TAD16K: AN ENHANCED BENCHMARK FOR AUTONOMOUS DRIVING

Yuming Li, Jue Wang, Tengfei Xing, Tianlu Liu, Chengjun Li, Kuifeng Su

Autonomous Driving Lab, Tencent, Beijing, China

ABSTRACT

Although promising results have been achieved in the areas of object detection and classification, few works have provided an end-to-end solution to the perception problems in the autonomous driving field. In this paper, we make two contributions. Firstly, we fully enhanced our previously released TT100K benchmark and provide 16,817 elaborately labeled Tencent Street View panoramas. This newly created benchmark, we call it Tencent Autonomous Driving 16K (TAD16K), not only contains previously labeled traffic-signs (221 types), but also creates annotations for three new objects, which are traffic lights (6 types), vehicles and pedestrians. Secondly, we provide the evaluation results of two state-of-the-art object detection algorithms (SSD and DetectNet) on our benchmark, which can be used as the baseline for future comparison purpose. Finally, we also demonstrate that the network trained on our benchmark can be directly deployed for practical application. The TAD16K, relevant additions and the source codes are publicly available [1,2].

Index Terms—Autonomous driving benchmark, Convolutional neural network, Object detection, Object classification.

1. INTRODUCTION

Object detection and classification in the scene are two important sub-tasks of scene understanding, which plays a major role in the autonomous driving. Recently, deep learning methods have shown superior performance for many tasks such as image classification and speech recognition. Since the deep neural network is a kind of data-driven tool and highly depended on the labeled data, the improved performance requires both good models and good-quality data.

Conventional benchmarks, such as PASCAL VOC [1], Microsoft COCO [2] and ImageNet ILSVRC [3], are usually designed for general purpose detection tasks. Since they are not aiming at self-driving purpose, many everyday objects, such as furniture, clothing, gadgets, food, and animals, are insignificant to self-driving cars. Besides, there also exists some specific benchmarks, the most relevant one to self-driving is the German traffic-sign detection and classification benchmark data [4,5]. However, these two benchmarks also exist some limitations. First, they are only designed for traffic-sign purpose and the image quantity is small. In addition, current methods achieve perfect or near perfect results for both tasks, with 100% recall and precision or detection and 99.67% precision for classification. Unfortunately, the high performance on these two benchmarks do not mean that the problems are solved since the benchmark data are not representative of that encountered in real tasks. In the GTSDB detection benchmark task, the benchmark only contains totally 900 images and 4 major categories. In the GTSRB classification benchmark, the benchmark provides only 40 classes and many traffic-signs occupy most of the image which is quite different from the real scenario. In real world tasks, the main difficulty when detecting and classifying traffic-signs in an ordinary image is their very small size, often less than 1% of the image. There are also no negative samples disrupting the classification in GTSRB. The KITTI Vision Benchmark [6] is most relevant to the autonomous driving. However, it is a comprehensive benchmark not only concentrates on object detection, but also focuses on stereo, optical flow, visual odometry and tracking. For the object detection part, it only provides images contains cars and pedestrians. In this paper, we have created a new, more realistic benchmark dedicated to solve the perception problems commonly concerned in the autonomous driving field. The contributions of this paper are as follows.

- We have created a new, more realistic benchmark for autonomous driving purpose, which is called TAD16K. It contains 16,817 elaborately labeled Tencent Street View panoramas with the resolution of 2048 × 2048. The images in TAD16K cover real world conditions in China, with large variations in such aspects as size, illuminance and weather conditions, also including examples with occlusion. Unlike existing benchmarks, TAD16K is annotated with 221 types of traffic-signs, 6 types of traffic lights, various vehicles and pedestrians.
- We have trained two state-of-the-art object detection deep neural networks, which are SSD and DetectNet. We provide the trained models and the evaluation results of these two methods on TAD16K. These models can not only be used as the baseline for future comparison, but also be directly deployed for real applications.

The rest of the paper is organized as follows: In Section 2, we detail our TAD16K benchmark. In Section 3, we present

[1]TAD16K: http://autopilot.qq.com/ICIP2017/
[2]Source code: https://github.com/lymhust/TAD16K_source

the deep neural networks we used. We give experimental results in Section 4 and conclusions in Section 5.

2. TAD16K BENCHMARK

TAD16K is an updated version of our previously released traffic-sign benchmark TT100K [7]. The TT100K is a large traffic-sign benchmark. The images in it come from 100,000 Tencent Street View panoramas. which choses 10 regions from 5 different cities in China (including both downtown regions and suburbs for each city) and downloaded from the Tencent Data Center. Presently, Tencent Street Views cover about 300 Chinese cities and the road networks linking them. The original panoramas were captured by 6 SLR cameras and then stitched together. Image processing techniques such as exposure adjustment were also used. Images were captured both from vehicles and shoulder-mounted equipment, at intervals of about 10 meters. These capture and processing methods of panoramas in Tencent Street View guarantees the high definition and resolution of our TT100K and TAD16K. Besides, panoramas in Tencent Street Views are captured in different seasons and weather, which further ensures the variety of our datasets.

TT100K provides 100,000 images containing 30,000 traffic-sign instances. It contains about 16k images which have been annotated with a pixel mask for each traffic-sign, as well as giving its bounding box and class. The rest of the images in TT100K are negative samples which have no traffic-signs in them.

The TAD16K benchmark is created based on the annotated 16k images in TT100K, which retained the traffic-sign annotations (221 types) and created annotations for three new objects, which are traffic lights (6 types), vehicles and pedestrians. All the images in TAD16K have the resolution of 2048 × 2048. In previous benchmarks such as GTSRB or KITTI objects, objects were extracted from a video sequence which leads to many very similar images. Since the images in TAD16K are extracted from panoramas, the objects in it vary significantly.

All the images in TAD16K are annotated by hand. Fig. 1 shows the examples of annotations for one image. The annotation rules for four different objects are as follows.

Traffic sign. Traffic signs in China follow international patterns, and can be classified into three categories: warnings (mostly yellow triangles with a black boundary and information), prohibitions (mostly white surrounded by a red circle and also possibly having a diagonal bar), and mandatory (mostly blue circles with white information). Other signs exist that resemble traffic-signs but are in fact not, such signs are placed in an 'other' class of a particular category. Fig. 2 shows the Chinese traffic-sign classes in TAD16K. Signs in yellow, red and blue boxes are warning, prohibitory and mandatory signs respectively. Each traffic sign has a unique label. Some signs are representative of a family, such as speed limit signs for different speeds. Such signs are generically denoted as shown in Fig. 2. Using 'pl*' as an example, the unique label is determined by replacing '*' by a specific value (e.g. 'pl40' for a 40 km/h speed limit sign).

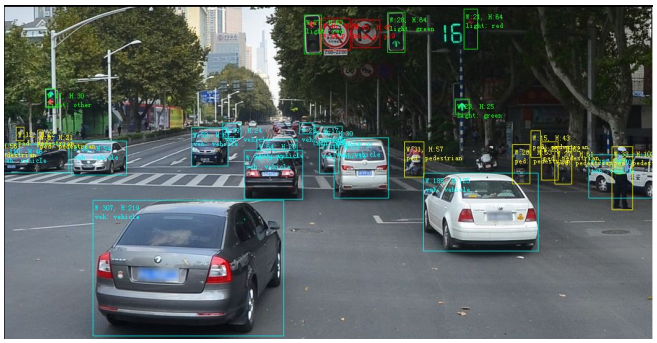

Fig. 1: Example of annotations for one image.

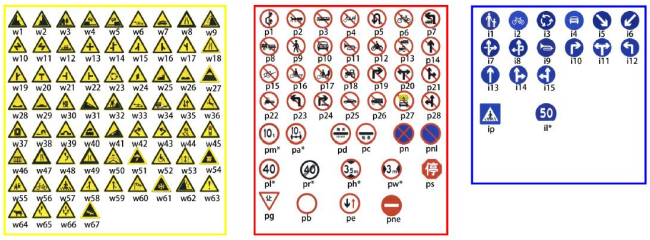

Fig. 2: Chinese traffic-sign classes.

Fig. 3: Traffic light classes.

Table 1: Statistics of the four types of objects in TAD16K.

Object	(Hmin, Hmax)	(Wmin, Wmax)	Number
Sign	(7, 495)	(7, 440)	27,253
Light	(6, 512)	(4, 732)	17,369
Vehicle	(3, 1955)	(3, 2049)	114,717
Pedestrian	(5, 712)	(3, 1164)	43,046

Traffic light. Traffic lights can be classified into six classes, which are red, yellow, green, person, bike and other. The ordinary traffic lights which indicating motor vehicles can be classified into red, yellow and green three types. Traffic lights indicating pedestrians, no matter it is red or green, are classified as pedestrian class. Similarly, those indicating non-motor vehicles, such as bicycles, are classified as bike class without considering their colors. Further, we classify other types as other class, for example, the color of the light cannot be seen or distinguished, or the mean of the light is not unique. All these situations are listed in Fig. 3.

Vehicle. We classify common motor vehicles as vehicle class which includes cars, buses and trucks etc. Besides, the

three-wheeled transport vehicle is also classified as vehicle class.

Pedestrian. The pedestrian class includes common pedestrians, cyclists and motorcyclists.

Table 1 shows the statistics of the four types of objects in TAD16K. We can see that the object size in TAD16K has a wide range, which means this benchmark contains both small size objects and large size ones. In addition, the total number of all these four types of objects are more than 15k. For vehicle class, it is up to 110k. The quantity of these objects are also enough to train a good classifier. It will hopefully provide a suitable basis for research into both detecting and classifying small objects.

Images in TAD16K are distributed into three folders which are train folder (6,105 images), test folder (3,071 images) and other folder (7,641 images). Images in train and test folders can be used for training and testing purpose, those in other folders can be for validation or other purpose. To make the TAD16K easy to use, we further divided these three folders into several sub-folders. Each sub-folder contains 500 images and a corresponding json file. This json file contains all the annotations of these images in this sub-folder. For example, the json file *train_1_label.json* is related to sub-folder *train_1*.

The source codes of TAD16K in our Github repository include the demo functions to parse these json files. In addition, detailed information and supplementary materials about our TAD16K benchmark can be found on our TAD16K website provided in the first page.

3. BENCHMARK ALGORITHM

After the successful usage of convolutional neural networks (CNNs) in image classification, they were quickly adapted to object detection. Current state-of-the-art object detection and recognition systems are variants of the deep neural networks especially CNNs.

To facilitate the usage of TAD16K and provide a baseline for future comparison of different algorithms. We trained two state-of-the-art object detection algorithms using the TAD16K benchmark and provide the evaluation results of them.

3.1. DetectNet

DetectNet [8] comes from the industry, which is proposed by NVIDIA and is now fully supported by DIGITS [9], TensorRT [10] and DriveWorks [11]. The general idea of DetectNet is very similar with our previously proposed method in [7]. In our previous method, AlexNet [12] is used as the main part of the architecture. In DetectNet, they apply GoogLeNet [13] as the fully-convolutional network (FCN) part. The DetectNet has three main features.

Firstly, since the FCN sub-network of DetectNet has the same structure as GoogLeNet, the DetectNet can be initialized using a pre-trained GoogLeNet model, thereby reducing training time and improving final model accuracy.

Secondly, since the FCN sub-network in DetectNet has no fully-connected layers, it can accept input images with varying sizes and effectively applies a CNN in a strided sliding window fashion.

Finally, the output of DetectNet has two branches, one predicts the coverage map and another predicts the bounding boxes. Since the coverage map is a real-valued multi-dimensional array, DetectNet can be easily extended to multi-objects detection.

3.2. SSD

Single Shot MultiBox Detector (SSD) [14] comes from the academia, which is a method for detecting objects in images using a single deep neural network. The features of SSD are as follows.

Firstly, the sub-network in SSD is VGG-16 [15]. Similar to DetectNet, SSD can be also initialized using a pre-trained VGG-16 model.

Secondly, the architecture design of SSD follows the mainstream which considers different aspect ratios and scales in feature map location. Besides, at prediction time, the network generates scores for the presence of each object category in each default box and produces adjustments to the box to better match the object shape.

Additionally, the network also combines predictions from multiple feature maps with different resolutions to naturally handle objects of various sizes. The SSD model is simple relative to methods that require object proposals because it completely eliminates proposal generation and subsequent pixel or feature resampling stage and encapsulates all computation in a single network. This makes SSD easy to train and straightforward to integrate into systems that require a detection component.

Finally, SSD has been proven to be better than Faster-RCNN and YOLO both in accuracy and speed.

4. EXPERIMENTAL RESULTS

In the experimental evaluation of DetectNet and SSD, both training and testing were done on an Ubuntu 14.04 PC with an Intel Xeon E5-1620 CPU, one NVIDIA TITAN X GPU with 12GB memory. DetectNet is directly trained in DIGITS, while SSD is trained using the script provided by its source code. Before training, we use the pre-trained BVLC GoogLeNet to initialize DetectNet, which is provided in Caffe framework. SSD is initialized by VGG16, which is pre-trained on the ILSVRC CLS-LOC dataset.

We train these two methods using all the images in the train folder and test them using the test folder. Validation is done using other folder. When testing, images are resized into 1024 × 1024 and we only consider the objects whose width or height is larger than a threshold. Table 2 shows the evaluation results of DetectNet and SSD when the Intersect-

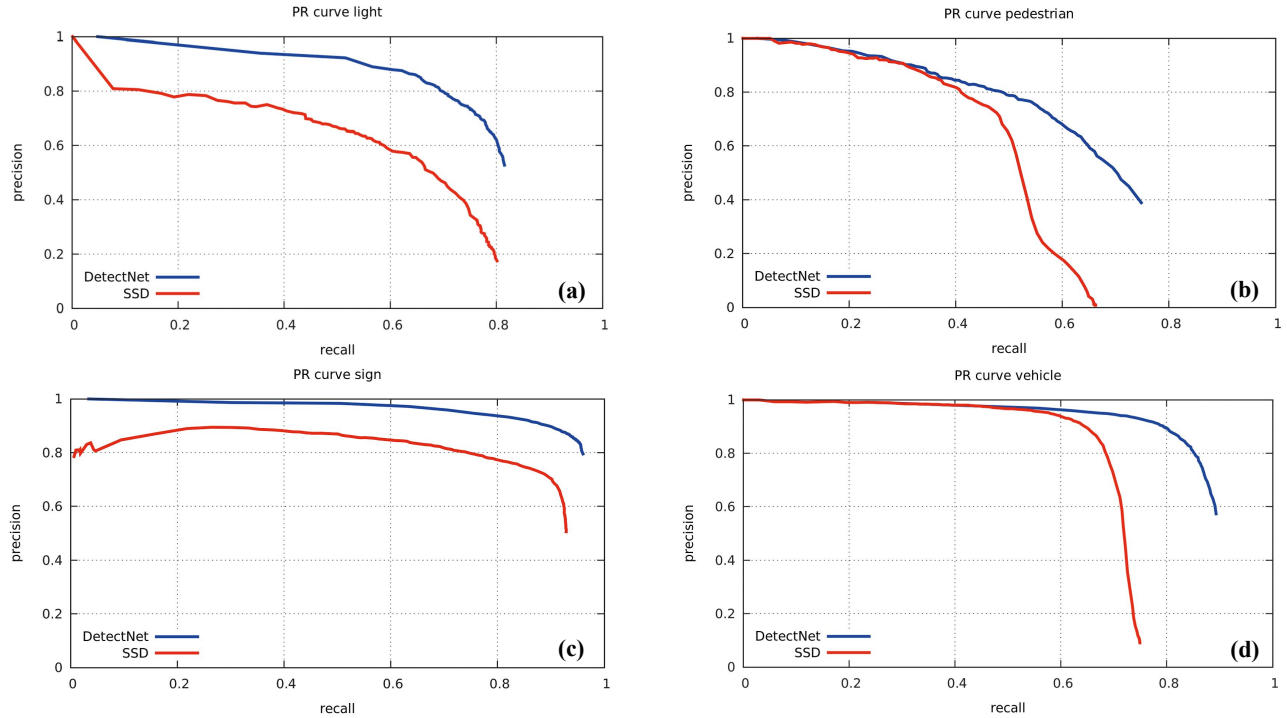

Fig. 4: PR curve of DetectNet and SSD for four types of objects in TAD16K. (a) Traffic light. (b) Pedestrian. (c) Traffic sign. (d) Vehicle.

Table 2: Detection results for two methods on TAD16K.

	mAP			
	Sign (T=20)	**Light (T=20)**	**Vehicle (T=50)**	**Pedestrian (T=30)**
DetectNet	0.93	0.74	0.85	0.61
SSD	0.77	0.54	0.70	0.49

ion over Union (IoU) threshold of the Non-Maximum Suppression (NMS) is set as 0.5. The T value is the threshold used for each class. Fig. 4 shows the Precision-Recall (PR) curve of both DetectNet and SSD for four types of objects in TAD16K. We can see that the performance of DetectNet is better than SSD for all of these four objects.

TAD16K benchmark is designed to solve the perception problems in autonomous driving and the trained models (both DetectNet and SSD) using TAD16K can be directly deployed for real applications. Besides, since all of the components in DetectNet are constructed using cuDNN [16], when deployed for real-time application, it can be further accelerated by TensorRT. Although object detection and recognition can be integrated by DetectNet and SSD, in the real application, we found that it is better to separate these two tasks. We apply DetectNet and SSD only as the detector and trained a simple network such as LeNet or IDSIA [17] as the dedicated classifier. In the TAD16K website, we have provided some testing results on images captured by different devices, which includes PointGrey industrial camera, Basler industrial camera, common car driving recorders and common web cameras. Although both the image quality and color captured by these devices vary significantly compared with the images in TAD16K, the neural networks trained on TAD16K still performs well on these new images. These results, from another point of view, demonstrate the quality and diversity of the samples in TAD16K.

5. CONCLUSIONS

In this paper, we have created a new benchmark for simultaneously detecting and classifying objects concerned by self-driving cars, which are traffic signs, traffic lights, vehicles and pedestrians. Compared with previous benchmarks related to autonomous driving, images in this benchmark are more variable and have a higher resolution, and the object size in these images is in a large range. In addition, we have trained two networks (DetectNet and SSD) on this benchmark and the evaluation results of them are provided. These two methods can be used as a baseline for future research. To assist research in this field, we make this benchmark, trained models and source code public available. We hope this benchmark can act as a new challenge for the object detection and recognition community as well as the autonomous driving community.

In the future, we plan to provide the ground truth of more objects for pixel-wised and instance-aware semantic segmentation in TAD16K, and the baseline algorithms will also be provided.

6. ACKNOWLEDGEMENTS

We thank NVIDIA to open source the DetectNet, DIGITS, TensorRT, cuDNN and DriveWorks. We also thank the authors of SSD to make their source code public available.

7. REFERENCES

[1] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9), 1627-1645.J

[2] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In *European Conference on Computer Vision* (pp. 740-755). Springer International Publishing.

[3] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

[4] Stallkamp, J., Schlipsing, M., Salmen, J., & Igel, C. (2011, July). The German traffic sign recognition benchmark: a multi-class classification competition. In *Neural Networks (IJCNN), The 2011 International Joint Conference on* (pp. 1453-1460). IEEE.

[5] Stallkamp, J., Schlipsing, M., Salmen, J., & Igel, C. (2012). Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32, 323-332.

[6] Geiger, A., Lenz, P., & Urtasun, R. (2012, June). Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 3354-3361). IEEE.

[7] Zhu, Z., Liang, D., Zhang, S., Huang, X., Li, B., & Hu, S. (2016). Traffic-sign detection and classification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2110-2118).

[8] DetectNet: Deep Neural Network for Object Detection in DIGITS. https://devblogs.nvidia.com/parallelforall/detectnet-deep-neural-network-object-detection-digits/.

[9] Interactive Deep Learning GPU Training System. https://developer.nvidia.com/digits.

[10] High performance deep learning inference for production deployment. https://developer.nvidia.com/tensorrt.

[11] Developer Tools for Self-Driving Cars. http://www.nvidia.com/object/driveworks.html.

[12] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

[13] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-9).

[14] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., & Reed, S. (2015). SSD: Single Shot MultiBox Detector. *arXiv preprint arXiv:1512.02325*.

[15] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[16] NVIDIA cuDNN GPU Accelerated Deep Learning. https://developer.nvidia.com/cudnn.

[17] CireşAn, D., Meier, U., Masci, J., & Schmidhuber, J. (2012). Multi-column deep neural network for traffic sign classification. *Neural Networks*, 32, 333-338.