MULTI-SCALE OBJECT DETECTION IN SATELLITE IMAGERY BASED ON YOLT

Wentong Li[1,2,4], *Wanyi Li*[3], *Feng Yang*[1,2,4*], *Peng Wang*[3]

[1]Northwestern Polytechnical University, Xi'an, China
[2]Key Laboratory of Information Fusion Technology, Ministry of Education, Xi'an, China
[3]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[4]CETC Key Laboratory of Data Link Technology, No. 20 Institute of CETC, Xi'an, China

ABSTRACT

Multi-scale object detection (MOD) is one of the remaining challenges for satellite imagery. To improve the performance of MOD task, YOLT (You Only Look Twice) has achieved a good accuracy in high resolution remote sensing images. Motivated by the state-of-art object detection method for satellite imagery, we explored and achieved the state-of-the-art accuracy based on the standard YOLT for MOD task by providing a novel method with enough experimental results and model comparison on the typical multi-scale satellite imagery dataset. First, we divide objects into three categories according to the scale of objects. Then, different training strategies are used to train the classifier and detector for different scale objects. Finally, multi-scale detection chips are stitched and fused to get more accurate localization and classification as the final predicted results for MOD in satellite imagery. Experiments have been conducted over dataset from the second stage of AIIA[1] Cup Competition of Typical Object Recognition for Satellite Imagery in Small Samples compared with the standard YOLT and Faster R-CNN, which demonstrates the effectiveness and the comparable detection performance of our proposed pipeline.

Index Terms— Multi-scale object detection, Satellite Imagery, YOLT

1. INTRODUCTION

Remarkable progresses have been made in object detection of satellite imagery recently due to the convolutional neural networks (CNNs) [1]. However, MOD is one of the remaining challenge tasks [2]. As shown in Fig.1, small and densely clustered object detection (including boat, oilcan, airplane) and large and dispersed object detection (including harbor, airport) in satellite imagery are two typical MOD problems in real world, which contain different scales of objects and the scale ratio is high. One of most impressive and successful frameworks for remote sensing object detection is YOLT [3,4]. The standard YOLT has achieved a good accuracy in high resolution sensing images for MOD tasks.

Fig.1: Examples of multi-scale object detection. Small and densely clustered object detection (including boat, oilcan, airplane) and large and dispersed object detection (including harbor, airport) in satellite imagery.

In this paper, we aim to explore YOLT framework to improve the performance of MOD task in satellite imagery. In principle, YOLT implements classification and regression by an end-to-end pipeline, which is inspired by YOLO (You Only Look Once) [5,6,7] owing to the speed, accuracy, and flexibility. Specifically, YOLT implements a unique network architecture with a denser final prediction grid to help differentiate between classes by yielding finer grained features. However, there are two main problems for MOD tasks with YOLT. First, in YOLT method, the single network architecture is applied to capture feature maps for different scales of objects, which is more effective to one kind of scale of objects knowledge but insufficient to capture all kinds of scale of objects in geospatial imagery.

[1]AIIA is China Artificial Intelligence Industry Development Alliance.

*Corresponding to yangfeng@nwpu.edu.cn

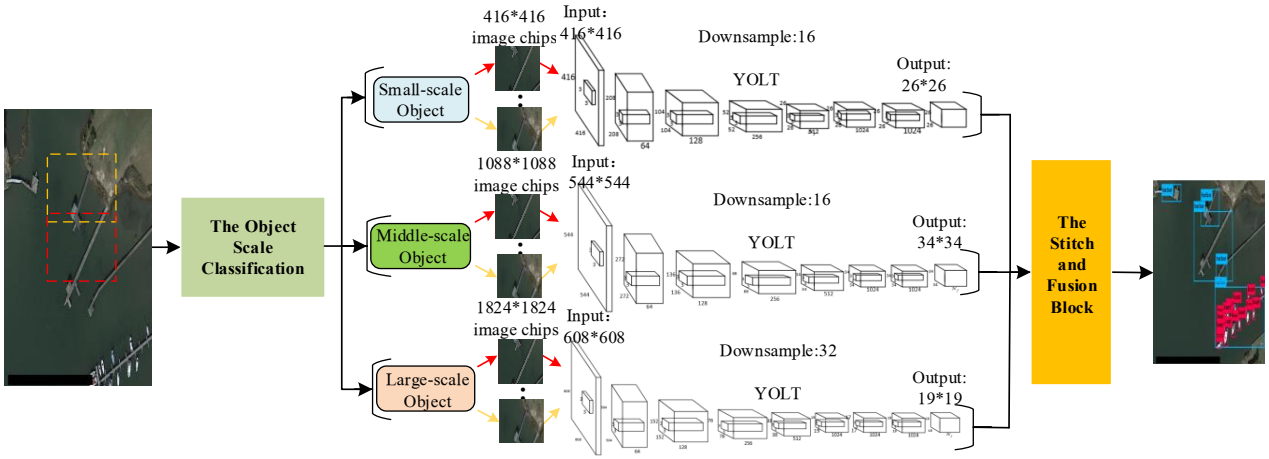

Fig.2: The overview framework of our MOD-YOLT method for satellite imagery.

Therefore, using the single network model, it is difficult to get the precise location of all objects. Second, the image size can vary from 1500×1500 to 20000×20000 in remote sensing images. YOLT partitions images of arbitrary size into manageable cutouts. But the size of pixel cutouts is fixed which is infeasible for different scales of objects. If the size of pixel cutouts is large, it is hard to capture fine-grained spatial details for small object. If the size of pixel cutouts is small, it is easy to destroy the large object, which makes the feature extracted incomplete for large and dispersed objects.

To tackle the problems discussed above, we propose a novel MOD method (MOD-YOLT) based on the standard YOLT framework by introducing a scale classification criteria and different training strategies for different scales of objects. Firstly, to get the fine-grained knowledge for different scales of objects, we divide objects into three categories according to our classification criteria. TridentNet [8] has exploited and demonstrated that the way of three branches/categories is a better setting. Then, different training strategies including the size of pixel cutouts, the architecture of network model are used to train the classifier and detector for different scale objects. Finally, multi-scale detection chips are stitched and fused to get more accurate localization and classification as the final predicted results for MOD in satellite imagery. To provide meaningful performance evaluation, our team takes part in a competition named AIIA Cup Competition of Typical Object Recognition for Satellite Imagery in Small Samples (AIIA) and gets a good mark. Meanwhile, the performance evaluation is also conducted to compare with the standard YOLT and Faster R-CNN [9] in AIIA2018_2nd dataset.

2. A NOVEL MOD METHOD BASED ON THE STANDARD YOLT FRAMEWORK

In this section, the pipeline of our proposed MOD-YOLT framework is presented and described in detail.

2.1. The pipeline of MOD-YOLT

As introduced above, our proposed MOD-YOLT method is based on the standard YOLT framework [3,4]. From Fig.2, our MOD-YOLT method mainly consists of 4 parts: object scale classification module, image splitting module, Multi-YOLT Network (MYN), and stitch and fusion module. The object scale classification module, the image splitting module and the MYN module are designed to get better feature representation for different scales objects of satellite imagery. The final stitch and fusion module are proposed to get more accurate localization and classification for MOD in satellite imagery.

2.2. MOD-YOLT method in details

The details about our MOD-YOLT method will be introduced in this section.

First, considering the complexity and performance, we make a reasonable classification for input images with annotations by the ratio of the object to the whole image (ROWI) to utilize the scale information present in satellite imagery according to Table 1. The category of objects includes large-scale objects, middle-scale objects and small-scale objects.

Second, it's a practical way to partition images of arbitrary size into manageable cutouts to feed into network models by a sliding window going from left to right. For small, densely packed objects, we opt to partition images of arbitrary size into 416×416 cutouts and 15% overlap. For middle-scale objects, we design to partition images of arbitrary size into 1088×1088 cutouts and 25% overlap. For

high density scenes, we attempt to partition images of arbitrary size into 1824×1824 cutouts and 35% overlap.

Third, MYN is designed to address the issue that a single model, YOLT, is not enough to recognize all three scales above. For small, densely packed objects, we opt to implement a network architecture that downsamples by a factor of 16 with a 416×416 input image pixel yielding a 26×26 prediction grid. For middle-scale objects, we design to implement a network architecture that downsamples by a factor of 16 with a 544×544 input image pixel yielding a 34×34 prediction grid. For high density scenes, we attempt to implement a network architecture that downsamples by a factor of 32 with a 608×608 input image pixel yielding a 19×19 prediction grid. More and better feature representation for different scales objects of satellite imagery will be got by the MYN.

The final step in the object detection pipeline seeks to stitch and fuse together the dozens or hundreds of chips' results into one final image strip by computing the scale ratio of each predicted object to the whole image, then select the best results (including predicted class, confidence, coordinates) whose confidence value is best as the final predicted results for the same predicted object under each scale model of MYN. Among which, the overlap (15%,25% or 35%) ensures all regions will be analyzed, then we apply non-maximum suppression (NMS) to the global matrix of bounding box predictions to alleviate overlapping detections. The detailed criteria of the object scale classification, image splitting and input size of different scales model of MYN is shown in Table 1.

Table 1: The detailed criteria of the object scale classification, image splitting and input size of different scales model.

	Scale	Cutouts	Overlap	Input size
ROWI≤1/150	Small-scale Object	416×416	15%	416×416
1/150<ROWI≤1/20	Middle-scale Object	1088×1088	25%	544×544
1/20<ROWI	Large-scale Object	1824×1824	35%	608×608

3. EXPERIMENTS

In order to demonstrate the effectiveness of our proposed MOD-YOLT method, we conduct experiments on the typical multi-scale dataset of satellite remote sensing images, AIIA2018_2nd. Mean Average Precision is used as the evaluation metric followed by the standard PASCAL VOC criteria, i.e. IoU > 0.5 between ground truths and predicted boxes [10]. The proposed MOD-YOLT network is trained with Stochastic Gradient Descent (SGD), where momentum is 0.9, and weight decay is 0.0005, on a single NVIDIA GeForce GTX 1080Ti GPU with 12GB memory. The learning rate is set of 0.0001 for 40k mini-batches, 0.00001 for the next 20k mini-batches, 0.000001 for the next 20k mini-batches and 0.0000001 for the final 20k mini-batches.

3.1. Dataset description

The AIIA2018_2nd dataset of satellite remote sensing images is provided from the second stage of AIIA, which covers six classes: airport, airplane, harbor, boat, oilcan, bridge. The dataset includes 2421 images whose size varies from 512×512 pixels to 5120 × 3584, mainly 90% from 512×512 pixels to 2800×2800. 450 images are randomly selected as the test data and the remaining ones are used for training. Evaluating the images in AIIA2018_2nd, it can be seen that the size of remote sensing object varies greatly from 5×4 to 2441×938, which is a great challenge for object detection. Some examples of AIIA2018_2nd are given in Fig.1.

3.2. Experimental results

Table 2: Detection results of Faster R-CNN, YOLT and MOD-YOLT on TP, FP, FN, Precision, Recall, F1-score and mAP.

Method	TP	FP	FN	Precision	Recall	F1-score	mAP
Faster R-CNN	1855	429	2672	0.8122	0.4098	0.54471	0.62
YOLT	1589	**113**	1182	**0.9336**	0.5734	0.71048	0.76
MOD-YOLT	**2077**	502	**944**	0.8054	**0.6875**	**0.74179**	**0.78**

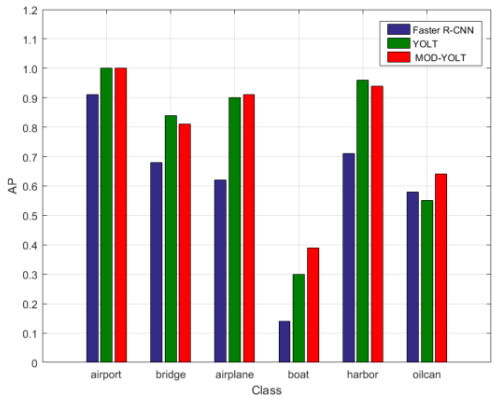

Fig.3: Detection results of Faster R-CNN, YOLT and MOD-YOLT on AP of each class.

In our experiments, we trained the-state-of-the-art algorithms, like Faster R-CNN and standard YOLT with hyper parameter architecture for the purpose of comparison. Experimental results are shown in on the Fig 3 and Table 2. From Table 2, we can see that the number of true positives is largest, the number of false negatives is least, which is 2077, 944 with MOD-YOLT method respectively. According to the number of TP, FP and FN, we can get the results of Precision, Recall, F1-core. As shown in Table 2, we can see

that though the Precision has dropped slightly, the Recall has risen sharply in MOD-YOLT method. Our MOD-YOLT method achieves F1-score 0.74179, which is 0.197085 and 0.03131 higher than that of Faster R-CNN(ResNet-101) and YOLT, respectively. At the same time, the Average Precision (AP) of each class is shown with Fig 3. For small-scale objects including boat, airplane and oilcan, the AP of MOD-YOLT method has the large relative performance advantage. For middle-scale objects including bridge, the AP of MOD-YOLT method has dropped slightly. For large-scale objects including airport and harbor, the AP of MOD-YOLT method has near performance. According to Fig 3, we can get the results of mAP in Table2. Our MOD-YOLT method achieves mAP 78%, which is 16% and 2% higher than that of Faster R-CNN(ResNet-101) and YOLT, respectively. Fig 4 shows visual detection results of ground truth, Faster R-CNN, YOLT and MOD-YOLT on validation dataset. These experimental results validate the effectiveness of our proposed MOD-YOLT method for satellite imagery.

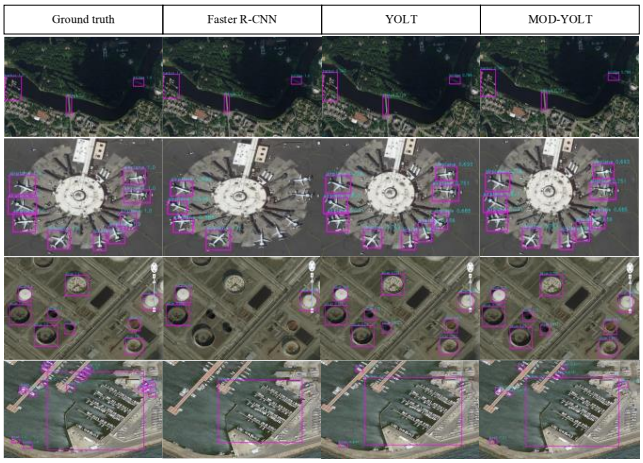

Fig.4: Visual detection results on validation dataset. Columns from left to right are ground truth, Faster R-CNN, YOLT and MOD-YOLT.

4. CONCLUSION

Aiming to improve the localization performance of multi-scale objects, this paper presents an effective MOD method in satellite imagery based on the standard YOLT. The experimental results on AIIA2018_2nd images demonstrate the effectiveness of the proposed method. Our proposed MOD-YOLT pipeline exhibits strong competency in handling multi-scale object detection tasks, achieving 16% and 2% higher than Faster R-CNN(ResNet-101) and YOLT on mAP, respectively. For future work, we will focus on the further tasks of multi-scale objects with YOLOv3 [7].

5. ACKNOWLEDGMENT

The work was supported by National Natural Science Foundation of China (No. 61771471, No. 61374159), the Foundation of CETC Key Laboratory of Data Link Technology (No. 20182316, No. 20182203), Natural Science Foundation of Shaanxi province (No. 2018MJ6048), and the Seed Foundation of Innovation and Creation for Graduate Students in Northwestern Polytechnical University (No. ZZ2019178).

6. REFERENCES

[1] Yuan Yao, Zhiguo Jiang, Haopeng Zhang, Bowen Cai, Gang Meng, and Deshan Zuo, “Chimney and condensing tower detection based on faster r-cnn in high resolution remote sensing images,” in 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). pp.3329-3332, 2017.

[2] Qingle Guo, Junping Zhang, Tong Li, and Xiaochen Lu. “Change detection for high-resolution remote sensing imagery based on multi-scale segmentation and fusion,” in 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). pp.1919–1922, 2017.

[3] Adam Van Etten, “You only look twice: Rapid multi-scale object detection in satellite imagery,” in ArXiv Preprint, ArXiv:1805.09512, 2018.

[4] Adam Van Etten, “Satellite imagery multiscale rapid detection with windowed networks,” in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 735–743, 2019

[5] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. “You only look once: Unified, real-time object detection,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788, 2016.

[6] Joseph Redmon, and Ali Farhadi. “Yolo9000: Better, faster, stronger,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6517–6525, 2017.

[7] Joseph Redmon, and Ali Farhadi. “Yolov3: An incremental improvement,” in ArXiv Preprint, ArXiv:1804.02767, 2018.

[8] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. “Scale-aware trident networks for object detection,” in ArXiv Preprint. ArXiv:1901.01892, 2019.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster r-cnn: Towards real-time object detection with region proposal networks,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, 39 (6):1137–1149, 2017.

[10] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. “The pascal visual object classes challenge: A retrospective,” in International Journal of Computer Vision, 111 (1): 98–136, 2015.