Optimizing Results in Aerial Images through Post-Processing Techniques on YOLOv7

Nguyen D. Vo
nguyenvd@uit.edu.vn
Multimedia Communications Laboratory, University of Information Technology
Vietnam National University, Ho Chi Minh city
Ho Chi Minh city, Vietnam

Thu M. Nguyen
21520472@gm.uit.edu.vn
Faculty of Computer Science, University of Information Technology
Vietnam National University, Ho Chi Minh city
Ho Chi Minh city, Vietnam

Triet H. M. Nguyen
21520497@gm.uit.edu.vn
Faculty of Computer Science, University of Information Technology
Vietnam National University, Ho Chi Minh city
Ho Chi Minh city, Vietnam

Khang Nguyen
khangnttm@uit.edu.vn
Faculty of Software Engineering, University of Information Technology
Vietnam National University, Ho Chi Minh city
Ho Chi Minh city, Vietnam

ABSTRACT

Object detection in aerial images has garnered significant attention from the research community in recent years. The challenges posed by small objects, diverse orientations, and complex backgrounds have spurred extensive research efforts. In this paper, we focus on object detection in a YOLOv7-based framework, highlighting its limitations and proposing a post-processing method to enhance object detection results in cases of overlapping predicted regions. The analyses are extended and demonstrated to be effective on the UCAS AOD dataset.

CCS CONCEPTS

• **Computer Vision → Object Detection**.

KEYWORDS

object detection in aerial images, ucas aod, yolov7

ACM Reference Format:
Nguyen D. Vo, Thu M. Nguyen, Triet H. M. Nguyen, and Khang Nguyen. 2022. Optimizing Results in Aerial Images through Post-Processing Techniques on YOLOv7 . In *The 12th International Symposium on Information and Communication Technology (SoICT 2023), December 7–8, 2023, Ho Chi Minh, Vietnam.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Aerial images are pictures taken from above using unmanned aerial devices such as drones, flycams, and so on. In recent years, aerial imagery has been utilized to assist humans in various fields such as surveillance, security, or border patrolling [1]. Therefore, the demand for developing systems based on inputs from aerial photographs is increasing. The task of object detection in aerial images is also a topic of great interest within the community.

However, object detection in aerial images still faces many limitations. Images captured from above pose more challenges compared to natural images in the past, such as variations in viewing angles, small sizes of objects, diversity in orientations, and differing scales. Despite numerous published studies, the results remain limited due to the specific challenges posed by aerial imagery. The object detection task can be divided into two main methodological groups: single-stage methods and two-stage methods. Two-stage methods have an advantage in terms of detection performance, while single-stage methods excel in execution time.

In single-stage methods, YOLO (You Only Look Once) [2] is often used for object detection due to its advantages such as high speed and accuracy. Among them, YOLOv7 [3], recently introduced at the CVPR 2023 conference, stands out as a method that achieves competitive results compared to many two-stage methods. However, when applying this method to aerial images, there are still some existing issues such as bounding box overlap and object omissions. Thus, in this paper, we propose a post-processing technique verified on the UCAS AOD dataset [4]to optimize the results.

The rest of the paper is organized as follows. In Section 2, we present a survey of previous work in object detection in aerial images. We propose a formula for selecting bounding boxes with high confidence, and the post-processing system is briefly described in Section 3. The experiments and results are reported and discussed in Section 4. Finally, we conclude the paper in Section 5.

2 RELATED WORKS

2.1 Object Detection Overview

Object detection is a field in computer vision that focuses on identifying, localizing, and classifying objects in digital images. It combines two crucial tasks: classification, which predicts the class of the object, and object localization, which determines the object's position through a bounding box.

SoICT'23, December 07–08, 2023, Hochiminh, Vietnam

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

In general, current approaches to object detection are divided into two main groups: single-stage and two-stage methods. Two-stage methods use a Region Proposal Network (RPN) [5] to extract regions likely to contain objects before performing classification and localization, achieving high accuracy but requiring more resources for training and prediction. A representative model of this type is Faster R-CNN [5] (In the first phase, RPN is used instead of the Selective Search algorithm to generate Region Proposals. In the second phase, the model performs object detection similar to Fast R-CNN, but using the proposals generated by RPN).

One-stage methods do not use region extraction for object localization and treat it as a regression problem. They tend to be faster but have lower accuracy compared to two-stage methods. Representative models in this category include SSD [6], RetinaNet [7], RepPoints [8], and YOLO (which uses a single neural network to directly predict bounding boxes and class probabilities for the entire image in one evaluation), and are often used for object detection.

2.2 Object Detection in Aerial Images Overview

The task of object detection in aerial images still faces many challenges, especially due to the difficulties in collecting images from above, which present a range of issues not previously encountered when capturing natural ground-level images. These challenges include adjusting the viewing angle appropriately, accurately handling small-sized objects, as well as dealing with their diverse orientations and scales [1, 9]. This was highlighted in the VisDrone competition [10], where participating teams utilized renowned models such as RCNN, Fast R-CNN, Faster R-CNN (with a one-stage approach), and YOLO models with a two-stage approach. The choice of YOLOv7 for application in aerial image analysis, particularly with challenging datasets like UCAS-AOD, may demonstrate the adaptability of YOLOv7.

2.3 Non Max Suppression

Non-Maximum Suppression (NMS) is a crucial component in computer vision and image processing applications. It is employed to reduce redundant or unnecessary predictions after a model generates initial predictions. The NMS process consists of three fundamental steps. Firstly, predictions are sorted based on their probabilities or scores. Next, predictions with scores below a threshold are removed from the list. Finally, bounding boxes that overlap beyond a permissible threshold (measured by Intersection over Union - IoU) lead to the elimination of extraneous predictions. The ultimate outcome of NMS is a list of unique and accurate bounding boxes representing objects in the image. NMS plays a pivotal role in various scenarios, from object detection to image segmentation, ensuring that only the most pertinent predictions are retained.

3 METHODOLOGY

3.1 YOLOv7 method

YOLOv7 was introduced at the CVPR 2023 conference by the authors of YOLOv4 and YOLOR. At that time, YOLOv7 outperformed other object detectors in both speed and accuracy, achieving a range of 5 to 160 frames per second (*FPS*). Notably, it achieved the highest accuracy of 56.8% Average Precision (*AP*) among all real-time object detectors running on a V100 GPU at speeds of 30 *FPS* or higher. In particular, the YOLOv7-E6 detector (56 *FPS* on V100, 55.9% *AP*) surpassed both the SWIN-L Cascade-Mask R-CNN (9.2 *FPS* on A100, 53.9% *AP*) based on transformers, with a speed improvement of over 509% and a higher accuracy of 2%. It also outperformed the ConvNeXt-XL Cascade-Mask R-CNN (8.6 *FPS* on A100, 55.2% *AP*) based on convolutions, with a speed improvement of over 551% and a higher accuracy of 0.7% *AP*. YOLOv7 also demonstrated superior performance compared to other object detectors such as YOLOR, YOLOX, Scaled-YOLOv4, YOLOv5, DETR, Deformable DETR, DINO-5scale-R50, ViT-Adapter-B, and many others, in terms of both speed and accuracy. It is worth noting that the authors trained YOLOv7 from scratch only on the MS COCO dataset, without using any additional data or pre-trained weights. This demonstrates the effectiveness and feasibility of the model in real-world applications [11].

3.2 Models

In YOLOv7, the authors provide three types of models tailored for different types of GPUs. Among them, YOLOv7-tiny is a lightweight version designed specifically for edge GPUs. YOLOv7-l, on the other hand, is the standard version for regular GPUs. We perform meticulous optimizations on the "neck" section and utilize the proposed combined optimization method to enhance the depth and width of the entire model, resulting in the creation of YOLOv7-x.

3.3 Suggested improvements

This paper opted for SOFT-NMS [12] as the post-processing technique for the network model. Unlike the conventional NMS approach, which identifies the highest-scoring detection box in the test outcomes, assesses if the neighboring detection box should be kept based on the overlap threshold, and assigns a score of zero to the adjacent detection box if it surpasses the threshold value, SOFT-NMS follows a more nuanced approach.

The reset function for fractions in the conventional NMS algorithm is demonstrated in

$$S_i = \begin{cases} S_i & \text{if } IOU(M, b_i) < threshold \\ 0 & \text{if } IOU(M, b_i) \geq threshold \end{cases} \tag{1}$$

However, there is a problem with the traditional NMS algorithm. Only the scores of nearby detection boxes surpassing the threshold in densely packed circumstances, such as natural pineapple fields, are instantly turned to zero. This may cause important detections to be missed. In light of this, this work presented the SOFT-NMS method. Instead of directly zeroing neighboring detection boxes with scores higher than the threshold, a punishment mechanism was implemented via altering the score reset function. Because of this, even some high-scoring detection boxes may continue to function as precise detection boxes in later calculations even if their scores drop during the NMS stage. This significantly improves recall rate and detection accuracy. A Gaussian penalty function was also used to solve the issue of score continuity. The score reset function of the SOFT-NMS algorithm utilized in this study is outlined there after.

$$S_i = S_i \times \exp\left(-\left(\frac{\sigma \cdot IOU(b_i, b_j)}{2}\right)^2\right) \tag{2}$$

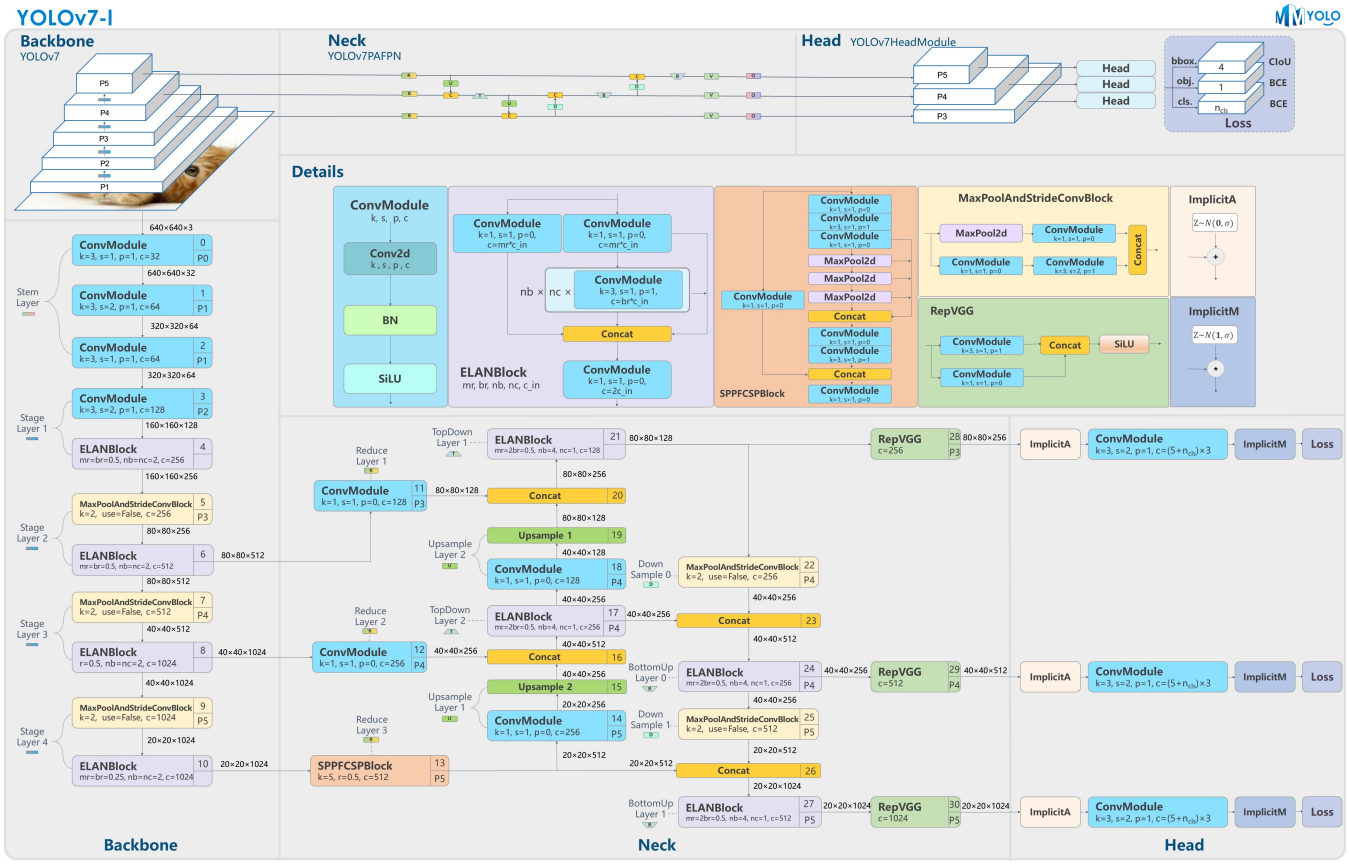

Figure 1: The YOLOv7 architecture consists of a backbone, neck, and head.

$$S_i = \begin{cases} S_i & \text{if } IOU(M, B_i) < \text{threshold} \\ S_i \times (1 - IOU(M, B_i)) & \text{if } IOU(M, B_i) \geq \text{threshold} \end{cases} \quad (3)$$

Where:

S_i : Score of detection box i

$IOU(b_i, b_j)$: Intersection over Union between bbox b_i and b_j

M : Reference bounding box

σ : Parameter for adjusting the strength of the penalty function

To optimize results in spatial images using YOLOv7, we conducted two main stages: training and prediction. In the training stage, we utilized the UCAS_AOD dataset to train the backbones of YOLOv7, including YOLOv7-x, YOLOv7-l, and YOLOv7-tiny. Through this process, the models were optimized to effectively detect objects in spatial images. Following the completion of the training stage, we proceeded to the prediction stage. Here, we employed the previously trained model to perform bounding box predictions on spatial images. The results of this stage were then stored in JSON files, facilitating convenient storage and processing in subsequent steps. Subsequently, based on the JSON files containing bounding boxes, we applied the SOFT-NMS method to eliminate unnecessary bounding boxes. This ensured that only essential and accurate objects were retained, thereby providing more precise classification results. As such, this sequence of procedures combines the training of YOLOv7 models with the utilization of post-processing techniques, resulting in optimized outcomes in object detection and localization in spatial images (2).

4 IMPLEMENTATION AND EXPERIMENTAL RESULTS

4.1 Dataset Description

We utilized the UCAS High Resolution Aerial Object Detection Dataset (UCAS AOD). This dataset comprises over 1500 aerial images captured from above, featuring two classes: airplanes and cars (3). These images were collected in various situations and under different conditions.

4.2 Implementation

We utilized the MMYOLO framework [13], which supports all three versions: YOLOv5, YOLOv6, and YOLOv7. Initially, we used the default parameters to conduct a comprehensive and thorough evaluation. The entire experiment was carried out on an Ubuntu 20.04 machine with an Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz, 63998 MB of RAM, and 2 RTX 2080Ti GPUs.

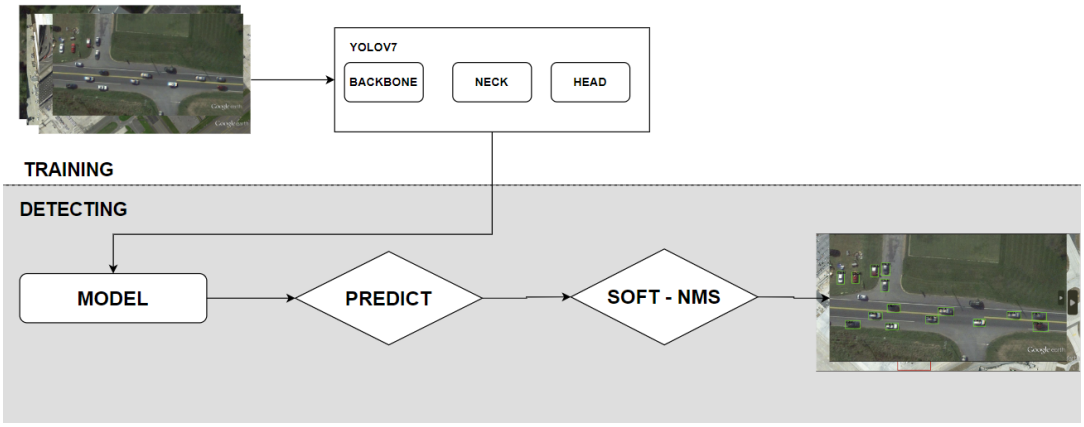

Figure 2: The framework of the proposed method.

Figure 3: Some images from the UCAS AOD dataset

We employ the MS-COCO standard for evaluating object detection tasks, using the Average Precision (AP) metric. The Mean Average Precision (mAP) value is utilized to assess the overall performance across the entire dataset.

4.3 Discussion

Through Table 1 and Figure 4, YOLOv7-x achieves the highest mAP result (0.486) among all supported YOLO backbones. However, the results still have limitations due to multiple proposed bounding boxes around a single object. The proposed post-processing method helps eliminate redundant bounding boxes, thereby improving the results (Figure 4d). The suggested approach employing post-processing techniques yielded the highest result. While YOLOv7 shows relatively promising results, it still faces challenges with aerial images data. Proposing post-processing methods proves crucial in enhancing results by leveraging the outcomes of advanced techniques, which is essential.

5 CONCLUSION

In conclusion, this study proposes an optimization approach for object detection in aerial image using YOLOv7, achieved through a post-processing technique. This opens up numerous applications and holds significant significance in the field of object detection with YOLOv7 and post-processing techniques for aerial image data. Experiments were conducted on the UCAS AOD dataset, and the results were compared with different backbones of YOLOv7. The experimental outcomes distinctly demonstrate that the proposed post-processing technique brings about notable advancements in both performance and accuracy.

ACKNOWLEDGMENTS

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number B2023-26-01.

REFERENCES

[1] P. Zhu *et al.*, "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021.
[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
[3] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464–7475.
[4] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *2015 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2015, pp. 3735–3739.
[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.
[6] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, Springer, 2016, pp. 21–37.
[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
[8] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9657–9666.
[9] K. Nguyen, N. T. Huynh, P. C. Nguyen, K.-D. Nguyen, N. D. Vo, and T. V. Nguyen, "Detecting objects from space: An evaluation of deep-learning modern approaches," *Electronics*, vol. 9, no. 4, p. 583, 2020.
[10] Y. Cao *et al.*, "Visdrone-det2021: The vision meets drone object detection challenge results," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 2847–2854.
[11] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint arXiv:2207.02696*, 2022.
[12] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms–improving object detection with one line of code," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5561–5569.
[13] M. Contributors, *Mmyolo: Openmmlab yolo series toolbox and benchmark*, 2022.

Table 1: The results of evaluating the models on the UCAS dataset.

Backbone	mAP	mAP50	mAP75	mAPs	mAPm	mAPl
YOLOv7-tiny	0.393	0.791	0.339	0.194	0.402	0.525
YOLOv7-l	0.471	0.859	0.468	0.204	0.490	0.641
YOLOv7-x	0.486	0.872	0.534	0.221	0.497	0.667
YOLOv7-x-SNMS	0.501	0.884	0.543	0.227	0.512	0.667

(a) YOLOv7-tiny

(b) YOLOv7-l

(c) YOLOv7-x

(d) YOLOv7-SNMS

Figure 4: Visual results of the four models YOLOv7-tiny, YOLOv7-l, YOLOv7-x, and YOLOv7-x-SNMS.