Improved YOLOv5 Algorithm for Small Object Detection in Drone Images

Yitong Lin and Yiguang Liu

Vision and Image Processing Lab, College of Computer Science, Sichuan University,
Chengdu 610065, China
liuyg@scu.edu.cn

Abstract. The object detection in the context of drone is a hot topic in the field of computer vision in recent years. In response to the challenge of limited image feature information and the presence of numerous small and densely packed objects in drone-captured images, this paper proposes a novel feature fusion detection model, HTH-YOLOv5, based on YOLOv5. Firstly, we enhance the detection capability of small objects by adding a detection channel from high-resolution feature maps and propose a Hybrid Transformer Head (HTH) that incorporates a hybrid Transformer module, aiming to improve the network's focus on small objects by fusing global and local feature information. Secondly, we introduce a Convolutional Attention Feature Fusion module(CA-FF) based on CBAM. This module dynamically adjusts attention weights for the allocation of original feature maps in both channel and spatial dimensions, aiming to enhance the feature extraction capability for small objects. Finally, to better capture global and contextual information, we introduce the Hybrid Transformer module into the backbone and enhance its original feature fusion method using the CA-FF module. Experiments on the Vis-Drone 2021 dataset demonstrate that, compared to the baseline YOLOv5s model, the improved model shows an increase of 7.2% in mAP_{50} and 6.3% in mAP_{75}. The model trained with an input resolution of 1540×1540 achieves an mAP_{50} of 57.1%, marking a 12.4% improvement over YOLOv5. The improved HTH-YOLOv5 achieves increased accuracy while maintaining a detection speed of 45 FPS, making it more suitable for small object detection in drone scenarios.

Keywords: Drone · Small object detection · Attention mechanism · Feature fusion

1 Introduction

In recent years, with extensive research into artificial intelligence technology, object detection technology in drone captured scenes has found widespread applications in various fields such as transportation, defense, wildlife conservation, and plant protection. In this article, we focus on improving the performance of small object detection in drone-captured images.

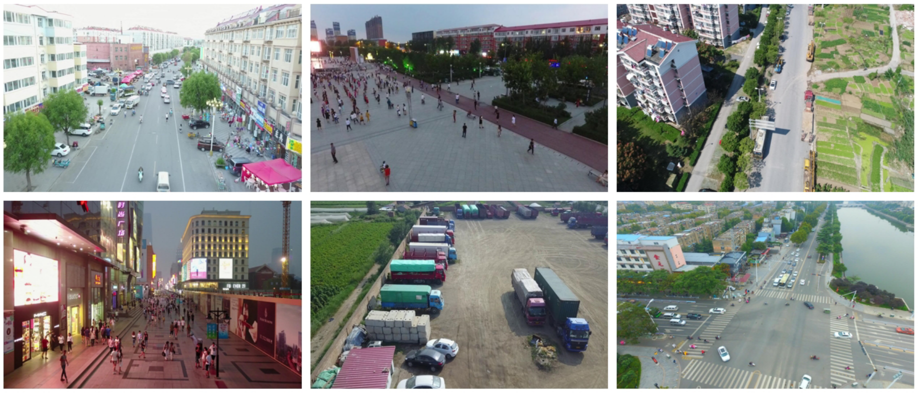

Fig. 1. The distribution of objects in images captured by drones. The first, second, and third columns examples respectively illustrate the dense distribution of objects, significant variations in object sizes, and complex backgrounds captured on drone-captured images.

Since the integration of Convolutional Neural Networks (CNN) into object detection in 2014, there has been remarkable progress in this field. Nevertheless, the majority of preceding deep convolutional neural networks were tailored for natural scene images. The detection of small objects is an inevitable challenge in drone scene detection tasks and has consistently posed a difficulty in object detection missions. This is primarily due to the fact that small objects suffer from (1) insufficient image resolution, (2) limited feature information, and (3) a small proportion of the overall image, making their detection more challenging compared to conventional objects. Furthermore, the close clustering and considerable size variations of small objects when viewed from a high altitude, combined with a wide field of vision and intricate geographical factors, lead to the loss of detailed information and insufficient feature extraction. Consequently, this lowers the accuracy of detection, placing greater demands on object detection technology. Some examples in Figure 1 also intuitively illustrate this issue: the first column displays densely distributed crowds and vehicles in street scenes; the second column shows significant variations in object sizes, even within the same image, with instances of objects appearing larger or smaller depending on their distance from the viewer; the third column demonstrates the wide aerial perspective of the drone, capturing backgrounds that include lakes, roads, houses, and more.

Addressing the aforementioned issues, this paper introduces an improved small object detection model, HTH-YOLOv5, based on YOLOv5 for drone scenarios. In the head section, we first propose a Hybrid Transformer Head (HTH) as the detection module to enhance attention on small object regions, achieving more efficient and accurate prediction capabilities. HTH-YOLOv5 comprises four detection heads designed for detecting micro, small, medium, and large objects, respectively. Subsequently, we incorporate the Convolutional Block Attention Module (CBAM[27]) into YOLOv5, embedding the CBAM module after each convolutional feature extraction in the backbone network and within the feature pyramid network[15]. Leveraging CBAM, we devise CA-FF to enhance the feature pyramid structure, improving its adaptability to small objects. Furthermore, we propose CAH-Transformer to enhance the feature fusion capability at the end of the backbone network. Compared to YOLOv5s, our improved HTH-YOLOv5 demonstrates superior performance in handling images captured by drones.

The main contributions of this paper are as follows:

1) We propose a Hybrid Transformer Head (HTH) and integrate it into YOLOv5 to capture global and local information.
2) We integrate CBAM into YOLOv5, which is a lightweight and efficient module that can generate attention graphs sequentially along channel and spatial directions.
3) We propose a Convolutional Attention Feature Fusion (CA-FF) module, which can improve the ability of the model to extract features from small objects.
4) We propose the CAH-Transformer module to help further focus the effective feature areas.

The structure of the remaining sections of this paper is as follows: In the second section, we briefly introduce several related works. The third section provides a detailed description of our designed HTH-YOLOv5. The fourth section presents the experimental results, and the fifth section concludes the paper.

2 Related Work

2.1 Object Detection

The current mainstream object detection algorithms can be divided into two types: one-stage detector and two-stage detector. The two-stage algorithm requires initially proposing regions of interest(ROI[9]) through selective search method or RPN (Region Proposal Network[9]). These regions represent coarse estimations of where objects might be located, and then a CNN is employed for classification and fine-grained boundary regression. Representative algorithms include: R-CNN[9], SPP-Net[11], Fast R-CNN[8], Faster R-CNN[22], feature pyramid networks (FPN)[15], and Mask R-CNN[10]. The development of two-stage object detection algorithms has been rapid, and detection accuracy continues to improve. However, the inherent architectural limitations pose constraints on detection speed, preventing it from meeting the real-time detection requirements in drone scenarios.

The main difference between one-stage object detection algorithms and two-stage object detection algorithms lies in the fact that the former lacks a candidate region proposal stage, making the training process relatively simpler. One-stage algorithms treat classification and localization as regression problems, generating detection results directly through a single network. This end-to-end detection approach achieves high accuracy and a detection speed of up to 45 frames per second (FPS), meeting the basic requirements for drone scene detection. Representative algorithms include YOLO (You Only Look Once)[19], SSD (Single Shot MultiBox Detector)[26], and RetinaNet[16].

2.2 Small Object Detection

In the context of small object detection research, literature [30] proposed a cascaded sparse query structure to accelerate small object detection in high-resolution images, achieving a 1.0 increase in mAP on the COCO dataset, with an average improvement in high-resolution inference speed by a factor of 3.0. The literature[35] builds upon YOLOv5 by introducing the Transformer Prediction Head (TPH) to replace the original prediction head, proposing the TPH-YOLO model. Additionally, a self-trained classifier is employed to enhance the classification capability for certain confusing categories. The literature[25] conducts pruning experiments on the basis of SSD for model compression, while simultaneously improving feature fusion methods to obtain more beneficial information for detecting small objects. The literature[12] developed a novel lightweight small object segmentation network by integrating various convolutional modules. This approach, combined with clustering algorithms and object feature adjustment strategies, achieved a reduction in parameter count and an improvement in detection accuracy.

Based on the characteristics of drone images and the research difficulties at this stage, literature[34] proposes a Dense Cropping and Local Attention Object Detector Network (DCLANet) specifically designed for small objects in drone scenarios. This approach enhances the network's focus on small objects by incorporating dense cropping and local attention mechanisms.The literature[14] introduced a bidirectional feature pyramid (BiFPN) to enhance the feature extraction ability of small targets in the image, by adding a small target detection layer based on YOLOv5 and fusing feature information from different scales. In order to address the issue of semantic loss during the detection of small targets, literature[7] incorporated the convolutional block attention module (CBAM) into YOLOv5 and also introduced a small target detection layer. To retain more feature information of small targets, literature[17] incorporated the efficient channel attention (ECA) module into the backbone network of YOLOv5l and replaced the sampling method with transposed convolution. However, the enhanced model based on YOLOv5l possesses a considerable number of parameters, which poses challenges for deploying it on edge devices like UAVs. Literature[29] proposes the magnifying glass method for image preprocessing to increase the feature information that can be used for learning. Literature[31] proposes a fast and accurate real-time small target detection system based on a two-stage architecture, which combines traditional algorithms and deep learning algorithms. Literature[23] proposes a new method based on graph neural network (GNN) to refine the detection results generated by the target detector. However, due to the low confidence score, some real predictions are easy to be selected as negative samples. In the literature[3], based on YOLOv7, a large convolutional kernel architecture is used to design the backbone network of the model in order to expand the effective sensitivity field of the convolutional model, but the additional computation caused by the large kernel architecture still needs to be further reduced.

From the above studies, it can be observed that deep learning holds significant research value in small object detection and has achieved notable results. However, further research is warranted to address small object detection in the context of drone scenarios, adapting solutions for more practical application scenarios.

2.3 YOLOv5

The YOLO series, representing a typical example of single-stage object detection algorithms, includes YOLOv1 through YOLOv8. YOLOv3[21], an improvement upon YOLOv1[19] and YOLOv2[20], replaced the base classification network with Darknet-53, leading to a significant improvement in inference speed compared to R-CNN[9] and Fast R-CNN[8].

Combining various improvements, Bochkovskiy et al. proposed YOLOv4[1], which can be trained on a regular GPU (1080Ti), meeting real-time requirements and deployable in production environments. YOLOv5 incorporates the advantages and addresses the drawbacks of previous versions, further enhancing both detection accuracy and speed. Meituan's Visual AI Department introduced YOLOv6[5], a detector without anchor points. Wang et al. introduced YOLOv7[4], featuring the E-ELAN and MPConv structures, achieving speeds and accuracy surpassing all known object detectors within the range of 5 FPS to 160 FPS. Subsequently, Alibaba Group released DAMO-YOLO[28], with the best model achieving 50.0% AP at 233 FPS on NVIDIA V100. This year, Ultralytics released YOLOv8, an anchor-free model that accelerates the speed of Non-Maximum Suppression (NMS).

YOLOv5 comprises three parts: the backbone, neck, and head, as illustrated in Figure 2. The backbone is primarily responsible for extracting features from input images, the neck handles multi-scale feature fusion on the feature maps, and transmits these feature details to the head. The head receives features from the neck and performs regression predictions. YOLOv5 has four versions: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. These versions share a consistent structure but correspond to different network widths and depths. Among them, YOLOv5s has the smallest network parameters, the fastest speed, and the lowest AP accuracy.

To validate the effectiveness of the algorithm in terms of both speed and accuracy and to meet the requirements for real-time detection on drone and deployment on mobile devices, YOLOv5s was chosen as the baseline model for improvement in this paper.

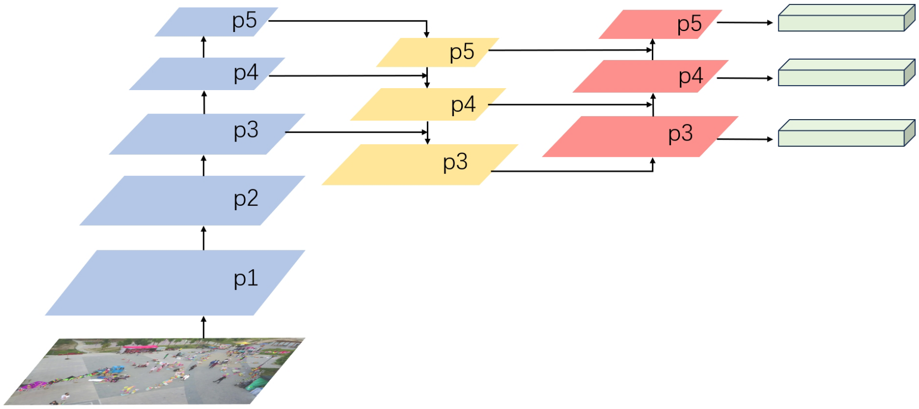

Fig. 2. The YOLOv5 architecture consists of three components: the backbone, neck, and head.

3 HTH-YOLOv5

The structure diagram of the improved model HTH-YOLOv5 proposed in this paper is shown in Figure 3, and the improvement measures of specific modules are introduced in the following sections to make the model more suitable for small objects and drone scenarios.

3.1 Hybrid Transformer Head

In the small object detection task of drone scenarios, the complex background is easy to block the small object, which interferes with the model's understanding of effective object and background. In recent years, models based on Self-Attention structures have gained quite good performance in the field of computer vision. The Self-Attention structure adopts the weighted average operation based on the input feature context, and the similarity function is used to dynamically calculate the attention weight between the relevant pixel pairs, so that the attention module can self-adapt to pay attention to different regions in the global and capture more effective features. This weight distribution allows the model to focus more on the effective object rather than the irrelevant background, so it is suitable for capturing the features of the effective object in the complex background. The Self-Attention structure calculates the self-attention weight as follows:

$$Z = Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_K}})V \tag{1}$$

Where, Z denotes the self-attention weight, QK^T describes the calculation of the correlation degree between each image block and other image blocks, $\sqrt{d_K}$ describes the scaling factor, the weight coefficient is normalized by softmax, and

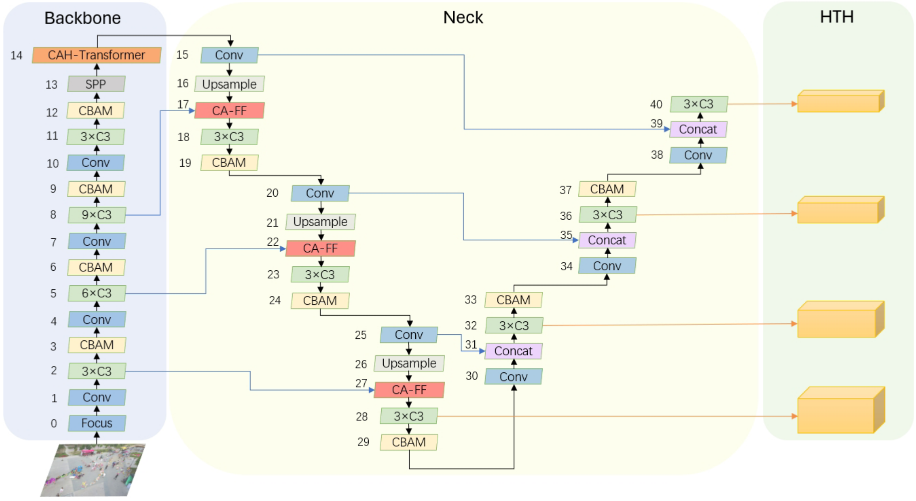

Fig. 3. The architecture of the HTH-YOLOv5. The number of each block is marked with a number on the left side of the block.

finally the weight coefficient and V are weighted and summed to obtain the self-attention weight matrix of each image block.

In addition, Q, K and V are three matrices with dimensions d_Q, d_K and d_V respectively (generally set $d_Q = d_K = d_V$), which can be calculated by multiplying the input sequence X by three random initialization matrices W^Q, W^K and W^V respectively:

$$Q = XW^Q, K = XW^K, V = XW^V \tag{2}$$

Inspired by high efficiency hybrid transformers[24], this paper proposes a Hybrid Transformer Head (HTH) based on Self-Attention for detection. The Hybrid Transformer module is divided into two sub-layers. The first layer captures the global context with multi-head attention block, introduces the convolutional layer to extract the local context, and then aggregates the global and local context to obtain a stronger feature representation. The second layer is a feedforward neural network, which is mainly composed of a multi-layer perceptron (MLP). LayerNorm is applied before each sublayer and DropPath is applied after each sublayer. The comparison of the structure of the standard Transformer module and the Hybrid Transformer module is shown in Figure 4.

The main module of the hybrid Transformer module is the global-local attention structure, which is a hybrid structure that uses linear multi-head self-attention to capture the global context and convolutional layers to extract the local context. Finally, an addition operation is applied to the global context and local context to extract the global-local context. The details are shown in Figure 5.

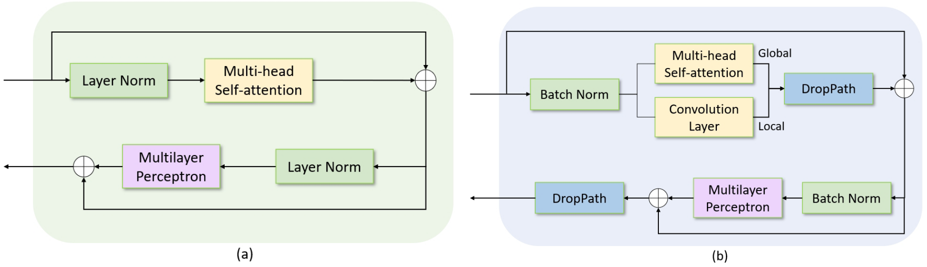

Fig. 4. (a) Standard Transformer module and (b) Hybrid Transformer module

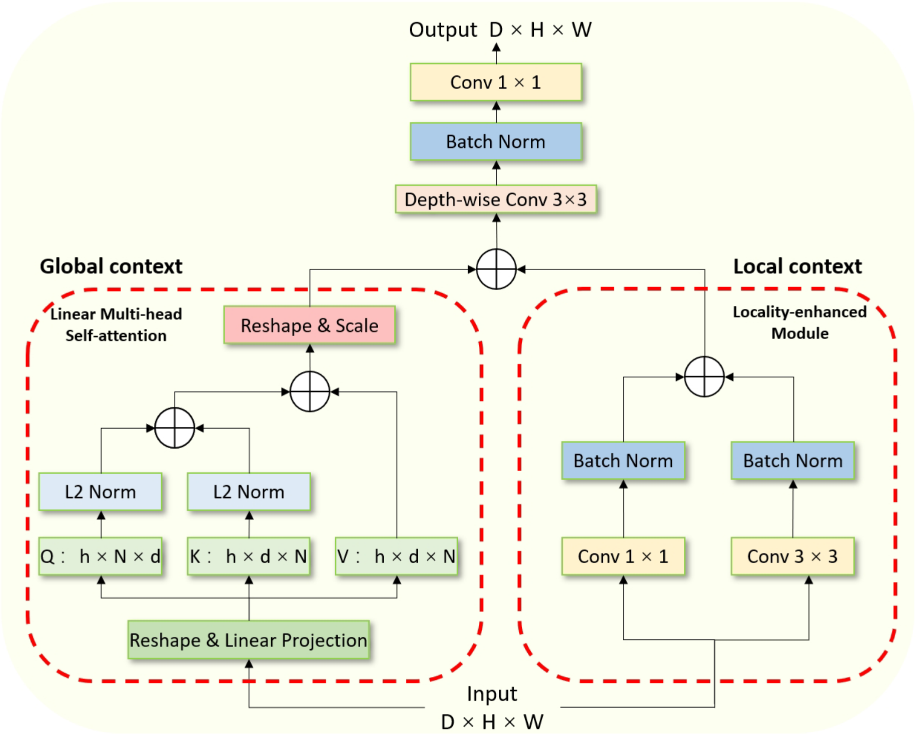

Fig. 5. Global-local attention structure of the Hybrid Transformer module

Where, H and W represent the resolution of the feature map, D and h represent the number of channels and the number of heads, respectively.

$$N = H \times W \tag{3}$$

$$d = D/h \tag{4}$$

In the global-local attention structure, the global structure uses a linear multi-head self-attention mechanism to improve efficiency and enhance the ability of sequence modeling, and the local enhancement module uses two parallel convolution layers and then performs batch normalization operations to extract the local context. Further deep convolution, batch normalization operations, and 1×1 convolution are performed on the generated global-local context to enhance generalization.

3.2 Convolutional Attention Feature Fusion Module

In order to extract and fuse the features of small projects effectively, a Convolutional Attention Feature Fusion (CA-FF) module based on CBAM[27] is designed from the perspective of feature fusion.

CBAM is a simple and effective attention module that is trained end-to-end and can be integrated into most CNN architectures. CBAM consists of two main modules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). CAM pays more attention to semantic features. For feature map Y, whose input size is $H \times W \times C$, CAM will use average pooling to aggregate spatial information and maximum pooling to obtain more detailed object feature information. By combining these two pooling methods, CAM can reduce the computation of feature maps and improve the expression of the network. The two one-dimensional vectors obtained after pooling are calculated at the fully connected layer, and 1×1 convolution kernel is used when the weights are shared between the eigenvectors. The process of generating channel attention Z_c is:

$$Z_c = sigmoid(MLP(AvgPool(Y)) + MLP(MaxPool(Y))) \tag{5}$$

SAM pays more attention to the location of features in the feature map, that is, the region with many effective features. By means of average pooling and maximum pooling, SAM compresses the feature map Y_c in channel dimension, and then obtains two two-dimensional feature maps. Then these two two-dimensional feature maps are concat together to get a feature map with two channels. Finally, a hidden layer containing a single convolution kernel is used to convolve the feature graph, and the process of generating the spatial attention weight Z_s through sigmoid operation is as follows:

$$Z_s = sigmoid(conv(AvgPool(Y), MaxPool(Y))) \tag{6}$$

When given a feature map, CBAM can independently infer attention maps along both channel and spatial dimensions. Subsequently, it refines features adaptively by multiplying the attention map with the input feature map. According to experiments in the paper [27], integrating CBAM into various models on different classification and detection datasets significantly improves the performance, demonstrating the effectiveness of this module.

The core idea of the CA-FF module proposed in this paper is to add attention mechanism on the basis of the feature fusion structure, and carry out feature refinement from the two dimensions of channel and space, so as to improve the feature extraction ability of the model. The structure diagram of CA-FF module is shown in Figure 6, and the contents in the dotted box are the original feature fusion structure. The feature fusion process of CA-FF module for feature maps of different scales can be expressed as:

$$Z = CBAM(X \oplus Y) \otimes X + (1 - CBAM(X \oplus Y)) \otimes Y \tag{7}$$

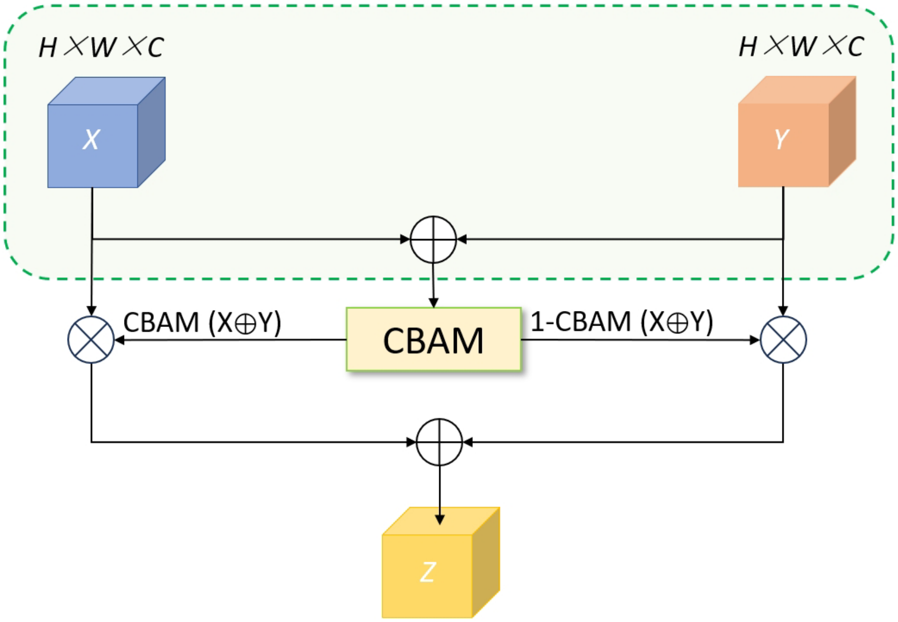

Fig. 6. Structure of the CA-FF module. X and Y are the feature maps before processing. Z is the feature map after processing.

Where, Z denotes the feature map after feature fusion processing; X denotes the low-level high-resolution feature map in the feature pyramid; Y denotes the feature map that is obtained by up-sampling the high-level, high-semantic feature map; $CBAM(X \oplus Y)$ represents the attention weight matrix obtained from the CBAM module after performing an element-wise sum of X and Y.

In this paper, CA-FF module is used to replace Concat module in YOLOv5 feature pyramid network, and the replaced structure diagram is shown in Figure 7.

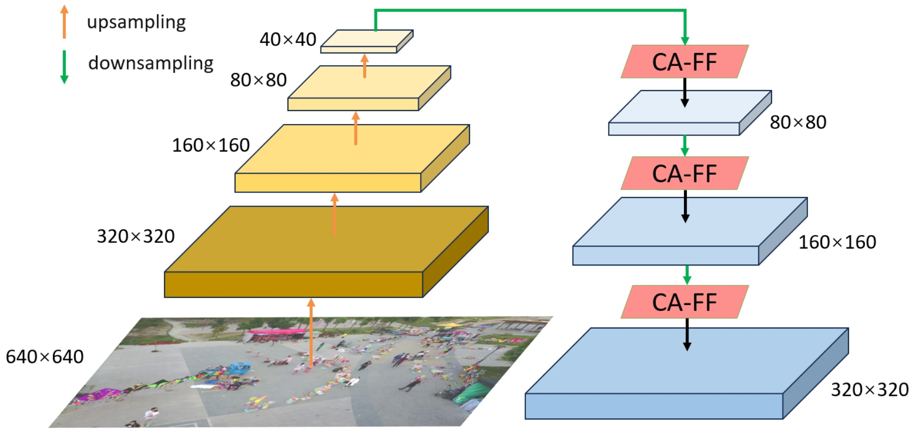

Fig. 7. FPN structure after replacing CA-FF. The addition of CA-FF module can better integrate the small object features in the feature map after upper and lower sampling.

The down-sampling operation of the backbone network will reduce the resolution of the feature map and lose a large number of small object features, while the upsampling can not bring more feature information. The modified feature fusion module is designed to more effectively integrate features of small objects within the feature map following both upsampling and downsampling processes, thereby minimizing the loss of small object features during the fusion process.

3.3 CAH-Transformer Module

This paper introduces further improvements to the Hybrid Transformer module by replacing the Hybrid Transformer's residual connection feature fusion module with a CA-FF module. Figure 8 illustrates the structure of the proposed CAH-Transformer module. By inserting the CAH-Transformer module at the end of the YOLOv5 backbone network, it further enhances the network's feature fusion capabilities across both channel and spatial dimensions.

4 Experiment

4.1 Data Sets and Evaluation Metrics

The model proposed in this paper is implemented in Pytorch 1.8.1, and the epoch is trained 300 times on NVIDIA RTX 3080Ti GPU with an initial learning rate

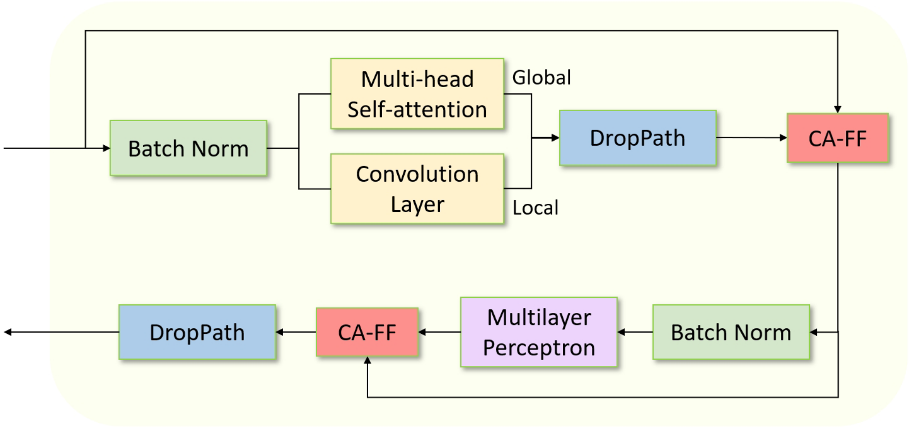

Fig. 8. Structure of the CAH-Transformer module. Compared with the Hybrid Transformer module, two feature fusion modules are replaced with CA-FF.

of 0.01. The experiment was conducted on the VisDrone2021 dataset and COCO dataset.

VisDrone2021 dataset was collected by the AISKYEYE team at the Machine Learning and Data Mining Laboratory of Tianjin University, and the baseline dataset included 288 video clips consisting of 261,908 frames and 10,209 still images. The dataset was collected using different drones in different scenarios, weather and lighting conditions, and included 10 types of images including pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus and motor. Figure 9 (a) shows the number of labels for each category. The horizontal and vertical coordinates in Figure 9 (a) represent the number of label instances and label categories, respectively. The horizontal and vertical coordinates in Figure 9 (b) respectively represent the width and height of the label box. The lower left corner of the figure has a high aggregation degree, indicating that the data set contains a high content of small objects, which can fully represent the general situation of object size in the drone capture scene.

MS COCO(Microsoft common objects in context) is one of the most authoritative and high-profile competitions in the field of machine vision. The dataset, which is mostly taken from complex everyday scenes, contains more than 330,000 images covering 80 different target categories, including people, animals, vehicles, food, furniture and more. COCO datasets are widely used in computer vision research and algorithm evaluation, providing an important benchmark for tasks such as object detection, segmentation, and key point detection.

To validate the performance of the proposed improved algorithm, this study employs mAP_{50}, mAP_{75}, $mAP_{50:95}$, Params, GFLOPs and Frames Per Second (FPS) as evaluation metrics for model performance. mAP_{50} and mAP_{75} represent the average detection accuracy of all object categories at IoU thresholds of 0.5 and 0.75, respectively. mAP_{50} reflects the algorithm's comprehensive de-

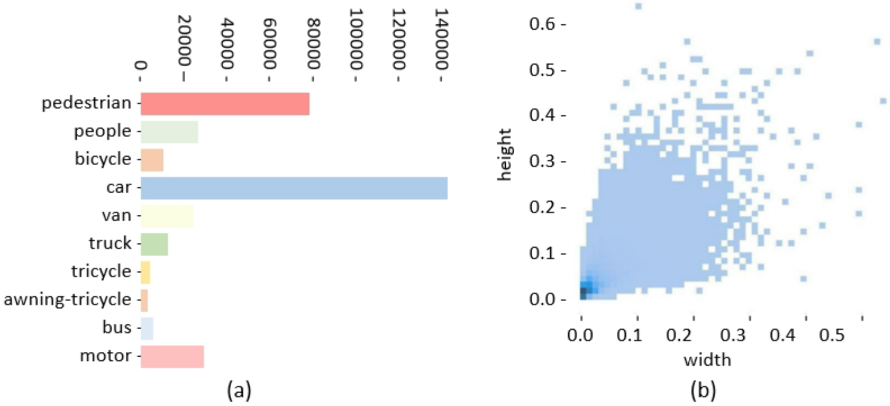

Fig. 9. VisDrone2021 data set (a) distribution of the number of labels in different categories (b) distribution of label sizes

tection capability for different object categories, while mAP_{75} demonstrates the algorithm's ability in bounding box regression. $mAP_{50:95}$ calculates the average accuracy for all IoU thresholds from 0.5 to 0.95 with a step size of 0.05. FPS stands for Frames Per Second, representing the number of frames that the algorithm can detect per second. It reflects the detection speed or real-time capability of the algorithm. Since the images captured in drone scenarios often have high resolutions, and the detection speed decreases with higher resolutions, the FPS measurements in this paper are conducted at a high resolution of 1504 × 1504.

4.2 Comparison with Existing Methods

To validate the superiority of the improved object detection algorithm proposed in this paper compared to other algorithms, we conducted comparative experiments with various advanced object detection algorithms, and the specific results are presented in Table 1. First, we compare with some classical object detection algorithms, then with YOLOv3, YOLOv4 and YOLOv5 models, and finally with the latest small object detection models to verify the progressiveness of our proposed method. Conclusions drawn from the data in Table 1 indicate that our proposed algorithm exhibits excellent performance in object detection accuracy, with mAP50 surpassing YOLOv4 by 14.1%, reaching 57.1%, and surpassing YOLOv5 by 12.4%. Moreover, our accuracy surpasses that of the most recent papers [18] and [6]. Furthermore, in terms of detection speed, our algorithm achieves an FPS_{1504} of 45, which is twice that of Faster-RCNN, 45.2% higher than YOLOv3, and 28.6% higher than both YOLOv4 and YOLOv5, only slightly lower than the performance reported in paper [6]. This suggests that the algorithm proposed in this paper not only demonstrates a significant improvement in detection accuracy but also has a good performance in real-time, making it more suitable for object detection tasks in drone capture scenarios.

Table 1. Comparison experiments of different object detection algorithms.

Methods	mAP_{50}(%)	mAP_{75}(%)	$mAP_{50:95}$(%)	FPS_{1504}
RetinaNet	28.7	11.6	11.8	–
RetfineDet[33]	28.8	14.1	14.9	–
Cascade-RCNN[2]	31.9	15.6	16.1	–
FPN	32.2	14.9	16.5	–
Light-RCNN[32]	32.8	15.1	16.5	–
Faster-RCNN	33.2	15.2	17.0	15
CornerNet[13]	34.1	15.9	17.4	33
YOLOv3	41.7	22.9	24.5	31
YOLOv4	43.0	25.2	24.9	35
YOLOv5	44.7	26.8	26.4	35
paper[18]	52.2	–	32.4	–
paper[6]	54.5	33.1	32.0	46
HTH-YOLOv5	**57.1**	**35.3**	**34.7**	45

4.3 Ablation experiment

In order to verify the effectiveness of HTH, CA-FF and CAH-Transformer modules proposed in this paper, ablation experiments are conducted to evaluate the influence of different modules on the performance of object detection algorithms under the same experimental conditions. The results of ablation experiments are shown in Table 2.

Table 2. The ablation experiments of HTH, CA-FF and CAH-Transformer modules proposed in this paper are carried out, and the original Transformer, CBAM and $CBAM(X \oplus Y)$ are also included.

Model	Methods						mAP_{50}(%)	mAP_{75}(%)	$mAP_{50:95}$(%)	Params(M)	GFLOPs
	HTH	CA-FF	CAH-Transformer	CBAM	Transformer	$CBAM(X \oplus Y)$					
A							33.0	14.8	16.5	7.0371	15.9
B	✓						36.6	17.4	18.4	8.4112	19.0
C		✓					34.5	16.2	17.7	7.4098	17.0
D			✓				36.1	17.1	18.2	8.4464	19.1
E					✓		35.6	16.8	17.9	8.3998	19.2
F						✓	33.7	15.8	17.3	7.2230	16.4
G	✓	✓					38.3	18.6	19.5	8.7831	19.8
H		✓	✓				37.7	18.2	19.4	8.8184	20.0
I	✓		✓				39.8	19.6	20.7	9.8205	21.4
J	✓	✓	✓				41.7	20.3	21.4	10.1925	22.2
K	✓	✓	✓	✓			42.3	20.7	21.7	10.7502	23.3

In the ablation experiment, Ultralytics 5.0 version of YOLOv5s was selected as the benchmark model. The input image resolution was 640 × 640. After training for 300 epochs, the results were shown in model A. Model B uses the HTH module in the detection head, which introduces some computation, but the mAP_{50}, mAP_{75} and $mAP_{50:95}$ are respectively 3.6%, 2.6% and 1.9% higher than the baseline of YOLOv5s, indicating that the HTH proposed in this paper can be used as the detection head to better improve the detection effect. The CA-FF module proposed in this paper is added to the neck network in model C. Compared with model A, mAP_{50}, mAP_{75} and $mAP_{50:95}$ are 1.5%, 1.4% and 1.2% higher than the baseline of YOLOv5s, respectively, reflecting the superiority of CA-FF module in feature fusion. After introducing the CAH-Transformer module into the model D backbone, the detection accuracy significantly improved, showing a 3.1% increase compared to YOLOv5s, which demonstrates the effectiveness of the CAH-Transformer. However, the parameter count increased by 1.4093 million and GFLOPs increased by 3.2. The analysis indicates that the CAH-Transformer itself requires a large amount of computing resources to calculate the correlation weights among each pixel in every feature map. Additionally, the detection speed exhibits a clear negative correlation with the number of parameters and the computational complexity. Therefore, considering the practical application scenario, this paper focuses on reducing computational costs, improving training efficiency, and enhancing the accuracy and speed of the model detection. To achieve these goals, the proposed approach only integrates the module at the end of the backbone network. To demonstrate the superiority of our proposed CAH-Transformer, we designed the model E. Model E differs from model D only in that the attention module at the end of the trunk uses the original Transformer. It can be seen from the data that the mAP_{50} of model D is 36.1%, 0.5% higher than the 35.6% of model E, which can prove that our proposed CAH-Transformer is more helpful to model detection effect in context acquisition ability. Compared to model C, the feature fusion module in model F is simply replaced by $CBAM(X \oplus Y)$, resulting in a significant decrease in accuracy. This further demonstrates the excellent feature fusion capability of the CA-FF module proposed in this paper. We combine the proposed modules with each other to test the effectiveness of the modules proposed in this paper. Model G uses HTH and CA-FF. Compared to models B and C, which solely utilize individual modules, model G exhibits significant improvements in mAP_{50}, mAP_{75} and $mAP_{50:95}$. This enhancement more effectively demonstrates the efficacy of embedding HTH and CA-FF modules within the overall network. Model H is embedded with CA-FF and CAH-Transformer in the network, and the effect is further improved compared with models C and D, indicating that the fusion of these two modules is better than the single use. Model I uses both HTH module and CAH-Transformer and obtains 39.8% mAP_{50}, which proves the excellent effect of the combination of the two modules. Then, we tested the model with simultaneous use of the three modules, and the results, as demonstrated by model J, surpassed those of all previously mentioned models. This suggests that integrating the three modules proposed in this paper yields superior detection performance. Finally, we added some CBAM modules to the backbone and neck networks. As a lightweight and effective attention module, after embedding CBAM, the effect of the model was further improved. In the final model K, the mAP_{50}, mAP_{75} and $mAP_{50:95}$ of the network size s were 42.3%, 20.7% and 21.7%, respectively. Improvements over YOLOv5s baseline were 9.3%, 5.9% and 5.2%, respectively.

4.4 Experimental analysis of COCO dataset

In order to further verify the performance of the model proposed in this paper, we conducted experiments on the COCO dataset, and compared the experimental results with those of YOLOv5s, SSD, YOLOv4-Tiny, YOLOX-Tiny, YOLOv6-N, and YOLOv7-Tiny. Detailed results are shown in Table 3. Although our algorithm introduces a certain amount of parameters and computation, it still enables it to meet most lightweight target detection tasks and mobile deployment requirements. The experimental results show that at $mAP_{50:95}$, our algorithm is obviously superior to other target detection algorithms, which indicates that our method has certain advantages in performance.

Table 3. Experimental analysis of COCO dataset.

Methods	Params(M)	GFLOPs	$mAP_{50:95}$(%)
YOLOv5s(2020)	7.2	16.5	37.2
SSD	36.1	–	25.1
YOLOv4-Tiny(2022)	6.1	–	21.7
YOLOX-Tiny(2021)	6.5	5.1	32.8
YOLOv6-N(2022)	4.3	11.7	35.9
YOLOv7-Tiny(2022)	6.2	13.7	37.4
HTH-YOLOv5	10.8	23.3	**37.5**

4.5 Algorithm Effectiveness Analysis

In order to directly reflect the detection effect of the improved algorithm in the actual scene, this paper uses four representative pictures in the VisDrone2021 test set for testing, and makes visual comparison with the test results of YOLOv5. As shown in Figure 10, the first row is the test result of HTH-YOLOv5 in this paper, and the second row is the test result of YOLOv5. These four pictures correspond to different detection difficulties. The first column shows the scene with large changes in light. There are a large number of pedestrians in both bright and dim areas, accompanied by partial occlusion. In this figure, (a1) shows that the algorithm in this paper is less affected by light and can accurately identify pedestrians and some bicycles in distant dim areas, while (a2) has many defects in the detection of pedestrians in the upper part and many missed detection of pedestrians in the right side. The second column shows the overhead shooting perspective. There are some vehicles and many pedestrians on the way, and the pixels occupied by pedestrians are very small. It can be seen from (b1) that for pedestrians, the algorithm in this paper can almost recognize them, which shows that the algorithm in this paper has outstanding detection ability for small objects. Compared with (b1), (b2) has lost many detection boxes for pedestrians. The third column is a high-altitude image of the street scene, with a large number of vehicles of different models distributed on the street and a large number of trees to shield it. Compared with (c2), (c1) has better anti-occlusion ability and can accurately identify pedestrians riding in the middle and rear, indicating that the algorithm in this paper can better handle occlusion and small object scenes. The fourth column is a blurry and distorted image, possibly caused by the shaking of the drone. For this graph, (d1) can still perform well on fuzzy objects in the graph, indicating that the algorithm in this paper has certain robustness and can better cope with actual scenes. In general, although the algorithm introduced a certain amount of computation, FPS can be maintained at 45 to meet the real-time needs. In addition, it can be seen from the detection effect diagram that the proposed algorithm has a good performance in the drone capture scenario, and the increased calculation amount is also worthwhile. Therefore, the algorithm in this paper is more suitable for the application and deployment of drones in practical scenarios.

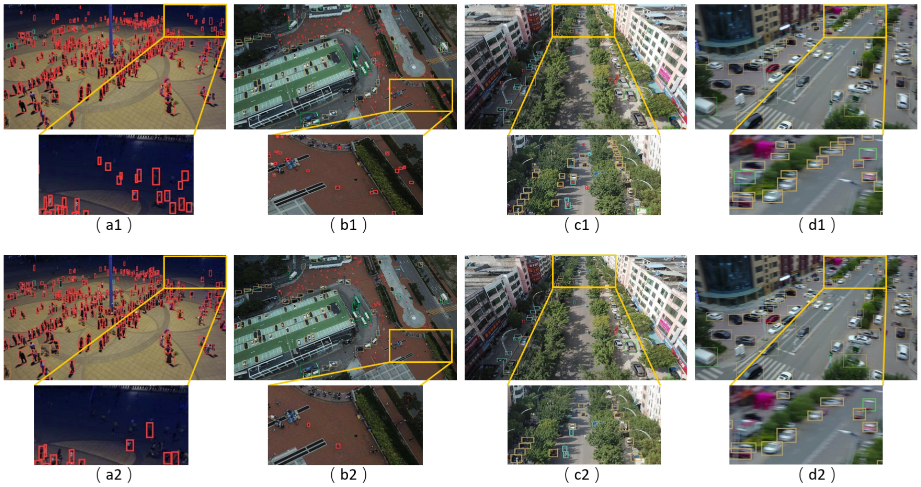

Fig. 10. Comparison of detection results of HTH-YOLOv5 and YOLOv5 on VisDrone2021.

5 Conclusion

This paper presents an enhanced approach based on YOLOv5 to boost the accuracy of detecting small objects within drone-captured scenarios. We propose the HTH detection head in YOLOv5 based on a Hybrid Transformer to enhance focus on small objects. Subsequently, we introduce the CBAM module and propose the Convolutional Attention Feature Fusion module (CA-FF) based on it to further improve feature fusion efficiency. Finally, we use CA-FF to enhance the structure of the Hybrid Transformer in the backbone, enabling better capture of global and contextual information. The effectiveness and real-time performance of these improvements are validated on the VisDrone2021 dataset. Experimental results demonstrate that HTH-YOLOv5, along with its modules, achieves a higher mAP for object detection in drone scenarios compared to the original YOLOv5s. The algorithm in this paper introduces a certain amount of computation, which can continue to carry out the research on the lightweight of the improved YOLOv5 network.

Acknowledgments. This work is supported by NSFC under grants 61860206007 and U19A2071, as well as the funding from Sichuan University under grant 2020SCUNG205.

References

1. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv:2004.10934 (2020). https://doi.org/10.48550/arxiv.2004.10934
2. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6154–6162 (2018). https://doi.org/10.1109/CVPR.2018.00644
3. Chen, H., Wang, J., Li, J., Qiu, Y., Zhang, D.: Small object detection for drone image based on advanced yolov7. In: 2023 42nd Chinese Control Conference (CCC). pp. 7453–7458 (2023). https://doi.org/10.23919/CCC58697.2023.10239784
4. Chien Yao Wang, Alexey Bochkovskiy, H.Y.M.L.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv:2207.02696 (2022). https://doi.org/10.48550/arxiv.2207.02696
5. Chuyi Li, Lulu Li, H.J.K.W.Y.G.L.L.e.a.: Yolov6: A single-stage object detection framework for industrial applications. arXiv:2209.02976 (2022). https://doi.org/10.48550/arxiv.2209.02976
6. Feng, Z., Xie, Z., Bao, Z., Chen, K.: Real-time dense small object detection algorithm for uav based on improved yolov5. Hang kong xue bao **44**(7), 251 (2023)
7. Gao, T., Wushouer, M., Tuerhong, G.: Small object detection method based on improved yolov5. In: 2022 International Conference on Virtual Reality, Human-Computer Interaction and Artificial Intelligence (VRHCIAI). pp. 144–149 (2022). https://doi.org/10.1109/VRHCIAI57205.2022.00032
8. Girshick, R.: Fast r-cnn. In: International Conference on Computer Vision (2015). https://doi.org/10.1109/ICCV.2015.169
9. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Computer Society (2014). https://doi.org/10.48550/arxiv.1311.2524
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. IEEE Transactions on Pattern Analysis & Machine Intelligence (2017). https://doi.org/10.1109/ICCV.2017.322
11. Kaiming, He, Xiangyu, Zhang, Shaoqing, Ren, Jian, Sun: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence (2015). https://doi.org/10.1109/TPAMI.2015.2389824
12. Kou, R., Wang, C., Yu, Y., Peng, Z., Yang, M., Huang, F., et al.: Lw-irstnet: Lightweight infrared small target segmentation network and application deployment. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–13 (2023). https://doi.org/10.1109/TGRS.2023.3314586

13. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. International Journal of Computer Vision **128**(3), 642–656 (2020). https://doi.org/10.1007/s11263-019-01204-1
14. Li, S., Yang, X., Lin, X., Zhang, Y., Wu, J.: Real-time vehicle detection from uav aerial images based on improved yolov5. Sensors **23**(12) (2023). https://doi.org/10.3390/s23125634, https://www.mdpi.com/1424-8220/23/12/5634
15. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. arXiv e-prints (2016). https://doi.org/10.1109/CVPR.2017.106
16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2999–3007 (2017). https://doi.org/10.1109/ICCV.2017.324
17. Liu, S., Liang, P., Duan, Y., Zhang, Y., Feng, J.: Small target detection for unmanned aerial vehicle images based on yolov5l. In: 2022 10th International Conference on Information Systems and Computing Technology (ISCTech). pp. 210–214 (2022). https://doi.org/10.1109/ISCTech58360.2022.00042
18. Liu, S., Liang, P., Duan, Y., Zhang, Y., Feng, J.: Small target detection for unmanned aerial vehicle images based on yolov5l. In: 2022 10th International Conference on Information Systems and Computing Technology (ISCTech). pp. 210–214 (2022). https://doi.org/10.1109/ISCTech58360.2022.00042
19. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Computer Vision & Pattern Recognition (2016). https://doi.org/10.1109/cvpr.2016.91
20. Redmon, J., Farhadi, A.: Yolo9000: Better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6517–6525 (2017). https://doi.org/10.1109/CVPR.2017.690
21. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv e-prints (2018). https://doi.org/10.48550/arxiv.1804.02767
22. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence **39**(6), 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031
23. Tang, Z., Liu, Y., Shang, Y.: A new gnn-based object detection method for multiple small objects in aerial images. In: 2023 IEEE/ACIS 23rd International Conference on Computer and Information Science (ICIS). pp. 14–19 (2023). https://doi.org/10.1109/ICIS57766.2023.10210246
24. Wang, L., Fang, S., Zhang, C., Li, R., Duan, C.: Efficient hybrid transformer: Learning global-local context for urban sence segmentation. arXiv.2109.08937 (2021). https://doi.org/10.48550/arXiv.2109.08937
25. Wang, Q., Zhang, H., Hong, X., Zhou, Q.: Small object detection based on modified fssd and model compression. In: 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP). pp. 88–92 (2021). https://doi.org/10.1109/ICSIP52628.2021.9688896
26. Wei, L., Dragomir, A., Dumitru, E., Christian, S., Scott, R., Cheng-Yang, F., et al.: Ssd: Single shot multibox detector. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
27. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1

28. Xianzhe Xu, Yiqi Jiang, W.C.Y.H.Y.Z.X.S.: Damo-yolo : A report on real-time object detection design. arXiv:2211.15444 (2023). https://doi.org/10.48550/arxiv.2211.15444
29. Yan, X., Shen, B., Li, H.: Small objects detection method for uavs aerial image based on yolov5s. In: 2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT). pp. 61–66 (2023). https://doi.org/10.1109/ICEICT57916.2023.10245156
30. Yang, C., Huang, Z., Wang, N.: Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13658–13667 (2022). https://doi.org/10.1109/CVPR52688.2022.01330
31. Yu, M., Leung, H.: Small-object detection for uav-based images. In: 2023 IEEE International Systems Conference (SysCon). pp. 1–6 (2023). https://doi.org/10.1109/SysCon53073.2023.10131084
32. Zeming Li, Chao Peng, G.Y.X.Z.Y.D.J.S.: Light-head r-cnn: In defense of two-stage object detector. arXiv:1711.07264 (2017). https://doi.org/10.48550/arxiv.1711.07264
33. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Single-shot refinement neural network for object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018). https://doi.org/10.1109/CVPR.2018.00442
34. Zhang, X., Feng, Y., Zhang, S., Wang, N., Mei, S.: Finding nonrigid tiny person with densely cropped and local attention object detector networks in low-altitude aerial images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **15**, 4371–4385 (2022). https://doi.org/10.1109/JSTARS.2022.3175498
35. 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 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 2778–2788 (2021). https://doi.org/10.1109/ICCVW54120.2021.00312