Can Tho University Journal of Science
website: ctujs.ctu.edu.vn

DOI: 10.22144/ctu.jen.2022.042

A Vietnamese benchmark for vehicle detection and real-time empirical evaluation

Truc Trinh, and Khang Nguyen*
University of Information Technology, VNU-HCM, Viet Nam
Vietnam National University, Ho Chi Minh City, Viet Nam
**Correspondence: Khang Nguyen (email: khangnttm@uit.edu.vn)*

Article info.

Received 25 Feb 2022
Revised 04 May 2022
Accepted 20 Jun 2022

Keywords

Vehicle detection, real-time object detection, dataset, UIT-VinaDeveS22

ABSTRACT

The current situation of traffic in Vietnam has many outstanding problems, especially traffic congestion, since the supply of infrastructure has often not been able to keep up with the growth in mobility. Thus, proposing monitoring plans to support authorities to make suitable and prompt decisions has always received large attention from the community. Meanwhile, applying information technology, especially advanced models which could process or analyze traffic data in real time is recently considered to be a priority solution due to the time, accuracy, and cost saving that it can potentially achieve. Therefore, this paper outlines research on three advanced real-time object detection methods: YOLOX, YOLOF, and YOLACT and the development of the newest Vietnamese traffic dataset named UIT-VinaDeveS22. The work contains both theoretical and empirical analysis, which are expected to create premises for further studies into addressing problems such as traffic density management, traffic separation, and traffic congestion.

1. INTRODUCTION

It is the reality that economic growth leads to the increasing demand for the adaptation of traffic models and transportation infrastructures. This consequently places a lot of pressure on traffic management agencies with problems such as traffic accidents, traffic congestion, vehicle statistics, and urban transportation planning; therefore, timely and correct responses are urgently needed. The Vietnamese government also has orientations for national modernization by applying information systems to traffic management activities that aim to run the transportation industry towards digital transformation during the period of 2020 – 2025 with a vision towards 2030[1]. Especially, a project entitled "Applying information technology to traffic management and administration, focusing on the road sector" which deploys enhanced intelligent transport system (ITS) has been recently approved[2].

ITS is a system that integrates communication and data processing technologies and aims to create a control and information center. The system improves the mobility of people, goods and also the experience of using transportation systems. ITS also includes a flow control system, traffic management system, accident alarm system, etc., which almost all of them approach the problems by object detection, especially, detecting vehicles

[1] https://mt.gov.vn/vn/tin-tuc/67496/ung-dung-cntt-trong-quan-ly--dieu-hanh-giao-thong-van-tai--tap-trung-linh-vuc-duong-bo.aspx

[2] https://vanban.chinhphu.vn/default.aspx?pageid=27160&docid=200384

traveling on the road. Vehicle detection is a challenging problem since detections could be misunderstood, overlapped, or ignored. The reasons may be due to the increase in traffic volume during rush hour which makes the road section too crowded to observe individual vehicle; there are still many roads in Vietnam where all types of vehicles use same lane without a clear separation between different directions of vehicles, and consequently the appearance of vehicles could be overlapped when captured by closed-circuit television cameras (CCTV); The problems of redesigning private vehicles which is very popular in Vietnam could make common models changed and lead to the confusion among vehicles; In addition, the training process or the model architecture is also important factors that significantly affect the performance. Moreover, real time is also the most important factor because, although the analysis is right, if the opportunity is missed, everything will become meaningless, i.e., rescue situations where accidents have occurred or anomalies that were detected in the traffic environment.

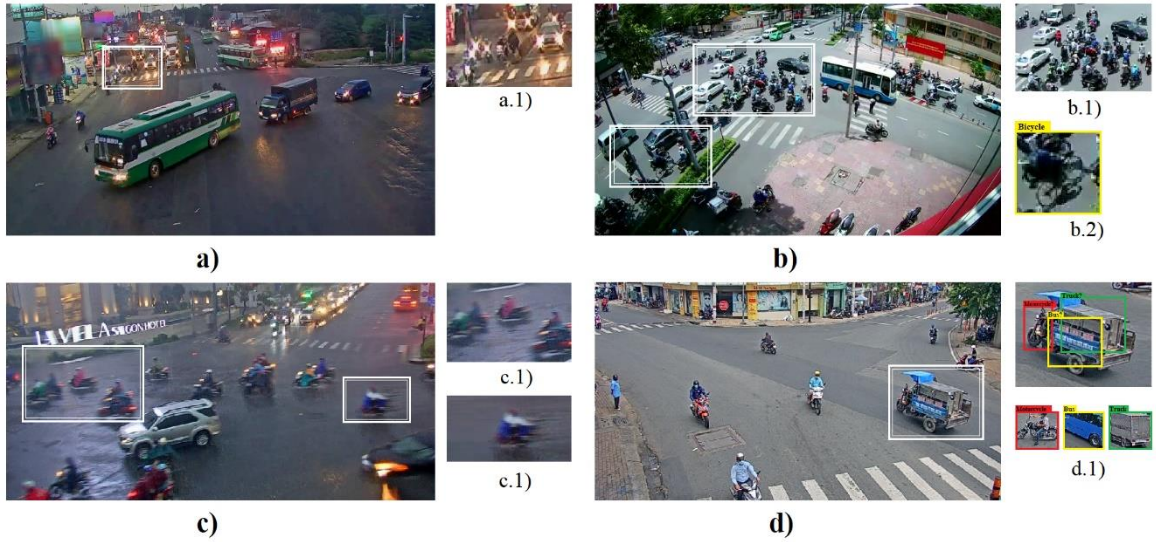

Figure 1. Samples from the dataset UIT-VinaDeveS22 highlighting possible challenges on the streets of Vietnam from CCTV view

Figure a shows the difficulty in detecting each motorbike when it's dark and the headlight of vehicle is on (Figure a.1). Figure b illustrates heavy traffic at crossroads where it is hard to observe clearly each vehicle (Figure b.1), besides, the bicycle which is a thin, small frame vehicle would be very hard to find if it is surrounded by other vehicles even by human eye (Figure b.2). In Figure c when it's dark and raining, most motorbikes tend to go fast which make its shape deformed when the video is captured (Figure c.1 and Figure c.3). In Figure d, although 'tricycle' is not include in the training set of UIT-VinaDeveS22, there are still some cargo motorized tricycles appearing in test set, which creates many misunderstandings with other vehicles, namely motorbike (red area), bus (yellow area), truck (green area) (Figure d.1).

To overcome the mentioned challenges, in addition to researching, surveying, and choosing a good object detection method, it is also necessary to build a dataset which is close to the context, so that the model could learn and process Vietnamese traffic data effectively. Thus, this study first focuses on researching three advanced real-time detection methods: YOLOX, YOLOF, YOLACT, then developing a Vietnam traffic dataset entitled UIT-VinaDeveS22 and finally conducting experiments and evaluations of the above methods within the dataset. This research will analyze the dataset and experimental result in detail which help to build a foundation for further applied research on addressing traffic problems in Vietnam.

2. RELATED WORKS

2.1. Traffic datasets in Vietnam

Thai et al. (2014) suggested a method for detecting motorbikes from the scenes, and they also built a dataset to evaluate the proposed method. The dataset only has a single type of object (motorbike) with a fixed camera from the vertical view port of a crossroad with a small angle of 15 degrees. Such settings allowed Thai et al. to capture the various illumination and 3D viewpoint changes of the motorbike. The dataset was also taken at different times of day with the intention to collect different levels of occlusion between motorbikes.

Huynh et al. (2016) introduced a Vietnamese vehicle dataset that was collected from two different data sources: one of them is selected from the article (Thai et al., 2014) and the other was cut from the video[3]. With a fixed camera angle, taken at top-down and to the right of objects, the five-minute video has captured a small area, but dense flow of vehicles in daytime and includes 1600 frames.

Dinh et al. (2016) have collected motors and cars to analyze the traffic jam and then solve detection problems by their new methods. They set fixed angle cameras to record the flow of moving vehicles in two different straight ways in daytime.

Trinh et al. (2021) have published a ten thousand – sample dataset about Vietnamese traffic which was captured by drone. Although this dataset seemed to be the latest and presents quite well Vietnamese vehicles (daytime, night, clear, foggy) and different contexts (samples are extracted from 18 videos), it still has some limitations: Firstly, the dataset was especially made for the classification task which means that each sample contains only one vehicle and this is completely not suitable for addressing the object detection problem; Secondly, although this dataset was built for Vietnamese context, its categories only include bus, car, truck, van without any motorcycles, while this remains the most popular vehicle in Vietnam.

2.2. Advanced real-time object detection

2.2.1. YOLOX

In 2021, Ge et al. (2021) introduced YOLOX as a superior method that improves the YOLO series. They have switched YOLO detector to an anchor-free mechanism and also conduct other advanced detection techniques which are a decoupled head and the leading label assignment SimOTA.

Anchor-free is a mechanism that solves the previous problems of anchor-based detectors: First, to get the best results, the previous methods conduct a clustering analysis to determine a set of optimal anchor boxes before training. However, these optimal anchor boxes are actually only good for domain specific data which is less generalized; Second, the anchor box mechanism increases the number of predictions for each image that completely makes detection heads more complex. Moreover, anchor-free mechanism also reduces a considerable number of parameters that need heuristic or many tricks involved (e.g., Clustering (Redmon & Farhadi, 2017), Grid Sensitive (Huang et al., 2021)), which could make training and decoding phase significantly simpler (Tian et al., 2019).

Moreover, YOLOX separates detection into two different tasks which correspond to two branches of the decouple head: classification and localization. These two tasks are a well-known conflict because features suitable for classification and localization are not the same and bounding boxes with high classification confidences are not sure that will bring a high intersection over unions (IoU) between the predicted location and the groundtruth (Jiang et al., 2018; S. Wu et al., 2019; Song et al., 2020; Y. Wu et al., 2020).

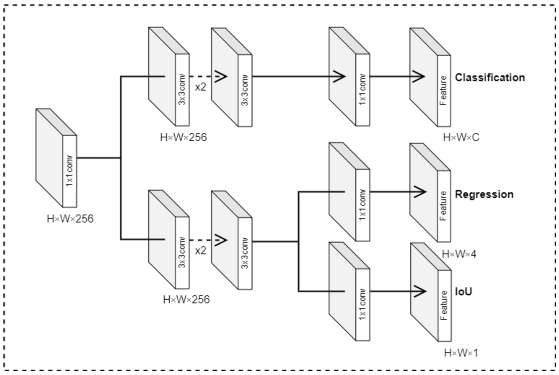

Figure 2. Decouple head architecture of YOLOX

In regards to SimOTA, in the old version – OTA (Ge et al., 2021), it analyzes the label assignment from a global perspective and formulates the assigning procedure as an Optimal Transport (OT) problem. Authors have switched to a dynamic top-k strategy to solve the OT problem since it could avoid the increase of training time by 25% compared to the old algorithm Sinkhorn – Knopp.

YOLOX has many variations which depend on the complexity of architecture such as YOLOX-tiny, YOLOX-s, YOLOX-l, and YOLO-x.

[3] https://www.youtube.com/watch?v=Op1hdgzmhXM

2.2.2. YOLOF

Chen et al. (2021) have indicated that the success of feature pyramid networks (FPN) comes from the principle of divide and conquer rather than multi-scale feature fusion. From that perspective, they introduced YOLOF (You Only Look One-level Feature) by trying to conduct an alternative method that uses only one single level feature for detection instead of using a complex pyramid feature extraction as FPN, but still achieves comparable results compared to RetinaNet (Lin et al., 2017) – a counterpart which also uses feature pyramid.

However, it is easy to realize that using just a simple level feature could drop the performance intensively. Authors also pointed out 2 reasons for that: First, the range of scales matching to one feature's receptive field is limited, so that after training with just a level feature, models may not recognize other objects whose scale is not matched to the scale that model has learned; Second, sparse anchors in just a single-level feature could lead to the imbalance problem on positive anchors.

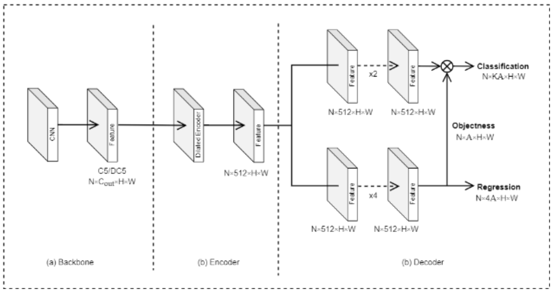

Figure 3. YOLOF architecture includes a backbone, a dilated encoder and a decoder

To address the problems mentioned above, YOLOF was built by two key components: Dilated Encoder and Uniform Matching. In particular, Dilated Encoder is responsible for enlarging the receptive field of the C5 feature by stacking standard and dilated convolutions through residual blocks with dilations on the middle 3 × 3 convolution layer in order to generate an output feature with various receptive fields, compensating for the lack of multiple-level features. About Uniform Matching, it is a mechanism to solve the imbalance problem in positive anchors which make sure that all ground-truth boxes participate in training and contribute equally.

2.2.3. YOLACT

YOLACT (Bolya et al., 2019) was introduced as a one stage approach for real-time instance segmentation which aids performance more effectively over time, but still has a competitive accuracy. YOLACT uses RetinaNet as a baseline, but improves it by adding a branch for predicting coefficients and applying fast NMS. However, the problem in this study is just approached at the bounding box level, which leads to the removal of the mask branch, but still applies Fast NMS for the experiment.

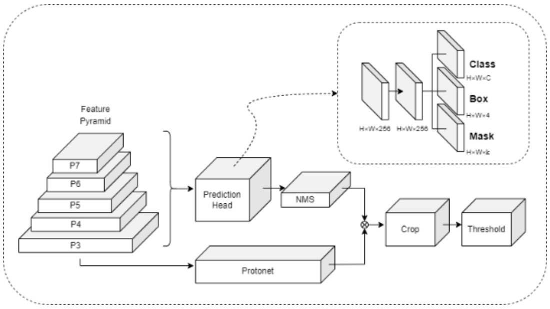

Figure 4. Full YOLACT architecture includes components to predict class, box, and mask of the objects

NMS is an algorithm that eliminates predictions that are overlapped. In the previous works, NMS was approached in a sequence that considers eliminating a prediction by comparing it with each box for each class. However, to enhance the processing speed, FastNMS tries to keep or discard boxes in parallel, which is presented below:

- First, sort all predictions and choose top n detections that have the highest confidence score.
- Then, initialize a $c \times n \times n$ pairwise IoU matrix X, which c is a number of classes and n is number of predictions. Values in matrix X were the IOU scores of each determined detection couple and was sorted descending by score for each of c classes.
- After initialization, predictions are removed if its corresponding IoU is higher than a threshold t. IoU is a term which could show how much two boxes are overlapped. But before that, it has to set the value in the lower triangle and diagonal of X to 0. Because in the lower triangle, there are comparisons among predictions with itself so the IoU will always be 1 and be eliminated if its value was not set to be 0 first.

After the matrix processing steps, there is finally a set of detections with a high confidence score, while its overlap has been discarded simultaneously.

3. UIT-VINADEVES22 DATASET

3.1. Overview

UIT-VinaDeveS22 has 1364 images captured from frames in videos which were collected by CCTV. The resolution of this dataset is about 553×1012 pixel which its smallest is 354×630 pixel and largest is 720×1280 pixels. Data was collected at daytime and night which weather could be clear or rain; The traffic density might be dense, intermediate or sparse depending on time of day or kind of road; The type of collected vehicles are bicycle, motorcycle, car, van, truck, bus, fire truck.

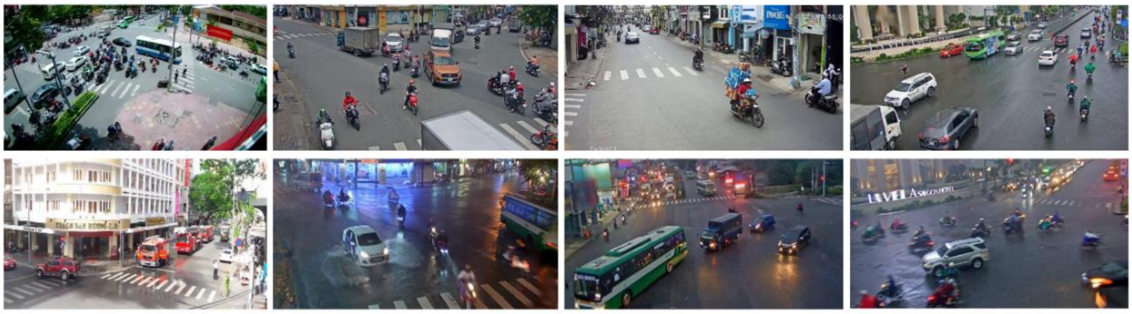

Figure 5. Some samples in UIT-VinaDeveS22 dataset

UIT-VinaDeveS22 is a fairly rare dataset in the present time that collects data from the Vietnamese Department of Transportation portal which presents the most truthfully and also the most easy to apply for the reality of the Vietnamese traffic situations in the near future. In UIT-VinaDeveS22:

- Number of videos captured: UIT-VinaDeveS22 shows various scenes, and the number of recorded sources is 6 different videos. This gives our dataset a comprehensive view of Vietnamese streets from many types of road and road junctions, such as crossroads, single carriageways or dual carriageways, in which the traffic density absolutely performs a big gap.
- Illuminate/Weather Conditions: Our image data is also dominated by illumination and weather conditions i.e., daytime, night, clear and rain, which pose many challenges in detecting phase, but represent the performance in real situations.

Although the scale mentioned above has many limitations (the number of collected videos less the previous works or not including fog in the weather conditions), UIT-VinaDeveS22 is still potential for enlarging and applying in the way it was exploited on an existing monitoring system.

3.2. Analysis

UIT-VinaDeveS22 includes:

- 1364 images: 653 in train set for training models; 173 in validation set for validating each epoch through the training; 538 in test set for testing final models.
- 15418 objects: 7724 in train set; 1765 in validation set; 5929 in test set; the number of objects in each class was distributed randomly (Table 1).

Table 1. Number of objects in each category for each set

Class	Test	Train	Valid	Total
Bicycle	35	163	35	233
Motorcycle	4002	4505	951	9458
Car	1357	2099	511	3967
Van	77	206	60	343
Truck	270	287	74	631
Bus	75	226	53	354
Fire truck	113	238	81	432
Total	5929	7724	1765	15418

Images were extracted from the frames of 6 videos which have different background landscapes, weather and illuminating conditions. Samples in the dataset were clearly separated into the training set and test set to get the fairest assessment. In particular, both training and test set contain images at daytime and night but the daytime and night images of two sets were extracted from different videos which have completely different scenes. Samples in validation set and training set were extracted from same videos but were different frames.

4. EXPERIMENT

4.1. Metrics

The mAP, which is a popular metric to measure the performance of models in terms of accuracy, is used for the evaluation. mAP (mean Average Precision) is calculated by taking the mean AP of all over the classes when AP gets the average AP based on a specific threshold value in the range of IoU threshold from 50 (AP_{50}) to 75 (AP_{50}) for a specific class. Moreover, AP is also known as a metric that represents the Precision-Recall curve by calculating the weighted mean of precision that is achieved at each threshold, with the increase in recall at the current threshold n from the previous threshold ($n -$ 1) which is used as the weight:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n$$

Where R_n and P_n are Precision and Recall at the threshold n, respectively. In the object detection problem, Precision and Recall are calculated by using IoU value for a given IoU threshold.

IoU is the Intersection over Union which is a standard for evaluating overlap between predicted bounding box and the ground truth which is represented in Figure 5.

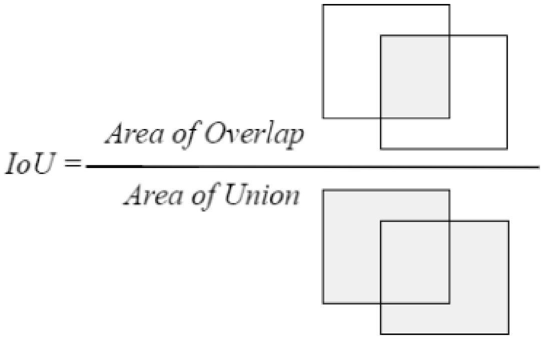

Figure 5. Illustration of IoU

4.2. Implementation

The experiments were conducted on YOLOF, YOLACT and YOLOX to evaluate the real time and accuracy of these three methods on the reality Vietnam traffic situations via UIT-VinaDeveS22. The original RetinaNet, which is a baseline as known as the theoretical basis of the above methods, was also employed to the experiment for a clear comparative analysis.

In detail, ResNet50 (He et al., 2016) was used as a backbone for YOLOF and ResNet101 (He et al., 2016) for RetinaNet and YOLACT. However, due to the fact that YOLOX was built on YOLOv3 (Redmon & Farhadi, 2018) toward to YOLOv4, YOLOv5 baseline so the architecture of Darknet CSP (Bochkovskiy et al., 2020) is adopted as backbone for it. About the neck of method architecture, both YOLACT and YOLOX are based on RetinaNet that use Feature Pyramid Network as a module to extract multiscale features. However, YOLOF has simplified the extract feature process in multi-level feature by just using one single level feature, so that the neck of YOLOF applies a Dilated Encoder aiming to enlarging scale of reception field for learning more scale of objects in image. The entire experiments were trained on MMDetection framework (Chen et al., 2019) with its default hyperparameters. The detailed schedule training is represented in Table 2.

About machine configuration, the implementation was run on Google Colaboratory which use: 1) CPU: 1×single core hyper threaded Xeon Processors @2.3Ghz; 2) GPU: 1×Tesla K80, compute 3.7, having 2496 CUDA cores, 12GB GDDR5 VRAM; 3) RAM: ~12.6 GB; 4) Disk: ~33 GB.

Table 2. Detail of schedule training

Method	Optimizer function	Learning rate	Momentum	Weight_decay	Epoch
YOLOX	*SGD*	0.01	0.9	0.0005	300
YOLOF	*SGD*	0.12	0.9	0.0001	12
YOLACT	*SGD*	0.001	0.9	0.0005	55
RetinaNet	*SGD*	0.001	0.9	0.0001	12

5. RESULTS AND DISCUSSION

In terms of real time, YOLOX gave the fastest performance due to its anchor free mechanism and an advanced assigning strategy – SimOTA. Particularly, the processing speed of YOLOX is 25.1 fps, followed by 6.4 fps of YOLACT and 4.9 fps of YOLOF. To compare with results of the original baseline, FastNMS of YOLACT proved its significant efficiency in time which was at least three times as fast as that for RetinaNet (2.1 fps), besides, YOLOF also was two times faster than RetinaNet by using a dilated encoder replacing a complex multi-level feature extraction (Table 3).

In regards to accuracy, considering mAP of all methods, it is not hard to see that although using a full feature pyramid network architecture as RetinaNet, YOLOX still trade off its accuracy against the 10 times faster speed (compared to 2.1

fps of RetinaNet) when using anchor free mechanism. Especially, mAP of YOLOX was 0.175, which was just a half for 0.284 of YOLACT, 0.306 of YOLOF and 0.278 of RetinaNet (Table 3). Although not using multi-level to extract features as FPN, the Dilated and Uniform matching mechanism of YOLOF have done a great job when obtain a best performance as having the highest mAP and being in a third place of the processing speed (Table 3).

To go into details of each class, about bicycle, this is the class containing the fewest number of training samples which was just 163 compared to 7724 samples in the overall the dataset. Moreover, this type of vehicle is also small, thin and very difficult to find (even by human eye) when captured from CCTV (Table 1). However, YOLOF's dilated and uniform matching mechanism worked quite well, which was the only method that correctly detected bicycles in some cases (0.001 AP on bicycle). Details are in Table 4.

Table 3. Experimental results of accuracy (mAP) and time (fps) evaluating 4 methods of real-time object detection YOLOX, YOLOF, YOLACT, RetinaNet on UIT-VinaDeveS22 dataset where mAP is mean Average Precision and fps is frame per second

Method	mAP	fps	
YOLOX	0.175	*25.1*	
YOLOF	0.306	*4.9*	
YOLACT	0.284	*6.4*	
RetinaNet		*0.278*	*2.1*

Table 4. Specific accuracy of 4 methods YOLOX, YOLOF, YOLACT, RetinaNet for each class (AP)

AP **Method**	bicycle	van	fire truck	motorcycle	truck	car	bus
YOLOX	0	0.038	0.621	0.096	0.131	0.288	0.052
YOLOF	0.001	0.137	0.707	0.216	0.333	0.547	0.198
YOLACT	0	0.121	0.595	0.183	0.325	0.488	0.276
RetinaNet	0	0.102	0.635	0.267	0.313	0.523	0.104

6. CONCLUSION

This paper outlines the development of UIT-VinaDeveS22, which was built as the latest Vietnam traffic dataset, and its potential in the way of collecting from the existing CCTV system was analyzed. Moreover, four experiments were conducted to evaluate and compare three real-time advance methods with one baseline (RetinaNet) to assess the challenges, as well as the feasibility of applying advanced traffic analysis systems in Vietnam.

The results show that YOLOF and YOLACT with bounding box approaching version have the best trade-off between time and accuracy compared to YOLOX which is five-time faster but showed the worst accuracy performance.

In the future, we hope to expand the dataset to include characteristics of Vietnam traffic as realistic as possible. Moreover, from what has been discussed above, we hope to combine those advanced techniques to create a better architecture to improve the performance not only in time but also in accuracy.

ACKNOWLEDGMENT

We would appreciate the Multimedia Processing Lab (MMLab) at the University of Information Technology, VNUHCM for their support. We are also grateful for anonymous reviewers who read and leave comments on this paper.

REFERENCES

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *ArXiv Preprint ArXiv:2004.10934*.

Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2019). Yolact: Real-time instance segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9157–9166.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., & Xu, J. (2019). MMDetection: Open mmlab detection toolbox and benchmark. *ArXiv Preprint ArXiv:1906.07155*.

Chen, Q., Wang, Y., Yang, T., Zhang, X., Cheng, J., & Sun, J. (2021). You only look one-level feature. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13039–13048.

Dinh, V.-T., Luu, N.-D., & Trinh, H.-H. (2016). Vehicle classification and detection based coarse data for warning traffic jam in VietNam. *2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS)*, 223–228.

Ge, Z., Liu, S., Li, Z., Yoshie, O., & Sun, J. (2021). Ota: Optimal transport assignment for object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 303–312.

Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). Yolox: Exceeding yolo series in 2021. *ArXiv Preprint ArXiv:2107.08430*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Huang, X., Wang, X., Lv, W., Bai, X., Long, X., Deng, K., Dang, Q., Han, S., Liu, Q., & Hu, X. (2021). PP-YOLOv2: A practical object detector. *ArXiv Preprint ArXiv:2104.10419*.

Huynh, C.-K., Le, T.-S., & Hamamoto, K. (2016). Convolutional neural network for motorbike detection in dense traffic. *2016 IEEE Sixth International Conference on Communications and Electronics (ICCE)*, 369–374.

Jiang, B., Luo, R., Mao, J., Xiao, T., & Jiang, Y. (2018). Acquisition of localization confidence for accurate object detection. *Proceedings of the European Conference on Computer Vision (ECCV)*, 784–799.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *ArXiv Preprint ArXiv:1804.02767*.

Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7263–7271.

Song, G., Liu, Y., & Wang, X. (2020). Revisiting the sibling head in object detector. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11563–11572.

Thai, N. D., Le, T. S., Thoai, N., & Hamamoto, K. (2014). Learning bag of visual words for motorbike detection. *2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*, 1045–1050.

Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9627–9636.

Wu, S., Yang, J., Wang, X., & Li, X. (2019). Iou-balanced loss functions for single-stage object detection. *ArXiv Preprint ArXiv:1908.05641*.

Wu, Y., Chen, Y., Yuan, L., Liu, Z., Wang, L., Li, H., & Fu, Y. (2020). Rethinking classification and localization for object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10186–10195.