Real-time Vehicle Detection from UAV Imagery

Xuemei Xie*, Wenzhe Yang, Guimei Cao, Jianxiu Yang, Zhifu Zhao, Shu Chen, Quan Liao, Guangming Shi
School of Artificial Intelligence
Xidian University, Xi'an, China 710071
xmxie@mail.xidian.edu.cn

***Abstract*—Fast and accurate vehicle detection in unmanned aerial vehicle (UAV) imagery is a meaningful but challenging task, playing an important role in a wide range of applications. Due to its tiny size, few features, variable scales and imbalance vehicle sample problems in UAV imagery, current deep learning methods used in this task cannot achieve a satisfactory performance both in accuracy and speed, which is obvious a classical trade-off problem. In this paper, we propose a single-shot vehicle detector, which focuses on accurate and real-time vehicle detection in UAV imagery. We make contributions in the following two aspects: 1) presenting a multi-scale feature fusion module to combine the high resolution but semantically weak features with the low resolution but semantically strong features, aiming to introduce context information to enhance the feature representation of the small vehicles; 2) proposing a dynamic training strategy (DTS) which constructs the network to learn more discriminative features of hard examples, via using cross entropy and focal loss function alternately. Experimental results show that our method can achieve 90.8% accuracy in UAV images and can run at 59 FPS on a single NVIDIA 1080Ti GPU for the small vehicle detection in UAV images.**

***Index Terms*—vehicle detection, unmanned aerial vehicle imagery, feature fusion, dynamic training strategy**

I. Introduction

Nowadays, vehicle detection in unmanned aerial vehicles (UAV) imagery plays a significant role for a wide range of applications [1]–[3]. However, there are some negative characteristics in real-time vehicle detection from UAV imagery, tiny objects, various orientation of the targets, and imbalance samples, which lead to unsatisfactory performance both in speed and accuracy.

Traditional methods are mainly based on the handcrafted features [4], [5] and sliding window search algorithms [6], [7]. The handcrafted features cannot extract good semantic representation. Some following studies [8], [9] exploit deep learning methods to improve the feature representation capability compared with handcrafted ones, bringing certain improvement in detection accuracy. But there is still a gap to real-time detection. Faster R-CNN [10], one of CNN-based detectors, has achieved a good performance in UAV imagery [11]–[14]. While, it has a limitation in speed due to its detection mechanism. Subsequently, YOLOs [15], [16] are employed to achieve real-time detection with lower accurate [17]. Due to the wide range of view of UAV images, the vehicle objects are usually small, occluded and with complex background. In the context of the situations, accurately detecting the vehicles from UAV imagery is quite difficult.

In this paper, we propose a single shot network using multi-level feature fusion method which utilizes context information efficiently and effectively, make a certain progress in accuracy and achieve real-time vehicle detection simultaneously. Moreover, the extremely hard-easy class imbalance in UAV dataset causes two problems as follows: 1) model training is insufficient for the categories which with a small amount of examples, so that it is hard for the network to extract representative features [18], [19]; 2) most easy samples will overwhelm the total loss and gradients computation so the network cannot learn the discriminative features well [20]. To solve these, we design a **dynamic training strategy** (DTS) to solve the imbalance problem and improve the network detection performance.

To summarize, we present a single-shot detector, which focuses on accurate and real-time vehicle detection from UAV imagery. Specifically, our main contributions are as follows:

- We present a multi-scale feature fusion module to combine the high resolution but semantically weak features with the low resolution but semantically strong features, which aims to introduce context information to enhance feature representation of the small vehicles;
- We propose a dynamic training strategy (DTS) which instruct the network to learn more discriminative features of hard examples, via using cross entropy and focal loss function alternately;

Experimental results show that our method can achieve 90.8% accuracy which is 7.5% and 3.1% higher than SSD [21] and RefineDet [22] respectively in UAV images. And the proposed network can run at 59 FPS on a single NVIDIA 1080Ti GPU for the small vehicle detection.

II. Related Work

A. UAV Vehicle Detector

Vehicle detection from UAV imagery has attracted extensive research attention in past years. Moranduzzo et al. [23], Shao et al. [4] and Kembhavi et al. [6] explore the vehicle detection by using handcrafted features (e.g., Haar, HOG, SIFT, local binary pattern, etc.) and intersection kernel SVM, which make some progress. Xu et al. [14] improves original Viola-Jones object detection scheme for better performance from low-altitude UAV imagery. However, traditional handcrafted

*This work is supported by Natural Science Foundation (NSF) of China (Nos.61472301, 61632019), the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (No. 61621005), Ministry of Education project (No. 6141A02011601).

978-1-5386-5321-0/18/$31.00 ©2018 IEEE

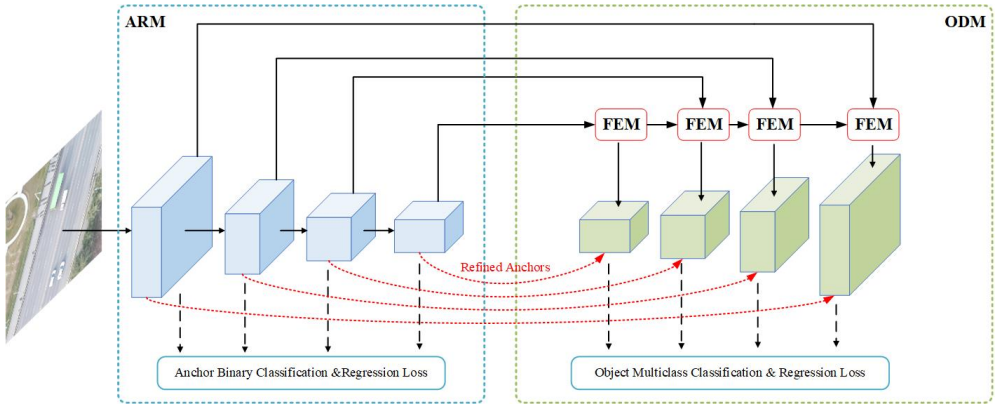

Fig. 1. The overview of the proposed architecture, including two sub-modules, i.e. Anchor Refinement Module (ARM) and Objects Detection Module (ODM). ARM generates dense default boxes and filter out easy negative positions. Then ARM will pass on the selected boxes to the ODM. ODM refines the selected boxes positions to suit the targets.

features are not representative enough for the high accuracy detection.

Additionally, vehicle detection in UAV imagery has inherited some achievements from generic object detection tasks. Xu et al. [11] and Sommer et al. [12] use simple Faster RCNN for different scene and improve the detection accuracy, but the detection speed is still slow. Lin et al. [13] segments the UAV images based on SegNet [24] to filter the small vehicle objects and then detect them accurately. Xu et al. [14] is the first research to investigate the use of YOLOv2 [16] for vehicle detection in UAV imagery, which can reach 21 FPS and 77.12% mAP in their own dataset.

B. Real-time Detector for Small Objects

Recent years there are many studies aiming to improve the detection speed and accuracy. YOLOs [15], [16] use a single feed-forward convolutional network to directly predict objects locations and categories, which is extremely fast. But the detection accuracy of these methods is not satisfactory. SSD [21] spreads out anchors of different scales to multiple layers and gets a good performance both in accuracy and speed. But both YOLOs and SSD perform weakly on small objects detection. To improve this, Zhang et al. [19], [22] and Hu et al. [25] use the top-down feature fusion model to enhance the small objects feature representation, which lead to better performance on small objects.

C. Class Imbalance

There are many strategies used in recent detectors to solve the class imbalance problem. Ren et al. [10], Uijlings et al. [26] and Zitnick et al. [27] rapidly narrow down the number of bounding boxes to a small number (i.e. 1∼2k), filtering out most background samples. In SSD [21], the network samples the boxes as a fixed foreground-to-background ratio (i.e. 1:3), and uses the online hard example mining [18]. Lin et al. [20] and Badrinarayanan et al. [24] present novel loss functions to deal with class imbalance, which can dynamic avoid the influence of easy negative bounding boxes.

III. Method

The overview framework of our proposed vehicle detector is shown in Figure 1. It contains two main components: feature-enhanced framework and dynamic training strategy.

A. Feature-Enhanced Framework

The proposed feature-enhanced framework for UAV vehicle detection is based on the state-of-the-art general detector RefineDet [22]. To improve the robustness of small vehicles with various orientation, we design suitable small default boxes which efficiently discretize the searching space for possible small vehicle shapes, so that the small size vehicles have enough matched default boxes for sufficient training. Moreover, the objects from UAV imagery are usually with weak features because of the small size, various orientation and the complex background. We propose an effective feature fusion module to enhance the feature representation capacity by exploiting the context information smartly. The enhanced powerful feature representation guarantees the significant improvement in detection accuracy.

TABLE I
THE ASPECT RATIO SETTINGS

Detection layers	scales	Aspect Radios
Conv3_3	16	1:1, 1:2, 2:1
Conv4_3	32	1:1, 1:2, 2:1
Conv5_3	64	1:1, 1:2, 2:1, 4:1
Fc7	128	2:1, 4:1

Constructing architecture. As shown in Figure 1, we construct our architecture based on the state-of-the-art RefineDet framework, and then design proper detection layers and default boxes settings, which are essential for high detection accuracy.

- Base RefineDet network. The backbone network based on RefineDet includes Anchor Refinement Module (ARM) and Object Detection Module (ODM). We keep the convolutional layers from 'conv1_1' to 'conv7' and remove the remaining layers both in ARM and ODM because

the too deeper convolutional layers behind is helpless for small objects detection.

- Detection layers. We select conv3_3, conv4_3, conv5_3 and fc7 as the detection convolution layers, which are associated with different scales of default boxes to predict detections.
- Suitable boxes design. Refer to [19], the scales in each detection layer should match the effective receptive filed (ERF) other than the theoretical receptive field (TRF). Given this, we reset the scales in our architecture according to the ERF computed by network visualization method [28]. Moreover, we adopt k-means clustering method to analyse the most suitable aspect ratios of the vehicles. The result is shown in TABLE I.

Feature-enhanced module. Instead of original detection layers directly generate default boxes, we design Feature-Enhanced Module (FEM) to each detection layers, enhanced feature maps and refine the object locations. The FEM will introduce context information smartly by combining inherent multi-level features and incorporate the useful multi-level features from high spatial resolution but low semantic and low spatial resolution but high semantic features, leading to more accurate object detection.

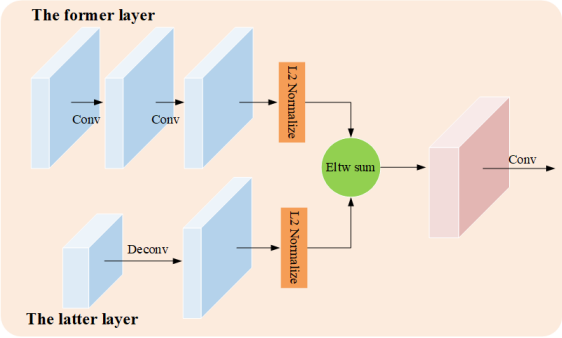

Fig. 2. Feature-Enhanced Module. An eltwise_sum layer combines two different transferred spatial features.

As shown in Figure 2, we use a deconvolution layer on the latter layer to make it the same resolution against the former layer. Besides, two convolution layers with same resolutions are added to the former layer, which extract more semantic features and transform its feature into the uniform space with the features. At last, an eltwise-sum layer sums different layers and output the enhanced features for the following detection.

B. Dynamic Training Strategy

In this paper, we proposed **dynamic training strategy** (DTS), which alternately train models with cross entropy loss and focal loss, can extract more discriminative features and lighten the easy examples influence. As mentioned in Section I, there are two imbalance problems during training in previous detectors. The previous detectors must process a large set of candidate object locations ($\sim$100k) which are regularly sampled across an image and these candidates contain a large amount of easy classified locations. Hence, the sum of these vastly easy examples will overwhelm the total loss and the computed gradients, resulting in insufficient training for those hard examples with weak features or small numbers.

Inspired from Lin et al. [20], we first try the α-balanced variant focal loss (FL) which can automatically down weighting the contribution of well-classified examples to the total loss and focus on hard examples during training. The focal loss function is defined as:

$$FL(p_t) = \alpha(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

where $\alpha \geq 0$, $\gamma \geq 0$ are tunable focusing parameters, $(1-p_t)^{\gamma}$ is a modulating factor by γ, and p_t is defined as:

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise.} \end{cases} \tag{2}$$

$y \in \{\pm 1\}$ specifies the ground truth class and $p \in [0,1]$ is the estimated probability for the class with label $y = 1$.

Unfortunately, it does not work well directly applying the FL on our framework, because most hard examples lack enough informative features for training. So, we propose DTS, which first train models with CE loss to construct the training process without paying more attention to hard examples, in order to learn representative features of each category. Then, in order to enhance the discriminability of the hard examples' extracted features, we finetune the trained models by using focal loss to decrease the contribution of well-classified examples, increase the contribution of miss-classified examples and make the hard examples training sufficiently. The following experiments show that the effect of DTS is better for the small vehicle detection in UAV images.

IV. Experiments

In this section, we evaluate the performance of each model on our UAV image dataset. All experiments are implemented based on Caffe, and we train all models with batch size 16 on a single NVIDIA GTX-1080Ti GPU and Intel Core i7-7700K@4.2GHz. We first evaluate existing architectures [16], [21], [22] on UAV dataset, and the results are shown in TABLE II. Moreover, we evaluate proposed models' performance and analyze the effectiveness of FEM and DTS. As examples, some results produced by the best performance model are shown in Figure 4.

A. Data Preparation

We capture some UAV vehicle videos from four simple scenes. In each scene, UAV is in a stationary height (i.e. 100m). We totally get 4344 images from all videos and annotate 6 categories of the on-road vehicles (i.e. car, bus, truck, tanker, motor and bicycle), and the number of each class is shown in TABLE III. We divide them to train dataset (3474 images) and test dataset (869 images), Figure 3 shows some image examples. We also use some data augmentation strategies while training [21].

TABLE II
DETECTION RESULTS EVALUATED BY EXISTING MODELS

Method	Input Size	mAP	AP of each class						FPS
			Car	Bus	Truck	Motor	Bicycle	Tanker	
Yolo [16]	448 × 448	53.7	66.4	73.1	77.0	29.3	12.2	64.1	**64**
SSD [21]	300 × 300	83.3	90.7	90.0	90.3	78.3	60.7	89.60	59
RefineDet [22]	320 × 320	87.1	**90.8**	**90.3**	90.6	80.2	80.5	90.6	40
Our proposed	320 × 320	**90.8**	**90.8**	**90.3**	**90.7**	**89.4**	**87.4**	**95.9**	59

Fig. 3. Illustrative examples of the UAV dataset.

B. Overall Performance

Firstly, we only use the focal loss (FL) during total training stage to find out the best parameters (i.e. α and γ) for our model. We set $\alpha = 1.0$, evaluate different models with the varying γ. As shown in TABLE V, the best γ should be set to 1.0. On the other hand, when γ is set to the best value (i.e. $\gamma = 1.0$), we can find in TABLE V that with α close to 1.0, the models performance become better and better. As a consequence, when $\alpha = 1$ and $\gamma = 1$, the accuracy of vehicle detection can come to 89.4%.

TABLE III
THE NUMBER OF EACH CLASS' OBJECTS

Class	Car	Bus	Truck	Motor	Bicycle	Tanker
Number	33841	2690	2848	6656	2024	173

Moreover, as mentioned in Section III, DTS which alternately uses CE and FL function during training can learn more discriminative features. To explore how different parameters influence the final result, we train the model using CE loss for 100k iterations and 20k iterations using FL. From TABLE VI we can find that the best parameters (i.e. $\alpha = 1, \gamma = 1$) can help the model achieve 90.8% accuracy. The results shown in TABLE IV use different training iterations with CE and FL, and 100k CE with 20k FL iterations is the best choice. To summarize, we can achieve 90.8% accuracy which is 1.4% higher using DTS than only using FL loss.

TABLE IV
VARYING ITERATIONS VIA DTS

CE iters FL iters	120k 0	100k 20k	80k 40k	60k 60k	40k 80k
mAP	89.3	**90.8**	89.7	89.6	89.6

C. Model Comparison

We train the images using existing fast and accurate frameworks comparing with our proposed architecture. All models are trained for 120k iterations.

TABLE V
VARYING α,γ VIA FOCAL LOSS(CE = 0, FL = 120K)

α	γ	mAP
1.0	0.0	89.3
	1.0	**89.4**
	2.0	87.5
	3.0	87.5
0.5	1.0	88.1
0.75		88.6
0.9		88.7
1.0		**89.4**

TABLE VI
VARYING α,γ VIA DTS(CE = 100K, FL = 20K)

α	γ	mAP
1.0	0.0	89.3
	1.0	**90.8**
	2.0	89.3
	3.0	89.3
0.5	1.0	89.8
0.75		89.9
0.9		90.6
1.0		**90.8**

For first 80k iterations, we use 10^{-3} learning rate, and then decay it to 10^{-4} for the following 40k iterations. All models are finetuned from the VGG16 model which is pretrained on ILSVRC dataset.

As shown in TABLE II, our method achieves 90.8% mAP, which is 37.1%, 7.5% and 3.7% better than YOLOv2 [16], SSD [21] and RefineDet [22] respectively. In particularly, the progress of small objects (i.e. 'motor' and 'bicycle') can prove that our proposed network can learn more discriminative features of hard examples which are benefit for detection. Moreover, our method processes an image in 17ms (59 FPS) via a single NVIDIA GTX 1080Ti GPU, achieving the real-time UAV detection with high accuracy.

V. CONCLUSION

In this paper, we present a single-shot object detection network which aims to detect UAV vehicles with high accuracy. We first propose a multi-scale feature-enhanced module, which combines the high resolution, semantically weak features with

Fig. 4. Examples of detection results evaluated by purposed model.

the low resolution, semantically strong features, utilizing more context information to enrich the feature representation of the small vehicles. We also introduce DTS which uses cross entropy and focal loss function alternately so that the network can learn more discriminative features of hard examples, leading to higher detection accuracy. Experimental results show that the proposed architecture can achieve state-of-the-art results with real-time detection.

REFERENCES

[1] Albert Yu-Min Lin, Alexandre Novo, Shay Har-Noy, Nathan D Ricklin, and Kostas Stamatiou. Combining geoeye-1 satellite remote sensing, uav aerial imaging, and geophysical surveys in anomaly detection applied to archaeology. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4(4):870–876, 2011.

[2] B Kršák, P Blišťan, A Pauliková, P Puškárová, Ľ Kovanič, J Palková, and V Zelizňaková. Use of low-cost uav photogrammetry to analyze the accuracy of a digital elevation model in a case study. *Measurement*, 91:276–287, 2016.

[3] Shaodan Li, Hong Tang, Shi He, Yang Shu, Ting Mao, Jing Li, and Zhihua Xu. Unsupervised detection of earthquake-triggered roof-holes from uav images using joint color and shape features. *IEEE Geoscience and Remote Sensing Letters*, 12(9):1823–1827, 2015.

[4] Thomas Moranduzzo and Farid Melgani. Detecting cars in uav images with a catalog-based approach. *IEEE Transactions on Geoscience and Remote Sensing*, 52(10):6356–6367, 2014.

[5] Tao Zhao and Ram Nevatia. Car detection in low resolution aerial images. *Image and Vision Computing*, 21(8):693–703, 2003.

[6] Wen Shao, Wen Yang, Gang Liu, and Jie Liu. Car detection from high-resolution aerial imagery using multiple features. In *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, pages 4379–4382. IEEE, 2012.

[7] Xueyun Chen, Shiming Xiang, Cheng-Lin Liu, and Chun-Hong Pan. Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geoscience and remote sensing letters*, 11(10):1797–1801, 2014.

[8] Matija Radovic, Offei Adarkwa, and Qiaosong Wang. Object recognition in aerial images using convolutional neural networks. *Journal of Imaging*, 3(2):21, 2017.

[9] Nassim Ammour, Haikel Alhichri, Yakoub Bazi, Bilel Benjdira, Naif Alajlan, and Mansour Zuair. Deep learning approach for car detection in uav imagery. *Remote Sensing*, 9(4):312, 2017.

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[11] Yongzheng Xu, Guizhen Yu, Yunpeng Wang, Xinkai Wu, and Yalong Ma. Car detection from low-altitude uav imagery with the faster r-cnn. *Journal of Advanced Transportation*, 2017, 2017.

[12] Lars W Sommer, Tobias Schuchert, and Jürgen Beyerer. Deep learning based multi-category object detection in aerial images. In *Automatic Target Recognition XXVII*, volume 10202, page 1020209. International Society for Optics and Photonics, 2017.

[13] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images. *Remote Sensing*, 9(4):368, 2017.

[14] Yongzheng Xu, Guizhen Yu, Xinkai Wu, Yunpeng Wang, and Yalong Ma. An enhanced viola-jones vehicle detection method from unmanned aerial vehicles imagery. *IEEE Transactions on Intelligent Transportation Systems*, 18(7):1845–1856, 2017.

[15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[16] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint*, 2017.

[17] Tianyu Tang, Zhipeng Deng, Shilin Zhou, Lin Lei, and Huanxin Zou. Fast vehicle detection in uav images. In *Remote Sensing with Intelligent Processing (RSIP), 2017 International Workshop on*, pages 1–5. IEEE, 2017.

[18] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016.

[19] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. *arXiv preprint arXiv:1708.05237*, 2017.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[22] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Single-shot refinement neural network for object detection. *arXiv preprint arXiv:1711.06897*, 2017.

[23] Aniruddha Kembhavi, David Harwood, and Larry S Davis. Vehicle detection using partial least squares. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1250–1265, 2011.

[24] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[25] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1522–1530. IEEE, 2017.

[26] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[27] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pages 391–405. Springer, 2014.

[28] Matthew D. Zeiler and Rob Fergus. *Visualizing and Understanding Convolutional Networks*. Springer International Publishing, 2014.