Automotive Innovation (2021) 4:400–412
https://doi.org/10.1007/s42154-021-00157-x

VCANet: Vanishing-Point-Guided Context-Aware Network for Small Road Object Detection

Guang Chen[1,2] · Kai Chen[1] · Lijun Zhang[1] · Liming Zhang[3] · Alois Knoll[2]

Received: 22 November 2020 / Accepted: 4 June 2021 / Published online: 13 September 2021

Abstract
Advanced deep learning technology has made great progress in generic object detection of autonomous driving, yet it is still challenging to detect small road hazards in a long distance owing to lack of large-scale small-object datasets and dedicated methods. This work addresses the challenge from two aspects. Firstly, a self-collected long-distance road object dataset (TJ-LDRO) is introduced, which consists of 109,337 images and is the largest dataset so far for the small road object detection research. Secondly, a vanishing-point-guided context-aware network (VCANet) is proposed, which utilizes the vanishing point prediction block and the context-aware center detection block to obtain semantic information. The multi-scale feature fusion pipeline and the upsampling block in VCANet are introduced to enhance the region of interest (ROI) feature. The experimental results with TJ-LDRO dataset show that the proposed method achieves better performance than the representative generic object detection methods. This work fills a critical capability gap in small road hazards detection for high-speed autonomous vehicles.

Keywords Autonomous driving · Road hazard · Object detection · Deep learning · Vanishing point

Abbreviations

ROI	Region of interest
TJ-LDRO	Tongji long-distance road object
VCANet	Vanishing-point-guided context-aware network
VPT	Vanishing point

1 Introduction

Detecting small road hazards in a long distance, such as lost cargo on highway, is a very demanding capability for high-speed autonomous vehicles. According to the US Department of Transportation, nearly 150 people were killed annually due to the traffic accidents involving lost hazardous cargo [1]. Although small road hazards detection is very critical for traffic safety, it is rarely addressed for autonomous vehicles. In order to fill this gap, a highly efficient small object detection system is proposed together with a dedicated long-distance road object dataset (TJ-LDRO dataset).

It has been accepted in past research [2–5] that the detection of small road object is quite an enormous challenge. Radar and Lidar are two of the most widely used active sensors applied in autonomous vehicles for target detection and tracking. Although they could provide high-accuracy measurement of point-wise distance and velocity, a low angular resolution is almost inevitable and leads to the failure of the detection of small objects. For instance, a typical Velodyne HDL-64E Lidar has a vertical angular resolution of about 0.4°. Assuming that three consecutive points are the minimum requirement for a true detection, the maximum detecting distance of this Lidar is shorter than 15 m for a small vertical object of 20 cm. Thus, when the vehicle is at a high speed, it is rather dangerous to detect this kind of small obstacles even at a long distance. Although dynamic vision sensors show great potential in detecting and tracking high-speed objects, the limited spatial resolution may lead to the failure of small object detections [6]. However, cameras often provide very high spatial resolution at a relatively low cost, which could help cope with this challenge.

Since a small object covers very limited image area and provides a rather small amount of texture information, the

✉ Guang Chen
guangchen@tongji.edu.cn

✉ Lijun Zhang
tjedu_zhanglijun@tongji.edu.cn

[1] Tongji University, Shanghai, China

[2] Technical University of Munich, Munich, Germany

[3] Geely Research Institute, Hangzhou, China

detection of which is more difficult than that of large ones such as cars and trucks in close range. Recently, most of state-of-the-art object detection methods [7–9] follow the region-based paradigm. The detection performance highly relies on the discriminative capabilities of detecting the feature lying in the bounding box, which requires sufficient training data of the objects. Whereas, the inadequate feature representation from object regions gives rise to difficulty in the detection of small objects. Therefore, on the condition that the methods are applied to detect the small road hazards, the performance would suffer significant reduction.

Many works have addressed the challenges in several aspects. Considering the quite limited data for small road object detection, the *Lost and Found* dataset [1] is introduced. To the authors' best knowledge, it is the only public dataset dedicated to the small road object detection. The *Lost and Found* dataset consists of 2104 annotated frames of 112 raw video sequences. Nevertheless, compared with other generic object detection datasets such as PASCAL VOC [10] and MSCOCO [11], the size of the *Lost and Found* dataset is still far from satisfying the standard deep learning approaches. Aiming at tackling the challenges with limited appearance information of the small objects, many approaches are proposed based on state-of-the-art object detection frameworks [12–14] , which further revise the network architecture by putting special emphasis on embedding the multi-scale representation [15–17], context information [18, 19] and super-resolution [20, 21]. However, these methods exaggerate the importance of the recognition of each separate region and pay little attention to the crucial semantic correlation among objects, regions and visual patterns.

To meet the challenge and solve the issue, a vanishing-point-guided method is proposed in order to focus more on the small objects as well as filter other disturbance, through which the feature is upsampled and fused in a multi-scale feature fusion pipeline. Moreover, to enrich the feature of small objects, a context-aware method is designed to encode the context information as a part of the small objects and then discuss the inference of the context. The major contributions of this paper can be summarized in the following three aspects:

- A new large-scale dataset dedicated to small road hazard detection[1];
- A new context-aware method for small object detection, which encodes the context information as a part of the small objects and enriches the features;
- A vanishing-point-guided method to focus more on the small objects, through which the feature maps can be upsampled and fused in a multi-scale feature fusion pipeline.

[1] https://github.com/ispc-lab/VCANet.

The main content of this work will cover six parts. Section 2 briefly reviews the related works about the small object detection methods and datasets. Section 3 introduces the self-collected TJ-LDRO dataset. Section 4 presented the proposed vanishing-pointguided context-aware network (VCANet). Section 5 gives the experimental results and ablation studies of VCANet on TJ-LDRO dataset. Section 6 concludes this work.

2 Related Works

Most earlier research on small object detection are on detecting vehicles in aerial images [22], flying objects in the sky [23, 24] or obstacles on the road [1, 2], using traditional methods with hand-crafted feature and shallow classifiers. Recently, some small object detection methods adapt deep learning technology as well as improve the detection performance of small objects by modifying the generic object detection frameworks on aspects including multi-scale representation, context information, image resolution, candidate region extraction and so on. This section first illustrates the related works of existing small object detection methods, and then describes the few public datasets tailored for small object detection.

2.1 Detection Methods with Multi-scale Representation

Multi-scale representation has been proved to be useful for many recognition tasks, such as semantic segmentation [25–27] and object detection [13, 18, 28], especially for the small object detection. Most of the two-stage detectors, such as RCNN [29], Fast-RCNN [12] and Faster-RCNN [8], merely use the last layer of feature maps to classify and locate the target object. As a result, although these state-of-the-art algorithms are good at detecting generic objects, they present poor performance as applied to detect small objects. To address this matter, multi-scale representation and fuse multi-layer feature maps are applied in Refs. [13, 17, 28, 30–33] to improve the detection performance of small objects. MR-CNN [17] performs a multi-scale deconvolutional operation upsampling the feature maps of deep layers and concatenates them with those of shallow layers. SSD [13] uses feature maps from shallow layers for small object detection, and it could exploit feature maps from deeper layers for larger object detection. For further improvement in the detection accuracy, DSSD [28] adds additional deconvolutional layers to the end of SSD [13], which combines the prediction layers and their deconvolutional layers to ensure more accurate detection of small objects. MDSSD [15] also presents a deconvolutional fusion block and uses skip connection to fuse more context information. A multi-level

feature fusion method based on SSD is proposed in Ref. [30], in which two modules of concatenation and element-sum are added to the basic SSD backbone to fuse the information of different feature layers in diverse ways. For avoiding mechanical stacking of multi-scale feature maps, Ref. [16] introduces a channel-aware deconvolutional network to study the relationship among feature maps in various channels.

2.2 Detection Methods with Context Information

In the physical world, visual objects appear in a specific environment and usually coexist with other related objects. Sufficient evidence in neuroscience [34, 35] has illustrated that context plays a crucial role in human recognition of target objects. Detection of small objects has been stuck in a dilemma, mainly because small objects contain too little information to be detected accurately only by their own feature. Ref. [36] argued that one could utilize image evidence beyond the object extent, formulating as "context." And it also presents a simple human experiment where users attempt to discriminate the true and false positive faces [36]. Obviously, humans need context to accurately classify tiny faces. Moreover, previous studies [37–39] have shown that appropriate context information is beneficial for object detection and recognition, especially when the feature of target object is insufficient for the small size, occlusion or poor image quality. Ref. [19] first introduced ContextNet to encode the context clue around small object proposal. Ref. [18] proposed inside-outside net (ION), an object detector that exploits information both inside and outside the region of interest (ROI). Additionally, Ref. [40] considered each column of feature maps as a spatial sequence. Then, a novel LSTM-based encoder–decoder, adding an attention mechanism, is used to explore detailed contextual information.

2.3 Detection Methods with Image Resolution and Others

Another common way to detect small objects is to upscale the resolution of raw images. For instance, Ref. [41] upscaled the input images and generated high-resolution feature maps. Ref. [20] introduced generative adversarial network (GAN) to reconstruct the super-resolution image, which significantly enriches the information of small objects. Ref. [21] constructed a perceptual generative adversarial network (PGAN) model to improve detection performance via narrowing the representation difference between small objects and the large ones.

In addition, a novel region context network (RCN) was designed in Ref. [42], which is used to generate the most likely candidate regions with small objects and their semantic information. Classification and regression are performed on these most optimal regions to reduce the memory usage and improve the location accuracy. As a result of the downsampling operation, the small object feature is reduced so that the recall rate becomes much lower. Ref. [43] put forward an atrous region proposal network (ARPN) to explore object contexts in multiple scales and incorporate atrous convolution into Fast R-CNN, thus improving detection rate of small object. Ref. [44] established Shifted SSD, which relieves the influence of the discreteness of the anchor method by moving the feature maps. Ref. [45] presented small-object-sensitive-CNN for small traffic signs detection where the large input image is cropped into small patches.

2.4 Related Dataset for Small Object Detection

In addition to object detectors, the large and high-quality dataset is another critical factor of deep learning technology. Although there are excellent datasets such as MSCOCO [11] and PASCAL VOC [10], focusing on generic objects detection task, datasets dedicated to small object detection are rather rare. Moreover, the small object detection datasets either have a small number of images or are inconsistent with our research scenes, so they could not meet the requirement of small road hazards detection. For example, a lost-cargo dataset was employed in Ref. [1] comprising only 2104 frames with pixel-level annotations of obstacle and free-space, which are collected from 13 different street scenarios, and involve 37 different combinations of objects. A dataset was applied in Ref. [46] for road garbage detection, consisting of 801 images and 966 bounding boxes. Ref. [47] built the Tsinghua-Tencent 100K dataset, including 100,000 images in 100 classes and 30,000 traffic sign instances. Ref. [19] established a small object detection dataset using a subset images from the MSCOCO dataset [11] and the Scene UNderstanding database (SUN) [48].

3 TJ-LDRO: TongJi Long-Distance Road Object Dataset

Research based on deep learning technology requires a large amount of data for tests, and the performance of generic detectors would degrade significantly when there exists huge bias of datasets which mainly result from different application domain. For the long-distance road object detection, a new dataset is collected in this work from both real and virtual simulation environment, called "TJ-LDRO Dataset." ZED binocular cameras and test vehicles are employed to capture real-world data. AirSim [49] and Unreal Engine4 [50] are applied to create virtual data in a simulation environment. The dataset has collected a total of 109,337 images,

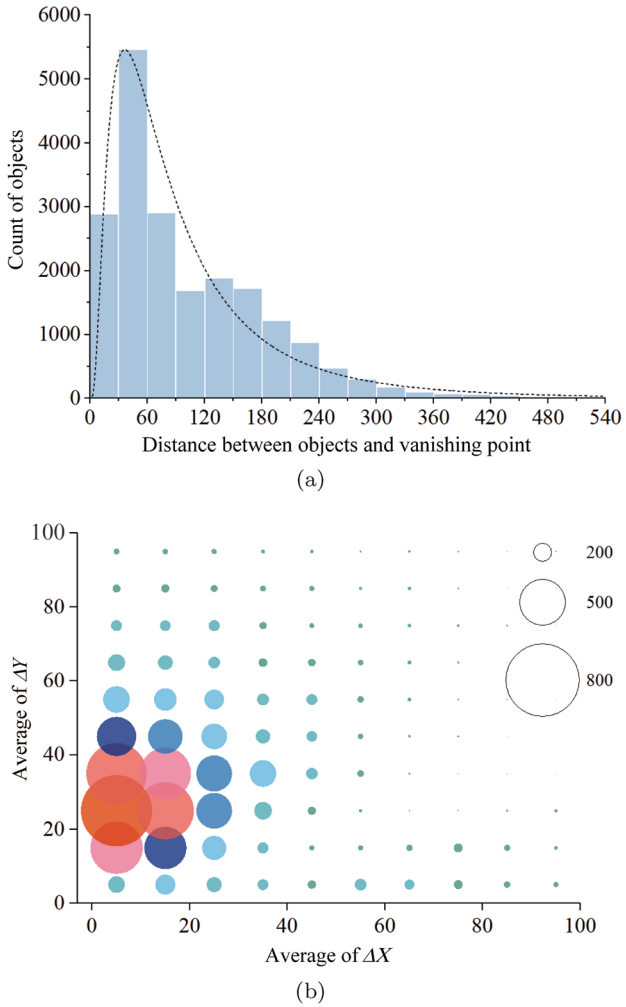

Fig. 1 The statistics of the TJ-LDRO Dataset: **a** The relationships between the count of objects and the distance from them to the vanishing points; **b** The distribution of objects from the vanishing point. The area of the dots is proportional to the number of the objects

each of which is labeled in detail. Furthermore, this dataset has been tested with the proposed small object detection framework in the following sections, providing basis for final superior detection performance. The statistics of the TJ-LDRO Dataset is shown in Table 1. The pixel distance statistics between the object and the road vanishing point are shown in Fig. 1.

3.1 TJ-LDRO Dataset Collected in Real-World Environment

In order to include as many scenes as possible in the real environment and to ensure the robustness of following test results, the real-world dataset is collected in various weather and light environments such as sunny, cloudy, rainy, and evening. What's more, different disturbance is estimated in real scenes such as shade, leaves, straights, curves, street lights, car lights and so on. Some of the weather and scenes are shown in Fig. 2, where Fig. 2a–f are selected from the real-word environment under six different conditions, Fig. 2g–l are selected of six distances from the objects, Fig. 2m–r are selected from the virtual environment under six different conditions.

As the road detection targets, 32 selected objects own the most probability to appear on the road, including road cones, warning signs, plastic box, children's cars, etc., as shown in Fig. 3. These objects are composed of 60 combinations of different number or different types of objects. The combinations vary greatly from each other in order to cover the real-world situation as much as possible. To obtain multi-scale images of each combination in the same field and scene, comprehensive data are collected from different distance from the camera, as a result getting 798 pieces of videos. It is premised that the maximum distance between the camera and the target objects is about 80 m, and the minimum distance is about 10 m. Images are taken in the interval of 15 frames, and each image is acquired when vehicle moves forward about 7 m. Figure 2 shows 6 images with six distances from the objects. The selected 32 objects are classified into 18 categories corresponding to 18 labels respectively, as shown in Table 1. The ground-truth of the objects as well as the road vanishing point in each image are annotated by using bounding box, as shown in Fig. 4. The vanishing point is a point where the parallel lines in a three-dimensional space converge into a two-dimensional plane by a graphical perspective. Herein, the road vanishing point is defined as the intersection of road edges. Images with a

Table 1 Statistics of the TJ-LDRO dataset in real world. Relative area of each instance is computed as the ratio of the bounding box area over the image area

Category	Tyre	Bucket	Chair	Plastic box	Traffic cone	Fire extinguisher	Warning sign	Car parts	Bicycle
Number of images	655	3265	826	3754	1152	440	495	1061	477
Median relative area	0.011	0.010	0.023	0.013	0.014	0.009	0.018	0.001	0.026
Category	Trunk	Scooter	Tricycle	Pedestrian	Dog	Deer	Woven bag	Carton	Plank
Number of images	570	278	874	711	830	707	412	384	73
Median relative area	0.013	0.032	0.032	0.024	0.029	0.031	0.019	0.013	0.013

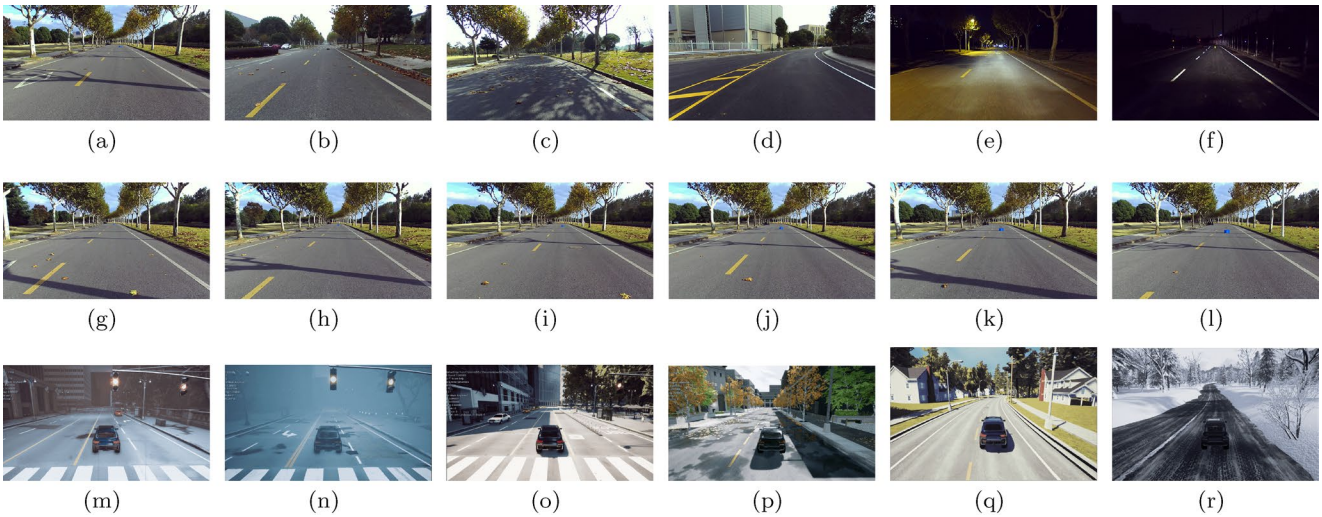

Fig. 2 Some examples of weather and scenes under different disturbance

Fig. 3 Collection of small objects included in the TJ-LDRO dataset

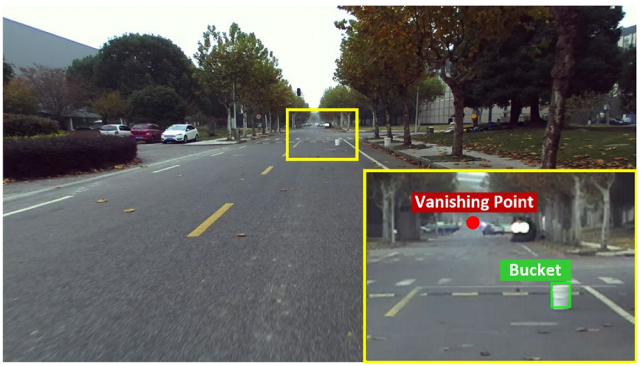

Fig. 4 Annotation of small object and vanishing point

bounding box larger than 96 × 96 pixels are discarded, considering the size of small objects. After deleting the unqualified images, 19,337 images are obtained, including 14,474 positive samples containing 17,284 bounding boxes of small objects and 4,863 negative samples with no objects. The size of each image is 2208 × 1242.

3.2 TJ-LDRO Dataset Collected in Virtual Environment

The TJ-LDRO dataset is enriched in the virtual environment, in view of the safety and difficulty of collecting a large amount of the TJ-LDRO images in the real environment, the advantages of convenient, diverse scenes, and high efficiency in data collection in the virtual environment.

In the virtual simulation environment, 7 scene models are selected from different places such as urban city, suburb, small town, village, mountain roads, etc., as well as several different weather such as sunny, cloudy, rainy, snowy, fogy, dusty and so on. Part of the virtual places and weathers are shown in Fig. 2. 187 object models are carefully identified to form 400 combinations. Similar to the real environment, the same collection method is adopted in the virtual environment. The distance between objects and the virtual camera is from 20 to 80 m. Each image is collected at a fixed distance of about 7 m, and a total of 15 images are collected in each photographing period. A total of 90,000 images are collected, and the size of each image is 2208 × 1242. Each image includes three formats: original image, depth image and segmentation image. In the segmentation image, each object corresponds to one label, and both of the ground and non-ground parts have their unique labels.

In addition, a random swing of 0° to 10° is added to the virtual vehicle to simulate the real vehicle as much as possible. In order to avoid the singleness of objects position, 5 positions are selected on each collection road, and each object could randomly select these positions. For ensuring the diversity of object angles, a horizontal random rotation of 0° - 360° is set for each object at each position. At each image taken, the virtual vehicle randomly swings an angle to left or right, then the object would randomly pick a position while rotating horizontally at an angle.

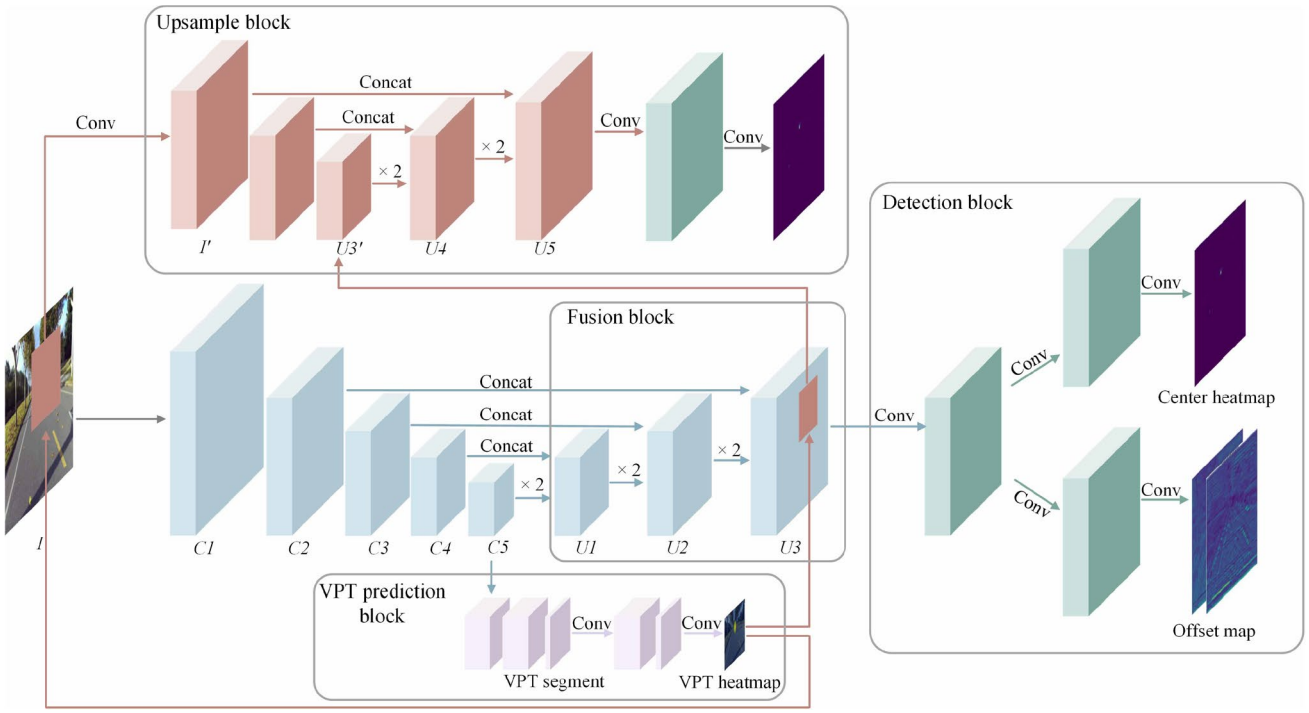

Fig. 5 The overall architecture of the proposed VCANet

4 Vanishing-Point-Guided Context-Aware Network

The overall architecture of the proposed VCANet is illustrated in Fig. 5. Our VCANet has four blocks including the multi-scale features fusion block (shown as blue), the vanishing point (VPT) prediction block (shown as purple), the vanishing-point-guided upsampling block (shown as red) and the context-aware object center detection block (shown as green). Firstly, the backbone network, a standard network pretrained on ImageNet [59] (e.g., ResNet [51]), takes an image I as input and outputs the 32× downsampled feature map $C5$. Then, features in $C5$ enters into the VPT prediction block to obtain the vanishing point (see Sect. 4.1). Next, in order to fuse the low-level detailed information and the high-level semantic information, the fusion block merges each series of earlier multi-scale features into stage-by-stage upsampled layers in corresponding size, and outputs the 8× upsampled feature map $U3$. After that, guided by vanishing point, a ROI of fixed size is generated and mapped to the corresponding area of original image I and feature map $U3$. In this case, the feature maps of ROI are further upsampled 4× and finally matched with the resolution of original image I via the upsampling block (see Sect. 4.4). The multi-scale features fusion pipeline is described in detail in Sect. 4.2.

According to the principle of perspective, most of the small road hazards near vanishing points are tiny, while the size of that grows larger as the hazard moves away from the vanishing point. The proposed method uses vanishing point feature to sense tinier and further objects on a deeper yet limited region of feature maps, thus achieving a trade-off between speed and accuracy. Otherwise, the larger and nearer ones can also be detected satisfyingly from the rich representation provided by the fusion block simultaneously. The end of VCANet is the context-aware center detection block, which takes the multi-scale fused feature maps as input and predicts the center point of each object (see Sect. 4.3).

4.1 VPT Prediction Block

The VPT prediction block is based on the work of Ref. [52], and a quadrant mask is used to divide the whole image into four sections. The VPT is the intersection of these four quadrant sections and could be inferred with the structure information of full global scene. Besides, an end-to-end method is proposed to generate the location of a VPT, different from the probability-based method proposed in Ref. [52]. In consideration of the distinctive feature maps of quadrant mask (illustrated in Fig. 6), a new VPT is presented, predicting the location of VPT from the mask directly. Accordingly, the VPT prediction is divided into two subtasks: quadrant mask generation task and VPT prediction task.

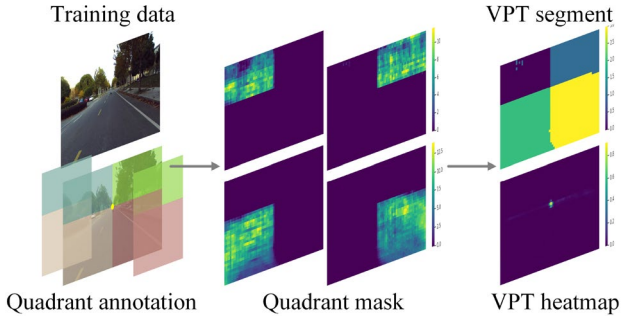

Fig. 6 Output visualization of quadrant based VPT prediction block

4.1.1 Quadrant Mask Generation Task

An illustration example of quadrant annotation and quadrant mask is depicted in Fig. 6. The quadrant mask generation task is formulated as semantic segmentation via the cross-entropy loss L_{seg}. Four channels for the outputs are defined to stand for the four sections divided by VPT, each keeps the same size as $C5$.

4.1.2 VPT Prediction Task

The VPT prediction task is regarded as a classification task. As shown in Fig. 6, every pixel in the VPT heatmap is identified to be VPT or not. Note that determining an "exact" VPT is difficult, thus it is harder to train the VPT utilizing the hard-designation of positives. Hence, similar to Ref. [53], a 2D Gaussian mask $G(.)$ is applied to reduce the ambiguity of negatives surrounding the positives at the center of each positive VPT:

$$G(i,j) = e^{-\frac{(i-x_0)^2+(j-y_0)^2}{2\sigma^2}} \tag{1}$$

where (x_0, y_0) is the coordinate of VPT, and the variance σ is relative to the diameter of Gaussian mask. Here, the diameter d is set to 9 and the σ to $\frac{d}{6}$. Thus, the VPT prediction loss L_{VPT} can be defined as

$$L_{\text{VPT}} = -\sum_{i=1}^{H}\sum_{j=1}^{W}\hat{p}_{ij}, \tag{2}$$

$$\hat{p}_{ij} = \begin{cases} (1-p_{ij})^{\alpha}\log(p_{ij}), & \text{if } y_{ij}=1 \\ (1-y_{ij})^{\beta}(p_{ij})^{\alpha}\log(1-p_{ij}), & \text{otherwise} \end{cases} \tag{3}$$

where p_{ij} is the confidence score indicating whether there is a VPT at location (i, j), and y_{ij} is the ground-truth label computed by Eq. 1, note that $y_{ij} = 1$ represents the positive location. α and β are hyper-parameters and $\alpha = 2$ and $\beta = 4$ during training as suggested in Ref. [53].

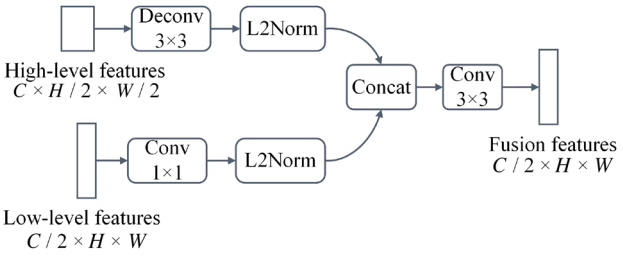

Fig. 7 The architecture of the multi-scale features fusion pipeline

4.2 Multi-scale Features Fusion Pipeline

It is generally believed that low-level feature maps reserve more localization information, while the high-level ones have more semantic information because of their larger receptive fields. Therefore, inspired by Refs. [54, 55], a multi-scale feature fusion pipeline is presented to introduce the semantic information of high-level feature maps to the low-level ones, so that the fusion feature maps have both rich sematic information and abundant localization information.

Figure 7 shows the multi-scale feature fusion pipeline used in both features fusion block and upsampling block. Firstly, the high-level feature maps are upsampled by a deconvolutional layer to match the height and width of the low-level ones, and a 1×1 convolutional layer is applied to the low-level ones to match the channels. What is more, since the low-level and high-level feature maps have different scales, a L2-normalization is used to rescale their norms to 10, as mentioned in Ref. [55]. Then, the two feature maps are concatenated together and the final fusion feature is generated by a 3×3 convolutional layer. The L2 Norm is formulized as follows:

$$\hat{\mathbf{X}} = \frac{\mathbf{X}}{||\mathbf{X}||_2} \tag{4}$$

where $\boldsymbol{X} = (x_1, x_2, .., x_c)$ is a c-dimensional input, $||\boldsymbol{X}||_2 = (\sum_{i=1}^{c}|x_i|^2)^{\frac{1}{2}}$ is the L2 norm of $\boldsymbol{X}$.

4.3 Context-Aware Center Detection Block

4.3.1 Context-Aware Ground Truth Generation

As shown in Fig. 8, a target object is too small in the images to provide enough feature to be identified, which gives rise to a great challenge of learning the representation of a small object only from itself. Previous studies such as Refs. [19, 56] indicate that the context can provide additional information of objects, helping make representations. In order to bridge context and objects, the related area is divided into

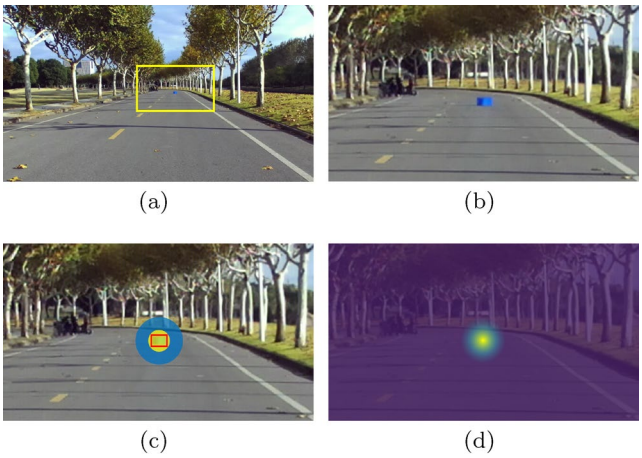

Fig. 8 Illustration of the Context-aware Ground Truth Generation: **a** is the original image; **b** is the 4 times upsampled result of the area covered by the yellow rectangle in **a**; **c** shows 3 different area related to the target object: the red rectangle is the ground-truth bounding box, the yellow circle represents the object area, which is the circumcircle of the red rectangle, and the blue annulus represents the context area; **d** shows the generated ground-truth as a heatmap

two parts: the object area and the context area. Based on this, Eq. 1 is modified into a Context-aware Ground Truth Generation method as follows:

$$G(m^2) = e^{-\frac{m^2}{2\sigma^2}} \tag{5}$$

in which

$$m^2 = M(d^2) = \begin{cases} pd^2, & \text{if } d^2 \le r^2 \\ qd^2 - (q-p)r^2, & \text{if } d^2 > r^2 \end{cases} \tag{6}$$

where m^2 is used to distinguish the context area from the object area, r is the circumradius of the ground-truth bounding box, and $d^2 = (i - x_0)^2 + (j - y_0)^2$. p and q are the hyper-parameters to adjust the influence of the object and the context.

4.3.2 Context-Aware Center Detection Block

Referring to the context-aware ground truth generation method, the context-aware center detection block is designed. It takes the fused feature maps *U3* as inputs and has two branches including the center detection branch and the offset regression branch. The loss function L_{cen} of the center detection branch is the same as Eq. 2. More importantly, it is found that the center heatmap outputted by center detection branch is downsampled as shown in Fig. 5, leading the location $P = (x, y)$ in the original image to correspond to the location $P' = ([\frac{x}{n}], [\frac{y}{n}])$ in the center heatmap, where n is the downsampling factor. Hence, once P' is remapped to P, the precision would be lost, thus affecting the accuracy of center detection. To address this issue, an offset regression branch is applied to predict the offset before remapping center prediction result to original image. The offset is defined as follows:

$$o_k = \left(\frac{x_k}{n} - \left[\frac{x_k}{n}\right], \frac{y_k}{n} - \left[\frac{y_k}{n}\right]\right) \tag{7}$$

where o_k, x_k and y_k are, respectively, the offset, x and y coordinating for center k. Specifically, the smooth L1 loss L_{off} is adopted only at the ground-truth center locations.

4.4 Upsampling Block

ROI feature maps and ROI images are generated, using the upsampling block as inputs by mapping the location of VPT to fused feature map *U3* and the original image *I*. Due to the small size of target object, even 4× downsampling would cause serious loss of information. In order to introduce more detailed information and enhance the model performance on rather small object, the 4× upsampled ROI feature maps and original ROI images are fused. The loss function L_{up} is the same as L_{cen}. But when generating the ground truth, only the objects smaller than 32×32 are selected while ignoring the others. Moreover, since the final feature map keeps the same resolution with original ROI image, the offset of center detection would not exist and corresponding loss L_{off} is of no significance.

4.5 Implementation Details

According to the task of small road hazard detection, the VCANet is proposed as an end-to-end and fully convolutional neural network architecture. The convolutional layer is only used to extract feature and the downsampling is realized by pooling layer. In order to enlarge the receptive field, dilated convolution is applied to both VPT prediction block and center detection block, which can provide more global and semantic information. Final loss function is the sum of loss functions from all mentioned blocks, which is clarified as follow:

$$L = L_{\text{seg}} + L_{\text{VPT}} + L_{\text{cen}} + L_{\text{off}} + L_{\text{up}} \tag{8}$$

5 Experiments of VCANet

In this section, the VCANet is evaluated on the TJ-LDRO dataset. The experimental settings are introduced first. Then experiments are compared with the state-of-the-art technologies. Finally, ablation studies and experimental results are discussed.

5.1 Experimental Settings

5.1.1 Experimental Dataset: TJ-LDRO Dataset

The VCANet is comprehensively evaluated on the TJ-LDRO dataset. TJ-LDRO dataset includes 14,474 annotated images of 18 classes of objects. As all the objects are rather small in the dataset, it is almost impossible to identify the types of objects. Besides, the autodrive task focuses more on localizing the obstacles on the road rather than distinguishing their classes. Herein, when training and testing, all the objects are regarded as a single class, as only the localization performance is stressed instead of the classification.

5.1.2 Evaluation Metrics for Small Object Detection

The basic metrics for object detection is mAP, which is defined in MSCOCO [11]. Average precision (AP) computes the average precision value for recall value over 0 to 1. More specifically, a 101-point interpolated AP definition, in which the recall value is from 0 to 1 at intervals of 0.1, is used in MSCOCO [11]. Moreover, mean average precision (mAP) is the average AP over multiple IoU threshold (the minimum IoU to consider a positive match). However, as for the small object detection task, using IoU threshold to distinguish the result of the true and false detection is no longer appropriate because of the tiny size of objects. When the size is small enough, even a little difference between the prediction location and ground-truth location can cause huge IoU difference. Therefore, instead of the IoU threshold, the distance between their center points is used to determine whether the result is true or not and build up the mean distance average precision (mDAP) and mean distance average recall (mDAR) metrics. Besides, follow the symbol in Ref. [11], DAP_4 means the DAP calculated with the distance threshold of 4 and DAP_S represents the DAP computed among all small objects (the size of which is less than 32×32). Considering the practical application in autonomous driving, the scope of distance threshold is set from 4 pixels to 16 pixels with the stride of 2 pixels. Smaller distance threshold means higher localization precision but lower recall rate.

5.1.3 Data Augmentation

Because of the particularity of small road hazards detection, using the original images to train the model directly has several disadvantages. Firstly, the size of images in TJ-LDRO is 2208×1242, which is too large for training. Meanwhile, the images could not be resized because objects are quite small. In addition, the vanishing points are always in the center of the original images, which may cause serious data bias.

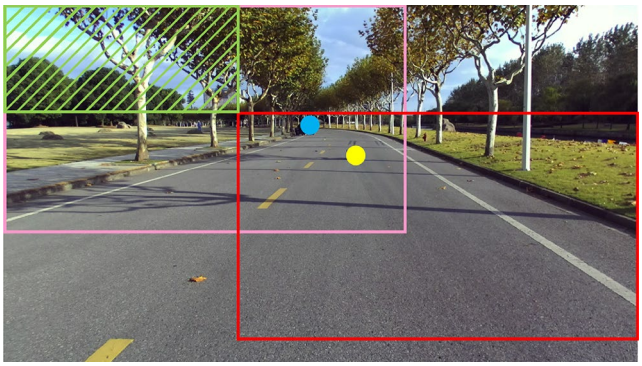

Fig. 9 Illustration of the randomly cropped region in data augmentation

Therefore, as illustrated in Fig. 9, a data augmentation is conducted by randomly cropping original images. The blue circle represents the vanishing point, and the yellow circle represents the object. The pink and red rectangles are the most top-left and bottom-right new images, respectively. The green dash area is the randomly cropped region, from which the top-left point of the cropped image is randomly chosen. This would make the size of images equal to 1344×768 and guarantee that the vanishing points and objects are all included.

5.1.4 Training Details

The VCANet is implemented in PyTorch [57] based on the project mmdetection [58]. ResNet-50 [51] pretrained on ImageNet [59] is used as the backbone unless otherwise stated. Adam [60] is chosen as the optimizer, and the learning rate is set to 10^{-3}. A mini-batch contains 16 images with proposed VCANet trained on 4 GPUs (GTX 1080ti), and the whole training process will be stopped after 10k iterations. The training pipeline of the VCANet is mainly divided into 2 stages: firstly, the VPT prediction block is trained independently and the weights of all other parts are frozen. Secondly, the features fusion block, detection block and upsampling block are jointly trained. Note that the weights of backbone are frozen during the whole training stages.

5.2 VPT Prediction Experiment

The VPT, generated from the VPT prediction block, is used to guide where the ROI is. The ROI is mapped to image and feature maps and generate the ROI image and ROI feature maps, which takes upsampling block as inputs. Thus, the performance of VPT prediction has a great influence on the upsampling block. In order to evaluate whether the VPT prediction is useful to upsampling block, the object in ROI

Table 2 The object in ROI as a percentage of all objects in different size threshold

Object size threshold	8	16	24	32
Number of objects in ROI	186	1523	2703	3337
Number of all annotated objects	224	1754	3010	3703
Percentage	0.830	0.868	0.898	0.901

is measured as a percentage of all objects in different size threshold on TJ-LDRO dataset.

As the results shown in Table 2, more than 90% objects smaller than 32×32 pixels are in the ROI guided by VPT. Although the ROI cloud does not include all the small objects, the current performance is enough for the upsampling block. Consequently, when generating the ground truth for upsampling block, only the objects smaller than 32×32 are adopted while ignoring others.

5.3 Ablation Experiment

In the proposed VCANet, 3 main blocks are of great essence for small object detection: the feature fusion block, the upsampling block and the detection block. The aim of the former 2 blocks is using the low-level detailed information to upsample the high-level feature maps. And the latter one is to use the context-aware method to assist the detection task. As a result, for the purpose of evaluating the influence of the proposed context-aware method and upsampling operation to the performance of VCANet, an ablative analysis is conducted on TJ-LDRO dataset.

5.3.1 Is the Context Information Important for Small Object Detection?

The results of this ablation study on the proposed context-aware method are presented in Tables 3 and 4. In this experiment, d is set to $2r$ and σ is set to $\frac{d}{3}$. In addition, the VCANet without VPT prediction block and upsampling block is used. Four groups of the hyper-parameters p and q are used to evaluate the performance. As discussed in Sect. 4.3.1, p is used to adjust the weight of object and q is used to adjust context environment. $(p = 1, q = 1)$ is used as the baseline parameters and tuned around 1 to compare the differences. Different p and q bring different results, which means that the weights of objects and context of the ground truth have a great influence on the network. As the results shown, $(p = 1, q = 1.5)$ works the best both in DAP and DAR. Therefore, for the next experiment in Sect. 5.3.2, the configuration $(p = 1, q = 1.5)$ is adopted.

Table 3 Ablation study on the proposed context-aware method (Bold numbers indicate the best performance)

Context parameters	mDAP	DAP_4	DAP_8	DAP_S
$p = 0.7, q = 1.0$	0.656	0.498	0.659	0.612
$p = 0.7, q = 1.5$	0.642	0.503	0.649	0.594
$p = 1.0, q = 1.0$	0.658	0.518	0.668	0.627
$p = 1.0, q = 1.5$	**0.673**	**0.548**	**0.683**	**0.641**

Table 4 Ablation study on the proposed context-aware method (Bold numbers indicate the best performance)

Context parameters	mDAR	DAR_S
$p = 0.7, q = 1.0$	0.736	0.727
$p = 0.7, q = 1.5$	**0.742**	**0.735**
$p = 1.0, q = 1.0$	0.732	0.730
$p = 1.0, q = 1.5$	0.735	0.725

Table 5 Ablation study on the upsampling operation (Bold numbers indicate the best performance)

Method	mDAP	DAP_4	DAP_8	DAP_S
B	0.525	0.168	0.514	0.452
B+F	0.673	**0.548**	0.683	0.641
B+F+U	**0.685**	0.540	**0.700**	**0.671**

Table 6 Ablation study on the upsampling operation (Bold numbers indicate the best performance)

Method	mDAR	DAR_S
B	0.629	0.600
B+F	0.735	0.725
B+F+U	**0.736**	**0.738**

5.3.2 Is the Upsampling Operation Important for Small Object Detection?

The results of this ablation study on the upsampling operation are demonstrated in Tables 5 and 6. The upsampling operations on both the feature fusion block and the upsampling block take advantages of the multi-scale feature fusion pipeline, mentioned in Sect. 4.2. Three combinations are selected including the backbone network(B), the feature fusion block(F) and the upsampling block(U) to explore the inference of different blocks. As can be seen, the feature fusion block has a huge impact on small object detection, because the detailed low-level features would be reserved to

Table 7 Comparison of the VCANet with state of the art methods on TJ-LDRO (Bold numbers indicate the best performance)

Method	mDAP	DAP_4	DAP_8	DAP_S	mDAR	DAR_S
Faster-RCNN [8]	0.612	0.553	0.616	0.573	0.702	0.656
Cascade-RCNN [61]	0.585	0.548	0.588	0.544	0.624	0.585
RetinaNet [62]	0.652	**0.585**	0.658	0.616	**0.746**	0.721
FCOS [63]	0.584	0.493	0.592	0.560	0.706	0.679
VCANet	**0.685**	0.540	**0.700**	**0.671**	0.736	**0.738**

Fig. 10 The visual results of our methods

enhance the high-level features. Besides, a further upsampling operation on upsampling block can improve the performance much more, revealing that once the architecture of network is elaborately designed, the network would demonstrate unprecedented potential to encode the features of small objects.

5.4 Compare with the State-of-the-Arts

The proposed VCANet is compared with the state-of-the-arts on the TJ-LDRO dataset, and the results are displayed in Table 7. Four representative models are compared, including Faster-RCNN [8], Cascade-RCNN [61], RetinaNet [62] and FCOS [63], all of which are typical models of two-stage, multi-stage, one-stage and anchor-free object detection, respectively. The implementations of all the models are from mmdetection [58], and no parameters are modified when training and testing. As the results disclose, although FCOS and Cascade-RCNN perform well on generic object detection dataset such as MSCOCO [11], they are not suitable for the small object detection task. However, the RetinaNet demonstrates the best performance in all of the four baseline models. And compared with the RetinaNet in Table 7, the VCANet has an improvement of 3.3% on mDAP and even 5.5% on DAP_S.

The visualization of the results is illustrated in Fig. 10. The first row shows the detection results of five sampled images, in which the yellow and green circles represent the vanishing points and target objects, respectively. The second and third rows exhibit the vanishing point heatmap and center heatmap corresponding to the first row. The last row is the ground truth, in which blue and red circles are the annotated vanishing points and target objects. As shown in Fig. 10, even quite small objects are distinct enough in the heatmap and could be detected by the proposed method.

6 Conclusions

Focusing on the small road hazards detection, this paper is the first attempt to establish a large TJ-LDRO dataset, which consists of 109,337 images from real and virtual simulation environment, labeled in detail. Besides, the vanishing-point-guided context-aware network (VCANet) is introduced, which is an architecture that leverages the vanishing point,

multi-scale features and contextual information for small object detection. This architecture utilizes the VPT prediction block and the context-aware center detection block to obtain more semantic information. Also, the multi-scale feature fusion pipeline and the upsampling block are introduced to gain more ROI feature. Aiming at evaluating the collected dataset and proposed architecture, extensive tests are conducted such as the objects percentage in ROI of different object size, different weights of object and context, different combination of blocks and other variations. Experimental results show that the proposed VCANet achieves an improvement of 3.3% on mDAP and 5.5% on DAPs compared with the state-of-the-art approaches on TJ-LDRO dataset. The future plan is to further extend the self-collected dataset and test the VCANet with autonomous vehicles in real word such as highway scenarios.

Acknowledgements This research has received funding from the National Natural Science Foundation of China (No. 61906138), National Key Research and Development Program of China (No.2016YFB0100901), Shanghai AI Innovative Development Project 2018, and Shanghai Rising Star Program (No. 21QC1400900). We would like to thank Mingyuan Chen for the support of small objects collection in developing the TJ-LDRO dataset.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Pinggera, P., Ramos, S., Gehrig, S., Franke, U., Rother, C., Mester, R.: Lost and found: detecting small road hazards for self-driving vehicles. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (2016)
2. Creusot, C., Munawar, A.: Real-time small obstacle detection on highways using compressive rbm road reconstruction. In: IEEE Intelligent Vehicles Symposium (2015)
3. Leng, J., Liu, Y., Du, D., Zhang, T., Quan, P.: Robust obstacle detection and recognition for driver assistance systems. IEEE Trans. Intell. Transp. Syst. **21**(4), 1560–1571 (2019)
4. Liu, Y., Chen, G., Knoll, A.: Globally optimal vertical direction estimation in atlanta world. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
5. Ramos, S., Gehrig, S., Pinggera, P., Franke, U., Rother, C.: Detecting unexpected obstacles for self-driving cars: Fusing deep learning and geometric modeling. In: IEEE Intelligent Vehicles Symposium (2017)
6. Chen, G., Cao, H., Conradt, J., Tang, H., Rohrbein, F., Knoll, A.: Event-based neuromorphic vision for autonomous driving: a paradigm shift for bio-inspired visual sensing and perception. IEEE Signal Process. Mag. **37**(4), 34–49 (2020). https://doi.org/10.1109/MSP.2020.2985815
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
8. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
9. Wu, Y., Chen, Y., Yuan, L., Liu, Z., Wang, L., Li, H., Fu, Y.: Rethinking classification and localization for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
10. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vision **111**(1), 98–136 (2015)
11. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (2014)
12. Girshick, R.: Fast r-cnn. In: IEEE International Conference on Computer Vision (2015)
13. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision (2016)
14. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
15. Cui, L., Ma, R., Lv, P., Jiang, X., Gao, Z., Zhou, B., Xu, M.: Mdssd: multi-scale deconvolutional single shot detector for small objects. SCIENCE CHINA Inf. Sci. **63**(2), 120113 (2020)
16. Duan, K., Du, D., Qi, H., Huang, Q.: Detecting small objects using a channel-aware deconvolutional network. IEEE Trans. Circuits Syst. Video Technol. **30**(6), 1639–1652 (2019)
17. Liu, Z., Du, J., Tian, F., Wen, J.: Mr-cnn: A multi-scale region-based convolutional neural network for small traffic sign recognition. IEEE Access **7**, 57120–57128 (2019)
18. Bell, S., Lawrence Zitnick, C., Bala, K., Girshick, R.: Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
19. Chen, C., Liu, M.Y., Tuzel, O., Xiao, J.: R-cnn for small object detection. In: Asian Conference on Computer Vision (2016)
20. Bai, Y., Zhang, Y., Ding, M., Ghanem, B.: Sod-mtgan: Small object detection via multi-task generative adversarial network. In: European Conference on Computer Vision (2018)
21. Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., Yan, S.: Perceptual generative adversarial networks for small object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
22. Kembhavi, A., Harwood, D., Davis, L.S.: Vehicle detection using partial least squares. IEEE Trans. Pattern Anal. Mach. Intell. **33**(6), 1250–1265 (2011)
23. Ma, J., Pan, Q., Hu, J., Zhao, C., Guo, Y., Wang, D.: Small object detection with random decision forests. In: IEEE International Conference on Unmanned Systems (2017)
24. Zhang, H., Niu, Y., Zhang, H.: Small target detection based on difference accumulation and gaussian curvature under complex conditions. Infrared Phys. Technol. **87**, 55–64 (2017)
25. Bansal, A., Chen, X., Russell, B., Gupta, A., Ramanan, D.: Pixelnet: Towards a general pixel-level architecture. arXiv preprint arXiv:1609.06694 (2016)
26. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
27. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
28. Fu, C.Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C.: Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659 (2017)

29. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)
30. Cao, G., Xie, X., Yang, W., Liao, Q., Shi, G., Wu, J.: Feature-fused SSD: fast detection for small objects. In: International Conference on Graphic and Image Processing (2018)
31. Hu, G.X., Yang, Z., Hu, L., Huang, L., Han, J.M.: Small object detection with multiscale features. Int. J. Digital Multimedia Broadcast. **2018**,(2018)
32. Liang, Z., Shao, J., Zhang, D., Gao, L.: Small object detection using deep feature pyramid networks. In: Pacific Rim Conference on Multimedia (2018)
33. Liu, Z., Li, D., Ge, S.S., Tian, F.: Small traffic sign detection from large image. Appl. Intell. **50**(1), 1–13 (2020)
34. Bar, M.: Visual objects in context. Nat. Rev. Neurosci. **5**(8), 617 (2004)
35. Biederman, I.: Perceiving real-world scenes. Science **177**(4043), 77–80 (1972)
36. Hu, P., Ramanan, D.: Finding tiny faces. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
37. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
38. Torralba, A.: Contextual priming for object detection. Int. J. Comput. Vis. **53**(2), 169–191 (2003)
39. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: European Conference on Computer Vision Workshops (2018)
40. Yuan, Y., Xiong, Z., Wang, Q.: Vssa-net: vertical spatial sequence attention network for traffic sign detection. IEEE Trans. Image Process. **28**(7), 3423–3434 (2019)
41. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A.G., Ma, H., Fidler, S., Urtasun, R.: 3d object proposals for accurate object class detection. In: Advances in Neural Information Processing Systems (2015)
42. Bosquet, B., Mucientes, M., Brea, V.M.: STDnet: a convnet for small target detection. In: British Machine Vision Conference (2018)
43. Guan, T., Zhu, H.: Atrous faster r-cnn for small scale object detection. In: International Conference on Multimedia and Image Processing (2017)
44. Fang, L., Zhao, X., Zhang, S.: Small-objectness sensitive detection based on shifted single shot detector. Multimedia Tools Appl. **78**(10), 13227–13245 (2019)
45. Meng, Z., Fan, X., Chen, X., Chen, M., Tong, Y.: Detecting small signs from large images. In: International Conference on Information Reuse and Integration (2017)
46. Zhang, R., Yin, D., Ding, J., Luo, Y., Liu, W., Yuan, M., Zhu, C., Zhou, Z.: A detection method for low-pixel ratio object. Multimedia Tools Appl. **78**(9), 11655–11674 (2019)
47. Zhu, Z., Liang, D., Zhang, S., Huang, X., Li, B., Hu, S.: Traffic-sign detection and classification in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
48. Xiao, J., Ehinger, K.A., Hays, J., Torralba, A., Oliva, A.: Sun database: exploring a large collection of scene categories. Int. J. Comput. Vis. **119**(1), 3–22 (2016)
49. Shah, S., Dey, D., Lovett, C., Kapoor, A.: Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In: Hutter, M., Siegwart, R. (eds) Field and Service Robotics. Springer Proceedings in Advanced Robotics, vol 5. Springer, Cham.
50. Qiu, W., Yuille, A.: Unrealcv: Connecting computer vision to unreal engine. In: European Conference on Computer Vision (2016)
51. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
52. Lee, S., Kim, J., Shin Yoon, J., Shin, S., Bailo, O., Kim, N., Lee, T.H., Seok Hong, H., Han, S.H., So Kweon, I.: Vpgnet: Vanishing point guided network for lane and road marking detection and recognition. In: IEEE International Conference on Computer Vision (2017)
53. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: The European Conference on Computer Vision (2018)
54. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
55. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579 (2015)
56. Guan, L., Wu, Y., Zhao, J.: Scan: semantic context aware network for accurate small object detection. Int. J. Comput. Intell. Syst. **11**(1), 951–961 (2018)
57. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
58. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
59. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
60. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015)
61. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
62. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
63. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: IEEE International Conference on Computer Vision (2019)