Object Classification Based on Enhanced Evidence Theory: Radar–Vision Fusion Approach for Roadside Application

Pengfei Liu, Guizhen Yu, Zhangyu Wang, Bin Zhou, and Peng Chen, *Member, IEEE*

Abstract—**Roadside object detection and classification provide a good understanding of driving scenarios in regard to over-the-horizon perception. However, typical roadside sensors are insufficient when used separately. The fusion of the millimeter-wave (MMW) radar and monovision camera serves as an efficient approach. Unfortunately, the uncertain and conflicting data in extreme light conditions pose challenges to the fusion process. To this end, this study proposed an evidential framework to fuse the radar and camera data. A novel modeling approach for basic belief assignments (BBAs) was proposed, which took the uncertainty of convolutional neural network (CNN) model into consideration. Moreover, the single-scan and multiscan fusion methods were developed based on the enhanced evidence theory, which utilized different weighted coefficients by referring to the reinforced belief (RB) divergence measure and belief entropy (BE). Both numerical and empirical experiments were conducted to investigate the method performance. Specifically, in numerical experiments, the belief value of actual classification increased to 99.01%. For empirical experiments, based on the real datasets collected by roadside devices, the proposed method was demonstrated to outperform the state-of-the-art ones in terms of 71.06% and 87.23% precisions for bright light and low illumination conditions, respectively. The results verify that the proposed method is effective in fusing the conflicting and uncertain data.**

Index Terms—**Evidence theory, object classification, roadside sensor, uncertainty estimation.**

I. Introduction

ROAD traffic safety and efficiency are the key challenges in modern transportation. With the rapid development of 5G communication technology, the growth of the related technology of intelligent transportation systems has accelerated, such as intelligent roadside units (RSUs). RSUs can achieve over-the-horizon perception and be used for collision prevention in blind areas to improve traffic safety. Moreover, the essential task of intelligent RSUs is to build a stable and reliable perception system [1].

Different types of sensors have been employed for roadside object detection, such as cameras [2]–[7], millimeter-wave (MMW) radars [8], ultrasonic sensors [9], and light detection and rangings (LiDARs) [10]–[14]. Owing to the advantages of cost-effectiveness and delivery of high-quality texture information, vision processing methods have become prominent in recent years. With the development of deep learning, its potential applications in object detection have been identified. Although great progress has been achieved in vision object detection, extreme light conditions cause significant reductions in the detection robustness.

In contrast, an MMW radar can operate reasonably in all weather and light conditions. It also achieves high-precision speed measurements. However, its applications are limited by its sparse spatial resolution [15]. In summary, the detection capabilities of a vision sensor and an MMW radar could complement each other. Fusing these two sensors is considered as an efficient approach to increase the detection accuracy.

Nevertheless, the undesirable performance of the camera in extreme light conditions and the radar subject to noise leads to uncertain and conflicting data, which poses challenges to the fusion of radars and cameras. Previous research [16]–[18] used the evidence theory to manage uncertain data. However, it has drawbacks in handling conflicting data [19]. In addition, mapping the detection results to the basic belief assignment (BBA) of evidence theory is still an open issue, especially for the output of the deep-learning model.

To address these problems, this study proposed an evidential architecture, as illustrated in Fig. 1. First, radar points were clustered to obtain an object list. Concurrently, an image was input to the convolutional neural network (CNN) model for detection and determining the corresponding uncertainty. Based on the detection results, the BBAs for the radar and the camera were modeled, respectively. Notably, a novel method for mapping the deep-learning output to BBA was proposed, which was integrated with the uncertainty estimation of the CNN detection output. To the best of our knowledge, this is the first study to introduce the uncertainty of the deep-learning methods to BBA modeling. Subsequently, the BBAs of the single scan and multiple scans were fused based on the enhanced evidence theory, which modified the evidence body

Manuscript received November 9, 2021; revised January 19, 2022; accepted February 1, 2022. Date of publication February 28, 2022; date of current version March 14, 2022. This work was supported in part by the National Key Technologies Research and Development Program of China under Grant 2020YFB1600301 and in part by the National Natural Science Foundation of China under Grant 52072020. The Associate Editor coordinating the review process was Huan Liu. *(Corresponding author: Zhangyu Wang.)*

Pengfei Liu and Peng Chen are with the School of Transportation Science and Engineering, Beihang University, Beijing 100191, China (e-mail: liupengfei2019@buaa.edu.cn; cpeng@buaa.edu.cn).

Guizhen Yu and Bin Zhou are with the School of Transportation Science and Engineering, Beihang University, Beijing 100191, China, and also with the Hefei Innovation Research Institute, Beihang University, Hefei 230012, China (e-mail: yugz@buaa.edu.cn; binzhou@buaa.edu.cn).

Zhangyu Wang is with the Research Institute for Frontier Science, Beihang University, Beijing 100191, China, and also with the Hefei Innovation Research Institute, Beihang University, Hefei 230012, China (e-mail: zywang@buaa.edu.cn).

Digital Object Identifier 10.1109/TIM.2022.3154001

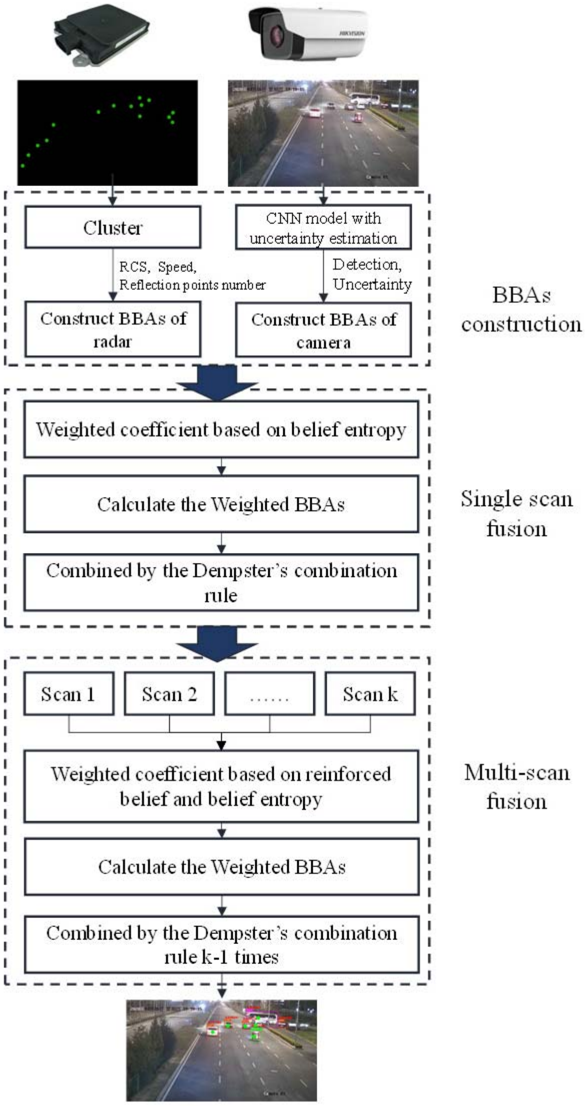

Cluster
CNN model with uncertainty estimation
RCS, Speed, Reflection points number
Detection, Uncertainty
Construct BBAs of radar
Construct BBAs of camera
BBAs construction
Weighted coefficient based on belief entropy
Calculate the Weighted BBAs
Combined by the Dempster's combination rule
Single scan fusion
Scan 1
Scan 2
……
Scan k
Weighted coefficient based on reinforced belief and belief entropy
Calculate the Weighted BBAs
Combined by the Dempster's combination rule k-1 times
Multi-scan fusion

Fig. 1. Architecture of the evidential framework.

by different weighted coefficients. Specifically, the single scan and multiple scans represent the conditions of two or more evidence bodies, respectively. The weighted coefficient of the single scan was calculated with two evidence bodies based on the belief entropy (BE) to measure the information volume. The BBAs were weighted by the coefficient and combined by Dempster's combination rule. For multiple scans with three or more evidence bodies, in addition to the BE, the credibility was obtained based on the reinforced belief (RB) divergence measurement. Similarly, the BBAs were weighted and combined. Finally, the proposed fusion method was implemented in roadside perception devices in an actual application. Empirical datasets in bright light and low illumination scenarios were collected to verify the performance of the proposed method.

The major contributions of this study are summarized as follows.

1) An evidential architecture for radar–vision fusion is designed to manage the uncertain and conflicting data. According to the number of evidence bodies for single-scan and multiscan fusion, different weighted coefficients of evidence theory are designed.
2) A novel modeling approach for the BBAs based on the deep-learning classification results is proposed, which considers the uncertainty of deep-learning model.
3) An enhanced weighted approach of evidence theory is proposed, which weights the evidence body by the credibility based on RB divergence and information volume based on BE. Such approach helps derive more reasonable BBAs by considering the relationships between belief functions and subsets of the sets of belief functions, as well as the effect of the evidence itself on the weight.

The remainder of this article is organized as follows. Section II summarizes the related studies. Section III briefly introduces the preliminaries of the Dempster–Shafer enhanced evidence theory. Object classification based on the enhanced evidence theory is presented in Section IV. The performance evaluation of the proposed method is extensively discussed in Section V. Section VI concludes this article.

II. Related Work

Numerous methods have been proposed to fuse heterogeneous data of roadside radars and cameras by fully utilizing their complementary advantages. In [20], a high time–frequency resolution of the signals of a radar was obtained by employing the method of time–frequency reassignment. Subsequently, the 3-D tracking information obtained from a calibrated video camera was fused with the information from the radar to detect targets within a surveillance region. Fu *et al.* [21] used YOLOv3 to process camera data and the density-based spatial clustering of applications with noise (DBSCAN) clustering method to process radar data, in order to obtain the position, speed, and category of a detected target. A Kalman filter was used to fuse the detection by the radar and camera. Bai *et al.* [22] proposed an optimal attribute fusion algorithm for target detection and tracking based on an improved Gaussian mixture probability hypothesis density filter.

Concurrently, fusion detection using radars and cameras has achieved desirable performance in autonomous driving. These studies are important references and guides for the application of cameras and radar sensors installed at roadsides. Jiang *et al.* [23] proposed employing an image detection method in the region of interest (ROI) generated by an MMW radar to eliminate false alarms. This fusion scheme helps reduce the computational burden of vision computing. However, the detection accuracy was influenced by the number of effective radar points in the ROI. In addition, some studies focused on decision fusion, which achieved more reliability and robustness in practice. Chen *et al.* [24] proposed a two-level association structure combining regional collision association and weighted track association. Moreover, they used a nonreset federated filter to fuse the associated tracks. Cho *et al.* [25] employed the sequential sensor method, which treated different sensors independently and fed them sequentially to the estimation process of an extended Kalman filter. Chavez-Garcia and Aycard [16] and Bouain *et al.* [17] proposed an evidential framework to fuse data from different sensors and evaluated the method based on the actual dataset. Owing to the desirable performance of the evidence theory in uncertainty management, this study focused on uncertain and conflicting data management based on evidence theory in the fusion process.

As aforementioned, open issues remain in evidence theory, namely, how to fuse conflicting evidence and map the result of sensor detection to BBAs. Most studies addressed the first issue from two perspectives. One is to modify the combination rules. The related studies contain Smets's [26] unnormalized combination rule and Yager's [27] combination rule as used in [16] and [17]. However, the modification of the combination rule would has no effect when dealing with the counterintuitive results caused by the sensor failure [28].

Thus, increasing studies resort to preprocessing the bodies of evidence when fusing conflicting data. One typical approach is Murphy's method, which computes the average of the evidence bodies [29]. Deng *et al.* [30] improved the method by the weighted average of the evidence based on the Jousselme distance. Similarly, Awogbami *et al.* [31] used Jousselme distance to measure the credibility degree of evidence, and the reliability factor was further applied to modify the evidence. Fan *et al.* [32] proposed the group evidence based on the conflict coefficient and the BE. The group evidence was weighted based on the Jousselme distance. Khan and Anwar [33] penalized the sensor by the accuracy and weighted the evidence by credibility degree based on Euclidean distance. Xiao [28] made an improvement by introducing the RB divergence and considered the relationships between subsets of the sets of belief functions. However, the effect of the evidence itself on the weight was ignored. In addition, some research was proposed to combine the above two types of methods [34].

Another issue relates to the BBAs modeling, which serves as the basis for the application of evidence theory. Chavez-Garcia and Aycard [16] and Bouain *et al.* [17] presented the modeling approaches of BBAs and conducted verifications based on the actual datasets. Chavez-Garcia and Aycard [16] assumed the class-related factors and discount factor to indicate the performance of LiDAR and camera detector. The BBAs of radar were decided by the prior information of speed for vehicle and person. Bouain *et al.* [17] built the BBAs of the camera in a similar way and modeled the BBAs of radar by considering radar cross section (RCS). As the potential applications of deep learning in object detection have been identified, its combination with the fusion method would be promising. Hence, this study will explore the relationship between the output of the deep learning approach and BBA.

III. Preliminaries

In this section, the preliminaries of the Dempster–Shafer evidence theory are first introduced.

A. Dempster–Shafer Evidence Theory

The Dempster–Shafer evidence theory [35], [36] can model imprecise and uncertain data with less prior information. Hence, it has additional flexibility and effectiveness in managing uncertain data. The basics of the evidence theory are introduced below.

A set of mutually exclusive and collectively exhaustive events are denoted as U, which is also referred to as a frame of discernment. A mass function mapping m from 2^U to [0, 1] is called as BBA, which meets the following conditions:

$$m(\emptyset) = 0 \quad \text{and} \quad \sum_{A \in 2^U} m(A) = 1. \tag{1}$$

For a proposition $A \subseteq U$, a belief function is defined as

$$\mathrm{Bel}(A) = \sum_{B \subseteq A} m(B). \tag{2}$$

Assuming that two BBAs m_1 and m_2 in the frame of discernment U are independent, Dempster's rule of combination, denoted by $m = m_1 \oplus m_2$, is defined as follows:

$$\begin{aligned} m(A) &= \begin{cases} \dfrac{1}{1-K} \displaystyle\sum_{X_i \cap Y_i = A} m_1(X_i) m_2(Y_i), & A \neq \emptyset \\ 0, & A = \emptyset \end{cases} \\ K &= \sum_{X_i \cap Y_j = \emptyset} m_1(X_i) m_2(Y_j) \end{aligned} \tag{3}$$

where X_i and Y_i are the elements of 2^U and K is a constant representing the conflicts between the BBAs.

In this study, we define a set of BBAs $m(2^U)$ for all possible hypotheses of object classification. In a road environment, $U = \{\text{car,truck,bus,van,motorcycle,pedestrian}\}$ is the frame of discernment representing the classes of interest.

B. RB Divergence Measure

The RB divergence measure was proposed by Xiao [28] to characterize the degree of discrepancy and conflict among evidence by considering the correlations between belief functions and subsets of the sets of belief functions.

For m_1 and m_2 in the frame of discernment U, let A_i and A_j be the hypotheses of m_1 and m_2, respectively. The belief divergence measure between the two BBAs is denoted by and defined as follows:

$$\begin{aligned} &\mathcal{B}(m_1, m_2) \\ &= \frac{1}{2} \sum_{i=1}^{2^U} \sum_{j=1}^{2^U} m_1(A_i) \log \frac{m_1(A_i)}{\frac{1}{2} m_1(A_i) + \frac{1}{2} m_2(A_j)} \frac{|A_i \cap A_j|}{|A_j|} \\ &\quad + \frac{1}{2} \sum_{i=1}^{2^U} \sum_{j=1}^{2^U} m_2(A_j) \log \frac{m_2(A_j)}{\frac{1}{2} m_1(A_i) + \frac{1}{2} m_2(A_j)} \frac{|A_i \cap A_j|}{|A_j|} \end{aligned} \tag{4}$$

where is the cardinality of set A_i.

The RB divergence measure for the BBAs is devised based on the $\mathcal{B}$ divergence

$$\mathfrak{RB}(m_1, m_2) = \sqrt{\frac{|\mathcal{B}(m_1, m_1) + \mathcal{B}(m_2, m_2) - 2\mathcal{B}(m_1, m_2)|}{2}}. \tag{5}$$

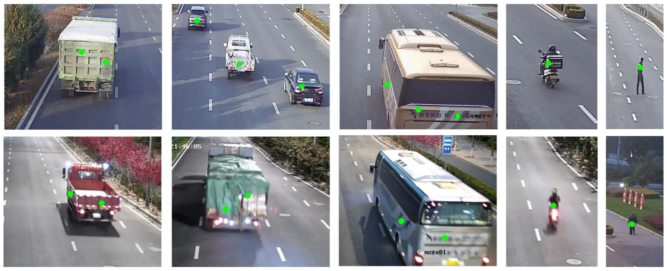


Fig. 2. Radar reflection points transformed into pixel coordinates.

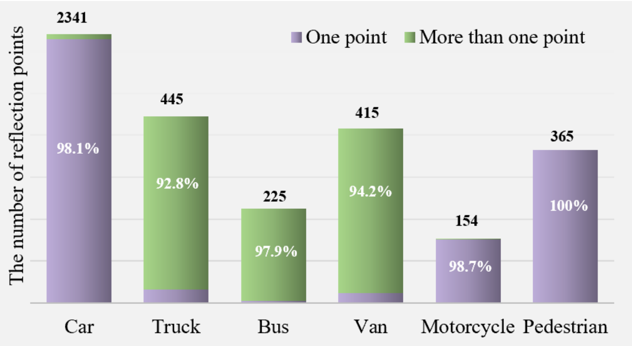


Fig. 3. Proportions of different numbers of reflecting points.

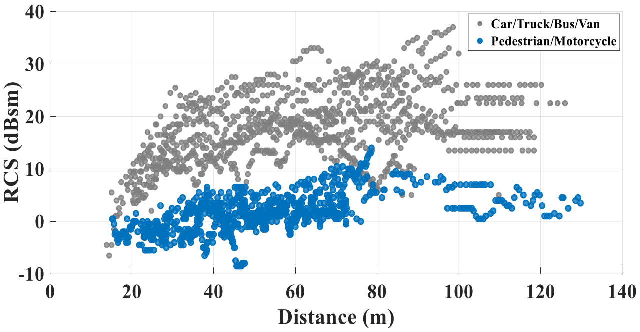


Fig. 4. RCS values of different objects.

C. Belief Entropy

Deng [37] proposed the BE, also called the Deng entropy, to measure information volume effectively. Deng entropy E_d of a set A_i is defined as follows:

$$E_d = -\sum_i m(A_i) \log \frac{m(A_i)}{2^{|A_i|} - 1}. \tag{6}$$

A larger value of the Deng entropy indicates that the evidence contains more information and thus plays a more important role in the final combination.

IV. Proposed Approach

In the proposed radar–vision fusion, the parameters of the BBAs for the MMW radar and the camera are first estimated separately. Subsequently, the single-scan data of the radar and the camera are fused based on the enhanced evidence theory. On this basis, the method of multiscan data fusion is proposed to improve the classification performance.

A. Parameter Estimation of BBAs for MMW Radar

Object classification based on an MMW radar is difficult in practice owing to the sparse resolution of the radar point. However, in view of the prominent performance in speed measurement and the robustness in different weather conditions, radar-based classification needs further investigation. Elements of a radar detection include the range, azimuth, relative speed, and RCS. The speed information can be used for classification purpose based on empirical data. The RCS is influenced by various factors including the object material, object size, and angle. In addition, practical analyses show that trucks, vans, and buses reflect more multiple radar points than cars, motorcycles, and pedestrians. Thus, we propose obtaining the BBAs for radar based on the number of reflection points, RCS, and speed.

1) Number of Reflection Points: A radar sensor uses a built-in mechanism to detect moving obstacles; however, it yields imprecise clustering results. Therefore, experiments were conducted to collect radar data for exploiting the original output (which will be introduced in detail in Section V). The data were manually classified based on the number of reflection points of the detected object, which is shown in Fig. 2. It can be found that the *truck*, *van*, and *bus* classes reflect relatively more radar points than the *car*, *motorcycle*, and *pedestrian* classes. The empirical analysis results are presented in Fig. 3. It can be found that the proportions of more than one reflecting point reach 92.8%, 97.9%, and 94.2% for the *truck*, *bus*, and *van* classes, respectively. Moreover, the proportions of one reflecting point reach 98.1%, 98.7%, and 100% for the *car*, *motorcycle*, and *pedestrian* classes, respectively. Thus, a truck, bus, or van can be distinguished from other types of objects (a car, pedestrian, or motorcycle) based on the number of reflection points. In the detection process, the original output is clustered again based on the Euclidean distance. The numbers of reflection points of each object are saved for building the BBAs

2) Radar Cross Section: An MMW radar can obtain accurate position and speed information of an object; however, object classification is difficult owing to sparse point clouds. Based on the experimental results (which will be introduced in detail in Section V), the RCS values of the MMW radar vary with the detection distance and object type. As shown in Fig. 4, when the detection distance is over 20 m, the RCS values of *cars/trucks/buses/vans* are notably larger than those of *motorcycles/pedestrians*. Thus, an initial classification can be conducted by referring to the RCS values at a detection distance exceeding the threshold of 20 m.

3) BBA Modeling of Radar: $m_r(A)$ is defined for each class, which describes the evidence distribution for the class of the object detected by the radar. Parameters α_n, α_s, α_r, and α_e are used to represent the probability to detect the *truck/bus/van*, *car*, *motorcycle/pedestrian*, and other classes, respectively. They all vary in the range of [0, 1]. Based on the above-mentioned experimental results, if the number of reflection points, n, is larger than the threshold, n_{thresh} (defined as 1 in this study), the possibility that the detected object belongs to

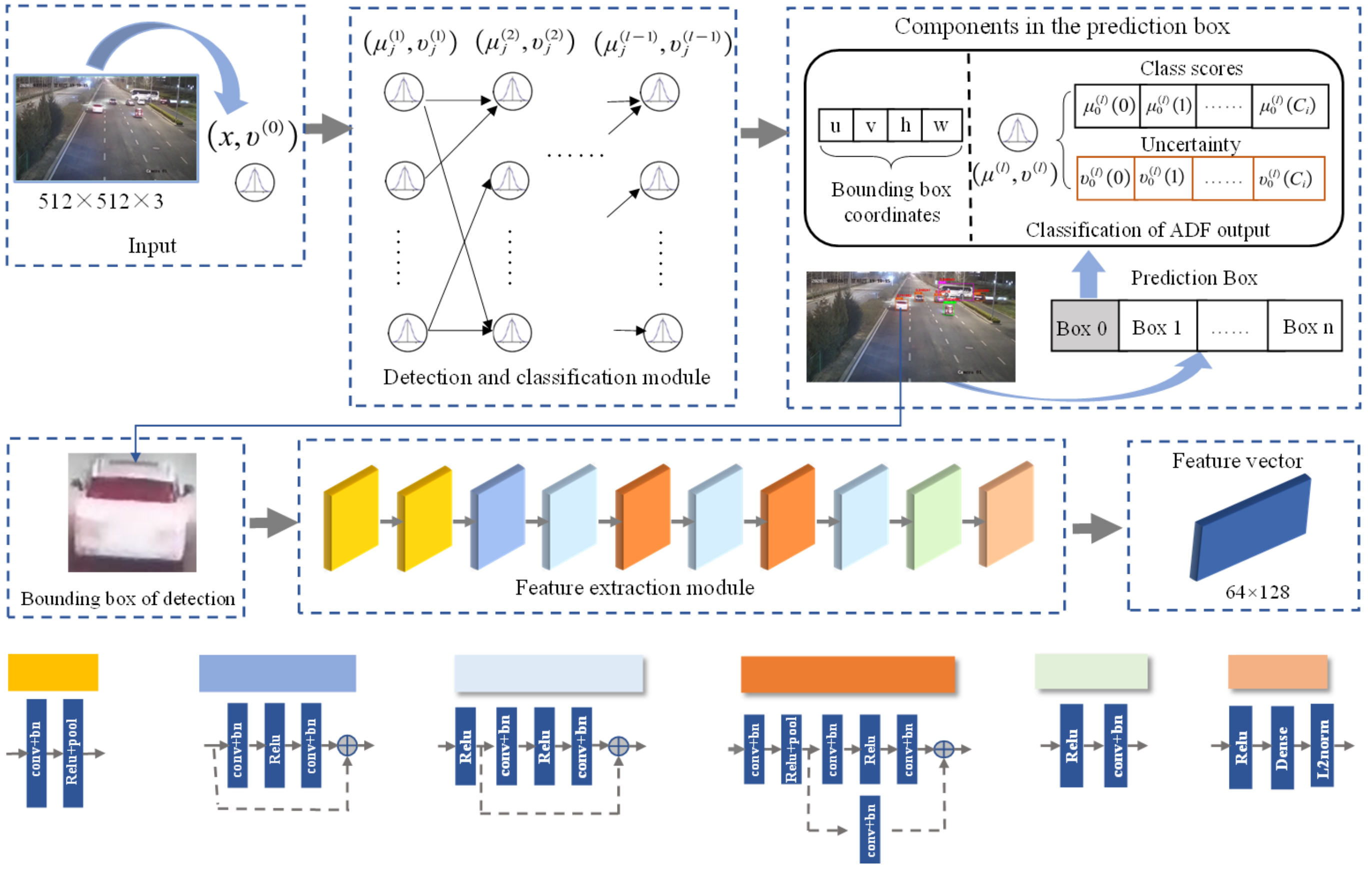

$(x, \upsilon^{(0)})$
512×512×3
Input
$(\mu_j^{(1)}, \upsilon_j^{(1)})$ $(\mu_j^{(2)}, \upsilon_j^{(2)})$ $(\mu_j^{(l-1)}, \upsilon_j^{(l-1)})$
Detection and classification module
Components in the prediction box
u v h w
Bounding box coordinates
$(\mu^{(l)}, \upsilon^{(l)})$
Class scores
$\mu_0^{(l)}(0)$ $\mu_0^{(l)}(1)$ …… $\mu_0^{(l)}(C_i)$
Uncertainty
$\upsilon_0^{(l)}(0)$ $\upsilon_0^{(l)}(1)$ …… $\upsilon_0^{(l)}(C_i)$
Classification of ADF output
Prediction Box
Box 0 Box 1 …… Box n
Bounding box of detection
Feature extraction module
Feature vector
64×128
conv+bn
Relu+pool
conv+bn
Relu
conv+bn
Relu
conv+bn
Relu
conv+bn
conv+bn
Relu+pool
conv+bn
Relu
conv+bn
conv+bn
Relu
conv+bn
Relu
Dense
L2norm

Fig. 5. Structure of CNN model comprising detection and feature extraction modules.

the *truck/bus/van class*, i.e., α_n, is close to 1. In contrast, if the number of reflection points, n, is less than or equal to n_{thresh} and the detected speed exceeds the threshold, S_{thresh}, there is a high possibility, i.e., α_s, that the object belongs to the *car* class. If n is less than or equal to n_{thresh} and the RCS is less than the threshold, R_{thresh}, the detected object may be a motorcycle or a pedestrian, as denoted by α_r. The BBAs for the detections of radar are represented in (7) and will be fused based on the enhanced evidence theory as in Section IV-C

$$\begin{cases} m_r(\text{truck,van,bus}) = \alpha_n, & \text{if } n > n_{\text{thresh}} \\ m_r(\text{others}) = 1 - \alpha_n \\ m_r(\text{car}) = \alpha_s \\ \qquad\qquad\qquad \text{if } n \le n_{\text{thresh}} \text{ and speed} > S_{\text{thresh}} \\ m_r(\text{others}) = 1 - \alpha_s \\ m_r(\text{pedestrian,motorcycle}) = \alpha_r \\ \qquad\qquad\qquad \text{if } n \le n_{\text{thresh}} \text{ and RCS} < R_{\text{thresh}} \\ m_r(\text{others}) = 1 - \alpha_r \\ m_r(\text{car}) = \alpha_e, & \text{else} \\ m_r(\text{others}) = 1 - \alpha_e. \end{cases} \tag{7}$$

B. Parameter Estimation of BBAs for Camera

An uncertainty-aware CNNs model for object detection is presented in this section. Its architecture can be divided into two parts: a detection and classification module and a feature extraction module, as shown in Fig. 5. The input of the module is combined with the image data and the noise. The backbone of CenterNet [38] is adopted in the detection and classification module. The input of the module is combined with the image data x and the noise $\upsilon^{(o)}$. Its structure is modified using assumed density filter (ADF) to propagate the output μ_j^l and uncertainties υ_j^l of unit j through the network during inference. The output of each detection box contains the coordinate of the bounding box and classification results generated by the ADF-based network. The coordinate of the bounding box contains the position of the center (u, υ), width, and height. The classification results contain the scores of each class and corresponding uncertainties. The details of the uncertainty estimation are described as follows.

1) Uncertainty Estimation: Uncertainty comprises aleatoric and epistemic uncertainties. Aleatoric uncertainties capture the inherent noise in observations, e.g., sensor noise, whereas epistemic uncertainties relate to the uncertainties in the model parameters, which can also be referred to as model uncertainties. Beluch *et al.* [39] argued that it is more effective to model aleatoric uncertainties because epistemic uncertainties can be explained if provided with sufficient data.

ADF is a general method for approximating posteriors in statistical models. By referring to the lightweight probabilistic deep networks proposed by Gast and Roth [40], each activation in the network was replaced by a probability distribution. Accordingly, the ADF-based forward propagation of the neural network eventually produces output predictions with uncertainties.

In order to apply the ADF, the joint density distribution should be factored into the product of some simple terms. The input is factored as $p(z^{(0)})$, and others are factored as conditional probability

$$p(z^{(0:l)}) = p(z^{(0)}) \prod_{i=1}^{l} p(z^{(i)}|z^{(i-1)}) \tag{8}$$

$$p(z^{(i)}|z^{(i-1)}) = \delta(z^{(i)} - f^{(i)} z^{(i-1)}) \tag{9}$$

where δ is the Dirac delta and $f^{(i)}$ is the ith network layer.

Because the joint density distribution is intractable, an approximated distribution should be introduced

$$p(z^{(0:l)}) \approx q(z^{(0:l)}) = p(z^{(0)}) \prod_{i=1}^{l} q(z^{(i)}). \tag{10}$$

The exponential family distribution is chosen to reduce the complexity. It is assumed that the input is corrupted by white Gaussian noise and the subsequent layer activations are subject to independent Gaussian distributions

$$q(z^{(i)}) = \prod_{j} \mathrm{N}\left(z_j^{(i)} | \mu_j^{(i)}, \nu_j^{(i)}\right) \tag{11}$$

where represent the activation output and variance of neural unit j, respectively.

Activation $z^{(i-1)}$ is processed by $f^{(i)}$ and transformed into the distribution

$$\tilde{p}(z^{(0:i)}) = p(z^{(i)}|z^{(i-1)}) \prod_{j=0}^{i-1} q(z^{(j)}). \tag{12}$$

Then, $\tilde{p}(z^{(0:i)})$ of each layer was approximated by the assumed exponential distribution $q(z^{(0:i)})$ by minimizing the Kullback–Leibler (KL) divergence

$$\arg\min_{\tilde{q}(z^{(0:i)})} \mathrm{KL}(\tilde{p}(z^{(0:i)}) || q(z^{(0:i)})). \tag{13}$$

Minka [41] showed that under the normality assumptions, (13) is equivalent to

$$\mu^{(i)} = \mathrm{E}_{q(z^{(i-1)})}[f^{(i)} z^{(i-1)}] \tag{14}$$

$$\nu^{(i)} = \mathrm{V}_{q(z^{(i-1)})}[f^{(i)} z^{(i-1)}] \tag{15}$$

where E and V are the first and second moments of the posterior distribution of each layer output, respectively. Most functions applied in neural networks, e.g., convolution deconvolution and rectified linear unit, can compute the solutions of (14) and (15) analytically.

2) BBA Modeling of Camera: The ADF enables generating not only output prediction classification $\mu^{(l)}$ but also the classification uncertainty $\nu^{(l)}$ in the forward pass of the neural network. Understandably, a high uncertainty implies a low confidence of the classification result. Confidence is defined as

$$\mathrm{conf}(C_i) = 1 - \frac{e^{\nu^{(l)}(C_i)}}{\sum_{i=1}^{k} e^{\nu^{(l)}(C_i)}}. \tag{16}$$

After the training of the model, the average precisions of different classes can be evaluated for the dataset, which are denoted as $\alpha_c, \alpha_b, \alpha_v, \alpha_t, \alpha_p,$ and α_m for the *car*, *bus*, *van*, *truck*, *pedestrian*, and *motorcycle* classes, respectively. The corresponding BBAs for camera detections are defined as in (17), which will be fused based on the enhanced evidence theory in Section IV-C

$$m_c(A) = \begin{cases} m_c(\{\text{car}\}) = \text{conf}(\text{car})\alpha_c / \Sigma \\ m_c(\{\text{bus}\}) = \text{conf}(\text{bus})\alpha_b / \Sigma \\ m_c(\{\text{van}\}) = \text{conf}(\text{van})\alpha_v / \Sigma \\ m_c(\{\text{truck}\}) = \text{conf}(\text{truck})\alpha_t / \Sigma \\ m_c(\{\text{pedestrian}\}) = \text{conf}(\text{pedestrian})\alpha_p / \Sigma \\ m_c(\{\text{motorcycle}\}) = \text{conf}(\text{motorcycle})\alpha_m / \Sigma \\ m_c(\{\text{pedestrian,motorcycle}\}) \\ \quad = ((1 - \text{conf}(\text{pedestrian}))\alpha_p \\ \qquad + (1 - \text{conf}(\text{motorcycle}))\alpha_m) / \Sigma \\ m_c(\{\text{bus, van, truck}\}) \\ \quad = ((1 - \text{conf}(\text{bus}))\alpha_b + (1 - \text{conf}(\text{van}))\alpha_v \\ \qquad + (1 - \text{conf}(\text{truck}))\alpha_t) / \Sigma \end{cases}$$

$$\Sigma = \sum \alpha_{C_i} + \text{conf}(\text{car}) \times \alpha_c. \tag{17}$$

3) Feature Extraction: The structure of the feature extraction module is shown in Fig. 5. Each bounding box is resized and input to the feature extraction module. After a series of convolutions, pooling, batch normalization, full connections, and other processing, the output of the module is a feature vector, which is used subsequently [as in (23)] to associate the objects in multiple scans.

C. Object Classification Based on Single-Scan Data Fusion

A spatial–temporal calibration should be conducted before fusion. The spatial relationship between the detection results of the MMW radar and the corresponding image was described by a perspective transformation. The transformation matrix was calibrated as described in [15]. A temporal calibration was performed by a trigger from the camera, allowing fusion with simultaneous MMW radar detections. Both sources (the radar and the camera) provide a series of detections. The global nearest neighbor matching algorithm was used to associate the detections of the radar and the camera. Subsequently, the BBAs for the associated radar and camera detections were fused.

This study focused on preprocessing the evidence body to deal with the fusion of conflicting data. The input of single scan fusion is the BBAs of radar and camera as in (7) and (17). It is noted that weighting the evidence by the relationship between evidence is not applicable to the single-scan fusion with two evidence bodies. The weighted coefficient is determined by the effect of evidence itself. The detailed processing is shown in Fig. 6.

First, information volume IV_s based on the BE is calculated to measure the relative importance of each evidence m_i as

$$\mathrm{IV}_i = e^{E_d} = e^{-\sum_s m_i(A_s) \log \frac{m_i(A_s)}{2^{|A_s|-1}}}, \quad 1 \le i \le k. \tag{18}$$

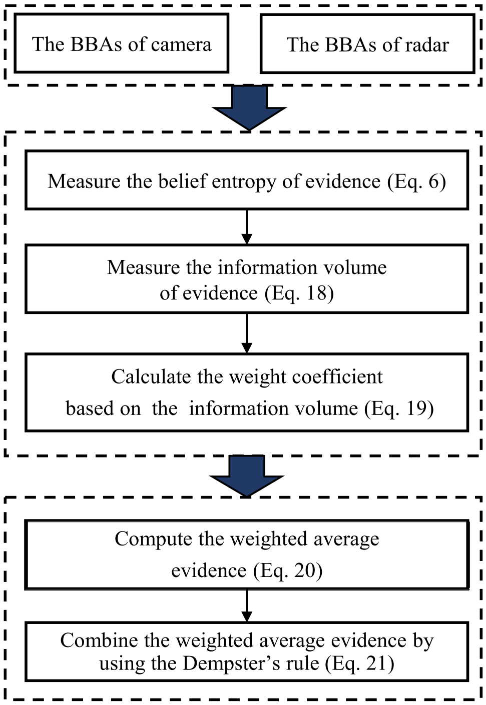

The BBAs of camera
The BBAs of radar
Measure the belief entropy of evidence (Eq. 6)
Measure the information volume of evidence (Eq. 18)
Calculate the weight coefficient based on the information volume (Eq. 19)
Compute the weighted average evidence (Eq. 20)
Combine the weighted average evidence by using the Dempster's rule (Eq. 21)

Fig. 6. Flowchart of the single-scan fusion.

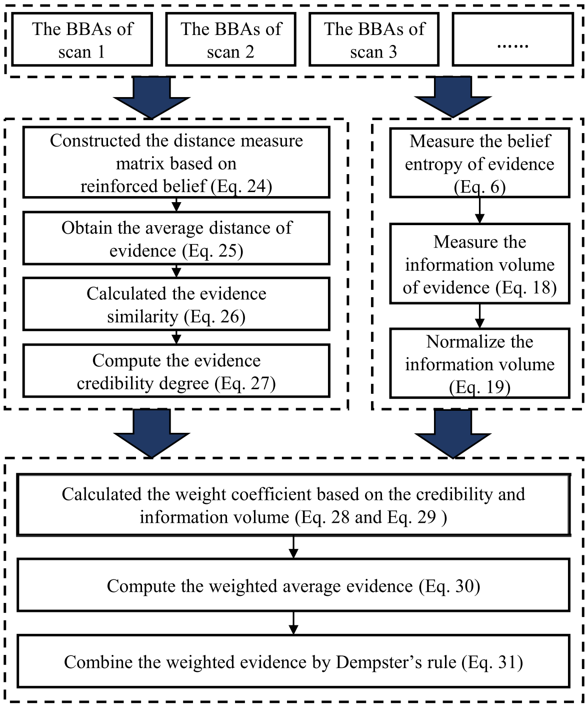

The BBAs of scan 1
The BBAs of scan 2
The BBAs of scan 3
......
Constructed the distance measure matrix based on reinforced belief (Eq. 24)
Obtain the average distance of evidence (Eq. 25)
Calculated the evidence similarity (Eq. 26)
Compute the evidence credibility degree (Eq. 27)
Measure the belief entropy of evidence (Eq. 6)
Measure the information volume of evidence (Eq. 18)
Normalize the information volume (Eq. 19)
Calculated the weight coefficient based on the credibility and information volume (Eq. 28 and Eq. 29)
Compute the weighted average evidence (Eq. 30)
Combine the weighted evidence by Dempster's rule (Eq. 31)

Fig. 7. Flowchart of the multiscan fusion.

The information volume of the evidence is normalized as

$$\bar{\mathrm{IV}}_i = \frac{\mathrm{IV}_i}{\sum_{s=1}^{k} \mathrm{IV}_s}. \tag{19}$$

The weighted evidence is calculated by the weighted coefficient $\bar{\mathrm{IV}}_i$ as

$$m_{\text{weight}} = \bar{\mathrm{IV}}_r \times m_r + \bar{\mathrm{IV}}_c \times m_c. \tag{20}$$

Finally, the weighted average evidence, m_{weight}, is fused using Dempster's combination rule. The BBA with the maximum value is taken as the classification result of the object

$$m_k(C) = (m_{\text{weight}} \oplus m_{\text{weight}}). \tag{21}$$

D. Object Classification Based on Multiple Scans Data Fusion

In practice, sensors suffer from malfunctions due to various reasons, which make it unreliable for object detection solely based on single-scan data. Thus, we proposed fusing multiple scans data acquired by the camera and the radar, in order to improve the detection accuracy and reliability. The key is to associate the detections of adjacent scans. Aeberhard and Bertram [18] calculated the classification similarity of two objects in the data association process at the fusion level. Motivated by it, an association approach for the multiscan detections of the camera and the radar was proposed.

The association process is conducted in an 8-D state space $(u, v, \lambda, h, \dot{u}, \dot{v}, \dot{\lambda}, \dot{h})$ that contains bounding box center position $(u,\ v)$, aspect ratio λ, height h, and their derivatives. Kalman filter with a constant velocity motion and linear observation model is used to predict the locations of the detections in the subsequent scan.

The Euclidean distance is calculated to determine the detected position and predicted position of a detection, as denoted by $\mathbf{p}_{i,d}^{uv}$ and $\mathbf{p}_{j,p}^{uv}$, respectively, as follows:

$$d_{i\to j}^2 = \left(\mathbf{p}_{i,d}^{uv} - \mathbf{p}_{j,p}^{uv}\right)^T \left(\mathbf{p}_{i,d}^{uv} - \mathbf{p}_{j,p}^{uv}\right). \tag{22}$$

The cosine distance relates to the appearance information, which is particularly useful for reidentifying an obstructed object in dense traffic conditions. The appearance descriptor is obtained using a pretrained CNN model, as mentioned in Section IV-B. The cosine distance is used to measure the similarity between feature descriptor $\mathbf{F}_{i,d}$ of the detection bounding box and feature descriptor $\mathbf{F}_{j,p}$ of the predicted bounding box using the following:

$$\mathrm{sim}_{i\to j} = 1 - \frac{||\mathbf{F}_{i,d}||_2||\mathbf{F}_{i,p}||_2 - \mathbf{F}_{i,d}\mathbf{F}_{i,p}}{||\mathbf{F}_{i,d}||_2||\mathbf{F}_{i,p}||_2}. \tag{23}$$

The Euclidean distance will be set as infinite when the corresponding similarity is more than ϖ. Subsequently, the Hungarian method is employed to associate the detections of multiple scans. The associated detection results of multiple scans are treated as the different evidence bodies.

The evidence body will be weighted by the coefficient to fuse the conflicting data. In addition to the information volume considering the effect of evidence itself, the distance between evidence is calculated to obtain the coefficient based on the RB divergence. The weighted coefficient will help obtain more reasonable BBAs by considering the relationships between belief functions and subsets of the sets of belief functions, as well as the effect of the evidence itself on the weight. The detail process is shown in Fig. 7.

Step 1: The distance between k evidences can be obtained based on the RB divergence measure, and a distance measurement matrix M can be constructed

$$M = \begin{bmatrix} 0 & \cdots & \Re\mathcal{B}_{1I} & \cdots & \Re\mathcal{B}_{1k} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ \Re\mathcal{B}_{i1} & \cdots & 0 & \cdots & \Re\mathcal{B}_{ik} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ \Re\mathcal{B}_{k1} & \cdots & \Re\mathcal{B}_{ki} & \cdots & 0. \end{bmatrix} \tag{24}$$

Step 2: The average distance of evidence m_i is calculated

$$\overline{\Re\mathcal{B}_i} = \frac{\sum_{j=1,j\neq i}^{k} \Re\mathcal{B}_{ij}}{k-1}. \tag{25}$$

Step 3: The evidence similarity is defined as

$$\mathrm{Sim}\frac{1}{\overline{\Re\mathcal{B}_i}}, \quad 1 \leq i \leq k. \tag{26}$$

Step 4: The credibility degree derived from the similarity is obtained as Crd_i in (27). The credibility represents the degree of conflict between the evidence. A high credibility implies a small degree of conflict, and a large weight should be assigned to the corresponding evidence

$$\mathrm{Crd}_i = \frac{\mathrm{Sim}(m_i)}{\sum_{s=1}^{k} \mathrm{Sim}(m_s)}, \quad 1 \leq i \leq k. \tag{27}$$

Step 5: The information volume $\mathrm{I}\bar{\mathrm{V}}_i$ is calculated using (18) and (19). The weighted coefficient is obtained based on $\mathrm{I}\bar{\mathrm{V}}_i$ and Crd_i. It is denoted as ACrd_i and normalized as $\mathrm{F}\bar{\mathrm{C}}\mathrm{rd}_i$

$$\mathrm{ACrd}_i = \mathrm{Crd}_i \times \mathrm{I}\bar{\mathrm{V}}_i, \quad 1 \leq i \leq k \tag{28}$$

$$\mathrm{F}\bar{\mathrm{C}}\mathrm{rd}_i = \frac{\mathrm{ACrd}_i}{\sum_{s=1}^{k} \mathrm{ACrd}_s}, \quad 1 \leq i \leq k. \tag{29}$$

Step 6: The body of the evidence is preprocessed by the weighted coefficient. Based on weight $\mathrm{F}\bar{\mathrm{C}}\mathrm{rd}_i$, the weighted average evidence is generated

$$m_{\text{weight}} = \sum_{i=1}^{k} (\mathrm{F}\bar{\mathrm{C}}\mathrm{rd}_i \times m_i), \quad 1 \leq i \leq k. \tag{30}$$

Step 7: The weighted average evidence, m_{weight}, is fused by $k-1$ times using Dempster's combination rule

$$m(C) = ((((m_{\text{weight}} \oplus m_{\text{weight}})_1 \oplus m_{\text{weight}})_2 \oplus \cdots) \oplus m_{\text{weight}})_{k-1}. \tag{31}$$

V. Experiments

This section presents the experiments conducted to evaluate the performance of the proposed method.

To verify the radar–vision fusion approach for objection classification, challenging scenes captured under extreme light conditions, i.e., bright light and low illumination, were selected. Both datasets were manually tagged to provide the ground truth reference. The details of the class distributions in the datasets are summarized in Table II. The images of the training set were used to train the CNN model with both detection and feature extraction modules. The radar data were selected to examine the number of reflection points and investigate the RCS values of different objects.

TABLE I

Specifications of MMW Radar and Camera

Camera		MMW radar	
Version	Hikvision DS-2CD1201	Version	Continental ARS-408
Resolution	1280 × 720	Max Range	170 m
Fps	20	Modulation Type	FMCW
Focal	8mm	Frequency Bands	76-77GHz
HFOV	±55.2°	HFOV	±60° (0-10m) ±45°(>10m)

TABLE II

Class Distributions in Datasets

Class	Training set	Testing set
car	23415	5540
truck	4452	1394
bus	2256	968
van	4187	996
motorcycle	1546	588
pedestrian	3876	1178

TABLE III

BBAs of Different Evidences

BBA	{A}	{B}	{C}	{A, C}
m1	0.40	0.28	0.30	0.02
m2	0.01	0.90	0.08	0.01
m3	0.63	0.06	0.01	0.30
m4	0.60	0.09	0.01	0.30
m5	0.60	0.09	0.01	0.30

A. Numerical Experiments and Analyses

This section reports numerical experiments to examine the performance of the enhanced evidence theory when fusing conflicting and uncertain data. Experimental data in [28] were utilized to compare the proposed method with the state-of-the-art ones.

Suppose that there are three categories $\{A, B, C\}$ in the frame of discernment. Moreover, there are five sensors to detect the targets. The output of these sensors has been modeled as BBAs, defined as $m1$, $m2$, $m3$, $m4$, and $m5$, as shown in Table III. Sensor 1 outputs an uncertain result with a similar belief assignment for $\{A\}$, $\{B\}$, and $\{C\}$. An extremely high belief value (0.9) of Sensor 2 was assigned to $\{B\}$. The belief values of Sensors 3–5 were assigned to $\{A\}$ with relatively high confidence, which apparently conflict with Sensor 2.

Table IV presents the fusion results generated by different methods. It can be identified that Yager's method [27] obtained the lowest belief value for the actual class $\{A\}$, whereas the unknown event Ω was assigned the highest value. The essential reason is that Yager's method modified the combination rule by assigning the conflicting data to the unknown event Ω to avoid the counterintuitive results. Lin and Xie's [34] method made an improvement by introducing the compatibility coefficient

TABLE IV
FUSION RESULTS GENERATED BY DIFFERENT METHODS

Method	{A}	{B}	{C}	{A, C}	{Ω}
Yager [27]	0.0063	0.0001	0.0009	0	0.9911
Lin & Xie [34]	0.3369	0.1361	0.0230	0.0843	0.4196
Murphy [29]	0.9694	0.0175	0.0110	0.0021	/
Deng et al. [30]	0.9885	0.0013	0.0079	0.0023	/
Fan et al. [32]	0.9885	0.0013	0.0079	0.0023	/
Xiao [28]	0.9888	0.0015	0.0073	0.0024	/
Proposed	**0.9901**	0.0006	0.0068	0.0024	/

TABLE V
PERFORMANCE OF DIFFERENT METHODS FOR SINGLE-SCAN FUSION

Method	Bright light		Low illumination	
	Precision	Recall	Precision	Recall
Jiang et al. [15]	64.74%	66.09%	82.29%	80.91%
Bouain et al. [17]	65.47%	65.24%	82.87%	78.92%
Proposed	**67.90%**	**66.10%**	**84.89%**	**81.19%**

and conflict degree in the combination rule. The estimated belief value for class $\{A\}$ was higher than Yager's method; however, the unknown event Ω was still assigned a higher value. It failed to meet the practical application.

For the methods modifying the evidence body, Murphy's [29] method and Deng *et al.*'s [30] method recognized class $\{A\}$ with belief values of 0.9694 and 0.9885, respectively. It is worth mentioning that Fan *et al.*'s [32] method obtained the same result as Deng *et al.*'s. Because the grouping conditions based on BE and conflict coefficient in [32] were not satisfied, these two methods were identical with the same weighted coefficients. Xiao's [28] method obtained desirable results due to the RB divergence, which considers the correlations between belief functions and subsets of the sets of belief functions. By comparison, the proposed method additionally considers the effect of evidence itself, thus obtaining the highest belief value. The results help verify that the proposed method outperforms the state-of-the-art evidence theory-based ones when fusing conflicting and uncertain data.

B. Empirical Experiments and Analyses

The performance of the proposed method was further evaluated based on the empirical dataset. Both single-scan and three-scan experiments were conducted to verify the proposed method from two perspectives. First, the performance of the proposed evidential architecture was compared with other fusion architectures. Second, the effectiveness of integrating the enhanced evidence theory in the proposed architecture was demonstrated.

Bouain *et al.* [17] and Wang *et al.* [15] constructed fusion frameworks with different single-scan fusion methods. They were selected for comparison purpose in single-scan experiments. Table V summarizes the comparison results of different methods. It shows that the recalls obtained by Jiang *et al.*'s method in both bright light and low illumination conditions are close to the ones by our proposed method. It can be attributed to the fact that Jiang *et al.*'s method eliminates false alarms based on the ROI generated by the radar. However, the precisions are relatively lower because its fusion method ignores the region when the ROI has limited or no radar points in certain cases. Bouain *et al.*'s method obtained higher precisions, however, lower recalls compared to Jiang *et al.*'s methods. It is mainly because Bouain *et al.*'s method adopted Yager's combination rule to deal with the highly conflicting data. By comparison, the proposed method achieved desirable performance in terms of both precision and recall. It helps highlight the advantages of the proposed fusion architecture: 1) constructing BBAs based on the probability distributions of all categories derived by the uncertainty-aware CNN model and 2) preprocessing the evidence body instead of adopting the modified combination rule. To examine the effectiveness of integrating the enhanced evidence theory in the proposed architecture, the methods [28], [32] as investigated in the numerical experiments were selected for comparison. They were integrated with the proposed architecture. Note that the methods [28], [32] calculated the weighted coefficient based on the distance measurement between evidence, which is not applicable for single-scan fusion with two evidence bodies. Thus, three-scan fusion experiments were conducted for illustration purpose. The output of single-scan fusion would be the input of these methods. In addition, the effects of RB divergence and BE were scrutinized. Specifically, Xiao's [28] method calculated the weighted coefficient based on RB. BE was solely integrated with the proposed fusion architecture for performance evaluation.

TABLE VI
PERFORMANCE OF DIFFERENT METHODS FOR THREE-SCAN FUSION

Method	Bright light		Low illumination	
	Precision	Recall	Precision	Recall
Fan et al. [32]	68.94%	69.39%	86.35%	82.53%
Xiao [28]	70.88%	70.17%	86.06%	85.19%
BE	68.06%	69.10%	85.69%	83.27%
RB+BE	**71.06%**	**72.35%**	**87.23%**	**86.09%**

Note: BE: belief entropy-based, RB: reinforced belief divergence

The results are shown in Table VI. It can be found that the proposed method (RB + BE) obtained better results of precision and recall than Fan *et al.*'s [32] method, Xiao's [28] method, and BE. Moreover, Xiao's method (RB) performs better than BE. It implies that the RB divergence contributes more to the improvement of precision and recall than the BE. This is because the RB divergence considers the relationships among different evidences by the divergence measurement of both evidence and the subsets of the sets of the belief functions. In the case of multiple scans, the relationship between evidence has a more critical impact on the final weight. Concurrently, the BE is important because it considers the information volume of evidence itself.

Tables VII and VIII provide the detailed comparison results of precision and recall for each class. The improvement by the single-scan fusion for the *truck*, *bus*, and *van* is insignificant

TABLE VII
COMPARISON RESULTS OF PRECISION FOR EACH CLASS

		Car	Truck	Bus	Van	Motorcycle	Pedestrian
Bright light	Camera-based [38]	77.10%	47.08%	44.63%	51.53%	74.59%	91.21%
	Single-scan fusion	78.32%	47.86%	56.20%	56.12%	77.35%	91.53%
	Three-scan fusion	79.85%	51.95%	58.68%	56.63%	94.46%	84.81%
		Car	Truck	Bus	Van	Motorcycle	Pedestrian
Low illumination	Camera-based [38]	90.68%	68.41%	78.79%	83.77%	79.79%	90.27%
	Single-scan fusion	93.49%	70.57%	80.72%	84.77%	82.45%	97.35%
	Three-scan fusion	96.03%	71.59%	82.92%	87.75%	86.88%	98.23%

TABLE VIII
COMPARISON RESULTS OF RECALL FOR EACH CLASS

		Car	Truck	Bus	Van	Motorcycle	Pedestrian
Bright light	Camera-based [38]	85.59%	56.54%	36.73%	51.01%	69.95%	87.77%
	Single-scan fusion	92.43%	56.42%	35.98%	50.23%	76.92%	84.64%
	Three-scan fusion	96.18%	65.76%	44.94%	56.06%	73.31%	91.63%
		Car	Truck	Bus	Van	Motorcycle	Pedestrian
Low illumination	Camera-based [38]	93.11%	68.80%	81.02%	75.75%	68.46%	89.46%
	Single-scan fusion	97.92%	70.81%	80.38%	74.53%	70.51%	93.00%
	Three-scan fusion	98.56%	82.89%	88.01%	75.28%	77.35%	94.41%

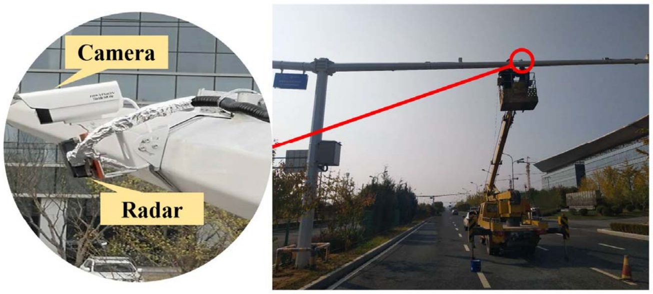


Fig. 8. Roadside camera and radar deployment.

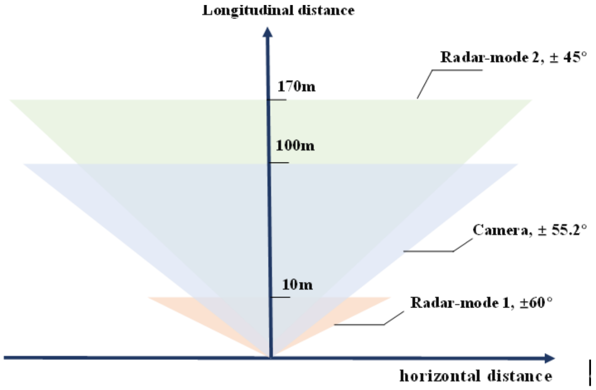
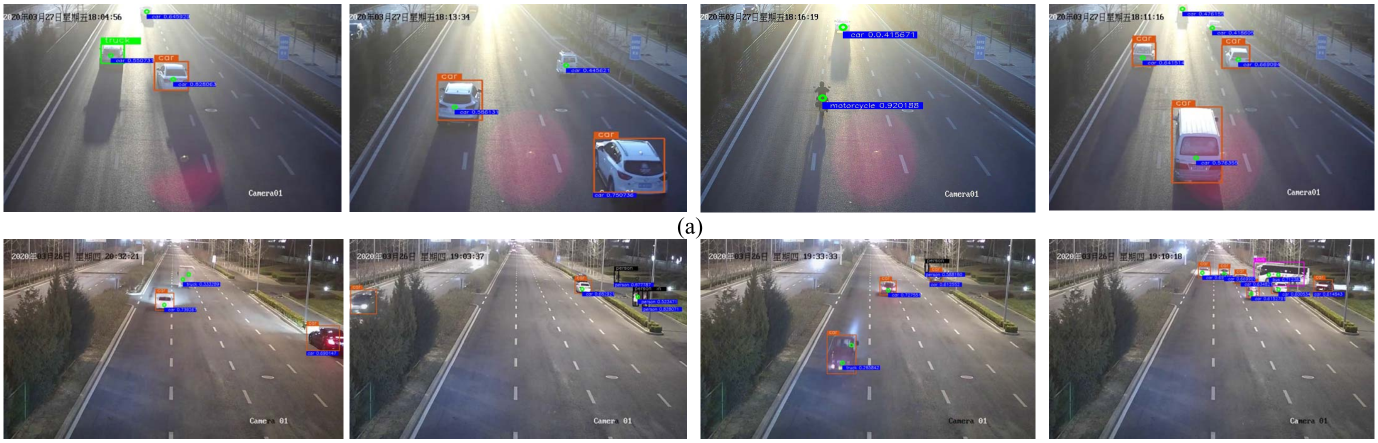


Fig. 9. Detection ranges of radar and camera.

as the three-scan fusion. This is because the radar cannot offer sufficient features to classify these categories. In such a case, the classification performance can only be improved by obtaining new evidence. Different from the precision, the recalls of different categories present discrepant tendencies. For the *car* class, the improvement in the recall is more significant than for the other categories. One essential reason is that the number of reflection points, RCS, and speed information detected by the radar are complementary to the image for object classification. In detail, the *car* class can be distinguished from the *motorcycle* and *person* classes based on the RCS and the speed and from the *truck*, *bus*, and *van* classes based on the number of reflection points. Thus, the most of the false detection for *car* are eliminated. Conversely, for the *truck*, *bus*, and *van* classes, the recall of the single-scan fusion is lower than that of the camera-based method in most scenarios. It can be explained that the objects that are wrongly classified as other categories (*car*, *motorcycle*, and *pedestrian*) by the camera-based method are reidentified as these three categories. However, the radar cannot provide sufficient evidence to distinguish between the three categories, resulting in false detections among the three categories. The recall improves until new data are obtained. The results of the *motorcycle* and *pedestrian* classes are similar. Overall, the improvement in the precision by the single-scan fusion method is more prominent than the decrease in the recall. In addition, it can be found that the performance of the fusion method is closely related to the precision of camera detection. The bright light has a significant influence on the image, and its precision and recall are lower than those in low illumination conditions. As a result, the fusion method performs better in low illumination conditions.

Fig. 10 shows several examples of the output classification results obtained by the proposed three-scan fusion. Note that the bounding box represents the output of camera detection,



(a)

(b)

Fig. 10. Outputs of radar–vision fusion under extreme light conditions. (a) Bright light conditions. (b) Low illumination conditions.

and the label at the top left side of the box indicates the classification result by the camera-based method. The green point represents the radar detection, which is converted to the image plane via the transformation. The classification result by the fusion-based method is presented in the blue rectangle, with the corresponding evidence confidence. As can be seen from Fig. 10, there exist inconsistencies in the classification results by the camera- and fusion-based methods. The radar–vision fusion approach helps achieve more reliable classifications of moving objects.

VI. Conclusion

In this study, a radar–vision fusion approach based on the enhanced evidence theory was developed for object classification in roadside challenging scenes. First, the BBAs of radar and camera were modeled to map the detection result to BBAs. Subsequently, single-scan fusion and multiscan fusion algorithms were devised based on the enhanced evidence theory. Both numerical and empirical experiments were conducted for verification purpose. The results of numerical experiments show that the proposed method obtained the highest belief value of 99.01%. The results of empirical experiments based on real roadside data show that the proposed method achieved 71.06% and 87.23% precisions for bright light and low illumination conditions, respectively. It should be noted that the detection performance for the *truck*, *bus*, and *van* classes remains undesirable. In practice, the radar cannot effectively distinguish these three categories owing to the sparsity of the radar points. In future work, a LiDAR point cloud is supposed to be fused with image and radar information in order to obtain denser expressions of object characteristics, thus making the detection more robust.

References

[1] E. Marti, M. A. de Miguel, F. Garcia, and J. Perez, "A review of sensor technologies for perception in automated driving," *IEEE Intell. Transp. Syst. Mag.*, vol. 11, no. 4, pp. 94–108, Sep. 2019.

[2] S. A. Ahmed *et al.*, "Query-based video synopsis for intelligent traffic monitoring applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3457–3468, Aug. 2020.

[3] H. T. Nguyen, S.-W. Jung, and C. S. Won, "Order-preserving condensation of moving objects in surveillance videos," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 9, pp. 2408–2418, Sep. 2016.

[4] S. R. E. Datondji, Y. Dupuis, P. Subirats, and P. Vasseur, "A survey of vision-based traffic monitoring of road intersections," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 10, pp. 2681–2698, Oct. 2016.

[5] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, Jul. 2018.

[6] J. Baek, J. Kim, and E. Kim, "Fast and efficient pedestrian detection via the cascade implementation of an additive kernel support vector machine," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 902–916, Apr. 2017.

[7] C. Liu, R. Fujishiro, L. Christopher, and J. Zheng, "Vehicle–bicyclist dynamic position extracted from naturalistic driving videos," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 734–742, Apr. 2017.

[8] E. Hyun, Y. S. Jin, and J. H. Lee, "Design and development of automotive blind spot detection radar system based on ROI pre-processing scheme," *Int. J. Automot. Technol.*, vol. 18, no. 1, pp. 165–177, Feb. 2017.

[9] G. Li, S. E. Li, R. Zou, Y. Liao, and B. Cheng, "Detection of road traffic participants using cost-effective arrayed ultrasonic sensors in low-speed traffic situations," *Mech. Syst. Signal Proc.*, vol. 132, pp. 535–545, Oct. 2019.

[10] Z. Zhang, J. Zheng, H. Xu, X. Wang, X. Fan, and R. Chen, "Automatic background construction and object detection based on roadside LiDAR," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4086–4097, Oct. 2020.

[11] J. Wu, H. Xu, and J. Zheng, "Automatic background filtering and lane identification with roadside LiDAR data," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2017, pp. 1–6.

[12] J. Wu, H. Xu, and J. Zhao, "Automatic lane identification using the roadside LiDAR sensors," *IEEE Intell. Transp. Syst. Mag.*, vol. 12, no. 1, pp. 25–34, Sep. 2020.

[13] J. Zhao, H. Xu, H. Liu, J. Wu, Y. Zheng, and D. Wu, "Detection and tracking of pedestrians and vehicles using roadside LiDAR sensors," *Transp. Res. C, Emerg. Technol.*, vol. 100, pp. 68–87, Mar. 2019.

[14] B. Li *et al.*, "Enhancing 3-D LiDAR point clouds with event-based camera," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[15] X. Wang, L. Xu, H. Sun, J. Xin, and N. Zheng, "On-road vehicle detection and tracking using MMW radar and monovision fusion," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 2075–2084, Jul. 2016.

[16] R. O. Chavez-Garcia and O. Aycard, "Multiple sensor fusion and classification for moving object detection and tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 525–534, Feb. 2016.

[17] M. Bouain, D. Berdjag, N. Fakhfakh, and R. B. Atitallah, "Multi-sensor fusion for obstacle detection and recognition: A belief-based approach," in *Proc. 21st Int. Conf. Inf. Fusion (FUSION)*, Jul. 2018, pp. 1217–1224.

[18] M. Aeberhard and T. Bertram, "Object classification in a high-level sensor data fusion architecture for advanced driver assistance systems," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 416–422.

[19] L. A. Zadeh, “A simple view of the Dempster–Shafer theory of evidence and its implication for the rule of combination,” *AI Mag.*, vol. 2, no. 7, p. 85, 1986.

[20] A. Roy, N. Gale, and L. Hong, “Automated traffic surveillance using fusion of Doppler radar and video information,” *Math. Comput. Model.*, vol. 54, nos. 1–2, pp. 531–543, 2011.

[21] Y. Fu *et al.*, “A Camera–Radar fusion method based on edge computing,” in *Proc. IEEE Int. Conf. Edge Comput. (EDGE)*, Oct. 2020, pp. 9–14.

[22] J. Bai, S. Li, H. Zhang, L. Huang, and P. Wang, “Robust target detection and tracking algorithm based on roadside radar and camera,” *Sensors*, vol. 21, no. 4, p. 1116, Feb. 2021.

[23] S. Lu, Z. Bao, Y. Zhi, and S. Zhang, “Target detection algorithm based on MMW radar and camera fusion,” in *Proc. 2nd Int. Conf. Comput. Data Sci.*, Jan. 2021, pp. 1–6.

[24] B. Chen, X. Pei, and Z. Chen, “Research on target detection based on distributed track fusion for intelligent vehicles,” *Sensors*, vol. 20, no. 1, p. 56, Dec. 2019.

[25] H. Cho, Y.-W. Seo, B. V. K. V. Kumar, and R. R. Rajkumar, “A multi-sensor fusion system for moving object detection and tracking in urban driving environments,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 1836–1843.

[26] P. Smets, “The combination of evidence in the transferable belief model,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 5, pp. 447–458, May 1990.

[27] R. R. Yager, “On the Dempster–Shafer framework and new combination rules,” *Inf. Sci.*, vol. 41, no. 2, pp. 93–137, Mar. 1987.

[28] F. Xiao, “A new divergence measure for belief functions in D–S evidence theory for multisensor data fusion,” *Inf. Sci.*, vol. 514, pp. 462–483, Apr. 2020.

[29] C. K. Murphy, “Combining belief functions when evidence conflicts,” *Decis. Support Syst.*, vol. 29, no. 1, pp. 1–9, Jul. 2000.

[30] Y. Deng, W. Shi, Z. Zhu, and Q. Liu, “Combining belief functions based on distance of evidence,” *Decis. Support Syst.*, vol. 38, pp. 489–493, Dec. 2004.

[31] G. Awogbami, N. Agana, S. Nazmi, X. Yan, and A. Homaifar, “An evidence theory based multi sensor data fusion for multiclass classification,” in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2018, pp. 1755–1760.

[32] X. Fan, Y. Guo, Y. Ju, J. Bao, and W. Lyu, “Multisensor fusion method based on the belief entropy and DS evidence theory,” *J. Sensors*, vol. 2020, pp. 1–16, Jan. 2020.

[33] N. Khan and S. Anwar, “Improved Dempster–Shafer sensor fusion using distance function and evidence weighted penalty: Application in object detection,” in *Proc. 16th Int. Conf. Informat. Control, Autom. Robot.*, Jan. 2019, pp. 664–671.

[34] Z. Lin and J. Xie, “Research on improved evidence theory based on multi-sensor information fusion,” *Sci. Rep.*, vol. 11, no. 1, pp. 1–6, Dec. 2021.

[35] A. P. Dempster, “Upper and lower probabilities induced by a multivalued mapping,” *Ann. Math. Statist.*, vol. 38, no. 2, pp. 325–339, Apr. 1967.

[36] G. Shafer, “A mathematical theory of evidence,” *Technometrics*, vol. 1, no. 20, p. 106, Feb. 1978.

[37] Y. Deng, “Deng entropy,” *Chaos, Solitons Fractals*, vol. 91, pp. 549–553, Oct. 2016.

[38] X. Zhou, D. Wang, and P. Krähenbühl, “Objects as points,” 2019, *arXiv:1904.07850*.

[39] W. H. Beluch, T. Genewein, A. Nurnberger, and J. M. Kohler, “The power of ensembles for active learning in image classification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9368–9377.

[40] J. Gast and S. Roth, “Lightweight probabilistic deep networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3369–3378.

[41] T. P. Minka, “A family of algorithms for approximate Bayesian inference,” Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2001.

Pengfei Liu received the B.E. degree in transportation engineering from Beihang University, Beijing, China, in 2018, where she is currently pursuing the Ph.D. degree with the School of Transportation Science and Engineering.

Her research interests include computer vision and data fusion.

Guizhen Yu received the Ph.D. degree from Jilin University, Changchun, China, in 2003.

He is currently a Professor with the School of Transportation Science and Engineering, Beihang University, Beijing, China, the Hefei Innovation Research Institute, Beihang University, Hefei, China. His research interests include intelligent vehicle, urban traffic operation, and intelligent transportation systems.

Zhangyu Wang received the Ph.D. degree in transportation information engineering and control from Beihang University, Beijing, China, in 2021.

He is currently an Assistant Professor with the Research Institute for Frontier Science, Beihang University, and the Hefei Innovation Research Institute, Beihang University, Hefei, China. His research interests include intelligent vehicle and computer vision.

Bin Zhou received the Ph.D. degree from Beihang University, Beijing, China, in 2018.

He is currently an Assistant Professor with the School of Transportation Science and Engineering, Beihang University, and the Hefei Innovation Research Institute, Beihang University, Hefei, China. His research interests include deep-learning algorithms, intelligent connected vehicle, big data analysis, and traffic network modeling.

Peng Chen (Member, IEEE) received the Ph.D. degree from Nagoya University, Nagoya, Japan, in 2012.

He is currently an Associate Professor with the School of Transportation Science and Engineering, Beihang University, Beijing, China. His research interests include intelligent transportation systems; urban traffic operation and control; and traffic flow modeling and simulation.

Dr. Chen is a member of the IEEE Intelligent Transportation Systems Society.