Al-YOUSOFI: A Novel Benchmark for Multi-Vehicle Detection and Tracking

Ahmad Al-Omari

Department of Information Systems, Yarmouk University, Jordan

Abstract

Vehicle detection in aerial imagery has been instrumental in a wide range of applications. Lately, as a result of the robust feature representations, convolutional neural networks (CNN) based detection methods have achieved prodigious performance in computer vision. The diversity of dataset sources that relate to vehicle images is numerous, but it is not sufficient in some cases due to the different types of vehicles from one country to another. In this paper, we propose a new dataset of vehicle images called AL-YOUSOFI taken in the Hashemite Kingdom of Jordan, Irbid city. The AL-YOUSOFI benchmark dataset consists of 40 challenging videos captured from real-world traffic scenes (over 158,000 frames with rich annotations, including vehicle type and vehicle bounding boxes) for multi-object detection and tracking. The state-of-the-art algorithms that have been previously trained on AL-YOUSOFI, has been evaluated. The results showed a variation in efficiency between the algorithms, and this is due to how each works. The full dataset is available at this URL.

Keywords: vehicle detection dataset; computer vision; object tagging; data collection; object detection and tracking benchmark

1. INTRODUCTION

Vehicle detection is a significant process in numerous areas, as for example: remote sensing, intelligent transportation and statistical surveys. For the increasing image-based datasets, vehicle detection is becoming a challenge, and has recently attracted wide attention. With the great development of the computer vision, it is becoming easier to develop a detector for moving and stationary vehicles. Vehicle detection is an essential task in computer vision and is widely applied in some real-world applications, such as smart traffic light [1], traffic surveillance [2,3], and driving assistant system [4,5].

One of the most important challenges facing the vehicle detector and classifier is that vehicles pass at different speeds and with different sizes and distances. For example, the results of detecting a vehicle passing at a low speed and its size in the image is large (close to the camera) better than discovering a vehicle that passes at high speed and its size is small (away from the camera).

Additionally, the vehicle image analysis includes classification of the image, based on its features. The classification of vehicles is essential for accurately determining the length of the vehicle. In general, the methods proposed for the automated detection and classification of these images require a large dataset and a variety of classes, which are cropped from the original images and further analyzed.

In this work, we propose a new large-scale dataset. The AL-YOUSOFI dataset includes 40 challenging videos with more than 158,000 frames of real-world traffic scenes.

Table 1 compares the existing vehicle detection or tracking datasets, first four columns: the number of training/testing data (1k = 1000) indicating the number of images and the number of objects bounding boxes, objects annotated, image resolution and the year of dataset creation.

2. RELATED WORK

Several datasets have been developed for vehicle detection. These datasets are mainly developed for vehicle detection in single images that can be used to train detectors for traffic management and monitoring systems. In this section, we will discuss the related work that developed image-based vehicle detection datasets, divided into ground images, aerial images, and Computer-generated imagery (CGI).

1. Aerial images

Some studies are concerned only with aerial images [6, 7, 8] and were built by aerial photography of cities and extracting composite images from them to benefit from them.

Razakarivony and Jurie [6] create another well-annotated dataset, called the VEDAI dataset, by cropping the very large satellite images into more than 1200 smaller images with two different resolutions, 512 × 512, and 1024 × 1024, and a ground sampling distance of around 25 and 12.5 cm/pixel, respectively. The VEDAI-512 is just the downscaled version of VEDAI-1024, which makes the target vehicles smaller and more challenging. VEDAI consists of nine different classes of sparse small vehicles all along with various backgrounds and confused objects captured from the same distance to the ground and with no oblique views. VEDAI is basically used to train and test small-size vehicle detection algorithms. Many state-of-the-art studies are based on the VEDAI dataset as a baseline for object detection algorithm development.

DOTA [7] was created recently by Xia et al., which contains 15 object categories, including vehicles of different scales and orientations. Moreover, each object is annotated manually by experts with an oriented bounding box. It consists of 2806 aerial images captured through different platforms in multiple cities. Thus, the size of each image is varying between (800 × 800) and (4000 × 4000) pixels with multiple GSDs. DOTA dataset provides a good balance between small- and middle-size objects, which makes it very similar to real-world scenes.

COWC [8] was designed by Mundhenk et al., the Cars Overhead with Context (COWC) dataset, which consists of 53 TIFF images, from six various locations, with different resolutions varying from (2000 × 2000) to (19 000 × 19 000) pixels, and a GSD of around 15 cm/pixel. Instead of labeling the images with bounding boxes, the authors followed another style of annotation, where they used the center pixel point of the vehicles. COWC dataset can be used for vehicle detection and counting tasks. However, it is a very challenging dataset due to the small size of cars that vary between 24 and 48 pixels.

2. Ground images

UA-DETRAC, MOT16, KITTI, and TIV datasets [9, 10, 11, 12] were used only cameras to collect data. The work is different with the position and height of the camera, the resolution of frames, and the considered objects.

The UA-DETRAC dataset [9] proposed by a University at Albany for comprehensive performance evaluation of detectors for the multi-object detection systems. The UA-DETRAC benchmark dataset consists of 100 challenging videos captured from real-world traffic scenes (over 140,000 frames with rich annotations, including illumination, vehicle type, occlusion, truncation ratio, and vehicle bounding boxes) for multi-object detection and tracking. The videos were recorded at 25 frames per second (fps) with a resolution of 960 × 540 pixels. The UA-DETRAC was evaluated and the results showed the effects of detection accuracy on detection system performance.

The MOT16 dataset [10] was developed for collecting existing and new data and creating a framework for the standardized evaluation of multiple object tracking. The MOT16 dataset consists of 14 very different types of sequences such as moving or static camera, several heights of camera, and all possible weather conditions. The sequences were recorded at 30 frames per second (fps) with a resolution of 1920 × 1080 pixels. The MOT16 dataset has proven to be a benchmark for evaluating multi-object detection systems, and the researchers stated that they will continue to address challenges multi-object detection and tracking systems.

The KITTI [11] dataset was developed by using the autonomous driving platform, which are acquired from a moving vehicle with viewpoint of the driver. The KITTI dataset is designed for 3D object detection and 3D tracking. They used four high resolution video cameras with a resolution of 1392 × 512 pixels. Results from state-of-the-art algorithms showed that methods ranking high.

The TIV [12] was developed for infrared video benchmark, for various visual analysis tasks, which include single object tracking, multi-object tracking in single or multiple views, analysing motion patterns of large groups, and censusing wild animals in flight. The TIV dataset contains real-world sequences with a resolution of 1024 × 1024 pixels. The TIV dataset improves and expands video datasets available to the research community with thermal infrared footage, which poses new and challenging video analysis problems.

3. CGI

SYNTHIA, GTA-V, and OPV2V [13, 14, 15] were all assembled from unrealistic frames, for example, as a result of simulation of the real world, or frames extracted from games.

The SYNTHIA [13] was released by Ros et al. in 2016 to address the challenge of semantic segmentation and to improve understanding of scenes related to driving. With over 213,400 synthetic images captured from different viewpoints, seasons, weather conditions and lighting, it provides pixel-level annotations for 13 categories, including sky, building, road, sidewalk, fence, vegetation, lane-marking, pole, car, traffic signs, pedestrians, cyclists and miscellaneous. Simulation results show that the models trained on SYNTHIA and fine-tuned on the real dataset achieve improved semantic accuracy. In addition, SYNTHIA-San Francisco (SYNTHIA-SF) [16] and SYNTHIA-AL [17] have been also released on the basis of SYNTHIA. SYNTHIA-SF contains 2,224 synthetic images with accurate depth information and 19 classes of semantic annotations making it ideal for evaluating the accuracy of depth and semantic segmentation. SYNTHIA-AL, on the other hand, includes annotations for instance segmentation, 2D and 3D bounding boxes, and depth information, specifically designed for evaluating active learning in road scenes for video target detection.

Grand Theft Auto V (GTA-V) [14], published by Richter et al. in 2016, is a pixel-level semantic segmentation dataset that relies on the realism of commercial video games. It features accurate simulations of material appearance, light transmission, player interaction, and realistic placement of objects and environments. By incorporating GTA-V into existing datasets such as CamVid and KITTI [11], model training accuracy can be greatly improved while minimizing the need for expensive manual annotation of real-world data. Some state-of-the-art models such as MIC [18], HRDA [19], SePiCo [20], have shown great performance in semantic segmentation of GTA-V.

The Open Dataset for Perception with V2V communication (OPV2V) [15] was proposed by Xu et al. in 2022, which is a large-scale vehicle-to-vehicle collaborative perception dataset. The dataset contains 11,464 frames and 232,913 annotated 3D box diagrams of vehicles and provides a comprehensive benchmark of up to 16 models to evaluate fusion strategies and advanced radar target detection algorithms. The dataset sensors include 4 RGB cameras, a 64-channel LiDAR, a GPS, and an IMU, providing researchers with annotated RGB images, lidar point cloud data, and BEV perception images. The OPV2V dataset supports collaborative 3D object detection, BEV semantic segmentation, tracking, and prediction using cameras or LiDAR sensors. At the same time, users can define tasks by adding additional sensors, such as depth estimation, sensor data fusion, etc. The domain shifts cover 70 different scenes, and the map is derived from geographic information of 8 towns in CARLA and Culver City, a digital town in Los Angeles.

3. DATASET CREATION

The AL-YOUSOFI dataset consists of 158,394 images of real-world traffic views acquired by a Canon EOS 7D camera with a resolution of 1280x720. The AL-YOUSOFI dataset are divided into 103,155 images for training and 55,239 images for testing. The dataset consists of four types of vehicles (classes) as follows: (a) car, (b) bus, (c) truck, and (d) motorbike.

A. Collection Criteria

As collection criteria, we set three basic goals, which are as follows: 1) the dataset must be open and fully distributable, so that everyone can use it for further studies, 2) it must contain a large number of images of all vehicles to be as effective as possible, because it is known that the greater number of images in the dataset, the greater the efficiency of the learning process, and 3) it must be expandable in the future, to be the first step to building a huge traffic dataset in the Hashemite Kingdom of Jordan.

In order to construct a new traffic dataset in the city of Irbid in the Hashemite Kingdom of Jordan. We obtained the approval of the officials of Yarmouk University and the approval of the Governor of Irbid City. It was called AL-YOUSOFI because it was collected in the AL-YOUSOFI intersection area.

B. Data Collection

AL-YOUSOFI dataset was constructed from real-world traffic videos acquired by a Canon EOS 7D camera with a resolution of 1280x720. The camera is placed at a height of 3 meters at 4 different locations. Figure 1 shows how the camera is installed on a street.

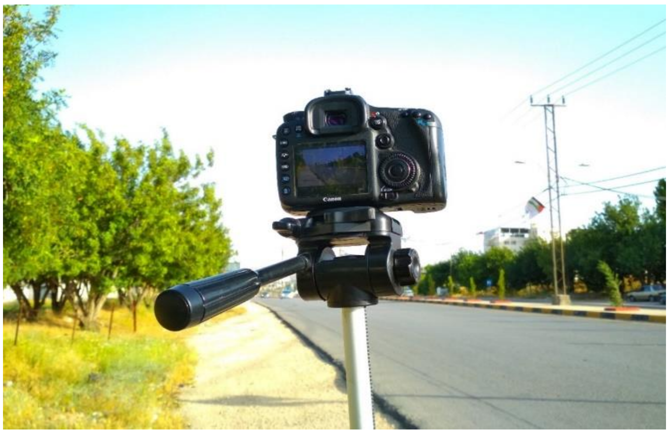

Figure 1 Installing the camera to record traffic videos.

Table 1 Comparison of existing vehicle detection or tracking datasets [B]: target object is Bus, [C]: target object is Car, [M]: target object is Motorbike, [P]: target object is Pedestrian, [T]: target object is Truck, [V]: target object is Van, and [O]: other objects not related to vehicles.

Dataset	Training set		Testing set		Properties		
	Frames	Boxes	Frames	Boxes	Classes	Resolution	Year
KITTI [11]	7.5k	40.6k	7.5k	39.7k	P, C, O	1392 × 512	2012
TIV [12]	57.2k	-	6.5k	-	P, C, M, O	1024 × 1024	2014
MOT16 [10]	5.3k	110k	5.9k	182k	P, C, M, O	1920 × 1080	2016
GTA-V [14]	24.9k	-	0	0	B, C, T, M	1914×1052	2016
SYNTHIA [13]	213k	-	0	0	C, O	1280 × 960	2016
UA-DETRAC [9]	84k	578k	56k	632k	B, C, V	960 × 540	2020
AL-YOUSOFI	103.1k	372.8k	55.2k	226.9k	B, C, T, M	1280 × 720	2023

C. Data Preparation

1) Converting videos into images

All video clips were converted into a group of images, so that five frames were extracted from each second. After completing the conversion process, the number of frames was (158,394). The frames are divided into (103,155) images for training and (55,239) images for testing. We took 65% of the data for training and 35% for testing. Figure 2 shows a sample of the dataset.

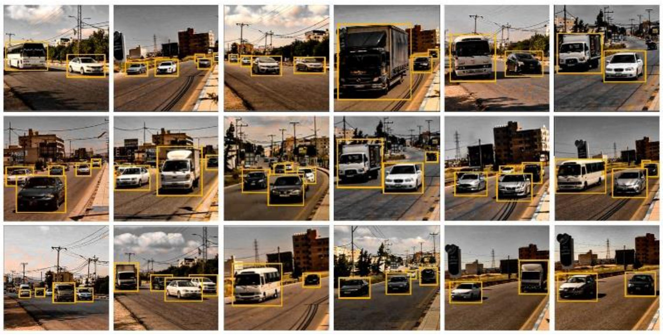

Figure 2 Sample of the dataset

2) Exclude irrelevant images

The images that not related to our study, such as empty images of vehicles or images containing people's faces, were excluded. In addition to making sure that vehicle license plates are not readable.

3) Data annotation

We used Visual Object Tagging Tool (Vott) to assign tags to the objects in each image in the dataset. We created four tags (vehicle types) Car, Bus, Truck and Motorbike. In each image, the coordinates (x1, y1, x2, y2) were determined for each vehicle, in addition to determining its type (tag) as shown in Figure 3 The output was csv file, which contains the following attributes:

- class_name: the vehicle type
- x1: the left coordinate of the bounding box
- y1: the top coordinate of the bounding box

- width: the width of the bounding box
- height: the height of the bounding box
- image_name: the name of the image file
- image_width: the width of original image
- image_height: the height of original image
-

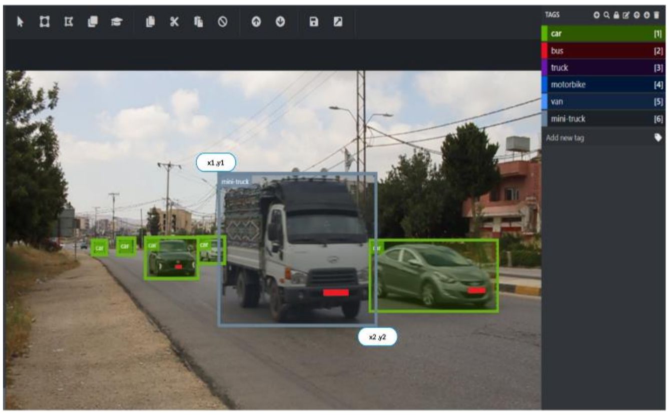

Figure 3 Assigning tags to vehicles using Vott tool

In this process all images were manually annotated using Vott with a total of (372,899) labeled bounding boxes for training and (226,966) labeled bounding boxes for testing. Figure 4 shows the number of bounding boxes for each class in the training and testing parts.

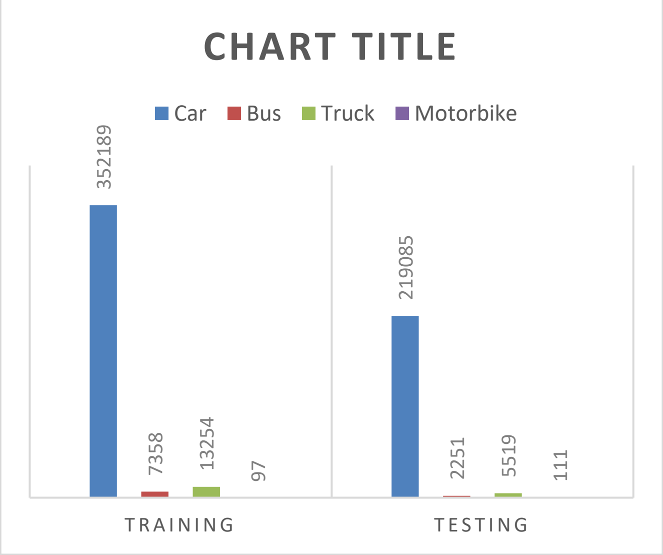

Figure 4 number of bounding boxes for each class in the training and testing

4. EVALUATION

Overall performance

We tested ten state-of-the-art object detectors on the AL-YOUSOFI dataset to calculate the Average Precision for each class and calculate the mean Average Precision (mAP) for all classes for each detector. In this research, the prediction is correct if (IoU $\geq$ 0.5). The results shown in Figure 5 with the mAP scores, indicate that there remains much room for improvement for object detection algorithms. Specifically, as there is significant and rapid development in the YOLO family. Unlike some algorithms, it doesn't do the job of detecting objects very well, specifically, the R-CNN and Fast R-CNN algorithms can't detect vehicles well with only 36% and 42% mAP scores respectively. SSD and RetinaNet algorithms perform slightly better than Mask R-CNN and Faster R-CNN algorithms by 4%. Back to YOLO family, the YOLO4, YOLO5 and YOLO6 algorithms achieve more accurate results with 79%, 85% and 85% mAP scores respectively. The highest score achieved by the YOLO7 algorithm was 88% mAP.

Vehicle type

As shown in Figure 6, the detectors work relatively well on cars compared to other types of vehicles. The reason is due to the large variation in the number of images for each type of vehicle in AL-YOUSOFI dataset. The largest part of the AL-YOUSOFI dataset was for the "car" because it is the most existing in the street, unlike other types of vehicles. Figure 4 illustrates this variation in the number of images in both the training and testing sets.

Scale

It is worth noting that the relative size of the vehicle in the image is an important factor that greatly affects the efficiency of the detector during the detection process.

Vehicle speed

The speed of the vehicle during capturing is also one of the factors affecting the efficiency of the detector, so the higher the speed of the vehicle, the less visible its features in the image.

5. CONCLUSION AND FUTURE WORK

In this paper, the publicly available AL-YOUSOFI image-based dataset is introduced. It contains of vehicles images, which passed on AL-YOUSOFI intersection. The images are divided into four categories of car, bus, truck and motorbike. Ten state-of-the-art algorithms have been trained and evaluated on the AL-YOUSOFI dataset. The results were dissimilar between the detectors, and this is due to the nature of the process of detecting the objects in each of them. The results were analysed based on some important criteria that influence the discovery process.

The results also showed that some algorithms achieved tremendous and promising results in vehicle detection, which indicates the efficiency of the dataset in terms of size and distribution. However, the dataset lacks a balance between the number of images for each type of vehicle, or in other words, some types of vehicles were very few compared to others, for example, the number of motorbike images is very small, and this clearly reflected the results. As a future work, this data set will be expanded and new vehicle types will be added and work to satisfy all vehicle types with a sufficient number of images.

ACKNOWLEDGMENTS

Necessary approvals were taken from the officials in Irbid and Yarmouk University to collect the dataset.

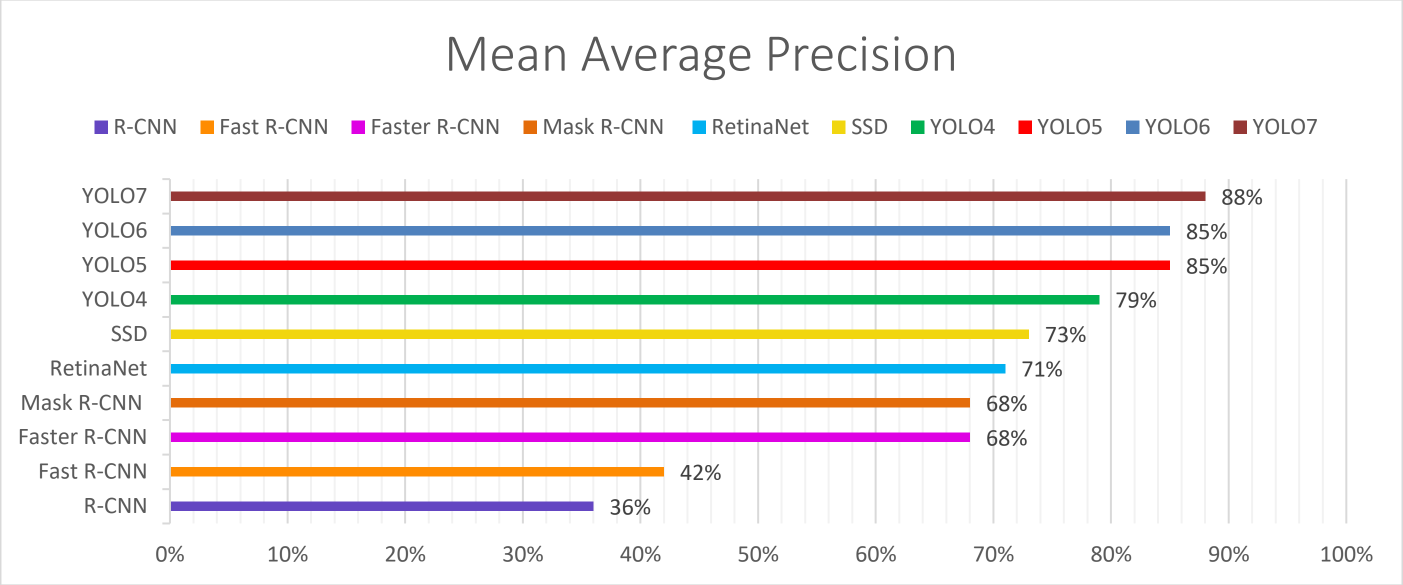

Figure 5 Mean Average Precision for state-of-art algorithms trained on AL-YOUSOFI Dataset

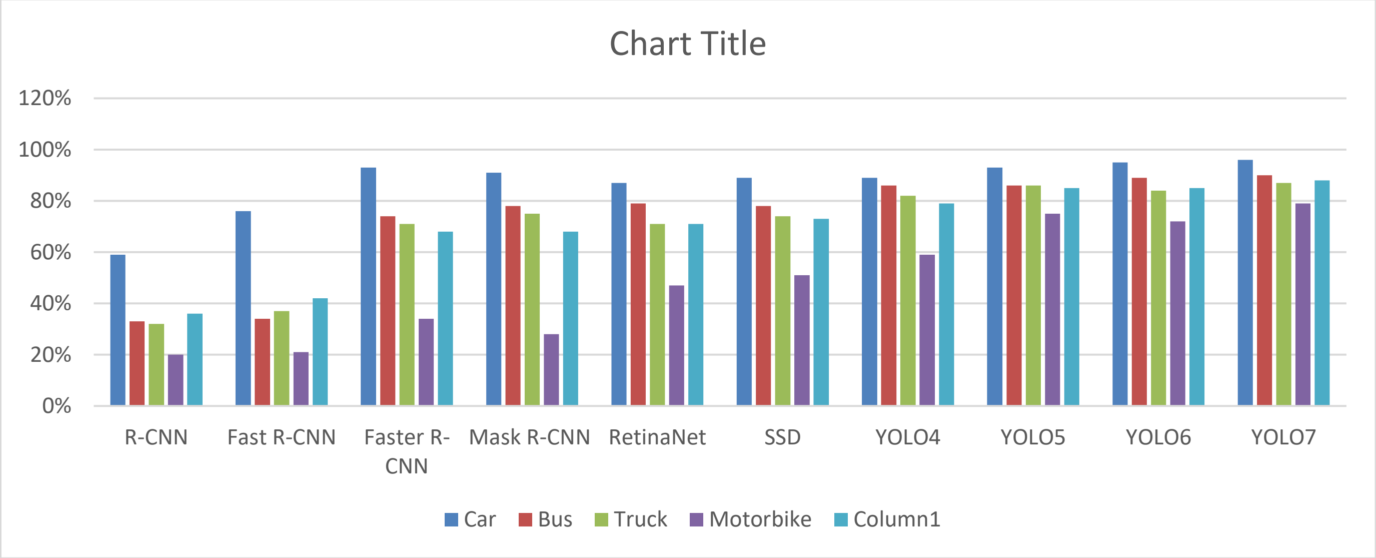

Figure 6 Average Precision for all classes for each algorithm

REFERENCES

1. Ottom, M.; Al-Omari, A. An Adaptive Traffic Lights System using Machine Learning. The International Arab Journal of Information Technology (IAJIT) Vol. 20, No. 3, 407- 418, 2023.
2. Tang, Y.; Zhang, C.; Gu, R.; Li, P.; Yang, B. Vehicle detection and recognition for intelligent traffic surveillance system. Multimedia tools and applications Vol. 76, 5817-5832, 2017.
3. Wen, X.; Shao, L.; Fang, W.; Xue, Y. Efficient feature selection and classification for vehicle detection. IEEE Transactions on Circuits and Systems for Video Technology Vol. 25, No. 3, 508-517, 2014.
4. Xu, H.; Zhou, Z.; Sheng, B.; Ma, L. Fast vehicle detection based on feature and real-time prediction. In: Proceedings of the IEEE International Symposium on Circuits and Systems, 2860-2863, 2013.

5. Gu, Q.; Yang, J.; Zhai, Y.; Kong, L. Vision-based multi-scaled vehicle detection and distance relevant mix tracking for driver assistance system. Optical Review Vol. 22, 197-209, 2015.
6. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery (vedai): a benchmark. Journal of Visual Communication and Image Representation Vol. 34, 2015.
7. Xia, G. S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 3974-3983, 2018.
8. Mundhenk, T. N.; Konjevod, G.; Sakla, W. A.; Boakye, K. A large contextual dataset for classification, detection and counting of cars with deep learning. In: Proceedings of Computer Vision–ECCV 2016: 14th European Conference, 785-800, 2016.
9. Wen, L.; Du, D.; Cai, Z.; Lei, Z.; Chang, M. C.; Qi, H.; Lyu, S. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. Computer Vision and Image Understanding Vol. 193, 2020.
10. Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; Schindler, K. MOT16: A benchmark for multi-object tracking. arXiv preprint, 2016.
11. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In: Proceedings of 2012 IEEE conference on computer vision and pattern recognition, 3354-3361, 2012.
12. Wu, Z.; Fuller, N.; Theriault, D.; Betke, M. A thermal infrared video benchmark for visual analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 201-208, 2014.
13. Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; Lopez, A. M. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 3234-3243, 2016.
14. Richter, S. R.; Vineet, V.; Roth, S.; Koltun, V. Playing for data: Ground truth from computer games. In: Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, 102-118, 2016.
15. Xu, R.; Xiang, H.; Xia, X.; Han, X.; Li, J.; Ma, J. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In: Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), 2583-2589, 2022.
16. Hernandez-Juarez, D.; Schneider, L.; Espinosa, A.; Vázquez, D.; López, A. M.; Franke, U.; Moure, J. C. Slanted stixels: Representing San Francisco's steepest streets. arXiv preprint, 2017.
17. Bengar, J. Z.; Gonzalez-Garcia, A.; Villalonga, G.; Raducanu, B.; Aghdam, H. H.; Mozerov, M.; Van de Weijer, J. Temporal coherence for active learning in videos. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop, 914-923, 2019.
18. Hoyer, L.; Dai, D.; Wang, H.; Van Gool, L. MIC: Masked Image Consistency for Context-Enhanced Domain Adaptation. arXiv preprint, 2022.
19. Hoyer, L.; Dai, D.; Van Gool, L. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In: Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, 372-391, 2022
20. Xie, B.; Li, S.; Li, M.; Liu, C. H.; Huang, G.; Wang, G. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 2023.