Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 4592124, 14 pages
https://doi.org/10.1155/2021/4592124

Research Article

Traffic Flow Parameters Collection under Variable Illumination Based on Data Fusion

Shaojie Jin,[1] **Ying Gao**,[1] **Shoucai Jing**,[1] **Fei Hui**,[1] **Xiangmo Zhao**,[1] **and Jianzhen Liu**[2]

[1]*School of Information and Engineering, Chang'an University, Xi'an, Shaanxi, China*
[2]*Jiaoke Transport Conultants LTD, Beijing, China*

Correspondence should be addressed to Fei Hui; feihui@chd.edu.cn

Received 1 June 2021; Revised 16 July 2021; Accepted 27 July 2021; Published 16 August 2021

Academic Editor: Xinqiang Chen

Accurate traffic flow parameters are the supporting data for analyzing traffic flow characteristics. Vehicle detection using traffic surveillance pictures is a typical method for gathering traffic flow characteristics in urban traffic scenes. In complicated lighting conditions at night, however, neither classical nor deep-learning-based image processing algorithms can provide adequate detection results. This study proposes a fusion technique combining millimeter-wave radar data with image data to compensate for the lack of image-based vehicle detection under complicated lighting to complete all-day parameters collection. The proposed method is based on an object detector named CenterNet. Taking this network as the cornerstone, we fused millimeter-wave radar data into it to improve the robustness of vehicle detection and reduce the time-consuming postcalculation of traffic flow parameters collection. We collected a new dataset to train the proposed method, which consists of 1000 natural daytime images and 1000 simulated nighttime images with a total of 23094 vehicles counted, where the simulated nighttime images are generated by a style translator named CycleGAN to reduce labeling workload. Another four datasets of 2400 images containing 20161 vehicles were collected to test the proposed method. The experimental results show that the method proposed has good adaptability and robustness at natural daytime and nighttime scenes.

1. Introduction

Traffic flow parameters are the most fundamental and critical physical factors that characterize traffic flow characteristics. In recent years, an increasing number of traffic video surveillance systems along the metropolitan curbside have been deployed to collect traffic flow parameters. Vehicle detection based on video images can achieve satisfactory detection results under normal daylight conditions. However, the precision of this method is reduced at nighttime due to the lack of natural light. Furthermore, streetlights and other light sources such as automobile lights are interlaced and complicated, resulting in an uneven image brightness distribution, poor image visibility and contrast, and a lack of required details and context. As a result, this study offers a millimeter-wave radar and camera fusion sensing approach for achieving improved detection results throughout the day and collecting more precise traffic flow parameters.

Generally speaking, data fusion can be categorized as original-data level, feature level, and decision level, based on the multiple abstraction levels of input data. Data layer fusion offers the advantage of retaining source domain information to the highest degree possible. It does, however, have stringent data transmission requirements, limited real-time efficiency, and significant processing costs. Although decision level fusion can successfully reduce transmission capacity, it does so at the expense of some of the original information. Feature layer fusion can fuse features extracted from multiple sensors while keeping data integrity and processing complexities.

Microwave radar, LiDAR, and cameras are the most popular sensors used to detect long-term and short-term traffic. Each sensor has its benefits and drawbacks. Although a camera is one of the most commonly utilized sensors, its detection capability deteriorates at night and in inclement weather [1]. The distance is calculated by measuring the time

between the laser pulse and its reflection and scattering by the target in a LiDAR system. It has a wide range and angular resolutions, but poor weather degrades its performance. The receiving antenna of a millimeter-wave radar radiates electromagnetic waves, which are collected by the receiving antenna, and the target information is acquired by processing a sequence of signals. A millimeter-wave radar has a longer wavelength than a camera or a LiDAR, allowing it to operate in adverse weather situations such as rain, snow, and fog [2].

The integration of data from multiple sensors has been demonstrated to increase a system's ability to perceive the outside world. Advanced Driving Assistance System (ADAS) [3] incorporates this technology. In numerous research articles, such as [4–8], millimeter-wave radar was coupled with a camera for vehicle detection. A few systems [9, 10] combined LiDAR and camera. In [11], the authors combined three sensors. While the combination of camera and LiDAR can produce outstanding detection results, it adds to the computational complexity, and both sensors' detection performance will be harmed in bad weather. As a result, for fusion detection in this study, we use millimeter-wave radar and a camera. Despite the fact that there has been a lot of study on sensor fusion, the most of it has been focused on vehicle onboard sensor data. The two techniques are not fully consistent in vehicle detection algorithms due to the difference in viewpoint between vehicles and roadside.

For fusion detection, this research utilizes millimeter-wave radar and a camera positioned on the roadside to increase the accuracy of traffic flow information collecting throughout the day. To unify the representation of the two sensors, the first step in accomplishing fusion is to map the observed points in millimeter-wave radar to pixel coordinates. Furthermore, the radar detection point must be linked to the appropriate item. In this work, we present a data association technique based on 3D areas of interest for this purpose. The modified Deep Layer Aggregation (DLA) [12] is used as the backbone to extract the picture characteristics, with CenterNet [13] as the baseline. Using the CycleGAN to transfer style and generate synthetic nighttime pictures from daytime photos, the annotation effort is reduced during training. Finally, the image and radar characteristics are combined to produce additional object properties.

Using the millimeter-wave radar and camera positioned along the roadside for the whole day, a new dataset is collected to validate the proposed technique. We manually label each vehicle at daytime and fake nighttime images to train the proposed model. Comparative tests are conducted using two standard image processing approaches and a deep learning method. The findings demonstrate that when compared to the three approaches previously stated, the proposed method can accomplish superior detection in various scenarios.

In this paper, the system structure of our proposed CenterNet-based framework has been described, and its feasibility and effectiveness are tested and analyzed. The rest of the paper is organized as follows: a literature review is illustrated in Section 2. Data fusion alongside traffic flow parameters collection methodology is elaborated in Section 3. Section 4 presents the experimental settings and results and then is followed by a conclusion with some future work in Section 5.

2. Related Work

The research state of existing work is examined in this part for the four categories of computer vision-based vehicle identification during the day and night, millimeter-wave radar and camera fusion method, and traffic parameters collecting.

2.1. Computer Vision-Based Vehicle Detection. Traffic videos play a crucial role in monitoring the current situation of roads and provide an efficient means of collecting traffic flow parameters on urban roads. Thanks to the rapid progress of deep learning in image processing, picture-based object recognition has seen encouraging outcomes in recent years. A contemporary detector typically comprises of two parts: (1) a backbone that extracts characteristics from pictures, and (2) a head that detects the object's class and bounding box. Furthermore, newly developed detectors place a layer called the neck between the backbone and the head. The neck gathers features from different scales and fuses them and fully uses all the features extracted from the backbone to achieve better detection performance.

(i) Backbone: VGG [14], ResNet [15], and ResNeXt [16] can serve as the backbone of detectors operating on a GPU platform. MobileNet [17] or ShuffleNet [18, 19] can be the backbone for detectors operating on CPU platforms.

(ii) Head: The head part is mainly divided into one-stage and two-stage object detectors. The R-CNN [20] series, including the Fast R-CNN [21], Faster R-CNN [22], R-FCN [23], and Libra R-CNN [24], are the most representative two-stage object detectors. A two-stage object detector can also be converted into an anchor-free object detector, such as RepPoints [25]. As for the one-stage object detector, YOLO [26–29], SSD [30], and RetinaNet [31] are the most representative versions. In recent years, anchor-free detectors have been created for one-stage objects. CenterNet [32], CornerNet [33], FCOS [34], and so on are detectors of this kind.

(iii) Neck: A neck is usually composed of several bottom-up and top-down paths. Function Pyramid Network (FPN) [35], Path Aggregation Network (PAN) [36], BiFPN [37], and NAS-FPN [38] are networks fitted with this mechanism.

2.2. Computer Vision-Based Vehicle Detection at Nighttime. Most techniques, according to Chen et al. [39], cannot consistently monitor traffic conditions at night. To achieve vehicle detection, Kosaka and Ohashi [40] retrieved the brightness, geometric information, and color characteristics of the front or tail lighting and categorized them using Support Vector Machine (SVM). At night, however,

complicated light reflection may impair feature extraction. Inside the cars of interest, Vancea et al. [41] initially identified and searched for suitable taillight pairings. The authors then segmented taillights using deep learning. Different end-to-end deep learning frameworks for identifying items of interest were created by Ruimin [42] and Yu et al. [43]. Deep learning-based approaches, according to these researches, can minimize false positive and false negative detection errors and are more accurate in complicated real-world situations than standard image processing methods. These investigations, on the other hand, are reliant on image processing technology and have difficulty providing accurate monitoring in a complicated lighting environment. Therefore, multisensor detection can be used to compensate for the uncertainty present in image detection.

2.3. Millimeter-Wave Radar and Camera Fusion Algorithm. In recent years, millimeter wave and camera fusion research has made tremendous progress. To decrease the image detection area, Kadow et al. [44] employed millimeter-wave radar, followed by the AdaBoost algorithm to detect road targets. Wang et al. [45] reported an experiment and calculations for millimeter-wave radar and vision point alignment for obstacle detection that were simple to execute. For millimeter-wave radar and monocular vision fusion, the authors developed a three-level fusion method based on visual attention mechanisms and the driver's visual consciousness. To integrate image and radar data, Nobis et al. [5] proposed the CRF-net architecture. Its design can automatically figure out what amount of sensor data fusion is best for detecting outcomes. Jiang [8] first defogged the image captured by the camera in foggy weather and filtered the effective target using millimeter-wave radar. The authors then mapped the result to the camera image to get the corresponding RoI region and finally used the weighted method to combine the two findings. To convert radar measurements to camera pictures and classify radar measurements using camera image datasets, Lekic and Babic [6] utilized Generative Adversarial Networks. Following that, the authors utilized radar to improve the camera's resilience and used camera to improve the radar precision.

2.4. Traffic Parameter Collection. As stated in [46], a variety of devices and techniques are frequently used to gather traffic flow characteristics. Loop detectors, video cameras, unmanned aerial vehicles (UAVs), radio frequency identification (RFID) detectors, Bluetooth, GPS devices on vehicles, float cars, light detection and ranging sensors, and other devices are common examples. The aforementioned equipment and technologies may extract a number of traffic flow characteristics, including speed, density, and quantity. The high-precision traffic flow parameters gathering technique based on video [42] is being applied with the rapid advancement of computer vision. However, because of the difference in light between day and night, vehicle identification accuracy would be significantly reduced if a daytime detection model was simply applied to nighttime. As a result, in a night environment, vehicle recognition and exact traffic flow data collecting are required. In this paper, the traffic flow parameters are collected by integrating millimeter-wave radar and camera data.

3. Methodology

In this section, we present a traffic flow parameters collection method based on millimeter-wave radar and camera fusion. We first construct a fusion detection framework and then unify the representation and alignment acquisition times of the two sensors. Afterward, a 3D region-of-interest-based target association method is proposed, and the fusion detection is completed. Finally, the traffic flow parameters collection method is presented.

3.1. Framework. The metrics of traffic flow are gathered via sensor fusion. In sensor fusion, the primary goal should be to unify the object representations of two sensors. After temporal alignment, the millimeter-wave radar detection data should be converted to pixel coordinates. It is worth noting that the radar return points used in this experiment are the object preliminary detection results instead of the radar point cloud. By directly obtaining the radar detection results, the information bandwidth can be minimized, and the anti-interference ability is more robust. Subsequently, the 3D region of interest method is proposed to achieve one-to-one object association between the radar and the image. The pillar expansion is introduced during object association to address the inaccurate height information problem. Based on the above processing steps, the available radar features map can be generated. As for image processing, CenterNet is adopted as a baseline network that uses keypoint estimation to find the center point and then regresses to other attributes of the object. In addition, it avoids screening all possible bounding boxes of the object by modeling the target as the center point of the bounding box. The first image processing step is annotating original images under normal lighting during the daytime. Second, CycleGAN is applied to transfer the annotated images for simulating the nighttime lighting environment and reducing the annotation workload. Third, the DLA is adopted as the backbone to extract image features and concatenate the extracted image features with the generated radar features. Fourth, the head is used to regress other object attributes to complete vehicle detection. Last, the required traffic flow parameters can be collected. Figure 1 shows the pipeline of the method proposed in this paper.

3.2. Data Calibration

3.2.1. Time Alignment. In order to accomplish temporal fusion, the millimeter radar and camera acquisition timings must be synchronized. The sample frequencies of the radar and camera, on the other hand, are not usually the same, but they are about 13 and 20 frames per second, respectively. The measurement time of millimeter-wave radar is used in this study to make camera data backward compatible, and radar and camera time registration is used to choose the picture

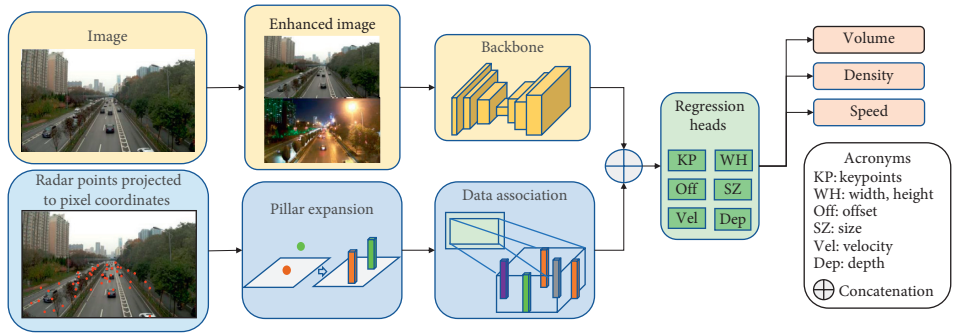


Figure 1: Pipeline of the proposed method. For image processing, the CycleGAN is used first to generate fake nighttime images based on annotated daytime images. Then, the features of images are extracted by the DLA backbone. For radar processing, the 3D region of interest is utilized for associating the radar detection points with the corresponding object after pillar expansion. Last, the image features are concatenated with the radar features and regressed to other object attributes.

that is closest to the radar acquisition time. Because the data collecting area's traffic speed restriction is 70 km/h, this study limits the sampling time difference between the two sensors within 20 ms to avoid errors. The associated pair of data points will be discarded, if the sample time between the nearest camera and the radar exceeds 20 ms.

3.2.2. Millimeter Wave Radar Coordinate Projection. Camera calibration is usually a time-consuming procedure that involves multiple variables. The final calibration result will include additional errors due to the complexity of coordinate translation using the traditional approach. Therefore, we adopt a calibration method similar to [45].

The target position perceived by the millimeter-wave radar and the pixel position of the target are represented as (x_r, y_r) and (u, v), respectively. The relationship between the two is as follows:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = T_I^R \begin{bmatrix} x_r \\ x_r \\ 1 \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{bmatrix} \begin{bmatrix} x_r \\ y_r \\ 1 \end{bmatrix}, \tag{1}$$

where T_I^R is a 3×3 transformation matrix. The above expression can be used to directly convert radar coordinates into pixel coordinates. The transformation matrix can be calculated based on the following calculations:

$$\begin{cases} T_i = \begin{bmatrix} t_{i1} & t_{i2} & t_{i3} \end{bmatrix}', \\ U = \begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix}', \\ V = \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix}', \\ V = \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix}', \end{cases} \tag{2}$$

$$I_{n\times 1} = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}',$$

$$P = \begin{bmatrix} x_r^1 & y_r^1 & 1 \\ \vdots & \vdots & \vdots \\ x_r^n & y_r^n & 1 \end{bmatrix},$$

where n is the number of aligned points, and (x_r^j, y_r^j) with $j = 1, 2, \dots, n (n \geq 4)$ represents the position of the aligned point in the radar coordinate system. The transformation matrix $T_I^R = \begin{bmatrix} T_1' & T_2' & T_3' \end{bmatrix}'$ is obtained using the linear least squares (LS) method as follows:

$$\begin{cases} T_1 = \left(P^T P\right)^{-1} P^T U, \\ T_2 = \left(P^T P\right)^{-1} P^T V, \\ T_3 = \left(P^T P\right)^{-1} P^T I_{n\times 1}. \end{cases} \tag{3}$$

Using the above expression, the conversion can be completed, and the mapping point of the radar detection in the pixel coordinate can be obtained. Ten groups of data with each group containing 20 pairs of points are selected to obtain accurate T_I^R. These groups are used to calculate different matrices and subsequently calculate the average value.

3.3. Center Point Detection. For preliminary image detection, we utilize the CenterNet [12] detection network, which was suggested in 2019. This network was chosen since the radar's return result is usually a point. For anticipating the object's center point, the CenterNet employs keypoint detection. The usage of CenterNet facilitates the fusion of data from these two sensors. The step of removing multiple overlapping prediction bounding boxes using Non-Maximum Suppression (NMS) can be avoided by simplifying the object to a single point.

Let $I \in R^{W\times H\times 3}$ be an image input to the CenterNet, which generates a keypoint heatmap $\hat{Y} \in [0,1]^{W/R\times H/R\times C}$, where W and H are the width and height of the image, respectively, R is the output stride, and C is the number of object types. The prediction from an input image is indicated by $\hat{Y}_{x,y,c}$, where $\hat{Y}_{x,y,c} = 1$ corresponds to a detected keypoint; otherwise, it corresponds to the background. For each ground truth keypoint $p \in R^2$ of class c in the input image, downsampling is used to obtain a low-resolution equivalent $\tilde{p} = \lfloor P/R \rfloor$. A ground truth keypoint heatmap $Y \in [0,1]^{W/R\times H/R\times C}$ is generated using the Gaussian kernel $Y_{xyc} = \exp\left(-\left((x - \check{p}_x)^2 + (y - \check{p}_y)^2\right)/2\sigma_p^2\right)$, where σ_P is an object size-adaptive standard deviation. The maximum value of keypoint heatmap Y is taken if two Gaussian distributions of the same class overlap.

A fully convolutional encoder-decoder network is used to predict $\widehat{Y}$ from the input image. Given the annotated objects $p_0, p_1 \ldots$ in an image, the training objective based on the focal loss is defined as follows:

$$L_k = \frac{-1}{N} \sum_{xyc} \begin{cases} \left(1 - \widehat{Y}_{xyc}\right)^{\alpha} \log\left(\widehat{Y}_{xyc}\right), & \text{if } Y_{xyc} = 1, \\ \left(1 - Y_{xyc}\right)^{\beta} \left(\widehat{Y}_{xyc}\right)^{\alpha} \log\left(1 - \widehat{Y}_{xyc}\right), & \text{otherwise.} \end{cases} \tag{4}$$

The image features are extracted by a fully convolutional encoder-decoder backbone network during the process of center point extraction based on an image. We follow CenterNet and adopt the modified Deep Layer Aggregation (DLA) network as the backbone. Then, the extracted image features are used to predict the center points and finally return to other object attributes, as well as the object's 2D size, i.e., width and height, and center offset.

3.4. Style Transfer from Daytime to Nighttime. Generally speaking, for object detection and recognition, it is necessary to label datasets, and these labeling workloads are often labor intensive. In this study, our data collection equipment is fixed in one place, and the shooting angle of the image data and the resolution of the image data are all the same. The different places mainly focus on the characteristics of same objects reflected in the image data in different time period (for example, daytime and nighttime). Therefore, we choose a cycleGAN of the GAN network to convert the labeled daytime dataset into a nighttime dataset to reduce the workload of manual annotation.

The aim of the work presented in this subsection is to achieve image-to-image conversion between the daytime source domain S and the nighttime target domain T. In this conversion, a training set of aligned image pairs is generally used to learn the mapping between the input and the output images. However, as paired training data will not be available for many tasks, this paper introduces a method proposed by Zhu et al. [47] in 2017 that learns to translate images from source domain X to target domain Y.

The goal of this method is to learn a mapping $G: X \longrightarrow Y$. It introduces the adversarial loss to differentiate the data distribution of the actual Y domain image from the data distribution of the image $G(X)$ converted from the X domain, which provides the foundation for creating the antagonistic network. An inverse mapping $F: Y \longrightarrow X$ is added, because this mapping is highly underconstrained, and cycle consistency loss is introduced to strengthen $F(G(X)) \approx X$. This means that after the image of X domain is mapped to Y domain and then mapped back to X domain, it should be consistent with the original image as much as possible, and vice versa. The Y domain image mapped to the X domain should also be consistent with the original image.

The total loss function in the style transfer architecture is defined as

$$\mathscr{L}\left(G, F, D_X, D_Y\right) = \mathscr{L}_{\text{GAN}}\left(G, D_Y, X, Y\right) + \mathscr{L}_{\text{GAN}}\left(F, D_X, Y, X\right) + \lambda \mathscr{L}_{\text{cyc}}\left(G, F\right), \tag{5}$$

where λ, $\mathscr{L}_{\text{cyc}}$ and $\mathscr{L}_{\text{GAN}}$ are the balanced weight, cycle consistency loss in the cycle architecture, and the adversarial training loss, respectively. The cycle consistency loss is used to regularize the GAN training. The two losses are defined as follows:

$$\begin{aligned} \mathscr{L}_{\text{GAN}}\left(G, D_Y, X, Y\right) &= E_{y \sim p_{\text{data}}(y)}\left[\log D_Y(y)\right] + E_{x \sim p_{\text{data}}(x)}\left[\log\left(1 - D_Y(G(X))\right)\right], \\ \mathscr{L}_{\text{cyc}}\left(G, F\right) &= E_{x \sim p_{\text{data}}(x)}\left[\left|\left|F(G(x)) - x\right|\right|_1\right] + E_{y \sim p_{\text{data}}(y)}\left[\left|\left|G(F(y)) - y\right|\right|_1\right]. \end{aligned} \tag{6}$$

The following expression should be solved to train these generators and discriminators:

$$G^*, F^* = \arg \min_{G,F} \max_{D_x, D_Y} \mathscr{L}\left(G, F, D_X, D_Y\right). \tag{7}$$

To solve equation (7) during training, the network parameters of the two generators are updated alternately, and the ADAM optimization algorithm is used in two discriminators.

3.5. Fusion Methodology. The center point detection network utilizes the object's center feature to regress to other characteristics of the object, as shown in Section 3.1. Instead of using the original radar point cloud, this article uses object detection findings, which minimizes the amount of data processing and therefore reduces information transmission bandwidth while improving anti-interference capabilities. The description of radar point cloud processing is beyond the scope of this paper, and it will not be elaborated here.

In order to make full use of radar data during the target detection process, we first need to associate the radar detection results with the objects on the image. Data association via 2D bounding boxes is not a reliable approach. When there is occlusion between items, for example, it is difficult to differentiate them using simply 2D plane information. Inspired by the frustum-based association method that is proposed by Nabati and Qi [7] and considering the difference of collection perspective, we propose the 3D region of interest (RoI) to complete the associated task. These RoIs are generated from the 2D ground truth of

the object and the estimated depth δ, as shown in Figure 2, where δ is defined as follows:

$$\delta = \frac{ld}{h - l}, \tag{8}$$

where d is the distance measured by the millimeter-wave radar, h is the height of the equipment, and l is the height of ground truth. They are all pixel values. The conversion factor c between real-word distance and pixel value is defined as follows:

$$c = \frac{1000 \times l}{l_w}, \tag{9}$$

where l_w and l are real-world and image-world lengths in millimeters and pixels, respectively. The value of c only needs to be calculated once in advance. The corresponding values of l_w and l can be obtained when the size of an image remains unchanged.

To achieve more accurate and rapid data association, a processing step called pixel expansion is introduced, where each radar point is expanded to a pillar. The pillars create a better representation of the physical objects detected by the radar, as these detections are now associated with a dimension in the 3D space. Based on this new representation, the radar detection point is considered available when the whole or a portion of the pillar is within the 3D RoI.

Using these techniques, we can solve the problem of association overlap owing to occlusion between objects in an improved manner. The radar feature map is created using the available radar detection points and then combined with image features. The head is used to regress to other attributes of the objects.

3.6. Traffic Flow Parameter Collection. The three most important parameters that can describe the traffic are volume, speed, and density in the traffic flow theory. Their relationship is given by the following equation:

$$Q = V \times K, \tag{10}$$

where Q denotes the volume in the same direction, V denotes the velocity, and K denotes the density, and K is defined as vehicle counts in the designated road section. As shown in Figure 3, we only collect traffic parameters in the red bounding box due to the comprehensive consideration of obstacle occlusion and limited device perception field of view. The total length of the experimental area is 130 meters.

Detailed bounding boxes of vehicles in each frame can be obtained according to the object detection result based on the image. The density can be obtained by counting the vehicles in a certain segment as follows:

$$K = \frac{N}{L}, \tag{11}$$

where N and L are the number of vehicles in the road segment and length of the road segment in kilometers, respectively.

There are two methods for obtaining speed in the suggested technique in this article, which are obtained from millimeter-wave radar and camera. We utilize the speed obtained by the millimeter-wave radar as the object speed for objects that can be detected by millimeter-wave radar, but we use the speed measured by the camera if the millimeter-wave radar does not detect the object.

The radar system sends out a series of continuous frequency modulation millimeter waves and receives the reflected signal from the target. According to the modulation voltage law, the frequency of the transmitted wave varies with time. The modulation signal is usually a triangle wave signal. The reflected wave has the same form as the transmitted wave, but there is a temporal delay. The frequency of the intermediate frequency signal generated by the mixer is proportional to the frequency difference between the transmitted signal and the reflected signal at a given instant, and the target distance is proportional to the intermediate frequency output by the front end. The reflected signal contains a Doppler shift induced by the relative movement of the target when it originates from a relatively moving object. The target distance and relative speed of the target may be determined using the Doppler principle. Millimeter-wave radar is one of the tools used in this article that can directly output object speed.

For camera speed measurement, we use the same target's pixel transformation between each frame for speed calculation. We assume $\overrightarrow{d_{(i,x)}}$ and $\overrightarrow{d_{(i,y)}}$ as the pixel changes of the i-th target in the horizontal and vertical directions. The overall motion magnitude d_i of the i-th motion vector in pixels/frame can be calculated by the following equation:

$$d_p = \sqrt{\overrightarrow{d_{(i,x)}}^2 + \overrightarrow{d_{(i,y)}}^2}, \tag{12}$$

where d_p donates the pixel transformation value of the target between two frames. It can be converted into a unit of measurement of length in the real world using equation (9).

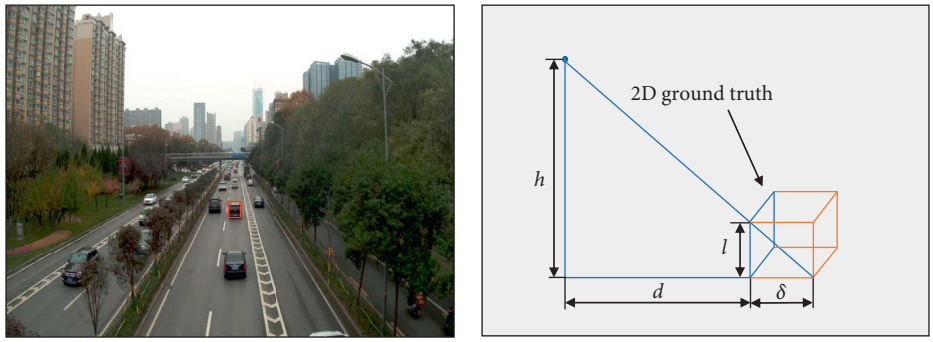
The speed refers to the space mean speed (SMS) that is the average speed of all vehicles driving within a certain length of road at a certain moment. It is used to evaluate the service level of the road and is defined as follows:

$$V = \frac{1}{1/n\left(\sum_{i=1}^{n} 1/v_i\right)}, \tag{13}$$

where n denotes the number of times traveled over the length of the road segment, and v_i is the travel speed of the i-th vehicle that is provided by the millimeter-wave radar. The volume can be calculated using equation (10) once the density and speed are available.

4. Experiments

In this section, the traffic dataset and experimental parameter settings are first used to test the proposed approach. Subsequently, the object detection performance is compared on two traditional algorithms and a deep learning method. Finally, we present and discuss the experimental results.



(a) (b)

Figure 2: 3D RoI association. (a) Annotated ground truth. (b) 3D region generated using the estimated value δ. The orange area is the estimated RoI.

Figure 3: Selected area of experiment in this paper.

4.1. Dataset. The data is collected using self-developed equipment in this study. To accomplish dual-sensor sensing, this device combines millimeter-wave radar with a camera. The equipment is mounted on a tripod on the pedestrian bridge, which is 6.73 meters above the ground. As illustrated in Figure 4, we collect data throughout the day. We record from all angles and are committed to the restoration of the data acquisition process.

The dataset of this paper is comprised of the millimeter wave-radar data at 4400 moments and the corresponding 4400 images. The image resolution is 2288×1712. The dataset is grouped into training and testing sets. The training set includes 2000 manually labeled traffic images in the daytime and simulated traffic images in the nighttime. There are 31708 vehicles in total. In this experiment, the unlabeled nighttime traffic images are defined as the Target Domain T, and the labeled daytime traffic images are defined as the Source Domain S. We use the CycleGAN described earlier to transfer the labeled daytime images and simulate the nighttime images in order to minimize the labeling workload. Each image in the testing set is manually labeled for performance evaluation purposes only. The testing set has data for four scenarios including 2400 images in total. It is divided into four subsets, each subset containing 600 pictures, which are divided into daytime traffic 1 (less vehicles), daytime traffic 2 (more vehicles), nighttime traffic 1 (more vehicles), and nighttime traffic 2 (less vehicles). Table 1 gives the details of our dataset. Figures 5 shows the sample pictures from four testing sets.

4.2. Experimental Settings. This experiment consists of two parts: vehicle detection and traffic flow parameters collection, which are both used to validate the accuracy of the proposed method. Two different methods are considered separately in the vehicle detection experiment:

(1) Method I: The images and manually labeled ground truth are used to train a CenterNet model on the training dataset. The learned model is then tested on Scenario I (less vehicles at daytime) and Scenario IV (less vehicles at nighttime).

(2) Method II: The proposed method fuses the data from camera and millimeter-wave radar. The images, points detected by millimeter-wave radar, and manually labeled ground truth are used to train the proposed model on the training dataset. The learned model is then tested on Scenario I (less vehicles at daytime) and Scenario IV (less vehicles at nighttime). The information detected by the millimeter-wave radar is used as a supplement to the image information, which can compensate for the performance loss based on image detection at nighttime. Besides, the trained CenterNet mentioned in method I is taken as the comparison methods.

In addition, two classical image processing algorithms for vehicle recognition based on background removal are used as comparative approaches. Multi-Layer background subtraction method (MultiLayer) [48] and Mixture of Gaussians algorithm based on Adaptive Gaussian Mixture Model (MOG) [49] are the two algorithms in question. In the tests, CenterNet [12] is used as a comparison technique. To reflect the sensor fusion effectiveness, all the images are used directly after labeling, without any self-defined preprocessing. The experiments are conducted

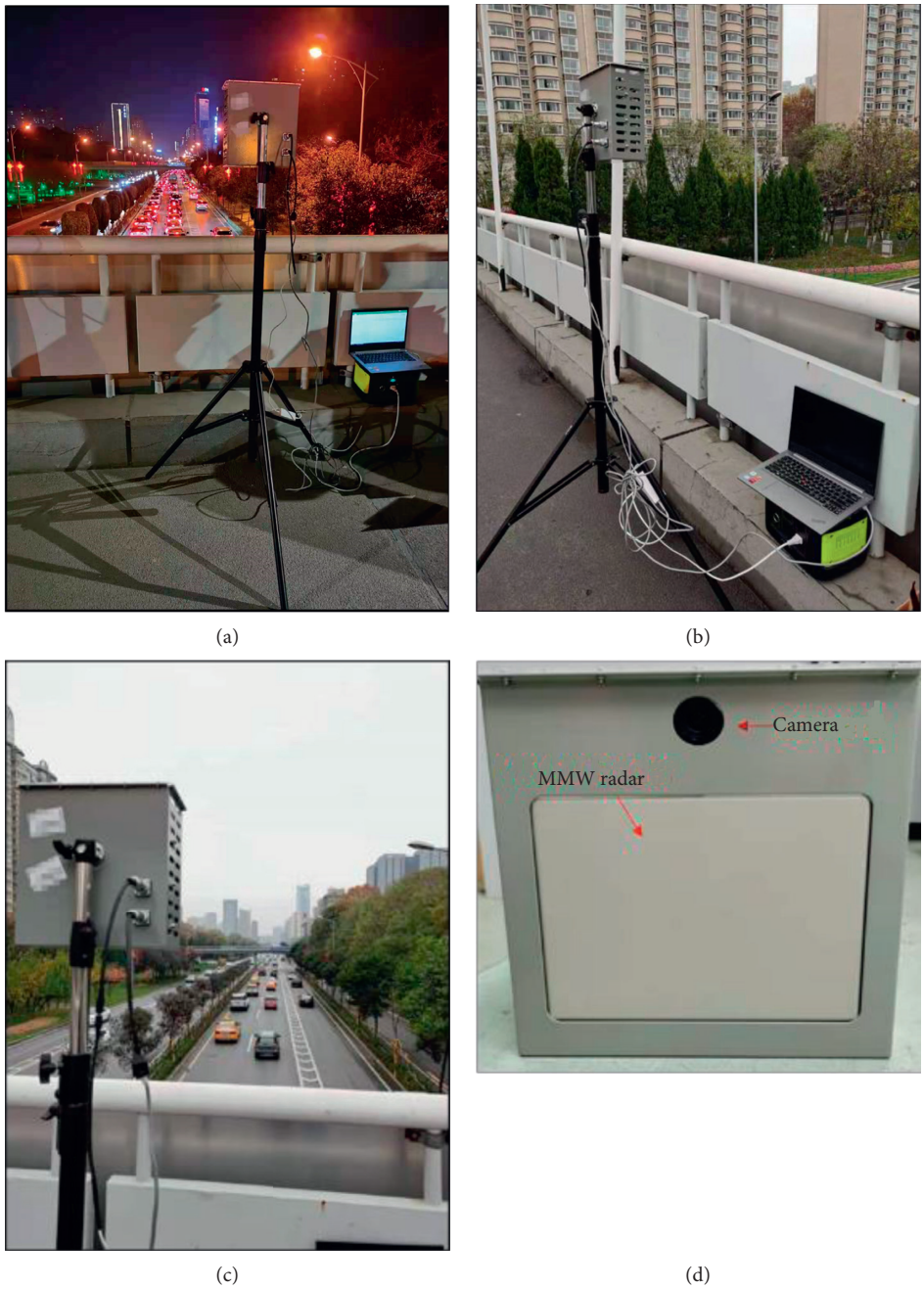


Figure 4: Taken during (a) nighttime and (b) daytime, respectively, (c) shooting the process of image acquisition from a closer perspective, and (d) composition diagram of the equipment.

using Python 3.6, PyTorch 1.1, and Cuda 10.1 in Windows 10 system. The batch size and training epoch are set as four images and 140 in the experiments. All these experiments are conducted on a workstation with a CPU of 2.6 GHz, and a NVIDIA GTX 2080 TI GPU with 12 GB memory. Six metrics are used to evaluate the detection performance of the aforementioned methods, including mean Average Precision (mAP), Precision, Recall, F-measure, Number of False Positives per image (N_{FP} error/image), and Number of False Negatives per image (N_{FN} error/image). The definitions of Precision, Recall, and F-measure are as follows:

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ F - \text{measure} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precison} + \text{Recall}}, \end{aligned} \tag{14}$$

where TP, FP, and FN refer to true positive, false positive, and false negative, respectively. The F-measure is an overall metric combining precision and recall; therefore, we use it to

Table 1: Details of the collected dataset in the experiments.

Training set	Image number	Vehicle number	Date
Daytime-training	1000	11547	11/20/2020
Fake nighttime-training	1000	11547	11/20/2020
Daytime-testing 1	600	5856	01/10/2020
Daytime-testing 2	600	7690	01/12/2020
Nighttime-testing 1	600	4023	01/10/2020
Nighttime-testing 2	600	2592	01/12/2020

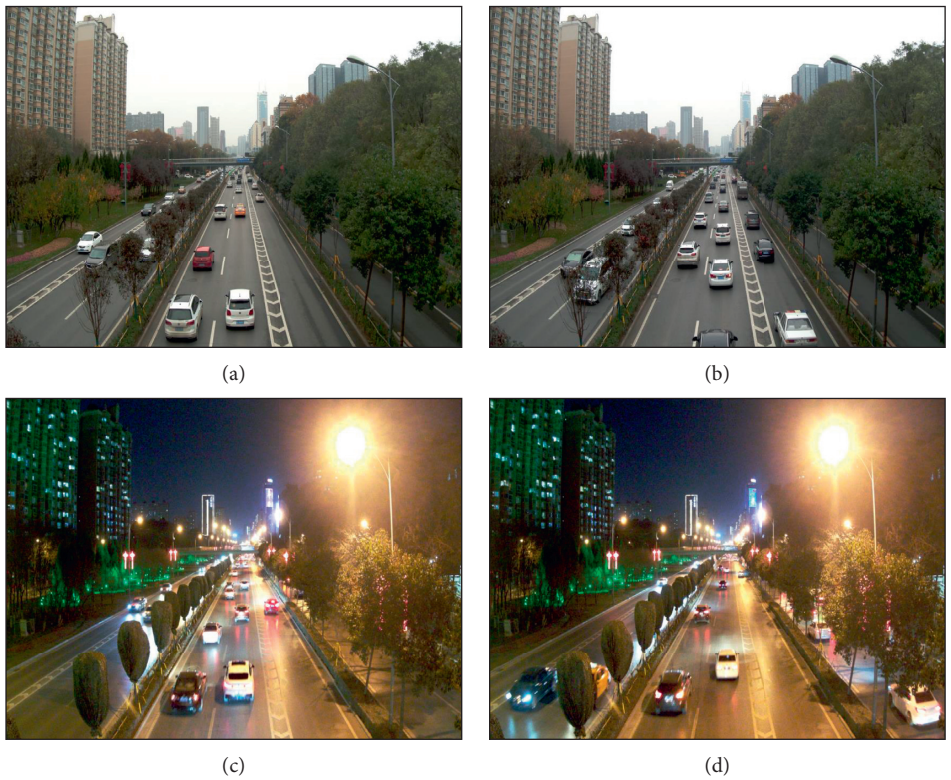


Figure 5: Sample pictures of four testing sets. (a) Daytime-testing 1. (b) Daytime-testing 2. (c) Nighttime-testing 1. (d) Nighttime-testing 2.

report the overall performance. The mAP (%) metric is the precision value averaged across all recall values between 0 and 1 for the vehicles, which is considered a comprehensive metric to effectively demonstrate the detection performance. For all the methods, the performance evaluation uses a uniform threshold of 0.5 for the Intersection over Union (IoU) between the predicted bounding box and the ground truth.

In the experiment on traffic flow parameters collection, the vehicle speed is provided by the millimeter-wave radar, and vehicle count is evaluated. Accuracy is used as a metric to evaluate the vehicle count. The mean absolute error (MAE) is taken as the metric to evaluate the count of vehicle. It is the average value of the absolute error, which can well reflect the actual situation of the predicted value error and can be calculated using the following equation:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|f_i - y_i\right|, \tag{15}$$

where f_i is the estimation value and y_i denotes the ground truth value. Smaller value of MAE indicates better performance.

4.3. Results. The improved vehicle detection performance can provide more accurate traffic flow parameters in practical applications. This study offers a technique based on the combination of millimeter-wave radar and camera to address the performance loss of vehicle detection at night using just traffic surveillance. The performance loss of image-based vehicle detection may be mitigated in the complicated illumination environment at night using millimeter-wave radar's all-weather operating features. We employ millimeter-wave radar to get object representations before performing the fusion in our detection solution. A uniform threshold of 0.5 is determined for the IoU between the predicted bounding boxes and ground truth, and the experimental results in this paper will be displayed in this section. The experimental results of vehicle detection of four

methods mentioned above will be compared in the performance evaluation in this paper.

The CycleGAN is used to achieve image-to-image conversion between the daytime source domain S and the nighttime target domain T. The result of a thousand rounds of transfer processing is stored, and the best result is displayed in Figure 6.

The findings of the complete detection performance evaluation are provided in Table 2. This table shows that, in terms of these six criteria, deep learning techniques to vehicle recognition outperform standard image processing methods. The detection performance of two classic image processing approaches, MOG and MultiLayer, is much worse when compared to deep-learning-based methods. Traditional image processing algorithms also perform worse at night than during the day in terms of detection. The possible explanation may be that, at night, the light spot formed by the lights of the vehicle on the ground will be recognized as a foreground object. Despite the fact that conventional approaches have a significantly lower recall than deep learning techniques, MultiLayer has outperformed another old method. On the average of four testing sets, MultiLayer receives 78.40% of accuracy and 71.29% of *F*-measure.

Deep learning performs better than the traditional method due to the strong discriminative feature extraction capabilities of the Convolutional Neural Network (CNN) framework. Out of all detection methods, our proposed method performs the best with the values of Precision and *F*-measure achieving 88.03% and 91.10%, respectively. followed by CenterNet, which achieves 86.17% and 89.97% on these two evaluation indicators. The proposed method and CenterNet are comparable and similar. Nevertheless, the proposed method has slightly better efficiency on our testing dataset.

Only the results based on MultiLayer, CenterNet, and the proposed approach for scenarios I and IV are given in Figure 7, which illustrates the visual results for vehicle recognition on actual daylight and nighttime pictures. During the night, the MultiLayer method displays several missed detections. CenterNet does a better job, but it still has a lot of false positives and false negatives. The proposed method gets the least false positive and false negative errors, which demonstrates improved vehicle detection in the nighttime. This improvement is achieved thanks to the proposed method's compensation of the image-based target detection by the millimeter-wave radar.

As CenterNet is the baseline model chosen for the experiments, we only compare it with the proposed method during traffic flow parameters collection. The performance of the two methods is similar during the daytime. The recall values of CenterNet at the daytime are 92.75% and 91.65%, respectively, while those of the proposed algorithm are 93.26% and 94.34%, respectively. Both methods have similar performance in the daytime. However, the proposed method improves the performance of collecting the number of vehicles at nighttime. Table 3 shows that, in the presence of vehicles at nighttime, the accuracy provided by proposed method increases by almost 8% and 5% in the two nighttime scenarios, respectively, compared with CenterNet.

As mentioned above, the testing set is divided into daytime traffic 1 (less vehicles), daytime traffic 2 (more vehicles), nighttime traffic 1 (more vehicles), and nighttime traffic 2 (less vehicles). It can be seen from Table 3 that, compared with CenterNet, the estimation accuracy of the proposed method of the counts of vehicles in the four scenarios is increased by 0.51%, 2.69%, 7.96%, and 3.98%, respectively. From this set of data, we can see that, in terms of the estimation accuracy of the counts of vehicles, the increase in the daytime is lower than that in the nighttime, and the increase in the sparse vehicle scene is lower than that in the dense vehicle scene. The possible reason is that the lighting environment at night is complicated, and other light sources such as street lights and car lights are interlaced and complicated, resulting in irregular image brightness distribution, poor image visibility and contrast, and lack of required details and context, resulting in a decrease in the accuracy of target detection based on image. The accuracy of vehicle counts estimation will also decrease. For the difference between the sparse vehicle scene and the dense vehicle scene, in the data collection method of this article, one possible explanation is that when the density of vehicles is high, the vehicles will overlap to a greater extent, which is not conducive to image-based target detection. For radar, radar relies on echoes for target judgment. When targets overlap, the loss of detection performance is smaller than that of the image.

The proposed method achieved a satisfactory performance in the vehicle count collection in daytime and nighttime as shown in Table 3. In this study, the deep learning model we trained did not use any nighttime manual labels as supervisions, and the great accuracy improvement during nighttime is quite promising.

Figure 8 shows the visual results of the estimated and ground-truth counts for the whole day traffic conditions. The position and internal parameters of our device are fixed during the data collection. It can be intuitively observed that the proposed method improves the vehicle counting accuracy. The experimental results show that the significant accuracy increases by the proposed method compared to CenterNet especially at nighttime.

Table 4 displays the vehicle speed estimation findings for the four scenarios in the collected dataset. We compared CenterNet and the proposed method with ground truth. During the daytime, the average MAE of the CenterNet is 1.63, and the average MAE during nighttime is 4.04. The proposed method has a MAE of 0.995 during the daytime and 1.295 during nighttime. Compared with CenterNet, the proposed method improves by 0.64 and 2.745 during the daytime and nighttime, respectively.

In addition, we were able to deduce the density and volume estimated by the proposed method in the selected road section in the collected dataset, which is shown in Table 5. It can be seen from Table 5 that when the number and density of vehicles are small, the vehicle speed is higher. The traffic flow is quite different during the day and night,

Figure 6: Image-to-image conversion between the daytime and nighttime by CycleGAN.

Table 2: Results of detection performance evaluation on the testing sets. On average of four testing subsets, the mean values of [Precision, *F*-measure] obtained by different methods are as follows: MOG [76.94%, 14.93%], MultiLayer [78.40%, 71.29%], CenterNet [86.17%, 89.97%], proposed method [88.03%, 91.10%].

Method	Precision	Recall	*F*-measure	N_{FP}/image	N_{FN}/image	mAP
			Daytime-testing 1			
MOG	83.27%	9.83%	16.76%	0.13	47.21	74.86%
MultiLayer	85.66%	78.29%	77.42%	6.59	15.62	79.93%
CenterNet	96.32%	92.75%	94.36%	0.95	2.79	97.57%
Proposed	97.37%	93.26%	93.21%	0.83	2.46	98.1%
			Daytime-testing 2			
MOG	86.18%	9.33%	17.14%	0.55	50.73	73.24%
MultiLayer	85.37%	69.86%	74.36%	7.48	17.23	87.4%
CenterNet	94.79%	91.65%	97.6%	0.58	2.49	98.2%
Proposed	96.3%	94.34%	96.83%	0.73	2.52	97.93%
			Nighttime-testing 1			
MOG	73.49%	7.28%	11.45%	1.90	43.69	58.42%
MultiLayer	74.28%	61.16%	62.13%	3.51	2.24	54.03%
CenterNet	77.21%	87.48%	83.93%	2.23	0.85	75.99%
Proposed	79.10%	95.44%	86.82%	1.89	0.75	88.39%
			Nighttime-testing 2			
MOG	64.81%	6.72%	14.36%	2.13	50.73	52.56%
MultiLayer	68.27%	64.62%	71.26%	2.87	2.62	51.78%
CenterNet	76.36%	89.14%	83.97%	1.81	1.10	75.37
Proposed	79.33%	93.12%	87.54%	1.34	1.82	86.27%

Figure 7: Detection results in Scenarios I and IV. The order from left to right is MultiLayer, CenterNet, and the proposed method.

Table 3: Results of vehicle counting accuracy during daytime and nighttime.

Method	Scenario 1 (%)	Scenario 2 (%)	Scenario 3 (%)	Scenario 4 (%)
CenterNet	92.75	91.65	87.48	89.14
Proposed	93.26	94.34	95.44	93.12

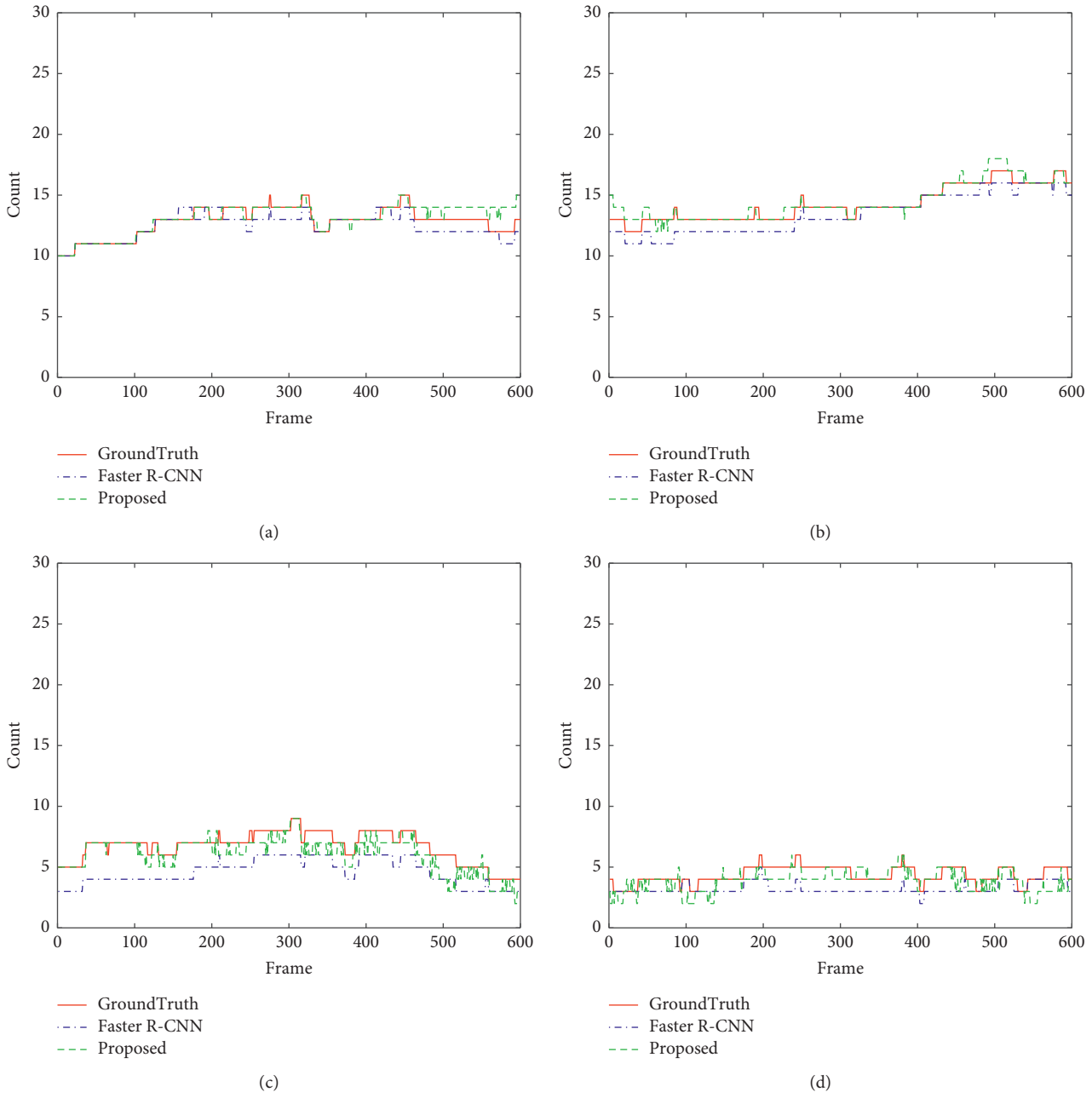


Figure 8: Daytime and nighttime vehicle count estimation in each scenario. (a) Daytime-testing 1. (b) Daytime-testing 2. (c) Nighttime-testing 1. (d) Nighttime-testing 2.

Table 4: Results in vehicle speed estimation of four scenarios in the collected dataset. MAE is the metric to evaluate the count of vehicle, which is mentioned in section 4.2.

Method	Metric	Scenario 1	Scenario 2	Scenario 3	Scenario 4
CenterNet	MAE	1.56	1.69	3.36	4.72
Proposed	MAE	0.96	1.03	1.36	1.23

Table 5: Summary of the estimated traffic flow parameters by the proposed method in four scenarios in the collected dataset. Pc donates passenger cars.

Traffic parameter	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Count (pc/frame)	13.6	15.2	7.2	5.1
Speed (km/h)	46.9	37.5	42.6	46.3
Density (pc/km)	104.6	116.9	55.4	39.2
Volume (pc/h)	4905.74	4383.75	2360.04	1826.72

showing a low-density high-speed overall at nighttime, and a high-density and vulgar flow during the daytime.

5. Conclusion

In this paper, a method for combining millimeter-wave radar and picture data is provided. The CenterNet serves as the baseline for our proposed approach, which is then augmented with millimeter-wave radar data. A set of data was obtained to train the suggested technique, and four more sets of data were collected to verify and analyze the method. The experimental results show that the proposed method improves the accuracy of vehicle detection and traffic flow parameters collection during the whole day.

In the future, we will concentrate on the following topics. First, various weather characteristics, such as rain, snow, and fog, will be considered during the traffic flow parameters collection. Second, based on the detection results, the next step of vehicle tracking and trajectory extraction will provide more data for road management and control. Last, using actual traffic flow parameters to maximize traffic efficiency in the face of mixed traffic flow will be considered in the coming era of autonomous vehicles.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was funded by National Key Research and Development Program of China (No. 2018YFB1600604) S&T Program of Hebei (NO. 20470801D).

References

[1] C. Michaelis, "Benchmarking robustness in object detection: autonomous driving when winter is coming," 2020, https://arxiv.org/abs/1907.07484.

[2] Z. Wang, Y. Wu, and Q. Niu, "Multi-sensor fusion in automated driving: a survey," *IEEE Access*, vol. 8, pp. 2847–2868, 2020.

[3] X. Jia, Z. Hu, and H. Guan, "A new multi-sensor platform for adaptive driving assistance system (ADAS)," in *Proceedings of the 2011 9th World Congress on Intelligent Control and Automation*, Taipei, Taiwan, June 2011.

[4] B. Yang, R. Guo, M. Liang, S. Casas, and R. Urtasun, "RadarNet: exploiting radar for robust perception of dynamic objects," 2020, https://arxiv.org/abs/2007.14366.

[5] M. G. Nobis, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," *Sensor Data Fusion: Trends, Solutions,Applications (SDF)*, vol. 99, pp. 1–7, 2019.

[6] V. Lekic and Z. Babic, "Automotive radar and camera fusion using Generative Adversarial Networks," *Computer Vision and Image Understanding*, vol. 184, pp. 1–8, 2019.

[7] R. Nabati and H. Qi, "CenterFusion: center-based radar and camera fusion for 3D object detection," 2020, https://arxiv.org/pdf/2011.04841.

[8] L. Jiang, "Target detection algorithm based on MMW radar and camera fusion," in *Proceedings of the 24th IEEE International Conference on Intelligent Transportation Systems*, Rhodes, Greece, May 2019.

[9] A. Asvadi, L. Garrote, C. Premebida, P. Peixoto, and U. Nunes, "Multimodal vehicle detection: fusing 3D-LIDAR and color camera data," *Pattern Recognition Letters*, vol. 115, pp. 20–29, 2018.

[10] T. E. Wu, C. C. Tsai, and J. I. Guo, "LiDAR/camera sensor fusion technology for pedestrian detection," 2017.

[11] R. O. Chavez-Garcia and O. Aycard, "Multiple sensor fusion and classification for moving object detection and tracking," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 525–534, 2016.

[12] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," 2019, https://arxiv.org/abs/1707.06484.

[13] X. Zhou, D. Wang, and P. Krahenb, "Objects as points," 2019, https://arxiv.org/abs/1904.07850.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition (CVPR)*, Las Vegas, NV, USA, July 2016.

[16] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," 2017, https://arxiv.org/abs/1611.05431.

[17] A. Howard, "Searching for MobileNetV3," in *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, South Korea, October 2019.

[18] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: practical guidelines for efficient CNN architecture design," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Glasgow, UK, August 2018.

[19] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: an extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, May 2018.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, June 2014.

[21] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 2015.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2016.

[23] J. Da, Y. Li, K. He, and J. Sun, "R-FCN: object detection via region-based fully convolutional networks," *Neural Information Processing Systems (NIPS)*, vol. 40, 2016.

[24] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: towards balanced learning for object detection," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, June 2019.

[25] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: point set representation for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Long Beach, CA, USA, June 2019.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, May 2016.

[27] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[28] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvemen," 2018, https://arxiv.org/abs/1804.02767.

[29] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: optimal speed and accuracy of object detection," 2020, https://arxiv.org/abs/2004.10934.

[30] W. Liu, "SSD: single shot MultiBox detector," 2016, https://arxiv.org/abs/1512.02325.

[31] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 318–327, 2018.

[32] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: keypoint triplets for object detection," 2019, https://arxiv.org/abs/1904.08189.

[33] H. Law and J. Deng, "CornerNet: detecting objects as paired keypoints," *International Journal of Computer Vision*, vol. 128, no. 3, pp. 642–656, 2019.

[34] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: fully convolutional one-stage object detection," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, Seoul, Korea, October 2019.

[35] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detectio," 2017, https://arxiv.org/abs/1612.03144.

[36] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, June 2018.

[37] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: scalable and efficient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, June 2020.

[38] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: learning scalable feature pyramid architecture for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, May 2019.

[39] Y.-L. Chen, B.-F. Wu, H.-Y. Huang, and C.-J. Fan, "A real-time vision system for nighttime vehicle detection and traffic surveillance," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 5, pp. 2030–2044, 2011.

[40] N. Kosaka and G. Ohashi, "Vision-based nighttime vehicle detection using CenSurE and SVM," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2599–2608, 2015.

[41] F. I. Vancea, A. D. Costea, and S. Nedevschi, "Vehicle taillight detection and tracking using deep learning and thresholding for candidate generation," 2017.

[42] D. Ruimin, "Real-time traffic flow parameter estimation from UAV video based on ensemble classifier and optical flow," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, pp. 54–64, 2019.

[43] H. Yu, D. Guo, Z. Yan, L. Fu, and S. Wang, "Weakly supervised easy-to-hard learning for object detection in image sequences," *Neurocomputing*, vol. 398, 2020.

[44] U. Kadow, G. Schneider, and A. Vukotich, "Radar-vision based vehicle recognition with evolutionary Optimized.pdf," *IEEE Intelligent Vehicles Symposium*, vol. 99, 2007.

[45] T. Wang, N. Zheng, J. Xin, and Z. Ma, "Integrating millimeter wave radar with a monocular vision sensor for on-road obstacle detection applications," *Sensors*, vol. 11, no. 9, pp. 8992–9008, 2011.

[46] B. Seo, "Traffic state estimation on highway: a comprehensive survey," *Annual Reviews in Control*, vol. 43, pp. 128–151, 2017.

[47] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*, IEEE, Venice, Italy, 2017.

[48] J. Yao and J.-M. Odobez, "Multi-layer background subtraction based on color and texture," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, MN, USA, June 2007.

[49] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, 2006.