YOLO-SNH: Target Detection Algorithm for Remote Sensing Data Sets

Haonan Zhang

***Abstract*—In recent years, the detection of targets by unmanned aerial vehicles (UAVs) has gained significant attention in research. However, aerial photography with drones often presents challenges such as object occlusion, large-scale transformations, and the detection of small objects. These difficulties pose significant obstacles for existing deep learning-based target detection algorithms. To address these issues, we propose a novel object detection algorithm called YOLOX-SNH. This algorithm builds upon the transformer structure of the core component, which effectively mitigates object occlusion and preserves essential global context information. Additionally, we have incorporated a specialized detection head to enhance the detection performance for small objects. To evaluate the effectiveness of YOLOX-SNH, we conducted experiments on the VisDrone2021 dataset, comparing it with state-of-the-art object detection methods such as ViT-YOLO. The results demonstrate that YOLOX-SNH outperforms these existing methods, achieving an impressive interpretability in drone capture scenarios. Specifically, when applied to the VisDrone2021 dataset, YOLOX-SNH achieved an average accuracy of 67.00%, surpassing the ViT-YOLO method by 1.11%.**

***Index Terms*—Object detection;YOLO;Visdrone**

I. Introduction

With the advancement of convolutional neural networks (CNNs[1]), CNN-based algorithms have emerged as popular approaches for object detection. Notable examples include R-CNN, Fast-RCNN[2], YOLO series, and SSD. However, when it comes to images captured by drones, general CNN object detection algorithms face challenges in achieving satisfactory results.

One of the primary difficulties arises from the clustering characteristics of tiny objects and the high resolution of drone-captured images. Factors such as flight height and focal length contribute to smaller object sizes in these images. For instance, in the VisDrone2021[6] dataset, approximately 10% of the objects have resolutions within 100 pixels. Moreover, variations in flying height and shooting angle result in significant scale transformations for the same object. Additionally, occlusion occurs when objects of interest overlap or when objects of interest and non-interest overlap. The large size of the images further compounds the detection challenges, as drones capture images with higher resolutions. In the VisDrone2021 dataset, the maximum resolution can reach 2000 x 1500 pixels. However, when these images are scaled down to fit the detector and network, the resolution of smaller objects is further reduced.

Currently, object detection algorithms can be broadly categorized into two types: single-stage and multi-stage. Multi-stage algorithms, such as R-CNN, Fast-RCNN, and Faster-RCNN[3], typically involve two independent steps: candidate frame extraction and subsequent classification and regression. While these algorithms exhibit high accuracy, their complex network structures and cumbersome detection steps result in slower detection speeds, making them unsuitable for real-time performance. In contrast, single-stage object detection algorithms, including SSD, RetinaNet, and the YOLO[4] series, perform object detection with a single feature extraction step on the detection object. These algorithms are widely used in various domains due to their simplicity and fast detection.

II. Related Work

A. YOLOX

The YOLO series is a well-known family of one-stage object detection algorithms. YOLOX[7] represents the latest advancement in this series and introduces several novel structures, including Focus, Spatial Pyramid Pooling (SPP), PANet, and residual networks. The largest YOLOX model achieves an impressive mean Average Precision (mAP) of 69.6% on the MS COCO dataset, while maintaining a fast detection speed of 57.8 frames per second (FPS) on the Tesla V100.

YOLOX offers six different options for network width and depth, namely YOLOX-nano, YOLOX-tiny, YOLOX-s, YOLOX-m, YOLOX-l, and YOLOX-x. As an enhancement over YOLOV3[5], YOLOX utilizes the CSP-Darknet as its backbone, which consists of multiple residual blocks. The feature fusion layer incorporates the Path Aggregation Network (PANet) to combine feature information effectively. In the prediction stage, YOLOX employs three separate heads for classification and regression tasks, unlike previous YOLO versions where these tasks were combined within a single convolution. This separation of classification and regression in the YOLOX Head yields improved predictions by avoiding potential negative effects on network recognition that may arise from the joint implementation of these tasks.

Manuscript received March 20, 2024
Haonan Zhang, School of computer science and technology, Tiangong University, Tianjin, China

B. *Vision Transformer*

In 2017, a transformer model was introduced by V. Aswani et al., primarily used for machine translation and later applied to natural language processing (NLP). This model demonstrated significant advancements in language-related tasks. In 2020, the Google team introduced the Vision Transformer (ViT), which extended the transformer architecture to the field of computer vision. ViT processed image inputs as patches and achieved remarkable performance in various vision tasks. In the same year, the Facebook team presented a comprehensive object detection algorithm called TransformerDETR[9]. Based on the transformer encoder-decoder structure, DETR made predictions by analyzing object relationships and global background information. Building upon previous literature, ViT-YOLO drew inspiration from these developments and incorporated the multi-head self-attention mechanism from the transformer architecture into the YOLO framework. This integration aimed to enhance the performance of object detection tasks by leveraging the powerful attention mechanisms of the transformer.

C. *Multi-scale feature fusion*

In the context of object detection, feature fusion plays a vital role. When utilizing convolutional neural networks (CNNs) for computation, the extracted features exhibit variations in scale and depth within the network. By solely relying on deep-level features, we risk losing valuable information and encountering challenges in detecting small objects. To enhance the efficiency and effectiveness of object detection models, it is essential to leverage features of different scales for prediction.

Currently, the mainstream approach for multi-scale feature fusion is the Feature Pyramid Network (FPN), which adopts a top-down architecture based on a feature pyramid structure. Building upon this concept, PANet[8] introduced an additional bottom-up path aggregation network on top of FPN to further improve feature fusion.This enables the model to determine the relative importance of different content regions and enhances the effectiveness of feature fusion.

III. Dataset Analysis

In this study, we utilized the VisDrone2021 UA V dataset, which was curated by the AISKYEYE team at Tianjin University in China. The dataset comprises 6,471 images in the training set, 548 images in the validation set, and 3,190 images in the test set. Each image in the dataset is annotated with labels from 10 different categories. However, the objects captured by the drones exhibit variations in size and occlusion due to differences in flying heights and focal lengths during data collection. To gain a comprehensive understanding of the VisDrone2021 dataset, we conducted a thorough analysis and identified three key issues associated with it.

A. *Many small goals*

Upon analyzing the training set, we conducted a comprehensive data analysis that resulted in a distribution map showcasing the sizes of all labeled objects. The map, depicted in Figure 1(a), represents the width and height of the label boxes using horizontal and vertical coordinates, respectively. Notably, the map reveals a higher density of points in the lower left corner. This indicates that the VisDrone2019 dataset predominantly consists of small objects.

Furthermore, Figure 1(b) illustrates the distribution of object pixels, indicating that 60% of the pixels fall within the range of 1000, while the remaining 40% exceed this threshold. This finding further emphasizes the prevalence of small objects within the dataset. To visually support this observation, Figure 1(c) presents an example highlighting the abundance of small objects in the dataset.

B. *Objects occlude each other*

During UAV flights at high altitudes, occlusion between objects can occur due to the flight height and shooting angle. Three individuals sitting in an awning-tricycle are marked within the red box, and they occlude each other. Additionally, there is occlusion between the awning tricycle and the people.

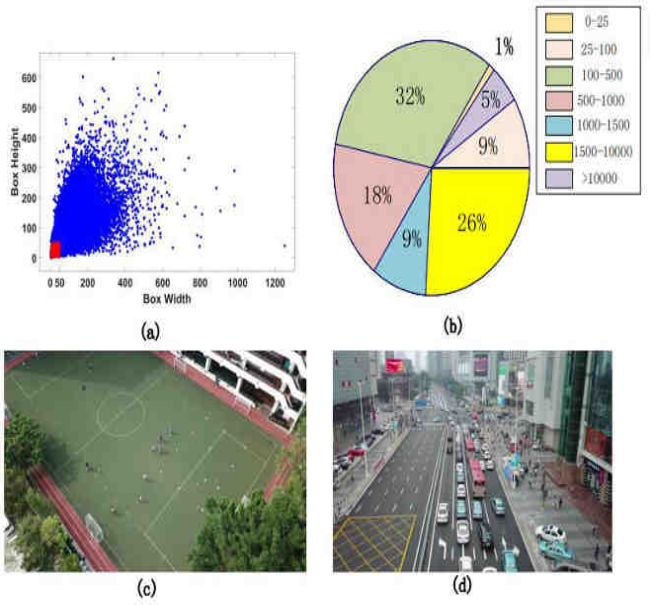
To address this challenge, our study examines the occurrences of occlusion among all objects in the training set. Remarkably, occluded objects make up 61.4% of the entire dataset. This significant presence of occluded objects poses a challenge for YOLO series networks in accurately identifying them. During the non-maximum suppression (NMS) process, severe occlusion can result in missed or falsely detected objects.

C. *Large scale transformation between objects*

When employing drones for image capture, the size of objects of the same type can vary significantly due to variations in flight height. For example, a pedestrian captured in a close-up shot may appear to be similar in size to a car captured in a long-range shot. In close-up views, cars may appear as large objects, whereas they appear much smaller in long-range views.

This substantial variation in object scale and size presents a challenge when extracting features with the model. The extracted features differ due to these variations, making it difficult to effectively fuse features for multi-scale feature fusion. This poses significant challenges in achieving accurate and robust object detection across different scales.

Figure 1. Visdrone data



IV. METHODS

A. *Transformer-Darknet*

The scenes captured by UA Vs often exhibit complex environments and dense distributions of objects. However, traditional convolutional neural networks (CNNs) can suffer from limitations in capturing both global and local context information due to their convolutional operations. To address this issue, a recent paper (Paper 22) proposed the use of a transformer model for object detection, which has shown superior performance in scenarios with high-density severe occlusion and disturbance compared to CNN-based models.
The transformer model replaces the convolutional network with a transformer architecture, which provides more informative object information. The key advantage of the transformer lies in its self-attention mechanism. This mechanism allows the model to assign varying degrees of attention to each pixel in the image, enabling it to focus more on important pixels and reduce the impact of occluded objects on the model's performance. In the implementation, the input image is divided into small tiles, and the self-attention mechanism calculates the relationships and similarities between these tiles. Based on these calculations, a weight matrix is generated. The model then applies this weight matrix to weight and aggregate the information from each tile, resulting in a vector representing the entire image. This vector can be utilized for tasks such as classification and object detection.

To balance the trade-off between network depth and preserving global information, the transformer is incorporated into the last Resblock body, rather than using it throughout the entire network. Each Transformer Encoder module consists of a multi-head attention mechanism and a feedforward neural network (MLP) connected by residuals. This design reduces computational complexity compared to convolutional networks. Additionally, the Transformer Encoder module enables the model to focus on different aspects of information, resulting in a more informative and feature-rich network architecture.

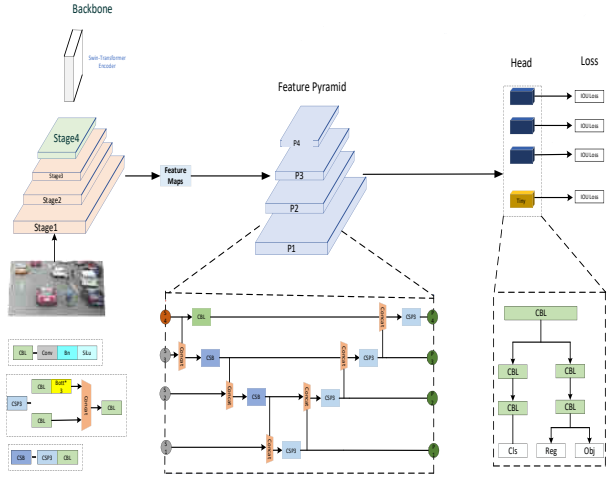
B. *YOLO Head*

In this study, a thorough analysis of the VisDrone 2021 dataset was conducted. The statistical analysis revealed the presence of a significant number of small targets within the dataset, some of which may be challenging to detect visually.

To address this challenge, the paper adopts the TPH-YOLOv5[12] approach (Paper 23), which introduces a specialized prediction head that is highly sensitive to tiny objects. This prediction head operates on a low-level, high-resolution feature map, allowing for improved detection of small targets. By incorporating four prediction heads, the approach mitigates the negative effects of drastic changes in object scale, enabling better detection of targets of different sizes.

It is important to note that the introduction of the prediction head does increase computational effort and the number of model parameters. However, the enhanced detection of small targets achieved through this approach justifies these additional costs.

Figure 2. YOLO-SNH



V. EXPERIMENT

A. *Experiment Details*

The experimental setup for this study involved using Ubuntu 18.04.5 LTS as the operating system. The hardware configuration consisted of an Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz processor and an NVIDIA Tesla A100 GPU. The network analysis was performed using the PyTorch framework (version 1.8.0) with CUDA version 10.2 for GPU acceleration.

To expedite the training process, the paper utilized pre-trained weights from YOLOX-l during the training phase. Since YOLOX-SNH and YOLOX share a significant portion of their backbone architecture, many of the weights from YOLOX-l could be transferred to YOLOX-SNH, reducing the need for extensive re-training.

A series of validation experiments were conducted on publicly available aerial remote sensing data sets. This dataset VisDrone2021 is specially designed for UAV vision tasks, providing a variety of image and video data suitable for training and evaluation of target detection, target tracking, and behavior analysis tasks.

B. *General Performance*

The qualitative experiments presented in Figure 2 showcase the performance of the proposed framework on the VisDrone dataset. The results demonstrate that our method outperforms other models in terms of target classification and regression. Notably, our model exhibits exceptional capability in accurately identifying small targets even in scenarios where occlusion occurs, surpassing the performance of alternative models. These figures serve as visual evidence that our method achieves superior accuracy in object positioning.

C. *Experimental results on Visdrone*

Due to the unavailability of the evaluation server, we followed the established practice of utilizing the validation set for performance evaluation. Table 1 showcases the test results obtained by our proposed approach on the VisDrone2021 validation set. Notably, among the various state-of-the-art object detection algorithms assessed, the ones belonging to the YOLO series demonstrate exceptional performance. Specifically, our YOLOX-SNH model achieves a mean

Average Precision (mAP) of 67.00%, which represents a significant improvement of 5.47% compared to the baseline model. These results highlight the effectiveness and superiority of our approach in object detection tasks.

Table1. Comparison of model performance on Visdrone2021 dataset

Method	Years	mAP（%）
YOLOv5	2020	55.33
Cascade R-CNN [11]	2018	37.84
DPNetv3	2020	62.05
Swin-T	2021	63.91
VIT-Adapter	2022	62.50
YOLOX-L	2021	61.53
VIT-YOLO[10]	2022	65.89
YOLO-SNH(ours)	2024	67.00

Table2. Ablation experiment

Method	mAP
YOLOX	60.86
+ Swin	64.27
+ Swin + head	67.00

VI. CONCLUSION

This paper tackles the problem of low detection accuracy for small objects in unmanned aerial vehicles (UAVs) using conventional algorithms. To address this issue, a comprehensive examination of the VisDrone2021 dataset is conducted, which sheds light on the prevalence of small objects and object occlusion within the dataset. In order to overcome these challenges, a novel feature pyramid structure is devised, and the transformer encoder and detection head are seamlessly integrated into this model for independent learning. This innovative approach aims to enhance the detection performance specifically for small objects in UAV imagery.

REFERENCES

[1] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014:580–587.
[2] Girshick R. Fast r-cnn. In: 2015:1440–1448.
[3] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems. 2015;28.
[4] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: 2016:779–788.
[5] Redmon J, Farhadi A. Y olov3: An incremental improvement. arXiv preprint arXiv:1804.02767. 2018.
[6] Cao Y , He Z, Wang L, et al. VisDrone-DET2021: The vision meets drone object detection challenge results. In: 2021:2847–2854.
[7] Ge Z, Liu S, Wang F, Li Z, Sun J. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430. 2021.
[8] Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. In: 2018:8759–8768.
[9] Naseer MM, Ranasinghe K, Khan SH, Hayat M, Shahbaz Khan F, Yang MH. Intriguing properties of vision transformers. Advances in Neural Information Processing Systems. 2021;34:23296–23308.
[10] Zhang Z, Lu X, Cao G, Y ang Y , Jiao L, Liu F. ViT-YOLO: Transformer-based YOLO for object detection. In: 2021:2799–2808.
[11] Meethal A, Granger E, Pedersoli M. Cascaded Zoom-In Detector for High Resolution Aerial Images. In: 2023:2045-2054.
[12] Zhu X, Lyu S, Wang X, Zhao Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In: 2021:2778–2788.