IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 16, 2019, accepted April 8, 2019, date of publication April 18, 2019, date of current version April 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2911562

CDBV: A Driving Dataset With Chinese Characteristics From a Bike View

YING HE[1,2], HAN-BO YANG[3], AND SU-JING WANG[1], (Member, IEEE)
[1]Key Laboratory of Behavior Sciences, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
[2]College of Information Engineering, North China University of Science and Technology, Tangshan 063210, China
[3]Monash University-Southeast University Joint Research Institute (Suzhou), Southeast University, Nanjing 211189, China

Corresponding author: Su-Jing Wang (wangsujing@psych.ac.cn)

This paper is supported in part by grants from the National Natural Science Foundation of China (61772511) and in part by the Director Fund of National Engineering Laboratory for Public Security Risk Perception and Control by Big Data (18112403).

ABSTRACT Public datasets are valuable for the development of the learning-based algorithms. In the field of autonomous driving, almost all the existing public datasets are collected by manually driving cars, and most of them are contributed by American and European researchers and represent the traffic scenarios typical for these countries. To diversify the public driving datasets and to provide data support for researches on the self-driving vehicles moving on bike lanes, such as wheelchairs, bikes, and some robots, we present the Chinese Driving from a Bike View (CDBV), which is a driving dataset collected by a bike in Beijing, China. The CDBV dataset contains annotated images, raw videos, and code tools. The annotations are for the object detection tasks and especially focus on the categories with Chinese characteristics, including the express tricycles, the motorbikes specializing in takeouts, the bicycle landmarks, the zebra crossings, and the traffic lights. We totally annotated 40 064 traffic elements in 13 427 images. The CDBV dataset is publicly available at http://sujingwang.name/CDBV.html.

INDEX TERMS Driving dataset, China, autonomous driving, self-driving wheelchair, collected by a bike, object detection.

I. INTRODUCTION

The autonomous driving technology has been attracting much attention especially for the recent decades. With the development of the artificial intelligence and machine learning, in recent years, learning-based approaches are more prevailing than model-based approaches in many areas. Especially for the deep learning approach, it not only continually becomes the state of the art in lots of computer vision tasks, but also penetrates into the field of autonomous driving and shows its great potential [1]. For example, [2] presents an end-to-end method to produce a generic vehicle motion model by learning from large scale video datasets, and it reveals very good performance. Many other researches, such as [3]–[5], have also explored the application of deep learning methods in autonomous driving. Besides the end-to-end model learning, there are many other sub-tasks in the field of autonomous driving, such as object detection, lane following, semantic understanding, SLAM (simultaneous localization and mapping), which are also well accomplished by learning methods [6], [7].

Due to the prevalence of learning algorithms, the datasets used for learning have become very critical resources. The autonomous driving technology has great commercial value, and the datasets are often the key factor that restricts the performance of algorithms. Therefore, most datasets are private, and publicly available datasets are limited. Yin and Berger [8] endeavored to collect all the openly available driving datasets until 2017 and they have only collected 27 relevant datasets. These datasets include the KITTI Vision Benchmark Suite [9] and its other versions [10]–[12], the Cityscape dataset [13], the comma.ai driving dataset [14], the DIPLECS autonomous driving datasets [15], the automotive multi-sensor dataset (AMUSE) [16], the Daimler Pedestrian Benchmarks, the Karlsruhe dataset [17], the Dr(eye)ve [18], the Ground Truth Stixel dataset (Stixel) [19], etc. More recently, the BDDV dataset is presented [20], which provides more driving videos with diverse kinds of annotations for researches. Besides, some visual perception tasks like pedestrian detection, traffic light detection and traffic sign detection

The associate editor coordinating the review of this manuscript and approving it for publication was Shirui Pan.

etc are also the classic tasks of autonomous driving, so there are datasets dedicated to these tasks specially, such as the Caltech Pedestrian Detection Benchmark [21], the Bosch Small Traffic Lights Dataset [7] and the German Traffic Sign Detection Benchmark [22]. Some general object detection datasets like the Pascal VOC [23] also contain some annotated objects related to the autonomous driving, such as cars, buses, bikes and motorbikes.

However, most attentions are paid to the autonomous technology of cars, but many other vehicles that also need the technology have been neglected. As a result, almost all the existing public driving datasets are captured by cars equipped with a variety of sensors, which causes the datasets are all observed from a car view. Nevertheless, there are other kinds of vehicles on the road, in which the applications of the autonomous driving technology are also of great value. For example, a self-driving wheelchair will facilitate the disabled and the old people who have lost part of their abilities, and there exist several researches on it [24]–[26]. For another example, if we create a self-driving bike, it will help people a lot due to the flexibility. And robotics is an important research field, in which some types of robots also require the autonomous driving technology. Different from the cars, the vehicles like these usually go on the bike lanes rather than on the motor way, and their normal speed is much lower than that of the cars, which leads that the scenes they see are different from what cars see. Datasets collected from their own views are necessary but very scarce for relevant researches.

Meanwhile, the diversity of datasets is important for a robust driving learning model. There has been enough diversity of weather conditions, lights, scenes, sensors, etc in the existing driving datasets. But as for countries, most public datasets are contributed by American and European researchers and represent the traffic scenarios typical for these countries, which only cover a tiny portion of the world map [8]. For many other countries, their contributions to public driving datasets are insufficient, like China, the largest developing country. As far as we know, among the public driving datasets, there are only two containing scenarios captured in China. One is the Daimler Pedestrian Benchmarks provided by German researchers, which is used for pedestrian-related vision tasks. Another is the ApolloScape [27], a large-scale dataset released by Baidu Research most recently, which is collected by car platforms. But obvious discrepancy exists between traffic conditions in different countries and from different views. For example, China's traffic regulations stipulate that vehicles must run on the right, which is contrary to the regulations of Britain, Japan, Australia and many other countries. For another example, some Chinese characteristic vehicles like express tricycles and takeout motorbikes usually run faster than bikes, which brings about different scenarios seen by bikes and by cars. Therefore, a driving dataset collected in China and observed from a distinctive view is meaningful for dataset diversity.

It is in the above context that we present the CDBV (Chinese Driving from a Bike View), which is a novel public driving dataset with Chinese characteristics from a bike view. The main original contributions of this paper are as follows:

1) The CDBV dataset has been recorded from a moving bike with a common camera, which differs from previous datasets. It provides more targeted data for researches on autonomous vehicles running on bike lanes.
2) Our presented videos and images are captured in China, and thus they reflect the Chinese traffic conditions. More importantly, we provide a large number of annotations (annotate 40064 objects in 13427 images) for object detection tasks, and we especially annotate the traffic elements with scarce annotations or with Chinese characteristics, which aren't contained by previous datasets.
3) We report experimental results of several state-of-the-art detection algorithms on the CDBV dataset, which provides baselines for other researchers to refer to.

II. DATA COLLECTION

In the CDBV dataset, source data consist of two parts: driving videos and annotated images. We captured and prepared them in Beijing, China, from July 10, 2018 to October 6, 2018. That is, the seasons varied from summer to autumn in China, the Northern Hemisphere country.

A. VIDEO CAPTURE

Different from the previous datasets, the CDBV aims at providing driving data for the autonomous vehicles like wheelchairs, bikes, some robots, etc, which usually run on bike lanes. These kinds of vehicles are smaller, slower, cheaper and more flexible than cars generally. We choose the bike representing such vehicles as the moving platform. And as for sensors, we choose the common camera to capture driving videos, considering visual perception is relatively cheap, more informative and practical for such autonomous vehicles. Our recording platform is shown in Fig.1. The camera is 130 centimeters above the ground.

For the place of the video shoot, we choose the city of Beijing, the capital of China. We choose this city mainly for the following two reasons. First, Beijing can represent most large and medium-sized cities in China, in which public driving datasets from a bike view are relatively scarce. Traffic conditions in these cities share common characteristics. For example, all these areas have the same traffic regulations, mainstream vehicles, road conditions, types of traffic lights, types of traffic signs and traffic rush hours. Driving datasets collected in Beijing can reflect the traffic conditions in most of China. Second, most demand for autonomous driving technology in China is in this kind of cities. It is because these cities have a higher level of economic development, have more perfect laws and regulations, and cover a larger area. Our shooting zone is shown in Fig.2, which is part of the Beijing map. The zone with red color is where we captured

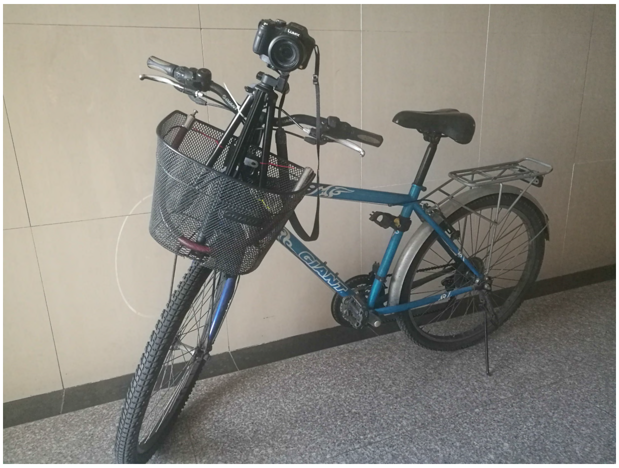


FIGURE 1. The data recording platform: a bike with a common camera.

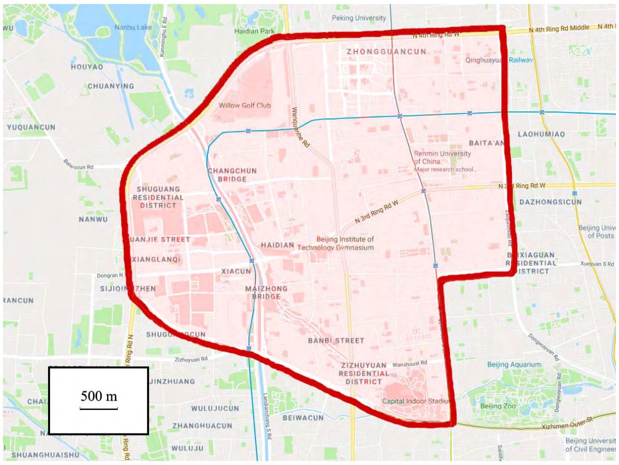


FIGURE 2. The shooting zone: a part of Beijing.

the videos in the three months, covering about 22 square kilometers in Beijing, and including various streets, residential areas and shopping malls.

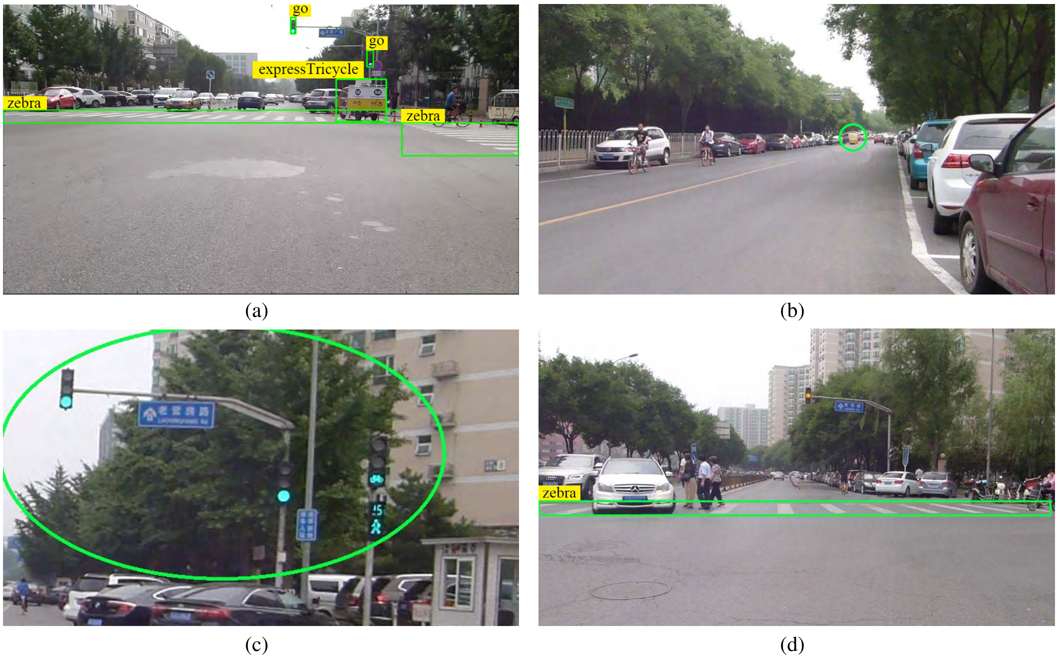
For other aspects of the traffic conditions, we take full account of the diversity of the dataset, so we shot the videos in different kinds of weather (sunny, cloudy, after rain), in different periods of a day (morning, middle of the day, afternoon, rush hours when people commute), and in different light conditions (lighting from the front, lighting from behind). The diversity is necessary for a learning algorithm to get and test the robustness of the model.

To get the natural videos that reflect the scenarios when human beings use the bike normally, we decided routes first, and rode the bike from the sources to the destinations naturally, as if we were on an ordinary journey instead of collecting data. Our normal riding speed was about 12 km/h, and we might speed up, slow down, stop, turn and etc according to the actual traffic situations. In this way, we captured several videos which reflect the actual driving in China from a bike view. We call these videos ‘‘long videos’’ and put them into a folder named ‘‘long videos’’ in the dataset. The lengths of the videos are more than ten minutes and less than one hour. The traffic elements in these long videos obey the actual distribution in reality. However, in the actual distribution, the two kinds of vehicles: express tricycles and motorbikes specializing in takeouts, which are the important traffic elements of modern China, account for a small proportion. To increase the samples of them, we shot some short videos specially. It is when we met them that we turned on the camera, rode the bike and shot videos containing them. We call these videos ‘‘short videos’’ and put them into a folder named ‘‘short videos’’ in the dataset. Meanwhile, we also put the other relatively short driving videos into the set of ‘‘short videos’’. The lengths of the short videos are less than ten minutes and most of them are less than one minute.

B. ANNOTATE IMAGES

The above is how we captured the videos in CDBV, including two kinds: ‘‘long videos’’ and ‘‘short videos’’. Besides, object detection is a classic task in both computer vision and autonomous driving. Adequate annotated images are necessary and important for the researches. Our CDBV dataset provides a number of detection-related annotations for the driving images. Considering the traffic elements like pedestrians, signs, cars, buses, bikes and etc have been well annotated in the existing datasets [21]–[23], we focus on other elements with scarce annotations or with Chinese characteristics. We choose the six categories to annotate: the express tricycles, the motorbikes specializing in takeouts, the bicycle landmarks, the zebra crossings, the green traffic lights and the red traffic lights. We put the detailed description of these categories in Section III-A, and here we introduce how we annotated them.

The images come from the captured videos described in Section II-A. First, we watched the videos one by one. When we found a video clip that contains the objects to be annotated, we took images of this video clip at ten-second intervals. Thus we got all the images to be annotated in the driving videos. Second, we designed the annotating protocols. The protocols include: (1) annotate all the target objects belonging to the six categories in every image, marking the bounding box that exactly frames the object, and giving which category the object belongs to (e.g., Fig.3(a) is a standard annotated image); (2) do not annotate the object that cannot be seen clearly and cannot be categorized surely (e.g. the vehicle in the green circle in Fig.3(b)); (3) all kinds of red and green traffic lights should be annotated, including the motor vehicle traffic lights, the non-motor vehicle traffic lights and the pedestrian crossing traffic lights (e.g., the four traffic lights in the green circle in Fig.3(c) all should be annotated); (4) when a target object is occluded, draw the bounding box according to the following rules: if one side of the target is completely occluded but another side is not, only frame the visible side; if a part of the target is occluded but the whole contour is not, frame the whole target with occlusion; if a zebra crossing is cut into two segments by the occlusion, we still regard



(a) (b) (c) (d)

FIGURE 3. Examples about the annotating protocols: (a) a standard annotated image; (b) do not annotate the vehicle in the green circle because it cannot be categorized surely; (c) the four traffic lights in the green circle in all should be annotated; (d) the zebra crossing occluded by the car should be framed by one bounding box.

it as one target and draw one bounding box framing the two segments (e.g. Fig.3(d)). Thirdly, after the protocols had been designed, we started to organize personnels to annotate the images. We trained them to use annotating software and to learn the annotating protocols, and then they annotated all the images, paying much industrious work. Fourthly and lastly, we checked all the annotated images and picked out the non-standard annotations. Then the non-standard annotations were revised correctly by relevant personnels.

The above is how we prepared the annotated images. We totally annotated 13427 images including 40064 bounding boxes. We separate the images into two sets that correspond to the ones coming from the "long videos" and the ones coming from the "short videos" respectively, in which the traffic elements obey two different distributions.

III. DATASET

We provide the CDBV dataset publicly, in which there are images, annotations (mainly for object detection tasks), raw videos and code tools. The total size of the dataset is 26.2 GB, in which images and annotations take up 8.2 GB, and raw videos take up 18 GB. All of them reflect the Chinese driving from a bike view.

A. ANNOTATED IMAGES

We annotated six categories of traffic elements. They are express tricycles, motorbikes specializing in takeouts, bicycle landmarks, zebra crossings, green traffic lights and red traffic lights, marked with "expressTricycle", "takeoutMotorbike", "bicycleLandmark", "zebraCrossing", "go" and "stop" as names in annotations respectively.

Here we introduce the six categories and our original contributions about them first. With the development of China, express delivery and takeout industries have been rising and thriving. As a result, express tricycles and motorbikes specializing in takeouts have become two kinds of characteristic and common vehicles in modern China. An express tricycle is a vehicle like the one shown in Fig.4(a). This kind of vehicle always has three wheels, a driver's cab and a packing box with designs and logos, though they may vary in style due to different companies. As far as we know, there have been no datasets providing annotations for this kind of vehicles except our CDBV, although more general tricycles that cover all kinds of three-wheeled vehicles have been annotated most recently [27]. A motorbike specializing in takeouts is like the one in Fig.4(b). It always has a storage box used for containing takeouts, which is different from the general motorbikes in existing datasets like the Pascal VOC [23]. This requires the algorithms to learn relevant concepts and features to discriminate the two kinds of motorbikes. Road marking detection is a classic task of autonomous driving. The relevant datasets are very scarce because a road marking usually has a single color and traditional image processing and computer vision methods are usually used. But the actual road environment is complex and changeable, which results in that most of the traditional methods are only suitable for the ideal situation. Therefore, learning methods and relevant datasets are necessary. To solve the issue of scarcity, our CDBV provides annotations for two kinds of road markings: bicycle

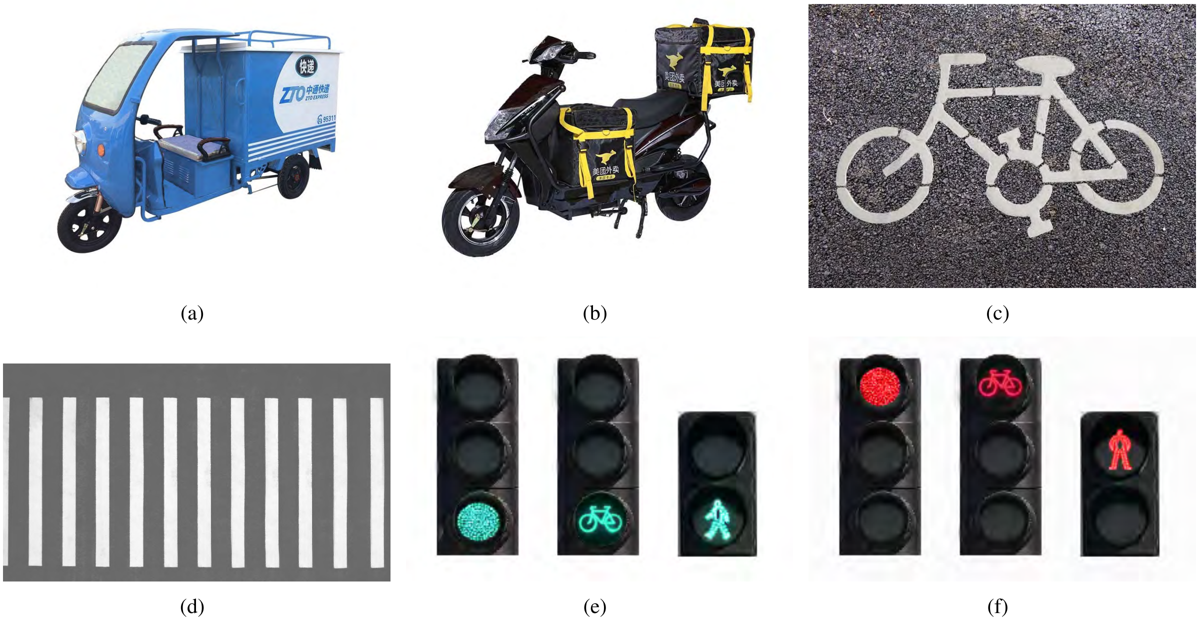


FIGURE 4. Instances of the six annotated categories in CDBV: (a) an express tricycle; (b) a motorbike specializing in takeouts; (c) a bicycle landmark; (d) a zebra crossing; (e) green traffic lights; (f) red traffic lights.

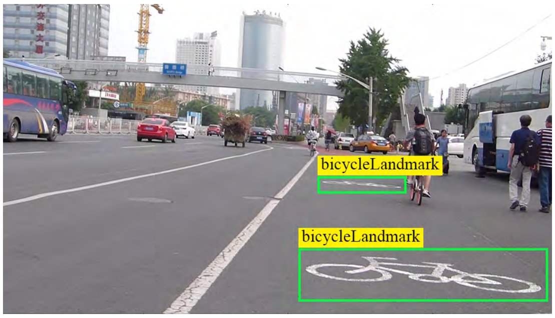


FIGURE 5. An image with bicycle landmarks in CDBV.

landmarks (see Fig.4(c)) and zebra crossings (see Fig.4(d)). The Fig.5 is an image with bicycle landmarks in the CDBV dataset (the green annotations are added only in this paper for readers). From it we can see the challenges for detection: the change of the view leads to the change of visual appearance, and the detection of the far away bicycle landmark with blurry contours even requires the algorithm to have the ability of logical reasoning from the whole image and traffic situation. Two other annotated categories in CDBV are green traffic lights (three instances in Fig.4(e)) and red traffic lights (three instances in Fig.4(f)). We annotated all kinds of traffic lights including the ones for motor vehicles, non-motor vehicles and pedestrians. Although there exist several datasets providing annotations for traffic lights, their images aren't captured in China, and so our CDBV is meaningful for enhancing diversity. We can see the differences in styles by comparing the Fig.3(c) which is from our CDBV and the Fig.6 which is from the Traffic Lights Recognition (TLR) public dataset.

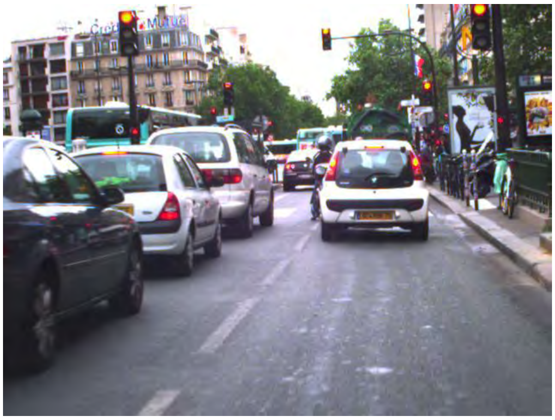


FIGURE 6. An image in the Traffic Lights Recognition (TLR) public dataset.

We provide 13427 annotated images including 40064 annotated traffic elements in total. The images are in the JPG format. The size of each image is 1920×1080. The annotations mainly provide the information of the category and the coordinates of the upper left vertex and the lower right vertex of the bounding box for every traffic element. The annotations are provided in three different formats: TXT, XLSX and XML, for convenience of different needs. Images and annotations are divided into two sets, respectively corresponding to the "long videos" and the "short videos" described in Section II-A. In the first image set, there are 9183 images and 28228 annotated traffic elements in total. The distribution of categories is illustrated by the bar graph in the Fig.7(a). It is the natural distribution in the actual traffic

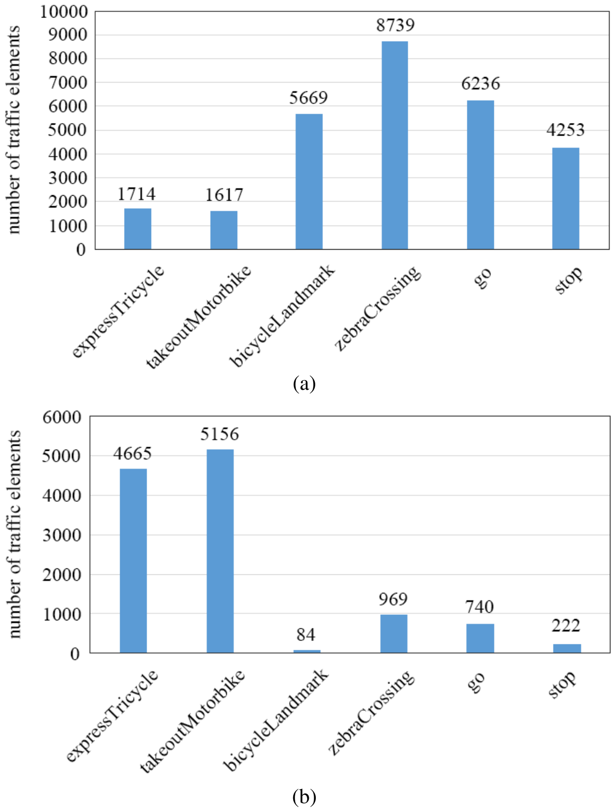


FIGURE 7. **The distributions of traffic elements: (a) in the first image set; (b) in the second image set.**

conditions. In the second image set, there are 4244 images and 11836 annotated traffic elements in total. The distribution of categories is illustrated by the bar graph in the Fig.7(b). It isn't the actual distribution because we consciously captured more express tricycles and motorbikes specializing in takeouts in order to balance the samples of each category in the whole dataset.

B. RAW VIDEOS

Raw data without annotations are also valuable resources for autonomous driving researches. There are several public driving datasets providing only raw data, such as CCSAD [28], Cheddar Gorge [29], EISATS [30] and Heidelberg [31]. Therefore, our CDBV dataset releases the raw driving videos for relevant researches. Compared with the images, the videos contain more information about driving, such as the vehicle speed and the driving behaviors, with the time dimension. Raw videos can be used for various tasks like traffic video prediction (refer to [14]), driving model learning and some unsupervised learning tasks. More importantly, our CDBV aims at autonomous driving technology of the vehicles on the bike lane, so the videos in CDBV are observed from the perspective of a bike, which makes them contain some characteristics that other datasets do not have.

Now we introduce some characteristics of the videos in CDBV. Captured in China, the videos reflect the actual Chinese traffic situations nowadays. And from a bike view, there are many characteristic scenes that other driving datasets do not have. Some examples are in the Fig.8. Fig.8(a) is the scene of waiting to cross the road on the bike lane. Some large intersections will have such lanes. Fig.8(b) is the scene of crossing the road through the zebra crossing. The vehicles such as wheelchairs and bikes need to cross the road through the zebra crossing at relatively small intersections without bike lanes. Fig.8(c) is the scene seen when we rode the bike ahead on the road. Bike lanes are the special lanes for such vehicles. The Chinese traffic regulations require us to run on the right lanes, so the videos in our CDBV contain a lot of such scenes: going ahead on the bike lane, with running cars on the motor vehicle lane on the left, and with parked cars in the parking space (if there is) on the right, or with the sidewalk with curbs on the right (if there isn't parking space on the road). All the scenes like above cannot appear in the videos captured by cars generally. Fig.8(d) is the scene when we used the bike to travel in the mall flexibly. It is also a common scene in which people use bikes or wheelchairs, but not cars. In addition to traffic scenes, there are various actual driving behaviors reflected from the driving videos, among which we select several characteristic behaviors in CDBV shown in the Fig.9. The four frames, Fig.9(a)-Fig.9(d), are taken from a CDBV video at a certain time interval. We can see the corresponding video clip reveals the driving behavior of lane selection: selecting a bike lane when crossing the intersection. It is unlikely to happen in the videos captured by cars. The frames Fig.9(e)-Fig.9(h) reveal the behavior of crossing the narrow gap between cars. If it had been a car that saw the scene of the Fig.9(e), the next driving behavior would have been "stop to wait for the front cars to drive away". But it was a bike that occupied very little space and had great flexibility, so its next behavior was "cross the narrow gap between the front cars", which is reflected in the next frames of the video. Similarly, the frames Fig.9(i)-Fig.9(l) reflect how the vehicle behaved when it met a group of people blocking the way ahead. A car might have stopped, but we can see the bike chose to bypass the crowd through the right side. Besides, bypassing the water on the road is also an interesting driving behavior reflected in our CDBV video, as is shown in the frames Fig.9(m)-Fig.9(p). In the videos, there are many other typical examples, including those that can or cannot be perceived by human beings. All in all, the driving videos contain abundant driving scenes, driving behaviors, and other driving information. They were captured from the view of a bike in the actual Chinese environment, which gives novel challenges to relevant algorithms.

We provide 47 raw videos taking up 18 GB of storage in total. These videos are divided into two sets: "long videos" and "short videos", corresponding to the description in Section II-A. In the "long videos" set, there are 5 long videos taking up 14.1 GB. And in the "short videos" set, there are 42 short videos taking up 3.9 GB. All videos are in the MP4 format. They are driving videos shot naturally by a common camera on a bike, in Beijing, China.

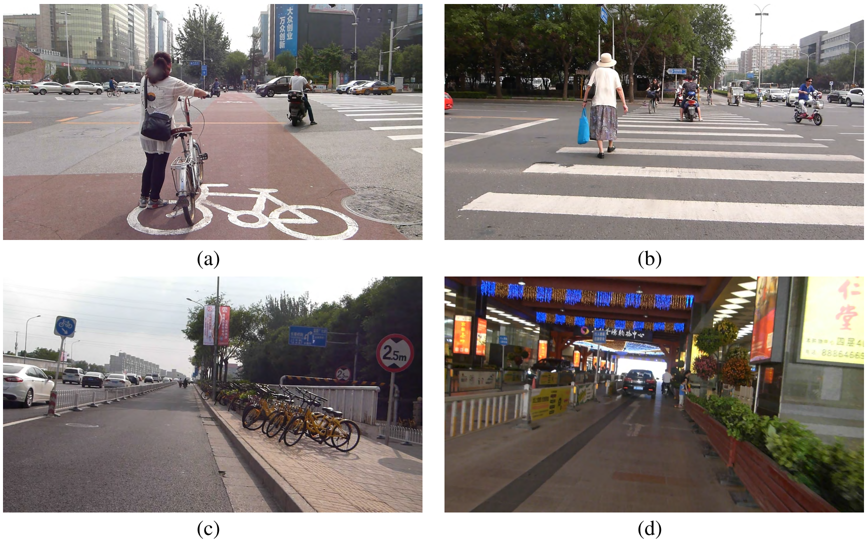


FIGURE 8. Some characteristic scenes from a bike view: (a) waiting to cross the road on the bike lane; (b) crossing the road through the zebra crossing; (c) driving ahead on the bike lane; (d) flexibly traveling in the mall.

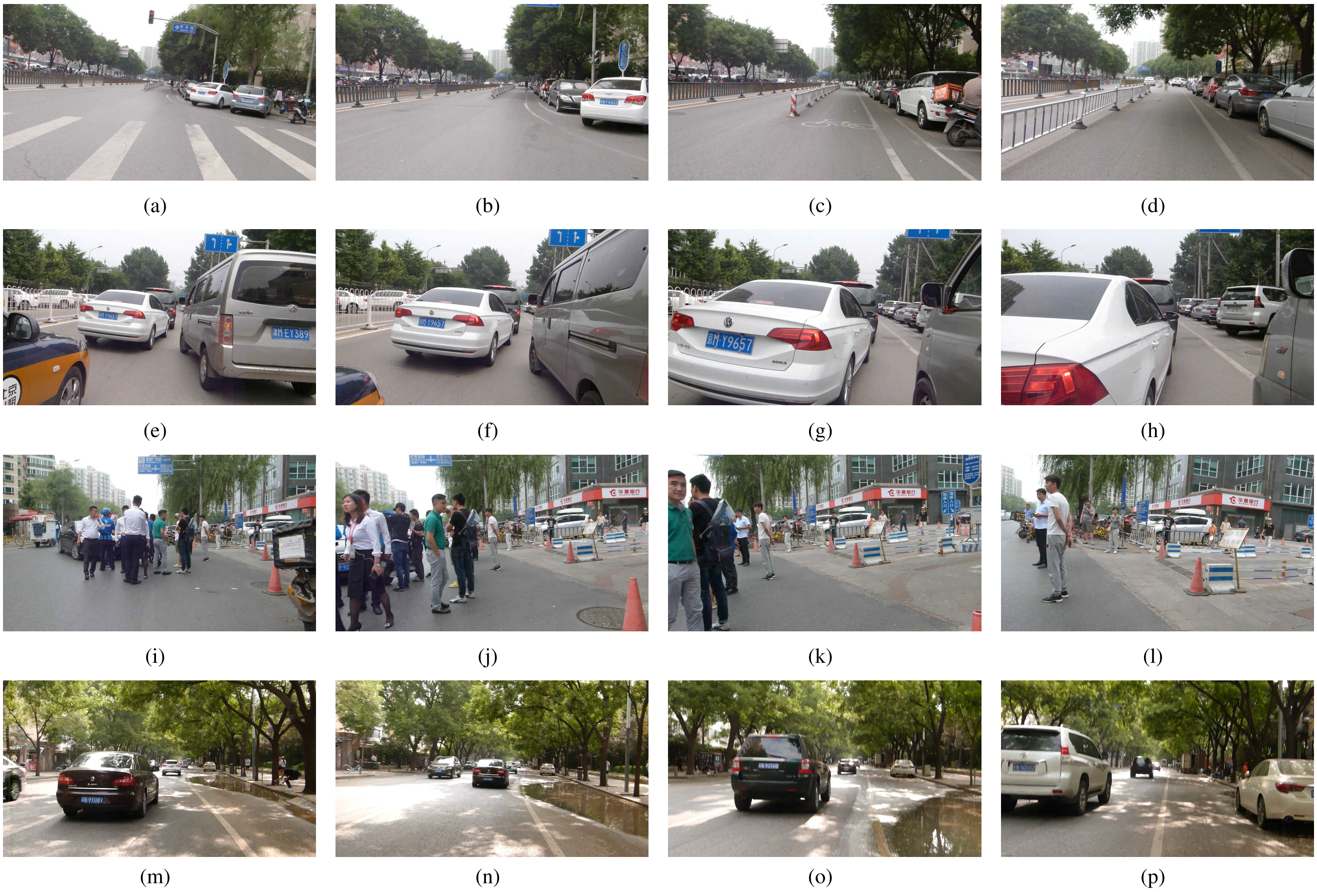


FIGURE 9. Some characteristic video clips from a bike view: (a)-(d) shows the behavior of lane selection: selecting a bike lane; (e)-(h) shows the behavior of crossing the narrow gap between cars; (i)-(l) shows the behavior of bypassing the crowd; (m)-(p) shows the behavior of bypassing the water on the road.

C. CODE TOOLS

We provide code tools to facilitate researchers to use the CDBV dataset. The file “txt_to_VOCxml.m” (MATLAB code) is used to convert the annotation file from the TXT format to the XML format. The file “split_train_val_test.py” (PYTHON code) is used to split the annotated images into the training set, the validation set and the test set according to the ratio set by users. The file “draw_picture_boxes.py”

TABLE 1. The details of the training set and the test set in our experiments.

dataset	image	bounding box	expressTricycle	takeoutMotorbike	bicycleLandmark	zebraCrossing	go	stop
training set	8057	24004	3835	4002	3437	5852	4178	2700
test set	5370	16060	2544	2771	2316	3856	2798	1775

TABLE 2. Quantitative results of CDBV baselines for object detection using several state-of-the-art methods.

method	mAP	expressTricycle	takeoutMotorbike	bicycleLandmark	zebraCrossing	go	stop
Faster R-CNN (ZF) [33]	59.6	80.2	74.8	68.1	72.5	36.1	25.9
Faster R-CNN (VGG) [33]	73.1	90.2	80.5	78.4	86.8	57.2	45.2
R-FCN (ResNet50) [34]	63.2	77.4	67.1	70.6	63.1	56.0	44.9
R-FCN (ResNet101) [34]	55.2	68.1	55.1	66.3	54.1	51.7	36.1
SSD300 [35]	65.7	86.1	76.9	84.5	82.3	38.1	26.4
SSD512 [35]	84.2	89.9	86.1	88.6	86.3	81.5	73.0
YOLOv3-320 [36]	79.0	93.4	87.1	84.7	85.0	66.7	57.4
YOLOv3-416 [36]	87.8	95.8	91.0	90.5	89.3	83.7	76.6
YOLOv3-608 [36]	91.5	95.9	90.9	92.8	90.1	91.6	87.9

(PYTHON code) is used to display an annotated image with bounding boxes and categories drawn, facilitating observing the image and annotations intuitively.

IV. BASELINES

We have done baseline experiments on the annotated images in our CDBV dataset, relying on several state-of-the-art methods. In our experiments, the whole annotated images are split into the training set and the test set at the ratio of 6:4. More details about the number of images, total bounding boxes and traffic elements per category in the training and test sets are shown in the Table.1.

Deep learning methods have shown great potential in the object detection task in recent years. They have achieved far better results than traditional methods and significantly advanced the state of the art. These deep learning detectors can be divided into two types: proposal-based detectors (two-stage detectors) and proposal-free detectors (one-stage detectors). The representative proposal-based detectors include R-CNN [36], Fast R-CNN [37], SPPnet [38], Faster R-CNN [32], R-FCN [33], Mask R-CNN [39], etc. And the representative proposal-free detectors include OverFeat [40], SSD [34], YOLOv3 [35], RetinaNet [41], etc. In our baseline experiments, we adopt the Faster R-CNN, R-FCN, SSD and YOLOv3 which can represent the state-of-the-art detectors running on our CDBV datasets, and we report their mAP (Mean Average Precision) and AP (Average Precision) of each category which are the common metrics to assess the detector performance.

The experiment results are given in the Table.2, providing the baselines on CDBV for other researches to refer to. Experiment details are introduced as follows: For all detectors, we did 120000 iterations on the training set and snapshot the model every 10000 iterations. We tried to utilize the default setup of the detectors, and modified hyper parameters only when necessary. All modified setup will be introduced next. We tested all the 12 snapshot models on the test set and took the best result as the final result for every detector. For Faster R-CNN, we chose the optimization way of approximate joint training, and we did two experiments using the ZF net [42] and the VGG16 net [43] as the feature extractor respectively. For R-FCN, we chose the way of end-to-end training with online hard example mining [44], and we did two experiments using the ResNet50 and the ResNet101 [45] as the feature extractor respectively. For the R-FCN using ResNet50, we modified the settings of "stepsize" and "iter_size" to 50000 and 4. For the R-FCN using ResNet101, we modified the settings of "base_lr", "stepsize" and "iter_size" to 0.0005, 50000 and 4. For SSD, we did experiments with its two versions: SSD300 and SSD512, which correspond to two different sizes of image inputs, and we modified the setting of "base_lr" to 0.0001 to avoid exploding gradients. For YOLOv3, we did experiments with its three versions: YOLOv3-320, YOLOv3-416 and YOLOv3-608, which correspond to the input image sizes of 320×320, 416×416 and 608×608, and we modified the settings of "batch", "subdivisions" and "steps" to 64, 16 and "40000, 80000".

From studying the results in the Table.2, we can get some findings: (1) The AP of express tricycles is much higher than the AP of any other category significantly for all detectors. (2) The AP of traffic lights is much lower than the AP of any other category significantly except for the YOLOv3-608 detector. And the AP of red traffic lights is the lowest for all detectors. It dramatically lower the value of mAP. Comparing the results of different versions of SSD and YOLOv3, we can find that the AP of traffic lights is greatly improved with the increase of the input sizes. It can be concluded that the operation of resizing input images results in serious loss of the traffic light information. It may be because traffic lights belong to small target objects. In addition, there are some other possible reasons for the above phenomena, such as the insufficient number of samples (from the Fig.7 we can see that the number of red traffic lights is the smallest), the lack of pre-training for the categories in the feature extractor, and so on. (3) Among all the detectors in our experiments, the YOLOv3-608 achieves the highest mAP, and it has very excellent performance (although it's at the cost of memory and computation).

Overall, the experiment results unveil the performance of several state-of-the-art detectors on CDBV, providing baselines for other researches. The categories annotated in CDBV are unique or with Chinese characteristics, which provides novel support for relevant detection researches.

V. CONCLUSION

We propose a novel driving dataset, CDBV, which is freely available for academic researches. The significance of CDBV is all the more apparent from three aspects. First, we annotate a large number of traffic elements for object detection tasks. Most of these categories aren't annotated by previous public datasets. It addresses the lack of relevant annotations. Second, all the driving videos and images are captured by a common camera on a running bike. So CDBV provides more targeted data for researches on the autonomous vehicles running on bike lanes. Third, CDBV is collected in Beijing, China, which reflects the Chinese traffic conditions. It enhances the diversity of driving datasets in terms of countries. Finally, baseline experiments are done to test how the state-of-the-art detectors perform on the CDBV dataset. We hope that CDBV can complement other datasets and contribute to the development of autonomous driving technology.

REFERENCES

[1] J. Janai, F. Güney, A. Behl, and A. Geiger. (2017). "Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art." [Online]. Available: *arXiv preprint*, 2017.

[2] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3530–3538.

[3] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2722–2730.

[4] M. Bojarski *et al.* (2016). "End to end learning for self-driving cars." [Online]. Available: https://arxiv.org/abs/1604.07316

[5] E. Santana and G. Hotz. (2016). "Learning a driving simulator." [Online]. Available: https://arxiv.org/abs/1608.01230

[6] Y. Zhu, C. Zhang, D. Zhou, X. Wang, X. Bai, and W. Liu, "Traffic sign detection and recognition using fully convolutional network guided proposals," *Neurocomputing*, vol. 214, pp. 758–766, Nov. 2016.

[7] K. Behrendt, L. Novak, and R. Botros, "A deep learning approach to traffic lights: Detection, tracking, and classification," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2017, pp. 1370–1377.

[8] H. Yin and C. Berger, "When to use what data set for your self-driving car algorithm: An overview of publicly available driving datasets," in *Proc. Intell. Transp. Syst. Conf.*, Oct. 2017, pp. 1–8.

[9] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[10] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[11] J. Fritsch, T. Andreas, and K. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *Proc. Int. Conf. Intell. Transp. Syst.*, Mar. 2013, pp. 1693–1700.

[12] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3061–3070.

[13] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3213–3223.

[14] E. Santana and G. Hotz'. (2016). "Learning a driving simulator." [Online]. Available: https://arxiv.org/abs/1608.01230

[15] N. Pugeault and R. Bowden, "How much of driving is pre-attentive?" *IEEE Trans. Veh. Technol.*, vol. 64, no. 12, pp. 5424–5438, Jun. 2015.

[16] P. Koschorrek, T. Piccini, P. Oberg, M. Felsberg, L. Nielsen, and R. Mester, "A multi-sensor traffic scene dataset with omnidirectional video," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 727–734.

[17] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *Proc. Intell. Vehicles Symp.*, Jun. 2011, pp. 963–968.

[18] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara, "DR (eye) VE: A Dataset for attention-based tasks with applications to autonomous and assisted driving," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2016, pp. 54–60.

[19] D. Pfeiffer, S. Gehrig, and N. Schneider, "Exploiting the power of stereo confidences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Mar. 2013, pp. 297–304.

[20] F. Yu *et al.* (2018). "BDD100K: A diverse driving video database with scalable annotation tooling." [Online]. Available: https://arxiv.org/abs/1805.04687

[21] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Mar. 2009, pp. 304–311.

[22] 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.*, Aug. 2013, pp. 1–8.

[23] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2014.

[24] M. Burhanpurkar, M. Labbe, C. Guan, F. Michaud, and J. Kelly, "Cheap or robust? the practical realization of self-driving wheelchair technology," in *Proc. IEEE Int. Conf. Rehabil. Robot.*, Jul. 2017, pp. 1079–1086.

[25] U. Borgolte, H. Hoyer, C. Bühler, H. Heck, and R. Hoelper, "Architectural concepts of a semi-autonomous wheelchair," *J. Intell. Robotic Syst.*, vol. 22, nos. 3–4, pp. 233–253, Jul. 1998.

[26] E. Demeester, A. Hüntemann, D. Vanhooydonck, G. Vanacker, H. V. Brussel, and M. Nuttin, "User-adapted plan recognition and user-adapted shared control: A bayesian approach to semi-autonomous wheelchair driving," *Auto. Robots*, vol. 24, no. 2, pp. 193–211, Feb. 2008.

[27] X. Huang *et al.*, "The apolloscape dataset for autonomous driving," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Aug. 2018, pp. 1067–1076.

[28] R. Guzmán, J.-B. Hayet, and R. Klette, "Towards ubiquitous autonomous driving: The CCSAD dataset," in *Proc. Int. Conf. Comput. Anal. Images Patterns*, Sep. 2015, pp. 582–593.

[29] R. Simpson, J. Cullip, and J. Revell, "The Cheddar Gorge data set," BAE Systems, Farnborough, U.K., Tech. Rep., 2011.

[30] Z. Liu, "Performance evaluation of stereo and motion analysis on rectified image sequences," The University of Auckland, New Zealand, Tech. Rep., 2007.

[31] S. Meister and B. Jähne, and D. Kondermann, "Outdoor stereo camera system for the generation of real-world benchmark data sets," *Opt. Eng.*, vol. 51, no. 2, 2012, Art. no. 021107.

[32] 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.

[33] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Feb. 2016, pp. 379–387.

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

[35] J. Redmon and A. Farhadi. (2018). "YOLOv3: An incremental improvement." [Online]. Available: https://arxiv.org/abs/1804.02767

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

[37] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Aug. 2015, pp. 1440–1448.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[39] K. He, G. Gkioxari, and P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Mar. 2017, pp. 2980–2988.

[40] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. (2013). "OverFeat: Integrated recognition, localization and detection using convolutional networks." [Online]. Available: https://arxiv.org/abs/1312.6229

[41] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, no. 99, pp. 2999–3007, 2017.

[42] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1235–1386.

[44] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Mar. 2016, pp. 761–769.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Mar. 2016, pp. 770–778.

YING HE received the master's degree in computer applied technology from the College of Information Engineering, North China University of Science and Technology, China, in 2019. She majored in computer science and minored in e-commerce before pursuing the master's degree, from 2011 to 2016. During the period of her master's degree, she has experiences of researches on computational mathematics, in the College of Science, North China University of Science and Technology, and researches on object detection based on deep learning, in the Institute of Psychology, Chinese Academy of Sciences, from 2016 to 2019. Her research interests include machine learning, computer vision, and approximation theory.

HAN-BO YANG received the B.E. degree in information security from Wuhan University, China. He is currently pursuing the master's degree in information technology systems with Monash University and the master's degree in computer technology with Southeast University, Nanjing. His research interests include data analysis, deep learning, and emulation techniques.

SU-JING WANG (M'12) received the master's degree from the Software College, Jilin University, Changchun, China, in 2007, and the Ph.D. degree from the College of Computer Science and Technology, Jilin University, in 2012. He was a Postdoctoral Researcher with the Institute of Psychology, Chinese Academy of Sciences, from 2012 to 2015, where he is currently an Associate Researcher. He has published more than 40 scientific papers. He is one of the ten selectees of the Doctoral Consortium at the International Joint Conference on Biometrics 2011. He was called the Chinese Hawkin by the Xinhua News Agency. His current research interests include pattern recognition, computer vision, and machine learning. He serves as an Associate Editor of *Neurocomputing* (Elsevier). For more information, visit http://sujingwang.name.