2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Performance Comparison of Small Object Detection Algorithms of UAV based Aerial Images

Hao Xu[1], Yuan Cao[1], Qian Lu[2], Qiang Yang[1]
[1]College of Electrical Engineering, Zhejiang University, Hangzhou, China 310027
[2]State Grid Jiangsu Electric Power CO., Ltd. Research Institute, Nanjing, China 211103
Email: qyang@zju.edu.cn

***Abstract*— Traffic controls in modern society are part of urban management. With the assistance of unmanned aerial vehicles (UAVs) equipped with mounted cameras, researchers could capture aerial (bird-view) images from appropriate altitude. The perspective in aerial images makes appearances of objects squat, although aerial images can supply more contextual information about the environment by a broader view angle, the object instances may be detected by mistake. This fact diminishes the aerial images that can be fed to a network with higher dimensions that increases the computational cost to prevent the diminishing of pixels belonging to small objects. To compare model performance on small objects with aerial images, this study trains and tests two object detectors, i.e. YOLOv4 and YOLOv3, on the AU-AIR dataset, and exploited the parameterization of YOLO based models for small object detection. Finally, the key numerical results and observations are presented.**

Keywords: AU-AIR dataset; small object detection; YOLO;

I. INTRODUCTION

Unmanned aerial vehicles (UAVs) are extensively used for management and control in different areas of cities, such as traffic surveillance [1] and urban environmental management. UAVs collect the visual data of the surrounding environment through the mounted camera and adopt the computer vision-based object detection for different purposes. Urban traffic surveillance has the characteristics of high space complexity and congestion lag. Considering real-time applicability and detection accuracy, this study finally selects the YOLO object detector, which is widely used in the industry. This study researches small object detection from the aerial view with the support of the AU-AIR dataset [4].

To the authors' best knowledge, the AU-AIR dataset is the first multi-modal UAV dataset for object detection. Compared with general deep learning datasets (such as MSCOCO and PASCAL VOC), most of the images in traditional datasets are taken at a nearly horizontal angle with a handheld camera, and most images only have a side-view. In the natural image, there are challenges of object detection such as occlusion, illumination changes, rotation, low resolution, and crowd existence of instances, which may cause detection mistakes. But in the aerial image, the possibility of the object being blocked is greatly reduced. However, there are some drawbacks to the aerial photography process of UAVs. Due to perspectives, the height information of the object may be abandoned, which makes appearances of object squat, different from the general dataset in object feature extraction.

This study selects the latest version of the YOLOv4 model [2] as the research object. As an extended version of YOLOv3 [3], YOLOv4 serves as an efficient and powerful object detection model. YOLOv4 is a regression algorithm based on deep learning and reduces the hardware cost of training a fast and accurate object detector. Taking into account the real-time performance and detection rate, the YOLOv4 algorithm is verified that the most advanced object detection methods are used in the detector training process, such as Bag-of-Freebies and Bag-of-Specials [2]. YOLOv4 algorithm adapts the most advanced methods for the YOLO model, including CBN, PAN, SAM, etc.

In the context of aerial photography traffic control, this work uses the YOLO object detection model to propose a visual detection method. The main technical contributions of this work can be summarized as follows: (1) the performance in terms of detection speed and bounding box positioning accuracy of the YOLOv4 and YOLOv3 algorithms are compared under the condition of small object detection; and (2) the model for small vehicle detection in the aerial images of UAVs is provided and assessed as a case study.

The remaining of the paper is organized as follows: Section II briefly overviews the YOLOv4 and YOLOv3 algorithms; Section III explains the image augmentation and data modification of the AU-AIR dataset; Followed by Section IV carrying the performance comparison and analysis through numerical experiments; finally, the conclusive remarks are given in Section V.

II. MODEL COMPARISON

The network structure diagram of YOLOv3 [3], as shown in Fig. 1, consists of three basic components: (1) CBL: The smallest component in the YOLOv3 network structure, which consists of three activation functions: Conv + BN + Leaky_ReLU; (2) Res unit: Use the residual structure in the Resnet to make the deep learning network deeper; and (3) ResX: It is composed of a piece of CBL and numbers of residual components, which is the large component in YOLOv3. Component CBL in front of each Res module plays the role of downsampling (image reduction), so after 5 layers of the Res module, the size of the obtained feature map respectively decreases to 1, 1/2, 1/4, 1/8, 1/16, 1/32 of the original feature map.

After being proposed in 2018, YOLOv3 has become a very classic algorithm in one-stage object detection, including the Darknet-53 network structure, anchor frame, FPN, and other excellent structures. The integral structure of YOLOv4 is similar to YOLOv3, but the structure CSP and PAN are added, with sub-structures improved, which may impact detection precision and prediction accuracy [2].

The network structure diagram of YOLOv4, as shown in Fig. 2, mainly composed of the following components. (1) CBM: The smallest component in the YOLOv4 network structure, which is composed of Conv + BN + Mish activation functions. (2) CSPX: Refers to the CSPNet

2473-3636/20/$31.00 ©2020 IEEE
DOI 10.1109/DCABES50732.2020.00014

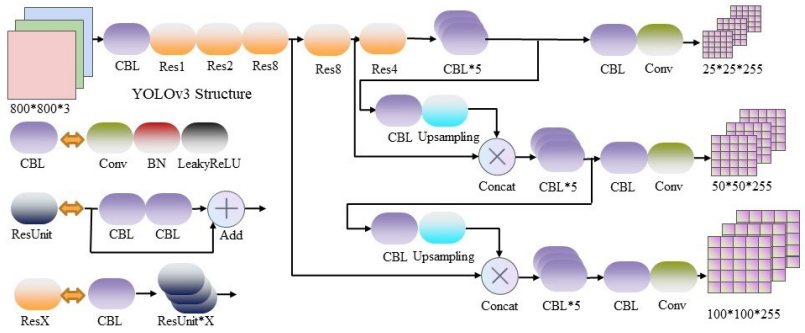


Figure 1. Network Structure of YOLOv3

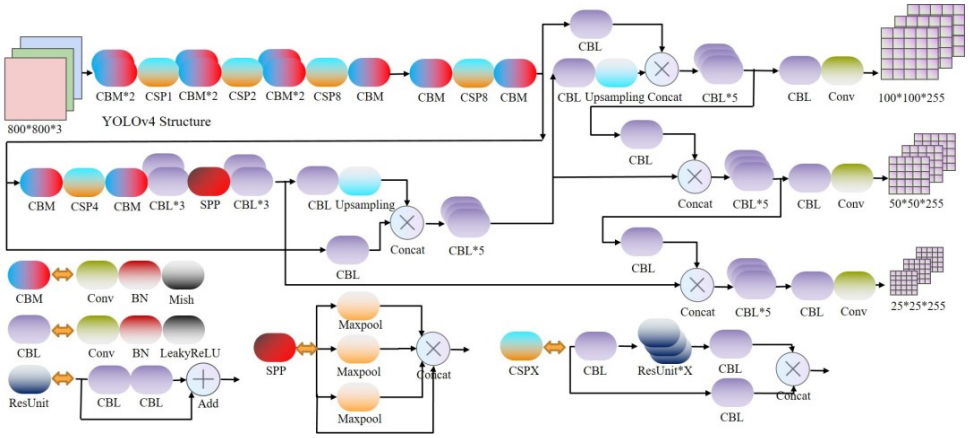


Figure 2. Network Structure of YOLOv4

TABLE 1. QUANTITATIVE ANALYSIS OF ANNOTATIONS IN AU-AIR DATASET

Classes	Instance	Max-Height	Min-Height	Mid-Height (Percentage)	Avg-Height (Percentage)	Max-Width	Min-Width	Mid-Width (Percentage)	Avg-Width (Percentage)
Human	5152	929	3	**68/6.30%**	**74.16/6.87%**	1856	3	**63/5.83%**	**73.70/3.84%**
Car	102581	1071	3	**61/5.65%**	**75.65/7.00%**	1920	3	**90/8.33%**	**116.45/6.07%**
Van	9540	1074	3	**131/12.13%**	**154.23/14.28%**	1875	3	**177/16.39%**	**226.34/11.79%**
Truck	9992	776	3	**92/8.52%**	**107.09/9.92%**	1157	3	**131.5/12.18%**	**152.86/7.96%**
Bike	318	465	3	**58.5/5.42%**	**70.40/6.52%**	527	3	**70/6.48%**	**85.31/4.44%**
Motorbike	1128	680	3	**60/5.56%**	**68.47/6.34%**	873	3	**68/6.30%**	**80.15/4.17%**
Bus	729	774	3	**141/13.06%**	**144.47/13.38%**	1293	11	**241/22.31%**	**265.80/13.84%**
Trailer	2537	874	3	**103/9.54%**	**130.85/12.12%**	1567	3	**141/13.06%**	**201.54/10.50%**

network structure, which is composed of three convolutional layers and X Res unit modules concatenated. (3) SPP: Adopts the maximum pooling method of 1×1, 5×5, 9×9, 13×13 to perform multi-scale integration.

YOLOv3 model uses a variant of Darknet and initially trained a 53-layer network. For object detection, 53 layers are stacked on top to provide YOLOv3 with a 106-layer fully convolutional bottom layer architecture, while the YOLOv4 model network has a total of 161 layers. With the resolution of 800 * 800 used in this article, the total amount of calculation is 280.703 BFLOPS, and YOLOv3 is 241.699 BFLOPS. YOLOv4 algorithm uses more neural networks with Mish, Leaky_ReLU, and other non-linear units, which is mathematically equivalent to a piecewise linear function. It can be observed that more linear regions can lead to stronger nonlinearity of the neural network, and hence the better detection performance in practice. Therefore, when the total number of neurons is equivalent, increasing the network depth can cause the network to produce more linear regions [8].

III. IMAGE PREPROCESSING

Reading and analyzing the distribution of annotations in the AU-AIR dataset can be obtained as Table I. Under the original pixels of 1920 * 1080, the average percentage and median percentage of the pixels identified by the dataset are generally below 15%, which means most of the detection objects belong to the small object category.

YOLOv4 retains heads of YOLOv3, but changes the backbone network to CSPDarknet53, adopts the idea of SPP (Spatial Pyramid Pooling) to expand the receptive field, and PANet as the neck. To improve the mAP and positioning accuracy of small object detection on the AU-AIR dataset, this study uses some tricks during training:

- Improve grid subdivision (resolution) from 416 * 416 to 800 * 800 by conventional datasets such as

MSCOCO, and appropriately increase the amount of training calculation and training time to better adapt to small object detection.

- Use the anchor mechanism in Faster-RCNN. To improve accuracy and positioning accuracy, the shape and scale of the object in the AU-AIR dataset are calculated offline through the k-means algorithm.
- Use the new image augmentation technology in YOLOv4, such as Mixup, Cutmix, Mosaic and Blur. Enable image adjustment parameters such as angle, saturation, exposure, and hue to enhance the data. At the same time, the activations, i.e. Swish, Mish, Norm_Chan, Norm_Chan_Softmax, are added.

IV. Experimental Process

During the evaluation, this study considers achieving a lower detection error rate and higher positioning accuracy, while also considers real-time performance. Therefore, YOLOv4, YOLOv3 and YOLOv3-tiny models as selected for the comparison study in this work.

A. Experiment Parameter Settings

The experimental environment is Ubuntu 19.10, the graphics cards are 4 pieces of NVIDIA TITAN XP PASCAL, and the CUDA version is 10.1.

Authors adjust the image resolution from 416*416 to 800*800, enable optimized memory allocation during network resizing, adjust image augmentation parameters, and appropriately reduce the learning rate to prevent overfitting. Taking 90% of the dataset as the training set and 10% as the test set, the object detectors are adapted to the total number of classes in the AU-AIR dataset (8 classes in total) by changing their YOLO layers, convolutional filters, and upsample strides.

B. Comparison of Indicators

For benchmarking, authors train the YOLOv4 and YOLOv3 model with the AU-AIR Dataset. Authors use the following parameters: the batch size of 64 with subdivisions=64 (mini_batch=1), set Adam optimizer with the initial default parameters (learning rate=0.001, beta1=0.8, beta2=0.9), and enable image augmentation. The training is stopped when the validation error starts to increase. The training process after curve fitting is shown in Fig. 3, and the compare results are shown in Table 2.

C. Experimental Results

This study noticed that the AP of motorbike and bike are significantly smaller than other categories, which may be caused by the imbalanced distribution of categories. These two categories have fewer instances and smaller sizes. On the other hand, although the target size of humans is smaller than others too, the training effect is close to the mAP due to the sufficient number of samples.

The qualitative analysis of the training results of YOLOv4 and YOLOv3 samples is shown in Fig. 4. After 300,000 iterations, detection mAP and Loss tend to be stable. Through curve fitting of data scatter, the mAP of YOLOv4 is 67.35%, which is nearly 7% higher than that of YOLOv3. The average detection rate of YOLOv4 is 24 FPS, and YOLOv3 is 29 FPS. Although YOLOv4 has higher network complexity, to the authors' knowledge, activations have a stronger correlation to runtime on hardware accelerators than flops [9].

Considering the perspective of positioning accuracy and detection rate, YOLOv4 model is slightly higher than YOLOv3, and much higher than YOLOv3 Tiny in detection rate, . Meanwhile, because YOLOv4 uses CIoU as the object detection regression loss function, compared with the IoU loss function used by YOLOv3. YOLOv4 adds the determination of the bounding box and object frame intersection method, the judgment of overlapping area and distance of a center point, and the comparison of the aspect ratio of the bounding box and fitting object frame. Therefore, in the prediction of the same detection object, the positioning of YOLOv4 is more accurate, and the size and position of the bounding box are closer to the fitting object frame. As shown in Fig. 4, in some cases, YOLOv4 has better performance in bounding box position than YOLOv3 but also exists object omission detection. The authors also found that YOLOv3 was slightly better than YOLOv4 in the detection of incomplete samples, possibly due to underfitting of YOLOv4.

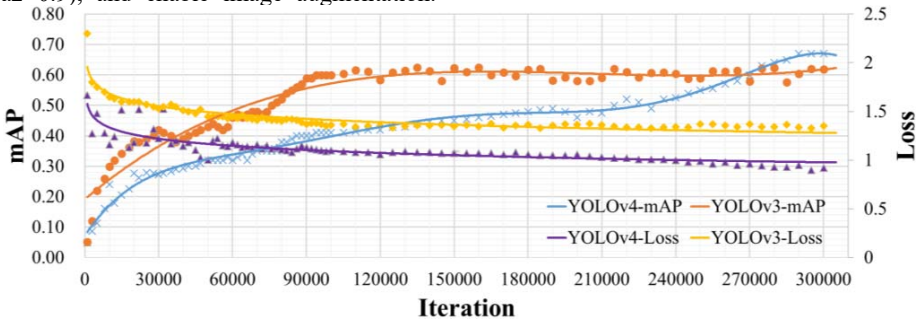
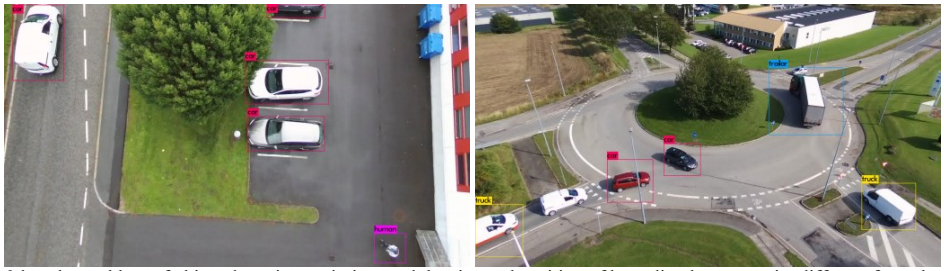


Figure 3. Loss and Mean Average Precision during Training

TABLE 2. Comparison of Three Models for Mean Average Precision After Iterations

Model	Iteration	Human	Car	Van	Truck	Bike	Motorbike	Bus	Trailer	mAP	Precision	Recall
YOLOv4	300000	56.36%	70.51%	79.04%	82.58%	64.13%	61.91%	62.62%	61.62%	67.35%	0.81	0.54
YOLOv3	300000	44.94%	52.44%	74.63%	75.33%	53.45%	52.53%	73.18%	52.13%	59.83%	0.69	0.45
YOLOv3-Tiny [4]	N/A	34.05%	36.30%	41.47%	47.13%	12.34%	4.80%	51.78%	13.95%	30.22%	N/A	N/A



(a) YOLOv3 has the problem of object detection omission, and the size and position of bounding box are quite different from the actual one.

(b) YOLOv4 also has the problem of object detection omission on some frame figure, maybe caused by inadequate iteration.

Figure 4. Small object detection performance comparison: (a) YOLOv3, (b)YOLOv4

V. Conclusion And Discussion

In summary, this study proposes an experimental verification based on the AU-AIR dataset to compare and test the algorithmic performance on small object detection of YOLOv4 and YOLOv3. YOLOv4 adapts the most advanced method for the YOLO model, enables activation functions such as Mish, and creatively uses CSPDarknet53 as the backbone network. It has better performance on common datasets, such as MSCOCO and PASCAL VOC datasets. However, in mainly composed of small object dataset, the mAP of YOLOv4 model is better than that of YOLOv3, there are still some problems such as detection omission, positioning and size deviation of bounding box. Our analysis is mainly due to the following reasons:

- The YOLO algorithm zooms out the original image. Because of the receptive field, reduction makes it difficult to detect objects with relatively small sizes. Compared with YOLOv3, YOLOv4 has deeper network layers and the shallow features, which is particularly significant for small object detection, some of shallow features are abnegated;
- YOLOv4 algorithm adds some tricks like Bag of Specials, increasing the SPP of the receptive field and the activation function Mish, etc. Bag of Specials reduces the calculation speed of YOLOv4 algorithm to some extent, but plays a positive role in increasing the accuracy of small object detection;
- It is also a common problem of the YOLO algorithm. Classification and regression operations are performed on the feature layer after several downsampling layers. The receptive field of the small object feature that is mapped back to the original image may be larger than the size of the small object in the original image, resulting in poor small object detection effect.

The shortcomings of this experiment are: Firstly, considering the accuracy requirements of small object detection algorithms in complex environments, YOLOv4 may not be the most suitable algorithm for small object detection accuracy and speed. You could consider other algorithms, which are specific for small object detection, such as DetNet [5], Cascade R-CNN [6], SNIP [7], and FPN [10]. Secondly, the above steps and methods are only used for the detection of the small object on vehicles and pedestrians, which may not apply to other object detection. Thirdly, due to the limitation of the graphics card's calculation efficiency and RAM, relevant experiments need to be further verified, and experimental results may be biased. This also further shows that in the field of small object detection such as pedestrians and vehicles based on aerial images of UAVs, further research is still needed.

References

[1] Puri, A., A survey of unmanned aerial vehicles for traffic surveillance. Department of Computer Science and Engineering, University of South Florida, Tech. Rep, 2004.

[2] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934, 2020.

[3] Redmon, J., & Farhadi, A. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[4] Bozcan, I., & Kayacan, E. AU-AIR: A Multi-modal Unmanned Aerial Vehicle Dataset for Low Altitude Traffic Surveillance. arXiv preprint arXiv:2001.11737, 2020.

[5] Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y., & Sun, J. Detnet: A backbone network for object detection. arXiv preprint arXiv:1804.06215, 2018.

[6] Cai, Z., & Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6154-6162, 2018.

[7] Singh, B., & Davis, L. S. An analysis of scale invariance in object detection snip. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3578-3587, 2018.

[8] Montufar, G. F., Pascanu, R., Cho, K., & Bengio, Y. On the number of linear regions of deep neural networks. Advances in neural information processing systems, pp. 2924-2932, 2014.

[9] Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., & Dollár, P.Designing network design spaces. arXiv preprint arXiv:2003.13678, 2020.

[10] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117-2125, 2017.