Small Target Detection Model in Aerial Images Based on YOLOv7X+

Shiqin Li, Weisheng Liu

Abstract—**The detection of small objects in UAV (Unmanned Aerial Vehicle) target acquisition is a persistent difficulty, mainly because of the abundance of small targets, significant mutual obstruction, limited feature statistics, and the intricate background in the target area. These factors collectively result in low detection accuracy. In order to tackle this challenge, we provide a small target recognition model dubbed YOLOv7X+ (You Only Look Once v7X+), which aims to improve the capability of detecting small targets by UAVs while also ensuring its operational effectiveness. The model effectively addresses the issue of significant mutual occlusion among small objects by integrating the Conv2Former module, which enhances the extraction of spatial information features with more precision. Concurrently, we present a Bi-level routing attention technique that is based on cavity convolution. This mechanism regulates the process of extracting and transmitting features at various levels, hence expanding the range of perception. It improves the model's ability to understand and recognize the connections between closely grouped small targets by analyzing individual pixels. This makes the model more resistant to recognizing these targets within complex and changing scenarios. Moreover, it saves crucial contextual details to enhance the precision of differentiation. Finally, the model incorporates multiple dimensions of input data obtained from beneath the UAV by dynamically modulating the size of the convolution kernel. This improves the algorithm's ability to adjust to the specific demands of real-world scenes. The YOLOv7X+ model attains a mean average precision (mAP50) of 60.3% with a validation test Intersection over Union (IoU) criterion of 0.5. The mean average precision (mAP) of the enhanced YOLOv7X model has been increased by 4.9%. and surpasses five other modern sophisticated detection methods. The data from experiments clearly shows that the model effectively handles the identification of small targets in complicated circumstances, achieving a good compromise between accuracy and efficiency.**

Index Terms—**UAV small target detection, YOLOv7X, attention mechanism, Conv2Former, Dynamic Conv.**

I. Introduction

THE maturation of UAV technology with its convenience, high flexibility, efficiency, and multi-perspective capabilities, has become indispensable in numerous industries. For instance, in disaster prevention and investigation, drones' efficient, multi-perspective view enables rapid disaster area detection and real-time data acquisition, significantly enhancing the effectiveness and precision of disaster prevention efforts. Regarding circuit maintenance, drones can substantially reduce manual inspection time and costs, minimizing human resource wastage. In military applications, drones excel in conducting prolonged and extensive surveillance, thereby enhancing defense capabilities. Small target detection has consistently posed a challenging issue in UAVs' myriad target detection tasks. The limited feature information of small targets, coupled with complex backgrounds and other factors in the target area, leads to suboptimal detection accuracy. Furthermore, UAV-based target detection exhibits characteristics such as a high density of small targets, significant scale variations between different categories, and intricate backdrops, all of which substantially impact the detection process. Traditional target detection methods rely on manual features like color, shape, and texture, which can yield satisfactory results. However, these techniques often require a substantial amount of time and suffer from limited and non-uniform accuracy of detection, especially in complex scenes with significant occlusion. As a result, it is vital to develop a highly effective network model for improving the precision of detecting small targets in UAV-related contexts.

In the past few years, the advancements in deep learning theories and technologies have rapidly progressed, and they have surpassed traditional methods in general-purpose target detection tasks. Deep learning-based general-purpose target detection algorithms can be classified into two-stage algorithms like RCNN [1] and single-stage algorithms like YOLO [2] and SSD [3]. Whereas single-stage detectors excel in providing end-to-end efficiency, they usually exhibit compromised accuracy when it comes to localizing and identifying small targets. On the other hand, two-stage target detectors, which first localize and then recognize, achieve superior accuracy but may lag in real-time execution. Integrating generic target detection algorithms directly into UAV-based target detection tasks often results in poor model generalization and a decline in detection performance. In response to UAV images featuring numerous small targets and significant occlusions, researchers have primarily focused on enhancing small target detection accuracy. They have pursued targeted improvements in areas such as data augmentation, anchor-less methods optimization, and lightweight networks.

Sommer leveraged Fast-RCNN, a two-stage network known for its efficacy in detecting small objects, to detect vehicles in drone photos. They fine-tuned anchor frame sizes and feature map resolutions to enhance small object detection. Nonetheless, the precision of detecting small objects remained subpar due to extensive feature map detail loss caused by convolution and pooling techniques [4]. To enable feature extraction across various scales and enhance small target detection performance, Lin devised a feature pyramid structure. This network enables the integration of detailed lower-level characteristics with high-level semantic information from top features, ultimately improving small object

Manuscript received July 7, 2023; revised December 23, 2023.

This work was supported by the Special Fund for Scientific Research Construction of University of Science and Technology Liaoning.

Shiqin Li is a postgraduate student at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: Lsq01022023@hotmail.com).

Weisheng Liu is a professor of the College of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, CO 114051, China (corresponding author to provide fax: 0412-5929809; e-mail: succman@163.com).

recognition capabilities [5]. Liang introduced a perceptual generative adversarial network (GAN) to produce high-resolution representations of small objects. This approach leverages the association between large and small objects to strengthen the visual depiction of small objects, making them resemble illustrations of larger objects [6]. Hu identified that pooling operations can distort the configuration of small objects. As a remedy, they introduced a novel context-aware Region of Interest (ROI) pooling approach [7]. Chen introduced the ResNeXt-d fusion architecture, which enhances small object perception and subsequently enhances the detection of small and densely packed objects [8]. Tang designed a coarse anchor-free detector called CPEN to address the challenge of detecting densely packed small objects, achieving remarkable results [9]. Yang introduced the SCRNet++ model, pioneering the integration of denoising concepts into target detection. This model conducts instance-level denoising on feature maps, resulting in improved small target detection in drone images [10]. Liao introduced the UGGNet model, which employs a local localization LLM to predict target distributions. It subsequently generates dense target areas of interest using an uncontrolled grouping module for detection, significantly reducing the time required for small target detection [11]. Singh presented a significant enhancement to the YOLOv5-based algorithm. They incorporated a novel feature fusion layer with a compact field of receptivity into the YOLOv5 feature pyramid, enabling the capture of fine details in the feature map. Additionally, this layer introduced a horizontal connection from the network's shallow portion to preserve feature map resolution and prevent information loss in the deeper layers. These combined measures markedly improved small target detection performance [12]. Xu devised a locally aware Swin-Transformer backbone that amalgamates Transformer and Convolutional Neural Network advantages to enhance the performance of the small object detection [13]. While the aforementioned algorithms have enhanced target detection in drone applications, they endure encounter substantial challenges.

As a consequence, this article presents an advanced version of the YOLOv7X model. It addresses the challenges of significant target occlusion and dense small targets effectively by incorporating Conv2Former, RES-DBAConv, and ODConv into the feature extraction process. These challenges are commonly encountered in complex backgrounds of high-angle images captured by drones. Furthermore, we validated the efficiency and real-time capabilities of the suggested approach by using the VisDrone2019-DET dataset.

The chapter structure is as follows: Chapter 2 offers an overview of the structure and characteristics of the original YOLOv7X model. Chapter 3 introduces the network structure of our enhanced YOLOv7X model. In the final chapter, Chapter 4, we offer a comparative analysis of our model with YOLOv7X and other deep learning-based models, highlighting its superior performance.

II. Relate work

Among the YOLO series target detection algorithms, YOLOv7X [14] stands out for enhancing the fine details of small targets when compared to its predecessors. YOLOv7X incorporates novel algorithms, network architectures, and introduces methods like multi-scale feature fusion and cross-scale feature transfer. These innovations enhance its ability to capture intricate characteristics of small targets and elevate detection accuracy. YOLOv4 [15] and YOLOv5 [16] have also implemented new techniques to enhance the detection accuracy of small targets, including the utilization of techniques like SPP Block and PANet [17]. Furthermore, distinctions in the nuanced characteristics of small targets among these algorithms are evident in their data set selection, data processing methods, and parameter configurations employed during the training phase. Additionally, distinctions among these algorithms concerning the nuanced characteristics of small targets are manifest in their selection of datasets, methods of data processing, and parameters chosen for training. Overall, although all of the aforementioned algorithms, while all of these algorithms possess certain advantages in detecting small targets, YOLOv7X excels in the capture and processing of fine-grained details of small targets.

YOLOv7X network can be dissected into three key segments for a clearer understanding of its architecture. These segments consist of the backbone network, neck network, and head network. Moving on to the backbone network, YOLOv7X adopts a state-of-the-art structure, perhaps building upon the success of previous YOLO versions or introducing entirely novel architectures. This backbone network is designed to effectively generate feature layers through the integration of various structures, potentially including Conv2D_BN_SiLU (CBS), ELAN, MP and SPPCSPC structures. These components work together to produce feature maps that will be crucial for object detection. The neck network in YOLOv7X plays a pivotal role in enhancing the feature layers and facilitating their fusion. This is achieved through the construction of a Feature Pyramid Network (FPN) or a similar mechanism, ensuring that the model can effectively utilize the features extracted by the backbone network. Finally, in the head network of YOLOv7X, the model fine-tunes the number of channels and makes predictions based on the processed features.

III. Improved YOLOv7X Algorithm

Although the YOLOv7X model offers advantages in drone-based target detection, it has certain limitations in capturing specific details and ensuring accuracy when dealing with small targets. Despite YOLOv7 incorporating technologies like multiple scales merging of features and cross-scale feature transmission to enhance the precision and efficiency of small target detection. However, in drone-based target detection, factors such as severe target occlusion, dense target distribution, and high-angle shooting can lead to significant variations in target size, shape, posture, and motion direction. Traditional convolutional neural networks are unable to address problems like missing information in local regions and target occlusion. Consequently, issues related to false or missed detections persist in the results, necessitating further optimization of the algorithm and model structure. Furthermore, in more complex scenarios, YOLOv7's accuracy might be compromised, demanding superior data support and additional model refinements to enhance its performance concerning detailed features and accuracy.

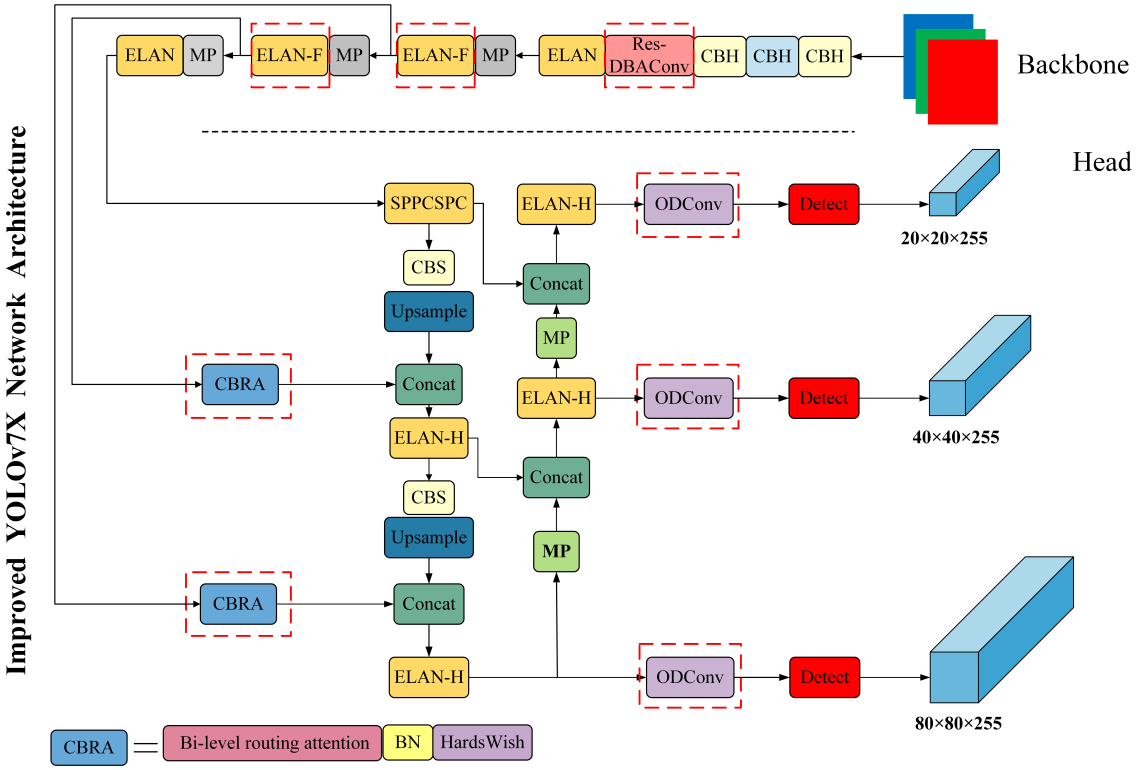

Fig. 1. Network structure of the enhanced YOLOv7 model

A. *YOLOv7X+ model architecture*

In response to these challenges, we enhanced YOLOv7X, leading to the development of a novel model named YOLOv7X+.

The introduction of the Conv2Former [18] element enables the extraction of spatial linkages and global details. which can automatically adapt various representations of features, effectively addressing object occlusion issues and enhancing object detection accuracy and efficiency. Furthermore, the Mixed Residual Hole Convolutional Attention Module (RES-DBAConv) can dynamically select and merge features from various levels, leveraging distinctions in local and global information within drone images. Particularly for the detection of densely populated small targets, the utilization of the Bi-level routing attention mechanism [19] allows for improved extraction of pertinent information from small targets, achieved through the amalgamation of dual-layer attention mechanisms. Lastly, the introduction of ODConv [20] to cater to targets of varying sizes enables adaptive adjustments to the convolutional kernel size, enhancing the capability to capture small target features and thereby improving detection accuracy and robustness. Figure 1 illustrates the enhanced model, YOLOv7X+.

B. *Conv2Former*

Due to the abundance of high-resolution photos containing obscured small targets in the UAV dataset, we have opted to utilize the Conv2Former module for feature extraction instead of the conventional convolution module. Its architecture is illustrated in Figure 2, representing a transformer-style convolutional network with a pyramid-like configuration that includes varying numbers of convolutional blocks across four stages. Each stage presents unique feature map resolutions, with a patch-embedding block incorporated between Successive phases to lower the resolution. The core of this approach revolves around the convolutional modulation operation, shown in Figure 4 and Equation (1). It employs deep convolutional features exclusively as weights for representation modulation, combined with the Hadamard product to streamline the self-attention mechanism and enhance the effectiveness of large kernel convolution. Conv2Former replaces the ELAN-F convolution block in the original YOLOv7 Backbone. Compared to the original structure, Conv2Former can more effectively capture the network's global information and contextual semantic information, thereby obtaining rich features for fusion operations and improving network performance.

$$\begin{aligned} z &= A \odot V \\ A &= D\,\mathrm{Conv}_{k\times k}\left(W_1 X\right) \\ V &= W_2 X \end{aligned} \tag{1}$$

Here $\odot$ represents the Hadamard product, W_1 and W_2 denote the weight matrices of the two linear layers, and $DConv_{k\times k}$ signifies depth convolution with a kernel size of $k \times k$. This convolutional modulation operation permits each spatial location, denoted as (h, w), to be linked with all pixel levels within a $k \times k$ square region centered at (h, w). Information interplay among channels is accomplished via the linear layers. The result at each spatial point is the total of the pixel values inside the specified square region, with each pixel value weighted accordingly.

Considering the superior performance of Conv2Former-L over the other five variants of Conv2Former, we opted to combine it with the benchmark model YOLOv7X. Our findings indicate that the Conv2Former module excels at dealing with occluded targets. Moreover, this module enhances computing efficiency, particularly when handling catastrophic images like those found in VisDrone2019.

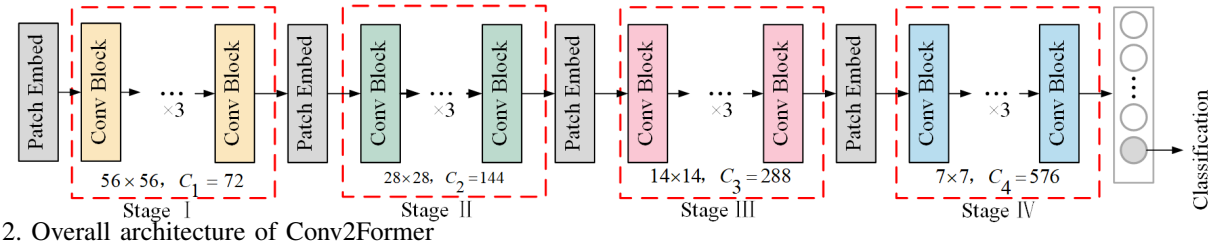

Fig. 2. Overall architecture of Conv2Former

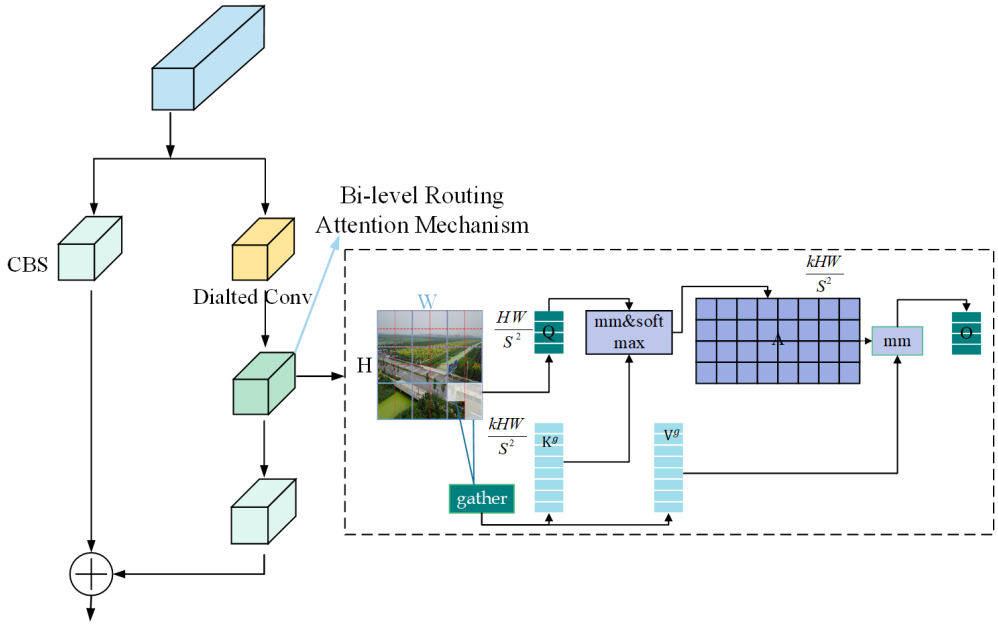

Fig. 3. RES-DBAConv network structure

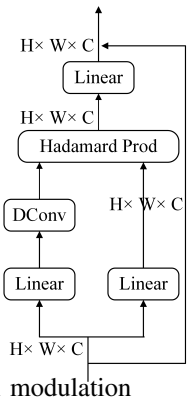

Fig. 4. Convolutional modulation

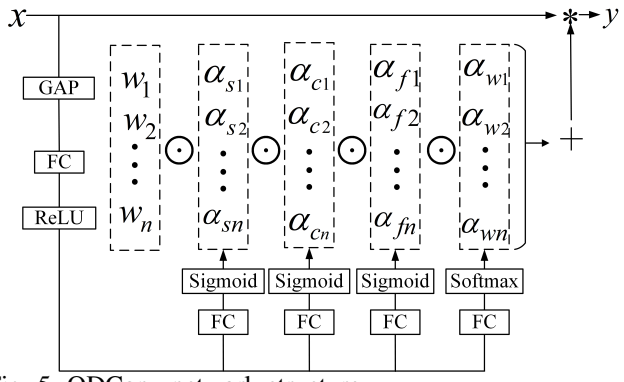

Fig. 5. ODConv network structure

C. RES-DBAConv

Conventional convolution calculations primarily capture information from nearby regions and can lack sensitivity in the detection of small objects. Dilated convolution separates pixels within the convolution kernel, thereby enlarging the receptive field. This allows the network to effectively gather distant information. Dilated convolutional neural networks offer improved accuracy in analyzing local image regions, mitigating the issue of local information loss in small target detection and enhancing small target information capture. Furthermore, during convolutional feature generation, two-layer routing the Bi-Level Routing attention technique can able to applied to selectively exclude less relevant regions at a broader level while preserving essential routing locations. This enhances the precision of detecting small targets detection but lowers detection error and missed detections for densely packed small targets, all while maintaining network reliability. By leveraging the strengths of dilated convolution with Bi-level routing attention, this study introduces a novel hybrid residual void convolution attention module (RES-DBAConv), illustrated in Figure 3.

Bi-Level Routing is a technique that dynamically and efficiently handles attention in a sparse manner, allowing for querying and retrieval of information. Its fundamental concept is as follows: for a feature map $X \in R^{H\times W\times C}$ produced by a null convolution, it is initially separated into non-overlapping $S\times S$ regions. Subsequently, the feature map is converted to $X^r \in R^{S^2\times \frac{H\times W}{S^2}\times C}$. Next, linear projections are applied to derive Q, K, V. Utilizing these linear projections of Q, K, the Q^k, K^r values of key regions are obtained. This information is used to deduce the adjacency matrix of these key regions, denoted as $A^r = Q^r(K^r)^T$. As serves as a metric for measuring semantic similarity between two regions. Subsequently, the routing index matrix I is generated to selectively keep the initial K connections in each of the regions. To circumvent memory constraints, the gather operation is employed to combine the tensors K and V.

This results in $K^g = gather(K, I^r), V^g = gather(V, I^r)$. Lastly, K^g and V^g are utilized in the attention mechanism $O = Attention(Q, K^g, V^g) + LCE(V)$, with the local context enhancement mechanism $LCE(V)$ assisting the network in preserving finer, details at lower-level. This proves particularly beneficial for detecting information in tasks involving densely clustered small targets in images. Given that low-level feature mappings share a comparable resolution with the input image, offering higher spatial resolution and often containing precise accurate details, it enhances the perception of small targets' details and provides more exact location information.

Consequently, we integrate this concept by creating a novel residual structure, RES-DBAConv. This structure is formed via a Bi-level routing attention mechanism combined with a dilated convolution (rate of 3) applied to the YOLOv7X backbone network's shallow layer. This integration enhances YOLOv7X+ in learning feature dependencies and comprehensively understanding small targets, ultimately rendering it better suited for detecting densely packed small targets. This results in enhanced detection accuracy and efficiency.

D. ODConv

In UAV scenarios, the varying positions and angles of UAV shots result in images of the same target having diverse sizes, shapes, and complex background clutter. This presents challenges for detecting and recognizing small targets. Traditional CNN models are typically limited to processing multi-dimensional image data, making it challenging to accommodate the multi-dimensional input data from UAVs.

ODConv(Onmi-Dimensional Dynamic Convolution) adjusts the dimensions of the convolutional kernel to accommodate varying target sizes. This capability enables ODConv to effectively manage the complexity of UAV scenarios and changing lighting conditions. This is particularly crucial for small target detection under UAV conditions, where small targets are susceptible to environmental disturbances. Consequently, full-dimensional dynamic convolution can significantly aid in small target detection under UAV conditions, enhancing accuracy and efficiency of detection, and bolstering model versatility. ODConv can be described as illustrated in Figure 5 and defined by Equation (2).

$$\begin{aligned} y = &(\alpha_{wi} \odot \alpha_{f1} \odot \alpha_{c1} \odot \alpha_{s1} \odot W_i + \dots \\ &+ \alpha_{wn} \odot \alpha_{fn} \odot \alpha_{cn} \odot \alpha_{sn} \odot W_n) * x \end{aligned} \tag{2}$$

ODConv incorporates a multi-dimensional attention mechanism to comprehend the convolution kernel space's four dimensions. This is achieved through a parallel approach that combines these four categories of attention, making them mutually reinforcing. By sequentially applying convolution with varying attention across location, channel, filter, and kernel dimensions, the convolution process is able to adapt to the input dimensions, enhancing its ability to capture rich contextual information. This significantly enhances the feature extraction capabilities of convolution. We opt to integrate ODConv and PANet at the neck, enabling the adaptive capture of spatial structure and object features through the convolutional kernel's dynamic size and shape. This enhances the model's capability to recognize intricate scenes and objects, subsequently enhancing its resilience to multi-angle data.

E. Improvement of Activation Function

The selection of an adequate activation function is crucial for enhancing the model's performance, considering the computational efficiency. The activation function Hardswish was first introduced in MobileNetV3 [21]. Compared to the SiLU activation function, Hardswish offers advantages such as lacking lower and upper bounds, flatness, and non-monotonicity. These characteristics enhance the expressive capacity of the neural network. Hardswish is preferred over SiLU because of its substitution of the computationally intensive sigmoid function with a piecewise linear approximation that is less taxing on computational resources than SiLU. The Hardswish activation function, represented by Equation (3), utilizes the input value x.

$$HardsWish = \begin{cases} 0, \text{ if } x \leq -3 \\ x, \text{ if } x \geq +3 \\ \frac{x(x+3)}{6}, \text{ otherwise} \end{cases} \tag{3}$$

IV. Experiments and analysis

This experiment assesses the improved YOLOv7X model's enhanced detection performance for identifying densely packed small targets, such as vehicles and pedestrians, in highly complex backgrounds. Initially, we compared the improved YOLOv7X model with the original YOLOv7X model to assess the impact of our modifications. Subsequently, we evaluated the enhanced YOLOv7X model against various other Deep Learning-based target detection models to derive comprehensive experimental findings.

A. Dataset Processing

Training target detection model algorithms necessitates an ample dataset, and prominent UAV image datasets for this purpose comprise VisDrone [22], UAVDAT [23], DOTA [24], and others.

In this study, we utilize the VisDrone2019-DET dataset for our experimentation. The VisDrone dataset consists of photos acquired by several unmanned aerial vehicle (UAV) cameras. The dataset was acquired by the AISKYEYE team from the Machine Learning and Data Mining Laboratory at Tianjin University. It comprises photos obtained from a diverse range of drone cameras. It encompasses 141 urban landscapes in China and encompasses 10,000 images with 2.6 million annotations. Several images in this dataset boast resolutions as high as 2000×1500 pixels. The training dataset contains 6471 images with corresponding annotations, while the validation set comprises 1115 images with annotations, and the test set includes 547 original images. The dataset encompasses 10 categories, notably, the pedestrians and people classes, which are prone to confusion. Additionally, as depicted in Figure 6, the VisDrone2019-DET dataset features a substantial quantity of detected objects, including a substantial quantity of diminutive, scattered large targets, an imbalanced dataset scatter, a substantial count of densely packed objects, and significant occlusion of objects. These factors present considerable challenges in algorithm design. Furthermore, it remains a demanding dataset for detection and tracking tasks.

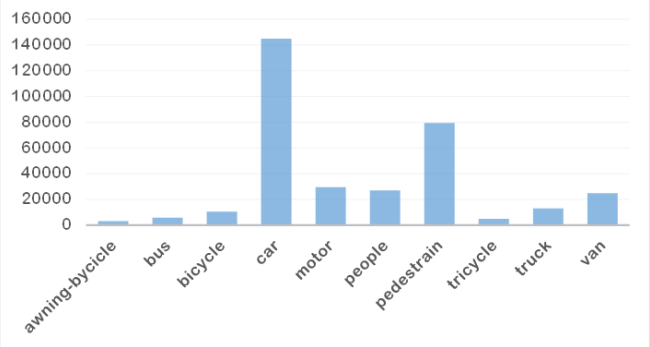

Fig. 6. Statistics of VisDrone dataset category instances

B. Experimental environment

The experimental server was equipped with an Intel(R) i7-6850K CPU, 64GB of RAM, and an NVIDIA GeForce RTX 3090 GPU, running on the operating system Windows 10. Our experimental code is built upon an enhanced initial release of Ultralytics' YOLOv7 project. It offers compatibility with both cfg and yaml model files, and all algorithms used here are adapted from the official model in this project. During the training process, we iterated the model for 300 epochs, with a batch size of 16 and a learning rate of 0.01, using a one-cycle learning rate decay strategy. While keeping other parameters at their default values. The enhanced YOLOv7X model was subsequently employed to train the VisDrone2019 training dataset, and its performance was evaluated through validation on the VisDrone2019 validation dataset.

C. Evaluation Criteria

To effectively access the detection ability of deep convolutional models on images captured in UAV circumstances, this work employs Mean Average Precision (mAP), a widely established metric for for evaluating target detection. The mAP is calculated as the average of precision (P) and recall (R) across various dataset instances. Definitions for these metrics are provided in Equation (4), where TP represents the count of true positives.

$$\begin{aligned} R &= \frac{TP}{TP+FP} \\ P &= \frac{TP}{TP+FN} \\ \text{map} &= \frac{1}{N}\sum_{i-1}^{N} AP_i \end{aligned} \tag{4}$$

We employed the COCO [25] evaluation criteria along with the Pyecotools tool for evaluating and analyzing the detection results. If the Intersection over Union (IOU) between the detection box and the ground truth exceeds 0.5, we consider the target as accurately detected.

D. Ablation Experiments

1) Conv2Former module comparison experiment: To assess the Conv2Former module's impact on detection accuracy, this study employs YOLOv7X as the benchmark model. A comparative test was conducted by integrating the Conv2Former module into the ELAN module, utilizing default experimental parameters and resolution settings. Table I clearly evident that the Conv2Former module achieved a 2.5% increase in accuracy and a 3.6% increase in mAP compared to the original YOLOv7X model, as observed in rows 1 and 3.

2) Comparison Experiment of the RES-DBAConv Module: To validate the performance of the RES-DBAConv module, we maintain YOLOv7X as the reference model and incorporate it into the shallow network of YOLOv7X to collect precise data on small targets for comparative trials. Furthermore, integrating Bi-Level Routing attention with the feature fusion module, it minimized the chances of incorrect identifications and overlooked detections for small targets. The circumstances of the experiment remain consistent with the previous setup. Table II demonstrates that RES-DBAConv improves the model's mean average precision (mAP) by 1.1% and recall by 2.1% in rows 1 and 4.

3) Comparison Experiment of the ODConv Module: To assess the impact of the ODConv module on the accuracy of model detection, this study employs YOLOv7X as the benchmark model. The ODConv module is integrated into PANet to address the challenge of various scale variations in the target. As indicated in rows 1 and 4 of Table I, ODConv enhances the mAP by 0.4%, and there is a slight improvement in core retrieval precision.

E. Comparative Experiments

TABLE II
PERFORMANCE OF THE ENHANCED YOLOV7X+ MODEL COMPARED WITH THE OTHER MODELS

Method	ImageSize(x)	mAP	BFLOPS
YOLOv5	640 × 640	49.33	50.4
THP-YOLOv5 [26]	640 × 640	57.31	-
YOLOvX-S [27]	640 × 640	53.5	-
ClusDet [28]	640 × 640	56.2	-
MobileNetv3 [29]	640 × 640	55.4	23.8
MobileViT [30]	640 × 640	55.5	-
YOLOv7X+	640 × 640	60.3	56.8

TABLE I
ABLATION EXPERIMENTAL RESULTS OF THE IMPROVED YOLOV7X+ ALGORITHM

Method	Conv2Former	RES-DBAConv	ODConv	mAP	Precision/%	Recall/%
YOLOv7	-	-	-	55.4	63.3	53.3
YOLOv7	√	-	-	57.9	66.9	54.4
YOLOv7	-	√	-	56.5	64.8	55.4
YOLOv7	-	-	√	55.8	63.6	53.5
YOLOv7	√	√	√	60.3	70.3	58.7

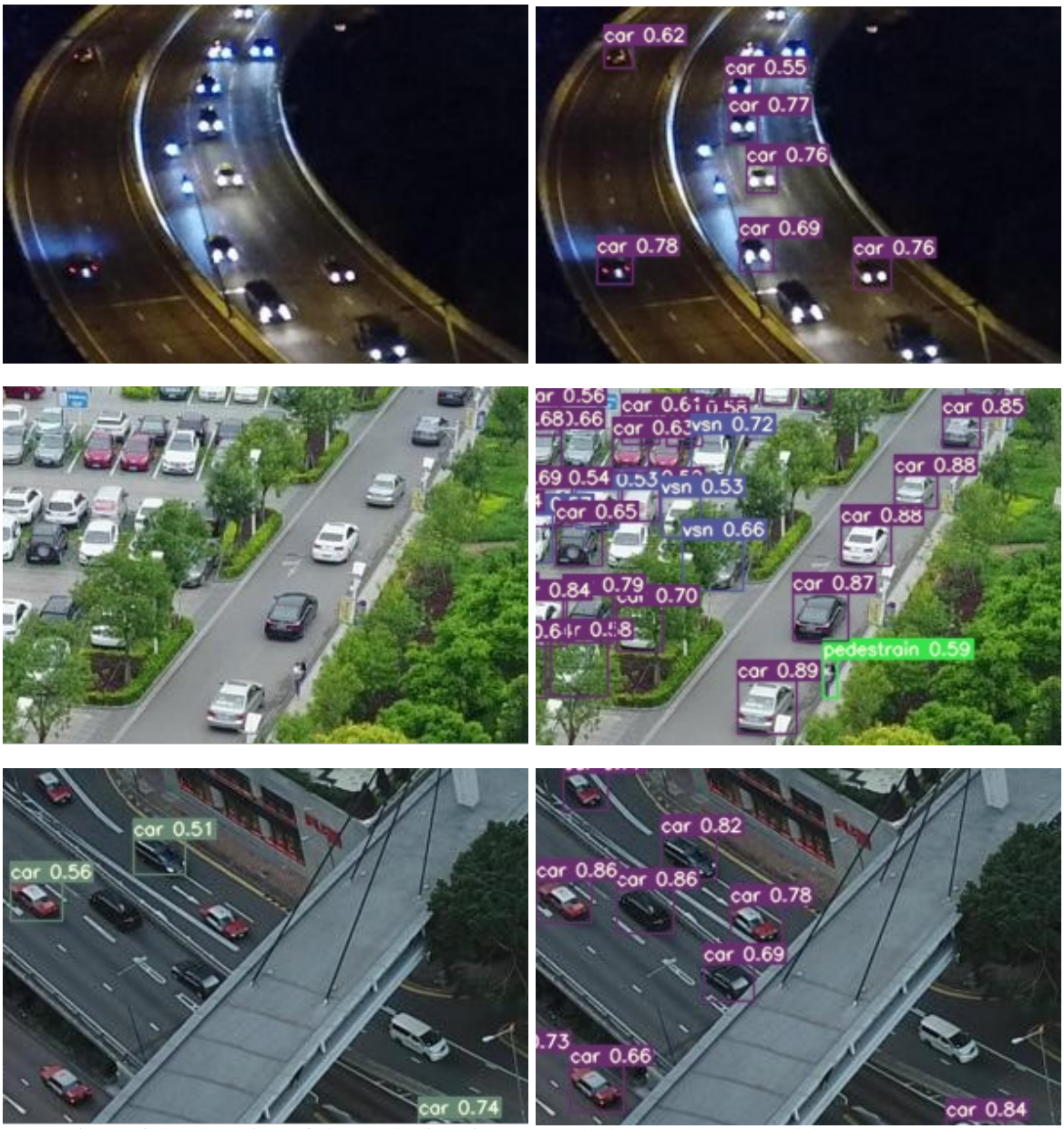

Fig. 7. Comparison of the detection of the model in this paper and the YOLOv7X model

With the intention of assessing the efficacy of the method suggested in this research, we performed comparisons with various popular algorithms to showcase the superiority of the YOLOv7X+ model. We compared YOLOv7X+ and other widely used models using the VisDrone dataset, and the results are presented in Table II. Upon comparison, it was found that the benchmark model YOLOv7X achieves a mAP@.50 accuracy of 55.4% at a resolution of 640 × 640. The YOLOv7X+ model that we introduced in this paper has shown superior capability in accurately detecting real targets within significantly overlapping target groups from a distance. Furthermore, the YOLOv7X+ model exhibits superior detection accuracy and robustness when identifying small distant targets.

F. Analysis of Model Visualization

YOLOv7X+ has demonstrated its effectiveness in detecting small objects in real-world scenarios by identifying representational and elaborate images within the VisDrone2019 dataset. This article evaluates and visually presents various approaches to detecting small targets. Figure 7 shows comparative detection results between the original YOLOv7X and YOLOv7X+ for dense small targets under varying conditions, including rapid camera rotation, occlusion and high-altitude photography. Compared to the original model, the enhanced YOLOv7X+ exhibits superior recognition and detection capabilities when handling small targets with diverse characteristics. In summary, the outcomes affirm that YOLOv7X+ is more robust in detecting small targets with diverse characteristics.

V. Conclusion

In summary, the outcomes affirm the enhanced robustness of YOLOv7X+ in detecting small targets with diverse characteristics. YOLOv7 is currently a widely used deep learning framework for object detection. This study introduces an improved model called YOLOv7X+, designed for object detection in drone target detection in complex scenarios using the YOLOv7X framework. Its goal is to exactness the

accuracy of detecting small and densely-packed targets in unmanned aerial vehicle circumstances, rendering it appropriate for precise aerial small target detection tasks. In our research, we analyzed the distribution patterns of targets within the VisDrone2019 dataset within the context of drone shooting situations. Firstly, the RES-DBAConv is integrated into the superficial layer of the YOLOv7X backbone network. This enhances the model's precision in identifying densely packed small targets and captures the finer features of the targets. Secondly, the Conv2Former architecture was introduced into the YOLOv7X backbone network. This addition generates feature maps with more extensive global information, strengthening the model's competence to learn from the entire feature space and improving the detection accuracy of obscured small targets. Lastly, through the fusion of ODConv convolution and the PANet structure, the model can effectively handle variations in target size and angle resulting from drone capture at diverse angles. Experimental results demonstrate that the enhanced YOLOv7X+ model notably enhances model accuracy, achieving an average accuracy increase of 4.9% as opposed to the benchmark YOLOv7X model on the VisDrone2019-DET dataset. However, YOLOv7X+ still encounters instances of missed detections and false positives for tiny targets. In future research, our focus will be further optimization of this model to improve small target detection results and exploration of methods for creating lightweight network models without compromising detection accuracy.

REFERENCES

[1] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2013.

[2] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2015.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, and C.-Y. F. et al, "SSD: Single Shot MultiBox Detector," 2016.

[4] L. W. Sommer, T. Schuchert, and J. Beyerer, "Fast Deep Vehicle Detection in Aerial Images," *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 311–319, 2017.

[5] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature Pyramid Networks for Object Detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2016.

[6] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual Generative Adversarial Networks for Small Object Detection," *CoRR*, vol. abs/1706.05274, pp. 1951–1959, 2017.

[7] X. Hu, X. Xu, Y. Xiao, H. Chen, S. He, and J. e. a. Qin, "SINet: A Scale-Insensitive Convolutional Neural Network for Fast Vehicle Detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 1010–1019, 2019.

[8] C. Chen, W. Gong, Y. Chen, and W. Li, "Object Detection in Remote Sensing Images Based on a Scene-Contextual Feature Pyramid Network," *Remote. Sens.*, vol. 11, p. 339, 2019.

[9] Z. Tang, X. Liu, G. Shen, and B. Yang, "PENet: Object Detection using Points Estimation in Aerial Images," *IEEE International Conference on Machine Learning and Applications (ICMLA)*, vol. abs/2001.08247, pp. 392–398, 2020.

[10] X. Yang, J. Yan, X. Yang, J. Tang, W. Liao, and T. He, "SCRDet++: Detecting Small, Cluttered and Rotated Objects via Instance-level Feature Denoising and Rotation Loss Smoothing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 2384–2399, 2020.

[11] J. Liao, Y. Piao, J. Su, G. Cai, X. Huang, and L. e. a. Chen, "Unsupervised Cluster Guided Object Detection in Aerial Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 11 204–11 216, 2021.

[12] G. M. Inderpreet Singh, "Improved YOLOv5 for Small Target Detection in Aerial Images," *SSRN Electronic Journal*, 2022.

[13] C. Xu, J. Wang, W. Yang, and L. Yu, "Dot Distance for Tiny Object Detection in Aerial Images," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 1192–1201.

[14] 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," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7464–7475, 2022.

[15] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *ArXiv*, vol. abs/2004.10934, 2020.

[16] G. R. Jocher, A. Stoken, J. Borovec, NanoCode, A. Chaurasia, and T. et al, "ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and Youtube integrations," 2021.

[17] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance Segmentation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. abs/1803.01534, pp. 8759–8768, 2018.

[18] Q. Hou, C. Lu, M.-M. Cheng, and J. Feng, "Conv2Former: A Simple Transformer-Style Convnet for Visual Recognition," *ArXiv*, vol. abs/2211.11943, 2022.

[19] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. W. H. Lau, "BiFormer: Vision Transformer with Bi-Level Routing Attention," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 323–10 333, 2023.

[20] C. Li, A. Zhou, and A. Yao, "Omni-Dimensional Dynamic Convolution," *ArXiv*, vol. abs/2209.07947, 2022.

[21] A. G. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, and M. T. et al, "Searching for MobileNetV3," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314–1324, 2019.

[22] P. Zhu, D. Du, L. Wen, and X. Bian, "VisDrone-VID2019: The Vision Meets Drone Object Detection in Video Challenge Results," *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 227–235, 2019.

[23] M. Mueller, N. Smith, and B. Ghanem, "A Benchmark and Simulator for UAV Tracking," vol. 9905, pp. 445–461, 2016.

[24] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. J. Belongie, and J. L. et al, "DOTA: A Large-Scale Dataset for Object Detection in Aerial Images," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, 2017.

[25] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, and P. D. et al, "Microsoft COCO Captions: Data Collection and Evaluation Server," *ArXiv*, vol. abs/1504.00325, 2015.

[26] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 2778–2788.

[27] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," *ArXiv*, vol. abs/2107.08430, 2021.

[28] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, "Deep Learning for Unmanned Aerial Vehicle-Based Object Detection and Tracking: A survey," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 1, pp. 91–124, 2022.

[29] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, and T. W. et al, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *ArXiv*, vol. abs/1704.04861, 2017.

[30] M. R. Sachin Mehta, "MobileVit: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," *ArXiv*, vol. abs/2110.02178, 2021.