Risk Quantification and Mitigation for Autonomous Driving

Liang Peng¹, Boqi Li², Wenhao Yu, Kai Yang³, Wenbo Shao⁴, and Hong Wang⁵, *Senior Member, IEEE*

Abstract—Autonomous driving confronts great challenges in complex traffic scenarios, where the SOTIF risk can be triggered by the dynamic operational environment and system insufficiencies. The SOTIF risk is reflected not only intuitively in the collision risk with objects outside the autonomous vehicles, but also inherently in the performance limitation risk of the implemented algorithms. How to minimize the SOTIF risk for autonomous driving is currently a critical, difficult, and unresolved issue. Therefore, this paper proposes the “Self-Surveillance and Self-Adaption System” as a systematic approach to online minimize the SOTIF risk, which aims to provide a systematic solution for monitoring, quantification, and mitigation of inherent and external risks. As a demonstration of the system, the risk monitoring of the perception algorithm is highlighted. Moreover, the inherent perception algorithm risk and external collision risk are jointly quantified via SOTIF entropy, which is then propagated downstream to the decision-making module and mitigated. Finally, Hardware-in-the-Loop experiments are conducted to verify the efficiency and effectiveness of the system. The results demonstrate that the system enables dependable online monitoring, quantification, and mitigation of SOTIF risk in real-time critical traffic environments.

Index Terms—Autonomous driving, SOTIF, inherent risk, artificial intelligence, perception uncertainty, entropy.

I. INTRODUCTION

ARTIFICIAL intelligence (AI) algorithms are widely adopted in autonomous driving systems to improve performance. However, AI algorithms typically provide black box solutions, and the randomness and unpredictability of those solutions will create inherent risk for autonomous vehicles (AVs). In addition, the risk of autonomous driving will also be triggered by complex external operational scenarios, such as extreme weather conditions and the stochastic behavior of road users. These types of risks fall within the scope of Safety of the

Manuscript received 18 September 2022; revised 9 April 2023 and 29 August 2023; accepted 7 September 2023. Date of publication 13 October 2023; date of current version 2 February 2024. This work was supported in part by the National Natural Science Foundation of China Project U1964203 and Project 52072215 and in part by the National Key Research and Development Program of China under Grant 2020YFB1600303. The Associate Editor for this article was B. Y. Chen. (*Corresponding author: Hong Wang.*)

Liang Peng, Wenhao Yu, Wenbo Shao, and Hong Wang are with the State Key Laboratory of Automotive Safety and Energy, School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China (e-mail: peng-l20@mails.tsinghua.edu.cn; wenhaoyu@mail.tsinghua.edu.cn; swb19@mails.tsinghua.edu.cn; hong_wang@mail.tsinghua.edu.cn).

Boqi Li is with the Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: boqili@umich.edu).

Kai Yang is with the College of Automotive Engineering, Chongqing University, Chongqing 400044, China (e-mail: kaiyang0401@gmail.com).

Digital Object Identifier 10.1109/TITS.2023.3322166

Intended Functionality (SOTIF), which refers to the absence of unreasonable risk due to a hazard caused by a performance limitation, functional insufficiency, or reasonably foreseeable misuse [1].

SOTIF is relatively a new concept, but not a new issue; it is derived from Functional Safety (FuSa) [2]. Most of the issues that FuSa can cover are hardware issues caused by electronic and electrical faults. With the application of AI algorithms in the autonomous driving system, FuSa is unable to address multiple arising safety-critical issues, which are SOTIF issues caused by algorithm limitations and functional insufficiencies. According to the reports of 3695 takeovers collected by the California Department of Motor Vehicles for the 2020 road test of BMW, Toyota, etc., 90.31% of the disengagements were due to software system SOTIF issues [3]. Among them, problems associated with perception, prediction, and planning accounted for 13.94%, 3.30%, and 35.23%, respectively.

SOTIF issues have had severe repercussions. One typical SOTIF case is that the perceptual system of one intelligent vehicle failed to distinguish a white truck from the sky on a sunny day in the year 2016, resulted in a fatal collision [4]. Similarly, this kind of intelligent vehicles hit three more white trucks in Florida, Taiwan, and Detroit due to the perceptual system’s performance limitations. One intelligent test vehicle collided with a woman crossing the road illegally in Tempe at night, which is another example of a fatal SOTIF case in the year 2018 [5]. This case was triggered by the functional insufficiency of its decision-making system, which failed to account for jaywalking pedestrians and respond to fluctuating perception results.

To sum up, the SOTIF issue necessitates two conditions: trigger conditions and system performance limitations or functional insufficiencies. Taking the perceptual subsystem as an instance, the first factor, which includes weather conditions and road conditions, is highly relevant to the operational design domain (ODD). Weather conditions, such as rainy, foggy, and snowy weather, as well as direct sunlight, will significantly degrade the performance of the perception system, leading to SOTIF issues. In addition, road conditions, such as ropes, drawbars, falling objects from surrounding vehicles (SVs), and other unanticipated occurrences, will pose significant obstacles for the autonomous driving system. The second necessary factor for SOTIF issues is related to the amount of data used to train the algorithms. Typically, the system’s performance degrades when exposed to long-tail and out-of-distribution scenarios.

In the dynamic operational environment of autonomous driving, many sources can trigger SOTIF threats. Especially, several of them can directly cause the perception system to misinterpret traffic scenarios, affecting the subsequent modules. Based on the theoretical examination of the physical and algorithmic principles of perception systems, Wu et al. formed a comprehensive list of factors and generated edge cases at the semantic level [6]. Further, Xing et al. proposed an analysis framework of perception system trigger conditions based on the chain of events model and another framework based on trigger conditions to support the safety analysis and verification of AVs [7].

In recent years, efforts to advance research related to SOTIF of autonomous driving algorithms have led to the emergence of various datasets aimed at collecting corner cases. The CODA dataset includes many unconventional traffic participants who may break into the road, such as dogs, wheelchairs, excavators, etc. [8]. The PeSOTIF dataset categorizes its collected scenarios based on perceptual triggering conditions and annotates objects with significant vulnerability [9]. Meanwhile, Hong et al. conducted experiments in controlled environments to analyze the performance limits of sensors and algorithms under different lighting and rainfall conditions [10]. Additionally, there are studies on failure detection undertaken to assess uncertainties in perception, prediction, planning modules, and information exchange among them [11], [12], [13]. These efforts have contributed to enhancing the recognition and mitigation strategies for SOTIF challenges in autonomous driving algorithms.

Regarding the whole system of AVs, extreme weather, poor lighting, road defects, temporary obstacles, unexpected target object behavior, and disturbing objects like animals and falling objects are all considered as trigger sources. As shown in Fig. 1, SOTIF risks triggered by these trigger sources can be divided into external and inherent risks. The external risk is the explicit possibility of colliding with entities on the road, such as target objects, obstacles, and road edges. The risk modeling of an external entity is often carried out based on its category, position, and velocity relative to the ego vehicle. The inherent risk is the implicit risk of algorithm performance degradation within the autonomous driving system, which is associated with the dynamic operational environment and the limitations of sensors and algorithms.

Most planning algorithms in existing autonomous driving systems are modeling the object-level results of perception and prediction algorithms as external risk without considering the inherent risk of AI algorithms failures. In practice, the entity category, relative position, relative velocity, and other information required for external risk modeling depend on upstream AI algorithms, whose unpredictability and black-box characteristics may result in the failure of external risk modeling. For instance, if the perceptual system fails to detect a target object, the vehicle will not respond. Therefore, the overall SOTIF risks, which includes both inherent and external risks, must be systematically considered. However, fundamental theories and systematic solutions are currently lacking for preventing the SOTIF risks in real-time. The detailed contributions of this paper are hence summarized as follows:

1) The Self-Surveillance and Self-Adaption System is a proposed framework for monitoring, quantification, and mitigation of both inherent and external risks of AVs.

2) This study investigates a demonstration of the proposed Self-Surveillance and Self-Adaption System for processing perception algorithm risk. The risk monitoring of the perception algorithm is highlighted in this paper. It includes monitoring the risk of YOLOv5 and quantifying the associated risk with SOTIF entropy, which can be mitigated with uncertainty-aware decision-making strategies.

3) A Perceptual Uncertainty-Aware Decision-Making (PUADM) method is proposed to mitigate the risks propagated by the perception algorithm. The effectiveness and efficiency are verified via Hardware-in-the-Loop experiments under multiple critical scenarios.

The remainder of this paper is organized as follows. The related work of quantification and mitigation of SOTIF risks for autonomous driving systems employing AI algorithms is presented in Section II. Section III outlines the Self-Surveillance and Self-Adaption System. Section IV describes a demonstration of handling the inherent perceptual risk in detail, including entropy quantification and risk mitigation modules. In Section V, comprehensive experiments are compared and analyzed under typical perceptual SOTIF-related scenarios, and in Section VI, conclusions are drawn.

II. RELATED WORK

According to the preceding analysis, it is essential to minimize the SOTIF risk systematically. In this regard, the systematic solutions for reducing the risk of FuSa are first examined. Then, the research on modeling the inherent SOTIF risk is investigated from the perspective of handling the uncertainty of AI algorithms.

A. System Solution of Safety Guidelines for AVs

Krzysztof Czarnecki et al. proposed an approach and architectural design for modifying the runtime representation of ODD based on system capability in [14]. This architecture consisted of a system layer and a supervisory layer. The system layer is responsible for executing the dynamic driving task, while the supervisory layer is responsible for monitoring the system and interacting with the system layer to respond to impairments. The proposed architecture could deal with most issues associated with running out of ODD in the functional safety concept, but not the SOTIF issues specifically.

In 2012, the project Stadpilot, which aims to achieve fully autonomous driving on the innercity ring road of Braunschweig, first proposed a systematic solution. Andreas Reschka et al. proposed a surveillance and safety system based on performance criteria and functional degradation for AVs in [15]. The surveillance system collected data from hardware sensors and software systems. Degradation actions and safety maneuvers would be executed according to the collected data, such as rain amount, temperature, sideslip angle, tire velocities, etc., to keep the vehicle safe.

However, the introduced systematic solutions were incapable of addressing SOTIF issues associated with trigger

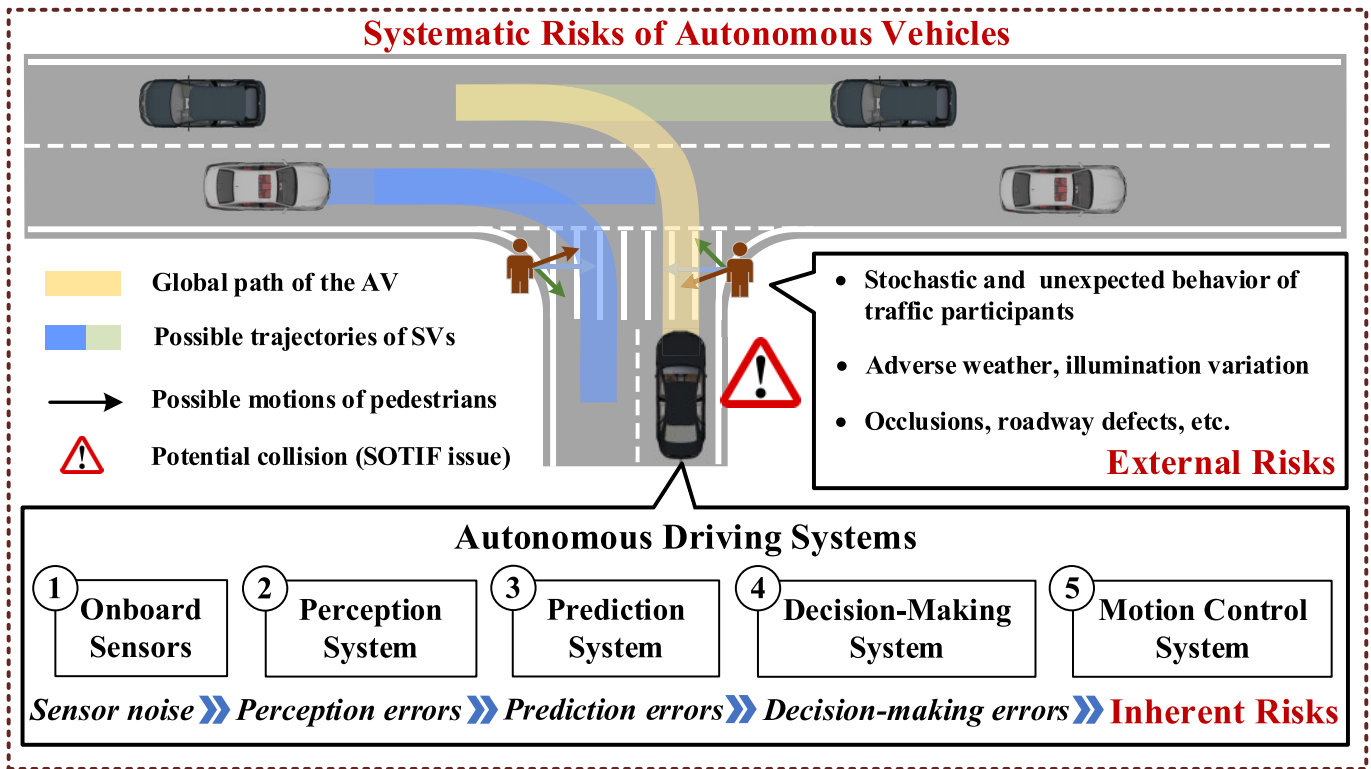


Fig. 1. Composition of the overall SOTIF risks of the autonomous vehicles, including external risks and inherent risks.

conditions, such as extreme weather, nor those associated with algorithm performance limitations and functional inefficiencies. Future systematic solutions for AVs must therefore consider monitoring the limitations of algorithms and making self-adaptive safe decisions.

B. Uncertainty Monitoring and Risk Quantification

Recent advancements have made the autonomous driving system, particularly the perception subsystem, highly dependent on artificial intelligence techniques. However, it is notoriously challenging to guarantee their safety, and they may exhibit unexpected behavior, which is unacceptable for safety-critical applications. To mitigate this issue, the systematic solution described in [14] suggested designing a supervisory layer to monitor the AI algorithms and interact in real-time with the system layer. In some subsequent studies, once a degraded performance is detected, the system layer will receive an early warning and modify its behavior [16].

As previously emphasized, the black-box nature of AI algorithms is one of the most important factors contributing to SOTIF risk. In recent years, this characteristic has been studied primarily via uncertainty estimation. The uncertainty of the algorithm can indicate the output's effectiveness and is suitable for monitoring. Perception is a safety-critical function of AVs, and AI algorithms play a crucial role in its implementation. Perceptual uncertainties usually result in SOTIF issues in the nominal performance of learning-based algorithms. Moreover, failures in perception usually result in unsafe maneuvers because the planning and control modules rely heavily on perception results.

Kendall et al. analyzed the role of uncertainty estimation in image classification and semantic segmentation, separating the output prediction uncertainty of deep neural networks into epistemic and aleatoric uncertainties [17]. The epistemic uncertainty refers to the algorithm's inherent cognitive ability, while the aleatoric uncertainty reflects the influence of input noise on the algorithm. Krzysztof et al. provided a summary of the perceptual uncertainties that may lead to perception failures during the development and operation phases of learning-based models, such as labeling uncertainty, model uncertainty, operational domain uncertainty, etc. [18]. The modeling and protection of inherent risk can be achieved by estimating and monitoring the AI algorithmic uncertainties.

The Bayesian model provides a mathematical framework for estimating uncertainties, but the increase in computation costs is a drawback [19]. Furthermore, due to the non-linearity of deep neural networks, the accurate posterior inference is difficult to achieve [20]. In terms of theoretical research, Laplace Approximation (LA), Variational Inference (VI), Markov Chain Monte Carlo (MCMC), and other similar methods can simplify the inference process of the Bayesian model. Nevertheless, the computational efficiency of these methods is still insufficient [21], [22], [23]. As research has progressed, more practical approaches have been proposed. Table I summarizes the pros and cons of various uncertainty estimation methods.

For epistemic uncertainty, Gal et al. demonstrated that utilizing the Monte Carlo Dropout (MCD) method in deep neural networks could be interpreted as a Bayesian approximation of Gaussian processes [24], [25], [26]. Peng et al. utilized

TABLE I
TYPICAL METHODS OF UNCERTAINTY ESTIMATION

Method	Basis	Uncertainty Type	Pros and Cons
VI	Bayesian theory	Epistemic uncertainty	Functional analysis [16]; precise theoretical derivation; high computational complexity.
MCMC	Bayesian theory	Epistemic uncertainty	Markov process [17]; convenient approximate inference; high computational complexity.
LA	Bayesian theory	Epistemic uncertainty	Gaussian posterior [15]; simple distribution form; high computational complexity.
MCD	Sampling-based	Epistemic uncertainty	Bernoulli weights [18]; easy to implement by dropout; need more samples than DE.
DE	Sampling-based	Epistemic uncertainty	Bootstrap combination [25]; efficient parallel inferences; only for white-box networks.
GOM	Gaussian distribution	Aleatoric uncertainty	Gaussian output [29]; efficient for single inference; need modifying layers and losses.
DOM	Dirichlet distribution	Aleatoric uncertainty	Dirichlet output [31]; feasible for black-box networks; need sampling like MCD.
DER	Evidence theory	Both	Evidential Prior [35]; efficient and scalable learning; complex evidence theory.
EP	Propagation mechanism	Both	Variance propagation [36]; limited modifications to layers; complex propagation mechanism.

the MCD method to obtain probabilistic object detectors and estimate the uncertainty for each object's bounding box (regression value) and category (classification value) [27], [28]. However, each model's weights should be sampled at the outset of the inference procedure, which would increase computational costs [29].

Deep Ensembles (DE) is another Bayesian approximation method that trains an ensemble of deterministic models with randomly initialized weights and random shuffling of data during training, as demonstrated by Osband and Van Roy [30], [31], [32]. Several studies trained ensembles of object detection models with different architectures to complement each other, reduce the number of missing objects, and enhance the mAP [33], [34]. However, the DE method for estimating the uncertainty of an object detector for safety evaluation has not been extensively studied. This paper adopts the DE technique because it is simple to be implemented, conducive to be parallel operated, and has the potential to satisfy real-time requirements. This paper adopts it in the uncertainty estimation and entropy quantification of the perception algorithm.

For aleatoric uncertainty, different distributions are usually imposed on the network output as the prior. The Gaussian Output Modeling (GOM) method typically assumes that each class score of the network output follows an independent Gaussian distribution, which does not require sampling and is widely adopted [35], [36]. In addition, Mena et al. assumed that the network output follows the Dirichlet distribution and created wrappers to estimate the uncertainty of black-box algorithms based on this assumption [37], [38]. However, the real-time performance of this Dirichlet Output Modeling (DOM) method cannot be guaranteed because it requires sampling and introduces an additional white-box network [39].

In general, different methods can be combined to quantify the overall uncertainty of a prediction. For example, Harakeh et al. used the MCD method and the GOM method, respectively, to measure epistemic uncertainty and aleatoric uncertainty [40]. In addition, some studies estimate both epistemic and aleatoric uncertainty at once. Amini et al. proposed the Deep Evidential Regression (DER) method to directly measure the prediction uncertainty, which places evidential priors over the original Gaussian likelihood function [41].

The Error Propagation (EP) method is also attracting practical applications. The batch normalization layers are viewed as noise-injection procedures, and the activation layers are transformed into layers of uncertainty propagation [42], [43]. This sampling-free method can propagate the uncertainty estimated in injection layers to the output layer via propagation layers in a single inference, which has high computational efficiency.

Given that the autonomous driving system comprises multiple modules, it is essential to quantify the risks based on the monitored uncertainties and propagate them downstream. The position and velocity of an object are continuous and convenient for the decision-making module to consider. Kahn et al. propagated the prediction results with uncertainty to the planning module, which can minimize dangerous collisions [44], [45]. However, the classification of an object is categorical, and the modeling of driving safety fields with traversable and untraversable obstacles is very different [46]. For instance, suppose a pedestrian is misclassified as a traffic cone and the perception module outputs only the category to the decision-making module, the vehicle may take dangerous actions with confidence [47], [48]. Ivanovic et al. took the class uncertainty into account. The prediction module receives all class scores instead of a single category to predict trajectories for various possible categories and produce a final prediction [49], [50]. Nonetheless, the uncertainty of the perception algorithm has not been evaluated.

III. SELF-SURVEILLANCE AND SELF-ADAPTION SYSTEM

As depicted in Fig. 2, this section proposes a systematic solution for monitoring, quantification, and mitigation of both inherent and external risks of AVs. As avoiding obstacles is the basic task of AVs, the external risk R_e can be modeled by the well-studied localization, perception, prediction, and planning algorithms. Moreover, the inherent risk R_i can be modeled by monitoring the operational status of AI algorithms through the Self-Surveillance and Self-Adaptive Safety System.

As previously emphasized, the safety of AI algorithms is particularly important for SOTIF. Therefore, each AI-powered module in the autonomous driving system must be monitored. Specifically, the inherent risk R_i mentioned in this paper includes three aspects, *i.e.*, the inherent risks of the

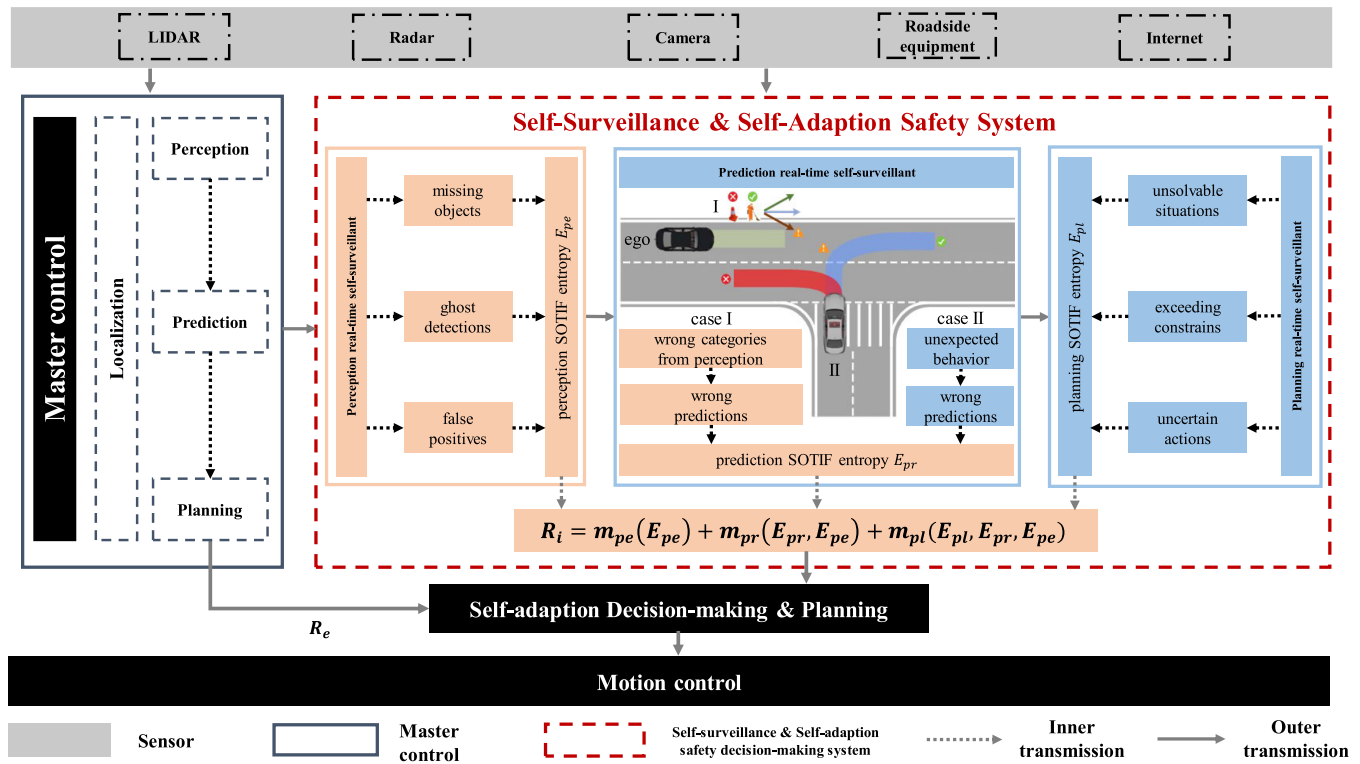


Fig. 2. The framework of the Self-Surveillance and Self-Adaption Safety System.

perception module, prediction module, and planning module, denoted as E_{pe} , E_{pr} , and E_{pl} . The inherent risk of the perception subsystem E_{pe} refers to the risk of missing objects, ghost deflections, and false positives of perception algorithms. The inherent risk of the prediction subsystem E_{pr} refers to the wrong predictions of prediction algorithms confronted with unexpected behavior or wrong perception results. The inherent risk of the planning subsystem E_{pl} refers to the unsolvable situations of rule-based decision-making algorithms, solutions exceeding constraints, and the uncertain actions of learning-based planning algorithms.

In addition to the impact caused by the inherent risk of each module itself, the risk of the module upstream will also affect the modules downstream [51], [52]. For instance, when the perception algorithm transmits incorrect object categories, the prediction algorithm will likely make incorrect predictions, thereby increasing the inherent risk. In summary, the implicit expression of the overall inherent risk is provided by the mappings related to the inherent risks of the subsystems.

Consequently, the overall risk of the AV system consists of inherent risk R_i from AI algorithms and external collision risk R_e . If the monitored and quantified overall risk falls within an acceptable range, the self-adaptive decision-making and planning algorithm will take it into account to implement motion control to mitigate the risk. However, if the overall risk is too high, it is recommended to activate the degradation mechanism or simply pull over.

Given that AI algorithms are widely adopted in the perception module, which is located upstream of the autonomous driving system so plenty of trigger conditions intuitively act on it. To demonstrate the effectiveness of the proposed frame-

work, a demonstration of perception algorithm risk processing is investigated. This paper focuses on the risk monitoring and protection measures of the perception algorithm, which is verified from the perspectives of perception performance and overall system performance.

IV. A DEMONSTRATION OF HANDLING THE INHERENT PERCEPTUAL RISK

The perceptual SOTIF risk of AVs may be triggered by environmental factors such as rain, snow, fog, and poor lighting conditions, as well as object characteristics such as unusual appearances or postures. In the aforementioned circumstances, learning-based perception algorithms often generate incorrect detection results. The causes can be summed up from two perspectives: the deterioration of sensor capabilities and the limited cognitive capacity of networks. However, current autonomous driving systems assume that the perceptual results are 100% accurate and transmit the results directly to the modules downstream, which could result in dangerous decisions and even fatal accidents [4], [5]. Therefore, it is crucial for safe driving to monitor the operational states of learning-based perception algorithms, quantify the associated risk, and carefully mitigate it in modules downstream.

This section displays a demonstration of the Self-Surveillance and Self-Adaption Safety System. In this demonstration, the perception module uses the YOLOv5 object detection algorithm, and the planning module is based on Model Predictive Control (MPC). Specifically, this paper utilizes Deep Ensembles to estimate the epistemic uncertainty of the AI perception algorithm to realize the self-monitoring of the perception module. In addition, the concept of entropy

is used to quantify the monitored perceptual risk [53], [54]. The quantified perception SOTIF entropy ultimately bypasses the prediction module. It is directly propagated to the planning module in real-time to generate safer trajectories where the overall risk is mitigated.

A. Monitoring the Perceptual Epistemic Uncertainty

Deep Ensembles is a sampling-based method for modeling epistemic uncertainty. The ensemble's sample networks are deterministic and adhere to the same architecture with slightly different weights. Moreover, it has been demonstrated that an ensemble of five networks is sufficient for operation, allowing different GPUs or threads to be used conveniently for parallel inference [31]. Therefore, this method only introduces the time cost of post-processing, which offers a potential practical implementation opportunity in the autonomous vehicle.

Applying the Deep Ensembles method, T object detectors are trained with the same training process and dataset but different random number seeds to generate an ensemble $E = \{O_1, O_2, \dots, O_T\}$, as shown in Fig. 3. For each input frame image I , a set of predictions $P = \{P_1, P_2, \dots, P_T\}$ can be obtained from E , where each P_i is the result after independently running the Non-Maximum Suppression (NMS) [30]. As shown in Algorithm 1, a set of clusters $L = \{L_1, L_2, \dots, L_N\}$ can be obtained from P through the Basic Sequential Algorithm Scheme with intra-sample exclusivity (BSAS excl.) based on the Intersection over Union (IoU) and the winning label (WL), in which each L_i corresponds to a detected object and contains at most T sampling tensors [34], [55]. The mean of all sample tensors for each cluster L_i is the final perception result. Then, the variances of coordinates of the bounding boxes are used to approximate the spatial uncertainty.

Algorithm 1 Basic Sequential Algorithm Scheme with Intra-Sample Exclusivity

Input: $Affinity = \{IoU \& WL\}$, threshold θ_{aff} , a set of predictions $P = \{P_1, P_2, \dots, P_T\}$;

Output: A set of clusters $L = \{L_1, L_2, \dots, L_N\}$;

- 1: Create a cluster for each object box in P_1 ;
 - 2: **for** $i \in \{2, 3, \dots, T\}$
 - 3: set $excl_flag = \mathbf{0}_n$, n is the current number of clusters
 - 4: **for** each object box B_j in P_i
 - 5: **for** $k \in \{1, 2, \dots, n\}$
 - 6: **if** $Affinity(B_j, L_k) \geq \theta_{aff}$ **and** $excl_flag(k) = 0$ **then**
 - 7: put B_j into L_k , and set $excl_flag(k) = 1$
 - 8: **if** B_j has not been processed yet **then**
 - 9: $n = n + 1$, create a new cluster L_n for B_j
 - 10: **return** L ;
-

B. Quantifying the Perception SOTIF Entropy

In probabilistic object detection, the sample mean and variance are frequently employed to approximate the corre-

sponding posteriori distribution statistics. Meanwhile, Shannon entropy is also a popular metric for estimating the quality of predictive uncertainty in classification tasks [56]. For a category label y of C categories, the Shannon entropy H is measured by:

$$H = - \sum_{c=1}^C p(y = c|x, D) \log(p(y = c|x, D)) \quad (1)$$

where x represents the input data and D is the training dataset. H reaches a minimum value $H_{\min} = 0$ when the network is completely certain in its prediction, *i.e.*, $p(y = c|x, D) = 0$ or 1 . When the prediction of the network follows a uniform distribution, *i.e.*, $p(y = c|x, D) = \frac{1}{C}$, H reaches a maximum value, *i.e.*, $H_{\max} = \log(C)$.

The above metric is applicable in single-label object detection, where the probability vector output is subject to category distribution, or $\sum_{c=1}^C p(y = c|x, D) = 1$. While YOLOv5 supports multi-label object detection, which has higher fault tolerance. It uses C independent logistic classifiers to predict the probabilities that the result belongs to a certain category. In summary, this paper quantifies the perception SOTIF entropy E_{pe} of the YOLOv5s network as:

$$p_c = p(y = c|x, D) \approx \frac{1}{T} \sum_{t=1}^T p(y = c|x, W_t) \quad (2)$$

$$E_{pe} = - \sum_{p_c}^C (p_c \log p_c + (1 - p_c) \log(1 - p_c)) \quad (3)$$

where T represents the number of networks in the ensemble, W_t denotes the corresponding weights of the t -th network, and C is the number of categories. For the probability p_c corresponding to the c -th category, $[p_c, 1 - p_c]$ is regarded as the probability vector output from the binary classification problem, and its Shannon entropy H_c takes $p_c = \frac{1}{2}$ as the axis of symmetry and is on a convex shape. Hence E_{pe} reaches a minimum value $E_{pe_min} = 0$ when the ensemble is completely certain in its all predictions, *i.e.*, $p_c = 0$ or 1 for each c . When the predictions of the ensemble are all ambiguous, *i.e.*, $p_c = \frac{1}{2}$ for each c , E_{pe} reaches a maximum value, *i.e.*, $E_{pe_max} = C \log 2$. Combined with (2), it can be concluded that E_{pe} tends to be large when there are errors or disagreements among the networks, no matter whether they are caused by environmental factors or object attributes. Therefore, E_{pe} can be used for self-monitoring and generating warning signals. The output of the perception algorithm modified in this paper includes three types of information, namely, spatial information with uncertainty, semantic information with uncertainty, and comprehensive uncertainty information. The spatial information reflects the uncertainty of the bounding box positioning. As shown in Fig. 3, two dashed boxes are added around each bounding box whose corner coordinates are three times the standard deviation apart. The semantic information reflects the uncertainty of the final perception results. The category with the highest probability within the class probability vector is referred to as the winning label, and the probability associated with it is recorded as the confidence score. The comprehensive uncertainty information reflects the

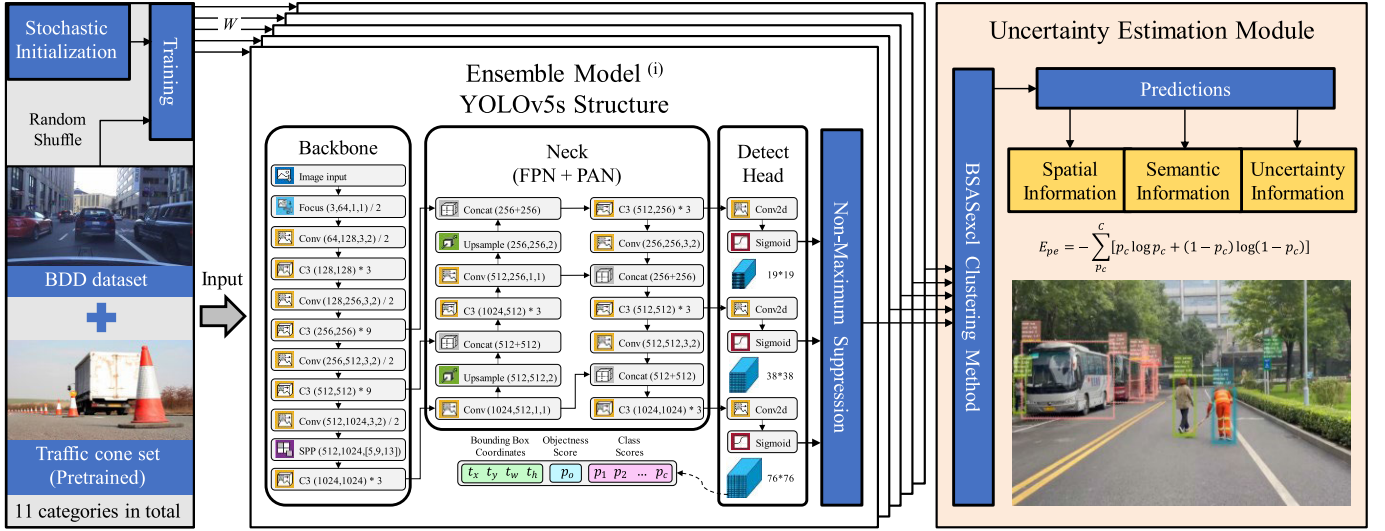


Fig. 3. Modeling the epistemic uncertainty of the AI perception algorithm.

macroscopic uncertainty results of the cluster corresponding to the object, including the number of object detectors that have detected the object, the prediction entropy of the final perception result, and the uncertainty level as determined by the prediction entropy. When only a portion of the ensemble detects an object, it may be a ghost detection or a missing object for a particular network. Consequently, an additional penalty is imposed on this object's prediction entropy:

$$E_{pe}^* = E_{pe} \times (1 + f_p \times (T - d)) \quad (4)$$

where E_{pe}^* refers to the final perception SOTIF entropy, f_p represents the penalty factor, and d refers to the number of networks that detects the object in the ensemble. Then, to output the monitoring results, three entropy levels are defined as:

$$l_u = \begin{cases} 0 & \text{if } E_{pe}^* < \theta_{lm} \quad \text{low, normal} \\ 1 & \text{if } \theta_{lm} \leq E_{pe}^* < \theta_{mh} \quad \text{medium, caution} \\ 2 & \text{if } \theta_{mh} \leq E_{pe}^* \quad \text{high, warning} \end{cases} \quad (5)$$

where l_u represents the uncertainty level, θ_{lm} refers to the uncertainty threshold between low and medium uncertainty, and θ_{mh} denotes the uncertainty threshold between medium and high uncertainty. If an object has a high level of uncertainty, a warning will be generated to alert the driver or subsequent planning module to pay attention to it.

C. Modeling and Mitigating the Overall Risk

As previously emphasized, it is essential to account for perceptual uncertainty in the downstream module, *i.e.*, the decision-making system. Consequently, a Perceptual Uncertainty-Aware Decision-Making (PUADM) method is proposed to handle the risks propagated by the perception system.

1) *Potential Field Incorporating Class Uncertainty*: To enable safe autonomous driving, AVs ought to avoid collision with surrounding road users such as vehicles and pedestrians. As aforementioned, the focus of this paper is to include the

extracted perception uncertainty into the decision-making process. Here, the potential field (PF) will be utilized to quantify collision risk between the AV and road users while considering the class uncertainty calculated in (3)-(5) [52]. Specifically, for road user i (category: c , characteristic length: $L_{x_{i,c}}$, characteristic width: $L_{y_{i,c}}$, and coordinates: $P_{x_{i,c}}, P_{y_{i,c}}, \theta_i$), the PF incorporating perception uncertainty is formulated as follows.

$$PF_{U_{i,c}}(X, Y) = a_c e^{\beta_{i,c}} \quad (6)$$

with

$$\beta_{i,c} = - \left(\frac{((X - P_{x_{i,c}})\cos\theta_i + (Y - P_{y_{i,c}})\sin\theta_i)^2}{2\lambda_{x_{i,c}}^2} + \frac{(-(X - P_{x_{i,c}})\sin\theta_i + (Y - P_{y_{i,c}})\cos\theta_i)^2}{2\lambda_{y_{i,c}}^2} \right) b_c \quad (7)$$

$$\lambda_{x_{i,c}} = L_{x_{i,c}} + E_{x_{i,c}} \quad (8)$$

$$\lambda_{y_{i,c}} = L_{y_{i,c}} + E_{y_{i,c}} \quad (9)$$

where X and Y represent the coordinates of ego vehicle, a_c and b_c are predefined intensity and shape factors of potential field subject to the object category c , respectively, $E_{x_{i,c}}$ and $E_{y_{i,c}}$ are the items that incorporate perception SOTIF entropy.

To prevent hazards caused by unexpected behavior of the high uncertainty road users, the $E_{x_{i,c}}$ and $E_{y_{i,c}}$ are calculated as:

$$E_{x_{i,c}} = \begin{cases} 0 & \text{if } l_u = 0 \\ L_{x_{i,person}} - L_{x_{i,c}} & \text{if } l_u = 1 \\ L_{x_{i,person}} - L_{x_{i,c}} + U_x & \text{if } l_u = 2 \end{cases} \quad (10)$$

$$E_{y_{i,c}} = \begin{cases} 0 & \text{if } l_u = 0 \\ L_{y_{i,person}} - L_{y_{i,c}} & \text{if } l_u = 1 \\ L_{y_{i,person}} - L_{y_{i,c}} + U_y & \text{if } l_u = 2 \end{cases} \quad (11)$$

where U_x and U_y are the safety margins set to prevent hazards caused by unexpected behavior of the high uncertainty road users.

From (6), it can be found that the closer the distance between the road user i and the AV, the higher the collision

risk PF is. Meanwhile, the perception uncertainty is also considered, *i.e.*, the greater the uncertainty, the higher the collision PF , which will force the AV to be more cautious when dealing with road users with high uncertainty. Fig.4 (middle part) demonstrates an example of the presented PF . It depicts the PF of a pedestrian and traffic cone, which reveal that if the object's class given by the YOLO v5 is certain, the PF is low. Conversely, the PF is enlarged when the object's class given by YOLO v5 is uncertain.

To sum up, when PUADM is establishing the potential field of the objects, if the uncertainty is low, it would directly establish the potential field according to the category; if the uncertainty is medium, the potential field would be established using parameters of persons to maintain safety; if the uncertainty is high, the potential field would be further expanded to keep the ego vehicle far away from the object. So far, the external collision risk based on location, speed, category, etc. and the inherent algorithm risk based on the estimated epistemic uncertainty of the perception algorithm have both been modeled, and the potential field of all objects in the scenario has been established.

On the other hand, AVs should also adhere to the road markings and not cross the road's boundary. As a result, until the lane change or overtaking instruction is sent, the AV will continue to drive in the current lane. Using a quadratic function, the potential field of road boundaries is designed to prevent unanticipated road crossings, embodied in the convex part of the two sides in the potential field in the middle bottom part of Fig. 4. The road potential field function is as follows:

$$PF_{R(X,Y)} = \begin{cases} a_q(S_{Rq}(X,Y) - D_a)^2 & S_{Rq}(X,Y) \leq D_a \\ 0 & S_{Rq}(X,Y) > D_a \end{cases} \quad (12)$$

where a_q represents the parameter of road boundaries PF_R , and S_{Rq} and D_a indicate the distance and safety threshold between the ego vehicle and road boundaries, respectively. The PF_R will influence the ego vehicle to maintain its position in the center of the lane.

Therefore, the overall potential field can be represented as follows:

$$PF = PF_U + PF_R \quad (13)$$

Fig. 4 demonstrates an example of the overall potential field that incorporates both PF_U and PF_R . The perception uncertainty is highlighted, *i.e.*, the tangerine part in the figure. In summary, based on the perceptual entropy and potential field, the perceptual uncertainty is quantified, which can be used in the decision-making process.

2) b. Decision-Making Paradigm Based on MPC:

First, a 3-Degree-of-Freedom (DOF) prediction model with multi-constraints is constructed to characterize the motion dynamics of the autonomous vehicle, as shown in [57] and [58].

$$m(\dot{u} - vr) = F_{xT} \quad (14)$$

$$m(\dot{v} + ur) = -C_{\alpha f}(\delta_f - \frac{v + l_f r}{u}) - C_{\alpha r}(-\frac{v - l_r r}{u}) \quad (15)$$

$$I_z \dot{r} = F_{yf} l_f - F_{yr} l_r \quad (16)$$

$$\dot{X} = u \cos \phi - v \sin \phi \quad (17)$$

$$\dot{Y} = u \sin \phi + v \cos \phi \quad (18)$$

where m refers to the vehicle mass, I_z refers to the vehicle's moment of inertia, l_f and l_r represent the distance from the vehicle CG (center of gravity) to the front and rear axles, respectively, and u and v denote the vehicle's longitudinal and lateral velocities, respectively. The vehicle yaw rate at CG is given by r , ϕ denotes the vehicle heading angle, X and Y are the vehicle longitudinal and lateral positions with respect to the global coordinate, respectively, $C_{\alpha f}$ and $C_{\alpha r}$ are the cornering stiffness of the front and rear tires, respectively, F_{xT} is the total longitudinal force of the tires, and δ_f is the front steering angle.

Subsequently, the model demonstrated in (15) can be denoted as the state space f form with linearization and discretization techniques, *i.e.*:

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{A}_d \mathbf{x}(k) + \mathbf{B}_d \mathbf{u}(k) \\ \mathbf{y}(k) &= \mathbf{C}_d \mathbf{x}(k) \end{aligned} \quad (19)$$

where $\mathbf{x} = [u, v, r, \phi, X, Y]^T$, $\mathbf{u} = [F_{xT}, \delta_f]^T$, $\mathbf{y} = [Y, u]^T$, \mathbf{A}_d , \mathbf{B}_d and \mathbf{C}_d denote the discrete matrices corresponding to (15), and k represents the time. The details of linearization and discretization techniques and calculation process of \mathbf{A}_d , \mathbf{B}_d , and \mathbf{C}_d can be found in [52].

$$\begin{aligned} \min_{\mathbf{u}} \sum_{k=1}^{N_p} (\|\Delta \mathbf{y}_k\|_Q^2 + \mathbf{P} \mathbf{F}_k) + \sum_{k=1}^{N_c} \|\mathbf{u}_k\|_R^2 + \|\Delta \mathbf{u}_k\|_S^2 \\ \text{s.t. } (k = 1, \dots, N_p) \\ \mathbf{x}(t+k) = \mathbf{A}_d \mathbf{x}(t+k-1) + \mathbf{B}_d \mathbf{u}(t+k-1) \\ \mathbf{y}(t+k) = \mathbf{C}_d \mathbf{x}(t+k) \\ \mathbf{u}_{\min}(t+k-1) \leq \mathbf{u}(t+k-1) \leq \mathbf{u}_{\max}(t+k-1) \\ \Delta \mathbf{u}_{\min}(t+k-1) \leq \Delta \mathbf{u}(t+k-1) \leq \Delta \mathbf{u}_{\max}(t+k-1) \\ \mathbf{u}(t+k) = \mathbf{u}(t+k-1), \quad \forall k \geq N_c \end{aligned} \quad (20)$$

Then, based on (15) and (19), and the constructed perceptual uncertainty-aware potential field, the decision-making process can be converted into solving the above optimal control issue (20), where $\Delta \mathbf{y}_k$ represents the relative lateral position and longitudinal velocity between ego vehicle and the object, N_p and N_c are the prediction and control horizons, respectively, $\mathbf{x}(t+k)$ represents the predicted state values of the system, $\mathbf{y}(t+k)$ denotes the predicted outputs of the system over the prediction horizon, \mathbf{u}_{\min} and \mathbf{u}_{\max} are the lower and upper bounds of the actuator, respectively, and $\Delta \mathbf{u}_{\min}$ and $\Delta \mathbf{u}_{\max}$ are the various ranges of control variables at each time, which can be found in [52]. In the cost function shown in (20), matrices \mathbf{Q} , \mathbf{R} , and \mathbf{S} are the weight matrices.

V. EXPERIMENTS AND RESULTS

For the sake of generality, this paper comes up with the study from one specific long-tail scenario related to the perceptual SOTIF problem. In this scenario, an AV travels on a straight road, while a sanitation worker sweeps on the right side of the lane with his back to the AV. Due to the

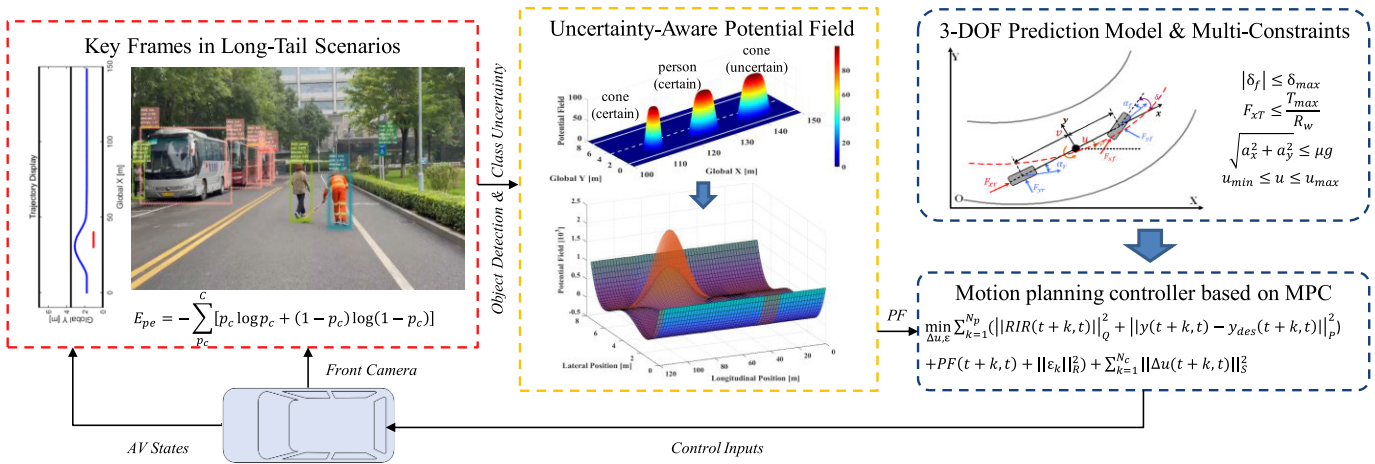


Fig. 4. Perceptual uncertainty-aware safe decision-making algorithm.

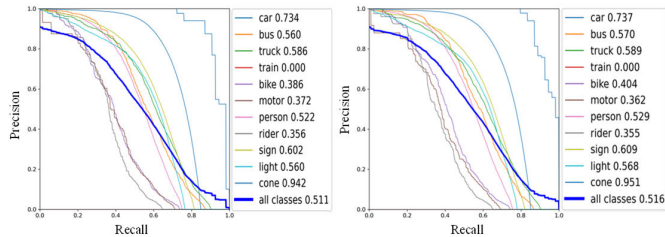


Fig. 5. The P-R curves of the worst and the best trained YOLOv5s networks on the BDD validation set.

similarity between the stripes and colors of sanitation workers' uniforms and traffic cones, the perception algorithm may be misled by certain trigger conditions. Once the AV erroneously identifies a sanitation worker as a traffic cone, which is considered a stationary obstacle, a collision is possible because sanitation workers may move laterally while working on the road. Therefore, it is necessary to monitor the uncertainty of perception algorithms in this scenario and account for it in algorithms that make safer decisions.

A. Uncertainty Monitoring and Risk Quantification

As the perception algorithm of the AV, a monocular object detection algorithm is utilized in this paper. To study this sanitation-worker-traffic-cone scenario, the object detection network must be capable of recognizing at least persons and traffic cones. Consequently, the YOLOv5s object detection network is trained with the BDD dataset (70k images) and a traffic cone dataset (0.27k images) to distinguish 11 categories of objects: car, bus, truck, train, bike, motor, person, rider, traffic sign, traffic light, and traffic cone [59], [60], [61]. As the BDD dataset is much significantly larger than the traffic cone dataset, the network is trained for 60 epochs with the BDD dataset after 300 epochs of pre-training with the traffic cone dataset.

In order to investigate the impact of the number of networks in the ensemble on redundancy performance enhancement and the effectiveness of uncertainty estimation, we iteratively executed the aforementioned procedure, training a total of 10 model weights with identical architecture but varying

TABLE II
THE TESTING OF ENSEMBLE PERFORMANCE IMPROVEMENT

Number of networks	Precision	Recall	mAP ₅₀
1	0.713	0.472	0.516
2	0.720	0.478	0.523
3	0.756	0.466	0.526
4	0.753	0.465	0.526
5	0.716	0.482	0.527
6	0.722	0.480	0.528
7	0.766	0.46	0.528
8	0.779	0.455	0.528
9	0.736	0.476	0.529
10	0.744	0.472	0.529

parameters. As shown in Fig. 5, the lowest mAP@0.5 of a single YOLOv5s network is 0.511 on the BDD validation set, which is comparable to the performance of ResNet50 (0.499) [16]. Subsequently, we extracted 1 to 10 sets of weights from the trained networks to form ensembles and conducted individual tests on the BDD validation dataset. As indicated in Table II, it can be observed that with the increase in the number of networks within the ensemble, the performance demonstrates enhancement due to redundancy. However, beyond a quantity of 5, the performance improvement becomes marginal. Consequently, considering the real-time and effectiveness requirements of the deployed algorithm, the Deep Ensembles approach is adopted based on the parameters listed in Table IV.

The sanitation-worker-traffic-cone scenario is established in the real world, and seven videos are filmed for testing. In a local video with 538 captured frames, the PyTorch-based algorithm infers at 20 fps, and the warning accuracy reaches 97.71%. Fig. 3 depicts the framework for employing the Deep Ensembles method to estimate the epistemic uncertainty of the YOLOv5s network and the detection results of one frame

TABLE III
 THE TESTING OF ENSEMBLE UNCERTAINTY QUALITY

Number of networks	mAP ₅₀	ACR	FAR	CQS	UQS
1	0.159	0.6370	0.1545	0.7250	3.7102
2	0.171	0.8025	0.1234	0.8061	4.2474
3	0.193	0.8623	0.1156	0.8375	4.3117
4	0.209	0.8894	0.1086	0.8539	4.3784
5	0.215	0.9018	0.1026	0.8631	4.4451
6	0.219	0.9131	0.1011	0.8689	4.3586
7	0.222	0.9195	0.0998	0.8726	4.3187
8	0.223	0.9241	0.0978	0.8758	4.3046
9	0.224	0.9290	0.0970	0.8789	4.2984
10	0.225	0.9298	0.0957	0.8799	4.2940

 TABLE IV
 THE PREDEFINED PARAMETERS IN THE DEMONSTRATION

Parameter	Meaning	Value
T	The total number of sampling networks in the ensemble.	5
C	The total number of categories.	11
θ_{aff}	The spatial affinity threshold in the BSAS excl. algorithm.	0.95
f_p	The additional penalty factor.	0.1
θ_{lm}	The threshold of entropy between low and medium levels.	1.2
θ_{mh}	The threshold of entropy between medium and high levels.	1.6
$L_{x_i, person}$	The characteristic length of the PF of a person.	15 m
$L_{y_i, person}$	The characteristic width of the PF of a person.	1.5 m
$L_{x_i, traffic cone}$	The characteristic length of the PF of a traffic cone.	8 m
$L_{y_i, traffic cone}$	The characteristic width of the PF of a traffic cone.	0.5 m
U_x	The length of the safety margin.	8 m
U_y	The width of the safety margin.	1 m
C_{af}	The cornering stiffness of the front tire.	90000 N/rad
C_{ar}	The cornering stiffness of the rear tire.	90000 N/rad
l_f	The distance between the mass point and the front axle.	1.18 m
l_r	The distance between the mass point and the rear axle.	1.77 m
m	The total mass of the vehicle.	1860 kg
I_z	The moment mass of the vehicle for Z axle.	3438.5 kg·m ²
v_e	The expected speed of the ego vehicle.	15 m/s
v_p	The forward speed of the sanitation worker.	1 m/s
w_l	The width of the road lane.	3.5 m
x_0	The original distance between the person and ego vehicle.	30 m
y_0	The original distance between the person and the lane.	1 m
Δt	The time step of the simulation in the experiment.	0.033 s

in the video, while Fig. 6 depicts the fluctuating prediction entropy of the sanitation worker in the video. When the sanitation worker faces the vehicle upright or sideways, the

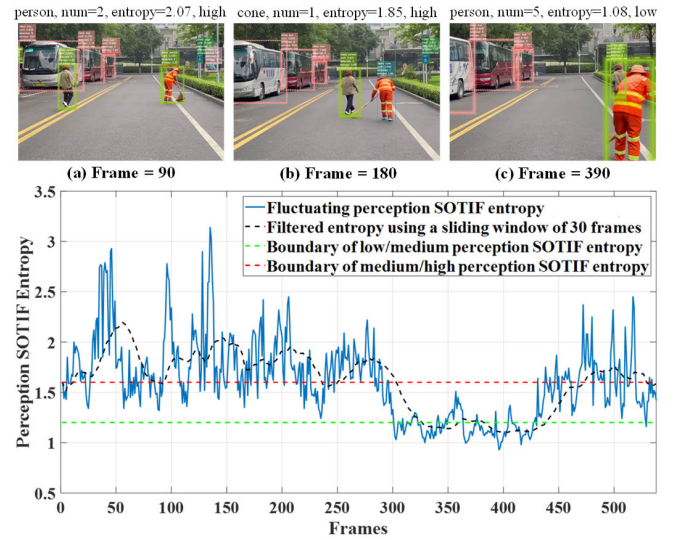


Fig. 6. Fluctuating perception SOTIF entropy in the test video.

perception algorithm can correctly identify the person with low uncertainty. However, when the sanitation worker bends or squats with the back towards the vehicle, there is a high likelihood that they will be mistaken for a traffic cone with high uncertainty. Especially when the sanitation worker is seated on the ground, even human drivers have difficulty distinguishing him.

After verifying continuous frames under the sanitation-worker-traffic-cone scenario, the effectiveness and scalability of uncertainty monitoring and risk quantification results are further verified through discrete keyframes of diverse scenarios. PeSOTIF is a labeled test dataset specifically established for vision-based probabilistic object detectors [9]. It has collected approximately 4000 objects in over 1000 keyframes of perceptual SOTIF scenarios from a variety of sources.

In addition, an evaluation protocol based on three dimensions of difficulty, precision, and uncertainty is also proposed by PeSOTIF, including metrics of alert coverage rate (ACR), false alert rate (FAR), classification quality score (CQS), and uncertainty quality score (UQS). Among them, ACR is the proportion of detected high uncertain objects in the total key objects, and FAR is the proportion of detected high uncertain objects that are actually not critical. We also extracted 1 to 10 sets of weights from the trained networks to form ensembles and conducted individual tests on the PeSOTIF dataset. As depicted in Table III, the improvement in uncertainty estimation effectiveness becomes gradual when the number of networks in the ensemble exceeds 5, aligning with the previously drawn conclusions.

The experimental results on the PeSOTIF dataset in accordance with its evaluation protocol demonstrate that the accuracy of identifying perceptual risk (ACR) reaches 90.2%, while the false alarm rate for monitoring high uncertainty objects (FAR) is 10.3%. In other words, 90.2% of the key objects affected by environment, appearance or posture in these scenarios can be successfully detected as high uncertainty objects. At the same time, only 10.3% of the objects

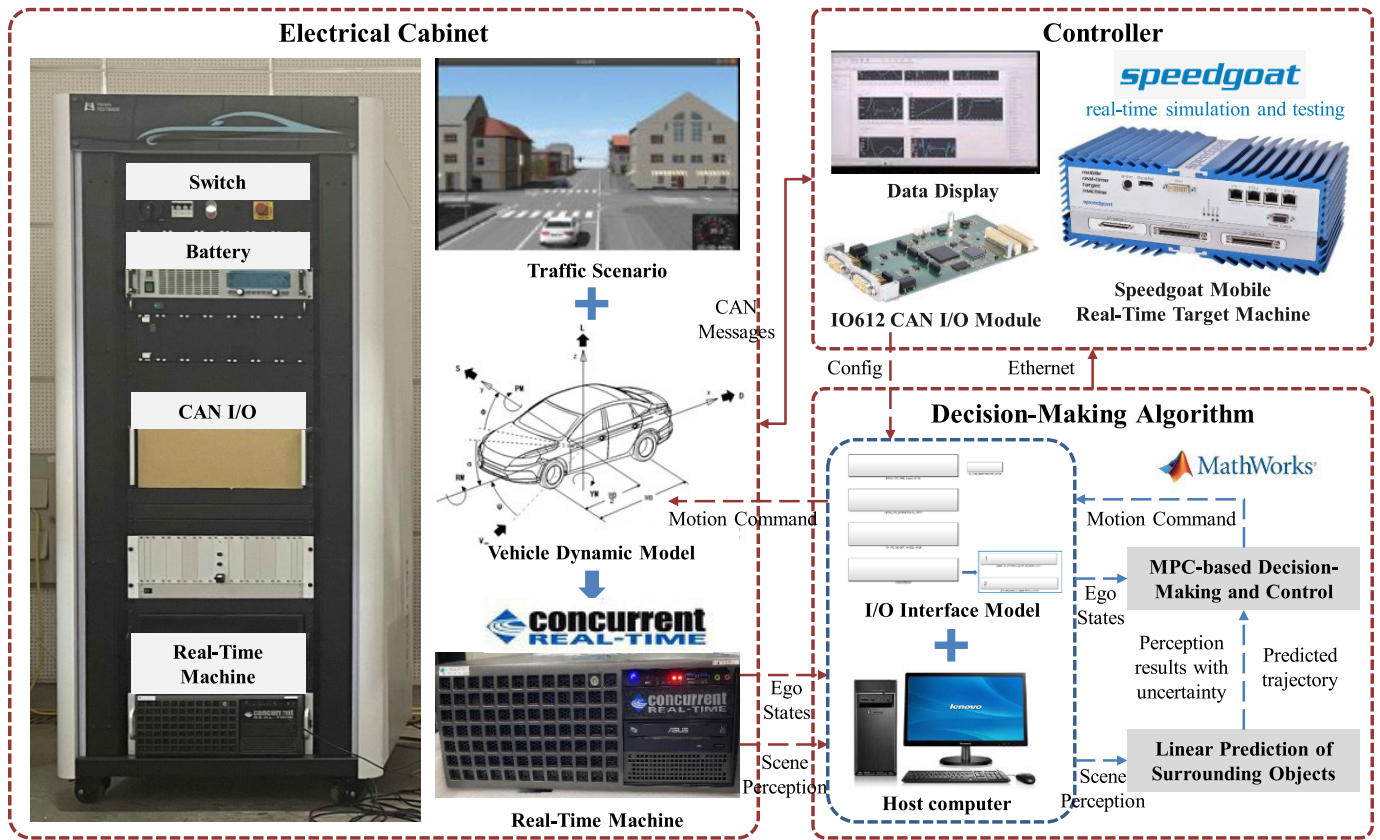


Fig. 7. Composition and architecture of the Hardware-in-the-Loop platform.

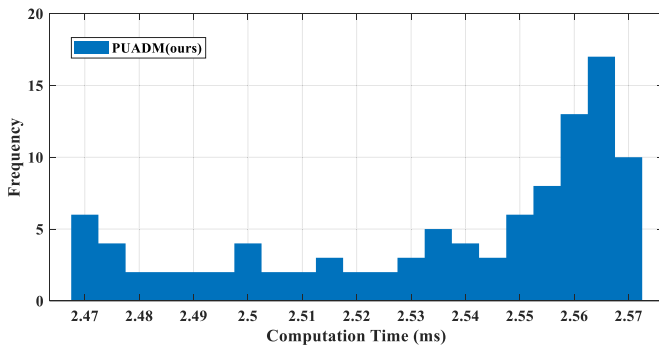


Fig. 8. Real-time performance of the proposed PUADM method in case 2.

considered by the algorithm to be highly uncertain are wrong, including most of the missing objects and ghost detections. Therefore, the perception module can effectively monitor itself in most cases without being too conservative.

In addition, the modified perception module is deployed to the AV system of the Apollo D-KIT, which utilizes an NVIDIA Geforce RTX2070S GPU, once it has been verified with a variety of perceptual SOTIF scenarios. Firstly, the Python implementation of the YOLOv5 algorithm based on PyTorch is converted to the C++ implementation based on TensorRT [62]. Secondly, the relevant necessary algorithms such as Deep Ensembles and BSAS excl. are rewritten into the C++ version. Finally, the multithreaded calling method is developed to parallelize the ensemble inference of the five networks.

Therefore, the algorithm that detects local videos can operate at 100 fps. In addition, the total speed can be stabilized at approximately 30 fps when acquiring and processing real-time images from the camera through CyberRT.

B. The Performance of Risk Mitigation Through PUADM

In the sanitation-worker-traffic-cone scenario, when the uniformed tester is normally sweeping along the road, the algorithm can correctly recognize the object as a person most of the time. However, the occasional error occurs when the tester bends over. When the tester squats or sits on the road, the algorithm frequently misidentifies him as a traffic cone, despite the fact that this situation rarely occurs in reality. As depicted in Fig. 6, the relevant data of the tester is extracted from many objects in the videos, and object tracking results are generated using a simple filtering method before being transmitted to the MPC-based decision-making algorithm downstream.

MATLAB/Simulink is utilized to implement the PUADM method introduced in Section IV-C. It establishes artificial potential fields for the perceived objects and then solves the optimal trajectory under the parameters in Table IV. This paper compared two models to demonstrate the significance of considering perceptual uncertainty in subsequent tasks. The baseline model, *i.e.*, MPC-YOLO, adheres to the logic of most common modular systems, which only receive the output categories from the perceptual module, and establishes different potential field sizes for each category. The potential

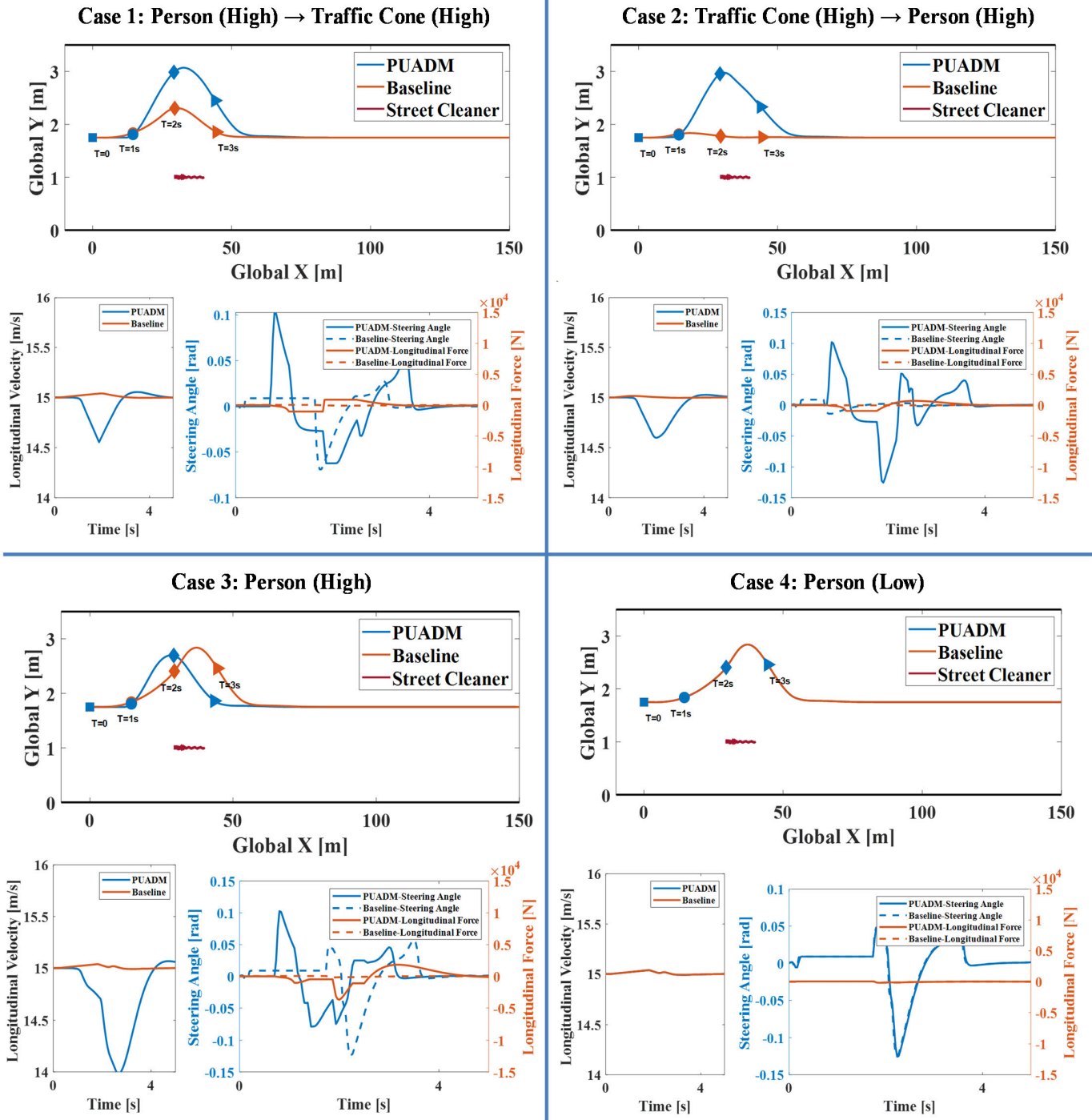


Fig. 9. Planning results in the test cases, including trajectory, longitudinal velocity, longitudinal force, and steering angle.

field of pedestrians, for instance, is the largest, followed by that of motorbikes and vehicles, and finally that of traffic cones and traffic signs, which is the smallest. PUADM receives not only the categories but also the prediction entropy as a measure of epistemic uncertainty. The artificial potential field is optimized based on uncertainty and according to the following rules: (1) low uncertainty: same as MPC-YOLO, the potential field is established based on the category; (2) Medium uncertainty: regardless of category, the potential field shall be established in accordance with the pedestrian standard; (3) High uncertainty:

further expand the maximum potential field in MPC-YOLO because it is difficult to predict the behavior of unknown objects.

The Hardware-in-the-Loop (HIL) experiment is conducted to verify the real-time performance of the decision-making algorithm synchronously. The hardware deployment is depicted in Fig. 7. In MATLAB/Simulink on the host computer, a scenario is established in which the ego vehicle advances at a constant speed of 15 m/s, and a sanitation worker sweeps near the lane centerline in front. The design of such a



Fig. 10. Perception results with entropy in key frames of some typical SOTIF scenarios in the PeSOTIF dataset. Objects with low, medium, and high entropy are labeled white, yellow, and red on the last line of the box information, respectively.

foundational scenario is depicted by the real-world image and trajectory curve on the left of Fig. 4. The perception results of the seven previously recorded videos are divided into four

cases and fed into the decision-making algorithm as scenario perception information from the concurrent real-time machine. The trajectory is then generated by the MPC-based PUADM

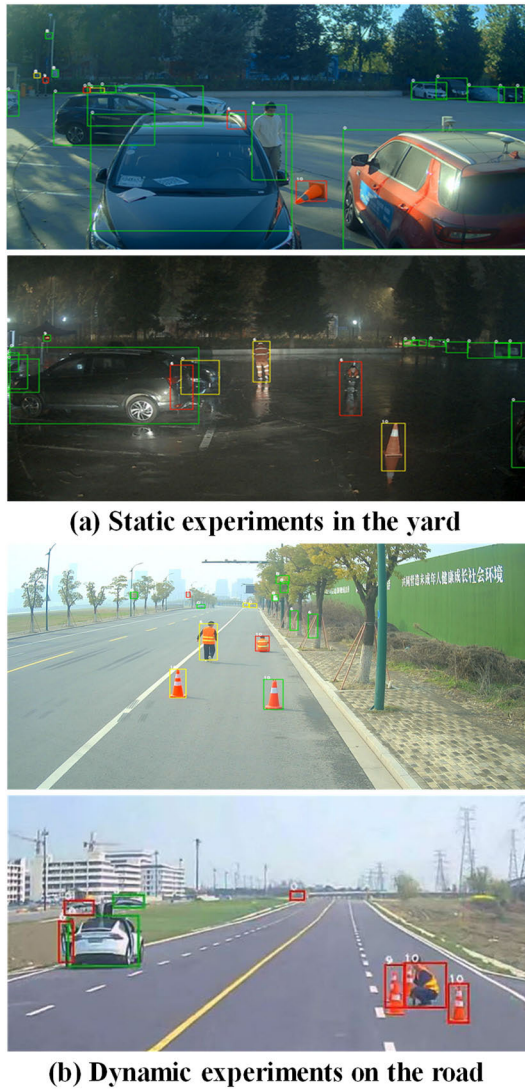


Fig. 11. Samples of perception results from filed experiments.

method according to Fig.4. Afterward, the motion command is transmitted via the I/O interface to the vehicle dynamic model on the real-time machine. The real-time machine then moves in circles after transmitting the ego vehicle states of the subsequent time step to the host computer. Meanwhile, the electrical cabinet and the controller could transmit data to each other by CAN. The variables in the MATLAB/Simulink model such as ego vehicle states can be observed through the software TCS on the host computer.

In the HIL experiments, the vehicle dynamic MATLAB/Simulink model initially causes the ego vehicle to travel straight and at a constant speed along the road. After the ego vehicle stabilizes, another MATLAB/Simulink model constructing the scenario is initiated so that the ego vehicle can conduct trajectory planning based on perception results and uncertainty. The fixed time step is set to 0.033s, and Fig. 8 demonstrates that the real-time performance satisfies the specifications.

The video sequence depicted in Fig. 6 is an instance of case 1. Due to the sanitation worker wearing a uniform that

shares similarities with the characteristics of a traffic cone, and given their frequent bending over to clean the road, the target's classification of the perception algorithm exhibits continuous shifts between person and traffic cone (frames 90 and 180). This phenomenon is particularly evident when the sanitation worker bends over to certain angles or sits on the ground, which usually leads to false detections. In such ambiguous situations, the self-surveillance of perception algorithms produces outputs with high uncertainty and transmits them to the decision-making module. It is only when the target is in close enough proximity to the ego vehicle that the perception algorithm yields a correct detection result with low uncertainty and high confidence (frame 390).

In case 1, the target's classification changed from person to traffic cone, while the entropy remained high throughout the process. For the baseline model, a proper evasive maneuver was made in the early stage. However, as soon as false detection occurred, the ego vehicle returned to the centerline, which is potentially dangerous. In case 2, the target's classification changed from traffic cone to person, and the filtered entropy remained high throughout the entire process. Due to the false detection in the baseline model, the ego vehicle did not evade at all and drove directly past the sanitation worker near the centerline of the road lane, which was extremely dangerous. However, the ego vehicle utilizing the PUADM model finished the complete evasive maneuver due to the high uncertainty of the target in the two cases.

In cases 3 and 4, the target was classified as a person with high and low entropy, respectively. The ego vehicle utilizing the baseline model completed a complete evasive maneuver after detecting a sanitation worker near the road's centerline. In case 3 of the PUADM model, where the target was determined to be an object with high entropy and whose behavior was difficult to predict, the ego vehicle executed a larger and earlier evasive maneuver. In case 4, where classification results were normal, the PUADM model performed similarly to the baseline model.

The simulation results depicted in Fig. 9 indicate that the PUADM method is safer and more aligned with the expectations of human drivers. Generally, when the uncertainty is low, the vehicle's behavior is consistent with the standard decision-making method MPC-YOLO, whereas when uncertainty is high, the vehicle will take safer actions. Specifically, the ego vehicle employing MPC-YOLO could not evade at all in case 2 due to the initial false detection. It drove directly past the pedestrian near the lane's centerline, which is extremely dangerous. Despite false detection, the ego vehicle utilizing PUADM completed the entire evasive maneuver due to the high level of uncertainty. Whether or not this method will result in overly cautious trajectory planning is dependent on whether or not the perception algorithm frequently produces high-level uncertain results. In addition, Moreover, it is related to the configuration of uncertainty thresholds and the performance of the perceptual module.

VI. CONCLUSION

This paper presented a systematic method for describing the overall SOTIF risk of AVs. A Self-Surveillance and

Self-Adaption System was proposed to monitor, quantify, and mitigate the SOTIF risk. Then, a demonstration system was developed by estimating the uncertainty of YOLOv5 to monitor the perceptual risk, quantifying the uncertainty as SOTIF entropy, and mitigating the entropy via a Perceptual Uncertainty-Aware Decision-Making (PUADM) technique. Ten networks were trained on the open traffic dataset BDD to construct various ensembles, with individual network mAP scores ranging from 0.511 to 0.516, while the mAP of the ensemble reached 0.529. Through experimentation, it was confirmed that an ensemble composed of five sets of weights exhibited advantages in terms of real-time performance and effectiveness. Diverse perceptual SOTIF scenarios were further evaluated to confirm the performance of the perception module, with an alert coverage rate of 90.2% and a false alert rate of 10.3%. Afterward, Hardware-in-the-Loop (HIL) experiments were conducted to confirm the system's real-time performance and effectiveness. Experimental experiments demonstrate that the estimated perception SOTIF entropy is dependable and the PUADM method is safer than the baseline method without being too conservative. Meanwhile, the computational speed of the above system meets the real-time requirement of 30 fps.

This paper quantified only the perception SOTIF entropy, disregarding the prediction SOTIF entropy and the planning SOTIF entropy, as well as the potential impact of prediction errors on planning in the demonstration. The system will become more complex if learning-based prediction and planning algorithms such as Long Short-Term Memory and Reinforcement Learning are considered. Future research on the overall SOTIF risk will focus on how to quantify the SOTIF entropy from learning-based prediction and planning algorithms and how to sort out the coupling relationship between them. Although the performance of the perception module and the PUADM method has been verified through simulation experiments on the Apollo platform and the HIL platform, respectively, this system has not been exposed to the actual perceptual SOTIF scenarios. The Self-Surveillance and Self-Adaption System will be deployed to an autonomous truck to conduct field experiments to verify its performance.

APPENDIX

Fig. 10 shows some perception results of the probabilistic object detector based on YOLOv5 and Deep Ensembles on the PeSOTIF dataset. Dotted lines and shadows are used to demonstrate the bounding box and its related spatial uncertainty. The information bar shows the class name, confidence score, number of detections among 5 networks, perception entropy, and uncertainty level.

In these scenarios, objects disturbed by the environmental factors or with abnormal appearance and posture are hard to locate and have high entropy. Meanwhile, other normal objects can usually be correctly detected and classified with low entropy. For example, the truck with a strange appearance in (h) is correctly classified, but it is meaningful to output high entropy because it seems weird to human drivers.

Furthermore, we conducted field experiments using a front-facing camera mounted on a truck, encompassing scenar-

ios characterized by environmental conditions such as daytime, nighttime, rain, and glare, as well as critical objects like sanitation workers, traffic cones, pedestrians with umbrellas, and overturned vehicles, etc. Fig. 11 shows some samples.

ACKNOWLEDGMENT

The authors would like to appreciate the contributions of the perception task group of the CAICV-SOTIF technical alliance in China.

REFERENCES

- [1] *Road Vehicles-Safety of the Intended Functionality*, Standard ISO/DIS 21448, Geneva, Switzerland, 2021.
- [2] *Road Vehicles-Functional Safety*, Standard ISO 26262, Geneva, Switzerland, 2011.
- [3] *California Department of Motor Vehicles. Disengagement Reports*. Accessed: Feb. 2021. [Online]. Available: <https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/>
- [4] National Highway Traffic Safety Administration. (Jan. 2017). *PE16007. Tesla Crash Preliminary Evaluation Report*. [Online]. Available: <https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.PDF>
- [5] National Transportation Safety Board. (Mar. 2018). *Preliminary Report Highway: HWY18MH010. Highway Accident Report Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian Tempe, Arizona*. [Online]. Available: <https://www.nts.gov/investigations/AccidentReports/Reports/HAR1903.pdf>
- [6] X. Wu, H. Meng, X. Xing, and J. Chen, "Edge test case generation of automatic driving system," *J. Tongji Univ., Natural Sci.*, vol. 10, no. 46, pp. 111–115, 2018.
- [7] A. Huang, X. Xing, T. Zhou, and J. Chen, "A safety analysis and verification framework for autonomous vehicles based on the identification of triggering events," SAE Tech. Paper 2021-01-5010, 2021.
- [8] K. Li et al., "CODA: A real-world road corner case dataset for object detection in autonomous driving," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 406–423.
- [9] L. Peng, J. Li, W. Shao, and H. Wang, "PeSOTIF: A challenging visual dataset for perception SOTIF problems in long-tail traffic scenarios," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2023, pp. 1–8.
- [10] W. Hong, P. Liang, L. Jun, Y. Wenhao, and X. Xiong, "Safety decision of running speed based on real-time weather," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2022, pp. 212–217.
- [11] J. Liu, H. Wang, L. Peng, Z. Cao, D. Yang, and J. Li, "PNNUAD: Perception neural networks uncertainty aware decision-making for autonomous vehicle," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24355–24368, Dec. 2022.
- [12] W. Shao, Y. Xu, L. Peng, J. Li, and H. Wang, "Failure detection for motion prediction of autonomous driving: An uncertainty perspective," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 12721–12728.
- [13] K. Yang, B. Li, W. Shao, X. Tang, X. Liu, and H. Wang, "Prediction failure risk-aware decision-making for autonomous vehicles on signalized intersections," *IEEE Trans. Intell. Transp. Syst.*, early access, Jul. 17, 2023, doi: [10.1109/TITS.2023.3288507](https://doi.org/10.1109/TITS.2023.3288507).
- [14] I. Colwell, B. Phan, S. Saleem, R. Salay, and K. Czarnecki, "An automated vehicle safety concept based on runtime restriction of the operational design domain," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1910–1917.
- [15] A. Reschka, J. R. Böhmer, T. Nothdurft, P. Hecker, B. Lichte, and M. Maurer, "A surveillance and safety system based on performance criteria and functional degradation for an autonomous vehicle," in *Proc. 15th Int. IEEE Conf. Intell. Transp. Syst.*, Sep. 2012, pp. 237–242.
- [16] Q. M. Rahman, N. Sünderhauf, and F. Dayoub, "Per-frame mAP prediction for continuous performance monitoring of object detection during deployment," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2021, pp. 152–160.
- [17] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5574–5584.
- [18] K. Czarnecki and R. Salay, "Towards a framework to manage perceptual uncertainty for safe automated driving," in *Proc. Int. Conf. Comput. Saf., Rel., Secur.* Cham, Switzerland: Springer, Sep. 2018, pp. 439–445.

- [19] D. Feng, A. Harakeh, S. L. Waslander, and K. Dietmayer, "A review and comparative study on probabilistic object detection in autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 9961–9980, Aug. 2022.
- [20] J. Mena, O. Pujol, and J. Vitria, "A survey on uncertainty estimation in deep learning classification systems from a Bayesian perspective," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–35, Dec. 2022.
- [21] D. J. C. MacKay, "A practical Bayesian framework for backpropagation networks," *Neural Comput.*, vol. 4, no. 3, pp. 448–472, May 1992.
- [22] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *J. Mach. Learn. Res.*, vol. 14, pp. 1303–1347, Jan. 2013.
- [23] T. Chen, E. Fox, and C. Guestrin, "Stochastic gradient Hamiltonian Monte Carlo," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2014, pp. 1683–1691.
- [24] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.
- [25] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1019–1027.
- [26] Y. Gal, J. Hron, and A. Kendall, "Concrete dropout," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3581–3590.
- [27] T. Azevedo, R. de Jong, M. Mattina, and P. Maji, "Stochastic-YOLO: Efficient probabilistic object detection under dataset shifts," 2020, *arXiv:2009.02967*.
- [28] L. Peng, H. Wang, and J. Li, "Uncertainty evaluation of object detection algorithms for autonomous vehicles," *Automot. Innov.*, vol. 4, no. 3, pp. 241–252, Aug. 2021.
- [29] J. Mukhoti and Y. Gal, "Evaluating Bayesian deep learning methods for semantic segmentation," 2018, *arXiv:1811.12709*.
- [30] I. Osband and B. Van Roy, "Bootstrapped Thompson sampling and deep exploration," 2015, *arXiv:1507.00300*.
- [31] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6402–6413.
- [32] A. Kleiner, A. Talwalkar, P. Sarkar, and M. Jordan, "The big data bootstrap," 2012, *arXiv:1206.6415*.
- [33] K. Doshi and Y. Yilmaz, "Road damage detection using deep ensemble learning," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2020, pp. 5540–5544.
- [34] D. Miller, N. Sünderhauf, H. Zhang, D. Hall, and F. Dayoub, "Benchmarking sampling-based probabilistic object detectors," in *Proc. CVPR Workshops*, vol. 3, Jun. 2019, p. 6.
- [35] S. Choi, K. Lee, S. Lim, and S. Oh, "Uncertainty-aware learning from demonstration using mixture density networks with sampling-free variance modeling," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 6915–6922.
- [36] J. Choi, D. Chun, H. Kim, and H.-J. Lee, "Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 502–511.
- [37] J. Mena, O. Pujol, and J. Vitrià, "Uncertainty-based rejection wrappers for black-box classifiers," *IEEE Access*, vol. 8, pp. 101721–101746, 2020.
- [38] J. Mena, O. Pujol, and J. Vitrià, "Dirichlet uncertainty wrappers for actionable algorithm accuracy accountability and auditability," 2019, *arXiv:1912.12628*.
- [39] P. Sadowski and P. Baldi. (Sep. 2019). *Neural Network Regression with Beta, Dirichlet, and Dirichlet-Multinomial Outputs*. [Online]. Available: <https://openreview.net/forum?id=BJeRg205Fm>
- [40] A. Harakeh, M. Smart, and S. L. Waslander, "BayesOD: A Bayesian approach for uncertainty estimation in deep object detectors," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 87–93.
- [41] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 14927–14937.
- [42] J. Postels, F. Ferroni, H. Coskun, N. Navab, and F. Tombari, "Sampling-free epistemic uncertainty estimation using approximated variance propagation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2931–2940.
- [43] J. Gast and S. Roth, "Lightweight probabilistic deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3369–3378.
- [44] G. Kahn, A. Villafior, V. Pong, P. Abbeel, and S. Levine, "Uncertainty-aware reinforcement learning for collision avoidance," 2017, *arXiv:1702.01182*.
- [45] B. Lütjens, M. Everett, and J. P. How, "Safe reinforcement learning with model uncertainty estimates," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 8662–8668.
- [46] J. Wang, J. Wu, and Y. Li, "The driving safety field based on driver-vehicle-road interactions," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2203–2214, Aug. 2015.
- [47] H. Wang, Y. Huang, A. Khajepour, D. Cao, and C. Lv, "Ethical decision-making platform in autonomous vehicles with lexicographic optimization based model predictive controller," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8164–8175, Aug. 2020.
- [48] H. Wang, A. Khajepour, D. Cao, and T. Liu, "Ethical decision making in autonomous vehicles: Challenges and research progress," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 1, pp. 6–17, Jan. 2022.
- [49] B. Ivanovic et al., "Heterogeneous-agent trajectory forecasting incorporating class uncertainty," 2021, *arXiv:2104.12446*.
- [50] X. Tang et al., "Prediction-uncertainty-aware decision-making for autonomous vehicles," *IEEE Trans. Intell. Vehicles*, vol. 7, no. 4, pp. 849–862, Dec. 2022.
- [51] K. Rezaee, P. Yadmellat, and S. Chamorro, "Motion planning for autonomous vehicles in the presence of uncertainty using reinforcement learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 3506–3511.
- [52] H. Wang, Y. Huang, A. Khajepour, Y. Zhang, Y. Rasekhipour, and D. Cao, "Crash mitigation in motion planning for autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 9, pp. 3313–3323, Sep. 2019.
- [53] X. Zhang, M. Zhou, W. Shao, T. Luo, and J. Li, "The architecture of the intended safety system for intelligent driving," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2019, pp. 1–4.
- [54] T. F. Cai and J. L. Zhang, "Discussion on operational mechanism of safety system-degree of safety and entropy of safety," *China Saf. Sci. J.*, vol. 6, pp. 4–7, Jan. 2006.
- [55] D. Miller, F. Dayoub, M. Milford, and N. Sünderhauf, "Evaluating merging strategies for sampling-based uncertainty techniques in object detection," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 2348–2354.
- [56] Y. Gal, "Uncertainty in deep learning." Ph.D. thesis, Dept. Eng., Univ. Cambridge, Cambridge, U.K., 2016.
- [57] S. Park, K. Oh, Y. Jeong, and K. Yi, "Model predictive control-based fault detection and reconstruction algorithm for longitudinal control of autonomous driving vehicle using multi-sliding mode observer," *Microsyst. Technol.*, vol. 26, no. 1, pp. 239–264, Jan. 2020.
- [58] K. Yang, X. Tang, Y. Qin, Y. Huang, H. Wang, and H. Pu, "Comparative study of trajectory tracking control for automated vehicles via model predictive control and robust H-infinity state feedback control," *Chin. J. Mech. Eng.*, vol. 34, no. 1, pp. 1–14, Dec. 2021.
- [59] D. Thuan, "Evolution of YOLO algorithm and YOLOv5: The state-of-the-art object detection algorithm," Ph.D. thesis, Oulu Univ. Appl. Sci., Oulu, Finland, 2021.
- [60] F. Yu et al., "BDD100K: A diverse driving video database with scalable annotation tooling," 2018, *arXiv:1805.04687*.
- [61] MarkDana. (2018). *Realtime Cone Detection: Traffic Cone Dataset*. [Online]. Available: <https://github.com/MarkDana/RealtimeConeDetection>
- [62] X. Wang. (2020). *TensorRTx*. [Online]. Available: <https://github.com/wang-xinyu/tensortrx>



Liang Peng received the B.Eng. degree in vehicle engineering from the School of Vehicle and Mobility, Tsinghua University, Beijing, China, in 2020, where he is currently pursuing the Ph.D. degree in vehicle engineering. He is a member of the Tsinghua Intelligent Vehicle Design and Safety (IVDAS) Research Institute and supervised by Prof. Jun Li and Assoc. Prof. Hong Wang. His research interests include evaluation and uncertainty analysis of perceptual algorithms and safety of autonomous driving vehicles.



Boqi Li received the B.S. degree in mechanical engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2015, and the M.S. degree in mechanical engineering from Stanford University, Stanford, CA, USA, in 2017. He is currently pursuing the Ph.D. degree in mechanical engineering with the University of Michigan, Ann Arbor, MI, USA.



Wenbo Shao received the B.E. degree in vehicle engineering from Tsinghua University, Beijing, China, in 2019. He is currently pursuing the Ph.D. degree in mechanical engineering with Tsinghua University. He is a member of the Tsinghua Intelligent Vehicle Design and Safety Research Institute (IVDAS) and supervised by Prof. Jun Li and Assoc. Prof. Hong Wang. His research interests include safety of the intended functionality of autonomous driving, prediction, decision-making, uncertainty theory, and applications.



Wenhao Yu received the Ph.D. degree from Jiangsu University, China, in 2020. He is currently a Research Associate with the School of Vehicle and Mobility, Tsinghua University. His research interests include decision-making, path planning and following control of autonomous vehicles, model predictive control, and safety of the intended functionality of autonomous vehicles.



Kai Yang received the B.E. degree in vehicle engineering from the Wuhan University of Technology in 2018. He is currently pursuing the Ph.D. degree with the Department of Automotive Engineering, Chongqing University, Chongqing, China. He researches as a Joint Ph.D. Student with the School of Vehicle and Mobility, Tsinghua University, Beijing, China. His research interests include motion prediction and decision-making of autonomous vehicles.



Hong Wang (Senior Member, IEEE) received the Ph.D. degree from the Beijing Institute of Technology, China, in 2015. From 2015 to 2019, she was a Research Associate of mechanical and mechatronics engineering with the University of Waterloo. She is currently a Research Associate Professor with Tsinghua University. She has published more than 60 articles on top international journals. Her research interests include the safety of the on-board AI algorithm, the safe decision-making for intelligent vehicles, and the test and evaluation of SOTIF.

Her domestic and foreign academic part-time includes an Associate Editor of *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, *IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY*, and *IEEE TRANSACTIONS ON INTELLIGENT VEHICLES*, a Young Communication Expert of Engineering, and the Lead Guest Editor of Special Issues on Intelligent Safety of *IEEE Intelligent Transportation Systems Magazine*.