Deep-Learning and Depth-Map Based Approach for Detection and 3-D Localization of Small Traffic Signs

Lirong Liu, Xinming Tang, Junfeng Xie, Xiaoming Gao, Wenji Zhao, Fan Mo, and Gang Zhang

***Abstract*—The three-dimensional (3-D) geographic locations of street furniture, such as traffic signs, comprise the basic content of 3-D city construction, and such information is indispensable for periodic statistics for road management and maintenance. This article presents a novel solution for acquiring 3-D information on small traffic signs based on mobile mapping system (MMS) data. First, a lightweight backbone network called VGG-L under an optimized faster region-based convolutional neural network detection framework is proposed for the detection of small traffic signs. An urban traffic sign detection (UTSD) dataset is created based on panoramic images obtained from test areas. Detection results from the UTSD dataset show that VGG-L outperforms other popular networks and achieves a mean average precision of 75.4%, which is 4.2%–14.8% higher than those of VGG16, MobileNet, ResNet, and YOLOv3. Second, a novel depth-map-based 3-D spatial geolocation method is proposed for obtaining the 3-D geographic locations of the objects. Then, a center-based method is proposed to automatically extract the final 3-D vector of the target. Experimental results illustrate that the proposed method performs 3-D positioning and vectorization of the milestones and circular and triangular traffic signs, accurately and effectively, achieving greater than 86% recall and precision for the three types of targets in the test areas. The experiments demonstrate that the overall 3-D information acquisition scheme is feasible and has great application potential.**

***Index Terms*—Convolution neural network (CNN), depth map, mobile mapping, three-dimensional (3-D) localization, traffic sign detection.**

I. Introduction

HIGH precision three-dimensional (3-D) information on street furniture plays an important role in road safety inspection, road facility management, maintenance, and 3-D city modeling [1]–[4]. As an important component of road facilities, traffic signs occur in multiple types of differing size and shape. Traditional methods based on manual statistics for collecting 3-D information on traffic signs are time consuming and labor intensive. Therefore, developing technology for automatic acquisition of 3-D information on traffic signs in order to meet the needs of regular monitoring and maintenance tasks is critical.

A mobile mapping system (MMS) is a multisensor system that integrates laser scanners, cameras, and a navigation system mounted on a vehicle. Such systems can acquire synchronized 3-D laser point clouds, images, and navigation data. An MMS can continuously and reliably scan the road surface and objects on both sides of a road, thus providing detailed elements of the urban model, such as building facades, road surfaces, and facilities [5]–[8]. Thus, an MMS is an effective tool for capturing 3-D road scene data for the automatic 3-D information acquisition of traffic signs.

Extensive studies regarding automatic object extraction from 3-D laser point cloud data of an MMS in order to identify road surfaces, curbs, street lamps, poles, trees, or buildings are currently on-going [9]–[14]. The 3-D extraction of large traffic signs based on laser point cloud data is possible, and some studies have achieved good results in terms of identifying such signs [15]–[20]. However, it remains challenging to automatically extract relatively small traffic signs, such as milestones and prohibitory and mandatory signs, based on the point cloud, due to the relative scarcity of point data on these small targets. Nevertheless, these small traffic signs are relatively easy to identify from the high-resolution panoramic images acquired simultaneously by the MMS. It is therefore possible to make full use of the images, point clouds, navigation data, and other multisource data acquired by an MMS, to explore a feasible solution for automatic extraction of small traffic signs. Therefore, we can adopt a two-step approach for obtaining 3-D information on small signs. The first step is to capture the targets' 2-D positions using image-based detection, and the second step is 3-D localization using auxiliary MMS data.

The most common task in traffic sign recognition consists of two main stages: detection and recognition [21]. This study focuses on the detection methods. Traffic sign detection is currently a well-studied and broad field of research. A number of approaches have been proposed for detecting traffic signs from images, which can be briefly divided into three categories according to [22]: color-based, shape-based, and CNN-based methods. Color-based methods often use a threshold to separate traffic signs from backgrounds [23]. Research works in [24] and [25] extracted color information of traffic signs using the

Manuscript received August 27, 2019; revised December 23, 2019; accepted January 10, 2020. Date of publication April 27, 2020; date of current version May 26, 2020. This work was supported in part by the National Key Research and Development Program under Grant 2017YFB0503004 and Grant 2017YFB0504201, in part by Active and passive composite mapping and application technology with visible, infrared and laser sensors under Grant D040106, in part by the National Natural Science Foundation of China under Grant 41971426 and Grant 41971425, and in part by the China Postdoctoral Science Foundation Funded Project under Grant 2019M650601. *(Corresponding author: Junfeng Xie.)*

Lirong Liu, Xinming Tang, Junfeng Xie, Xiaoming Gao, and Fan Mo are with the Land Satellite Remote Sensing Application Center, Ministry of Natural Resources, Beijing 100048, China (e-mail: liulirong1125@163.com; tangxinming99@qq.com; junfeng_xie@163.com; gaoxm@sasmac.cn; surveymofan@163.com).

Wenji Zhao is with the College of Resource Environment and Tourism, Capital Normal University, Beijing 100048, China (e-mail: zhwenji1215@163.com).

Gang Zhang is with the Chinese Academy of Surveying and Mapping, Beijing 100830, China (e-mail: 43131904@qq.com).

Digital Object Identifier 10.1109/JSTARS.2020.2966543

HSV color model. Some researchers use HSI color space or Eigen color space, based on Karhunen–Loève rather than RGB for better performance of traffic sign detection [26], [27]. Gao *et al.* [28] proposed a traffic sign detection method based on the extraction of red and blue color regions in the CIECAM97 color model. Yang *et al.* [29] proposed a convolutional pose machines model to detect signs followed by a support vector machine to filter out the background. Because of diverse natural lighting conditions, many heuristics have been applied to the treatment of color [30], [31]. The main disadvantage of these color-based methods is that it is difficult to set the threshold value, because of the complexity of the real-world environment. With different lighting conditions, the color information also changes, thus color-based methods have limited ability in dealing with variations in illumination.

Many shape-based methods, such as Hough transform [32], [33], corner detection, or radial symmetry voting [34]–[36], have also been widely used. For circular signs, a method using smoothness and Laplacian filters was proposed in [37]. For triangular signs, a method using gradient and orientation information was designed in [38]. Hoferlin and Zimmermann[39] introduced the Hough transform for detecting traffic signs. Generalized Hough transform can be used to detect circle, triangle, or rectangle shapes. Loy and Barnes [34] proposed a method that uses local radial symmetry to highlight the place of interest in each image, thereby detecting octagonal, square, and triangular traffic signs. Hough-like and Viola-Jones-like methods are the two paradigms of shape-based methods, and the Viola-Jones-like detectors compute a number of fast and robust features and try to identify trained patterns by different classifiers [40]. Most shape-based methods rely on gradient information, which is very sensitive to noise. Some studies combine both color and shape features for traffic sign detection. Fleyeh and Dougherty [41] used color segmentation to roughly locate the signs and then rule out false candidates using shape information. Greenhalgh and Mirmehdi [42] proposed a traffic sign detection algorithm using a novel application of maximally stable extremal regions (MSERs), and demonstrated that the MSERs are robust to variations in both lighting and contrast. However, these methods still have limited ability to deal with rotation, illumination, and scale variations.

Recently, deep learning methods have shown superior performance for many tasks, such as image classification, detection, and recognition. Since the introduction of the R-CNN by Girshick *et al.* [43] in 2014, the application of deep learning to object detection has attracted considerable interest. By various visual recognition challenges based on public datasets, such as ImageNet [44], [45], PASCAL VOC [46], or COCO [47], a number of high-precision and high-efficiency object detection algorithms have been proposed, such as faster R-CNN [48], single shot detector (SSD) [49], and the "You only look once" (YOLO) [50]–[52] series. The detection of small objects remains an open challenge in computer vision, due to the influences of image resolution, change in scale, and the context. Several advanced object-detection algorithms have been optimized to detect small objects. Faster R-CNN [48] introduces anchors of different scales and ratios as region proposals to accommodate multiscale detection. SSD [49] proposes a network working on feature maps of different layers to obtain a range of resolutions, but the bottom features of high resolution that are useful for detecting small objects were abandoned. The feature pyramid network [53], which combines multiple feature layers, is integrated into YOLOv3 [50] for better performance with small objects. With their strong robustness and increasing speed and accuracy, deep learning-based approaches have also been applied in traffic sign detection tasks. Aghdam *et al.* [54] proposed a lightweight and accurate ConvNet with a sliding window detector to detect traffic signs on German Traffic Sign Detection Benchmark (GTSDB). There are also some studies [55]–[60] focused on Chinese traffic sign detection based on CNN directly. Zhang *et al.* [61] provided an end-to-end method to detect Chinese traffic signs inspired by YOLOv2, which can be applied to a real-time system.

Panoramic images, such as those used in the present study, suffer greatly from object distortion, viewpoint variations, motion-blur, and illumination variations; in addition, the targets are very small. Although the abovementioned algorithms have yielded excellent performance in public competitions, these exiting models still require adjustments for practical applications, such as successfully identifying small traffic signs from panoramic images. Thus, this study makes appropriate adjustments to faster R-CNN and proposes a simple CNN, termed VGG-L, as the backbone network for improved detection of small traffic signs from panoramic images.

A considerable amount of research work has been devoted to benchmarking datasets of traffic signs, such as the German Traffic Sign Recognition Benchmark (GTSRB) [62], GTSDB [63], and Tsinghua-Tencent 100K benchmark [57]. Wang *et al.* [64] presented a traffic sign detection method that uses the histogram of oriented gradient (HOG) and a coarse-to-fine sliding window scheme, which achieved outstanding results in GTSDB competition. Wu *et al.* [22] proposed an approach based on the combination of color transformation and CNN, and achieved competitive results in GTSDB. The Boolean convolutional neural networks (BCNN) [65] also employed the HOG features to detect traffic signs on GTSDB. Zhu *et al.* [57] created the Tsinghua-Tencent 100K benchmark and proposed a CNN for simultaneously detecting and classifying traffic signs. However, the objects in images from these public datasets are relatively large, and are, therefore, not suitable for direct use as training data in detecting small traffic signs in the present study.

Many recent studies have focused on the detection and recognition of traffic signs, but few have addressed their accurate 3-D localization. Timofte *et al.* [66] proposed a multiview scheme, combining 2-D and 3-D analyses, for traffic sign detection and recognition, employing minimum description length (MDL) optimization for 3-D localization of traffic signs. Balali *et al.* [67], [68] proposed an automated computer-vision-based method that detects, classifies, and localizes traffic signs via street-level image-based 3-D point cloud models. An improved structure-from-motion procedure was developed to create a 3-D point cloud for the street level imagery, and 3-D points corresponding to the traffic sign in question were labeled and visualized in 3-D by using camera pose information. Wen *et al.* [69] developed a registration algorithm that projects those

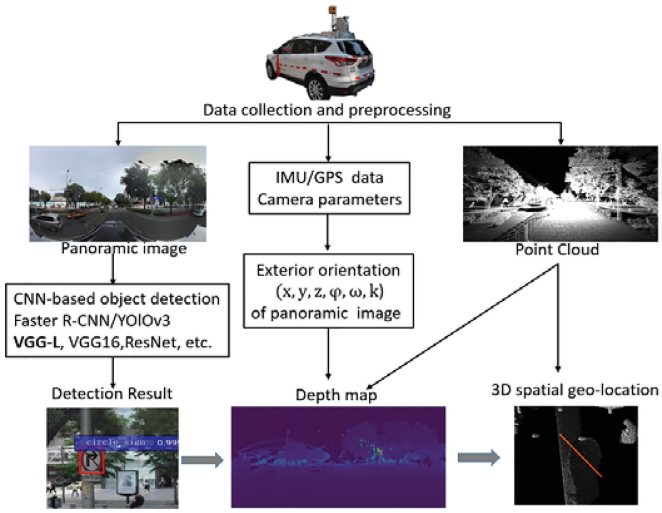


Fig. 1. Flowchart of small traffic signs detection and 3-D Localization.

traffic sign board points onto an image plane based on coordinate transformation for 3-D localization of traffic signs. Krylov and Dahyot [70] proposed an automatic object geolocation technique by an MRF information fusion approach defined over irregular grid, which is capable of fusing information from street level imagery and LiDAR data. However, these methods are either not suitable for the application of real complex scenarios or not highly automated. More effective solutions for automatic 3-D information acquisition of small traffic signs still need to be developed.

This article focuses on the detection and 3-D positioning of small traffic signs represented in panoramic images and point cloud data acquired by the Shou Shi Si Wei (SSW) MMS. The contributions of this article are as follows.

1) This article develops a deep-learning and depth-map based framework for object detection and 3-D localization based on MMS data, which provides a novel solution for automatic collection of 3-D information on small traffic signs.
2) The faster R-CNN model is finely adjusted. A lightweight network, called VGG-L, is proposed as the new backbone network and several parameters are optimized for a better performance of the deep model on detecting small traffic signs from large panorama images.
3) A new concept of the depth-map-based precise 3-D spatial geolocation of small urban targets is proposed. It is the first attempt to use a depth map as an intermediary for automatic 3-D positioning of small traffic signs. A practical method, termed center-diagonal-distance vectorization (CDDV), is proposed for obtaining the final vectors in the 3-D space.

II. Method

This article presents the overall workflow for small traffic sign detection and 3-D positioning based on MMS data. The main steps are as follows (see Fig. 1).

1) *MMS data collection and preprocessing*. The point cloud, panoramic images, and corresponding attitude data are acquired for method validation.
2) *Small traffic signs detection in panoramic images via the CNN-based deep learning method*. Using the deep learning algorithm, faster R-CNN, YOLOv3, etc., for object detection; comparing the detection accuracy of different CNN models; finally a simplified VGG network, called VGG-L, is proposed to extract small traffic signs from panoramic images.
3) *3-D localization and vectorization of small traffic signs*. A depth map corresponding to the panoramic image is generated based on the attitude of the panoramic image exposure moment and the synchronously acquired point cloud data. The depth map is taken as an intermediary to convert the (x, y) pixel coordinates of the detected target on the panoramic image to 3-D geodetic coordinates (X, Y, Z). Then, a center-based method is used to automatically extract the 3-D vector of the target. These steps are described in detail in the following sections.

A. Fine Adjustment of Faster R-CNN

Our method follows the deep learning framework of faster R-CNN, which was shown to be an advanced object detection network for generic object detection [10]. Essentially, it consists of two modules: first, a fully convolutional region proposal network (RPN) that proposes a list of candidate regions likely to contain objects of interest; and second, a downstream fast R-CNN [71] classifier that uses the proposed regions for classifying, refining the boundaries of those regions. Many popular CNNs, such as VGG, ResNet, or MobileNet, can be used as the backbone network under the faster R-CNN framework. Both modules share a common set of convolutional layers of the backbone network and convolutional features of the whole image.

For better performance of faster R-CNN on the detection of small traffic signs, a lightweight VGG network, called VGG-L, is proposed in this article as the backbone network. In the literature [72], it was demonstrated that deeper layers are beneficial for large-scale image classification, while the error rate of the VGG architecture saturates when the depth reaches 19 layers. Therefore, the data volume should be fully considered when designing a CNN model, rather than simply pursuing a deeper network. With increasing number of CNN layers, deeper networks tend to be more robust and the semantic features output by higher network layers will be richer. However, this also results in the loss of high-resolution details in the bottom layers, which are very important for the detection of small targets. As the traffic signs to be detected in this study are very small, and the training dataset is not large, we assumed that if the amount of data is small, then the use of fewer network layers will be helpful for improving small-target detection. We therefore designed a lightweight feature extraction CNN (VGG-L), which follows the VGG architecture and consists of six convolutional layers combined with three maxpool layers.

The network structure is shown in Fig. 2, where "conv3-k" represents the convolution kernel size (3×3) and the number of convolution kernels is k.

As shown in Fig. 2, compared with VGG16, VGG-L keeps the front 5 convolutional layers and reduced 7 convolution layers

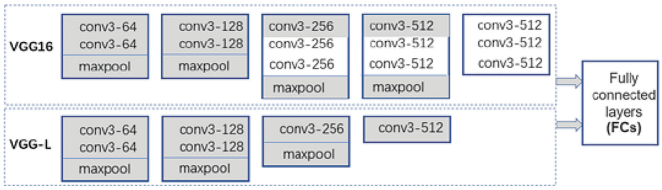


Fig. 2. Network structures of VGG16 and VGG-L.

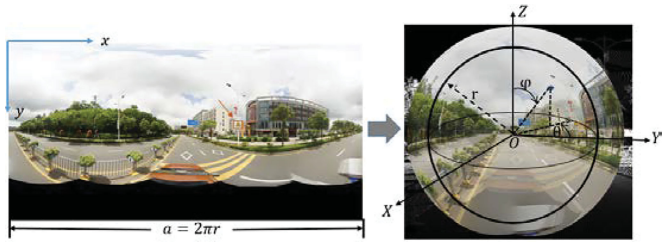


Fig. 3. Relationship of three coordinate systems.

with more channels (256/512) in deeper positions. The pooling layers have also been reduced from 4 to 3. The number of parameters is decreased accordingly [72]. These modifications consider that the traffic signs to be detected from panoramic images are very small. The use of fewer convolutional and pooling layers can result in less information loss on small objects.

In addition, the parameters affecting the detection of small targets in the faster R-CNN model include anchors, normalized input size of images, ROIs for training, proposal threshold, strides, etc. These parameters are optimized in faster R-CNN for detection of small traffic signs. The specific meaning of each parameter and the adjustments are as follows.

1) Anchors: At each sliding-window location of the output feature map, an RPN simultaneously predicts k region proposals, which are parameterized and called anchors. Each anchor is centered at the corresponding sliding window, and is assigned a scale and aspect ratio. In the original implementation of faster R-CNN [48], when scales = [8, 16, 32] and ratios = [0.5, 1, 2], the proposal regions on the corresponding input images are $[128^2, 256^2, 512^2]$ pixels. Thus, we simply set scales = [2, 4, 8] with region size of $[32^2, 64^2, 128^2]$ pixels for better performance in detecting small traffic signs.
2) The normalized input size of the images represents the uniformly scaled input image size of the model. The larger the size, the larger the GPU memory required by the detection model. The normalized input size and the anchors jointly determine the actual candidate region of the target to be detected. Therefore, for the large panoramic images (4096 $\times$ 8192 pixels) and relatively small traffic signs acquired by SSW in our experiment, original images have been divided into blocks (e.g., 512 $\times$ 512 pixels) for training and detection, to avoid the loss of small object details caused by scaling.
3) ROIs for training denote the number of regions sampled for training the region classifier. As gradient accumulation across multiple batches is slow, following the improvement of faster R-CNN in [50], we use 256 ROIs sampled from an image as a minibatch of the stochastic gradient descent in the training stage rather than 128 ROIs from two images. Proposal threshold contains both the size threshold and confidence value threshold of the proposals. The original faster R-CNN removes small proposals ($<$16 pixels in original scale) at the detection stage and those proposal regions with low confidence values ($<$0.05) prior to the nonmaximum suppression stage. Redmon and Farhadi [50] demonstrated that those steps hindered the performance of small objects in both precision and recall to some degree. Thus, we retain small proposals and set the confidence threshold to 0 (instead of 0.05) for our detection process.
4) Strides refers to the ratio of the input image size to the output feature map of a given layer. The value of strides is related to the number of pooling layers in the backbone network. For the same input image, smaller strides leads to a larger feature map and denser anchor boxes. The strides for layer conv5_3 of VGG16 is 16, whereas the value for the last layer of VGG-L is 8, which denotes that one pixel on the output feature map corresponding to an 8 $\times$ 8 region on the original image rather than 16 $\times$ 16 as in VGG16.

For detecting small traffic signs from panoramic images, the above details, which seem to affect the performance of faster R-CNN, have been fully considered in order to facilitate the detection of small objects to some extent.

B. Depth Map Generation

The processing of panoramic images involves three types of coordinate systems: image, spherical, and 3-D space.

The relationship between the three coordinate systems is shown in Fig. 3. Here, (x, y) denotes the image coordinates of the panoramic image, and (X, Y, Z) denotes the sphere-centered 3-D space coordinates. In the spherical coordinate system (SCS), a point is specified by the triplet (r, θ, φ), where r is the point's distance from the origin (the radius), θ is the angle of rotation from the positive direction of the Y-axis, and φ is the angle from the positive direction of the Z-axis. According to the field of view (FOV) of the panoramic image, θ ranges from 0° to 360° and φ ranges from 0° to 180°. The aspect ratio of the panoramic image is 2:1 and the width a of the image is $2\pi r$.

The conversion between the image, spherical, and 3-D space coordinate systems of the panoramic image is shown in following equations:

$$\left.\begin{aligned} \theta &= \tfrac{x}{r} \\ \varphi &= \tfrac{y}{r} \\ a &= 2\pi r \end{aligned}\right\} \tag{1}$$

$$\left.\begin{aligned} X &= r \cdot \sin\theta \cdot \sin\varphi \\ Y &= r \cdot \cos\theta \cdot \sin\varphi \\ Z &= r \cdot \cos\theta \end{aligned}\right\}. \tag{2}$$

A depth map (also known as range image) is an image that contains information about the distance between the surfaces of objects from a given viewpoint. A depth map can be created

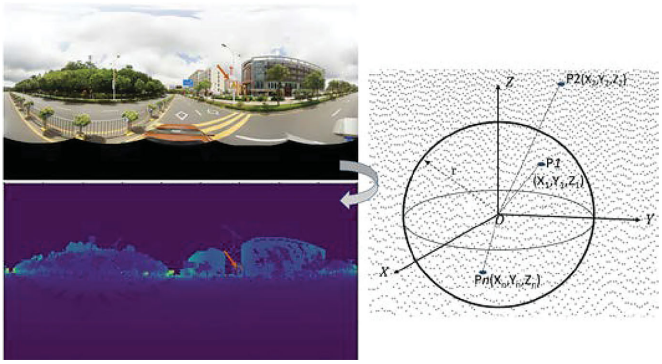


Fig. 4. Depth map generated from point cloud.

from stereo images or generated from the 3-D point cloud with necessary ancillary information. Based on the inertial measurement unit (IMU)/global position system (GPS) data and camera exposure information obtained by the MMS, we can obtain the corresponding external orientation (EO) of each panoramic image, including exposure position (*x, y, z*) and attitude angles (φ, ω, k). Thus, we can generate the corresponding depth map based on the EO of the image and the laser point cloud that were acquired synchronously.

As shown in Fig. 4, the rays emitted from the origin of the SCS of a panoramic image intersect with the laser point cloud, and the 3-D spatial coordinates of the intersection points $(P_1, P_2, \ldots, P_3)$ can be obtained. The distance d_n (3) between the origin and the intersection point is called depth, and we then obtain the depth map by calculating the depth values corresponding to all pixels of the panoramic image. Therefore, each pixel on the depth map has a corresponding depth value given as

$$d_n = \sqrt{X_n^2 + Y_n^2 + Z_n^2}. \tag{3}$$

C. 3-D Localization and Vectorization of Small Traffic Signs

Since this article employs deep learning methods to detect small traffic signs on panoramic images and to obtain the real geographic coordinates or world coordinates of these targets, a depth-map-based method is proposed for the 3-D positioning of detected targets. As shown in Fig. 5, the 2-D coordinates of the targets detected in the panoramic images are converted to 3-D world coordinates using the depth map as an intermediary.

First, the bounding boxes of the traffic signs on the panoramic images are detected by the proposed VGG-L model under Faster R-CNN framework. The depth maps are generated from the laser point cloud corresponding to the EOs of the panoramic images. Then the position (x, y) on the panoramic image can be mapped to the corresponding depth map with coordinates $(x_d, y_d, \mathrm{depth})$, which will be further converted to the SCS and the spherical centered 3-D coordinate system (SC3S) according to (1) and (2). Finally we obtain the world coordinates $(X_w, Y_w, Z_w)$of the detected object by a transformation matrix R from SC3S to the world coordinate system (WCS).

According to the abovementioned process, we can calculate the world coordinates of each pixel in the bounding boxes of the target, while the extracted traffic signs need to be vectorized in the form of points, lines, or polygons for further application. As this study focuses on the extraction and 3-D localization of objects, a simple yet effective approach for vectorizing in the 3-D space is introduced. We propose a CDDV method to connect the diagonals of the bounding boxes of the objects based on their world coordinates.

As illustrated in Fig. 6, CDDV is mainly used to convert the 2-D bounding boxes into 3-D vector lines in the WCS.

In Fig. 6(a), the four corners of the detected bounding box are outside the panel of the traffic sign and the world coordinates of the corners may differ greatly; thus, the choice of the two endpoints of the final vector line is very critical. The vectorization steps are as follows.

1) Take the center pixel of the bounding box, and calculate its corresponding world coordinates $P_0(X_0, Y_0, Z_0)$.
2) Determine the two endpoints of the vector. As shown in Fig. 6(b), we draw the diagonal of a rectangular box from the left top (LT) corner to the right bottom (RB) corner. Then, we take one point every k pixels along the diagonal from both ends, and each point is calculated in 4). The points $P_{\mathrm{LT}}(X_i, Y_i, Z_i)$ and $P_{\mathrm{RB}}(X_j, Y_j, Z_j)$, which satisfy (4), are finally taken as the two endpoints of the vector line.
3) The two endpoints P_{LT} and P_{RB} are connected to form a 3-D vector line in the WCS

$$\begin{cases} (X_i - X_0) < d_{\mathrm{thresh}} \ \text{ and } \ (Y_i - Y_0) < d_{\mathrm{thresh}} \\ \quad \text{and} \ \ (Z_i - Z_0) < d_{\mathrm{thresh}} \\ (X_j - X_0) < d_{\mathrm{thresh}} \ \text{ and } \ (Y_j - Y_0) < d_{\mathrm{thresh}} \\ \quad \text{and} \ \ (Z_j - Z_0) < d_{\mathrm{thresh}}. \end{cases} \tag{4}$$

Here, d_{thresh}denotes the threshold value of the coordinate difference between a given point (X_n, Y_n, Z_n) along the diagonal and the center point (X_0, Y_0, Z_0). The value of d_{thresh} is related to the actual size of the traffic signs. d_{thresh}= 1 m is found appropriate in our experiments.

III. Experimental Data and Setup

A. SSW Data Collection and Preprocessing

In order to evaluate the feasibility and performance of the proposed deep-learning-based detection and 3-D spatial geolocation method, experimental data were collected by the SSW MMS. As shown in Fig. 7, the SSW system is integrated with four microsingle cameras (SONY α7) with automatic exposure control, an RTW laser scanner, an IMU, a GPS antenna, and a wheel-mounted odometer. These sensors are all mounted on the same vehicle. The four cameras (A–D) capture synchronized images, which are then stitched together to form a panoramic image that provides a full visual field of 360° × 180° information. The RTW laser scanner provides a 360° FOV and scans at laser pulse repetition rates up to 550 kHz, 200 scan lines per

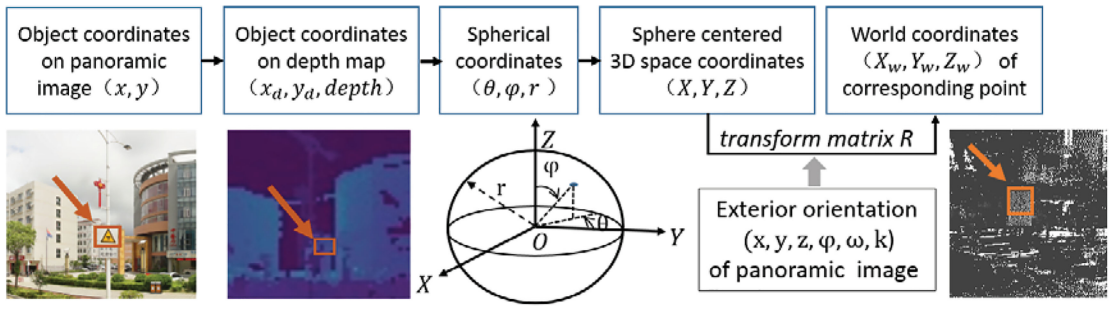


Fig. 5. Depth-map-based 3-D spatial geolocation.

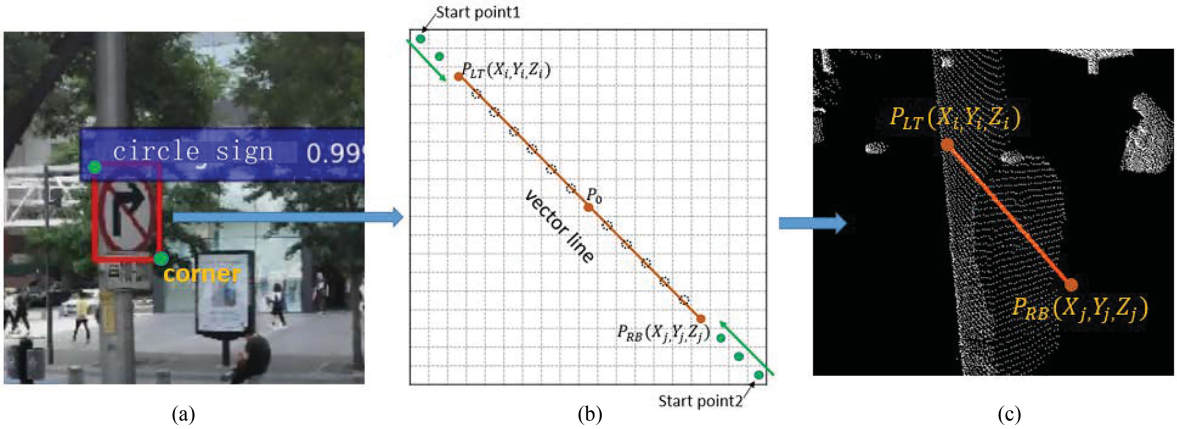


Fig. 6. CDDV method. (a) Detected 2-D bounding box. (b) Determine the endpoints of the vector. (c) 3-D vector line.

Fig. 7. SSW-IV data collection system.

Fig. 8. Data acquisition trajectory. (a) Part of Shenzhen test area. (b) Ningxia test area.

second, and the angle between scanning plane and road surface is 45°.

The test areas are mainly located in the Chinese cities of Shenzhen and Ningxia (see Fig. 8), covering hundreds of kilometers of urban expressways and highways. The data were collected at a vehicle speed of 55 km/h, the image resolution is less than 5 cm, and the exposure spacing of the cameras is 6–10 m. At the ground surface near the scanning center, the distance between adjacent points in the same scan line is 6 mm, and the distance between scan lines is 7.5 cm (see Fig. 7).

As presented in the workflow (see Fig. 9), before conducting the experiments in object detection and 3-D localization, SWDY supporting software (by SSW) was used to carry out some preprocessing work including processing the navigation data, panoramic image stitching, and point cloud processing, etc., as follows.

1) *Navigation data processing*: Based on the attitude data $(\text{roll}, \text{pitch}, \text{yaw})$obtained by the IMU and the high-precision positioning coordinates(x_0, y_0, z_0)acquired by the GPS, combined with the system calibration procedure, the six elements of exterior orientation (EO) for each image can be obtained. The positioning accuracy of the images is $<$10 cm in the plane and $<$5 cm in elevation.
2) *Panoramic image stitching*: The original images are simultaneously recorded by four SONYα7 microsingle

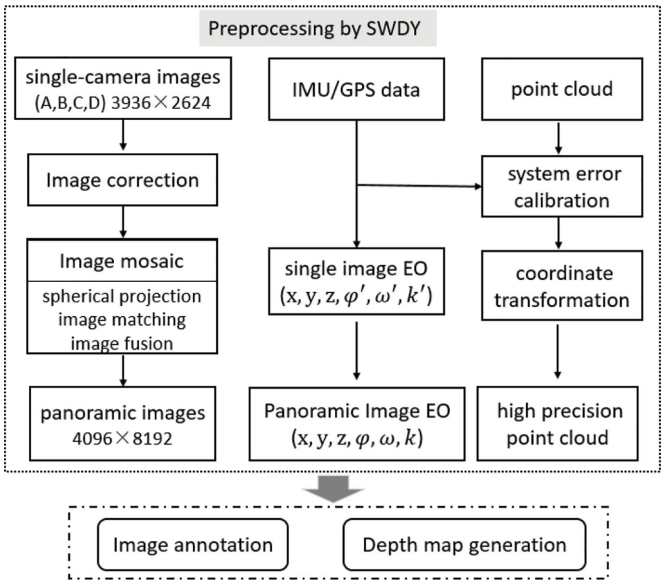


Fig. 9. Workflow for SSW data preprocessing.

cameras (image size 3936 × 2624 pixels) and the cameras' internal parameters are used to correct image distortion. The four corrected single images (from cameras A–D) are then stitched into a whole panoramic image by spherical projection, image matching, and fusion. The size of the output differentially rectified panoramic image is 8192 × 4096 pixels. We also obtain the EO$(x, y, z, \varphi, \omega, k)$of the panoramic image under the WCS, based on the EOs of the four corresponding single images.

3) *Point cloud processing*: Combined with the processing of navigation data, high-precision 3-D spatial coordinates of the laser points can be calculated and calibrated according to the previously established spatial posture relationships between multiple sensors. The final laser point cloud is then obtained by coordinate transformation from the local independent coordinate system to the WCS. The accuracy of the point cloud can be controlled under 2 cm in elevation in road surveys.

Depth maps are generated (as described in Section III-A) based on the preprocessed panoramic images, corresponding EOs, and point cloud data. The resolution of the depth maps can be adjusted by setting the angular resolution in both the horizontal and vertical planes. Considering the large amount of image data: since the panoramic image has 360° × 180° FOV, we set the angular resolution in both directions to 0.2° in the subsequent experiments and set the output depth map size as 1800 × 900 pixels.

B. Dataset Annotation and Statistics

Object detection based on deep learning usually requires amounts of annotated data for training and validating the proposed model. For the detection experiment, we created a small urban traffic sign detection (UTSD) dataset containing three classes: milestones, circular signs, and triangular signs. This dataset was derived from panoramic images acquired by the SSW system. As this study focuses on comparatively small traffic signs, the choice of the three classes was made to meet the needs of both practical application and the current experiments. Traffic signs can be divided into different categories according to function, and each category may be further subdivided into subclasses with different shapes or details [57]. In view of the numerous types of traffic signs, we classify small traffic signs according to their shapes. Milestones include hundred-meter and kilometer piles; circular signs include prohibitory signs and some mandatory signs; and triangular signs include danger signs and also some mandatory signs.

TABLE I
STATISTICS OF UTSD DATASET

Class	Trainval	Test_ Object	Test_ Back	Size (pixel)	Percentage (%)
milestone	816	272	14632	24-165	0.15-5.27
circular sign	477	158			
triangular sign	113	38			

Examples of panoramic images and labeled targets are shown in Fig.10. The images were labeled using the open-source image annotation tool LabelImg [73] and saved as XML files in the widely used PASCAL VOC format. Based on the data acquired in test areas, the labeled images cover some variations in illuminance and weather conditions. More than a quarter of the images in the UTSD were collected at sunset when the light was low, or on cloudy days. However, given the complex types of traffic signs, our labeling process attempts to achieve better detection performance by excluding any signs that are partially occluded.

Owing to the limited available computer GPU memory, all the panoramic images of 4096 × 8192 pixels were divided into smaller sections for further training and testing. One panoramic image can be divided into 128 tiles, with 100 pixels of overlap between adjacent small tiles, to avoid the segmentation of a whole traffic sign while dividing. Thus, the size of image tiles ranges from 512 × 512 to 612 × 612 pixels. In the following experiments, the normalized input size is set to 600 pixels in the faster R-CNN framework as a tradeoff between input resolution and GPU memory usage.

Statistics for the three types of annotated objects are shown in Table I. The UTSD dataset contains 1088 mileage piles, 635 circular signs, and 151 triangular signs. To ensure the multiscale of the targets in the training data, if the same traffic sign appears on multiple panoramic images, it can be counted several times in our statistics, except for the ones with extremely small size (<10 pixels). The ratio of the number of training (Trainval) to testing (Test_Object) samples was set to 3:1 for all three types of signs in the experiment. The sizes of the traffic signs in the images vary between 24 and 165 pixels with regard to the longer edge, and about 80% are of sizes between 30 and 70 pixels.

In addition, on the expressways or highways of test areas, panoramic images were acquired by SSW every 6–10 m, which is very dense. Most of the images contain no objects of interest, and most of the 128 tiles in a panoramic image were occupied by the background without traffic signs. Therefore, we added

Fig. 10. Panoramic images and annotation samples of traffic signs.

a large number of background images without targets (14632 in Test_Back) to the test process, to evaluate the background error [false positives (FPs)] of the proposed CNN model. In Table I, Trainval represents the number of labeled traffic signs for training and validation; Test_Object and Test_Back denote the numbers of test images with and without targets, respectively; Size denotes the statistical size range of the objects on the panoramic images; and Percentage is the percentage range of object sizes in the divided image. It is clear that traffic signs account for less than 6% of the 512 × 512 input images used in the experiment, and most objects only constituted 1%–3% of the images.

C. Experimental Setup

The fine adjustments of faster R-CNN were implemented based on the Tensorflow-based open-source code [74]. The experiments to detect small traffic signs use the UTSD dataset to evaluate the performance of the proposed VGG-L model, as compared with other popular networks, including VGG16, MobileNet, ResNet, and YOLOv3. The UTSD dataset was divided into training and testing sets. The experiments were carried out on a Linux PC with an Intel Xeon E5-1620 CPU, 16 GB of memory, and one NVIDIA Titan Xp GPU with 12 GB of memory.

To assess the detection experiments, we used the universally-agreed standard average precision (AP) of each class as the evaluation metric. In a binary decision problem, a classifier labels examples as either positive or negative. The decision made by the classifier can be represented by four categories [75]: true positives (TPs), FPs, true negatives (TN), and false negatives (FN). TP and FP refer to the number of objects detected, respectively, correctly and falsely in all bounding boxes. TN corresponds to negatives correctly labeled as negative. FN denotes the number of positive objects that are not detected. A bounding box can usually be identified as TP when the intersection over union (IoU) between the proposed bounding box and the true box is $\geq$0.5. Conversely, it will be considered an FP when the IoU is $<$0.5.

Average precision is a measure that combines recall and precision for ranked retrieval results, where a higher AP value reflects better performance in object detection. Precision refers to the ratio of objects detected correctly relative to the total number of detected boxes. Recall denotes the ratio of objects labeled as true targets relative to the real number of true targets. The definitions of precision, recall, and AP are shown in the following equations:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{AP} = \sum_{\text{n}} \left(R_n - R_{n-1}\right) P_n \tag{7}$$

where P_n and R_n denote the precision and recall at the nth threshold.

In addition, for multiclass object detection, we use mean AP (mAP) as the evaluation metric, which is calculated by computing the mean of AP values of each category [76]. For the 3-D localization experiments conducted in Shenzhen and Ningxia, the small circular and triangular signs were mainly detected and geolocated based on the data collected on Shenzhen highways, and the milestones were tested on the data from Ningxia expressways. The recall and precision indicators were used for assessing the 3-D localization experiments.

IV. Experimental Results

A. Small Traffic Signs Detection Experiments and Analysis

This experiment adapted the join-training scheme of faster R-CNN [48], and used VGG-L, VGG16, MobileNet-v1, ResNet-50, ResNet-101, and ResNet-152 as CNN feature extraction networks for comparison. We used the same anchor scales [2, 4, 8] and ratios [0.5, 1, 2] as illustrated in Section II-B for all experiments under the faster R-CNN framework with different backbone networks. Considering the size of the UTSD dataset and the number of categories it contains, each network was trained on 50000 iterations. The learning rate was adjusted according to the number of iterations. For the first 35000 iterations, the learning rate was initialized to 10^{-3} (VGG16) or 5×10^{-4} (VGG-L, MobileNet, and ResNet) according to experience, whereas for the remaining 15000 iterations, it was set to 10^{-4}. The other parameters in the proposed detection network were the same as those in the official faster R-CNN [48]: momentum of 0.9 and weight decay of 10-4 were used, and we only performed horizontal flipping on the training data as image augmentation.

Using the UTSD dataset, we also compare our detection framework with the state-of-the-art YOLOv3 deep object detector with the backbone network of Darknet-53 inside. The experiments used the open-source implementation of YOLOv3 in Keras with the Tensorflow backend [77]. Different from faster R-CNN, *K*-means clustering method was adopted in YOLOv3 for calculating anchors, and the number of anchors was set to 9 in our experiment. Due to the dataset size and limited GPU memory, the normalized input size of images was set to 608 × 608 pixels, the learning rate was set to 10-4, and the batch size during training was set to 4. Based on experience, the training epoch was set to 100 in the Keras implementation of YOLOv3, and the iteration stopped at the 41st epoch when the total loss met a given threshold for the UTSD dataset in the actual experiment.

In this experiment, two tests were performed on all the networks, one using Test_Object data containing 468 small traffic sign samples and the other using Test_Object + Test_Back data, which added 14 632 background images without targets for testing.

The AP values of each category and mAP of the VGG-L and other object detection models are presented in Table II. It can be seen that VGG-L shows the highest accuracy in the detection experiment based on the faster R-CNN framework. The overall mAP of the three categories of urban traffics signs reached 97.8% in Test1, which was 0.9%–9% higher than that of VGG16, ResNet, and MobileNet. The AP values of milestones differ only slightly (97.6%–100%) between the models, whereas the

TABLE II
ACCURACY (%) AND SPEED (FPS) OF DIFFERENT MODELS ON UTSD DATASET

Framework	Model	Learning rate	Test1（Test_Object）				Test2（Test_Object+ Test_Back）				Speed (FPS)
			milestone (AP)	circular sign	triangular sign	mAP (%)	milestone	circular sign	triangular sign	mAP (%)	
Adjusted Faster R-CNN	**VGG-L**	35k:0.0005 15K:0.0001	100	96.6	96.5	**97.8**	99.8	62.8	63.5	**75.4**	**20**
	VGG16	35k: 0.001 15K:0.0001	99.3	95.1	96.3	**96.9**	96.9	60.4	56.3	**71.2**	16
	MobileNet-v1	35k:0.0005 15K:0.0001	97.6	87.4	**81.4**	**88.8**	93.7	44.5	62.9	**67.0**	**25**
	ResNet-50		99.8	94.9	94.9	**96.5**	93.4	50.4	59.7	**67.8**	14
	ResNet-101		99.6	92.3	90.4	**94.1**	88.9	43.0	54.1	**62.0**	12
	ResNet-152		98.8	92.9	**96.8**	**96.2**	84.6	43.0	54.3	**60.6**	10
YOLOv3	Darknet-53	0.0001	**99.8**	95.6	95.9	97.1	92.1	46.5	55.7	**64.8**	12

The bold and underlined entities in Table II are the experimental results that require the reader to focus on, which correspond to the experimental analysis in the following paragraphs of section IV-A.

detection results for circular and triangular signs differ greatly (81.4%–96.8%). The mAP of VGG-L reached 75.4% on the test dataset with a large number of background images added (Test2), which was 14.8% higher than that of the lowest ResNet-152 network. In Test2, the APs of all three types of urban traffics signs differ greatly.

It is obvious that the overall test result of Test2 was much lower than that of Test1; that is, after adding a large number of background tests, the false detection rate is greatly increased, and the mAP drops by more than 20% for all detection models. Especially for ResNet series, the detection precision decreased greatly in Test2, resulting in up to 35.2% decline in mAP values. As most of the panoramic images are background in the actual application, the results in Test2 are taken to represent realistic detection levels for small traffic signs.

In terms of speed, Mobilenet-v1 is slightly faster than VGG-L, but the mAP is 9% and 8.4% smaller in the two tests. As seen in Table II, VGG-L also has advantages in accuracy and speed compared with the advanced YOLOv3 detection method, especially for the data in Test2.

As shown in Fig. 11, the mAP values of ResNet-50, ResNet-101, and ResNet-152 show declining trends in Test2. VGG-L also outperforms VGG16 on the two detection tests of the UTSD dataset. Thus, we can infer that a deeper network does not ensure better detection results, especially for a small dataset with limited samples. Since the proposed VGG-L network performs better than other existing models, the detection results of VGG-L under fine-adjusted faster R-CNN in Test2 were used for further 3-D localization experiments in Section IV-B.

Some experimental detection results of the proposed fine adjusted faster R-CNN method regarding various scenes are shown in Fig. 12. The most common mistake is that wheels are easily mistaken as circular traffic signs in complex urban environments. The bottom group of Fig. 12 presents that the same milestone has been detected in several consecutive images, which indicates the robustness of the detection algorithm to scale changes to some extent, while the repetitive problem of targets should be considered in subsequent statistical processing.

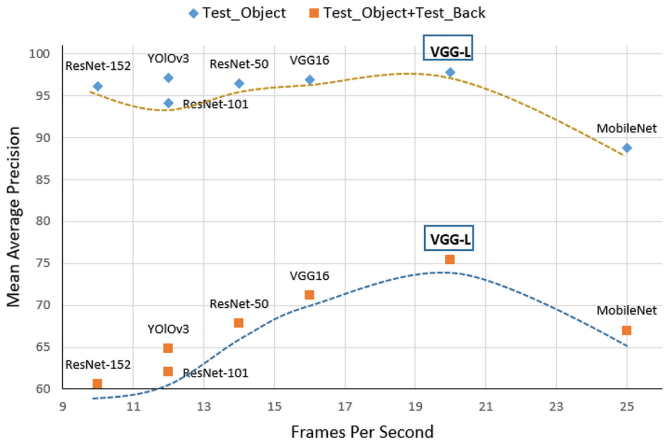


Fig. 11. Results of detection experiment based on UTSD dataset.

For further comparison, we also tested our detector on the Tsinghua-Tencent 100K (TT100K) benchmark, which contains 100000 cropped street view images with 30000 traffic-signs of 45 classes in total. The traffic signs are divided into three categories according to their size: small (area $<32^2$ pixels), medium ($32^2 <$ area $<96^2$ pixels), and large (area $>96^2$ pixels). The average precision and recall for the different sizes of traffic signs are given in Table III. The test results on the TT100K benchmark show that the proposed model, based on adjusted faster R-CNN with VGG-L backbone, performs obviously better than the network used in the TT100K [57]. Therefore, the proposed method is robust and not only applicable in the UTSD dataset of this article, but also performs well in the open dataset.

B. 3-D Localization Experiments and Analysis

Based on the VGG-L network and finely adjusted faster R-CNN detector proposed here, using the UTSD dataset for model training, selected data from the Shenzhen and Ningxia test areas were chosen for 3-D localization experiments on small urban traffic signs. According to the target distribution of the test areas,

Fig. 12. Experimental results regarding various scenes. (a) Challenging scene test results in downtown area or poor lighting environment. (b) Test results in suburban expressways.

about 25 km of Shenzhen expressway data were selected for the detection and 3-D spatial geo-location of circular and triangular signs. More than 10 km of expressway data from the Ningxia test area were mainly chosen to detect and locate milestones. The test data used for this experiment were excluded from the UTSD dataset.

The test process mainly comprises two steps. The first employs the proposed deep network to detect three types of traffic signs from the panoramic images; the second employs the proposed depth-map-based 3-D spatial geolocation method to obtain the real 3-D locations of the targets. The latter includes 3-D localization and vectorization of the detected 2-D bounding boxes, error elimination (EE) and repetition removal, etc. The test results were statistically analyzed by the evaluation indicators of recall and precision.

1) Performance of the CDDV Method: According to the process described in Section III, we can obtain the 3-D world coordinates of each pixel in the 2-D bounding boxes of the target. For vectorization of the targets, as shown in Fig. 6(b), we first directly connected start points 1 and 2 of the bounding box as the

(a) (b) (c)

Fig. 13. Vectorization results of the two methods.

TABLE III
DETECTION RESULTS FOR DIFFERENT SIZES OF TRAFFIC SIGNS USING TT100K AND OUR APPROACH

Method	Category	AR	AP
TT100K	small (0,32]	87%	82%
	medium (32,96]	94%	91%
	large (96,400]	88%	91%
Proposed	small (0,32]	90%	88%
	medium (32,96]	96%	93%
	large (96,400]	91%	91%

TT100K: Method in benchmark of Tsinghua-Tencent 100K.
Proposed: Adjusted faster R-CNN with VGG-L backbone.
AR: average recall. AP: average precision.

TABLE IV
EVALUATION OF THE CDDV METHOD

Method	Class	Total Detections	Targets	False Vectorization	VER
Direct	milestone	72	68	43	63.2%
	circular sign	204	109	67	61.5%
	triangular sign	88	49	33	67.3%
CDDV	milestone	72	68	2	2.9%
	circular sign	204	109	3	2.7%
	triangular sign	88	49	1	2.0%

vector line in the experiment, which is called "direct" method and produced many errors. We then used the proposed CDDV approach, in which the d_{thresh} value was set to 1 m.

Fig. 13 compares some results from the two methods: Panel (a) presents the diagonals of the detected bounding boxes on the panoramic images, and the error vector results produced by directly connecting both ends of the boxes are shown in (b). In most cases, the two endpoints of the bounding box are outside the plane of the traffic sign, thus the corresponding point coordinates may differ greatly in 3-D spatial, which may result in a much longer connecting vector line than the true scale of a traffic sign, or else an entirely wrong position far from the real target. Panel (c) presents the correct vectors produced by the CDDV method.

In order to evaluate the accuracy of the proposed CDDV, as shown in Table IV, "Total Detections" denotes the detected 2-D bounding boxes on images, "Targets" denotes the detected targets belong to the three categories (IoU $\geq$ 0.5), "False Vectorization" denotes the number of targets that are wrongly 3-D vectorized by "direct" or "CDDV" methods, and "VER" denotes vectorization error rate.

The calculation results of the two methods show that, the VER of the "direct" method is very high (all over 60%), whereas the "CDDV" method greatly reduced the vectorization error rate; all the VER value have been controlled under 3%. In addition to the adopted vectorization method, the calibration error of images and laser data also has certain impacts on the VER value here. Inaccurate 2-D image coordinates or 3-D point locations will cause the objects detected on the panorama images to not be mapped to the point cloud space correctly, which may lead to vectorization errors. By further improving the calibration accuracy, the VER can be reduced to some extent.

2) Error Elimination (EE): Taking the Shenzhen test area as an example, the CDDV method is used to vectorize the detection results for the entire test area of 25 km. The superposition results of the vectors and point cloud are shown in Fig. 14. It can be seen that there are still some relatively long erroneous vector lines.

Analysis indicates that these errors may be caused by the following issues: The first error type is FPs output in the detection phase using deep learning methods. These FPs, such as vehicles wheels that are mistakenly classified as circular signs, may not be suitable for the center-based CDDV vectorization methods and are eventually converted into incorrect vectors. Secondly, the obtained depth maps may contain inaccuracies due to point cloud noise, which further affects the 3-D positioning accuracy of detected targets. The third issue is that, due to some existing shortcomings of the CDDV method, a small number of detected TPs are still converted to longer vectors.

According to the actual sizes of the three types of small traffic signs detected in this study, generated vectors longer than 2 meters were simply deleted in order to eliminate most errors.

3) Removal of Repetitive Targets (RRT): Since SSW images are acquired every 6–10 m, the same object may appear on multiple panoramic images and be detected. As shown in Fig. 15,

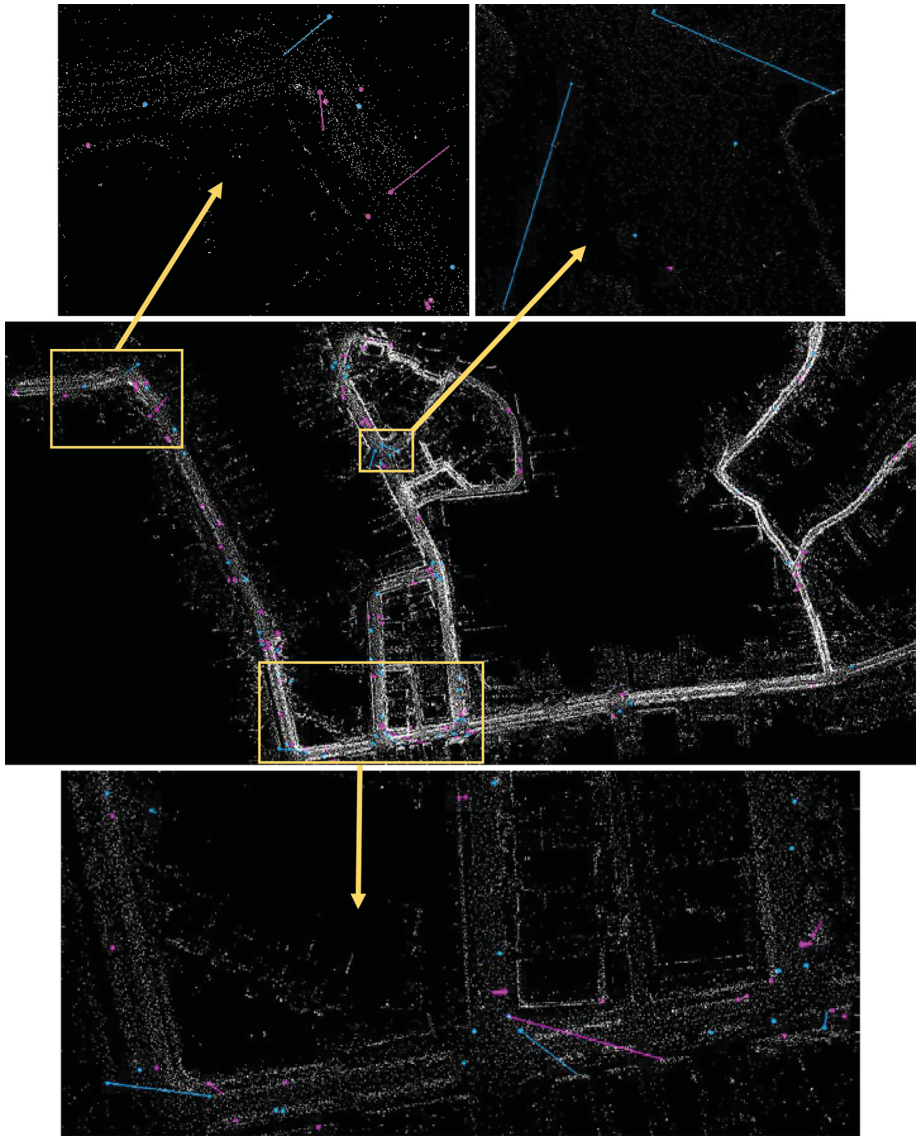


Fig. 14. Detected vector results in Shenzhen test area. Red = circular signs and blue = triangular signs. The enlarged panel shows an area in greater detail.

Fig. 15. Samples of repetitive targets detected at the same position.

the same traffic sign occurring on both images was detected and converted into very closely spaced vectors. In view of this situation, the distance judgment approach was used to remove repetitive targets. If the distance between the centers of two vectors is less than a specified threshold d_v in the 3-D WCS, only one vector will ultimately be preserved. The experiment shows that d_v value within the range 0.2–0.3 m can effectively eliminate repeated targets, and the final results of the experiments in this section used a distance threshold of 0.25 m.

4) Result Statistics and Analysis: Samples of the final test results for the milestones, circular, and triangular traffic signs are shown in Fig. 16. With the metric of IoU = 0.5, the numbers of FP, TP, and FN of each category were counted separately, and the Recall and Precision parameters were calculated for evaluation. The statistical values of the detection results are derived after the above process of eliminating errors and removing repetitive targets. "CDDV + EE + RRT" denotes the statistic is based on

TABLE V
STATISTICAL SUMMARY OF 3-D LOCALIZATION RESULTS FOR THE THREE TARGET TYPES

Method	Class	False Positive	True Positive	False Negative	Recall	Precision
CDDV + EE + RRT	milestone	2	51	1	98.08%	96.2%
	circular sign	7	59	9	86.7%	89.4%
	triangular sign	5	31	4	88.6%	86.1%

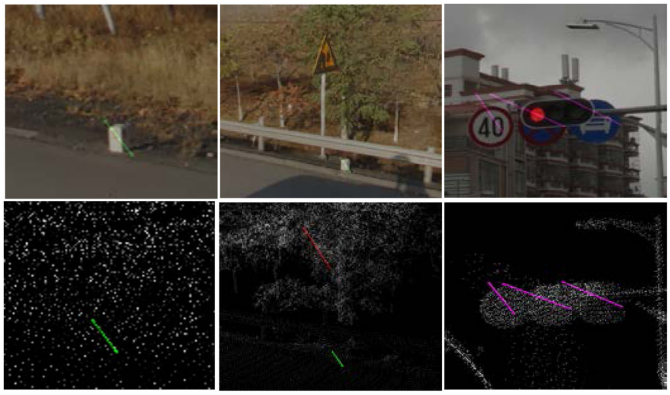


Fig. 16. Samples of the final 3-D spatial geolocation results for small traffic signs.

the CDDV method with eliminating errors (EE) and removing repetitive targets (RRT) in the 3-D WCS.

The results (see Table V) show that Recall and Precision for the three types of targets in the test areas exceeded 86% using the proposed fully automatic detection, 3-D positioning, and vectorization scheme. There was previously no effective solution for the automatic 3-D localization of these small traffic signs. However, the results demonstrate that the overall proposed scheme is feasible and has great application value in practical projects.

V. CONCLUSION

This study tackles the problem of automatic extraction of 3-D information for small traffic signs based on MMS data. The faster R-CNN framework was optimized for improved performance, including a proposed backbone network called VGG-L and the optimization of several parameters. Experiments using the generated UTSD dataset illustrate the superiority of the proposed model over other existing deep networks, both in terms of accuracy and speed. Based on the detected 2-D bounding boxes of the small traffic signs from panoramic images, the depth-map-based 3-D spatial geolocation method is proposed to obtain the real 3-D locations of targets, which includes the processes of depth map generation, coordinate transformation, and vectorization using the CDDV method, EE, and RRTs.

Quantitative evaluation of the experimental 3-D localization results indicates that the proposed workflow provides advantageous performance, and the automatic detection and 3-D localization of small traffic signs both is efficient and helpful for the regular maintenance and management of the target signs. The proposed method can be used to extract various types of small, urban street furniture for constructing high-precision 3-D city representations based on MMS data.

Although extensive experiments and comprehensive evaluations have demonstrated the feasibility and superiority of the presented method, there is still much room for improvement in the accuracy of small traffic sign detection and 3-D positioning. In the future work, we will expand the UTSD dataset and study more effective CNN detection networks for systematic detection of traffic signs, and more robust algorithms will be explored to improve the accuracy of vectorizing various *traffic* signs.

REFERENCES

[1] A. Jaakkola, J. Hyyppä, H. Hyyppä, and A. Kukko, "Retrieval algorithms for road surface modelling using laser-based mobile mapping," *Sensors*, vol. 8, pp. 5238–5249, 2008.

[2] D. Li and S. O. Elberink, "Optimizing detection of road furniture (pole-like objects) in mobile laser scanner data," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci*, vol. 1, pp. 163–168, 2013.

[3] M. Lehtomäki, A. Jaakkola, J. Hyyppä, A. Kukko, and H. Kaartinen, "Detection of vertical pole-like objects in a road environment using vehicle-based laser scanning data," *Remote Sens.*, vol. 2, pp. 641–664, 2010.

[4] S. Pu, M. Rutzinger, G. Vosselman, and S. O. Elberink, "Recognizing basic structures from mobile laser scanning data for road inventory studies," *ISPRS J. Photogramm. Remote Sens.*, vol. 66, pp. S28–S39, 2011.

[5] L. Li, Y. Li, and D. Li, "A method based on an adaptive radius cylinder model for detecting pole-like objects in mobile laser scanning data," *Remote Sens. Lett.*, vol. 7, pp. 249–258, 2016.

[6] A. Kukko, H. Kaartinen, J. Hyyppä, and Y. Chen, "Multiplatform mobile laser scanning: Usability and performance," *Sensors*, vol. 12, pp. 11712–11733, 2012.

[7] Z. Shi, Z. Kang, Y. Lin, Y. Liu, and W. Chen, "Automatic recognition of pole-like objects from mobile laser scanning point clouds," *Remote Sens.*, vol. 10, 2018, Art. no. 1891.

[8] A. Aijazi, P. Checchin, and L. Trassoudaine, "Segmentation based classification of 3D urban point clouds: A super-voxel based approach with evaluation," *Remote Sens.*, vol. 5, pp. 1624–1650, 2013.

[9] Y. Yu, J. Li, H. Guan, C. Wang, and J. Yu, "Semiautomated extraction of street light poles from mobile LiDAR point-clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, pp. 1374–1386, 2015.

[10] B. Yang, L. Fang, Q. Li, and J. Li, "Automated extraction of road markings from mobile LiDAR point clouds," *Photogramm. Eng. Remote Sens.*, vol. 78, pp. 331–338, 2012.

[11] Z. Wang *et al.*, "A multiscale and hierarchical feature extraction method for terrestrial laser scanning point cloud classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2409–2425, May 2015.

[12] R. Schnabel, R. Wahl, and R. Klein, "Efficient RANSAC for point-cloud shape detection," *Comput. Graph. Forum*, vol. 26, pp. 214–226, 2007.

[13] R. B. Rusu and S. Cousins, "3D is here: Point cloud library (PCL)," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2011, pp. 1–4.

[14] A. Burt, M. Disney, and K. Calders, "Extracting individual trees from lidar point clouds using *treeseg*," *Methods Ecol. Evol.*, vol. 10, pp. 438–445, 2019.

[15] H. Guan, W. Yan, Y. Yu, L. Zhong, and D. Li, "Robust traffic-sign detection and classification using mobile LiDAR data with digital images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1715–1724, May 2018.

[16] Y. Yu, J. Li, H. Guan, and C. Wang, "Automated extraction of urban road facilities using mobile laser scanning data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2167–2181, Aug. 2015.

[17] H. Wang *et al.*, "Object detection in terrestrial laser scanning point clouds based on Hough forest," *IEEE Geosci. Remote Sens. Letters*, vol. 11, no. 10, pp. 1807–1811, Oct. 2014.

[18] P. Huang, M. Cheng, Y. Chen, H. Luo, C. Wang, and J. Li, "Traffic sign occlusion detection using mobile laser scanning point clouds," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 9, pp. 2364–2376, Sep. 2017.

[19] Y. Yu, J. Li, C. Wen, H. Guan, H. Luo, and C. Wang, "Bag-of-visual-phrases and hierarchical deep models for traffic sign detection and recognition in mobile laser scanning data," *ISPRS J. Photogramm. Remote Sens.*, vol. 113, pp. 106–123, 2016.

[20] J. Jung, E. Che, M. J. Olsen, and C. Parrish, "Efficient and robust lane marking extraction from mobile lidar point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 1–18, 2019.

[21] X. Yuan, X. Hao, H. Chen, and X. Wei, "Robust traffic sign recognition based on color global and local oriented edge magnitude patterns," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 4, pp. 1466–1477, Aug. 2014.

[22] Y. Wu, Y. Liu, J. Li, H. Liu, and X. Hu, "Traffic sign detection based on convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, 2013, pp. 1–7.

[23] W. Ritter, "Traffic sign recognition in color image sequences," in *Proc. Intell. Vehicles'92 Symp.*, 1992, pp. 12–17.

[24] S. Maldonado-Bascón, S. Lafuente-Arroyo, P. Gil-Jimenez, H. Gómez-Moreno, and F. López-Ferreras, "Road-sign detection and recognition based on support vector machines," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 2, pp. 264–278, Jun. 2007.

[25] S. M. Bascón, J. A. Rodríguez, S. L. Arroyo, A. F. Caballero, and F. López-Ferreras, "An optimization on pictogram identification for the road-sign recognition task using SVMs," *Comput. Vision Image Understanding*, vol. 114, pp. 373–383, 2010.

[26] L.-W. Tsai, J.-W. Hsieh, C.-H. Chuang, Y.-J. Tseng, K.-C. Fan, and C.-C. Lee, "Road sign detection using eigen colour," *IET Comput. Vision*, vol. 2, pp. 164–177, 2008.

[27] U. L. Jau, C. S. Teh, and G. W. Ng, "A comparison of RGB and HSI color segmentation in real-time video images: A preliminary study on road sign detection," in *Proc. Int. Symp. Inf. Technol.*, 2008, pp. 1–6.

[28] X. W. Gao, L. Podladchikova, D. Shaposhnikov, K. Hong, and N. Shevtsova, "Recognition of traffic signs based on their colour and shape features extracted using human vision models," *J. Vis. Commun. Image Representation*, vol. 17, pp. 675–685, 2006.

[29] Y. Yang, H. Luo, H. Xu, and F. Wu, "Towards real-time traffic sign detection and classification," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 2022–2031, Jul. 2016.

[30] A. Arlicot, B. Soheilian, and N. Paparoditis, "Circular Road sign extraction from street level images using colour, shape and texture databases maps," in *Proc. Workshop Laserscanning*, 2009, pp. 205–210.

[31] S. Houben, "A single target voting scheme for traffic sign detection," in *Proc. IEEE Intell. Veh. Symp.*, 2011, pp. 124–129

[32] W.-J. Kuo and C.-C. Lin, "Two-stage road sign detection and recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2007, pp. 1427–1430.

[33] Á. González *et al.*, "Automatic traffic signs and panels inspection system using computer vision," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 2, pp. 485–499, Jun. 2011.

[34] G. Loy and N. Barnes, "Fast shape-based road sign detection for a driver assistance system," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2004, pp. 70–75.

[35] N. Barnes, A. Zelinsky, and L. S. Fletcher, "Real-time speed sign detection using the radial symmetry detector," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 2, pp. 322–332, Jun. 2008.

[36] Y. Gu, T. Yendo, M. P. Tehrani, T. Fujii, and M. Tanimoto, "Traffic sign detection in dual-focal active camera system," in *Proc. IEEE Intell. Veh. Symp.*, 2011, pp. 1054–1059.

[37] Y. Aoyagi and T. Asakura, "A study on traffic sign recognition in scene image using genetic algorithms and neural networks," in *Proc. IEEE 22nd Int. Conf. Ind. Electron., Control, Instrum.*, 1996, pp. 1838–1843.

[38] R. Belaroussi and J.-P. Tarel, "Angle vertex and bisector geometric model for triangular road sign detection," in *Proc. Workshop Appl. Comput. Vision*, 2009, pp. 1–7.

[39] B. Hoferlin and K. Zimmermann, "Towards reliable traffic sign recognition," in *Proc. IEEE Intell. Veh. Symp.*, 2009, pp. 324–329.

[40] K. Brkić, A. Pinz, and S. Šegvić, "Traffic sign detection as a component of an automated traffic infrastructure inventory system," in *Proc. 33rd Annual Workshop Austrian Assoc. Pattern Recognit.*, 2009, pp.1–12.

[41] H. Fleyeh and M. Dougherty, "Road and traffic sign detection and recognition," in *Proc. 16th Mini-EURO Conf. 10th Meeting of EWGT*, 2005, pp. 644–653.

[42] J. Greenhalgh and M. Mirmehdi, "Real-time detection and recognition of road traffic signs," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1498–1506, Dec. 2012.

[43] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 580–587.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.

[45] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, pp. 211–252, 2015.

[46] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes (VOC) challenge," *Int. J. Comput. Vision*, vol. 88, pp. 303–338, 2010.

[47] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 740–755.

[48] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[49] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 21–37.

[50] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv:1804.02767*, 2018. Available: https://arxiv.org/abs/1804.02767

[51] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6517–6525.

[52] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 779–788.

[53] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. Comput. Vision Pattern Recognition*, 2017, pp. 2117–2125.

[54] H. H. Aghdam, E. J. Heravi, and D. Puig, "A practical approach for detection and classification of traffic signs using convolutional neural networks," *Robot. Auton. Syst.*, vol. 84, pp. 97–112, 2016.

[55] R. Qian, B. Zhang, Y. Yue, Z. Wang, and F. Coenen, "Robust Chinese traffic sign detection and recognition with deep convolutional neural network," in *Proc. 11th Int. Conf. Natural Comput.*, 2015, pp. 791–796.

[56] X. Changzhen, W. Cong, M. Weixin, and S. Yanmei, "A traffic sign detection algorithm based on deep convolutional neural network," in *Proc. IEEE Int. Conf. Signal Image Process.*, 2016, pp. 676–679.

[57] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2110–2118.

[58] A. Jain, A. Mishra, A. Shukla, and R. Tiwari, "A novel genetically optimized convolutional neural network for traffic sign recognition: A new benchmark on Belgium and Chinese traffic sign datasets," *Neural Process. Lett.*, vol. 50, pp. 3019–3043, 2019.

[59] Y. Tian, J. Gelernter, X. Wang, J. Li, and Y. Yu, "Traffic sign detection using a multi-scale recurrent attention network," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 12, pp. 4466–4475, Dec. 2019.

[60] A. Vennelakanti, S. Shreya, R. Rajendran, D. Sarkar, D. Muddegowda, and P. Hanagal, "Traffic sign detection and recognition using a CNN ensemble," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2019, pp. 1–4.

[61] J. Zhang, M. Huang, X. Jin, and X. Li, "A real-time Chinese traffic sign detection algorithm based on modified YOLOv2," *Algorithms*, vol. 10, 2017, Art. no. 127.

[62] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German Traffic Sign Recognition Benchmark: A multi-class classification competition," in *Proc. Int. Joint Conf. Neural Netw.*, 2011, pp. 1453–1460.

[63] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The German traffic sign detection benchmark," in *Proc. Int. Joint Conf. Neural Netw.*, 2013, pp. 1–8.

[64] G. Wang, G. Ren, Z. Wu, Y. Zhao, and L. Jiang, "A robust, coarse-to-fine traffic sign detection method," in *Proc. Int. Joint Conf. Neural Netw.*, 2013, pp. 1–5.

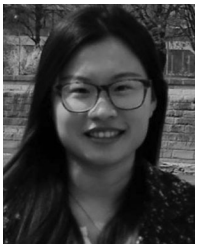
[65] Z. Xiao, Z. Yang, L. Geng, and F. Zhang, "Traffic sign detection based on histograms of oriented gradients and Boolean convolutional neural networks," in *Proc. Int. Conf. Mach. Vision Inf. Technol.*, 2017, pp. 111–115.

[66] R. Timofte, K. Zimmermann, and L. Van Gool, "Multi-view traffic sign detection, recognition, and 3D localisation," *Mach. Vision Appl.*, vol. 25, pp. 633–647, 2014.

[67] V. Balali, A. Jahangiri, and S. G. Machiani, "Multi-class US traffic signs 3D recognition and localization via image-based point cloud model using color candidate extraction and texture-based recognition," *Adv. Eng. Inform.*, vol. 32, pp. 263–274, 2017.

[68] V. Balali and M. Golparvar-Fard, "Recognition and 3D localization of traffic signs via image-based point cloud models," in *Proc. Int. Workshop Comput. Civil Eng.*, 2015, pp. 206–214.

[69] C. Wen *et al.*, "Spatial-related traffic sign inspection for inventory purposes using mobile laser scanning data," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 27–37, Jan. 2016.
[70] V. A. Krylov and R. Dahyot, "Object geolocation using mrf based multi-sensor fusion," in *Proc. 25th IEEE Int. Conf. Image Process.*, 2018, pp. 2745–2749.
[71] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
[72] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[73] Tzutalin, *LabelImg*, Git code. 2015. [Online]. Available: https://github.com/tzutalin/labelImg
[74] X. Chen and A. Gupta, "An implementation of faster RCNN with study for region sampling," 2017, *arXiv:1702.02138*.
[75] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 233–240.
[76] N. Mo, L. Yan, R. Zhu, and H. Xie, "Class-specific anchor based and context-guided multi-class object detection in high resolution remote sensing imagery with a convolutional neural network," *Remote Sens.*, vol. 11, 2019, Art. no. 272.
[77] Experiencor, keras-yolo3, 2018. Available: https://github.com/experiencor/keras-yolo3

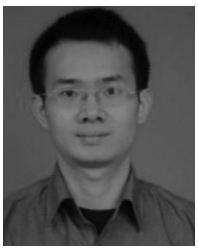


Lirong Liu received the Ph.D. degree in geographic information systems from Capital Normal University, Beijing, China, in 2018.

She is currently doing postdoctoral work with Land Satellite Remote Sensing Application Center, Ministry of Natural Resources, Beijing, China. Her research interests include deep learning, object detection, and satellite remote sensing.

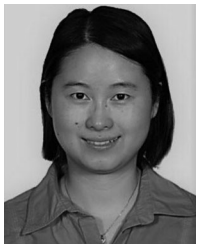
Xinming Tang received the M.S. degree in land administration from the Faculty of Geo-Information Science and Earth Observation (ITC), Enschede, the Netherlands, in 1998, and the Ph.D. degree in geoinformation science and computer application from the University of Twente, Enschede, the Netherlands, in 2004.

He is currently the Academic Deputy Director of the Land Satellite Remote Sensing Application Center, Ministry of Natural Resources, Beijing, China. His scientific interests are in spatial information science and technology, including remote sensing, geographic information systems, and their integration. He is currently the President of the Commission I Working Group V of the International Society for Photogrammetry and Remote Sensing.



Junfeng Xie received the M.S. and Ph.D. degrees in surveying and mapping engineering from Wuhan University, Wuhan, China, in 2006 and 2009, respectively.

He is currently a Professor with the Land Satellite Remote Sensing Application Center, Ministry of Natural Resources, Beijing, China. He has authored or coauthored several papers in *Remote Sensing*, *International Journal of Digital Earth*, etc. His research interests include attitude determination and satellite imaging model construction.



Xiaoming Gao received the M.S. and Ph.D. degrees in cartography and geographic information engineering from Wuhan University, Wuhan, China, in 2005 and 2017, respectively.

She is currently the Director of Research and Development Department, Land Satellite Remote Sensing Application Center, Ministry of Natural Resources, Beijing, China. Her research interests include high-resolution optical mapping and interferometric radar techniques.

Wenji Zhao received the Ph.D. degree in tectonic geology from Jilin University, Changchun, China, in 1998.

From 1998 to 2000, he was a Postdoctoral Fellow with the Institute of Remote Sensing and Geographic Information System (GIS), Peking University, Beijing, China. Since 2000, he has been with Capital Normal University, Beijing, China. His research interests include the applications of remote sensing and GIS on geoscience.

Fan Mo received the B.S. degree in surveying and mapping engineering from Shandong University of Technology, Zibo, China, in 2009, and the M.S. degree in surveying and mapping engineering from the Information Engineer University, China, in 2013.

He is currently an Assistant Research Fellow of LASAC, and employed in the Calibration Department. He has authored the laser altimeter calibration software. His research interest focuses on spaceborne laser altimetry. He also does some research in satellite attitude processing, including jitter frequency detection by different methods and attitude data correction.

Gang Zhang received the M.S. degree in software engineering from the University of Chinese Academy of Sciences, Beijing, China, in 2007.

He is currently with the Chinese Academy of Surveying and Mapping, Beijing, China. His research interests include three-dimensional spatial data obtaining, photogrammetry, and satellite remote sensing.